Contents

Basic edgeR results exploration

Project: edgeR PDF report.

Introduction

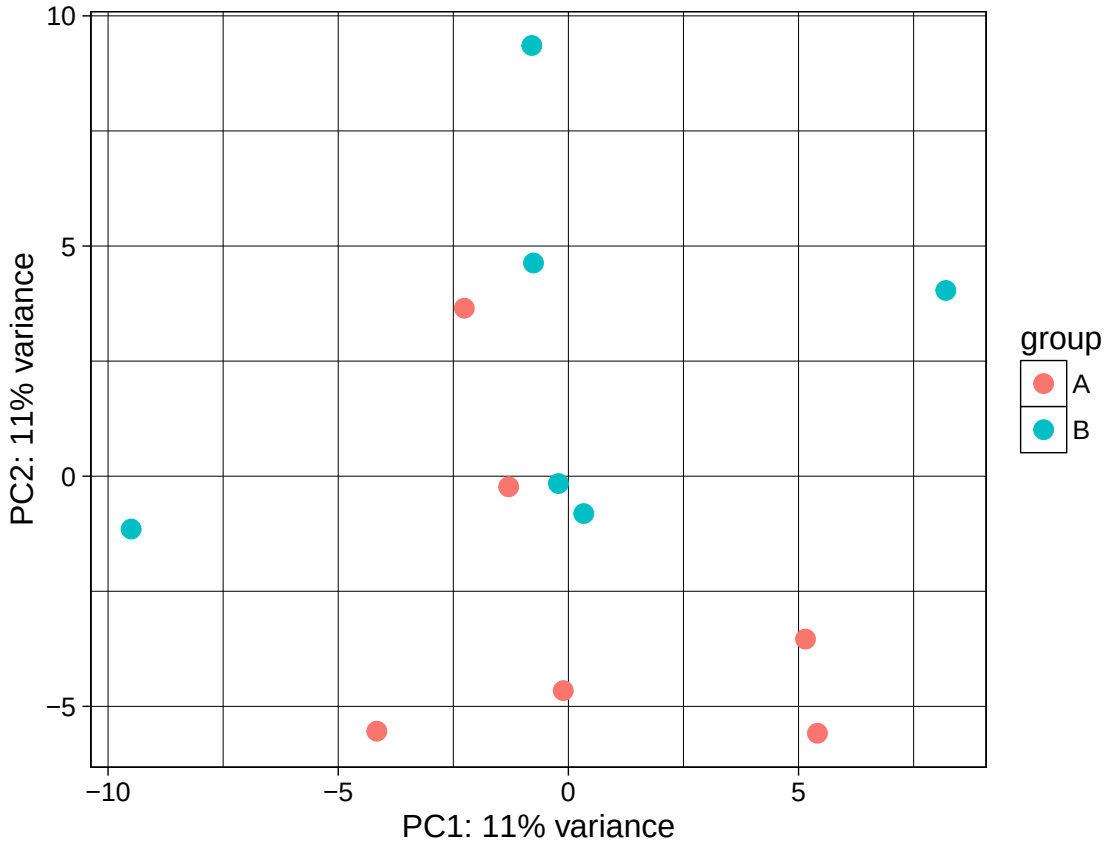
This report is meant to help explore edgeR (McCarthy, J., Chen, Yunshun, et al., 2012; Zhou, Lindsay, and Robinson, 2014; Chen, Lun, and Smyth, 2014) results and was generated using the `regionReport` (Collado-Torres, Jaffe, and Leek, 2015) package. While the report is rich, it is meant to just start the exploration of the results and exemplify some of the code used to do so. If you need a more in-depth analysis for your specific data set you might want to use the `customCode` argument. This report is based on the vignette of the `DESeq2` (Love, Huber, and Anders, 2014) package which you can find here.

Code setup

This section contains the code for setting up the rest of the report.

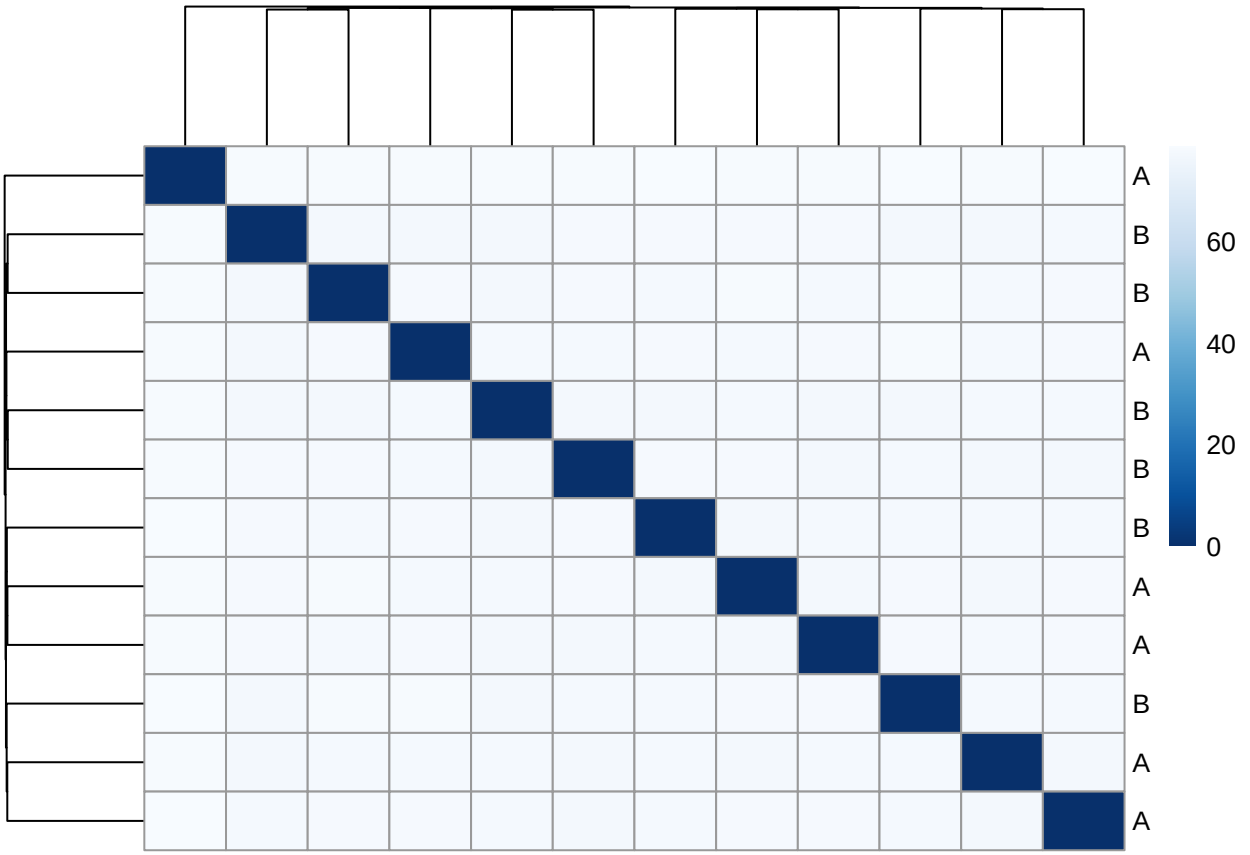
```
## knitrBoostrap and device chunk options
load_install('knitr')
opts_chunk$set(bootstrap.show.code = FALSE, dev = device)
if(!outputIsHTML) opts_chunk$set(bootstrap.show.code = FALSE, dev = device, echo = FALSE)
```

PCA



The above plot shows the first two principal components that explain the variability in the data using the regularized log count data. If you are unfamiliar with principal component analysis, you might want to check the Wikipedia entry or this interactive explanation. In this case, the first and second principal component explain 11 and 11 percent of the variance respectively.

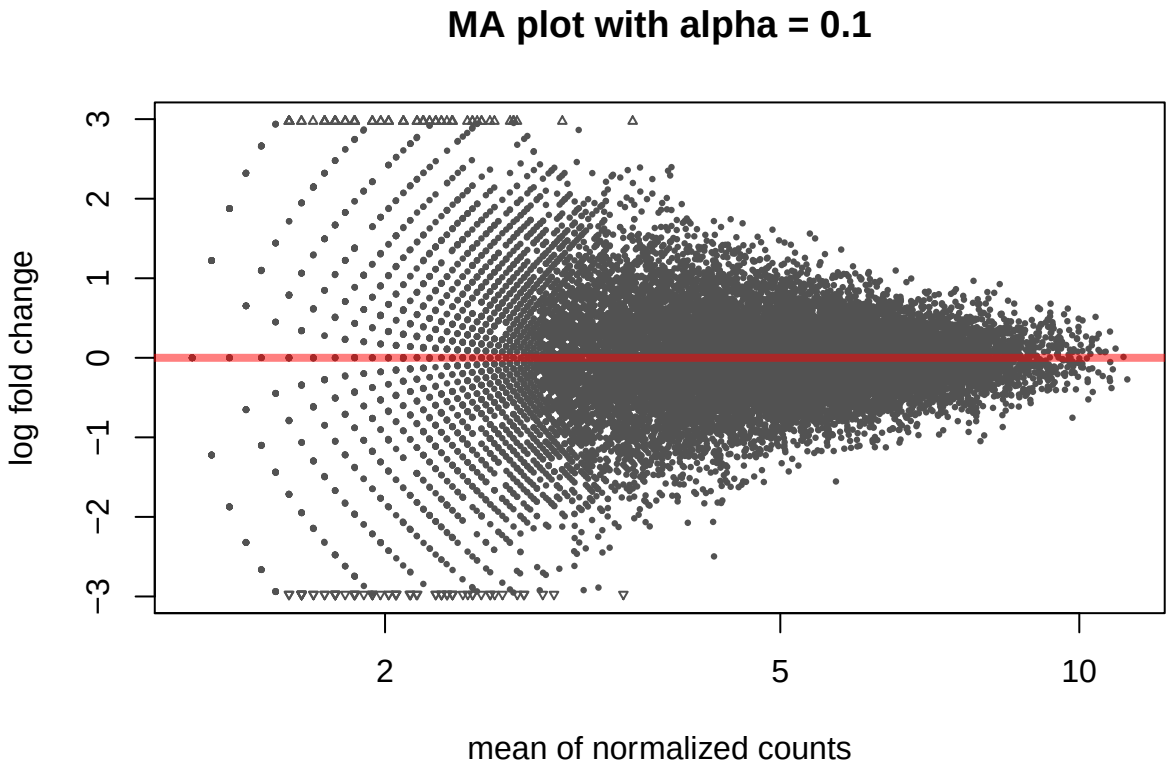
Sample-to-sample distances



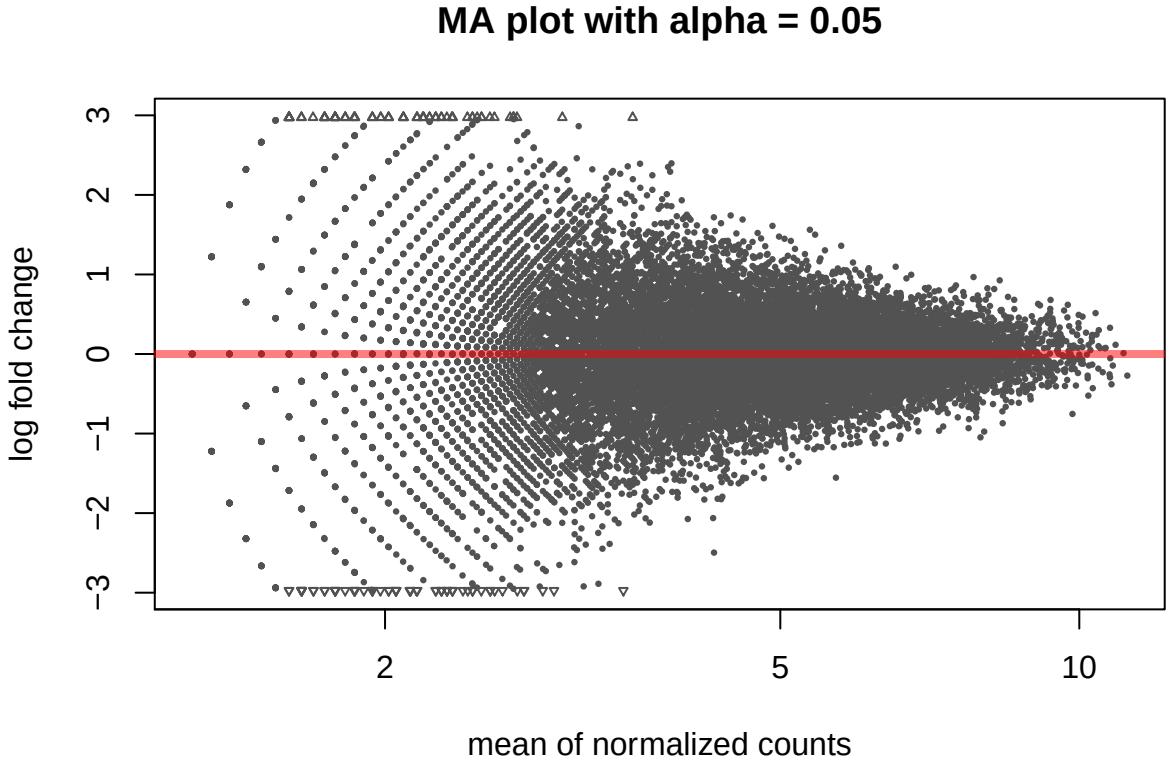
This plot shows how samples are clustered based on their euclidean distance using the regularized log transformed count data. This figure gives an overview of how the samples are hierarchically clustered. It is a complementary figure to the PCA plot.

