dpIEL-Microbiome project, Batch correction

Puspendu Sardar, Ph.D, Department of medicine, University of Cambridge, UK

Install required packages

```
cran.packages <- c('knitr', 'xtable', 'ggplot2', 'vegan', 'cluster',
                   'gridExtra', 'pheatmap', 'ruv', 'lmerTest', 'bapred')
#install.packages(cran.packages)
bioconductor.packages <- c('sva', 'mixOmics', 'affyPLM', 'limma',
'AgiMicroRna',
                           'variancePartition', 'pvca')
#install.packages(bioconductor.packages)

if (!requireNamespace('BiocManager', quietly = TRUE))
  install.packages('BiocManager')
#BiocManager::install(bioconductor.packages)

#install.packages("remotes")
#remotes::install_github("EvaYiwenWang/PLSDAbatch")
```

Load required packages

```
library(mixOmics)
library(sva) # For Combat
library(ggplot2)
library(gridExtra)
library(vegan)
library(pvca)
library(PLSDAbatch)
library(doParallel)
library(ConQuR)
```

Taxonomic analysis

Load taxonomic data

```
taxon_df <-
read.delim("combat_seq/batch_correct_after_rclr/rclr_nonbatch.txt", sep =
"\t",
                       row.names = 1, header = TRUE)
dim(taxon_df)

## [1]  127 4630
```

Load metadata

```
crohn_metadata <- read.delim("metadata_HBI.txt", header = TRUE, sep = "\t")
head(crohn_metadata)

##      Sample HBI_score HBI_3High HBI_3Low Country PRJNA_Study Study_ID
## 1 ERR209679     HBI_0       Low      Low   Spain   PRJEB1220  Study_4
```

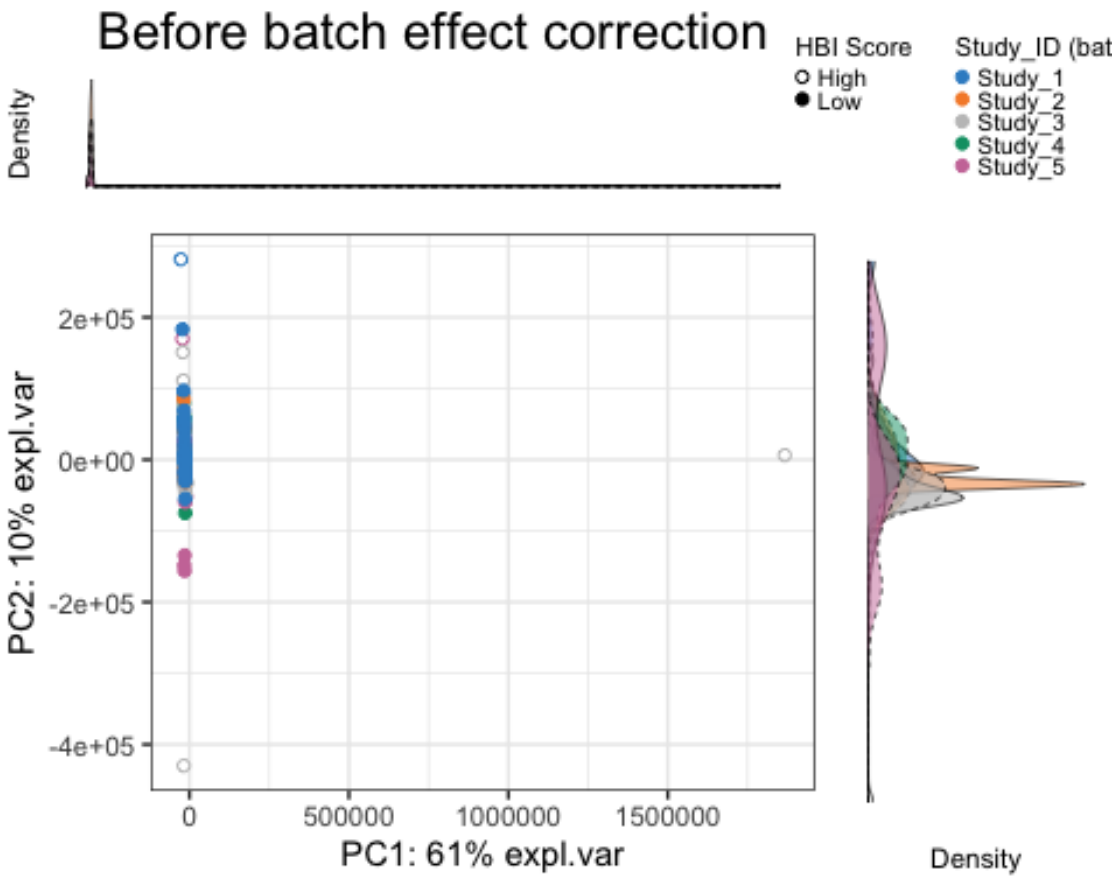
```
## 2 ERR209680     HBI_1       Low      Low   Spain   PRJEB1220  Study_4
## 3 ERR209684     HBI_1       Low      Low   Spain   PRJEB1220  Study_4
## 4 ERR209690     HBI_0       Low      Low   Spain   PRJEB1220  Study_4
## 5 ERR209695     HBI_0       Low      Low   Spain   PRJEB1220  Study_4
## 6 ERR209698     HBI_0       Low      Low   Spain   PRJEB1220  Study_4
```

