Using MBatch Corrections: EB_withParametricPriors

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

EB with 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_withParametricPriors(theBeaData, theBatchIdsNotToCorrect, theDoCheckPlotsFlag, theBatchType, theThreads = 1, thePath = NULL, theDataVersion=NULL, theTestVersion=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 cleanFilePathname 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.

```
{
  require(MBatch)

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

  # set the paths
  theGeneFile=cleanFilePath(inputDir, "matrix_data-Tumor.tsv")
  theBatchFile=cleanFilePath(inputDir, "batches-Tumor.tsv")
  theOutputDir=cleanFilePath(outputDir, "EB_withParametricPriors")
  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_withParametricPriors(theBeaData=myData,
                             theBatchIdsNotToCorrect=c(""),
                             theDoCheckPlotsFlag=TRUE,
                             theBatchType=theBatchType,
                             theThreads=1,
                             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 12:32:49.550 DEBUG qcprludev10 Changing LC_COLLATE to C for duration of run
## 2023 10 06 12:32:49.550 INFO qcprludev10 \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/
## 2023 10 06 12:32:49.551 INFO qcprludev10 Starting mbatchLoadFiles
## 2023 10 06 12:32:49.551 INFO qcprludev10 MBatch Version: BEA_VERSION_TIMESTAMP
## 2023 10 06 12:32:49.551 INFO qcprludev10 read batch file= /builds/BatchEffects_clean/BatchEffectsPac
## 2023 10 06 12:32:49.553 INFO qcprludev10 read gene file= /builds/BatchEffects_clean/BatchEffectsPack
## 2023 10 06 12:32:51.794 INFO qcprludev10 filter samples in batches using gene samples
## 2023 10 06 12:32:51.796 INFO qcprludev10 sort batches by gene file samples
## 2023 10 06 12:32:51.854 INFO qcprludev10 Finishing mbatchLoadFiles
## 2023 10 06 12:32:51.854 INFO qcprludev10 ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
## 2023 10 06 12:32:51.855 DEBUG qcprludev10 Changing LC_COLLATE to C for duration of run
## 2023 10 06 12:32:51.855 INFO qcprludev10 \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/
## 2023 10 06 12:32:51.856 INFO qcprludev10 mbatchTrimData Starting
## 2023 10 06 12:32:51.856 INFO qcprludev10 MBatch Version: BEA_VERSION_TIMESTAMP
## 2023 10 06 12:32:59.330 INFO qcprludev10 mbatchTrimData theMaxSize= 1e+05
## 2023 10 06 12:32:59.331 INFO qcprludev10 mbatchTrimData ncol(theMatrix)= 80
## 2023 10 06 12:32:59.332 INFO qcprludev10 mbatchTrimData nrow(theMatrix)= 1250
## 2023 10 06 12:32:59.332 INFO qcprludev10 mbatchTrimData Finishing
## 2023 10 06 12:32:59.332 INFO qcprludev10 ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
## 2023 10 06 12:32:59.333 INFO qcprludev10 EB_internal - starting
## 2023 10 06 12:32:59.536 DEBUG qcprludev10 starting BeaEB
## 2023 10 06 12:32:59.538 DEBUG qcprludev10 EB start
## 2023 10 06 12:32:59.538 DEBUG qcprludev10 EB theNumberOfThreads= 1
## 2023 10 06 12:32:59.539 DEBUG qcprludev10 convertDataFrameToSi start
## 2023 10 06 12:32:59.539 DEBUG qcprludev10 convertDataFrameToSi asmatrixWithIssues
## 2023 10 06 12:32:59.540 DEBUG qcprludev10 convertDataFrameToSi rownames
## 2023 10 06 12:32:59.540 DEBUG qcprludev10 convertDataFrameToSi colnames
## 2023 10 06 12:32:59.540 DEBUG qcprludev10 convertDataFrameToSi done
## 2023 10 06 12:32:59.544 DEBUG qcprludev10 EB check number of batches
## 2023 10 06 12:32:59.544 DEBUG qcprludev10 EB