MBatch 04-01
Using MBatch Assessments: SupervisedClustering_Batches_Structures
Tod Casasent
2017-10-17-1330

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.docx for instructions on downloading test data.

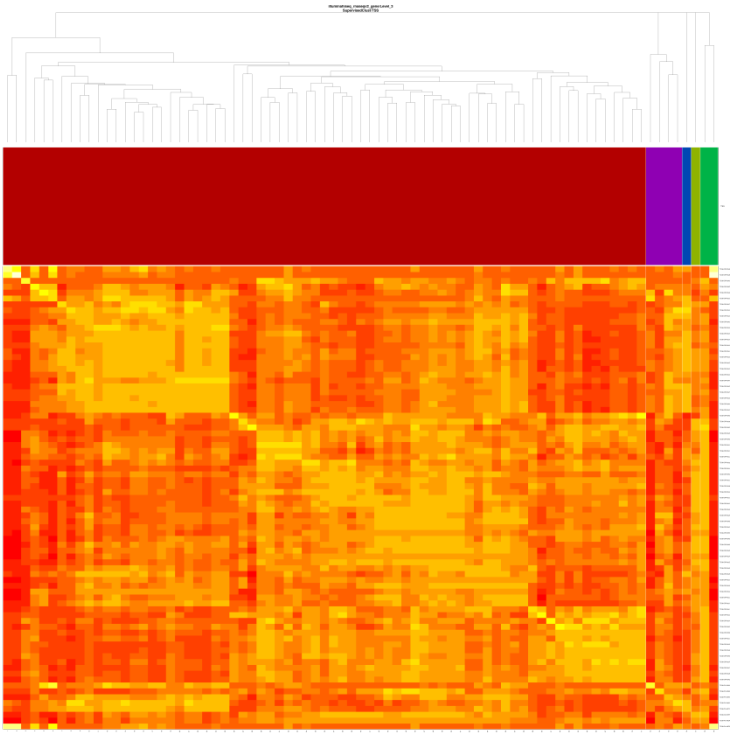
Algorithm

SupervisedClustering_Batches_Structures is a function used to perform batch effects assessments using the supervised clustering algorithm for each batch type available by default.

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_Batches_Structures(theData, theTitle, theOutputPath,

theDoHeatmapFlag,

theBatchTypeAndValuePairsToRemove = list(),

theBatchTypeAndValuePairsToKeep = list())

Arguments

theData An instance of BEA_DATA BEA_DATA-class. This is the MBatch Data Object (of class BEA_DATA) described in MBatch_03_UserData.docx, and returned from mbatchLoadFiles or mbatchLoadStructures.

theTitle Object of class "character". Title to use in PNG files.

theOutputPath Object of class "character". Directory in which to place output PNG files and related data files used by the Batch Effects Website.

theDoHeatmapFlag Object of class "logical" A flag indicating whether or not to create HC heatmap, where TRUE means to create heatmap.

theBatchTypeAndValuePairsToRemove Object of class "list". A list of vectors containing the batch type (or * for all types) and the value to remove. list() indicates do nothing, while NULL will cause an error. This type of list is described in MBatch_03_ParametersBatchTypesValues.docx.

theBatchTypeAndValuePairsToKeep Object of class "list". A list of vectors containing the batch type (or * for all types) and a vector of the value(s) to keep. list() indicates do nothing, while NULL will cause an error. This type of list is described in MBatch_03_ParametersBatchTypesValues.docx.

Example Call

The following code performs Supervised Clustering and is taken from the tests/SupervisedClustering_Batches_Structures.R file. Data used is from the testing data as per the MBatch_01_InstallLinux.docx document.

```
library(MBatch)

# set the paths

theGeneFile <- "/bea_testing/MATRIX_DATA/matrix_data-Tumor.tsv"

theBatchFile <- "/bea_testing/MATRIX_DATA/batches-Tumor.tsv"

theOutputDir <- "/bea_testing/output/SupervisedClustering_Batches_Structures"

# 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_Batches_Structures(theData=myData,

theTitle="Test Supervised Clustering",

theOutputPath=theOutputDir,

theDoHeatmapFlag=TRUE)
```

Command Line Output

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.

```
> # set the paths
```

