Direct inhibition of the NOTCH TF. Differenctial expression analysis

Fernando Freire

May 27, 2019

Contents

1 Array expression profiling: direct inhibition of the NOTCH transcription complex

Goals

We will try to reproduce some of the differential expression results obtained by the paper *Direct inhibition of the NOTCH transcription complex*. In this paper, Moellering et al. try to design and synthesize a peptide able to inhibit NOTCH transcription factor for the treatment of individuals affected by Acute lymphoblastic leukemia (T-ALL).

NOTCH proteins regulate signaling pathways involved in cellular differentiation, proliferation and apoptosis. Overactive Notch signaling as been observed in numerous cancers and has been extensively studied in the context of T-ALL where more than 50% of pateints have mutant NOTCH1. Small molecule modulators of these proteins would be important for understanding the role of NOTCH proteins in malignant and normal biological processes.

The authors measure the global changes in gene expression upon treatment of the human T-ALL cell lines HPB-ALL and KOPT-K1 with either vehicle alone dimethylsulphoxide (DMSO) control or the designed peptite SAHM1, an alpha-helical hydrocarbon peptide derived from the MAML1 co-activator protein.

Summary design

They triplicate cultures of KOPT-K1 or HPB-ALL cells that were treated with either DMSO or SAHM1 (20 uM) for 24 hours. Total RNA was extracted and hybridized to Affymetrix human U133 plus 2.0 microarrays (three arrays per treatment per cell line for a total of 12 arrays).

1.1 Pipeline

1.2 Python imports

Script 1.2.1 (python)

```python
import rpy2.rinterface
%reload_ext rpy2.ipython
```

1.3 R imports

Script 1.3.1 (R)

```r
%%R
##1. Load libraries
library("affy")
library("limma")
library("genefilter")
library(simpleaffy)
library(hgu133plus2.db)
wd <- "/Users/nandoide/misc_work/Desktop/uni/TRREP"
setwd(wd)
```

1.4 Functions

Script 1.4.1 (R)

```r
%%R

import_CEL <- function(pattern) {
    # Import CEL files into affiBatch object
    files <- list.files(pattern = pattern)
    names <- gsub(".CEL.gz", "", files)
    abatch <- ReadAffy(filenames = files,  compress = TRUE, sampleNames = names)
    return(abatch)
}

create_eset <- function(affyBatch) {
    # Generates object eset (class ExprSet),
    # expresso function provides intensities in log scale
    return(expresso(affyBatch,
             bg.correct = TRUE,
             bgcorrect.method="rma",
             normalize = TRUE,
             normalize.method="quantiles",
             pmcorrect.method="pmonly",
             summary.method="medianpolish",
             verbose = TRUE))
}

boxplots <- function(affyBatch, eset, title) {
    # Generate BOXPLOTS before and after normalization
    boxplot(affyBatch,
        main=paste0("Boxplot Before Normalization ", title),
        col = "lightgrey")
    df_eset <- as.data.frame(exprs(eset))

    boxplot(data.frame(df_eset),
        main=paste0("Boxplot After Normalization (log scale) ", title), col = "white")
}

create_TopTable <- function(eset, control_samples=c(1,1,1,0,0,0)) {
    # Generate Toptable with limma

    # Data filtering using IQR.
    esetIQR <- varFilter(eset, var.func=IQR, var.cutoff=0.5, filterByQuantile=TRUE)

    # Differential expression analysis.#######
    r_control_samples <- 1 - control_samples
    design <- cbind(DMSO=control_samples, SAHM1=r_control_samples)

    rownames(design) <- colnames(eset)

    #7. Contrasts matrix.
    cont.matrix <- makeContrasts(DMSO_SAHM1 = SAHM1 - DMSO, levels = design)

    #8. Obtaining differentially expressed genes (DEGs)
    #Linear model and eBayes
    fit <- lmFit(esetIQR, design)  ##getting DEGs from IQR
    fit2 <- contrasts.fit(fit, cont.matrix)
    fit2 <- eBayes(fit2)

    #Table with DEGs results
    toptableIQR <- topTable(fit2, number=dim(exprs(esetIQR))[1], adjust.method="BH", sort.by="p")
    return(toptableIQR)
}

anotate_TopTable <- function(toptable) {
    # Obtain gene names from probe names and chip symbol dataset
    probenames_toptable <- as.character(rownames(toptable ))
    genesymbols_toptable <- as.character(mget(probenames_toptable, hgu133plus2SYMBOL))
    # Annotated gene table
    toptable_anot <- cbind(Symbol = genesymbols_toptable, toptable)
    return(toptable_anot)
}

generank_table <- function(toptable, rnk.file) {
    # Generate rank of table top 50 upregulared and top 50 downregulated from 250 better
    # adjustes p-values
    more_significant = toptable[order(toptable$adj.P.Val, decreasing = FALSE),][1:250,]
    up_50 = more_significant[which(toptable$logFC > 0), ] [1:50,] # up reg top 50
    down_50 = more_significant[which(toptable$logFC < 0), ] [1:50,] # down reg top 50

    print("Down-regulated genes")
    print(down_50[order(down_50$logFC), c(1,2,6)])

    print("Up-regulated genes")
    print(up_50[order(up_50$logFC), c(1,2,6)])

    d <- rbind(down_50[order(down_50$logFC), c(1,2,6)], up_50[order(up_50$logFC), c(1,2,6)])

    df <- data.frame(d$Symbol, d$logFC)
    write.table(df,row.names=FALSE,col.names=FALSE,
                quote=FALSE,sep="\t",file=paste0(rnk.file, ".rnk"))
}
```

1.5 Quality control

Script 1.5.1 (R)

```r
%%R

setwd("GSE18198_data")
affyBatch = import_CEL("*")
setwd(wd)
```

```r
affyBatch_MAS5 <- call.exprs(affyBatch,"mas5")
qcs <- qc(affyBatch, affyBatch_MAS5)
plot(qcs)
qcs
```

```
An object of class "QCStats"
Slot "scale.factors":
 [1] 0.4624769 0.9923886 0.5749658 0.5299016 0.4725059 0.4445994 1.4066966
 [8] 1.2441075 1.3089160 1.9466663 2.1150710 2.3084671

Slot "target":
[1] 100

Slot "percent.present":
  HPB_DMSO_01.present   HPB_DMSO_02.present   HPB_DMSO_03.present
             45.75583              41.61500              44.97485
 HPB_SAHM1_01.present  HPB_SAHM1_02.present  HPB_SAHM1_03.present
             45.39369              45.37906              46.93004
 KOPT_DMSO_01.present  KOPT_DMSO_02.present  KOPT_DMSO_03.present
             39.18793              39.72016              40.17558
KOPT_SAHM1_01.present KOPT_SAHM1_02.present KOPT_SAHM1_03.present
             38.46182              38.39049              38.57888

Slot "average.background":
  HPB_DMSO_01   HPB_DMSO_02   HPB_DMSO_03  HPB_SAHM1_01  HPB_SAHM1_02
     71.40948      51.82766      67.15371      73.25355      78.06755
 HPB_SAHM1_03  KOPT_DMSO_01  KOPT_DMSO_02  KOPT_DMSO_03 KOPT_SAHM1_01
     64.75102      58.67351      58.61665      56.23122      53.46125
KOPT_SAHM1_02 KOPT_SAHM1_03
     44.98804      45.99328

Slot "minimum.background":
  HPB_DMSO_01   HPB_DMSO_02   HPB_DMSO_03  HPB_SAHM1_01  HPB_SAHM1_02
     67.78476      49.90441      64.95388      68.98329      74.59893
 HPB_SAHM1_03  KOPT_DMSO_01  KOPT_DMSO_02  KOPT_DMSO_03 KOPT_SAHM1_01
     61.72838      56.40174      54.42714      53.83356      50.28008
KOPT_SAHM1_02 KOPT_SAHM1_03
     42.42914      43.87834

Slot "maximum.background":
  HPB_DMSO_01   HPB_DMSO_02   HPB_DMSO_03  HPB_SAHM1_01  HPB_SAHM1_02
     73.33486      52.82101      69.57489      75.70203      79.83810
 HPB_SAHM1_03  KOPT_DMSO_01  KOPT_DMSO_02  KOPT_DMSO_03 KOPT_SAHM1_01
     66.69588      60.24642      60.27544      57.78016      54.91569
KOPT_SAHM1_02 KOPT_SAHM1_03
     45.78716      46.82307
```

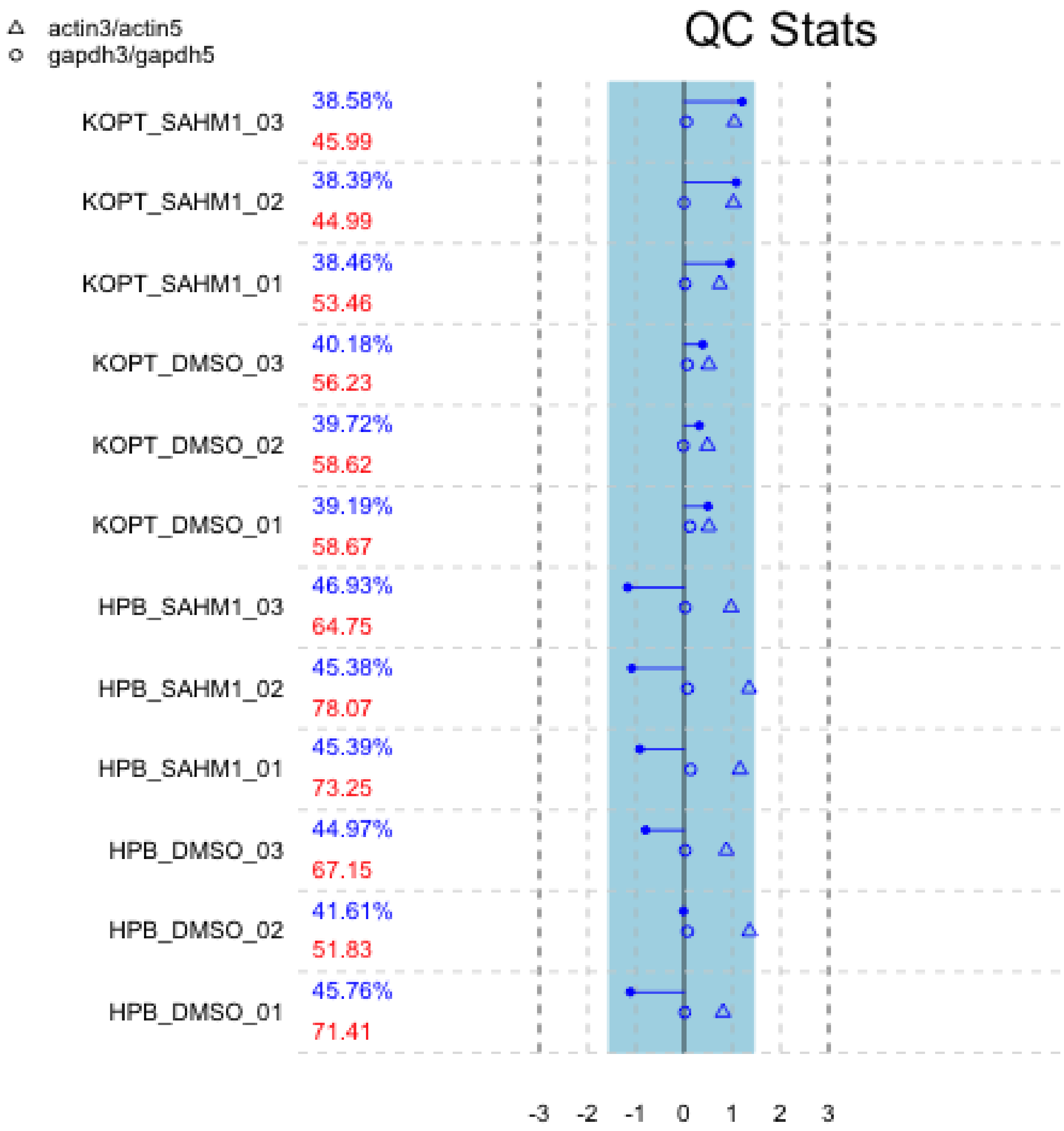
```
Slot "spikes":
              AFFX-r2-Ec-bioB-3_at AFFX-r2-Ec-bioC-3_at AFFX-r2-Ec-bioD-3_at
HPB_DMSO_01               8.248482             9.704242             12.15884
HPB_DMSO_02               9.488241            11.021774             13.39970
HPB_DMSO_03               8.174298             9.589539             12.00550
HPB_SAHM1_01              8.381753             9.772156             12.21394
HPB_SAHM1_02              8.159541             9.712199             12.13458
HPB_SAHM1_03              7.898193             9.390437             11.90994
KOPT_DMSO_01              8.920814             8.512096             12.74309
KOPT_DMSO_02              9.007404             8.715061             12.82763
KOPT_DMSO_03              8.978651             8.593016             12.75616
KOPT_SAHM1_01             9.710773             9.355043             13.32068
KOPT_SAHM1_02            10.167828             9.646911             13.70326
KOPT_SAHM1_03            10.414865             9.875183             13.89010
              AFFX-r2-P1-cre-3_at
HPB_DMSO_01              13.20270
HPB_DMSO_02              14.44458
HPB_DMSO_03              13.18265
HPB_SAHM1_01             13.31337
HPB_SAHM1_02             13.19213
HPB_SAHM1_03             12.97417
KOPT_DMSO_01             14.04051
KOPT_DMSO_02             14.10125
KOPT_DMSO_03             14.00066
KOPT_SAHM1_01            14.63294
KOPT_SAHM1_02            14.82687
KOPT_SAHM1_03            15.01834

Slot "qc.probes":
              AFFX-HSAC07/X00351_3_at AFFX-HSAC07/X00351_5_at
HPB_DMSO_01                  12.73093                11.92493
HPB_DMSO_02                  13.63221                12.27401
HPB_DMSO_03                  12.77779                11.90235
HPB_SAHM1_01                 12.79605                11.63738
HPB_SAHM1_02                 12.66274                11.31439
HPB_SAHM1_03                 12.56844                11.59679
KOPT_DMSO_01                 13.53358                13.02196
KOPT_DMSO_02                 13.49021                13.00410
KOPT_DMSO_03                 13.50476                12.99276
KOPT_SAHM1_01                13.78372                13.04395
KOPT_SAHM1_02                13.85151                12.82967
KOPT_SAHM1_03                13.88275                12.84090
              AFFX-HSAC07/X00351_M_at AFFX-HUMGAPDH/M33197_3_at
HPB_DMSO_01                  12.19455                  12.87855
HPB_DMSO_02                  12.84310                  13.87735
HPB_DMSO_03                  12.24595                  12.89411
HPB_SAHM1_01                 12.17076                  12.98237
HPB_SAHM1_02                 11.91872                  12.87606
```

```
HPB_SAHM1_03                 11.97960                  12.69856
KOPT_DMSO_01                 13.21759                  13.80409
KOPT_DMSO_02                 13.20014                  13.72550
KOPT_DMSO_03                 13.16413                  13.78359
KOPT_SAHM1_01                13.33575                  14.13775
KOPT_SAHM1_02                13.31856                  14.20714
KOPT_SAHM1_03                13.29794                  14.34094
              AFFX-HUMGAPDH/M33197_5_at AFFX-HUMGAPDH/M33197_M_at
HPB_DMSO_01                    12.86110                  12.72697
HPB_DMSO_02                    13.80521                  13.68863
HPB_DMSO_03                    12.86872                  12.70080
HPB_SAHM1_01                   12.84679                  12.78306
HPB_SAHM1_02                   12.80222                  12.66854
HPB_SAHM1_03                   12.67308                  12.53523
KOPT_DMSO_01                   13.68232                  13.59723
KOPT_DMSO_02                   13.74016                  13.55294
KOPT_DMSO_03                   13.71389                  13.56799
KOPT_SAHM1_01                  14.11439                  13.93807
KOPT_SAHM1_02                  14.20615                  13.97154
KOPT_SAHM1_03                  14.29181                  14.05688

Slot "bioBCalls":
  HPB_DMSO_01.present   HPB_DMSO_02.present   HPB_DMSO_03.present
                  "P"                   "P"                   "P"
 HPB_SAHM1_01.present  HPB_SAHM1_02.present  HPB_SAHM1_03.present
                  "P"                   "P"                   "P"
 KOPT_DMSO_01.present  KOPT_DMSO_02.present  KOPT_DMSO_03.present
                  "P"                   "P"                   "P"
KOPT_SAHM1_01.present KOPT_SAHM1_02.present KOPT_SAHM1_03.present
                  "P"                   "P"                   "P"

Slot "arraytype":
[1] "hgu133plus2cdf"
```



