analysis_polyA_vs_pseudo

Ariel Hippen

2022-07-31

Contents

Analyzing differential expression results

We have calculated differential expression genes in DE_ribo_vs_polyA.Rmd, now we will try to make sense of them. Our main workhorse will be Gene Set Enrichment Analysis (GSEA) across several reference sets.

```
suppressPackageStartupMessages({
  library(DESeq2)
  library(WebGestaltR)
  library(ggplot2)
  library(yaml)
})

params <- read_yaml("../../config.yml")
data_path <- params$data_path
local_data_path <- params$local_data_path
samples <- params$samples
```

```
# Load the DESeq2 object with the original count matrix
deseq_path <- paste(local_data_path, "deseq2_output", sep = "/")
dds <- readRDS(paste(deseq_path, "polyA_vs_pseudo_data.rds", sep = "/"))

# Load the DESeqResults object with differentially expressed genes, at FDR 0.1 and 0.05
res1 <- readRDS(paste(deseq_path, "polyA_vs_pseudo_FDR_0.1.rds", sep = "/"))
res05 <- readRDS(paste(deseq_path, "polyA_vs_pseudo_FDR_0.05.rds", sep = "/"))
```

Top genes

Let's look at the top 20 most upregulated and downregulated genes and see if we can find a pattern.

```
res1 <- subset(res1, res1$padj < 0.1)
res1 <- res1[order(res1$log2FoldChange), ]
as.data.frame(head(res1, n=20))
```

```
##              baseMean log2FoldChange     lfcSE       stat       pvalue
## AC000123.1 104.04962     -10.812481 0.8005574 -13.506190 1.437723e-41
## LAMC1-AS1  146.00598      -9.835680 0.7034012 -13.983030 1.978809e-44
## LIMS1-AS1   98.41865      -9.668729 0.7090164 -13.636819 2.418826e-42
## AC005104.1  53.02654      -9.415706 0.7845085 -12.002045 3.466253e-33
## AC011510.1  55.93759      -9.395642 0.7933071 -11.843638 2.321605e-32
## AC021683.1  74.80835      -9.049879 0.7903612 -11.450307 2.343137e-30
## NADK2-AS1   56.18247      -8.946004 0.7076319 -12.642172 1.235924e-36
## LINC02637   31.00691      -8.891071 0.8310229 -10.698948 1.029402e-26
## AL355075.2  92.56400      -8.792439 0.5993409 -14.670181 1.000778e-48
## AP002340.1  34.21704      -8.690400 0.7941363 -10.943209 7.161895e-28
## AP001793.1  35.96543      -8.605386 0.8157615 -10.548900 5.139547e-26
## AF235103.1  40.69350      -8.536563 0.7995084 -10.677265 1.300469e-26
## AC018695.4  53.63662      -8.536246 0.7998947 -10.671712 1.380594e-26
## AC005884.2  45.77090      -8.437622 0.7716925 -10.933918 7.934707e-28
## SPDYE2      50.93918      -8.355601 0.7641029 -10.935178 7.825257e-28
## ETV5-AS1    36.10791      -8.352391 0.7857165 -10.630285 2.154538e-26
## AL138921.2  67.98953      -8.324670 0.5971517 -13.940628 3.587702e-44
## AL139349.1  39.62756      -8.303501 0.7869751 -10.551161 5.017322e-26
## IGKV1-33    36.84540      -8.225531 0.8373767  -9.822975 8.965780e-23
## AL133406.2 119.17882      -8.192827 0.4586779 -17.861833 2.338310e-71
##                    padj
## AC000123.1 3.194682e-40
## LAMC1-AS1  4.884720e-43
## LIMS1-AS1  5.527571e-41
## AC005104.1 5.374178e-32
## AC011510.1 3.501896e-31
## AC021683.1 3.221607e-29
## NADK2-AS1  2.230600e-35
## LINC02637  1.186519e-25
## AL355075.2 2.827490e-47
## AP002340.1 8.716903e-27
## AP001793.1 5.710147e-25
## AF235103.1 1.491236e-25
## AC018695.4 1.579983e-25
## AC005884.2 9.641285e-27
## SPDYE2     9.512291e-27
## ETV5-AS1   2.444429e-25
## AL138921.2 8.729671e-43
## AL139349.1 5.580782e-25
## IGKV1-33   8.449153e-22
## AL133406.2 1.308504e-69
```

