Comparison by treatment

Contents

Intro **1**

Signal to noise analysis **1**

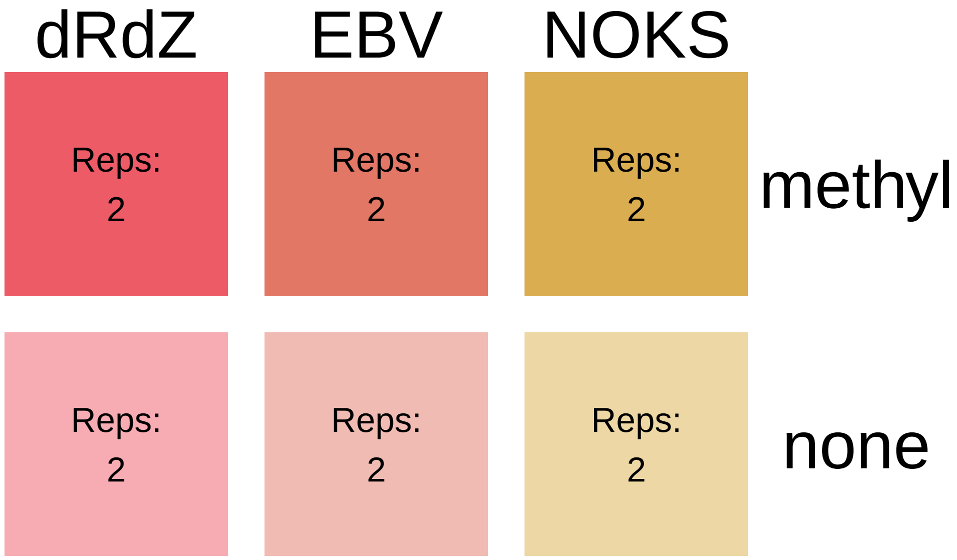
Pathway analysis **7**
GSEA plots for `MYC_TARGETS`: 9

Demo gene coefficients **13**

Intro

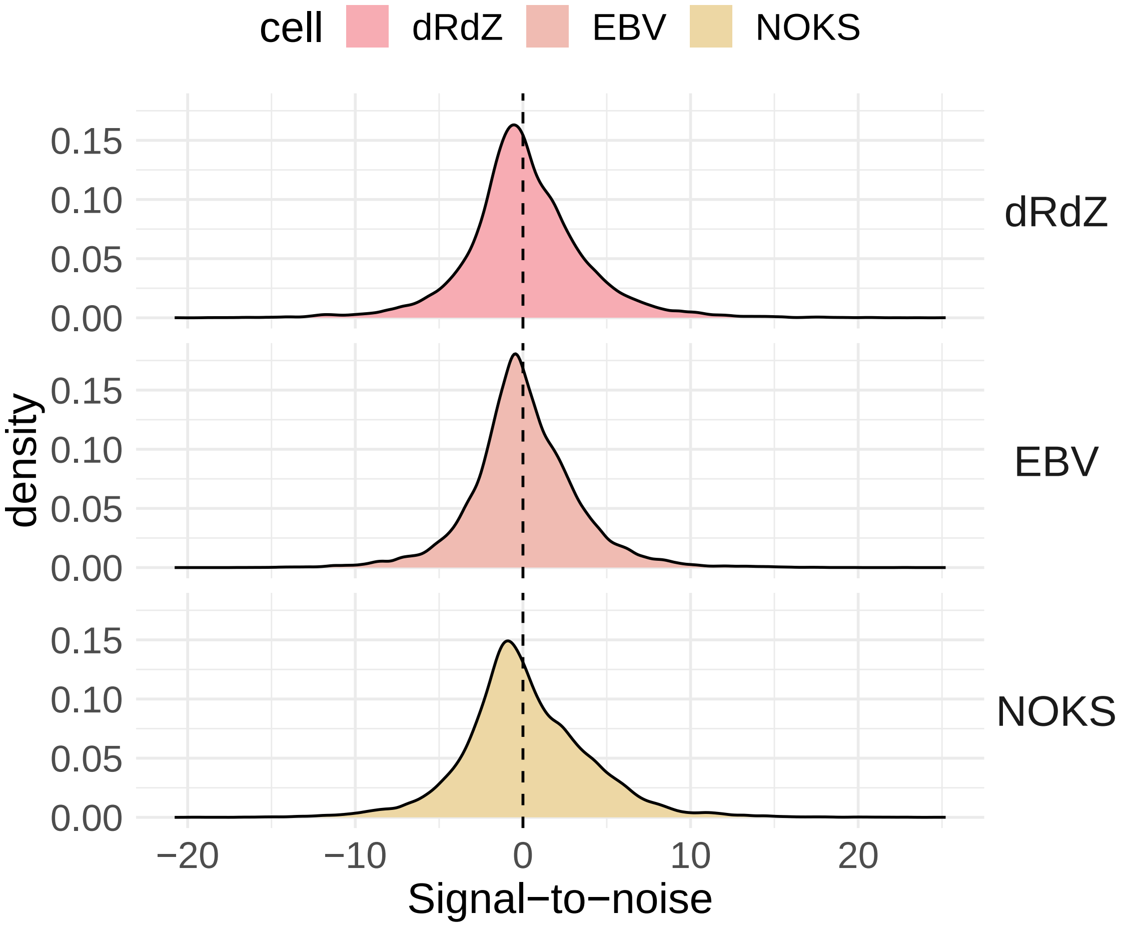
I guess the idea is to find genes such that show high expression under the treatment, but at the same time exhibit low expression without the treatment, i.e. we are going to focus on the genes that are on the Wald's statistic distribution's tails. For that purpose, we are going to:

1. Perform contrast of MC-treated vs untreated samples for each cell line.
2. Compare the Wald's t-statistic between the three cell lines.

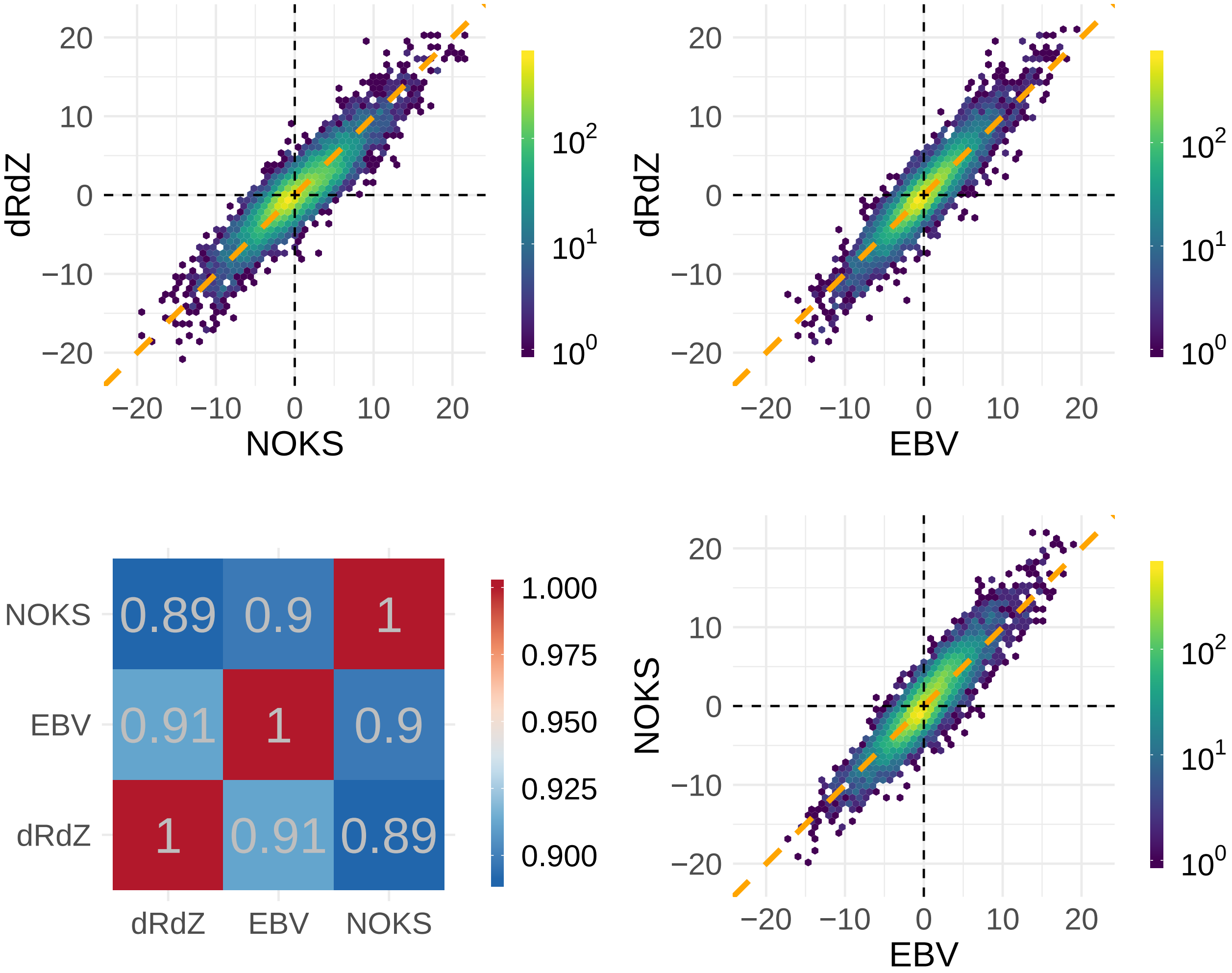


Signal to noise analysis

The figure below shows the densities of the signal-to-noise measures of the treated vs untreated contrast for all three cell lines. Clearly, all three cell lines resemble a similar pattern, with a slightly heavier tails for the NOKS case.



