Next-generation sequencing - analysis with DESeq2

Comparison of gene expression in DSCs compared to melanocytes

Marc Bender

2024-04-06

Table of Contents

# Resources

For additional info on analysis workflows check: <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>

# Thresholds, design formula and number of transcripts

|  | **Value** |
| --- | --- |
| **Thresholds** |  |
| P-value | 0.05 |
| Log-fold change | 0.585 |
| Minimum read count | 10 |
| Smallest group size | 5 |
| **Design Formula** | ~cell\_type |
| **Transcript number** |  |
| Total number of transcripts | 60609 |
| Filtered number of transcripts† | 15536 |
| †All transcripts which are above the minimum read count threshold (10) in at   least 5 samples. | |

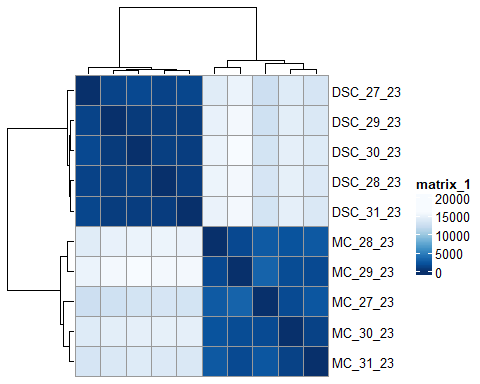
|  | **DSC\_27\_23** | **DSC\_28\_23** | **DSC\_29\_23** | **DSC\_30\_23** | **DSC\_31\_23** | **MC\_27\_23** | **MC\_28\_23** | **MC\_29\_23** | **MC\_30\_23** | **MC\_31\_23** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Number of reads (total) | 2.024e+07 | 2.298e+07 | 2.215e+07 | 1.996e+07 | 2.114e+07 | 2.132e+07 | 2.260e+07 | 2.663e+07 | 2.537e+07 | 2.360e+07 |
| Number of reads (filtered)† | 2.020e+07 | 2.295e+07 | 2.212e+07 | 1.993e+07 | 2.112e+07 | 2.129e+07 | 2.256e+07 | 2.660e+07 | 2.534e+07 | 2.358e+07 |
| †All transcripts which are above the minimum read count threshold (10) in at   least5 samples. | | | | | | | | | | |

# Sample sheet data

# Exploratory data analysis

## Heatmap of Poisson distances between samples

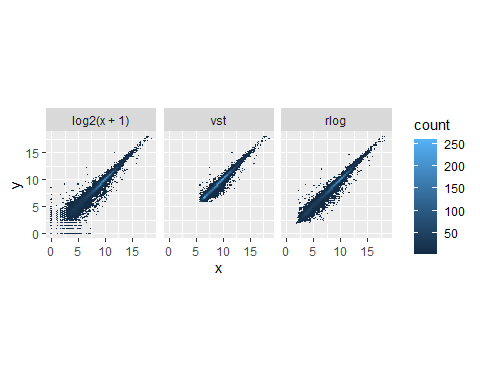
* IMPORTANT:
  + use Poisson distance for raw (non-normalized) count data
  + use Euclidean distance for data normalized by regularized-logarithm transformation (rlog) or variance stablization transfromation (vst)



## Transform data

Which transformation to choose? The VST is much faster to compute and is less sensitive to high count outliers than the rlog. The rlog tends to work well on small datasets (n < 30), potentially outperforming the VST when there is a wide range of sequencing depth across samples (an order of magnitude difference). We therefore recommend the VST for medium-to-large datasets (n > 30). You can perform both transformations and compare the meanSdPlot or PCA plots generated, as described below.

To show the effect of the transformation, in the figure below we plot the first sample against the second, first simply using the log2 function (after adding 1, to avoid taking the log of zero), and then using the VST and rlog-transformed values. For the log2 approach, we need to first estimate size factors to account for sequencing depth, and then specify normalized=TRUE. Sequencing depth correction is done automatically for the vst and rlog.



## Principal component analysis (PCA)

Die Hauptkomponentenanalyse oder *principal component analysis* (PCA) ist ein Verfahren zur Reduktion der Datendimensionalität und findet Anwendung bei komplexen Datensätzen (aus n Proben mit p gemessenen Merkmalen), um Muster oder Strukturen in den Daten zu erkennen und die Interpretation der Daten zu vereinfachen. Dies erfolgt durch Identifikation neuer unkorrelierter Variablen, den sog. Hauptkomponenten (PC = *principal component*), welche sukzessive die erklärte Varianz innerhalb eines Datensets maximieren. Die Berechnung dieser Hauptkomponenten erfolgt dabei durch Zentrieren der Daten und einen nachfolgenden iterativen Prozess, bei dem eine zufällige Gerade durch den Mittelwert der Datenpunkte gelegt und so lange rotiert wird, bis der Abstand der auf die Gerade projizierten Datenpunkte zum Zentrum maximal ist. Die Distanzen werden berechnet als:

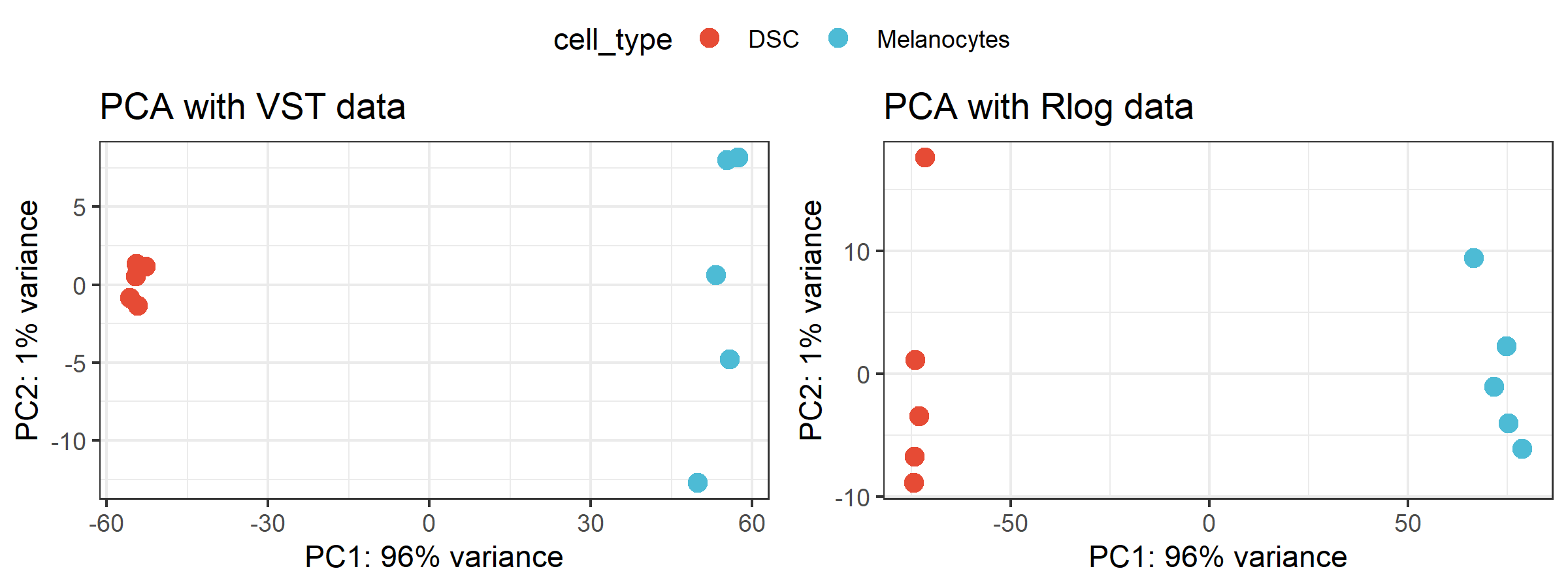
mit SS(distances) = Eigenwert (Quadratsumme der Distanzen), d = Distanz vom projizierten Datenpunkt zum Zentrum, n = Probengröße.

