CEGS Overall Expression Analysis

Table of Contents

[Reference Construction 1](#_Toc388213157)

[Exonic Regions 1](#_Toc388213158)

[Non-Redundant Exonic Regions 2](#_Toc388213159)

[All Possible Junctions 2](#_Toc388213160)

[FASTQ Preparation 2](#_Toc388213161)

[Quality Control 2](#_Toc388213162)

[Duplicate Removal 3](#_Toc388213163)

[Alignments 3](#_Toc388213164)

[Junctions 3](#_Toc388213165)

[Genome 3](#_Toc388213166)

[Exonic Regions 3](#_Toc388213167)

[Coverage Counts 3](#_Toc388213168)

[Junctions 4](#_Toc388213169)

[Genome 4](#_Toc388213170)

[Exonic Regions 4](#_Toc388213171)

[Normalization 4](#_Toc388213172)

[Replicate Level Normalizations 4](#_Toc388213173)

[Remove problem samples 4](#_Toc388213174)

[Remove problem exonic regions 4](#_Toc388213175)

[Upper Quartile Normalization 5](#_Toc388213176)

[Centering 6](#_Toc388213177)

[Mahalanobis Distance 7](#_Toc388213178)

[**Line Level Normalizations** 8](#_Toc388213179)

[**Remove problem samples** 8](#_Toc388213180)

[**Remove problem exonic regions** 9](#_Toc388213181)

[Collapse replicates to line level 9](#_Toc388213182)

[Upper Quartile Normalization 9](#_Toc388213183)

[Centering 10](#_Toc388213184)

[Data Given to USC 11](#_Toc388213185)

[Sample Level Data 11](#_Toc388213186)

[Sample Level Variable names 11](#_Toc388213187)

[Line Level Data 12](#_Toc388213188)

[Sample Level Variable names 13](#_Toc388213189)

# Reference Construction

The FASTA formatted reference and GFF annotations were downloaded from FlyBase (fb5.51). Basic summary of fb5.51:

* Number of genes: 15355
* Number of single-exon genes: 2731
* Number of single-isoform genes: 7737
* Total number of exons: 80895
* Total number of introns: 68910
* Mean exon length: 524.411224427
* Mean intron length: 1599.14391235
* Mean number of isoforms per gene: 1.87671768154
* Mean number of exons per isoform: 5.95277093382
* Number of genes with intron retention: 590
* Number of genes with exon cassettes: 1206
* Number of genes with alt donor: 3163
* Number of genes with alt acceptor: 2530
* Number of genes with alt donor,<10 bases difference: 418
* Number of genes with alt acceptor,<10 bases difference: 730
* Number of genes with alt promoter: 3511
* Number of genes with alt 3'-most exon: 2221
* Number of genes with an internal, sometimes-absent exon of size <50: 520
* Number of genes with an internal, sometimes-absent exon of size between 50 and 100: 1055
* Number of genes with an internal, sometimes-absent exon of size between 100 and 200: 1886
* Min intron length: intron\_FBgn0039307:1\_FBgn0039307:2, length 2
* Max intron length: intron\_FBgn0263780:15\_FBgn0263780:13, length 142627
* Min exon length: FBgn0050423:3, length 1
* Max exon length: FBgn0261836:29, length 28074
* Max isoforms per gene: FBgn0033159: 75
* Max exons per isoform: FBgn0053196, FBtr0332345: 82

## Exonic Regions

Overlapping exons were collapsed into exonic regions (a.k.a. fusions, **Figure 1** red boxes). A unique identifier (fusion\_id) was created for each exonic region (e.g. F10001\_SI). If the first character is ‘S’, the exonic region corresponds to a single exon. If the first character is ‘F’, the exonic region is total region after combining multiple overlapping exons. The ‘SI’ stands for ‘strand independent’ meaning that we did not take into account the strand exons were on. Exonic regions have been grouped into three classes (constitutive, common, and alternative). Constitutive exonic regions contain a single exon that is present in all isoforms (e.g. **Figure 1** exon 4). Common exonic regions have an exon at a given location in all isoforms, but this exon can have alternative donor and acceptor sites (e.g. **Figure 1** exons 5 and 6). Alternative exonic regions are missing an exon in at least one isoform. Finally, given the compactness of the fly genome, some exonic regions contain exons from multiple genes (**Figure 1** blue box). There are a total of 63,706 exonic regions in fb5.51.

