# Both datasets exploratory

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.
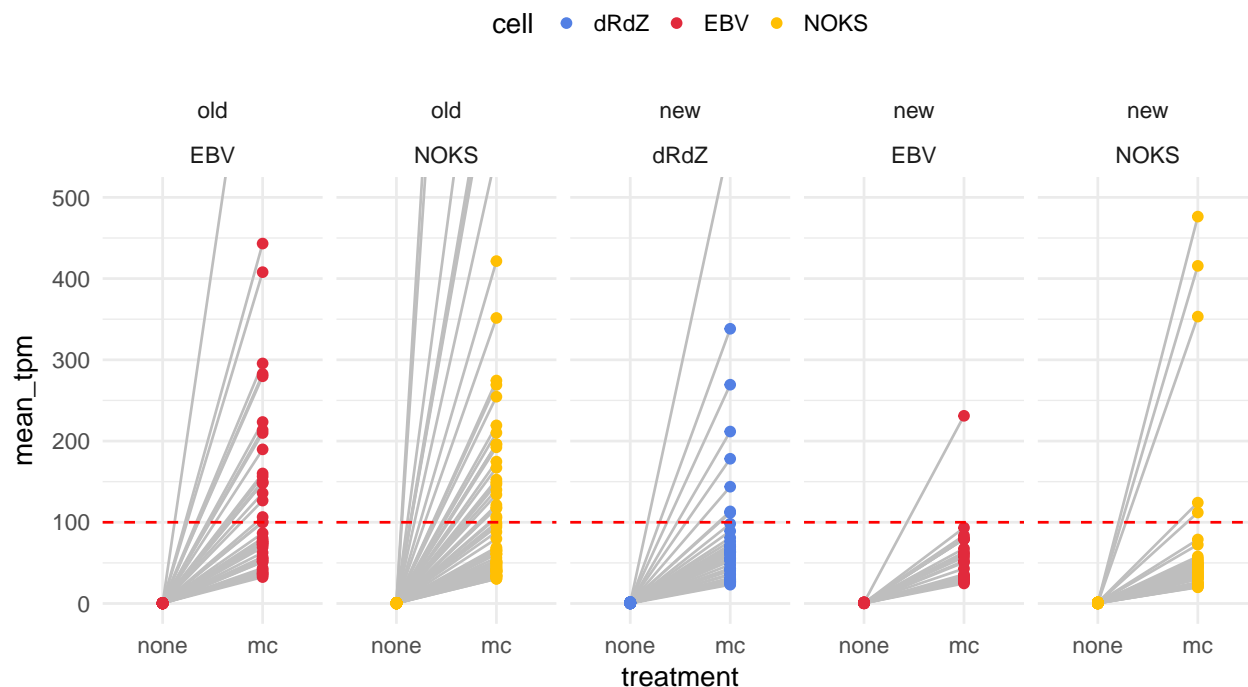
TPM summaries

Both data batches are grouped, so we are going to rank the genes by the mean TPM.
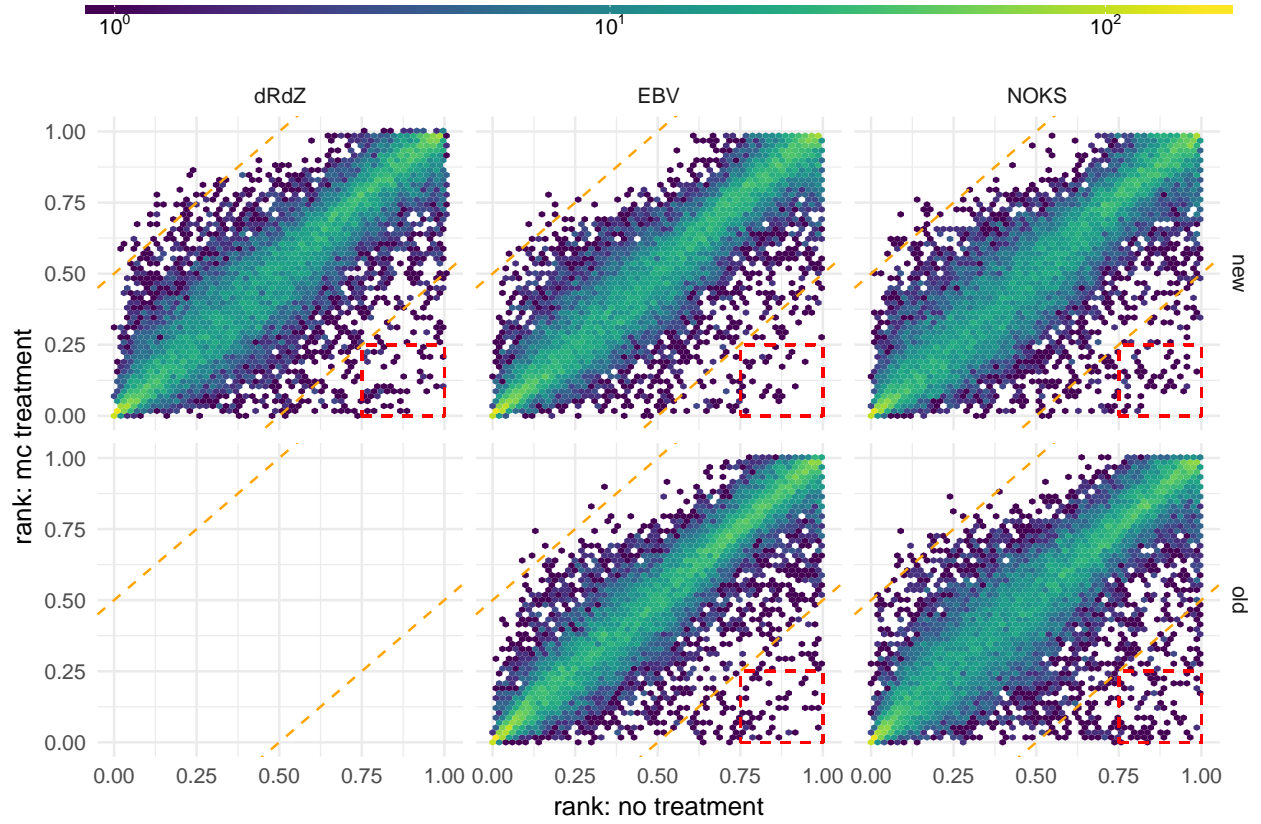
We further explore the genes in the red-square there are the genes that go from being unexpressed without treatment to be very expressed with treatment. In total, there are:

batch	cell	genes_in_square
new	dRdZ	53
new	EBV	25
new	NOKS	47
old	EBV	41
old	NOKS	73

In the figure above, it is shown that as expected when no treatment applied, those genes are not expressed and very likely there is no signal if we observe the tracks. On the other hand, the average tpm for the genes after applying the mc treatment, we can notice that there is some expression. To provide context, we observe the mean TPM distribution:
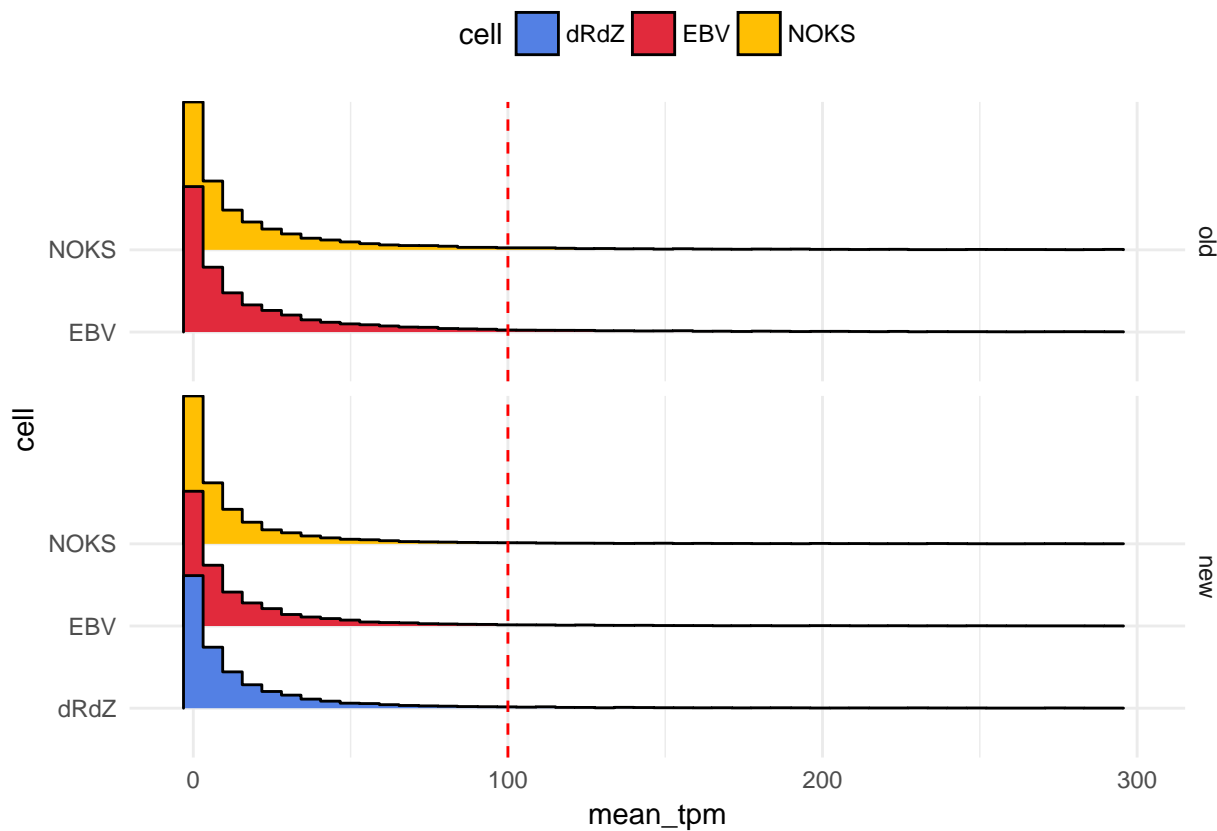
Filtering out genes

We filtered out genes for the GSEA analysis, by considering only the genes inside the orange lines that are differentially expressed (defined as genes with adj. p.value ≤ 0.01):

batch	cell	total_genes	diff_genes
old	EBV	16057	2302
old	NOKS	16057	4230
new	EBV	15876	1624
new	NOKS	15876	2445
new	dRdZ	15876	2025

GSEA analysis

Hallmark analysis
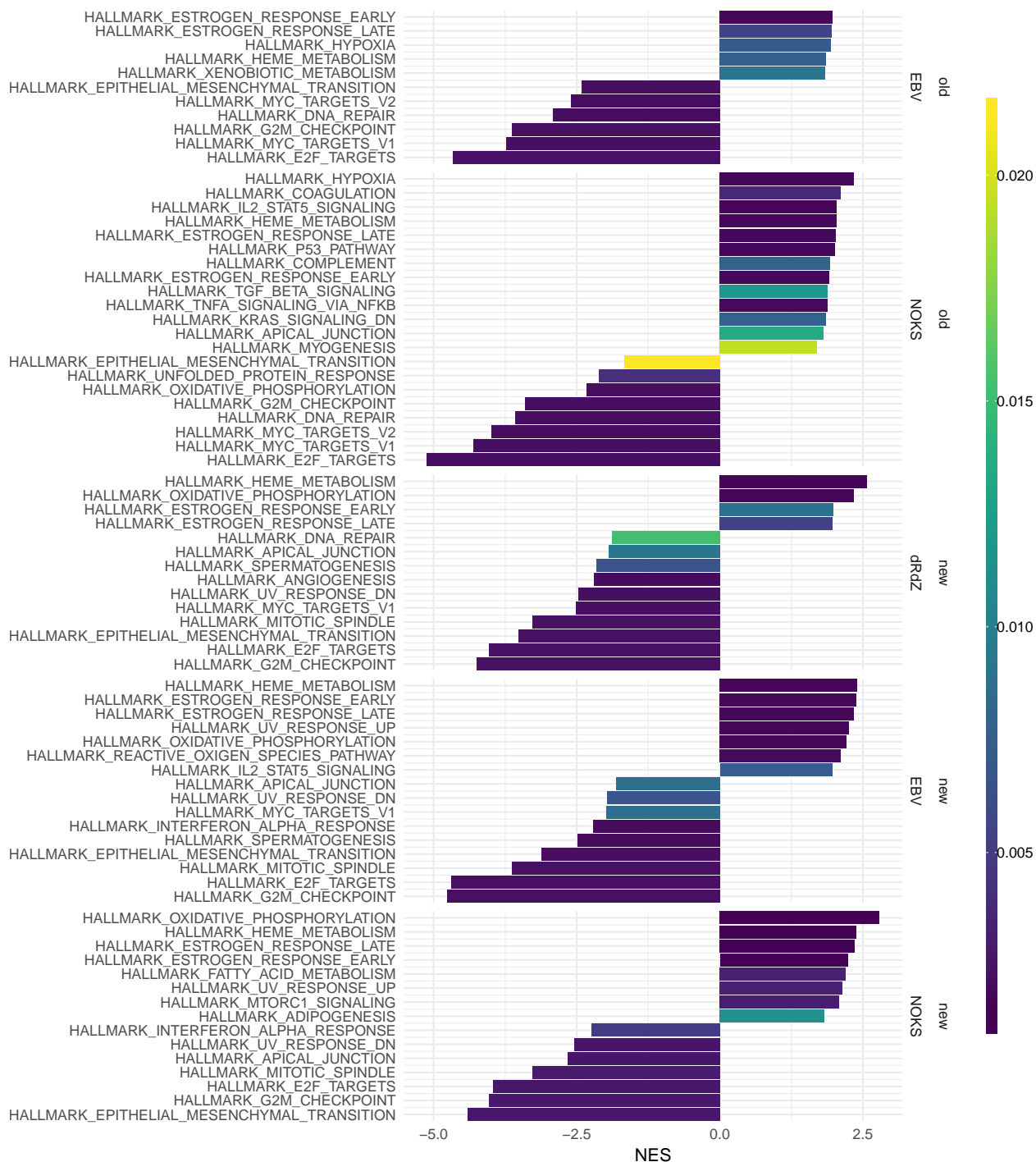
Curated analysis
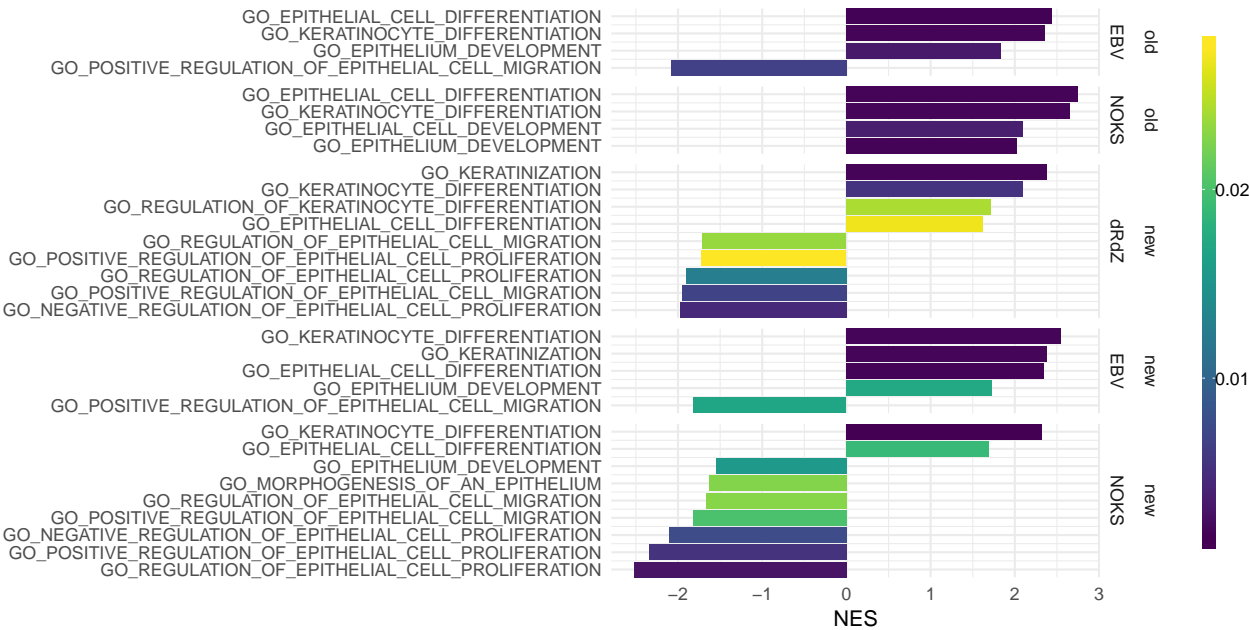
Further analysis