1.6 Load raw data

Script 1.6.1 (R)

```
%%R
setwd("GSE18198_data")
affyBatch_HPB = import_CEL("HPB*")
affyBatch_KOPT = import_CEL("KOPT*")
setwd(wd)
```

1.7 Create expression sets

Script 1.7.1 (R)

```r
%%R
eset_HPB <- create_eset(affyBatch_HPB)
eset_KOPT <- create_eset(affyBatch_KOPT)
```

```
background correction: rma
normalization: quantiles
PM/MM correction : pmonly
expression values: medianpolish
background correcting...done.
normalizing...done.
54675 ids to be processed
|                    |
|####################|
background correction: rma
normalization: quantiles
PM/MM correction : pmonly
expression values: medianpolish
background correcting...done.
normalizing...done.
54675 ids to be processed
|                    |
|####################|
```

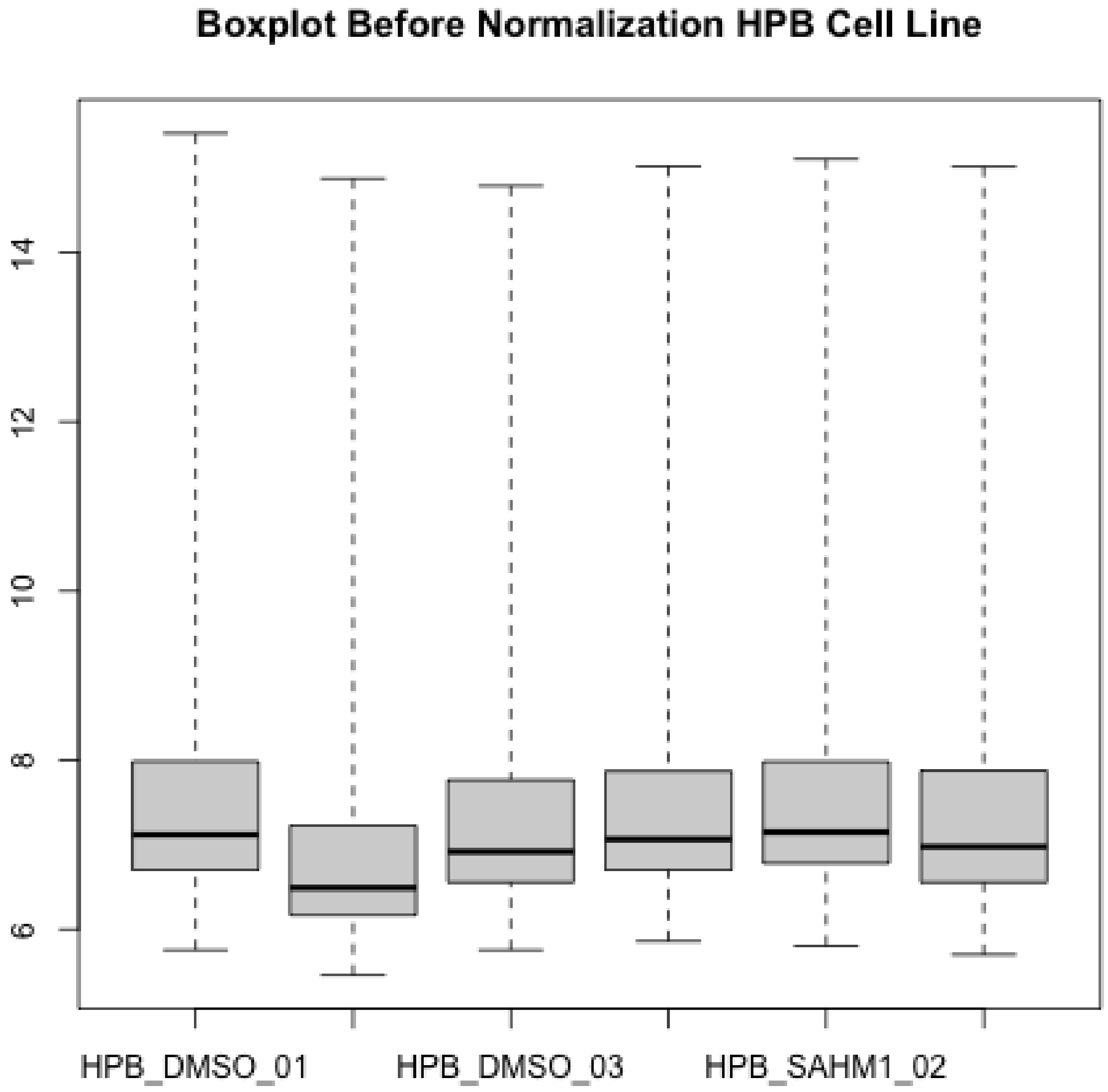
Script 1.7.2 (R)

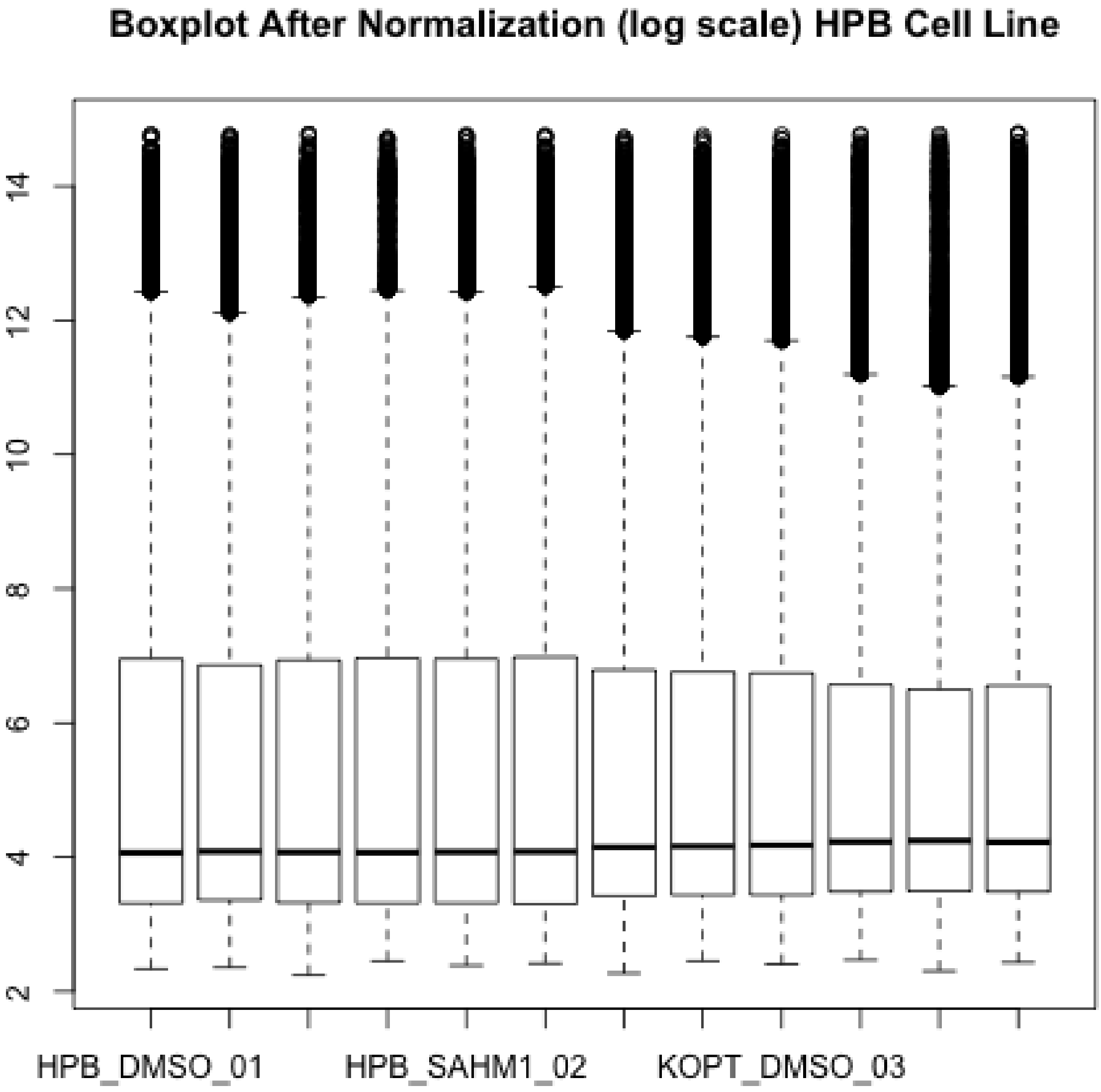
```r
%%R
save(eset_HPB, file="eset_HPB.RData")
save(eset_KOPT, file="eset_KOPT.RData")
```

