

Pathway analysis - Fig5

Contents

Objective	1
TPM summary	1
Differentially expressed genes under this filter	3
GSEA analysis	3
All genes	4
Only differentially expressed genes	5
Further exploration of pathways obtained with DE genes	5

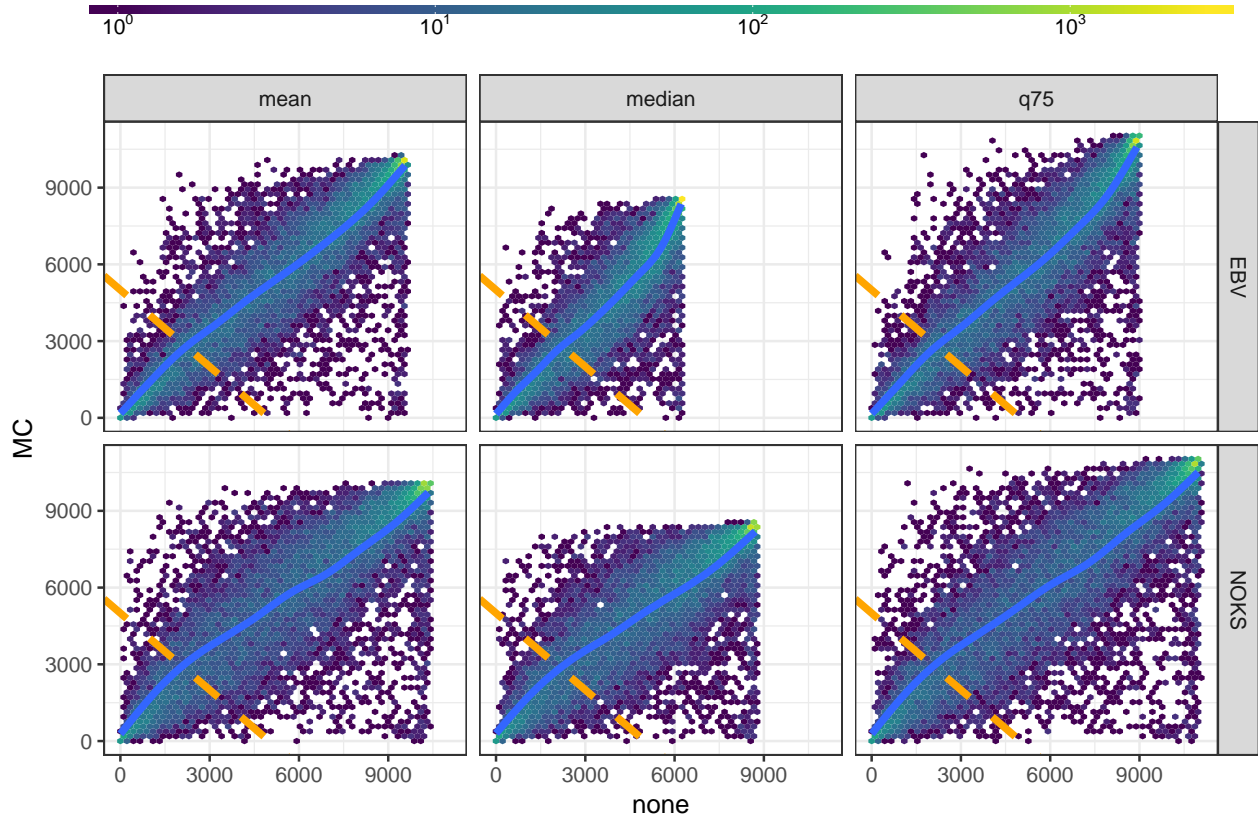
Objective

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).

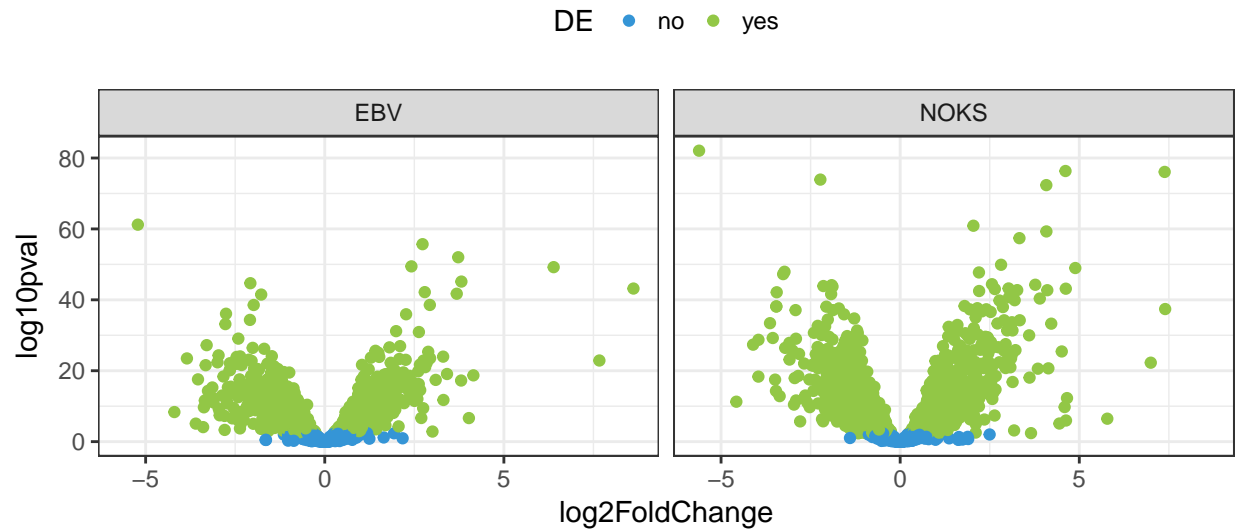


Two ideas:

- 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.
- Take the genes in the data cloud around the blue line. The genes that are differentially expressed are not going to occur 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 expressed genes under this filter



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

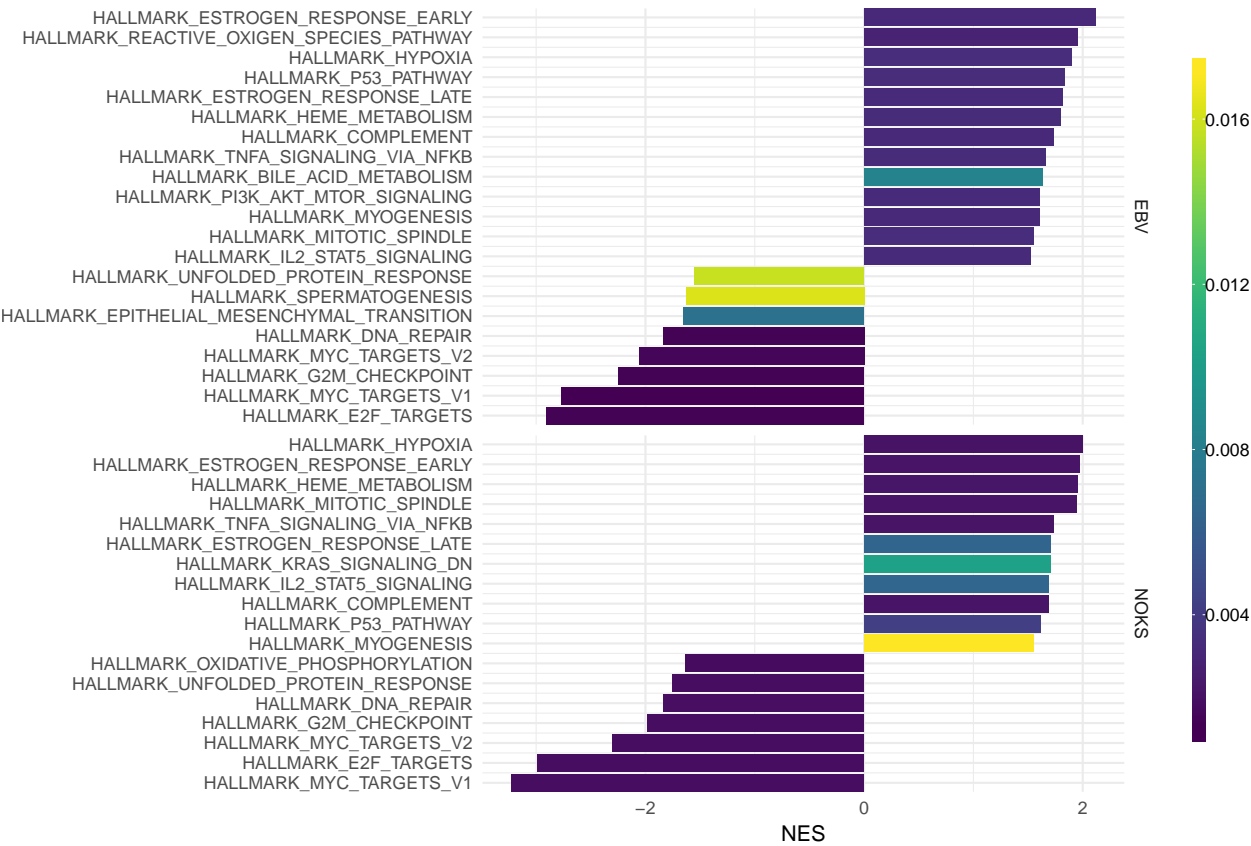
```
## # A tibble: 2 x 3
##   cell    no  yes
##   <chr> <int> <int>
## 1 EBV   2786 1756
## 2 NOKS  1887 2338
```

