WormBase_Tissue_Specific_Genes

Robert Williams

2/20/2023

Install Libraries

```
# install.packages("tidyverse")
# if (!require("BiocManager", quietly = TRUE))
#     install.packages("BiocManager")
# BiocManager::install(version = "3.14")
# BiocManager::install("InterMineR")
```

Load Libraries

```
library(tidyverse)
library(InterMineR)
library(ComplexHeatmap)
library(biomaRt)
```

load genes lists from WormBase

```
tissues <- c("intestine", "pharyngeal-intestinal-valve", "rectum", "coelomic-system", "reproductive-sys
tissue_paths <- c("../01_input/genes_direct_and_inferred_for_WBbt_0005772_intestine.txt",
                  "../01_input/genes_direct_and_inferred_for_WBbt_0005767_pharyngeal-intestinal-valve.t
                  "../01_input/genes_direct_and_inferred_for_WBbt_0005773_rectum.txt",
                  "../01_input/genes_direct_and_inferred_for_WBbt_0005749_coelomic-system.txt",
                  "../01_input/genes_direct_and_inferred_for_WBbt_0005747_reproductive-system.txt",
                  "../01_input/genes_direct_and_inferred_for_WBbt_0005736_excretory-system.txt",
                  "../01_input/genes_direct_and_inferred_for_WBbt_0005735_nervous-system.txt",
                  "../01_input/genes_direct_and_inferred_for_WBbt_0005730_epithelial-system.txt",
                  "../01_input/genes_direct_and_inferred_for_WBbt_0005737_muscular-system.txt"
                  )
gene_tissue_annotations <- data.frame()
for(i in 1:length(tissues)){
  # print(i)
  # print(tissues[i])
  # print(tissue_paths[i])

  gene_tissue_annotations <- data.frame(read_tsv(file = tissue_paths[i],
                        c("WBGeneID", "Sequence.name", "Species"),
                        show_col_types = FALSE
                        ),
                        tissue = tissues[i]) %>%
    bind_rows(gene_tissue_annotations)
}
table(gene_tissue_annotations$tissue)
```

```
##
##             coelomic-system           epithelial-system
##                        1002                        5272
##            excretory-system                   intestine
##                        1097                        7094
##             muscular-system              nervous-system
##                        7100                       14057
## pharyngeal-intestinal-valve                      rectum
##                         327                         711
##         reproductive-system
##                        8773
```

Add ubiquitous genes

```
ub_genes <- read_csv(file = "../01_input/Rechtsteiner_et_al_2010_Table_S2.csv", col_names = "wormbase_g
```

```
## Rows: 2580 Columns: 1
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): wormbase_gseq
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
paramart <- biomaRt::useMart("parasite_mart", dataset = "wbps_gene", host = "https://parasite.wormbase.

ub_genes <- biomaRt::getBM(
  mart = paramart,
  filter = c("wormbase_gseqname"),
  value = ub_genes$wormbase_gseq,
  attributes = c("wbps_gene_id","wormbase_gseq", "wikigene_name"))

ub_genes<- ub_genes %>% rename(WBGeneID = "wbps_gene_id")
```

Remove ubiquitous genes from the tissue specific list and then add on ubiquitous genes

```
gene_tissue_annotations <- gene_tissue_annotations %>% filter(!(WBGeneID %in% ub_genes$WBGeneID)) %>%
  bind_rows(data.frame(WBGeneID = ub_genes[,colnames(ub_genes) == "WBGeneID"], Sequence.name = ub_genes

head(gene_tissue_annotations)
```

```
##         WBGeneID Sequence.name                Species          tissue
## 1 WBGene00001028        dnj-10 Caenorhabditis elegans muscular-system
## 2 WBGene00004802       sir-2.3 Caenorhabditis elegans muscular-system
## 3 WBGene00000584         cog-1 Caenorhabditis elegans muscular-system
## 4 WBGene00006832       unc-105 Caenorhabditis elegans muscular-system
## 5 WBGene00008389       D1086.2 Caenorhabditis elegans muscular-system
## 6 WBGene00012476     Y18D10A.3 Caenorhabditis elegans muscular-system
```

Select for genes with one tissue annottaion

```
tissue_specific_genes <- gene_tissue_annotations %>% group_by(WBGeneID) %>% summarise(sum_tissues = n_d

head(tissue_specific_genes)
```

```
## # A tibble: 6 x 5
##   WBGeneID       sum_tissues Sequence.name Species                 tissue
##   <chr>                <int> <chr>         <chr>                   <chr>
## 1 WBGene00000004           1 aat-3         Caenorhabditis elegans ubiquitous
## 2 WBGene00000007           1 aat-6         Caenorhabditis elegans intestine
## 3 WBGene00000008           1 aat-7         Caenorhabditis elegans intestine
## 4 WBGene00000016           1 abf-5         Caenorhabditis elegans nervous-system
## 5 WBGene00000028           1 abu-5         Caenorhabditis elegans muscular-syst~
## 6 WBGene00000029           1 abu-6         Caenorhabditis elegans muscular-syst~
```

