

LONDON  
SCHOOL of  
HYGIENE  
& TROPICAL  
MEDICINE



LSHTM Research Online

Recker, M; Laabei, M; Toleman, MS; Reuter, S; Saunderson, RB; Blane, B; Török, ME; Ouadi, K; Stevens, E; Yokoyama, M; +7 more... Steventon, J; Thompson, L; Milne, G; Bayliss, S; Bacon, L; Peacock, SJ; Massey, RC; (2017) Clonal differences in *Staphylococcus aureus* bacteraemia-associated mortality. *Nature microbiology*. ISSN 2058-5276 DOI: <https://doi.org/10.1038/s41564-017-0001-x>

Downloaded from: <http://researchonline.lshtm.ac.uk/4258946/>

DOI: <https://doi.org/10.1038/s41564-017-0001-x>

**Usage Guidelines:**

Please refer to usage guidelines at <http://researchonline.lshtm.ac.uk/policies.html> or alternatively contact [researchonline@lshtm.ac.uk](mailto:researchonline@lshtm.ac.uk).

Available under license: <http://creativecommons.org/licenses/by-nc-nd/2.5/>

<https://researchonline.lshtm.ac.uk>

1 Clonal differences in *Staphylococcus aureus* bacteraemia-associated mortality.

2  
3

4 Mario Recker<sup>1\*</sup>, Maisem Laabei<sup>2\*</sup>, Michelle S. Toleman<sup>3</sup>, Sandra Reuter<sup>3</sup>, Rebecca B.

5 Saunderson<sup>3</sup>, Beth Blane<sup>3</sup>, M. Estee Török<sup>3</sup>, Khadija Ouadi<sup>2</sup>, Emily Stevens<sup>2</sup>, Maho

6 Yokoyama<sup>2</sup>, Joseph Steventon<sup>2</sup>, Luke Thompson<sup>2</sup>, Gregory Milne<sup>2</sup>, Sion Bayliss<sup>2</sup>, Leann

7 Bacon<sup>2</sup>, Sharon J. Peacock<sup>3,4</sup> and Ruth C. Massey<sup>2,5∅</sup>.

8

9 1: Centre for Mathematics & the Environment, University of Exeter, Penryn Campus, Penryn

10 TR10 9EZ, UK.

11 2: Dept. of Biology and Biochemistry and the Milner Centre for Evolution, University of Bath,

12 Bath, BA2 7AY, UK.

13 3: Department of Medicine, University of Cambridge, Cambridge, UK.

14 4: London School of Hygiene and Tropical Medicine, London, WC1E 7HT, UK.

15 5: School of Cellular and Molecular Medicine, University of Bristol, BS8 1TD, UK.

16 \* contributed equally to this work.

17 ∅ Corresponding author: [ruth.massey@bristol.ac.uk](mailto:ruth.massey@bristol.ac.uk)

18

19

20

21

22

23

24

25

26

27

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26

## Abstract

The bacterium *Staphylococcus aureus* is a major human pathogen, where the emergence of antibiotic resistance is a global public-health concern. Infection severity, and in particular bacteraemia-associated mortality, has been attributed to several host-related factors, such as age and the presence of co-morbidities. The role of the bacterium in infection severity is less well understood as it is complicated by the multi-faceted nature of bacterial virulence, which has so far prevented a robust mapping between genotype, phenotype and infection outcome. To investigate the role of bacterial factors in contributing to bacteraemia-associated mortality we phenotyped a collection of sequenced clinical *S. aureus* isolates from patients with bloodstream infections, representing two globally important clonal types (CC22 and CC30). By adopting a GWAS approach we identified and functionally verified several genetic loci that affect the expression of cytolytic toxicity and biofilm formation. By analysing the pooled data comprising bacterial genotype and phenotype together with clinical meta-data within a machine-learning framework we found significant clonal differences in the determinants most predictive of poor infection outcome. Whereas elevated cytolytic toxicity in combination with low levels of biofilm formation was predictive of an increased risk of mortality in infections by strains of a CC22 background, these virulence-specific factors had little influence on mortality rates associated with CC30 infections. Our results therefore suggest that different clones may have adopted different strategies to overcome host responses and cause severe pathology. Our study further demonstrates the use of a combined genomics and data analytic approach to enhance our understanding of bacterial pathogenesis at the individual level, which will be an important step towards personalised medicine and infectious disease management.

## 1 Introduction

2 *Staphylococcus aureus* bacteraemia (SAB) is a significant global health problem<sup>1</sup> and is  
3 exacerbated by the emergence and widespread circulation of drug resistant strains, such as  
4 methicillin-resistant *S. aureus* (MRSA)<sup>2</sup>. Mandatory surveillance of SAB has been implemented  
5 in several countries, with many reporting a decline in the incidence of methicillin-resistant  
6 SAB<sup>3-5</sup>. However, in the UK the incidence of methicillin-susceptible SAB has been increasing  
7 year on year, with an overall increase of more than 15% since reporting became mandatory  
8 in 2011/2012<sup>3</sup>. Furthermore, the 30 day (all-cause) mortality rate for SAB has not significantly  
9 changed over the last two decades and appears to have plateaued at approximately 20%<sup>6</sup>.  
10 This strongly suggests that existing infection control and treatment options are insufficient to  
11 tackle this important health problem and that a better understanding of the factors that  
12 contribute to bacteraemia-associated morbidity is crucially needed.

13

14 To date, many host risk factors have been identified for both the occurrence of and treatment  
15 failure following SAB<sup>6</sup>. However, the contribution of the bacterium is only partially  
16 understood and is largely informed by experimental animal studies. These model systems  
17 come with their own set of limitations, and many observations contrast with those from  
18 human studies. For example, whereas cytolytic toxins have previously been shown to enhance  
19 disease severity in animal models of SAB<sup>7,8</sup>, isolates from invasive diseases in humans, such  
20 as bacteraemia and pneumonia, were recently found to be significantly less toxic than those  
21 isolated from skin and soft tissue infections or even those of healthy volunteers<sup>9-12</sup>. This raises  
22 the question as to whether animal models are adequate to study bacterial virulence in human  
23 SAB infections, or whether there is an important distinction between the role of toxicity in  
24 the development of bacteraemia and its pathogenic effect once bacteraemia has been

1 established. Either way, human-based approaches are essential to close this gap in our  
2 knowledge.

