TMT16plex titration dataset - FragPipe search processed on Percolator - Evaluate performance of FDR correction on subset

Miguel Cosenza

14 October, 2021

Table of Contents

[1 Load data 2](#_Toc85104746)

[2 Questions/aims: 2](#_Toc85104747)

[3 Processing 3](#_Toc85104748)

[3.1 Annotate peptides 3](#_Toc85104749)

[3.2 Label peptide with free N-terminus (no TMT at N-terminus) 3](#_Toc85104750)

[3.3 Merge TMT-integrator report with annotated peptide IDs 4](#_Toc85104751)

[3.4 Label peptides as E.coli or human 4](#_Toc85104752)

[4 Summarized count results 5](#_Toc85104753)

[5 Distribution of E.coli quant per feature category 6](#_Toc85104754)

[6 Differential Abundance analysis performance 9](#_Toc85104755)

[6.0.1 Prep abundance matrix and design matrix and comparisons 9](#_Toc85104756)

[6.0.2 Prep annotation table mapping peptide/index to specificity category 10](#_Toc85104757)

[6.0.3 Prep PSM count information for weighted correction and DEqMS 10](#_Toc85104758)

[6.1 Approach 1 11](#_Toc85104759)

[6.2 Approach 2 12](#_Toc85104760)

[6.3 Approach 3 12](#_Toc85104761)

[6.4 Approach 4 13](#_Toc85104762)

[6.5 Approach 5 13](#_Toc85104763)

[6.6 Approach 5.2 14](#_Toc85104764)

[6.7 Approach 6 14](#_Toc85104765)

[6.8 Approach 7 15](#_Toc85104766)

[6.9 Approach 8 16](#_Toc85104767)

[6.10 Approach 9 16](#_Toc85104768)

[6.11 Approach 10 17](#_Toc85104769)

[7 Distribution of adjusted p-values per approach 17](#_Toc85104770)

[7.1 Observations: 18](#_Toc85104771)

# 1 Load data

tmt\_75\_pept <- read\_tsv(here("data/tmt16plex\_variable-Nter\_0-75\_peptide\_lev\_norm\_fp\_v16\_msfrag3.3\_percolator\_shuffled\_tag/tmt-report/abundance\_peptide\_MD.tsv")) %>%  
 clean\_names()  
  
tmt\_75\_prot <- read\_tsv(here("data/tmt16plex\_variable-Nter\_0-75\_peptide\_lev\_norm\_fp\_v16\_msfrag3.3\_percolator\_shuffled\_tag/tmt-report/abundance\_protein\_MD.tsv")) %>%   
 clean\_names()  
  
ids\_75\_pept <- read\_tsv(here("data/tmt16plex\_variable-Nter\_0-75\_peptide\_lev\_norm\_fp\_v16\_msfrag3.3\_percolator\_shuffled\_tag/MC-FragPipe/peptide.tsv")) %>%   
 clean\_names() %>%   
 mutate(Peptide = peptide, Genes = protein\_id)  
  
ids\_75\_prot <- read\_tsv(here("data/tmt16plex\_variable-Nter\_0-75\_peptide\_lev\_norm\_fp\_v16\_msfrag3.3\_percolator\_shuffled\_tag/MC-FragPipe/protein.tsv")) %>%   
 clean\_names()  
  
psms\_75 <- read\_tsv(here("data/tmt16plex\_variable-Nter\_0-75\_peptide\_lev\_norm\_fp\_v16\_msfrag3.3\_percolator\_shuffled\_tag/MC-FragPipe/psm.tsv")) %>%   
 clean\_names()  
  
fasta75 <- read.fasta(here("data/tmt16plex\_variable-Nter\_0-75\_peptide\_lev\_norm\_fp\_v16\_msfrag3.3\_percolator\_shuffled\_tag/MC-FragPipe/protein.fas"),  
 seqtype = "AA", as.string = TRUE)

# 2 Questions/aims:

This report is intended to explore and show how to extract (filter-in) N-terminal peptides from FragPipe and annotate as potentially canonical or non-canonical based on Uniprot processing information.

#### 2.0.0.1 Objectives:

* Extract N-terminal peptides from FragPipe output.
* Extract annotated processing information from Uniprot.
* Use this information to mark the identified N-terminal peptides as potentially canonical or not.

