Next-generation sequencing - analysis with DESeq2

Comparison of gene expression in DSCs compared to melanocytes - differential gene expression analysis

Marc Bender

2024-04-08

Table of Contents

# Resources and introduction

This document includes the differential expression analysis for the comparison of RNASeq data from dermal stem cells (DSCs) and Melanocytes. The exploratory data analysis and functional analysis (pathway analysis and GO terms) are included in separate files. Statistical methods are described in a separate .docx file.

For additional info on analysis workflows check: DESEQ workflow: <https://www.bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html> <https://bioconductor.org/packages/release/workflows/vignettes/rnaseqGene/inst/doc/rnaseqGene.html>  
<https://www.bigbioinformatics.org/r-and-rnaseq-analysis> <https://github.com/bigbioinformatics/r-programming-and-rnaseq-workshop> ClusterProfiler: <https://pubmed.ncbi.nlm.nih.gov/34557778/>

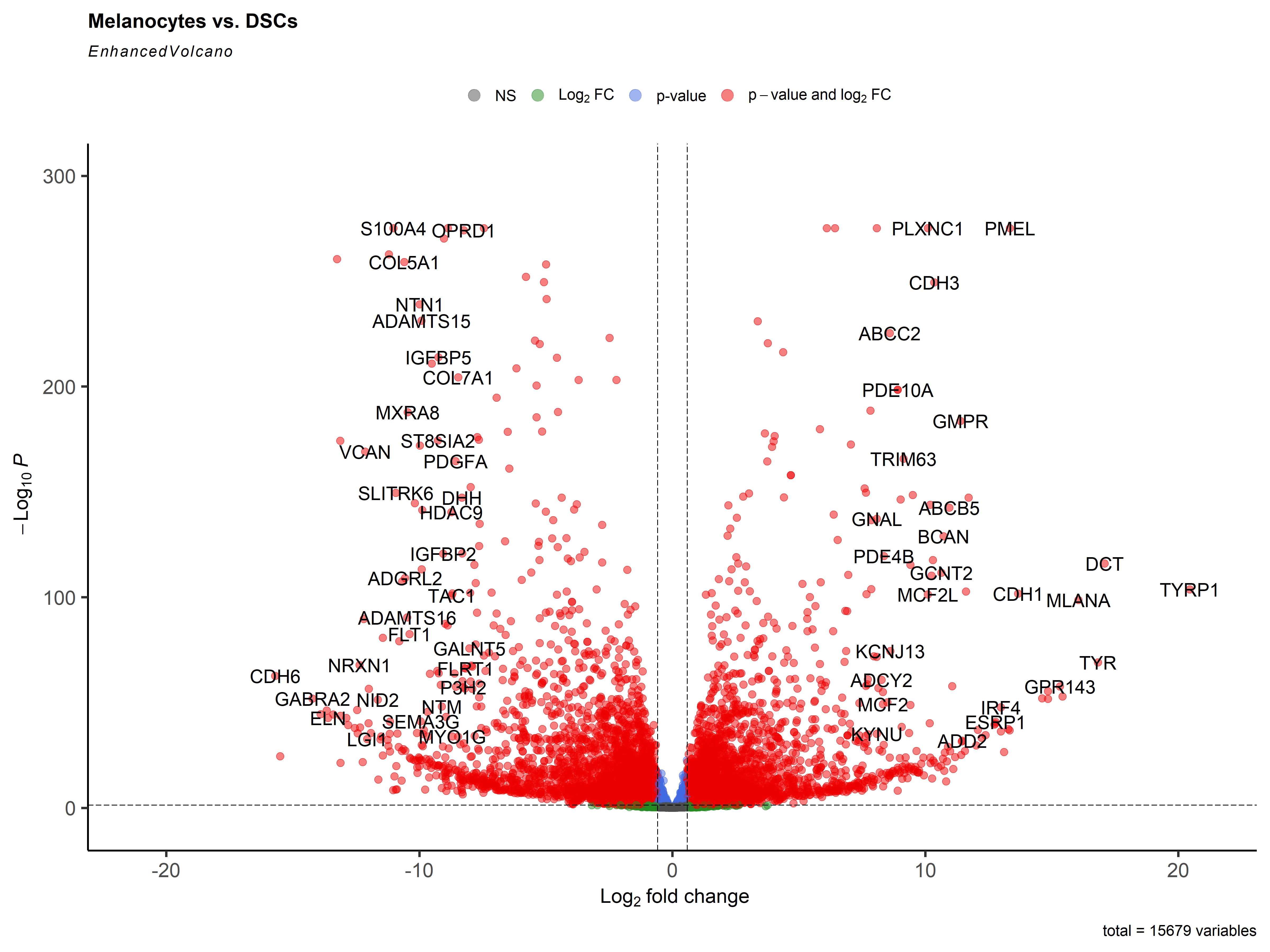
# Differential expression analysis

## Overview

##   
## out of 15536 with nonzero total read count  
## adjusted p-value < 0.05  
## LFC > 0.58 (up) : 2068, 13%  
## LFC < -0.58 (down) : 2265, 15%  
## outliers [1] : 36, 0.23%  
## low counts [2] : 0, 0%  
## (mean count < 6)  
## [1] see 'cooksCutoff' argument of ?results  
## [2] see 'independentFiltering' argument of ?results