3

4 Our understanding of the pathogenesis of SAB has been further limited by the fact that that  
5 most studies so far have focused on only a single or small number of factors in isolation. For  
6 example, several studies have found increased mortality rates associated with methicillin-  
7 resistant SAB compared to methicillin-susceptible SAB<sup>13,14</sup>. However, patients with co-  
8 morbidities are more likely to develop methicillin-resistant bacteraemia due to their impaired  
9 health and longer time spent in healthcare facilities compared to those without co-  
10 morbidities. When subsequent studies considered both methicillin resistance and the  
11 presence of co-morbidities together no difference in mortality was associated with the  
12 methicillin resistance status of the infecting bacterium<sup>15</sup>. This clearly illustrates that multiple  
13 host and bacterial factors, as well as their possible interactions, have to be taken into  
14 consideration when investigating the outcome of an individual infection.

15

16 We have previously demonstrated how genotype-phenotype mapping in *S. aureus* has the  
17 potential to provide sufficient information to enable predictions of the level of virulence  
18 expressed by the infecting microorganism<sup>16</sup>. Here, we have expanded this whole-genome  
19 approach to a set of clinical isolates from *S. aureus* bacteraemic patients, representing two  
20 different clonal backgrounds (CC22 and CC30), and determined a number of genetic loci that  
21 affect the bacteria's cytolytic toxicity and biofilm formation, both of which are important  
22 virulence factors. Using a predictive modelling framework on pooled data comprising genetic,  
23 phenotypic and clinical factors we identified stark differences in the bacterial and host factors  
24 most associated with patient death. Our results imply that the clonal background of the

1 infecting microorganism has a more important role in determining the outcome of an  
2 infection than previously recognised.

3

#### 4 **Material and Methods.**

5 **Strain and clinical metadata collection.** All isolates were collected from patients admitted to  
6 a single hospital with their first episode of SAB between 2006 and 2012, and were stored in  
7 glycerol at -80°C. Samples were collected during an observational cohort study of adults with  
8 SAB at Addenbrooke's Hospital, Cambridge, UK between 2006 and 2012. These isolates  
9 represented the genetic diversity present across the British Society for Antimicrobial  
10 Chemotherapy collection of sequenced CC22 and CC30 isolates collected from multiple sites  
11 across the UK and Ireland since 2008. Written informed consent was not required as the study  
12 was conducted as part of a service evaluation of the management of SAB. Ethical approval  
13 was obtained from the University of Cambridge Human Biology Research Ethics Committee  
14 (reference HBREC.2013.05) and the Cambridge University Hospitals NHS Foundation Trust  
15 Research and Development Department (reference A092869). Study definitions have been  
16 defined previously<sup>17</sup> and were used to determine the focus of the bacteraemia, classify the  
17 bacteraemia as community-acquired, hospital-acquired or healthcare-associated, and to  
18 report outcomes, including death at 30 days. The previously defined definitions<sup>17</sup> were placed  
19 into eight categories; intravascular catheter infection (SAB caused from a central or peripheral  
20 venous catheter), bone and joint infection (SAB due to a native joint infection), deep tissue  
21 abscess (infection due to lung infection, deep tissue abscesses, mycotic aneurysms and  
22 lesions in the central nervous system), infective endocarditis, infection due to a prosthetic  
23 device (including prosthetic joints, implanted vascular devices, prosthetic cardiac valves), skin

1 and soft tissue infection, disease of the urogenital tract, or an unknown focus of infection.

2 The presence of comorbidities was assessed using the Charlson comorbidity index (CCI) and  
3 were dichotomized into scores of  $<3$  or  $\geq 3$ <sup>18,19</sup>, as detailed in Supplementary Table 1.

4 **Genome Sequencing.** Bacterial DNA extraction was carried out on a QIAextractor (Qiagen), and  
5 library preparation was performed as previously described<sup>20</sup>. Index-tagged libraries were  
6 created, and 96 separated libraries were sequenced in each of eight channels using the  
7 Illumina HiSeq platform (Illumina) to generate 100-bp paired-end reads at the Wellcome Trust  
8 Sanger Institute, UK. Paired-end reads for these isolates were mapped to either the  
9 ST22/EMRSA15 reference strain, HO 5096 0412<sup>21</sup>, or the CC30/EMRSA16 reference strain  
10 MRSA252<sup>22</sup> and SNPs were identified as described previously<sup>21</sup>. The accession number for the  
11 sequence data for each of these isolates is listed in Supplementary Table 1.

12 **Cytotoxicity Assay.** Overnight *S. aureus* cultures were diluted 1:1000 into fresh tryptone soya  
13 broth and incubated for 18 h at 37°C with shaking at 180 rpm in 30 mL glass tubes. *S. aureus*  
14 supernatants were harvested from 18 h cultures by centrifugation at 14,600 rpm for 10 min.

15 The THP-1 human monocyte-macrophage cell line (ATCC#TIB-202, which was authenticated  
16 and if they were tested for mycoplasma) was routinely grown in suspension in 30 mL of of  
17 RPMI-1640, supplemented with 10% heat-inactivated fetal bovine serum (FBS), 1  $\mu$ M L-  
18 glutamine, 200 units/mL penicillin, and 0.1 mg/mL streptomycin at 37°C in a humidified  
19 incubator with 5% CO<sub>2</sub>. Cells were routinely viewed microscopically every 48–60 h and  
20 harvested by centrifugation at 1,000 rpm for 10 min at room temperature and re-suspended  
21 to a final density of 1–1.2 x 10<sup>6</sup> cells/mL in tissue-grade phosphate buffered saline. This  
22 procedure typically yielded >95% viability of cells as determined by trypan blue exclusion and  
23 easyCyte flow cytometry. To evaluate *S. aureus* toxicity, for the CC30 isolates undiluted  
24 supernatant was used, but due to the higher level of toxicity of the CC22 isolated the

1 supernatant was diluted to 30% of the original volume in TSB. The neat and diluted bacterial  
2 supernatants (20 $\mu$ L) were incubated with 20  $\mu$ L of washed THP-1 cells for 12 min at 37°C  
3 under static conditions. Cell death was quantified by easyCyte (Millipore) flow cytometry  
4 using the Guava Viability stain (Millipore) according to manufacturer's instructions, with the  
5 toxicity of each isolate quantified in triplicate and the mean of this presented (listed in  
6 Supplementary Table 1).

7 **Biofilm assay.** Biofilm formation was quantified using a 1:40 dilution from overnight cultures  
8 into 100  $\mu$ L of fresh tryptic soy broth supplemented with 0.5% sterile filtered glucose (TSBG)  
9 in 96-well polystyrene plate (Costar). Perimeter wells of the 96-well plate were filled with  
10 sterile H<sub>2</sub>O and plates were placed in a separate plastic container inside a 37°C incubator and  
11 grown for 24 h under static conditions. For the transposon mutants, erythromycin (5  $\mu$ g/mL)  
12 was added to the growth medium. Semi-quantitative measurements of biofilm formation on  
13 96-well polystyrene plates was determined based on the method of Ziebuhr et al<sup>23</sup>. Following  
14 24-h growth, plates were washed vigorously five times in PBS, dried and stained with 150  $\mu$ L  
15 of 1% crystal violet for 30 min at room temperature. Following five washes of PBS, wells were  
16 re-suspended in 200  $\mu$ L of 7% acetic acid, and optical density at 595 nm was recorded using a  
17 Fluorimeter plate reader (BMG Labtech). To control for day to day variability, for the clinical  
18 isolates a control strain (E-MRSA15) was included on each plate in triplicate, and absorbance  
19 values were normalised against this (listed in Additional Table 1). For the transposon mutants,  
20 as JE2 is the wild type strain this was used as the control strain, and the effect of mutating the  
21 loci made relative to this. For this experiment the assays were performed in triplicate on each  
22 plate and repeated four times.

