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

Boxplot_AllSamplesData_Structures is a function used to perform batch effects assessments using the boxplots on all samples without modification.

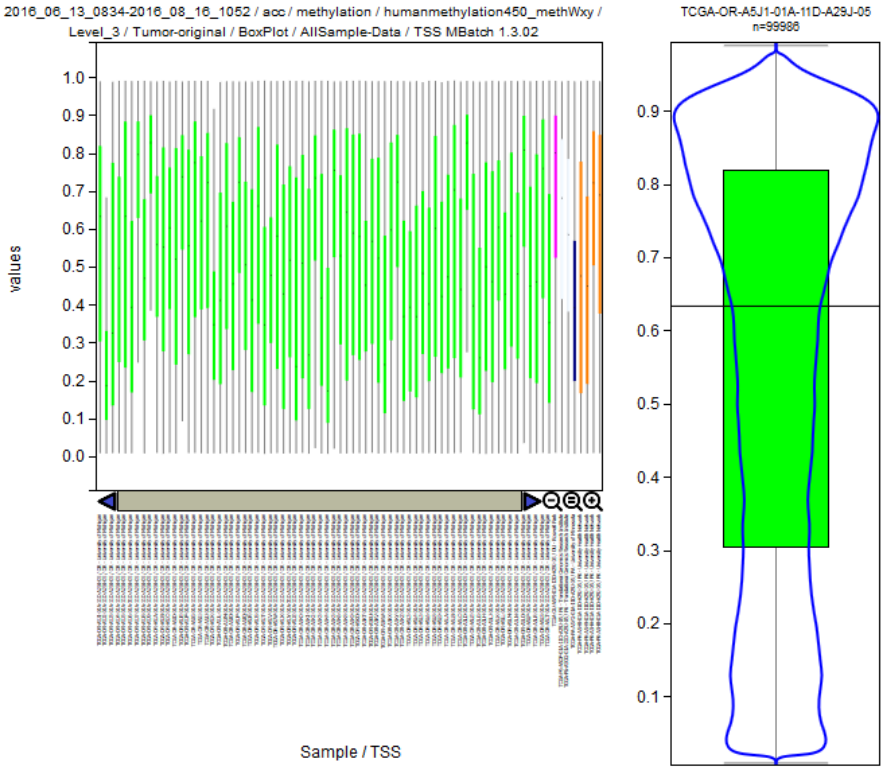
Output

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.

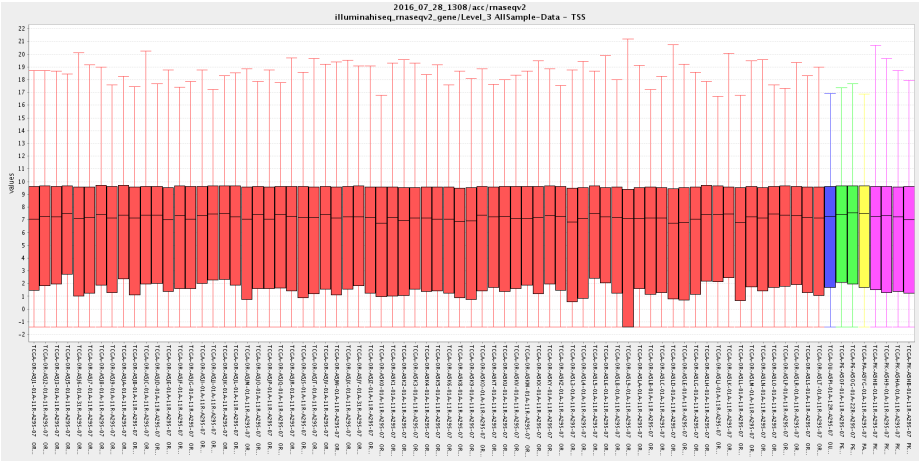
Graphical output is a set of boxplots where each boxplot (also called a box and whisper plot) represent a single sample. For datasets with many samples, the static PNG may be so dense as to be unusable.

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.



Usage

Boxplot_AllSamplesData_Structures(theData, theTitle, theOutputPath, theBatchTypeAndValuePairsToRemove, theBatchTypeAndValuePairsToKeep, theMaxGeneCount=20000)

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.

##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.

