DESeq2 Analysis of Test counts.table

CSF-NGS Bioinformatics

May 19, 2014

sample	conditions	sizeFactor	gene.counts
sample1	wt	1.89	8155855
sample2	ko	0.53	2131298

Table 1: Column names in original table, condition names and size factors for normalization

parameter	value
comparison	wt vs ko
filter lowest	FALSE
filter cook	TRUE
p-value cutoff	0.05

Table 2: Parameters used in this study.

column name	explanation
id	input id
baseMean	mean normalised counts, averaged over all conditions from both conditions
log2FoldChange	the logarithm (to basis 2) of the moderated fold change
lfcSE	standard error of the log2FoldChange
stat	the statistic for p-value calculation
pval	p value for the statistical significance of this change
padj	p value adjusted for multiple testing with the Benjamini-Hochberg procedure
zeroExp	if the gene had 0 counts in all samples
meanFilter	if gene was removed because of expression filtering
cookFilter	if gene was removed because of outlier detection (Cook distance)
...	the following columns are the normalized input counts for each input sample

Table 3: Explanation for Columns in Output Table

DESeq2 [1] was used to compare 2 samples separated into 2 conditions (Table 1). The data of each column given in the input file was normalized[1] to an effective library size by the sizeFactors (Table 1). A clustered heat map of the spearman correlation coefficient of the sample read counts was generated to detect labeling errors regarding the sample/condition combinations (Figure 2). Spearman correlation was used to prevent very highly and lowly expressed genes to dominate the correlation. From the normalized data the variance for each gene across the samples was estimated (Figure **??**). A statistical test of wt vs ko was carried out using the estimated gene variances. The resuting p-values were adjusted for multiple testing by the Benjamini-Hochberg procedure[2]. The number of significantly up and downregulated genes based on a significant treshold and a log2 FC treshold is shown in Table 5. A tab delimited file was produced showing for each gene multiple statistics. The columns are explained in table 3.

[1]The size factors are determined using the median of the deviations from the logarithmic mean from each gene across all samples

countName	filterDFSummary
total	23985
no reads	4088
p-value filter	0
outlier filter	0

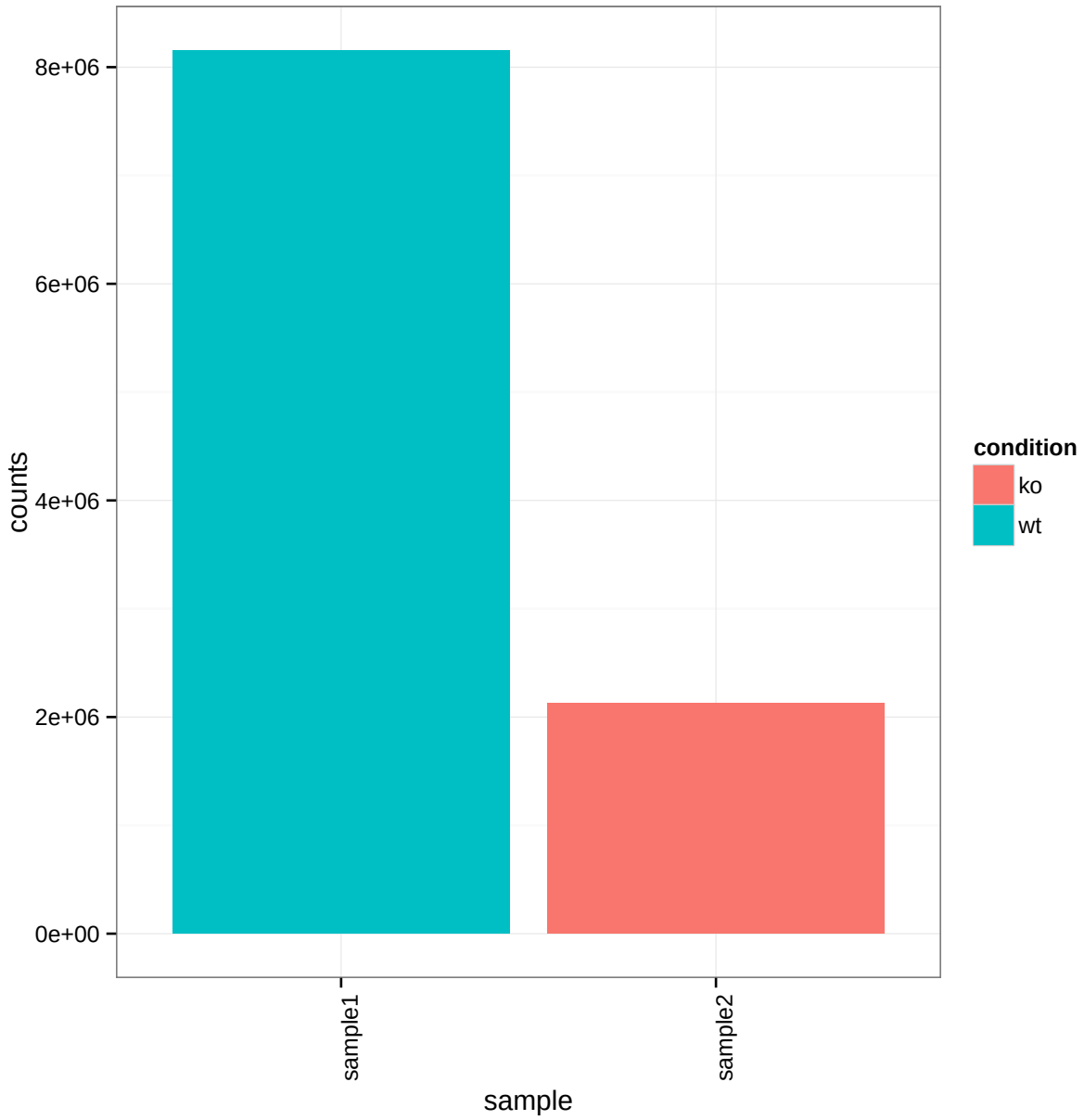
Table 4: Count of genes with adjusted p-values set to NA because of no detectable expression, mean value filter removal with treshold <0 and outlier filter (Cooks Distance).