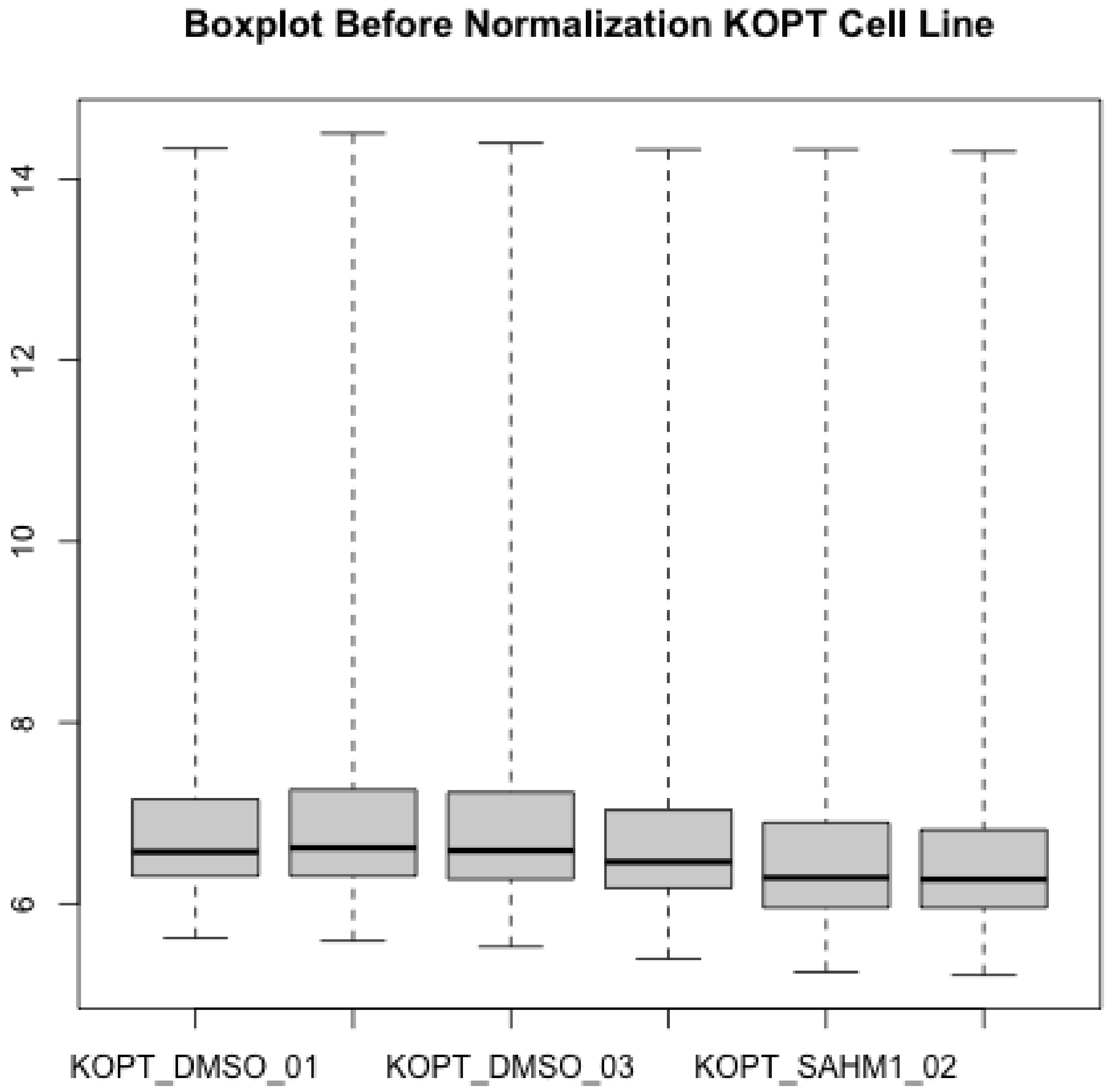
1.8 Quality plots

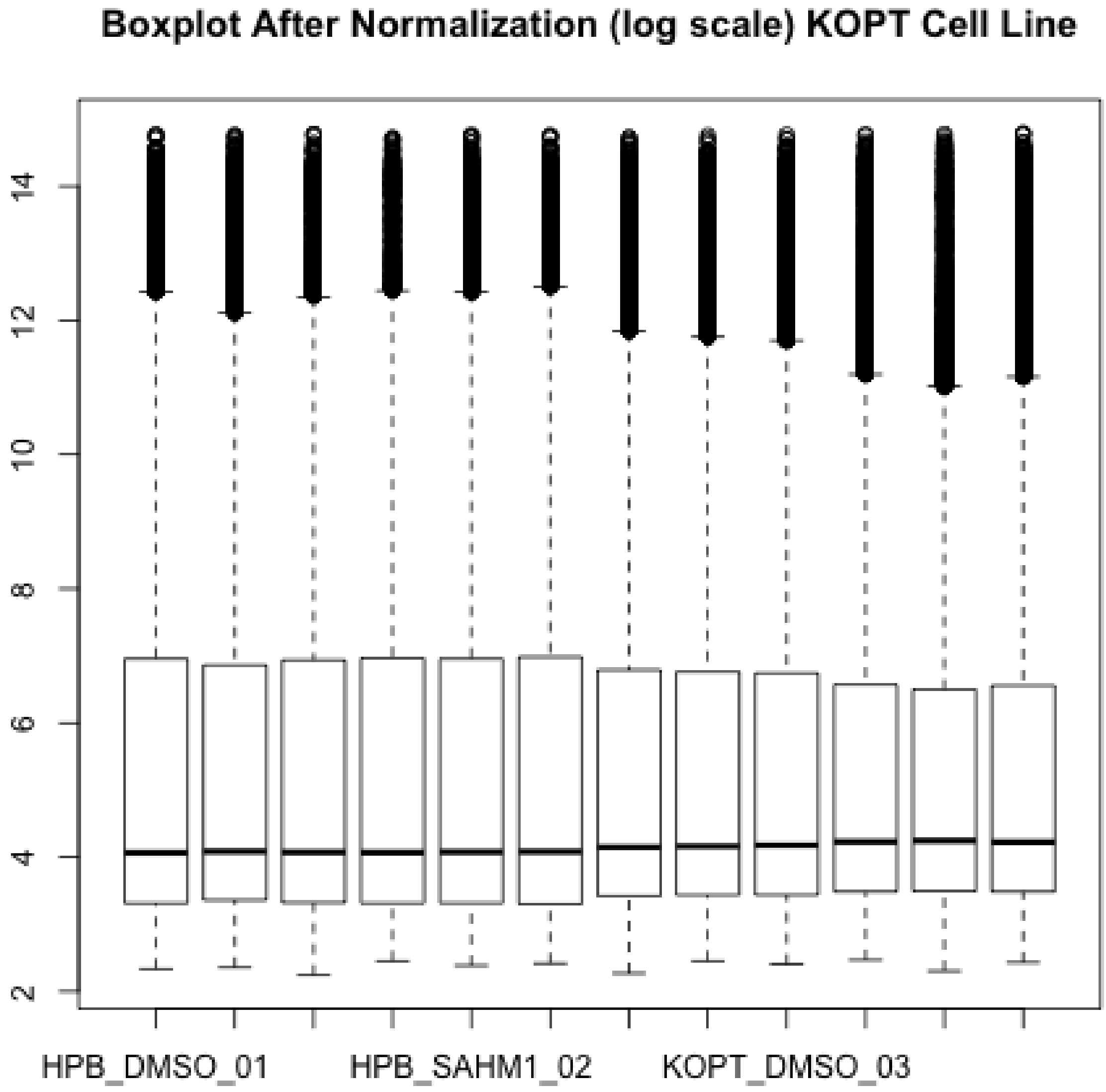
Script 1.8.1 (R)

```r
%%R
boxplots(affyBatch_HPB, eset, "HPB Cell Line")
boxplots(affyBatch_KOPT, eset, "KOPT Cell Line")
```


Boxplot Before Normalization HPB Cell Line
6
8
10
12
14
HPB_DMSO_01
HPB_DMSO_03
HPB_SAHM1_02


Boxplot After Normalization (log scale) HPB Cell Line
2
4
6
8
10
12
14
HPB_DMSO_01
HPB_SAHM1_02
KOPT_DMSO_03


Boxplot Before Normalization KOPT Cell Line
14
12
10
8
6
KOPT_DMSO_01
KOPT_DMSO_03
KOPT_SAHM1_02



1.9 Differential expressed genes

Script 1.9.1 (R)

```
%%R
toptable_HPB <- create_TopTable(eset_HPB)
toptable_anot_HPB <- anotate_TopTable(toptable_HPB)
generank_table(toptable_anot_HPB, "generank_HPB")
```

```
[1] "Down-regulated genes"
               Symbol      logFC    adj.P.Val
225342_at         AK4 -2.1078858 1.128982e-05
230710_at     MIR210HG -1.9761492 7.688286e-06
227336_at        DTX1 -1.3841443 8.636324e-05
201842_s_at    EFEMP1 -1.3678548 6.004820e-05
204348_s_at       AK4 -1.3491612 3.350727e-04
227347_x_at      HES4 -1.2640712 1.228528e-04
200953_s_at     CCND2 -1.2357909 7.997079e-05
202464_s_at    PFKFB3 -1.1937490 1.228528e-04
202022_at       ALDOC -1.1753644 1.248092e-04
240546_at   LINC01120 -1.1131135 1.552628e-04
227337_at     ANKRD37 -1.1079168 2.110727e-04
200894_s_at     FKBP4 -1.0737202 5.300330e-04
217078_s_at    CD300A -1.0719092 8.239876e-04
201170_s_at   BHLHE40 -1.0402130 2.549012e-04
202934_at         HK2 -1.0377714 2.104071e-04
201848_s_at     BNIP3 -1.0052154 6.744018e-04
218051_s_at    NT5DC2 -1.0049990 2.851346e-03
203394_s_at      HES1 -0.9993973 7.294799e-04
219371_s_at      KLF2 -0.9800918 3.918155e-04
201251_at         PKM -0.9787897 4.328581e-03
201849_at       BNIP3 -0.9724936 6.167814e-04
213746_s_at      FLNA -0.9705676 1.430899e-03
201516_at         SRM -0.9428084 1.627743e-03
203523_at        LSP1 -0.9360851 2.263449e-03
225944_at         NLN -0.9317256 1.421369e-03
214183_s_at     TKTL1 -0.9290362 2.452591e-03
236180_at          NA -0.9287828 3.049302e-03
201194_at     SELENOW -0.9285463 6.167814e-04
231922_at      ZNF276 -0.9269927 4.098501e-03
209933_s_at    CD300A -0.9049911 8.239876e-04
214752_x_at      FLNA -0.9018607 1.636952e-03
226348_at       FUT11 -0.8976507 1.075839e-03
201212_at        LGMN -0.8972886 1.381530e-03
218305_at        IPO4 -0.8783786 1.923264e-03
205544_s_at       CR2 -0.8502424 1.206402e-03
202145_at        LY6E -0.8406636 2.534145e-03
200859_x_at      FLNA -0.8368746 2.851346e-03
202887_s_at     DDIT4 -0.8271607 1.813467e-03
200965_s_at    ABLIM1 -0.8229490 1.522219e-03
203504_s_at     ABCA1 -0.8211483 1.430899e-03
208116_s_at    MAN1A1 -0.7931451 4.037398e-03
202472_at         MPI -0.7831169 2.929531e-03
207543_s_at     P4HA1 -0.7783781 2.263449e-03
201563_at        SORD -0.7663457 4.281959e-03
222150_s_at      GSAP -0.7532244 4.098501e-03
207539_s_at       IL4 -0.7404704 4.098501e-03
```

```
205895_s_at     NOLC1 -0.7357033 4.328581e-03
219389_at       SUSD4 -0.7347418 3.852969e-03
201562_s_at      SORD -0.7343969 4.098501e-03
218984_at        PUS7 -0.7076878 4.281959e-03
[1] "Up-regulated genes"
                  Symbol     logFC    adj.P.Val
204962_s_at           NA 0.8016614 1.927646e-03
222670_s_at         MAFB 0.8314472 1.414249e-03
244075_at             NA 0.8555977 1.927646e-03
205047_s_at         ASNS 0.8566562 1.373571e-03
236153_at             NA 0.8615325 1.373571e-03
228999_at           CHD2 0.8640573 1.731300e-03
202847_at           PCK2 0.8839066 1.373571e-03
242388_x_at        TAGAP 0.8933876 1.116736e-03
243368_at             NA 0.9077303 1.731300e-03
1558212_at            NA 0.9381646 1.381530e-03
212907_at        SLC30A1 0.9572462 5.305252e-04
241505_at             NA 0.9630797 1.430899e-03
230659_at             NA 0.9742286 8.892420e-04
203279_at          EDEM1 0.9785129 3.944123e-04
218923_at           CTBS 0.9839137 1.116736e-03
1558920_at    SLC8A1-AS1 0.9839933 1.272470e-03
215071_s_at    HIST1H2AC 0.9867407 1.634677e-03
217678_at        SLC7A11 1.0002602 6.787587e-04
235795_at           PAX6 1.0066337 6.744018e-04
1556294_at         FXYD2 1.0219109 1.381530e-03
229538_s_at       IQGAP3 1.0315469 5.922793e-04
206864_s_at          HRK 1.0324569 1.969355e-03
243495_s_at       ZNF652 1.0513033 1.522219e-03
218145_at          TRIB3 1.0673820 2.149263e-04
219892_at         TM6SF1 1.0684116 7.475097e-04
244377_at         SLC1A4 1.0880295 2.042060e-04
201010_s_at        TXNIP 1.1062690 1.552628e-04
209921_at        SLC7A11 1.1333321 1.248092e-04
209822_s_at        VLDLR 1.1350680 2.104071e-04
230795_at             NA 1.1699268 2.104071e-04
213931_at             NA 1.1702639 3.804126e-04
201009_s_at        TXNIP 1.2328092 1.663743e-04
244042_x_at           NA 1.2501748 1.080676e-03
218559_s_at         MAFB 1.2553755 5.305252e-04
225957_at         CREBRF 1.2981952 4.382754e-04
222853_at          FLRT3 1.3254635 1.138577e-04
219270_at          CHAC1 1.3362724 3.651273e-05
202672_s_at         ATF3 1.3481402 3.651273e-05
218280_x_at           NA 1.3495011 1.272470e-03
207076_s_at         ASS1 1.4918960 3.651273e-05
201008_s_at        TXNIP 1.4920023 1.663743e-04
243871_at   LOC100130476 1.5055489 4.060560e-04
```

```
201464_x_at          JUN 1.5541848 7.688286e-06
236962_at             NA 1.5966520 1.522219e-03
235412_at        ARHGEF7 1.5978442 2.999663e-04
229541_at             NA 1.6143542 2.042060e-04
229147_at         RASSF6 1.6458323 7.688286e-06
235638_at         RASSF6 1.7804932 7.997079e-05
201466_s_at          JUN 2.1333931 2.108830e-06
201465_s_at          JUN 2.2743608 1.053822e-05
```

Script 1.9.2 (R)

```
%%R
toptable_KOPT <- create_TopTable(eset_KOPT)
toptable_anot_KOPT <- anotate_TopTable(toptable_KOPT)
generank_table(toptable_anot_KOPT, "generank_KOPT")
```

```
[1] "Down-regulated genes"
               Symbol     logFC    adj.P.Val
209921_at     SLC7A11 -2.916468 2.066610e-11
205047_s_at      ASNS -2.757747 2.764706e-11
209369_at       ANXA3 -2.661459 2.066610e-11
219270_at       CHAC1 -2.264373 7.569090e-10
226517_at       BCAT1 -2.216157 2.899776e-10
214452_at       BCAT1 -2.166267 3.402665e-09
225285_at       BCAT1 -2.072317 2.630231e-10
217678_at     SLC7A11 -2.058710 7.569090e-10
230748_at     SLC16A6 -2.004791 6.729951e-09
219892_at      TM6SF1 -1.957664 1.012031e-09
220892_s_at     PSAT1 -1.938100 4.136414e-10
204351_at       S100P -1.902944 5.924315e-09
223195_s_at     SESN2 -1.889417 7.569090e-10
214079_at       DHRS2 -1.873374 5.924315e-09
209822_s_at     VLDLR -1.863734 4.214321e-09
212290_at      SLC7A1 -1.860735 2.786337e-08
202847_at        PCK2 -1.828534 1.164949e-09
225520_at          NA -1.790127 1.147212e-09
223062_s_at     PSAT1 -1.786577 1.164949e-09
223196_s_at     SESN2 -1.730910 3.463629e-08
200924_s_at    SLC3A2 -1.698282 2.186020e-08
1553972_a_at      CBS -1.682036 3.595785e-09
207076_s_at      ASS1 -1.675254 2.113541e-08
229787_s_at       OGT -1.662192 1.743772e-07
222632_s_at    LZTFL1 -1.643945 9.106004e-08
212816_s_at       CBS -1.599989 1.068633e-07
224839_s_at      GPT2 -1.553796 1.734549e-08
```

```
240983_s_at      CARS -1.552793 5.615021e-08
215411_s_at  TRAF3IP2 -1.540251 2.113541e-08
221539_at    EIF4EBP1 -1.533442 1.149078e-08
201195_s_at    SLC7A5 -1.511384 1.336464e-08
214390_s_at     BCAT1 -1.459915 2.424870e-07
204999_s_at      ATF5 -1.451634 2.113541e-08
210512_s_at     VEGFA -1.448871 5.654616e-08
200878_at       EPAS1 -1.435966 3.072051e-07
212501_at       CEBPB -1.428290 2.844253e-08
224580_at     SLC38A1 -1.428164 1.068633e-07
204744_s_at      IARS -1.419418 2.844253e-08
208693_s_at      GARS -1.406661 4.090508e-08
223059_s_at   FAM107B -1.392809 5.070038e-08
218437_s_at    LZTFL1 -1.332751 1.769993e-07
1558212_at         NA -1.308392 1.982077e-07
205653_at        CTSG -1.279331 2.502696e-07
217078_s_at    CD300A -1.278830 2.424870e-07
203627_at       IGF1R -1.275358 3.751987e-07
221933_at      NLGN4X -1.272235 4.113388e-07
231894_at        SARS -1.263321 2.424870e-07
214095_at       SHMT2 -1.235111 1.982077e-07
201263_at        TARS -1.197960 3.751987e-07
226181_at       TUBE1 -1.189891 3.344154e-07
[1] "Up-regulated genes"
                           Symbol     logFC    adj.P.Val
224429_x_at                    NA 0.9259213 6.376426e-06
220725_x_at                 DNAH3 0.9423323 6.447180e-06
210686_x_at              SLC25A16 0.9631964 4.491576e-06
213605_s_at                    NA 0.9632616 6.833714e-06
1556206_at              LINC00408 0.9653901 4.353339e-06
244114_x_at                    NA 0.9679377 6.750428e-06
241632_x_at                    NA 0.9686426 4.885508e-06
239017_at                      NA 0.9834736 7.316357e-06
1558496_at              LINC02053 0.9853247 5.091816e-06
211585_at                    NPAT 0.9926606 6.429684e-06
236389_x_at                    NA 1.0027588 4.491576e-06
208120_x_at                FKSG49 1.0184438 2.815681e-06
206323_x_at                 OPHN1 1.0212889 5.601191e-06
224284_x_at                FKSG49 1.0251308 3.950099e-06
201464_x_at                   JUN 1.0348832 3.490668e-06
220828_s_at             LINC01949 1.0397500 4.491576e-06
81737_at             LOC100505915 1.0462560 5.662707e-06
210800_at                  TIMM8A 1.0479516 4.730744e-06
215182_x_at                    NA 1.0559763 4.491576e-06
243489_at                      NA 1.0626878 4.070717e-06
240988_x_at                    NA 1.0629756 4.779299e-06
224288_x_at                FKSG49 1.0813463 1.113017e-06
AFFX-r2-Ec-bioB-5_at           NA 1.1020958 5.322529e-06
```

```
242862_x_at                         NA 1.1113007 1.113017e-06
1563674_at                       FCRL2 1.1127799 6.945728e-06
AFFX-BioC-5_at                      NA 1.1199525 6.430788e-06
210718_s_at                         NA 1.1396478 6.670158e-06
1562755_at                          NA 1.1425399 2.235176e-06
232964_at                           NA 1.1460961 1.656041e-06
220232_at                         SCD5 1.1545960 1.183928e-06
1569940_at                     SLC6A16 1.1781696 8.464513e-07
211454_x_at                     FKSG49 1.1800267 5.834686e-07
AFFX-BioB-M_at                      NA 1.1849198 6.376426e-06
209700_x_at                    PDE4DIP 1.1871538 8.406263e-07
1566145_s_at                        NA 1.1918754 3.003518e-07
234949_at                       FRG1BP 1.1922760 1.231065e-06
1560144_at                          NA 1.2077566 1.790306e-06
1553185_at                       RASEF 1.2129105 6.160739e-07
231597_x_at                         NA 1.2169048 1.477715e-06
242619_x_at                         NA 1.2257718 4.302148e-07
1553186_x_at                     RASEF 1.2330332 6.881976e-07
227952_at                       ZNF595 1.2408024 5.034675e-07
1561754_at                          NA 1.2457988 9.800827e-07
224159_x_at                      TRIM4 1.2471076 7.904835e-07
243689_s_at                     FRG1BP 1.3748167 6.160739e-07
231598_x_at                         NA 1.4418759 1.469974e-07
228919_at                           NA 1.4589605 1.682330e-06
1562527_at                   LOC441666 1.4936054 7.062932e-08
1558048_x_at                        NA 1.5005510 2.186020e-08
211565_at                       SH3GL3 1.5671795 2.844253e-08
```

1.10 Generate GSEA gct, cls files

To send raw data, we need to process the expression data from affiBatch:

```
exprs2_HPB <- cbind(featureNames(affyBatch_HPB), c(" "), exprs(affyBatch_HPB))
write.table(exprs2_HPB,row.names=FALSE,col.names=FALSE,quote=FALSE,file="eset_HPB.tsv", sep = "\t")
exprs2_KOPT <- cbind(featureNames(affyBatch_KOPT), c(" "), exprs(affyBatch_KOPT))
write.table(exprs2_KOPT,row.names=FALSE,col.names=FALSE,quote=FALSE,file="eset_KOPT.tsv", sep = "\t")
```

We decide to compute from expression set generated by expresso, in order to consume less processing time in GSEA. But we need to transform the expression amount to no log quantities.

