Metagenomic evidence for a polymicrobial signature of sepsis

Cedric Chih Shen Tan[1*], Mislav Acman[1], Lucy van Dorp[1&], Francois Balloux[1&]

[1] UCL Genetics Institute, University College London, Gower Street, London, WC1E 6BT, United Kingdom

* Corresponding Author

E-mail: cedriccstan@gmail.com

[&] Co-last authors.

Abstract

Background

The last two decades have seen significant progress in our understanding of the human host component of sepsis. However, detailed study of the microbial species present in sepsis, including whether sepsis infections derive from single pathogen infections or wider microbial communities, has only received limited attention.

Methods

Using various rigorous computational methods, we analysed publicly available metagenomic datasets, comparing the patterns of microbial DNA in the blood plasma of septic patients relative to that of healthy individuals. We evaluated the performance of gradient-boosted tree classifiers in determining if microbial taxonomic assignments of a blood metagenomic sample can be used to distinguish a septic from a healthy individual. Additionally, we demonstrated a novel application of SHapley Additive exPlanations (SHAP), a recently developed model interpretation approach, to computationally remove putative contaminant genera present in metagenomics data. We then characterised the microbial component of sepsis using SHAP-based model interpretations of trained classifiers and co-occurrence networks. Finally, we used blood metagenomic sequencing data from three other sepsis cohorts to validate our findings.

Results

Classifiers had high performance in delineating septic and healthy individuals when provided only with information on the abundance of microbial genera (AUROC = 0.995), including when putative contaminant genera were excluded (AUROC = 0.942). Following the removal of specific pathogens ‘confirmed’ through clinical microbiological testing, septic samples could still be readily distinguished from healthy samples. Models trained on single genera as features did not perform as well as those making

use of a full polymicrobial model. Lastly, our microbial networks showed multiple co-occurrence relationships between microbial genera identified in septic samples that are not present in healthy controls.

Conclusion

While the prevailing clinical diagnostic paradigms seek to identify single causative agents, our results provide evidence for a polymicrobial signature of sepsis. We anticipate that this polymicrobial signature can provide insights into the nature of microbial infection during sepsis and yield predictions which may guide biological insights and ultimately be leveraged as a clinically useful tool.

Introduction

Sepsis poses a significant challenge to public health and was listed as a global health priority by the World Health Organisation (WHO) in 2017. In the same year, 48.9 million cases of sepsis and 11 million deaths were recorded worldwide (1). The burden of sepsis weighs heavier on low and lower-middle income countries due to poorer healthcare infrastructures (2).

Current research efforts have predominately focused on understanding the host's response to sepsis and developing faster diagnostics, predictive tools and treatments. Indeed, all contemporary definitions of sepsis focus on the host's response and resulting systemic complications. The 1991 Sepsis-1 definition described sepsis as a systemic inflammatory response syndrome (SIRS) caused by infection, with patients being diagnosed with sepsis if they fulfil at least two SIRS criteria and have a clinically confirmed infection (3). The 2001 Sepsis-2 definition then expanded the scope of SIRS to include more symptoms (4). More recently, the 2016 Sepsis-3 definition sought to differentiate between mild and severe cases of dysregulated host responses, describing sepsis as a life-threatening organ dysfunction as a result of infection (5). Significant progress has been made in understanding how dysregulation occurs (6) and the long-term impacts of sepsis (7,8). Additionally, early-warning tools have been developed based on patient health-care records (9–11) and clinical checklists (12,13). However, the focus on the host component of sepsis may overlook the important role of microbial composition in the pathogenesis of the disease.

Due to the severity of sepsis, current practice considers identification of a single pathogen sufficient to warrant a diagnosis, without consideration of other, potentially relevant, species in the bloodstream. Upon diagnosis, infections are rapidly treated with broad spectrum antibiotics. However, blood cultures, the current recommended method of diagnosis before antimicrobial treatment (14), are known to yield false negatives due to certain microorganisms failing to grow in culture (15), particularly in samples with low

microbial loads (16). Culture-based methods, while useful in a clinical context, may therefore under-estimate the true number of causative pathogens infecting septic patients.

Sepsis is a highly heterogeneous disease which consists of both a host component and a microbial component. While the former has been widely studied, the latter appears to represent a largely untapped source of information that could further advance our understanding of sepsis. Several diseases manifest as a result of interactions in a polymicrobial community. For example, microbial interactions in lung, urinary tract and wound infections are all known to contribute to differing disease outcomes (reviewed by Tay *et al.* (17)). These findings suggest that the microbial component of sepsis may also be crucial to understanding its pathogenesis.

Current technologies to investigate the presence of polymicrobial communities have some major limitations. As noted previously, culture-based methods have a high false negative rate. Further, without knowledge of the range of microorganisms that infect blood, co-culture experiments to study microbial interactions prove difficult. For polymerase chain reaction-based technologies, the use of species-specific primers (e.g. SeptiFast (18)) necessitates *a priori* knowledge of microbial sequences endogenous to septic blood. Lastly, metagenomic sequencing is ubiquitously prone to environmental contamination. This can include DNA from viable cells introduced during sample collection, sample processing, or DNA present in laboratory reagents (19–21); the so called 'kitome'. As such, it can be difficult to determine which microorganisms are truly endogenous to the sample, and at what abundance.

In this study, we sought to expand our understanding of the full microbial component of sepsis. Multiple statistical and state-of-the-art machine learning techniques were applied to metagenomic sequencing data published by Blauwkamp *et al*. (22) (henceforth Karius study). To circumvent the problem of potential contamination in metagenomic data, we developed and applied a novel computational contamination reduction technique. We also validated our findings using a pooled dataset comprising three other

independent sepsis cohorts. Taken together, our results provide strong evidence for a polymicrobial signature of sepsis. In addition, we detail several recommendations for the use of metagenomic sequencing to investigate blood-borne infections.

Methods

Datasets

Our primary analysis involved published shotgun metagenomic sequence data from the Karius study. Patients were diagnosed with sepsis if they presented with a temperature > 38°C or < 36°C, at least one other Systemic Inflammatory Response Syndrome (SIRS) criterion, and evidence of bacteraemia. Bacteraemia was confirmed via clinical microbiological testing performed within seven days after collection of the blood samples. The list of pathogens identified by such tests (henceforth 'confirmed' pathogens) can be found in Supplementary Table 5 of the Karius study, under the 'definite' adjucation. This included tissue, fluid and blood cultures, serology and nucleic acid testing. The clinical outcome of each patient was not reported in the original study. Seven of the 117 septic patients were found to have more than one 'confirmed' pathogen identified by microbiological testing (Supplementary Table 5; Karius study). According to the Karius study, healthy individuals were "screened for common health conditions including infectious diseases through a questionnaire and standard blood donor screening assays". We believe this to be reasonable grounds for ruling out bloodstream infections in healthy patients (i.e. of non-septic origin).

