KIDS24: GSE247821

Standardized Differentially Binding Peaks identification pipeline v1.0.1

Lead analyst : Wojciech Rosikiewicz

Report generated on : August 29, 2024

Additional analysis description : Re-analysis of the ATACseq data from the GEO dataset with the accession number GSE247821, PubMed ID 38802751. To make the report smaller, we only include one contrast.

Reviewed by:..... Date:.....

Thank you for using the Center for Applied Bioinformatics (CAB) for your analysis. We have provided the results of our standardized pipeline below for your review. After reviewing these results, please feel free to contact us at cab.helpdesk@stjude.org with "[Epigenetics]" in email subject, with any questions or concerns. We are also looking forward to discuss any follow-up, customized analyses.

–Center for Applied Bioinformatics Team

TABLE OF CONTENTS

LIST OF FIGURES

SECTION 1

OVERVIEW

Project: KIDS24
Task ID: CAB
PI: Gang Wu
Project Lead(s): Wojciech Rosikiewicz
Department: CAB

1.1 CAB Analysis Team

Lead Analyst(s): Wojciech Rosikiewicz
Group Lead: Beisi Xu
Contact E-mail: wojciech.rosikiewicz@stjude.org
CAB Pipeline: Standardized Differentially Binding Peaks identification pipeline v1.0.1

1.2 Results directory

```
HPCF: /research_jude/rgs01_jude/groups/cab/projects/automapper/common/wrosikie/DEV_
↪projects/gptHelper/KIDS24_publicDataPreprocessing/GSE247821_ATACseq_PIK3CA_hotspots_
↪breat_cancer/diffPeak/report_oneContrast

Mac: smb://jude.stjude.org/groups/cab/projects/automapper/common/wrosikie/DEV_projects/
↪gptHelper/KIDS24_publicDataPreprocessing/GSE247821_ATACseq_PIK3CA_hotspots_breat_cancer/
↪diffPeak/report_oneContrast

PC: \\jude.stjude.org\groups\cab\projects\automapper\common\wrosikie\DEV_projects\
↪gptHelper\KIDS24_publicDataPreprocessing\GSE247821_ATACseq_PIK3CA_hotspots_breat_cancer\
```

(continues on next page)

(continued from previous page)

```
↪diffPeak\report_oneContrast

PCZ: Z:/ResearchHome/Groups/cab/projects/automapper/common/wrosikie/DEV_projects/
↪gptHelper/KIDS24_publicDataPreprocessing/GSE247821_ATACseq_PIK3CA_hotspots_breat_cancer/
↪diffPeak/report_oneContrast
```

The results presented in this report are all based on the human reference genome (**hg38**; GRCh38.p12), and gene annotations were based on Gencode v31 (PMID: 30357393).

SECTION 2

METHODS

Interactive version of this report, which is html-based, might be accessed with `GSE247821.InteractiveReport.20240829.html` file , under *Results directory*.

While analytical methods and laboratory technologies continue to evolve, various Next Generation Sequencing analysis, including ChIP-seq, ATAC-seq or Cut-and-Run, has now matured to the point where a number of best practices have emerged. The CAB has developed standardized analysis protocols, incorporating community best practices as well as accumulated in-house expertise. The procedures used to generate results in this report are outlined below, with additional details available in referenced publications and linked CAB Wiki pages.

2.1 ATAC-seq data processing

Preprocessing, and mapping of the reads from ATAC-seq experiments is using the pipeline described in detail on Wiki page of Automapper. Briefly, raw reads in fastq format were processed with Trim-Galore tool (v0.4.4, Krueger F. (2012), Availible online), in order to remove potential adapters and quality trim 3' end of reads with cutadapt program (DOI:10.14806/ej.17.1.200), followed by FastQC analysis (Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online). A quality score cutoff of Q20 was used and the first 15bp of each reads were also clipped to reduce the GC bias. Next, reads were mapped to human reference genome (**hg38**; GRCh38.p12) with BWA mem (0.7.17-r1188, PMID: 19451168), then converted to BAM format and deduplicated with fq2bam (v3.0.0.6). Subsequently, uniquely mapped properly paired reads were extracted with samtools (v1.2, PMID: 19505943) and fragments were extracted with bedtools (v2.24.0, PMID: 20110278).

2.2 Reproducible peaks identification

Peaks called from various samples might not be reproducible due to the natural and technical variation. Before any downstream analysis, the peaks from those individual samples, combined in groups, were compiled in reproducible peak sets following the "Reproducible Sharp" strategy, visualized on the graph below. More detailed description might be found at our Wiki page.

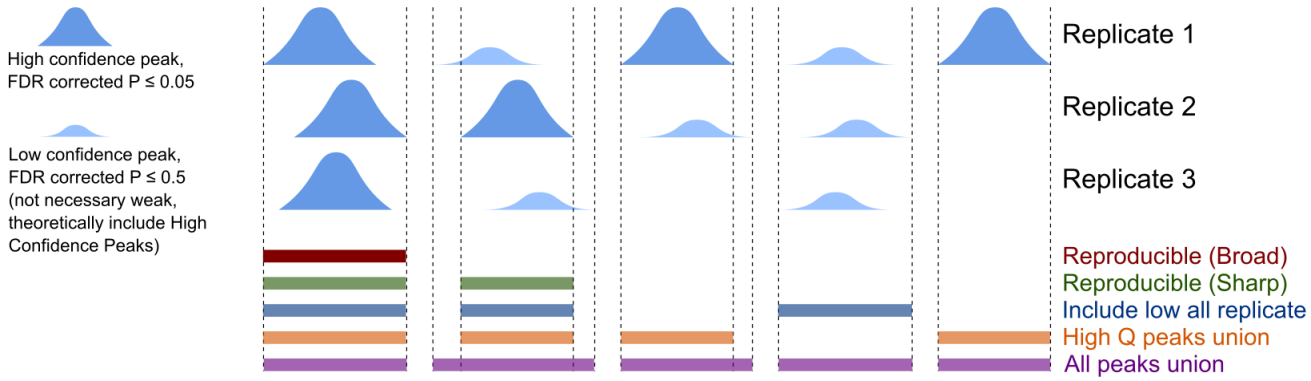

Fig. 2.1: Alternative strategies for identification of reproducible peaks from experiments with replicates

Note: Reproducible peaks were generated independently for each of the grouping levels described in Samples metadata section. The assumption for this level-wise separation is that these levels would generally reflect different subgroups of the IP experiment(s) conducted (e.g. lvl #1 would consist of Active Histone marks, while lvl. #2 would have IPs targeted at transcription factor).

2.3 Peak annotation

Standardized peak annotation can be separated into two independent stages: (1) Annotation of peaks to genes, and (2) annotation of peaks to genomic features / genomic context. Please note that both of these approaches are computed independently.

2.3.1 Peak annotation to genes

Regions were first assigned to genes, with reference gene annotation from Gencode v31 (PMID: 30357393), if they overlap the gene promoters, which was done with bedtools (v2.24.0, PMID: 20110278), one region could be assigned to multiple genes. These are considered **putative promoter-related regions**. By default, the promoter region is here defined as TSS +- 2 kbp (see the Fig. 2.2 schematics of peak to gene annotation). Then, regions not assigned to any promoter, are assigned to gene if their distance to gene's TSS were within a threshold of +- 50 kbp (by default), excluding the promoter region (see the Fig. 2.2 schematics of peak to gene annotation), one region could be assigned to multiple genes. These are considered **putative enhancer-related regions**. At last, we also report the closest gene by TSS and their distance to the region, one region would only be assigned to one gene.

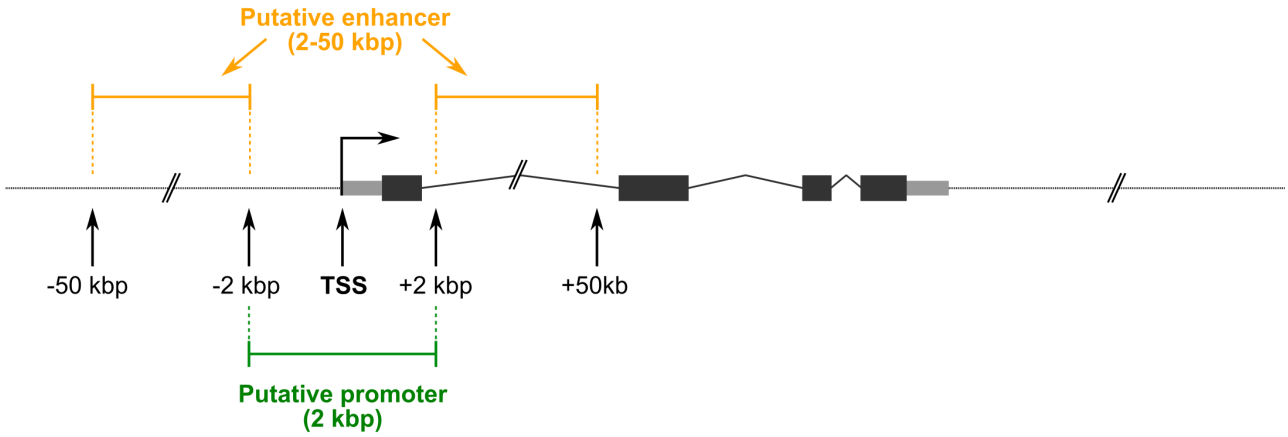

Fig. 2.2: Prioritization order for peak to genomic context annotation

2.3.2 Peak annotation to genomic context

Regions (i.e. reproducible peaks) were overlaped with genomic context one-by-one, with the following prioritization order, as visualized on the Fig. 2.3:

1. **Promoter.Up** - Region up to 2 kbp upstream from the TSS.
2. **Promoter.Down** - Region down to 2 kbp downstream from the TSS.
3. **Exon** - all exons, from any isoform.
4. **Intron** - all introns, from any isoform.
5. **TES (transcription end sites)** - region spanning TES (transcription end site), also known as TTS (transcription termination site), +- 2 kbp.
6. **Dis5 (5' distal regions)** - region up to 50 kbp upstream from TSS, excluding promoter region.
7. **Dis3 (3' distal regions)** - region down to 50 kbp downstream from TES.
8. **Intergenic** - all remaning regions, excluding alternative chromosomes that lack any known reference annotation.

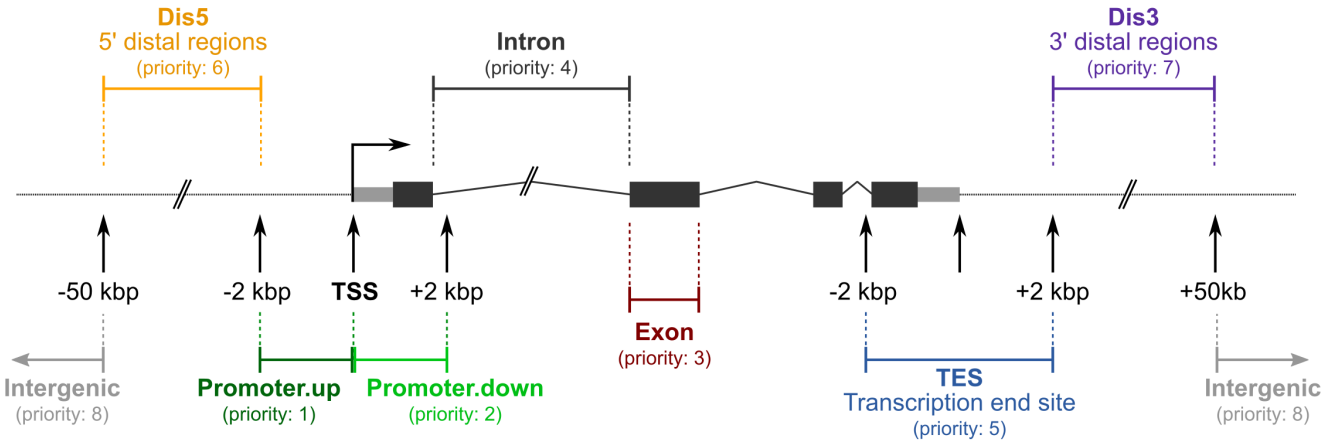

Fig. 2.3: Prioritization order for peak to genomic context annotation

The one-by-one overlap was computed based on reference gene annotation from Gencode v31 (PMID:

30357393) with pybedtools (v0.8.1; PMIDs: 21949271, 20110278), requiring at least 1 bp of overlap to assign the genomic context to peak. In order to ensure that each region would be assigned to only one genomic feature, once the peak was annotated, it was excluded from the pool, and subsequently all remaining/unassigned peaks were attempted to be overlapped with the next, lower prioritization group of the genomic context(s).

Overall this peak-to-context annotation could be informative as sanity check for the analyzed data, e.g. H3K4me3 reproducible peaks should be enriched for Promoter while H3K36me3 would be more for Intron/Exon.