MA plots

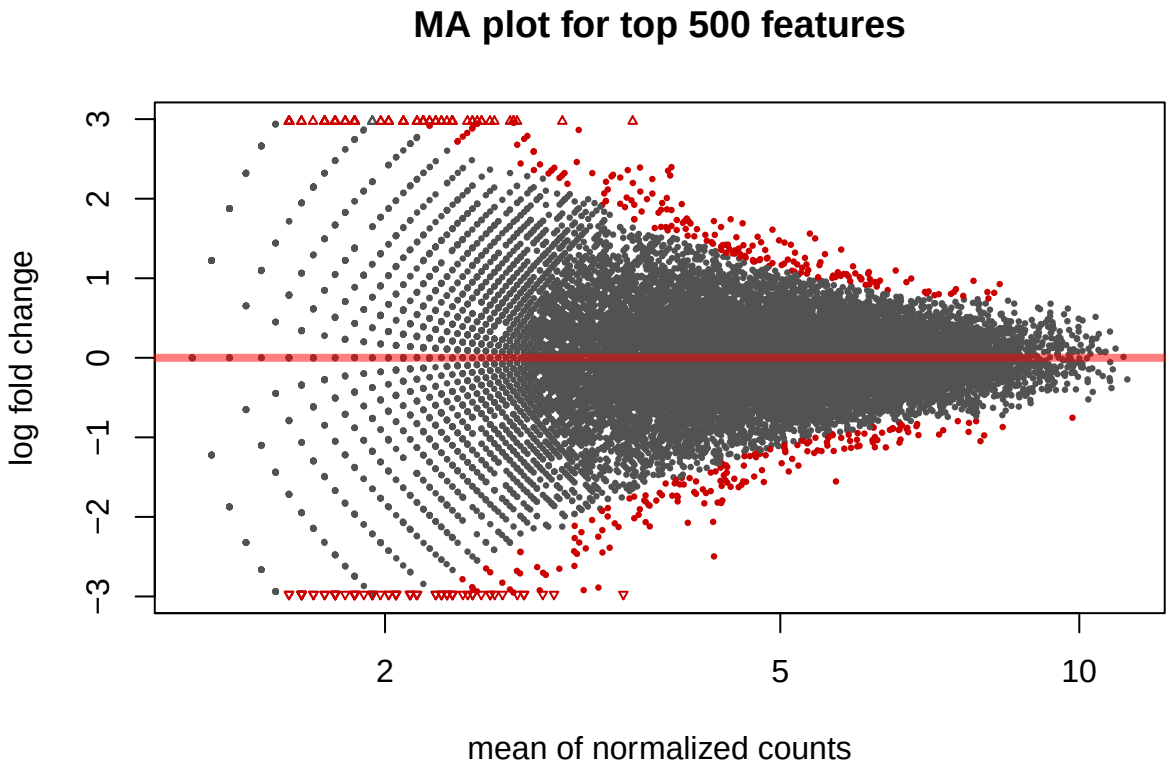
This section contains three MA plots (see Wikipedia) that compare the mean of the normalized counts against the log fold change. They show one point per feature. The points are shown in red if the feature has an adjusted p-value less than `alpha`, that is, the statistically significant features are shown in red.



This first plot shows uses `alpha` = 0.1, which is the `alpha` value used to determine which resulting features were significant when running the function `DESeq2::results()`.

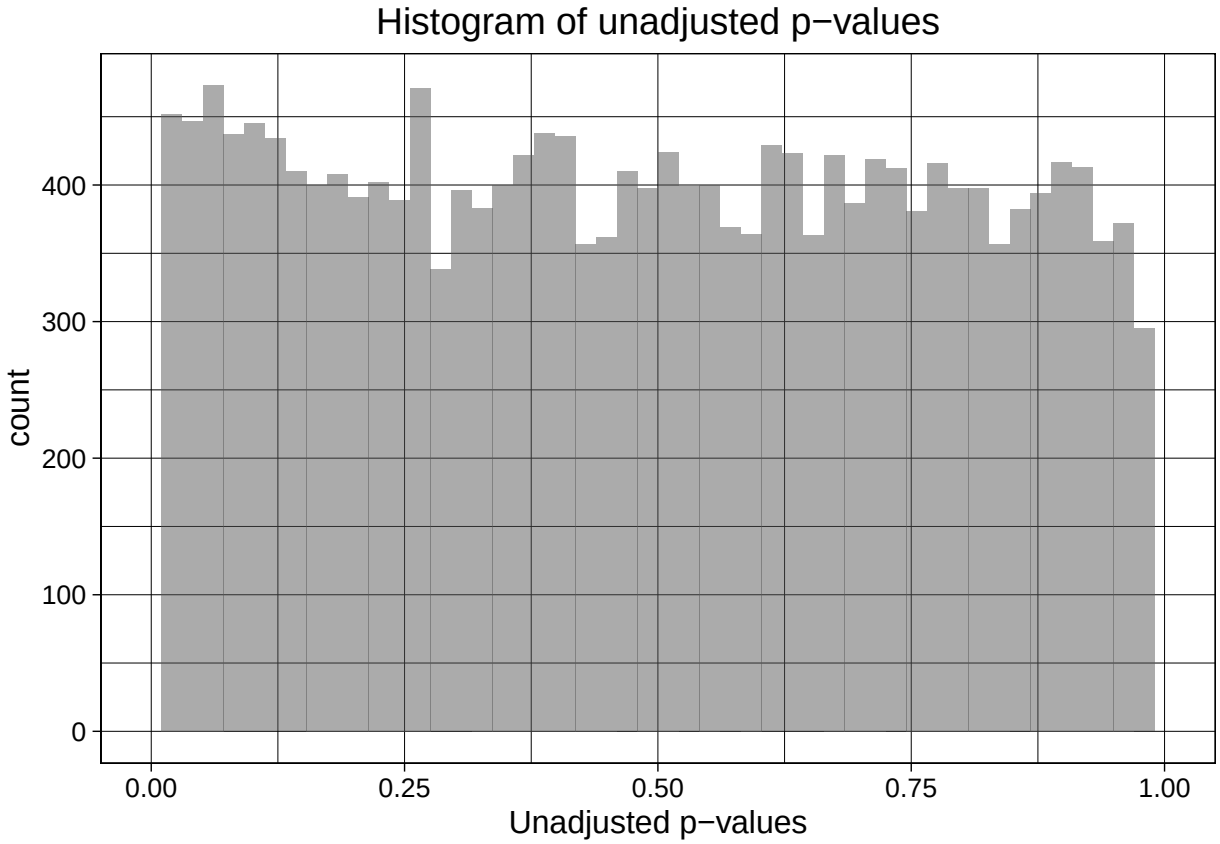


This second MA plot uses `alpha` = 0.05 and can be used agains the first MA plot to identify which features have adjusted p-values between 0.05 and 0.1.



The third and final MA plot uses an alpha such that the top 500 features are shown in the plot. These are the features that whose details are included in the *top features* interactive table.

P-values distribution



This plot shows a histogram of the unadjusted p-values. It might be skewed right or left, or flat as shown in the Wikipedia examples. The shape depends on the percent of features that are differentially expressed. For further information on how to interpret a histogram of p-values check David Robinson's post on this topic.

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.0000116 0.2349000 0.4901000 0.4918000 0.7446000 1.0000000
```

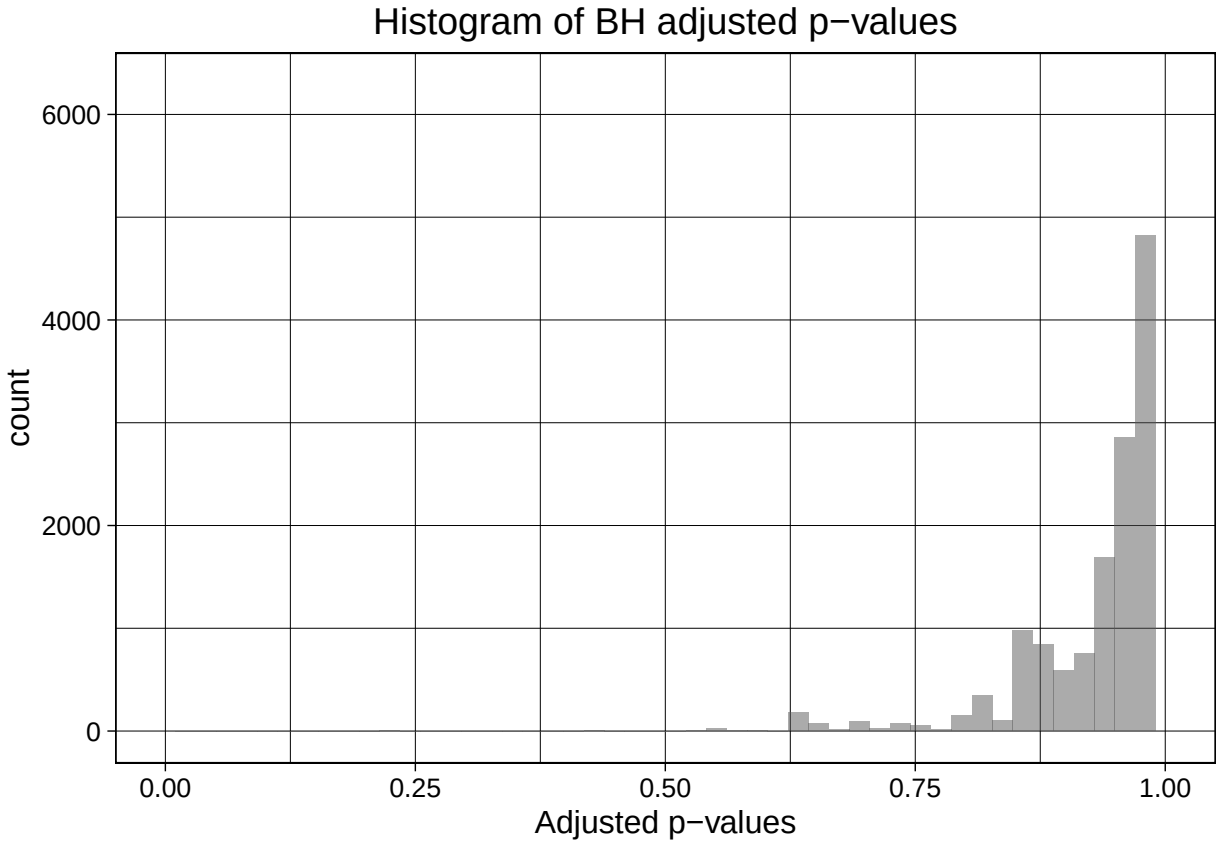
This is the numerical summary of the distribution of the p-values.

Cut	Count
0.0001	3
0.0010	35
0.0100	298
0.0250	616
0.0500	1174
0.1000	2283
0.2000	4315
0.3000	6261
0.4000	8286
0.5000	10194
0.6000	12114
0.7000	14095
0.8000	16078
0.9000	17999

Cut	Count
1.0000	20000

This table shows the number of features with p-values less or equal than some commonly used cutoff values.

Adjusted p-values distribution



This plot shows a histogram of the BH adjusted p-values. It might be skewed right or left, or flat as shown in the Wikipedia examples.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2327  0.9383  0.9795  0.9508  0.9923  1.0000
```

This is the numerical summary of the distribution of the BH adjusted p-values.

Cut	Count
0.0001	0
0.0010	0
0.0100	0
0.0250	0
0.0500	0
0.1000	0
0.2000	0

Cut	Count
0.3000	1
0.4000	1
0.5000	2
0.6000	36
0.7000	402
0.8000	603
0.9000	3381
1.0000	20000

This table shows the number of features with BH adjusted p-values less or equal than some commonly used cutoff values.

