Multi-omics approach identifies novel pathogen-derived prognostic biomarkers in patients with *Pseudomonas aeruginosa* bloodstream infection

Matthias Willmann[1,2]*, Stephan Götting[3], Daniela Bezdan[4,5], Boris Maček[6], Ana Velic[6], Matthias Marschal[1], Wichard Vogel[7], Ingo Flesch[8], Uwe Markert[9], Annika Schmidt[1], Pierre Kübler[1], Maria Haug[1], Mumina Javed[1,2], Benedikt Jentzsch[1,2], Philipp Oberhettinger[1], Monika Schütz[1,2], Erwin Bohn[1,2], Michael Sonnabend[1,2], Kristina Klein[1,2], Ingo B Autenrieth[1,2], Stephan Ossowski[4,5,10], Sandra Schwarz[1], and Silke Peter[1,2]

[1]Institute of Medical Microbiology and Hygiene, University of Tübingen, Tübingen, Germany

[2]German Center for Infection Research (DZIF), partner site Tübingen, Tübingen, Germany

[3]Institute for Medical Microbiology and Infection Control, University Hospital, Goethe-University, Frankfurt am Main, Germany

[4]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain

[5]Universitat Pompeu Fabra (UPF), Barcelona, Spain

[6]Proteome Center Tübingen, Auf der Morgenstelle, Tübingen, Germany

[7]Medical Center, Department of Hematology, Oncology, Immunology, Rheumatology & Pulmonology, University of Tübingen, Tübingen, Germany

[8]BG Trauma Center, University of Tübingen, Tübingen, Germany

[9]Clinic for General, Visceral and Vascular Surgery, Zollernalb Hospital, Albstadt, Germany

[10]Institute of Medical Genetics and Applied Genomics, University of Tübingen, Tübingen, Germany

*Address correspondence to Matthias Willmann: Institute of Medical Microbiology and Hygiene, Elfriede-Aulhorn-Str 6, 72076, Tübingen, Germany
matthias.willmann@med.uni-tuebingen.de

Abstract

Pseudomonas aeruginosa is a human pathogen that causes health-care associated blood stream infections (BSI). Although *P. aeruginosa* BSI are associated with high mortality rates, the clinical relevance of pathogen-derived prognostic biomarker to identify patients at risk for unfavorable outcome remains largely unexplored. We found novel pathogen-derived prognostic biomarker candidates by applying a multi-omics approach on a multicenter sepsis patient cohort. Multi-level Cox regression was used to investigate the relation between patient characteristics and pathogen features (2298 accessory genes, 1078 core protein levels, 107 parsimony-informative variations in reported virulence factors) with 30-day mortality. Our analysis revealed that presence of the *helP* gene encoding a putative DEAD-box helicase was independently associated with a fatal outcome (hazard ratio 2.01, $p = 0.05$). *helP* is located within a region related to the pathogenicity island PAPI-1 in close proximity to a *pil* gene cluster, which has been associated with horizontal gene transfer. Besides *helP*, elevated protein levels of the bacterial flagellum protein FliL (hazard ratio 3.44, $p < 0.001$) and of a bacterioferritin-like protein (hazard ratio 1.74, $p = 0.003$) increased the risk of death, while high protein levels of a putative aminotransferase were associated with an improved outcome (hazard ratio 0.12, $p < 0.001$). The prognostic potential of biomarker candidates and clinical factors was confirmed with different machine learning approaches using training and hold-out datasets. The *helP* genotype appeared the most attractive biomarker for clinical risk stratification due to its relevant predictive power and ease of detection.

Introduction

Blood stream infections (BSI) are a frequent and often fatal occurrence in hospitalized patients, particularly under immunosuppression (ECDC 2015). According to the European Detailed Mortality Database (http://data.euro.who.int/dmdb/), more than 40,000 deaths in Europe can be attributed to sepsis in 2014. *Pseudomonas aeruginosa* is an important pathogen causing up to 15.4% of all BSI cases (ECDC 2015). Mortality rates of up to 42% even in advanced settings are reported (McCarthy and Paterson 2017), especially when appropriate antibiotic treatment is delayed (Skaar 2010).

The search for appropriate biomarkers is linked with the prospect of improving early diagnosis and prognosis prediction in sepsis. To date, C-reactive protein, interleukin-6 and procalcitonin are the only well-established diagnostic biomarkers, despite extensive evaluation of more than 100 biomarkers (Pierrakos and Vincent 2010). The majority of these biomarkers is host-derived. This is in line with the current paradigm of sepsis pathophysiology that explains lethal septic shock and multi-organ failure primarily as a result of the host's pro- and anti-inflammatory reaction to pathogen components like carbohydrate and fatty acids, termed pathogen-associated molecular patterns (PAMP) (Walton et al. 2014; Gotts and Matthay 2016). It is indeed well known that the genetic diversity in human genes encoding for pathogen recognition receptors as well as for pro- and anti-inflammatory mediators explains in part the variability in the clinical course of septic patients (Khor et al. 2007; Lehmann et al. 2009; Thompson et al. 2014). However, the role of the nature and characteristics of the infecting pathogen is frequently neglected (Lisboa et al. 2010; Angus and van der Poll 2013). Given the huge diversity of bacterial genomes and functional capacities even within one species, the pathogen itself could account for

unexplained heterogeneity in the clinical course and outcome of sepsis. Recently, the pangenome of the species *P. aeruginosa* was estimated to contain more than 16,000 non-redundant genes, while only 15% of these genes were present in all strains forming the core genome (Mosquera-Rendon et al. 2016). Of particular interest are prognostic bacterial biomarkers that can indicate the risk of a fatal outcome in septic patients, thus providing guidance in therapy and improved management of patient monitoring.

While bacterial virulence factors have been extensively explored in *P. aeruginosa* (Veesenmeyer et al. 2009), investigations have almost exclusively been carried out in *in vitro* experimentations or in animal models providing no evidence of their relevance and utility as prognostic biomarkers in humans. The type 3 secretion bacterial effector proteins ExoS, ExoT, and ExoU are an exception (Lisboa et al. 2010), with some authors reporting an association between expression level and sepsis outcome (El-Solh et al. 2012; Hattemer et al. 2013). In addition, one recent study presented evidence that the presence of the *exoU* gene is an independent predictor of early sepsis mortality in *P. aeruginosa* BSI (Pena et al. 2015).

In a multicenter study, we applied a multi-omics approach to identify pathogen factors that contribute to differential mortality outcomes in patients with *P. aeruginosa* bloodstream infection. We first used genomics and proteomics to characterize *P. aeruginosa* strains from sepsis patients. Next, we integrated these omics data from bacterial isolates with treatment- and patient-related data to gain a broader understanding of the complex interactions between host and pathogen during blood stream infections. We thereby screened for pathogen factors which are independently linked to 30-day mortality and which would consequently be attractive

prognostic biomarker candidates. Finally, we confirmed biomarker candidates identified by our statistical model using different machine learning algorithms.

Results

Clinical and patient-related risk factors for 30-day mortality

From 175 eligible patients, 166 (94.86%) patients with *P. aeruginosa* BSI were included into the final analysis (Figure S1). The basic demographic, clinical and infection-related characteristics of the patient study population are presented in table S1. An investigation of factors that had an impact on the mortality rate was initially performed on clinical and patient-related variables (Table S2). Multivariate Cox regression modelling revealed that immunosuppression as well as a rise in the SAPS II score increased 30-day mortality while administration of appropriate antibiotic treatment and a genitourinary infection source decreased the risk of a fatal outcome (Table S3).

