# Gene identifiers and pathway annotation

## Contents



## Intro

This markdown aims to collect entrez ids, ensemble protein ids and gene symbols in a single table

```
library(readr)
library(tidyr)
library(dplyr)
library(tidylog)
library(magrittr)
```

## Neurotransmitter pathways annotation

### Defining selected pathways

```
pathways <- tribble(
  ~pathway_id,      ~pathway_name,
  "path:hsa04724",  "glutamatergic",
  "path:hsa04725",  "cholinergic",
  "path:hsa04726",  "serotonergic",
  "path:hsa04727",  "gabaergic",
  "path:hsa04728",  "dopaminergic",
  "path:hsa04721",  "vesicle"
)
```

### Genes to pathways relationships

Downloading table containing all genes and pathways relationships

```
if (!file.exists("entrez_to_pathway.tsv")) {
  download.file("http://rest.kegg.jp/link/pathway/hsa","entrez_to_pathway.tsv")
}

entrez_to_pathway <- read.table(
  "entrez_to_pathway.tsv"
  ,header = F
  ,stringsAsFactors = F
  ,col.names = c("entrez_id", "pathway_id")
  ,sep="\t"
)

# removing hsa prefix
entrez_to_pathway[, "entrez_id"] %<>% substring(5)

# exporting for package use
usethis::use_data(entrez_to_pathway, overwrite = TRUE)
```

```
## <U+2714> Setting active project to 'C:/R/neurotransmission'
## <U+2714> Saving 'entrez_to_pathway' to 'data/entrez_to_pathway.rda'
```

Filtering for genes in the selected pathways

```
gene_pathways <- inner_join(entrez_to_pathway, pathways) %>%
  mutate(n = 1) %>%
  pivot_wider(
    id_cols = entrez_id,
    names_from = pathway_name,
    values_from = n,
    values_fn = list(n = length),
    values_fill = list(n = 0)
  ) %>%
  # filling 1's in all systems for synaptic vesicle genes
  mutate_at(pathways[["pathway_name"]], ~ ifelse(vesicle == 1, 1L, .)) %>%
  # neurotransmitter systems count for each gene (>=1, <= 5)
  mutate(system_count = rowSums(select(., -entrez_id, -vesicle)))
```

```
## Joining, by = "pathway_id"

## inner_join: added one column (pathway_name)

##             > rows only in x  (31,062)

##             > rows only in y  (     0)

##             > matched rows        639

##             >                 ========
```

```
##             > rows total          639

## mutate: new variable 'n' with one unique value and 0% NA

## mutate_at: changed 69 values (18%) of 'glutamatergic' (0 new NA)

##            changed 75 values (20%) of 'cholinergic' (0 new NA)

##            changed 73 values (19%) of 'serotonergic' (0 new NA)

##            changed 70 values (18%) of 'gabaergic' (0 new NA)

##            changed 73 values (19%) of 'dopaminergic' (0 new NA)

## select: dropped 2 variables (entrez_id, vesicle)

## mutate: new variable 'system_count' with 5 unique values and 0% NA
```

```
usethis::use_data(gene_pathways, overwrite = TRUE)
```

```
## <U+2714> Saving 'gene_pathways' to 'data/gene_pathways.rda'
```

```
gene_pathways
```

```
## # A tibble: 382 x 8
##    entrez_id vesicle glutamatergic cholinergic serotonergic gabaergic
##    <chr>       <int>         <int>       <int>        <int>     <int>
##  1 10312           1             1           1            1         1
##  2 10497           1             1           1            1         1
##  3 10814           1             1           1            1         1
##  4 10815           1             1           1            1         1
##  5 112755          1             1           1            1         1
##  6 1173            1             1           1            1         1
##  7 1175            1             1           1            1         1
##  8 1211            1             1           1            1         1
##  9 1212            1             1           1            1         1
## 10 1213            1             1           1            1         1
## # ... with 372 more rows, and 2 more variables: dopaminergic <int>,
## #   system_count <dbl>
```