> theGeneFile <- "/bea_testing/MATRIX_DATA/matrix_data-Tumor.tsv"

> theBatchFile <- "/bea_testing/MATRIX_DATA/batches-Tumor.tsv"

> theOutputDir <- "/bea_testing/output/SupervisedClustering_Batches_Structures"

>

> # 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)

2017 10 03 09:16:04.998 DEBUG MachineName Changing LC_COLLATE to C
for duration of run

2017 10 03 09:16:04.998 INFO MachineName \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/
\/ \/

2017 10 03 09:16:04.999 INFO MachineName Starting mbatchLoadFiles

2017 10 03 09:16:04.999 INFO MachineName MBatch Version: 2017-09-19-1530

2017 10 03 09:16:05.000 INFO MachineName read batch file= /bea_testing/MATRIX_DATA/batches-
Tumor.tsv

2017 10 03 09:16:05.028 INFO MachineName read gene file= /bea_testing/MATRIX_DATA/matrix_data-
Tumor.tsv

Read 100000 records

2017 10 03 09:16:18.321 INFO MachineName filter samples in batches using gene
samples

2017 10 03 09:16:18.323 INFO MachineName sort batches by gene file samples

2017 10 03 09:16:18.472 INFO MachineName Finishing mbatchLoadFiles

2017 10 03 09:16:18.474 INFO MachineName ^^

> myData@mData <- mbatchTrimData(myData@mData, 100000)

2017 10 03 09:16:18.511 DEBUG MachineName Changing LC_COLLATE to C
for duration of run

2017 10 03 09:16:18.511 INFO MachineName \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/
\/ \/

2017 10 03 09:16:18.512 INFO MachineName mbatchTrimData Starting

2017 10 03 09:16:18.512 INFO MachineName MBatch Version: 2017-09-19-1530

2017 10 03 09:16:26.799 INFO MachineName mbatchTrimData Finishing

2017 10 03 09:16:26.800 INFO MachineName ^^

>

> # here, we take most defaults

> SupervisedClustering_Batches_Structures(theData=myData,

+ theTitle="Test Supervised Clustering",

+ theOutputPath=theOutputDir,

+ theDoHeatmapFlag=TRUE)

2017 10 03 09:16:26.804 DEBUG MachineName Changing LC_COLLATE to C for duration of run

2017 10 03 09:16:26.804 INFO MachineName \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/

2017 10 03 09:16:26.805 INFO MachineName mbatchFilterData Starting

2017 10 03 09:16:26.805 INFO MachineName MBatch Version: 2017-09-19-1530

2017 10 03 09:16:26.806 DEBUG MachineName rows pre filter 1250

2017 10 03 09:16:27.127 DEBUG MachineName rows post filter 1250

2017 10 03 09:16:27.128 DEBUG MachineName mbatchFilterData Prefilter, gene data had 1250 while post filter 1250

2017 10 03 09:16:27.129 DEBUG MachineName mbatchFilterData Prefilter, batch data had 80 while post filter 80

2017 10 03 09:16:27.129 INFO MachineName mbatchFilterData Finishing

2017 10 03 09:16:27.130 INFO MachineName ^^

2017 10 03 09:16:27.130 DEBUG MachineName checkCreateDir: /bea_testing/output/SupervisedClustering_Bat

2017 10 03 09:16:27.131 INFO MachineName createBatchEffectsOutput_SupervisedClustering_batches - batchTypeName = BatchId

2017 10 03 09:16:27.133 INFO MachineName makeBiasClust - starting

2017 10 03 09:16:27.322 INFO MachineName makeBiasClust - quantile dat dim = 1250,80

2017 10 03 09:16:27.327 INFO MachineName makeBiasClust - quantile U.data is.data.frame = FALSE

2017 10 03 09:16:27.328 INFO MachineName makeBiasClust - quantile U.data is.array = TRUE

2017 10 03 09:16:27.328 INFO MachineName makeBiasClust - quantile U.data is.list = FALSE

2017 10 03 09:16:27.329 INFO MachineName makeBiasClust - quantile U.data nrow = 312

2017 10 03 09:16:27.334 INFO MachineName makeBiasClust - quantile U.data ncol = 80

2017 10 03 09:16:27.334 INFO MachineName makeBiasClust - quantile U.data length = 24960

