The whole blood microbiome of Indonesians reveals translocated and pathogenic microbiota

Katalina Bobowik[1,2,3,*], Muhamad Fachrul[2,4,*], Chelzie Crenna Darusallam[5], Pradiptajati Kusuma[5], Herawati Sudoyo[5], Clarissa A. Febinia[5], Safarina G. Malik[5], Christine Wells[3], Irene Gallego Romero[1,2,3,4,6,†]

1 Melbourne Integrative Genomics, University of Melbourne, Parkville, VIC, Australia

2 School of BioSciences, The University of Melbourne, Parkville, VIC, Australia

3 The Centre for Stem Cell Systems, Faculty of Medicine, Dentistry and Health Sciences, The University of Melbourne, Parkville, VIC, Australia

4 St Vincent's Institute of Medical Research, Fitzroy, VIC, Australia

5 Genome Diversity and Diseases Division, Mochtar Riady Institute for Nanotechnology, Tangerang, Banten, Indonesia

6 Center for Genomics, Evolution and Medicine, Institute of Genomics, University of Tartu, Riia 23b, 51010 Tartu, Estonia

* These authors contributed equally

† Correspondence: irene.gallego@unimelb.edu.au

Abstract

Pathogens found within local environments are a major cause of morbidity and mortality. This is particularly true in Indonesia, where infectious diseases such as malaria or dengue are a significant part of the disease burden. Unequal investment in medical funding throughout Indonesia, particularly in rural areas, has resulted in under-reporting of cases, making surveillance challenging. Here, we use transcriptome data from 117 healthy individuals living on the islands of Mentawai, Sumba, and the Indonesian side of New Guinea Island to explore which pathogens are present within whole blood. We detect a range of taxa within RNA-sequencing data generated from whole blood and find that two pathogens—Flaviviridae and Plasmodium—are the most predominantly abundant, both of which are most pronounced in the easternmost island within our Indonesian dataset. We also compare the Indonesian data to two other cohorts from Mali and UK and find a distinct microbiome profile for each group. This study provides a framework for RNA-seq as a possible retrospective surveillance tool and an insight to what makes up the transient human blood microbiome.

Introduction

Pathogens are a major cause of morbidity and mortality, especially in the Global South [1–3]. Current knowledge of which taxa are present within remote regions of the world, along with how they impact health outcomes, remains limited. Not only is surveillance complex in these settings, but identifying which pathogens are responsible for disease symptoms can be challenging. For instance, although a pathogen may be identified in a population, it might not be the causative agent of disease due to indistinguishable symptoms and cross-reactivity of multiple pathogens [4]. Having a more detailed understanding of which pathogens are the major causes of morbidity across different global populations can focus elimination efforts on specific pathogens and aid in more targeted disease therapeutics.

Blood transcriptome data can be used to empirically test which blood-borne pathogens are present within an individual. Along with pathogenic organisms that infect blood cells, such as arthropod-borne pathogens [5, 6] and various viruses [7, 8], emerging research has shown that even bacteria and fungi can release DNA and RNA into blood [9]. For example, commensal bacteria [10, 11], viruses [12, 13], fungi [14], and archaea [15] have all been identified independently in multiple studies of human blood. While not yet common, the use of blood as a surveillance tool is growing. For instance, Kafetzopoulou et al. [16] used plasma samples from Lassa fever patients to identify the emergence of new strains, while two recent studies used whole blood samples from critically endangered mammals [17] and songbirds [18] to aid in the characterisation of diverse blood parasites.

Still, the topic of whether consistent microbial communities exist across healthy individuals remains highly debatable. Recent studies have found no evidence of a core microbiome circulating in the blood of healthy individuals [19]. This suggests that the blood microbiome is more transient in nature, comprising of either commensal microbes translocated from other body sites or those involved in pathogenic activity and other disease states [20]. However, not enough studies have been done to confirm whether this lack of a core microbiome is also consistent in regions where infectious diseases are endemic.

Indonesia is a country with large numbers of endemic and emerging infectious diseases [21], making it a crucially important location to monitor and understand the effects of pathogens on human hosts. While several endemic diseases have been successfully reduced or eliminated in Indonesia [22], pathogen abundance can still be high in more rural areas, which tend to have less access to medical resources [22–24]. We have previously sampled individuals from three remote islands in Indonesia—Mentawai, Sumba, and the Indonesian side of New Guinea Island—and showed that individuals from the easternmost side of Indonesia (New Guinea Island) show widespread differences in immune gene expression levels compared

to individuals from western (Mentawai) or central (Sumba) Indonesian islands [25]. While some of this variation is likely attributable to the different genetic ancestries of individuals in these islands [25, 26], another significant contributor may be environmental differences, such as pathogenic load. Indeed, both *Plasmodium falciparum* and *Plasmodium vivax* are detectable at low levels within whole blood of some of these individuals [27], with a higher Plasmodium abundance within individuals from New Guinea Island. This observation suggests that pathogen loads are variable across the country, and that a non-targeted, transcriptomic approach can be used to capture these differences.

To characterise blood-borne microorganisms within Indonesia, this study utilises transcriptomic data collected from whole blood within these three previously described groups: the peoples of Mentawai and Sumba, and the Korowai. These populations span a gradient from west to east across Indonesia, thus capturing pathogens along the main geographical axis of the country. Unlike more populous regions within Indonesia, these three islands serve as models to understand pathogen load in areas with limited resources and where reporting and traditional surveillance methods can be challenging. This can therefore provide a valuable resource from under-represented areas.

Methods

Datasets

The Indonesian dataset consists of 101 base-pair, paired-end RNA-seq data from the whole blood of 117 healthy individuals living on the Indonesian islands of Sumba (n = 49), Mentawai (n = 48), and on the Indonesian side of New Guinea Island (n = 20, as described in [25]; all Indonesian data are available from the European Genome-phenome Archive study EGAS00001003671). All collections and analyses followed protocols for the protection of human subjects established by institutional review boards at the Eijkman Institute (EIREC #90 and #126); the analyses in this publication were additionally approved by University of Melbourne's Human Ethics Advisory Group (1851639.1). In the original Natri et al. study, additional 6 libraries were generated to serve as technical replicates between sequencing batches, however for our study we only retained the replicate with the highest read depth. Samples for the dataset were collected using Tempus Blood RNA Tubes (Applied Biosystems) and RNA-Seq libraries were prepared using Illumina's Globin-Zero Gold rRNA Removal Kit. Samples were then sequenced on an Illumina HiSeq 2500, resulting in an average read depth of 30 million read pairs per individual (Supplementary File 1).

To compare the Indonesian dataset to other global populations, we searched for multiple publicly available transcriptomic datasets of whole blood from self-described healthy human donors. To control for technical covariates, we limited ourselves to datasets prepared using a globin depletion method and collected using Tempus Blood RNA Tubes, the same process followed by our own Indonesian dataset. We identified two publicly available datasets as controls. The first dataset comes from Tran et al. [28, 29], and consists of 101-bp human whole blood RNA-seq data, hereafter referred to as the Mali study. As described in [29], samples were collected from individuals living in the rural village of Kalifabougou, Mali, an area where there is a high rate of seasonal *P. falciparum* transmission. Raw sequence reads for this study were downloaded from SRA study GSE52166 and only samples which were collected pre-infection (n = 54) were used. The second dataset comes from Singhania et al. [30] consisting of 75-bp human whole blood RNA-seq data, collected from volunteers at the MRC National Institute for Medical Research in London, UK, hereafter referred to as the UK study. Raw sequence reads for this study were downloaded from SRA study GSE107991 and only healthy control samples (n = 12; all of European ethnicity) were used.

RNA sequencing data processing

To investigate the metatranscriptome of whole blood, we put all reads through a stringent quality control pipeline. RNA-seq reads from all datasets went through an initial sample quality analysis using FastQC v. 0.11.5 [31]. To ensure reads were of high quality and free from artefacts, leading and trailing bases below a Phred quality score of 20 were removed and universal Illumina adapter sequences were trimmed (TruSeq3-PE.fa) using Trimmomatic v. 0.36 [32]. For comparisons between the Indonesian, Malian, and UK populations, the Malian and Indonesian datasets were trimmed to 75-bp, which is the read length of the UK dataset. We did this to control for differences in mappability and taxa identification associated with read length.

Paired-end RNA-seq reads were first aligned to the human genome (GRCh38, Ensembl release 90: August 2017) with STAR v. 2.5.3a [33] using the two-pass alignment mode and default parameters, and only reads that did not map to the human genome were retained for further analysis. This step was performed to reduce the total library size to only pathogen candidates, and significantly decreases subsequent processing time. Unmapped sequencing reads were then processed using KneadData v. 0.7.4, which uses BMTagger [34] and Tandem Repeats Finder (TRF) [35] to remove human contaminant reads and tandem repeats, respectively. Using Kneaddata, BMtagger and TRF were run with default parameters. This resulted in a mean of 39,863 and 58,424 reads per sample for the 101-bp (Supplementary Table 1) and 75-bp (Supplementary Table 2) Indonesian datasets, respectively. For the 75-bp Malian (Supplementary Table 3) and UK (Supplementary Table 4) datasets, this resulted in a mean of 300,123 and 422,404 reads per sample, respectively.

Mapping and metagenomic classification

Processed metagenomic reads were mapped using KMA v. 1.2.21 [36] against a filtered NCBI nt reference database, where artificial sequences and environmental sequences without valid taxonomic IDs were excluded [37] (downloaded on June 28, 2019 from https://researchdata.edu.au/indexed-reference-databases-kma-ccmetagen/1371207). We mapped paired-end reads using default settings and the following additional flags: -ef (extended features) was used to calculate reads as the total number of fragments, -1t1 was used for one read to one template (no splicing allowed in the reads), and -apm was set to p which rewards pairing of reads. After mapping, we performed read classification using CCMetagen v. 1.2.2 [38] with default settings for paired-end reads. Read depth was calculated using the number of fragments with the read depth set to 1 so that we could analyse all possible matches. For the Indonesian dataset, these steps resulted in a mean of 6,480 reads per sample, which dropped to 4,579 when we trimmed reads to 75-bp (Supplementary Table

2). For the 75-bp Malian (Supplementary Table 3) and UK (Supplementary Table 4) datasets, this resulted in a mean of 8,129 and 15,494 reads, respectively.

Data filtering

After removing singletons to prevent spurious identification of taxa, we filter out reads mapped to the kingdoms Viridiplantae as these likely represented misassignments or poor quality annotation (Supplementary Figure 1, A-D) and further investigated the metazoan reads. We found that the majority of these mapped to the phylum Chordata (Supplementary Figure 1, E-H). We therefore decided to discard all reads mapping to Metazoa from subsequent analysis, as BLAST analysis of confirmed that these were reads that mapped equally well to the human genome. In addition, we also chose to remove taxa with no taxonomic rank assigned at the superkingdom level, as these taxa could not be linked to any known species. After removing Viridiplantae, Metazoa, and taxa with no taxonomic rank assigned at the superkingdom level, we obtained a mean of 905 reads in the Indonesian dataset (a mean of 694 for the 75-bp Indonesian reads; Supplementary Table 2), 546 for the 75-bp Malian dataset (Supplementary Table 3), and 5,230 for the 75-bp UK dataset (Supplementary Table 4; Supplementary Figure 2).

