MDstatsDIAMS: Real Data Exploration Using Spectronaut Report

Namgil Lee*, Hojin Yoo†, Juhyoung Kim‡, Heejung Yang§

June 14, 2025

Contents

This vignette explores and visualizes real DIA-MS data given in a Spectronaut report. This vignette reproduces figures and tables in Sections S1 and S2 of the Supplementary Material of the paper "*A Shrinkage-based Statistical Method for Testing Group Mean Differences in Quantitative Bottom-up Proteomics*" written by the authors.

Preparation

1. Set parameters to reduce analysis time

Because the real data size can be too large, users can choose parameters to reduce analysis time in this vignette.

```
# n_protein:
#   Size of a subset of proteins randomly selected.
#   If -1 or Inf, include all proteins. Default to 100.
n_protein <- 100

# remove_intermediate_reports:
#   If TRUE, remove report from environment if not used in further steps.
#   Default is TRUE.
remove_intermediate_reports <- TRUE
```

2. Load packages.

```
library(dplyr)
```

3. Load Report Data

Load a Spectronaut report from a remote repository.

```
df_real <- arrow::read_parquet(
  paste0(
    "/Users/namgil/Documents/Projects/MDstatsDIAMS/data/",
```

*Kangwon National University, and Bionsight Inc., namgil.lee@kangwon.ac.kr
†Bionsight Inc.
‡Kangwon National University
§Kangwon National University, and Bionsight Inc., heejyang@kangwon.ac.kr

```
    "lip_quant_staurosporine_hela_sn_report.parquet"
  )
)
```

4. Filter the report and compute log10-transformed precursor quantity.

```
df_filtered <- df_real %>%
  filter(
    F.ExcludedFromQuantification == 'False'
    & EG.Qvalue < 0.01
    & F.NormalizedPeakArea > 1
  ) %>%
  mutate(F.Log10NormalizedPeakArea = log10(F.NormalizedPeakArea)) %>%
  select(
    R.Condition, R.Replicate, PG.ProteinGroups, PG.ProteinNames,
    EG.ModifiedSequence, FG.Charge, F.FrgIon, F.FrgLossType, F.Charge,
    EG.Qvalue, F.Log10NormalizedPeakArea
  )

print(dim(df_filtered))
```

```
## [1] 13904962       11
```

5. Set the order of the condition values so that "DMSO" comes before the other conditions.

```
conditions = unique(df_filtered$R.Condition)

df_filtered <- df_filtered %>%
  mutate(
    R.Condition = factor(
      R.Condition,
      labels = c("DMSO", conditions[-which(conditions == "DMSO")])
    )
  )

print(levels(df_filtered$R.Condition))
```

```
## [1] "DMSO"  "100pM" "1nM"   "10nM"  "100nM" "1mM"   "10mM"  "100mM"
```

6. Randomly select a subset of the predefined number of proteins to reduce analysis time.

```
if (
  n_protein < 0
  || is.infinite(n_protein)
  || n_protein >= n_distinct(df_filtered$PG.ProteinGroups)
) {
  df_subset = df_filtered
} else {
  set.seed(111)
  protein_subset = sample(unique(df_filtered$PG.ProteinGroups), n_protein)
  df_subset = df_filtered %>% filter(PG.ProteinGroups %in% protein_subset)
}

print(dim(df_subset))
```

```
## [1] 246385     11
```

7. Remove the large report data from the R environment that will not be used in the next steps.

Summary Statistics

Print summary statistics by condition.