Hallmark pathways

HALLMARK_ESTROGEN_RESPONSE_EARLY

HALLMARK_DNA_REPAIR
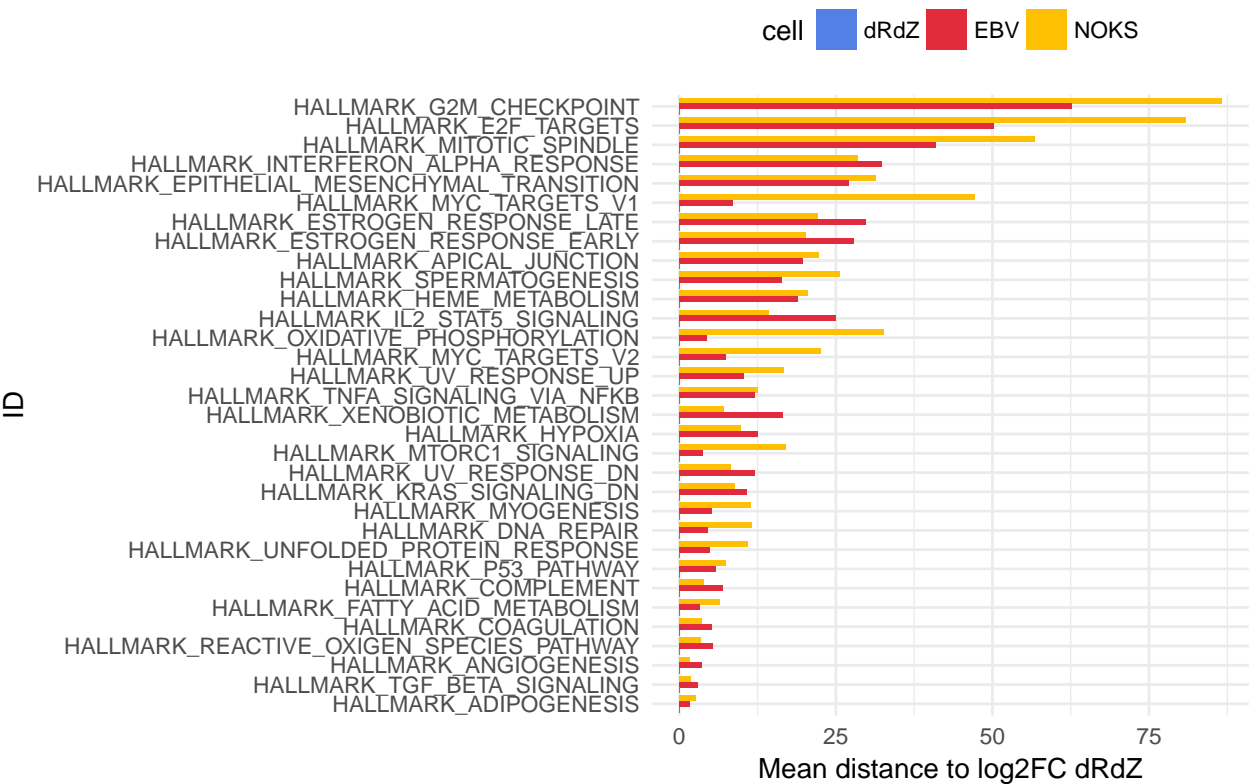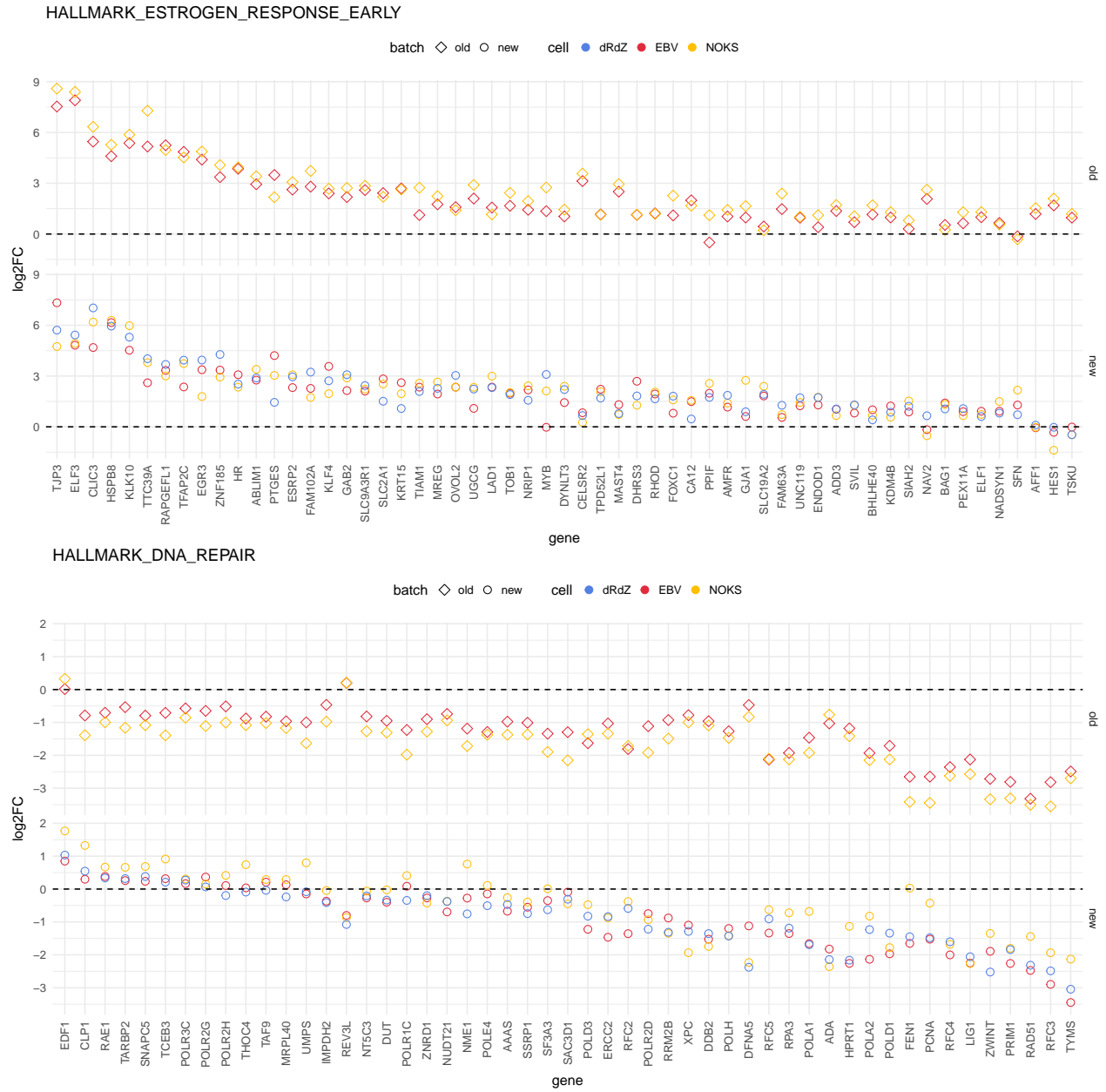
HALLMARK_MYC_TARGETS_V2
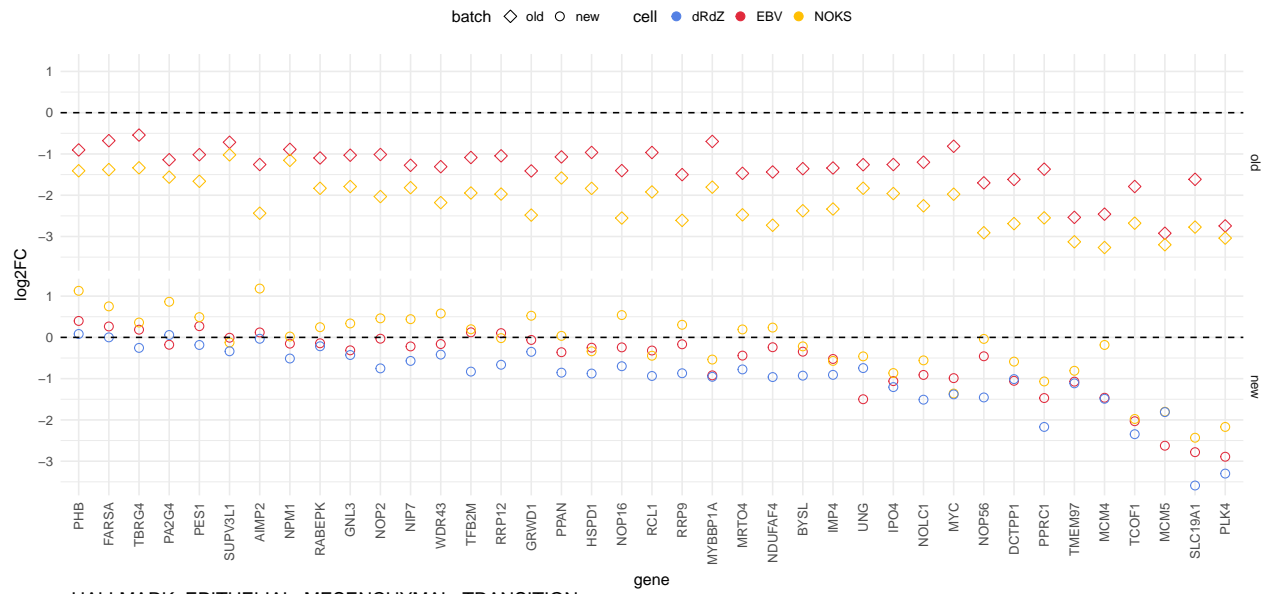
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION
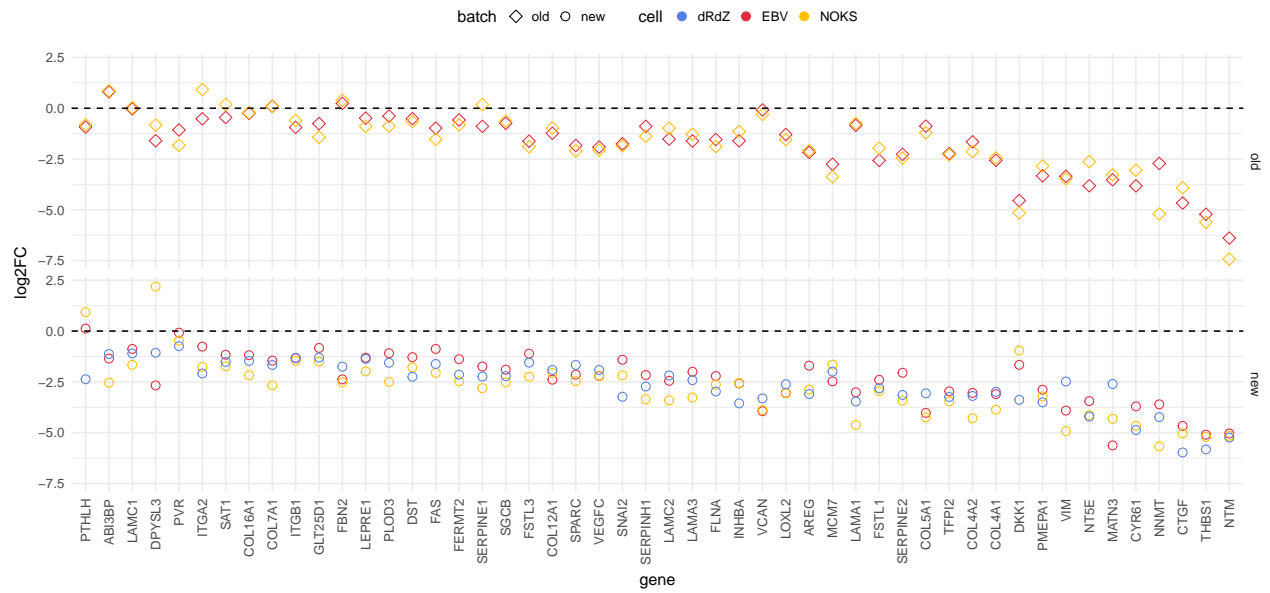
HALLMARK_MYC_TARGETS_V1
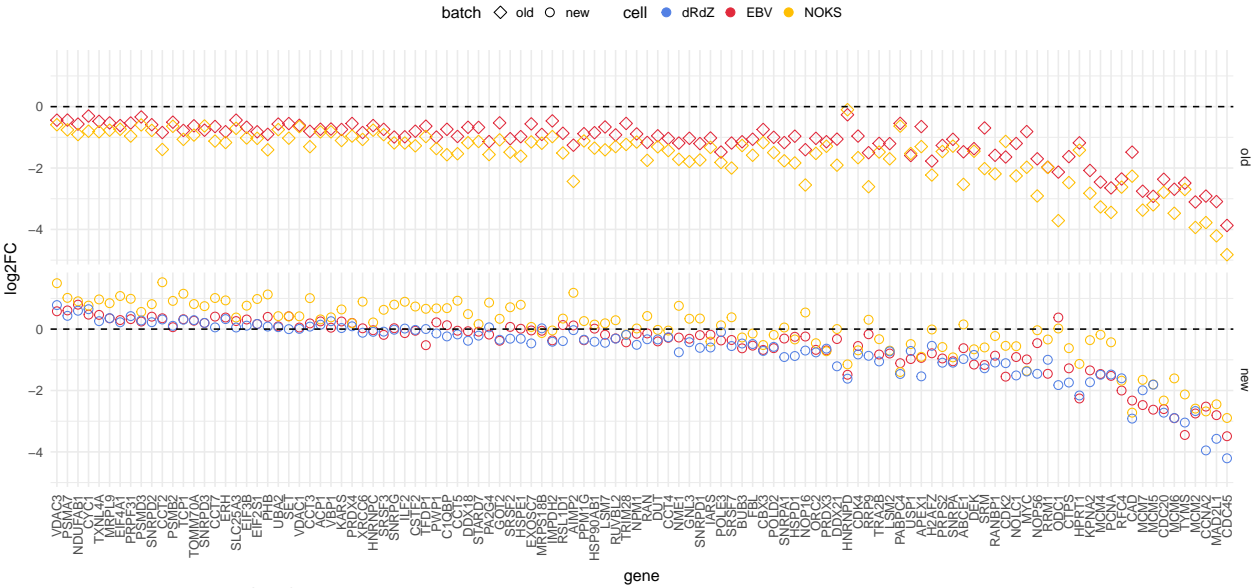
HALLMARK_E2F_TARGETS
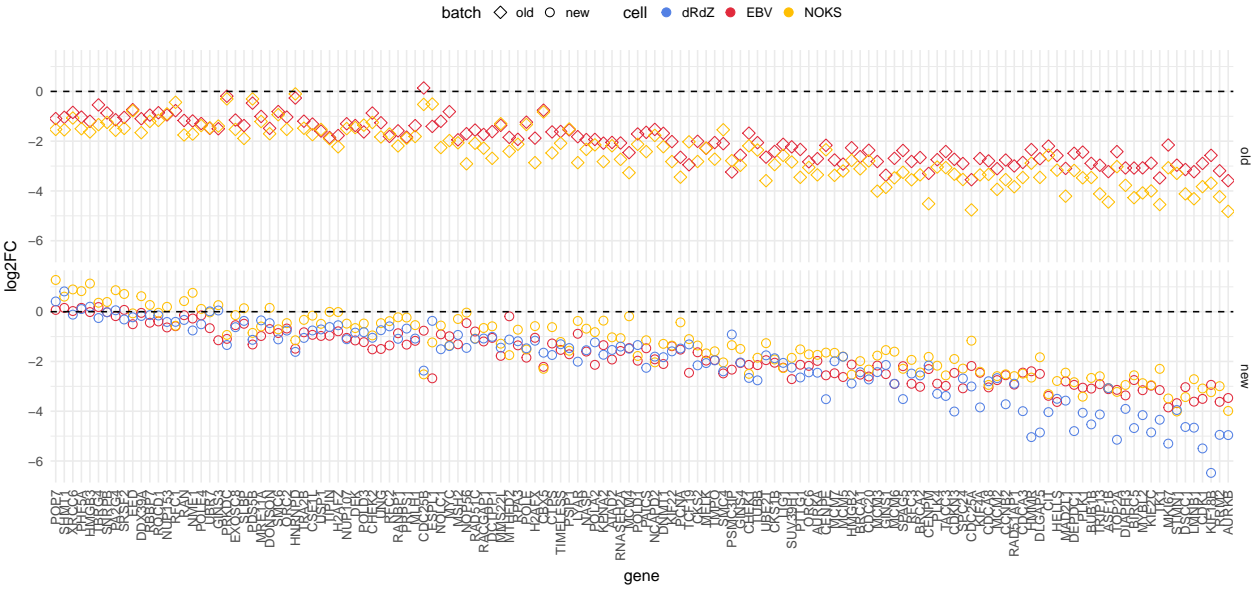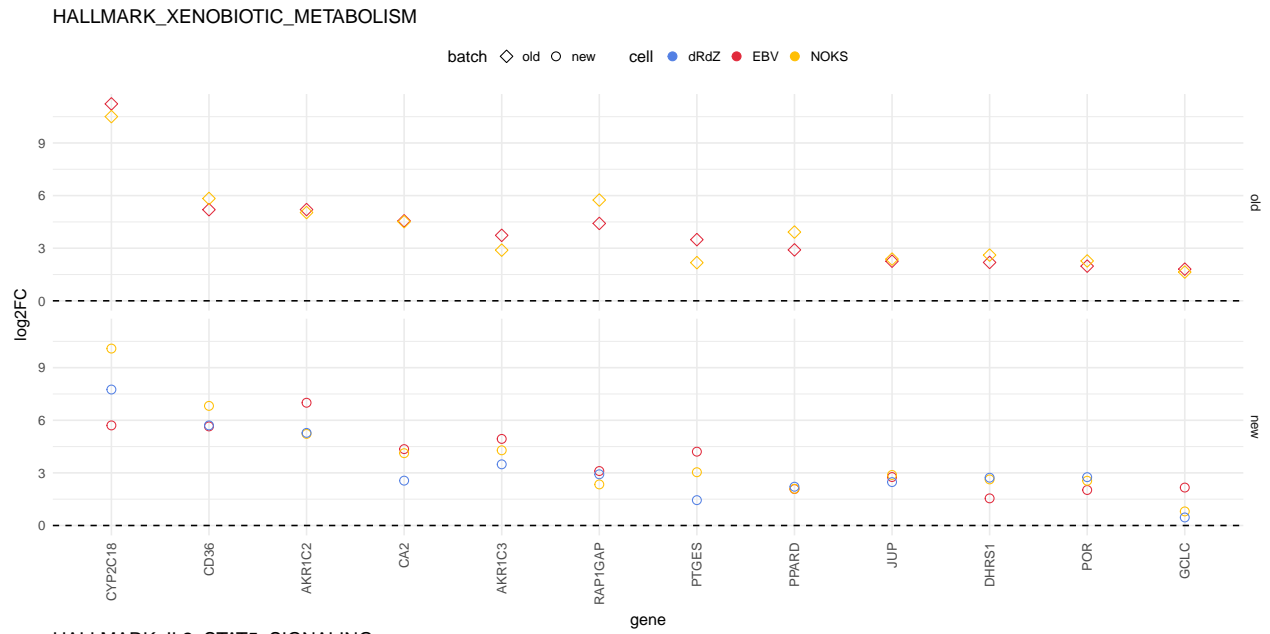
HALLMARK_G2M_CHECKPOINT
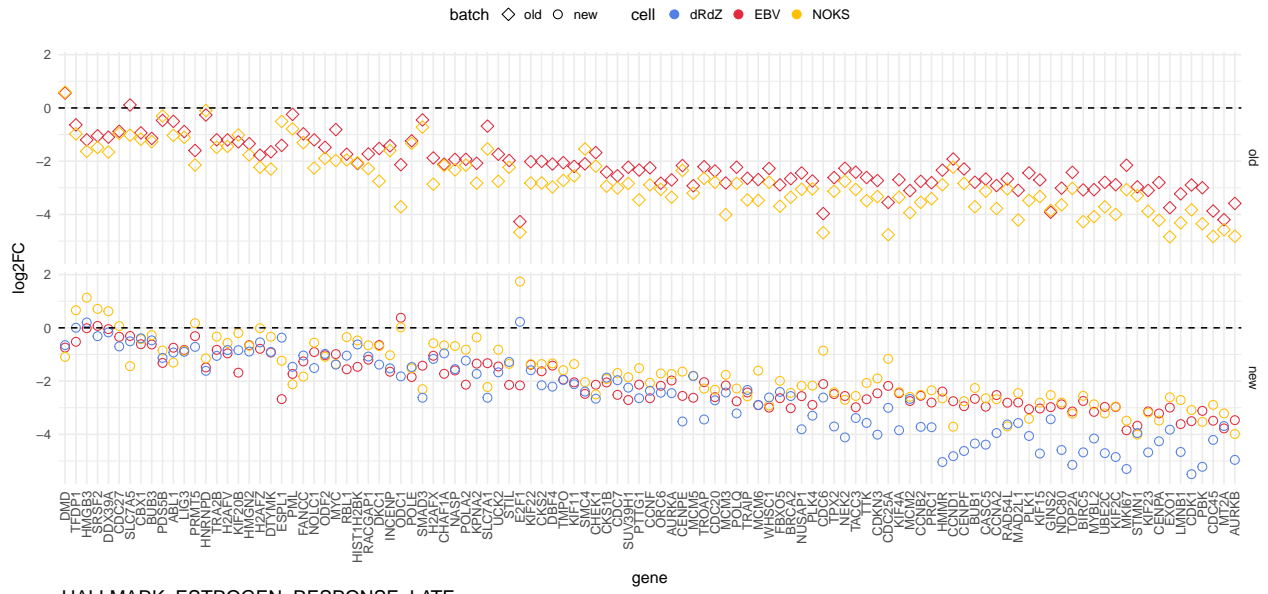
HALLMARK_ESTROGEN_RESPONSE_LATE

