Using MBatch Corrections: AN_Unadjusted

Tod Casasent

2018-06-21

1 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 for instructions on downloading test data.

2 Algorithm

AN Adjusted performs an ANOVA Unadjusted correction taking a BEA_DATA object (with data matrix and batch dataframe) and returning either a corrected matrix or a string containing the path to where the data file was written.

3 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.

4 Usage

AN_Unadjusted(theBeaData, theBatchType, thePath = NULL, theWriteToFile = FALSE)

5 Arguments

5.1 theBeaData

BEA_DATA objects can be created by calls of the form new("BEA_DATA", theData, theBatches, theCovariates). If you have no covariate data, use an empty data.frame created with data.frame()

mData: Object of class "matrix" A matrix where the colnames are sample ids and the rownames are gene equivalents. All names should be strings, not factors.

mBatches: Object of class "data.frame" A data.frame where the column "names" are batch types. The first batch "type" is "Sample". All names and values should be strings, not factors or numeric.

mCovariates: Object of class "data.frame" A data.frame where the column "names" are covariate types. The first covariate "type" is "Sample". All names and values should be strings, not factors or numeric.

5.2 theBatchType

A string identifying the batch type to correct.

5.3 thePath

Output path for any files.

5.4 theWriteToFile

TRUE to write the corrected data to file and return the file pathname instead of the corrected matrix.

6 Example Call

The following code is adapted from the tests/AN_Unadjusted.R file. Data used is from the testing data as per the MBatch_01_InstallLinux document. 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

  theGeneFile="/bea_testing/MATRIX_DATA/matrix_data-Tumor.tsv"
  theBatchFile="/bea_testing/MATRIX_DATA/batches-Tumor.tsv"
  theOutputDir="/bea_testing/output/AN_Unadjusted"
  theRandomSeed=314
  theBatchType="TSS"

  # make sure the output dir exists and is empty
  unlink(theOutputDir, recursive=TRUE)
  dir.create(theOutputDir, showWarnings=FALSE, recursive=TRUE)
  # load data
  myData <- mbatchLoadFiles(theGeneFile, theBatchFile)
  myData@mData <- mbatchTrimData(myData@mData, 100000)
  # call
  outputFile <- AN_Unadjusted(theBeaData=myData,
                              theBatchType=theBatchType,
                              thePath=theOutputDir,
                              theWriteToFile=TRUE)
  correctedMatrix <- readAsGenericMatrix(outputFile)
  print(correctedMatrix[1:4, 1:4])
}
```

```
## 2018 06 21 10:35:45.639 DEBUG megazone23 Changing LC_COLLATE to C for duration of run
## 2018 06 21 10:35:45.640 INFO megazone23 \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/
## 2018 06 21 10:35:45.640 INFO megazone23 Starting mbatchLoadFiles
## 2018 06 21 10:35:45.640 INFO megazone23 MBatch Version: 2017-09-19-1530
## 2018 06 21 10:35:45.641 INFO megazone23 read batch file= /bea_testing/MATRIX_DATA/batches-Tumor.tsv
## 2018 06 21 10:35:45.642 INFO megazone23 read gene file= /bea_testing/MATRIX_DATA/matrix_data-Tumor.ts
## 2018 06 21 10:35:55.645 INFO megazone23 filter samples in batches using gene samples
## 2018 06 21 10:35:55.647 INFO megazone23 sort batches by gene file samples
```

