Analysis of phosphoproteomic data with MSstatsPTM: user's view

Michal Kloc (Bentires lab)

FMI Seminar, Feb 6, 2024

Outline

- Biological background, experimental design
- Data acquisition
- MSstatsPTM workflow
  - PTM site identification + proper formatting
  - Data cleaning, normalisation, imputation, summarization
  - Statistical modeling
- My thoughts

Experimental setup

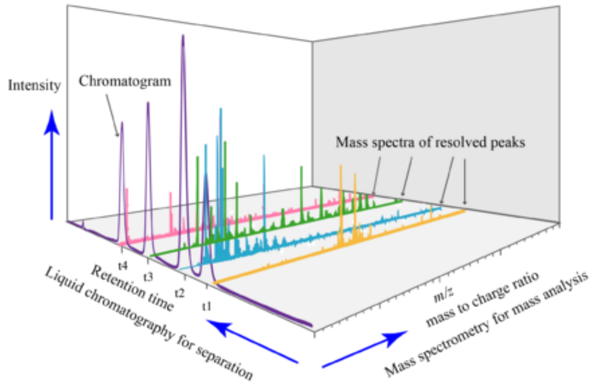
- Goal: molecular mechanisms of drug resistance
- PDX models from ER+ breast cancer patients
- 4 treatment groups: *Fulvestrant, cdk4/6i, Combo* with acquired resistance
- *Control* group from each patient (vehicle “treated”)
- Also initially *resistant patients*
- Multi-layer data sets from the tumor pieces
  RNAseq, proteomics, **phosphoproteomics**
- Each set >70 samples

Phosphoproteomics, LC-MS/MS measurements

Label-free data independent acquisition (DIA), peptide identification via Spectronaut

At the beginning

1) Digestion of the proteins (bottom-up strategy)
2) phosphopeptide enrichment



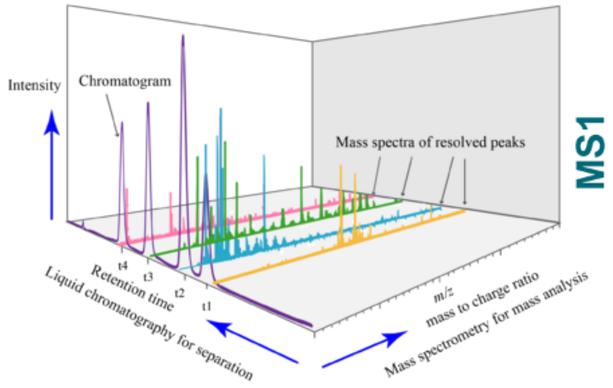
wikipedia.org

Phosphoproteomics, LC-MS/MS measurements

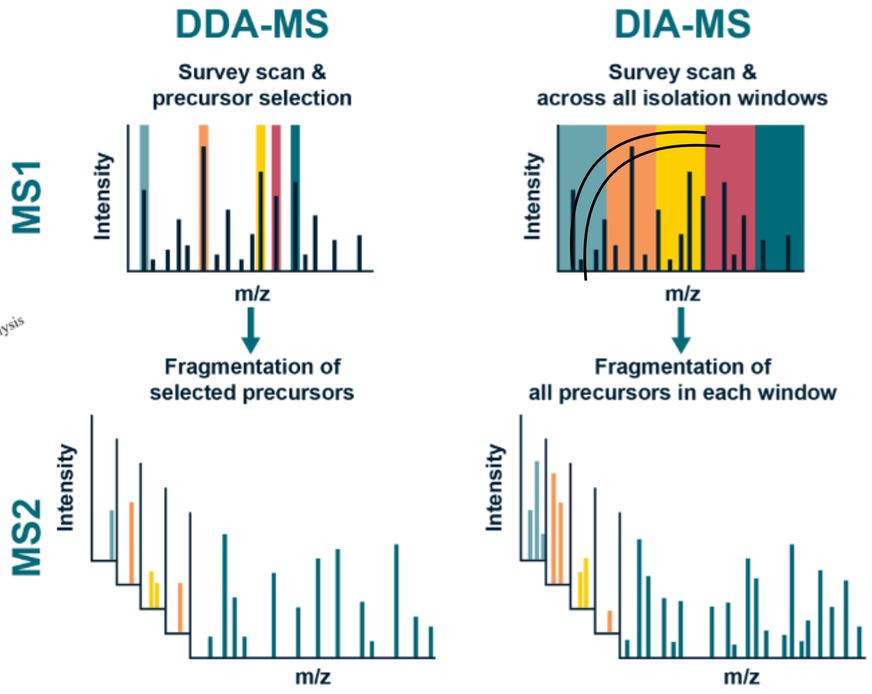
Label-free data independent acquisition (DIA), peptide identification via Spectronaut