# Base ID lookup table

## Entrez to STRING

Downloading entrez to string mapping table directly from STRING https://string-db.org/mapping_files/entrez/

```r
if (!file.exists("human.entrez_2_string.2018.tsv.gz")) {
  download.file(
    "https://string-db.org/mapping_files/entrez/human.entrez_2_string.2018.tsv.gz",
    "human.entrez_2_string.2018.tsv.gz"
  )
}

entrez_to_string <- read_tsv(
  gzfile("human.entrez_2_string.2018.tsv.gz"),
  skip = 1,
  col_names = c("entrez_id", "string_id"),
  col_types = cols_only("-", "c", "c")
)

gene_ids <- gene_pathways %>% select(entrez_id) %>% left_join(entrez_to_string)
```

```
## select: dropped 7 variables (vesicle, glutamatergic, cholinergic, serotonergic, gabaergic, ...)

## Joining, by = "entrez_id"

## left_join: added one column (string_id)

##            > rows only in x       11

##            > rows only in y  (18,222)

##            > matched rows        371

##            >                 ========

##            > rows total          382
```

```r
gene_ids
```

```
## # A tibble: 382 x 2
##    entrez_id string_id
##    <chr>     <chr>
##  1 10312     9606.ENSP00000265686
##  2 10497     9606.ENSP00000367756
##  3 10814     9606.ENSP00000352544
##  4 10815     9606.ENSP00000305613
##  5 112755    9606.ENSP00000215095
##  6 1173      9606.ENSP00000292807
##  7 1175      9606.ENSP00000263270
##  8 1211      9606.ENSP00000242285
##  9 1212      9606.ENSP00000309415
## 10 1213      9606.ENSP00000479606
## # ... with 372 more rows
```

### STRING names

Downloading string names mapping table directly from STRING https://string-db.org/mapping_files/entrez/

```
if (!file.exists("human.name_2_string.tsv.gz")) {
  download.file(
    "https://string-db.org/mapping_files/STRING_display_names/human.name_2_string.tsv.gz",
    "human.name_2_string.tsv.gz"
  )
}

string_names <- read_tsv(
  gzfile("human.name_2_string.tsv.gz"),
  skip = 1,
  col_names = c("string_name", "string_id"),
  col_types = cols_only("-", "c", "c")
)

gene_ids %<>% left_join(string_names)
```

```
## Joining, by = "string_id"

## left_join: added one column (string_name)

##            > rows only in x       11

##            > rows only in y  (18,724)

##            > matched rows        371

##            >                 ========

##            > rows total          382
```

```
gene_ids
```

```
## # A tibble: 382 x 3
##    entrez_id string_id            string_name
##    <chr>     <chr>                <chr>
##  1 10312     9606.ENSP00000265686 TCIRG1
##  2 10497     9606.ENSP00000367756 UNC13B
##  3 10814     9606.ENSP00000352544 CPLX2
##  4 10815     9606.ENSP00000305613 CPLX1
##  5 112755    9606.ENSP00000215095 STX1B
##  6 1173      9606.ENSP00000292807 AP2M1
##  7 1175      9606.ENSP00000263270 AP2S1
##  8 1211      9606.ENSP00000242285 CLTA
##  9 1212      9606.ENSP00000309415 CLTB
## 10 1213      9606.ENSP00000479606 CLTC
## # ... with 372 more rows
```