Reminder, these ˆ are the ones that are much more expressed in true bulk than in pseudobulk. This looks like a lot of lncRNAs. None of the genes jump out as significant (except for maybe the immunoglobulin gene IHGJ3P).

```
as.data.frame(tail(res1, n=20))
```

```
##                baseMean log2FoldChange     lfcSE     stat        pvalue
## ICAM4          71.61889       8.931028 0.8025448 11.12838  9.127426e-29
## AP005329.3     73.00113       9.193496 0.7837153 11.73066  8.876524e-32
## AC005523.2     64.39783       9.218436 0.7812961 11.79890  3.954423e-32
## CIDEB          68.12531       9.320185 0.7825104 11.91062  1.041985e-32
## AL391121.1    104.98773       9.466573 0.7703664 12.28840  1.045498e-34
## LINC02591     153.19954       9.518911 0.7586782 12.54670  4.144025e-36
## MAFIP          74.03265       9.749916 0.7826940 12.45687  1.283120e-35
## RPS10-NUDT3    80.45463       9.790623 0.7955178 12.30723  8.281369e-35
## AC092069.1    185.49030       9.876999 0.7676609 12.86636  6.960854e-38
## AL133453.1    198.81626      10.155562 0.7585382 13.38833  7.075115e-41
## HIST1H2AK     190.71820      10.391822 0.7689105 13.51500  1.275629e-41
## DDT          5426.62151      10.578877 0.3302520 32.03274 3.818486e-225
## AC087190.1    258.54147      10.594199 0.7694527 13.76848  3.943974e-43
## SFT2D3        189.19177      10.660077 0.7623820 13.98259  1.991004e-44
## AC020656.1    562.90434      11.351167 0.7730433 14.68374  8.194215e-49
## EVA1B         772.85972      11.691198 0.7500108 15.58804  8.778475e-55
## HYI           508.34796      12.111364 0.7697252 15.73466  8.752764e-56
## ZNF593        914.83142      12.617987 0.7581973 16.64209  3.453547e-62
## C19orf33     3002.60065      13.981387 0.7680679 18.20332  4.857007e-74
## MIF         12368.57330      14.850642 1.0777692 13.77906  3.406942e-43
##                      padj
## ICAM4        1.168949e-27
## AP005329.3   1.300946e-30
## AC005523.2   5.900228e-31
## CIDEB        1.589124e-31
## AL391121.1   1.727430e-33
## LINC02591    7.310414e-35
## MAFIP        2.204391e-34
## RPS10-NUDT3  1.372999e-33
## AC092069.1   1.329270e-36
## AL133453.1   1.515101e-39
## HIST1H2AK    2.843238e-40
## DDT         2.569130e-222
## AC087190.1   9.284224e-42
## SFT2D3       4.910633e-43
## AC020656.1   2.337937e-47
## EVA1B        3.056198e-53
## HYI          3.126249e-54
## ZNF593       1.520770e-60
## C19orf33     2.867716e-72
## MIF          8.046225e-42
```

These ˆ are the genes that are much more expressed in pseudobulk than true bulk. There are two histone genes, which I find interesting. I'm also interested that there's more MIF and DDT, which are apparently closely related inflammatory cytokines. It could be a coincidence though if there just happened to be more macrophages in that set.

GeneCards says both EVA1B and SFT2D3 are "predicted to be an integral component of membrane"… is that anything?

