Nanopore Sequenced Transcriptome Analysis Resource

NanoSTAR DGE Report

Report created: 18 September 2021

This NanoSTAR DGE report was generated automatically based on user input, reflected below. The principles of reproducible research have been implemented, and a full list of software used (including version numbers) and user defined parameters is included at the end of this document.

Study design

Table 1: Groups, sequence files, and sample identifiers used for analysis.

Group	Files	Identifiers
Group A	barcode01.fastq, barcode02.fastq, barcode03.fastq, barcode04.fastq, barcode08.fastq, barcode11.fastq, 2barcode05.fastq	A1, A2, A3, A4, A5, A6, A7
Group B	barcode05.fastq, barcode06.fastq, barcode10.fastq, 2barcode03.fastq, 2barcode06.fastq, 2barcode07.fastq, 2barcode08.fastq, 2barcode10.fastq, 2barcode11.fastq	B1, B2, B3, B4, B5, B6, B7, B8, B9
Group C	barcode07.fastq, barcode09.fastq, barcode12.fastq, 2barcode02.fastq, 2barcode01.fastq, 2barcode04.fastq, 2barcode09.fastq, 2barcode12.fastq	C1, C2, C3, C4, C5, C6, C7, C8

Differential gene expression analysis

Statistical analysis performed using **`edgeR`** (Robinson, McCarthy, and Smyth (2010), McCarthy et al. (2012)) on gene counts from **`Salmon`** (Patro et al. (2017)), filtered by **`DRIMSeq`** (Nowicka and Robinson (2016)), using the `TMM` method for normalisation, and correcting for false discovery rate (*FDR*) using the method of Benjamini & Hochberg (Benjamini and Hochberg (1995)).

Using a Log_2 fold change threshold of **±1** and *p*-value threshold of **0.05**: **3771** genes were differentially expressed between Control and at least one experimental group. **826** genes had increased expression in at least one experimental group, and **3254** had reduced expression. **309** genes had increased expression and decreased expression in different experimental groups.

Using a Log_2 fold change threshold of **±1** and FDR threshold of **0.1**: **868** genes were differentially expressed between Control and at least one experimental group. **221** genes had increased expression in at least one experimental group, and **732** had reduced expression. **85** genes had increased expression and decreased expression in different experimental groups.

Table 2: Table showing the top 50 differentially expressed genes, ranked by adjusted p-value, from the edgeR analysis.

