Using MBatch Assessments: Boxplot_AllSamplesRLE_Structures

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

Boxplot_AllSamplesRLE_Structures is a function used to perform batch effects assessments using the boxplots on all samples using RLE (run length encoding).

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 set of boxplots where each boxplot (also called a box and whisper plot) represent a single sample. For datasets with many samples, the static PNG may be so dense as to be unusable.

The All Samples Boxplots plot the value for each feature (genes or probes) for a sample, with the samples grouped and colored by batch. So the vertical axis is based on the values of the original data and the points plotted are features. The actual meaning of the data used, such as expression, read counts, and the like, will vary based on the data being processed.

Here is an example of a smallish dynamic boxplot. (See Batch Effects Viewer documentation for more details.)

Here is an example of the static plot for a medium-sized dataset.

4 Usage

Boxplot_AllSamplesRLE_Structures(theData, theTitle, theOutputPath, theBatchTypeAndValuePairsToRemove, theBatchTypeAndValuePairsToKeep, theDataVersion, theTestVersion, theMaxGeneCount=20000)

5 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.

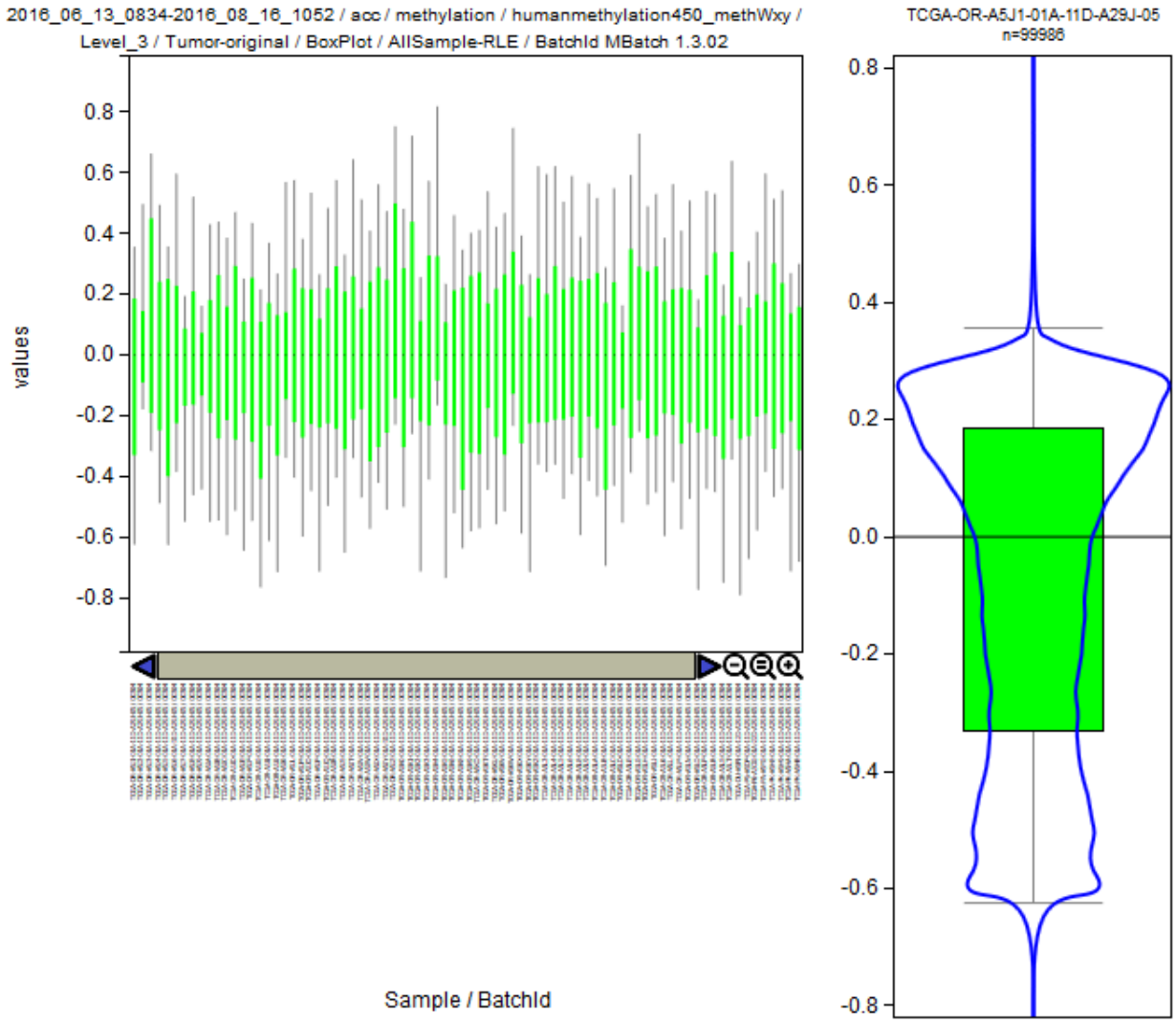

Figure 1: Dynamic Boxplot Example

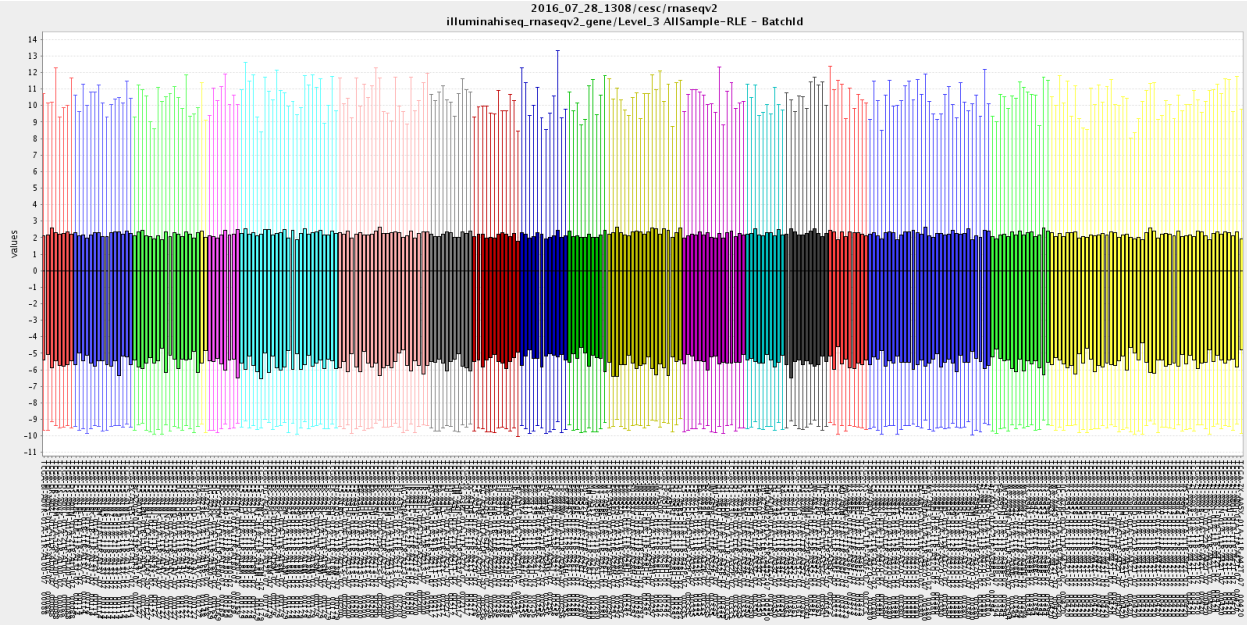

Figure 2: Static Boxplot Example

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.

##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.

##theMaxGeneCount
Integer giving maximum number of features (genes) to keep. Default is 20000. 0 means keep all.

