2019 APRIL Biosciences The University of Melbourne

Katalina Bobowik

2019/05/14

Our understanding of function within the human genome has been largely limited to populations of European ancestry. In order to better characterise human functional diversity, this study aims to analyse regulatory variation in three isolated populations within the Indonesian archipelago. No study has yet analysed regulatory variation in Indonesia and the small, isolated nature of these populations makes them ideal to identify variants that have risen to higher frequencies. In this study, I identify variants influencing gene expression through the analysis of RNA sequencing data with a focus on pathways associated with immune function and disease.

Contents

List of Figures

1

Introduction

Humans, during their migration through diverse ecological settings, have experienced unique pathogenic and environmental pressures, consequently shaping genomes through microevolutionary processes. Although population genetic studies and genome-wide scans have provided new insight into the demographic and selection history of our species, the majority of variants found have been in non-coding regions of the genome [1, 19, 28] making the biological processes they affect difficult to identify. In the past decade, the endeavour to understand genome biology and function has been greatly aided by gene expression studies in which regulatory variants can be matched to RNA expression data. This has facilitated our understanding of complex traits [1, 7], enabled us to identify variants associated with genetic disease

[13], and understand the relationship between genetic variation and immune phenotypic diversity [5, 10]. To date, the vast majority of these studies have focused on populations of primarily European descent. Given the diverse demographic history of our species, this fails to capture the full landscape of human regulatory variation [17, 21]. This study will focus on identifying regulatory variation in small, traditional populations of Southeast Asia in order to better understand phenotypic diversity and traits associated with immune function and disease.

1.1 Study group

Our study groups come from three small island populations within Indonesia: Mentawai, Mappi, and Sumba. The first two islands, Mentawai and Mappi, capture two major genetic ancestries within Indonesia—Austronesians (Mentawai) and Melanesians (Mappi)—and also provide possible source populations to the last island, Sumba. This island is host to a network of traditional communities that derive from pre-existing Melanesians, arriving on the island 50,000 years ago, and incoming Asian farming cultures, arriving 4,000 years ago

1.2 Previous studies analysing regulatory variation outside of Europe

Although previous gene expression studies have identified factors influencing population-level variation in gene regulation [2, 26], few have analysed them in non-European populations [16, 21]. One such study to analyse transcriptome variation across a globally distributed set of populations was performed by Martin et al. [16]. In this analysis, genome, exome, and transcriptome sequencing data for seven global populations was generated from lymphoblastoid cell lines derived from 45 individuals in the Human Genome Diversity Panel (4-7 samples per population). The authors found that approximately 25% of the variation in gene expression could be attributed to population differences and were able to identify novel gene structures. Although the small sample size of the study lacked the power to detect additional population-specific variants, it highlighted the role of demography in understanding functional genetic variation. Furthermore, this is one of the only studies to analyse RNA sequencing (RNA-seq) data on a Southeast Asian population. The other such study to do so to our knowledge, and the only study to analyse RNA-seq data from a population within Indonesia, was conducted by Yamagishi et al. [27]. This study analysed expression data collected from peripheral blood from 116 individuals infected with malaria. Despite having a population of over 260 million people, hosting hundreds of different ethnic groups, and covering the same area as continental Europe [12], no study has analysed expression data from healthy Indonesian samples.

Along with a lack of studies analysing variation outside of Europe, most studies use large cohorts of well-studied populations for transcriptomic analysis. This fails to account for the genomic variation in small, structured populations that experience different evolutionary dynamics. Furthermore, small populations allow for the detection of rare variants that rise in frequency by chance, providing more statistical power for identification of functional alleles [25].

1.3 Project significance and Aims

Understanding how genes and regulatory networks operate is one of the biggest challenges in genomics and public health science. Not only do we have a limited understanding of the function of the human genome, but the regions we have characterised have been represented by only a small subset of all human genetic variation. Characterising this variation can ultimately lead to more accurate diagnosis and specific disease treatment, as well as a more thorough understanding of biological processes. Here, I analyses regulatory variation in 117 individuals who share similar environmental pressures yet have different ancestries in order to better understand how ancestry-related variation contributes to immune gene regulation.

2

Methods

2.1 Sample collection and quality control

A general overview of the differential expression pipeline is outlined in Figure 2.1. Whole blood was collected by trained phlebotomists from the Eijkman Institute from over 300 Indonesian males. Samples were collected across multiple villages in the three islands using Tempus Blood RNA Tubes (Applied Biosystems) and extracted according to the manufacturer's protocol. All extraction and sequencing batches were randomised with respect to village and island. RNA quality was assessed with a Bioanalyzer 2100 (Agilent) and concentration through Qubit (Life Technologies). On the basis of these results, we selected the five villages in our study for having the highest average RIN and the largest number of individuals with a RIN >= 6.5. Library prep was done by Macrogen (South Korea), using 750ng of RNA and the Globin-Zero Gold rRNA Removal Kit (Illumina).

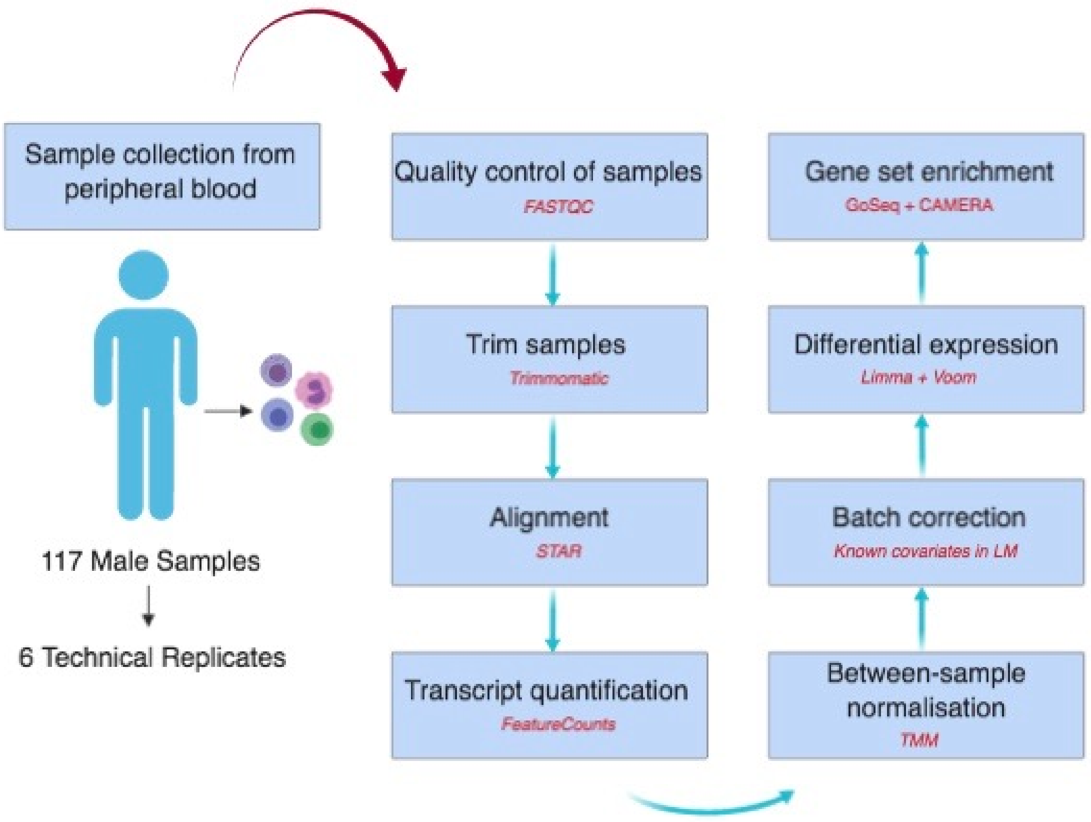

Figure 2.1: Peripheral blood samples from 117 Indonesian males (plus 6 technical replicates) were collected by trained phlebotomists from the Eijkman Institute. My sample pipeline consisted of an initial quality control step using FastQC, trimmiming samples using the software Trimmomatic, aligning reads using STAR, and then counting the number of transcripts using featureCounts. I used a Limma+Voom differential expression pipeline that used TMM normalisation for between-sample normalisation, batch correction incorporating covariates known to affect expression levels into a statistical model, and then gene ontology testing using GoSeq and Camera.

101-bp paired-end reads were generated by Illumina sequencing (HiSeq 25000). This resulted in a total of 117 individuals, including six technical replicates making a total of 123 sequences, sequenced in three batches (Supplementary Table 5.1). This consisted of 48 individuals from the island of Mentawai (plus 1 replicate), 49 individuals from the island of Sumba (plus 4 replicates), and 20 individuals from the island of West Papua (plus one replicate). Upon data retrieval, all RNA sequencing reads went through an initial sample quality analysis using FastQC V0.11.5 (Babraham Institute, available for download at http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Base quality decreased near the 3' end of reads, as is common with Illumina sequencing, and therefore all leading and trailing bases below a Phred quality score of 20 were removed using Trimmomatic version 0.36 [4].

2.2 Alignment and transcript quantification

RNA-seq reads were aligned to the human genome (GrCh38, Ensembl release 90: August 2017) with STAR version 2.5.3a [8]. I used a two-pass alignment to improve identification of novel splice junctions which resulted in a mean of 29 million uniquely-mapped read pairs per sample (Figure 2.2, Supplementary Table 5.2). Next, I performed read quantification with featureCounts version 1.5.3 [15] and used a filtered GTF file for GRCh38.p10 that I created using the biomaRt R package. As some studies have suggested that prefiltering improves the performance of differential expression analysis by controlling the false discovery rate [24], I set filtering parameters to only include GENCODE basic annotation and transcript support levels 1-3 (i.e., transcripts with at least one non-suspect mRNA or expressed sequence tag). Since many transcripts are poorly supported, this method provides a way of including transcript annotations that are actually expressed in humans. I used the default settings for featureCounts paired end reads, where features of exons are grouped into meta-features of genes. On average, I successfully assigned 15 million paired-end reads to each sample (Figure 2.2, Supplementary Table 5.2).

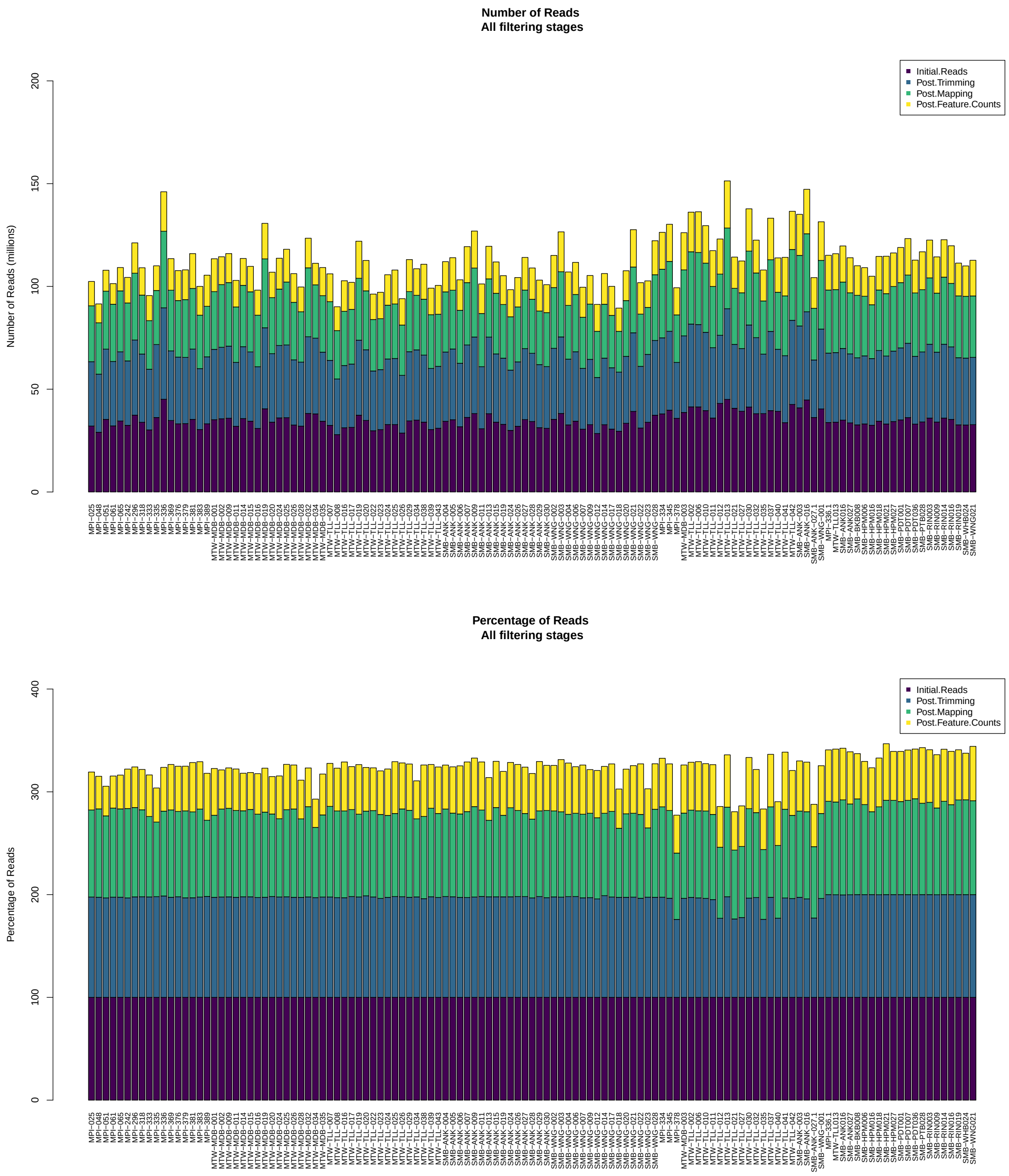

Figure 2.2: Number (top) and percentage (bottom) of reads for each sample at each step of the RNASeq pipeline.

2.3 Technical replicate analysis

Six replicates were used in this study to ensure batch effects could be measured and removed: one sample was replicated twice in batches two and three (SMB-ANK-027) and the other four samples were replicated in batch three (MPI-381, MTW-TLL-013, SMB-ANK-016, SMB-WNG-021; Supplementary Table 5.1). I compared all samples to their corresponding replicate by plotting the correlation between the log2-CPM normalised counts after removal of lowly-expressed genes and then fitting a best-fit regression line (Figure 2.3). The correlation between samples after log2 transformation and removal of lowly-expressed genes ranged from an R-squared value of 0.97-0.99, suggesting high similarity between all replicates.

