Covariate Adjustment of Expression Data

Adjust gene expression data by batch, sex, and RIN and analyze cell type surrogate proportion variables

The analysis for cell type surrogate proportion variables uses the CellCODE library(https://github.com/mchikina/CellCODE) based on Chikina, Zaslavsky, & Sealfon, (2015) Bioinformatics, 10, 1584-1591. Paper can be found here: https://academic.oup.com/bioinformatics/article/31/10/1584/177237.

Setup

```
# load libraries
library(here)
install.packages(here("code","CellCODE"), repos=NULL, type="source")
library(easypackages)
libraries("limma","CellCODE")
options(stringsAsFactors = FALSE)
```

Read in data

```
# load gene expression data, gene information, and labels
load(here("data","tidy","exprData.Rdata"))

# construct model
cov_columns = c("batch2","batchWG6","sex","RIN")
full_model = model.matrix(~0+as.factor(Dx) +
                            as.factor(batch) +
                            as.factor(sex) +
                            RIN,
                          data=labelData)
colnames(full_model) = c("ASD", "TD",cov_columns)

# fit model ----------------------------------------------------------
fit = lmFit(exprData,full_model)

# remove batch, sex, and RIN
beta1 = fit$coefficients[, cov_columns, drop = FALSE]
beta1[is.na(beta1)] = 0
exprDataAdj = exprData - beta1 %*% t(full_model[,cov_columns])

# save adjusted expression data
save(exprDataAdj,geneInfo,labelData,
     file = here("data","processed","exprDataAdj.Rdata"))
```

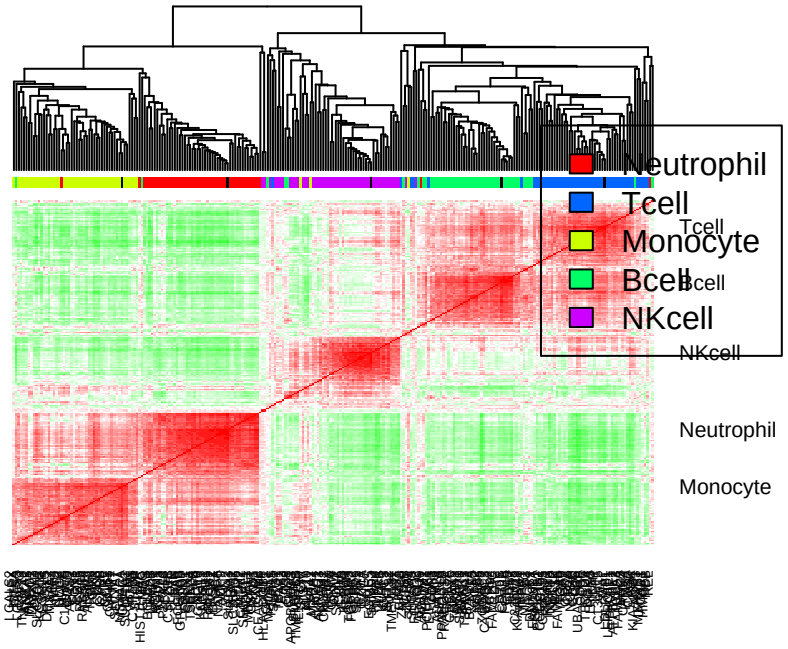
Estimate surrogate proportion variables for leukocyte cell types

```
data("IRIS")

ref_data = exprDataAdj
rownames(ref_data) = geneInfo$geneSymbol

cellTypes2use = c("Neutrophil-Resting","CD4Tcell-N0",
                  "Monocyte-Day0", "Bcell-naïve","NKcell-control")
cellTypeNames = c("Neutrophil","Tcell", "Monocyte","Bcell", "NKcell")
irisTag = tagData(IRIS[,cellTypes2use],
                  cutoff = 2,
                  max = 50,
                  ref = ref_data,
                  ref.mean = F)
colnames(irisTag)= cellTypeNames

SPVs=getAllSPVs(data = ref_data,
                grp = labelData$subgrp2,
                dataTag = irisTag,
                method =  "mixed" ,
                plot = TRUE)
```



Test SPVs for group-difference

```
SPVs = data.frame(SPVs)
labelData$Neutrophil = SPVs$Neutrophil
labelData$Tcell = SPVs$Tcell
labelData$Monocyte = SPVs$Monocyte
labelData$Bcell = SPVs$Bcell
labelData$NKcell = SPVs$NKcell
```

```r
cols2use = c("Fstat","pval")
aov_res = data.frame(matrix(nrow = length(cellTypeNames),
                            ncol = length(cols2use)))
rownames(aov_res) = cellTypeNames
colnames(aov_res) = cols2use

for (i in 1:length(cellTypeNames)){
  form2use = as.formula(sprintf("%s ~ subgrp2",cellTypeNames[i]))
  mod2use = lm(formula = form2use, data = labelData)
  res = anova(mod2use)
  aov_res[cellTypeNames[i],"Fstat"] = res["subgrp2","F value"]
  aov_res[cellTypeNames[i],"pval"] = res["subgrp2","Pr(>F)"]
}
aov_res
```

```
##                Fstat      pval
## Neutrophil 1.2569308 0.2884049
## Tcell      0.8111317 0.4468806
## Monocyte   1.9553526 0.1461911
## Bcell      1.3541728 0.2622459
## NKcell     1.0345787 0.3586596
```