2017 10 03 09:16:27.334 INFO MachineName makeBiasClust - quantile U.data dim = 312,80

2017 10 03 09:16:27.335 INFO MachineName makeBiasClust - quantile U.data is.null = FALSE

2017 10 03 09:16:27.335 INFO MachineName makeBiasClust - data frame

2017 10 03 09:16:27.336 INFO MachineName makeBiasClust - U.dend1 <- bias.clust

2017 10 03 09:16:27.349 INFO MachineName makeBiasClust new.dis size - 80-80

2017 10 03 09:16:27.354 INFO MachineName makeBiasClust orig - 80-80

2017 10 03 09:16:27.355 INFO MachineName makeBiasClust is.na - 80-80

2017 10 03 09:16:27.355 INFO MachineName makeBiasClust is.infinite - 80-80

2017 10 03 09:16:29.770 DEBUG MachineName mbatchStandardLegend - Calling .jinit /home/linux/R/x86_64-pc-linux-gnu-library/3.4/MBatch/LegendJava/LegendJava.jar:/home/linux/R/x86 pc-linux-gnu-library/3.4/MBatch/LegendJava/jcommon-1.0.17.jar:/home/linux/R/x86_64-pc-linux-gnu-library/3.4/MBatch/LegendJava/jfreechart-1.0.14.jar

2017 10 03 09:16:31.308 DEBUG MachineName mbatchStandardLegend - .jinit complete

2017 10 03 09:16:31.309 DEBUG MachineName mbatchStandardLegend - theTitle BatchId

2017 10 03 09:16:31.309 DEBUG MachineName mbatchStandardLegend - theVersion MBatch 1.4.16

2017 10 03 09:16:31.310 DEBUG MachineName mbatchStandardLegend - theFilenamePath /bea_testing/output/SupervisedClustering_Batches_Structures/Batches/SupervisedClust_Legend-BatchId.png

2017 10 03 09:16:31.311 DEBUG MachineName mbatchStandardLegend - theLegendNames 00304 (80)

2017 10 03 09:16:31.311 DEBUG MachineName mbatchStandardLegend - theLegendNames 1

2017 10 03 09:16:31.312 DEBUG MachineName mbatchStandardLegend - theLegendColors 1

2017 10 03 09:16:31.312 DEBUG MachineName mbatchStandardLegend - theLegendSymbols 0

2017 10 03 09:16:31.313 DEBUG MachineName mbatchStandardLegend - myColors #b30000

2017 10 03 09:16:31.314 DEBUG MachineName mbatchStandardLegend before java

LegendJava 2013_05_03_0823

writeLegendWithSymbols theTitle = BatchId

writeLegendWithSymbols theVersion = MBatch 1.4.16

writeLegendWithSymbols theFilenamePath = /bea_testing/output/SupervisedClustering_Batches_Structures/
BatchId.png

Colors is non-null

writeLegendWithSymbols write

writeLegendWithSymbols done

2017 10 03 09:16:37.321 DEBUG MachineName mbatchStandardLegend after java

2017 10 03 09:16:37.322 INFO MachineName createBatchEffectsOutput_SupervisedClustering_batches - batchTypeName = PlateId

2017 10 03 09:16:37.423 INFO MachineName makeBiasClust - starting

2017 10 03 09:16:37.630 INFO MachineName makeBiasClust - quantile dat dim = 1250,80

2017 10 03 09:16:37.631 INFO MachineName makeBiasClust - quantile U.data is.data.frame = FALSE

2017 10 03 09:16:37.632 INFO MachineName makeBiasClust - quantile U.data is.array = TRUE

2017 10 03 09:16:37.632 INFO MachineName makeBiasClust - quantile U.data is.list = FALSE

2017 10 03 09:16:37.633 INFO MachineName makeBiasClust - quantile U.data nrow = 312

2017 10 03 09:16:37.633 INFO MachineName makeBiasClust - quantile U.data ncol = 80

2017 10 03 09:16:37.633 INFO MachineName makeBiasClust - quantile U.data length = 24960

2017 10 03 09:16:37.634 INFO MachineName makeBiasClust - quantile U.data dim = 312,80

2017 10 03 09:16:37.634 INFO MachineName makeBiasClust - quantile U.data is.null = FALSE

