The whole blood microbiome of Indonesians reveals that environmental differences shape immune gene expression signatures

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

1 Melbourne Integrative Genomics, University of Melbourne, Royal Parade, 3010, Parkville, Victoria, Australia

2 School of BioSciences, The University of Melbourne, Royal Parade, 3010, Parkville, Australia

3 The Centre for Stem Cell Systems, Faculty of Medicine, Dentistry and Health Sciences, The University of Melbourne, 30 Royal Parade Parkville, Victoria 3010, Australia

4 Genome Diversity and Diseases Laboratory, Eijkman Institute for Molecular Biology, Jakarta, Indonesia

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

6 Present address: Genome Diversity and Diseases Division, Mochtar Riady Institute for Nanotechnology, Tangerang, Banten, Indonesia

* 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 within the country. One way to strengthen the control of infectious diseases is through better surveillance, however unequal investment in medical funding throughout Indonesia, particularly in rural areas, has resulted in under-reporting of cases. 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 are able to detect a broad range of taxa within RNA-sequencing data generated from whole blood, including bacteria, viruses, archaea, and eukaryotes. Using independent component analysis, we find that two of these pathogens—Flaviviridae and Plasmodium—have

the most noticeable effects on expression profiles. We also identify specific genes linked with Plasmodium and Flavivirus abundance and find that both of these infections are most pronounced in the easternmost island within our Indonesian dataset. This study provides a framework for novel applications of RNA-seq as surveillance and a better understanding of environmental contributors affecting gene expression within Indonesia.

Introduction

Pathogens are a major cause of morbidity and mortality, especially in the Global South. 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. Having a more detailed understanding of which pathogens are the major causes of morbidity across different global populations and how they affect host responses to disease can focus elimination efforts on specific pathogens and aid in more targeted disease therapeutics.

Using blood transcriptome data serves as a way 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 [1, 2] and various viruses [3, 4], emerging research has shown that even bacteria and fungi can release DNA and RNA into blood [5]. For example, commensal bacteria [6, 7], viruses [8, 9], fungi [10], and archaea [11] have all been identified independently in multiple studies of human blood, suggesting blood has its own unique microbiome. For health diagnostics, this has been exploited to identify relationships between the blood microbiome and celiac disease [12] and to explore connections between the blood microbiome and brain disorders [7]. While not yet common, the use of blood as a surveillance tool is growing. For instance, Kafetzopoulou et al. [13] 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 [14] and songbirds [15] to aid in the characterisation of diverse blood parasites.

Indonesia is a country with large numbers of endemic and emerging infectious diseases [16], 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 [17], pathogen abundance can still be high in more rural areas, which tend to have less access to medical resources [17–19]. 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 [20]. While some of this variation is likely attributable to the different genetic ancestries of individuals in these islands [20, 21], another contributor may be differences in pathogenic loads between them. Indeed, both *Plasmodium falciparum* and *Plasmodium vivax* are detectable within whole blood of these individuals [22], 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 and investigate the relationship between gene expression in whole blood and pathogen abundance, 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, as well as show the relationship between local environments and immune gene expression.

Methods

Samples

The Indonesian dataset consists of 100 base-pair, paired-end data from whole blood collected by members of the Eijkman Institute from 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 [20]; 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, 6 samples were sequenced twice as technical replicates, however for our study we only retained the replicate with the highest read depth. Samples 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 Table 1).

In order to compare our samples to other global populations, we searched 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 criteria as in our own samples. We identified two publicly-available datasets as controls. The first dataset comes from Tran et al. [23, 24], and consists of 100-bp human whole blood RNA-seq data, hereafter referred to as the Mali study. As described in [24], 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. [25] 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

In order 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 [26]. In order 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 [27]. 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.

RNA-seq reads were first aligned to the human genome (GrCh38, Ensembl release 90: August 2017) with STAR v. 2.5.3a [28] using a two-pass alignment 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 sequence reads were then processed using KneadData v. 0.7.4, which uses BMTagger [29] and Tandem Repeats Finder (TRF) [30] 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 100,000 reads per sample for the Indonesian dataset (both 75 and 100-bp; Supplementary Table 1). For the 75-bp Malian and UK datasets, this resulted in a mean of 330,000 and 3,000,000 reads per sample, respectively (Supplementary Table 1). Read depths after each filtering step are available in Supplementary Table 1.

Mapping and metagenomic classification

Processed metagenomic reads were mapped using KMA v. 1.2.21 [31] against a filtered NCBI nt reference database, where artificial sequences and environmental sequences without valid taxonomic IDs were excluded [32] (downloaded from https://researchdata.edu.au/indexed-reference-databases-kma-ccmetagen/1371207). We mapped reads using default settings and the following additional flags as recommended on the CCMetagen page: -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 false so that matches could be made against sequences that were not significantly over-represented. We attempted read mapping using both paired and single-end configurations. Single-end mapping resulted in a much larger proportion of successfully mapped reads than paired-end. Upon investigation, we found that, although pairwise correlations between reads from the same mate pair were higher than between reads outside of a mate

pair (Supplementary Figure 1, A), read pairs had over 35% dissimilarity, on average, at the species level (Supplementary Figure 1, B), resulting in an excess of unmapped reads. We therefore decided to perform mapping on single-ended reads, using the forward strand only from each dataset. After mapping, we performed read classification using CCMetagen v. 1.2.2 [33] with default settings for single-ended 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 24,000 reads per sample, which increased to 44,000 when we trimmed reads to 75-bp (Supplementary Table 1). For the 75-bp Malian and UK datasets, this resulted in a mean of 52,000 and 2,200,000 reads, respectively (Supplementary Table 1).

Data filtering

After removing singletons to prevent spurious identification of taxa, we observed a large proportion of the remaining reads mapped to the kingdoms Viridiplantae, which contains green algaea and plants, and Metazoa. As we are interested in pathogenic microorganisms, we decided to remove the entire kingdom of Viridiplantae, reasoning that these likely represented misassignments or poor quality annotation (Supplementary Figure 2, A-C) and further investigated the metazoan reads. We found that the majority of these mapped to the phylum Chordata (Supplementary Figure 2, D-F) although some potentially pathogenic taxa, such as helminths (Platyhelminthes, Nematoda), were present in the data. Upon further investigation, we found that every individual, including samples within the UK dataset where we would not expect to observe widespread helminth infection, had reads mapping to helminth species (Supplementary Figure 2, G-I). BLAST analysis of helminth reads also confirmed that these were reads that mapped equally well to the human genome. We therefore decided to discard all reads mapping to Metazoa from subsequent analysis. 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 8,120 reads in the Indonesian dataset (a mean of 8,466 for the 75-bp Indonesian reads; Supplementary Table 1), 20,096 for the 75-bp Malian dataset, and 966,195 for the 75-bp UK dataset (Supplementary Table 1; Supplementary Figure 3).

Sample clustering

To correct for uneven library depth between samples and the compositional nature of microbiome data [34], we applied a center log ratio (CLR) transformation [35] 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 [36], 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) [37–39] to test for differences in species composition between populations, which applies CLR-transformation to correct for uneven library depth and data compositionality [38]. 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 [40], which expects the presence of singletons in order to model species richness [40].

Independent component analysis of expression data

To estimate source signals within expression data, we applied independent components analysis (ICA) to the whole blood expression data using the Bioconductor package MineICA v. 1.26.0 [41]. To compute the independent components (ICs), we applied the default JADE algorithm [42] to the data using 5 ICs. As suggested by Jutten et al. [43], we chose the number of ICs to compute based on the amount of variance explained by principal component analysis (PCA). We found that the first 5 components captured the most variance within the data (18% of variance in the first component, down to 3.5% of variance in the fifth component), with all subsequent components contributing only a small amount of variance (less than 3% per component; Supplementary Figure 4). To test whether populations and pathogen abundances were differentially distributed on the components, we respectively performed Kruskal-Wallis and Pearson correlation tests using MineICA and corrected for multiple testing using the Benjamini-Hochberg method [44]. To draw correlations between sample contributions and pathogen load, we used the CLR-transformed pathogen abundance matrix at the phylum level. Finally, in order to test for enrichment of contributing genes within each IC against Gene Ontology (GO) [45] and Kyoto Encyclopedia of Genes and Genomes (KEGG) [46] pathways, we used GOseq [47], which corrects for gene length bias.

Code for all analyses is available at https://gitlab.unimelb.edu.au/igr-lab/Epi_Study

Results

The blood microbiome of Indonesians

In order 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 in the Indonesian side of New Guinea Island. The human samples have been extensively described [20, 21]. After extensive quality control, we obtained a mean library size of 8,146 taxonomically informative reads (range: 221 - 403,796), which was further reduced to 8,120 reads after the removal of singletons (range: 196 - 403,771; Supplementary Table 1). We assigned these reads to a total of 1,390 taxa across all phylogenetic levels, including 271 distinct taxa at the family level. We found these reads were predominantly assigned to the families Plasmodiidae (86.4% of the total read pool across all individuals) and Flaviviridae (4.0% of reads), and to various species of bacteria, the most abundant being Enterobacteriaceae (2.9%; Figure 1, A). In order to control for sparsity in the abundance matrix, which is crucial when performing CLR-transformation [36], 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 33 taxa, with Apicomplexa (85.9% of reads, within which 99.9% of reads mapped to the family Plasmodiidae), Proteobacteria (5.8% of reads), Kitrinoviricota (4.0% of reads, within which 100% of reads mapped to Flaviviridae), Actinobacteria (1.9% of reads), and Firmicutes (0.8% of reads) making up the majority. These estimates of Apicomplexa load are higher than our previous estimates of Plasmodium burden [22], where we used a different, more conservative approach. We observed that the microbiome composition varied substantially between islands. This was most pronounced in the Korowai population, where the majority of samples had reads assigned to either Apicomplexa (65% of reads) or Kitrinoviricota (30% of reads).

PCA of the CLR-transformed taxonomic matrix showed sample clustering clearly driven by the phyla Apicomplexa and Kitrinoviricota (Figure 1, B and C). We found that PC1, which captured over 34% of the variation, separated individuals by their abundance of either of these pathogens, as well as separating the Korowai from the populations of Mentawai and Sumba (Figure 1, B 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 1, B and C). Although other taxa contributed less to sample clustering, we did

observe that some bacteria had a significant correlation with sample clustering, such as Actinobacteria in PC3 which captured 9.9% of the variation (Supplementary Figure 5; For a full table of p-values for each taxa from ANOVA, see Supplementary Table 2).

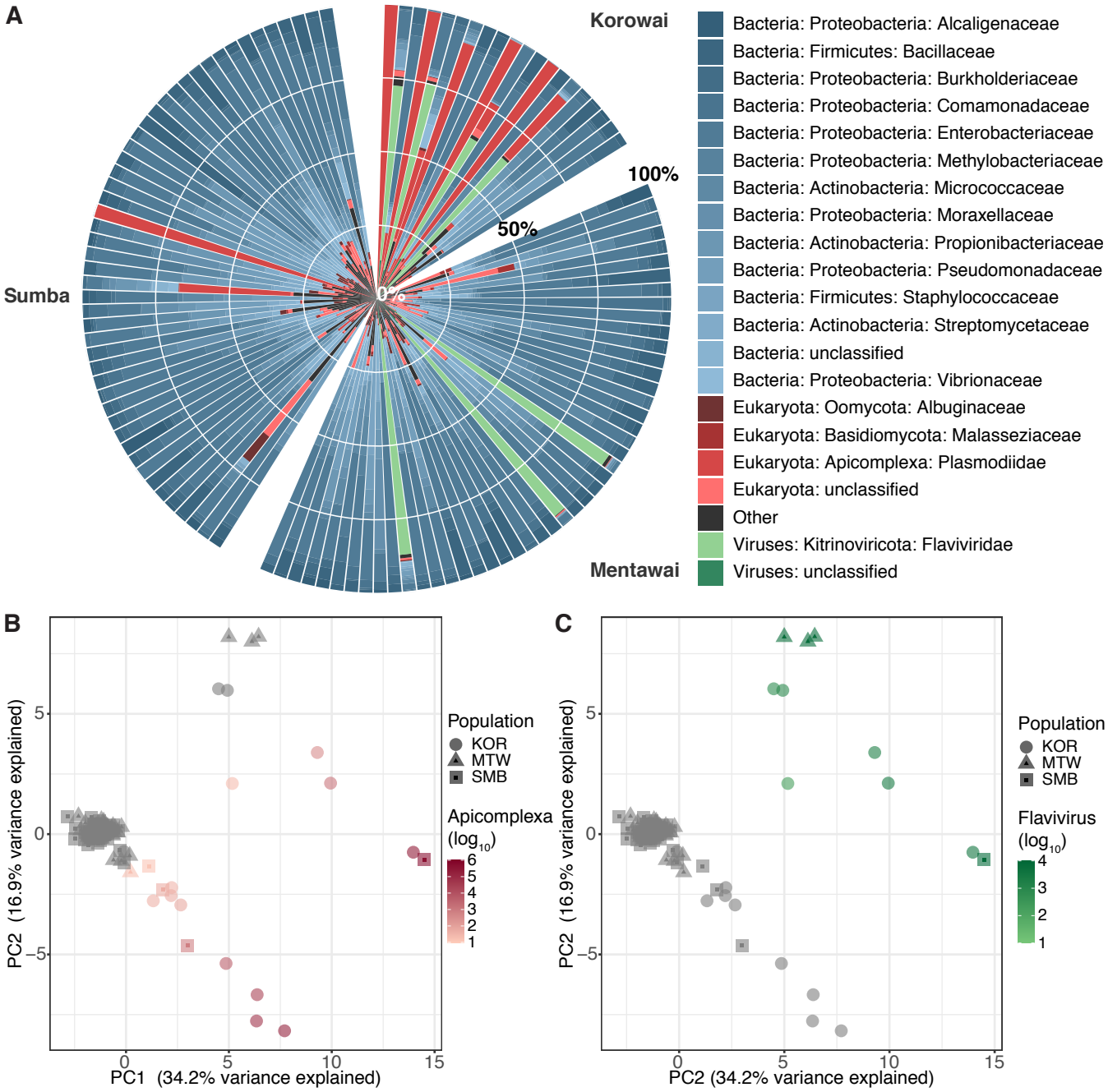


Fig 1. The blood metatranscriptome of the Indonesian populations. A) Circular barplot showeing relative abundance (as % of reads) of the top 20 taxa within each individual in the Indonesian dataset, resolved at the family level. Bacteria are shown in blue, eukaryotes in red, and viruses in green. KOR = 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 $\log_1 0$ Apicomplexa abundance is indicated in red and C) green for Kitrinoviricota.