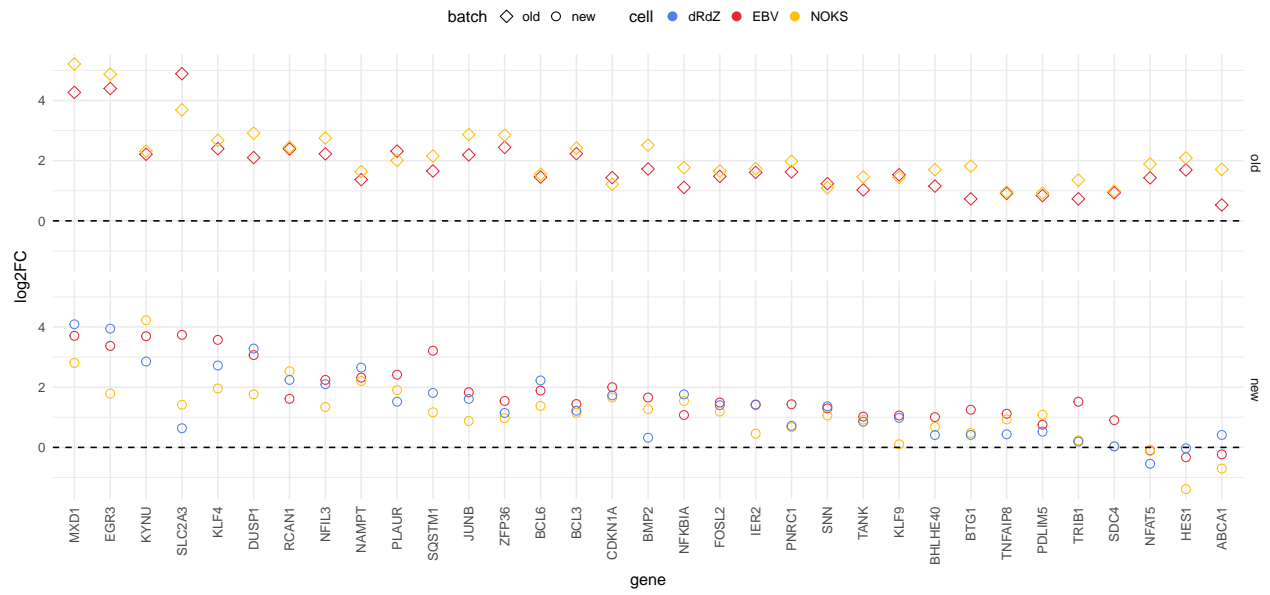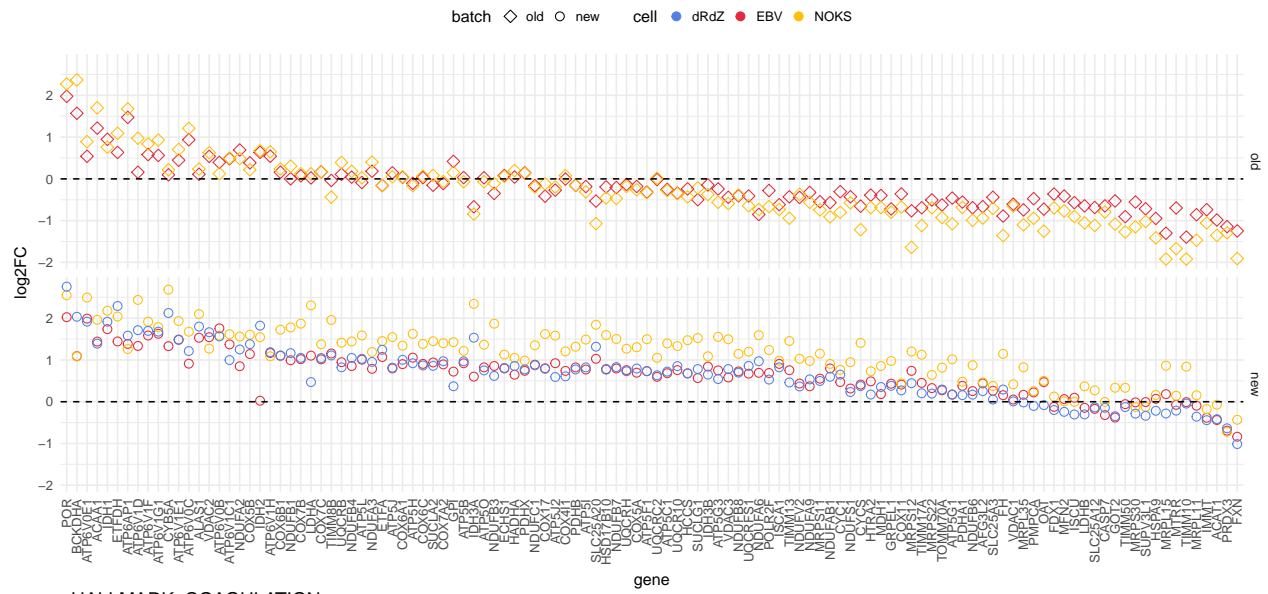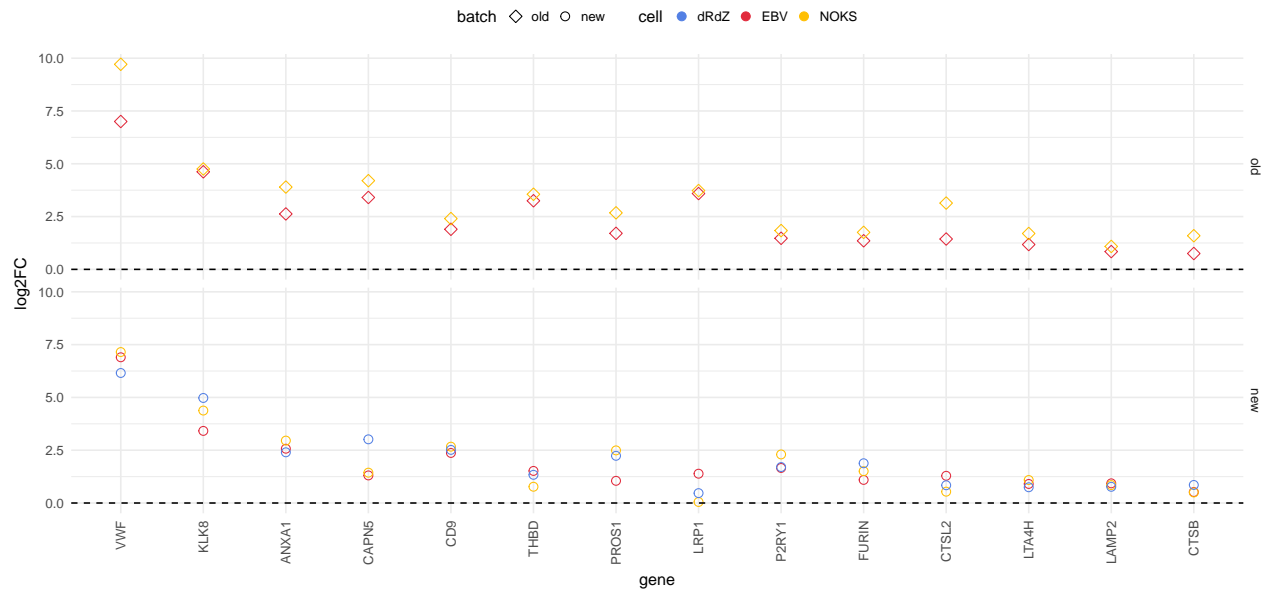# HALLMARK_HYPOXIA