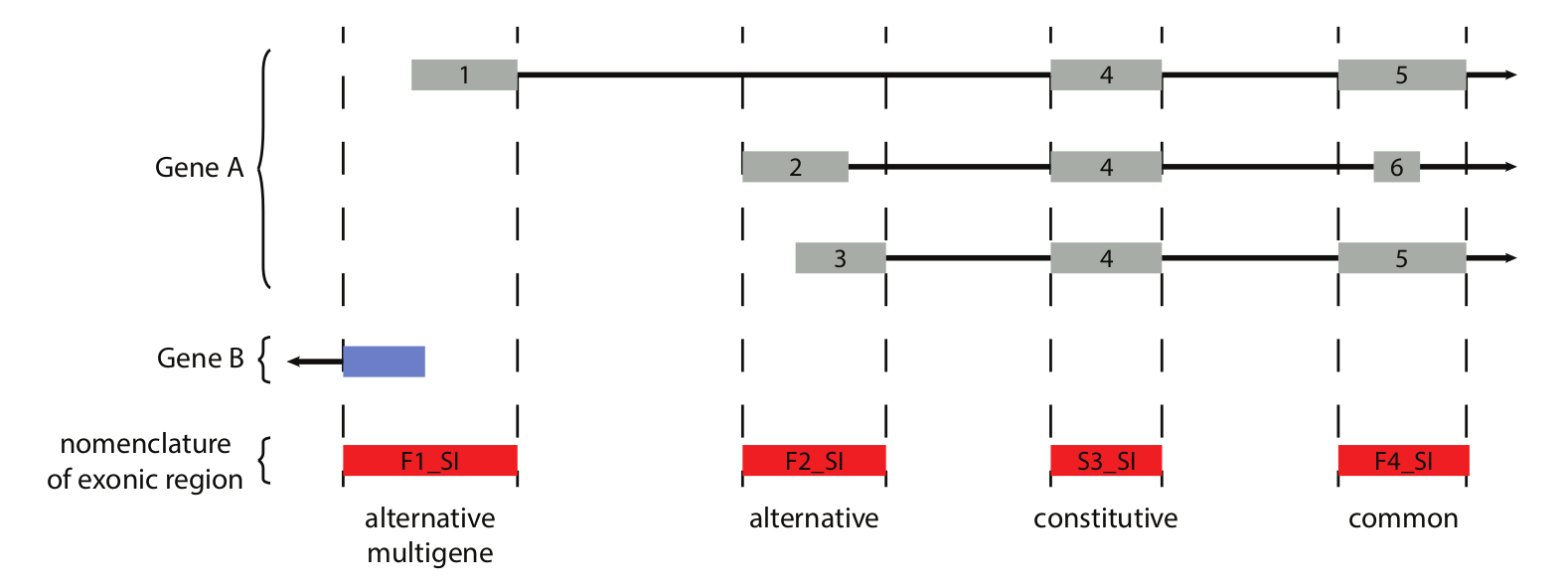


Figure : Exonic regions were created by combining overlapping exons together.

## Non-Redundant Exonic Regions

To reduce mapping ambiguity we identified 762 identical exonic regions. A non-redundant reference was created were only a single representative of each identical set was included. The non-redundant reference contains a total of 63,181 exonic regions.

## All Possible Junctions

Given the size of the CEGS project, instead of creating a set of *de novo* junctions, we created a set ‘logically’ possible junction using flybase annotations. For each gene, 200bp junctions were created using ‘linear’ pairwise combinations of exons. For example, if a gene has 3 exons (E1, E2, E3) we created junctions of (E1E2, E1E3, E2E3). We did not create junctions of (E2E1, E3E1, or E3E2). This junction set has been labeled as “Canonical Junctions”, but “all possible/logical junctions” is probably more accurate.

# FASTQ Preparation

## Quality Control

A large variety of quality control metrics were run on each fastq file. I will not go into these here.

## Duplicate Removal

QC metrics showed a large number of duplicate reads possibly from the library creation process or optical duplicates created by sequencing. We wrote a script to remove reads with identical sequences. For pair-end reads we used both mate-pairs to identify duplicates. A distinct set of reads were created containing all unique reads along with a single representative of duplicate reads.

# Alignments

We performed only single-end alignments. We decided this as a group because of the mixture of single and pair-end reads. Initially we had grouped samples into ‘complete’ and ‘incomplete’ based on the number of bio-reps.

## Junctions

First distinct reads were aligned uniquely to the junction reference using only bowtie (-m1 –v3). Alignment files can be found at:

/auto/cmb-panasas2/cegs/uf\_sync/bam\_fb551\_canonical\_junctions\_nodup  
/auto/cmb-panasas2/cegs/uf\_sync/bam\_fb551\_canonical\_junctions\_incomplete\_nodup

## Genome

Unaligned reads from junction alignments were aligned uniquely to fb5.51 genome reference using bowtie (-m1 –v3) and LAST (-l 25). Alignment files can be found at:

/auto/cmb-panasas2/cegs/uf\_sync/bam\_fb551\_genome\_nodup  
/auto/cmb-panasas2/cegs/uf\_sync/bam\_fb551\_genome\_incomplete\_nodup

## Exonic Regions

Unaligned reads from junction alignments were aligned uniquely to the non-redundant exonic regions using bowtie (-m1 –v3) and LAST (-l 25). Alignments can be found at:

/auto/cmb-panasas2/cegs/uf\_sync/bam\_fb551\_non-redundant\_fusions\_nodup  
/auto/cmb-panasas2/cegs/uf\_sync/bam\_fb551\_non-redundant\_fusions\_incomplete\_nodup

# Coverage Counts

Coverage was calculated for each set of alignments. The coverage tables contain the following information:

* fusion\_id = region that coverage was calculated on. Typically this is the exonic region.
* mapped\_reads = Total number of reads that mapped.
* read\_length = Maximum read length found in a sample (should be ~95bp)
* region\_length = The length of the region that coverage is being calculated on. Typically this is the length of the exonic region
* region\_depth = Sum of depth across each base in a region.
* reads\_in\_region = Estimate of the number of reads are in that region (region\_depth / read\_length).
* apn = Average per nucleotide coverage over the region (region\_detph / region\_length).
* rpkm = region coverage adjusting for number of mapped reads ([reads\_in\_region \* 10^9]/ [mapped\_reads \* region\_legnth])

## Junctions

Coverage was calculated on junctions.

/auto/cmb-panasas2/cegs/uf\_sync/coverage\_count\_fb551\_canonical\_junctions\_incomplete\_nodup  
/auto/cmb-panasas2/cegs/uf\_sync/coverage\_count\_fb551\_canonical\_junctions\_nodup

## Genome

Coverage was calculated on exonic regions.

/auto/cmb-panasas2/cegs/uf\_sync/coverage\_count\_fb551\_genome\_on\_fusions\_incomplete\_nodup  
/auto/cmb-panasas2/cegs/uf\_sync/coverage\_count\_fb551\_genome\_on\_fusions\_nodup

## Exonic Regions

Coverage was calculated on exonic regions.

/auto/cmb-panasas2/cegs/uf\_sync/coverage\_count\_fb551\_non-redundant\_fusions\_on\_fusions\_incomplete\_nodup  
/auto/cmb-panasas2/cegs/uf\_sync/coverage\_count\_fb551\_non-redundant\_fusions\_on\_fusions\_nodup

# Normalization

With the inclusion of plate k we have decided to split the normalization into two parts. (1) excludes plate k and replicates will be maintained separately. (2) includes plate k but everything has been summarized to the line level. **We are only doing normalizations on coverage counts on exonic regions from alignments to exonic regions.**

## Replicate Level Normalizations

### Remove problem samples

**Flag\_sample\_on.** There were a total of 565 samples prior to normalization. 97 samples were removed because of low coverage, while the remaining 468 samples had acceptable coverage. Acceptable coverage was defined as have an APN > 0 in more than 29,300 exonic regions (**Figure 2**).

