Assignment RNAseq analysis, Systems Biology II, April, 2019

Mrinal Vashisth

in_class_ex_1

Log report from alignment of mouse rnaseq data to human reference

HISAT2 summary stats:

Total reads: 5000000

Aligned 0 time: 4230166 (84.60%)

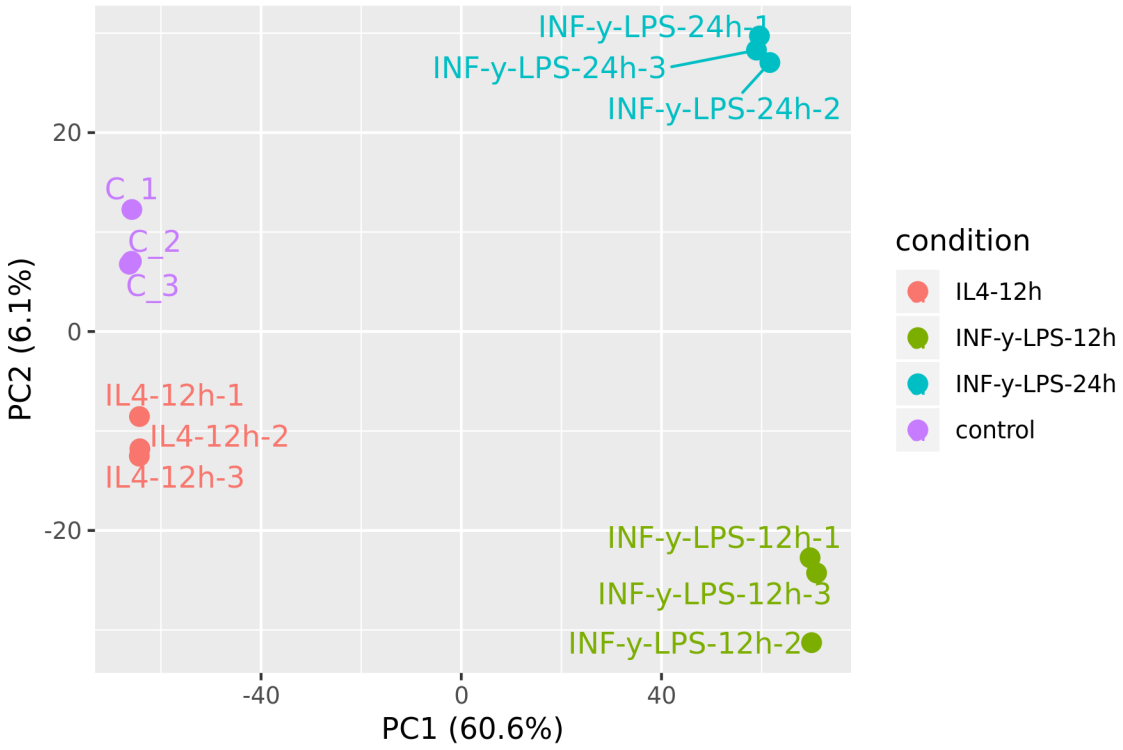
Aligned 1 time: 361379 (7.23%)

Aligned >1 times: 408455 (8.17%)

Overall alignment rate: 15.40%

in_class_ex_2

plotting PCA for the given data


INF-y-LPS-24h-1
INF-y-LPS-24h-3
INF-y-LPS-24h-2
C_1
C_2
C_3
IL4-12h-1
IL4-12h-2
IL4-12h-3
INF-y-LPS-12h-1
INF-y-LPS-12h-3
INF-y-LPS-12h-2
20
0
-20
-40
0
40
PC2 (6.1%)
PC1 (60.6%)
condition
IL4-12h
INF-y-LPS-12h
INF-y-LPS-24h
control

dataset_1

Mrinal Vashisth
22 April 2019

read the dataset into R

library(GEOquery) library(limma) #library(org.Mm.eg.db) library(org.Hs.eg.db)

for collapseBy

source("//home/manu/Documents/assignment/dataset_3_neuro/functions.r")

dir.create("cache")

res <- getGEO("GSE53890", AnnotGPL = TRUE, destdir="cache")[[1]] # info: Age effect on normal adult brain: frontal cortical region

Collapsing the data

for collapseBy

source("//home/student/deseq/functions.R")

str(experimentData(res)) str(pData(res)) head(fData(res)) res$Sex:ch1

res$condition <- gsub("\\+", "_", res$Sex:ch1) res$condition

res <- collapseBy(res, fData(res)$Gene symbol, FUN=median) res <- res[!grepl("///", rownames(res)),] res <- res[rownames(res) != "",]

there is a lot of garbage there

fData(res) <- data.frame(row.names = rownames(res))

fData(res)$entrez <- row.names(fData(res))

fData(res) $org.Hs.eg.db, keys=fData(res)$symbol<-mapIds¿ entrez, keytype="SYMBOL", column="ENTREZID")

res.qnorm <- res

summary(exprs(res.qnorm)) exprs(res.qnorm) <- normalizeBetweenArrays(log2(exprs(res.qnorm)+1), method="quantile") summary(exprs(res.qnorm))

res.qnorm.top12K <- res.qnorm res.qnorm.top12K <- res.qnorm.top12K[head(order(apply(exprs(res.qnorm.top12K), 1, mean), decreasing = TRUE), 12000),]

res.design <- model.matrix(~0+condition, data=pData(res.qnorm.top12K))

fit <- lmFit(res.qnorm.top12K, res.design)

fit2 <- contrasts.fit(fit, makeContrasts(conditionFemale-conditionMale, levels=res.design))

fit2 <- eBayes(fit2) de <- topTable(fit2, adjust.method="BH", number=Inf) head(de) library(data.table) de <- as.data.table(de, keep.rownames=TRUE) de[entrez == "REST"]

FGSEA

de2 <- data.frame(de $entrez, de$ t) colnames(de2) <- c('ENTREZ', 'stat')

library(fgsea)

ranks <- deframe(de2) head(ranks, 20)

Load the pathways into a named list

library(msigdbr)

m_df <- msigdbr(species = "Homo sapiens") m_df pathways <- split(m_df $human_gene_symbol, m_df$ gs_name)

filter the list to include only hallmark pathways

library(data.table)

pathways.hallmark <- m_df[m_df$gs_name %like% “HALLMARK_”,]
pathways.hallmark <- split(pathways.hallmark
$human_{g}ene_{s}ymbol, pathways.hallmark$ gs_name)

Show the first few pathways, and within those, show only the first few genes.