	abs(log2FC)> 1	abs(log2FC)> 2	abs(log2FC)> 3	abs(log2FC)> 4
up	0	0	0	0
down	0	0	0	0

Table 5: Count of up and downregulated genes with a log2 FC cutoff and an adjusted p-value <0.05

0.1 R Software Versions

- R version 3.0.2 (2013-09-25), `x86_64-unknown-linux-gnu`
- Locale: `LC_CTYPE=en_US.UTF-8`, `LC_NUMERIC=C`, `LC_TIME=en_US.UTF-8`, `LC_COLLATE=en_US.UTF-8`, `LC_MONETARY=en_US.UTF-8`, `LC_MESSAGES=en_US.UTF-8`, `LC_PAPER=en_US.UTF-8`, `LC_NAME=C`, `LC_ADDRESS=C`, `LC_TELEPHONE=C`, `LC_MEASUREMENT=en_US.UTF-8`, `LC_IDENTIFICATION=C`
- Base packages: base, datasets, graphics, grDevices, grid, methods, parallel, stats, utils
- Other packages: AnnotationDbi 1.24.0, Biobase 2.22.0, BiocGenerics 0.8.0, DBI 0.2-7, DESeq2 1.2.10, genefilter 1.44.0, GenomicRanges 1.14.4, ggplot2 0.9.3.1, gplots 2.12.1, hexbin 1.26.3, hwriter 1.3, IRanges 1.20.7, knitr 1.5, lattice 0.20-27, latticeExtra 0.6-26, RColorBrewer 1.0-5, Rcpp 0.11.1, RcppArmadillo 0.4.100.2.1, ReportingTools 2.2.0, reshape2 1.2.2, RSQLite 0.11.4, scales 0.2.3, stringr 0.6.2, xtable 1.7-3, XVector 0.2.0
- Loaded via a namespace (and not attached): annotate 1.40.1, AnnotationForge 1.4.4, biomaRt 2.18.0, Biostrings 2.30.1, biovizBase 1.10.8, bitops 1.0-6, BSgenome 1.30.0, Category 2.28.0, caTools 1.16, cluster 1.15.1, colorspace 1.2-4, dichromat 2.0-0, digest 0.6.4, edgeR 3.4.2, evaluate 0.5.1, formatR 0.10, Formula 1.1-1, gdata 2.13.2, GenomicFeatures 1.14.3, ggbio 1.10.12, GO.db 2.10.1, GOstats 2.28.0, graph 1.40.1, gridExtra 0.9.1, GSEABase 1.24.0, gtable 0.1.2, gtools 3.3.1, Hmisc 3.14-3, KernSmooth 2.23-10, labeling 0.2, limma 3.18.13, locfit 1.5-9.1, MASS 7.3-30, Matrix 1.1-2-2, munsell 0.4.2, PFAM.db 2.10.1, plyr 1.8.1, proto 0.3-10, R.methodsS3 1.6.1, R.oo 1.18.0, R.utils 1.29.8, RBGL 1.38.0, RCurl 1.95-4.1, Rsamtools 1.14.3, rtracklayer 1.22.4, splines 3.0.2, stats4 3.0.2, survival 2.37-7, tools 3.0.2, VariantAnnotation 1.8.12, XML 3.98-1.1, zlibbioc 1.8.0




```
## Error:  'x' must have at least 2 rows and 2 columns
```

References

[1] Love, Michael I. and Huber, Wolfgang and Anders, Simon: Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. bioRxiv, February 19, 2014. doi:10.1101/002832. http://biorxiv.org/content/early/2014/02/19/002832

[2] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society B, 57:289300, 1995.