The Korowai drive differences in 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 [37–39]. Differential abundance testing at the phylum level resulted in significant differences in Apicomplexa abundances between population comparisons with the Korowai (FDR adjusted Welch's t-test: Mentawai versus Korowai $p = 0.011$; Sumba versus Korowai $p = 0.028$; Figure 2, A and B). When we performed the same test on individuals from Mentawai versus individuals from Sumba, we found no differentially abundant phyla (Figure 2, C).

The diversity and types of microbes within human tissues can be an indicator of the overall health of an individual, and of a population [7, 48]. We therefore analysed levels of alpha (within individual) and beta (between individual) diversity within the three islands using DivNet [40], again at the phylum level. We found that while alpha diversity estimates were overall largely similar between individuals from Mentawai and Sumba, they were slightly lower in individuals from the Korowai population. This was true for both estimates of Shannon diversity (mean Shannon KOR = 0.92; MTW = 1.15; SMB = 1.18; Supplementary Figure 6, A) and inverse Simpson diversity indices (mean inverse Simpson KOR = 0.42; MTW = 0.53; SMB = 0.54; Supplementary Figure 6, B). On average, however, this observation was likely due to 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. To confirm this, we focused on Korowai individuals with some of the lowest rates of diversity and found they were samples with a high number of reads mapping to either Apicomplexa or Kitrinoviricota (Supplementary Figure 7). In support of this, we found that if we excluded individuals with the highest Apicomplexa and Kitrinoviricota loads when calculating these statistics, diversity estimates increased. We also found that samples within the Korowai population had the greatest levels of dissimilarity from each other in estimates of beta diversity (Figure 2, D). Indeed, most comparisons that involved Korowai individuals resulted in higher estimates of Bray-Curtis dissimilarity than comparisons with either the Sumba or Mentawai populations (Figure 2, D), demonstrating the range of microbial diversity within this population.

Microbiomes are distinct between global populations

In order 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 [23, 24], which represents the microbiome

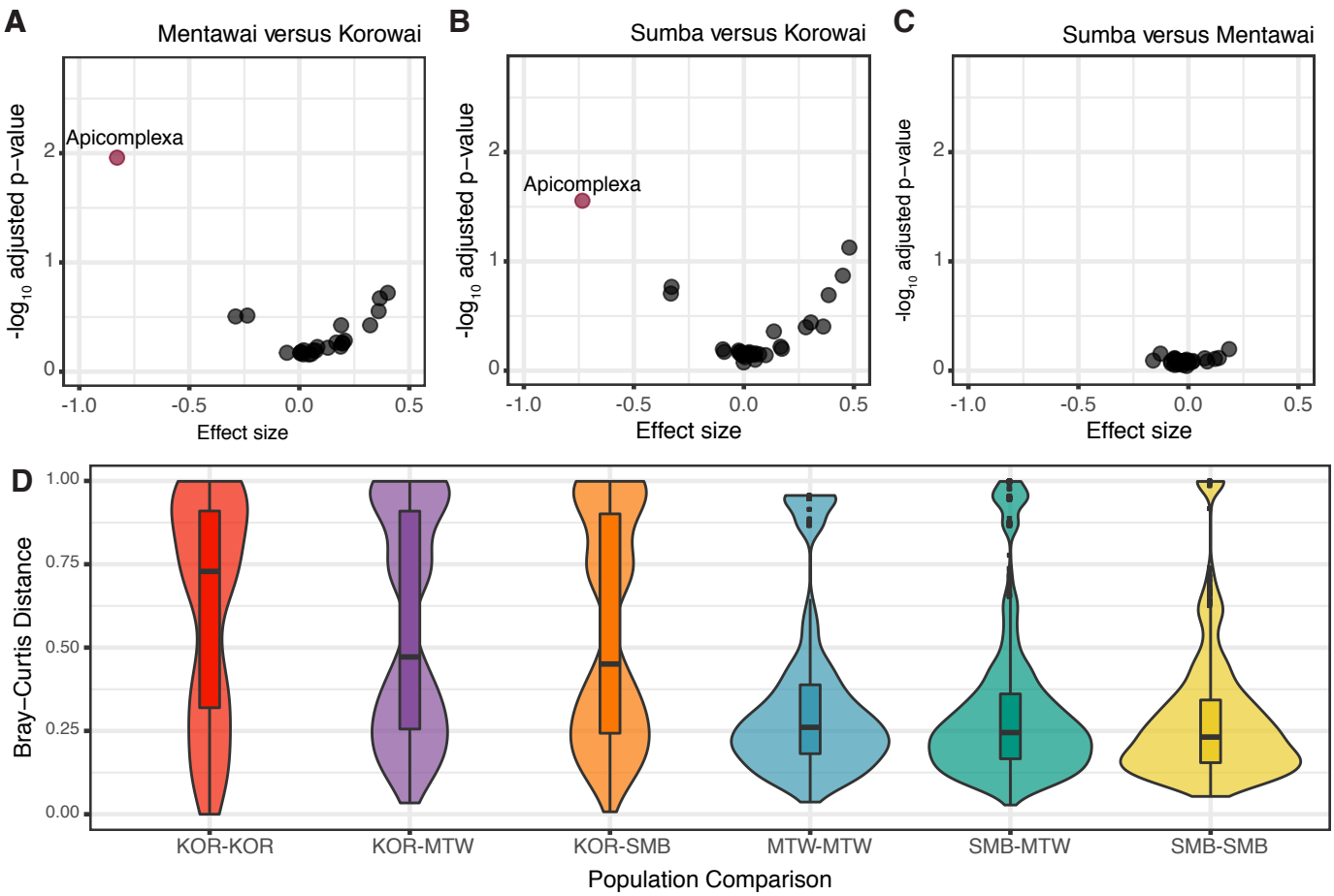


Fig 2. The Korowai drive differences between island populations. A) Volcano plot of BH adjusted p-values from Welch's t-test and the effect size for each taxa at the phylum level, in Mentawai versus Korowai B) Sumba versus Korowai and C) Sumba versus Mentawai. Taxa with a BH-corrected p-value less than 0.05 are coloured in red. D) Bray-Curtis distance estimates for each population comparison at the phylum level.

of individuals living in rural environments, and 12 healthy individuals collected from the city of London in the United Kingdom [25], 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 [49].

After processing of reads as above, we obtained a mean library size of 966,203 reads (range: 403,414 - 1,422,632) for the UK dataset and 20,119 (range: 2,268 - 180,484) for the Malian dataset. This was further reduced to a mean of 966,195 reads (range: 403,402 - 1,422,623) and 20,096 reads (range: 2,241 - 180,468) for the UK and Malian datasets after the removal of singletons, respectively (Supplementary Table 1). 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. In particular, the Indonesian dataset loses significant numbers of reads when we filter reads assigned to either Viridiplantae or Metazoa, and the UK dataset has a lot fewer reads mapping to repetitive regions than either the Malian or Indonesian

datasets (Supplementary Table 1; Supplementary Figure 3). In the UK dataset, we identified a total of 745 distinct taxa across all phylogenetic levels. These were predominantly bacterial in origin, with the majority of reads assigned to Proteobacteria (75.1% of the total read pool across all individuals) and Actinobacteria (22.3% of reads; Supplementary Figure 8). Within the Malian dataset, we found a much greater variation in taxa (2,193 distinct taxa across all phylogenetic levels), the majority of which were Apicomplexa (42.6% of reads), followed by Euryarchaeota (20.3% of reads), Actinobacteria (12.7% of reads), Firmicutes (12.3% of reads), Proteobacteria (7.8% of reads), and Artverviricota (1.1% of reads; Supplementary Figure 8). Although archaea have previously been identified in whole blood [7, 50], we found the high number of Euryarchaeota within the Malian dataset surprising. However, investigation of the archaeal reads confirmed they mapped to several different loci (Supplementary Figure 9), and BLAST searches only returned archaea as best matches. Although there is a substantial difference in read depths between all three data sets, saturation curves show that at the same read depth the UK samples are systematically less diverse than the Indonesian or Mali samples (Supplementary Figure 10).

We performed differential abundance testing between the Indonesian, Malian, and UK datasets (Supplementary Figure 11). Twenty-nine phyla were significantly differentially abundant between Indonesian and Malian individuals (Supplementary Table 3). The strongest signal was driven by Euryarchaeota (FDR adjusted Welch's t-test $p = 2.9 \times 10^{-70}$), which was completely absent from the Indonesian population, as well as higher abundances of Artverviricota and Apicomplexa in the Malian population (FDR adjusted Welch's t-test $p = 5.3 \times 10^{-12}$ and 7.5×10^{-12}, respectively; Supplementary Table 3). Euryarchaeota is an archaeal taxa that contains methanogens, halophiles, and hyperthermophiles [51], and has been previously observed in the blood of Korean and Dutch populations [7, 50]. When comparing blood microbiomes between the UK and Indonesian populations, we found 4 differentially abundant phyla, the most significant being Proteobacteria and Actinobacteria, both of which were more abundant in the UK population (FDR adjusted Welch's t-test $p = 1.3 \times 10^{-10}$ and 5.7×10^{-14}, respectively; Supplementary Table 3).

Since our analyses above suggested that the Korowai are the most differentiated out of the three Indonesian island populations (Figure 2), we next repeated differential abundance testing using only the Korowai as the Indonesian comparison group. We found that in comparisons between the UK and Korowai population, Apicomplexa and Kitrinoviricota were significantly more abundant within the Korowai (FDR adjusted Welch's t-test $p = 1.4 \times 10^{-3}$ and 0.049, respectively; Figure 3, A; Supplementary Table 3). In comparisons between the Korowai and Malian groups, 9 taxa were significantly diferentially abundant, with no significant difference in Apicomplexa abundance and a higher abundance of Kitrinoviricota in the

Korowai population (FDR adjusted Welch's t-test p = 0.043; Figure 3, B; Supplementary Table 3). We found that 8 of these taxa were shared with the Indonesian versus Malian comparison at the country-level, however 21 taxa which were significantly differentially abundant at the country-level did not near significance (before and after FDR correction) in tests between the Korowai and Malian populations, suggesting closer similarity of taxa abundance between these two populations.

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, with PCA yielding a separate cluster for each dataset (Figure 3, C). These differences were most apparent in comparisons with the Malian population, where PC1 separated the UK and Indonesian samples from the Malian samples. Indeed, this was recapitulated by Bray-Curtis distance estimates, where population comparisons with Mali showed higher estimates of Bray-Curtis distance compared to comparisons with the Indonesian or UK populations (Supplementary Figure 12). In PC2, we found samples to be further separated into two clusters, with the Malian and Indonesian samples were separated from the UK samples (Figure 3, D).

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 and inverse Simpson diversity values (mean Shannon = 0.61; mean inverse Simpson = 0.36), followed by individuals from Indonesia, then Mali (Indonesian mean Shannon = 1.04; Indonesian mean inverse Simpson = 0.48; Malian mean Shannon = 1.41; Malian mean inverse Simpson = 0.63; Figure 3, E and F). Although alpha diversity estimates were highest in the Malian population, we expect that the Indonesian and Malian populations would have similar estimates of diversity if read abundances were higher in the Indonesian dataset (Supplementary Table 1), which did not reach a full saturation of reads (Supplementary Figure 10). In order to ensure that diversity estimates were not driven by differences in sample size, we also subsampled the Malian and Indonesian datasets to 12 samples and repeated this test 1,000 times. We found that after subsampling, each population had similar diversity estimates (Indonesian mean Shannon = 1.02; Indonesian mean inverse Simpson = 0.48; Malian mean Shannon = 1.41; Malian mean inverse Simpson = 0.63). We also note that the UK population has the highest library depth out of the three populations and consequently the greatest power to detect rare taxa, and therefore these estimates likely reflect true rates of lower diversity within the UK population.