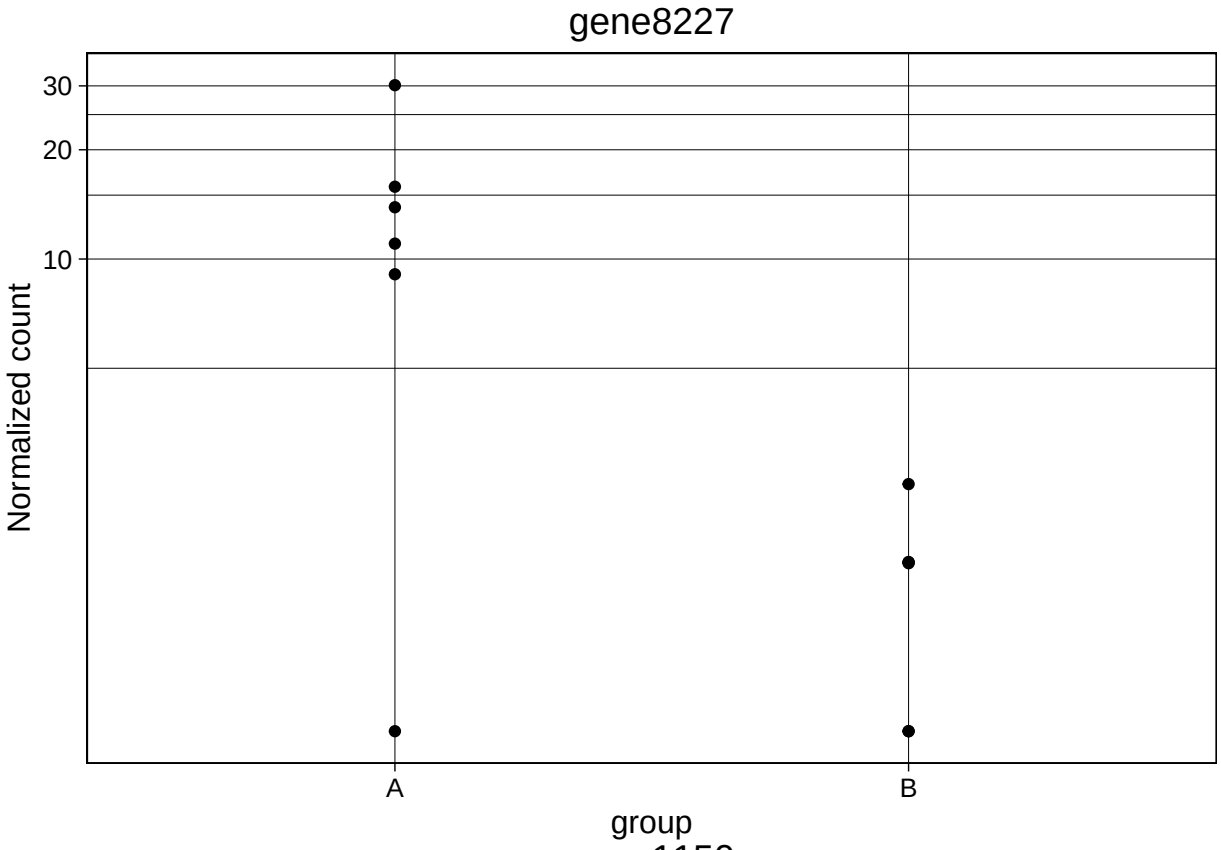
Top features

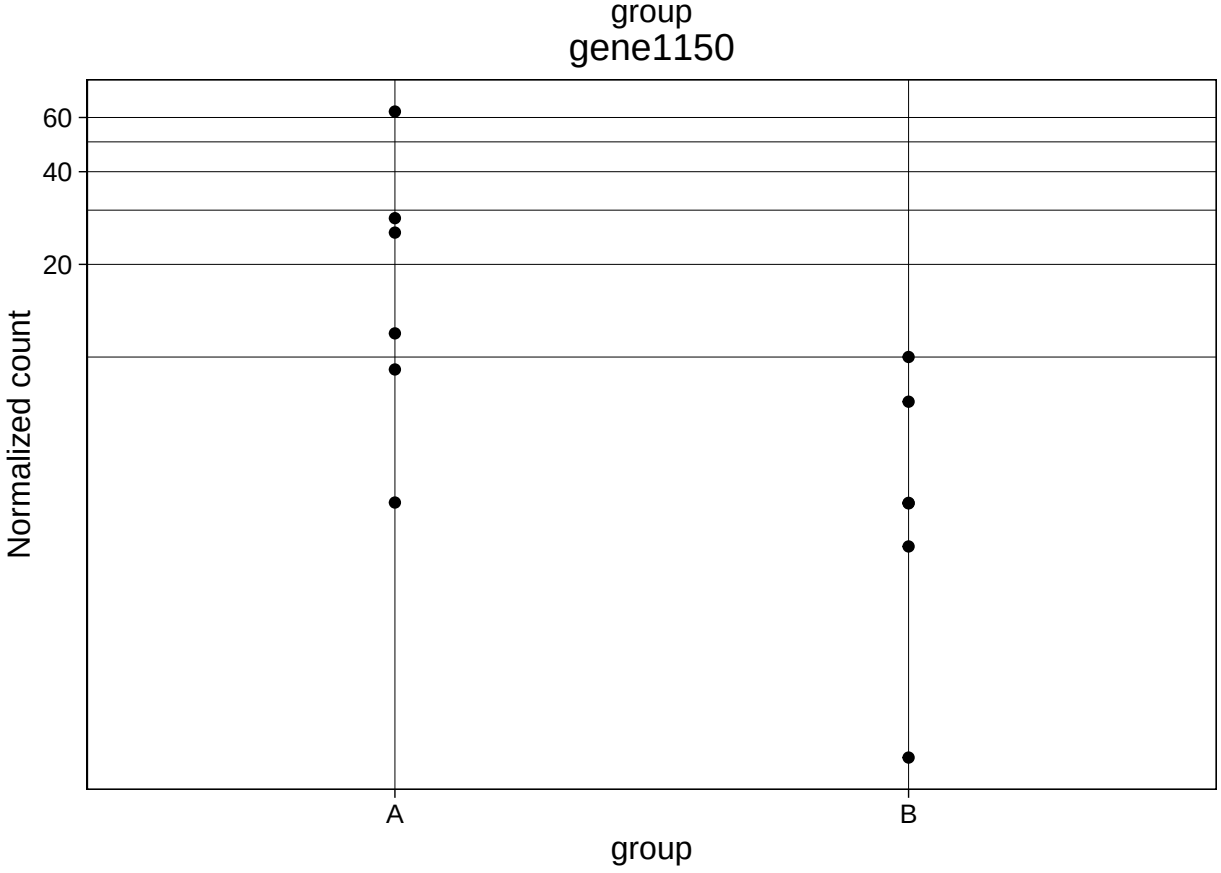
This table shows the top 500 features ordered by their BH adjusted p-values. Since the report is in PDF format, only the top 20 features are shown.

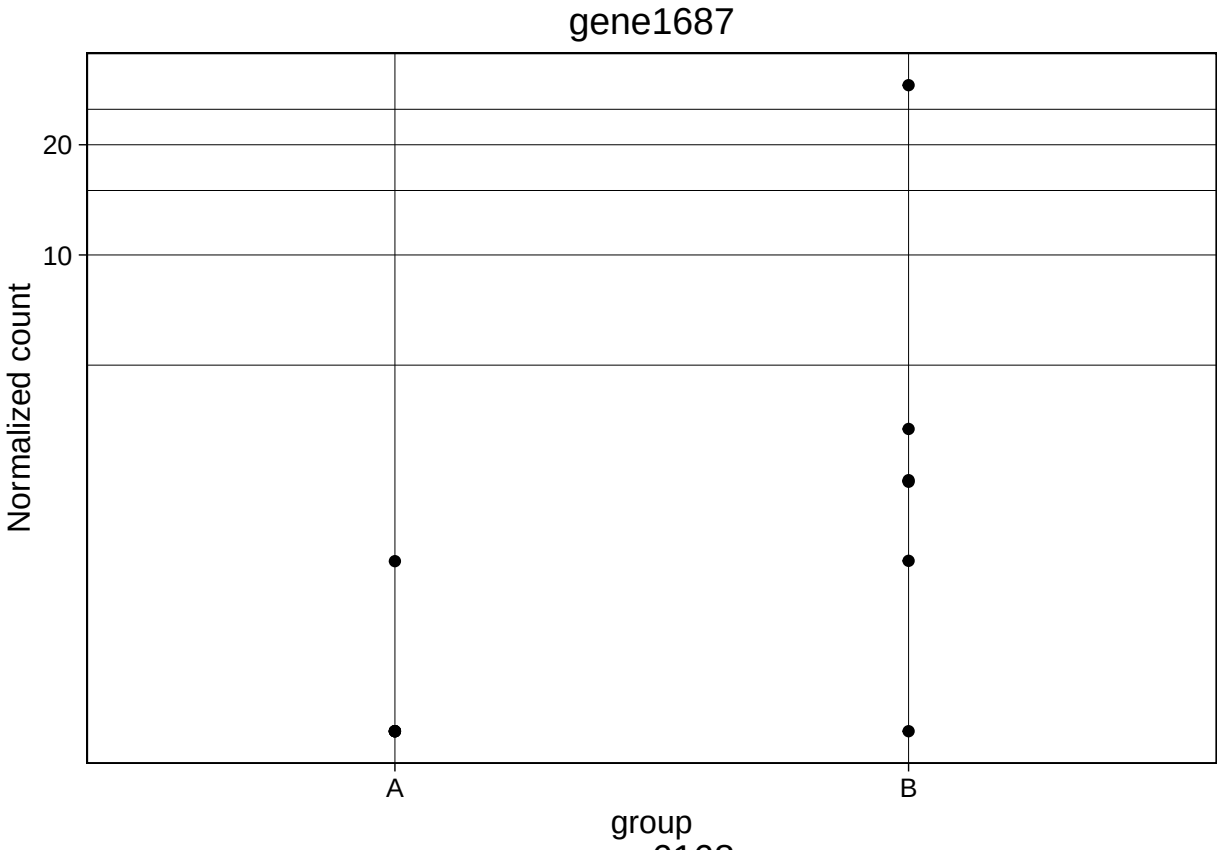
Feature	log2FoldChange	baseMean	LR	pvalue	padj
gene8227	-3.83	3.48	19.22	1.163289e-05	2.326579e-01
gene1150	-2.50	4.29	16.77	4.217729e-05	4.217729e-01
gene1687	4.47	2.67	13.70	2.146930e-04	5.376917e-01
gene6168	-5.05	2.28	13.69	2.150767e-04	5.376917e-01
gene8748	4.50	2.69	13.96	1.867848e-04	5.376917e-01
gene10003	3.10	3.55	14.87	1.151489e-04	5.376917e-01
gene11742	-1.55	5.69	14.34	1.522844e-04	5.376917e-01
gene17664	-5.26	2.40	15.22	9.570850e-05	5.376917e-01
gene1428	4.09	2.45	10.90	9.599831e-04	5.485618e-01
gene1558	-2.06	4.28	11.96	5.445798e-04	5.485618e-01
gene1987	4.86	2.19	12.48	4.103770e-04	5.485618e-01
gene3012	-3.38	2.96	11.20	8.168382e-04	5.485618e-01
gene3192	2.35	3.85	11.55	6.780950e-04	5.485618e-01
gene5006	4.18	2.50	11.54	6.799674e-04	5.485618e-01
gene5588	4.86	2.19	12.47	4.131867e-04	5.485618e-01
gene5621	-4.79	2.15	12.07	5.124740e-04	5.485618e-01
gene6735	2.40	3.89	12.18	4.839640e-04	5.485618e-01
gene9636	1.36	5.76	11.46	7.109156e-04	5.485618e-01
gene10330	4.65	2.09	11.34	7.571205e-04	5.485618e-01
gene11082	4.65	2.09	11.35	7.538685e-04	5.485618e-01

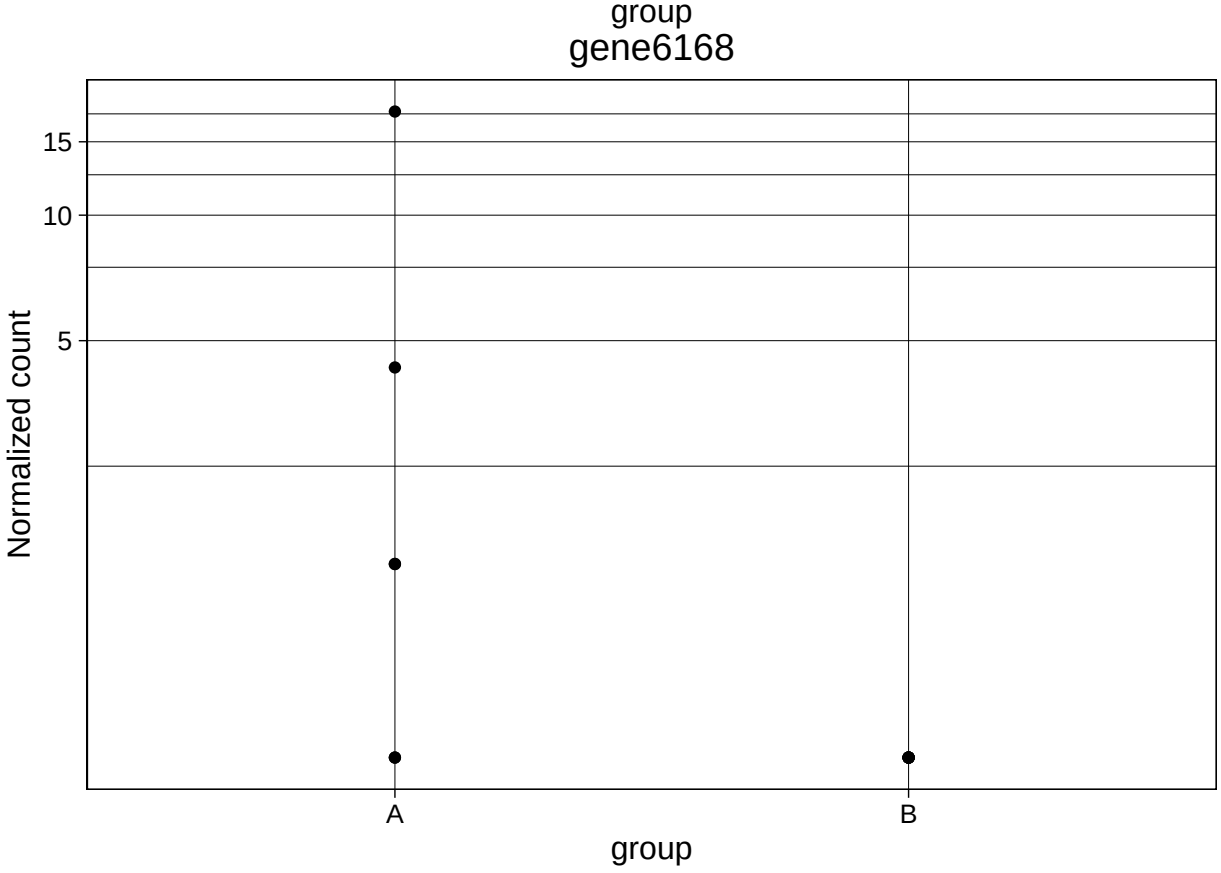
Count plots top features

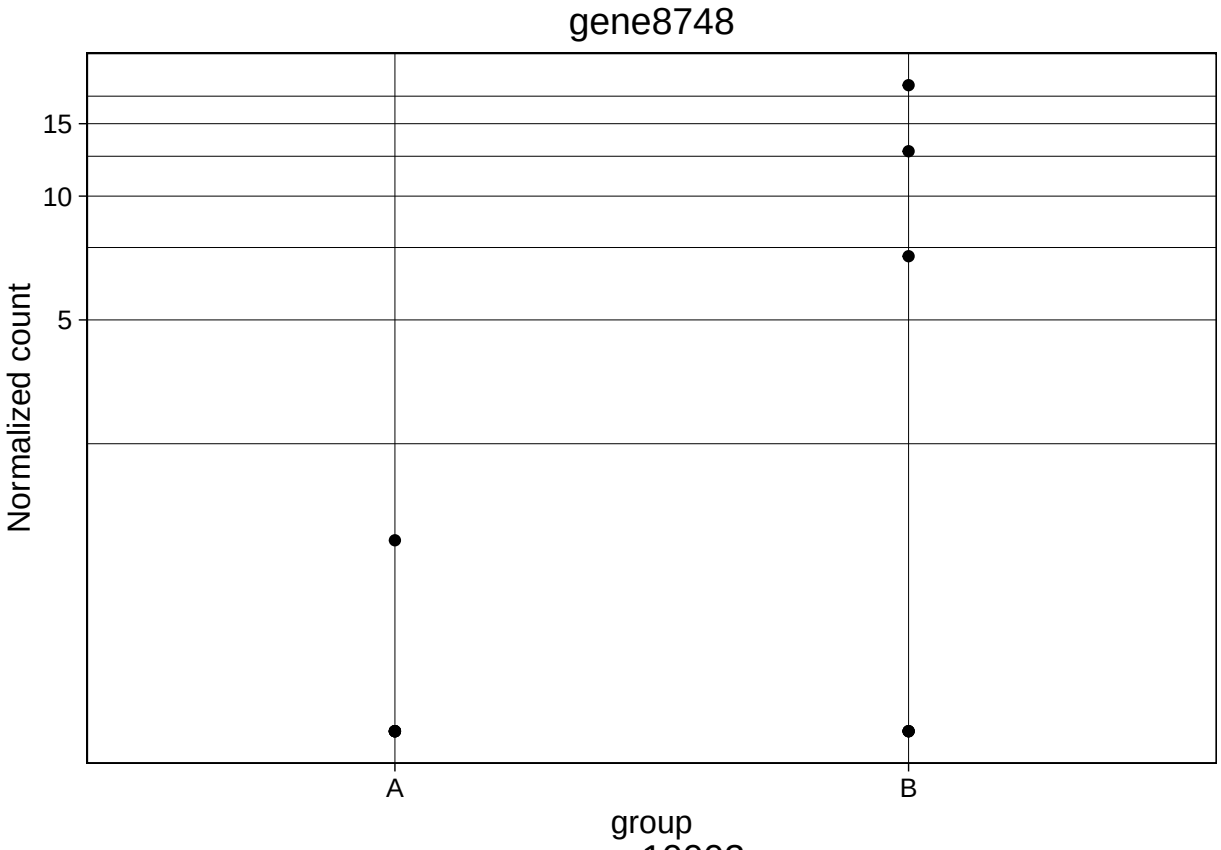
This section contains plots showing the normalized counts per sample for each group of interest. Only the best 20 features are shown, ranked by their BH adjusted p-values. The Y axis is on the log10 scale and the feature name is shown in the title of each plot.