6 Example Call

The following code is adapted from the tests/Boxplot_AllSamplesRLE_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.

```
{
  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, "Boxplot_AllSamplesRLE_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
  Boxplot_AllSamplesRLE_Structures(myData, "Disease/Data Type/Platform/Data Level", theOutputDir, list()
                                   theDataVersion="DATA_2022-09-09-1600", theTestVersion="TEST_2022-10-
}
```

```
## 2023 10 06 12:31:42.501 DEBUG qcprludev10 Changing LC_COLLATE to C for duration of run
## 2023 10 06 12:31:42.505 INFO qcprludev10 \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/
## 2023 10 06 12:31:42.505 INFO qcprludev10 Starting mbatchLoadFiles
## 2023 10 06 12:31:42.506 INFO qcprludev10 MBatch Version: BEA_VERSION_TIMESTAMP
## 2023 10 06 12:31:42.506 INFO qcprludev10 read batch file= /builds/BatchEffects_clean/BatchEffectsPacl
## 2023 10 06 12:31:42.507 INFO qcprludev10 read gene file= /builds/BatchEffects_clean/BatchEffectsPacka
## 2023 10 06 12:31:44.897 INFO qcprludev10 filter samples in batches using gene samples
## 2023 10 06 12:31:44.899 INFO qcprludev10 sort batches by gene file samples
## 2023 10 06 12:31:44.965 INFO qcprludev10 Finishing mbatchLoadFiles
## 2023 10 06 12:31:44.965 INFO qcprludev10 ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
## 2023 10 06 12:31:44.966 DEBUG qcprludev10 Changing LC_COLLATE to C for duration of run
## 2023 10 06 12:31:44.966 INFO qcprludev10 \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/
## 2023 10 06 12:31:44.967 INFO qcprludev10 mbatchTrimData Starting
## 2023 10 06 12:31:44.967 INFO qcprludev10 MBatch Version: BEA_VERSION_TIMESTAMP
## 2023 10 06 12:31:52.432 INFO qcprludev10 mbatchTrimData theMaxSize= 1e+05
## 2023 10 06 12:31:52.433 INFO qcprludev10 mbatchTrimData ncol(theMatrix)= 80
## 2023 10 06 12:31:52.434 INFO qcprludev10 mbatchTrimData nrow(theMatrix)= 1250
## 2023 10 06 12:31:52.434 INFO qcprludev10 mbatchTrimData Finishing
## 2023 10 06 12:31:52.434 INFO qcprludev10 ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
## 2023 10 06 12:31:52.435 DEBUG qcprludev10 Changing LC_COLLATE to C for duration of run
## 2023 10 06 12:31:52.436 INFO qcprludev10 \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/
## 2023 10 06 12:31:52.436 INFO qcprludev10 mbatchFilterData Starting
## 2023 10 06 12:31:52.437 INFO qcprludev10 MBatch Version: BEA_VERSION_TIMESTAMP
## 2023 10 06 12:31:52.437 DEBUG qcprludev10 rows pre filter  1250
## 2023 10 06 12:31:52.663 DEBUG qcprludev10 rows post filter  1250
## 2023 10 06 12:31:52.664 DEBUG qcprludev10 mbatchFilterData Prefilter, gene data had  1250  while post
## 2023 10 06 12:31:52.665 DEBUG qcprludev10 mbatchFilterData Prefilter, batch data had  80  while post
## 2023 10 06 12:31:52.665 INFO qcprludev10 mbatchFilterData Finishing
## 2023 10 06 12:31:52.666 INFO qcprludev10 ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
## 2023 10 06 12:31:52.666 DEBUG qcprludev10 Changing LC_COLLATE to C for duration of run
## 2023 10 06 12:31:52.667 INFO qcprludev10 \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/
## 2023 10 06 12:31:52.667 INFO qcprludev10 mbatchTrimData Starting
## 2023 10 06 12:31:52.668 INFO qcprludev10 MBatch Version: BEA_VERSION_TIMESTAMP
## 2023 10 06 12:31:52.668 INFO qcprludev10 mbatchTrimData theMaxSize= 1600000
```

