MBatch 05-03
Using MBatch Assessments: RBN_Replicates
Tod Casasent
2017-11-10-1455

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

RBN_Replicates is a function used to perform the RBN correction algorithm on two datasets, using replicates. This function takes structures (matrices) and returns a corrected matrix of data.

Output

The primary output method for MBatch is to view results in the Batch Effects Website. Correction algorithms generally do not create graphical output and instead create TSV output files.

Usage

RBN_Replicates(theInvariantMatrix, theVariantMatrix, theInvariantGroupId = "",

theVariantGroupId = "", theMatchedReplicatesFlag = TRUE,

theCombineOnlyFlag = FALSE, thePath = NULL, theWriteToFile = FALSE)

Arguments

theInvariantMatrix Matrix with sample names in colnames and features (like genes) in rownames. This matrix is invariant.

theVariantMatrix Matrix with sample names in colnames and features (like genes) in rownames. This matrix is variant.

theInvariantGroupId Group name used for labelling invariant features when combining matrixes. This defaults to "", but the user should generally provide a value.

theVariantGroupId Group name used for labelling variant features when combining matrixes. This defaults to "", but the user should generally provide a value.

theMatchedReplicatesFlag If TRUE, indicates that NAs should be added for missing replicates. Defaults to TRUE.

theCombineOnlyFlag If TRUE, only combined the matrixes, do not correct. Defaults to FALSE.

thePath Location for output. Defaults to NULL. If NULL, no output file is created.

theWriteToFile TRUE means write corrected data to thePath. Only works if thePath is given. Defaults to FALSE.