GSEA analysis

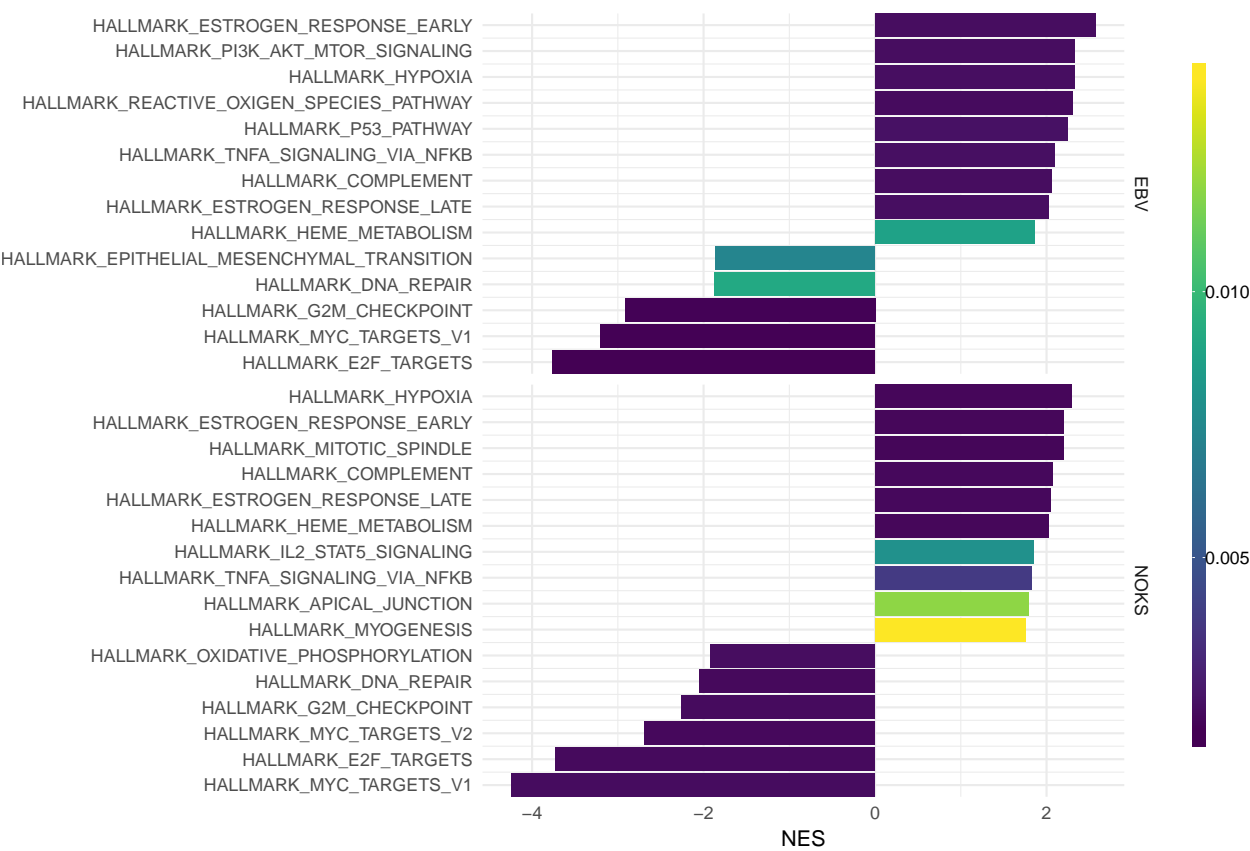
There are two alternatives to do the GSEA analysis:

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

All genes

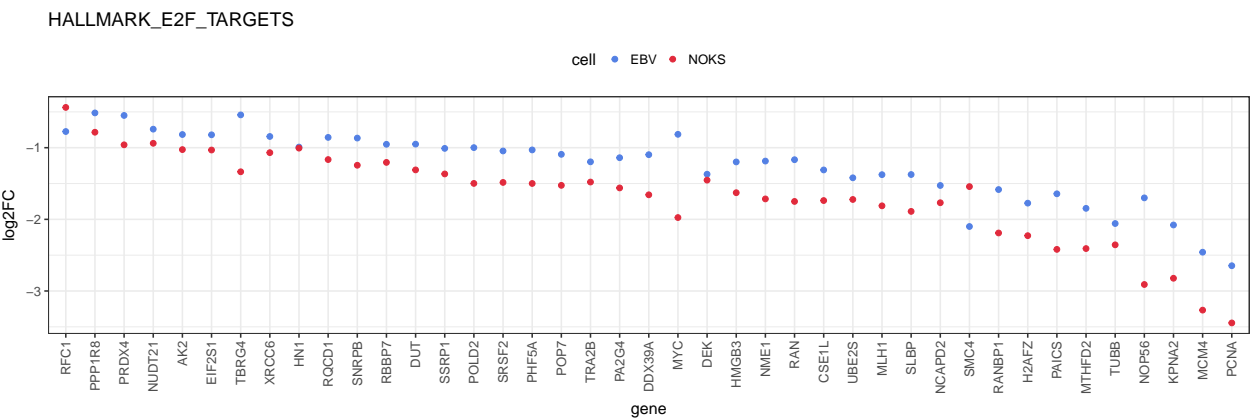


Only differentially expressed genes

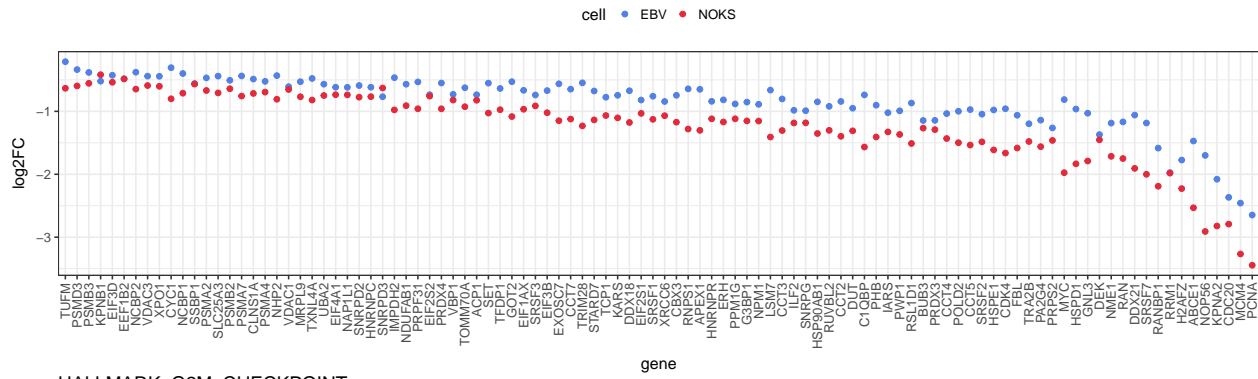


Further exploration of pathways obtained with DE genes

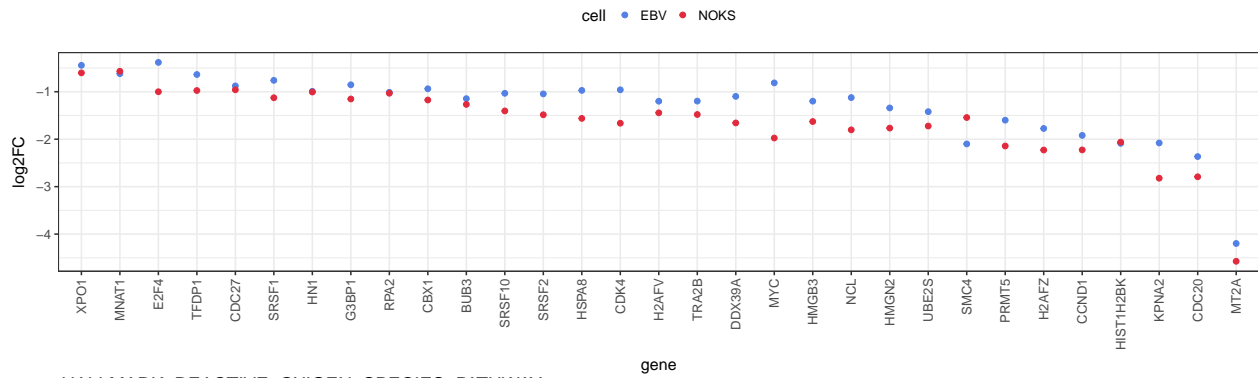
A lot of the pathways coincide, hence I am going to focus on the ones obtained when using the DE genes explicitly.