	logFC Group B	logFC Group C	logCPM	F	p-Value	FDR
CORO6	-5.59	-5.59	6.22	20.23789	1.626231e-09	3.554454e-05
BEND7	-5.46	-5.18	6.60	18.61305	8.256272e-09	9.022866e-05
ADAM32	-4.95	-4.87	6.43	16.20495	9.173308e-08	5.104771e-04
GPR161	-4.77	-4.77	5.89	16.18672	9.342126e-08	5.104771e-04
ST18	-4.75	-4.75	5.88	15.83467	1.328398e-07	5.806958e-04
RIMS2	2.53	-4.31	7.33	15.56628	1.737315e-07	6.328750e-04
SLC6A11	-4.80	-4.54	6.60	15.27999	2.313153e-07	7.222656e-04
KIAA0556	-4.16	-4.97	5.97	14.76799	3.859664e-07	1.054508e-03
MAN2A2	-4.53	-4.53	5.81	14.61826	4.483018e-07	1.088726e-03
PLEKHA5	-0.17	-6.86	7.77	14.40377	5.555429e-07	1.214250e-03
RYR2	-0.55	-6.97	7.70	14.23027	6.607926e-07	1.312995e-03
LZTR1	-4.40	-4.40	5.77	14.07554	7.713584e-07	1.368500e-03
RBM23	1.42	5.13	6.09	14.02180	8.139500e-07	1.368500e-03
CREB3L3	0.00	4.29	5.76	13.87224	9.452426e-07	1.385590e-03
TCEA2	-5.93	-2.30	6.53	13.82299	9.929625e-07	1.385590e-03
MED12L	-3.12	-5.78	6.39	13.80173	1.014294e-06	1.385590e-03
LOC101114261	-0.28	-6.66	7.56	13.41218	1.497394e-06	1.925209e-03
TPRG1	-0.79	-6.76	7.45	13.24024	1.778296e-06	2.147927e-03
RANBP3L	-4.79	-3.56	5.93	13.19147	1.867164e-06	2.147927e-03
CAPG	1.55	4.98	6.05	13.12841	1.988689e-06	2.173339e-03
DDX51	-4.13	-4.13	5.70	12.70579	3.034564e-06	3.158404e-03
SEC31B	0.00	4.01	5.68	12.54202	3.574515e-06	3.537743e-03
ZNF583	-4.09	-4.09	5.69	12.46137	3.874755e-06	3.537743e-03
CPEB2	-4.45	-3.73	6.44	12.45883	3.884606e-06	3.537743e-03
AK4	-4.07	-4.07	5.68	12.32237	4.452533e-06	3.755366e-03
PCLO	4.60	1.08	5.95	12.31908	4.467197e-06	3.755366e-03
PHF21A	-2.94	-4.47	8.06	12.21071	4.978499e-06	3.893875e-03
PPHLN1	-4.07	1.54	6.49	12.20875	4.988265e-06	3.893875e-03
SLC4A11	4.12	0.11	6.15	12.10690	5.523060e-06	4.162673e-03
UNC13A	-3.63	-4.44	5.79	12.02752	5.979300e-06	4.356319e-03
MAD2L1BP	-3.27	-4.85	5.96	11.95586	6.423487e-06	4.490778e-03
KLHL12	-2.74	-5.33	6.17	11.90632	6.749714e-06	4.490778e-03
CHL1	1.77	-4.20	6.79	11.90181	6.780239e-06	4.490778e-03
SUSD6	-2.18	-5.81	6.49	11.81747	7.376821e-06	4.641078e-03
INPP4A	2.94	-2.06	6.15	11.81004	7.431840e-06	4.641078e-03
PKHD1	1.80	-4.11	6.74	11.76618	7.765018e-06	4.642975e-03
STX16	-5.21	-2.18	7.04	11.75406	7.859728e-06	4.642975e-03
LOC105609280	-1.61	3.05	6.28	11.66315	8.607703e-06	4.951015e-03
PGPEP1	-3.91	-3.91	5.64	11.56239	9.520123e-06	5.168174e-03
LOC105603128	-0.60	-5.39	8.09	11.55043	9.634660e-06	5.168174e-03
LOC105605231	-3.99	-3.99	5.67	11.54423	9.694613e-06	5.168174e-03
ARFIP1	-2.15	-5.73	6.45	11.47761	1.036242e-05	5.242830e-03
OTUD3	-1.38	-6.13	6.83	11.47579	1.038129e-05	5.242830e-03
CHRDL2	-3.50	-4.46	5.81	11.45927	1.055426e-05	5.242830e-03
PRPF3	-5.19	-1.93	6.17	11.42691	1.090131e-05	5.294887e-03
LOC101121441	-4.33	-3.28	5.77	11.34172	1.187068e-05	5.640380e-03
ITGA11	3.85	0.00	5.69	11.29659	1.241864e-05	5.775198e-03
MUC19	0.22	-5.37	7.63	11.23352	1.322706e-05	6.022997e-03
ARMC3	-3.19	-4.57	6.32	11.15468	1.431211e-05	6.384076e-03
NADSYN1	-0.83	-6.06	6.92	11.10485	1.504320e-05	6.544450e-03

logFC = Log_2 fold change between experimental conditions. logCPM = Log_2 counts per million.

Full DEG analysis results and a complete DEG list have been saved here:

```
Analysis/Results/DiffExpr_Results.xlsx
Analysis/Results/GeneList.xlsx
```

Gene Expression Heatmap

Shown below is a heat map of the 50 genes showing the most significant differential expression, plotted against z-score. Genes are ordered from top to bottom by increasing p-value. Plot has been saved as:

`Analysis/Results/Images/GeneExprHeatmap.pdf`

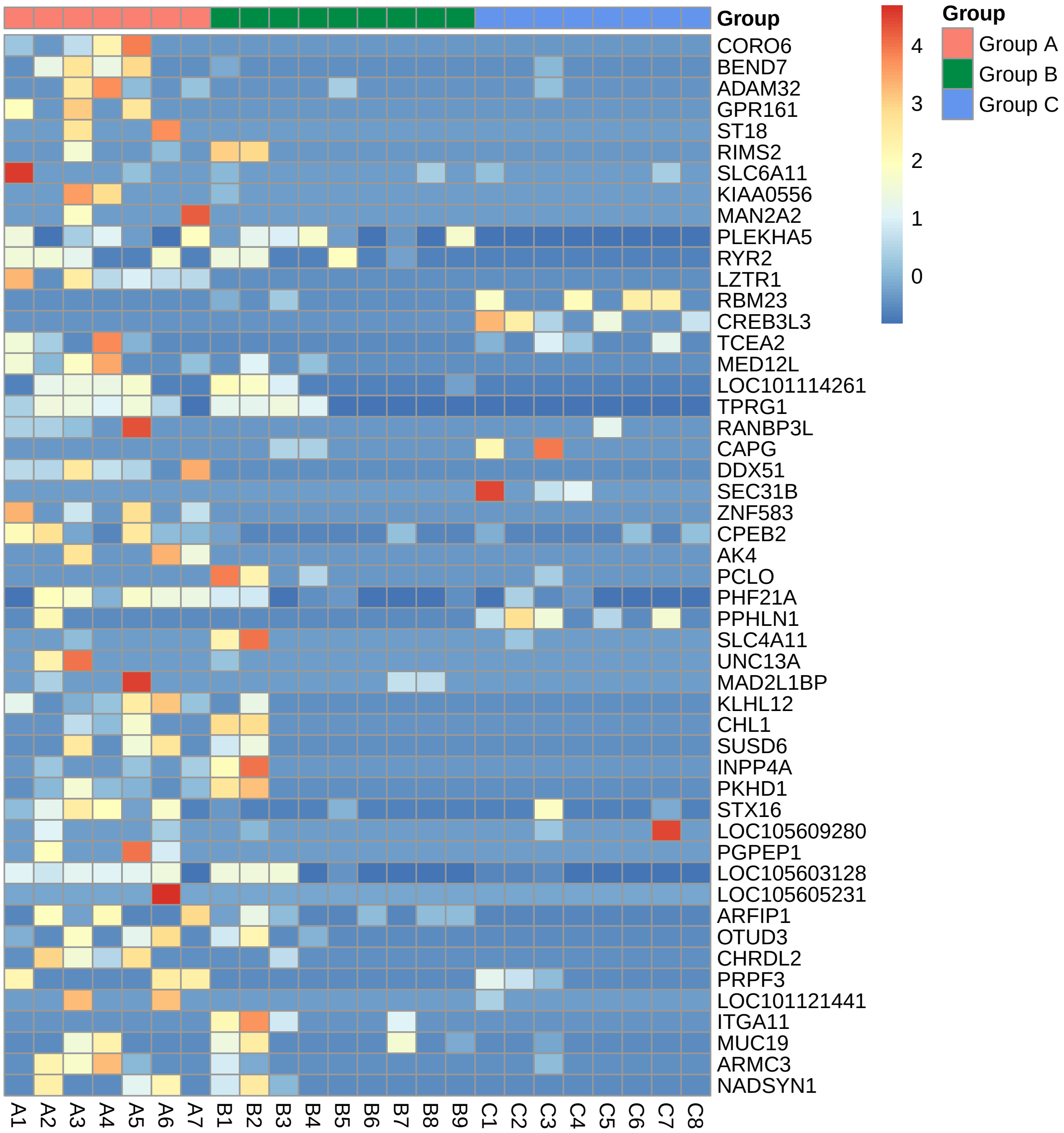

Figure 1: 50 most significant DEGs

Principal Component Analysis

PCA analysis performed by **`pcaMethods`** showing the distribution of sample data for the first two principal components. The first principal component is shown on the x-axis; the second on the y. The total amount of variation explained is shown on the axis legends. The plot has been saved to:

`Analysis/Results/Images/PCAplot.pdf`

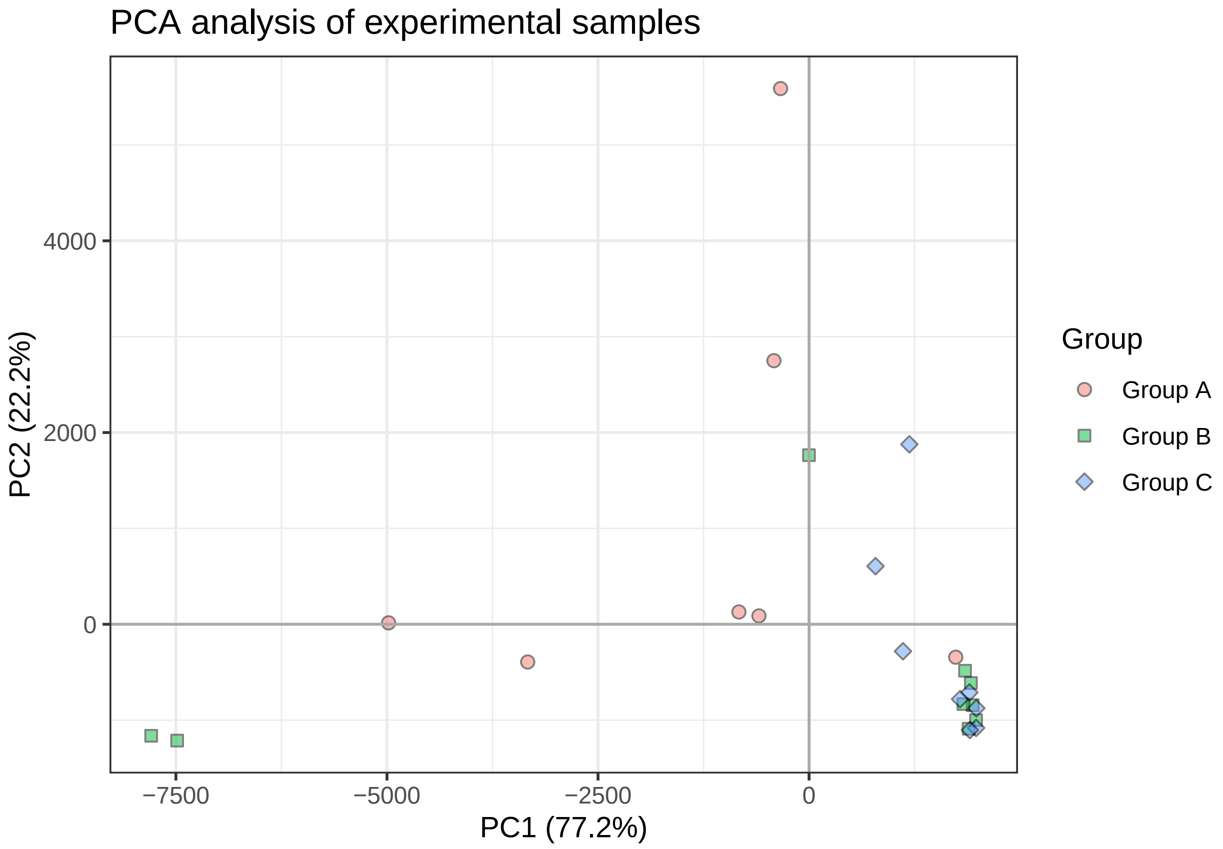

Figure 2: PCA plot.

Hierarchial clustering

Shown below is a heatmap of the 1000 most variably expressed genes, clustered for similarity of gene expression with regards to inter-gene and inter-sample expression. Plot has been saved as:

`Analysis/Results/Images/heatmap_cluster.pdf`

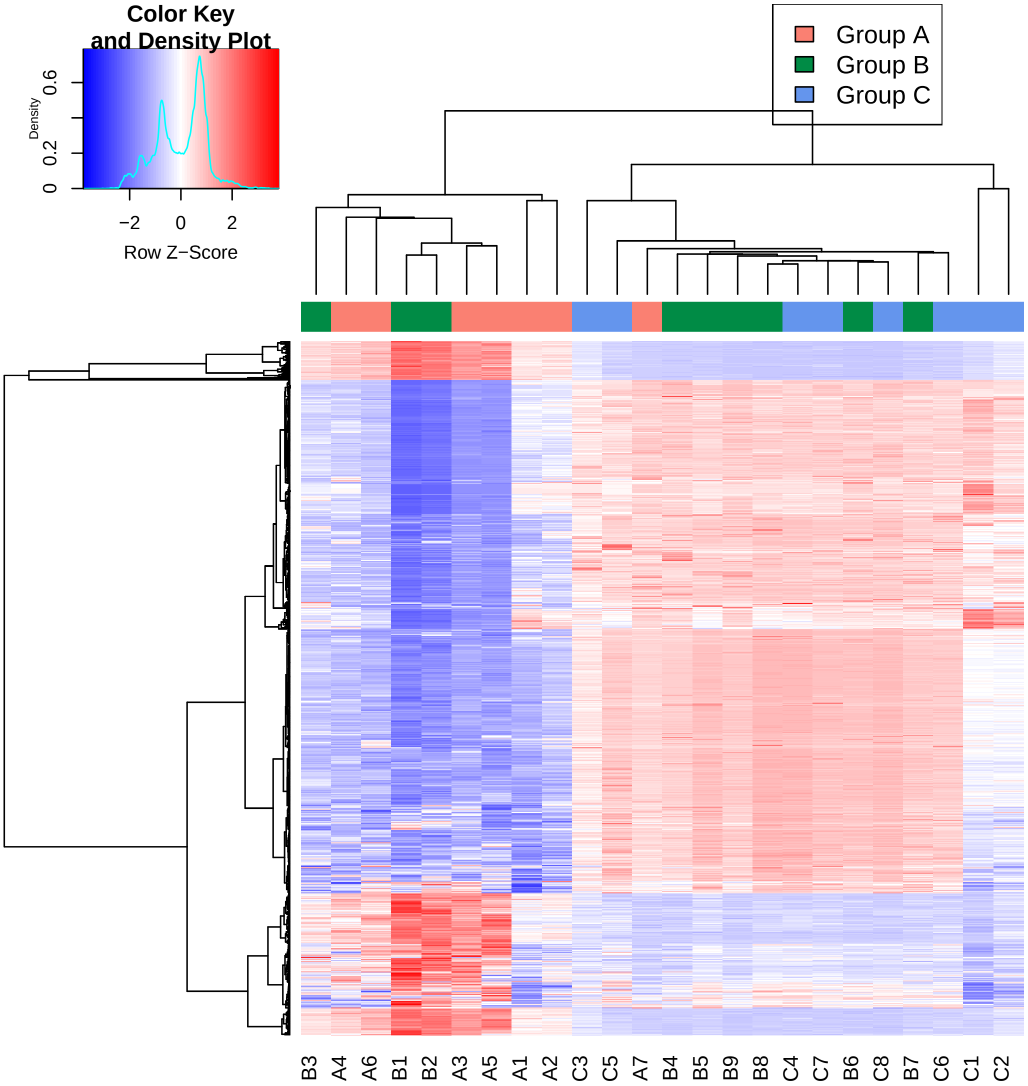

Figure 3: Heatmap with gene and sample dendrograms.

Heatmap below shows the similarity of 250 genes with the most significant differential expression arranged based on similarity into 6 clusters. Plot has been saved as:

`Analysis/Results/Images/heatmap_DEG_clusters.pdf`

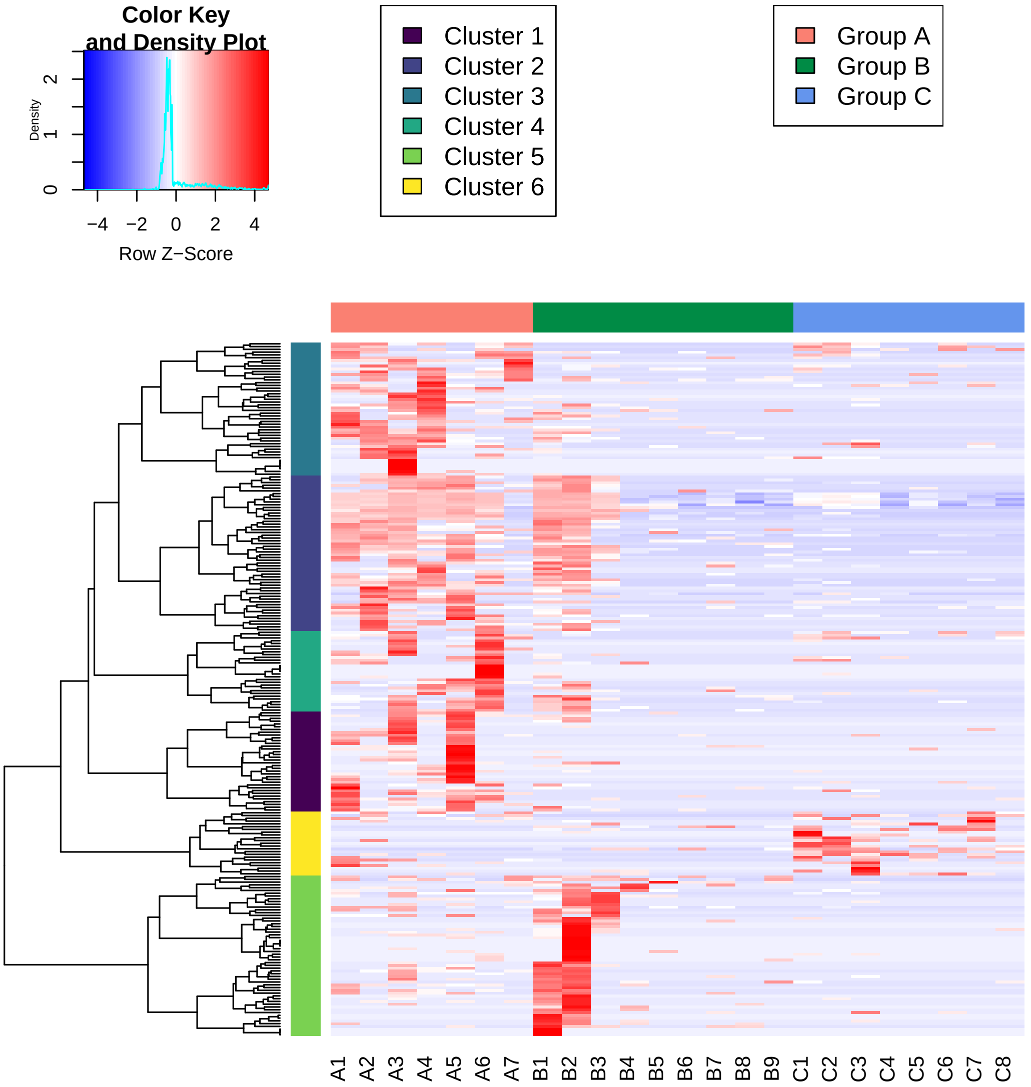

Figure 4: Heatmap with gene dendrogram into DEG clusters.