23 **Genome wide association study (GWAS).** The genome of the reference strains HO 5096 0412  
24 and MRSA252 was split loci corresponding to coding region and intergenic regions, and loci



1 containing SNPs relative to the reference genome were identified. Annotated intergenic  
2 elements such as miscellaneous RNAs were considered separate loci. Synonymous SNPs in  
3 coding regions and SNPs in known mobile genetic elements and repeat regions were not  
4 considered. The resulting loci were named allele\_X where X refers to the position of the SNP  
5 at the 5' end of that block, relative to the origin of replication. Each of these alleles in each  
6 isolate was scored as 1 if it differed from the reference and 0 if it didn't. These allele scores  
7 for each isolate were used as the genotypic information for the following analysis. Significant  
8 associations between bacterial genotype and either phenotype (toxicity and biofilm  
9 formation) were identified by fitting an analysis of variance model (ANOVA) in R<sup>24</sup> and using  
10 a minor allele frequency cut-off of 5%. The reported *P* values are not corrected for multiple  
11 testing; the Bonferroni statistical significance threshold is instead provided in fig. 3.

12 **Predictive modelling.** We employed a *Random Forests*<sup>25</sup> machine learning approach, using  
13 the *randomForest* package in R<sup>26</sup>, to identify the most important and clone-specific  
14 determinants for host mortality based on the genotype, phenotype and clinical meta-data  
15 (see supplementary file S1\_RF\_analysis.pdf for full details). The input variables were thus  
16 taken from the pooled set of SNP data, relative cytolytic toxicity, relative biofilm formation  
17 and infection/patient-specific data for each isolate. Due to the genetic differences between  
18 CC22 and CC30, where only a small proportion of SNP positions overlapped, we carried out  
19 the analysis for both clones separately.

20 **Predictive accuracy.** To assess the models' predictive accuracies we employed two different  
21 measures: (i) the receiver operating characteristic (ROC) curve, which is generated by plotting  
22 the true positive rate against the false positive rate (i.e. the observed incidence against the  
23 false predicted incidence) at various threshold settings and where the area under the curve  
24 (AUC) is a measure of predictive accuracy, with an AUC=1 equating to zero error and an

1 AUC=0.5 equating to random guessing; and (ii) by means of a confusion matrix, which  
2 contrasts the instances of the predicted classes (*alive* or *death*) against the actually observed  
3 classes. The misclassification-rates reported here are based on the so-called out-of-bag  
4 errors<sup>26</sup>, which are derived by iteratively testing the models' performances against subsets of  
5 data left out during the fitting processes and thus provide a measure of how well the models  
6 would fare against unknown data.

7 **Feature selection.** Due to the high dimensionality of the combined dataset including  
8 genotype, phenotype and clinical data, where the number of features significantly  
9 outweighed the number of clinical samples, we performed a feature selection procedure  
10 using the VSURF<sup>27</sup> R package nested within a 10-fold cross-validation procedure. The VSURF  
11 algorithm selects features based on step-wise introduction of predictors in order of their  
12 importance ranking with regards to their contribution to model performance. At each fold a  
13 small number of features (in our case between 3 and 20) are selected, from which we then  
14 took only those that were selected more than once to form our final feature set. This  
15 procedure reduces the risk of overfitting and ensures that only those features are considered  
16 that are important across a wide range of isolates. The Supplementary Note provides a  
17 detailed overview of our analysis, including the feature selection procedure and parameter  
18 settings.

19 **Capsule dot-blot.** Bacterial were grown overnight in TSB at 37°C with shaking. 5µl of  
20 overnight culture was spotted onto nitrocellulose membranes and the membranes dried for  
21 10 minutes at 65°C. The membranes were washed three times in PBS and then once for 1hr  
22 at 37°C in PBS with 2mg/ml trypsin. Membranes were then blocked in 0.05% skim milk for  
23 1hr, and washed three times in PBS with 0.05% Tween (PBS-tween). Anti-Cap5 antiserum (a  
24 generous gift from Prof. Jean Lee, Harvard) was diluted 1:1000 in the PBS-tween buffer and

1 the membrane was incubated in this with gentle agitation for 1hr at room temperature. The  
2 membranes were washed three time in PBS-tween. The protein A-HRP conjugate was diluted  
3 1:1000 in PBS-tween and the membrane was incubated in this with gentle agitation for 1hr at  
4 room temperature. The membranes were washed and the reactivity of the antiserum  
5 detected using the colorimetric substrate 4-chloro-1-naphthol. The blots were scanned and  
6 density of capsule anti-serum reactivity quantified using the ImageJ software. Each strain was  
7 blotted three times and the average density of the three repeats for each isolate is presented.

8 **Opsinophagocytosis assays.** Polymorphonuclear leukocytes (PMNs) from healthy human  
9 volunteers were separated from heparinized blood by density centrifugation. The cells were  
10 washed and resuspended at a density of  $10^7$  cells per ml. The opsinophagocytosis killing  
11 assays were performed in  $5 \times 10^6$  PMNs, 10% (vol/vol) human serum and  $10^6$  cfu of each of  
12 the *S. aureus* isolates. This was incubated with gentle agitation for 2hr at 37°C followed by  
13 diluting and plating onto TSA plates to enumerate the live bacteria. The data presented is the  
14 % killed bacteria for each of the 24 clinical isolates in triplicate.

15 **Data availability.** All of the data collected and analysed in this study here can be found in  
16 the supplementary material.

## 1 **Results**

2 **Virulence-specific traits show significant intra- and inter-clonal variation.** To elucidate the  
3 role of bacterial factors in human SAB we studied a collection of 300 *S. aureus* isolates  
4 belonging to either the multi-locus defined clonal complex 22 or 30 (CC22, n=135; CC30,  
5 n=165) and containing both methicillin-resistant (MRSA) and methicillin-susceptible (MSSA)  
6 isolates. Each isolate was quantitatively phenotyped with respect to cytolytic activity (the  
7 ability to lyse a monocyte cell line, THP-1) and biofilm formation. Both traits are major  
8 virulence determinants implicated in *S. aureus* disease<sup>1,2</sup>: cytolytic toxins enable the evasion  
9 of cellular aspects of host immunity, release nutrient from host cells and are responsible for  
10 much of the purulent tissue damage associated of *S. aureus* infections<sup>1,2</sup>, whereas biofilm  
11 formation enables the bacteria to colonise foreign material and medical devices, protects it  
12 from many aspects of host immunity and renders some antibiotics less effective<sup>1,2</sup>.

