Adult vs. Fetal Prefrontal Cortex Differential Gene Expression and Epigenetic Roadmap Correlation

M. D'Amour

2/12/2017

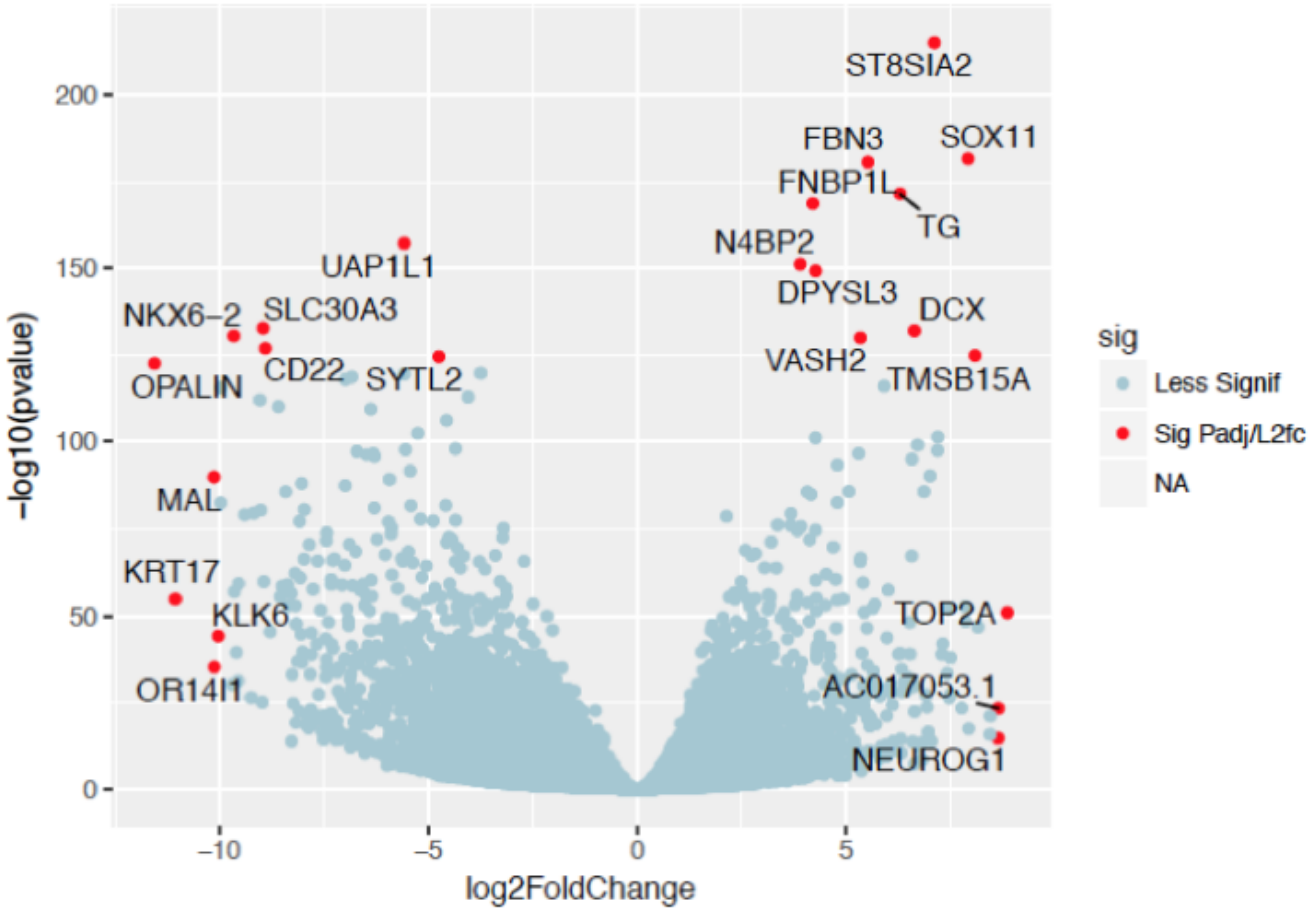


Figure 1: Adult vs Fetal Brain Differentially Expressed Genes

Study Intent

DNA deep sequencing provides an exquisite view of the DNA in the chromosomes of an organism. It provides a read-out of the organism's "code". RNA sequencing, on the other hand, provides a relative count of transcribed RNA in the cell. It gives a snapshot of RNA molecules present in the cell at a particular moment in time. Extending the earlier analogy, RNA-seq provides "debug trace" information that can be used to determine the current functioning of the particular cell. That is, even though each cell has identical DNA, the RNA sample can show which of the genes in this particular cell are being expressed to give its particular character, e.g. fetal brain cell. Comparative analysis of expressed genes between cells from different organs, of different ages, of siblings, of diseased vs. normal highlights pathways to potential repair and health.

This study involves the analysis of RNA-seq data for twelve (12) samples of post-mortem dorso-lateral prefrontal cortex tissue taken from six (6) second-trimester fetuses and six (6) adults of age ~40 years. The primary purpose of the project is to determine whether and which genes are differentially expressed in brain tissue comparing fetus

to adult. This data can then show genes involved in the developing brain of the fetus as compared to the active maintenance of the adult brain. Correlation with Epigenomic Roadmap ChIP-seq data for H3K4me3 in adult brain vs. fetal brain vs. adult liver is done as a data and analysis sanity check.

The secondary purpose of this study is to elaborate a complete pipeline using open-source tools to efficiently generate differential gene expression data for further study of cell function, differentiation, and disease. Here we study changing gene expression by age of organism. This same pipeline may be expanded to study tumor/normal/sibling gene for cancer pathways.

Conclusions

Comparisons of the adult vs. fetal brain cohorts show a large number of differentially expressed genes, as expected. Of the 32,683 named gene/RNA present in the samples, 10,114 or nearly 1/3 showed significant differential expression - both up and down regulated. The fetal brain tissue shows up regulation of over 250 times for brain-development-related genes such as SOX11 and NEUROG1, as well as TOP2A (topoisomerase 2). This data is available for further study. The comparison of expressed genes between the cohorts and H3K4me3 methylation peaks show significant age and organ correlation.

The open-source tool chain used to perform this analysis was very satisfactory. Over the course of the work, performance of alignment between Tophat2 and Hisat2 improved by more than an order of magnitude. Importantly, Hisat2 has a direct interface to the NCBI Sequence Read Archive that uses a fast pipe to download data as needed, speeding up the process a lot. FastQC and MultiQC provided excellent performance and good visualization of data, allowing easy preparation of the data. The R Bioconductor environment provides easy availability and installation of tools. The *Rsubread featureCount* and *DESeq2* R packages made the analyses quite manageable. Work was completed on a contemporary mid-range laptop with few individual runtimes of over an hour for the most compute-intensive tasks.

Reproducibility

This report is executable R-markdown and is fully reproducible. It may be accessed in Github repository mikedamour/CortexDEGenes as Linux command line scripts and R-markdown with all code, data, methods, and program versions available.

