Using MBatch Assessments: SupervisedClustering_Pairs_Structures

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

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

3 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.

4 Usage

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

5 Arguments

5.1 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.

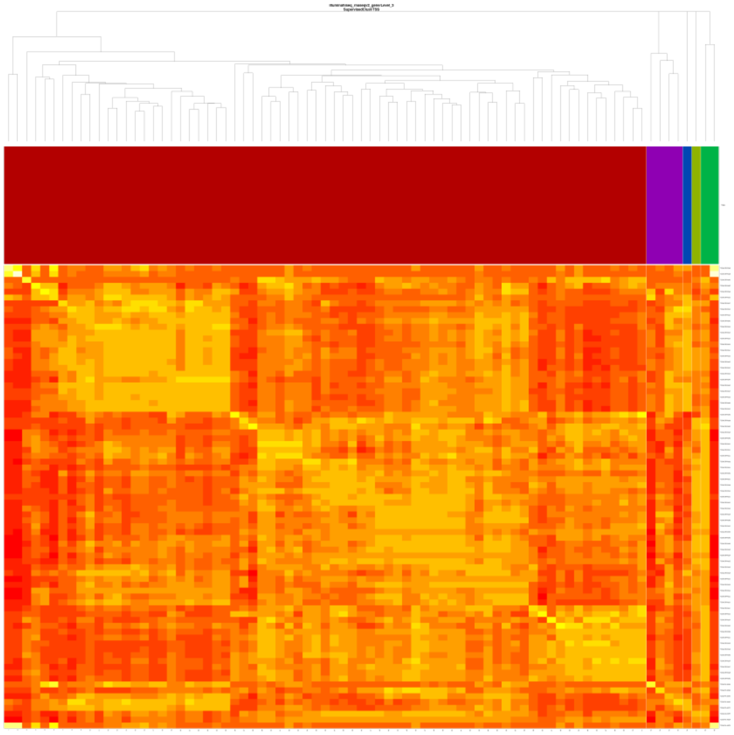

Figure 1: Supervised Clustering Example

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 theTitle

A string title to use in PNG files.

5.3 theOutputPath

String giving directory in which to place output PNG files.

5.4 theDoHeatmapFlag

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

5.5 theListOfBatchPairs

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

5.6 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.

5.7 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.