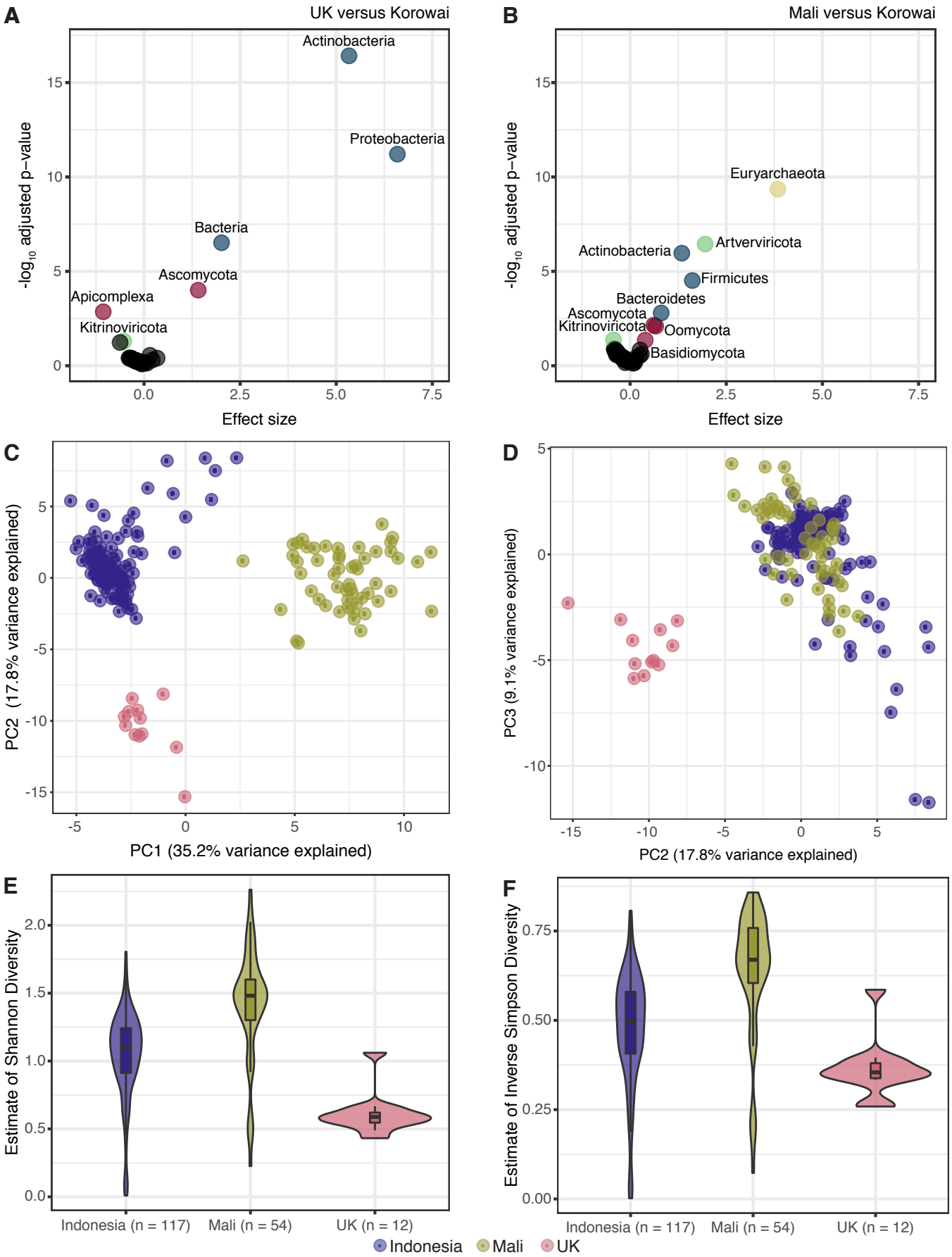


Fig 3. (caption on next page)

Fig 3. Taxa differences between Korowai individuals and other global populations. A) Volcano plot of BH adjusted p-values from Welch's t-test for each phyla in Malian versus Korowai individuals and B) UK versus Korowai individuals. Taxa with a BH-corrected p-value below 0.05 for are coloured by superkingdom (red: eukaryotes; blue: bacteria; green: viruses; yellow: archaea). C) Principal component analysis of the CLR-normalised taxa abundance data at the phylum level for PC1 versus PC2 and D) PC2 versus PC3. E) Violin plots of Shannon diversity and F) inverse Simpson diversity for each population.

ICA reveals Apicomplexa and Kitrinoviricota are contributors to gene expression

Transcriptomic data is influenced by various technical and biological factors. The ability to separate these mixed signals into independent sources can be solved by using ICA [52]. At the most basic level, ICA works by uncovering independent source signals from a set of expression data by creating a set of linear mixtures across genes and samples, which when applied to gene expression can reflect distinct biological processes [53,54], and the contributions of specific genes to each component. Here we apply it to the original human RNA-seq dataset from these same individuals, to explore whether human gene expression in these individuals is influenced by pathogen load as quantified in our analyses above. ICA across 5 independent components identified hundreds of significant contributing human genes (Supplementary Table 4), which we then tested for enrichment against GO and KEGG pathways to interpret which biological processes these components represent. With the exception of IC5 which had no significantly enriched pathways, we were able to investigate the source signals associated with the first four ICs.

In IC1, the most enriched GO categories and KEGG pathways were those related to cell adhesion and the general immune response. Some of the most highly enriched GO categories included granulocyte activation and leukocyte activation involved in the immune response (Supplementary Table 5), while KEGG pathways were enriched in cell adhesion molecules, cell surface proteins involved in the immune response, and inflammation [55] (Table Supplementary Table 6). We next assessed correlations between microbial taxa load and individual sample contributions to the IC and found that six microbial taxa were significantly associated with IC1 (Supplementary Figure 13). The strongest correlation in this IC was that of total CLR-transformed Apicomplexa (Plasmodiidae) load and sample contribution, with a negative sample contribution correlating with a higher Apicomplexa load (R = -0.51; FDR-adjusted p = 5.4×10^{-9}; Figure 4A). This seemed to be driven by inter-island differences: Korowai individuals had a much larger negative contribution to IC1, while individuals from Mentawai and Sumba had, on average, positive contributions (Figure 4, B). Since Apicomplexa levels are, on average, higher in Korowai individuals, we cannot distinguish between this component being driven by Apicomplexa abundance or by other island-level differences; however, we note that this IC was the most significant in its differential distribution of island and sample contributions

(FDR adjusted Kruskal-Wallis $p = 1.3 \times 10^{-18}$). Some of the most significant contributing genes included genes we have previously identified as differentially expressed between these island populations [20], such as *MARCO*, a macrophage surface receptor involved in antigen presentation [56] and which has been shown to clear various microorganisms within the host [57] (Figure 4, C; Supplementary Table 4). While this gene did have a significant negative correlation between Apicomplexa load and expression (Figure 4, C), we found that this gene is heavily-stratified by island, with the majority of Korowai individuals having lower *MARCO* expression. This was true even for Korowai individuals with low levels of Apicomplexa abundance. Indeed, we have recently [21] identified rs13425622 as an eQTL that strongly impacts expression levels of MARCO in these populations ($p = 3.10 \times 10^{-14}$). All Korowai individuals in our dataset are fixed for the minor G allele, which is associated with lower expression, while the majority of individuals from both Mentawai and Sumba are at least heterozygotes for the major T allele (Figure 4, D).

In IC2, we found 215 contributing genes, which were broadly associated with multiple cardiomyopathies and regulation of cardiac muscle (Supplementary Table Supplementary Table 5; Supplementary Table 6). Within this component, there was a significant difference between island and sample contributions (FDR adjusted Kruskal-Wallis $p = 2.2 \times 10^{-8}$), with individuals from Mentawai observed to drive differences between the populations (Supplementary Figure 14, A). Although the abundance of Apicomplexa was significantly correlated with sample contribution to IC2, only 4 Mentawai individuals had reads mapping to Apicomplexa, all at low levels (range: 4-18 reads; Supplementary Figure 14, B). This suggests that taxa we could identify within whole blood are likely not a driver of the signal within this component.

Apicomplexa load was also weakly associated with IC3 contributions (Supplementary Figure 14, D), a component which we found to be involved in the response to malaria. Some of the most highly enriched GO categories associated with this IC were those involved in heme metabolic processes (Supplementary Table 5); the only significantly enriched KEGG pathway was malaria (Table Supplementary Table 6). While both IC1 and IC3 had an association between sample contributions and Apicomplexa (Plasmodiidae) abundance, we found that there were differences between contributing genes of the two components. Rather than cell adhesion and a general immune response, many of the genes in IC3 were well-characterised genes related to a malaria response. These included *SLC4A1*, the gene with the highest contribution to IC3, and which was not found in IC1. *SLC4A1* is involved in Southeast Asian Ovalocytosis, a red blood cell disorder that is protective against malaria infection and most commonly found in the Southeast Asia and the Southwest Pacific region [58]. Other genes which were unique to IC3 and that have been implicated in the response to malaria included *HBB*, *HBA1*, *HBA2*, *ACKR1*, and the glycophorins *GYPA*, *GYPB*, and *GYPC*. However,

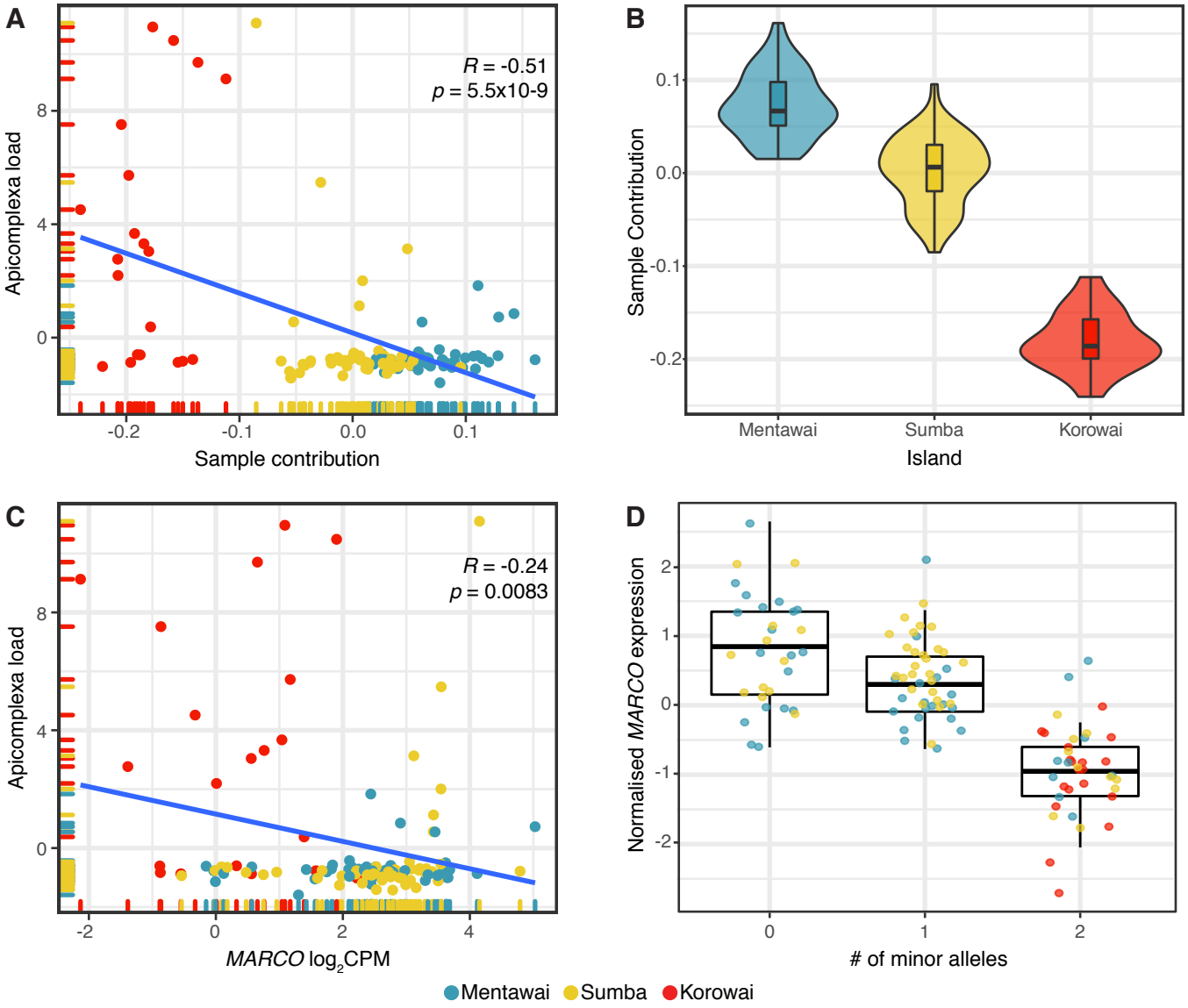


Fig 4. Independent component analysis on the CLR-normalised Indonesian expression data. A) Correlation of CLR-normalised Apicomplexa load and sample contribution for IC1 B) Distribution of sample contributions for each island within IC1 C) Correlation of CLR-normalised Apicomplexa load and *MARCO* $\log_2$ CPM values across all samples D) Relationship between genotype at rs13425622 and *MARCO* expression levels in these populations.

although many of the contributing genes and enriched pathways within this component were involved in malaria processes, we did not find a strong correlation between total Apicomplexa load and sample contributions. While we did find a significant difference between island populations for this entire component (FDR adjusted Kruskal-Wallis $p = 2.2 \times 10^{-8}$), with individuals from Sumba having a positive sample contribution on average, and Korowai and Mentawai individuals having a negative sample contribution on average (Supplementary Figure 14, E), it was not as clearly stratified as in IC1. Furthermore, major contributing genes such as *SLC4A1* did not have significant differences in levels of expression between islands (Supplementary Figure 14, F). Indeed, individuals with Southeast Asian Ovalocytosis are reported

to have decreased levels of *SLC4A1* due to a nonfunctional copy of the gene [59], and we did not find any evidence of decreased *SLC4A1* expression to be correlated with higher Apicomplexa abundance ($R = 0.04$; $p = 0.64$). While this component as a whole does seem to be involved in malaria-associated processes, the signal may instead come from multiple individuals across islands, and is more challenging to interpret.