Brain Tissue RNA-seq Data

Data generated by Jaffe *et al.*[1] was used for this project. Each sample was sequenced twice on Illumina HiSeq 2000 equipment, post-processed with Illumina software, providing twenty-four (24) libraries of paired-end reads in .fastq format. (1.03TB in .fastq and 328GB in post-aligned .bam.). Data may be accessed at the NCBI SRA repository URL https://www.ncbi.nlm.nih.gov/sra/?term=SRXxxxxxx using the experiment numbers in the table below.

Phenotype data was compiled in Excel from meta data provided by NCBI for each of the SRX samples and SRR files. Reads data was added after alignment for statistical analysis. The data, in the form of the table below, was saved from Excel in tab-delimited text format (.txt).

Sample	Cohort	Experiment	File	Age	RIN	Sex	Race	Replct	rds_map'd	map_%
R3452	fetal	SRX683795	SRR1554537	-0.38	9.6	F	AA	1	129220081	98.8%
R3452	fetal	SRX683795	SRR2071348	-0.38	9.6	F	AA	2	243191549	90.7%
R3462	fetal	SRX683796	SRR1554538	-0.40	6.4	F	AA	1	156236147	98.9%
R3462	fetal	SRX683796	SRR2071349	-0.40	6.4	F	AA	2	462286753	92.9%
R3485	fetal	SRX683799	SRR1554541	-0.38	5.7	M	AA	1	171773460	98.9%
R3485	fetal	SRX683799	SRR2071352	-0.38	5.7	M	AA	2	192826059	91.7%
R4706	fetal	SRX683824	SRR1554566	-0.50	8.3	M	HISP	1	123431458	98.7%
R4706	fetal	SRX683824	SRR2071377	-0.50	8.3	M	HISP	2	134015735	94.0%
R4707	fetal	SRX683825	SRR1554567	-0.40	8.6	M	AA	1	143917249	98.9%
R4707	fetal	SRX683825	SRR2071378	-0.40	8.6	M	AA	2	155804075	93.9%

Sample	Cohort	Experiment	File	Age	RIN	Sex	Race	Replct	rds_map'd	map_%
R4708	fetal	SRX683826	SRR1554568	-0.50	8.0	M	AA	1	113018703	98.9%
R4708	fetal	SRX683826	SRR2071379	-0.50	8.0	M	AA	2	233805731	91.7%
R2869	adult	SRX683793	SRR1554535	41.58	8.7	M	AA	1	96258319	99.1%
R2869	adult	SRX683793	SRR2071346	41.58	8.7	M	AA	2	118061708	81.6%
R3098	adult	SRX683794	SRR1554536	44.17	5.3	F	AA	1	49744953	99.5%
R3098	adult	SRX683794	SRR2071347	44.17	5.3	F	AA	2	77925728	92.5%
R3467	adult	SRX683797	SRR1554539	36.50	9.0	F	AA	1	80649750	99.2%
R3467	adult	SRX683797	SRR2071350	36.50	9.0	F	AA	2	88041049	83.8%
R3969	adult	SRX683814	SRR1554556	36.98	8.5	M	AA	1	113119523	99.2%
R3969	adult	SRX683814	SRR2071367	36.98	8.5	M	AA	2	123291026	90.6%
R4166	adult	SRX683819	SRR1554561	43.88	8.7	M	AA	1	93639705	99.0%
R4166	adult	SRX683819	SRR2071372	43.88	8.7	M	AA	2	106177991	84.6%
R3969	adult	SRX683792	SRR1554534	40.42	8.4	M	AA	1	67363924	98.9%
R3969	adult	SRX683792	SRR2071345	40.42	8.4	M	AA	2	77828716	85.2%

Regarding the data, these samples are from the Jaffe DERfinder "discovery set". That set also includes samples of four (4) other ages ranging from early childhood through old age. Each age is represented equally as the samples above - six (6) samples each with two (2) technical duplicate runs each. In addition, an equally sized and sequenced "verification set" is available. These data sets were used to to explore differentially expressed and highly-conserved RNA outside of genes, developing the DERfinder R package to explore that data.

Hisat2 Sequence Alignment, QC, and Stats at Linux Command Line

Hisat2 was used to align the reads to the *H. sapiens* UCSC GRCh38 human genome (hg38). Hisat2 alignment software, as available in the end of 2016, is demonstrating performance improvements of as much as 50 X over Tophat2, making alignment time on a contemporary laptop quite acceptable. The publishers of Hisat2 make available at https://ccb.jhu.edu/software/hisat2/index.shtml a pre-indexed version of hg38 downloadable by web interface. Hisat2 also provides a direct SRA interface option that obviates the need for separate download. The SRA reads files were accessed directly from Hisat2 using this option.

Hisat2, compression to .bam, FastQC, and stats analysis were run at the command line from the following scripts.

```
GENEDIR=/pathToGenomeData
ARCVDIR=/pathToArchive
LOGDIR=/pathToLog
WRKDIR=/pathToWorkDir

# Create file listing all the SRR numbers for files that are to be analyzed
printf '%s\n' 'SRR1554534' 'SRR1554535' 'SRR1554536' 'SRR1554537' 'SRR1554538' 'SRR1554539' \
'SRR1554541' 'SRR1554556' 'SRR1554561' 'SRR1554566' 'SRR1554567' 'SRR1554568' 'SRR2071345' \
'SRR2071346' 'SRR2071347' 'SRR2071348' 'SRR2071349' 'SRR2071350' 'SRR2071352' 'SRR2071367' \
'SRR2071372' 'SRR2071377' 'SRR2071378' 'SRR2071379' >srrList.txt

# Hisat2 only outputs .sam files, so separately compress to .bam
for i in `cat srrList.txt`; do
    hisat2 -p 4 -t -x $GENEDIR/hg38_ht/genome --sra-acc i -S $ARCVDIR/i.sam >& $LOGDIR/i.log
    samtools view -b -@ 4 -o $WRKDIR/i.bam $ARCVDIR/i.sam
    # Check quality of Hisat2 output .bam files
    fastqc -o $WRKDIR/fastqcOut -t 4 $WRKDIR/i.bam
    samtools stats $WRKDIR/i.bam > $WRKDIR/statsOut/i.stats
done
```

The (example) lines below were extracted from each .stats file for reporting. The "reads mapped" values were manually inserted into the phenotype table for later FPMR calculation for plotting.

```
# SN raw total sequences: 132911310
# SN reads mapped: 106284374
# SN average quality: 27.9
```

Review of QC Data

The QC data output by FastQC above was analyzed using MultiQC. Directing MultiQC to the main working directory, it locates subdirectories containing FastQC output and logs, consolidating them into .html and text output.

```
# Run MultiQC on all QC and log data
multiqc -f  $WRKDIR/readsData
```

After review of the MultiQC .html output file and, since this analysis is for gene expression read counts, not for SNPs, no further trimming was done for base quality, read quality, or contamination.

