Using MBatch Assessments: Boxplot_AllSamplesData_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_AllSamplesData_Structures is a function used to perform batch effects assessments using the boxplots on all samples without modification.

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 RLE Boxplots plot the value for each feature (genes or probes) for a sample, with the samples grouped and colored by batch. In this case, RLE is used to move the mean of each sample to zero on the vertical axis. 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_AllSamplesData_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()

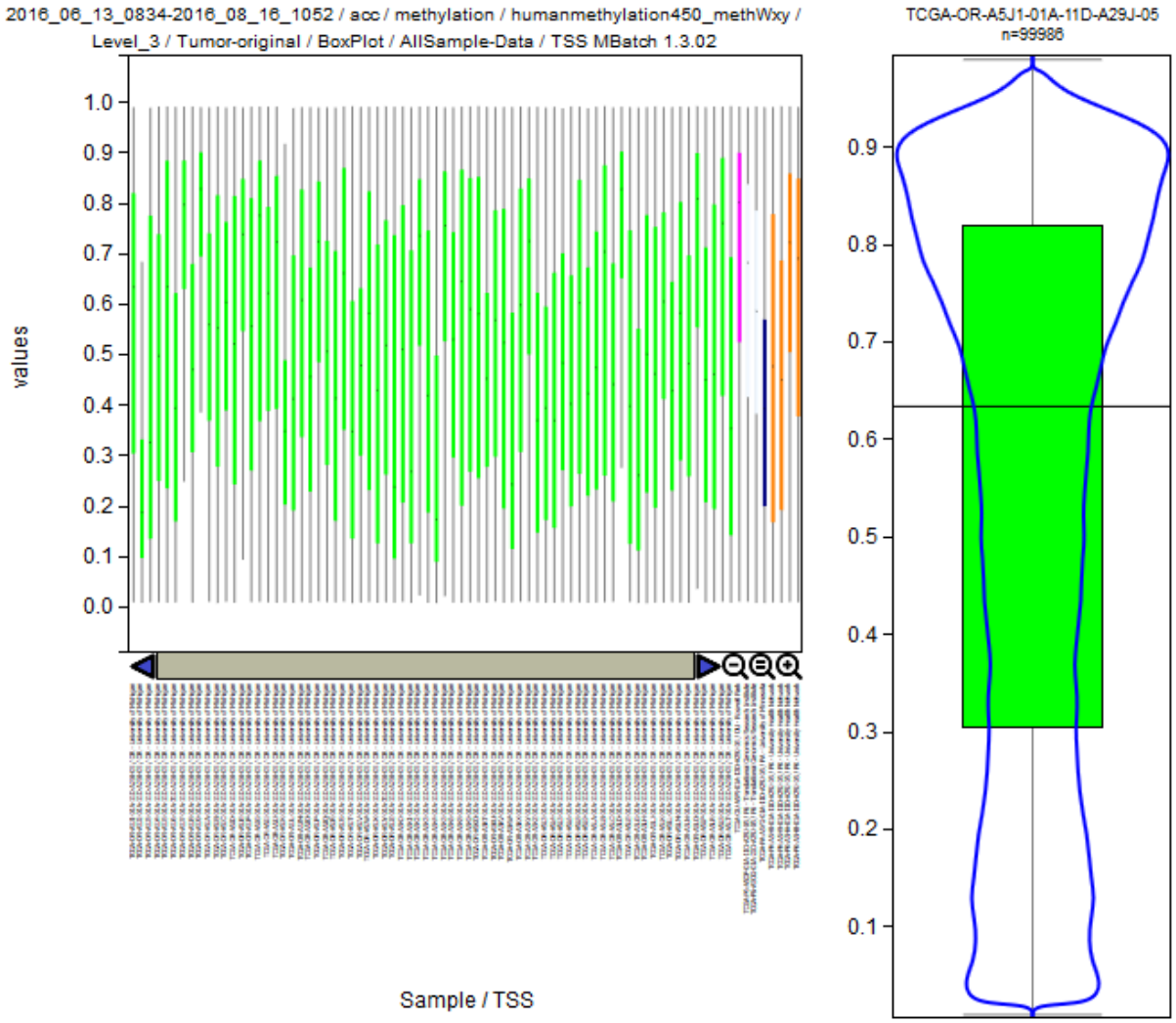

Figure 1: Dynamic Boxplot Example

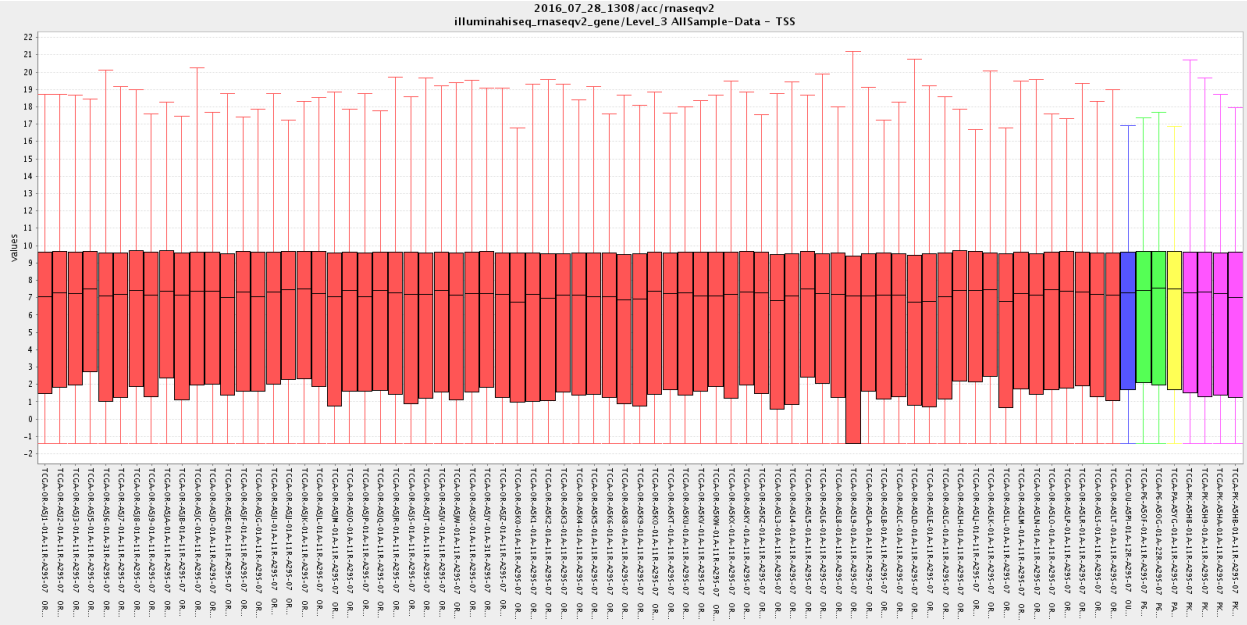

Figure 2: Static Boxplot Example

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.

##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_AllSamplesData_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_AllSamplesData_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_AllSamplesData_Structures(myData, "Disease/Data Type/Platform/Data Level", theOutputDir, list
                                    theDataVersion="DATA_2022-09-09-1600", theTestVersion="TEST_2022-10
}
```

