GSEA Chunk vs dissociated

Ariel Hippen

2022-07-28

Contents

Analyzing differential expression results

We have calculated differential expression genes in DE_chunk_vs_dissociated.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(rtracklayer)
  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, "chunk_vs_dissociated_data.rds", sep = "/"))

# Load the DESeqResults object with differentially expressed genes, at FDR 0.1 and 0.05
res1 <- readRDS(paste(deseq_path, "chunk_vs_dissociated_FDR_0.1.rds", sep = "/"))
res05 <- readRDS(paste(deseq_path, "chunk_vs_dissociated_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
## HBB        2264.208007      -5.274949 0.5652136 -9.332664 1.032427e-20
## HBA2        673.751104      -4.757030 0.6046143 -7.867875 3.607153e-15
## CXCR1        81.309358      -4.189585 0.8399832 -4.987701 6.110198e-07
## HBA1        181.662594      -3.916449 0.8980385 -4.361115 1.294011e-05
## AC008083.2    2.076658      -3.586494 1.0980669 -3.266189 1.090056e-03
## ADIPOQ     1911.854070      -3.576924 1.5643494 -2.286525 2.222357e-02
## LINC01355   346.657721      -3.542691 0.3813332 -9.290276 1.538885e-20
## LEP         659.431225      -3.501804 0.8992838 -3.893993 9.860768e-05
## AP003031.1    3.682507      -3.423470 1.1016482 -3.107589 1.886204e-03
## AL512303.1    2.164243      -3.361063 1.0989848 -3.058334 2.225711e-03
## PLIN1      4600.183400      -3.360219 0.9704688 -3.462470 5.352406e-04
## AC009019.1    2.931348      -3.344863 1.0812273 -3.093580 1.977572e-03
## CIDEC       886.538783      -3.343077 0.9366554 -3.569164 3.581216e-04
## AC007991.4    5.651049      -3.250859 1.2545567 -2.591241 9.563052e-03
## AC069023.1    6.893661      -3.150502 0.8467974 -3.720491 1.988355e-04
## AL034397.3  213.851165      -3.116796 0.5526555 -5.639673 1.703737e-08
## TEX48         2.977500      -3.111579 0.9807365 -3.172696 1.510306e-03
## FCGR3B      193.803406      -2.994792 0.5669353 -5.282423 1.274861e-07
## AL109936.6   41.516596      -2.952083 0.3777630 -7.814642 5.511941e-15
## AC091117.2    4.983689      -2.933680 0.8210291 -3.573175 3.526791e-04
##                    padj
## HBB        9.875591e-18
## HBA2       1.107237e-12
## CXCR1      2.141483e-05
## HBA1       2.709565e-04
## AC008083.2 9.179539e-03
## ADIPOQ     8.694829e-02
## LINC01355  1.386635e-17
## LEP        1.399598e-03
## AP003031.1 1.390167e-02
## AL512303.1 1.574343e-02
## PLIN1      5.274116e-03
## AC009019.1 1.442662e-02
## CIDEC      3.829733e-03
## AC007991.4 4.705386e-02
## AC069023.1 2.419020e-03
## AL034397.3 9.752147e-07
## TEX48      1.172056e-02
## FCGR3B     5.661478e-06
## AL109936.6 1.655538e-12
## AC091117.2 3.786082e-03
```

So these are more upregulated in the chunks than in the dissociated cells. HBB, HBA1, and HBA2 are all hemoglobin genes. CXCR1 is a chemokine receptor. Several of these genes are secreted by fat cells/adipocytes

(ADIPOQ, LEP, PLIN1, CIDEC). There's a gene (TEX48) that's very lowly expressed but is apparently only expressed in testes????? Is that cancer being funky or a sign of contamination? And then FCCGR3B is a component of IgG.

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