Counting Reads for Gene Expression

The Gencode release 25 GCRh38.p7 in .gtf format was downloaded from https://www.gencodegenes.org/releases/current.html using the web interface. The featureCount program was run at the command line, as follows.

```
featureCounts -T 4 -t gene -p -a $GENES/hg38genes/gencode.v25.annotation.gtf \
-o $WKDIR/expGenes/aln.pg.fcnt \
$WRKDIR/SRR1554535.bam $WRKDIR/SRR2071346.bam $WRKDIR/SRR1554536.bam \
$WRKDIR/SRR2071347.bam $WRKDIR/SRR1554539.bam $WRKDIR/SRR2071350.bam \
$WRKDIR/SRR1554556.bam $WRKDIR/SRR2071367.bam $WRKDIR/SRR1554561.bam \
$WRKDIR/SRR2071372.bam $WRKDIR/SRR1554534.bam $WRKDIR/SRR2071345.bam \
$WRKDIR/SRR1554537.bam $WRKDIR/SRR2071348.bam $WRKDIR/SRR1554538.bam \
$WRKDIR/SRR2071349.bam $WRKDIR/SRR1554541.bam $WRKDIR/SRR2071352.bam \
$WRKDIR/SRR1554566.bam $WRKDIR/SRR2071377.bam $WRKDIR/SRR1554567.bam \
$WRKDIR/SRR2071378.bam $WRKDIR/SRR1554568.bam $WRKDIR/SRR2071379.bam
```

Exploratory Analysis of Gene Expression

From this point, all analysis was performed in R. Initializing environment.

```
library(Rsubread);  library(GenomicRanges);   library(BiocGenerics)
library(DESeq2);    library(AnnotationDbi);   library(SummarizedExperiment)
library(Biobase);   library(stringr);         library(ChIPpeakAnno)
library(dplyr);     library(RColorBrewer);    library(AnnotationHub)
library(pheatmap);  library(rtracklayer);     library(ggplot2);   library(ggrepel)
```

Data Preparation - RangedSummarizedExperiment Object

Followed RNA-Seq workflow published by Love *et al.*[2]

```
bamDir = c('/Volumes/MacExp/geneData/hisatBam/')
rsltDir = c('/Users/mikedamour/Genomics/gdsCap/gdsCapData/readsData/expGenes/')
geneDir = c('/Users/mikedamour/Genomics/genomes/hg38genes/')

# Import data from command line featureCounts
expGenesCL = read.table (paste0(rsltDir, 'aln.pg.fcnt.txt'),
                         header = TRUE, sep = '\t', skip = 1)

# Clean col names to include only the file id
```

```
colnames(expGenesCL) = c(colnames(expGenesCL[1:6]),
                         str_extract(colnames(expGenesCL[7:30]), "SRR[0-9]+"))
```

Annotate Gene Names

The inbuilt ENSEMBL to SYMBOL annotation does not match the Gencode v25 GTF and, even when forced, produces names for only 20K of 36K dif-exp RNA. Code below was used to extract the gene/RNA names directly from the GTF file, then mapped names back to genes. (Would be appropriate for featureCounts to carry gene names through with counts data.)

```
# Pull gene symbols out of the GTF file used for featureCounts - takes about 3 minutes
# Write into a file first time.  Read from file after.
hg38GeneSymsRaw = read.table(paste0(geneDir, "gencode.v25.annotation.gtf"),
                             fill = TRUE, skip = 5, stringsAsFactors = TRUE)
hg38GeneSyms = select(geneSymsRaw[hg38GeneSymsRaw$V3 == "gene",], V10, V19)
row.names(hg38GeneSyms) = NULL; colnames(hg38GeneSyms) = c("ENSEMBL", "SYMBOL")
write.table(hg38GeneSyms, paste0(geneDir, "hg38GeneSyms.txt"), quote = FALSE)
```

```
# Recover previously written gene symbol file
hg38GeneSyms = read.table(paste0(geneDir, "hg38GeneSyms.txt"), stringsAsFactors = FALSE)
expGenesCL$hg38Sym = hg38GeneSyms$SYMBOL[match(expGenesCL[ ,1], hg38GeneSyms$ENSEMBL)]
expGenesCL = select(expGenesCL, Geneid, hg38Sym, Chr:SRR2071379)
dim(expGenesCL)  # 58037  31  # All the genes and 31 columns
```

```
## [1] 58037    31
```

```
write.table(expGenesCL, paste0(rsltDir, "dlpfcDEGFull.txt"), sep = '\t', quote = FALSE)

# Extract assay information and row name
expData = DataFrame(expGenesCL[,8:31])
row.names(expData) = expGenesCL$Geneid

# Make a GRanges for the row ranges
expGr = makeGRangesFromDataFrame(expGenesCL[,1:6], keep.extra.columns = TRUE)

# Bring in phenotype data
pdata = read.table (paste0(rsltDir, 'phenoData2_1-18-17.txt'), header = TRUE, sep = '\t')
pdata = pdata[1:24,]
pdata$Sample = as.character(pdata$Sample)
pdata$Sample = as.factor(str_extract(pdata$Sample, "R[0-9]+"))
pdata$Replicate = as.factor(pdata$Replicate)
row.names(pdata) = pdata$File

# Construct the Ranged SE
pfcSe = SummarizedExperiment(assays = list(counts = as.matrix(expData)),
                             rowRanges = expGr, colData = pdata)
# Use adult as the reference level for the cohort variable
pfcSe$Cohort = relevel(pfcSe$Cohort, 'adult')  # Already alphabetic, but to be sure
```

Transform of Data for Visual Exploration

```
nrow(pfcSe)  # 58037
```

```
## [1] 58037
```

