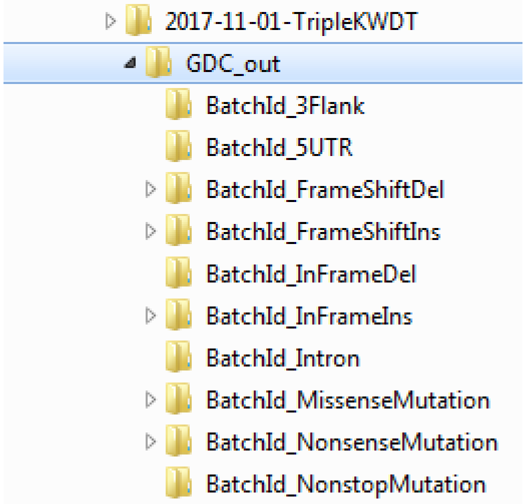
Mutation Batch Output
Tod Casasent
2019-10-15-1100

I ran two DCC runs and a GDC run through the Mutation Batch code. I then used the index.html "report" files to create the index.pdf files.

The end of this document also discusses the ViewBatches.html report.

If you open the GDC_out directory in the browser and look for directories with subdirectories, those are batch-type/mutation-type pairs that had batches called as significant by the Kruskal-Wallis and Dunn's Test pairing.



2017-11-01-TripleKWDT
GDC_out
BatchId_3Flank
BatchId_5UTR
BatchId_FrameShiftDel
BatchId_FrameShiftIns
BatchId_InFrameDel
BatchId_InFrameIns
BatchId_Intron
BatchId_MissenseMutation
BatchId_NonsenseMutation
BatchId_NonstopMutation

The ideas behind the index.html were 1) to provide a simpler way to view the output to determine if the batch effects called by KWDT are verifiable and 2) provide a way to save the evaluation, by printing the HTML to a PDF (which also allows you to zoom in on the thumbnails clearly in the PDF).

Within each batch-type/mutation-type pairs I write the batch type and mutation type ("PlateId Total" in the screenshot below); the disease, platform, and other identifiers for the batches called ("TCGA-BRCA MuTect2VariantAggregationandMasking_HG38"); and the batches called ("A12Q, A22N").

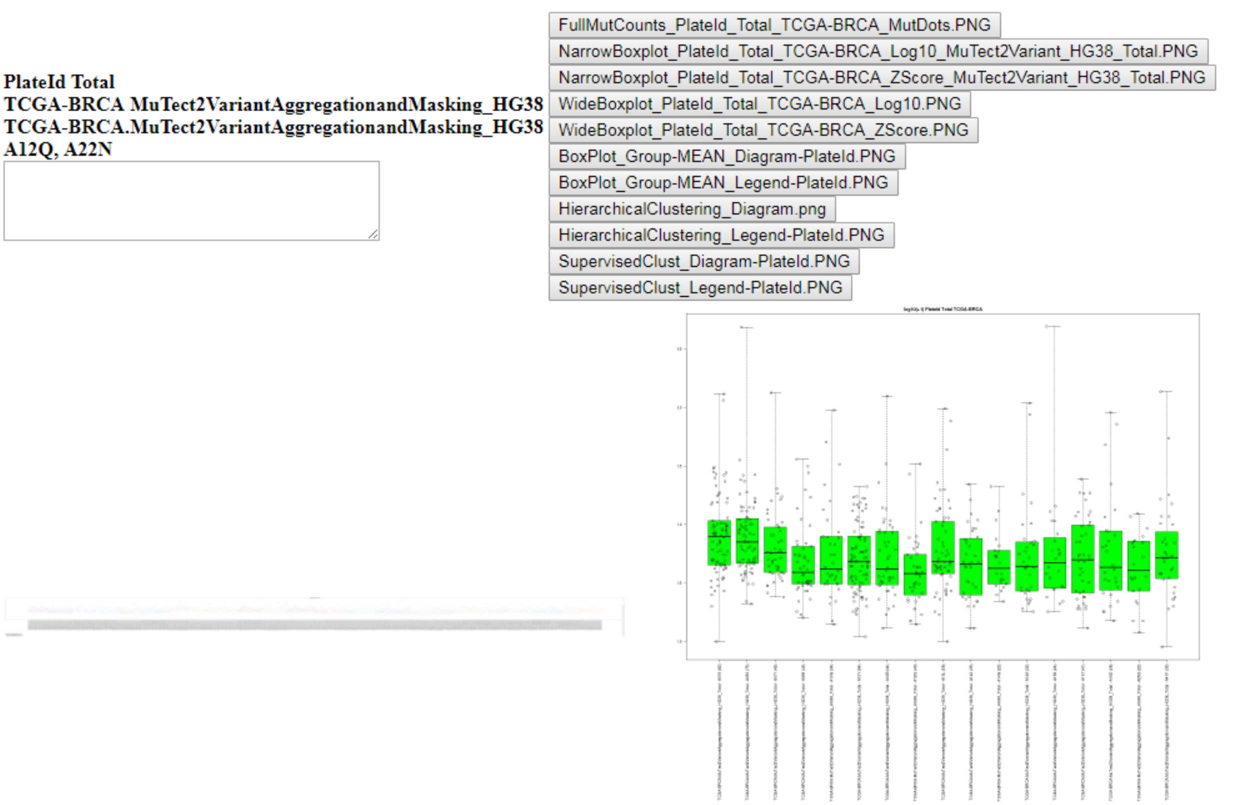
Below this information, I provide a box to place notes.

In boxplots, we see four types of batch effects.

- A batch with the mean (with box) above or below other boxes.
- A batch with the only boxes in the plot.
- A batch with the same mean, but only one with a box.

- A batch with its box with mean at min or max (both for box and for plot).

I also provide 11 plots. By clicking on the grey boxes, plots can be hidden or shown in the HTML. By clicking on a diagram, you can show the diagram full size or shrink it back to the thumbnail size.



