

detected missing protein unique peptides analysis

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', sheet = 1)
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 AOA0A0MTM1_HUMAN AOA0A0MTM1         Isoform of Q5JVB7_ Un~
## 2 C1orf141    3363 AOA0A0MTM1_HUMAN AOA0A0MTM1         Isoform of Q5JVB7_ Un~
## 3 C1orf141    3363 AOA0A0MTM1_HUMAN AOA0A0MTM1         Isoform of Q5JVB7_ Un~
## 4 C1orf141    3363 AOA0A0MTM1_HUMAN AOA0A0MTM1         Isoform of Q5JVB7_ Un~
## 5 C1orf141    3363 AOA0A0MTM1_HUMAN AOA0A0MTM1         Isoform of Q5JVB7_ Un~
## 6 C1orf141    3363 AOA0A0MTM1_HUMAN AOA0A0MTM1         Isoform of Q5JVB7_ 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$unique <- ifelse((MP_final_peptide$unique == '.' | MP_final_peptide$unique == 'N'), 0
```

Import peptides of 177 identified missing proteins

```
# MP_204_products <- readxl::read_xlsx('Supplementary file 2 identified missing protein details.xlsx',
MP_177_products <- read_csv('detected_177_MP_products_with_RNA.csv')
```

```
## Rows: 177 Columns: 6
## -- Column specification -----
## Delimiter: ","
## chr (4): accession, IPAS, type, gene_symbol
## dbl (2): NSAF, TPM
##
## 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.
```

```
MP_177_products$IPAS <- paste0('IPAS', str_extract(MP_177_products$IPAS, '^(?<=IP).+(?=_)'))
```

Calculate the number of unique peptides for each detected MPs
(and identified MPs)

```
unique_peptide_summary <- MP_final_peptide %>% filter(unique == 1) %>% group_by(protein.Accession, Source)

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

Get Summary of number of unique peptides for each of the 177 MPs

```
unique_peptide_summary <- MP_177_products %>% left_join(unique_peptide_summary, by=c('accession' = 'protein.Accession'))
# write_csv(unique_peptide_summary, 'missing_protein_unique_peptides.csv')
```