pathways.hallmark %>% head() %>% lapply(head)

running the fgsea algorithm on hallmark.pathways

fgseaRes <- fgsea(pathways=pathways.hallmark, stats=ranks, nperm=1000)

fgseaResTidy <- fgseaRes %>% as_tibble() %>% arrange(desc(NES)) # ggplotting for hallmark pathways

pdf(“fgseaResTidy.pdf”, width = 100, height = 1000)

ggplot(fgseaResTidy, aes(reorder(pathway, NES), NES)) + geom_col(aes(fill=pval<0.05)) + coord_flip() + labs(x=“Pathway”, y=“Normalized Enrichment Score”, title=“Hallmark pathways NES from GSEA”) + theme_minimal()

dev.off()

let’s look at all pathways and select for

pathways associated with brain in significance threshold

running the fgsea algorithm on all pathways

fgseaRes.all <- fgsea(pathways=pathways, stats=ranks, nperm=1000)

number <- data.frame(grep(“REST”, fgseaRes.all
$fgseaRes.all, rownames(fgseaRes.all)$
$leadingEdge¿¿colnames(number)<-c('row_{n}umber')REST<-subset¿$
row_number)

using tidy to view pretty results

fgseaResTidy <- REST %>% as_tibble() %>% arrange(desc(NES)) # Show in a nice table for all pathways fgseaResTidy %>% dplyr::select(-leadingEdge, -ES, -nMoreExtreme) %>% arrange(padj) %>% DT::datatable()

ggplotting for all pathways

pdf("fgseaResTidy_REST.pdf", width=100, height=1000)

ggplot(fgseaResTidy, aes(reorder(pathway, NES), NES)) + geom_col(aes(fill=pval<0.05)) + coord_flip() + labs(x="Pathway", y="Normalized Enrichment Score", title="All pathways NES from GSEA") + theme_minimal()

dev.off()

We can see that two pathways are significant, associated with

ion channel activity

scale rows

xt<-t(as.matrix(res.qnorm.top12K)) xts<-scale(xt) xtst<-t(xts) xtst <-na.omit(xtst)

only grab top 1000 by p-value

h<-head(xtst, n = 1000L)

set layout options - adjust if labels get cut off

pdf("heatmap.pdf",width=500, height=500)

draw heatmap allowing larger margins and adjusting row label font size

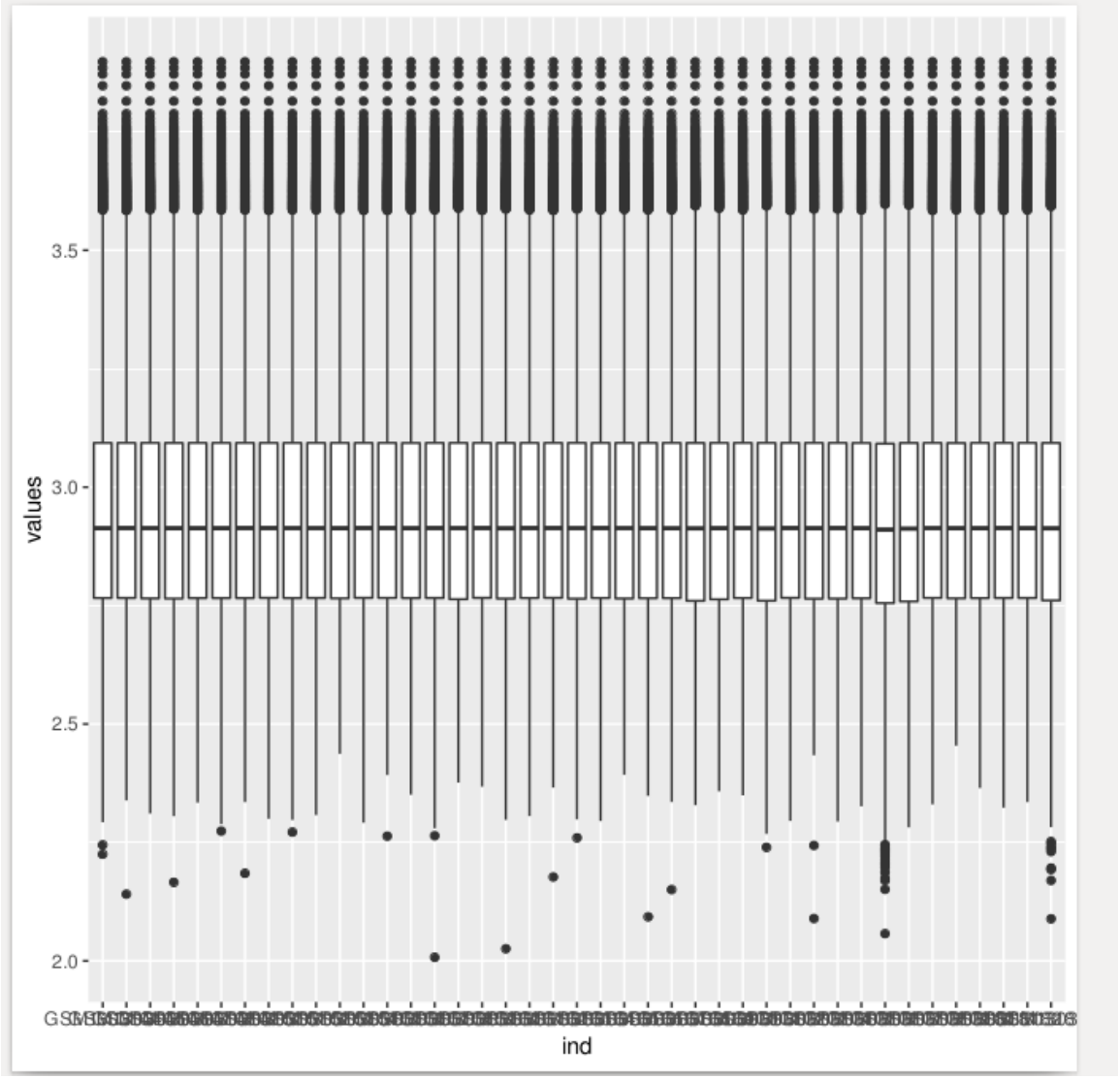
heatmap(h)

