Introduction

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.

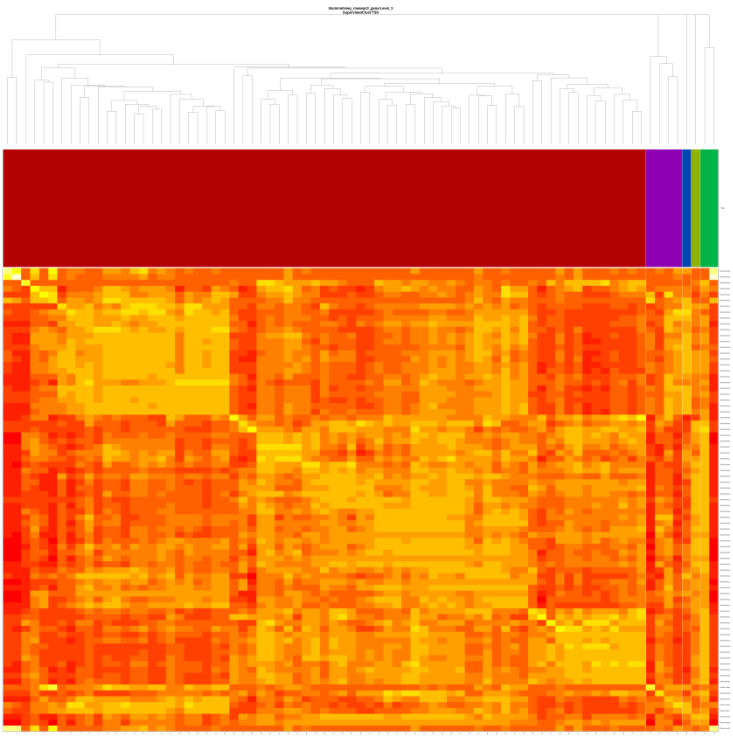
Algorithm

SupervisedClustering_Pairs_Structures is a function used to perform batch effects assessments using the supervised clustering algorithm for each pair of batch types provided.

Output

The primary output method for MBatch is to view results in the Batch Effects Website, described elsewhere. The PNG files are rough versions of the website output.

Graphical output is a heatmap of the correlation values, topped by a covariate bar with the batch information, and at the top dendrograms for the clustering. The columns are batch values for a single batch type. The rows are sample ids.



Usage

SupervisedClustering_Pairs_Structures(theData, theTitle, theOutputPath, theDoHeatmapFlag, theListOfBatchPairs, theBatchTypeAndValuePairsToRemove=list(), theBatchTypeAndValuePairsToKeep=list())

Arguments

##theData An instance of BEA_DATA.

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.

##theTitle A string title to use in PNG files.

##theOutputPath String giving directory in which to place output PNG files.

##theDoHeatmapFlag
A flag indicating whether or not to create HC heatmap, where TRUE meants to create heatmap.

##theListOfBatchPairs
A vector of strings, where pairs of strings give batch types to use for pairs assessment.

##theBatchTypeAndValuePairsToRemove A list of vectors containing the batch type (or * for all types) and the value to remove. list() indicates none while NULL will cause an error.

##theBatchTypeAndValuePairsToKeep A list of vectors containing the batch type (or * for all types) and a vector of the the value(s) to keep. list() indicates none while NULL will cause an error.

Example Call

The following code is adapted from the tests/SupervisedClustering_Pairs_Structures 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.

