# Metagenomic evidence for a healthy human blood microbiome

Cedric C.S. Tan

Department of Structural and Molecular Biology

University College London

Gower Street, London, WC1E 6BT, United Kingdom

# Graphical abstract

# Summary (150)

150 words

# Introduction (553)

Human blood has traditionally been considered to be sterile, where the occasional entry of microorganisms is believed to cause bloodstream infections (BSIs) associated with high mortality and morbidity (Goto and Al-Hasan, 2013). However, there have been an increasing number of studies reporting bacterial DNA or even culturable bacteria in the blood of apparently ‘healthy’ individuals (*i.e.* with no clinical symptoms of disease) (reviewed by Castillo *et al.* (2019)). These studies sought to investigate two main questions: is there a blood microbiome? If there is, what does the microbial profile look like? To do so, most studies have employed culture-based or sequencing approaches. The former approach involves the use of microbiological cultures of blood samples; the latter involves either targeted 16S marker-gene, or non-targeted ‘shotgun’ metagenomic sequencing of microbial DNA extracted from blood. Both types of approaches have their advantages and limitations. Culture-based methods demonstrate the viability of detected microorganisms. However, they cannot detect ‘unculturable’ microorganisms and have a notoriously low sensitivity, especially when the bacterial load is low (Benjamin and Wagner, 2007). Additionally, the microorganisms detected are largely biased by the type of culture medium used. This approach therefore prevents a comprehensive characterisation of the breadth of blood microorganisms.

