Pathway analysis - Fig5

Contents

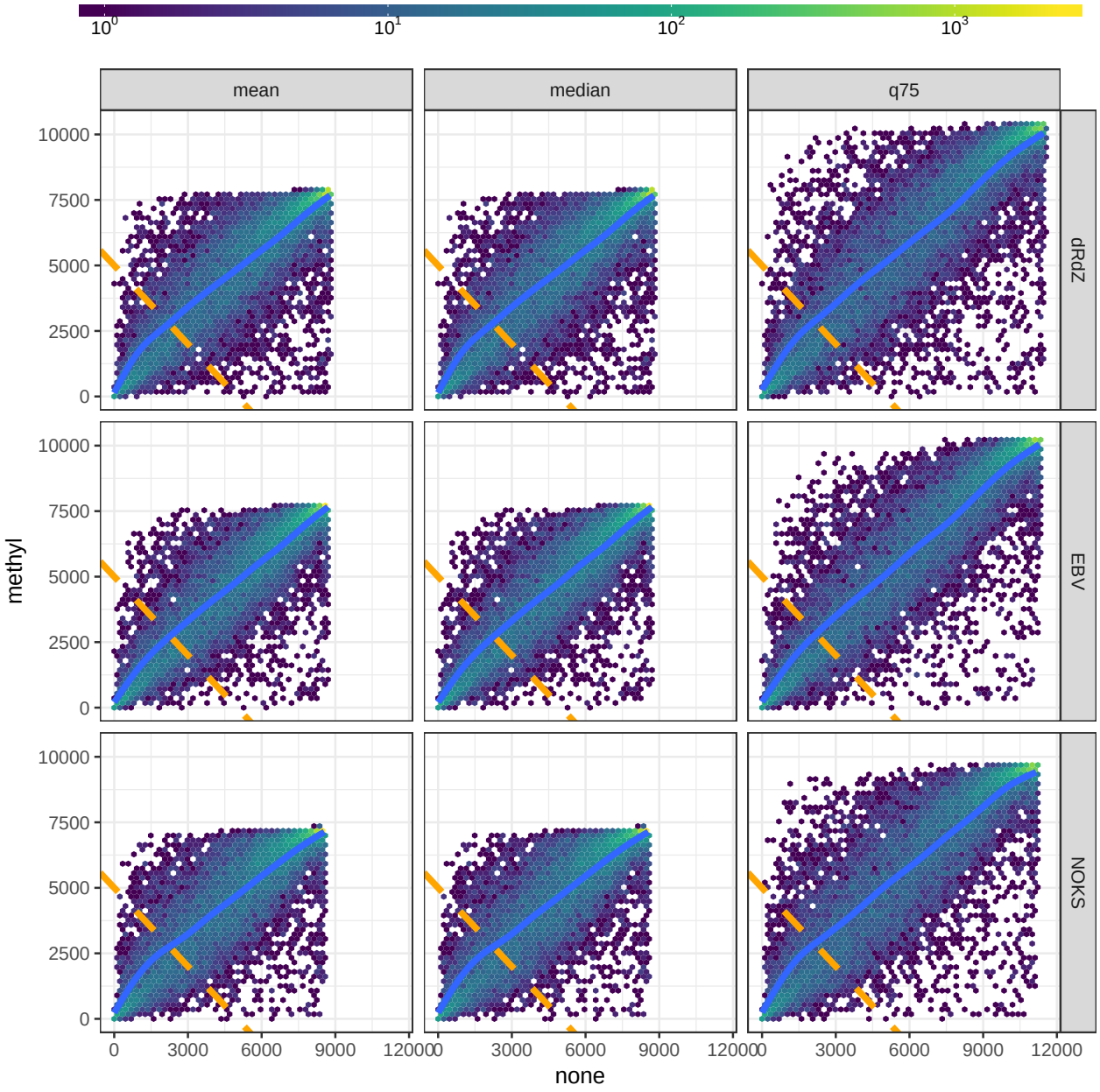
Objective

The idea of this document is to draft what we need to do to integrate the GSEA pathway analysis with the diff. expression analysis we did with DESeq2. We propose the following pipeline:

1. We fit a model with a formula of the type `Count_ij ~ Treat_i + Cell_j + Interaction_ij`
2. For every gene, test an hypothesis if there is a treatment effect in the cell specific expression.
3. For every gene, summarize the TPM level of each treatment, cell and interaction.
4. For every cell, take the genes that are differentially expressed and are among the top K most expressed genes for both treatments (the assumption here is to avoid cases where the `log2FC` value is extreme due to a very low quantity of reads in either treatment).
5. Using the t.stat values as a signal-2-noise metric, do a pathway analysis with GSEA.

TPM summary

We ranked three summary metrics by its decreasing order: `mean`, `median` and the third quartile `q75`. In the figure below, we can see that regardless of the metric we pick, we are going to observe several genes with a low amount of reads without the treatment but well expressed with treatment. There is a clear trend in the TPM levels for both treatment, and the gene surrounding the blue line are the more likely to be differentially expressed (not exactly a neighborhood around the line but the data cloud surrounding it).



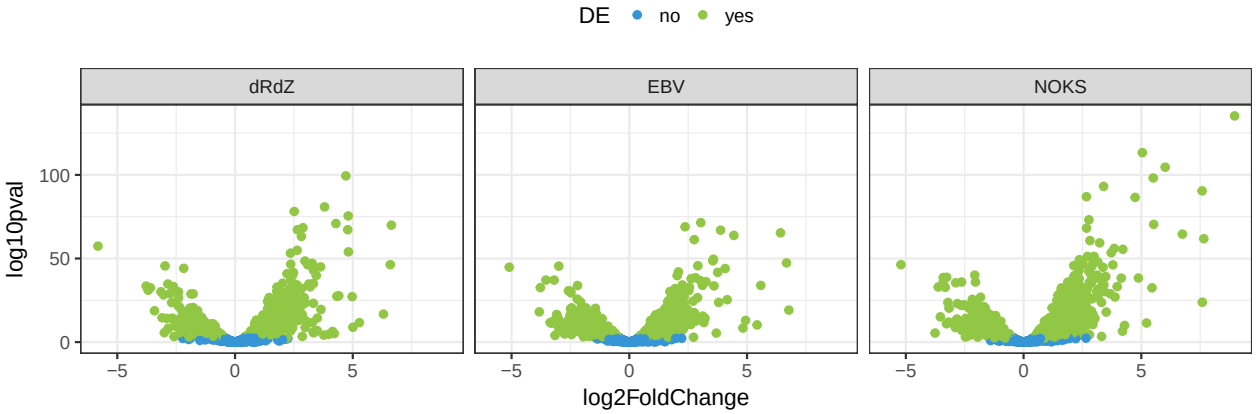
Two ideas:

a. Take the genes that are below the orange line. Those are the genes with the highest TPM, and we can expect for those differentially expressed genes to be part of the cell footprint.

b. Take the genes in the data cloud around the blue line. The genes that are differentially expressed are not going to ocurr due to the read imbalance in the treatment (i.e. the bias we had been dealing with), but the genes that are on the top of the line are more likely to not be important, as those are genes with the lowest TPM.