At the beginning

1) Digestion of the proteins (bottom-up strategy)
2) phosphopeptide enrichment



wikipedia.org



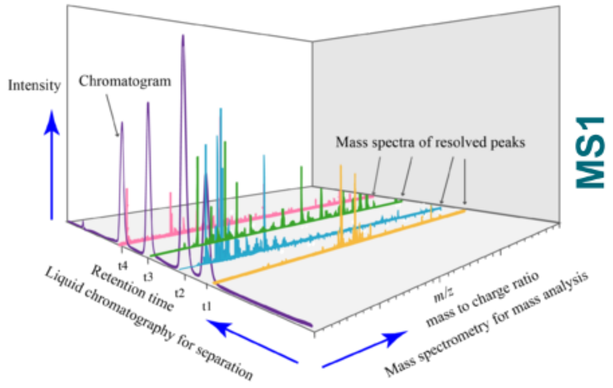
crownbio.com

Phosphoproteomics, LC-MS/MS measurements

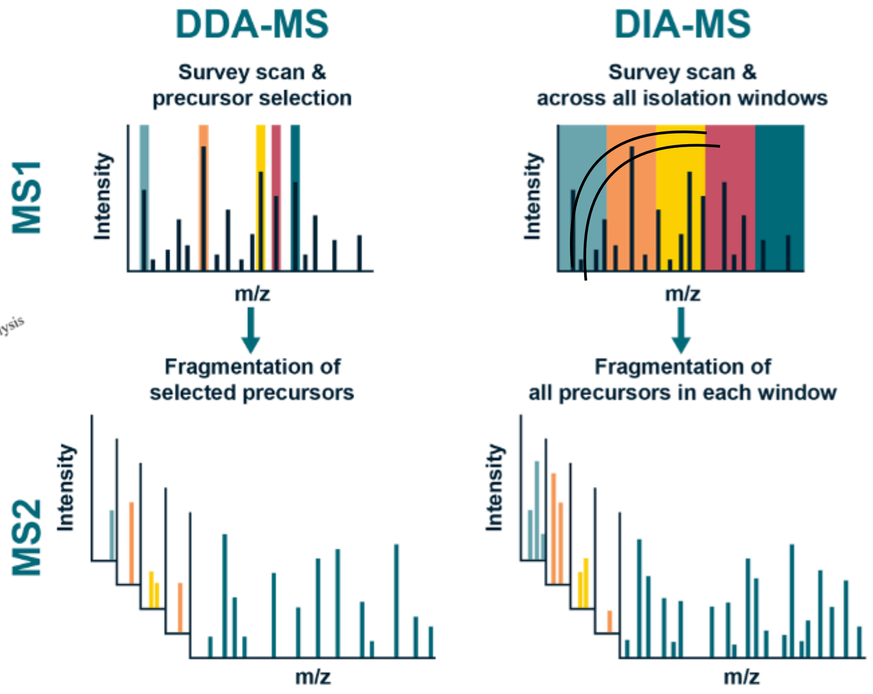
Label-free data independent acquisition (DIA), peptide identification via Spectronaut

At the beginning

1) Digestion of the proteins (bottom-up strategy)
2) phosphopeptide enrichment



wikipedia.org



crownbio.com

Raw Spectronaut Output

```
> colnames(SNoutput)
 [1] "R.FileName"                  "PG.ProteinAccessions"
 [3] "PG.ProteinGroups"            "PG.Cscore"
 [5] "PG.Qvalue"                   "PG.RunEvidenceCount"
 [7] "PG.Database"                 "PG.OrganismId"
 [9] "PG.Protein Existence"        "PG.Sequence Version"
[11] "PG.FASTAName"                "PG.FASTAHeader"
[13] "PEP.Quantity"                "EG.ModifiedSequence"
[15] "EG.Qvalue"                   "FG.Charge"
[17] "F.Charge"                    "F.FrgIon"
[19] "F.FrgLossType"               "F.ExcludedFromQuantification"
[21] "F.PeakArea"
```

MSstatsPTM

1) Convert to MSstats format

```
PTMraw <- SpectronauttoMSstatsPTMFormat(input = SNoutput,
                                        annotation = annotPTM,
                                        use_unmod_peptides=FALSE,
                                        mod_id = "\\[Phospho \\(STY\\)\\]",
                                        removeProtein_with1Feature = FALSE,
                                       fasta_path = "s_HSapiens_UP000005640_20220222.fasta")
```

```
> annotPTM
# A tibble: 70 × 3
   Run                                          Condition BioReplicate
   <chr>                                        <chr>     <chr>
 1 001_Phospho1_P04_DIA_D22                     Control   BCM.15057
 2 002_Phospho1_P04_DIA_D22                     Control   BCM.15057
 3 003_Phospho1_P04_DIA_D22                     Control   BCM.15057
 4 004_Phospho1_P04_DIA_D22                     Control   BCM.15057
 5 005_Phospho1_P04_DIA_D22                     Ribo      BCM.15057
 6 006_Phospho1_P04_DIA_D22                     Ribo      BCM.15057
 7 007_Phospho1_P04_DIA_20220322194039_D22      Ribo      BCM.15057
 8 008_Phospho1_P04_DIA_20220322215824_D22      Ribo      BCM.15057
 9 009_Phospho1_P04_DIA_D22                     Fulves    BCM.15057
10 010_Phospho1_P04_DIA_D22                     Fulves    BCM.15057
```

MSstatsPTM

1) Convert to MSstats format

```
PTMraw <- SpectronauttoMSstatsPTMFormat(input = SNoutput,
                                        annotation = annotPTM,
                                        use_unmod_peptides=FALSE,
                                        mod_id = "\\[Phospho \\(STY\\)\\]",
                                        removeProtein_with1Feature = FALSE,
                                       fasta_path = "s_HSapiens_UP000005640_20220222.fasta")
```

```
> annotPTM
# A tibble: 70 × 3
   Run                                                Condition BioReplicate
   <chr>                                              <chr>     <chr>
 1 001_Phospho1_P04_DIA_D22                           Control   BCM.15057
 2 002_Phospho1_P04_DIA_D22                           Control   BCM.15057
 3 003_Phospho1_P04_DIA_D22                           Control   BCM.15057
 4 004_Phospho1_P04_DIA_D22                           Control   BCM.15057
 5 005_Phospho1_P04_DIA_D22                           Ribo      BCM.15057
 6 006_Phospho1_P04_DIA_D22                           Ribo      BCM.15057
 7 007_Phospho1_P04_DIA_20220322194039_D22            Ribo      BCM.15057
 8 008_Phospho1_P04_DIA_20220322215824_D22            Ribo      BCM.15057
 9 009_Phospho1_P04_DIA_D22                           Fulves    BCM.15057
10 010_Phospho1_P04_DIA_D22                           Fulves    BCM.15057
```

```
     ProteinName                          PeptideSequence PrecursorCharge
141 Q8WXD9_S1364 AAAAAAAAAAAPPAPPEGAS[Phospho (STY)]PGDSAR               3
142 Q8WXD9_S1364 AAAAAAAAAAAPPAPPEGAS[Phospho (STY)]PGDSAR               3
143 Q8WXD9_S1364 AAAAAAAAAAAPPAPPEGAS[Phospho (STY)]PGDSAR               3
144 Q8WXD9_S1364 AAAAAAAAAAAPPAPPEGAS[Phospho (STY)]PGDSAR               3
145 Q8WXD9_S1364 AAAAAAAAAAAPPAPPEGAS[Phospho (STY)]PGDSAR               3
146 Q8WXD9_S1364 AAAAAAAAAAAPPAPPEGAS[Phospho (STY)]PGDSAR               3
    FragmentIon ProductCharge IsotopeLabelType Condition BioReplicate
141          b6             1                L   Control    BCM.15057
142          b6             1                L   Control    BCM.15057
143          b6             1                L   Control    BCM.15057
144          b6             1                L   Control    BCM.15057
145          b6             1                L      Ribo    BCM.15057
146          b6             1                L      Ribo    BCM.15057
                        Run Fraction Intensity
141 001_Phospho1_P04_DIA_D22        1         0
142 002_Phospho1_P04_DIA_D22        1         0
143 003_Phospho1_P04_DIA_D22        1         0
144 004_Phospho1_P04_DIA_D22        1         0
145 005_Phospho1_P04_DIA_D22        1         0
146 006_Phospho1_P04_DIA_D22        1         0
```

