Using MBatch Corrections: MP_ByBatch

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_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 cleanFilePathname 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.

```
{
  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_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,
                           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 12:33:15.351 DEBUG qcprludev10 Changing LC_COLLATE to C for duration of run
## 2023 10 06 12:33:15.351 INFO qcprludev10 \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/
## 2023 10 06 12:33:15.352 INFO qcprludev10 Starting mbatchLoadFiles
## 2023 10 06 12:33:15.352 INFO qcprludev10 MBatch Version: BEA_VERSION_TIMESTAMP
## 2023 10 06 12:33:15.353 INFO qcprludev10 read batch file= /builds/BatchEffects_clean/BatchEffectsPack
## 2023 10 06 12:33:15.354 INFO qcprludev10 read gene file= /builds/BatchEffects_clean/BatchEffectsPacka
## 2023 10 06 12:33:18.071 INFO qcprludev10 filter samples in batches using gene samples
## 2023 10 06 12:33:18.073 INFO qcprludev10 sort batches by gene file samples
## 2023 10 06 12:33:18.137 INFO qcprludev10 Finishing mbatchLoadFiles
## 2023 10 06 12:33:18.138 INFO qcprludev10 ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
```

```
## 2023 10 06 12:33:18.138 DEBUG qcprludev10 Changing LC_COLLATE to C for duration of run
## 2023 10 06 12:33:18.139 INFO qcprludev10 \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/
## 2023 10 06 12:33:18.139 INFO qcprludev10 mbatchTrimData Starting
## 2023 10 06 12:33:18.140 INFO qcprludev10 MBatch Version: BEA_VERSION_TIMESTAMP
## 2023 10 06 12:33:25.579 INFO qcprludev10 mbatchTrimData theMaxSize= 1e+05
## 2023 10 06 12:33:25.579 INFO qcprludev10 mbatchTrimData ncol(theMatrix)= 80
## 2023 10 06 12:33:25.580 INFO qcprludev10 mbatchTrimData nrow(theMatrix)= 1250
## 2023 10 06 12:33:25.580 INFO qcprludev10 mbatchTrimData Finishing
## 2023 10 06 12:33:25.581 INFO qcprludev10 ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
## 2023 10 06 12:33:25.581 INFO qcprludev10 MP_Internal - starting
## 2023 10 06 12:33:25.783 DEBUG qcprludev10 starting BeaMP
## 2023 10 06 12:33:25.784 DEBUG qcprludev10 starting MP
## 2023 10 06 12:33:25.785 DEBUG qcprludev10 MP batch
## 2023 10 06 12:33:25.785 DEBUG qcprludev10 convertDataFrameToSi start
## 2023 10 06 12:33:25.786 DEBUG qcprludev10 convertDataFrameToSi asmatrixWithIssues
## 2023 10 06 12:33:25.786 DEBUG qcprludev10 convertDataFrameToSi rownames
## 2023 10 06 12:33:25.786 DEBUG qcprludev10 convertDataFrameToSi colnames
## 2023 10 06 12:33:25.787 DEBUG qcprludev10 convertDataFrameToSi done
## 2023 10 06 12:33:26.904 DEBUG qcprludev10 finishing BeaMP
## 2023 10 06 12:33:26.905 TIMING qcprludev10    1.113   1.12200000000001    MPByBatch   /BEA/BatchEffe
## 2023 10 06 12:33:26.905 DEBUG qcprludev10 Write to file  /BEA/BatchEffectsPackage_data/testing_dynam
## 2023 10 06 12:33:27.013 DEBUG qcprludev10 Finished write to file  /BEA/BatchEffectsPackage_data/test
## 2023 10 06 12:33:27.013 INFO qcprludev10 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: 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.