breast_cancer_practice

Lang Liu

12/10/2022

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.5      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.1.4      v stringr 1.4.0
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
df1 <- read_csv("../breast_cancer1.csv")
```

```
## Rows: 151 Columns: 32
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): type
## dbl (31): samples, 222859_s_at, 243182_at, 221157_s_at, 211521_s_at, 223297_...
##
## 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.
```

```
df2 <- read_csv("../breast_cancer2.csv")
```

```
## Rows: 151 Columns: 32
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): type
## dbl (31): samples, 235630_at, 208858_s_at, 203313_s_at, 1566695_at, 201585_s...
##
## 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.
```

```
code <- read_tsv("../GPL570.annot",skip=27)
```

```
## Warning: One or more parsing issues, call ‘problems()‘ on your data frame for details,
## e.g.:
##   dat <- vroom(...)
##   problems(dat)
## Rows: 54676 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (17): ID, Gene title, Gene symbol, Gene ID, UniGene title, UniGene symbo...
## dbl  (1): GI
## lgl  (3): Platform_CLONEID, Platform_ORF, Platform_SPOTID
##
## 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.
```

```
df1 %>% head()
```

```
## # A tibble: 6 x 32
##   samples type  222859~1 24318~2 22115~3 21152~4 22329~5 21175~6 22451~7 24247~8
##     <dbl> <chr>    <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1      84 basal     7.22    6.45    4.08    5.63    9.36   10.3     9.81    6.18
## 2      85 basal     8.17    5.90    3.71    5.53    7.73    8.49    9.66    5.68
## 3      87 basal     6.93    6.67    4.25    5.34    8.48    9.64   10.2     5.89
## 4      90 basal     7.80    7.10    3.83    5.67    9.55    9.42   10.3     6.02
## 5      91 basal     7.32    7.63    4.00    5.31    8.64    9.64   10.9     5.32
## 6      92 basal     5.65    5.80    4.59    5.45    8.43    9.09   10.2     5.53
## # ... with 22 more variables: ‘1560877_a_at‘ <dbl>, ‘204812_at‘ <dbl>,
## #   ‘209934_s_at‘ <dbl>, ‘239421_at‘ <dbl>, ‘236616_at‘ <dbl>,
## #   ‘214718_at‘ <dbl>, ‘1564439_a_at‘ <dbl>, ‘214065_s_at‘ <dbl>,
## #   ‘228048_at‘ <dbl>, ‘209945_s_at‘ <dbl>, ‘230539_at‘ <dbl>,
## #   ‘229195_at‘ <dbl>, ‘225733_at‘ <dbl>, ‘1561685_a_at‘ <dbl>,
## #   ‘241363_at‘ <dbl>, ‘242249_at‘ <dbl>, ‘1567179_at‘ <dbl>,
## #   ‘1554004_a_at‘ <dbl>, ‘244161_at‘ <dbl>, ‘213071_at‘ <dbl>, ...
```

```
df2 %>% head()
```

```
## # A tibble: 6 x 32
##   samples type  235630~1 20885~2 20331~3 15666~4 20158~5 24368~6 15613~7 15555~8
##     <dbl> <chr>    <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1      84 basal     5.99    7.72    7.23    2.75    8.96    6.35    4.52    7.63
## 2      85 basal     6.19    7.82    9.34    2.85    9.46    5.66    3.91    8.12
## 3      87 basal     6.15    8.26    8.65    3.33    9.51    5.65    4.12    8.72
## 4      90 basal     6.39    7.63    8.73    3.06    8.97    6.00    4.02    7.64
## 5      91 basal     6.05    7.97    8.54    3.22    9.03    6.36    3.98    8.00
## 6      92 basal     6.33    8.46    8.95    3.04    9.84    5.26    3.81    8.15
## # ... with 22 more variables: ‘223169_s_at‘ <dbl>, ‘37966_at‘ <dbl>,
## #   ‘228374_at‘ <dbl>, ‘227638_at‘ <dbl>, ‘236413_at‘ <dbl>,
## #   ‘1570009_at‘ <dbl>, ‘1553936_a_at‘ <dbl>, ‘1558785_a_at‘ <dbl>,
## #   ‘220431_at‘ <dbl>, ‘1562080_at‘ <dbl>, ‘209920_at‘ <dbl>,
## #   ‘238070_at‘ <dbl>, ‘237115_at‘ <dbl>, ‘1557022_at‘ <dbl>,
## #   ‘220489_s_at‘ <dbl>, ‘218301_at‘ <dbl>, ‘211570_s_at‘ <dbl>,
## #   ‘203806_s_at‘ <dbl>, ‘243905_at‘ <dbl>, ‘226591_at‘ <dbl>, ...
```