PlateId Total
TCGA-BRCA MuTect2VariantAggregationandMasking_HG38
TCGA-BRCA.MuTect2VariantAggregationandMasking_HG38
A12Q, A22N

FullMutCounts_PlateId_Total_TCGA-BRCA_MutDots.PNG
NarrowBoxplot_PlateId_Total_TCGA-BRCA_Log10_MuTect2Variant_HG38_Total.PNG
NarrowBoxplot_PlateId_Total_TCGA-BRCA_ZScore_MuTect2Variant_HG38_Total.PNG
WideBoxplot_PlateId_Total_TCGA-BRCA_Log10.PNG
WideBoxplot_PlateId_Total_TCGA-BRCA_ZScore.PNG
BoxPlot_Group-MEAN_Diagram-PlateId.PNG
BoxPlot_Group-MEAN_Legend-PlateId.PNG
HierarchicalClustering_Diagram.png
HierarchicalClustering_Legend-PlateId.PNG
SupervisedClust_Diagram-PlateId.PNG
SupervisedClust_Legend-PlateId.PNG

Output is divided into <BatchType>_<MutationType> subdirectories. Datasets without Kruskal-Wallis-called significant results will only have a PNG (FullMutCounts_<BatchType>_<MutationType>.PNG) giving the Kruskal-Wallis/Dunn's Test results.

For datasets with significant calls, the aforementioned PNG will exist, as will files using the following patterns. These files are unique to the Mutation Batch code.

- FullMutCounts_<BatchType>_<MutationType>_<DiseaseType>_MutDots.PNG, which contains all samples sorted by batch with mutations for all platforms plotted
- NarrowBoxplot_<BatchType>_<MutationType>_<DiseaseType>_<Log10|ZScore>_<Platform_Gen which contains mutations for the given platform plotted by batches (with

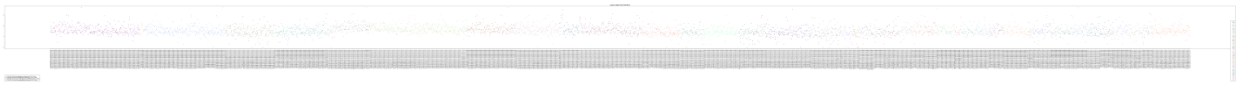
more than the average number of samples) and plots are provided both in terms of Log10 and ZScores.

- WideBoxplot_<BatchType>_<MutationType>_<DiseaseType>_<Log10|ZScore>.PNG, which contains mutations for all platforms plotted by batches and plots are provided both in terms of Log10 and ZScores.
- callReference.tsv is a tab delimited file with a header of MutationType, BatchTuype, MutationFile, and called Batches. Batches are comma delimited within parenthesis.

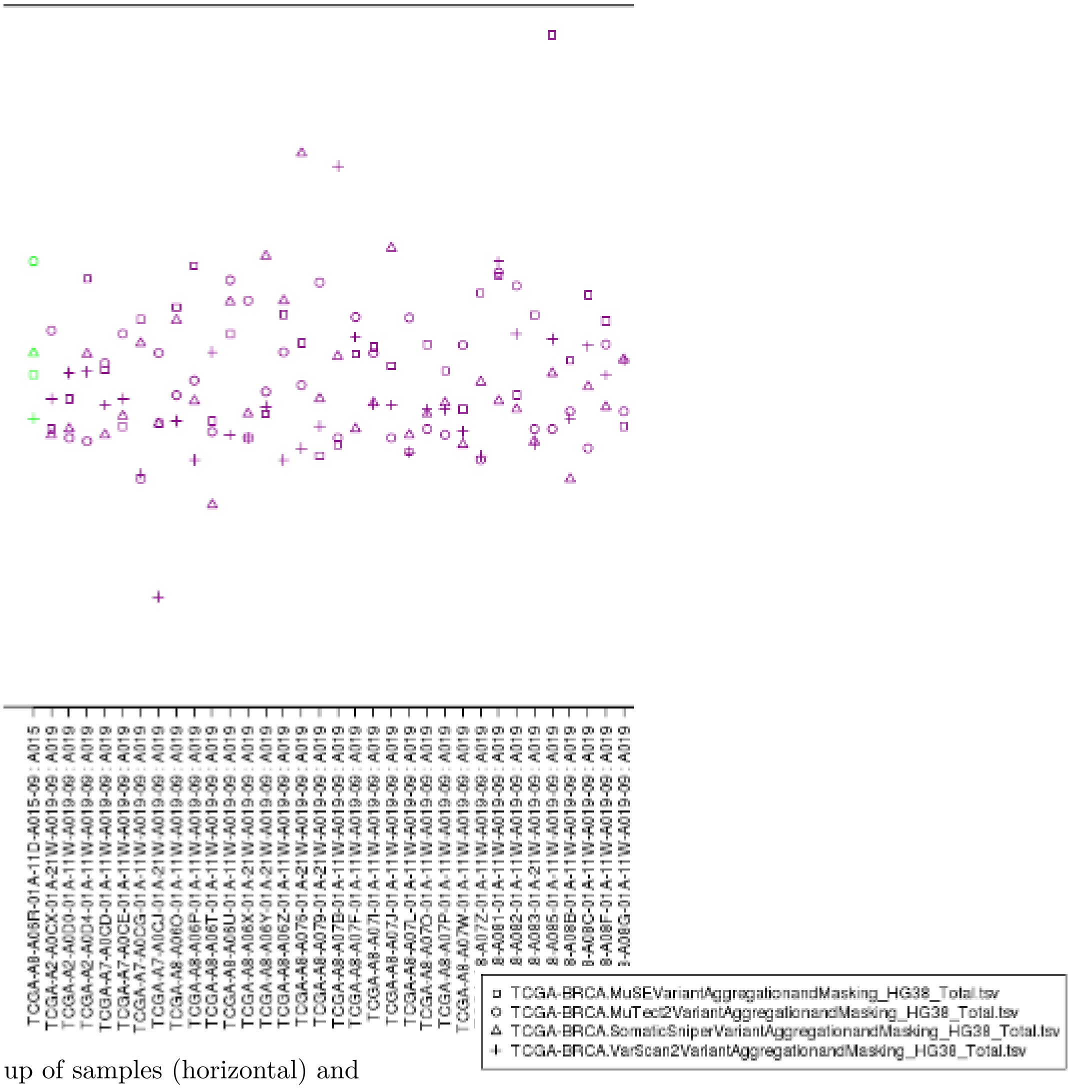
We also provide boxplots, hierarchical clustering, and supervised clustering diagrams from the MBatch package. While we ran PCAplus for this data, PCAplus does not generally provide useful insights into mutation batch effects, so they are not part of the report. (Supervised clustering did not create any output that I recall, so it is not described here.)

