Protein_Domain_RNA_Localization

Robert Williams

3/10/2021

```
# BiocManager::install("biomaRt")
# install.packages("tidyverse")
```

Load Libraries

```
library(biomaRt)
```

```
## Warning: package 'biomaRt' was built under R version 4.1.1
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

```
## Warning: package 'tidyr' was built under R version 4.1.2
```

```
## Warning: package 'readr' was built under R version 4.1.2
```

```
## Warning: package 'dplyr' was built under R version 4.1.2
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x dplyr::select() masks biomaRt::select()
```

```
library(openxlsx)
library(ComplexHeatmap)
```

```
## Loading required package: grid
```

```
## ========================================
## ComplexHeatmap version 2.8.0
## Bioconductor page: http://bioconductor.org/packages/ComplexHeatmap/
## Github page: https://github.com/jokergoo/ComplexHeatmap
## Documentation: http://jokergoo.github.io/ComplexHeatmap-reference
##
## If you use it in published research, please cite:
## Gu, Z. Complex heatmaps reveal patterns and correlations in multidimensional
##   genomic data. Bioinformatics 2016.
##
## The new InteractiveComplexHeatmap package can directly export static
## complex heatmaps into an interactive Shiny app with zero effort. Have a try!
##
## This message can be suppressed by:
```

```
##   suppressPackageStartupMessages(library(ComplexHeatmap))
## ========================================
```

Download information from WormBase ParaSite BioMart.
Guide is located here: https://parasite.wormbase.org/info/Tools/biomart.html

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

protdomain_df <- getBM(
  mart = paramart,
  filter = c("species_id_1010","biotype"),
  values = list(species_id_1010 = "caelegprjna13758", biotype = "protein_coding"),
  attributes = c("production_name_1010", "wormbase_gseq","wbps_gene_id", "wikigene_name", "interpro_id"
)
head(protdomain_df)
```

```
##                 production_name_1010 wormbase_gseq   wbps_gene_id wikigene_name
## 1 caenorhabditis_elegans_prjna13758    Y110A7A.10 WBGene00000001         aap-1
## 2 caenorhabditis_elegans_prjna13758    Y110A7A.10 WBGene00000001         aap-1
## 3 caenorhabditis_elegans_prjna13758    Y110A7A.10 WBGene00000001         aap-1
## 4 caenorhabditis_elegans_prjna13758    Y110A7A.10 WBGene00000001         aap-1
## 5 caenorhabditis_elegans_prjna13758    Y110A7A.10 WBGene00000001         aap-1
## 6 caenorhabditis_elegans_prjna13758    Y110A7A.10 WBGene00000001         aap-1
##   interpro_id interpro_short_description   interpro_description interpro_start
## 1   IPR036860                 SH2_dom_sf SH2 domain superfamily              8
## 2   IPR036860                 SH2_dom_sf SH2 domain superfamily            333
## 3   IPR036860                 SH2_dom_sf SH2 domain superfamily             14
## 4   IPR036860                 SH2_dom_sf SH2 domain superfamily            358
## 5   IPR000980                        SH2             SH2 domain            360
## 6   IPR000980                        SH2             SH2 domain             20
##   interpro_end
## 1          125
## 2          450
## 3          115
## 4          443
## 5          422
## 6           94
```

```
protdomain_df %>% group_by(interpro_description) %>% count() %>% nrow
```

```
## [1] 8512
```

