Given the fairly robust separation by group by PC2 in last week's report, I hypothesized that you might be able to define genes to separate **Adult vs Fetal** samples ***without*** **including covariates**. If there is a highly confounded variable, then I believe that it may not be possible to confidently separate that effect in the statistical model. However, I use a heatmap for visual inspection of replicate consistency.

I think it is important to do some methods testing for every project. So, I will choose 1 set of criteria for generating the requested volcano plot, but I will first show a summary table under various conditions:

Method	Criteria	Adult-Up (+)	Fetal-Up (-)	SOX11 Adult-Down?
edgeR	FDR < 0.01	5,176 genes	4,963 genes	Yes
	FDR < 0.01 Max Quant CPM ≥ 5	4,135 genes	4,151 genes	Yes
	FDR < 0.01 Max Quant CPM ≥ 5 |fold-change| ≥ 2	3,335 genes	2,951 genes	Yes
DESeq2	FDR < 0.01	7,344 genes	6,825 genes	Yes
	FDR < 0.01 Max Quant CPM ≥ 5	4,528 genes	4,266 genes	Yes
	FDR < 0.01 Max Quant CPM ≥ 5 |fold-change| ≥ 2	3,515 genes	2,853 genes	Yes
limma-voom	FDR < 0.01	4,574 genes	4,977 genes	Yes
	FDR < 0.01 Max Quant CPM ≥ 5	3,454 genes	4,396 genes	Yes
	FDR < 0.01 Max Quant CPM ≥ 5 |fold-change| ≥ 2	2,797 genes	3,195 genes	Yes

False Discovery Rate (FDR) values are calculated using the method of Benjamini and Hochberg 1995 for all 3 programs. In order to make volcano plot creation easier, log2ratio values are saved (where **|log2ratio| ≥ 1** is comparable to |fold-change| ≥ 2).

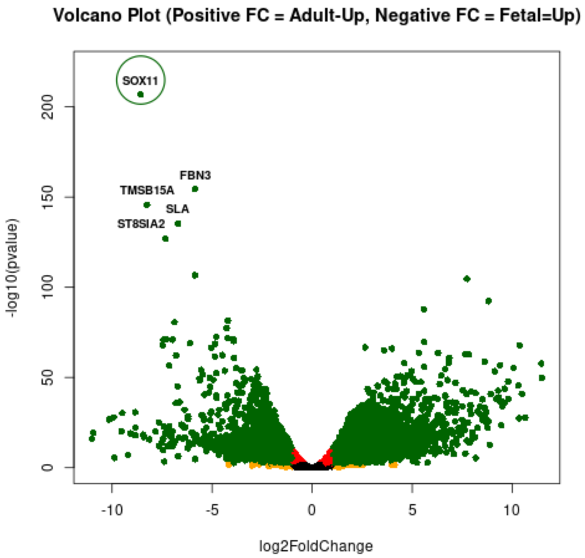
I would typically use an independently calculated fold-change value, but I have simplified the code (written more from scratch) by using the fold-change or log2ratio provided by the given program. Since SOX11 was mentioned in Figure 1 (and showed good visual inspection of the alignment), I am using that as a positive control that I think should be in the gene list.

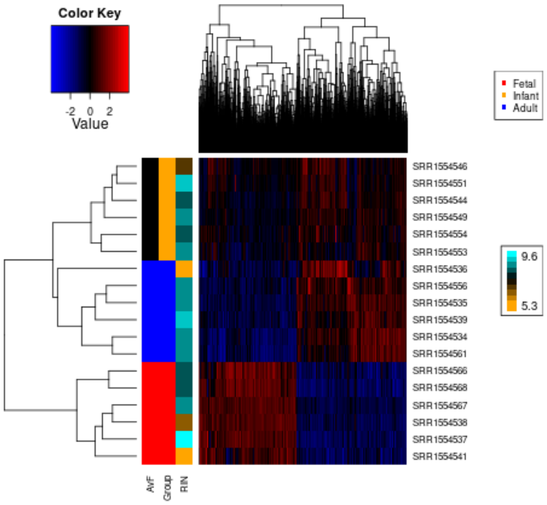
There are usually multiple ways to calculate p-values within a package, so you would have to check the code (and pre-processing steps) to see if results should be similar or identical. For example, I used TMM normalization (Robinson and Oshlack 2010) when testing edgeR, but that is not a strict requirement. So, along with preparation to Week 10, code is hosted for all weeks and this week.

While not directly shown in the table above, I think using FDR < 0.01 (versus FDR < 0.05) helped some with getting the filtered gene sets to be even more similar. Nevertheless, given the slightly more symmetric gene list and the expectation that TMM normalization should help if the underlying distribution changes (even if it might also be a slightly over-correction), I chose to use **edgeR**.

As a general rule of thumb, I think having 100s or ~1000 genes for each up-regulated and down-regulated gene list is helpful for gene enrichment. In this situation, I think there really are a large number of true differences. Nevertheless, assuming the possibility of having more specific enrichment values could help, I am using **FDR < 0.01**, **Max Quantified CPM ≥ 5**, **|fold-change| ≥ 2**.

It can be helpful to use filters beyond just a FDR threshold, and that is part of the basis for the volcano plot comparing the -log10(p-value) and the log2ratio (shown **below, left**):





Whether or not the p-values are "reasonable" is arguably difficult in the sense that I think the RNA-Seq p-values generally tend to be on the lower side (which likely helps with getting significant results with an FDR correction). Again, I believe needing filters like fold-change and expression relates to this. **However, relative to other RNA-Seq datasets and the expected differences, I think this looks reasonable.**

That said, it might be worth noting that the gene illustrated in *Figure 1* (SOX11) is the gene with the lowest p-value, and could be identified using an annotation-based strategy (*without* using derfinder).

You could also export pseudocounts from edgeR, instead of visualizing quantified Count-Per-Million (CPM) values. However, you can see a heatmap from log2(CPM+1) values (**above, right**) that **concordance between replicates is good**. There is some extra variability in the outlier SRR1554536. However, infant samples are more similar to adult samples than fetal samples (matching what I expect from the paper), and I think it is clear that adding a covariate was not needed to achieve this.

If I still had too many genes for enrichment, perhaps this means I could take the top *n* genes ranked by p-value (as a subset of the current gene list). However, the next step is more about looking for differential histone binding in liver among this set of genes, rather than traditional gene set enrichment.

Nevertheless, I believe the clear and consistent clustering within groups is reasonably good evidence to use this as a candidate list for the RNA-Seq portion of analysis (if using only this set of samples).