Moreover, to provide the theoretical background distribution of the peaks with the approach applied here, the human hg38 genome was converted into 3,209,513 windows/bins, each 1 kbp long. Next, these regions were annotated, resulting in the following distribution:

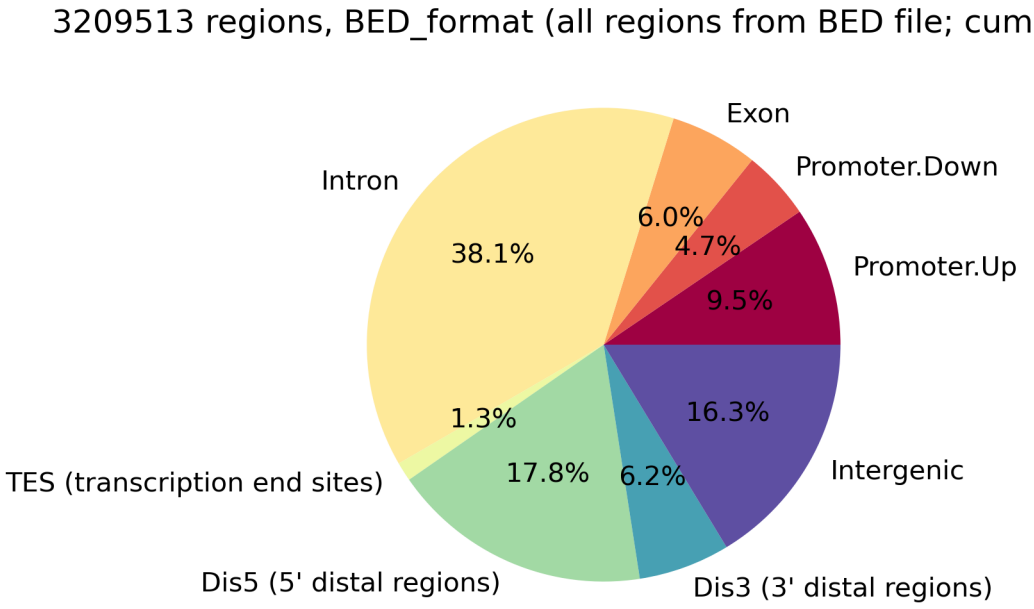

Fig. 2.4: Theoretical background distribution of peak to context annotation

The exact number of regions annotated with each genomic context:

Category:	Number of regions
All regions from hg38 genome with 1 kbp bin window	3,209,513
Promoter.Up	293,117
Promoter.Down	146,118
Exon	185,299
Intron	1,177,773
TES (transcription end sites)	40,077
Dis5 (5' distal regions)	549,355
Dis3 (3' distal regions)	192,714
Intergenic	503,845
Regions excluded from annotation (those on alternative chromosomes)	121,215

2.4 Differentially accessible regions identification

To perform statistical test between experimental groups, we first counted the number of fragments for each reference peak using with intersect command from pybedtools (v0.8.1; PMIDs: 21949271, 20110278). Next, the number of raw reads mapping per peak was converted to FPKM unit (Fragments Per Kilo base per Million mapped reads), and TMM (trimmed mean of M-values) from edgeR (PMID: 19910308), followed by limma-voom approach (PMIDs: 25605792, 24485249) was used to asses the significance of the differential peak binding / accessibility. More detailed description, including code sample, i available under section '1.5 Perform statistical tests' at our Wiki page. Next, the regions are categorized into **one** of the *differential* categories, each representing a different threshold:

1. **Up2** - FC $>$ 2 (log2(FC) $>$ 1) and FDR $<$ 0.05.
2. **Up2NoFDR** - FC $>$ 2 (log2(FC) $>$ 1) and p-value $<$ 0.05.
3. **Up** - FC $>$ 1 (log2(FC) $>$ 0) and FDR $<$ 0.05.
4. **UpNoFDR** - FC $>$ 1 (log2(FC) $>$ 0) and p-value $<$ 0.05.
5. **Control** - least changing regions, with 1/1.05 $<$ FC $<$ 1.05 and p.value $>$ 0.5 and average enrichment in both groups $>$ 0.
6. **DownNoFDR** - FC $<$ 1 (log2(FC) $<$ 0) and p-value $<$ 0.05.
7. **Down** - FC $<$ 1 (log2(FC) $<$ 0) and FDR $<$ 0.05.
8. **Down2NoFDR** - FC $<$ 0.5 (log2(FC) $<$ -1) and p-value $<$ 0.05.
9. **Down2** - FC $<$ 0.5 (log2(FC) $<$ -1) and FDR $<$ 0.05.

Up2 and Down2 categories are the most stringent and its usually recommended to use those thesholds in order to identify the most statistically significant, as well as likely the most biologically relevant changes. Nevertheless, because in real biology there usually are no "universal cutoffs", dependently on the experimental design and the hypothesis, it might be insightful to examine less stringent thresholds.

Note: FDR - The false discovery rate (FDR) is a statistical approach used in multiple hypothesis testing to correct for multiple comparisons. It is typically used in high-throughput experiments in order to correct for random events that falsely appear significant (source). Read more about multiple hypothesis testing and why its important here.

If the experiment had only one sample in each of the groups, the statistical testing is impossible, thus the following, simplified, categories are used:

1. **Up2FC** - FC $>$ 2 (log2(FC) $>$ 1).
2. **Control** - least changing regions, with 1/1.05 $<$ FC $<$ 1.05 and enrichment in both conditions $>$ 0.
3. **Down2FC** - FC $<$ 0.5 (log2(FC) $<$ -1).

Regions not classified to one of the above categories (categories 1-9 or 1-3, dependently on the experimental design), are classified as "**Other**".

2.4.1 Genes associated with differential peaks

After peaks are classified into the appropriate categories, e.g. UpNoFDR, and annotated to genes, the list of genes associated with particular category of differential peaks is subsequently extracted in *.gmt format. These files might be found inside `supplementaryFiles` subdirectory under *Results directory*. This allows investigators to have and easy access to the list of genes associated with differential peaks. This also allows one more easily to run tools like GSEA (Gene Set Enrichment Analysis), either with additional expression matrix (e.g. from RNA-seq experiments), or with list of genes preranked based on the ChIP-seq / ATAC-seq / Cut-and-Run data from this experiment; for more details see *Gene ranking based on differential peaks*.

Tip: When regions are classified into differential categories, one region is assigned to one threshold. However, e.g. UpNoFDR threshold, which is much more lenient than e.g. Up2 category, will actually also include all regions categorized into the latter, more stringent category. Therefore, to accommodate this relation, genes in GMT format are also extracted including all higher threshold categories;

2.4.2 Gene ranking based on differential peaks

In order to provide investigators with even more possibilities to examine the underlying data from as many angles as possible, ChIP-seq / ATAC-seq / Cut-and-Run data might also be used to calculate metric of genes annotated with regions, which might subsequently be used to rank genes. More specifically, the following procedure is applied:

1. **For each peak, the metric was calculated following four alternative approaches:**
    a. **FCRank metric**: $\log_2$(FC)
    b. **PRank metric**: direction of change * -$\log_{10}$(p-value)
    c. **FCPRank metric**: $\log_2$(FC) * -$\log_{10}$(p-value)
    d. **FCPERank metric**: $\log_2$(FC) * -$\log_{10}$(p-value) * $\log_{10}$(Mean Enrichment + 1)
2. Peaks annotated with genes by **putative promoter-related** regions. *Putative enhancer-related* related regions were excluded from the ranking.
3. Because each gene might be annotated with many peaks, for each gene, the representative peak was selected, based on which the metric for that gene was calculated. The representative peak per gene is the one which absolute metric is the highest.

Ranked gene lists are extracted to individual files which might be found inside `supplementaryFiles` subdirectory under *Results directory*. Those files are extracted in *.rnk format.

2.5 Computational environment

In this section we list all the essential libraries and modules that were used during the analysis, with separation to Python, R and HPCF.

2.5.1 HPCF modules loaded

Module	Version

2.5.2 Python packages

Below we list only essential package names and their versions, the full list of all the packages and their versions can be found in the `environment.conda_Py.txt` file, which is located in the `supplementaryFiles` subdirectory under *Results directory*.

library	version	source
adjusttext	1.0.4	pyhd8ed1ab_0
cairosvg	2.7.1	pyhd8ed1ab_0
chart-studio	1.1.0	pyh9f0ad1d_0
deeptools	3.5.4	pyhdfd78af_1
gseapy	1.1.1	py310hbee2dd9_0
joblib	1.3.2	pyhd8ed1ab_0
markdown	3.5.2	pyhd8ed1ab_0
matplotlib	3.6.2	py310hff52083_0
networkx	3.2.1	pyhd8ed1ab_0
numpy	1.23.5	py310h53a5b5f_0
pandas	2.1.4	py310hcc13569_0
pathlib	1.0.1	py310hff52083_7
pillow	10.0.1	py310ha6cbd5a_0
plotly	5.18.0	pyhd8ed1ab_0
pybedtools	0.9.1	py310h2b6aa90_0
python	3.10.8	h257c98d_0_cpython
scikit-learn	1.3.2	py310h1fdf081_2
scipy	1.11.4	py310hb13e2d6_0
seaborn	0.13.1	hd8ed1ab_0
tabulate	0.9.0	pyhd8ed1ab_1
tqdm	4.66.1	pyhd8ed1ab_0

2.5.3 R packages

```
R version 4.3.2 (2023-10-31)
Platform: x86_64-conda-linux-gnu (64-bit)
Running under: Red Hat Enterprise Linux 8.4 (Ootpa)

Matrix products: default
BLAS/LAPACK: /research_jude/rgs01_jude/groups/cab/projects/Control/common/cab_epi/
```

(continues on next page)

(continued from previous page)

```
↪condaenv/rShared2401/lib/libopenblasp-r0.3.25.so;  LAPACK version 3.11.0

locale:
[1] C

time zone: America/Chicago
tzcode source: system (glibc)

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] reshape2_1.4.4 pROC_1.18.5    ggplot2_3.4.4  edgeR_3.42.4   limma_3.56.2

loaded via a namespace (and not attached):
 [1] vctrs_0.6.5      cli_3.6.2        rlang_1.1.3      stringi_1.8.3
 [5] glue_1.7.0       colorspace_2.1-0 plyr_1.8.9       locfit_1.5-9.8
 [9] scales_1.3.0     fansi_1.0.6      grid_4.3.2       munsell_0.5.0
[13] tibble_3.2.1     lifecycle_1.0.4  stringr_1.5.1    compiler_4.3.2
[17] Rcpp_1.0.12      pkgconfig_2.0.3  farver_2.1.1     lattice_0.22-5
[21] R6_2.5.1         utf8_1.2.4       pillar_1.9.0     magrittr_2.0.3
[25] ggsci_3.0.0      tools_4.3.2      withr_3.0.0      gtable_0.3.4
```

SECTION 3

ACKNOWLEDGEMENTS

The Center for Applied Bioinformatics is supported by the National Cancer Institute, Cancer Center Support Grant P30 CA21765 and ALSAC. All manuscripts used from this report must acknowledge these two funding resources. We would also appreciate the title and journal name of the resulting manuscript to be forwarded to cab.helpdesk@stjude.org to help both non- and competitive renewals of the SJCRH Cancer Center grant.

Once again, thank you for using Center for Applied Bioinfromatics for your analyses. If you had any questions, comments or suggestions, never hesitate to follow up with us.

SECTION 4

SAMPLES METADATA

The differentially binding peaks identification analysis was conducted on a total of 4 samples, which metadata was described as follows:

4.1 Grp. lvl.: ATAC

Sample ID	Sample name	Grp. lvl.: ATAC
GSE247821-GSM7901246-MCF10A_WT_rep1	WT_rep1	WT
GSE247821-GSM7901247-MCF10A_WT_rep2	WT_rep2	WT
GSE247821-GSM7901248-MCF10A_E545K_rep1	E545K_rep1	E545K
GSE247821-GSM7901249-MCF10A_E545K_rep2	E545K_rep2	E545K

Note: *NoGrp* indicates the sample not included at particular grouping level into calculation of reproducible peaks for that level.

SECTION 5

REPRODUCIBLE PEAKS

5.1 Number of reproducible peaks

The following is the number of peaks called in GSE247821 experiment, as well as the number of reproducible and all-reproducible peaks combined from each condition and grouping level:

Sample name	#called peaks	#reproducible peaks	#All reproducible peaks
WT_rep1	184060	179968	197049
WT_rep2	180365	179968	197049
E545K_rep1	146327	149738	197049
E545K_rep2	160773	149738	197049

Tip: Reproducible peaks for each group (`*.reproduciblePeaks.bed`), as well as all reproducible peaks per grouping level (`*.allRepro.bed`) are availible under `reproduciblePeakCounts` subdirectory under *Results directory*.

SECTION 6

PCA OF ALL SAMPLES

Principal Component Analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set [Source].

Here, PCA of all samples included into each grouping level of the analysis, was based on top 3000 most variable peaks, identified independently on each of these levels.

6.1 Grouping: ATAC

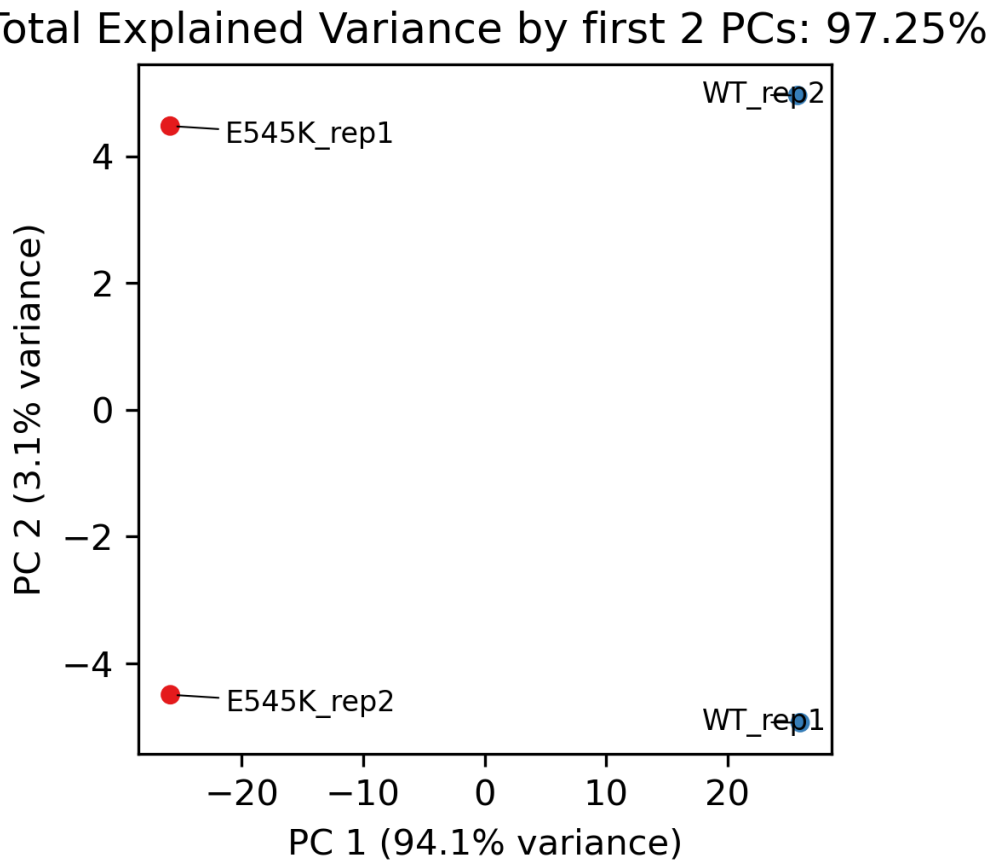

Fig. 6.1: Principal Component Analysis for ATAC grouping level

Note: The above PCA was calcualted for all the samples that were assigned to the same grouping level, so i.e. if the grouping level had more than two conditions, all of them will be displayed here, while the individual contrasts will display another PCA, that only contain the samples that were contrasted against each other. Moreover, the PDF version of the above PCA plots are available with the `*allSamples.PCA.2PCs.pdf` suffix under `visualization` subdirectory of *Results directory*. Side by side, one may also find there the interactive HTML version of this plot.

SECTION 7

CONTRASTS SUMMARY

The total of 4 samples, grouped on 1 levels, was used to compute 1 contrasts. They were listed in the following table:

#	Grouping	Target	Control	Up/Down diff. regions*
1	ATAC	E545K (2 samples)	WT (2 samples)	16538 up, 15288 down

* - The number of differentially binding regions listed in this summary table lists regions from Up2 and Down2 categories (FDR<0.05 and abs(log2FC)>1), in case if more than 1 sample was available for each tested condition; More thresholds are availible below. The table will display Up2FC and Down2FC categories (abs(log2FC)>1), in case if only 1 sample per condition was present.