```
## 2023 10 06 12:31:52.669 INFO qcprludev10 mbatchTrimData ncol(theMatrix)= 80
## 2023 10 06 12:31:52.669 INFO qcprludev10 mbatchTrimData nrow(theMatrix)= 1250
## 2023 10 06 12:31:52.669 INFO qcprludev10 mbatchTrimData Finishing
## 2023 10 06 12:31:52.670 INFO qcprludev10 ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
## 2023 10 06 12:31:52.670 DEBUG qcprludev10 createBatchEffectsOutput_BoxPlot_AllSampleRLE - theOutputDi
## 2023 10 06 12:31:52.671 DEBUG qcprludev10 dim(theMatrixGeneData) 1250, dim(theMatrixGeneData) 80
## 2023 10 06 12:31:52.671 DEBUG qcprludev10 length(colnames(theMatrixGeneData)) 80
## 2023 10 06 12:31:52.672 DEBUG qcprludev10 length(rownames(theMatrixGeneData)) 1250
## 2023 10 06 12:31:52.672 DEBUG qcprludev10 dim(theDataframeBatchData) 80, dim(theDataframeBatchData) 5
## 2023 10 06 12:31:52.672 DEBUG qcprludev10 length(names(theDataframeBatchData)) 5
## 2023 10 06 12:31:52.673 DEBUG qcprludev10 batchTypeName =  BatchId
## 2023 10 06 12:31:52.673 DEBUG qcprludev10 theBatchType= BatchId
## 2023 10 06 12:31:52.674 DEBUG qcprludev10 calcAndWriteBoxplot - theBoxDataFile= /BEA/BatchEffectsPacl
## 2023 10 06 12:31:52.675 DEBUG qcprludev10 calcAndWriteBoxplot - dim(theData)[1]= 1250
## 2023 10 06 12:31:52.675 DEBUG qcprludev10 calcAndWriteBoxplot - dim(theData)[2]= 80
## 2023 10 06 12:31:52.676 DEBUG qcprludev10 writeTitleFile - pre title  Disease/Data Type/Platform/Data
## 2023 10 06 12:31:52.676 DEBUG qcprludev10 writeTitleFile - theTitle  Disease / Data Type / Platform /
## 2023 10 06 12:31:52.677 DEBUG qcprludev10 writeTitleFile - titleFile  /BEA/BatchEffectsPackage_data/t
## 2023 10 06 12:31:52.677 DEBUG qcprludev10 calcAndWriteBoxplot - before calcAndWriteBoxDataFile
## 2023 10 06 12:31:52.678 DEBUG qcprludev10 calcAndWriteBoxDataFile theFile= /BEA/BatchEffectsPackage_d
## 2023 10 06 12:31:52.678 DEBUG qcprludev10 calcAndWriteBoxDataFile thePngFile= /BEA/BatchEffectsPackag
## 2023 10 06 12:31:52.716 DEBUG qcprludev10 calcAndWriteBoxDataFile CairoPNG= /BEA/BatchEffectsPackage_
## 2023 10 06 12:31:52.717 DEBUG qcprludev10 calcAndWriteBoxDataFile call boxplot

## 2023 10 06 12:31:52.754 DEBUG qcprludev10 calcAndWriteBoxDataFile call text
## 2023 10 06 12:31:52.765 DEBUG qcprludev10 calcAndWriteBoxDataFile done
## 2023 10 06 12:31:52.807 DEBUG qcprludev10 calcAndWriteBoxplot - after calcAndWriteBoxDataFile
## 2023 10 06 12:31:52.807 DEBUG qcprludev10 calcAndWriteBoxplot - before calcAndWriteHistogramFile
## 2023 10 06 12:31:52.808 DEBUG qcprludev10 calcAndWriteHistogramFile  /BEA/BatchEffectsPackage_data/te
## 2023 10 06 12:31:52.898 DEBUG qcprludev10 calcAndWriteBoxplot - after calcAndWriteHistogramFile
## 2023 10 06 12:31:52.899 DEBUG qcprludev10 calcAndWriteBoxplot - before calcAndWriteAnnotationsFile
## 2023 10 06 12:31:52.899 DEBUG qcprludev10 calcAndWriteAnnotationsFile theFile= /BEA/BatchEffectsPacka
## 2023 10 06 12:31:52.916 DEBUG qcprludev10 calcAndWriteBoxplot - after calcAndWriteAnnotationsFile
## 2023 10 06 12:31:52.917 DEBUG qcprludev10 batchTypeName =  