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

[¶] These authors contributed equally to this work.

[&] Co-last authors.

Abstract

Sepsis is defined as a life-threatening organ dysfunction arising from a dysregulated host response to infection. The last two decades have seen significant progress in our understanding of the host component of sepsis. However, detailed study of the composition of the microbial community present in sepsis, or indeed, the finer characterisation of this community as a predictive tool to identify sepsis cases, has only received limited attention. The microbial component of sepsis has been attributed to a heterogenous array of pathogens, usually identified through targeted culture. Metagenomic sequencing of microbial cell-free DNA (cfDNA) offers opportunities to instead classify the full spectrum of microorganisms in a clinical sample. In this study, we statistically characterise the microbial component of sepsis using microbial taxonomic assignments from metagenomes generated from blood samples collected in septic and healthy patients. Using gradient-boosted tree classifiers, we demonstrate the remarkable performance of microbial abundances alone to identify patients with sepsis (AUROC = 0.999; 95% CI 0.997-1.000). Additionally, we demonstrate the promising application of Shapley Additive exPlanations (SHAP), a recently developed model interpretation approach, to reduce sepsis into a set of 23 genera sufficient for diagnosis (AUROC = 0.959; 95% CI 0.903-0.991) and to aid in determining which pathogens contribute most to each diagnosis. While the prevailing clinical diagnostic paradigms seek to identify single causative agents, we instead find evidence for a polymicrobial signature of sepsis using unsupervised clustering, gradient-boosted tree classifiers and microbial network analysis. We anticipate that this polymicrobial signature can provide insights into the nature of microbial infection during sepsis and yield predictions which can ultimately be leveraged as a clinically useful tool. In light of these results, we provide some recommendations for future work involving metagenomic sequencing of blood samples.

Author summary

Sepsis is a leading cause of death globally, responsible for an estimated six million deaths every year. Current definitions and research efforts have predominately focused on understanding the host's response to sepsis and developing faster diagnostics, predictive tools and treatments. The role of the resident and pathogenic microbial community in contributing to disease status is more poorly characterised. The advent of large scale metagenomic sequencing of clinical samples offers new opportunities to characterise the species contributing to systemic infections, and unlike culture-based methods is not limited to organisms that are fast-growing or culturable. We analysed a publicly available metagenomic dataset, comparing the patterns of microbial DNA in the blood plasma of septic patients relative to that of healthy individuals. Our results provide evidence that septic infections tend to be polymicrobial in nature and that this polymicrobial signature can be leveraged for diagnosis. Our results open paths to more microbial focused models of sepsis, with long run potential to inform early detection, clinical interventions and improve patient outcomes.

Introduction

Sepsis poses a significant challenge to public health and was listed as a global health priority by the World Health Assembly and 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].

All contemporary definitions of sepsis focus on the host response to disease 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 pathogenesis of the disease.

Due to the severity of sepsis, current practice considers identification of a single pathogen sufficient to warrant a diagnosis, without considering the co-occurrence and interactions of other species. 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 the failure of culture-based methods to cultivate certain types of microorganisms [15], particularly in samples with low microbial loads [16]. Culture-based methods, while useful in a clinical

context, may therefore under-represent the true number of causative pathogens infecting septic patients. As such, it is possible that most diagnoses of 'unimicrobial' sepsis may in fact be polymicrobial in nature.

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 polymicrobial infections. For example, microbial interactions in lung, urinary tract and wound infections are all known to impact disease outcome (reviewed by Tay *et al.* [17]). These findings suggest that the microbial component of sepsis may also be crucial in 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]) necessitate *a priori* knowledge of microbial sequences endogenous to septic blood. Lastly, metagenomic sequencing is ubiquitously prone to environmental contamination. This could 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 is 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 microbiological component of sepsis and consequently to improve on current prognostic and diagnostic markers. Multiple rigorous statistical and state-of-the-art machine learning techniques were applied to metagenomic sequencing data obtained from the published study by Blauwkamp et al. [22] (henceforth Karius study). Noting the problem of contamination, we developed and successfully applied a novel computational contamination reduction technique. Additionally, several levels of statistical stringency were applied to reduce the likelihood of

detecting spurious signals. Our results provide evidence for a polymicrobial signature of sepsis and highlight the promising diagnostic potential of a metagenomic approach. In addition, we detail several recommendations for the use of metagenomic sequencing to investigate blood-borne infections.

Results

Performance of metagenomics-based diagnostics

To develop a diagnostic model based on metagenomic data, gradient-boosted tree classifiers were trained and optimised to predict if the taxonomic classification of a blood metagenome belonged to a septic patient (n = 117) or a healthy control (n = 167). Taxonomic *k*-mer-based classification of metagenomes in *KrakenUniq* [23] resulted in an abundance matrix with rows as samples and genera as columns (i.e. features). To maximise classification performance, models were trained and evaluated with four different representations of this abundance matrix (i.e. feature spaces). Genera abundance was either represented as the number of unique *k*-mer counts assigned by *KrakenUniq* ('absolute' abundance) or as relative abundance (*Raw-?* and *RA-?* models, respectively). Models were also trained on feature spaces with or without application of our contamination reduction step (*?-None* or *?-CR* models respectively). The model names, normalisation techniques and number of features used are summarised in Table 1. Generally, all classifiers had a high classification performance with areas under the receiver operating characteristic curve (AUROC) > 0.9 in predicting if a patient suffered from sepsis (Table 1). The classifier trained on 'absolute' abundance without any decontamination steps performed (*Raw-None* model) had the highest performance (AUROC = 0.999, 95% CI: 0.997-1.000). In general, classifiers trained on 'absolute' abundance (*Raw-?* models) performed the best whether or not contamination reduction was applied.

Table 1. Model summary. Total *k*-mer count refers to the total number of unique *k*-mers classified for a particular sample. The prefix and suffix of each model name corresponds to the normalisation and contamination reduction technique applied respectively. *Raw*, *RA*, *None* and *CR* denote 'absolute' abundance, relative abundance, no contamination reduction applied and contamination reduction applied, respectively.