Script 1.10.1 (R)

```
%%R
exprs2_HPB <- cbind(c(" "), 2 ** exprs(eset_HPB))
write.table(exprs2_HPB,row.names=TRUE,col.names=FALSE,quote=FALSE,file="eset_HPB.tsv", sep =
↪  "\t")
exprs2_KOPT <- cbind(c(" "), 2 ** exprs(eset_KOPT))
```

```
write.table(exprs2_KOPT,row.names=TRUE,col.names=FALSE,quote=FALSE,file="eset_KOPT.tsv", sep = "\t")
```

Script 1.10.2 (bash)

```
%%bash
echo "#1.2" > gct.head.HPB
echo "$(cat eset_HPB.tsv | wc -l) 6" >> gct.head.HPB
echo "GID         NAME        HPB_DMSO_01         HPB_DMSO_02         HPB_DMSO_03         HPB_SAHM1_01        HPB_SAHM1_02        HPB_SAHM1_03" >> gct.head.HPB

echo "#1.2" > gct.head.KOPT
echo "$(cat eset_KOPT.tsv | wc -l) 6" >> gct.head.KOPT
echo "GID         NAME        KOPT_DMSO_01        KOPT_DMSO_02        KOPT_DMSO_03        KOPT_SAHM1_01       KOPT_SAHM1_02       KOPT_SAHM1_03" >> gct.head.KOPT

cat gct.head.HPB eset_HPB.tsv > eset_HPB.gct
cat gct.head.KOPT eset_KOPT.tsv > eset_KOPT.gct

echo "6        2        1" > phenotypes.cls
echo "#DMSO SAHM1" >> phenotypes.cls
echo "0        0        0        1        1        1" >> phenotypes.cls
```

1.11 Processing all samples

Script 1.11.1 (R)

```
%%R
setwd("GSE18198_data")
affyBatch = import_CEL("*")
setwd(wd)
eset <- create_eset(affyBatch)
toptable <- create_TopTable(eset, control_samples=c(1,1,1,0,0,0,1,1,1,0,0,0))
toptable_anot <- anotate_TopTable(toptable)
generank_table(toptable_anot, "generank")
save(eset, file="eset.RData")
```

```
background correction: rma
normalization: quantiles
PM/MM correction : pmonly
expression values: medianpolish
background correcting...done.
normalizing...done.
54675 ids to be processed
|                    |
|####################|
[1] "Down-regulated genes"
               Symbol      logFC      P.Value
227347_x_at      HES4 -1.2391846 4.541483e-05
227336_at        DTX1 -1.0787341 1.772044e-03
230263_s_at     DOCK5 -1.0589452 2.174812e-04
218051_s_at    NT5DC2 -0.9507155 1.126824e-03
205544_s_at       CR2 -0.9438572 5.961088e-05
202464_s_at    PFKFB3 -0.9408542 1.077045e-04
226452_at        PDK1 -0.8325181 2.019305e-05
223364_s_at     DHX37 -0.8207408 1.605856e-03
206686_at        PDK1 -0.8172945 9.936732e-04
203627_at       IGF1R -0.8050123 1.277490e-03
203867_s_at      NLE1 -0.7963720 7.746991e-04
204513_s_at     ELMO1 -0.7885410 1.839908e-04
207543_s_at     P4HA1 -0.7663495 3.287902e-05
212063_at        CD44 -0.7621344 1.769410e-03
227337_at     ANKRD37 -0.7620441 3.022164e-04
239410_at         HK2 -0.7615742 2.557385e-03
1554918_a_at    ABCC4 -0.7263246 1.912084e-03
231094_s_at        NA -0.7232606 9.131629e-04
200965_s_at    ABLIM1 -0.7085777 6.009393e-04
210625_s_at     AKAP1 -0.6987526 1.183409e-04
231310_at      TRIM71 -0.6679915 6.024183e-04
219253_at    TMEM185B -0.6494920 1.313180e-03
215195_at       PRKCA -0.6258312 2.255091e-04
228205_at         TKT -0.6126859 8.915350e-04
218806_s_at      VAV3 -0.5936019 2.050231e-03
206923_at       PRKCA -0.5924819 2.694018e-04
201367_s_at   ZFP36L2 -0.5879670 1.331411e-03
236180_at          NA -0.5535885 1.051495e-03
221989_at          NA -0.5531493 1.744579e-04
223058_at     FAM107B -0.5499447 2.198129e-03
208858_s_at     ESYT1 -0.5411513 2.844208e-04
1555434_a_at SLC39A14 -0.5382917 2.440482e-03
1553138_a_at   ANKLE1 -0.5306391 5.104033e-04
227099_s_at  C11orf96 -0.5289498 2.627354e-03
203612_at        BYSL -0.5257364 1.818973e-04
226938_at       DCAF4 -0.5161954 1.874088e-03
214484_s_at   SIGMAR1 -0.5153579 1.465927e-03
206653_at      POLR3G -0.5067532 2.177288e-03
226498_at        FLT1 -0.5041981 2.319967e-03
208758_at        ATIC -0.4943349 1.053895e-03
208997_s_at      UCP2 -0.4905291 2.479831e-03
209461_x_at     WDR18 -0.4862286 8.241993e-04
201161_s_at      YBX3 -0.4839871 9.730003e-04
225883_at     ATG16L2 -0.4745402 2.113101e-03
201692_at     SIGMAR1 -0.4727785 1.380736e-03
217139_at       VDAC1 -0.4553310 2.588750e-03
```

```
229236_s_at     SFXN4 -0.4251656 1.858039e-03
224824_at       COX20 -0.4173264 2.375617e-03
204027_s_at    METTL1 -0.4116541 2.046282e-03
201250_s_at    SLC2A1 -0.3947427 2.225655e-03
[1] "Up-regulated genes"
                   Symbol     logFC      P.Value
232059_at         DSCAML1 0.3346000 0.0068497056
209392_at           ENPP2 0.3540105 0.0052561029
205381_at          LRRC17 0.3645748 0.0064690771
214710_s_at         CCNB1 0.3671989 0.0073863284
1569680_at             NA 0.3814353 0.0033810217
1559023_a_at      EFCAB14 0.3845018 0.0052968042
236353_at              NA 0.3952903 0.0082999985
206448_at          ZNF365 0.4055256 0.0032538740
226936_at           CENPW 0.4121128 0.0055882829
243469_at              NA 0.4133233 0.0048561664
201896_s_at         PSRC1 0.4219794 0.0084642203
1568596_a_at        TROAP 0.4261581 0.0017179111
238875_at              NA 0.4288226 0.0078367183
242966_x_at          RFX2 0.4321379 0.0027616154
236253_at          ZNF546 0.4657924 0.0075436998
1557290_at             NA 0.4742477 0.0041564711
243992_at              NA 0.4761079 0.0040639088
204641_at            NEK2 0.4770373 0.0034259514
220167_s_at            NA 0.4845848 0.0045882849
241685_x_at          PURA 0.4879466 0.0058069390
239735_at              NA 0.4897705 0.0030891034
202644_s_at       TNFAIP3 0.4934166 0.0004891236
242476_at              NA 0.5030038 0.0056605223
238595_at              NA 0.5058669 0.0059647137
213605_s_at            NA 0.5068902 0.0050211914
244427_at           KIF23 0.5209074 0.0015710943
242637_at              NA 0.5222779 0.0047282688
232953_at     LINC00266-1 0.5259613 0.0041673250
1559156_at             NA 0.5325258 0.0057922093
228390_at           RAB30 0.5332846 0.0071758418
238407_at              NA 0.5390014 0.0066182068
216756_at              NA 0.5406353 0.0019487381
239248_at      SDCBP2-AS1 0.5441412 0.0010999281
213544_at            ING2 0.5513826 0.0046805162
239531_at              NA 0.5536821 0.0026133238
243709_at         SLC38A9 0.5569113 0.0009694726
241745_at    LOC100507557 0.5600510 0.0027569686
1557813_at             NA 0.5691647 0.0029037060
215599_at              NA 0.5931984 0.0012579337
244114_x_at            NA 0.6411913 0.0034400322
213281_at             JUN 0.6419108 0.0061333797
228834_at            TOB1 0.6427144 0.0006729806
```

```
244532_x_at             NA 0.6470301 0.0046383310
230795_at               NA 0.6509000 0.0039752814
216094_at               NA 0.6647132 0.0040965858
234759_at     LOC100287497 0.6758224 0.0008985553
244075_at               NA 0.6914067 0.0071383024
215071_s_at      HIST1H2AC 0.7357238 0.0038095135
210718_s_at             NA 0.7879121 0.0013751119
201465_s_at            JUN 1.5137055 0.0061879051
```

Script 1.11.2 (R)

```r
%%R
# Create files for GSEA
exprs2 <- cbind(c(" "), 2 ** exprs(eset))
write.table(exprs2, row.names=TRUE,col.names=FALSE, quote=FALSE,file="eset.tsv", sep = "\t")
```

Script 1.11.3 (bash)

```bash
%%bash
echo "#1.2" > gct.head
echo "$(cat eset.tsv | wc -l) 12" >> gct.head
echo "GID        NAME        HPB_DMSO_01        HPB_DMSO_02        HPB_DMSO_03        HPB_SAHM1_01        HPB_SAHM1_02        HPB_SAHM1_03        KOPT_DMSO_01        KOPT_DMSO_02        KOPT_DMSO_03        KOPT_SAHM1_01        KOPT_SAHM1_02        KOPT_SAHM1_03" >> gct.head
#head gct.head
cat gct.head eset.tsv > eset.gct

echo "12        2        1" > phenotypes_all.cls
echo "#DMSO SAHM1" >> phenotypes_all.cls
echo "0        0        0        1        1        1        0        0        0        1        1        1" >> phenotypes_all.cls
```

1.12 GSEA results

See figures 1 to 8.

1.13 Methods

Pipeline

We have created the working environment under an i-python notebook of the *jupyter* platform configured to be able to execute R in code cells that start with *%%R*. To do so we have used the python package *rpy2*. This allows us to keep the documentation unified with the execution pipeline. It also becomes a good environment to launch hybrid pipelines with steps in R, python or even bash. It would not be difficult to develop on top a checkpoint and restart system for those developments highly time consuming.