Note: The above group names are also used as a part of the prefix of the results from particular contrast, following the pattern of `<TARGET>__VS__<CONTROL>`, which is how one may easily identify the appropriate supplementary files with the list of differential regions or genes associated with them, all located under `supplementaryFiles` subdirectory of *Results directory*.

SECTION 8

CONTRAST: E545K VS. WT

8.1 PCA plot

Principal component analysis was calculated for the samples included in this contrast, based on top 3000 most variable peaks.

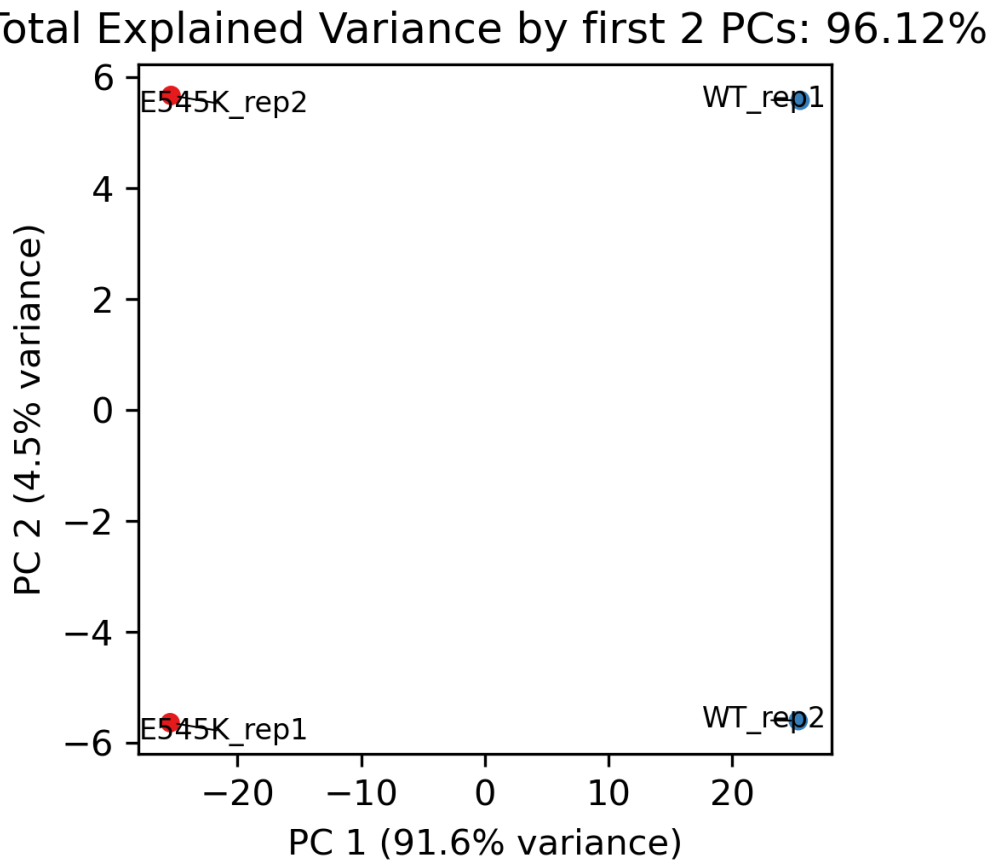

Fig. 8.1: Principal Component Analysis for E545K vs. WT

8.2 Summary plot and table

In order to identify which peaks are differentially binding, we have applied several non-exclusive thresholds. Results for each of those are summarized in the tables and graphs below, and might be further explored with the supplementary data files located under `supplementaryFiles` subdirectory of *Results directory*.

The results might be interactively explored in the HTML version of this report. Alternatively, the same may be done with MS Excel program, by opening the *GSE247821.E545K__VS__WT.allRegions.xlsx* file, located under `supplementaryFiles` subdirectory of *Results directory*.

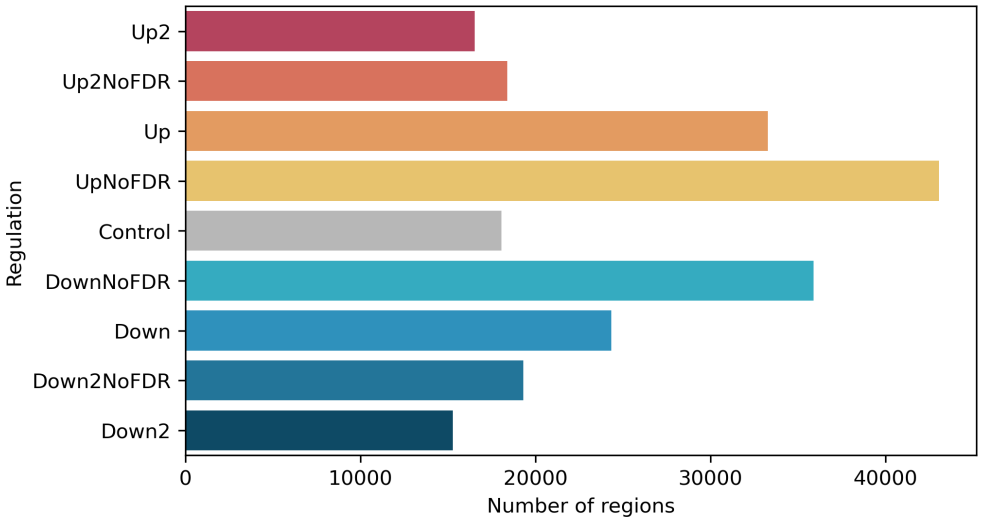

Fig. 8.2: Number of differential regions in E545K vs. WT contrast

Summary of the number of differential regions (various thresholds) and the number of genes annotated with them, based on either overlap with promoter region, or enhancer:

Category	Threshold	#Regions	#Gene-promoters	#Gene-enhancers
Up2	log2(FC) > 1, FDR < 0.05	16538	2216	11157
Up2NoFDR	log2(FC) > 1, p < 0.05	18379	2412	12075
Up	FDR < 0.05	33266	8005	14944
UpNoFDR	p < 0.05	43051	9726	17351
Control	Control	18054	5160	13171
DownNoFDR	p < 0.05	35884	4879	20157
Down	FDR < 0.05	24323	3394	16320
Down2NoFDR	log2(FC) < -1, p < 0.05	19310	2249	14762
Down2	log2(FC) < -1, FDR < 0.05	15288	1868	12475

8.3 Genomic contexts

This section described the genomic contexts within which the differential and control peaks are located. Visual summary:

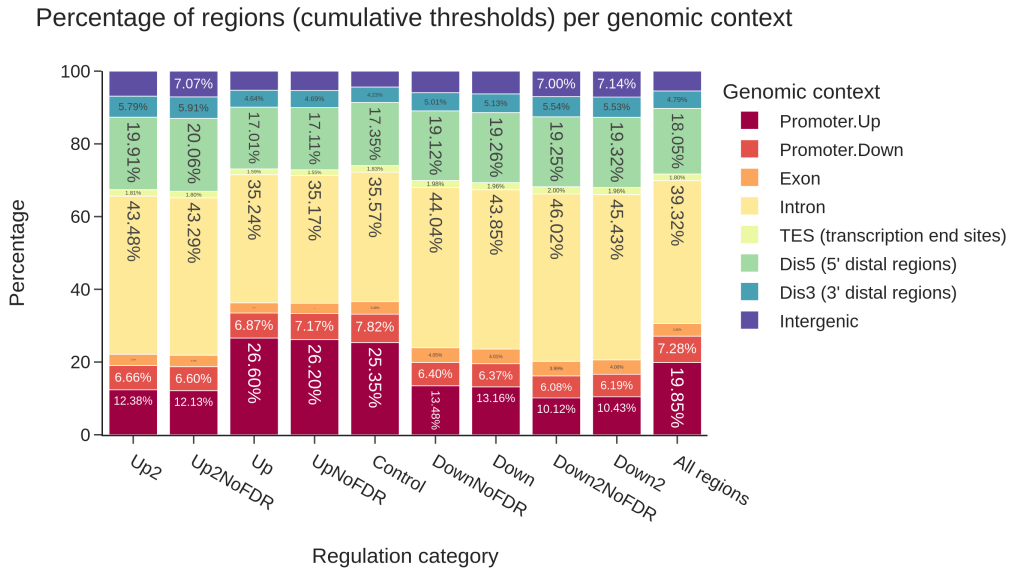

Fig. 8.3: Number of differential regions in E545K vs. WT contrast

The summary table with the details about the genomic context of the differential regions is availible in the supplementary files, under the name of *GSE247821.E545K__VS__WT.GenomicFeaturesAnnotation.tsv*.

8.4 Volcano plot

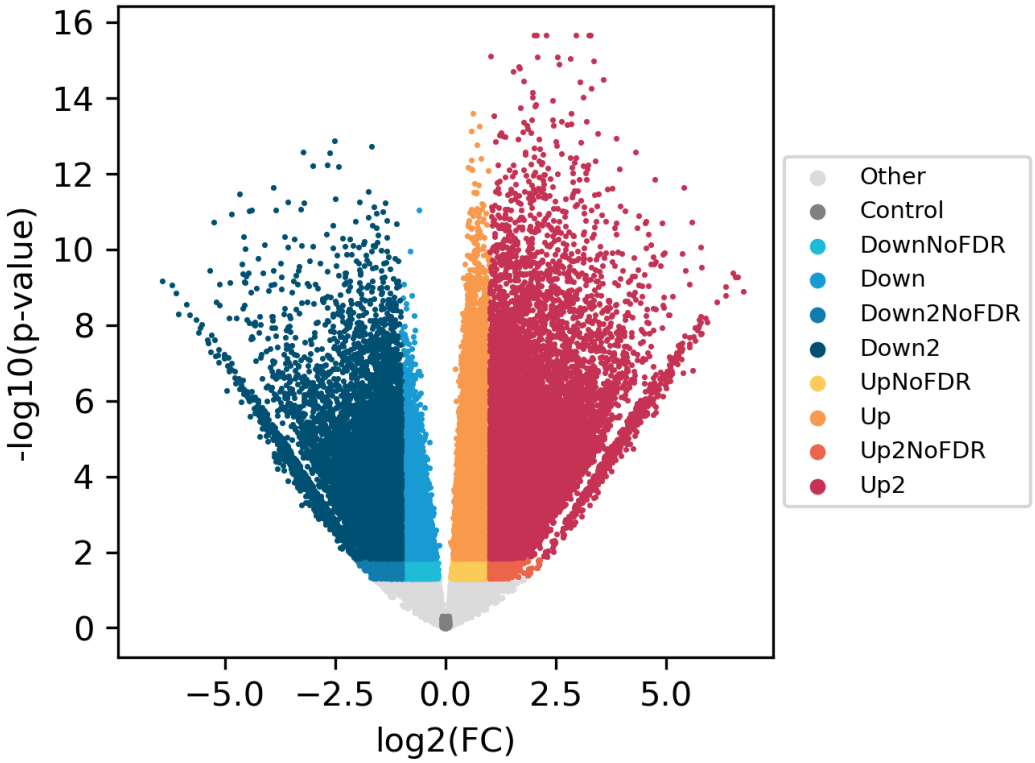

Fig. 8.4: Volcano plot for E545K vs. WT contrast

8.5 MA plot

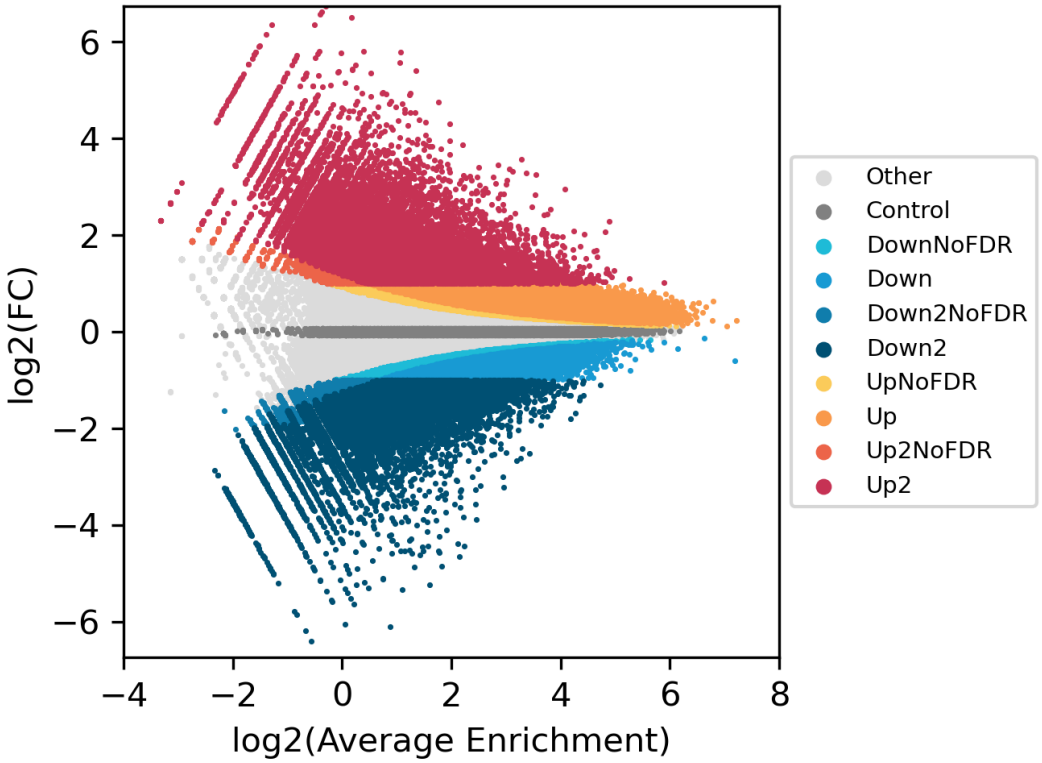

Fig. 8.5: MA plot for E545K vs. WT contrast

8.6 Signal enrichment heatmap

Heatmap of the signal enrichment at the most stringent threshold with at least 10 differential regions - Up2 and Down2 (abs(log2FC)>1, FDR < 0.05). Similar heatmaps were also computed for all less stringent thresholds, and are availible under `visualization/deeptoolsHeatmaps` subdirectory of the *Results directory*. The `deeptools heatmap` was generated with computeMatrix and plotHeatmap from Deeptools (PMID: 27079975). The PlotGrid below, shows the Pearson and Spearman correlation coefficients of the differential regions based on the signal enrichments from the heatmap.