This output can generally be skipped as very long and generally obscure. After the output is an explanation of files and directories created.

```
{
  library(MBatch)

  inputDir <- getTestInputDir()
  outputDir <- getTestOutputDir()
  compareDir <- getTestCompareDir()

  # set the paths
  theGeneFile=file.path(inputDir, "matrix_data-Tumor.tsv")
  theBatchFile=file.path(inputDir, "batches-Tumor.tsv")
  theOutputDir=file.path(outputDir, "SupervisedClustering_Pairs_Structures")
  theRandomSeed=314

  # make sure the output dir exists and is empty
  unlink(theOutputDir, recursive=TRUE)
  dir.create(theOutputDir, showWarnings=FALSE, recursive=TRUE)
```

```
  # load the data and reduce the amount of data to reduce run time
  myData <- mbatchLoadFiles(theGeneFile, theBatchFile)
  myData@mData <- mbatchTrimData(myData@mData, 100000)

  # here, we take most defaults
  SupervisedClustering_Pairs_Structures(theData=myData,
    theTitle="Test Data Title",
    theOutputPath=theOutputDir,
    theDoHeatmapFlag=TRUE,
    theListOfBatchPairs=c("PlateId", "TSS", "BatchId", "TSS"),
    theBatchTypeAndValuePairsToRemove=list(),
    theBatchTypeAndValuePairsToKeep=list() )
}
```

```
## 2020 11 18 16:18:17.106 DEBUG ab7c64738d52 Changing LC_COLLATE to C for duration of run
## 2020 11 18 16:18:17.107 INFO ab7c64738d52 \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/
## 2020 11 18 16:18:17.107 INFO ab7c64738d52 Starting mbatchLoadFiles
## 2020 11 18 16:18:17.107 INFO ab7c64738d52 MBatch Version: BEA_VERSION_TIMESTAMP
## 2020 11 18 16:18:17.107 INFO ab7c64738d52 read batch file= /builds/BatchEffects_clean/Bat
## 2020 11 18 16:18:17.108 INFO ab7c64738d52 read gene file= /builds/BatchEffects_clean/Batc
## 2020 11 18 16:18:19.559 INFO ab7c64738d52 filter samples in batches using gene samples
## 2020 11 18 16:18:19.560 INFO ab7c64738d52 sort batches by gene file samples
## 2020 11 18 16:18:19.622 INFO ab7c64738d52 Finishing mbatchLoadFiles
## 2020 11 18 16:18:19.622 INFO ab7c64738d52 ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
## 2020 11 18 16:18:19.622 DEBUG ab7c64738d52 Changing LC_COLLATE to C for duration of run
## 2020 11 18 16:18:19.623 INFO ab7c64738d52 \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/
## 2020 11 18 16:18:19.623 INFO ab7c64738d52 mbatchTrimData Starting
## 2020 11 18 16:18:19.623 INFO ab7c64738d52 MBatch Version: BEA_VERSION_TIMESTAMP
## 2020 11 18 16:18:27.028 INFO ab7c64738d52 mbatchTrimData theMaxSize= 1e+05
## 2020 11 18 16:18:27.028 INFO ab7c64738d52 mbatchTrimData ncol(theMatrix)= 80
## 2020 11 18 16:18:27.028 INFO ab7c64738d52 mbatchTrimData nrow(theMatrix)= 1250
## 2020 11 18 16:18:27.028 INFO ab7c64738d52 mbatchTrimData Finishing
## 2020 11 18 16:18:27.028 INFO ab7c64738d52 ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
## 2020 11 18 16:18:27.029 DEBUG ab7c64738d52 Changing LC_COLLATE