```
## 2023 10 06 12:31:56.849 DEBUG qcprludev10 Changing LC_COLLATE to C for duration of run
## 2023 10 06 12:31:56.850 INFO qcprludev10 \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/
## 2023 10 06 12:31:56.850 INFO qcprludev10 Starting mbatchLoadFiles
## 2023 10 06 12:31:56.851 INFO qcprludev10 MBatch Version: BEA_VERSION_TIMESTAMP
## 2023 10 06 12:31:56.851 INFO qcprludev10 read batch file= /builds/BatchEffects_clean/BatchEffectsPacl
## 2023 10 06 12:31:56.852 INFO qcprludev10 read gene file= /builds/BatchEffects_clean/BatchEffectsPacka
## 2023 10 06 12:31:59.093 INFO qcprludev10 filter samples in batches using gene samples
## 2023 10 06 12:31:59.095 INFO qcprludev10 sort batches by gene file samples
## 2023 10 06 12:31:59.161 INFO qcprludev10 Finishing mbatchLoadFiles
## 2023 10 06 12:31:59.162 INFO qcprludev10 ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
## 2023 10 06 12:31:59.163 DEBUG qcprludev10 Changing LC_COLLATE to C for duration of run
## 2023 10 06 12:31:59.163 INFO qcprludev10 \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/
## 2023 10 06 12:31:59.163 INFO qcprludev10 mbatchTrimData Starting
## 2023 10 06 12:31:59.164 INFO qcprludev10 MBatch Version: BEA_VERSION_TIMESTAMP
## 2023 10 06 12:32:06.563 INFO qcprludev10 mbatchTrimData theMaxSize= 1e+05
## 2023 10 06 12:32:06.563 INFO qcprludev10 mbatchTrimData ncol(theMatrix)= 80
## 2023 10 06 12:32:06.564 INFO qcprludev10 mbatchTrimData nrow(theMatrix)= 1250
## 2023 10 06 12:32:06.564 INFO qcprludev10 mbatchTrimData Finishing
## 2023 10 06 12:32:06.565 INFO qcprludev10 ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
## 2023 10 06 12:32:06.566 DEBUG qcprludev10 Changing LC_COLLATE to C for duration of run
## 2023 10 06 12:32:06.566 INFO qcprludev10 \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/
## 2023 10 06 12:32:06.567 INFO qcprludev10 mbatchFilterData Starting
## 2023 10 06 12:32:06.567 INFO qcprludev10 MBatch Version: BEA_VERSION_TIMESTAMP
## 2023 10 06 12:32:06.567 DEBUG qcprludev10 rows pre filter  1250
## 2023 10 06 12:32:06.777 DEBUG qcprludev10 rows post filter  1250
## 2023 10 06 12:32:06.777 DEBUG qcprludev10 mbatchFilterData Prefilter, gene data had  1250  while pos
## 2023 10 06 12:32:06.778 DEBUG qcprludev10 mbatchFilterData Prefilter, batch data had  80  while post
## 2023 10 06 12:32:06.779 INFO qcprludev10 mbatchFilterData Finishing
## 2023 10 06 12:32:06.779 INFO qcprludev10 ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
## 2023 10 06 12:32:06.780 DEBUG qcprludev10 Changing LC_COLLATE to C for duration of run
## 2023 10 06 12:32:06.780 INFO qcprludev10 \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/
## 2023 10 06 12:32:06.781 INFO qcprludev10 mbatchTrimData Starting
```