We then compare those three vectors, and we can notice that for most of the genes the signal-to-noise (i.e the log2FoldChange) share the same sign, which means that the MC treatment is affecting the majority of the genes in a similar fashion. This figure shows that the most differentially expressed genes are shared across cell lines, hence it is unlikely to find genes that are differentially expressed by the treatment in one cell line but not the other.

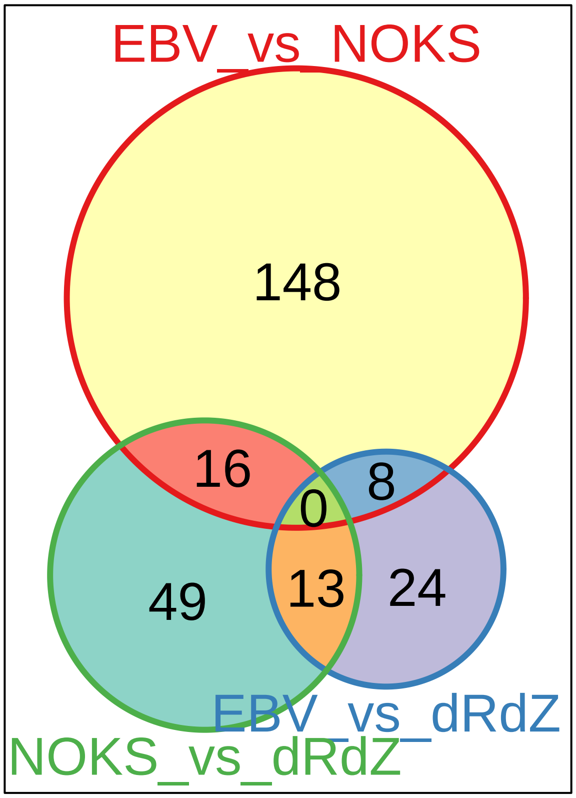


The figure above exhibits that in the case of NOKS vs dRdZ, the signal to noise metrics are relatively symmetric around the diagonal, but in the case of the other two pairs (EBV vs NOKS, and EBV vs dRdZ) we can notice a slightly tilted distribution towards NOKS, and dRdZ respectively. Searching in the Bioconductor forums, I found the following answer by the creator of DESeq2, which states a framework to test ratio of ratio (which appears to be a common problem in RIP-seq / CLIP-seq), where the IP counts are normalized by the Input counts (very similar to ChIP-seq).

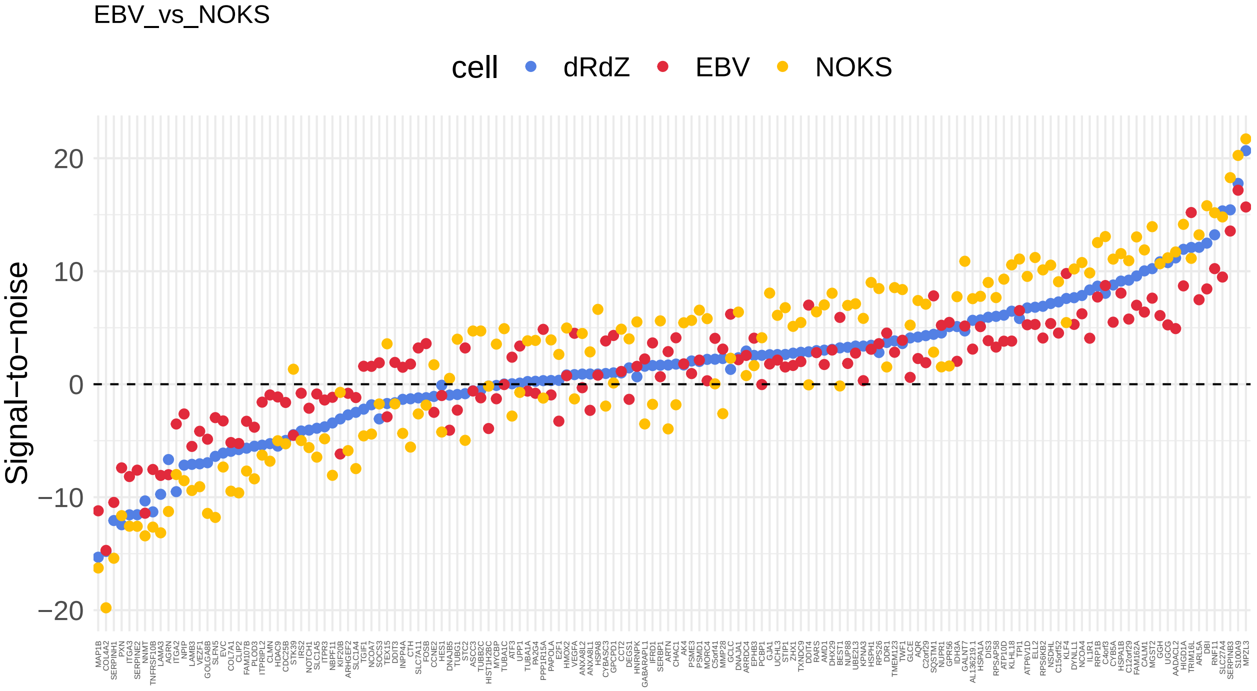
For that purpose, we fit a different models for each pair of cell lines, which means that we are testing the `cell:treatment` effect, which indicates that we are testing the `treatment` effect across `cell` lines.

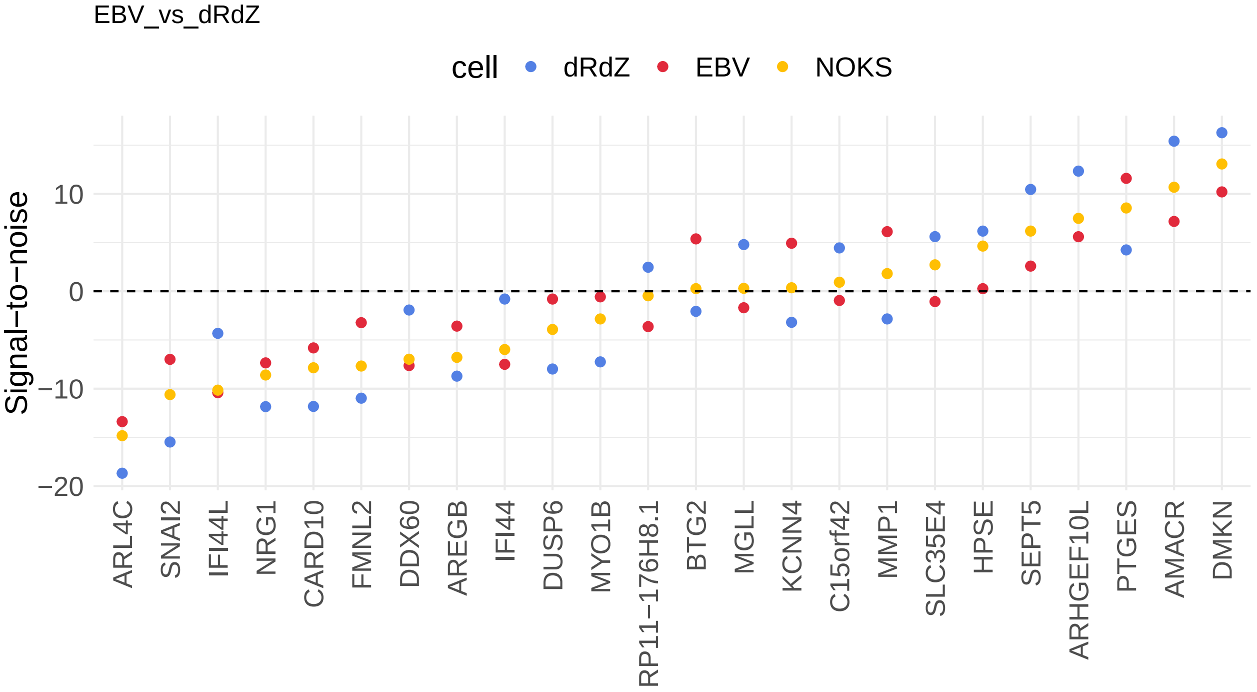
```
ratio_of_ratios_deseq <- function(rsem_data,thr = 20)
{
  count_matrix =  rsem_data %>%
    dplyr::select(file,rsem) %>%
    unnest() %>%
    dplyr::select(file,gene_id,expected_count) %>%
    mutate(
      expected_count = floor(expected_count)
      ) %>%
    spread(file,expected_count) %>%
    as_matrix()

  coldata = rsem_data %>%
    dplyr::select(file,cell,treatment) %>%
    mutate(interac = paste(cell,treatment, sep = ".")) %>%
    as.data.frame() %>%
    tibble::remove_rownames() %>%
    tibble::column_to_rownames("file")

  deseq = DESeqDataSetFromMatrix(
    count_matrix,colData = coldata,
    design = ~ cell + treatment + cell:treatment)

  deseq = deseq[ rowSums(assay(deseq) ) > thr,]
  deseq = DESeq(deseq, test = "LRT", reduced =  ~ cell + treatment)

  deseq

}
```