output plot to file

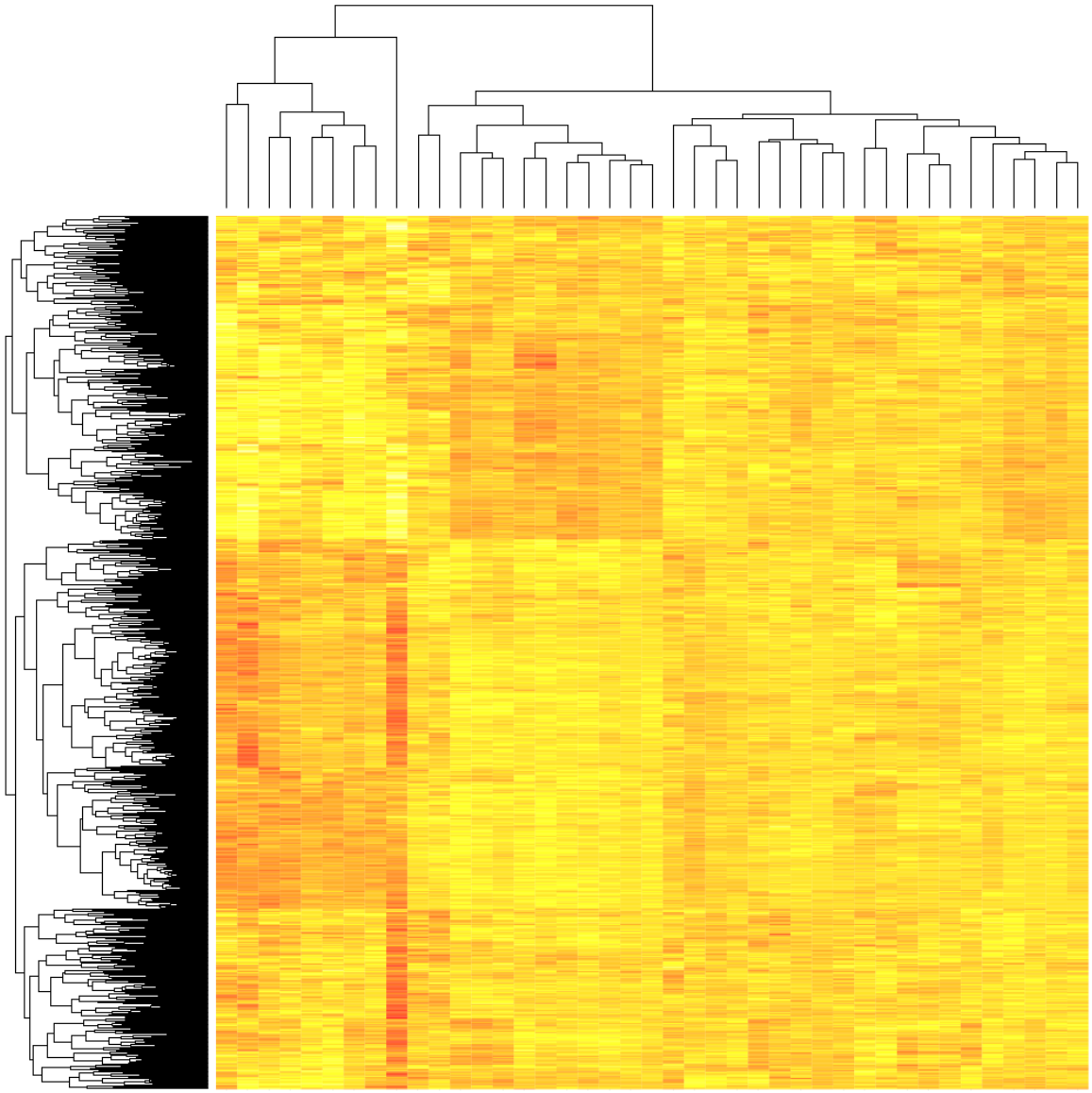
```
dev.off()
```

install.packages(‘devtools’)

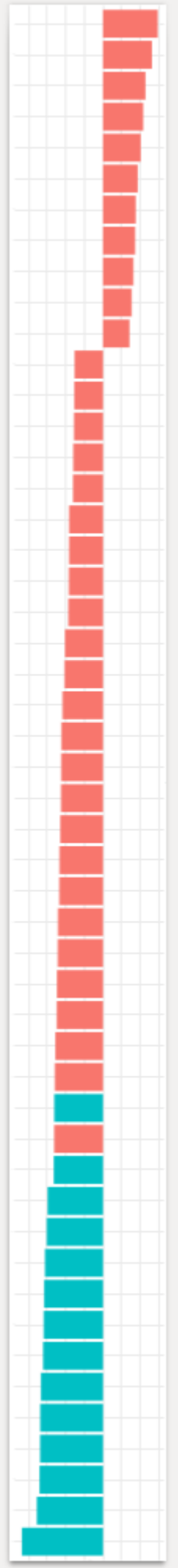
```
library(devtools) # devtools::install_github(“sinhrks/ggfortify”)
library(ggfortify)

pdf(‘box_dataset1.pdf’)

ggplot(stack(data.frame(t(xt))), aes(x = ind, y = values)) +
geom_boxplot()

dev.off()
```



Box plot of all samples



Heatmap of top 1000 genes



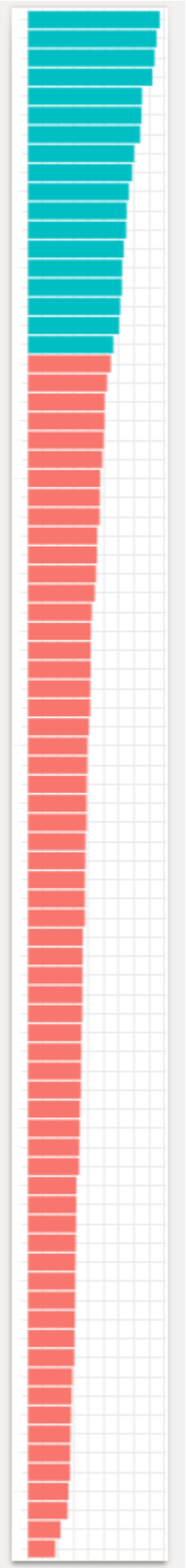
Expression analysis

dataset_2

Mrinal Vashisth
22 April 2019

read the dataset into R

library(GEOquery) library(limma)
library(org.Mm.eg.db)
library(org.Hs.eg.db)



for collapseBy

source(“//home/manu/Documents/assignment/dataset_3_neuro/functions.r”)

dir.create(“cache”)

res <- getGEO(“GSE28790”, AnnotGPL = TRUE, destdir=“cache”)[[1]] # info: SIRT1 impact on global gene expression in the brain