Finally, viral signatures were the main signal within IC4. We found that the most enriched GO categories included defense response to virus ($p = 4.6 \times 10^{-30}$) and negative regulation of viral genome replication ($p = 6.8 \times 10^{-12}$), while KEGG pathways were enriched in hepatitis C genes (FDR adjusted $p = 1.3 \times 10^{-5}$), RIG-I-like receptor signaling pathway (FDR adjusted $p = 1.1 \times 10^{-4}$), and cytosolic DNA-sensing pathway (FDR adjusted $p = 2.3 \times 10^{-4}$; Table Supplementary Table 6), all well-documented responses to viruses [60, 61]. Similar to IC1 and IC3, there was a significant negative correlation between Kitrinoviricota (Flaviviridae) load and sample contributions (FDR adjusted Kruskal-Wallis $p = 4.1 \times 10^{-15}$; Supplementary Figure 14, G), with Korowai individuals having on average mostly negative contributions (Supplementary Figure 14, H). While we did find a significant difference between islands and sample contributions, once again sample contributions were not as clearly stratified as in IC1. Rather, certain individuals, predominantly within the Korowai, were found to drive the correlation between Kitrinoviricota load and sample contributions (Supplementary Figure 14, G). However, we did find that multiple genes involved in response to viruses, such as *RSAD2*, were genes which we have previously found to be significantly differentially expressed between comparisons involving Korowai individuals [21]. *RSAD2*, or viperin (virus inhibitory protein, endoplasmic reticulum associated, interferon inducible), is a well-characterised gene involved in antiviral activity [62] and its upregulation has been associated with multiple viruses, including Flaviviruses [63–66]. Indeed, we found that *RSAD2* was the main contributing gene to this component, with higher levels of expression, on average, in Korowai individuals (Supplementary Figure 14, I). Other notable genes implicated in antiviral activity within this component, which we have previously found to be significantly differentially expressed between comparisons with Korowai individuals, include the genes *OAS1-3*. These genes are implicated in antiviral activity [67], with both *OAS1* and *OAS3* being implicated in protection against Flavivirus infection [68, 69].

Discussion

Our understanding of pathogens found within remote regions within Indonesia, along with their impact on gene expression, is limited. Here, we have investigated which microbial taxa can be detected within whole blood and the influence they have on blood expression profiles. Although we did not reach full saturation of reads (Supplementary Figure 10), we found a combination of multiple taxonomic kingdoms that constitute the Indonesian whole blood microbiome. As described in previous research on blood microbiomes from other global populations [70–72], bacteria were some of the most abundant taxa found within Indonesian samples (Figure 1, A). Although bacteria found within blood have most commonly been associated with sepsis, mounting evidence suggests that some bacteria are normal inhabitants of whole blood, likely originating from the gut and oral cavities [72,73], although they may also represent leakage from other parts of the body. We also found evidence for the presence of eukaryotes, archaea, and viruses, all of which have previously been characterised in blood transcriptomes [74]. This study supports a growing body of research suggesting that rather than being a sterile environment, a variety of taxa naturally reside within whole blood, and understanding the roles of these microbes in future studies may help facilitate better understanding of healthy and disease states in different populations.

Although we found that the majority of these microorganisms did not have a detectable association with gene expression, two phyla——Apicomplexa (driven nearly exclusively by the family Plasmodiidae) and Kitrinoviricota (driven by the family Flaviviridae)——did have noticeable effects. This was supported by ICA, which showed that contributing genes were enriched in responses to malaria and viruses (Supplementary Table 5). Indeed, genes such as *SLC4A1*, which is involved in Southeast Asian Ovalocytosis—a protective polymorphism against severe malaria [58]—and *ACKR1* which encodes the Duffy antigen/chemokine receptor (*DARC*) [75], were some of the highest contributing genes in IC3. Responses to viral infections were also apparent in IC4, where *RSAD2*, a well-characterised gene involved in the antiviral response [62], was the main contributing gene and multiple pathways were enriched for viral responses (Supplementary Table 5). From taxonomic profiling, we could attribute Kitrinoviricota viral signals to the family Flaviviridae, which is a family of primarily found in mosquitos and ticks, and is responsible for multiple human illnesses including Zika, Dengue, and Yellow Fever [76,77], alhough we were unable to refine this assignment further. For Apicomplexa, we could attribute 99.9% of reads to the family Plasmodiidae, of which *Plasmodium falciparum* and *Plasmodium vivax* are endemic throughout Indonesia [78].

Of all the Indonesian island populations in this study, we found that the Korowai not only had the highest abundance of both of these two pathogens, but were also a driver of differences between islands in ICA. The

Indonesian side of New Guinea Island is documented to have the highest rates of malaria in Indonesia [79], as well as the lowest number of healthcare facilities [80]; our results corroborate existing observations of a high endemic pathogen load within this region. The Korowai were the biggest drivers of difference within IC1 (FDR adjusted Kruskal-Wallis $p = 1.2 \times 10^{-18}$), which we identified to be associated with a general immune response. Supporting this, we have shown in our previous data [20] that multiple immune genes are differentially expressed between both Sumba and Mentawai and the Korowai. This includes *MARCO*, a macrophage receptor gene which is activated upon infection by bacteria and parasites [81,82], and which we found to be one of the main contributing genes to IC1. Together, this suggests that differences in exposure to pathogens contribute to gene expression differences between populations.

Although this study focuses on pathogens within Indonesian blood microbiomes, the ability to detect signals relating to cardiomyopathies and cardiovascular pathways in IC2 (Supplementary Table 5; Supplementary Table 6) demonstrates the ability of ICA to differentiate multiple biological signals into their constituent parts. Indeed, Indonesia faces multiple health burdens including not only infectious diseases, but also non-communicable diseases, such as cardiovascular disease [83,84]. Applications of ICA on expression profiles could therefore be especially useful when multiple illnesses are comorbid within a population in order to discriminate between diseases and apply more targeted interventions.

In addition to characterising Indonesian whole blood microbiomes, we have also shown that these are distinct to those of other global populations, although these findings are limited by the fact that all three datasets we considered were generated by different groups in different places. Nevertheless, Bray-Curtis distance estimates showed that the Indonesian, Malian, and UK populations had high levels of dissimilarity from one another (Supplementary Figure 12, C), and multiple taxa were differentially abundant between the three global populations (Figure 3, A and B; Supplementary Table 3). Intriguingly, we found that within the Malian population, every individual had high abundances of the archaeal phylum Euryarchaeota. Although we found this result surprising, previous studies have documented not only archaea within whole blood [7], but also the same phylum of archaea [50]. Furthermore, an independent study conducted on Malian individuals found Euryarchaeota within oral cavities of patients, [85] and archaea are well-characterised to naturally inhabit the human body [86].

Apart from differences between the three global populations, we also found differences in diversity between populations comparisons to the UK. Indeed, alpha diversity indices were higher in Malian and Indonesian populations, although the UK population had the highest read depth out of all three populations. This may suggest that rural and urban microbiomes differ, with a more microbially-rich microbiome in

rural populations. Supporting this, a study of gut microbiomes within hunter gatherers found similar results: the Hadza, a small hunter gatherer group living in Tanzania, had more diverse gut microbiomes than Italian urban controls [87]. In addition, a different 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 [88]. This raises the question of what healthy whole blood microbiomes look like, and calls for further research into the influence of lifestyle, geography, and pathogenic load on this tissue type.

Although we have shown that whole blood transcriptomes can be exploited for diagnostic purposes, some limitations remain. For one, we note that although our total sample sizes are high—which is rare in studies utilising understudied 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 and did not reach full saturation (Supplementary Figure 10). 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. Therefore, for studies utilising whole blood RNA for diagnostic purposes, care should be taken to understand the influence of these factors on pathogen detection.

A better understanding of which pathogens affect remote populations and the impact they have on the immune response is crucial. Whole blood is one of the most abundant tissue types in RNA-seq analysis due to its relative ease of collection [89], 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 [16, 90], however proper surveillance in rural areas still remains limited. This study therefore provides valuable surveillance information on blood-borne microorganisms within the region, which is a crucial step in limiting the spread of endemic and emerging diseases, as well as a readily-adaptable approach that can be applied to already existing RNA-seq datasets from anywhere in the globe.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

KSB and IGR designed the study and wrote the manuscript with input from all authors, with help from CW. KSB performed analyses . CCD, PK, HS and SM generated raw data for all anaylses. PK performed additional analyses. CAF 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 and members of the Gallego Romero group for helpful comments on the manuscript.

References

1. 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.

2. 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.

3. 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.

4. 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.

5. 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.

6. 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.

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

8. 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.

9. 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.

10. 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.

11. 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.

12. Gloria Serena, Camron Davies, Murat Cetinbas, Ruslan I Sadreyev, and Alessio Fasano. Analysis of blood and fecal microbiome profile in patients with celiac disease. *Human Microbiome Journal*, 11:100049, 2019.

13. 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.

14. 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.

15. 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.

16. 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.

17. Yodi Mahendradhata, Laksono Trisnantoro, Shita Listyadewi, Prastuti Soewondo, Tiara Marthias, Pandu Harimurti, and John Prawira. The Republic of Indonesia health system review. 2017.

18. 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.

19. World Health Organization et al. WHO country cooperation strategy 2014–2019: Indonesia. 2016.

20. 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. 16(5):e1008749. Publisher: Public Library of Science.

21. 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.

22. 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.

23. 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.

24. 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.

25. 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.

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

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

28. 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.

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

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

31. 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.

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

33. 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.

34. 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.

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

36. 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.

37. 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.

38. 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.

39. 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.

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

41. Anne Biton. MineICA: Analysis of an ICA decomposition obtained on genomics data, 2019.

42. Jean-François Cardoso and Antoine Souloumiac. Blind beamforming for non-Gaussian signals. In *IEE proceedings F (Radar and Signal Processing)*, volume 140, pages 362–370. IET, 1993.

43. Christian Jutten, Saïd Moussaoui, and Frédéric Schmidt. How to apply ICA on actual data? example of Mars hyperspectral image analysis. In *15th International Conference on Digital Signal Processing*, pages 3–12. IEEE, 2007.

44. Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

45. Gene Ontology Consortium. The Gene Ontology resource: 20 years and still GOing strong. *Nucleic Acids Research*, 47(D1):D330–D338, 2019.

46. Minoru Kanehisa and Susumu Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

47. Matthew D Young, Matthew J Wakefield, Gordon K Smyth, and Alicia Oshlack. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biology*, 11(2):R14, 2010.

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

49. 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.

50. Yeojun Yun, Han-Na Kim, Yoosoo Chang, Yunho Lee, Seungho Ryu, Hocheol Shin, Won-Serk Kim, Hyung-Lae Kim, and Jae-Hui Nam. Characterization of the blood microbiota in Korean females with rosacea. *Dermatology*, 235(3):255–259, 2019.

51. John A Leigh, Sonja-Verena Albers, Haruyuki Atomi, and Thorsten Allers. Model organisms for genetics in the domain Archaea: methanogens, halophiles, Thermococcales and Sulfolobales. *FEMS Microbiology Reviews*, 35(4):577–608, 2011.

52. Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13(4-5):411–430, 2000.

53. Wei Kong, Xiaoyang Mou, Xing Zhi, Xin Zhang, and Yang Yang. Dynamic regulatory network reconstruction for Alzheimer's disease based on matrix decomposition techniques. *Computational and Mathematical Methods in Medicine*, 2014, 2014.

54. Chun-Hou Zheng, De-Shuang Huang, Xiang-Zhen Kong, and Xing-Ming Zhao. Gene expression data classification using consensus independent component analysis. *Genomics, Proteomics & Bioinformatics*, 6(2):74–82, 2008.

55. Guangwen Ren, Arthur I Roberts, and Yufang Shi. Adhesion molecules: key players in mesenchymal stem cell-mediated immunosuppression. *Cell Adhesion & Migration*, 5(1):20–22, 2011.

56. Annabelle Grolleau, David E Misek, Rork Kuick, Samir Hanash, and James J Mulé. Inducible expression of macrophage receptor Marco by dendritic cells following phagocytic uptake of dead cells uncovered by oligonucleotide arrays. *The Journal of Immunology*, 171(6):2879–2888, 2003.

57. Nguyen Thuy Thuong Thuong, Trinh Thi Bich Tram, Tran Dinh Dinh, Phan Vuong Khac Thai, Dorothee Heemskerk, Nguyen Duc Bang, Tran Thi Hong Chau, David G Russell, Guy E Thwaites, Thomas R Hawn, et al. MARCO variants are associated with phagocytosis, pulmonary tuberculosis susceptibility and Beijing lineage. *Genes and Immunity*, 17(7):419, 2016.

58. Jason A Wilder, Jonathan A Stone, Elizabeth G Preston, Lauren E Finn, Hannah L Ratcliffe, and Herawati Sudoyo. Molecular population genetics of SLC4A1 and Southeast Asian Ovalocytosis. *Journal of Human Genetics*, 54(3):182–187, 2009.

59. Philip W Fowler, Mark SP Sansom, and Reinhart AF Reithmeier. Effect of the Southeast Asian ovalocytosis deletion on the conformational dynamics of signal-anchor transmembrane segment 1 of red cell anion exchanger 1 (AE1, band 3, or SLC4A1). *Biochemistry*, 56(5):712–722, 2017.

60. Yueh-Ming Loo and Michael Gale Jr. Immune signaling by RIG-I-like receptors. *Immunity*, 34(5):680–692, 2011.

61. Hyun-Cheol Lee, Kiramage Chathuranga, and Jong-Soo Lee. Intracellular sensing of viral genomes and viral evasion. *Experimental & molecular medicine*, 51(12):1–13, 2019.

62. Sandy Mattijssen and Ger JM Pruijn. Viperin, a key player in the antiviral response. *Microbes and Infection*, 14(5):419–426, 2012.

63. Richard Lindqvist, Chaitanya Kurhade, Jonathan D Gilthorpe, and Anna K Överby. Cell-type-and region-specific restriction of neurotropic flavivirus infection by viperin. *Journal of Neuroinflammation*, 15(1):80, 2018.

64. Richard Lindqvist and Anna K Överby. The role of viperin in antiflavivirus responses. *DNA and Cell Biology*, 37(9):725–730, 2018.

