Using MBatch Corrections: EB_withNonParametricPriors

Tod Casasent

2018-05-03

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])
}
```

```
## 2018 05 03 13:38:16.113 DEBUG megazone23 Changing LC_COLLATE to C for duration of run
## 2018 05 03 13:38:16.113 INFO megazone23 \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/
## 2018 05 03 13:38:16.114 INFO megazone23 Starting mbatchLoadFiles
## 2018 05 03 13:38:16.114 INFO megazone23 MBatch Version: 2017-09-19-1530
## 2018 05 03 13:38:16.114 INFO megazone23 read batch file= /bea_testing/MATRIX_DATA/batches-Tumor.tsv
## 2018 05 03 13:38:16.115 INFO megazone23 read gene file= /bea_testing/MATRIX_DATA/matrix_data-Tumor.ts
## 2018 05 03 13:38:24.814 INFO megazone23 filter samples in batches using gene samples
## 2018 05 03 13:38:24.815 INFO megazone23 sort batches by gene file samples
## 2018 05 03 13:38:24.961 INFO megazone23 Finishing mbatchLoadFiles
## 2018 05 03 13:38:24.962 INFO megazone23 ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
## 2018 05 03 13:38:24.962 DEBUG megazone23 Changing LC_COLLATE to C for duration of run
## 2018 05 03 13:38:24.962 INFO megazone23 \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/
## 2018 05 03 13:38:24.963 INFO megazone23 mbatchTrimData Starting
## 2018 05 03 13:38:24.963 INFO megazone23 MBatch Version: 2017-09-19-1530
## 2018 05 03 13:38:31.817 INFO megazone23 mbatchTrimData Finishing
## 2018 05 03 13:38:31.817 INFO megazone23 ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
## 2018 05 03 13:38:31.818 INFO megazone23 EB_internal - starting
## 2018 05 03 13:38:31.818 DEBUG megazone23 checkCreateDir:  /bea_testing/output/EB_withNonParametricPr
## 2018 05 03 13:38:32.030 DEBUG megazone23 starting BeaEB
## 2018 05 03 13:38:32.030 DEBUG megazone23 EB start
## 2018 05 03 13:38:32.032 DEBUG megazone23 convertDataFrameToSi start
## 2018 05 03 13:38:32.032 DEBUG megazone23 convertDataFrameToSi asmatrixWithIssues
## 2018 05 03 13:38:32.033 DEBUG megazone23 convertDataFrameToSi rownames
## 2018 05 03 13:38:32.033 DEBUG megazone23 convertDataFrameToSi colnames
## 2018 05 03 13:38:32.033 DEBUG megazone23 convertDataFrameToSi done
## 2018 05 03 13:38:32.036 DEBUG megazone23 EB check number of batches
## 2018 05 03 13:38:32.037 DEBUG megazone23 EB Check for missing values
## 2018 05 03 13:38:32.037 DEBUG megazone23 Check for genes with whole batch missing or no variation
## 2018 05 03 13:38:32.307 DEBUG megazone23 Standardizing Data across genes
## 2018 05 03 13:38:32.384 DEBUG megazone23 Standarization Model
## 2018 05 03 13:38:32.412 DEBUG megazone23 stand.mean
## 2018 05 03 13:38:32.415 DEBUG megazone23 Fitting L/S model and finding priors
## 2018 05 03 13:38:32.415 DEBUG megazone23 with NAs
## 2018 05 03 13:38:32.797 DEBUG megazone23 Find priors
## 2018 05 03 13:38:32.800 DEBUG megazone23 Plot empirical and parametric priors
## 2018 05 03 13:38:32.801 DEBUG megazone23 Find EB batch adjustments
## 2018 05 03 13:38:32.801 DEBUG megazone23 Finding nonparametric adjustments
## 2018 05 03 13:39:00.844 DEBUG megazone23 Adjusting the Data
## 2018 05 03 13:39:00.850 DEBUG megazone23 add back the removed genes with missing data in whole batch
## 2018 05 03 13:39:00.850 DEBUG megazone23 EB done
## 2018 05 03 13:39:00.851 DEBUG megazone23 finishing BeaEB
## 2018 05 03 13:39:00.851 TIMING megazone23     0.864000000000033   28.8220000000001    EBwithNonParam
## 2018 05 03 13:39:00.852 DEBUG megazone23 Write to file  /bea_testing/output/EB_withNonParametricPrior
## 2018 05 03 13:39:00.958 DEBUG megazone23 Finished write to file  /bea_testing/output/EB_withNonParam
## 2018 05 03 13:39:00.959 INFO megazone23 EB_internal - completed
##                            TCGA-OR-A5J1-01A-11D-A29J-05
## ABR-cg06968724-17-1012579                    0.02838688
## ABR-cg23568341-17-1011974                    0.03132255
## ABR-cg24479027-17-1012576                    0.03458342
## ACOT7-cg16034168-1-6336711                   0.94492247
##                            TCGA-OR-A5J2-01A-11D-A29J-05
## ABR-cg06968724-17-1012579                    0.03024093
## ABR-cg23568341-17-1011974                    0.03830954
## ABR-cg24479027-17-1012576                    0.03849226
## ACOT7-cg16034168-1-6336711                   0.08633366
##                            TCGA-OR-A5J3-01A-11D-A29J-05
## ABR-cg06968724-17-1012579                    0.87728806
## ABR-cg23568341-17-1011974                    0.81419391
## ABR-cg24479027-17-1012576                    0.89438833
## ACOT7-cg16034168-1-6336711                   0.08859015
##                            TCGA-OR-A5J4-01A-11D-A29J-05
## ABR-cg06968724-17-1012579                     0.9018025
## ABR-cg23568341-17-1011974                     0.8916279
## ABR-cg24479027-17-1012576                     0.9017489
## ACOT7-cg16034168-1-6336711                    0.9089835
```

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.