I am going to focus on **a** as it seems more conservative and try the mean ranking, as the median appears to shrink everything and the quantile seems to behave similarly to the mean.

Differentially expresed genes under this filter



The plot above is a bit missleading because, even though we are filtering tons of genes it is not noticeable in the plot. So, we have in total:

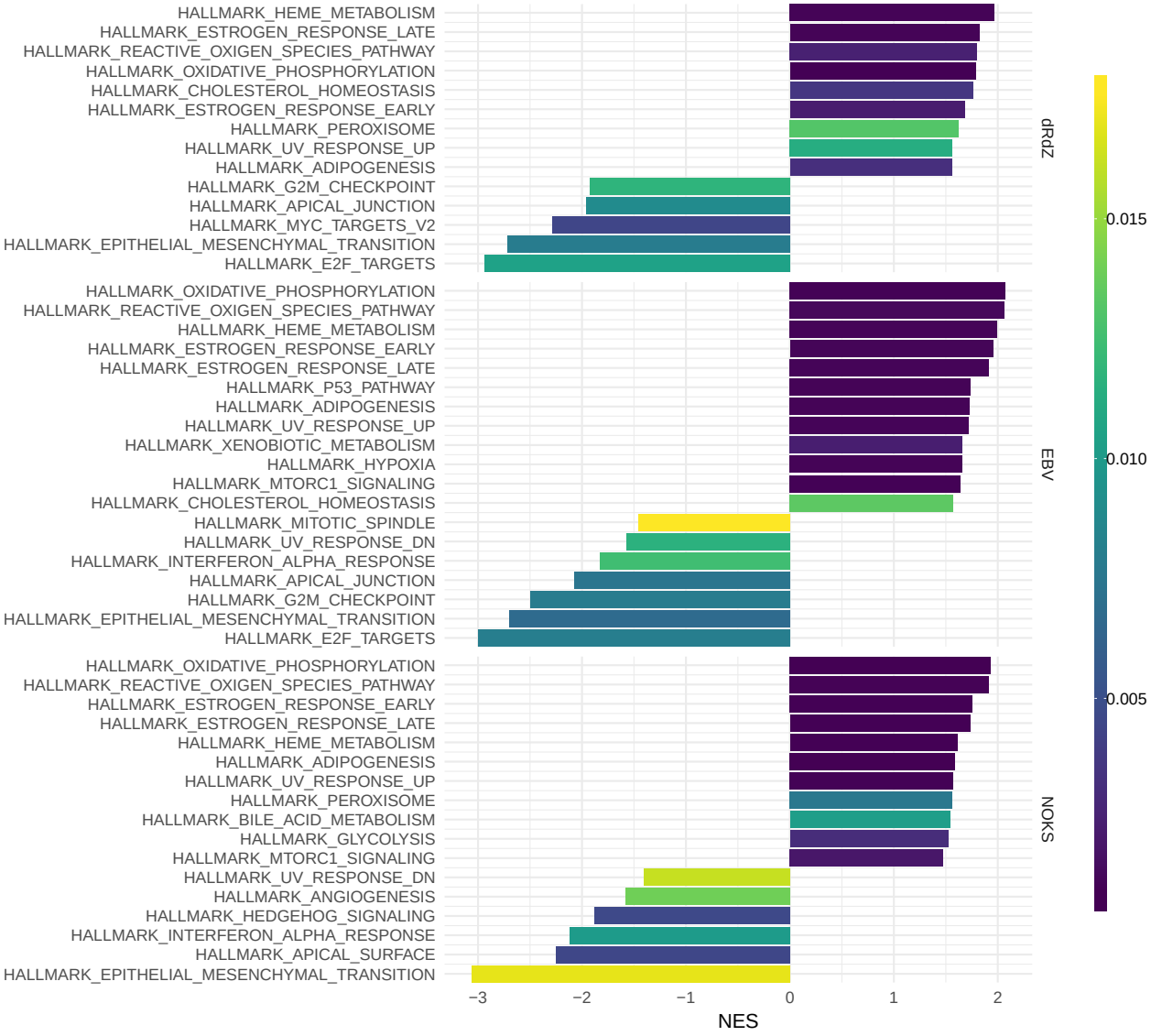
```
## # A tibble: 3 x 3
##   cell     no   yes
##   <chr> <int> <int>
## 1 dRdZ   3095  1772
## 2 EBV    3384  1567
## 3 NOKS   2695  2322
```

