ET_week45

Jilong

11/10/2021

R load the data with tidyverse

```
genes <- read_tsv("genes.txt")
```

```
## Rows: 24 Columns: 7
```

```
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: "\t"
## chr (4): Symbol, Class, Tissue_expression, X_linked_Homologue
## dbl (3): coorStart, coorEnd, Length
```

```
## 
## 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.
```

```
variants <- read_tsv("variants.txt")
```

```
## Rows: 25397 Columns: 5
```

```
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: "\t"
## chr (2): Type, Region
## dbl (3): Position, Count, minorAllele
```

```
## 
## 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.
```

```
copies <- read_tsv("copyNumbers.txt")
```

```
## Rows: 62 Columns: 26
```

```
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: "\t"
## chr  (2): Haplogroup, Ind
## dbl (24): AMELY, BPY2, CDY, DAZ, DBY, EIF1AY, HSFY, KDM5D, NLGN4Y, PRKY, RBM...
```

```
## 
## 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.
```

Questions

```
## Question 1
# get the number of snps vector for all genes
nsnps<-c()
for (i in seq(1,NROW(genes),1)){
  gene <- genes$Symbol[i]
  snps <- NROW(variants%>%filter(Position>= genes$coorStart[i] & Position <= genes$coorEnd[i])%>%filter(Type == "
SNP"))
  nsnps <- c(nsnps,snps)
}
# add the vector as a column
stat_genes <- genes %>% add_column(N_snps = nsnps)
# calculate the mean frequency
stat_genes%>%mutate(snp_f = N_snps/Length)%>%select(snp_f)%>%summarise(across(everything(),
                                                                             mean))
```

```
## # A tibble: 1 × 1
##      snp_f
##      <dbl>
## 1 0.000303
```

```
## Question 2
stat_genes%>%mutate(snp_f = N_snps/Length)%>%arrange(desc(snp_f))%>%head(1)
```

```
## # A tibble: 1 × 9
##   Symbol Class Tissue_expressi… X_linked_Homolo… coorStart coorEnd Length N_snps
##   <chr>  <chr> <chr>            <chr>                <dbl>   <dbl>  <dbl>  <int>
## 1 TSPY   Ampl… Testis           NaN                9466955 9469749   2794      5
## # … with 1 more variable: snp_f <dbl>
```

```
## Question 3
N_nsnps<-c()
for (i in seq(1,NROW(genes),1)){
  gene <- genes$Symbol[i]
  snps <- NROW(variants%>%filter(Position>= genes$coorStart[i] & Position <= genes$coorEnd[i])%>%filter(Type != "
SNP"))
  N_nsnps <- c(N_nsnps,snps)
}
# add the vector as a column
stat_genes <- stat_genes %>% add_column(N_nsnps = N_nsnps)
# calculate the mean frequency
stat_genes%>%mutate(nsnp_f = N_nsnps/Length)%>%select(nsnp_f)%>%summarise(across(everything(),
                                                                              mean))
```

```
## # A tibble: 1 × 1
##     nsnp_f
##      <dbl>
## 1 0.000160
```

```
## Question 4
stat_genes%>%mutate(nsnp_f = N_nsnps/Length)%>%arrange(desc(nsnp_f))%>%head(1)
```

```
## # A tibble: 1 × 10
##   Symbol Class Tissue_expressi… X_linked_Homolo… coorStart coorEnd Length N_snps
##   <chr>  <chr> <chr>            <chr>                <dbl>   <dbl>  <dbl>  <int>
## 1 RPS4Y2 X-de… Ubiquitous       RPS4X             20756068  2.08e7  24988     10
## # … with 2 more variables: N_nsnps <int>, nsnp_f <dbl>
```

```
## Question 5
copies %>% gather(key = "gene", value = "copy_number", -Haplogroup,-Ind)%>%
  group_by(Ind)%>%summarise(mean_cp = mean(copy_number))%>%arrange(desc(mean_cp))%>%head(1)
```

```
## Warning: One or more parsing issues, see `problems()` for details
```

```
## # A tibble: 1 × 2
##   Ind     mean_cp
##   <chr>     <dbl>
## 1 1054-01    2.80
```

```
## Question 6
copies %>% gather(key = "gene", value = "copy_number", -Haplogroup,-Ind)%>%
  group_by(gene)%>%summarise(mean_cp = mean(copy_number))%>%arrange(desc(mean_cp))%>%head(1)
```

```
## # A tibble: 1 × 2
##   gene  mean_cp
##   <chr>   <dbl>
## 1 TSPY     22.0
## Question 7
var_across_gene <- copies %>% gather(key = "gene", value = "copy_number", -Haplogroup,-Ind)%>%