Sample clustering

To correct for uneven library depth between samples and the compositional nature of microbiome data [39], we applied a center log ratio (CLR) transformation [40] to the taxa abundance matrix when performing principal component analysis (PCA). Since a high number of zeros were present in the data, which CLR transformation is sensitive to [41], we chose to merge the abundance matrix at the phylum level. For this reason, we also performed analyses at the phylum level for all subsequent analyses utilising CLR-transformation. Throughout, analyses are reported at the taxonomic level at which they were carried out, unless otherwise noted.

Differential abundance testing and diversity estimation

We used ANOVA-like differential expression (ALDEx2) [42–44] to test for differences in species composition between populations, which applies CLR-transformation to correct for uneven library depth and data compositionality [43]. We performed differential abundance testing at the phylum level using the default Welch's t-test and default 128 Monte Carlo simulations. For alpha and beta diversity estimates, we used count abundances at the phylum level without removing singletons using the package DivNet v. 0.3.6 [45],

which expects the presence of singletons in order to model species richness [45].

Code for all analyses is available at https://gitlab.svi.edu.au/muhamad.fachrul/indo_blood_microbiome

Results

The blood microbiome of Indonesians

To provide a more comprehensive understanding of the blood microbiome of remote populations within Indonesia, we analysed unmapped reads from previously published whole blood transcriptomes, collected from 117 Indonesian individuals living on the islands of Mentawai (MTW) in western Indonesia, and Sumba (SMB) in central Indonesia, as well as the Korowai (KOR), a group living on the Indonesian side of New Guinea Island. The human samples have been extensively described [25, 26]. After extensive quality control, we obtained a mean library size of 6,480 taxonomically informative reads after the removal of singletons (range: 2,212 - 48,471; Supplementary Table 1). We assigned these reads to a total of 50 taxa across all phylogenetic levels, including 25 distinct taxa at the family level. As reads were predominantly assigned to Metazoan taxa, including *Homo sapiens*, further filtering was done; this resulted in an average of 3,923 reads across 27 samples that passed filtering, mapping to 15 distinct taxa at family level. *Plasmodiidae* (54.5% of the total read pool across all individuals) and *Flaviviridae* (40.8% of reads) were families with most reads assigned (Figure 1A). To control for sparsity in the abundance matrix, which is crucial when performing CLR-transformation [41], we also analysed the abundance of taxa at the phylum level in tests applying a CLR transformation to the data. Analysis of microbial reads at the phylum level resulted in the identification of 9 taxa, with Apicomplexa (54.3% of reads, within which 99.9% of reads mapped to the family *Plasmodiidae*), Kitrinoviricota (41% of reads, within which 100% of reads mapped to *Flaviviridae*), Ascomycota (1.6% of reads), and Pseudomonadota (0.8% of reads) making up the majority. These estimates of Apicomplexa load are higher than our previous estimates of Plasmodium burden [27], where we used a different, more conservative approach. We observed that the microbiome composition varied substantially between islands. In Korowai and Sumba populations, the majority of samples had reads assigned to either Apicomplexa (71.3% and 67.3% of reads) or Kitrinoviricota (28.1% of and 26.6% of reads, respectively), whereas majority of reads in Mentawai samples mapped to Kitrinoviricota (91.8%).

PCA of the CLR-transformed taxonomic matrix showed sample clustering clearly driven by the phyla Apicomplexa (Figure 1B) and Kitrinoviricota (Figure 1C). We found that PC1, which captured over 40% of the variation, separated individuals by their abundance of either of these pathogens, as well as separating

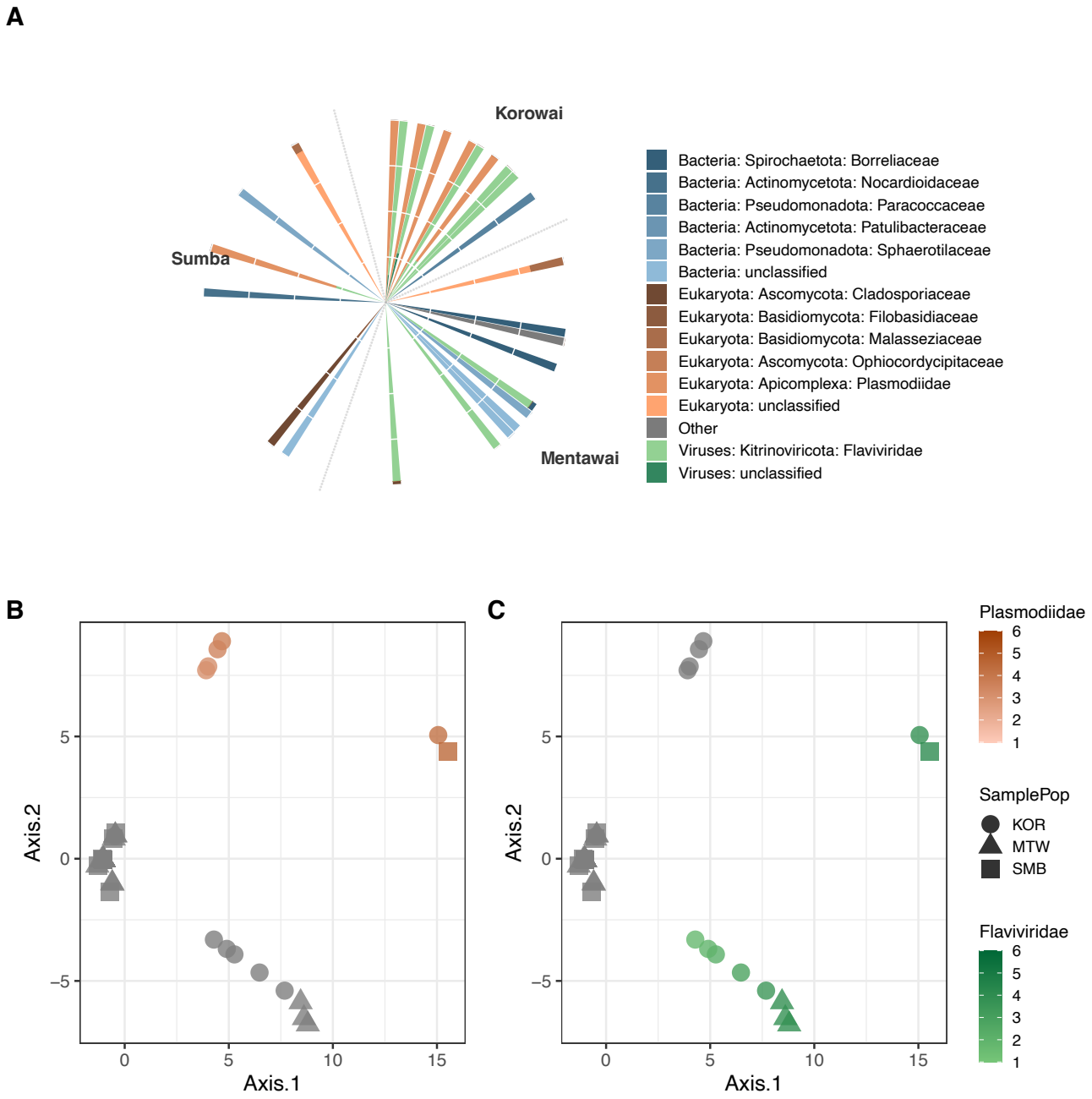


Fig 1: The blood metatranscriptome of the Indonesian populations. A) Circular barplot showing relative abundance (as percentage of reads) of the detected taxa within each individual in the Indonesian dataset, resolved at the family level. Bacteria are shown in blue, eukaryotes in orange, and viruses in green. KOR = Korowai; MTW = Mentawai; SMB = Sumba. Taxon labels include both phylum and family information. B) Principal component analysis of the CLR-normalised taxa abundance data at the phylum level. Plotting shapes indicate population while $\log_1$ *Plasmodiidae* abundance is indicated in orange and C) green for *Flaviviridae*.

the Korowai from most of the populations of Mentawai and Sumba (Figure 1B and C). PC2 could further be seen to separate samples with a high abundance of Apicomplexa from samples with a high abundance of

Kitrinoviricota (Figure 1B and C).

Microbiome diversity between island populations

As we are interested in whether there are observable differences in blood microbiomes between Indonesian island populations, we next performed differential abundance testing between the three groups using the ALDEx2 package [42–44]. Despite Apicomplexa having the largest abundance differences between sites, differential abundance testing at the phylum level did not result in significant differences in between islands, either before or after BH adjustment (Figure 2A, B, and C).

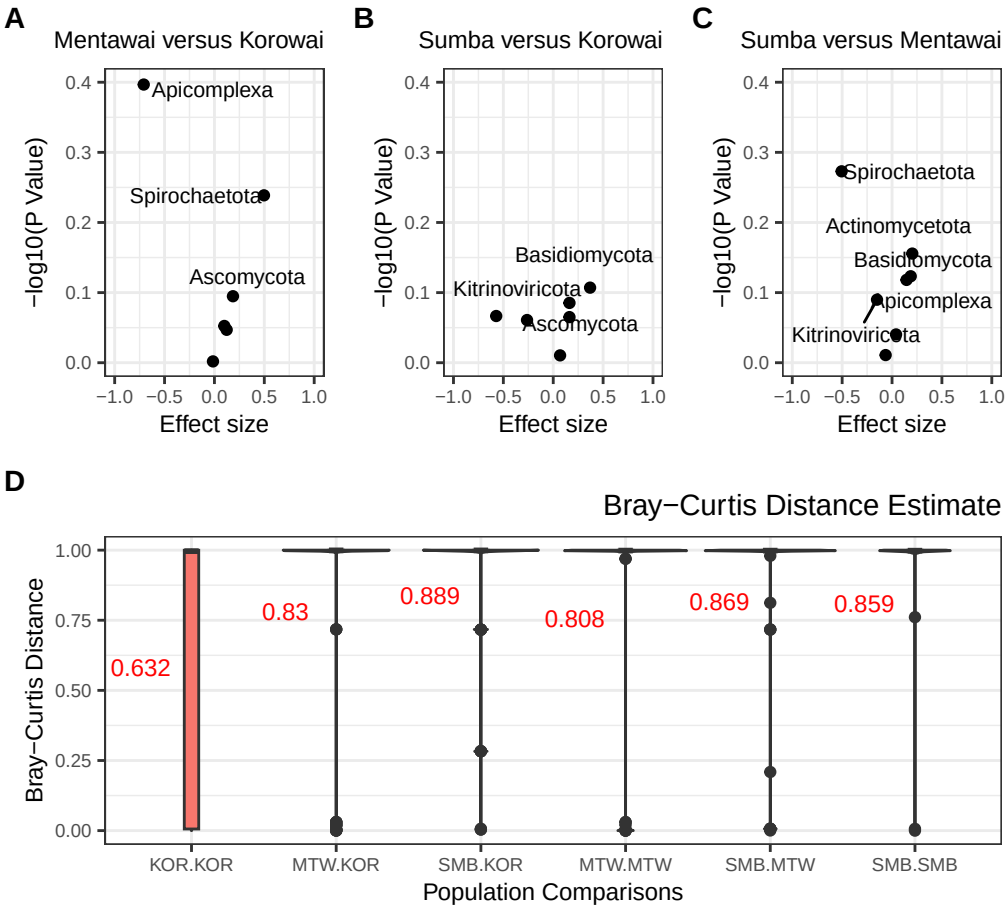


Fig 2: Blood microbiomes are not statistically different between island populations. A) Volcano plot of BH-adjusted p-values from Welch's t-test and the effect size for each taxa at the phylum level, in Korowai versus Mentawai B) Korowai versus Sumba and C) Mentawai versus Sumba. D) Bray-Curtis distance estimates for each population comparison at the phylum level.

