

# Gene Set Enrichment

Antonio Osamu Katagiri Tanaka - A01212611@itesm.mx

May 15, 2020

## Setup

```
# Clear all objects (from the workspace)
rm(list = ls())

# Suppress Warning messages
options(warn = -1)

# Turn off scientific notation like 1e+06
# options(scipen=999)

options(stringsAsFactors = F)

# Load Libs

# # INSTALL with:
# if (!requireNamespace("BiocManager", quietly = TRUE))
#   install.packages("BiocManager")
# BiocManager::install("DESeq2")

library(limma)
library(webshot)

# Use local functions
source("../functions_enrichment.R")
```

## Directives

- Vamos a trabajar con la base de datos de Low Grade Glioma de TCGA comparando los subtipos IDHmut-codel vs IDHmut-non-codel.
- Archivos necesarios: TCGA\_LGG\_data\_GSEA.RData, c6.all.v7.1.symbols.gmt, gsea2-2.2.4.jar y script\_enrichment.R.
- Correr las instrucciones en script\_enrichment.R
- Objetivo: comparar los gene sets más significativos al usar la prueba hipergeométrica y GSEA.
- Tomar en cuenta que al correr GSEA se debe de generar un directorio donde se guardan los resultados.
- Entregar un reporte en blackboard.

**Cargar datos 3 objetos: lgg\_datos (datos de expresión), lgg\_subtypes (clases de las muestras) y onco\_gs (lista de gene sets)**

```
load("../TCGA_LGG_data_GSEA.RData")
```

## Limma: comparar respecto la clase IDHmut-non-codel

```
aux_classes <- ifelse(lgg_subtypes == "IDHmut-non-codel", 1, 0)
lgg_limma <- limma4DS_fdr(lgg_datos, aux_classes, c("IDHmut-non-codel", "IDHmut-codel"))
```

## toptable() is deprecated and will be removed in the future version of limma. Please use topTable() instead

## Obtener genes up y down regulated

```
lgg_up_genes <- rownames(subset(lgg_limma, FC > 0 & q.value < 1e-06))
lgg_dw_genes <- rownames(subset(lgg_limma, FC < 0 & q.value < 1e-06))
```

## Enrichment analysis usando hipergeometrica

```
lgg_up_gs <- gene_set_stats(rownames(lgg_datos), onco_gs, lgg_up_genes)
lgg_dw_gs <- gene_set_stats(rownames(lgg_datos), onco_gs, lgg_dw_genes)
```

## GSEA

### Ordenar genes por signo FC \* -log10(p-value)

```
lgg_limma$Cor <- sign(lgg_limma$FC) * -log10(lgg_limma$p.value)
lgg_cor <- data.frame(Gene=rownames(lgg_limma), Cor=lgg_limma$Cor)
lgg_cor <- lgg_cor[order(lgg_cor$Cor, decreasing=T),]
#Guardar en archivo de texto
write.table(lgg_cor, sep="\t", quote=F, row.names=F, col.names=F, file="LGG_subtypes.rnk")
```

## Java

```
rnk_files <- dir(pattern="*.rnk")
rnk_names <- gsub(pattern="*.rnk", replacement="", rnk_files)
gsea_jar <- "./gsea2-2.2.4.jar"
```

### Crear directorio

```
aux_dir <- "./GSEAscores/"
aux_gmt <- "./c6.all.v7.1.symbols.gmt"
nplot <- 30
```

```
for(i in 1:length(rnk_names)){
  java_command <- paste("java -cp", gsea_jar, "xtools.gsea.GseaPreranked")
  java_command <- paste(java_command, "-gm", aux_gmt, "-collapse false -mode Max_probe -norm meandiv")
  java_command <- paste(java_command, rnk_files[i], "-scoring_scheme weighted -rpt_label", rnk_names[i])
  java_command <- paste(java_command, "-make_sets true -plot_top_x", nplot, "-rnd_seed timestamp -set_rnd")
  java_command <- paste(java_command, aux_dir, "-gui false")
  system(java_command)
}
```

## Hypergeometric Test & GSEA

### Dataset details

- The dataset has 20501 features (genes)
- No probe set => gene symbol collapsing was requested, so all 20501 features were used

### Gene set details

- Gene set size filters (min=5, max=500) resulted in filtering out 0 / 189 gene sets
- The remaining 189 gene sets were used in the analysis

