Proteomics Data Analysis

Simon Andrews

V2024-09

Babraham Bioinformatics

Course Content

- Principles of Mass Spectrometry
- Types of Quantitative MS
- Processing MS Data
  - Running searches
  - Evaluating Quality Control
- Analysing MS Data
  - MSstats Shiny
  - MSstats in R
    - Data import
    - Quantitation and normalisation
    - Differential abundance

Related Courses

- Introduction to R
- Advanced R
- GGplot
- Statistics with R

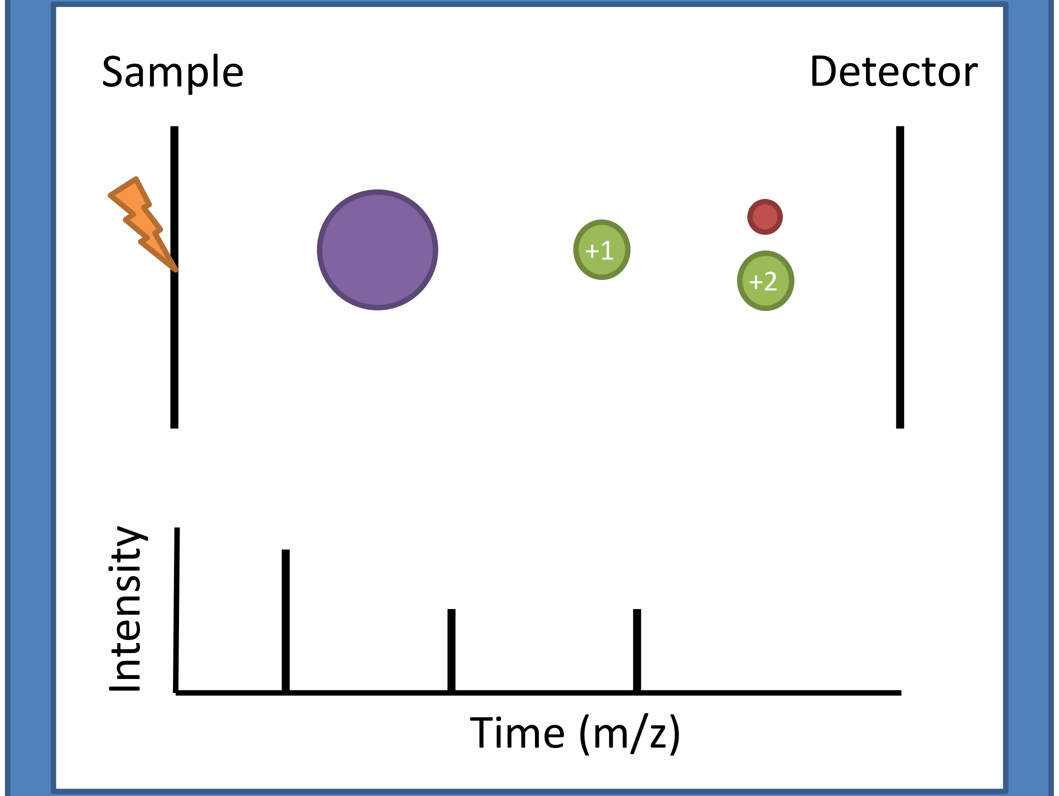
- Interpreting Gene Lists

Principles of Proteomics
Mass Spec

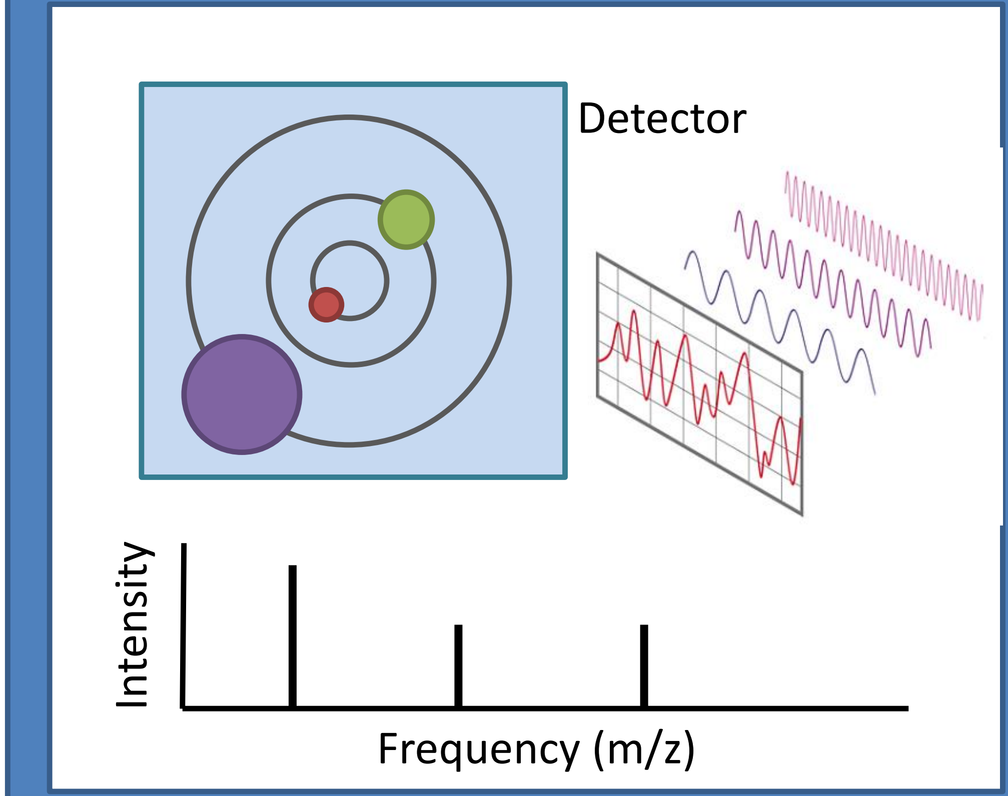
Babraham
Bioinformatics

How mass spectrometers work

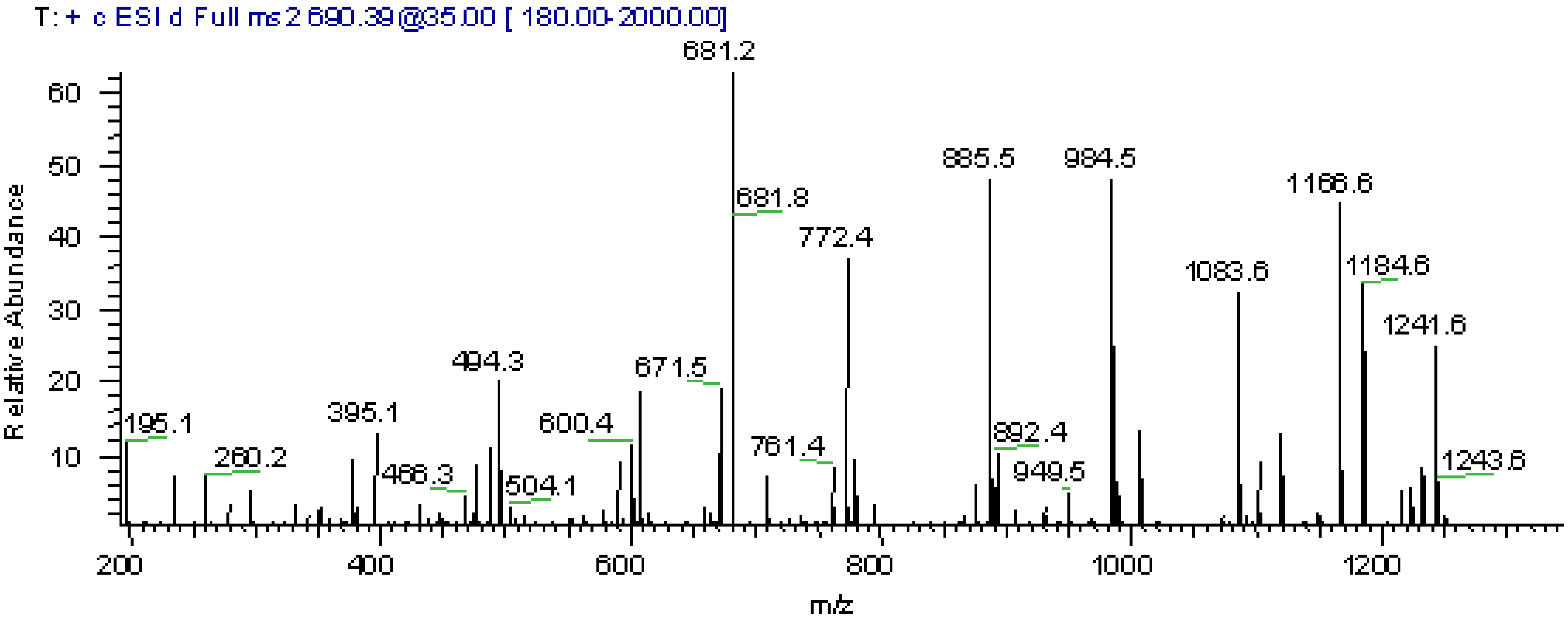
Time of Flight (TOF)



Fourier Transform Ion Cyclotron Resonance (FT-ICR)