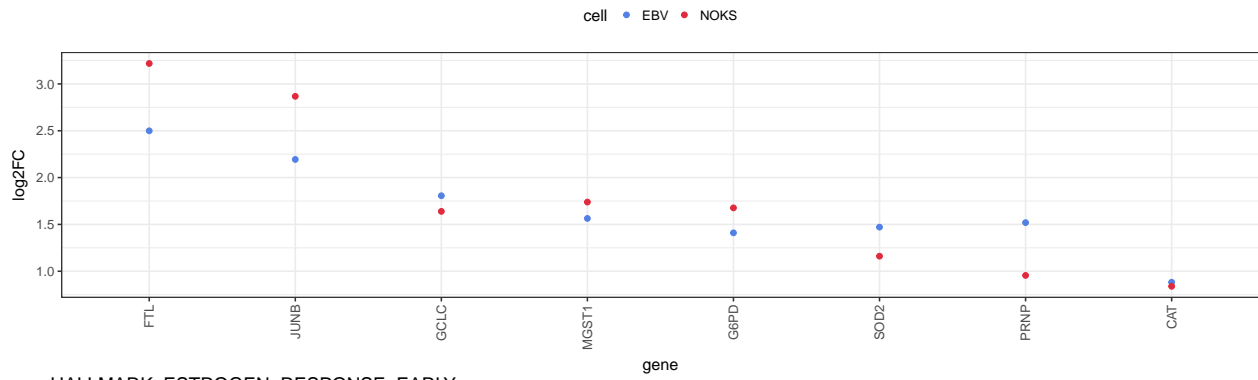
HALLMARK_MYC_TARGETS_V1



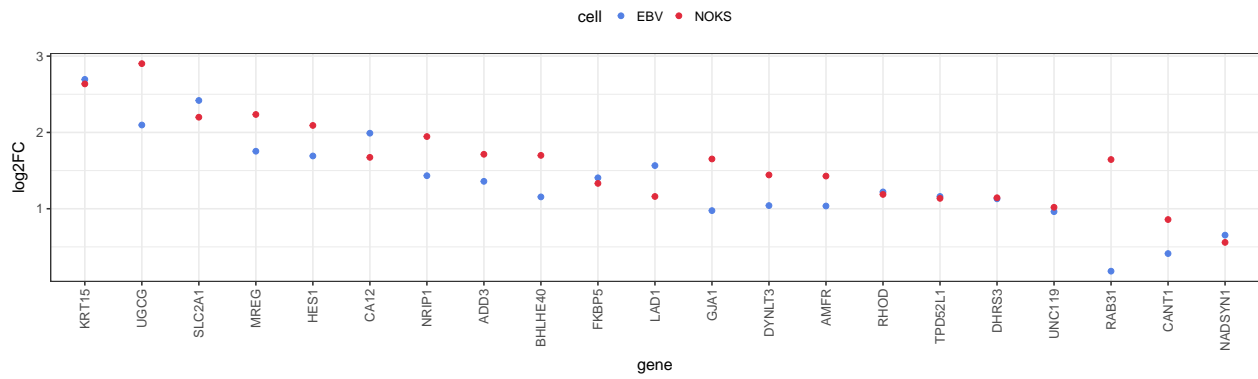
HALLMARK_G2M_CHECKPOINT



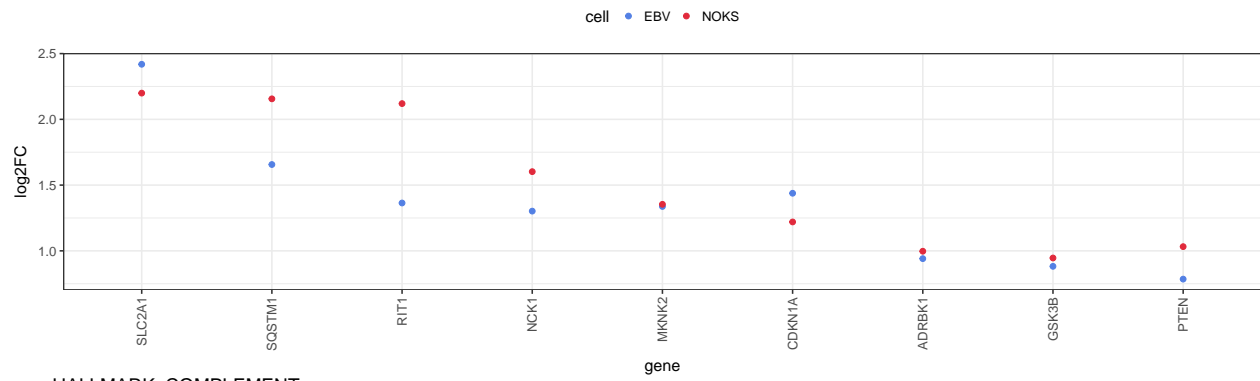
HALLMARK_REACTIVE_OXYGEN_SPECIES_PATHWAY