The diversity and types of microbes within human tissues can be an indicator of the overall health of an individual, and of a population [11, 46]. We therefore analysed levels of alpha (within individual) and beta (between individual) diversity within the three islands using DivNet [45], again at the phylum level.

We found that while alpha diversity estimates were overall comparably lower in individuals from Korowai and Mentawai than in individuals from Sumba, they were still lowest in individuals from the Korowai population. This was true for both estimates of Shannon diversity (median Shannon KOR = 0.006; MTW = 0.023; SMB = 0.304; Supplementary Figure 3A) and inverse Simpson diversity indices (mean inverse Simpson KOR = 0.001; MTW = 0.005; SMB = 0.19; Supplementary Figure 3B). This observation was likely driven by the high abundance of Apicomplexa reads amongst the Korowai, which account for the majority of the available read pool in these individuals, and therefore drive overall diversity rates down. Other than within Korowai population, we found that most comparisons between populations resulted in similarly high estimates of Bray-Curtis dissimilarity (Figure 2D), which again mostly reflects the sparsity of the dataset, even between samples of the same island group.

Microbiomes are distinct between global populations

To test whether blood microbiomes in Indonesia differ from those of other global populations, we also analysed microbiome data from two other publicly available datasets of whole blood transcriptomes. This includes 54 healthy individuals living in Kalifabougou, Mali [28, 29], which represents the microbiome of individuals living in rural environments, and 12 healthy individuals collected from the city of London in the United Kingdom [30], representing the blood microbiome of individuals living in a highly urbanised environment. Similar to our Indonesian datasets, Kalifabougou is a malaria-endemic region and the majority of residents engage in subsistence farming practices [47].

After processing of reads as above, we obtained a mean library size of 15,494 reads (range: 4,493 - 33,711) for the UK dataset (Supplementary Table 4) and 8,129 (range: 1,637 - 180,484) for the Malian dataset (Supplementary Table 3) after the removal of singletons, respectively. This difference in depths is attributable to different numbers of reads being filtered out at different processing stages in the three datasets, as all three had similar starting read depths. All datasets lost significant numbers of reads when we filter reads assigned to either Viridiplantae or Metazoa (Supplementary Tables 1-4; Supplementary Figure 2). In the UK dataset, we identified a total of 101 distinct taxa across all phylogenetic levels. The majority of reads assigned to the bacterial phylum Pseudomonadota (81.2% of the total read pool across all individuals) and the fungal phylum Basidiomycota (11.2% of reads; Supplementary Figure 4). Within the Malian dataset, we found 41 distinct taxa across all phylogenetic levels, the majority of which were Actinomycetota (41.2% of reads), followed by Artverviricota (18.4% of reads), Apicomplexa (10.7% of reads), Kitrinoviricota (9.6% of reads), Bacillota (5.9% of reads), and Ascomycota (4.2% of reads; Supplementary

Figure 4). Although there is a substantial difference in read depths between all three data sets, saturation curves show systematic similarity in diversity between the Indonesian and Mali samples (Supplementary Figure 5).

We performed differential abundance testing between the Indonesian, Malian, and UK datasets. Only Actinomycetota (FDR adjusted Welch's t-test $p = 0.026$) was found to be significantly differentially abundant between Malian and Indonesian individuals (Figure 3A; Supplementary Table 5). Kitrinoviricota was found to be significantly differentially abundant prior to FDR correction (Welch's t-test $p = 0.011$; Supplementary Table 5). When comparing blood microbiomes between the UK and Indonesian populations, we found 2 differentially abundant phyla, the most significant being Pseudomonadota and Kitrinoviricota, the former more abundant in the UK population and the latter in the Indonesian population (FDR adjusted Welch's t-test $p = 3.76 \times 10^{-9}$ and 0.02, respectively; Figure 3B; Supplementary Table 6).

We next repeated differential abundance testing using only the Korowai as the Indonesian comparison group due to them containing the most pathogenic reads. We found that the comparisons yielded very similar results, with only Actinomycetota being significantly differentially abundant between Mali and Korowai samples (FDR adjusted Welch's t-test $p = 0.036$; Supplementary Table 7) and Pseudomonadota between UK and Korowai samples (FDR adjusted Welch's t-test $p = 3.2 \times 10^{-6}$; Supplementary Figure 6; Supplementary Table 8).

To identify overall trends between whole blood microbiomes of Indonesians and that of other populations, we next performed PCA on the CLR-transformed abundance matrix containing the Indonesian, UK, and Malian samples. Microbiomes clearly differed between countries as shown in PCs 1-2, yielding a separate cluster for each dataset (Figure 3C). PC2 in particular separated the Malian samples from the rest, with a clearer separation from the UK samples. This was recapitulated by the Bray-Curtis distance estimates, where population comparisons with Malian samples showed the greatest dissimilarity (Supplementary Figure 7). PCs 3 and 4 did not show any clear clustering of the populations and instead were driven by *Plasmodiidae* (Figure 3D) and *Flaviviridae* loads (Figure 3E).

Finally, to understand species richness in blood microbiomes between populations, we again analysed levels of alpha diversity in each of the three global datasets. We found that the UK samples had the lowest Shannon (mean Shannon = 0.195; Figure 3F) and inverse Simpson diversity values (mean inverse Simpson = 0.098; Figure 3G), followed by individuals from Mali (mean Shannon = 0.208, mean inverse Simpson = 0.11), then Indonesia (mean Shannon = 0.27, mean inverse Simpson = 0.116). We also note that the UK population has the highest sequencing depth out of the three populations (Supplementary Table 4) and

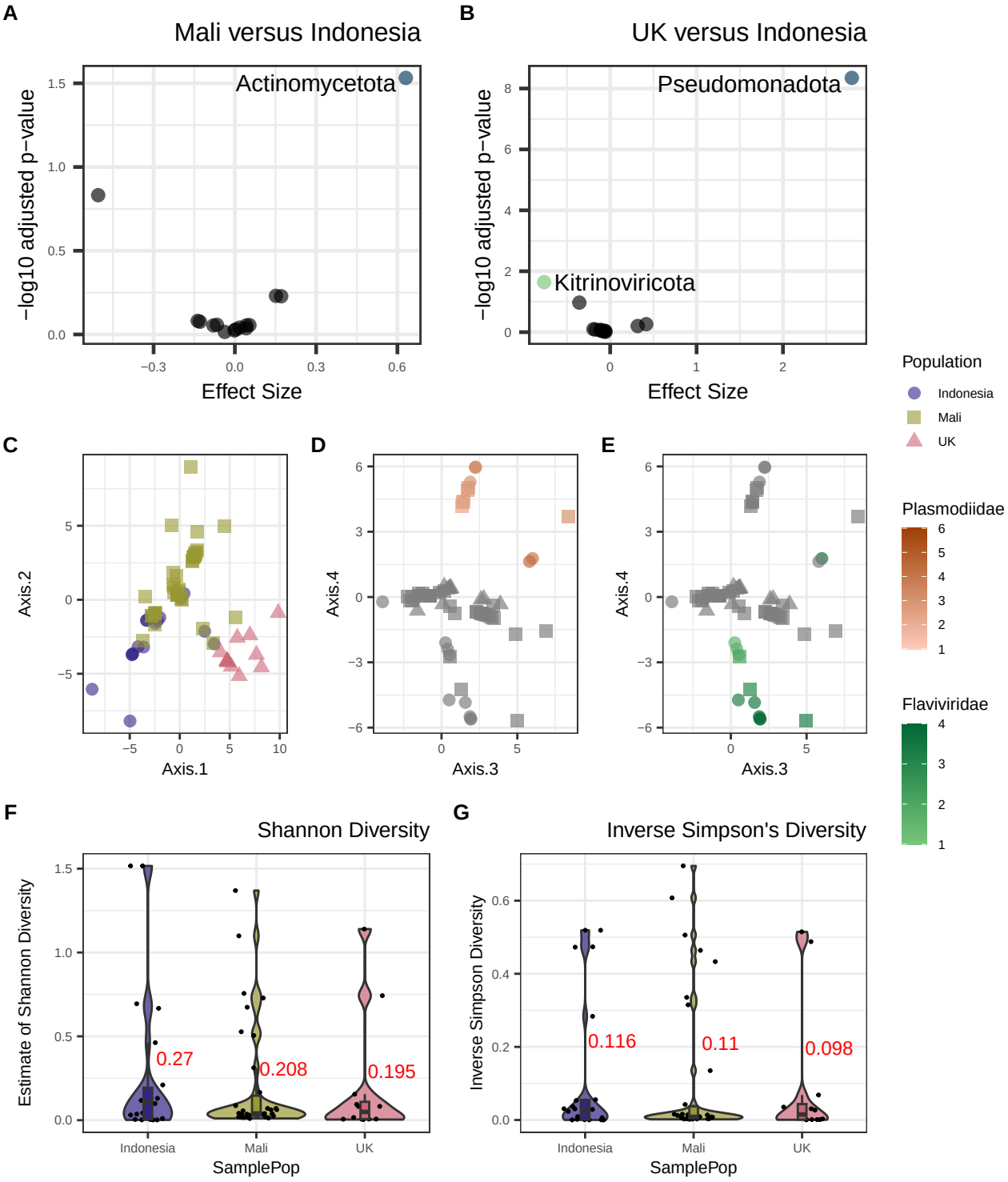


Fig 3: (caption on next page)

Fig 3: Taxa differences between Indonesian individuals and other global populations. A) Volcano plot of BH adjusted p-values from Welch's t-test for each phyla in Malian versus Indonesian individuals and B) UK versus Indonesian individuals. Taxa with a BH-corrected p-value below 0.05 for are coloured by superkingdom (blue: bacteria; green: viruses). C) Principal components (PCs) 1 and 2 of the CLR-normalised taxa abundance data at the phylum level, colored by population. D) PCs 3 and 4 of the same data colored by *Plasmodiidae* and E) *Flaviviridae* loads. F) Violin plots of Shannon diversity and G) inverse Simpson diversity for each population.

consequently the greatest power to detect rare taxa, and therefore these estimates likely reflect true rates of lower diversity within the UK population.