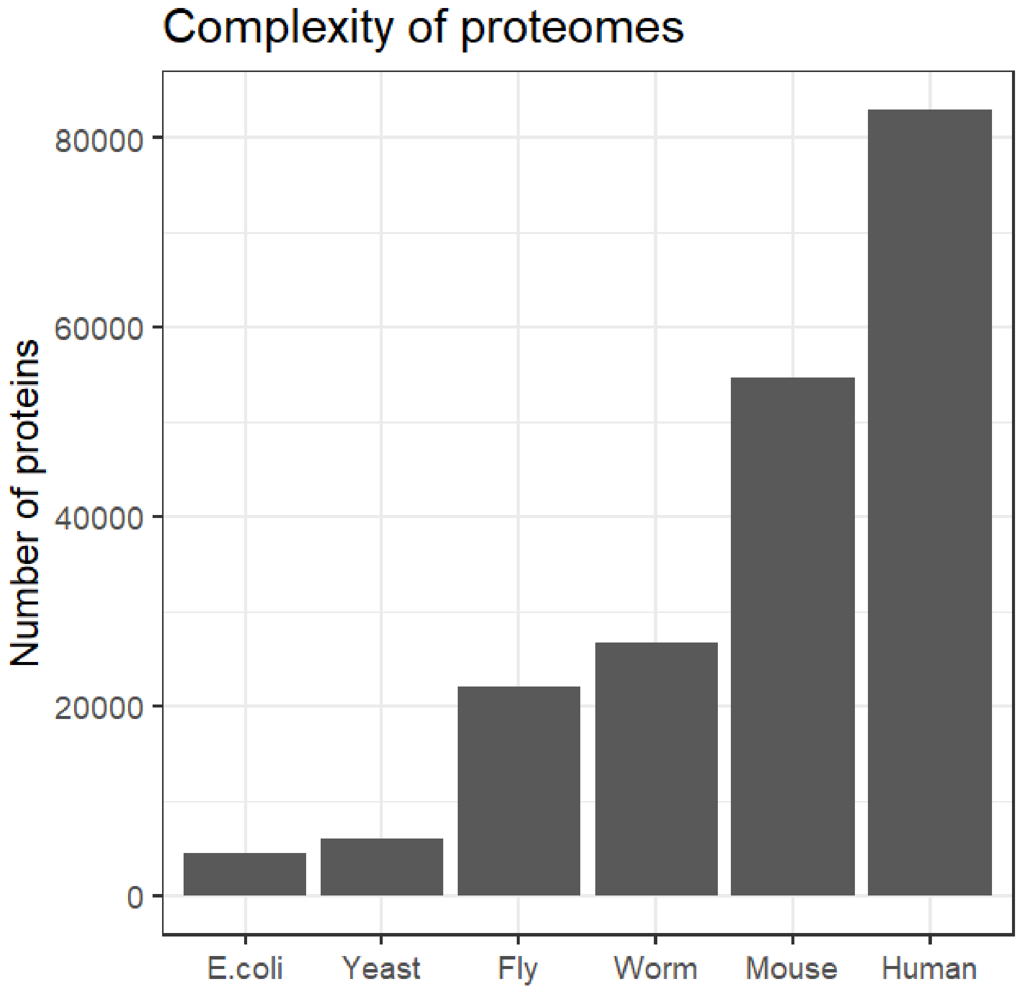


A typical mass spectrum



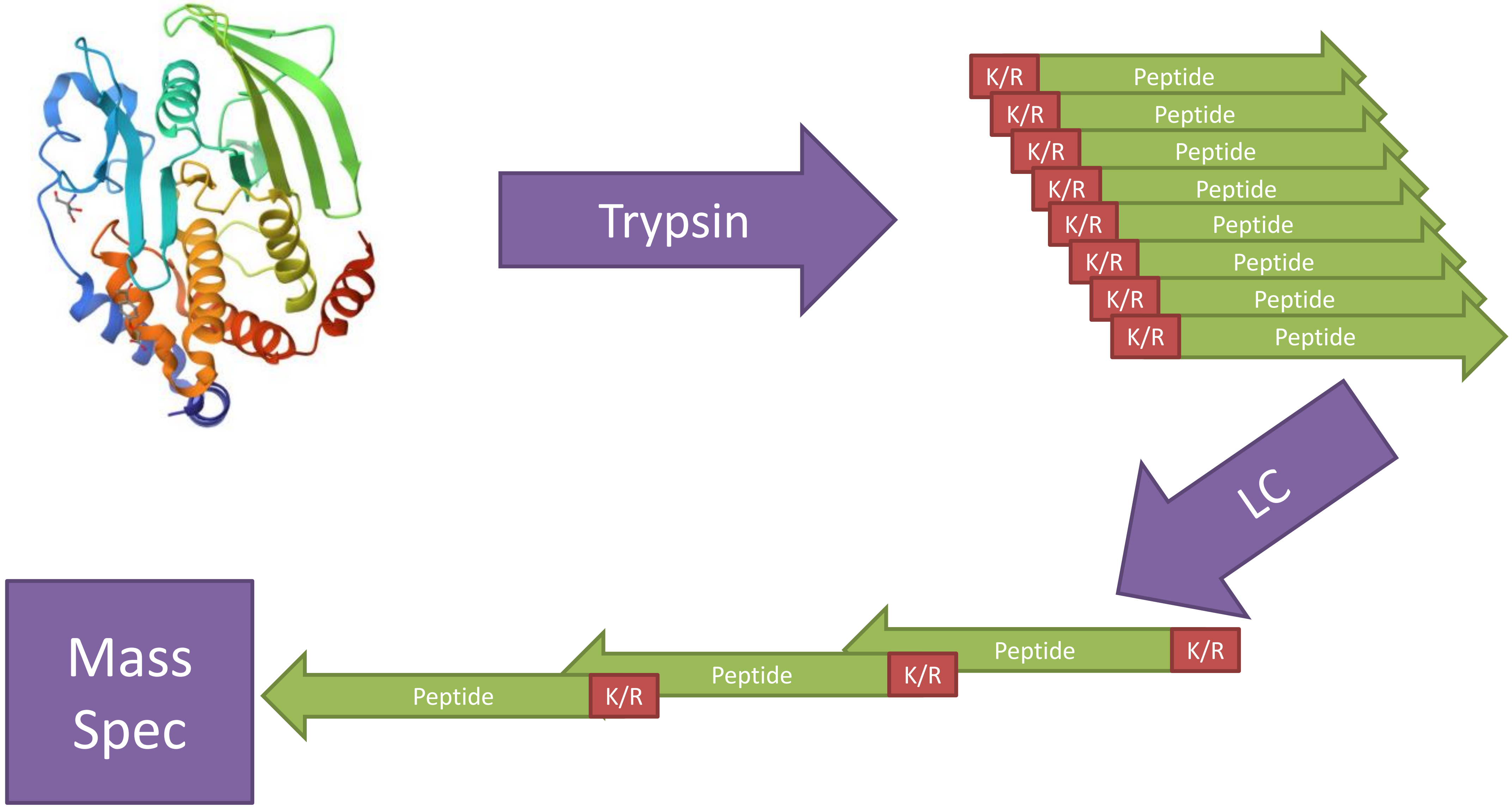
http://www.ionsource.com/tutorial/protID/spectralmatching_mascot.htm

Measuring whole proteomes



- Whole proteins are too big and can't be identified from their mass
- Proteome samples are too complex to put all proteins into the machine at the same time
- Need to find a way to measure data for a complex proteome

"Bottom-up" proteomics



Trypsin

K/R Peptide
K/R Peptide
K/R Peptide
K/R Peptide
K/R Peptide
K/R Peptide
K/R Peptide
K/R Peptide

LC

Peptide K/R
Peptide K/R
Peptide K/R

Mass Spec

Mass Spectrometry

`SILAGVK`	686Da
`KVGALIS`	686Da
`VLAGISK`	686Da

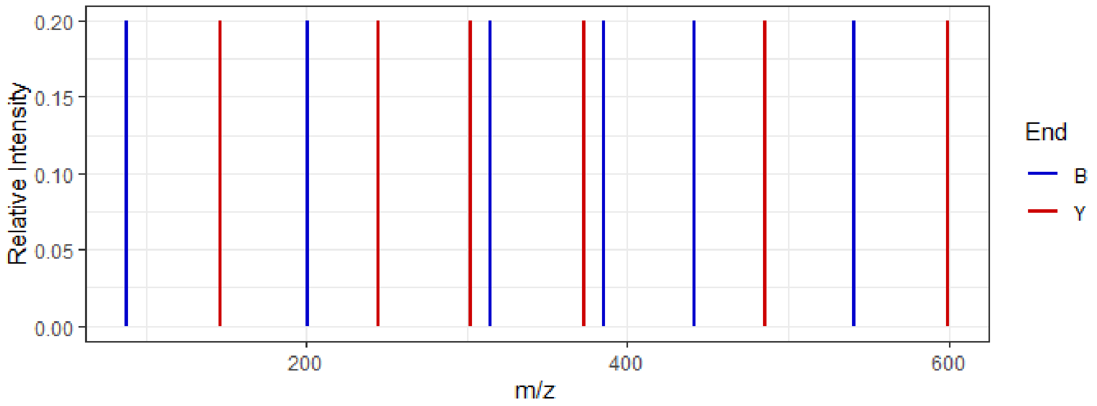
Just knowing a peptide's mass isn't enough to identify it

Tandem Mass Spectrometry

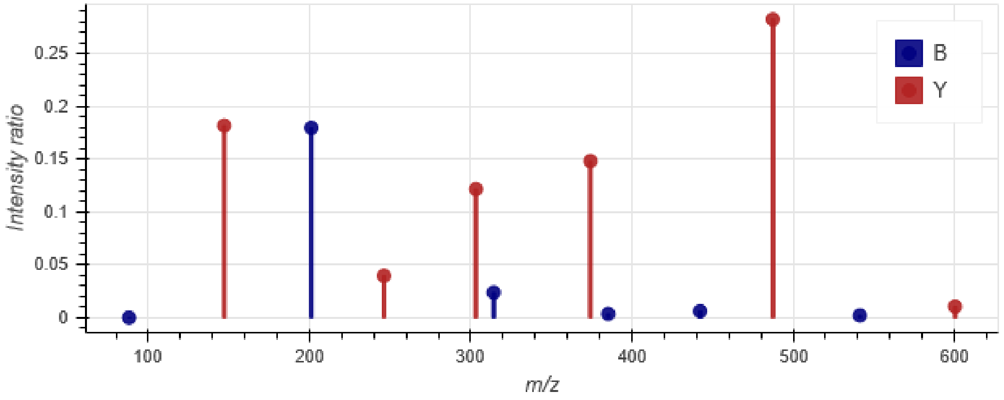
686Da	**SILAGVK**		
541Da	SILAGV	K	147Da
442Da	SILAG	VK	246Da
385Da	SILA	GVK	303Da
314Da	SIL	AGVK	374Da
201Da	SI	LAGVK	487Da
88Da	S	ILAGVK	600Da

Peptide MS2 Spectra

Theoretical



Observed

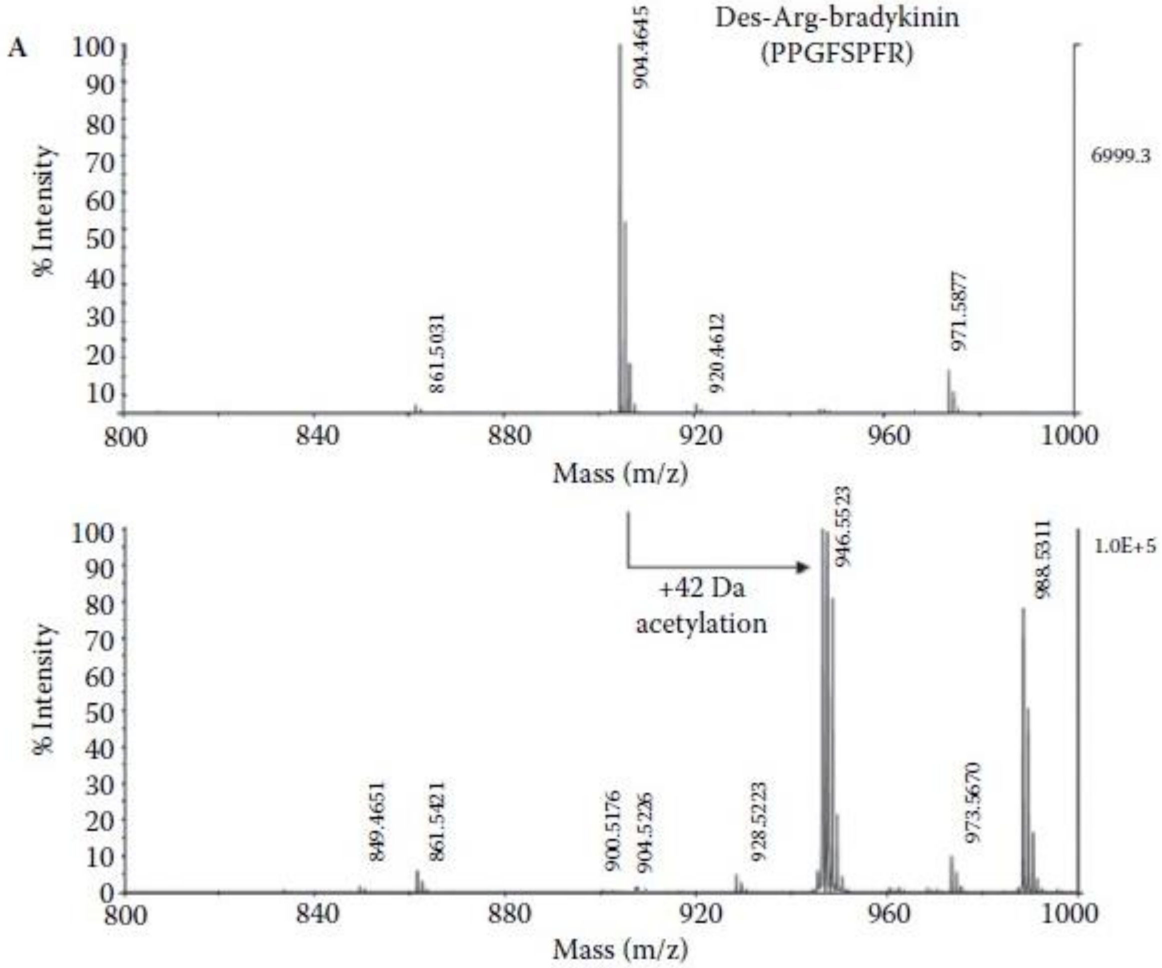


Searches are not performed by inferring sequence from spectra, but by scoring matches to predicted spectra

https://iomics.ugent.be/ms2pip/

Measuring Modifications

- Acetylation
- Formylation
- Met Oxidation
- Phosphorylation
- Ubiquitination
- Glycosylation



Problems with bottom up proteomics

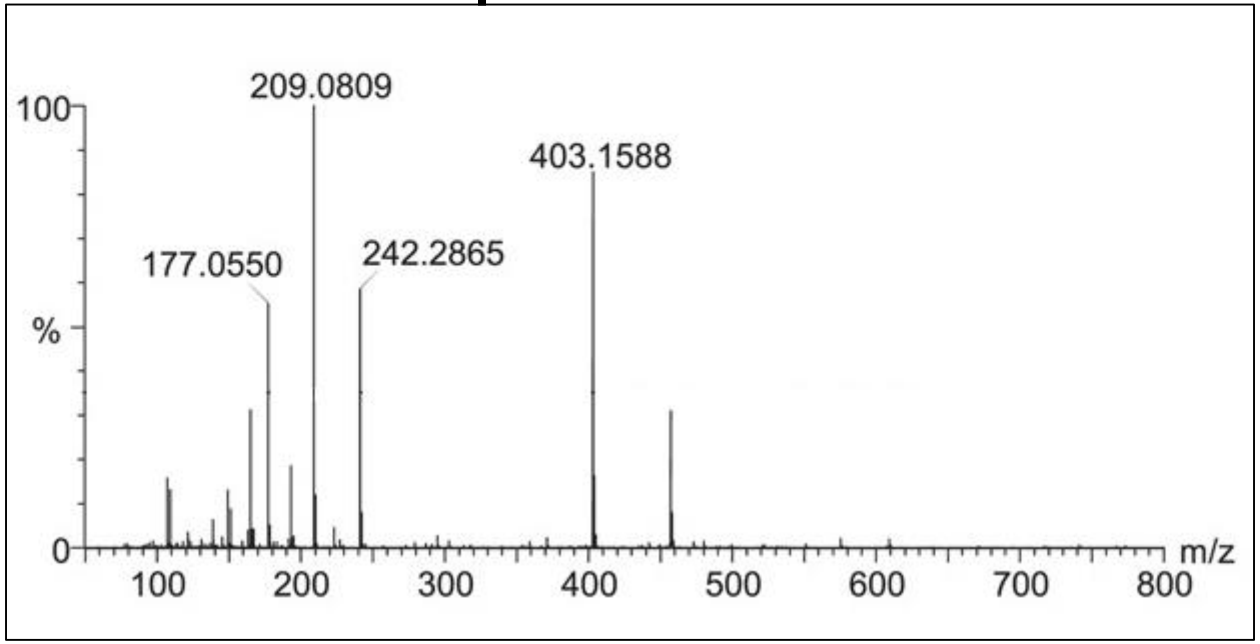
Peptide Mix → LC Column → MS1 Peptide Spectrum → MS2 Fragment Spectra