65. Kirstin Vonderstein, Emma Nilsson, Philipp Hubel, Larsård Nygård Skalman, Arunkumar Upadhyay, Jenny Pasto, Andreas Pichlmair, Richard Lundmark, and Anna K Överby. Viperin targets flavivirus virulence by inducing assembly of noninfectious capsid particles. *Journal of Virology*, 92(1), 2018.

66. Theodore C Pierson and Michael S Diamond. The continued threat of emerging flaviviruses. *Nature Microbiology*, pages 1–17, 2020.

67. Luis B Barreiro and Lluis Quintana-Murci. Evolutionary and population (epi) genetics of immunity to infection. *Human Genetics*, 139(6):723–732, 2020.

68. Aaron J Sams, Anne Dumaine, Yohann Nédélec, Vania Yotova, Carolina Alfieri, Jerome E Tanner, Philipp W Messer, and Luis B Barreiro. Adaptively introgressed Neandertal haplotype at the OAS locus functionally impacts innate immune responses in humans. *Genome Biology*, 17(1):1–15, 2016.

69. Yize Li, Shuvojit Banerjee, Yuyan Wang, Stephen A Goldstein, Beihua Dong, Christina Gaughan, Robert H Silverman, and Susan R Weiss. Activation of RNase L is dependent on OAS3 expression during infection with diverse human viruses. *Proceedings of the National Academy of Sciences*, 113(8):2241–2246, 2016.

70. Richard W McLaughlin, Hojatollah Vali, Peter CK Lau, Roger GE Palfree, Angela De Ciccio, Marc Sirois, Darakhshan Ahmad, Richard Villemur, Marcel Desrosiers, and Eddie CS Chan. Are there naturally occurring pleomorphic bacteria in the blood of healthy humans? *Journal of Clinical Microbiology*, 40(12):4771–4775, 2002.

71. Kosei Moriyama, Chie Ando, Kosuke Tashiro, Satoru Kuhara, Seiichi Okamura, Shuji Nakano, Yasumitsu Takagi, Takeyoshi Miki, Yoshiyuki Nakashima, and Hideki Hirakawa. Polymerase chain reaction detection of bacterial 16S rRNA gene in human blood. *Microbiology and Immunology*, 52(7):375–382, 2008.

72. 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.

73. 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.

74. 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.

75. Richard Horuk. The Duffy antigen receptor for chemokines DARC/ACKR1. *Frontiers in Immunology*, 6:279, 2015.

76. 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.

77. 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.

78. 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.

79. 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.

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

81. Jintao Xu, Adam Flaczyk, Lori M Neal, Zhenzong Fa, Alison J Eastman, Antoni N Malachowski, Daphne Cheng, Bethany B Moore, Jeffrey L Curtis, John J Osterholzer, et al. Scavenger receptor MARCO orchestrates early defenses and contributes to fungal containment during cryptococcal infection. *The Journal of Immunology*, 198(9):3548–3557, 2017.

82. D Gowda and Xianzhu Wu. Parasite recognition and signaling mechanisms in innate immune responses to malaria. *Frontiers in immunology*, 9:3006, 2018.

83. Dyah Purnamasari. The emergence of non-communicable disease in Indonesia. *Acta Medica Indonesiana*, 50(4):273, 2019.

84. Wiku Adisasmito, Vilda Amir, Anila Atin, Amila Megraini, and Dian Kusuma. Geographic and socioeconomic disparity in cardiovascular risk factors in Indonesia: analysis of the Basic Health Research 2018. *BMC Public Health*, 20(1):1–13, 2020.

85. Elisabeth Sogodogo, Ogobara Doumbo, Gérard Aboudharam, Bourema Kouriba, Ousseynou Diawara, Hapssa Koita, Souleymane Togora, and Michel Drancourt. First characterization of methanogens in oral cavity in malian patients with oral cavity pathologies. *BMC Oral Health*, 19(1):1–6, 2019.

86. H-P Horz and G Conrads. The discussion goes on: what is the role of euryarchaeota in humans? *Archaea*, 2010, 2010.

87. 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.

88. 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.

89. 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.

90. 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.

Supplementary materials

Supplementary Figure 1 Analysis of forward and reverse reads to investigate the higher proportion of mapped single-ended reads.

Supplementary Figure 2 Summary of reads mapping to filtered taxa for the Indonesian, Malian, and UK populations.

Supplementary Figure 3 Read depth per individual library across all filtering steps.

Supplementary Figure 4 Scree plot of variance explained per principal component.

Supplementary Figure 5 PCA of taxa abundance at the phylum level, highlighted by Actinobacteria abundance.

Supplementary Figure 6 Shannon and inverse Simpson diversity estimates in the Indonesian population.

Supplementary Figure 7 Plasmodium abundance versus Shannon diversity for each Korowai individual.

Supplementary Figure 8 Relative abundance of the top 20 taxa within the Indonesian, Malian, and UK dataset at the family level.

Supplementary Figure 9 Processed, unmapped Malian reads spanning the *Methanocaldococcus jannaschii* genome, visualised using IGV.

Supplementary Figure 10 Saturation curves for each global population after singleton removal.

Supplementary Figure 11 Volcano plots of differentially abundant taxa between Indonesian and other global populations.

Supplementary Figure 12 Bray-Curtis distance estimates for Indonesian, Malian, and UK population comparisons at the phylum level.

Supplementary Figure 13 Significant pathogens from BH-adjusted Pearson correlation p-values for each IC.

Supplementary Figure 14 Distribution of sample contributions for each island for ICs 2-5.

Supplementary Table 1 The number of reads at each filtering stage for 100-bp Indonesian reads, as well as 75-bp reads of all populations.

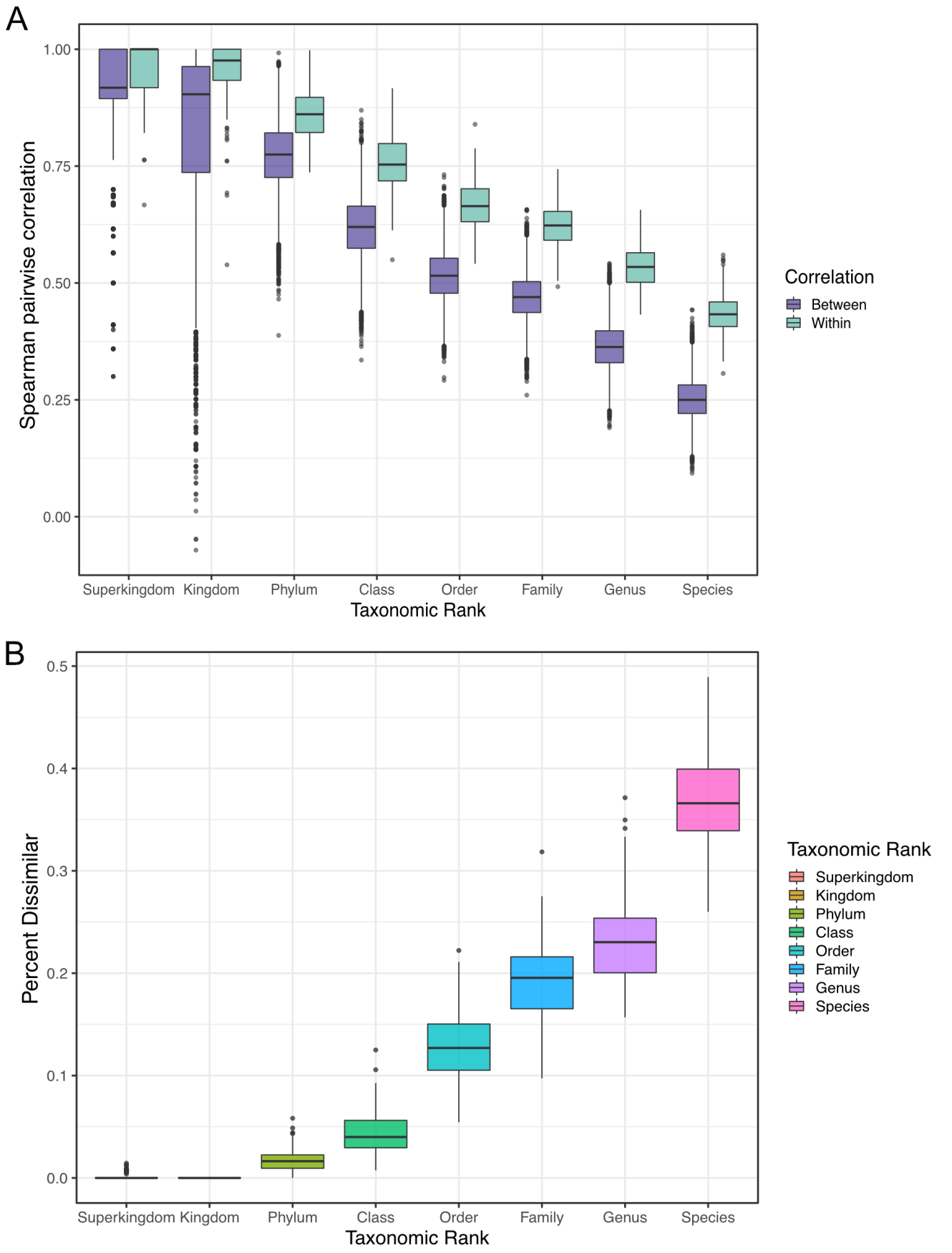
Supplementary Table 2 P-values from ANOVA tests between PCA of taxa abundances and logged abundances of individual taxa.

Supplementary Table 3 Significantly deferentially abundant taxa (Welch's t-test BH-adjusted p = 0.05) at the phylum level for all populations.

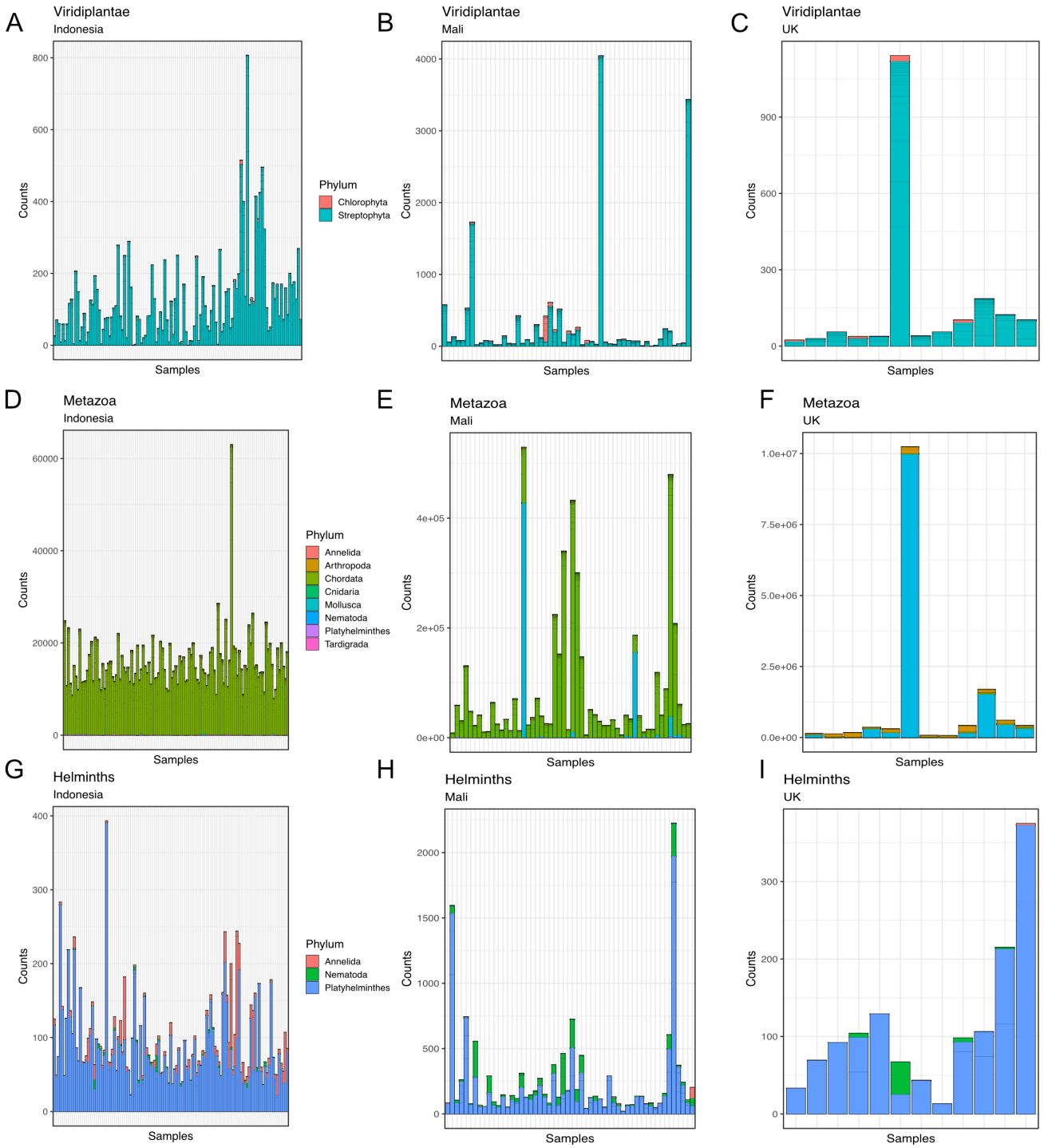
Supplementary Table 4 Contributing genes for each IC for the Indonesian population.

Supplementary Table 5 GO enrichment testing results for contributing genes within each IC.