# 3 Processing

## 3.1 Annotate peptides

## PIF 0.75   
  
if(!file.exists(here("results/peptide\_annotation\_pif75\_32search.tsv"))){  
   
 peptide\_annot\_75 <- annotate\_peptides(ids\_75\_pept, fasta75)  
   
 write\_tsv(peptide\_annot\_75, here("results/peptide\_annotation\_pif75\_32search.tsv"))  
   
} else {  
   
 peptide\_annot\_75 <- read\_tsv(here("results/peptide\_annotation\_pif75\_32search.tsv"))  
   
}  
  
protein\_length <- ids\_75\_prot %>%  
 dplyr::select(protein\_id, length)  
  
pept\_75\_annot <- left\_join(ids\_75\_pept, peptide\_annot\_75,  
 by = c("protein\_id", "Peptide")) %>%  
 left\_join(., protein\_length)

## 3.2 Label peptide with free N-terminus (no TMT at N-terminus)

nterannot\_75 <- pept\_75\_annot %>%   
 mutate(nterm = case\_when(str\_detect(assigned\_modifications, "N-term\\(304.2072\\)") ~ "TMT-labelled",  
 str\_detect(assigned\_modifications, "N-term\\(42.0106\\)") ~ "acetylated",  
 TRUE ~ "free")) %>%  
 mutate(tmt\_tag = case\_when(str\_detect(assigned\_modifications, "N-term\\(304.2072\\)") ~ "nterm",  
 str\_detect(assigned\_modifications, "K\\(304.2072\\)") ~ "lysine",  
 str\_detect(assigned\_modifications, "K\\(304.2072\\)",  
 negate = TRUE) & str\_detect(assigned\_modifications, "N-term\\(42.0106\\)") ~ "untagged\_acetylated",  
 str\_detect(assigned\_modifications, "K\\(304.2072\\)",  
 negate = TRUE) & nterm == "acetylated" ~ "untagged\_acetylated",  
 str\_detect(assigned\_modifications, "K\\(304.2072\\)",  
 negate = TRUE) & nterm == "free" ~ "untagged\_free",  
 TRUE ~ "untagged")) %>%  
 mutate(specificity = case\_when(nterm == "acetylated" & str\_detect(last\_aa, "R|K") ~ "tryptic",  
 nterm == "acetylated" & str\_detect(last\_aa, "R|K", negate = TRUE) ~ "unspecific",  
 TRUE ~ specificity),  
 semi\_type = case\_when(nterm == "acetylated" & str\_detect(last\_aa, "R|K") ~ "tryptic\_nterm",  
 nterm == "acetylated" & str\_detect(last\_aa, "R|K", negate = TRUE) ~ "unspecific\_nterm",  
 TRUE ~ semi\_type)) %>%   
 dplyr::select(-c(matches("hek\_"), matches("x1"))) # eliminate columns with quant information (not normalized)

## 3.3 Merge TMT-integrator report with annotated peptide IDs

## PIF 0.75  
tmt\_reprt\_75\_annot <- left\_join(tmt\_75\_pept, nterannot\_75)  
  
excluded\_tmt\_reprt\_75\_annot <- anti\_join(nterannot\_75, tmt\_75\_pept)

## 3.4 Label peptides as E.coli or human

zero2na <- function(x){  
 y <- ifelse(x == 0,  
 yes = NA,  
 no = x)  
   
 return(y)  
}  
  
#zero2na(0.223)   
  
prequant\_75 <- tmt\_reprt\_75\_annot %>%   
 mutate(human\_ecoli = ifelse(test = str\_detect(entry\_name, "HUMAN"),  
 yes = "human",  
 no = "ecoli")) %>%  
 mutate(across(matches("x1\_"),zero2na)) %>%  
 dplyr::select(peptide, matches("^x1\_"), peptide, specificity, nterm, semi\_type, tmt\_tag, human\_ecoli) %>%  
 na.omit() %>%  
 rowwise() %>%  
 mutate(`FC\_0.06/0.02` = mean(x1\_17\_8, x1\_17\_10, x1\_17\_12, x1\_17\_14, na.rm = TRUE)/mean(x1\_50\_15, x1\_50\_17, x1\_50\_19, x1\_50\_21, na.rm = TRUE),  
 `FC\_0.14/0.06` = mean(x1\_7\_16, x1\_7\_18, x1\_7\_20,x1\_7\_22, na.rm = TRUE)/mean(x1\_17\_8, x1\_17\_10, x1\_17\_12, x1\_17\_14, na.rm = TRUE),  
 `FC\_0.14/0.02` = mean(x1\_7\_16, x1\_7\_18, x1\_7\_20,x1\_7\_22,na.rm = TRUE)/mean(x1\_50\_15, x1\_50\_17, x1\_50\_19, x1\_50\_21, na.rm = TRUE))  
  