Check for missing values
## 2023 10 06 12:32:59.545 DEBUG qcprludev10 Check for genes with whole batch missing or no variation
## 2023 10 06 12:32:59.692 DEBUG qcprludev10 Standardizing Data across genes
## 2023 10 06 12:32:59.796 DEBUG qcprludev10 Standarization Model
## 2023 10 06 12:32:59.815 DEBUG qcprludev10 stand.mean
## 2023 10 06 12:32:59.817 DEBUG qcprludev10 Fitting L/S model and finding priors
## 2023 10 06 12:32:59.818 DEBUG qcprludev10 with NAs
## 2023 10 06 12:32:59.965 DEBUG qcprludev10 Find priors
## 2023 10 06 12:32:59.967 DEBUG qcprludev10 Plot empirical and parametric priors
## 2023 10 06 12:32:59.968 DEBUG qcprludev10 Print prior plots at  /BEA/BatchEffectsPackage_data/testing

## 2023 10 06 12:33:00.031 DEBUG qcprludev10 finished prior plots
## 2023 10 06 12:33:00.124 DEBUG qcprludev10 Find EB batch adjustments
## 2023 10 06 12:33:00.124 DEBUG qcprludev10 Finding parametric adjustments
## 2023 10 06 12:33:00.125 DEBUG qcprludev10 Parametric batch num  1  of  5
## 2023 10 06 12:33:00.142 DEBUG qcprludev10 Finding parametric adjustments
## 2023 10 06 12:33:00.143 DEBUG qcprludev10 Parametric batch num  2  of  5
## 2023 10 06 12:33:00.157 DEBUG qcprludev10 Finding parametric adjustments
## 2023 10 06 12:33:00.158 DEBUG qcprludev10 Parametric batch num  3  of  5
## 2023 10 06 12:33:00.175 DEBUG qcprludev10 Finding parametric adjustments
## 2023 10 06 12:33:00.175 DEBUG qcprludev10 Parametric batch num  4  of  5
## 2023 10 06 12:33:00.186 DEBUG qcprludev10 Finding parametric adjustments
## 2023 10 06 12:33:00.187 DEBUG qcprludev10 Parametric batch num  5  of  5
## 2023 10 06 12:33:00.204 DEBUG qcprludev10 Adjusting the Data
## 2023 10 06 12:33:00.206 DEBUG qcprludev10 add back the removed genes with missing data in whole batch
## 2023 10 06 12:33:00.207 DEBUG qcprludev10 EB done
## 2023 10 06 12:33:00.207 DEBUG qcprludev10 finishing BeaEB
## 2023 10 06 12:33:00.208 TIMING qcprludev10    1.12    0.671000000000006   EBwithParametricPriors
## 2023 10 06 12:33:00.208 DEBUG qcprludev10 Write to file  /BEA/BatchEffectsPackage_data/testing_dynami
## 2023 10 06 12:33:00.317 DEBUG qcprludev10 Finished write to file  /BEA/BatchEffectsPackage_data/testi
## 2023 10 06 12:33:00.317 INFO qcprludev10 EB_internal - completed
##                            TCGA-OR-A5J1-01A-11D-A29J-05
## ABR-cg06968724-17-1012579                    0.02849061
## ABR-cg23568341-17-1011974                    0.03145030
## ABR-cg24479027-17-1012576                    0.03469008
## ACOT7-cg16034168-1-6336711                   0.94492167
##                            TCGA-OR-A5J2-01A-11D-A29J-05
## ABR-cg06968724-17-1012579                    0.03034156
## ABR-cg23568341-17-1011974                    0.03843565
## ABR-cg24479027-17-1012576                    0.03859779
## ACOT7-cg16034168-1-6336711                   0.08647875
##                            TCGA-OR-A5J3-01A-11D-A29J-05
## ABR-cg06968724-17-1012579                     0.8772416
## ABR-cg23568341-17-1011974                     0.8141922
## ABR-cg24479027-17-1012576                     0.8943430
## ACOT7-cg16034168-1-6336711                    0.0887355
##                            TCGA-OR-A5J4-01A-11D-A29J-05
## ABR-cg06968724-17-1012579                     0.9017516
## ABR-cg23568341-17-1011974                     0.8916133
## ABR-cg24479027-17-1012576                     0.9017020
## ACOT7-cg16034168-1-6336711                    0.9089891
```

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.