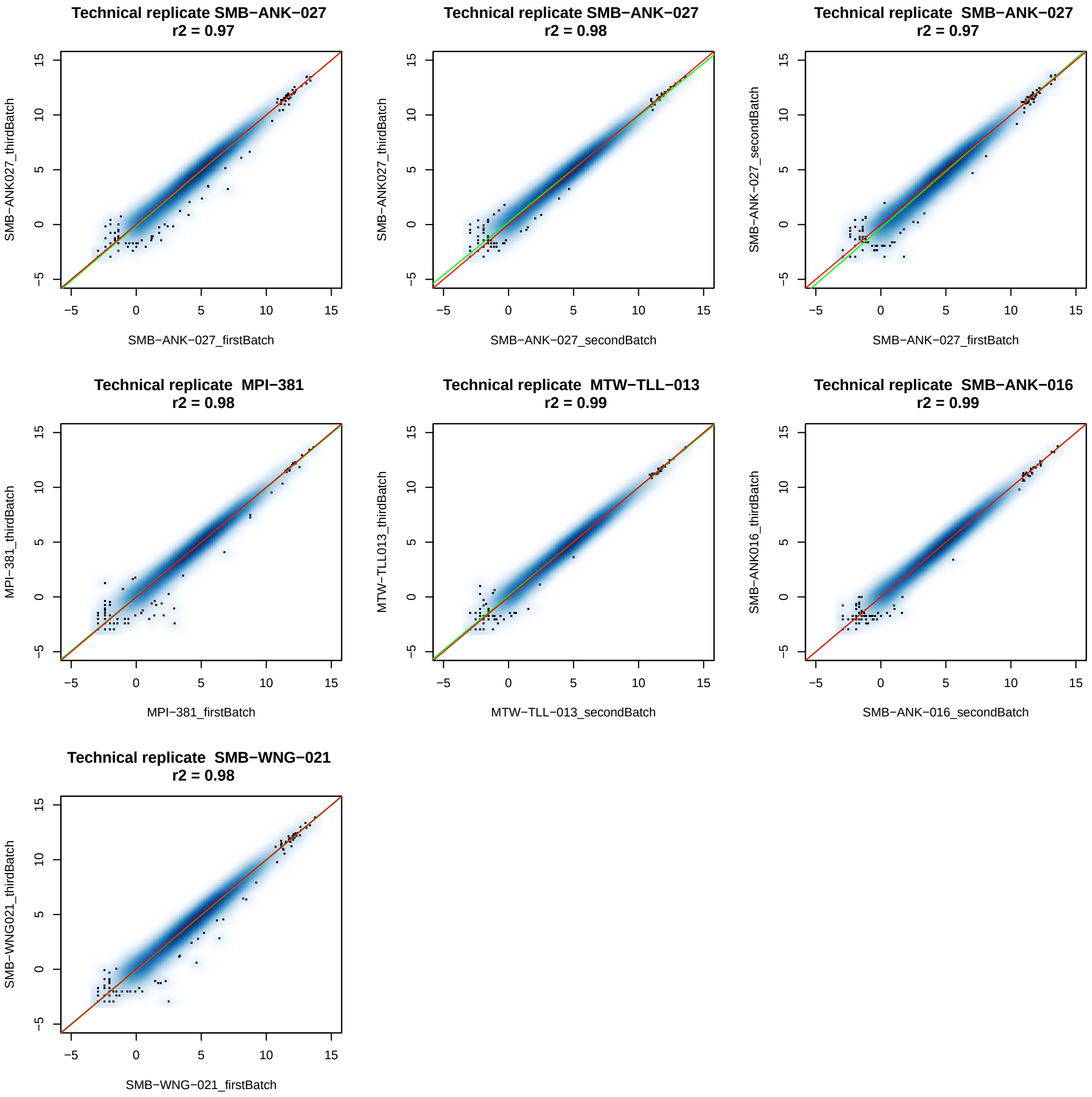

Figure 2.3: Performance of each control compared to its corresponding technical replicate sample after removal of lowly-expressed genes. All controls have one corresponding technical replicate with the exception of SMB-ANK-027, which has two (shown in panes 1, 2, and 3).

2.4 Removal of lowly-expressed genes and data normalization

To account for library size differences, I transformed expression counts to log2-counts per million using a prior count of 0.25 in order to avoid taking the log of zero. I then removed lowly expressed genes by only keeping genes expressed at or over a CPM of one in at least one of the island groups, resulting in a total of 12,975 genes (Figure 2.4, Supplementary Figure 5.1). After filtering, sample library sizes ranged from roughly 9 million to 23 million reads (Figure 2.5).

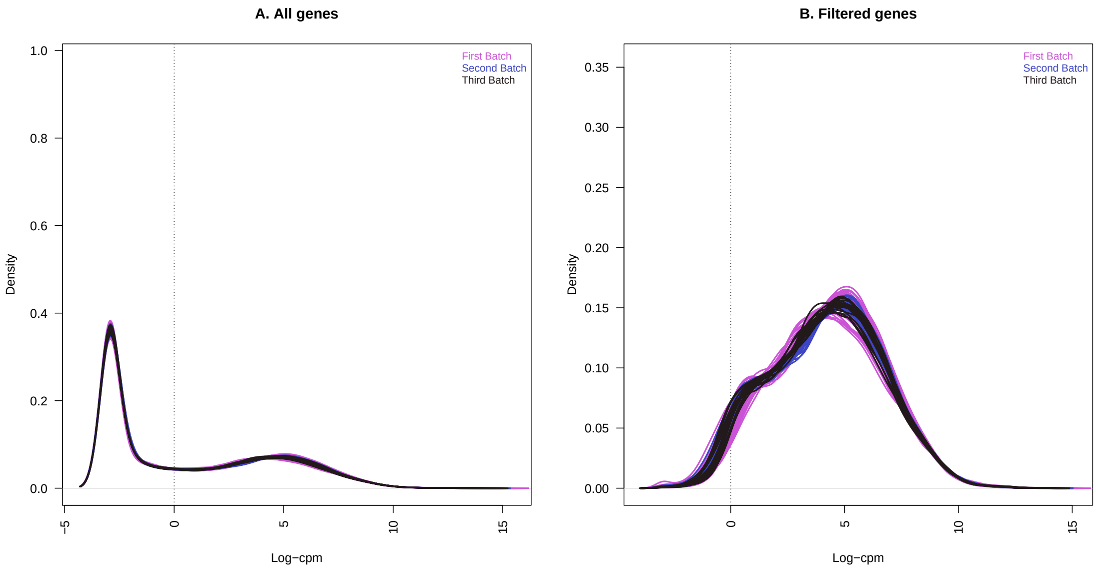

Figure 2.4: Gene expression distributions before and after filtering for lowly-expressed genes. The density of log2 counts per million (CPM) are shown for each sample with pink lines indicating samples from batch 1, purple lines indicating samples from batch 2, and black lines indicating samples from batch 3. The left panel shows data from 27,413 genes and the right panel shows data from 12,975 genes after filtering (i.e., only keeping genes expressed at or over a CPM of one in at least one of the island groups).

In order to eliminate composition bias between libraries, I tested three different normalization methods: upperquartile, trimmed mean of M values (TMM), and relative log expression (RLE). Visual inspection of all three methods showed similar performance of normalization (Supplementary Figure 5.2). A test of the TMM normalization was therefore chosen as it has been shown to have a low false discovery rate and high true positive rate [20]. Furthermore, it is also recommended for RNA-Seq data where over half of the genes are not believed to be differentially expressed [6]. In order to ensure composition bias was removed, I examined the performance of TMM-normalisation by plotting mean-difference (MD) plots for each sample. Most genes for each sample were centered around a log-fold change of zero, indicating that the composition bias was successfully removed (Figure 2.6).

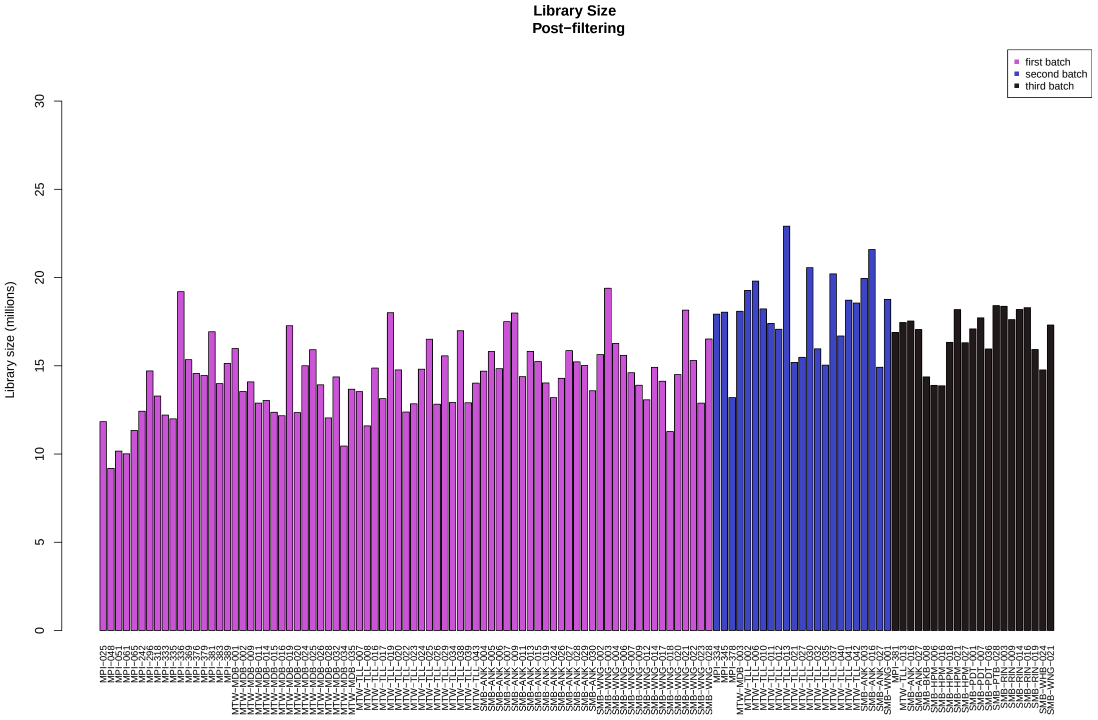

Figure 2.5: Total number of reads after filtering for lowly-expressed genes. The total library size ranged from 9 million 23 million read pairs.

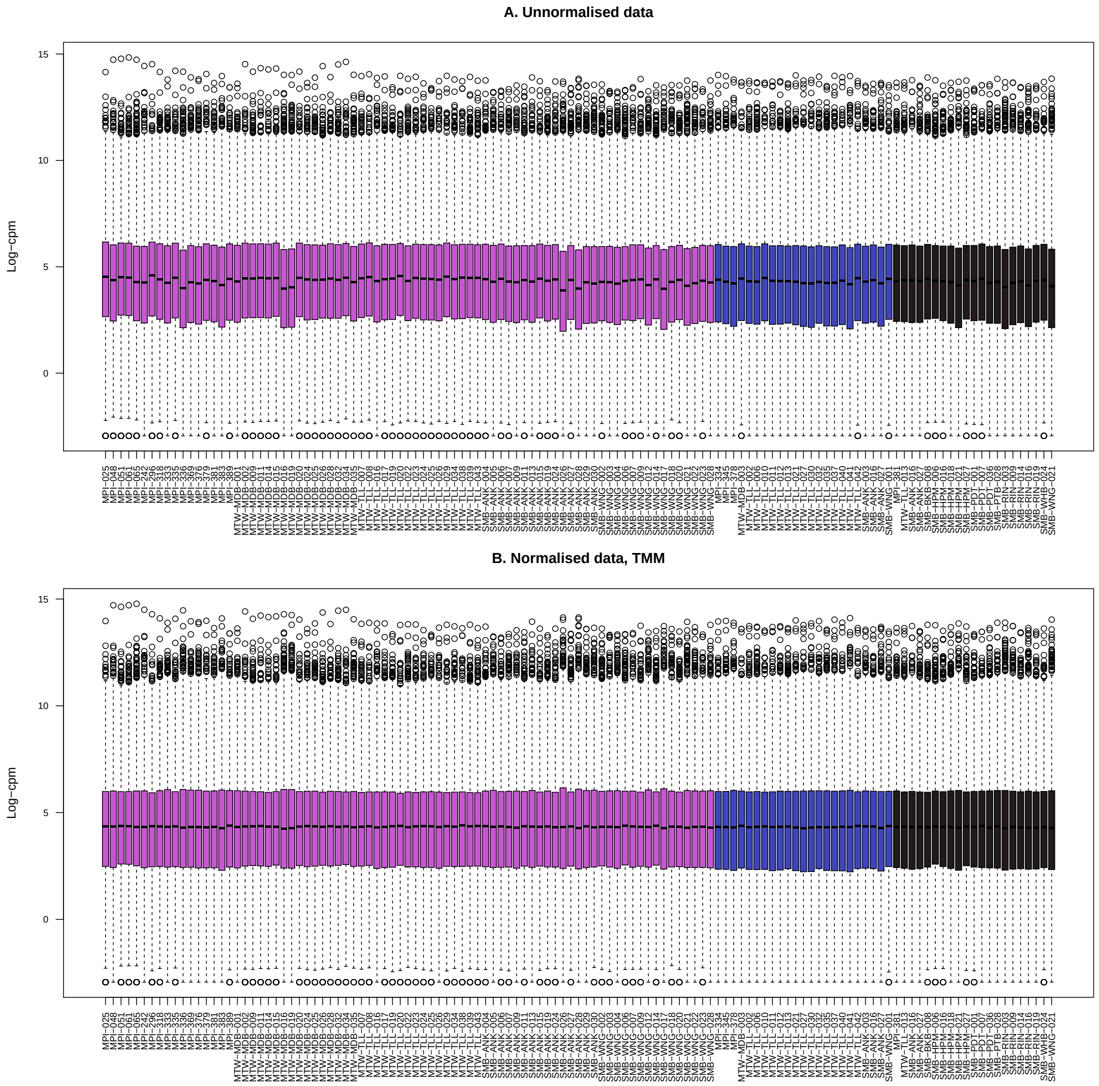

Figure 2.6: Performance of TMM normalisation on the log2-CPM transformed samples. Pink indicates sample from batch 1, purple indicates samples from batch 2, and black indicates samples from batch 3.

2.5 Sample variation analysis and data exploration

In order to identify outlier samples and assess which covariates might be associated with expression levels, I used principal component analysis on the TMM-transformed log2 expression values (for a full list of covariates, see Supplementary

Table 5.3). Associations between covariates and expression levels were analysed by inspecting the first ten dimensions in the data for clustering within each group. To identify significant associations ($p<0.05$) between covariates and expression levels, I performed ANOVA tests between covariates and each dimension of the PCA. The first PCA was most strongly associated with batch (Figure 2.7), followed by significant associations with island, sampling site, RIN score, CD8T cells, and granulocytes (Figure 2.8; for a full list of significant associations, see Supplementary Table 5.4). Language, sampling date, sequencing pool, the Mappi tribe, and season also had significant associations, however these covariates were found to be confounded with other variables. In the second dimensions, blood cell type—in particular, granulocytes—had the highest correlation with expression levels.

Hierarchical clustering by Euclidean on the TMM-normalised, log2-transformed samples was performed in order to identify any sample outliers and analyse sample grouping (Figure 2.9). Samples were shown to mostly cluster by island and no clear outliers were identified. Furthermore, between and within-island variation was analysed using Pearson correlations between log2-CPM expression values of samples (Figure 2.10). As expected, most variation was seen between islands. Mappi was shown to have the highest amount of within-island variation and island populations compared to Mappi also had a higher degree of variation.

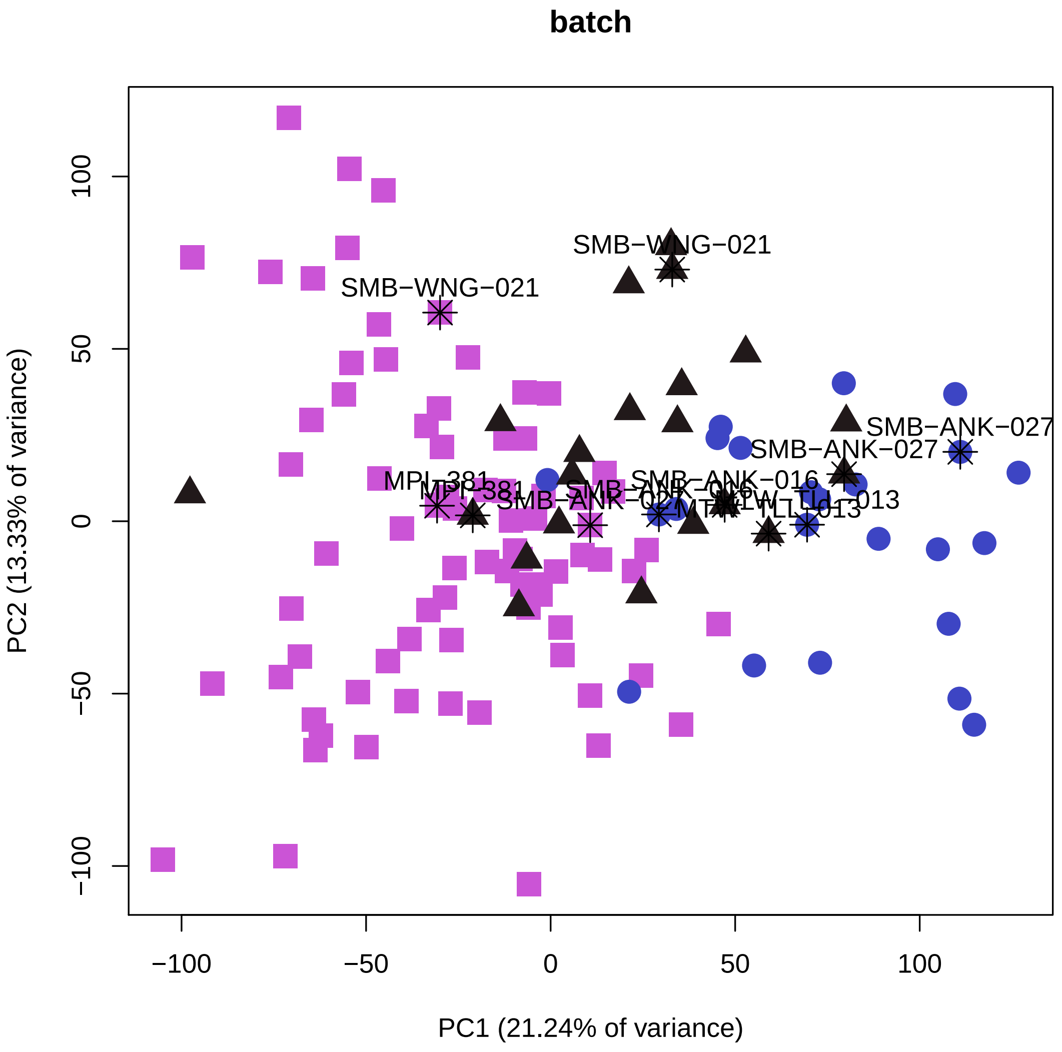

Figure 2.7: Principal component analysis of the TMM-normalised, log2-transformed counts. Pink indicates sample from batch 1, purple indicates samples from batch 2, and black indicates samples from batch 3. Separation by batch can be seen in the first and second dimension of the PCA. Replicate samples are labelled with stars.

