Using MBatch Corrections: MP_Overall

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

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 cleanFilePathname 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.

```
{
  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, "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])
}
```

```
## 2023 10 06 12:33:02.510 DEBUG qcprludev10 Changing LC_COLLATE to C for duration of run
## 2023 10 06 12:33:02.510 INFO qcprludev10 \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/
## 2023 10 06 12:33:02.510 INFO qcprludev10 Starting mbatchLoadFiles
## 2023 10 06 12:33:02.511 INFO qcprludev10 MBatch Version: BEA_VERSION_TIMESTAMP
## 2023 10 06 12:33:02.511 INFO qcprludev10 read batch file= /builds/BatchEffects_clean/BatchEffectsPac
## 2023 10 06 12:33:02.512 INFO qcprludev10 read gene file= /builds/BatchEffects_clean/BatchEffectsPack
## 2023 10 06 12:33:04.957 INFO qcprludev10 filter samples in batches using gene samples
## 2023 10 06 12:33:04.959 INFO qcprludev10 sort batches by gene file samples
## 2023 10 06 12:33:05.017 INFO qcprludev10 Finishing mbatchLoadFiles
## 2023 10 06 12:33:05.018 INFO qcprludev10 ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
## 2023 10 06 12:33:05.018 DEBUG qcprludev10 Changing LC_COLLATE to C for duration of run
## 2023 10 06 12:33:05.019 INFO qcprludev10 \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/
## 2023 10 06 12:33:05.019 INFO qcprludev10 mbatchTrimData Starting
## 2023 10 06 12:33:05.020 INFO qcprludev10 MBatch Version: BEA_VERSION_TIMESTAMP
## 2023 10 06 12:33:12.474 INFO qcprludev10 mbatchTrimData theMaxSize= 1e+05
## 2023 10 06 12:33:12.474 INFO qcprludev10 mbatchTrimData ncol(theMatrix)= 80
## 2023 10 06 12:33:12.475 INFO qcprludev10 mbatchTrimData nrow(theMatrix)= 1250
```

```
## 2023 10 06 12:33:12.475 INFO qcprludev10 mbatchTrimData Finishing
## 2023 10 06 12:33:12.476 INFO qcprludev10 ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
## 2023 10 06 12:33:12.476 INFO qcprludev10 MP_Internal - starting
## 2023 10 06 12:33:12.673 DEBUG qcprludev10 starting BeaMP
## 2023 10 06 12:33:12.674 DEBUG qcprludev10 starting MP
## 2023 10 06 12:33:12.675 DEBUG qcprludev10 MP overall
## 2023 10 06 12:33:13.045 DEBUG qcprludev10 finishing BeaMP
## 2023 10 06 12:33:13.046 TIMING qcprludev10    0.369   0.373000000000005   MPOverall   /BEA/BatchEffe
## 2023 10 06 12:33:13.046 DEBUG qcprludev10 Write to file  /BEA/BatchEffectsPackage_data/testing_dynam
## 2023 10 06 12:33:13.154 DEBUG qcprludev10 Finished write to file  /BEA/BatchEffectsPackage_data/testi
## 2023 10 06 12:33:13.154 INFO qcprludev10 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: 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.