GSEA analysis

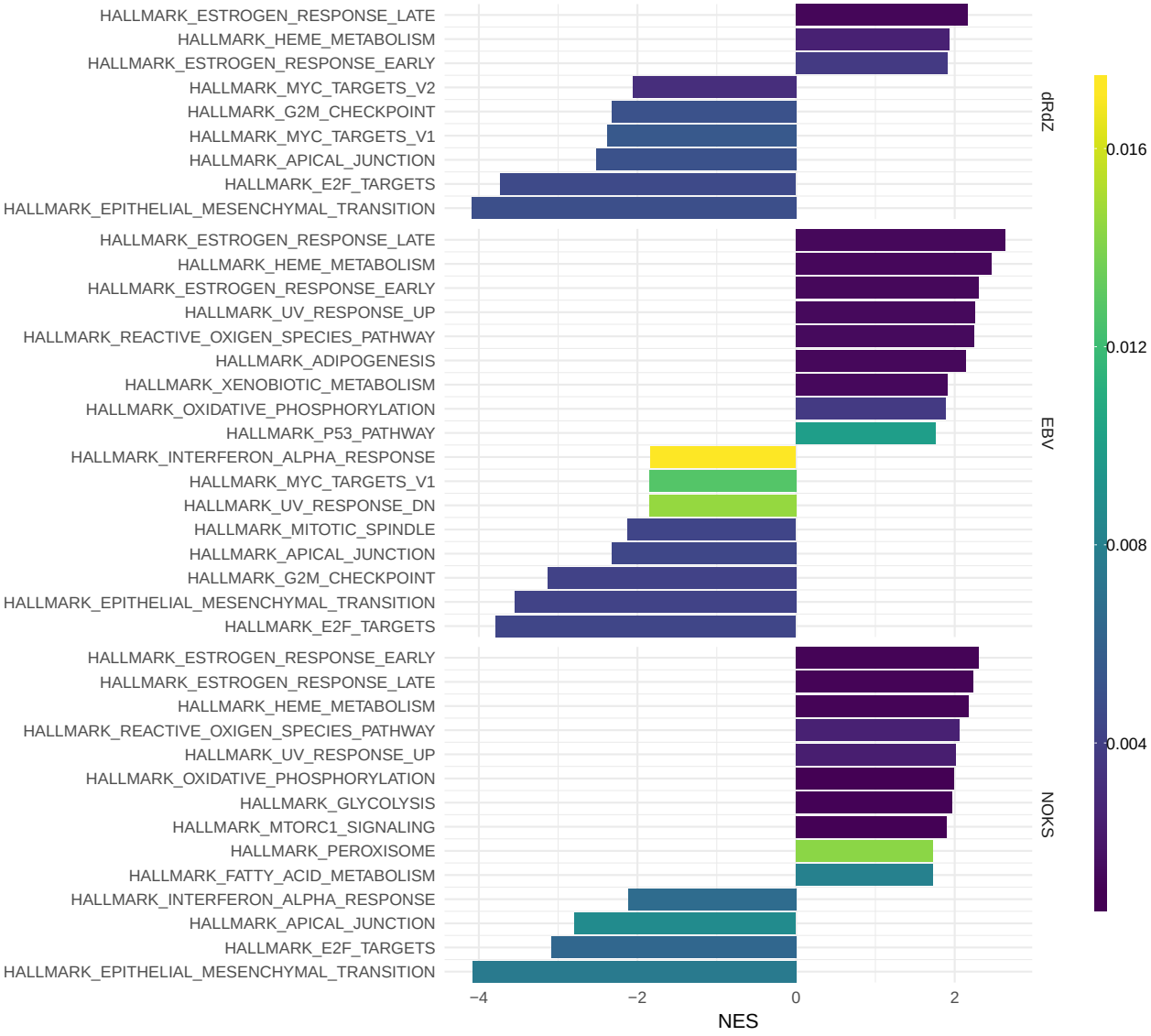
There are two alternatives to do the GSEA analysis:

i. Use all the genes in the list
ii. Use only the differentially expressed genes

All genes


HALLMARK_HEME_METABOLISM
HALLMARK_ESTROGEN_RESPONSE_LATE
HALLMARK_REACTIVE_OXIGEN_SPECIES_PATHWAY
HALLMARK_OXIDATIVE_PHOSPHORYLATION
HALLMARK_CHOLESTEROL_HOMEOSTASIS
HALLMARK_ESTROGEN_RESPONSE_EARLY
HALLMARK_PEROXISOME
HALLMARK_UV_RESPONSE_UP
HALLMARK_ADIPOGENESIS
HALLMARK_G2M_CHECKPOINT
HALLMARK_APICAL_JUNCTION
HALLMARK_MYC_TARGETS_V2
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION
HALLMARK_E2F_TARGETS
dRdZ
HALLMARK_OXIDATIVE_PHOSPHORYLATION
HALLMARK_REACTIVE_OXIGEN_SPECIES_PATHWAY
HALLMARK_HEME_METABOLISM
HALLMARK_ESTROGEN_RESPONSE_EARLY
HALLMARK_ESTROGEN_RESPONSE_LATE
HALLMARK_P53_PATHWAY
HALLMARK_ADIPOGENESIS
HALLMARK_UV_RESPONSE_UP
HALLMARK_XENOBIOTIC_METABOLISM
HALLMARK_HYPOXIA
HALLMARK_MTORC1_SIGNALING
HALLMARK_CHOLESTEROL_HOMEOSTASIS
HALLMARK_MITOTIC_SPINDLE
HALLMARK_UV_RESPONSE_DN
HALLMARK_INTERFERON_ALPHA_RESPONSE
HALLMARK_APICAL_JUNCTION
HALLMARK_G2M_CHECKPOINT
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION
HALLMARK_E2F_TARGETS
EBV
HALLMARK_OXIDATIVE_PHOSPHORYLATION
HALLMARK_REACTIVE_OXIGEN_SPECIES_PATHWAY
HALLMARK_ESTROGEN_RESPONSE_EARLY
HALLMARK_ESTROGEN_RESPONSE_LATE
HALLMARK_HEME_METABOLISM
HALLMARK_ADIPOGENESIS
HALLMARK_UV_RESPONSE_UP
HALLMARK_PEROXISOME
HALLMARK_BILE_ACID_METABOLISM
HALLMARK_GLYCOLYSIS
HALLMARK_MTORC1_SIGNALING
HALLMARK_UV_RESPONSE_DN
HALLMARK_ANGIOGENESIS
HALLMARK_HEDGEHOG_SIGNALING
HALLMARK_INTERFERON_ALPHA_RESPONSE
HALLMARK_APICAL_SURFACE
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION
NOKS
−3
−2
−1
0
1
2
NES
0.015
0.010
0.005