```
## 2023 10 06 12:32:06.781 INFO qcprludev10 MBatch Version: BEA_VERSION_TIMESTAMP
## 2023 10 06 12:32:06.782 INFO qcprludev10 mbatchTrimData theMaxSize= 1600000
## 2023 10 06 12:32:06.782 INFO qcprludev10 mbatchTrimData ncol(theMatrix)= 80
## 2023 10 06 12:32:06.783 INFO qcprludev10 mbatchTrimData nrow(theMatrix)= 1250
## 2023 10 06 12:32:06.783 INFO qcprludev10 mbatchTrimData Finishing
## 2023 10 06 12:32:06.784 INFO qcprludev10 ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
## 2023 10 06 12:32:06.785 DEBUG qcprludev10 dim(theMatrixGeneData) 1250, dim(theMatrixGeneData) 80
## 2023 10 06 12:32:06.785 DEBUG qcprludev10 length(colnames(theMatrixGeneData)) 80
## 2023 10 06 12:32:06.785 DEBUG qcprludev10 length(rownames(theMatrixGeneData)) 1250
## 2023 10 06 12:32:06.786 DEBUG qcprludev10 dim(theDataframeBatchData) 80, dim(theDataframeBatchData) 5
## 2023 10 06 12:32:06.786 DEBUG qcprludev10 length(names(theDataframeBatchData)) 5
## 2023 10 06 12:32:06.787 DEBUG qcprludev10 batchTypeName =  BatchId
## 2023 10 06 12:32:06.787 DEBUG qcprludev10 theBatchType= BatchId
## 2023 10 06 12:32:06.789 DEBUG qcprludev10 calcAndWriteBoxplot - theBoxDataFile= /BEA/BatchEffectsPacl
## 2023 10 06 12:32:06.789 DEBUG qcprludev10 calcAndWriteBoxplot - dim(theData)[1]= 1250
## 2023 10 06 12:32:06.790 DEBUG qcprludev10 calcAndWriteBoxplot - dim(theData)[2]= 80
## 2023 10 06 12:32:06.791 DEBUG qcprludev10 writeTitleFile - pre title  Disease/Data Type/Platform/Data
## 2023 10 06 12:32:06.792 DEBUG qcprludev10 writeTitleFile - theTitle  Disease / Data Type / Platform /
## 2023 10 06 12:32:06.793 DEBUG qcprludev10 writeTitleFile - titleFile  /BEA/BatchEffectsPackage_data/t
## 2023 10 06 12:32:06.794 DEBUG qcprludev10 calcAndWriteBoxplot - before calcAndWriteBoxDataFile
## 2023 10 06 12:32:06.795 DEBUG qcprludev10 calcAndWriteBoxDataFile theFile= /BEA/BatchEffectsPackage_d
## 2023 10 06 12:32:06.795 DEBUG qcprludev10 calcAndWriteBoxDataFile thePngFile= /BEA/BatchEffectsPackag
## 2023 10 06 12:32:06.833 DEBUG qcprludev10 calcAndWriteBoxDataFile CairoPNG= /BEA/BatchEffectsPackage_
## 2023 10 06 12:32:06.834 DEBUG qcprludev10 calcAndWriteBoxDataFile call boxplot

## 2023 10 06 12:32:06.862 DEBUG qcprludev10 calcAndWriteBoxDataFile call text
## 2023 10 06 12:32:06.870 DEBUG qcprludev10 calcAndWriteBoxDataFile done
## 2023 10 06 12:32:06.911 DEBUG qcprludev10 calcAndWriteBoxplot - after calcAndWriteBoxDataFile
## 2023 10 06 12:32:06.911 DEBUG qcprludev10 calcAndWriteBoxplot - before calcAndWriteHistogramFile
## 2023 10 06 12:32:06.912 DEBUG qcprludev10 calcAndWriteHistogramFile  /BEA/BatchEffectsPackage_data/te
## 2023 10 06 12:32:06.996 DEBUG qcprludev10 calcAndWriteBoxplot - after calcAndWriteHistogramFile
## 2023 10 06 12:32:06.997 DEBUG qcprludev10 calcAndWriteBoxplot - before calcAndWriteAnnotationsFile
## 2023 10 06 12:32:06.998 DEBUG qcprludev10 calcAndWriteAnnotationsFile theFile= /BEA/BatchEffectsPacka
## 2023 10 06 12:32:07.012 DEBUG qcprludev10 calcAndWriteBoxplot - after calcAndWriteAnnotationsFile
## 2023 10 06 12:32:07.012 DEBUG qcprludev10 batchTypeName =  PlateId
## 2023 10 06 12:32:07.013 DEBUG qcprludev10 theBatchType= PlateId
## 2023 10 06 12:32:07.013 DEBUG qcprludev10 calcAndWriteBoxplot - theBoxDataFile= /BEA/BatchEffectsPacl
## 2023 10 06 12:32:07.014 DEBUG qcprludev10 calcAndWriteBoxplot - dim(theData)[1]= 1250
## 2023 10 06 12:32:07.014 DEBUG qcprludev10 calcAndWriteBoxplot - dim(theData)[2]= 80
## 2023 10 06 12:32:07.015 DEBUG qcprludev10 writeTitleFile - pre title  Disease/Data Type/Platform/Data
## 2023 10 06 12:32:07.016 DEBUG qcprludev10 writeTitleFile - theTitle  Disease / Data Type / Platform /
## 2023 10 06 12:32:07.017 DEBUG qcprludev10 writeTitleFile - titleFile  /BEA/BatchEffectsPackage_data/t
## 2023 10 06 12:32:07.018 DEBUG qcprludev10 calcAndWriteBoxplot - before calcAndWriteBoxDataFile
## 2023 10 06 12:32:07.018 DEBUG qcprludev10 calcAndWriteBoxDataFile theFile= /BEA/BatchEffectsPackage_d
## 2023 10 06 12:32:07.019 DEBUG qcprludev10 calcAndWriteBoxDataFile thePngFile= /BEA/BatchEffectsPackag
## 2023 10 06 12:32:07.053 DEBUG qcprludev10 calcAndWriteBoxDataFile CairoPNG= /BEA/BatchEffectsPackage_
## 2023 10 06 12:32:07.055 DEBUG qcprludev10 calcAndWriteBoxDataFile call boxplot

