MBatch: Assessment Statistical Documentation for Non-Statisticians
Tod Casasent
2019-07-17-0820

Introduction

This document is aimed at the non-statistician and is intended to explain how to use the diagrams generated by MBatch and to understand the statistics behind the diagrams, with an emphasis on those created by the MD Anderson "pipeline" runs. A broad overview of each analysis method, with examples, and links to detailed reading is provided.

Principal Component Analysis (PCA)

Principal Component Analysis transforms a data matrix such that the larger the data variance provided by an element, the larger the resulting transformed values. **The values along each axis are unit-less principal component scores.**

In terms of Batch Effects Analysis, this means we can use this to detect variances of the data attributable to batches. For Batch Effects Analysis, we are less concerned with the individual mapping of samples to principal components, and more concerned with mapping between batches. To this effect, we created PCA-Plus, which adds a centroid to each set of batch data and color-codes members of a batch.

By providing batch information as part of the diagram, we highlight certain aspects of the data relevant to Batch Effects Analysis. In the BRCA diagram below, the batch circled in blue is a definite outlier or Batch Effect. By examining this batch, we discover these are all control analyte samples processed together in a single batch. (We discover this by hovering the mouse over an outlier data point, which reports in the Datapoint log the type is a Control Analyte--see the Datapoint Log picture below the diagram.)

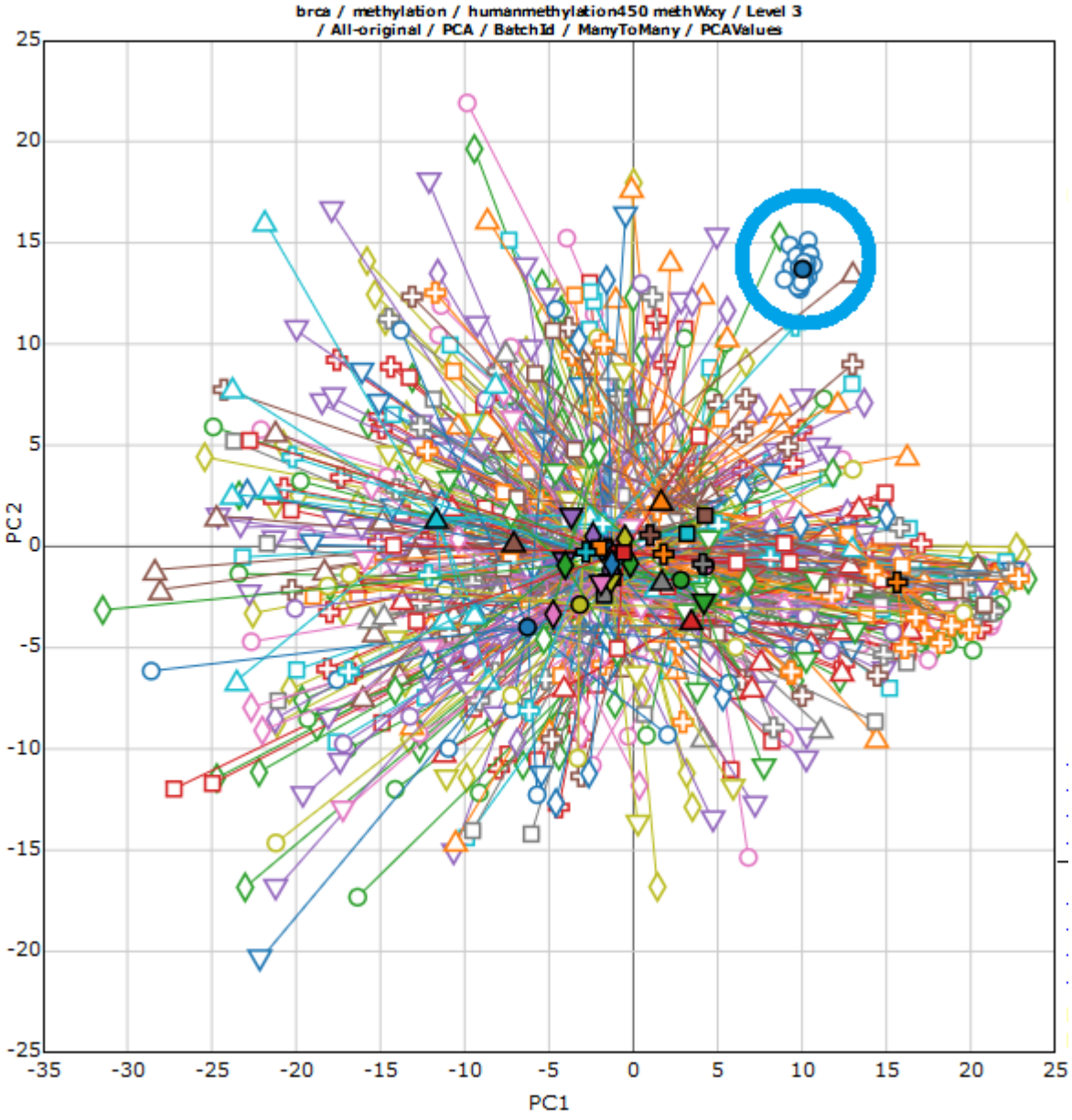
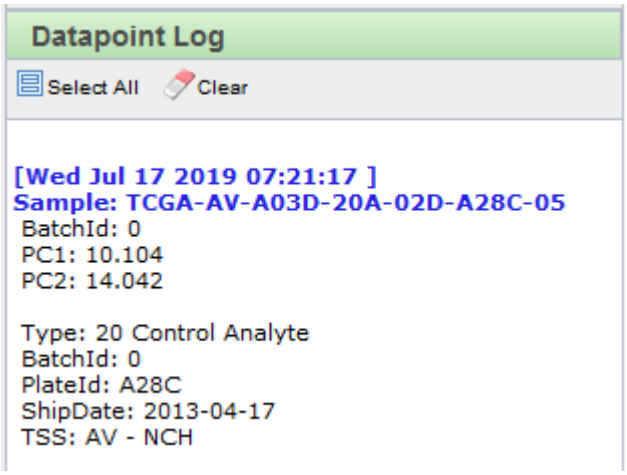


Diagram: BRCA Data with Analyte Outliers



Datapoint Log: Showing Control Analyte Type for Data Point

In the below diagram, the KIRC methylation data includes sex-based chromosomes, which generates a dichotomy in the data. Data on one side of the blue line is one sex, and data on the other side of the line is from another sex--this split can be horizontal or vertical, depending on the data.