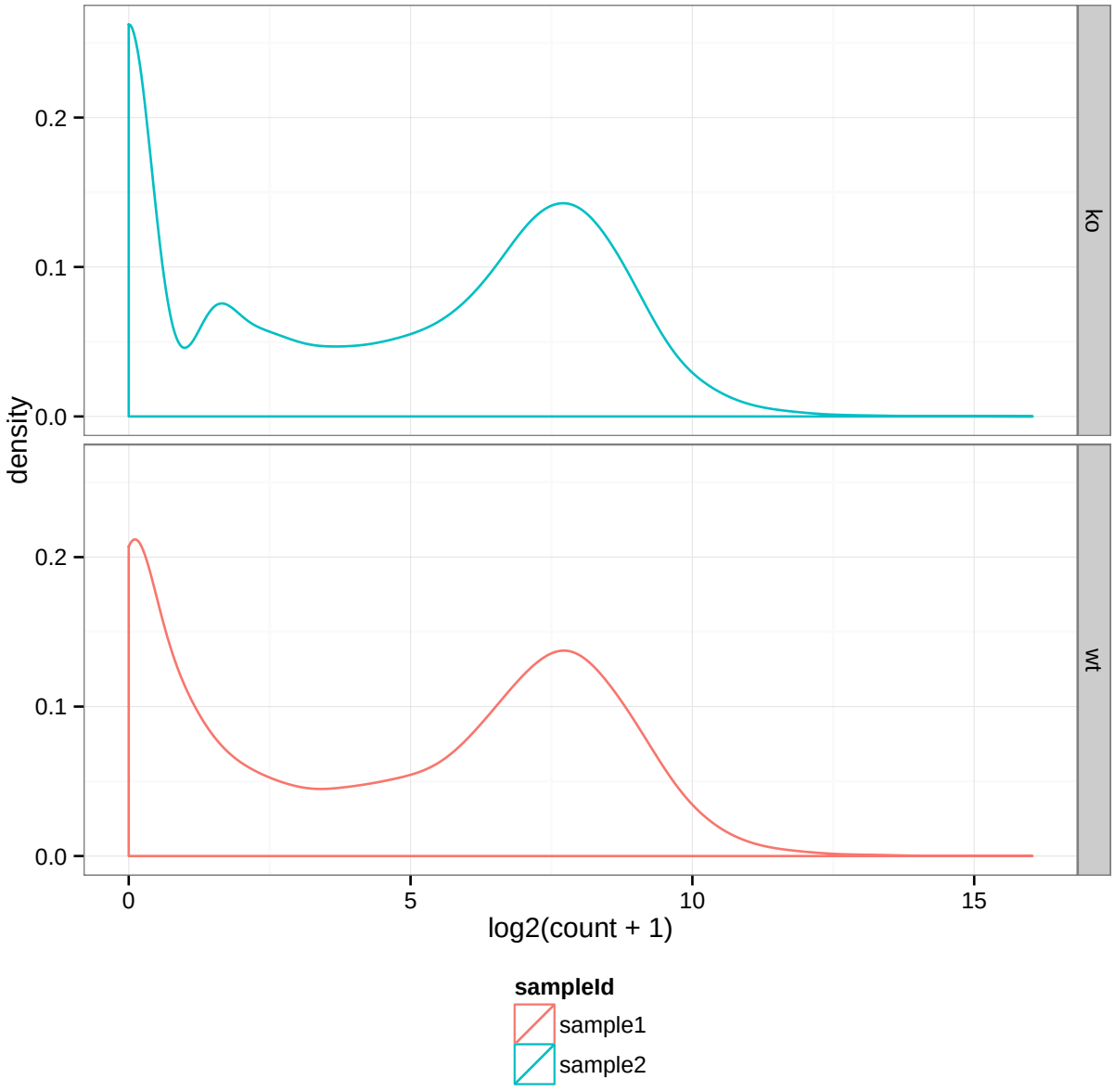


Figure 1: Denstiy plot of the normalized counts split by the conditions. Well normalised data should have similar shapes.