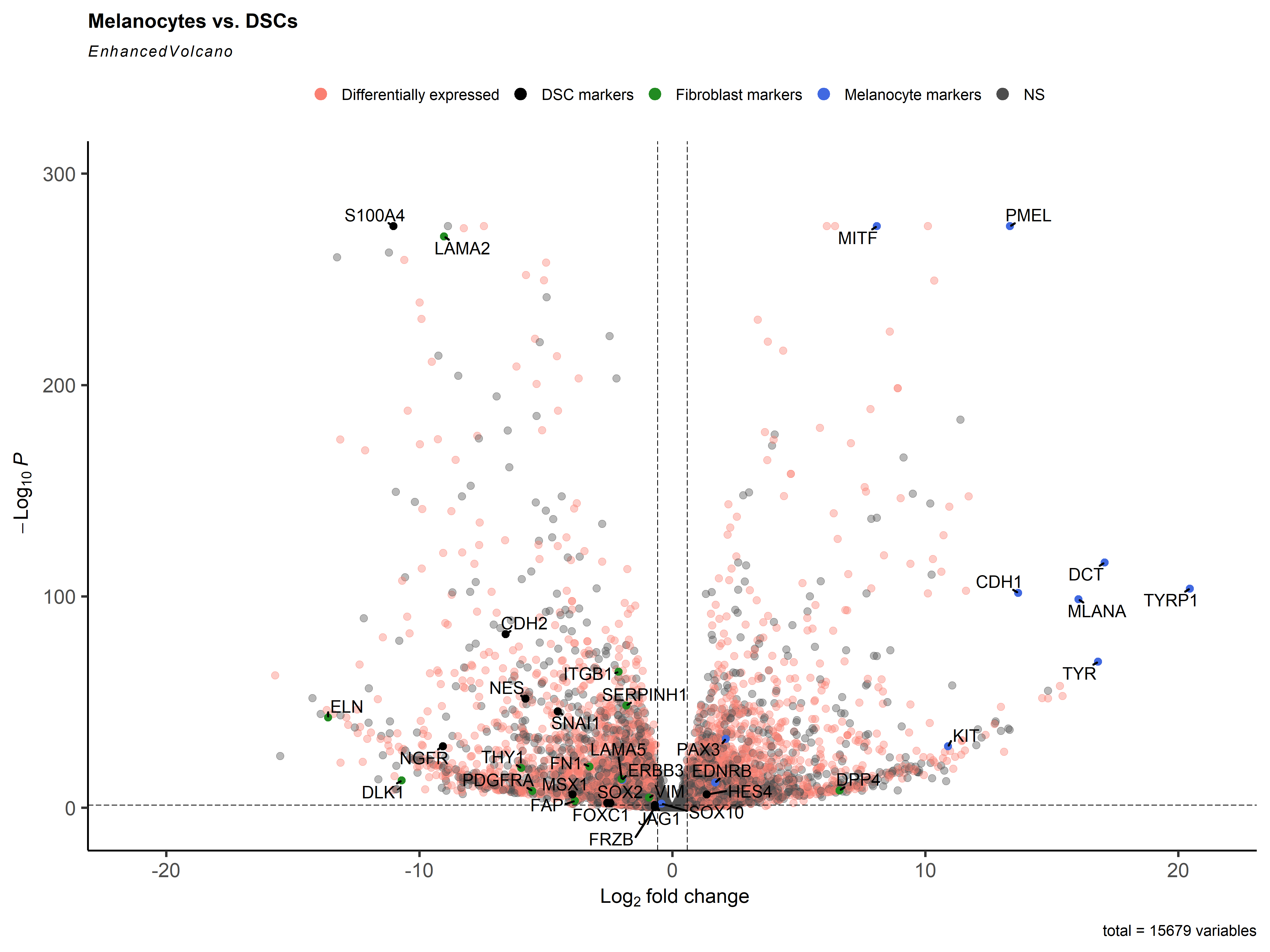
## Volcano Plot (most differentially expressed genes labelled)

This plot shows the most significantly changed genes (<= 1e-30, logFC >= |8|).



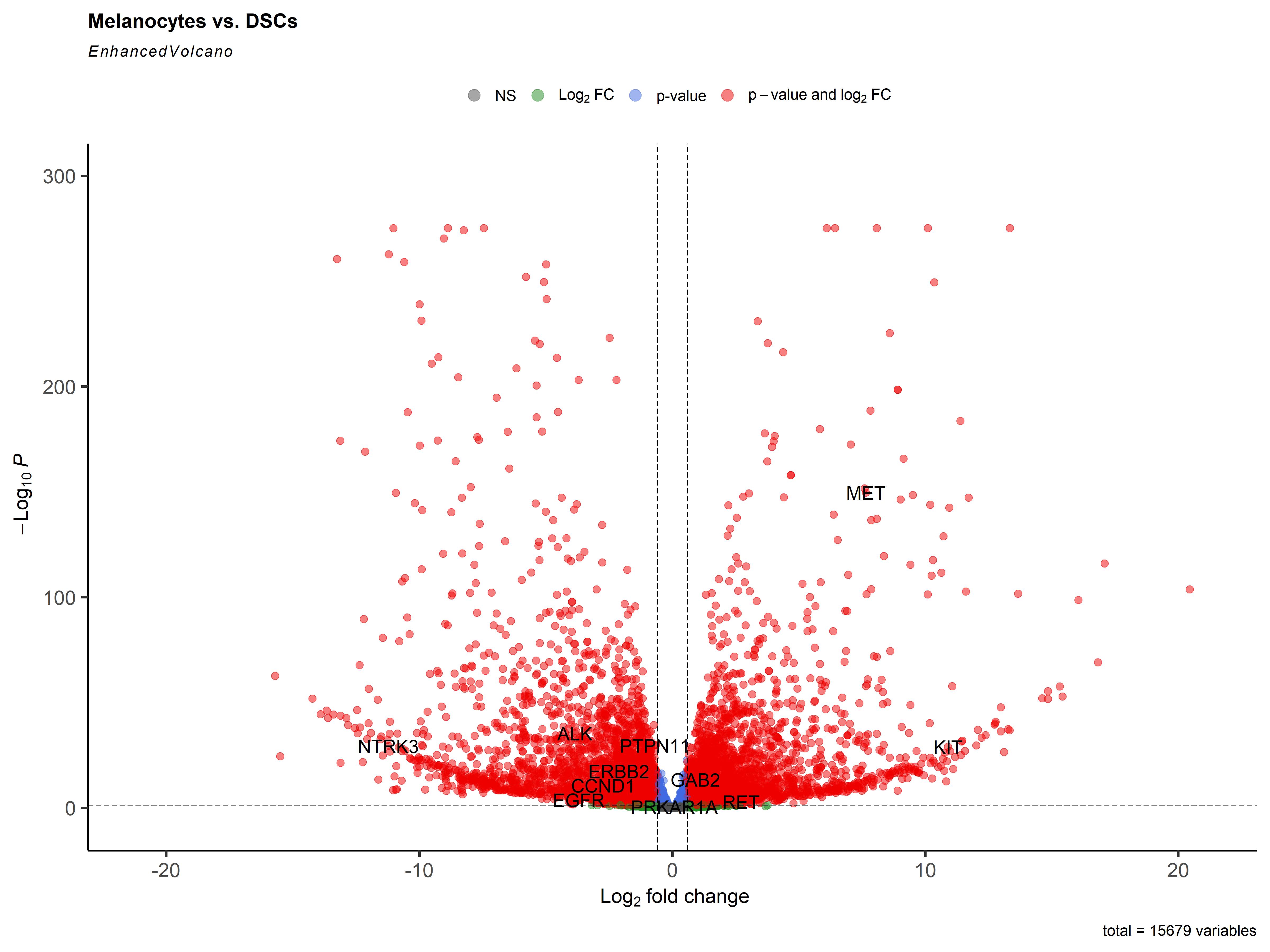
## Volcano Plot (skin cell marker genes labelled)

This plot shows 24 selected genes from a publication assessing differences between Melanocytes (12 genes) and DSCs (12 genes) (Zabierowski et al. 2012) as well as 12 genes characteristic for Fibroblasts.



## Volcano Plot (melanoma marker genes labelled)

This plot shows genes frequently mutated in melanoma (taken from Cherepakhin et. al (2022)).

