Capstone Final Report

Janet Barroso

8/20/2020

Capstone Final Report

The main question to answer is *What is the difference between gene expression in fetal and adult brains?*

The data used was deposited into the Short Read Archive (SRA) as BioProject PRJNA245228, provided by *Jaffe et. al* and published in the paper:link

6 samples were used in this analysis; 3 fetal samples: *SRR1554537*, *SRR1554567*, *SRR1554538* and 3 adult samples: *SRR1554534*, *SRR1554535*, *SRR1554536*

Phenotype Data was collected accesing to the SRA website. Search for the experiment and in *"Related Information"* click *"BioSample"* and collect information as: AGE, sex, tissue, disease, race, RIN, Fraction

For accessing the data and perform the alignment I use the **Galaxy platform (Galaxy tool version 2.1.0+galaxy5)**. Since I am located in Germany, I used the Europe Galaxy platform which slighted differs from the USA platform. How ever, for the basic tools remain the same.

To access the data I used **"Faster Download and Extract Reads in FASTQ format from NCBI SRA"** I download 6 samples corresponding to 3 adults(*SRR1554534*, *SRR1554535*, *SRR1554536*) and 3 fetal(*SRR1554537*, *SRR1554567*, *SRR1554538*). Samples were chosen based on sex and RIN parameter. I choose 3 females and 3 males and a RIN of (8.4,8.7,5.3) for adults and (9.6,8.6,6.4) for fetals

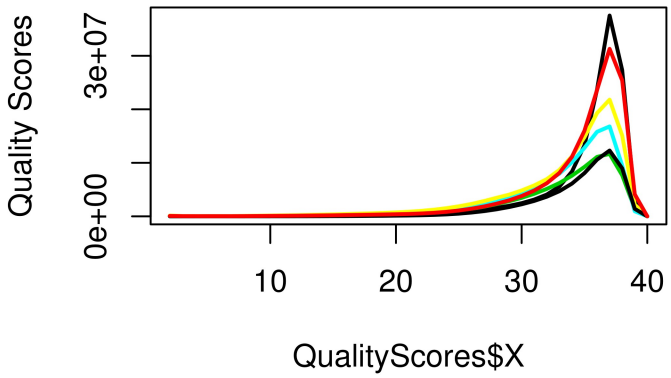
```
phenoData<-read.delim("C:/Users/janet/Documents/R/Cursera/Genomic Data/Capstone/phenoData.txt", header =
phenoData
```

Phenotypic information

```
##      EXP_ID   sapmleID ISOLATE    AGE GROUP provider    sex tissue disease race
## 1 SRX683792 SRR1554534   DLPFC 40.420 Adult     LIBD   male  DLPFC Control   AA
## 2 SRX683793 SRR1554535   R2869 41.580 Adult     LIBD   male  DLPFC Control   AA
## 3 SRX683794 SRR1554536   R3098 44.170 Adult     LIBD female  DLPFC Control   AA
## 4 SRX683795 SRR1554537   R3452 -0.384 Fetal     LIBD female  DLPFC Control   AA
## 5 SRX683825 SRR1554567   R4707 -0.400 Fetal     LIBD   male  DLPFC Control   AA
## 6 SRX683796 SRR1554538   R3462 -0.400 Fetal     LIBD female  DLPFC Control   AA
##   RIN Fraction  TotalSeq Alignment_Rate QualityScore GC_content
## 1 8.4    Total 138944420          99.73        35.54         51
## 2 8.7    Total  81919618          99.76        33.54         47
## 3 5.3    Total 104753495          99.90        33.96         46
## 4 9.6    Total 134583162          99.80        33.95         48
## 5 8.6    Total  67711557          94.92        34.72         46
## 6 6.4    Total 149267911          99.48        34.91         47
```

Alignment For the Alignment I used **HISAT2** since I needed something robust that used fewer computational resources. Alignment of the reads was done to the version hg19 of the human genome.

Quality control Quality control on the alignments was done using **FastQC**Read Quality reports (Galaxy Version 0.72+galaxy1)



One thing to notice is that all of my alignments failed at the "Per base sequence content" in which the showed a high variation at the beginning like the following plot (I will show just one as an example).