6 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.

```r
{
  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_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() )
}
```

```
## 2019 10 10 11:16:07.650 DEBUG megazone23 Changing LC_COLLATE to C for duration of run
## 2019 10 10 11:16:07.651 INFO megazone23 \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/
## 2019 10 10 11:16:07.651 INFO megazone23 Starting mbatchLoadFiles
## 2019 10 10 11:16:07.651 INFO megazone23 MBatch Version: 2019-09-04-1100
## 2019 10 10 11:16:07.664 INFO megazone23 read batch file= /bea_testing/MATRIX_DATA/batches-Tumor.tsv
## 2019 10 10 11:16:07.671 INFO megazone23 read gene file= /bea_testing/MATRIX_DATA/matrix_data-Tumor.ts
## 2019 10 10 11:16:13.221 INFO megazone23 filter samples in batches using gene samples
## 2019 10 10 11:16:13.222 INFO megazone23 sort batches by gene file samples
## 2019 10 10 11:16:13.306 INFO megazone23 Finishing mbatchLoadFiles
## 2019 10 10 11:16:13.306 INFO megazone23 ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
## 2019 10 10 11:16:13.306 DEBUG megazone23 Changing LC_COLLATE to C for duration of run
## 2019 10 10 11:16:13.307 INFO megazone23 \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/
## 2019 10 10 11:16:13.307 INFO megazone23 mbatchTrimData Starting
## 2019 10 10 11:16:13.307 INFO megazone23 MBatch Version: 2019-09-04-1100
## 2019 10 10 11:16:20.783 INFO megazone23 mbatchTrimData theMaxSize= 1e+05
## 2019 10 10 11:16:20.784 INFO megazone23 mbatchTrimData ncol(theMatrix)= 80
## 2019 10 10 11:16:20.784 INFO megazone23 mbatchTrimData nrow(theMatrix)= 1250
## 2019 10 10 11:16:20.784 INFO megazone23 mbatchTrimData Finishing
## 2019 10 10 11:16:20.784 INFO megazone23 ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
## 2019 10 10 11:16:20.785 DEBUG megazone23 Changing LC_COLLATE to C for duration of run
## 2019 10 10 11:16:20.785 INFO megazone23 \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/
## 2019 10 10 11:16:20.786 INFO megazone23 mbatchFilterData Starting
## 2019 10 10 11:16:20.786 INFO megazone23 MBatch Version: 2019-09-04-1100
## 2019 10 10 11:16:20.786 DEBUG megazone23 rows pre filter  1250
## 2019 10 10 11:16:21.036 DEBUG megazone23 rows post filter  1250
## 2019 10 10 11:16:21.037 DEBUG megazone23 mbatchFilterData Prefilter, gene data had  1250  while post
## 2019 10 10 11:16:21.038 DEBUG megazone23 mbatchFilterData Prefilter, batch data had  80  while post
## 2019 10 10 11:16:21.038 INFO megazone23 mbatchFilterData Finishing
## 2019 10 10 11:16:21.038 INFO megazone23 ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
## 2019 10 10 11:16:21.039 INFO megazone23 createBatchEffectsOutput_SupervisedClustering_pairs - batchTy
## 2019 10 10 11:16:21.039 INFO megazone23 createBatchEffectsOutput_SupervisedClustering_pairs - batchTy
## 2019 10 10 11:16:21.040 DEBUG megazone23 checkCreateDir:  /bea_testing/output/SupervisedClustering_Pa
## 2019 10 10 11:16:21.041 INFO megazone23 makeBiasClust - starting
## 2019 10 10 11:16:21.143 INFO megazone23 makeBiasClust - quantile dat dim =  1250,80
## 2019 10 10 11:16:21.145 INFO megazone23 makeBiasClust - quantile U.data is.data.frame =  FALSE
## 2019 10 10 11:16:21.145 INFO megazone23 makeBiasClust - quantile U.data is.array =  TRUE
## 2019 10 10 11:16:21.145 INFO megazone23 makeBiasClust - quantile U.data is.list =  FALSE
## 2019 10 10 11:16:21.145 INFO megazone23 makeBiasClust - quantile U.data nrow =  312
## 2019 10 10 11:16:21.145 INFO megazone23 makeBiasClust - quantile U.data ncol =  80
## 2019 10 10 11:16:21.146 INFO megazone23 makeBiasClust - quantile U.data length =  24960
## 2019 10 10 11:16:21.146 INFO megazone23 makeBiasClust - quantile U.data dim =  312,80
## 2019 10 10 11:16:21.146 INFO megazone23 makeBiasClust - quantile U.data is.null =  FALSE
## 2019 10 10 11:16:21.146 INFO megazone23 makeBiasClust - data frame
## 2019 10 10 11:16:21.147 INFO megazone23 makeBiasClust - U.dend1 <- bias.clust
## 2019 10 10 11:16:21.151 INFO megazone23 makeBiasClust new.dis size -  80-80
## 2019 10 10 11:16:21.151 INFO megazone23 makeBiasClust orig -  80-80
## 2019 10 10 11:16:21.151 INFO megazone23 makeBiasClust is.na -  80-80
## 2019 10 10 11:16:21.152 INFO megazone23 makeBiasClust is.infinite -  80-80
```

```
## 2019 10 10 11:16:23.489 DEBUG megazone23 mbatchStandardLegend - Calling .jinit  /tmp/RtmpGYrlWy/temp_
## 2019 10 10 11:16:23.493 DEBUG megazone23 mbatchStandardLegend - .jinit complete
## 2019 10 10 11:16:23.493 DEBUG megazone23 mbatchStandardLegend - theTitle  PlateId
## 2019 10 10 11:16:23.494 DEBUG megazone23 mbatchStandardLegend - theVersion  MBatch 1.5.2
## 2019 10 10 11:16:23.494 DEBUG megazone23 mbatchStandardLegend - theFilenamePath  /bea_testing/output/
## 2019 10 10 11:16:23.494 DEBUG megazone23 mbatchStandardLegend - theLegendNames  A29J (80)
## 2019 10 10 11:16:23.494 DEBUG megazone23 mbatchStandardLegend - theLegendNames  1
## 2019 10 10 11:16:23.494 DEBUG megazone23 mbatchStandardLegend - theLegendColors  1
## 2019 10 10 11:16:23.495 DEBUG megazone23 mbatchStandardLegend - theLegendSymbols  0
## 2019 10 10 11:16:23.495 DEBUG megazone23 mbatchStandardLegend - myColors  #b30000
## 2019 10 10 11:16:23.495 DEBUG megazone23 mbatchStandardLegend before java
## 2019 10 10 11:16:23.563 DEBUG megazone23 mbatchStandardLegend after java
## 2019 10 10 11:16:23.578 DEBUG megazone23 mbatchStandardLegend - Calling .jinit  /tmp/RtmpGYrlWy/temp_
## 2019 10 10 11:16:23.585 DEBUG megazone23 mbatchStandardLegend - .jinit complete
## 2019 10 10 11:16:23.586 DEBUG megazone23 mbatchStandardLegend - theTitle  TSS
## 2019 10 10 11:16:23.586 DEBUG megazone23 mbatchStandardLegend - theVersion  MBatch 1.5.2
## 2019 10 10 11:16:23.586 DEBUG megazone23 mbatchStandardLegend - theFilenamePath  /bea_testing/output/
## 2019 10 10 11:16:23.586 DEBUG megazone23 mbatchStandardLegend - theLegendNames  OR - University of Mi
## (72), OU - Roswell Park (1), P6 - Translational Genomics
## Research Institute (2), PA - University of Minnesota
## (1), PK - University Health
## Network (4)
## 2019 10 10 11:16:23.587 DEBUG megazone23 mbatchStandardLegend - theLegendNames  5
## 2019 10 10 11:16:23.587 DEBUG megazone23 mbatchStandardLegend - theLegendColors  5
## 2019 10 10 11:16:23.587 DEBUG megazone23 mbatchStandardLegend - theLegendSymbols  0
## 2019 10 10 11:16:23.587 DEBUG megazone23 mbatchStandardLegend - myColors  #b30000,#8fb300,#00b347,#0
## 2019 10 10 11:16:23.587 DEBUG megazone23 mbatchStandardLegend before java
## 2019 10 10 11:16:23.666 DEBUG megazone23 mbatchStandardLegend after java
## 2019 10 10 11:16:23.667 INFO megazone23 createBatchEffectsOutput_SupervisedClustering_pairs - batchT
## 2019 10 10 11:16:23.667 INFO megazone23 createBatchEffectsOutput_SupervisedClustering_pairs - batchT
## 2019 10 10 11:16:23.668 DEBUG megazone23 checkCreateDir:  /bea_testing/output/SupervisedClustering_P
## 2019 10 10 11:16:23.669 INFO megazone23 makeBiasClust - starting
## 2019 10 10 11:16:23.793 INFO megazone23 makeBiasClust - quantile