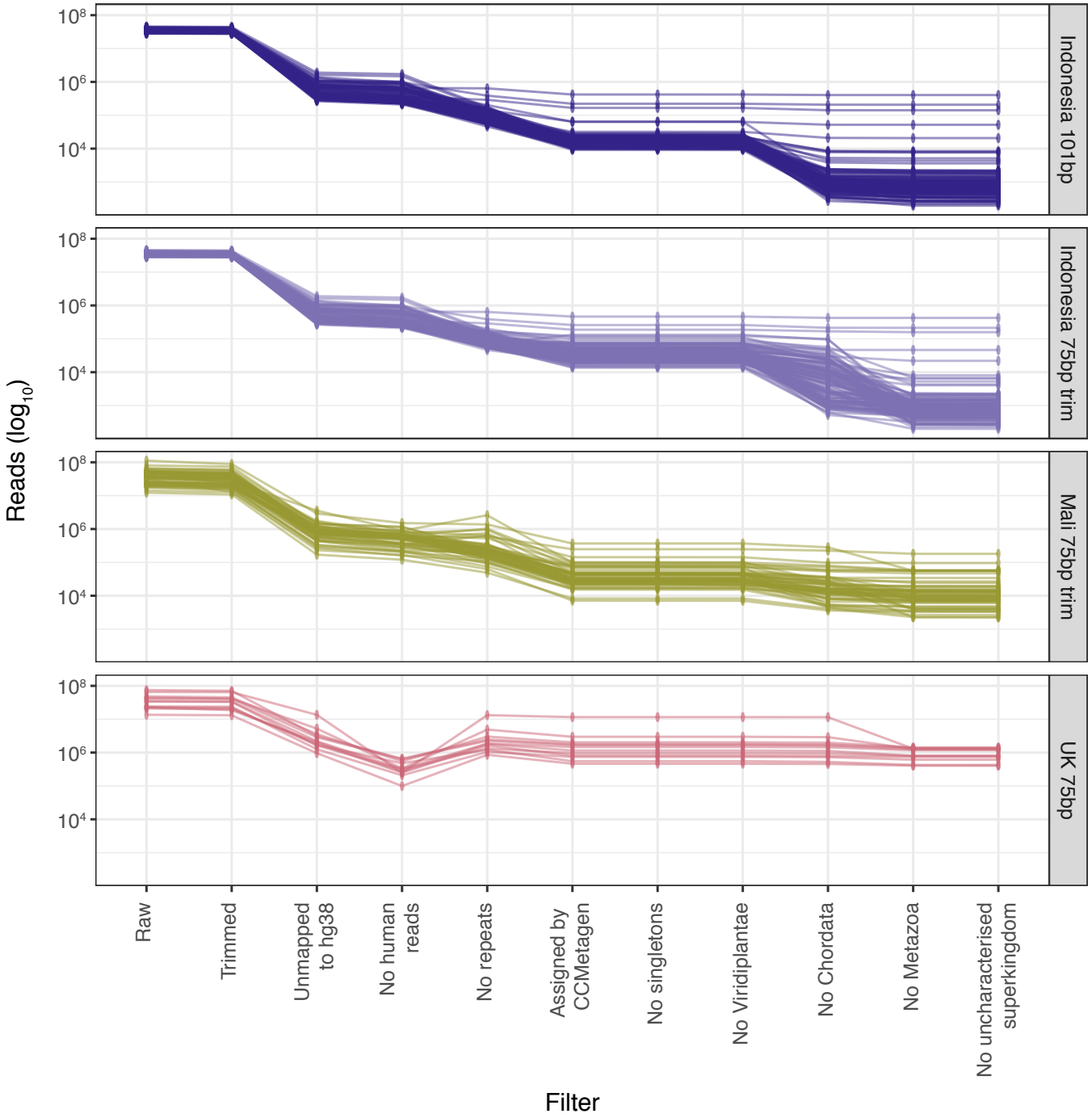
Supplementary Table 6 KEGG enrichment testing results for contributing genes within each IC.



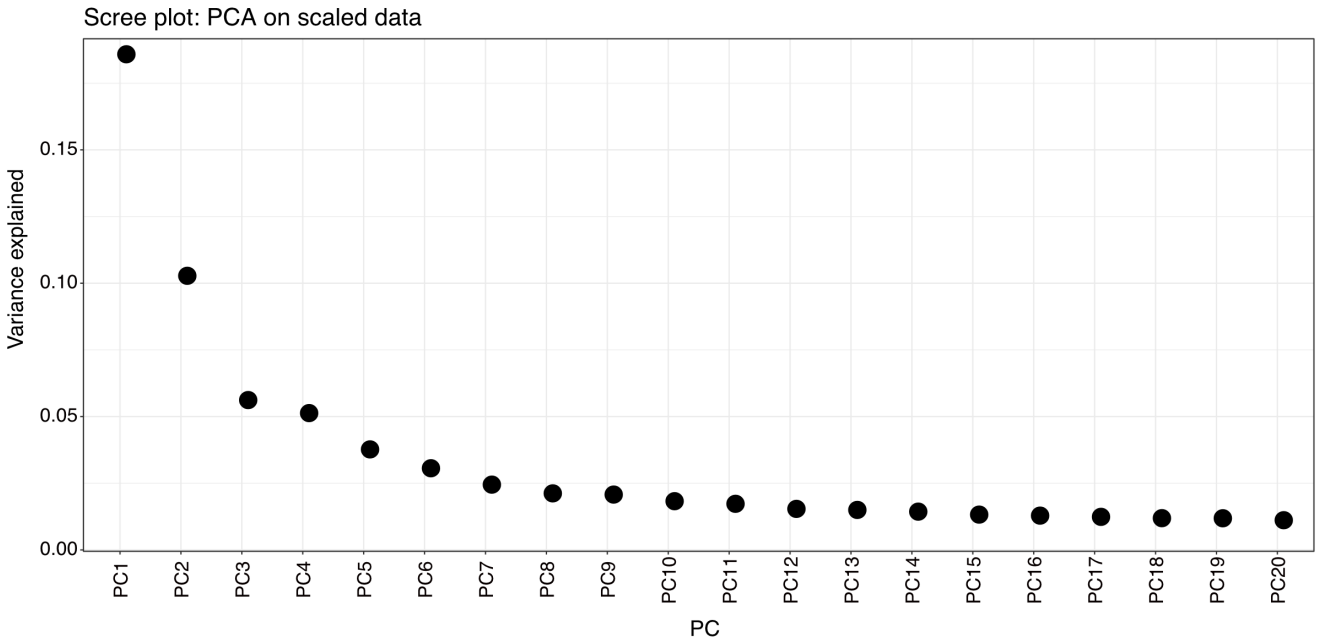
Supplementary Figure 1. Forward versus reverse read comparison. A) Pairwise Pearson correlations between reads from the same mate pair (in blue) and reads outside of their mate pair (purple) for each taxonomic rank. B) Proportion of dissimilar reads between forward and reverse reads at each taxonomic rank.



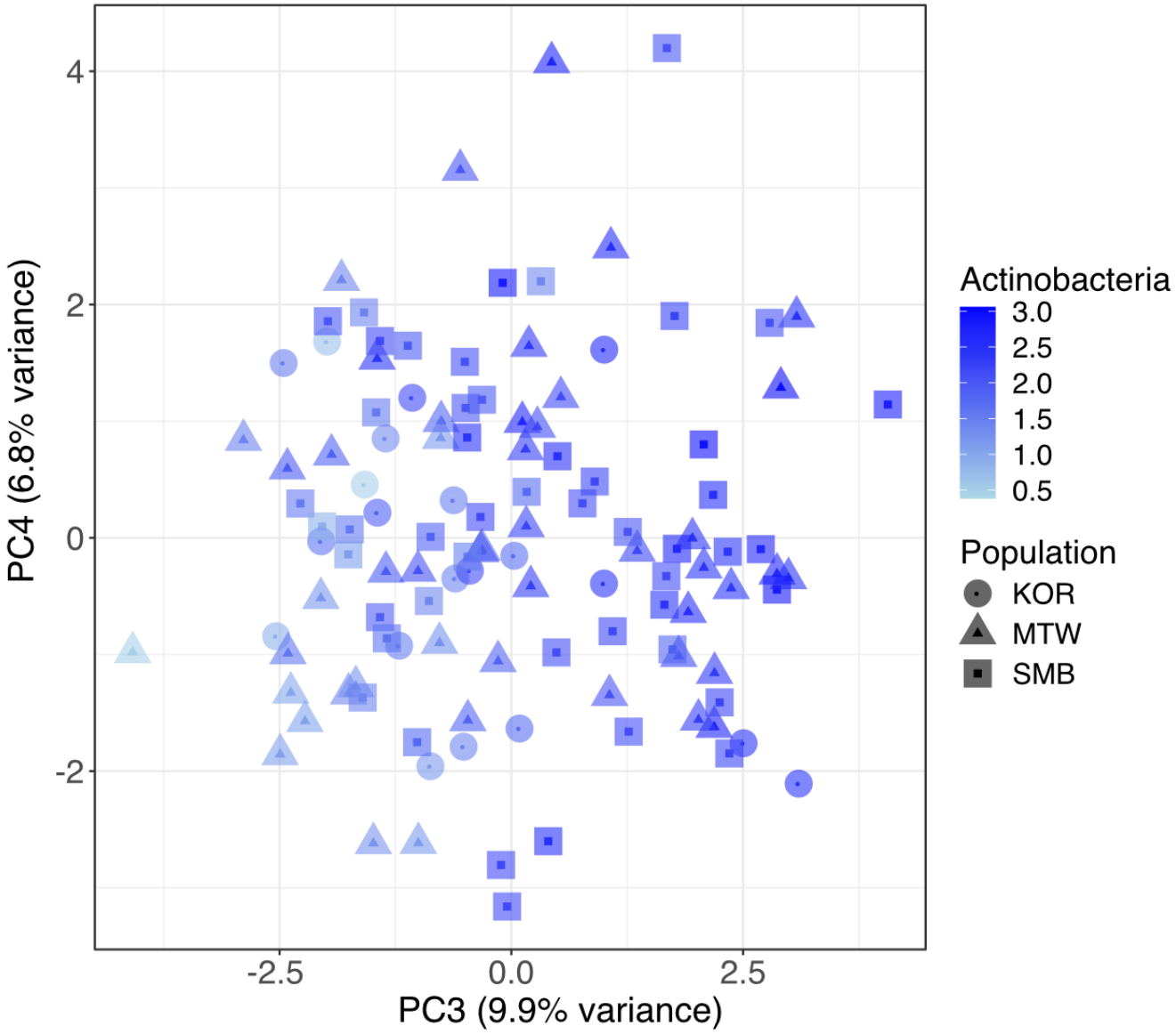
Supplementary Figure 2. Summary of reads mapping to filtered taxa for the Indonesian, Malian, and UK populations. A-C) Reads mapping to the Viridiplantae D-F) Metazoa G-I) and helminths.



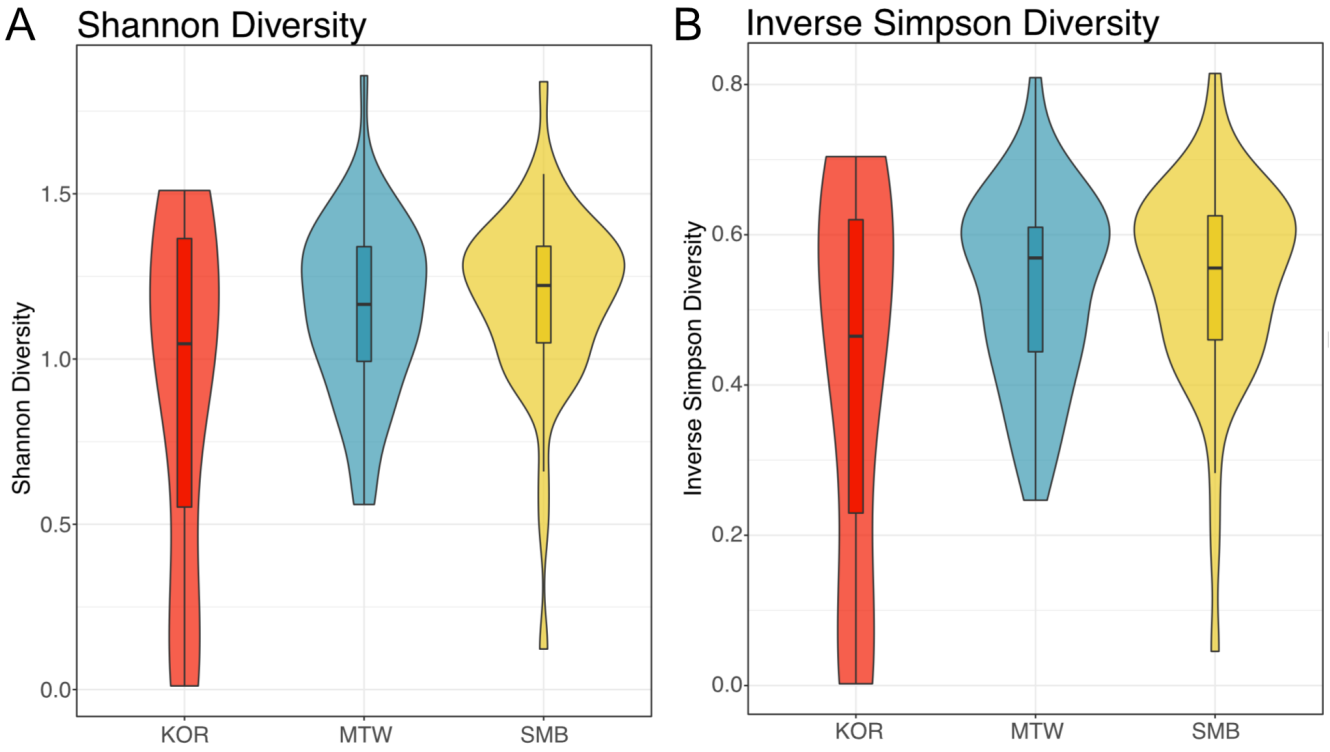
Supplementary Figure 3. Read depth per individual library across all filtering steps.