Expression profiles for genes within the 6 gene clusters are shown below. Plot has been saved as: `Analysis/Results/Images/Clusters.pdf`

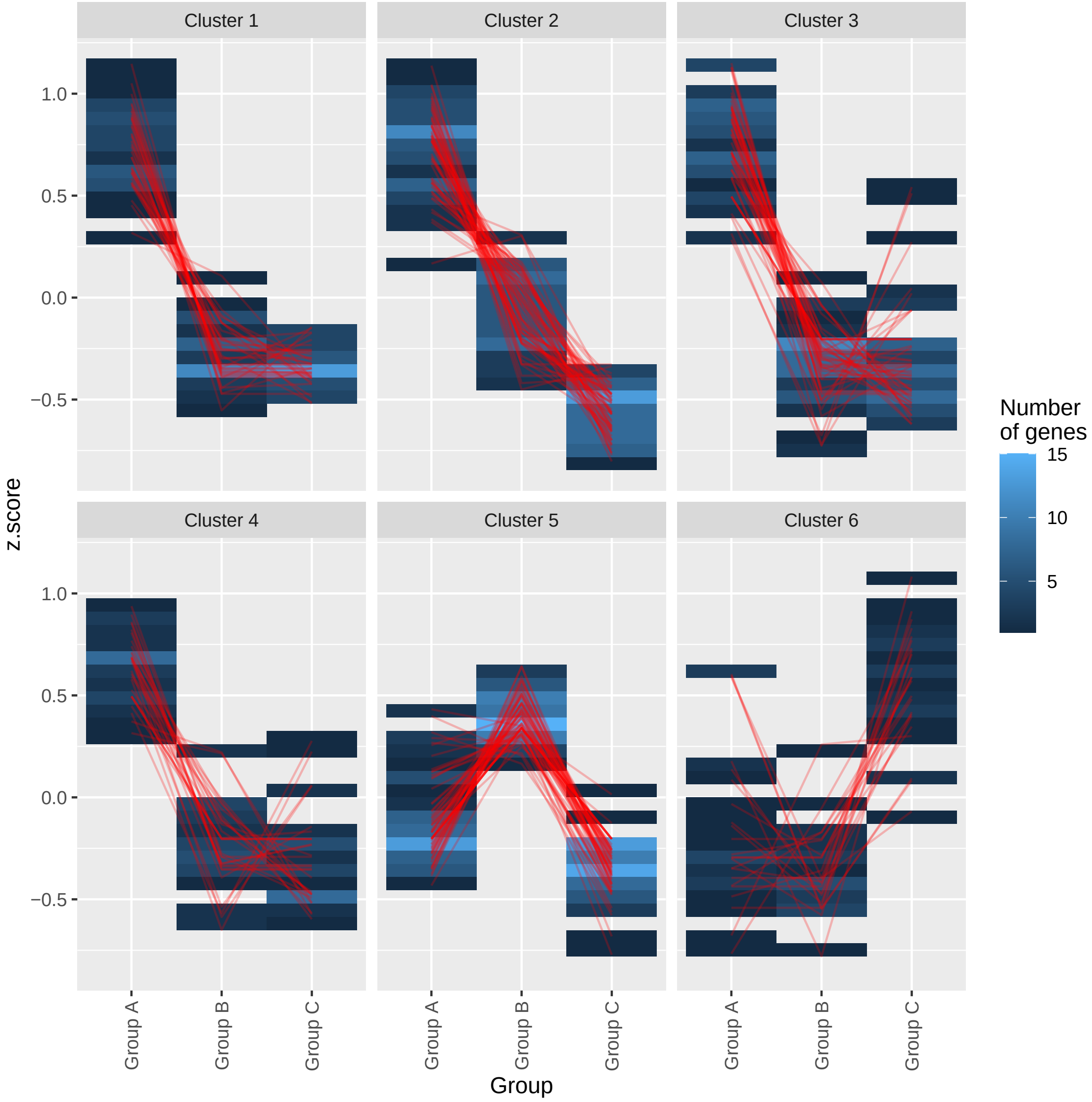

Figure 5: Plot showing the expression profiles of clustered genes.