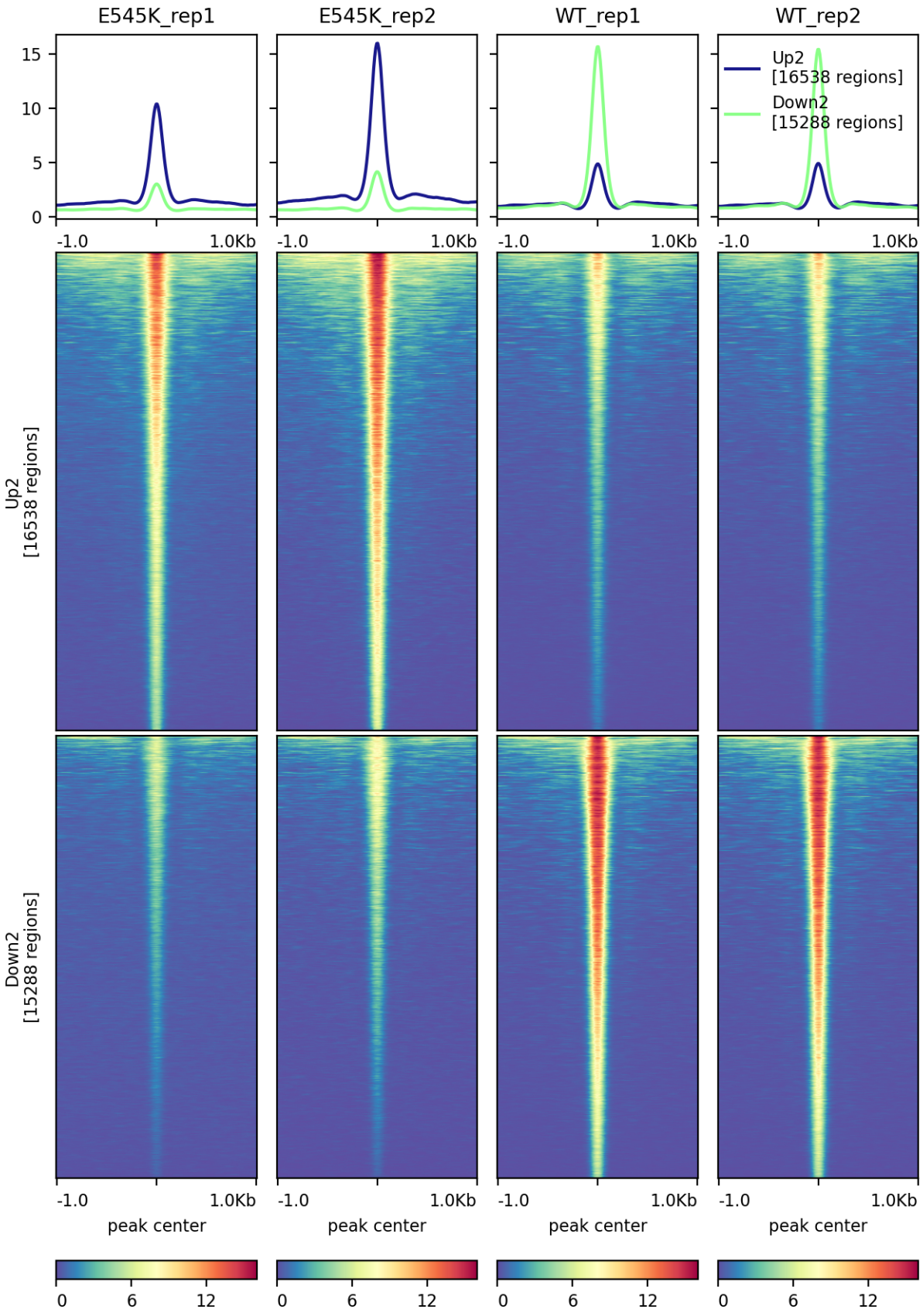

Fig. 8.6: Deeptools heatmap for E545K vs. WT contrast

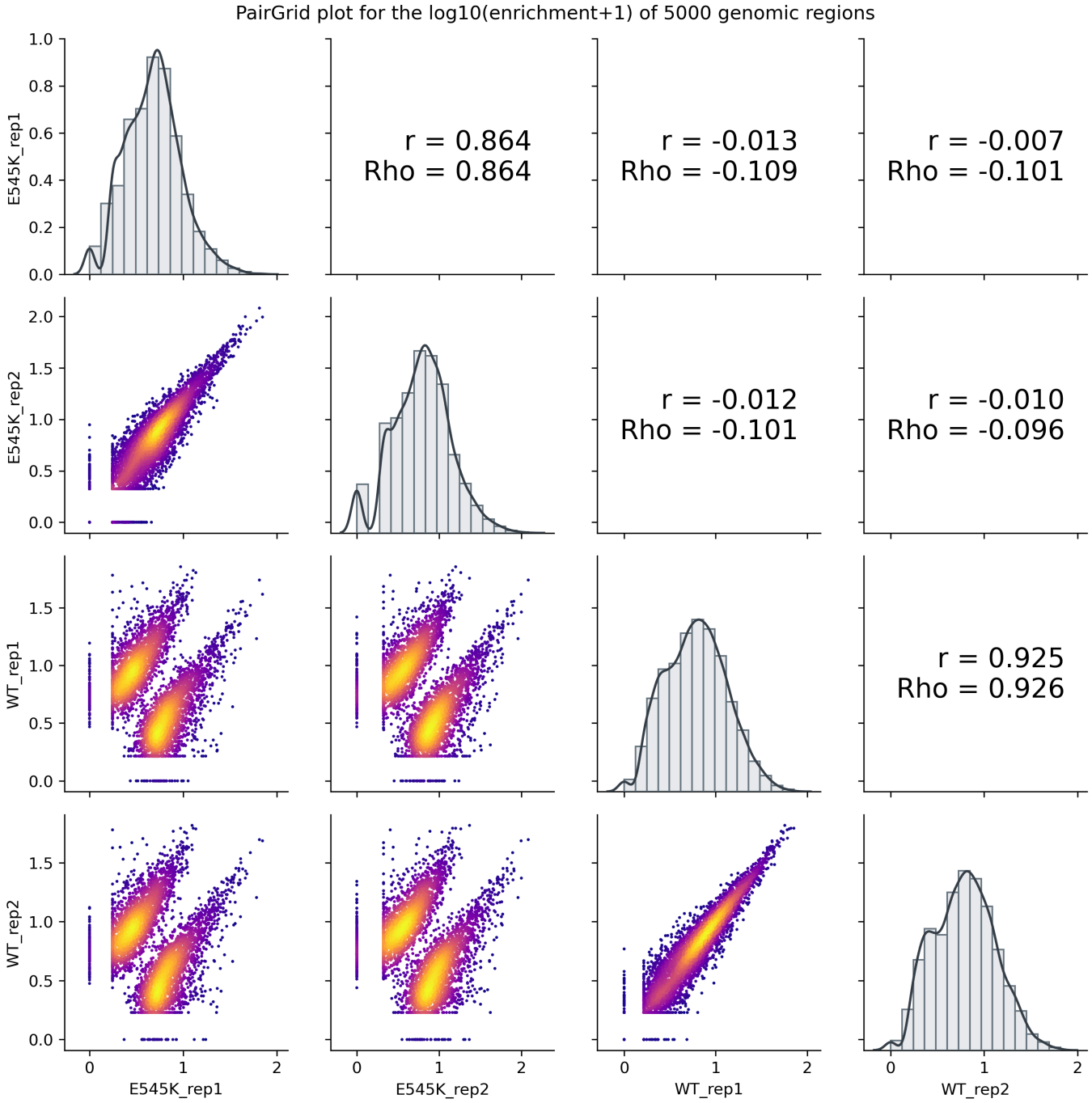

Fig. 8.7: PairGrid plot for E545K vs. WT contrast

8.7 Signal enrichment heatmap - control regions

Heatmap of the signal enrichment, and their Pearson and Spearman correlation coefficients, in *Control* regions. Those are the regions which after voom normalization shown the least changes, thus as an semi-independent validation, here we plot their enrichment using bigWig files normalized to the same sequencing depth. The rationalle here is that each sample, irrespectively of the condition, should in those regions have roughly similar enrichment and relatively high correlation. The control and differential peaks are plotted separately to avoid biasing the of heatmap color-scale to e.g. only the control regions, if those would display much stronger enrichment than differential regions (or vice versa).

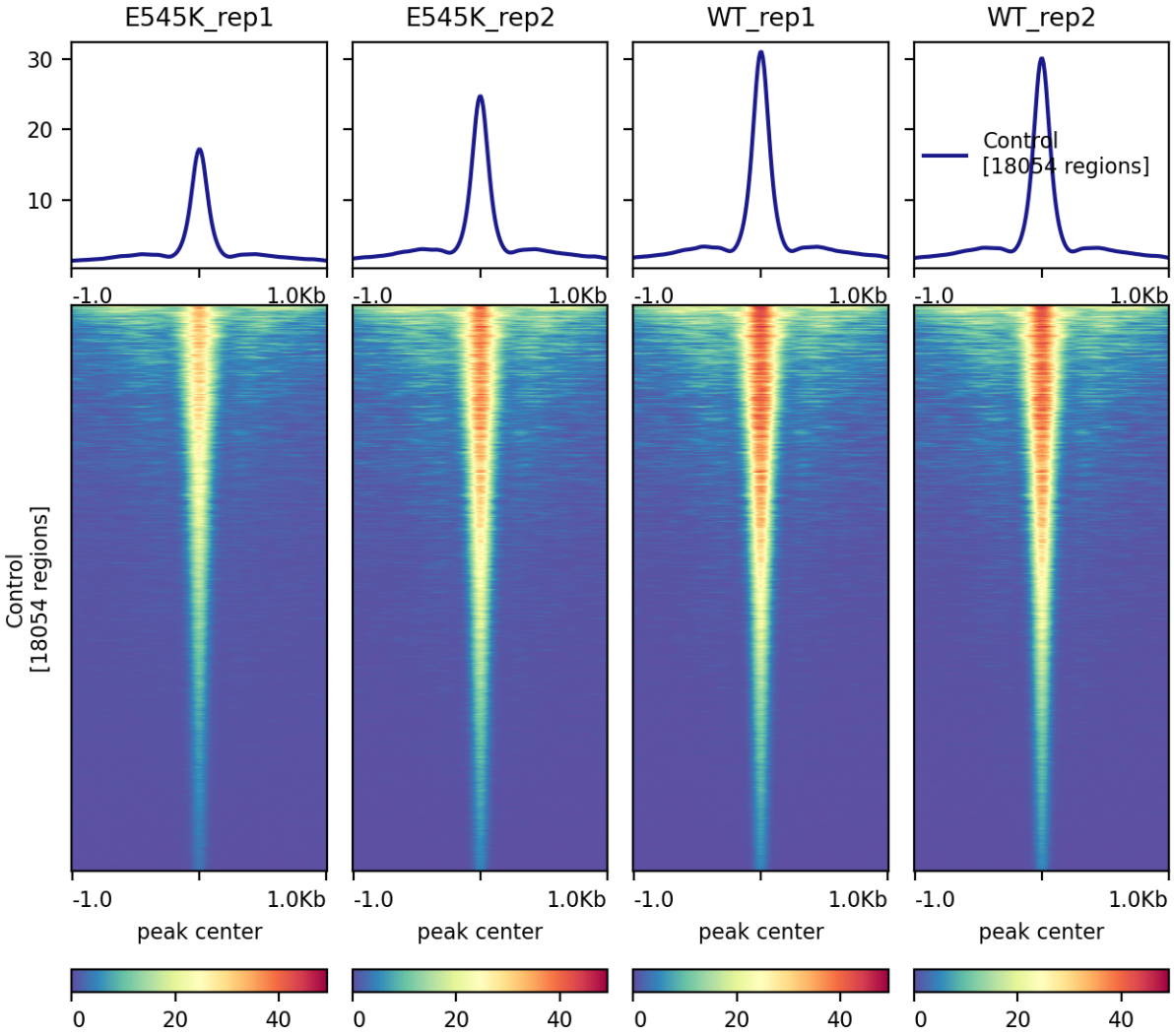

Fig. 8.8: Deeptools heatmap for E545K vs. WT contrast

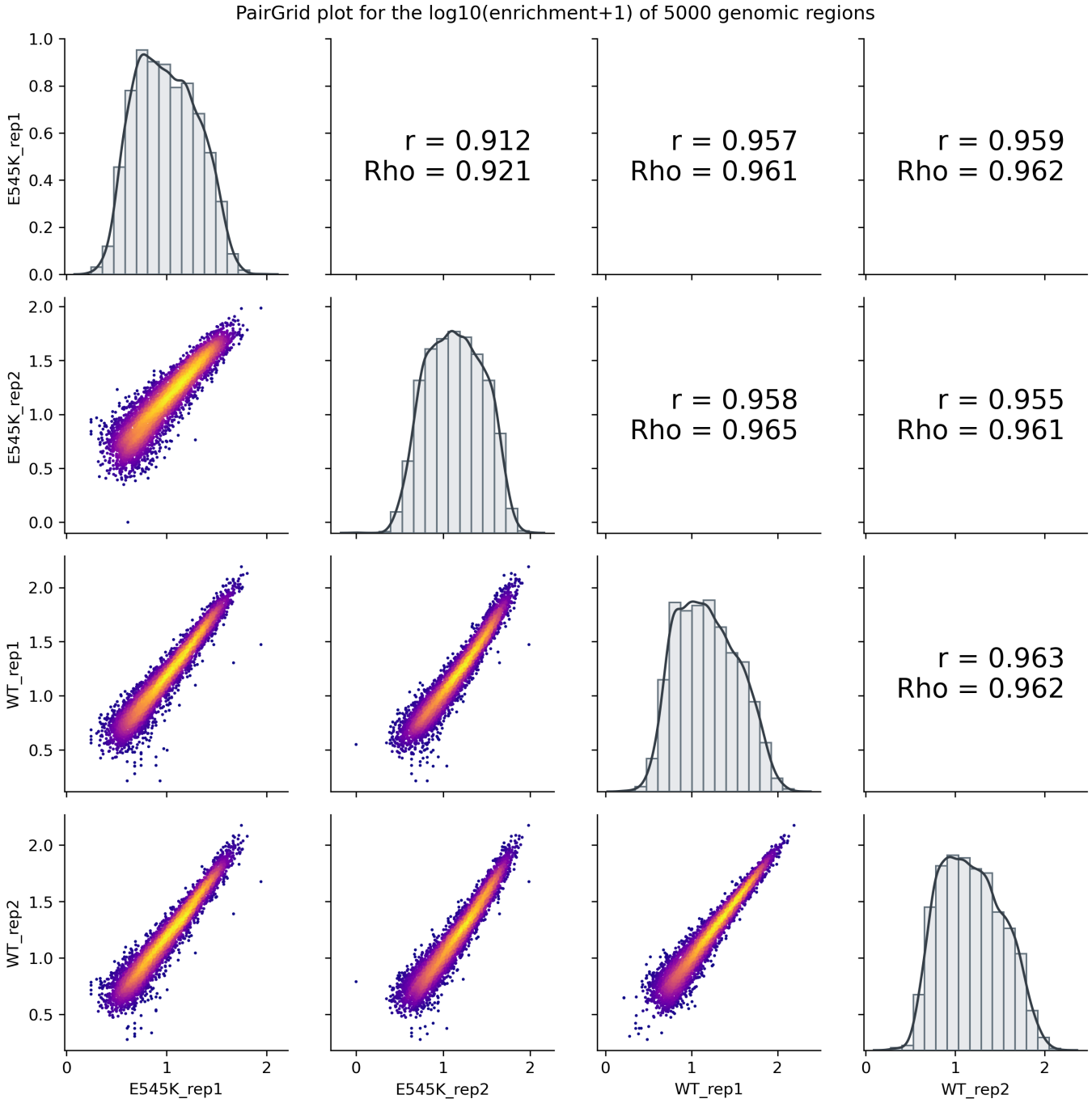

Fig. 8.9: PairGrid plot for E545K vs. WT contrast

8.8 Heatmap

Heatmap of top 30 differential regions annotated with gene promoters, ranked by p-value.

