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

Mario Recker[1*], Maisem Laabei[2*], Michelle S. Toleman[3], Sandra Reuter[3], Rebecca B. Saunderson[3], Beth Blane[3], M. Estee Török[3], Khadija Ouadi[2], Emily Stevens[2], Maho Yokoyama[2], Joseph Steventon[2], Luke Thompson[2], Gregory Milne[2], Sion Bayliss[2], Leann Bacon[2], Sharon J. Peacock[3,4] and Ruth C. Massey[2,5ø].

1: Centre for Mathematics & the Environment, University of Exeter, Penryn Campus, Penryn TR10 9EZ, UK.

2: Dept. of Biology and Biochemistry and the Milner Centre for Evolution, University of Bath, Bath, BA2 7AY, UK.

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

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

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

* contributed equally to this work.

ø Corresponding author: ruth.massey@bristol.ac.uk

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.

Introduction

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

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

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

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

Material and Methods.

Strain and clinical metadata collection. All isolates were collected from patients admitted to a single hospital with their first episode of SAB between 2006 and 2012, and were stored in glycerol at -80°C. Samples were collected during an observational cohort study of adults with SAB at Addenbrooke's Hospital, Cambridge, UK between 2006 and 2012. These isolates represented the genetic diversity present across the British Society for Antimicrobial Chemotherapy collection of sequenced CC22 and CC30 isolates collected from multiple sites across the UK and Ireland since 2008. Written informed consent was not required as the study was conducted as part of a service evaluation of the management of SAB. Ethical approval was obtained from the University of Cambridge Human Biology Research Ethics Committee (reference HBREC.2013.05) and the Cambridge University Hospitals NHS Foundation Trust Research and Development Department (reference A092869). Study definitions have been defined previously[17] and were used to determine the focus of the bacteraemia, classify the bacteraemia as community-acquired, hospital-acquired or healthcare-associated, and to report outcomes, including death at 30 days. The previously defined definitions[17] were placed into eight categories; intravascular catheter infection (SAB caused from a central or peripheral venous catheter), bone and joint infection (SAB due to a native joint infection), deep tissue abscess (infection due to lung infection, deep tissue abscesses, mycotic aneurysms and lesions in the central nervous system), infective endocarditis, infection due to a prosthetic device (including prosthetic joints, implanted vascular devices, prosthetic cardiac valves), skin and soft tissue infection, disease of the urogenital tract, or an unknown focus of infection. The presence of comorbidities was assessed using the Charlson comorbidity index (CCI) and were dichotomized into scores of <3 or ≥3 [18,19], as detailed in Supplementary Table 1.

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

Cytotoxicity Assay. Overnight *S. aureus* cultures were diluted 1:1000 into fresh tryptone soya broth and incubated for 18 h at 37°C with shaking at 180 rpm in 30 mL glass tubes. *S. aureus* supernatants were harvested from 18 h cultures by centrifugation at 14,600 rpm for 10 min. The THP-1 human monocyte-macrophage cell line (ATCC#TIB-202, which was authenticated and if they were tested for mycoplasma) was routinely grown in suspension in 30 mL of of RPMI-1640, supplemented with 10% heat-inactivated fetal bovine serum (FBS), 1 µM L-glutamine, 200 units/mL penicillin, and 0.1 mg/mL streptomycin at 37°C in a humidified incubator with 5% CO_2. Cells were routinely viewed microscopically every 48–60 h and harvested by centrifugation at 1,000 rpm for 10 min at room temperature and re-suspended to a final density of 1–1.2 x 10^6 cells/mL in tissue-grade phosphate buffered saline. This procedure typically yielded >95% viability of cells as determined by trypan blue exclusion and easyCyte flow cytometry. To evaluate *S. aureus* toxicity, for the CC30 isolates undiluted supernatant was used, but due to the higher level of toxicity of the CC22 isolated the supernatant was diluted to 30% of the original volume in TSB. The neat and diluted bacterial supernatants (20µL) were incubated with 20 µL of washed THP-1 cells for 12 min at 37°C under static conditions. Cell death was quantified by easyCyte (Millipore) flow cytometry using the Guava Viability stain (Millipore) according to manufacturer's instructions, with the toxicity of each isolate quantified in triplicate and the mean of this presented (listed in Supplementary Table 1).