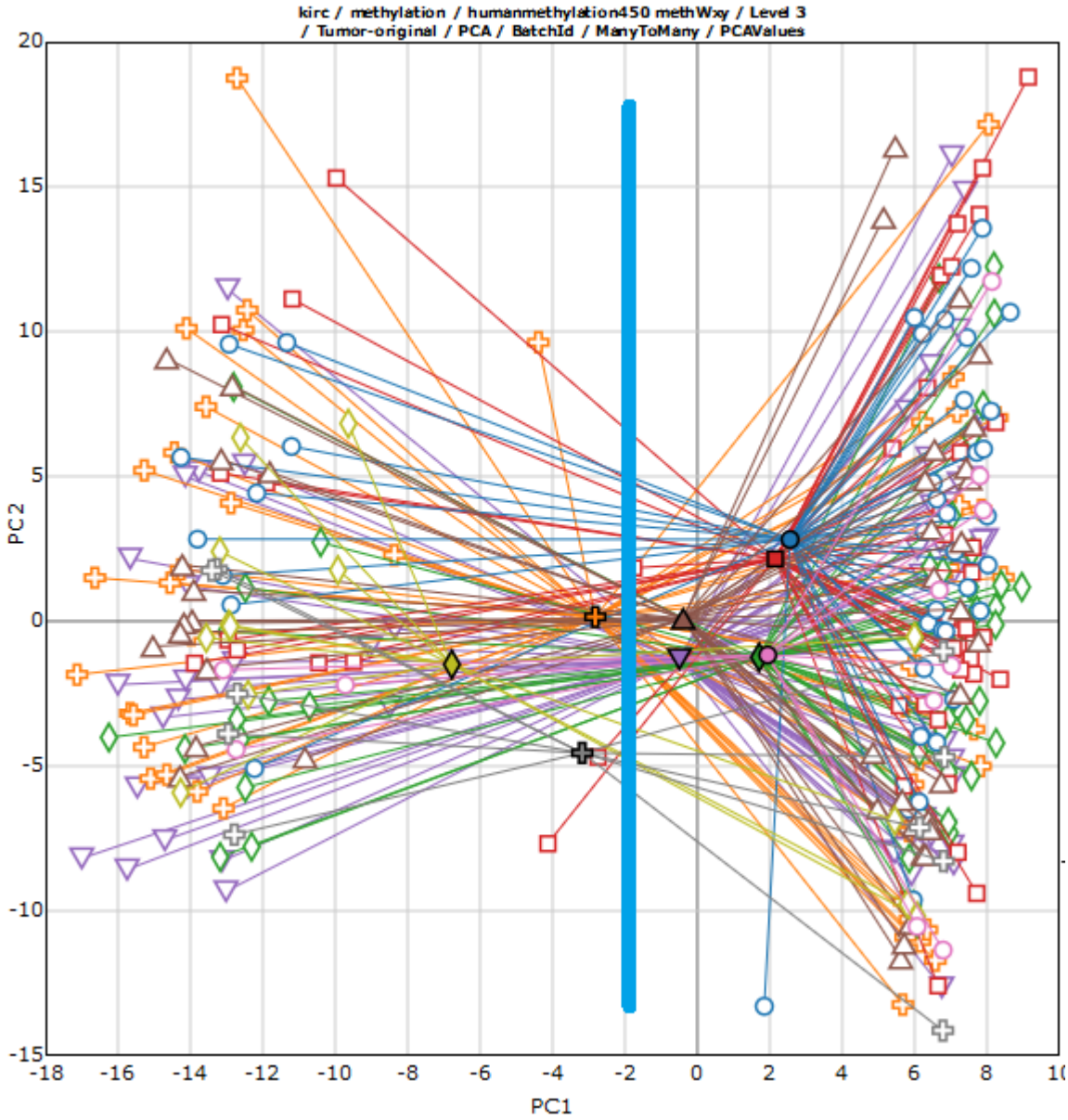


Diagram: KIRC Methylation Data Illustrating Sex Dichotomy

The diagram below shows the same data with sex chromosomes removed, eliminating the sex dichotomy.

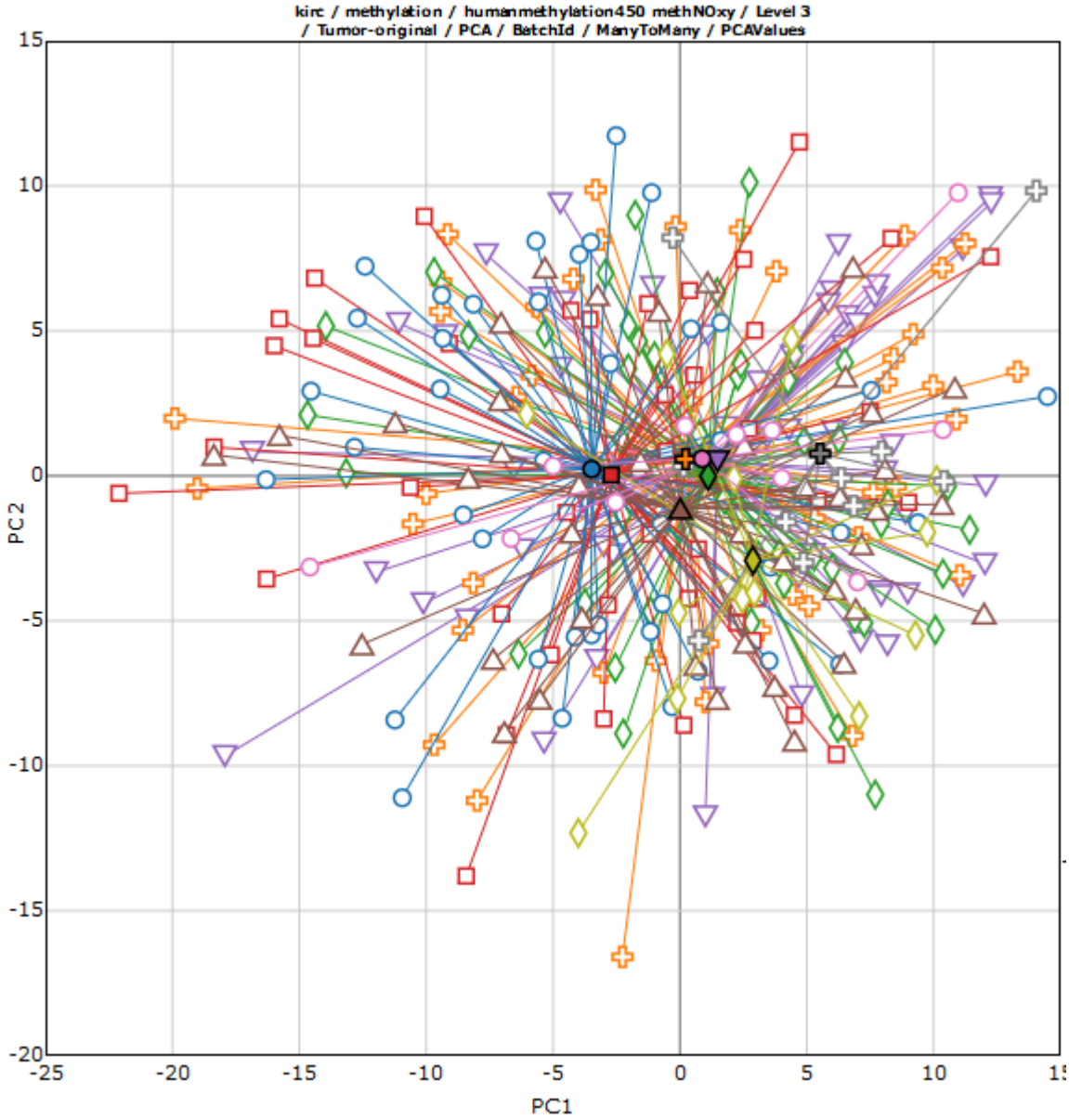


Diagram: KIRC Methylation Data with Sex Dichotomy Removed

The below diagram illustrates BRCA methylation data. BRCA data tends to be all female and lacks the sex dichotomy, while still showing the Control Analyte group.