Die Gerade mit dem größten Eigenwert wird als PC1 bezeichnet und trägt am meisten zur Varianzaufklärung bei. Im Anschluss wird die nächste Gerade gesucht, die orthogonal zur ersten Gerade ist und ebenfalls durch den Mittelwert der Datenpunkte verläuft. Nach diesem Grundsatz können (bis zur n-ten Hauptkomponente) weitere Hauptkomponenten gefunden werden, wobei diese stets orthogonal zu den bisherigen Hauptkomponenten stehen müssen.

Die Reduktion der Dimensionalität beruht darauf, dass die jeweiligen Hauptkomponenten einen unterschiedlich großen (absteigend von PC1 zu PCn) Anteil der Datenvarianz erklären und es daher in den meisten Fällen ausreicht die ersten drei Hauptkomponenten zu analysieren (welche jeweils als 2-dimensionaler Plot gegeneinander aufgetragen werden), um Gemeinsamkeiten innerhalb der Proben zu identifizieren.  
Der Anteil der Varianz, der durch die jeweilige Hauptkomponente erklärt wird (), kann anhand folgender Formel berechnet werden:

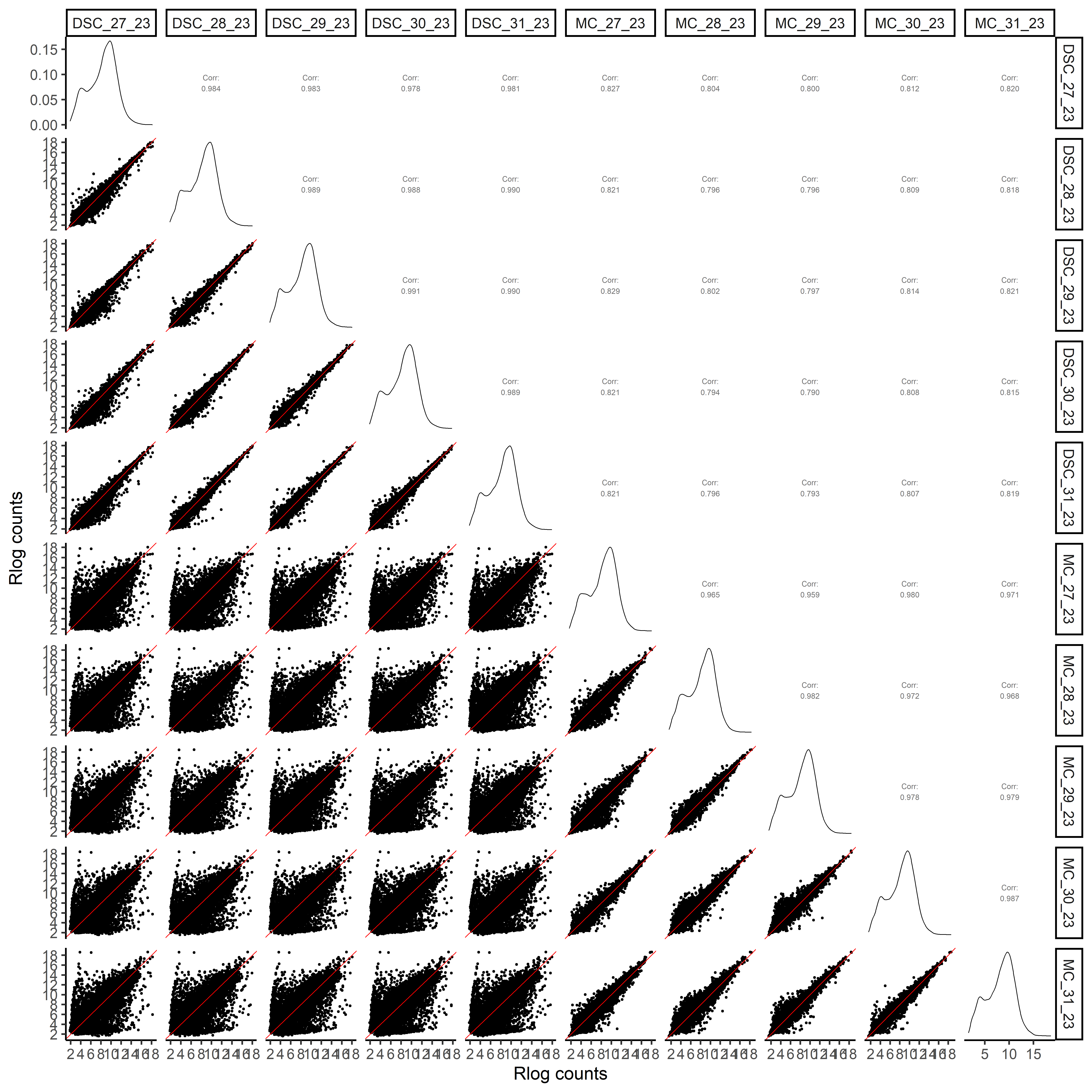
mit = Hauptkomponente i.

Die PCA wurde in dem Statistikprogramm R mithilfe der Pakete *factoextra* und *FactoMineR* durchgeführt. Die Ermittlung relevanter Hauptkomponenten erfolgte mittels Skree Plot. Dort werden entweder die Eigenwerte oder die erklärte Varianz gegen die Nummer der jeweiligen Hauptkomponente aufgetragen und mittels einer zunächst steil abfallenden Linie, die sich asymptotisch der Abzisse annähert, verbunden. An dem sog. *elbow* ist ein deutlicher Knick in der Linie zu erkennen. Faktoren, die links dieses Punktes liegen reichen größtenteils aus, um Muster in den Daten zu erklären. Die anderen Faktoren unterscheiden sich oft nicht deutlich von Zufallskorrelationen und sind daher nicht von Bedeutung.