to C for duration of run
## 2020 11 18 16:18:27.030 INFO ab7c64738d52 \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/
## 2020 11 18 16:18:27.030 INFO ab7c64738d52 mbatchFilterData Starting
## 2020 11 18 16:18:27.030 INFO ab7c64738d52 MBatch Version: BEA_VERSION_TIMESTAMP
## 2020 11 18 16:18:27.030 DEBUG ab7c64738d52 rows pre filter  1250
## 2020 11 18 16:18:27.234 DEBUG ab7c64738d52 rows post filter  1250
## 2020 11 18 16:18:27.234 DEBUG ab7c64738d52 mbatchFilterData Prefilter, gene data had  125
## 2020 11 18 16:18:27.235 DEBUG ab7c64738d52 mbatchFilterData Prefilter, batch data had  80
## 2020 11 18 16:18:27.235 INFO ab7c64738d52 mbatchFilterData Finishing
## 2020 11 18 16:18:27.235 INFO ab7c64738d52 ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
## 2020 11 18 16:18:27.236 INFO ab7c64738d52 createBatchEffectsOutput_SupervisedClustering_p
## 2020 11 18 16:18:27.236 INFO ab7c64738d52 createBatchEffectsOutput_SupervisedClustering_p
```

```
## 2020 11 18 16:18:27.236 DEBUG ab7c64738d52 checkCreateDir:  /builds/BatchEffects_clean/Ba
## 2020 11 18 16:18:27.237 INFO ab7c64738d52 makeBiasClust - starting
## 2020 11 18 16:18:27.317 INFO ab7c64738d52 makeBiasClust - quantile dat dim =  1250,80
## 2020 11 18 16:18:27.319 INFO ab7c64738d52 makeBiasClust - quantile U.data is.data.frame =
## 2020 11 18 16:18:27.319 INFO ab7c64738d52 makeBiasClust - quantile U.data is.array =  TRU
## 2020 11 18 16:18:27.319 INFO ab7c64738d52 makeBiasClust - quantile U.data is.list =  FALS
## 2020 11 18 16:18:27.319 INFO ab7c64738d52 makeBiasClust - quantile U.data nrow =  312
## 2020 11 18 16:18:27.319 INFO ab7c64738d52 makeBiasClust - quantile U.data ncol =  80
## 2020 11 18 16:18:27.320 INFO ab7c64738d52 makeBiasClust - quantile U.data length =  24960
## 2020 11 18 16:18:27.320 INFO ab7c64738d52 makeBiasClust - quantile U.data dim =  312,80
## 2020 11 18 16:18:27.320 INFO ab7c64738d52 makeBiasClust - quantile U.data is.null =  FALS
## 2020 11 18 16:18:27.320 INFO ab7c64738d52 makeBiasClust - data frame
## 2020 11 18 16:18:27.321 INFO ab7c64738d52 makeBiasClust - U.dend1 <- bias.clust
## 2020 11 18 16:18:27.324 INFO ab7c64738d52 makeBiasClust new.dis size -  80-80
## 2020 11 18 16:18:27.325 INFO ab7c64738d52 makeBiasClust orig -  80-80
## 2020 11 18 16:18:27.325 INFO ab7c64738d52 makeBiasClust is.na -  80-80
## 2020 11 18 16:18:27.325 INFO ab7c64738d52 makeBiasClust is.infinite -  80-80

## 2020 11 18 16:18:29.580 DEBUG ab7c64738d52 mbatchStandardLegend - Calling .jinit  /tmp/Rt
## 2020 11 18 16:18:29.581 DEBUG ab7c64738d52 mbatchStandardLegend - .jinit complete
## 2020 11 18 16:18:29.582 DEBUG ab7c64738d52 mbatchStandardLegend - theTitle  PlateId
## 2020 11 18 16:18:29.582 DEBUG ab7c64738d52 mbatchStandardLegend - theVersion  MBatch 1.7.