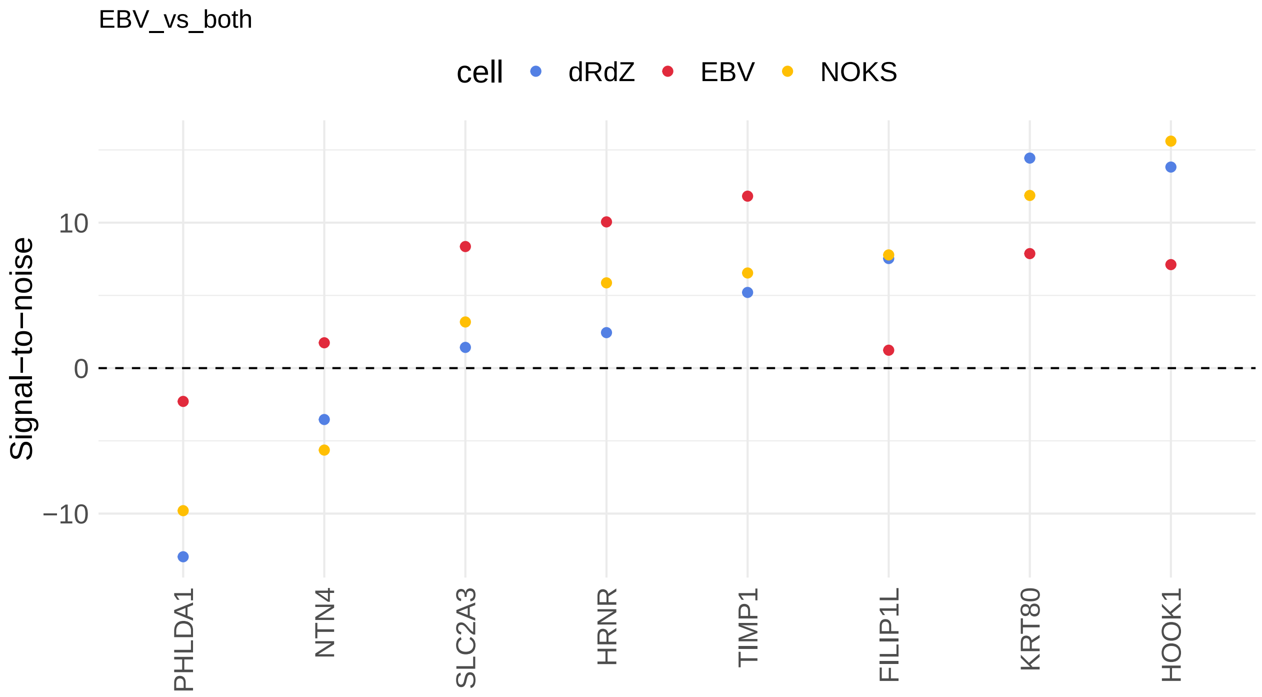
This test returns a smaller list of genes that are differentially expressed. For example, if we consider the genes that are differentially expressed with adjusted p.value ≤ 0.01, we can notice that the number of genes that are differentially expressed is much larger when testing the MC effect across the EBV and NOKS cell lines, than when any of them is the mutant type dRdZ.