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.

```
{
  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, "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", theOutp
}
```

```
## 2020 11 18 16:18:43.046 DEBUG ab7c64738d52 Changing LC_COLLATE to C for duration of run
## 2020 11 18 16:18:43.047 INFO ab7c64738d52 \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/
## 2020 11 18 16:18:43.047 INFO ab7c64738d52 Starting mbatchLoadFiles
## 2020 11 18 16:18:43.047 INFO ab7c64738d52 MBatch Version: BEA_VERSION_TIMESTAMP
## 2020 11 18 16:18:43.047 INFO ab7c64738d52 read batch file= /builds/BatchEffects_clean/Bat
## 2020 11 18 16:18:43.048 INFO ab7c64738d52 read gene file= /builds/BatchEffects_clean/Batc
## 2020 11 18 16:18:45.158 INFO ab7c64738d52 filter samples in batches using gene samples
## 2020 11 18 16:18:45.160 INFO ab7c64738d52 sort batches by gene file samples
## 2020 11 18 16:18:45.274 INFO ab7c64738d52 Finishing mbatchLoadFiles
## 2020 11 18 16:18:45.274 INFO ab7c64738d52 ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
## 2020 11 18 16:18:45.275 DEBUG ab7c64738d52 Changing LC_COLLATE to C for duration of run
## 2020 11 18 16:18:45.275 INFO ab7c64738d52 \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/
## 2020 11 18 16:18:45.275 INFO ab7c64738d52 mbatchTrimData Starting
## 2020 11 18 16:18:45.275 INFO ab7c64738d52 MBatch Version: BEA_VERSION_TIMESTAMP
## 2020 11 18 16:18:52.953 INFO ab7c64738d52 mbatchTrimData theMaxSize= 1e+05
## 2020 11 18 16:18:52.953 INFO ab7c64738d52 mbatchTrimData ncol(theMatrix)= 80
## 2020 11 18 16:18:52.954 INFO ab7c64738d52 mbatchTrimData nrow(theMatrix)= 1250
## 2020 11 18 16:18:52.954 INFO ab7c64738d52 mbatchTrimData Finishing
## 2020 11 18 16:18:52.954 INFO ab7c64738d52 ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
## 2020 11 18 16:18:52.955 DEBUG ab7c64738d52 Changing LC_COLLATE to C for duration of run
## 2020 11 18 16:18:52.955 INFO ab7c64738d52 \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/
```

```
## 2020 11 18 16:18:52.956 INFO ab7c64738d52 mbatchFilterData Starting
## 2020 11 18 16:18:52.956 INFO ab7c64738d52 MBatch Version: BEA_VERSION_TIMESTAMP
## 2020 11 18 16:18:52.956 DEBUG ab7c64738d52 rows pre filter  1250
## 2020 11 18 16:18:53.165 DEBUG ab7c64738d52 rows post filter  1250
## 2020 11 18 16:18:53.165 DEBUG ab7c64738d52 mbatchFilterData Prefilter, gene data had  125
## 2020 11 18 16:18:53.166 DEBUG ab7c64738d52 mbatchFilterData Prefilter, batch data had  80
## 2020 11 18 16:18:53.166 INFO ab7c64738d52 mbatchFilterData Finishing
## 2020 11 18 16:18:53.166 INFO ab7c64738d52 ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
## 2020 11 18 16:18:53.167 DEBUG ab7c64738d52 Changing LC_COLLATE to C for duration of run
## 2020 11 18 16:18:53.167 INFO ab7c64738d52 \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/
## 2020 11 18 16:18:53.167 INFO ab7c64738d52 mbatchTrimData Starting
## 2020 11 18 16:18:53.167 INFO ab7c64738d52 MBatch Version: BEA_VERSION_TIMESTAMP
## 2020 11 18 16:18:53.167 INFO ab7c64738d52 mbatchTrimData theMaxSize= 1600000
## 2020 11 18 16:18:53.168 INFO ab7c64738d52 mbatchTrimData ncol(theMatrix)= 80
## 2020 11 18 16:18:53.168 INFO ab7c64738d52 mbatchTrimData nrow(theMatrix)= 1250
## 2020 11 18 16:18:53.168 INFO ab7c64738d52 mbatchTrimData Finishing
## 2020 11 18 16:18:53.168 INFO ab7c64738d52 ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
## 2020 11 18 16:18:53.169 DEBUG ab7c64738d52 checkCreateDir:  /builds/BatchEffects_clean/Ba
## 2020 11 18 16:18:53.169 DEBUG ab7c64738d52 dim(theMatrixGeneData) 1250, dim(theMatrixGene
## 2020 11 18 16:18:53.169 DEBUG ab7c64738d52 length(colnames(theMatrixGeneData)) 80
## 2020 11 18 16:18:53.169 DEBUG ab7c64738d52 length(rownames(theMatrixGeneData)) 1250
## 2020 11 