## 2023 10 06 12:32:07.082 DEBUG qcprludev10 calcAndWriteBoxDataFile call text
## 2023 10 06 12:32:07.090 DEBUG qcprludev10 calcAndWriteBoxDataFile done
## 2023 10 06 12:32:07.131 DEBUG qcprludev10 calcAndWriteBoxplot - after calcAndWriteBoxDataFile
## 2023 10 06 12:32:07.132 DEBUG qcprludev10 calcAndWriteBoxplot - before calcAndWriteHistogramFile
## 2023 10 06 12:32:07.132 DEBUG qcprludev10 calcAndWriteHistogramFile  /BEA/BatchEffectsPackage_data/te
## 2023 10 06 12:32:07.217 DEBUG qcprludev10 calcAndWriteBoxplot - after calcAndWriteHistogramFile
## 2023 10 06 12:32:07.218 DEBUG qcprludev10 calcAndWriteBoxplot - before calcAndWriteAnnotationsFile
```

```
## 2023 10 06 12:32:07.218 DEBUG qcprludev10 calcAndWriteAnnotationsFile theFile= /BEA/BatchEffectsPacka
## 2023 10 06 12:32:07.231 DEBUG qcprludev10 calcAndWriteBoxplot - after calcAndWriteAnnotationsFile
## 2023 10 06 12:32:07.232 DEBUG qcprludev10 batchTypeName =  ShipDate
## 2023 10 06 12:32:07.232 DEBUG qcprludev10 theBatchType= ShipDate
## 2023 10 06 12:32:07.233 DEBUG qcprludev10 calcAndWriteBoxplot - theBoxDataFile= /BEA/BatchEffectsPacl
## 2023 10 06 12:32:07.233 DEBUG qcprludev10 calcAndWriteBoxplot - dim(theData)[1]= 1250
## 2023 10 06 12:32:07.234 DEBUG qcprludev10 calcAndWriteBoxplot - dim(theData)[2]= 80
## 2023 10 06 12:32:07.234 DEBUG qcprludev10 writeTitleFile - pre title  Disease/Data Type/Platform/Data
## 2023 10 06 12:32:07.235 DEBUG qcprludev10 writeTitleFile - theTitle  Disease / Data Type / Platform /
## 2023 10 06 12:32:07.235 DEBUG qcprludev10 writeTitleFile - titleFile  /BEA/BatchEffectsPackage_data/t
## 2023 10 06 12:32:07.236 DEBUG qcprludev10 calcAndWriteBoxplot - before calcAndWriteBoxDataFile
## 2023 10 06 12:32:07.236 DEBUG qcprludev10 calcAndWriteBoxDataFile theFile= /BEA/BatchEffectsPackage_d
## 2023 10 06 12:32:07.236 DEBUG qcprludev10 calcAndWriteBoxDataFile thePngFile= /BEA/BatchEffectsPackag
## 2023 10 06 12:32:07.271 DEBUG qcprludev10 calcAndWriteBoxDataFile CairoPNG= /BEA/BatchEffectsPackage_
## 2023 10 06 12:32:07.272 DEBUG qcprludev10 calcAndWriteBoxDataFile call boxplot

