Finding Differentially Expressed Genes in the Kennerdell et al. (2018) RNA-seq data.

Jason Kennerdell

11/15/2017

Synopsis:

Here I am calling genes as different differentially expressed in the *E(z)* mutant *Drosophila* brains, and genes that are differentially expressed with age. The csv files are used to create the Supplemental Data 1 spreadsheet.

Input the data and metadata:

```
inpath <- "~/Desktop/brain"
outpath <- "~/Desktop/brain/CalledDE"
setwd(inpath)
library(RColorBrewer)
colSet <- brewer.pal(12, "Paired")
files <- list.files()
htseq_files <- files[grepl("^JKL.*txt$", files)]
EzNames <- read.csv("EzSampleNames.csv")
PclNames <- read.csv("PclSampleNames.csv")
sampleNames <- rbind(EzNames, PclNames)
sampleTable <- data.frame(fileName = htseq_files,
                          stringsAsFactors=FALSE)
sampleTable$Library <- gsub("-counts.txt", "", sampleTable$fileName)
sampleTable$Library <- gsub("b", "", sampleTable$Library)
sampleTable$seq.batch <- ifelse(grepl("b", sampleTable$fileName), "B", "A")
sampleTable$seq.batch <- paste(sampleTable$Library, sampleTable$seq.batch)
sampleTable <- merge(sampleTable, sampleNames[,1:3], by = "seq.batch")
sampleTable$genotype <- gsub("-.*$", "", sampleTable$Sample)
sampleTable$genotype <- gsub("JKLY.... ", "", sampleTable$genotype)
sampleTable$genotype <- gsub("\\[1118\\]", "", sampleTable$genotype)
sampleTable$genotype <- factor(sampleTable$genotype, levels = c("w", "Ez"))
sampleTable <- sampleTable[!is.na(sampleTable$genotype),]
sampleTable$Temp <- gsub("[A-Z, a-z, -]", "", sampleTable$Sample)
sampleTable$Temp <- gsub("^11.*$", "25", sampleTable$Temp)
sampleTable$Temp <- factor(sampleTable$Temp, levels = c("25", "29"))
sampleTable$age <- gsub("JKLY.*$", "3d", sampleTable$Sample)
sampleTable$age <- gsub("^[a-zA-z].*$", "20d", sampleTable$age)
sampleTable$age <- factor(sampleTable$age, levels = c("3d", "20d"))
sampleTable$condition <- paste(sampleTable$genotype, sampleTable$Temp, sampleTable$age, sep = "-")
sampleTable$color <- c(rep(colSet[7], 4), rep(colSet[9], 2),
                       rep(colSet[8], 2), rep(colSet[10], 2), rep(colSet[8], 2), rep(colSet[10], 2),
                       rep(colSet[8], 2), rep(colSet[10], 2), rep(colSet[8], 2),
                       rep(colSet[1], 2),
                       rep(colSet[10], 2), rep(colSet[8], 2), rep(colSet[10], 2),
```

```
                      rep(colSet[1], 4), rep(colSet[7], 2), rep(colSet[9], 4))
sampleTable
```

```
##    seq.batch          fileName Library           Sample batch genotype
## 3    JKL10 A  JKL10-counts.txt   JKL10          w-25-II     A        w
## 4    JKL10 B JKL10b-counts.txt   JKL10          w-25-II     B        w
## 5    JKL11 A  JKL11-counts.txt   JKL11         w-25-III     A        w
## 6    JKL11 B JKL11b-counts.txt   JKL11         w-25-III     B        w
## 7    JKL12 A  JKL12-counts.txt   JKL12        Ez-25-III     A       Ez
## 8    JKL12 B JKL12b-counts.txt   JKL12        Ez-25-III     B       Ez
## 9    JKL13 A  JKL13-counts.txt   JKL13           w-29-I     A        w
## 10   JKL13 B JKL13b-counts.txt   JKL13           w-29-I     B        w
## 11   JKL14 A  JKL14-counts.txt   JKL14          Ez-29-I     A       Ez
## 12   JKL14 B JKL14b-counts.txt   JKL14          Ez-29-I     B       Ez
## 13   JKL15 A  JKL15-counts.txt   JKL15          