Filtered data based on Qvalue,
PTM assignment (multiplicities reported)
Other modifications than STY removed

MSstatsPTM

2) Data summarization `dataSummarizationPTM`

Goal: Get summarised intestines for a given peptide with PTM in a sample

A. Log2 transformation
B. Normalization (Median)
C. Imputation (via accelerate failure time model - AFT)
D. Summation (Tuckey median polish - TMP)

MSstatsPTM

2) Data summarization `dataSummarizationPTM`

Goal: Get summarised intestines for a given peptide with PTM in a sample

A. Log2 transformation
B. Normalization (Median)
C. Imputation (via accelerate failure time model - AFT)
D. Summation (Tuckey median polish - TMP)

Whole plot

Subplot	$Condition_1$						...				...	$Condition_I$						...			
	$Subject_1$			$Subject_2$			...	$Subject_J$			...	$Subject_{(I-1)J+1}$			$Subject_{(I-1)J+2}$			...	$Subject_{IJ}$		
	Run_1	Run_2	Run_3	Run_4	Run_5	Run_6	...	Run_{JK-2}	Run_{JK-1}	Run_{JK}	...	$Run_{(I-1)JK+1}$	$Run_{(I-1)JK+2}$	$Run_{(I-1)JK+3}$	$Run_{(I-1)JK+4}$	$Run_{(I-1)JK+5}$	$Run_{(I-1)JK+6}$	...	Run_{IJK-2}	Run_{IJK-1}	Run_{IJK}
$Feature_1$	y	y	y	y	y	y	...	y	y	y	...	y	y	y	y	y	y	...	y	y	y
$Feature_2$	y	y	y	y	y	NA	...	y	y	y	...	y	y	y	y	y	y	...	y	y	y
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
$Feature_L$	y	Cens	y	Cens	Cens	y	...	Cens	y	y	...	NA	y	y	y	y	y	...	y	Cens	y

AFT:

$$y_{ijkl} = \mu + Run_{ijk} + Feature_l + \sigma_\epsilon x_{ijkl}, \text{ where } x_{ijkl} \overset{iid}{\sim} \mathcal{N}(0, 1), \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} Run_{ijk} = 0, \sum_{l=1}^{L} Feature_l = 0$$

$$\hat{y}_{ijkl,imp} = \hat{\mu} + \widehat{Run}_{ijk} + \widehat{Feature}_l$$

$$\delta_{ijkl} = \begin{cases} 1 & \text{observed} \\ 0 & \text{censored} \end{cases}$$

$$L(\mu, Run_{ijkl}, Feature_{ijkl}, \sigma_\epsilon | y_{ijkl}) = \prod_{i,j,k,l} f(y_{ijkl})^{\delta_{ijkl}} \times \prod_{i,j,k,l} F(y_{ijkl})^{1-\delta_{ijkl}}$$

where f is the probability density function and F is the cumulative density function of the Normal distribution with expected value $\mu + Run_{ijk} + Feature_l$ and variance σ_ϵ^2. The

MSstatsPTM

2) Data summarization

`dataSummarizationPTM`

Goal: Get summarised intestines for a given peptide with PTM in a sample

A. Log2 transformation
B. Normalization (Median)
C. Imputation (via accelerate failure time model - AFT)
D. Summation (Tuckey median polish - TMP)

Whole plot

Subplot	$Condition_1$						...				...	$Condition_I$										
	$Subject_1$			$Subject_2$			...	$Subject_J$			...	$Subject_{(I-1)J+1}$			$Subject_{(I-1)J+2}$			...	$Subject_{IJ}$		
	Run_1	Run_2	Run_3	Run_4	Run_5	Run_6	...	Run_{JK-2}	Run_{JK-1}	Run_{JK}	...	$Run_{(I-1)JK+1}$	$Run_{(I-1)JK+2}$	$Run_{(I-1)JK+3}$	$Run_{(I-1)JK+4}$	$Run_{(I-1)JK+5}$	$Run_{(I-1)JK+6}$	...	Run_{IJK-2}	Run_{IJK-1}	Run_{IJK}
$Feature_1$	y	y	y	y	y	y	...	y	y	y	...	y	y	y	y	y	y	...	y	y	y
$Feature_2$	y	y	y	y	y	**NA**	...	y	y	y	...	y	y	y	y	y	y	...	y	y	y
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
$Feature_L$	y	**Cens**	y	**Cens**	**Cens**	y	...	**Cens**	y	y	...	**NA**	y	y	y	y	y	...	y	**Cens**	y

TMP:

TMP : Parameter estimation by robust method

$$y_{ijkl} = \mu + Run_{ijk} + Feature_l + \epsilon_{ijkl} \text{ , where}$$

$$median_{ijk}(Run_{ijk}) = 0,\ median_l(Feature_l) = 0,\ and\ median_{ijk}(\epsilon_{ijkl}) = median_l(\epsilon_{ijkl}) = 0$$