PERMANOVA before batch correction

```
before_mat <- as.matrix(taxon_df)
before.dist <- vegdist(before_mat, method='bray')
before.div <- adonis2(before.dist ~ Country + Study_ID + HBI_3Low, data =
crohn_metadata,
                      method='bray')
```

PCA plot before batch correction

```
pca_before <- pca(taxon_df, ncomp = 3)
pca_plot_before <- Scatter_Density(pca_before, batch =
crohn_metadata$Study_ID,
                trt = crohn_metadata$HBI_3Low,
                batch.legend.title = 'Study_ID (batch)',
                trt.legend.title = 'HBI Score',
                title = 'Before batch effect correction')
```



Batch correction using ConQur

```
batchid = factor(crohn_metadata[, 'Study_ID'])
#tax_corrected = ConQuR(tax_tab = taxon_df, batchid = batchid, batch_ref =
"Study_3",
#covariates = crohn_metadata$HBI_3Low)
#tax_corrected_df <- as.data.frame(tax_corrected)
```

PERMANOVA before batch correction

```
#after_mat <- as.matrix(tax_corrected_df)
#after.dist <- vegdist(after_mat, method='bray')
#after.div <- adonis2(after.dist ~ Country + Study_ID + HBI_3Low, data =
crohn_metadata, method='bray')
```

PCA plot after batch correction

```
#pca_after <- pca(tax_corrected, ncomp = 3)
#pca_plot_after <- Scatter_Density(pca_after, batch =
crohn_metadata$Study_ID,
                                  #trt = crohn_metadata$HBI_3Low,
                                  #batch.legend.title = 'Study_ID (batch)',
                                  #trt.legend.title = 'HBI Score',
                                  #title = 'After batch effect correction')
```

Save the plots

```
# pdf(file = "Taxonomy_batch_effect_before_after.pdf", width = 16, height =
6)
#grid.arrange(pca_plot_before, pca_plot_after,
#              ncol = 2)
# dev.off()
```

Functional analysis

Load functional data

```
function_df <-
read.delim("functional/KEGG_rowSum_batchCorrected/read_count_EC.txt", sep =
"\t",
                           row.names = 1, header = TRUE)
#dim(function_df)
```