13

14 Despite the close genetic and geographic relationship between the isolates there was  
15 considerable variability with regards to cytolytic toxicity and biofilm formation within the two  
16 clonal backgrounds (fig. 1). We mapped the level of toxicity and biofilm onto maximum  
17 likelihood trees for both collections, which revealed no obvious clustering with respect to  
18 genetic relationship and virulence traits (Suppl. Figs. 1 and 2). The distribution in biofilm  
19 formation, here normalised relative to a single control strain, was mostly comparable  
20 between the two clones and no apparent difference between methicillin resistant (MRSA)  
21 and susceptible (MSSA) strains was observed (fig. 1a and b).

22

23 Toxicity, on the other hand, was found to be significantly higher in CC22 isolates. We  
24 therefore had to dilute the CC22 supernatants to 30% in fresh TSB to see the variability within

1 this clone, whereas the toxicity assay for CC30 isolates was performed on undiluted  
2 supernatant. Although toxicity could not be compared directly between the two clones, it is  
3 clear that the distribution in this virulence-associated trait is much more bimodal for bacteria  
4 of the CC22 background, with strains showing either very high or very low levels of cytolytic  
5 toxicity (fig. 1c), which contrasts with the more uniform distribution in toxicity of CC30  
6 isolates (fig. 1d). As with biofilm, no association was found between methicillin susceptibility  
7 and this particular virulence phenotype in the CC22 isolates. It has previously been reported  
8 that methicillin resistance conferred by the type II *SCCmec* element in the CC30 clonal  
9 background significantly alters the ability of the bacteria to secrete toxins and form  
10 biofilm<sup>28,29</sup>; here we could only confirm its effect on toxicity, with MSSA isolates showing  
11 significantly elevated levels of cytolytic activity (means of 42% for MRSA and 74% for MSSA,  
12  $P < 1E-8$ , Welch's two sample t-test).

13

14 **Genome-wide association scan identifies bacterial virulence-affecting loci.** As the genome  
15 sequence for each isolate in these two collections was available, we sought to examine  
16 whether the phenotypic and genotypic data presented here would contain sufficient  
17 information to enable us to identify toxicity and biofilm affecting loci. We performed a  
18 genome-wide association study (GWAS) on both the CC22 and CC30 collection, where  
19 associations were tested at an uncorrected ( $P < 0.05$ ) and a Bonferroni corrected ( $P < 4.6 \times 10^{-5}$ )  
20 significance threshold (fig. 2a and b). For biofilm formation, no loci reached significance levels  
21 above the Bonferroni threshold for the CC30, whereas 20 loci reached this threshold within  
22 the CC22 collection, (fig. 2a and b). Of these, we were able to functionally verify the effect of  
23 five loci using transposon mutant from the Nebraska library<sup>32</sup> (fig. 2c; Table 1). For toxicity no  
24 loci reached significance levels above the Bonferroni threshold for the CC22 collection, and

1 95 loci reached it for the CC30 collection, however, of these we were only able to verify the  
2 effects of five (fig. 2d; Table 1).

3

4 **Mortality-associated determinants crucially depend on clonal background.** For 160 of the  
5 300 bacteraemic patients we had access to their 30-day mortality data, as well as access to  
6 some of the established host risk factors for mortality including age, comorbidity index and  
7 whether the bacteraemia developed in the community or a healthcare setting (Supp. Table  
8 1). With these data and our bacterial genotypic and phenotypic data we sought to determine  
9 which factors were predictive of infection outcome in those 160 patients. Due to the high  
10 dimensionality and high number of possible interactions between the various host and  
11 bacterial factors we employed a random forests machine learning approach<sup>25,26</sup>. We analysed  
12 the data both in a stratified and in a combined way, i.e. we fitted models to each of the  
13 phenotype, genotype and clinical / patient data separately before fitting a model to the  
14 pooled data.

15 Overall, we found that good predictive accuracy could be achieved by fitting a model to the  
16 various subsets of data for both the CC22 and CC30 isolates, as shown by receiver operator  
17 characteristic (ROC) curves (fig. 3a and b). However, the particular features, or predictors,  
18 that the models revealed as most important for distinguishing cases where patients died or  
19 survived for at least 30 days were crucially dependent on clonal background of the infecting  
20 strain.

21 For the CC22 collection we found that the bacterial phenotype was more predictive than the  
22 host risk factors for which we had data (AUC=0.75 compared to AUC=0.66), whereas the  
23 model including all available data showed the best performance, with misclassification rates  
24 as low as 22% (fig 3c). Looking at the variable importance scores (fig. 3e), a measure of the

1 contribution of the individual features to model performance, confirms that both virulence  
2 factors (toxicity and biofilm) of the infecting bacteria together with the patient's co-  
3 morbidities are the principal predictors of poor infection outcome. Interestingly, although  
4 patient age was selected as one of the more important variables by our feature selection  
5 procedure, its power to differentiate between outcomes of infection by strains of this clonal  
6 lineage was significantly less than that of the bacteria's phenotype. Supplementary figure 3 shows  
7 how the interactions between toxicity, biofilm, co-morbidity and patient age determine the  
8 risk of SAB-associated mortality for this clone, which clearly highlights how the combination  
9 of high toxicity or low biofilm with patient age and co-morbidities significantly increase the  
10 risk of poor infection outcomes in this collection of isolates.

11 For our CC30 collection we also found good predictive accuracy but in this case the factors  
12 contributing to model performance were very different. In this case, neither the bacterial  
13 phenotype nor the data relating to common host risk factors were sufficient to predict 30-  
14 day mortality (fig. 3b). The model performance significantly improved by considering the  
15 bacteria genotype, either by itself or in combination with the other data, and resulted in  
16 relatively low misclassification errors, although specificity with respect to patient death was  
17 not as high as what we achieved for the CC22 collection (64% vs. 74%). The lack of importance of the  
18 phenotype and clinical data in predicting the outcome of CC30-associated bacteraemia is  
19 further highlighted in the variable importance plot in fig. 3f, with the majority of features  
20 being related to the bacterial genotype.

