MBatch 04-05
Using MBatch Assessments: CDP_Structures
Tod Casasent
2017-10-17-1330

# Introduction

These instructions are aimed at people familiar with R and familiar with TCGA/GDC platforms and data types. They are intended to introduce the reader to producing the given assessment. These instructions will only rarely, if ever, touch on the appropriateness of the assessment algorithm or interpretation of output. See MBatch_01_InstallLinux.docx for instructions on downloading test data.

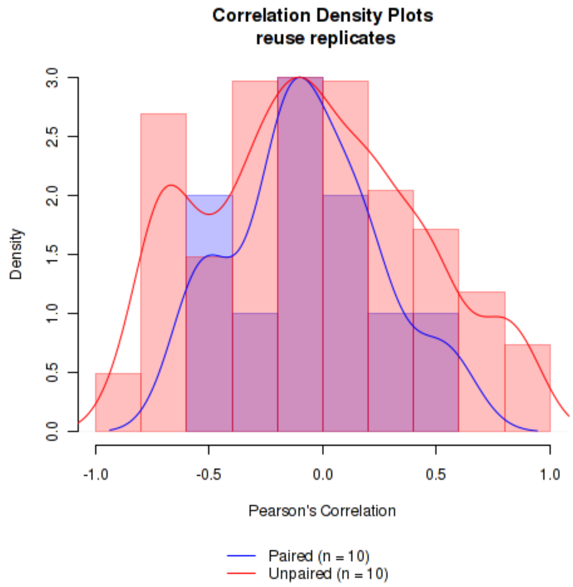
# Algorithm

CDP_Structures is a function used to perform batch effects assessments to create a correlation density plot of original versus corrected data.

# Output

The primary output method for MBatch is to view results in the Batch Effects Website, described elsewhere. The PNG files are rough versions of the website output.

Graphical output is a set overlaid correlation density plots for paired and unpaired values between two data sets.



## Usage

CDP_Structures(theFilePath, theData1, theData2, theSubTitle,

theUnmatchedCount = 1000, theMethod = "pearson",

theUse = "pairwise.complete.obs", theSeed = NULL,

theUseReplicatesUnpaired=FALSE,

theLinePlot=TRUE, theHistPlot=TRUE, theBinWidth=NULL)

## Arguments

**theFilePath** Full path and filename for PNG output file

**theData1** Matrix with columns as samples and rows as features.

**theData2** Matrix with columns as samples and rows as features.

**theSubTitle** Subtitle for image, giving data type being displayed.

**theUnmatchedCount** Number of iterations for unpaired samples.

**theMethod** Defaults to pearson. Valid values are: concordance, pearson, kendall, spearman.

**theUse** Defaults to pairwise.complete.obs. Valid values are accepted by the method parameter to cor.

**theSeed** Default to NULL.

**theUseReplicatesUnpaired** Defaults to FALSE. If TRUE, use both the replicates and non-replicates for the unpaired plot.

**theLinePlot** Default to TRUE. TRUE means plot the lines for Correlation Density Plots.

**theHistPlot** Default to TRUE. TRUE means plot the histogram for Correlation Density Plots.

**theBinWidth** Default to NULL. Non-null means to use the given wide for bins. Otherwise, use default for hist.

# Example Call

The following code performs correlation density plots and is taken from the tests/CDP_Structures.R file. Data used is from the testing data as per the MBatch_01_InstallLinux.docx document.

```
library(MBatch)

# set the paths

theGeneFile1="/bea_testing/MATRIX_DATA/CDP_reuserep_data1.tsv"

theGeneFile2="/bea_testing/MATRIX_DATA/CDP_reuserep_data2.tsv"

theOutputDir="/bea_testing/output/CDP_Structures"

theRandomSeed=314

# make sure the output dir exists and is empty

unlink(theOutputDir, recursive=TRUE)

dir.create(theOutputDir, showWarnings=FALSE, recursive=TRUE)

# load the data and reduce the amount of data to reduce run time

theData1 <- readAsGenericMatrix(theGeneFile1)

theData2 <- readAsGenericMatrix(theGeneFile2)

theUseReplicatesUnpaired <- FALSE

theUnmatchedCount <- 1000

# here, we take most defaults

CDP_Structures(file.path(theOutputDir, "CDP_Plot.png"), theData1, theData2,

theSubTitle="reuse replicates", theMethod="pearson",

theUse="pairwise.complete.obs", theSeed=theRandomSeed,
theLinePlot=TRUE, theHistPlot=TRUE, theBinWidth=NULL,
theUseReplicatesUnpaired=TRUE)
```

## Command Line Output

In the future, we plan to make the output from MBatch more user friendly, but currently, this produces the following output at the command line.

```
> library(MBatch)
>
> # set the paths
> theGeneFile1="/bea_testing/MATRIX_DATA/CDP_reuserep_data1.tsv"
> theGeneFile2="/bea_testing/MATRIX_DATA/CDP_reuserep_data2.tsv"
> theOutputDir="/bea_testing/output/CDP_Structures"
> theRandomSeed=314
>
> # make sure the output dir exists and is empty
> unlink(theOutputDir, recursive=TRUE)
> dir.create(theOutputDir, showWarnings=FALSE, recursive=TRUE)
>
> # load the data and reduce the amount of data to reduce run time
> theData1 <- readAsGenericMatrix(theGeneFile1)
Read 6 records
> theData2 <- readAsGenericMatrix(theGeneFile2)
Read 6 records
>
> theUseReplicatesUnpaired <- FALSE
> theUnmatchedCount <- 1000
>
> # here, we take most defaults
> CDP_Structures(file.path(theOutputDir, "CDP_Plot.png"), theData1, theData2,
+ theSubTitle="reuse replicates", theMethod="pearson",
+ theUse="pairwise.complete.obs", theSeed=theRandomSeed,
+ theLinePlot=TRUE, theHistPlot=TRUE, theBinWidth=NULL,
+ theUseReplicatesUnpaired=TRUE)
2017 10 17 12:36:39.742 INFO megazone23 CDP_Plot theFilePath=
/bea_testing/output/CDP_Structures/CDP_Plot.png
2017 10 17 12:36:39.742 INFO megazone23 CDP_Plot theData1PairedReplicates=
10
2017 10 17 12:36:39.743 INFO megazone23 CDP_Plot theData2PairedReplicates=
10
2017 10 17 12:36:39.743 INFO megazone23 CDP_Plot theData1UnmatchedReplicates=
1000
2017 10 17 12:36:39.744 INFO megazone23 CDP_Plot theData2UnmatchedReplicates=
1000
2017 10 17 12:36:39.782 INFO megazone23 CDP_Plot pairedCorr= 10
2017 10 17 12:36:39.782 INFO megazone23 CDP_Plot unmatchedCorr= 1000
2017 10 17 12:36:39.788 INFO megazone23 CDP_Plot pairedDensity$x= 512
2017 10 17 12:36:39.788 INFO megazone23 CDP_Plot pairedDensity$y= 512
2017 10 17 12:36:39.789 INFO megazone23 CDP_Plot pairedDensity$bw=
0.13641688445874
2017 10 17 12:36:39.789 INFO megazone23 CDP_Plot unmatchedDensity$x=
512
2017 10 17 12:36:39.790 INFO megazone23 CDP_Plot unmatchedDensity$y=
512
2017 10 17 12:36:39.790 INFO megazone23 CDP_Plot unmatchedDensity$bw=
0.104939276458641
```

## Example File Output

The above code creates the following output files. Files are named using the following naming convention:

CDP_Plot.png