PERMANOVA before batch correction

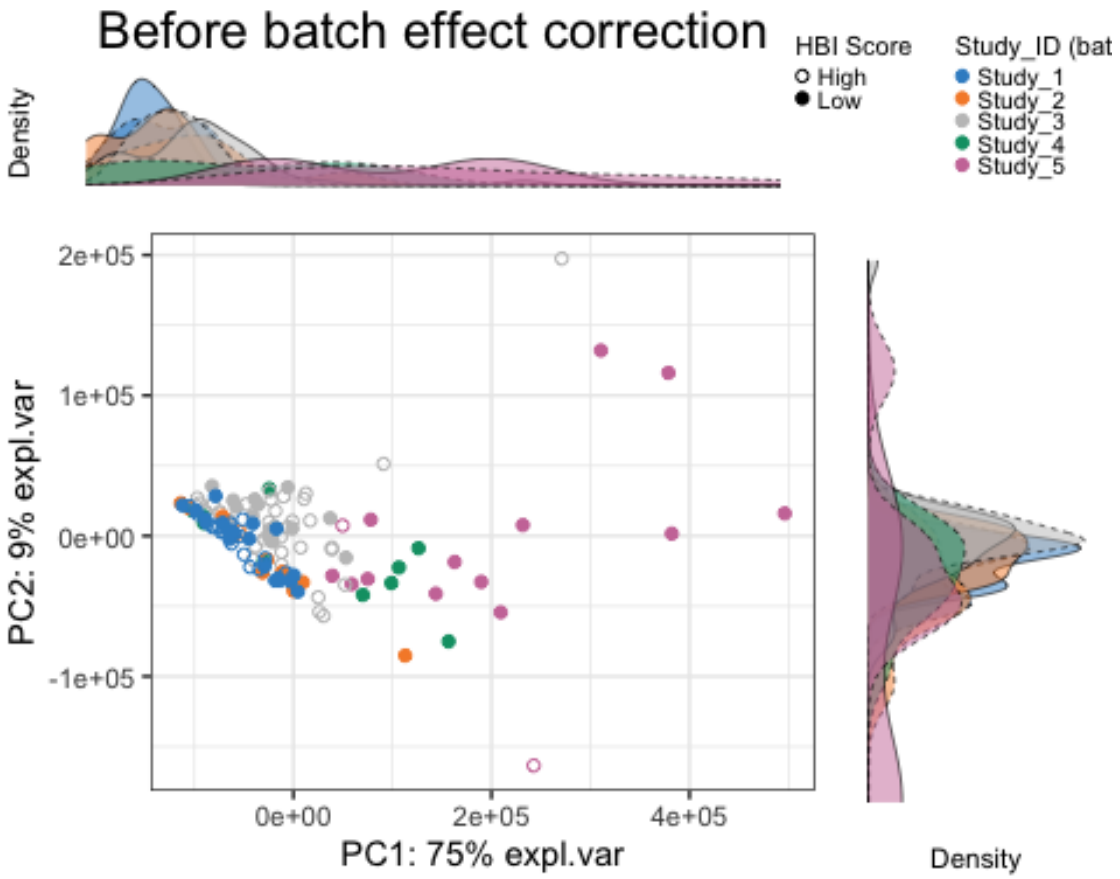
```
before_mat_fun <- as.matrix(t(function_df))
before.dist_fun <- vegdist(before_mat_fun, method='bray')
before.div_fun <- adonis2(before.dist_fun ~ Country + Study_ID + HBI_3Low,
data = crohn_metadata,
                          method='bray')
before.div_fun

## Permutation test for adonis under reduced model
## Terms added sequentially (first to last)
## Permutation: free
## Number of permutations: 999
```

```
##
## adonis2(formula = before.dist_fun ~ Country + Study_ID + HBI_3Low, data =
crohn_metadata, method = "bray")
##           Df SumOfSqs      R2       F Pr(>F)
## Country    1   2.4711 0.14173 22.7130  0.001 ***
## Study_ID   3   1.7379 0.09968  5.3246  0.001 ***
## HBI_3Low   1   0.0614 0.00352  0.5644  0.677
## Residual 121  13.1643 0.75506
## Total    126  17.4346 1.00000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

PCA plot before batch correction

```
pca_before_fun <- pca(t(function_df), ncomp = 3)
pca_plot_before_fun <- Scatter_Density(pca_before_fun, batch =
crohn_metadata$Study_ID,
                trt = crohn_metadata$HBI_3Low,
                batch.legend.title = 'Study_ID (batch)',
                trt.legend.title = 'HBI Score',
                title = 'Before batch effect correction')
```



Batch correction using ComBat-Seq

```
count_mat <- as.matrix(function_df)
batchid = factor(crohn_metadata[, 'Study_ID'])
groupid = factor(crohn_metadata[, 'HBI_3Low'])
adjusted <- ComBat_seq(count_mat, batch=batchid, group=groupid,
full_mod=TRUE)

## Found 5 batches
## Using full model in ComBat-seq.
## Adjusting for 1 covariate(s) or covariate level(s)
## Estimating dispersions
## Fitting the GLM model
## Shrinkage off - using GLM estimates for parameters
## Adjusting the data
```

Normalization of corrected counts