Collapsing the data

for collapseBy

source(“//home/student/deseq/functions.R”)

str(experimentData(res)) str(pData(res)) head(fData(res)) res$title

res$condition <- c("BSKO", "BSKO", "BSKO", "BSKO", "WT", "WT", "WT", "WT") res$condition

res <- collapseBy(res, fData(res)$Gene symbol, FUN=median) res <- res[!grepl(“///”, rownames(res)),] res <- res[rownames(res) != “”,]

there is a lot of garbage there

fData(res) <- data.frame(row.names = rownames(res))

fData(res)$entrez <- row.names(fData(res))

fData(res) $org.Mm.eg.db, keys=fData(res)$ symbol<-mapIds¿ entrez, keytype=“SYMBOL”, column=“ENTREZID”)

res.qnorm <- res

summary(exprs(res.qnorm)) exprs(res.qnorm) <- normalizeBetweenArrays(log2(exprs(res.qnorm)+1), method=“quantile”) summary(exprs(res.qnorm))

res.qnorm.top12K <- res.qnorm res.qnorm.top12K <- res.qnorm.top12K[head(order(apply(exprs(res.qnorm.top12K), 1, mean), decreasing = TRUE), 12000),]

res.design <- model.matrix(~0+condition, data=pData(res.qnorm.top12K))

fit <- lmFit(res.qnorm.top12K, res.design)

fit2 <- contrasts.fit(fit, makeContrasts(conditionBSKO-conditionWT, levels=res.design))

fit2 <- eBayes(fit2) de <- topTable(fit2, adjust.method="BH", number=Inf) head(de)

library(data.table) de <- as.data.table(de, keep.rownames=TRUE) de[entrez == "Sirt2"] de[entrez == "Sirt3"] de[entrez == "Sirt4"] de[entrez == "Sirt7"]

FGSEA

de2 <- data.frame(de $entrez, de$ t) colnames(de2) <- c('ENTREZ', 'stat')

library(fgsea)

ranks <- deframe(de2) head(ranks, 20)

Load the pathways into a named list

library(msigdbr)

m_df <- msigdbr(species = "Homo sapiens") m_df pathways <- split(m_df $human_gene_symbol, m_df$ gs_name)

filter the list to include only hallmark pathways

library(data.table)

pathways.hallmark <- m_df[m_df$gs_name %like% "HALLMARK_",] pathways.hallmark <- split(pathways.hallmark $human_gene_symbol, pathways.hallmark$ gs_name)

Show the first few pathways, and within those, show only the first few genes.

pathways.hallmark %>% head() %>% lapply(head)

running the fgsea algorithm on hallmark.pathways

```
fgseaRes <- fgsea(pathways=pathways.hallmark, stats=ranks, nperm=1000)

fgseaResTidy <- fgseaRes %>% as_tibble() %>% arrange(desc(NES)) # ggplotting for hallmark pathways

pdf("fgseaResTidy.pdf", width = 20, height = 20)

ggplot(fgseaResTidy, aes(reorder(pathway, NES), NES)) + geom_col(aes(fill=pval<0.05)) + coord_flip() + labs(x="Pathway", y="Normalized Enrichment Score", title="Hallmark pathways NES from GSEA") + theme_minimal()

dev.off()
```

We can see that two of the hallmark pathways are significant

let's look at all pathways and select for

pathways associated with SIRT in significance threshold

running the fgsea algorithm on all pathways

```
fgseaRes.all <- fgsea(pathways=pathways, stats=ranks, nperm=1000)
```

number <- data.frame(grep(“Sirt”, fgseaRes.all$leadingEdge)) # colnames(number) <- c('row_number') # Sirt <- subset(fgseaRes.all, rownames(fgseaRes.all) %in% number$row_number)

using tidy to view pretty results

fgseaResTidy <- fgseaRes.all %>% as_tibble() %>% arrange(desc(NES)) # Show in a nice table for all pathways fgseaResTidy %>% dplyr::select(-leadingEdge, -ES, -nMoreExtreme) %>% arrange(padj) %>% DT::datatable()

ggplotting for all pathways

pdf(“fgseaResTidy_all.pdf”, width=20, height=500)

ggplot(fgseaResTidy, aes(reorder(pathway, NES), NES)) + geom_col(aes(fill=pval<0.05)) + coord_flip() + labs(x=“Pathway”, y=“Normalized Enrichment Score”, title=“All pathways NES from GSEA”) + theme_minimal()

dev.off()

plot pca

df_12k <- data.frame(res.qnorm.top12K@assayData$exprs) colnames(df_12k) <- c(‘BSKO’, ‘BSKO’, ‘BSKO’, ‘BSKO’, ‘WT’, ‘WT’, ‘WT’, ‘WT’) p <- prcomp(na.omit(df_12k))

devtools::install_github(“sinhrks/ggfortify”) library(ggfortify)

ggplot2::autoplot(p, label = FALSE, shape = FALSE, loadings.label = TRUE)

ggplot(stack(df_12k), aes(x = ind, y = values)) + geom_boxplot()

We can see that there are various activated pathways

 # SIRT1 is a NAD dependent deacetylase that provides coping

mechanism against nutritional changes in cell.

It is shown that by activation of genes encoding for

MAO-A (monoamine oxidase). SSRIs are inhibited.

SSRIs play a key role in uptake of serotonin and

thus manifestation of symptoms of depression.

 # In analysis we can see that indeed expression

of this gene is increased. Further, MAO-A inhibitors or SSRIs

(selective serotonin reuptake inhibitors)

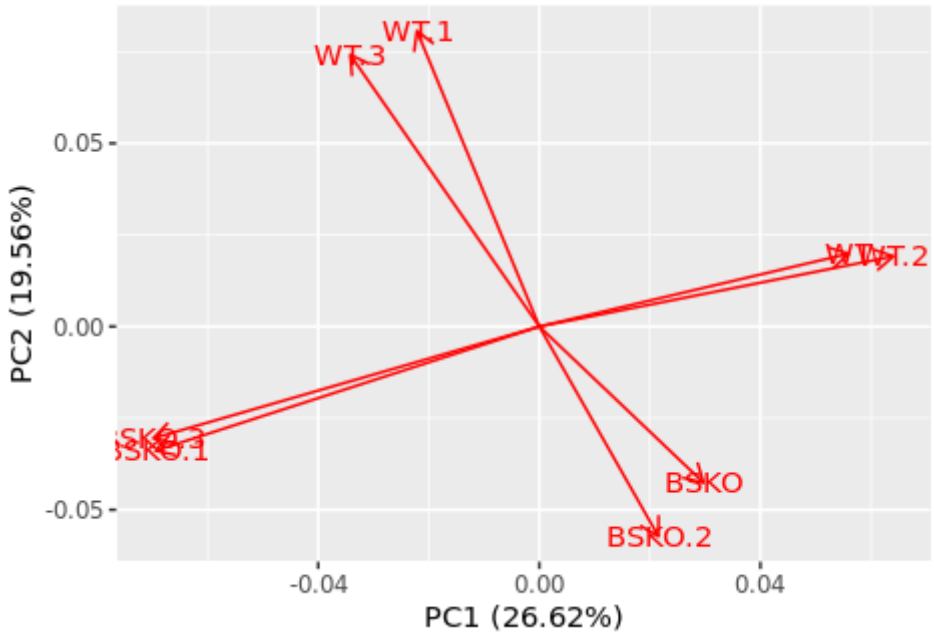
normalise anxiety differences between wild-type and mutant animals.