## Scatter plot matrices

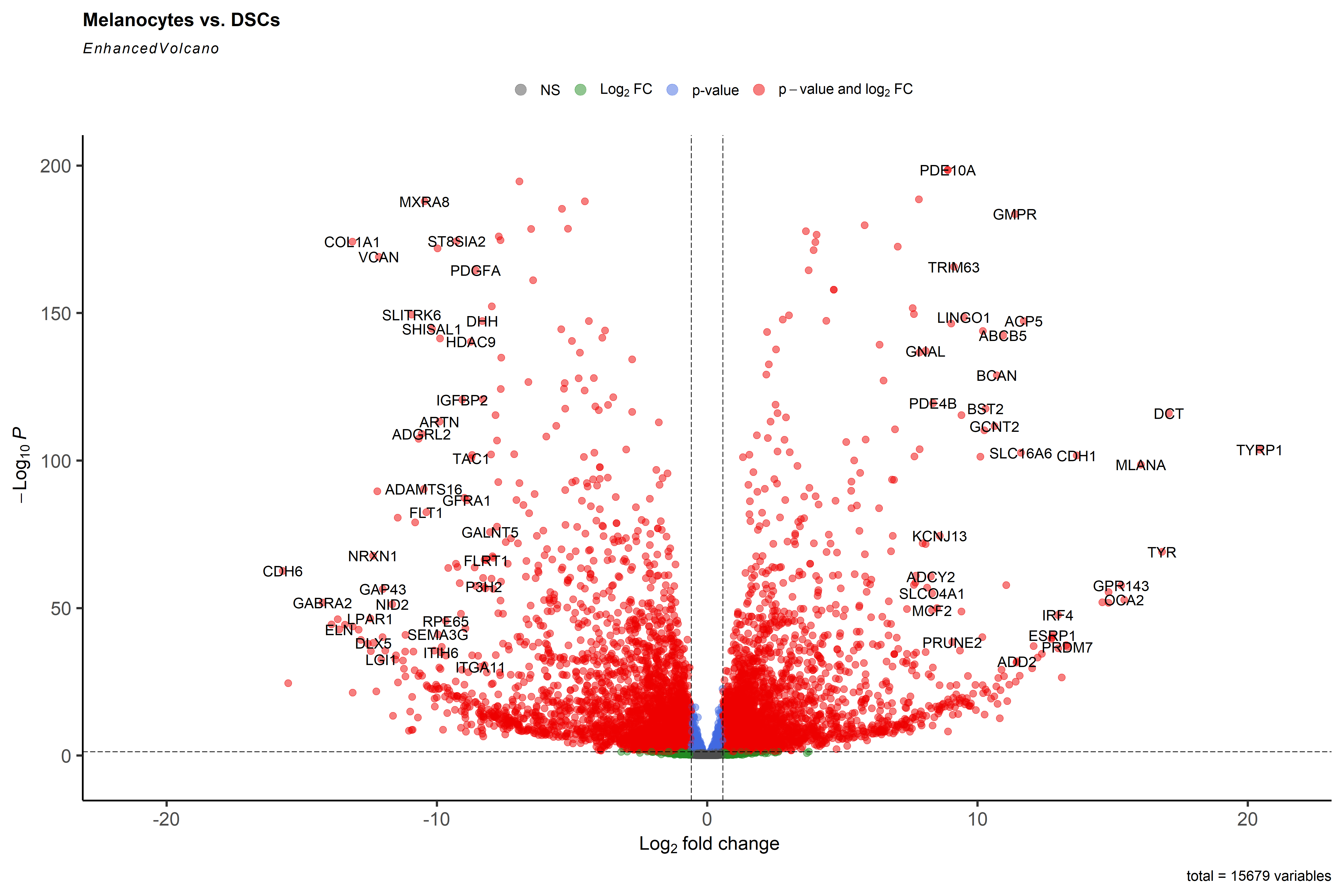
These types of plots were suggested in <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-2968-1>. They depict the correlation pairs of each sample with other samples. The spread of points should be closer to the identity line (red line) in technical replicates than in biological replicates because a larger spread indicates higher differences in expression values most often associated with differential expression. Divergence of the expected pattern (as described in the aforemented publication) should also warrant a closer inspection.



# Differential expression analysis

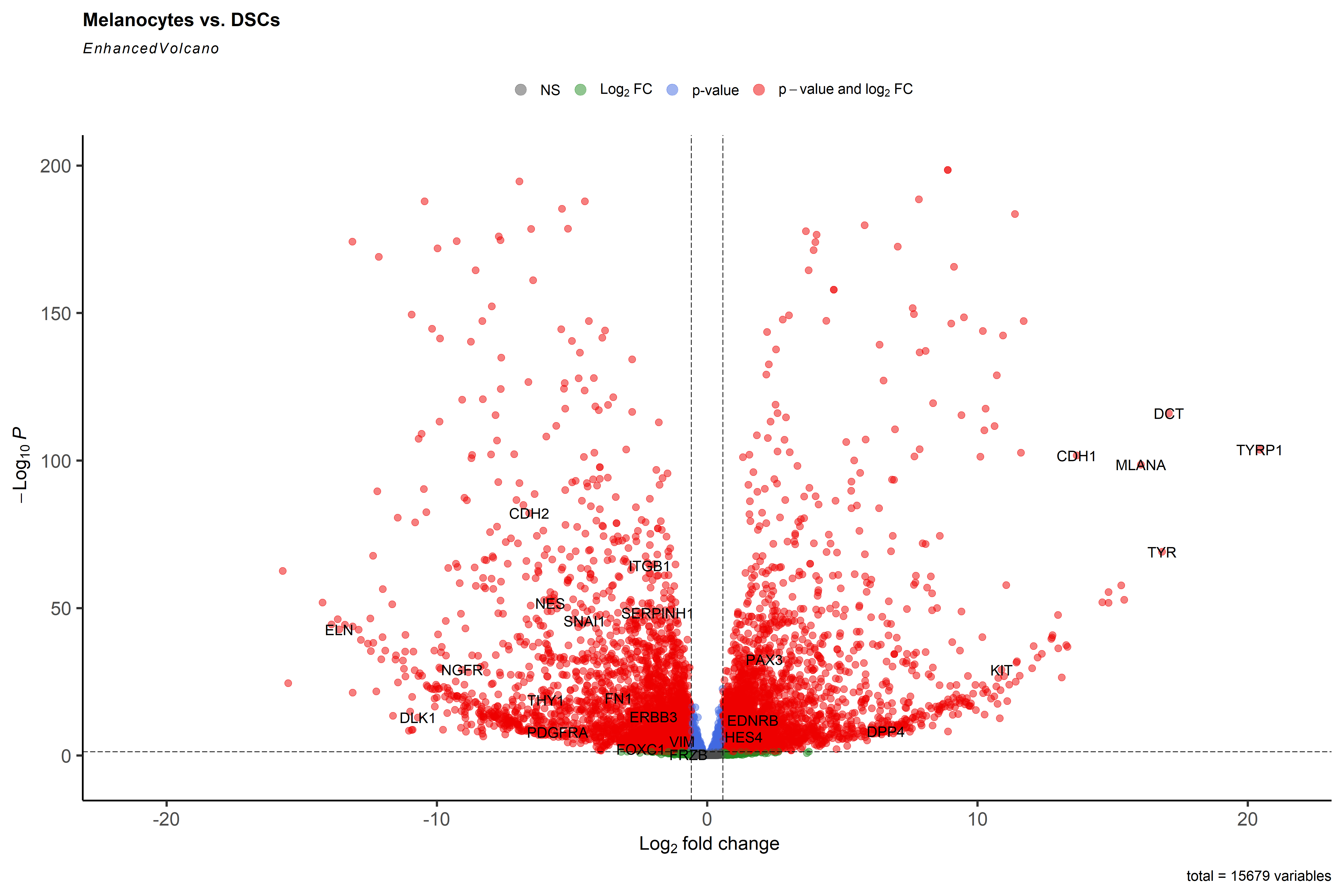
## Volcano Plot (most differentially expressed genes labelled)