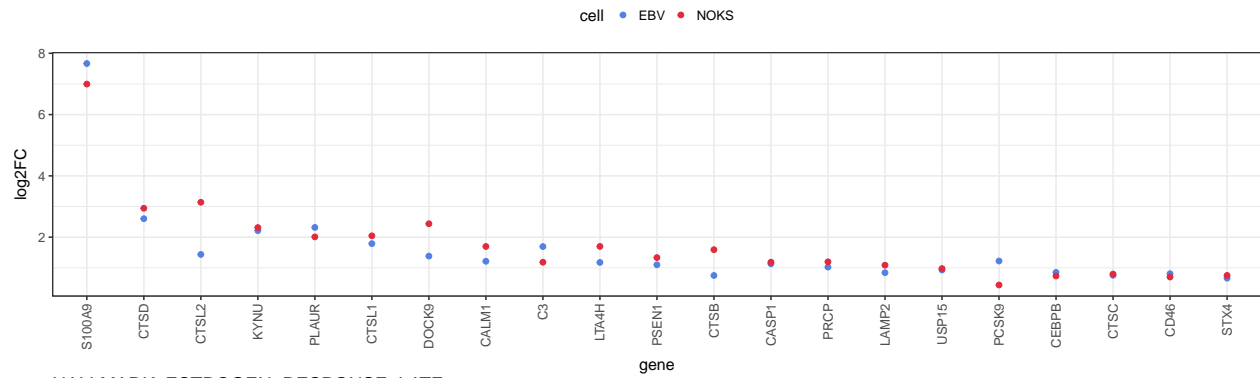
HALLMARK_ESTROGEN_RESPONSE_EARLY



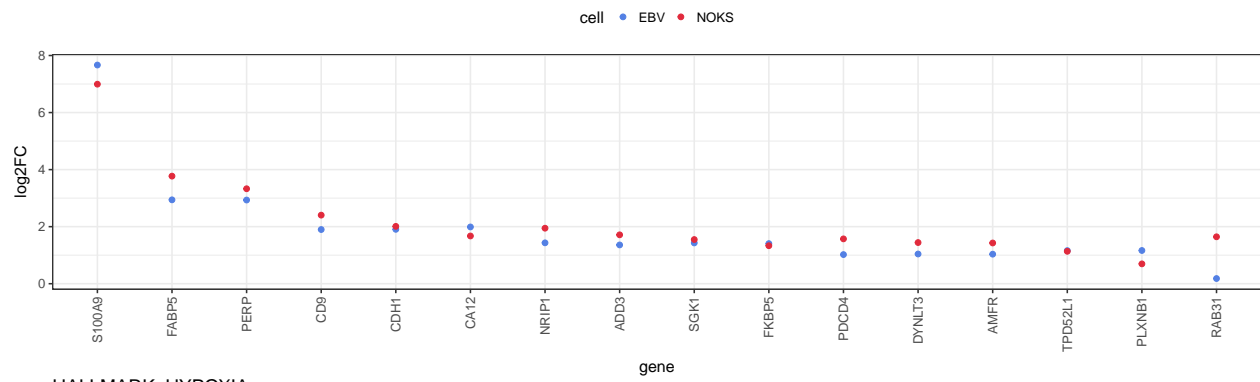
HALLMARK_PI3K_AKT_MTOR_SIGNALING



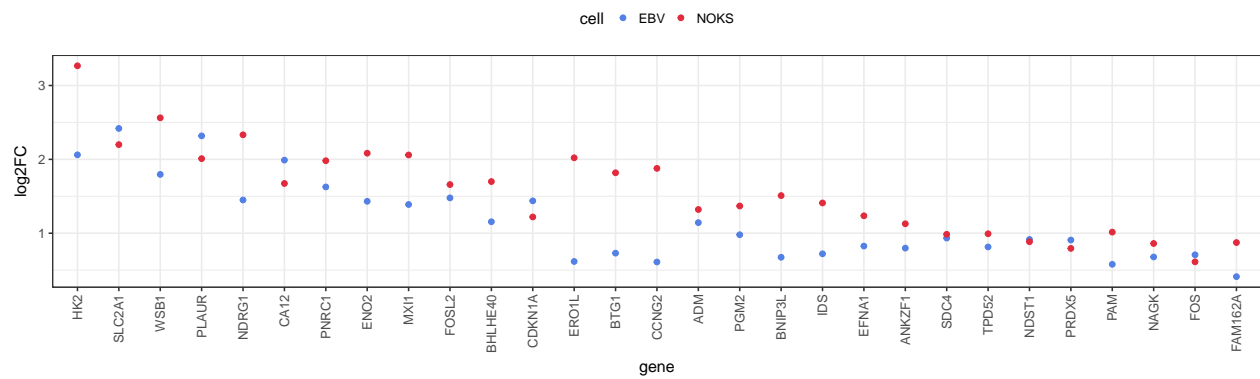
HALLMARK_COMPLEMENT



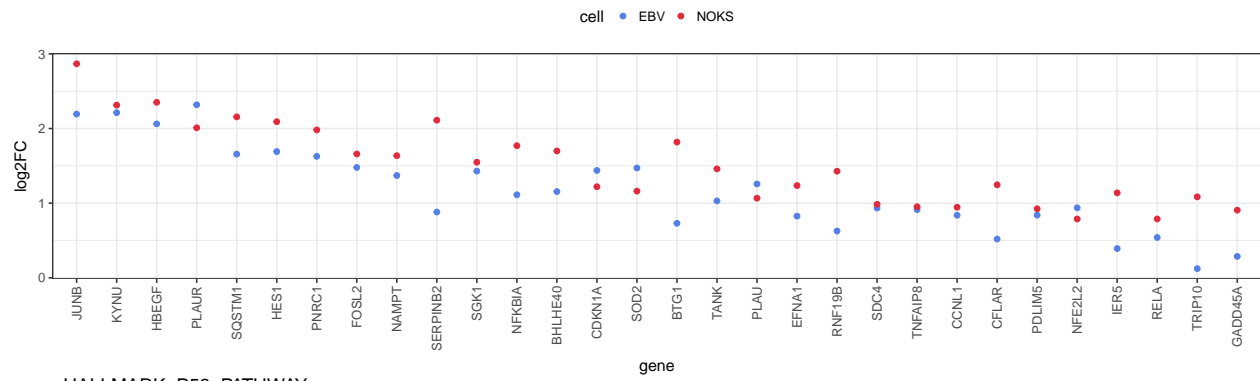
HALLMARK_ESTROGEN_RESPONSE_LATE