- Too many peaks for MS2
  - More LC Separation

    (longer run time)
  - Select some peaks

    (ignore others)
  - Mix peaks for MS2

    (messy data)



DDA vs DIA

Data Dependent Acquisition (DDA)

- Pick the strongest peaks from MS1
- Pass them individually to MS2

- Clean MS2 spectra

- Smaller peaks missed – lower coverage
- Different peaks picked in each run
  - Missing values
  - Noise

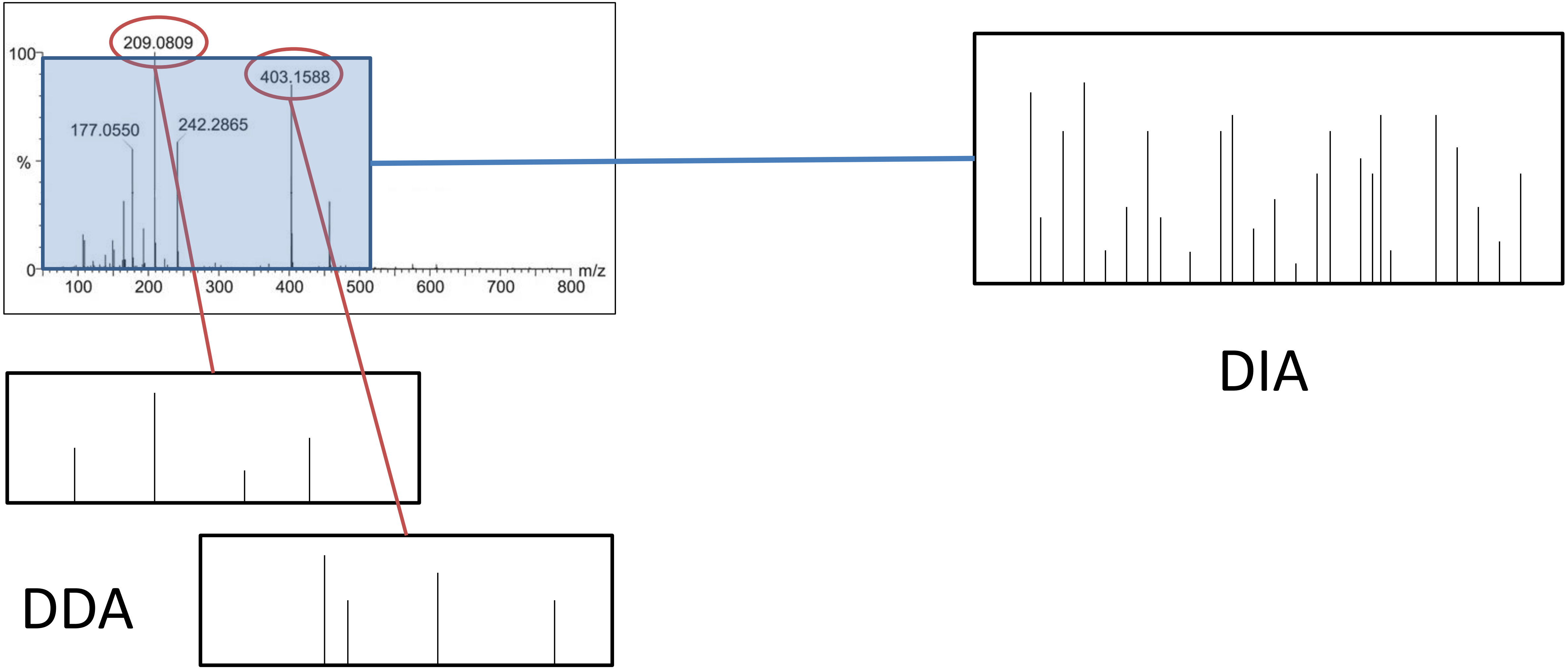
Data Independent Acquisition (DIA)

- Pick all peaks from MS1 (MZ range)
- Pass them simultaneously to MS2

- Mixed MS2 spectra
- More difficult spectrum matching

- Higher coverage
- More complete coverage

DIA vs DDA



DDA

DIA

Identifying Proteins from spectra

Database Searching

Spectronaut®
powered by Pulsar

PROTEOME DISCOVERER

MQ

DIA-NN

- Protein Identification (with confidence)
- Abundance Quantitation
- Downstream analysis

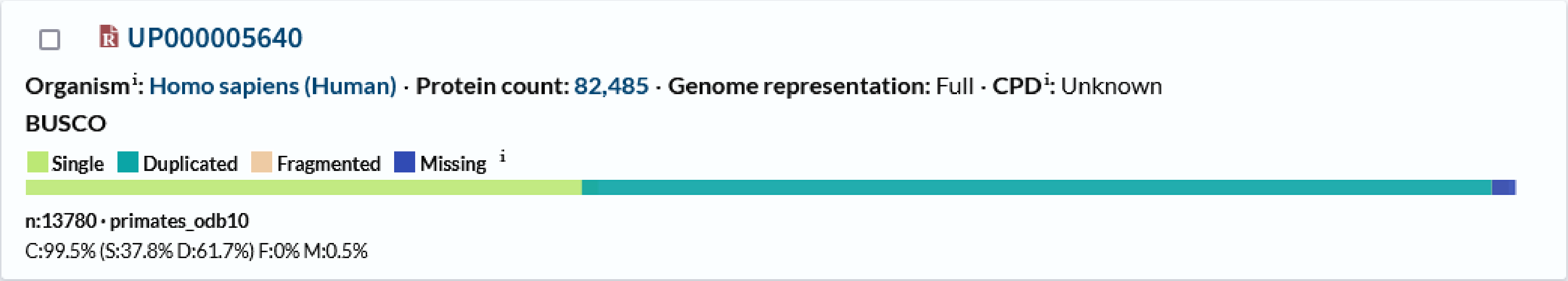
https://www.uniprot.org/proteomes

Proteins
UniProt Knowledgebase
Reviewed (Swiss-Prot) 570,830
Unreviewed (TrEMBL) 249,751,891

Species
Proteomes
Protein sets for species with sequenced genomes from across the tree of life

Protein Clusters
UniRef
Clusters of protein sequences at 100%, 90% & 50% identity

Sequence Archive
UniParc
Non-redundant archive of publicly available protein sequences seen across different databases



☐ UP000005640

Organismⁱ: Homo sapiens (Human) · **Protein count:** 82,485 · **Genome representation:** Full · **CPD**ⁱ: Unknown

BUSCO

n:13780 · primates_odb10
C:99.5% (S:37.8% D:61.7%) F:0% M:0.5%

gpm cRAP protein sequences

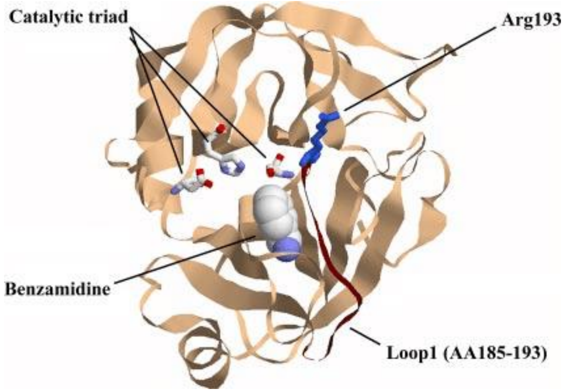
The **c**ommon **R**epository of **A**dventitious **P**roteins

Keratin

(human, sheep)

Cow Proteins

(Cell Culture Medium, BSA)



Trypsin

(or Lys-C)

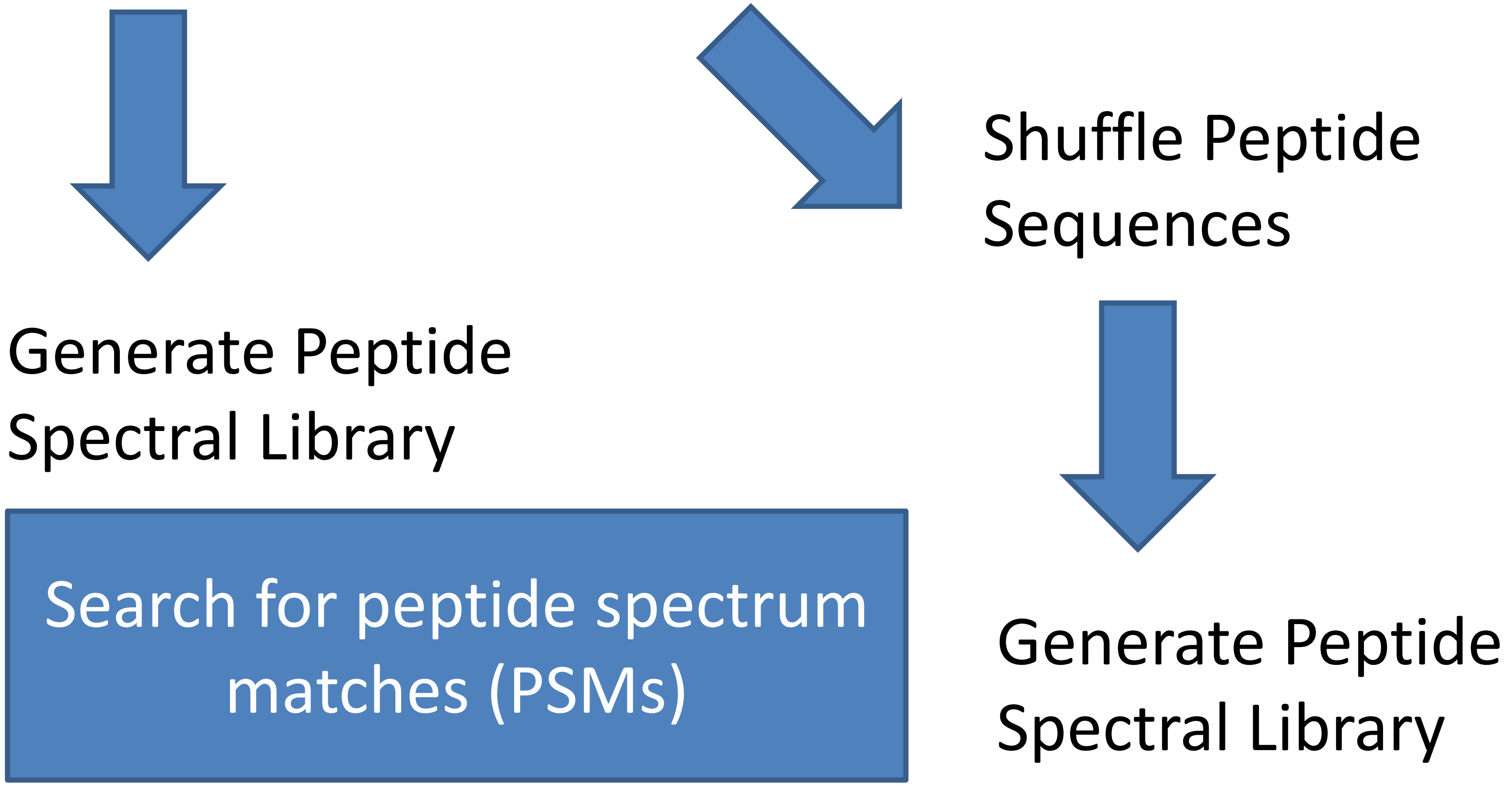
Amylase (Saliva)   Rubber Proteins (gloves)   Weight Markers   Proteomics Standards   Pepsin   Caesein   FLAG/HA   Streptavidin

https://www.thegpm.org/crap/

Database Searching

Spectronaut®
powered by Pulsar

DIA-NN



Take all proteins from your species of interest

UniProt

Shuffle Peptide Sequences

Generate Peptide Spectral Library

Search for peptide spectrum matches (PSMs)