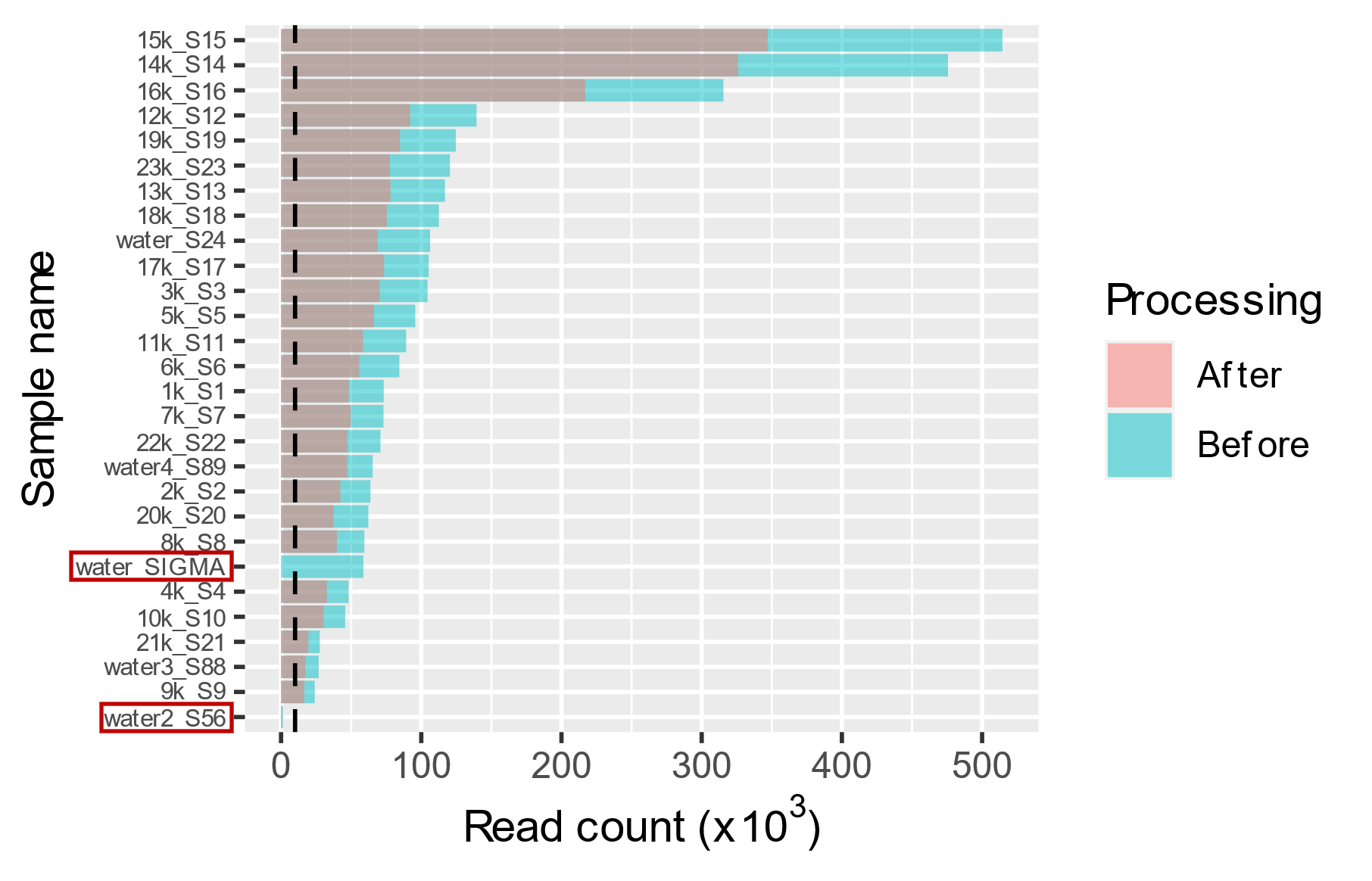
On the other hand, metagenomic sequencing can detect unculturable or low abundance microorganisms, and is therefore a more sensitive approach for profiling the blood microbiome. The main trade-off for this increased sensitivity is that sequencing is highly prone to environmental contamination. Contaminant DNA from laboratory reagents, referred to as the ‘kitome’, can be introduced during sample processing (Salter *et al.*, 2014; Weiss *et al.*, 2014; Glassing *et al.*, 2016). Alternatively, contaminant DNA may be introduced when using non-sterile equipment or poor aseptic technique during sample collection. These contaminating sequences are known to skew biological inferences of metagenomic sequencing data, particularly for sample types with low microbial biomass such as blood (Eisenhofer *et al.*, 2019). Delineating contaminant DNA sequences from those endogenous to blood is therefore crucial in lending weight to arguments for a blood microbiome. However, as Castillo *et al.* (2019) noted, experimental controls for addressing the contamination problem are often omitted in blood microbiome studies. Even when controls are included, they are often only used for quantifying the overall level of contamination introduced by the experimental protocols used (Païssé *et al.*, 2016; Gosiewski *et al.*, 2017). Such controls should ideally be used to identify and remove putative contaminant taxa before making any biological inference of the microbial profiles observed.

In this study, 16S marker-gene metagenomic sequencing (henceforth ‘16S sequencing’) was performed on microbial DNA extracted from the blood of 23 clinically healthy individuals to characterise the hypothetical human blood microbiome. To address the problem of contamination, five negative controls of sterile water were also sequenced, and their taxonomic profiles compared at the genus rank. Since the microbial taxa observed in the negative controls are assumed to be contaminants, these taxa can be selectively omitted to produce a more accurate depiction of the microbial composition in blood. The results suggest that the number and relative abundance of microbial taxa detected differed between the two sample types. This implies that the microbial sequences found in healthy human blood are more likely to originate from endogenous bacteria rather than from contamination sources during library preparation, providing support for the existence of a human blood microbiome.

# Results (1151)

## 16S metagenomic sequencing

Sequencing of the hypervariable region spanning V3-V4 of the bacterial 16S ribosomal RNA gene was peformed to characterise the microbial composition of 23 healthy human blood samples and five negative water controls. A total of 3,205,568 paired-end reads were generated for all 28 samples in a single sequencing run, and 2,120,549 reads were retained after all read processing steps. The distribution of reads per sample before and after these processing steps (**Figure 1**) suggests that most samples had a reasonable sequencing depth and overall read quality. Two negative controls with less than 10,000 reads are considered potentially problematic samples but were not excluded from further analysis to avoid introducing additional biases.



