Using MBatch Corrections: EB_withNonParametricPriors

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

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.

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

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

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

5.3 theDoCheckPlotsFlag

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

5.4 theBatchType

A string identifying the batch type to correct.

5.5 theThreads

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

5.6 thePath

Output path for any files.

5.7 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/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)

  # 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/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])
}
```

```
## 2019 10 10 11:17:17.403 DEBUG megazone23 Changing LC_COLLATE to C for duration of run
## 2019 10 10 11:17:17.412 INFO megazone23 \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/
## 2019 10 10 11:17:17.412 INFO megazone23 Starting mbatchLoadFiles
## 2019 10 10 11:17:17.412 INFO megazone23 MBatch Version: 2019-09-04-1100
## 2019 10 10 11:17:17.413 INFO megazone23 read batch file= /bea_testing/MATRIX_DATA/batches-Tumor.tsv
## 2019 10 10 11:17:17.418 INFO megazone23 read gene file= /bea_testing/MATRIX_DATA/matrix_data-Tumor.ts
## 2019 10 10 11:17:22.551 INFO megazone23 filter samples in batches using gene samples
## 2019 10 10 11:17:22.553 INFO megazone23 sort batches by gene file samples
## 2019 10 10 11:17:22.690 INFO megazone23 Finishing mbatchLoadFiles
## 2019 10 10 11:17:22.691 INFO megazone23 ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
## 2019 10 10 11:17:22.691 DEBUG megazone23 Changing LC_COLLATE to C for duration of run
## 2019 10 10 11:17:22.692 INFO megazone23 \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/
## 2019 10 10 11:17:22.692 INFO megazone23 mbatchTrimData Starting
## 2019 10 10 11:17:22.692 INFO megazone23 MBatch Version: 2019-09-04-1100
## 2019 10 10 11:17:31.699 INFO megazone23 mbatchTrimData theMaxSize= 1e+05
## 2019 10 10 11:17:31.699 INFO megazone23 mbatchTrimData ncol(theMatrix)= 80
## 2019 10 10 11:17:31.700 INFO megazone23 mbatchTrimData nrow(theMatrix)= 1250
## 2019 10 10 11:17:31.700 INFO megazone23 mbatchTrimData Finishing
## 2019 10 10 11:17:31.701 INFO megazone23 ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
## 2019 10 10 11:17:31.701 INFO megazone23 EB_internal - starting
## 2019 10 10 11:17:31.702 DEBUG megazone23 checkCreateDir:  /bea_testing/output/EB_withNonParametricPri
## 2019 10 10 11:17:31.936 DEBUG megazone23 starting BeaEB
## 2019 10 10 11:17:31.937 DEBUG megazone23 EB start
## 2019 10 10 11:17:31.938 DEBUG megazone23 convertDataFrameToSi start
## 2019 10 10 11:17:31.938 DEBUG megazone23 convertDataFrameToSi asmatrixWithIssues
## 2019 10 10 11:17:31.938 DEBUG megazone23 convertDataFrameToSi rownames
## 2019 10 10 11:17:31.939 DEBUG megazone23 convertDataFrameToSi colnames
## 2019 10 10 11:17:31.939 DEBUG megazone23 convertDataFrameToSi done
## 2019 10 10 11:17:31.943 DEBUG megazone23 EB check number of batches
## 2019 10 10 11:17:31.943 DEBUG megazone23 EB Check for missing values
## 2019 10 10 11:17:31.944 DEBUG megazone23 Check for genes with whole batch missing or no variation
## 2019 10 10 11:17:32.267 DEBUG megazone23 Standardizing Data across genes
## 2019 10 10 11:17:32.346 DEBUG megazone23 Standarization Model
## 2019 10 10 11:17:32.366 DEBUG megazone23 stand.mean
## 2019 10 10 11:17:32.369 DEBUG megazone23 Fitting L/S model and finding priors
## 2019 10 10 11:17:32.369 DEBUG megazone23 with NAs
## 2019 10 10 11:17:32.547 DEBUG megazone23 Find priors
## 2019 10 10 11:17:32.549 DEBUG megazone23 Plot empirical and parametric priors
## 2019 10 10 11:17:32.549 DEBUG megazone23 Find EB batch adjustments
```

```
## 2019 10 10 11:17:32.550 DEBUG megazone23 Finding nonparametric adjustments
## 2019 10 10 11:18:00.300 DEBUG megazone23 Adjusting the Data
## 2019 10 10 11:18:00.304 DEBUG megazone23 add back the removed genes with missing data in whole batch
## 2019 10 10 11:18:00.304 DEBUG megazone23 EB done
## 2019 10 10 11:18:00.304 DEBUG megazone23 finishing BeaEB
## 2019 10 10 11:18:00.305 TIMING megazone23     0.659999999999997   28.37   EBwithNonParametricPriors
## 2019 10 10 11:18:00.305 DEBUG megazone23 Write to file  /bea_testing/output/EB_withNonParametricPrio
## 2019 10 10 11:18:00.429 DEBUG megazone23 Finished write to file  /bea_testing/output/EB_withNonParame
## 2019 10 10 11:18:00.430 INFO megazone23 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
```

7 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.