```
code %>% head()
```

```
## # A tibble: 6 x 21
##   ID      Gene ~1 Gene ~2 Gene ~3 UniGe~4 UniGe~5 UniGe~6 Nucle~7     GI GenBa~8
##   <chr>   <chr>   <chr>   <chr>   <chr>   <chr>   <chr>   <chr>    <dbl> <chr>
## 1 1007_s~ microR~ MIR464~ 100616~ <NA>    <NA>    <NA>    Human ~ 1.75e6 U48705
## 2 1053_at replic~ RFC2    5982    <NA>    <NA>    <NA>    Human ~ 1.59e6 M87338
## 3 117_at  heat s~ HSPA6   3310    <NA>    <NA>    <NA>    Human ~ 3.52e4 X51757
## 4 121_at  paired~ PAX8    7849    <NA>    <NA>    <NA>    H.sapi~ 3.84e4 X69699
## 5 1255_g~ guanyl~ GUCA1A  2978    <NA>    <NA>    <NA>    Homo s~ 6.23e5 L36861
## 6 1294_at microR~ MIR519~ 100847~ <NA>    <NA>    <NA>    Homo s~ 5.21e5 L13852
## # ... with 11 more variables: Platform_CLONEID <lgl>, Platform_ORF <lgl>,
## #   Platform_SPOTID <lgl>, ‘Chromosome location‘ <chr>,
## #   ‘Chromosome annotation‘ <chr>, ‘GO:Function‘ <chr>, ‘GO:Process‘ <chr>,
## #   ‘GO:Component‘ <chr>, ‘GO:Function ID‘ <chr>, ‘GO:Process ID‘ <chr>,
## #   ‘GO:Component ID‘ <chr>, and abbreviated variable names 1: ‘Gene title‘,
## #   2: ‘Gene symbol‘, 3: ‘Gene ID‘, 4: ‘UniGene title‘, 5: ‘UniGene symbol‘,
## #   6: ‘UniGene ID‘, 7: ‘Nucleotide Title‘, 8: ‘GenBank Accession‘
```

```
#Merge breast_cancer1 and breast_cancer2
df <- merge(df1,df2,by = c("samples"))
#first way to avoid duplicate columns
#df <- merge(df1,df2,by = c("samples","type"))




#second way to avoid duplicate columns
#colnames(df[df %>% colnames() %>% str_detect("y")])#detect if a column is duplicated when merging
df <- df %>% select(-c("type.y")) %>% rename(type = type.x) #deselct the column and rename the type
```

```
#Replace the probe name with gene symbol in GPL570.annot
code_sub <- code %>% select(ID, "Gene symbol")
df_columns <- tibble(ID=colnames(df)[-c(1:2)])
df_code <- merge(df_columns,code_sub,by = "ID",sort=FALSE)
#NA present in gene symbols
df_code %>% summarise_all(~sum(is.na(.)))
```

```
##   ID Gene symbol
## 1  0           8
```

```
df_code %>% filter(is.na(df_code$`Gene symbol`))
```

```
##             ID Gene symbol
## 1    243182_at        <NA>
## 2 1560877_a_at        <NA>
## 3    244161_at        <NA>
## 4    235630_at        <NA>
## 5   1566695_at        <NA>
## 6    243682_at        <NA>
## 7    236413_at        <NA>
## 8    237115_at        <NA>
```

```
#replace these missing values with IDs
df_code <- df_code %>%
  mutate(`Gene symbol` = ifelse(is.na(df_code$`Gene symbol`),
                                      df_code$ID,df_code$`Gene symbol`))
```

```r
#check if missing values are present again
df_code %>% summarise_all(~sum(is.na(.)))
```

```
##   ID Gene symbol
## 1  0           0
```

```r
#rename the columns
colnames(df)[3:length(colnames(df))] <- df_code[,2]
```

```r
#get top 10 genes that are expressed the highest in basal type
df %>%
  group_by(type) %>%
  select(-samples) %>%
  summarise_all(list(avg=mean)) %>%
  pivot_longer(cols = !type,names_to = 'gene') %>%
  pivot_wider(id_cols = gene, names_from=type) %>%
  arrange(desc(basal)) %>%
  slice(1:10) %>%
  select(gene,basal)
```

```
## # A tibble: 10 x 2
##    gene         basal
##    <chr>        <dbl>
##  1 TXNDC17_avg  10.2
##  2 PHB_avg       9.75
##  3 TXNDC9_avg    9.27
##  4 HPCAL1_avg    9.06
##  5 SFPQ_avg      8.61
##  6 AMMECR1L_avg  8.58
##  7 TGIF1_avg     8.57
##  8 GSK3B_avg     8.46
##  9 WDR45_avg     8.34
## 10 ESYT1_avg     8.17
```