  
fc\_quant\_eval1\_75 <- dplyr::select(prequant\_75,  
 peptide, specificity, nterm, semi\_type, tmt\_tag, human\_ecoli,  
 `FC\_0.06/0.02`, `FC\_0.14/0.06`,`FC\_0.14/0.02`) %>%   
 pivot\_longer(cols = c(`FC\_0.06/0.02`, `FC\_0.14/0.06`,`FC\_0.14/0.02`),  
 values\_to = "FCs",  
 names\_to = "Ratios") %>%  
 pivot\_longer(cols = c(specificity, nterm, semi\_type, tmt\_tag),  
 values\_to = "feature\_type",  
 names\_to = "category") %>%  
 mutate(FCs = ifelse(is.nan(FCs),  
 yes = NA,  
 no = FCs),  
 Expected\_ratio = str\_remove(Ratios, "^FC\_")) %>%  
 mutate(Expected\_FC = eval(parse(text = Expected\_ratio))) %>%  
 mutate(log2\_FC = log2(FCs),  
 log2\_Expected\_FC = log2(Expected\_FC))  
  
df2 <- fc\_quant\_eval1\_75 %>%  
 group\_by(feature\_type, Ratios) %>%  
 summarise(Expected = log2\_Expected\_FC)

# PIF 0.75  
to\_count\_info\_75 <- prequant\_75 %>%   
 dplyr::select(peptide, specificity, nterm, semi\_type, tmt\_tag)  
  
to\_count\_info\_acetyl <- to\_count\_info\_75 %>%   
 filter(nterm == "acetylated")

# 4 Summarized count results

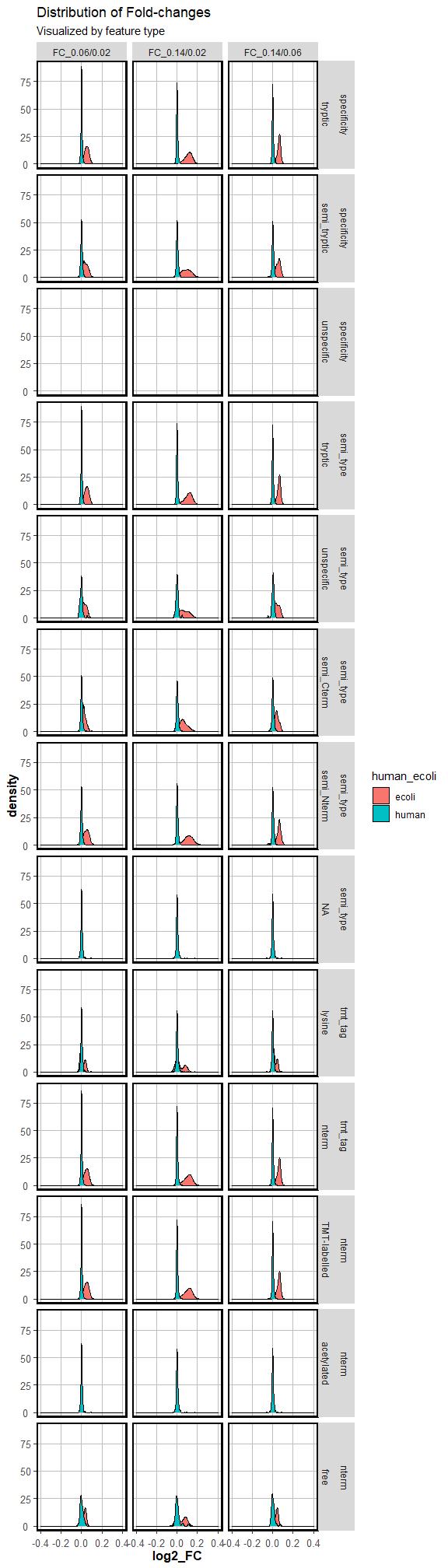
## PIF 75  
n\_semi <- dplyr::count(to\_count\_info\_75, specificity) %>%   
 dplyr::rename(feature\_type = specificity) %>%  
 dplyr::mutate(category = "specificity")  
  