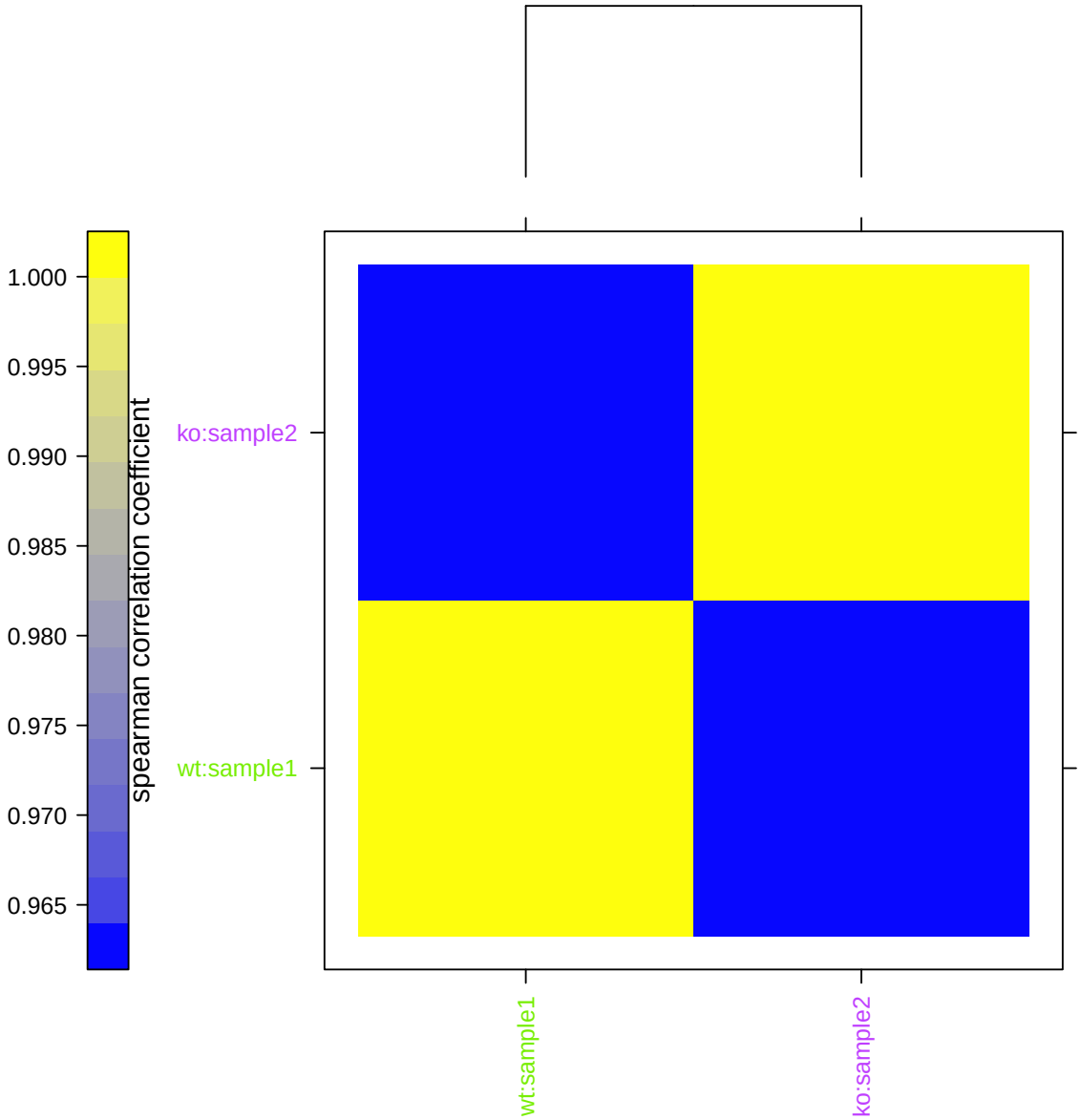


Figure 2: Correlation of Replicates. The normalized samples were clustered by their pairwise spearman correlation coefficient. Samples with the same condition should cluster together on the x-axis.

