Gene Sample Gene Interaction Network

Brian Jo Hsuan Lee

5/5/2022

Read in a sample BRCA file to get an idea of what it looks like. May clean it a little.

```
file_path = "~/Downloads/TCGA/"
tumor_f = list.files(path = file_path, pattern = "T.tsv")

# 60660 BRCA genes available, use TPM Unstranded reads
sample_brca_data =
  read_delim(file = paste(file_path, tumor_f[1], sep = ""),
             skip = 1, delim = "\t", quote = "") %>%
  filter(!row_number() %in% c(1:4)) %>%
  select(gene_name, tpm_unstranded) %>%
  data.frame()
```

STRING general network maps close to 19000 genes in the BRCA file. Save the list of mapped genes

```
string_db = STRINGdb$new(version="11", species=9606, score_threshold=200, input_directory="")

# 18812 distinct BRCA genes can be mapped
mappable_genes =
  string_db$map(sample_brca_data, "gene_name", removeUnmappedRows = T) %>%
  distinct(gene_name, .keep_all = TRUE) %>%
  distinct(STRING_id, .keep_all = TRUE)
```

```
## Warning:  we couldn't map to STRING 68% of your identifiers
```

```
gene_list = mappable_genes$gene_name
write_lines(gene_list, file = "~/Desktop/Columbia/Spring_2022/P8139-Statistical_Genetic_Modeling/Projec
```

Randomly sample 2000 mapped genes from the sample BRCA file, and save the indices of those sampled

```
# identify the 18812 distinct genes on the BRCA data
sample_brca_data =
  sample_brca_data %>%
  mutate(gene_name = toupper(gene_name)) %>%
  filter(gene_name %in% gene_list) %>%
  distinct(gene_name, .keep_all = TRUE) %>%
  arrange(gene_name)

# sample 2000 from the 18812 distinct genes
set.seed(2022)
sample_ind = sample(1:nrow(sample_brca_data), 2000)
write_lines(sample_ind, file = "~/Desktop/Columbia/Spring_2022/P8139-Statistical_Genetic_Modeling/Proje
```

Attached the STRING ids onto the 2000 sampled BRCA genes

```
sample_brca_data =
  inner_join(sample_brca_data, mappable_genes) %>%
  arrange(gene_name) %>%
  filter(row_number() %in% sample_ind)
```

```
## Joining, by = c("gene_name", "tpm_unstranded")
```

Find the interaction terms for the 2000 genes

```
gene_int = string_db$get_interactions(sample_brca_data$STRING_id) %>% distinct(.keep_all = TRUE)
```

Because this sample BRCA data contains all 2000 genes and their reference STRING ids, make use of it and create a dictionary to translate STRING_ids to gene name

```
h = hash()
for(i in seq_len(nrow(sample_brca_data))){
  h[[sample_brca_data[i, 3]]] = sample_brca_data[i, 1]
}
```

Translate the interaction matrix and save

```
for(i in seq_len(nrow(gene_int))){
  gene_int[i, 1] = h[[gene_int[i, 1]]]
  gene_int[i, 2] = h[[gene_int[i, 2]]]
}

write_csv(gene_int, file = "~/Desktop/Columbia/Spring_2022/P8139-Statistical_Genetic_Modeling/Project/g
```

Build an empty sample gene correlation matrix

```
# create an empty column
S_0 = matrix(0, nrow=length(sample_brca_data$gene_name), ncol=length(sample_brca_data$gene_name)) %>% da

# name the columns and rows
col_names = sort(sample_brca_data$gene_name)
names(S_0) = col_names
rownames(S_0) = col_names

write_csv(S_0, file = "~/Desktop/Columbia/Spring_2022/P8139-Statistical_Genetic_Modeling/Project/s0_zer
```