missing-protein-unique-peptides

2023-04-06

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.5
## v tibble  3.2.1     v dplyr   1.1.2
## v tidyr   1.2.1     v stringr 1.4.1
## 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()
```

```
if (!require("protools", quietly = TRUE))
   devtools::install_github("https://github.com/FDUguchunhui/protools")
library(protools)
```

Import peptides of identified missing proteins

```
MP_final_peptide <- readxl::read_xlsx('Supplementary file 2 identified missing protein details.xlsx', sh
head(MP_final_peptide)
```

```
## # A tibble: 6 x 93
##   Gene     protein.key protein.Entry    protein.Accession protein.Description
##   <chr>          <dbl> <chr>            <chr>             <chr>
## 1 C1orf141        3363 A0A0A0MTM1_HUMAN A0A0A0MTM1        Isoform of Q5JVX7_ Un~
## 2 C1orf141        3363 A0A0A0MTM1_HUMAN A0A0A0MTM1        Isoform of Q5JVX7_ Un~
## 3 C1orf141        3363 A0A0A0MTM1_HUMAN A0A0A0MTM1        Isoform of Q5JVX7_ Un~
## 4 C1orf141        3363 A0A0A0MTM1_HUMAN A0A0A0MTM1        Isoform of Q5JVX7_ Un~
## 5 C1orf141        3363 A0A0A0MTM1_HUMAN A0A0A0MTM1        Isoform of Q5JVX7_ Un~
## 6 C1orf141        3363 A0A0A0MTM1_HUMAN A0A0A0MTM1        Isoform of Q5JVX7_ Un~
## # i 88 more variables: protein.dataBaseType <chr>, protein.score <dbl>,
## #   protein.falsePositiveRate <dbl>, protein.avgMass <dbl>,
## #   protein.MatchedProducts <dbl>, protein.matchedPeptides <dbl>,
## #   protein.digestPeps <dbl>, `protein.seqCover(%)` <dbl>,
## #   protein.MatchedPeptideIntenSum <dbl>,
## #   protein.top3MatchedPeptideIntenSum <dbl>,
## #   protein.MatchedProductIntenSum <dbl>, protein.fmolOnColumn <lgl>, ...
```

Import uniqueness checking

```
uniqueness_checking <- readxl::read_xlsx('Supplementary file 2 identified missing protein details.xlsx'
head(uniqueness_checking)
```

```
## # A tibble: 6 x 7
##   peptide   UniquenessWithoutVar~1 countIsoMatchedWitho~2 listIsoMatchedWithou~3
##   <chr>     <chr>                                  <dbl> <chr>
## 1 HTGSGILS~ N                                          0 <NA>
## 2 RPAFPVIH~ Y                                          1 NX_Q8N687-1
## 3 VYGPAESQ~ Y                                          2 NX_Q9UIL4-1 NX_Q9UIL4~
## 4 GHVGIFFI~ Y                                          2 NX_Q8N5U1-2 NX_Q8N5U1~
## 5 PLLPSTVG~ Y                                          1 NX_Q9H3Y0-1
## 6 ILQKEEEA~ Y                                          1 NX_A6NFK2-1
## # i abbreviated names: 1: UniquenessWithoutVariant,
## #   2: countIsoMatchedWithoutVariant, 3: listIsoMatchedWithoutVariant
## # i 3 more variables: UniquenessWithVariant <chr>,
## #   countAdditionalIsoMatchedWithVariant <dbl>,
## #   listAdditionalIsoMatchedWithVariant <chr>
```

Processing data before computing the number of unique peptides for each identified missing proteins

```
MP_final_peptide$Source <- str_extract(MP_final_peptide$Source, '^IPAS[0-9]+(?=_)')
MP_final_peptide$unique <- ifelse(is.na(MP_final_peptide$unique) | MP_final_peptide$unique == 'N', 0 , 1
```

Import peptides of 204 identified missing proteins

```
MP_204_products <- readxl::read_xlsx('Supplementary file 2 identified missing protein details.xlsx', sh
MP_204_products$IPAS <- paste0('IPAS', str_extract(MP_204_products$IPAS, '(?<=IP).+(?=_)'))
```

The uniqueness checker contains information about the identified 298 missing proteins (without SpC >=2 constraint), and it is a superset of the identified 204 MPs. The following code extract uniqueness checker information only for those 204 MPs.