No. of Features	Normalisation	Model Name	Model Performance		
			Sensitivity	Specificity	AUROC
1056	-	*Raw-None*	1.000	0.980	0.999
1056	$\frac{k-mer\ count}{Total\ k-mer\ count}$	*RA-None*	0.943	0.941	0.993
23	-	*Raw-CR*	0.914	0.941	0.943
23	$\frac{k-mer\ count}{Total\ k-mer\ count}$	*RA-CR*	0.886	0.922	0.959

Contamination reduction

Our contamination reduction technique was built upon SHapely Additive exPlanations (SHAP) – a state-of-the-art machine learning technique for interpreting 'black-box' classifiers such as gradient-boosted trees [24]. Briefly, each feature in a single sample is assigned a SHAP value, which corresponds to the change in a sample's predicted probability (diagnostic) score when the feature is either present or absent. Using SHAP values allow the decomposition of predicted probability scores for each sample into the sum of contributions from individual genera (demonstrated in Fig 1a).

Fig 1. Model interpretation using SHAP values. (a) Force plot providing the decomposition of predicted probability scores for a septic patient (run accession: SRR828875) with a 'confirmed' *E. coli* infection. Each genus is assigned a SHAP value which increases (red) or decreases (blue) the predicted probability score. (b) Plot summarising the SHAP values across all samples for the 23 most important features ranked by the mean absolute

SHAP value (highest at the top) for *Raw-None*, and for (c) *Raw-CR* 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 value (i.e. unique *k*-mer counts).

The premise behind our technique is that pathogens should occur at higher abundance in septic patients relative to healthy controls. Given this, the 'absolute' abundance of contaminant taxa would demonstrate a negative Spearman's correlation with their corresponding SHAP values. That is, a contaminant genus would contribute to a lower predicted probability score at high abundance. This allows putative contaminant genera to be computationally removed. We observed that without employment of this contamination reduction step, 14 of the 23 most important genera had negative SHAP values at a high relative abundance (*Raw-None* model; Fig 1a). These genera were known sequencing contaminants (e.g. *Mesorhizobium* and *Ralstonia*) [19] and/or did not contain probable blood pathogens. After contamination reduction, 1033 genera were removed and the remaining 23 genera all had positive SHAP values at high abudance. Despite the large reduction in dimensionality of the feature space, classifiers performed remarkably well (*?-CR* models; AUROC > 0.9). Further, 10 of the 23 remaining genera (*Raw-CR* model; Fig 1b) matched that of the 'confirmed' list of causative pathogens determined via microbiological testing in the Karius study (summarised in Fig 2). Following this contamination reduction step the best performing model was *RA-CR* (AUROC = 0.959; 95% CI 0.903-0.991).

Fig 2. Barplot of 'confirmed' pathogens in 117 sepsis patients following microbiological tests. This data was obtained from Supplementary Table 5 of the Karius study. Bars provide the assigned genus (coloured legend), with the number of samples (y-axis) provided atop.

Evidence for a polymicrobial community

SHAP values were also used to determine the extent to which the genera containing 'confirmed' pathogens drove classifier predictions. Most septic patients had 'confirmed' *E. coli* infections. Since SHAP values represent the change in probability when a feature is absent or present, we subtracted SHAP values for the

genus *Escherichia* from predicted probability scores for the *Raw-CR* model. Doing so resulted in a modest decrease in AUROC score from 0.943 to 0.877, suggesting that these predictions were not solely driven by a single taxon.

Additionally, to visualise how similar septic metagenomes were, unsupervised clustering of the relative genera abundance was performed using t-distributed stochastic neighbour embedding (t-SNE) [25]. t-SNE attempts to project the data on a two-dimensional latent space where data points with similar Euclidian distances are clustered and those with dissimilar distances further apart. Septic patients with 'confirmed' infections by bacteria from the genera *Escherichia*, *Proteus*, *Enterococcus* and *Herpesviridae* family had metagenomic taxa profiles that appeared as visually distinct clusters (Fig 3a). However, these genera often account for a large proportion of microbial reads (S1 Fig), which could largely drive the clustering signal observed. Despite this, even after exclusion of these genera from the analysis, septic metagenomes could still be distinguished from that of healthy controls (Fig 3b).

Fig 3. Scatter plots of t-SNE projections. (a) Genera corresponding to the feature space used in the *RA-CR* model and (b) the same feature space with *Escherichia*, *Proteus*, *Enterococcus*, *Lymphocryptovirus* and *Cytomegalovirus* removed. Each point represents the projection of the full metagenomic profile of a patient. Healthy samples were coloured orange. A subset of the septic samples were colored according to the genera of the 'confirmed' causative pathogens.

Lastly, microbial networks were used to determine if microbial genera detected in septic samples co-occured. 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. 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 (Fig 4). This network was constructed by subtracting all edges of the healthy network from the sepsis network. Multiple co-occurrence relationships remained in the corrected network, including genera containing 'confirmed' pathogens such as *Escherichia*,

Aeromonas and *Enterococcus*. Interestingly, we observed co-occurrence clusters within genera broadly associated with the mouth, lungs and gut (Fig 4).

Fig 4. Corrected microbial network for genera assigned in sepsis metagenomes. Input data corresponds to the features used for the *RA-CR* classifier. The edges in this network represent those in the septic network that were not present in the healthy network. Co-occurrence and co-exclusion relationships are represented by green and red edges respectively.

Discussion

Computational contamination reduction

Contamination from environmental sources poses one of the greatest challenges for metagenomic investigations of microbial communities, particularly in low biomass and clinical samples [20]. Our results demonstrate the efficacy of our novel post-hoc contamination reduction technique in removing redundancy in the feature space while selectively retaining taxa involved in sepsis. These results also suggest the possibility that sepsis can be reduced to a distinct set of pathogens sufficient for diagnosis. Further, removal of putative contaminant genera is necessary to protect against spurious signals. As such, the effectiveness of our contamination reduction technique provides greater confidence that the polymicrobial signals we observe were not driven by contamination. However, 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 along the lines of those proposed by Jervis-Bardy et al. [26] could be used.