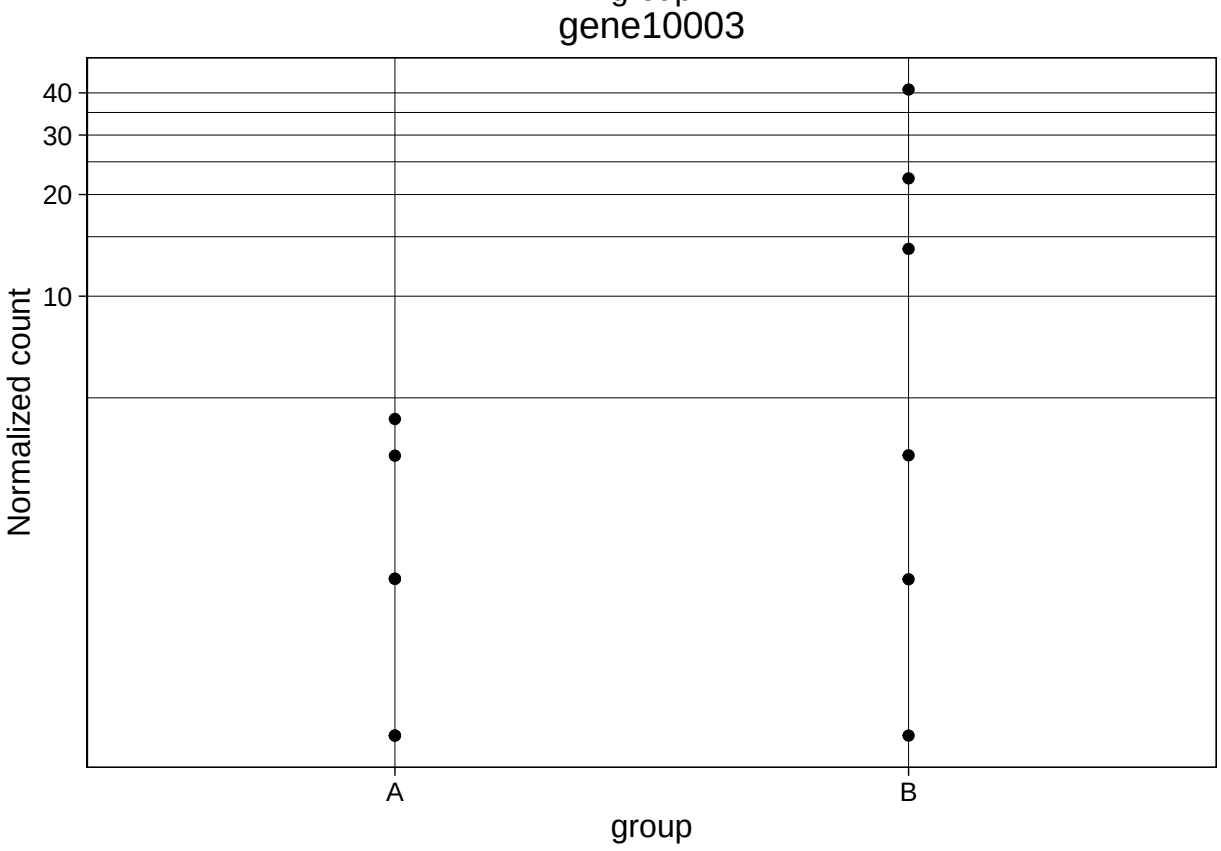

gene8227
30
20
10
Normalized count
A
B
group

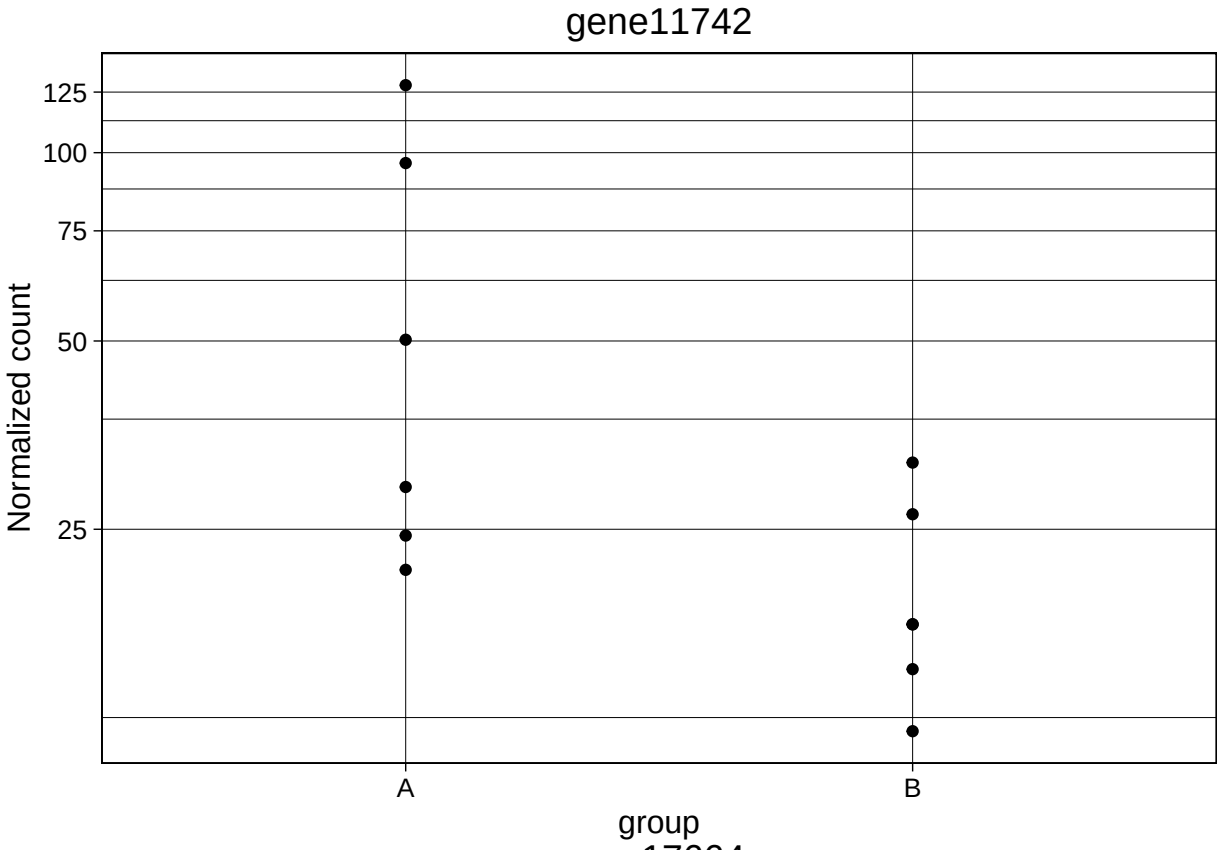

gene1150
60
40
20
Normalized count
A
B
group

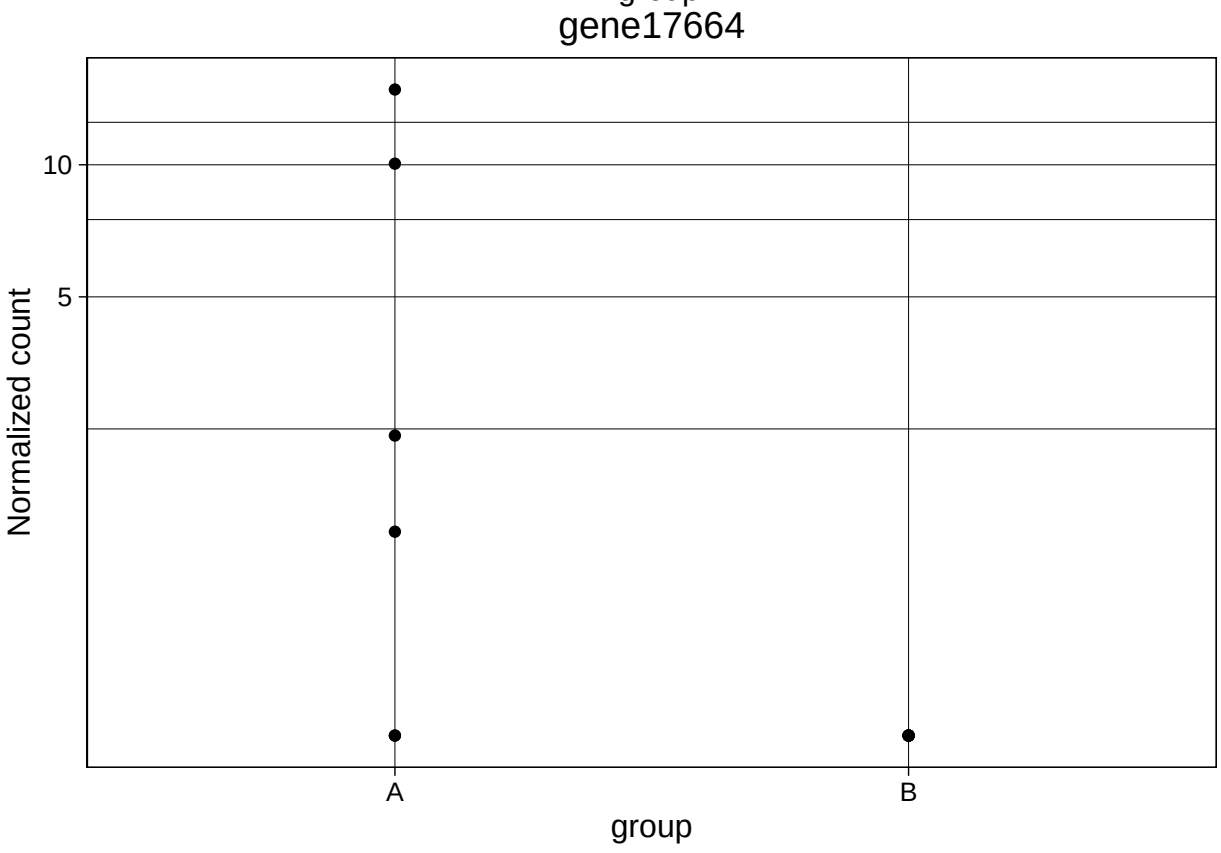

gene1687
20
10
Normalized count
A
B
group

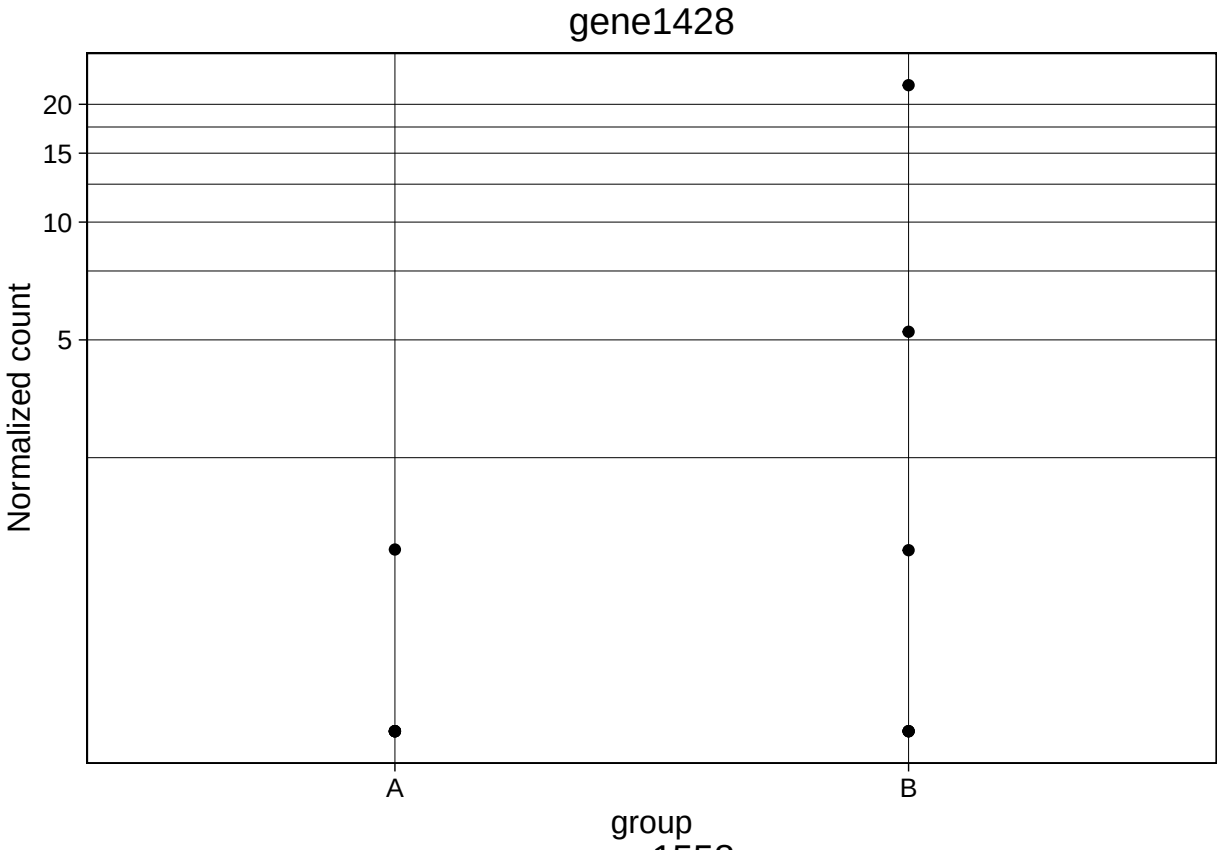

gene6168
15
10
5
Normalized count
A
B
group

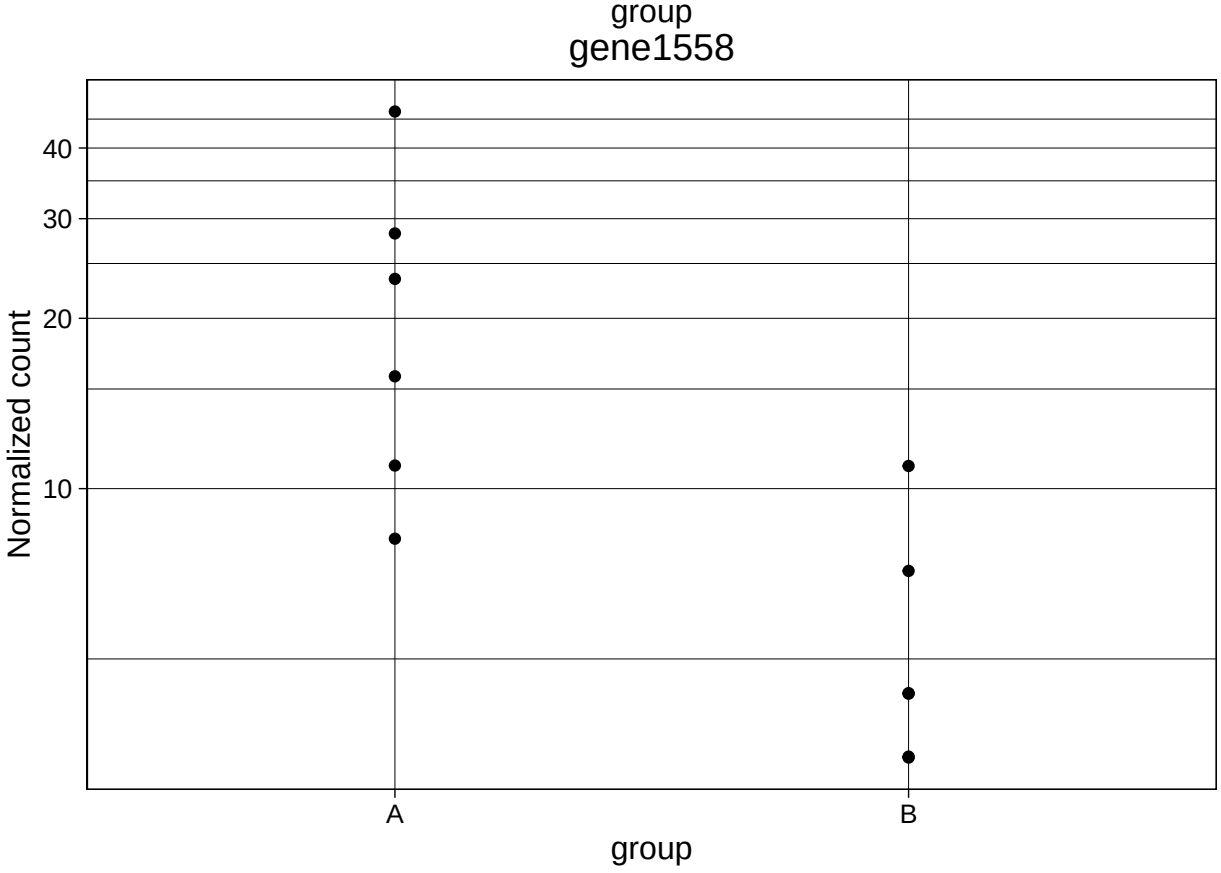

gene8748
15
10
5
Normalized count
A
B
group

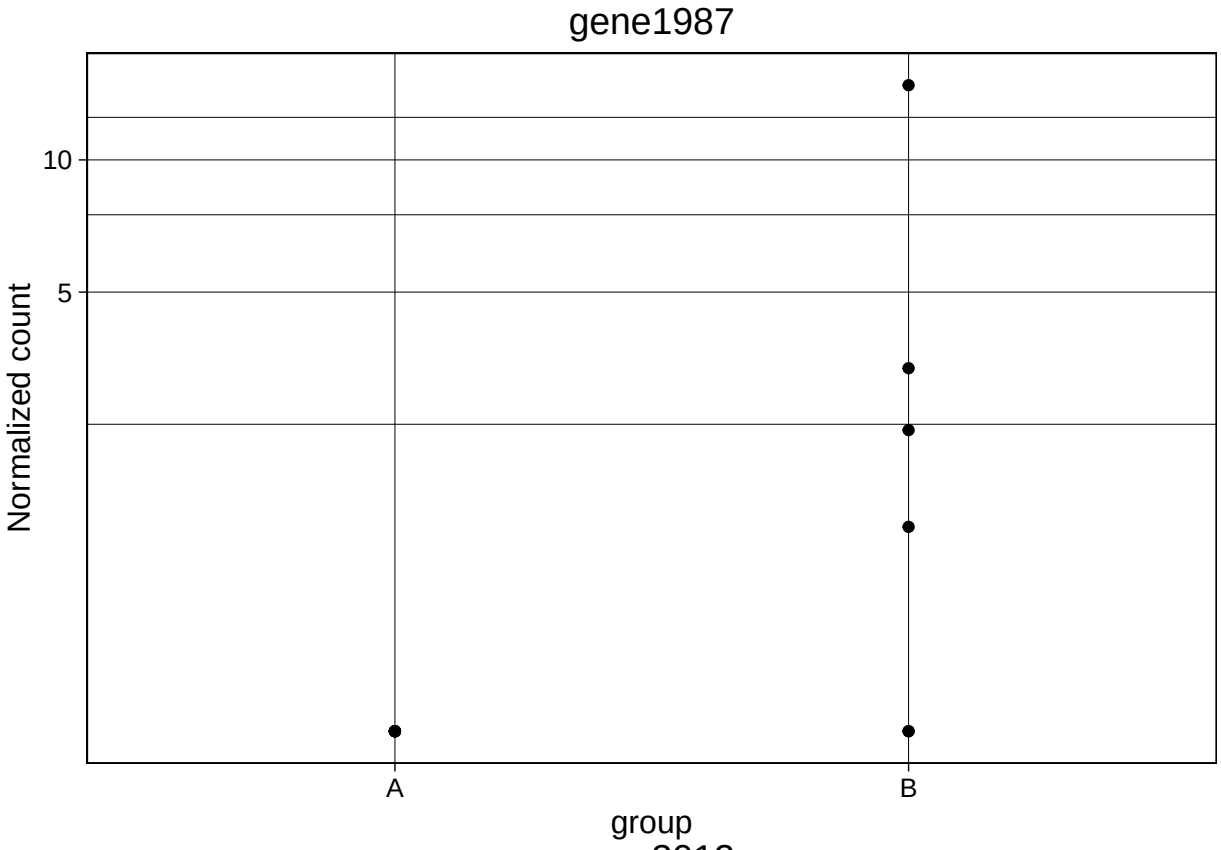

gene10003
40
30
20
10
Normalized count
A
B
group

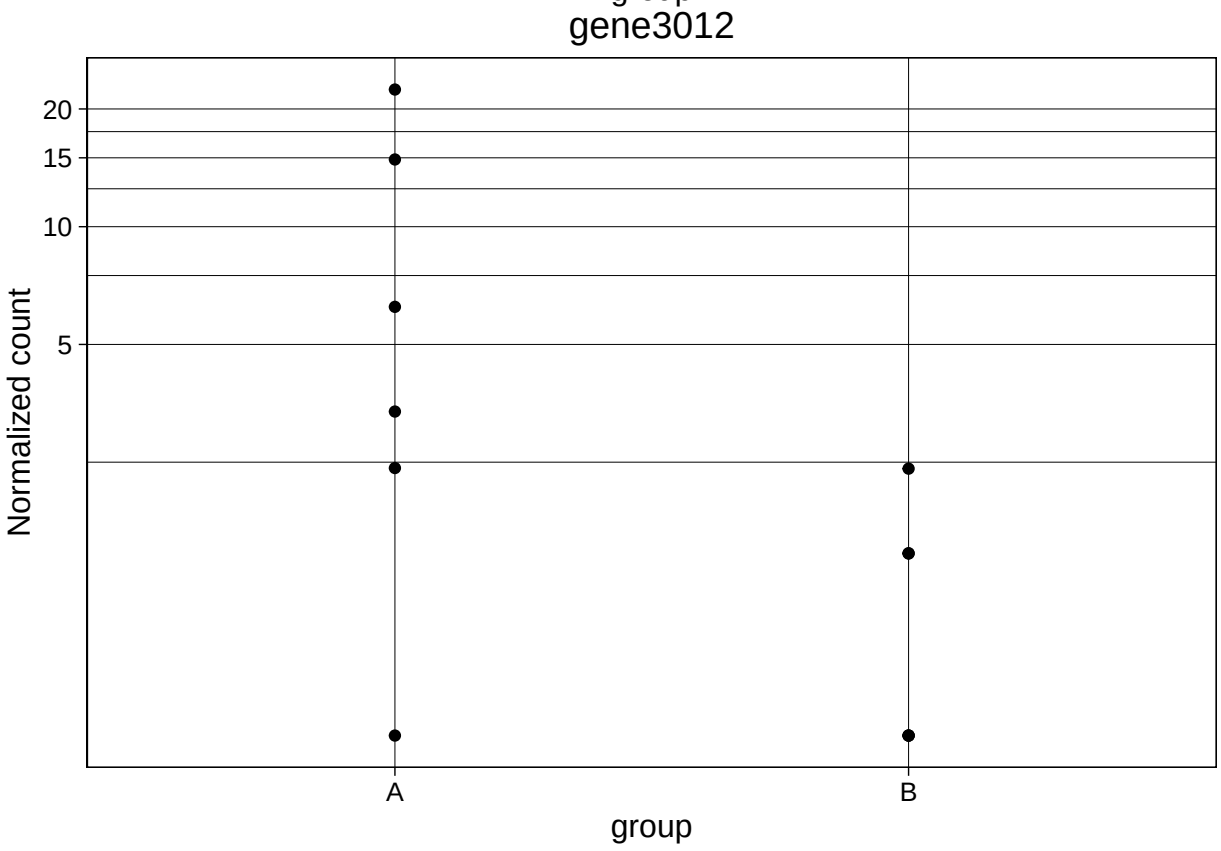

gene11742
125
100
75
50
25
Normalized count
A
B
group

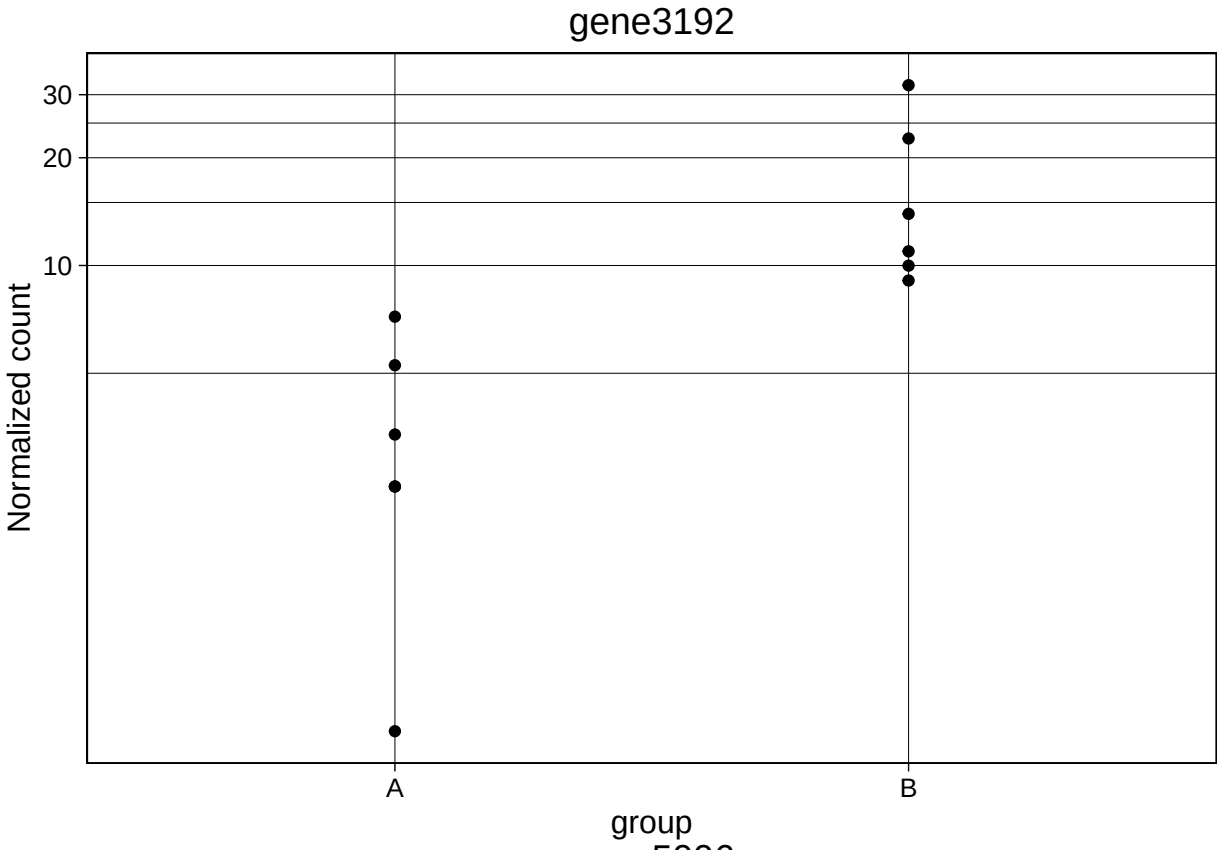

gene17664
10
5
Normalized count
A
B
group

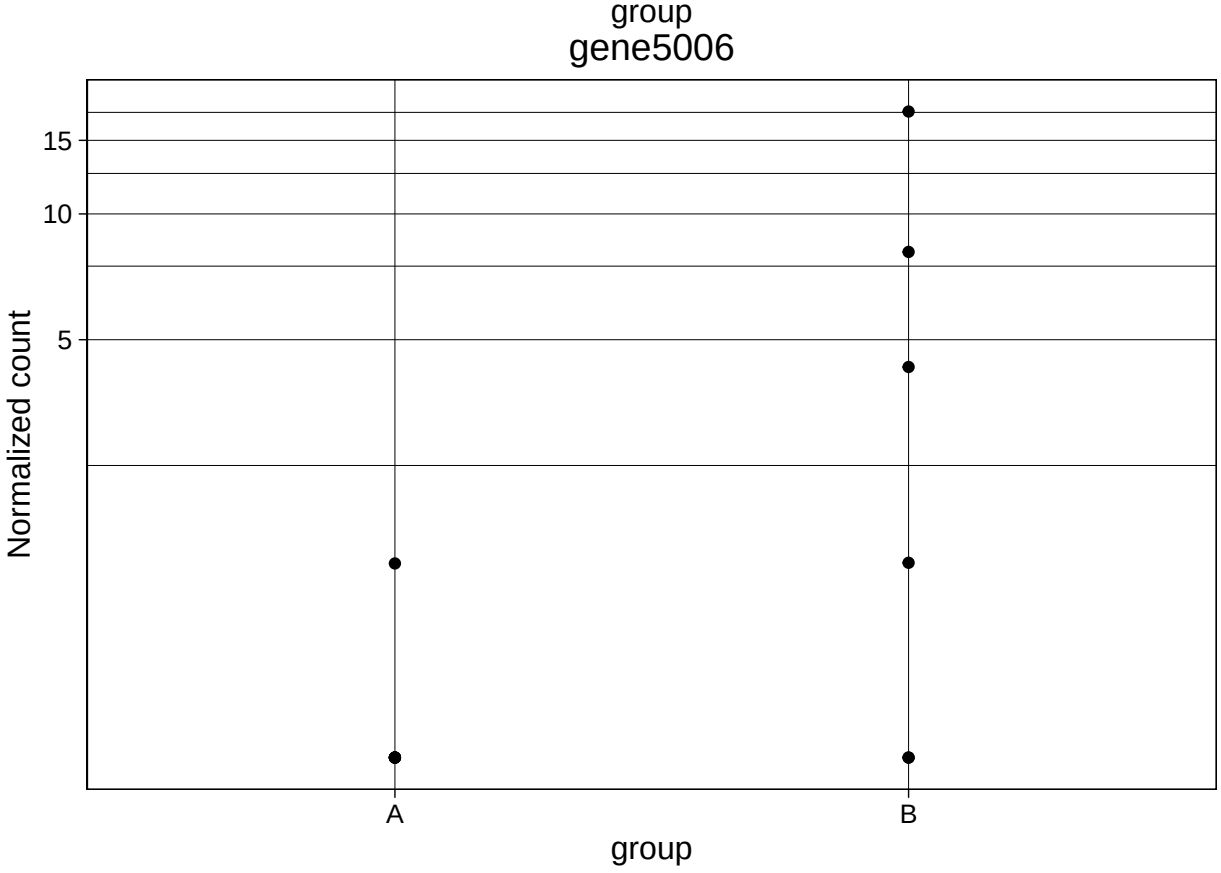

gene1428
Normalized count
20
15
10
5
A
B
group

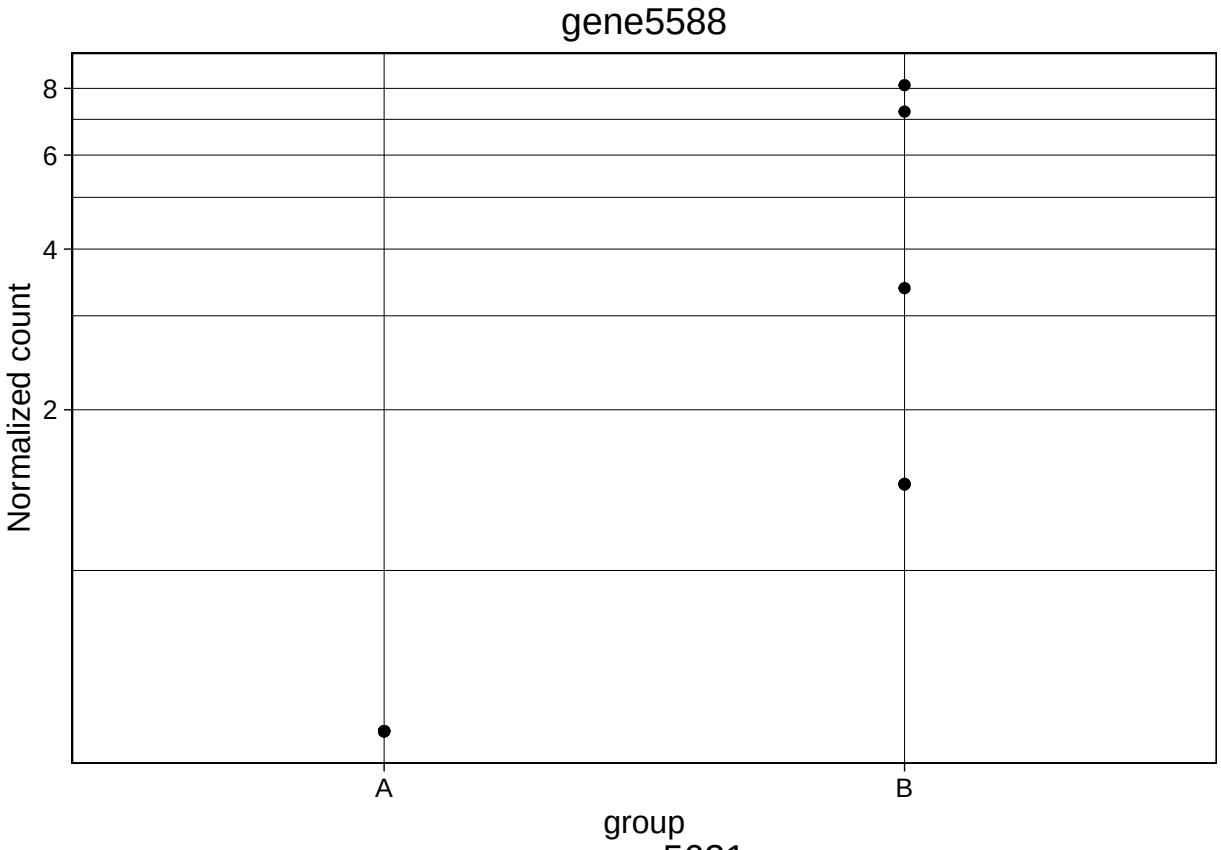

gene1558
Normalized count
40
30
20
10
A
B
group

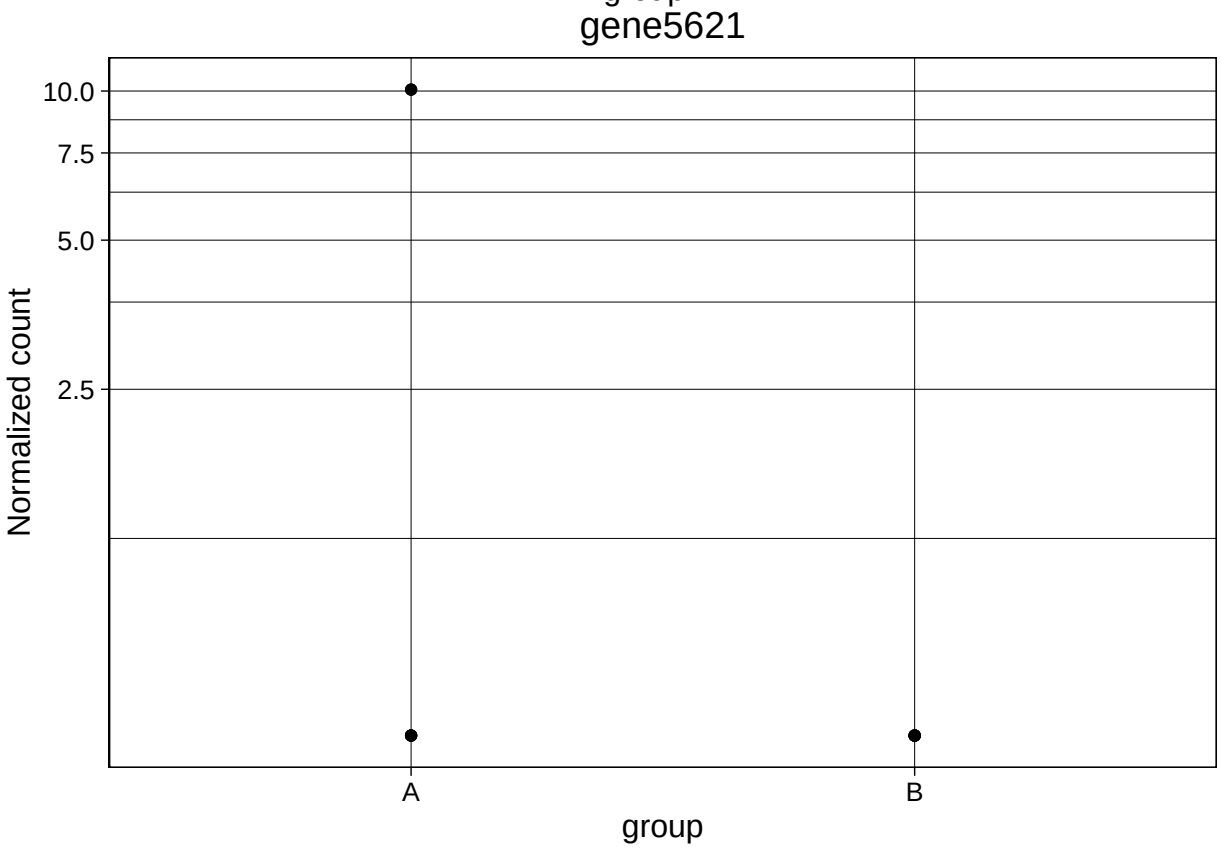

gene1987
Normalized count
10
5
A
B
group

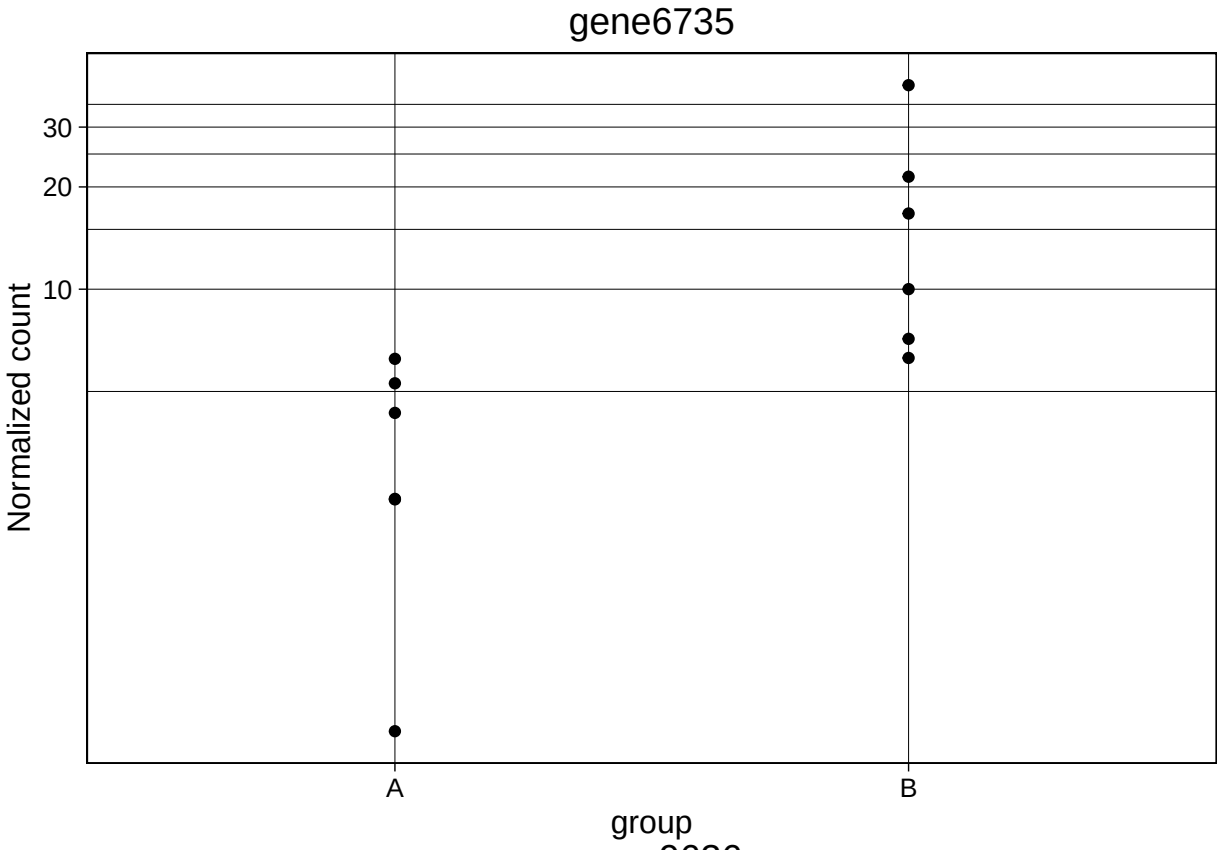

gene3012
Normalized count
20
15
10
5
A
B
group

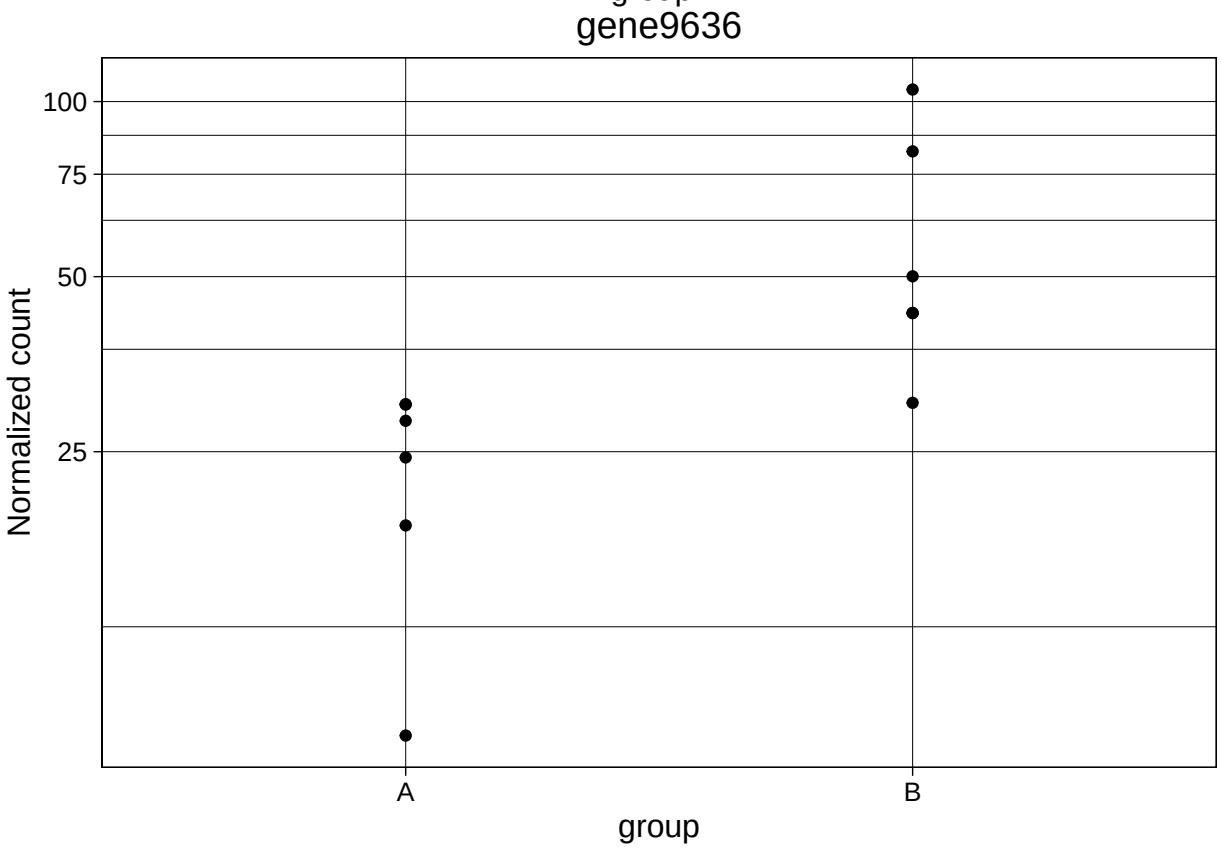

gene3192
30
20
10
Normalized count
A
B
group

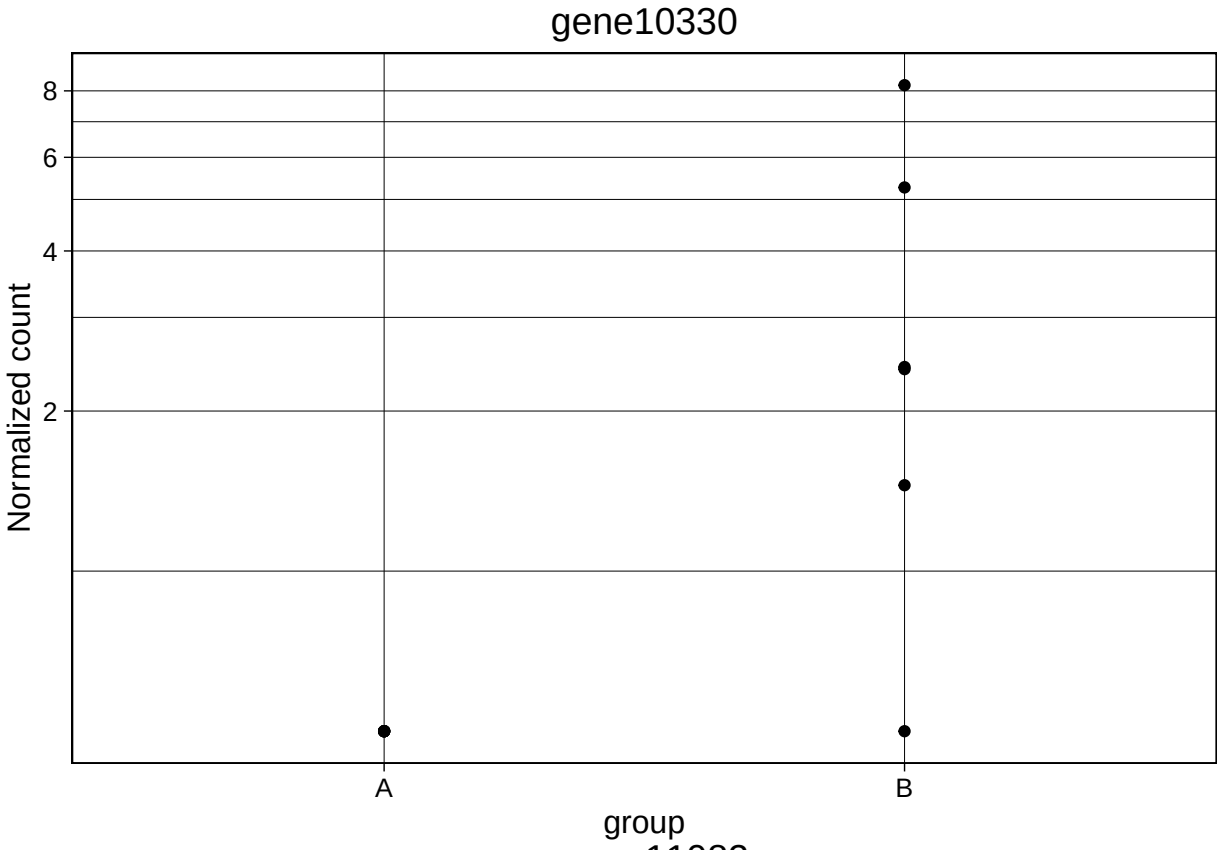

gene5006
15
10
5
Normalized count
A
B
group

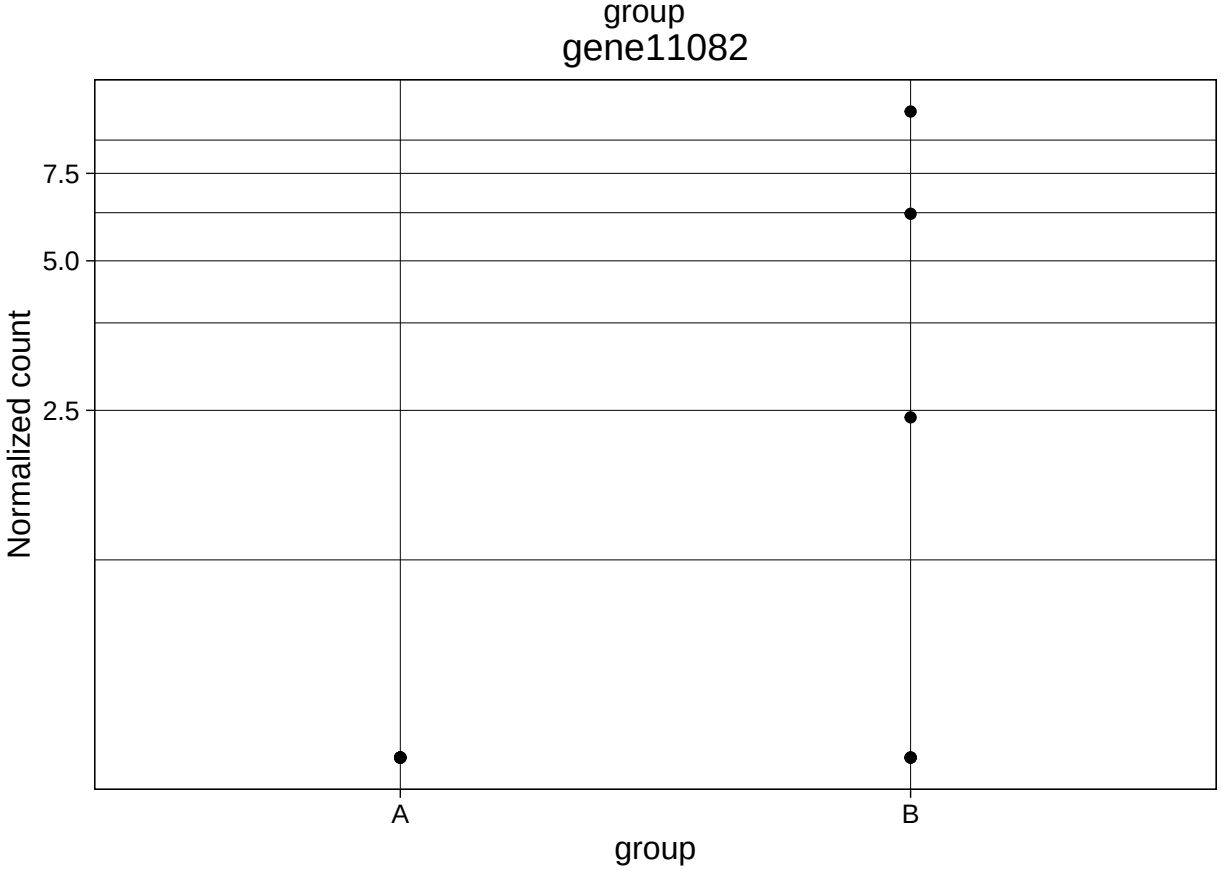

gene5588
8
6
4
2
Normalized count
A
B
group


gene5621
10.0
7.5
5.0
2.5
Normalized count
A
B
group


gene6735
Normalized count
30
20
10
A
B
group

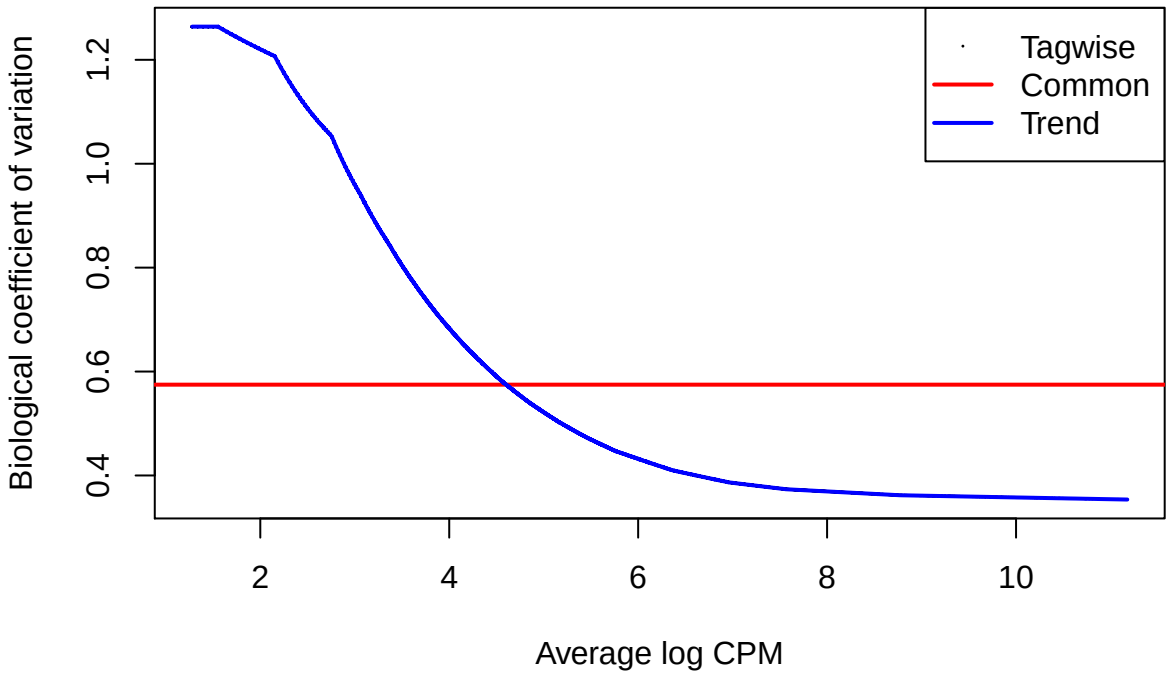

gene9636
Normalized count
100
75
50
25
A
B
group


gene10330
Normalized count
8
6
4
2
A
B
group


gene11082
Normalized count
7.5
5.0
2.5
A
B
group

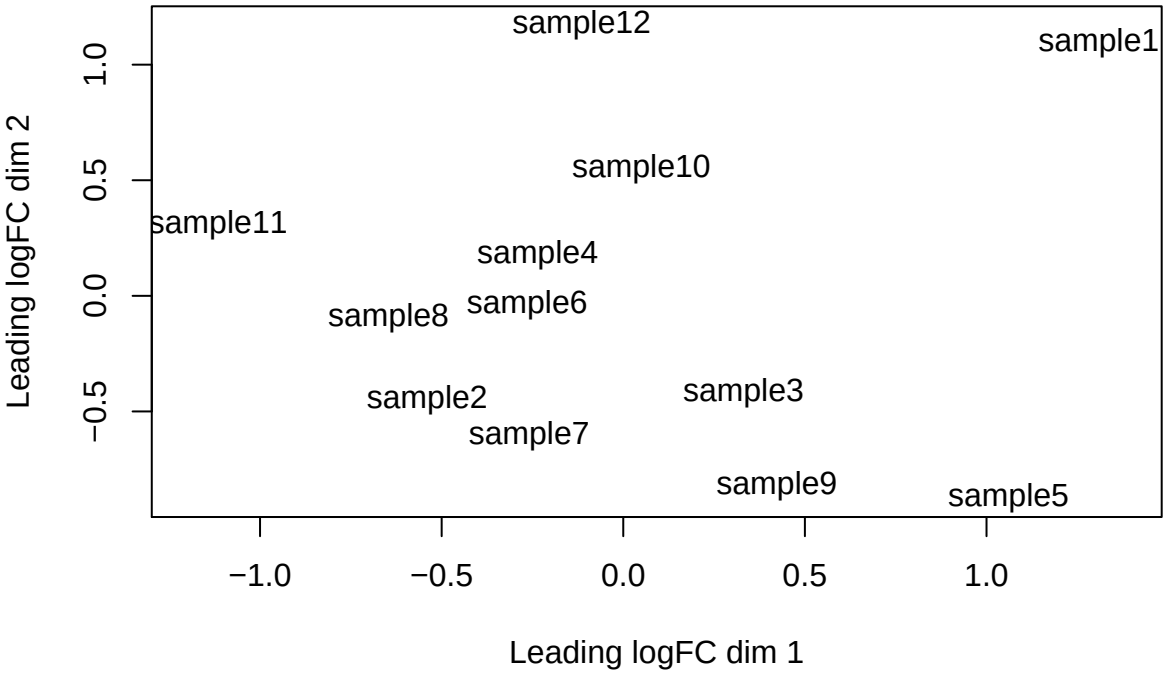
edgeR specific plots

Biological coefficient of variation



This plot shows the feature-wise biological coefficient of variation (BCV) against the feature abundances in log2 counts per million. The plot shows the common, trended and feature-wise BCV estimates. If using *edgeR-robust* only the trend and tagwise are shown. Check the edgeR vignette for further details regarding this plot.

MDS plot of distances



This plot is a multidimensional scaling plot of distances between feature expression profiles. It shows the names of the samples in a two-dimensional scatterplot such that the distances are approximately the log2 fold changes between samples. Check the edgeR vignette for further details regarding this plot.

Reproducibility