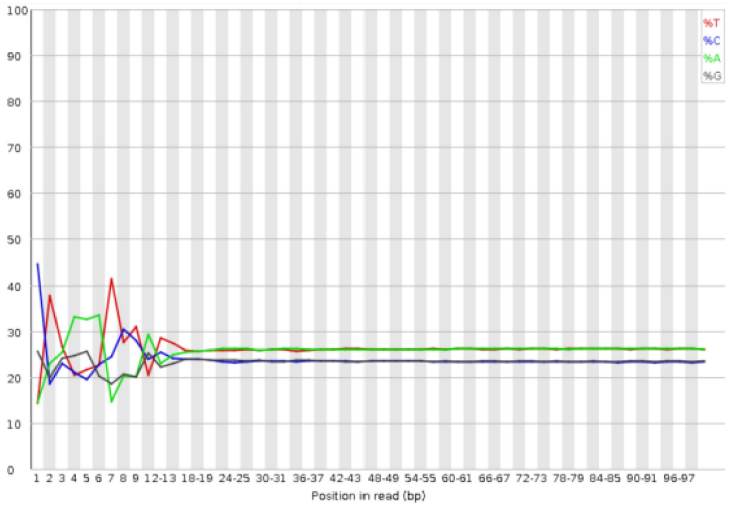


Figure 1: Per base sequence content

However, as discussed in link this bias could be due to the pseudo random primers which are used in the library generation. It was also notices that this does **NOT** affect results

Expression measurements at gene level To Measure gene expression from the BAM files ***feature counts***(Galaxy Version 1.6.4+galaxy2) was used and then all"gtf" files (featureCounts: Counts with locations) were used in R to perform the rest of the analysis.

Complete Galaxy workflow is imported for clarification

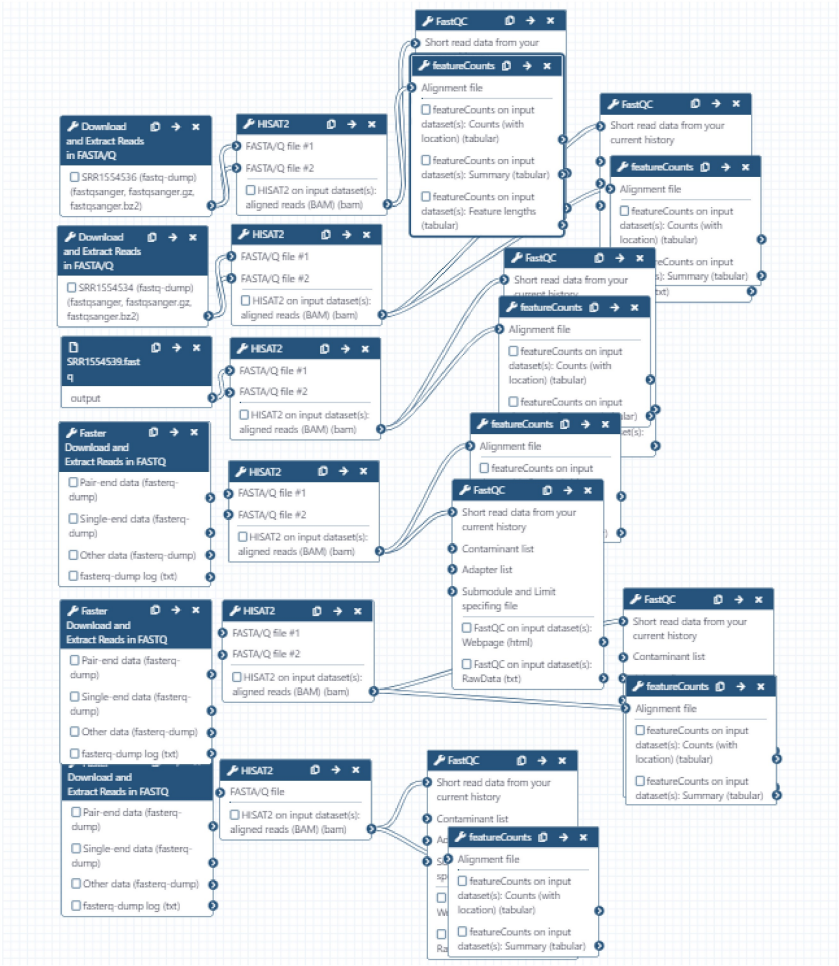


Figure 2: Worflow Galaxy

```
featureCounts_AllSamples <- read.delim("~/R/Cursera/Genomic Data/Capstone/featureCounts_AllSamples.txt"
```