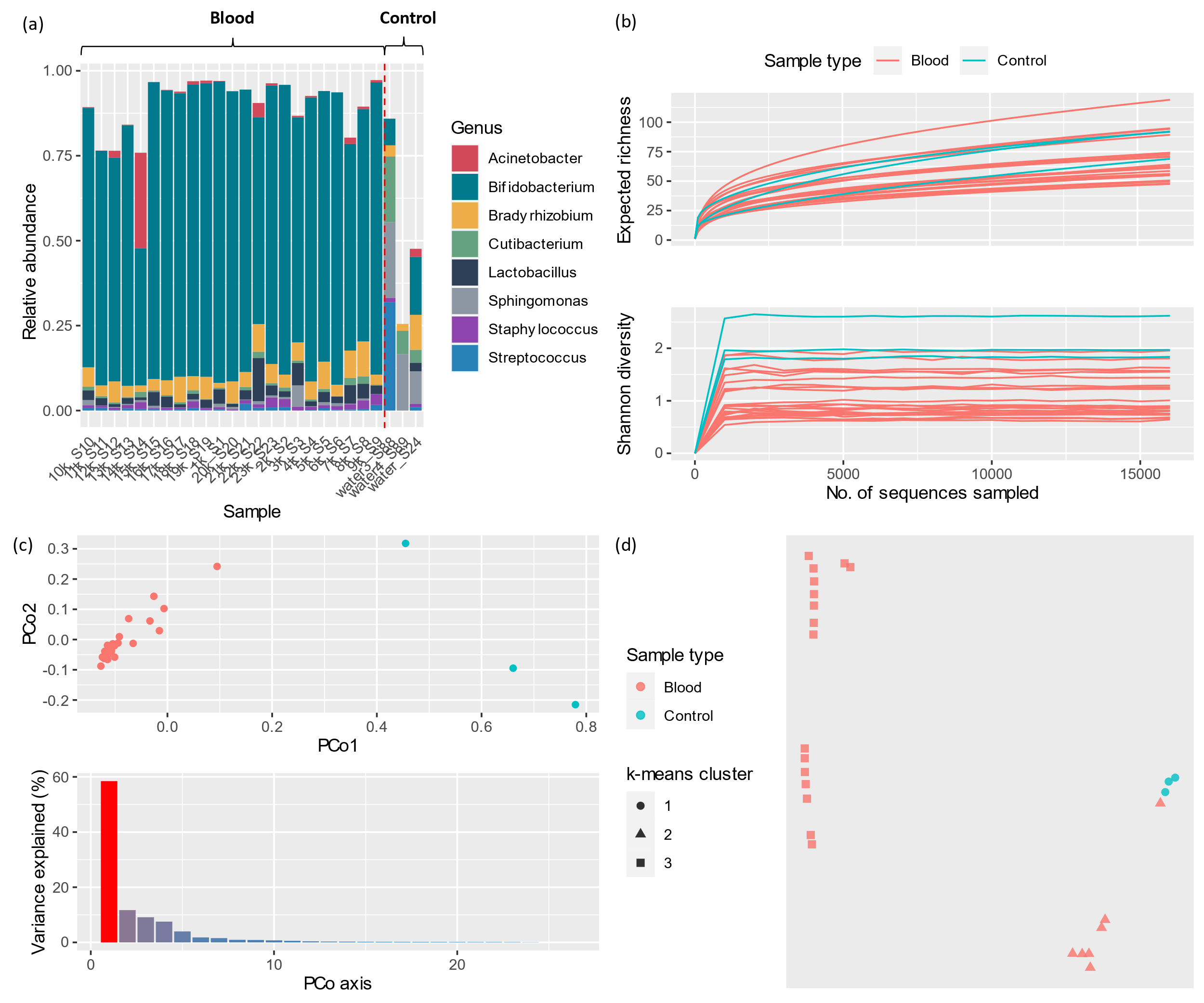
**Figure 1. Per-sample distribution of reads.** The read counts of samples before and after all read processing steps were extracted from the respective FASTQ files and visualised. The black dashed line indicates the read count threshold of 10,000 reads. Negative controls that were excluded are denoted by red boxes.