n\_term <- dplyr::count(to\_count\_info\_75, nterm) %>%   
 dplyr::rename(feature\_type = nterm) %>%  
 dplyr::mutate(category = "N-term")  
  
n\_semi\_type <- dplyr::count(to\_count\_info\_75, semi\_type) %>%   
 dplyr::rename(feature\_type = semi\_type) %>%  
 dplyr::mutate(category = "Semi type")  
  
n\_tmt\_tag <- dplyr::count(to\_count\_info\_75, tmt\_tag) %>%   
 dplyr::rename(feature\_type = tmt\_tag) %>%  
 dplyr::mutate(category = "TMT location")  
  
n\_total <- tibble(feature\_type = "Total",  
 n = nrow(to\_count\_info\_75),  
 category = "Total")  
   
summary\_count\_75 <- bind\_rows(n\_semi,  
 n\_term,  
 n\_semi\_type,  
 n\_tmt\_tag,  
 n\_total)

# 5 Distribution of E.coli quant per feature category

These quant values were extracted based on the list of peptides summarized after TMT-integrator filtering.

fc\_quant\_eval2\_75 <- mutate(fc\_quant\_eval1\_75,  
 feature\_type = factor(feature\_type,  
 levels = c("tryptic","semi\_tryptic", "unspecific",  
 "semi\_Cterm", "semi\_Nterm",  
 "TMT-labelled", "acetylated","free",  
 "lysine", "nterm")),  
 category = factor(category,  
 levels = c("specificity", "semi\_type",  
 "tmt\_tag", "nterm")))  
  
df2 <- fc\_quant\_eval2\_75 %>% # this is generating a df that can be used to plot the expected log2 ratio  
 group\_by(feature\_type, Ratios) %>%  
 summarise(Expected = log2\_Expected\_FC)

ggplot(fc\_quant\_eval2\_75,  
 aes(x = log2\_FC, fill = human\_ecoli)) +   
 geom\_density() +  
 #geom\_vline(data = df2, mapping = aes(xintercept = Expected), color = "red",  
 # linetype = "dashed") +  
 facet\_grid(category+feature\_type~Ratios) +  
 labs(title = "Distribution of Fold-changes",  
 subtitle = "Visualized by feature type") +  
 theme(axis.text.x = element\_text(hjust = 0.5, vjust = 0.1, size = 10),  
 axis.text.y = element\_text(hjust = 0.5, size = 10),  
 panel.background = element\_blank(),  
 panel.grid.major = element\_line(color = "grey"),  
 panel.border = element\_rect(colour = "black", fill=NA, size=1.5),  
 axis.title=element\_text(size=12,face="bold"))



# 6 Differential Abundance analysis performance

This section is intended to evaluate the performance for differential abundant semi-tryptic peptides based on different approaches.

1. Limma Robust on whole dataset + Global FDR correction on all peptides.
2. Limma Robust on whole dataset + FDR only on semi-tryptic peptides.
3. Limma Robust on whole dataset + weighted FDR correction on all peptides.
4. Limma Robust on whole dataset + weighted FDR on semi-tryptic peptides.
5. Subset semi-tryptic peptides + Limma Robust (Global FDR correction, on the subset)
6. DEqMS on whole dataset + Global FDR correction.
7. DEqMS on whole dataset + FDR only on semi-tryptic peptides.
8. DEqMS on whole dataset + weighted FDR correction on all peptides.
9. DEqMS on whole dataset + weighted FDR on semi-tryptic peptides.
10. Subset semi-tryptic peptides + DEqMS (Global FDR correction, on the subset).

### 6.0.1 Prep abundance matrix and design matrix and comparisons

For this test, I will focus on the comparison that yield the biggest expected theoretical differences: 0.16 vs 0.02 (*E. coli*:HEK ratios).

The subset of peptides will be based on N-term peptides because these show the best separation in the density plots.

tomat <- tmt\_75\_pept %>%  
 dplyr::select(index, where(is.numeric)) %>%  
 dplyr::select(-c(max\_pep\_prob, reference\_intensity)) %>%  
 dplyr::select(-starts\_with("hek\_only")) %>%  
 dplyr::select(index, starts\_with("x1\_7"), starts\_with("x1\_50")) %>%   
 na.omit()  
  
tomat2 <- dplyr::rename(tomat, ID = index)  
  
write\_delim(tomat2, file = here("data/input\_limma.txt"), delim = "\t")  
  
mat <- tomat %>%  
 column\_to\_rownames("index") %>%  
 as.matrix()  
  
ratio <- colnames(mat) %>%  
 str\_sub(end = -4)  
  
simple\_annotation <- tibble(Sample\_ID = colnames(mat),  
 Group = ratio)  
  
write\_delim(simple\_annotation, file = here("data/annotation.txt"), delim = "\t")  
  
  
design <- model.matrix(~ratio)  
row.names(design) <- colnames(mat)

### 6.0.2 Prep annotation table mapping peptide/index to specificity category

An annotation table is necessary for this.