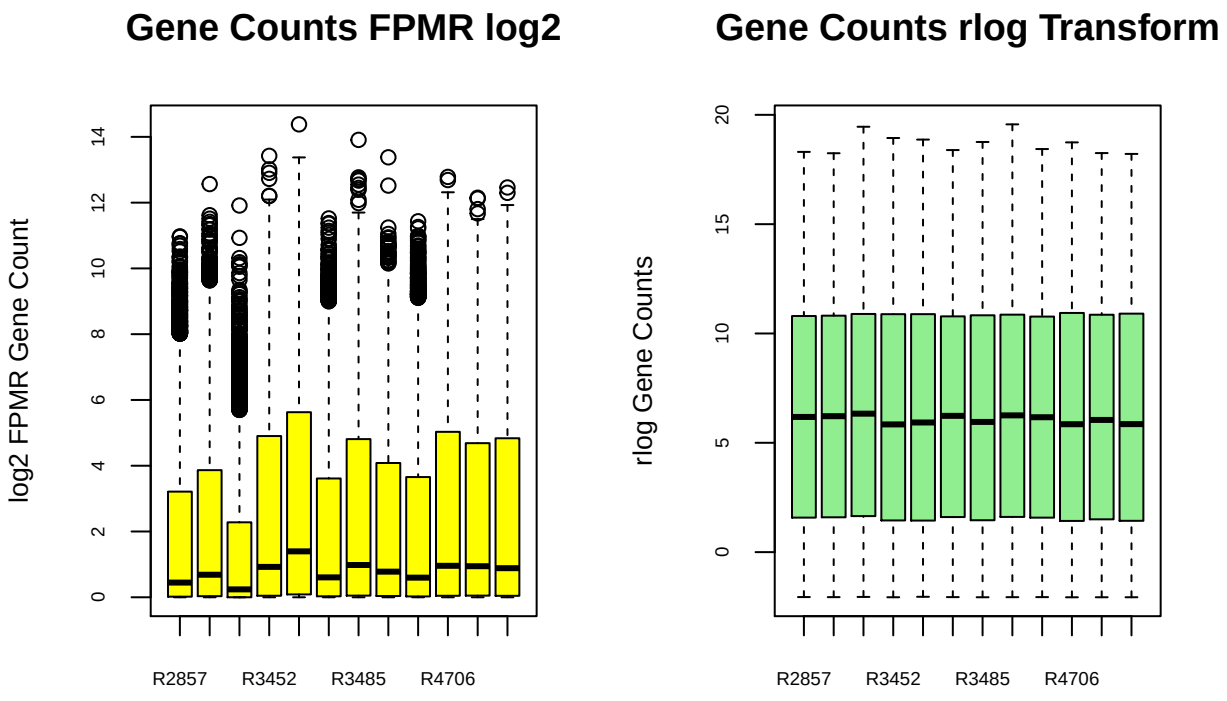
```
pfcSe = pfcSe[rowSums(assay(pfcSe)) > 1,]  # Remove empty rows and chrM
pfcSe = subset(pfcSe, seqnames != 'chrM')
nrow(pfcSe)  # 32683
```

```
## [1] 32683
```

```
# Make a collapsed sample SE with values as fragments per million reads, not FPKM
pfcSeColl = collapseReplicates(pfcSe, pfcSe$Sample, renameCols = TRUE)
# Make an fpmr assay for plotting
fpmrAssay = log2(assay(pfcSeColl) / (pfcSeColl$Reads_Mapped/1e6) + 1)

# Cannot convert se to dds due to bug? Ensure pfcSE built above with matrix not DataFrame
pfcDds = DESeqDataSet(pfcSe, ~ RIN + Sex + Replicate + Race + Dup_Lev + Contam + Cohort)
# Assay must be integers, so must wait to collapseReplicates and any FPMR stuff
pfcDdsColl = DESeqDataSet(pfcSeColl, ~ RIN + Sex + Race + Dup_Lev + Contam + Cohort)
pfcDdsColl$Replicate = droplevels(pfcDdsColl$Replicate)  # After subset columns

# Reg log transform
pfcSeRL = rlog(pfcDds, blind = FALSE)  # Keep replicates separate for PCA
pfcSeCollRL = rlog(pfcDdsColl, blind = FALSE)  # Replicates collapsed for boxplot
```

Plot Sample Data to Explore Transform

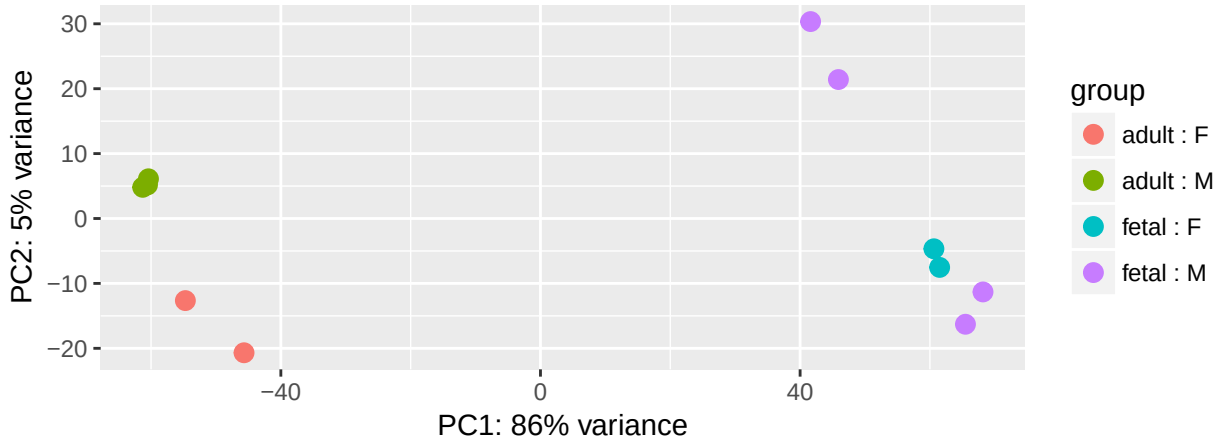
```
par(mfrow = c(1,2))
boxplot(fpmrAssay, col = "yellow", main = "Gene Counts FPMR log2",
        ylab = "log2 FPMR Gene Count", cex.axis = 0.6, cex.lab = 0.8)
boxplot(assay(pfcSeCollRL), col = "lightgreen", main = "Gene Counts rlog Transform",
        ylab = "rlog Gene Counts",cex.axis = 0.6, cex.lab = 0.8)
```



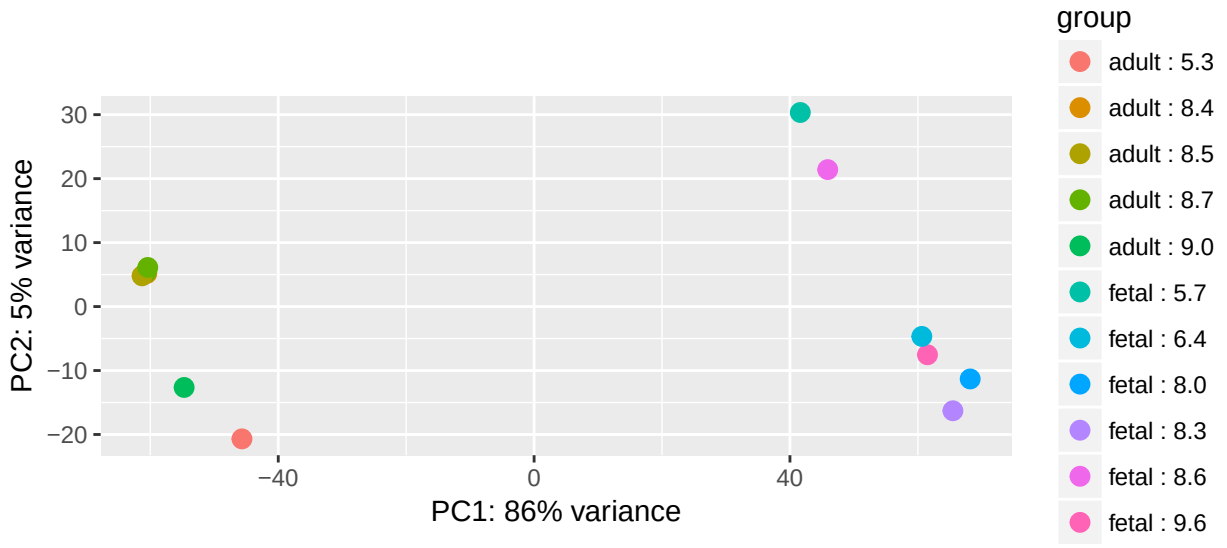
Primary Component Analysis

The first plot highlights Cohort vs. Sex and in fact shows that these variables represent the primary and secondary component of variance. Jaffe mentioned in his lecture that RIN was an important component of variance. PCA plot does not bear this out. Suspicion of artificats in the technical replicates was plotted and proved to be unfounded. Note that one fetal sample was assigned sex of male when all data seem to indicate - including the chrY expression data - that the sample is female. See the bottom right corner of the first PCA plot.

