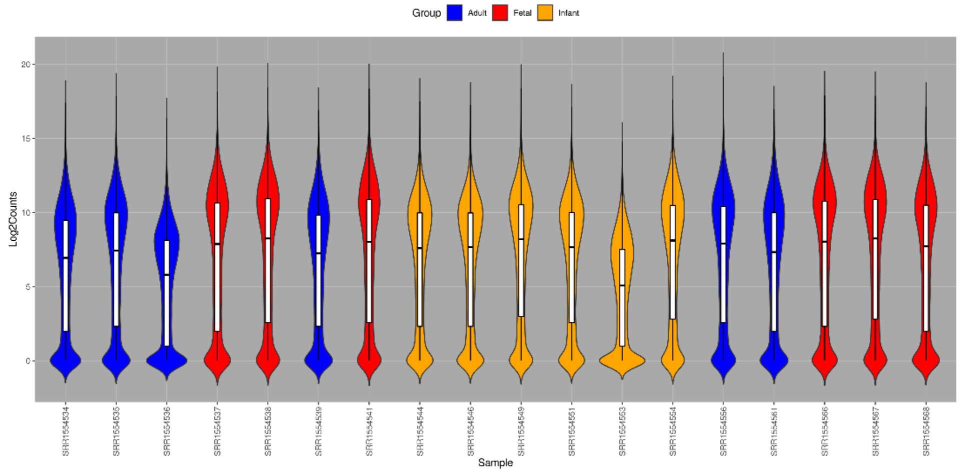
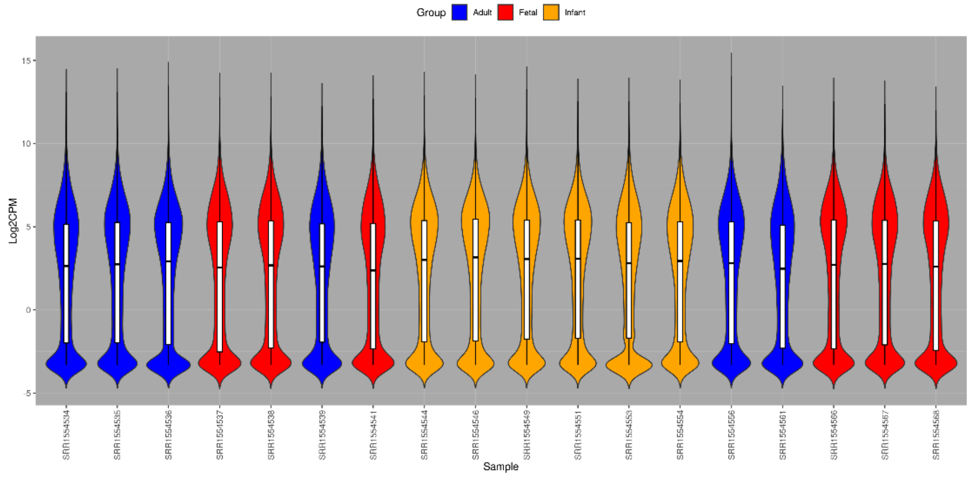
I would usually plot a mix of boxplots and density distributions, in part because log2-transformed expression often has a bimodal distribution (not really captured in the regular boxplot). So, as a compromise, if I only show 1 plot per normalization, then I am using **violin plots** instead of boxplots. I still ***plotted a regular boxplot within the violin plot***, but the boxplot part is relatively smaller.

In the plots below, you can see the effect of normalizing based upon the total quantified count of reads per sample (in millions), referred to as Count-Per-Million (CPM).

Violin Plots for raw log2(count+1) values



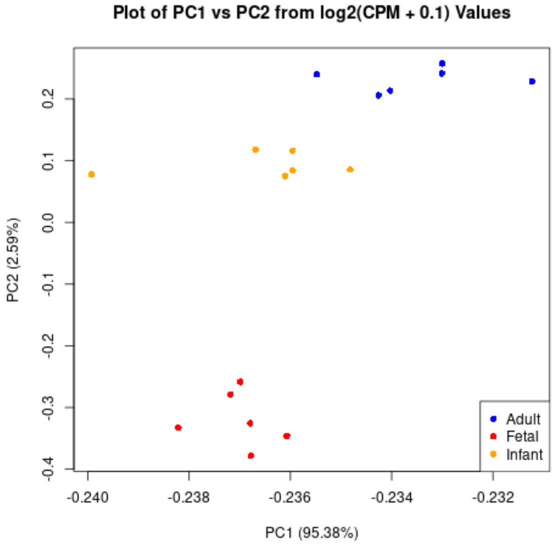
Violin Plots for log2(CPM+0.1) values



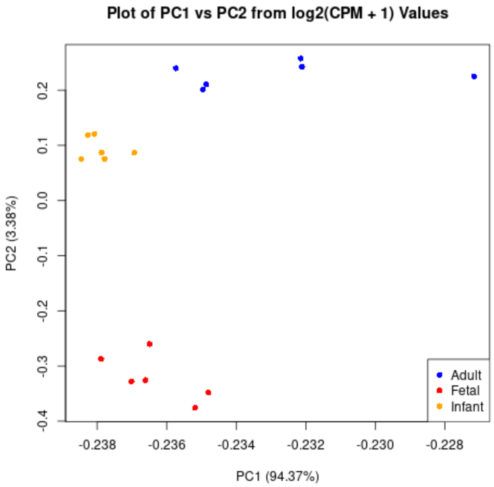
As you can see, the distributions are more similar after normalizing to the total million reads.