Metagenomics-based diagnostics

Our results demonstrate the clear promise of metagenomics-based diagnostic models. To put this 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 [27]. Notably, it is difficult to compare between 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 blood metagenomic data alone contains sufficient information for the diagnosis of sepsis.

The best classification performance was achieved when no normalisation or computational contamination reduction was applied. Additionally, models trained on relative abundance had poorer classification performance (Table 1). The only difference between 'absolute' and relative abundance is information regarding the total genera abundance in each sample. As such, microbial load appears important for classification, which is in line with the prevailing understanding that microbial infection and colonisation is characterised by microbial proliferation [28]. Models also performed more poorly when contaminant genera were computationally removed. A possible explanation is that the abundance and types of contaminants sequenced increase with lower microbial loads [26]. Acording to the Karius study, the microbial cfDNA concentration in healthy samples is markedly lower than that of septic samples. Hence, levels of contamination could be useful predictors of sepsis and removal of such information would lower classification performance. Nevertheless, even when classifiers were provided with only 23 curated genera, they performed remarkably well. Considering that our contamination reduction technique was adequate in removing contaminant genera, classifiers would have been forced to classify samples mainly from the abundance of genera of likely blood pathogens. Thus, the results cumulatively suggest the presence of a distinct set of sepsis-related pathogens.

As demonstrated in Fig 2a, SHAP values allow interpretation of which features contribute the most to each classification instance. This information has potential value for clinicians in diagnosing the likely causes of bacteraemia and could be used to inform personalised patient care. However, without applying contamination reduction, causal interpretations based on SHAP values are prone to high false positive rates which comes as a trade-off for increased diagnostic sensitivity. Conversely, applying contamination reduction provides a higher precision of causal interpretations at the expense of sensitivity. These considerations are crucial in future studies that seek to develop metagenomic-based diagnostic models. A more comprehensive evaluation of the effectiveness of SHAP values in a clinical context is necessary but goes beyond the scope of this study.

Evidence for a polymicrobial community

In the Karius study, only seven of the 117 septic patients were found to have more than one causative pathogen identified by microbiological testing (Supplementary Table 5; Karius study). Here, we provide three main lines of evidence suggesting that these infections were more likely to be polymicrobial in nature.

Firstly, we would expect that if the metagenomes obtained from sepsis patients have certain genera assigned at a markedly higher relative abundance, as would be expected due to single pathogen infections, these features would drive the generation of distinct clusters under t-SNE unsupervised clustering. Instead, we observed that even after the removal of predicted 'unimicrobial' infecting genera *Escherichia*, *Proteus*, *Enterococcus*, *Cytomegalovirus* and *Lymphocryptovirus*, a distinct cluster of septic samples remain (Fig 3). This suggests that the relative abundance of 'non-causative' genera in septic samples are more similar to each other than to that of healthy samples.

Secondly, SHAP values were used to interpret the contribution of features (genera) to model predictions. Of note, meaningful inference from SHAP values was only possible given the high classification

performance of all models (Table 1). With this approach, we observed multiple genera contributing significantly to each predicted probability score. Additionally, even after removal of *Escherichia*, the most prevalent 'confirmed' cause of infection in the Karius dataset, classifiers were still able to distinguish septic from healthy metagenomes reasonably well. This implies a considerable association between the 'non-causative' genera present and patient health status.

Thirdly, microbial networks were employed to identify the co-occurrence or co-exclusion of genera in health and disease (Fig 4). Despite the multiple levels of stringency employed, the large number of associations between genera in the final network suggest that many genera tend to co-occur only in septic samples.

Given that pathogens associated with sepsis are highly heterogenous, the multiple co-occurrence relationships between genera observed suggest that there may be a set of pathogen communities common to the disease. Although our networks were inferred computationally, there is evidence in the literature supporting the possibility of synergies between some of the co-occuring genera. For example, although we inferred *Proteus* and *Enterococcus* to be part of a cluster of gut pathogens in the corrected sepsis network (Fig 4) , these two genera are also known to contain synergistic uropathogens. Indeed, *E. faecalis* has been shown to increase urease activity of *P. mirabilis*, a major factor driving tissue damage and ingress to blood in urinary tract infections [29]. Separately, *Stenotrophomonas*, *Burkholderia* and *Pandoraea* are known to play a collective role in the pathogenesis of cystic fibrosis. Interestingly, *P. aeruginosa*, although removed as a contaminant in our analyses, is known to facilitate colonisation of *S. maltophilia* [30] and upregulate virulence factors in *B. cepacia complex* [31]. An important point here is that this indirect association between the two species mediated by *P. aeruginosa* could appear as a direct association in microbial networks. However, experimental validation is required to link these findings to the *Stenotrophomonas-Burkholderia* association detected in the corrected septic network (Fig 4).

The presence of oral, lung and gut clusters may point to potential infection sources in sepsis. In addition, it suggests the possibility of opportunistic infections and dysbioses that could affect disease severity. Considering that commensals such as *Blautia spp.* are known to produce short-chain fatty acids and aid systemic immunity [32], co-infection by these 'protective species' might exacerbate present illness due to current infections resulting in patient deterioration. In this way, bioinformatic and machine learning approaches can assist the generation of hypotheses for work to functionally elucidate the complexities of polymicrobial communities in sepsis.

Limitations

There are several limitations to this study, which can broadly be divided into the limitations intrinsic to our analyses and technical limitations of the Karius study. The former are difficult to overcome given the theoretical basis of the approaches employed. The latter could be circumvented via improved experimental designs.

We identified several limitations intrinsic to our analyses. Firstly, metagenomic sequencing involves measurements of cfDNA 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 [33] or shotgun metagenomic sequencing with culture [22,34,35]. 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 aetiological contributions of these genera and their ecological relationships cannot be inferred. Such hypotheses must be confirmed experimentally and we outline some

potential approaches below. 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 of sepsis. However, furthering our understanding of the microbial component of sepsis may prove useful in the development of better prognostic tools.