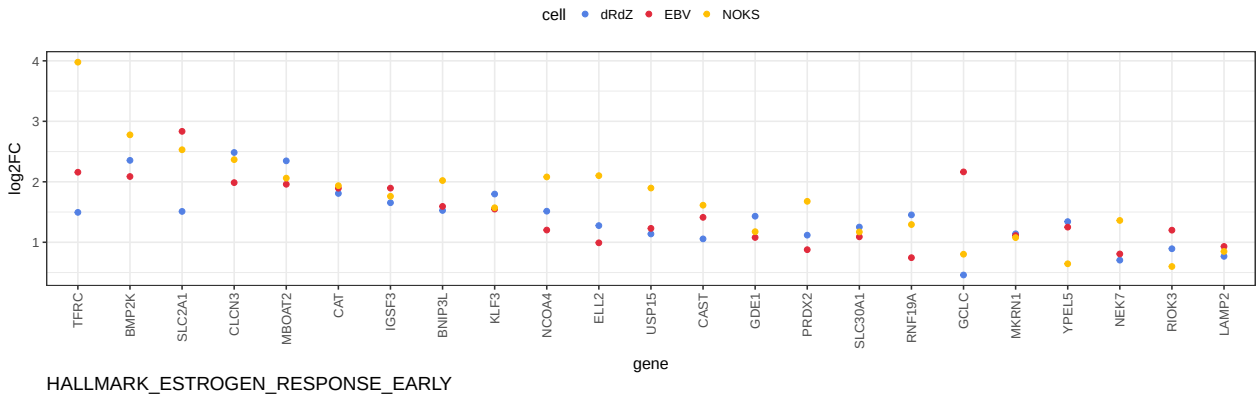
Only differentially expressed genes



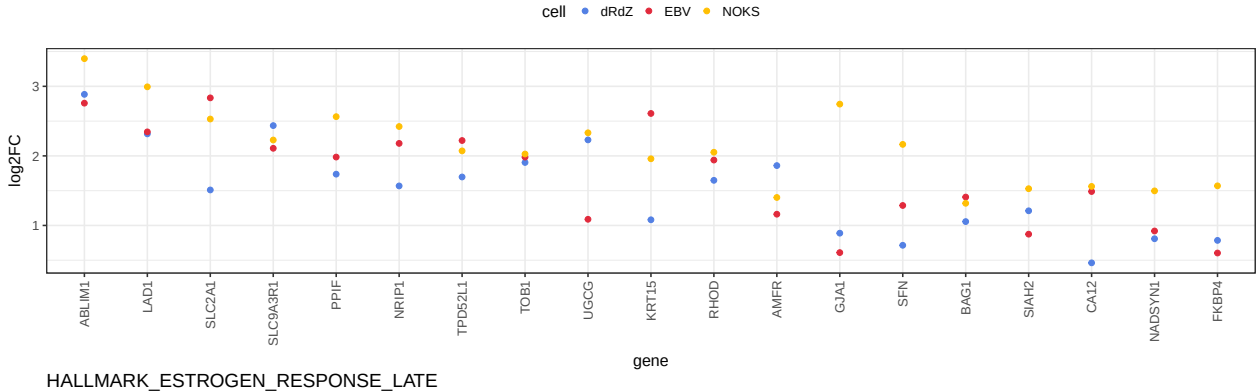
Further exploration

A lot of the pathways coincide, hence I am going to focus on the ones obtained when using the DE genes explicitly.