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



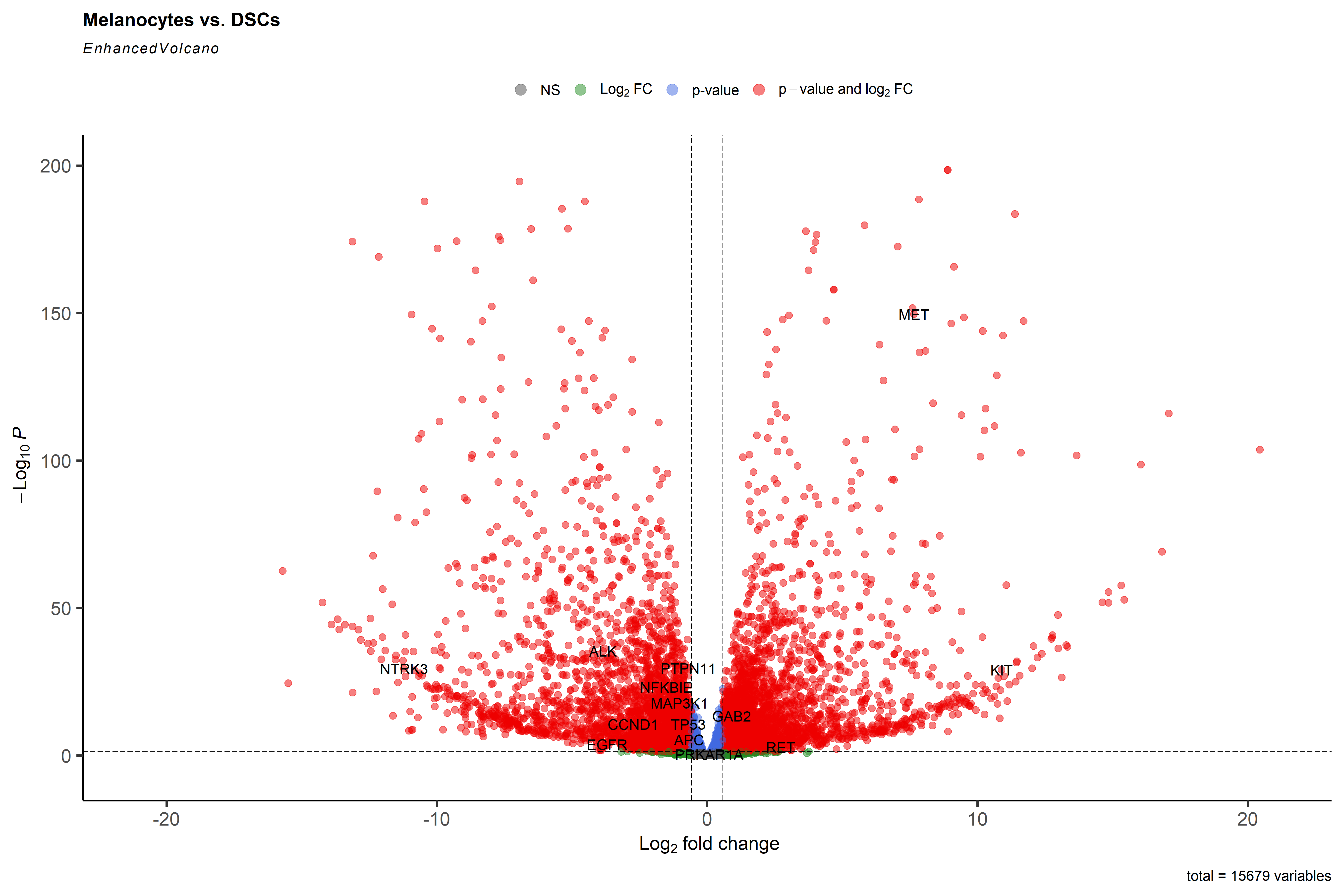
## Volcano Plot (skin cell marker genes labelled)

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)).

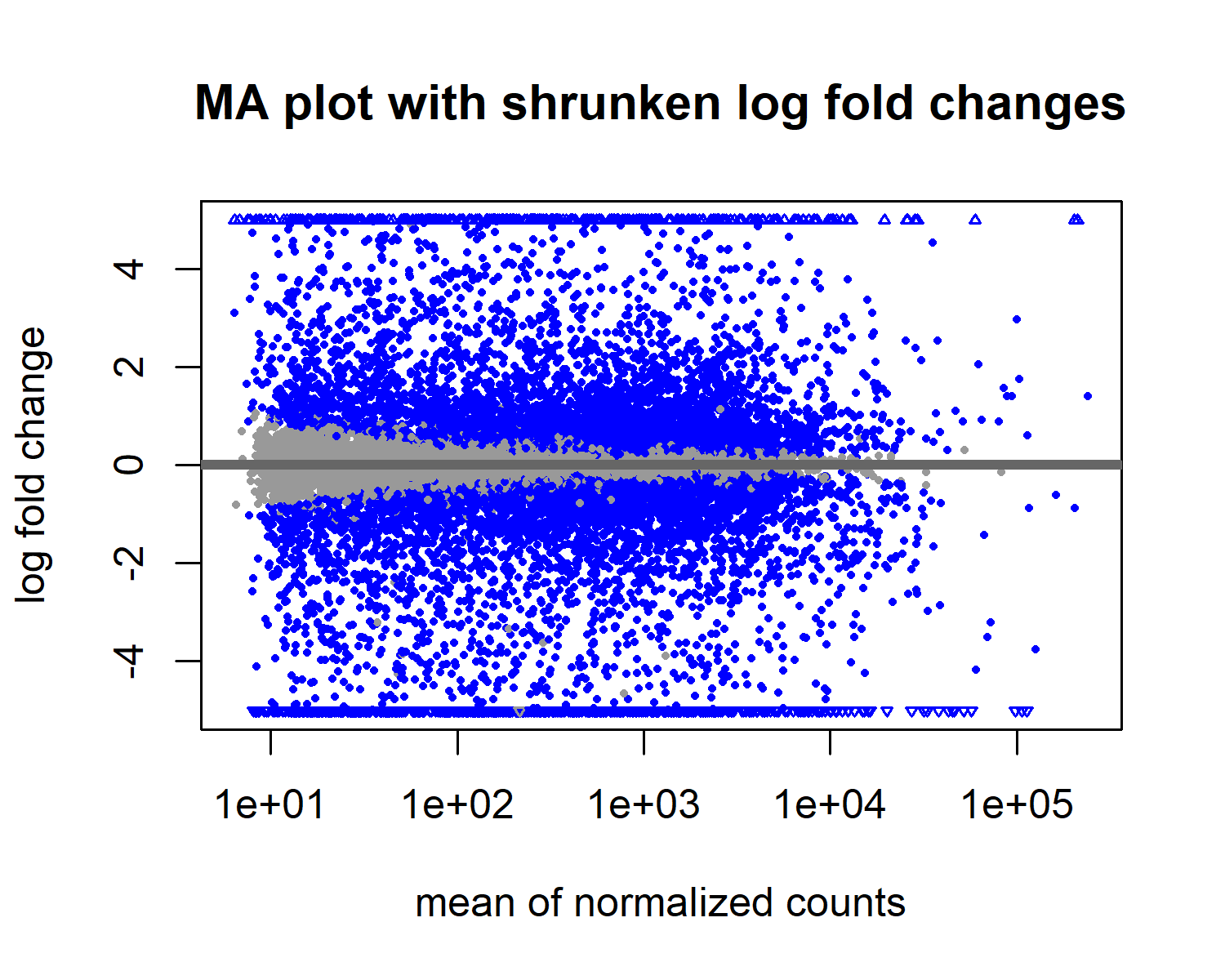
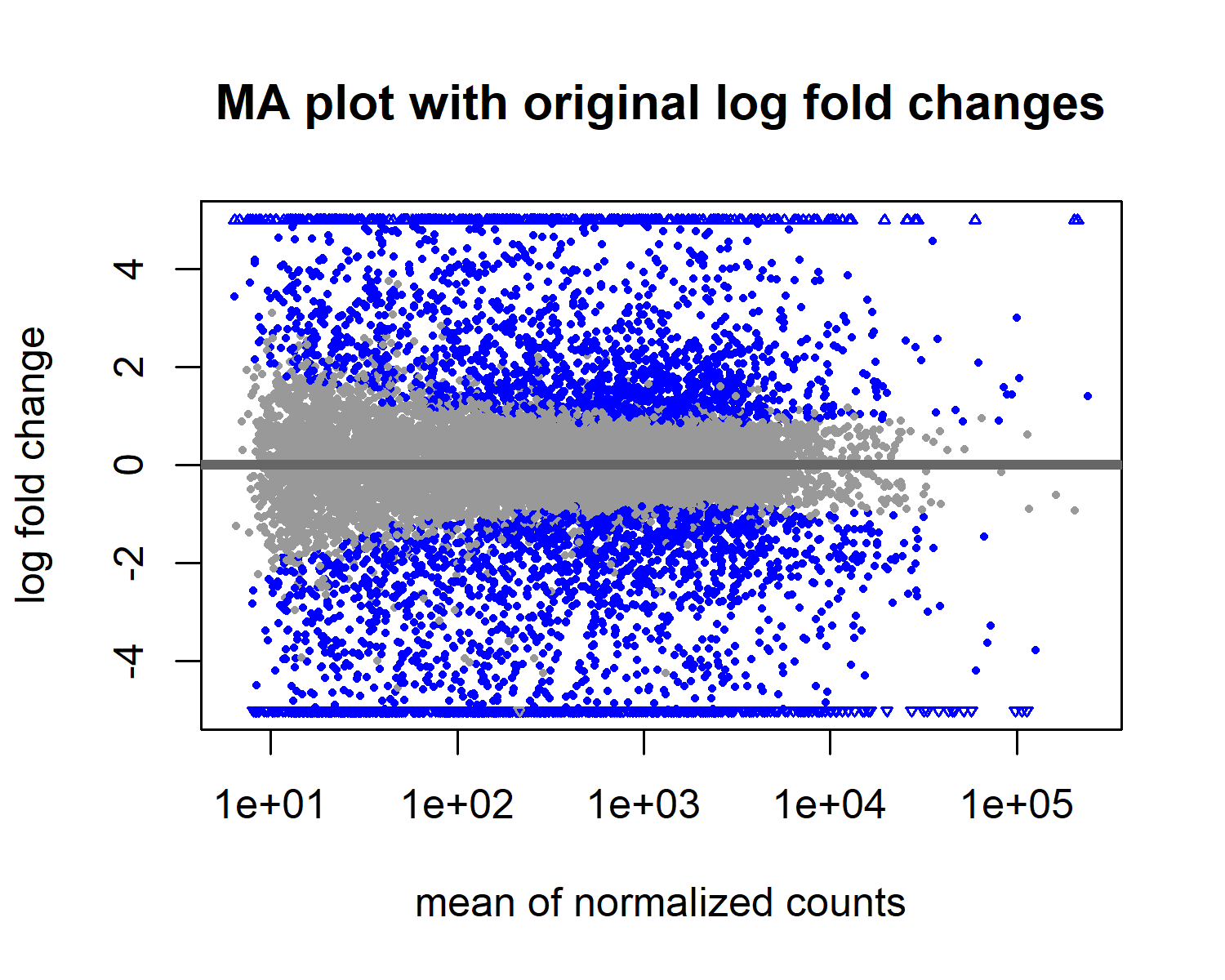