```
##                baseMean log2FoldChange     lfcSE      stat       pvalue
## AC082651.3     3.298485       4.014912 1.2315989  3.259918 1.114443e-03
## AL450322.1    11.335458       4.016885 0.8575780  4.683988 2.813470e-06
## AL031432.1     2.978587       4.030684 1.5135530  2.663061 7.743333e-03
## TRIM72        51.103458       4.034666 0.5397907  7.474502 7.749675e-14
## IL1B        5013.558651       4.040173 0.5288699  7.639256 2.184798e-14
## AL512603.2    96.416357       4.153359 0.5912721  7.024447 2.149162e-12
## SELE        1265.716907       4.235680 1.0213159  4.147277 3.364535e-05
## AC103591.3   174.459490       4.247204 0.4036049 10.523172 6.756033e-26
## AP002008.3    23.015276       4.275042 0.6825899  6.262973 3.777057e-10
## AC084262.1    15.831221       4.282993 0.6963939  6.150245 7.736331e-10
## HSPB3         13.932159       4.284497 0.8446749  5.072363 3.929069e-07
## ASTL         483.145817       4.295618 0.4594194  9.350101 8.756428e-21
## LINC02404      8.229731       4.313546 0.7646813  5.640972 1.690925e-08
## CSF3         840.783269       4.324277 1.1762588  3.676297 2.366439e-04
## EREG        1495.336882       4.372911 0.5682361  7.695587 1.408457e-14
## AL450322.2    40.149822       4.380950 0.5388039  8.130880 4.261850e-16
## OVOL1-AS1     28.077768       4.405757 0.5093041  8.650543 5.125512e-18
## AGXT          42.060301       4.605439 0.7499662  6.140863 8.207455e-10
## PRR35          5.705105       4.920525 1.0650368  4.620052 3.836435e-06
## PYDC1         16.081991       5.920640 1.0858810  5.452384 4.969895e-08
##                    padj
## AC082651.3 9.314705e-03
## AL450322.1 7.652256e-05
## AL031432.1 4.034575e-02
## TRIM72     1.915679e-11
## IL1B       5.868073e-12
## AL512603.2 3.660520e-10
## SELE       5.918690e-04
## AC103591.3 1.572635e-22
## AP002008.3 3.552342e-08
## AC084262.1 6.568356e-08
## HSPB3      1.493207e-05
## ASTL       8.735476e-18
## LINC02404  9.718646e-07
## CSF3       2.777384e-03
## EREG       3.934243e-12
## AL450322.2 1.653420e-13
## OVOL1-AS1  2.753287e-15
## AGXT       6.884650e-08
## PRR35      9.804939e-05
## PYDC1      2.510381e-06
```

These are more upregulated in the dissociated cells. The top ones are a mess. AGXT is supposedly only expressed in the liver, OVOL1-AS1 is a lncRNA. EREG is "a member of the epidermal growth factor (EGF) family of proteins. The encoded protein may be involved in a wide range of biological processes including inflammation, wound healing, oocyte maturation, and cell proliferation." https://www.genecards.org/cgi-

bin/carddisp.pl?gene=EREG CSF3 "is produced by endothelium, macrophages, and a number of other immune cells." https://en.wikipedia.org/wiki/Granulocyte_colony-stimulating_factor HSPB3 "This gene encodes a muscle-specific small heat shock protein." https://www.ncbi.nlm.nih.gov/gene/8988 SELE "The protein encoded by this gene is found in cytokine-stimulated endothelial cells" https://www.genecards.org/cgi-bin/carddisp.pl?gene=SELE

Let's try again but with a filter to only genes that are decently expressed.

