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

EB with non-Parametric Priors performs Empirical Bayes 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

EB_withNonParametricPriors(theBeaData, theBatchIdsNotToCorrect, theDoCheckPlotsFlag, theBatchType, theThreads = 1, 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.

theBatchIdsNotToCorrect

A vector of strings giving batch names/ids within the batch type that should not be corrected

theDoCheckPlotsFlag

Defaults to FALSE. TRUE indicates a prior plots image should be created.

theBatchType

A string identifying the batch type to correct.

theThreads

Integer defaulting to 1. Number of threads to use for calculating priors.

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/EB_withNonParametricPriors.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, "EB_withNonParametricPriors")
  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 <- EB_withNonParametricPriors(theBeaData=myData,
                             theBatchIdsNotToCorrect=c(""),
                             theDoCheckPlotsFlag=TRUE,
                             theBatchType=theBatchType,
                             theThreads=1,
                             thePath=theOutputDir,
                             theWriteToFile=TRUE)
  correctedMatrix <- readAsGenericMatrix(outputFile)
  print(correctedMatrix[1:4, 1:4])
}
```

```
## 2020 11 18 16:18:54.092 DEBUG ab7c64738d52 Changing LC_COLLATE to C for duration of run
## 2020 11 18 16:18:54.092 INFO ab7c64738d52 \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/
## 2020 11 18 16:18:54.092 INFO ab7c64738d52 Starting mbatchLoadFiles
## 2020 11 18 16:18:54.093 INFO ab7c64738d52 MBatch Version: BEA_VERSION_TIMESTAMP
## 2020 11 18 16:18:54.093 INFO ab7c64738d52 read batch file= /builds/BatchEffects_clean/Bat
## 2020 11 18 16:18:54.094 INFO ab7c64738d52 read gene file= /builds/BatchEffects_clean/Batc
## 2020 11 18 16:18:56.199 INFO ab7c64738d52 filter samples in batches using gene samples
## 2020 11 18 16:18:56.200 INFO ab7c64738d52 sort batches by gene file samples
## 2020 11 18 16:18:56.558 INFO ab7c64738d52 Finishing mbatchLoadFiles
## 2020 11 18 16:18:56.559 INFO ab7c64738d52 ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
## 2020 11 18 16:18:56.559 DEBUG ab7c64738d52 Changing LC_COLLATE to C for duration of run
## 2020 11 18 16:18:56.559 INFO ab7c64738d52 \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/
## 2020 11 18 16:18:56.559 INFO ab7c64738d52 mbatchTrimData Starting
## 2020 11 18 16:18:56.560 INFO ab7c64738d52 MBatch Version: BEA_VERSION_TIMESTAMP
## 2020 11 18 16:19:04.024 INFO ab7c64738d52 mbatchTrimData theMaxSize= 1e+05
## 2020 11 18 16:19:04.024 INFO ab7c64738d52 mbatchTrimData ncol(theMatrix)= 80
## 2020 11 18 16:19:04.025 INFO ab7c64738d52 mbatchTrimData nrow(theMatrix)= 1250
## 2020 11 18 16:19:04.025 INFO ab7c64738d52 mbatchTrimData Finishing
## 2020 11 18 16:19:04.025 INFO ab7c64738d52 ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
## 2020 11 18 16:19:04.026 INFO ab7c64738d52 EB_internal - starting
## 2020 11 18 16:19:04.026 DEBUG ab7c64738d52 checkCreateDir:  /builds/BatchEffects_clean/Ba
## 2020 11 18 16:19:04.264 DEBUG ab7c64738d52 starting BeaEB
## 2020 11 18 16:19:04.265 DEBUG ab7c64738d52 EB start
## 2020 11 18 16:19:04.266 DEBUG ab7c64738d52 convertDataFrameToSi start
## 2020 11 18 16:19:04.266 DEBUG ab7c64738d52 convertDataFrameToSi asmatrixWithIssues
## 2020 11 18 16:19:04.267 DEBUG ab7c64738d52 convertDataFrameToSi rownames
## 2020 11 18 16:19:04.267 DEBUG ab7c64738d52 convertDataFrameToSi colnames
## 2020 11 18 16:19:04.267 DEBUG ab7c64738d52 convertDataFrameToSi done
## 2020 11 18 16:19:04.270 DEBUG ab7c64738d52 EB check number of batches
## 2020 11 18 16:19:04.270 DEBUG ab7c64738d52 EB Check for missing values
## 2020 11 18 16:19:04.271 DEBUG ab7c64738d52 Check for genes with whole batch missing or no
## 2020 11 18 16:19:04.416 DEBUG ab7c64738d52 Standardizing Data across genes
## 2020 11 18 16:19:04.477 DEBUG ab7c64738d52 Standarization Model
## 2020 11 18 16:19:04.506 DEBUG ab7c64738d52 stand.mean
## 2020 11 18 16:19:04.508 DEBUG ab7c64738d52 Fitting L/S model and finding priors
## 2020 11 18 16:19:04.508 DEBUG ab7c64738d52 with NAs
## 2020 11 18 16:19:04.632 DEBUG ab7c64738d52 Find priors
## 2020 11 18 16:19:04.634 DEBUG ab7c64738d52 Plot empirical and parametric priors
## 2020 11 18 16:19:04.634 DEBUG ab7c64738d52 Find EB batch adjustments
## 2020 11 18 16:19:04.634 DEBUG ab7c64738d52 Finding nonparametric adjustments
## 2020 11 18 16:19:29.602 DEBUG ab7c64738d52 Adjusting the Data
## 2020 11 18 16:19:29.608 DEBUG ab7c64738d52 add back the removed genes with missing data i
## 2020 11 18 16:19:29.608 DEBUG ab7c64738d52 EB done
## 2020 11 18 16:19:29.608 DEBUG ab7c64738d52 finishing BeaEB
## 2020 11 18 16:19:29.608 TIMING ab7c64738d52   0.441999999999993   25.345      EBwithNonPa
## 2020 11 18 16:19:29.609 DEBUG ab7c64738d52 Write to file  /builds/BatchEffects_clean/Batc
## 2020 11 18 16:19:29.713 DEBUG ab7c64738d52 Finished write to file  /builds/BatchEffects_c
## 2020 11 18 16:19:29.713 INFO ab7c64738d52 EB_internal - completed
##                            TCGA-OR-A5J1-01A-11D-A29J-05
## ABR-cg06968724-17-1012579                    0.02838625
## ABR-cg23568341-17-1011974                    0.03132256
## ABR-cg24479027-17-1012576                    0.03458332
## ACOT7-cg16034168-1-6336711                   0.94492191
##                            TCGA-OR-A5J2-01A-11D-A29J-05
## ABR-cg06968724-17-1012579                      0.03023752
## ABR-cg23568341-17-1011974                      0.03830906
## ABR-cg24479027-17-1012576                      0.03849171
## ACOT7-cg16034168-1-6336711                     0.08633289
##                               TCGA-OR-A5J3-01A-11D-A29J-05
## ABR-cg06968724-17-1012579                      0.87728768
## ABR-cg23568341-17-1011974                      0.81419310
## ABR-cg24479027-17-1012576                      0.89438773
## ACOT7-cg16034168-1-6336711                     0.08859002
##                               TCGA-OR-A5J4-01A-11D-A29J-05
## ABR-cg06968724-17-1012579                       0.9018020
## ABR-cg23568341-17-1011974                       0.8916270
## ABR-cg24479027-17-1012576                       0.9017480
## ACOT7-cg16034168-1-6336711                      0.9089832
```

Example File Output

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