FullMutCounts_PlateId_Total_TCGA-BRCA_MutDots.PNG

This plot has all samples with number of calls for given mutation type for each platform (colors). Samples are grouped by batch (color) and sorted by sample id (sorted order may indicate time). For MutDots plots, the vertical axis is the number of mutations per sample with a log10(+1) modification/normalization.

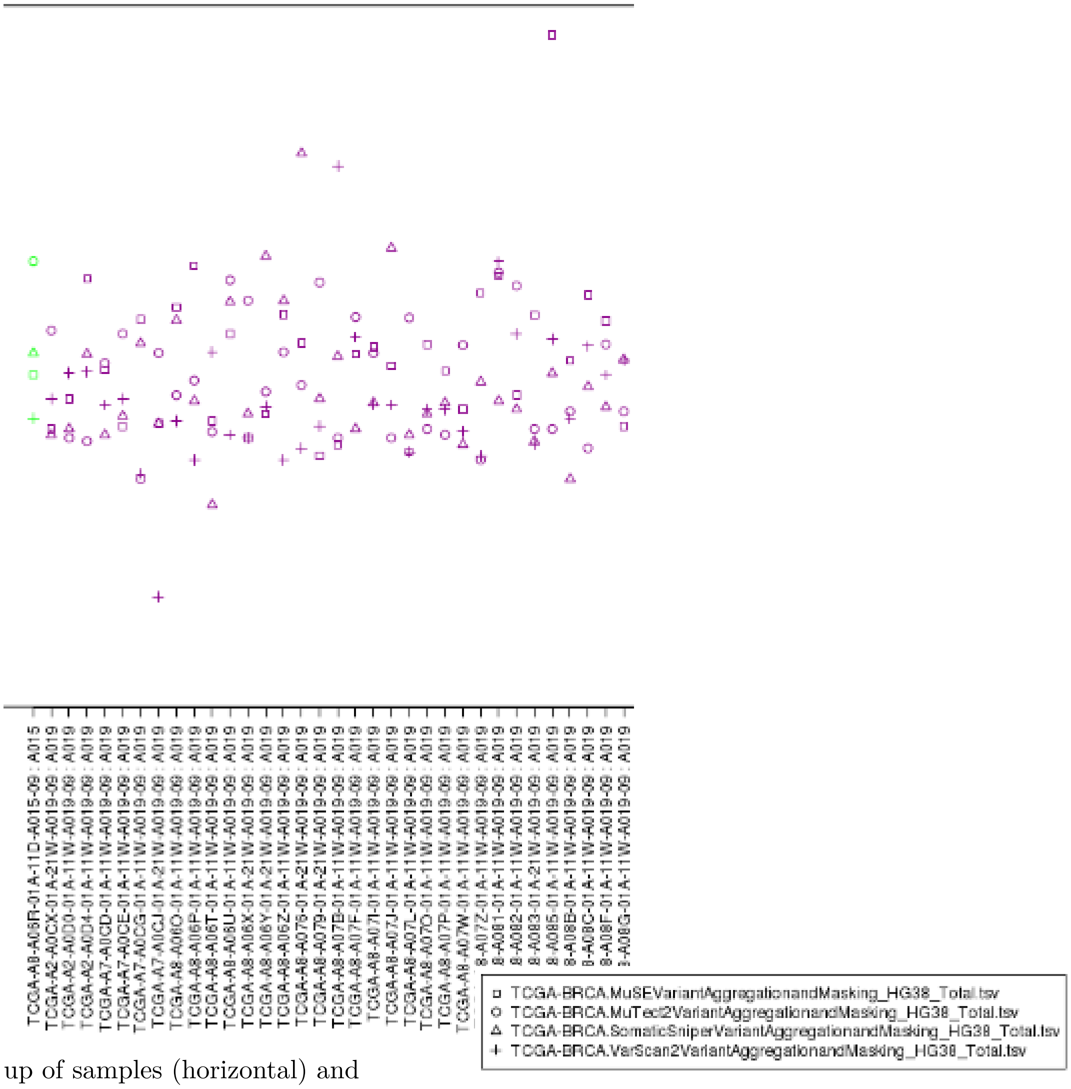


Full Plot



up of samples (horizontal) and number of calls by platform (vertical)