w-29-II     A        w
## 14   JKL15 B JKL15b-counts.txt   JKL15          w-29-II     B        w
## 15   JKL16 A  JKL16-counts.txt   JKL16         Ez-29-II     A       Ez
## 16   JKL16 B JKL16b-counts.txt   JKL16         Ez-29-II     B       Ez
## 17   JKL17 A  JKL17-counts.txt   JKL17         w-29-III     A        w
## 18   JKL17 B JKL17b-counts.txt   JKL17         w-29-III     B        w
## 19   JKL18 A  JKL18-counts.txt   JKL18        Ez-29-III     A       Ez
## 20   JKL18 B JKL18b-counts.txt   JKL18        Ez-29-III     B       Ez
## 21   JKL19 A  JKL19-counts.txt   JKL19          w-29-IV     A        w
## 22   JKL19 B JKL19b-counts.txt   JKL19          w-29-IV     B        w
## 23    JKL2 A   JKL2-counts.txt    JKL2 JKLY1125 w[1118]     A        w
## 24    JKL2 B  JKL2b-counts.txt    JKL2 JKLY1125 w[1118]     B        w
## 25   JKL20 A  JKL20-counts.txt   JKL20         Ez-29-IV     A       Ez
## 26   JKL20 B JKL20b-counts.txt   JKL20         Ez-29-IV     B       Ez
## 27   JKL21 A  JKL21-counts.txt   JKL21           w-29-V     A        w
## 28   JKL21 B JKL21b-counts.txt   JKL21           w-29-V     B        w
## 29   JKL22 A  JKL22-counts.txt   JKL22          Ez-29-V     A       Ez
## 30   JKL22 B JKL22b-counts.txt   JKL22          Ez-29-V     B       Ez
## 33    JKL4 A   JKL4-counts.txt    JKL4 JKLY1127 w[1118]     A        w
## 34    JKL4 B  JKL4b-counts.txt    JKL4 JKLY1127 w[1118]     B        w
## 37    JKL6 A   JKL6-counts.txt    JKL6 JKLY1129 w[1118]     A        w
## 38    JKL6 B  JKL6b-counts.txt    JKL6 JKLY1129 w[1118]     B        w
## 39    JKL7 A   JKL7-counts.txt    JKL7           w-25-I     A        w
## 40    JKL7 B  JKL7b-counts.txt    JKL7           w-25-I     B        w
## 41    JKL8 A   JKL8-counts.txt    JKL8          Ez-25-I     A       Ez
## 42    JKL8 B  JKL8b-counts.txt    JKL8          Ez-25-I     B       Ez
## 43    JKL9 A   JKL9-counts.txt    JKL9         Ez-25-II     A       Ez
## 44    JKL9 B  JKL9b-counts.txt    JKL9         Ez-25-II     B       Ez
##    Temp age condition   color
## 3    25 20d  w-25-20d #FDBF6F
## 4    25 20d  w-25-20d #FDBF6F
## 5    25 20d  w-25-20d #FDBF6F
## 6    25 20d  w-25-20d #FDBF6F
## 7    25 20d Ez-25-20d #CAB2D6
## 8    25 20d Ez-25-20d #CAB2D6
## 9    29 20d  w-29-20d #FF7F00
## 10   29 20d  w-29-20d #FF7F00
## 11   29 20d Ez-29-20d #6A3D9A
## 12   29 20d Ez-29-20d #6A3D9A
## 13   29 20d  w-29-20d #FF7F00
```

```
## 14   29 20d  w-29-20d #FF7F00
## 15   29 20d Ez-29-20d #6A3D9A
## 16   29 20d Ez-29-20d #6A3D9A
## 17   29 20d  w-29-20d #FF7F00
## 18   29 20d  w-29-20d #FF7F00
## 19   29 20d Ez-29-20d #6A3D9A
## 20   29 20d Ez-29-20d #6A3D9A
## 21   29 20d  w-29-20d #FF7F00
## 22   29 20d  w-29-20d #FF7F00
## 23   25  3d   w-25-3d #A6CEE3
## 24   25  3d   w-25-3d #A6CEE3
## 25   29 20d Ez-29-20d #6A3D9A
## 26   29 20d Ez-29-20d #6A3D9A
## 27   29 20d  w-29-20d #FF7F00
## 28   29 20d  w-29-20d #FF7F00
## 29   29 20d Ez-29-20d #6A3D9A
## 30   29 20d Ez-29-20d #6A3D9A
## 33   25  3d   w-25-3d #A6CEE3
## 34   25  3d   w-25-3d #A6CEE3
## 37   25  3d   w-25-3d #A6CEE3
## 38   25  3d   w-25-3d #A6CEE3
## 39   25 20d  w-25-20d #FDBF6F
## 40   25 20d  w-25-20d #FDBF6F
## 41   25 20d Ez-25-20d #CAB2D6
## 42   25 20d Ez-25-20d #CAB2D6
## 43   25 20d Ez-25-20d #CAB2D6
## 44   25 20d Ez-25-20d #CAB2D6
```