## 2020 11 18 16:18:29.582 DEBUG ab7c64738d52 mbatchStandardLegend - theFilenamePath  /build
## 2020 11 18 16:18:29.582 DEBUG ab7c64738d52 mbatchStandardLegend - theLegendNames  A29J (8
## 2020 11 18 16:18:29.582 DEBUG ab7c64738d52 mbatchStandardLegend - theLegendNames  1
## 2020 11 18 16:18:29.583 DEBUG ab7c64738d52 mbatchStandardLegend - theLegendColors  1
## 2020 11 18 16:18:29.583 DEBUG ab7c64738d52 mbatchStandardLegend - theLegendSymbols  0
## 2020 11 18 16:18:29.583 DEBUG ab7c64738d52 mbatchStandardLegend - myColors  #b30000
## 2020 11 18 16:18:29.583 DEBUG ab7c64738d52 mbatchStandardLegend before java
## 2020 11 18 16:18:29.598 DEBUG ab7c64738d52 mbatchStandardLegend after java
## 2020 11 18 16:18:29.599 DEBUG ab7c64738d52 mbatchStandardLegend - Calling .jinit  /tmp/Rt
## 2020 11 18 16:18:29.601 DEBUG ab7c64738d52 mbatchStandardLegend - .jinit complete
## 2020 11 18 16:18:29.601 DEBUG ab7c64738d52 mbatchStandardLegend - theTitle  TSS
## 2020 11 18 16:18:29.601 DEBUG ab7c64738d52 mbatchStandardLegend - theVersion  MBatch 1.7.
## 2020 11 18 16:18:29.601 DEBUG ab7c64738d52 mbatchStandardLegend - theFilenamePath  /build
## 2020 11 18 16:18:29.602 DEBUG ab7c64738d52 mbatchStandardLegend - theLegendNames  OR - Un
## (72), OU - Roswell Park (1), P6 - Translational Genomics
## Research Institute (2), PA - University of Minnesota
## (1), PK - University Health
## Network (4)
## 2020 11 18 16:18:29.602 DEBUG ab7c64738d52 mbatchStandardLegend - theLegendNames  5
## 2020 11 18 16:18:29.602 DEBUG ab7c64738d52 mbatchStandardLegend - theLegendColors  5
## 2020 11 18 16:18:29.602 DEBUG ab7c64738d52 mbatchStandardLegend - theLegendSymbols  0
## 2020 11 18 16:18:29.602 DEBUG ab7c64738d52 mbatchStandardLegend - myColors  #b30000,#8fb3
## 2020 11 18 16:18:29.603 DEBUG ab7c64738d52 mbatchStandardLegend before java
## 2020 11 18 16:18:29.661 DEBUG ab7c64738d52 mbatchStandardLegend after java
```

```
## 2020 11 18 16:18:29.661 INFO ab7c64738d52 createBatchEffectsOutput_SupervisedClustering_p
## 2020 11 18 16:18:29.661 INFO ab7c64738d52 createBatchEffectsOutput_SupervisedClustering_p
## 2020 11 18 16:18:29.661 DEBUG ab7c64738d52 checkCreateDir:  /builds/BatchEffects_clean/Ba
## 2020 11 18 16:18:29.662 INFO ab7c64738d52 makeBiasClust - starting
## 2020 11 18 16:18:29.740 INFO ab7c64738d52 makeBiasClust - quantile dat dim =  1250,80
## 2020 11 18 16:18:29.741 INFO ab7c64738d52 makeBiasClust - quantile U.data is.data.frame =
## 2020 11 18 16:18:29.742 INFO ab7c64738d52 makeBiasClust - quantile U.data is.array =  TRU
## 2020 11 18 16:18:29.742 INFO ab7c64738d52 makeBiasClust - quantile U.data is.list =  FALS
## 2020 11 18 16:18:29.742 INFO ab7c64738d52 makeBiasClust - quantile U.data nrow =  312
## 2020 11 18 16:18:29.742 INFO ab7c64738d52 makeBiasClust - quantile U.data ncol =  80
## 2020 11 18 16:18:29.742 INFO ab7c64738d52 makeBiasClust - quantile U.data length =  24960
## 2020 11 18 16:18:29.743 INFO ab7c64738d52 makeBiasClust - quantile U.data dim =  312,80
## 2020 11 18 16:18:29.743 INFO ab7c64738d52 makeBiasClust - quantile U.data is.null =  FALS
## 2020 11 18 16:18:29.743 INFO ab7c64738d52 makeBiasClust - data frame
## 2020 11 18 16:18:29.743 INFO ab7c64738d52 makeBiasClust - U.dend1 <- bias.clust
## 2020 11 18 16:18:29.746 INFO ab7c64738d52 makeBiasClust new.dis size -  80-80
## 2020 11 18 16:18:29.746 INFO ab7c64738d52 makeBiasClust orig -  80-80
## 2020 11 18 16:18:29.747 INFO ab7c64738d52 makeBiasClust is.na -  80-80
## 2020 11 18 16:18:29.747 INFO ab7c64738d52 makeBiasClust is.infinite -  80-80

## 2020 11 18 16:18:31.982 DEBUG ab7c64738d52 mbatchStandardLegend - Calling .jinit  /tmp/Rt
## 2020 11 18 16:18:31.984 DEBUG ab7c64738d52 mbatchStandardLegend - .jinit complete
## 2020 11 18 16:18:31.984 DEBUG ab7c64738d52 mbatchStandardLegend - theTitle  BatchId
## 2020 11 18 16:18:31.984 DEBUG ab7c64738d52 mbatchStandardLegend - theVersion  MBatch 1.7.
## 2020 11 18 16:18:31.985 DEBUG ab7c64738d52 mbatchStandardLegend - theFilenamePath  /build
## 2020 11 18 16:18:31.985 DEBUG ab7c64738d52 mbatchStandardLegend - theLegendNames  00304 (
## 2020 11 18 16:18:31.985 DEBUG ab7c64738d52 mbatchStandardLegend - theLegendNames  1
## 2020 11 18 16:18:31.985 DEBUG ab7c64738d52 mbatchStandardLegend - theLegendColors  1
## 2020 11 18 16:18:31.985 DEBUG ab7c64738d52 mbatchStandardLegend - theLegendSymbols  0
## 2020 11 18 16:18:31.986 DEBUG ab7c64738d52 mbatchStandardLegend - myColors  #b30000
## 2020 11 18 16:18:31.986 DEBUG ab7c64738d52 mbatchStandardLegend before java
## 2020 11 18 16:18:32.000 DEBUG ab7c64738d52 mbatchStandardLegend after java
## 2020 11 18 16:18:32.002 DEBUG ab7c64738d52 mbatchStandardLegend - Calling .jinit  /tmp/Rt
## 2020 11 18 16:18:32.003 DEBUG ab7c64738d52 mbatchStandardLegend - .jinit complete
## 2020 11 18 16:18:32.004 DEBUG ab7c64738d52 mbatchStandardLegend - theTitle  TSS
## 2020 11 18 16:18:32.004 DEBUG ab7c64738d52 mbatchStandardLegend - theVersion  MBatch 1.7.
## 2020 11 18 16:18:32.004 DEBUG ab7c64738d52 mbatchStandardLegend - theFilenamePath  /build
## 2020 11 18 16:18:32.004 DEBUG ab7c64738d52 mbatchStandardLegend - theLegendNames  OR - Un
## (72), OU - Roswell Park (1), P6 - Translational Genomics
## Research Institute (2), PA - University of Minnesota
## (1), PK - University Health
## Network (4)
## 2020 11 18 16:18:32.004 DEBUG ab7c64738d52 mbatchStandardLegend - theLegendNames  5
## 2020 11 18 16:18:32.005 DEBUG ab7c64738d52 mbatchStandardLegend - theLegendColors  5
## 2020 11 18 16:18:32.005 DEBUG ab7c64738d52 mbatchStandardLegend - theLegendSymbols  0
## 2020 11 18 16:18:32.005 DEBUG ab7c64738d52 mbatchStandardLegend - myColors  #b30000,#8fb3
```

```
## 2020 11 18 16:18:32.005 DEBUG ab7c64738d52 mbatchStandardLegend before java
## 2020 11 18 16:18:32.064 DEBUG ab7c64738d52 mbatchStandardLegend after java
```

