Baseline

Gupta & Anand

October 27, 2016

Raw <- read.csv("~/Downloads/nbt.3154-S3-raw.csv", row.names = 1)  
Nor <- read.csv("~/Downloads/nbt.3154-S3.csv", row.names = 1)  
Norm <- as.matrix(Nor)  
desc1=unlist(strsplit(rownames(Norm),"\_"))  
desc=desc1[seq(1,7867,2)]  
gr4sfg=which(substr(rownames(Norm),1,5)=="4SFGA")  
gr4sf=which(substr(rownames(Norm),1,4)=="4SGA")  
gr1=which(substr(rownames(Norm),1,2)=="PS")  
gr2=which(substr(rownames(Norm),1,2)=="NP")  
gr3=which(substr(rownames(Norm),1,2)=="HF")  
colscells=c("blue","green","orange","red","purple")  
colnb=rep(0,3934)  
colnb[gr1]=1  
colnb[gr2]=2  
colnb[gr3]=3  
colnb[gr4sf]=4  
colnb[gr4sfg]=5  
typesort=rep(0,3934)  
typesort[which(nchar(desc)<5 & substr(rownames(Norm),3,3)=="A")]="sortA"  
typesort[which(nchar(desc)<5 & substr(rownames(Norm),3,3)=="B")]="sortB"  
typesort[which(nchar(desc)>4)]="sortA"  
ftable(typesort)

## typesort sortA sortB  
##   
## 3175 759

celltypes=as.factor(c("PS","NP","HF","4G","4GF")[colnb])  
cellcol=colscells[colnb]

Compute centroids of fate types

fourg <- Nor[celltypes == "4G",]  
fourgf <- Nor[celltypes == "4GF",]  
  
fourg\_centroid <- colMeans(fourg)  
fourgf\_centroid <- colMeans(fourgf)

Assign each cell to closest centroid

assign <- function(profile) {  
 dist\_fourg <- sqrt(sum((profile - fourg\_centroid)^2))  
 dist\_fourgf <- sqrt(sum((profile - fourgf\_centroid)^2))  
 return(if(dist\_fourg < dist\_fourgf) "4G" else "4GF")  
}  
  
assignments = apply(Nor, 1, assign)  
sum(assignments == "4G")

## [1] 1100

sum(assignments == "4GF")

## [1] 2834

endothelial <- c("Cdh5", "Erg", "HoxB4", "Sox7", "Sox17")  
erythroid <- c("Gata1", "Gfi1b", "HbbbH1", "Ikaros", "Myb", "Nfe2")  
  
fourgf\_rel <- t(data.frame(fourgf\_centroid))  
fourg\_rel <- t(data.frame(fourg\_centroid))  
  
erythroid\_distance <- function(profile) {  
 return(sqrt(sum((profile - fourgf\_rel[,erythroid])^2)))  
}  
  
endothelial\_distance <- function(profile) {  
 return(sqrt(sum((profile - fourg\_rel[,endothelial])^2)))  
}  
  
erythroid\_distances <- apply(Nor[,erythroid], 1, erythroid\_distance)  
endothelial\_distances <- apply(Nor[,endothelial], 1, endothelial\_distance)  
  
assignment\_error <- (sum(erythroid\_distances[assignments == "4G"]) + sum(endothelial\_distances[assignments=="4GF"]))/nrow(Nor)  
non\_assignment\_error <- (sum(erythroid\_distances[assignments == "4GF"]) + sum(endothelial\_distances[assignments == "4G"]))/nrow(Nor)

The mean assignment error is 21.1119312, and the mean non-assignment error is 8.2270436. Ideally, the assignment error should be much smaller than the non-assignment error, which would indicate that a cell is much closer to its assigned centroid in terms of the transcription factors relevant to that fate than to the centroid it was not assigned to in terms of transcription factors relevant to the fate it did not achieve.