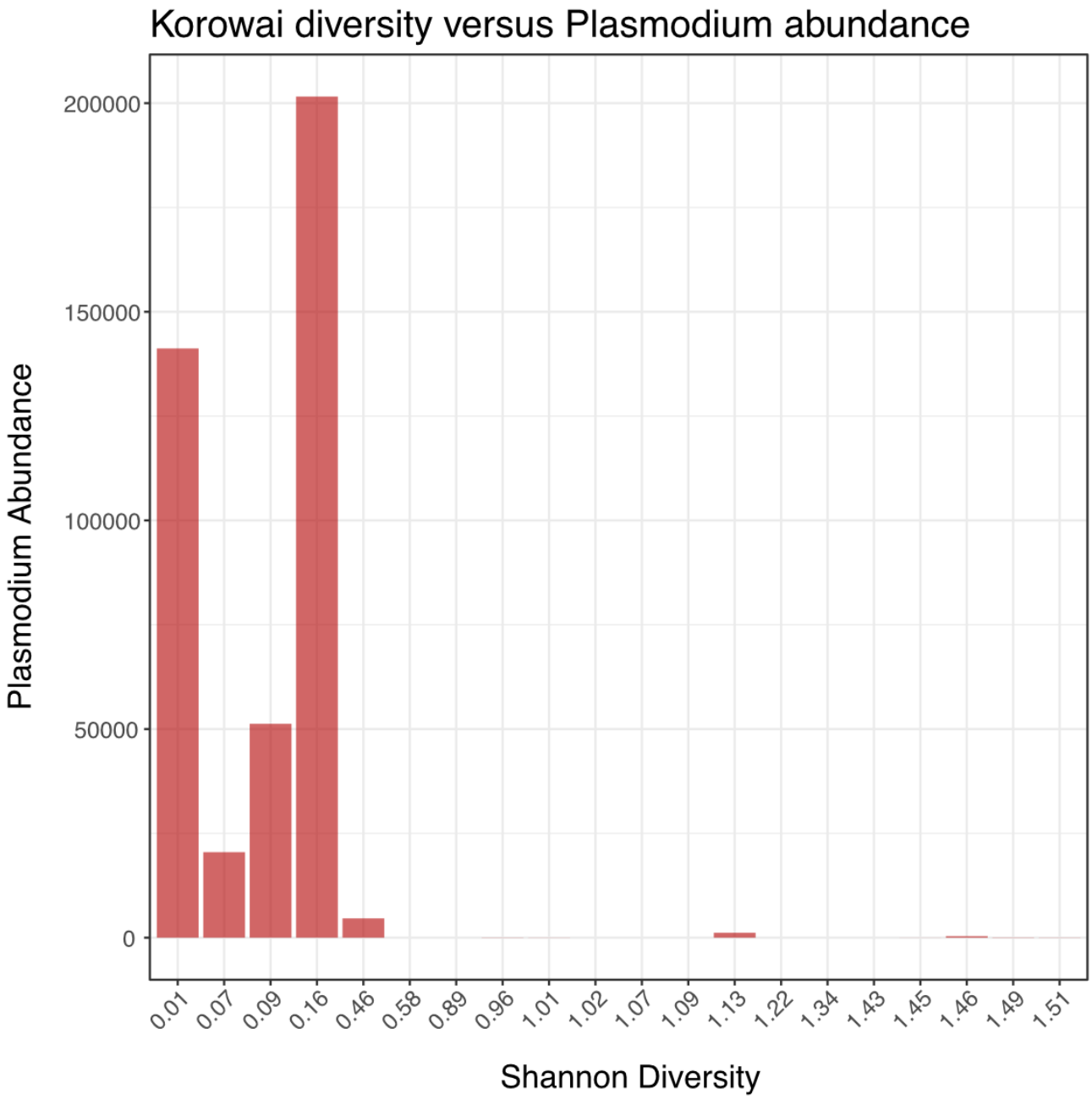
Supplementary Figure 4. Scree plot showing the percentage of variance explained for the first 20 principal components.



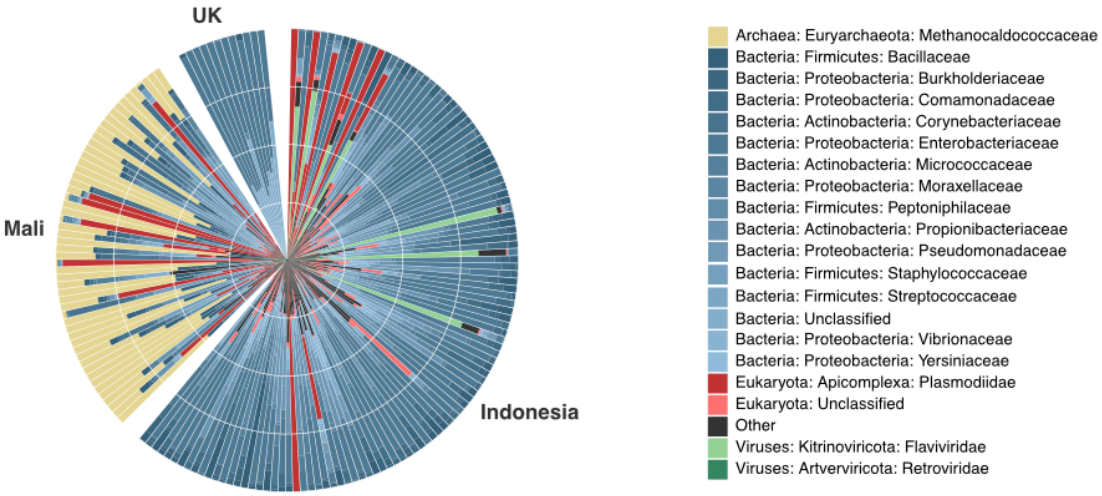
Supplementary Figure 5. PCA of taxa abundance at the phylum level, highlighted by logged abundance of Actinobacteria.



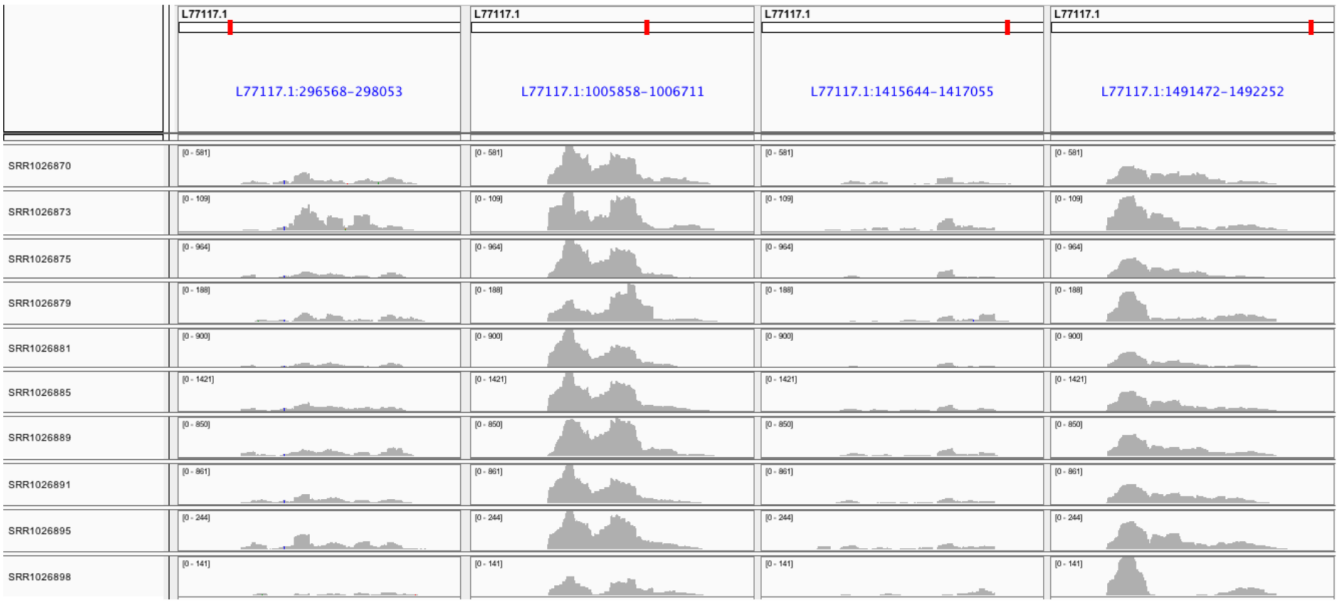
Supplementary Figure 6. Alpha diversity estimates for Indonesian island populations. A) Estimates of Shannon and B) inverse Simpson diversity within each population. KOR = Korowai; MTW = Mentawai; SMB = Sumba



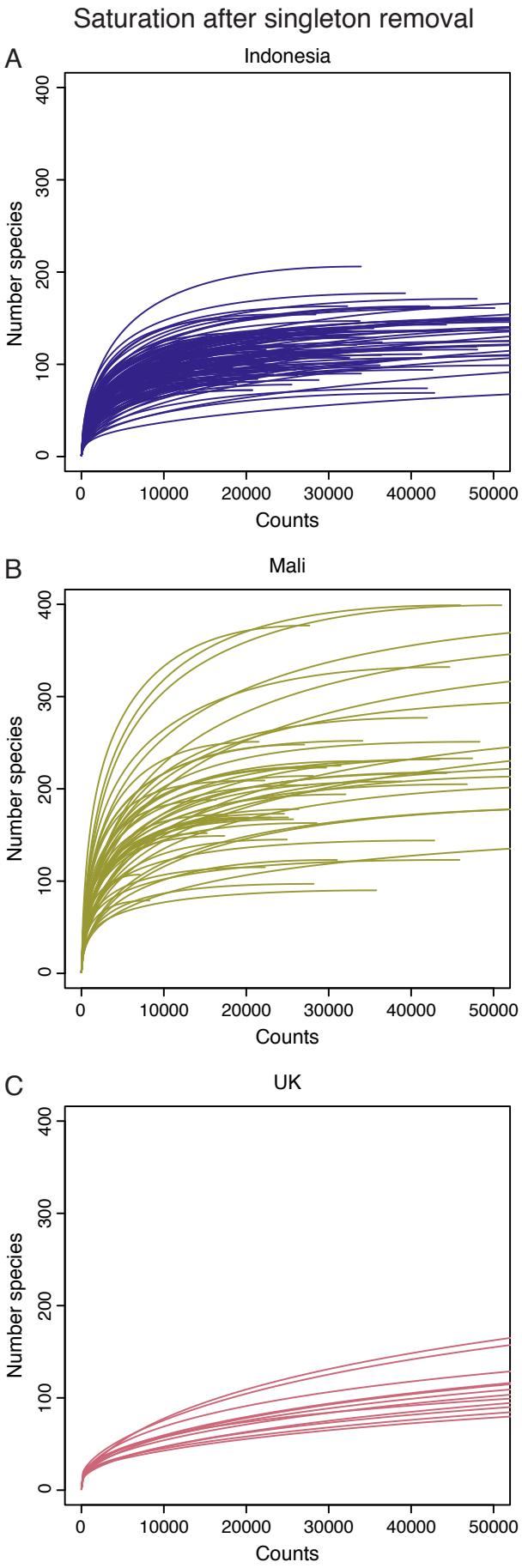
Supplementary Figure 7. Plasmodium abundance for each Korowai individual. Individuals are ranked from lowest Shannon diversity (on the left), to highest diversity.



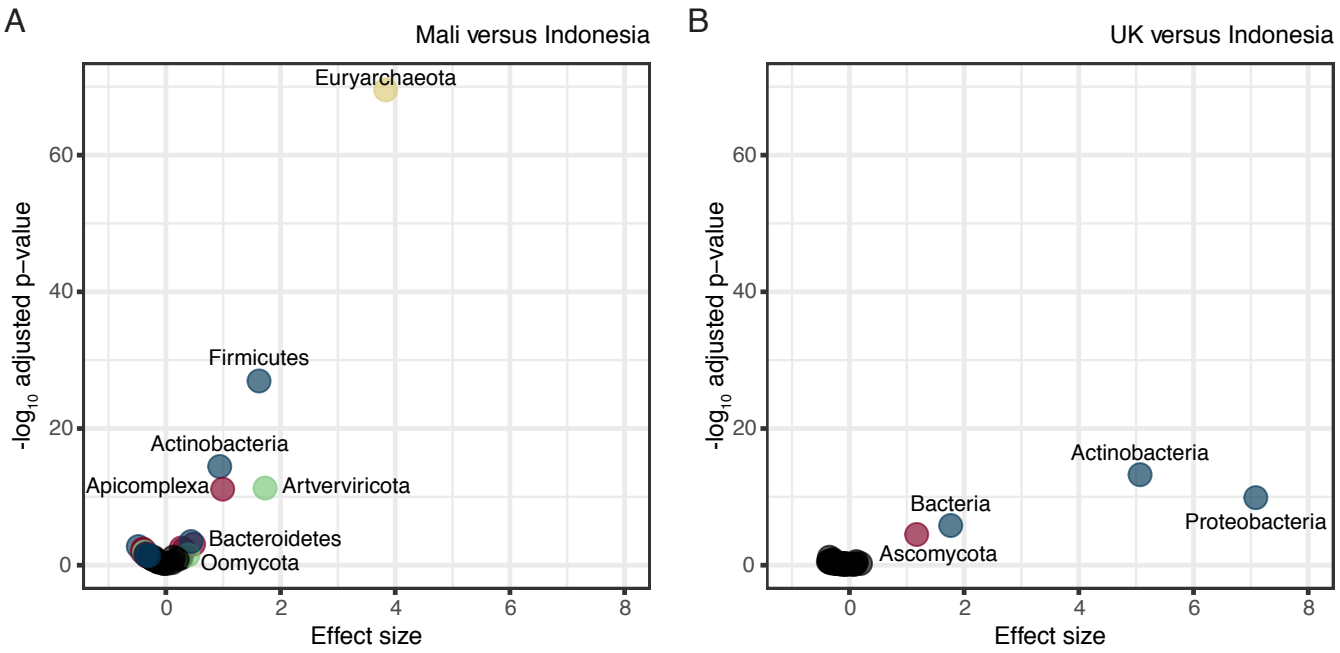
Supplementary Figure 8. Relative abundance of the top 20 taxa within the Indonesian, Malian, and UK dataset at the superkingdom, phylum, and family level. Bacteria are shown in blue, eukaryotes in red, viruses in green, and archaea in yellow.