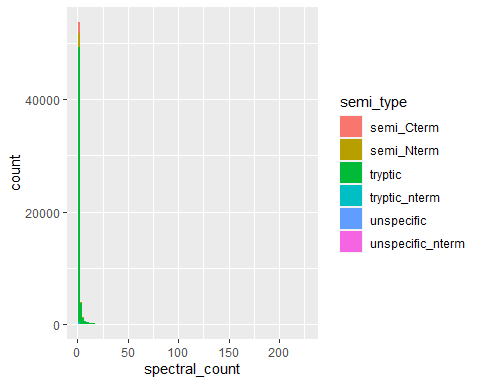
index2semitype <- tmt\_reprt\_75\_annot %>%  
 dplyr::select(index, peptide, semi\_type)

### 6.0.3 Prep PSM count information for weighted correction and DEqMS

peptide2spectralcount <- ids\_75\_pept %>%  
 dplyr::select(peptide, spectral\_count)  
   
annotation\_correction <- left\_join(index2semitype, peptide2spectralcount)  
  
annotation\_correction2 <- dplyr::select(annotation\_correction,  
 ID = index, feature\_category = semi\_type)  
  
write\_delim(annotation\_correction2, file = here("data/annotation\_features.txt"), delim = "\t")

#### 6.0.3.1 Distribution of spectral counts per peptide

ggplot(annotation\_correction,  
 aes(x = spectral\_count, fill = semi\_type)) +   
 geom\_histogram(binwidth = 2)



Semi tryptic peptides tend to have lower spectral counts, represented by the higher amount of peptides with count 1-2.

## 6.1 Approach 1

Limma Robust on whole dataset + Global FDR correction on all peptides.

fit1 <- lmFit(mat, # expression matrix  
 design = design, # design matrix  
 method = "robust", # set to robust regression   
 maxit = 50)   
  
fit1 <- eBayes(fit1)  
  
fit\_tab1 <- topTable(fit1,   
 coef = "ratiox1\_7",   
 number = Inf,   
 adjust.method = "BH") %>% # FDR correction of DE analysis happens here  
 mutate(algorithm = "Limma",  
 correction = "bh\_global") %>% # just adding a column indicating which model was applied.  
 rownames\_to\_column("index")  
  
fit\_tab1\_2 <- left\_join(fit\_tab1, annotation\_correction)

fit\_tab1\_3 <- filter(fit\_tab1\_2,  
 semi\_type == "semi\_Nterm")

## 6.2 Approach 2

Limma Robust on whole dataset + FDR only on semi-tryptic peptides.

fit\_tab2 <- topTable(fit1,   
 coef = "ratiox1\_7",   
 number = Inf,   
 adjust.method = "BH") %>% # FDR correction of DE analysis happens here  
 mutate(algorithm = "Limma",  
 correction = "bh\_subset") %>% # just adding a column indicating which model was applied.  
 rownames\_to\_column("index")  
  
fit\_tab2\_2 <- left\_join(fit\_tab2, annotation\_correction)

#### 6.2.0.1 Apply adjustment only on subset

fit\_tab2\_3 <- filter(fit\_tab2\_2,  
 semi\_type == "semi\_Nterm") %>%  
 mutate(adj.P.Val = p.adjust(p = P.Value, method = "BH"))

## 6.3 Approach 3

Limma Robust on whole dataset + weighted FDR global correction.