Generate Peptide Spectral Library

Protein Libraries

REAL

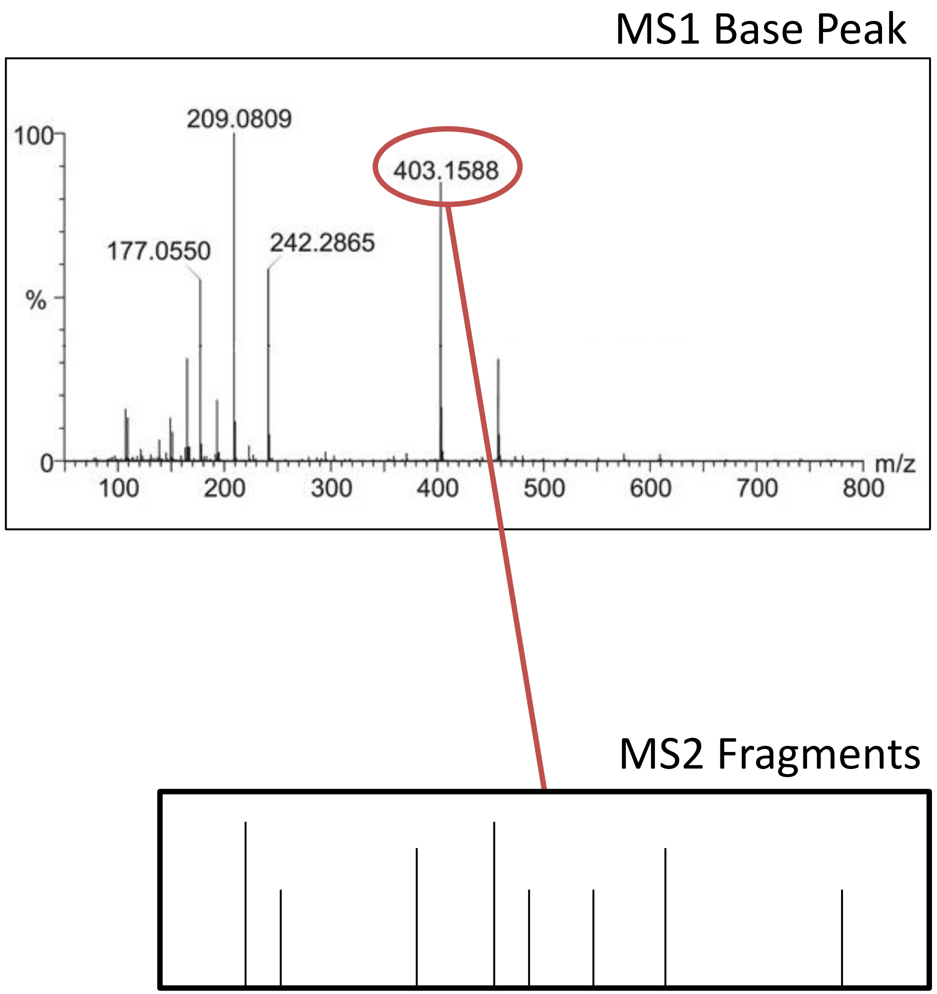
```
>P05067 Amyloid-beta precursor protein
MLPGLALLLLAAWTARALEVPTDGNAGLLAEPQIAMFCGRLNMHMNVQNGKWDSDPSG
TKTCIDTKEGILQYCQEVYPELQITNVVEANQPVTIQNWCKRGRKQCKTHPHFVIPYR
CLVGEFVSDALLVPDKCKFLHQERMDVCETHLHWHTVAKETCSEKSTNLHDYGMLLPC
GIDKFRGVEFVCCPLAEESDNVDSADAEEDDSDVWWGGADTDYADGSEDKVVEVAEEE
EVAEVEEEEADDDEDDEDGDEVEEEAEEPYEEATERTTSIATTTTTTTESVEEVVREV
CSEQAETGPCRAMISRWYFDVTEGKCAPFFYGGCGGNRNNFDTEEYCMAVCGSAMSQS
LLKTTQEPLARDPVKLPTTAASTPDAVDKYLETPGDENEHAHFQKAKERLEAKHRERM
SQVMREWEEAERQAKNLPKADKKAVIQHFQEKVESLEQEAANERQQLVETHMARVEAM
LNDRRRLALENYITALQAVPPRPRHVFNMLKKYVRAEQKDRQHTLKHFEHVRMVDPKK
AAQIRSQVMTHLRVIYERMNQSLSLLYNVPAVAEEIQDEVDELLQKEQNYSDDVLANM
ISEPRISYGNDALMPSLTETKTTVELLPVNGEFSLDDLQPWHSFGADSVPANTENEVE
PVDARPAADRGLTTRPGSGLTNIKTEEISEVKMDAEFRHDSGYEVHHQKLVFFAEDVG
SNKGAIIGLMVGGVVIATVIVITLVMLKKKQYTSIHHGVVEVDAAVTPEERHLSKMQQ
NGYENPTYKFFEQMQN
```

DECOY

```
>P05067_REV
NQMQEFFKYTPNEYGNQQMKSLHREEPTVAADVEVVGHHISTYQKKKLMVLTIVIVTA
IVVGGVMLGIIAGKNSGVDEAFFVLKQHHVEYGSDHRFEADMKVESIEETKINTLGSG
PRTTLGRDAAPRADVPEVENETNAPVSDAGFSHWPQLDDLSFEGNVPLLEVTTKTETL
SPMLADNGYSIRPESIMNALVDDSYNQEKQLLEDVEDQIEEAVAPVNYLLSLSQNMRE
YIVRLHTMVQSRIQAAKKPDVMRVHEFHKLTHQRDKQEARVYKKLMNFVHRPRPPVAQ
LATIYNELALRRRDNLMAEVRAMHTEVLQQRENAAEQELSEVKEQFHQIVAKKDAKPL
NKAQREAEEWERMVQSMRERHKAELREKAKQFHAHENEDGPTELYKDVADPTSAATTP
LKVPDRALPEQTTKLLSQSMASGCVAMCYEETDFNNRNGGCGGYFFPACKGETVDFYW
RSIMARCPGTEAQESCVERVVEEVSETTTTTTTAISTTRETAEEYPEEAEEEVEDGDE
DDEDDDAEEEEVEAVEEEEAVEVVKDESGDAYDTDAGGWWVDSDDEEADASDVNDSEE
ALPCCVFEVGRFKDIGCPLLMGYDHLNTSKESCTEKAVTHWHLHTECVDMREQHLFKC
KDPVLLADSVFEGVLCRYPIVFHPHTKCQKRGRKCWNQITVPQNAEVVNTIQLEPYVE
QCYQLIGEKTDICTKTGSPDSDWKGNQVNMHMNLRGCFMAIQPEALLGANGDTPVELA
RATWAALLLLALGPLM
```

Decoy libraries can be reversed or shuffled

Peptide Spectrum Matches



Find peptides with masses close
to the parent peak

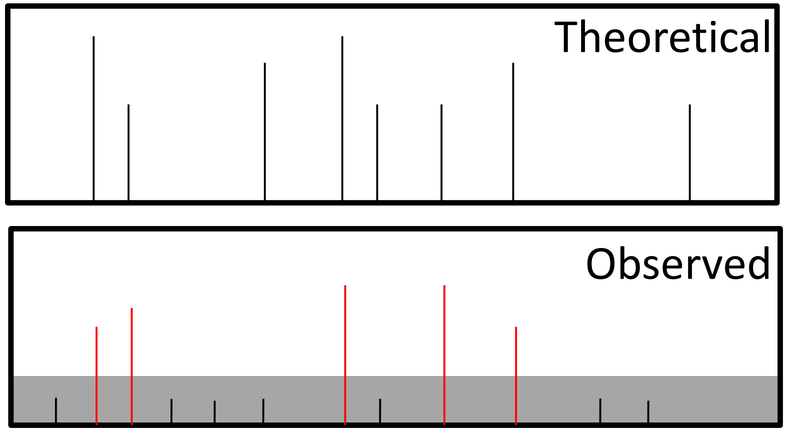
>P05067
MLPGLALLLLAAWTARALEVPTDGNAGLLAEPQIAMFCGRLNMHMNV
QNGKWDSDP**SGTKTCIDTKEGIL**QYCQEVYPELQITNVVEANQPVTI
QNWCKRGRKQCKTHPHFVIPYRCLVGEFVSDALLVPDKCKFLHQERM
DVCETHLHW

>P90210
MAVCGSAMSQSLLKTTQEPLARDPVKLPTTAASTPDAVDKYLETPGD
ENEHAHFQKAKERLEAKHRERMSQVMREWEEAERQAKNLPKADKKAV
IQHFQEKVESLEQEAANERQQLVET**HMARVEAMLNDRR**RLALENYIT
ALQAVPPRPRHVFNMLKKYVRAEQKDRQHTLKHFEH

Hundreds of candidates

Scoring a PSM match

Count Overlaps (Andromeda - MQ)



$$s(q,\text{loss}) = -10\log_{10}\sum_{j=k}^{n}\left[\binom{n}{j}\left(\frac{q}{100}\right)^{j}\left(1-\frac{q}{100}\right)^{n-j}\right]$$

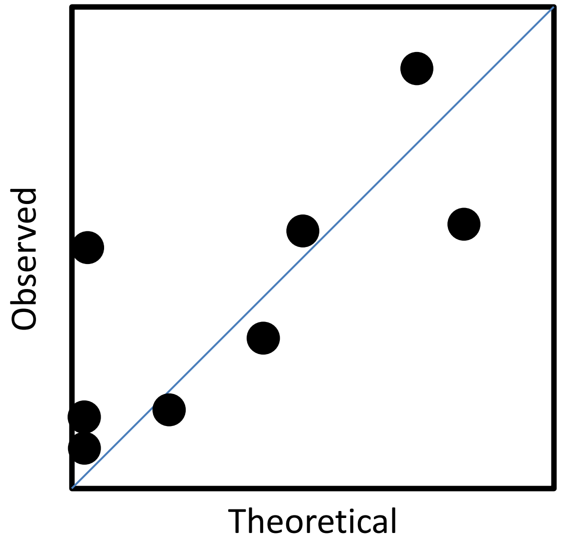
Optimize inclusion of losses

$$s(q) = \max_{\text{loss = true/false}} s(q,\text{loss})$$

Probability of finding *n* matching peaks out of *k* theoretical peaks when taking the top *p* peaks in the spectrum

Correlate intensities (Perseus – PD)

Correlate intensities by mass for true masses and masses shifted +/- 75Da



$$\text{xcorr} = R_0 - \left(\sum_{\tau=-75}^{\tau=+75} R_\tau\right) \Big/ 151$$

$$R_\tau = \sum x[i] \cdot y[i+\tau]$$

Difference between the true correlation and the average mass shifted correlation

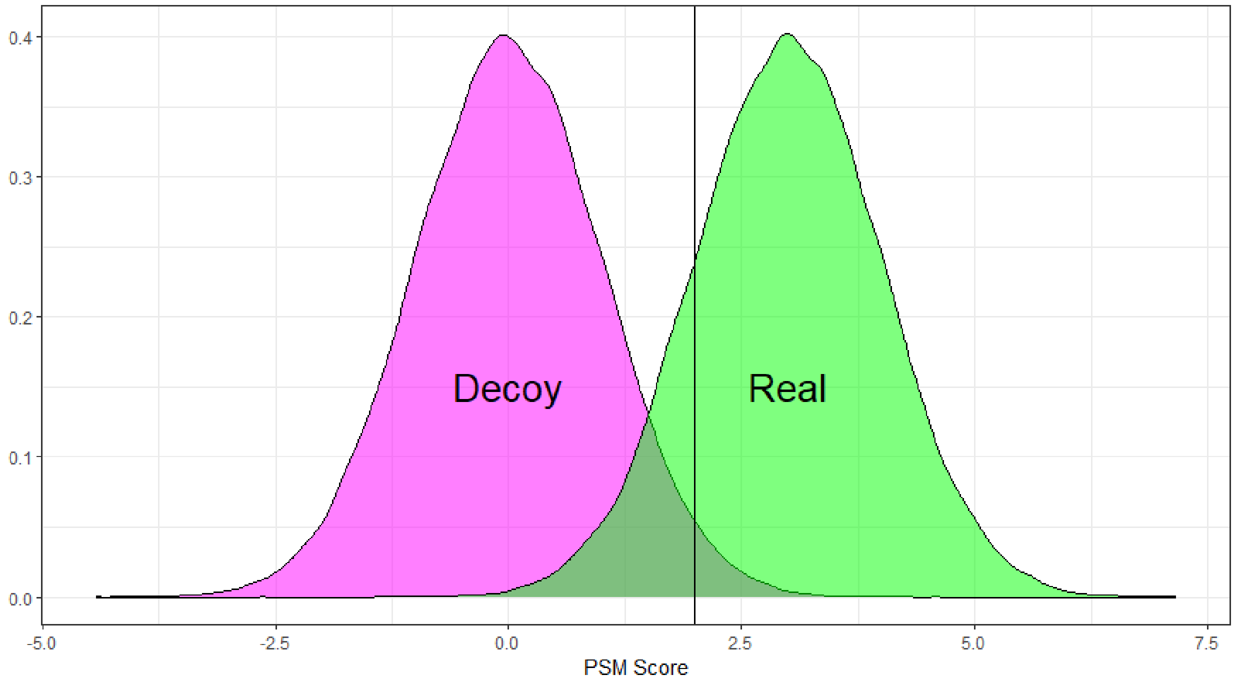
https://pubs.acs.org/doi/10.1021/pr101065j