HALLMARK_HEME_METABOLISM

HALLMARK_XENOBIOTIC_METABOLISM

HALLMARK_IL2_STAT5_SIGNALING
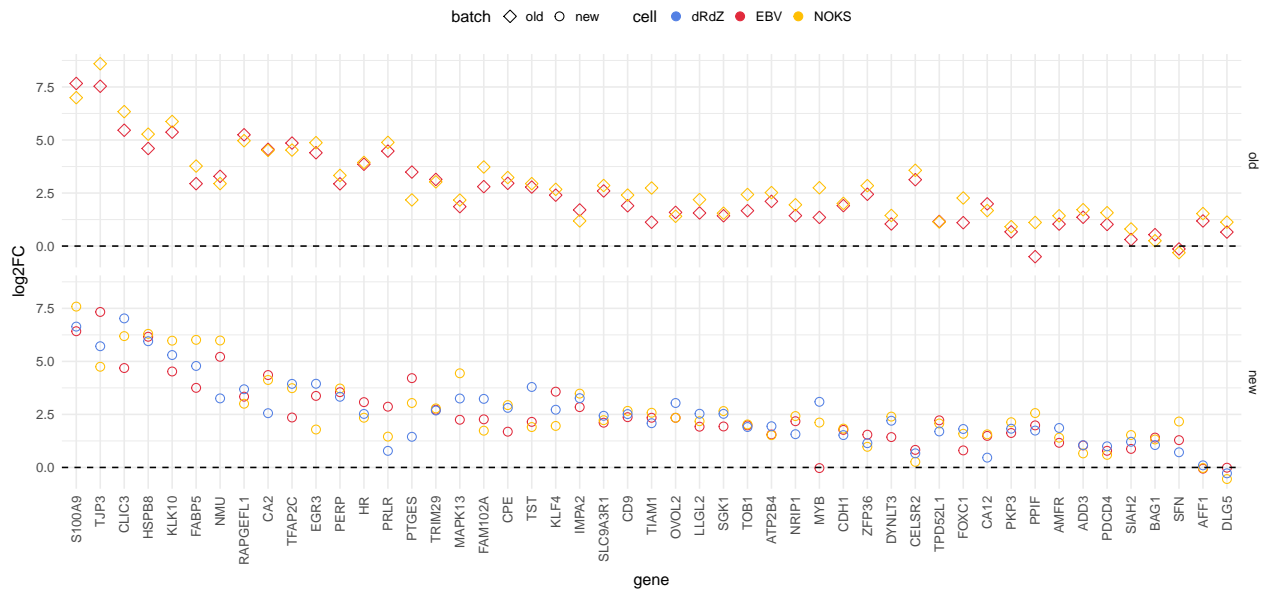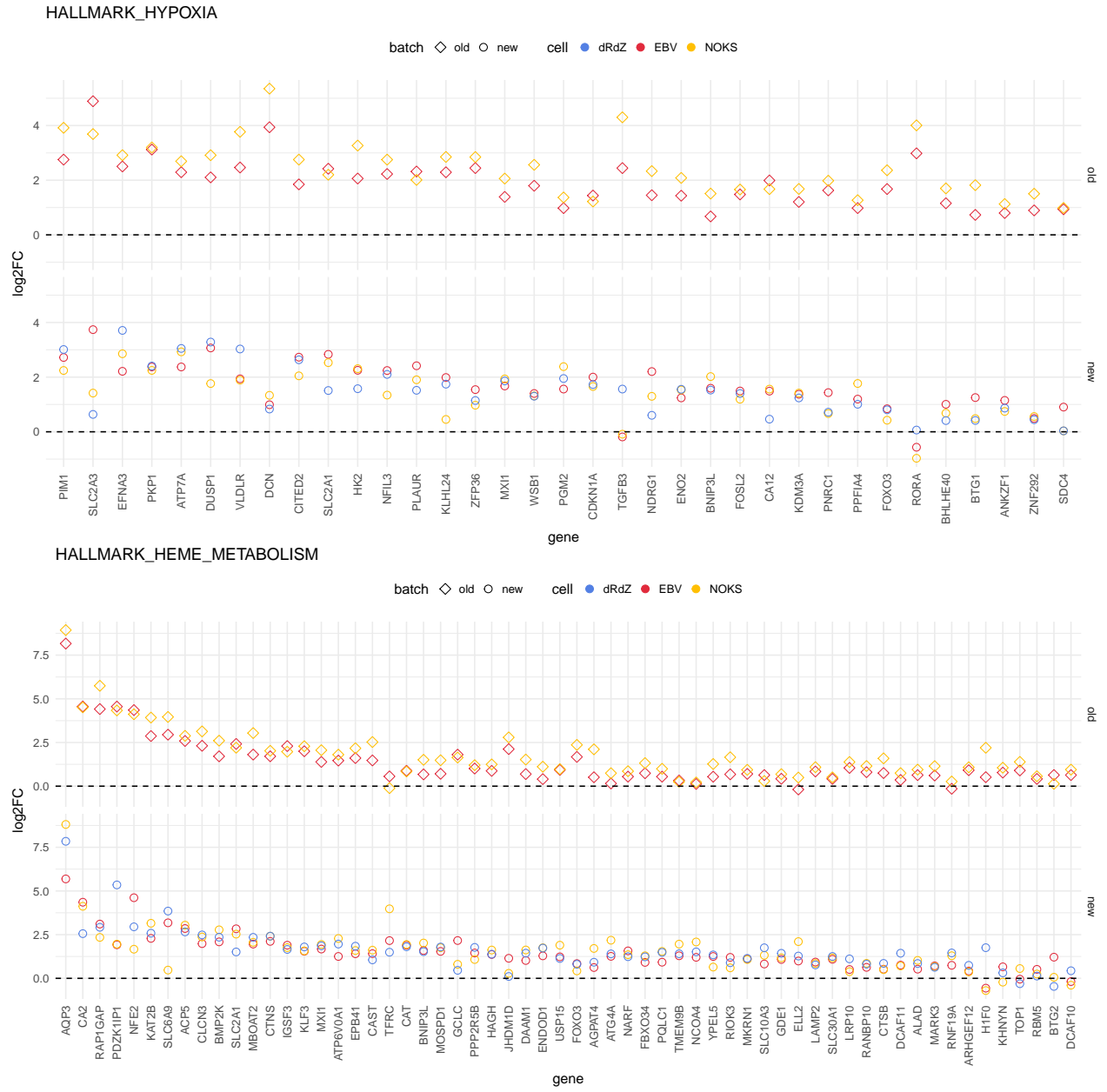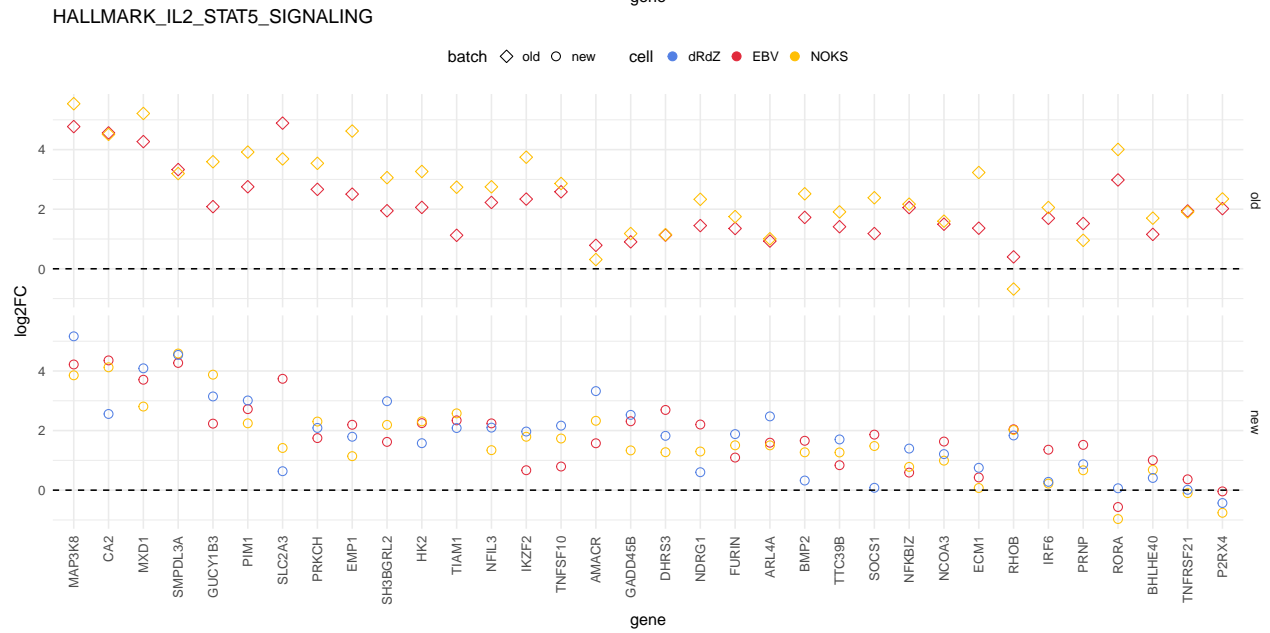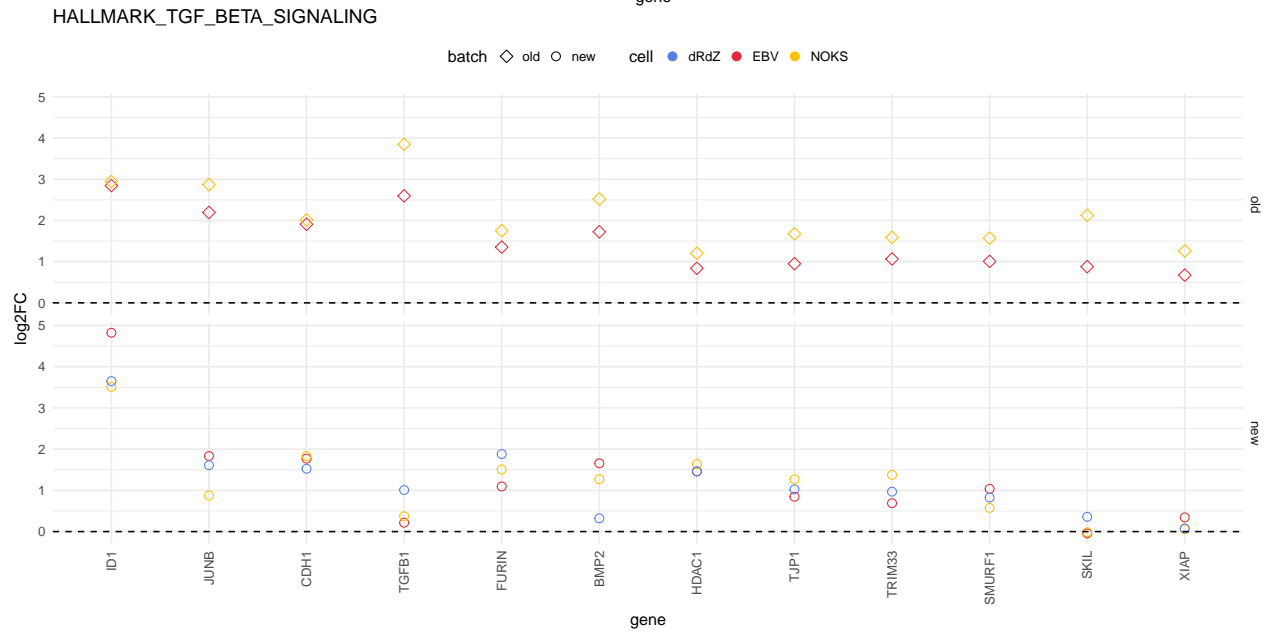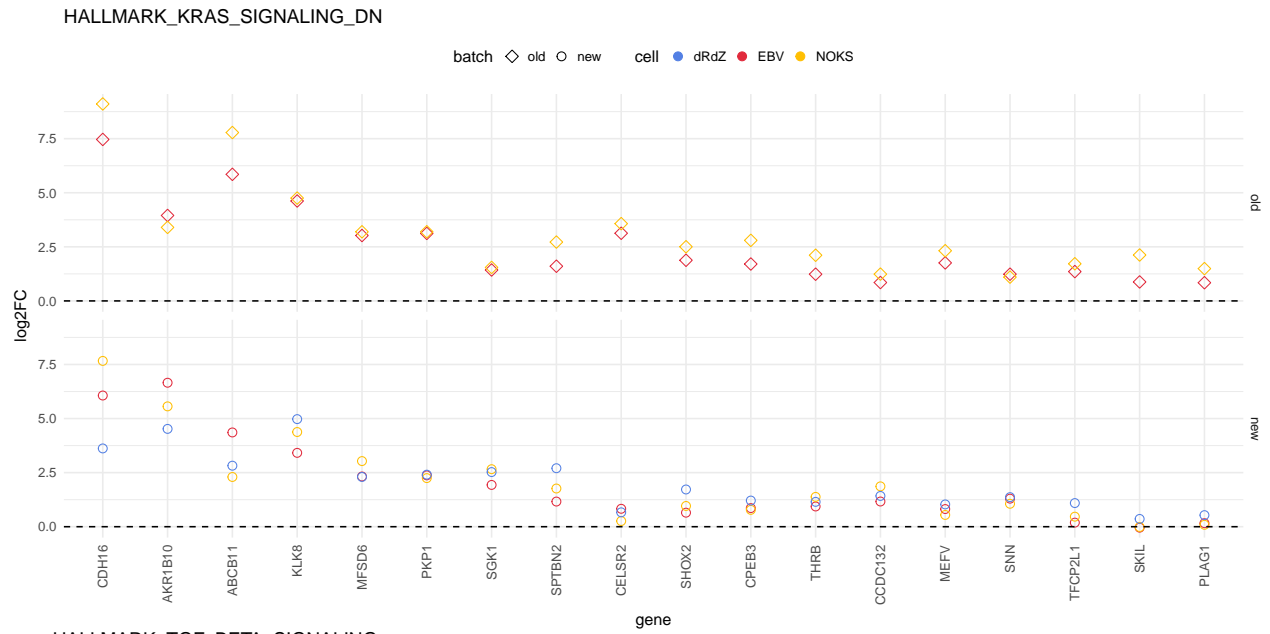
HALLMARK_P53_PATHWAY
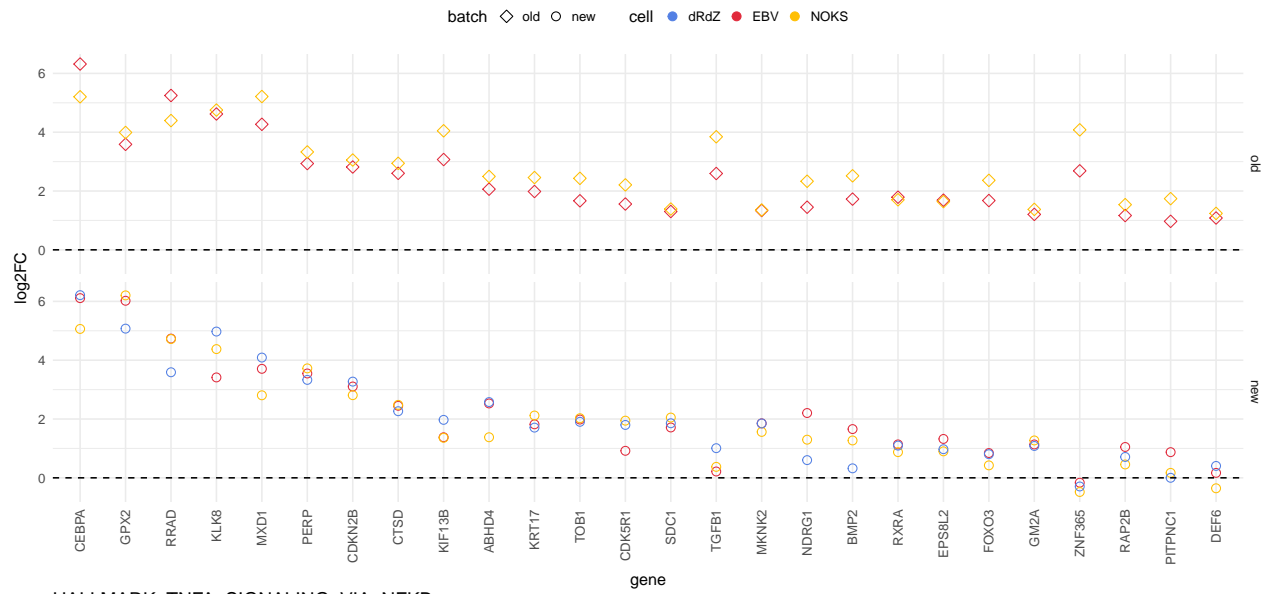
HALLMARK_TNFA_SIGNALING_VIA_NFKB