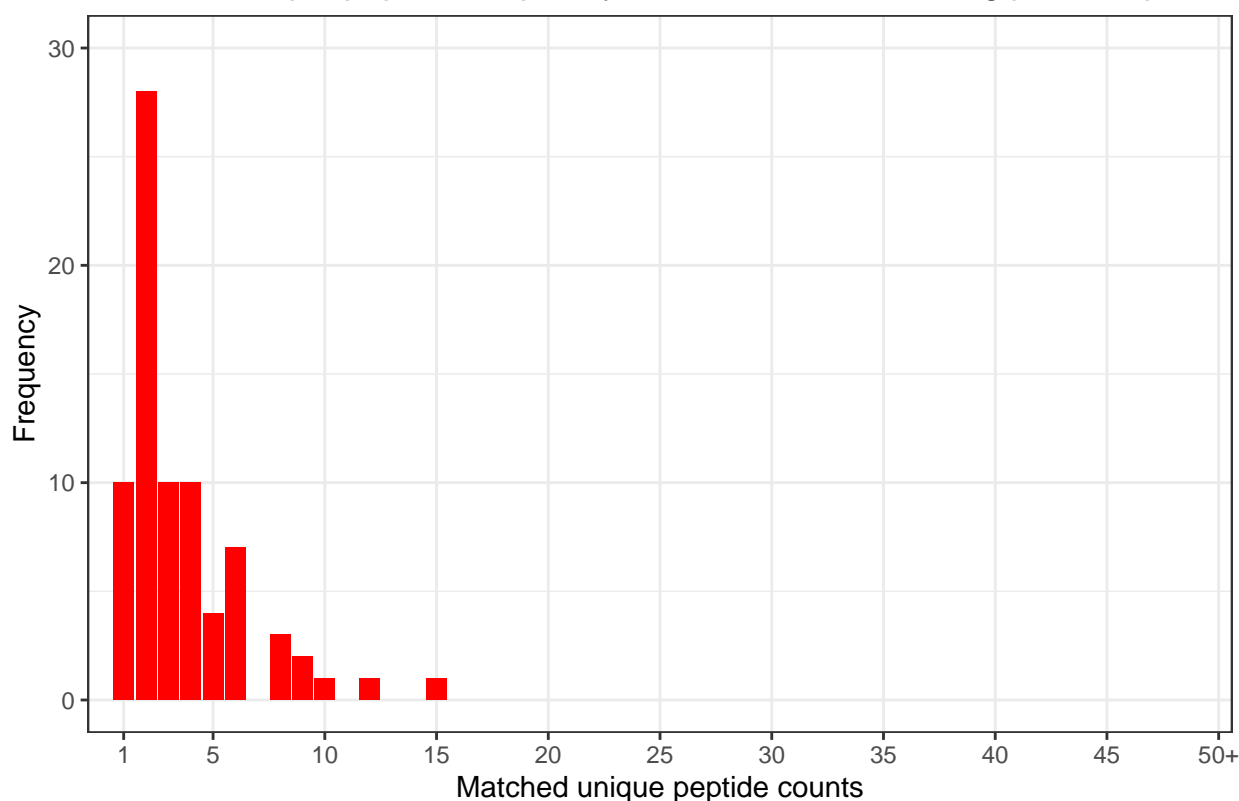
plot unique peptide frequency for MP with/without RNA expression

```
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.character(1:49), '50+', '50+')))

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 177 missing proteins products') +
  scale_x_discrete(breaks=c('1', seq(5, 49, 5), '50+'), drop=FALSE) +
  scale_y_continuous(limits=c(0, 30))
```

```
## Warning: Removed 1 rows containing missing values (geom_bar).
```

Matched unique peptide frequency of identified 177 missing proteins produc



```
unique_peptide_summary %>% summary()
```

```
##   accession           IPAS           NSAF           type
## Length:177      Length:177      Min.   : 9.264   Length:177
## Class :character Class :character 1st Qu.: 48.719   Class :character
## Mode  :character Mode  :character Median : 78.349   Mode  :character
##                                     Mean  :138.911
##                                     3rd Qu.:159.265
##                                     Max.   :950.963
##
## gene_symbol      TPM              n           unique_peptide
## Length:177      Min.   : 0.0000   Min.   : 0.000   Length:177
## Class :character 1st Qu.: 0.0000   1st Qu.: 0.000   Class :character
## Mode  :character Median : 0.0000   Median : 0.000   Mode  :character
##                                     Mean  : 7.3663   Mean  : 1.565
##                                     3rd Qu.: 0.7085   3rd Qu.: 2.000
##                                     Max.   :321.2116   Max.   :15.000
##
## count_discrete
## 0      :100
## 2      : 28
## 1      : 10
## 3      : 10
## 4      : 10
## 6      : 7
## (Other): 12
```

```
unique_peptide_summary %>% filter(n >= 2)
```

```
## # A tibble: 67 x 9
##   accession IPAS      NSAF type gene_symbol      TPM      n unique_peptide
##   <chr>      <chr>    <dbl> <chr> <chr>      <dbl> <int> <chr>
## 1 AOA075B6T7 IPAS7105 193.  MP   TRAV6      11.6      2 QSLFHITASQPADSAT~
## 2 AOA087WT02 IPAS7100 375.  MP   TRAV9-2    84.9      5 GSVQVSDSAVYFCALS~
## 3 AOA096LP55 IPAS7105 560.  MP   UQCRHL    321.      5 SHTEEDCTEELFDFLH~
## 4 AOA0A6YYG3 IPAS0995 146.  MP   TRBV6-8      0      6 QDPGMGLRLIYYSAAA~
## 5 AOA0B4J237 IPAS0995 218.  MP   TRAV8-2      0      2 SETSFHLTKPSAHMSD~
## 6 AOA0J9YX75 IPAS0982  78.3 MP   TRBV6-9      0      4 MSIGLLCCVAFSLLWA~
## 7 A4D1E1      IPAS7105  56.7 MP   ZNF804B      0     10 ISECLDEFSSLEPSEQ~
## 8 A6NFK2      IPAS0982  72.0 MP   GRXCR2    0.0128     8 ILQKEEEEAEESLMNK~
## 9 A6NFK2      IPAS7100 203.  MP   GRXCR2      0      8 QVFEDGQELESPKEEY~
## 10 A6NFN9     IPAS0999  67.6 MP   ANKUB1    0.00357    2 VALYIAAFCGYIELTE~
## # i 57 more rows
## # i 1 more variable: count_discrete <fct>
```