PlateId
## 2023 10 06 12:31:52.917 DEBUG qcprludev10 theBatchType= PlateId
## 2023 10 06 12:31:52.918 DEBUG qcprludev10 calcAndWriteBoxplot - theBoxDataFile= /BEA/BatchEffectsPacl
## 2023 10 06 12:31:52.918 DEBUG qcprludev10 calcAndWriteBoxplot - dim(theData)[1]= 1250
## 2023 10 06 12:31:52.918 DEBUG qcprludev10 calcAndWriteBoxplot - dim(theData)[2]= 80
## 2023 10 06 12:31:52.919 DEBUG qcprludev10 writeTitleFile - pre title  Disease/Data Type/Platform/Data
## 2023 10 06 12:31:52.919 DEBUG qcprludev10 writeTitleFile - theTitle  Disease / Data Type / Platform /
## 2023 10 06 12:31:52.920 DEBUG qcprludev10 writeTitleFile - titleFile  /BEA/BatchEffectsPackage_data/t
## 2023 10 06 12:31:52.921 DEBUG qcprludev10 calcAndWriteBoxplot - before calcAndWriteBoxDataFile
## 2023 10 06 12:31:52.921 DEBUG qcprludev10 calcAndWriteBoxDataFile theFile= /BEA/BatchEffectsPackage_d
## 2023 10 06 12:31:52.921 DEBUG qcprludev10 calcAndWriteBoxDataFile thePngFile= /BEA/BatchEffectsPackag
## 2023 10 06 12:31:52.962 DEBUG qcprludev10 calcAndWriteBoxDataFile CairoPNG= /BEA/BatchEffectsPackage_
## 2023 10 06 12:31:52.964 DEBUG qcprludev10 calcAndWriteBoxDataFile call boxplot

## 2023 10 06 12:31:52.998 DEBUG qcprludev10 calcAndWriteBoxDataFile call text
## 2023 10 06 12:31:53.006 DEBUG qcprludev10 calcAndWriteBoxDataFile done
## 2023 10 06 12:31:53.048 DEBUG qcprludev10 calcAndWriteBoxplot - after calcAndWriteBoxDataFile
## 2023 10 06 12:31:53.049 DEBUG qcprludev10 calcAndWriteBoxplot - before calcAndWriteHistogramFile
## 2023 10 06 12:31:53.049 DEBUG qcprludev10 calcAndWriteHistogramFile  /BEA/BatchEffectsPackage_data/te
## 2023 10 06 12:31:53.144 DEBUG qcprludev10 calcAndWriteBoxplot - after calcAndWriteHistogramFile
## 2023 10 06 12:31:53.145 DEBUG qcprludev10 calcAndWriteBoxplot - before calcAndWriteAnnotationsFile
## 2023 10 06 12:31:53.145 DEBUG qcprludev10 calcAndWriteAnnotationsFile theFile= /BEA/BatchEffectsPacka
```

```
## 2023 10 06 12:31:53.162 DEBUG qcprludev10 calcAndWriteBoxplot - after calcAndWriteAnnotationsFile
## 2023 10 06 12:31:53.164 DEBUG qcprludev10 batchTypeName =  ShipDate
## 2023 10 06 12:31:53.165 DEBUG qcprludev10 theBatchType= ShipDate
## 2023 10 06 12:31:53.166 DEBUG qcprludev10 calcAndWriteBoxplot - theBoxDataFile= /BEA/BatchEffectsPacl
## 2023 10 06 12:31:53.166 DEBUG qcprludev10 calcAndWriteBoxplot - dim(theData)[1]= 1250
## 2023 10 06 12:31:53.167 DEBUG qcprludev10 calcAndWriteBoxplot - dim(theData)[2]= 80
## 2023 10 06 12:31:53.167 DEBUG qcprludev10 writeTitleFile - pre title  Disease/Data Type/Platform/Data
## 2023 10 06 12:31:53.168 DEBUG qcprludev10 writeTitleFile - theTitle  Disease / Data Type / Platform /
## 2023 10 06 12:31:53.168 DEBUG qcprludev10 writeTitleFile - titleFile  /BEA/BatchEffectsPackage_data/t
## 2023 10 06 12:31:53.169 DEBUG qcprludev10 calcAndWriteBoxplot - before calcAndWriteBoxDataFile
## 2023 10 06 12:31:53.170 DEBUG qcprludev10 calcAndWriteBoxDataFile theFile= /BEA/BatchEffectsPackage_d
## 2023 10 06 12:31:53.170 DEBUG qcprludev10 calcAndWriteBoxDataFile thePngFile= /BEA/BatchEffectsPackag
## 2023 10 06 12:31:53.210 DEBUG qcprludev10 calcAndWriteBoxDataFile CairoPNG= /BEA/BatchEffectsPackage_
## 2023 10 06 12:31:53.211 DEBUG qcprludev10 