To determine if our findings were applicable beyond the Karius dataset, we pooled the Karius dataset with metagenomic information from three other independent sepsis cohorts (23–25). The final pooled dataset contains sequence data from multiple sources, sepsis definitions and sequencing techniques (summarised Table 1). We will henceforth refer to individual datasets by their dataset alias as shown in Table 1.

Table 1. Summary of metagenomic datasets. Sample sizes indicated here are those after all quality control steps have been applied.

Study	Dataset alias	Sepsis definition	Sequencing technique	Sample size	
				Septic	Healthy
Grumaz *et al.* (2019)	Grumaz-19	Sepsis-2	Shotgun	50	-
Grumaz *et al.* (2016)	Grumaz-16	Sepsis-2	Shotgun	7	15
Gosiewski *et al.* (2017)	Gosiewski-17	Sepsis-1	16S (paired-end)	56	23
Blauwkamp *et al.* (2019)	Karius	Sepsis-1	Shotgun	117	170

Data pre-processing

As described in the Karius study, input circulating free DNA was sequenced using NextSeq500 (75-cycle PCR, 1 x 75 nucleotides). Raw Illumina sequencing reads were demultiplexed by bcl2fastq (v2.17.1.14; default parameters) and quality trimmed using Trimmomatic (v0.32) (26) retaining reads with a quality (Q-score) above 20. Mapping and alignment were performed using Bowtie (v2.2.4) (27). Human reads were identified by mapping to the human reference genome and removed prior to deposition in NCBI's Sequence Read Archive (PRJNA507824).

For Grumaz-16 and Grumaz-19, *BBMap* (v38.79) (28) was used to trim adapter sequences, remove reads with a Q-score below 20 and remove reads mapping to the masked hg19 reference (https://tinyurl.com/yya4xmrg). For the Gosiewski-17 dataset, we performed the same pre-processing steps as reported in the associated study (24). Briefly, primers and adapters were removed using *Cutadapt* (v1.18) (29), paired reads merged using *ea-utils* (v1.1.2.537) (30), merged reads and forward unmerged fastq files concatenated, and reads with a Q-score below 20 removed using *BBMap*.

Taxonomic classification of all shotgun sequencing data was performed using *Kraken 2* (v2.0.9-beta; default parameters) (31) with the *maxikraken2_1903_140GB* database (https://tinyurl.com/y7zfg9kr). For the Gosiewski-17 dataset, *Kraken 2* with a *Kraken 2*-built *Silva* database was used instead of conventional 16S amplicon metagenomic classification methods (32). Read assignments for all 'confirmed' bacterial pathogens using the *maxikraken2_1903_140GB* and *Kraken 2*-built *Silva* databases are shown in Fig. S1. While the relative number of reads assigned to each bacterial genus showed some inconsistencies, this hardly affected the classifier performance of septic and healthy patients (Fig. S2; training protocol described below). This suggests that our model is fairly robust to any heterogeneity which may be introduced by the metagenomic classification step. For the subsequent analyses, we present results using the taxonomic assignments based on the *Kraken 2*-built *Silva* database for the 16S Gosiewski-17 samples.

Kraken 2 taxonomic assignment provides the total number of reads assigned to each taxon, which we loosely refer to as 'abundance'. The final data matrices used in downstream analyses had samples represented in rows and taxa in columns (*i.e.* features) and excluded unclassified reads. The set of taxa used in each analysis will henceforth be referred to as the 'feature space' and the final data matrices obtained here represent the *Neat* feature spaces.

Unexpectedly, for the Karius dataset, some reads were assigned to the genus *Homo* which was possibly due to misclassification. Mapping of all reads in the Karius sequencing data found just 873 bases with 96% identity to the masked human reference. Since human reads were already removed in the bioinformatic workflow of the Karius study, we did not perform an additional human read removal step to avoid introducing biases into the data.

Model training, optimisation and evaluation

We compared the performance of gradient-boosted tree classifiers trained on different feature spaces. The task of each classifier was to predict if the taxonomic assignments of a blood metagenome belonged to a

septic patient or a healthy control. Classifiers were trained with a binary-logistic loss function and implemented using *XGBoost* API (v0.90) (33). Model optimisation was performed using a randomised hyperparameter optimisation protocol (34) (1000 samples) implemented using *RandomizedSearchCV* in the *Scikit-learn* API (v0.23.1) (35). The test error of each model was evaluated using a nested, stratified, 10 x 10-fold cross-validation procedure. The best performing sets of hyperparameters that maximise the receiver operating characteristic curve (AUROC) of each model were determined and used for downstream analyses.

Model interpretation

To interpret classifiers, we used SHapley Additive exPlanations (SHAP) – a state-of-the-art machine learning technique for interpreting 'black-box' classifiers (36). Briefly, each feature in a single sample is assigned a SHAP value, which corresponds to the change in a sample's predicted probability score (*i.e.* probability of sepsis) when the feature is either present or absent. Using SHAP values therefore allows the decomposition of predicted probability scores for each sample into the sum of contributions from individual genera. SHAP values were computed using *TreeExplainer*, part of the *shap* library (v0.34.0) (36). The relative importance of each feature (*i.e.* the read count assigned to a genus) was inferred via the mean absolute SHAP value across all samples for a particular feature. A higher mean absolute SHAP value implies that the feature has a larger impact on the model predictions. For every model, SHAP values were computed for the whole dataset by setting the *feature_pertubation* parameter to 'interventional'.