## Gene markers for the na\_pos versus na\_neg comparison

- The dataset has 20501 features (genes) Detailed rank ordered gene list for all features in the dataset Global statistics and plots Plot of p-values vs. NES Global ES histogram

## GSEA Report for Dataset LGG\_subtypes

- 157 / 189 gene sets are upregulated in phenotype na\_pos
- 5 gene sets are significant at FDR < 25%
- 6 gene sets are significantly enriched at nominal pvalue < 1%
- 17 gene sets are significantly enriched at nominal pvalue < 5%

```
# https://wch.github.io/webshot/articles/intro.html
```

```
# https://github.com/wch/webshot
```

```
invisible(capture.output(webshot(  
  "file:///C:/Users/oskat/OneDrive%20-%20Instituto%20Tecnologico%20y%20de%20Estudios%20Superiores%20de%20  
  "/img/pos_snapshot_0.png",  
  cliprect=c(30, 10, 655, 1077))))
```

```
invisible(capture.output(webshot(  
  "file:///C:/Users/oskat/OneDrive%20-%20Instituto%20Tecnologico%20y%20de%20Estudios%20Superiores%20de%20  
  "/img/pos_snapshot_1.png",  
  cliprect=c(1106, 10, 655, 1050))))
```

```
invisible(capture.output(webshot(  
  "file:///C:/Users/oskat/OneDrive%20-%20Instituto%20Tecnologico%20y%20de%20Estudios%20Superiores%20de%20  
  "/img/gsea_report_for_na_pos.png",  
  cliprect=c(0, 0, 1000, 715))))
```

- 32 / 189 gene sets are upregulated in phenotype na\_neg
- 6 gene sets are significantly enriched at FDR < 25%
- 7 gene sets are significantly enriched at nominal pvalue < 1%
- 7 gene sets are significantly enriched at nominal pvalue < 5%

```
invisible(capture.output(webshot(  
  "file:///C:/Users/oskat/OneDrive%20-%20Instituto%20Tecnologico%20y%20de%20Estudios%20Superiores%20de%20  
  "/img/neg_snapshot_0.png",  
  cliprect=c(30, 10, 655, 1077))))
```

```
invisible(capture.output(webshot(  
  "file:///C:/Users/oskat/OneDrive%20-%20Instituto%20Tecnologico%20y%20de%20Estudios%20Superiores%20de%20  
  "/img/neg_snapshot_1.png",  
  cliprect=c(1106, 10, 655, 1050))))
```

```
invisible(capture.output(webshot(  
  "file:///C:/Users/oskat/OneDrive%20-%20Instituto%20Tecnologico%20y%20de%20Estudios%20Superiores%20de%20  
  "/img/gsea_report_for_na_neg.png",  
  cliprect=c(0, 0, 1000, 715))))
```