dat dim =  1250,80
## 2019 10 10 11:16:23.794 INFO megazone23 makeBiasClust - quantile U.data is.data.frame =  FALSE
## 2019 10 10 11:16:23.795 INFO megazone23 makeBiasClust - quantile U.data is.array =  TRUE
## 2019 10 10 11:16:23.795 INFO megazone23 makeBiasClust - quantile U.data is.list =  FALSE
## 2019 10 10 11:16:23.795 INFO megazone23 makeBiasClust - quantile U.data nrow =  312
## 2019 10 10 11:16:23.795 INFO megazone23 makeBiasClust - quantile U.data ncol =  80
## 2019 10 10 11:16:23.796 INFO megazone23 makeBiasClust - quantile U.data length =  24960
## 2019 10 10 11:16:23.796 INFO megazone23 makeBiasClust - quantile U.data dim =  312,80
## 2019 10 10 11:16:23.796 INFO megazone23 makeBiasClust - quantile U.data is.null =  FALSE
## 2019 10 10 11:16:23.796 INFO megazone23 makeBiasClust - data frame
## 2019 10 10 11:16:23.797 INFO megazone23 makeBiasClust - U.dend1 <- bias.clust
## 2019 10 10 11:16:23.799 INFO megazone23 makeBiasClust new.dis size -  80-80
## 2019 10 10 11:16:23.800 INFO megazone23 makeBiasClust orig -  80-80
## 2019 10 10 11:16:23.801 INFO megazone23 makeBiasClust is.na -  80-80
## 2019 10 10 11:16:23.801 INFO megazone23 makeBiasClust is.infinite -  80-80
```

```
## 2019 10 10 11:16:26.068 DEBUG megazone23 mbatchStandardLegend - Calling .jinit  /tmp/RtmpGYrlWy/temp
## 2019 10 10 11:16:26.072 DEBUG megazone23 mbatchStandardLegend - .jinit complete
## 2019 10 10 11:16:26.073 DEBUG megazone23 mbatchStandardLegend - theTitle  BatchId
## 2019 10 10 11:16:26.073 DEBUG megazone23 mbatchStandardLegend - theVersion  MBatch 1.5.2
## 2019 10 10 11:16:26.073 DEBUG megazone23 mbatchStandardLegend - theFilenamePath  /bea_testing/output/
## 2019 10 10 11:16:26.074 DEBUG megazone23 mbatchStandardLegend - theLegendNames  00304 (80)
## 2019 10 10 11:16:26.074 DEBUG megazone23 mbatchStandardLegend - theLegendNames  1
## 2019 10 10 11:16:26.074 DEBUG megazone23 mbatchStandardLegend - theLegendColors  1
## 2019 10 10 11:16:26.074 DEBUG megazone23 mbatchStandardLegend - theLegendSymbols  0
## 2019 10 10 11:16:26.075 DEBUG megazone23 mbatchStandardLegend - myColors  #b30000
## 2019 10 10 11:16:26.075 DEBUG megazone23 mbatchStandardLegend before java
## 2019 10 10 11:16:26.089 DEBUG megazone23 mbatchStandardLegend after java
## 2019 10 10 11:16:26.092 DEBUG megazone23 mbatchStandardLegend - Calling .jinit  /tmp/RtmpGYrlWy/temp
## 2019 10 10 11:16:26.097 DEBUG megazone23 mbatchStandardLegend - .jinit complete
## 2019 10 10 11:16:26.097 DEBUG megazone23 mbatchStandardLegend - theTitle  TSS
## 2019 10 10 11:16:26.097 DEBUG megazone23 mbatchStandardLegend - theVersion  MBatch 1.5.2
## 2019 10 10 11:16:26.098 DEBUG megazone23 mbatchStandardLegend - theFilenamePath  /bea_testing/output/
## 2019 10 10 11:16:26.098 DEBUG megazone23 mbatchStandardLegend - theLegendNames  OR - University of M
## (72), OU - Roswell Park (1), P6 - Translational Genomics
## Research Institute (2), PA - University of Minnesota
## (1), PK - University Health
## Network (4)
## 2019 10 10 11:16:26.098 DEBUG megazone23 mbatchStandardLegend - theLegendNames  5
## 2019 10 10 11:16:26.098 DEBUG megazone23 mbatchStandardLegend - theLegendColors  5
## 2019 10 10 11:16:26.099 DEBUG megazone23 mbatchStandardLegend - theLegendSymbols  0
## 2019 10 10 11:16:26.099 DEBUG megazone23 mbatchStandardLegend - myColors  #b30000,#8fb300,#00b347,#00
## 2019 10 10 11:16:26.099 DEBUG megazone23 mbatchStandardLegend before java
## 2019 10 10 11:16:26.172 DEBUG megazone23 mbatchStandardLegend after java
```