Genomic characteristics of clinical *P. aeruginosa* strains

The genome of the first isolate recovered from each patient with a *P. aeruginosa* blood stream infection was sequenced (166 strains). The pangenome consisted of 23917 genes with 4354 core genes shared by > 99% of isolates, 639 soft core genes shared by 95% - 99%, 1762 shell genes shared by 15% - 95%, and 17162 cloud genes shared by < 15%. The high number of accessory genes underlines the stupendous genomic diversity and plasticity of *P. aeruginosa* species represented by our study dataset.

The phylogenetic tree based on the core genome SNP alignment shows a highly diverse population structure of our clinical isolates with distinct clades of similar branch length within two major phylogenetic clusters along with a discriminative cluster formed by only 3 strains (Figure 1). Recombination had occurred at a median rate of 0.07 (SNPs inside recombination/SNPs outside recombination, interquartile range: 0.03 - 0.27), demonstrating that recombination events have only slightly contributed to shaping the diversity of our strain set. Closely related isolates were most commonly found in only one hospital in a narrow time frame, providing evidence

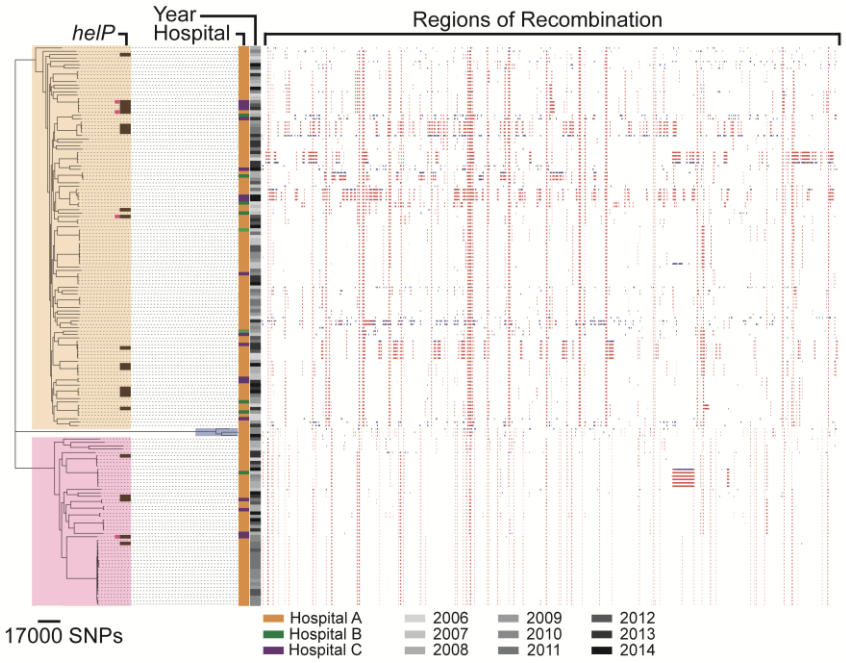

Figure 1. Core genome phylogenetic, temporal and spatial relationship of clinical *P. aeruginosa* strains recovered from patients with blood stream infection.

A maximum-likelihood phylogeny of the core genome SNP alignment reveals three major genomic clusters (Core-cluster 1 = bright yellow; Core-cluster 2 = light rose; Core-cluster 3 = grey). A high diversity within these clusters is reflected by numerous subgroups and distinct single isolates. Location and year of isolation is provided for each strain. Regions of predicted recombinations are shown by the right-sided panels of blocks. Red blocks indicate recombinations on internal branches, therefore shared by several strains through common descent. Blue blocks indicate recombinations that take place on terminal branches, thus are specific to individual isolates. Presence of the dead box helicase gene *helP* is shown by brown blocks beside the phylogenetic tree, with pink-colored squares that illustrate strains where *helP* location was predicted to be on a plasmid (plasmidSPAdes). The scale beneath demonstrates a distance of 17.000 point mutations.

for a spatial and temporal clustering. In some cases, strains from the same cluster have been isolated in different hospitals during the entire study period, suggesting either a transfer from one to the other hospital or a circulation of the particular strain within the community and sporadic reintroduction in our hospitals.

Antibiotic susceptibility profiles are shown in figure S2, demonstrating a wide range from broadly susceptible to extensively drug resistant (XDR) strains. XDR strains

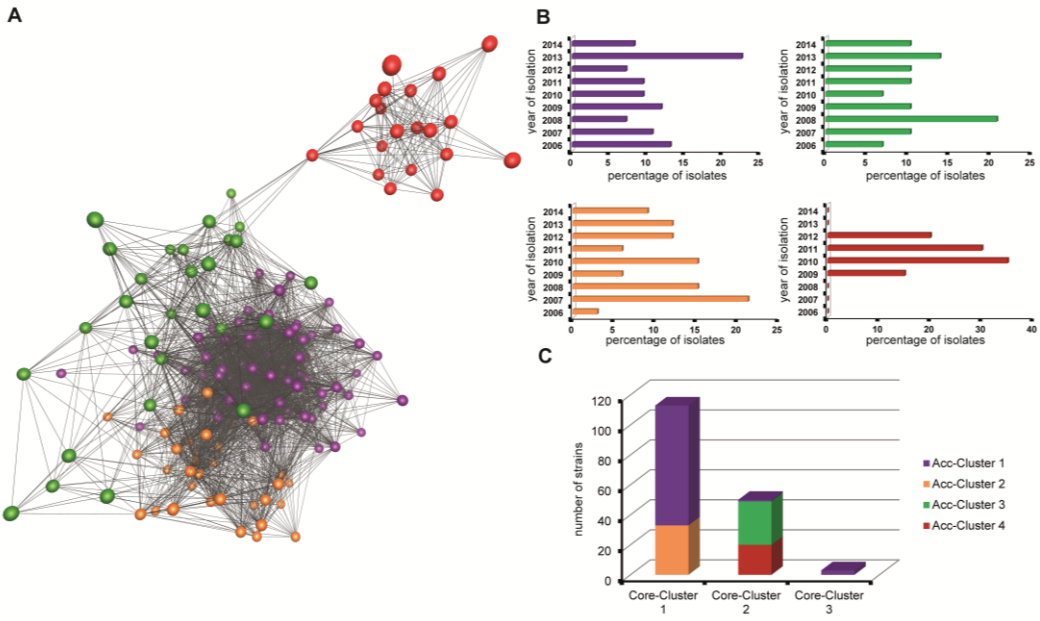

Figure 2. Genomic and temporal clustering of accessory genes and its link to core genome clusters.

(A) Population structure of the accessory genome using the Ward cluster algorithm with the Pearson similarity index and displayed as three-dimensional correlation network by Biolayout. Each node represents a study isolate, whereas the connecting edges reflect similarity based on the input gene presence-absence matrix. Four accessory genome clusters (acc-clusters) were revealed, coded by different colors (acc-cluster 1 = purple; acc-cluster 2 = orange; acc-cluster 3 = green, acc-cluster 4 = red). (B) The histograms illustrate the distribution of isolation time over the study period for each acc-cluster. All acc-clusters were evenly isolated with the exception of acc-cluster 4 which has only been found consecutively in four years (2009 - 2012). (C) The bars display overlaps between core- and acc-clusters. Core-cluster 1 and 2 were mainly partitioned in two distinct types of accessory genomes, hence rising evidence for a deeper structural genomic disparity than the one shown by the core genome maximum-likelihood phylogeny. Only one isolate from core-cluster 2 had an accessory genome that is grouped with acc-cluster 1. The three isolates of core-cluster 3 belong exclusively to acc-cluster 1, pointing to a closer genomic relation to core-cluster 1 than core cluster 2.

were only susceptible to colistin and were usually phylogenetically clustered, suggesting outbreak situations as previously described (Willmann et al. 2015).