Output

A number of files have been output.

Complete transcript mapping counts and parental gene assignments:

`Analysis/Results/ExpressedGenesTranscripts(Raw).xlsx`

Filtered gene counts and gene isoform counts:

`Analysis/Results/GeneCounts(Filtered).xlsx`
`Analysis/Results/FeatureCounts(Filtered).xlsx`

Log_2 normalised counts:

`Analysis/Results/LogNormCounts.xlsx`

The results of the differential expression analysis:

`Analysis/Results/DiffExpr_Results.xlsx`

A list of deferentially expressed genes:

`Analysis/Results/GeneList.xlsx`

Gene lists from clustering results:

`Analysis/Results/GeneClusteringResults.xlsx`

The session data produced in the production of this report, which can be used for further analysis of the dataset:

`Analysis/Results/NanoSTAR_DGE_Report.Rdata`

Reproducible research

This analysis used publicly available **`Linux`** software, which are listed below with their version numbers.

```
# packages in environment at /home/chris/miniconda3:
#
# Name                    Version                   Build  Channel
bioconductor-annotationdbi 1.54.0            r41hdfd78af_0    bioconda
bioconductor-deseq2       1.32.0            r41h399db7b_0    bioconda
bioconductor-dexseq       1.38.0            r41hdfd78af_0    bioconda
bioconductor-drimseq      1.20.0            r41hdfd78af_0    bioconda
bioconductor-edger        3.34.0            r41h399db7b_0    bioconda
bioconductor-genomicfeatures 1.44.0            r41hdfd78af_0    bioconda
bioconductor-pcamethods   1.84.0            r41h399db7b_0    bioconda
bioconductor-shortread    1.50.0            r41h399db7b_0    bioconda
bioconductor-stager       1.14.0            r41hdfd78af_0    bioconda
filtlong                  0.2.1                h9a82719_0    bioconda
porechop                  0.2.4           py39h7cff6ad_2    bioconda
r-devtools                2.4.2             r41hc72bb7e_0    conda-forge
r-digest                  0.6.27            r41h03ef668_0    conda-forge
r-dplyr                   1.0.7             r41h03ef668_0    conda-forge
r-ggplot2                 3.3.5             r41hc72bb7e_0    conda-forge
r-gplots                  3.1.1             r41hc72bb7e_0    conda-forge
```

```
r-gridextra               2.3             r41hc72bb7e_1003    conda-forge
r-kableextra              1.3.4             r41hc72bb7e_0    conda-forge
r-pheatmap                1.0.12            r41hc72bb7e_2    conda-forge
r-plotrix                 3.8_2             r41hc72bb7e_0    conda-forge
r-reshape2                1.4.4             r41h03ef668_1    conda-forge
r-rstudioapi              0.13              r41hc72bb7e_0    conda-forge
r-tidyr                   1.1.3             r41h03ef668_0    conda-forge
r-tidyverse               1.3.1             r41hc72bb7e_0    conda-forge
r-viridis                 0.6.1             r41hc72bb7e_1    conda-forge
r-viridislite             0.4.0             r41hc72bb7e_0    conda-forge
r-writexl                 1.4.0             r41hcfec24a_0    conda-forge
r-yaml                    2.2.1             r41hcfec24a_1    conda-forge
Parsing /var/lib/dpkg/status... completed.
apt-show-versions:all/focal 0.22.11 uptodate
minimap2:amd64/focal 2.17+dfsg-2 uptodate
pandoc:amd64/focal 2.5-3build2 uptodate
salmon:amd64/focal 0.12.0+ds1-1 uptodate
samtools:amd64/focal 1.10-3 uptodate
texlive-fonts-recommended:all/focal 2019.20200218-1 uptodate
texlive-latex-base:all/focal 2019.20200218-1 uptodate
texlive-latex-extra:all/focal 2019.202000218-1 uptodate
texlive-latex-recommended:all/focal 2019.20200218-1 uptodate
```