SHAP decontamination

Accurate characterisation of the microbial component of sepsis requires discrimination between a true biological signal and that arising from putative environmental contamination in metagenomes. We developed a procedure to remove biologically irrelevant genera from the feature space, which we will refer to as SHAP decontamination. The main premise behind SHAP decontamination is that pathogens

should occur at higher abundance in septic patients relative to healthy controls. This is because we expect most infections to be characterised by the proliferation of microorganisms (37,38) and, as such, true pathogenic genera should contribute to a higher predicted probability of sepsis at higher abundances. Consequently, the abundance of contaminant taxa would demonstrate a negative Spearman's correlation with their corresponding SHAP values. This allows putative contaminant genera to be computationally detected and removed. SHAP decontamination was performed in two main steps. Firstly, genera that are not currently identified as known human pathogens were first removed. This selection was based on a study by Shaw *et al*. (39), who considered as a 'human pathogen' any microbial species for which there is evidence in the literature that it can cause infection in humans, sometimes in a single patient. Secondly, a classifier was optimised and trained on genera abundance (*Neat* feature spaces). SHAP values for model predictions on the dataset were then calculated. Genera with a negative Spearman's correlation between their corresponding SHAP values and abundances were removed. Spearman's correlations were calculated using *spearmanr* as part of the *SciPy* library (v1.4.1) (40). A new classifier was then retrained using the previously optimised set of parameters but with this new reduced feature space. This process was repeated iteratively until the number of genera retained remained constant. The resultant feature space is denoted by *CR*.

Simple decontamination

We also employed a more direct, model-free contaminant removal technique (Simple decontamination) that follows the same underlying premise of SHAP decontamination. In this procedure, genera in the *Neat* feature space that were significantly ($p < 0.05$) more abundant in healthy controls than septic samples were considered contaminants and removed. The resultant feature space is denoted by *SD*.

Microbial networks

Microbial co-occurrence networks were constructed using the *SparCC* algorithm (41), implemented in the *SpiecEasi* package (v1.1.0) (42) and visualised using *Igraph* (v1.2.5) (43). *SparCC* was used so as to account for compositionality that could lead to spurious correlations. Separate networks were constructed for the genera assignments of septic and healthy metagenomes. To determine the microbial associations present exclusive to septic samples, a corrected sepsis network was produced. This network was constructed by subtracting all edges of the healthy network from the sepsis network. Only co-occurrence relationships where the *SparCC* correlations exceed 0.2 were retained. The *Karius-SD* feature space was used as input.

Data and code availability

Data parsing, model training, optimisation, and evaluation were performed with *Python* 3.7.3. Microbial network analysis was performed with *R* v3.6.0. All relevant source code and parsed datasets can be found on GitHub (https://github.com/cednotsed/Polymicrobial-Signature-of-Sepsis).

Results

Metagenomic sequencing can be used to discriminate septic from healthy samples

The performance of all classifiers are summarised in Table 2. Using the raw feature space parsed from the *Kraken 2* taxonomic assignments, classifiers had a very high classification performance (*Karius-Neat* model; AUROC = 0.995) in predicting if a patient suffered from sepsis. This was similarly observed when using the pooled dataset (*Pooled-Neat* model; AUROC = 0.982).

Table 2. Summary of models trained. The prefix and suffix of each model name corresponds to the dataset and contamination reduction technique applied, respectively. *Neat*, *SD*, and *CR* refer to the feature spaces with no, Simple decontamination, and SHAP decontamination applied, respectively. *Karius-Without* corresponds to the SHAP decontaminated feature space after claimed 'confirmed' pathogens are excluded. *Karius-Only* refers to the feature space containing only genera with 'confirmed' pathogens as features.

No. of Features	Feature Space	Model Performance		
		Precision	**Recall**	**AUROC**
1564	*Karius-Neat*	0.976	0.983	0.995
111	*Karius-SD*	0.896	0.787	0.942
25	*Karius-CR*	0.883	0.810	0.942
22	*Karius-Without*	0.803	0.727	0.915
22	*Karius-Only*	0.929	0.862	0.950
685	*Pooled-Neat*	0.950	0.939	0.982
21	*Pooled-CR*	0.870	0.796	0.904

Removing putative contaminants

Visualisation of SHAP values allows us to determine how a feature influences model predictions. We inspected the SHAP values and read counts assigned to the genera corresponding to cases of each type of 'confirmed' infection (*e.g.* SHAP value/read count assigned to *Escherichia* for only *Escherichia*-positive samples) using the *Karius-Neat* feature space. The SHAP values were all at greater or equal to zero with the exception of a single sample which had a negative SHAP value for *Mycobacterium* (Fig. S3). The assigned read counts were non-zero except for one sample with a 'confirmed' fungal *Candida glabrata* infection reported (SRR8288759). This reiterates previous suggestions that metagenomic sequence data of blood samples can be used to identify pathogens involved in bacteraemia (22,23). Additionally, these findings suggest that SHAP values indeed reflect the prior that genera containing true pathogens should be positively associated with sepsis.

For the Karius dataset, application of SHAP decontamination resulted in 25 genera after three iterations while Simple decontamination resulted in 111 genera. The *Karius-CR* and *Karius-SD* feature spaces shared 21 genera. Classifiers trained on either the *Karius-CR* or *Karius-SD* feature space had similarly high performance (*Karius-CR/SD*; AUROC = 0.942), despite the large reduction in the number of features, suggesting that computational decontamination efficiently removes redundancy in the metagenomic feature space. Since the application of both decontamination techniques offer equivalent classification performance, these findings also suggest that SHAP decontamination is more stringent in removing putative contaminants than Simple decontamination with minimal detriment to classification performance.

Separately, we observed that the *Karius-CR* model comprised almost all genera associated to sepsis at higher abundance. Additionally, genera such as *Sphingobium*, *Mesorhizobium* and *Ralstonia*, were highly important features in the *Karius-Neat* feature space (Fig. 1a), though not present in either the *Karius-CR* or *Karius-SD* feature space (Fig. 1b and c). These genera are likely to be contaminants since they

contribute negatively to the predicted probability of sepsis at high abundance, and have been previously ascribed as common sequencing contaminants (19). Of the 25 genera in the *Karius-CR* feature space, eight corresponded to genera containing 'confirmed' pathogens (see Methods). Notably, *Escherichia* and *Enterobacter*, which are both 'confirmed' pathogens but also common contaminants (19), were retained in both decontaminated feature spaces. These findings suggest that computational decontamination procedures were removing putative contaminants while selectively retaining biologically important genera.