Biofilm assay. Biofilm formation was quantified using a 1:40 dilution from overnight cultures into 100 µL of fresh tryptic soy broth supplemented with 0.5% sterile filtered glucose (TSBG) in 96-well polystyrene plate (Costar). Perimeter wells of the 96-well plate were filled with sterile H_2O and plates were placed in a separate plastic container inside a 37°C incubator and grown for 24 h under static conditions. For the transposon mutants, erythromycin (5 µg/mL) was added to the growth medium. Semi-quantitative measurements of biofilm formation on 96-well polystyrene plates was determined based on the method of Ziebuhr et al[23]. Following 24-h growth, plates were washed vigorously five times in PBS, dried and stained with 150 µL of 1% crystal violet for 30 min at room temperature. Following five washes of PBS, wells were re-suspended in 200 µL of 7% acetic acid, and optical density at 595 nm was recorded using a Fluorimeter plate reader (BMG Labtech). To control for day to day variability, for the clinical isolates a control strain (E-MRSA15) was included on each plate in triplicate, and absorbance values were normalised against this (listed in Additional Table 1). For the transposon mutants, as JE2 is the wild type strain this was used as the control strain, and the effect of mutating the loci made relative to this. For this experiment the assays were performed in triplicate on each plate and repeated four times.

Genome wide association study (GWAS). The genome of the reference strains HO 5096 0412 and MRSA252 was split loci corresponding to coding region and intergenic regions, and loci containing SNPs relative to the reference genome were identified. Annotated intergenic elements such as miscellaneous RNAs were considered separate loci. Synonymous SNPs in coding regions and SNPs in known mobile genetic elements and repeat regions were not considered. The resulting loci were named allele_X where X refers to the position of the SNP at the 5' end of that block, relative to the origin of replication. Each of these alleles in each isolate was scored as 1 if it differed from the reference and 0 if it didn't. These allele scores for each isolate were used as the genotypic information for the following analysis. Significant associations between bacterial genotype and either phenotype (toxicity and biofilm formation) were identified by fitting an analysis of variance model (ANOVA) in R[24] and using a minor allele frequency cut-off of 5%. The reported *P* values are not corrected for multiple testing; the Bonferroni statistical significance threshold is instead provided in fig. 3.

Predictive modelling. We employed a *Random Forests*[25] machine learning approach, using the *randomForest* package in R[26], to identify the most important and clone-specific determinants for host mortality based on the genotype, phenotype and clinical meta-data (see supplementary file S1_RF_analysis.pdf for full details). The input variables were thus taken from the pooled set of SNP data, relative cytolytic toxicity, relative biofilm formation and infection/patient-specific data for each isolate. Due to the genetic differences between CC22 and CC30, where only a small proportion of SNP positions overlapped, we carried out the analysis for both clones separately.

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

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

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

Capsule dot-blots. Bacterial were grown overnight in TSB at 37°C with shaking. 5μl of overnight culture was spotted onto nitrocellulose membranes and the membranes dried for 10 minutes at 65°C. The membranes were washed three times in PBS and then once for 1hr at 37°C in PBS with 2mg/ml trypsin. Membranes were then blocked in 0.05% skim milk for 1hr, and washed three times in PBS with 0.05% Tween (PBS-tween). Anti-Cap5 antiserum (a generous gift from Prof. Jean Lee, Harvard) was diluted 1:1000 in the PBS-tween buffer and the membrane was incubated in this with gentle agitation for 1hr at room temperature. The membranes were washed three time in PBS-tween. The protein A-HRP conjugate was diluted 1:1000 in PBS-tween and the membrane was incubated in this with gentle agitation for 1hr at room temperature. The membranes were washed and the reactivity of the antiserum detected using the colorimetric substrate 4-chloro-1-napthol. The blots were scanned and density of capsule anti-serum reactivity quantified using the ImageJ software. Each strain was blotted three times and the average density of the three repeats for each isolate is presented.

