Using MBatch Corrections: MP_Overall

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_Overall(theBeaData, 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 thePath

Output path for any files.

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

```
## 2019 10 10 11:18:16.071 DEBUG megazone23 Changing LC_COLLATE to C for duration of run
## 2019 10 10 11:18:16.073 INFO megazone23 \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/
## 2019 10 10 11:18:16.074 INFO megazone23 Starting mbatchLoadFiles
## 2019 10 10 11:18:16.074 INFO megazone23 MBatch Version: 2019-09-04-1100
## 2019 10 10 11:18:16.075 INFO megazone23 read batch file= /bea_testing/MATRIX_DATA/batches-Tumor.tsv
## 2019 10 10 11:18:16.080 INFO megazone23 read gene file= /bea_testing/MATRIX_DATA/matrix_data-Tumor.ts
## 2019 10 10 11:18:21.339 INFO megazone23 filter samples in batches using gene samples
## 2019 10 10 11:18:21.346 INFO megazone23 sort batches by gene file samples
## 2019 10 10 11:18:21.505 INFO megazone23 Finishing mbatchLoadFiles
## 2019 10 10 11:18:21.506 INFO megazone23 ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
## 2019 10 10 11:18:21.506 DEBUG megazone23 Changing LC_COLLATE to C for duration of run
## 2019 10 10 11:18:21.507 INFO megazone23 \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/
## 2019 10 10 11:18:21.508 INFO megazone23 mbatchTrimData Starting
## 2019 10 10 11:18:21.509 INFO megazone23 MBatch Version: 2019-09-04-1100
```

```
## 2019 10 10 11:18:29.682 INFO megazone23 mbatchTrimData theMaxSize= 1e+05
## 2019 10 10 11:18:29.683 INFO megazone23 mbatchTrimData ncol(theMatrix)= 80
## 2019 10 10 11:18:29.683 INFO megazone23 mbatchTrimData nrow(theMatrix)= 1250
## 2019 10 10 11:18:29.683 INFO megazone23 mbatchTrimData Finishing
## 2019 10 10 11:18:29.683 INFO megazone23 ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
## 2019 10 10 11:18:29.684 INFO megazone23 MP_Internal - starting
## 2019 10 10 11:18:29.684 DEBUG megazone23 checkCreateDir:  /bea_testing/output/MP_Overall
## 2019 10 10 11:18:29.884 DEBUG megazone23 starting BeaMP
## 2019 10 10 11:18:29.884 DEBUG megazone23 starting MP
## 2019 10 10 11:18:29.884 DEBUG megazone23 MP overall
## 2019 10 10 11:18:30.356 DEBUG megazone23 finishing BeaMP
## 2019 10 10 11:18:30.356 TIMING megazone23     0.451999999999998   0.47399999999999    MPOverall   /b
## 2019 10 10 11:18:30.357 DEBUG megazone23 Write to file  /bea_testing/output/MP_Overall/ANY_Correctio
## 2019 10 10 11:18:30.472 DEBUG megazone23 Finished write to file  /bea_testing/output/MP_Overall/ANY_
## 2019 10 10 11:18:30.473 INFO megazone23 MP_Internal - completed
##                              TCGA-OR-A5J1-01A-11D-A29J-05
## ABR-cg06968724-17-1012579                    -0.3515316
## ABR-cg23568341-17-1011974                    -0.3788064
## ABR-cg24479027-17-1012576                    -0.3478635
## ACOT7-cg16034168-1-6336711                    0.4332530
##                              TCGA-OR-A5J2-01A-11D-A29J-05
## ABR-cg06968724-17-1012579                    0.4391718
## ABR-cg23568341-17-1011974                    0.4171611
## ABR-cg24479027-17-1012576                    0.4449490
## ACOT7-cg16034168-1-6336711                   0.3414584
##                              TCGA-OR-A5J3-01A-11D-A29J-05
## ABR-cg06968724-17-1012579                   0.973778288
## ABR-cg23568341-17-1011974                   0.878712547
## ABR-cg24479027-17-1012576                   0.988616482
## ACOT7-cg16034168-1-6336711                  0.009950896
##                              TCGA-OR-A5J4-01A-11D-A29J-05
## ABR-cg06968724-17-1012579                    0.5607078
## ABR-cg23568341-17-1011974                    0.5198881
## ABR-cg24479027-17-1012576                    0.5579590
## ACOT7-cg16034168-1-6336711                   0.4131724
```

7 Example File Output

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