The input for this report was generated with edgeR (McCarthy, J., Chen, Yunshun, et al., 2012; Zhou, Lindsay, and Robinson, 2014; Chen, Lun, and Smyth, 2014) and the resulting features were called significantly differentially expressed if their BH adjusted p-values were less than `alpha` = 0.1. This report was generated in path /Users/lcollado/Dropbox/JHSPH/Code/regionReportSupp using the following call to edgeReport():

```
## edgeReport(dge = dge, object = lrt, project = "edgeR PDF report",
##     intgroup = "group", outdir = "edgeR-example", output = "edgeReport",
##     device = "pdf", theme = theme_linedraw(), output_format = "pdf_document")
```

Date the report was generated.

```
## [1] "2016-04-11 22:15:15 EDT"
```

Wallclock time spent generating the report.

```
## Time difference of 1.027 mins
```

`R` session information.

```
## Session info ------------------------------------------------------------------------------------
```

```
##  setting  value
##  version  R version 3.3.0 alpha (2016-03-23 r70368)
##  system   x86_64, darwin13.4.0
##  ui       X11
##  language (EN)
##  collate  en_US.UTF-8
##  tz       America/New_York
##  date     2016-04-11

## Packages ------------------------------------------------------------------------------------------------------

##  package              * version  date       source
##  acepack                1.3-3.3  2014-11-24 CRAN (R 3.3.0)
##  annotate               1.49.1   2016-02-06 Bioconductor
##  AnnotationDbi          1.33.8   2016-04-10 Bioconductor
##  backports              1.0.2    2016-03-18 CRAN (R 3.3.0)
##  bibtex                 0.4.0    2014-12-31 CRAN (R 3.3.0)
##  Biobase              * 2.31.3   2016-01-14 Bioconductor
##  BiocGenerics         * 0.17.4   2016-04-07 Bioconductor
##  BiocParallel           1.5.21   2016-03-23 Bioconductor
##  biomaRt                2.27.2   2016-01-14 Bioconductor
##  Biostrings             2.39.12  2016-02-21 Bioconductor
##  bitops                 1.0-6    2013-08-17 CRAN (R 3.3.0)
##  BSgenome               1.39.4   2016-02-21 Bioconductor
##  bumphunter             1.11.5   2016-03-29 Bioconductor
##  checkmate              1.7.4    2016-04-08 CRAN (R 3.3.0)
##  cluster                2.0.3    2015-07-21 CRAN (R 3.3.0)
##  codetools              0.2-14   2015-07-15 CRAN (R 3.3.0)
##  colorspace             1.2-6    2015-03-11 CRAN (R 3.3.0)
##  DBI                    0.3.1    2014-09-24 CRAN (R 3.3.0)
##  DEFormats            * 0.99.8   2016-03-31 Bioconductor
##  derfinder              1.5.30   2016-03-25 Bioconductor
##  derfinderHelper        1.5.3    2016-03-23 Bioconductor
##  DESeq2               * 1.11.42  2016-04-10 Bioconductor
##  devtools             * 1.10.0   2016-01-23 CRAN (R 3.3.0)
##  digest                 0.6.9    2016-01-08 CRAN (R 3.3.0)
##  doRNG                  1.6      2014-03-07 CRAN (R 3.3.0)
##  DT                   * 0.1      2015-06-09 CRAN (R 3.3.0)
##  edgeR                * 3.13.8   2016-04-08 Bioconductor
##  evaluate               0.8.3    2016-03-05 CRAN (R 3.3.0)
##  foreach                1.4.3    2015-10-13 CRAN (R 3.3.0)
##  foreign                0.8-66   2015-08-19 CRAN (R 3.3.0)
##  formatR                1.3      2016-03-05 CRAN (R 3.3.0)
##  Formula                1.2-1    2015-04-07 CRAN (R 3.3.0)
##  genefilter             1.53.3   2016-03-23 Bioconductor
##  geneplotter            1.49.0   2016-01-14 Bioconductor
##  GenomeInfoDb         * 1.7.6    2016-01-29 Bioconductor
##  GenomicAlignments      1.7.20   2016-02-25 Bioconductor
##  GenomicFeatures        1.23.29  2016-04-05 Bioconductor
##  GenomicFiles           1.7.9    2016-02-22 Bioconductor
##  GenomicRanges        * 1.23.25  2016-03-31 Bioconductor
##  ggplot2              * 2.1.0    2016-03-01 CRAN (R 3.3.0)
##  gridExtra              2.2.1    2016-02-29 CRAN (R 3.3.0)
##  gtable                 0.2.0    2016-02-26 CRAN (R 3.3.0)
```

```
##  highr                  0.5.1    2015-09-18 CRAN (R 3.3.0)
##  Hmisc                  3.17-3   2016-04-03 CRAN (R 3.3.0)
##  htmltools              0.3.5    2016-03-21 CRAN (R 3.3.0)
##  htmlwidgets            0.6      2016-02-25 CRAN (R 3.3.0)
##  httr                   1.1.0    2016-01-28 CRAN (R 3.3.0)
##  IRanges              * 2.5.43   2016-04-10 Bioconductor
##  iterators              1.0.8    2015-10-13 CRAN (R 3.3.0)
##  knitcitations          1.0.7    2015-10-28 CRAN (R 3.3.0)
##  knitr                * 1.12.3   2016-01-22 CRAN (R 3.3.0)
##  knitrBootstrap         1.0.0    2016-03-24 Github (jimhester/knitrBootstrap@cdaa4a9)
##  labeling               0.3      2014-08-23 CRAN (R 3.3.0)
##  lattice                0.20-33  2015-07-14 CRAN (R 3.3.0)
##  latticeExtra           0.6-28   2016-02-09 CRAN (R 3.3.0)
##  limma                * 3.27.14  2016-03-23 Bioconductor
##  locfit                 1.5-9.1  2013-04-20 CRAN (R 3.3.0)
##  lubridate              