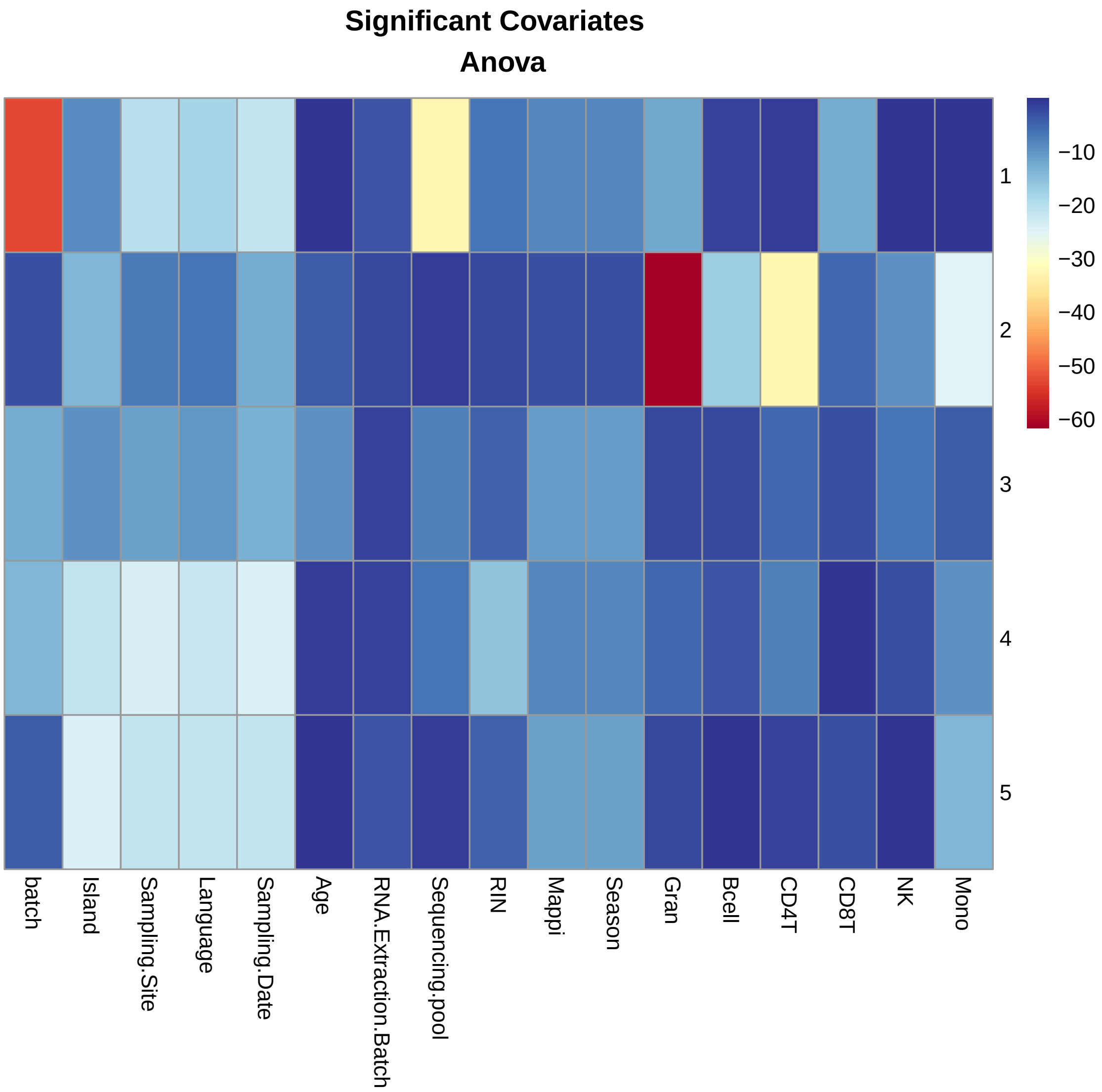

Figure 2.8: Heatmap of significant covariates of the first 5 PCA dimensions(p=0.05). The covariate that had the strongest association with expression levels was batch, followed by granulocytes, CD8T cells, island, and RIN score.In the second dimension, all blood cell types(CD8T, CD4T, NK, B cells, monocytes, granulocytes) and island had the highest correlation with expression levels.

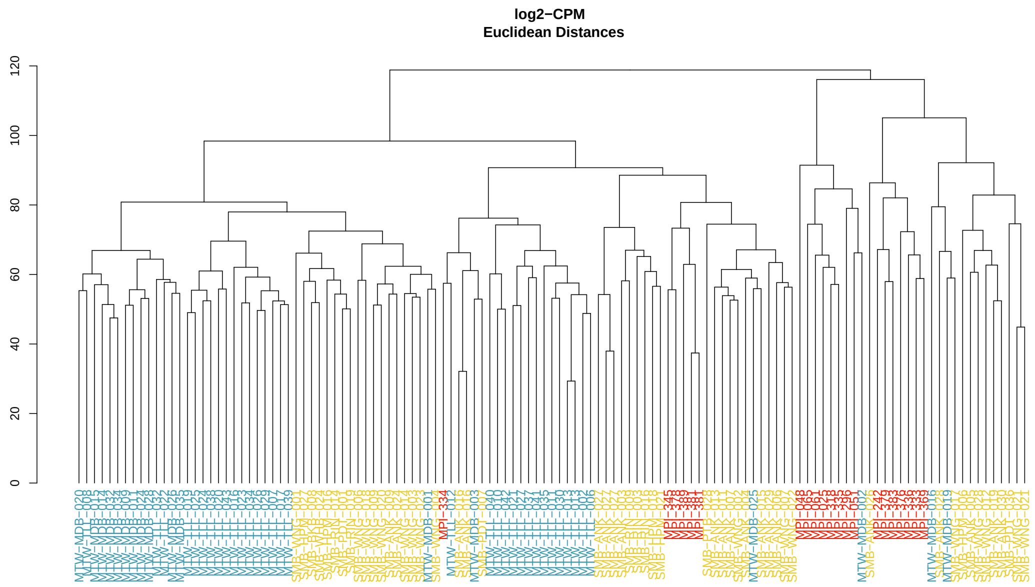

Figure 2.9: Hierarchical clustering of log2-transformed CPM values of TMM-normalised counts. Yellow indicates samples from Sumba, red indicates samples from Mappi, and blue indicates samples from Mentawai. Clustering can be seen predominantly by island.

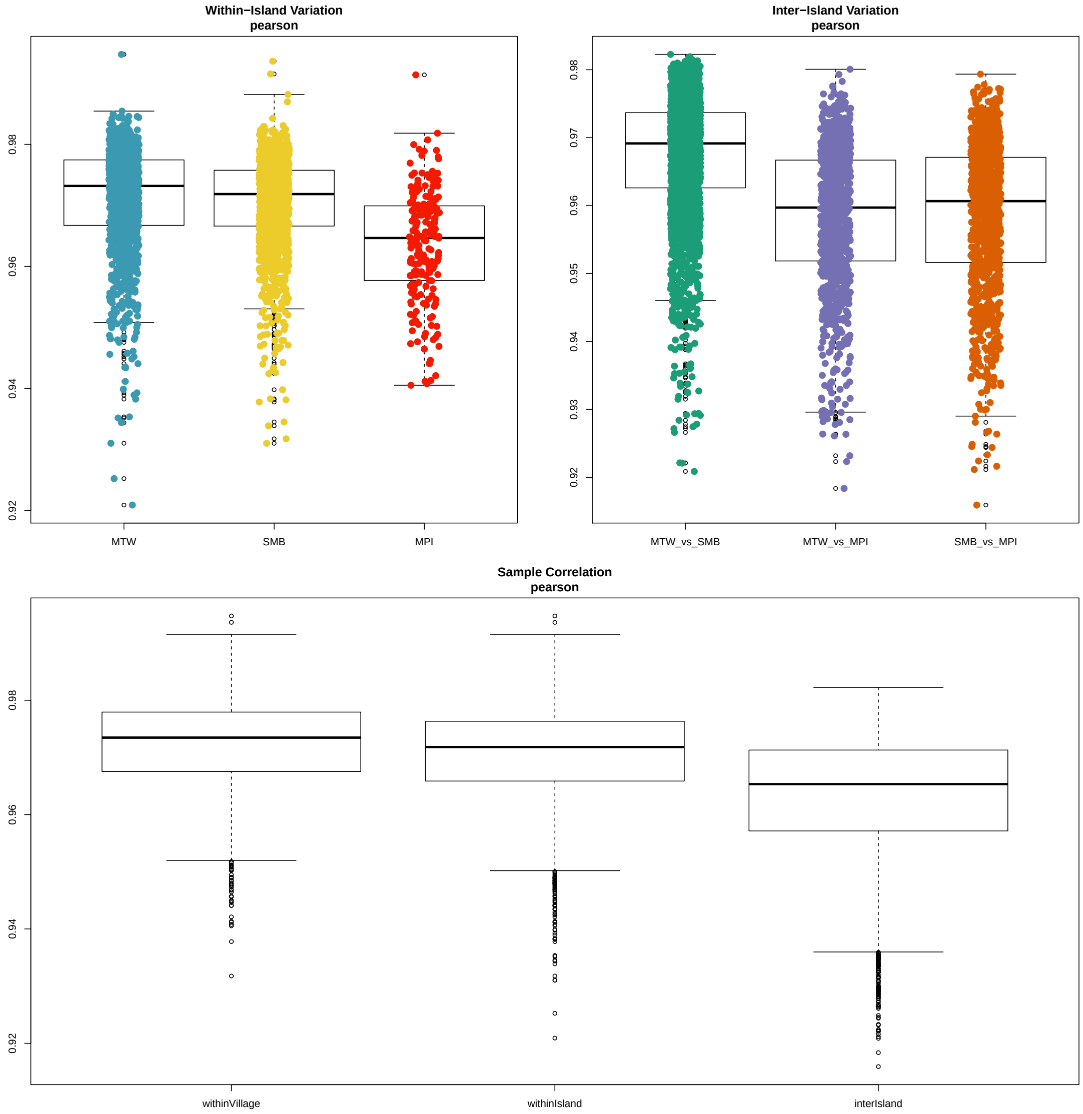

Figure 2.10: Within and between island variation in all three island populations using Pearson correlations of the log2-transformed, TMM-normalised CPM values.

2.6 Differential expression analysis

Because clear clustering was observed by batch, age, RIN score, and blood type (CD8T, CD4T, NK, B cells, monocytes, and granulocytes), I incorporated these known covariates into the design matrix of my linear model. The final model used to test for differential expression was therefore:

0 + Island + Age + batch + RIN + CD8T + CD4T + NK + Bcell + Mono + Gran

where island is the covariate of interest and Age, batch, RIN, and blood types are the covariates affecting expression levels. I then set up a list of contrasts comparing each island population, which consisted of the following contrasts: Sumba versus Mentawai, Sumba versus Mappi, and Mentawai versus Mappi. To test the number of differentially expressed genes at the village level, I replaced the covariate of interest with sampling site and set up my contrasts so that each village was contrasted with every other village. Furthermore, 16 samples from the Sumba population were removed which had 5 or fewer samples in each village.

I removed high sample variability from the count data using Voom-normalisation on the TMM-normalised sample counts. Voom estimates the mean-variance relationship of log2 CPM-normalised counts and generates a precision weight for each sample [14]. This approach enables poor-quality samples to be down-weighted and effectively eliminates the mean-variance relationship (Figure 2.11). Since technical replicates were used in the study design, the limma function duplicateCorrelation was used in order to estimate the correlation between expression measurements made on the same subject [23]. DuplicateCorrelation takes information from replicated samples and uses an empirical Bayes method to moderate the standard deviation between genes [23]. After running duplicateCorrelation on the samples, I then ran Voom a second time to incorporate the technical replicates as blocking variables and to feed in the estimated correlation within the blocks. A simple linear regression model was then fit to the voom output for each gene and an empirical Bayes moderated t-statistic was used to test each individual contrast from the design matrix equal to zero. When fitting the models, I used the 'robust' parameter in Limma's 'eBayes' function in order to deal with outlier genes that have abnormally high or low variance. I then identified significant genes using an adjusted p-value of 0.01 (Benjamini–Hochberg corrected [3]) and an absolute log-fold change of 1.

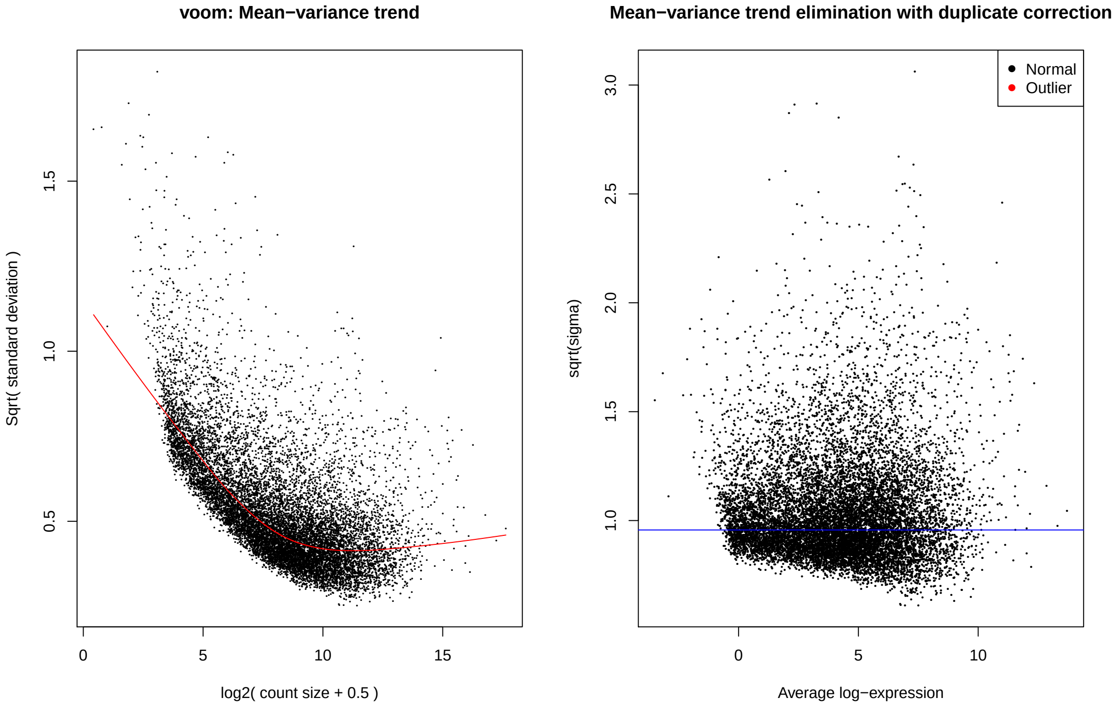

Figure 2.11: Mean-variance trend and elimination of the trend after applying Voom-normalisation to the TMM-normalised count data.

2.7 RUVs

In addition to removing batch effects by regressing out known variables affecting gene expression, I also tested an alternative unsupervised batch-correction method using RUV [9, 22]. RUV uses factor analysis to control for technical effects based on control genes or samples. For this analysis, I chose RUVs which utilizes counts of negative control samples. Unlike RUVg which uses control genes to correct for technical effects, RUVs is less sensitive to the control genes used in the factor analysis [22]. Because of this and the availability of technical replicates, RUVs was chosen as the RUV correction method. The first step of RUVs is to set up a count matrix using raw read counts. I did this using all raw reads filtered for lowly-expressed genes. I then performed upperquartile normalization on the dataset and checked the performance before and after correction (Figure 2.12).