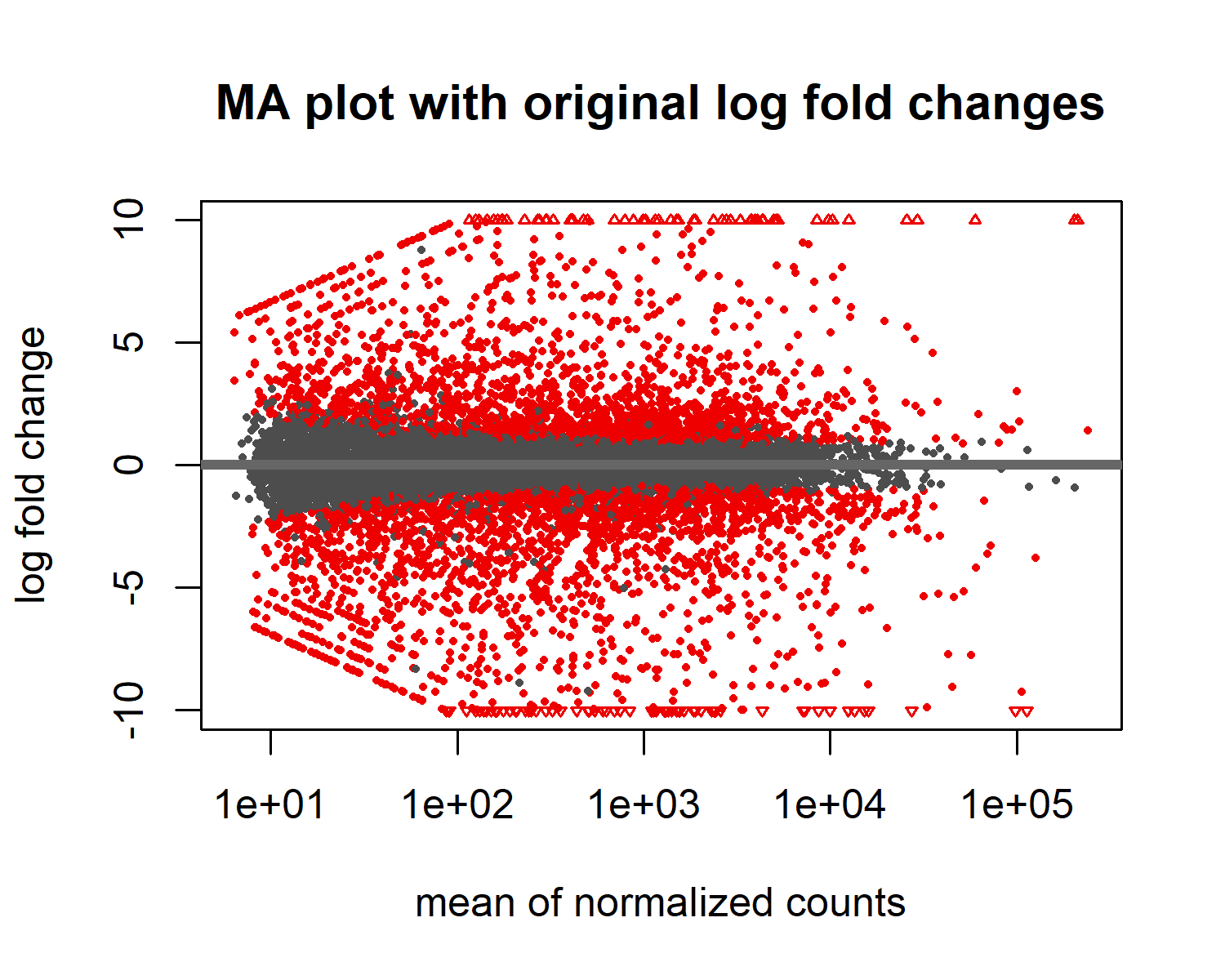


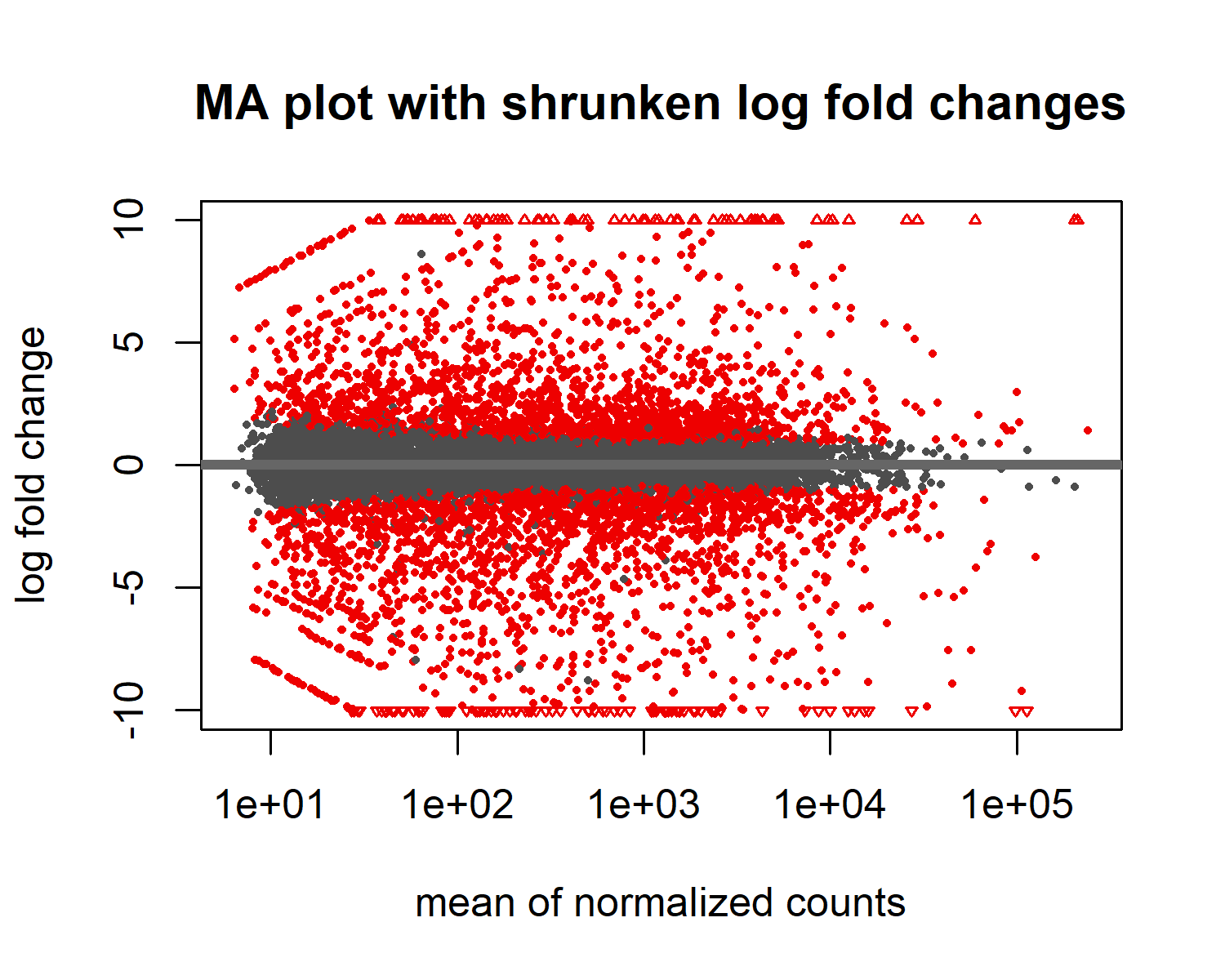
## MA plot

An MA-plot (Dudoit et al. 2002) provides a useful overview for the distribution of the estimated coefficients in the model, e.g. the comparisons of interest, across all genes. On the y-axis, the “M” stands for “minus” – subtraction of log values is equivalent to the log of the ratio – and on the x-axis, the “A” stands for “average”. You may hear this plot also referred to as a mean-difference plot, or a Bland-Altman plot.