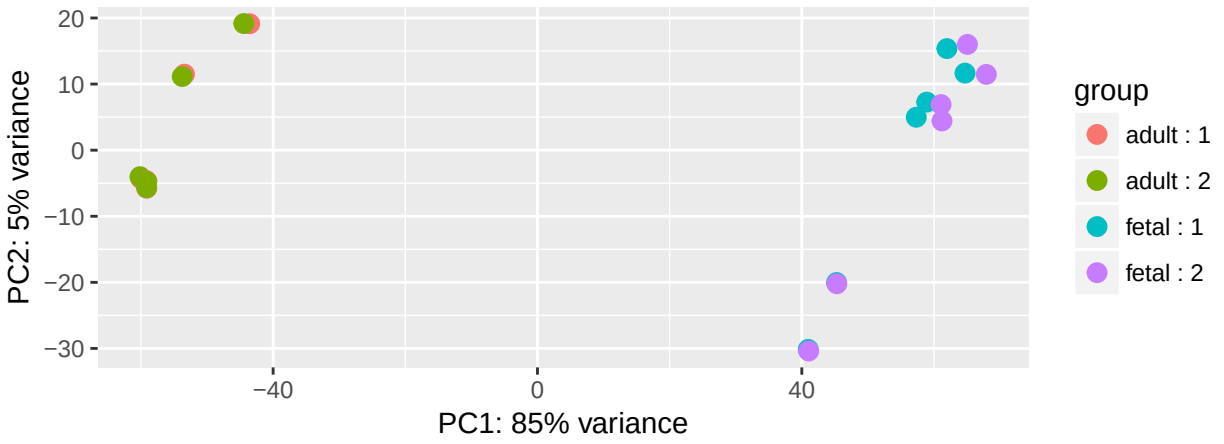
```
plotPCA(pfcSeCollRL, intgroup = c("Cohort", "Sex"))
```



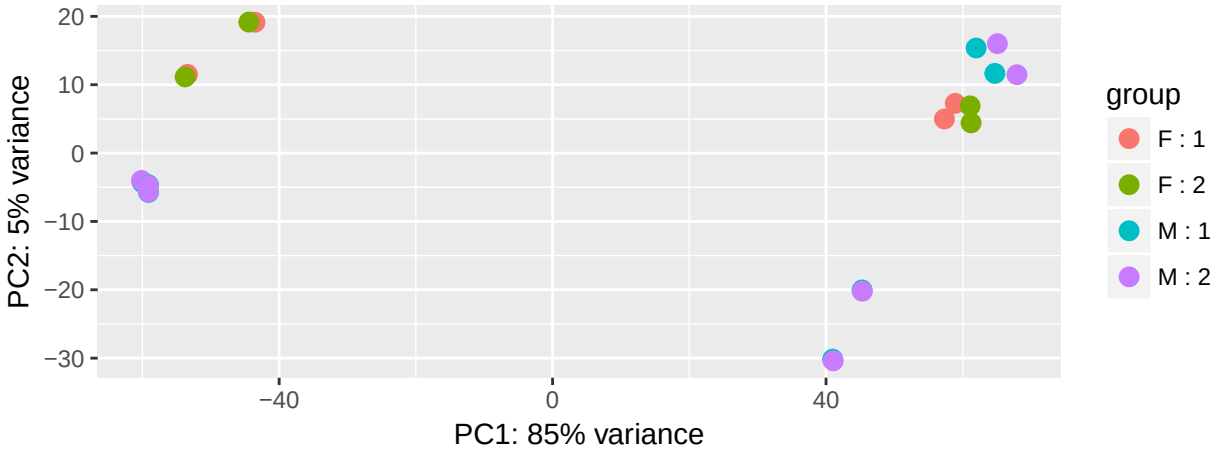

```
plotPCA(pfcSeCollRL, intgroup = c("Cohort", "RIN"))
```




```
plotPCA(pfcSeRL, intgroup = c("Cohort", "Replicate"))
```




```
plotPCA(pfcSeRL, intgroup = c("Sex", "Replicate"))
```



Statistical Analysis of DLPFC Genes Data Set - Adult vs Fetal

The data at this point is in the DESeqDataSet format. Although DESeq2 *rlog* normaization was used previously for plotting, the raw data, filtered only of "0" count and mitochondrial rows, is provided to DESeq2 as specified by the tool authors[1].

Null hypothesis

```
There is no differential gene expression between the adult (base) cohort vs. the fetal
cohort, controlling for the Sex and RIN variables.
```

Alternative hypothesis

```
Genes are up regulated or down regulated based on age difference between sampled
individuals with a false discovery rate (FDR) of 1% using the BH adjustment.
```

Methods

1. The preprocessed sample data including raw (filtered for 0 and chrM) counts data and phenotype table is formatted into a DESeqDataSet with the formula ~ RIN + Sex + Cohort. This means that we want to test for the effect of Cohort membership (age) controlling for RIN quality and Sex.

2. The *DESeq* function of the DESeq2 package processes the data using normalization, GLM and BH adjustment on a per gene basis. Output is a DESeqDataSet object including the analysis data.

3. The *results* function of the DESeq2 package extracts data from the returned DESeq object to build the appropriate experiment contrasts.

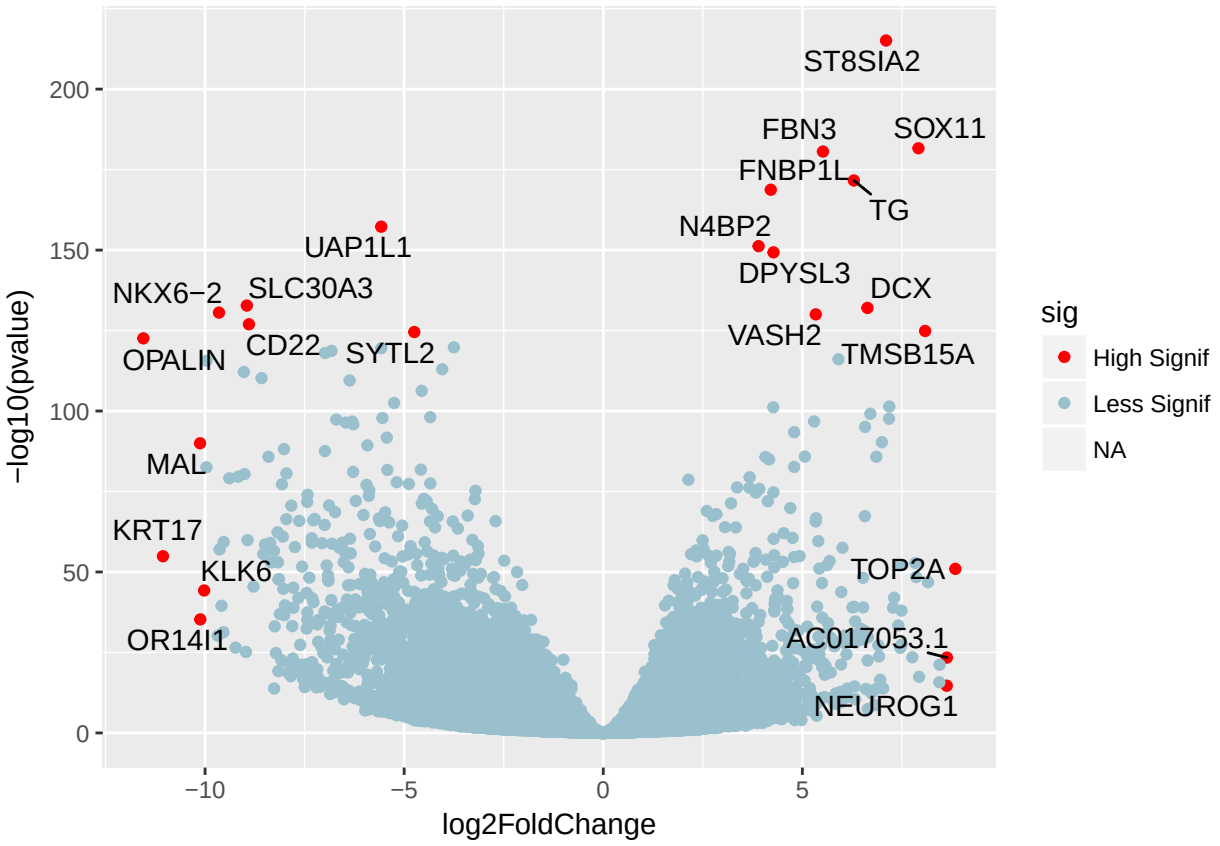
```
# Simplify formula and run analysis
ddsRS = DESeqDataSet(pfcSeColl, ~ RIN + Sex + Cohort)
ddsRS = DESeq(ddsRS)

# Extract results and annotate with gene symbols
resRS = results(ddsRS, alpha = 0.01, contrast = c("Cohort", "fetal", "adult"))
resRS$hg38Sym = hg38GeneSyms$SYMBOL[match(row.names(resRS), hg38GeneSyms$ENSEMBL)]
```