https://pubs.acs.org/doi/10.1021/pr800420s

Estimating PSM confidence

Search against combined real + decoy database

Use the distribution of decoy hits to calculate a false discovery background

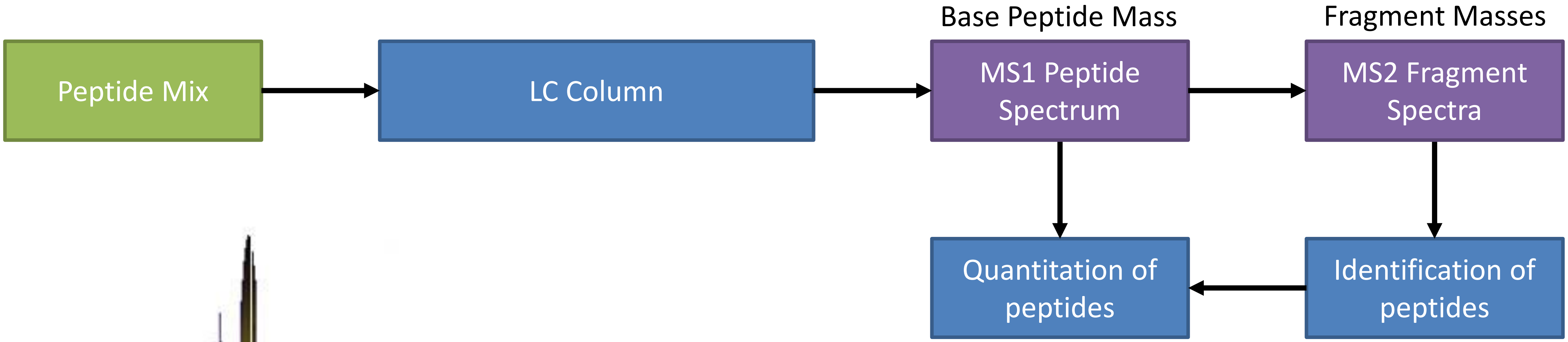


PEP Score
Probability of peptide being wrongly identified

Q-value
Ratio of best Real hit to best Decoy hit

Quantitating Proteins

Label Free Quantitation



Base Peptide Mass

Fragment Masses

Peptide Mix

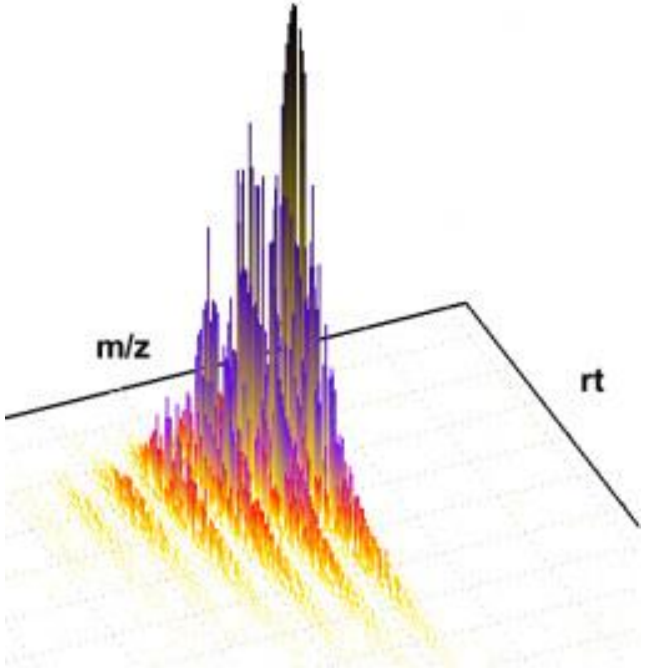
LC Column

MS1 Peptide Spectrum

MS2 Fragment Spectra

Quantitation of peptides

Identification of peptides



- Variation in m/z (isotopes)
- Variation in retention time (adjacent windows)
- Build a "3D" peak – measure peak area

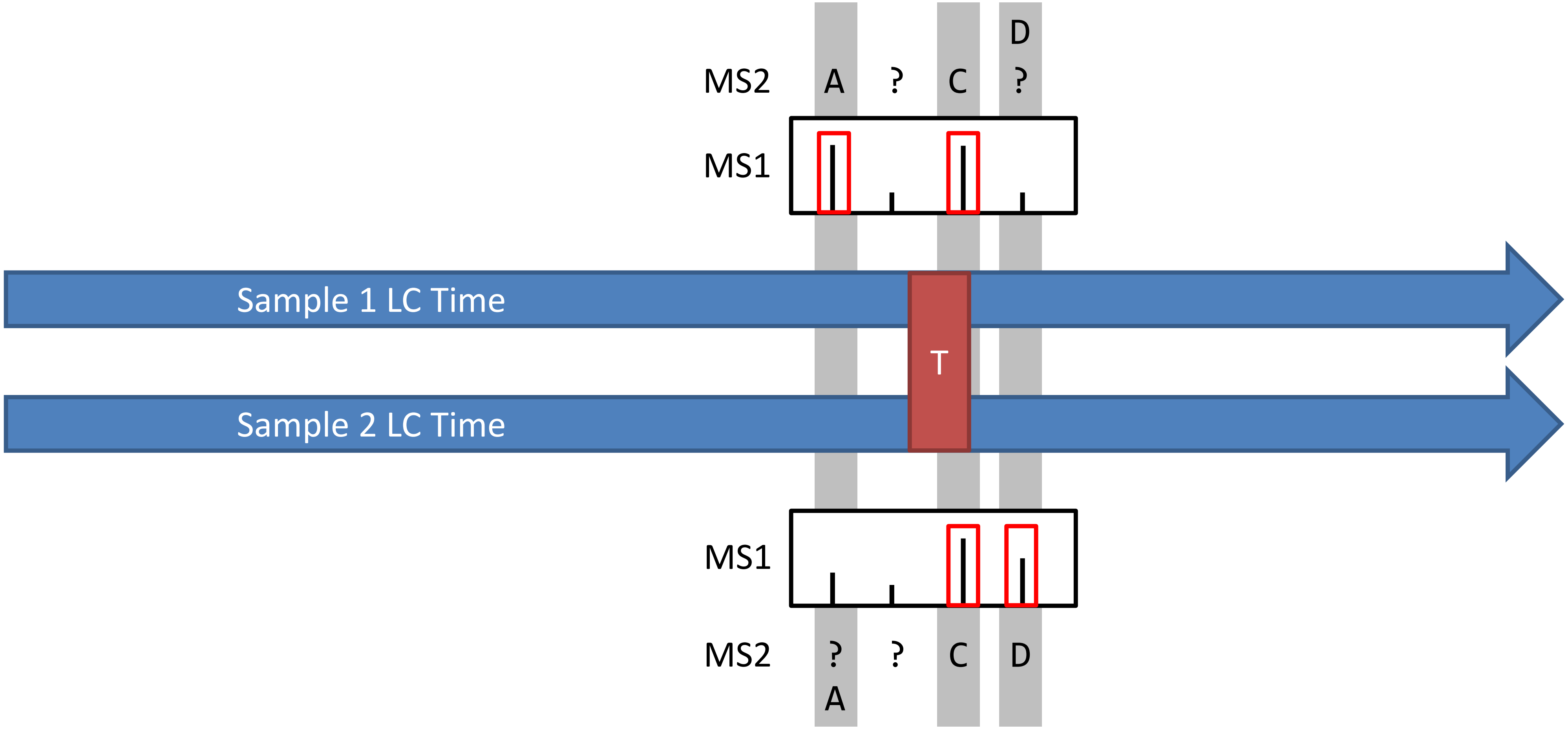
DOI:10.1186/1756-0381-2-4

Measuring multiple samples

- Variability in LC performance / time
- DDA selects different peaks
- Different peptides identified
- Missing values

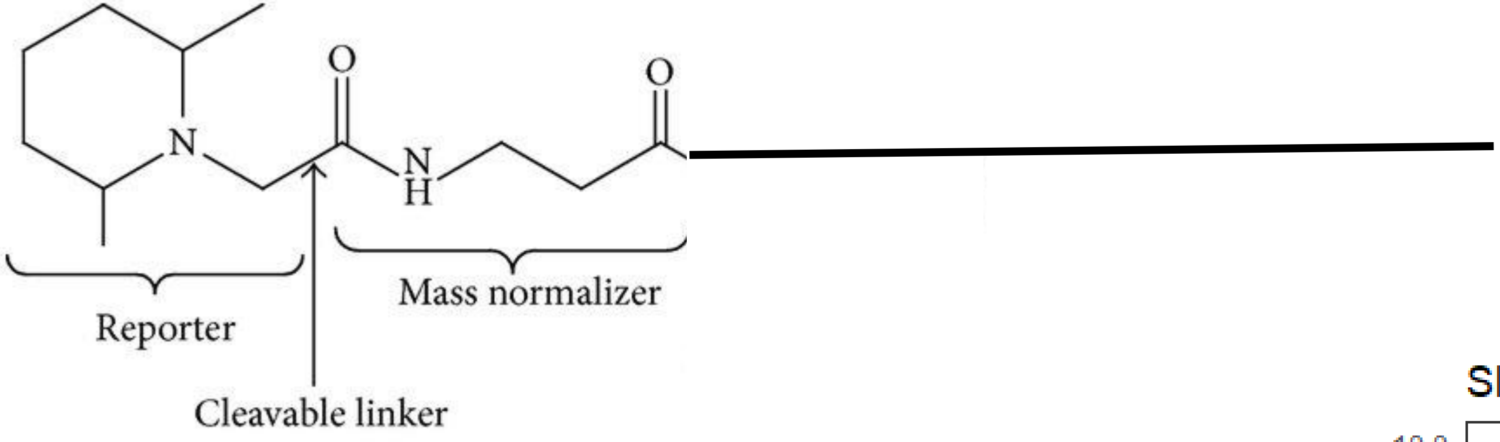
- How to measure consistently across samples?

Finding missing label free MS2 peaks

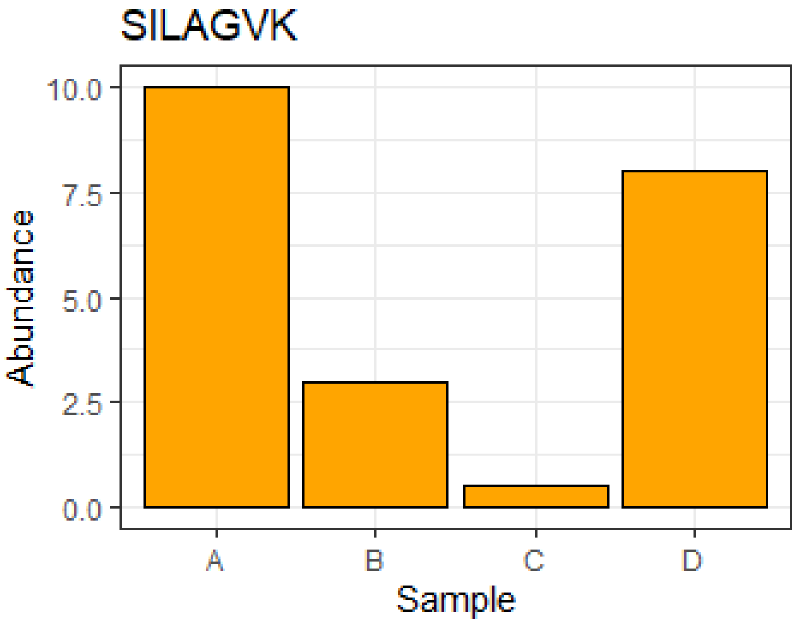


Matching MS1 base peaks based on LC time and M/Z allows more consistent data collection.

Tandem Mass Tagging (TMT)