Before making the MA-plot, we use the lfcShrink function to shrink the log2 fold changes for the comparison of DSCs vs. Melanocytes. There are three types of shrinkage estimators in DESeq2, which are covered in the DESeq2 vignette. Here we specify the apeglm method for shrinking coefficients, which is good for shrinking the noisy LFC estimates while giving low bias LFC estimates for true large differences (Zhu, Ibrahim, and Love 2018). To use apeglm we specify a coefficient from the model to shrink, either by name or number as the coefficient appears in resultsNames(dds).

WARNING: DO NOT ALWAYS SHRINK EVERYTHING BY DEFAULT. Look at MA plot and Volcano plot to see if it makes sense.

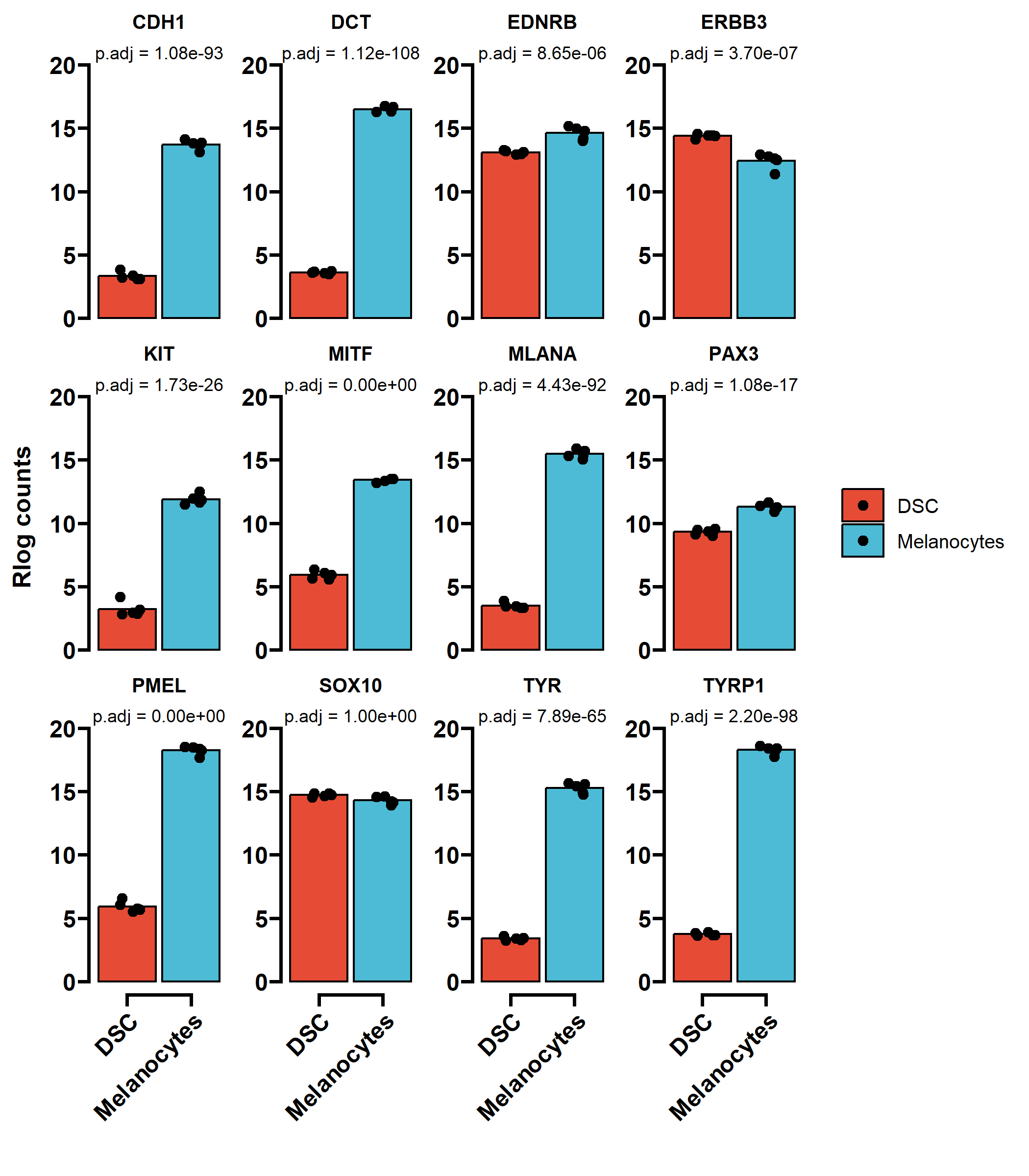




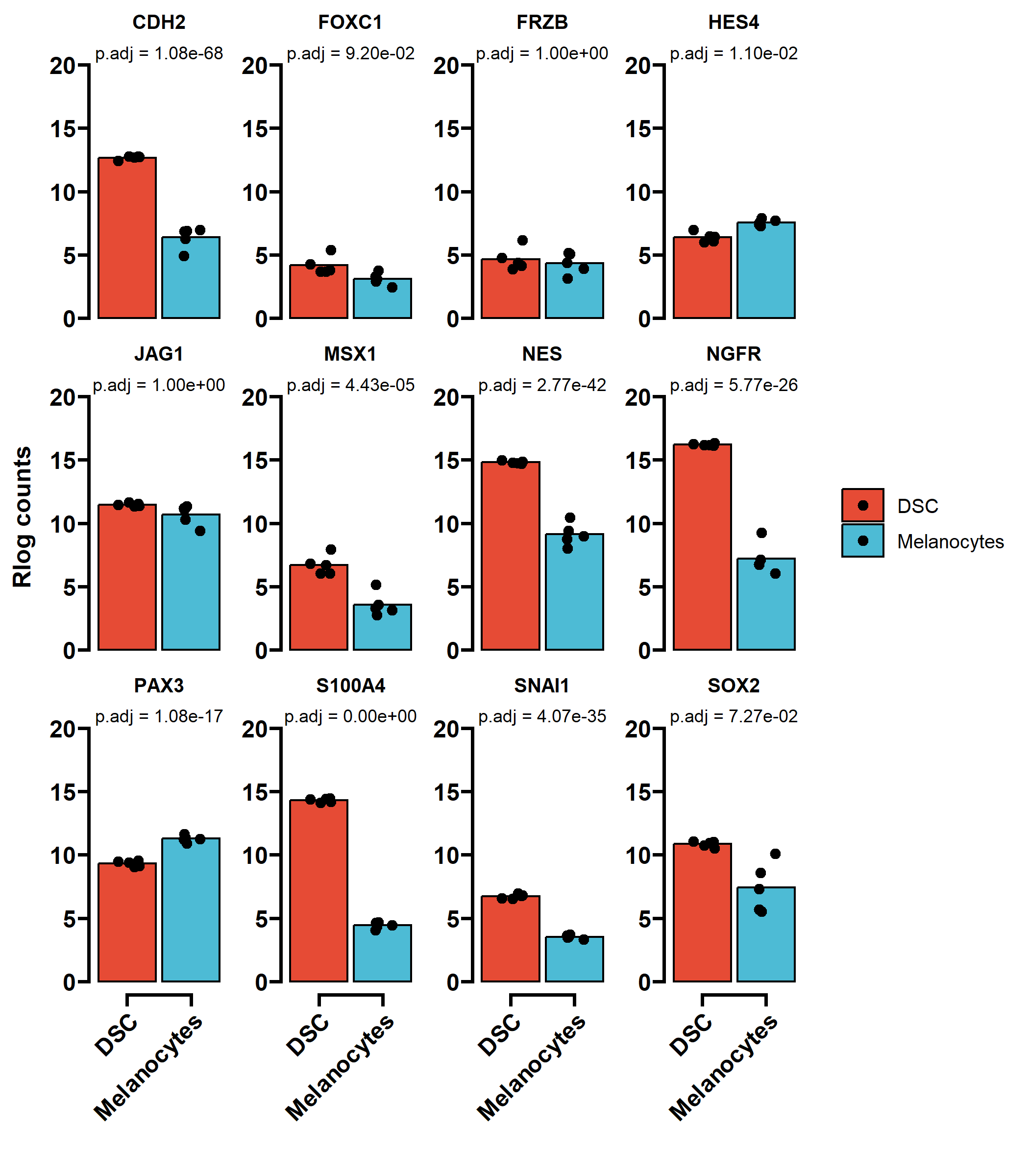
## Comparison of marker genes

## Expression of melanocyte markers

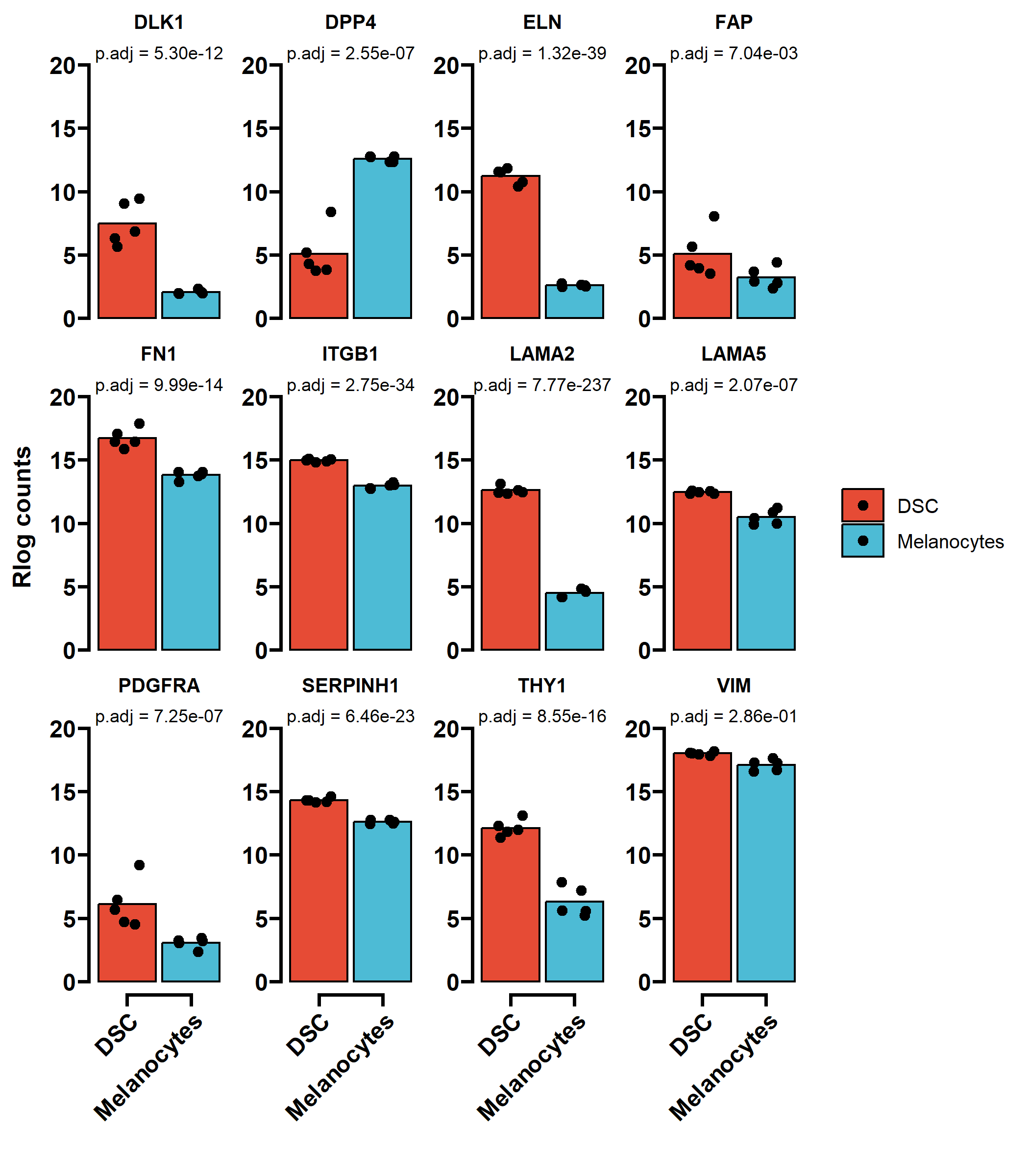
Marker genes were labelled according to the publication of Zabierowski 2012, lab practice and <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4278762/>.