**Flag\_contaminated**. The group has generated a list of samples that should be removed due to various issues including contamination and matedness. Of the 468 samples with acceptable coverage 123 samples dropped because of contamination, **leaving 345 samples**.

### Remove problem exonic regions

**Flag\_drop\_fusion*.*** Next we removed exonic regions that had low coverage across the remaining samples. An exonic region was considered expressed if it had an APN > 0. An exonic region was on in a line\*mating\_status if it was expressed in >50% of replicates. Fusions that were not on in 90% of line\*mating\_status were removed. Of the 63,181 non-redundant exonic regions; 32,102 exonic regions were removed from mated and 30,459 exonic regions were removed from virgin. **Leaving a total of 31,079 (32,722) exonic regions for mated (virgin)**.

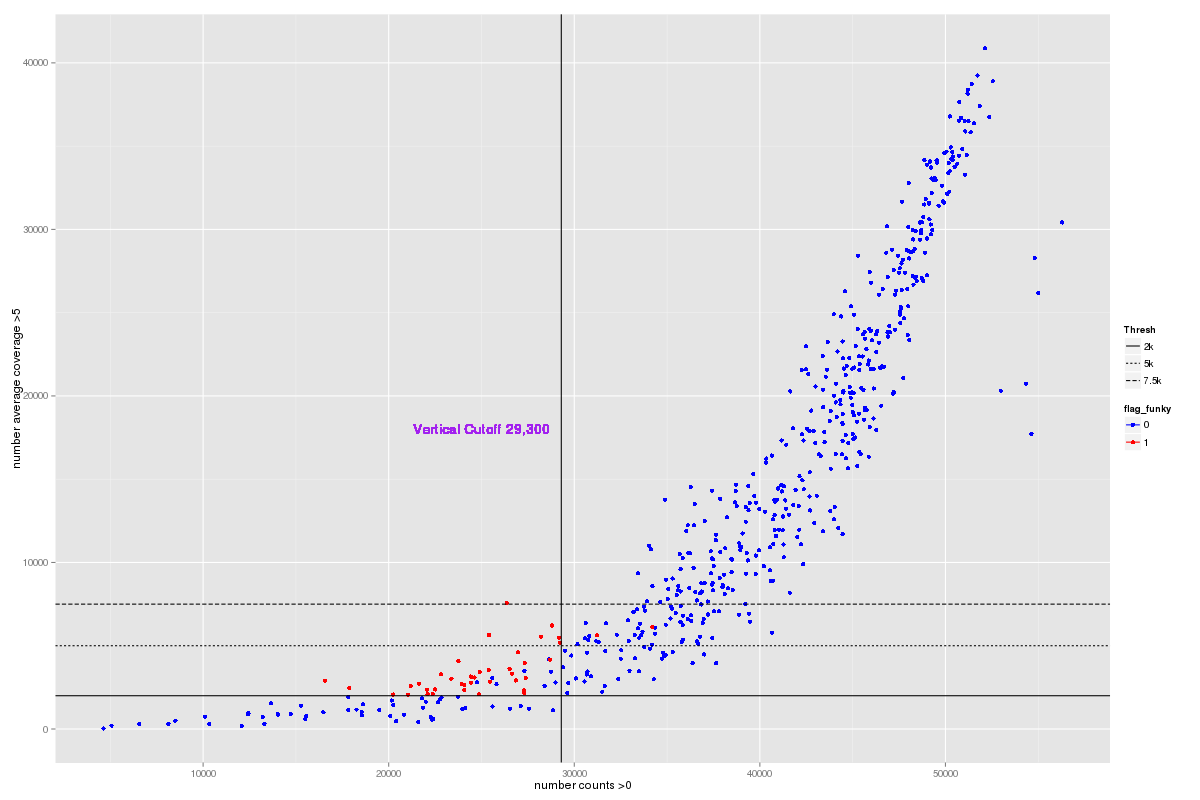


Figure : Number of expressed exons. Each sample is plotted according to the number of exonic regions with and APN>0 (x-axis) vs APN >5 (y-axis). A coverage cutoff at 29,300 exons with APN >0 (vertical line) was established by the group.

### Upper Quartile Normalization

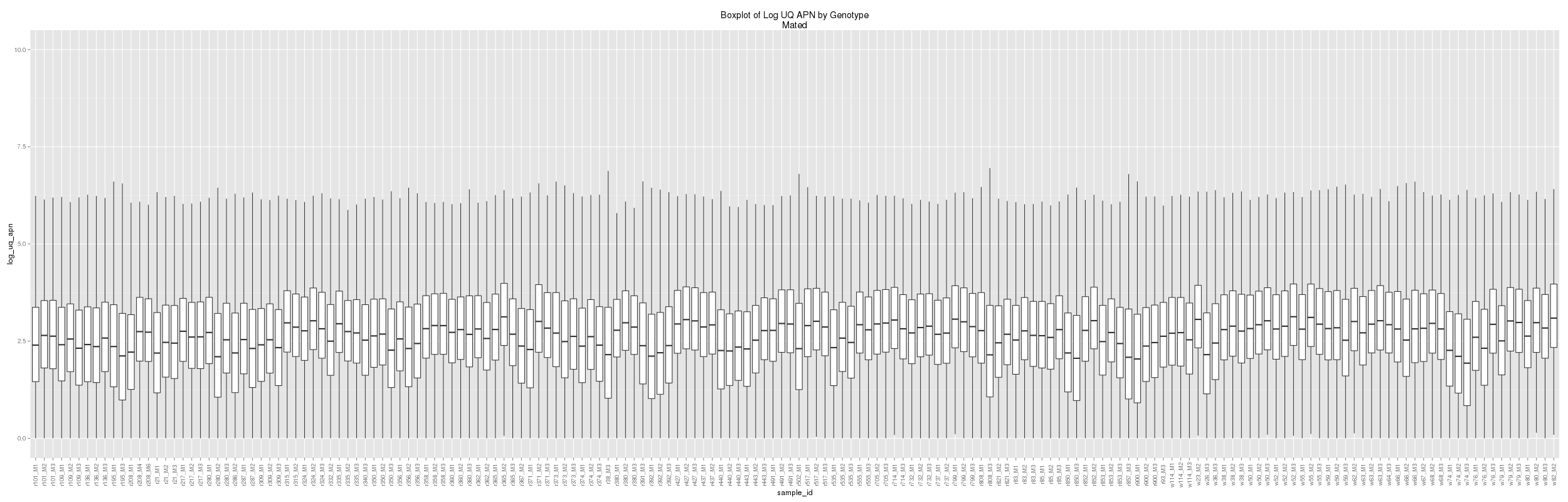
For each sample *i*, we calculate the upper quartile (q3i) for the number of reads in each exonic region *j* (reads\_in\_region). The median of the q3i within mating status was then calculated ( and ). Normalized coverage counts were calculated by:

uq\_apn­ij

Then the log was taken:

log\_uq\_apnij = log2(uq\_apnij + 1)

Uq normalization does a good job at normalizing sample variances; however plots show a need for centering (**Figure 3**).



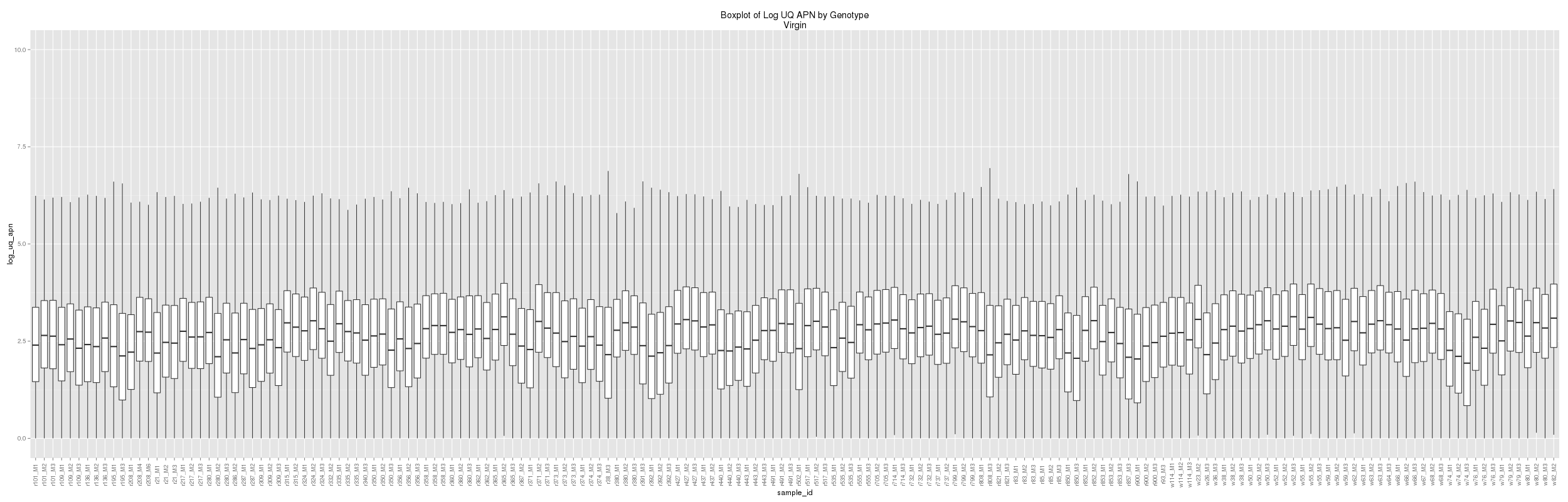


Figure : Sample distributions of log\_uq\_apn for Mated (top) and Virgin (bottom).

### Centering

We centered samples using the upper quartile of log\_uq\_apn.

**uq\_log\_uq\_centerij** = log\_uq\_apnij – q3(log\_uq\_apnij)

The majority of the samples look good after centering. There are a few samples that look a little off.