In order to investigate the genomic relatedness of the accessory genome between the 166 strains, we only considered accessory genes with a prevalence ≥ 10% and ≤ 90%. Using this criterion, a subset of 2298 accessory genes was tested. Ward analysis revealed that the 166 isolates can be divided in four accessory genome (acc-) clusters (Figure 2A). Except for acc-cluster 4, which was confined to a 4-year period, the appearance of strains from the other acc-clusters was evenly distributed over time (Figure 2B). Acc-clusters showed a strong affiliation to the three major core-genome clusters, underlining a further structural distinction within the core genome phylogeny (Figure 2C). Acc-clusters were included in the clinical Cox regression model. The analysis showed that acc-cluster 2 was independently associated with 30-day mortality (HR 1.95, p = 0.048, Wald test), suggesting the presence of genomic pathogen factors that negatively influence patient survival (Table S4).

Subsequently, we investigated whether certain gene ontology (GO) terms and thereby gene functions are over- or under-represented in acc-cluster 2. Compared to the three remaining acc-clusters that served as a reference, acc-cluster 2 was enriched with the GO terms “peptidyl-histidine modification” (GO:0118202, FDR = 0.033) and “peptidyl-histidine phosphorylation” (GO:0018106, FDR = 0.033) (Figure S3). Both GO terms involve sensor histidine kinase genes that usually function in two-component systems. These bacterial regulatory systems, designed to sense external stimuli and to facilitate an appropriate adaptive response to stressors and changes in environmental and growth conditions, modulate the transcription of genes including virulence factors and antimicrobial resistance genes in *P. aeruginosa*

(Gooderham and Hancock 2009). Such systems could have a significant influence on a strain's survival chance during infection (Mikkelsen et al. 2011).

Protein level characteristics of clinical *P. aeruginosa* strains

After determining the genomic features, we next defined the cellular proteome of all 166 isolates. A total of 7757 unique proteins were identified in the proteomics analysis, with a subset of 1078 proteins (13.9%) synthesized by all study strains (core proteome). Principal component analysis of protein level profiles of the strains

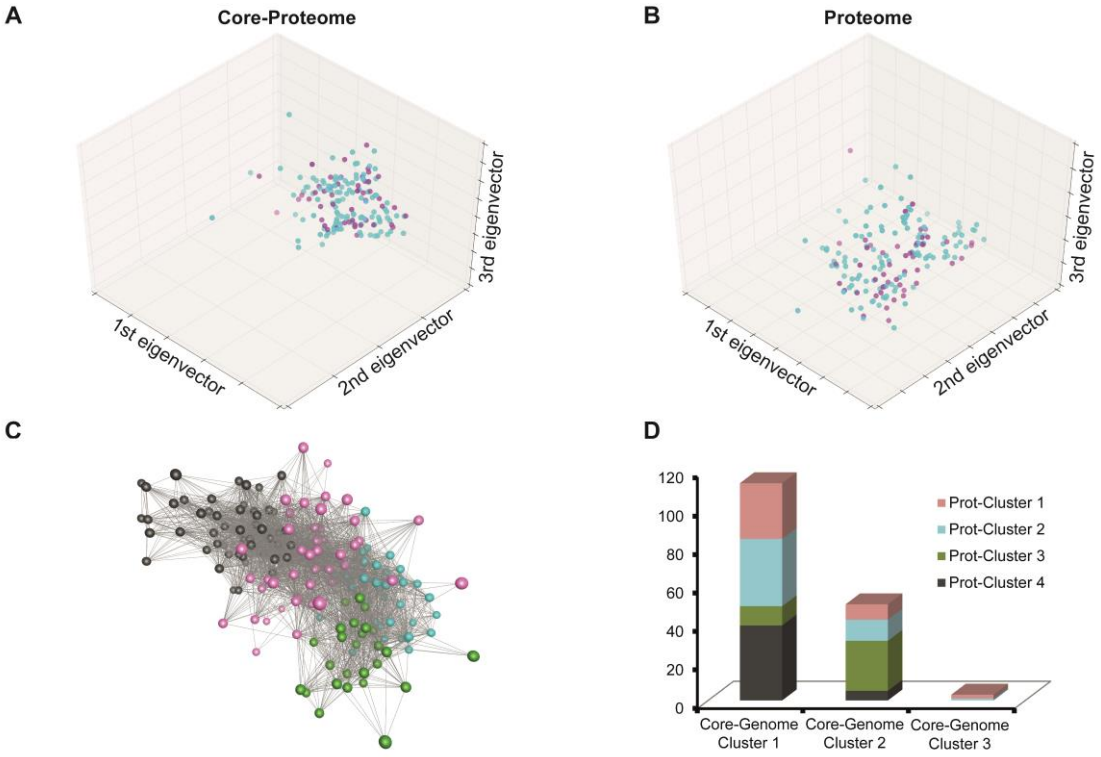

Figure 3. Basic proteome and core proteome characteristics and comparison with genomic features.

A principle component analysis with the first three eigenvectors is presented for the core proteome (A) and the whole proteome (B). Any data point stands for a strain isolated from survivors (blue) and decedents (purple). In both cases, no clustering according to survival status was observed. The first three eigenvectors comprised 57.1% of the overall variance for the core proteome, and 68.51% for the proteome. (C) Four core-proteome clusters were revealed, each of them representing a highly correlated expression profile. Visualization was done using Biolayout. Clusters are labelled as indicated by the color bar in D. (D) The fraction of the four identified core proteome clusters within the three core genome clusters showed the presence of all closely related protein level profiles in both the numerical greatest core genome clusters. This suggested that the protein level profiles are independent from the underlying core genome structure.

did not show clustering according to the survival status of patients, neither for the core-proteome (Figure 3A) nor for the whole proteome (Figure 3B). Ward cluster analysis of the core proteome resulted in four core proteome clusters (prot-clusters) of strains with closely related protein level profiles (Figure 3C). All prot-clusters were found in both the two numerically greatest core genome clusters (Figure 3D). This demonstrates that even phylogenetically distinct strains can share similar profiles of core protein levels. Since the presence of different accessory genes could also have an impact on the protein level pattern, we ordered the strains according to their acc-clusters and compared the protein levels of the core proteome (Figure S4A). We observed no distinct patterns associated with an acc-cluster or a strong relationship in a cross-comparison of prot- and acc-clusters (Figure S4B).

Inclusion of prot-clusters into the clinical Cox regression model did not reveal a link between prot-clusters and 30-day mortality (Table S4). These results suggest that the risk of a fatal outcome is not determined by complex core proteome clusters. But since individual protein levels could still have an impact, we performed a multi-level Cox regression analysis of single genomic and protein level factors on patient outcome.

Multi-level Cox regression and prediction model

Figure S5 provides an overview of the statistical workflow up to the final prediction model using a multi-level Cox regression analysis approach. We conducted an in-depth analysis of a pre-assigned accessory genome subset comprising 2298 genes and the natural log-normalized protein level data set of 1078 core proteins. Four pathogen-derived factors (one genomic and three proteomic factors) were independently associated with mortality in the final Cox regression prediction model (Table 1).

Table 1. Final multivariate Cox regression model including virulence factor candidates

Parameter	Annotation	Hazard ratio	95% CI	P-value
Appropriate antibiotic treatment	-	0.24	0.12 - 0.45	0.0001
SAPSII (index day)*	-	1.072	1.0465 - 1.0981	<0.0001
helP	DEAD/DEAH box helicase	2.01	1.03 - 3.94	0.05
log-Prot7*	FliL	3.44	1.67 - 7.07	0.0006
log-Prot214*	Bacterioferritin	1.74	1.18 - 2.57	0.003
log-Prot330*	Probable aminotransferase	0.12	0.03 - 0.44	0.0009

* Continuous variable. The hazard ratio reflects the increase/decrease in mortality risk per unit increase.
95% CI, 95% confidence interval; SAPSII, simplified acute physiology score II; Prot, protein
LFQ intensities were natural log-transformed for core proteomic data.