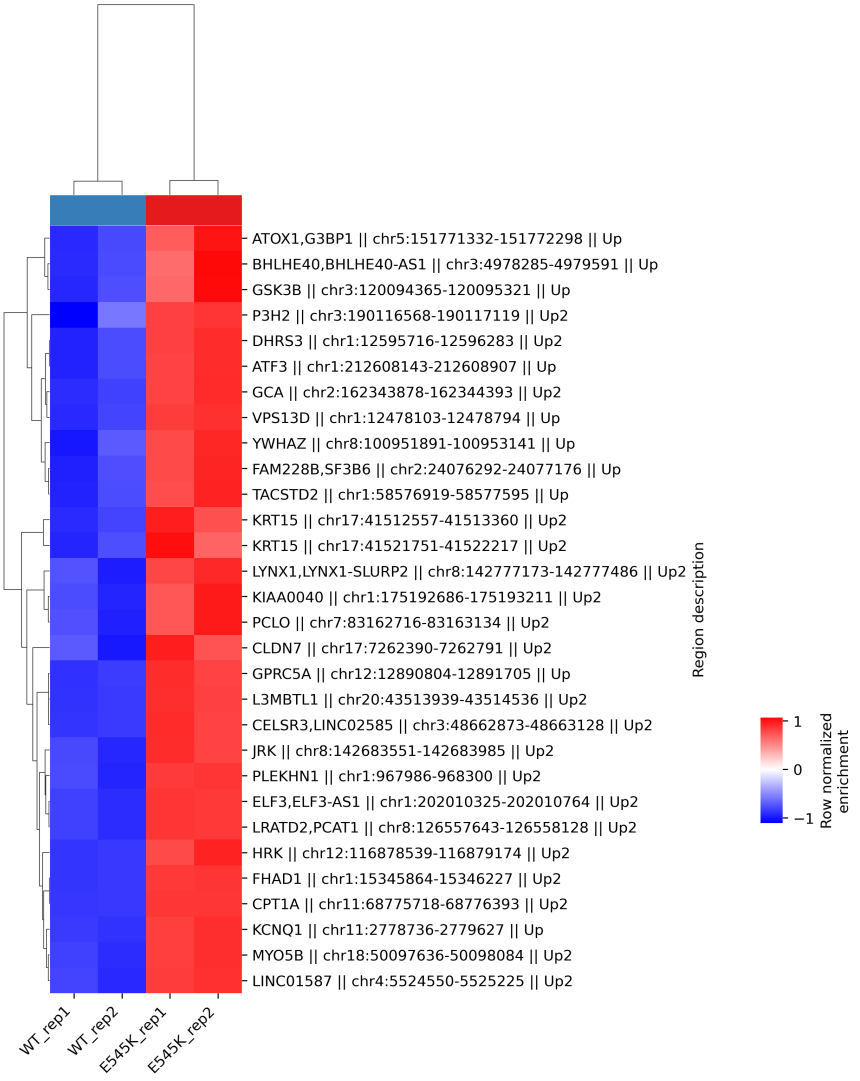

Fig. 8.10: Heatmap of top 30 genes for E545K vs. WT contrast

8.9 Gene Cloud

"Gene clouds", are novel visual representation of the relation between gene rank and metric (in here **log2FC metric was used to assign weights to genes**). Gene clouds are using a concept of Tag clouds, and was implemented in-house with wordcloud library from Python programming language.

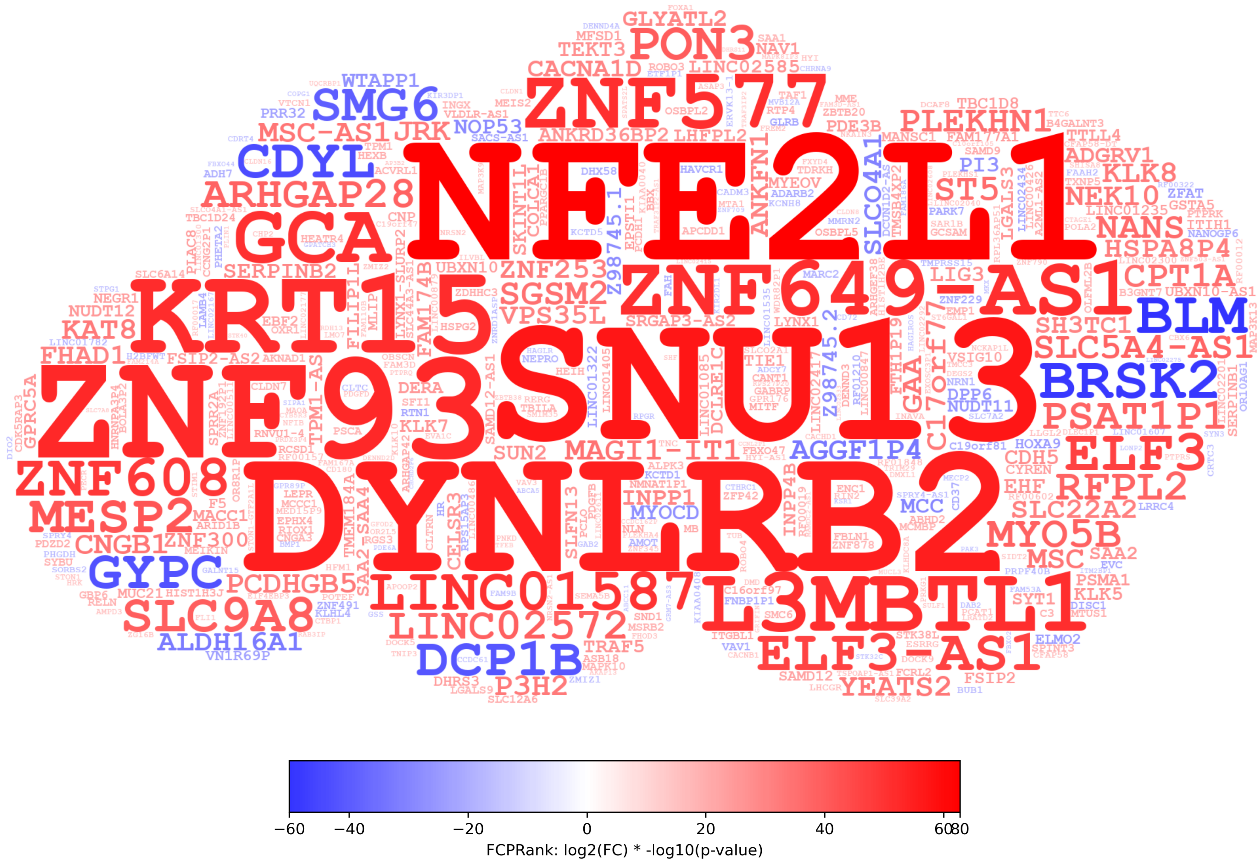

Fig. 8.11: log2FC based Gene Cloud for E545K vs. WT contrast

Note: Interpretation: Although the color of the gene names is directly based on the metric, the size of the gene is based both on the metric and the rank of the gene in relation to other genes. The Gene Cloud is intended not for direct interpretation, but rather as an visual aid, in order to for example more easily identify top changing genes or potential target genes, which are worth more in-depth investigation. Moreover, please keep in mind that genes were assigned with log2FC metric based on the overlap of reproducible peaks with promoter regions, which might not always be best choice, dependently on the biological question, protein studied and experimental design. To learn more on the approach used for assigning rank to gene go back to *Gene ranking based on differential peaks* methods section.

SECTION 9

MEA FOR E545K VS. WT

Motif enrichment analysis (MEA) was calculated using two alternative approaches, first of which utilize Homer software (v4.10, PMID: 20513432). Description of the results might be found on the project's website.

The strategy applied here attempted to identify motifs that are overrepresented among *target* sequences as compared with *background*. For that purpose, differential peaks, both up- and down- regulated, as well as non-differential control peaks, were compared as follows:

1. Target: Up-regulated vs. Background: Control non-diff. peaks
2. Target: Up-regulated vs. Background: Down (various thresholds)
3. Target: Down-regulated vs. Background: Control non-diff. peaks
4. Target: Down-regulated vs. Background: Up (various thresholds)

, but also:

5. Target: Up-regulated vs. Background: auto-selected by Homer
6. Target: Down-regulated vs. Background: auto-selected by Homer

For each of the above, Homer tool was used to identify *known motifs* as well as the *de novo motifs*. IF you would like to get more information on how Homer tool selects background control peaks on its own, read more about this on Homer's page, step no. 5 of section 'How findMotifsGenome.pl works'.

Note: In this report we only include the results from the most stringent category available for the contrast. However, the motif enrichment analysis was precomputed for all thresholds (e.g. Up vs. Down, Up2NoFDR vs. Down2NoFDR peaks, etc.). Those results are located under `MotifEnrichmentAnalysis_Homer` subdirectory of *Results directory*. Moreover, a second motif enrichment analysis approach using FIMO was applied to these data, but although those results were not yet incorporated into this report, they are available under `MotifEnrichmentAnalysis_FIMO` subdirectory of *Results directory*.

9.1 Up2 vs. Control - known motifs

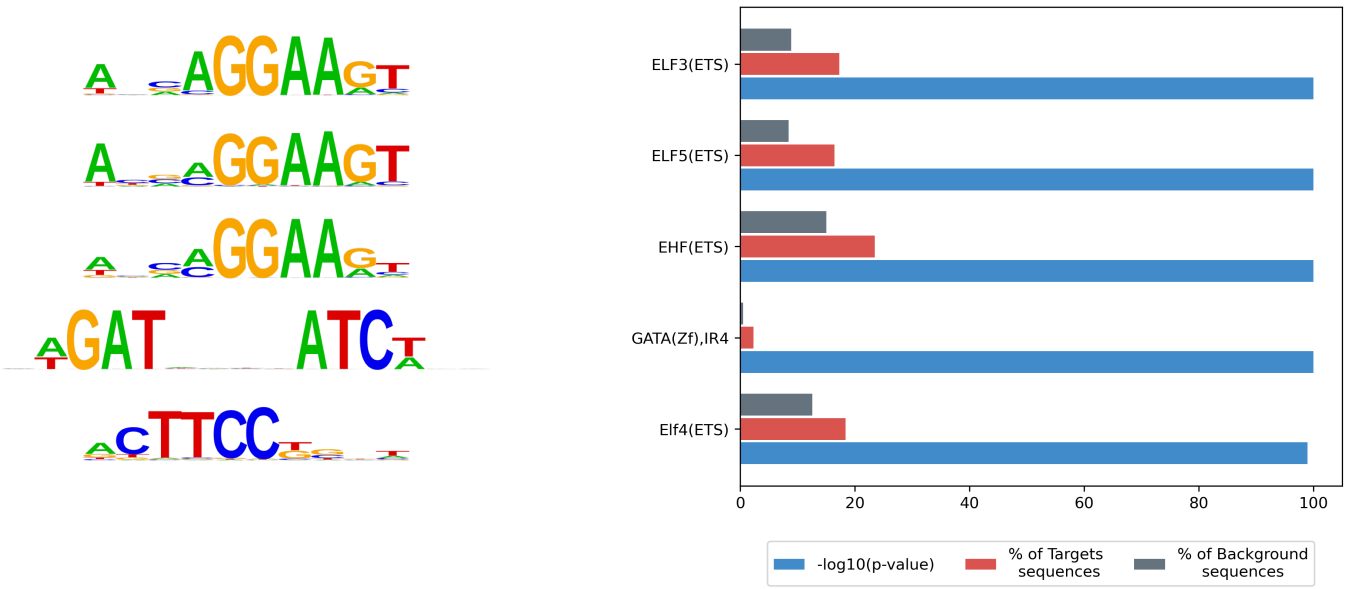

Fig. 9.1: Top known motifs enriched for E545K vs. WT contrast

Detailed Homer results may be explored under *MotifEnrichmentAnalysis_Homer* results subfolder, by opening *knownResults.html* file for the contrast and category that you wish to explore more.

9.2 Up2 vs. Control - de novo motifs

Please note, that here motif search was done in de novo mode, and although the motif itself is novel, the figure indicates the *best match* to known motifs. Detailed Homer results may be explored under *MotifEnrichmentAnalysis_Homer* results subfolder, by opening *homerResults.html* file for the contrast and category that you wish to explore more.