```
res1high <- subset(res1, res1$baseMean > 100)
as.data.frame(head(res1high, n=20))
```

```
##              baseMean log2FoldChange     lfcSE      stat       pvalue
## HBB        2264.2080      -5.274949 0.5652136 -9.332664 1.032427e-20
## HBA2        673.7511      -4.757030 0.6046143 -7.867875 3.607153e-15
## HBA1        181.6626      -3.916449 0.8980385 -4.361115 1.294011e-05
## ADIPOQ     1911.8541      -3.576924 1.5643494 -2.286525 2.222357e-02
## LINC01355   346.6577      -3.542691 0.3813332 -9.290276 1.538885e-20
## LEP         659.4312      -3.501804 0.8992838 -3.893993 9.860768e-05
## PLIN1      4600.1834      -3.360219 0.9704688 -3.462470 5.352406e-04
## CIDEC       886.5388      -3.343077 0.9366554 -3.569164 3.581216e-04
## AL034397.3  213.8512      -3.116796 0.5526555 -5.639673 1.703737e-08
## FCGR3B      193.8034      -2.994792 0.5669353 -5.282423 1.274861e-07
## CXCR2       123.2648      -2.902115 0.6040217 -4.804654 1.550197e-06
## GPD1       2110.0034      -2.864240 0.7833736 -3.656288 2.558936e-04
## THRSP       111.3514      -2.806047 0.5185531 -5.411301 6.256841e-08
## ACVR1C      402.7838      -2.561761 0.6224156 -4.115836 3.857782e-05
## SLC7A10     173.5175      -2.541003 0.6983780 -3.638436 2.742989e-04
## DGAT2       657.6586      -2.379434 0.5416326 -4.393078 1.117572e-05
## SLC19A3     783.9477      -2.147723 0.4350801 -4.936385 7.958396e-07
## NAV2-AS5    107.3732      -2.112261 0.4127966 -5.116954 3.105091e-07
## EDDM13      118.7541      -2.098489 0.3144391 -6.673754 2.493419e-11
## PRH2        118.4492      -2.089533 0.3042387 -6.868073 6.507514e-12
##                    padj
## HBB        9.875591e-18
## HBA2       1.107237e-12
## HBA1       2.709565e-04
## ADIPOQ     8.694829e-02
## LINC01355  1.386635e-17
## LEP        1.399598e-03
## PLIN1      5.274116e-03
## CIDEC      3.829733e-03
## AL034397.3 9.752147e-07
## FCGR3B     5.661478e-06
## CXCR2      4.611948e-05
## GPD1       2.950040e-03
## THRSP      3.066181e-06
## ACVR1C     6.530874e-04
## SLC7A10    3.104534e-03
## DGAT2      2.403167e-04
## SLC19A3    2.630791e-05
## NAV2-AS5   1.211376e-05
## EDDM13     3.180305e-09
## PRH2       1.004278e-09
```

Okay, a couple more adipose tissue things (THRSP, DGAT2).