Exploratory Data Analysis In order to create a complete summarized experiment, BAM files should be imported

```
SRR1554534 <- read.delim("~/R/Cursera/Genomic Data/Capstone/SRR1554534.tabular")
SRR1554535 <- read.delim("~/R/Cursera/Genomic Data/Capstone/SRR1554535.tabular")
SRR1554536 <- read.delim("~/R/Cursera/Genomic Data/Capstone/SRR1554536.tabular")
SRR1554537 <- read.delim("~/R/Cursera/Genomic Data/Capstone/SRR1554537.tabular")
SRR1554567 <- read.delim("~/R/Cursera/Genomic Data/Capstone/SRR1554567.tabular")
SRR1554538 <- read.delim("~/R/Cursera/Genomic Data/Capstone/SRR1554538.tabular")

FullExperiment<- cbind(SRR1554534, SRR1554535$HISAT2.on.data.46..aligned.reads..BAM., SRR1554536$HISAT2
```

There is 2 correction that need to be done:

1. Exp_list$Chr contains several times the chromosome number according to the transcripts of the gene, here I will extract only one chromosome name since all of share the same chromosome

```
library(dplyr)
library(tidyr)
library(purrr)
colnames(FullExperiment)[7:12]<-c("SRR1554534","SRR1554535","SRR1554536","SRR1554537","SRR1554567","SRR
FullExperiment$Chr<-as.character(FullExperiment$Chr)
string <- as.character(FullExperiment$Chr)
string <- as.data.frame(string)
FullExperiment$Chr <- string[1] %>% extract(string[1], ";")
```

2. Strand factor is also repeated as many times as transcripts in the gene, and here I will extract the strand

```
dna_strand = matrix(, nrow= dim(FullExperiment)[1], ncol=1)
for (i in 1:length(FullExperiment$Strand)){
      if (grepl("-", FullExperiment$Strand[i])){
            dna_strand[i] = "-"
      } else{
            dna_strand[i] = "+"
      }
}
FullExperiment$Strand <- dna_strand
rm(dna_strand)
```

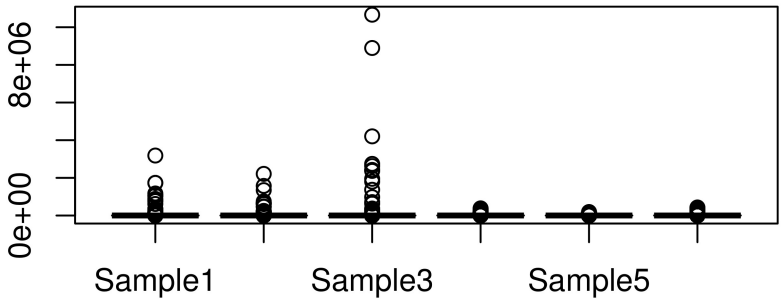
```
phenoData<-read.delim("C:/Users/janet/Documents/R/Cursera/Genomic Data/Capstone/phenoData.txt", header =

counts <- as.matrix(cbind(FullExperiment$SRR1554534, FullExperiment$SRR1554535, FullExperiment$SRR15545

library(SummarizedExperiment)
library(GenomicRanges)
library(edgeR)

counts <- counts[rowMeans(counts)>10,]
se <- SummarizedExperiment(assays=list(counts=counts), colData = phenoData)
group <- rep(c("adult", "fetal"),c(3,3))

Exp_list <- DGEList(counts=counts, group=group)
Exp_list$samples
```

```
##         group  lib.size norm.factors
## Sample1 adult 101180337            1
## Sample2 adult  60940542            1
## Sample3 adult  74797493            1
## Sample4 fetal  85463039            1
## Sample5 fetal  43392540            1
## Sample6 fetal  91521717            1
```

```
keep<-filterByExpr(Exp_list)
Exp_list <- calcNormFactors(Exp_list)
Exp_list$samples
```

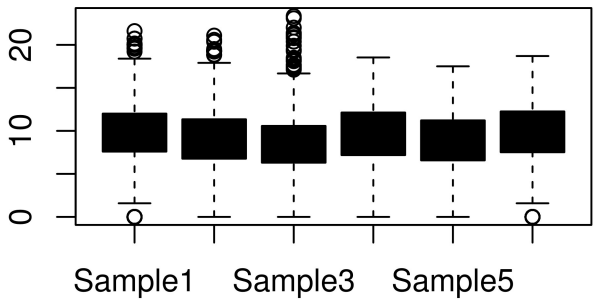
```
##         group  lib.size norm.factors
## Sample1 adult 101180337    1.0052739
## Sample2 adult  60940542    0.9945536
## Sample3 adult  74797493    0.5009252
## Sample4 fetal  85463039    1.2075906
## Sample5 fetal  43392540    1.3236132
## Sample6 fetal  91521717    1.2492049
```

```
boxplot(Exp_list$counts, col=1)
```