Volcano Plot Differential Gene Expression Fetal vs Adult(ref) Brain

After Turner[3].

```
# Add results data to DESeq object for tidy
mcols(ddsRS) = as(mutate(as.data.frame(mcols(ddsRS)), FvAl2fc = resRS$log2FoldChange,
                         FvApvalue = resRS$pvalue,FvApadj = resRS$padj), "DataFrame")
resRSdf = mutate(as.data.frame(resRS),
                 sig = ifelse((resRS$padj<10e-120) |
                                (resRS$log2FoldChange > 8.5) |
                                (resRS$log2FoldChange < -10),
                              "High Signif", "Less Signif"))
p = ggplot(resRSdf, aes(log2FoldChange, -log10(pvalue))) +
  geom_point(aes(col=sig)) +
  scale_color_manual(values=c("red", "lightblue3")) +
  geom_text_repel(data=filter(resRSdf, sig == "High Signif"), aes(label=hg38Sym))
p
```

```
## Warning: Removed 5070 rows containing missing values (geom_point).
```



Most Highly Differentially Expressed Genes

The top 25/25 differentially expressed genes are displayed below. The entire list of expressed gense are in to file *dlpfcDEGSig.txt*.

```
print('Number of expressed RNA with sum of row counts > 1, of 58,037 potential')
dim(resRS)[1]  # 32683
resRS = resRS[order(-resRS$log2FoldChange), ]
write.table(resRS, paste0(rsltDir, "dlpfcDEGFull.txt"), sep = '\t', quote = FALSE)

# Subset DEGs at 1% FDR and 1.0 l2fc, sort up to down regulation, and store
resRSSig = subset(resRS, padj < 0.01)
resRSSig = subset(resRSSig, abs(log2FoldChange) > 1.0)
print('Number of significantly expressed genes subset to 0.01 FDR and 1.0 log2 fold change.')
dim(resRSSig)[1]  # 10,114
resRSSig = resRSSig[order(-resRSSig$log2FoldChange), ]
resRSPrnt = select(as.data.frame(resRSSig), hg38Sym, log2FoldChange, padj)
write.table(resRSSig, paste0(rsltDir, "dlpfcDEGSig.txt"), sep = '\t', quote = FALSE)
```

```
## [1] "Number of expressed RNA with sum of row counts > 1, of 58,037 potential"
## [1] 32683
## [1] "Number of significantly expressed genes subset to 0.01 FDR and 1.0 log2 fold change."
## [1] 10114
```

Top Differentially Expressed Genes - Adult over Fetal

```
head(resRSPrnt, 25)
```

```
##                       hg38Sym log2FoldChange          padj
## ENSG00000131747.14      TOP2A       8.841992  1.469919e-49
## ENSG00000234275.1  AC017053.1       8.630268  9.677531e-23
## ENSG00000181965.5     NEUROG1       8.625824  2.437800e-14
## ENSG00000232597.5  AC013727.1       8.446001  1.372294e-20
## ENSG00000159217.9     IGF2BP1       8.439728  2.289812e-15
## ENSG00000148773.13      MKI67       8.156270  1.615671e-45
## ENSG00000158164.6     TMSB15A       8.082819 2.697042e-122
## ENSG00000178999.12      AURKB       7.931203  5.345294e-17
## ENSG00000176887.6       SOX11       7.915969 3.215214e-178
## ENSG00000180269.7      GPR139       7.856278  4.635682e-47
## ENSG00000242540.2  AC010729.1       7.840274  2.526689e-51
## ENSG00000109674.3       NEIL3       7.761886  7.873926e-23
## ENSG00000123307.3     NEUROD4       7.490886  1.187699e-26
## ENSG00000112984.11     KIF20A       7.489532  7.116027e-37
## ENSG00000109193.10    SULT1E1       7.459384  1.083850e-25
## ENSG00000224747.1    MTCYBP21       7.423688  1.316174e-50
## ENSG00000175063.16      UBE2C       7.409758  2.028511e-32
## ENSG00000171848.13       RRM2       7.304505  8.882820e-41
## ENSG00000168078.9         PBK       7.278611  8.536721e-38
## ENSG00000254726.2       MEX3A       7.183619  3.725119e-99
## ENSG00000085552.16      IGSF9       7.177192  1.970326e-95
## ENSG00000140557.11    ST8SIA2       7.102355 2.153628e-211
## ENSG00000126787.12     DLGAP5       7.028951  1.550094e-13
## ENSG00000178403.3     NEUROG2       6.997468  3.047848e-88
## ENSG00000163508.12      EOMES       6.955922  4.298022e-16
```