1.5.6    2016-04-06 CRAN (R 3.3.0)
##  magrittr               1.5      2014-11-22 CRAN (R 3.3.0)
##  markdown               0.7.7    2015-04-22 CRAN (R 3.3.0)
##  Matrix                 1.2-4    2016-03-02 CRAN (R 3.3.0)
##  matrixStats            0.50.1   2015-12-15 CRAN (R 3.3.0)
##  memoise                1.0.0    2016-01-29 CRAN (R 3.3.0)
##  munsell                0.4.3    2016-02-13 CRAN (R 3.3.0)
##  nnet                   7.3-12   2016-02-02 CRAN (R 3.3.0)
##  pheatmap             * 1.0.8    2015-12-11 CRAN (R 3.3.0)
##  pkgmaker               0.22     2014-05-14 CRAN (R 3.3.0)
##  plyr                   1.8.3    2015-06-12 CRAN (R 3.3.0)
##  qvalue                 2.3.2    2016-01-14 Bioconductor
##  R6                     2.1.2    2016-01-26 CRAN (R 3.3.0)
##  RColorBrewer         * 1.1-2    2014-12-07 CRAN (R 3.3.0)
##  Rcpp                   0.12.4   2016-03-26 CRAN (R 3.3.0)
##  RCurl                  1.95-4.8 2016-03-01 CRAN (R 3.3.0)
##  RefManageR             0.10.13  2016-04-04 CRAN (R 3.3.0)
##  regionReport         * 1.5.45   2016-04-12 Bioconductor
##  registry               0.3      2015-07-08 CRAN (R 3.3.0)
##  reshape2               1.4.1    2014-12-06 CRAN (R 3.3.0)
##  RJSONIO                1.3-0    2014-07-28 CRAN (R 3.3.0)
##  rmarkdown              0.9.5    2016-02-22 CRAN (R 3.3.0)
##  rngtools               1.2.4    2014-03-06 CRAN (R 3.3.0)
##  rpart                  4.1-10   2015-06-29 CRAN (R 3.3.0)
##  Rsamtools              1.23.8   2016-04-10 Bioconductor
##  RSQLite                1.0.0    2014-10-25 CRAN (R 3.3.0)
##  rtracklayer            1.31.10  2016-04-07 Bioconductor
##  S4Vectors            * 0.9.46   2016-04-07 Bioconductor
##  scales                 0.4.0    2016-02-26 CRAN (R 3.3.0)
##  stringi                1.0-1    2015-10-22 CRAN (R 3.3.0)
##  stringr                1.0.0    2015-04-30 CRAN (R 3.3.0)
##  SummarizedExperiment * 1.1.23   2016-04-06 Bioconductor
##  survival               2.38-3   2015-07-02 CRAN (R 3.3.0)
##  VariantAnnotation      1.17.23  2016-04-07 Bioconductor
##  XML                    3.98-1.4 2016-03-01 CRAN (R 3.3.0)
##  xtable                 1.8-2    2016-02-05 CRAN (R 3.3.0)
##  XVector                0.11.8   2016-04-06 Bioconductor
##  yaml                   2.1.13   2014-06-12 CRAN (R 3.3.0)
##  zlibbioc               1.17.1   2016-03-19 Bioconductor
```

Pandoc version used: 1.17.0.3.

Bibliography

