Exploratory Data Analysis on Enhancers Project

Liwen Yin

2025-03-25

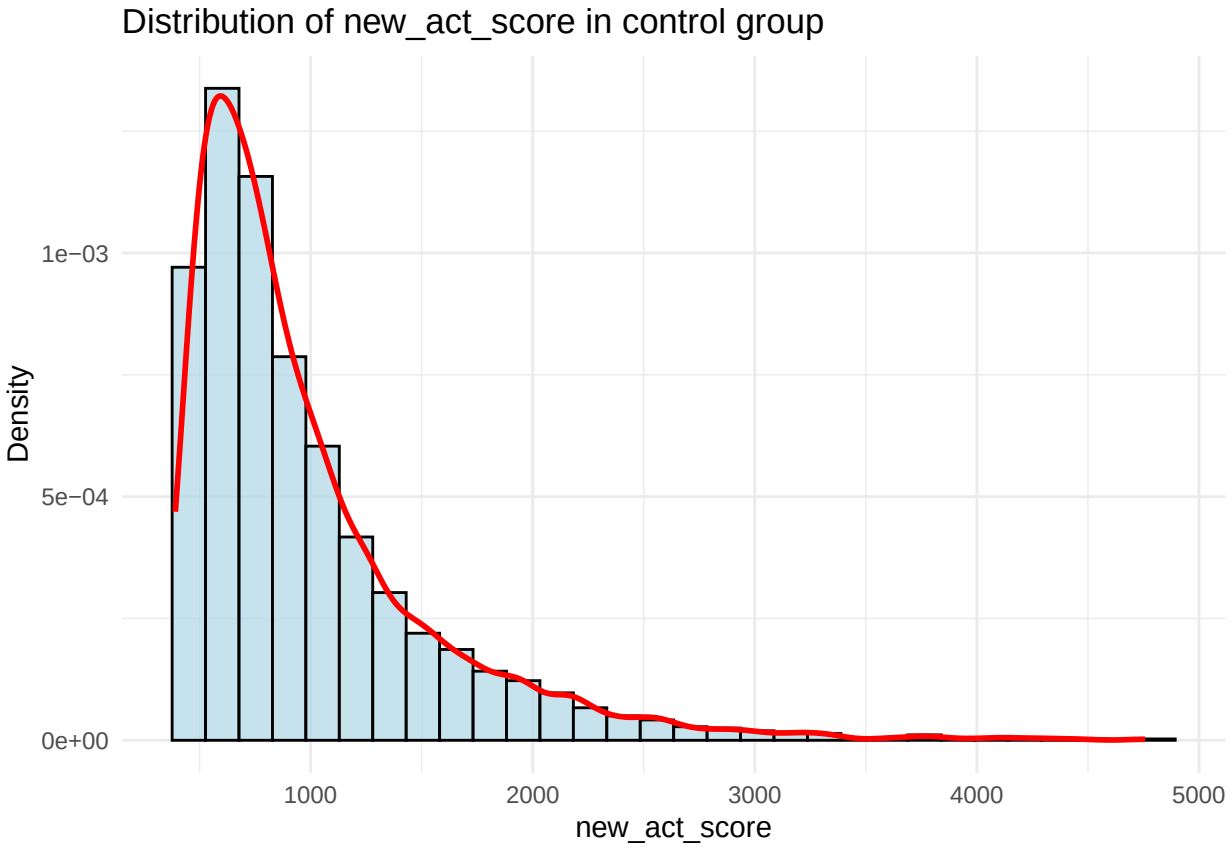
```
knitr::opts_chunk$set(echo = FALSE)
```