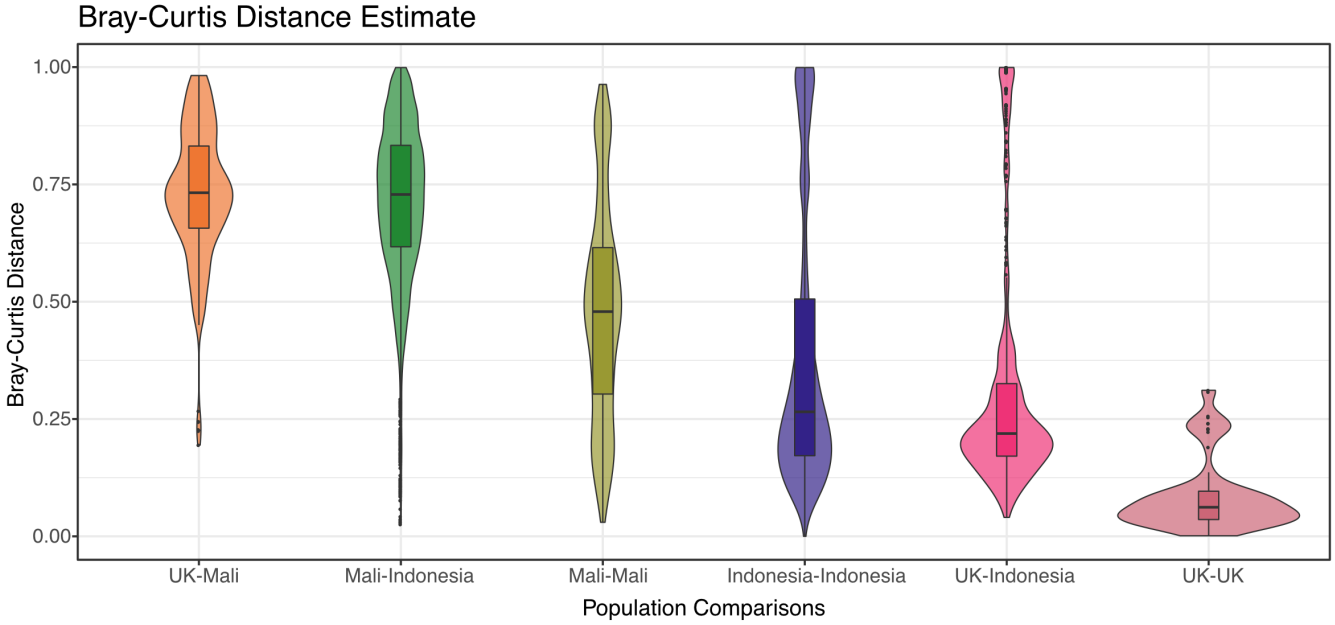
Supplementary Figure 9. Processed, unmapped Malian reads spanning the *Methanocaldococcus jannaschii* genome, visualised using IGV. Each row indicates read coverage for the first ten Malian samples in the dataset, while each pane (column) indicates a genomic region.



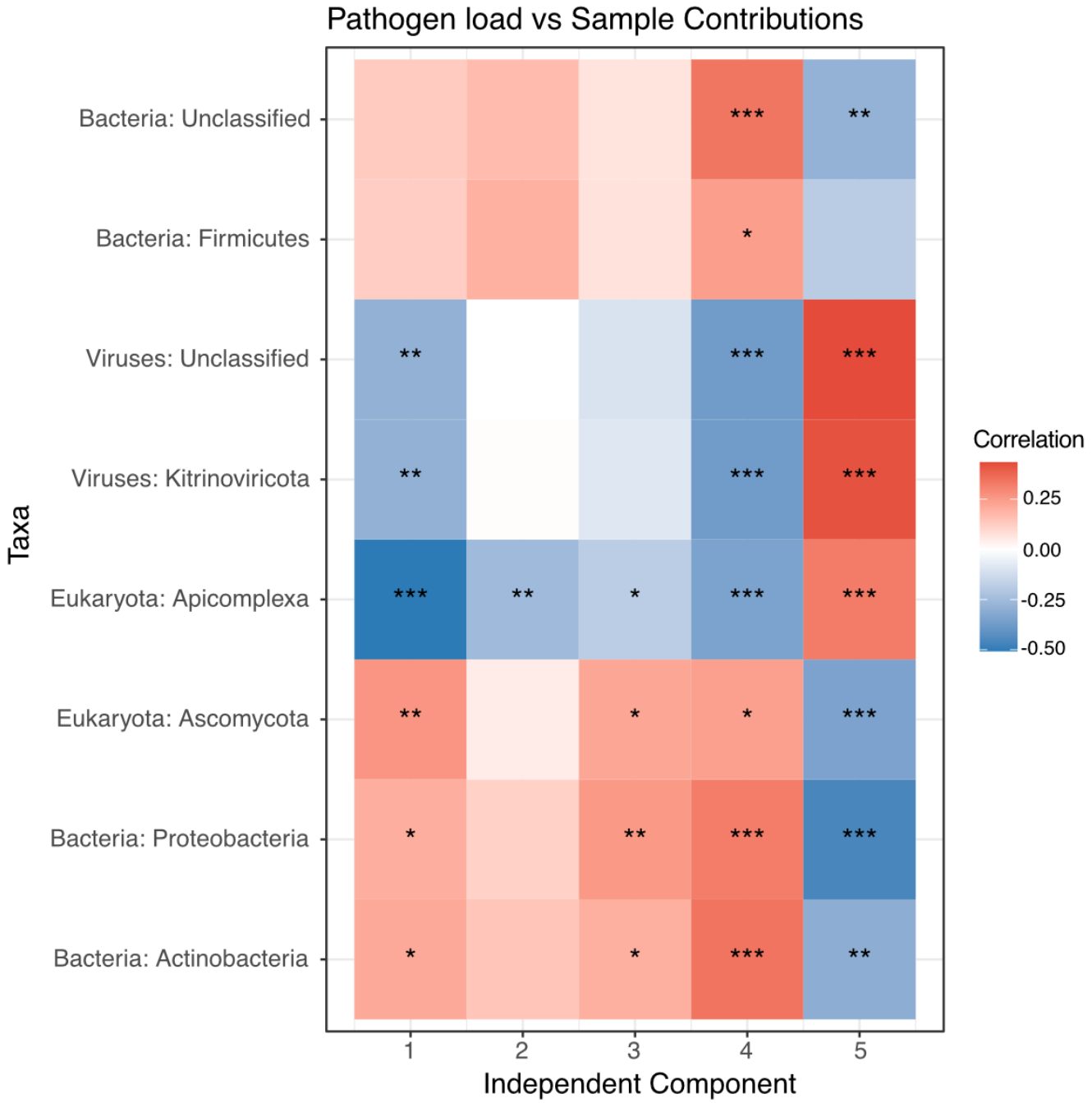
Supplementary Figure 10. Rarefaction curves of species saturation per individual at varying read depths for the A) Indonesian B) Malian and C) UK populations.



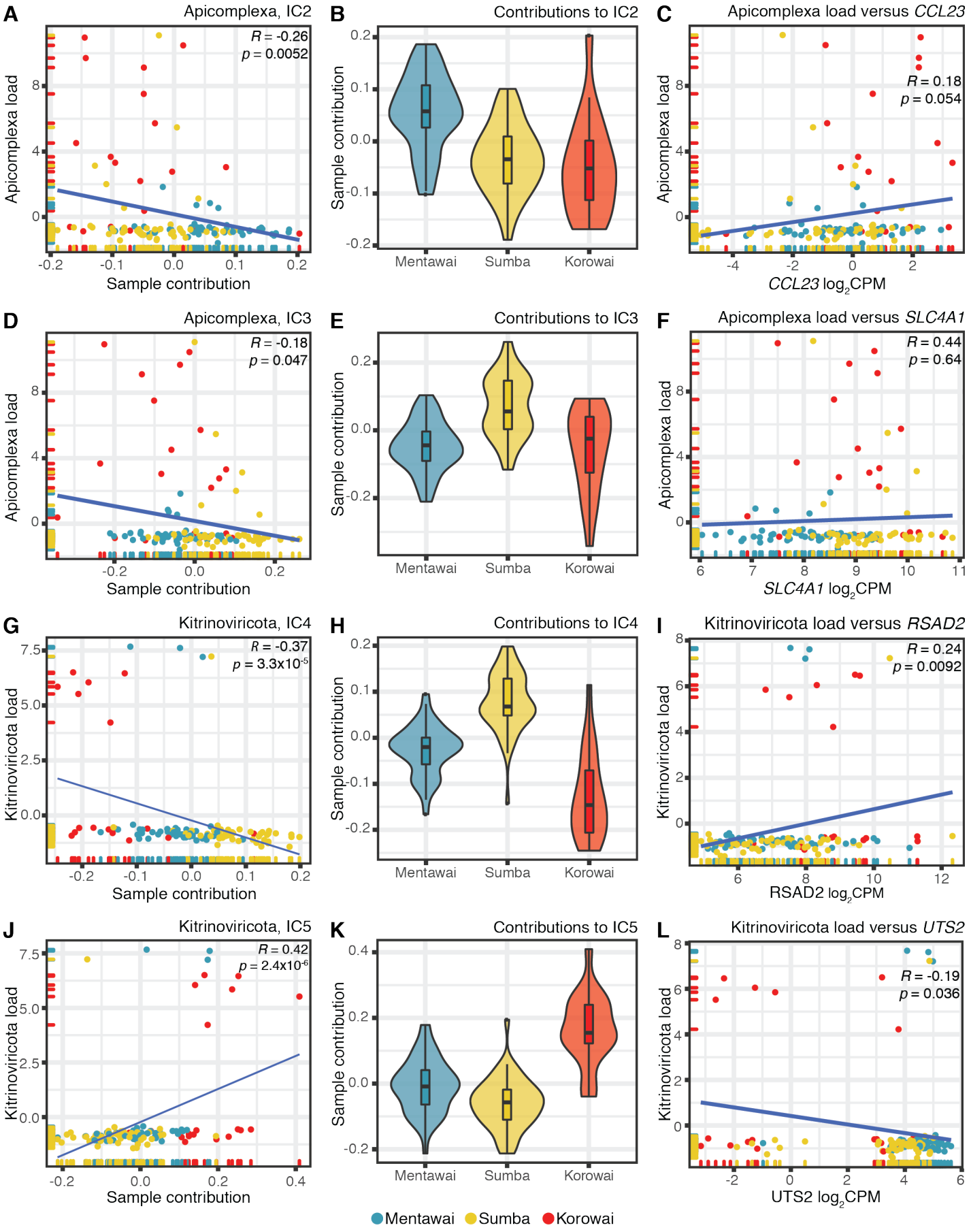
Supplementary Figure 11. Taxa differences between Indonesian and other global populations. A) Volcano plot of BH adjusted p-values from Welch's t-test for each phyla in the Malian versus Indonesian populations and B) UK versus Korowai populations. Taxa with a BH-corrected p-value below 0.05 for are coloured by superkingdom (red: eukaryotes; blue: bacteria; green: viruses; yellow: archaea).



Supplementary Figure 12. Bray-Curtis distance estimates for Indonesian, Malian, and UK population comparisons at the phylum level.



Supplementary Figure 13. Significant pathogens within each IC. Each row shows a pathogen with a significant correlation between pathogen load and sample contribution for each IC (in columns). Positive correlations are shown in red, while negative correlations are shown in blue. BH-adjusted Pearson's p-values are indicated by stars (0.05, 0.01, and 0.001 for one, two, and three stars, respectively).



Supplementary Figure 14. Caption on next page

Supplementary Figure 14. Independent component analysis for IC2 and IC5. A) Distribution of sample contributions for each island within IC2 B) Correlation of CLR-normalised Apicomplexa load and sample contributions for IC2 C) Correlation of CLR-normalised Apicomplexa load and *CCL23* $\log_2$ CPM values, the gene with one of the highest contributions to this IC, across all samples D) Correlation of Apicomplexa load and sample contributions for IC3 E) Distribution of sample contributions for each island within IC3 F) Correlation of CLR-normalised Apicomplexa load and *SLC4A1* $\log_2$ CPM values across all samples G) Correlation of Kitrinoviricota load and sample contribution to IC4 H) Distribution of sample contributions for each island within IC4 I) Correlation of CLR-normalised Kitrinoviricota load and *RSAD2* $\log_2$ CPM values across all samples. J) Distribution of sample contributions for each island within IC5 K) Correlation of Kitrinoviricota load and sample contributions for IC5 L) Correlation of CLR-normalised Kitrinoviricota load and *UTS2* $\log_2$ CPM values, the gene with the highest contribution to this IC, across all samples