Table: Snapshot of enrichment results

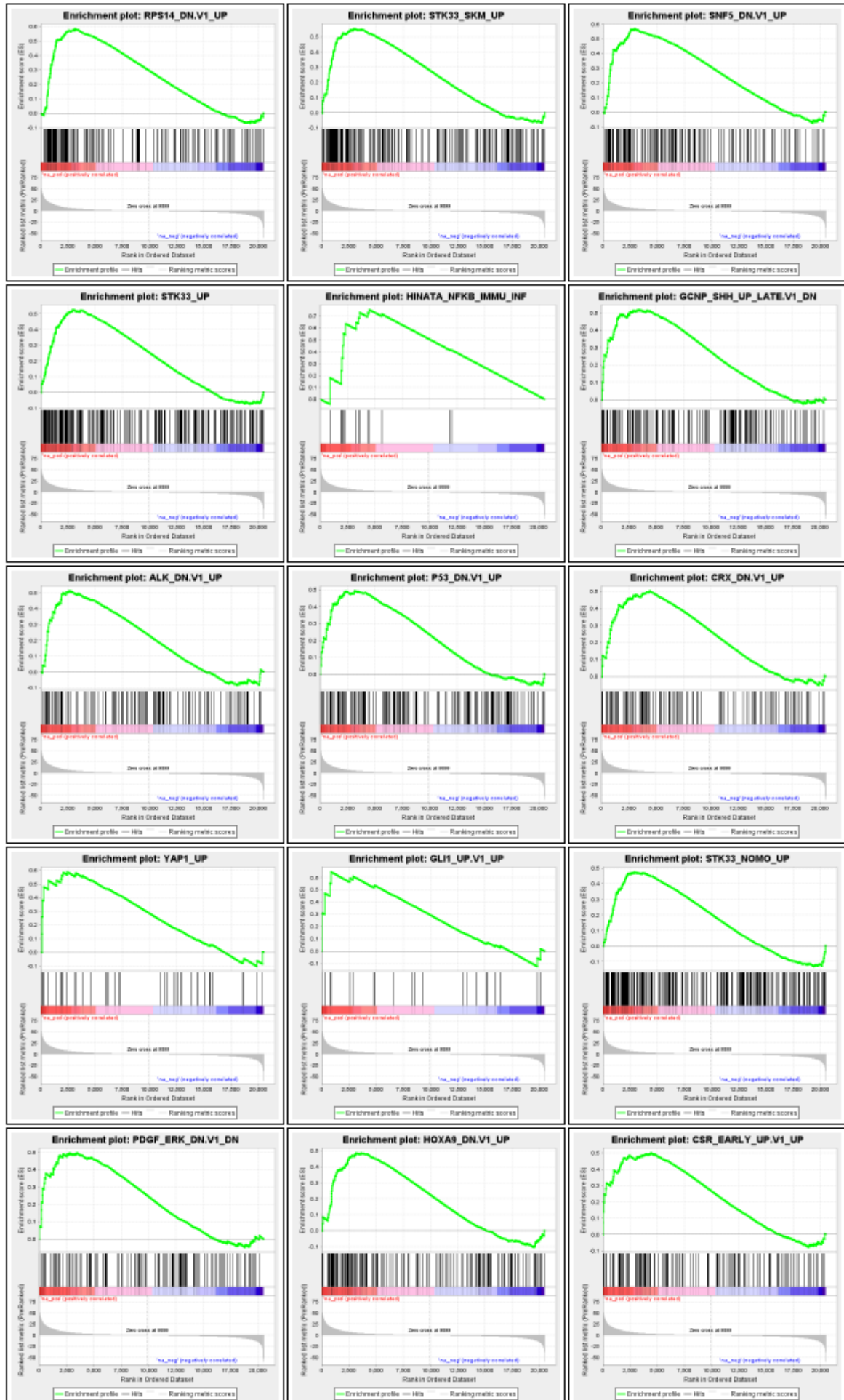


Figure 1: Snapshot of positive enrichment results : part 1

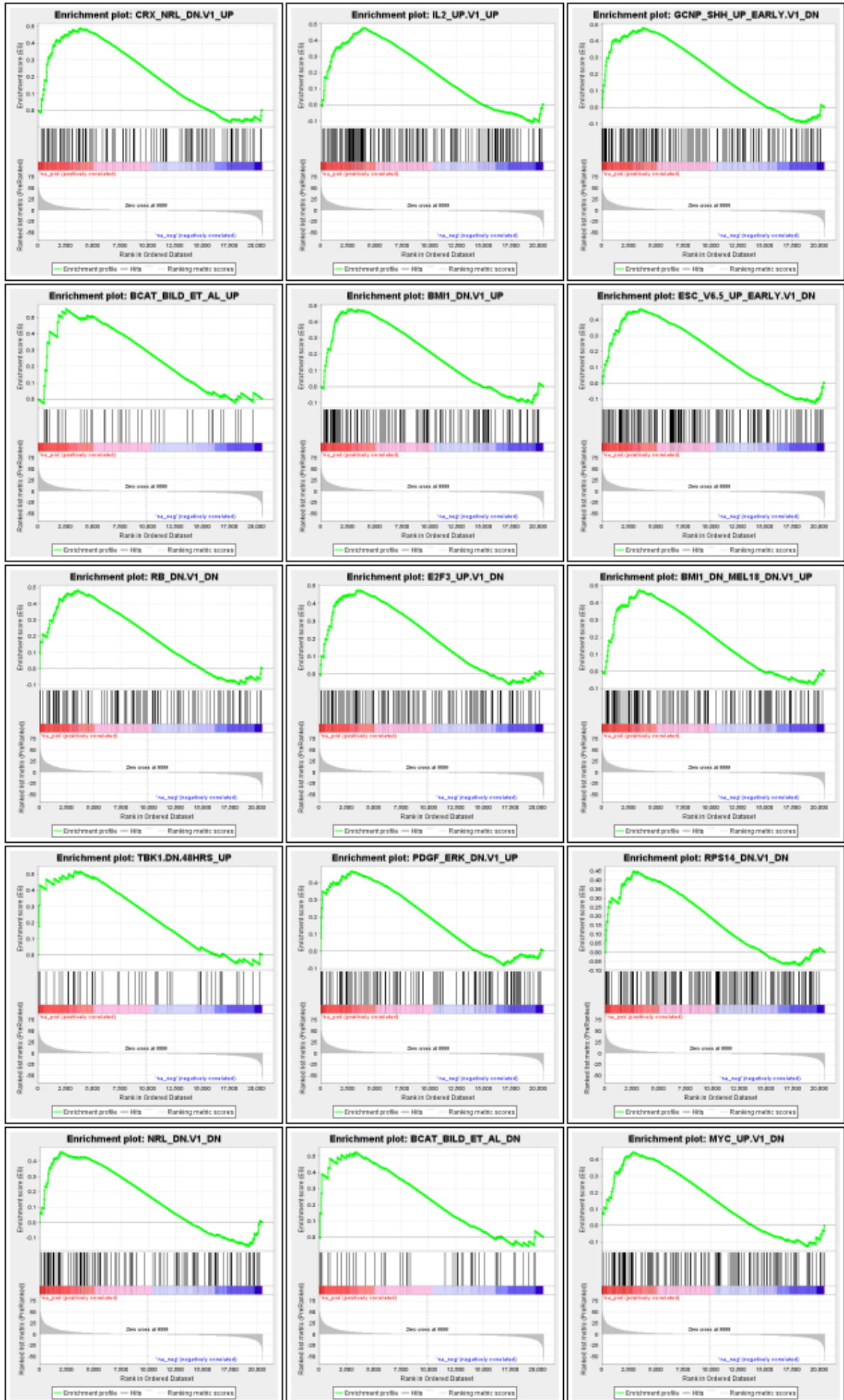


Figure 2: Snapshot of positive enrichment results : part 1

Table: Gene sets enriched in phenotype na [\[plain text format\]](#)

	GS follow link to MSigDB	GS DETAILS	SIZE	ES	NES	NOM p- val	FDR q- val	FWER p- val	RANK AT MAX	LEADING EDGE
1	<a href="#">RPS14_DN.V1_UP</a>	<a href="#">Details ...</a>	179	0.59	1.62	0.000	0.049	0.049	3135	tags=41%, list=15%, signal=48%
2	<a href="#">STK33_SKM_UP</a>	<a href="#">Details ...</a>	236	0.55	1.58	0.000	0.057	0.112	3025	tags=32%, list=15%, signal=37%
3	<a href="#">SNF5_DN.V1_UP</a>	<a href="#">Details ...</a>	154	0.57	1.56	0.001	0.065	0.180	2919	tags=30%, list=14%, signal=35%
4	<a href="#">STK33_UP</a>	<a href="#">Details ...</a>	254	0.53	1.49	0.000	0.142	0.455	3025	tags=31%, list=15%, signal=36%
5	<a href="#">HINATA_NFKB_IMMUN_INF</a>	<a href="#">Details ...</a>	14	0.75	1.48	0.033	0.150	0.549	4492	tags=79%, list=22%, signal=101%
6	<a href="#">GCNP_SHH_UP_LATE.V1_DN</a>	<a href="#">Details ...</a>	169	0.52	1.43	0.005	0.258	0.810	3505	tags=24%, list=17%, signal=28%