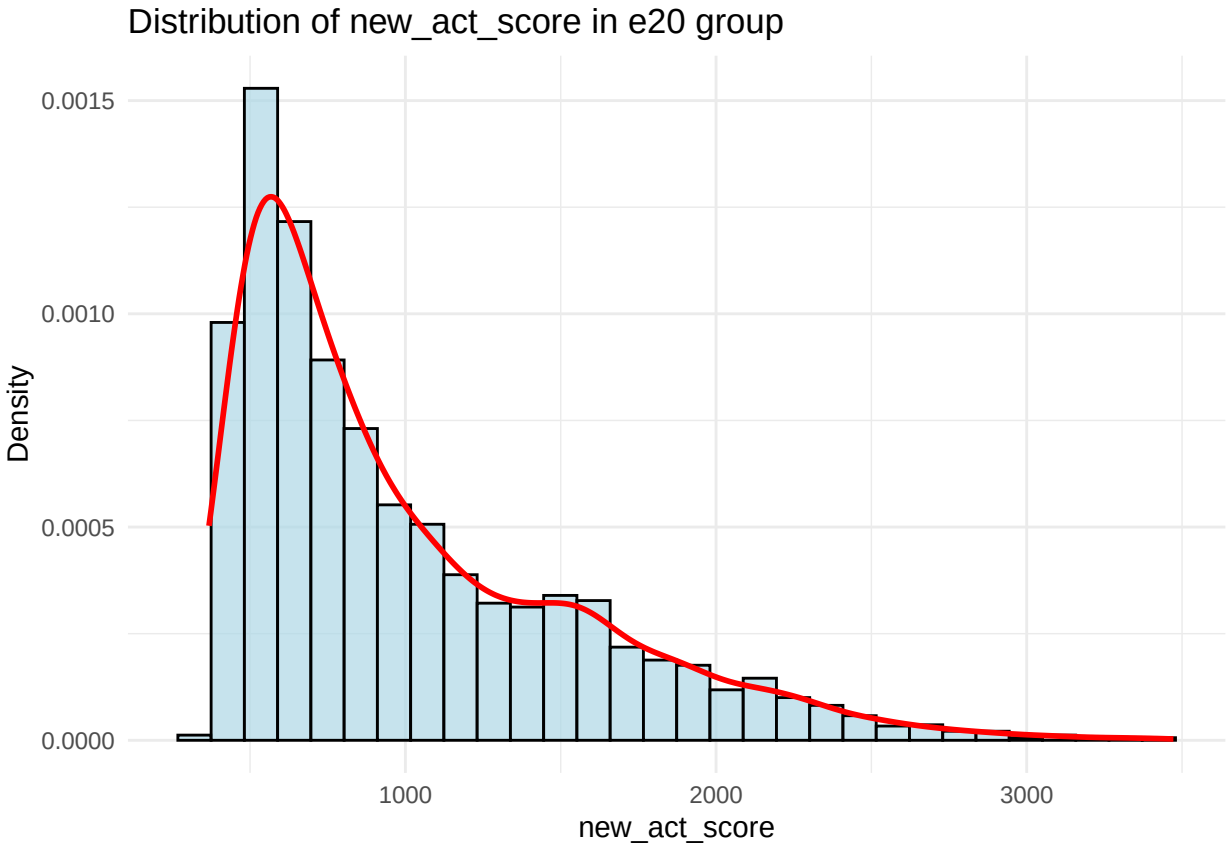
EDA

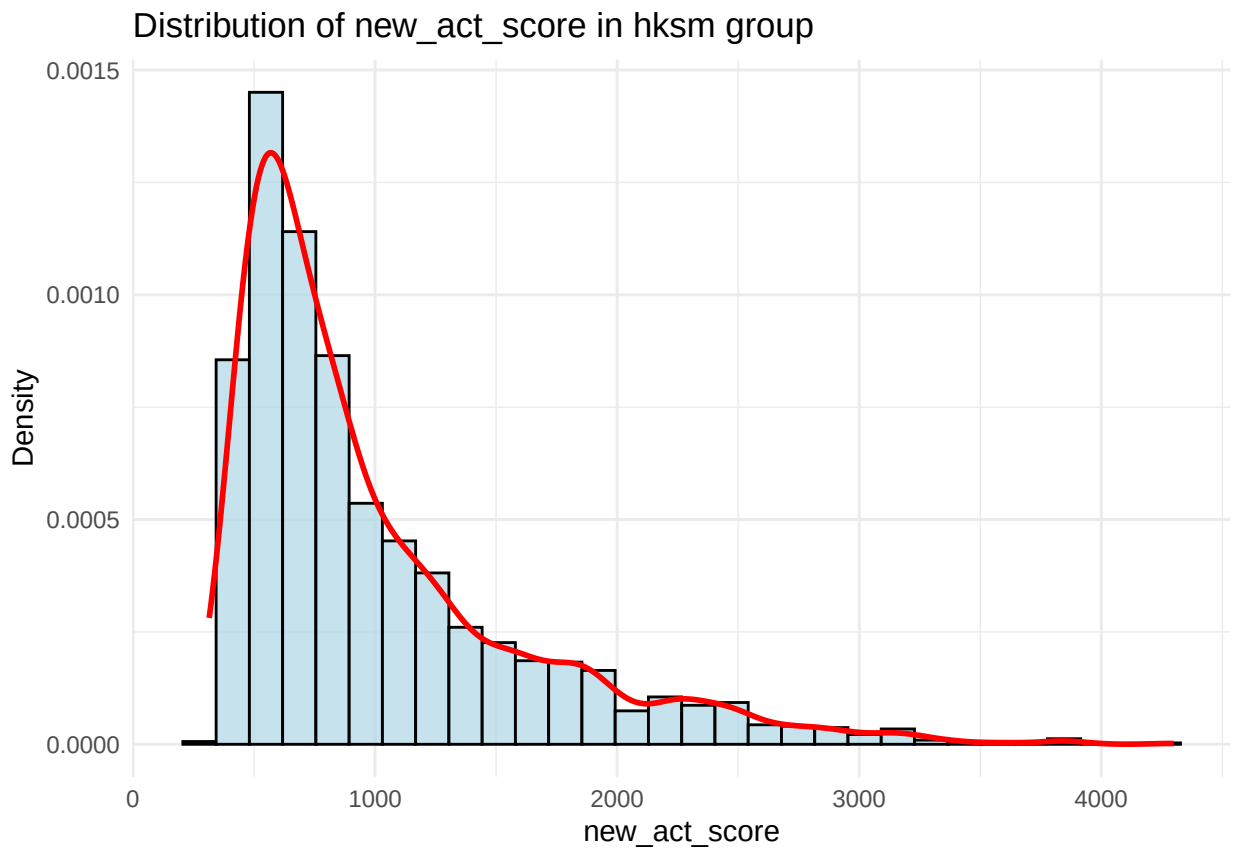
Distribution of activity score in three treatment groups