## MA plot

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).

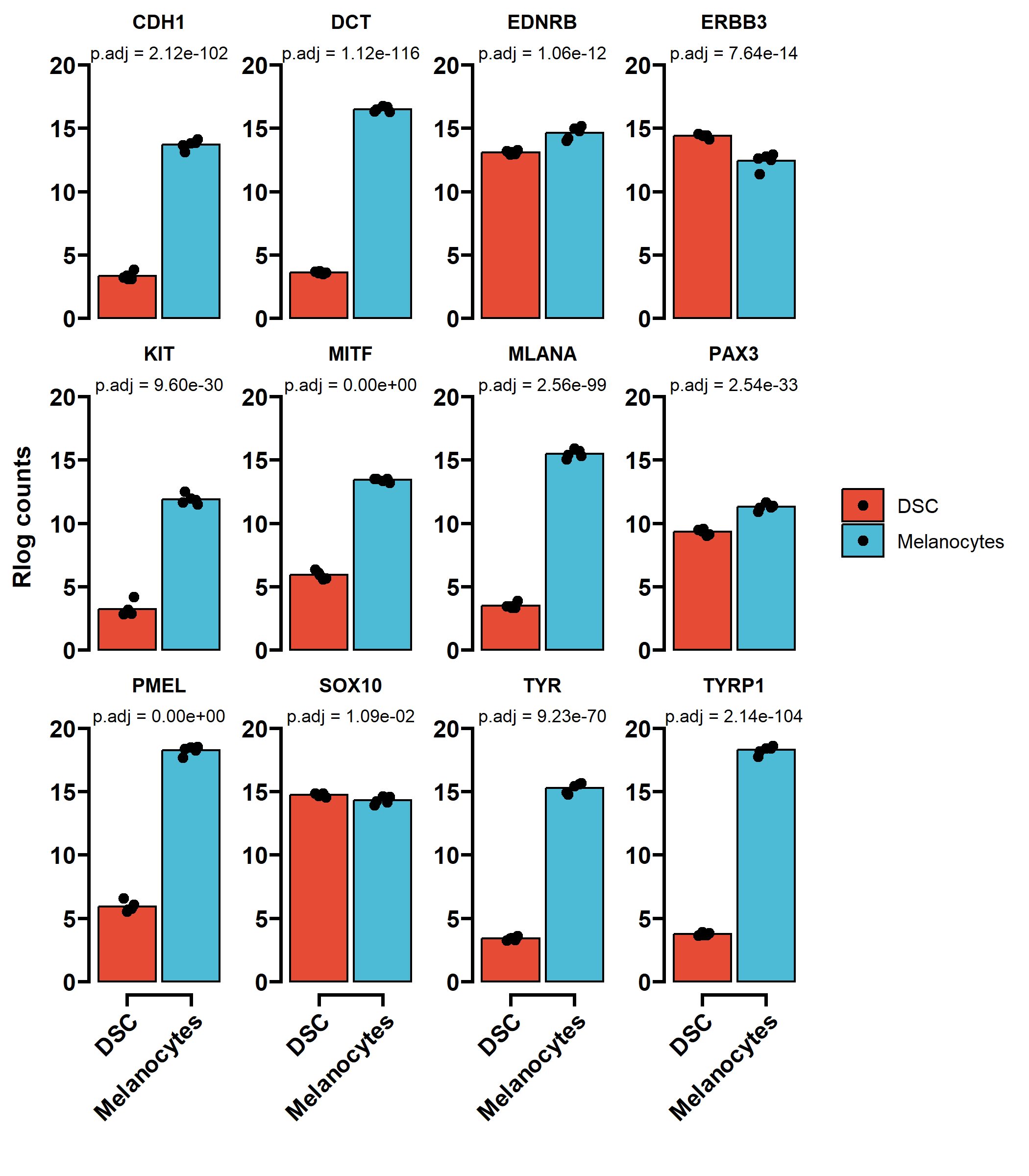
WARNING: DO NOT ALWAYS SHRINK EVERYTHING BY DEFAULT. Look at MA plot and Volcano plot to see if it makes sense.