Top Differentially Expressed Genes - Fetal over Adult (ascending)

```
tail(resRSPrnt, 25)
```

```
##                       hg38Sym log2FoldChange          padj
## ENSG00000117152.13        RGS4      -8.539851  4.514555e-54
## ENSG00000138100.13      TRIM54      -8.580865 6.669599e-108
## ENSG00000206195.10      DUXAP8      -8.782298  3.547264e-44
## ENSG00000012124.15        CD22      -8.898261 2.376694e-124
## ENSG00000174403.15 C20orf166-AS1      -8.933835  2.842479e-58
## ENSG00000115194.10      SLC30A3      -8.951123 5.210725e-130
## ENSG00000233123.1     LINC01007      -8.972681  1.925066e-24
## ENSG00000260328.1   RP11-416I2.1      -9.009978  1.790199e-78
## ENSG00000197971.14          MBP      -9.026870 8.524793e-110
## ENSG00000165643.10      SOHLH1      -9.170984  1.186308e-77
## ENSG00000261710.1  RP11-953B20.1      -9.232365  9.449684e-26
## ENSG00000136541.14         ERMN      -9.389305  3.453096e-77
## ENSG00000256193.5      LINC00507      -9.530709  1.043446e-57
## ENSG00000047936.10         ROS1      -9.537419  2.224502e-30
## ENSG00000117594.9       HSD11B1      -9.585054  2.313748e-38
## ENSG00000268038.1     AC011516.2      -9.635432  1.847215e-55
## ENSG00000148826.8        NKX6-2      -9.650306 6.704218e-128
## ENSG00000265015.1   RP11-454P7.3      -9.681137  2.504866e-29
## ENSG00000105695.14          MAG      -9.960438 2.821375e-113
## ENSG00000135426.15       TESPA1      -9.963298  1.488158e-80
## ENSG00000167755.13         KLK6     -10.023017  5.499527e-43
## ENSG00000189181.4         OR14I1     -10.118783  3.143183e-34
## ENSG00000172005.10          MAL     -10.125209  6.481707e-88
```

```
## ENSG00000128422.15         KRT17     -11.057066  2.147082e-53
## ENSG00000197430.10        OPALIN     -11.549359 4.508466e-120
```

Epigenome Roadmap Data

The H3K4me3 peak data was identified at the Epigenome Roadmap website with URL given below. Files were loaded through Bioconductor AnnotationHub already in R's GRanges format.

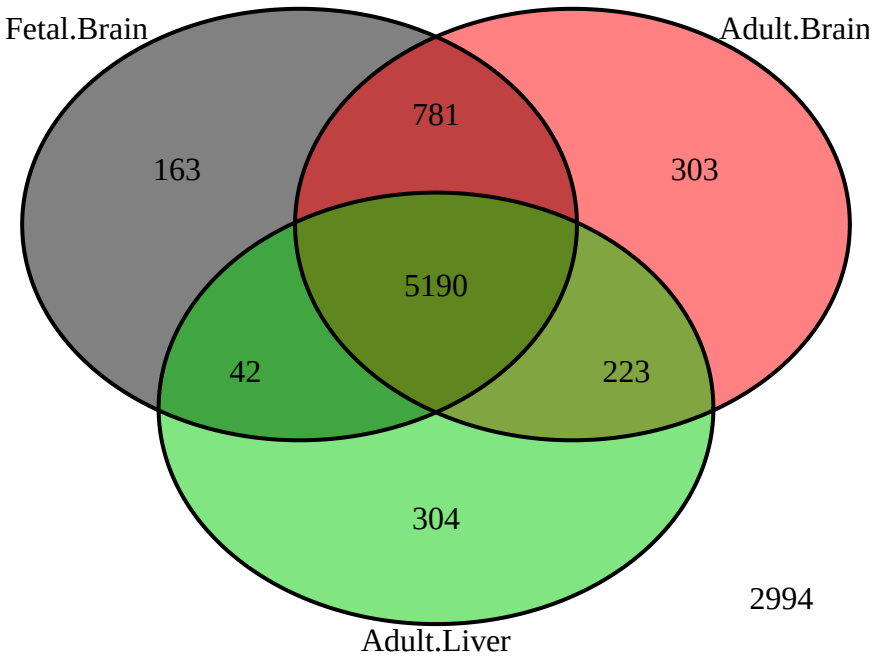
All gene expression data was aligned to the hg38 genome but the Epigenome Roadmap data was aligned to hg19. This required doing liftOver from hg19 to hg38 to proceed.

```
# Roadmap data references at http://egg2.wustl.edu/roadmap/web_portal/meta.html
# Go to Annotation hub for data - all hg19!  Load liftOver chain file.
loChain = import.chain(paste0(geneDir, "hg19ToHg38.over.chain"))
ahub = AnnotationHub()
ahub = subset(ahub, species == "Homo sapiens")

# Get roadmap data and liftover to hg38 from hg19, start with fetal brain data
qh = query(ahub, c('h3k4me3', 'Homo sapiens', 'narrowPeak', 'brain', 'E082'))
fb19 = qh[['AH30479']]  # Female fetal E082
fb = liftOver(fb19, loChain); fb = unlist(fb)

# Adult brain data
qh = AnnotationHub::query(ahub, c('h3k4me3', 'Homo sapiens', 'narrowPeak', 'brain', 'E073'))
ab19 = qh[['AH30413']]  # Male mixed adult E073
ab = liftOver(ab19, loChain); ab = unlist(ab)

# Adult liver data
qh = query(ahub, c('h3k4me3', 'Homo sapiens', 'narrowPeak', 'liver', 'E066'))
al19 = qh[['AH30367']]  # Adult liver E066
al = liftOver(al19, loChain); al = unlist(al)
```

Analysis of Brain Gene Expression Overlaps with H3K4me3 Marks

The brain tissue differential expression data is checked for overlap with fetal brain, adult brain and liver H3K4me3 peaks. Only differentially-expressed genes above FDR 0.01 and log2 fold change > 1 were considered.

```
# Compare to highly expressed genes
fdr = 0.01; l2fc = 1
ddsRSSig = subset(ddsRS, FvApadj < fdr & abs(FvAl2fc) > l2fc)
fbOlap = subsetByOverlaps(promoters(rowRanges(ddsRSSig)), fb)
abOlap = subsetByOverlaps(promoters(rowRanges(ddsRSSig)), ab)
alOlap = subsetByOverlaps(promoters(rowRanges(ddsRSSig)), al)

# Plot Venn diagram of fetal brain vs adult brain vs adult liver
par(mfrow=c(1,1))
makeVennDiagram(list(fbOlap, abOlap, alOlap), ignore.strand = T,
                NameOfPeaks = c('Fetal Brain', 'Adult Brain', 'Adult Liver'),
                totalTest=10000, scaled=F, euler.d=F, fill=c(1,2,3))
```




```
## $p.value
##      Fetal.Brain Adult.Brain Adult.Liver pval
## [1,]           0           1           1    0
## [2,]           1           0           1    0
## [3,]           1           1           0    0
##
## $vennCounts
##      Fetal.Brain Adult.Brain Adult.Liver Counts
## [1,]           0           0           0   2994
## [2,]           0           0           1    304
## [3,]           0           1           0    303
## [4,]           0           1           1    223
## [5,]           1           0           0    163
## [6,]           1           0           1     42
## [7,]           1           1           0    781
## [8,]           1           1           1   5190
## attr(,"class")
## [1] "VennCounts"
```