21 It is interesting to note that when the methicillin resistance status of the bacteria was  
22 considered in isolation, it was significantly associated with mortality (26% death amongst the  
23 MRSA Vs 14% death amongst the MSSA infected patients for the CC22 collection and 22% Vs  
24 15% for the CC30 collection). However, the model did not identify it as predictive when

1 considered alongside host features, such as co-morbidity. What this suggests is that patients  
2 with higher co-morbidity scores are more likely to be infected by MRSA, but their mortality is  
3 affected more by their underlying health than the methicillin resistance status of the infecting  
4 bacteria, demonstrating the importance of adopting a multi-variant approach. More  
5 importantly however, is that what is very clear from this work is that the clonal background  
6 of the infecting bacterium has a much greater influence on the severity of infection, and  
7 bacteraemia-associated mortality in particular, than previously appreciated.

8

9 **Functional verification of mortality-associated locus.** Although the feature selection  
10 procedure we employed should reduce the false-discovery rate, false-positive findings cannot  
11 be ruled out without an independent validation dataset. We therefore sought to functionally  
12 verify true virulence affecting polymorphisms. Of the 16 loci selected by the model to be  
13 predictive of mortality within both the CC22 and CC30 collections, only one encoded a known  
14 virulence factors for *S. aureus*, *CapA*. The *capA* gene encodes an enzyme involved in the  
15 capsule biosynthesis, which is responsible for protecting the bacteria from many aspect of  
16 host immunity, and has been shown in animal models to significantly affect disease severity  
17 following bacteraemia<sup>30,31</sup>. Given its role in protecting the bacteria during infection we  
18 hypothesised that variability in *capA* may affect capsule production and as a consequence  
19 affect infection outcome.

20 Within the CC22 collection there were six non-synonymous and one synonymous SNP in this  
21 gene. Of these, only one SNP (at position 142543) was distributed disproportionately  
22 between those isolates from patients who died and those who survived, thus providing the  
23 mortality-related signal detected by the model. SNP142543 affects the substitution of a  
24 proline at position 146 in the protein with a serine. That all six incidences occurred in isolates



1 from patients who survived their bacteraemia suggests that this substitution may alter  
2 capsule quantity or quality to lessen its protective effect. To test this we performed dot-blot  
3 on whole bacterial cells using anti-sera raised against capsule (see Material & Methods).  
4 Alongside the appropriate controls we examined capsule production of 12 isolates from  
5 patients who died, all of which had the proline at position 146, and 12 isolates from patients  
6 who survived, half of which had the serine substitution. As demonstrated in fig. 4a and b,  
7 isolates expressing CapA with a serine at position 146 showed a lack of reactivity to the  
8 antisera, suggesting that these isolates either produce less capsule or that the capsule is  
9 sufficiently different in structure to evade antibody recognition.

10 As capsule is known to protect *S. aureus* from killing by phagocytes, we also performed  
11 opsinophagocytosis killing (OPK) assays on these 24 clinical isolates, and demonstrate that  
12 the isolates with a serine at position 146 of CapA were killed more effectively than those with  
13 a proline at this site. While a definitive demonstration of the effect of this substitution would  
14 require the construction of isogenic mutants, our aim here was to verify the findings of the  
15 predictive model. That capsule production and survival upon exposure to phagocytes,  
16 established features that protects the bacteria during infection, are affected in all of the  
17 strains with this substitution and none of the 18 without it provides functional support for  
18 the sensitivity of the predictive model.

19

## 20 **Discussion**

21 In this study we have demonstrated how *S. aureus* bacteraemia-associated mortality is  
22 crucially influenced by a variety of host and bacteria-related factors, where our findings  
23 suggest that of the host and bacterial features we had access to, the bacterial phenotype and  
24 genotype were the most predictive of infection outcome. From the host perspective, age and

1 co-morbidities were the only factors that we found to be predictive of mortality of patients  
2 infected with bacteria of the CC22 clonal background. For the CC30 strains, on the other hand,  
3 patient age or co-morbidities only played a marginal role for correctly distinguishing infection  
4 outcomes. The lack of a greater effect of co-morbidities for these isolates may be explained  
5 by a difference in their prevalence in patients, as within our collection we found that patients  
6 with CC22-associated bacteraemia had higher CCI scores on average than those with a  
7 bacteraemia caused by a CC30 isolate (3.7 and 2.4 respectively,  $P=0.005$ , Welch's Two Sample  
8 t-test). Furthermore, patients infected with CC22 strains were also significantly older on  
9 average (43 years compared to 32 years,  $P<1E-4$ , Welch's Two Sample t-test). It is therefore  
10 possible that amongst those infected with a CC30 strain the existence of co-morbidities was  
11 less discriminating than for those infected with a CC22 isolates.

12

13 Of the bacterial factors we examined we found significant variability in biofilm formation  
14 within the CC22 and CC30 clones but not between them. Animal studies on the role of biofilm  
15 in virulence have demonstrated that it is specific to certain types of infection. That is, it may  
16 contribute positively to infectious processes involving indwelling catheters but negatively to  
17 virulence in sepsis models. Our CC22 data supports these findings, where low levels of biofilm  
18 was associated with mortality. Unlike with the CC30 collection, where neither high or low  
19 levels of biofilm were predictive of patient mortality.

20

21 Bacterial toxicity was highly predictive of mortality amongst the CC22 infected patients,  
22 whereas for the CC30 infected patient it was not. This may be explained by the relatively low  
23 level of toxicity we and others have reported for this clone<sup>28,34,35</sup>. We have previously  
24 reported that the antibiotic resistant conferring type II SCC*mec* element affects the ability of

1 MRSA to secrete toxins<sup>28</sup>, and we see this effect again here where the mean toxicity for the  
2 CC30 MRSA was significantly lower than for MSSA. However, even the MSSA belonging to this  
3 clone were less toxic than either methicillin-resistant or susceptible isolates from the CC22  
4 collection, which could be a consequence of clonally associated polymorphisms in regulators  
5 of toxicity, such as *Agr*<sup>34</sup>. This clone appears therefore to be inherently constrained in its  
6 ability to produce toxins, and as such must utilize alternative and as yet to be elucidated  
7 means of causing damage to their host. Our result from the two clones studied here contrasts  
8 with those of another study which also looked at toxicity (although they used the term “agr  
9 dysfunction”), where they found that low levels of toxicity was associated with increased  
10 mortality amongst the most severely ill patients in their study<sup>36</sup>. This study contained multiple  
11 *S. aureus* clones, and so it would be interesting to see if the effect they found was specific to  
12 individual clones.

13

14 The multivariate approach taken here also provides some clues on other bacterial factors  
15 linked to mortality of bacteraemic patients. We functionally verified the effect of one of  
16 those: the polymorphism in the *capA* gene that appears to affect capsule production and  
17 consequently the ability of the bacteria to protect itself from host immunity. For the CC30  
18 collection the polymorphisms in the *drp35* gene were the most predictive of mortality, and  
19 this gene has been shown to be up regulated in response to exposure to specific antibiotics<sup>37</sup>,  
20 suggesting that enhanced ability to survive such pressures may increase the ability of the  
21 bacteria to kill their host.