HALLMARK_OXIDATIVE_PHOSPHORYLATION

HALLMARK_COAGULATION

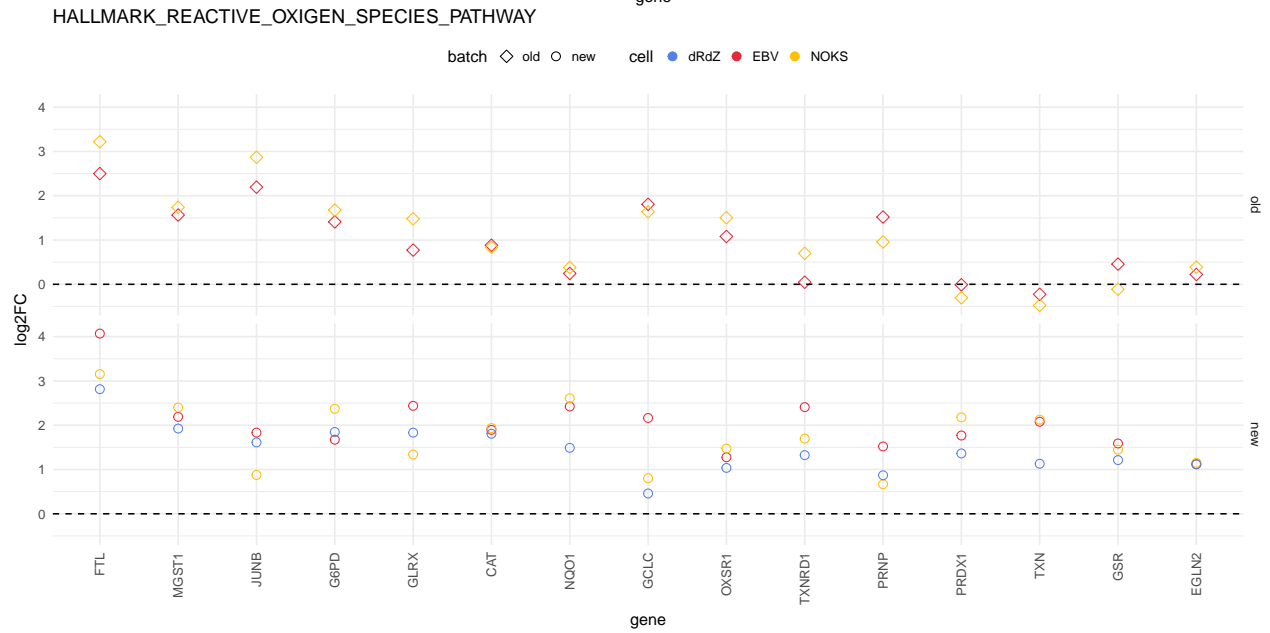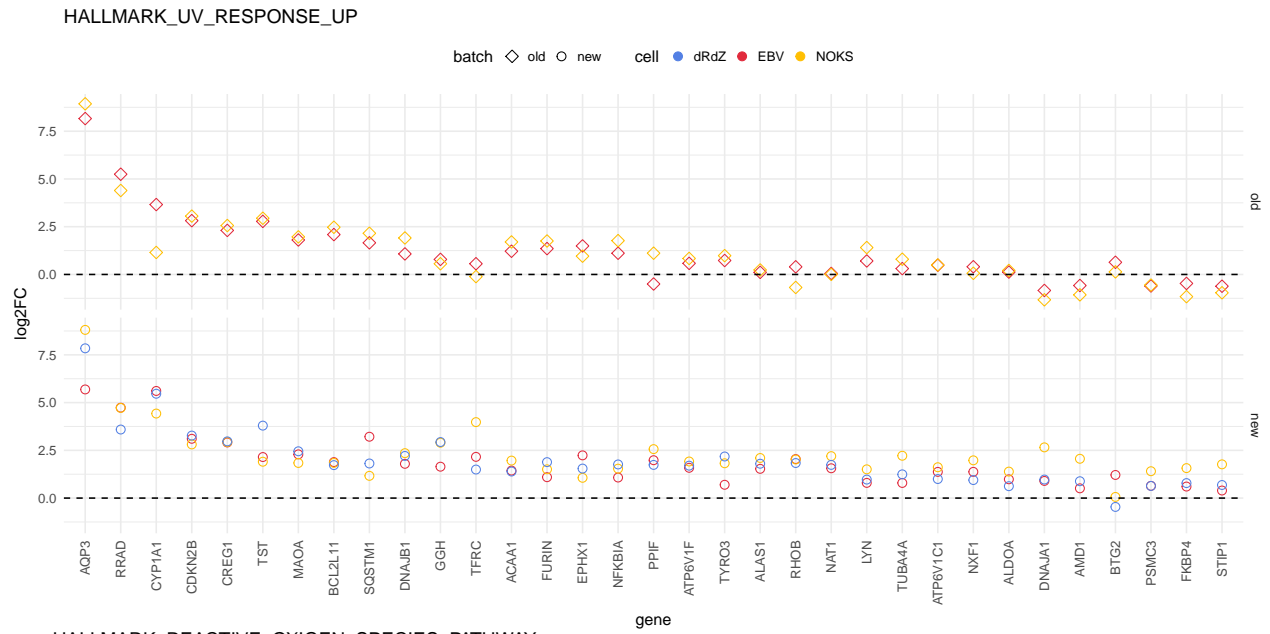# HALLMARK_UNFOLDED_PROTEIN_RESPONSE