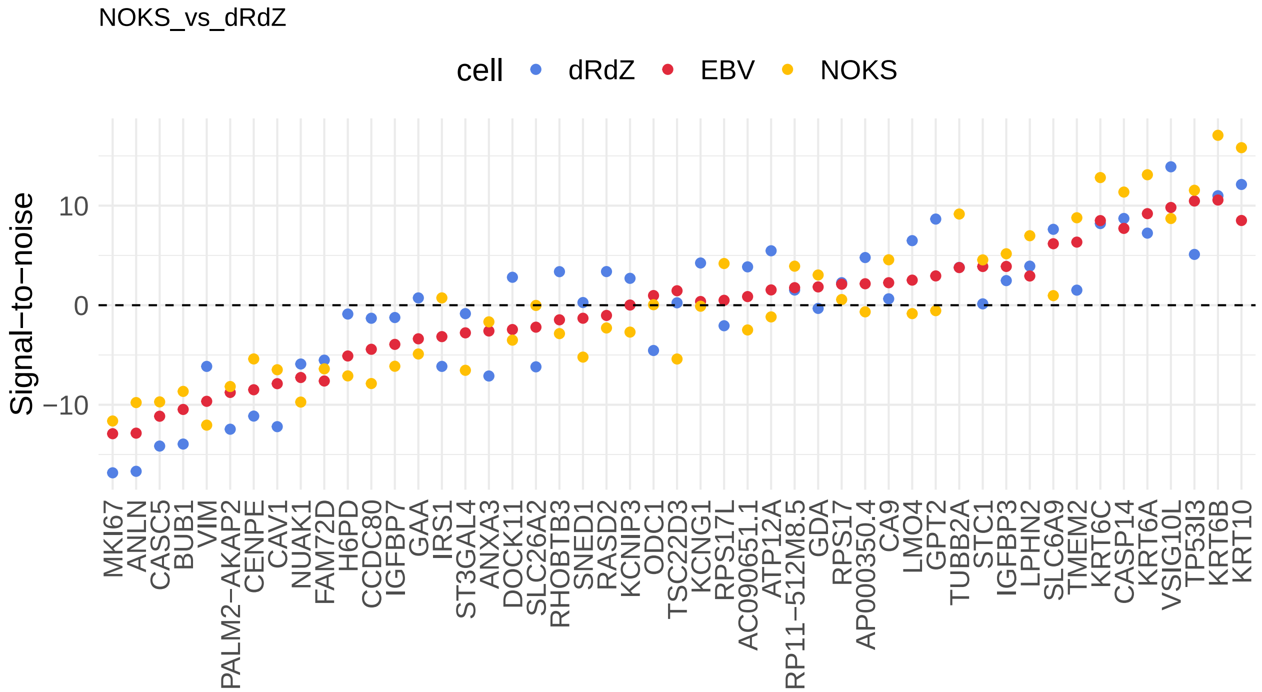


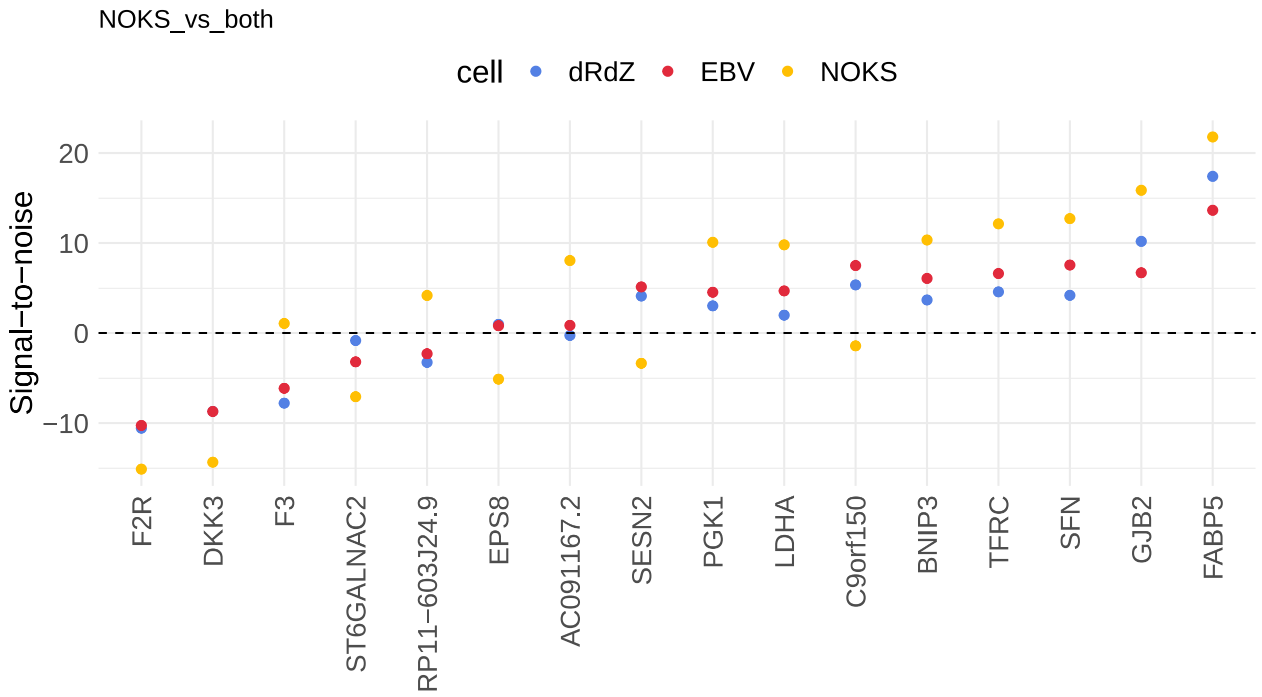
Furthermore, we can examine the genes in these subgroups. For example, we can notice that the intersection of `EBV_vs_dRdZ` and `EBV_vs_NOKS` have 8, and in that group there are genes such that the distance between EBV and the other cell lines are maximized, thus it results in genes where the signal-to-noise is close between them (i.e. NOKS and dRdZ). Alternatively, in the regions of the Venn diagram where the genes are only differentiated in one category, we can observe that the cell line that was not considered is usually in the middle.

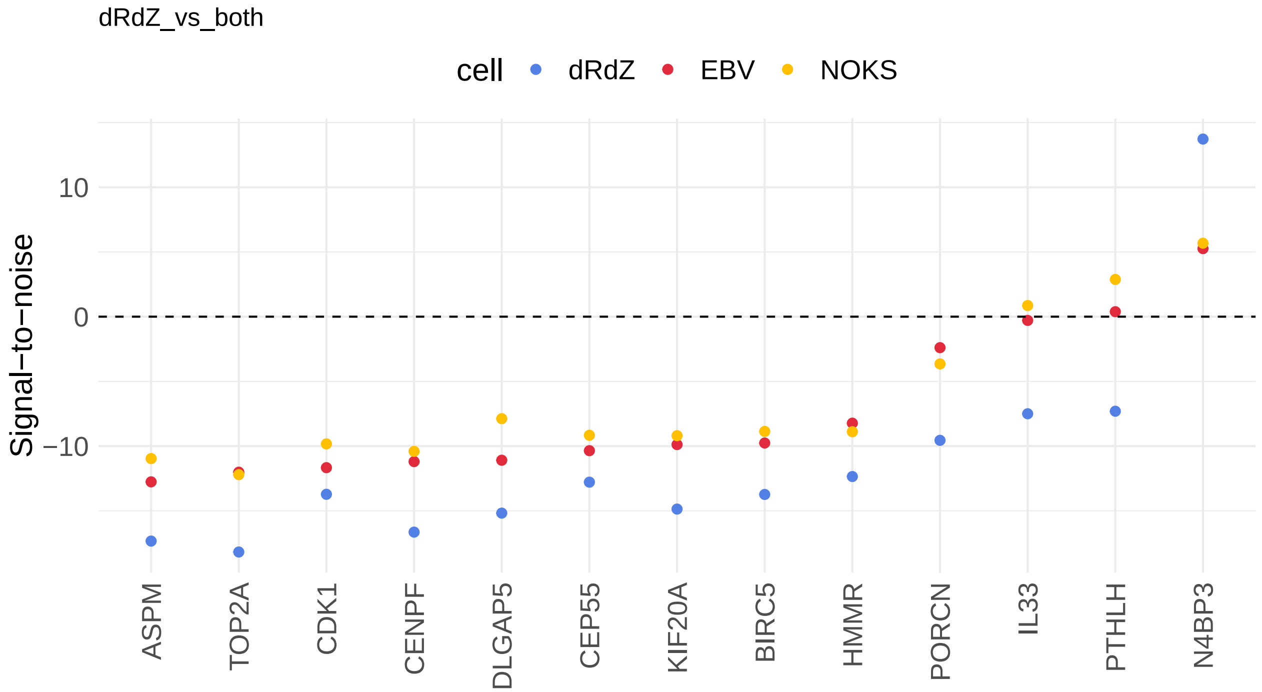

EBV_vs_NOKS
cell
● dRdZ
● EBV
● NOKS
Signal−to−noise
20
10
0
−10
−20
MAP1B
COL4A2
SERPINH1
PXN
ITGA3
SERPINE2
NNMT
TNFRSF10B
LAMA3
AGRN
ITGA2
NPIP
LAMB3
VEZF1
GOLGA8B
SLFN5
EVC
COL7A1
CLIP2
FAM107B
PLOD3
ITPRIPL2
CLMN
HDAC9
CDC25B
STK39
IRS2
NOTCH1
SLC1A5
ITPR3
NBPF11
KIF20B
ARHGEF2
SLC1A4
TGIF1
NCOA7
SOCS3
TEX15
DDIT3
INPP4A
CTH
SLC7A11
FOSB
CCNE2
HES1
DNAJB5
TUBG1
STC2
ASCC3
TUBB2C
HIST1H2BC
MYCBP
TUBA1C
ATF3
UPP1
TUBA1A
PA2G4
PPP1R15A
PAPOLA
E2F1
HMOX2
VEGFA
ANXA8L2
ANXA8L1
HSPA8
CYBASC3
GPCPD1
CCT2
DEGS1
HNRNPK
GABARAPL1
IFRD1
SERBP1
ARTN
CHAC1
AK4
PSME3
PSMD1
MORC4
C5orf41
MMP28
GCLC
DNAJA1
ARRDC4
EPHB3
PCBP1
GJA1
UCHL3
STIP1
ZHX1
TXNDC9
DDIT4
RARS
AMD1
DHX29
BEST1
NUP88
UBE2L3
KPNA3
HSPH1
RPS26
DDR1
TMEM123
TWF1
GLCE
AQR
C2orf29
SQSTM1
NUPR1
GPR56
IDH3A
GALNT7
AL136219.1
HSPA1A
DIS3
RPSAP58
ATP10D
KLHL18
TPI1
ATP6V1D
ELL2
RPS6KB2
NSDHL
C15orf52
KLF4
DYNLL1
NCOA4
IL1R1
RRP1B
C4orf3
CYB5A
HSPA1B
C12orf29
FAM162A
CALM1
MGST2
GGH
UGCG
AADACL2
HIGD1A
TRIM16L
ARL5A
DBI
RNF11
SLC27A4
SERPINB3
S100A9
MPZL3


EBV_vs_dRdZ
cell
● dRdZ
● EBV
● NOKS
Signal−to−noise
10
0
−10
−20
ARL4C
SNAI2
IFI44L
NRG1
CARD10
FMNL2
DDX60
AREGB
IFI44
DUSP6
MYO1B
RP11−176H8.1
BTG2
MGLL
KCNN4
C15orf42
MMP1
SLC35E4
HPSE
SEPT5
ARHGEF10L
PTGES
AMACR
DMKN