2017 10 03 09:16:37.635 INFO MachineName makeBiasClust - data frame

2017 10 03 09:16:37.635 INFO MachineName makeBiasClust - U.dend1 <- bias.clust

2017 10 03 09:16:37.678 INFO MachineName makeBiasClust new.dis size - 80-80

2017 10 03 09:16:37.679 INFO MachineName makeBiasClust orig - 80-80

2017 10 03 09:16:37.680 INFO MachineName makeBiasClust is.na - 80-80

2017 10 03 09:16:37.680 INFO MachineName makeBiasClust is.infinite - 80-80

2017 10 03 09:16:40.407 DEBUG MachineName mbatchStandardLegend - Calling .jinit /home/linux/R/x86_64-pc-linux-gnu-library/3.4/MBatch/LegendJava/LegendJava.jar:/home/linux/R/x86 pc-linux-gnu-library/3.4/MBatch/LegendJava/jcommon-1.0.17.jar:/home/linux/R/x86_64-pc-linux-gnu-library/3.4/MBatch/LegendJava/jfreechart-1.0.14.jar

2017 10 03 09:16:40.528 DEBUG MachineName mbatchStandardLegend - .jinit complete

2017 10 03 09:16:40.528 DEBUG MachineName mbatchStandardLegend - theTitle PlateId

2017 10 03 09:16:40.531 DEBUG MachineName mbatchStandardLegend - theVersion MBatch 1.4.16

2017 10 03 09:16:40.531 DEBUG MachineName mbatchStandardLegend - theFilenamePath /bea_testing/output/SupervisedClustering_Batches_Structures/Batches/SupervisedClust_Legend-PlateId.png

2017 10 03 09:16:40.531 DEBUG MachineName mbatchStandardLegend - theLegendNames A29J (80)

2017 10 03 09:16:40.532 DEBUG MachineName mbatchStandardLegend - theLegendNames 1

2017 10 03 09:16:40.532 DEBUG MachineName mbatchStandardLegend - theLegendColors 1

2017 10 03 09:16:40.532 DEBUG MachineName mbatchStandardLegend - theLegendSymbols 0

2017 10 03 09:16:40.533 DEBUG MachineName mbatchStandardLegend - myColors #b30000

2017 10 03 09:16:40.533 DEBUG MachineName mbatchStandardLegend before java

LegendJava 2013_05_03_0823

writeLegendWithSymbols theTitle = PlateId

writeLegendWithSymbols theVersion = MBatch 1.4.16

writeLegendWithSymbols theFilenamePath = /bea_testing/output/SupervisedClustering_Batches_Structures/
PlateId.png

Colors is non-null

writeLegendWithSymbols write

writeLegendWithSymbols done

2017 10 03 09:16:40.592 DEBUG MachineName mbatchStandardLegend after java

2017 10 03 09:16:40.602 INFO MachineName createBatchEffectsOutput_SupervisedClustering_batches - batchTypeName = ShipDate

2017 10 03 09:16:40.603 INFO MachineName makeBiasClust - starting

2017 10 03 09:16:40.863 INFO MachineName makeBiasClust - quantile dat dim = 1250,80

2017 10 03 09:16:40.865 INFO MachineName makeBiasClust - quantile U.data is.data.frame = FALSE

2017 10 03 09:16:40.865 INFO MachineName makeBiasClust - quantile U.data is.array = TRUE

2017 10 03 09:16:40.866 INFO MachineName makeBiasClust - quantile U.data is.list = FALSE

2017 10 03 09:16:40.866 INFO MachineName makeBiasClust - quantile U.data nrow = 312

2017 10 03 09:16:40.866 INFO MachineName makeBiasClust - quantile U.data ncol = 80

2017 10 03 09:16:40.867 INFO MachineName makeBiasClust - quantile U.data length = 24960

2017 10 03 09:16:40.867 INFO MachineName makeBiasClust - quantile U.data dim = 312,80

2017 10 03 09:16:40.867 INFO MachineName makeBiasClust - quantile U.data is.null = FALSE