HALLMARK_COMPLEMENT

HALLMARK_KRAS_SIGNALING_DN

HALLMARK_TGF_BETA_SIGNALING

HALLMARK_APICAL_JUNCTION

HALLMARK_MYOGENESIS

HALLMARK_UV_RESPONSE_UP

HALLMARK_REACTIVE_OXIGEN_SPECIES_PATHWAY

HALLMARK_SPERMATOGENESIS
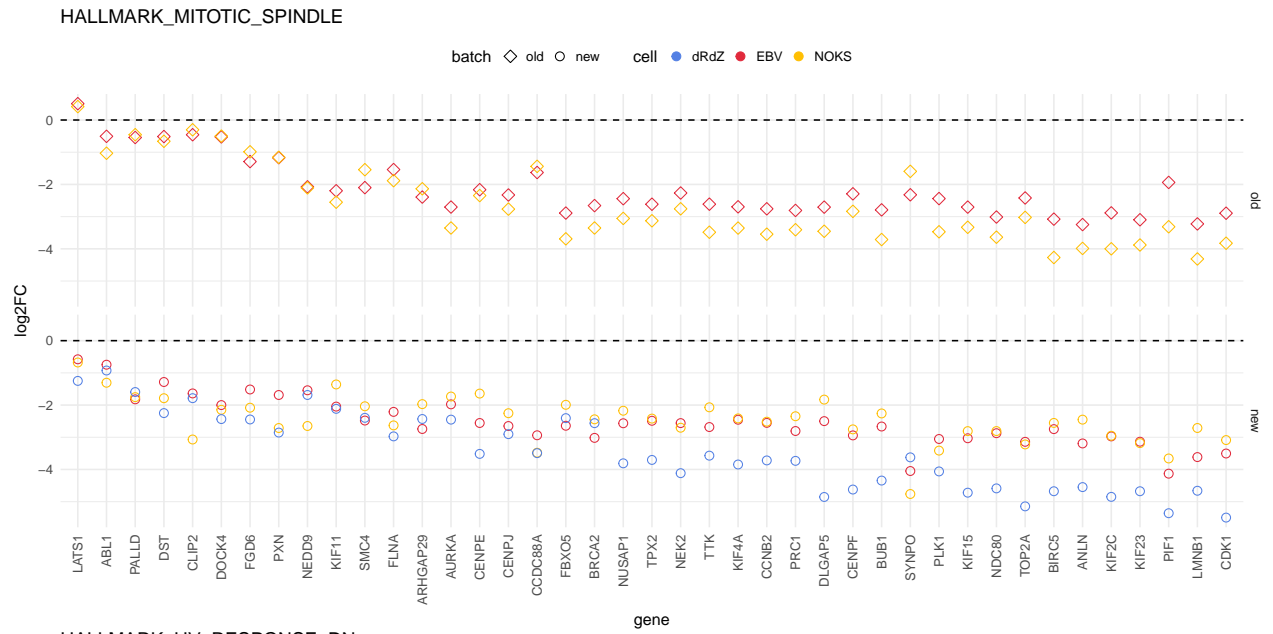

HALLMARK_INTERFERON_ALPHA_RESPONSE

19

HALLMARK_MITOTIC_SPINDLE

HALLMARK_UV_RESPONSE_DN
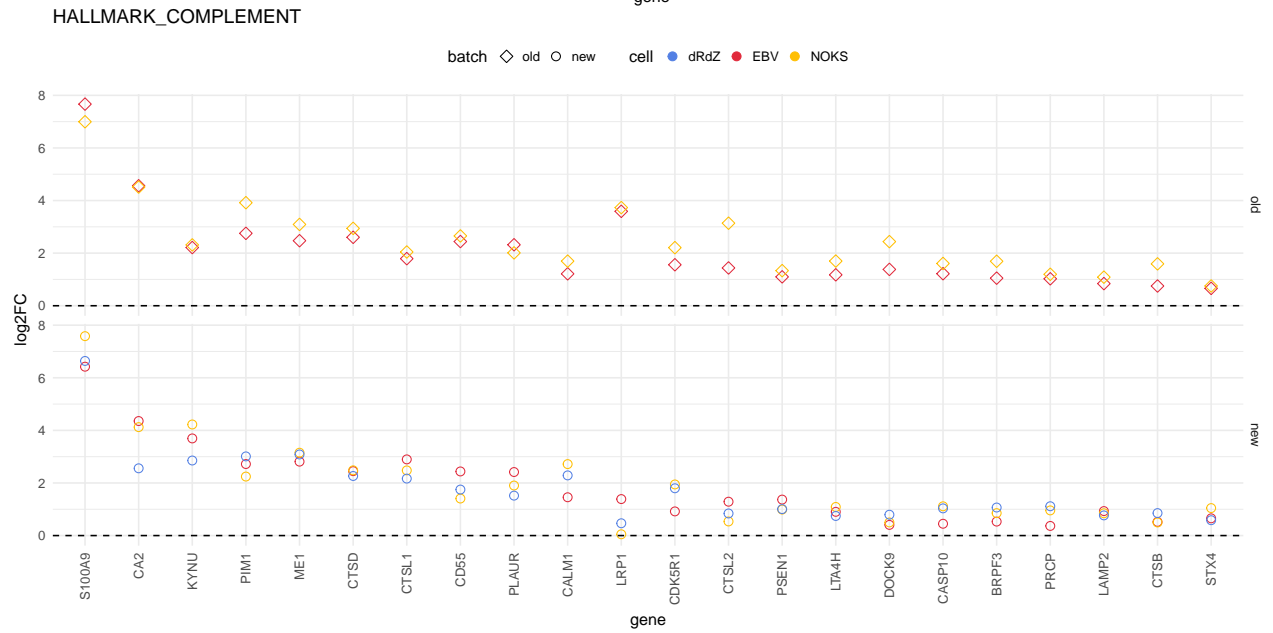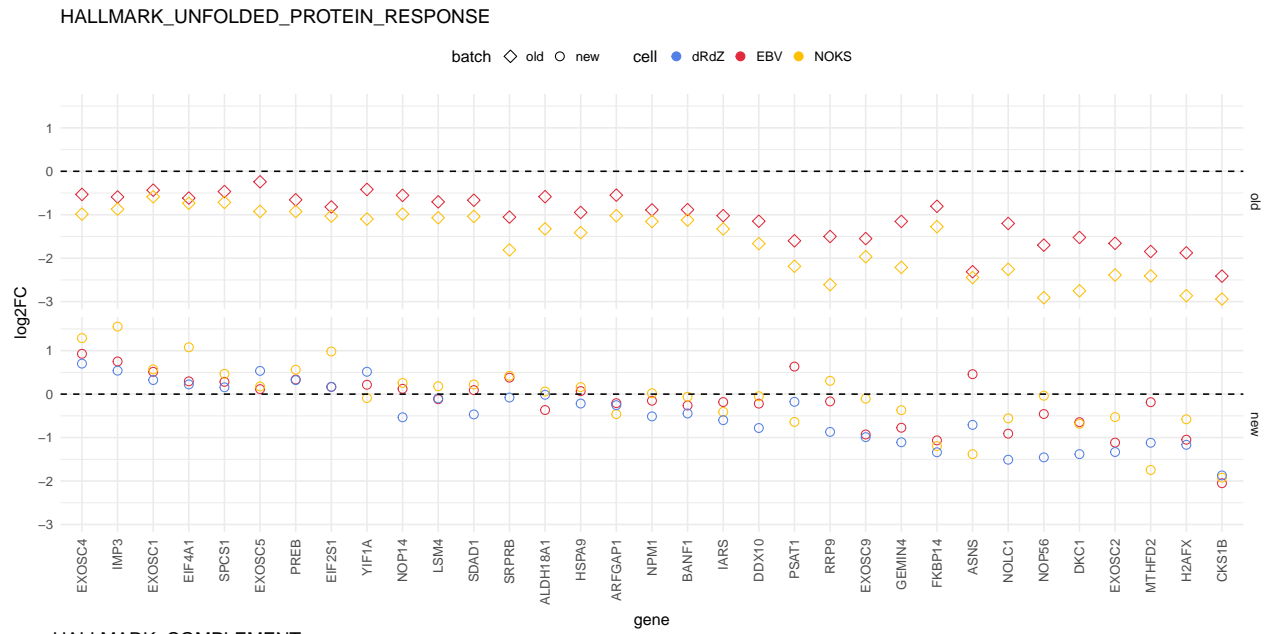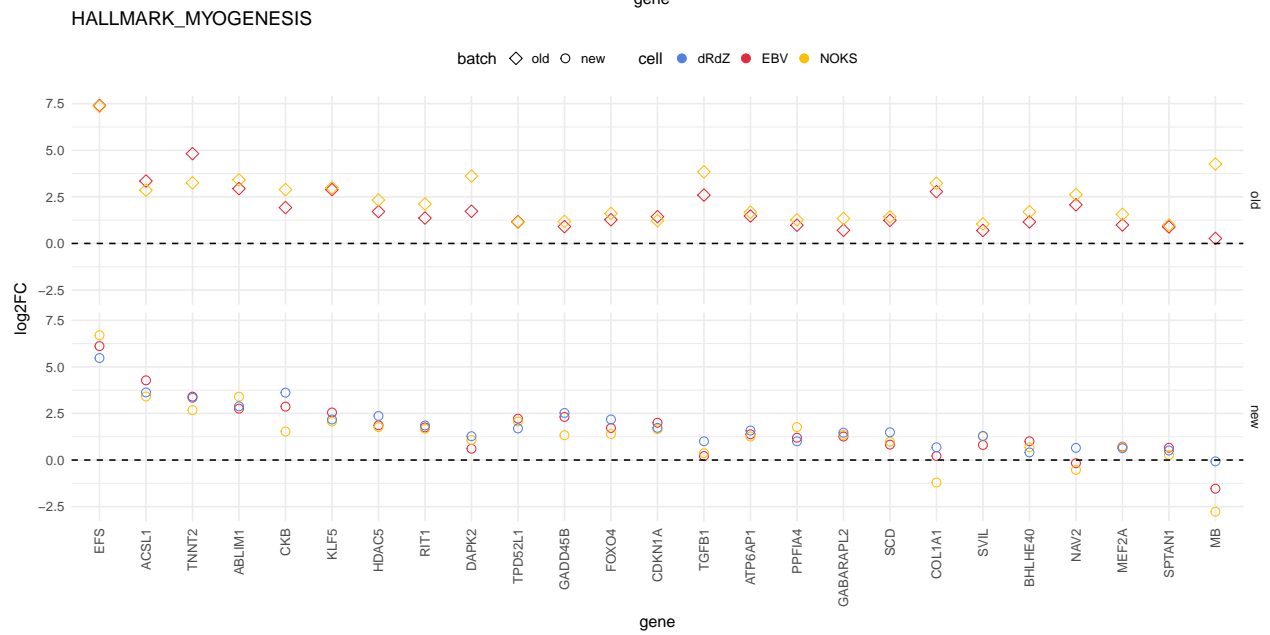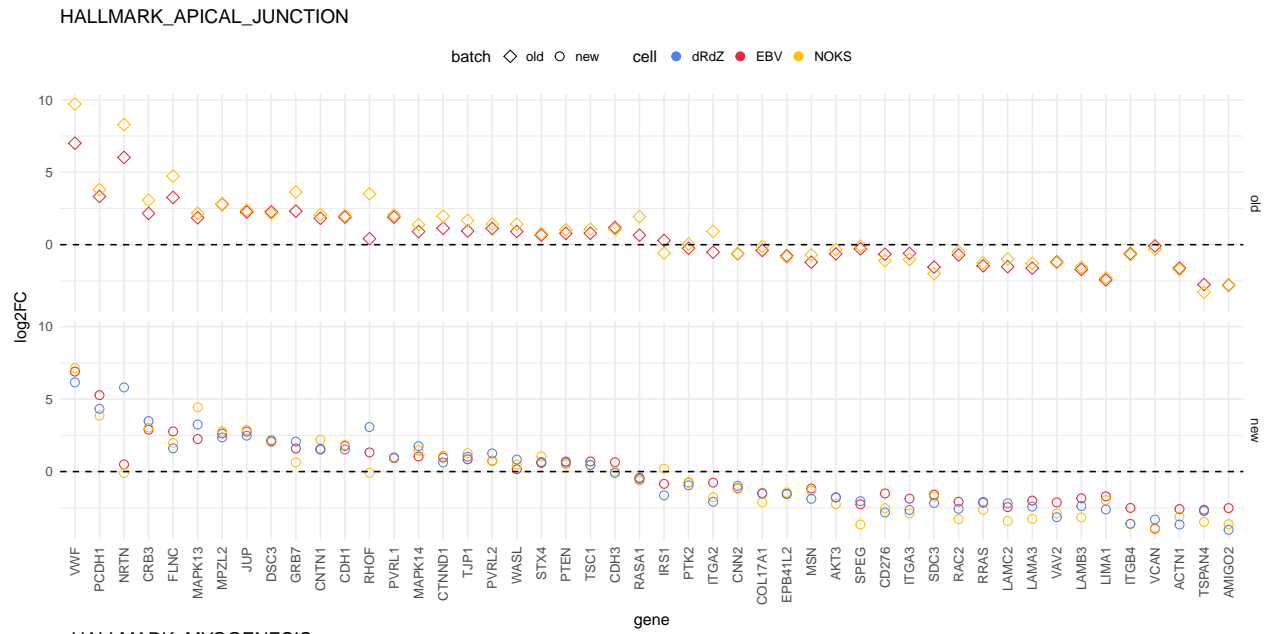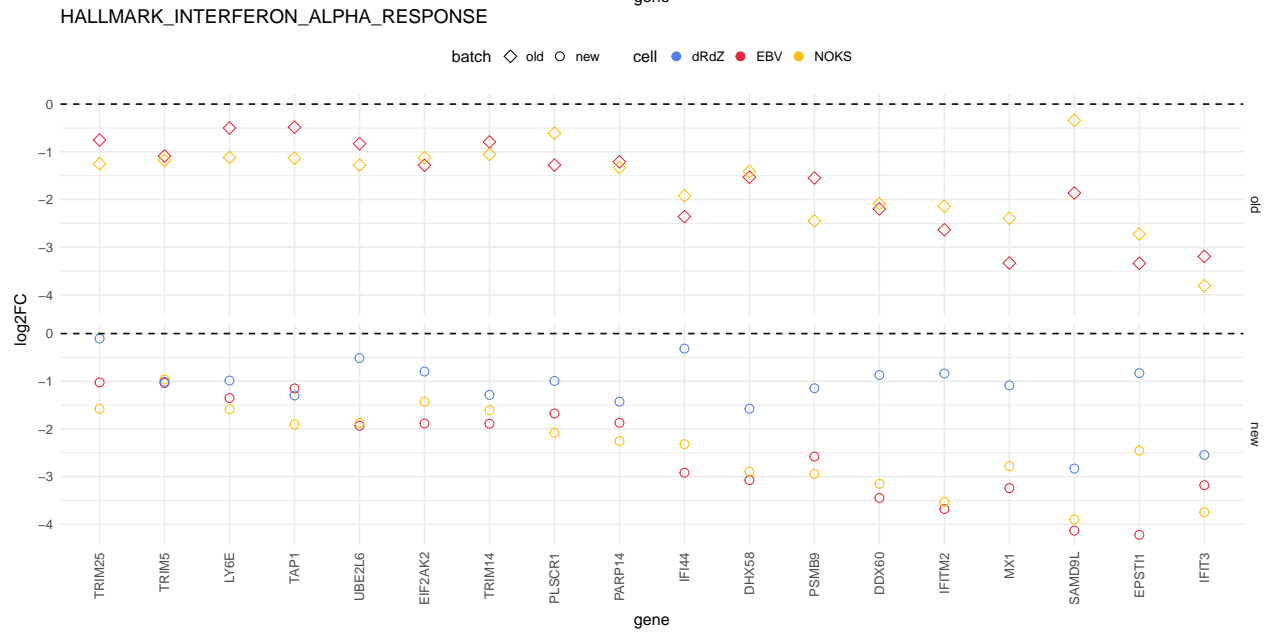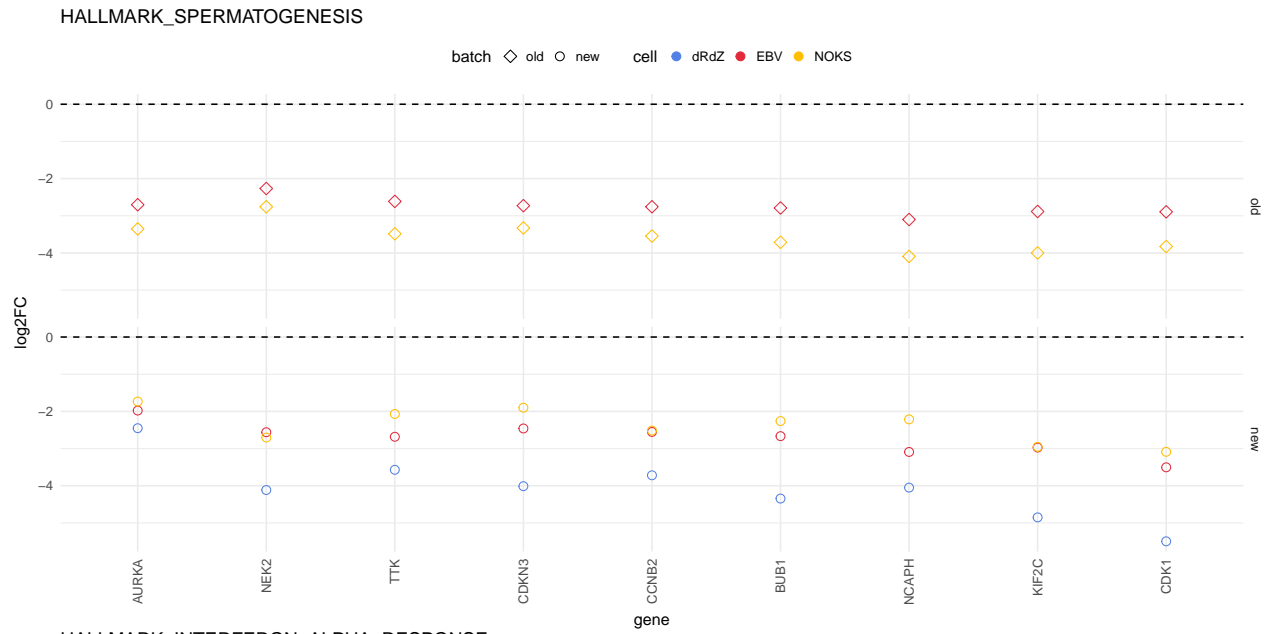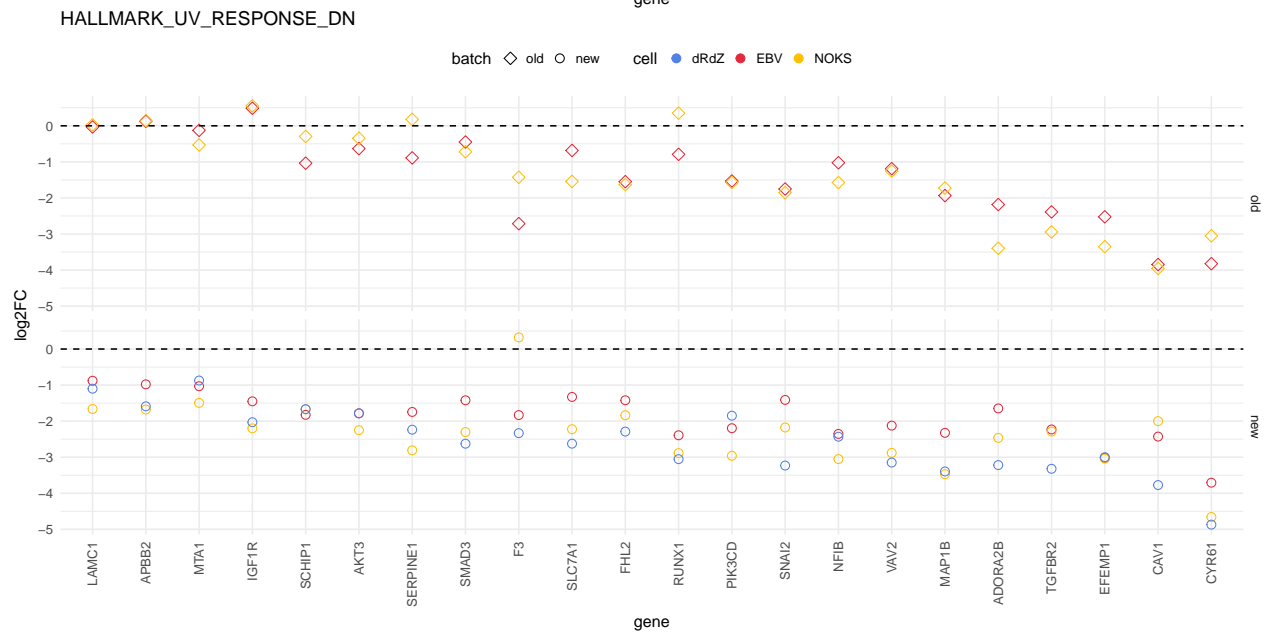
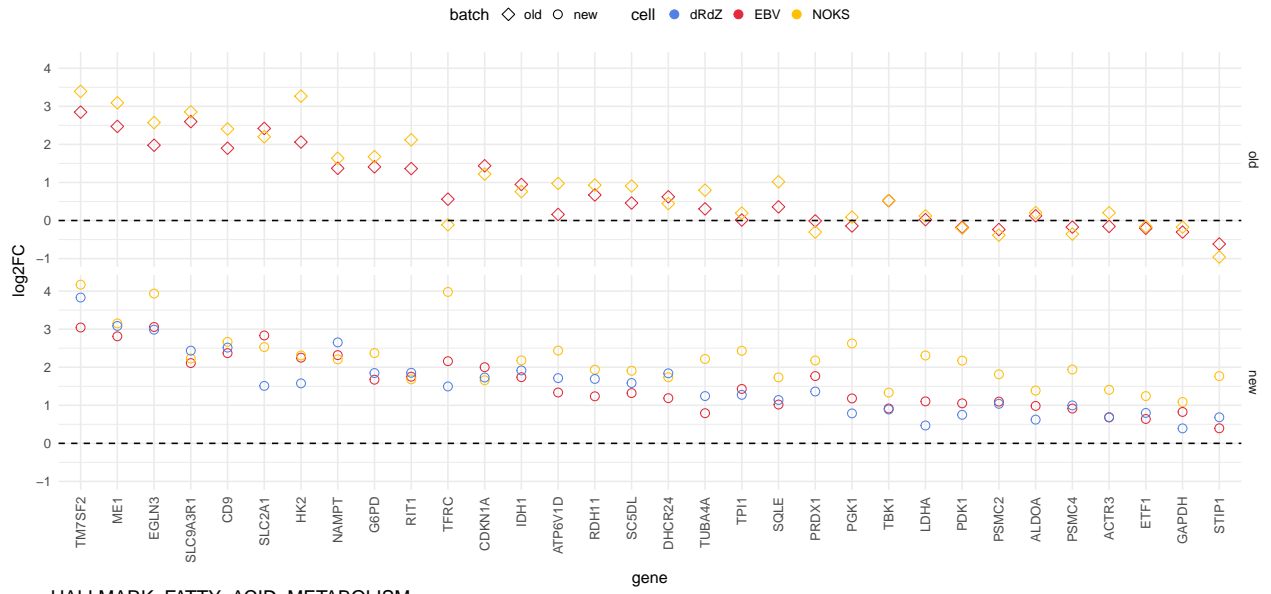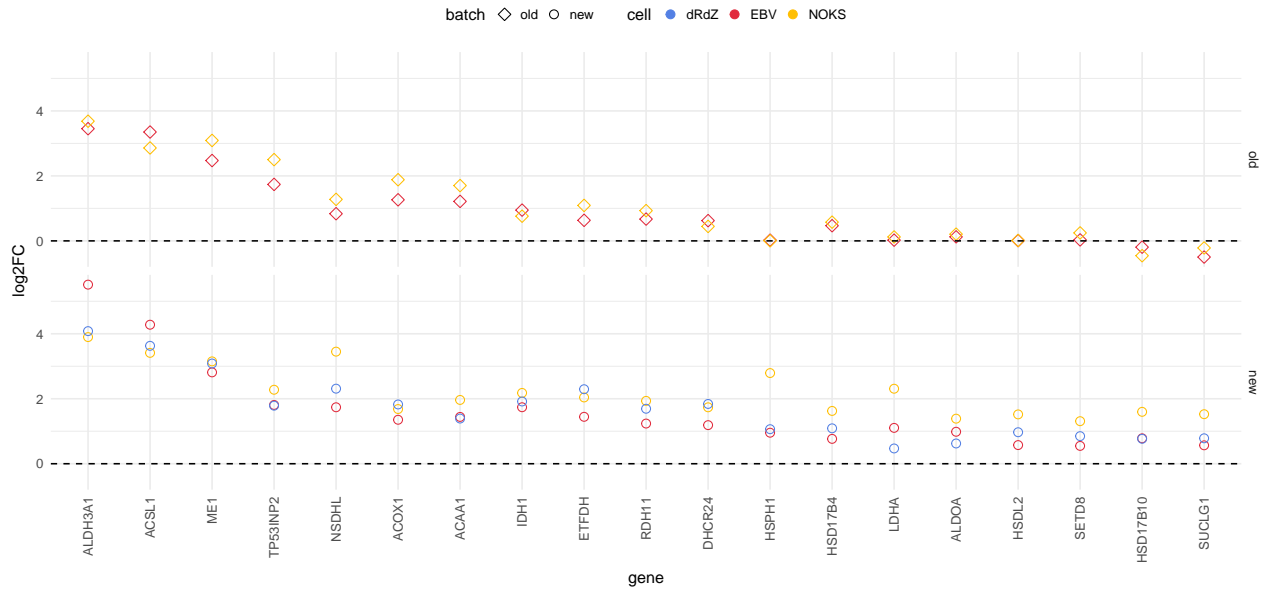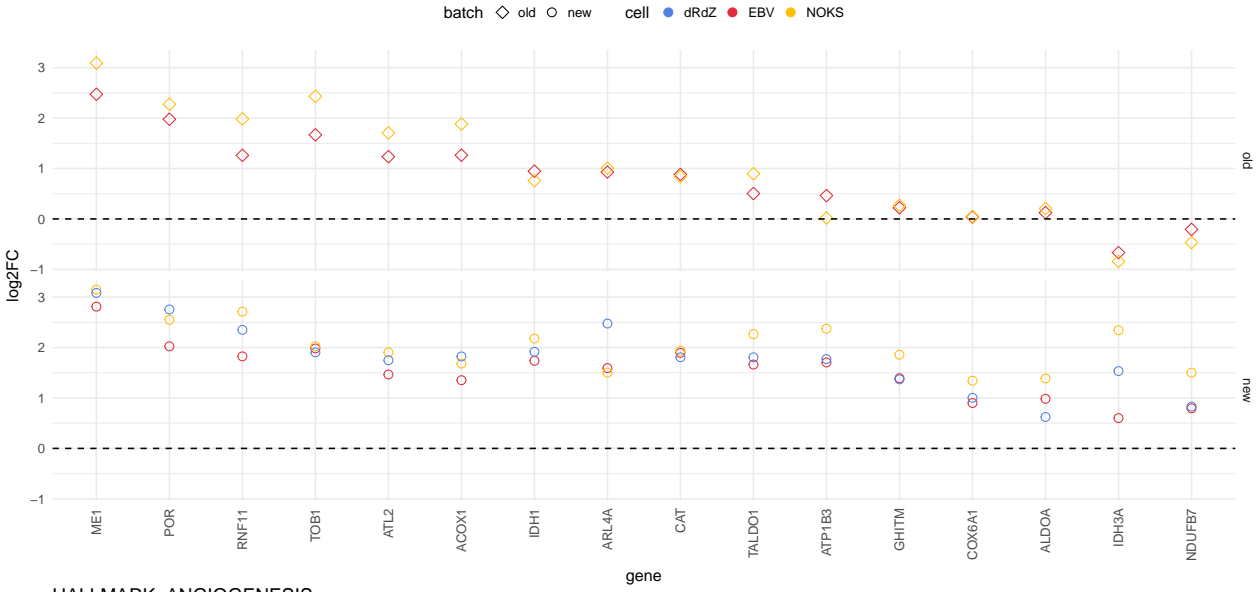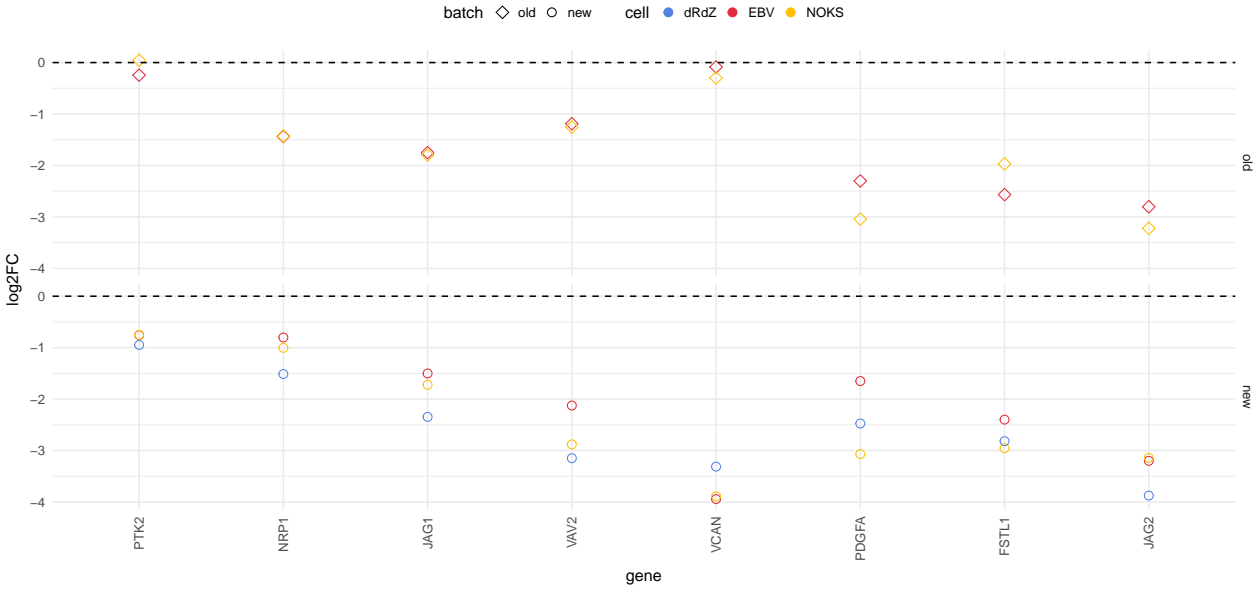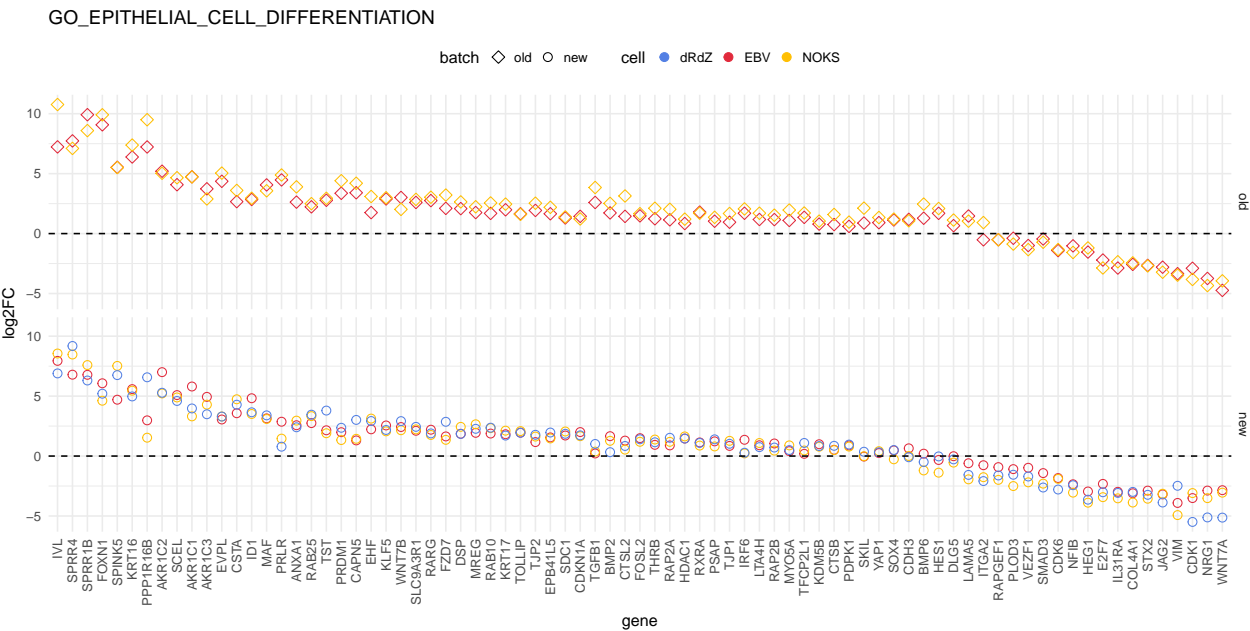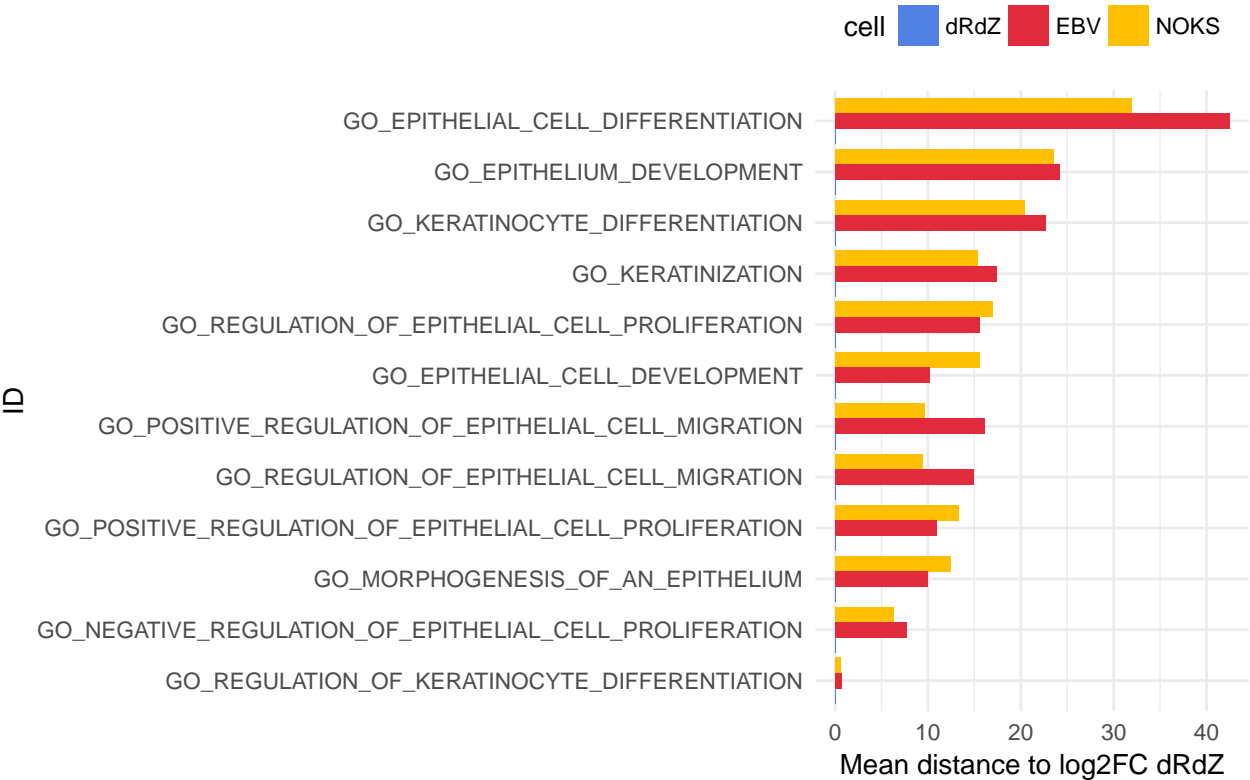## HALLMARK_MTORC1_SIGNALING
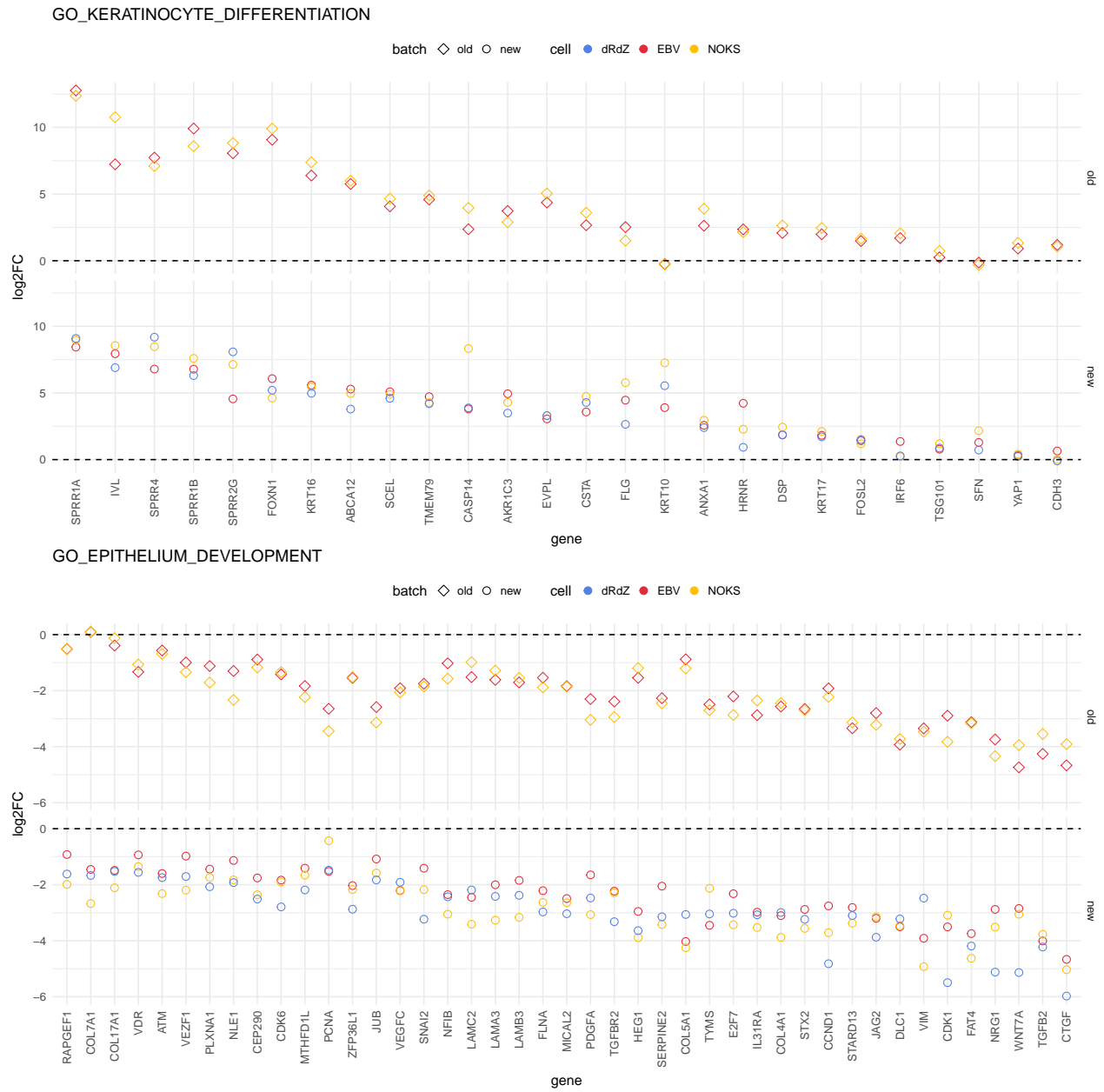