Discussion

Our understanding of pathogens found within remote regions of Indonesia, along with their impact on gene expression, is limited. Here, we have investigated to what extent microbial taxa can be detected within whole blood, and whether a core blood microbiome could be profiled. We did not detect taxa that constitute a core Indonesian whole blood microbiome, yet found evidence of strong pathogenic signals. This is consistent with recent findings of how the blood of healthy individuals do not support a consistent core microbial community [19]. We found evidence for the presence of eukaryotes and viruses, all of which have previously been characterised in blood transcriptomes [48]. This study supports a growing body of research suggesting that rather than being a sterile environment, a variety of taxa reside transiently within whole blood, and understanding their occurrence may facilitate better understanding of diseases and conditions in different populations.

Despite our attempts to remove contaminant human reads and tandem repeats prior to classification, we still identified significant numbers of reads mapping to *Homo sapiens* and had to filter out reads mapping to Metazoa. This reflects the issue raised by a recent study where Gihawi et al. reanalyzed a large-scale tumor microbiome study and found most bacterial reads to be misclassified human reads, largely due to the human genome being excluded from the classification reference database [49]. They also highlighted how the inclusion of draft bacterial genomes, which are often contaminated with human reads, contributed to the overestimation of bacterial species. Our results serve as another example of the importance of careful quality control measures to minimize false assignment of microbial reads, particularly by including the human genome and when possible using only complete microbial genomes during the classification process.

Despite not finding a core whole blood microbiome in the population, we identified two phyla that were dominant in multiple samples, namely Apicomplexa (driven nearly exclusively by the family Plasmodiidae)

and Kitrinoviricota (driven by the family Flaviviridae). From taxonomic profiling, we could attribute Kitrinoviricota viral signals to the family Flaviviridae, which is a family of viruses primarily found in mosquitos and ticks, and is responsible for multiple human illnesses including Dengue in Indonesia [50–52]. Around 3.6% of the reads could be further specified as belonging to the *Pegivirus* genus, yet we were unable to refine this assignment for the majority of the reads. The *Pegivirus* genus includes the human pegivirus (HPgV-1), a non-cytopathic lymphotropic virus previously associated with increased potential risk of lymphoma and reduction of disease progression caused by HIV-1 during co-infection [53]. For Apicomplexa, we could attribute 99.9% of reads to the family Plasmodiidae, of which *Plasmodium falciparum* and *Plasmodium vivax* are endemic throughout Indonesia [54].

Of all the Indonesian island populations in this study, we found that the Korowai had the highest abundance of the two pathogens. The Indonesian side of New Guinea Island is documented to have the highest rates of malaria in Indonesia, contributing up to 94% of all national cases [55–57], as well as the lowest number of healthcare facilities [58]; our results corroborate existing observations of a high endemic pathogen load within this region.

We also profiled and compared the blood microbiome of Malian and UK populations to the Indonesian samples. Bray-Curtis distance estimates showed that the Indonesian, Malian, and UK populations had high levels of dissimilarity from one another (Supplementary Figure 7). We also found differences in diversity between Indonesian and Malian populations (rural) compared to the UK (urban). Alpha diversity indices were higher in Malian and Indonesian populations, although the UK population had the highest read depth out of all three populations; diversity in the UK samples was driven primarily by bacterial taxa whereas the other two sites were characterised by widespread presence of pathogen-derived reads. Previous studies have reported similar findings when it comes to diversity between rural and urban populations: the Hadza, a small hunter gatherer group in Tanzania was found to have more diverse gut microbiomes than Italian urban controls [59]. Another study comparing gut microbiomes of rural and urban environments found that urban microbiomes were distinct, and that urbanisation led to a loss of certain bacterial taxa [60].

Our findings are limited by the fact that all three datasets we considered were generated by different groups in different places, where biological variations might be affected by differences during sampling and processing. Although our total sample sizes for the Indonesian samples are high—which is rare in studies of underrepresented populations, or more broadly, populations outside an urban, "western" environment—our total read depth is low, limiting the taxa we can detect in the population. Indeed, out of all three global populations, the Indonesian dataset had the lowest read depth (Supplementary Figure 5). However, in

opportunistic studies such as this, meeting the conditions required for high sequencing depth is rare; sequencing depth of unmapped reads is sensitive to multiple factors, including sequencing platform, sample collection and processing strategy, and only two publicly available datasets that we could find met the requirements needed to withstand total microbiome depletion.

Mounting evidence suggests that some microorganisms are common inhabitants of whole blood yet are likely originating from the gut and oral cavities [61, 62], as well as representing leakage from other parts of the body. We found stronger evidence of this in our analyses of the Malian and UK datasets. Taxa of the Actinomycetota phylum were found to be the most abundant in the Malian cohort, and around 84.6% of which could be further specified as *Corynebacterium tuberculostearicum*: a bacterium commonly found on human skin that is generally harmless, yet may play a role in skin health and disease [63, 64]. In the UK cohort Pseudomonadota was found to be the most abundant, up to 51.3% of which could be further specified as Enterobacteriaceae, a bacterial family that encompasses species commonly found in the human gut such as *E. coli* and linked to inflammatory bowel disease [65]. Additionally, up to 25.% of the Pseudomonadota reads were also defined as part of Xanthomonadaceae, a bacterial family that has been previously reported to colonize various hosts including the human skin [66, 67]. Interestingly, the water-borne Xanthomonadaceae has also been reported as contaminants in DNA extraction kits and reagents ("kitomes"), including in a study identifying the placental microbiome [68].This further challenges the notion of a core human blood microbiome; our findings reaffirm how the blood microbiome is comprised of transient microbiota originating from other body sites and/or from pathogenic infections, and how as a diagnostic medium how it may be hindered by limitations and variations of technical aspects.

A better understanding of which pathogens affect remote populations is crucial. Whole blood is one of the most abundant tissue types in RNA-seq analysis due to its relative ease of collection [69], and therefore its ability to provide information on environmental factors influencing disease phenotypes is ripe for investigation. In Indonesia, this is particularly important; Indonesia has a growing number of emerging infections [2, 21], however proper surveillance in rural areas remains limited. Our study demonstrates the use of whole blood RNA for microbiome-based diagnostic purposes that perhaps may be more suited in a retrospective context. Profiling microbiome from whole blood RNA may not be the most efficient approach as a first-line diagnostic tool due to the time-intensive process it requires. Nevertheless, this study provides valuable retrospective surveillance information on blood-borne microorganisms within the region, which is a valuable step in understanding and eventually limiting the spread of endemic and emerging diseases. Extra care should be taken to understand the influences on both environmental and technical factors while using

such approach for pathogen detection.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

KSB and IGR designed the study. KSB, IGR, and MF and wrote the manuscript with input from all authors. KSB performed all presented analyses. MF reperformed all presented analyses to improve robustness. CCD, PK, HS and SM generated raw data for all anaylses. CAF and PK assisted with validation of results.

Acknowledgements

We would like to acknowledge all of the study participants who generously consented to genome sequencing in the original study, as well as Emily R. Davenport, Murray P. Cox and members of the Gallego Romero group for helpful comments on the manuscript. St Vincent's Institute acknowledges the infrastructure support it receives from the National Health and Medical Research Council Independent Research Institutes Infrastructure Support Program and from the Victorian Government through its Operational Infrastructure Support Program. PK is supported by the Wellcome Trust International Training Fellowship (no. 222992/Z/21/Z).

References

1. Jennifer H. McQuiston, Joel M. Montgomery, and Christina L. Hutson. Ten Years of High-Consequence Pathogens—Research Gains, Readiness Gaps, and Future Goals - Volume 30, Number 4—April 2024 - Emerging Infectious Diseases journal - CDC. *Emerging Infectious Diseases*, 2024.

2. Richard J Coker, Benjamin M Hunter, James W Rudge, Marco Liverani, and Piya Hanvoravongchai. Emerging infectious diseases in southeast Asia: regional challenges to control. *The Lancet*, 377(9765):599–609, 2011.

3. GBD 2019 Antimicrobial Resistance Collaborators. Global mortality associated with 33 bacterial pathogens in 2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet*, 400(10369):2221–2248, December 2022. Publisher: Elsevier.

4. Buddha Basnyat. Typhoid versus typhus fever in post-earthquake Nepal. *The Lancet Global Health*, 4(8):e516–e517, August 2016. Publisher: Elsevier.

5. Kannan Venugopal, Franziska Hentzschel, Gediminas Valkiūnas, and Matthias Marti. *Plasmodium* asexual growth and sexual development in the haematopoietic niche of the host. *Nature Reviews Microbiology*, pages 1–13, 2020.

6. Byron E Martina, Luisa Barzon, Gorben P Pijlman, José de la Fuente, Annapaola Rizzoli, Linda J Wammes, Willem Takken, Ronald P van Rij, and Anna Papa. Human to human transmission of arthropod-borne pathogens. *Current Opinion in Virology*, 22:13–21, 2017.

7. Jack T Stapleton, Donna Klinzman, Warren N Schmidt, Michael A Pfaller, Ping Wu, Douglas R LaBrecque, Jian-qiu Han, Mary Jeanne Perino Phillips, Robert Woolson, and Beth Alden. Prospective comparison of whole-blood-and plasma-based hepatitis C virus RNA detection systems: improved detection using whole blood as the source of viral RNA. *Journal of Clinical Microbiology*, 37(3):484–489, 1999.

8. Céline Couturier, Atsuhiko Wada, Karen Louis, Maxime Mistretta, Benoit Beitz, Moriba Povogui, Maryline Ripaux, Charlotte Mignon, Bettina Werle, Adrien Lugari, et al. Characterization and analytical validation of a new antigenic rapid diagnostic test for ebola virus disease detection. *PLOS Neglected Tropical Diseases*, 14(1):e0007965, 2020.

9. Mark Kowarsky, Joan Camunas-Soler, Michael Kertesz, Iwijn De Vlaminck, Winston Koh, Wenying Pan, Lance Martin, Norma F Neff, Jennifer Okamoto, Ronald J Wong, et al. Numerous uncharacterized and highly divergent microbes which colonize humans are revealed by circulating cell-free DNA. *Proceedings of the National Academy of Sciences*, 114(36):9623–9628, 2017.

10. Emma Whittle, Martin O Leonard, Rebecca Harrison, Timothy W Gant, and Daniel Paul Tonge. Multi-method characterization of the human circulating microbiome. *Frontiers in Microbiology*, 9:3266, 2019.