fit\_tab3 <- topTable(fit1,   
 coef = "ratiox1\_7",   
 number = Inf,   
 adjust.method = "BH") %>% # FDR correction of DE analysis happens here  
 mutate(algorithm = "Limma",  
 correction = "ihw\_global") %>% # just adding a column indicating which model was applied.  
 rownames\_to\_column("index")  
  
fit\_tab3\_2 <- left\_join(fit\_tab3, annotation\_correction)

ihw\_corr\_3 <- ihw(P.Value ~ spectral\_count, data = fit\_tab3\_2, alpha = 0.05)

ihw\_corr\_3\_df <- dplyr::rename(ihw\_corr\_3@df,  
 adj.P.Val = adj\_pvalue,  
 P.Value = pvalue)  
  
fit\_tab3\_21 <- fit\_tab3\_2 %>% dplyr::select(-adj.P.Val)  
  
fit\_tab3\_22 <- left\_join(fit\_tab3\_21, ihw\_corr\_3\_df, by = "P.Value")  
  
fit\_tab3\_3 <- filter(fit\_tab3\_22,  
 semi\_type == "semi\_Nterm")  
  
fit\_tab3\_4 <- dplyr::select(fit\_tab3\_3,  
 names(fit\_tab2\_3))

## 6.4 Approach 4

Limma Robust on whole dataset + weighted FDR on semi-tryptic peptides.

fit\_tab4\_2 <- fit\_tab3\_2 %>%  
 mutate(correction = "ihw\_subset") %>%  
 filter(semi\_type == "semi\_Nterm")

ihw\_corr\_4 <- ihw(P.Value ~ spectral\_count, data = fit\_tab4\_2, alpha = 0.05)

fit\_tab4\_3 <- dplyr::select(fit\_tab4\_2,  
 names(fit\_tab2\_3))

## 6.5 Approach 5

Subset semi-tryptic peptides + Limma Robust (Global FDR correction, on the subset)

**Filter abundance matrix**

semi\_peptides <- annotation\_correction %>%  
 filter(semi\_type == "semi\_Nterm") %>%  
 pull(index)  
  
tomat\_fil <- tomat %>%  
 filter(index %in% semi\_peptides)  
  
mat\_fil <- tomat\_fil %>%  
 column\_to\_rownames("index") %>%  
 as.matrix()

**Limma fit**

fit5 <- lmFit(mat\_fil, # expression matrix  
 design = design, # design matrix  
 method = "robust", # set to robust regression   
 maxit = 50)   
  
fit5 <- eBayes(fit5)  
  
fit\_tab5 <- topTable(fit5,   
 coef = "ratiox1\_7",   
 number = Inf,   
 adjust.method = "BH") %>% # FDR correction of DE analysis happens here  
 mutate(algorithm = "Limma",  
 correction = "prefiltered\_bh") %>% # just adding a column indicating which model was applied.  
 rownames\_to\_column("index")  
  
fit\_tab5\_2 <- left\_join(fit\_tab5, annotation\_correction)

## 6.6 Approach 5.2

Subset semi-tryptic peptides + Limma Robust + Weighted FDR correction (on the subset)

fit\_tab52\_2 <- fit\_tab5\_2 %>%  
 mutate(correction = "prefiltered\_ihw") %>%  
 filter(semi\_type == "semi\_Nterm")

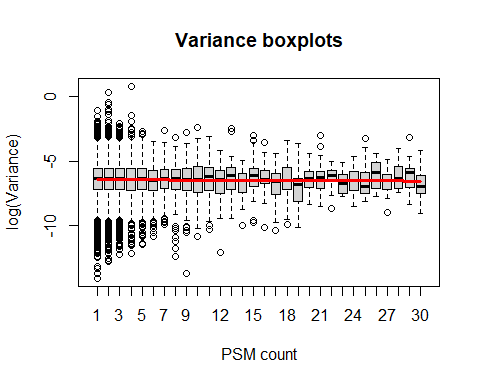
ihw\_corr\_52 <- ihw(P.Value ~ spectral\_count, data = fit\_tab52\_2, alpha = 0.05)

## 6.7 Approach 6

DEqMS on whole dataset + Global FDR correction.

fit6 <- fit1  
  
fit6\_pretab <- topTable(fit6,   
 coef = "ratiox1\_7",   
 number = Inf,   
 adjust.method = "BH") %>%  
 rownames\_to\_column("index") %>%  
 left\_join(., annotation\_correction)  
  
fit6$count <- fit6\_pretab$spectral\_count  
  
fit6 = spectraCounteBayes(fit6)  
  
fit\_tab6 <- topTable(fit6,   
 coef = "ratiox1\_7",   
 number = Inf,   
 adjust.method = "BH") %>% # FDR correction of DE analysis happens here  
 mutate(algorithm = "deqms",  
 correction = "bh\_global") %>% # just adding a column indicating which model was applied.  
 rownames\_to\_column("index")   
  
fit\_tab6\_2 <- left\_join(fit\_tab6, annotation\_correction)