Prior screening model values of the four pathogen-derived predictors are presented in table S5, while all factors that were included into the multivariate models are shown in table S6. Two variables of the clinical Cox regression model were removed from the final and comprehensive Cox regression prediction model: immunosuppression due to a p-value > 0.05 and genitourinary infection source due to failing the internal bootstrap validation process. Three of the four pathogen-derived predictors shown in table 1 increase the risk of death (hazard ratio > 1), either when present in the bacterial genome (the genomic factor *helP*) or in case of high protein levels (proteins Prot7 and Prot214). All three were considered pathogen-derived risk factors for a fatal outcome.

In contrast, high levels of Prot330, a putative aminotransferase, turned out to have an anti-virulence effect (hazard ratio 0.12, 95% CI 0.03 - 0.44, p = 0.0023). Table S7 illustrates the functional annotation of genomic and proteomic predictors and their GO-terms from the UniProtKB database (http://www.uniprot.org/).

Besides protein levels and presence of genes, the existence of structural variations in putative pathogen virulence factors within the core genome set could have further contributed to mortality and was thus specifically explored. A total of 92 reported putative virulence factors were identified in the dataset; 59 of those were present in the core genome (Table S8). Using PAO1 (accession number: NC_002516.2) and PA14 (accession number: NC_008463.1) as genetic references, multivariate analyses of 107 parsimony-informative SNPs causing amino acid replacements detected only one candidate (LasA, A111V) linked to mortality (hazard ratio 2.18; 95% confidence interval 1.16 - 4.09; p = 0.012) in the variant screening model. However, this SNP candidate failed statistical significance in the final Cox regression prediction model (p = 0.28), suggesting that structural variations in putative pathogen virulence factors did not impact 30-day mortality in our study population.

In-depth characterization of the prognostic biomarker candidate *helP*

One of the identified prognostic biomarker candidates, the 1866th accessory genome gene that we named *helP* (GenBank accession number: KY940721), is particularly interesting. The presence of *helP* in the genome of the study strains was estimated to double the risk of a fatal outcome (hazard ratio 2.01, 95% confidence interval 1.03 - 3.94, p = 0.05). Its gene product is highly similar to RL063, a protein whose gene sequence is located on the pathogenicity island I in PA14 (98.3% protein sequence similarity, UniProt accession number Q7WXZ7). The amino acid sequence of HelP is identical to a protein named PSPA7_4493 (UniProt accession number: A6V9V7), a predicted DEAD/DEAH box helicase from the *P. aeruginosa* strain PA7. This prediction is primarily based on a helicase conserved C-terminal domain (PF00271, domain boundary positions: 570 - 685) and a DEXDc domain (SM00487, domain boundary positions: 57 - 410). *helP* appeared in 22 of our study strains, but much

more frequently in the high-risk acc-cluster 2 strains (27.27% in acc2 vs 9.77% in the other three clusters, p = 0.008, chi-squared test). A maximum-likelihood phylogeny showed that HelP groups together with other predicted helicases from Pseudomonas sp., thereby most closely related to the class of DEAD-box helicases within the superfamily 2 (Figure S6, table S9).

Generally, *helP* was evenly distributed among all strains in the core phylogeny and was found in different hospitals (Figure 1), reflecting the gene's integration in many different phylogenetic groups rather than just in one. We hypothesized that *helP* might be transmittable via horizontal gene transfer, which is another important aspect apart from virulence capabilities. DEAD/DEAH box helicases are non-essential bacterial genes that might be acquired through horizontal gene transfer. The recycler tool (Rozov et al. 2016) and plasmidSPAdes (Antipov et al. 2016) were used to predict plasmids from the short Illumina sequence reads of all *helP* positive strains. Plasmids predicted by plasmidSPAdes harbored *helP* in strain ID 26, ID 93, ID 101, and ID 138 (Figure 1) while plasmids predicted by the recycler tool did not. In the remaining strains, *helP* location was predicted to be on the bacterial chromosome. Since genome assembly from short reads can be prone to errors, particularly in the detection and characterization of mobile genetic elements, we sequenced the four strains including strain ID50 on a PacBio instrument to improve assembly quality. In all strains, *helP* was located on a contig with a size > 800 kb. This makes a location on a plasmid very unlikely, indicating that plasmidSPAdes provided a false positive rating.

Nevertheless, *helP* can be found in 12 strains that originate from four different phylogentic clusters as well as in 10 strains that are genetically distinct from all other *helP* positive strains. This genomic diversity of *helP* positive strains suggests a

horizontal transfer of *helP* in the past. Interestingly, the genomic environment of *helP* on the five large PacBio contigs resembled the pathogenicity island PAPI-1 from PA14 (Figure S7), where a homologous gene of *helP* is located (RL063). Particularly upstream of *helP*, we found PAPI-1 well conserved. Of special interest is the 10-gene cluster of a type IV pilus (T4P) apparatus that is located in close proximity to *helP* (Figure S7). This T4P system has been described as a conjugative apparatus genetically closely related to genes on the enterobacterial plasmid R64. The system has been reported of being capable of transferring PAPI-1 into recipient *P. aeruginosa* (Carter et al. 2010). When mapping the short Illumina sequence reads of the 22 *helP* positive strains against PAPI-1, we found a similar picture with a few differences between single strains and clusters (Figure S8). The structure of this genomic environment suggests a past exchange of *helP* between different *P. aeruginosa* strains via conjugation machineries but not via plasmids.

It was recently reported that a RNA helicase (Uniprot accession: Q9I003) in *P. aeruginosa* affected expression of ExoS (Tan et al. 2016). For this reason, we explored protein levels of known *P. aeruginosa* exotoxins, secretion system effectors and factors of the T4P in all clinical isolates (Table S10). Strains that were positive for *helP* had a 6.57-fold higher expression of *exoU* compared to *helP* negative strains ($p = 0.04$), but were not distinct in their ExoS levels. This is likely due to different structures of both helicases. In a pairwise alignment comparison, HelP had only a 20% amino acid sequence similarity with the respective RNA helicase (Uniprot accession: Q9I003), suggesting that both putative helicases do not necessarily operate with the same mode of action. However, our findings indicate a potential connection between the *helP* genotype and ExoU expression, and could play a role

in sepsis when considering the clinical impact of the *exoU* genotype (Pena et al. 2015).

Predictive performance of identified prognostic factors in machine learning algorithms

The following datasets were submitted for further evaluation using different machine learning strategies: datasets including all features of the three screening models (genomic, phenotypic, or SNP features), a dataset containing all pathogen-derived factors from the screening models ("ALL") and the dataset with the variables from the final Cox regression model ("Final"). All datasets consisted of the clinical predictors identified in the clinical Cox regression model (Table S3).

Performance specifications of the estimators from each tested dataset are presented in supplement table S11. Values for the area under the receiver operating characteristic curve (AUC) indicate each estimator's potential to discriminate patients at high risk of a fatal outcome from those with a lower risk. In most cases, AUC values were below 0.8, indicating a rather weak discriminatory power of the estimators. Exceptions were estimators from the dataset of the final cox regression model which gained higher AUC values compared with the estimators from the other datasets (median AUC 0.829 vs 0.736, $p = 0.0009$). The best estimator from the dataset of the final cox regression model was a linear support vector classifier that showed no sign of overfitting in its learning curve (Figure 4A). In contrast, most estimators were prone to overfitting and were difficult to regularize. However, using the best 5% of features generally increased performance significantly and often removed overfitting. This indicates a high background of uninformative features that disturb the predictive potential of the estimators. It underlines the importance of

feature selection in datasets with a high number of features and a comparative smaller number of instances, as is usually the case in multi-omics studies.