7	<a href="#">ALK_DN.V1_UP</a>	<a href="#">Details ...</a>	130	0.51	1.38	0.018	0.393	0.950	2582	tags=22%, list=13%, signal=25%
8	<a href="#">P53_DN.V1_UP</a>	<a href="#">Details ...</a>	182	0.49	1.37	0.014	0.412	0.975	3148	tags=24%, list=15%, signal=28%
9	<a href="#">CRX_DN.V1_UP</a>	<a href="#">Details ...</a>	122	0.50	1.37	0.023	0.380	0.977	4424	tags=39%, list=22%, signal=50%
10	<a href="#">YAP1_UP</a>	<a href="#">Details ...</a>	40	0.59	1.37	0.054	0.343	0.977	2430	tags=25%, list=12%, signal=28%
11	<a href="#">GLI1_UP.V1_UP</a>	<a href="#">Details ...</a>	21	0.64	1.37	0.090	0.313	0.977	843	tags=19%, list=4%, signal=20%
12	<a href="#">STK33_NOMO_UP</a>	<a href="#">Details ...</a>	260	0.48	1.35	0.005	0.353	0.994	2915	tags=29%, list=14%, signal=33%
13	<a href="#">PDGF_ERK_DN.V1_DN</a>	<a href="#">Details ...</a>	140	0.50	1.34	0.029	0.372	0.996	3443	tags=21%, list=17%, signal=26%
14	<a href="#">HOXA9_DN.V1_UP</a>	<a href="#">Details ...</a>	167	0.49	1.34	0.026	0.352	0.996	3156	tags=31%, list=15%, signal=37%
15	<a href="#">CSR_EARLY_UP.V1_UP</a>	<a href="#">Details ...</a>	135	0.50	1.34	0.026	0.334	0.997	4426	tags=36%, list=22%, signal=45%

