No evidence for a common blood microbiome based on a population study of 9,770 healthy humans

Cedric C.S. Tan[1,2], Minghao Chia[1], Karrie K.K. Ko[1,3,4,5], Hui Chen[1], Jianjun Liu[1], Marie Loh[6,7,8], Niranjan Nagarajan[1,5]

1) Genome Institute of Singapore, A*STAR, Singapore.
2) UCL Genetics Institute, University College London, UK.
3) Department of Microbiology, Singapore General Hospital, Singapore.
4) Department of Molecular Pathology, Singapore General Hospital, Singapore.
5) Yong Loo Lin School of Medicine, National University of Singapore, Singapore.
6) Population and Global Health, Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore.
7) Department of Epidemiology and Biostatistics, Imperial College London, United Kingdom.
8) National Skin Centre, Singapore.

Correspondence: cedriccstan@gmail.com (C.C.S. Tan) or nagarajann@gis.a-star.edu.sg (N. Nagarajan)

Abstract

Human blood is conventionally considered sterile. Recent studies have challenged this, suggesting the presence of a blood microbiome in healthy humans. We present the largest investigation to date of microbes in blood, based on shotgun sequencing libraries from 9,770 healthy subjects. Leveraging the availability of data from multiple cohorts, we stringently filtered for laboratory contaminants to identify 117 microbial species detected in the blood of sampled individuals, some of which had signatures of DNA replication. These primarily comprise of commensals associated with human body sites such as the gut (*n*=40), mouth (*n*=32), and genitourinary tract (*n*=18), which are species that are distinct from common pathogens detected in clinical blood cultures based on more than a decade of records from a tertiary hospital. Contrary to the expectations of a shared blood microbiome, no species were detected in 84% of individuals, while only a median of one microbial species per individual was detected in the remaining 16%. Futhermore, microbes of the same species were detected in <5% of individuals, no co-occurrence patterns similar to microbiomes in other body sites was observed, and no associations between host phenotypes (e.g. demographics and blood parameters) and microbial species could be established. Overall, these results do not support the hypothesis of a consistent core microbiome endogenous to human blood. Rather, our findings support the transient and sporadic translocation of commensal microbes from other body sites into the bloodstream.

Introduction

In recent years, there has been considerable interest regarding the existence of a microbiome in the blood of healthy individuals, and its links to health and disease. Human blood is traditionally considered a sterile environment (i.e., devoid of viable microbes), where the occasional entry and proliferation of pathogens in blood can trigger a dysregulated host response, resulting in severe clinical sequelae such as sepsis, septic shock or death[1]. Asymptomatic transient bacteraemia (i.e., bacterial presence in blood) in blood donors is also known to be a major cause of transfusion-related sepsis[2]. Recent studies have suggested the presence of a blood microbiome, providing evidence for microbes circulating in human blood for healthy individuals[3–7] (reviewed in Castillo *et al*[8]). However, most of these studies were either done in relatively small cohorts or lacked rigorous checks to distinguish true biological measurements from different sources of contamination[8]. In this work, we analysed blood DNA sequencing data from a population study of healthy individuals, comprising of multiple cohorts processed by different laboratories with varied sequencing kits. By leveraging the large dataset (*n*=9,770) complete with batch information in our systematic differential analyses for potential contaminants, our aim was to determine whether a blood microbiome truly exists in the general population.

For meaningful discourse, it is useful to formalise what the presence of a hypothetical 'blood microbiome' entails. Berg et al.[9] concluded that the term microbiome should refer to a community of microbes that interact with each other and with the environment in their ecological niche, which in our context is human blood. Therefore in a blood microbiome, the presence of microbial cells in blood from healthy individuals should exhibit community structures indicated by co-occurrence or mutual exclusion of species[10] as seen in the microbiomes of other sites such as the gut[11] or mouth[12]. Furthermore, we may expect the presence of core microbial species, which can be defined as species that are frequently observed and shared across individuals[13,14], such as *Staphylococcus epidermidis* on human skin[15]. More precisely, taxa that are found in a substantial fraction of samples from distinct individuals (i.e. with high prevalence) may be considered 'core'. Notably, the prevalence threshold for defining core taxa is

arbitrary, with previous microbiome studies using values ranging from 30-100% and many of these studies opting for 100%[14]. Regardless, identifying core microbes in blood would form the basis for associating microbiome changes with human health.

Existing studies have provided evidence for the presence of microbes in the blood of healthy individuals using both culture-based[3,4] and culture-independent[5–7] approaches. The former approach involves blood culture experiments while the latter involves one or a combination of the following molecular methods: 16S ribosomal RNA (rRNA) quantitative polymerase chain reaction (qPCR), 16S rRNA amplicon sequencing, and shotgun sequencing of RNA or DNA. Depending on the study design, these results should be interpreted with caution due to several methodological and technical limitations which include small sample sizes, limited taxonomic resolution, difficulties in distinguishing cell-free microbial DNA from live microbial cells, and the ubiquity of environmental contamination[8,16–19]. In particular, contaminating DNA must be accounted for in order to characterize the blood microbiome. The workflow of sample processing, from skin puncture during phlebotomy, to microbial detection, is rife with opportunities for microbes or microbial nucleic acids to be introduced. Contaminating microbial cells introduced due to poor aseptic technique or insufficient disinfection of the skin puncture site[20] affects both culture-dependent and culture-independent approaches. Sequencing-based approaches are especially sensitive to contaminant microbial DNA native to laboratory reagent kits (i.e., the 'kitome')[19], exacerbated by the low microbial biomasses in blood, accompanied by high host background which increases the noise-to-signal ratio[17]. Correspondingly, comprehensive profiling of the breadth and prevalence of microbial species in blood after accounting for external sources of contamination has not yet been done and several aspects of the 'blood microbiome' remain unclear. For instance, are the detected microbes endogenous to blood or translocated from other body sites? Is there a core set of microbes that circulates in human blood? Is there a microbial community whose structure and function could influence host health?

To address these questions, we performed the largest scale analysis of a blood sequencing dataset to date, based on DNA libraries for 9,770 healthy individuals from six distinct cohorts (**Supplementary Table 1**). We applied various bioinformatic techniques to differentiate DNA signatures of microbes in blood from potential reagent

contaminants and sequence analysis artefacts, leveraging the differences in reagent kits used to process each cohort. We detected 117 microbial species in the blood of these healthy individuals, most of which are commensals associated with the microbiomes of other body sites. Additionally, we identified DNA signatures of replicating bacteria in blood using coverage-based peak-to-trough ratio analyses[21,22], providing a culture-independent survey that has not been achieved previously. Despite this, we found no evidence for microbial co-occurrence relationships, core species, or associations with host phenotypes. These findings challenge the paradigm of a 'blood microbiome' and instead support a model whereby microbes from other body sites (e.g. gut, oral) sporadically translocate into the bloodstream of healthy individuals, albeit more commonly than previously assumed. Overall, our observations serve to establish a much needed baseline for the use of clinical metagenomics in investigating bloodstream infections.

Results

Robust inference of microbial DNA signatures in blood based on multi-cohort analysis

Blood samples from healthy individuals typically contain low microbial biomass accompanied by high host DNA background[17], making it difficult to discriminate between biologically relevant signals from artefactual ones. We first addressed artefacts arising during bioinformatic sequence analysis by performing stringent quality control on samples (**Figure 1a**), comprising of read quality trimming and filtering, removal of low complexity sequences that are of ambiguous taxonomic origin, exclusion of reads that likely originate from human DNA (**Methods**), and removal of samples with low number of reads (<100 read pairs) of microbial origin after taxonomic classification with *Kraken2*[23]. This provided a species-level characterisation of microbial DNA signatures in blood for most (*n*=8,892) samples. To minimise noise due to false positive taxonomic assigments, we applied an abundance-cutoff based filter to discriminate between species that are likely present from those that could be misclassification artefacts (**Methods**). Additionally, we validated the reliability of the microbial species detected via *Kraken2* by comparison to read alignment analysis using reference genomes, where

recovery of large fractions of a microbial reference covered uniformly by mapped reads improves our confidence that they are true positives as opposed to sequencing or analysis artefacts[24,25]. We validated 93% of the species detected using this mapping-based approach. We further observed an excellent linear relationship between the number of *Kraken2*-assigned read pairs and the number of aligned read pairs on the log10 scale (slope=1.05; F=228, $d.f.$=118, p<0.001; **Supplementary Figure 1**), suggesting that *Kraken2* taxonomic assignments are a reliable proxy for the more precise and stringent read alignment approach. These findings collectively provide confidence that the microbial species detected in our blood sequencing libraries are not likely sequence analysis artefacts.

To address artefactual signals arising due to reagent and handling contamination during sample processing, we used a series of stringent decontamination filters (**Figure 1a**). These filters are based on the idea that contamination artefacts will lead to false positive detections that are often correlated with each other and biased towards specific cohorts[26], and such analysis was found to be highly effective for *in silico* decontamination in previous studies[27–29] (**Methods**). Additionally, the identification of batch-specific contaminants in this study was greatly aided by the availability of multiple large cohorts of healthy individuals (**Supplementary Table 1**), and corresponding rich batch information, including reagent kit types and lot numbers. Application of reagent and handling contamination filters resulted in a final list of 117 microbial species that were detected in the whole blood samples of 8,892 individuals (**Supplementary Table 2**). The list of 117 confidently detected microbial species spanned 56 genera, and comprised of 110 bacteria, 5 viruses and 2 fungi.

To estimate the effectiveness of our filtering strategy in improving biological signal while reducing contamination noise, we examined the types of microbial species detected in our dataset before (870 species) and after (117 species) all filters were applied (**Figure 1b-d**). Firstly, the microbial species were cross-referenced against a published list of common genera seen as contaminants in sequencing data as curated by Poore *et al*[30] and derived from the list published by Salter *et al*[19]. In this list, genera were either classified as likely contaminants, mixed-evidence (i.e., both a pathogen and common contaminant), or potential pathogens/commensals. Following decontamination, the

proportion of detected species that are classified as contaminants decreased from 21% to 10% (**Figure 1b**). Next, the microbial species were compared against human blood culture records spanning more than a decade (2011-2021) from a tertiary hospital (**Figure 1c**). These blood cultures were typically ordered if clinical indications of bacteraemia were present, and therefore represent the range of microbial species that are known to cause symptomatic infection as detected in a clinical setting. The proportion of species that have been cultured from blood increased from 12% to 27% after decontamination, suggesting that our filtering procedures enriched for microbial species which are capable of invading the bloodstream. Finally, we compared the proportion of human-associated microbes before and after decontamination using a host-pathogen association database describing the host range of pathogens[31] (**Figure 1d**). For species that were not found in this database, a systematic PubMed search (**Methods**) was performed to determine if there was at least one past report of human infection. The proportion of human-associated species increased from 40% to 78% after decontamination, indicating that they are more likely to be biologically relevant. These results collectively suggest that by using a set of contaminant-identification heuristics, our filters effectively retain a higher proportion of biologically relevant taxa while removing likely contaminants.