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

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

Results

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

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

Toxicity, on the other hand, was found to be significantly higher in CC22 isolates. We therefore had to dilute the CC22 supernatants to 30% in fresh TSB to see the variability within this clone, whereas the toxicity assay for CC30 isolates was performed on undiluted supernatant. Although toxicity could not be compared directly between the two clones, it is clear that the distribution in this virulence-associated trait is much more bimodal for bacteria of the CC22 background, with strains showing either very high or very low levels of cytolytic toxicity (fig. 1c), which contrasts with the more uniform distribution in toxicity of CC30 isolates (fig. 1d). As with biofilm, no association was found between methicillin susceptibility and this particular virulence phenotype in the CC22 isolates. It has previously been reported that methicillin resistance conferred by the type II SCC*mec* element in the CC30 clonal background significantly alters the ability of the bacteria to secrete toxins and form biofilm[28,29]; here we could only confirm its effect on toxicity, with MSSA isolates showing significantly elevated levels of cytolytic activity (means of 42% for MRSA and 74% for MSSA, *P*<1E-8, Welch's two sample t-test).

Genome-wide association scan identifies bacterial virulence-affecting loci. As the genome sequence for each isolate in these two collections was available, we sought to examine whether the phenotypic and genotypic data presented here would contain sufficient information to enable us to identify toxicity and biofilm affecting loci. We performed a genome-wide association study (GWAS) on both the CC22 and CC30 collection, where associations were tested at an uncorrected ($P<0.05$) and a Bonferroni corrected ($P<4.6x10^{-5}$) significance threshold (fig. 2a and b). For biofilm formation, no loci reached significance levels above the Bonferroni threshold for the CC30, whereas 20 loci reached this threshold within the CC22 collection, (fig. 2a and b). Of these, we were able to functionally verify the effect of five loci using transposon mutant from the Nebraska library[32] (fig. 2c; Table 1). For toxicity no loci reached significance levels above the Bonferroni threshold for the CC22 collection, and

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

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

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

For the CC22 collection we found that the bacterial phenotype was more predictive than the host risk factors for which we had data (AUC=0.75 compared to AUC=0.66), whereas the model including all available data showed the best performance, with misclassification rates as low as 22% (fig 3c). Looking at the variable importance scores (fig. 3e), a measure of the contribution of the individual features to model performance, confirms that both virulence factors (toxicity and biofilm) of the infecting bacteria together with the patient's co-morbidities are the principal predictors of poor infection outcome. Interestingly, although patient age was selected as one of the more important variables by our feature selection procedure, its power to differentiate between outcomes of infection by strains of this clonal lineage was significantly less that of the bacteria's phenotype. Supplementary figure 3 shows how the interactions between toxicity, biofilm, co-morbidity and patient age determine the risk of SAB-associated mortality for this clone, which clearly highlights how the combination of high toxicity or low biofilm with patient age and co-morbidities significantly increase the risk of poor infection outcomes in this collection of isolates.

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

It is interesting to note that when the methicillin resistance status of the bacteria was considered in isolation, it was significantly associated with mortality (26% death amongst the MRSA Vs 14% death amongst the MSSA infected patients for the CC22 collection and 22% Vs 15% for the CC30 collection). However, the model did not identify it as predictive when considered alongside host features, such as co-morbidity. What this suggests is that patients with higher co-morbidity scores are more likely to be infected by MRSA, but their mortality is affected more by their underlying health than the methicillin resistance status of the infecting bacteria, demonstrating the importance of adopting a multi-variant approach. More importantly however, is that what is very clear from this work is that the clonal background of the infecting bacterium has a much greater influence on the severity of infection, and bacteraemia-associated mortality in particular, than previously appreciated.

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

Within the CC22 collection there were six non-synonymous and one synonymous SNP in this gene. Of these, only one SNP (at position 142543) was distributed disproportionately between those isolates from patients who died and those who survived, thus providing the mortality-related signal detected by the model. SNP142543 affects the substitution of a proline at position 146 in the protein with a serine. That all six incidences occurred in isolates from patients who survived their bacteraemia suggests that this substitution may alter capsule quantity or quality to lessen its protective effect. To test this we performed dot-blots on whole bacterial cells using anti-sera raised against capsule (see Material & Methods). Alongside the appropriate controls we examined capsule production of 12 isolates from patients who died, all of which had the proline at position 146, and 12 isolates from patients who survived, half of which had the serine substitution. As demonstrated in fig. 4a and b, isolates expressing CapA with a serine at position 146 showed a lack of reactivity to the antisera, suggesting that these isolates either produce less capsule or that the capsule is sufficiently different in structure to evade antibody recognition.