## Expression of dsc markers

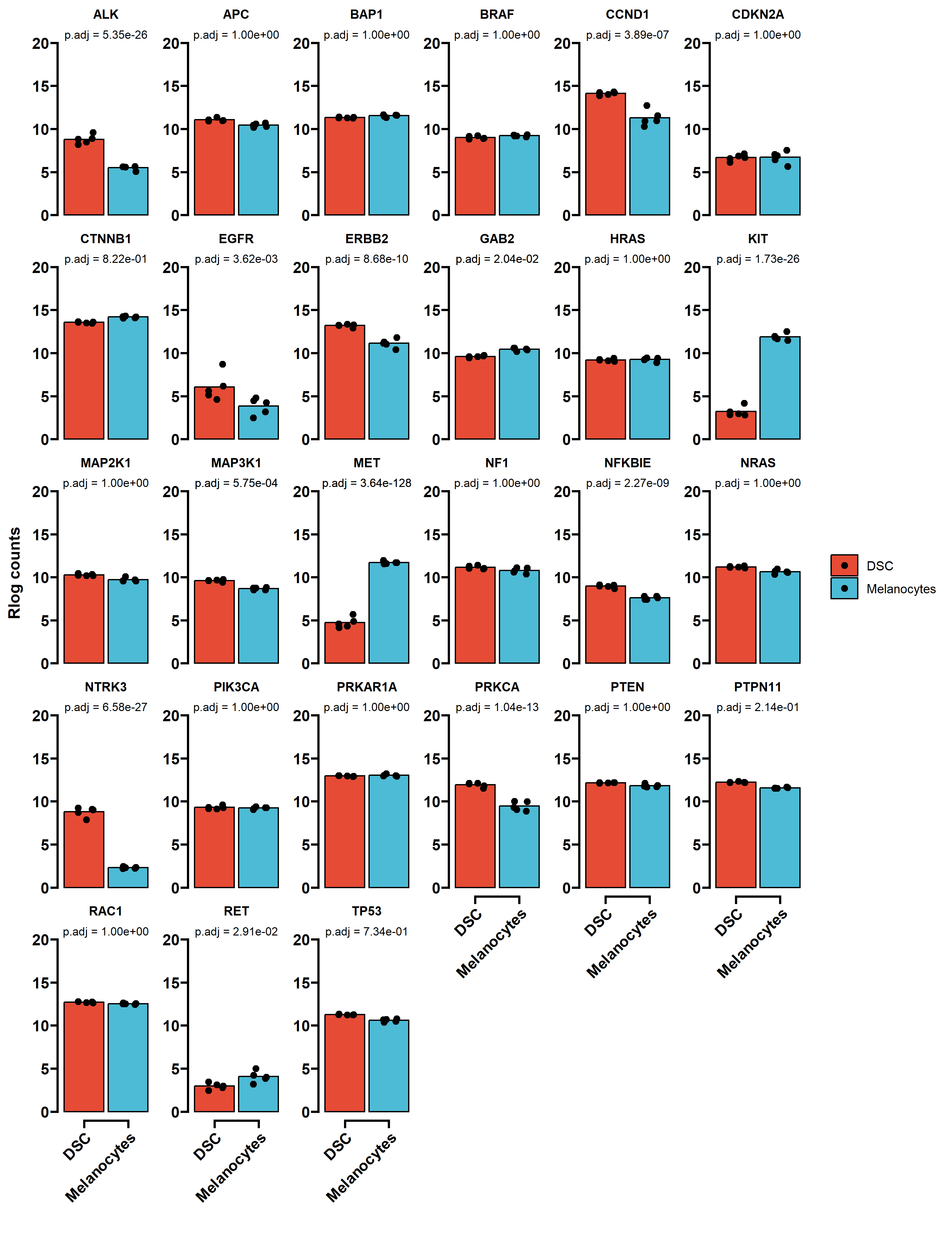


## Expression of fibroblast markers



## Expression of genes associated with melanomgenesis

The following genes were extracted from a publication of Cherepakhin et. al (2022) (<https://doi.org/10.3390/cancers14010123>) for cutaneous melanoma entities.