	Run_1	Run_2	Run_3	Run_4	Run_5	Run_6	...	Run_{JK-2}	Run_{JK-1}	Run_{JK}	...	$Run_{(I-1)JK+1}$	$Run_{(I-1)JK+2}$	$Run_{(I-1)JK+3}$	$Run_{(I-1)JK+4}$	$Run_{(I-1)JK+5}$	$Run_{(I-1)JK+6}$	...	Run_{IJK-2}	Run_{IJK-1}	Run_{IJK}
Summarized	$\hat{y}$	$\hat{y}$	$\hat{y}$	$\hat{y}$	$\hat{y}$	$\hat{y}$	...	$\hat{y}$	$\hat{y}$	$\hat{y}$	...	$\hat{y}$	$\hat{y}$	$\hat{y}$	$\hat{y}$	$\hat{y}$	$\hat{y}$	...	$\hat{y}$	$\hat{y}$	$\hat{y}$

MSstatsPTM

2) Data summarization

```
MSstatsPTM.summary <- dataSummarizationPTM(raw.input.PTM, verbose = FALSE,
                                           use_log_file = FALSE, append = FALSE,censoredInt = "0")
```

Checking the summed intensities

```
> sum(MSstatsPTM.summary.PTM$ProteinLevelData$NumImputedFeature)
[1] 852550
```

66% of the measured PTMs were to some degree imputed

MSstatsPTM

2) Data summarization

```
MSstatsPTM.summary <- dataSummarizationPTM(raw.input.PTM, verbose = FALSE,
                                           use_log_file = FALSE, append = FALSE,censoredInt = "0")
```

Checking the summed intensities

```
> sum(MSstatsPTM.summary.PTM$ProteinLevelData$NumImputedFeature)
[1] 852550
```

66% of the measured PTMs were to some degree imputed

```
> summary(MSstatsPTM.summary.PTM$ProteinLevelData$LogIntensities)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-5866499       11       13        9       15       33
```

MSstatsPTM

2) Data summarization

```
MSstatsPTM.summary <- dataSummarizationPTM(raw.input.PTM, verbose = FALSE,
                                           use_log_file = FALSE, append = FALSE,censoredInt = "0")
```

Checking the summed intensities

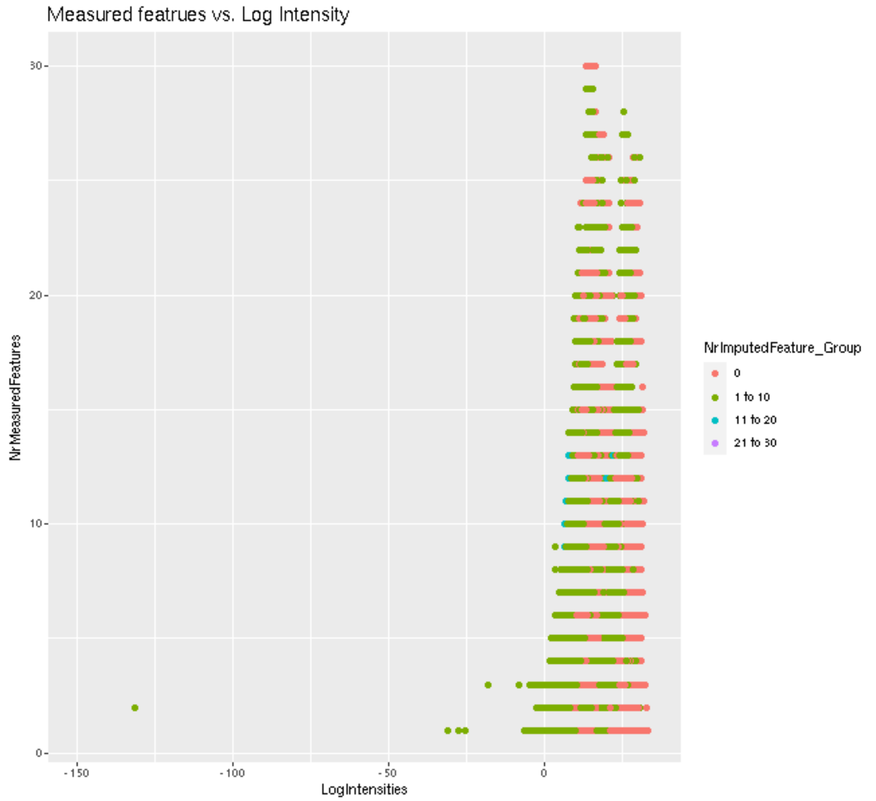
```
> sum(MSstatsPTM.summary.PTM$ProteinLevelData$NumImputedFeature)
[1] 852550
```

66% of the measured PTMs were to some degree imputed

```
> summary(MSstatsPTM.summary.PTM$ProteinLevelData$LogIntensities)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-5866499       11       13        9       15       33
```

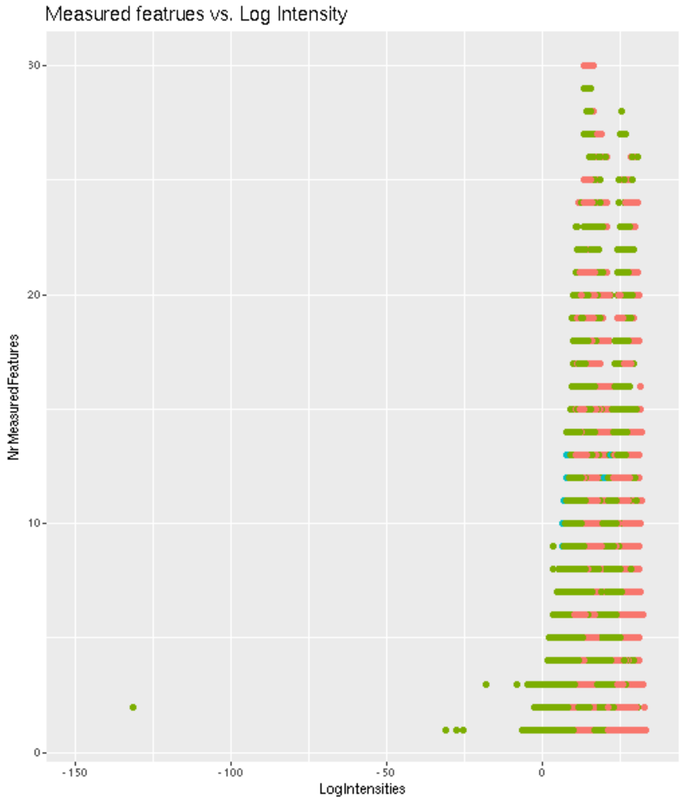
Each point: one measured p-site in a sample