Blood microbial signatures from healthy individuals reflect sporadic translocation of commensals

We next determined the fraction of distinct, healthy individuals for which microbes could be detected (i.e., prevalence). Notably, the most prevalent microbial species, *C. acnes*, was observed in 4.7% of individuals (**Figure 2a**), suggesting that none of the 117 microbes can be considered ‘core’ species that are consistently detected across most healthy individuals. Additionally, we did not detect any microbial species in most (82%) of the samples after decontamination (**Figure 2b**), whereas the remaining 18% of samples had a median of only one microbial species per sample. This low number of species detected per sample was not due to insufficient sequencing depth since there was a weak negative correlation between the number of confidently detected species per sample and the microbial read depth (Spearman’s ρ=-0.232, p<0.001). Furthermore, some samples containing no microbial species had a microbial read count of up to ~2.1

million (median=6,187 reads; distribution shown in **Supplementary Figure 2**). That is, even though a considerable number of reads were classified as microbial, they were all assigned to contaminant species. These results suggest that the presence of microbes in the blood of healthy and apparently asymptomatic individuals, as estimated by our detection methods, is infrequent and sporadic.

Given past reports of bacterial translocation from the mouth[32] or gut[33] into blood, we asked if the microbes we detected could have originated from various body sites. To do so, we assigned potential body site origins to the 117 microbial species detected in blood based on microbe-to-body-site mappings extracted from the Disbiome database[34]. We found that many (*n*=59; 50%) of these confidently detected species are indeed human commensals that are present at various human body sites (**Figure 2c**). This, together with their low prevalence, suggests that the microbial DNA of these species may have transiently translocated from other locations in the body rather than being endogenous to blood. We further categorised the microbial species based on their growth environments (**Figure 2d**). A significant portion (*n*=42; 36%) of the species were obligate anaerobes or obligate intracellular microbes, atypical of skin-associated microbes that may be introduced during phlebotomy[2], indicating that they are not likely to be sampling artefacts. All in all, the diverse origins of the microbes detected in blood, together with their low prevalence across a healthy population, is consistent with sporadic translocation of commensals into the bloodstream.

Microbial presence in blood (i.e., bacteraemia) is typically associated with a range of clinical sequelae from mild fevers to sepsis. As such, we asked if the common microbes identified in patients with disease-associated bacteraemia are different from those detected in our cohorts of healthy individuals. To do so, we compared the prevalence of microbes detected in the sequenced blood samples against observations from 11 years of hospital blood culture records. The prevalence of microbial genera detected in the hospital blood culture records clearly differed from that in our sequenced blood samples, despite the overlap in detected taxa (**Figure 2e**). For example, while *Staphylococcus*, *Escherichia* and *Klebisiella* were the predominant genera identified in blood cultures, they were rarely detected in our blood sequencing libraries. These findings may be explained by the potentially higher virulence of pathogens detected in

the clinic, which are more likely to cause clinical symptoms in individuals that would result in exclusion during our recruitment process. Conversely, our findings suggest that the microbes detected in the blood of healthy individuals are potentially better tolerated by the immune system (e.g. *Bifidobacterium* spp.[35] and *Faecalibacterium prausnitizii*[36] with immunomodulatory properties as gut commensals; **Figure 2a**).

Evidence for replicating microbial cells but without community structure or host associations

To better characterise the microbial DNA signatures detected in blood, we asked if they reflect the presence of viable microbial cells as opposed to circulating cell-free DNA. This is because the former would allow for complex microbe-microbe or microbe-host interactions that would be of greater and more direct clinical relevance. In contrast to previous approaches that used microbial cultures[3,37], we looked for more broad-based evidence of live bacterial growth in by applying replication rate analyses[21,22] on our sequenced blood samples. This approach is based on the principle that DNA sequencing of replicating bacteria would yield an increased read coverage (i.e., peak) nearer to the origin of replication (*Ori*) and decreased coverage (i.e., trough) nearer to the terminus (*Ter*)[22]. A coverage peak-to-trough ratio (PTR) greater than one is indicative of bacterial replication. Through this analysis, we found evidence for replication of 11 bacterial species out of the 20 that were sufficiently abundant to do this analysis (**Figure 3a)**. The median-smoothed coverage plots of the replicating species all exhibited the sinusoidal coverage pattern (in black; **Figure 3b**) characteristic of replicating bacterial cells[22]. This contrasts with the even coverage patterns of three representative contaminants identified during the decontamination steps: *Achromobacter xylosoxidans*, *Pseudomonas mendocina* and *Alcaligenes faecalis* (**Figure 3c**). The *Ori* and *Ter* positions determined using coverage biases largely corresponded with an orthogonal method based on the GC-skew[38] of bacterial genomes, suggesting that the replication rate analyses are reliable. Additionally, all but one of these replicating species are present in hospital blood culture records and in previous reports of bacteraemia[39–48] (**Figure 3a**), indicating their ability to replicate in human blood. Overall, beyond the detection of microbial DNA, we present the first culture-independent evidence for replicating bacterial cells in blood.

Given the presence of live bacteria, we investigated if the microbial species detected showed patterns of microbe-microbe interactions as would be expected from a microbial community. To do so, we computed pairwise *SparCC* correlations[49] between species, where positive and negative values indicate co-occurrence and mutual-exclusion, respectively. *SparCC* correlation is a reliable metric for assessing co-occurrence since it accounts for the sparse and compositional nature of microbial taxonomic profiles that confound standard correlation inference techniques[49]. We visualised *SparCC* correlations of the 117 microbial species confidently detected in blood sequencing libraries using network graphs, where each node is a species and each edge represents the co-occurrence/exclusion associations between two species (**Figure 4a**). We could not detect strong community co-occurrence/exclusion patterns, with most associations being weak (SparCC correlation<0.05), and only 19 pairwise associations exceeding a correlation value of 0.2, with four exceeding a value of 0.3 (**Figure 4a**). To determine if this result is a function of our stringent decontamination filters, we generated independent network graphs for the five adult cohorts before decontamination filtering and examined the co-occurrence/exclusion associations shared across cohorts. With an already lenient SparCC correlation threshold of 0.2, we identified no associations common to all the network graphs (**Figure 4b**), indicating that there were no consistent detectable microbial community associations in blood typical of microbiomes in various human body sites.

Previous studies have demonstrated the use of blood microbial DNA as a biomarker for disease, demonstrating associations with cancer[30], type II diabetes[50] and periodontal disease[51]. In a similar vein, we investigated if the presence of microbes was associated with host phenotypes in our dataset. We first examined if microbes were detected more frequently in infants relative to adults. Given that the still-developing immune systems of infants puts them at greater risk of infection relative to healthy adults[52], we reasoned that the prevalence of microbes in blood may differ within a birth cohort (GUSTO) relative to adult cohorts. Indeed, samples from GUSTO appeared to have a higher prevalence of microbes associated with most human body sites (**Supplementary Figure 3a**). This was in part, driven by genitourinary tract-associated microbes, *Fannyhessea vaginae*, *Lactobacillus jensenii, Lactobacillus crispatus*, *Lactobacillus*

iners, and *Gardnerella vaginalis* (**Supplementary Figure 3b**). Similarly, we found enrichment of gut-associated bacteria such as *Bifidobacterium* spp. in GUSTO (**Supplementary Figure 3c**). These findings suggest that bacterial translocation may be more frequent in infants relative to adults, though differences in sample collection (umbilical cord *versus* venipuncture) could also partially explain them.

Next, we systematically tested for pairwise associations between eight host phenotypes that were documented on the day of blood collection and the presence of each of the 117 microbial species detected in blood. These host phenotypes attributes were: sex, ancestry, age, body mass index (BMI), blood total cholesterol (TC), blood triglycerides (TG), systolic and diastolic blood pressure (SBP and DBP). Given the multiple large independent cohorts, we could perform statistical tests on each cohort separately, which allowed us to assess the consistency of identifed association patterns across the different cohorts. Since these cohorts were sampled from a homogenous population, true association patterns are expected to be detected repeatedly regardless of cohort. Using this statistical testing approach, we found only five significant microbe-phenotype associations ($p<0.05$; **Supplementary Table 3**) after adjusting for multiple comparisons. Notably, all but one of the significant associations were present in only one cohort. The exception was *C. acnes*, which was significantly associated with ancestry in two cohorts. However, while *C. acnes* was more prevalent in individuals of Malay ancestry within the SEED cohort, it was more prevalent in Chinese individuals within the MEC cohort (**Supplementary Figure 4**). These cohort specific differences could be due to other demographic variables that were not recorded in this study, or perhaps from *C. acnes* subspecies differences. To ensure that we did not miss any associations due to the possible non-linearity of host-phenotype and microbial relationships, we also derived categorical phenotypes based on the recorded phenotypic information. These include being elderly (age>=65), and other measures of ‘poorer health’, such as being obese (BMI>30), having high blood triglycerides (TG>2.3 mmol/L), high total cholesterol (TC>=6.3 mmol/L), or high blood pressure (SBP>=130 and DBP>=80). We then tested for pairwise associations between these derived phenotypes and the presence of *any* bacteria but found no significant associations ($p>0.05$; **Supplementary Table 4**). Collectively, these results suggest no consistent associations between the presence of

microbes in blood and the host phenotypes tested within a healthy population of individuals.

Discussion

We present the largest scale analysis, to date, of microbial signatures in human blood with rigorous accounting for computational and contamination artefacts and found no evidence for a common blood microbiome in a healthy population. Instead, we observed mostly sporadic instances of blood harbouring DNA from single microbial species of diverse bodily origins, some of which might be actively replicating. Our findings hint at the possibility that the bloodstream represents a route for movement of microbes between different body sites in healthy individuals. However, the low prevalence of the detected species suggest that this movement is likely to be infrequent and transient. Unresolved questions remain about how interconnected the microbiomes at various body sites are, and whether these processes are altered during disease or throughout a person's lifetime. Can perturbations to the microbial community at one body site affect that at another site, and how does the host immune system asymptomatically regulate microbial presence in blood? Our study lays the groundwork for future investigations into these questions, which may pave the way for a systemic understanding of the human microbiome across body sites in relation to human health and disease.