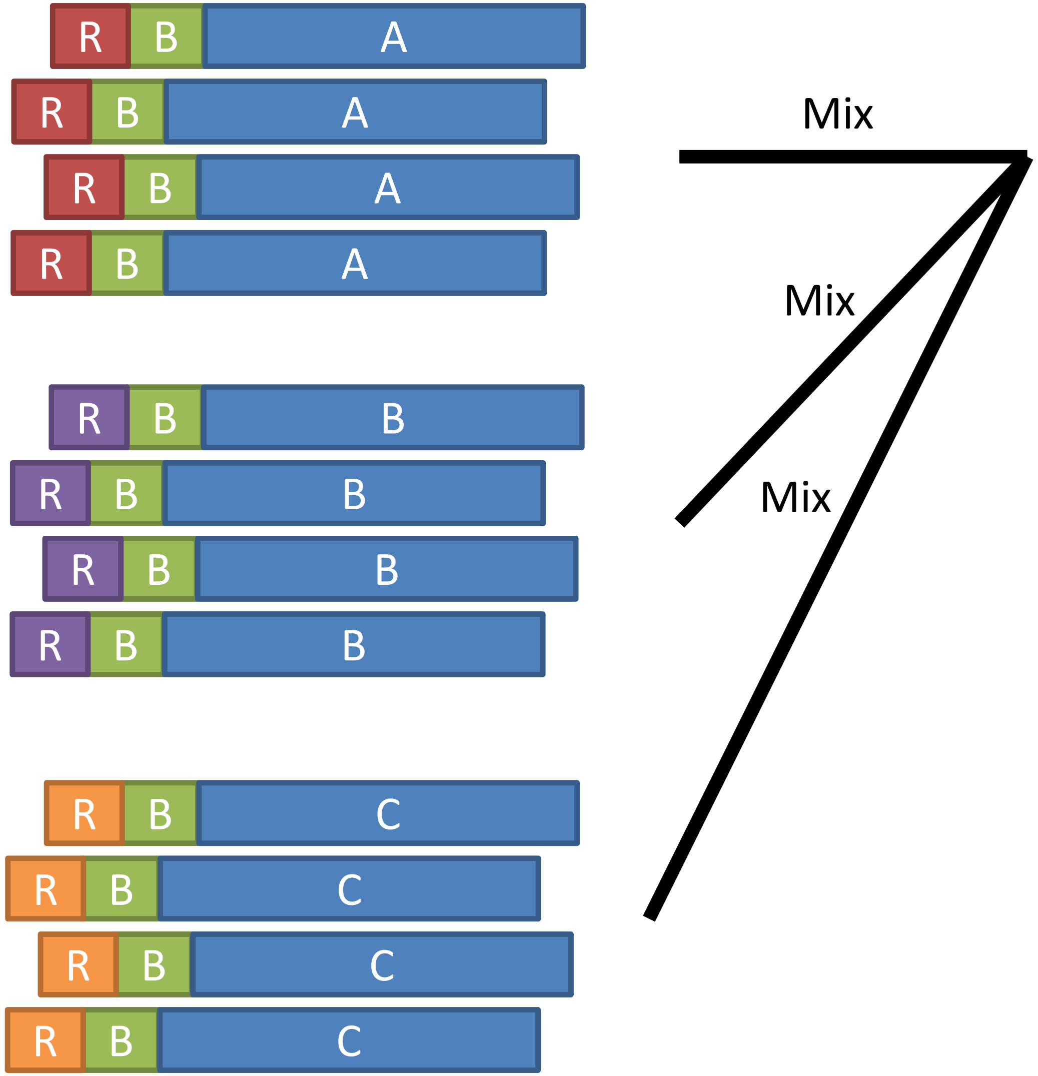
SILAGVR



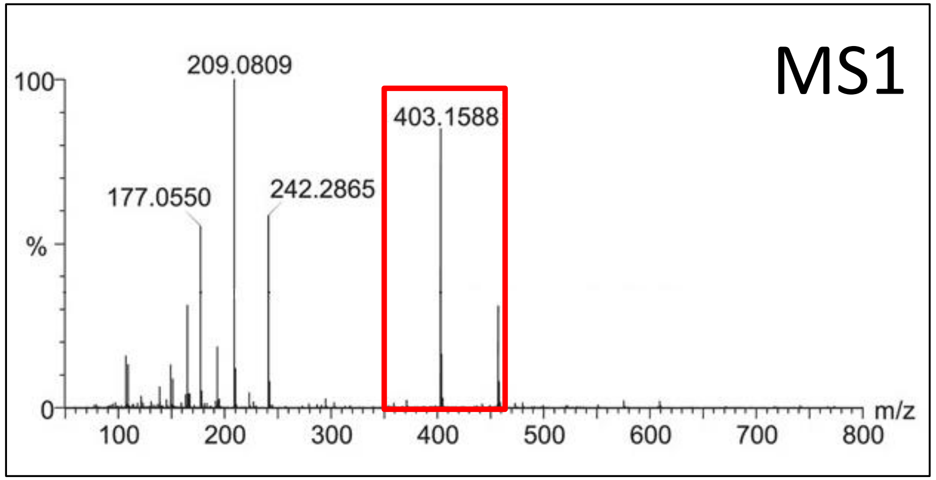
Multiple tags with different reporter masses
Normalisers ensure total tag masses are identical

DOI:10.1155/2014/767812

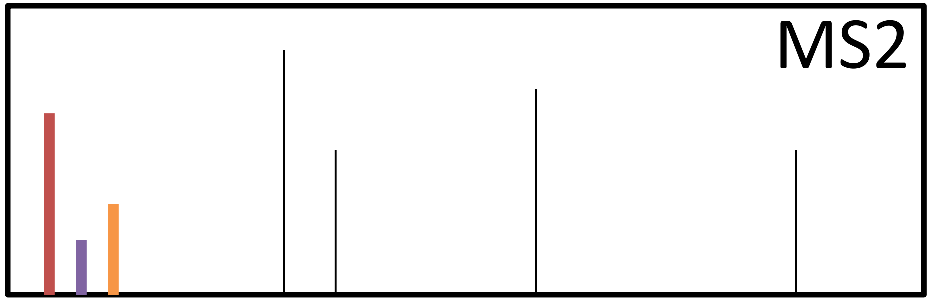
Tandem Mass Tagging



>15 reporters available



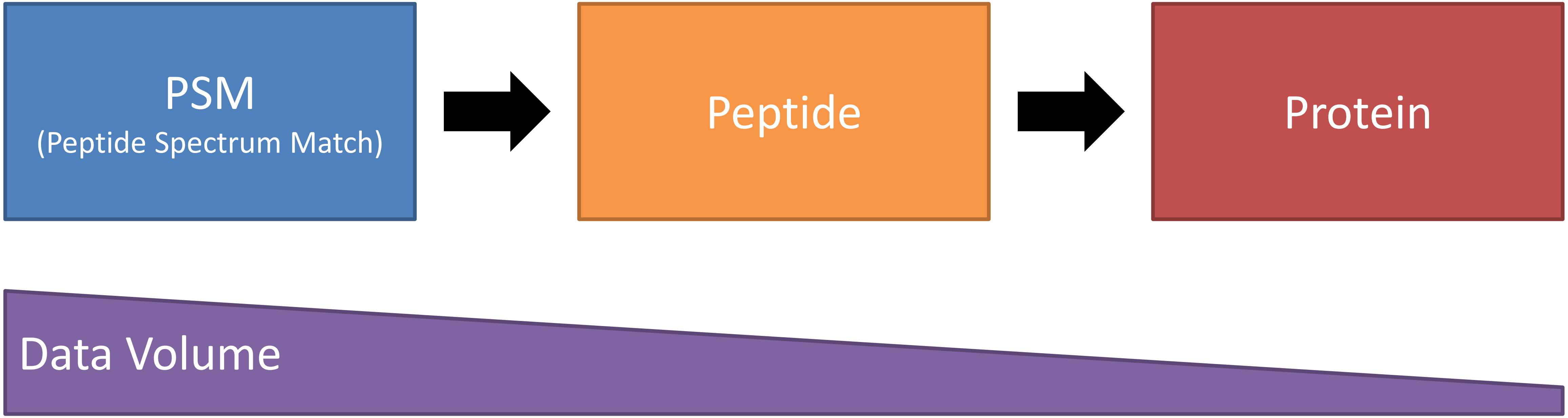
Peptides from all samples run together with a fixed mass shift



Reporters detach leaving a separately quantifiable signal

Moving from peptides to proteins

Levels of quantitation



PSM

(Peptide Spectrum Match)

Peptide

Protein

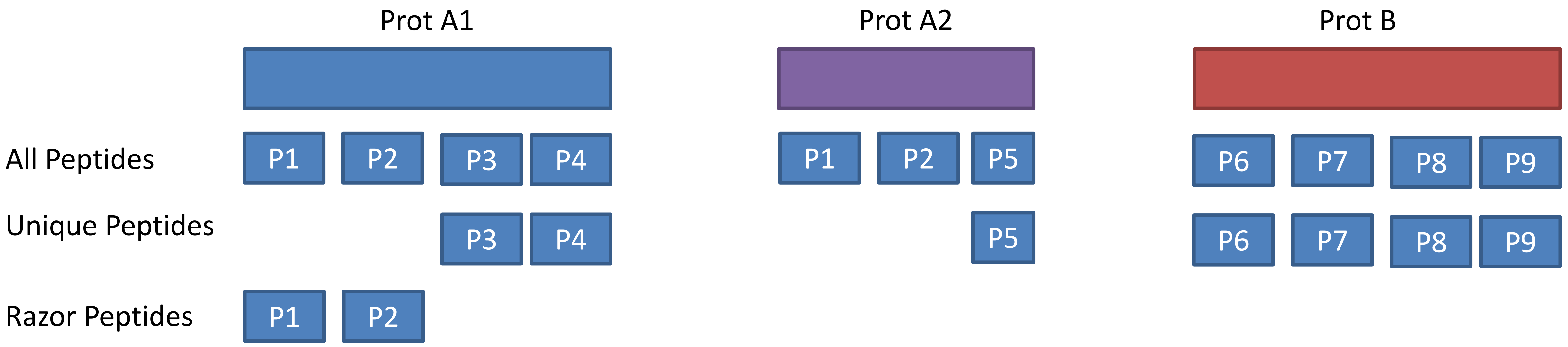
Data Volume

PSMs to Peptides

- One peptide can produce multiple PSMs
  - Different charge states
  - Different modifications
  - Missed cleavage sites

- Combine the intensities for all PSMs for the same peptide
  - Mean
  - Trimmed mean
  - Sum

Peptides to Proteins



	Prot A1	Prot A2	Prot B
All Peptides	P1 P2 P3 P4	P1 P2 P5	P6 P7 P8 P9
Unique Peptides	P3 P4	P5	P6 P7 P8 P9
Razor Peptides	P1 P2		

Assigning Razor Peptides

- Protein with most unique evidence
- Protein with highest molecular weight

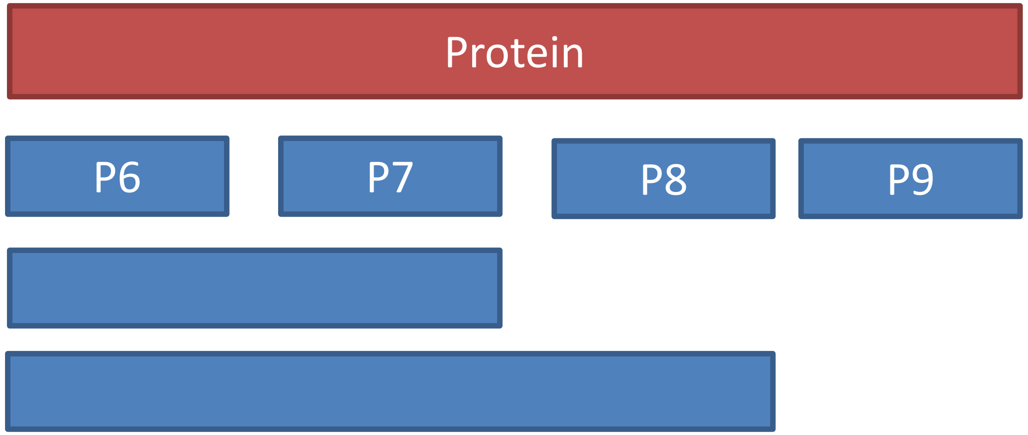
Quantitative value assignment

- Mean of peptide quantitation
- Sum of peptide quantitation
- Highest peptide quantitation

Grouping Proteins

- Multiple proteins which share the same peptides are grouped together
- Different groups can share peptides (Razor Peptides)

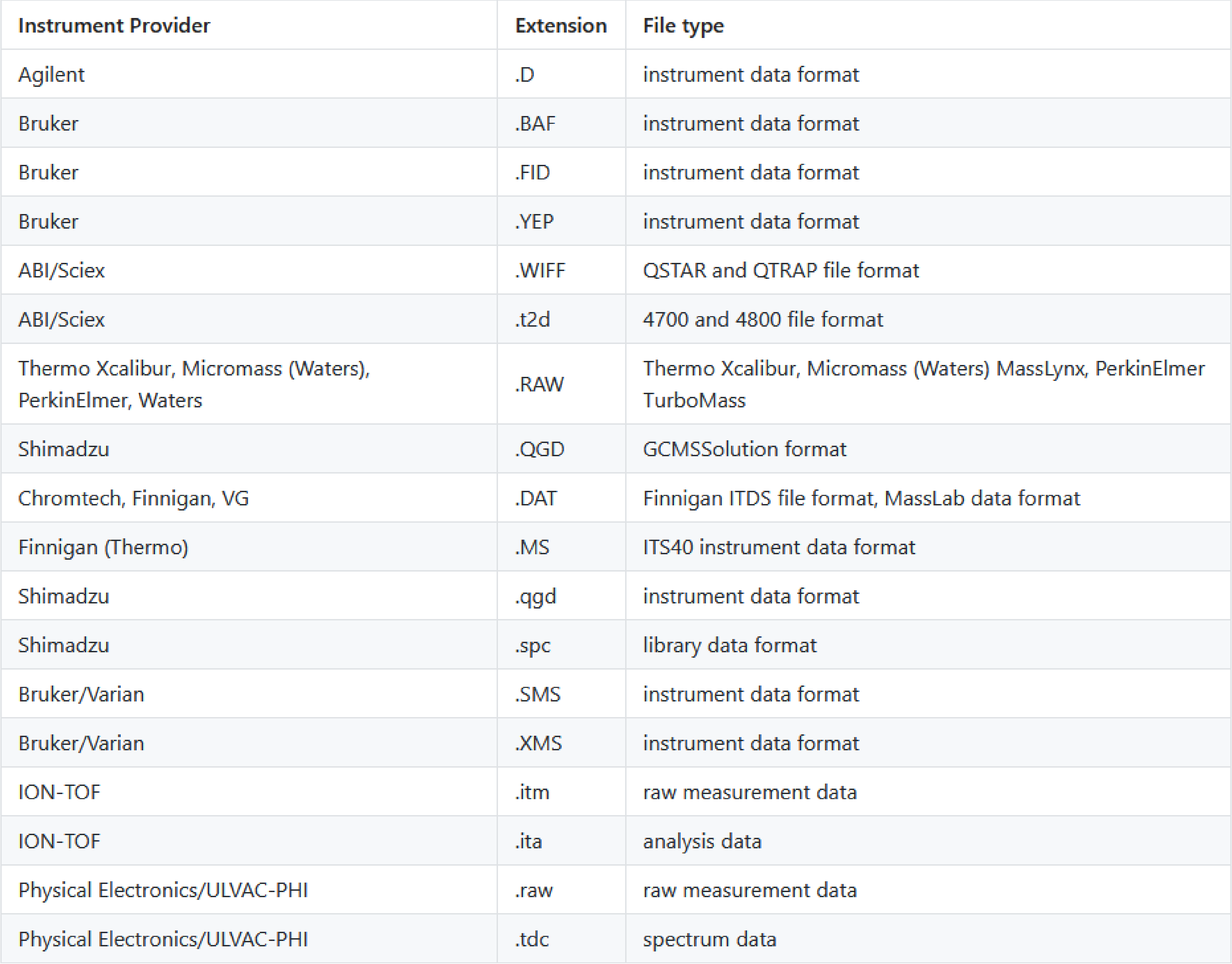
Reported Values



- How many peptides were observed (unique or with razor)
- What percentage of peptides were observed (coverage)
- Missed Cleavages