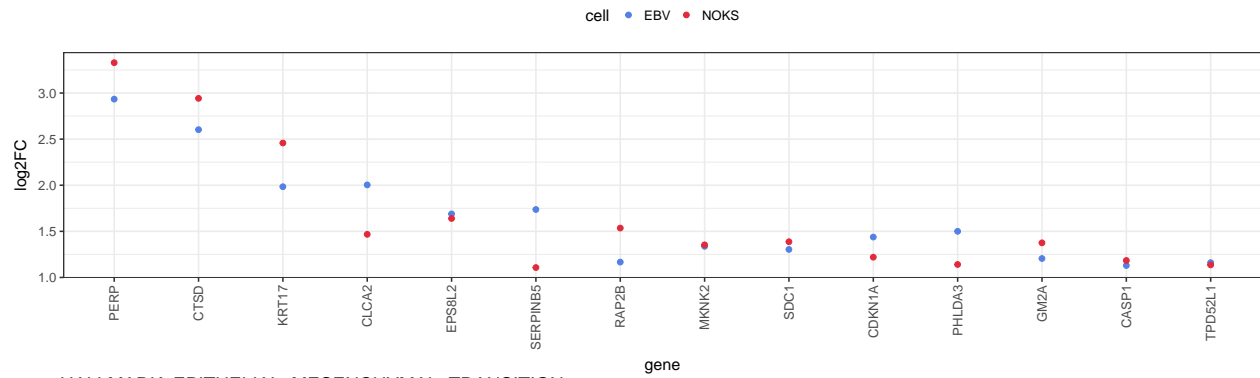
HALLMARK_HYPOXIA



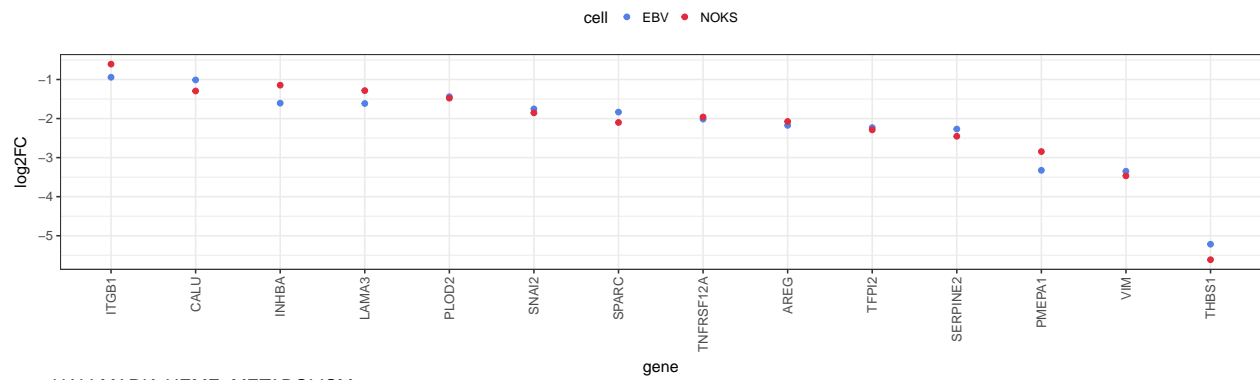
HALLMARK_TNFA_SIGNALING_VIA_NFKB



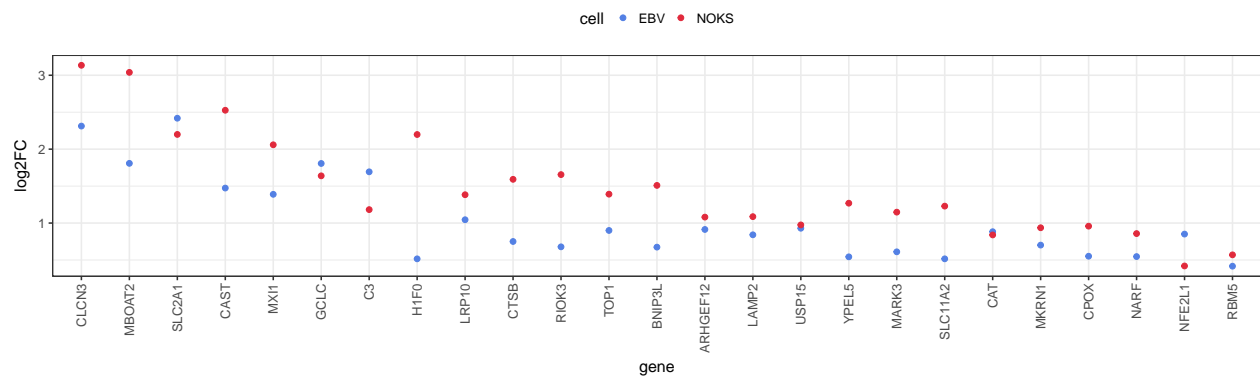
HALLMARK_P53_PATHWAY



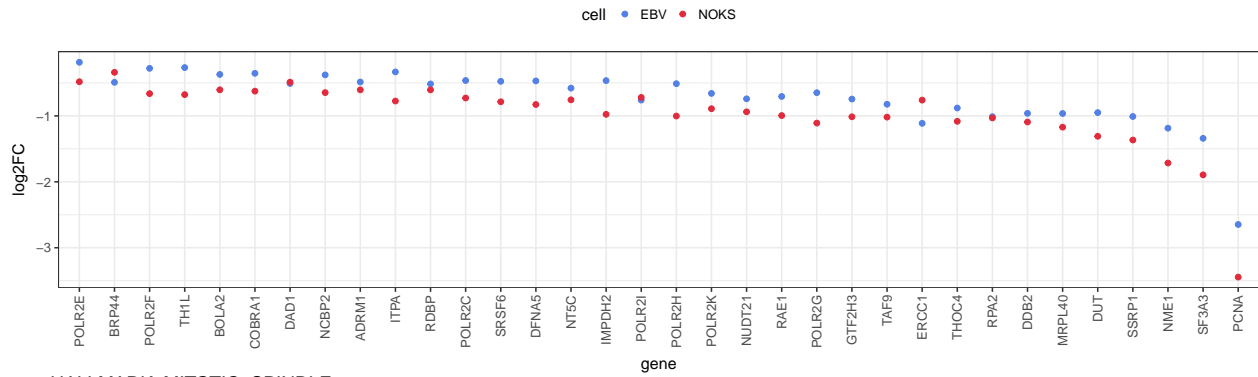
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION



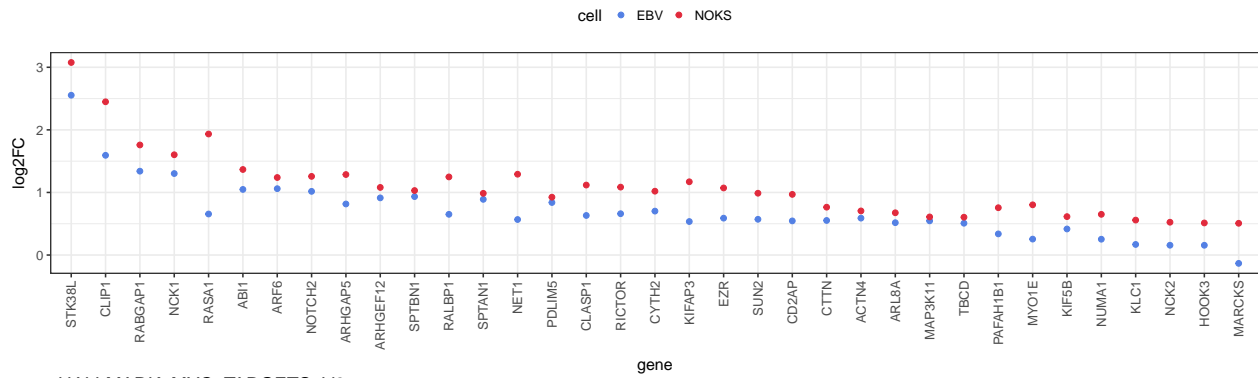
HALLMARK_HEME_METABOLISM



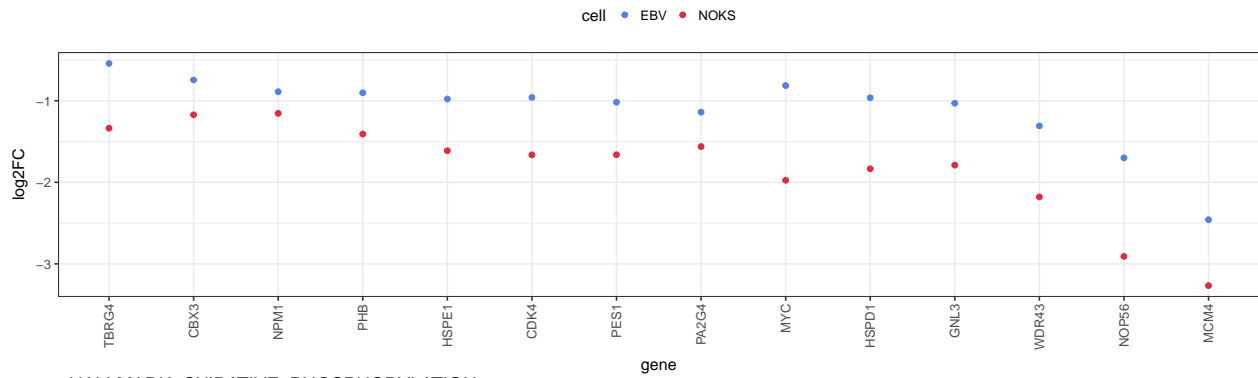
HALLMARK_DNA_REPAIR



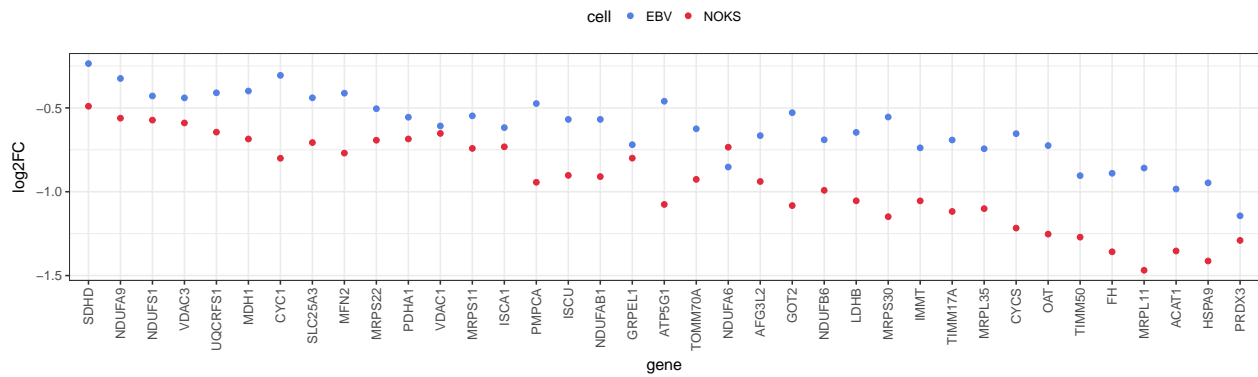
HALLMARK_MITOTIC_SPINDLE



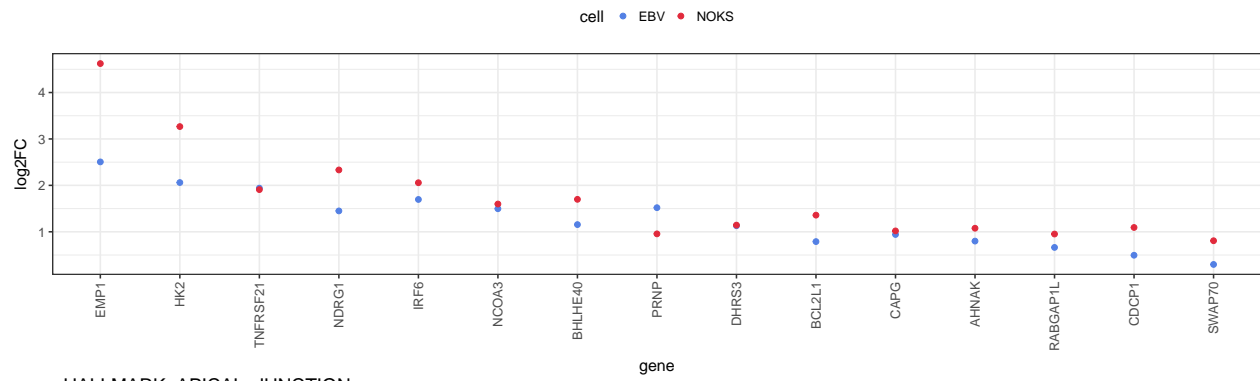
HALLMARK_MYC_TARGETS_V2



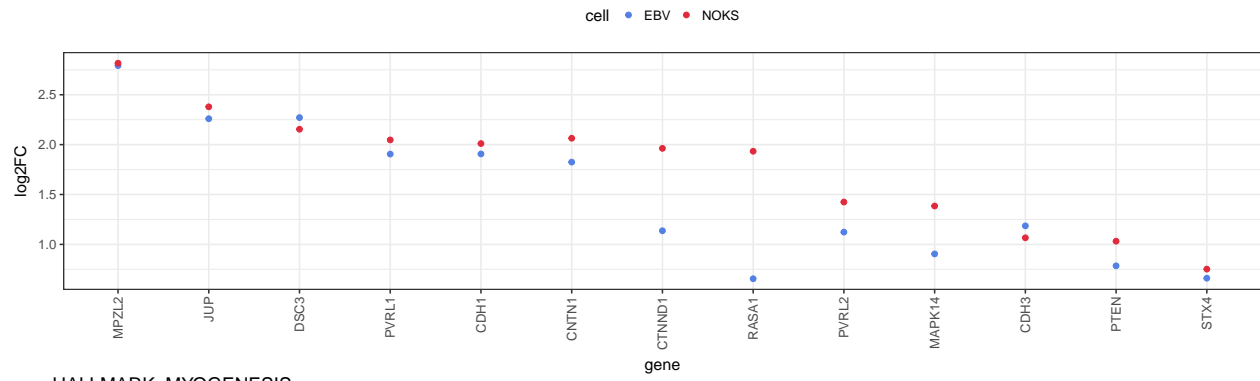
HALLMARK_OXIDATIVE_PHOSPHORYLATION



HALLMARK_IL2_STAT5_SIGNALING



HALLMARK_APICAL_JUNCTION



HALLMARK_MYOGENESIS