## Microbial composition of healthy human blood is distinct from that of negative controls

Taxonomic classification of the processed reads for each sample was performed using Kraken 2 (Wood, Lu and Langmead, 2019), and with the SILVA database (Quast *et al.*, 2012). The taxonomic assignments were used to calculate the relative abundance of genera in each sample (see *Methods*). The relative abundance of the eight most abundant genera in blood samples and in controls were visually distinct (**Figure 2a**). For example, *Bifidobacterium* predominates in blood samples, but not in the negative controls. To determine if the diversity of microbial sequences differed between the sample types, the genus richness and Shannon index (Shannon, 1948) for each sample was compared. Rarefaction was performed to account for the different sequencing depths for each sample (**Figure 1**). Blood samples and controls had comparable genus richness, but blood samples had a visibly lower Shannon index (**Figure 2b**). The Shannon index is a measure of both the number of genera (*i.e.* richness) and how equal the relative abundance of genera are (*i.e.* evenness). A higher Shannon index indicates a higher richness and/or evenness. Since the genus richness for the two sample types were comparable, the lower Shannon index for blood samples indicate a lower evenness, which is likely driven by the excess of *Bifidobacterium* sequences.

To further determine whether the microbial composition of the two sample types could be distinguished, a principal coordinates analysis (PCoA) was performed. The first principal coordinate axis separated the two sample types clearly (**Figure 2c**; upper panel) and accounted for a major proportion of variation in the relative abundance data (58.5%; lower panel). Additionally, *t*-distributed stochastic neighbour embedding (*t*-SNE) was used to identify groups of samples with similar microbial compositions in an unsupervised fashion. In this context, *t*-SNE projects the relative genera abundance for each sample on a two-dimensional space, where samples with more similar microbial compositions are brought closer together, and those with dissimilar compositions further apart. Echoing the results of the PCoA, blood samples could be spatially separated from the negative controls (**Figure 2d**). The results of the two analyses collectively suggest that the two sample types have distinct microbial compositions. Interestingly, the blood samples appear to form more than one cluster on the *t*-SNE projection (**Figure 2d**). *k*-means clustering, which can classify samples with similar microbial compositions into groups, confirmed that blood samples could be classified into two groups, in addition to a separate negative control group. This suggests that the hypothetical healthy blood microbiome may be categorised into different subtypes.