HALLMARK_FATTY_ACID_METABOLISM
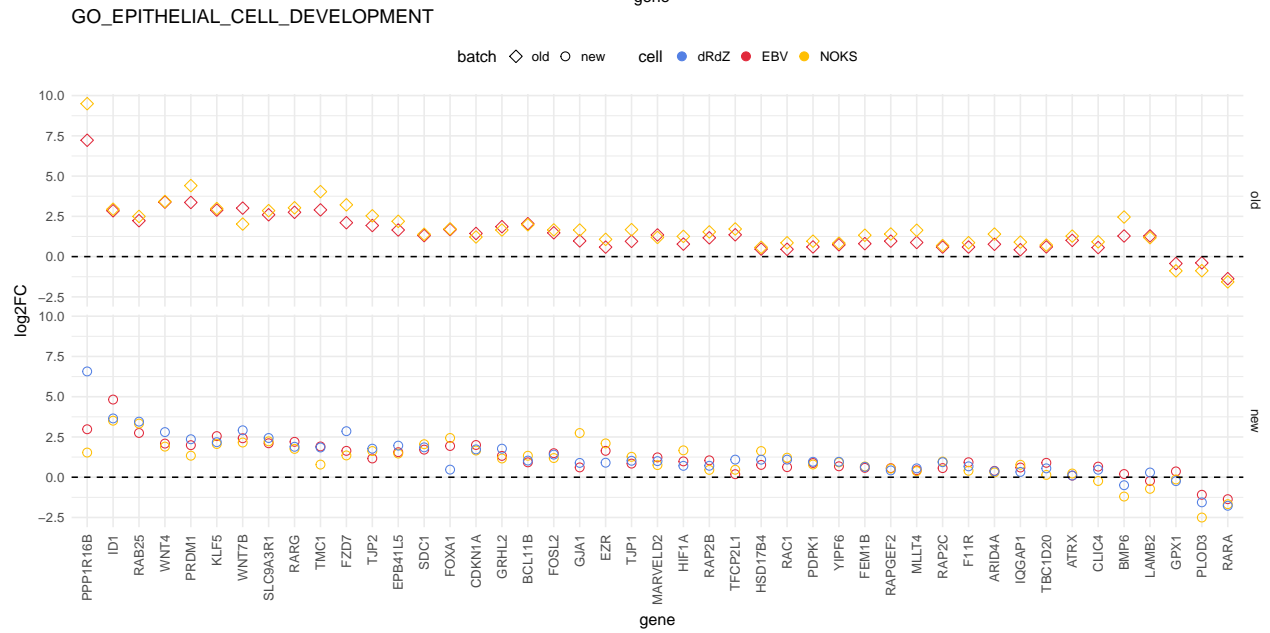
HALLMARK_ADIPOGENESIS