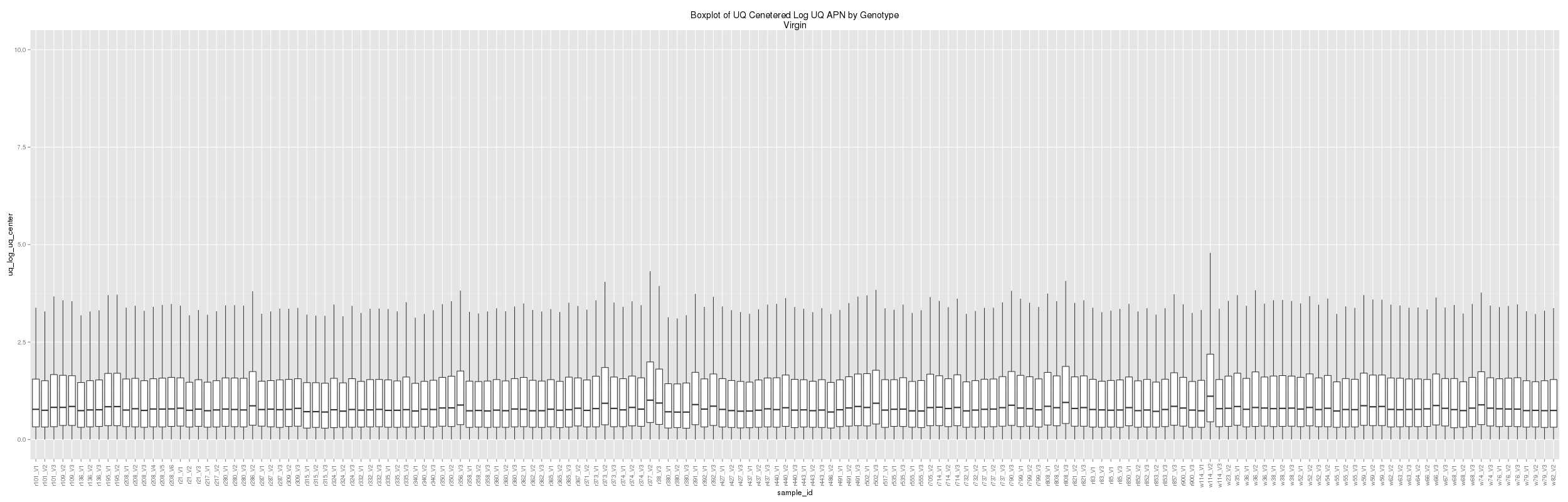
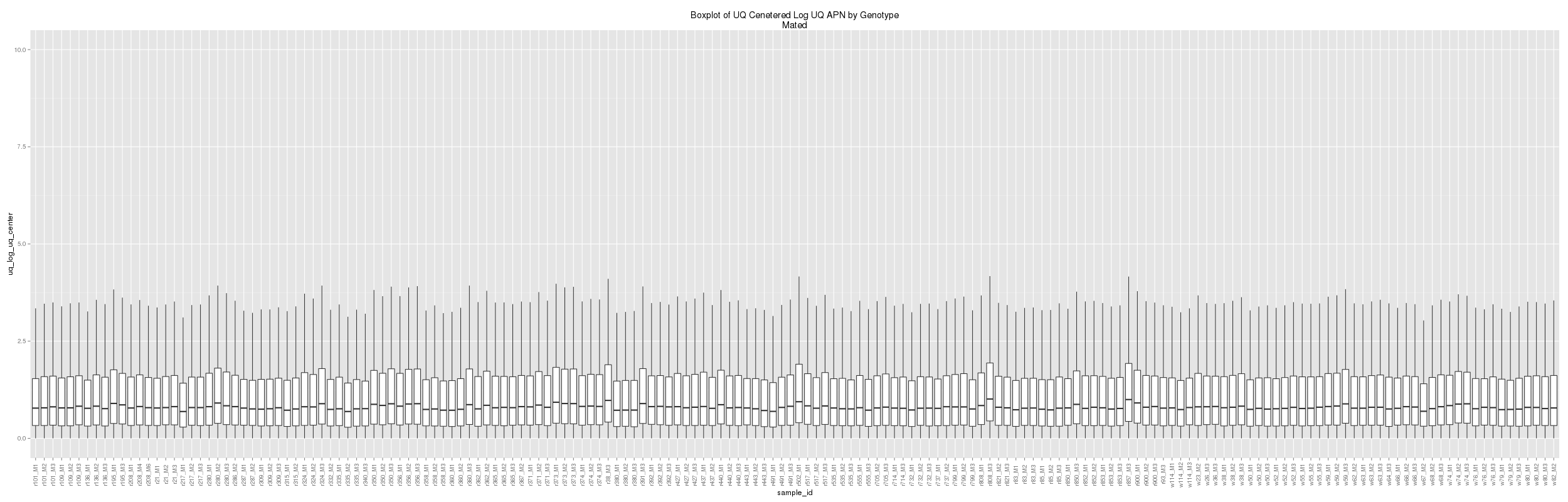


Figure : Sample distributions of uq\_log\_uq\_center for Mated (top) and Virgin (bottom).

### Mahalanobis Distance

Next I looked at the Mahalanobis distance (MD) to try to identify any potential outliers. MD was calculated across samples and plotted (**Figure 5** and **Figure 6**). There were 47 samples that were outliers according to the MD cutoff ~4. I created an outlier flag (**flag\_mahalanobis\_outlier**), just so we know that these samples behaved strangely. I plotted this outlier flag against centered values (uq\_log\_uq\_center) and did not see any distributional differences between outlier and non-outliers (**Figure 7**).

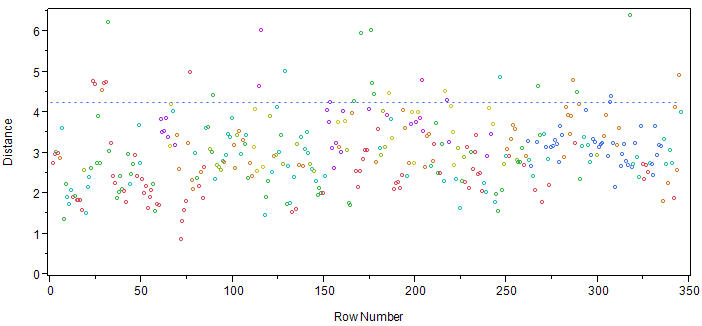


Figure : Mahalanobis distance by sample.

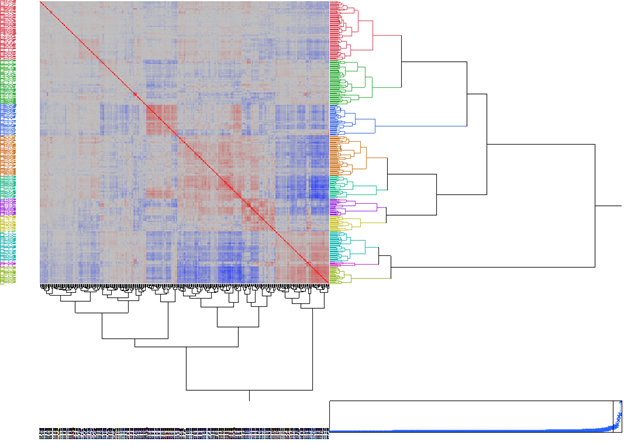


Figure : Heatmap of Mahalanobis distances sorted using hierarchical clustering.