Legend for Platform Shapes

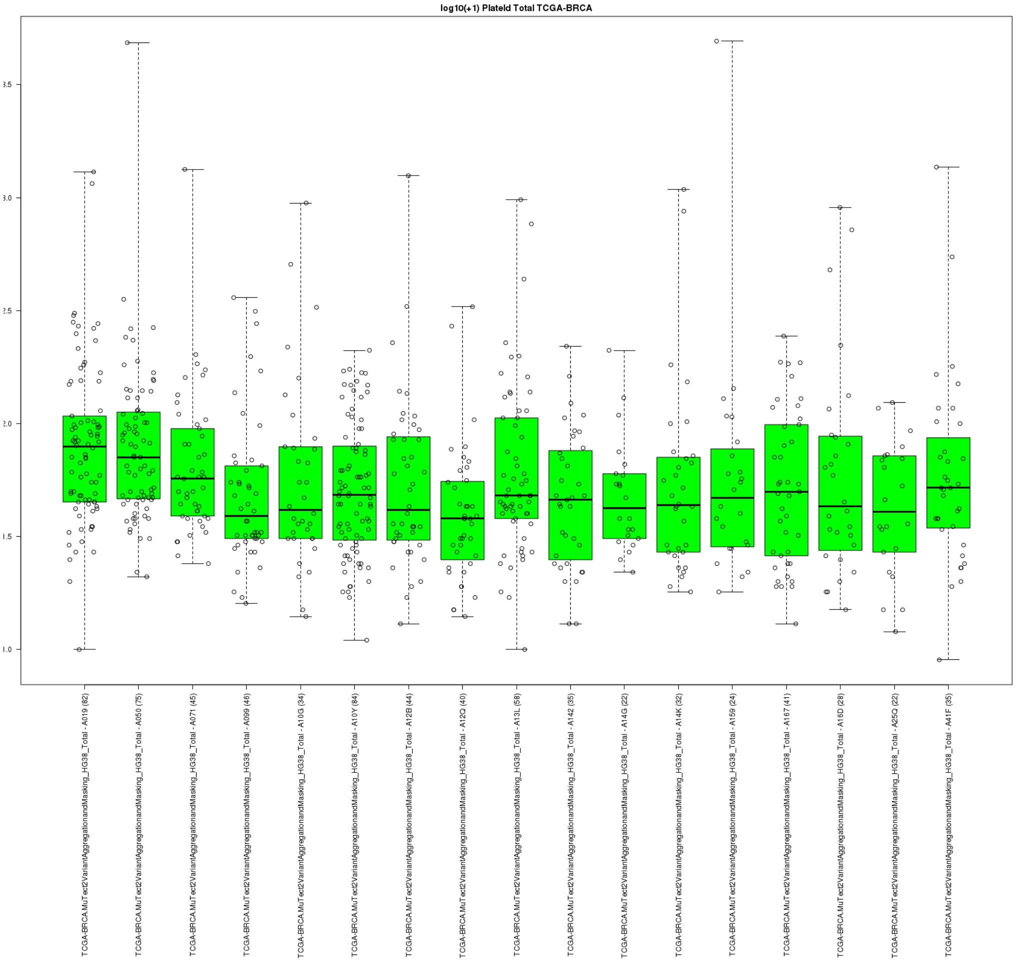


up of samples (horizontal) and number of calls by platform (vertical)

Legend for Platform Shapes

NarrowBoxplot__PlateId__Total__TCGA-BRCA__Log10__MuTect2Variant__HG38__Total.PNG

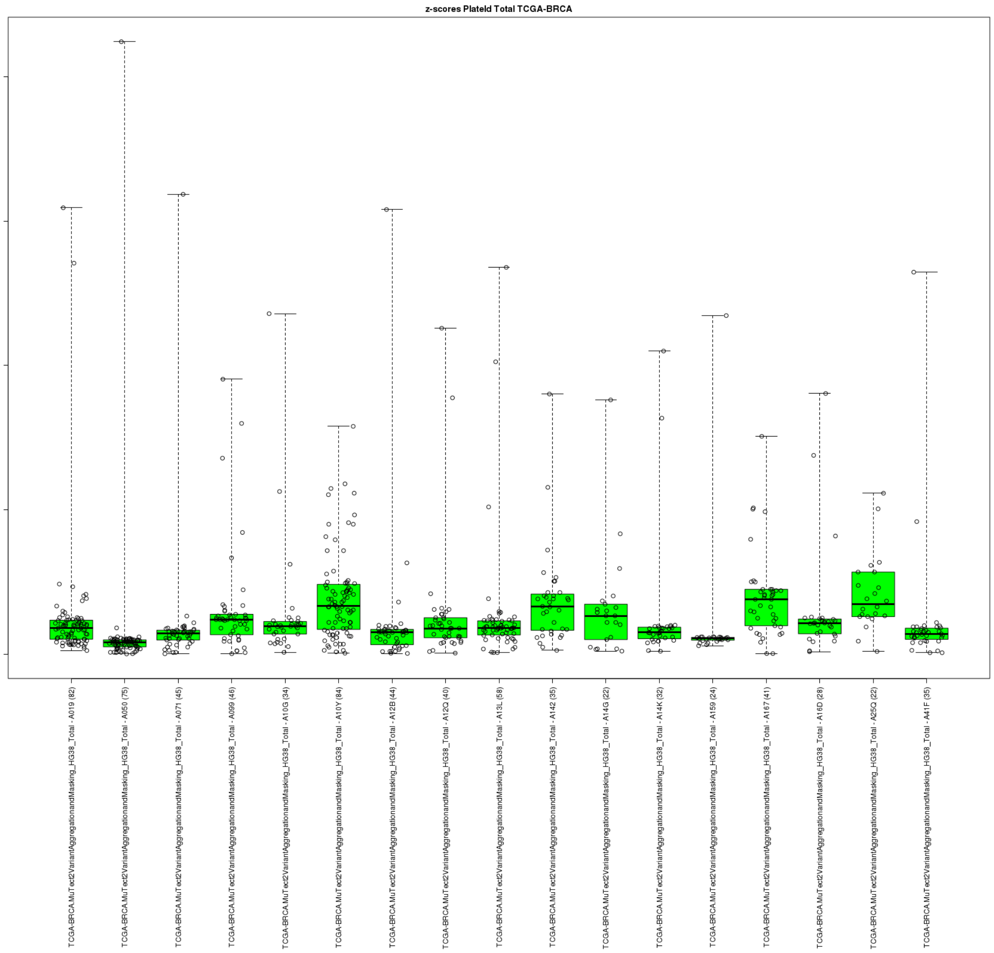
This plot has batches that contain at least the average number of samples and only batches within the given platform. (Filtering is to make "narrow" plots usable. For complete data, see the "wide" plots.) Boxplot data is in terms of Log10 data (to bring out small data differences in mutation counts). For Narrow and Wide Boxplots, the vertical axis is the number of mutations per sample, with the appropriate (Z-Score or Log10(+1)) modification/normalization.



Full Plot

NarrowBoxplot__PlateId__Total__TCGA-BRCA__ZScore__MuTect2Variant__HG38__Total.PNG

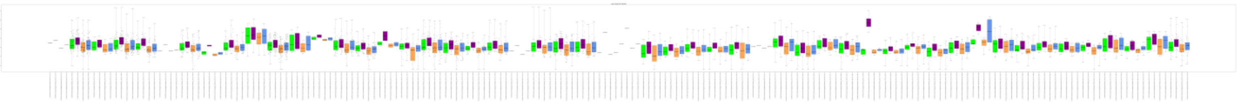
This plot has batches that contain at least the average number of samples and only batches within the given platform. (Filtering is to make "narrow" plots usable. For complete data, see the "wide" plots.) Boxplot data is in terms of ZScore data (to bring out data differences in mutation counts not shown by Log10 or raw scores). For Narrow and Wide Boxplots, the vertical axis is the number of mutations per sample, with the appropriate (Z-Score or Log10(+1)) modification/normalization.