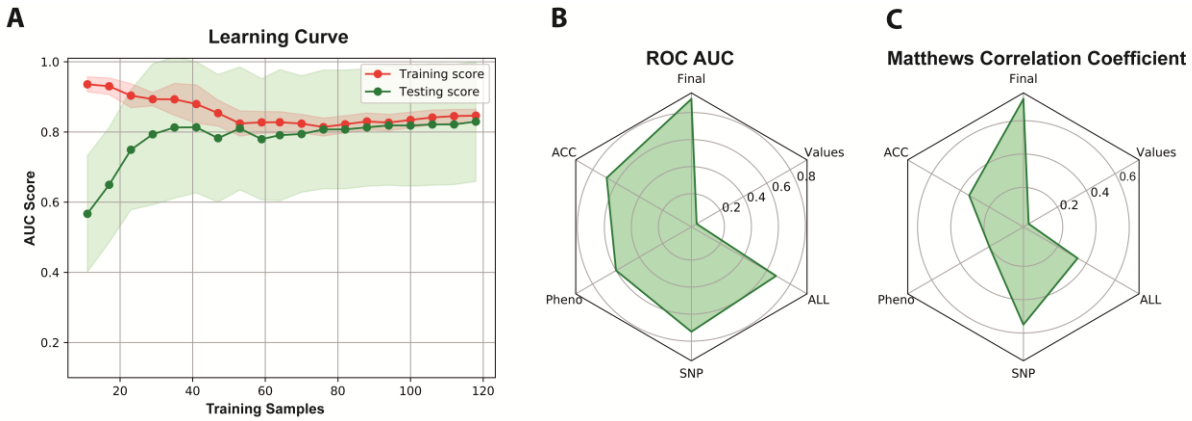

Figure 4. Machine learning estimator validation of the multi-omics datasets

A) The learning curve of the linear support vector machine estimator from the dataset containing the features from the final Cox regression model ("Final") determined cross-validated training and test area under the receiver operating characteristic curves scores (AUC scores) for different training set sizes. The faded areas indicate standard deviations of the respective test scores. Both curves approach each other with increasing training set size, reflecting evidence against overfitting. (B) The radial graph demonstrates the area under the receiver operating characteristic curve (ROC AUC) values of the final validation step, when the best estimators from each dataset were tested against the hold-out data. The prediction performance from estimators of the following datasets was examined: ACC = accessory genome features, Pheno = protein levels and antibiotic susceptibility features, SNP = single nucleotide polymorphisms in reported virulence factors, ALL = combination of the three models above. These datasets included the clinical risk factors from the respective clinical Cox regression model. "Final" marks the estimator that consists of the dataset with the features from the final Cox regression model. The label "Values" does not indicate a dataset but the AUC values at the grid in the radial graph. The estimator from the Final dataset showed superior performance (AUC = 0.895), particularly over the Pheno dataset estimator which performed quite weak (AUC = 0.595). (C) Matthews correlation coefficients are provided for the same estimators as mentioned in B. The label "Values" indicates the correlation coefficients but not a dataset. Again, the estimator from the Final dataset shows the highest coefficient (0.726), reflecting a high number of correct predictions.

For each dataset, we subsequently evaluated the most promising estimator on the hold-out dataset that contained 20% of the total data (Table S12). The estimator trained on the dataset with the features from the final Cox regression model was clearly superior to the estimators of the other datasets (AUC = 0.895, Matthews correlation coefficient 0.726, Figure 4B and 4C). It was the only model that predicted all fatal cases correctly and thus did not miss one patient at high risk of a fatal outcome (Table S12). This suggests a significant relation between the identified

clinical risk factors and prognostic biomarker candidates with fatal outcome. It also demonstrates that a feature selection based on classical epidemiological methodology (here Cox regression) can be a powerful tool when combined with machine learning algorithms in multi-omics approaches.

Discussion

Our approach integrated genomic and proteomic pathogen data into a clinical multicenter cohort study with a wide range of sepsis conditions. This was followed by an evaluation of promising prognostic biomarker candidates using machine learning. Our results support that certain *P. aeruginosa* pathogen factors significantly contribute to the risk of a fatal outcome in bloodstream infections independently of the physiological patient status and administration of appropriate antibiotic therapy.

In terms of a more detailed characterization, we have focused on the genomic candidate *helP* since its presence increased the risk of death by two-fold in our study and since it can be easily measured in a routine diagnostic setting (e.g. by PCR), which makes it an interesting prognostic biomarker. Moreover, its genomic environment and its detection in different genetic lineages suggest that *helP* has been acquired by horizontal gene transfer.

We found *helP* in close genomic proximity to a type IV pili (T4P) gene complex within a genetic environment that is similar to the pathogenicity island I of PA14. These T4P systems are involved in motility and adhesion to host cells during infection (Hahn 1997; Bieber et al. 1998). T4P-deficient *P. aeruginosa* mutants were reported to have a lower cytotoxicity, potentially due to the loss of cell contact and therefore an inefficiently working type III secretion system (T3SS) (Comolli et al. 1999), whose

importance as a prognostic biomarker in *P. aeruginosa* bacteremia has been repeatedly shown (El-Solh et al. 2012; Hattemer et al. 2013; Pena et al. 2015). Besides the possibility of an interaction of *helP* with its flanking T4P-system, there might be other mechanisms involved in how DEAD-box helicases could affect virulence. Tan et al. have reported on a DEAD-box helicase of *P. aeruginosa* that was essential for virulence and bacterial cytotoxicity in a mouse pneumonia model (Tan et al. 2016). Deletion of the DEAD-box helicase resulted in significantly lower expression levels of the T3SS effector protein ExoS and in a decreased production of proinflammatory cytokines and neutrophil infiltration in infected mice. However, we did not observe a differential expression of ExoS in *helP* positive isolates, but of ExoU levels, suggesting another potential linkage with the T3SS effectors.

Protein level analysis has also identified putative virulence and anti-virulence factors in *P. aeruginosa* bloodstream infection. Although strains were immediately conserved after detection and protein levels were determined in the first subculture after thawing, and therefore close to the conditions in the blood culture bottle, it is unknown how such protein level profiles would mirror pathogen protein levels in a patient's bloodstream. Because of this limitation, we focused on *helP* as genomic biomarker due to its stability even under different pre-analytical conditions. We also hypothesized that protein level analysis can be a valuable tool in detecting additional virulence markers. Thus, we included factors arisen from this phenotype screening model into our final Cox regression prediction model.

High protein level of the flagellum basal body protein FliL was associated with increased mortality. The flagellum apparatus has been widely reported to be vital for virulence in *P. aeruginosa* (Kazmierczak et al. 2015). It is mainly important for swimming motility and attachment to host cells. Flagella components are known to

bind to Toll-like receptor 5, thereby activating a mostly proinflammatory immune response (Zhang et al. 2003). During chronic infection in cystic fibrosis patients, flagellum expression is often downregulated to reduce inflammation (Mahenthiralingam et al. 1994). Our observation, together with the reported success of an anti-flagella vaccine in a clinical trial (Doring et al. 2007), makes the flagellum apparatus an interesting target for therapeutic virulence blockers in sepsis.

Another prognostic biomarker candidate is Prot214 which is annotated as bacterioferritin. Its role as risk factor for fatal outcome remains elusive. It is well known that an important line of defense against bacterial infection is the withholding of free iron since bacterial pathogens essentially depend on iron for replication and their pathogenic actions. In order to ensure sufficient iron levels in the bacterial cytosol and to also prevent iron-induced toxicity, cellular levels of free iron need to be highly regulated. In *P. aeruginosa*, two ferritin-like molecules are known to store iron intracellularly (bacterial ferritin A and bacterioferritin B) and both are considered obligatory for iron metabolism (Rivera 2017). These iron stores, suggested to be an important source for the heme prosthetic group of KatA, can increase resistance against hydrogen peroxide (Ma et al. 1999). This could be crucial for the rapid adaptation of invasive strains to new environments like the human blood and could augment pathogen survival against innate immune defense mechanisms. Nonetheless, the function of the bacterioferritin-like protein Prot214 as well as the anti-virulence capacity of the putative aminotransferase Prot330 during bacteraemia needs to be further investigated. This also applies to the discovery that GO-terms for peptidyl-histidine phosphorylation as part of two-component systems were enriched in the high risk accessory genome group (acc-cluster 2). The enrichment could reflect an improved ability for an immediate response to external stimuli. This could be

advantageous in terms of a pronounced growth of invasive strains in different human body sites.