Proteomics Data Files

Babraham Bioinformatics

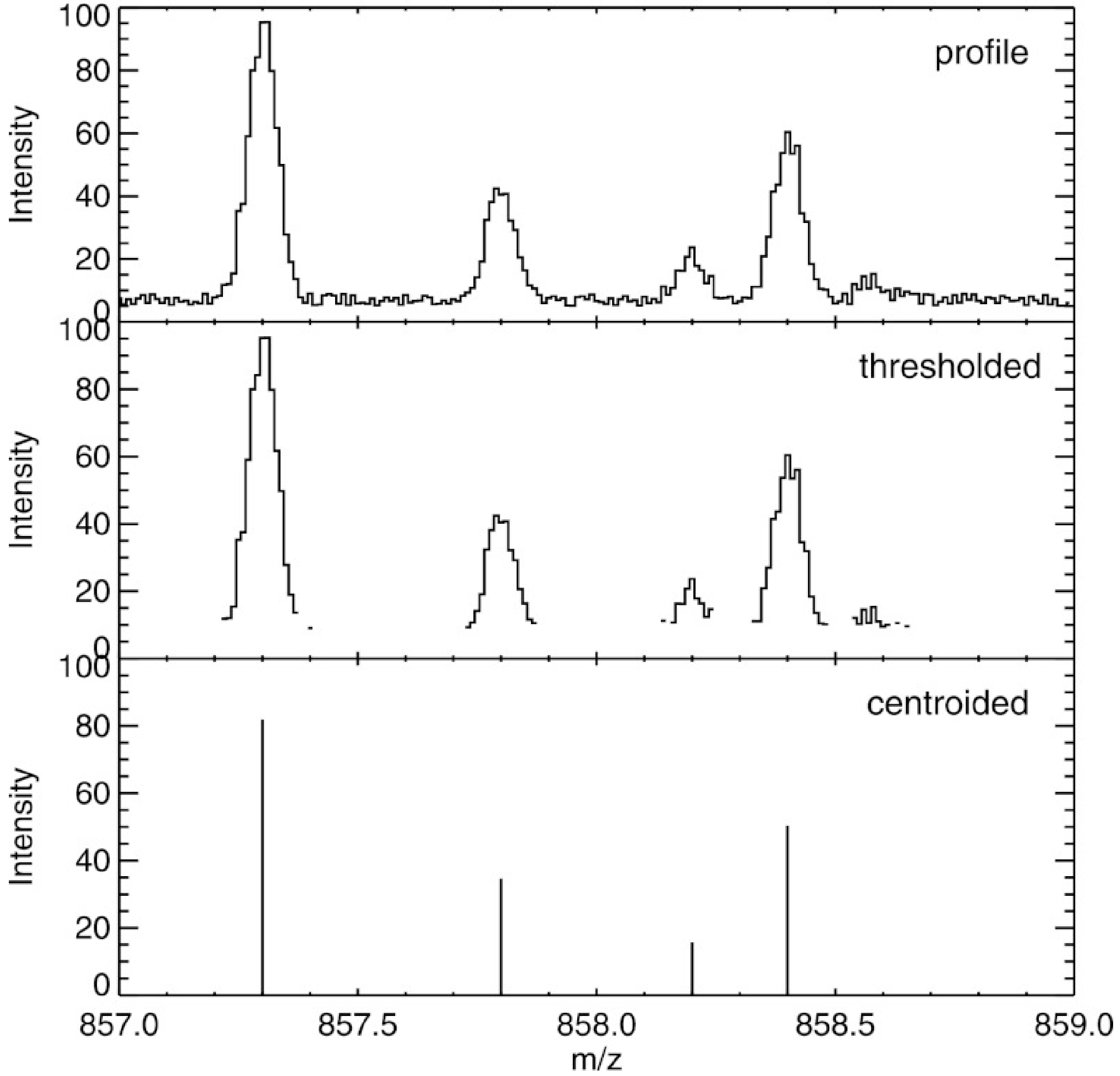


Instrument Provider	Extension	File type
Agilent	.D	instrument data format
Bruker	.BAF	instrument data format
Bruker	.FID	instrument data format
Bruker	.YEP	instrument data format
ABI/Sciex	.WIFF	QSTAR and QTRAP file format
ABI/Sciex	.t2d	4700 and 4800 file format
Thermo Xcalibur, Micromass (Waters), PerkinElmer, Waters	.RAW	Thermo Xcalibur, Micromass (Waters) MassLynx, PerkinElmer TurboMass
Shimadzu	.QGD	GCMSSolution format
Chromtech, Finnigan, VG	.DAT	Finnigan ITDS file format, MassLab data format
Finnigan (Thermo)	.MS	ITS40 instrument data format
Shimadzu	.qgd	instrument data format
Shimadzu	.spc	library data format
Bruker/Varian	.SMS	instrument data format
Bruker/Varian	.XMS	instrument data format
ION-TOF	.itm	raw measurement data
ION-TOF	.ita	analysis data
Physical Electronics/ULVAC-PHI	.raw	raw measurement data
Physical Electronics/ULVAC-PHI	.tdc	spectrum data

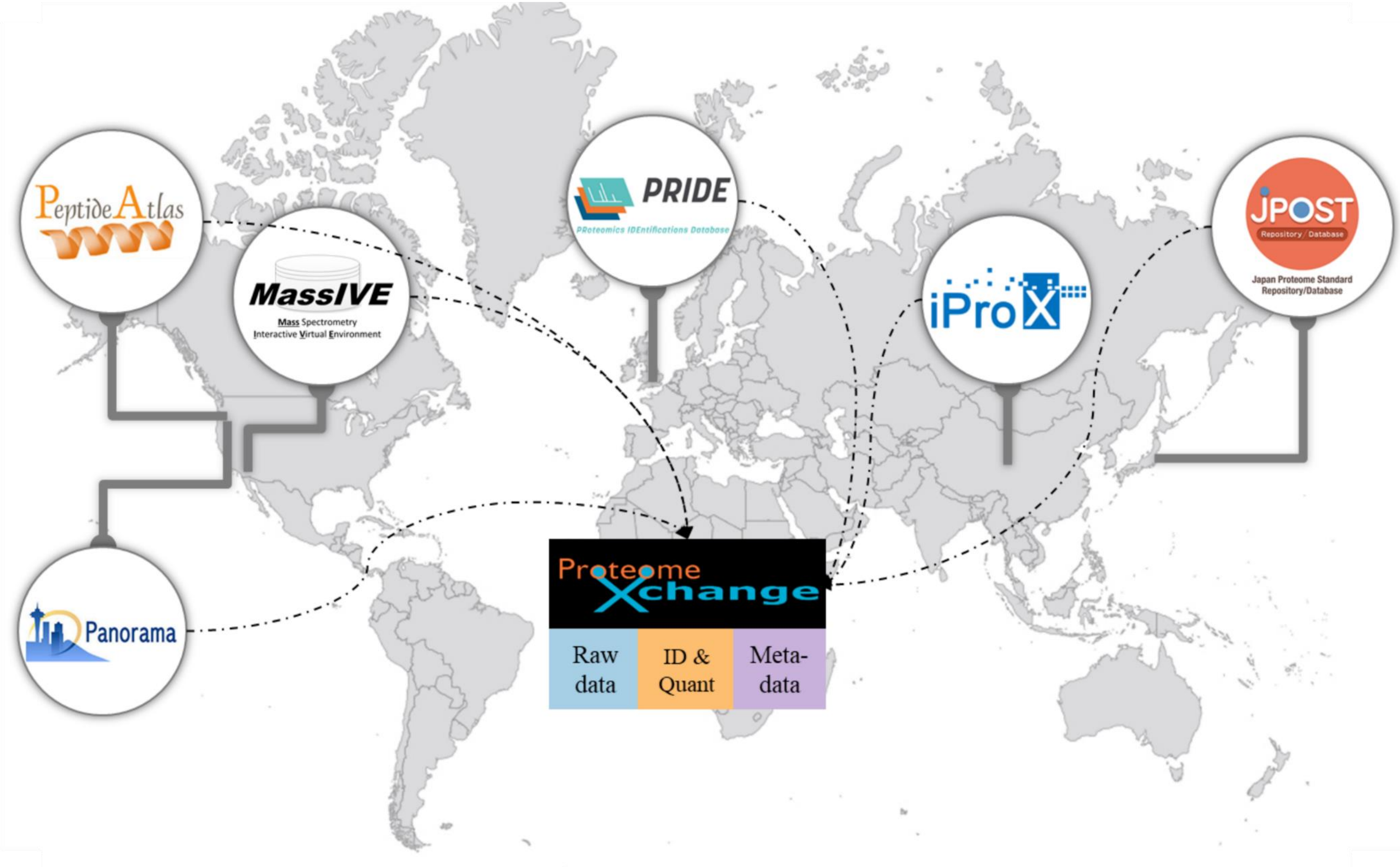
Most common format (>70% of PRIDE)

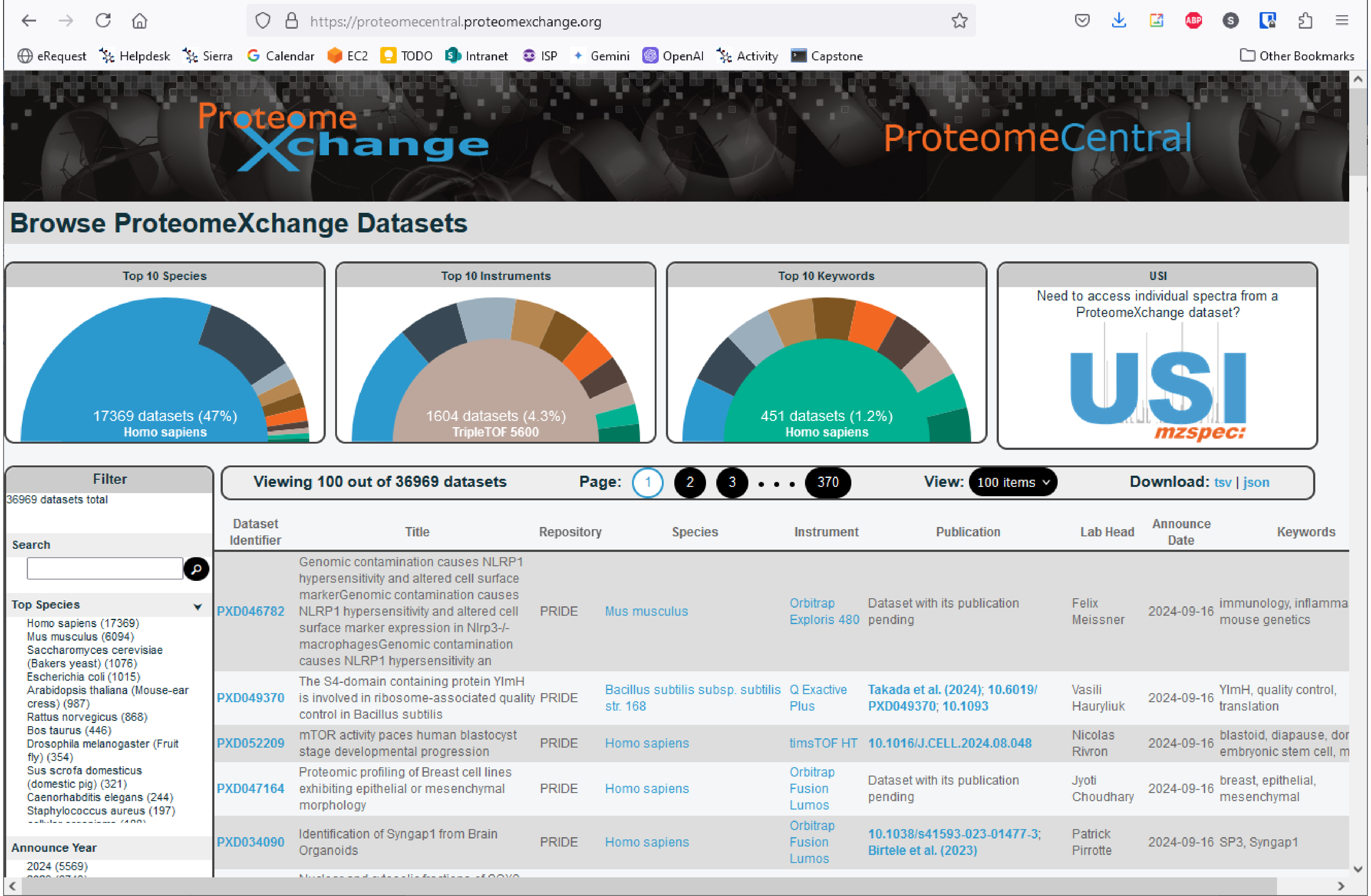
Information in RAW files

- Chromatography times
- Instrument settings
- Spectra (with details)
  - MS1
  - MS2



Data Repositories for Proteomics Mass Spec





Browse ProteomeXchange Datasets

USI: Need to access individual spectra from a ProteomeXchange dataset?