In order to perform RUVs on the raw read counts, RUVs needs four pieces of information: 1) the object containing read counts, 2) a list of control genes, 3) k, the model parameter indicating the number of hidden factors of variation, and 4) replicate information. For the object containing read counts, I used the upperquartile-normalised reads filtered for lowly-expressed genes. I then used all genes within the dataset for control genes, as this has been shown to perform well in an RUVs pipeline [22]. For k, the number of hidden factors of variation, I tested values from one to six (the highest value of k is the total number of replicates in the study) and analysed the total amount of variation through RLE and PCA plots (Figure 2.13). A k of three to six was most effective in reducing the total amount of variation (Figure 2.13, panels e, g, i, and k) and minimizing distances between technical replicates (Figure 2.13, panels f, h, j, and l). In addition to looking at total variation, I also plotted the p-value distribution and total number of differentially expressed genes at an FDR of 0.01 and log fold change of 1 (Figure 2.14). For each population comparison, I chose the best value of k as the inflection point in the total number of differentially expressed genes. This changed for each population comparison (Sumba vs Mentawai = 4, Sumba vs Mappi = 5, Mentawai vs Mappi = 3) and therefore chose a k of 5 based on its performance in the PCA and RLE plots.

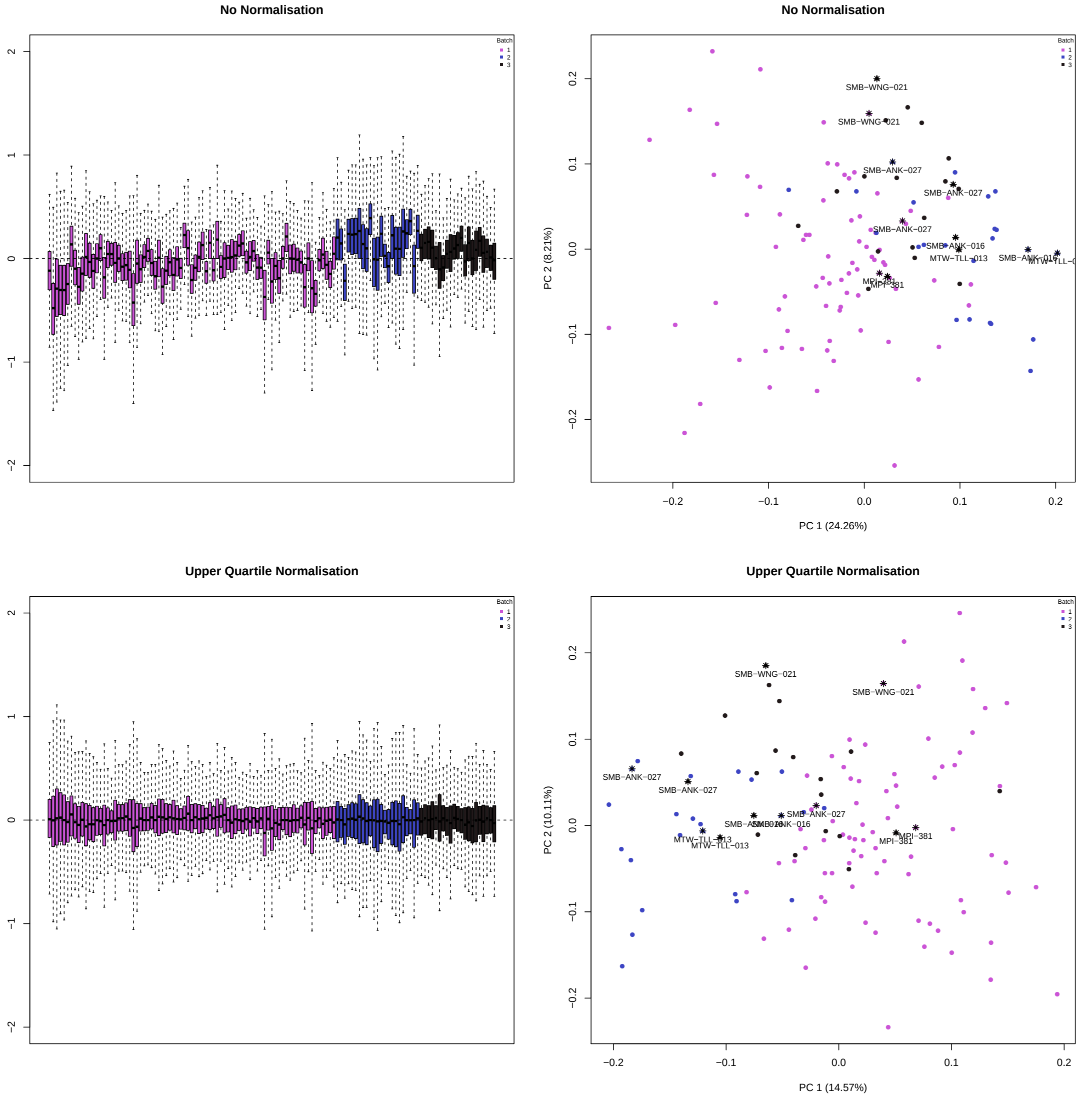

Figure 2.12: Sample variability of raw read counts before and after upperquartile normalisation. An RLE plot showing the total amount of variability in each batch (batch 1=pink,batch 2=purple, batch 3=black) can be seen before and after upperquartile normalisation (top and bottom left, respectively). Principal component analysis of the first two PCs before and after upperquartile normalisation (top and bottom right,respectively) are shown to cluster by batch. Sample replicates are highlighted with stars.

RUVs outputs estimated factors of unwanted variation and normalised counts, which are obtained by regressing out the original counts on the unwanted factors. After obtaining the factors of unwanted variation from RUV, I used the same negative binomial GLM approach, as done in the Limma pipeline, using known covariates. In my model, I included my covariate of interest as island and the estimated values of k from all 5 hidden factors of unwanted variation (for all estimated weights generated by RUVs, see Supplementary Table 5.5). I then converted my RUVs-corrected dataset to a DGE-list object, normalised the data with upperquartile normalisation, and applied Voom to eliminate the mean-variance trend (Figure 2.15). As in the pipeline using known covariates (above), the duplicateCorrelation function was used was used in order to estimate the correlation between expression measurements made on the same subject, then Voom was run a second time to incorporate the technical replicates as blocking variables. The 'robust' parameter was used using Limma's eBayes function and all significant genes were identified using an adjusted p-value of 0.01 (Benjamini–Hochberg corrected [3]) and an absolute log-fold change of 1.

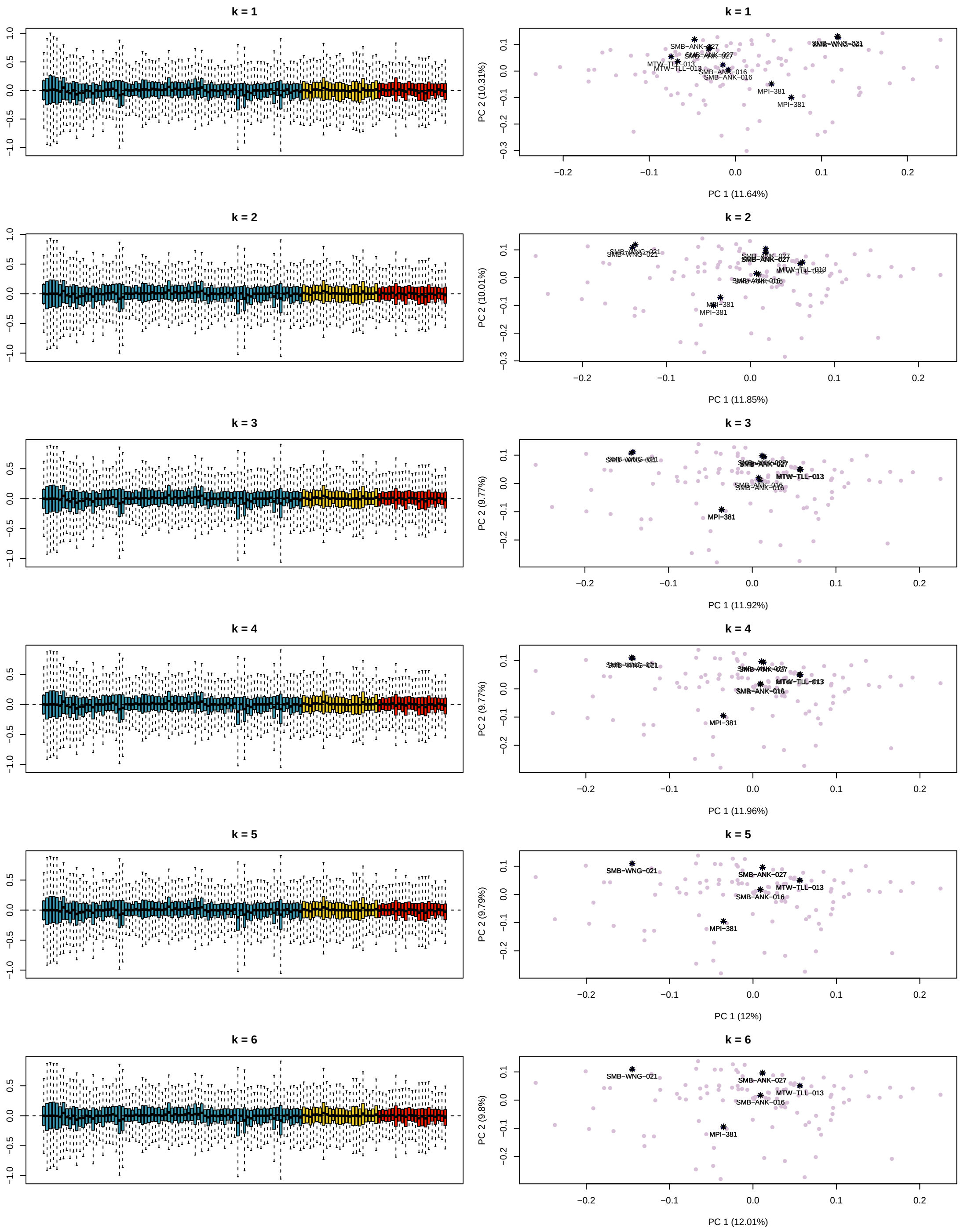

Figure 2.13: Performance of varying levels of k. Left: RLE plots highlighted by island show the amount of variability for varying levels of k. Right: Principal component analysis of the first two dimensions can be seen for a k of one to six. A k of three and higher results in reduced variability as well as technical replicates sitting closer together.

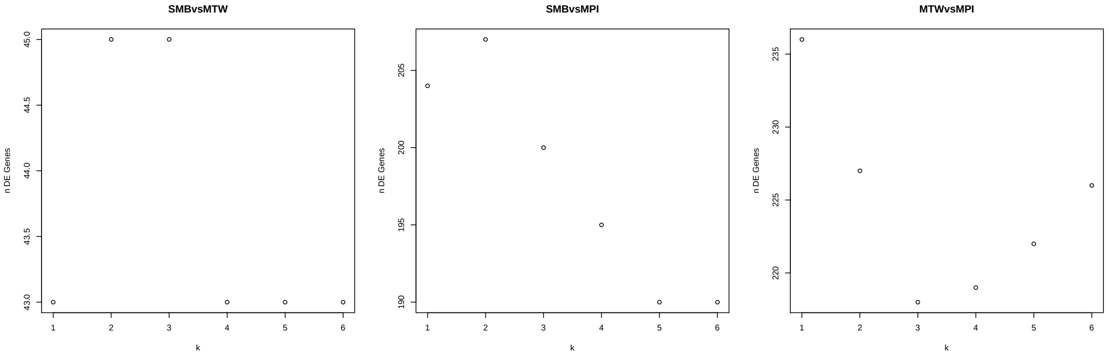

Figure 2.14: Total number of differentially expressed genes for varying levels of k. The significance threshold was set to an FDR of 0.01 and a log fold change of one. The inflection point for the number of differentially expressed genes can be seen to vary for each population comparison.

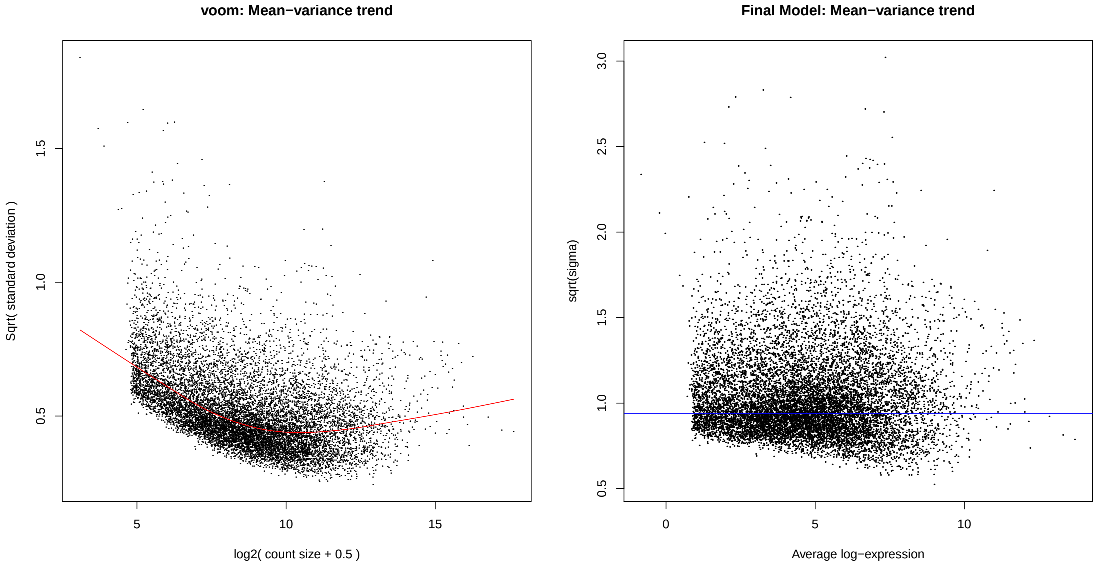

Figure 2.15: Mean-variance trend and elimination of the trend after applying Voom-normalisation to the upperquartile-normalised, RUVs-corrected count data.

2.8 Comparing both

After correcting for batch effects by regressing out known covariates influencing gene expression and with RUVs, I visually inspected grouping by batch through PCA analysis. The RUVs-corrected data had a higher percentage of variance in the first and second PCAs, whereas the limma-corrected data had less clustering by batch (Figure 2.16). Analysis of variance (ANOVA) between each dimension of the PCA and known covariates showed a stronger association between batch and the RUVs-corrected data than the limma-corrected data (Figure 2.17, Supplementary Table 5.6, Supplementary Table 5.7).

