

# Unsupervised Hierarchical Clustering of Non-diabetic Single Cell Ensemble Transcriptomes Without Hormone Marker Genes

## Introduction

This file will detail the steps used to perform unsupervised hierarchical clustering analysis on the non-diabetic single cell transcriptomes without using pancreatic cell hormone marker genes (INS, GCG, SST, PPY, GHRL, COL1A1, PRSS1, and KRT19) as shown in Supplemental Fig S7.

## Hierarchical Clustering

```
suppressPackageStartupMessages(library(Biobase))
suppressPackageStartupMessages(library(edgeR))
suppressPackageStartupMessages(library(ape))
suppressPackageStartupMessages(library(gplots))
suppressPackageStartupMessages(library(dendextend))
suppressPackageStartupMessages(library(RColorBrewer))
library(edgeR)
library(Biobase)
library(gplots)
library(dendextend)
library(ape)
library(RColorBrewer)
rm(list=ls())
set.seed(53079239)
# File name
fname = "NonT2D.log2cpm.no.hormones"
setwd("/Users/lawlon/Documents/Final_RNA_Seq_3/Data/")
# Load single cell data
load("nonT2D.rdata")
p.anns <- featureData(cnts.eset)
probe.anns <- as(p.anns,"data.frame")
ND.anns <- pData(cnts.eset)
# Remove multiples and keep all other groups
ND.sel <- ND.anns[ND.anns$cell.type %in% c("INS", "PPY", "GCG", "SST",
                                           "COL1A1", "KRT19", "PRSS1", "none"),]
ND.counts <- exprs(cnts.eset)
cpms <- cpm(x = ND.counts)
data <- log2(cpms+1)
data <- data[,rownames(ND.sel)]
# Combine sample anns and expression data
s.anns.sel <- ND.sel
r.max <- apply(data,1,max)
# Use highly expressed genes
data.sel <- data[r.max > 10.5,]
ND.data.sel<- data.sel[, rownames(ND.sel)]
```

```

# Remove hormonal genes from list
horm <- which(probe.anns$Associated.Gene.Name %in% c("INS", "GCG", "PPY", "SST", "GHRL",
                                                    "COL1A1", "PRSS1", "KRT19"))

ids <- rownames(probe.anns)[horm]
indices <- which(rownames(ND.data.sel) %in% ids)
ND.data.sel <- ND.data.sel[-indices,]

# Save a copy of the data
exp.sel <- ND.data.sel
# Change column name labels to cell type
colnames(ND.data.sel)[1:dim(ND.data.sel)[1]] <- ND.data.sel$cell.type

# Change name of one KRT19 cell to ghrelin cell
g <- which(probe.anns$Associated.Gene.Name == "GHRL")
ghrl <- data[g,]
samp <- which(ghrl > 15)
g.idx <- which(rownames(s.anns.sel) == names(samp))
colnames(ND.data.sel)[g.idx] <- "GHRL"

p.res <- probe.anns[rownames(ND.data.sel),]
# Combine probe anns with selected cpm values
ND.data.sel.exp <- cbind(p.res, ND.data.sel)
# Write genes used for clustering to file
write.csv(ND.data.sel.exp, paste(fname, "genes_selected_for_cing.csv", sep = "."))

# Dendrogram of samples using hclust
d <- dist(t(ND.data.sel))
hc.final <- hclust(d, method="ward.D2")

# Change hclust object to dendrogram
dend1 <- as.dendrogram(hc.final)
groupCodes <- s.anns.sel$cell.type

# Color Schema
grey <- brewer.pal(n=9, name="Greys")
colorCodes <- c(INS="#e41a1c", GCG = "#377eb8", SST = "#4daf4a",
                PPY = "#984ea3", GHRL = "#ff7f00",
                COL1A1 = grey[9], PRSS1 = grey[7], KRT19 = grey[5],
                none = grey[3])

namelist <- c("Beta", "Alpha", "Delta", "Gamma", "Epsilon",
              "Stellate", "Acinar", "Ductal", "none")
# Change label colors
labels_colors(dend1) <- colorCodes[groupCodes][order.dendrogram(dend1)]

# Change dend to phylo object
dend2 <- as.phylo(dend1)

# Match up colors and labels
cols = NULL
for (i in 1:length(labels(dend2))) {
  if ((dend2$tip.label[i] %in% names(colorCodes)) == TRUE) {
    cols <- c(cols, colorCodes[dend2$tip.label[i]])
  }
}

```

```

}
}

#Use the long hyphen or the minus sign instead of regular hyphen symbol
labels(dend2) <- rep(x = "-", length(labels(dend2)))

# Create high resolution tiff of dendrogram
tiff(file=paste(fname, "dendrogram.tiff", sep = "."),
     width = 9000, height = 9000, units = "px", res = 800)
plot(dend2, type = "fan", tip.color = cols, cex = 6.5, label.offset = 0)
legend("bottomleft", title = "Cell Types", title.col = "black",
      legend = c(expression(bold("Beta (INS)")), expression(bold("Alpha (GCG)")),
                  expression(bold("Delta (SST)")), expression(bold("Gamma (PPY)")),
                  expression(bold("Epsilon (GHRL)")),
                  expression(bold("Stellate (COL1A1)")), expression(bold("Acinar (PRSS1)")),
                  expression(bold("Ductal (KRT19)")), expression(bold("None"))), text.col = colorCodes,
      cex = 0.75, xjust=0, yjust=0)
dev.off()

```

## Session Information

```

suppressPackageStartupMessages(library(Biobase))
suppressPackageStartupMessages(library(edgeR))
suppressPackageStartupMessages(library(ape))
suppressPackageStartupMessages(library(gplots))
suppressPackageStartupMessages(library(dendextend))
suppressPackageStartupMessages(library(RColorBrewer))
library(edgeR)
library(Biobase)
library(gplots)
library(dendextend)
library(ape)
library(RColorBrewer)
sessionInfo()

## R version 3.3.0 (2016-05-03)
## Platform: x86_64-apple-darwin13.4.0 (64-bit)
## Running under: OS X 10.11.3 (El Capitan)
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] parallel stats graphics grDevices utils datasets methods
## [8] base
##
## other attached packages:
## [1] RColorBrewer_1.1-2 dendextend_1.1.8 gplots_3.0.1
## [4] ape_3.5 edgeR_3.14.0 limma_3.28.7
## [7] Biobase_2.32.0 BiocGenerics_0.18.0
##
## loaded via a namespace (and not attached):

```

## [1]	Rcpp_0.12.5	knitr_1.13	whisker_0.3-2
## [4]	magrittr_1.5	lattice_0.20-33	stringr_1.0.0
## [7]	caTools_1.17.1	tools_3.3.0	grid_3.3.0
## [10]	nlme_3.1-128	KernSmooth_2.23-15	htmltools_0.3.5
## [13]	gtools_3.5.0	yaml_2.1.13	digest_0.6.9
## [16]	formatR_1.4	bitops_1.0-6	evaluate_0.9
## [19]	rmarkdown_0.9.6	gdata_2.17.0	stringi_1.1.1