There are 8533 unique protein domain IDs

These are all the protein domains associated with “erm-1”, “frm-7”, and “imb-2”

```
# protdomain_df %>% filter(wikigene_name == "erm-1")
# protdomain_df %>% filter(wikigene_name == "frm-7")
# protdomain_df %>% filter(interpro_short_description %in% c("PH_domain", "FERM_domain", "PH-like_dom_s

goi_domains <- protdomain_df %>% filter(wikigene_name %in% c("erm-1")) %>% group_by(wikigene_name, inte

goi_domains
```

```
## # A tibble: 15 x 4
## # Groups:   wikigene_name, interpro_short_description, interpro_description
## #   [15]
##    wikigene_name interpro_short_description interpro_description               n
##    <chr>          <chr>                      <chr>                          <int>
##  1 erm-1          Band_41_domain             Band 4.1 domain                   10
##  2 erm-1          ERM                        Ezrin/radixin/moesin               2
##  3 erm-1          ERM_C_dom                  Ezrin/radixin/moesin, C-termi~     2
##  4 erm-1          ERM_FERM_C                 ERM family, FERM domain C-lobe     2
##  5 erm-1          Ez/rad/moesin-like         Ezrin/radixin/moesin-like         16
##  6 erm-1          FERM_2                     FERM superfamily, second doma~     2
##  7 erm-1          FERM_central               FERM central domain                4
##  8 erm-1          FERM_CS                    FERM conserved site                4
##  9 erm-1          FERM_domain                FERM domain                        2
## 10 erm-1          FERM_N                     FERM, N-terminal                   2
## 11 erm-1          FERM_PH-like_C             FERM, C-terminal PH-like doma~     4
## 12 erm-1          FERM/acyl-CoA-bd_prot_sf   FERM/acyl-CoA-binding protein~     2
## 13 erm-1          Moesin_tail_sf             Moesin tail domain superfamily     4
## 14 erm-1          PH-like_dom_sf             PH-like domain superfamily         2
## 15 erm-1          Ubiquitin-like_domsf       Ubiquitin-like domain superfa~     2
```

```
domain_hits <- protdomain_df %>% filter(interpro_short_description %in% goi_domains$interpro_short_desc
  !(interpro_short_description %in% c("Ubiquitin-like_domsf"))
  ) %>%
  select(wbps_gene_id, wikigene_name, interpro_description, interpro_short_description) %>%
  group_by(wbps_gene_id,wikigene_name, interpro_description, interpro_short_description) %>%
  count(name = "domain_count") %>%
  ungroup()
head(domain_hits)
```

```
## # A tibble: 6 x 5
##   wbps_gene_id   wikigene_name interpro_descripti~ interpro_short_~ domain_count
##   <chr>          <chr>         <chr>               <chr>                   <int>
## 1 WBGene00000102 akt-1         PH-like domain sup~ PH-like_dom_sf              2
## 2 WBGene00000103 akt-2         PH-like domain sup~ PH-like_dom_sf              1
## 3 WBGene00000149 apl-1         PH-like domain sup~ PH-like_dom_sf              2
## 4 WBGene00000420 ced-6         PH-like domain sup~ PH-like_dom_sf              1
## 5 WBGene00000426 ced-12        PH-like domain sup~ PH-like_dom_sf              2
## 6 WBGene00000564 cnk-1         PH-like domain sup~ PH-like_dom_sf              1
```

```
length(unique(domain_hits$wbps_gene_id))
```

```
## [1] 144
```

```
domain_hits_totals <- protdomain_df %>% filter(interpro_short_description %in% goi_domains$interpro_shor
  !(interpro_short_description %in% c("Ubiquitin-like_domsf"))
  ) %>%  group_by(interpro_short_description, interpro_description) %>% count(name = "domain_count") %>
domain_hits_totals
```

```
## # A tibble: 14 x 3
##    interpro_short_description interpro_description                   domain_count
##    <chr>                      <chr>                                         <int>
##  1 PH-like_dom_sf             PH-like domain superfamily                      282
##  2 Band_41_domain             Band 4.1 domain                                 101
##  3 FERM_central               FERM central domain                              67
##  4 Ez/rad/moesin-like         Ezrin/radixin/moesin-like                        47
##  5 FERM_domain                FERM domain                                      42
##  6 FERM_PH-like_C             FERM, C-terminal PH-like domain                  40
##  7 FERM_2                     FERM superfamily, second domain                  37
##  8 FERM/acyl-CoA-bd_prot_sf   FERM/acyl-CoA-binding protein superf~            36
##  9 FERM_CS                    FERM conserved site                              15
## 10 FERM_N                     FERM, N-terminal                                 15
## 11 Moesin_tail_sf             Moesin tail domain superfamily                   10
## 12 ERM_C_dom                  Ezrin/radixin/moesin, C-terminal                  6
## 13 ERM_FERM_C                 ERM family, FERM domain C-lobe                    3
## 14 ERM                        Ezrin/radixin/moesin                              2
```

```
present_sub <- read.xlsx(xlsxFile ="S1_Dataset_AB_P1_Transcriptome.xlsx",
          sheet = "present_subset",
          startRow = 2) %>% select(WBID)
AB_enr_sub <- read.xlsx(xlsxFile ="S1_Dataset_AB_P1_Transcriptome.xlsx",
          sheet = "AB-enriched_subset",
          startRow = 2) %>% select(WBID)
P1_enr_sub <- read.xlsx(xlsxFile ="S1_Dataset_AB_P1_Transcriptome.xlsx",
          sheet = "P1-enriched_subset",
          startRow = 2) %>% select(WBID)
symm_sub <- read.xlsx(xlsxFile ="S1_Dataset_AB_P1_Transcriptome.xlsx",
          sheet = "symm_subset",
          startRow = 2) %>% select(WBID)
c(nrow(present_sub), nrow(AB_enr_sub), nrow(P1_enr_sub), nrow(symm_sub))
```

```
## [1] 7945   80  201 3974
```