### Entrez names

Downloading gene symbols mapping table directly from NCBI's FTP ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/GENE_INFO/Mammalia/

```
if (!file.exists("Homo_sapiens.gene_info.gz")) {
  download.file(
    "ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/GENE_INFO/Mammalia/Homo_sapiens.gene_info.gz",
    "Homo_sapiens.gene_info.gz"
  )
}

entrez_names <- read_tsv(
  gzfile("Homo_sapiens.gene_info.gz"),
  skip = 1,
  col_names = c("entrez_id", "entrez_name"),
  col_types = cols_only("-", "c", "c")
)
```

```
## Warning: 61645 parsing failures.
## row col  expected     actual         file
##   1  -- 3 columns 16 columns <connection>
##   2  -- 3 columns 16 columns <connection>
##   3  -- 3 columns 16 columns <connection>
##   4  -- 3 columns 16 columns <connection>
##   5  -- 3 columns 16 columns <connection>
## ... ... ......... .......... ............
## See problems(...) for more details.
```

```
gene_ids %<>% left_join(entrez_names)
```

```
## Joining, by = "entrez_id"
```

```
## left_join: added one column (entrez_name)
```

```
##            > rows only in x        0
```

```
##            > rows only in y  (61,263)
```

```
##            > matched rows        382
```

```
##            >                 ========
```

```
##            > rows total          382
```

```
gene_ids
```

```
## # A tibble: 382 x 4
##    entrez_id string_id            string_name entrez_name
##    <chr>     <chr>                <chr>       <chr>
```

```
##  1 10312     9606.ENSP00000265686 TCIRG1      TCIRG1
##  2 10497     9606.ENSP00000367756 UNC13B      UNC13B
##  3 10814     9606.ENSP00000352544 CPLX2       CPLX2
##  4 10815     9606.ENSP00000305613 CPLX1       CPLX1
##  5 112755    9606.ENSP00000215095 STX1B       STX1B
##  6 1173      9606.ENSP00000292807 AP2M1       AP2M1
##  7 1175      9606.ENSP00000263270 AP2S1       AP2S1
##  8 1211      9606.ENSP00000242285 CLTA        CLTA
##  9 1212      9606.ENSP00000309415 CLTB        CLTB
## 10 1213      9606.ENSP00000479606 CLTC        CLTC
## # ... with 372 more rows
```

## Updating missing info

Printing incomplete rows

```
gene_ids[!complete.cases(gene_ids),]
```

```
## # A tibble: 11 x 4
##    entrez_id string_id string_name entrez_name
##    <chr>     <chr>     <chr>       <chr>
##  1 9296      <NA>      <NA>        ATP6V1F
##  2 100137049 <NA>      <NA>        PLA2G4B
##  3 85358     <NA>      <NA>        SHANK3
##  4 8681      <NA>      <NA>        JMJD7-PLA2G4B
##  5 1139      <NA>      <NA>        CHRNA7
##  6 107987478 <NA>      <NA>        LOC107987478
##  7 107987479 <NA>      <NA>        LOC107987479
##  8 1564      <NA>      <NA>        CYP2D7
##  9 801       <NA>      <NA>        CALM1
## 10 805       <NA>      <NA>        CALM2
## 11 808       <NA>      <NA>        CALM3
```

Incomplete rows are filled manually

```
complete_info <- tribble(
  ~entrez_id,    ~string_id,                ~string_name,      ~entrez_name,
   "9296",        "9606.ENSP00000417378",    "ATP6V1F",         "ATP6V1F",
   "100137049",   "9606.ENSP00000396045",    "PLA2G4B",         "PLA2G4B",
   "85358",       NA,                        NA,                "SHANK3",
   "8681",        "9606.ENSP00000371886",    "JMJD7-PLA2G4B",   "JMJD7-PLA2G4B",
   "1139",        "9606.ENSP00000407546",    "CHRNA7",          "CHRNA7",
   "107987478",   NA,                        NA,                "LOC107987478",
   "107987479",   NA,                        NA,                "LOC107987479",
   "1564",        NA,                        NA,                "CYP2D7",
   "801",         "9606.ENSP00000349467",    "CALM1",           "CALM1",
   "805",         "9606.ENSP00000272298",    "CALM2",           "CALM2",
   "808",         "9606.ENSP00000291295",    "CALM3",           "CALM3"
)

# removing incomplete cases and adding updated ones
gene_ids %<>% na.omit %>% bind_rows(complete_info)
```

Exporting base table for package use

```
usethis::use_data(gene_ids, overwrite = TRUE)
```

```
## <U+2714> Saving 'gene_ids' to 'data/gene_ids.rda'
```