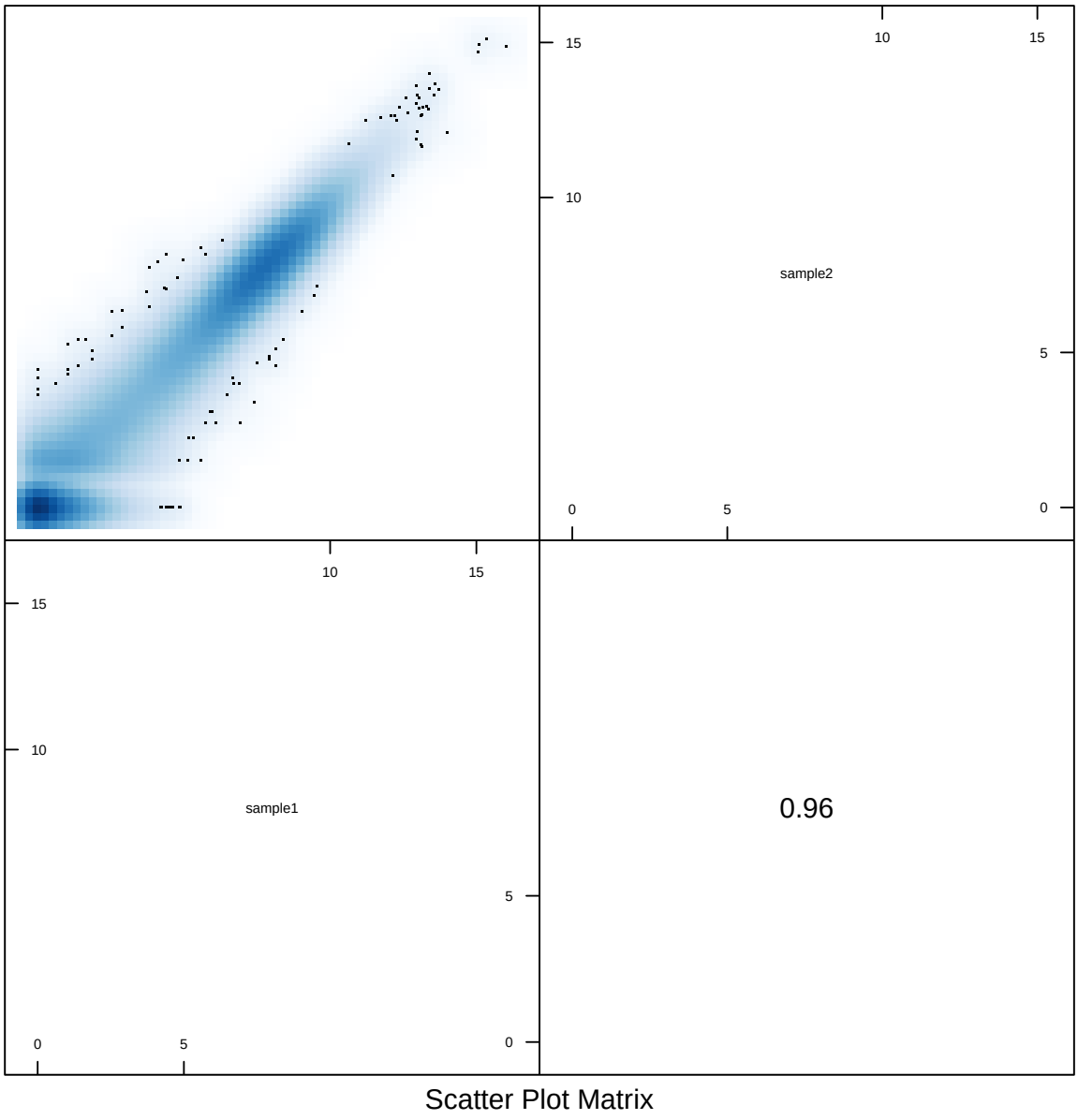


Figure 3: Scatter plot matrix of replicates. The scatter plots show log2(x+1) values. The spearman correlation in the boxes is calculated on the non-logged normalized counts.