We found no core species in human blood on the basis of low prevalence across individuals in our population-level dataset. The prevalence estimates provided in this study are contingent on the sensitivity of detecting microbes through sequencing. Previous studies have shown that untargeted shotgun sequencing is highly sensitive for the detection of microbes in blood at a total sequencing depth of 20-30 million reads per sample[53–55], perhaps even more so than culture-based methods[56,57]. In contrast, a median of 373 million reads was generated per sample for our sequencing libraries, suggesting that our methods do not lack sensitivity. Our prevalence estimates are also affected by the abundance thresholds used to determine whether a species is present in a single sample (i.e., abundance filter; **Figure 1a**). We defined these thresholds in terms of both absolute read count and relative abundance, which were determined based on simulation experiments (see **Methods**). Overly stringent abundance thresholds would

lead to the erroneous masking of genuine signals, leading to an underestimation of microbial prevalence. However, even when relaxing the threshold to just a relative abundance of 0.001, none of the species, whether flagged as a contaminant or not, had more than 52% prevalence (**Supplementary Table 5**). Furthermore, the 20 most prevalent species at this threshold are all environmental microbes, and mostly comprise of *Sphingomonas* and *Bradyrhizobium* species, which are known to be common sequencing-associated contaminants[19]. This suggests that independent of our decontamination filters, none of the species detected qualify as core members.

In addition to not being able to detect any core species, we could not detect any strong co-occurrence or mutual exclusion associations between species regardless of whether our decontamination filters were applied. These associations generally reflect cooperation or competition between species, respectively[58]. Indeed, within a microbial community, metabolic dependencies of species and the ability of different species to complement these dependencies have been shown to be a key driver of microbial co-occurrence[59]. On the other hand, competitive behaviours such as nutrient sequestration to deprive potential competitors of nutrients or producing adhesins to bind and occupy favourable sites in an environment[60] can lead to mutual exclusion between species. The fact that we could not detect any strong associations therefore points to the absence of an interacting microbial community in healthy humans. Of note, since our dataset was derived from circulating venous blood, we are, in principle, not able to detect microbial interactions that may be occurring at other sites of the bloodstream such as the inner endothelial lining of blood vessels. Experiments investigating the adherence of bacteria to blood vessel linings may provide further insight into this.

The availability of 11 years of blood culture records from the same country of origin as our blood samples enabled a reliable comparison of the prevalence of microbes in the healthy population and in the clinic. This is because the frequency of infections caused by different microbial species is known to differ from country to country[61]. Despite this, we expect that some of the variation in prevalence estimates may be due to the differences in detection methods. That said, previous studies have shown a strong concordance between culture and sequencing-based detection[53,54,56,57], suggesting that the distinction between the prevalence of microbes found in healthy individuals and in

the clinic is not due to the differences in detection methods. Our results support the conclusion that microbial presence in blood (i.e., bacteraemia) does not always lead to disease. These results are consistent with our other observation that microbes detected in our cohorts of asymptomatic individuals tend to be commensals, which may inherently be less virulent and better tolerated by the host compared to disease-causing pathogens. Indeed, the long-standing co-evolution of humans and colonizing microbes, places a selective pressure against high virulence phenotypes in these microbes to maintain host viability[62]. Simultaneously, there is a selective pressure for immunomodulatory phenotypes in commensals to improve their fitness, evidenced by the wealth of immunomodulatory activities found in the gut microbiome[63]. This agrees with previous findings that colonisation by commensals modulate early development of the immune system[64], which would allow a measured and regulated response against translocated commensals. By extension, the immunomodulatory properties of bacteria and their links to host tolerance to bacteraemia may be key factors in determining clinical outcomes. Perhaps, the presence (or lack) of these properties may determine whether an individual with bacteraemia is asymptomatic or septic. For example, abundant gut bacterial species such as *Bacteroides* spp. were not commonly detected in blood. Further exploration into the immunomodulatory activities of commensals vis-à-vis common blood culture pathogens may be the key to design therapeutics to manage or prevent the dysregulated host response that defines sepsis[1].

We found no convincing associations between both measured (e.g. TC, SBP) and derived (e.g. obesity) host phenotypes with microbial presence. This suggests that the risk of transient microbial translocation, at least across our cohorts of healthy adults, is fairly consistent. In contrast, this risk may increase in individuals with more severe disease. In fact, variable microbial DNA profiles in blood have been used to delineate health and disease states. This has most prominently been shown for sepsis[53–57,65], where the presence of viable microbes is expected, but also for cancer[30], periodontal disease[51], and chronic kidney disease[66], which are unrelated to bloodstream infections. These studies highlight the promise of metagenomic sequencing of blood for developing diagnostic, prognostic, or therapeutic tools. Our characterisation of the species breadth in healthy individuals forms a crucial baseline for comparison with that in diseased

individuals. Indeed, our findings open new doors to understanding why and how blood microbial profiles correlate with health status. One possible hypothesis is that mucosal integrity is compromised in a disease state, leading to higher translocation rates of microbes into the bloodstream. This is consistent with findings of increased intestinal permeability (i.e., 'leaky gut') in disease or even during physiological stress[67]. Future studies testing this hypothesis may consider a focus on the gut-associated bacteria that were detected in our study (e.g. *Bifidobacterium adolescentis, Faecalibacterium prausnitzii*). Further experimental investigations into the mechanisms of microbial translocation and the modulatory effects of the microbiomes present at other body sites may shed light on the relationship between microbial presence in blood and health status.

If we take the definition of a 'microbiome' as a microbial community whose member species interact amongst themselves and with their ecological niche[9], our findings lead to the conclusion that there is no consistent circulating blood microbiome. Sporadic and transient translocation of commensals from other body sites into the bloodstream (**Figure 5**) is the more parsimonious explanation for the observation that most of the microbes detected are commensals from other body sites. Furthermore, the relatively low prevalence of microbes in blood suggests rapid clearance of translocated microbes rather than prolonged colonisation in blood. Based on these findings, we advocate against the use of the term 'blood microbiome' or 'circulating microbiome', which are potentially misleading, when referring to the detection of microbial DNA or of microbial cells in blood due to transient translocation events.

Methods

Datasets

Our sequencing dataset, also known as the SG10K_Health dataset (https://www.npm.sg/collaborate/partners/sg10k/), comprises of shotgun sequencing libraries of DNA extracted from the whole blood or umbilical cord blood of 9,770 healthy Singaporean individuals[68] who were recruited as part of six independent cohorts. Whole blood for sequencing was collected via venipuncture from the five adult cohorts (median age=49; interquartile range=16): Health for Life in Singapore (HELIOS; *n*=2,286), SingHealth Duke-NUS Institute of Precision Medicine (PRISM, *n*=1,257), Tan Tock Seng Hospital Personalised Medicine Normal Controls (TTSH, *n*=920), Singapore Epidemiology of Eye Diseases (SEED, *n*=1,436)[69,70], and the Multi-Ethnic Cohort (MEC, *n*=2,902)[71]. Additionally, cord blood was collected for the birth cohort Growing Up in Singapore Towards healthy Outcomes (GUSTO; *n*=969)[72]. Measurement of host phenotypes was performed on the day of blood collection, except for the GUSTO cohort where measurements were taken at a later timepoint when the children were at a median age of 6.1 (interquartile range=0.1). Using nearest neighbor approaches to reference genotypes[73], individuals were broadly categorised into four ethnic categories representing distinct genetic ancestries: Chinese (59%), Malays (19%), Indians (21%) and Others (1%). All individuals were deemed healthy at the point of recruitment if they did not include any self-reported diseases in the recruitment questionnaires. All cohort studies were approved by relevant institutional ethics review boards. A summary of the cohort demographics and the ethics review approval reference numbers are provided in **Supplementary Table 1**.

Additionally, we retrieved anonymised blood culture records from Singapore General Hospital, the largest tertiary hospital in Singapore. These records span the years 2011-2021 and include aerobic, anaerobic and fungal blood cultures taken from 282,576 unique patients. These blood cultures were ordered as part of routine clinical management, that is, when clinically indicated for the investigation of bacteremia or fungemia. Blood cultures were performed and analysed as per hospital standard operating procedures. In brief, blood samples were collected aseptically and inoculated

into BD™ BACTEC™ bottles at the bedside (BD™ BACTEC™ Plus Aerobic/F Culture vials Plastic [catalogue number 442023] for aerobic blood culture, BD™ BACTEC™ Plus Anaerobic/F Culture vials Plastic [catalogue number 442022] for anaerobic blood culture and Myco/F Lytic [catalogue number 42288] for fungal blood culture). The inoculated bottles were transported to the diagnostic laboratory at ambient temperature and incubated in the BD™ BACTEC™ FX Blood Culture System on arrival. Aerobic and anaerobic blood culture bottles were incubated for a maximum of five days, and fungal blood culture bottles were incubated for a maximum of 28 days. Blood culture bottles that were flagged positive by the BD™ BACTEC™ FX Blood Culture System were inoculated onto solid media, and the resultant colonies were identified using a combination of biochemical tests and matrix assisted laser desorption ionization-time of flight mass spectrometry (MALDI-TOF MS) (Bruker® microflex LRF).

Sample preparation and batch metadata

DNA from whole blood was extracted using one of six different DNA extraction kits. Paired-end 151bp sequencing with an insert size of 350bp was performed up to 15-fold or 30-fold coverage of the human genome. Library preparation was performed using one of three library preparation kits. Sequencing was performed on the Illumina HiSeq X platform with HiSeq PE Cluster Kits and HiSeq SBS Kits. The type of extraction kits and library preparation kits used, and lot numbers for the SBS Kits, PE Cluster Kits, and sequencing flow cells used are provided as batch metadata. All reagent kits used, the number of batches and the number of samples processed per batch are provided in **Supplementary Table 6**.