Keep only protein-coding genes

```
transcript_type <- read_csv(file = "../01_input/biomaRt_elegans_transcript_biotype.csv")
```

```
## Rows: 59897 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (4): Gene stable ID, Genome project, Gene name, Transcript biotype
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
colnames(transcript_type) <- c("WBGeneID", "genome_id", "gene_name", "biotype")
transcript_type %>% distinct(WBGeneID, .keep_all = TRUE) %>%group_by(biotype) %>% summarise(n())
```

```
## # A tibble: 13 x 2
##    biotype         `n()`
##    <chr>           <int>
##  1 antisense_RNA     100
##  2 lincRNA           193
##  3 ncRNA            7809
##  4 piRNA           15363
##  5 pre_miRNA         260
##  6 protein_coding  19952
##  7 pseudogene       1916
##  8 rRNA               22
##  9 rRNA_pseudogene     1
## 10 snoRNA            346
## 11 snRNA             129
## 12 tRNA              634
## 13 tRNA_pseudogene   209
```

```
tissue_specific_genes_protein <- tissue_specific_genes %>%
  filter(WBGeneID %in% (transcript_type %>%
                          filter(biotype == "protein_coding") %>%
                          pull(WBGeneID)
                        )
         )
nrow(tissue_specific_genes_protein)
```

```
## [1] 6873
```

export the tissue specific gene dataframe

```
write_csv(tissue_specific_genes_protein %>% dplyr::select(-sum_tissues, -Species), file = "../03_output
```

Session info

```
sessionInfo()
```

```
## R version 4.1.0 (2021-05-18)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS Catalina 10.15.7
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.1/Resources/lib/libRblas.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.1/Resources/lib/libRlapack.dylib
##
## 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      stats     graphics  grDevices utils     datasets  methods
## [8] base
##
## other attached packages:
##  [1] biomaRt_2.48.3       ComplexHeatmap_2.8.0 InterMineR_1.14.1
##  [4] forcats_0.5.1        stringr_1.4.0        dplyr_1.0.8
##  [7] purrr_0.3.4          readr_2.1.2          tidyr_1.2.0
## [10] tibble_3.1.6         ggplot2_3.3.5        tidyverse_1.3.1
##
## loaded via a namespace (and not attached):
##   [1] colorspace_2.0-3            rjson_0.2.21
##   [3] ellipsis_0.3.2              circlize_0.4.14
##   [5] XVector_0.32.0              GenomicRanges_1.44.0
##   [7] GlobalOptions_0.1.2         fs_1.5.2
##   [9] clue_0.3-60                 rstudioapi_0.13
##  [11] bit64_4.0.5                 AnnotationDbi_1.54.1
##  [13] fansi_1.0.3                 lubridate_1.8.0
##  [15] sqldf_0.4-11                xml2_1.3.3
##  [17] codetools_0.2-18            doParallel_1.0.17
##  [19] cachem_1.0.6                knitr_1.38
##  [21] jsonlite_1.8.0              Cairo_1.5-15
##  [23] broom_0.8.0                 cluster_2.1.3
##  [25] dbplyr_2.1.1                png_0.1-7
##  [27] compiler_4.1.0              httr_1.4.2
##  [29] backports_1.4.1             assertthat_0.2.1
##  [31] Matrix_1.4-1                fastmap_1.1.0
##  [33] cli_3.2.0                   htmltools_0.5.2
##  [35] prettyunits_1.1.1           tools_4.1.0
##  [37] igraph_1.3.0                gtable_0.3.0
##  [39] glue_1.6.2                  GenomeInfoDbData_1.2.6
##  [41] rappdirs_0.3.3              Rcpp_1.0.8.3
##  [43] Biobase_2.52.0              cellranger_1.1.0
##  [45] vctrs_0.4.0                 Biostrings_2.60.2
##  [47] RJSONIO_1.3-1.6             iterators_1.0.14
##  [49] xfun_0.30                   proto_1.0.0
##  [51] rvest_1.0.2                 lifecycle_1.0.1
##  [53] XML_3.99-0.9                zlibbioc_1.38.0
##  [55] scales_1.2.0                vroom_1.5.7
##  [57] hms_1.1.1                   MatrixGenerics_1.4.3
##  [59] parallel_4.1.0              SummarizedExperiment_1.22.0
##  [61] RColorBrewer_1.1-3          curl_4.3.2
##  [63] yaml_2.3.5                  memoise_2.0.1
##  [65] stringi_1.7.6               RSQLite_2.2.12
##  [67] S4Vectors_0.30.2            foreach_1.5.2
##  [69] filelock_1.0.2              BiocGenerics_0.38.0
##  [71] shape_1.4.6                 chron_2.3-56
##  [73] GenomeInfoDb_1.28.4         rlang_1.0.2
##  [75] pkgconfig_2.0.3             matrixStats_0.61.0
##  [77] bitops_1.0-7                evaluate_0.15
##  [79] lattice_0.20-45             bit_4.0.4
##  [81] tidyselect_1.1.2            magrittr_2.0.3
##  [83] R6_2.5.1                    IRanges_2.26.0
##  [85] generics_0.1.2              DelayedArray_0.18.0
##  [87] DBI_1.1.2                   gsubfn_0.7
##  [89] pillar_1.7.0                haven_2.4.3
##  [91] withr_2.5.0                 KEGGREST_1.32.0
##  [93] RCurl_1.98-1.6              modelr_0.1.8
##  [95] crayon_1.5.1                utf8_1.2.2
##  [97] BiocFileCache_2.0.0         tzdb_0.3.0
##  [99] rmarkdown_2.13              GetoptLong_1.0.5
## [101] progress_1.2.2              readxl_1.4.0
## [103] blob_1.2.3                  reprex_2.0.1
## [105] digest_0.6.29               stats4_4.1.0
## [107] munsell_0.5.0
```