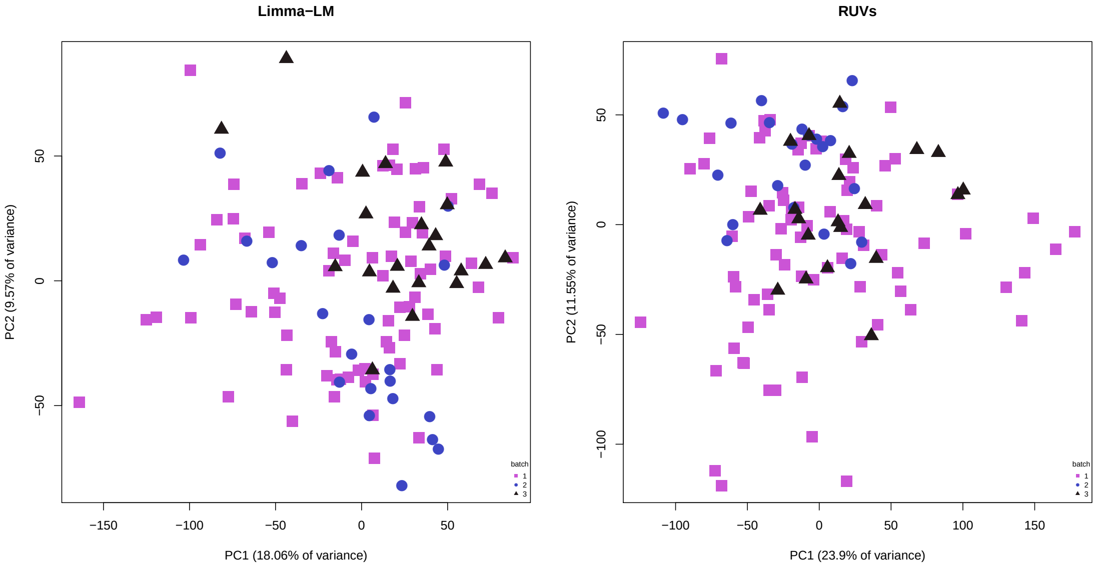

Figure 2.16: PCA plots of batch-corrected, log2-transformed CPM. The plot on the left is the limma corrected data, generated by the ‘Remove Batch effect’ function from the limma package. The second figure is generated from the RUVs-corrected counts and transformed to the log2 scale. Data corrected by the RUVs method shows more grouping by batch in PC1 and PC2, as well as a higher amount of variance

In order to see which method had a higher amount of unexplained variation, I used the package variance Partition [11] which uses a linear mixed model to

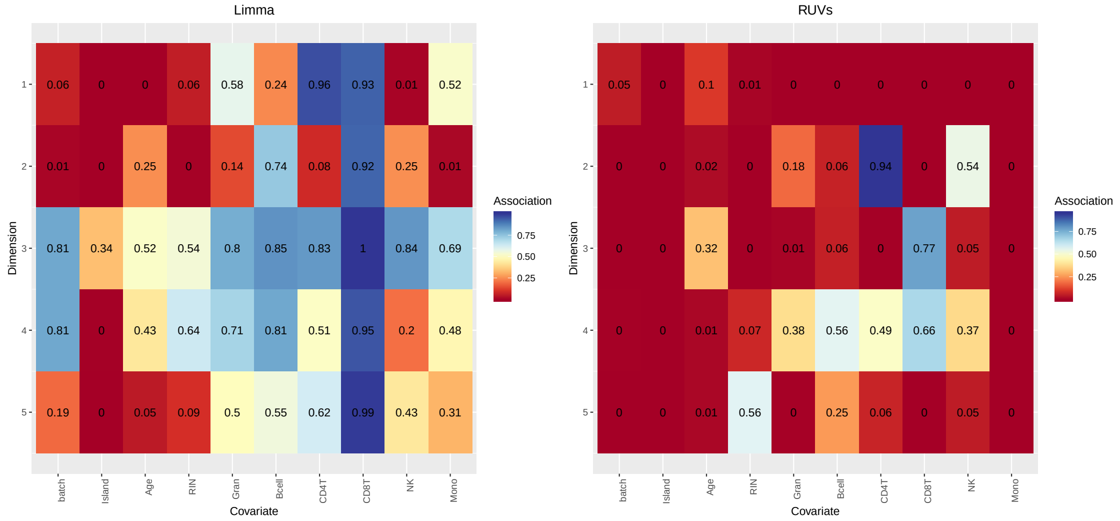

Figure 2.17: Analysis of variance (ANOVA) between each PCA dimension and covariates included in the design matrix for Limma-corrected and RUVs-corrected data. Count data corrected with the RUVs method can be seen to have a stronger association between most covariates and expression levels.

partition the amount of variance attributable to each variable. The amount of variance explained by each predictor was higher in the linear modelling method, whereas the RUVs correction method had a higher amount of residual variance (Figure 2.18).

Since the batch correction method including known covariates influencing gene expression into a model had less residual variation and a lower association between expression levels and sequencing batch, this method was chosen to correct for batch effects. Furthermore, estimating batch effects through packages such as RUV may introduce incorrect group differences in downstream analyses, particularly when groups are not evenly distributed between the batches [18].

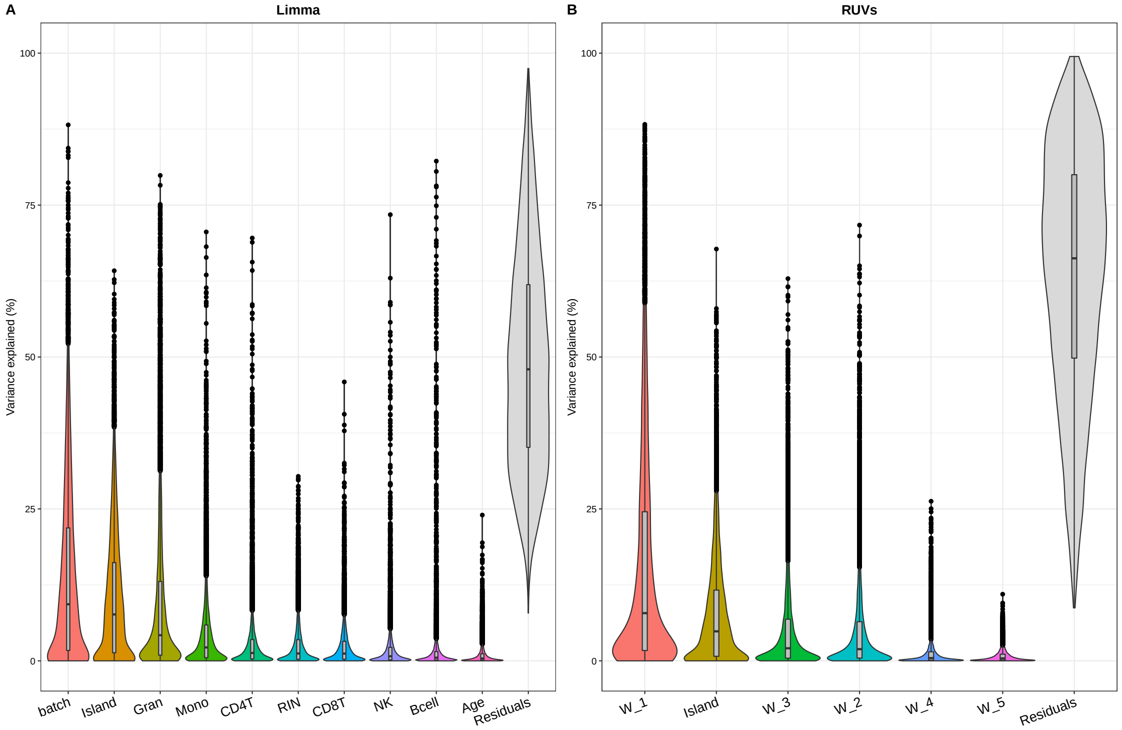

Figure 2.18: Total amount of variance explained. variancePartition from the Bioconductor R package was used to assign the amount of variance from each variable in the data. (A) The total amount of variance from all of the covariates shown to influence expression levels, as well as the total amount of residual variation. (B) The total amount of variance explained by each of the five weights calculated by RUVs, as well as island (the covariate of interest) and residual variance.

2.9 Gene enrichment analyses

example: Gene enrichment analyses were performed using the topGO package in R, with a Fisher exact test. Differentially expressed genes between MTL-locals and SAG-locals were compared against the 15,632 genes expressed in the CARTaGENE cohort that were retained after QC filters (background). Reactome enrichment analyses were conducted with R the package reactomePA, and here again, the background set of genes was defined as the 15,632 genes expressed in blood that pass our filters (Supplementary Fig. 7 and Supplementary Table 3).

3

Results

3.1 Ancestry is associated with expression profiles

After normalisation of the data and correcting for batch effects, expression profiles from each island formed unique clusters (Figure 3.1). PC1 and PC2 was most strongly correlated with island and village, as well as language and sampling date (two variables which were confounded with village). Mappi and season also had significant associations with PC1 (Figure 3.2).

I tested the number of differentially expressed genes (FDR < 0.01) at differing log fold thresholds (0, 0.5, and 1; Table 3.1). The majority of differentially expressed genes had a log fold change below one, particularly in Sumba versus Mentawai

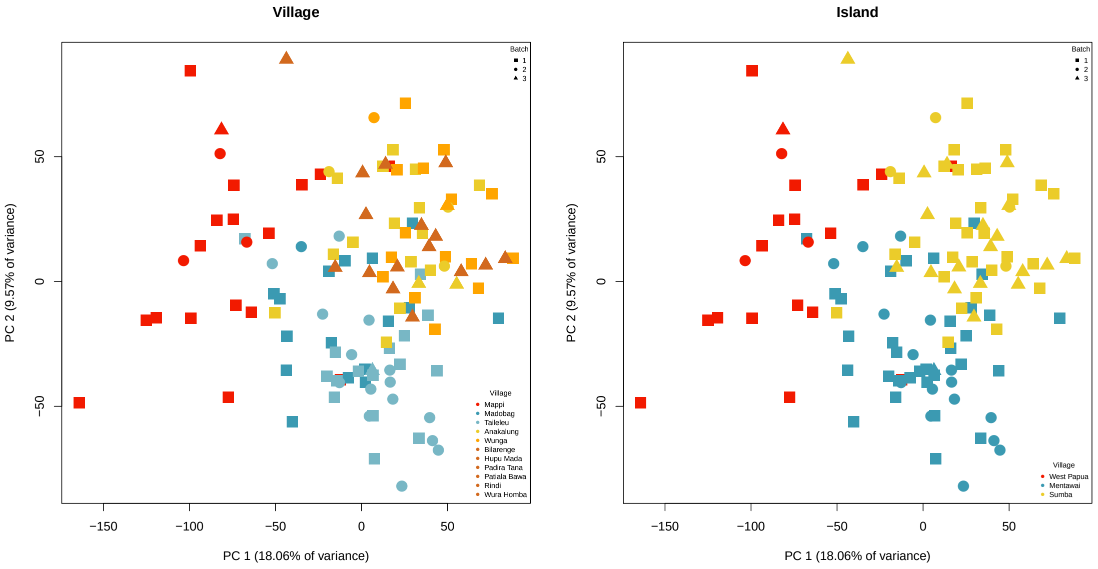

Figure 3.1

(Figure 3.3). At all thresholds, comparisons between Sumba and Mappi had the highest number of DE genes (n = 4496 at 0 LFC, n = 1398 at 0.5 LFC, and n = 313 at 1 LFC), followed by comparisons between Mentawai and Mappi (n = 4088 at 0 LFC, n = 1017 at 0.5 LFC, and n = 227 at 1 LFC). Sumba versus Mentawai consistently had the least number of differentially expressed genes (n = 1545 at 0 LFC, n = 314 at 0.5 LFC, and n = 35 at 1 LFC). Populations compared to Mappi had the highest number of DE genes: at the most stringent LFC of one (FDR < 0.01), I found 135 DE genes. In comparison, populations compared to Sumba at the same significance threshold had 16 DE genes and populations compared to Mentawai had 10 DE genes (Figure 3.4). This continued to hold true at differing LFC thresholds. A total of five DE genes were shared with all populations, most of which were involved in cell adhesion and cell growth.

When looking at the total number of differentially expressed genes at a finer scale, I found that many of the genes were driven by the Taillelou tribe (within Sumba) and Wunga tribe (within Mappi; Figure 3.5). As expected, comparisons

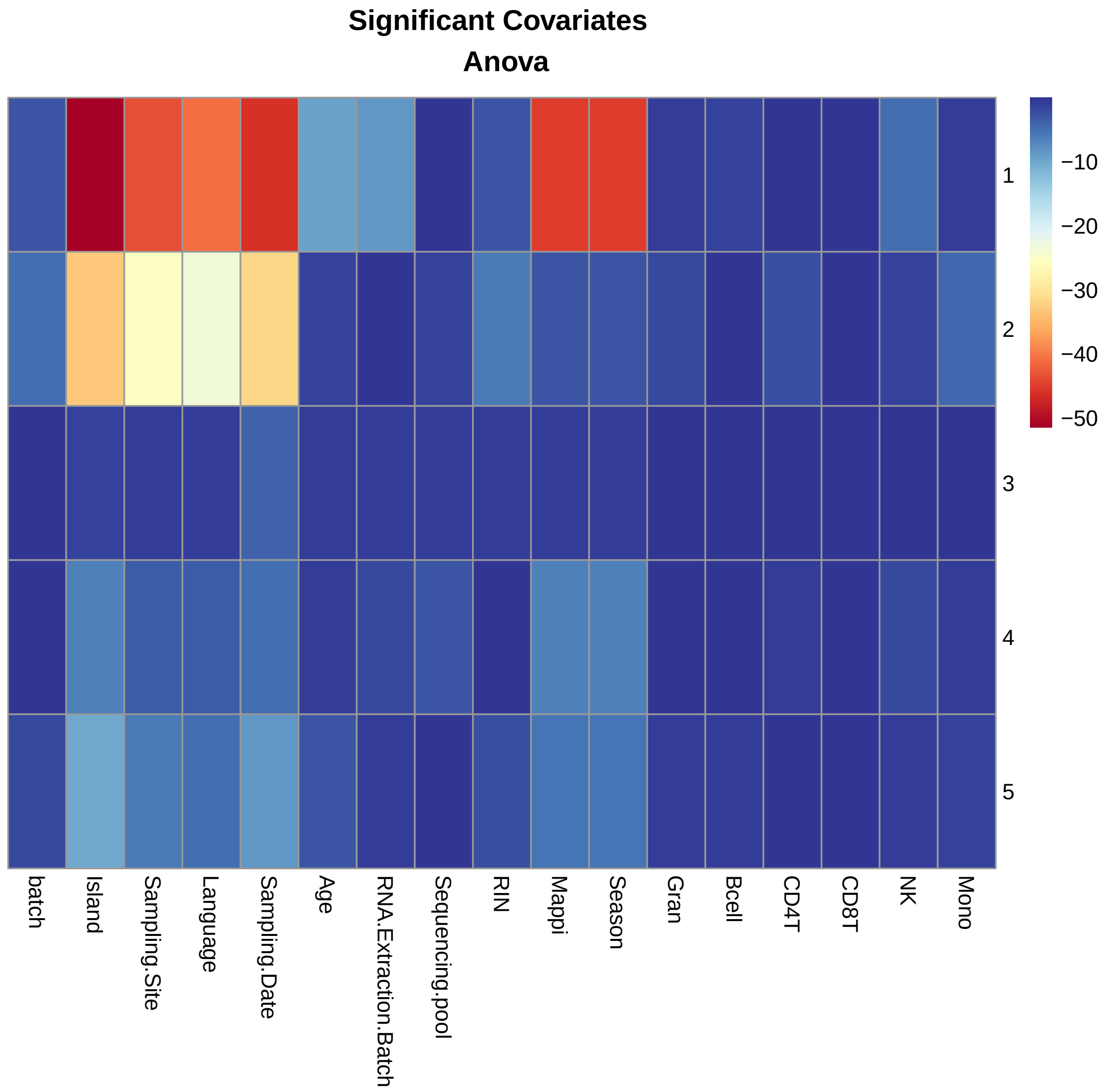

Figure 3.2

Table 3.1: Number of significantly DE genes at different logFC values. All DE genes are calculated with a pvalue cutoff of 0.01

LogFC	SMBvsMTW	SMBvsMPI	MTWvsMPI
0	1545	4496	4088
0.5	314	1398	1017
1	35	313	227

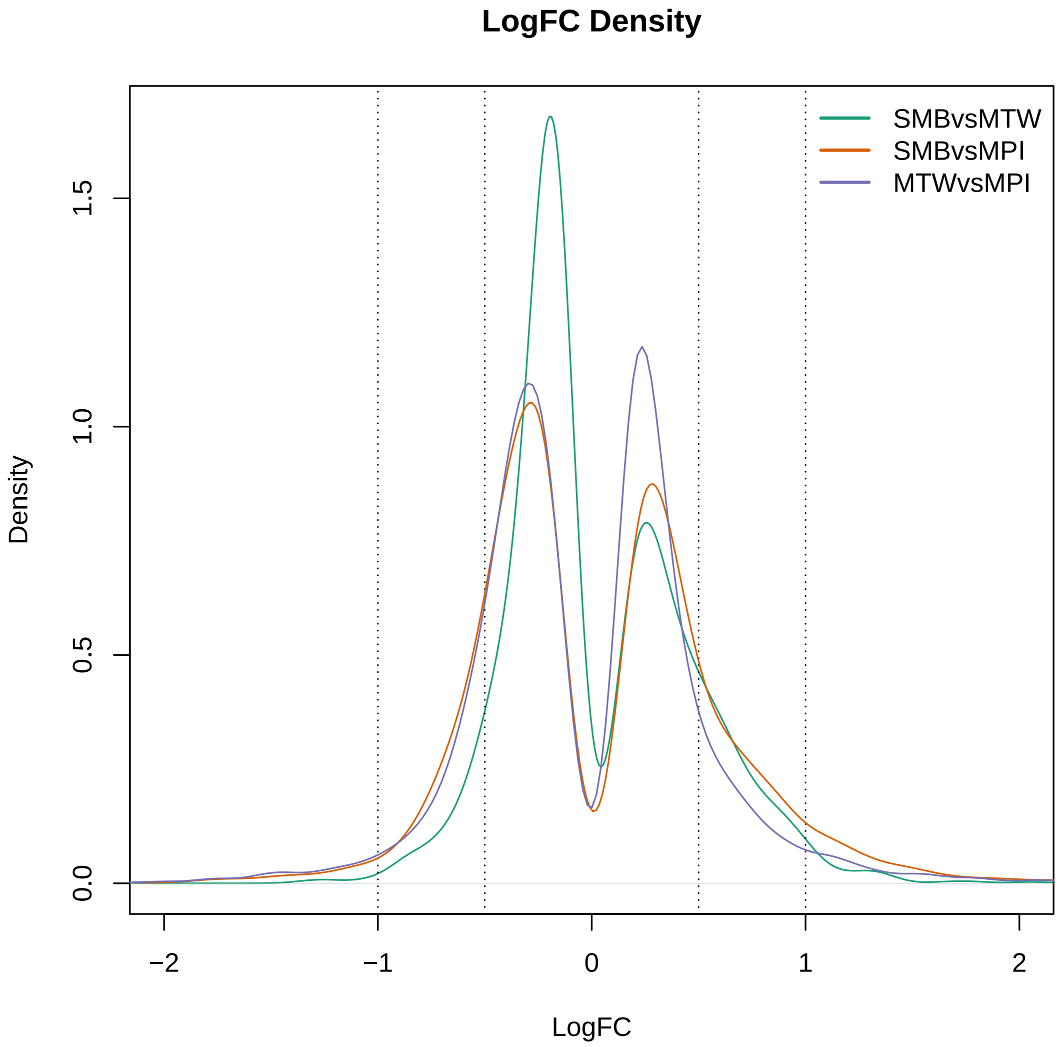

Figure 3.3: Density of the total number of differentially expressed genes with an FDR of 0.01 for all three population comparisons. Most of the differentially expressed genes with an FDR of 0.01 had an absolute log fold change less than one, particularly in Sumba versus Mentawai.

between villages had the least number of DE genes (Madobag and Tallelou within Mentawai: n = 471 at LFC 0; n = 53 at LFC of 0.5; n = 6 at LFC of l; Anakalung and Wunga within Sumba: n = 2 at LFC 0; n = 1 at LFC of 1; n = 1 at LFC of 1).