Data pre-processing and quality control

The bioinformatic processing steps applied to the sequencing libraries are summarised in **Figure 1a**. Read alignment of sequencing reads to the GRCh38 human reference genome was already performed as part of a separate study[68] using *BWA-MEM v0.7.17*[74]. We retrieved read pairs where both members of the pair did not map to the human genome. Following which, we performed quality control of the sequencing reads. We trimmed low quality bases at the ends of reads with quality <Q10 (base quality trimming) and discarded reads with average read quality less than Q10 (read quality

filter). We also discarded low complexity sequences with an average entropy less than 0.6, with a sliding window of 50 and k-mer length of five (low complexity read filter). All basic quality control steps were performed using *bbduk* from the *BBTools suite* v37.62 (sourceforge.net/projects/bbmap/).

Taxonomic classification of blood sequencing libraries

Taxonomic classification of non-human reads was done using *Kraken2* v2.1.2[23] with the '—paired' flag. We used the *PlusPF* database (17th May 2021 release) maintained by Ben Langmead (https://genome-idx.s3.amazonaws.com/kraken/k2_pluspf_20210517.tar.gz), which includes archaeal, bacterial, viral, protozoan, and fungal references. Of all non-human read pairs, 72% were classified as microbial at the species level, yielding 8,890 species. Samples with less than 100 microbial read pairs were removed, resulting in a final dataset comprising 8,892 samples, with a median microbial read-pair count of 6187.

To minimise noise in the taxonomic assignments, we defined a set of abundance thresholds whereby species with abundance values less than or equal to these thresholds (i.e., relative abundance≤0.05, read pairs assigned≤10) were counted as absent (set to zero read counts). We performed simulations to systematically determine a relative abundance threshold that minimizes false positive species assignments. Sequencing reads were simulated using *InSilicoSeq v1.5.4*[75] with error models trained on the SG10K_Health sequencing libraries and processed using the same bioinformatic steps as per the SG10K_Health dataset to obtain microbial taxonomic profiles. We simulated 373 million reads equivalent to the median library read count of all samples, comprising reads from the GRCh38 human reference and ten microbial genomes (*Yersinia enterocolitica*, *Leclercia adecarboxylata*, *Moraxella osloensis*, *Streptococcus pneumoniae*, *Pasteurella multocida*, *Staphylococcus epidermidis*, *Actinomyces viscosus*, *Torque teno virus*, *Human betaherpesvirus 6A*, *Candida albicans*) at various proportions. Due to read misclassification, some of the simulated reads were erroneously assigned to another species and produced false positives. A final relative abundance threshold of 0.005 that delineated these false positive assignments from true positives was selected (**Supplementary Figure 5**). Relative abundances were

calculated by dividing the microbial read count in a sample by the total number of microbial reads assigned to that sample.

Decontamination filters

After application of the presence/absence filter, we identified and removed putative contaminants using established decontamination heuristics[26] that have been validated in previous studies[27,28], prior to our downstream analyses. These rules were applied using eight types of batch information: source cohort, DNA extraction kit type, library preparation kit type, and lot numbers for sequencing-by-synthesis kit (box 1, box 2), paired-end cluster kit (box 1, box 2) and sequencing flow cell used. Other batch information such as the pipettes and consumables used, or storage location and duration were not recorded and could potentially contribute to some level of batch-specific contamination. However, these batches are expected to be correlated with the other types of batch information available, and so the resultant contaminants could in theory be accounted for using our filters. We describe the four decontamination filters used, as shown in **Figure 1a**, in sequential order:

(1) *Prevalence filter*. A microbial species is considered a contaminant specific to a batch if it is present at greater than 25% prevalence in that batch and has greater than a two-fold higher prevalence than that for any other batch. Batches with less than 100 samples were excluded from this analysis. This filter is based on the principle that species which are highly prevalent in some batches but lowly prevalent or absent in others are likely contaminants[26]. We illustrate this for an example species in **Supplementary Figure 6a**.

(2) *Correlation filter*. A microbial species is considered a contaminant if it is highly correlated (Spearman's ρ>0.7) with any contaminant within the same batch, as identified by the prevalence filter. This filter is based on the principle that contaminants are highly correlated within the same batch[26]. Spearman's ρ was calculated using centred log-ratio (CLR) transformed[76] microbial relative abundances. CLR transformations and Spearman's ρ were calculated using the *clr* function as part of the *compositions* package[77] and *cor.test* function in *R*. We

illustrate this within-batch correlation for an example species in **Supplementary Figure 6b**.

(3) *Batch filter*. A non-contaminant microbial species must be detected in samples processed by at least two reagent kit batches or reagent types. That is, any species that is only detected in a single batch for any of the reagent kits used (**Supplementary Table 6**) are considered contaminants. This filter is based on the principle that species that can be repeatedly observed across different reagent batches are more likely to reflect genuine non-contaminant signals[26]. Library preparation kit type was excluded from this analysis since only three kit types were used, with 86% of samples processed using one of the kits.

(4) *Read count filter*. A microbial species is considered a sequencing or analysis artefact if it is not assigned at least 100 reads in at least one sample. This filter is based on the principle that species that are always assigned a low number of read pairs, never exceeding the background noise within sequencing libraries, are more likely to be artefactual rather than genuine signals. An example of an artefactual species is *Candidatus Nitrosocosmicus franklandus*, which was assigned at most 22 read pairs by *Kraken2* across 21 sequenced samples.

Characterisation of microbial species

We classified microbial species as human-associated or not based on a published host-pathogen association database[78]. In this database, host-pathogen associations are defined by the presence of at least one documented infection of the host by the pathogen[31]. For species that were not found in this database, we performed a systematic PubMed search using the search terms: (microbial species name) AND (human) AND ((infection) OR (commensal)). Similarly, species that had at least one published report of human colonisation/infection were considered human-associated. Additionally, we classified the potential body site origins for each microbial species using the *Disbiome* database, which collects data and metadata of published microbiome studies in a standardised way[34]. We extracted the information for all microbiome experiments in the database using the URL: ‘https://disbiome.ugent.be:8080/experiment’ (accessed 26th April 2022). We first

extracted microbe-to-sample type mappings from this information (e.g. *C. acnes*→skin swab). We then manually classified each sample type into different body sites (e.g. skin swab→skin). This allowed us to generate microbe-to-body site mappings. Sample types with ambiguous body site origins (e.g. abscess pus) were excluded. The range of sample types within the Disbiome database used to derive the microbe-body-site mappings are provided in **Supplementary Table 7**. Finally, we classified microbial species based on their growth requirements, with reference to a clinical microbiology textbook[79]. Viruses were classified as obligate intracellular. The microbiological classifications for each species are provided in **Supplementary Table 2.**

Estimating coverage breadth and bacterial replication rates

We performed read alignment of sequencing libraries to microbial reference genomes using *Bowtie v2.4.5*[80] with default parameters. In total we used references for 27 of the 117 microbial species detected in blood, comprising the 20 bacterial species with the highest number of reads in a sample, all fungal (*n*=2) and all viral species (*n*=5). For each species, we aligned the microbial reads of five sample libraries with the most reads assigned to that species, to the reference genome of that species. For each sample and microbial genome, the genome coverage per position was computed using the *pileup* function as part of the *Rsamtools v2.8.0* package[81] in *R*. In principle, recovery of a larger fraction of a microbial genome provides a higher confidence that it is truly present in a sample[24,25]. We could recover at least 10% of the microbial genomes for 25/27 (93%) of the species. For the replication rate analyses, PTR values were calculated using the *bPTR* function in *iRep v1.1.0*[21], which is based on the method proposed by Korem et al.[22]. The *Ori* and *Ter* positions were determined based on the coverage peaks and troughs (in red and blue, respectively; **Figure 3**). *Ori* and *Ter* positions were also calculated using a cumulative GC-skew line, which is expected to be in anti-phase with the sinusoidal coverage pattern across the genome[38] (in green; **Figure 3**).

Microbial networks

Microbial co-occurrence/mutual exclusion associations were computed using the *SparCC* algorithm[49], implemented in the *SpiecEasi v1.1.2* package[82] in *R* and the

microbial networks were visualized using *Igraph v1.2.9*[83]. We excluded the birth cohort GUSTO since it is of a different demographic that may possess a distinct set of microbial associations.

Detecting associations between microbial taxonomic profiles and host phenotypes

We tested for microbe-host phenotype associations within individual cohorts separately. For the two categorical host phenotypes, genetic sex and ancestry, we tested for differences in the prevalence of each microbial species between the different categories using a two-sided Fisher's exact test (*fisher.test* function in *R*). For the continuous variables (age, BMI, TC, TG, SBP and DBP) we used a two-sided Mann-Whitney U test (*wilcox.test* function in *R*) to test for differences in the distributions of the variables when a species was present or absent. Benjamini-Hochberg multiple-testing correction was applied only after consolidating the *p*-values from both tests and for all cohorts using the *p.adjust* function in *R*. Statistical tests were only performed if a species was present in at least 50 samples in total. Separately, for derived phenotypes (i.e., being elderly or measures of 'poorer health'), we used the Fisher's exact test before applying Benjamini-Hochberg multiple-testing correction. In all cases, samples with missing host phenotypes were excluded.

Data analysis and visualisation

All data analyses were performed using *R* v4.1.0 or using *Python* v3.9.12. Visualisations were performed using *ggplot v3.3.5*[84]. **Figure 5** was created using BioRender.com under an academic subscription.

Data availability

Requests for the sequencing data used in this study should be made through the National

Precision Medicine (NPM) Programme Data Access Committee (contact_npco@gis.a-star.edu.sg). The accession numbers for all genome references used are provided in **Supplementary Table 8**.

Code availability

All custom code used to perform the analyses reported here are hosted on GitHub (https://github.com/cednotsed/blood_microbial_signatures.git).

Acknowledgements

The computational work for this article was partially performed on resources of the National Supercomputing Centre, Singapore (https://www.nscc.sg). This study made use of data generated as part of the Singapore National Precision Medicine program funded by a grant from the Industry Alignment Fund (Pre-Positioning) (IAF-PP: H17/01/a0/007).