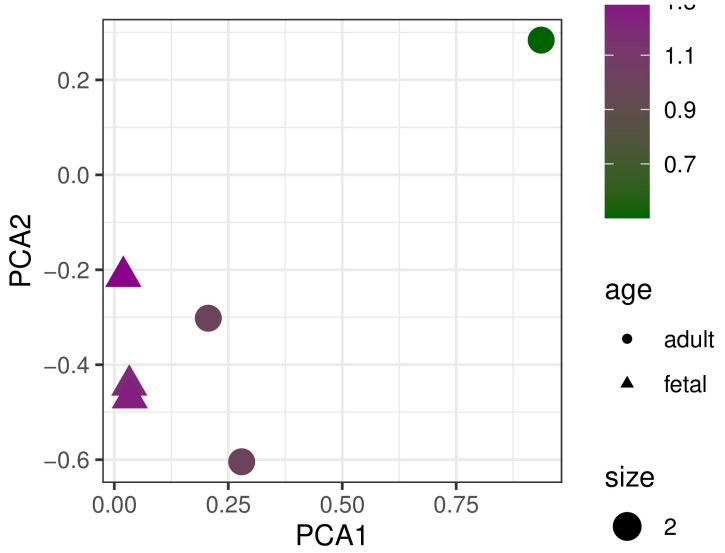



```
boxplot(log2(Exp_list$counts+1), col =1)
```

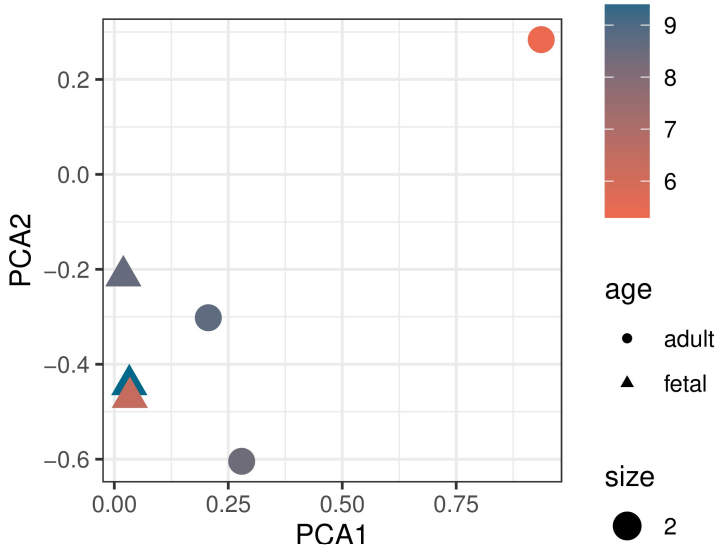
A correction on the scale needs to be done



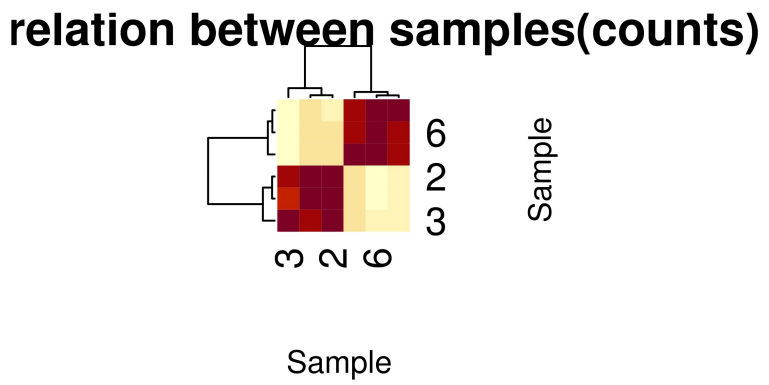
```
pr <- prcomp(Exp_list$counts)
DT <- data.frame(PCA1 = pr$rotation[,1], PCA2 = pr$rotation[,2],age = group, RIN = phenoData$RIN, libsi

library(ggplot2)

par(mfrow = c(1,2))
ggplot(data = DT, aes(x= PCA1, y = PCA2, color = libsize, shape= age, size=2)) +geom_point()+theme_bw()
```




```
ggplot(data = DT, aes(x= PCA1, y = PCA2, color = RIN, shape= age, size=2)) +geom_point()+theme_bw()+sca
```




```
r<- cor(counts,use = "all.obs")
heatmap(r, main = "Correlation between samples(counts)", xlab = "Sample ", ylab = "Sample ")
```