We implemented several stringency measures to protect against spurious signals arising from contamination. This could have resulted in the exclusion of taxa that have a role in the pathogenesis of sepsis. Further, species within genera including *Escherichia* and *Pseudomonas* could contain both biologically-relevant and contaminant reads. As such it is expected that a proportion of DNA molecules, and hence sequencing reads, may have come from contamination rather than endogenous microorganisms in blood. The abundance of these microorganisms, as detected by metagenomic approaches, may differ from the true abundance.

Lastly, *k*-mer based approaches may be less accurate for taxonomic classification compared to, for example, Bayesian sequence read-assignment methods [36]. 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 [37]. However, we appreciate that *k*-mer based classification approaches are significantly faster [38], which may provide clinically relevant turnaround times that are important in sepsis diagnostics.

We further identified a number of technical limitations to the Karius study. The definition of sepsis used by the Karius study corresponds to that put forth by Bone et al. [3], which is the fulfilment of at least two SIRS criteria and a confirmed blood infection. As this definition is broad, meeting these criteria does not necessarily lead to adverse long-term clinical outcomes [5]. Therefore, classifiers trained on labels generated by this definition cannot discriminate between mild and severe cases of sepsis. The corollary is that the diagnostic predictions of such classifiers must be used with caution since a positive diagnosis may prompt unnecessary clinical interventions. The ideal classifier must instead be able to discriminate

between healthy patients, those with mild symptoms and those with severe clinical complications as a result of infection. Though, to date metagenomic data from septic patients with detailed curation of symptoms and clinical outcomes is lacking, preventing the development of such models.

Healthy blood cultures in the Karius dataset were obtained from third-party sources (Serologix and Stemexpress), and it is unclear whether samples have gone through adequately stringent tests for bacteraemia. As such we cannot formally rule out that a subset of 'healthy' controls may have been bacteraemic. Further, observed microbial cfDNA signatures delineating septic and healthy samples could in part arise from differences in sample preparation techniques.

According to the Karius study, the microbial load in septic blood was much higher than that of healthy controls. If contaminant signals are amplified in samples with lower biomass, higher levels of contamination would be present in healthy metagenomes. This poses a challenge when comparing septic and healthy blood samples. Finally, we acknowledge the relatively small size of the dataset 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.

Irrespective of these limitations, our results nonetheless demonstrate the importance of the microbial component of sepsis and suggest that taking a metagenomics-based approach may provide biological insights and allow the future development of diagnostic and potentially prognostic 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 aetiologies. Finding evidence for this potentially allows us to redefine sepsis to include a microbial component. However, to do so, a much larger sepsis cohort must be recruited for adequate statistical

power, particularly considering the true number of sepsis subtypes is unknown. The associations between polymicrobial communities with the continuum of host response severity from mild to lethal could also be investigated. To do this, anonymised health-care records with detailed curation of the clinical outcomes of each patient would be required. In addition, pre-infection data from animal models holds promise to identify taxa important in identifying the early onset through to outcomes of disease (prognosis).

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).

Methods

Dataset

The published metagenomic sequence data from the Karius study consists of 117 septic patients and 170 healthy controls [22]. 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 microbiological testing performed within seven days after collection of the blood samples. 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. In all cases,

metagenomic DNA was extracted from blood plasma. The list of 'confirmed' sepsis cases and the causal pathogens can be found in Supplementary Table 5 of the Karius study, under the 'definite' adjucation.

Data pre-processing

According to the Karius study, the input 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). The reads were then quality trimmed using Trimmomatic (v0.32; parameters unknown) retaining reads with a quality (Q-score) above 20. Mapping and alignment were performed using Bowtie (v2.2.4; parameters unknown). Human reads were identified by mapping to the human reference genome and removed prior to deposition in NCBI's Sequence Read Archive (SRA). In our analyses, healthy samples corresponding to the run accessions: SRR8288965, SRR8289465 and SRR8289466 were excluded due to low sequencing depth (2331, 494 and 1770 bases respectively). This resulted in a final dataset of 117 septic samples and 167 healthy controls. No further quality control steps were performed on the sequencing data prior to our analyses.

Metagenomic assignment was performed using *KrakenUniq* (v0.5.8; default parameters) [23]. The *standard database* used in all analyses was built using *kraken-build* on January 20th, 2020. *KrakenUniq* matches short sequences from reads (*k*-mers) to a reference database and subsequently assigns each *k*-mer to a taxon. The resultant classification output contains the number of unique *k*-mers mapping to each taxon (unique *k*-mer count). Unclassified *k*-mers were not included in any of the analyses. The final data matrix used in downstream analyses had samples represented in rows and genera in columns.

Computational contamination reduction

Computational contamination reduction was performed in three steps. Firstly, a classifier was optimised (*Raw-None* model) and the spearman's correlation between the SHAP values of each feature and corresponding unique *k*-mer counts were computed. Features with significantly ($p < 0.05$) negative

correlations were removed. Subsequently a new model was retrained using the previously optimised set of parameters but with this new reduced feature space. This process was repeated for a total of five iterations (see Supplementary material for list of genera remaining at each iteration), resulting in 28 genera remaining. No multiple-testing correction was performed to increase the stringency of contaminant removal. Spearman's correlations and *p*-values were calculated using *spearmanr* as part of the *SciPy* library (v1.4.1).

Secondly, genera not currently identified as known human pathogens were removed, yielding 25 genera. The selection was based on a preprint 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.

Lastly, by visual inspection, *Agrobacterium* and *Cellulomonas* were removed due to irregular patterns of SHAP values and because both represent unlikely bloodborne pathogens (S2 Fig). Normalisation to relative abundance was only applied after all three contamination reduction steps (*RA-CR* model).