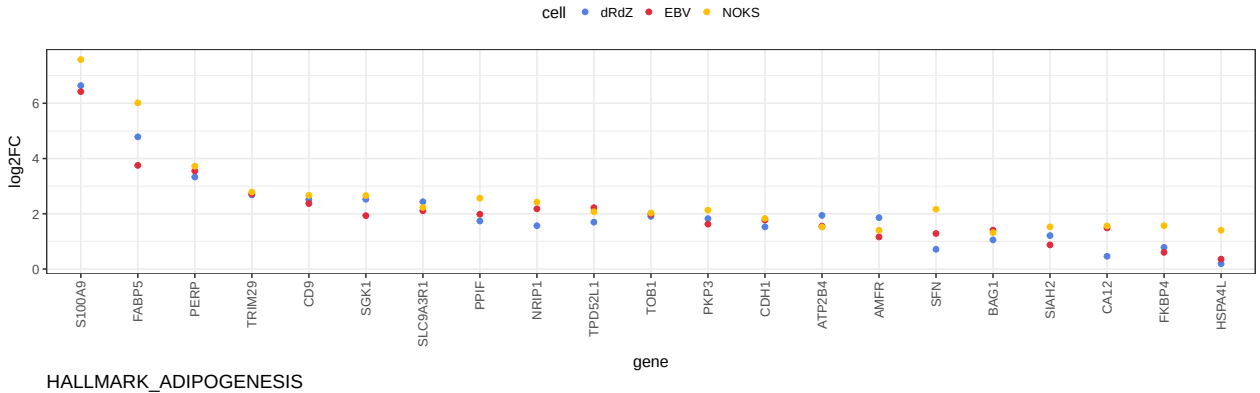
HALLMARK_HEME_METABOLISM



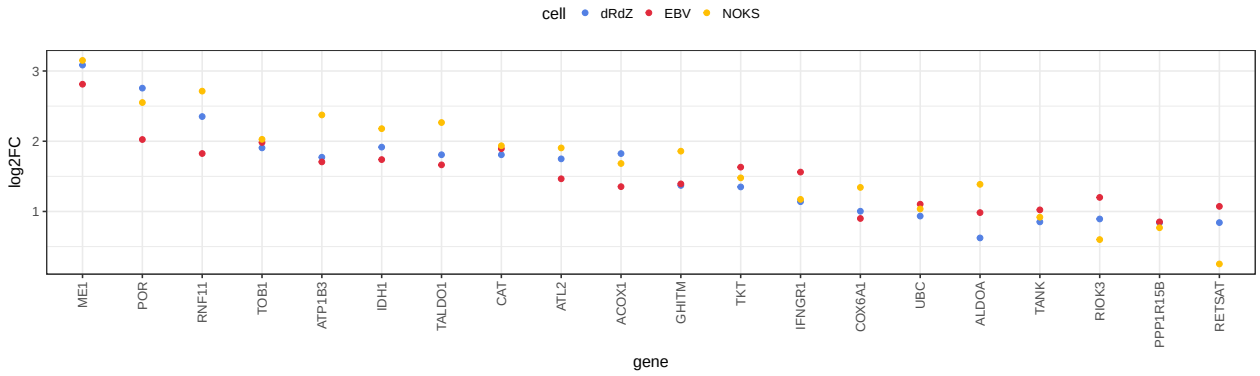
HALLMARK_ESTROGEN_RESPONSE_EARLY



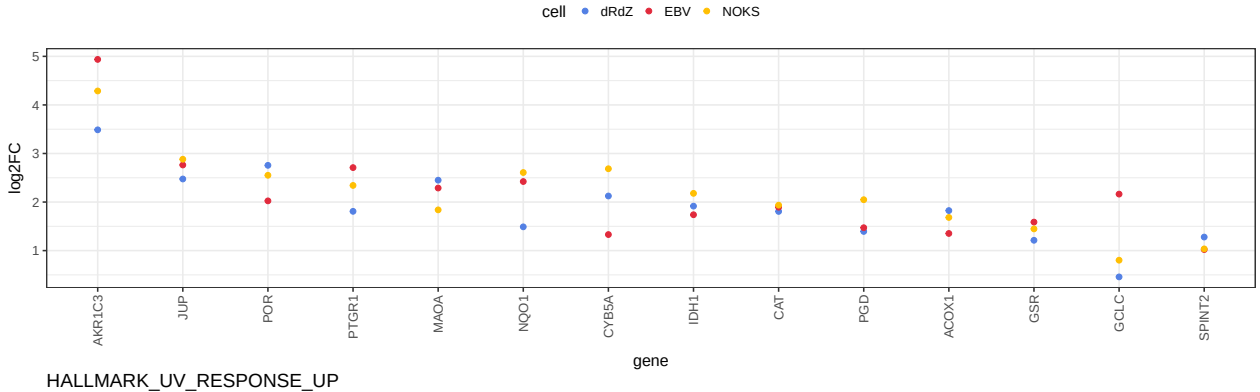
HALLMARK_ESTROGEN_RESPONSE_LATE



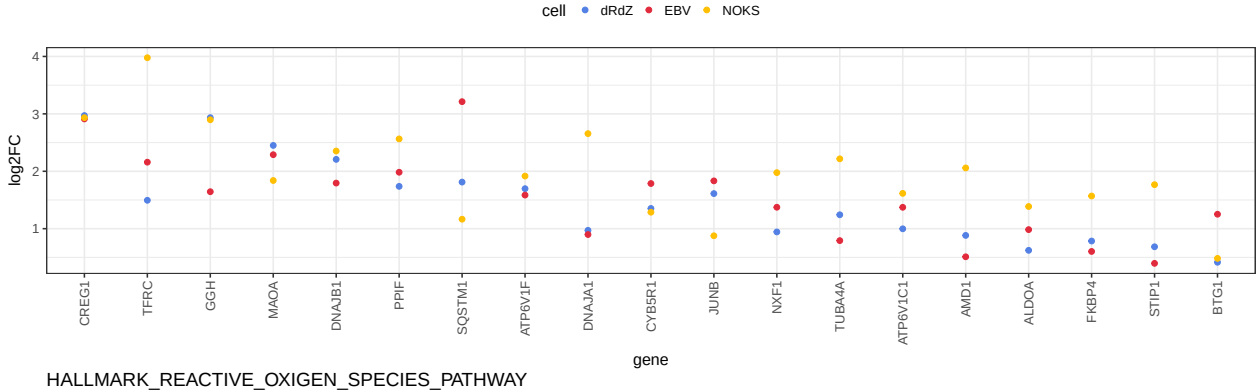
HALLMARK_ADIPOGENESIS



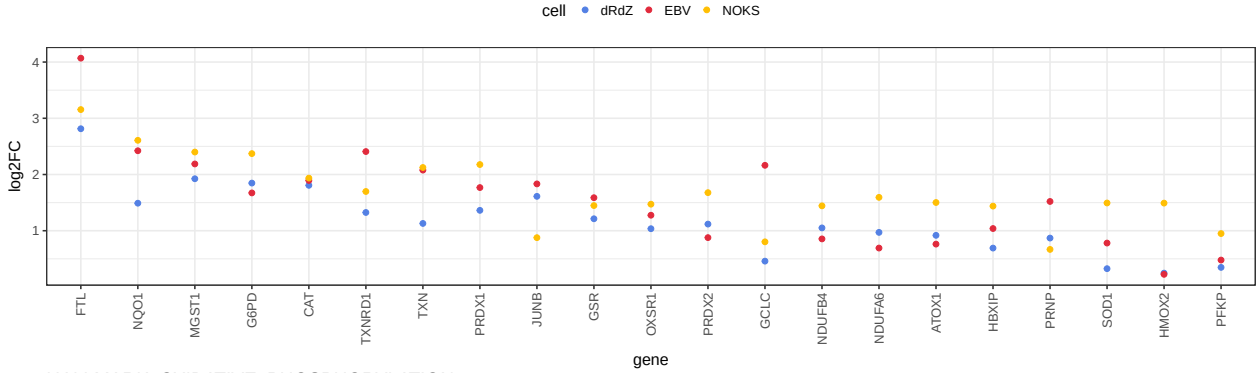
HALLMARK_XENOBIOTIC_METABOLISM



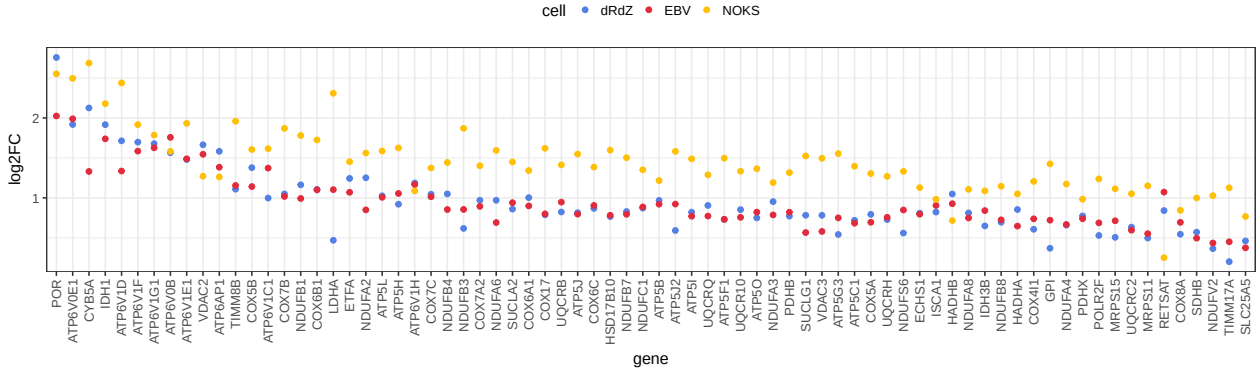
HALLMARK_UV_RESPONSE_UP



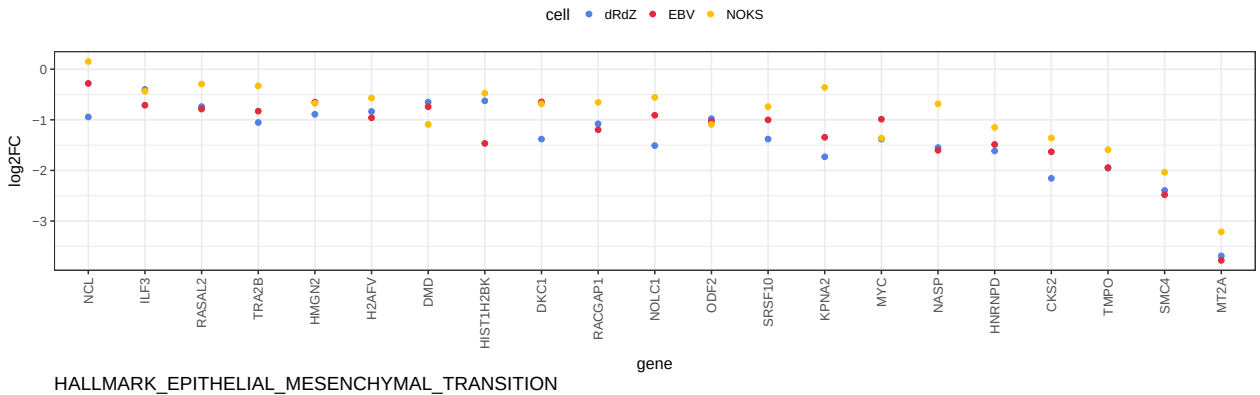
HALLMARK_REACTIVE_OXIGEN_SPECIES_PATHWAY