Here we can see that sample 1, 2 and 3 form a cluster, while sample 5 and 6 form another. This makes total sense since sample 1, 2 and 3 are adult samples while samples 4, 5, and 6 are fetal samples.

```
edata = assay(se)
edata = log2(as.matrix(edata) + 1)
edata = edata[rowMeans(edata) > 10, ]
```

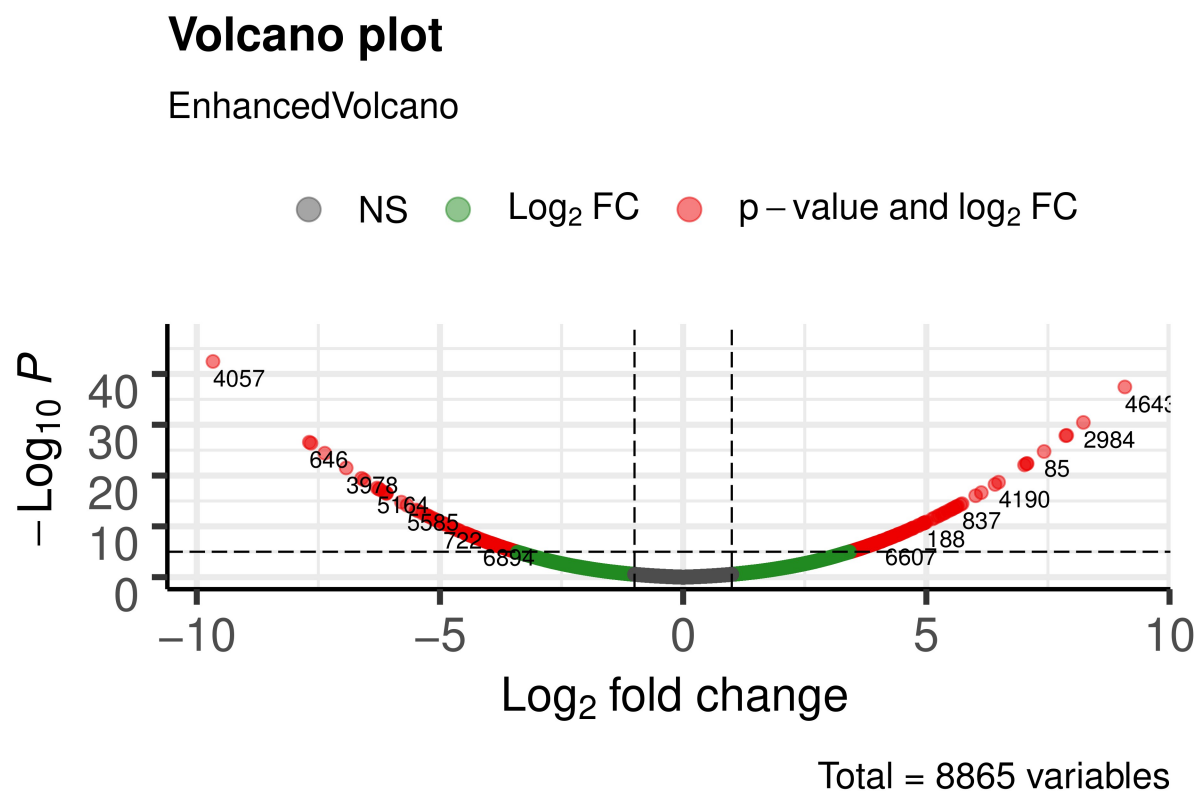
Statistical Analysis Statistical model was fitted by GROUP (factors: adult/fetal)

```
mod = model.matrix(~ se$GROUP)
fit_limma = lmFit(edata,mod)
ebayes_limma = eBayes(fit_limma)
limma_toptable = topTable(ebayes_limma,number=dim(edata)[1])
limma_table_output = limma_toptable[,c(1,4,5)]
```

```
head(limma_toptable)
```

```
##          logFC  AveExpr         t      P.Value    adj.P.Val        B
## 4057 -9.667069 14.12116 -14.44108 3.883792e-47 3.442982e-43 95.00083
## 4643  9.080950 12.45373  13.56551 8.156671e-42 3.615445e-38 83.07888
## 2984  8.228444 10.55299  12.29200 1.176564e-34 3.476747e-31 67.06909
## 596   7.887273 10.93518  11.78235 5.522757e-32 1.223981e-28 61.10385
## 8763  7.867660 13.47194  11.75305 7.804885e-32 1.383806e-28 60.76861
## 646  -7.686166 10.38121 -11.48192 1.841051e-30 2.720153e-27 57.70589
```

```
library(EnhancedVolcano)
EnhancedVolcano(limma_toptable, lab = rownames(limma_table_output),x = 'logFC',y = 'adj.P.Val')
```




```
print(paste("Number of Genes differencially expressed (adult/fetal):",sum(limma_table_output$adj.P.Val<
```

```
## [1] "Number of Genes differencially expressed (adult/fetal): 2107"
```

```
print(paste("Number of Genes upregulated (adult vs.fetal):",sum(limma_table_output$adj.P.Val<0.05 & limm
```

```
## [1] "Number of Genes upregulated (adult vs.fetal): 1060"
```

```
print(paste("Number of Genes downregulated (adult vs.fetal):",sum(limma_table_output$adj.P.Val<0.05 & li
```

```
## [1] "Number of Genes downregulated (adult vs.fetal): 377"
```