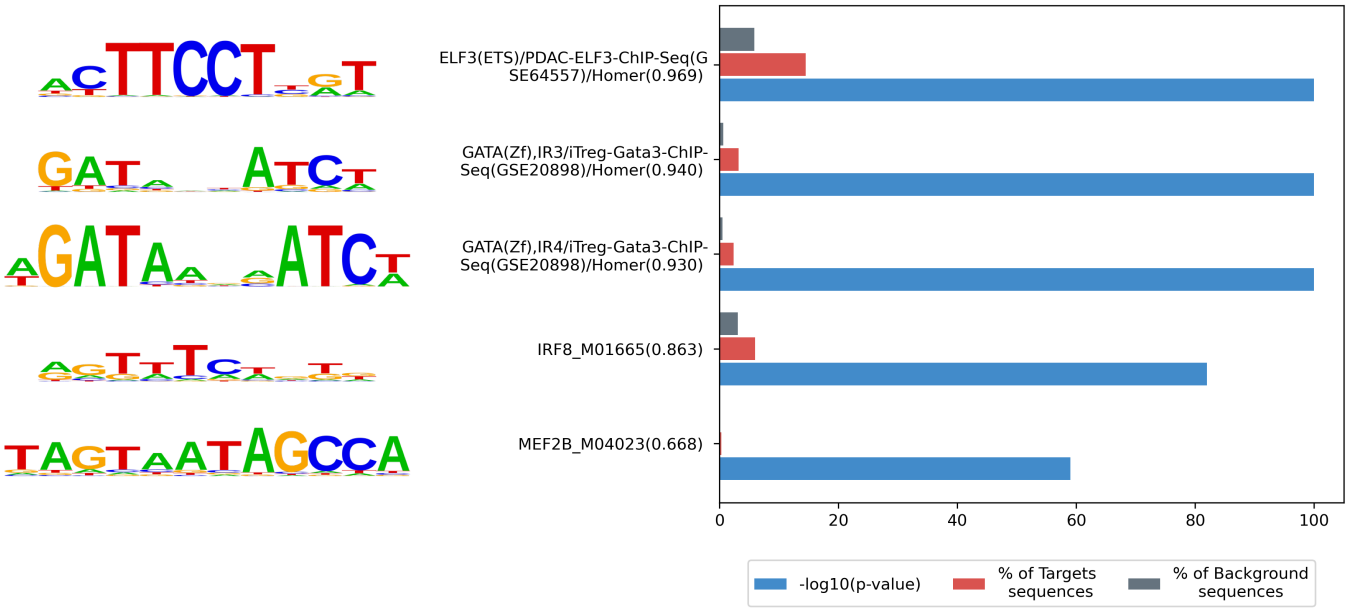

Fig. 9.2: Top de novo motifs for E545K vs. WT contrast (TF name indicate closest match)

9.3 Down2 vs. Control - known motifs

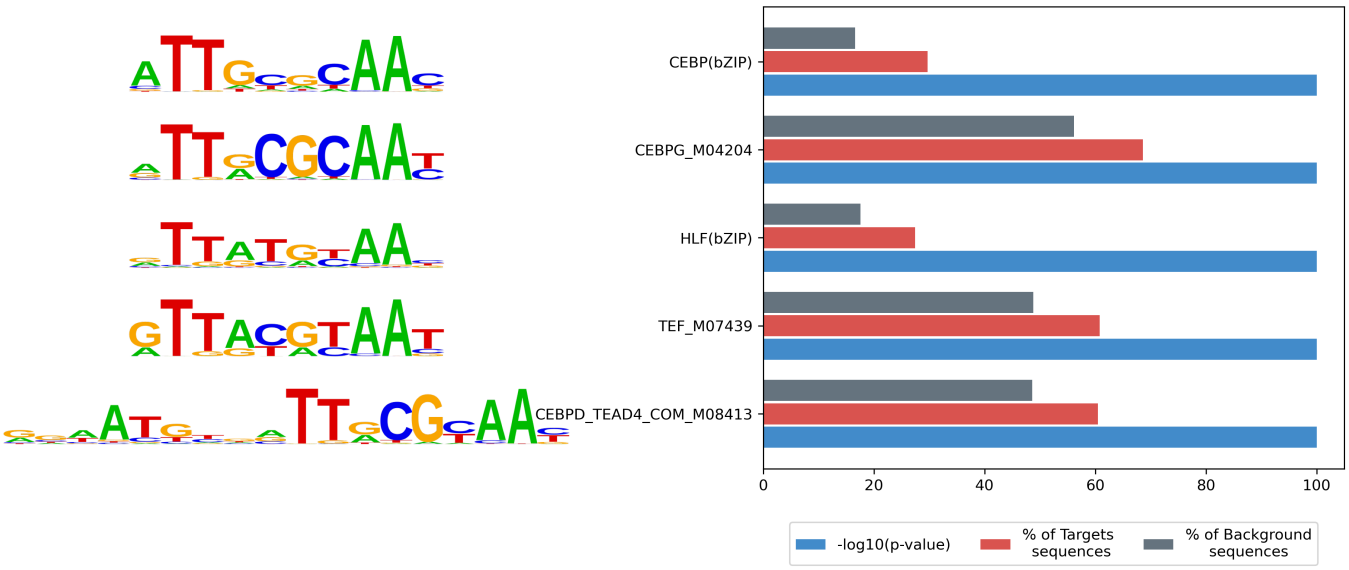

Fig. 9.3: Top known motifs enriched for E545K vs. WT contrast

Detailed Homer results may be explored under *MotifEnrichmentAnalysis_Homer* results subfolder, by opening *knownResults.html* file for the contrast and category that you wish to explore more.

9.4 Down2 vs. Control - de novo motifs

Please note, that here motif search was done in de novo mode, and although the motif itself is novel, the figure indicates the *best match* to known motifs. Detailed Homer results may be explored under *MotifEnrichmentAnalysis_Homer* results subfolder, by opening *homerResults.html* file for the contrast and category that you wish to explore more.

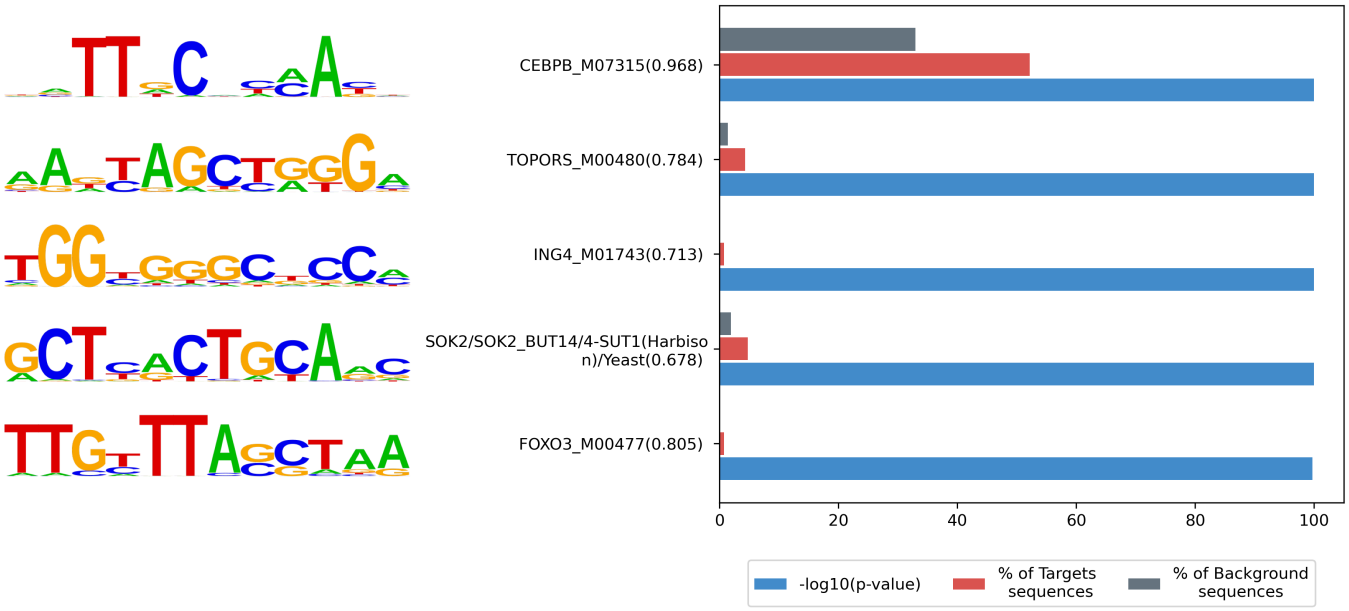

Fig. 9.4: Top de novo motifs for E545K vs. WT contrast (TF name indicate closest match)

9.5 Up2 vs. Down2 - known motifs

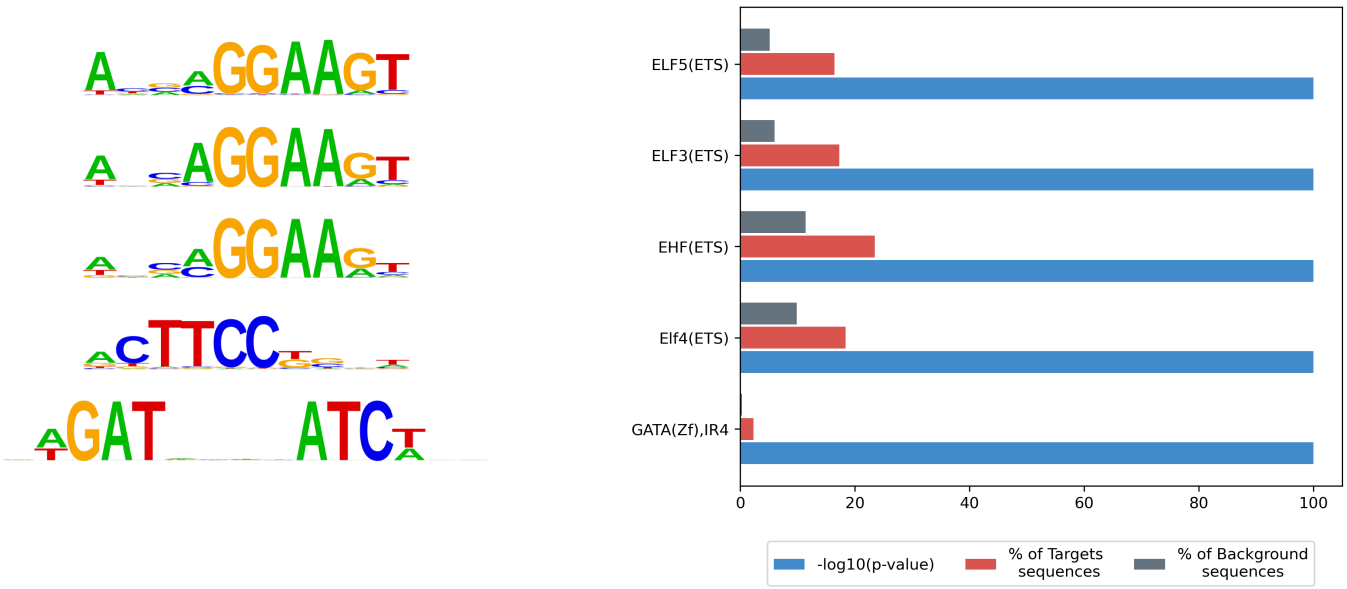

Fig. 9.5: Top known motifs enriched for E545K vs. WT contrast

Detailed Homer results may be explored under *MotifEnrichmentAnalysis_Homer* results subfolder, by opening *knownResults.html* file for the contrast and category that you wish to explore more.

9.6 Up2 vs. Down2 - de novo motifs

Please note, that here motif search was done in de novo mode, and although the motif itself is novel, the figure indicates the *best match* to known motifs. Detailed Homer results may be explored under *MotifEnrichmentAnalysis_Homer* results subfolder, by opening *homerResults.html* file for the contrast and category that you wish to explore more.

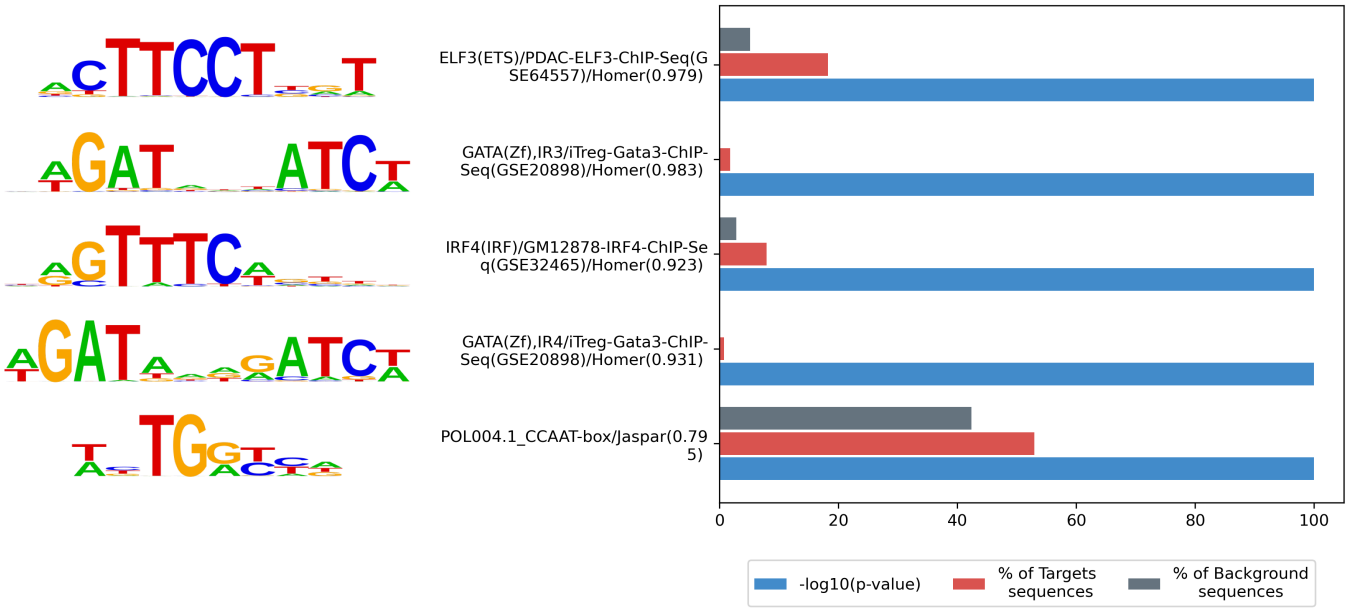

Fig. 9.6: Top de novo motifs for E545K vs. WT contrast (TF name indicate closest match)

9.7 Down2 vs. Up2 - known motifs

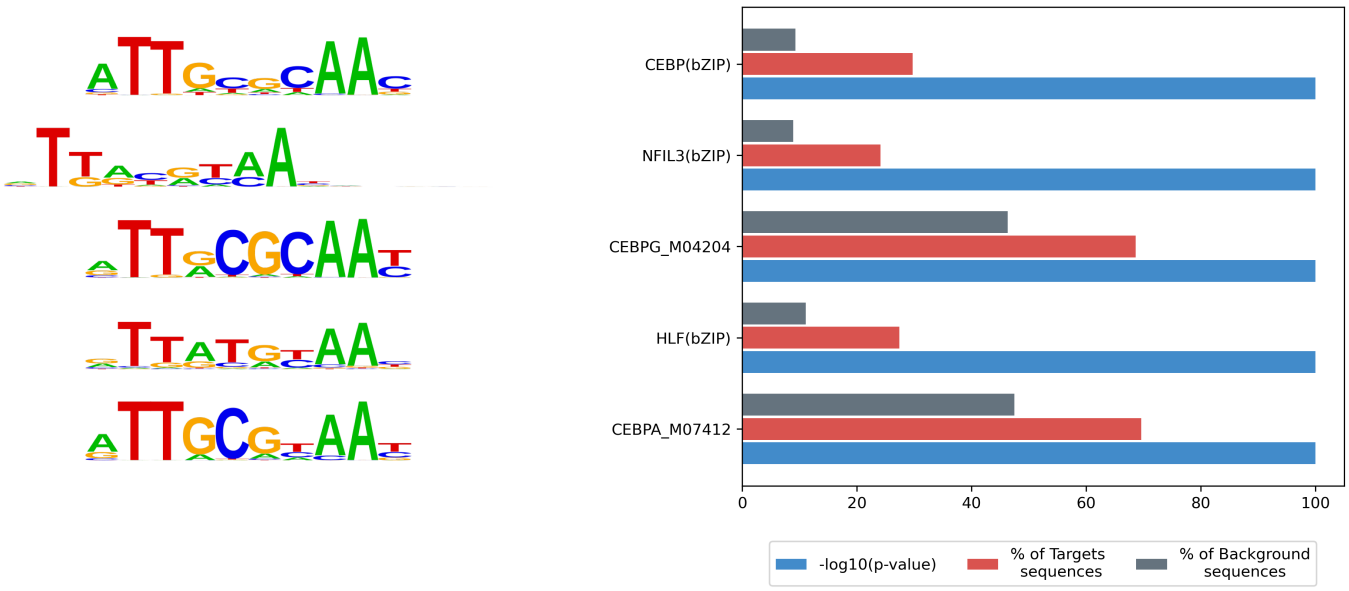

Fig. 9.7: Top known motifs enriched for E545K vs. WT contrast

Detailed Homer results may be explored under *MotifEnrichmentAnalysis_Homer* results subfolder, by opening *knownResults.html* file for the contrast and category that you wish to explore more.