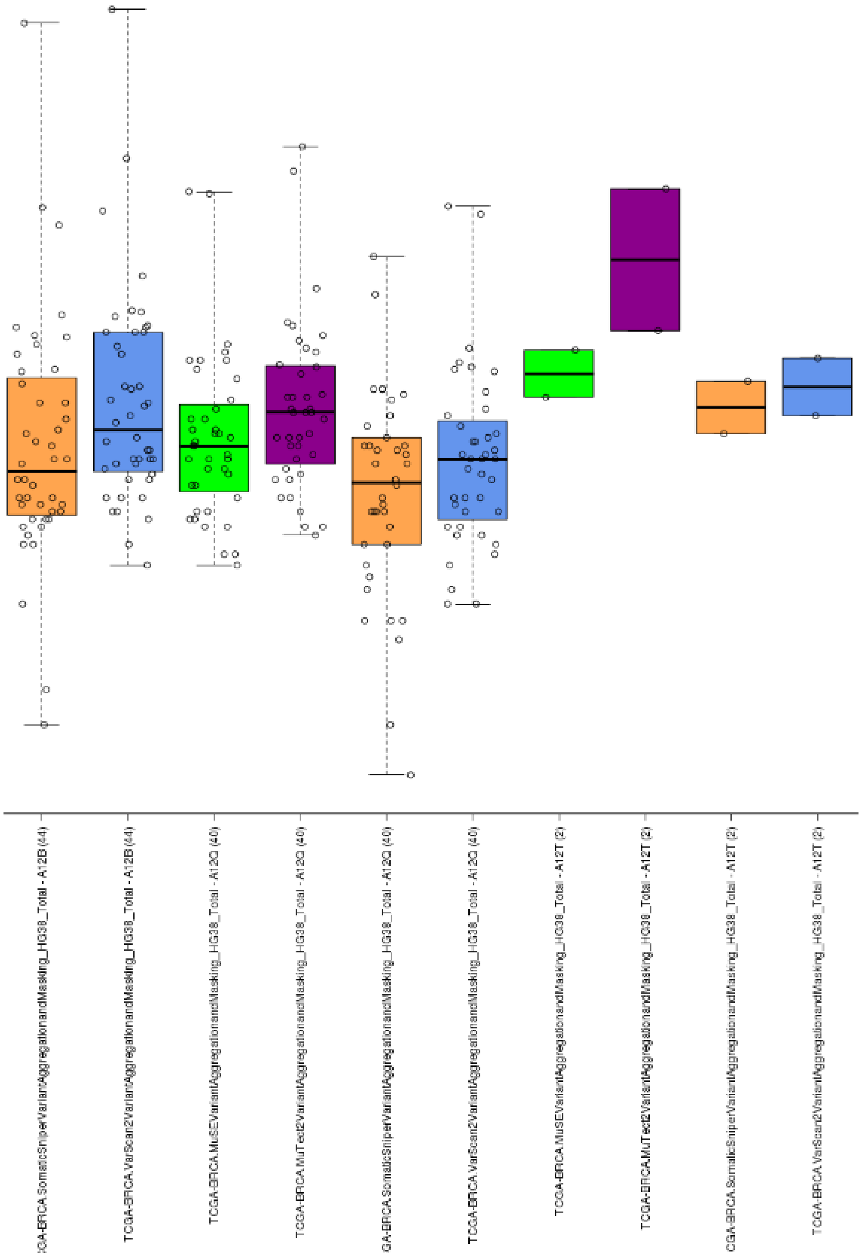
Full Plot

WideBoxplot__PlateId__Total__TCGA-BRCA__Log10.PNG

This plot has all batches for all platforms. (This is to provide all data and to allow comparisons between platforms.) Boxplot data is in terms of Log10 data (to bring out small data differences in mutation counts). For Narrow and Wide Boxplots, the vertical axis is the number of mutations per sample, with the appropriate (Z-Score or Log10(+1)) modification/normalization.