22

23 In addition to identifying factors that affect mortality, we have also used our bacterial  
24 phenotypic and genotypic data to identify effectors of toxicity and biofilm by adopting a

1 GWAS approach, and verified their effect using transposon mutants (fig. 2; Table 1). Helicases  
2 are involved in unwinding DNA for processes such as transcription. Given that the loss of this  
3 gene (NE513) results in a decrease in biofilm formation, this suggests that this helicase may  
4 be responsible for unwinding genes involved in increasing the amount of biofilm a bacterium  
5 produces. The involvement of the quinolone efflux pump NorA (NE1034) is intriguing, given  
6 the co-incidence of biofilm with increased antibiotic resistance. The inactivation of the  
7 peptidase (NE1455) also causes an increase in biofilm formation, and it is thus possible that  
8 it may have a direct effect on breaking down of the proteinaceous components of biofilm.  
9 The role of the other genes in toxicity and biofilm formation is currently unclear, but work to  
10 elucidate this is underway.

11

12 GWAS in any species is prone to produce spurious associations, and on this dataset our  
13 approach generated a 70% (14/20) false positive rate for biofilm formation and 95% for  
14 toxicity (90/95). It is clear that there is significant variability in success rates of this approach,  
15 and this may be due to the presence of loci with varying effects on the phenotype in question.  
16 However, given that these can be readily identified and discounted in bacteria using mutants  
17 from transposon libraries, we believe that it is still a suitable approach to identify candidate  
18 loci whilst avoiding the need to screen entire mutant libraries consisting of thousands of  
19 individual mutants.

20

21 It is likely that other factors not considered here may also contribute to a patient's risk of  
22 death following SAB. Apart from patient care, host genotype and other bacterial phenotypes  
23 could have a significant effect on infection outcome. A larger, more detailed dataset may  
24 enable us to fully identify these factors and unravel their interactions to predict mortality with

1 even higher accuracy. However, given the general complexity of this type of disease and its  
2 epidemiology, where older patients with co-morbidities are the most susceptible, it might  
3 never be feasible to fully predict infection outcome.

4

5 With the growing global problem of antimicrobial resistance, alternative intervention and  
6 control strategies are needed. These include the development of vaccines and identification  
7 of drugs that attenuate the virulence of pathogens. However, without a full understanding of  
8 how the bacterial targets for these strategies are involved in causing disease in humans, there  
9 is a significant risk of investing in and pursuing unsuccessful lines of therapeutic development.

10 For example, our findings suggest that cytolytic toxins, components of biofilm and capsules  
11 are unlikely to be good targets, due to their disparate roles in different stages of disease and  
12 their highly variable expression between and even within closely related bacterial clones.

13 With the move towards the introduction of microbial genome sequencing into routine  
14 diagnostic settings, the ability to use such information to inform clinicians on the likely  
15 outcome of an infection for individual patients would allow them to tailor treatment to that  
16 individual. Our work therefore has the potential to make a significant contribution towards  
17 the implementation of personalised medicine and infectious disease management.

18

## 19 **References.**

- 20 1. Lowy FD. *Staphylococcus aureus* infections. *N. Engl. J. Med.* 1998;339:520-32.
- 21 2. Gordon, RJ & Lowy FD. Pathogenesis of methicillin-resistant *Staphylococcus aureus*  
22 infection. *Clin Infect Dis.* 2008;46:S350-9.

- 1 3. [https://www.gov.uk/government/uploads/system/uploads/attachment\\_data/file/53563](https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/53563)  
2 5/AEC\_final.pdf
- 3 4. Okon KO, Shittu AO, Kudi AA, Umar H, Becker K, Schaumburg F. Population dynamics of  
4 Staphylococcus aureus from Northeastern Nigeria in 2007 and 2012. *Epidemiol Infect.*  
5 2014;142:1737-40.
- 6 5. Walter J, Haller S, Blank HP, Eckmanns T, Abu Sin M, Hermes J. Incidence of invasive  
7 methicillin-resistant Staphylococcus aureus infections in Germany, 2010 to 2014. *Euro*  
8 *Surveill.* 2015;20(46).
- 9 6. van Hal SJ, Jenson SO, Vaska VL, Espedido BA, Pateron DL, Gosbell IB. Predictors of  
10 mortality in Staphylococcus aureus Bacteraemia. *Clin Microbiol Rev.* 2012;25:362-86.
- 11 7. Jenkins A, Diep BA, Mai TT, Vo NH, Warrenner P, Suzich J, et al. Differential expression  
12 and roles of Staphylococcus aureus virulence determinants during colonization and  
13 disease. *MBio.* 2015;6:e02272-14.
- 14 8. Crémieux AC, Saleh-Mghir A, Danel C, Couzon F, Dumitrescu O, Lilin T, et al.  $\alpha$ -  
15 Hemolysin, not Panton-Valentine leukocidin, impacts rabbit mortality from severe sepsis  
16 with methicillin-resistant Staphylococcus aureus osteomyelitis. *J Infect Dis.* 2014;209:  
17 1773-80.
- 18 9. Sharma-Kuinkel BK, Wu Y, Tabor DE, Mok H, Sellman BR, Jenkins A, et al.  
19 Characterization of alpha-toxin hla gene variants, alpha-toxin expression levels, and  
20 levels of antibody to alpha-toxin in hemodialysis and postsurgical patients with  
21 Staphylococcus aureus bacteraemia. *J Clin Microbiol.* 2015;53:227-36.
- 22 10. Laabei M, Uhlemann AC, Lowy FD, Austin ED, Yokoyama M, Ouadi K, et al. Evolutionary  
23 Trade-Offs Underlie the Multi-faceted Virulence of Staphylococcus aureus. *PLoS Biol.*  
24 2015;13:e1002229.

