Using MBatch Corrections: MP_ByBatch

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

MP Overall performs a Median Polish Overall 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

MP_ByBatch(theBeaData, theBatchType, 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 theBatchType

A string identifying the batch type to correct.

5.3 thePath

Output path for any files.

5.4 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/MP_ByBatch.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/MP_ByBatch"
  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 <- MP_ByBatch(theBeaData=myData,
                           theBatchType=theBatchType,
                           thePath=theOutputDir,
                           theWriteToFile=TRUE)
  correctedMatrix <- readAsGenericMatrix(outputFile)
  print(correctedMatrix[1:4, 1:4])
}
```

```
## 2019 10 10 11:18:31.780 DEBUG megazone23 Changing LC_COLLATE to C for duration of run
## 2019 10 10 11:18:31.780 INFO megazone23 \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/
## 2019 10 10 11:18:31.781 INFO megazone23 Starting mbatchLoadFiles
## 2019 10 10 11:18:31.781 INFO megazone23 MBatch Version: 2019-09-04-1100
## 2019 10 10 11:18:31.781 INFO megazone23 read batch file= /bea_testing/MATRIX_DATA/batches-Tumor.tsv
## 2019 10 10 11:18:31.800 INFO megazone23 read gene file= /bea_testing/MATRIX_DATA/matrix_data-Tumor.ts
## 2019 10 10 11:18:37.200 INFO megazone23 filter samples in batches using gene samples
## 2019 10 10 11:18:37.201 INFO megazone23 sort batches by gene file samples
## 2019 10 10 11:18:37.465 INFO megazone23 Finishing mbatchLoadFiles
## 2019 10 10 11:18:37.465 INFO megazone23 ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
## 2019 10 10 11:18:37.465 DEBUG megazone23 Changing LC_COLLATE to C for duration of run
## 2019 10 10 11:18:37.465 INFO megazone23 \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/
## 2019 10 10 11:18:37.466 INFO megazone23 mbatchTrimData Starting
## 2019 10 10 11:18:37.466 INFO megazone23 MBatch Version: 2019-09-04-1100
## 2019 10 10 11:18:44.890 INFO megazone23 mbatchTrimData theMaxSize= 1e+05
## 2019 10 10 11:18:44.890 INFO megazone23 mbatchTrimData ncol(theMatrix)= 80
## 2019 10 10 11:18:44.891 INFO megazone23 mbatchTrimData nrow(theMatrix)= 1250
## 2019 10 10 11:18:44.891 INFO megazone23 mbatchTrimData Finishing
## 2019 10 10 11:18:44.891 INFO megazone23 ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
## 2019 10 10 11:18:44.891 INFO megazone23 MP_Internal - starting
## 2019 10 10 11:18:44.892 DEBUG megazone23 checkCreateDir:  /bea_testing/output/MP_ByBatch
## 2019 10 10 11:18:45.081 DEBUG megazone23 starting BeaMP
## 2019 10 10 11:18:45.081 DEBUG megazone23 starting MP
## 2019 10 10 11:18:45.082 DEBUG megazone23 MP batch
## 2019 10 10 11:18:45.082 DEBUG megazone23 convertDataFrameToSi start
## 2019 10 10 11:18:45.082 DEBUG megazone23 convertDataFrameToSi asmatrixWithIssues
## 2019 10 10 11:18:45.083 DEBUG megazone23 convertDataFrameToSi rownames
## 2019 10 10 11:18:45.083 DEBUG megazone23 convertDataFrameToSi colnames
## 2019 10 10 11:18:45.083 DEBUG megazone23 convertDataFrameToSi done
## 2019 10 10 11:18:46.504 DEBUG megazone23 finishing BeaMP
## 2019 10 10 11:18:46.504 TIMING megazone23     1.36399999999999    1.42400000000001    MPByBatch   /be
## 2019 10 10 11:18:46.505 DEBUG megazone23 Write to file  /bea_testing/output/MP_ByBatch/ANY_Correctio
## 2019 10 10 11:18:46.634 DEBUG megazone23 Finished write to file  /bea_testing/output/MP_ByBatch/ANY_
## 2019 10 10 11:18:46.634 INFO megazone23 MP_Internal - completed
##                            TCGA-OR-A5J1-01A-11D-A29J-05
## ABR-cg06968724-17-1012579                    -0.3519384
## ABR-cg23568341-17-1011974                    -0.3830757
## ABR-cg24479027-17-1012576                    -0.3482736
## ACOT7-cg16034168-1-6336711                    0.4385596
##                            TCGA-OR-A5J2-01A-11D-A29J-05
## ABR-cg06968724-17-1012579                     0.4393411
## ABR-cg23568341-17-1011974                     0.4134679
## ABR-cg24479027-17-1012576                     0.4451149
## ACOT7-cg16034168-1-6336711                    0.3473411
##                            TCGA-OR-A5J3-01A-11D-A29J-05
## ABR-cg06968724-17-1012579                    0.97342791
## ABR-cg23568341-17-1011974                    0.87449968
## ABR-cg24479027-17-1012576                    0.98826271
## ACOT7-cg16034168-1-6336711                   0.01531393
##                            TCGA-OR-A5J4-01A-11D-A29J-05
## ABR-cg06968724-17-1012579                     0.5599392
## ABR-cg23568341-17-1011974                     0.5152570
## ABR-cg24479027-17-1012576                     0.5571871
## ACOT7-cg16034168-1-6336711                    0.4181173
```

7 Example File Output

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