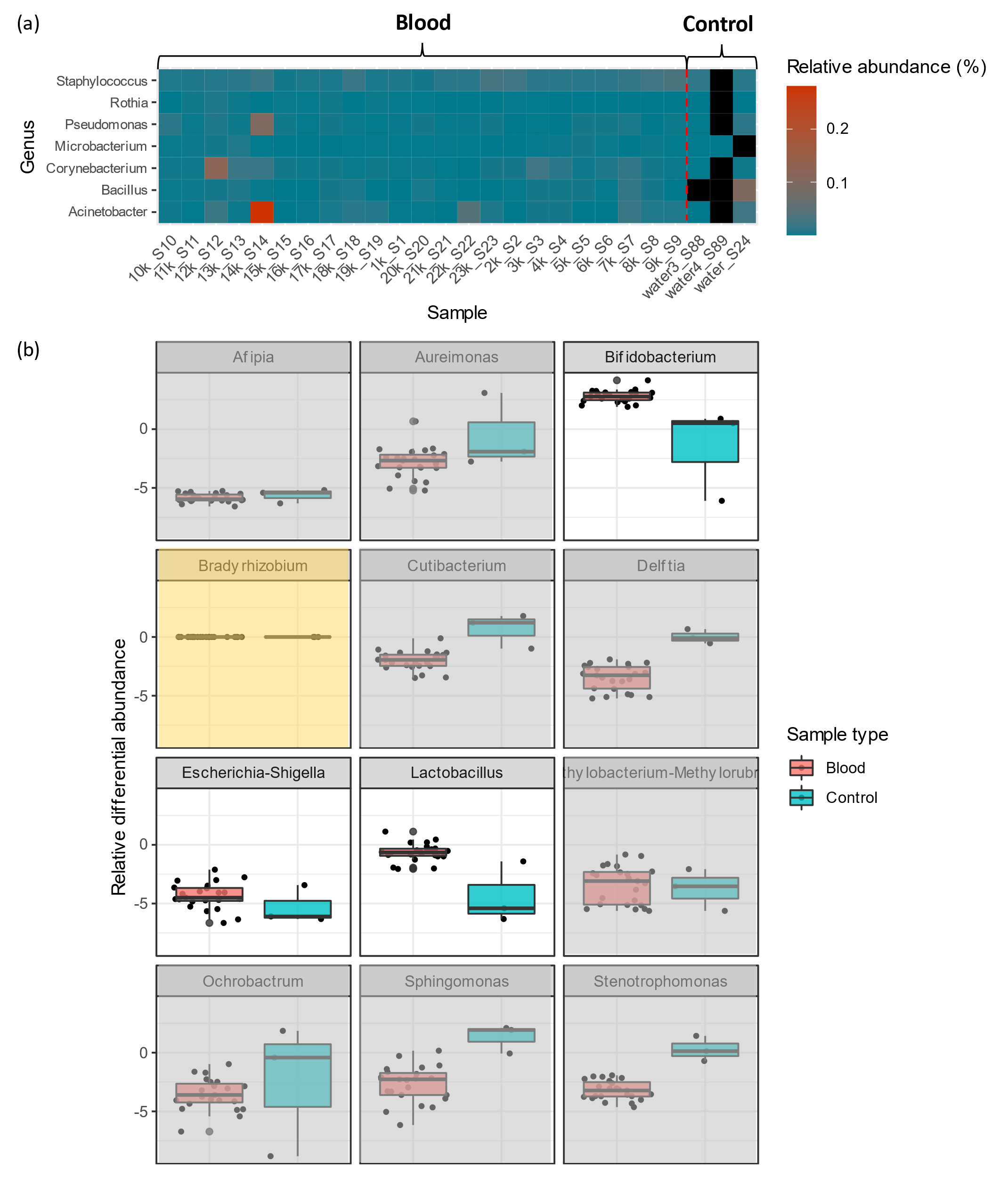
Taken together, the results indicate that the number and composition of microbial genera in blood samples can be distinguished from that in the negative controls. This implies that the microbial sequences in blood are not solely due to contamination from sequencing reagents, suggesting the presence of DNA from potentially viable bacteria in healthy human blood.



**Figure 2. Microbial composition and diversity of healthy blood and negative controls.** (a) Relative abundance of the eight most abundant genera (ranked by their median values across all samples) are visualised in a stacked bar-plot. (b) Rarefaction plots of genus richness and Shannon index. (c) Upper panel: PCoA plot of all samples on the first two PCo axes. Each point represents a single sample after ordination using PCoA. Lower panel: Percentage of variance explained by each PCo axis. (d) *t-*SNE plot projecting the relative genera abundance of samples on a two-dimensional space. Each point represents a single sample after ordination using *t*-SNE. The shape of each point represents the clusters identified using the *k*-means clustering algorithm.

## Candidate members of the human blood microbiome

Since all samples were processed and sequenced in the same batch, the taxonomic assignments for the negative controls could be used to identify potential microbiome-associated genera while excluding putative contaminants. Firstly, seven genera were found in all blood samples, but absent in at least one negative control (**Figure 3a**). Species within these genera present as preliminary targets for members of the blood microbiome. Notably, some genera are present in all samples, but may also be part of the human blood microbiome. One way to determine whether such genera are microbiome-associated or contaminants is to compare their relative abundance in blood and in controls (Salter *et al.*, 2014). This is because taxa that are present in a true microbial community are likely to be highly abundant in microbiome samples and less abundant in negative controls. Conversely, contaminant taxa are likely to be highly abundant in controls but not in samples. Therefore, the measured relative abundance of taxa is a good proxy for selecting candidate members of the human blood microbiome. However, if there is a microbial community in blood, blood samples should have a higher microbial load than negative controls. This differing microbial load is known to confound comparisons of relative abundance (Morton *et al.*, 2019). As such, relative differential abundance (RDA), where the relative abundance of a particular taxon is normalised by a reference taxon in the same sample (see Methods), was compared. RDA controls for the differing microbial loads in samples when comparing relative abundances. Using *Bradyrhizobium* as the reference genus, the RDA of genera that were found in all samples were computed. Of the twelve genera that were present in all samples, only *Bifidobacterium*, *Lactobacillus* and *Escherichia-Shigella* appeared to be more abundant in blood than in controls (**Figure 3b**). These genera are therefore candidate members of the human blood microbiome.