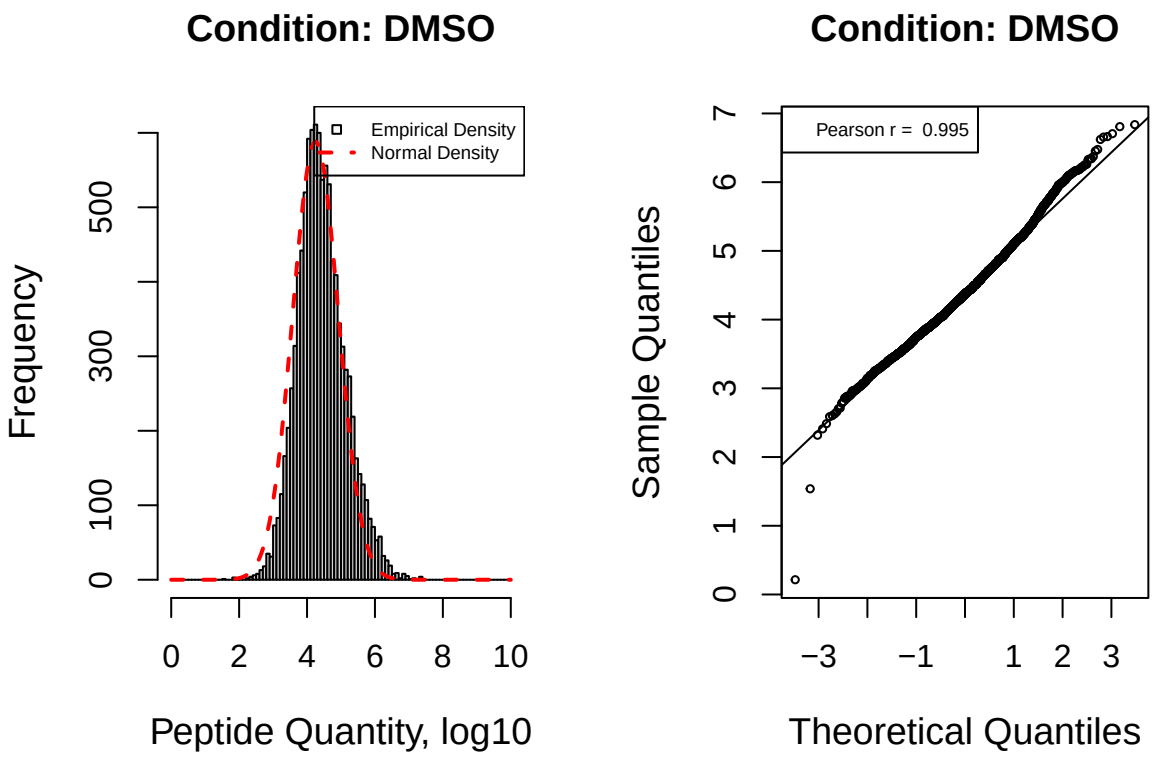
```
## # A tibble: 8 x 4
##   R.Condition num_protein num_precursor num_fragment_ion
##   <fct>             <int>         <int>            <int>
## 1 DMSO                 99          2871             9064
## 2 100pM               100          2841             8972
## 3 1nM                  99          2822             8907
## 4 10nM                 98          2818             8861
## 5 100nM                99          2850             8967
## 6 1mM                  99          2822             8895
## 7 10mM                 98          2836             8951
## 8 100mM                99          2854             8912
```

```
## # A tibble: 8 x 3
##   R.Condition precursor_mean_quantity precursor_sd_quantity
##   <fct>                         <dbl>                 <dbl>
## 1 DMSO                           4.43                 0.707
## 2 100pM                          4.44                 0.712
## 3 1nM                            4.45                 0.710
## 4 10nM                           4.45                 0.721
## 5 100nM                          4.44                 0.709
## 6 1mM                            4.44                 0.713
## 7 10mM                           4.45                 0.714
## 8 100mM                          4.44                 0.729
```