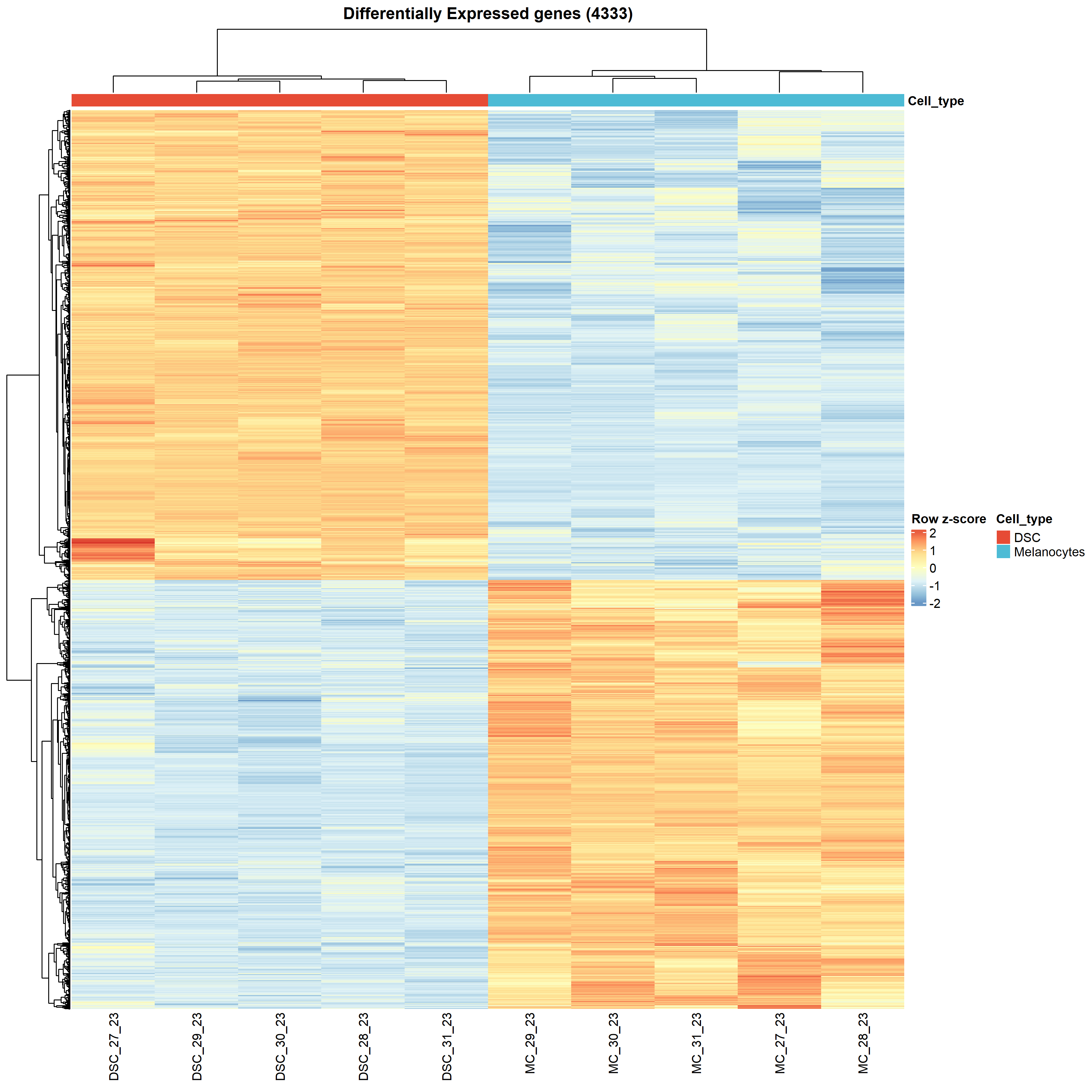


## Heatmap

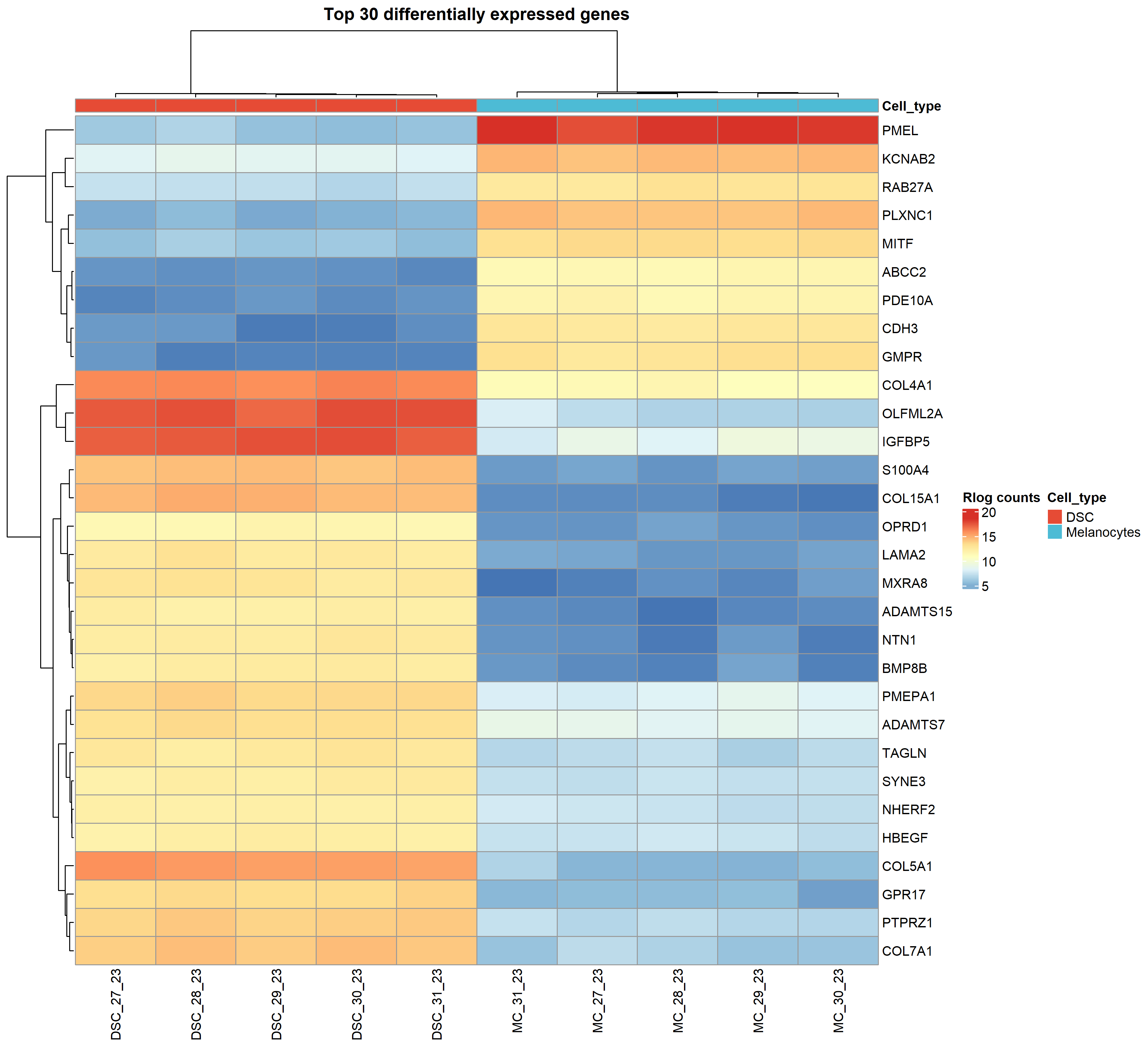
If using either of Euclidean distance or Pearson correlation, data should follow a Gaussian / normal (parametric) distribution. So, if coming from a microarray, anything from RMA normalisation is fine, whereas, if coming from RNA-seq, any data deriving from a transformed normalised count metric should be fine, such as variance-stabilised, regularised log, or log CPM expression levels.

If you are performing clustering on non-normal data, like ‘normalised’ [non-transformed] RNA-seq counts, FPKM expression units, etc., then use Spearman correlation (non-parametric).