HALLMARK_ANGIOGENESIS

Curated pathways

GO_KERATINOCYTE_DIFFERENTIATION

GO_EPITHELIUM_DEVELOPMENT

GO_POSITIVE_REGULATION_OF_EPITHELIAL_CELL_MIGRATION

GO_EPITHELIAL_CELL_DEVELOPMENT

GO_KERATINIZATION

GO_REGULATION_OF_EPITHELIAL_CELL_PROLIFERATION

GO_POSITIVE_REGULATION_OF_EPITHELIAL_CELL_PROLIFERATION

GO_NEGATIVE_REGULATION_OF_EPITHELIAL_CELL_PROLIFERATION
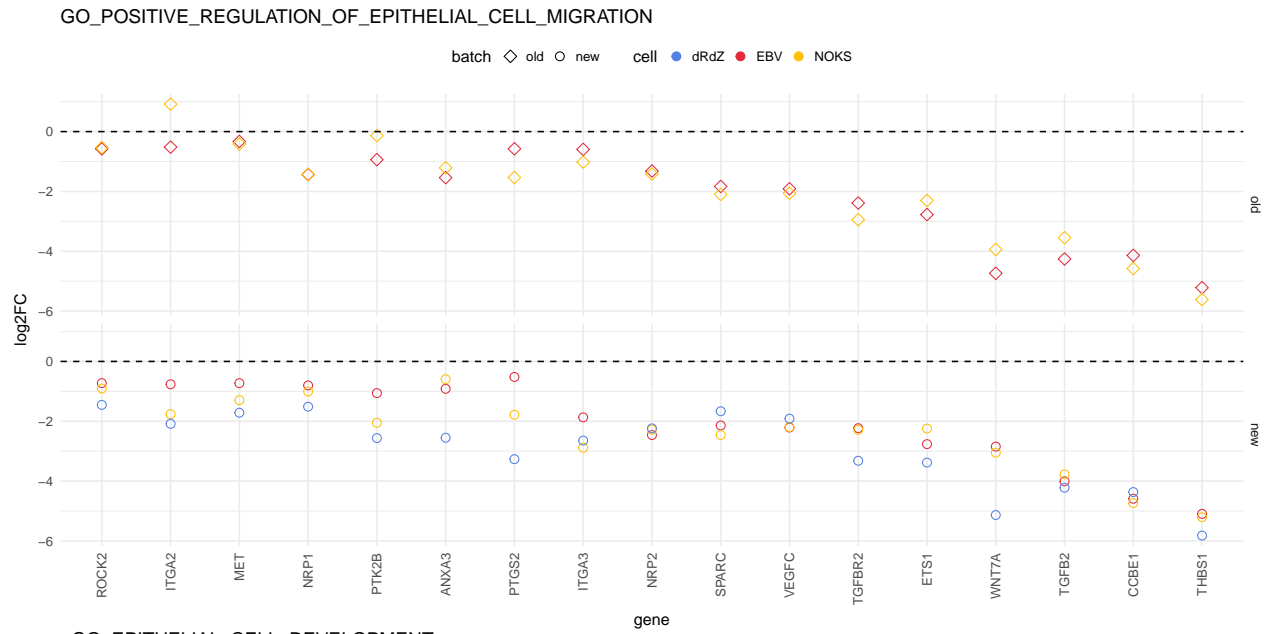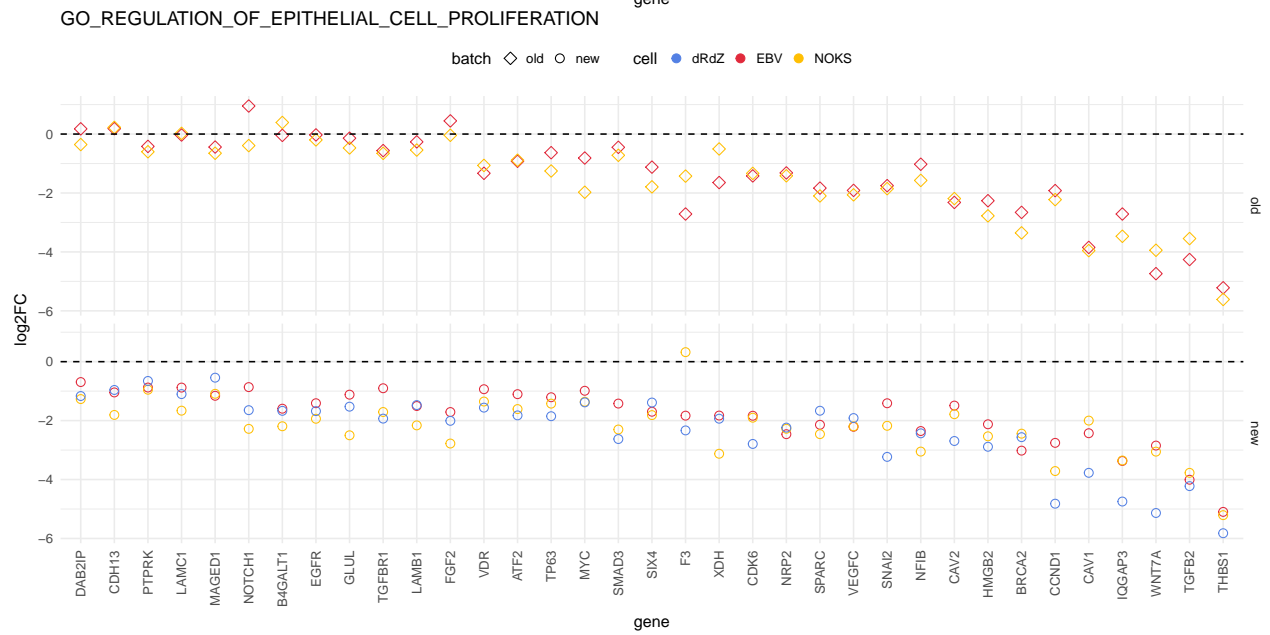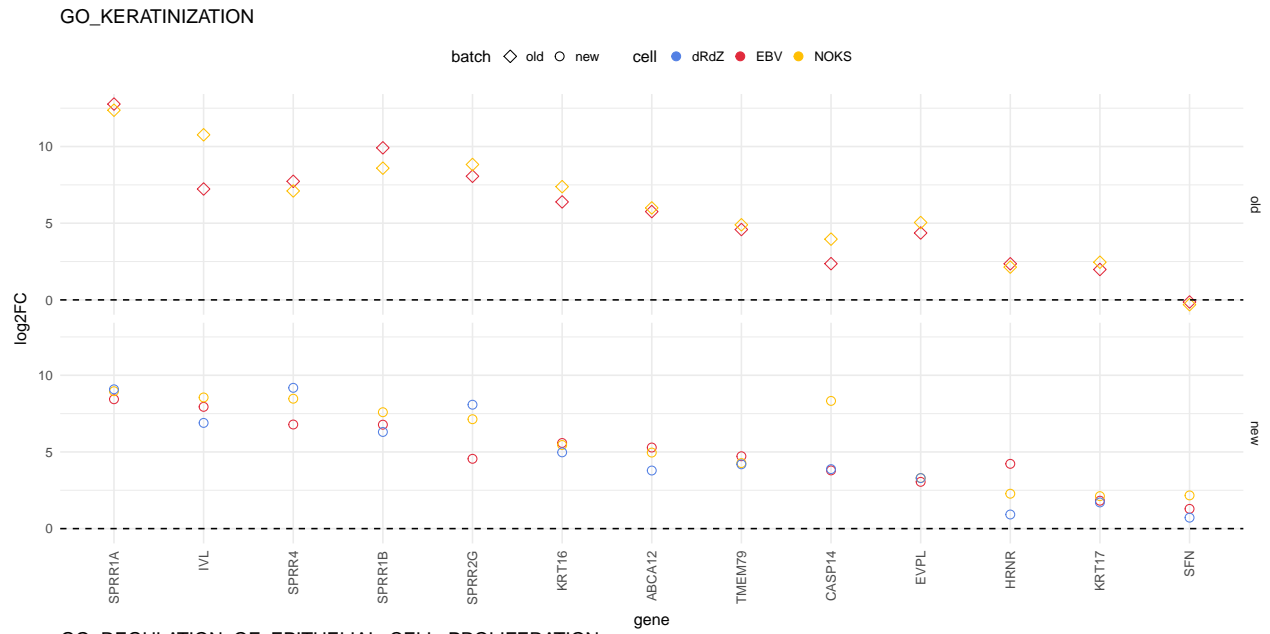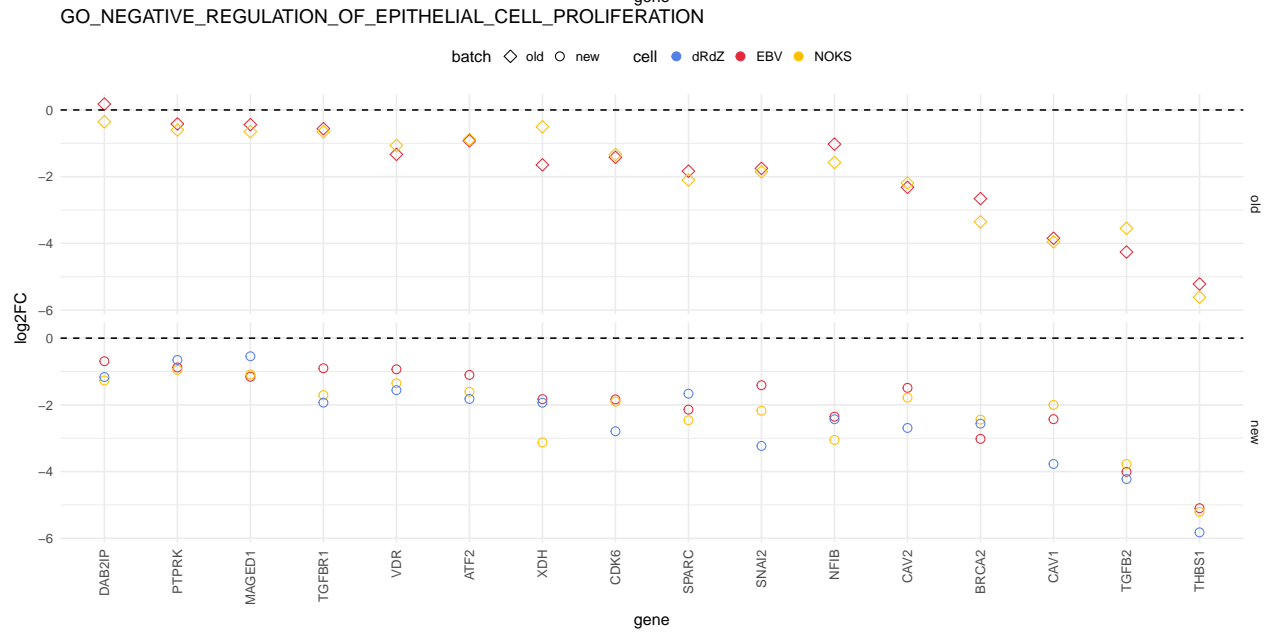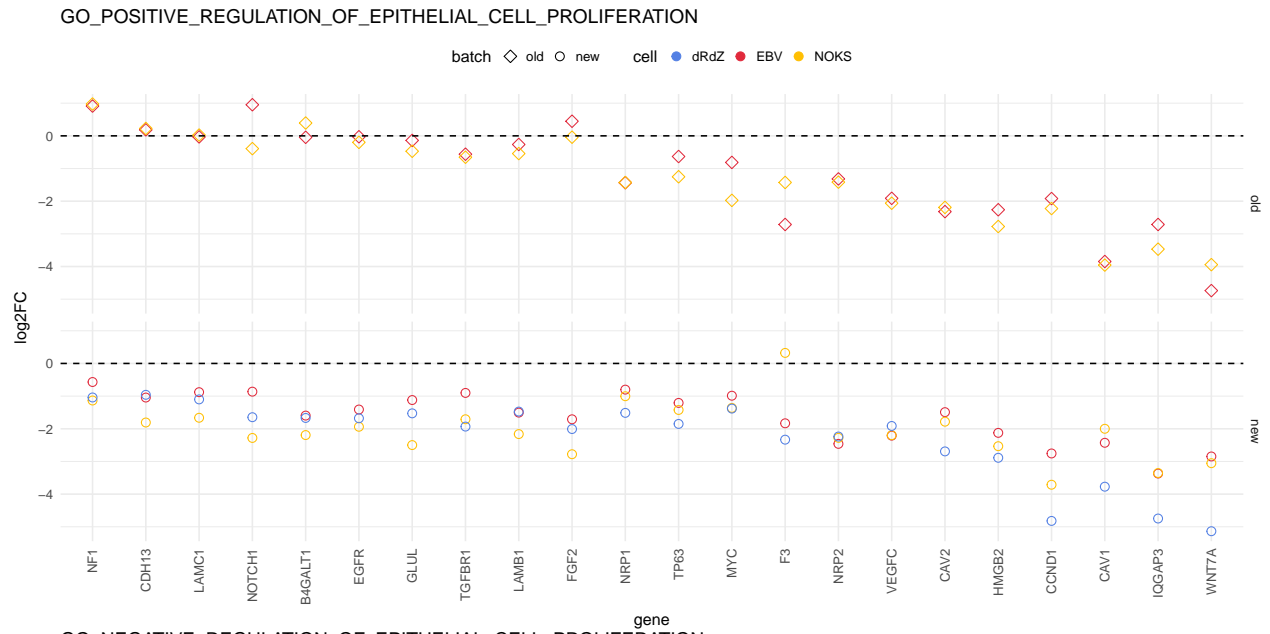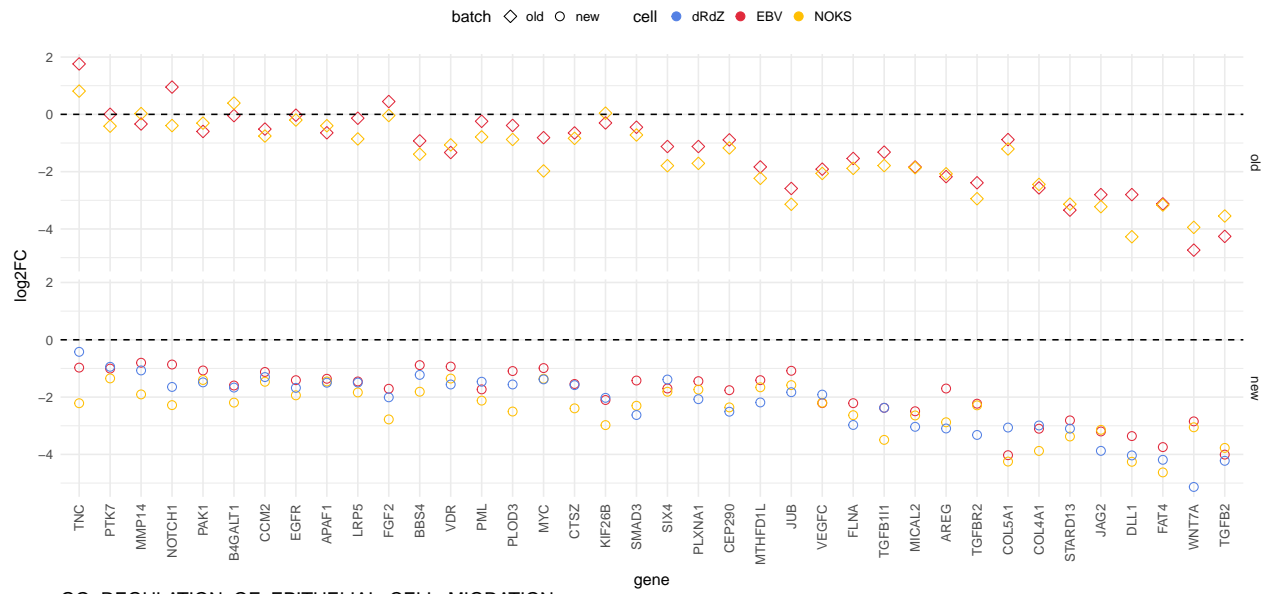
GO_MORPHOGENESIS_OF_AN_EPITHELIUM

GO_REGULATION_OF_EPITHELIAL_CELL_MIGRATION

GO_REGULATION_OF_KERATINOCYTE_DIFFERENTIATION