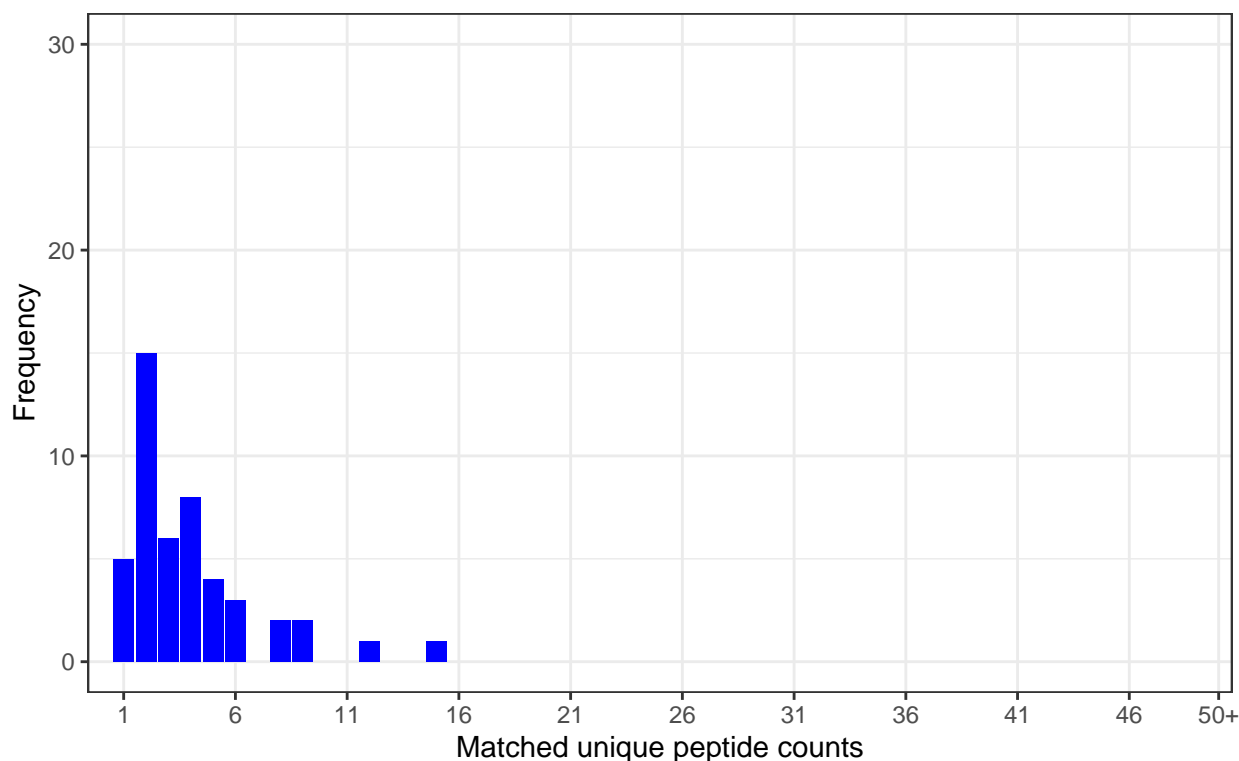
plot unique peptide frequency for MP with RNA expression

```
unique_peptide_summary_88 <- unique_peptide_summary %>% filter(TPM > 0)
unique_peptide_summary_88 <- unique_peptide_summary_88 %>% mutate(count_discrete=ifelse(n >= 50, '50+',
unique_peptide_summary_88$count_discrete <- factor(unique_peptide_summary_88$count_discrete, levels=c(a

unique_peptide_summary_88 %>%
  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 88 missing proteins products
          with mRNA expression') +
  scale_x_discrete(breaks=c('1', seq(1, 49, 5), '50+'), drop=FALSE) +
  scale_y_continuous(limits=c(0, 30))
```

```
## Warning: Removed 1 rows containing missing values (geom_bar).
```