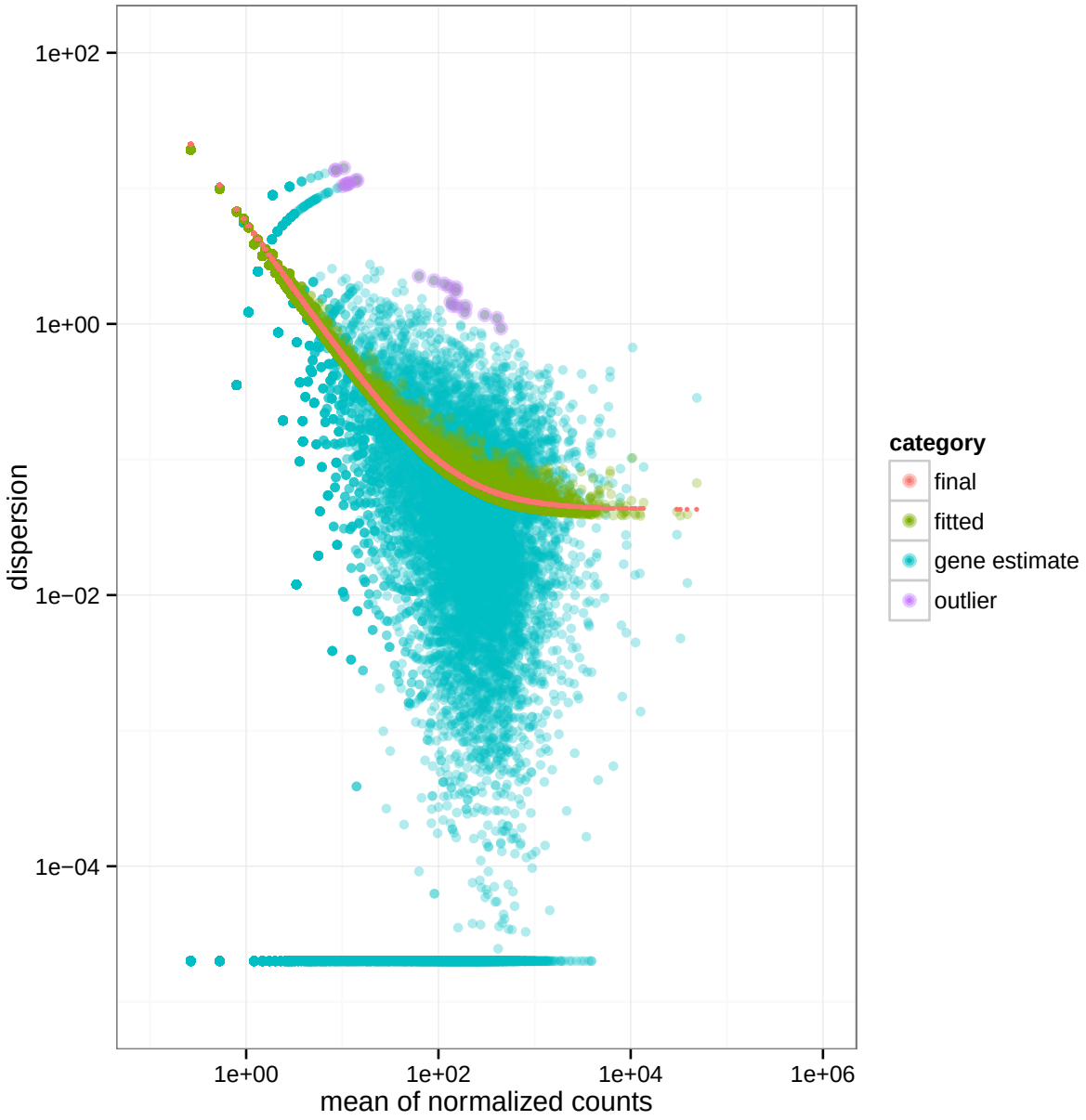


Figure 4: Estimate of the variance per gene. Density plot of empirical sample variance against means of normalized counts. The red line shows the fitted values.