18 16:18:53.169 DEBUG ab7c64738d52 dim(theDataframeBatchData) 80, dim(theDatafram
## 2020 11 18 16:18:53.170 DEBUG ab7c64738d52 length(names(theDataframeBatchData)) 5
## 2020 11 18 16:18:53.170 DEBUG ab7c64738d52 batchTypeName =  BatchId
## 2020 11 18 16:18:53.170 DEBUG ab7c64738d52 theBatchType= BatchId
## 2020 11 18 16:18:53.170 DEBUG ab7c64738d52 calcAndWriteBoxplot - theBoxDataFile= /builds/
## 2020 11 18 16:18:53.171 DEBUG ab7c64738d52 calcAndWriteBoxplot - theMedian= 0
## 2020 11 18 16:18:53.171 DEBUG ab7c64738d52 calcAndWriteBoxplot - dim(theData)[1]= 1250
## 2020 11 18 16:18:53.171 DEBUG ab7c64738d52 calcAndWriteBoxplot - dim(theData)[2]= 80
## 2020 11 18 16:18:53.171 DEBUG ab7c64738d52 checkCreateDir:  /builds/BatchEffects_clean/Ba
## 2020 11 18 16:18:53.172 DEBUG ab7c64738d52 calcAndWriteBoxplot - before calcAndWriteBoxDa
## 2020 11 18 16:18:53.172 DEBUG ab7c64738d52 calcAndWriteBoxDataFile theFile= /builds/Batch
## 2020 11 18 16:18:53.172 DEBUG ab7c64738d52 calcAndWriteBoxDataFile thePngFile= /builds/Ba
## 2020 11 18 16:18:53.200 DEBUG ab7c64738d52 calcAndWriteBoxDataFile CairoPNG= /builds/Batc
## 2020 11 18 16:18:53.202 DEBUG ab7c64738d52 calcAndWriteBoxDataFile call boxplot
## 2020 11 18 16:18:53.247 DEBUG ab7c64738d52 calcAndWriteBoxDataFile call text
## 2020 11 18 16:18:53.257 DEBUG ab7c64738d52 calcAndWriteBoxDataFile done
## 2020 11 18 16:18:53.303 DEBUG ab7c64738d52 calcAndWriteBoxplot - after calcAndWriteBoxDat
## 2020 11 18 16:18:53.303 DEBUG ab7c64738d52 calcAndWriteBoxplot - before calcAndWriteHisto
## 2020 11 18 16:18:53.303 DEBUG ab7c64738d52 calcAndWriteHistogramFile  /builds/BatchEffect
## 2020 11 18 16:18:53.382 DEBUG ab7c64738d52 calcAndWriteBoxplot - after calcAndWriteHistog
## 2020 11 18 16:18:53.382 DEBUG ab7c64738d52 calcAndWriteBoxplot - before calcAndWriteAnnot
## 2020 11 18 16:18:53.382 DEBUG ab7c64738d52 calcAndWriteAnnotationsFile theFile= /builds/B
## 2020 11 18 16:18:53.388 DEBUG ab7c64738d52 calcAndWriteBoxplot - after calcAndWriteAnnota
## 2020 11 18 16:18:53.388 DEBUG ab7c64738d52 batchTypeName =  PlateId
```

```
## 2020 11 18 16:18:53.389 DEBUG ab7c64738d52 theBatchType= PlateId
## 2020 11 18 16:18:53.389 DEBUG ab7c64738d52 calcAndWriteBoxplot - theBoxDataFile= /builds/
## 2020 11 18 16:18:53.389 DEBUG ab7c64738d52 calcAndWriteBoxplot - theMedian= 0
## 2020 11 18 16:18:53.389 DEBUG ab7c64738d52 calcAndWriteBoxplot - dim(theData)[1]= 1250
## 2020 11 18 16:18:53.390 DEBUG ab7c64738d52 calcAndWriteBoxplot - dim(theData)[2]= 80
## 2020 11 18 16:18:53.390 DEBUG ab7c64738d52 checkCreateDir:  /builds/BatchEffects_clean/Ba
## 2020 11 18 16:18:53.390 DEBUG ab7c64738d52 calcAndWriteBoxplot - before calcAndWriteBoxDa
## 2020 11 18 16:18:53.390 DEBUG ab7c64738d52 calcAndWriteBoxDataFile theFile= /builds/Batch
## 2020 11 18 16:18:53.390 DEBUG ab7c64738d52 calcAndWriteBoxDataFile thePngFile= /builds/Ba
## 2020 11 18 16:18:53.418 DEBUG ab7c64738d52 calcAndWriteBoxDataFile CairoPNG= /builds/Batc
## 2020 11 18 16:18:53.419 DEBUG ab7c64738d52 calcAndWriteBoxDataFile call boxplot

## 2020 11 18 16:18:53.448 DEBUG ab7c64738d52 calcAndWriteBoxDataFile call text
## 2020 11 18 16:18:53.459 DEBUG ab7c64738d52 calcAndWriteBoxDataFile done
## 2020 11 18 16:18:53.505 DEBUG ab7c64738d52 calcAndWriteBoxplot - after calcAndWriteBoxDat
## 2020 11 18 16:18:53.505 DEBUG ab7c64738d52 calcAndWriteBoxplot - before calcAndWriteHisto
## 2020 11 18 16:18:53.505 DEBUG ab7c64738d52 calcAndWriteHistogramFile  /builds/BatchEffect
## 2020 11 18 16:18:53.585 DEBUG ab7c64738d52 calcAndWriteBoxplot - after calcAndWriteHistog
## 2020 11 18 16:18:53.586 DEBUG ab7c64738d52 calcAndWriteBoxplot - before calcAndWriteAnnot
## 2020 11 18 16:18:53.586 DEBUG ab7c64738d52 calcAndWriteAnnotationsFile theFile= /builds/B
## 2020 11 18 