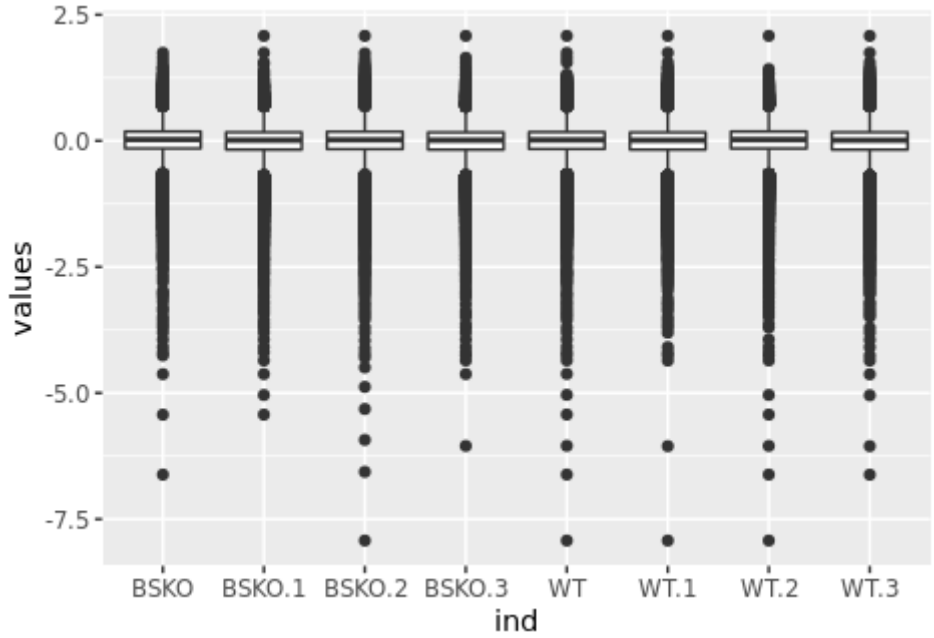
 # From network analysis we can see the involvement

in Sleep Cycle. Perhaps a disrupted sleep cycle and

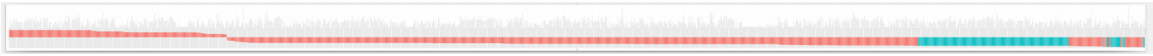
altered cell signalling response is promoting

anxious behaviour in these mice.