Matched unique peptide frequency of identified 88 missing proteins products with mRNA expression



```
unique_peptide_summary_88 %>% summary()
```

```
##   accession           IPAS           NSAF           type
## Length:88           Length:88         Min.    : 9.264   Length:88
## Class :character     Class :character 1st Qu.: 43.362   Class :character
## Mode  :character     Mode  :character Median : 77.131   Mode  :character
##                                     Mean  :139.252
##                                     3rd Qu.:155.888
##                                     Max.  :950.963
##
##   gene_symbol      TPM           n           unique_peptide
## Length:88         Min.    : 0.0036   Min.    : 0.000   Length:88
## Class :character   1st Qu.: 0.1063   1st Qu.: 0.000   Class :character
## Mode  :character   Median : 0.7315   Median : 1.000   Mode  :character
##                                     Mean  : 14.8163   Mean  : 2.091
##                                     3rd Qu.: 7.5493   3rd Qu.: 3.000
##                                     Max.  :321.2116   Max.  :15.000
##
## count_discrete
## 0      :41
## 2      :15
## 4      : 8
## 3      : 6
## 1      : 5
## 5      : 4
## (Other): 9
```

```
unique_peptide_summary_88 %>% filter(n >= 2)
```

```
## # A tibble: 42 x 9
##   accession IPAS      NSAF type gene_symbol      TPM      n unique_peptide
##   <chr>      <chr>    <dbl> <chr> <chr>      <dbl> <int> <chr>
## 1 A0A075B6T7 IPAS7105 193.  MP   TRAV6      11.6      2 QSLFHITASQPADSAT~
## 2 A0A087WT02 IPAS7100 375.  MP   TRAV9-2    84.9      5 GSVQVSDSAVYFCALS~
## 3 A0A096LP55 IPAS7105 560.  MP   UQCRHL    321.      5 SHTEEDCTEELFDLH~
## 4 A6NFK2      IPAS0982 72.0  MP   GRXCR2     0.0128     8 ILQKEEEEESLMNK~
## 5 A6NFN9      IPAS0999 67.6  MP   ANKUB1     0.00357    2 VALYIAAFCGYIELTE~
## 6 A6NH13      IPAS7100 227.  MP   DNAJC9-AS1 0.269      2 PGGDTTPEEAAAPSCA~
## 7 A6NHG4      IPAS0982 400.  MP   DDTL       0.0891     5 FPTVLSTSPA AHGGPR~
## 8 A6NJJ6      IPAS0995 68.9  MP   C19orf67   0.906      2 MATEQWFEGSLPLDPG~
## 9 A6NK53      IPAS0982 26.7  MP   ZNF233     1.01       4 FQEMVTFKDVAVVFTR~
## 10 A6NK53     IPAS7105 76.1  MP   ZNF233     10.3      4 ESSQHSIIQSGEQTSD~
## # i 32 more rows
## # i 1 more variable: count_discrete <fct>
```