11. Loes M Olde Loohuis, Serghei Mangul, Anil PS Ori, Guillaume Jospin, David Koslicki, Harry Taegyun Yang, Timothy Wu, Marco P Boks, Catherine Lomen-Hoerth, Martina Wiedau-Pazos, et al. Tran-

scriptome analysis in whole blood reveals increased microbial diversity in schizophrenia. *Translational Psychiatry*, 8(1):1–9, 2018.

12. Matthew H Stremlau, Kristian G Andersen, Onikepe A Folarin, Jessica N Grove, Ikponmwonsa Odia, Philomena E Ehiane, Omowunmi Omoniwa, Omigie Omoregie, Pan-Pan Jiang, Nathan L Yozwiak, et al. Discovery of novel rhabdoviruses in the blood of healthy individuals from West Africa. *PLOS Neglected Tropical Diseases*, 9(3):e0003631, 2015.

13. Rika A Furuta, Hirotaka Sakamoto, Ayumu Kuroishi, Kazuta Yasiui, Harumichi Matsukura, and Fumiya Hirayama. Metagenomic profiling of the viromes of plasma collected from blood donors with elevated serum alanine aminotransferase levels. *Transfusion*, 55(8):1889–1899, 2015.

14. Stefan Panaiotov, Georgi Filevski, Michele Equestre, Elena Nikolova, and Reni Kalfin. Cultural isolation and characteristics of the blood microbiome of healthy individuals. *Advances in Microbiology*, 8(5):406–421, 2018.

15. Vasudevan Dinakaran, Andiappan Rathinavel, Muthuirulan Pushpanathan, Ramamoorthy Sivakumar, Paramasamy Gunasekaran, and Jeyaprakash Rajendhran. Elevated levels of circulating DNA in cardiovascular disease patients: metagenomic profiling of microbiome in the circulation. *PLOS One*, 9(8):e105221, 2014.

16. LE Kafetzopoulou, ST Pullan, P Lemey, MA Suchard, DU Ehichioya, M Pahlmann, A Thielebein, J Hinzmann, L Oestereich, DM Wozniak, et al. Metagenomic sequencing at the epicenter of the Nigeria 2018 Lassa fever outbreak. *Science*, 363(6422):74–77, 2019.

17. Peter A Larsen, Corinne E Hayes, Cathy V Williams, Randall E Junge, Josia Razafindramanana, Vanessa Mass, Hajanirina Rakotondrainibe, and Anne D Yoder. Blood transcriptomes reveal novel parasitic zoonoses circulating in Madagascar's lemurs. *Biology Letters*, 12(1):20150829, 2016.

18. Spencer C Galen, Janus Borner, Jessie L Williamson, Christopher C Witt, and Susan L Perkins. Metatranscriptomics yields new genomic resources and sensitive detection of infections for diverse blood parasites. *Molecular Ecology Resources*, 20(1):14–28, 2020.

19. Cedric C. S. Tan, Karrie K. K. Ko, Hui Chen, Jianjun Liu, Marie Loh, Minghao Chia, and Niranjan Nagarajan. No evidence for a common blood microbiome based on a population study of 9,770 healthy humans. *Nature Microbiology*, pages 1–13, 2023. Publisher: Nature Publishing Group.

20. Hong Sheng Cheng, Sin Pei Tan, David Meng Kit Wong, Wei Ling Yolanda Koo, Sunny Hei Wong, and Nguan Soon Tan. The Blood Microbiome and Health: Current Evidence, Controversies, and Challenges. *International Journal of Molecular Sciences*, 24(6), March 2023. Publisher: Multidisciplinary Digital Publishing Institute (MDPI).

21. Wesley de Jong, Musofa Rusli, Soerajja Bhoelan, Sofie Rohde, Fedik A Rantam, Purwati A Noeryoto, Usman Hadi, Eric CM van Gorp, and Marco Goeijenbier. Endemic and emerging acute virus infections in Indonesia: an overview of the past decade and implications for the future. *Critical Reviews in Microbiology*, 44(4):487–503, 2018.

22. Yodi Mahendradhata, Laksono Trisnantoro, Shita Listyadewi, Prastuti Soewondo, Tiara Marthias, Pandu Harimurti, and John Prawira. The Republic of Indonesia health system review. Technical report, WHO Regional Office for South-East Asia, 2017.

23. Rina Agustina, Teguh Dartanto, Ratna Sitompul, Kun A Susiloretni, Endang L Achadi, Akmal Taher, Fadila Wirawan, Saleha Sungkar, Pratiwi Sudarmono, Anuraj H Shankar, et al. Universal health coverage in Indonesia: concept, progress, and challenges. *The Lancet*, 393(10166):75–102, 2019.

24. World Health Organization et al. WHO country cooperation strategy 2014–2019: Indonesia. Technical report, World Health Organization. Regional Office for South-East Asia, 2016.

25. Heini M. Natri, Katalina S. Bobowik, Pradiptajati Kusuma, Chelzie Crenna Darusallam, Guy S. Jacobs, Georgi Hudjashov, J. Stephen Lansing, Herawati Sudoyo, Nicholas E. Banovich, Murray P. Cox, and Irene Gallego Romero. Genome-wide DNA methylation and gene expression patterns reflect genetic ancestry and environmental differences across the indonesian archipelago. *PLOS Genetics*, 16(5):e1008749, 2020. Publisher: Public Library of Science.

26. Heini M. Natri, Georgi Hudjashov, Guy Jacobs, Pradiptajati Kusuma, Lauri Saag, Chelzie Crenna Darusallam, Mait Metspalu, Herawati Sudoyo, Murray P. Cox, Irene Gallego Romero, and Nicholas E. Banovich. Genetic architecture of gene regulation in Indonesian populations identifies QTLs associated with global and local ancestries. *The American Journal of Human Genetics*, 109(1):50–65, 2022.

27. Katalina S Bobowik, Din Syafruddin, Chelzie Crenna Darusallam, Herawati Sudoyo, Christine A Wells, and Irene Gallego Romero. Transcriptomic profiles of *Plasmodium falciparum* and *Plasmodium vivax*-infected individuals in Indonesia. *bioRxiv*, pages 2021–01, 2021.

28. Tuan M Tran, Marcus B Jones, Aissata Ongoiba, Else M Bijker, Remko Schats, Pratap Venepally, Jeff Skinner, Safiatou Doumbo, Edwin Quinten, Leo G Visser, et al. Transcriptomic evidence for modulation of host inflammatory responses during febrile *Plasmodium falciparum* malaria. *Scientific Reports*, 6:31291, 2016.

29. Tuan M Tran, Rajan Guha, Silvia Portugal, Jeff Skinner, Aissata Ongoiba, Jyoti Bhardwaj, Marcus Jones, Jacqueline Moebius, Pratap Venepally, Safiatou Doumbo, et al. A molecular signature in blood reveals a role for p53 in regulating malaria-induced inflammation. *Immunity*, 51(4):750–765, 2019.

30. Akul Singhania, Raman Verma, Christine M Graham, Jo Lee, Trang Tran, Matthew Richardson, Patrick Lecine, Philippe Leissner, Matthew PR Berry, Robert J Wilkinson, et al. A modular transcriptional signature identifies phenotypic heterogeneity of human tuberculosis infection. *Nature Communications*, 9(1):1–17, 2018.

31. Simon Andrews, Felix Krueger, Anne Segonds-Pichon, Laura Biggins, Christel Krueger, and Steven Wingett. FastQC. Babraham Institute, January 2012.

32. Anthony M Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014.

33. Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. Star: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

34. Kirill Rotmistrovsky and Richa Agarwala. BMTagger: Best Match Tagger for removing human reads from metagenomics datasets. *Unpublished*, 2011.

35. Gary Benson. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research*, 27(2):573–580, 1999.

36. Philip TLC Clausen, Frank M Aarestrup, and Ole Lund. Rapid and precise alignment of raw reads against redundant databases with KMA. *BMC Bioinformatics*, 19(1):1–8, 2018.

37. Vanessa Rossetto Marcelino, Jan Buchmann, and Philip Clausen. Indexed reference databases for KMA and CCMetagen, 2019.

38. Vanessa R Marcelino, Philip TLC Clausen, Jan P Buchmann, Michelle Wille, Jonathan R Iredell, Wieland Meyer, Ole Lund, Tania C Sorrell, and Edward C Holmes. CCMetagen: comprehensive and accurate identification of eukaryotes and prokaryotes in metagenomic data. *Genome Biology*, 21(1):1–15, 2020.

39. Gregory B Gloor, Jean M Macklaim, Vera Pawlowsky-Glahn, and Juan J Egozcue. Microbiome datasets are compositional: and this is not optional. *Frontiers in Microbiology*, 8:2224, 2017.

40. John Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160, 1982.

41. David R Lovell, Xin-Yi Chua, and Annette McGrath. Counts: an outstanding challenge for log-ratio analysis of compositional data in the molecular biosciences. *NAR Genomics and Bioinformatics*, 2(2):lqaa040, 2020.

42. Andrew D Fernandes, JM Macklaim, TG Linn, G Reid, and GB Gloor. ANOVA-like differential gene expression analysis of single-organism and meta-RNA-seq. *PLOS one*, 8(7):e67019, 2013.

43. Andrew D Fernandes, Jennifer Ns Reid, Jean M Macklaim, Thomas A McMurrough, David R Edgell, and Gregory B Gloor. Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome*, 2(1):15, 2014.

44. Gregory B Gloor, Jean M Macklaim, and Andrew D Fernandes. Displaying variation in large datasets: plotting a visual summary of effect sizes. *Journal of Computational and Graphical Statistics*, 25(3):971–979, 2016.

45. Amy D Willis and Bryan D Martin. Estimating diversity in networked ecological communities. *Biostatistics*, 2020.

46. Mark L Heiman and Frank L Greenway. A healthy gastrointestinal microbiome is dependent on dietary diversity. *Molecular metabolism*, 5(5):317–320, 2016.

47. Safiatou Doumbo, Tuan M Tran, Jules Sangala, Shanping Li, Didier Doumtabe, Younoussou Kone, Abdrahamane Traore, Aboudramane Bathily, Nafomon Sogoba, Michel E Coulibaly, et al. Co-infection of long-term carriers of *Plasmodium falciparum* with *Schistosoma haematobium* enhances

protection from febrile malaria: a prospective cohort study in Mali. *PLOS Neglected Tropical Diseases*, 8(9):e3154, 2014.

48. Diego J Castillo, Riaan F Rifkin, Don A Cowan, and Marnie Potgieter. The healthy human blood microbiome: fact or fiction? *Frontiers in Cellular and Infection Microbiology*, 9:148, 2019.

49. Abraham Gihawi, Yuchen Ge, Jennifer Lu, Daniela Puiu, Amanda Xu, Colin S. Cooper, Daniel S. Brewer, Mihaela Pertea, and Steven L. Salzberg. Major data analysis errors invalidate cancer microbiome findings. *bioRxiv*, 2023. Pages: 2023.07.28.550993 Section: Contradictory Results.