We conducted a systematic search for prognostic biomarker candidates in patients with *P. aeruginosa* bloodstream infection. Routine detection of highly virulent strains could result in administering high-dose combination therapy to those patients who need it most, providing a fair rationale for the additive toxicity. This is especially the case in *P. aeruginosa* bloodstream infection where combination therapy is thought to be superior over monotherapy, particularly when patients are at a higher risk of fatal outcome (Safdar et al. 2004; Park et al. 2012; Kim et al. 2014). Beside therapeutic management, detection of high-risk strains could also be followed by infection control measures like contact isolation. This would allow a reduction in the spread of virulent strains, an objective that is neglected by current infection control guidelines that tend to focus solely on a strain's antibiotic susceptibility profile. Such practices could help in significantly reducing the more than 40,000 annual sepsis deaths alone in Europe and the hundred thousands more throughout the world. Our multi-omics approach has produced genomic and proteomic data identifying pathogen-derived prognostic biomarker candidates that are interwoven with treatment- and patient-related risk factors in a complex interplay. Our results reveal the importance of multi-omics approaches, which allow us to investigate multiple pathogen and host factors at the same time. Future studies that aim to validate these findings and to move confirmed pathogen-derived prognostic biomarkers into clinics are warranted.

Methods

Setting

We conducted a multicenter genomic cohort study in a 1500-bed tertiary teaching hospital, a 500-bed district hospital, and a 300-bed trauma center in Tübingen, Germany, and the surrounding community. The broad spectrum of medical services provided by these hospitals includes multiple medical and surgical specialties, pediatric units, dialysis and a maternity ward. Organ transplantations are carried out at the tertiary teaching hospital. The study is reported pursuant to the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) and Strengthening the reporting of Genetic Risk Prediction Studies (GRIPS) statements (Janssens et al. 2011; Collins et al. 2015). Our study was approved by the local research ethics committee of the University of Tübingen (reference number: 364/2013R).

Study design, patients and definitions

Adult patients (≥ 18 years) admitted to one of the participating hospitals were considered eligible when they were suffering from a blood stream infection (BSI) with ≥ 1 blood culture positive for *P. aeruginosa*. Patients were included once at the time of the first positive blood culture (index culture). Thirty-day mortality for any cause was the clinical endpoint while patient- and pathogen-related factors were regarded potential predictors of outcome.

Relevant patient data variables were defined as follows: site of infection (primary, secondary, vascular catheter-related) as classified by the International Sepsis Forum (Calandra and Cohen 2005); Charlson comorbidity score (Charlson et al. 1987); nosocomial infection (any infection ≥ 48 hours after hospital admission);

immunosuppression (HIV and/or neutropenia with a neutrophil count ≤ 1000 cells/µl, and/or immunosuppressive chemotherapy within the previous two months and/or receipt of prednisolone ≥ 10 mg/daily or equivalent steroid dose). The physiological patient condition was determined using the simplified acute physiology score II (SAPS II) at the index culture day (Le Gall et al. 1993). A systemic administration of at least one antimicrobial agent to which the isolate was susceptible *in vitro* was defined an appropriate antimicrobial treatment.

Species identification and susceptibility testing

Species identification was carried out using MALDI-TOF mass spectrometry and the Vitek 2 system (bioMérieux, Marcy l'Etoile, France). Minimum inhibitory concentrations were assessed by antibiotic gradient strips (MIC Test Strip, Liofilchem, Italy) and interpreted according to EUCAST breakpoints (version 8.0, 2018).

Genomic data acquisition and analysis

Genomic DNA of *P. aeruginosa* has been sheared into 450 bp fragments using a focus-ultrasonicator (Covaris, Woburn, USA). Preparation of DNA libraries was done with the NEXTflex^{TM} DNA Sequencing Kit (Bioo Scientific, Austin, USA) followed by sequencing at 2 x 125 bp on a HiSeq2500 platform (Illumina, San Diego, USA). SPAdes (version 3.7.0) has been selected as *de novo* assembly tool and assembled scaffolds were annotated using Prokka (version 1.11) (Bankevich et al. 2012; Seemann 2014). A gene presence-absence-matrix of all study isolates has been generated by Roary (version 1.006924) using a 95% minimum percentage identity for blastp (Page et al. 2015). Structure of the accessory genome was assessed using the Ward cluster algorithm with the Pearson similarity index (Stata version 12.1, Stat

Corp., College Station, USA). Clustering was visualized using Biolayout (Theocharidis et al. 2009). Core genome construction was performed with Spine (version 0.1.2), and SNP were called using samtools (version 0.1.19) and GATK tools (version 3.2-2) (Li et al. 2009; Van der Auwera et al. 2013). The core genome maximum-likelihood phylogeny was generated using Gubbins (version 2.1.0) to account for genomic regions which had undergone homologous recombination (Croucher et al. 2015). A maximum of 10 iterations were used and a generalized time reversible (GTR) substitution model with a gamma distribution of rates. Gene ontology (GO) term enrichment analysis was conducted with Blast2GO (version 4.0.7) after genes from each group were clustered using CD-HIT-EST (version 4.6) with a 90% similarity threshold to remove redundancies (Conesa et al. 2005; Fu et al. 2012). The analysis was performed using a two-tailed Fisher's exact test. The maximum false discovery rate (FDR) was set to 0.05 for reporting significant GO-terms and the Benjamini-Hochberg correction was used. Genomes from five strains were further determined on a PacBio *RS* II instrument. Each strain was sequenced in one SMRT cell resulting in coverage rates between 9-fold and 149-fold. Assembly of PacBio long reads was done using Canu version 1.5 (Koren et al. 2017), and contigs were subsequently polished with Pilon (version 1.22) to improve accuracy (Walker et al. 2014). For ID50, due to the low coverage of 9-fold, we used the SPAdes assembler version 3.9.0 (Bankevich et al. 2012) with Illumina short reads and the --pacbio option. With this hybrid approach, we improved N50 statistic from 255,540 bp to 634,760 bp in this strain.

Proteomic data acquisition and analysis

P. aeruginosa strains were grown overnight, and proteins were extracted as described elsewhere (Krug et al. 2013). Protein extracts were precipitated overnight

with acetone and approximately 10 µg were loaded onto a NuPAGE Bis-Tris 4-12 % gradient gel (Thermo Fisher Scientific, Waltham, USA). Samples were let run approximately 10 mm into the gel and cut out as a single slice. In-gel digestion and peptide extraction were performed essentially as described previously (Borchert et al. 2010). Peptides were desalted using C18 StageTips (Rappsilber et al. 2007). LC-MS/MS analyses were performed on an EasyLC II nano-HPLC coupled to an LTQ Orbitrap Elite mass spectrometer (Thermo Fisher Scientific, Waltham, USA). LTQ Orbitrap Elite was operated in the positive ion mode. Samples were randomized before injection and a custom-made standard was measured in regular intervals to assess long-term performance of the MS.