```
unique_peptide_summary <- MP_final_peptide %>% filter(unique == 1) %>% group_by(Gene, Source) %>% summa
```

```
## `summarise()` has grouped output by 'Gene'. You can override using the
## `.groups` argument.
```

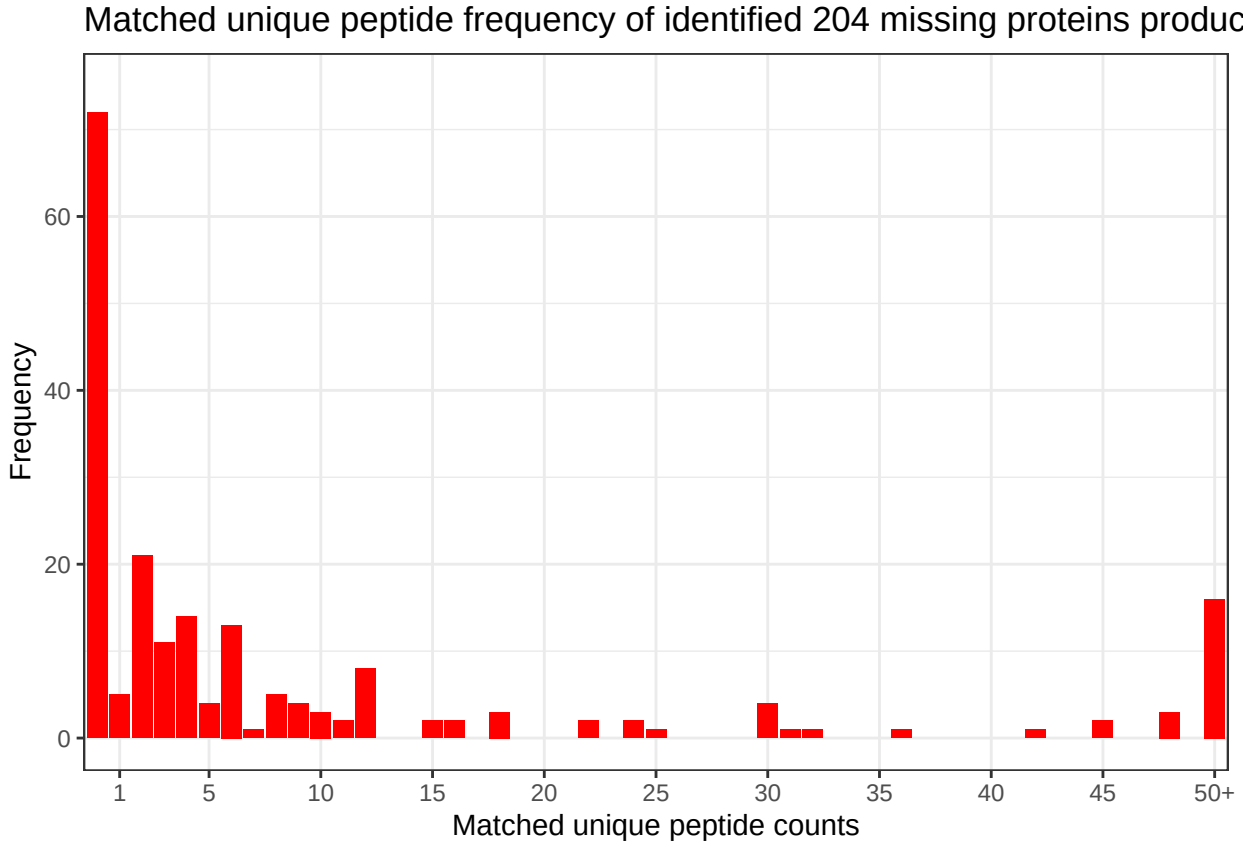
Get Summary of number of unique peptides for each of the 204 MPs

```
unique_peptide_summary <- MP_204_products %>% left_join(unique_peptide_summary, by=c('accession' = 'Gen
# write_csv(unique_peptide_summary, 'missing-protein-project/output/missing_protein_unique_peptides.csv
```

plot unique peptide frequency for MP with/without RNA expression

```
unique_peptide_summary <-  unique_peptide_summary %>% mutate(count_discrete=ifelse(n >= 50, '50+', n))
unique_peptide_summary$count_discrete <- factor(unique_peptide_summary$count_discrete, levels=c(as.chara

unique_peptide_summary %>%
  ggplot(aes(x=count_discrete)) +
  geom_bar(position = 'identity', fill='red') +
theme_bw() +
  xlab('Matched unique peptide counts') +
  ylab('Frequency') +
  ggtitle('Matched unique peptide frequency of identified 204 missing proteins products') +
  scale_x_discrete(breaks=c('1', seq(5, 49, 5), '50+'), drop=FALSE) +
  scale_y_continuous(limits=c(0, 75))
```



plot unique peptide frequency for MP with RNA expression

```
unique_peptide_summary <- unique_peptide_summary %>% filter(TPM > 0)
unique_peptide_summary <-  unique_peptide_summary %>% mutate(count_discrete=ifelse(n >= 50, '50+', n))
unique_peptide_summary$count_discrete <- factor(unique_peptide_summary$count_discrete, levels=c(as.chara
```

```
unique_peptide_summary %>%
  ggplot(aes(x=count_discrete)) +
  geom_bar(position = 'identity', fill='blue') +
theme_bw() +
  xlab('Matched unique peptide counts') +
  ylab('Frequency') +
  ggtitle('Matched unique peptide frequency of identified 119 missing proteins products
          with mRNA expression') +
  scale_x_discrete(breaks=c('0', seq(1, 49, 5), '50+'), drop=FALSE) +
  scale_y_continuous(limits=c(0, 75))
```