```
as.data.frame(tail(res1high, n=20))
```

```
##              baseMean log2FoldChange     lfcSE      stat       pvalue
## AC015912.3  115.7652       3.441191 0.5106750  6.738514 1.600142e-11
## TEX14       506.7056       3.458933 0.4128350  8.378487 5.361261e-17
## AL645608.7  251.0051       3.561417 0.3339811 10.663530 1.507648e-26
## ARC         468.9911       3.582926 0.6197149  5.781572 7.400601e-09
## PRSS22      847.8038       3.592108 0.4997242  7.188182 6.565986e-13
## KRT16      1301.5981       3.604455 0.6626794  5.439214 5.351601e-08
## NLRP3      4526.4309       3.719893 0.4310242  8.630359 6.115979e-18
## PEAK3       247.3214       3.792340 0.3163026 11.989595 4.028711e-33
## AC253572.2 2103.7996       3.799999 0.3781945 10.047737 9.400152e-24
## CCL3L1     1773.0335       3.852048 0.5126728  7.513659 5.749744e-14
## ADRA2B      329.1694       3.857407 0.3385773 11.392986 4.531907e-30
## PMAIP1     1407.2301       3.904967 0.4278432  9.127098 7.036027e-20
## KRT17      7909.5372       3.956979 0.5389221  7.342395 2.098052e-13
## CXCL8      9524.6068       4.014204 0.5531115  7.257495 3.943256e-13
## IL1B       5013.5587       4.040173 0.5288699  7.639256 2.184798e-14
## SELE       1265.7169       4.235680 1.0213159  4.147277 3.364535e-05
## AC103591.3  174.4595       4.247204 0.4036049 10.523172 6.756033e-26
## ASTL        483.1458       4.295618 0.4594194  9.350101 8.756428e-21
## CSF3        840.7833       4.324277 1.1762588  3.676297 2.366439e-04
## EREG       1495.3369       4.372911 0.5682361  7.695587 1.408457e-14
##                    padj
## AC015912.3 2.180330e-09
## TEX14      2.303940e-14
## AL645608.7 4.375544e-23
## ARC        4.796310e-07
## PRSS22     1.300764e-10
## KRT16      2.683775e-06
## NLRP3      3.106139e-15
## PEAK3      1.875567e-29
## AC253572.2 1.544556e-20
## CCL3L1     1.466380e-11
## ADRA2B     1.808425e-26
## PMAIP1     5.615352e-17
## KRT17      4.803680e-11
## CXCL8      8.538524e-11
## IL1B       5.868073e-12
## SELE       5.918690e-04
## AC103591.3 1.572635e-22
## ASTL       8.735476e-18
## CSF3       2.777384e-03
## EREG       3.934243e-12
```

Oh boy, another testis-expressed gene (TEX14). Oooh upregulated keratin (KRT16, KRT17). And we've got a brain-specific protease (PRSS22). PEAK3 is "Involved in regulation of actin cytoskeleton organization." https://www.genecards.org/cgi-bin/carddisp.pl?gene=PEAK3 NLRP3 is "NLRP3 is expressed predominantly in macrophages and as a component of the inflammasome." https://en.wikipedia.org/wiki/NLRP3 ARC is "Activity-regulated cytoskeleton-associated protein" and apparently really important for learning? https://en.wikipedia.org/wiki/Activity-regulated_cytoskeleton-associated_protein A couple more chemokines (CCL3L1, CXCL8) Something involved in apoptosis (PMAIP1), https://www.genecards.org/cgi-bin/carddisp.pl?gene=PMAIP1

Volcano plot

We'll now filter down to only protein-coding genes, using info from the gtf file downloaded from the Cellranger website.

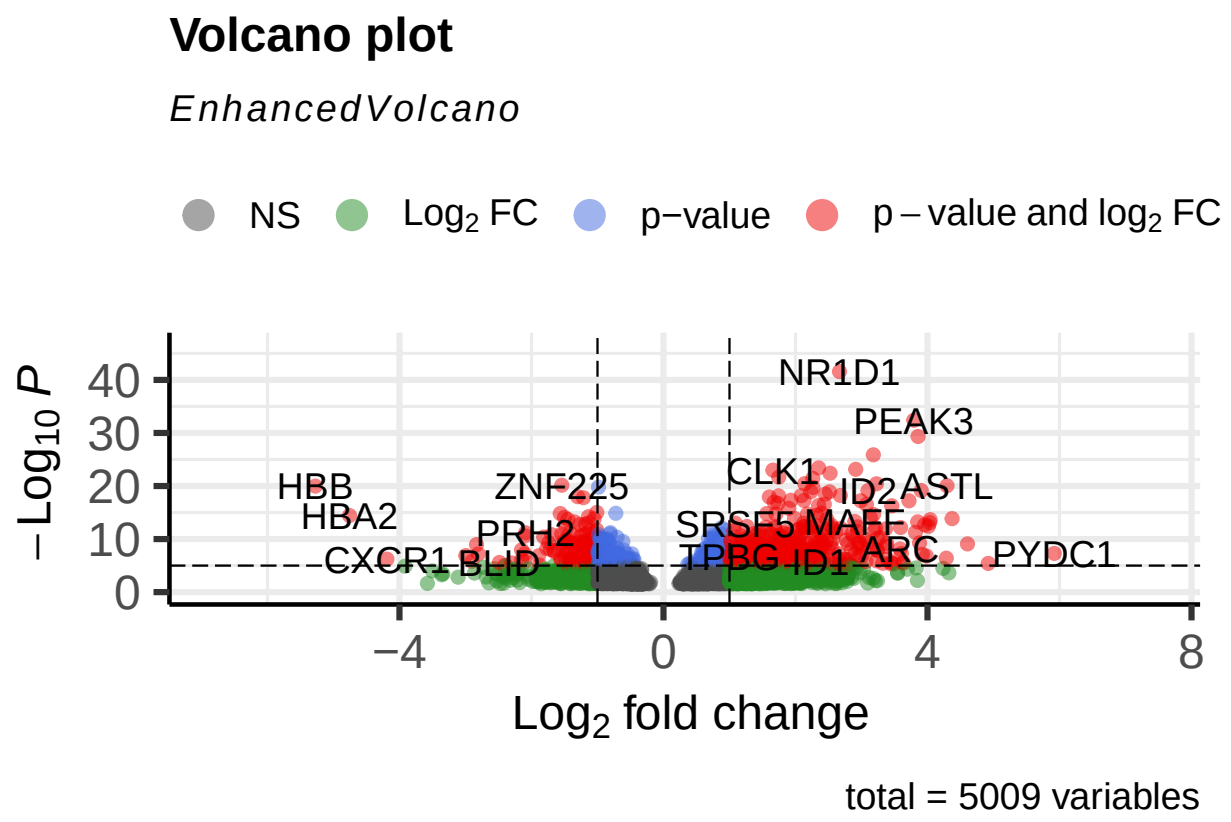
```
genefile <- paste(data_path,"index/refdata-gex-GRCh38-2020-A/genes/genes.gtf",sep = "/")
gff <- readGFF(genefile)
protein_coding <- subset(gff, gff$gene_type=="protein_coding")

