Introduction

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.

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.

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

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

Arguments

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.

theBatchType

A string identifying the batch type to correct.

thePath

Output path for any files.

theWriteToFile

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

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.

```
{
  library(MBatch)

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

  # set the paths
  theGeneFile=file.path(inputDir, "matrix_data-Tumor.tsv")
  theBatchFile=file.path(inputDir, "batches-Tumor.tsv")
  theOutputDir=file.path(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,
                            theWriteToFile=TRUE)
  correctedMatrix <- readAsGenericMatrix(outputFile)
  print(correctedMatrix[1:4, 1:4])
}
```

```
## 2020 11 18 16:20:02.546 DEBUG ab7c64738d52 Changing LC_COLLATE to C for duration of run
## 2020 11 18 16:20:02.547 INFO ab7c64738d52 \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/
## 2020 11 18 16:20:02.547 INFO ab7c64738d52 Starting mbatchLoadFiles
## 2020 11 18 16:20:02.547 INFO ab7c64738d52 MBatch Version: BEA_VERSION_TIMESTAMP
## 2020 11 18 16:20:02.547 INFO ab7c64738d52 read batch file= /builds/BatchEffects_clean/Bat
## 2020 11 18 16:20:02.548 INFO ab7c64738d52 read gene file= /builds/BatchEffects_clean/Batc
## 2020 11 18 16:20:04.972 INFO ab7c64738d52 filter samples in batches using gene samples
## 2020 11 18 16:20:04.973 INFO ab7c64738d52 sort batches by gene file samples
## 2020 11 18 16:20:05.276 INFO ab7c64738d52 Finishing mbatchLoadFiles
## 2020 11 18 16:20:05.276 INFO ab7c64738d52 ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
## 2020 11 18 16:20:05.277 DEBUG ab7c64738d52 Changing LC_COLLATE to C for duration of run
## 2020 11 18 16:20:05.277 INFO ab7c64738d52 \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/
## 2020 11 18 16:20:05.277 INFO ab7c64738d52 mbatchTrimData Starting
## 2020 11 18 16:20:05.277 INFO ab7c64738d52 MBatch Version: BEA_VERSION_TIMESTAMP
## 2020 11 18 16:20:12.748 INFO ab7c64738d52 mbatchTrimData theMaxSize= 1e+05
## 2020 11 18 16:20:12.748 INFO ab7c64738d52 mbatchTrimData ncol(theMatrix)= 80
## 2020 11 18 16:20:12.749 INFO ab7c64738d52 mbatchTrimData nrow(theMatrix)= 1250
## 2020 11 18 16:20:12.749 INFO ab7c64738d52 mbatchTrimData Finishing
## 2020 11 18 16:20:12.749 INFO ab7c64738d52 ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
## 2020 11 18 16:20:12.749 INFO ab7c64738d52 AN_Internal - starting
## 2020 11 18 16:20:12.750 DEBUG ab7c64738d52 checkCreateDir:  /builds/BatchEffects_clean/Ba
## 2020 11 18 16:20:12.991 DEBUG ab7c64738d52 starting BeaAN
## 2020 11 18 16:20:12.991 DEBUG ab7c64738d52 AN names
## 2020 11 18 16:20:12.992 DEBUG ab7c64738d52 convertDataFrameToSi start
## 2020 11 18 16:20:12.992 DEBUG ab7c64738d52 convertDataFrameToSi asmatrixWithIssues
## 2020 11 18 16:20:12.992 DEBUG ab7c64738d52 convertDataFrameToSi rownames
## 2020 11 18 16:20:12.992 DEBUG ab7c64738d52 convertDataFrameToSi colnames
## 2020 11 18 16:20:12.993 DEBUG ab7c64738d52 convertDataFrameToSi done
## 2020 11 18 16:20:12.993 DEBUG ab7c64738d52 AN all
## 2020 11 18 16:20:12.993 DEBUG ab7c64738d52 AN cbin
## 2020 11 18 16:20:12.993 DEBUG ab7c64738d52 AN function
## 2020 11 18 16:20:12.993 DEBUG ab7c64738d52 AN check number of batch
## 2020 11 18 16:20:12.994 DEBUG ab7c64738d52 AN Check for missing values
```

```
## 2020 11 18 16:20:12.994 DEBUG ab7c64738d52 AN Check for genes with whole batch missing or
## 2020 11 18 16:20:13.117 DEBUG ab7c64738d52 AN design
## 2020 11 18 16:20:13.117 DEBUG ab7c64738d52 AN build.X
## 2020 11 18 16:20:13.118 DEBUG ab7c64738d52 AN NAs
## 2020 11 18 16:20:13.118 INFO ab7c64738d52 NAs & var.adj
## 2020 11 18 16:20:13.118 INFO ab7c64738d52 is.matrix dat TRUE
## 2020 11 18 16:20:14.640 INFO ab7c64738d52 transpose
## 2020 11 18 16:20:14.640 INFO ab7c64738d52 is.matrix ANdat TRUE
## 2020 11 18 16:20:14.641 INFO ab7c64738d52 check nulls
## 2020 11 18 16:20:14.833 INFO ab7c64738d52 sum nulls
## 2020 11 18 16:20:14.834 DEBUG ab7c64738d52 finishing BeaAN
## 2020 11 18 16:20:14.834 TIMING ab7c64738d52   1.79299999999999    1.845   ANAdjusted
## 2020 11 18 16:20:14.834 DEBUG ab7c64738d52 Write to file  /builds/BatchEffects_clean/Batc
## 2020 11 18 16:20:14.936 DEBUG ab7c64738d52 Finished write to file  /builds/BatchEffects_c
## 2020 11 18 16:20:14.936 INFO ab7c64738d52 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
```

Example File Output

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