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

Discussion

In this study we have demonstrated how *S. aureus* bacteraemia-associated mortality is crucially influenced by a variety of host and bacteria-related factors, where our findings suggest that of the host and bacterial features we had access to, the bacterial phenotype and genotype were the most predictive of infection outcome. From the host perspective, age and co-morbidities were the only factors that we found to be predictive of mortality of patients infected with bacteria of the CC22 clonal background. For the CC30 strains, on the other hand, patient age or co-morbidities only played a marginal role for correctly distinguishing infection outcomes. The lack of a greater effect of co-morbidities for these isolates may be explained by a difference in their prevalence in patients, as within our collection we found that patients with CC22-associated bacteraemia had higher CCI scores on average than those with a bacteraemia caused by a CC30 isolate (3.7 and 2.4 respectively, *P*=0.005, Welch's Two Sample t-test). Furthermore, patients infected with CC22 strains were also significantly older on average (43 years compared to 32 years, *P*<1E-4, Welch's Two Sample t-test). It is therefore possible that amongst those infected with a CC30 strain the existence of co-morbidities was less discriminating than for those infected with a CC22 isolates.

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

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

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

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

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

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

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

It is likely that other factors not considered here may also contribute to a patient’s risk of death following SAB. Apart from patient care, host genotype and other bacterial phenotypes could have a significant effect on infection outcome. A larger, more detailed dataset may enable us to fully identify these factors and unravel their interactions to predict mortality with even higher accuracy. However, given the general complexity of this type of disease and its epidemiology, where older patients with co-morbidities are the most susceptible, it might never be feasible to fully predict infection outcome.

With the growing global problem of antimicrobial resistance, alternative intervention and control strategies are needed. These include the development of vaccines and identification of drugs that attenuate the virulence of pathogens. However, without a full understanding of how the bacterial targets for these strategies are involved in causing disease in humans, there is a significant risk of investing in and pursuing unsuccessful lines of therapeutic development. For example, our findings suggest that cytolytic toxins, components of biofilm and capsules are unlikely to be good targets, due to their disparate roles in different stages of disease and their highly variable expression between and even within closely related bacterial clones. With the move towards the introduction of microbial genome sequencing into routine diagnostic settings, the ability to use such information to inform clinicians on the likely outcome of an infection for individual patients would allow them to tailor treatment to that individual. Our work therefore has the potential to make a significant contribution towards the implementation of personalised medicine and infectious disease management.

References.

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

3. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/535635/AEC_final.pdf

4. Okon KO, Shittu AO, Kudi AA, Umar H, Becker K, Schaumburg F. Population dynamics of Staphylococcus aureus from Northeastern Nigeria in 2007 and 2012. Epidemiol Infect. 2014;142:1737-40.

5. Walter J, Haller S, Blank HP, Eckmanns T, Abu Sin M, Hermes J. Incidence of invasive meticillin-resistant Staphylococcus aureus infections in Germany, 2010 to 2014. Euro Surveill. 2015;20(46).

6. van Hal SJ, Jenson SO, Vaska VL, Espedido BA, Pateron DL, Gosbell IB. Predictors of mortality in Staphylococcus aureus Bacteraemia. Clin Microbiol Rev. 2012;25:362-86.

7. Jenkins A, Diep BA, Mai TT, Vo NH, Warrener P, Suzich J, et al. Differential expression and roles of Staphylococcus aureus virulence determinants during colonization and disease. MBio. 2015;6:e02272-14.

8. Crémieux AC, Saleh-Mghir A, Danel C, Couzon F, Dumitrescu O, Lilin T, et al. α-Hemolysin, not Panton-Valentine leukocidin, impacts rabbit mortality from severe sepsis with methicillin-resistant Staphylococcus aureus osteomyelitis. J Infect Dis. 2014;209: 1773-80.

9. Sharma-Kuinkel BK, Wu Y, Tabor DE, Mok H, Sellman BR, Jenkins A, et al. Characterization of alpha-toxin hla gene variants, alpha-toxin expression levels, and levels of antibody to alpha-toxin in hemodialysis and postsurgical patients with Staphylococcus aureus bacteraemia. J Clin Microbiol. 2015;53:227-36.