This report has been created for reproducibility, using **Rmarkdown**, publicly available **R** packages, and the **LaTeX** document typesetting software. Packages and their version numbers are listed below.

```
R version 4.1.1 (2021-08-10)
Platform: x86_64-conda-linux-gnu (64-bit)
Running under: Ubuntu 20.04.3 LTS

Matrix products: default
BLAS/LAPACK: /home/chris/miniconda3/lib/libopenblasp-r0.3.17.so

attached base packages:
 [1] tools     stats4    parallel  grid      stats     graphics  grDevices utils     datasets
[10] methods   base

loaded via a namespace (and not attached):
  [1] readxl_1.3.1           backports_1.2.1        BiocFileCache_2.0.0    systemfonts_1.0.2
  [5] plyr_1.8.6             splines_4.1.1          htmltools_0.5.2        fansi_0.5.0
  [9] magrittr_2.0.1         memoise_2.0.0          tzdb_0.1.2             remotes_2.4.0
 [13] annotate_1.70.0        modelr_0.1.8           svglite_2.0.0          prettyunits_1.1.1
 [17] jpeg_0.1-9             colorspace_2.0-2       blob_1.2.2             rvest_1.0.1
 [21] rappdirs_0.3.3         haven_2.4.3            xfun_0.25              callr_3.7.0
 [25] crayon_1.4.1           RCurl_1.98-1.4         jsonlite_1.7.2         genefilter_1.74.0
 [29] survival_3.2-13        glue_1.4.2             gtable_0.3.0           zlibbioc_1.38.0
 [33] webshot_0.5.2          DelayedArray_0.18.0    pkgbuild_1.2.0         scales_1.1.1
 [37] DBI_1.1.1              Rcpp_1.0.7             xtable_1.8-4           progress_1.2.2
 [41] bit_4.0.4              httr_1.4.2             ellipsis_0.3.2         farver_2.1.0
 [45] pkgconfig_2.0.3        XML_3.99-0.7           dbplyr_2.1.1           locfit_1.5-9.4
 [49] utf8_1.2.2             labeling_0.4.2         tidyselect_1.1.1       rlang_0.4.11
 [53] munsell_0.5.0          cellranger_1.1.0       cachem_1.0.6           cli_3.0.1
 [57] generics_0.1.0         RSQLite_2.2.5          broom_0.7.9            evaluate_0.14
 [61] fastmap_1.1.0          processx_3.5.2         knitr_1.34             bit64_4.0.5
 [65] fs_1.5.0               caTools_1.18.2         KEGGREST_1.32.0        xml2_1.3.2
 [69] biomaRt_2.48.0         compiler_4.1.1         rstudioapi_0.13        filelock_1.0.2
 [73] curl_4.3.2             png_0.1-7              testthat_3.0.4         reprex_2.0.1
 [77] statmod_1.4.36         geneplotter_1.70.0     stringi_1.7.4          ps_1.6.0
 [81] desc_1.3.0             lattice_0.20-44        Matrix_1.3-4           vctrs_0.3.8
 [85] pillar_1.6.2           lifecycle_1.0.0        bitops_1.0-7           rtracklayer_1.52.0
 [89] latticeExtra_0.6-29    R6_2.5.1               BiocIO_1.2.0           hwriter_1.3.2
 [93] KernSmooth_2.23-20     sessioninfo_1.1.1      gtools_3.9.2           assertthat_0.2.1
 [97] pkgload_1.2.2          rprojroot_2.0.2        rjson_0.2.20           withr_2.4.2
[101] GenomeInfoDbData_1.2.6 hms_1.1.0              rmarkdown_2.10         lubridate_1.7.10
[105] restfulr_0.0.13
```

Defined parameters for this analysis were:

Genes expressed in a minimum of 3 samples.

Transcripts expressed in a minimum of 1 samples.

Minimum gene counts of 10.

Minimum transcript counts of 3

Log_2 fold change threshold of ± 1.

Adjusted p-value threshold of 0.05.

False discovery rate threshold of 0.1.

References and citations

Benjamini, Yoav, and Yosef Hochberg. 1995. "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. http://www.jstor.org/stable/2346101.

McCarthy, Davis J., Chen, Yunshun, Smyth, and Gordon K. 2012. "Differential Expression Analysis of Multifactor RNA-Seq Experiments with Respect to Biological Variation." *Nucleic Acids Research* 40 (10): 4288–97.

Nowicka, Malgorzata, and Mark D. Robinson. 2016. "DRIMSeq: A Dirichlet-Multinomial Framework for Multivariate Count Outcomes in Genomics [Version 2; Referees: 2 Approved]." *F1000Research* 5 (1356). https://doi.org/10.12688/f1000research.8900.2.

Patro, Robert, Geet Duggal, Michael I Love, Rafael A Irizarry, and Carl Kingsford. 2017. "Salmon Provides Fast and Bias-Aware Quantification of Transcript Expression." *Nature Methods* 14 (March). https://doi.org/10.1038/nmeth.4197.

Robinson, Mark D, Davis J McCarthy, and Gordon K Smyth. 2010. "edgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data." *Bioinformatics* 26 (1): 139–40.