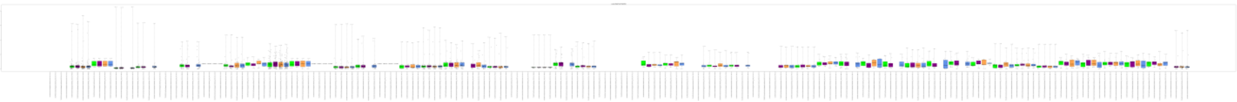
Full Plot



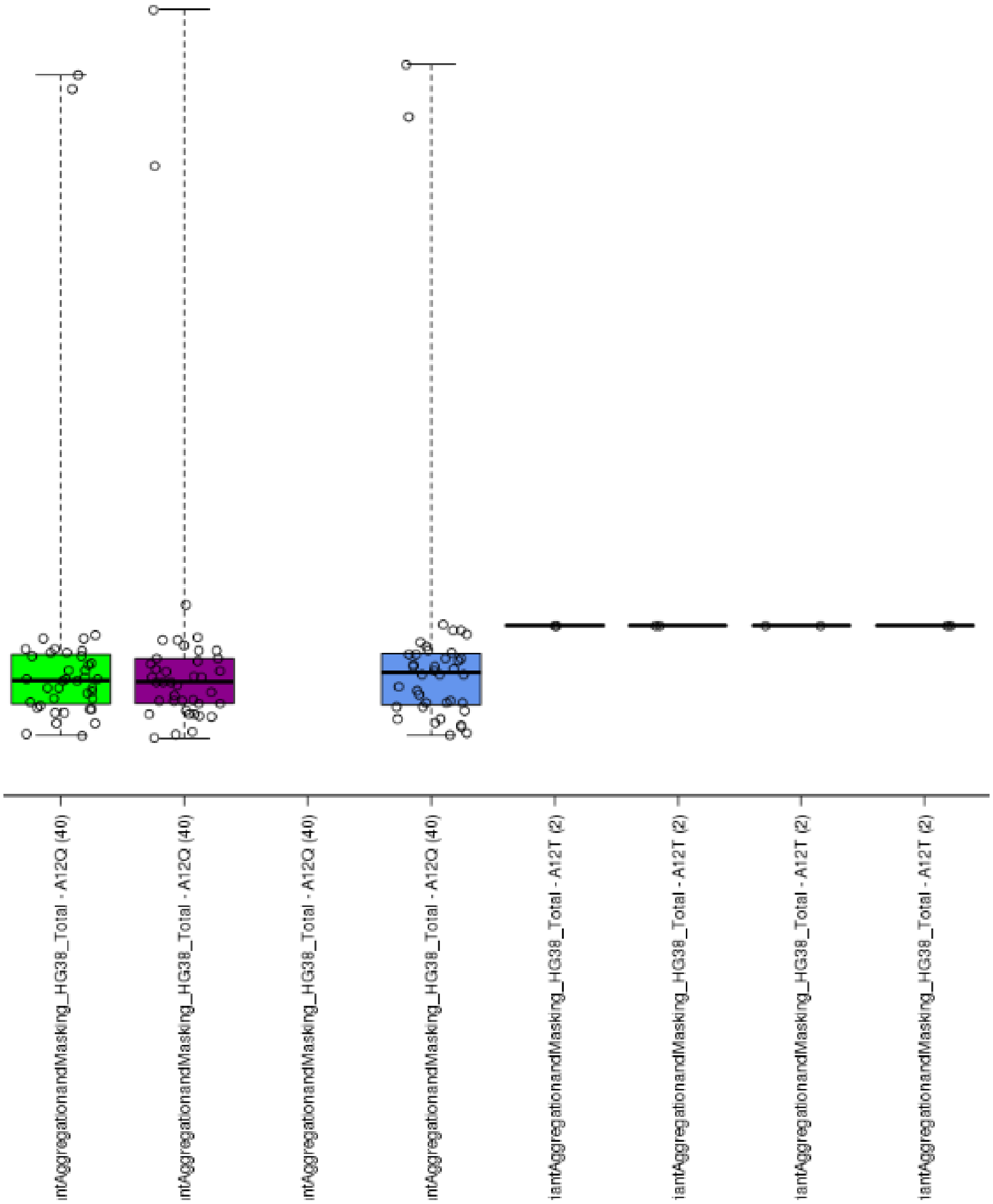
Close up of batches (horizontal) and
number of calls by sample (vertical)

WideBoxplot__PlateId__Total__TCGA-BRCA__ZScore.PNG

This plot has all batches for all platforms. (This is to provide all data and to allow comparisons between platforms.) Boxplot data is in terms of ZScore data (to bring out data differences in mutation counts not visible in Log10 or raw scores). For Narrow and Wide Boxplots, the vertical axis is the number of mutations per sample, with the appropriate (Z-Score or Log10(+1)) modification/normalization.



Full Plot

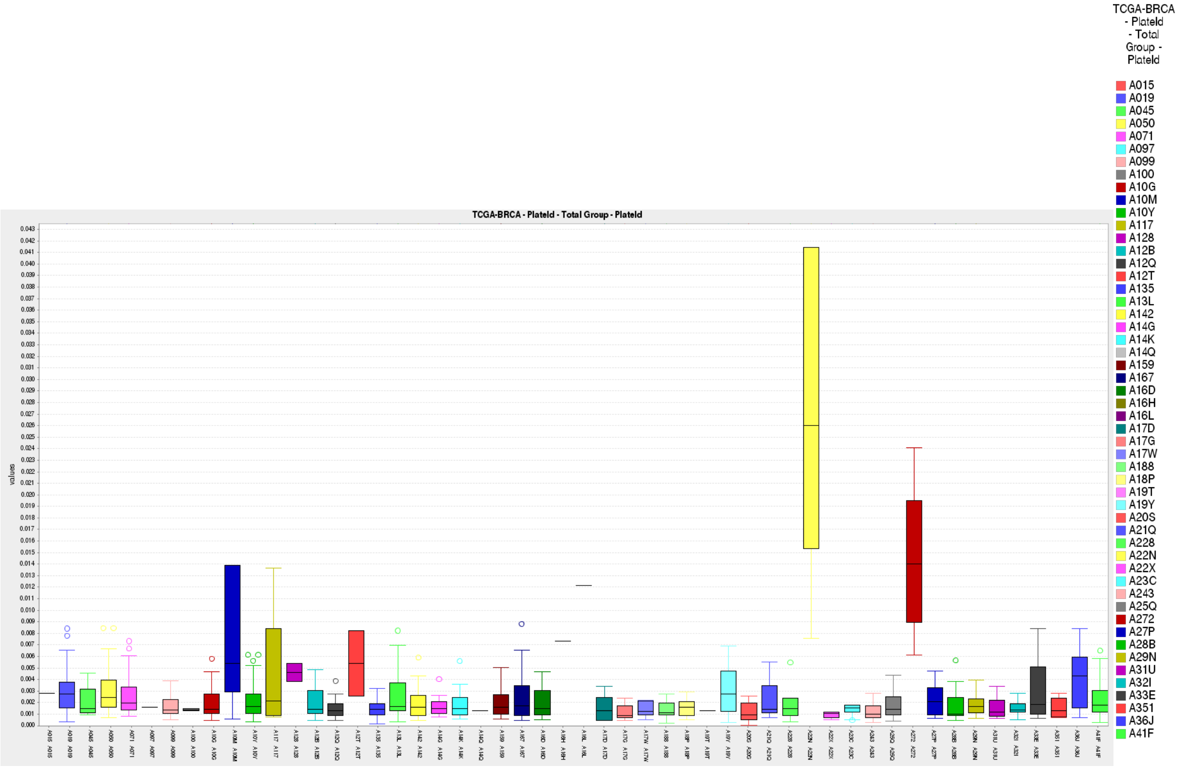


Close up of batches (horizontal) and number of calls by sample (vertical)