Set up the statistical model to test for Differentially Expressed genes in E(z) mutants:

```
design <- formula(~ Temp + age + genotype)
```

DESeq2 Statistics

```
dds <- DESeqDataSetFromHTSeqCount(sampleTable = sampleTable, directory = inpath, design = design)
# Combine the technical replicates (different runs) by adding the count
# totals for each gene across the two runs:
dds <- collapseReplicates(dds, groupby=dds$Library, run = dds$batch)
dds <- DESeq(dds)
# What does the data look like?
head(assay(dds))
```

```
##             JKL10 JKL11 JKL12 JKL13 JKL14 JKL15 JKL16 JKL17 JKL18 JKL19
## FBgn0000003     0     0     0     0     0     0     0     0     0     0
## FBgn0000008  1444  1874  1687  1305  1453  1725  1529  1856  1500  1558
## FBgn0000014     0     0     0     0     0     0     1     0     1     0
## FBgn0000015     1     6     1     3     0     1     1     1     2     0
## FBgn0000017  9186 10798  9189  6877  8808  9121  9834 11313  9272  8564
## FBgn0000018   262   306   286   336   288   317   272   346   274   285
##             JKL2 JKL20 JKL21 JKL22 JKL4  JKL6  JKL7  JKL8  JKL9
## FBgn0000003    0     1     0     0    0     0     0     0     1
## FBgn0000008 1218  1353  1589  1367 1277  1573  1980  1861  1884
```

```
## FBgn0000014    0     1     1     1    0     0     0     3     0
## FBgn0000015    2     0     1     0    1     2     0     0     2
## FBgn0000017 9134  8697  8196  8612 9526 11793 11453 11360 11139
## FBgn0000018  204   285   238   284  158   202   286   333   333
```

```
# What are the columns?
colData(dds)
```

```
## DataFrame with 19 rows and 10 columns
##           Library           Sample    batch genotype     Temp      age
##       <character>         <factor> <factor> <factor> <factor> <factor>
## JKL10       JKL10          w-25-II        A        w       25      20d
## JKL11       JKL11         w-25-III        A        w       25      20d
## JKL12       JKL12        Ez-25-III        A       Ez       25      20d
## JKL13       JKL13           w-29-I        A        w       29      20d
## JKL14       JKL14          Ez-29-I        A       Ez       29      20d
## ...           ...              ...      ...      ...      ...      ...
## JKL4         JKL4 JKLY1127 w[1118]        A        w       25       3d
## JKL6         JKL6 JKLY1129 w[1118]        A        w       25       3d
## JKL7         JKL7           w-25-I        A        w       25      20d
## JKL8         JKL8          Ez-25-I        A       Ez       25      20d
## JKL9         JKL9         Ez-25-II        A       Ez       25      20d
##         condition       color runsCollapsed sizeFactor
##       <character> <character>   <character>  <numeric>
## JKL10    w-25-20d     #FDBF6F           A,B  0.9301385
## JKL11    w-25-20d     #FDBF6F           A,B  1.1761379
## JKL12   Ez-25-20d     #CAB2D6           A,B  1.0267534
## JKL13    w-29-20d     #FF7F00           A,B  1.0565380
## JKL14   Ez-29-20d     #6A3D9A           A,B  0.9791047
## ...           ...         ...           ...        ...
## JKL4      w-25-3d     #A6CEE3           A,B  0.7871747
## JKL6      w-25-3d     #A6CEE3           A,B  0.9912254
## JKL7     w-25-20d     #FDBF6F           A,B  1.1662988
## JKL8    Ez-25-20d     #CAB2D6           A,B  1.1349590
## JKL9    Ez-25-20d     #CAB2D6           A,B  1.1534079
```

```
resultsEz <- results(dds, alpha=0.05) # this gives identical for E(z) vs wt contrast
resultsEz$ensembl <- rownames(resultsEz)
summary(resultsEz)
```

```
##
## out of 15383 with nonzero total read count
## adjusted p-value < 0.05
## LFC > 0 (up)     : 1197, 7.8%
## LFC < 0 (down)   : 1136, 7.4%
## outliers [1]     : 111, 0.72%
## low counts [2]   : 2944, 19%
## (mean count < 1)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