16:18:53.592 DEBUG ab7c64738d52 calcAndWriteBoxplot - after calcAndWriteAnnota
## 2020 11 18 16:18:53.592 DEBUG ab7c64738d52 batchTypeName =  ShipDate
## 2020 11 18 16:18:53.592 DEBUG ab7c64738d52 theBatchType= ShipDate
## 2020 11 18 16:18:53.593 DEBUG ab7c64738d52 calcAndWriteBoxplot - theBoxDataFile= /builds/
## 2020 11 18 16:18:53.593 DEBUG ab7c64738d52 calcAndWriteBoxplot - theMedian= 0
## 2020 11 18 16:18:53.593 DEBUG ab7c64738d52 calcAndWriteBoxplot - dim(theData)[1]= 1250
## 2020 11 18 16:18:53.593 DEBUG ab7c64738d52 calcAndWriteBoxplot - dim(theData)[2]= 80
## 2020 11 18 16:18:53.593 DEBUG ab7c64738d52 checkCreateDir:  /builds/BatchEffects_clean/Ba
## 2020 11 18 16:18:53.594 DEBUG ab7c64738d52 calcAndWriteBoxplot - before calcAndWriteBoxDa
## 2020 11 18 16:18:53.594 DEBUG ab7c64738d52 calcAndWriteBoxDataFile theFile= /builds/Batch
## 2020 11 18 16:18:53.594 DEBUG ab7c64738d52 calcAndWriteBoxDataFile thePngFile= /builds/Ba
## 2020 11 18 16:18:53.622 DEBUG ab7c64738d52 calcAndWriteBoxDataFile CairoPNG= /builds/Batc
## 2020 11 18 16:18:53.623 DEBUG ab7c64738d52 calcAndWriteBoxDataFile call boxplot

## 2020 11 18 16:18:53.651 DEBUG ab7c64738d52 calcAndWriteBoxDataFile call text
## 2020 11 18 16:18:53.662 DEBUG ab7c64738d52 calcAndWriteBoxDataFile done
## 2020 11 18 16:18:53.709 DEBUG ab7c64738d52 calcAndWriteBoxplot - after calcAndWriteBoxDat
## 2020 11 18 16:18:53.709 DEBUG ab7c64738d52 calcAndWriteBoxplot - before calcAndWriteHisto
## 2020 11 18 16:18:53.709 DEBUG ab7c64738d52 calcAndWriteHistogramFile  /builds/BatchEffect
## 2020 11 18 16:18:53.802 DEBUG ab7c64738d52 calcAndWriteBoxplot - after calcAndWriteHistog
## 2020 11 18 16:18:53.802 DEBUG ab7c64738d52 calcAndWriteBoxplot - before calcAndWriteAnnot
## 2020 11 18 16:18:53.802 DEBUG ab7c64738d52 calcAndWriteAnnotationsFile theFile= /builds/B
## 2020 11 18 16:18:53.808 DEBUG ab7c64738d52 calcAndWriteBoxplot - after calcAndWriteAnnota
## 2020 11 18 16:18:53.808 DEBUG ab7c64738d52 batchTypeName =  TSS
## 2020 11 18 16:18:53.808 DEBUG ab7c64738d52 theBatchType= TSS
## 2020 11 18 16:18:53.809 DEBUG ab7c64738d52 calcAndWriteBoxplot - theBoxDataFile= /builds/
## 2020 11 18 16:18:53.809 DEBUG ab7c64738d52 calcAndWriteBoxplot - theMedian= 0
```

```
## 2020 11 18 16:18:53.809 DEBUG ab7c64738d52 calcAndWriteBoxplot - dim(theData)[1]= 1250
## 2020 11 18 16:18:53.809 DEBUG ab7c64738d52 calcAndWriteBoxplot - dim(theData)[2]= 80
## 2020 11 18 16:18:53.810 DEBUG ab7c64738d52 checkCreateDir:  /builds/BatchEffects_clean/Ba
## 2020 11 18 16:18:53.810 DEBUG ab7c64738d52 calcAndWriteBoxplot - before calcAndWriteBoxDa
## 2020 11 18 16:18:53.810 DEBUG ab7c64738d52 calcAndWriteBoxDataFile theFile= /builds/Batch
## 2020 11 18 16:18:53.810 DEBUG ab7c64738d52 calcAndWriteBoxDataFile thePngFile= /builds/Ba
## 2020 11 18 16:18:53.837 DEBUG ab7c64738d52 calcAndWriteBoxDataFile CairoPNG= /builds/Batc
## 2020 11 18 16:18:53.838 DEBUG ab7c64738d52 calcAndWriteBoxDataFile call boxplot

## 2020 11 18 16:18:53.866 DEBUG ab7c64738d52 calcAndWriteBoxDataFile call text
## 2020 11 18 16:18:53.885 DEBUG ab7c64738d52 calcAndWriteBoxDataFile done
## 2020 11 18 16:18:53.933 DEBUG ab7c64738d52 calcAndWriteBoxplot - after calcAndWriteBoxDat
## 2020 11 18 16:18:53.934 DEBUG ab7c64738d52 calcAndWriteBoxplot - before calcAndWriteHisto
## 2020 11 18 16:18:53.934 DEBUG ab7c64738d52 calcAndWriteHistogramFile  /builds/BatchEffect
## 2020 11 18 16:18:54.012 DEBUG ab7c64738d52 calcAndWriteBoxplot - after calcAndWriteHistog
## 2020 11 18 16:18:54.012 DEBUG ab7c64738d52 calcAndWriteBoxplot - before calcAndWriteAnnot
## 2020 11 18 16:18:54.012 DEBUG ab7c64738d52 calcAndWriteAnnotationsFile theFile= /builds/B
## 2020 11 18 16:18:54.018 DEBUG ab7c64738d52 calcAndWriteBoxplot - after calcAndWriteAnnota

## [1] TRUE
```