Model training, optimisation and evaluation

Gradient-boosted tree models were trained with a binary-logistic loss function and implemented using XGBoost API (v0.90) [40]. The dataset was randomly split into training and test datasets consisting of 82 septic and 116 healthy samples (70%), and 35 septic and 51 healthy samples (30%) respectively. For each model, the training dataset was used to optimize model parameters, namely: *n_estimators*, *max_depth*, *gamma*, *subsample* and *colsample_bytree*. This was implemented using a stratified 5-fold cross-validation protocol in *GridSearchCV* as part of the *Scikit-learn* library (v0.22.1). The best performing sets of hyperparameters that maximise the AUROC of each model were determined. Subsequently, all models were trained on the full training dataset with their corresponding optimised hyperparameters. Finally, the models were evaluated using the test dataset.

Confidence intervals for AUROC scores were generated using a non-parametric bootstrap method. The test dataset was resampled with replacement to generate a dataset of equivalent number of samples. AUROC was then computed for each optimised model. This process was repeated 1001 times and the lower and higher confidence interval bounds were taken to be the 25th and 975th score respectively.

Model interpretation

TreeExplainer, part of the *shap* library (v0.34.0) [24], was used to determine the relative importance of each feature (*i.e.* the unique *k*-mer count of a genus) in tested models. For example, SHAP values for the *Raw-CR* model were generated from the test dataset by setting the *feature_pertubation* parameter to 'interventional'. SHAP values corresponding to *Escherichia* were retrieved and subtracted from the predicted probability scores for the test dataset. The model was subsequently re-evaluated. For contamination reduction and for model interpretation, SHAP values were computed using the training dataset and testing dataset respectively. Relative feature importance 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 model predictions.

t-SNE

The Barnes-Hut implementation of t-SNE was performed using the *R* package *Rtsne* (v0.15) [25]. The number of iterations for each t-SNE run was chosen to ensure convergence (10,000 iterations). Perplexity was set to 15. The input features corresponded to that for the *RA-CR* model. The feature columns of *Escherichia*, *Proteus*, *Enterococcus*, *Lymphocryptovirus* and *Cytomegalovirus* were only removed after normalisation to yield relative abundance.

Microbial networks

Microbial networks were generated using *seqgroup* (v0.1.0), an *R* implementation of *CoNet* (https://github.com/hallucigenia-sparsa/seqgroup) [41]. Four levels of stringency were applied to

minimise the likelihood of detecting spurious associations: (A) An ensemble of Spearman's correlation, Kullback-Leibler and Bray-Curtis dissimilarity measures was used. *p*-values were generated for each measure and merged via Fisher's method. An association was only retained if at least two measures agreed on whether the taxa co-occur or co-exclude. This allowed for more reliable detection of microbial associations. (B) The REBOOT method was used for computing a *p*-value for Spearman's correlation to protect against spurious correlations when analysing relative taxon abundance [42]. These spurious correlations arise because, given that the relative abundance of each taxon in a metagenome must sum to one, the decrease in abundance of one taxon would result in an increased abundance of other taxa. (C) Multiple-testing correction via the Benjamini-Hochberg procedure [43] was used to limit the false discovery rate to 0.05. This was necessary since there were $(n-1)!$ possible pairwise associations between the genera, where n is the number of genera. (D) Our contamination reduction technique was applied before performing this analysis. This increases the power for detecting biologically relevant associations by removing putative contamination signals.

The *init.edge.num* parameter was set to 40. Permutation tests and REBOOT were performed over 1000 iterations. Vertices of the network represent individual genera. Edges represent co-occurrence (positive correlation / high similarity) or co-exclusion (negative correlation / low similarity). The corrected septic network was constructed by subtracting edges present in the healthy network from the septic network. All input data used to construct the networks corresponds to that used in the *RA-CR* classifier.

Hardware and software

All analyses with the exception of metagenomic classification were performed on a local machine with 12 processors and 32Gb RAM. KrakenUniq was run on a compute node in a remote server using 300Gb RAM and 12 processes. Data parsing, model training, optimisation, and evaluation were performed with *Python* 3.7. Microbial network analysis and t-SNE was performed with *R* 3.6.0.

Data and code availability

All relevant source code can be found on Git (https://github.com/cednotsed/Polymicrobial-Signature-of-Sepsis). Metagenomic sequencing reads for the Karius dataset are available at BioProject identifier PRJNA507824. The parsed data matrix for the *Raw-None* model can be found in the supplementary materials (Table S1).

Acknowledgements

We would like to thank Dr. Sivan Bercovici from Karius Inc. for his helpful comments.

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. Breitwieser FP, Baker DN, Salzberg SL. KrakenUniq: confident and fast metagenomics classification using unique k-mer counts. Genome Biol [Internet]. 2018;19(1):198. Available from: https://doi.org/10.1186/s13059-018-1568-0

24. 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

25. Maaten L van der, Hinton G. Visualizing data using t-SNE. J Mach Learn Res. 2008;9(Nov):2579–605.

26. 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.

27. 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.

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

29. Armbruster CE, Smith SN, Johnson AO, DeOrnellas V, Eaton KA, Yep A, et al. The Pathogenic Potential of Proteus mirabilis Is Enhanced by Other Uropathogens during Polymicrobial Urinary Tract Infection. Infect Immun [Internet]. 2017 Jan 26;85(2):e00808-16. Available from: https://pubmed.ncbi.nlm.nih.gov/27895127

30. Wahab AA, Janahi IA, Marafia MM, El-Shafie S. Microbiological identification in cystic fibrosis patients with CFTR I1234V mutation. J Trop Pediatr. 2004;50(4):229–33.

31. Riedel K, Hentzer M, Geisenberger O, Huber B, Steidle A, Wu H, et al. N-acylhomoserine-lactone-mediated communication between Pseudomonas aeruginosa and Burkholderia cepacia in mixed biofilms.

Microbiology. 2001;147(12):3249–62.

32. Rajilić-Stojanović M, de Vos WM. The first 1000 cultured species of the human gastrointestinal microbiota. FEMS Microbiol Rev. 2014;38(5):996–1047.

33. 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

34. 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.

35. 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).

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

37. 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.

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

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. BioRxiv. 2019;670315.

40. 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.