This study made use of data / samples collected in the following cohorts in Singapore:

(1) The Health for Life in Singapore (HELIOS) study at the Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore (supported by grants from a Strategic Initiative at Lee Kong Chian School of Medicine, the Singapore Ministry of Health (MOH) under its Singapore Translational Research Investigator Award (NMRC/STaR/0028/2017) and the IAF-PP: H18/01/a0/016);

(2) The Growing up in Singapore Towards Healthy Outcomes (GUSTO) study, which is jointly hosted by the National University Hospital (NUH), KK Women's and Children's Hospital (KKH), the National University of Singapore (NUS) and the Singapore Institute for Clinical Sciences (SICS), Agency for Science Technology and Research (A*STAR) (supported by the Singapore National Research Foundation under its Translational and Clinical Research (TCR) Flagship Programme and administered by the Singapore Ministry of Health's National Medical Research Council (NMRC), Singapore - NMRC/TCR/004-NUS/2008; NMRC/TCR/012-NUHS/2014. Additional funding was provided by SICS and IAF-PP H17/01/a0/005);

(3) The Singapore Epidemiology of Eye Diseases (SEED) cohort at Singapore Eye Research Institute (SERI) (supported by NMRC/CIRG/1417/2015; NMRC/CIRG/1488/2018; NMRC/OFLCG/004/2018);

(4) The Multi-Ethnic Cohort (MEC) cohort (supported by NMRC grant 0838/2004; BMRC grant 03/1/27/18/216; 05/1/21/19/425; 11/1/21/19/678, Ministry of Health, Singapore, National University of Singapore and National University Health System, Singapore);

(5) The SingHealth Duke-NUS Institute of Precision Medicine (PRISM) cohort (supported by NMRC/CG/M006/2017_NHCS; NMRC/STaR/0011/2012, NMRC/STaR/ 0026/2015, Lee Foundation and Tanoto Foundation);

(6) The TTSH Personalised Medicine Normal Controls (TTSH) cohort (supported by NMRC/CG12AUG17 and CGAug16M012).

The views expressed are those of the authors, are not necessarily those of the National Precision Medicine investigators, or institutional partners. We thank all investigators, staff members and study participants who made the National Precision Medicine Project possible.

Ethics declaration

All individuals in the participating cohorts were recruited with signed informed consent from the participating individual or parent/guardian in the case of minors. All studies were approved by relevant institutional ethics review boards detailed in **Supplementary Table 1**.

Figure legends

Figure 1: Decontamination results. (a) Summary of pre-processing steps and filters applied to taxonomic profiles (*n*=9,770 individuals) and the number of species retained after each filter. Pie charts showing the proportion of microbial species that are (b) human-associated, (c) common sequencing contaminants, and (d) detected in blood culture records, before and after applying the decontamination filters.

Figure 2: Microbial signatures in human blood from healthy individuals. (a) Bar chart showing the prevalence of the top 50 confidently detected microbial species in all 8,892 blood sequencing libraries. (b) Histogram of the number of microbial species per sample. (c) Bar chart of the human body sites that the 117 confidently detected species are associated with, as determined using the Disbiome database[34]. Species are classified as 'multiple' if they are associated with more than one body site and classified otherwise if they are only associated with a single body site. (d) Piechart showing the microbiological classification of the 117 confidently detected species. (e) Bar chart showing prevalence of genera in blood culture records and in the blood sequencing libraries before and after decontamination.

Figure 3: Evidence for replicating bacteria in blood samples from healthy individuals. (a) Summary statistics for samples where bacterial species were deemed to be replicating using *iRep*[21] (i.e., peak-to-trough ratio (PTR)>1). The number of reads assigned to the species by *Kraken2* [23], the possible body sites the species are associated with, whether they were previously reported in published studies of bacteraemia, the overall prevalence of the species across all 8,892 individuals in our study and the calculated PTR values, are indicated. Coverage plots of (b) three representative confidently detected species and (c) three representative contaminant species, showing the expected patterns of *Ori* to *Ter* coverage skew only where expected i.e. confidently detected species.

Figure 4: Microbial co-occurrence networks. (a) *SparCC*[49] co-occurrence networks computed from all samples with at least two microbial species following decontamination at different *SparCC* correlation thresholds (0.05, 0.2, 0.3). Only associations with a magnitude of *SparCC* correlation greater than the respective

thresholds are retained. (b) *SparCC* networks for individual cohorts at a correlation threshold of 0.2. No co-occurrence associations were retained after taking the intersection of edges between all cohort networks. For (a) and (b), each node represents a single microbial species, and each edge a single association between a pair of microbial species. Edge thickness is scaled by the magnitude of correlation. The number of samples used to compute each network and the correlation thresholds used are annotated. Positive and negative *SparCC* correlations are indicated in green and blue respectively.

Figure 5: **Potential models for microbes in blood.** Our findings suggest that there is no consistent circulating blood microbiome (i.e., the blood microbiome model). The more likely model is where microbes from other body sites transiently and sporadically translocate into blood. Created with BioRender.com under an academic subscription.

References

1. Singer, M. *et al.* The third international consensus definitions for sepsis and septic shock (Sepsis-3). *Jama* **315**, 801–810 (2016).

2. Brecher, M. E. & Hay, S. N. Bacterial contamination of blood components. *Clin. Microbiol. Rev.* **18**, 195–204 (2005).

3. Damgaard, C. *et al.* Viable bacteria associated with red blood cells and plasma in freshly drawn blood donations. *PLoS One* **10**, e0120826 (2015).

4. Schierwagen, R. *et al.* Circulating microbiome in blood of different circulatory compartments. *Gut* **68**, 578–580 (2019).

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

6. Whittle, E., Leonard, M. O., Harrison, R., Gant, T. W. & Tonge, D. P. Multi-method characterization of the human circulating microbiome. *Front. Microbiol.* 3266 (2019).

7. D'Aquila, P. *et al.* Microbiome in Blood Samples From the General Population Recruited in the MARK-AGE Project: A Pilot Study. *Front. Microbiol.* 2055 (2021).

8. Castillo, D. J., Rifkin, R. F., Cowan, D. A. & Potgieter, M. The healthy human blood microbiome: Fact or fiction? *Front. Cell. Infect. Microbiol.* **9**, 148 (2019).

9. Berg, G. *et al.* Microbiome definition re-visited: old concepts and new challenges. *Microbiome* **8**, 1–22 (2020).

10. Faust, K. *et al.* Microbial co-occurrence relationships in the human microbiome. *PLoS Comput. Biol.* **8**, (2012).

11. Das, P., Ji, B., Kovatcheva-Datchary, P., Bäckhed, F. & Nielsen, J. In vitro co-cultures of human gut bacterial species as predicted from co-occurrence network analysis. *PLoS One* **13**, e0195161 (2018).

12. Relvas, M. *et al.* Relationship between dental and periodontal health status and

the salivary microbiome: bacterial diversity, co-occurrence networks and predictive models. *Sci. Rep.* **11**, 1–22 (2021).

13. Risely, A. Applying the core microbiome to understand host–microbe systems. *J. Anim. Ecol.* **89**, 1549–1558 (2020).

14. Neu, A. T., Allen, E. E. & Roy, K. Defining and quantifying the core microbiome: Challenges and prospects. *Proc. Natl. Acad. Sci.* **118**, e2104429118 (2021).

15. Consortium, H. M. P. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207 (2012).

16. Johnson, J. S. *et al.* Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nat. Commun.* **10**, 5029 (2019).

17. Glassing, A., Dowd, S. E., Galandiuk, S., Davis, B. & Chiodini, R. J. Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. *Gut Pathog.* **8**, 24 (2016).

18. Hornung, B. V. H., Zwittink, R. D. & Kuijper, E. J. Issues and current standards of controls in microbiome research. *FEMS Microbiol. Ecol.* **95**, fiz045 (2019).

19. Salter, S. J. *et al.* Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* **12**, 87 (2014).

20. Doern, G. V *et al.* A comprehensive update on the problem of blood culture contamination and a discussion of methods for addressing the problem. *Clin Microbiol Rev* **33**, e00009-19 (2019).

21. Brown, C. T., Olm, M. R., Thomas, B. C. & Banfield, J. F. Measurement of bacterial replication rates in microbial communities. *Nat. Biotechnol.* **34**, 1256–1263 (2016).

22. Korem, T. *et al.* Growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples. *Science (80-.).* **349**, 1101–1106 (2015).

23. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).

24. Hillmann, B. *et al.* SHOGUN: a modular, accurate and scalable framework for microbiome quantification. *Bioinformatics* **36**, 4088–4090 (2020).

25. Al-Ghalith, G. & Knights, D. BURST enables mathematically optimal short-read alignment for big data. *bioRxiv* 2020.09.08.287128 (2020) doi:10.1101/2020.09.08.287128.

26. de Goffau, M. C. *et al.* Recognizing the reagent microbiome. *Nat. Microbiol.* **3**, 851–853 (2018).

27. Chia, M. *et al.* Shared signatures and divergence in skin microbiomes of children with atopic dermatitis and their caregivers. *J. Allergy Clin. Immunol.* (2022) doi:https://doi.org/10.1016/j.jaci.2022.01.031.

28. Chng, K. R. *et al.* Cartography of opportunistic pathogens and antibiotic resistance genes in a tertiary hospital environment. *Nat. Med.* **26**, 941–951 (2020).

29. de Goffau, M. C. *et al.* Human placenta has no microbiome but can contain potential pathogens. *Nature* **572**, 329–334 (2019).

30. Poore, G. D. *et al.* Microbiome analyses of blood and tissues suggest cancer diagnostic approach. *Nature* **579**, 567–574 (2020).

31. Shaw, L. P. *et al.* The phylogenetic range of bacterial and viral pathogens of vertebrates. *Mol. Ecol.* **n/a**, (2020).

32. Tomás, I., Diz, P., Tobías, A., Scully, C. & Donos, N. Periodontal health status and bacteraemia from daily oral activities: systematic review/meta-analysis. *J. Clin. Periodontol.* **39**, 213–228 (2012).

33. Wells, C. L., Maddaus, M. A. & Simmons, R. L. Proposed mechanisms for the translocation of intestinal bacteria. *Rev. Infect. Dis.* **10**, 958–979 (1988).

34. Janssens, Y. *et al.* Disbiome database: linking the microbiome to disease. *BMC Microbiol.* **18**, 50 (2018).

35. Ruiz, L., Delgado, S., Ruas-Madiedo, P., Sánchez, B. & Margolles, A. Bifidobacteria and their molecular communication with the immune system. *Front.*

Microbiol. **8**, 2345 (2017).

36. Sokol, H. *et al.* Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. *Proc. Natl. Acad. Sci.* **105**, 16731–16736 (2008).

37. Domingue, G. J. & Schlegel, J. U. Novel bacterial structures in human blood: cultural isolation. *Infect. Immun.* **15**, 621–627 (1977).