## 2023 10 06 12:32:07.316 DEBUG qcprludev10 calcAndWriteBoxDataFile call text
## 2023 10 06 12:32:07.326 DEBUG qcprludev10 calcAndWriteBoxDataFile done
## 2023 10 06 12:32:07.367 DEBUG qcprludev10 calcAndWriteBoxplot - after calcAndWriteBoxDataFile
## 2023 10 06 12:32:07.368 DEBUG qcprludev10 calcAndWriteBoxplot - before calcAndWriteHistogramFile
## 2023 10 06 12:32:07.368 DEBUG qcprludev10 calcAndWriteHistogramFile  /BEA/BatchEffectsPackage_data/te
## 2023 10 06 12:32:07.451 DEBUG qcprludev10 calcAndWriteBoxplot - after calcAndWriteHistogramFile
## 2023 10 06 12:32:07.452 DEBUG qcprludev10 calcAndWriteBoxplot - before calcAndWriteAnnotationsFile
## 2023 10 06 12:32:07.452 DEBUG qcprludev10 calcAndWriteAnnotationsFile theFile= /BEA/BatchEffectsPacka
## 2023 10 06 12:32:07.465 DEBUG qcprludev10 calcAndWriteBoxplot - after calcAndWriteAnnotationsFile
## 2023 10 06 12:32:07.466 DEBUG qcprludev10 batchTypeName =  TSS
## 2023 10 06 12:32:07.466 DEBUG qcprludev10 theBatchType= TSS
## 2023 10 06 12:32:07.467 DEBUG qcprludev10 calcAndWriteBoxplot - theBoxDataFile= /BEA/BatchEffectsPacl
## 2023 10 06 12:32:07.467 DEBUG qcprludev10 calcAndWriteBoxplot - dim(theData)[1]= 1250
## 2023 10 06 12:32:07.468 DEBUG qcprludev10 calcAndWriteBoxplot - dim(theData)[2]= 80
## 2023 10 06 12:32:07.468 DEBUG qcprludev10 writeTitleFile - pre title  Disease/Data Type/Platform/Data
## 2023 10 06 12:32:07.469 DEBUG qcprludev10 writeTitleFile - theTitle  Disease / Data Type / Platform /
## 2023 10 06 12:32:07.469 DEBUG qcprludev10 writeTitleFile - titleFile  /BEA/BatchEffectsPackage_data/t
## 2023 10 06 12:32:07.470 DEBUG qcprludev10 calcAndWriteBoxplot - before calcAndWriteBoxDataFile
## 2023 10 06 12:32:07.470 DEBUG qcprludev10 calcAndWriteBoxDataFile theFile= /BEA/BatchEffectsPackage_d
## 2023 10 06 12:32:07.471 DEBUG qcprludev10 calcAndWriteBoxDataFile thePngFile= /BEA/BatchEffectsPackag
## 2023 10 06 12:32:07.504 DEBUG qcprludev10 calcAndWriteBoxDataFile CairoPNG= /BEA/BatchEffectsPackage_
## 2023 10 06 12:32:07.506 DEBUG qcprludev10 calcAndWriteBoxDataFile call boxplot