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-
-rw-r--r-- 1 linux linux   7552 Jun 19 15:12 BoxPlot_AllSample-Data_CatData-BatchId-TCGA-OR-
#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
```

```
-rw-r--r-- 1 linux linux   5461 Jun 19 15:13 BoxPlot_AllSample-Data_CatData-TSS-TCGA-PK-A5HB
-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  Uppe
TCGA-OR-A5J1-01A-11D-A29J-05    NA  NA  -0.020527642802858643   -0.8467955227772493 -0.40564
TCGA-OR-A5J2-01A-11D-A29J-05    NA  NA  -0.002911079872554134   -0.039930119705853896   -0.0
TCGA-OR-A5J3-01A-11D-A29J-05    NA  NA  -0.035001758602725926   -0.3988124487830225 -0.34987
TCGA-OR-A5J4-01A-11D-A29J-05    NA  NA  -0.017185120892053492   -0.8247218460963763 -0.31831
TCGA-OR-A5J5-01A-11D-A29J-05    NA  NA  -0.03364073791133153    -0.8079754846584357 -0.64425
TCGA-OR-A5J6-01A-31D-A29J-05    NA  NA  -0.0034328681890936023  -0.04187597080189986    -0.0
```

```
TCGA-OR-A5J7-01A-11D-A29J-05    -0.865878813052012  -0.18821669173503153    -0.0038894121418
```

###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
TCGA-OR-A5J1-01A-11D-A29J-05    12  -0.8064387185053226 193.0   -0.7257251099614688 44.0
TCGA-OR-A5J2-01A-11D-A29J-05    79  -0.033911616995144944   168.0   -0.02187461157372705
TCGA-OR-A5J3-01A-11D-A29J-05    7   -0.32982819164709853    520.0   -0.19185967737525045
```

##Diagram

Here is a diagram generated from this code.