Precursor Quantity Distribution

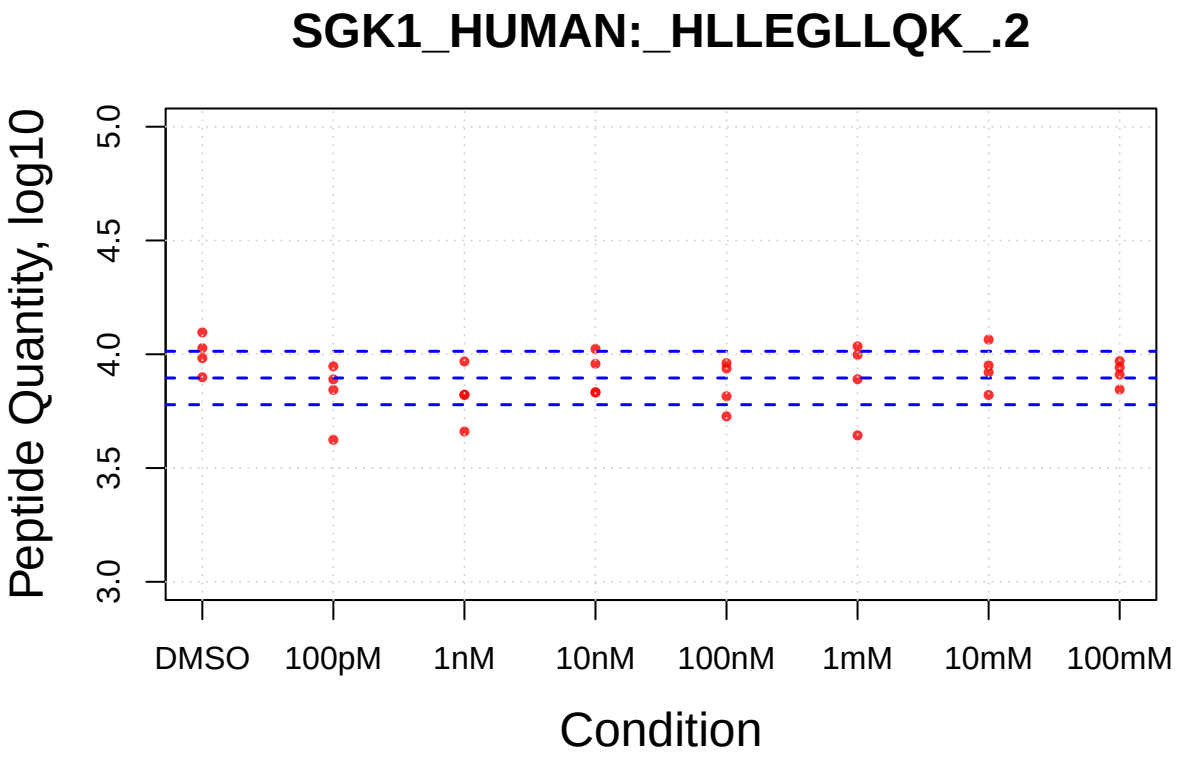
1. Precursor quantity distribution at fixed condition

Distribution of precursor quantity: all precursors measured in four replicates under fixed condition.