41. Faust K, Raes J. CoNet app: inference of biological association networks using Cytoscape. F1000Research. 2016;5.

42. Faust K, Sathirapongsasuti JF, Izard J, Segata N, Gevers D, Raes J, et al. Microbial co-occurrence relationships in the human microbiome. PLoS Comput Biol. 2012;8(7).

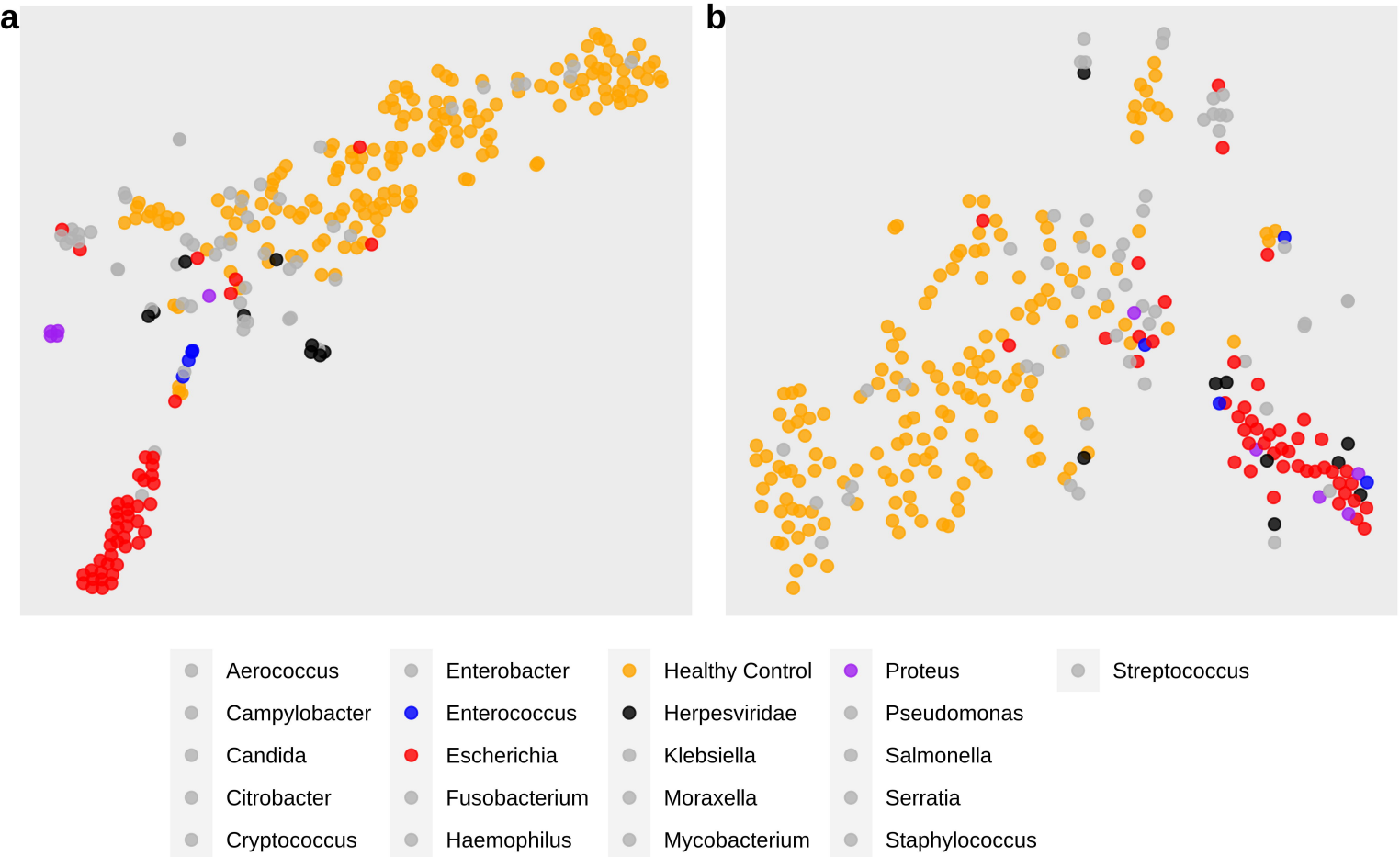
43. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B. 1995;57(1):289–300.