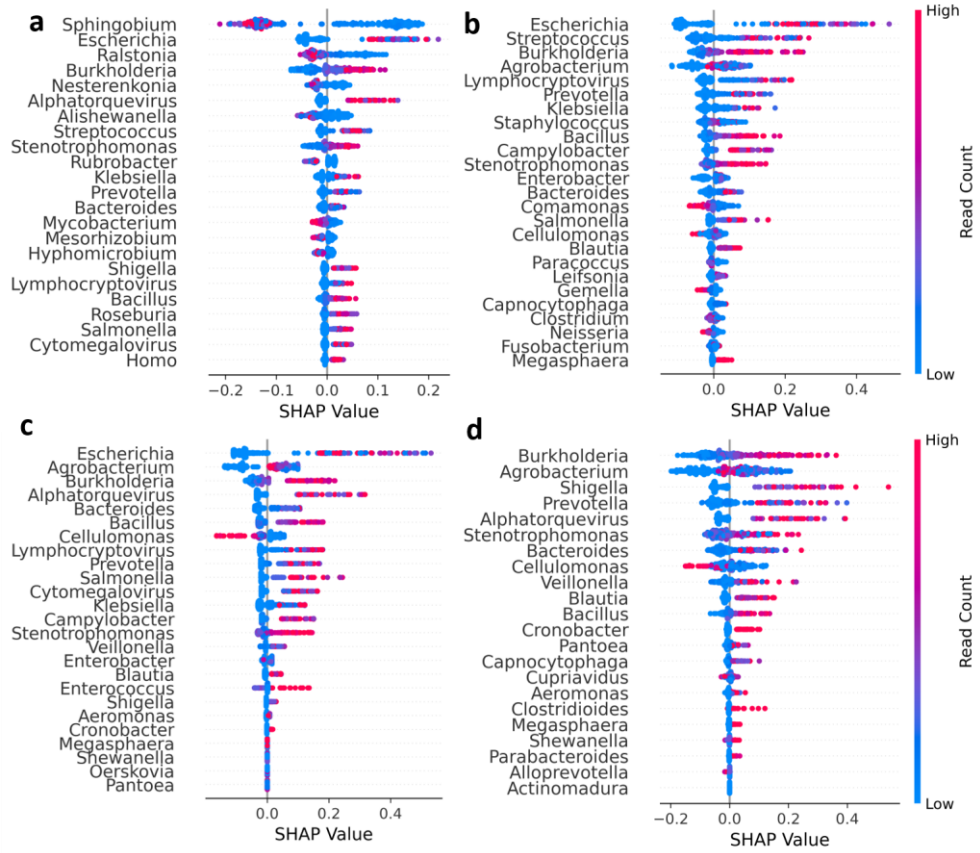

Figure 1. Model interpretation and performance. (a) Plot summarising the SHAP values across all samples for the most important features ranked by the mean absolute SHAP value (highest at the top) for *Karius-Neat*, (b) *Karius*-SD, (c) *Karius-CR* and (d) *Karius-Without* models. Each point represents a single sample. Points with similar SHAP values were stacked vertically for visualisation of point density and were coloured according to the magnitude of the feature values (*i.e.* read counts).

Evidence for a polymicrobial community

We provide several pieces of evidence supporting the polymicrobial nature of sepsis. Most notably, a classifier trained on the Karius dataset using the SHAP decontaminated feature space but with all genera containing 'confirmed' pathogens removed, performed well. This suggests that the abundance of microbial genera that were not amongst the 'confirmed' pathogens also provides a good predictor of sepsis (*Karius-Without* model; AUROC = 0.915). Visualisation of the SHAP values for this model (Fig. 1d) confirmed that most genera had positive associations with sepsis at higher abundances, and are therefore considered biologically relevant. We then trained and optimised multiple single-feature classifiers with each genus in the *Karius-Without* feature space. Fig. 2 shows the performance of the multi-feature *Karius-Neat*, *Karius-Without* and *Karius-Only* models compared to single-feature models. The classifier trained on genera containing 'confirmed' pathogens as features only (*Karius-Only*) performed well. However, the classifier trained on the SHAP decontaminated feature space with said genera removed (*Karius-Without*) also performed well, suggesting that genera that do not contain 'confirmed' pathogens can also be used to delineate septic and healthy samples. Furthermore, the single-feature models performed relatively poorly, highlighting that no genus is solely responsible for the high classification performance of the *Karius-Without* model, further supporting the polymicrobial nature of sepsis infections.