2. An Example of Dose-Response Plot for A Selected Precursor

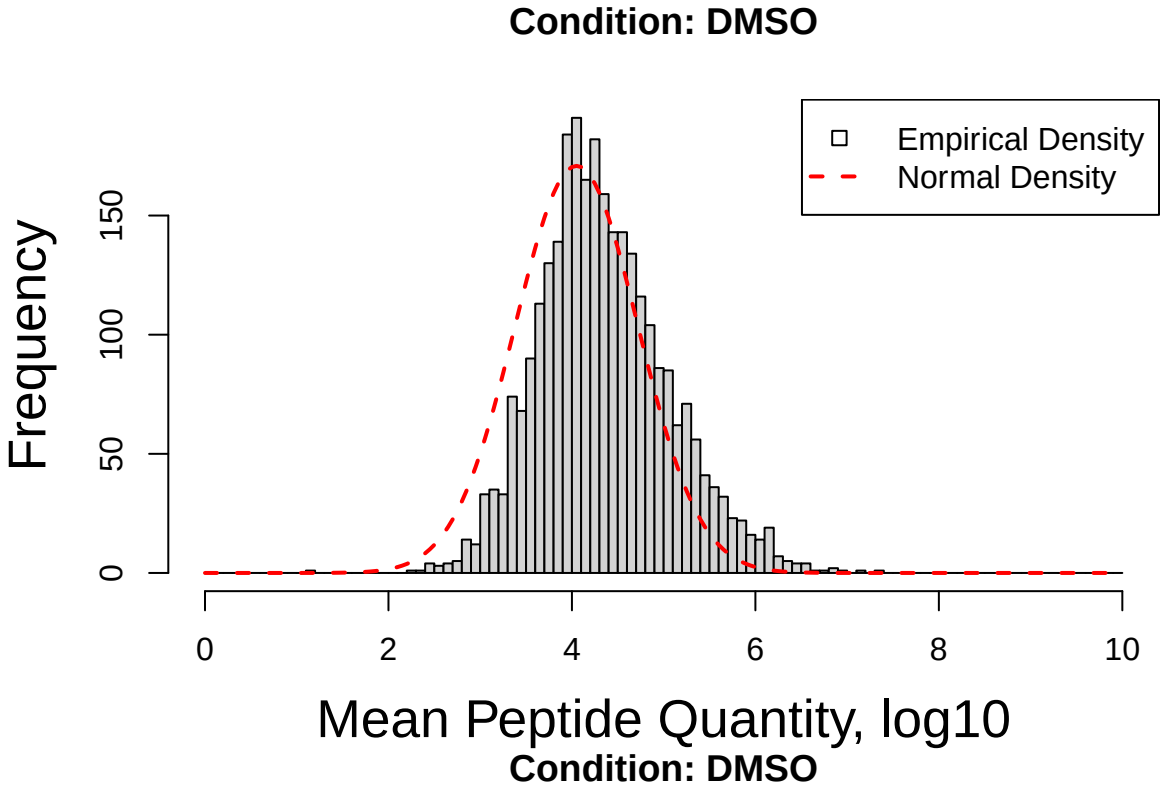
Distribution of precursor quantity: a selected precursor across all conditions.