**Figure 3.** (a) Heatmap showing the relative abundance of genera that are found in all blood samples but not in all negative controls. Each tile represents the relative abundance of a single genus for a particular sample. Black tiles indicate a relative abundance of zero. (b) Boxplots of RDA for genera that were present in all samples. Each point represents a single sample. Panels that are not likely candidate members of the human blood microbiome are shaded grey, and the reference frame is shaded yellow.

# Discussion

## The blood microbiome

The results suggest that the microbial genera sequenced from healthy human blood samples differs markedly from that of negative controls.

-Negative control is for sequencing only. Phlebotomy or DNA extraction. Povidone control.

- Other methods, ensemble for better stringency

-Microbiome subtypes. Reference other examples. Emphasise preliminary

-Small data size, not meaningful for statistical tests.

-16S so bacteria only. There is also virome

- Possible sources of blood bacteria

- Transient or permanent?

- DNAemia or viable bacteria?

## Addressing the problem of contamination

Contamination from environmental sources poses one of the greatest challenges for metagenomic investigations of microbial communities, particularly in low biomass and clinical samples (Glassing *et al.*, 2016; Bharucha *et al.*, 2020). It is therefore crucial to discriminate between contaminants and biologically relevant taxa and to remove putative contaminants to protect against spurious signals.

As an alternative to our contamination reduction technique, statistical decontamination techniques identifying inverse relationships between the assigned abundance of taxa and sample DNA concentration (Jervis-Bardy *et al.*, 2015; Davis *et al.*, 2018) could be used. However, this method was not applicable for our study since the sample DNA concentrations in the datasets used were not reported.

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 (Salipante *et al.*, 2013) or shotgun metagenomic sequencing with culture (Brenner *et al.*, 2018; Blauwkamp *et al.*, 2019; Grumaz *et al.*, 2020). This suggests some level of agreement between the presence of microbial cells and their DNA in blood. 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 were not trained to 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 (Morfopoulou and Plagnol, 2015). 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 (McIntyre *et al.*, 2017). We also appreciate that *k*-mer based classification approaches are significantly faster (Simon *et al.*, 2019), 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 a metagenomics-based approach may provide biological and clinical insights supporting the future development of rapid diagnostic tools.

## Clinical implications of the blood microbiome

# Methods

The taxonomic assignments for the reads were aggregated to generate an abundance matrix where samples are represented as rows and genera as columns. Each element in the matrix represents the number of reads assigned to a particular genus for a single sample. This abundance matrix was used

## Sample collection

Phlebotomies were carried out on 13 female and 10 male volunteers (median age = 59). These individuals had no clinical symptoms of bloodstream infections and had normal levels of the C-reactive protein inflammatory marker. All volunteers were not undergoing antibiotic treatment prior to sample collection. Blood samples were collected in 4ml Vacutainer K3E (BectonDickinson) tubes.

## DNA extraction and quantitation

Microbial DNA was extracted from blood using the protocol described previously according to the method described by Gosiewski et al. with the employment of a ready-to-use Blood Mini kit (A&A Biotechnology).

The concentration and purity of total DNA isolates in the samples were measured spectrophotometrically (NanoDrop, Thermo Scientific) at wavelengths of A260 and A280.