Code related to Hansen[4].

```
# Check for H3K4me3 mark enrichment in fetal brain vs. adult brain vs. adult liver
methylation = data.frame(MethOlaps = c("FetalBrainOlaps", "AdultBrainOlaps", "AdultLiverOlaps"),
                         OddsRatio = c(0,0,0))
refPeak = c('fb', 'ab', 'al')
prom = reduce(promoters(rowRanges(ddsRSSig), ignore.strand = TRUE))
for(i in seq_along(refPeak))  {
  peaks = reduce(get(refPeak[i]))
  both <- GenomicRanges::intersect(prom, peaks, ignore.strand = TRUE)
  only.prom <- BiocGenerics::setdiff(prom, both)
  only.peaks <- BiocGenerics::setdiff(peaks, both)
  overlapMat <- matrix(0, ncol = 2, nrow = 2)
  colnames(overlapMat) <- c("in.peaks", "out.peaks")
  rownames(overlapMat) <- c("in.promoters", "out.promoter")
  overlapMat[1,1] <- sum(width(both))
  overlapMat[1,2] <- sum(width(only.prom))
```

```
  overlapMat[2,1] <- sum(width(only.peaks))
  overlapMat[2,2] <- 1.5*(10^9) - sum(overlapMat)
  round(overlapMat / 10^6, 2)

  print(as.character(methylation$MethOlaps[i]))
  print(round(overlapMat/ 10^6, 1))
  print("----------------------------")
  oddsRatio <- overlapMat[1,1] * overlapMat[2,2] / (overlapMat[2,1] * overlapMat[1,2])
  methylation$OddsRatio[i] <-oddsRatio
}
```

```
## [1] "FetalBrainOlaps"
##              in.peaks out.peaks
## in.promoters      6.4      22.2
## out.promoter     44.9    1426.4
## [1] "----------------------------"
## [1] "AdultBrainOlaps"
##              in.peaks out.peaks
## in.promoters      6.2      22.2
## out.promoter     47.7    1423.9
## [1] "----------------------------"
## [1] "AdultLiverOlaps"
##              in.peaks out.peaks
## in.promoters      5.1      22.2
## out.promoter     48.7    1423.9
## [1] "----------------------------"
```

```
# Show odds ratios for fetal brain, adult brain, and adult liver overlaps with DE genes
print(methylation)
```

```
##         MethOlaps OddsRatio
## 1 FetalBrainOlaps  9.218440
## 2 AdultBrainOlaps  8.349017
## 3 AdultLiverOlaps  6.745786
```

Promoter Comparison - Fetal Brain, Adult Brain, Adult Liver; Verification

The table above shows the odds ratio of overlaps with highly expressed genes. There are significant differences in H3K4me3 marks between fetal and adult brain as seen with the Venn diagram and contingency table. Fetal brain shows 205 highly expressed gene promoters marked H3K4me3 separate from adult brain, with adult brain having 526 such separately marked genes.

There are significantly fewer methylation peaks from the liver ChIP-seq data that overlap with either the fetal brain or adult brain differentially expressed genes, as shown by the tables and diagram above.

References

1. Jaffe, A.E., Shin, J., Collado-Torres, L., Leek, J.T., Tao, R., Li, C., Gao, Y., Jia, Y., Maher, B.J., Hydel, T.M., Kleinman, J.E., Weinberger, D.R. Developmental regulation of human cortex transcription and its clinical relevance at single base resolution. *Nature Neuroscience* **18**, 154-161 (2015)
2. Love, M., Anders, S., Huber, W., Morgan M. RNA-Seq workflow: gene-level exploratory analysis and differential expression. *Bioconductor* http://www.bioconductor.org/help/workflows/rnaseqGene/ (2017)
3. Turner, S. *Getting Genetics Done*, http://www.gettinggeneticsdone.com/2014.
4. Hansen, K.D. Usecase - Basic GRanges and AnnotationHub. http://kasperdanielhansen.github.io/genbioconductor/html/Usecase_AnnotationHub_GRanges.html

```
sessionInfo()
```

```
## R version 3.3.2 (2016-10-31)
## Platform: x86_64-apple-darwin13.4.0 (64-bit)
## Running under: macOS Sierra 10.12.5
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
##  [1] grid      parallel  stats4    stats     graphics  grDevices utils
##  [8] datasets  methods   base
##
## other attached packages:
##  [1] ggrepel_0.6.5              ggplot2_2.2.1
##  [3] rtracklayer_1.34.1         pheatmap_1.0.8
##  [5] AnnotationHub_2.6.4        RColorBrewer_1.1-2
##  [7] dplyr_0.5.0                ChIPpeakAnno_3.8.9
##  [9] VennDiagram_1.6.17         futile.logger_1.4.3
## [11] Biostrings_2.42.1          XVector_0.14.0
## [13] stringr_1.1.0              AnnotationDbi_1.36.2
## [15] DESeq2_1.14.1              SummarizedExperiment_1.4.0
## [17] Biobase_2.34.0             GenomicRanges_1.26.2
## [19] GenomeInfoDb_1.10.3        IRanges_2.8.1
## [21] S4Vectors_0.12.1           BiocGenerics_0.20.0
## [23] Rsubread_1.24.1
##
## loaded via a namespace (and not attached):
##  [1] bitops_1.0-6                  matrixStats_0.51.0
##  [3] httr_1.2.1                    rprojroot_1.2
##  [5] tools_3.3.2                   backports_1.0.5
##  [7] R6_2.2.0                      rpart_4.1-10
##  [9] Hmisc_4.0-2                   DBI_0.5-1
## [11] lazyeval_0.2.0                colorspace_1.3-2
## [13] ade4_1.7-5                    nnet_7.3-12
## [15] gridExtra_2.2.1               graph_1.52.0
## [17] htmlTable_1.9                 labeling_0.3
## [19] scales_0.4.1                  checkmate_1.8.2
## [21] genefilter_1.56.0             RBGL_1.50.0
## [23] digest_0.6.12                 Rsamtools_1.26.1
## [25] foreign_0.8-67                rmarkdown_1.3
## [27] base64enc_0.1-3               htmltools_0.3.5
## [29] ensembldb_1.6.2               limma_3.30.11
## [31] BSgenome_1.42.0               regioneR_1.6.2
## [33] htmlwidgets_0.8               RSQLite_1.1-2
## [35] BiocInstaller_1.24.0          shiny_1.0.0
## [37] BiocParallel_1.8.1            acepack_1.4.1
## [39] RCurl_1.95-4.8                magrittr_1.5
## [41] GO.db_3.4.0                   Formula_1.2-1
## [43] Matrix_1.2-8                  Rcpp_0.12.9
## [45] munsell_0.4.3                 stringi_1.1.2
## [47] yaml_2.1.14                   MASS_7.3-45
## [49] zlibbioc_1.20.0               plyr_1.8.4
## [51] lattice_0.20-34               splines_3.3.2
## [53] multtest_2.30.0               GenomicFeatures_1.26.2
## [55] annotate_1.52.1               locfit_1.5-9.1
```

```
## [57] knitr_1.15.1                 seqinr_3.3-3
## [59] geneplotter_1.52.0           biomaRt_2.30.0
## [61] futile.options_1.0.0         XML_3.98-1.5
## [63] evaluate_0.10                latticeExtra_0.6-28
## [65] lambda.r_1.1.9               data.table_1.10.4
## [67] idr_1.2                      httpuv_1.3.3
## [69] gtable_0.2.0                 assertthat_0.1
## [71] mime_0.5                     xtable_1.8-2
## [73] survival_2.40-1              tibble_1.2
## [75] GenomicAlignments_1.10.0     memoise_1.0.0
## [77] cluster_2.0.5                interactiveDisplayBase_1.12.0
```