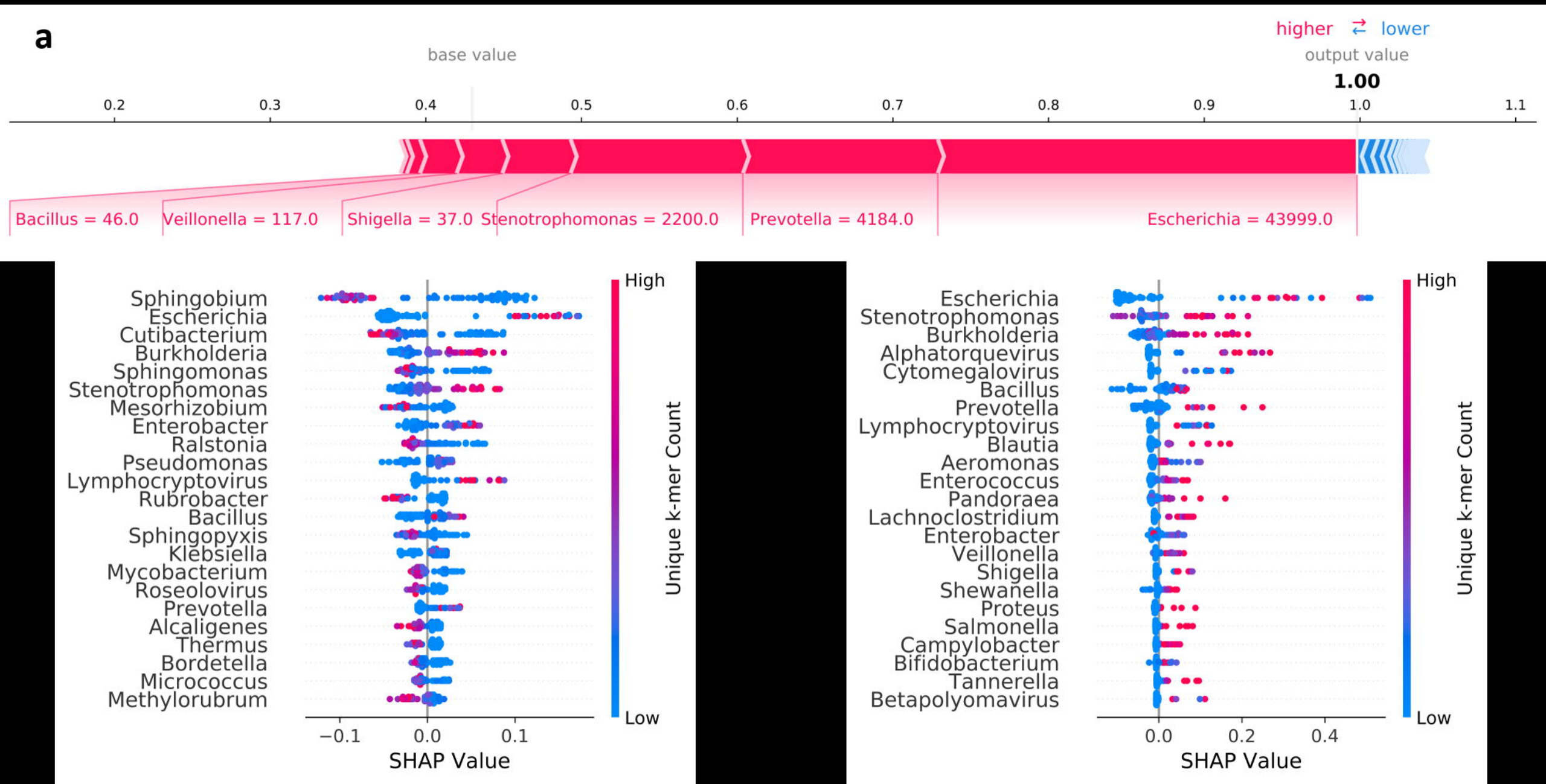
Supporting information captions

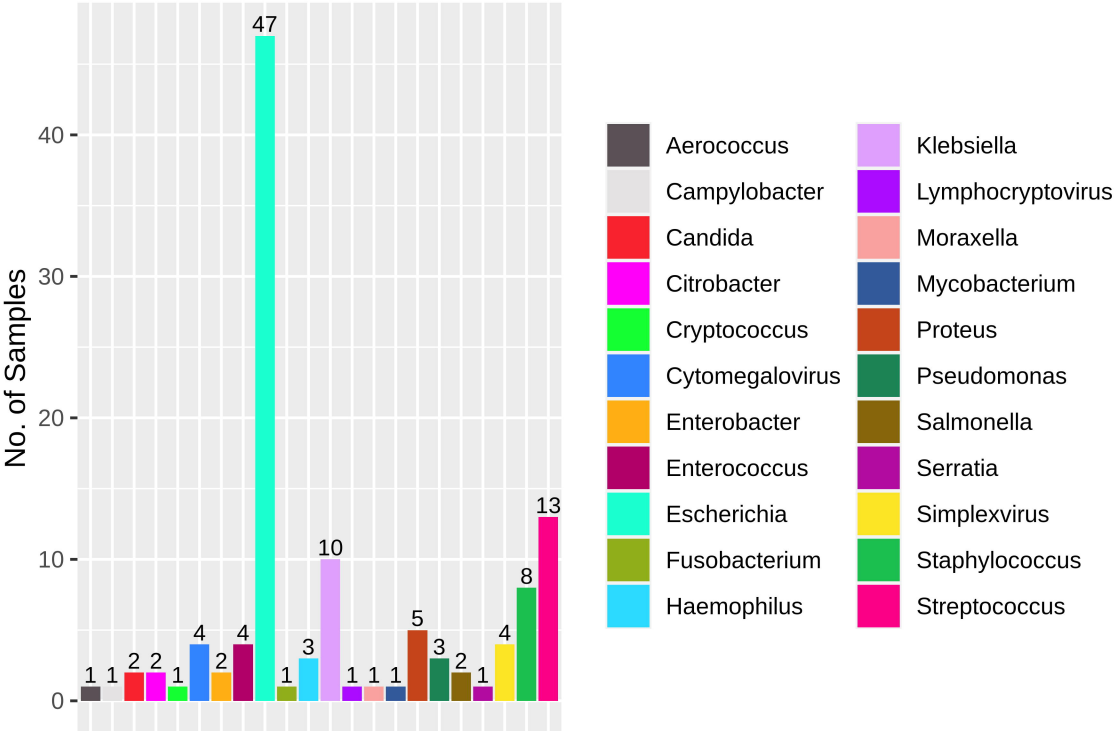
S1 Fig. Boxplot of relative genera abundance. Relative abundance was computed across patients with different 'confirmed' infections. Genera containing the 'confirmed' causative pathogen generally occupies a large proportion of the total number of unique *k*-mers assigned for each sample (*ie.* high relative abundance). The relative abundance of only three features were plotted for clarity.