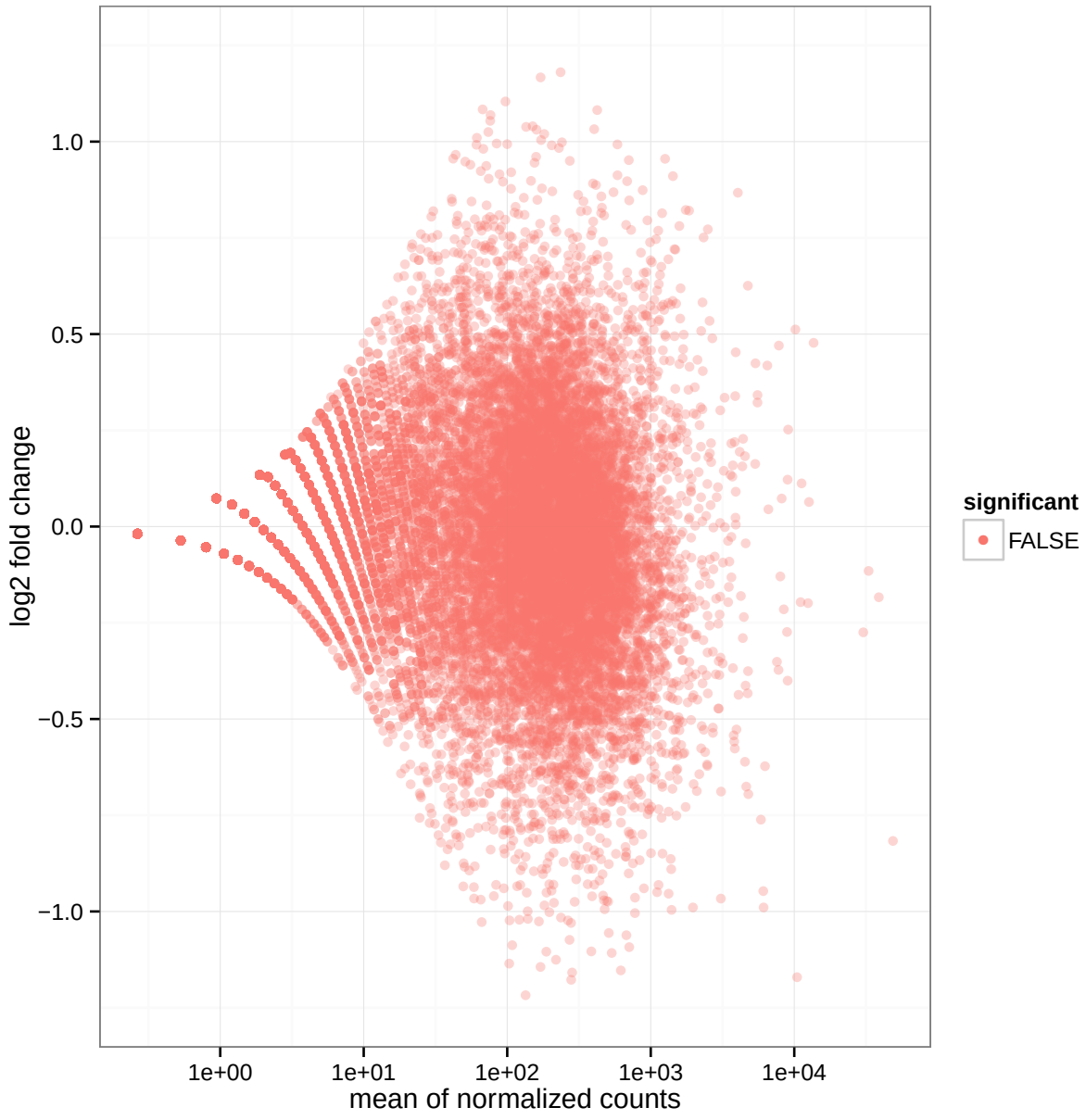


Figure 5: Differential expression and normalization. Scatter plot of log2 fold change against means of normalized counts. Genes with an adjusted p-value <0.05 are highlighted in red. Deviations of the bulk from the horizontal line indicate normalization problems. Significantly changed genes should not aggregate at a certain expression value (e.g. lowly expressed genes). Espression values are normalised especially at lower read counts, where overplotting results in grey dots.

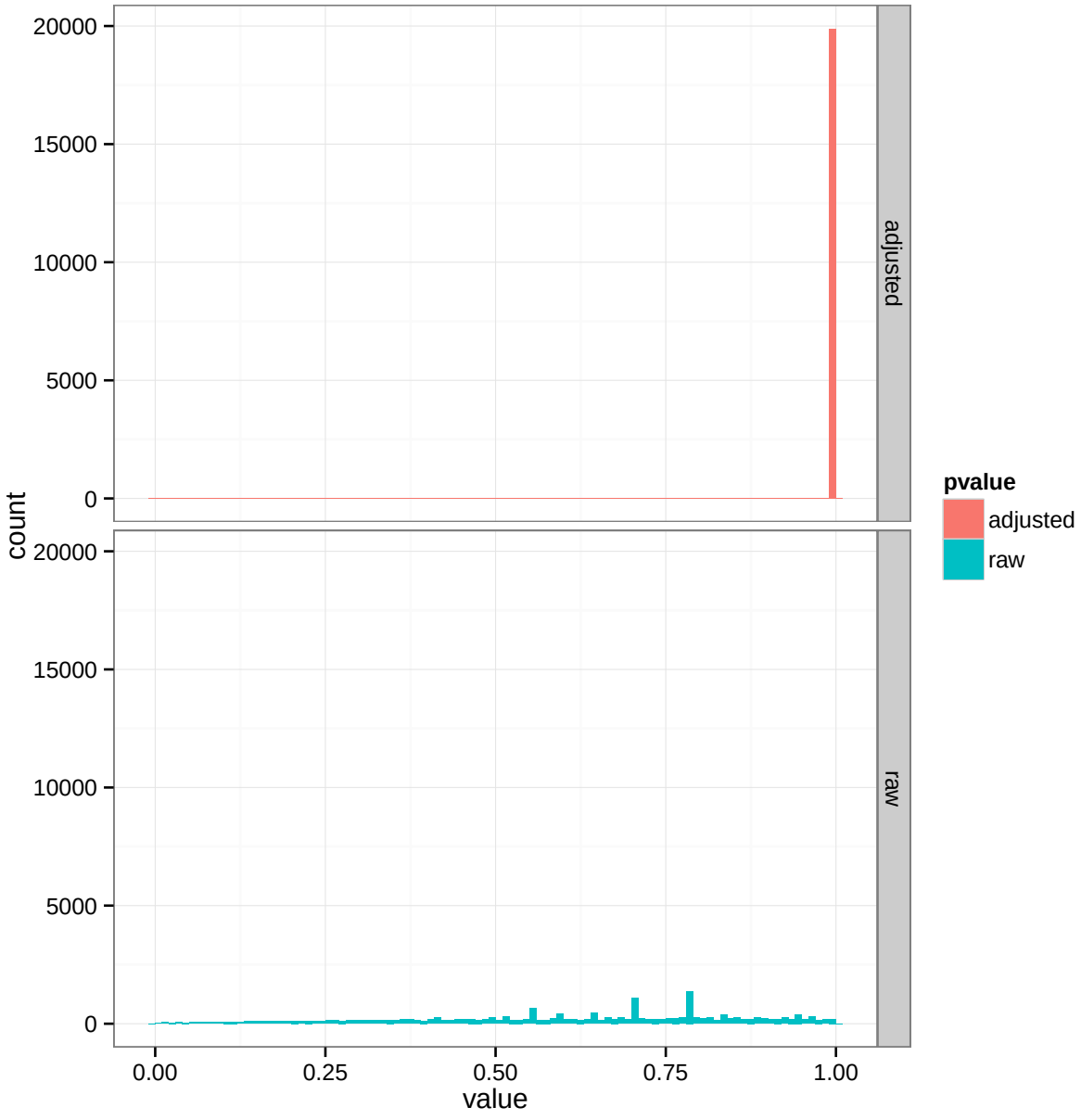


Figure 6: Distribution of p-values raw and Benjamini-Hochberg adjusted [2].