res1 <- subset(res1, rownames(res1) %in% protein_coding$gene_name)
```

```
library(EnhancedVolcano)
```

```
## Loading required package: ggrepel
```

```
EnhancedVolcano(res1, lab = rownames(res1), x = 'log2FoldChange', y = 'pvalue')
```



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] 5009
```

```
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] 5836
```

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] 291
```

```
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] 261
```

```
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 normalizedEnrichmentScore
## 1 GO:0043062     extracellular structure organization                   2.558967
## 2 GO:0048705            skeletal system morphogenesis                   2.534183
## 3 GO:0045165                     cell fate commitment                   2.523143
## 4 GO:0007389            pattern specification process                   2.443704
## 5 GO:0002237 response to molecule of bacterial origin                   2.428714
## 6 GO:0033002                muscle cell proliferation                   2.424678
##   pValue FDR size
## 1      0   0  109
## 2      0   0   63
## 3      0   0   66
## 4      0   0   93
## 5      0   0   99
## 6      0   0   56
```

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

```
##       geneSet                                          description
## 43 GO:0033865           nucleoside bisphosphate metabolic process
## 52 GO:0140053                        mitochondrial gene expression
## 8  GO:0009593                       detection of chemical stimulus
## 9  GO:0050906 detection of stimulus involved in sensory perception
## 10 GO:0016999                         antibiotic metabolic process
## 11 GO:0006399                               tRNA metabolic process
##    normalizedEnrichmentScore pValue          FDR size
## 43                -2.575540      0 0.0002897701   22
## 52                -2.594618      0 0.0003477242   41
## 8                 -2.655345      0 0.0000000000   33
## 9                 -2.662161      0 0.0000000000   40
## 10                -2.695368      0 0.0000000000   22
## 11                -2.703201      0 0.0000000000   43
```

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] 339
```