38. Lobry, J. R. Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol. Biol. Evol.* **13**, 660–665 (1996).

39. Yang, C.-C. *et al.* Characteristics and outcomes of Fusobacterium nucleatum bacteremia—a 6-year experience at a tertiary care hospital in northern Taiwan. *Diagn. Microbiol. Infect. Dis.* **70**, 167–174 (2011).

40. DEMMLER, G. J., COUCH, R. S. & TABER, L. H. Neisseria subflava bacteremia and meningitis in a child: report of a case and review of the literature. *Pediatr. Infect. Dis. J.* **4**, 286–288 (1985).

41. Oill, P. A., Chow, A. W. & Guze, L. B. Adult bacteremic Haemophilus parainfluenzae infections: seven reports of cases and a review of the literature. *Arch. Intern. Med.* **139**, 985–988 (1979).

42. Chan, J. F. W. *et al.* First report of spontaneous intrapartum Atopobium vaginae bacteremia. *J. Clin. Microbiol.* **50**, 2525–2528 (2012).

43. Mendes, R. E. *et al.* Assessment of linezolid resistance mechanisms among Staphylococcus epidermidis causing bacteraemia in Rome, Italy. *J. Antimicrob. Chemother.* **65**, 2329–2335 (2010).

44. Choi, J. Y. *et al.* Mortality risk factors of Acinetobacter baumannii bacteraemia. *Intern. Med. J.* **35**, 599–603 (2005).

45. Wertlake, P. T. & Williams, T. W. Septicaemia caused by Neisseria flavescens. *J. Clin. Pathol.* **21**, 437–439 (1968).

46. Shah, S. S., Ruth, A. & Coffin, S. E. Infection due to Moraxella osloensis: case report and review of the literature. *Clin. Infect. Dis.* **30**, 179–181 (2000).

47. Felten, A., Barreau, C., Bizet, C., Lagrange, P. H. & Philippon, A. Lactobacillus species identification, H2O2 production, and antibiotic resistance and correlation with human clinical status. *J. Clin. Microbiol.* **37**, 729–733 (1999).

48. JeŽek, P. *et al.* Corynebacterium imitans isolated from blood culture in a patient with suspected bacteremia-the first isolation from human clinical material in the Czech Republic. *Klin. Mikrobiol. Infekc. Lek.* **20**, 98–101 (2014).

49. Friedman, J. & Alm, E. J. Inferring correlation networks from genomic survey data. *PLoS Comput Biol* **8**, e1002687 (2012).

50. Anhê, F. F. *et al.* Type 2 diabetes influences bacterial tissue compartmentalisation in human obesity. *Nat. Metab.* **2**, 233–242 (2020).

51. Emery, D. C. *et al.* Comparison of blood bacterial communities in periodontal health and periodontal disease. *Front. Cell. Infect. Microbiol.* **10**, 799 (2021).

52. Simon, A. K., Hollander, G. A. & McMichael, A. Evolution of the immune system in humans from infancy to old age. *Proc. R. Soc. B Biol. Sci.* **282**, 20143085 (2015).

53. Blauwkamp, T. A. *et al.* Analytical and clinical validation of a microbial cell-free DNA sequencing test for infectious disease. *Nat. Microbiol.* **4**, 663–674 (2019).

54. Grumaz, C. *et al.* Rapid Next-Generation Sequencing–Based Diagnostics of Bacteremia in Septic Patients. *J. Mol. Diagnostics* **22**, 405–418 (2020).

55. Tan, C. C. S., Acman, M., van Dorp, L. & Balloux, F. Metagenomic evidence for a polymicrobial signature of sepsis. *Microb. genomics* **7**, (2021).

56. Grumaz, S. *et al.* Next-generation sequencing diagnostics of bacteremia in septic patients. *Genome Med.* **8**, 73 (2016).

57. Grumaz, S. *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.* **47**, e394 (2019).

58. Faust, K. & Raes, J. Microbial interactions: from networks to models. *Nat. Rev. Microbiol.* **10**, 538–550 (2012).

59. Zelezniak, A. *et al.* Metabolic dependencies drive species co-occurrence in

diverse microbial communities. *Proc. Natl. Acad. Sci.* **112**, 6449–6454 (2015).

60. Hibbing, M. E., Fuqua, C., Parsek, M. R. & Peterson, S. B. Bacterial competition: surviving and thriving in the microbial jungle. *Nat. Rev. Microbiol.* **8**, 15–25 (2010).

61. Cross, A. & Levine, M. M. Patterns of bacteraemia aetiology. *Lancet Infect. Dis.* **17**, 1005–1006 (2017).

62. Dethlefsen, L., McFall-Ngai, M. & Relman, D. A. An ecological and evolutionary perspective on human–microbe mutualism and disease. *Nature* **449**, 811–818 (2007).

63. Geva-Zatorsky, N. *et al.* Mining the human gut microbiota for immunomodulatory organisms. *Cell* **168**, 928–943 (2017).

64. Gensollen, T., Iyer, S. S., Kasper, D. L. & Blumberg, R. S. How colonization by microbiota in early life shapes the immune system. *Science (80-.).* **352**, 539–544 (2016).

65. Brenner, 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).* **97**, (2018).

66. Shah, N. B. *et al.* Blood microbiome profile in CKD: a pilot study. *Clin. J. Am. Soc. Nephrol.* **14**, 692–701 (2019).

67. Camilleri, M. Leaky gut: mechanisms, measurement and clinical implications in humans. *Gut* **68**, 1516–1526 (2019).

68. Wu, D. *et al.* Large-scale whole-genome sequencing of three diverse Asian populations in Singapore. *Cell* **179**, 736–749 (2019).

69. Foong, A. W. P. *et al.* Rationale and methodology for a population-based study of eye diseases in Malay people: The Singapore Malay eye study (SiMES). *Ophthalmic Epidemiol.* **14**, 25–35 (2007).

70. Lavanya, R. *et al.* Methodology of the Singapore Indian Chinese Cohort (SICC) eye study: quantifying ethnic variations in the epidemiology of eye diseases in Asians. *Ophthalmic Epidemiol.* **16**, 325–336 (2009).

71. Tan, K. H. X. *et al.* Cohort profile: the Singapore multi-ethnic cohort (mec) study. *Int. J. Epidemiol.* **47**, 699-699j (2018).

72. Soh, S.-E. *et al.* Cohort profile: Growing Up in Singapore Towards healthy Outcomes (GUSTO) birth cohort study. *Int. J. Epidemiol.* **43**, 1401–1409 (2014).

73. Teo, Y.-Y. *et al.* Singapore Genome Variation Project: a haplotype map of three Southeast Asian populations. *Genome Res.* **19**, 2154–2162 (2009).

74. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *bioinformatics* **25**, 1754–1760 (2009).

75. Gourlé, H., Karlsson-Lindsjö, O., Hayer, J. & Bongcam-Rudloff, E. Simulating Illumina metagenomic data with InSilicoSeq. *Bioinformatics* **35**, 521–522 (2019).

76. Aitchison, J. The statistical analysis of compositional data. *J. R. Stat. Soc. Ser. B* **44**, 139–160 (1982).

77. Van den Boogaart, K. G. & Tolosana-Delgado, R. “Compositions”: a unified R package to analyze compositional data. *Comput. Geosci.* **34**, 320–338 (2008).

78. Shaw, L. The phylogenetic range of bacterial and viral pathogens of vertebrates: dataset and supplementary material. (2020) doi:10.6084/m9.figshare.8262779.v2.

79. Jorgensen, J. *et al. Manual of Clinical Microbiology*. (American Society for Microbiology Press, 2015). doi:10.1128/9781555817381.

80. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

81. Morgan, M., Pagès, H., Obenchain, V. & Hayden, N. Rsamtools: Binary alignment (BAM), FASTA, variant call (BCF), and tabix file import. R package version 2.8.0. (2021).

82. Kurtz, Z. D. *et al.* Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput Biol* **11**, e1004226 (2015).

83. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal, complex Syst.* **1695**, 1–9 (2006).