GSEA

WebGestaltR expects a data frame with two columns, gene name and fold change.

```
res1$gene <- rownames(res1); rownames(res1) <- NULL
res1 <- subset(res1, select=c("gene","log2FoldChange"))
res1 <- as.data.frame(res1)
nrow(res1)
```

```
## [1] 19820
```

```
res05 <- subset(res05, res05$padj < 0.05)
res05$gene <- rownames(res05); rownames(res05) <- NULL
res05 <- subset(res05, select=c("gene","log2FoldChange"))
res05 <- as.data.frame(res05)
nrow(res05)
```

```
## [1] 18433
```

GO Biological process

Our first try at GSEA will use the same reference set we used for overrepresentation analysis in the single-cell data, GO Biological process.

```
GO_bp <- suppressWarnings(WebGestaltR(enrichMethod = "GSEA",
                          enrichDatabase = "geneontology_Biological_Process_noRedundant",
                          interestGene = res1,
                          interestGeneType = "genesymbol",
                          isOutput = FALSE))
```

```
## Loading the functional categories...
## Loading the ID list...
## Performing the enrichment analysis...
## 1000 permutations of score complete...
```

```
nrow(GO_bp)
```

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

```
GO_bp <- GO_bp[order(GO_bp$normalizedEnrichmentScore, decreasing = TRUE),]
head(subset(GO_bp, select=c("geneSet","description","normalizedEnrichmentScore","pValue","FDR","size"))
```

```
##      geneSet                                      description
## 1 GO:0033108 mitochondrial respiratory chain complex assembly
## 2 GO:0010257             NADH dehydrogenase complex assembly
## 3 GO:0017004                     cytochrome complex assembly
## 4 GO:0140053                    mitochondrial gene expression
## 5 GO:0006414                          translational elongation
## 6 GO:0009141        nucleoside triphosphate metabolic process
##   normalizedEnrichmentScore pValue          FDR size
## 1                  3.180506      0 0.0000000000   83
## 2                  2.862794      0 0.0000000000   52
## 3                  2.818635      0 0.0000000000   32
## 4                  2.559436      0 0.0000000000  137
## 5                  2.520298      0 0.0000000000  112
## 6                  2.336421      0 0.0001532636  231
```

```
tail(subset(GO_bp, select=c("geneSet","description","normalizedEnrichmentScore","pValue","FDR","size"))
```

```
##       geneSet                                               description
## 22 GO:0014074                    response to purine-containing compound
## 13 GO:0035249                      synaptic transmission, glutamatergic
## 12 GO:0007606                     sensory perception of chemical stimulus
## 10 GO:0051932                          synaptic transmission, GABAergic
## 14 GO:0098742 cell-cell adhesion via plasma-membrane adhesion molecules
## 15 GO:0043062                        extracellular structure organization
##    normalizedEnrichmentScore pValue         FDR size
## 22                 -1.980920      0 0.022446001  104
## 13                 -2.036757      0 0.013361557   53
## 12                 -2.076490      0 0.011134631   95
## 10                 -2.109470      0 0.009543969   22
## 14                 -2.111255      0 0.014315954  183
## 15                 -2.148563      0 0.015906615  294
```

```
GO_bp_05 <- suppressWarnings(WebGestaltR(enrichMethod = "GSEA",
                             enrichDatabase = "geneontology_Biological_Process_noRedundant",
                             interestGene = res05,
                             interestGeneType = "genesymbol",
                             isOutput = FALSE))
```

```
## Loading the functional categories...
## Loading the ID list...
## Performing the enrichment analysis...
## 1000 permutations of score complete...
```

```
nrow(GO_bp_05)
```

```
## [1] 32
```

```
GO_bp_05 <- GO_bp_05[order(GO_bp_05$normalizedEnrichmentScore, decreasing = TRUE),]
head(subset(GO_bp_05, select=c("geneSet","description","normalizedEnrichmentScore","pValue","FDR","size
```

```
##      geneSet                                      description
## 1 GO:0033108 mitochondrial respiratory chain complex assembly
## 2 GO:0010257               NADH dehydrogenase complex assembly
## 3 GO:0017004                       cytochrome complex assembly
## 4 GO:0006414                          translational elongation
## 5 GO:0140053                      mitochondrial gene expression
## 7 GO:0009141        nucleoside triphosphate metabolic process
##   normalizedEnrichmentScore pValue          FDR size
## 1                  3.208698      0 0.0000000000   83
## 2                  2.968032      0 0.0000000000   52
## 3                  2.770416      0 0.0000000000   32
## 4                  2.479687      0 0.0000000000  111
## 5                  2.478813      0 0.0000000000  136
## 7                  2.347948      0 0.0004560144  222
```

```
tail(subset(GO_bp_05, select=c("geneSet","description","normalizedEnrichmentScore","pValue","FDR","size
```

```
##       geneSet                                               description
## 20 GO:0014074                    response to purine-containing compound
## 12 GO:0098742 cell-cell adhesion via plasma-membrane adhesion molecules
## 11 GO:0007606                   sensory perception of chemical stimulus
## 13 GO:0035249                      synaptic transmission, glutamatergic
## 10 GO:0051932                          synaptic transmission, GABAergic
## 14 GO:0043062                      extracellular structure organization
##    normalizedEnrichmentScore pValue         FDR size
## 20                 -1.983351      0 0.025626879   97
## 12                 -2.097999      0 0.004911818  169
## 11                 -2.120663      0 0.004271146   84
## 13                 -2.121953      0 0.005338933   49
## 10                 -2.149639      0 0.003737253   21
## 14                 -2.159137      0 0.006406720  277
```