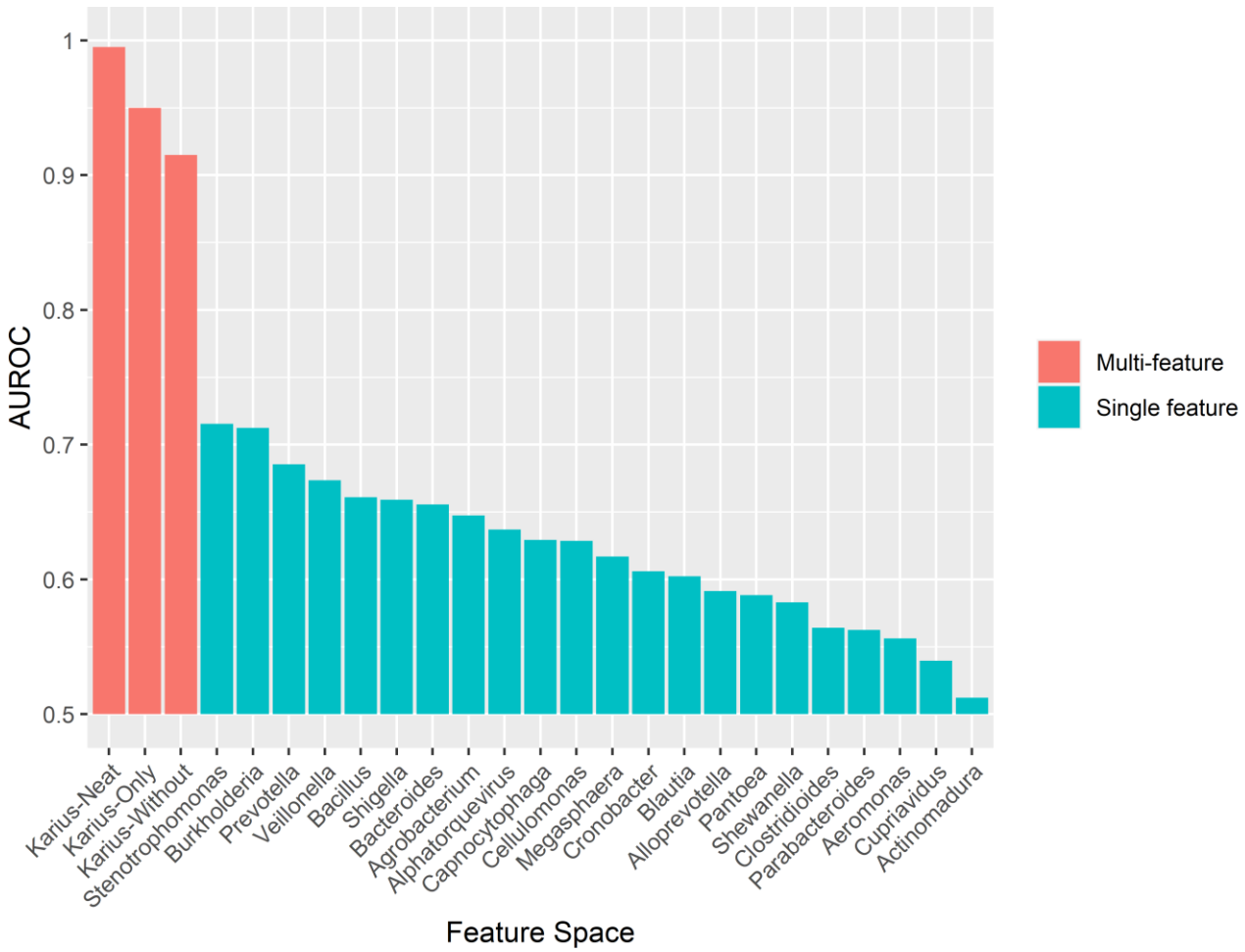

Figure 2. Comparison of performance (AUROC) for the multi-feature models (*Karius-Neat, Karius-Only, Karius-Without* feature space) and single-feature models (x-axis).

Further, we trained classifiers on the pooled dataset to determine if our results were portable to other sepsis cohorts beyond the Karius dataset. Despite the pooled dataset comprising multiple data sources from different studies, the classifiers were still able to determine a generalisable signature across the datasets (*Pooled-Neat* model, AUROC = 0.982; *Pooled-CR* model, AUROC = 0.904). This suggests that there is a generalisable signature which can be leveraged across metagenomic datasets. If we accept that SHAP decontamination can remove most contaminant genera, these results suggest that the generalisable signature captures, to some extent, the underlying microbial component of sepsis. 21 genera were retained after SHAP decontamination (*Pooled-CR* feature space) of which 12 were in common with that of the *Karius-CR* feature space.

Lastly, microbial co-occurrence networks were used to identify relationships between genera that were exclusive to septic samples. Two genera are said to co-occur if an increase in the abundance of one is associated with an increase in the abundance of the other. The presence of such relationships would lend weight to the polymicrobial nature of sepsis infections. The *Karius-SD* feature space was used in this analysis to corroborate previous analyses using the *Karius-CR* feature space. Multiple co-occurrence relationships between genera were present in the corrected network including those containing 10 of the 22 'confirmed' pathogens and 14 of the 25 genera in the *Karius-CR* feature space (Fig. 3). Interestingly, a co-occurrence cluster of genera associated to the oral cavity was detected (Fig. 3), as suggested by the Human Oral Microbiome Database (44) (accessed 15th July 2020) and the current literature (45–48). This was also present in the corrected network when the *Pooled-SD* feature space was used as input (Fig. S4). The multiple co-occurrence relationships between genera detected suggest that there may be a set of pathogens that tend to be present during infection.

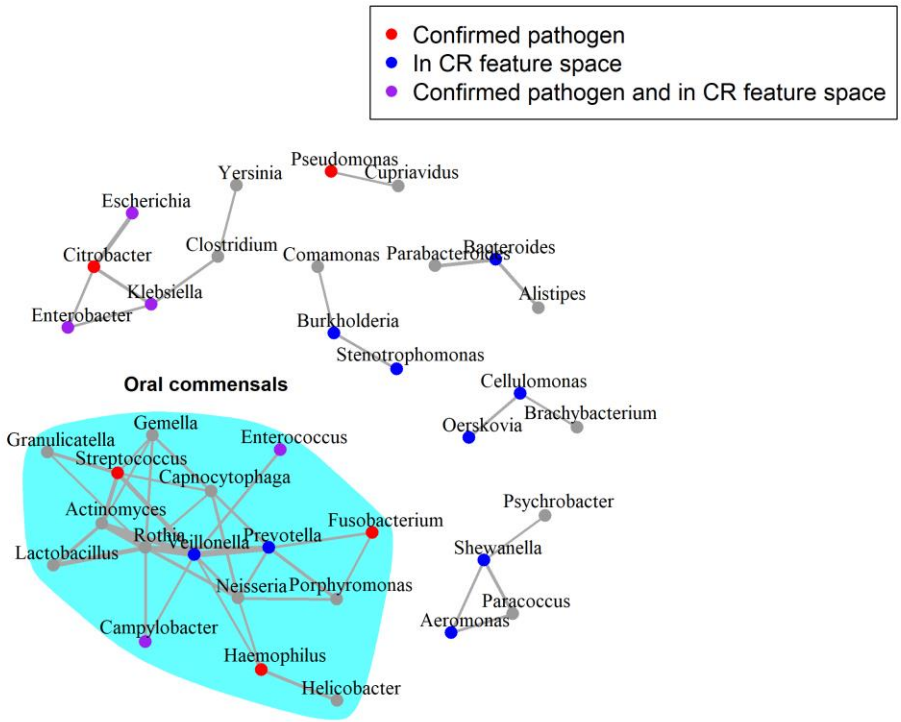

Figure 3. Corrected microbial co-occurrence network for genera assigned in sepsis metagenomes. Input data corresponds to the *Karius-SD* feature space. The edges in this network represent those in the septic network that were not present in the healthy network. The width of edges are weighted by the strength of the *SparCC* correlations. The layout of the graph was generated using the Fruchterman-Reingold algorithm.

Discussion

SHAP decontamination

Contamination from environmental sources poses one of the greatest challenges for metagenomic investigations of microbial communities, particularly in low biomass and clinical samples (20). It is therefore crucial to discriminate between contaminants and biologically relevant taxa. The removal of putative contaminant genera is necessary to protect against spurious polymicrobial signals. The underlying principle behind our proposed SHAP-based removal of contaminants is that clinically relevant microorganisms in blood should be associated with septic samples at higher microbial abundance. Our results demonstrate the efficacy of our post-hoc contamination reduction technique called SHAP decontamination in removing redundancy in the feature space while selectively retaining taxa involved in sepsis. It is likely that the taxa removed in this procedure would in principle include commensals and environmental contaminants introduced during sample collection or preparation; the “kitome”. As such, application of this technique provided greater confidence that the polymicrobial signals we observed were not driven by contaminant genera.