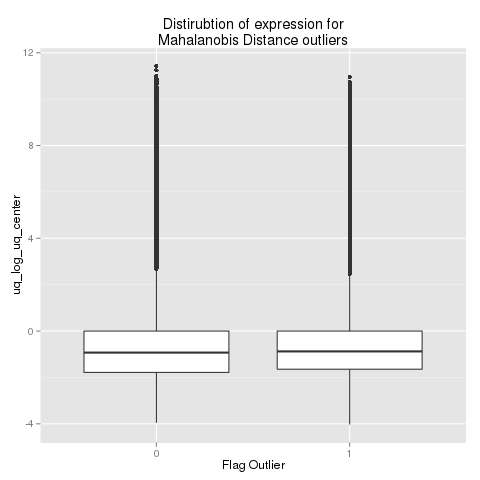


Figure : Distribution of centered values between samples identified as outliers and not identified as outliers.

**Line Level Normalizations**

For line level normalization I have added replicates from Plate K.

**Remove problem samples**

**Flag\_sample\_on.** There were a total of 632 samples prior to normalization. 104 samples were removed because of low coverage, while the remaining 528 samples had acceptable coverage. Acceptable coverage was defined as have an APN > 0 in more than 29,300 exonic regions (**Figure 2**).

**Flag\_contaminated**. The group has generated a list of samples that should be removed due to various issues including contamination and matedness. Of the 528 samples with acceptable coverage 123 samples dropped because of contamination, **leaving 405 samples**.

**Remove problem exonic regions**

**Flag\_drop\_fusion*.*** Next we removed exonic regions that had low coverage across the remaining samples. An exonic region was considered expressed if it had an APN > 0. An exonic region was on in a line\*mating\_status if it was expressed in >50% of replicates. Fusions that were not on in 90% of line\*mating\_status were removed. Of the 63,181 non-redundant exonic regions; 24,689 exonic regions were removed from mated and 26,377 exonic regions were removed from virgin. **Leaving a total of 37,442 (35,676) exonic regions for mated (virgin)**.

### Collapse replicates to line level

It is unclear which ‘biological replicates’ were run on plate k. We discussed this as a group and decided to collapse all replicates to line\*mating\_status level. For each exonic region I summed region\_depth across replicates and calculated APN = region\_depht/region\_length.

### Upper Quartile Normalization

For each line\*mating\_status *i*, we calculate the upper quartile (q3i) for the number of reads in each exonic region *j* (reads\_in\_region). The median of the q3i within mating status was then calculated ( and ). Normalized coverage counts were calculated by:

uq\_apn­ij

Then the log was taken:

log\_uq\_apnij = log2(uq\_apnij + 1)

Uq normalization does a good job at normalizing sample variances; however plots show a need for centering (**Figure 8**).

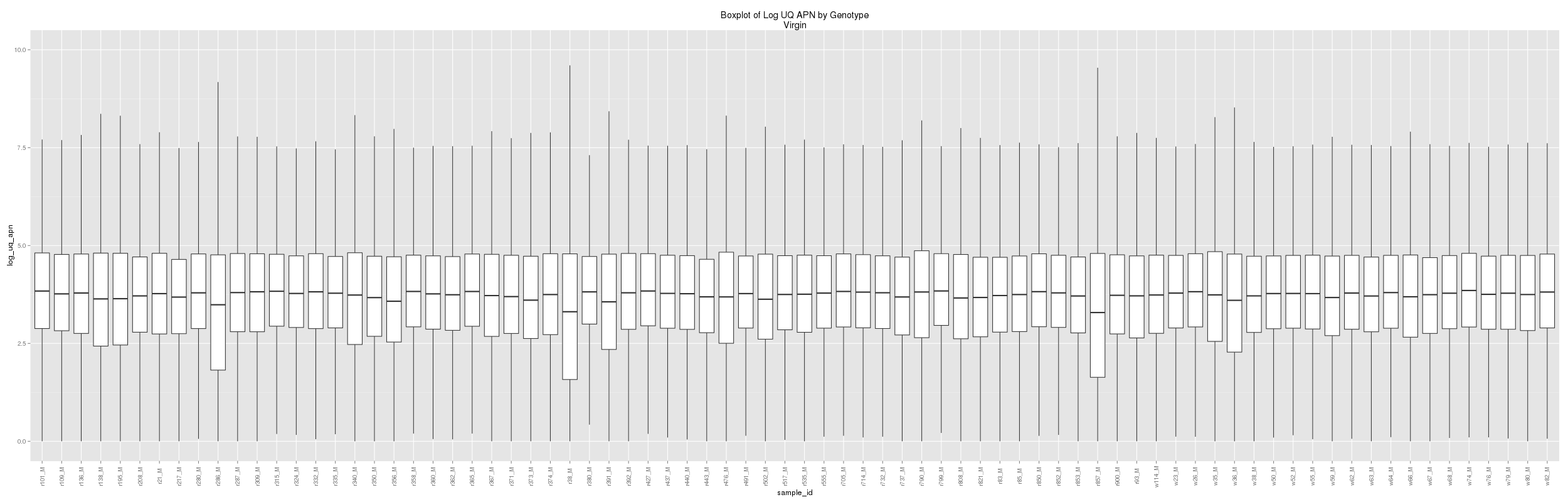
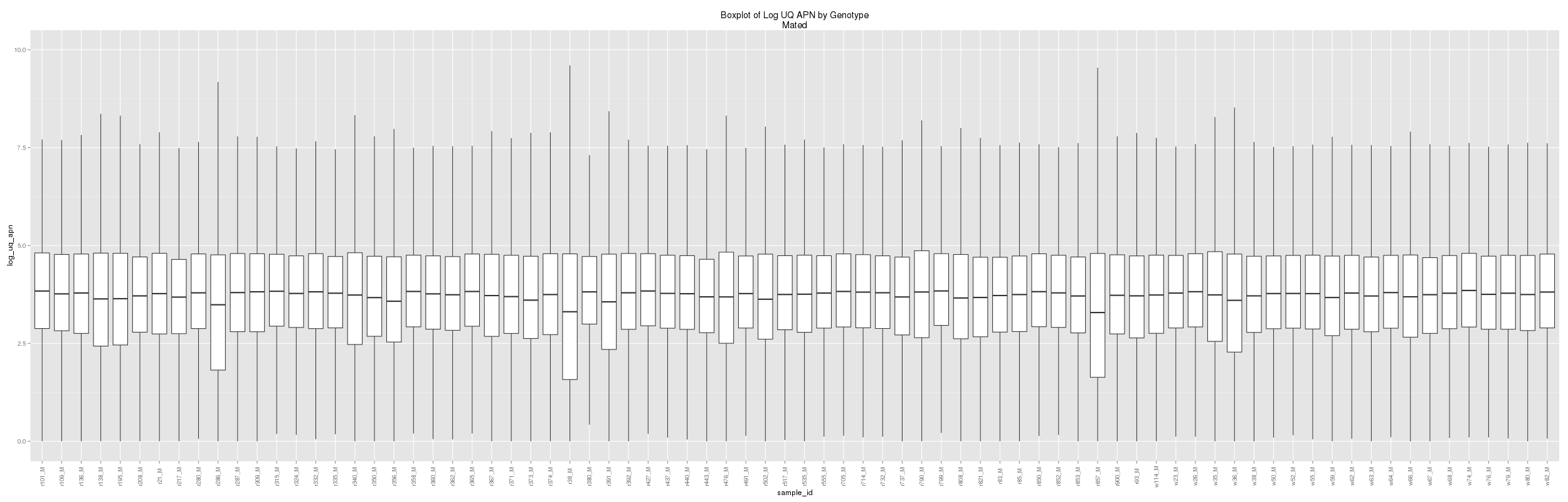


