BRCA Data Join

Brian Jo Hsuan Lee

5/5/2022

Load file paths and pre curated data, including sample indices and list of mapped genes (18812)

```
file_path = "~/Downloads/TCGA/"
tumor_f = list.files(path = file_path, pattern = "T.tsv")
norm_f = list.files(path = file_path, pattern = "N.tsv")
sample_ind = read_lines(file = "~/Desktop/Columbia/Spring_2022/P8139-Statistical_Genetic_Modeling/Proje
  as.numeric()
gene_list = read_lines(file = "~/Desktop/Columbia/Spring_2022/P8139-Statistical_Genetic_Modeling/Projec
```

Binding the 28 tumor cell data

```
# begin with the first tumor cell data
all_data =
  # 60600 x 9
  read_delim(file = paste(file_path, tumor_f[1], sep = ""),
             skip = 1, delim = "\t", quote = "") %>%
  # rid all read columns besides TSM unstranded and gene name
  select(gene_name, tpm_unstranded) %>%
  # the 18812 mapped genes are in uppercase, so case matching is important
  mutate(gene_name = toupper(gene_name)) %>%
  # drop the first 4 non-gene expression rows, and rid all gene off the mapped gene list
  filter(!row_number() %in% c(1:4),
         gene_name %in% gene_list) %>%
  # some genes are repeats (for some reasons). instead of tallying them, rid the repeats
  distinct(gene_name, .keep_all = TRUE) %>%
  # reorder them to ensure the sample indices always point to the same genes
  arrange(gene_name) %>%
  # rid the genes off the sampled gene list
  filter(row_number() %in% sample_ind) %>%
  data.frame() %>%
  # note whether the entry is a case or control
  rbind(c("data", 'case')) %>%
  # note the id of the entry for paired t test down the line
  rbind(c("id", substring(tumor_f[1], first=6, last=12)))

# continue with the other 27 tumor cell data
for(i in seq_len(length(tumor_f))[-1]){
  new_data =
    read_delim(file = paste(file_path, tumor_f[i], sep = ""),
               skip = 1, delim = "\t", quote = "") %>%
    select(gene_name, tpm_unstranded) %>%
    mutate(gene_name = toupper(gene_name)) %>%
    filter(!row_number() %in% c(1:4),
```

```
           gene_name %in% gene_list) %>%
    distinct(gene_name, .keep_all = TRUE) %>%
    arrange(gene_name) %>%
    filter(row_number() %in% sample_ind) %>%
    data.frame() %>%
    rbind(c("data", 'case')) %>%
    rbind(c("id", substring(tumor_f[i], first=6, last=12)))
  all_data = all_data %>% full_join(new_data, by = "gene_name")
}

# continue with the 28 paired adjacent normal tissue cell data
for(i in seq_len(length(norm_f))){
  new_data =
    read_delim(file = paste(file_path, norm_f[i], sep = ""),
               skip = 1, delim = "\t", quote = "") %>%
    select(gene_name, tpm_unstranded) %>%
    mutate(gene_name = toupper(gene_name)) %>%
    filter(!row_number() %in% c(1:4),
           gene_name %in% gene_list) %>%
    distinct(gene_name, .keep_all = TRUE) %>%
    arrange(gene_name) %>%
    filter(row_number() %in% sample_ind) %>%
    data.frame() %>%
    rbind(c("data", 'control')) %>%
    rbind(c("id", substring(norm_f[i], first=6, last=12)))
  all_data = all_data %>% full_join(new_data, by = "gene_name")
}

# cells to column, entries to rows
all_data_t =
  all_data %>%
  transpose() %>%
  janitor::row_to_names(row_number = 1) %>%
  dplyr::select("data", "id", everything())
```

Save the 28 pair tumor normal cell gene expression data to be combined the with rest of 244 files.

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