Gen Set Analysis Examine whether genes which are differentially associated between fetal and adult brain, are associated with changes in H3K4me3 in their promoters, between fetal and adult brain.

```
library(AnnotationHub)
ah <- AnnotationHub()
ah <- subset(ah, species == "Homo sapiens")
ah <- query(ah,"H3K4me3", "Roadmap Epigenomics")

fetal_query <- query(ah, c("H3K4me3", "EpigenomeRoadMap","E081")) # male fetal Brain sample_E081, AH294
fetal_data <- fetal_query[[2]] #retreiving narrow peaks dataset
```

```
adult_query <- query(ah, c("H3K4me3", "EpigenomeRoadMap","E073")) # Brain_DLPC_E073_AH29392
adult_data <- adult_query[[2]] #retreiving narrow peaks dataset
liver_query <- query(ah, c("H3K4me3", "EpigenomeRoadMap", "E066"))#liver sample_E066_AH30367
liver_data <- liver_query[[2]] #retreiving narrow peaks dataset
```

Extracting promoters of the differentially expressed genes between adult and fetal samples

```
library(TxDb.Hsapiens.UCSC.hg19.knownGene)
txdb <- TxDb.Hsapiens.UCSC.hg19.knownGene
txdb_genes <- genes(txdb)


diffExp_genes <- row.names(limma_toptable[limma_toptable$adj.P.Val < 0.05,])
diffExp_promoters <- promoters(txdb_genes[diffExp_genes %in% txdb_genes$gene_id])
```

Now let's find the overlap between the differentially expressed gene promoters and the fetal/adult/liver data sets.

```
fetalANDpromoters <-subsetByOverlaps(fetal_data, diffExp_promoters, ignore.strand=TRUE)
adultANDpromoters <- subsetByOverlaps(adult_data, diffExp_promoters, ignore.strand=TRUE)
liverANDpromoters <- subsetByOverlaps(liver_data, diffExp_promoters, ignore.strand=TRUE)
```

Percentage of overlap of H3K4me3 methilation of the Fetal/Adult/Liver samples with promoters of the differentialy expressed genes

```
FetalPercentage <- length(fetalANDpromoters)/length(fetal_data)*100
paste("Percentage of differentially expressed gene in fetal H3K4me3 narrowpeaks",round(FetalPercentage,
```

```
## [1] "Percentage of differentially expressed gene in fetal H3K4me3 narrowpeaks 25.48 %"
```

```
AdultPercentage <- length(adultANDpromoters)/length(adult_data)*100
paste("Percentage of differentially expressed gene in adult H3K4me3 narrowpeaks",round(AdultPercentage,
```

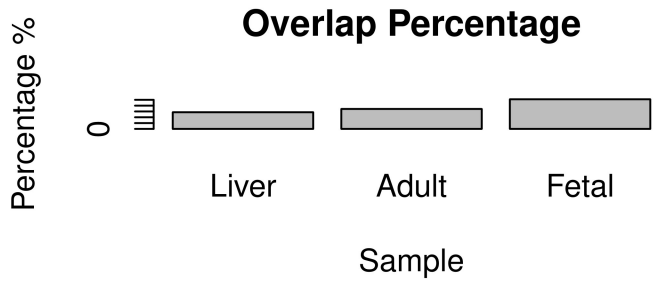
```
## [1] "Percentage of differentially expressed gene in adult H3K4me3 narrowpeaks 17.158 %"
```

```
LiverPercentage <- length(liverANDpromoters)/length(liver_data)*100
paste("Percentage of differentially expressed gene in liver H3K4me3 narrowpeaks",round(LiverPercentage,
```

```
## [1] "Percentage of differentially expressed gene in liver H3K4me3 narrowpeaks 14.34 %"
```