This report was created with `regionReport` (Collado-Torres, Jaffe, and Leek, 2015) using `rmarkdown` (Allaire, Cheng, Xie, McPherson, et al., 2016) while `knitr` (Xie, 2014) and `DT` (Xie, 2015) were running behind the scenes. `pheatmap` (Kolde, 2015) was used to create the sample distances heatmap. Several plots were made with `ggplot2` (Wickham, 2009).

Citations made with `knitcitations` (Boettiger, 2015). The BibTeX file can be found here.

[1] J. Allaire, J. Cheng, Y. Xie, J. McPherson, et al. rmarkdown: Dynamic Documents for R. R package version 0.9.5. 2016. URL: https://CRAN.R-project.org/package=rmarkdown.

[2] C. Boettiger. knitcitations: Citations for 'Knitr' Markdown Files. R package version 1.0.7. 2015. URL: https://CRAN.R-project.org/package=knitcitations.

[3] Y. Chen, A. T. L. Lun and G. K. Smyth. "Differential expression analysis of complex RNA-seq experiments using edgeR". In: Statistical Analysis of Next Generation Sequencing Data. Ed. by S. Datta and D. Nettleton. New York: Springer, 2014, pp. 51-74.

[4] L. Collado-Torres, A. E. Jaffe and J. T. Leek. regionReport: Generate HTML reports for exploring a set of regions. https://github.com/leekgroup/regionReport - R package version 1.5.45. 2015. URL: http://www.bioconductor.org/packages/regionReport.

[5] R. Kolde. pheatmap: Pretty Heatmaps. R package version 1.0.8. 2015. URL: https://CRAN.R-project.org/package=pheatmap.

[6] M. I. Love, W. Huber and S. Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: Genome Biology 15 (12 2014), p. 550. DOI: 10.1186/s13059-014-0550-8.

[7] McCarthy, D. J., Chen, Yunshun, et al. "Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation". In: Nucleic Acids Research 40.10 (2012), pp. -9.

[8] H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2009. ISBN: 978-0-387-98140-6. URL: http://ggplot2.org.

[9] Y. Xie. DT: A Wrapper of the JavaScript Library 'DataTables'. R package version 0.1. 2015. URL: https://CRAN.R-project.org/package=DT.

[10] Y. Xie. "knitr: A Comprehensive Tool for Reproducible Research in R". In: Implementing Reproducible Computational Research. Ed. by V. Stodden, F. Leisch and R. D. Peng. ISBN 978-1466561595. Chapman and Hall/CRC, 2014. URL: http://www.crcpress.com/product/isbn/9781466561595.

[11] X. Zhou, H. Lindsay and M. D. Robinson. "Robustly detecting differential expression in RNA sequencing data using observation weights". In: Nucleic Acids Research 42 (2014), p. e91.