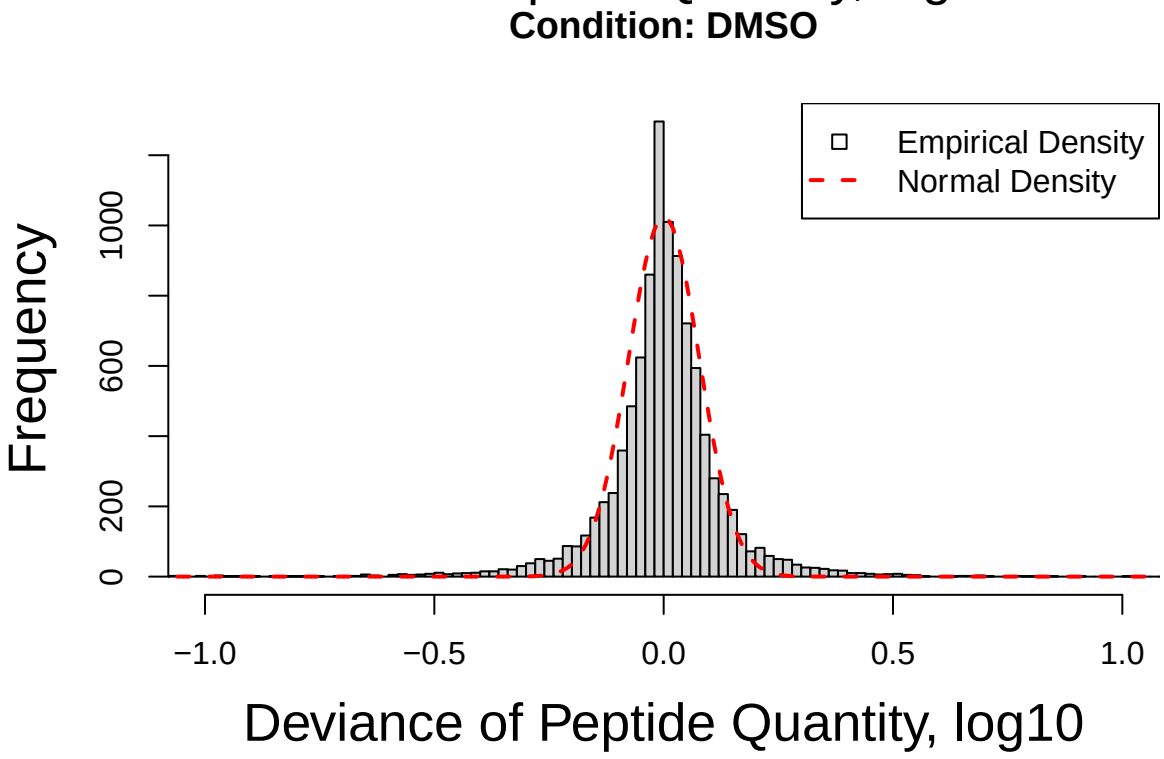


Sampling Distribution for Hierarchical Model

Distribution of precursor mean and deviance:

- precursor mean: all precursors averaged over replicates under fixed condition.
- precursor deviance: all precursors subtracted by the mean.