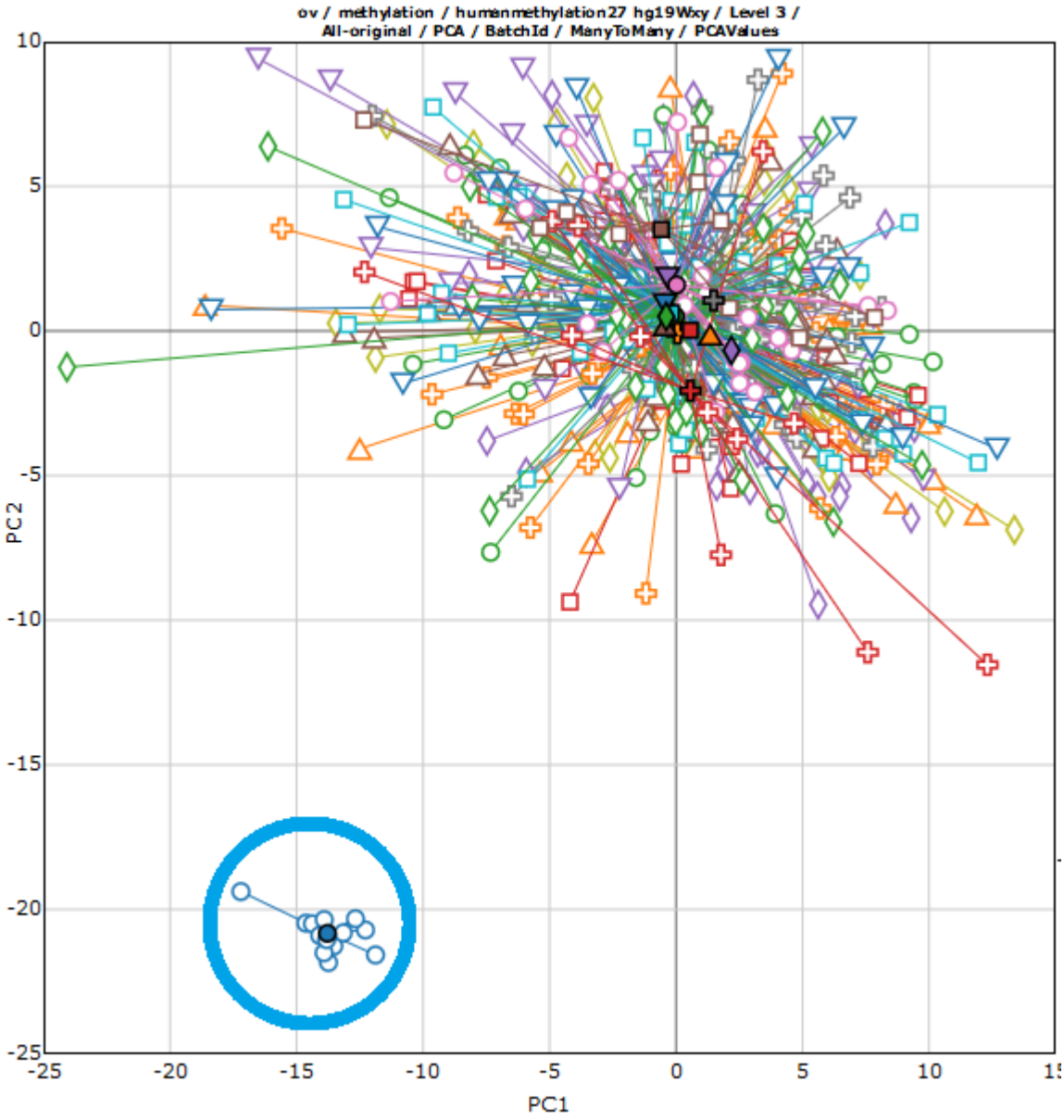


Diagram: Unfiltered BRCA Methylation Data with Sex Chromosome and No Dichotomy

Below illustrates an early Batch Effect in TCGA KIRC data. After removing sex chromosomes, the KIRC data from TCGA showed significant Batch Effects between one set of two and three other batches. (Image from https://www.genome.gov/Multimedia/Slides/TCGA2/15_Akbani.pdf.)

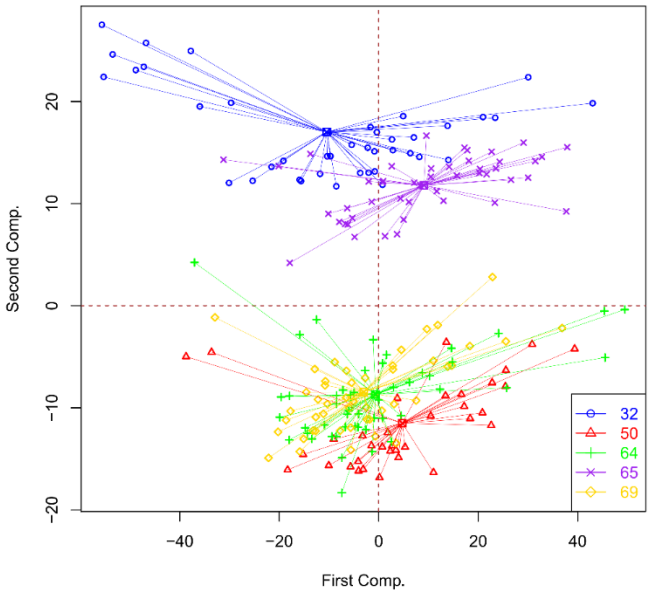
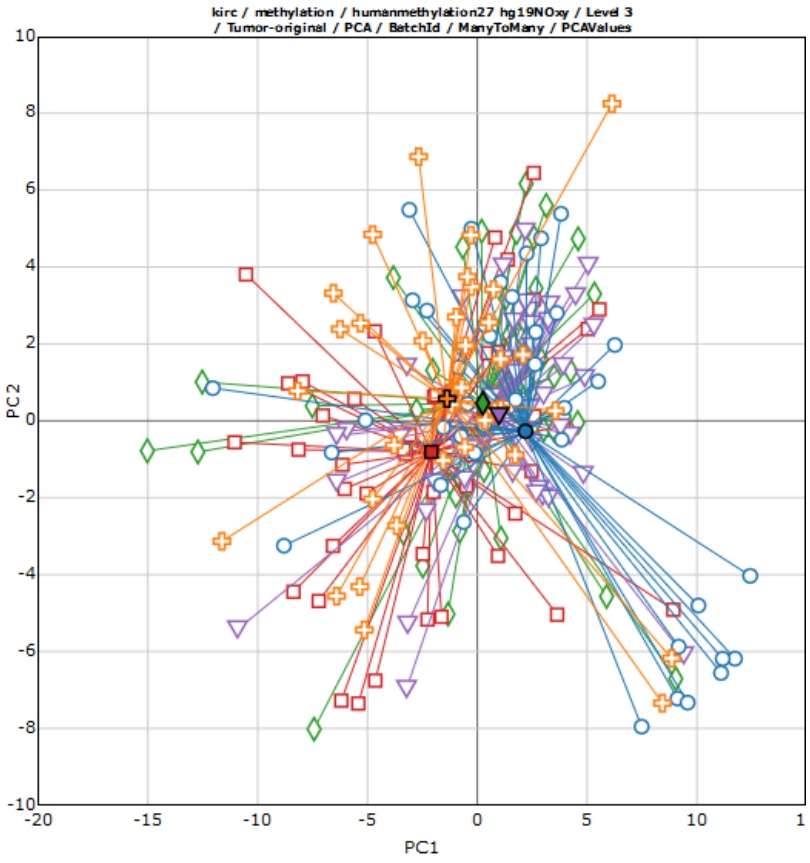


Diagram: Early Version of KIRC Batch Effects

Below is the same data, after the original batch-causing issues were resolved. (The scattered elements in the lower right corner are from a chromophobe sub-type. As a reminder, most cancer "types" consist of multiple sub-types that can show up in the PCA-Plus plots.)



PCA Details

For more details on PCA, Wikipedia provides several levels of explanation at https://en.wikipedia.org/wiki/Principal_component_analysis. For a more detailed statistical discussion, see Principal component analysis from Wold, Esbensen, and Geladi at https://doi.org/10.1016/0169-7439(87)80084-9.