9.8 Down2 vs. Up2 - de novo motifs

Please note, that here motif search was done in de novo mode, and although the motif itself is novel, the figure indicates the *best match* to known motifs. Detailed Homer results may be explored under *MotifEnrichmentAnalysis_Homer* results subfolder, by opening *homerResults.html* file for the contrast and category that you wish to explore more.

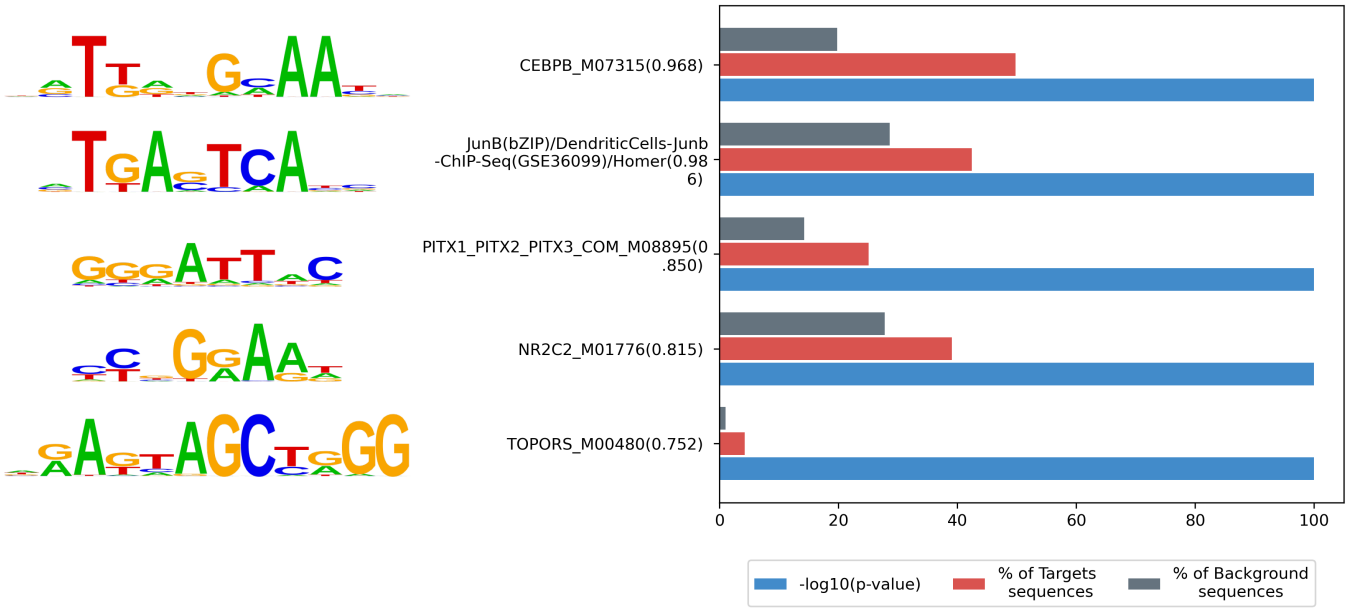

Fig. 9.8: Top de novo motifs for E545K vs. WT contrast (TF name indicate closest match)

9.9 Up2 vs. HomerBackground - known motifs

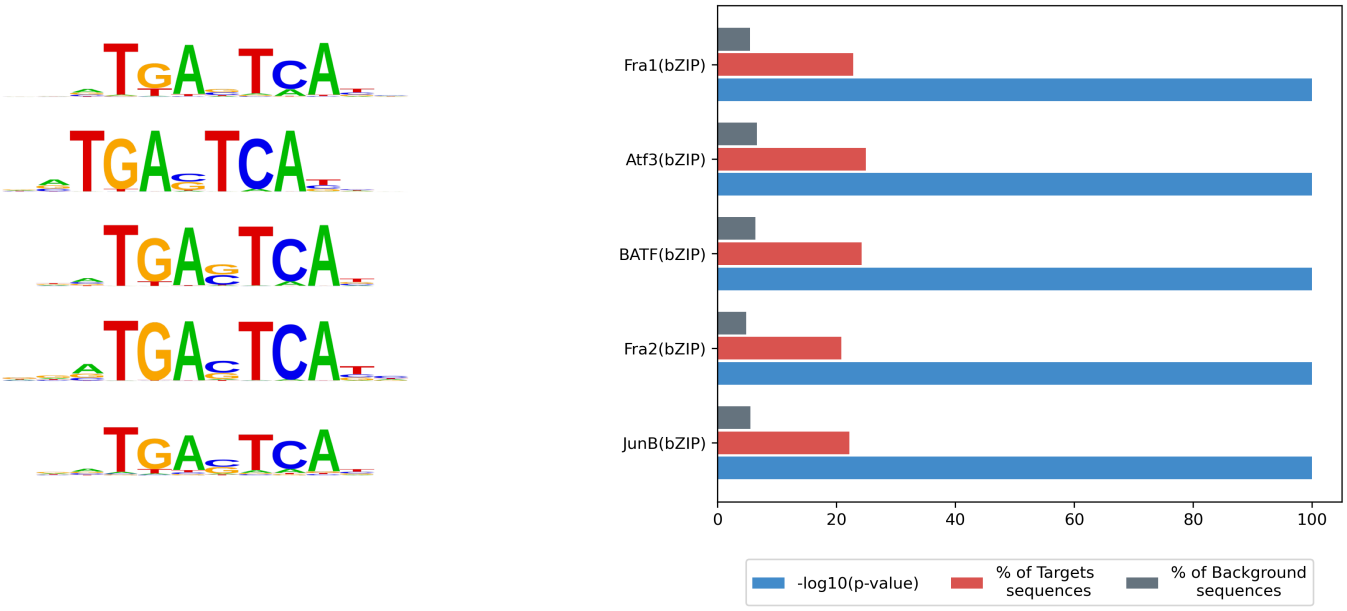

Fig. 9.9: Top known motifs enriched for E545K vs. WT contrast

Detailed Homer results may be explored under *MotifEnrichmentAnalysis_Homer* results subfolder, by opening *knownResults.html* file for the contrast and category that you wish to explore more.

9.10 Up2 vs. HomerBackground - de novo motifs

Please note, that here motif search was done in de novo mode, and although the motif itself is novel, the figure indicates the *best match* to known motifs. Detailed Homer results may be explored under *MotifEnrichmentAnalysis_Homer* results subfolder, by opening *homerResults.html* file for the contrast and category that you wish to explore more.

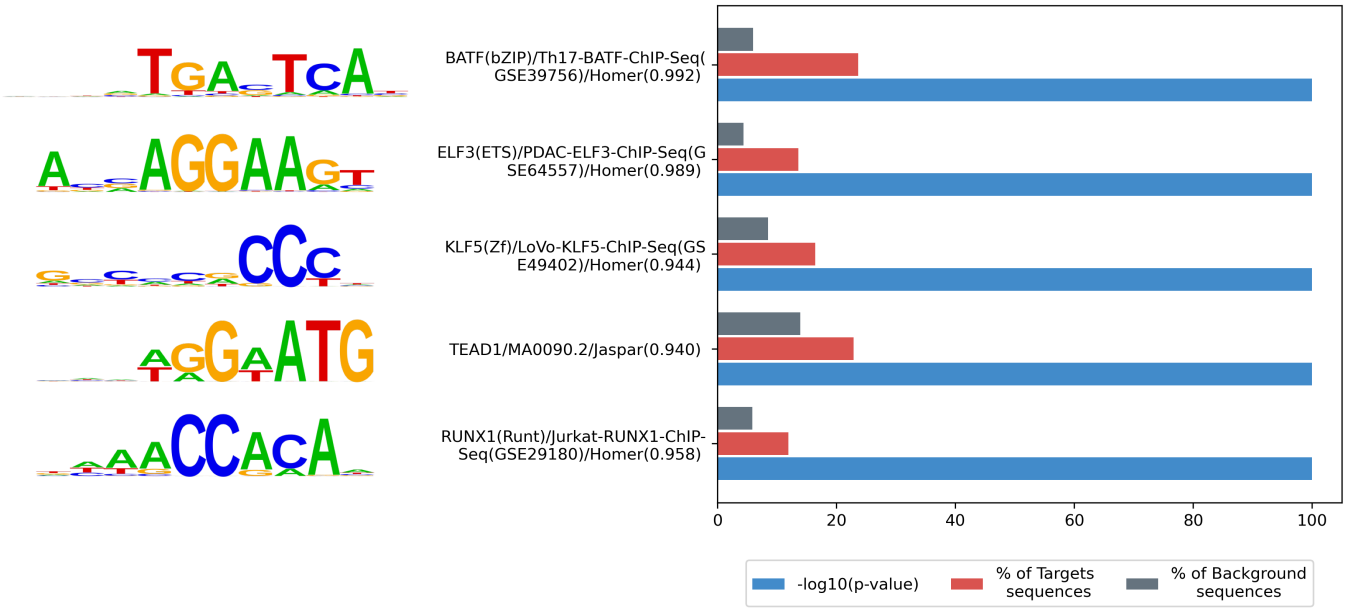

Fig. 9.10: Top de novo motifs for E545K vs. WT contrast (TF name indicate closest match)

9.11 Down2 vs. HomerBackground - known motifs

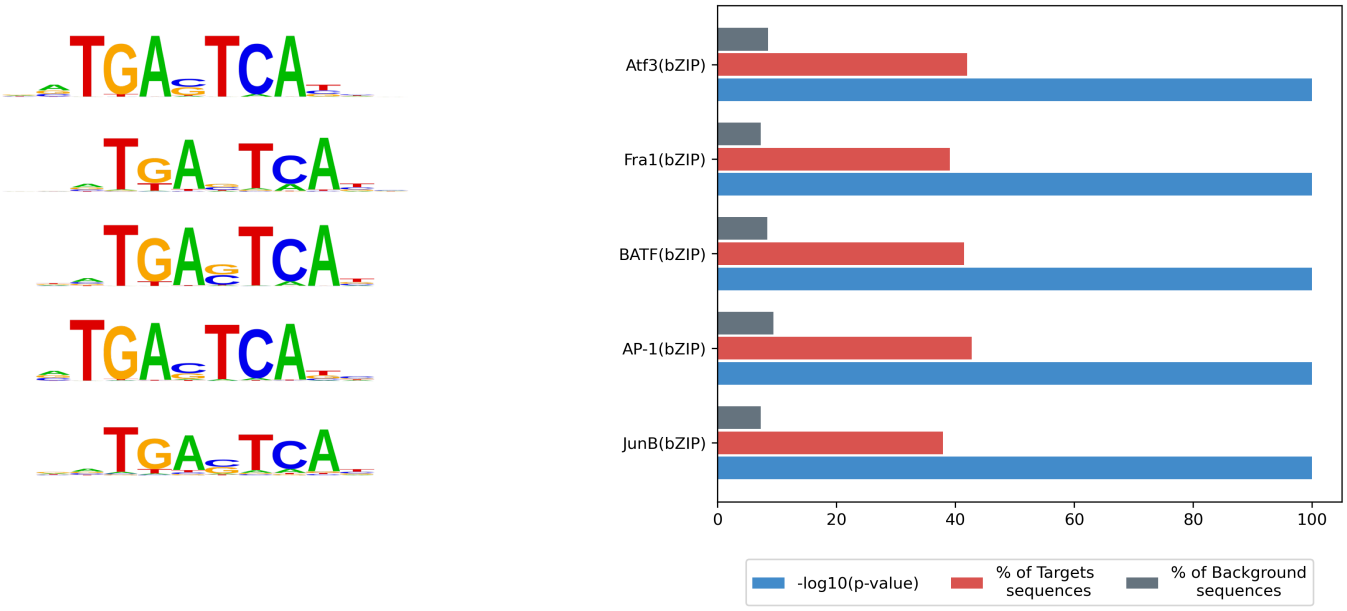

Fig. 9.11: Top known motifs enriched for E545K vs. WT contrast

Detailed Homer results may be explored under *MotifEnrichmentAnalysis_Homer* results subfolder, by opening *knownResults.html* file for the contrast and category that you wish to explore more.

9.12 Down2 vs. HomerBackground - de novo motifs

Please note, that here motif search was done in de novo mode, and although the motif itself is novel, the figure indicates the *best match* to known motifs. Detailed Homer results may be explored under *MotifEnrichmentAnalysis_Homer* results subfolder, by opening *homerResults.html* file for the contrast and category that you wish to explore more.