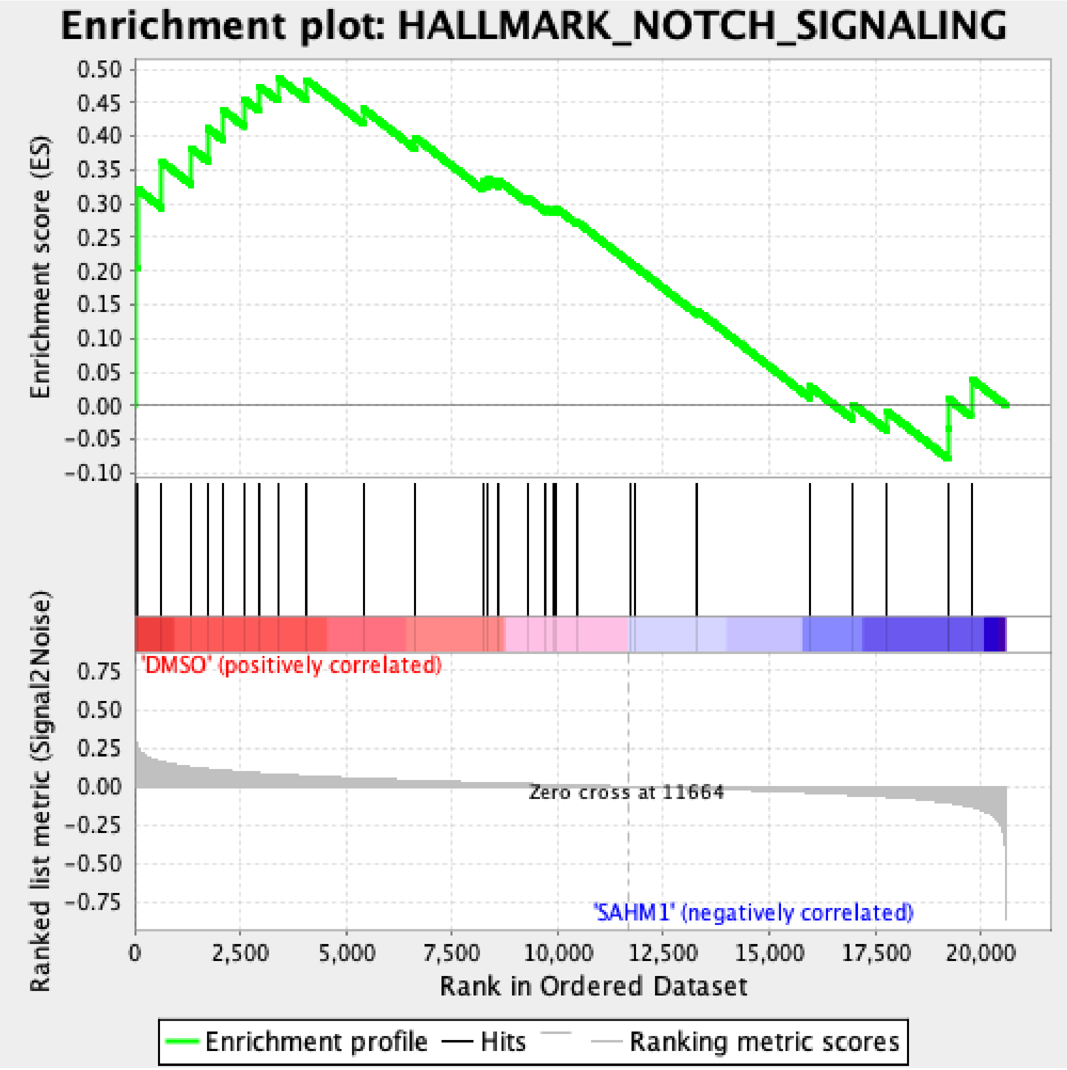
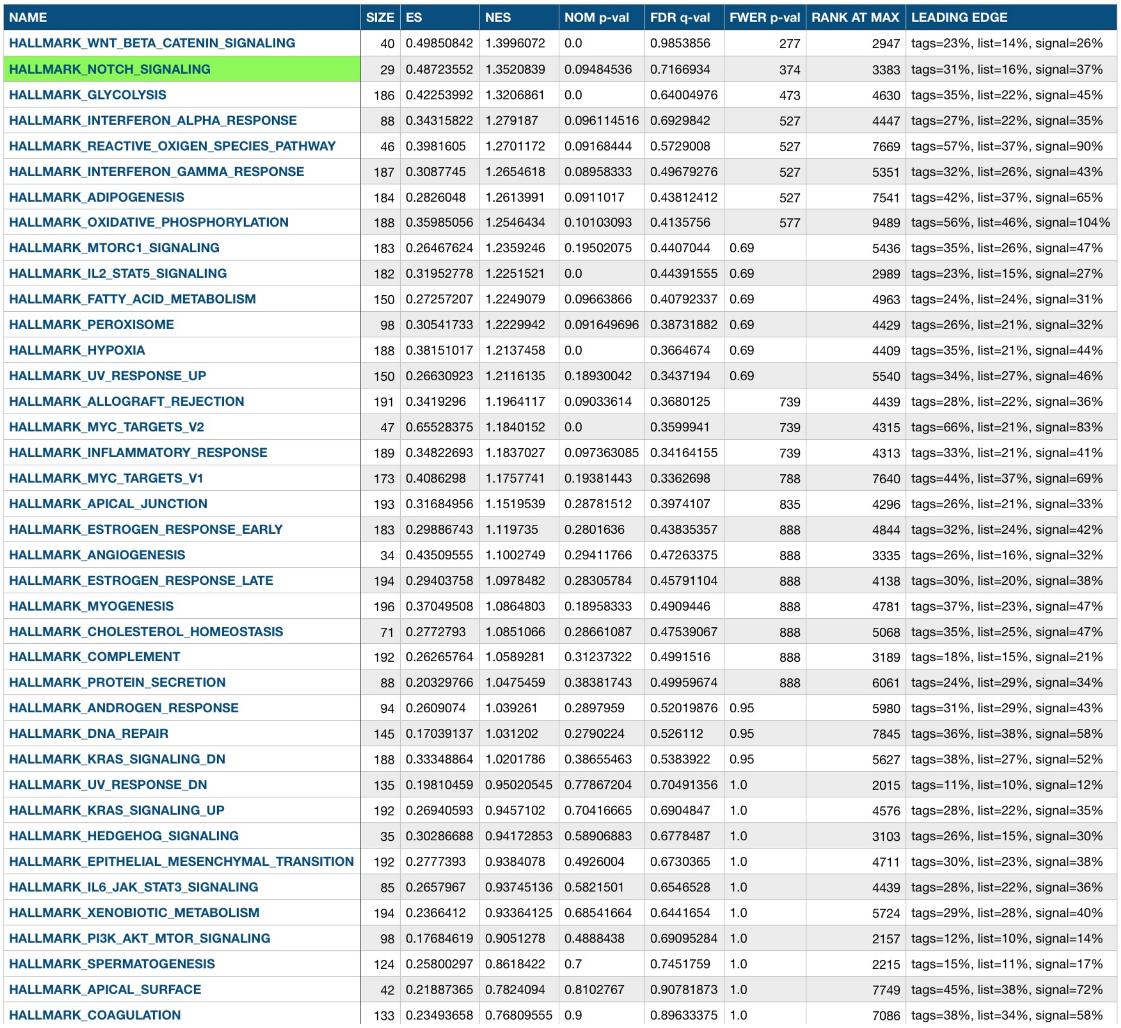


Figure 1: Enrichment plot Notch Signalling Pathway



NAME	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX	LEADING EDGE
HALLMARK_WNT_BETA_CATENIN_SIGNALING	40	0.49850842	1.3996072	0.0	0.9853856	277	2947	tags=23%, list=14%, signal=26%
HALLMARK_NOTCH_SIGNALING	29	0.48723552	1.3520839	0.09484536	0.7166934	374	3383	tags=31%, list=16%, signal=37%
HALLMARK_GLYCOLYSIS	186	0.42253992	1.3206861	0.0	0.64004976	473	4630	tags=35%, list=22%, signal=45%
HALLMARK_INTERFERON_ALPHA_RESPONSE	88	0.34315822	1.279187	0.096114516	0.6929842	527	4447	tags=27%, list=22%, signal=35%
HALLMARK_REACTIVE_OXIGEN_SPECIES_PATHWAY	46	0.3981605	1.2701172	0.09168444	0.5729008	527	7669	tags=57%, list=37%, signal=90%
HALLMARK_INTERFERON_GAMMA_RESPONSE	187	0.3087745	1.2654618	0.08958333	0.49679276	527	5351	tags=32%, list=26%, signal=43%
HALLMARK_ADIPOGENESIS	184	0.2826048	1.2613991	0.0911017	0.43812412	527	7541	tags=42%, list=37%, signal=65%
HALLMARK_OXIDATIVE_PHOSPHORYLATION	188	0.35985056	1.2546434	0.10103093	0.4135756	577	9489	tags=56%, list=46%, signal=104%
HALLMARK_MTORC1_SIGNALING	183	0.26467624	1.2359246	0.19502075	0.4407044	0.69	5436	tags=35%, list=26%, signal=47%
HALLMARK_IL2_STAT5_SIGNALING	182	0.31952778	1.2251521	0.0	0.44391555	0.69	2989	tags=23%, list=15%, signal=27%
HALLMARK_FATTY_ACID_METABOLISM	150	0.27257207	1.2249079	0.09663866	0.40792337	0.69	4963	tags=24%, list=24%, signal=31%
HALLMARK_PEROXISOME	98	0.30541733	1.2229942	0.091649696	0.38731882	0.69	4429	tags=26%, list=21%, signal=32%
HALLMARK_HYPOXIA	188	0.38151017	1.2137458	0.0	0.3664674	0.69	4409	tags=35%, list=21%, signal=44%
HALLMARK_UV_RESPONSE_UP	150	0.26630923	1.2116135	0.18930042	0.3437194	0.69	5540	tags=34%, list=27%, signal=46%
HALLMARK_ALLOGRAFT_REJECTION	191	0.3419296	1.1964117	0.09033614	0.3680125	739	4439	tags=28%, list=22%, signal=36%
HALLMARK_MYC_TARGETS_V2	47	0.65528375	1.1840152	0.0	0.3599941	739	4315	tags=66%, list=21%, signal=83%
HALLMARK_INFLAMMATORY_RESPONSE	189	0.34822693	1.1837027	0.097363085	0.34164155	739	4313	tags=33%, list=21%, signal=41%
HALLMARK_MYC_TARGETS_V1	173	0.4086298	1.1757741	0.19381443	0.3362698	788	7640	tags=44%, list=37%, signal=69%
HALLMARK_APICAL_JUNCTION	193	0.31684956	1.1519539	0.28781512	0.3974107	835	4296	tags=26%, list=21%, signal=33%
HALLMARK_ESTROGEN_RESPONSE_EARLY	183	0.29886743	1.119735	0.2801636	0.43835357	888	4844	tags=32%, list=24%, signal=42%
HALLMARK_ANGIOGENESIS	34	0.43509555	1.1002749	0.29411766	0.47263375	888	3335	tags=26%, list=16%, signal=32%
HALLMARK_ESTROGEN_RESPONSE_LATE	194	0.29403758	1.0978482	0.28305784	0.45791104	888	4138	tags=30%, list=20%, signal=38%
HALLMARK_MYOGENESIS	196	0.37049508	1.0864803	0.18958333	0.4909446	888	4781	tags=37%, list=23%, signal=47%
HALLMARK_CHOLESTEROL_HOMEOSTASIS	71	0.2772793	1.0851066	0.28661087	0.47539067	888	5068	tags=35%, list=25%, signal=47%
HALLMARK_COMPLEMENT	192	0.26265764	1.0589281	0.31237322	0.4991516	888	3189	tags=18%, list=15%, signal=21%
HALLMARK_PROTEIN_SECRETION	88	0.20329766	1.0475459	0.38381743	0.49959674	888	6061	tags=24%, list=29%, signal=34%
HALLMARK_ANDROGEN_RESPONSE	94	0.2609074	1.039261	0.2897959	0.52019876	0.95	5980	tags=31%, list=29%, signal=43%
HALLMARK_DNA_REPAIR	145	0.17039137	1.031202	0.2790224	0.526112	0.95	7845	tags=36%, list=38%, signal=58%
HALLMARK_KRAS_SIGNALING_DN	188	0.33348864	1.0201786	0.38655463	0.5383922	0.95	5627	tags=38%, list=27%, signal=52%
HALLMARK_UV_RESPONSE_DN	135	0.19810459	0.95020545	0.77867204	0.70491356	1.0	2015	tags=11%, list=10%, signal=12%
HALLMARK_KRAS_SIGNALING_UP	192	0.26940593	0.9457102	0.70416665	0.6904847	1.0	4576	tags=28%, list=22%, signal=35%
HALLMARK_HEDGEHOG_SIGNALING	35	0.30286688	0.94172853	0.58906883	0.6778487	1.0	3103	tags=26%, list=15%, signal=30%
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	192	0.2777393	0.9384078	0.4926004	0.6730365	1.0	4711	tags=30%, list=23%, signal=38%
HALLMARK_IL6_JAK_STAT3_SIGNALING	85	0.2657967	0.93745136	0.5821501	0.6546528	1.0	4439	tags=28%, list=22%, signal=36%
HALLMARK_XENOBIOTIC_METABOLISM	194	0.2366412	0.93364125	0.68541664	0.6441654	1.0	5724	tags=29%, list=28%, signal=40%
HALLMARK_PI3K_AKT_MTOR_SIGNALING	98	0.17684619	0.9051278	0.4888438	0.69095284	1.0	2157	tags=12%, list=10%, signal=14%
HALLMARK_SPERMATOGENESIS	124	0.25800297	0.8618422	0.7	0.7451759	1.0	2215	tags=15%, list=11%, signal=17%
HALLMARK_APICAL_SURFACE	42	0.21887365	0.7824094	0.8102767	0.90781873	1.0	7749	tags=45%, list=38%, signal=72%
HALLMARK_COAGULATION	133	0.23493658	0.76809555	0.9	0.89633375	1.0	7086	tags=38%, list=34%, signal=58%

Figure 2: Gene sets enriched in phenotype DMSO(Cell Line HPB-ALL)

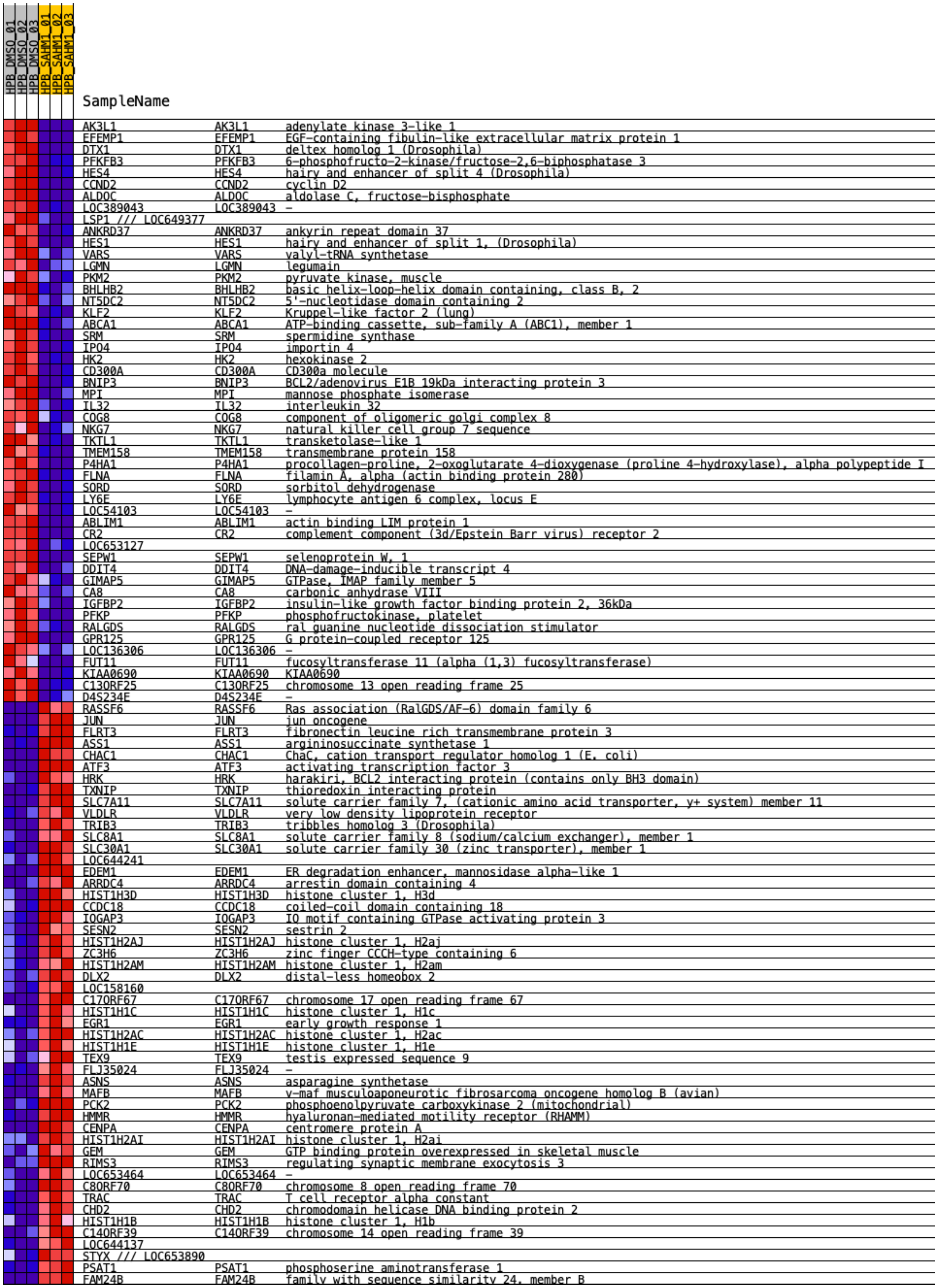
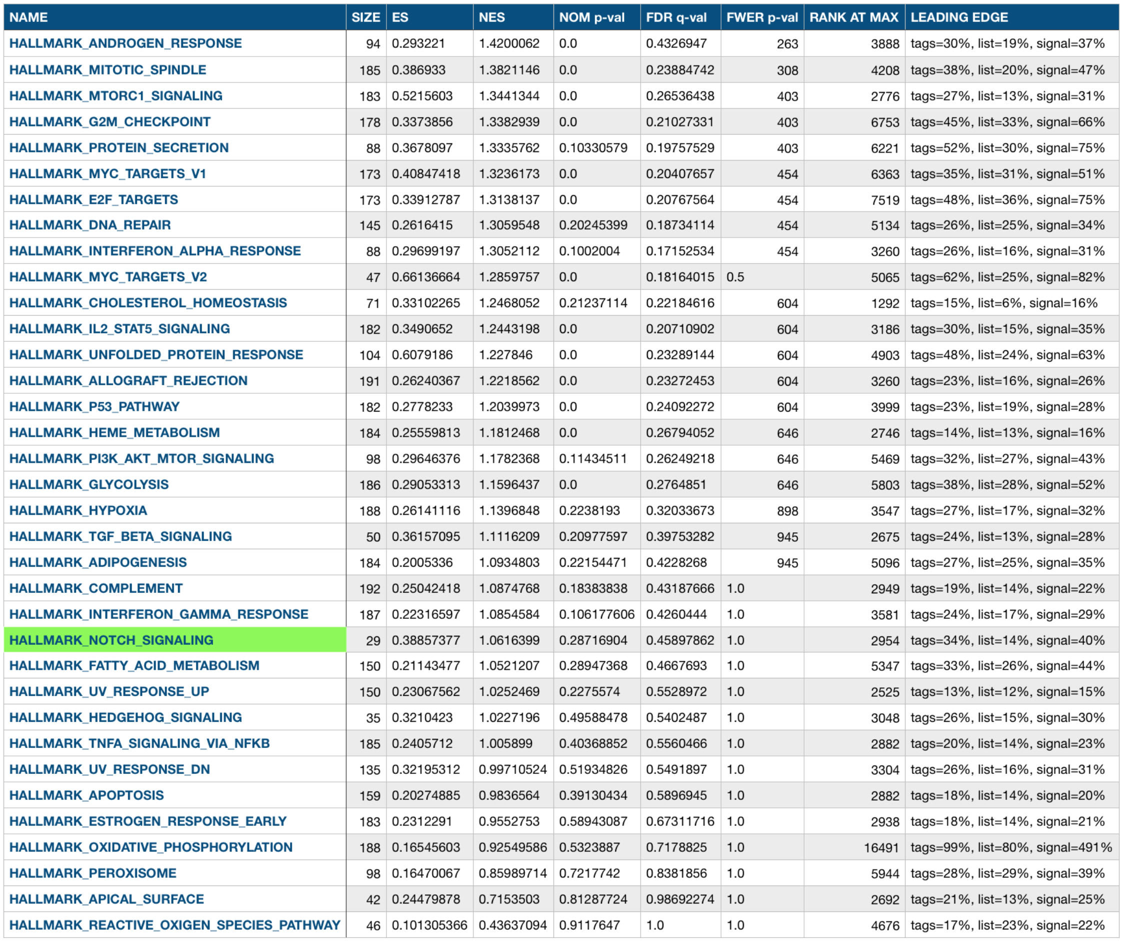
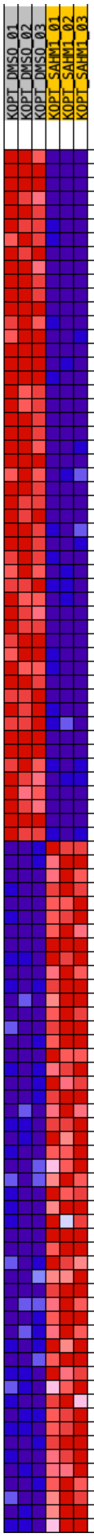