  group_by(Ind)%>%summarise(var_cp = var(copy_number))%>%select(var_cp)%>%summarise(across(everything(),
                                                                              mean))
var_across_ind <- copies %>% gather(key = "gene", value = "copy_number", -Haplogroup,-Ind)%>%
  group_by(gene)%>%summarise(var_cp = var(copy_number))%>%select(var_cp)%>%summarise(across(everything(),
                                                                              mean))
```

```
## Question 8
gene_mean_cp <- copies %>% gather(key = "Symbol", value = "copy_number", -Haplogroup,-Ind)%>%
  group_by(Symbol)%>%summarise(mean_cp = mean(copy_number))
final_gene <- right_join(stat_genes,gene_mean_cp,by="Symbol")%>%mutate(snp_f = N_snps/Length)
cor(final_gene$mean_cp,final_gene$snp_f)
```

```
## [1] 0.7942312
```

```
# ## Question 9
# final_gene%>%mutate(fix_snp_f = N_snps/(Length*mean_cp))%>%arrange(desc(fix_snp_f))%>%head(1)
# ## Question 1
# # get the number of snps vector for all genes
# nsnps<-c()
# for (i in seq(1,NROW(genes),1)){
#   gene <- genes$Symbol[i]
#   snps <- NROW(variants%>%filter(Position>= genes$coorStart[i] & Position <= genes$coorEnd[i])%>%filter(Type ==
"SNP"))
#   nsnps <- c(nsnps,snps)
# }
# # add the vector as a column
# stat_genes <- genes %>% add_column(N_snps = nsnps)
# # calculate the mean frequency
# stat_genes%>%mutate(snp_f = N_snps/Length)%>%select(snp_f)%>%summarise(across(everything(),
#                                                                                mean))
# stat_genes
# ## Question 2
# stat_genes%>%mutate(snp_f = N_snps/Length)%>%arrange(desc(snp_f))%>%head(1)
#
# ## Question 3
# N_nsnps<-c()
# for (i in seq(1,NROW(genes),1)){
#   gene <- genes$Symbol[i]
#   snps <- NROW(variants%>%filter(Position>= genes$coorStart[i] & Position <= genes$coorEnd[i])%>%filter(Type !=
"SNP"))
#   N_nsnps <- c(N_nsnps,snps)
# }
# # add the vector as a column
# stat_genes <- stat_genes %>% add_column(N_nsnps = N_nsnps)
# # calculate the mean frequency
# stat_genes%>%mutate(nsnp_f = N_nsnps/Length)%>%select(nsnp_f)%>%summarise(across(everything(),
#                                                                                mean))
#
# ## Question 4
# stat_genes%>%mutate(nsnp_f = N_nsnps/Length)%>%arrange(desc(nsnp_f))%>%head(1)
#
# ## Question 5
# copies %>% gather(key = "gene", value = "copy_number", -Haplogroup,-Ind)%>%
#   group_by(Ind)%>%summarise(mean_cp = mean(copy_number))%>%arrange(desc(mean_cp))%>%head(1)
#
# ## Question 6
# copies %>% gather(key = "gene", value = "copy_number", -Haplogroup,-Ind)%>%
#   group_by(gene)%>%summarise(mean_cp = mean(copy_number))%>%arrange(desc(mean_cp))%>%head(1)
#
# ## Question 7
# var_across_gene <- copies %>% gather(key = "gene", value = "copy_number", -Haplogroup,-Ind)%>%
#   group_by(Ind)%>%summarise(var_cp = var(copy_number))%>%select(var_cp)%>%summarise(across(everything(),
#                                                                                mean))
# var_across_ind <- copies %>% gather(key = "gene", value = "copy_number", -Haplogroup,-Ind)%>%
#   group_by(gene)%>%summarise(var_cp = var(copy_number))%>%select(var_cp)%>%summarise(across(everything(),
#                                                                                mean))
# ## Question 8
# gene_mean_cp <- copies %>% gather(key = "Symbol", value = "copy_number", -Haplogroup,-Ind)%>%
#   group_by(Symbol)%>%summarise(mean_cp = mean(copy_number))
# final_gene <- right_join(stat_genes,gene_mean_cp,by="Symbol")%>%mutate(snp_f = N_snps/Length)
# cor(final_gene$mean_cp,final_gene$snp_f)
#
# ## Question 9
# final_gene%>%mutate(fix_snp_f = N_snps/(Length*mean_cp))%>%arrange(desc(fix_snp_f))%>%head(1)
#
# ## Question 10
```