2017 10 03 09:16:40.868 INFO MachineName makeBiasClust - data frame

2017 10 03 09:16:40.868 INFO MachineName makeBiasClust - U.dend1 <- bias.clust

2017 10 03 09:16:40.872 INFO MachineName makeBiasClust new.dis size - 80-80

2017 10 03 09:16:40.872 INFO MachineName makeBiasClust orig - 80-80

2017 10 03 09:16:40.873 INFO MachineName makeBiasClust is.na - 80-80

2017 10 03 09:16:40.873 INFO MachineName makeBiasClust is.infinite - 80-80

2017 10 03 09:16:43.291 DEBUG MachineName mbatchStandardLegend - Calling .jinit /home/linux/R/x86_64-pc-linux-gnu-library/3.4/MBatch/LegendJava/LegendJava.jar:/home/linux/R/x86 pc-linux-gnu-library/3.4/MBatch/LegendJava/jcommon-1.0.17.jar:/home/linux/R/x86_64-pc-linux-gnu-library/3.4/MBatch/LegendJava/jfreechart-1.0.14.jar

2017 10 03 09:16:43.298 DEBUG MachineName mbatchStandardLegend - .jinit complete

2017 10 03 09:16:43.298 DEBUG MachineName mbatchStandardLegend - theTitle ShipDate

2017 10 03 09:16:43.298 DEBUG MachineName mbatchStandardLegend - theVersion MBatch 1.4.16

2017 10 03 09:16:43.299 DEBUG MachineName mbatchStandardLegend - theFilenamePath /bea_testing/output/SupervisedClustering_Batches_Structures/Batches/SupervisedClust_Legend-ShipDate.png

2017 10 03 09:16:43.299 DEBUG MachineName mbatchStandardLegend - theLegendNames 2013-05-08 (80)

2017 10 03 09:16:43.299 DEBUG MachineName mbatchStandardLegend - theLegendNames 1

2017 10 03 09:16:43.300 DEBUG MachineName mbatchStandardLegend - theLegendColors 1

2017 10 03 09:16:43.300 DEBUG MachineName mbatchStandardLegend - theLegendSymbols 0

2017 10 03 09:16:43.300 DEBUG MachineName mbatchStandardLegend - myColors #b30000

2017 10 03 09:16:43.301 DEBUG MachineName mbatchStandardLegend before java

LegendJava 2013_05_03_0823

writeLegendWithSymbols theTitle = ShipDate

writeLegendWithSymbols theVersion = MBatch 1.4.16

writeLegendWithSymbols theFilenamePath = /bea_testing/output/SupervisedClustering_Batches_Structures/
ShipDate.png

Colors is non-null

writeLegendWithSymbols write

writeLegendWithSymbols done

2017 10 03 09:16:43.359 DEBUG MachineName mbatchStandardLegend after java

2017 10 03 09:16:43.360 INFO MachineName createBatchEffectsOutput_SupervisedClustering_batches - batchTypeName = TSS

2017 10 03 09:16:43.361 INFO MachineName makeBiasClust - starting

2017 10 03 09:16:43.586 INFO MachineName makeBiasClust - quantile dat dim = 1250,80

2017 10 03 09:16:43.587 INFO MachineName makeBiasClust - quantile U.data is.data.frame = FALSE

2017 10 03 09:16:43.588 INFO MachineName makeBiasClust - quantile U.data is.array = TRUE

2017 10 03 09:16:43.588 INFO MachineName makeBiasClust - quantile U.data is.list = FALSE

2017 10 03 09:16:43.588 INFO MachineName makeBiasClust - quantile U.data nrow = 312

2017 10 03 09:16:43.589 INFO MachineName makeBiasClust - quantile U.data ncol = 80

2017 10 03 09:16:43.589 INFO MachineName makeBiasClust - quantile U.data length = 24960

2017 10 03 09:16:43.589 INFO MachineName makeBiasClust - quantile U.data dim = 312,80

2017 10 03 09:16:43.590 INFO MachineName makeBiasClust - quantile U.data is.null = FALSE