Add true/false for different AB/P1 category

```
twocell_domains <- domain_hits %>%
  mutate(present = case_when(wbps_gene_id %in% present_sub$WBID == TRUE ~ TRUE,
                                           wbps_gene_id %in% present_sub$WBID == FALSE ~ FALSE),
                       AB_enriched = case_when(wbps_gene_id %in% AB_enr_sub$WBID == TRUE ~ TRUE,
                                           wbps_gene_id %in% AB_enr_sub$WBID == FALSE ~ FALSE),
                       P1_enriched = case_when(wbps_gene_id %in% P1_enr_sub$WBID == TRUE ~ TRUE,
                                           wbps_gene_id %in% P1_enr_sub$WBID == FALSE ~ FALSE),
                       symmetric = case_when(wbps_gene_id %in% symm_sub$WBID == TRUE ~ TRUE,
                                           wbps_gene_id %in% symm_sub$WBID == FALSE ~ FALSE),
                       )
twocell_domains
```

```
## # A tibble: 273 x 9
##    wbps_gene_id   wikigene_name interpro_descript~ interpro_short_~ domain_count
##    <chr>          <chr>         <chr>              <chr>                   <int>
##  1 WBGene00000102 akt-1         PH-like domain su~ PH-like_dom_sf              2
##  2 WBGene00000103 akt-2         PH-like domain su~ PH-like_dom_sf              1
##  3 WBGene00000149 apl-1         PH-like domain su~ PH-like_dom_sf              2
##  4 WBGene00000420 ced-6         PH-like domain su~ PH-like_dom_sf              1
##  5 WBGene00000426 ced-12        PH-like domain su~ PH-like_dom_sf              2
##  6 WBGene00000564 cnk-1         PH-like domain su~ PH-like_dom_sf              1
##  7 WBGene00000565 cnt-1         PH-like domain su~ PH-like_dom_sf              3
##  8 WBGene00000894 dab-1         PH-like domain su~ PH-like_dom_sf              2
##  9 WBGene00001093 drp-1         PH-like domain su~ PH-like_dom_sf              2
## 10 WBGene00001116 dyc-1         PH-like domain su~ PH-like_dom_sf              2
## # ... with 263 more rows, and 4 more variables: present <lgl>,
## #   AB_enriched <lgl>, P1_enriched <lgl>, symmetric <lgl>
```

```
twocell_domain_genes <- twocell_domains %>% mutate(gene_type = case_when(
  present == TRUE & symmetric == FALSE & AB_enriched == FALSE ~ "no_sig_dif",
  present == TRUE & symmetric == TRUE ~ "symmetric",
  present == TRUE & AB_enriched == TRUE ~ "AB_enriched",
  present == FALSE ~ "not_detected",
)) %>% select(wbps_gene_id:domain_count, gene_type)
twocell_domain_genes
```

```
## # A tibble: 273 x 6
##    wbps_gene_id   wikigene_name interpro_descript~ interpro_short_~ domain_count
##    <chr>          <chr>         <chr>              <chr>                   <int>
##  1 WBGene00000102 akt-1         PH-like domain su~ PH-like_dom_sf              2
##  2 WBGene00000103 akt-2         PH-like domain su~ PH-like_dom_sf              1
##  3 WBGene00000149 apl-1         PH-like domain su~ PH-like_dom_sf              2
##  4 WBGene00000420 ced-6         PH-like domain su~ PH-like_dom_sf              1
##  5 WBGene00000426 ced-12        PH-like domain su~ PH-like_dom_sf              2
##  6 WBGene00000564 cnk-1         PH-like domain su~ PH-like_dom_sf              1
##  7 WBGene00000565 cnt-1         PH-like domain su~ PH-like_dom_sf              3
##  8 WBGene00000894 dab-1         PH-like domain su~ PH-like_dom_sf              2
##  9 WBGene00001093 drp-1         PH-like domain su~ PH-like_dom_sf              2
## 10 WBGene00001116 dyc-1         PH-like domain su~ PH-like_dom_sf              2
## # ... with 263 more rows, and 1 more variable: gene_type <chr>
```

```
# Number of unique genes in dataset
# Make sure the numbers match the plot above
table((twocell_domain_genes %>% distinct(wikigene_name, .keep_all = TRUE))$gene_type)
```

```
##
##  AB_enriched   no_sig_dif not_detected    symmetric
##            6           57           20           61
```