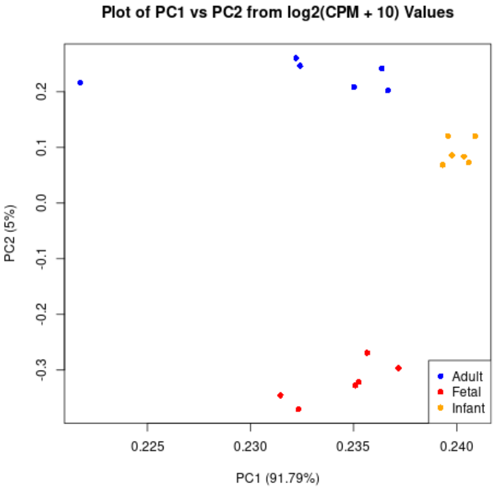
From this view, it looks like SRR1554536 (with a lower percent quantified reads as well as a lower GenomicAlignments alignment rate) is more similar with normalization. However, I will keep an eye on this sample in subsequent analysis.

At least in the PCA plot, SRR1554536 doesn't seem to be a strong outlier:



In the 7th course, I noticed that filtering for more highly expressed genes substantially affected the differential expression results. So, for comparison, I have also plotted PC1 and PC2 with higher rounding thresholds (of CPM + 1 on the left, and CPM + 10 on the right):

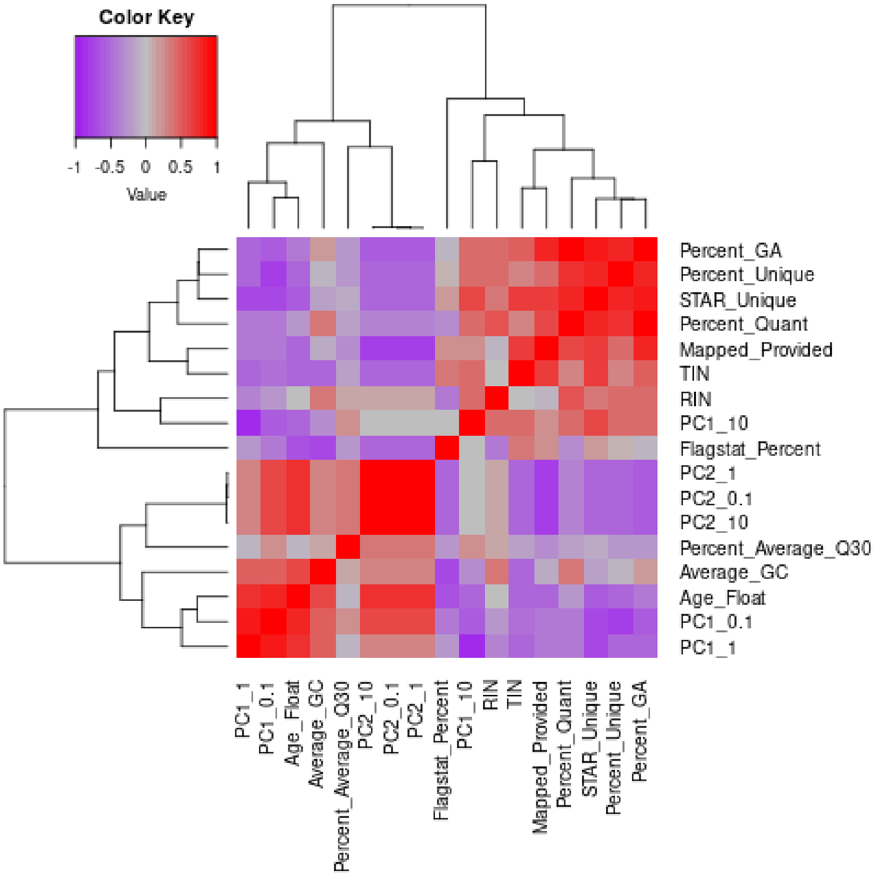




There is more of an Adult outlier in the log2(CPM+10) plot. SRR1554536 is an Adult sample, but this outlier is SRR1554553 (with SRR1554536 being the Adult sample with the highest PC1 value). However, as a little bit less extreme outlier, SRR1554536 is the sample with the highest PC1 value when using log2(CPM + 1) values.

Given the variably based upon the input file, I think continuing the watch for the effect on downstream analysis is important.

For comparison, the previously calculated variables can be tested for correlations with the 3 sets of PC1 and PC2 values:



There is a full legend for the values in the GitHub page for this week (if you scroll down), but this mostly shows what you would expect from the previous report and the current report. The GitHub content is also generally in preparation for providing reproducible content for Week 10.

Age is the variable that we want to study, and the PCA plots show a clear association with PC2 (with "*Age_Float*" in the correlation heatmap). The correlation coefficient for PC2 is positive in all situations, but higher for the first 2 PC1 values. This is probably largely because of the adult samples (where age is also the most different, in absolute terms). The 3rd PC1 value an outlier. So, if we filter for more highly expressed genes, perhaps emplacing those with expression whose **CPM is greater than 1 is sufficient**). That is admittedly where is the greatest outlier, but that sample could potentially be removed if it looks like it is biasing the differentially expressed gene lists.

For a given set of differentially expressed genes, I expect being able to view various variables along with expression in the heatmap (as well as the 6 infant samples not directly used in the statistical comparison) may help with troubleshooting in the next step.

TIN is more correlated with possible confounding variables. So, if a second variable is added for differential expression, perhaps I will use TIN instead of RIN.