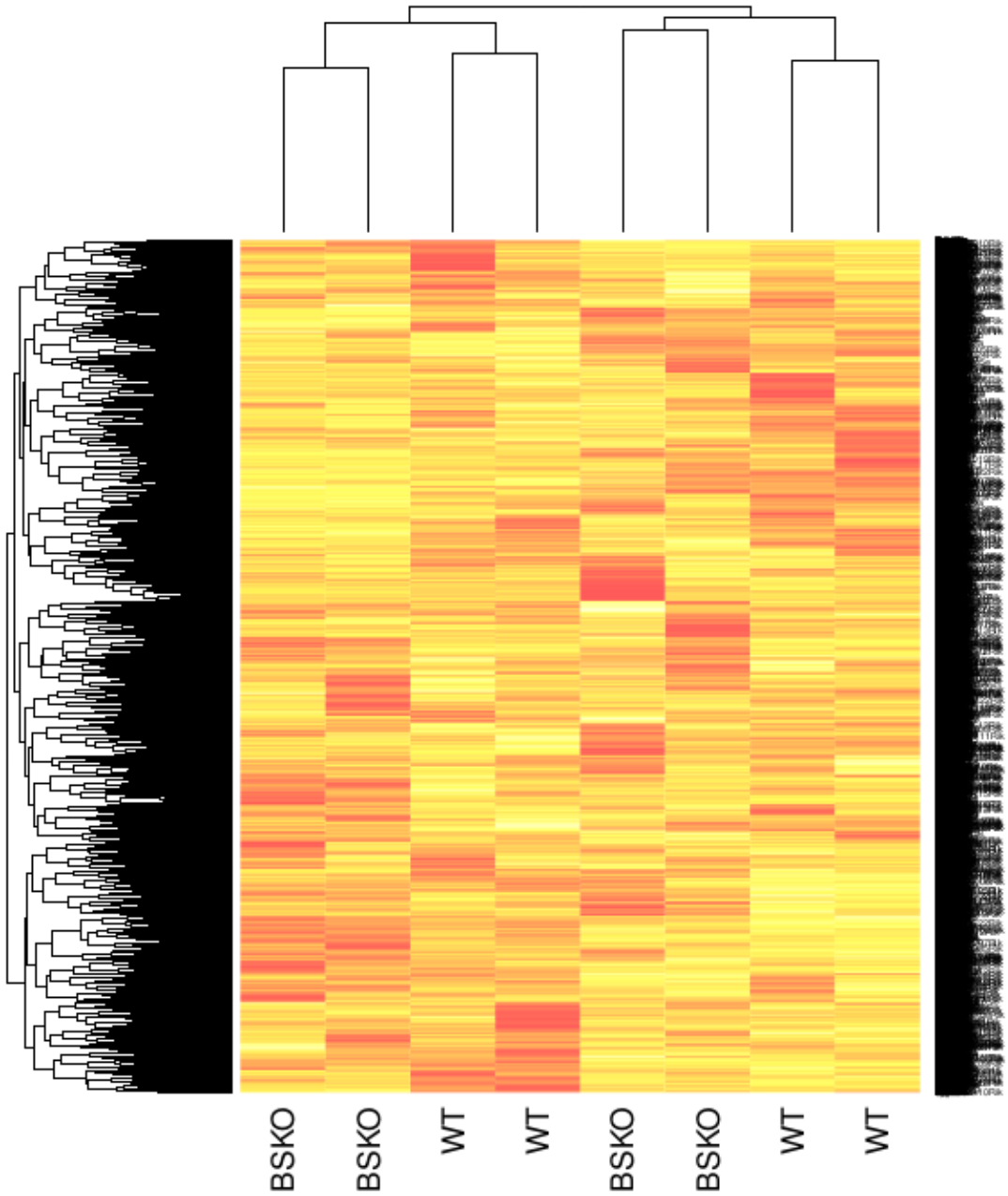




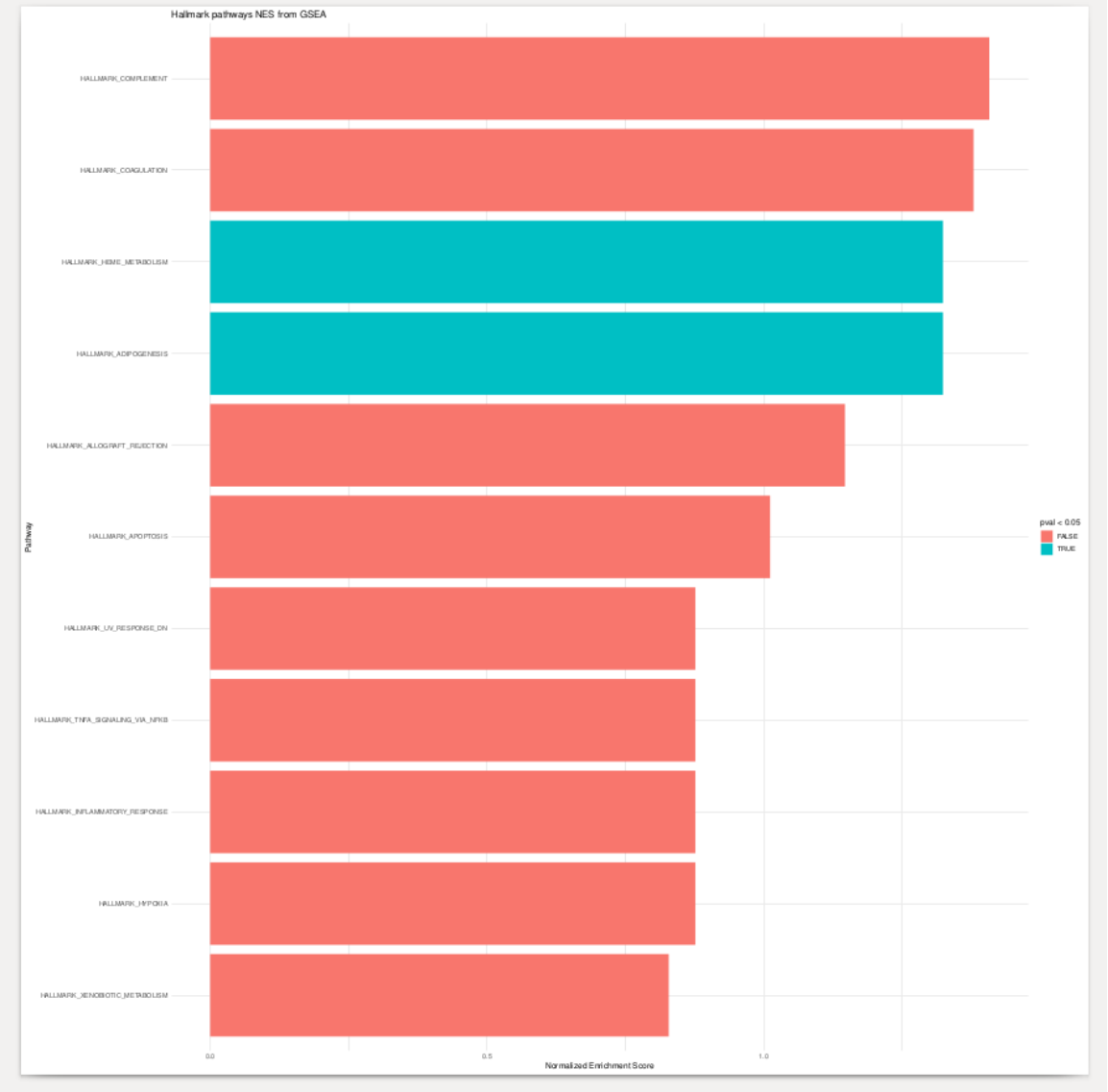
PCA and box plot for the data.



Enriched pathways for Sirt gene.



Heatmap of the data.



Enriched hallmark pathways for Sirt gene.

dataset_3

Mrinal Vashisth
23 April 2019

read the dataset into R

library(GEOquery) library(limma) # library(org.Mm.eg.db) library(org.Hs.eg.db)

for collapseBy

source("//home/manu/Documents/assignment/dataset_3_neuro/functions.r")

dir.create("cache")

to get the dataset uncomment the comment line below and execute

res <- getGEO("GSE19404", AnnotGPL = TRUE, destdir="cache")[[1]] # info: SIRT1 impact on global gene expression in the brain #

Collapsing the data

for collapseBy

str(experimentData(res)) str(pData(res)) head(fData(res)) res $characteristics_ch1 <- NULL$ res characteristics_ch1 <- t(annotate[1,])

res$title

library(readr) annotate <- read_delim("~/Desktop/annotate", " ", escape_double = FALSE, col_names = FALSE, trim_ws = TRUE) View(annotate)

annotate <- t(annotate) res $condition <- NULL$ res condition <- annotate res$condition

res <- collapseBy(res, fData(res)$Gene symbol, FUN=median) res <- res[!grepl("///", rownames(res)),] res <- res[rownames(res) != "",]

there is a lot of garbage there

fData(res) <- data.frame(row.names = rownames(res))

fData(res)$entrez <- row.names(fData(res))

fData(res) $org.Mm.eg.db, keys=fData(res)$ entrez, $symbol<-mapIds¿$ keytype="SYMBOL", column="ENTREZID")

res.qnorm <- res

summary(exprs(res.qnorm)) exprs(res.qnorm) <- normalizeBetweenArrays(log2(exprs(res.qnorm)+1), method="quantile") summary(exprs(res.qnorm))

res.qnorm.top12K <- res.qnorm res.qnorm.top12K <- res.qnorm.top12K[head(order(apply(exprs(res.qnorm.top12K), 1, mean), decreasing = TRUE), 12000),]

res.design <- model.matrix(~0+condition, data=pData(res.qnorm.top12K))

intermediate <- data.frame(res.design) colnames(intermediate) <- c("ATRT", "CNS", "Medulloblastoma", "Normal", "Pineoblastoma") rm(res.design) res.design <- as.matrix(intermediate)

fit <- lmFit(res.qnorm.top12K, res.design)

fit2 <- contrasts.fit(fit, makeContrasts(ATRT-Normal, CNS-Normal, Medulloblastoma-Normal, Pineoblastoma-Normal, levels=res.design))

fit2 <- eBayes(fit2) de <- topTable(fit2, adjust.method="BH", number=Inf) head(de)

library(data.table) de <- as.data.table(de, keep.rownames=TRUE) entry <- function(){ item <- data.frame("CDK4","SEC61G","TSPAN31","LANCL2", "EGFR","APC","ATM","BMPR1A","BRCA1", "BRCA2","CDK4","CDKN2A","CREBBP","EGFR", "EP300","ETV6","FHIT","FLT3","HRAS","KIT", "MET","MLH1","NTRK1","PAX8","PDGFRA", "PRCC","PRKAR1A","PTEN","RET","STK11", "TFE3","TP53","WWOX") item<- t(item) rownames(item) <- NULL

x<- for (i in item){ print(de[entrez == i])

}

return(x)

}

entry()