How much “evidence” I have for the detected intensity



MSstatsPTM

2) Data summarization

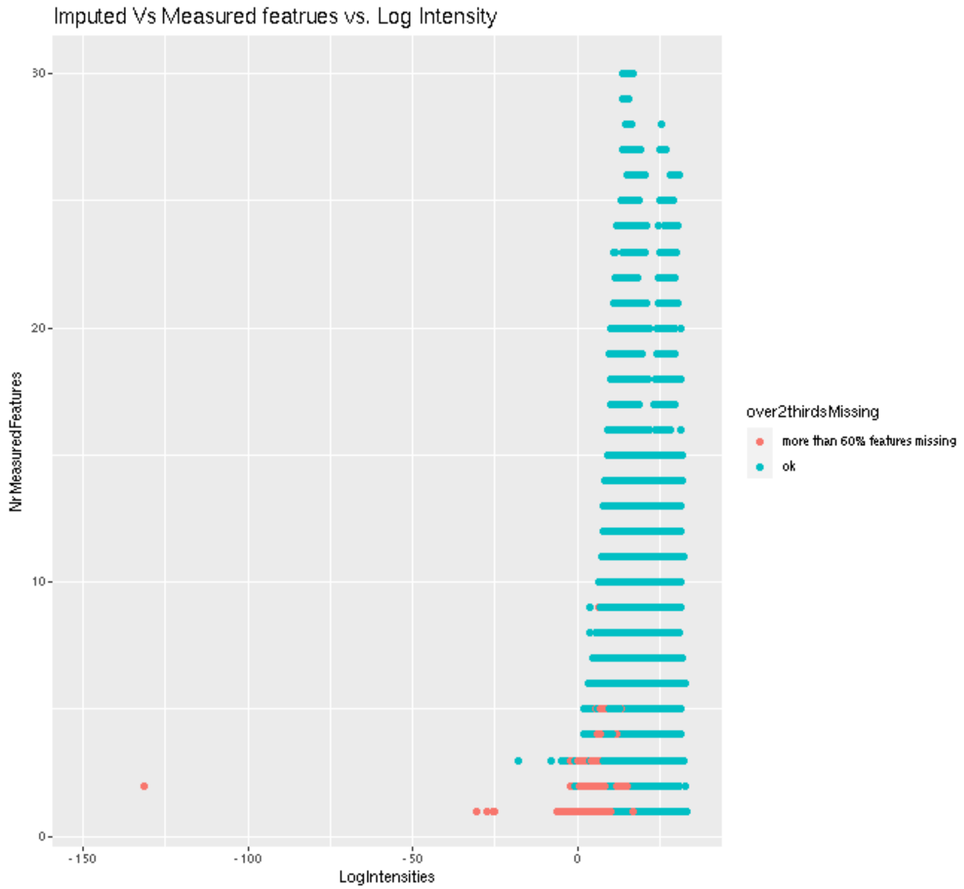



```
> unique(MSstatsPTM.summary.PTM$FeatureLevelData[MSstatsPTM.summary.PTM$FeatureLevelData$PROTEIN == "Q09666_S5841","FEATURE"]) #30
 [1] GHYEVTGS[Phospho (STY)]DDETGK_2_b3_1
 [2] GHYEVTGS[Phospho (STY)]DDETGK_2_b4_1
 [3] GHYEVTGS[Phospho (STY)]DDETGK_2_y3_1
 [4] GHYEVTGS[Phospho (STY)]DDETGK_3_b3_1
 [5] GHYEVTGS[Phospho (STY)]DDETGK_3_b4_1
 [6] GHYEVTGS[Phospho (STY)]DDETGKLQGSGVSLASK_3_b3_1
 [7] GHYEVTGS[Phospho (STY)]DDETGKLQGSGVSLASK_3_b4_1
 [8] GHYEVTGS[Phospho (STY)]DDETGKLQGSGVSLASK_3_y3_1
 [9] GHYEVTGS[Phospho (STY)]DDETGKLQGSGVSLASK_3_y5_1
[10] GHYEVTGS[Phospho (STY)]DDETGKLQGSGVSLASK_3_y9_1
[11] GHYEVTGS[Phospho (STY)]DDETGKLQGSGVSLASK_4_b3_1
[12] GHYEVTGS[Phospho (STY)]DDETGKLQGSGVSLASK_4_b4_1
[13] GHYEVTGS[Phospho (STY)]DDETGKLQGSGVSLASK_4_y5_1
[14] GHYEVTGS[Phospho (STY)]DDETGKLQGSGVSLASKK_3_b3_1
[15] GHYEVTGS[Phospho (STY)]DDETGKLQGSGVSLASKK_3_b4_1
[16] GHYEVTGS[Phospho (STY)]DDETGKLQGSGVSLASKK_3_y10_1
[17] GHYEVTGS[Phospho (STY)]DDETGKLQGSGVSLASKK_4_b4_1
[18] GHYEVTGS[Phospho (STY)]DDETGKLQGSGVSLASKK_4_y3_1
[19] GHYEVTGS[Phospho (STY)]DDETGKLQGSGVSLASKK_4_y4_1
[20] SKGHYEVTGS[Phospho (STY)]DDETGK_2_y3_1
[21] SKGHYEVTGS[Phospho (STY)]DDETGK_3_b4_1
[22] SKGHYEVTGS[Phospho (STY)]DDETGK_3_y3_1
[23] SKGHYEVTGS[Phospho (STY)]DDETGKLQGSGVSLASK_3_b7_1
[24] SKGHYEVTGS[Phospho (STY)]DDETGKLQGSGVSLASK_3_y3_1
[25] SKGHYEVTGS[Phospho (STY)]DDETGKLQGSGVSLASK_3_y5_1
[26] SKGHYEVTGS[Phospho (STY)]DDETGKLQGSGVSLASK_4_b4_1
[27] SKGHYEVTGS[Phospho (STY)]DDETGKLQGSGVSLASK_4_b6_1
[28] SKGHYEVTGS[Phospho (STY)]DDETGKLQGSGVSLASK_4_y3_1
[29] SKGHYEVTGS[Phospho (STY)]DDETGKLQGSGVSLASK_4_y5_1
[30] SKGHYEVTGS[Phospho (STY)]DDETGKLQGSGVSLASK_4_y9_1
```