7 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.

```
/bea_testing/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.

```
/bea_testing/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.

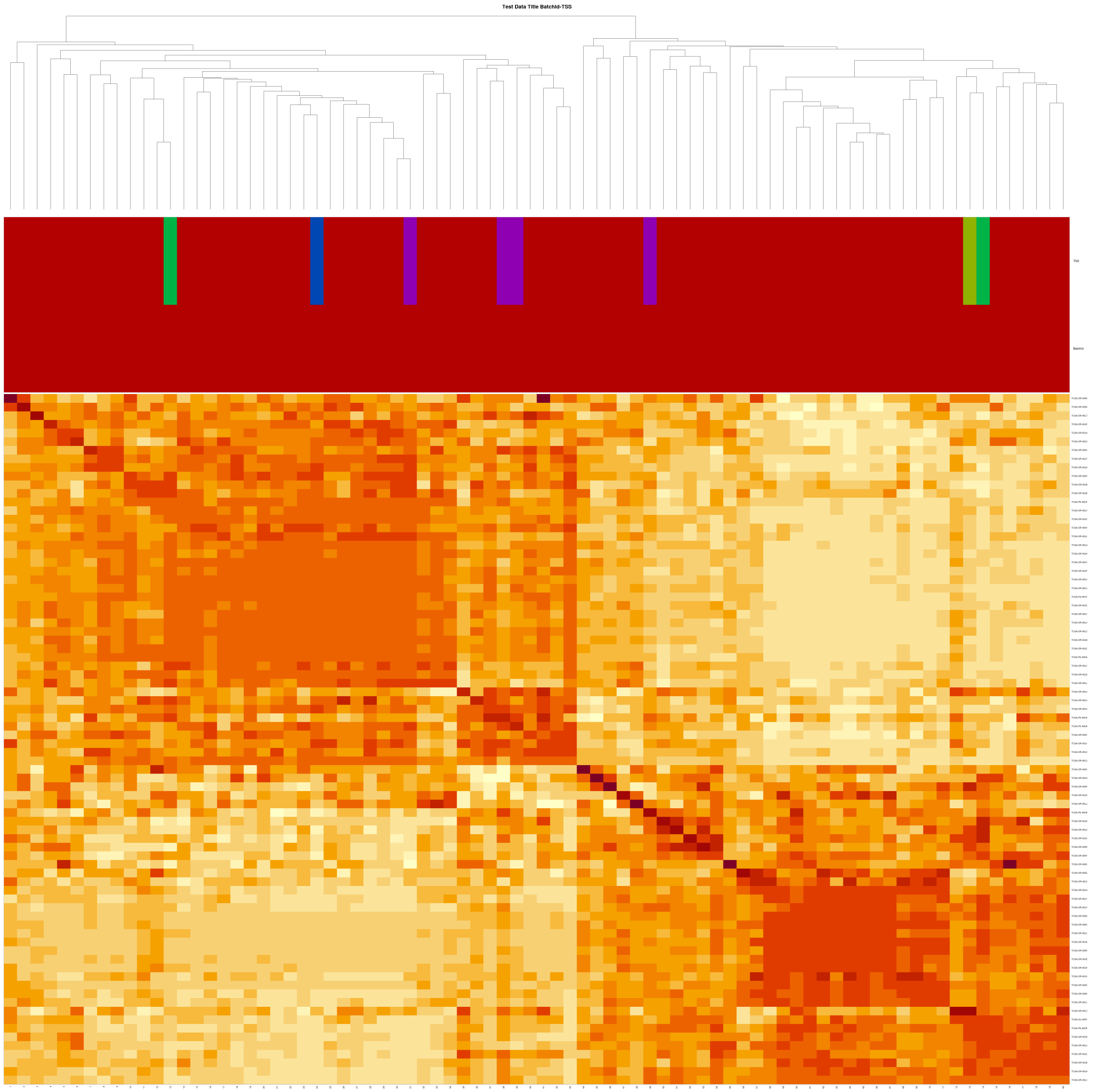

Figure 2: SupervisedClustering_Pairs_Structures Output