FGSEA

install.packages(‘fgsea’)

install.packages(‘tibble’)

library(fgsea) library(tibble)

de2 <- data.frame(de $entrez, de$ P.Value) colnames(de2) <- c(‘ENTREZ’, ‘stat’)

ranks <- deframe(de2) head(ranks, 20)

Load the pathways into a named list

install.packages(‘msigdbr’)

library(msigdbr)

m_df <- msigdbr(species = “Homo sapiens”) m_df pathways <- split(m_df $human_gene_symbol, m_df$ gs_name)

filter the list to include only hallmark pathways

library(data.table)

pathways.hallmark <- m_df[m_df$gs_name %like% “HALLMARK_”,] pathways.hallmark <- split(pathways.hallmark $human_gene_symbol, pathways.hallmark$ gs_name)

Show the first few pathways, and within those, show only the first few genes.

```
pathways.hallmark %>% head() %>% lapply(head)
```

running the fgsea algorithm on hallmark.pathways

```
fgseaRes <- fgsea(pathways=pathways.hallmark, stats=ranks, nperm=1000)

fgseaResTidy <- fgseaRes %>% as_tibble() %>% arrange(desc(NES)) # ggplotting for hallmark pathways

library(ggplot2) pdf("fgseaResTidy.pdf", width = 20, height = 20)

ggplot(fgseaResTidy, aes(reorder(pathway, NES), NES)) + geom_col(aes(fill=pval<0.05)) + coord_flip() + labs(x="Pathway", y="Normalized Enrichment Score", title="Hallmark pathways NES from GSEA") + theme_minimal()

dev.off()
```

We can sees seven significant hallmark pathways

let's look at all pathways and select for

pathways associated with genes of intrest in significance threshold

running the fgsea algorithm on all pathways

```
fgseaRes.all <- fgsea(pathways=pathways, stats=ranks, nperm=1000)
```

searching for the genes in pathway and appending the rownumbers

```
sink('numbers.txt')
```

options(max.print=2000) for(i in item){ print(grep(i, fgseaRes.all$leadingEdge)) }

sink()

we have to do a lot of cleaning of the data before importing it as csv

getting only unique values from all numbers

unique_vals <- data.frame(as.integer(unique(unlist(numbers))))

colnames(unique_vals) <- c('row_number') final <- subset(fgseaRes.all, rownames(fgseaRes.all) %in% unique_vals$row_number)

using tidy to view pretty results

install.packages('DT')

fgseaResTidy <- fgseaRes.all %>% as_tibble() %>% arrange(desc(NES))

Show in a nice table for all pathways

library(DT)

fgseaResTidy %>% dplyr::select(-leadingEdge, -ES, -nMoreExtreme) %>% arrange(padj) %>% DT::datatable()

ggplotting for all pathways

ggplot(fgseaResTidy, aes(reorder(pathway, NES), NES)) + geom_col(aes(fill=pval<0.05)) + coord_flip() + labs(x="Pathway", y="Normalized Enrichment Score", title="All pathways NES from GSEA") + theme_minimal()

plot pca

df_12k <- data.frame(res.qnorm.top12K@assayData$exprs) colnames(df_12k) <- annotate p <- prcomp(na.omit(df_12k))

install.packages(‘devtools’)

```
library(devtools) # devtools::install_github(“sinhrks/ggfortify”)
library(ggfortify)
```

```
ggplot2::autoplot(p, label = TRUE, shape = FALSE, loadings.label =
TRUE)
```

```
ggplot(stack(df_12k), aes(x = ind, y = values)) + geom_boxplot()
```

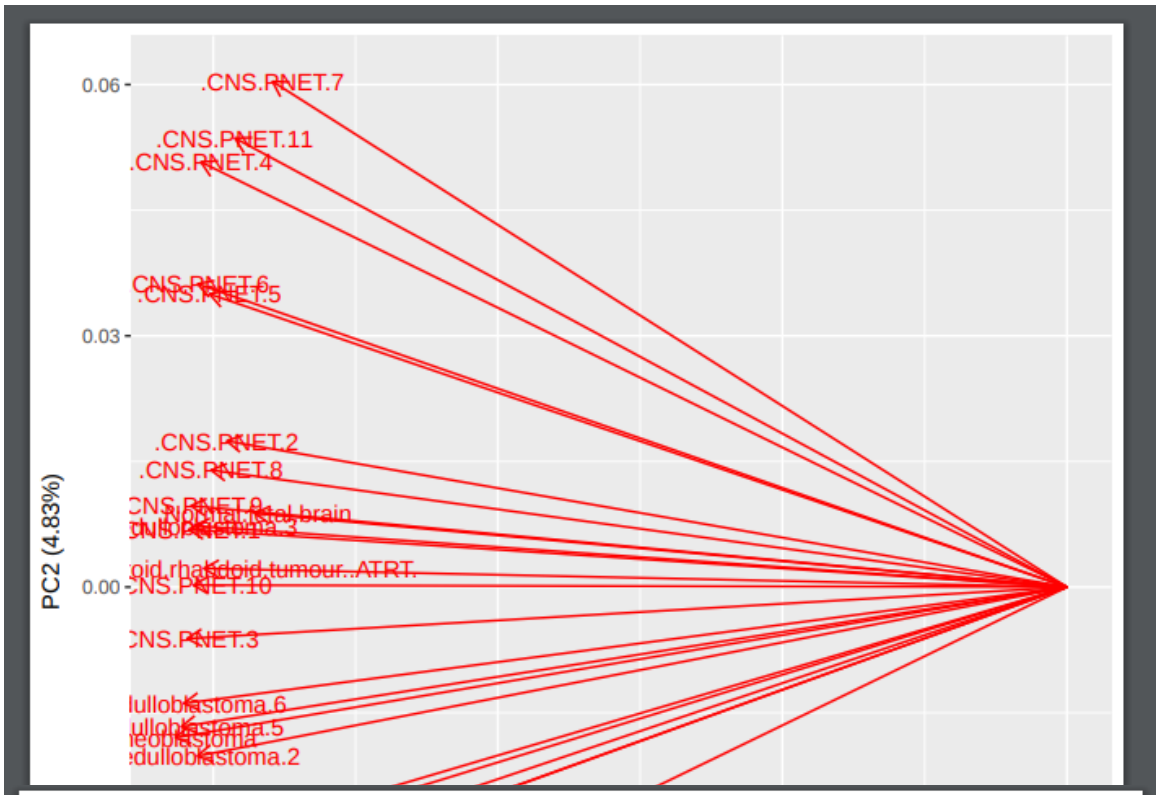
In the paper, authors identified similarities