Each plot shows a density histogram with an overlaid smoothed density curve. All groups show a right-skewed distribution, which means most enhancers have low activity scores, with a smaller number showing very high activity.

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.

## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(density)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```


Distribution of new_act_score in control group
Density
1e−03
5e−04
0e+00
1000
2000
3000
4000
5000
new_act_score


Distribution of new_act_score in e20 group
0.0015
0.0010
0.0005
0.0000
Density
1000
2000
3000
new_act_score

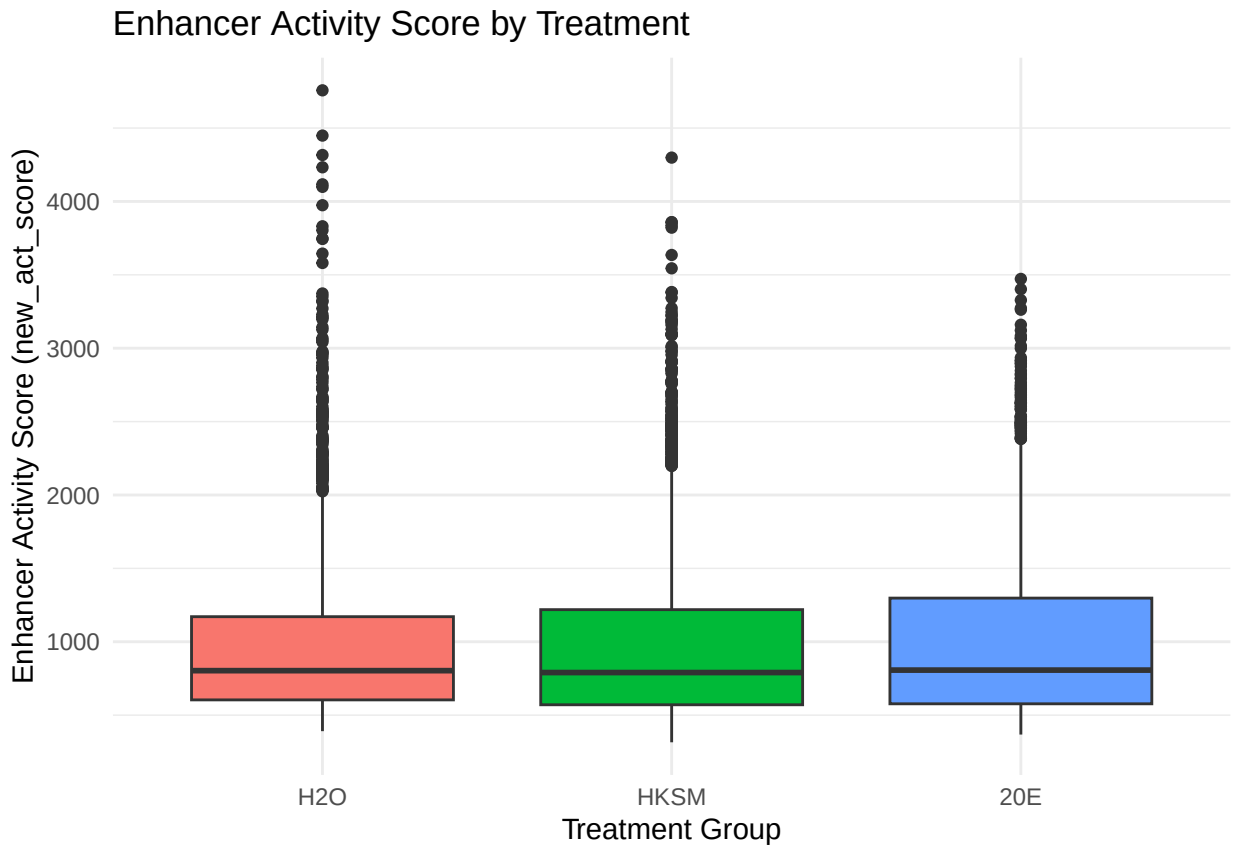


#if they share the same enhancer–no same enhancer among 3 groups

```
## [1] 0
```