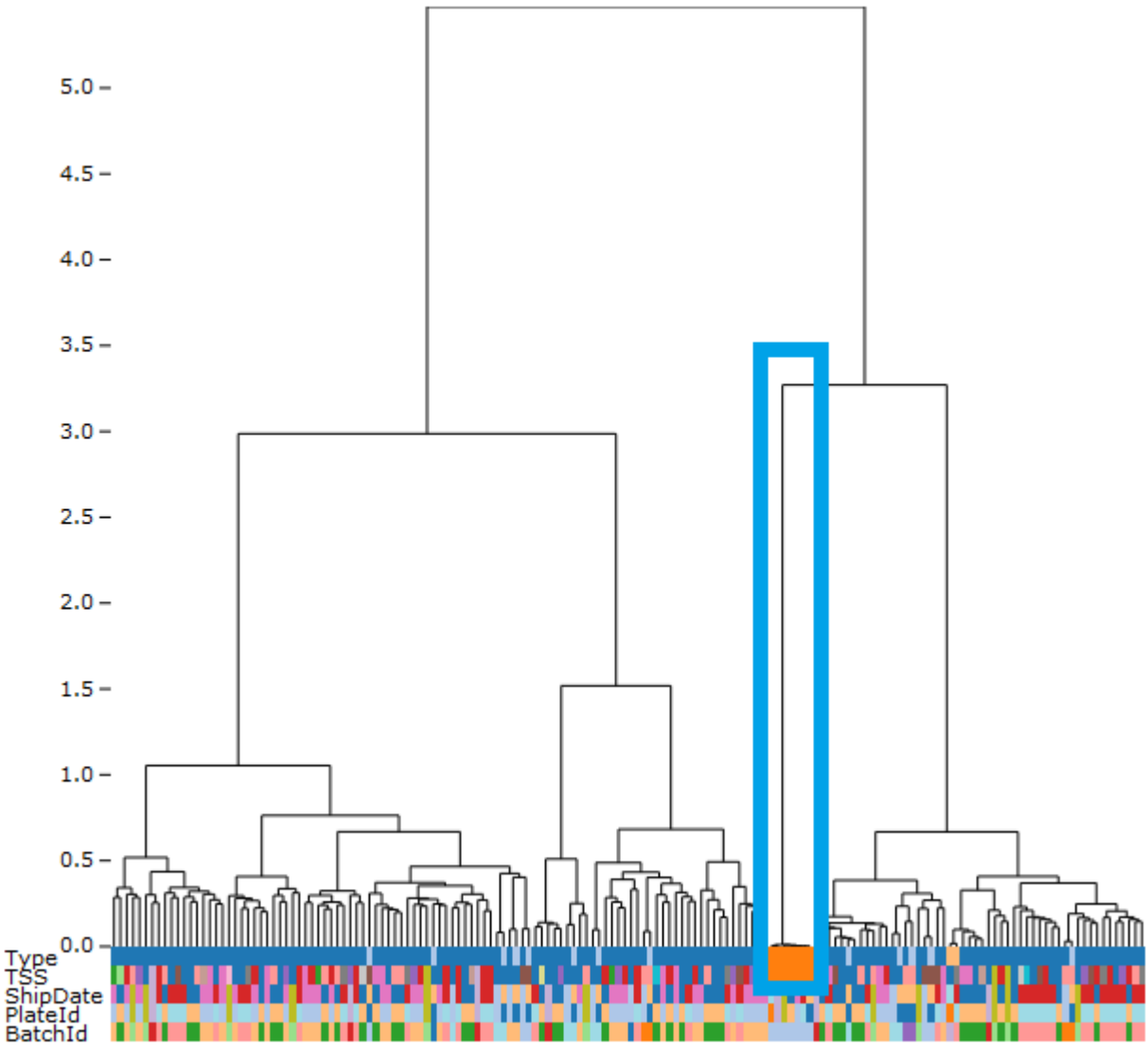
PCA-Plus: Regular and Dual Batch

The MBatch R package provides two kinds of PCA-Plus analysis. The "Regular" PCA illustrated above plots each batch with its own centroid and color/shape marker. The "Dual Batch" PCA takes pairs of batch types and plots them against each other in PCA-Plus.

Hierarchical Clustering

Hierarchical Clustering Analysis groups samples based on similarity, based on a correlation calculation and a clustering method. Below is the Hierarchical Clustering diagram for the GBM Methylation Data with the same Control Analyte Batch Effect.

The lines at the top of the diagram show how the samples cluster. There are two large clusters at the top, following by four smaller clusters, one of which is the Control Analyte Batch Effect, noticeable because for that cluster three of the Batch Types are all the same. The blue box highlights the cluster formed uniquely by the Control Analytes, seen by the orange Type and TSS annotations and the light blue Batch Id clustered together.



Hierarchical Clustering Diagram with Control Analyte

In other words, elements which are more closely correlated are closer together in the cluster. **In Hierarchical Clustering diagrams, the vertical axis is the "Height" and gives a measurement of the distance between elements or clusters. For "ward" clustering used in our pipeline runs, this is the amount the sum of squares grows when clusters are combined, which is shown by the height going from 0.0 to some larger number.**

Colors in the annotation bar have been selected to highlight vertical grouping of

related batches whenever possible.

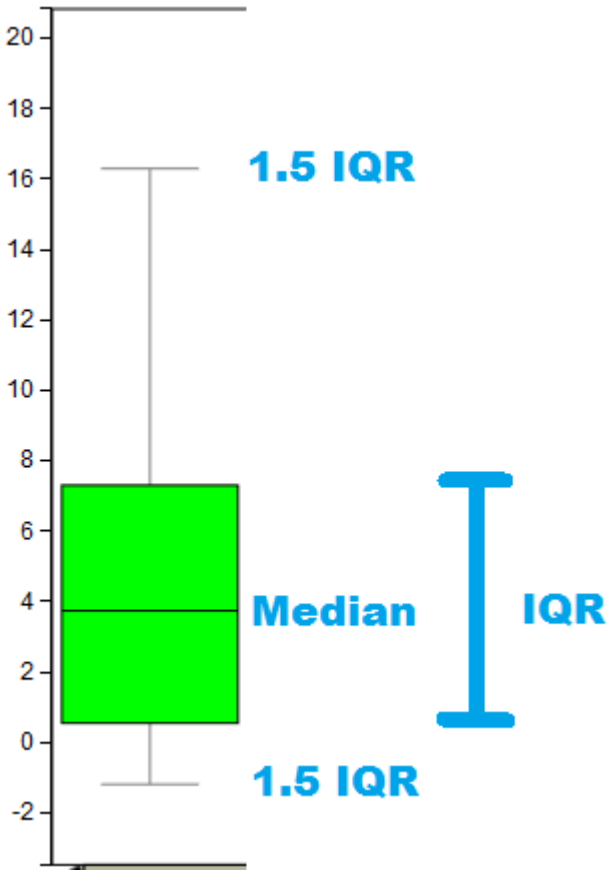
For details on reading clustering and dendrograms (a graphical bifurcating plot of the relationships between elements), see https://www.displayr.com/what-is-hierarchical-clustering/.

Hierarchical Clustering also lets us immediately see the unfortunate design where the control analytes were all processed by the same TSS in a single batch, confounding any non-biological batch effects on those samples.

Additional reading on the techniques used are available. Wikipedia has entries on Hierarchical clustering https://en.wikipedia.org/wiki/Hierarchical_clustering, Pearson correlation coefficient https://en.wikipedia.org/wiki/Pearson_correlation_coefficient, and Ward's method https://en.wikipedia.org/wiki/Ward%27s_method. (The ?hclust documentation in R notes "ward" and "ward.D" do not include the squaring of dissimilarities before updating clusters.)

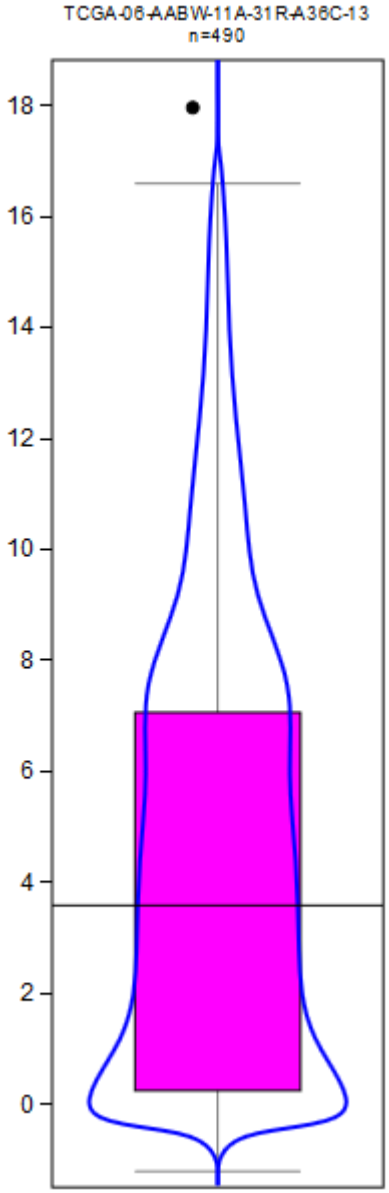
Boxplot

Boxplot Analysis is also called Box and Whiskers. Boxplots contain information for one or more samples. Summary information for the data includes a box indicating the interquartile range (IQR) covering 50% of the data points. A line within the box gives the median for the data, indicating half the data points are on each side of that line. Lines outside of each end of the box indicate the maximum and minimum non-outlier values, which are within 1.5 IQR of the median.