84. Wickham, H. ggplot2. *Wiley Interdiscip. Rev. Comput. Stat.* **3**, 180–185 (2011).

Figure 1

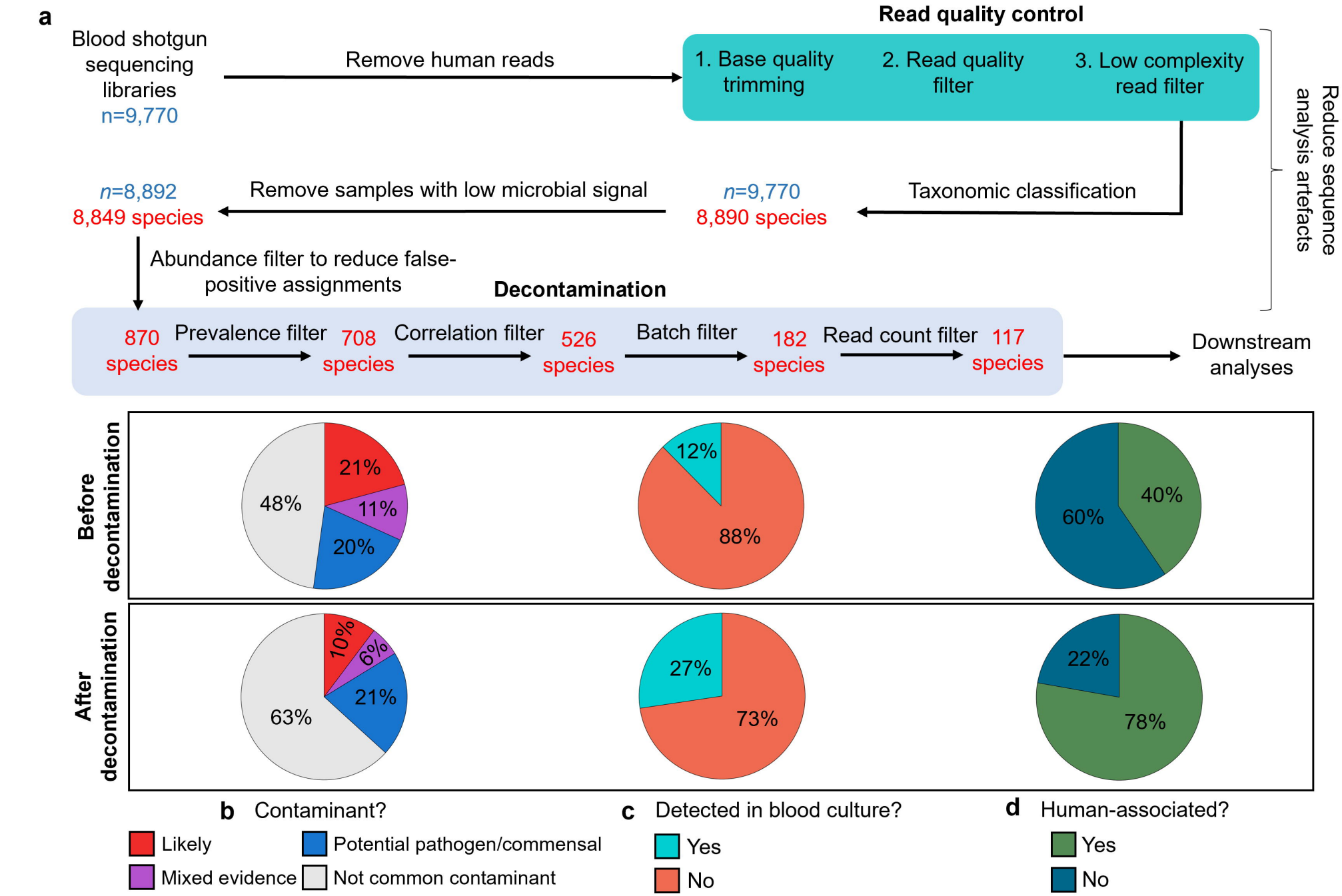

a
Blood shotgun sequencing libraries
n=9,770
Remove human reads
Read quality control
1. Base quality trimming
2. Read quality filter
3. Low complexity read filter
Taxonomic classification
n=9,770
8,890 species
Remove samples with low microbial signal
n=8,892
8,849 species
Abundance filter to reduce false-positive assignments
Decontamination
870 species
Prevalence filter
708 species
Correlation filter
526 species
Batch filter
182 species
Read count filter
117 species
Downstream analyses
Reduce sequence analysis artefacts
Before decontamination
After decontamination
21%
11%
20%
48%
12%
88%
40%
60%
10%
6%
21%
63%
27%
73%
22%
78%
b Contaminant?
Likely
Potential pathogen/commensal
Mixed evidence
Not common contaminant
c Detected in blood culture?
Yes
No
d Human-associated?
Yes
No

Figure 2

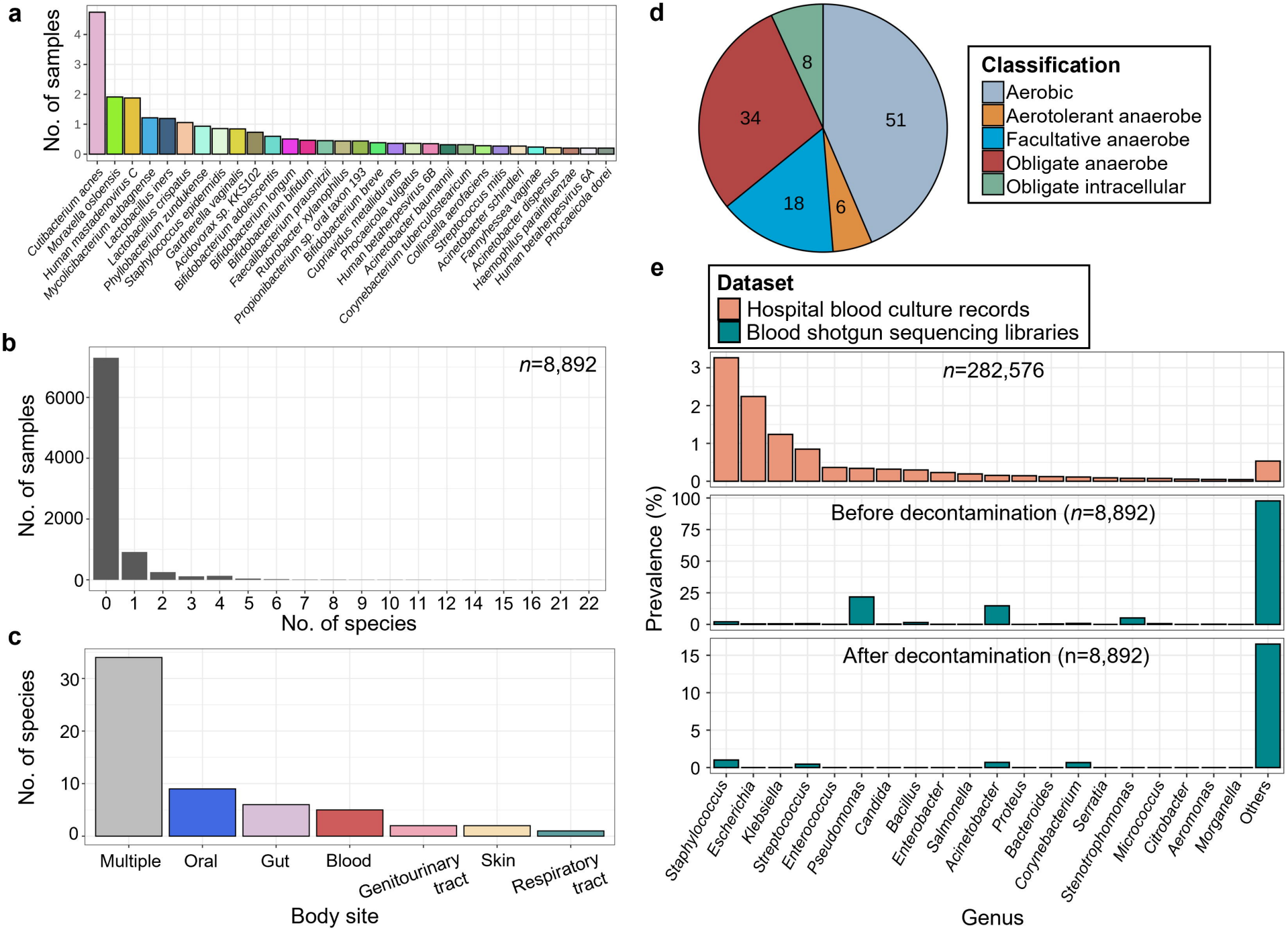

a
No. of samples
Cutibacterium acnes
Moraxella osloensis
Human mastadenovirus C
Mycolicibacterium aubagnense
Lactobacillus iners
Lactobacillus crispatus
Phyllobacterium zundukense
Staphylococcus epidermidis
Gardnerella vaginalis
Acidovorax sp. KKS102
Bifidobacterium adolescentis
Bifidobacterium longum
Faecalibacterium prausnitzii
Rubrobacter xylanophilus
Propionibacterium sp. oral taxon 193
Bifidobacterium breve
Bifidobacterium bifidum
Cupriavidus metallidurans
Phocaeicola vulgatus
Human betaherpesvirus 6B
Acinetobacter baumannii
Corynebacterium tuberculostearicum
Collinsella aerofaciens
Streptococcus mitis
Acinetobacter schindleri
Fannyhessea vaginae
Acinetobacter dispersus
Haemophilus parainfluenzae
Acinetobacter parvus
Human betaherpesvirus 6A
Phocaeicola dorei
b
n=8,892
No. of samples
No. of species
c
No. of species
Multiple
Oral
Gut
Blood
Genitourinary tract
Skin
Respiratory tract
Body site
d
Classification
Aerobic
Aerotolerant anaerobe
Facultative anaerobe
Obligate anaerobe
Obligate intracellular
51
6
18
34
8
e
Dataset
Hospital blood culture records
Blood shotgun sequencing libraries
n=282,576
Before decontamination (n=8,892)
After decontamination (n=8,892)
Prevalence (%)
Staphylococcus
Escherichia
Klebsiella
Streptococcus
Enterococcus
Pseudomonas
Candida
Bacillus
Enterobacter
Salmonella
Acinetobacter
Proteus
Bacteroides
Corynebacterium
Serratia
Stenotrophomonas
Micrococcus
Citrobacter
Aeromonas
Morganella
Others
Genus

Figure 3

a

Sample	Species	Body site(s)	Reported in blood	No. of read pairs assigned	Overall prevalence (%)	PTR
WHB4594	*Fusobacterium nucleatum*	genitourinary tract, gut, mouth	✓	194199	0.11	1.68
WHB9179	*Neisseria subflava*	gut, mouth	✓	15385	0.16	1.51
WHB9179	*Haemophilus parainfluenzae*	gut, mouth, respiratory tract	✓	12183	0.2	1.17
WHB4035	*Fannyhessea vaginae*	genitourinary tract	✓	10395	0.24	1.88
WHB6459	*Staphylococcus epidermidis*	gut, mouth, respiratory tract, skin	✓	9140	0.85	1.57
WHB10710	*Lactobacillus crispatus*	genitourinary tract, gut, mouth	✓	7799	1.06	1.57
0116-0053	*Acinetobacter baumannii*	mouth	✓	7673	0.31	1.9
WHB9179	*Neisseria flavescens*	mouth	✓	3787	0.06	1.38
WHB9978	*Rickettsia sp. Tillamook 23*		✗	2923	0.02	1.35
WHH1248	*Moraxella osloensis*	skin	✓	2402	1.91	1.33
WHB9812	*Corynebacterium imitans*		✓	1976	0.02	1.59