EBV_vs_both

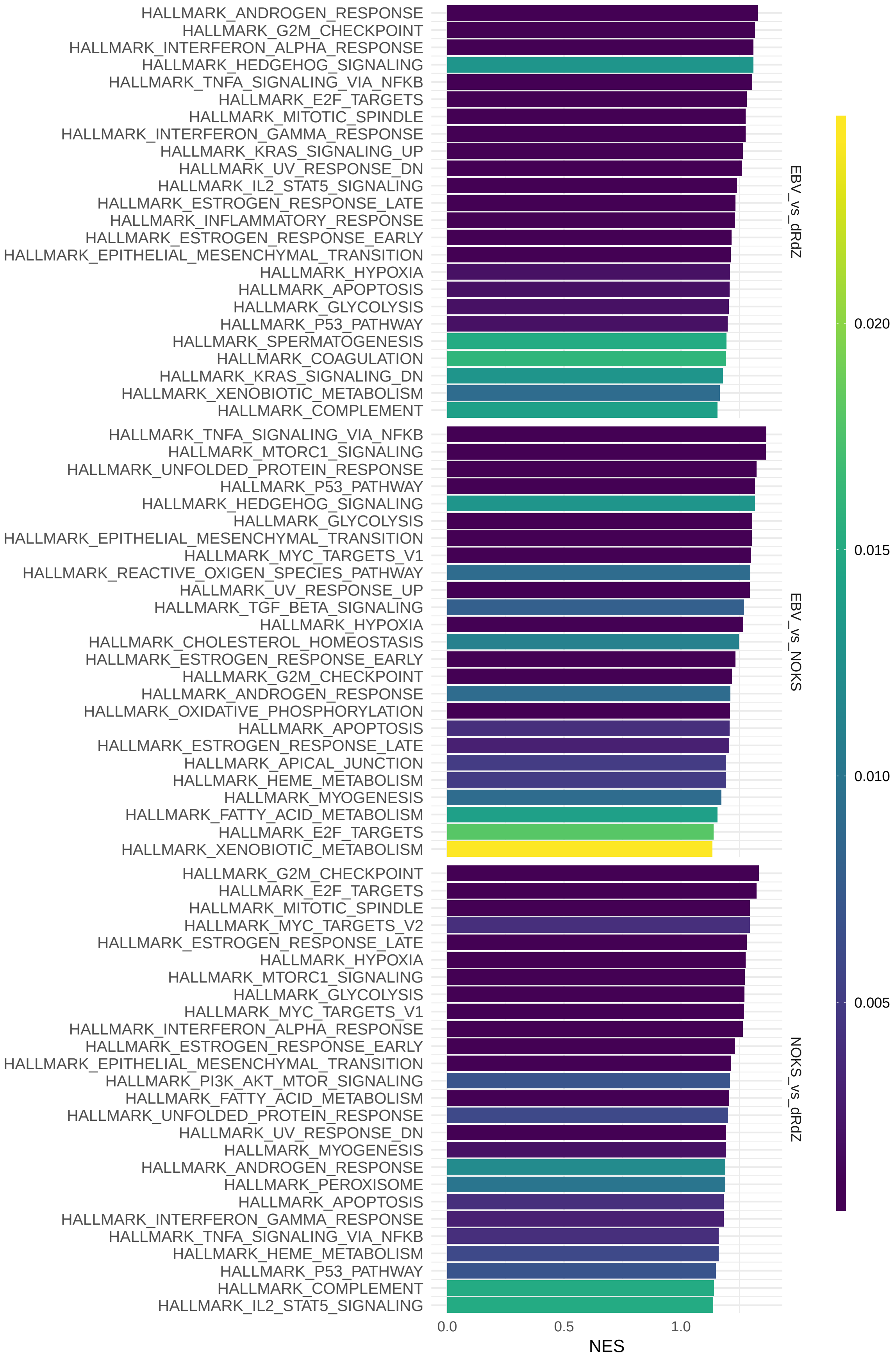

NOKS_vs_dRdZ



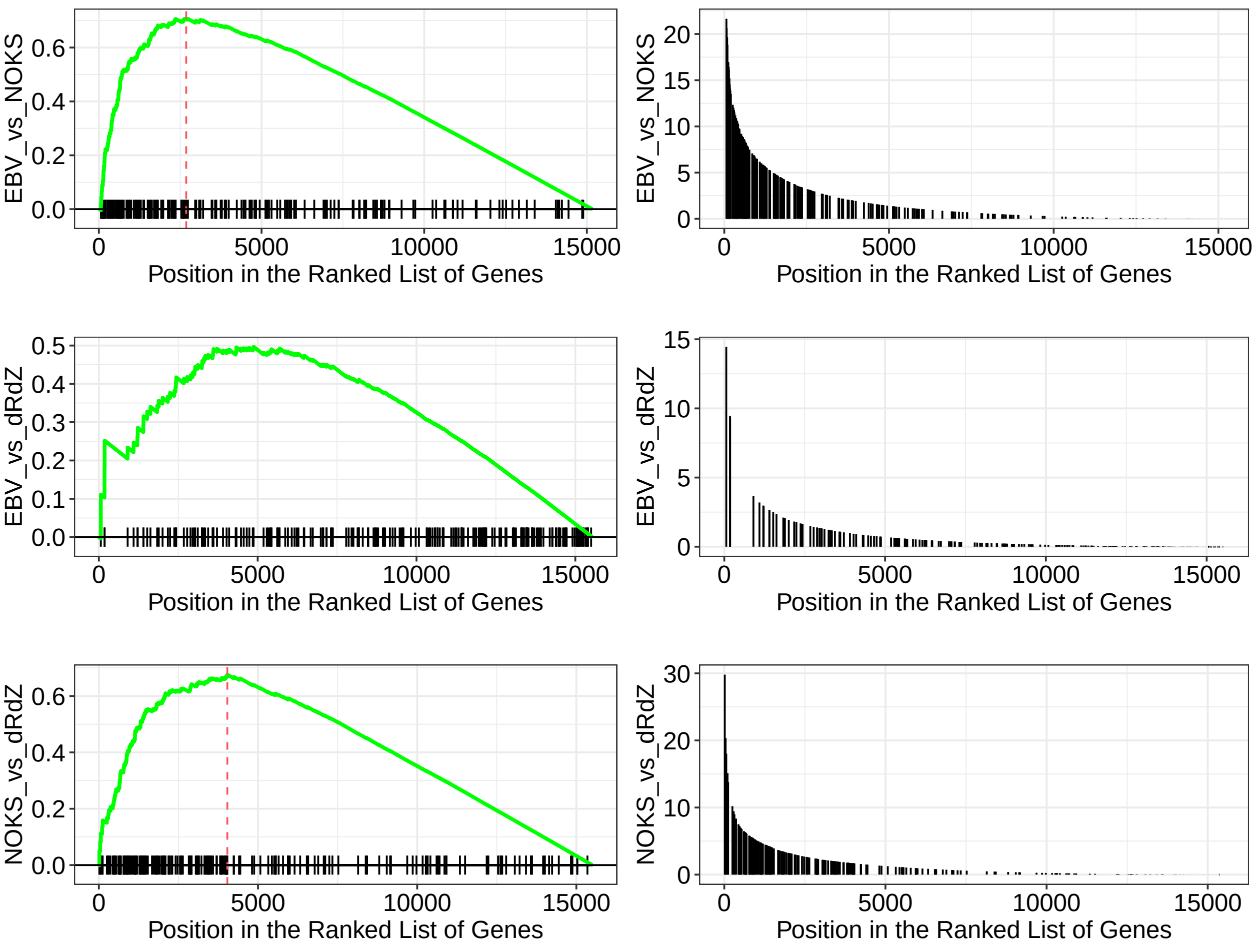


Pathway analysis

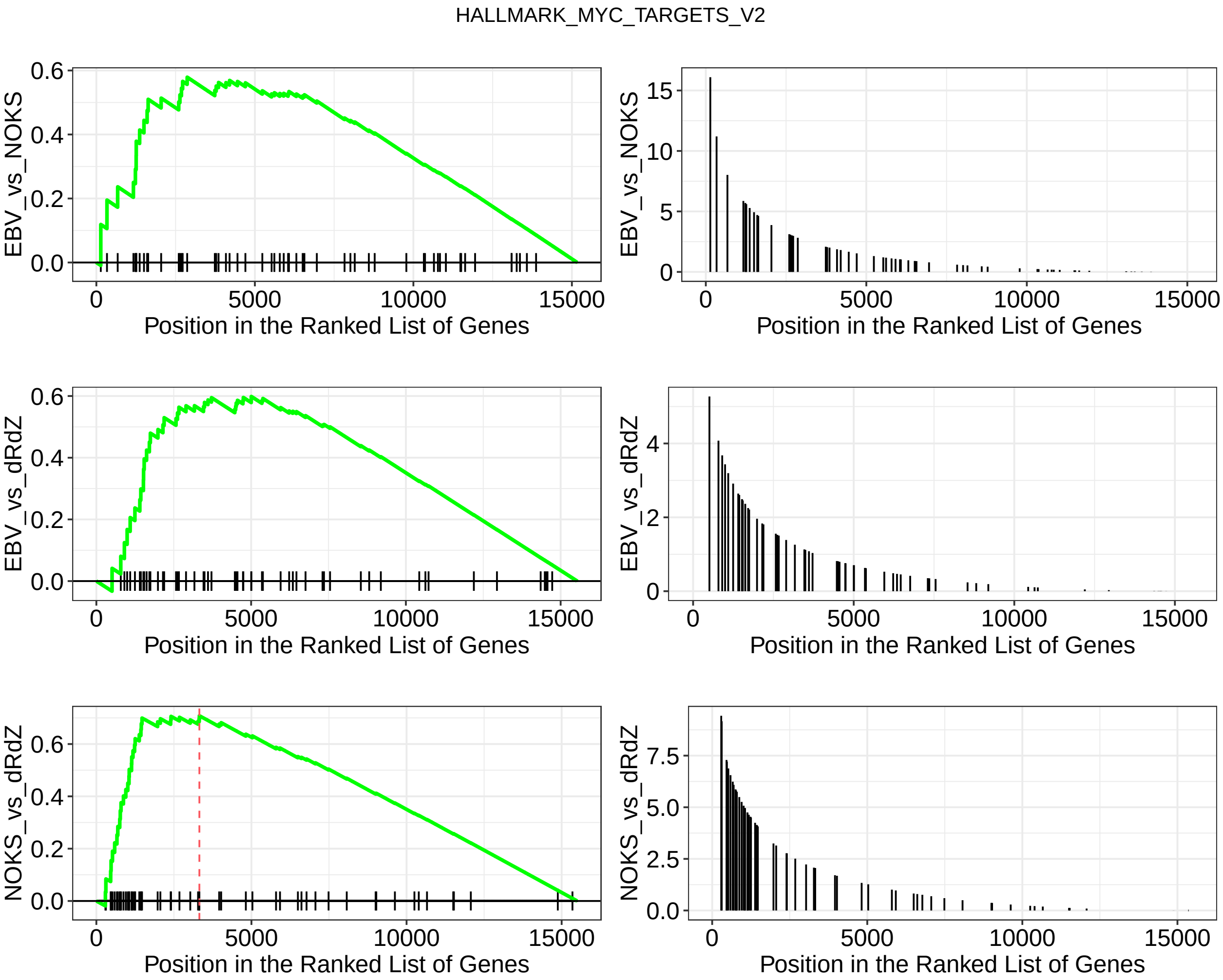
This analysis returns a different signal-to-noise metric for each cell line. Hence, we are capable of performing a pathway analysis too. For example, we can notice that the none of the `MYC_TARGETS_V1` or `MYC_TARGETS_V2` pathways are enriched in the `EBV_vs_dRdZ` test, but the first one is enriched in the `EBV_vs_NOKS` test and both are enriched in the `NOKS_vs_dRdZ` test.


HALLMARK_ANDROGEN_RESPONSE
HALLMARK_G2M_CHECKPOINT
HALLMARK_INTERFERON_ALPHA_RESPONSE
HALLMARK_HEDGEHOG_SIGNALING
HALLMARK_TNFA_SIGNALING_VIA_NFKB
HALLMARK_E2F_TARGETS
HALLMARK_MITOTIC_SPINDLE
HALLMARK_INTERFERON_GAMMA_RESPONSE
HALLMARK_KRAS_SIGNALING_UP
HALLMARK_UV_RESPONSE_DN
HALLMARK_IL2_STAT5_SIGNALING
HALLMARK_ESTROGEN_RESPONSE_LATE
HALLMARK_INFLAMMATORY_RESPONSE
HALLMARK_ESTROGEN_RESPONSE_EARLY
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION
HALLMARK_HYPOXIA
HALLMARK_APOPTOSIS
HALLMARK_GLYCOLYSIS
HALLMARK_P53_PATHWAY
HALLMARK_SPERMATOGENESIS
HALLMARK_COAGULATION
HALLMARK_KRAS_SIGNALING_DN
HALLMARK_XENOBIOTIC_METABOLISM
HALLMARK_COMPLEMENT
EBV_vs_dRdZ
HALLMARK_TNFA_SIGNALING_VIA_NFKB
HALLMARK_MTORC1_SIGNALING
HALLMARK_UNFOLDED_PROTEIN_RESPONSE
HALLMARK_P53_PATHWAY
HALLMARK_HEDGEHOG_SIGNALING