Boxplot Description

The above description is slightly different from the "Box plot" described in Wikipedia, as there are many variations of this plot. Next to the set of boxplots is a violin plot for one of the elements/batches. The violin plot is basically a box and whiskers plot with a histogram drawn on top of it.



Violin Plot Example

For additional reading on Boxplot topics, see:

Wikipedia Box plot https://en.wikipedia.org/wiki/Box_plot

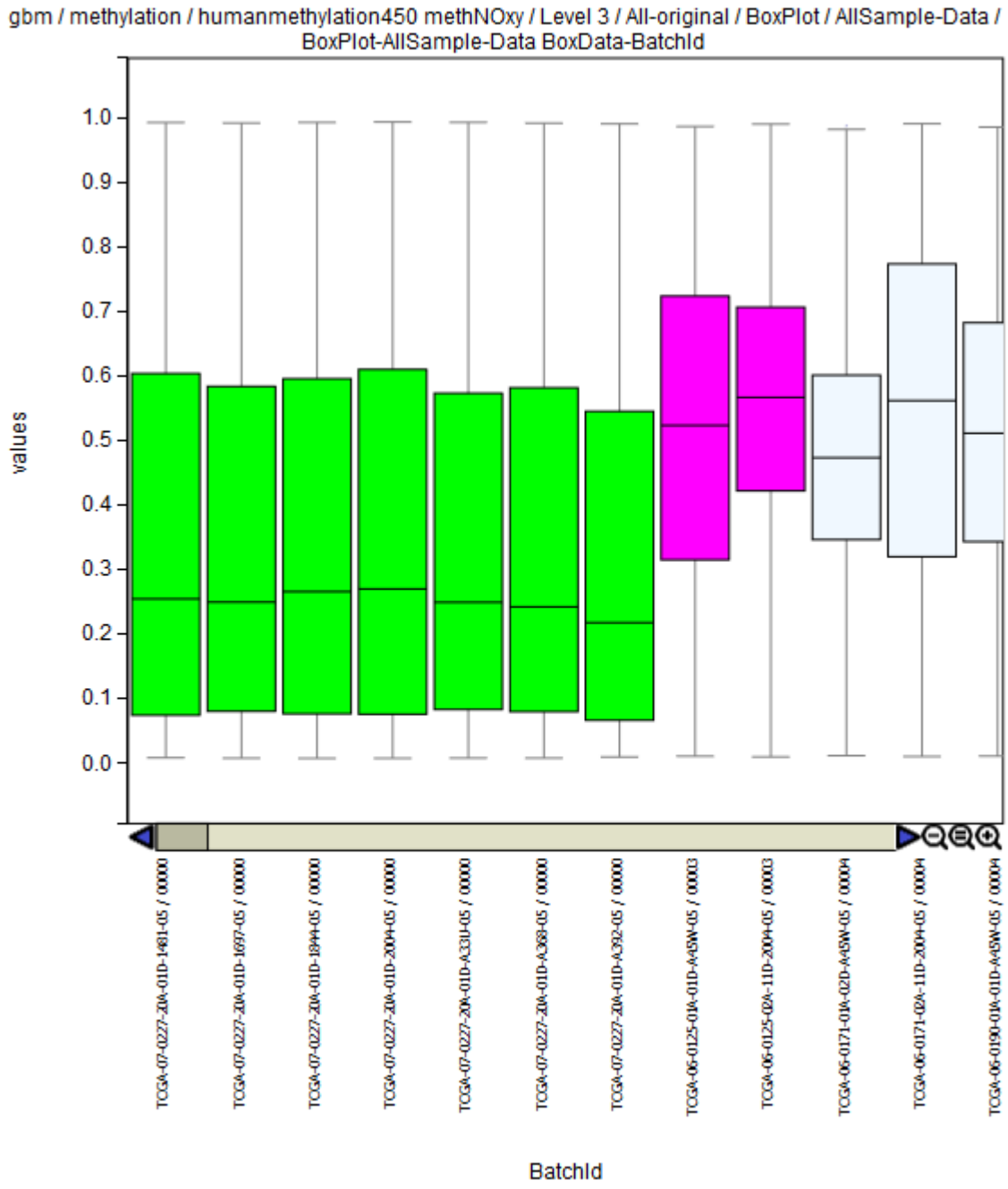
Wikipedia Interquartile Range (IQR) https://en.wikipedia.org/wiki/Interquartile_range

Wikipedia Violin Plot https://en.wikipedia.org/wiki/Violin_plot

Gandolfo LC, Speed TP (2018) RLE plots: Visualizing unwanted variation in high dimensional data. PLoS ONE 13(2): e0191629. https://doi.org/10.1371/journal.pone.0191629

Boxplot: All Sample Data

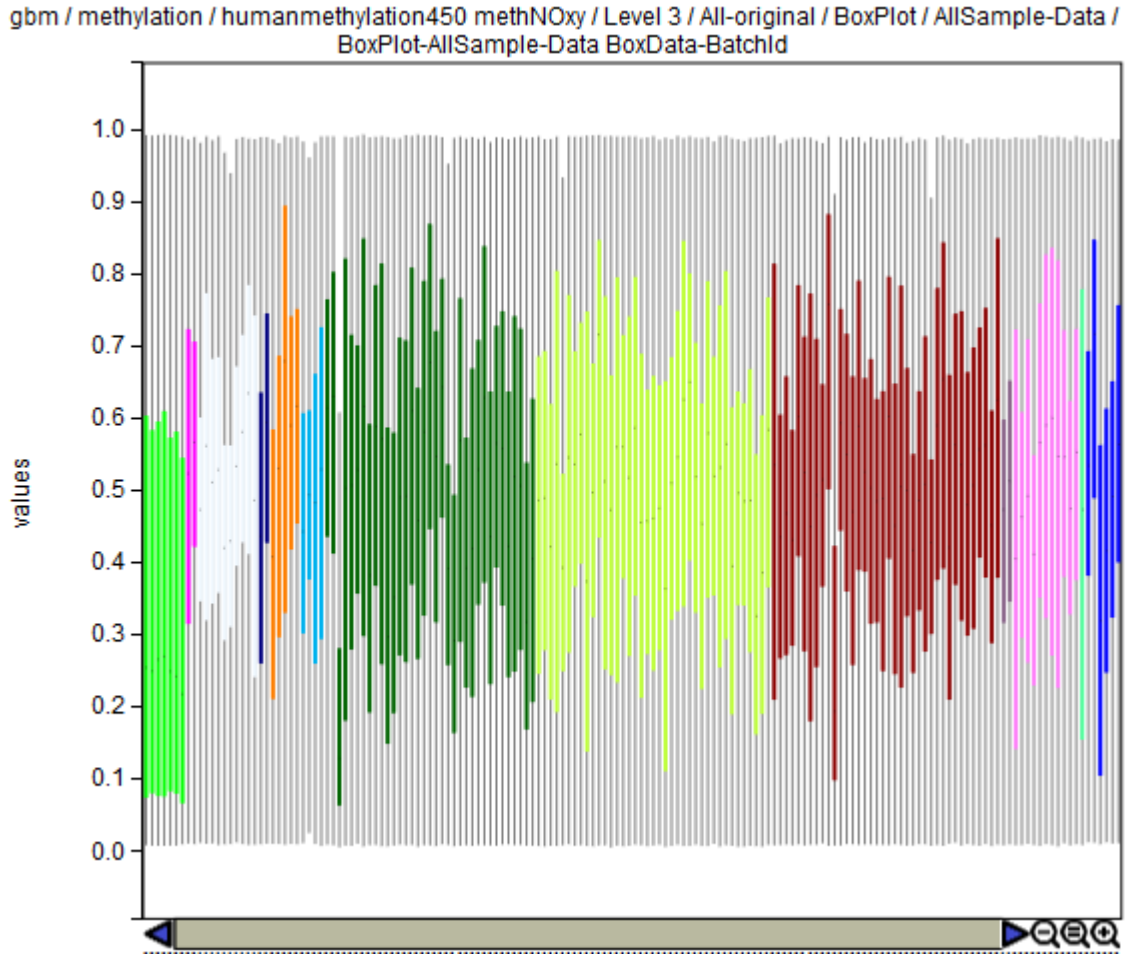
For the Boxplot: All Sample Data Analysis, each individual sample is plotted as a boxplot. Samples are grouped and colored by batch. The plot below illustrates a portion of the OV TCGA data. The green boxes with Batch Id "0" are the Control Analyte samples. The median for these boxes are significantly lower than the medians (or even the IRQ) for the other samples, suggesting Control Analyte samples are an outlier.