10. Laabei M, Uhlemann AC, Lowy FD, Austin ED, Yokoyama M, Ouadi K, et al. Evolutionary Trade-Offs Underlie the Multi-faceted Virulence of Staphylococcus aureus. PLoS Biol. 2015;13:e1002229.

11. Rose HR, Hilzman RS, Altman DR, Smyth DS, Wasserman GA, Kafer JM, et al. Cytotoxic Virulence Predicts Mortality in Nosocomial Pneumonia Due to Methicillin-Resistant Staphylococcus aureus. J Infect Dis. 2015;211:1862-74.

12. Das S, Lindemann C, Young BC, Muller J, Osterreich B, Ternette N, et al. Natural mutations in a Staphylococcus aureus virulence regulator attenuate cytotoxicity but permit bacteraemia and abscess formation. Proc Natl Acad Sci USA. 2016;113:E3101-10.

13. Cosgrove SE, Sakoulas G, Perencevich EN, Schwaber MJ, Karchmer AW, Carmeli Y. Comparison of mortality associated with methicillin-resistant and methicillin-susceptible Staphylococcus aureus bacteraemia: a meta-analysis. Clin Infect Dis 2003;36:53–59.

14. Whitby M, McLaws ML, Berry G. Risk of death from methicillin-resistant Staphylococcus aureus bacteraemia: a meta-analysis. Med J Aust. 2001;175:264 –267.

15. Melzer M, Eykyn SJ, Gransden WR, Chinn S. Is methicillin- resistant Staphylococcus aureus more virulent than methicillin- susceptible S. aureus? A comparative cohort study of British patients with nosocomial infection and bacteraemia. Clin Infect Dis. 2003;37:1453–1460.

16. Laabei M, Recker M, Rudkin JK, Aldeljawi M, Gulay Z, Sloan TJ et al. Predicting the virulence of MRSA from its genome sequence. Genome research, 2014;24:839-849.

17. Saunderson RB, Gouliouris T, Nickerson EK, Cartwright EJ, Kindey A, Aliyu SH, et al. Impact of routine bedside infectious disease consultation on clinical management and outcome of Staphylococcus aureus bacteraemia in adults. Clin Microbiol Infect. 2015;21:779-85.

18. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. J Chronic Dis. 1987;40:373–83.

19. Lesens O, Methlin C, Hansmann Y. Role of comorbidity in mortality related to Staphylococcus aureus bacteraemia: a prospective study using the Charlson weighted index of comorbidity. Infect Control Hosp Epidemiol.2003;24:890–96.

20. Köser CU, Holden MT, Ellington MJ, Cartwright EJ, Brown NM, Ogilvy-Stuart AL, et al. Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. N Engl J Med. 2012;366:2267-2275.

21. Holden MT, Hsu LY, Kurt K, Weinert LA, Mather AE, Harris SR, et al. A genomic portrait of the emergence, evolution, and global spread of a methicillin-resistant Staphylococcus aureus pandemic. Genome Res. 2013;23:653-64.

22. Holden MT, 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;101:9786-91

23. Ziebuhr W, Krimmer V, Rachid S, Lossner I, Gotz F, Hacker J. A novel mechanism of phase variation of virulence in Staphylococcus epidermidis: evidence for control of the polysaccharide intercellular adhesin synthesis by alternating insertion and excision of the insertion sequence element IS256. Mol Microbiol. 1999;32:345-56.

24. R Developement Core Team. R: A Language and Environment for Statistical Computing. R Found Stat Comput. 2015;1:409.

25. Breiman L. Random forests. Mach Learn. 2001;45:5–32.

26. Liaw A, Wiener M. Classification and Regression by random forest. R News. 2002;2:18–22.

27. Genuer R, Poggi J-M, Tuleau-Malot C. VSURF: An R Package for Variable Selection Using Random Forests. R J 2015;7:19–33.

28. 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(5):798-806.

29. Pozzi C, Waters EM, Rudkin JK, Schaeffer CR, Lohan AJ, et al. Methicillin resistance alters the biofilm phenotype and attenuates virulence in Staphylococcus aureus device-associated infections. PLoS Pathog. 2012;8(4):e1002626.