```
## 2018 06 21 10:35:55.743 INFO megazone23 Finishing mbatchLoadFiles
## 2018 06 21 10:35:55.746 INFO megazone23 ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
## 2018 06 21 10:35:55.747 DEBUG megazone23 Changing LC_COLLATE to C for duration of run
## 2018 06 21 10:35:55.747 INFO megazone23 \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/
## 2018 06 21 10:35:55.748 INFO megazone23 mbatchTrimData Starting
## 2018 06 21 10:35:55.749 INFO megazone23 MBatch Version: 2017-09-19-1530
## 2018 06 21 10:36:04.502 INFO megazone23 mbatchTrimData Finishing
## 2018 06 21 10:36:04.503 INFO megazone23 ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
## 2018 06 21 10:36:04.503 INFO megazone23 AN_Internal - starting
## 2018 06 21 10:36:04.504 DEBUG megazone23 checkCreateDir:  /bea_testing/output/AN_Unadjusted
## 2018 06 21 10:36:04.761 DEBUG megazone23 starting BeaAN
## 2018 06 21 10:36:04.763 DEBUG megazone23 AN names
## 2018 06 21 10:36:04.763 DEBUG megazone23 convertDataFrameToSi start
## 2018 06 21 10:36:04.764 DEBUG megazone23 convertDataFrameToSi asmatrixWithIssues
## 2018 06 21 10:36:04.765 DEBUG megazone23 convertDataFrameToSi rownames
## 2018 06 21 10:36:04.765 DEBUG megazone23 convertDataFrameToSi colnames
## 2018 06 21 10:36:04.765 DEBUG megazone23 convertDataFrameToSi done
## 2018 06 21 10:36:04.766 DEBUG megazone23 AN all
## 2018 06 21 10:36:04.766 DEBUG megazone23 AN cbin
## 2018 06 21 10:36:04.766 DEBUG megazone23 AN function
## 2018 06 21 10:36:04.766 DEBUG megazone23 AN check number of batch
## 2018 06 21 10:36:04.767 DEBUG megazone23 AN Check for missing values
## 2018 06 21 10:36:04.767 DEBUG megazone23 AN Check for genes with whole batch missing or no variation
## 2018 06 21 10:36:04.914 DEBUG megazone23 AN design
## 2018 06 21 10:36:04.914 DEBUG megazone23 AN build.X
## 2018 06 21 10:36:04.915 DEBUG megazone23 AN NAs
## 2018 06 21 10:36:04.923 DEBUG megazone23 finishing BeaAN
## 2018 06 21 10:36:04.924 TIMING megazone23     0.159999999999854   0.163000000000466   ANUnadjusted
## 2018 06 21 10:36:04.924 DEBUG megazone23 Write to file  /bea_testing/output/AN_Unadjusted/ANY_Correc
## 2018 06 21 10:36:05.026 DEBUG megazone23 Finished write to file  /bea_testing/output/AN_Unadjusted/A
## 2018 06 21 10:36:05.027 INFO megazone23 AN_Internal - completed
##                              TCGA-OR-A5J1-01A-11D-A29J-05
## ABR-cg06968724-17-1012579                     0.02710388
## ABR-cg23568341-17-1011974                     0.10753616
## ABR-cg24479027-17-1012576                     0.02863927
## ACOT7-cg16034168-1-6336711                    1.05951005
##                              TCGA-OR-A5J2-01A-11D-A29J-05
## ABR-cg06968724-17-1012579                     0.02900472
## ABR-cg23568341-17-1011974                     0.11469866
## ABR-cg24479027-17-1012576                     0.03264673
## ACOT7-cg16034168-1-6336711                    0.17891026
##                              TCGA-OR-A5J3-01A-11D-A29J-05
## ABR-cg06968724-17-1012579                      0.8974306
## ABR-cg23568341-17-1011974                      0.9100730
## ABR-cg24479027-17-1012576                      0.9101368
## ACOT7-cg16034168-1-6336711                     0.1812246
##                              TCGA-OR-A5J4-01A-11D-A29J-05
## ABR-cg06968724-17-1012579                      0.9225638
## ABR-cg23568341-17-1011974                      0.9894521
## ABR-cg24479027-17-1012576                      0.9176831
## ACOT7-cg16034168-1-6336711                     1.0226497
```

7 Example File Output

The above code creates the following output file. File is named using the following naming convention: ANY_Corrections-ANUnadjusted.tsv The TSV file with the corrected dataset is written by the MBatch package. The end of the output shows a snippet from the corrected matrix.