HALLMARK_GLYCOLYSIS
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION
HALLMARK_MYC_TARGETS_V1
HALLMARK_REACTIVE_OXIGEN_SPECIES_PATHWAY
HALLMARK_UV_RESPONSE_UP
HALLMARK_TGF_BETA_SIGNALING
HALLMARK_HYPOXIA
HALLMARK_CHOLESTEROL_HOMEOSTASIS
HALLMARK_ESTROGEN_RESPONSE_EARLY
HALLMARK_G2M_CHECKPOINT
HALLMARK_ANDROGEN_RESPONSE
HALLMARK_OXIDATIVE_PHOSPHORYLATION
HALLMARK_APOPTOSIS
HALLMARK_ESTROGEN_RESPONSE_LATE
HALLMARK_APICAL_JUNCTION
HALLMARK_HEME_METABOLISM
HALLMARK_MYOGENESIS
HALLMARK_FATTY_ACID_METABOLISM
HALLMARK_E2F_TARGETS
HALLMARK_XENOBIOTIC_METABOLISM
EBV_vs_NOKS
HALLMARK_G2M_CHECKPOINT
HALLMARK_E2F_TARGETS
HALLMARK_MITOTIC_SPINDLE
HALLMARK_MYC_TARGETS_V2
HALLMARK_ESTROGEN_RESPONSE_LATE
HALLMARK_HYPOXIA
HALLMARK_MTORC1_SIGNALING
HALLMARK_GLYCOLYSIS
HALLMARK_MYC_TARGETS_V1
HALLMARK_INTERFERON_ALPHA_RESPONSE
HALLMARK_ESTROGEN_RESPONSE_EARLY
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION
HALLMARK_PI3K_AKT_MTOR_SIGNALING
HALLMARK_FATTY_ACID_METABOLISM
HALLMARK_UNFOLDED_PROTEIN_RESPONSE
HALLMARK_UV_RESPONSE_DN
HALLMARK_MYOGENESIS
HALLMARK_ANDROGEN_RESPONSE
HALLMARK_PEROXISOME
HALLMARK_APOPTOSIS
HALLMARK_INTERFERON_GAMMA_RESPONSE
HALLMARK_TNFA_SIGNALING_VIA_NFKB
HALLMARK_HEME_METABOLISM
HALLMARK_P53_PATHWAY
HALLMARK_COMPLEMENT
HALLMARK_IL2_STAT5_SIGNALING
NOKS_vs_dRdZ
0.0
0.5
1.0
NES
0.020
0.015
0.010
0.005

GSEA plots for `MYC_TARGETS`:



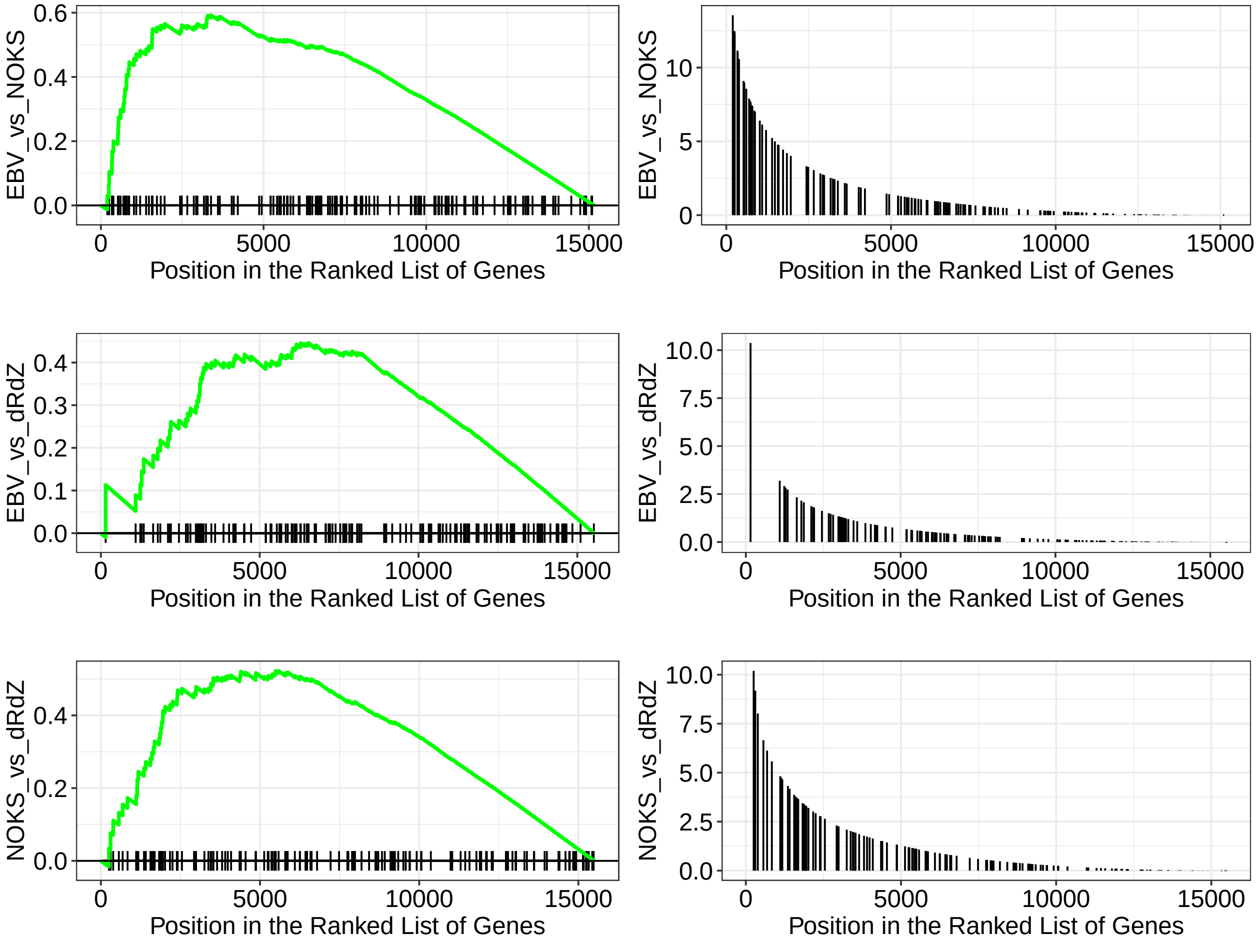
HALLMARK_MYC_TARGETS_V1

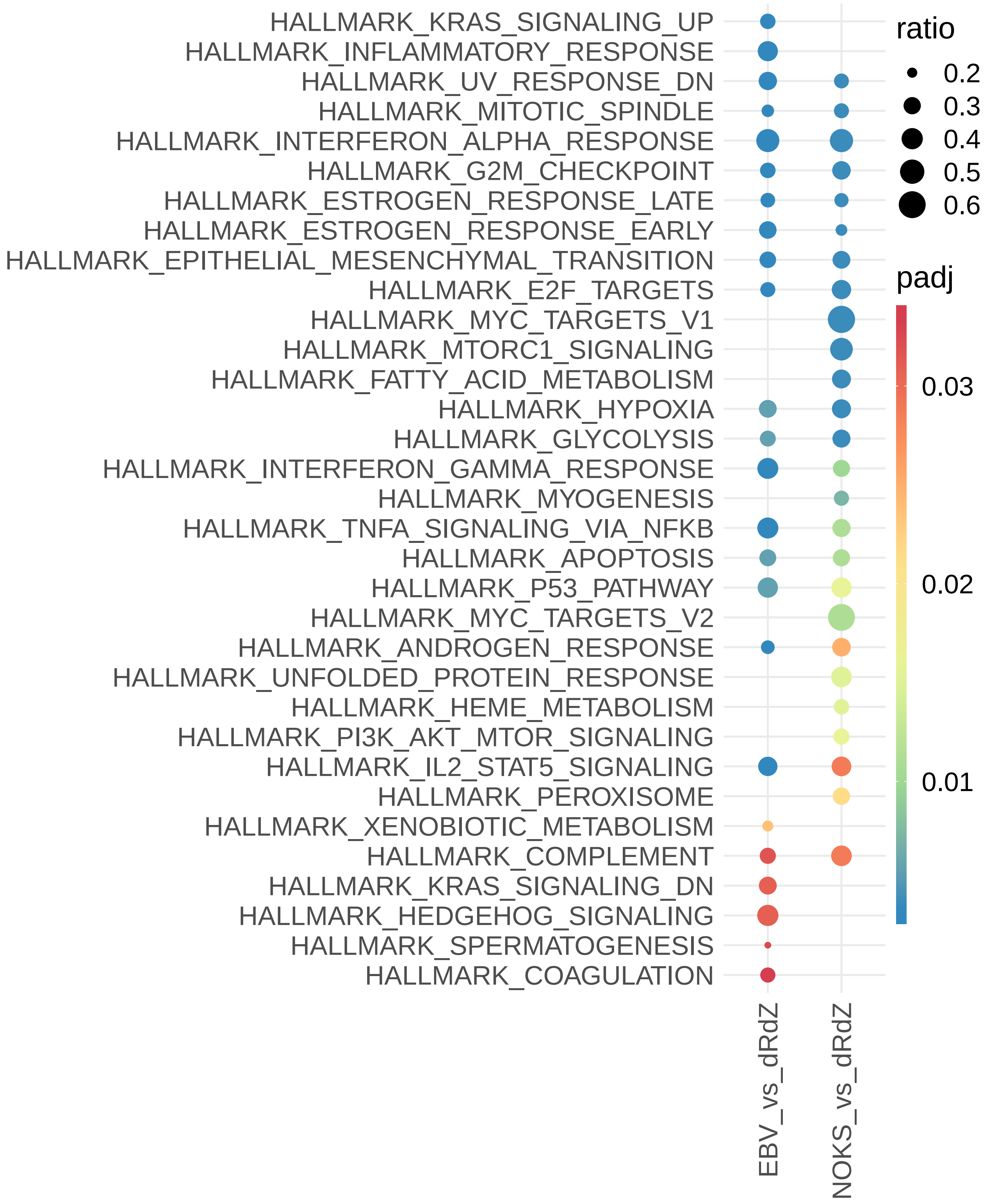


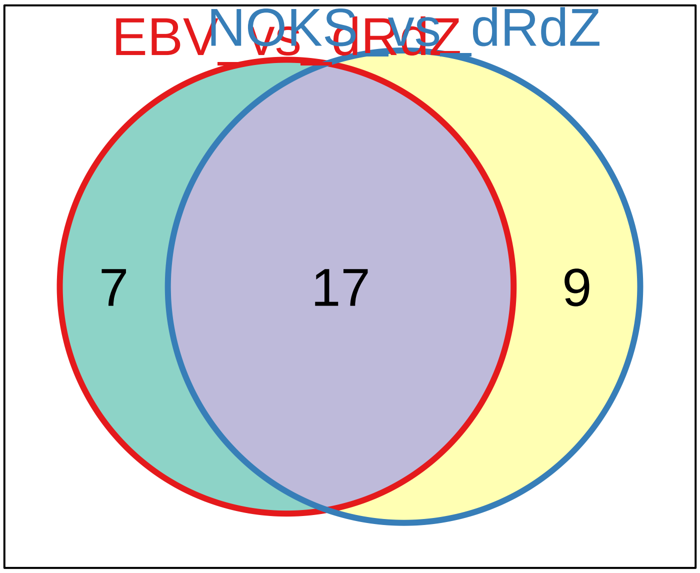
GSEA plot for `DNA_REPAIR`

I picked this pathway because it appear in both of the treated vs untreated pathways of Scott's cell lines, but in none of your data.