- 1 11. Rose HR, Hiltzman RS, Altman DR, Smyth DS, Wasserman GA, Kafer JM, et al. Cytotoxic  
2 Virulence Predicts Mortality in Nosocomial Pneumonia Due to Methicillin-Resistant  
3 Staphylococcus aureus. *J Infect Dis.* 2015;211:1862-74.
- 4 12. Das S, Lindemann C, Young BC, Muller J, Osterreich B, Ternette N, et al. Natural  
5 mutations in a Staphylococcus aureus virulence regulator attenuate cytotoxicity but  
6 permit bacteraemia and abscess formation. *Proc Natl Acad Sci USA.* 2016;113:E3101-10.
- 7 13. Cosgrove SE, Sakoulas G, Perencevich EN, Schwaber MJ, Karchmer AW, Carmeli Y.  
8 Comparison of mortality associated with methicillin-resistant and methicillin-susceptible  
9 Staphylococcus aureus bacteraemia: a meta-analysis. *Clin Infect Dis* 2003;36:53–59.
- 10 14. Whitby M, McLaws ML, Berry G. Risk of death from methicillin-resistant Staphylococcus  
11 aureus bacteraemia: a meta-analysis. *Med J Aust.* 2001;175:264 –267.
- 12 15. Melzer M, Eykyn SJ, Gransden WR, Chinn S. Is methicillin- resistant Staphylococcus  
13 aureus more virulent than methicillin- susceptible S. aureus? A comparative cohort  
14 study of British patients with nosocomial infection and bacteraemia. *Clin Infect Dis.*  
15 2003;37:1453–1460.
- 16 16. Laabei M, Recker M, Rudkin JK, Aldeljawi M, Gulay Z, Sloan TJ et al. Predicting the  
17 virulence of MRSA from its genome sequence. *Genome research,* 2014;24:839-849.
- 18 17. Saunderson RB, Gouliouris T, Nickerson EK, Cartwright EJ, Kindey A, Aliyu SH, et al.  
19 Impact of routine bedside infectious disease consultation on clinical management and  
20 outcome of Staphylococcus aureus bacteraemia in adults. *Clin Microbiol Infect.*  
21 2015;21:779-85.
- 22 18. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic  
23 comorbidity in longitudinal studies: Development and validation. *J Chronic Dis.*  
24 1987;40:373–83.

- 1 19. Lesens O, Methlin C, Hansmann Y. Role of comorbidity in mortality related to  
2 Staphylococcus aureus bacteraemia: a prospective study using the Charlson weighted  
3 index of comorbidity. *Infect Control Hosp Epidemiol.*2003;24:890–96.
- 4 20. Köser CU, Holden MT, Ellington MJ, Cartwright EJ, Brown NM, Ogilvy-Stuart AL, et al.  
5 Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. *N Engl*  
6 *J Med.* 2012;366:2267-2275.
- 7 21. Holden MT, Hsu LY, Kurt K, Weinert LA, Mather AE, Harris SR, et al. A genomic portrait of  
8 the emergence, evolution, and global spread of a methicillin-resistant Staphylococcus  
9 aureus pandemic. *Genome Res.* 2013;23:653-64.
- 10 22. Holden MT, et al. Complete genomes of two clinical Staphylococcus aureus strains:  
11 evidence for the rapid evolution of virulence and drug resistance. *Proc Natl Acad Sci U S*  
12 *A.* 2004;101:9786-91
- 13 23. Ziebuhr W, Krimmer V, Rachid S, Lossner I, Gotz F, Hacker J. A novel mechanism of phase  
14 variation of virulence in Staphylococcus epidermidis: evidence for control of the  
15 polysaccharide intercellular adhesin synthesis by alternating insertion and excision of  
16 the insertion sequence element IS256. *Mol Microbiol.* 1999;32:345-56.
- 17 24. R Development Core Team. R: A Language and Environment for Statistical Computing.  
18 *R Found Stat Comput.* 2015;1:409.
- 19 25. Breiman L. Random forests. *Mach Learn.* 2001;45:5–32.
- 20 26. Liaw A, Wiener M. Classification and Regression by random forest. *R News.* 2002;2:18–  
21 22.
- 22 27. Genuer R, Poggi J-M, Tuleau-Malot C. VSURF: An R Package for Variable Selection Using  
23 Random Forests. *R J* 2015;7:19–33.



- 1 28. Rudkin JK, Edwards AM, Bowden MG, Brown EL, Pozzi C, et al. Methicillin resistance  
2 reduces the virulence of healthcare-associated methicillin-resistant *Staphylococcus*  
3 *aureus* by interfering with the *agr* quorum sensing system. *J Infect Dis.* 2012; 205(5):798-  
4 806.
- 5 29. Pozzi C, Waters EM, Rudkin JK, Schaeffer CR, Lohan AJ, et al. Methicillin resistance alters  
6 the biofilm phenotype and attenuates virulence in *Staphylococcus aureus* device-  
7 associated infections. *PLoS Pathog.* 2012;8(4):e1002626.
- 8 30. Nilsson IM, Lee JC, Bremell T, Rydén C, Tarkowski A. The role of staphylococcal  
9 polysaccharide microcapsule expression in septicemia and septic arthritis. *Infect Immun.*  
10 1997;65:4216-21.
- 11 31. Tzianabos AO, Wang JY, Lee JC. Structural rationale for the modulation of abscess  
12 formation by *Staphylococcus aureus* capsular polysaccharides. *Proc Natl Acad Sci USA.*  
13 2001;98:9365-70.
- 14 32. Fey PD, Endres JL, Yajjala VK, Widhelm TJ, Boissy RJ, Bose JL, et al. A genetic resource for  
15 rapid and comprehensive phenotype screening of nonessential *Staphylococcus aureus*  
16 genes. *MBio.* 2013;4:e00537-12.
- 17 33. Fowler VG Jr, Nelson CL, McIntyre LM, Kreiswirth BN, Monk A, et al. Potential  
18 associations between hematogenous complications and bacterial genotype in  
19 *Staphylococcus aureus* infection. *J Infect Dis.* 2007;196(5):738-47.
- 20 34. DeLeo FR, Kennedy AD, Chen L, Bubeck Wardenburg J, Kobayashi SD, et al. Molecular  
21 differentiation of historic phage-type 80/81 and contemporary epidemic *Staphylococcus*  
22 *aureus*. *Proc Natl Acad Sci U S A.* 2011;108(44):18091-6

- 1 35. Cheung GY, Kretschmer D, Duong AC, Yeh AJ, Ho TV, et al. Production of an attenuated  
2 phenol-soluble modulin variant unique to the MRSA clonal complex 30 increases  
3 severity of bloodstream infection. *PLoS Pathog.* 2014;10(8):e1004298.
- 4 36. Schweizer ML, Furuno JP, Sakoulas G, Johnson JK, Harris AD, et al. Increased Mortality  
5 with Accessory Gene Regulator (*agr*) Dysfunction in *Staphylococcus aureus* among  
6 Bacteremic Patients. *Antimicrob Agents Chemother.* 2011;55(3): 1082–1087.
- 7 37. Murakami H, Matsumaru H, Kanamori M, Hayashi H, Ohta T. Cell wall-affecting  
8 antibiotics induce expression of a novel gene, *drp35*, in *Staphylococcus aureus*.  
9 *Biochem Biophys Res Commun.* 1999;264(2):348-51.

10

11 Correspondence and requests for materials should be addressed to Dr. Ruth Massey at  
12 [ruth.massey@bristol.ac.uk](mailto:ruth.massey@bristol.ac.uk).

