Using MBatch Corrections: AN_Adjusted

Tod Casasent

2023-10-06

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 Adjusted 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_Adjusted(theBeaData, theBatchType, thePath = NULL, theDataVersion=NULL, theTestVersion=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 cleanFilePathname instead of the corrected matrix.

6 Example Call

The following code is adapted from the tests/AN_Adjusted.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.

```
{
  require(MBatch)

  inputDir <- getTestInputDir()
  outputDir <- getTestOutputDir()
  compareDir <- getTestCompareDir()

  # set the paths
  theGeneFile=cleanFilePath(inputDir, "matrix_data-Tumor.tsv")
  theBatchFile=cleanFilePath(inputDir, "batches-Tumor.tsv")
  theOutputDir=cleanFilePath(outputDir, "AN_Adjusted")
  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_Adjusted(theBeaData=myData,
                         theBatchType=theBatchType,
                         thePath=theOutputDir,
                         theDataVersion="DATA_2022-09-09-1600",
                         theTestVersion="TEST_2022-10-10-1300",
                         theWriteToFile=TRUE)
  correctedMatrix <- readAsGenericMatrix(outputFile)
  print(correctedMatrix[1:4, 1:4])
}
```

```
## 2023 10 06 12:33:29.184 DEBUG qcprludev10 Changing LC_COLLATE to C for duration of run
## 2023 10 06 12:33:29.185 INFO qcprludev10 \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/
## 2023 10 06 12:33:29.185 INFO qcprludev10 Starting mbatchLoadFiles
## 2023 10 06 12:33:29.185 INFO qcprludev10 MBatch Version: BEA_VERSION_TIMESTAMP
## 2023 10 06 12:33:29.186 INFO qcprludev10 read batch file= /builds/BatchEffects_clean/BatchEffectsPac
## 2023 10 06 12:33:29.187 INFO qcprludev10 read gene file= /builds/BatchEffects_clean/BatchEffectsPacka
## 2023 10 06 12:33:31.645 INFO qcprludev10 filter samples in batches using gene samples
## 2023 10 06 12:33:31.647 INFO qcprludev10 sort batches by gene file samples
## 2023 10 06 12:33:31.867 INFO qcprludev10 Finishing mbatchLoadFiles
## 2023 10 06 12:33:31.868 INFO qcprludev10 ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
```

```
## 2023 10 06 12:33:31.868 DEBUG qcprludev10 Changing LC_COLLATE to C for duration of run
## 2023 10 06 12:33:31.869 INFO qcprludev10 \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/
## 2023 10 06 12:33:31.869 INFO qcprludev10 mbatchTrimData Starting
## 2023 10 06 12:33:31.870 INFO qcprludev10 MBatch Version: BEA_VERSION_TIMESTAMP
## 2023 10 06 12:33:39.306 INFO qcprludev10 mbatchTrimData theMaxSize= 1e+05
## 2023 10 06 12:33:39.307 INFO qcprludev10 mbatchTrimData ncol(theMatrix)= 80
## 2023 10 06 12:33:39.307 INFO qcprludev10 mbatchTrimData nrow(theMatrix)= 1250
## 2023 10 06 12:33:39.308 INFO qcprludev10 mbatchTrimData Finishing
## 2023 10 06 12:33:39.308 INFO qcprludev10 ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
## 2023 10 06 12:33:39.309 INFO qcprludev10 AN_Internal - starting
## 2023 10 06 12:33:39.513 DEBUG qcprludev10 starting BeaAN
## 2023 10 06 12:33:39.514 DEBUG qcprludev10 AN names
## 2023 10 06 12:33:39.515 DEBUG qcprludev10 convertDataFrameToSi start
## 2023 10 06 12:33:39.515 DEBUG qcprludev10 convertDataFrameToSi asmatrixWithIssues
## 2023 10 06 12:33:39.516 DEBUG qcprludev10 convertDataFrameToSi rownames
## 2023 10 06 12:33:39.516 DEBUG qcprludev10 convertDataFrameToSi colnames
## 2023 10 06 12:33:39.517 DEBUG qcprludev10 convertDataFrameToSi done
## 2023 10 06 12:33:39.517 DEBUG qcprludev10 AN all
## 2023 10 06 12:33:39.517 DEBUG qcprludev10 AN cbin
## 2023 10 06 12:33:39.518 DEBUG qcprludev10 AN function
## 2023 10 06 12:33:39.518 DEBUG qcprludev10 AN check number of batch
## 2023 10 06 12:33:39.518 DEBUG qcprludev10 AN Check for missing values
## 2023 10 06 12:33:39.519 DEBUG qcprludev10 AN Check for genes with whole batch missing or no variatio
## 2023 10 06 12:33:39.647 DEBUG qcprludev10 AN design
## 2023 10 06 12:33:39.647 DEBUG qcprludev10 AN build.X
## 2023 10 06 12:33:39.648 DEBUG qcprludev10 AN NAs
## 2023 10 06 12:33:39.649 INFO qcprludev10 NAs & var.adj
## 2023 10 06 12:33:39.649 INFO qcprludev10 is.matrix dat TRUE
## 2023 10 06 12:33:40.904 INFO qcprludev10 transpose
## 2023 10 06 12:33:40.905 INFO qcprludev10 is.matrix ANdat TRUE
## 2023 10 06 12:33:40.906 INFO qcprludev10 check nulls
## 2023 10 06 12:33:41.119 INFO qcprludev10 sum nulls
## 2023 10 06 12:33:41.119 DEBUG qcprludev10 finishing BeaAN
## 2023 10 06 12:33:41.120 TIMING qcprludev10    5.095   1.608   ANAdjusted      /BEA/BatchEffectsPackag
## 2023 10 06 12:33:41.120 DEBUG qcprludev10 Write to file  /BEA/BatchEffectsPackage_data/testing_dynam
## 2023 10 06 12:33:41.227 DEBUG qcprludev10 Finished write to file  /BEA/BatchEffectsPackage_data/testi
## 2023 10 06 12:33:41.228 INFO qcprludev10 AN_Internal - completed
##                              TCGA-OR-A5J1-01A-11D-A29J-05
## ABR-cg06968724-17-1012579                     0.03374912
## ABR-cg23568341-17-1011974                     0.11227690
## ABR-cg24479027-17-1012576                     0.03534363
## ACOT7-cg16034168-1-6336711                    1.04323618
##                              TCGA-OR-A5J2-01A-11D-A29J-05
## ABR-cg06968724-17-1012579                     0.03561054
## ABR-cg23568341-17-1011974                     0.11935952
## ABR-cg24479027-17-1012576                     0.03927605
## ACOT7-cg16034168-1-6336711                    0.19927863
##                              TCGA-OR-A5J3-01A-11D-A29J-05
## ABR-cg06968724-17-1012579                      0.8873025
## ABR-cg23568341-17-1011974                      0.9059178
## ABR-cg24479027-17-1012576                      0.9004334
## ACOT7-cg16034168-1-6336711                     0.2014973
##                              TCGA-OR-A5J4-01A-11D-A29J-05
## ABR-cg06968724-17-1012579                     0.9119511
```

```
## ABR-cg23568341-17-1011974                    0.9844169
## ABR-cg24479027-17-1012576                    0.9078389
## ACOT7-cg16034168-1-6336711                   1.0079099
```

7 Example File Output

The above code creates the following output file. File is named using the following naming convention: adjusted_matrix.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.