## 16S library preparation and sequencing

## **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) (Bolger, Lohse and Usadel, 2014) retaining reads with a quality (Q-score) above 20. Mapping and alignment were performed using *Bowtie* (v2.2.4) (Langmead *et al.*, 2009). 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) (Bushnell, 2014) was used to trim adapter sequences, remove reads with a Q-score below 20 and remove reads mapping to the masked human hg19 reference (https://tinyurl.com/yya4xmrg). For the Gosiewski-17 dataset, we performed the same pre-processing steps as reported in the associated study (Gosiewski *et al.*, 2017). Briefly, primers and adapters were removed using *Cutadapt* (v1.18) (Martin, 2011), paired reads merged using *ea-utils* (v1.1.2.537) (Aronesty, 2013), merged reads and forward unmerged *fastq* files concatenated, and reads with a Q-score below 20 removed using *BBMap*.

## Taxonomic classification

Taxonomic classification of all shotgun sequencing data was performed using *Kraken 2* (v2.0.9-beta; default parameters) (Wood, Lu and Langmead, 2019)⁠ with the *maxikraken2\_1903\_140GB* database (<https://tinyurl.com/y7zfg9kr>). To mitigate potential misclassification of closely related species (*e.g.* *E. coli* and *Shigella spp.)* during taxonomic assignment, we considered only microbial abundance at the genus rank for downstream analyses. For the Gosiewski-17 dataset, *Kraken 2* with a *Kraken 2*-built *Silva* database was used instead of conventional 16S amplicon metagenomic classification methods (Lu and Salzberg, 2020). Read assignments for all ‘culture-confirmed’ bacterial pathogens using the *maxikraken2\_1903\_140GB* and *Kraken 2*-built *Silva* databases are shown in Fig. S2. 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. S3). This suggests that our model is fairly robust to heterogeneity which may be introduced by the classification step. For downstream analyses, we use the genera assignments based on the *Kraken* *2*-built *Silva* database for the 16S Gosiewski-17 samples. Additionally, all unclassified reads were excluded from the analyses. The feature space obtained directly from *Kraken 2* taxonomic assignmentis denoted by *Neat*.

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, with an average of 0.3 reads per sample (range: 0-7 reads). 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.

## Microbial diversity

## Ordination and clustering n

## Relative differential abundance

# Data and code availability

All relevant source code and parsed datasets used can be found on GitHub (<https://github.com/cednotsed/Polymicrobial-Signature-of-Sepsis>). The raw sequence data for this study was obtained from the study by Gosiewski *et al.* (2017).

# References

Aronesty, E. (2013) ‘Comparison of sequencing utility programs’, *The open bioinformatics journal*, 7(1).

Benjamin, R. J. and Wagner, S. J. (2007) ‘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*. Wiley Online Library, 47(8), pp. 1381–1389.

Bharucha, T. *et al.* (2020) ‘STROBE-metagenomics: a STROBE extension statement to guide the reporting of metagenomics studies’, *The Lancet Infectious Diseases*. Elsevier.

Blauwkamp, T. A. *et al.* (2019) ‘Analytical and clinical validation of a microbial cell-free DNA sequencing test for infectious disease’, *Nature Microbiology*, 4(4), pp. 663–674. doi: 10.1038/s41564-018-0349-6.

Bolger, A. M., Lohse, M. and Usadel, B. (2014) ‘Trimmomatic: a flexible trimmer for Illumina sequence data’, *Bioinformatics*. Oxford University Press, 30(15), pp. 2114–2120.

Brenner, T. *et al.* (2018) ‘Next-generation sequencing diagnostics of bacteremia in sepsis (Next GeneSiS-Trial): study protocol of a prospective, observational, noninterventional, multicenter, clinical trial’, *Medicine*. Wolters Kluwer Health, 97(6).

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

Castillo, D. J. *et al.* (2019) ‘The healthy human blood microbiome: Fact or fiction?’, *Frontiers in cellular and infection microbiology*. Frontiers, 9, p. 148.

Davis, N. M. *et al.* (2018) ‘Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data’, *Microbiome*. BioMed Central, 6(1), p. 226.