Figure 8: Line distributions of log\_uq\_apn for Mated (top) and Virgin (bottom).

### Centering

We centered lines using the upper quartile of log\_uq\_apn.

**uq\_log\_uq\_centerij** = log\_uq\_apnij – q3(log\_uq\_apnij)

The majority of the lines look good after centering (**Figure 9**).

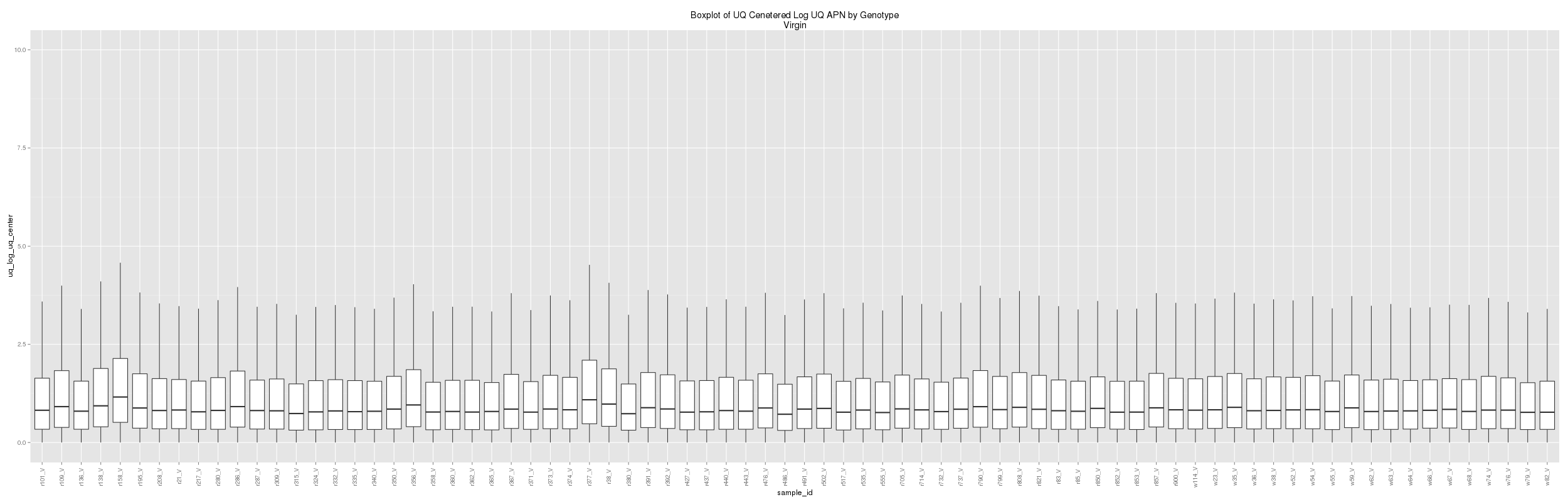
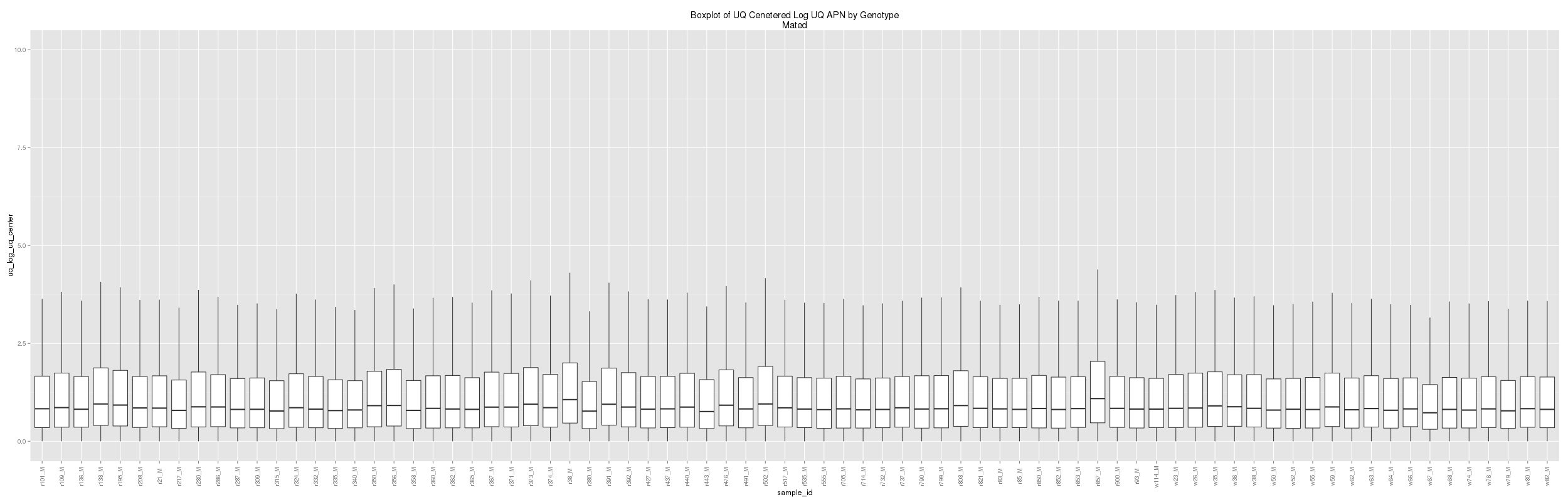


Figure 9: Line distributions of uq\_log\_uq\_center for Mated (top) and Virgin (bottom).