Figure 3: Heatmap(Cell Line HPB-ALL)



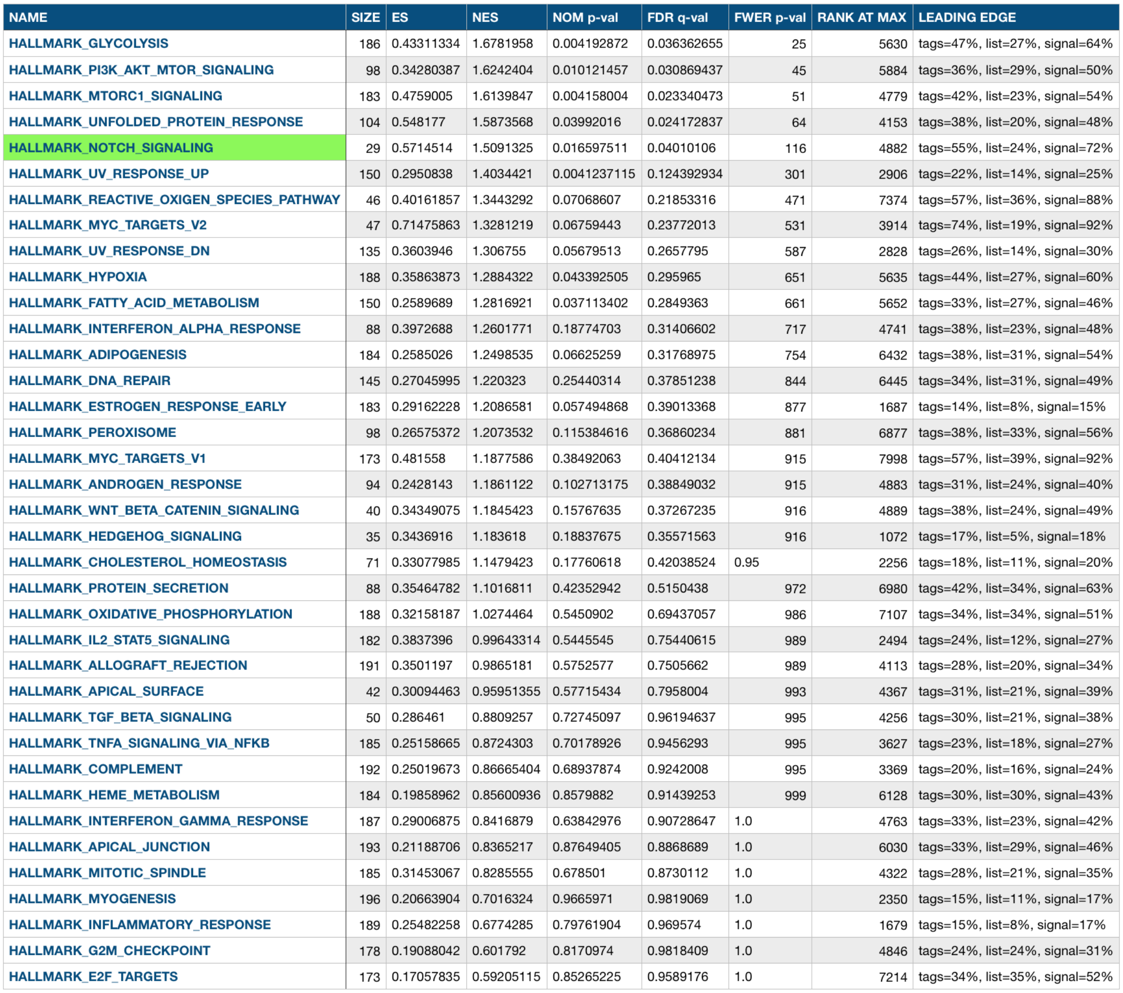
NAME	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX	LEADING EDGE
HALLMARK_ANDROGEN_RESPONSE	94	0.293221	1.4200062	0.0	0.4326947	263	3888	tags=30%, list=19%, signal=37%
HALLMARK_MITOTIC_SPINDLE	185	0.386933	1.3821146	0.0	0.23884742	308	4208	tags=38%, list=20%, signal=47%
HALLMARK_MTORC1_SIGNALING	183	0.5215603	1.3441344	0.0	0.26536438	403	2776	tags=27%, list=13%, signal=31%
HALLMARK_G2M_CHECKPOINT	178	0.3373856	1.3382939	0.0	0.21027331	403	6753	tags=45%, list=33%, signal=66%
HALLMARK_PROTEIN_SECRETION	88	0.3678097	1.3335762	0.10330579	0.19757529	403	6221	tags=52%, list=30%, signal=75%
HALLMARK_MYC_TARGETS_V1	173	0.40847418	1.3236173	0.0	0.20407657	454	6363	tags=35%, list=31%, signal=51%
HALLMARK_E2F_TARGETS	173	0.33912787	1.3138137	0.0	0.20767564	454	7519	tags=48%, list=36%, signal=75%
HALLMARK_DNA_REPAIR	145	0.2616415	1.3059548	0.20245399	0.18734114	454	5134	tags=26%, list=25%, signal=34%
HALLMARK_INTERFERON_ALPHA_RESPONSE	88	0.29699197	1.3052112	0.1002004	0.17152534	454	3260	tags=26%, list=16%, signal=31%
HALLMARK_MYC_TARGETS_V2	47	0.66136664	1.2859757	0.0	0.18164015	0.5	5065	tags=62%, list=25%, signal=82%
HALLMARK_CHOLESTEROL_HOMEOSTASIS	71	0.33102265	1.2468052	0.21237114	0.22184616	604	1292	tags=15%, list=6%, signal=16%
HALLMARK_IL2_STAT5_SIGNALING	182	0.3490652	1.2443198	0.0	0.20710902	604	3186	tags=30%, list=15%, signal=35%
HALLMARK_UNFOLDED_PROTEIN_RESPONSE	104	0.6079186	1.227846	0.0	0.23289144	604	4903	tags=48%, list=24%, signal=63%
HALLMARK_ALLOGRAFT_REJECTION	191	0.26240367	1.2218562	0.0	0.23272453	604	3260	tags=23%, list=16%, signal=26%
HALLMARK_P53_PATHWAY	182	0.2778233	1.2039973	0.0	0.24092272	604	3999	tags=23%, list=19%, signal=28%
HALLMARK_HEME_METABOLISM	184	0.25559813	1.1812468	0.0	0.26794052	646	2746	tags=14%, list=13%, signal=16%
HALLMARK_PI3K_AKT_MTOR_SIGNALING	98	0.29646376	1.1782368	0.11434511	0.26249218	646	5469	tags=32%, list=27%, signal=43%
HALLMARK_GLYCOLYSIS	186	0.29053313	1.1596437	0.0	0.2764851	646	5803	tags=38%, list=28%, signal=52%
HALLMARK_HYPOXIA	188	0.26141116	1.1396848	0.2238193	0.32033673	898	3547	tags=27%, list=17%, signal=32%
HALLMARK_TGF_BETA_SIGNALING	50	0.36157095	1.1116209	0.20977597	0.39753282	945	2675	tags=24%, list=13%, signal=28%
HALLMARK_ADIPOGENESIS	184	0.2005336	1.0934803	0.22154471	0.4228268	945	5096	tags=27%, list=25%, signal=35%
HALLMARK_COMPLEMENT	192	0.25042418	1.0874768	0.18383838	0.43187666	1.0	2949	tags=19%, list=14%, signal=22%
HALLMARK_INTERFERON_GAMMA_RESPONSE	187	0.22316597	1.0854584	0.106177606	0.4260444	1.0	3581	tags=24%, list=17%, signal=29%
HALLMARK_NOTCH_SIGNALING	29	0.38857377	1.0616399	0.28716904	0.45897862	1.0	2954	tags=34%, list=14%, signal=40%
HALLMARK_FATTY_ACID_METABOLISM	150	0.21143477	1.0521207	0.28947368	0.4667693	1.0	5347	tags=33%, list=26%, signal=44%
HALLMARK_UV_RESPONSE_UP	150	0.23067562	1.0252469	0.2275574	0.5528972	1.0	2525	tags=13%, list=12%, signal=15%
HALLMARK_HEDGEHOG_SIGNALING	35	0.3210423	1.0227196	0.49588478	0.5402487	1.0	3048	tags=26%, list=15%, signal=30%
HALLMARK_TNFA_SIGNALING_VIA_NFKB	185	0.2405712	1.005899	0.40368852	0.5560466	1.0	2882	tags=20%, list=14%, signal=23%
HALLMARK_UV_RESPONSE_DN	135	0.32195312	0.99710524	0.51934826	0.5491897	1.0	3304	tags=26%, list=16%, signal=31%
HALLMARK_APOPTOSIS	159	0.20274885	0.9836564	0.39130434	0.5896945	1.0	2882	tags=18%, list=14%, signal=20%
HALLMARK_ESTROGEN_RESPONSE_EARLY	183	0.2312291	0.9552753	0.58943087	0.67311716	1.0	2938	tags=18%, list=14%, signal=21%
HALLMARK_OXIDATIVE_PHOSPHORYLATION	188	0.16545603	0.92549586	0.5323887	0.7178825	1.0	16491	tags=99%, list=80%, signal=491%
HALLMARK_PEROXISOME	98	0.16470067	0.85989714	0.7217742	0.8381856	1.0	5944	tags=28%, list=29%, signal=39%
HALLMARK_APICAL_SURFACE	42	0.24479878	0.7153503	0.81287724	0.98692274	1.0	2692	tags=21%, list=13%, signal=25%
HALLMARK_REACTIVE_OXIGEN_SPECIES_PATHWAY	46	0.101305366	0.43637094	0.9117647	1.0	1.0	4676	tags=17%, list=23%, signal=22%

Figure 4: Gene sets enriched in phenotype DMSO(Cell Line KOPT-K1)