50. Camilo Guzmán, Alfonso Calderón, Salim Mattar, Luiz Tadeu-Figuereido, Jorge Salazar-Bravo, Nelson Alvis-Guzmán, Elias Zakzuk Martinez, and Marco González. Ecoepidemiology of alphaviruses and flaviviruses. In Moulay Mustapha Ennaji, editor, *Emerging and Reemerging Viral Pathogens*, pages 101–125. Academic Press, 2020.

51. Alton B. Farris, Martin K. Selig, and G. Petur Nielsen. Ultrastructural diagnosis of infection. In Richard L. Kradin, editor, *Diagnostic Pathology of Infectious Disease*, pages 77–98. W.B. Saunders, New York, 2010.

52. World Health Organization. Disease Outbreak News; Dengue – Global situation, December 2023.

53. Yaqi Yu, Zhenzhou Wan, Jian-Hua Wang, Xianguang Yang, and Chiyu Zhang. Review of human pegivirus: Prevalence, transmission, pathogenesis, and clinical implication. *Virulence*, 13(1):324–341, 2022.

54. Claudia Surjadjaja, Asik Surya, and J Kevin Baird. Epidemiology of *Plasmodium vivax* in Indonesia. *The American Journal of Tropical Medicine and Hygiene*, 95(6_Suppl):121–132, 2016.

55. Wulung Hanandita and Gindo Tampubolon. Geography and social distribution of malaria in Indonesian Papua: a cross-sectional study. *International Journal of Health Geographics*, 15(1):13, 2016.

56. World Health Organization. World malaria report 2022. Technical report, World Health Organization, Geneva, 2022.

57. Directorate General for Disease Prevention and Control. Annual Malaria Report 2022. Technical report, Ministry of Health RI, 2023.

58. World Health Organization et al. *State of health inequality: Indonesia*. World Health Organization, 2017.

59. Stephanie L Schnorr, Marco Candela, Simone Rampelli, Manuela Centanni, Clarissa Consolandi, Giulia Basaglia, Silvia Turroni, Elena Biagi, Clelia Peano, Marco Severgnini, et al. Gut microbiome of the Hadza hunter-gatherers. *Nature Communications*, 5(1):1–12, 2014.

60. Funmilola A Ayeni, Elena Biagi, Simone Rampelli, Jessica Fiori, Matteo Soverini, Haruna J Audu, Sandra Cristino, Leonardo Caporali, Stephanie L Schnorr, Valerio Carelli, et al. Infant and adult gut microbiome and metabolome in rural Bassa and urban settlers from Nigeria. *Cell Reports*, 23(10):3056–3067, 2018.

61. Marnie Potgieter, Janette Bester, Douglas B Kell, and Etheresia Pretorius. The dormant blood microbiome in chronic, inflammatory diseases. *FEMS Microbiology Reviews*, 39(4):567–591, 2015.

62. Sandrine Païssé, Carine Valle, Florence Servant, Michael Courtney, Rémy Burcelin, Jacques Amar, and Benjamin Lelouvier. Comprehensive description of blood microbiome from healthy donors assessed by 16S targeted metagenomic sequencing. *Transfusion*, 56(5):1138–1147, 2016.

63. V. Hinić, C. Lang, M. Weisser, C. Straub, R. Frei, and D. Goldenberger. Corynebacterium tuberculostearicum: a potentially misidentified and multiresistant corynebacterium species isolated from clinical specimens. *Journal of Clinical Microbiology*, 50(8):2561–2567, 2012.

64. Mohammed O. Altonsy, Habib A. Kurwa, Gilles J. Lauzon, Matthias Amrein, Anthony N. Gerber, Wagdi Almishri, and Paule Régine Mydlarski. Corynebacterium tuberculostearicum, a human skin colonizer, induces the canonical nuclear factor-κb inflammatory signaling pathway in human skin cells. *Immunity, Inflammation and Disease*, 8(1):62–79, 2020. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/iid3.284.

65. Valerio Baldelli, Franco Scaldaferri, Lorenza Putignani, and Federica Del Chierico. The role of enterobacteriaceae in gut microbiota dysbiosis in inflammatory bowel diseases. *Microorganisms*, 9(4):697, 2021.

66. Krzysztof Skowron, Justyna Bauza-Kaszewska, Zuzanna Kraszewska, Natalia Wiktorczyk-Kapischke, Katarzyna Grudlewska-Buda, Joanna Kwiecińska-Piróg, Ewa Wałecka-Zacharska, Laura Radtke, and Eugenia Gospodarek-Komkowska. Human skin microbiome: Impact of intrinsic and extrinsic factors on skin microbiota. *Microorganisms*, 9(3):543, 2021.

67. Bo Zeng, Jiangchao Zhao, Wei Guo, Siyuan Zhang, Yutong Hua, Jingsi Tang, Fanli Kong, Xuewu Yang, Lizhi Fu, Kun Liao, Xianqiong Yu, Guohong Chen, Long Jin, Surong Shuai, Jiandong Yang, Xiaohui Si, Ruihong Ning, Sudhanshu Mishra, and Ying Li. High-altitude living shapes the skin microbiome in humans and pigs. *Frontiers in Microbiology*, 8, 2017. Publisher: Frontiers.

68. Isoken Nicholas Olomu, Luis Carlos Pena-Cortes, Robert A. Long, Arpita Vyas, Olha Krichevskiy, Ryan Luellwitz, Pallavi Singh, and Martha H. Mulks. Elimination of "kitome" and "splashome" contamination results in lack of detection of a unique placental microbiome. *BMC Microbiology*, 20(1):157, 2020.

69. Duncan E Donohue, Aarti Gautam, Stacy-Ann Miller, Seshamalini Srinivasan, Duna Abu-Amara, Ross Campbell, Charles R Marmar, Rasha Hammamieh, and Marti Jett. Gene expression profiling of whole blood: a comparative assessment of RNA-stabilizing collection methods. *PloS ONE*, 14(10):e0223065, 2019.

Supplementary materials

Supplementary Figure 1 Summary of reads mapping to filtered taxa for the Indonesian (101BP and trimmed 75BP), Malian (75BP), and UK (75BP) populations. A-D) Reads mapping to the Viridiplantae E-H) and Metazoa.

Supplementary Figure 2 Read depth per individual library across all filtering steps.

Supplementary Figure 3 Alpha diversity estimates for Indonesian island populations. A) Estimates of Shannon and B) inverse Simpson diversity within each population (median in blue text). KOR = Korowai; MTW = Mentawai; SMB = Sumba

Supplementary Figure 4 Relative abundance of the top 20 taxa within the Indonesian, Malian, and UK dataset at the superkingdom, phylum, and family level. Bacteria are shown in blue, eukaryotes in red, and viruses in green.

Supplementary Figure 5 Rarefaction curves of species saturation per individual at varying read depths for the Indonesian, Malian, and UK populations.

Supplementary Figure 6 Taxa differences between samples from Korowai and other global populations. A) Volcano plot of BH adjusted p-values from Welch's t-test for each phyla in the Korowai versus Malian populations and B) Korowai versus UK populations. Taxa with a BH-corrected p-value below 0.05 for are coloured by superkingdom (blue: bacteria).

Supplementary Figure 7 Bray-Curtis distance estimates for Indonesian, Malian, and UK population comparisons at the phylum level (mean in red text).

Supplementary Table 1 Read depth of each individual in the Indonesian dataset (101BP) after each filtering step.

Supplementary Table 2 Read depth of each individual in the Indonesian dataset (75BP) after each filtering step.

Supplementary Table 3 Read depth of each individual in the Malian dataset (75BP) after each filtering step.

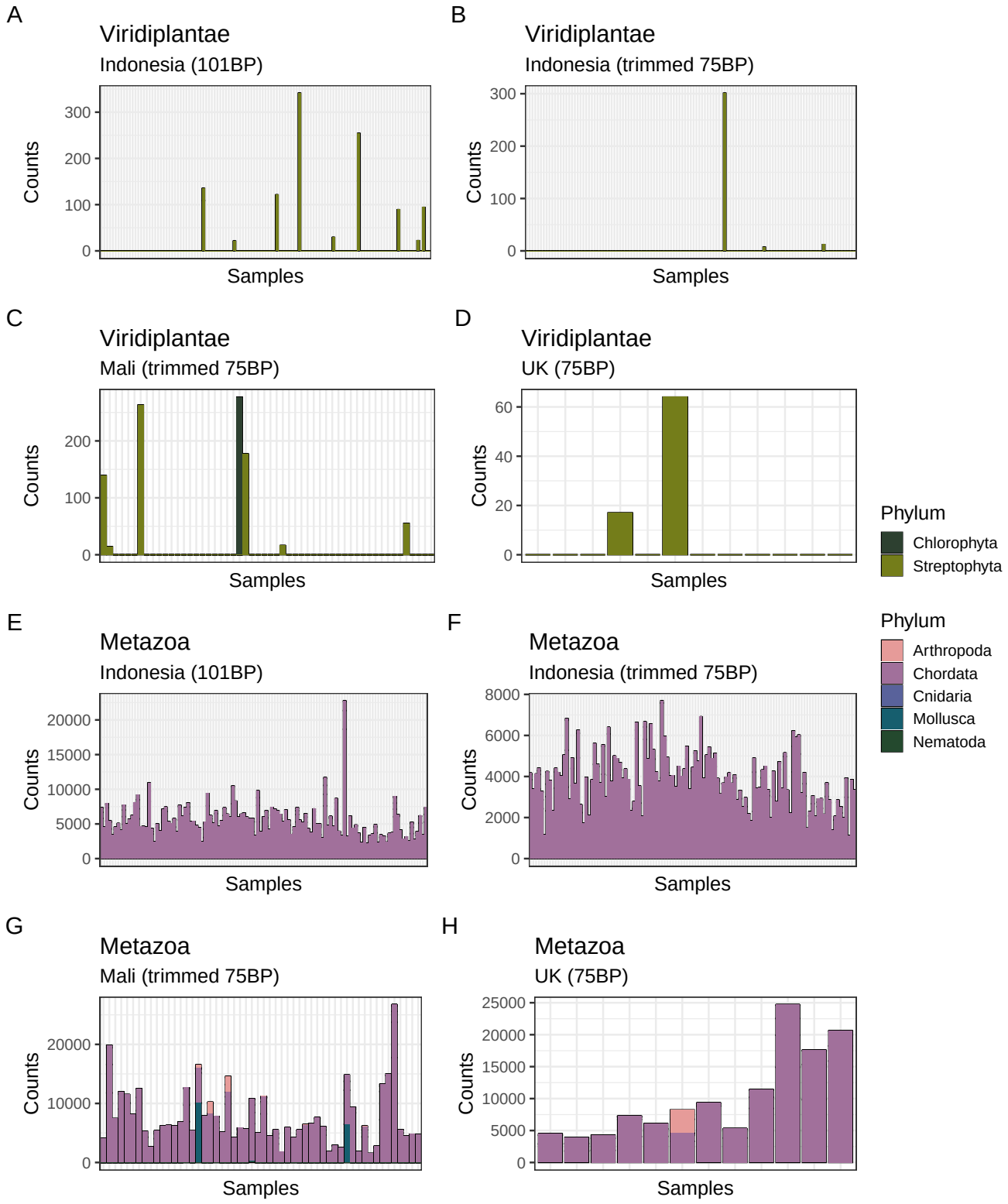
Supplementary Table 4 Read depth of each individual in the UK dataset (75BP) after each filtering step.