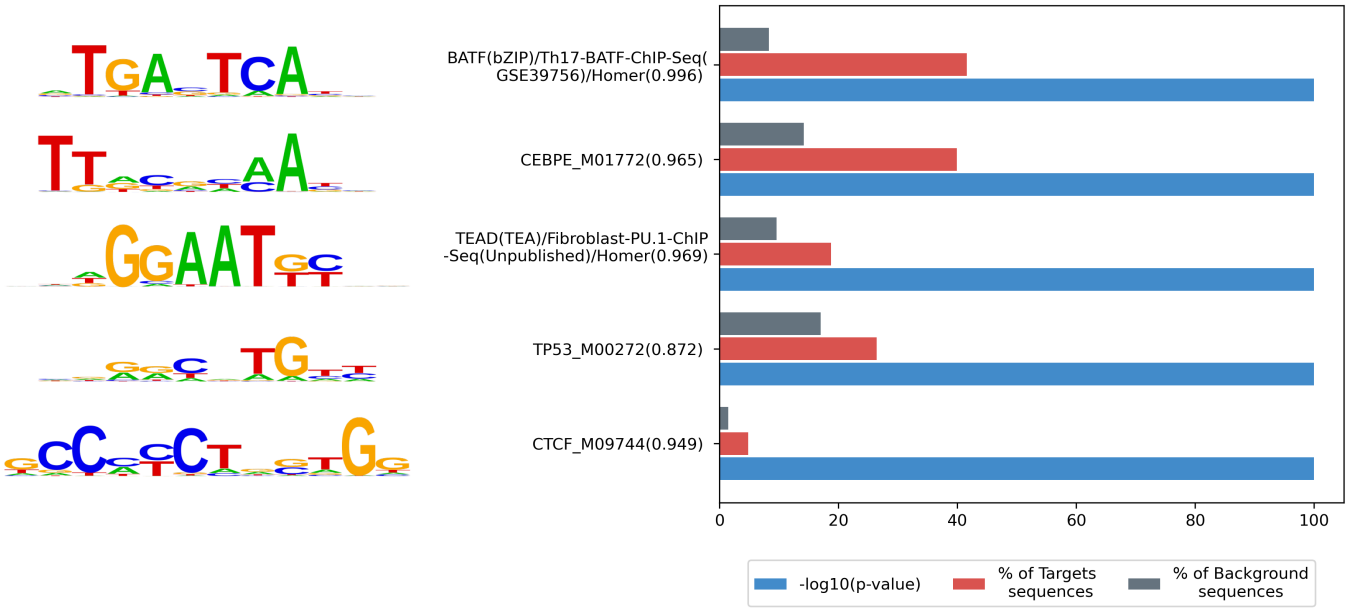

Fig. 9.12: Top de novo motifs for E545K vs. WT contrast (TF name indicate closest match)

SECTION 10

GSEA FOR E545K VS. WT

Gene Set Enrichment Analysis (GSEA; PMID: 16230612) was computed using preranked list of genes (Annotation of peaks-to-promoters of genes with FCPRank ranking strategy, more details in *Gene ranking based on differential peaks* methods section); with subcollections of gene sets from MSigDB v2023.1 and using GSEApy (version 1.1.1). The following gene sets subcollections were used for the analysis:

- H: Hallmark gene sets
- **C2: Curated gene sets**
  - C2 Biocarta
  - C2 KEGG
  - C2 Reactome
  - C2 WikiPathways
- **C3: Regulatory target gene sets**
  - C3: Transcription factor targets (GTRD)
  - C3: Transcription factor targets (TFT_LEGACY)
- C4: Computational gene sets
- **C5: Ontology gene sets**
  - C5: Gene ontology biological process (GO-BP)
  - C5: Gene ontology cellular component (GO-CC)
  - C5: Gene ontology molecular function (GO-MF)
- C6: Oncogenic signature gene sets
- C8: Cell type signature gene sets

Note: Detailed GSEA results may be interactively explored under GSEA results subfolder, by opening `GSEA_reports_combined.GSE247821.E545K__VS__WT.GeneNames_promoter_FCPRank.tsv.html` file located under *GSEA/summaryCombinedGSEA/processedMergedTables/* directory.

A total of 9380 gene signatures was analyzed, and the summary of the GSEA run is:

Description	Biased toward positive phenotype	Biased toward negative phenotype
Gene sets bias (all based on NES score)	7431	1949
Significant bias (FDR < 0.05)	21	2

GSEA enrichment volcano plot:

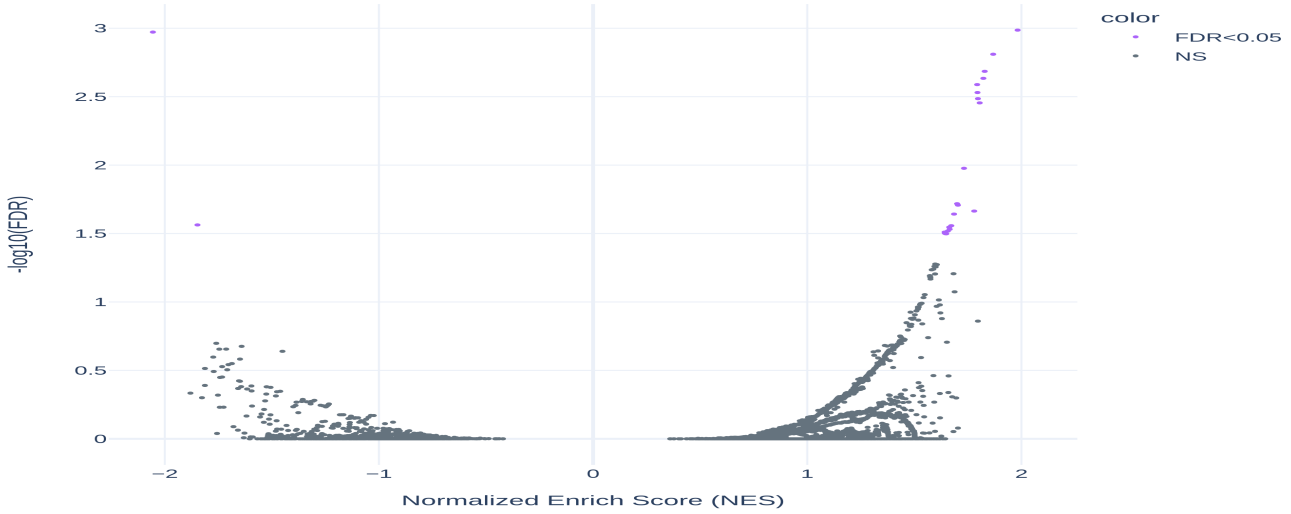

Fig. 10.1: Volcano plot for gene set enrichment in E545K vs. WT contrast

Network analysis results:

Network analysis presented here was calculated for top 10 gene sets enriched significantly by FDR < 0.05. The color in the figure below represents the NES score, while the size of the node represents *Tag %*, that is the number of genes in gene set being enriched in the data. Finally, the edge thickness represents the number of shared genes enriched in two gene sets, as calculated by jaccard coefficient. We recommend to explore the interactive version of the network plots generated for *top 30*, *top 50* and *top 100* gene sets, located under *GSEA/summaryCombinedGSEA/visualizations/NetworkPlots/interactiveNetwork.*.html* directory. Note however, that the more "top" gene sets are included, the slower the loading of the document into browser, with top 100 taking up to a minute to fully render.

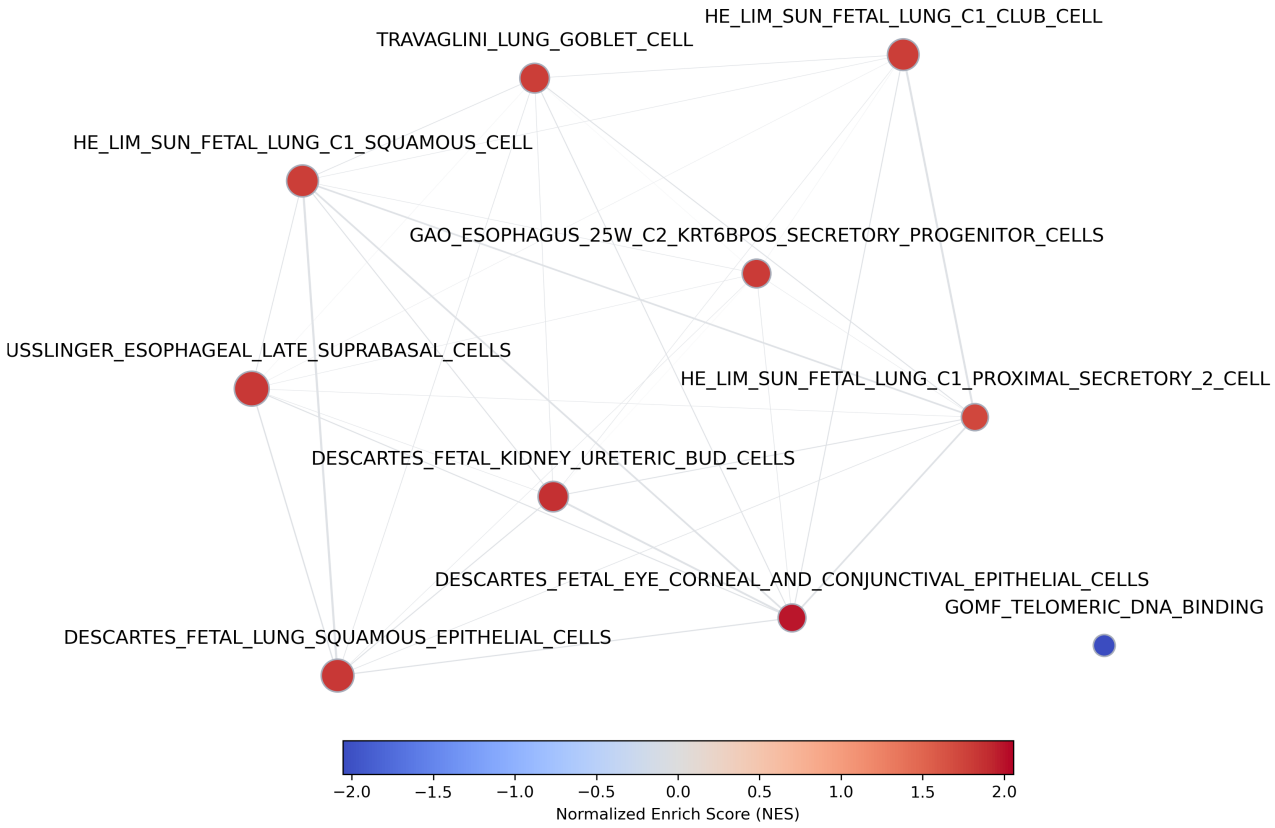

Fig. 10.2: GSEA-based network plot for top 10 gene sets enriched in E545K vs. WT contrast

The following table shows stats for the gene sets, for which the above network was calculated:

Term	NES	NOM p-val	FDR q-val	Tag %
DESCARTES_FETAL_EYE_CORNEAL_AND_CONJUNCTIVAL	1.9821	0	0.001	68 of 222
GOMF_TELOMERIC_DNA_BINDING	-2.0554	0	0.0011	7 of 37
DESCARTES_FETAL_KIDNEY_URETERIC_BUD_CELLS	1.8681	0	0.0015	85 of 234
BUSSLINGER_ESOPHAGEAL_LATE_SUPRABASAL_CELLS	1.8284	0	0.0021	62 of 130
DESCARTES_FETAL_LUNG_SQUAMOUS_EPITHELIAL_CEL	1.8223	0	0.0023	76 of 180
HE_LIM_SUN_FETAL_LUNG_C1_SQUAMOUS_CELL	1.7932	0	0.0026	100 of 250
HE_LIM_SUN_FETAL_LUNG_C1_CLUB_CELL	1.7948	0	0.0029	24 of 61
TRAVAGLINI_LUNG_GOBLET_CELL	1.7968	0	0.0033	46 of 133
GAO_ESOPHAGUS_25W_C2_KRT6BPOS_SECRETORY_PRO	1.8052	0	0.0035	22 of 68
HE_LIM_SUN_FETAL_LUNG_C1_PROXIMAL_SECRETORY_	1.7319	0	0.0106	75 of 261

Finally, the following are the combined enrichment plot for top 10 gene signatures. We recommend to open the interactive version of the plot, in order to explore the enrichment plots for individual gene sets.

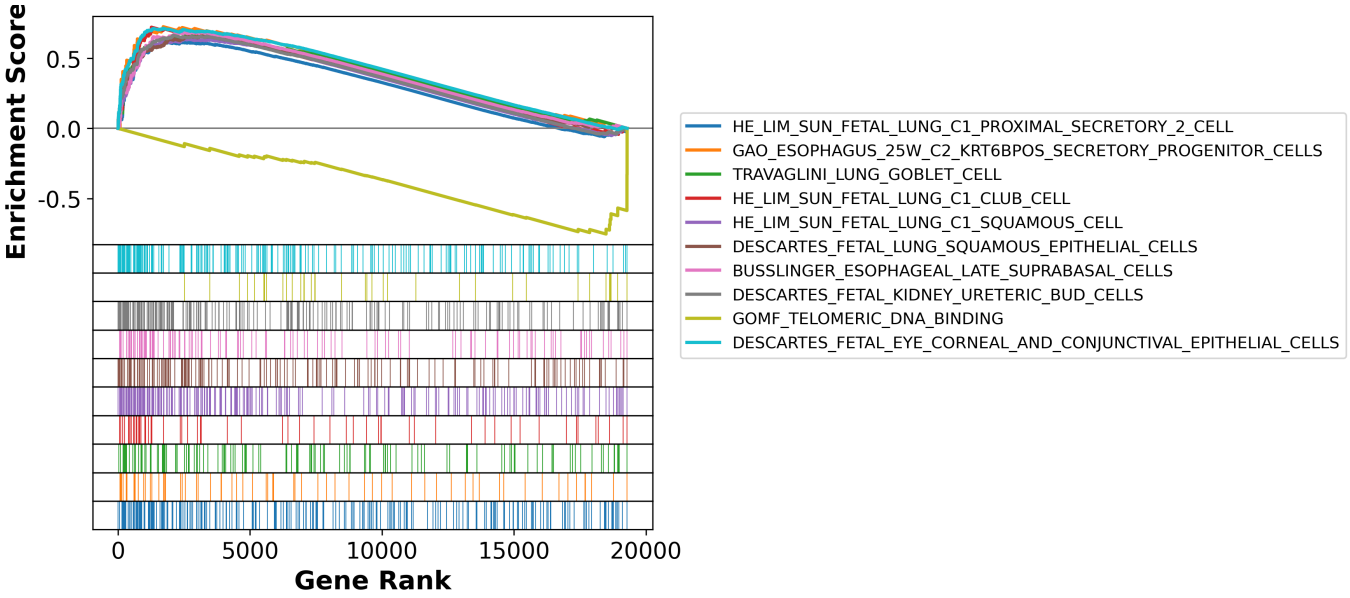

Fig. 10.3: Combined Enrichment Plot for top 10 gene sets enriched in E545K vs. WT contrast