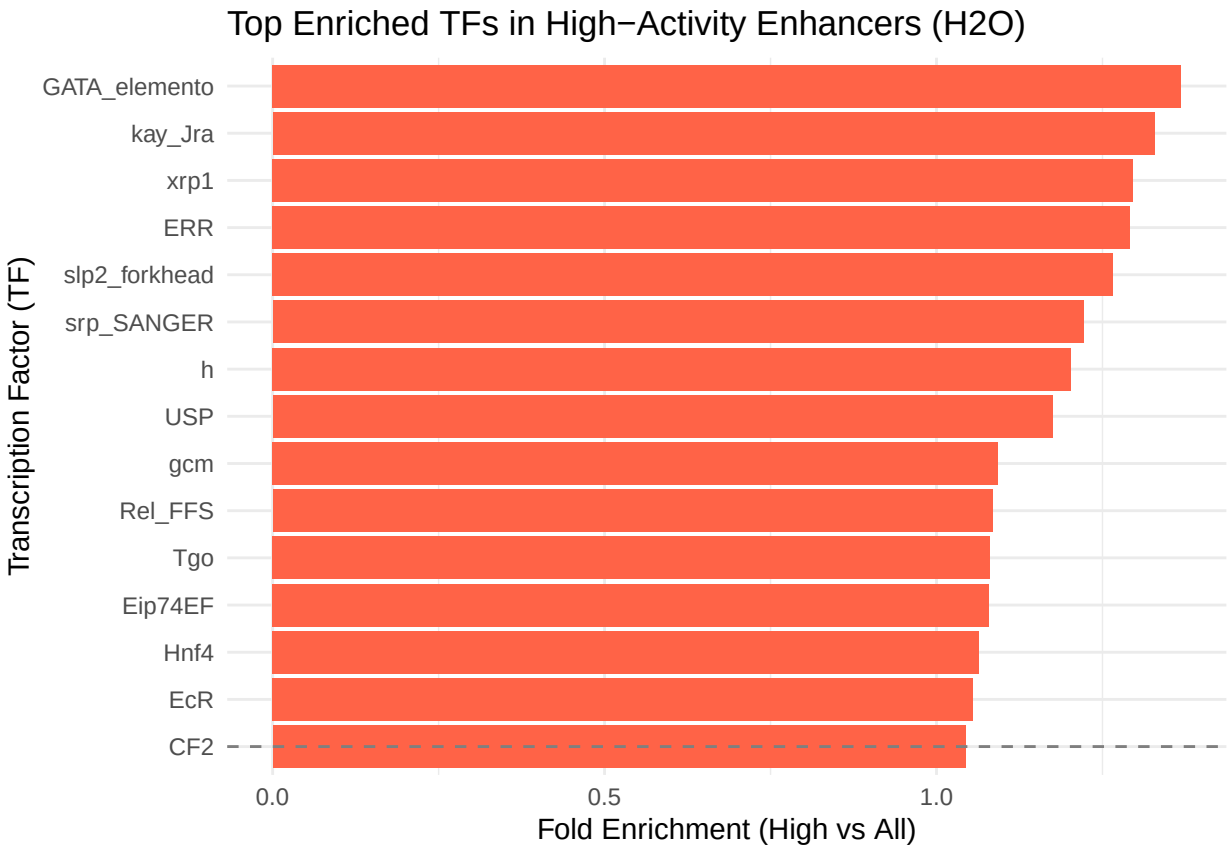
```
## character(0)
```

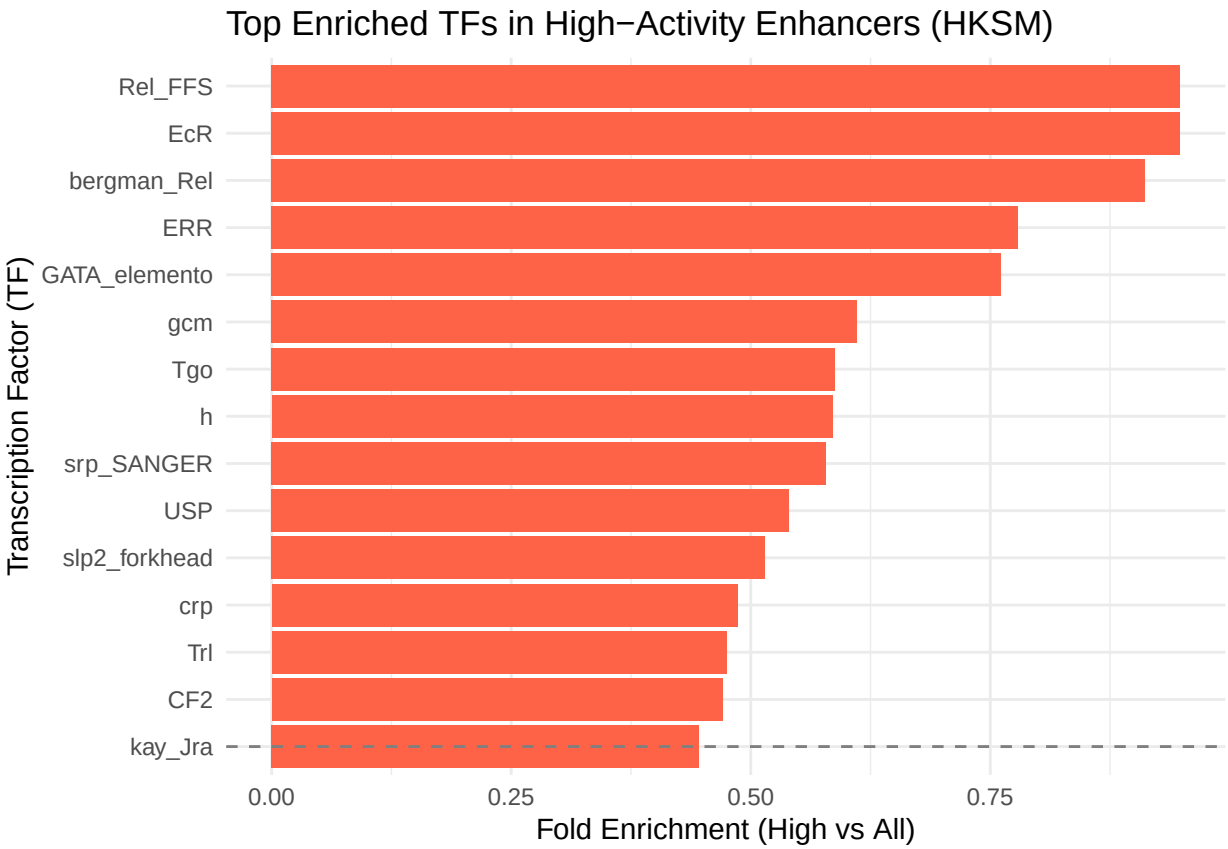
#Here is the boxplot. The box represents the interquartile range (IQR), which includes the middle 50% of the data. The dots outside the whiskers represent outliers, values that fall outside 1.5 times the IQR. These are higher activity scores that deviate from the general distribution.