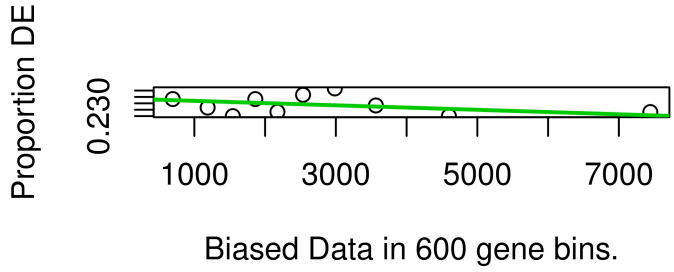
Promoters for the list of differentially expressed genes are less marked by H3K4me3 in liver than in the Brain(adult/fetal).

```
barplot(c(LiverPercentage, AdultPercentage, FetalPercentage),
main = "Overlap Percentage",
xlab = "Sample",
ylab = "Percentage %",
names.arg = c("Liver", "Adult", "Fetal"))
```



Gene Ontology analysis

```
new_genes <- as.integer(limma_toptable$adj.P.Val < 0.05)
not_na <- !is.na(genes)
names(new_genes) <- rownames(limma_toptable)

new_genes <- new_genes[not_na]
adj_pwf <- nullp(new_genes,"hg19","knownGene")
```




```
head(adj_pwf)
```

```
##      DEgenes bias.data       pwf
## 4057       1    2450.5 0.2394085
## 4643       1    4742.0 0.2354012
## 2984       1    2714.5 0.2389486
## 596        1    6495.0 0.2323223
## 8763       1    3059.0 0.2383479
## 646        1    4613.5 0.2356266
```

```
adj_GO.wall <- goseq(adj_pwf,"hg19","knownGene")
head(adj_GO.wall, n=10)
```

```
##         category over_represented_pvalue under_represented_pvalue numDEInCat
## 14743 GO:0071872           0.0008299286                0.9999175          8
## 13005 GO:0060218           0.0009331774                0.9996223         28
```

```
## 15435 GO:0090263            0.0012002594                0.9994597         33
## 14742 GO:0071871            0.0019871146                0.9997375          8
## 8813  GO:0035725            0.0021453880                0.9990461         29
## 10446 GO:0044794            0.0026536460                0.9997033          7
## 6440  GO:0030177            0.0026932197                0.9986433         37
## 9448  GO:0042287            0.0027615968                0.9993804         11
## 16766 GO:1900378            0.0031827318                1.0000000          4
## 19184 GO:2000628            0.0032703331                1.0000000          4
##       numInCat                                                           term
## 14743       11                       cellular response to epinephrine stimulus
## 13005       67                         hematopoietic stem cell differentiation
## 15435       84           positive regulation of canonical Wnt signaling pathway
## 14742       12                                        response to epinephrine
## 8813        74                               sodium ion transmembrane transport
## 10446       10                    positive regulation by host of viral process
## 6440       101                    positive regulation of Wnt signaling pathway
## 9448        20                                            MHC protein binding
## 16766        4 positive regulation of secondary metabolite biosynthetic process
## 19184        4                           regulation of miRNA metabolic process
##       ontology
## 14743       BP
## 13005       BP
## 15435       BP
## 14742       BP
## 8813        BP
## 10446       BP
## 6440        BP
## 9448        MF
## 16766       BP
## 19184       BP
```

```
sessionInfo()
```

```
## R version 3.6.1 (2019-07-05)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 18362)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_Germany.1252  LC_CTYPE=English_Germany.1252
## [3] LC_MONETARY=English_Germany.1252 LC_NUMERIC=C
## [5] LC_TIME=English_Germany.1252
##
## attached base packages:
## [1] parallel  stats4    stats     graphics  grDevices utils     datasets
## [8] methods   base
##
## other attached packages:
##  [1] DESeq2_1.26.0
##  [2] goseq_1.38.0
##  [3] geneLenDataBase_1.22.0
##  [4] BiasedUrn_1.07
##  [5] devtools_2.3.1
```

```