SampleName		
SLC7A11	SLC7A11	solute carrier family 7, (cationic amino acid transporter, y+ system) member 11
ASNS	ASNS	asparagine synthetase
ANXA3	ANXA3	annexin A3
CHAC1	CHAC1	ChaC, cation transport regulator homolog 1 (E. coli)
BCAT1	BCAT1	branched chain aminotransferase 1, cytosolic
SLC16A6	SLC16A6	solute carrier family 16, member 6 (monocarboxylic acid transporter 7)
S100P	S100P	S100 calcium binding protein P
SESN2	SESN2	sestrin 2
VLDLR	VLDLR	very low density lipoprotein receptor
PCK2	PCK2	phosphoenolpyruvate carboxykinase 2 (mitochondrial)
MTHFD1L	MTHFD1L	methylenetetrahydrofolate dehydrogenase (NADP+ dependent) 1-like
PSAT1	PSAT1	phosphoserine aminotransferase 1
SLC3A2	SLC3A2	solute carrier family 3 (activators of dibasic and neutral amino acid transport), member 2
DHRS2	DHRS2	dehydrogenase/reductase (SDR family) member 2
CBS	CBS	cystathionine-beta-synthase
ASS1	ASS1	argininosuccinate synthetase 1
LZTFL1	LZTFL1	leucine zipper transcription factor-like 1
GPT2	GPT2	glutamic pyruvate transaminase (alanine aminotransferase) 2
TRAF3IP2	TRAF3IP2	TRAF3 interacting protein 2
EIF4EBP1	EIF4EBP1	eukaryotic translation initiation factor 4E binding protein 1
SLC7A5	SLC7A5	solute carrier family 7 (cationic amino acid transporter, y+ system), member 5
VEGF	VEGF	vascular endothelial growth factor
CEBPB	CEBPB	CCAAT/enhancer binding protein (C/EBP), beta
EPAS1	EPAS1	endothelial PAS domain protein 1
IARS	IARS	isoleucine-tRNA synthetase
GARS	GARS	glycyl-tRNA synthetase
FAM107B	FAM107B	family with sequence similarity 107, member B
CST7	CST7	cystatin F (leukocystatin)
FLJ35024	FLJ35024	-
CTSG	CTSG	cathepsin G
NLGN4X	NLGN4X	neuroligin 4, X-linked
IGF1R	IGF1R	insulin-like growth factor 1 receptor
RAG1	RAG1	recombination activating gene 1
HES4	HES4	hairy and enhancer of split 4 (Drosophila)
TARS	TARS	threonyl-tRNA synthetase
TUBE1	TUBE1	tubulin, epsilon 1
SLC4A5	SLC4A5	solute carrier family 4, sodium bicarbonate cotransporter, member 5
SLC1A5	SLC1A5	solute carrier family 1 (neutral amino acid transporter), member 5
SLC7A3	SLC7A3	solute carrier family 7 (cationic amino acid transporter, y+ system), member 3
MARS	MARS	methionine-tRNA synthetase
SLC7A1	SLC7A1	solute carrier family 7 (cationic amino acid transporter, y+ system), member 1
H1F0	H1F0	H1 histone family, member 0
PHGDH	PHGDH	phosphoglycerate dehydrogenase
YARS	YARS	tyrosyl-tRNA synthetase
MGC29671	MGC29671	-
ULBP1	ULBP1	UL16 binding protein 1
JDP2	JDP2	-
CR2	CR2	complement component (3d/Epstein Barr virus) receptor 2
CARS	CARS	cysteinyl-tRNA synthetase
KIAA1211	KIAA1211	-
SH3GL3	SH3GL3	SH3-domain GRB2-like 3
LOC283027	LOC283027	-
TRIM4	TRIM4	tripartite motif-containing 4
ZNF595	ZNF595	zinc finger protein 595
SLC36A1	SLC36A1	solute carrier family 36 (proton/amino acid symporter), member 1
RASEF	RASEF	RAS and EF-hand domain containing
MGC72104	MGC72104	-
LOC644450	LOC644450	-
SLC6A16	SLC6A16	solute carrier family 6, member 16
SCD5	SCD5	stearoyl-CoA desaturase 5
WBSCR19	WBSCR19	Williams Beuren syndrome chromosome region 19
FCRL2	FCRL2	Fc receptor-like 2
LOC643749	LOC643749	-
LOC643675	LOC643675	-
FLJ11292	FLJ11292	-
JUN	JUN	jun oncogene
OPHN1	OPHN1	oligophrenin 1
MEFV	MEFV	Mediterranean fever
FKSG49	FKSG49	-
LOC644488		
OIT3	OIT3	oncoprotein induced transcript 3
DSPP	DSPP	dentin sialophosphoprotein
LOC389634	LOC389634	-
TMEFF2	TMEFF2	transmembrane protein with EGF-like and two follistatin-like domains 2
CYP1A2	CYP1A2	cytochrome P450, family 1, subfamily A, polypeptide 2
LOC643373 /// LOC653		
KIAA1217	KIAA1217	KIAA1217
LOC653117	LOC653117	-
SLC25A16	SLC25A16	solute carrier family 25 (mitochondrial carrier; Graves disease autoantigen), member 16
DNAH3	DNAH3	dynein, axonemal, heavy chain 3
LOC401131	LOC401131	-
RPL37A	RPL37A	ribosomal protein L37a
MGC10997	MGC10997	-
ZNF528	ZNF528	zinc finger protein 528
FLJ45803	FLJ45803	-
PLSCR4	PLSCR4	phospholipid scramblase 4
RP11-151A6.2	RP11-151A6.2	-
PDE3B	PDE3B	phosphodiesterase 3B, cGMP-inhibited
FRRS1	FRRS1	ferric-chelate reductase 1
RP11-262H14.4	RP11-262H14.4	-
DEFB106A	DEFB106A	defensin, beta 106A
C21ORF114	C21ORF114	chromosome 21 open reading frame 114
LOC645352		
RECK	RECK	reversion-inducing-cysteine-rich protein with kazal motifs
HCG2P7	HCG2P7	HLA complex group 2 pseudogene 7
ALDH1B1	ALDH1B1	aldehyde dehydrogenase 1 family, member B1
PRO1268	PRO1268	-
NLN	NLN	neurolysin (metallopeptidase M3 family)
C1ORF181	C1ORF181	chromosome 1 open reading frame 181
PRICKLE1	PRICKLE1	prickle homolog 1 (Drosophila)

Figure 5: Heatmap(Cell Line KOPT-K1)



NAME	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX	LEADING EDGE
HALLMARK_GLYCOLYSIS	186	0.43311334	1.6781958	0.004192872	0.036362655	25	5630	tags=47%, list=27%, signal=64%
HALLMARK_PI3K_AKT_MTOR_SIGNALING	98	0.34280387	1.6242404	0.010121457	0.030869437	45	5884	tags=36%, list=29%, signal=50%
HALLMARK_MTORC1_SIGNALING	183	0.4759005	1.6139847	0.004158004	0.023340473	51	4779	tags=42%, list=23%, signal=54%
HALLMARK_UNFOLDED_PROTEIN_RESPONSE	104	0.548177	1.5873568	0.03992016	0.024172837	64	4153	tags=38%, list=20%, signal=48%
HALLMARK_NOTCH_SIGNALING	29	0.5714514	1.5091325	0.016597511	0.04010106	116	4882	tags=55%, list=24%, signal=72%
HALLMARK_UV_RESPONSE_UP	150	0.2950838	1.4034421	0.0041237115	0.124392934	301	2906	tags=22%, list=14%, signal=25%
HALLMARK_REACTIVE_OXIGEN_SPECIES_PATHWAY	46	0.40161857	1.3443292	0.07068607	0.21853316	471	7374	tags=57%, list=36%, signal=88%
HALLMARK_MYC_TARGETS_V2	47	0.71475863	1.3281219	0.06759443	0.23772013	531	3914	tags=74%, list=19%, signal=92%
HALLMARK_UV_RESPONSE_DN	135	0.3603946	1.306755	0.05679513	0.2657795	587	2828	tags=26%, list=14%, signal=30%
HALLMARK_HYPOXIA	188	0.35863873	1.2884322	0.043392505	0.295965	651	5635	tags=44%, list=27%, signal=60%
HALLMARK_FATTY_ACID_METABOLISM	150	0.2589689	1.2816921	0.037113402	0.2849363	661	5652	tags=33%, list=27%, signal=46%
HALLMARK_INTERFERON_ALPHA_RESPONSE	88	0.3972688	1.2601771	0.18774703	0.31406602	717	4741	tags=38%, list=23%, signal=48%
HALLMARK_ADIPOGENESIS	184	0.2585026	1.2498535	0.06625259	0.31768975	754	6432	tags=38%, list=31%, signal=54%
HALLMARK_DNA_REPAIR	145	0.27045995	1.220323	0.25440314	0.37851238	844	6445	tags=34%, list=31%, signal=49%
HALLMARK_ESTROGEN_RESPONSE_EARLY	183	0.29162228	1.2086581	0.057494868	0.39013368	877	1687	tags=14%, list=8%, signal=15%
HALLMARK_PEROXISOME	98	0.26575372	1.2073532	0.115384616	0.36860234	881	6877	tags=38%, list=33%, signal=56%
HALLMARK_MYC_TARGETS_V1	173	0.481558	1.1877586	0.38492063	0.40412134	915	7998	tags=57%, list=39%, signal=92%
HALLMARK_ANDROGEN_RESPONSE	94	0.2428143	1.1861122	0.102713175	0.38849032	915	4883	tags=31%, list=24%, signal=40%
HALLMARK_WNT_BETA_CATENIN_SIGNALING	40	0.34349075	1.1845423	0.15767635	0.37267235	916	4889	tags=38%, list=24%, signal=49%
HALLMARK_HEDGEHOG_SIGNALING	35	0.3436916	1.183618	0.18837675	0.35571563	916	1072	tags=17%, list=5%, signal=18%
HALLMARK_CHOLESTEROL_HOMEOSTASIS	71	0.33077985	1.1479423	0.17760618	0.42038524	0.95	2256	tags=18%, list=11%, signal=20%
HALLMARK_PROTEIN_SECRETION	88	0.35464782	1.1016811	0.42352942	0.5150438	972	6980	tags=42%, list=34%, signal=63%
HALLMARK_OXIDATIVE_PHOSPHORYLATION	188	0.32158187	1.0274464	0.5450902	0.69437057	986	7107	tags=34%, list=34%, signal=51%
HALLMARK_IL2_STAT5_SIGNALING	182	0.3837396	0.99643314	0.5445545	0.75440615	989	2494	tags=24%, list=12%, signal=27%
HALLMARK_ALLOGRAFT_REJECTION	191	0.3501197	0.9865181	0.5752577	0.7505662	989	4113	tags=28%, list=20%, signal=34%
HALLMARK_APICAL_SURFACE	42	0.30094463	0.95951355	0.57715434	0.7958004	993	4367	tags=31%, list=21%, signal=39%
HALLMARK_TGF_BETA_SIGNALING	50	0.286461	0.8809257	0.72745097	0.96194637	995	4256	tags=30%, list=21%, signal=38%
HALLMARK_TNFA_SIGNALING_VIA_NFKB	185	0.25158665	0.8724303	0.70178926	0.9456293	995	3627	tags=23%, list=18%, signal=27%
HALLMARK_COMPLEMENT	192	0.25019673	0.86665404	0.68937874	0.9242008	995	3369	tags=20%, list=16%, signal=24%
HALLMARK_HEME_METABOLISM	184	0.19858962	0.85600936	0.8579882	0.91439253	999	6128	tags=30%, list=30%, signal=43%
HALLMARK_INTERFERON_GAMMA_RESPONSE	187	0.29006875	0.8416879	0.63842976	0.90728647	1.0	4763	tags=33%, list=23%, signal=42%
HALLMARK_APICAL_JUNCTION	193	0.21188706	0.8365217	0.87649405	0.8868689	1.0	6030	tags=33%, list=29%, signal=46%
HALLMARK_MITOTIC_SPINDLE	185	0.31453067	0.8285555	0.678501	0.8730112	1.0	4322	tags=28%, list=21%, signal=35%
HALLMARK_MYOGENESIS	196	0.20663904	0.7016324	0.9665971	0.9819069	1.0	2350	tags=15%, list=11%, signal=17%
HALLMARK_INFLAMMATORY_RESPONSE	189	0.25482258	0.6774285	0.79761904	0.969574	1.0	1679	tags=15%, list=8%, signal=17%
HALLMARK_G2M_CHECKPOINT	178	0.19088042	0.601792	0.8170974	0.9818409	1.0	4846	tags=24%, list=24%, signal=31%
HALLMARK_E2F_TARGETS	173	0.17057835	0.59205115	0.85265225	0.9589176	1.0	7214	tags=34%, list=35%, signal=52%

Figure 6: Gene sets enriched in phenotype DMSO(Cell Lines: HPB-ALL KOPT-K1)

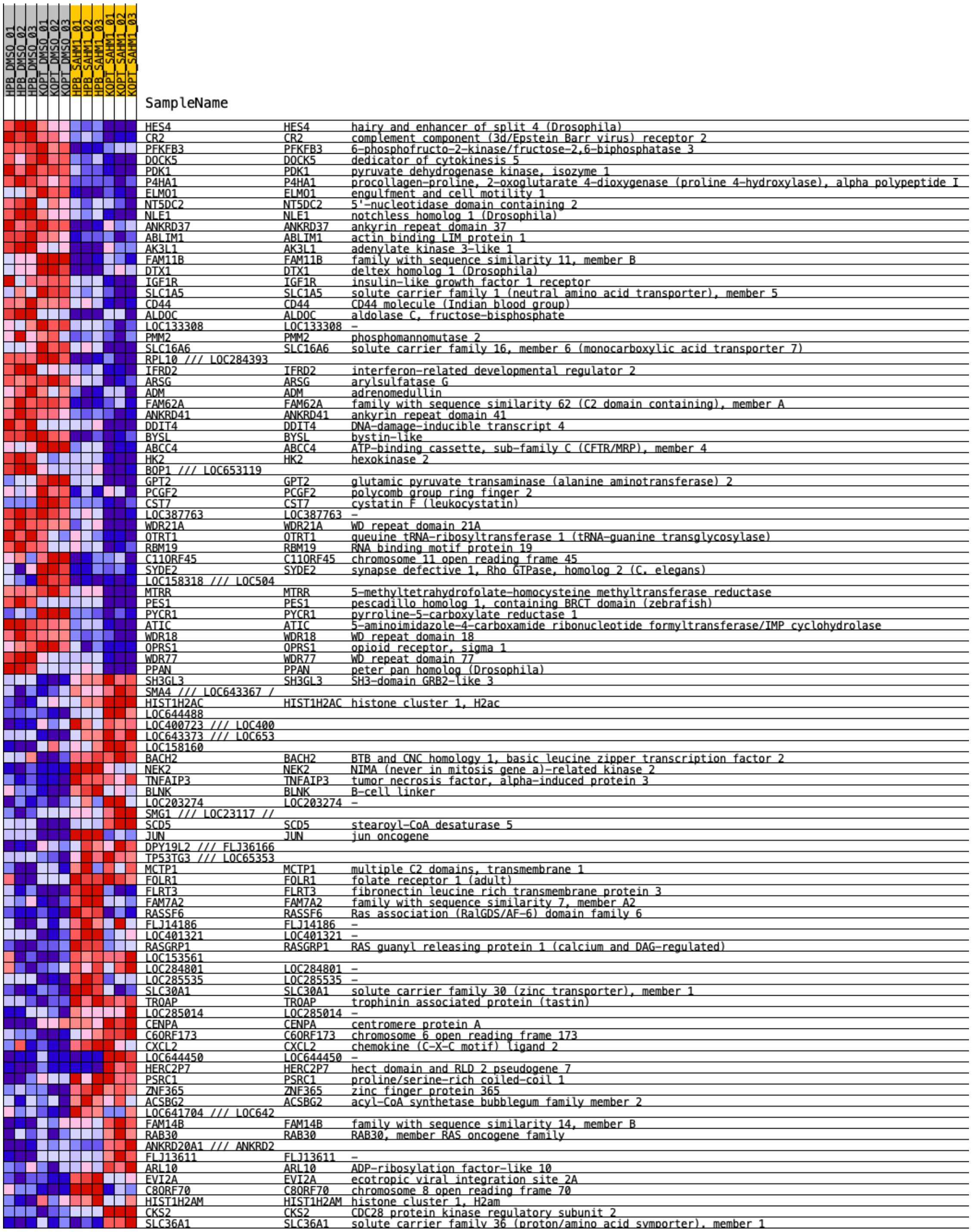


Figure 7: Heatmap(Cell Lines: HPB-ALL KOPT-K1)