We appreciate that a more rigorous evaluation of this technique, particularly with mock communities, will be required. As an alternative to our contamination reduction technique, statistical decontamination techniques identifying inverse relationships between the assigned abundance of taxa and sample DNA concentration (49,50) could be used. However, this method was not applicable for our study since sample DNA concentrations in all the datasets used were not reported.

Potential for metagenomics-based diagnostics

Although we do not claim to have developed a model sufficiently robust for diagnostics, our results highlight the clear promise of metagenomics-based diagnostic models, which have also been suggested by previous studies (22,51). To put the performance of our models in context, Mao et al. (9) reported that InSight, a model trained on vital signs of patients, had a diagnostic AUROC of 0.92 using Sepsis-2 as the ground truth. They also reported that MEWs, SOFA and SIRS had an AUROC of 0.76, 0.63 and 0.75 respectively. Additionally, a classifier trained on nasal metagenomes of septic and healthy samples had an AUROC of 0.89 with Sepsis-3 as the ground truth (52). Notably, it is difficult to compare the performance of models trained with labels generated by different definitions of sepsis, which is also inherently a highly heterogeneous disease. Further, the discrepancies in model performance could be due to differences in the size of training and testing datasets. However, at the very least, our results suggest that the microbial component of sepsis alone contains sufficient information for the diagnosis of sepsis. The crucial next step would be to generate larger datasets, from more diverse sources, to allow the training of more robust and generalisable diagnostic or prognostic models.

The polymicrobial signature of sepsis

Our work demonstrates a strong polymicrobial signal in sepsis, with multiple, co-occuring, microorganisms contributing to prediction of disease. Although our networks were inferred computationally, there is evidence in the literature supporting possible synergies between some of the co-occuring genera. For example, using fluorescence in-situ hybridisation, interspecies spatial associations were found between *Prevotella*, *Veillonella*, *Streptococcus*, *Gemella*, *Rothia* and *Actinomyces* (53), which were also the genera with the strongest correlations in the corrected sepsis network (Fig. 3). Separately, *Stenotrophomonas* and *Burkholderia* are known to play a collective role in the pathogenesis of cystic fibrosis (54). Lastly, *Klebsiella pneumoniae* was found to be able to transmit extended spectrum beta-

lactamase genes to *Citrobacter freundii* and *E. coli* (55), potentiating synergism during polymicrobial infections. These examples suggest the co-occurrence relationships we detected are genuine signals. Further investigation of the interactions between different clusters of genera in the corrected sepsis network, together with expanding to future datasets, may yield valuable insights into the underlying biology of sepsis infections and ultimately inform treatment.

The presence of a densely-connected cluster of oral colonisers may point to a potential reservoir of pathogens in sepsis. This also suggests the possibility of opportunistic infections from the human microbiota and dysbioses that could affect disease severity. This hypothesis is in line with the reported changes in nasal microbiomes in septic individuals (52) and the associations of intestinal dysbiosis with increased susceptibility to sepsis (56). If these hypotheses were true, microbiome profiles of patients may yet offer opportunities to assess a patient's risk of developing sepsis prior to onset.

Limitations

We identified several limitations in our study. Firstly, metagenomic sequencing involves measurements of circulating free DNA and not of viable microorganisms in blood. As such, the detection of DNA from multiple taxa does not necessarily represent the true number or abundance of active taxa present. However, multiple studies have demonstrated high concordance of targeted (57) or shotgun metagenomic sequencing with culture (22,51,58). This suggests some level of agreement between the presence of microbial cells and their DNA in blood samples. Additionally, given its higher sensitivity and throughput, metagenomic sequencing appears to be the best tool currently available for gaining insights into polymicrobial infections.

Though our results suggest the importance of multiple genera in delineating metagenomes of septic patients from that of healthy controls, the etiological contributions of these genera and their ecological relationships cannot be inferred. Such hypotheses must be confirmed experimentally. It is also important

to keep in mind that the models presented in this study are not prognostic in nature, in that they cannot predict the onset or progression of sepsis. However, furthering our understanding of the microbial component of sepsis may prove useful in the development of better prognostic tools.

Some genera such as *Escherichia* and *Enterobacter* contain both biologically-relevant genera and common sequencing contaminants. As such it is expected that a proportion of DNA molecules, and hence sequencing reads, may have come from contamination rather than microorganisms endogenous to blood. The abundance of these microorganisms, as detected by metagenomic approaches, may differ from the true abundance.

Additionally, *k*-mer based approaches may be less accurate for taxonomic classification compared to, for example, Bayesian sequence read-assignment methods (59). As such, we used taxonomic assignments at the genus level which were shown to be, in general, more reliable than that at the species level (60). We also appreciate that *k*-mer based classification approaches are significantly faster (61), which may provide clinically relevant turnaround times that are important in sepsis diagnostics.

Finally, we acknowledge the relatively small size of the datasets used in our analyses. As a result, the models presented in this study are not yet robust enough to be used in a clinical context. A larger and more diverse dataset is required to develop such models. This is to ensure that models can learn a more generalisable decision boundary for accurate sepsis diagnosis.

Irrespective of these limitations, our results nonetheless demonstrate the importance of considering the full polymicrobial component of sepsis and suggest that taking a metagenomics-based approach may provide biological and clinical insights supporting the future development of rapid diagnostic tools.

Future directions

A major next step forward would be to elucidate the role of polymicrobial communities identified in sepsis. One key hypothesis is whether there are different clusters of microbial communities in different sepsis etiologies. Finding evidence for this potentially allows us to formally redefine sepsis to include a microbial component. However, to do so, a much larger sepsis cohort must be recruited to provide adequate statistical power, particularly considering the true number of sepsis subtypes is unknown. The associations between polymicrobial communities and host response severity, from mild to lethal, could also be investigated. To do so, anonymised health-care records with detailed curation of the clinical outcomes and treatment history of each patient would be required. In addition, pre-infection data from animal models holds promise to identify taxa important in detecting early onset through to outcomes of disease.