Boxplot All Sample Data: Green is the Analyte Control Group

Below is a zoomed out version of the same boxplot. You can see how the bright green of the Control Analyte stands out from the other batches with a lower

median.



Boxplot Diagram: Overview Showing Control Analyte Outliers

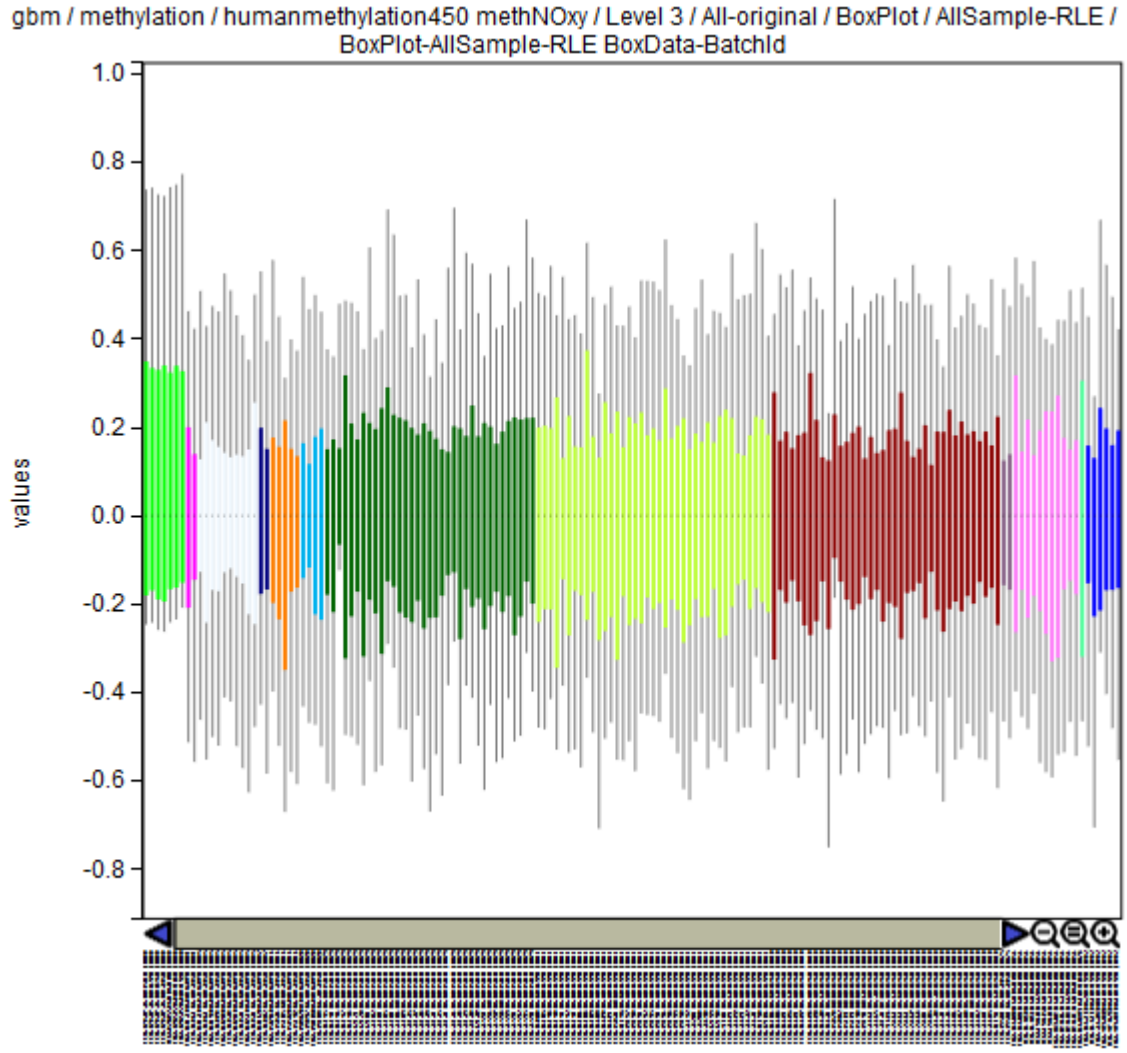
The vertical axis value is based on the values from the original data, which in this case is between zero and one.

Boxplot: All Samples RLE

For the Boxplot: All Sample RLE Analysis, plots are similar to the All Samples Data plots. Each box is still a different sample. We performed "relative log expression" (the RLE above) on the data by subtracting the median of the entire data set from each value. For some data sets, this removes variation that can hide other patterns.

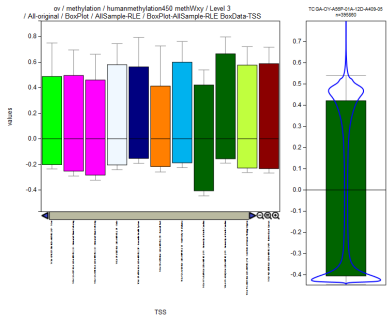
The diagram below illustrates the All Samples RLE diagram for the same GBM

data shown earlier. In this case, RLE tends to obscure the lower median of the Control Analyte group, except the top half of that group is taller than the other groups.



Boxplot for All Samples with RLE

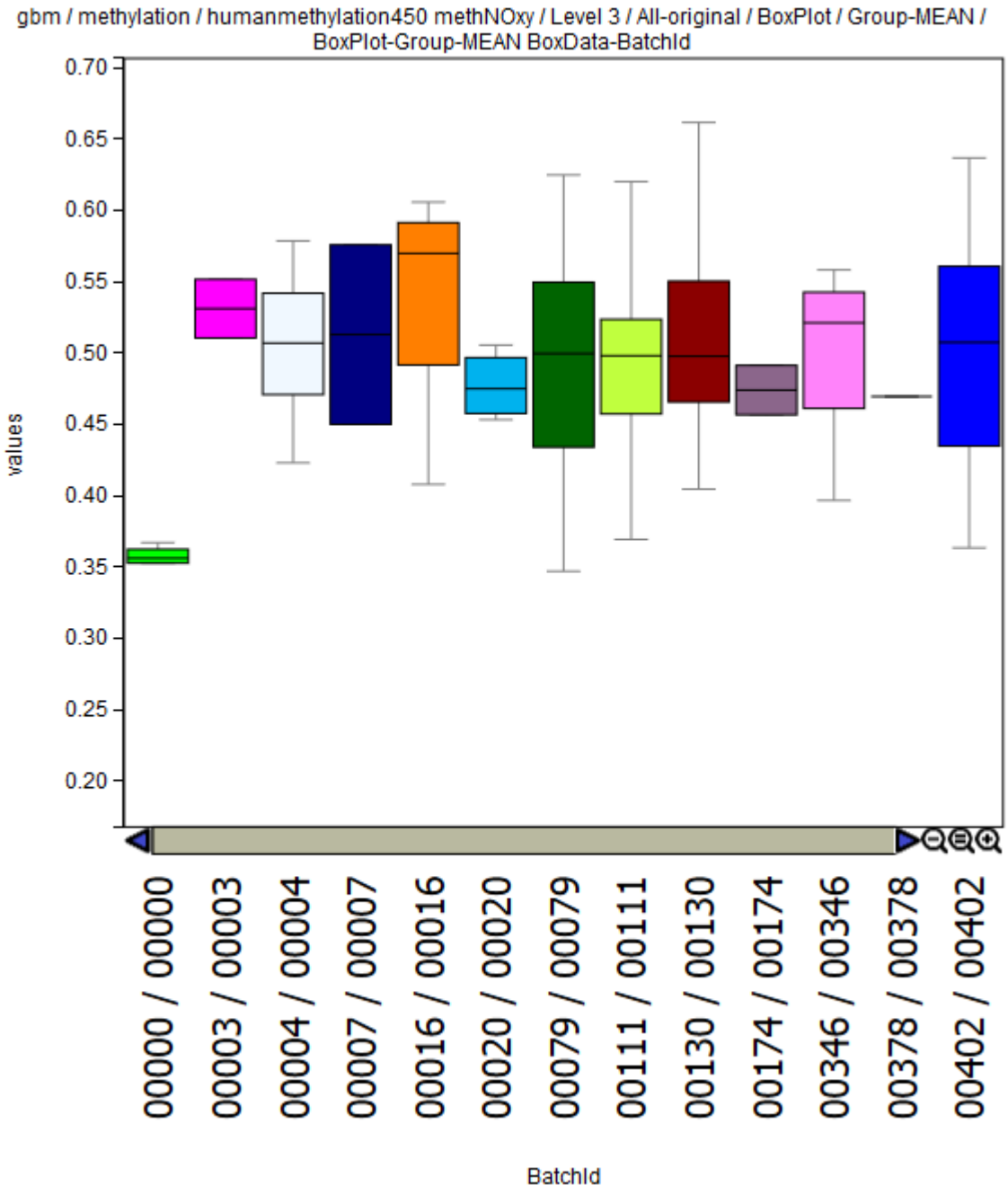
Looking at the same OV RPPA data with RLE, we note the Control Analyte batch id 0 boxes are generally shorter than the other boxes.