MSstatsPTM

2) Data summarization

Remove p-sites with more that 60% missing features

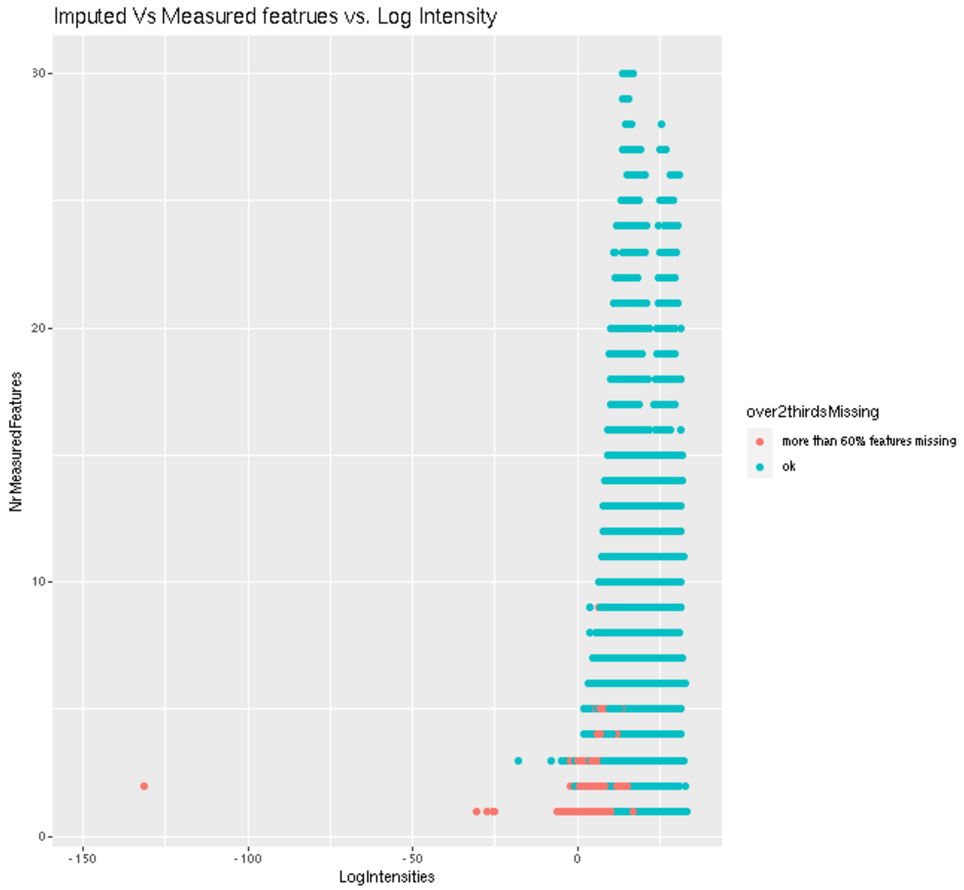



```
> summary(MSstatsPTM.summary.PTM$ProteinLevelData$LogIntensities)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-18.09   11.50   12.93   13.89   14.73   33.25
```

MSstatsPTM

2) Data summarization

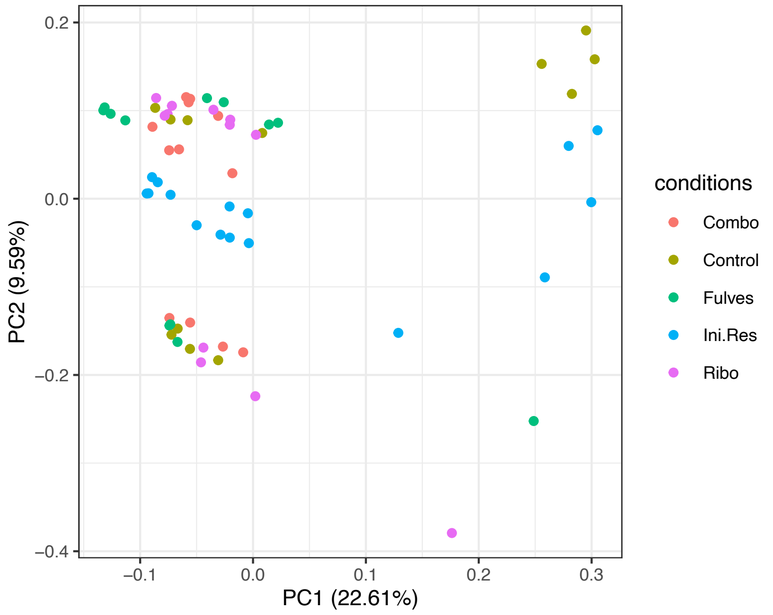
Remove p-sites with more that 60% missing features




```
> summary(MSstatsPTM.summary.PTM$ProteinLevelData$LogIntensities)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-18.09   11.50   12.93   13.89   14.73   33.25
```