Example File Output

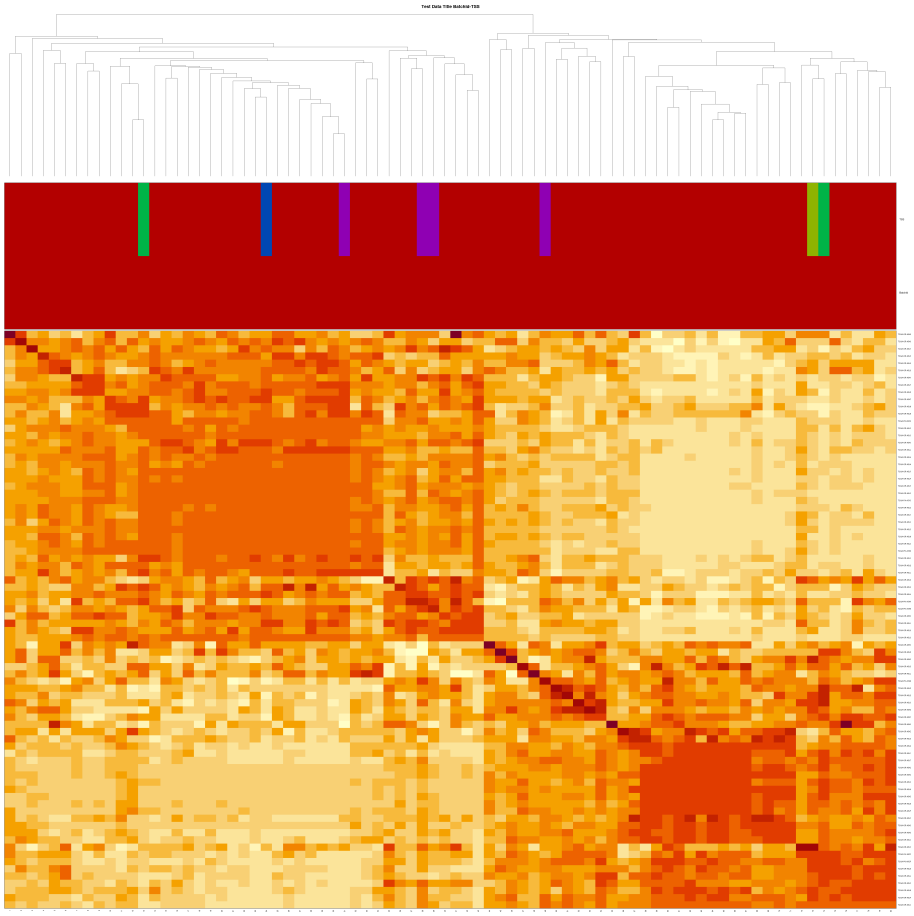
The above code creates the following subdirectories and files. The subdirectories correspond to the Batch Type Pairs on which assessments were requested.

```
/output/SupervisedClustering_Pairs_Structures$ ls -l
total 8
drwxr-xr-x 2 linux linux 4096 Jun 14 12:56 BatchId-TSS
drwxr-xr-x 2 linux linux 4096 Jun 14 12:56 PlateId-TSS
```

Looking at the "BatchId-TSS" subdirectory, it contains the following diagram and legend files. This algorithm does not currently generate data usable with dynamic displays.

```
/output/SupervisedClustering_Pairs_Structures/BatchId-TSS$ ls -l
total 276
-rw-r--r-- 1 linux linux 261168 Jun 19 09:58 SupervisedClust_Diagram.png
-rw-r--r-- 1 linux linux   2701 Jun 19 09:58 SupervisedClust_Legend-BatchId.png
-rw-r--r-- 1 linux linux  12899 Jun 19 09:58 SupervisedClust_Legend-TSS.png
```

Here is the diagram generated from this code.


Test Data Title Matched TSS
TSS
Batch