## 2023 10 06 12:32:07.533 DEBUG qcprludev10 calcAndWriteBoxDataFile call text
## 2023 10 06 12:32:07.548 DEBUG qcprludev10 calcAndWriteBoxDataFile done
## 2023 10 06 12:32:07.591 DEBUG qcprludev10 calcAndWriteBoxplot - after calcAndWriteBoxDataFile
## 2023 10 06 12:32:07.591 DEBUG qcprludev10 calcAndWriteBoxplot - before calcAndWriteHistogramFile
## 2023 10 06 12:32:07.592 DEBUG qcprludev10 calcAndWriteHistogramFile  /BEA/BatchEffectsPackage_data/te
## 2023 10 06 12:32:07.676 DEBUG qcprludev10 calcAndWriteBoxplot - after calcAndWriteHistogramFile
## 2023 10 06 12:32:07.677 DEBUG qcprludev10 calcAndWriteBoxplot - before calcAndWriteAnnotationsFile
## 2023 10 06 12:32:07.677 DEBUG qcprludev10 calcAndWriteAnnotationsFile theFile= /BEA/BatchEffectsPacka
## 2023 10 06 12:32:07.690 DEBUG qcprludev10 calcAndWriteBoxplot - after calcAndWriteAnnotationsFile

## writeBatchDataTsvForBoxplot

## /BEA/BatchEffectsPackage_data/testing_dynamic/MBatch/Boxplot_AllSamplesData_Structures/AllSample-Data

## 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_AllSamplesData_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_AllSamplesData_Structures/AllSample-Data$ ls -l
total 6228
-rw-r--r-- 1 linux linux   3873 Jun 19 15:12 BoxPlot_AllSample-Data_Annotations-BatchId.tsv
-rw-r--r-- 1 linux linux   3873 Jun 19 15:13 BoxPlot_AllSample-Data_Annotations-PlateId.tsv
-rw-r--r-- 1 linux linux   3873 Jun 19 15:13 BoxPlot_AllSample-Data_Annotations-ShipDate.tsv
-rw-r--r-- 1 linux linux   3873 Jun 19 15:13 BoxPlot_AllSample-Data_Annotations-TSS.tsv
-rw-r--r-- 1 linux linux  15072 Jun 19 15:12 BoxPlot_AllSample-Data_BoxData-BatchId.tsv
-rw-r--r-- 1 linux linux  15072 Jun 19 15:13 BoxPlot_AllSample-Data_BoxData-PlateId.tsv
-rw-r--r-- 1 linux linux  15072 Jun 19 15:13 BoxPlot_AllSample-Data_BoxData-ShipDate.tsv
-rw-r--r-- 1 linux linux  15072 Jun 19 15:13 BoxPlot_AllSample-Data_BoxData-TSS.tsv
-rw-r--r-- 1 linux linux      9 Jun 19 15:12 BoxPlot_AllSample-Data_CatData-BatchId-TCGA-OR-A5J1-01A-11
-rw-r--r-- 1 linux linux   7552 Jun 19 15:12 BoxPlot_AllSample-Data_CatData-BatchId-TCGA-OR-A5J2-01A-11
#snipped out "CatData" files for each sample for each batch type
-rw-r--r-- 1 linux linux   6469 Jun 19 15:13 BoxPlot_AllSample-Data_CatData-TSS-TCGA-PK-A5HA-01A-11D-A2
-rw-r--r-- 1 linux linux   5461 Jun 19 15:13 BoxPlot_AllSample-Data_CatData-TSS-TCGA-PK-A5HB-01A-11D-A2
-rw-r--r-- 1 linux linux  70954 Jun 19 15:12 BoxPlot_AllSample-Data_Diagram-BatchId.png
-rw-r--r-- 1 linux linux  70493 Jun 19 15:13 BoxPlot_AllSample-Data_Diagram-PlateId.png
-rw-r--r-- 1 linux linux  70713 Jun 19 15:13 BoxPlot_AllSample-Data_Diagram-ShipDate.png
-rw-r--r-- 1 linux linux  66492 Jun 19 15:13 BoxPlot_AllSample-Data_Diagram-TSS.png
-rw-r--r-- 1 linux linux 916490 Jun 19 15:13 BoxPlot_AllSample-Data_Histogram-BatchId.png
-rw-r--r-- 1 linux linux  44432 Jun 19 15:12 BoxPlot_AllSample-Data_Histogram-BatchId.tsv
-rw-r--r-- 1 linux linux 916490 Jun 19 15:13 BoxPlot_AllSample-Data_Histogram-PlateId.png
-rw-r--r-- 1 linux linux  44432 Jun 19 15:13 BoxPlot_AllSample-Data_Histogram-PlateId.tsv
-rw-r--r-- 1 linux linux 916490 Jun 19 15:13 BoxPlot_AllSample-Data_Histogram-ShipDate.png
-rw-r--r-- 1 linux linux  44432 Jun 19 15:13 BoxPlot_AllSample-Data_Histogram-ShipDate.tsv
-rw-r--r-- 1 linux linux 916490 Jun 19 15:13 BoxPlot_AllSample-Data_Histogram-TSS.png
-rw-r--r-- 1 linux linux  44432 Jun 19 15:13 BoxPlot_AllSample-Data_Histogram-TSS.tsv
-rw-r--r-- 1 linux linux   4431 Jun 19 15:12 BoxPlot_AllSample-Data_Legend-BatchId.png
-rw-r--r-- 1 linux linux   4450 Jun 19 15:13 BoxPlot_AllSample-Data_Legend-PlateId.png
-rw-r--r-- 1 linux linux   4521 Jun 19 15:13 BoxPlot_AllSample-Data_Legend-ShipDate.png
-rw-r--r-- 1 linux linux  13135 Jun 19 15:13 BoxPlot_AllSample-Data_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.