calcAndWriteBoxDataFile call boxplot

## 2023 10 06 12:31:53.243 DEBUG qcprludev10 calcAndWriteBoxDataFile call text
## 2023 10 06 12:31:53.253 DEBUG qcprludev10 calcAndWriteBoxDataFile done
## 2023 10 06 12:31:53.294 DEBUG qcprludev10 calcAndWriteBoxplot - after calcAndWriteBoxDataFile
## 2023 10 06 12:31:53.295 DEBUG qcprludev10 calcAndWriteBoxplot - before calcAndWriteHistogramFile
## 2023 10 06 12:31:53.295 DEBUG qcprludev10 calcAndWriteHistogramFile  /BEA/BatchEffectsPackage_data/te
## 2023 10 06 12:31:53.398 DEBUG qcprludev10 calcAndWriteBoxplot - after calcAndWriteHistogramFile
## 2023 10 06 12:31:53.398 DEBUG qcprludev10 calcAndWriteBoxplot - before calcAndWriteAnnotationsFile
## 2023 10 06 12:31:53.399 DEBUG qcprludev10 calcAndWriteAnnotationsFile theFile= /BEA/BatchEffectsPacka
## 2023 10 06 12:31:53.413 DEBUG qcprludev10 calcAndWriteBoxplot - after calcAndWriteAnnotationsFile
## 2023 10 06 12:31:53.413 DEBUG qcprludev10 batchTypeName =  TSS
## 2023 10 06 12:31:53.414 DEBUG qcprludev10 theBatchType= TSS
## 2023 10 06 12:31:53.414 DEBUG qcprludev10 calcAndWriteBoxplot - theBoxDataFile= /BEA/BatchEffectsPacl
## 2023 10 06 12:31:53.415 DEBUG qcprludev10 calcAndWriteBoxplot - dim(theData)[1]= 1250
## 2023 10 06 12:31:53.415 DEBUG qcprludev10 calcAndWriteBoxplot - dim(theData)[2]= 80
## 2023 10 06 12:31:53.416 DEBUG qcprludev10 writeTitleFile - pre title  Disease/Data Type/Platform/Data
## 2023 10 06 12:31:53.416 DEBUG qcprludev10 writeTitleFile - theTitle  Disease / Data Type / Platform /
## 2023 10 06 12:31:53.417 DEBUG qcprludev10 writeTitleFile - titleFile  /BEA/BatchEffectsPackage_data/t
## 2023 10 06 12:31:53.417 DEBUG qcprludev10 calcAndWriteBoxplot - before calcAndWriteBoxDataFile
## 2023 10 06 12:31:53.418 DEBUG qcprludev10 calcAndWriteBoxDataFile theFile= /BEA/BatchEffectsPackage_d
## 2023 10 06 12:31:53.418 DEBUG qcprludev10 calcAndWriteBoxDataFile thePngFile= /BEA/BatchEffectsPackag
## 2023 10 06 12:31:53.452 DEBUG qcprludev10 calcAndWriteBoxDataFile CairoPNG= /BEA/BatchEffectsPackage_
## 2023 10 06 12:31:53.454 DEBUG qcprludev10 calcAndWriteBoxDataFile call boxplot

## 2023 10 06 12:31:53.481 DEBUG qcprludev10 calcAndWriteBoxDataFile call text
## 2023 10 06 12:31:53.496 DEBUG qcprludev10 calcAndWriteBoxDataFile done
## 2023 10 06 12:31:53.539 DEBUG qcprludev10 calcAndWriteBoxplot - after calcAndWriteBoxDataFile
## 2023 10 06 12:31:53.540 DEBUG qcprludev10 calcAndWriteBoxplot - before calcAndWriteHistogramFile
## 2023 10 06 12:31:53.540 DEBUG qcprludev10 calcAndWriteHistogramFile  /BEA/BatchEffectsPackage_data/te
## 2023 10 06 12:31:53.623 DEBUG qcprludev10 calcAndWriteBoxplot - after calcAndWriteHistogramFile
## 2023 10 06 12:31:53.623 DEBUG qcprludev10 calcAndWriteBoxplot - before calcAndWriteAnnotationsFile
## 2023 10 06 12:31:53.623 DEBUG qcprludev10 calcAndWriteAnnotationsFile theFile= /BEA/BatchEffectsPacka
## 2023 10 06 12:31:53.637 DEBUG qcprludev10 calcAndWriteBoxplot - after calcAndWriteAnnotationsFile

## writeBatchDataTsvForBoxplot

## /BEA/BatchEffectsPackage_data/testing_dynamic/MBatch/Boxplot_AllSamplesRLE_Structures/AllSample-RLE/M

## writeAsGenericDataframe writeBatchDataTsvForBoxplot

## Writing BatchData.tsv now

## [1] TRUE
```

