Using MBatch Corrections: AN_Unadjusted

Tod Casasent

2019-10-10

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])
}
```

```
## 2019 10 10 11:19:04.425 DEBUG megazone23 Changing LC_COLLATE to C for duration of run
## 2019 10 10 11:19:04.426 INFO megazone23 \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/
## 2019 10 10 11:19:04.426 INFO megazone23 Starting mbatchLoadFiles
## 2019 10 10 11:19:04.426 INFO megazone23 MBatch Version: 2019-09-04-1100
## 2019 10 10 11:19:04.427 INFO megazone23 read batch file= /bea_testing/MATRIX_DATA/batches-Tumor.tsv
## 2019 10 10 11:19:04.442 INFO megazone23 read gene file= /bea_testing/MATRIX_DATA/matrix_data-Tumor.ts
## 2019 10 10 11:19:09.387 INFO megazone23 filter samples in batches using gene samples
## 2019 10 10 11:19:09.391 INFO megazone23 sort batches by gene file samples
```

```
## 2019 10 10 11:19:09.444 INFO megazone23 Finishing mbatchLoadFiles
## 2019 10 10 11:19:09.444 INFO megazone23 ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
## 2019 10 10 11:19:09.445 DEBUG megazone23 Changing LC_COLLATE to C for duration of run
## 2019 10 10 11:19:09.445 INFO megazone23 \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/
## 2019 10 10 11:19:09.445 INFO megazone23 mbatchTrimData Starting
## 2019 10 10 11:19:09.445 INFO megazone23 MBatch Version: 2019-09-04-1100
## 2019 10 10 11:19:16.807 INFO megazone23 mbatchTrimData theMaxSize= 1e+05
## 2019 10 10 11:19:16.807 INFO megazone23 mbatchTrimData ncol(theMatrix)= 80
## 2019 10 10 11:19:16.808 INFO megazone23 mbatchTrimData nrow(theMatrix)= 1250
## 2019 10 10 11:19:16.808 INFO megazone23 mbatchTrimData Finishing
## 2019 10 10 11:19:16.808 INFO megazone23 ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
## 2019 10 10 11:19:16.809 INFO megazone23 AN_Internal - starting
## 2019 10 10 11:19:16.809 DEBUG megazone23 checkCreateDir:  /bea_testing/output/AN_Unadjusted
## 2019 10 10 11:19:17.058 DEBUG megazone23 starting BeaAN
## 2019 10 10 11:19:17.059 DEBUG megazone23 AN names
## 2019 10 10 11:19:17.061 DEBUG megazone23 convertDataFrameToSi start
## 2019 10 10 11:19:17.061 DEBUG megazone23 convertDataFrameToSi asmatrixWithIssues
## 2019 10 10 11:19:17.062 DEBUG megazone23 convertDataFrameToSi rownames
## 2019 10 10 11:19:17.062 DEBUG megazone23 convertDataFrameToSi colnames
## 2019 10 10 11:19:17.062 DEBUG megazone23 convertDataFrameToSi done
## 2019 10 10 11:19:17.063 DEBUG megazone23 AN all
## 2019 10 10 11:19:17.063 DEBUG megazone23 AN cbin
## 2019 10 10 11:19:17.063 DEBUG megazone23 AN function
## 2019 10 10 11:19:17.063 DEBUG megazone23 AN check number of batch
## 2019 10 10 11:19:17.064 DEBUG megazone23 AN Check for missing values
## 2019 10 10 11:19:17.064 DEBUG megazone23 AN Check for genes with whole batch missing or no variation
## 2019 10 10 11:19:17.234 DEBUG megazone23 AN design
## 2019 10 10 11:19:17.234 DEBUG megazone23 AN build.X
## 2019 10 10 11:19:17.235 DEBUG megazone23 AN NAs
## 2019 10 10 11:19:17.243 DEBUG megazone23 finishing BeaAN
## 2019 10 10 11:19:17.243 TIMING megazone23     0.175999999999988   0.186000000000007   ANUnadjusted
## 2019 10 10 11:19:17.244 DEBUG megazone23 Write to file  /bea_testing/output/AN_Unadjusted/ANY_Correc
## 2019 10 10 11:19:17.360 DEBUG megazone23 Finished write to file  /bea_testing/output/AN_Unadjusted/A
## 2019 10 10 11:19:17.360 INFO megazone23 AN_Internal - completed
##                              TCGA-OR-A5J1-01A-11D-A29J-05
## ABR-cg06968724-17-1012579                     0.02710339
## ABR-cg23568341-17-1011974                     0.10753656
## ABR-cg24479027-17-1012576                     0.02863955
## ACOT7-cg16034168-1-6336711                    1.05951016
##                              TCGA-OR-A5J2-01A-11D-A29J-05
## ABR-cg06968724-17-1012579                     0.02900139
## ABR-cg23568341-17-1011974                     0.11469856
## ABR-cg24479027-17-1012576                     0.03264655
## ACOT7-cg16034168-1-6336711                    0.17891016
##                              TCGA-OR-A5J3-01A-11D-A29J-05
## ABR-cg06968724-17-1012579                      0.8974304
## ABR-cg23568341-17-1011974                      0.9100726
## ABR-cg24479027-17-1012576                      0.9101366
## ACOT7-cg16034168-1-6336711                     0.1812252
##                              TCGA-OR-A5J4-01A-11D-A29J-05
## ABR-cg06968724-17-1012579                      0.9225634
## ABR-cg23568341-17-1011974                      0.9894516
## ABR-cg24479027-17-1012576                      0.9176826
## ACOT7-cg16034168-1-6336711                     1.0226502
```

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.