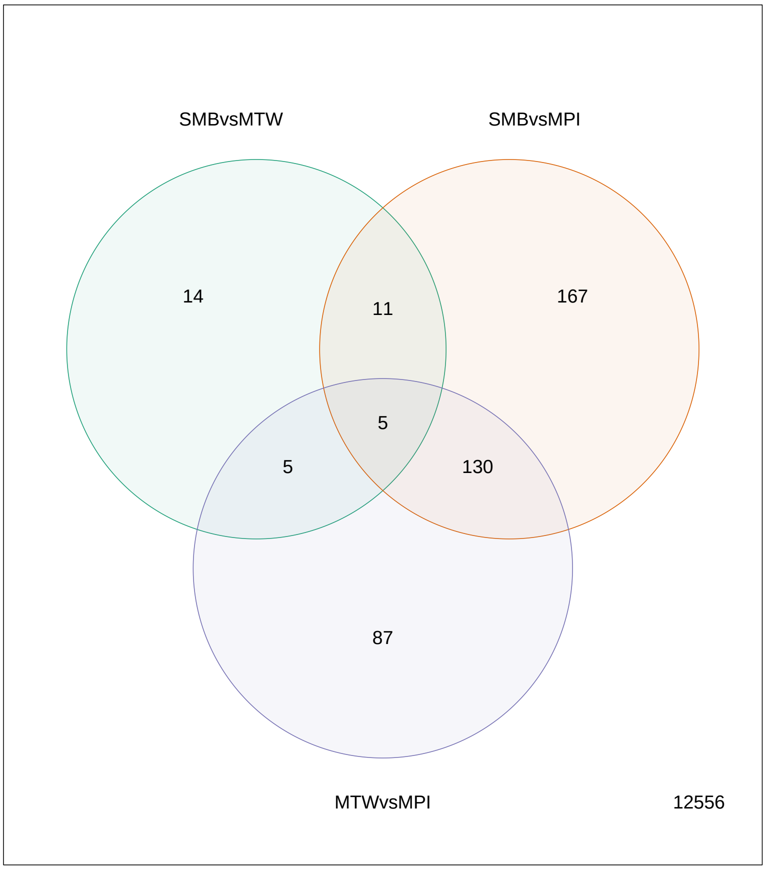

Figure 3.4: Number of significantly DE genes at an FDR of 0.01 and log fold change of one. The number of DE genes that are not in any population comparison are marked in the bottom-right.

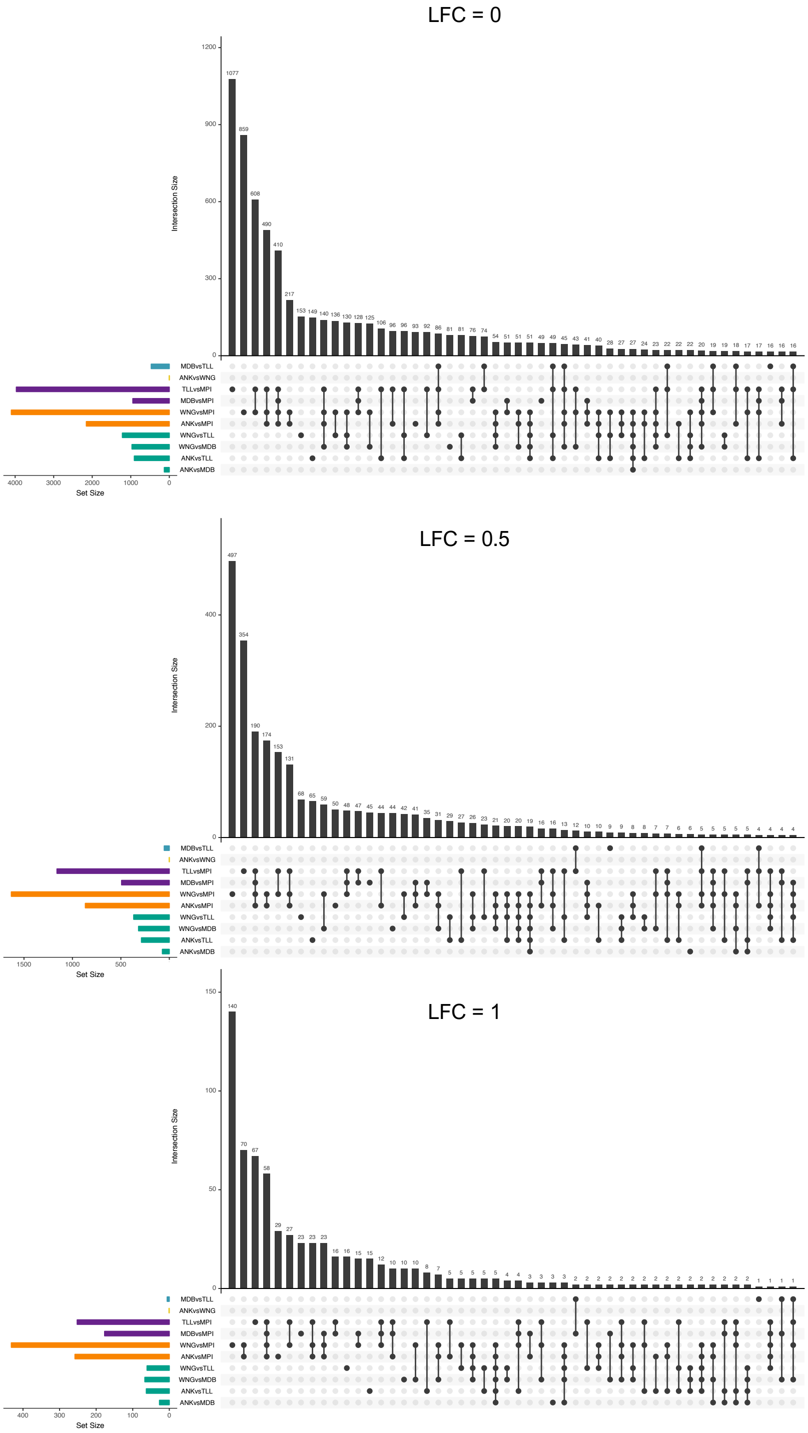

Figure 3.5: Number of significantly DE genes at an FDR of 0.01 and log fold change of one. The number of DE genes that are not in any population comparison are marked in the bottom-right.

3.2 Exploration of enriched gene sets in the Indonesian island populations

In order to see what pathways were involved in the genes differentially expressed between island populations, I used the goseq package in R which accounts for gene length bias in detection of over-represented genes. I used the default Wallenius dustribution to approximate the true distribution and all 12,975 genes which passed the filtering threshold as the background. I first identified all differentially expressed genes at a log fold change threshold of one and a p-value of 0.01. Since very few enriched GO terms remained after this threshold, I changed the LFC threshold to 0.05, which resulted in 200 enriched GO terms for Mentawai versus Mappi and 31 enriched GO terms for Sumba versus Mappi. No enriched GO terms were found for Mentawia versus Sumba. In population comparisons to Mappi, many gene sets were enriched in processes involved in inflammatory response, cell to cell signalling, and the defense response (Figure 3.6). This was especially apparent in Mentawai versus Mappi, which had the highest number of enriched gene sets out of all population comparisons.

In addition to enriched GO terms, I also looked at enriched pathways from c2 and KEGG databases using the EGSEA package in R. For c2 gene sets, many pathways were involved in interferon and immune activity, such as the JAK-STAT signaling pathway, regulation of IFNA signaling, and interferon gamma signaling 3.7. Interestingly, many c2 gene sets were also involved myeloid leukemia and other cancers. One of the highest enriched pathways was a c2 gene set associated with differences in mast cell production, of which the *SIGLEC6* gene was most one of the most highly differentially expressed 3.8. When searching the KEGG database, most significantly enriched pathways were involved in disease-associated phenotypes such as graft-versus-host disease, hepatitis C, viral myocarditis, glioma, and asthma.

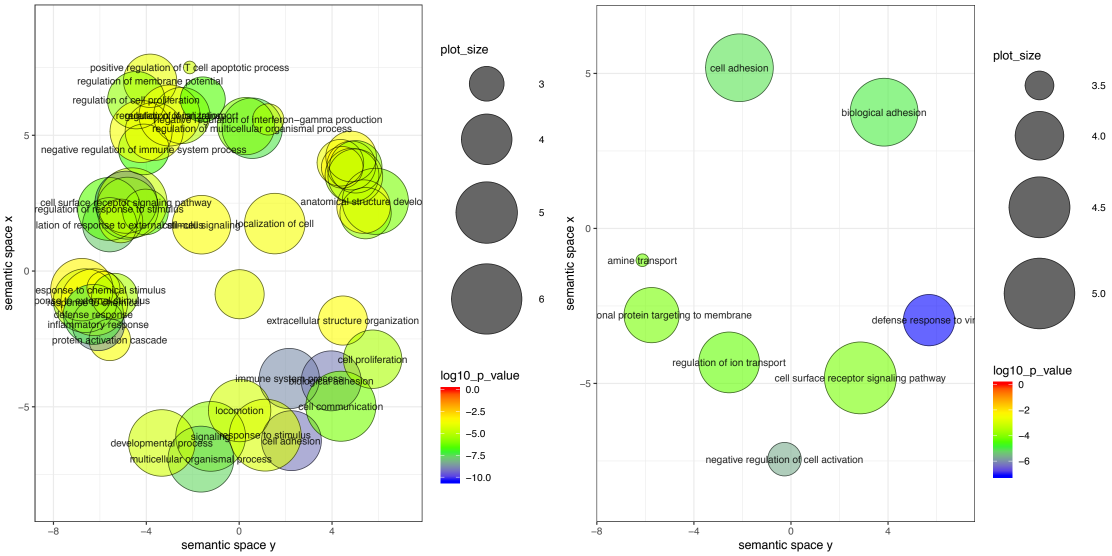

Figure 3.6: Visual representation of enriched GO terms (5% FDR) using REVIGO (Supek et al. 2011). For each GO ID, redundant GO terms are removed and grouped by semantically similar terms.

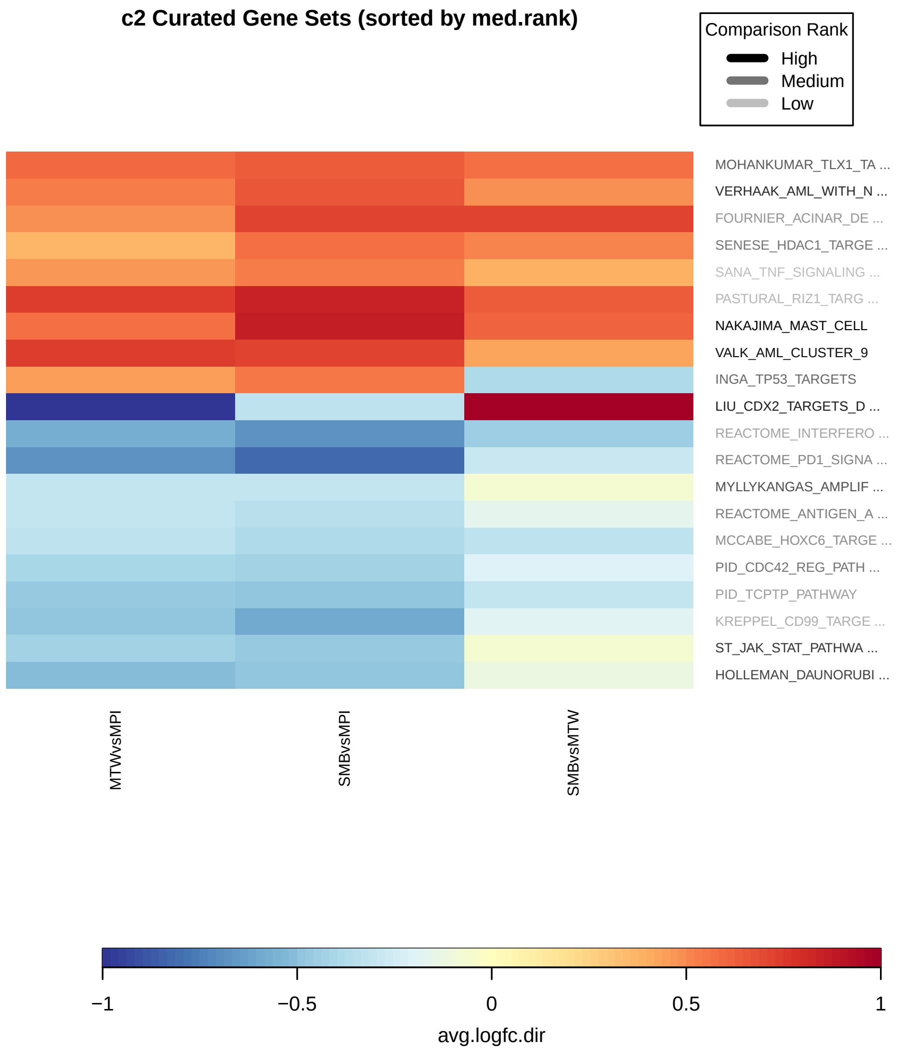

Figure 3.7: Bar plot of the -log10 adjusted p-values of the top 20 gene sets, ranked in order by the results of the gene set test CAMERA. Up-regulated genes are highlighted in orange and down-regulated genes are highlighted in blue.