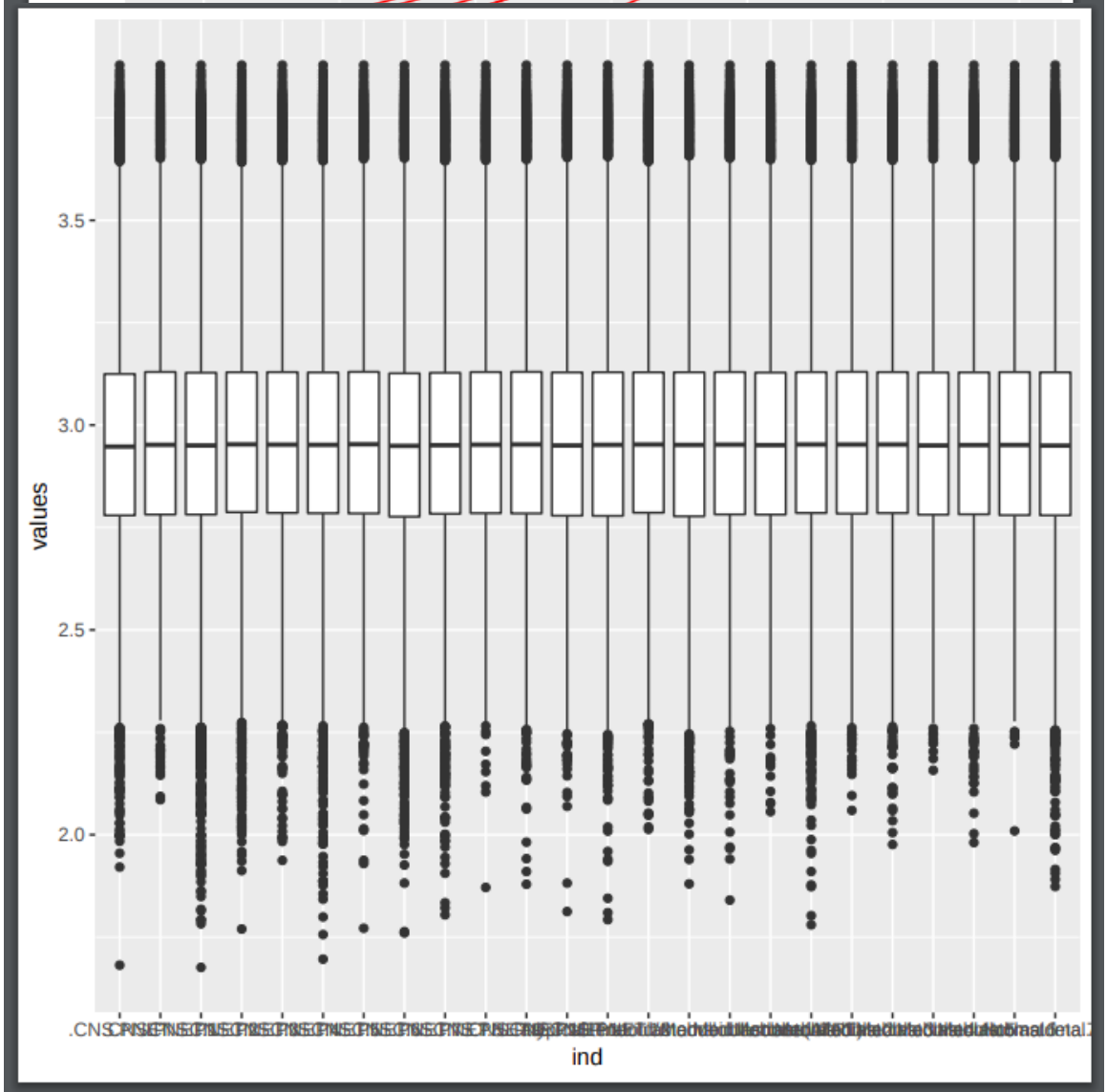
in transcriptomes of distinct human brain tumor,

specifically glioblastomas of TCGA classical

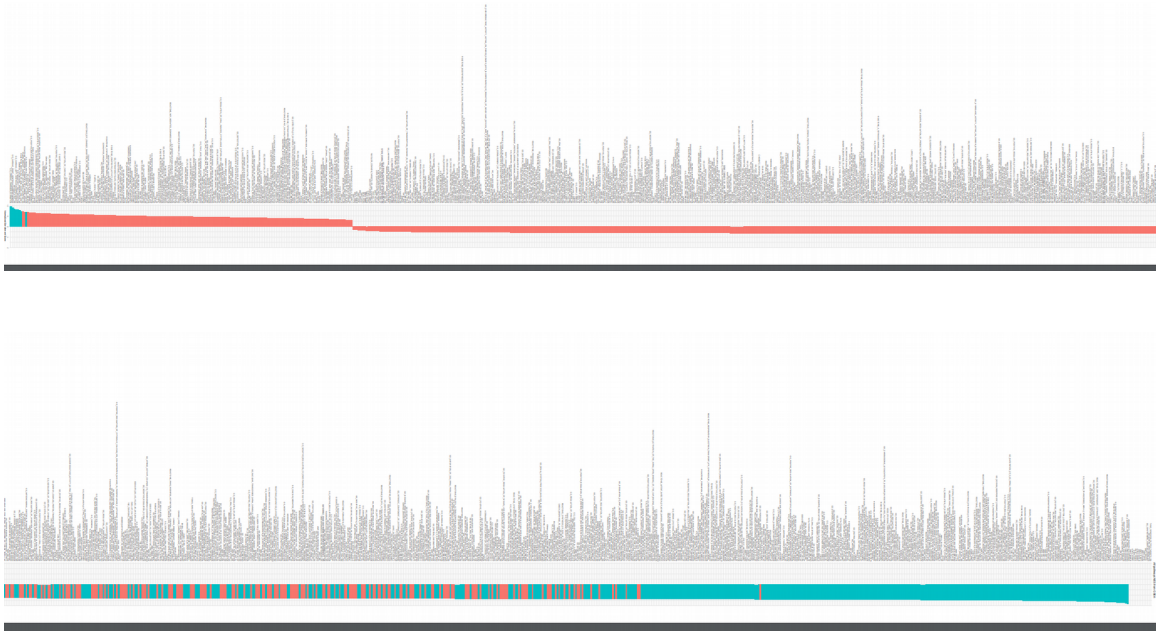
subtype and oligodendroglial tumors, Philips

proferative gliomas and AT/RT.


0.06
0.03
0.00
PC2 (4.83%)
.CNS.PNET.7
.CNS.PNET.11
.CNS.PNET.4
.CNS.PNET.5
.CNS.PNET.2
.CNS.PNET.8
.CNS.PNET.10
.CNS.PNET.3
.CNS.PNET.6


3.5
3.0
2.5
2.0
values
ind

PCA and box plot of the data



Pathway analysis for all 33 genes. A long list of ~ 5000 pathways.

Hallmark pathways