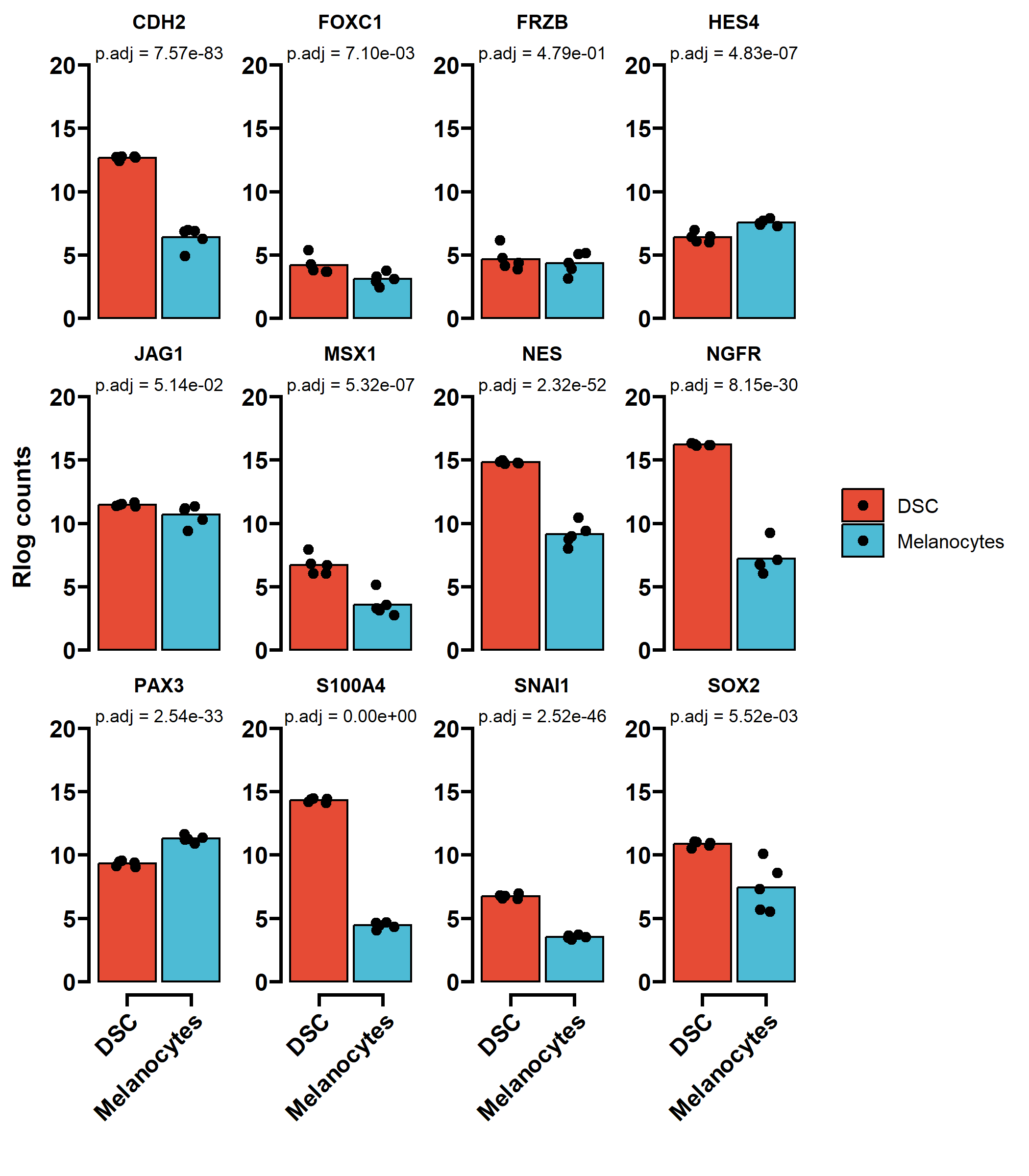
## Comparison of marker genes

## Expression of melanocyte markers

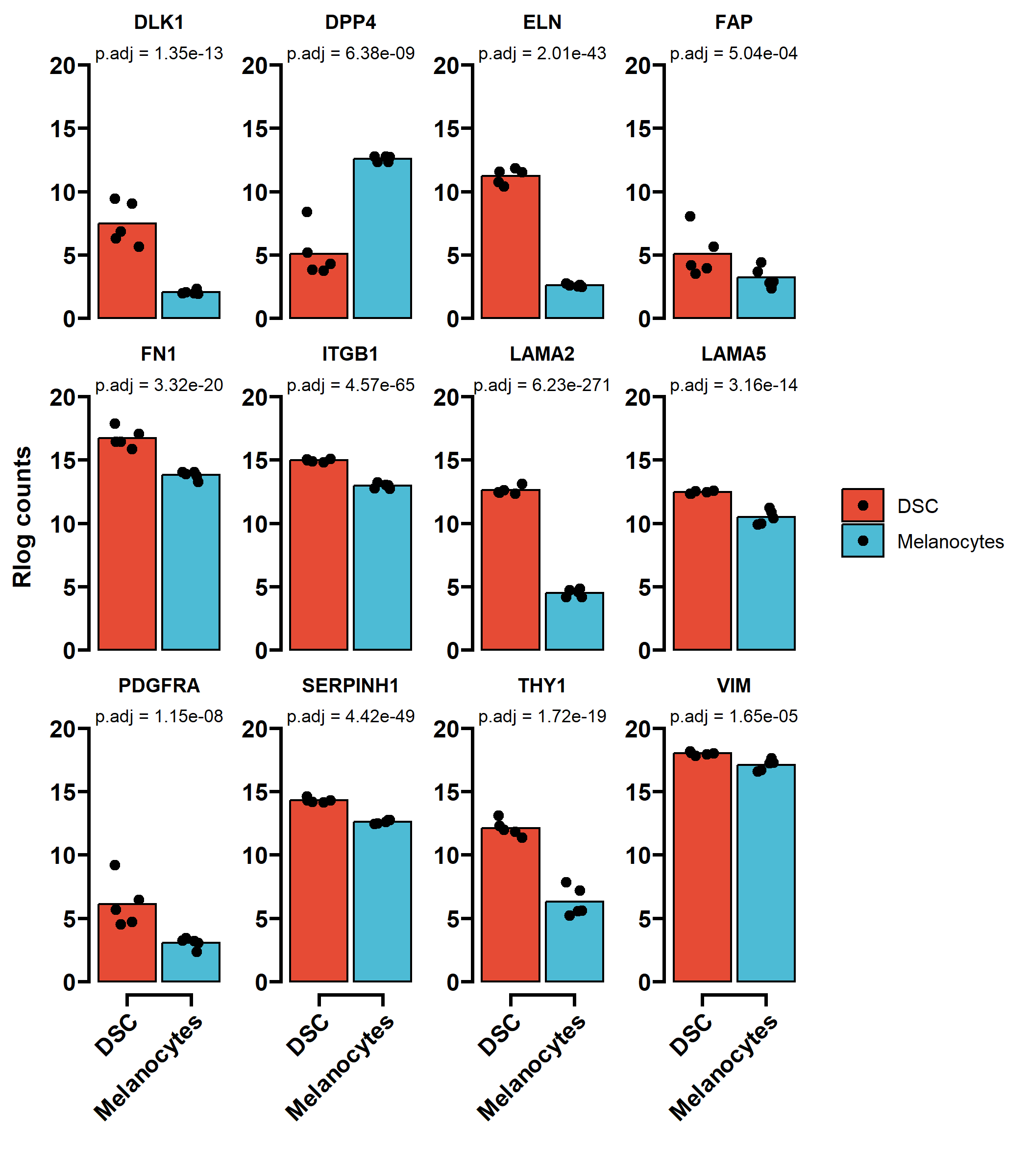
Marker genes were labelled according to the publication of Zabierowski 2012, lab practice and <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4278762/>.