Example Call

The following code is taken from the tests/RBN_Replicates.R file. Data used is from the testing data as per the MBatch_01_InstallLinux.docx document.

```
library(MBatch)

# set the paths

invariantFile="/bea_testing/MATRIX_DATA/rbn-test6-iset.tsv"

variantFile="/bea_testing/MATRIX_DATA/rbn-test6-vset.tsv"

theOutputDir="/bea_testing/output/RBN_Replicates"

theRandomSeed=314

resolveDuplicates <- function(theNames)

{

# keep first instance of a name

# number subsequent ones starting with .1

make.unique(theNames)

}

readRPPAdataAsMatrix_WithTab <- function(theFile)

{

# read RPPA data as a dataframe
# column rppaDF[,1] contains row names that may contain duplicates
rppaDF <- readAsGenericDataframe(theFile)
# resolve duplicates in row names here
myRownames <- rppaDF[,1]
myRownames <- resolveDuplicates(myRownames)
# convert to matrix
myMatrix <- data.matrix(rppaDF[,-1])
rownames(myMatrix) <- myRownames
t(myMatrix)
}
readRPPAdataAsMatrix_NoInitialTab <- function(theFile)
{
# read RPPA data as a dataframe
# column rppaDF[,1] contains row names that may contain duplicates
rppaDF <- read.table(theFile, header=TRUE, sep="\t", as.is=TRUE,
check.names=FALSE, stringsAsFactors=FALSE,
colClasses="character", na.strings="NA",
row.names=NULL)
# resolve duplicates in row names here
myRownames <- rppaDF[,1]
myRownames <- resolveDuplicates(myRownames)
# convert to matrix
myMatrix <- data.matrix(rppaDF[,-1])
rownames(myMatrix) <- myRownames
t(myMatrix)
}
# make sure the output dir exists and is empty
unlink(theOutputDir, recursive=TRUE)
dir.create(theOutputDir, showWarnings=FALSE, recursive=TRUE)
message("Reading invariant file")
invMatrix = readRPPAdataAsMatrix_WithTab(invariantFile)
```

```
message("Reading variant file")
varMatrix = readRPPAdataAsMatrix_WithTab(variantFile)
filename <- RBN_Replicates(theInvariantMatrix=invMatrix,
theVariantMatrix=varMatrix,
theInvariantGroupId="Grp1",
theVariantGroupId="Grp2",
theMatchedReplicatesFlag=TRUE,
theCombineOnlyFlag=FALSE,
thePath=theOutputDir,
theWriteToFile=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
> invariantFile="/bea_testing/MATRIX_DATA/rbn-test6-iset.tsv"
> variantFile="/bea_testing/MATRIX_DATA/rbn-test6-vset.tsv"
> theOutputDir="/bea_testing/output/RBN_Replicates"
> theRandomSeed=314
>
> resolveDuplicates <- function(theNames)
+ {
+ # keep first instance of a name
+ # number subsequent ones starting with .1
+ make.unique(theNames)
+ }
>
> readRPPAdataAsMatrix_WithTab <- function(theFile)
+ {
```

```
+ # read RPPA data as a dataframe
+ # column rppaDF[,1] contains row names that may contain duplicates
+ rppaDF <- readAsGenericDataframe(theFile)
+ # resolve duplicates in row names here
+ myRownames <- rppaDF[,1]
+ myRownames <- resolveDuplicates(myRownames)
+ # convert to matrix
+ myMatrix <- data.matrix(rppaDF[,-1])
+ rownames(myMatrix) <- myRownames
+ t(myMatrix)
+ }
>
> readRPPAdataAsMatrix_NoInitialTab <- function(theFile)
+ {
+ # read RPPA data as a dataframe
+ # column rppaDF[,1] contains row names that may contain duplicates
+ rppaDF <- read.table(theFile, header=TRUE, sep="\t", as.is=TRUE,
+ check.names=FALSE, stringsAsFactors=FALSE,
+ colClasses="character", na.strings="NA",
+ row.names=NULL)
+ # resolve duplicates in row names here
+ myRownames <- rppaDF[,1]
+ myRownames <- resolveDuplicates(myRownames)
+ # convert to matrix
+ myMatrix <- data.matrix(rppaDF[,-1])
+ rownames(myMatrix) <- myRownames
+ t(myMatrix)
+ }
>
> # make sure the output dir exists and is empty
> unlink(theOutputDir, recursive=TRUE)
> dir.create(theOutputDir, showWarnings=FALSE, recursive=TRUE)
>
> message("Reading invariant file")
Reading invariant file
> invMatrix = readRPPAdataAsMatrix_WithTab(invariantFile)
> message("Reading variant file")
Reading variant file
> varMatrix = readRPPAdataAsMatrix_WithTab(variantFile)
> filename <- RBN_Replicates(theInvariantMatrix=invMatrix,
+ theVariantMatrix=varMatrix,
+ theInvariantGroupId="Grp1",
+ theVariantGroupId="Grp2",
+ theMatchedReplicatesFlag=TRUE,
+ theCombineOnlyFlag=FALSE,
+ thePath=theOutputDir,
+ theWriteToFile=TRUE)
2017 10 18 12:30:38.107 DEBUG MachineName please note: internally, RBN
processes transposed the data, output (file and matrix) match the submitted
data with samples across columns and features down the rows
2017 10 18 12:30:38.108 INFO MachineName RBN_internal - starting
2017 10 18 12:30:38.108 DEBUG MachineName checkCreateDir: /bea_testing/output/RBN_Replicates
2017 10 18 12:30:38.112 INFO MachineName Found 213 common features in
both matrices.
2017 10 18 12:30:38.505 DEBUG MachineName Write to file /bea_testing/output/RBN_Replicates/ANY_Corre
RBN_Replicates.tsv
2017 10 18 12:30:38.982 DEBUG MachineName Finished write to file
/bea_testing/output/RBN_Replicates/ANY_Corrections-RBN_Replicates.tsv
2017 10 18 12:30:38.982 INFO MachineName RBN_internal - completed
>
```

Example File Output

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

ANY_Corrections-RBN_Replicates.tsv

The TSV file with the combined/corrected dataset is written by the MBatch package.