Eisenhofer, R. *et al.* (2019) ‘Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations’, *Trends in Microbiology*, 27(2), pp. 105–117. doi: https://doi.org/10.1016/j.tim.2018.11.003.

Glassing, A. *et al.* (2016) ‘Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples’, *Gut pathogens*. Springer, 8(1), p. 24.

Gosiewski, T. *et al.* (2017) ‘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’, *European Journal of Clinical Microbiology & Infectious Diseases*. Springer, 36(2), pp. 329–336.

Goto, M. and Al-Hasan, M. N. (2013) ‘Overall burden of bloodstream infection and nosocomial bloodstream infection in North America and Europe’, *Clinical Microbiology and Infection*. Elsevier, 19(6), pp. 501–509.

Grumaz, C. *et al.* (2020) ‘Rapid Next-Generation Sequencing–Based Diagnostics of Bacteremia in Septic Patients’, *The Journal of Molecular Diagnostics*. Elsevier, 22(3), pp. 405–418.

Jervis-Bardy, J. *et al.* (2015) ‘Deriving accurate microbiota profiles from human samples with low bacterial content through post-sequencing processing of Illumina MiSeq data’, *Microbiome*. Springer, 3(1), p. 19.

Langmead, B. *et al.* (2009) ‘Ultrafast and memory-efficient alignment of short DNA sequences to the human genome’, *Genome biology*. Springer, 10(3), p. R25.

Lu, J. and Salzberg, S. (2020) ‘Ultrafast and accurate 16S microbial community analysis using Kraken 2’, *bioRxiv*. Cold Spring Harbor Laboratory.

Martin, M. (2011) ‘Cutadapt removes adapter sequences from high-throughput sequencing reads’, *EMBnet. journal*, 17(1), pp. 10–12.

McIntyre, A. B. R. *et al.* (2017) ‘Comprehensive benchmarking and ensemble approaches for metagenomic classifiers’, *Genome biology*. BioMed Central, 18(1), p. 182.

Morfopoulou, S. and Plagnol, V. (2015) ‘Bayesian mixture analysis for metagenomic community profiling’, *Bioinformatics*. Oxford University Press, 31(18), pp. 2930–2938.

Morton, J. T. *et al.* (2019) ‘Establishing microbial composition measurement standards with reference frames’, *Nature Communications*, 10(1), p. 2719. doi: 10.1038/s41467-019-10656-5.

Païssé, S. *et al.* (2016) ‘Comprehensive description of blood microbiome from healthy donors assessed by 16 S targeted metagenomic sequencing’, *Transfusion*. Wiley Online Library, 56(5), pp. 1138–1147.

Quast, C. *et al.* (2012) ‘The SILVA ribosomal RNA gene database project: improved data processing and web-based tools’, *Nucleic acids research*. Oxford University Press, 41(D1), pp. D590–D596.

Salipante, S. J. *et al.* (2013) ‘Rapid 16S rRNA next-generation sequencing of polymicrobial clinical samples for diagnosis of complex bacterial infections’, *PloS one*. Public Library of Science, 8(5), pp. e65226–e65226. doi: 10.1371/journal.pone.0065226.

Salter, S. J. *et al.* (2014) ‘Reagent and laboratory contamination can critically impact sequence-based microbiome analyses’, *BMC biology*. Springer, 12(1), p. 87.

Shannon, C. E. (1948) ‘A mathematical theory of communication’, *The Bell system technical journal*. Nokia Bell Labs, 27(3), pp. 379–423.

Simon, H. Y. *et al.* (2019) ‘Benchmarking metagenomics tools for taxonomic classification’, *Cell*. Elsevier, 178(4), pp. 779–794.

Weiss, S. *et al.* (2014) ‘Tracking down the sources of experimental contamination in microbiome studies’, *Genome biology*. BioMed Central, 15(12), p. 564.

Wood, D. E., Lu, J. and Langmead, B. (2019) ‘Improved metagenomic analysis with Kraken 2’, *Genome biology*. Springer, 20(1), p. 257.