30. Nilsson IM, Lee JC, Bremell T, Rydén C, Tarkowski A. The role of staphylococcal polysaccharide microcapsule expression in septicemia and septic arthritis. Infect Immun. 1997;65:4216-21.

31. Tzianabos AO, Wang JY, Lee JC. Structural rationale for the modulation of abscess formation by Staphylococcus aureus capsular polysaccharides. Proc Natl Acad Sci USA. 2001;98:9365-70.

32. Fey PD, Endres JL, Yajjala VK, Widhelm TJ, Boissy RJ, Bose JL, et al. A genetic resource for rapid and comprehensive phenotype screening of nonessential Staphylococcus aureus genes. MBio. 2013;4:e00537-12.

33. 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(5):738-47.

34. DeLeo FR, Kennedy AD, Chen L, Bubeck Wardenburg J, Kobayashi SD, et al. Molecular differentiation of historic phage-type 80/81 and contemporary epidemic Staphylococcus aureus. *Proc Natl Acad Sci U S A.* 2011;108(44):18091-6

35. Cheung GY, Kretschmer D, Duong AC, Yeh AJ, Ho TV, et al. Production of an attenuated phenol-soluble modulin variant unique to the MRSA clonal complex 30 increases severity of bloodstream infection. *PLoS Pathog*. 2014;10(8):e1004298.

36. 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(3): 1082–1087.

37. Murakami H, Matsumaru H, Kanamori M, Hayashi H, Ohta T. Cell wall-affecting antibiotics induce expression of a novel gene, *drp35*, in *Staphylococcus aureus*. *Biochem Biophys Res Commun*. 1999;264(2):348-51.

Correspondence and requests for materials should be addressed to Dr. Ruth Massey at ruth.massey@bristol.ac.uk.

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

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

Figure legends

Figure 1. The toxicity and biofilm forming abilities of *S. aureus* bacteraemia isolates. (a, b) Biofilm forming abilities were quantified relative to a control included in each assay in a static 96 well format. A wide range of biofilm forming abilities was evident with no discernible difference between methicillin-resistant (MRSA, grey circles) and methicillin-susceptible (MSSA, black circles) isolates. Isolates of the CC22 (a, n=136) and CC30 (b, n=164) backgrounds had comparable distributions. **(c, d)** Toxicity for each isolate was determined by incubating bacterial supernatant with cultured human cells, using flow cytometry to quantify cell death (toxicity). Note, the supernatant of the CC22 isolates was diluted to 30%, whereas the supernatant of the CC30 isolates was used undiluted. Apart from differences in baseline toxicity between the two clones, there is a marked difference in their distribution, where CC22 isolates where toxicity was either very high or very low, compared to a more uniform distribution for CC30. No difference was observed between methicillin-resistant (MRSA, grey circles) and methicillin-susceptible (MSSA, black circles) isolates.

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

Figure 3. Predictive model performance and variable importance. Random forests were fitted against different subsets of data and the combined data for both the CC22 isolates (top) and CC30 isolates (bottom). **(a, b)** Receiver operator characteristic (ROC) curves of the *random forests* fit to four sets of variables consisting of host risk factors (clin, red line), genotype data (geno, blue line), phenotype data (pheno, black line) and all available variables (all, cyan line). As indicated by the area under curve (AUC), the model fit to a combination of all available data results in the highest predictive accuracy. The dotted black line denotes the expected result by random guessing (AUC=0.5). **(c, d)** Confusion matrices demonstrating the model's accuracies in predicting 30-day mortality when fitted against all data. The out-of-bag classification and misclassification rates are illustrated as dark-blue/diagonal and light-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-blots 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**
BIOFILM			
NE1034	SAEMRSA15_16820	*norA*	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
TOXICITY			
NE873	SAR2125	*agrC*	Response regulator of the Agr quorum sensing system
NE1532	SAR1216	*agrA*	Autoinducer sensor protein of the Agr quorum sensing system
NE569	SAR1221	*sucC*	Putative CoA synthetase protein
NE885	SAR1265		putative pyruvate flavodoxin/ferredoxin oxidoreductase

NE932	SAR0147		putative nucleotidase