Figure 7: Plot of genes ranked by the row sums of the normalized expression value vs negative logarithm of p-values. The significance threshold was adjusted p-value of <0.05.

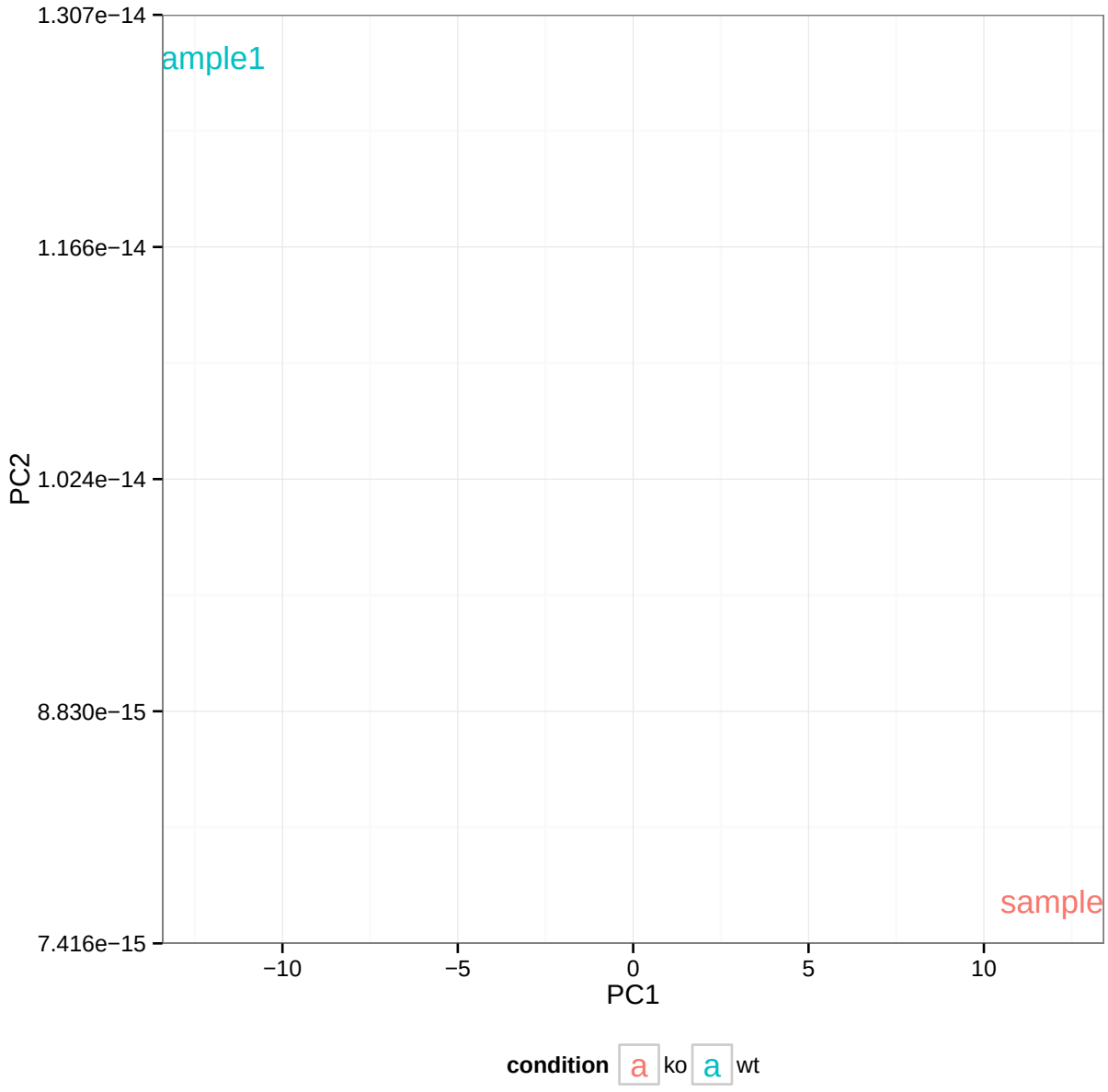


Figure 8: PCA plot of the vsn normalized expression values, normalized independently of the known conditions, from the top 500 genes with the higest variance. This plot illustrates distances between samples.