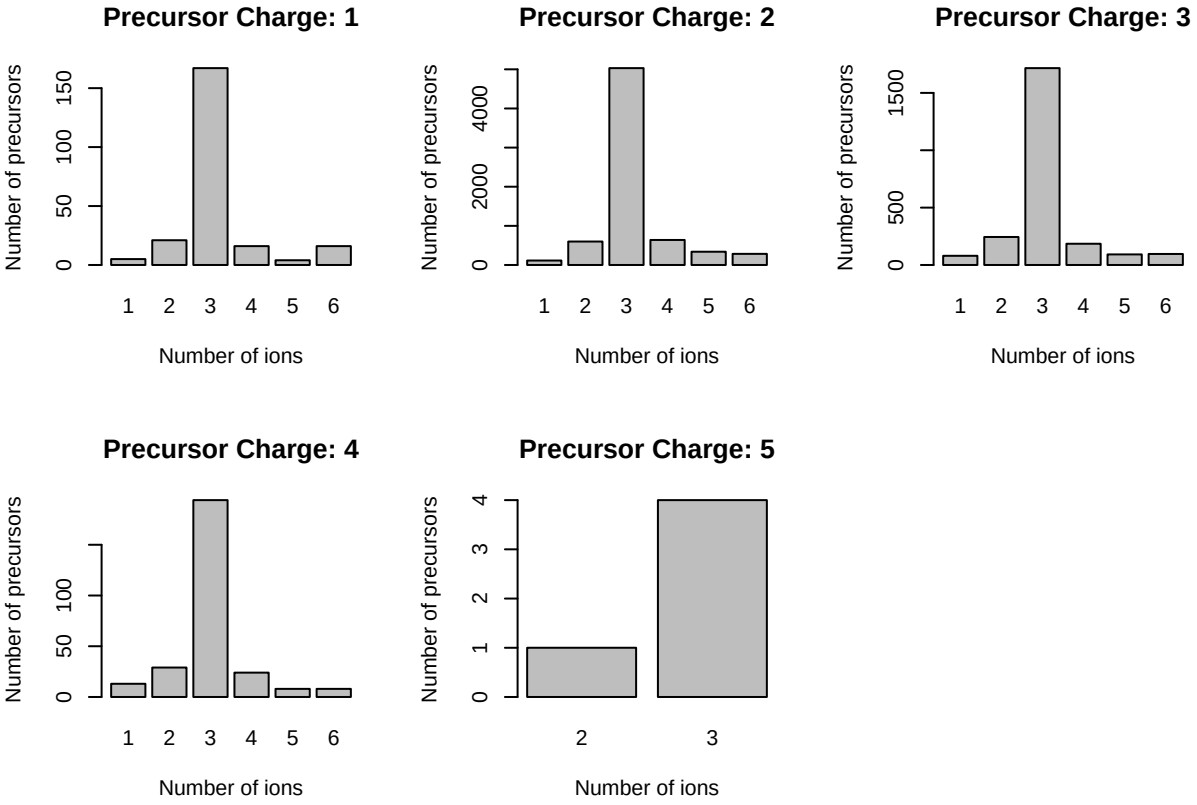




Distribution of Ionization Efficiency

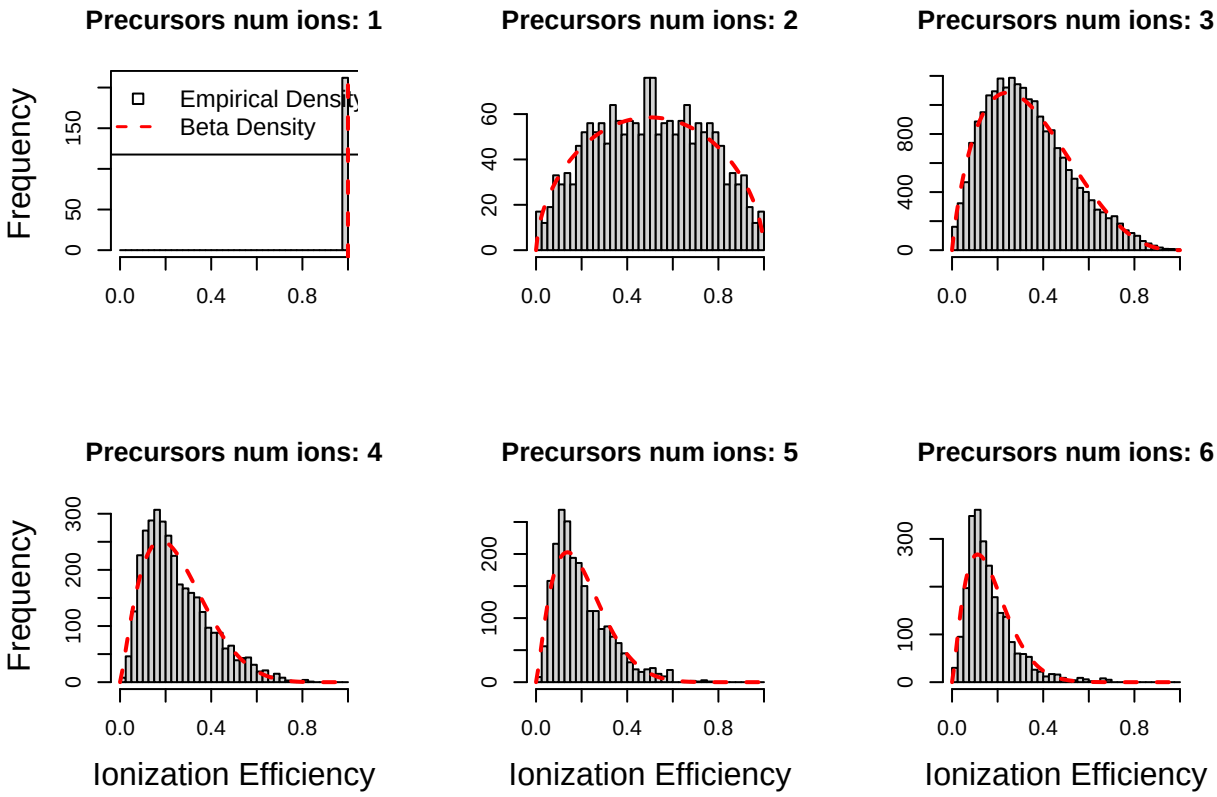
1. Distribution of the number of fragment ions of each precursor

- The number of fragment ions of each precursor is computed. The precursors are separated by their charges to inspect distribution specifically.



