Using MBatch Corrections: AN_Unadjusted

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 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 cleanFilePathname 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.

```
{
  require(MBatch)

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

  theGeneFile=cleanFilePath(inputDir, "matrix_data-Tumor.tsv")
  theBatchFile=cleanFilePath(inputDir, "batches-Tumor.tsv")
  theOutputDir=cleanFilePath(outputDir, "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,
                              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 12:33:43.441 DEBUG qcprludev10 Changing LC_COLLATE to C for duration of run
## 2023 10 06 12:33:43.441 INFO qcprludev10 \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/
## 2023 10 06 12:33:43.442 INFO qcprludev10 Starting mbatchLoadFiles
## 2023 10 06 12:33:43.442 INFO qcprludev10 MBatch Version: BEA_VERSION_TIMESTAMP
## 2023 10 06 12:33:43.442 INFO qcprludev10 read batch file= /builds/BatchEffects_clean/BatchEffectsPacl
## 2023 10 06 12:33:43.444 INFO qcprludev10 read gene file= /builds/BatchEffects_clean/BatchEffectsPacka
## 2023 10 06 12:33:46.000 INFO qcprludev10 filter samples in batches using gene samples
## 2023 10 06 12:33:46.001 INFO qcprludev10 sort batches by gene file samples
## 2023 10 06 12:33:46.072 INFO qcprludev10 Finishing mbatchLoadFiles
## 2023 10 06 12:33:46.073 INFO qcprludev10 ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
## 2023 10 06 12:33:46.073 DEBUG qcprludev10 Changing LC_COLLATE to C for duration of run
```

```
## 2023 10 06 12:33:46.074 INFO qcprludev10 \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/
## 2023 10 06 12:33:46.074 INFO qcprludev10 mbatchTrimData Starting
## 2023 10 06 12:33:46.074 INFO qcprludev10 MBatch Version: BEA_VERSION_TIMESTAMP
## 2023 10 06 12:33:53.619 INFO qcprludev10 mbatchTrimData theMaxSize= 1e+05
## 2023 10 06 12:33:53.620 INFO qcprludev10 mbatchTrimData ncol(theMatrix)= 80
## 2023 10 06 12:33:53.620 INFO qcprludev10 mbatchTrimData nrow(theMatrix)= 1250
## 2023 10 06 12:33:53.620 INFO qcprludev10 mbatchTrimData Finishing
## 2023 10 06 12:33:53.621 INFO qcprludev10 ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
## 2023 10 06 12:33:53.621 INFO qcprludev10 AN_Internal - starting
## 2023 10 06 12:33:53.819 DEBUG qcprludev10 starting BeaAN
## 2023 10 06 12:33:53.821 DEBUG qcprludev10 AN names
## 2023 10 06 12:33:53.821 DEBUG qcprludev10 convertDataFrameToSi start
## 2023 10 06 12:33:53.821 DEBUG qcprludev10 convertDataFrameToSi asmatrixWithIssues
## 2023 10 06 12:33:53.822 DEBUG qcprludev10 convertDataFrameToSi rownames
## 2023 10 06 12:33:53.822 DEBUG qcprludev10 convertDataFrameToSi colnames
## 2023 10 06 12:33:53.823 DEBUG qcprludev10 convertDataFrameToSi done
## 2023 10 06 12:33:53.823 DEBUG qcprludev10 AN all
## 2023 10 06 12:33:53.824 DEBUG qcprludev10 AN cbin
## 2023 10 06 12:33:53.824 DEBUG qcprludev10 AN function
## 2023 10 06 12:33:53.824 DEBUG qcprludev10 AN check number of batch
## 2023 10 06 12:33:53.825 DEBUG qcprludev10 AN Check for missing values
## 2023 10 06 12:33:53.825 DEBUG qcprludev10 AN Check for genes with whole batch missing or no variatio
## 2023 10 06 12:33:53.947 DEBUG qcprludev10 AN design
## 2023 10 06 12:33:53.948 DEBUG qcprludev10 AN build.X
## 2023 10 06 12:33:53.948 DEBUG qcprludev10 AN NAs
## 2023 10 06 12:33:53.957 DEBUG qcprludev10 finishing BeaAN
## 2023 10 06 12:33:53.957 TIMING qcprludev10    0.135999999999996   0.137   ANUnadjusted    /BEA/Batch
## 2023 10 06 12:33:53.958 DEBUG qcprludev10 Write to file  /BEA/BatchEffectsPackage_data/testing_dynam
## 2023 10 06 12:33:54.066 DEBUG qcprludev10 Finished write to file  /BEA/BatchEffectsPackage_data/test
## 2023 10 06 12:33:54.066 INFO qcprludev10 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: 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.