Summary

The goal of the study (and Capstone Project) was to look for gene expression differences that vary in the dorsolateral prefrontal cortex region of the brain, hopefully helping better understand neuronal developmental processes.

The steps for RNA-Seq analysis are outlined in detail, starting from raw FASTQ files. The project may be somewhat different than others were expecting, since I downloaded additional Infant samples for visual inspection. In other words, I have downloaded and re-analyzed 6 Adult samples, 6 Fetal samples, and 6 Infant samples (for a total of 18 re-analyzed samples).

While assessment of the results is somewhat difficult, I believe the filtered gene list in Week 9 shows enrichment for neuronal genes (using an analysis strategy based upon known annotations, rather than using *derfinder* as described in the associated paper).

Methods

I have provided a full set of code and small-sized files (excluding raw reads, alignment files, etc.) on GitHub:

https://github.com/cwarden45/JHU_Coursera_GDS_Capstone

However, to briefly describe the methods used for each step of the process:

When available, hyperlinks for each week's report are provided in the name for each week.

Week 4: Alignment

- Raw separate forward (R1) and reverse (R2) reads were downloaded using the "wget" function from the European Nucleotide Archive (ENA) database (for project **PRJNA245228**, https://www.ebi.ac.uk/ena/browser/view/PRJNA245228)
- The **hg19** genome reference was used for alignment, indexed by **STAR** (v2.7.2d) along with a GTF of RefSeq gene annotations (defined from Bioconductor).
- Metadata for samples were defined from the **Biosample** entries (linked in the SRA, as well as the ENA – recorded during the week of 1/25/2021 to 1/30/2021), in addition to the **phenotypeDat_n121_combinedPheno.txt** within the .zip file for Supplemental Data 1 of the paper (which includes the group labels for age bins).
- Aligned read were counting using 3 strategies:
    1. STAR alignment log
    2. samtools (v1.9-168-gb1e2c78) "flagstat", compared to the starting reads to calculate an alignment rate.
    3. GenomicAlignments (v 1.20.1) counts of unique aligned read names using **readGAlignments()**, compared to the starting reads to calculate an alignment rate.
- Plots were generated within R (v3.6.3) using R-base functions.
- **If you re-calculate the alignment rate using GenomicAlignments, SRR1554536 is an outlier with a notably lower alignment rate.**

- *So, this will be considered a possible confounding factor later on*. However, these results were robust with either STAR or TopHat2 as the splice-aware RNA-Seq alignment. This is also a good match for the quantification rate in Week 6.
- **Otherwise, I think these results looked OK, even if a couple other alignment rates seemed a little on the lower side.**

Week 5: Read / Alignment QC

- Alignment analysis from previous week was referenced.
- Statistics were calculated for the starting reads using **FastQC** (v0.11.9)
- The library strand was confirmed to be **unstranded** using **RSeQC** (v4.0.0)
- **TIN** (Transcript Integrity Number) scores were calculated to assess the quality of the aligned sequences using **RSeQC** (v4.0.0), using a set of downloaded housekeeping genes (as defined by the RSeQC developers)
- Plots were generated within R (v3.6.3) using R-base functions.
- **In short, candidate QC metrics to check at later steps were generated** (TIN, unique/duplicated percent, average GC, and percent average Q30)

Week 6: Feature Counts

- I used the **TxDb_hg19_gene.gtf** file for annotations, which is downloadable and originally defined from Bioconductor *TxDb.Hsapiens.UCSC.hg19.knownGene* and *org.Hs.eg.db* annotations (created on **9/22/2019**)
- Parameters "-p -B -C" were added to **featureCounts** (subread v2.0.1) for paired-end data.
- It is the default setting, but I also added "-s 0" to featureCounts (subread v2.0.1) because I checked that the libraries are unstranded.

Week 7: Exploratory Analysis

- Visualize the effect of normalizing raw counts to millions of quantified reads (Count Per Million, or CPM)
- Generate log2-transformed PCA plots with variation rounding factors for the log2-transformation: log2(CPM + 0.1), log2(CPM + 1), or log2(CPM + 10)
- Look for correlations between variable biological variables (such as age) and other variables (principal components, QC stats, alignment stats, etc.).

Week 8: Statistical Analysis / Differential Expression

- A benchmark of differentially expressed genes was compared for edgeR (v3.40.6), DESeq2 (v1.24.0), and limma-voom (3.40.6)
- The **edgeR** gene list was selected, and defined with the following criteria:
- Beyond the requirements, a heatmap was created using the **heatmap.3.R** (https://github.com/obigriffith/biostar-tutorials/blob/master/Heatmaps/heatmap.3.R) script. The gplots (v3.1.0)and RColorBrewer (v 1.1-2) packages were also used to create color schemes / gradients in this plot.
- Unless otherewise specified, plots were generated within R (v3.6.3) using R-base functions.

- Overall, **the replicates showed good concordance**, and the Infant samples were more similar to the Adult samples than the Fetal samples (among these differentially expressed genes). So, for this particular dataset, this looks like a reasonable strategy (without including any covariates).
- The gene in Figure 1 (**SOX11**) was identified with the traditional annotated gene differential expression analysis and this gene had the **lowest edgeR p-value**. To add emphasis for this in the volcano plot, plotrix (v3.8-1) was used to draw a circle around this gene.

Week 9: **Gene Set Enrichment**

- Compare gene list from Week 8 to Roadmap Epigenome H3K4me3 histone binding data for adult and fetal brains.
- As a control, consider how many genes have non-specific expression in liver.
- For both brain and liver H3K4me4 comparsions, use R packages for analysis (with R version 4.0.3)
- RNA-Seq gene symbols were mapped using *org.Hs.eg.db* (v2.22.0)
- Gene promoters were defined using *TxDb.Hsapiens.UCSC.hg19.knownGene* (v3.2.2)
- Roadmap Epigenomics data was downloaded using biomaRt (v2.46.0) and AnnotationHub (v2.22.0)
- In order to gain confidence in the filtered gene list, use Enrichr (https://maayanlab.cloud/Enrichr/, used on 2/23/2021) for gene set enrichment of the filtered gene list.
- I was skeptical about the analysis, and there are a lot of genes that also change in the liver. However, if you filter those genes, then I think the Enrichr enrichment analysis looks encouraging (and includes the SOX11 gene emphasized in the paper).

Limitations / Future Considerations

Setting up all of the dependencies can take some time, and I would guess peer reviewers will likely not re-run all of the analysis to test reproducibly (and possibly not even any of the analysis, for various reasons). In theory, setting up a Docker image could help with this. However, Docker images were not covered in this class. Also, if any steps use almost all of the computational resources for a machine and/or take a long time to run, then this could have a negative impact on running code within a Docker virtual environment.

I think the results are encouraging, but I would not say this is "complete" analysis. For example, I would say considerably more time for projects should be planned.

That said, if I do continue to look into the underlying data, then I will post updates on GitHub. I think post-publication review is something important that should be emphasized more. So, while I hope nothing is so wrong that it would be equivalent to a peer-reviewed paper being retracted, I do think updated comments are a good form of post-publication review that I hope to contribute after the Capstone Project is formally completed.