7 Example File Output

The above code creates the following subdirectories and files. The subdirectories correspond to the run type were requested.

```
/output/Boxplot_AllSamplesRLE_Structures$ ls -l
total 44
drwxr-xr-x 2 linux linux 40960 Jun 19 11:41 AllSample-RLE
```

Looking at the "AllSample-RLE" subdirectory, it contains the diagram and legend files, and data usable with dynamic displays.

```
/output/Boxplot_AllSamplesRLE_Structures/AllSample-RLE$ ls -l
total 2472
-rw-r--r-- 1 linux linux  3873 Jun 19 11:40 BoxPlot_AllSample-RLE_Annotations-BatchId.tsv
-rw-r--r-- 1 linux linux  3873 Jun 19 11:41 BoxPlot_AllSample-RLE_Annotations-PlateId.tsv
-rw-r--r-- 1 linux linux  3873 Jun 19 11:41 BoxPlot_AllSample-RLE_Annotations-ShipDate.tsv
-rw-r--r-- 1 linux linux  3873 Jun 19 11:41 BoxPlot_AllSample-RLE_Annotations-TSS.tsv
-rw-r--r-- 1 linux linux 15387 Jun 19 11:40 BoxPlot_AllSample-RLE_BoxData-BatchId.tsv
-rw-r--r-- 1 linux linux 15387 Jun 19 11:41 BoxPlot_AllSample-RLE_BoxData-PlateId.tsv
-rw-r--r-- 1 linux linux 15387 Jun 19 11:41 BoxPlot_AllSample-RLE_BoxData-ShipDate.tsv
-rw-r--r-- 1 linux linux 15387 Jun 19 11:41 BoxPlot_AllSample-RLE_BoxData-TSS.tsv
-rw-r--r-- 1 linux linux     9 Jun 19 11:40 BoxPlot_AllSample-RLE_CatData-BatchId-TCGA-OR-A5J1-01A-11D-
-rw-r--r-- 1 linux linux  7647 Jun 19 11:40 BoxPlot_AllSample-RLE_CatData-BatchId-TCGA-OR-A5J2-01A-11D-
#snipped out "CatData" files for each sample for each batch type
-rw-r--r-- 1 linux linux  6688 Jun 19 11:41 BoxPlot_AllSample-RLE_CatData-TSS-TCGA-PK-A5HA-01A-11D-A29J
-rw-r--r-- 1 linux linux  5583 Jun 19 11:41 BoxPlot_AllSample-RLE_CatData-TSS-TCGA-PK-A5HB-01A-11D-A29J
-rw-r--r-- 1 linux linux  60434 Jun 19 14:27 BoxPlot_AllSample-RLE_Diagram-BatchId.png
-rw-r--r-- 1 linux linux  59978 Jun 19 14:27 BoxPlot_AllSample-RLE_Diagram-PlateId.png
-rw-r--r-- 1 linux linux  60366 Jun 19 14:27 BoxPlot_AllSample-RLE_Diagram-ShipDate.png
-rw-r--r-- 1 linux linux  58667 Jun 19 14:27 BoxPlot_AllSample-RLE_Diagram-TSS.png
-rw-r--r-- 1 linux linux 819911 Jun 19 14:27 BoxPlot_AllSample-RLE_Histogram-BatchId.png
-rw-r--r-- 1 linux linux  45619 Jun 19 14:27 BoxPlot_AllSample-RLE_Histogram-BatchId.tsv
-rw-r--r-- 1 linux linux 819911 Jun 19 14:27 BoxPlot_AllSample-RLE_Histogram-PlateId.png
-rw-r--r-- 1 linux linux  45619 Jun 19 14:27 BoxPlot_AllSample-RLE_Histogram-PlateId.tsv
-rw-r--r-- 1 linux linux 819911 Jun 19 14:27 BoxPlot_AllSample-RLE_Histogram-ShipDate.png
-rw-r--r-- 1 linux linux  45619 Jun 19 14:27 BoxPlot_AllSample-RLE_Histogram-ShipDate.tsv
-rw-r--r-- 1 linux linux 819911 Jun 19 14:27 BoxPlot_AllSample-RLE_Histogram-TSS.png
-rw-r--r-- 1 linux linux  45619 Jun 19 14:27 BoxPlot_AllSample-RLE_Histogram-TSS.tsv
-rw-r--r-- 1 linux linux   4358 Jun 19 14:27 BoxPlot_AllSample-RLE_Legend-BatchId.png
-rw-r--r-- 1 linux linux   4378 Jun 19 14:27 BoxPlot_AllSample-RLE_Legend-PlateId.png
-rw-r--r-- 1 linux linux   4593 Jun 19 14:27 BoxPlot_AllSample-RLE_Legend-ShipDate.png
-rw-r--r-- 1 linux linux  13061 Jun 19 14:27 BoxPlot_AllSample-RLE_Legend-TSS.png
```

##Files

Example data may not match output from above.

##Annotations Files Looking at BoxPlot_AllSample-RLE_Annotations-TSS.tsv, we see it is a tab-delimited file, with two columns with the headers "key" nad "value". The first entry after that is the "Total-Data-Points", and then for each sample, we have the number of points available for that sample that are not NA. These two numbers will not always be equal, since some samples may have NAs for genes or probes where the other samples have values.

```
key value
Total-Data-Points   1250
Non-NA-Points-TCGA-OR-A5J1-01A-11D-A29J-05  1250
```

```
Non-NA-Points-TCGA-OR-A5J2-01A-11D-A29J-05  1250
Non-NA-Points-TCGA-OR-A5J3-01A-11D-A29J-05  1250
Non-NA-Points-TCGA-OR-A5J4-01A-11D-A29J-05  1250
Non-NA-Points-TCGA-OR-A5J5-01A-11D-A29J-05  1250
```