2017 10 03 09:16:43.590 INFO MachineName makeBiasClust - data frame

2017 10 03 09:16:43.591 INFO MachineName makeBiasClust - U.dend1 <- bias.clust

2017 10 03 09:16:43.594 INFO MachineName makeBiasClust new.dis size - 80-80

2017 10 03 09:16:43.594 INFO MachineName makeBiasClust orig - 80-80

2017 10 03 09:16:43.595 INFO MachineName makeBiasClust is.na - 80-80

2017 10 03 09:16:43.595 INFO MachineName makeBiasClust is.infinite - 80-80

2017 10 03 09:16:45.992 DEBUG MachineName mbatchStandardLegend - Calling .jinit /home/linux/R/x86_64-pc-linux-gnu-library/3.4/MBatch/LegendJava/LegendJava.jar:/home/linux/R/x86
pc-linux-gnu-library/3.4/MBatch/LegendJava/jcommon-1.0.17.jar:/home/linux/R/x86_64-
pc-linux-gnu-library/3.4/MBatch/LegendJava/jfreechart-1.0.14.jar

2017 10 03 09:16:45.999 DEBUG MachineName mbatchStandardLegend - .jinit complete

2017 10 03 09:16:45.999 DEBUG MachineName mbatchStandardLegend - theTitle TSS

2017 10 03 09:16:46.000 DEBUG MachineName mbatchStandardLegend - theVersion MBatch 1.4.16

2017 10 03 09:16:46.000 DEBUG MachineName mbatchStandardLegend - theFilenamePath /bea_testing/output/SupervisedClustering_Batches_Structures/Batches/SupervisedClust_Legend-TSS.png

2017 10 03 09:16:46.000 DEBUG MachineName mbatchStandardLegend - theLegendNames OR - University of Michigan

(72), OU - Roswell Park (1), P6 - Translational Genomics

Research Institute (2), PA - University of Minnesota

(1), PK - University Health

Network (4)

2017 10 03 09:16:46.001 DEBUG MachineName mbatchStandardLegend - theLegendNames 5

2017 10 03 09:16:46.001 DEBUG MachineName mbatchStandardLegend - theLegendColors 5

2017 10 03 09:16:46.001 DEBUG MachineName mbatchStandardLegend - theLegendSymbols 0

2017 10 03 09:16:46.002 DEBUG MachineName mbatchStandardLegend - myColors #b30000,#8fb300,#00b347,#0047b3,#8f00b3

2017 10 03 09:16:46.002 DEBUG MachineName mbatchStandardLegend before java

LegendJava 2013_05_03_0823

writeLegendWithSymbols theTitle = TSS

writeLegendWithSymbols theVersion = MBatch 1.4.16

writeLegendWithSymbols theFilenamePath = /bea_testing/output/SupervisedClustering_Batches_Structures/
TSS.png

Colors is non-null

writeLegendWithSymbols write

writeLegendWithSymbols done

2017 10 03 09:16:46.109 DEBUG MachineName mbatchStandardLegend after java

>

Example File Output

The above code creates the following output files. Note that since we called the "Batches" version of this function (there is a separate "Pairs" version) files are placed within a Batches subdirectory. Files are named using the following naming convention:

SupervisedClust_Diagram-<BatchType>.png

SupervisedClust_Legend-<BatchType>.png

The diagram file contains a Supervised Clustering plot for a single batch type (the columns from the batches.tsv file). The legend gives the list of batches within the given batch type, and the colors used for each batch within the covariate bar.

linux@MachineName:/bea_testing/output/SupervisedClustering_Batches_Structures/Batches$ ls -l

total 1060

-rw-r--r-- 1 linux linux 262001 Oct 3 09:16 SupervisedClust_Diagram-BatchId.png

-rw-r--r-- 1 linux linux 261610 Oct 3 09:16 SupervisedClust_Diagram-PlateId.png

-rw-r--r-- 1 linux linux 262158 Oct 3 09:16 SupervisedClust_Diagram-ShipDate.png

-rw-r--r-- 1 linux linux 262373 Oct 3 09:16 SupervisedClust_Diagram-TSS.png

-rw-r--r-- 1 linux linux 2701 Oct 3 09:16 SupervisedClust_Legend-BatchId.png

-rw-r--r-- 1 linux linux 2839 Oct 3 09:16 SupervisedClust_Legend-PlateId.png

-rw-r--r-- 1 linux linux 3196 Oct 3 09:16 SupervisedClust_Legend-ShipDate.png

-rw-r--r-- 1 linux linux 12899 Oct 3 09:16 SupervisedClust_Legend-TSS.png