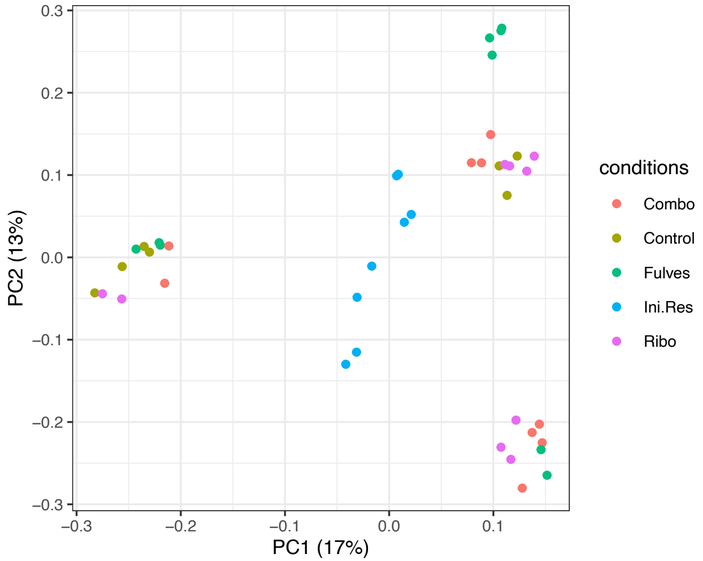
From long format to wide format -> **Intensity matrix**
still many NAs (38%, not equally among the samples)

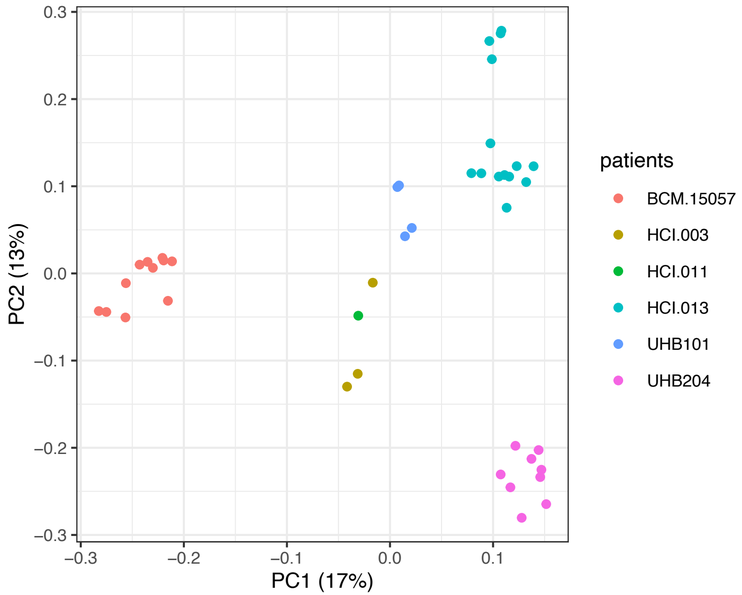
dim(**Intensity matrix**) = [29974 64]
Naively NA -> 0



MSstatsPTM

2) Data summarization



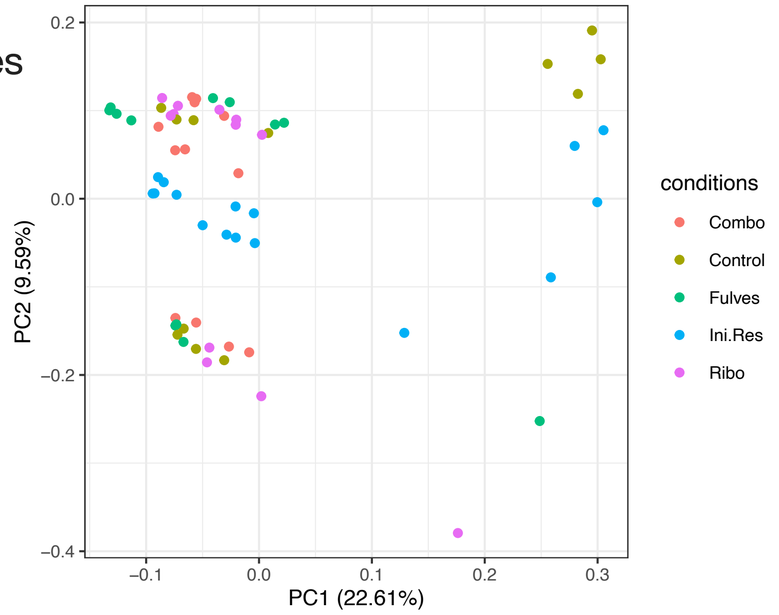


From long format to wide format -> **Intensity matrix**
still many NAs (37%, not equally among the samples)

dim(**Intensity matrix**) = [29974 65]
Naively NA -> 0

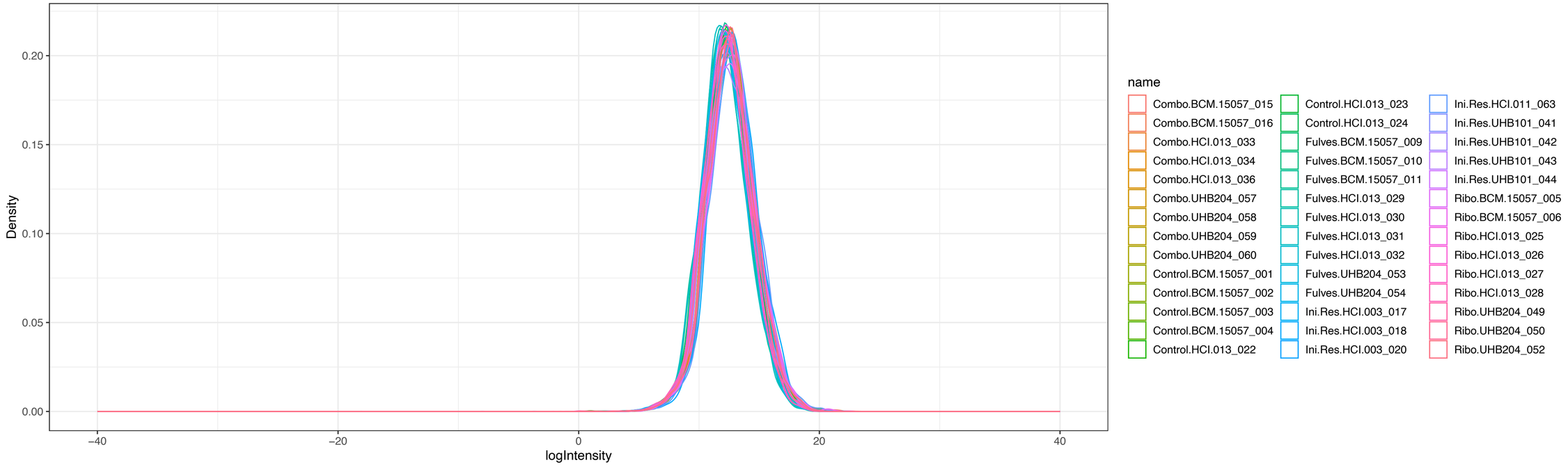
Keep samples that have
at least 60% of the detected p-sites