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

```
##                                                geneSet normalizedEnrichmentScore
## 1                 CUI_DEVELOPING_HEART_C5_VALVAR_CELL                   2.923429
## 2                          AIZARANI_LIVER_C29_MVECS_2                   2.903700
## 3           GAO_LARGE_INTESTINE_ADULT_CJ_IMMUNE_CELLS                   2.885018
## 4         TRAVAGLINI_LUNG_ADVENTITIAL_FIBROBLAST_CELL                   2.885015
## 5        CUI_DEVELOPING_HEART_VALVAR_ENDOTHELIAL_CELL                   2.859748
## 6 DESCARTES_FETAL_CEREBRUM_VASCULAR_ENDOTHELIAL_CELLS                   2.833196
##   pValue FDR size
## 1      0   0  102
## 2      0   0  163
## 3      0   0  214
## 4      0   0  109
## 5      0   0   81
## 6      0   0  237
```

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

```
##                                             geneSet normalizedEnrichmentScore
## 174              DESCARTES_MAIN_FETAL_ERYTHROBLASTS                  -2.355485
## 161 DESCARTES_FETAL_PLACENTA_AFP_ALB_POSITIVE_CELLS                  -2.403167
## 170            DESCARTES_FETAL_MUSCLE_ERYTHROBLASTS                  -2.419611
## 103           DESCARTES_FETAL_STOMACH_ERYTHROBLASTS                  -2.498829
## 104         DESCARTES_FETAL_INTESTINE_ERYTHROBLASTS                  -2.504469
## 105           DESCARTES_FETAL_ADRENAL_ERYTHROBLASTS                  -2.620851
##     pValue          FDR size
## 174      0 0.0004382483   52
## 161      0 0.0002629490   45
## 170      0 0.0003286862   30
## 103      0 0.0000000000   23
## 104      0 0.0000000000   44
## 105      0 0.0000000000   25
```

```r
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...
```

```r
nrow(C8_05)
```

```
## [1] 308
```

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

```
##                                             geneSet normalizedEnrichmentScore
## 1                 CUI_DEVELOPING_HEART_C5_VALVAR_CELL                  2.924764
## 2                         AIZARANI_LIVER_C29_MVECS_2                  2.895396
## 3           GAO_LARGE_INTESTINE_ADULT_CJ_IMMUNE_CELLS                  2.828090
## 4                         AIZARANI_LIVER_C10_MVECS_1                  2.819640
## 5 DESCARTES_FETAL_CEREBRUM_VASCULAR_ENDOTHELIAL_CELLS                  2.811425
## 6                            HAY_BONE_MARROW_STROMAL                  2.804545
##   pValue FDR size
## 1      0   0   90
## 2      0   0  131
## 3      0   0  179
## 4      0   0  105
## 5      0   0  199
## 6      0   0  207
```

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

```
##                                          geneSet normalizedEnrichmentScore
## 181              DESCARTES_MAIN_FETAL_ERYTHROBLASTS                 -2.307258
## 168 DESCARTES_FETAL_PLACENTA_AFP_ALB_POSITIVE_CELLS                 -2.368194
## 169            DESCARTES_FETAL_MUSCLE_ERYTHROBLASTS                 -2.400061
## 179         DESCARTES_FETAL_INTESTINE_ERYTHROBLASTS                 -2.400364
## 182           DESCARTES_FETAL_STOMACH_ERYTHROBLASTS                 -2.441366
## 142           DESCARTES_FETAL_ADRENAL_ERYTHROBLASTS                 -2.722040
##     pValue          FDR size
## 181      0 0.0018179226   38
## 168      0 0.0012119484   29
## 169      0 0.0012119484   24
## 179      0 0.0016159312   34
## 182      0 0.0018179226   17
## 142      0 0.0006059742   16
```

Conclusions

The main things being lost in dissociation are red blood cells and adipose tissue. The latter makes sense, because it's documented that it's very hard to get adipose tissue to dissociate cleanly, and as such it's much more prevalent in single-nucleus RNA-seq than single cell. Still need to figure out the reason for the loss of red blood cells in dissociation.

The dissociated samples seem to have an enrichment of fibroblasts, endothelial cells, and immune cells. I imagine this means they're being sequenced to a higher depth since they're not competing with adipose tissue.

Still not totally sure what this means for the cancer cells. Probably more analysis is needed there.