Supplementary Table 5 Differential abundance analysis results (Welch's t-test BH-adjusted p = 0.05) at the phylum level between Malian and Indonesian datasets.

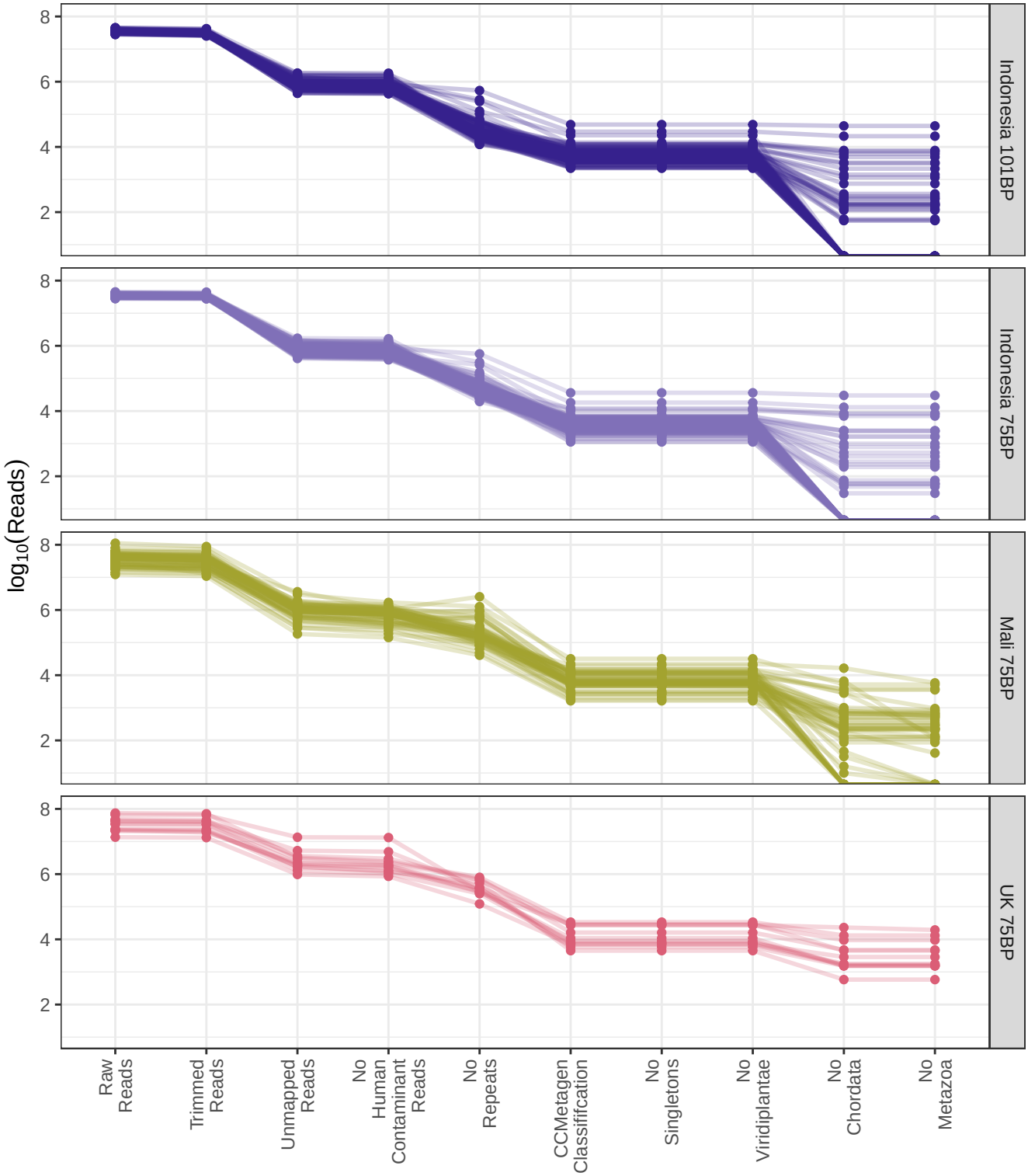
Supplementary Table 6 Differential abundance analysis results (Welch's t-test BH-adjusted p = 0.05) at the phylum level between UK and Indonesian datasets.

Supplementary Table 7 Differential abundance analysis results (Welch's t-test BH-adjusted p = 0.05) at the phylum level between Malian and Korowai samples.

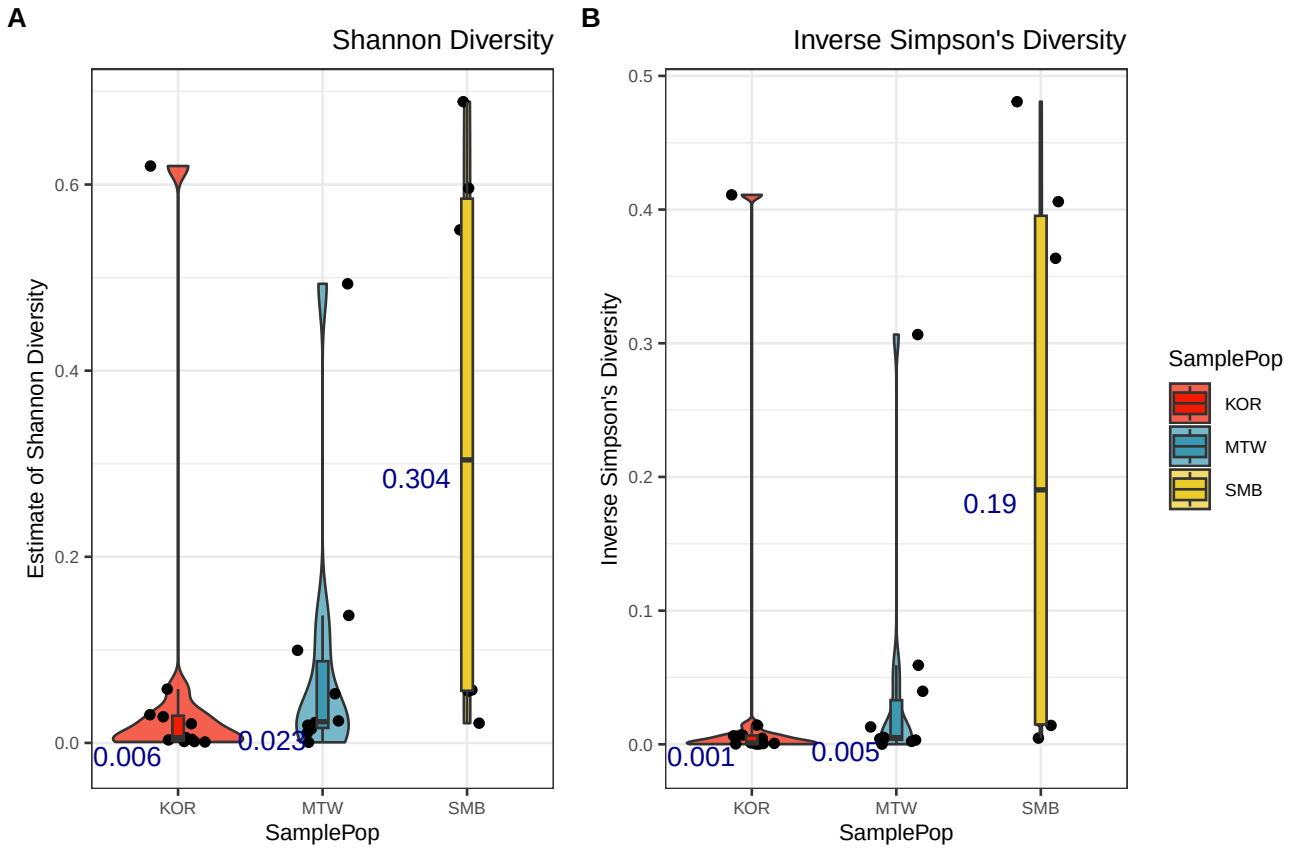
Supplementary Table 8 Differential abundance analysis results (Welch's t-test BH-adjusted p = 0.05) at the phylum level between UK and Korowai samples.



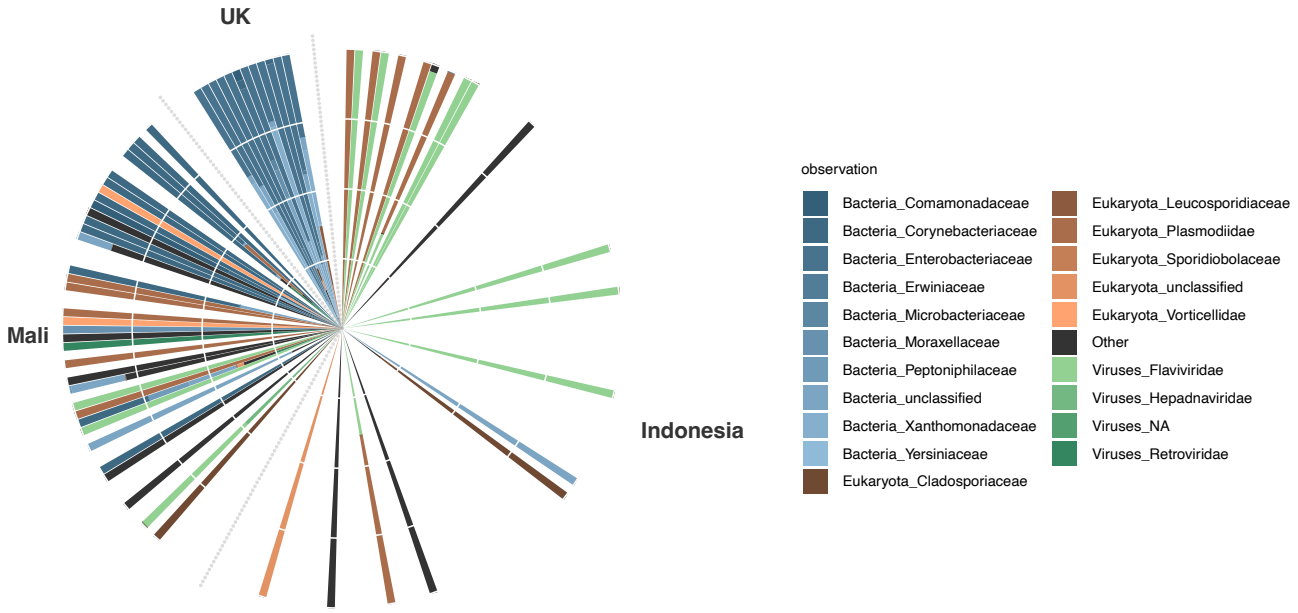
Supplementary Figure 1: Summary of reads mapping to filtered taxa for the Indonesian (101BP and trimmed 75BP), Malian (75BP), and UK (75BP) populations. A-D) Reads mapping to the Viridiplantae E-H) and Metazoa.