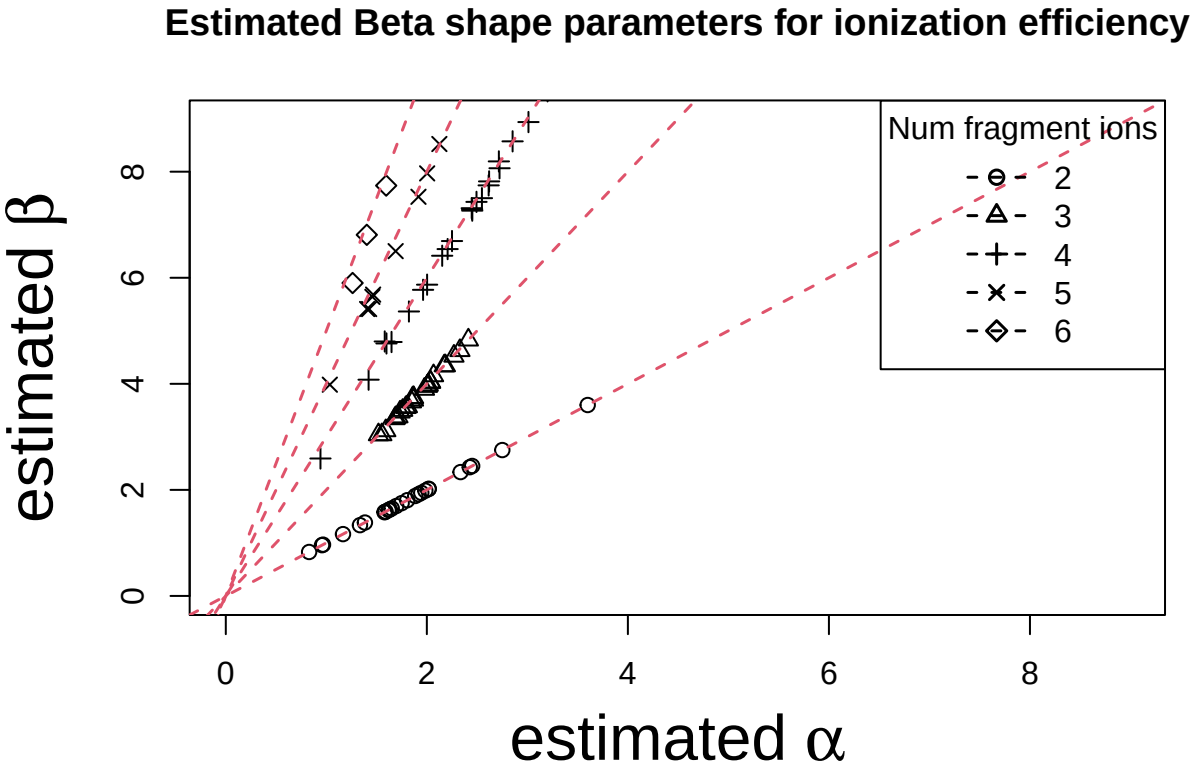
2. Calculate ionization efficiency distribution

- A fragment ion quantity is calculated by its normalized peak area.
- The ionization efficiency of a fragment ion is the proportion of the fragment ion quantity out of the sum of the fragment ion quantities in its corresponding precursor.
- For analysis, the precursors are separated by their numbers of fragment ions to inspect distribution specifically.