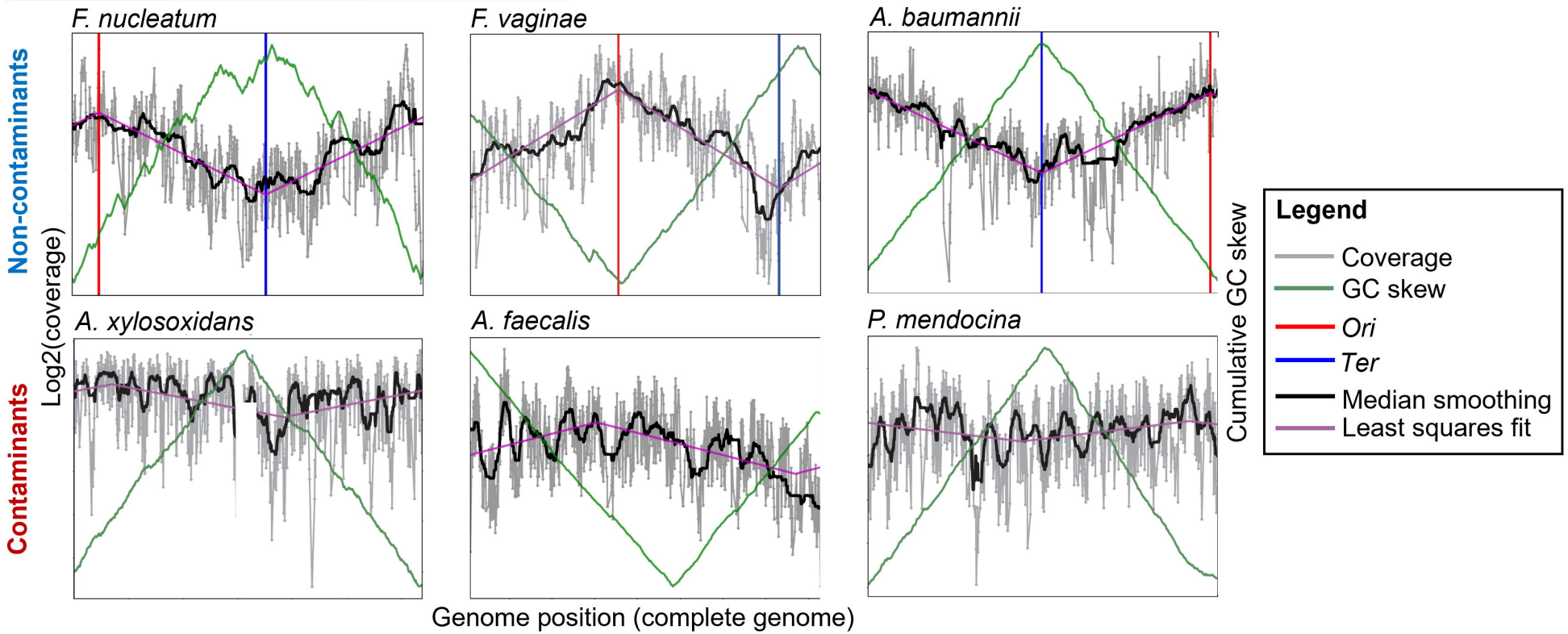


Figure 4

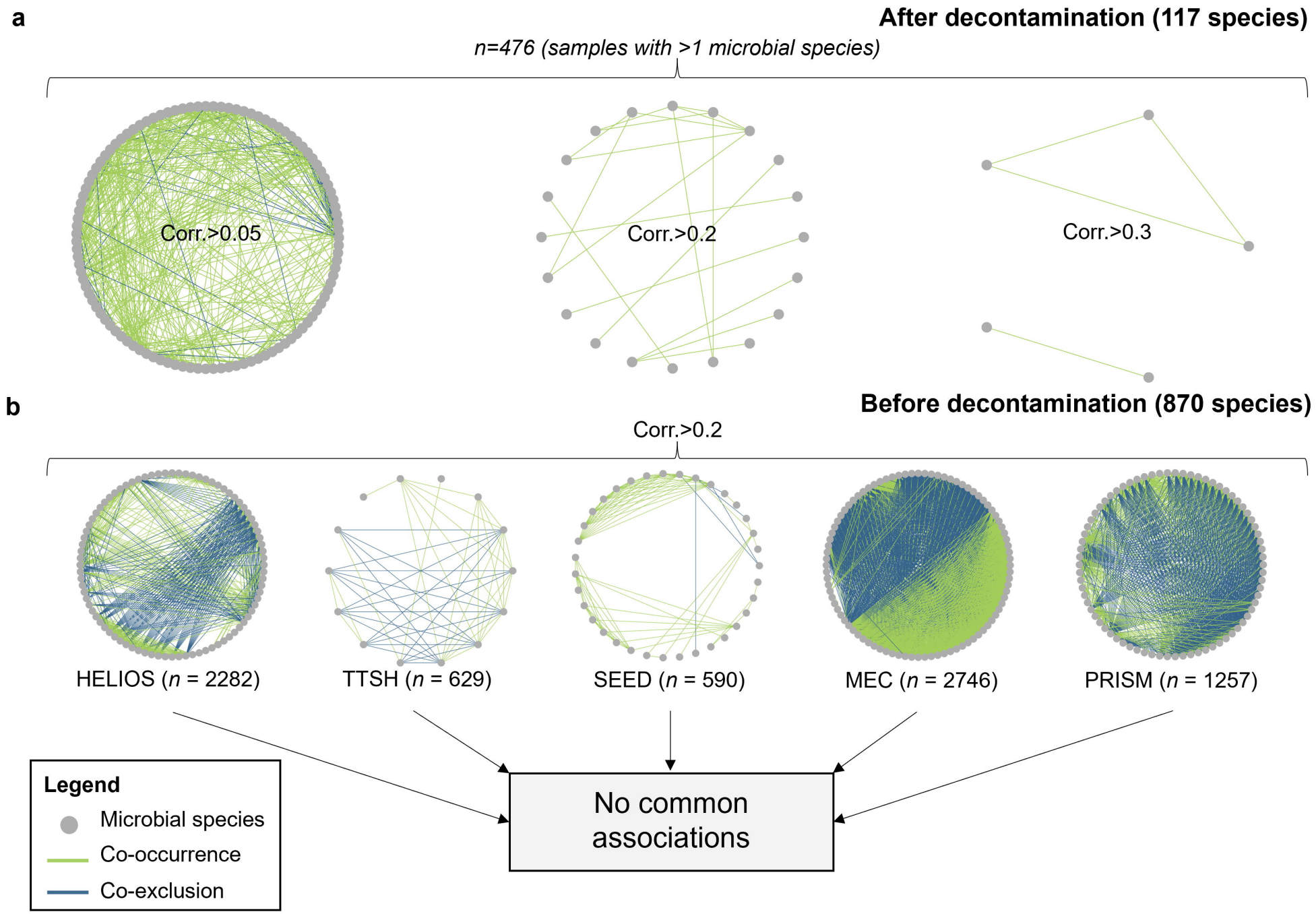
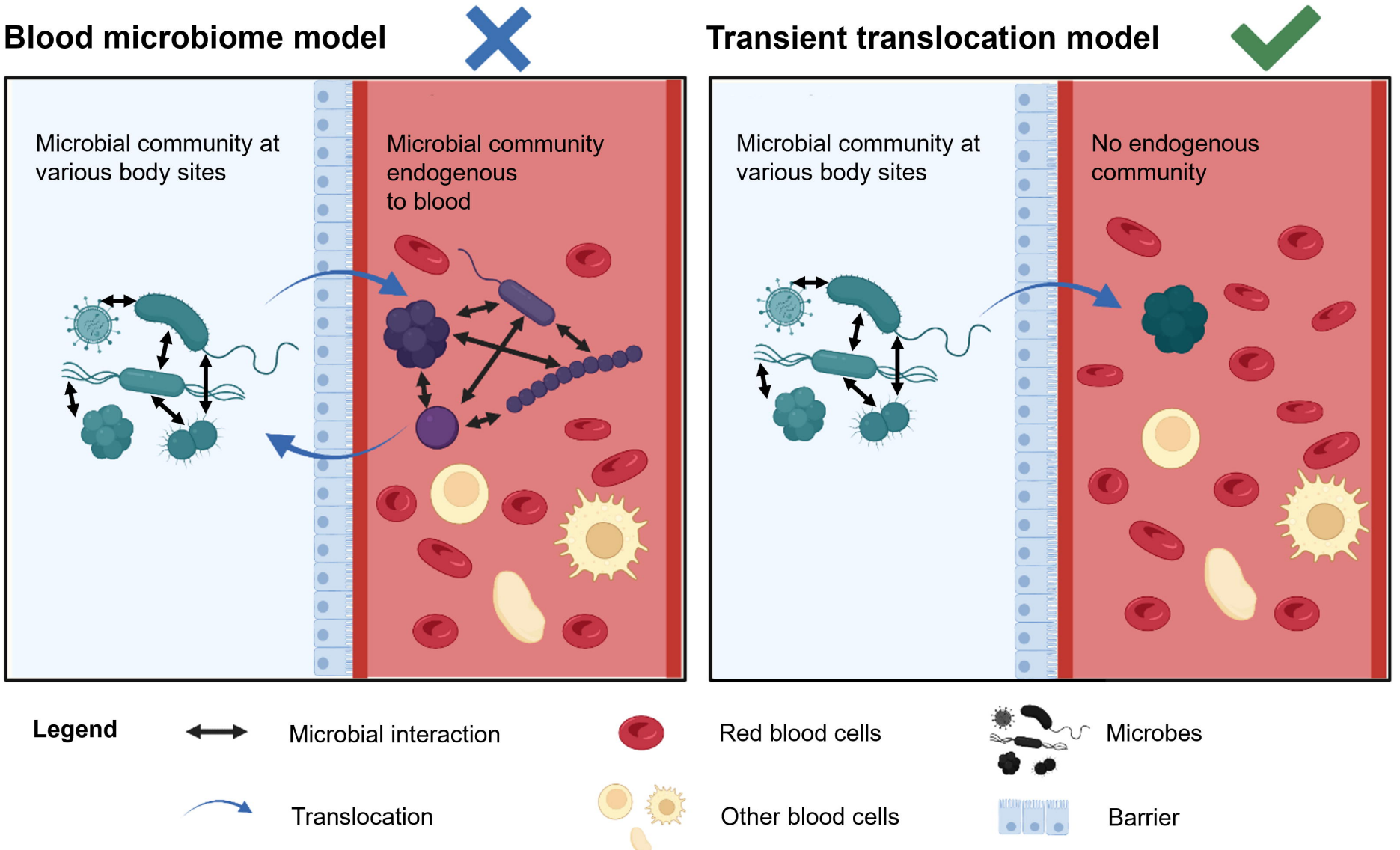

a
After decontamination (117 species)
n=476 (samples with >1 microbial species)
Corr.>0.05
Corr.>0.2
Corr.>0.3
b
Before decontamination (870 species)
Corr.>0.2
HELIOS (n = 2282)
TTSH (n = 629)
SEED (n = 590)
MEC (n = 2746)
PRISM (n = 1257)
No common associations
Legend
Microbial species
Co-occurrence
Co-exclusion

Figure 5


Blood microbiome model
Microbial community at various body sites
Microbial community endogenous to blood
Transient translocation model
Microbial community at various body sites
No endogenous community
Legend
Microbial interaction
Translocation
Red blood cells
Other blood cells
Microbes
Barrier