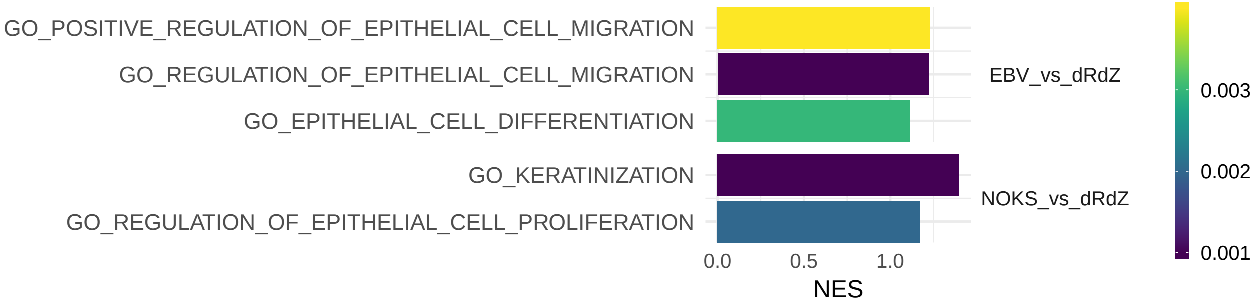
HALLMARK_DNA_REPAIR




HALLMARK_KRAS_SIGNALING_UP
HALLMARK_INFLAMMATORY_RESPONSE
HALLMARK_UV_RESPONSE_DN
HALLMARK_MITOTIC_SPINDLE
HALLMARK_INTERFERON_ALPHA_RESPONSE
HALLMARK_G2M_CHECKPOINT
HALLMARK_ESTROGEN_RESPONSE_LATE
HALLMARK_ESTROGEN_RESPONSE_EARLY
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION
HALLMARK_E2F_TARGETS
HALLMARK_MYC_TARGETS_V1
HALLMARK_MTORC1_SIGNALING
HALLMARK_FATTY_ACID_METABOLISM
HALLMARK_HYPOXIA
HALLMARK_GLYCOLYSIS
HALLMARK_INTERFERON_GAMMA_RESPONSE
HALLMARK_MYOGENESIS
HALLMARK_TNFA_SIGNALING_VIA_NFKB
HALLMARK_APOPTOSIS
HALLMARK_P53_PATHWAY
HALLMARK_MYC_TARGETS_V2
HALLMARK_ANDROGEN_RESPONSE
HALLMARK_UNFOLDED_PROTEIN_RESPONSE
HALLMARK_HEME_METABOLISM
HALLMARK_PI3K_AKT_MTOR_SIGNALING
HALLMARK_IL2_STAT5_SIGNALING
HALLMARK_PEROXISOME
HALLMARK_XENOBIOTIC_METABOLISM
HALLMARK_COMPLEMENT
HALLMARK_KRAS_SIGNALING_DN
HALLMARK_HEDGEHOG_SIGNALING
HALLMARK_SPERMATOGENESIS
HALLMARK_COAGULATION
EBV_vs_dRdZ
NOKS_vs_dRdZ
ratio
0.2
0.3
0.4
0.5
0.6
padj
0.03
0.02
0.01



Brief note: There are 17 shared pathways, 7 that are EBV vs dRdZ specific, and 9 that are NOKS vs dRdZ specific. This suggests, that dRdZ is more similar than NOKS as there is a smaller amount of enriched pathways that turn on by the genes that show a significant interaction between dRdZ and NOKS when the treatment in applied.



Demo gene coefficients

Lets recall the DESeq model, for gene i and sample j:

$$K_{ij} \sim \text{NB}(\mu_{ij}, \alpha_i)$$
$$\mu_{ij} = s_j q_{ij}$$
$$\log_2(q_{ij}) = \mathbf{x}_{j\cdot}\beta_i$$

where:

- K_{ij} are the counts of gene i and sample j
- μ_{ij} is the fitted mean, and α_i is a gene-specific dispersion parameter

The dispersion specific parameters are obtained by fitting a model of the following relationship:

$$\text{Var}(K_{ij}) = \mu_{ij} + \alpha_i \mu_{ij}^2$$

From this description of the model, we can notice that:

1. For every gene, we estimate a different model defined by its mean and dispersion parameters.

2. This model differs from a typical linear regression model in the estimation of the variance. From the previous equation, we can observe that the estimated variance of the counts for gene i and sample j depends on the estimated mean. Hence, when testing for the significance of more than one terms it is necessary to use different methods.
3. The method that we used (suggested in here) is the LRT. Which sort of works like an ANOVA test, as it always positive and we are comparing the likelihood of the full model `~ cell + treatment + cell:treatment` with the likelihood of the reduced model `~ cell + treatment`.

As an example, we are going to consider the test that we did for the ratios of EBV vs NOKS (respect to treatment), we can see that the size factors are a slight correction (i.e. close to 1) but we are still working in the count scale:

```
demo_deseq = ratio_of_ratios$deseq[[1]]  ## we already fit this one
colData(demo_deseq)
```

```
## DataFrame with 8 rows and 4 columns
##                                           cell treatment     interac
##                                       <factor>  <factor> <character>
## RNAseq-akata-noks-methyl_cell-rep1         EBV    methyl  EBV.methyl
## RNAseq-akata-noks-methyl_cell-rep2         EBV    methyl  EBV.methyl
## RNAseq-akata-noks-no_treatment-rep1        EBV      none    EBV.none
## RNAseq-akata-noks-no_treatment-rep2        EBV      none    EBV.none
## RNAseq-noks-methyl_cell-rep1              NOKS    methyl NOKS.methyl
## RNAseq-noks-methyl_cell-rep2              NOKS    methyl NOKS.methyl
## RNAseq-noks-no_treatment-rep1             NOKS      none   NOKS.none
## RNAseq-noks-no_treatment-rep2             NOKS      none   NOKS.none
##                                             sizeFactor
##                                              <numeric>
## RNAseq-akata-noks-methyl_cell-rep1   0.966044740953985
## RNAseq-akata-noks-methyl_cell-rep2   0.809050888573786
## RNAseq-akata-noks-no_treatment-rep1   1.63538799638813
## RNAseq-akata-noks-no_treatment-rep2   1.01831506952894
## RNAseq-noks-methyl_cell-rep1          0.94607812078984
## RNAseq-noks-methyl_cell-rep2         0.668767477283564
## RNAseq-noks-no_treatment-rep1         1.15666601909984
## RNAseq-noks-no_treatment-rep2         1.12423001513368
```

For every gene, we are testing the interaction effect between the reduced model `~ cell + treatment` and the full model `~ cell + treatment + cell:treatment`. The model matrices are:

- Reduced model `~ cell + treatment`

```
with(colData(demo_deseq), model.matrix( ~ cell + treatment))
```

```
##   (Intercept) cellNOKS treatmentnone
## 1           1        0             0
## 2           1        0             0
## 3           1        0             1
## 4           1        0             1
## 5           1        1             0
## 6           1        1             0
## 7           1        1             1
## 8           1        1             1
## attr(,"assign")
## [1] 0 1 2
## attr(,"contrasts")
## attr(,"contrasts")$cell
## [1] "contr.treatment"
##
## attr(,"contrasts")$treatment
## [1] "contr.treatment"
```

- Full model `~ cell + treatment + cell:treatment`

```
with(colData(demo_deseq), model.matrix( ~ cell + treatment + cell:treatment))
```

```
##   (Intercept) cellNOKS treatmentnone cellNOKS:treatmentnone
## 1           1        0             0                      0
## 2           1        0             0                      0
## 3           1        0             1                      0
## 4           1        0             1                      0
## 5           1        1             0                      0
## 6           1        1             0                      0
## 7           1        1             1                      1
## 8           1        1             1                      1
## attr(,"assign")
## [1] 0 1 2 3
## attr(,"contrasts")
## attr(,"contrasts")$cell
## [1] "contr.treatment"
##
## attr(,"contrasts")$treatment
## [1] "contr.treatment"
```

For example, we can check for a few genes:

```
count_matrix = assay(demo_deseq,"counts")

geneID = "MAP1B"
log2(count_matrix)[ rownames(count_matrix) %>%
          str_detect(geneID),]
```

```
##  RNAseq-akata-noks-methyl_cell-rep1  RNAseq-akata-noks-methyl_cell-rep2
##                            8.299208                            8.149747
## RNAseq-akata-noks-no_treatment-rep1 RNAseq-akata-noks-no_treatment-rep2
##                           11.332037                           10.831307
##        RNAseq-noks-methyl_cell-rep1        RNAseq-noks-methyl_cell-rep2
##                            7.531381                            7.321928
##       RNAseq-noks-no_treatment-rep1       RNAseq-noks-no_treatment-rep2
##                           11.608255                           11.178665
```

```
geneID = "MPZL3"
log2(count_matrix)[ rownames(count_matrix) %>%
          str_detect(geneID),]
```

```
##  RNAseq-akata-noks-methyl_cell-rep1  RNAseq-akata-noks-methyl_cell-rep2
##                           10.440869                           10.271463
## RNAseq-akata-noks-no_treatment-rep1 RNAseq-akata-noks-no_treatment-rep2
##                            8.366322                            7.954196
##        RNAseq-noks-methyl_cell-rep1        RNAseq-noks-methyl_cell-rep2
##                           11.276124                           10.723661
##       RNAseq-noks-no_treatment-rep1       RNAseq-noks-no_treatment-rep2
##                            7.734710                            7.569856
```

```
coeffs = coef(demo_deseq)

geneID = "MAP1B"
coeffs[ rownames(coeffs) %>%
          str_detect(geneID),]
```

```
##                Intercept         cell_NOKS_vs_EBV treatment_none_vs_methyl
##                8.4025145               -0.6442865                2.3136192
##   cellNOKS.treatmentnone
##                1.1452521
```

```
geneID = "MPZL3"
coeffs[ rownames(coeffs) %>%
          str_detect(geneID),]
```

```
##                Intercept         cell_NOKS_vs_EBV treatment_none_vs_methyl
##               10.5343668                0.7961216               -2.7425265
##   cellNOKS.treatmentnone
##               -1.1235614
```