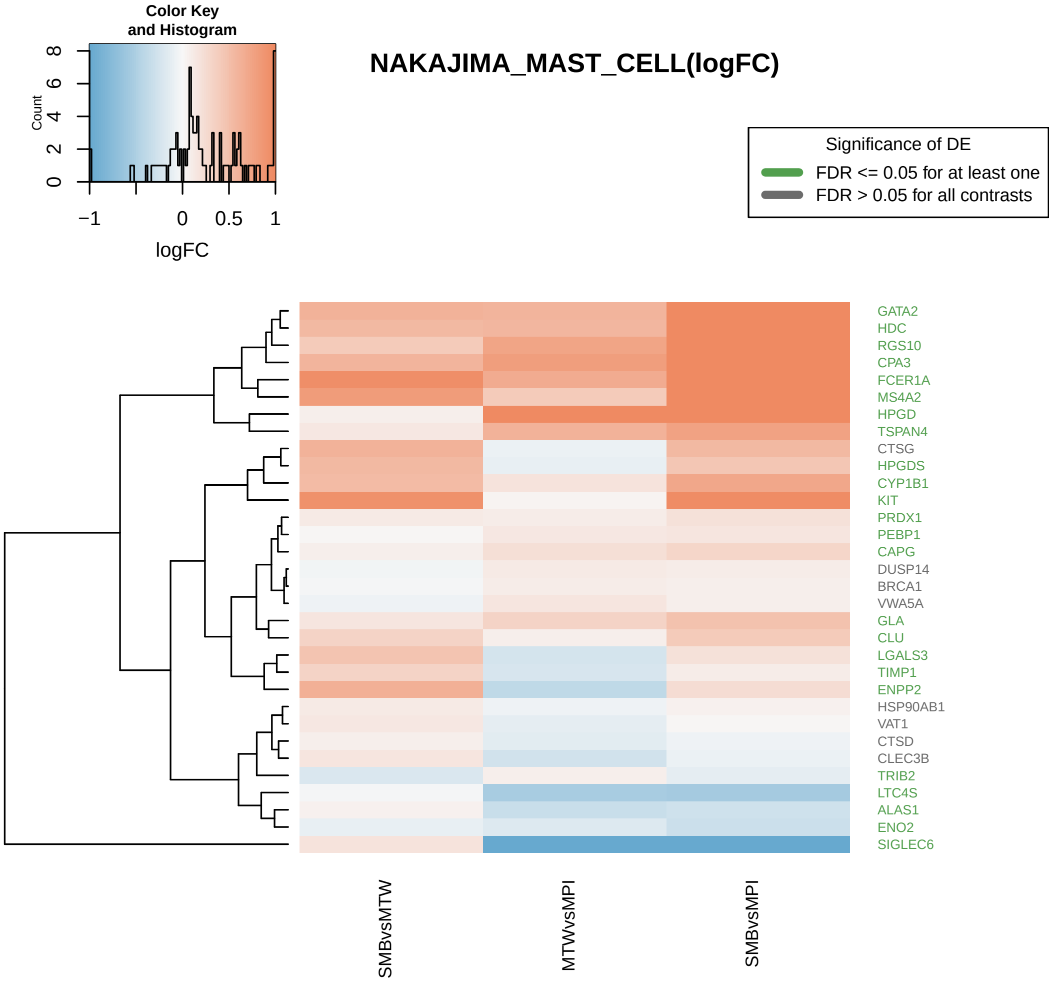

Figure 3.8: Bar plot of the -log10 adjusted p-values of the top 20 gene sets, ranked in order by the results of the gene set test CAMERA. Up-regulated genes are highlighted in orange and down-regulated genes are highlighted in blue.

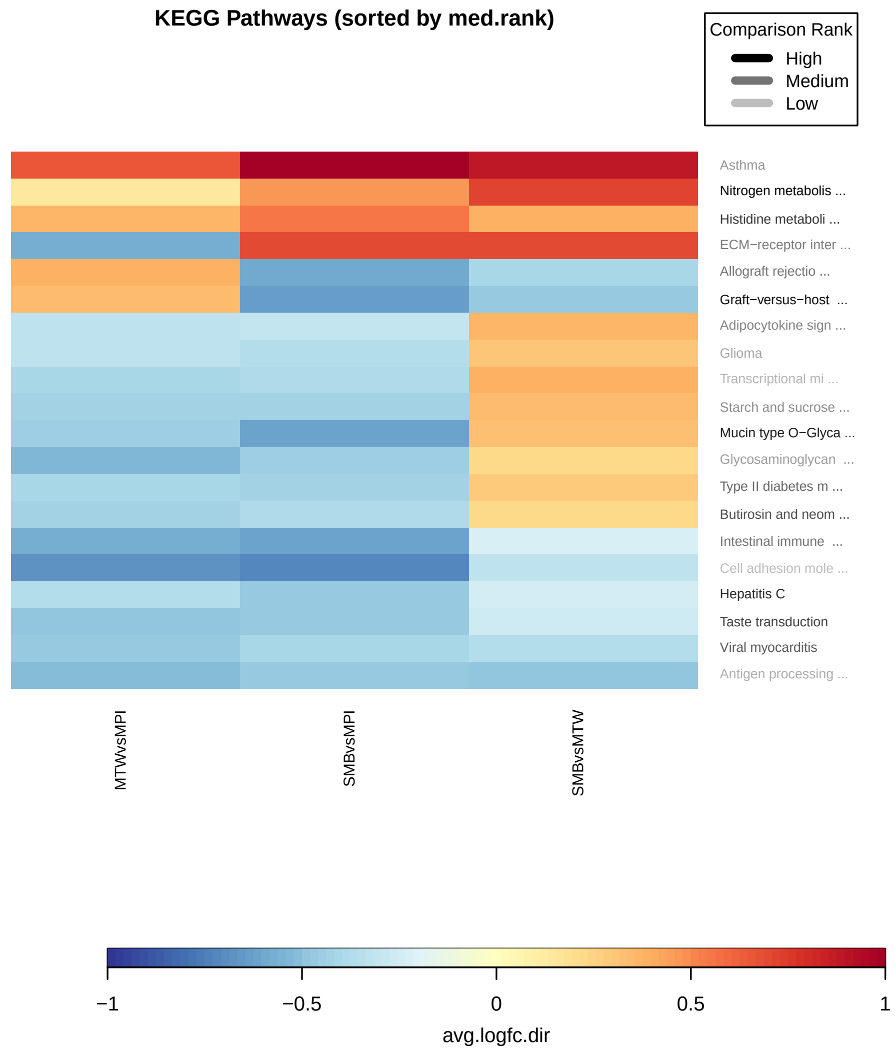

Figure 3.9: Bar plot of the -log10 adjusted p-values of the top 20 gene sets, ranked in order by the results of the gene set test CAMERA. Up-regulated genes are highlighted in orange and down-regulated genes are highlighted in blue.

3.3 Many DE Genes are involved in the immune response

After exploring the pathways involved in enriched genes, I looked at specific genes that were highly differentiated between populations. A heatmap of some of the most highly differentiated genes can be viewed in Figure 3.11. As in the pathway analysis, many of the most highly expressed genes were involved in immune-related processes, such as *SIGLEC6*, *SIGLEC7*, and *MARCO*. I analysed the distribution of all genes to see if significant genes were driven by individuals or uniform throughout the population. Samples from populations within each island generally had similar expression profiles for the most highly differentiated genes, as seen in Figure **??**.

Since the majority of differentially expressed genes were in comparisons to the Mappi population, I plotted the log fold cnahge values of all genes in common with Mappi in order to see if there were similarities in the direction of regulation. I found that every single DE gene (LFC=1, pvalue=0.01) was either concordantly up or down regulated in Mentawai and Sumba 3.12.

Two genes, SIGLEC6 and SOX5, had some of the highest adjusted pvalues in Mentawai verus Mappi and Sumba versus Mappi. Interestingly, SIGLEC6 was also found in to be involved in one of the most significant pathways involved in mast cell production.

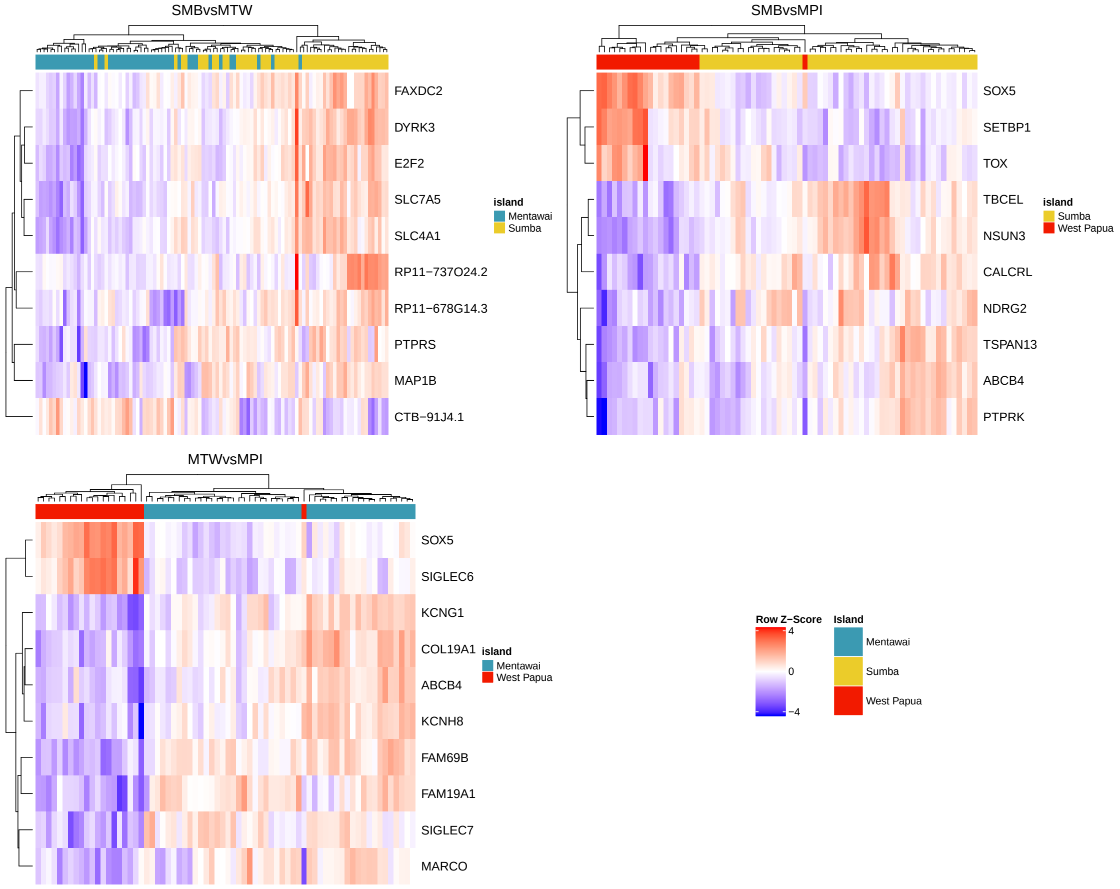

Figure 3.10: Heatmap of log-CPM values for the top ten differentially expressed genes for all population comparisons (sorted by adjusted p-value). Expression across each row is scaled so that the mean expression is zero and the standard deviation is one. Genes with high expression levels are highlighted in red and low expression highlighted in blue. Each column is a sample marked by its island colour (Sumba=yellow, Mappi=red, Mentawai=blue).

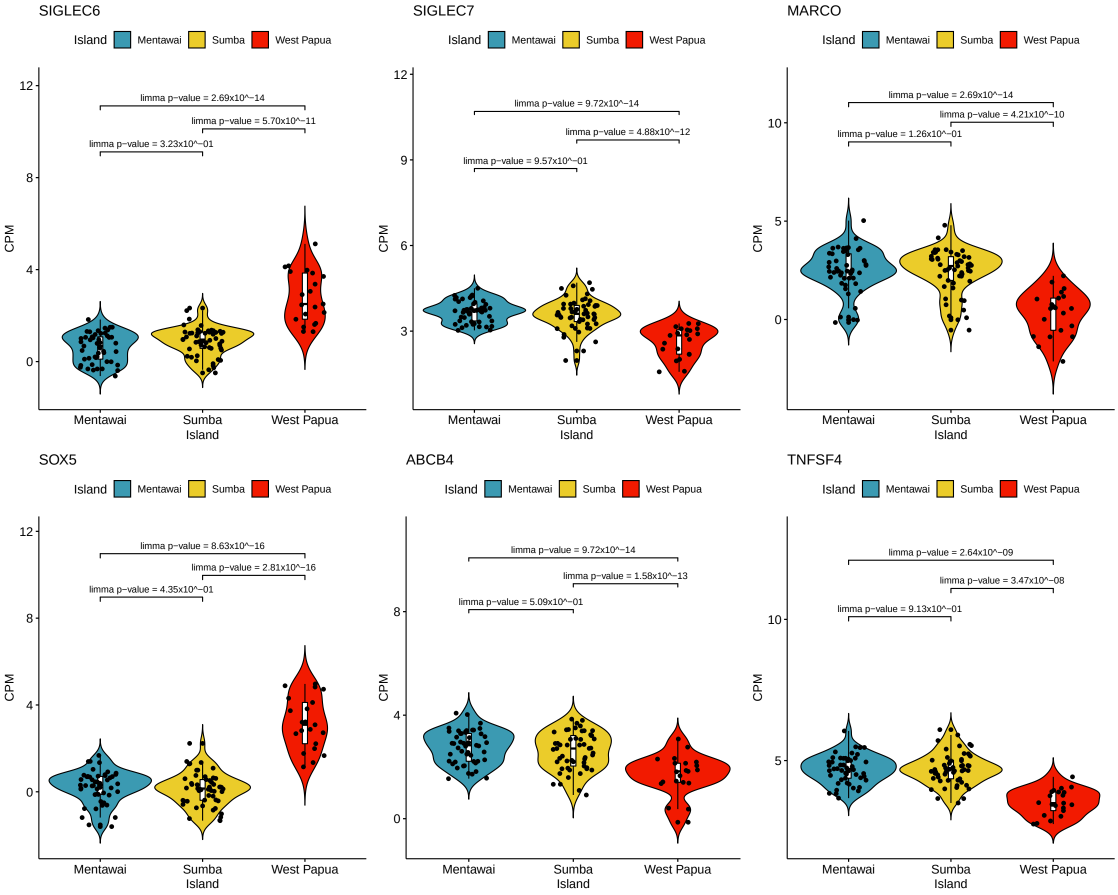

Figure 3.11: Expression levels of some of the top DE genes, all involved in the innate immune system. SIGLEC6 and SOX5 are upregulated in the Mappi population, whereas SIGLEC7, MARCO, ABCB4, and TNSF4 are downregulated in the Mappi population.

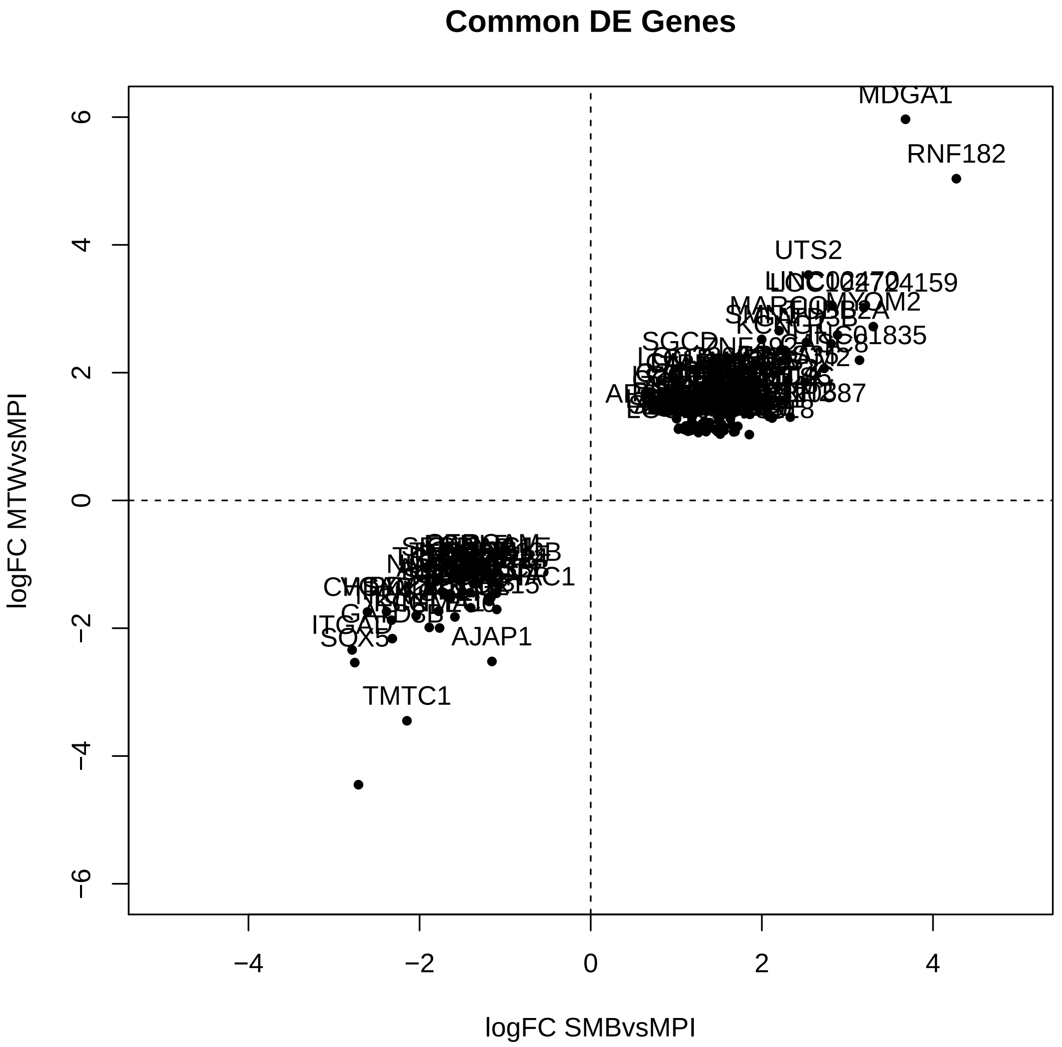

Figure 3.12: Log fold change of differentially expressed genes in Sumba versus Mappi and Mentawai versus Mappi. All significantly DE genes are shown to be both concordantly up or downregulated in both population comparisons.