## Expression of dsc markers

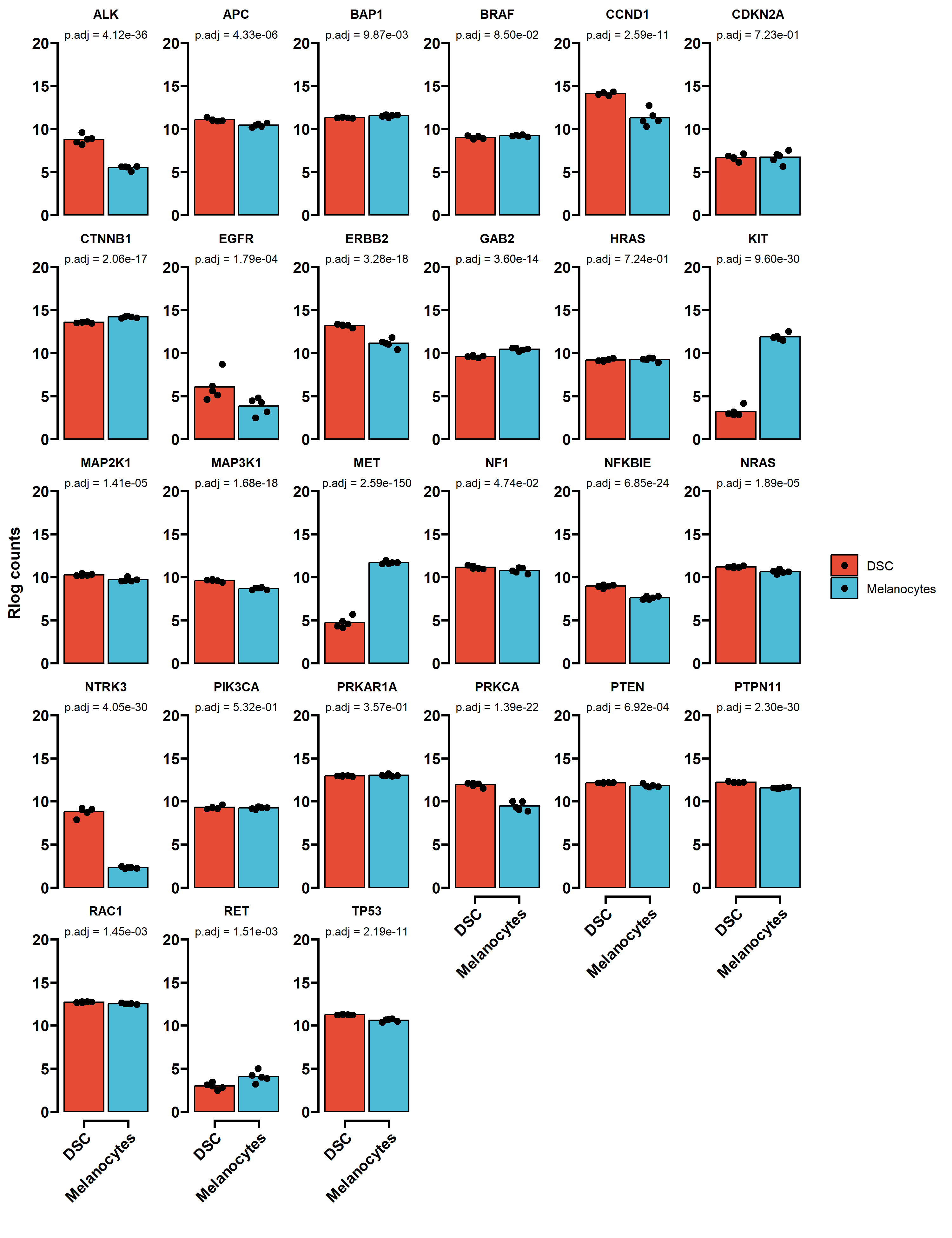


## Expression of fibroblast markers



## 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.

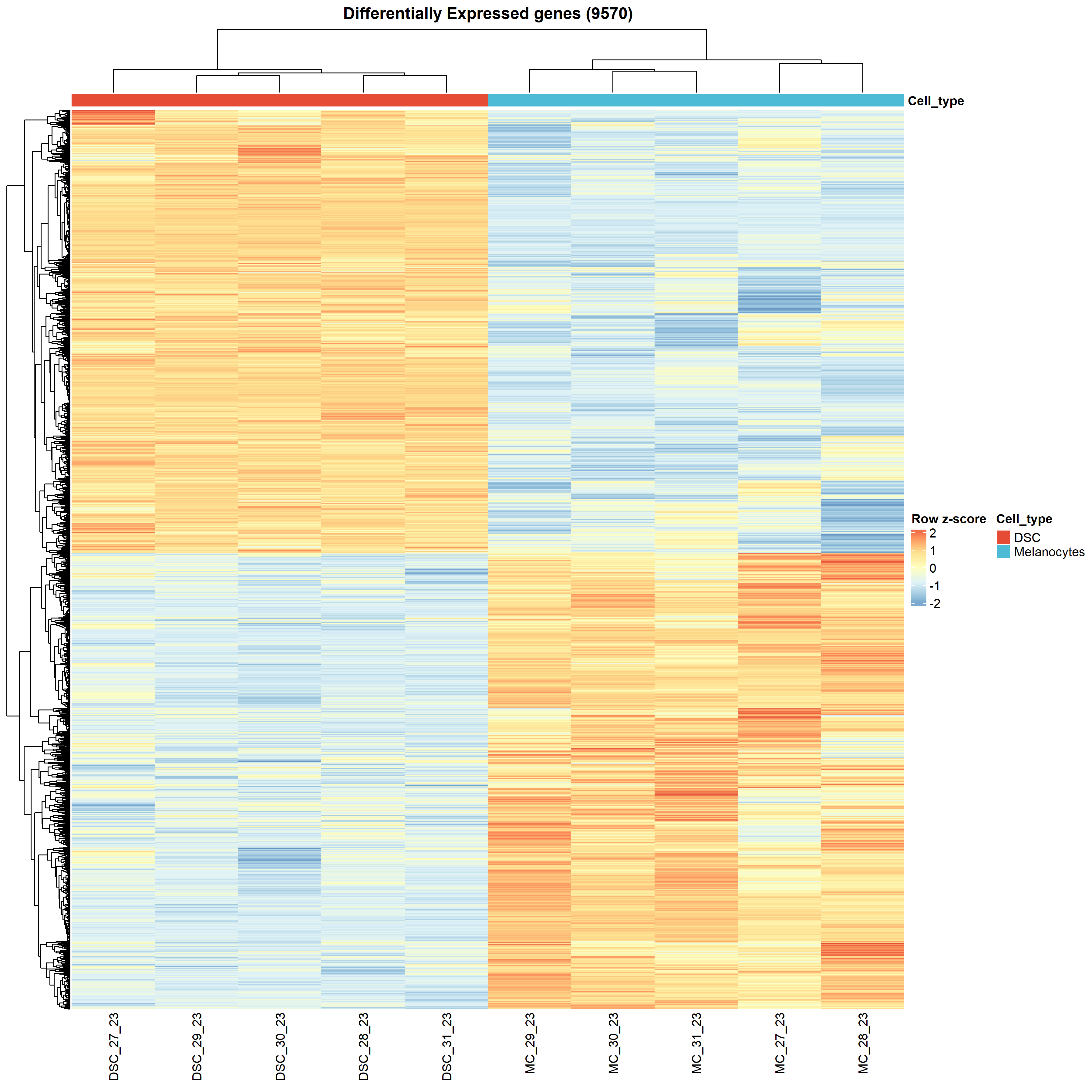


## Heatmap

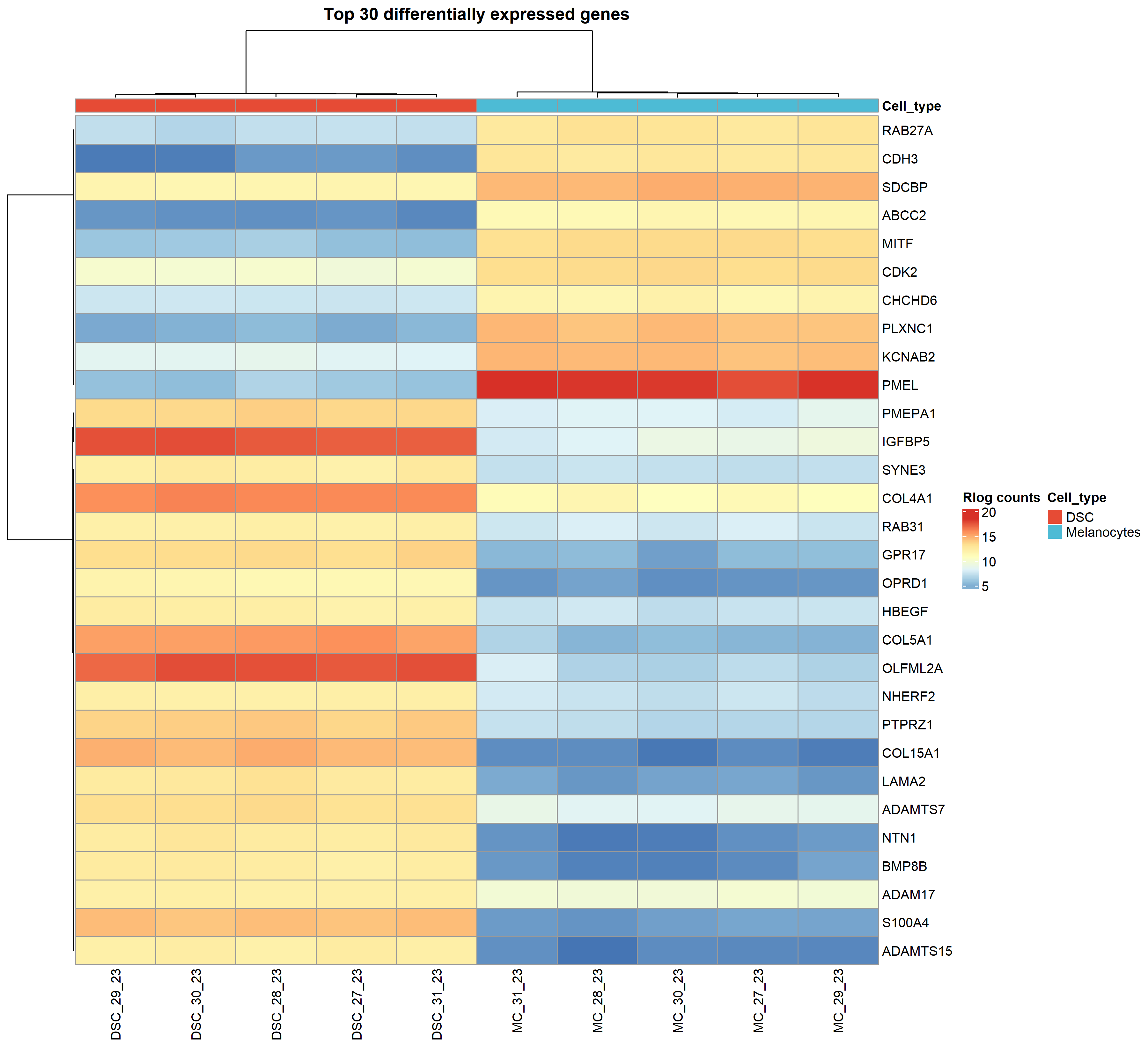
If using either of Euclidean distance or Pearson correlation, data should follow a Gaussian / normal (parametric) distribution. So, if