Examining a violin plot for methylation data, we see a "standard" violin plot with some features are highly methylated (up around .5) and some features are unmethylated (-.4). This is not unusual for methylation data.

Boxplot: Group - Mean

The final Boxplot analysis Boxplot Group, plots each batch in a single box using the mean for the batch. This makes our previous GBM outlier for the Control Analyte batch, shown below, even more obvious. This also helps illustrate why all three boxplots are included--each type of boxplot tends to be more effective at illustrating different kinds of Batch Effects.

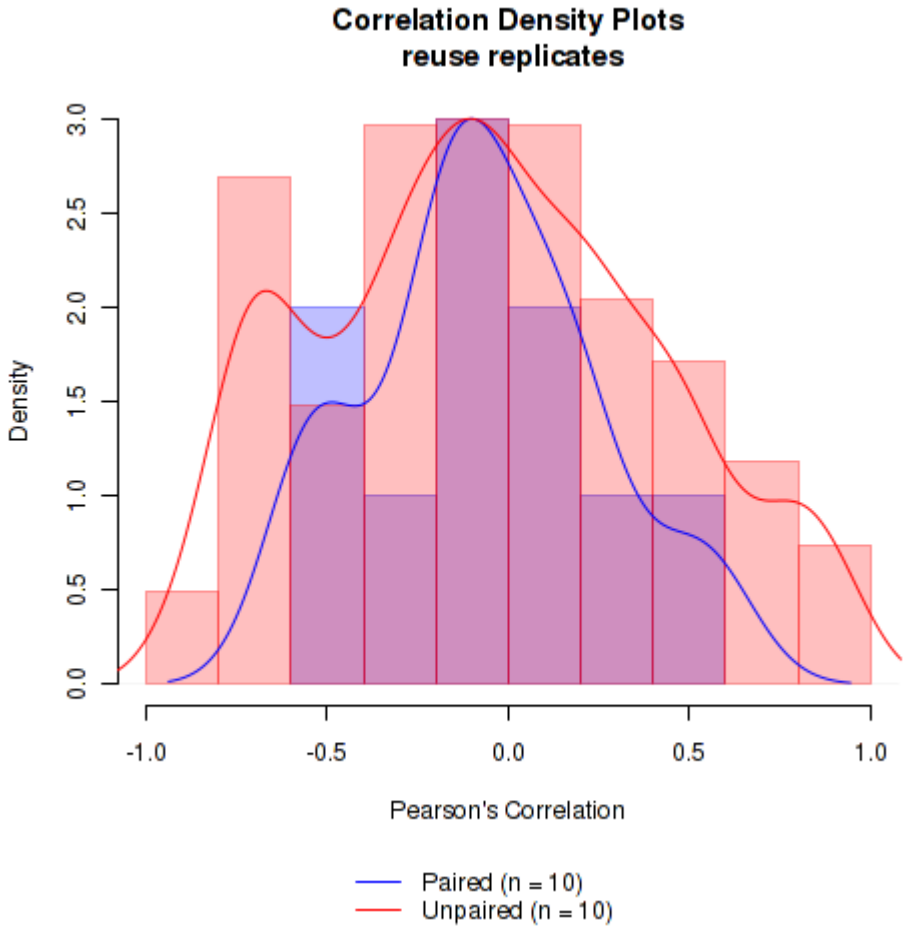

gbm / methylation / humanmethylation450 methNOxy / Level 3 / All-original / BoxPlot / Group-MEAN /
BoxPlot-Group-MEAN BoxData-BatchId
values
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
00000 / 00000
00003 / 00003
00004 / 00004
00007 / 00007
00016 / 00016
00020 / 00020
00079 / 00079
00111 / 00111
00130 / 00130
00174 / 00174
00346 / 00346
00378 / 00378
00402 / 00402
BatchId

Correlation Density Plot (CDP)

The Correlation Density Plot (CDP) Analysis is a Bayesian algorithm combining the concepts of a probability density function, histograms, and correlation. The probability density function and histogram both estimate the probability distribution of a continuous variable. The CDP uses both plots overlaying one atop the other.

The CDP plots paired versus unpaired replicates between two datasets--usually an uncorrected and corrected dataset. This lets the use check to see if the shape (distribution) of the correlation of the data as represented by the density plot has been changed by the correction. By default, Pearson's Correlation and pairwise complete checks are performed.

The paired and unpaired correlations should be similar. This data is from the CDP test included with MBatch.



For additional reading:

Probability Density Function https://en.wikipedia.org/wiki/Probability_density_function

Histogram https://en.wikipedia.org/wiki/Histogram

Correlation https://en.wikipedia.org/wiki/Correlation_and_dependence

Supervised Clustering

Supervised Clustering Analysis groups samples, based on a distance calculation related to batch membership and a clustering method. Similar to hierarchical clustering, elements which are more closely correlated are closer together in the cluster, but the clustering is "biased" (hence "supervised") to tend to cluster elements using batch information. Supervised Clustering also adds a simple heat map to show correlation between elements.

In Supervised Clustering diagrams, the vertical axis is the "Height" and gives a measurement of the distance between elements or clusters. For "ward" clustering, this is the amount the sum of squares grows when clusters are combined, which is shown by the height going from 0.0 to some larger number.

Colors in the annotation bar have been selected to highlight vertical grouping of related batches whenever possible.