##BoxData Files Looking at BoxPlot_AllSample-RLE_BoxData-TSS.tsv, we see it is a tab delimited file with headers indicating the Id (sample) and the different parts of the boxplot. Subsequent rows give the box settings for each sample. NAs are possible in this data.

```
Id  LowerOutMax LowerOutMin LowerNotch  LowerWhisker    LowerHinge  Median  UpperHinge  UpperWhisker
TCGA-OR-A5J1-01A-11D-A29J-05    NA  NA  -0.020527642802858643   -0.8467955227772493 -0.4056428985980960
TCGA-OR-A5J2-01A-11D-A29J-05    NA  NA  -0.002911079872554134   -0.039930119705853896   -0.021222413369
TCGA-OR-A5J3-01A-11D-A29J-05    NA  NA  -0.035001758602725926   -0.3988124487830225 -0.3498757664560811
TCGA-OR-A5J4-01A-11D-A29J-05    NA  NA  -0.017185120892053492   -0.8247218460963763 -0.3183107584949181
TCGA-OR-A5J5-01A-11D-A29J-05    NA  NA  -0.03364073791133153    -0.8079754846584357 -0.644252206301912
TCGA-OR-A5J6-01A-31D-A29J-05    NA  NA  -0.0034328681890936023  -0.04187597080189986    -0.022337461512
TCGA-OR-A5J7-01A-11D-A29J-05    -0.865878813052012  -0.18821669173503153    -0.003889412141825995   -0.
```

##CatData Files If we look at BoxPlot_AllSample-RLE_CatData-TSS-TCGA-PK-A5HB-01A-11D-A29J-05.tsv, we see it is a tab-delimited file with "id" and "value" as headers. The id is a feature (in this case a gene, probe, location) combination and then the value from the data for that id. This is used to populate the violin plot with a subset of outliers, if any.

```
id  value
ADCY4-cg14287235-14-24804339    -0.7667974166463363
ASCL2-cg12499235-11-2293173 -0.7077020078715286
BAI1-cg09968723-8-143545789 -0.8074333452970504
BNC1-cg06523224-15-83953883 -0.7850694441252194
```

##Histogram Data Files Looking at BoxPlot_AllSample-RLE_Histogram-TSS.tsv, we see it is a tab-delimited file. The first row is headers, with "entry" and "size" being the first two, followed by pairs of headers of the form "xN" and "yN", where they are pairs of X,Y coordinates for plotting the histogram. The entry column is the sample id and the size entry is the number of X,Y pairs.

```
entry   size    x0  y0  x1  y1  x2  y2  x3  y3  x4  y4  x5  y5  x6  y6  x7  y7  x8  y8  x9  y9  x10 y10
TCGA-OR-A5J1-01A-11D-A29J-05    12  -0.8064387185053226 193.0   -0.7257251099614688 44.0    -0.6450115
TCGA-OR-A5J2-01A-11D-A29J-05    79  -0.033911616995144944   168.0   -0.02187461157372705    253.0   -0.
TCGA-OR-A5J3-01A-11D-A29J-05    7   -0.32982819164709853    520.0   -0.19185967737525045    68.0    -0.
```

##Diagram Here is a diagram generated from this code.

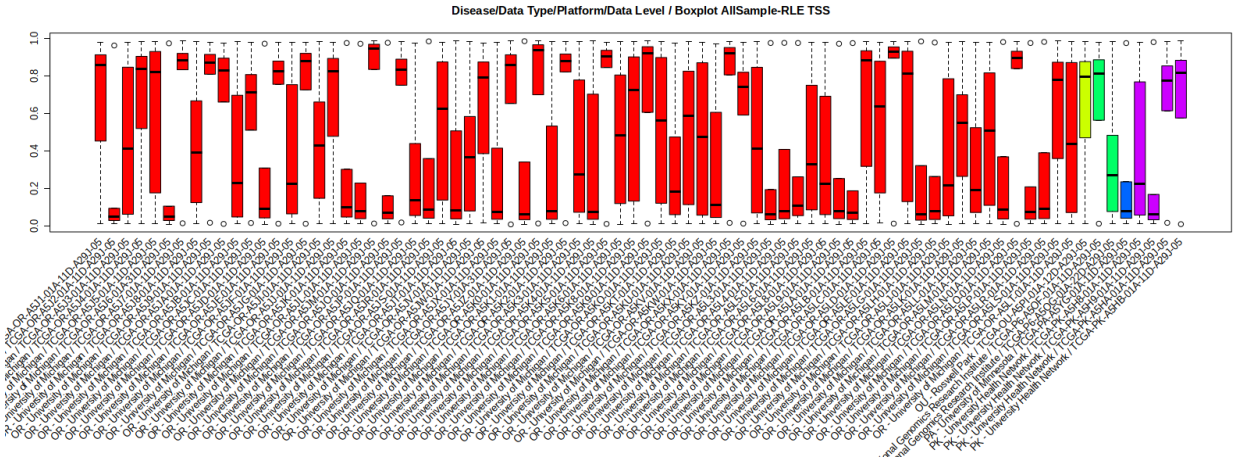

Figure 3: Boxplot All Samples RLE Output