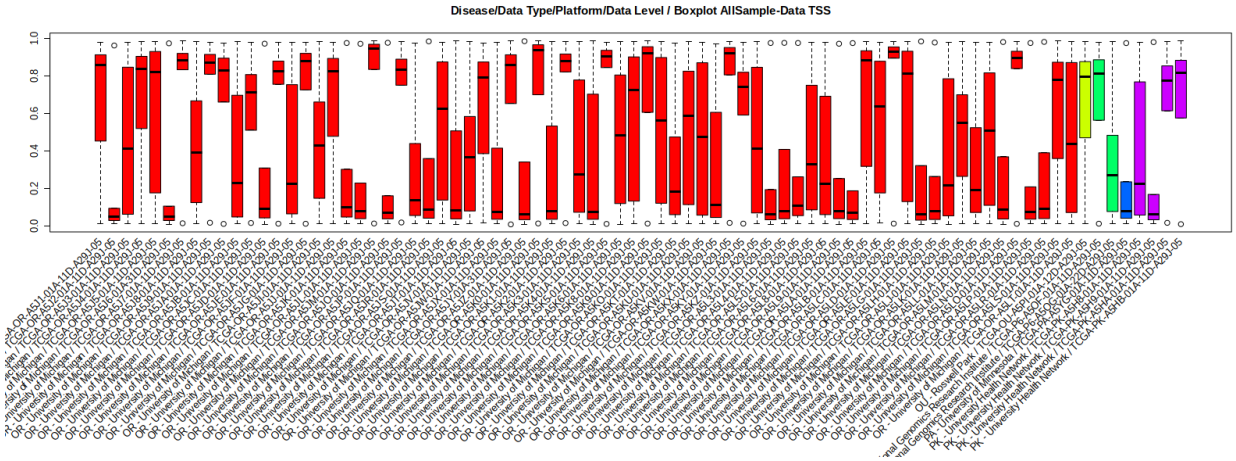

Figure 3: Boxplot All Samples Data Output