# Data Given to USC

## Sample Level Data

I have packaged and shipped the following:

/auto/cmb-panasas2/cegs/uf\_sync/normalization/rawAPN\_uqNormCenter\_plusFlags\_20140518.csv

Variable names are exactly the same as before except flag\_funky was removed and replaced with flag\_mahalanobis\_outlier (see below). We recommend that you use uq\_log\_uq\_center as your measure of gene expression.

All of the data is included in this file. So you will have to filter the file before running any analysis. Flags to filter on:

flag\_sample\_on = 1

flag\_drop\_fusion = 1

You may also consider removing flag\_mahalanobis\_outlier = 1, but these are included in the normalizations.

### Sample Level Variable names

line = genotype

mating\_status = {mated, virgin}

rep = replicate

flag\_raleigh = 1 if genotype is raleigh, 0 otherwise

flag\_incomplete = 1 if the line\*mating\_status\*rep is not >= 3, 0 otherwise

flag\_partial = 1 if a particaular line\*mating\_status\*rep is >= 3 (ie if theare are >= 3 reps for Mated OR Virgin)

flag\_complete = 1 if line\*mating\_status\*rep is >= 3 for mated AND virgin (ie all 6 reps).

fusion\_id = exonic region id

mapped\_reads = Number of mapped reads from alignment to the non-redundant exonic region set

read\_length = read lengh

region\_length = length of the exonic region

region\_depth = sum of the per base count across a exonic region

reads\_in\_region = region\_depth / read\_length (ie an estimate of coverage).

apn = Average Per Nucleotide Coverage (ie reads\_in\_region / region\_length).

rpkm = RPKM (ie APN \* 10^9/mapped\_reads)

flag\_sample\_on = 1 if >=29300 exonic regions have APN > 0

flag\_fusion\_on = 1 if exonic region was on for a line\*mating\_status. Exonic region was considered on for a line\*mating\_status if APN > 0 in 2/3 replicates

flag\_drop\_fusion = 1 if flag\_fusion\_on was 1 in <90% of replicates

flag\_mahalanobis\_outlier = 1 sample was an outlier

sum\_mapped = sum(reads\_in\_region) for each sample (line\*mating\_status\*rep)

q3 = q3(reads\_in\_region) for each sample (line\*mating\_status\*rep)

median = med(reads\_in\_region) for each sample (line\*mating\_status\*rep)

junc\_mapped\_reads = number of mapped reads when aligned to the canonical Junctions

total\_mapped\_reads = total number of mapped reads (ie mapped\_reads + junc\_mapped\_reads)

uq\_apn = APN \* uq\_ff

uq\_ff = UQ fudge factor (ie med(Q3\_matingStatus) / q3\_matingStatus\_line)

log\_uq\_apn = log2(uq\_apn)

mean\_log\_uq\_apn = mean(log\_uq\_apn) for each sample (line\*mating\_status\*rep)

median\_log\_uq\_apn = median(log\_uq\_apn) for each sample (line\*mating\_status\*rep)

uq\_log\_uq\_apn = q3(log\_uq\_apn) for each sample (line\*mating\_status\*rep)

mean\_log\_uq\_center = log\_uq\_apn - mean\_log\_uq\_apn

median\_log\_uq\_center = log\_uq\_apn - median\_log\_uq\_apn

uq\_log\_uq\_center = log\_uq\_apn - uq\_log\_uq\_apn

symbol\_cat = gene symbols associated with the exonic region

FBgn\_cat = FBgn numbers associated with the exonic region

Constitutive = 1 if an exonic region includes only a single exon that is in all transcript isoforms, 0 otherwise

Common = same as constitutive except exonic region includes multiple overlapping exons merged together, 0 otherwise

Alternative = 1 if exonic region is not in all transcript isoforms

## Line Level Data

I have packaged and shipped the following:

/auto/cmb-panasas2/cegs/uf\_sync/normalization/line\_level\_uq\_normCenter\_plusFlags\_20140518.csv

Unlike the sample data, these data are already filtered and have fewer variables in them. We recommend that you use uq\_log\_uq\_center as your measure of gene expression.

### Sample Level Variable names

line = genotype

mating\_status = {mated, virgin}

fusion\_id = exonic region id

Constitutive = 1 if an exonic region includes only a single exon that is in all transcript isoforms, 0 otherwise

Common = same as constitutive except exonic region includes multiple overlapping exons merged together, 0 otherwise

Alternative = 1 if exonic region is not in all transcript isoforms

Sum\_region\_depth = sum of the per base count across a exonic region (summed across reps)

Sum\_mapped\_reads = Number of mapped reads from alignment to the non-redundant exonic region set (summed across reps)

uq\_apn = APN \* uq\_ff

log\_uq\_apn = log2(uq\_apn)

mean\_log\_uq\_center = log\_uq\_apn - mean\_log\_uq\_apn

median\_log\_uq\_center = log\_uq\_apn - median\_log\_uq\_apn

uq\_log\_uq\_center = log\_uq\_apn - uq\_log\_uq\_apn

symbol\_cat = gene symbols associated with the exonic region

FBgn\_cat = FBgn numbers associated with the exonic region