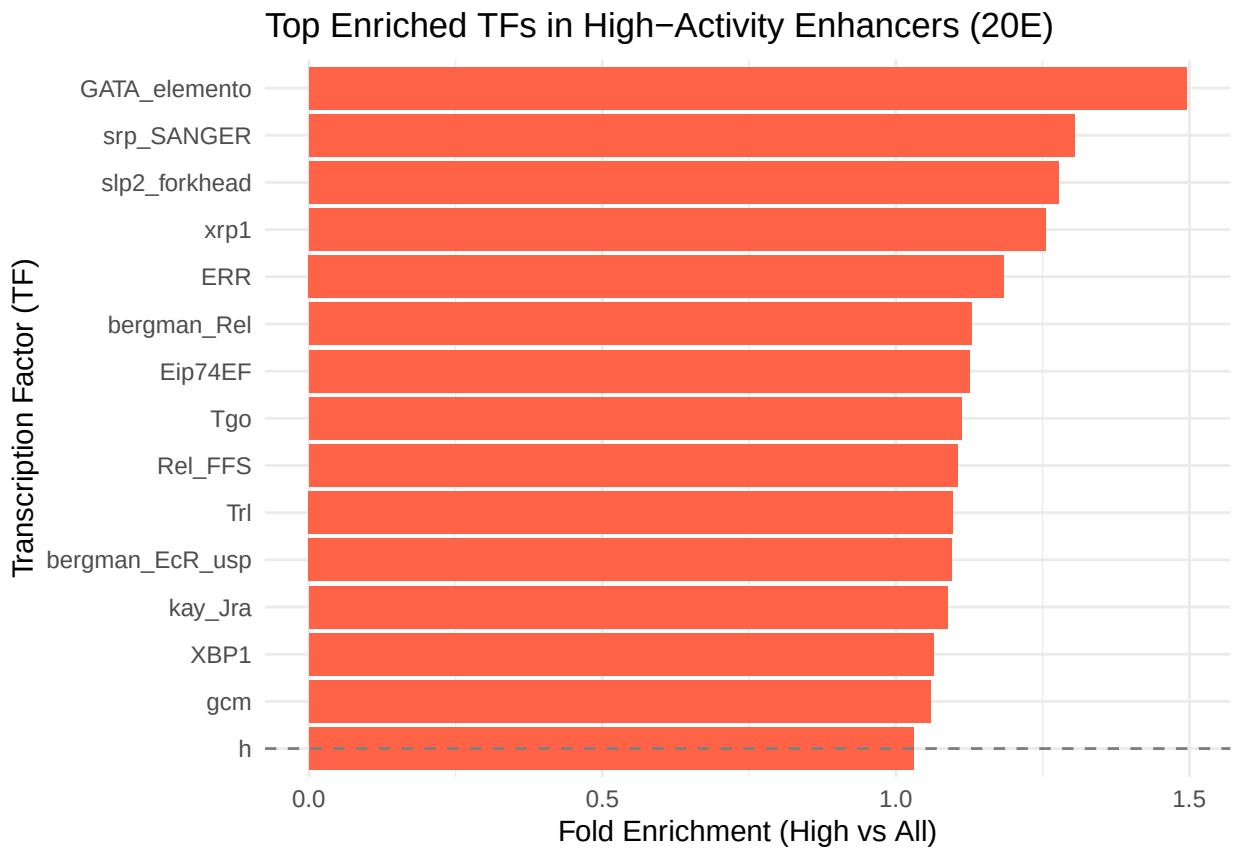


#TF enrichment analysis We are wondering which transcription factors (TFs) are more common in high activity enhancers(activity score over 1000)? The bar plot shown below represents the fold enrichment of transcription factors (TFs) in high-activity enhancers for three treatment groups. It displays the top enriched TFs, comparing their frequency in high-activity enhancers versus the entire dataset.

$$\text{Fold Enrichment} = \frac{\text{Mean TF motif count in high-activity enhancers}}{\text{Mean TF motif count in all enhancers}}$$


Top Enriched TFs in High−Activity Enhancers (H2O)
Transcription Factor (TF)
GATA_elemento
kay_Jra
xrp1
ERR
slp2_forkhead
srp_SANGER
h
USP
gcm
Rel_FFS
Tgo
Eip74EF
Hnf4
EcR
CF2
0.0
0.5
1.0
Fold Enrichment (High vs All)


Top Enriched TFs in High−Activity Enhancers (HKSM)
Transcription Factor (TF)
Rel_FFS
EcR
bergman_Rel
ERR
GATA_elemento
gcm
Tgo
h
srp_SANGER
USP
slp2_forkhead
crp
Trl
CF2
kay_Jra
0.00
0.25
0.50
0.75
Fold Enrichment (High vs All)



The higher the TF ranking, the more likely it is to appear on enhancers with higher activity A Fold Enrichment > 1 means the TF is more common in high-activity enhancers than in general potentially important for enhancer activity. A Fold Enrichment 1 means no real difference. A Fold Enrichment < 1 means the TF is less frequent in high-activity enhancers.