GO Cellular Component

```
GO_cc <- suppressWarnings(WebGestaltR(enrichMethod = "GSEA",
                             enrichDatabase = "geneontology_Cellular_Component_noRedundant",
                             interestGene = res05,
                             interestGeneType = "genesymbol",
                             isOutput = FALSE))
```

```
## Loading the functional categories...
## Loading the ID list...
## Performing the enrichment analysis...
## 1000 permutations of score complete...
```

```
nrow(GO_cc)
```

```
## [1] 21
```

```
GO_cc <- GO_cc[order(GO_cc$normalizedEnrichmentScore, decreasing = TRUE),]
head(subset(GO_cc, select=c("geneSet","description","normalizedEnrichmentScore","pValue","FDR","size"))
```

```
##      geneSet                   description normalizedEnrichmentScore pValue FDR
## 1 GO:0098798 mitochondrial protein complex                  3.338596      0   0
## 2 GO:0070469             respiratory chain                  3.191677      0   0
## 3 GO:0005743  mitochondrial inner membrane                  3.084723      0   0
## 4 GO:0030964    NADH dehydrogenase complex                  3.080469      0   0
## 5 GO:0005840                      ribosome                  3.058882      0   0
## 6 GO:0044455   mitochondrial membrane part                  3.006808      0   0
##   size
## 1  219
## 2   74
## 3  337
## 4   41
## 5  193
## 6  177
```

```
tail(subset(GO_cc, select=c("geneSet","description","normalizedEnrichmentScore","pValue","FDR","size"))
```

```
##       geneSet                 description normalizedEnrichmentScore      pValue
## 19 GO:0005844                    polysome                  1.720138 0.008948546
## 21 GO:0030684                 preribosome                  1.641510 0.023310023
## 20 GO:0045177         apical part of cell                 -1.859651 0.000000000
## 18 GO:0016323 basolateral plasma membrane                 -1.951507 0.000000000
## 14 GO:0005581              collagen trimer                 -2.198581 0.000000000
## 15 GO:0031012        extracellular matrix                 -2.238606 0.000000000
##             FDR size
## 19 0.0281917885   57
## 21 0.0499569907   52
## 20 0.0282576159  235
## 18 0.0126594119  129
## 14 0.0009042437   56
## 15 0.0009042437  305
```

```
GO_cc_05 <- suppressWarnings(WebGestaltR(enrichMethod = "GSEA",
                             enrichDatabase = "geneontology_Cellular_Component_noRedundant",
                             interestGene = res05,
                             interestGeneType = "genesymbol",
                             isOutput = FALSE))
```

```
## Loading the functional categories...
## Loading the ID list...
## Performing the enrichment analysis...
## 1000 permutations of score complete...
```

```
nrow(GO_cc_05)
```

```
## [1] 21
```

```
GO_cc_05 <- GO_cc_05[order(GO_cc_05$normalizedEnrichmentScore, decreasing = TRUE),]
head(subset(GO_cc_05, select=c("geneSet","description","normalizedEnrichmentScore","pValue","FDR","size
```

```
##      geneSet                  description normalizedEnrichmentScore pValue FDR
## 1 GO:0098798 mitochondrial protein complex                  3.391830      0   0
## 2 GO:0070469            respiratory chain                  3.207390      0   0
## 3 GO:0030964    NADH dehydrogenase complex                  3.126552      0   0
## 4 GO:0005743  mitochondrial inner membrane                  3.103100      0   0
## 5 GO:0005840                     ribosome                  3.047249      0   0
## 6 GO:0044455   mitochondrial membrane part                  3.017820      0   0
##   size
## 1  219
## 2   74
## 3   41
## 4  337
## 5  193
## 6  177
```

```
tail(subset(GO_cc_05, select=c("geneSet","description","normalizedEnrichmentScore","pValue","FDR","size
```

```
##       geneSet                 description normalizedEnrichmentScore      pValue
## 20 GO:0005844                    polysome                  1.692982 0.011709602
## 21 GO:0030684                 preribosome                  1.648616 0.008988764
## 19 GO:0045177         apical part of cell                 -1.848681 0.000000000
## 18 GO:0016323 basolateral plasma membrane                 -1.931832 0.000000000
## 12 GO:0031012        extracellular matrix                 -2.244989 0.000000000
## 13 GO:0005581             collagen trimer                 -2.282803 0.000000000
##           FDR size
## 20 0.03809260   57
## 21 0.04950971   52
## 19 0.02865283  235
## 18 0.01576664  129
## 12 0.00000000  305
## 13 0.00000000   56
```

