Subtype-Aware Batch Correction Retains Biological Signal of Integrated Breast Cancer Datasets

Gil Tomás
gil.tomas@igmm.ed.ac.uk

Contents


```
library(data.table)  # data frame manipulation
library(oligo)  # preprocessing oligonucleotide arrays
library(limma)  # differential expression analysis
library(AIMS)  # implements AIMS classifier (non-parametric version of PAM50)
library(sva)  # implements ComBat
library(ggplot2)  # sophisticated plotting framework
library(ggsignif)  # significance bars for ggplot2
library(irr)  # Cohen's Kappa
library(knitr)  # kable function for tables
library(kableExtra)  # format kable tables
library(xtable)
```

```
## Directories
rdsDir <- "../out/rds"
csvDir <- "../data/csv"
dsetDir <- "../data/rds"
libDir <- "../lib"
graphsDir <- "../out/pdf"
hrmnDataDir <- "../eddie/data/out"

## Affy Chips
chips <- c("p2", "a")

## Normalisation Methods
normMths <- c("frma", "mas5", "rma")

## Colours
cols <- c("Basal" = "red2",
          "Her2" = "purple",
          "Luminal B" = "cadetblue2",
          "Luminal A" = "dodgerblue4",
          "Normal" = "forestgreen")

## Seed
set.seed(42)
```

1—Dataset Acquisition

Raw CEL files from ten breast cancer gene expression dataset where downloaded from GEO and normalized with fRMA, RMA and MAS5. Demographics of each dataset are shown in **Table 1**.

```
## load datasets table
dsets.dfr <- read.csv(file.path(csvDir, "datasets.csv"), stringsAsFactors = FALSE)
dsets.dtb <- data.table(dsets.dfr)
dsets.dtb[, `:=`(platform, gsub("_", "", platform))]

## split datasets by chip
p2Dsets <- dsets.dtb[, id[grepl("p2", id)]]
aDsets <- dsets.dtb[, id[!grepl("p2", id)]]
## setkey(dsets.dtb, id)
dsets.dtb[, `:=`(dataset = NULL, from = NULL, notes = NULL)]
setnames(dsets.dtb, c("ref", "fracER", "fracHER2"), c("GSE", "fracER+", "fracHER2+"))
setcolorder(dsets.dtb, c("GSE", "id", "platform", "nSamples", "fracER+", "fracHER2+"))
setorder(dsets.dtb, "platform", "fracER+", "fracHER2+")
## sir-p2/GSE17907 has had 4 samples removed because they were not labeled
## HER2- (see bellow)
dsets.dtb[GSE == "GSE17907", `:=`(nSamples, 33)]
```