Figure 3: Top 15 detailed enrichment results: gsea report for positives



Table: Snapshot of enrichment results

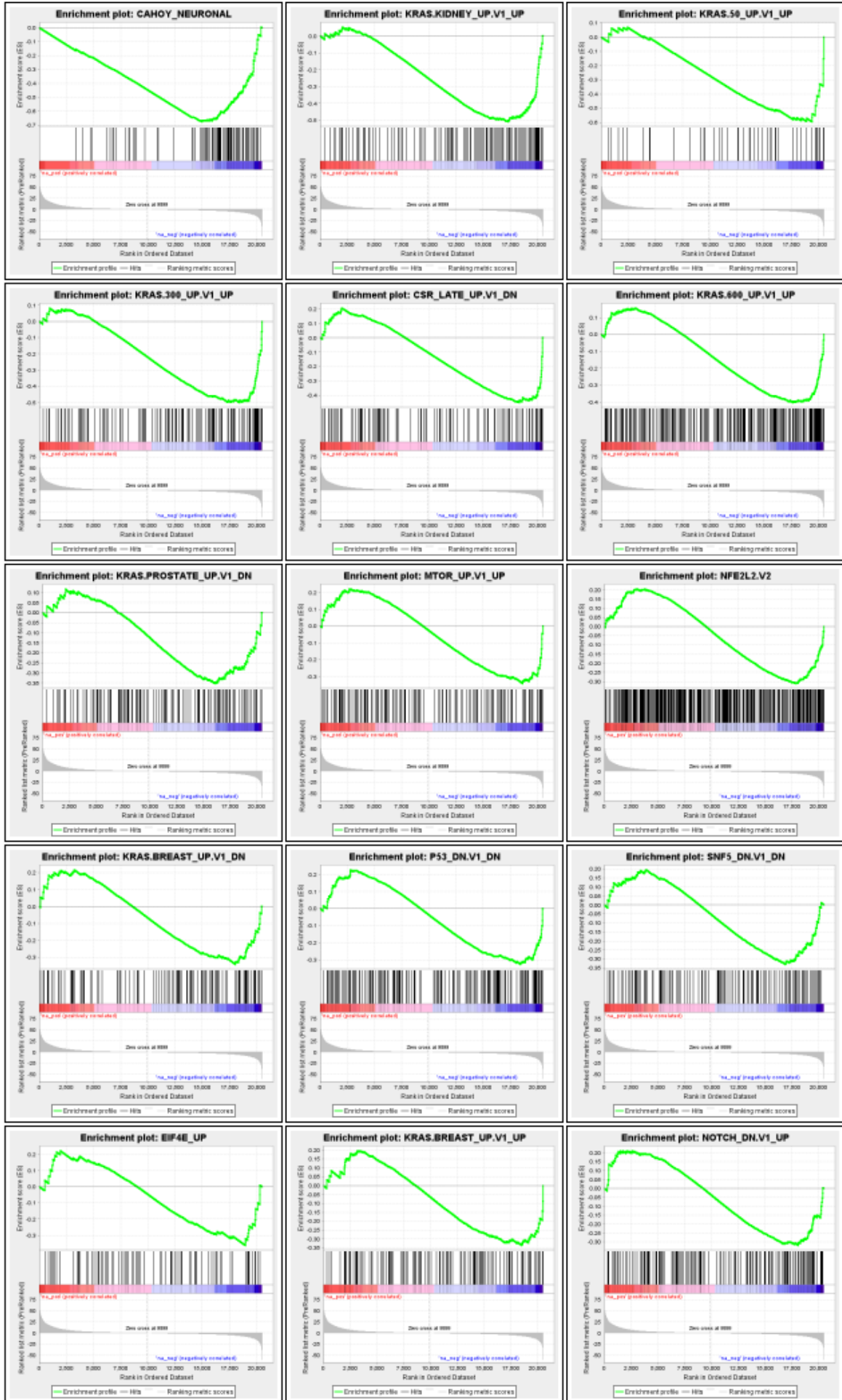


Figure 4: Snapshot of negative enrichment results : part 1

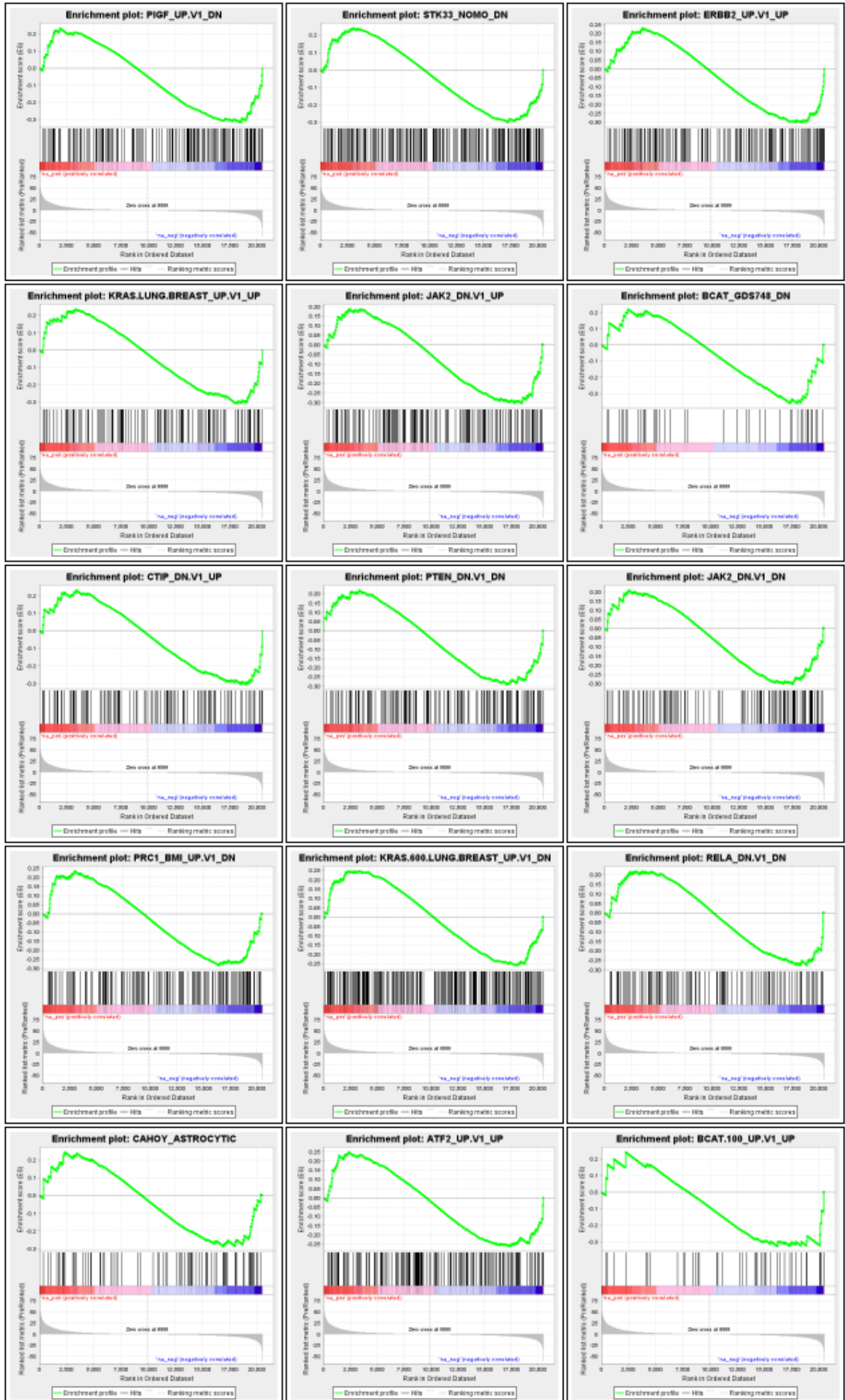


Figure 5: Snapshot of negative enrichment results : part 2



Table: Gene sets enriched in phenotype na [\[plain text format\]](#)

	GS follow link to MSigDB	GS DETAILS	SIZE	ES	NES	NOM p- val	FDR q- val	FWER p- val	RANK AT MAX	LEADING EDGE
1	<a href="#">CAHOY_NEURONAL</a>	<a href="#">Details ...</a>	94	-0.68	-2.14	0.000	0.000	0.000	5523	tags=74%, list=27%, signal=101%
2	<a href="#">KRAS.KIDNEY_UP.V1_UP</a>	<a href="#">Details ...</a>	137	-0.51	-1.73	0.000	0.006	0.009	3202	tags=34%, list=16%, signal=40%
3	<a href="#">KRAS.50_UP.V1_UP</a>	<a href="#">Details ...</a>	46	-0.60	-1.70	0.000	0.007	0.015	1088	tags=24%, list=5%, signal=25%