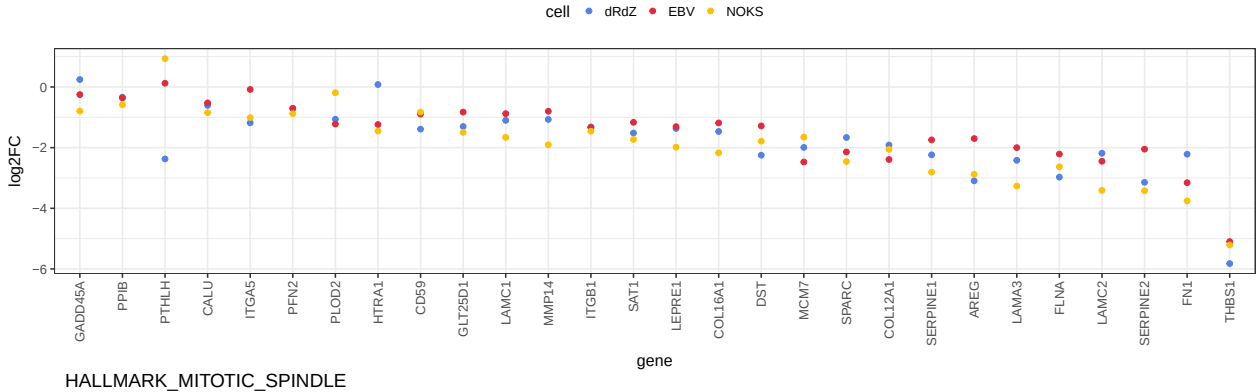
HALLMARK_OXIDATIVE_PHOSPHORYLATION



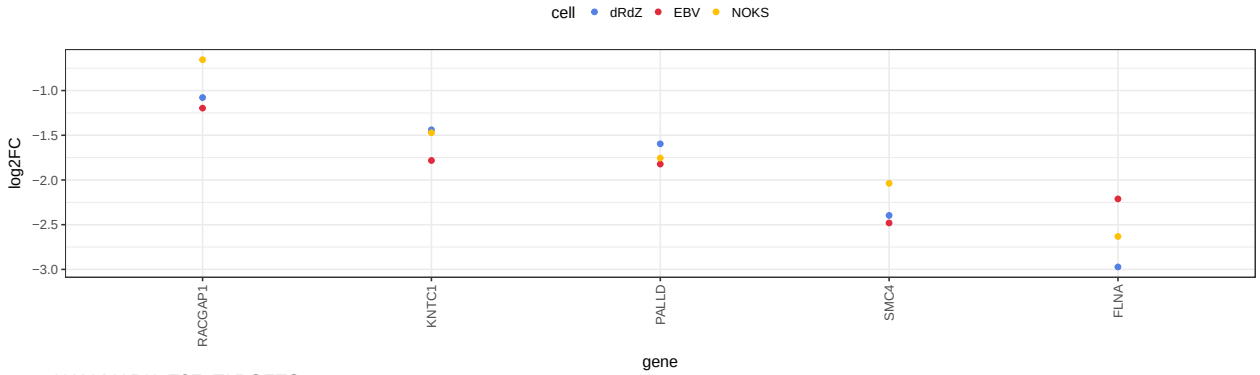
HALLMARK_G2M_CHECKPOINT



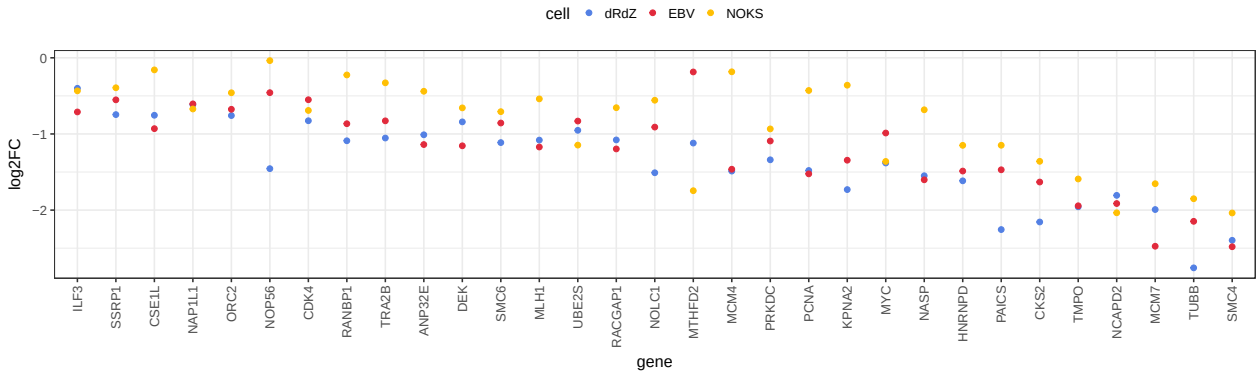
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION



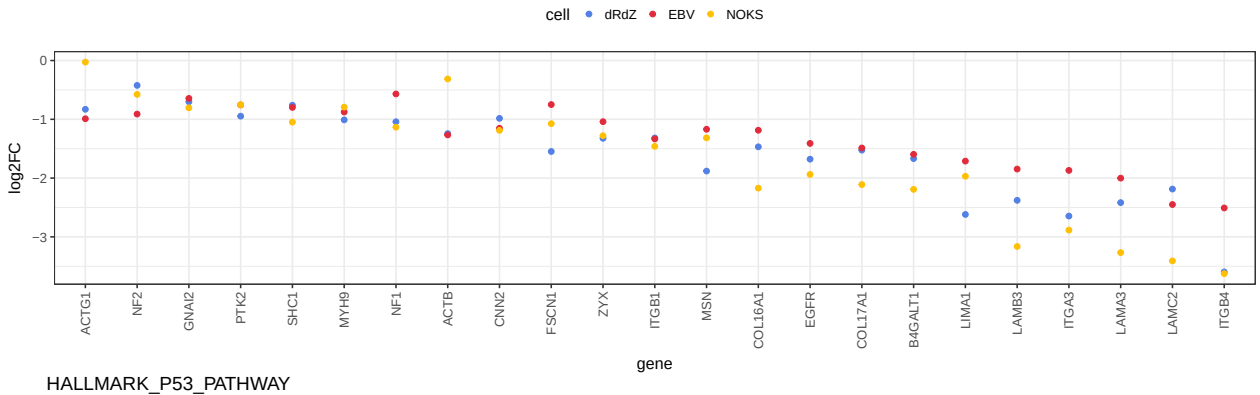
HALLMARK_MITOTIC_SPINDLE



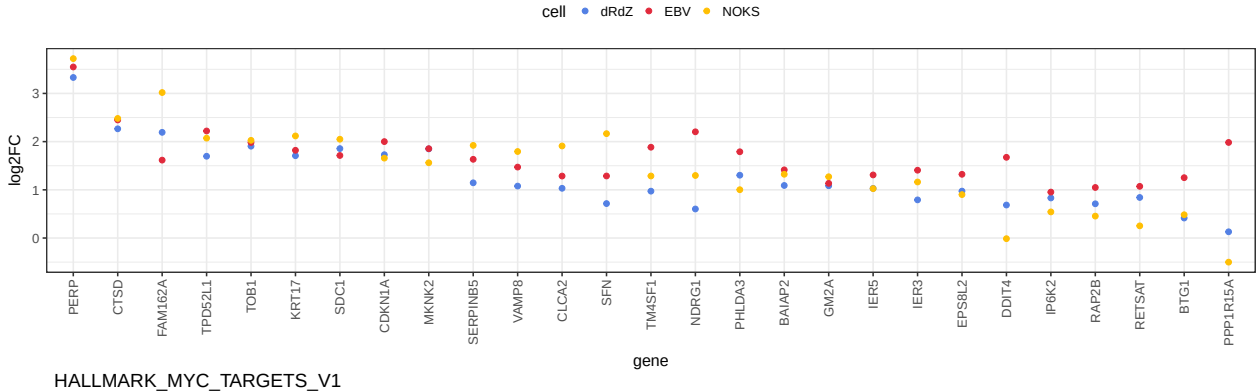
HALLMARK_E2F_TARGETS



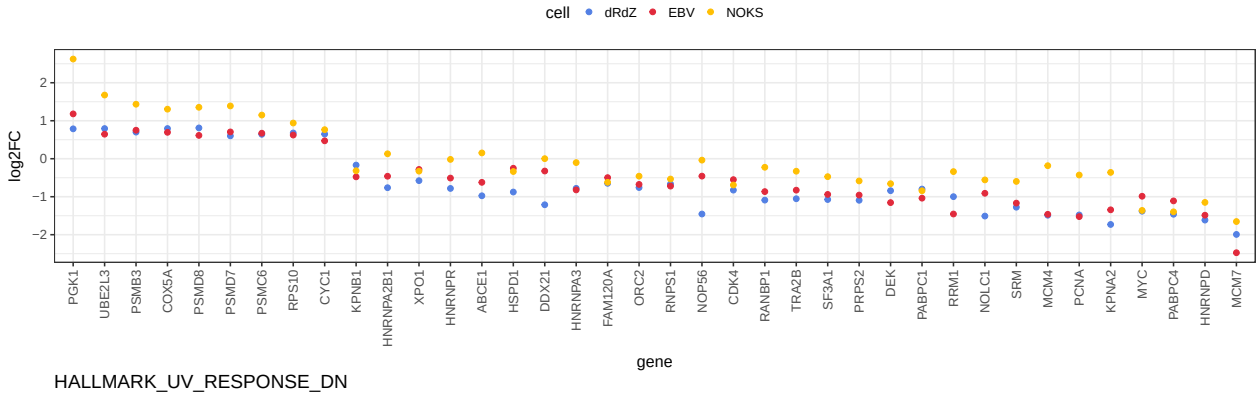
HALLMARK_APICAL_JUNCTION



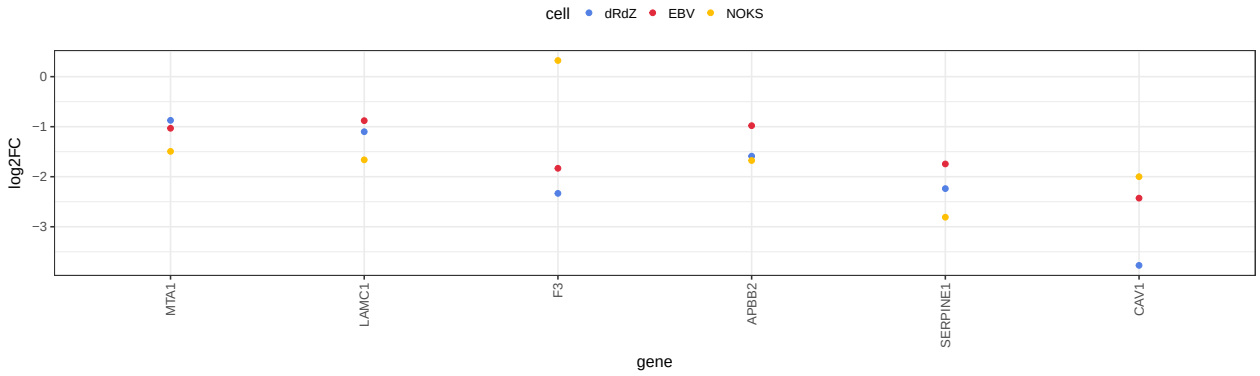
HALLMARK_P53_PATHWAY



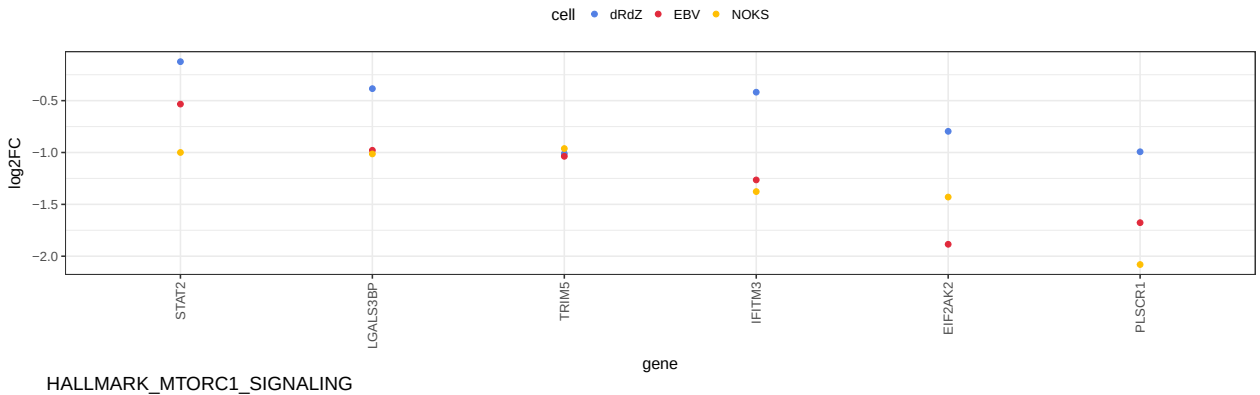
HALLMARK_MYC_TARGETS_V1