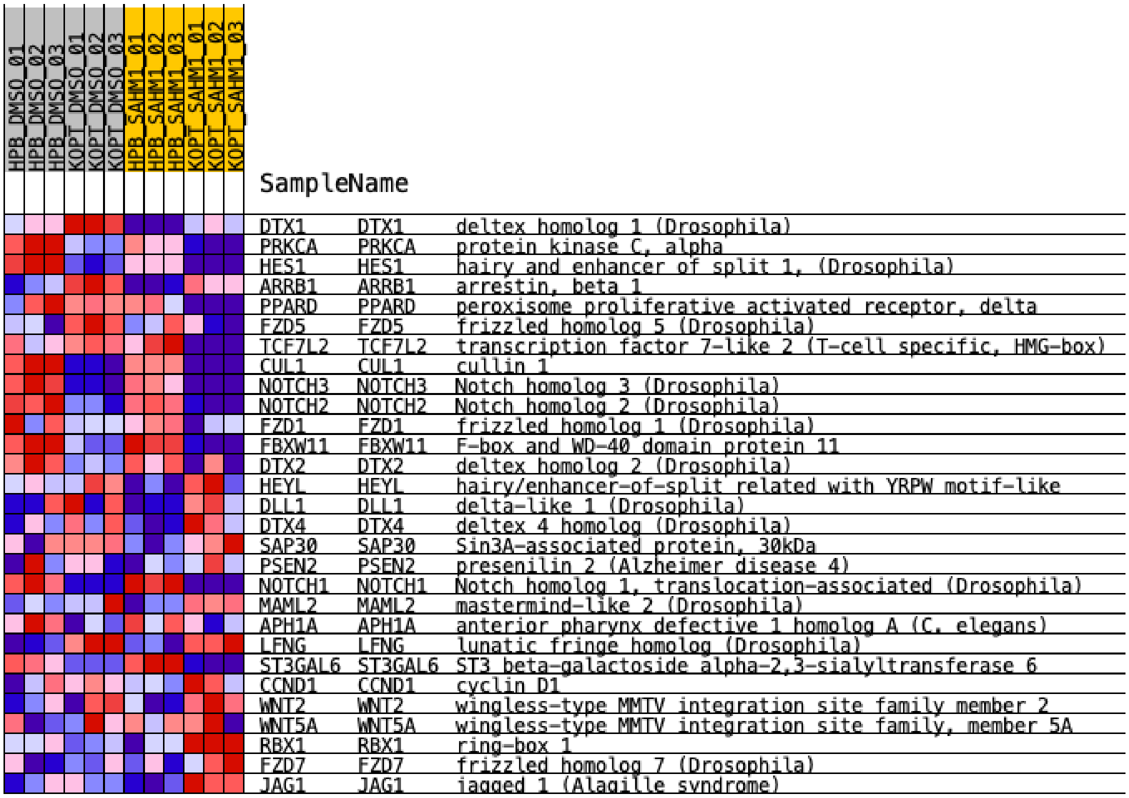


Figure 8: Heatmap(NOTCH signaling pathway)

The interface with python has not been necessary but we have developed in *bash* part of the conversions to the *gct* format and the generation of phenotype files (*.cls*). We use the expression set tranformed by *expresso* method from *affy* package as input on GSEA.

In R we used the *affy*, *limma*, and *simpleaffy* libraries and developed a pipeline similar to the one followed in the lectures. We group in functions the most used methods to be able to launch in a more compact way the different experiments.

Input data

We download from the GEO platform the raw data belonging to expression arrays Human Genome U133 Plus 2.0 with code *GSE18198*. These data have RNA information from cell cultures of the KOPT-K1 and HPB-ALL lines treated for 24h with SAHM1. Also their respective controls with the same amount of DMSO.

Analysis strategy

The quality analysis is carried out with the MAS 5.0 algorithm provided by simpleaffy.

The two cell lines were analyzed separately and jointly in R-limma and in GSEA.

In R-limma we use the Benjamini, Hochberg algorithm to control the false discovery rate derived from multiple testing and obtain adjusted p-values.

Before to fit the model in limma, we call the expresso command to got background correction, normalization and summarization to log_2 values.

The conversion to GSEA is done from the expression set obtained from the expresso command. Then we perform the conversions to the gct format.

The results in GSEA and R are coherent, at least in regard to our target: NOTCH signaling pathway on which we have concentrated exclusively.

1.14 Discusion

1.14.1 Quality

We use the simpleaffy R package that generates a series of metrics recommended by the manufacturer Affymetrix:

1. Average background
2. Scale factor
3. Percentage of genes called present.
4. 3′ to 5′ ratios (related with RNA degradation)

It is observed that all the indicators are within the acceptance margins (see graph of section 1.5), but that the patterns are clearly different between the samples of both cell lines.

1.14.2 Differential expression

We performed three different analysis to discover the effectiveness of SAHM1 in the inhibition of NOTCH.

1. Comparison between control (DMSO) and inhibitor (SAHM1) in the HBP-ALL cell line.
2. Comparison between control (DMSO) and inhibitor (SAHM1) in the KOPT-K1 cell line.
3. Comparison between control (DMSO) and inhibitor (SAHM1) joint for both cell lines.

It seems more correct to isolate each cell line separately in the analysis, according to the results of the quality analysis, where the expression patterns within cell lines appear more homogeneous than between. However, in figure 3 of the paper a heatmap is shown where the 12 samples seem to have been treated together, so we reproduced this analysis in case it could really show significant differences with the individual ones.

On the HPB-ALL cell line (1.9) and on the analysis of the two lines together (1.11) we found several direct targets of Notch TF among the 50 most significantly infra-regulated probes on inhibition scenario (lowest values of adjusted p-value and logFC <0): HES1, HES4, and DTX1, which are also investigated in the article. This result is also reproduced in the parallel analysis performed on GSEA. See figures 3 & 7.

GSEA also provides hallmarks that are overexpressed in the absence of inhibitory treatment, and among them we find Notch signaling (figures 2, 4 ,6)

The ES plots of the notch signalling pathway show that their gen-set is overrepresented in the high zone of the ranking in both cell lines. Figure 1 shows the one calculated for HPB-ALL.

These results are compatible with the expected effectiveness of SAHM1 as a NOTCH inhibitor.

In the analysis on the KOPT-K1 line, the hallmark NOTCH signaling still appears overexpressed in absence of inhibitor, but there are only traces of HES4 in the GSEA heatmap.

To get more insights we need to dive into notch hallmark in GSEA analysis.

So, for GSEA analysis performed in both cell-lines, we see (figure 8) that HES1, DTX1 and NOTCH homologous are under-expressed in treated samples, confirming the paper results.

1.14.3 Other results

We have not included them in this document, but we have also worked against GSEA from the raw experimental data, starting directly from the *affiBatch* object. The conversion to *gct* is somewhat different and is detailed in [1.7]. We have not time to analyze this results.

Another hallmark that I analyse are the MYC_targets (figures 9 and 10). MYC is mentioned at the paper and we have found articles that relate this with NOTCH signaling ("NOTCH1 directly regulates c-MYC and

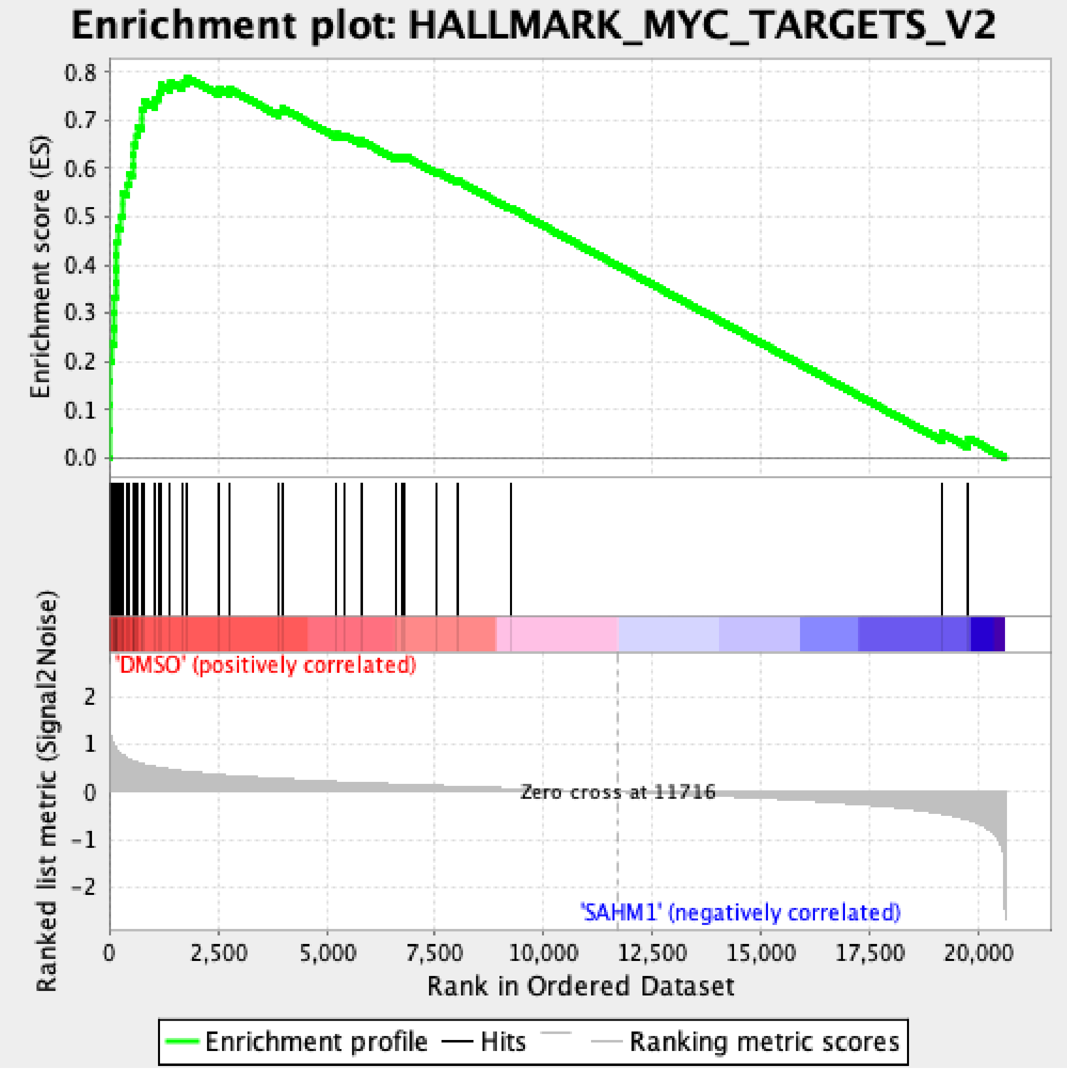


Figure 9: Enrichment plot MYC Target Pathway (HPB-ALL)

activates a feed-forward-loop transcriptional network promoting leukemic cell growth", Teresa Palomero et al.).

Also it is overrepresented in the upper area of the ranking related to control cells:

1.14.4 Conclusions

These results seem to confirm that the designed peptide is carrying out the desired functions related to the inhibition of the NOTCH signaling pathway, although we have not been able to delimit the targets in the same way as in the article, especially in the analysis of the cell line KOPT-K1.

It has become clear to me, as on other occasions, that differences between computational procedures can give rise to subtle and sometimes not so subtle differences between the data obtained. It is more than necessary to execute analysis with several tools, at least three. In this case we have done it with two: GSEA and R-limma, but in GSEA we do not input raw data, so being strict, we would not be following our own recommendations. Still, there are differences.

Another practice that we believe is advisable is to have a highly tested homemade version of the main algorithms. This can help to analyze the reliability of the software used. It's not necessary that this type of code be highly optimized.

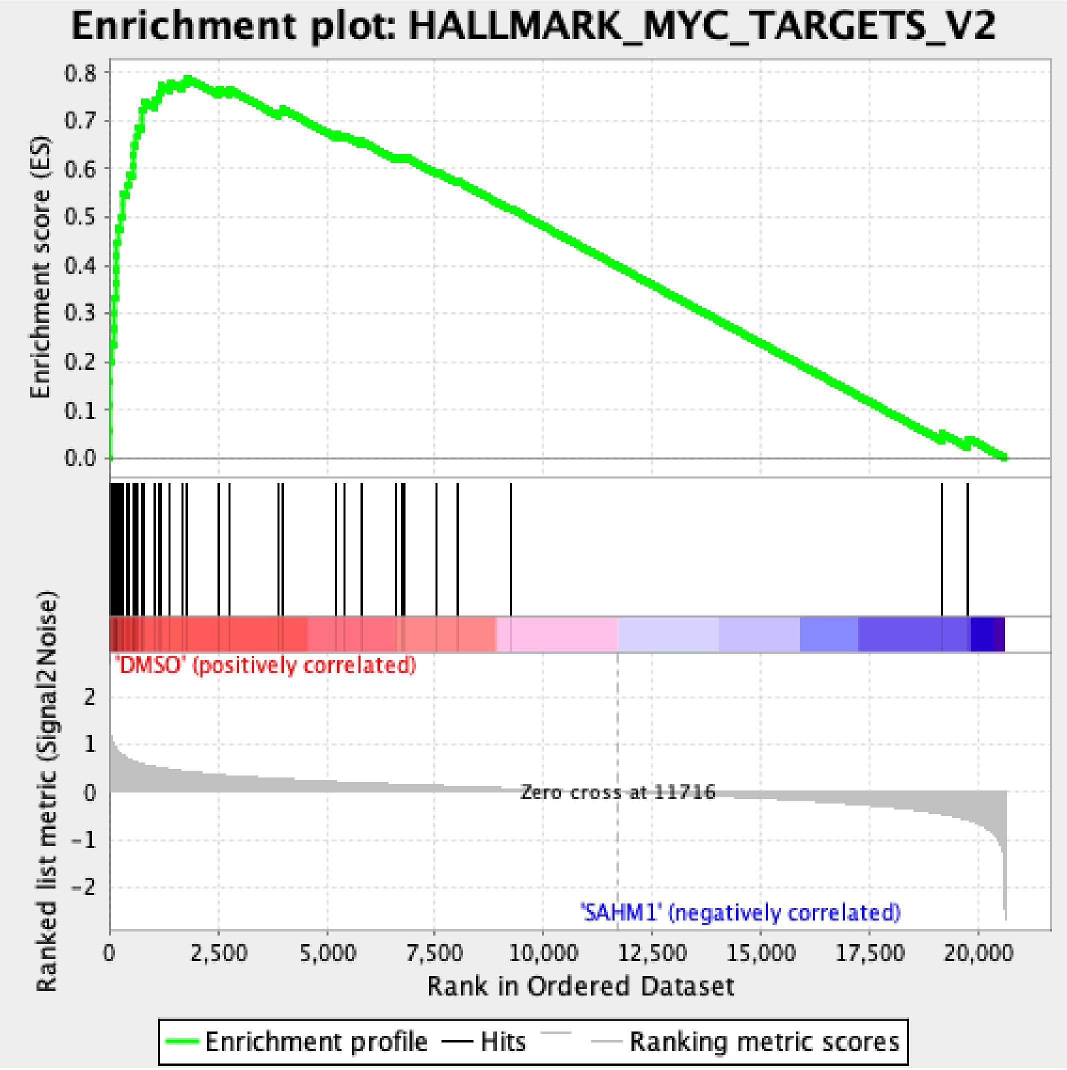


Figure 10: Enrichment plot MYC Target Pathway (KNOPT-K1)