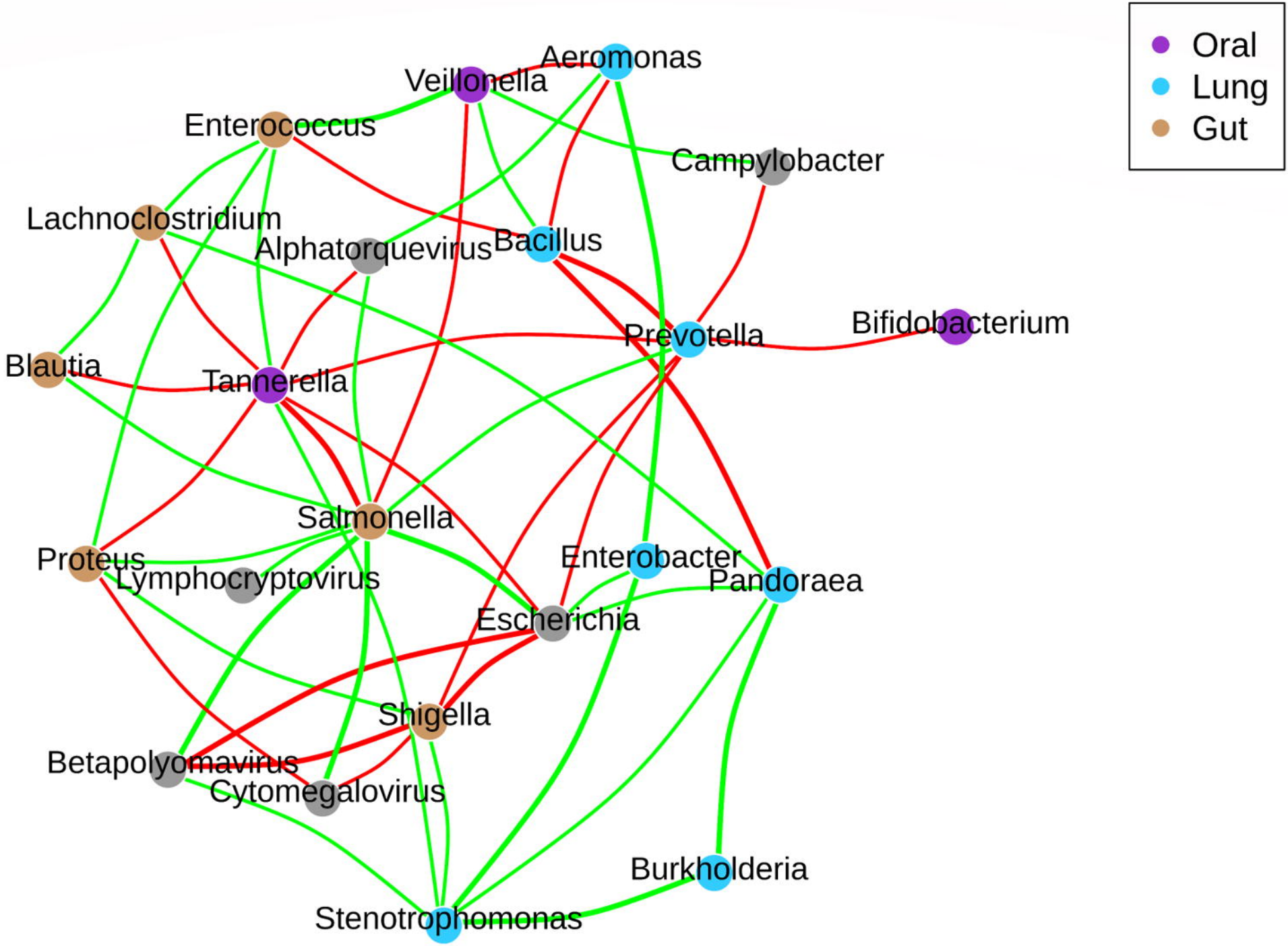
S2 Fig. SHAP summary plot before removal of *Cellulomonas* and *Agrobacterium*. High positive SHAP values at low abundances were observed for both genera. Note that this was also the case for *Enterobacter* but it was not removed as it contained 'confirmed' causative pathogens.

S1 Table. Data matrix (comma-separated) for the Raw-None model. This matrix includes 'confirmed' causative pathogens identified by microbiological testing, sample accession numbers as well as disease state for each sample.


a
b
Aerococcus
Campylobacter
Candida
Citrobacter
Cryptococcus
Enterobacter
Enterococcus
Escherichia
Fusobacterium
Haemophilus
Healthy Control
Herpesviridae
Klebsiella
Moraxella
Mycobacterium
Proteus
Pseudomonas
Salmonella
Serratia
Staphylococcus
Streptococcus


a
higher ⇄ lower
base value
output value
1.00
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1.1
Bacillus = 46.0
Veillonella = 117.0
Shigella = 37.0
Stenotrophomonas = 2200.0
Prevotella = 4184.0
Escherichia = 43999.0
Sphingobium
Escherichia
Cutibacterium
Burkholderia
Sphingomonas
Stenotrophomonas
Mesorhizobium
Enterobacter
Ralstonia
Pseudomonas
Lymphocryptovirus
Rubrobacter
Bacillus
Sphingopyxis
Klebsiella
Mycobacterium
Roseolovirus
Prevotella
Alcaligenes
Thermus
Bordetella
Micrococcus
Methylorubrum
−0.1
0.0
0.1
SHAP Value
High
Low
Unique k-mer Count
Escherichia
Stenotrophomonas
Burkholderia
Alphatorquevirus
Cytomegalovirus
Bacillus
Prevotella
Lymphocryptovirus
Blautia
Aeromonas
Enterococcus
Pandoraea
Lachnoclostridium
Enterobacter
Veillonella
Shigella
Shewanella
Proteus
Salmonella
Campylobacter
Bifidobacterium
Tannerella
Betapolyomavirus
0.0
0.2
0.4
SHAP Value
High
Low
Unique k-mer Count


No. of Samples
0
10
20
30
40
1
1
2
2
1
4
2
4
47
1
3
10
1
1
1
5
3
2
1
4
8
13
Aerococcus
Campylobacter
Candida
Citrobacter
Cryptococcus
Cytomegalovirus
Enterobacter
Enterococcus
Escherichia
Fusobacterium
Haemophilus
Klebsiella
Lymphocryptovirus
Moraxella
Mycobacterium
Proteus
Pseudomonas
Salmonella
Serratia
Simplexvirus
Staphylococcus
Streptococcus


Oral
Lung
Gut
Aeromonas
Veillonella
Enterococcus
Campylobacter
Lachnoclostridium
Alphatorquevirus
Bacillus
Bifidobacterium
Prevotella
Blautia
Tannerella
Salmonella
Proteus
Lymphocryptovirus
Enterobacter
Pandoraea
Escherichia
Shigella
Betapolyomavirus
Cytomegalovirus
Burkholderia
Stenotrophomonas