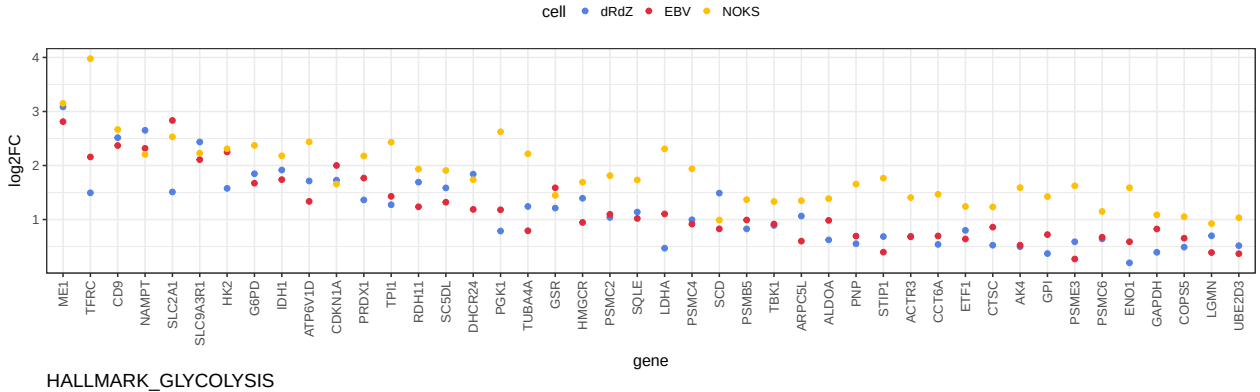
HALLMARK_UV_RESPONSE_DN



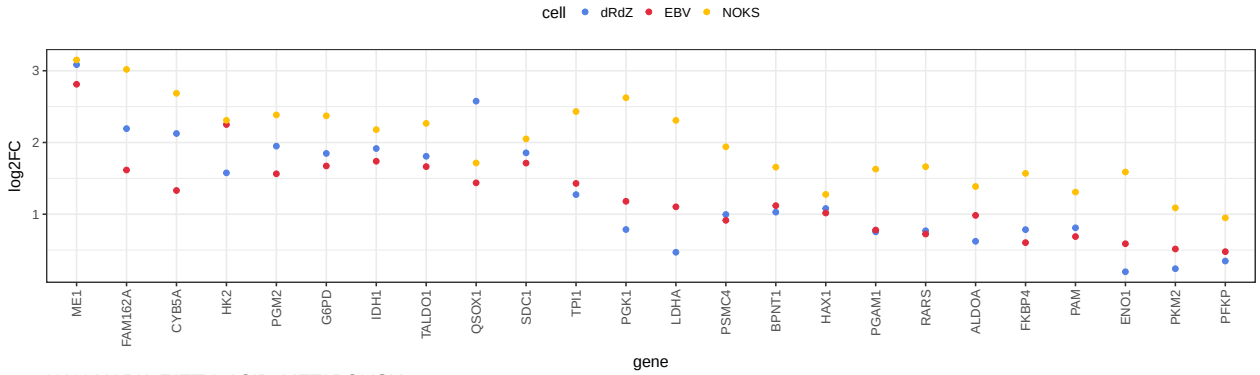
HALLMARK_INTERFERON_ALPHA_RESPONSE



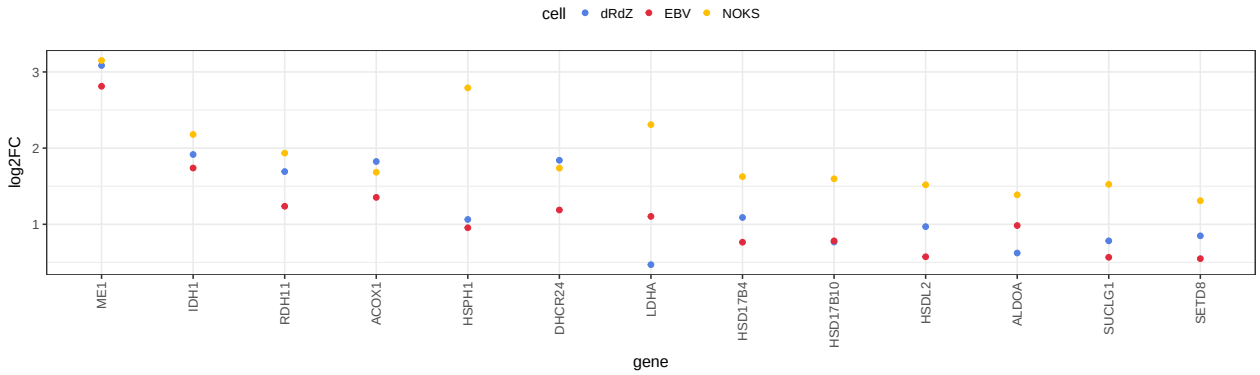
HALLMARK_MTORC1_SIGNALING



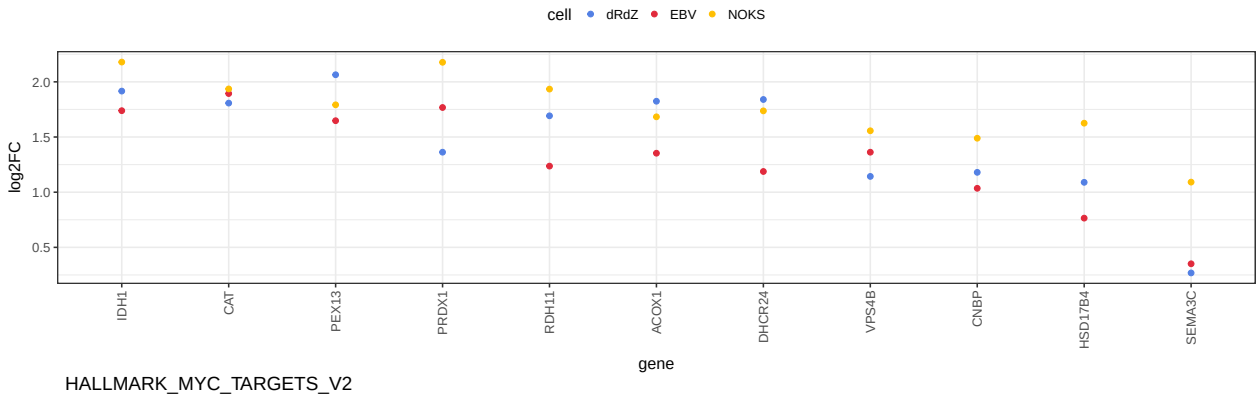
HALLMARK_GLYCOLYSIS



HALLMARK_FATTY_ACID_METABOLISM



HALLMARK_PEROXISOME



HALLMARK_MYC_TARGETS_V2