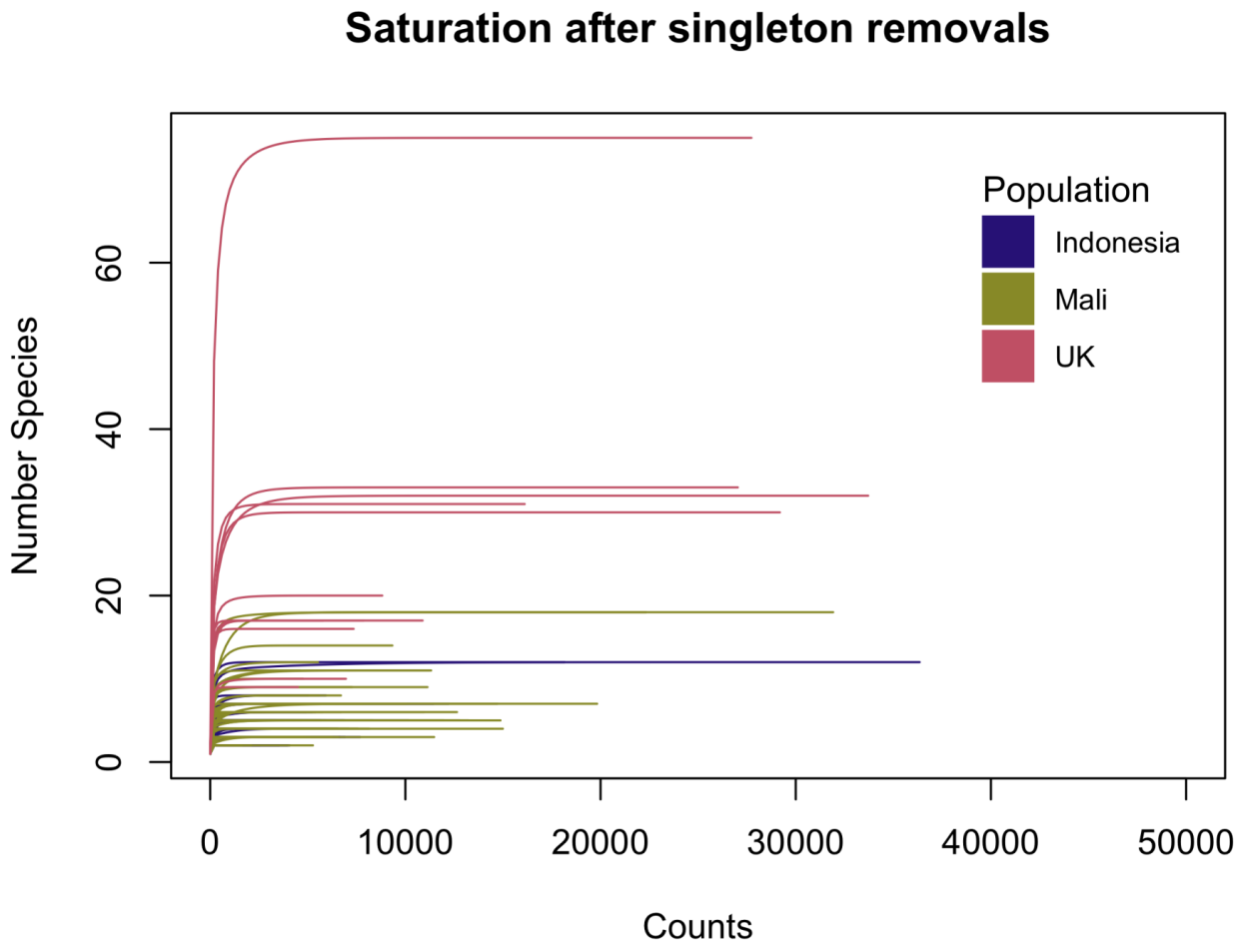
Supplementary Figure 2: Read depth per individual library across all filtering steps.



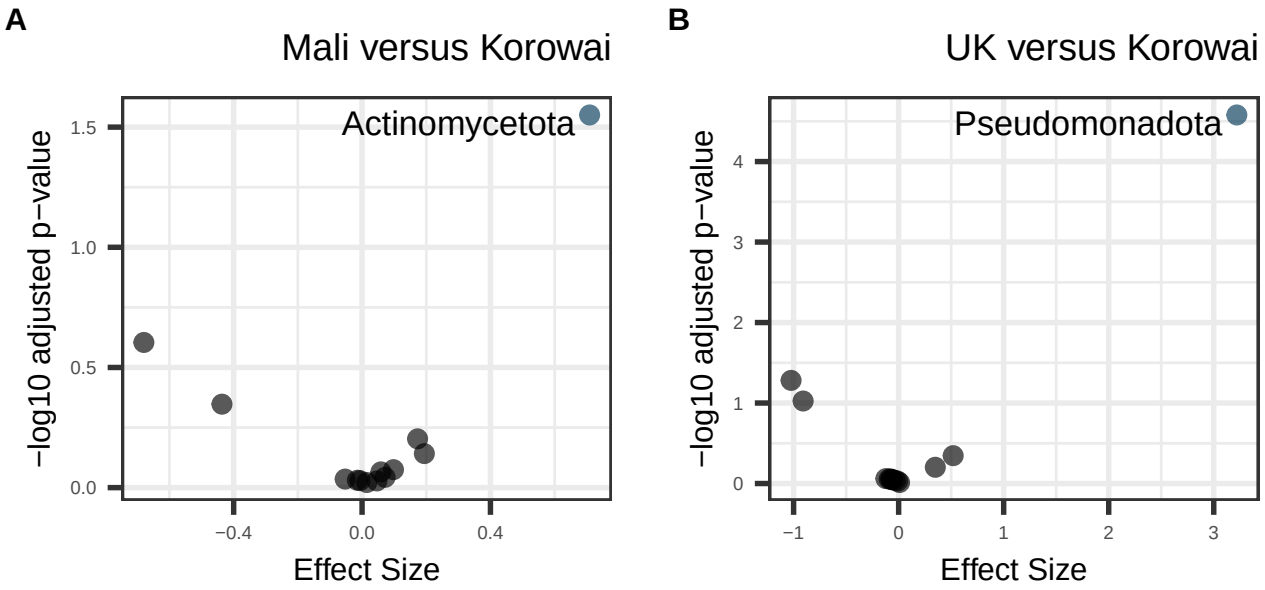
Supplementary Figure 3: Alpha diversity estimates for Indonesian island populations. A) Estimates of Shannon and B) inverse Simpson diversity within each population (median in blue text). KOR = Korowai; MTW = Mentawai; SMB = Sumba



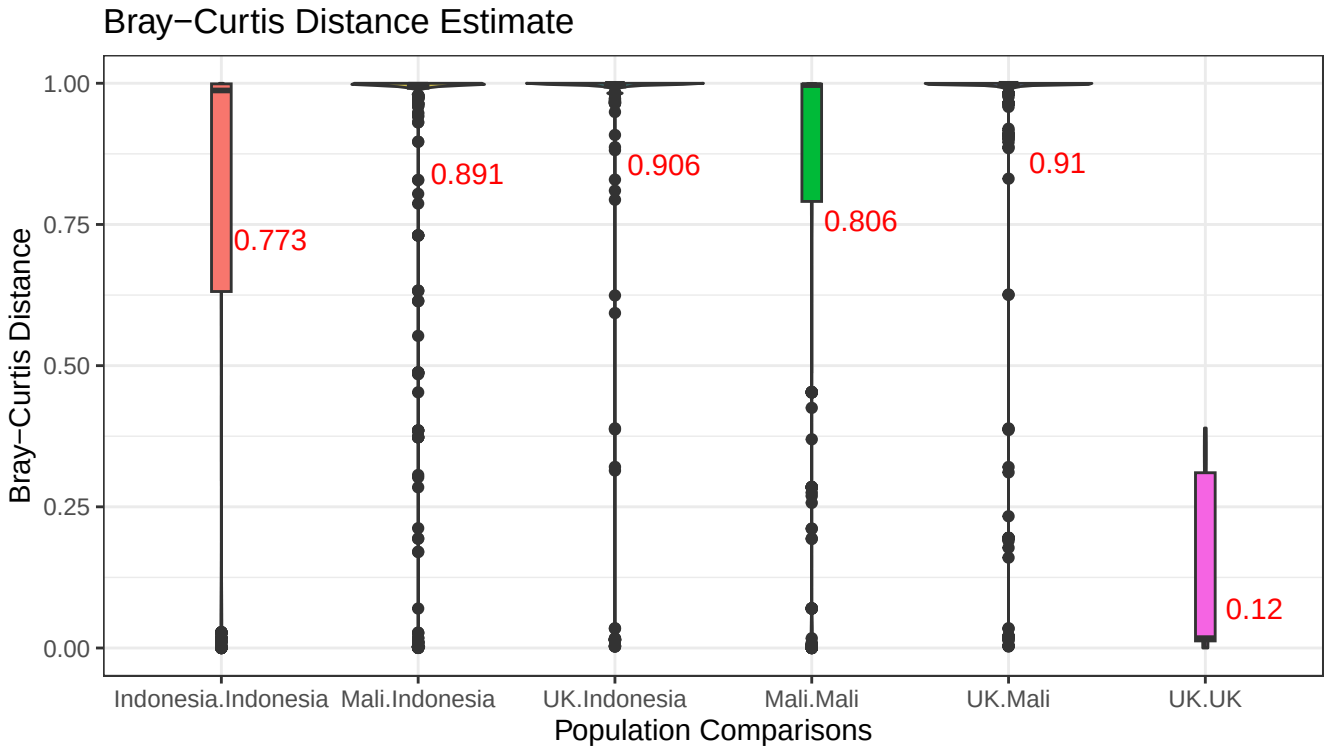
Supplementary Figure 4: Relative abundance of the top 20 taxa within the Indonesian, Malian, and UK dataset at the superkingdom, phylum, and family level. Bacteria are shown in blue, eukaryotes in red, and viruses in green.



Supplementary Figure 5: Rarefaction curves of species saturation per individual at varying read depths for the Indonesian, Malian, and UK populations.



Supplementary Figure 6: Taxa differences between samples from Korowai and other global populations. A) Volcano plot of BH adjusted p-values from Welch's t-test for each phyla in the Korowai versus Malian populations and B) Korowai versus UK populations. Taxa with a BH-corrected p-value below 0.05 for are coloured by superkingdom (blue: bacteria).



Supplementary Figure 7: Bray-Curtis distance estimates for Indonesian, Malian, and UK population comparisons at the phylum level (mean in red text).