4

Discussion

Here, we have analysed transcriptomic variation for the first time in healthy Indonesian samples. GIven the huge lack in diversity in current studies, this is important because...

It's cool that you can see fine-scale variation.

considerable expansion, the population remained linguistically and religiously isolated while remote regions were colonized by small numbers of settlers, such as SAG25,26 and contributed to the establishment of subpopulations. –> sumba had influence from.. and mappi.

Some interesting things: we have anlsyed how gene expressino looks in small populations in INdonesia. This has never been done in Indonesia, and aonly a handful of studies have analysed how populations differ in their trabscriptomic profile. Interstingly, we find that ancetrsy seems to correlate with differnces in genes expresion and that the populations with the two most disparate population histories have the highest number of differnetially expressed genes. We also find that the genes found to be most highly differentially expressed are primarily those involved in innate immune oathways. We annot tell whether this is due to biological differences, or whetheer this is caused by differences in regulatuoin due to encirinmental factors. These populations face significant environmetal pressures from pathogens. Many of which may be similar, but some can be different. Mappui in particular has a very different lifestyle- they live in treehouses and also have a lower amount of access to healthcare. This may affect their ability to repsond to infectoous diseases. It would be intersting ti see whether these are genes involved in protivtive immunity, o rif this is a resonse form teh enviornment.

We found some really interesting genes. In partivular, SIGLEC6 would be a really cool gene to follow up on. This gee is involved in.. SOm eother genes include. Inyerestingly, we also see Malaria gens, so this may be a clue to something that's going on from the envinrment SOm eother genes include SOX5, jsgdjs,

It's interesting that all the genes are either up or down regulated together...

5
Supplementary Information

5.1 Supplementary Figures

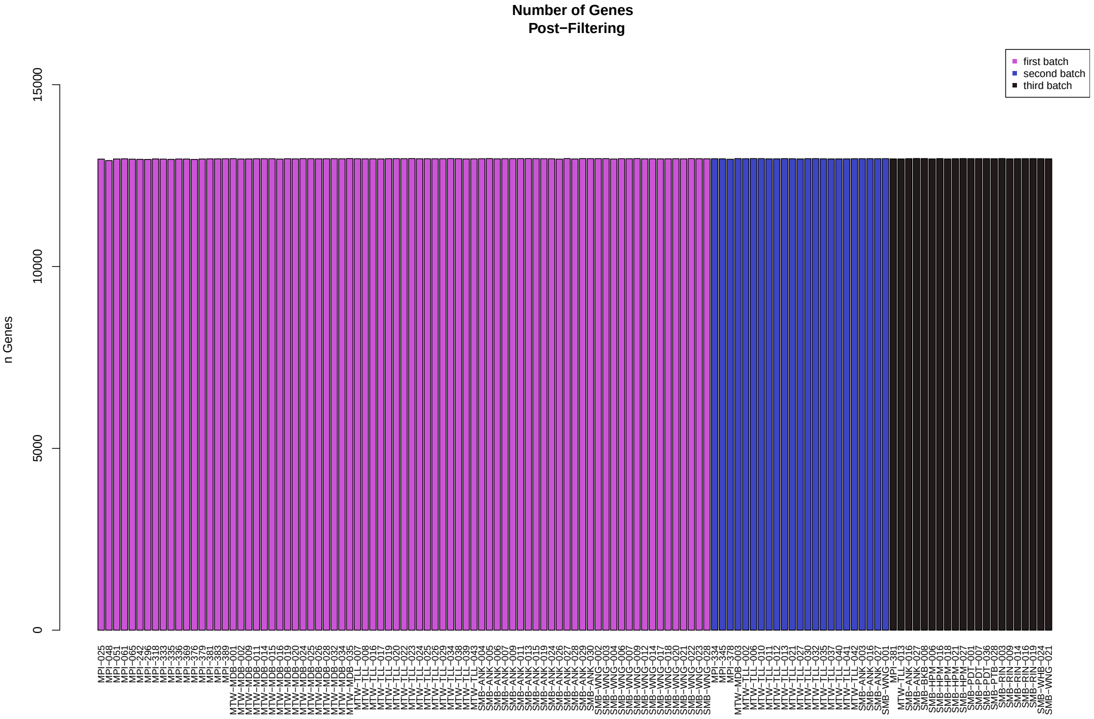

Figure 5.1: Lowly-expressed genes were removed by only retaining genes expressed at or over 1 in at least on of the island comparisons. This resulted in a total of 12,975 genes (12,914-12,971 genes per sample).

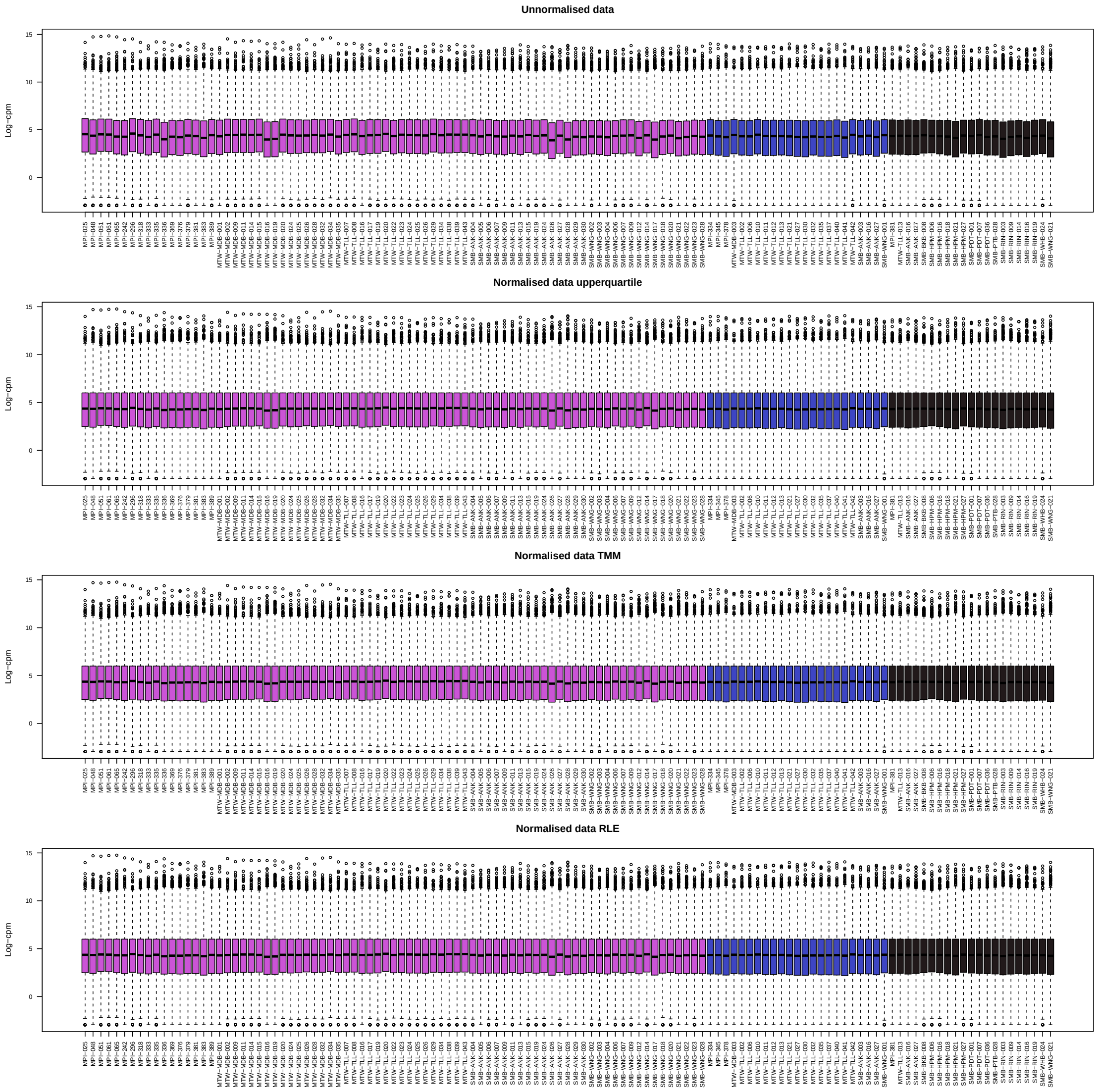

Figure 5.2: Performance of three different normalisation methods on the log2-transformed, filtered data. TMM normalisation was chosen due to its high performance in previous studies.

5.2 Supplementary Tables

Table 5.1: Batch and replicate information for all 117 samples (plus 6 technical replicates) used in the RNA-Seq analysis. (See supplementary file associated with this dissertation.)

Table 5.2: Total number of reads at each step of the pipeline. (See supplementary file associated with this dissertation.)

Table 5.3: Full list of covariates analysed in the RNA-seq analysis. (See supplementary file associated with this dissertation.)

Table 5.4: Associations between covariates and all PCA dimensions of the log2-transformed, TMM-normalised count data. (See supplementary file associated with this dissertation.)

Table 5.5: Weights from RUVs-corrected output. (See supplementary file associated with this dissertation.)

Table 5.6: Limma-corrected comparisons for each PCA. (See supplementary file associated with this dissertation.)

Table 5.7: RUVs-corrected comparisons for each PCA. (See supplementary file associated with this dissertation.)

Bibliography

[1] Frank W Albert and Leonid Kruglyak. The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics*, 16(4):197, 2015.

[2] Alexis Battle, Sara Mostafavi, Xiaowei Zhu, James B Potash, Myrna M Weissman, Courtney McCormick, Christian D Haudenschild, Kenneth B Beckman, Jianxin Shi, Rui Mei, et al. Characterizing the genetic basis of transcriptome diversity through rna-sequencing of 922 individuals. *Genome research*, 24(1): 14–24, 2014.

[3] 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.

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

[5] Lu Chen, Bing Ge, Francesco Paolo Casale, Louella Vasquez, Tony Kwan, Diego Garrido-Martín, Stephen Watt, Ying Yan, Kousik Kundu, Simone Ecker, et al. Genetic drivers of epigenetic and transcriptional variation in human immune cells. *Cell*, 167(5):1398–1414, 2016.

[6] Yunshun Chen, D McCarthy, M Robinson, and GK Smyth. edger: differential expression analysis of digital gene expression data. *Bioconductor User?s Guide*, pages 1–78, 2014.

[7] William Cookson, Liming Liang, Gonçalo Abecasis, Miriam Moffatt, and Mark Lathrop. Mapping complex disease traits with global gene expression. *Nature Reviews Genetics*, 10(3):184, 2009.

[8] 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.

[9] Johann A Gagnon-Bartsch and Terence P Speed. Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, 13(3):539–552, 2012.

[10] Simone Gupta, Shannon E Ellis, Foram N Ashar, Anna Moes, Joel S Bader, Jianan Zhan, Andrew B West, and Dan E Arking. Transcriptome analysis reveals dysregulation of innate immune response genes and neuronal activity-dependent genes in autism. *Nature communications*, 5:5748, 2014.

[11] Gabriel E Hoffman and Eric E Schadt. variancepartition: interpreting drivers of variation in complex gene expression studies. *BMC bioinformatics*, 17(1): 483, 2016.

[12] Georgi Hudjashov, Tatiana M Karafet, Daniel J Lawson, Sean Downey, Olga Savina, Herawati Sudoyo, J Stephen Lansing, Michael F Hammer, and Murray P Cox. Complex patterns of admixture across the indonesian archipelago. *Molecular biology and evolution*, 34(10):2439–2452, 2017.

[13] Laura S Kremer, Daniel M Bader, Christian Mertes, Robert Kopajtich, Garwin Pichler, Arcangela Iuso, Tobias B Haack, Elisabeth Graf, Thomas Schwarzmayr, Caterina Terrile, et al. Genetic diagnosis of mendelian disorders via rna sequencing. *Nature communications*, 8:15824, 2017.

[14] Charity W Law, Yunshun Chen, Wei Shi, and Gordon K Smyth. voom: Precision weights unlock linear model analysis tools for rna-seq read counts. *Genome biology*, 15(2):R29, 2014.

[15] Yang Liao, Gordon K Smyth, and Wei Shi. featurecounts: an efficient general

purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930, 2013.

[16] Alicia R Martin, Helio A Costa, Tuuli Lappalainen, Brenna M Henn, Jeffrey M Kidd, Muh-Ching Yee, Fabian Grubert, Howard M Cann, Michael Snyder, Stephen B Montgomery, et al. Transcriptome sequencing from diverse human populations reveals differentiated regulatory architecture. *PLoS genetics*, 10 (8):e1004549, 2014.

[17] Alicia R Martin, Christopher R Gignoux, Raymond K Walters, Genevieve L Wojcik, Benjamin M Neale, Simon Gravel, Mark J Daly, Carlos D Bustamante, and Eimear E Kenny. Human demographic history impacts genetic risk prediction across diverse populations. *The American Journal of Human Genetics*, 100(4):635–649, 2017.

[18] Vegard Nygaard, Einar Andreas Rødland, and Eivind Hovig. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics*, 17(1):29–39, 2016.

[19] Mauro Pala, Zachary Zappala, Mara Marongiu, Xin Li, Joe R Davis, Roberto Cusano, Francesca Crobu, Kimberly R Kukurba, Michael J Gloudemans, Frederic Reinier, et al. Population-and individual-specific regulatory variation in sardinia. *Nature genetics*, 49(5):700, 2017.

[20] Mariana Buongermino Pereira, Mikael Wallroth, Viktor Jonsson, and Erik Kristiansson. Comparison of normalization methods for the analysis of metagenomic gene abundance data. *BMC genomics*, 19(1):274, 2018.

[21] Hélène Quach, Maxime Rotival, Julien Pothlichet, Yong-Hwee Eddie Loh, Michael Dannemann, Nora Zidane, Guillaume Laval, Etienne Patin, Christine

Harmant, Marie Lopez, et al. Genetic adaptation and neandertal admixture shaped the immune system of human populations. *Cell*, 167(3):643–656, 2016.

[22] Davide Risso, John Ngai, Terence P Speed, and Sandrine Dudoit. Normalization of rna-seq data using factor analysis of control genes or samples. *Nature biotechnology*, 32(9):896, 2014.

[23] Gordon K Smyth, Natalie Thorne, and James Wettenhall. Limma: linear models for microarray data user?s guide. *Software manual available from http://www. bioconductor. org*, 2003.

[24] Charlotte Soneson, Michael I Love, and Mark D Robinson. Differential analyses for rna-seq: transcript-level estimates improve gene-level inferences. *F1000Research*, 4, 2015.

[25] Nicholas J Timpson, Celia MT Greenwood, Nicole Soranzo, Daniel J Lawson, and J Brent Richards. Genetic architecture: the shape of the genetic contribution to human traits and disease. *Nature Reviews Genetics*, 19(2):110, 2018.

[26] Harm-Jan Westra, Marjolein J Peters, Tõnu Esko, Hanieh Yaghootkar, Claudia Schurmann, Johannes Kettunen, Mark W Christiansen, Benjamin P Fairfax, Katharina Schramm, Joseph E Powell, et al. Systematic identification of trans eqtls as putative drivers of known disease associations. *Nature genetics*, 45 (10):1238, 2013.

[27] Junya Yamagishi, Anna Natori, Mohammed EM Tolba, Arthur E Mongan, Chihiro Sugimoto, Toshiaki Katayama, Shuichi Kawashima, Wojciech Makalowski, Ryuichiro Maeda, Yuki Eshita, et al. Interactive transcriptome analysis of malaria patients and infecting plasmodium falciparum. *Genome research*, 24(9):1433–1444, 2014.

[28] Zachary Zappala and Stephen B Montgomery. Non-coding loss-of-function variation in human genomes. *Human heredity*, 81(2):78–87, 2016.