42 samples left
(30% NAs)



MSstatsPTM

2) Data summarization



MSstatsPTM

3) Statistical modeling

```
contr.matrix <- makeContrasts(Combo.Vs.Ctr = Combo-Control,
                              Fulves.Vs.Ctr = Fulves-Control,
                              Ribo.Vs.Ctr = Ribo-Control,
                              ResVsSens =Ini.Res - Control,
                              levels = levels(as.factor(MSstatsPTM.summary$PTM$ProteinLevelData$GROUP)) )  %>% t
```

```
MSstatsPTM.model = groupComparisonPTM( MSstatsPTM.summary,
                                       data.type = "LabelFree",
                                       contrast.matrix = contr.matrix,
                                       use_log_file = FALSE, append = FALSE,
                                       verbose = TRUE)
```

- Goes one after another p-site separately and fits either linear or linear mixed-effect model, lm() or lmer() functions

MSstatsPTM

3) Statistical modeling

```
contr.matrix <- makeContrasts(Combo.Vs.Ctr = Combo-Control,
                              Fulves.Vs.Ctr = Fulves-Control,
                              Ribo.Vs.Ctr = Ribo-Control,
                              ResVsSens =Ini.Res - Control,
                              levels = levels(as.factor(MSstatsPTM.summary$PTM$ProteinLevelData$GROUP)) )  %>% t
```

```
MSstatsPTM.model = groupComparisonPTM( MSstatsPTM.summary,
                                       data.type = "LabelFree",
                                       contrast.matrix = contr.matrix,
                                       use_log_file = FALSE, append = FALSE,
                                       verbose = TRUE)
```

- Goes one after another p-site separately and fits either linear or linear mixed-effect model, lm() or lmer() functions

$$ABUNDANCE \sim \underbrace{GROUP}_{\text{treatment}} + \underbrace{(1 \mid SUBJECT)}_{\text{patient}} + \underbrace{(1 \mid GROUP{:}SUBJECT)}_{\text{patient-mouse interaction}}$$

effect **bio-replicas** **tech-replicas**

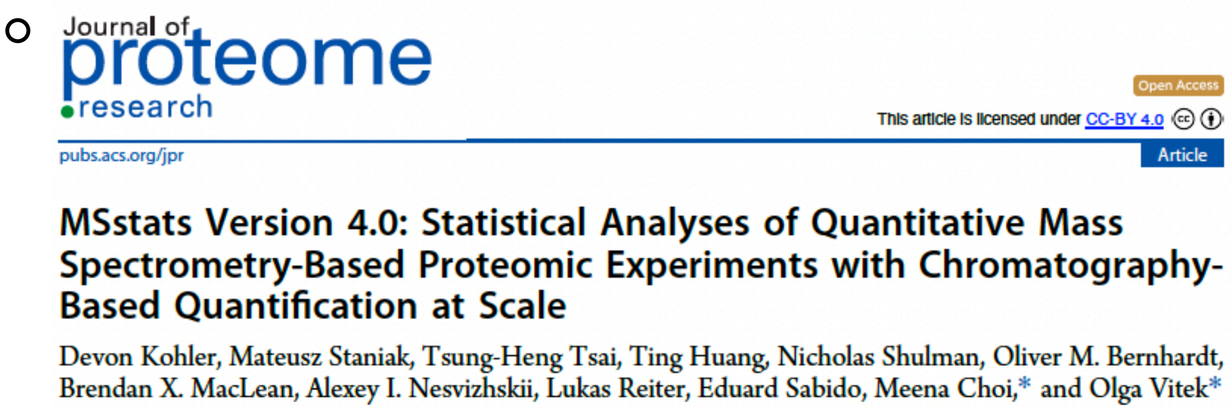
- Each p-site can be fitted using different model depending on how many of NAs there is in the Intensity matrix
- Comparisons to completely missing groups are annotated as “Inf” or “-Inf”

```
> annotPTM
# A tibble: 70 × 3
   Run                                                Condition BioReplicate
   <chr>                                              <chr>     <chr>
 1 001_Phospho1_P04_DIA_D22                           Control   BCM.15057
 2 002_Phospho1_P04_DIA_D22                           Control   BCM.15057
 3 003_Phospho1_P04_DIA_D22                           Control   BCM.15057
 4 004_Phospho1_P04_DIA_D22                           Control   BCM.15057
 5 005_Phospho1_P04_DIA_D22                           Ribo      BCM.15057
 6 006_Phospho1_P04_DIA_D22                           Ribo      BCM.15057
 7 007_Phospho1_P04_DIA_20220322194039_D22            Ribo      BCM.15057
 8 008_Phospho1_P04_DIA_20220322215824_D22            Ribo      BCM.15057
 9 009_Phospho1_P04_DIA_D22                           Fulves    BCM.15057
10 010_Phospho1_P04_DIA_D22                           Fulves    BCM.15057
```

My thoughts

- The package offers good functionalities for PTM annotation, imputation and summation, but must be checked!
- MSstats ecosystem is made to perform the analysis from the beginning to the end by running a few function
  -> restrictive and potentially dangerous (black-box results)

- Advice on the experimental design? Replication!! + Simple designs



Acknowledgement

- Bentires Lab
- Bioinfo Core Team DBM