### All differentially expressed (DE) genes

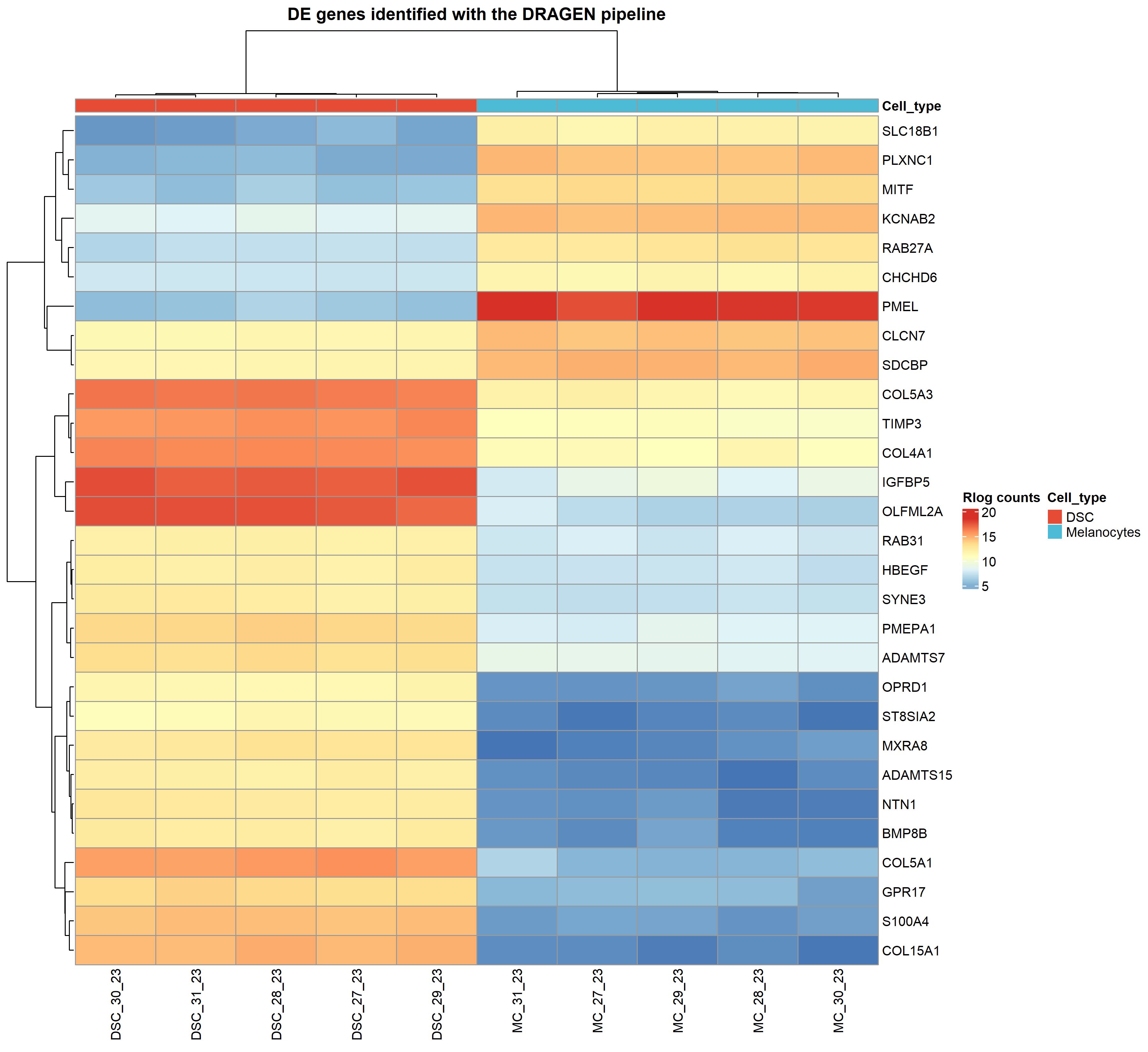


### Top 30 DE genes



### Comparison of genes identified with the DRAGEN pipeline

This plot includes the same 30 genes which are depicted in the Heatmap included in the output generated from the DRAGEN analysis software.



# Appenix

## Session Info

## R version 4.3.2 (2023-10-31 ucrt)  
## Platform: x86\_64-w64-mingw32/x64 (64-bit)  
## Running under: Windows 10 x64 (build 17763)  
##   
## Matrix products: default  
##   
##   
## locale:  
## [1] LC\_COLLATE=German\_Germany.1252 LC\_CTYPE=German\_Germany.1252   
## [3] LC\_MONETARY=German\_Germany.1252 LC\_NUMERIC=C   
## [5] LC\_TIME=German\_Germany.1252   
##   
## time zone: Europe/Berlin  
## tzcode source: internal  
##   
## attached base packages:  
## [1] grid stats4 stats graphics grDevices utils datasets   
## [8] methods base   
##   
## other attached packages:  
## [1] ggprism\_1.0.4 ComplexHeatmap\_2.18.0   
## [3] pheatmap\_1.0.12 EnhancedVolcano\_1.20.0   
## [5] ggrepel\_0.9.5 org.Hs.eg.db\_3.18.0   
## [7] AnnotationDbi\_1.64.1 lubridate\_1.9.3   
## [9] forcats\_1.0.0 stringr\_1.5.1   
## [11] dplyr\_1.1.4 purrr\_1.0.2   
## [13] readr\_2.1.5 tidyr\_1.3.0   
## [15] tibble\_3.2.1 ggplot2\_3.4.4   
## [17] tidyverse\_2.0.0 DESeq2\_1.42.0   
## [19] SummarizedExperiment\_1.32.0 Biobase\_2.62.0   
## [21] MatrixGenerics\_1.14.0 matrixStats\_1.2.0   
## [23] GenomicRanges\_1.54.1 GenomeInfoDb\_1.38.5   
## [25] IRanges\_2.36.0 S4Vectors\_0.40.2   
## [27] BiocGenerics\_0.48.1   
##   
## loaded via a namespace (and not attached):  
## [1] RColorBrewer\_1.1-3 rstudioapi\_0.15.0 jsonlite\_1.8.8   
## [4] shape\_1.4.6 magrittr\_2.0.3 magick\_2.8.3   
## [7] farver\_2.1.1 rmarkdown\_2.25 GlobalOptions\_0.1.2   
## [10] zlibbioc\_1.48.0 ragg\_1.2.7 vctrs\_0.6.5   
## [13] memoise\_2.0.1 RCurl\_1.98-1.14 askpass\_1.2.0   
## [16] htmltools\_0.5.7 S4Arrays\_1.2.0 curl\_5.2.0   
## [19] SparseArray\_1.2.3 cachem\_1.0.8 uuid\_1.1-1   
## [22] mime\_0.12 lifecycle\_1.0.4 iterators\_1.0.14   
## [25] pkgconfig\_2.0.3 Matrix\_1.6-5 R6\_2.5.1   
## [28] fastmap\_1.1.1 GenomeInfoDbData\_1.2.11 shiny\_1.8.0   
## [31] clue\_0.3-65 digest\_0.6.34 colorspace\_2.1-0   
## [34] textshaping\_0.3.7 RSQLite\_2.3.4 labeling\_0.4.3   
## [37] fansi\_1.0.6 timechange\_0.2.0 httr\_1.4.7   
## [40] abind\_1.4-5 compiler\_4.3.2 bit64\_4.0.5   
## [43] fontquiver\_0.2.1 withr\_2.5.2 doParallel\_1.0.17   
## [46] BiocParallel\_1.36.0 DBI\_1.2.1 highr\_0.10   
## [49] openssl\_2.1.1 DelayedArray\_0.28.0 rjson\_0.2.21   
## [52] ggsci\_3.0.0 gfonts\_0.2.0 tools\_4.3.2   
## [55] zip\_2.3.0 httpuv\_1.6.13 glue\_1.7.0   
## [58] promises\_1.2.1 cluster\_2.1.6 generics\_0.1.3   
## [61] gtable\_0.3.4 tzdb\_0.4.0 data.table\_1.14.10   
## [64] hms\_1.1.3 xml2\_1.3.6 utf8\_1.2.4   
## [67] XVector\_0.42.0 foreach\_1.5.2 pillar\_1.9.0   
## [70] later\_1.3.2 circlize\_0.4.15 lattice\_0.22-5   
## [73] bit\_4.0.5 ekbSeq\_0.0.2 tidyselect\_1.2.0   
## [76] fontLiberation\_0.1.0 locfit\_1.5-9.8 Biostrings\_2.70.1   
## [79] knitr\_1.45 fontBitstreamVera\_0.1.1 crul\_1.4.0   
## [82] xfun\_0.41 stringi\_1.8.3 yaml\_2.3.8   
## [85] evaluate\_0.23 codetools\_0.2-19 httpcode\_0.3.0   
## [88] officer\_0.6.5 gdtools\_0.3.6 cli\_3.6.2   
## [91] xtable\_1.8-4 systemfonts\_1.0.5 munsell\_0.5.0   
## [94] Rcpp\_1.0.12 png\_0.1-8 parallel\_4.3.2   
## [97] ellipsis\_0.3.2 blob\_1.2.4 bitops\_1.0-7   
## [100] scales\_1.3.0 crayon\_1.5.2 flextable\_0.9.4   
## [103] GetoptLong\_1.0.5 rlang\_1.1.3 KEGGREST\_1.42.0