coming from a microarray, anything from RMA normalisation is fine, whereas, if coming from RNA-seq, any data deriving from a transformed normalised count metric should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

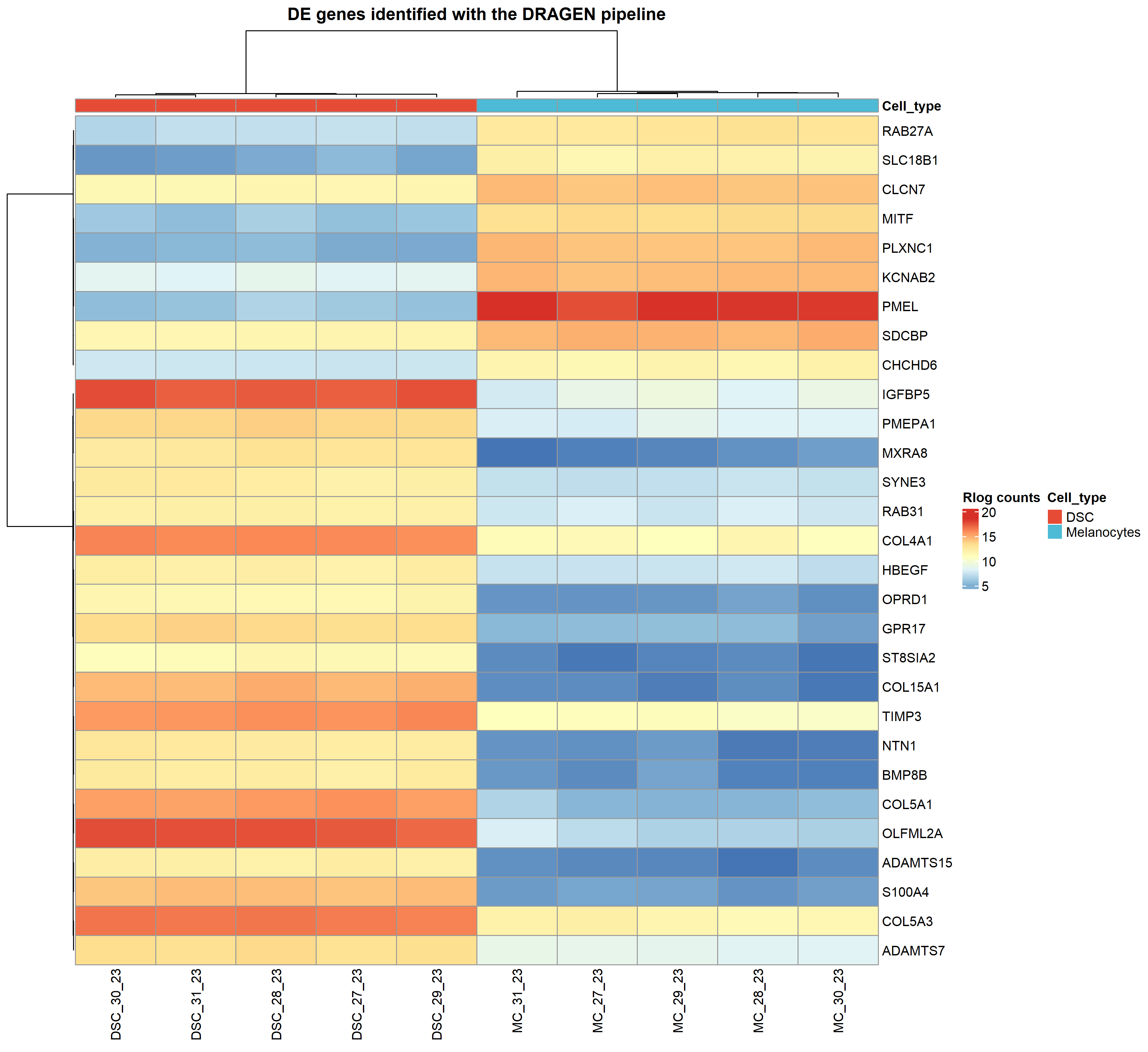


### Top 30 DE genes



### 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.



–> –> –> –> –> –> –>