It would also be valuable to investigate how polymicrobial communities change along the course of infection. This can be done via analysis of microbial community dynamics (32) using longitudinal metagenomic data. By monitoring the change in microbial composition along the course of infection, ecological relationships between pathogens can be inferred. Additionally, this would allow for a better identification of key taxa involved in sepsis at the level of the microbial species. Lastly, co-culture experiments (39) could be performed to elucidate interactions between pathogens. These could also be paired with metabolomic approaches which may be useful in identifying possible synergies or antagonisms between microbial species (40).

Conclusions

The advent of large-scale metagenomic sequencing of clinical samples offers new opportunities to better characterise the pathogens contributing to systemic infections, and unlike culture-based methods are not limited to organisms that are fast-growing or culturable. In this study, we demonstrate the promise of a metagenomics-based approach to sepsis. Our results provide evidence that septic infections should be

considered as polymicrobial in nature, comprising multiple co-occuring pathogens indicative of disease. Our findings thus pave the way for more microbial-focused models of sepsis, with long run potential to inform early detection, clinical interventions and improve patient outcomes.

References

1. Rudd KE, Johnson SC, Agesa KM, Shackelford KA, Tsoi D, Kievlan DR, et al. Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the Global Burden of Disease Study. Lancet. 2020;395(10219):200–11.
2. Kwizera A, Baelani I, Mer M, Kissoon N, Schultz MJ, Patterson AJ, et al. The long sepsis journey in low- and middle-income countries begins with a first step... but on which road? Springer; 2018.
3. Bone RC, Balk RA, Cerra FB, Dellinger RP, Fein AM, Knaus WA, et al. Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. Chest. 1992;101(6):1644–55.
4. Levy MM, Fink MP, Marshall JC, Abraham E, Angus D, Cook D, et al. 2001 sccm/esicm/accp/ats/sis international sepsis definitions conference. Intensive Care Med. 2003;29(4):530–8.
5. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The third international consensus definitions for sepsis and septic shock (Sepsis-3). Jama. 2016;315(8):801–10.
6. van der Poll T, van de Veerdonk FL, Scicluna BP, Netea MG. The immunopathology of sepsis and potential therapeutic targets. Nat Rev Immunol. 2017;17(7):407.
7. Venet F, Monneret G. Advances in the understanding and treatment of sepsis-induced immunosuppression. Nat Rev Nephrol. 2018;14(2):121.
8. Ammer-Herrmenau C, Kulkarni U, Andreas N, Ungelenk M, Ravens S, Huebner C, et al. Sepsis induces long-lasting impairments in CD4+ T-cell responses despite rapid numerical recovery of T-lymphocyte populations. PLoS One. 2019;14(2).
9. Mao Q, Jay M, Hoffman JL, Calvert J, Barton C, Shimabukuro D, et al. Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU. BMJ Open. 2018;8(1):e017833.
10. Taylor RA, Pare JR, Venkatesh AK, Mowafi H, Melnick ER, Fleischman W, et al. Prediction of in-hospital mortality in emergency department patients with sepsis: a local big data–driven, machine learning approach. Acad Emerg Med. 2016;23(3):269–78.

11. Nemati S, Holder A, Razmi F, Stanley MD, Clifford GD, Buchman TG. An interpretable machine learning model for accurate prediction of sepsis in the ICU. Crit Care Med. 2018;46(4):547–53.

12. Henry KE, Hager DN, Pronovost PJ, Saria S. A targeted real-time early warning score (TREWScore) for septic shock. Sci Transl Med. 2015;7(299):299ra122-299ra122.

13. Smith GB, Prytherch DR, Meredith P, Schmidt PE, Featherstone PI. The ability of the National Early Warning Score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death. Resuscitation. 2013;84(4):465–70.

14. Rhodes A, Evans LE, Alhazzani W, Levy MM, Antonelli M, Ferrer R, et al. Surviving sepsis campaign: international guidelines for management of sepsis and septic shock: 2016. Intensive Care Med. 2017;43(3):304–77.

15. Klaerner H-G, Eschenbach U, Kamereck K, Lehn N, Wagner H, Miethke T. Failure of an automated blood culture system to detect nonfermentative gram-negative bacteria. J Clin Microbiol. 2000;38(3):1036–41.

16. Benjamin RJ, Wagner SJ. The residual risk of sepsis: modeling the effect of concentration on bacterial detection in two-bottle culture systems and an estimation of false-negative culture rates. Transfusion. 2007;47(8):1381–9.

17. Tay WH, Chong KKL, Kline KA. Polymicrobial–host interactions during infection. J Mol Biol. 2016;428(17):3355–71.

18. Westh H, Lisby G, Breysse F, Böddinghaus B, Chomarat M, Gant V, et al. Multiplex real-time PCR and blood culture for identification of bloodstream pathogens in patients with suspected sepsis. Clin Microbiol Infect. 2009;15(6):544–51.

19. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol. 2014;12(1):87.

20. Glassing A, Dowd SE, Galandiuk S, Davis B, Chiodini RJ. Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. Gut Pathog. 2016;8(1):24.

21. Weiss S, Amir A, Hyde ER, Metcalf JL, Song SJ, Knight R. Tracking down the sources of experimental

contamination in microbiome studies. Genome Biol. 2014;15(12):564.

22. Blauwkamp TA, Thair S, Rosen MJ, Blair L, Lindner MS, Vilfan ID, et al. Analytical and clinical validation of a microbial cell-free DNA sequencing test for infectious disease. Nat Microbiol [Internet]. 2019;4(4):663–74. Available from: https://doi.org/10.1038/s41564-018-0349-6

23. Grumaz S, Stevens P, Grumaz C, Decker SO, Weigand MA, Hofer S, et al. Next-generation sequencing diagnostics of bacteremia in septic patients. Genome Med. 2016;8(1):73.

24. Gosiewski T, Ludwig-Galezowska AH, Huminska K, Sroka-Oleksiak A, Radkowski P, Salamon D, et al. Comprehensive detection and identification of bacterial DNA in the blood of patients with sepsis and healthy volunteers using next-generation sequencing method-the observation of DNAemia. Eur J Clin Microbiol Infect Dis. 2017;36(2):329–36.

25. Grumaz S, Grumaz C, Vainshtein Y, Stevens P, Glanz K, Decker SO, et al. Enhanced performance of next-generation sequencing diagnostics compared with standard of care microbiological diagnostics in patients suffering from septic shock. Crit Care Med. 2019;47(5):e394.

26. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20.

27. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10(3):R25.

28. Bushnell B. BBMap: a fast, accurate, splice-aware aligner. Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States); 2014.

29. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 2011;17(1):10–2.

30. Aronesty E. Comparison of sequencing utility programs. Open Bioinforma J. 2013;7(1).

31. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol. 2019;20(1):257.

32. Lu J, Salzberg S. Ultrafast and accurate 16S microbial community analysis using Kraken 2. bioRxiv. 2020;

33. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd

international conference on knowledge discovery and data mining. 2016. p. 785–94.

34. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. J Mach Learn Res. 2012;13(1):281–305.

35. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011;12:2825–30.

36. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell [Internet]. 2020;2(1):56–67. Available from: https://doi.org/10.1038/s42256-019-0138-9

37. Casadevall A, Pirofski L. Host-pathogen interactions: basic concepts of microbial commensalism, colonization, infection, and disease. Infect Immun. 2000;68(12):6511–8.

38. Balloux F, van Dorp L. Q&A: What are pathogens, and what have they done to and for us? BMC Biol. 2017;15(1):1–6.

39. Shaw LP, Wang AD, Dylus D, Meier M, Pogacnik G, Dessimoz C, et al. The phylogenetic range of bacterial and viral pathogens of vertebrates. Mol Ecol [Internet]. 2020 May 10;n/a(n/a). Available from: https://doi.org/10.1111/mec.15463

40. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17(3):261–72.

41. Friedman J, Alm EJ. Inferring correlation networks from genomic survey data. PLoS Comput Biol. 2012;8(9):e1002687.

42. Kurtz ZD, Müller CL, Miraldi ER, Littman DR, Blaser MJ, Bonneau RA. Sparse and compositionally robust inference of microbial ecological networks. PLoS Comput Biol. 2015;11(5):e1004226.

43. Csardi G, Nepusz T. The igraph software package for complex network research. InterJournal, complex Syst. 2006;1695(5):1–9.

44. Chen T, Yu W-H, Izard J, Baranova O V, Lakshmanan A, Dewhirst FE. The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information. Database. 2010;2010.

45. Pytko-Polonczyk J, Konturek SJ, Karczewska E, Bielański W, Kaczmarczyk-Stachowska A. Oral cavity as permanent reservoir of Helicobacter pylori and potential source of reinfection. J Physiol Pharmacol an Off J Polish Physiol Soc. 1996;47(1):121–9.

46. Periasamy S, Kolenbrander PE. Mutualistic biofilm communities develop with Porphyromonas gingivalis and initial, early, and late colonizers of enamel. J Bacteriol. 2009;191(22):6804–11.

47. Cephas KD, Kim J, Mathai RA, Barry KA, Dowd SE, Meline BS, et al. Comparative analysis of salivary bacterial microbiome diversity in edentulous infants and their mothers or primary care givers using pyrosequencing. PLoS One. 2011;6(8):e23503.

48. Chen H, Jiang W. Application of high-throughput sequencing in understanding human oral microbiome related with health and disease. Front Microbiol. 2014;5:508.

49. Jervis-Bardy J, Leong LEX, Marri S, Smith RJ, Choo JM, Smith-Vaughan HC, et al. Deriving accurate microbiota profiles from human samples with low bacterial content through post-sequencing processing of Illumina MiSeq data. Microbiome. 2015;3(1):19.

50. Davis NM, Proctor DM, Holmes SP, Relman DA, Callahan BJ. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. Microbiome. 2018;6(1):226.

51. Grumaz C, Hoffmann A, Vainshtein Y, Kopp M, Grumaz S, Stevens P, et al. Rapid Next-Generation Sequencing–Based Diagnostics of Bacteremia in Septic Patients. J Mol Diagnostics. 2020;22(3):405–18.

52. Tan X, Liu H, Long J, Jiang Z, Luo Y, Zhao X, et al. Septic patients in the intensive care unit present different nasal microbiotas. Future Microbiol. 2019;14(5):383–95.

53. Valm AM, Welch JLM, Rieken CW, Hasegawa Y, Sogin ML, Oldenbourg R, et al. Systems-level analysis of microbial community organization through combinatorial labeling and spectral imaging. Proc Natl Acad Sci. 2011;108(10):4152–7.

54. Saiman L, Chen Y, San Gabriel P, Knirsch C. Synergistic activities of macrolide antibiotics against Pseudomonas aeruginosa, Burkholderia cepacia, Stenotrophomonas maltophilia, and Alcaligenes xylosoxidans isolated from patients with cystic fibrosis. Antimicrob Agents Chemother. 2002;46(4):1105–7.

55. Sánchez MU, Bello HT, Domínguez MY, Mella SM, Zemelman RZ, González GR. Transference of extended-spectrum beta-lactamases from nosocomial strains of Klebsiella pneumoniae to other species of Enterobacteriaceae. Rev Med Chil. 2006;134(4):415–20.

56. Haak BW, Prescott HC, Wiersinga WJ. Therapeutic potential of the gut microbiota in the prevention and treatment of sepsis. Front Immunol. 2018;9:2042.

57. Salipante SJ, Sengupta DJ, Rosenthal C, Costa G, Spangler J, Sims EH, et al. Rapid 16S rRNA next-generation sequencing of polymicrobial clinical samples for diagnosis of complex bacterial infections. PLoS One [Internet]. 2013 May 29;8(5):e65226–e65226. Available from: https://pubmed.ncbi.nlm.nih.gov/23734239

58. Brenner T, Decker SO, Grumaz S, Stevens P, Bruckner T, Schmoch T, et al. Next-generation sequencing diagnostics of bacteremia in sepsis (Next GeneSiS-Trial): study protocol of a prospective, observational, noninterventional, multicenter, clinical trial. Medicine (Baltimore). 2018;97(6).

59. Morfopoulou S, Plagnol V. Bayesian mixture analysis for metagenomic community profiling. Bioinformatics. 2015;31(18):2930–8.

60. McIntyre ABR, Ounit R, Afshinnekoo E, Prill RJ, Hénaff E, Alexander N, et al. Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. Genome Biol. 2017;18(1):182.

61. Simon HY, Siddle KJ, Park DJ, Sabeti PC. Benchmarking metagenomics tools for taxonomic classification. Cell. 2019;178(4):779–94.