Cell types

Let's try a custom set for cell types, as curated by the folks at http://www.gsea-msigdb.org/

```
C8 <- suppressWarnings(WebGestaltR(enrichMethod = "GSEA",
                       enrichDatabaseFile = "GSEA_custom_sets/c8.all.v7.5.1.symbols.gmt",
                       enrichDatabaseType = "genesymbol",
                       interestGene = res1,
                       interestGeneType = "genesymbol",
                       isOutput = FALSE))
```

```
## Loading the functional categories...
## Loading the ID list...
## Performing the enrichment analysis...
## 1000 permutations of score complete...
```

```
nrow(C8)
```

```
## [1] 175
```

```
C8 <- C8[order(C8$normalizedEnrichmentScore, decreasing = TRUE),]
head(subset(C8, select=c("geneSet","normalizedEnrichmentScore","pValue","FDR","size")))
```

```
##                                                geneSet
## 1                  DESCARTES_FETAL_LUNG_LYMPHOID_CELLS
## 2              DESCARTES_FETAL_PANCREAS_LYMPHOID_CELLS
## 3                 DESCARTES_FETAL_HEART_LYMPHOID_CELLS
## 4 FAN_OVARY_CL10_PUTATIVE_EARLY_ATRESIA_GRANULOSA_CELL
## 5               DESCARTES_FETAL_ADRENAL_LYMPHOID_CELLS
## 6             DESCARTES_FETAL_INTESTINE_LYMPHOID_CELLS
##   normalizedEnrichmentScore pValue FDR size
## 1                  3.614286      0   0  115
## 2                  3.588763      0   0  108
## 3                  3.527898      0   0   81
## 4                  3.517152      0   0  233
## 5                  3.517135      0   0  116
## 6                  3.433727      0   0  119
```

```
tail(subset(C8, select=c("geneSet","normalizedEnrichmentScore","pValue","FDR","size")))
```

```
##                                                geneSet
## 89      DESCARTES_FETAL_EYE_VASCULAR_ENDOTHELIAL_CELLS
## 81          DESCARTES_FETAL_MUSCLE_SMOOTH_MUSCLE_CELLS
## 75                          AIZARANI_LIVER_C13_LSECS_2
## 64           GAO_LARGE_INTESTINE_ADULT_CJ_IMMUNE_CELLS
## 65 DESCARTES_FETAL_PLACENTA_VASCULAR_ENDOTHELIAL_CELLS
## 66                          AIZARANI_LIVER_C10_MVECS_1
##    normalizedEnrichmentScore pValue          FDR size
## 89                 -2.214887      0 0.0005080715   72
## 81                 -2.235256      0 0.0003048429   50
## 75                 -2.275383      0 0.0001905268  212
## 64                 -2.350739      0 0.0000000000  356
## 65                 -2.379345      0 0.0000000000   75
## 66                 -2.411041      0 0.0000000000  209
```

```
C8_05 <- suppressWarnings(WebGestaltR(enrichMethod = "GSEA",
                       enrichDatabaseFile = "GSEA_custom_sets/c8.all.v7.5.1.symbols.gmt",
                       enrichDatabaseType = "genesymbol",
                       interestGene = res05,
                       interestGeneType = "genesymbol",
                       isOutput = FALSE))
```

```
## Loading the functional categories...
## Loading the ID list...
## Performing the enrichment analysis...
## 1000 permutations of score complete...
```

```
nrow(C8_05)
```

```
## [1] 179
```