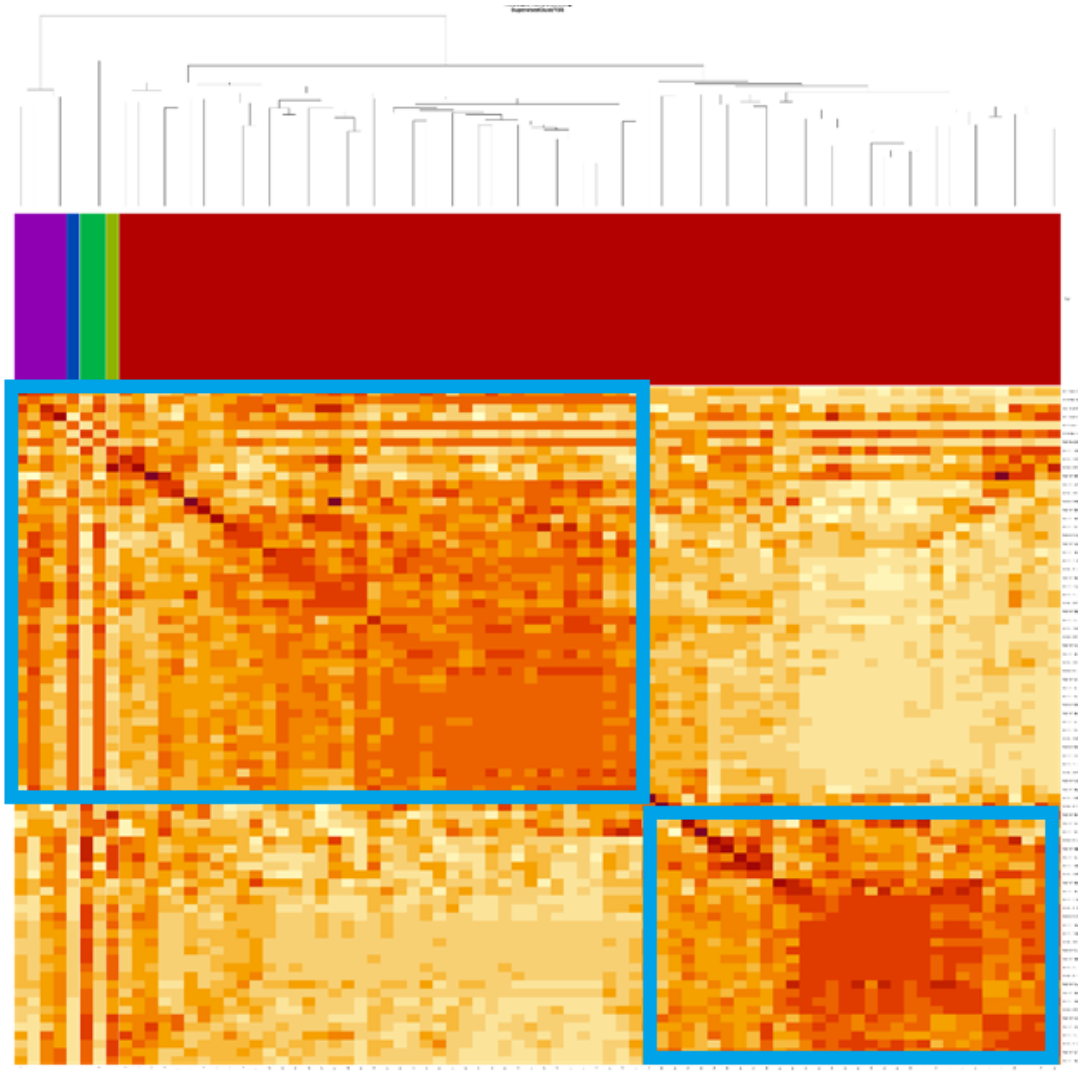
For details on reading clustering and dendrograms (a graphical bifurcating plot of the relationships between elements), see https://www.displayr.com/what-is-hierarchical-clustering/.

For general details on Supervised Clustering, see https://ieeexplore.ieee.org/document/1374270. For details on the average (UPGMA) agglomerative clustering method, see https://en.wikipedia.org/wiki/UPGMA.

Supervised Clustering: Batches

For the Supervised Clustering Batches analysis, each individual sample is plotted with biasing based on the batch values. Here, from the "tests" code, is a diagram of TSS batches.

Below, you can see how each TSS is grouped together, with the OR (University of Michigan) TSS generating two super-clusters (shown in blue) within its batch. In this case, the two super-clusters suggest biology or batch effects within that TSS.

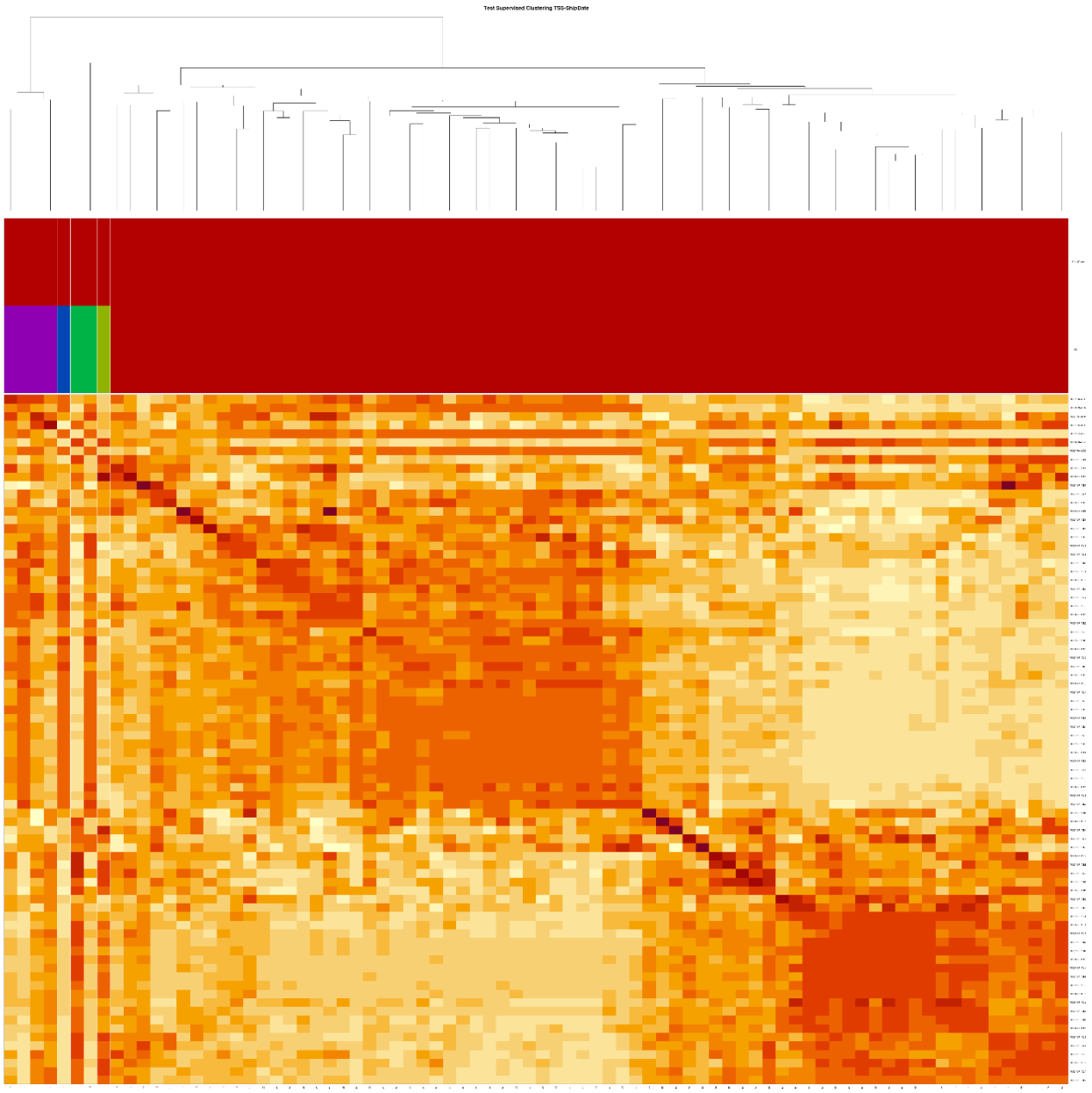


Supervised Clustering Super Clusters

Supervised Clustering: Pairs

For the Supervised Clustering Pairs analysis, each individual sample is plotted with biasing based on the batch values for membership in two different batches. Here, from the "tests" code, is a diagram of TSS batches versus Ship Date.

Below, you can see how the OR (University of Michigan) TSS generates the same two super-clusters within its batch. In this case, the two super-clusters suggest biology or batch effects within that TSS, which does not appear related to Ship Date, since there is only a single Ship Data.



Supervised Clustering with Clusters