VarianceBoxplot(fit6,n=30,main="Variance boxplots",xlab="PSM count")



DEqMS is developed for protein quant and using PSM counts to weight the protein variance in limma [DEqMS vignette](https://bioconductor.org/packages/release/bioc/vignettes/DEqMS/inst/doc/DEqMS-package-vignette.html)…

In terms of peptides (here), the number of spectral counts does not seem to have a consistent impact on variance.

fit\_tab6\_3 <- filter(fit\_tab6\_2,  
 semi\_type == "semi\_Nterm")

## 6.8 Approach 7

DEqMS on whole dataset + FDR only on semi-tryptic peptides.

fit\_tab7 <- topTable(fit6,   
 coef = "ratiox1\_7",   
 number = Inf,   
 adjust.method = "BH") %>% # FDR correction of DE analysis happens here  
 mutate(algorithm = "deqms",  
 correction = "bh\_subset") %>% # just adding a column indicating which model was applied.  
 rownames\_to\_column("index")   
  
fit\_tab7\_2 <- left\_join(fit\_tab7, annotation\_correction)

#### 6.8.0.1 Apply adjustment only on subset

fit\_tab7\_3 <- filter(fit\_tab7\_2,  
 semi\_type == "semi\_Nterm") %>%  
 mutate(adj.P.Val = p.adjust(p = P.Value, method = "BH"))

## 6.9 Approach 8

DEqMS on whole dataset + weighted FDR correction on all peptides.

fit\_tab8 <- topTable(fit6,   
 coef = "ratiox1\_7",   
 number = Inf,   
 adjust.method = "BH") %>% # FDR correction of DE analysis happens here  
 mutate(algorithm = "deqms",  
 correction = "ihw\_global") %>% # just adding a column indicating which model was applied.  
 rownames\_to\_column("index")  
  
fit\_tab8\_2 <- left\_join(fit\_tab8, annotation\_correction)

ihw\_corr\_8 <- ihw(P.Value ~ spectral\_count, data = fit\_tab8\_2, alpha = 0.05)

ihw\_corr\_8\_df <- dplyr::rename(ihw\_corr\_8@df,  
 adj.P.Val = adj\_pvalue,  
 P.Value = pvalue)  
  
fit\_tab8\_21 <- fit\_tab8\_2 %>% dplyr::select(-adj.P.Val)  
  
fit\_tab8\_22 <- left\_join(fit\_tab8\_21, ihw\_corr\_8\_df, by = "P.Value")  
  
fit\_tab8\_3 <- filter(fit\_tab8\_22,  
 semi\_type == "semi\_Nterm")   
  
fit\_tab8\_4 <- dplyr::select(fit\_tab8\_3,  
 names(fit\_tab7\_2))

## 6.10 Approach 9

DEqMS on whole dataset + weighted FDR on semi-tryptic peptides.

fit\_tab9\_2 <- fit\_tab8\_2 %>%  
 mutate(correction = "ihw\_subset") %>%  
 filter(semi\_type == "semi\_Nterm")

ihw\_corr\_9 <- ihw(P.Value ~ spectral\_count, data = fit\_tab9\_2, alpha = 0.05)

fit\_tab9\_3 <- dplyr::select(fit\_tab9\_2,  
 names(fit\_tab7\_2))

## 6.11 Approach 10

Subset semi-tryptic peptides + DEqMS (Global FDR correction, on the subset).