Number of protein domain types in each two cell embryo gene category

```
table(twocell_domain_genes$gene_type)
```

```
##
##  AB_enriched   no_sig_dif not_detected    symmetric
##           25          106           40          102
```

Get the names of AB enriched genes

```
unique((twocell_domain_genes %>% filter(gene_type == "AB_enriched"))$wikigene_name)
```

```
## [1] "akt-1"    "erm-1"    "frm-7"    "sma-1"    "unc-73"   "F22G12.5"
```

Get the names of symmetric enriched genes

```
unique((twocell_domain_genes %>% filter(gene_type == "symmetric"))$wikigene_name)
```

```
##  [1] "akt-2"     "cnk-1"     "cnt-1"     "dyc-1"     "ech-4"     "exc-5"
##  [7] "feh-1"     "frm-10"    "grp-1"     "hmg-3"     "hum-4"     "icln-1"
## [13] "ist-1"     "kin-32"    "max-1"     "mtm-1"     "mtm-6"     "mtm-9"
## [19] "nfm-1"     "npp-16"    "pdk-1"     "plc-1"     "ptp-1"     "ran-5"
## [25] "soc-1"     "sos-1"     "stn-1"     "tag-52"    "unc-31"    "unc-34"
## [31] "unc-104"   "vav-1"     "wsp-1"     "F07C6.4"   "vps-36"    "snx-17"
## [37] "F31D4.5"   "F38B7.3"   "T10B10.3"  "acbp-5"    "dkf-2"     "dkf-1"
## [43] "Y37D8A.25" "Y41E3.7"   "ani-3"     "obr-1"     "exoc-8"    "ZK632.12"
## [49] "obr-3"     "obr-4"     "shc-1"     "ani-2"     "K10B4.3"   "bris-1"
## [55] "M03A8.3"   "gtf-2H1"   "shc-2"     "dcap-1"    "prhg-1"    "ZC328.3"
## [61] "tbc-2"
```

Total number of genes

```
twocell_domain_genes %>% distinct(wbps_gene_id) %>% nrow()
```

```
## [1] 144
```

Output the list of protein domains

```
write.xlsx(twocell_domain_genes %>% select(wbps_gene_id:domain_count), file = "Protein_Domains_2-Cell_Em
```

Output the list of genes with 2 cell data annotation

```
write.xlsx(twocell_domain_genes %>% select(wbps_gene_id, wikigene_name, AB_vs_P1 = gene_type) %>% distin
```