MBatch 05-04 Using MBatch Assessments: RBN_Pseudoreplicates 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_Pseudoreplicates is a function used to perform the RBN correction algorithm on two datasets, using pseudoreplicates. 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\_Pseudoreplicates (the Invariant Matrix,\ the Variant Matrix,\ the Invariant Replicates,\ the Invariant Group Id="",\ the Variant Group Id="",\ the Variant Group Id="",\ the Matched Replicates Flag=TRUE,\ the Combine Only Flag=FALSE,\ the Path=NULL,\ the Write To File=FALSE)
```

Arguments

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

the VariantMatrix Matrix with sample names in colnames and features (like genes) in rownames. This matrix is variant. the Invariant Replicates Vector of feature ids indicating replicates for variant data

the Variant Replicates Vector of feature ids indicating replicates for invariant data.

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.

the Write To File TRUE means write corrected data to the Path. Only works if the Path is given. Defaults to FALSE.

Example Call

The following code is taken from the tests/RBN_Pseudoreplicates.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-pseudo-iset.tsv"
variantFile="/bea_testing/MATRIX_DATA/rbn-pseudo-vset.tsv"
theOutputDir="/bea_testing/output/RBN_Pseudoreplicates"
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)
}
# 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)
invPseudo <- c("BN", "BO", "BP", "BQ", "BR", "BS")
varPseudo <- c("AN", "AO", "AP", "AQ", "AR", "AS")
filename <- RBN_Pseudoreplicates(theInvariantMatrix=invMatrix,
theVariantMatrix=varMatrix,
theInvariantReplicates = invPseudo,
the Variant Replicates = var Pseudo,
theInvariantGroupId="Grp1",
the Variant Group Id="Grp2",
theMatchedReplicatesFlag=FALSE,
theCombineOnlyFlag=FALSE,
thePath=theOutputDir,
```

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-pseudo-iset.tsv"
> variantFile="/bea_testing/MATRIX_DATA/rbn-pseudo-vset.tsv"
> theOutputDir="/bea_testing/output/RBN_Pseudoreplicates"
> 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)
+ }
> # 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)
> invPseudo <- c("BN", "BO", "BP", "BQ", "BR", "BS")
>var
Pseudo<<br/>- c("AN", "AO", "AP", "AQ", "AR", "AS")
> filename <- RBN Pseudoreplicates(theInvariantMatrix=invMatrix,
+ the Variant Matrix=var Matrix,
+ the Invariant Replicates = inv Pseudo,
+ the Variant Replicates = var Pseudo,
+ theInvariantGroupId="Grp1",
+ the Variant Group Id="Grp2",
+\ the {\bf Matched Replicates Flag} {\bf = FALSE},
+ theCombineOnlyFlag=FALSE,
+ thePath=theOutputDir,
+ theWriteToFile=TRUE)
2017 10 18 12:51:16.693 DEBUG megazone23 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 12:51:16.694 INFO megazone23 RBN_internal - starting
2017 10 18 12:51:16.694 DEBUG megazone23 checkCreateDir: /bea_testing/output/RBN_Pseudoreplicates
2017 10 18 12:51:16.742 WARN megazone 23 Less than 30 replicates provided for
Invariant Matrix. RBN's performance may deteriorate.
```

 $2017\ 10\ 18\ 12:51:16.743$ WARN megazone 23 Less than 30 replicates provided for Variant Matrix. RBN's performance may deteriorate.

 $2017\ 10\ 18\ 12{:}51{:}16.743\ \mathrm{INFO}$ megazone 23 Found 213 common features in both matrices.

 $2017\ 10\ 18\ 12:51:16.763\ DEBUG\ megazone 23\ Write\ to\ file\ / bea_testing/output/RBN_Pseudoreplicates/ANY_CRBN_Pseudoreps.tsv$

2017 10 18 12:51:16.772 DEBUG megazone 23 Finished write to file /bea_testing/output/RBN_Pseudoreplicates/ANY_Corrections-RBN_Pseudoreps.tsv

2017 10 18 12:51:16.773 INFO megazone 23 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_Pseudoreps.tsv

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