```
resultsAge <- results(dds, alpha=0.05, contrast=c("age", "20d", "3d"))
resultsAge$ensembl <- rownames(resultsAge)
summary(resultsAge)
```

```
##
```

```
## out of 15383 with nonzero total read count
## adjusted p-value < 0.05
## LFC > 0 (up)     : 2219, 14%
## LFC < 0 (down)   : 2439, 16%
## outliers [1]     : 111, 0.72%
## low counts [2]   : 4103, 27%
## (mean count < 2)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

Add Annotations

```
# Add usefull gene names:
library(biomaRt)
mart = useMart("ENSEMBL_MART_ENSEMBL", host="aug2017.archive.ensembl.org")
#listDatasets(mart)
mart = useMart("ENSEMBL_MART_ENSEMBL", host="aug2017.archive.ensembl.org",
               dataset = "dmelanogaster_gene_ensembl")

# For the Ez gene list:
genemap <- getBM(attributes = c("ensembl_gene_id", "entrezgene", "external_gene_name",
                                "flybasecgid_gene"), filters = "ensembl_gene_id",
                 values = resultsEz$ensembl, mart = mart)
idx <- match(resultsEz$ensembl, genemap$ensembl_gene_id)
resultsEz$entrez <- genemap$entrezgene[idx]
resultsEz$geneSymbol <- genemap$external_gene_name[idx]
resultsEz$cg <- genemap$flybasecgid_gene[idx]
write.csv(as.data.frame(resultsEz), file = paste(outpath, "EzContrast_Benjp05.csv", sep="/"))
# For the Aging gene list:
genemap <- getBM(attributes = c("ensembl_gene_id", "entrezgene", "external_gene_name",
                                "flybasecgid_gene"), filters = "ensembl_gene_id",
                 values = resultsAge$ensembl, mart = mart)
idx <- match(resultsAge$ensembl, genemap$ensembl_gene_id)
resultsAge$entrez <- genemap$entrezgene[idx]
resultsAge$geneSymbol <- genemap$external_gene_name[idx]
resultsAge$cg <- genemap$flybasecgid_gene[idx]
write.csv(as.data.frame(resultsAge), file = paste(outpath, "AgeContrast_Benjp05.csv", sep="/"))
```

Additional Diagnostic Plots

Print out gene names of called diferentially expressed genes:

How about Bonferroni corrected data?

```
# For Ez Contrast:
resultsEzBonf <- results(dds, alpha = 0.05, pAdjustMethod = "bonferroni")
resultsEzBonf$ensembl <- rownames(resultsEzBonf)
idxEzBonf <- match(resultsEzBonf$ensembl, genemap$ensembl_gene_id)
resultsEzBonf$geneSymbol <- genemap$external_gene_name[idxEzBonf]
write.csv(as.data.frame(resultsEzBonf), file = paste(outpath, "EzContrast_Bonfp05.csv", sep="/"))
# For Aging Contrast:
```

```
resultsAgeBonf <- results(dds, alpha=0.05, contrast=c("age", "20d", "3d"), pAdjustMethod = "bonferroni")
resultsAgeBonf$ensembl <- rownames(resultsAgeBonf)
idxAgeBonf <- match(resultsAgeBonf$ensembl, genemap$ensembl_gene_id)
resultsAgeBonf$geneSymbol <- genemap$external_gene_name[idxAgeBonf]
write.csv(as.data.frame(resultsAgeBonf), file = paste(outpath, "AgeContrast_Bonfp05.csv", sep="/"))
```

Prepare lists of DE genes:

```
datEz <- as.data.frame(resultsEz)
datEz$Bonf.correction <- as.data.frame(resultsEzBonf)$padj
calledEz <- subset(datEz, padj < 0.05 & abs(log2FoldChange) > 0.5)
write.csv(calledEz, file = paste(outpath, "EzCalledDE.csv", sep="/"))
datAge <- as.data.frame(resultsAge)
datAge$Bonf.correction <- as.data.frame(resultsAgeBonf)$padj
calledAge <- subset(datAge, padj < 0.05 & abs(log2FoldChange) > 0.5)
write.csv(calledAge, file = paste(outpath, "AgeCalledDE.csv", sep="/"))
```