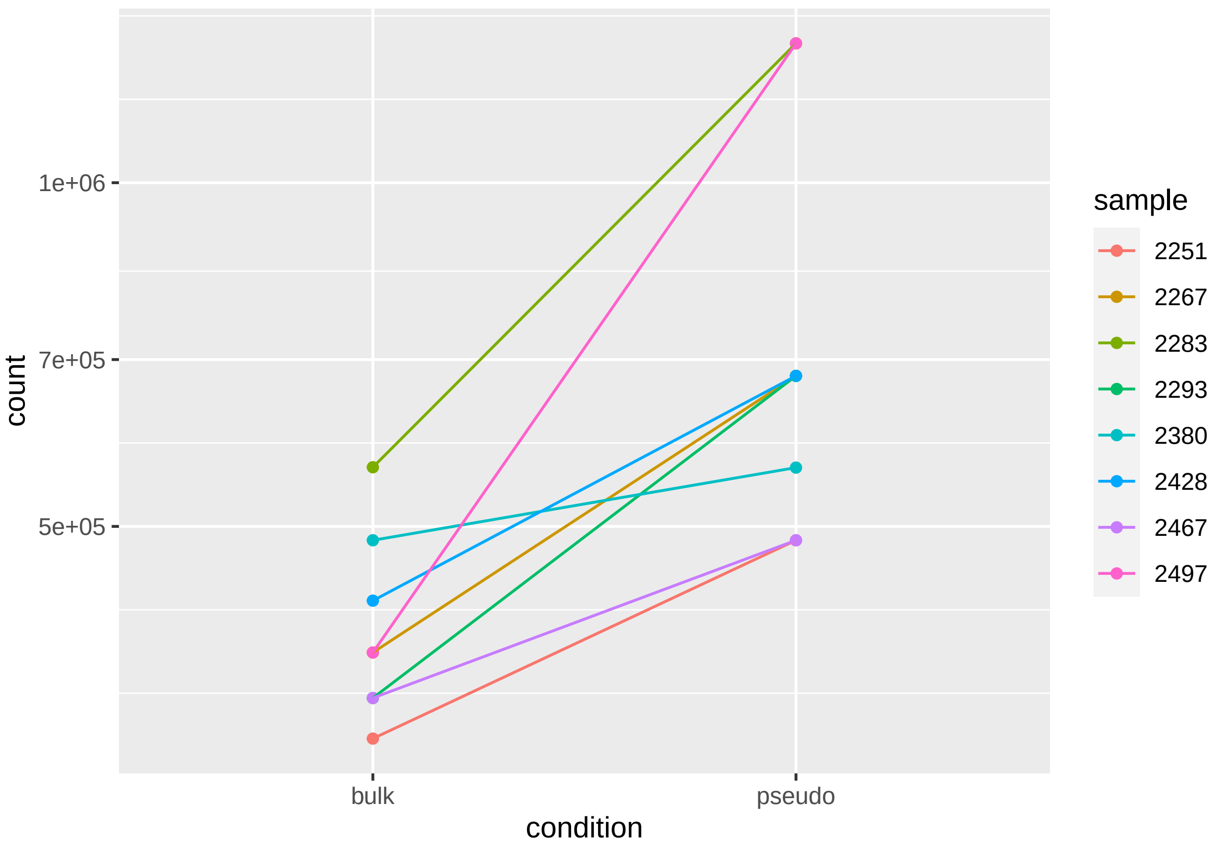
```
C8_05 <- C8_05[order(C8_05$normalizedEnrichmentScore, decreasing = TRUE),]
head(subset(C8_05, select=c("geneSet","normalizedEnrichmentScore","pValue","FDR","size")))
```

```
##                                                geneSet
## 1                  DESCARTES_FETAL_LUNG_LYMPHOID_CELLS
## 2                 DESCARTES_FETAL_HEART_LYMPHOID_CELLS
## 3               DESCARTES_FETAL_ADRENAL_LYMPHOID_CELLS
## 4 FAN_OVARY_CL10_PUTATIVE_EARLY_ATRESIA_GRANULOSA_CELL
## 5              DESCARTES_FETAL_PANCREAS_LYMPHOID_CELLS
## 6             DESCARTES_FETAL_INTESTINE_LYMPHOID_CELLS
##   normalizedEnrichmentScore pValue FDR size
## 1                  3.617819      0   0  109
## 2                  3.560492      0   0   77
## 3                  3.545725      0   0  108
## 4                  3.529760      0   0  227
## 5                  3.510799      0   0  104
## 6                  3.472424      0   0  115
```

```
tail(subset(C8_05, select=c("geneSet","normalizedEnrichmentScore","pValue","FDR","size")))
```

```
##                                                 geneSet
## 64          DESCARTES_FETAL_MUSCLE_SMOOTH_MUSCLE_CELLS
## 65    DESCARTES_FETAL_LIVER_VASCULAR_ENDOTHELIAL_CELLS
## 66                          AIZARANI_LIVER_C13_LSECS_2
## 67           GAO_LARGE_INTESTINE_ADULT_CJ_IMMUNE_CELLS
## 68 DESCARTES_FETAL_PLACENTA_VASCULAR_ENDOTHELIAL_CELLS
## 69                          AIZARANI_LIVER_C10_MVECS_1
##    normalizedEnrichmentScore pValue FDR size
## 64                 -2.277272      0   0   46
## 65                 -2.302453      0   0   85
## 66                 -2.356499      0   0  196
## 67                 -2.410707      0   0  338
## 68                 -2.423104      0   0   71
## 69                 -2.434836      0   0  202
```

```
gene <- "MT-ATP6"
d <- plotCounts(dds, gene=gene, intgroup=c("condition","sample"), returnData=TRUE)
ggplot(d, aes(x=condition, y=count, group=sample, color=sample)) +
  geom_point() + scale_y_log10() + geom_line()
```



Conclusions