BoxPlot__Group-MEAN__Diagram-PlateId.png

BoxPlot__Group-MEAN__Legend-PlateId.png

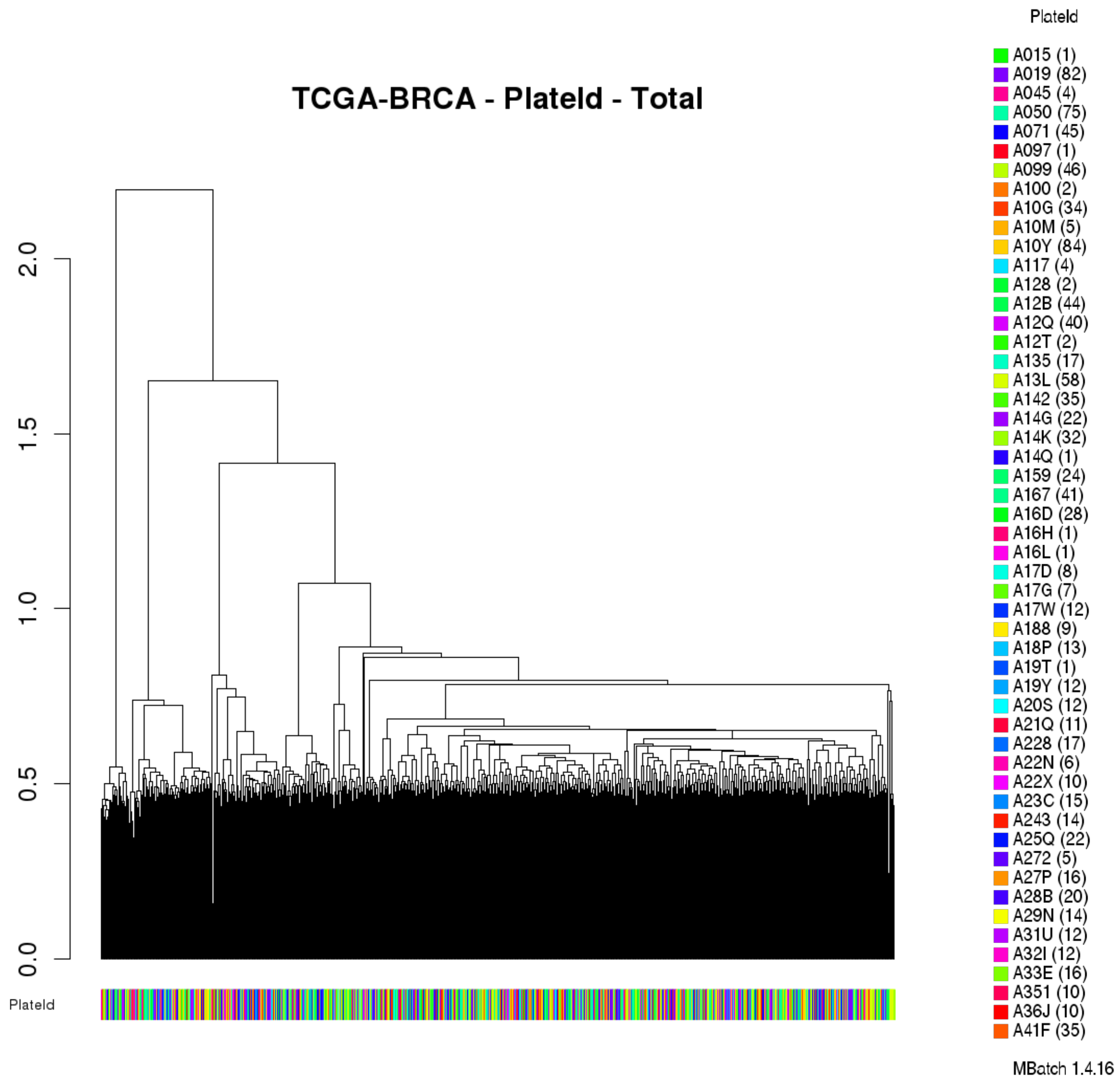
This is another version of the basic boxplot by batch with data using the mean for each group and sample. (Since the mean is used, and there are thousands of genes, the vertical scale is rather small numerically.) All batches are displayed. One PNG is the diagram and the other is the color legend for batches. The Group-MEAN Boxplots plot the mean for each feature within a batch (genes or probes). So the vertical axis is based on the number of mutations and the points plotted are features. Values used are mutation counts.