4	<a href="#">KRAS.300_UP.V1_UP</a>	<a href="#">Details ...</a>	135	-0.50	-1.68	0.000	0.007	0.020	2068	tags=23%, list=10%, signal=25%
5	<a href="#">CSR.LATE_UP.V1_DN</a>	<a href="#">Details ...</a>	130	-0.45	-1.49	0.000	0.049	0.175	2168	tags=26%, list=11%, signal=29%
6	<a href="#">KRAS.600_UP.V1_UP</a>	<a href="#">Details ...</a>	260	-0.40	-1.48	0.000	0.047	0.197	2824	tags=24%, list=14%, signal=27%
7	<a href="#">KRAS.PROSTATE_UP.V1_DN</a>	<a href="#">Details ...</a>	133	-0.35	-1.19	0.065	0.437	0.900	4325	tags=28%, list=21%, signal=35%
8	<a href="#">MTOR_UP.V1_UP</a>	<a href="#">Details ...</a>	153	-0.34	-1.17	0.090	0.450	0.937	1863	tags=15%, list=9%, signal=16%
9	<a href="#">NFE2L2.V2</a>	<a href="#">Details ...</a>	406	-0.31	-1.16	0.000	0.426	0.943	2574	tags=17%, list=13%, signal=19%
10	<a href="#">KRAS.BREAST_UP.V1_DN</a>	<a href="#">Details ...</a>	129	-0.34	-1.16	0.112	0.397	0.953	2438	tags=22%, list=12%, signal=24%
11	<a href="#">P53_DN.V1_DN</a>	<a href="#">Details ...</a>	184	-0.33	-1.13	0.107	0.450	0.976	2034	tags=16%, list=10%, signal=17%
12	<a href="#">SNF5_DN.V1_DN</a>	<a href="#">Details ...</a>	147	-0.33	-1.13	0.144	0.414	0.977	3601	tags=30%, list=18%, signal=36%
13	<a href="#">EIF4E_UP</a>	<a href="#">Details ...</a>	86	-0.36	-1.12	0.219	0.408	0.982	1534	tags=15%, list=7%, signal=16%
14	<a href="#">KRAS.BREAST_UP.V1_UP</a>	<a href="#">Details ...</a>	129	-0.34	-1.11	0.175	0.424	0.991	1998	tags=12%, list=10%, signal=14%
15	<a href="#">NOTCH_DN.V1_UP</a>	<a href="#">Details ...</a>	170	-0.32	-1.11	0.125	0.396	0.991	2385	tags=20%, list=12%, signal=22%

Figure 6: Top 15 detailed enrichment results: gsea report for NOT positives