##  [6] usethis_1.6.1
##  [7] org.Hs.eg.db_3.10.0
##  [8] TxDb.Hsapiens.UCSC.hg19.knownGene_3.2.2
##  [9] GenomicFeatures_1.38.2
## [10] AnnotationDbi_1.48.0
## [11] BSgenome.Hsapiens.UCSC.hg19_1.4.0
## [12] BSgenome_1.54.0
## [13] Biostrings_2.54.0
## [14] XVector_0.26.0
## [15] rtracklayer_1.46.0
## [16] AnnotationHub_2.18.0
## [17] BiocFileCache_1.10.2
## [18] dbplyr_1.4.4
## [19] EnhancedVolcano_1.4.0
## [20] ggrepel_0.8.2
## [21] ggplot2_3.3.2
## [22] edgeR_3.28.1
## [23] limma_3.42.2
## [24] SummarizedExperiment_1.16.1
## [25] DelayedArray_0.12.3
## [26] BiocParallel_1.20.1
## [27] matrixStats_0.56.0
## [28] Biobase_2.46.0
## [29] GenomicRanges_1.38.0
## [30] GenomeInfoDb_1.22.1
## [31] IRanges_2.20.2
## [32] S4Vectors_0.24.4
## [33] BiocGenerics_0.32.0
## [34] purrr_0.3.4
## [35] tidyr_1.1.1
## [36] dplyr_1.0.0
##
## loaded via a namespace (and not attached):
##   [1] backports_1.1.8              Hmisc_4.4-1
##   [3] splines_3.6.1                digest_0.6.25
##   [5] htmltools_0.5.0              GO.db_3.10.0
##   [7] fansi_0.4.1                  magrittr_1.5
##   [9] checkmate_2.0.0              memoise_1.1.0
##  [11] cluster_2.1.0                remotes_2.2.0
##  [13] annotate_1.64.0              askpass_1.1
##  [15] prettyunits_1.1.1            jpeg_0.1-8.1
##  [17] colorspace_1.4-1             blob_1.2.1
##  [19] rappdirs_0.3.1               xfun_0.15
##  [21] callr_3.4.3                  crayon_1.3.4
##  [23] RCurl_1.98-1.2               genefilter_1.68.0
##  [25] survival_3.2-3               glue_1.4.1
##  [27] gtable_0.3.0                 zlibbioc_1.32.0
##  [29] pkgbuild_1.1.0               scales_1.1.1
##  [31] DBI_1.1.0                    Rcpp_1.0.5
##  [33] xtable_1.8-4                 progress_1.2.2
##  [35] htmlTable_2.0.1              foreign_0.8-71
##  [37] bit_4.0.4                    Formula_1.2-3
##  [39] htmlwidgets_1.5.1            httr_1.4.2
##  [41] RColorBrewer_1.1-2           ellipsis_0.3.1
```

```
##  [43] pkgconfig_2.0.3                XML_3.99-0.3
##  [45] farver_2.0.3                   nnet_7.3-12
##  [47] locfit_1.5-9.4                 tidyselect_1.1.0
##  [49] labeling_0.3                   rlang_0.4.6
##  [51] later_1.1.0.1                  munsell_0.5.0
##  [53] BiocVersion_3.10.1             tools_3.6.1
##  [55] cli_2.0.2                      generics_0.0.2
##  [57] RSQLite_2.2.0                  evaluate_0.14
##  [59] stringr_1.4.0                  fastmap_1.0.1
##  [61] yaml_2.2.1                     processx_3.4.3
##  [63] knitr_1.29                     bit64_4.0.2
##  [65] fs_1.5.0                       nlme_3.1-140
##  [67] mime_0.9                       biomaRt_2.42.1
##  [69] rstudioapi_0.11                compiler_3.6.1
##  [71] curl_4.3                       png_0.1-7
##  [73] interactiveDisplayBase_1.24.0 testthat_2.3.2
##  [75] geneplotter_1.64.0             tibble_3.0.2
##  [77] stringi_1.4.6                  ps_1.3.4
##  [79] desc_1.2.0                     lattice_0.20-38
##  [81] Matrix_1.2-17                  vctrs_0.3.1
##  [83] pillar_1.4.6                   lifecycle_0.2.0
##  [85] BiocManager_1.30.10            data.table_1.13.0
##  [87] bitops_1.0-6                   httpuv_1.5.4
##  [89] R6_2.4.1                       latticeExtra_0.6-29
##  [91] promises_1.1.1                 gridExtra_2.3
##  [93] sessioninfo_1.1.1              assertthat_0.2.1
##  [95] pkgload_1.1.0                  openssl_1.4.2
##  [97] rprojroot_1.3-2                withr_2.2.0
##  [99] GenomicAlignments_1.22.1       Rsamtools_2.2.3
## [101] GenomeInfoDbData_1.2.2         mgcv_1.8-28
## [103] hms_0.5.3                      grid_3.6.1
## [105] rpart_4.1-15                   rmarkdown_2.3
## [107] shiny_1.5.0                    base64enc_0.1-3
```