HierarchicalClustering__Diagram.png

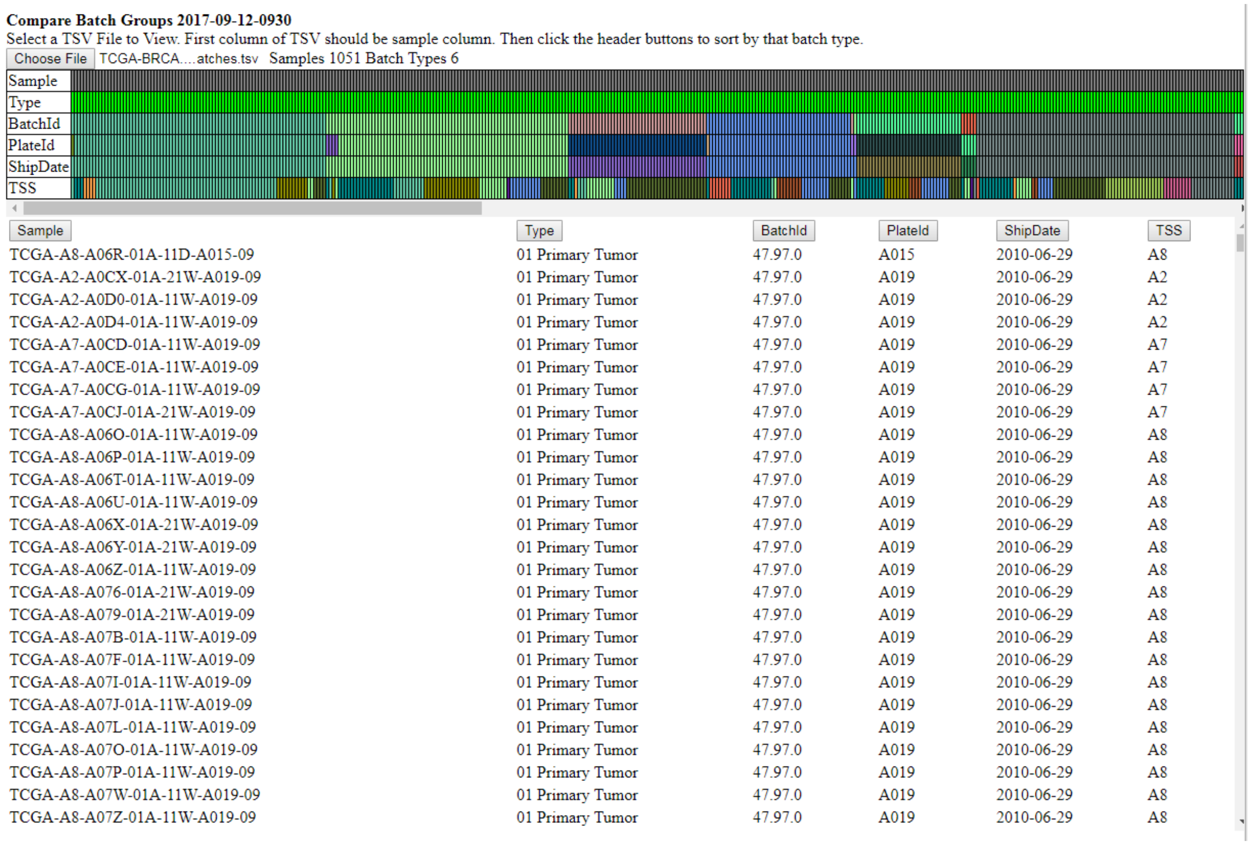
HierarchicalClustering__Legend-PlateId.png

This is hierarchical clustering diagram. This will not always be produced for all datasets. The clustering is not generally useful for what we are looking at, but Keith has noted "potentially interesting" clustering in some datasets.



ViewBatches.html

Here, we used ViewBatches.html and opened the file containing MuTect2 batch information MUT_BEA/2017-11-01-TripleKWDT/GDC_mut/TCGA-BRCA.MuTect2VariantAggregationandMasking_batches.tsv. I then clicked on PlateId to sort by plate id. We then see that much of the early data is confounded by BatchId, PlateId, and ShipDate.



Compare Batch Groups 2017-09-12-0930

Select a TSV File to View. First column of TSV should be sample column. Then click the header buttons to sort by that batch type.

Choose File TCGA-BRCA....atches.tsv Samples 1051 Batch Types 6

Sample	Type	BatchId	PlateId	ShipDate	TSS
TCGA-A8-A06R-01A-11D-A015-09	01 Primary Tumor	47.97.0	A015	2010-06-29	A8
TCGA-A2-A0CX-01A-21W-A019-09	01 Primary Tumor	47.97.0	A019	2010-06-29	A2
TCGA-A2-A0D0-01A-11W-A019-09	01 Primary Tumor	47.97.0	A019	2010-06-29	A2
TCGA-A2-A0D4-01A-11W-A019-09	01 Primary Tumor	47.97.0	A019	2010-06-29	A2
TCGA-A7-A0CD-01A-11W-A019-09	01 Primary Tumor	47.97.0	A019	2010-06-29	A7
TCGA-A7-A0CE-01A-11W-A019-09	01 Primary Tumor	47.97.0	A019	2010-06-29	A7
TCGA-A7-A0CG-01A-11W-A019-09	01 Primary Tumor	47.97.0	A019	2010-06-29	A7
TCGA-A7-A0CJ-01A-21W-A019-09	01 Primary Tumor	47.97.0	A019	2010-06-29	A7
TCGA-A8-A06O-01A-11W-A019-09	01 Primary Tumor	47.97.0	A019	2010-06-29	A8
TCGA-A8-A06P-01A-11W-A019-09	01 Primary Tumor	47.97.0	A019	2010-06-29	A8
TCGA-A8-A06T-01A-11W-A019-09	01 Primary Tumor	47.97.0	A019	2010-06-29	A8
TCGA-A8-A06U-01A-11W-A019-09	01 Primary Tumor	47.97.0	A019	2010-06-29	A8
TCGA-A8-A06X-01A-21W-A019-09	01 Primary Tumor	47.97.0	A019	2010-06-29	A8
TCGA-A8-A06Y-01A-21W-A019-09	01 Primary Tumor	47.97.0	A019	2010-06-29	A8
TCGA-A8-A06Z-01A-11W-A019-09	01 Primary Tumor	47.97.0	A019	2010-06-29	A8
TCGA-A8-A076-01A-21W-A019-09	01 Primary Tumor	47.97.0	A019	2010-06-29	A8
TCGA-A8-A079-01A-21W-A019-09	01 Primary Tumor	47.97.0	A019	2010-06-29	A8
TCGA-A8-A07B-01A-11W-A019-09	01 Primary Tumor	47.97.0	A019	2010-06-29	A8
TCGA-A8-A07F-01A-11W-A019-09	01 Primary Tumor	47.97.0	A019	2010-06-29	A8
TCGA-A8-A07I-01A-11W-A019-09	01 Primary Tumor	47.97.0	A019	2010-06-29	A8
TCGA-A8-A07J-01A-11W-A019-09	01 Primary Tumor	47.97.0	A019	2010-06-29	A8
TCGA-A8-A07L-01A-11W-A019-09	01 Primary Tumor	47.97.0	A019	2010-06-29	A8
TCGA-A8-A07O-01A-11W-A019-09	01 Primary Tumor	47.97.0	A019	2010-06-29	A8
TCGA-A8-A07P-01A-11W-A019-09	01 Primary Tumor	47.97.0	A019	2010-06-29	A8
TCGA-A8-A07W-01A-11W-A019-09	01 Primary Tumor	47.97.0	A019	2010-06-29	A8
TCGA-A8-A07Z-01A-11W-A019-09	01 Primary Tumor	47.97.0	A019	2010-06-29	A8