```
norm_colsum <- decostand(adjusted, method = "total", MARGIN = 2)
norm_CPM_count_fun <- norm_colsum * 1000000
```

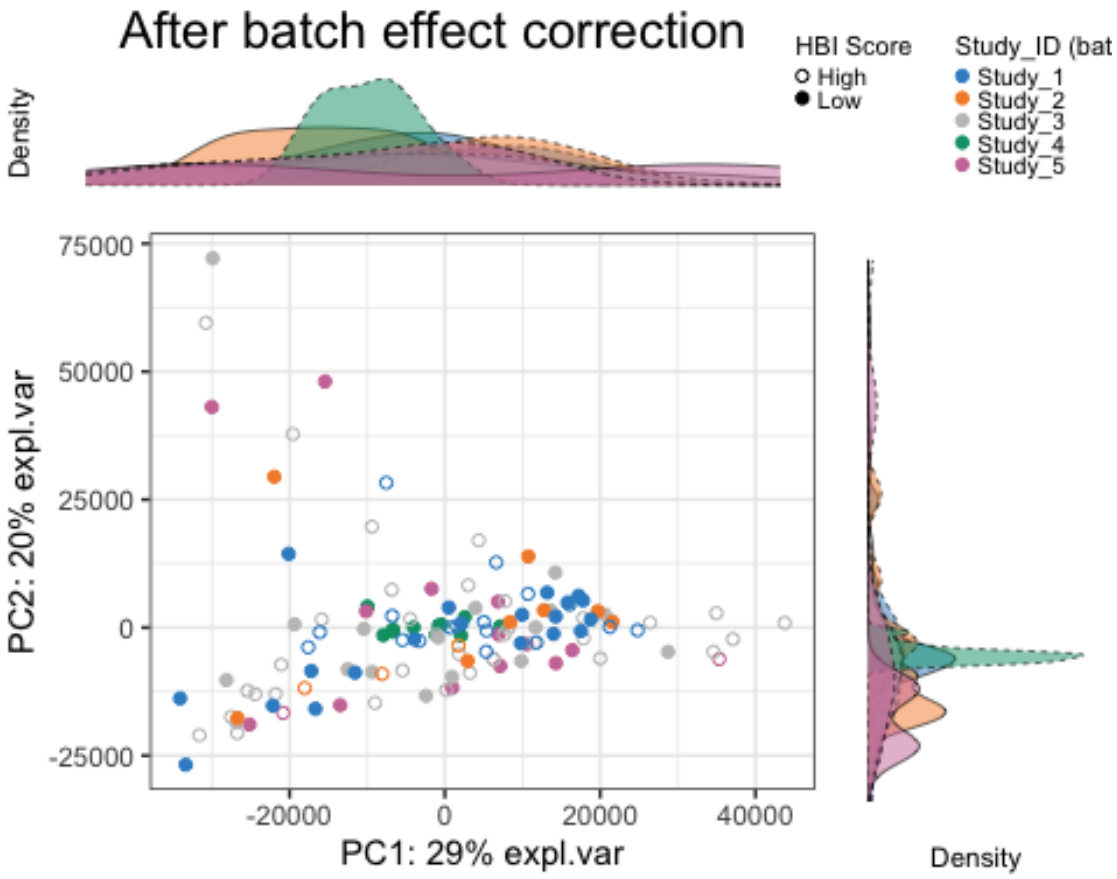
PERMANOVA before batch correction

```
after_mat_fun <- as.matrix(t(norm_CPM_count_fun))
after.dist.fun <- vegdist(after_mat_fun, method='bray')
after.div.fun <- adonis2(after.dist.fun ~ Country + Study_ID + HBI_3Low, data
= crohn_metadata,
                         method='bray')
after.div.fun

## Permutation test for adonis under reduced model
## Terms added sequentially (first to last)
## Permutation: free
## Number of permutations: 999
##
## adonis2(formula = after.dist.fun ~ Country + Study_ID + HBI_3Low, data =
crohn_metadata, method = "bray")
##           Df SumOfSqs      R2      F Pr(>F)
## Country    1   0.0148 0.00218 0.2664  0.997
## Study_ID   3   0.0343 0.00503 0.2053  1.000
## HBI_3Low   1   0.0252 0.00369 0.4517  0.919
## Residual 121   6.7368 0.98910
## Total    126   6.8111 1.00000
```

PCA plot after batch correction

```
pca_after_fun <- pca(t(norm_CPM_count_fun), ncomp = 3)
pca_plot_after_fun <- Scatter_Density(pca_after_fun, batch =
crohn_metadata$Study_ID,
                trt = crohn_metadata$HBI_3Low,
                batch.legend.title = 'Study_ID (batch)',
                trt.legend.title = 'HBI Score',
                title = 'After batch effect correction')
```



Save the plots

```
# pdf(file = "Taxonomy_batch_effect_before_after.pdf", width = 16, height =
6)
#grid.arrange(pca_plot_before_fun, pca_plot_after_fun,
#              ncol = 2)
# dev.off()
```