13

14 **Acknowledgements:** We would like to thank Dr Theodore Gouliouris, for his role in  
15 collecting and supplying the clinical data, and, Dr Emma Nickerson and Dr Sani Aliyu for their  
16 role in collection of clinical information. We thank the Sanger Institute’s core Pathogen  
17 Production Groups and Pathogen Informatics Group.

18

19 **Authors contribution:** MR, ML, MST, RBS, BB, MET, SJP and RCM collected the data, MR, ML  
20 and RCM designed the experiments, ML, KO, ES, MY, JS, LT, GM and performed  
21 experiments, MR, ML, MST, SR, SB, SJP and RCM analysed the data, MR, ML, SJP and RCM  
22 wrote the manuscript.

23

24

1

## 2 **Figure legends**

### 3 **Figure 1. The toxicity and biofilm forming abilities of *S. aureus* bacteraemia isolates. (a, b)**

4 Biofilm forming abilities were quantified relative to a control included in each assay in a static  
5 96 well format. A wide range of biofilm forming abilities was evident with no discernible  
6 difference between methicillin-resistant (MRSA, grey circles) and methicillin-susceptible  
7 (MSSA, black circles) isolates. Isolates of the CC22 (a, n=136) and CC30 (b, n=164) backgrounds  
8 had comparable distributions. **(c, d)** Toxicity for each isolate was determined by incubating  
9 bacterial supernatant with cultured human cells, using flow cytometry to quantify cell death  
10 (toxicity). Note, the supernatant of the CC22 isolates was diluted to 30%, whereas the  
11 supernatant of the CC30 isolates was used undiluted. Apart from differences in baseline  
12 toxicity between the two clones, there is a marked difference in their distribution, where CC22  
13 isolates where toxicity was either very high or very low, compared to a more uniform  
14 distribution for CC30. No difference was observed between methicillin-resistant (MRSA, grey  
15 circles) and methicillin-susceptible (MSSA, black circles) isolates.

16

17

18

### 19 **Figure 2. Genome-wide associations and functional validation of biofilm affecting**

20 **polymorphisms. (a, b)** Manhattan plots representing the results of the GWAS (performed  
21 using ANOVA in R) for both biofilm (left-hand side) and toxicity (right-hand side), performed  
22 on the bacteraemia isolates of the CC22 **(a)** and CC30 **(b)** background. For toxicity, only two  
23 polymorphic loci were significantly associated using an uncorrected threshold (indicated by  
24 the red vertical dashed line), one in the *agrC* gene and the other in a gene encoding a putative  
25 membrane protein. These loci are indicated with red circles, the top dot corresponding to *agrC*  
26 and the bottom the membrane protein. For biofilm, 20 loci were significantly associated with  
27 toxicity when Bonferroni was used to correct for multiple comparisons (indicated by the green  
28 vertical dotted line). **(c, d)** Functional validation of the effects of five biofilm-associated loci **(c)**  
29 and five toxicity-associated loci **(d)** in CC22 bacteria using transposon insertion and compared  
30 to the wild-type (JE2) (Welch's two-sided *t*-test, with \*:  $P < 2E-2$ , \*\*:  $P < 2E-4$ , \*\*\*:  $P < 2E-6$ ).  
31 Results are based on four biological replicates per strain that were repeated three times each.  
32 The medians are presented as horizontal bars, with the boxes and whiskers showing the 1st  
33 and 3rd quartile and interquartile ranges.

34

35

### 36 **Figure 3. Predictive model performance and variable importance.**

37 Random forests were fitted against different subsets of data and the combined data for both the CC22 isolates (top)  
38 and CC30 isolates (bottom). **(a, b)** Receiver operator characteristic (ROC) curves of the *random*  
39 *forests* fit to four sets of variables consisting of host risk factors (clin, red line), genotype data  
40 (geno, blue line), phenotype data (pheno, black line) and all available variables (all, cyan line).  
41 As indicated by the area under curve (AUC), the model fit to a combination of all available data  
42 results in the highest predictive accuracy. The dotted black line denotes the expected result  
43 by random guessing (AUC=0.5). **(c, d)** Confusion matrices demonstrating the model's  
44 accuracies in predicting 30-day mortality when fitted against all data. The out-of-bag  
45 classification and misclassification rates are illustrated as dark-blue/diagonal and light-  
46 blue/off-diagonal wedges, respectively. **(e, f)** Relative importance scores, as measures of a

variable's influence on the model's predictive performance, of the most predictive features (red circles: mean, errorbars: standard deviation); a detailed list of the genetic loci is provided as Supplementary Table 1 and 2.

**Figure 4. Capsule production is affected in CC22 isolates containing a mortality-predicting SNP.** **A:** The production of capsule was compared between 24 CC22 isolates using anti-serum raised against a type 5 *S. aureus* capsule by dot-blotting (a representative blot is presented). Twelve of the isolates were those from patients who survived their bacteraemia (alive) and 12 from patients who did not survive their bacteraemia (dead). Of the strains from patient who survived, 6 contained a proline (P) at position 146 of CapA, while the other 6 contained a serine (S). **B:** The dot-blotting were performed three times and the intensity of the reaction of each strain quantified. Mean intensity for each isolate is presented, where the isolates containing the serine were significantly less reactive to the antiserum. **C:** To compare the protective qualities of the capsule we performed opsinophagocytosis assays using human PMNs with each of the 24 isolates. The data presented is the % killed bacteria for each of the 24 clinical isolates which were performed in triplicate. (Welch's two-sided *t*-test, with \*:  $P < 2E-2$ , \*\*:  $P < 2E-4$ , \*\*\*:  $P < 2E-6$ ).

## TABLES

**Table 1:** Functionally verified loci associated with biofilm and toxicity.

Transposon Number	Locus Tag in Reference Genome	Gene name (where known)	Protein activity
<b>BIOFILM</b>			
NE1034	SAEMRSA15_16820	<i>norA</i>	A quinolone efflux protein
NE1149	SAEMRSA15_01530		a putative thiamine pyrophosphate enzyme/indole-3-pyruvate decarboxylase
NE1637	SAEMRSA15_02020		a putative inosine-uridine preferring nucleoside hydrolase
NE513	SAEMRSA15_23880		A putative helicase
NE20	SAEMRSA15_02030		A putative PTS multidomain regulator
<b>TOXICITY</b>			
NE873	SAR2125	<i>agrC</i>	Response regulator of the Agr quorum sensing system
NE1532	SAR1216	<i>agrA</i>	Autoinducer sensor protein of the Agr quorum sensing system
NE569	SAR1221	<i>sucC</i>	Putative CoA synthetase protein
NE885	SAR1265		putative pyruvate flavodoxin/ferredoxin oxidoreductase

NE932	SAR0147		putative nucleotidase
-------	---------	--	-----------------------

1