Using MBatch Corrections: EB_withNonParametricPriors

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 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 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)

  cat("directories")
  inputDir <- getTestInputDir()
  outputDir <- getTestOutputDir()
  compareDir <- getTestCompareDir()

  # set the paths
  cat("paths")
  theGeneFile=cleanFilePath(inputDir, "matrix_data-Tumor.tsv")
  theBatchFile=cleanFilePath(inputDir, "batches-Tumor.tsv")
  theOutputDir=cleanFilePath(outputDir, "EB_withNonParametricPriors")
  theRandomSeed=314
  theBatchType="TSS"

  # make sure the output dir exists and is empty
  cat("clean dirs")
  unlink(theOutputDir, recursive=TRUE)
  dir.create(theOutputDir, showWarnings=FALSE, recursive=TRUE)
  # load data
  cat("load data")
  myData <- mbatchLoadFiles(theGeneFile, theBatchFile)
  cat("trim data")
  myData@mData <- mbatchTrimData(myData@mData, 100000)
  # call
  cat("call EB")
  outputFile <- EB_withNonParametricPriors(theBeaData=myData,
                                 theBatchIdsNotToCorrect=c(""),
                             theDoCheckPlotsFlag=TRUE,
                             theBatchType=theBatchType,
                             theThreads=1,
                             thePath=theOutputDir,
                             theWriteToFile=TRUE)
  cat("outputFile")
  cat(outputFile)
  # cat("read corrected")
  # correctedMatrix <- readAsGenericMatrix(outputFile)
  # cat("print corrected")
  # print(correctedMatrix[1:4, 1:4])
}
```

```
## directoriespathsclean dirsload data2023 10 06 12:32:13.425 DEBUG qcprludev10 Changing LC_COLLATE to
## 2023 10 06 12:32:13.426 INFO qcprludev10 \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/
## 2023 10 06 12:32:13.426 INFO qcprludev10 Starting mbatchLoadFiles
## 2023 10 06 12:32:13.427 INFO qcprludev10 MBatch Version: BEA_VERSION_TIMESTAMP
## 2023 10 06 12:32:13.427 INFO qcprludev10 read batch file= /builds/BatchEffects_clean/BatchEffectsPac
## 2023 10 06 12:32:13.428 INFO qcprludev10 read gene file= /builds/BatchEffects_clean/BatchEffectsPack
## 2023 10 06 12:32:15.883 INFO qcprludev10 filter samples in batches using gene samples
## 2023 10 06 12:32:15.885 INFO qcprludev10 sort batches by gene file samples
## 2023 10 06 12:32:15.925 INFO qcprludev10 Finishing mbatchLoadFiles
## 2023 10 06 12:32:15.925 INFO qcprludev10 ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
## trim data2023 10 06 12:32:15.926 DEBUG qcprludev10 Changing LC_COLLATE to C for duration of run
## 2023 10 06 12:32:15.926 INFO qcprludev10 \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/
## 2023 10 06 12:32:15.927 INFO qcprludev10 mbatchTrimData Starting
## 2023 10 06 12:32:15.927 INFO qcprludev10 MBatch Version: BEA_VERSION_TIMESTAMP
## 2023 10 06 12:32:23.342 INFO qcprludev10 mbatchTrimData theMaxSize= 1e+05
## 2023 10 06 12:32:23.343 INFO qcprludev10 mbatchTrimData ncol(theMatrix)= 80
## 2023 10 06 12:32:23.343 INFO qcprludev10 mbatchTrimData nrow(theMatrix)= 1250
## 2023 10 06 12:32:23.343 INFO qcprludev10 mbatchTrimData Finishing
## 2023 10 06 12:32:23.344 INFO qcprludev10 ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
## call EB2023 10 06 12:32:23.344 INFO qcprludev10 EB_internal - starting
## 2023 10 06 12:32:23.550 DEBUG qcprludev10 starting BeaEB
## 2023 10 06 12:32:23.551 DEBUG qcprludev10 EB start
## 2023 10 06 12:32:23.551 DEBUG qcprludev10 EB theNumberOfThreads= 1
## 2023 10 06 12:32:23.552 DEBUG qcprludev10 convertDataFrameToSi start
## 2023 10 06 12:32:23.552 DEBUG qcprludev10 convertDataFrameToSi asmatrixWithIssues
## 2023 10 06 12:32:23.553 DEBUG qcprludev10 convertDataFrameToSi rownames
## 2023 10 06 12:32:23.554 DEBUG qcprludev10 convertDataFrameToSi colnames
## 2023 10 06 12:32:23.554 DEBUG qcprludev10 convertDataFrameToSi done
## 2023 10 06 12:32:23.557 DEBUG qcprludev10 EB check number of batches
## 2023 10 06 12:32:23.558 DEBUG qcprludev10 EB Check for missing values
## 2023 10 06 12:32:23.558 DEBUG qcprludev10 Check for genes with whole batch missing or no variation
## 2023 10 06 12:32:23.701 DEBUG qcprludev10 Standardizing Data across genes
## 2023 10 06 12:32:23.795 DEBUG qcprludev10 Standarization Model
## 2023 10 06 12:32:23.815 DEBUG qcprludev10 stand.mean
## 2023 10 06 12:32:23.817 DEBUG qcprludev10 Fitting L/S model and finding priors
## 2023 10 06 12:32:23.818 DEBUG qcprludev10 with NAs
## 2023 10 06 12:32:23.966 DEBUG qcprludev10 Find priors
## 2023 10 06 12:32:23.968 DEBUG qcprludev10 Plot empirical and parametric priors
## 2023 10 06 12:32:23.969 DEBUG qcprludev10 Find EB batch adjustments
## 2023 10 06 12:32:23.969 DEBUG qcprludev10 Finding nonparametric adjustments
## 2023 10 06 12:32:23.970 DEBUG qcprludev10 Nonparametric batch num  1  of  5
## 2023 10 06 12:32:30.967 DEBUG qcprludev10 Nonparametric batch num  2  of  5
## 2023 10 06 12:32:35.029 DEBUG qcprludev10 Nonparametric batch num  3  of  5
## 2023 10 06 12:32:39.151 DEBUG qcprludev10 Nonparametric batch num  4  of  5
## 2023 10 06 12:32:43.170 DEBUG qcprludev10 Nonparametric batch num  5  of  5
## 2023 10 06 12:32:47.270 DEBUG qcprludev10 Adjusting the Data
## 2023 10 06 12:32:47.293 DEBUG qcprludev10 add back the removed genes with missing data in whole batc
## 2023 10 06 12:32:47.294 DEBUG qcprludev10 EB done
## 2023 10 06 12:32:47.295 DEBUG qcprludev10 finishing BeaEB
## 2023 10 06 12:32:47.295 TIMING qcprludev10    0.932000000000002   23.746      EBwithNonParametricPri
## 2023 10 06 12:32:47.296 DEBUG qcprludev10 Write to file  /BEA/BatchEffectsPackage_data/testing_dynami
## 2023 10 06 12:32:47.405 DEBUG qcprludev10 Finished write to file  /BEA/BatchEffectsPackage_data/testi
## 2023 10 06 12:32:47.406 INFO qcprludev10 EB_internal - completed
## outputFile/BEA/BatchEffectsPackage_data/testing_dynamic/MBatch/EB_withNonParametricPriors/adjusted_ma
```

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.