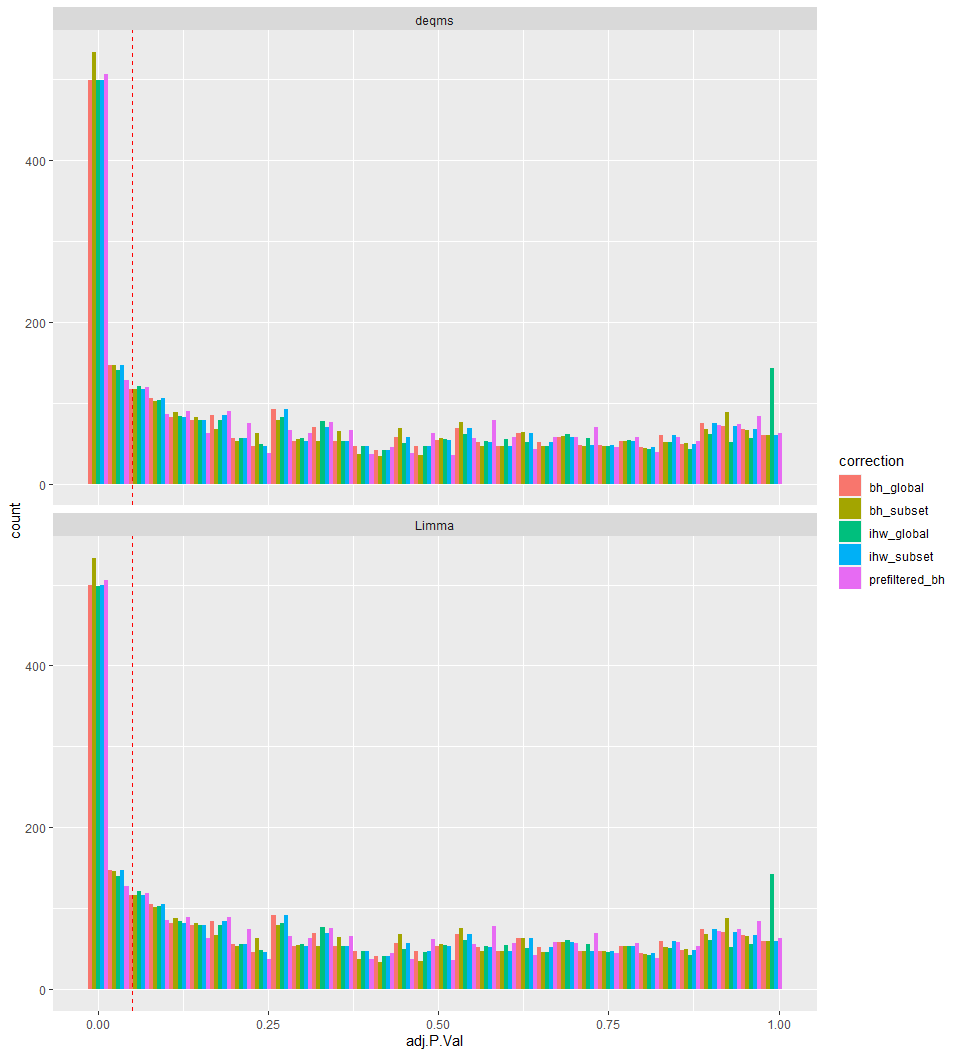
**DEqMS-Limma fit**

fit10 <- fit5  
  
fit10\_pretab <- topTable(fit10,   
 coef = "ratiox1\_7",   
 number = Inf,   
 adjust.method = "BH") %>%  
 rownames\_to\_column("index") %>%  
 left\_join(., annotation\_correction)  
  
fit10$count <- fit10\_pretab$spectral\_count  
  
fit10 = spectraCounteBayes(fit10)  
  
fit\_tab10 <- topTable(fit10,   
 coef = "ratiox1\_7",   
 number = Inf,   
 adjust.method = "BH") %>% # FDR correction of DE analysis happens here  
 mutate(algorithm = "deqms",  
 correction = "prefiltered\_bh") %>% # just adding a column indicating which model was applied.  
 rownames\_to\_column("index")   
  
fit\_tab10\_2 <- left\_join(fit\_tab10, annotation\_correction)

# 7 Distribution of adjusted p-values per approach

all\_semi\_tables <- bind\_rows(fit\_tab1\_3,  
 fit\_tab2\_3,  
 fit\_tab3\_4,  
 fit\_tab4\_3,  
 fit\_tab5\_2,  
 fit\_tab6\_3,  
 fit\_tab7\_3,  
 fit\_tab8\_4,  
 fit\_tab9\_3,  
 fit\_tab10\_2)

ggplot(all\_semi\_tables,  
 aes(x = adj.P.Val, fill = correction)) +   
 geom\_histogram(binwidth = 0.03, position = "dodge") +   
 geom\_vline(xintercept = 0.05, linetype = "dashed", color = "red") +   
 facet\_wrap(~algorithm, ncol = 1)



## 7.1 Observations:

Including peptide spectral counts as weights for limma (DEqMS) does not improve the sensitivity when doing peptide-level differential abundance analysis.

The sensitivity seems to be better when applying the BH correction on a subset of p-values.

Fitting the whole limma model directly on the subset is not necessarily better than the global FDR correction on the whole set of identified values and it is definitely worse than applying FDR correction on the subset after limma on the whole data set.

Using weighted correction is similar than applying global FDR correction.