3. Fit Beta distribution to an ionization efficiency distribution

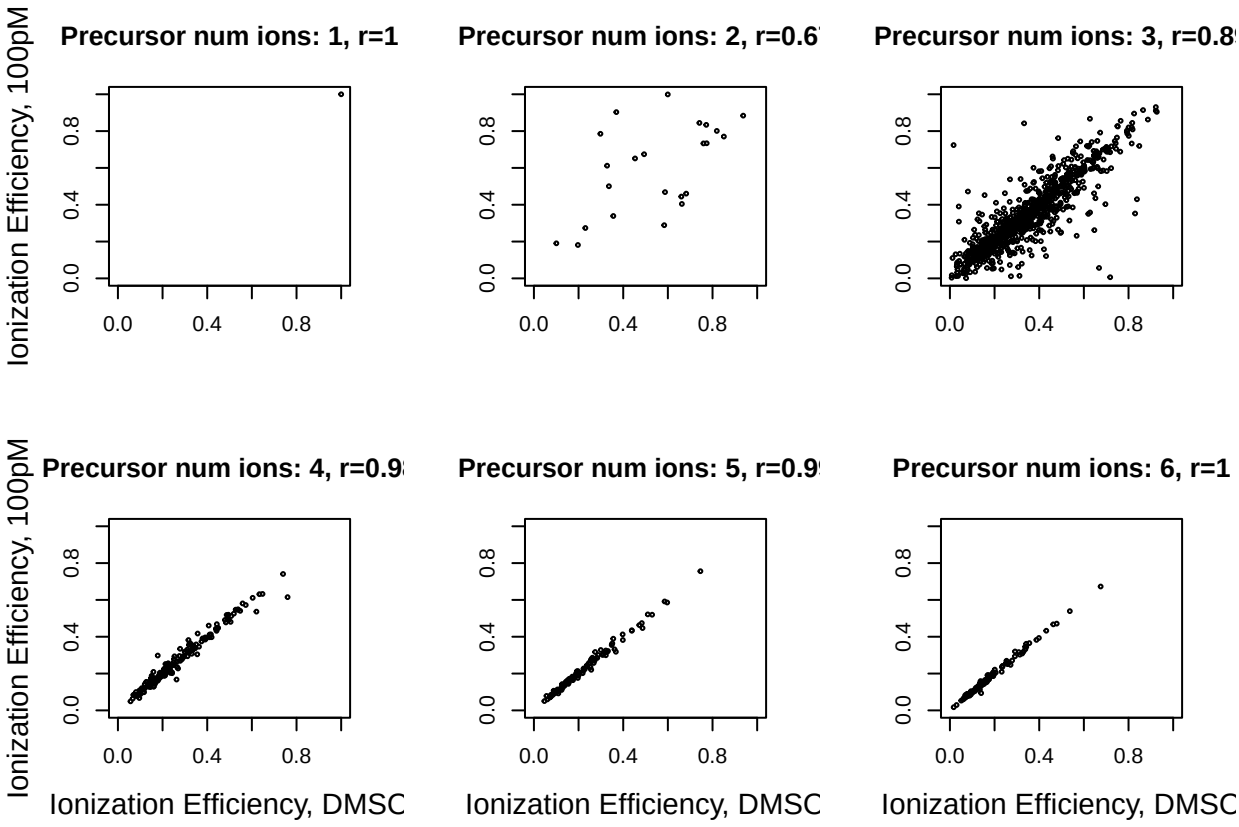
- A beta distribution has two shape parameters, alpha and beta. The mean of a Beta distribution is equal to alpha / (alpha + beta).
- In this analysis, 30 proteins are randomly selected among the proteins having at least 30 peptides. For each protein, alpha and beta are estimated by ionization efficiency distribution of its fragment ions.



- In the figure, each dot is a protein, and there are 30 points on each straight line.
- The dashed lines are $y = x$, $y = 2x$, ..., $y = 5x$.
- The figure shows that the estimated (alpha, beta) values agree with the dashed lines.

4. Correlation of Ionization Efficiency Between Conditions

- Investigate correlation between fragment ion quantities of two conditions with the same precursors, fragment ions, and replicate


Precursor num ions: 1, r=1
Precursor num ions: 2, r=0.6
Precursor num ions: 3, r=0.8
Precursor num ions: 4, r=0.9
Precursor num ions: 5, r=0.9
Precursor num ions: 6, r=1
Ionization Efficiency, 100pM
Ionization Efficiency, DMSC
0.0
0.4
0.8