In order to best match expectations, I started with using strategies similar to what I learned from the Bioconductor course. I found the assignment description to be somewhat confusing, so I have divided this report into sections to try and make that more clear in this report.

There is also code for reproducibity on GitHub:
https://github.com/cwarden45/JHU_Coursera_GDS_Capstone/tree/main/Week9

Roadmap Epigenomics Adult vs Fetal Brain H3K4me3 Analysis

At first, I had some difficulty finding the right samples with words like "brain," "adult" and "fetal." I am still not sure if I selected the absolute best samples, but I used this guide for the cell line IDs along with this forum discussion to select the 2 candidates:

Adult Brain: E073 (Brain Dorsolateral Prefrontal Cortex)

Fetal Brain: E081 (Fetal Brain Male)

While I have used broad peak analysis for histones in my own work, I selected the narrow peaks to match what was described in the earlier course.

I started without filtering the peaks (as shown on GitHub), but I wanted to see if I could improve the expected trend. So, I filtered for qValue > 5 (which I think is a little more stringent than FDR < 0.01):

	RNA-Seq Up (Adult Promoters)	**RNA-Seq Down** (Fetal Promoters)
H3K4me3 Adult Brain	8,607 promoter overlaps	8,109 promoter overlaps
H3K4me3 Fetal Brain	1,864 promoter overlaps	4,194 promoter overlaps

In general, the peak counts seem higher in the adult sample than the fetal sample. There is a trend in the direction that we expect, but there are certainly a large number of overlapping peaks for the opposite direction as well. **The difference in distribution between adult and fetal sample is significant with a Fisher's Exact Test p-value < 2.2 x 10^{-16}**. I think this may be inflated due to the large number of peak counts, but peak q-value filtering noticeably decreased the fetal peaks (even more strongly favoring the desired trend).

SOX11 (*Gene ID 6664*) was successfully mapped among the fetal peaks, which is relevant given that that sis the gene with the lowest p-value that was mentioned in Figure 1 of the paper. With or without the qValue filter, SOX11 is in both sets of peaks overlapping the fetal RNA-Seq gene list.

Roadmap Epigenomics Adult Liver H3K4me3 Analysis

If the same changes are observed in different adult vs fetal samples, then that undermines the goal of specifically isolating DLPFC (DorsoLateral PreFrontal Cortex) brain tissue. So, the goal of this section to either see if the currently filtered gene list is unique for the brain (versus liver), and/or see if there might be a benefit to additional filtering.

I don't believe that there are any fetal liver from the Roadmap Epigenomics Project. So, I picked the liver sample with a similar sounding name (beginning with an "E"):

Adult Liver: E066 (Liver)

Figuring out exactly what to compare is difficult. However, if I start with the peaks with the expected trend, I think that is good to check for liver overlap (and perhaps filtering):

	Adult **(RNA-Seq + H3K4me3)**	**Fetal** **(RNA-Seq + H3K4me3)**
Starting Count	8,607 promoter overlaps	4,194 promoter overlaps
Overlapping Liver	6,637 promoter overlaps (77.1% not specific)	3,991 promoter overlaps (95.2% not specific)

The previous fetal gene list looked a little better (in terms of having better overlap in the expected direction). **However, this looks like a large number of non-specific overlapping peaks.**

SOX11 is **not** overlapping the liver peaks (from the fetal gene list). **This might be good,** since we *don't* see a H3K4me3 mark for the more confident liver peaks (and we want something with **relatively unique expression in the brain**).

To be clear, I am not sure if this was really the best sort of validation to do for the dataset. However, with additional analysis, I think I can show that the **filtered** gene list may provide some plausible candidates.

Additional Analysis

To be clear, I am not saying that I can complete additional public analysis that is the best way to filter the gene list, within the schedule of completing this capstone project. Nevertheless, I wanted to pick at least 1 thing to try.

As something that I could do relatively quickly, I ran Enrichr on the filtered gene list set (**611 adult genes** and **69 fetal genes**) Without being a specialist in the area, it seems like there could be some relevant categories:

Selected Fetal Enrichment.

Pathways →BioPlanet 2019: "**axon guidance**" is the top category

Pathways →Elseivier Pathway Collection: "**Glioblastoma, Proneural Subtype**" is the top category (includes DLL3, which I recognize from other work).

Pathways →Reactome 2016: "**Neuronal System**" is the top category

Ontologies →GO Biological Process 2018: "**central nervous system development**" is the top category (including SOX11)

Cell Types →Human Gene Atlas: "**Fetalbrain**" is the top category (including SOX11 and DLL3)

I downloaded tab-delimited text files from Enrichr, and the hyperlinks above are for those same files that I on GitHub. However, please note that the exported tables are sorted by combined score (and I am noting the lowest p-value enrichment in the notes above).

Because I was looking for what happened with SOX11, I checked the fetal gene first (whereas I usually presented the Adult results first in most other situations). Again, with somewhat limited knowledge, I decided to check the top enrichment in those same 5 reference sets for the adult genes:

Selected Adult Enrichment.

Pathways → *BioPlanet 2019*: "**neuronal system**" is the top category

Pathways → *Elseivier Pathway Collection*: "**Proteins Involved in Epilepsy**" is the top category

Pathways → *Reactome 2016*: "**Neuronal System**" is the top category

Ontologies → *GO Biological Process 2018*: "**chemical synaptic transmission**" is the top category

Cell Types → *Human Gene Atlas*: "**Amygdala**" is the top category

For the "*Human Gene Atlas*" enrichment of the filtered Adult gene list, "Wholebrain" was the 2nd category and "PrefrontalCortex" was the 3rd category.

So, as far as I can tell, it seems like this is also a reasonable match if we wanted to see genes that might have specific expression in the brain.

I also did some additional research on publicly available data, with more extended notes on GitHub. However, for reasons of time, I will skip checking additional datasets at the moment.

Nevertheless, I think the Enrichr results provide some evidence that the filtered gene list may be useful for identifying genes with relevant function in the dorsolateral prefrontal cortex of the brain.