Acquired MS spectra were processed with MaxQuant software package (version 1.5.2.8), with integrated Andromeda search engine (Cox and Mann 2008; Cox et al. 2011). Database search was performed against a *P. aeruginosa* database obtained from Uniprot (all strains), containing 103,188 protein entries, together with the custom-made database containing 30 additional entries which were not represented in the main database. Trypsin (full specificity) was set as the protease and the maximum number of missed cleavages was set to two. False discovery rate of 0.01 was set at the peptide and protein level. The label-free algorithm was enabled and a minimum of two unmodified peptide counts were required for quantification. Core proteome architecture was explored by using the Ward cluster algorithm with the correlation coefficient index (Stata version 12.1, Stat Corp., College Station, USA)

Statistical analysis for virulence candidate assessment

A multi-level Cox regression analysis was applied to study the association between exposure (patient characteristics, geno- and phenotype of the pathogen) and the study endpoint (30-day all-cause mortality). Prior to testing, a variance range was set

for all binary variables of interest to reduce dimensionality. Only those variables with a frequency ≥ 10% and ≤ 90% were tested. The final Cox regression model was built in a stepwise procedure. First, patient characteristics were individually tested and any variable with a p-value of < 0.2 was included in a multivariate model, wherein only variables with a p-value of ≤ 0.05 were retained, generating the clinical Cox regression model. In a second step, pathogen-related features were individually incorporated into the clinical Cox regression model. This led to three different screening models which integrated each one of the following information: accessory genome information (accessory genome gene screening model), phenotypic properties (natural log-transformed LFQ intensities of the core proteome and MICs = phenotypic screening model) or information about SNPs in known virulence factors (variant screening model) (Table S5). For SNPs, linkage disequilibrium was assessed and an $R^2 > 0.8$ led to grouping and testing of one representative SNP for each group.

Integration of pathogen-related variables from each of the three screening models into the final multivariate model had to run through two selection processes: (i) Within each screening model, variables must have had a $p < 0.05$ and (ii) must have belonged to the 10% of variables with the lowest p-value from that model. In the final Cox regression prediction model, variables with a $p \leq 0.05$ were retained. Hypothesis testing was performed by using the likelihood ratio test. The proportional hazards assumption was verified on the basis of Schoenfeld residuals. The final Cox regression prediction model was internally validated by bootstrapping (10000 replicates) and the jackknife method. Computations were done using Stata version 12.1 (Stat Corp., College Station, USA).

Machine learning estimator search and optimization

The following were submitted for further evaluation using different machine learning approaches: datasets containing all features of the three screening models, a dataset containing the clinical risk factors and all pathogenic factors from the screening models (ALL) and the dataset with the variables from the final Cox regression prediction model (Final). For each algorithm, 30-day mortality was the outcome variable, and a model's ability to predict the risk of a fatal case was assessed through a receiver operating characteristics analysis (ROC) measuring the area under the ROC curve (AUC) and through Matthews correlation coefficient. The scikit-learn toolbox version 0.19.1 was used for all calculations (http://scikit-learn.org/stable/). The following classification algorithms were tested on each dataset: random forest, support vector classifier, linear support vector classifier, k-nearest neighbour, and multi-layer perceptron.

The best estimator was searched on a training dataset that was comprised of 80% of the whole dataset. On each training dataset, we performed (i) no feature modification, (ii) dimensionality reduction using a principle component analysis with a maximum of 100 components and (iii) feature selection of the 5% features with the lowest univariate p-values. An exception was the dataset with the features from the final Cox regression prediction model where all calculations were only performed on the unaltered set of features. Hyperparameter tuning was conducted using the exhaustive grid search function (GridSearchCV), and estimator performance was evaluated by a ten-fold cross-validation. Here, the training set was split into 10 smaller sets. Subsequently, a model was trained on 9 folds of the training data and validated on the remaining part. The reported performance was the average of values computed in the loop. Potential over- and underfitting was determined by learning

curves of the training datasets. Best estimators were finally evaluated on a hold-out dataset, which contained data the estimator has not seen before (remaining 20% of the data). This is to assess the likely “real-world” performance of the model estimator.

Declarations

Acknowledgements

We thank the directors, physicians, laboratory and nursing staff of the medical wards in all participating hospitals. We also extend our gratitude to Kerstin Fischer, Nadine Hoffmann, Sara Riedel-Christ, Silke Wahl, and Irina Droste-Borel for their excellent technical support. Our work was partially funded by the AKF fund of the University Hospital Tübingen (project number: E.03.43003) and the German Center for Infection Research (project number: TTU 08.702). The funders had no role in study design, data collection and analysis, in decision to publish, or preparation of the manuscript.

Disclosure declaration

The authors declare that they have no competing interests.

References

Angus DC, van der Poll T. 2013. Severe sepsis and septic shock. *The New England journal of medicine* **369**(21): 2063.

Antipov D, Hartwick N, Shen M, Raiko M, Lapidus A, Pevzner PA. 2016. plasmidSPAdes: assembling plasmids from whole genome sequencing data. *Bioinformatics (Oxford, England)* **32**(22): 3380-3387.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD et al. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology : a journal of computational molecular cell biology* **19**(5): 455-477.

Bieber D, Ramer SW, Wu CY, Murray WJ, Tobe T, Fernandez R, Schoolnik GK. 1998. Type IV pili, transient bacterial aggregates, and virulence of enteropathogenic Escherichia coli. *Science* **280**(5372): 2114-2118.

Borchert N, Dieterich C, Krug K, Schutz W, Jung S, Nordheim A, Sommer RJ, Macek B. 2010. Proteogenomics of Pristionchus pacificus reveals distinct proteome structure of nematode models. *Genome research* **20**(6): 837-846.

Calandra T, Cohen J. 2005. The international sepsis forum consensus conference on definitions of infection in the intensive care unit. *Crit Care Med* **33**(7): 1538-1548.

Carter MQ, Chen J, Lory S. 2010. The Pseudomonas aeruginosa pathogenicity island PAPI-1 is transferred via a novel type IV pilus. *Journal of bacteriology* **192**(13): 3249-3258.

Charlson ME, Pompei P, Ales KL, MacKenzie CR. 1987. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis* **40**(5): 373-383.

Collins GS, Reitsma JB, Altman DG, Moons KG. 2015. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD). *Annals of internal medicine* **162**(10): 735-736.

Comolli JC, Hauser AR, Waite L, Whitchurch CB, Mattick JS, Engel JN. 1999. Pseudomonas aeruginosa gene products PilT and PilU are required for cytotoxicity in vitro and virulence in a mouse model of acute pneumonia. *Infection and immunity* **67**(7): 3625-3630.

Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics (Oxford, England)* **21**(18): 3674-3676.

Cox J, Mann M. 2008. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature biotechnology* **26**(12): 1367-1372.

Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen JV, Mann M. 2011. Andromeda: a peptide search engine integrated into the MaxQuant environment. *Journal of proteome research* **10**(4): 1794-1805.

Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, Parkhill J, Harris SR. 2015. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic acids research* **43**(3): e15.

Doring G, Meisner C, Stern M, Flagella Vaccine Trial Study G. 2007. A double-blind randomized placebo-controlled phase III study of a Pseudomonas aeruginosa

flagella vaccine in cystic fibrosis patients. *Proceedings of the National Academy of Sciences of the United States of America* **104**(26): 11020-11025.

ECDC. 2015. Annual epidemiological report 2014. Antimicrobial resistance and healthcare-associated infections. ECDC, Stockholm.

El-Solh AA, Hattemer A, Hauser AR, Alhajhusain A, Vora H. 2012. Clinical outcomes of type III Pseudomonas aeruginosa bacteremia. *Crit Care Med* **40**(4): 1157-1163.

Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics (Oxford, England)* **28**(23): 3150-3152.

Gooderham WJ, Hancock RE. 2009. Regulation of virulence and antibiotic resistance by two-component regulatory systems in Pseudomonas aeruginosa. *FEMS microbiology reviews* **33**(2): 279-294.

Gotts JE, Matthay MA. 2016. Sepsis: pathophysiology and clinical management. *Bmj* **353**: i1585.

Hahn HP. 1997. The type-4 pilus is the major virulence-associated adhesin of Pseudomonas aeruginosa--a review. *Gene* **192**(1): 99-108.

Hattemer A, Hauser A, Diaz M, Scheetz M, Shah N, Allen JP, Porhomayon J, El-Solh AA. 2013. Bacterial and clinical characteristics of health care- and community-acquired bloodstream infections due to Pseudomonas aeruginosa. *Antimicrobial agents and chemotherapy* **57**(8): 3969-3975.

Janssens AC, Ioannidis JP, van Duijn CM, Little J, Khoury MJ, Group G. 2011. Strengthening the reporting of Genetic Risk Prediction Studies: the GRIPS statement. *Genetics in medicine : official journal of the American College of Medical Genetics* **13**(5): 453-456.

Kazmierczak BI, Schniederberend M, Jain R. 2015. Cross-regulation of Pseudomonas motility systems: the intimate relationship between flagella, pili and virulence. *Current opinion in microbiology* **28**: 78-82.

Khor CC, Chapman SJ, Vannberg FO, Dunne A, Murphy C, Ling EY, Frodsham AJ, Walley AJ, Kyrieleis O, Khan A et al. 2007. A Mal functional variant is associated with protection against invasive pneumococcal disease, bacteremia, malaria and tuberculosis. *Nature genetics* **39**(4): 523-528.

Kim YJ, Jun YH, Kim YR, Park KG, Park YJ, Kang JY, Kim SI. 2014. Risk factors for mortality in patients with Pseudomonas aeruginosa bacteremia; retrospective study of impact of combination antimicrobial therapy. *BMC infectious diseases* **14**: 161.

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome research* **27**(5): 722-736.

Krug K, Carpy A, Behrends G, Matic K, Soares NC, Macek B. 2013. Deep coverage of the Escherichia coli proteome enables the assessment of false discovery rates in simple proteogenomic experiments. *Molecular & cellular proteomics : MCP* **12**(11): 3420-3430.

Le Gall JR, Lemeshow S, Saulnier F. 1993. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. *JAMA* **270**(24): 2957-2963.

Lehmann LE, Book M, Hartmann W, Weber SU, Schewe JC, Klaschik S, Hoeft A, Stuber F. 2009. A MIF haplotype is associated with the outcome of patients with severe sepsis: a case control study. *Journal of translational medicine* **7**: 100.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* **25**(16): 2078-2079.

Lisboa T, Waterer G, Rello J. 2010. We should be measuring genomic bacterial load and virulence factors. *Crit Care Med* **38**(10 Suppl): S656-662.

Ma JF, Ochsner UA, Klotz MG, Nanayakkara VK, Howell ML, Johnson Z, Posey JE, Vasil ML, Monaco JJ, Hassett DJ. 1999. Bacterioferritin A modulates catalase A (KatA) activity and resistance to hydrogen peroxide in Pseudomonas aeruginosa. *Journal of bacteriology* **181**(12): 3730-3742.

Mahenthiralingam E, Campbell ME, Speert DP. 1994. Nonmotility and phagocytic resistance of Pseudomonas aeruginosa isolates from chronically colonized patients with cystic fibrosis. *Infection and immunity* **62**(2): 596-605.

McCarthy KL, Paterson DL. 2017. Long-term mortality following Pseudomonas aeruginosa bloodstream infection. *The Journal of hospital infection* **95**(3): 292-299.

Mikkelsen H, Sivaneson M, Filloux A. 2011. Key two-component regulatory systems that control biofilm formation in Pseudomonas aeruginosa. *Environmental microbiology* **13**(7): 1666-1681.

Mosquera-Rendon J, Rada-Bravo AM, Cardenas-Brito S, Corredor M, Restrepo-Pineda E, Benitez-Paez A. 2016. Pangenome-wide and molecular evolution analyses of the Pseudomonas aeruginosa species. *BMC genomics* **17**: 45.

Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, Fookes M, Falush D, Keane JA, Parkhill J. 2015. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics (Oxford, England)* **31**(22): 3691-3693.

Park SY, Park HJ, Moon SM, Park KH, Chong YP, Kim MN, Kim SH, Lee SO, Kim YS, Woo JH et al. 2012. Impact of adequate empirical combination therapy on

mortality from bacteremic Pseudomonas aeruginosa pneumonia. *BMC infectious diseases* **12**: 308.

Pena C, Cabot G, Gomez-Zorrilla S, Zamorano L, Ocampo-Sosa A, Murillas J, Almirante B, Pomar V, Aguilar M, Granados A et al. 2015. Influence of virulence genotype and resistance profile in the mortality of Pseudomonas aeruginosa bloodstream infections. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* **60**(4): 539-548.

Pierrakos C, Vincent JL. 2010. Sepsis biomarkers: a review. *Critical care* **14**(1): R15.

Rappsilber J, Mann M, Ishihama Y. 2007. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nature protocols* **2**(8): 1896-1906.

Rivera M. 2017. Bacterioferritin: Structure, Dynamics, and Protein-Protein Interactions at Play in Iron Storage and Mobilization. *Accounts of chemical research* **50**(2): 331-340.

Rozov R, Brown Kav A, Bogumil D, Shterzer N, Halperin E, Mizrahi I, Shamir R. 2016. Recycler: an algorithm for detecting plasmids from de novo assembly graphs. *Bioinformatics (Oxford, England)*.

Safdar N, Handelsman J, Maki DG. 2004. Does combination antimicrobial therapy reduce mortality in Gram-negative bacteraemia? A meta-analysis. *The Lancet Infectious diseases* **4**(8): 519-527.

Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics (Oxford, England)* **30**(14): 2068-2069.

Skaar EP. 2010. The battle for iron between bacterial pathogens and their vertebrate hosts. *PLoS pathogens* **6**(8): e1000949.

Tan H, Zhang L, Zhao Q, Chen R, Liu C, Weng Y, Peng Q, Bai F, Cheng Z, Jin S et al. 2016. DeaD contributes to Pseudomonas aeruginosa virulence in a mouse acute pneumonia model. *FEMS microbiology letters* **363**(20).

Theocharidis A, van Dongen S, Enright AJ, Freeman TC. 2009. Network visualization and analysis of gene expression data using BioLayout Express(3D). *Nature protocols* **4**(10): 1535-1550.

Thompson CM, Holden TD, Rona G, Laxmanan B, Black RA, O'Keefe GE, Wurfel MM. 2014. Toll-like receptor 1 polymorphisms and associated outcomes in sepsis after traumatic injury: a candidate gene association study. *Annals of surgery* **259**(1): 179-185.

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J et al. 2013. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al]* **11**(1110): 11.10.11-11.10.33.

Veesenmeyer JL, Hauser AR, Lisboa T, Rello J. 2009. Pseudomonas aeruginosa virulence and therapy: evolving translational strategies. *Crit Care Med* **37**(5): 1777-1786.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS one* **9**(11): e112963.

Walton AH, Muenzer JT, Rasche D, Boomer JS, Sato B, Brownstein BH, Pachot A, Brooks TL, Deych E, Shannon WD et al. 2014. Reactivation of multiple viruses in patients with sepsis. *PloS one* **9**(2): e98819.

Willmann M, Bezdan D, Zapata L, Susak H, Vogel W, Schroppel K, Liese J, Weidenmaier C, Autenrieth IB, Ossowski S et al. 2015. Analysis of a long-term outbreak of XDR Pseudomonas aeruginosa: a molecular epidemiological study. *The Journal of antimicrobial chemotherapy* **70**(5): 1322-1330.

Zhang J, Xu K, Ambati B, Yu FS. 2003. Toll-like receptor 5-mediated corneal epithelial inflammatory responses to Pseudomonas aeruginosa flagellin. *Investigative ophthalmology & visual science* **44**(10): 4247-4254.