Filter

36969 datasets total

Search

Top Species

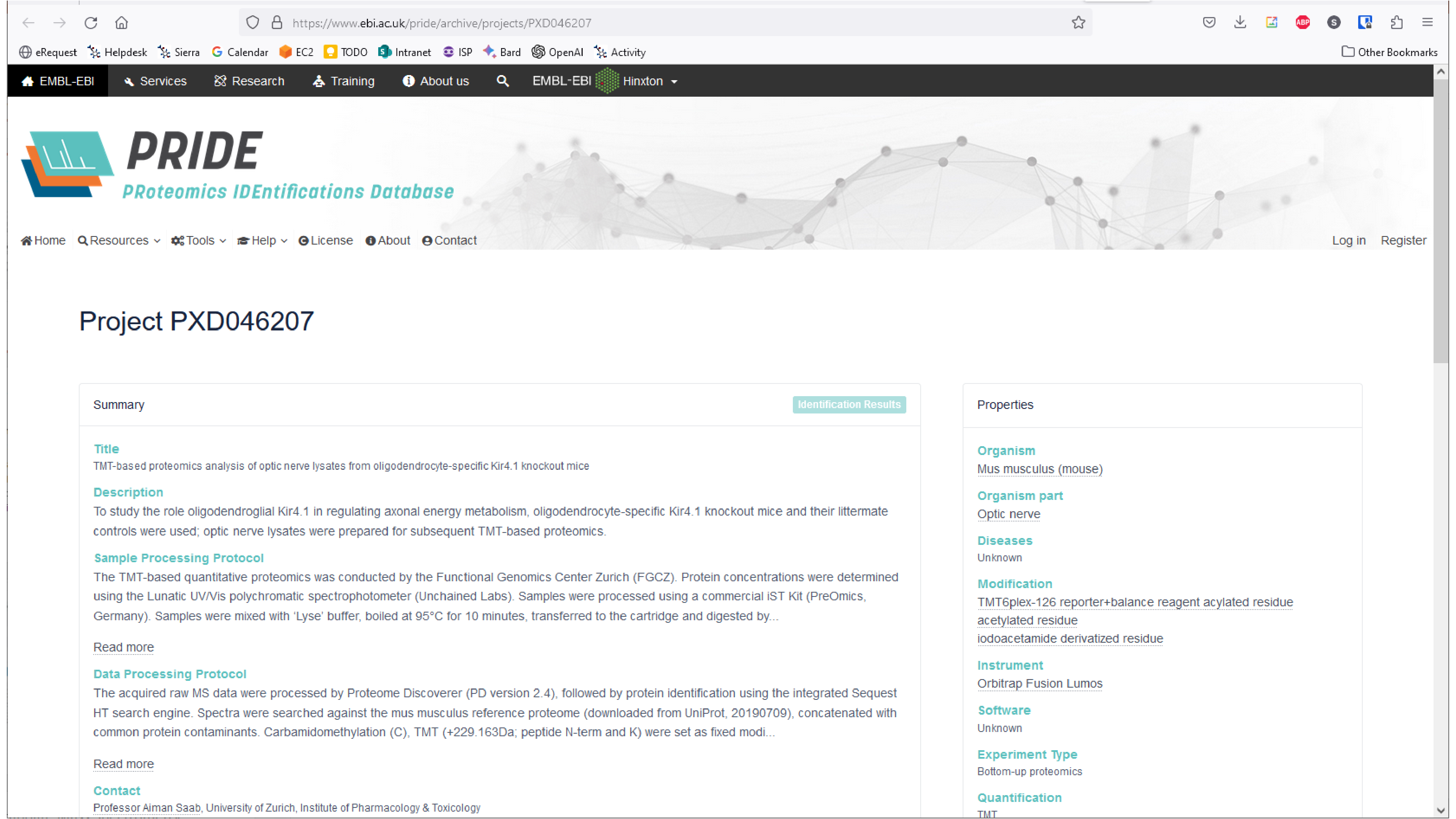
- Homo sapiens (17369)
- Mus musculus (6094)
- Saccharomyces cerevisiae (Bakers yeast) (1076)
- Escherichia coli (1015)
- Arabidopsis thaliana (Mouse-ear cress) (987)
- Rattus norvegicus (868)
- Bos taurus (446)
- Drosophila melanogaster (Fruit fly) (354)
- Sus scrofa domesticus (domestic pig) (321)
- Caenorhabditis elegans (244)
- Staphylococcus aureus (197)

Announce Year

- 2024 (5569)

Viewing 100 out of 36969 datasets **Page:** 1 2 3 . . . 370 **View:** 100 items **Download:** tsv | json

Dataset Identifier	Title	Repository	Species	Instrument	Publication	Lab Head	Announce Date	Keywords
PXD046782	Genomic contamination causes NLRP1 hypersensitivity and altered cell surface markerGenomic contamination causes NLRP1 hypersensitivity and altered cell surface marker expression in Nlrp3-/- macrophagesGenomic contamination causes NLRP1 hypersensitivity an	PRIDE	Mus musculus	Orbitrap Exploris 480	Dataset with its publication pending	Felix Meissner	2024-09-16	immunology, inflamma mouse genetics
PXD049370	The S4-domain containing protein YlmH is involved in ribosome-associated quality control in Bacillus subtilis	PRIDE	Bacillus subtilis subsp. subtilis str. 168	Q Exactive Plus	**Takada et al. (2024)**; **10.6019/PXD049370**; **10.1093**	Vasili Hauryliuk	2024-09-16	YlmH, quality control, translation
PXD052209	mTOR activity paces human blastocyst stage developmental progression	PRIDE	Homo sapiens	timsTOF HT	**10.1016/J.CELL.2024.08.048**	Nicolas Rivron	2024-09-16	blastoid, diapause, dor embryonic stem cell, m
PXD047164	Proteomic profiling of Breast cell lines exhibiting epithelial or mesenchymal morphology	PRIDE	Homo sapiens	Orbitrap Fusion Lumos	Dataset with its publication pending	Jyoti Choudhary	2024-09-16	breast, epithelial, mesenchymal
PXD034090	Identification of Syngap1 from Brain Organoids	PRIDE	Homo sapiens	Orbitrap Fusion Lumos	**10.1038/s41593-023-01477-3**; **Birtele et al. (2023)**	Patrick Pirrotte	2024-09-16	SP3, Syngap1



Project PXD046207

Summary

Title

TMT-based proteomics analysis of optic nerve lysates from oligodendrocyte-specific Kir4.1 knockout mice

Description

To study the role oligodendroglial Kir4.1 in regulating axonal energy metabolism, oligodendrocyte-specific Kir4.1 knockout mice and their littermate controls were used; optic nerve lysates were prepared for subsequent TMT-based proteomics.

Sample Processing Protocol

The TMT-based quantitative proteomics was conducted by the Functional Genomics Center Zurich (FGCZ). Protein concentrations were determined using the Lunatic UV/Vis polychromatic spectrophotometer (Unchained Labs). Samples were processed using a commercial iST Kit (PreOmics, Germany). Samples were mixed with 'Lyse' buffer, boiled at 95°C for 10 minutes, transferred to the cartridge and digested by...

Read more

Data Processing Protocol

The acquired raw MS data were processed by Proteome Discoverer (PD version 2.4), followed by protein identification using the integrated Sequest HT search engine. Spectra were searched against the mus musculus reference proteome (downloaded from UniProt, 20190709), concatenated with common protein contaminants. Carbamidomethylation (C), TMT (+229.163Da; peptide N-term and K) were set as fixed modi...

Read more

Contact

Professor Aiman Saab, University of Zurich, Institute of Pharmacology & Toxicology

Properties

Organism

Mus musculus (mouse)

Organism part

Optic nerve

Diseases

Unknown

Modification

TMT6plex-126 reporter+balance reagent acylated residue

acetylated residue

iodoacetamide derivatized residue

Instrument

Orbitrap Fusion Lumos

Software

Unknown

Experiment Type

Bottom-up proteomics

Quantification

TMT

Problems with public data

Things that are well recorded

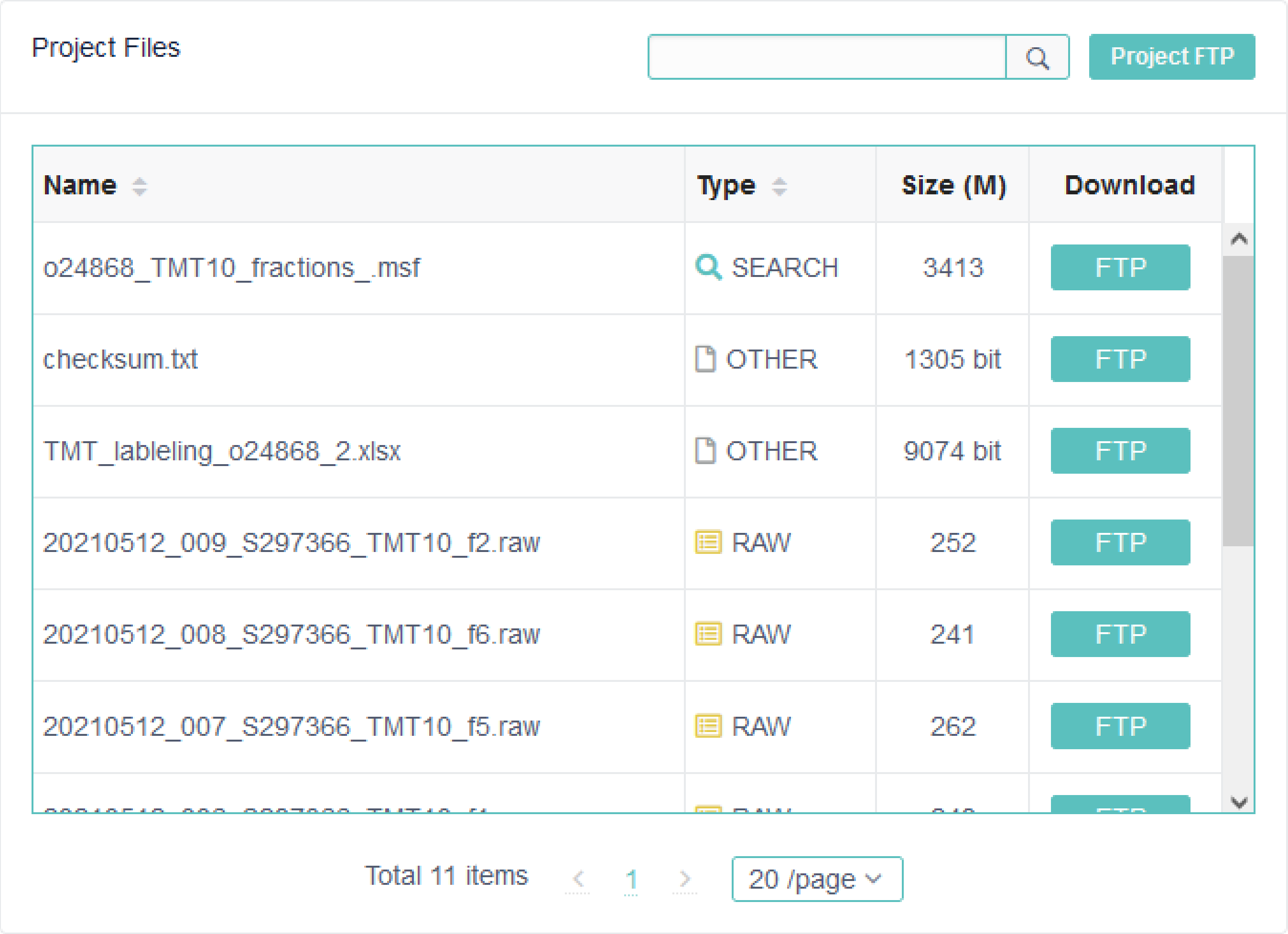
- Mass spec collection metrics
- Organism
- Modifications
- (Search method)

Things that are NOT recorded

- Sample details
- Experimental Conditions
- Link from RAW files to samples

Finding data is simple. Downloading RAW files is easy. Figuring out which sample is which can be a complete nightmare.

Files to download



Project Files

Name	Type	Size (M)	Download
o24868_TMT10_fractions_.msf	SEARCH	3413	FTP
checksum.txt	OTHER	1305 bit	FTP
TMT_lableling_o24868_2.xlsx	OTHER	9074 bit	FTP
20210512_009_S297366_TMT10_f2.raw	RAW	252	FTP
20210512_008_S297366_TMT10_f6.raw	RAW	241	FTP
20210512_007_S297366_TMT10_f5.raw	RAW	262	FTP

Total 11 items

Quantitated search results

Sample metadata

Raw spectrum files

Exercise

Finding Data in Public Repositories

Running a Database Search

Babraham
Bioinformatics

Main Information Required

- Which RAW file(s) are you analysing?
- Which sequences do you want to search against?
- Which type of quantitation are you using?
- How did you digest your peptides?
- What modifications do you expect to be present?

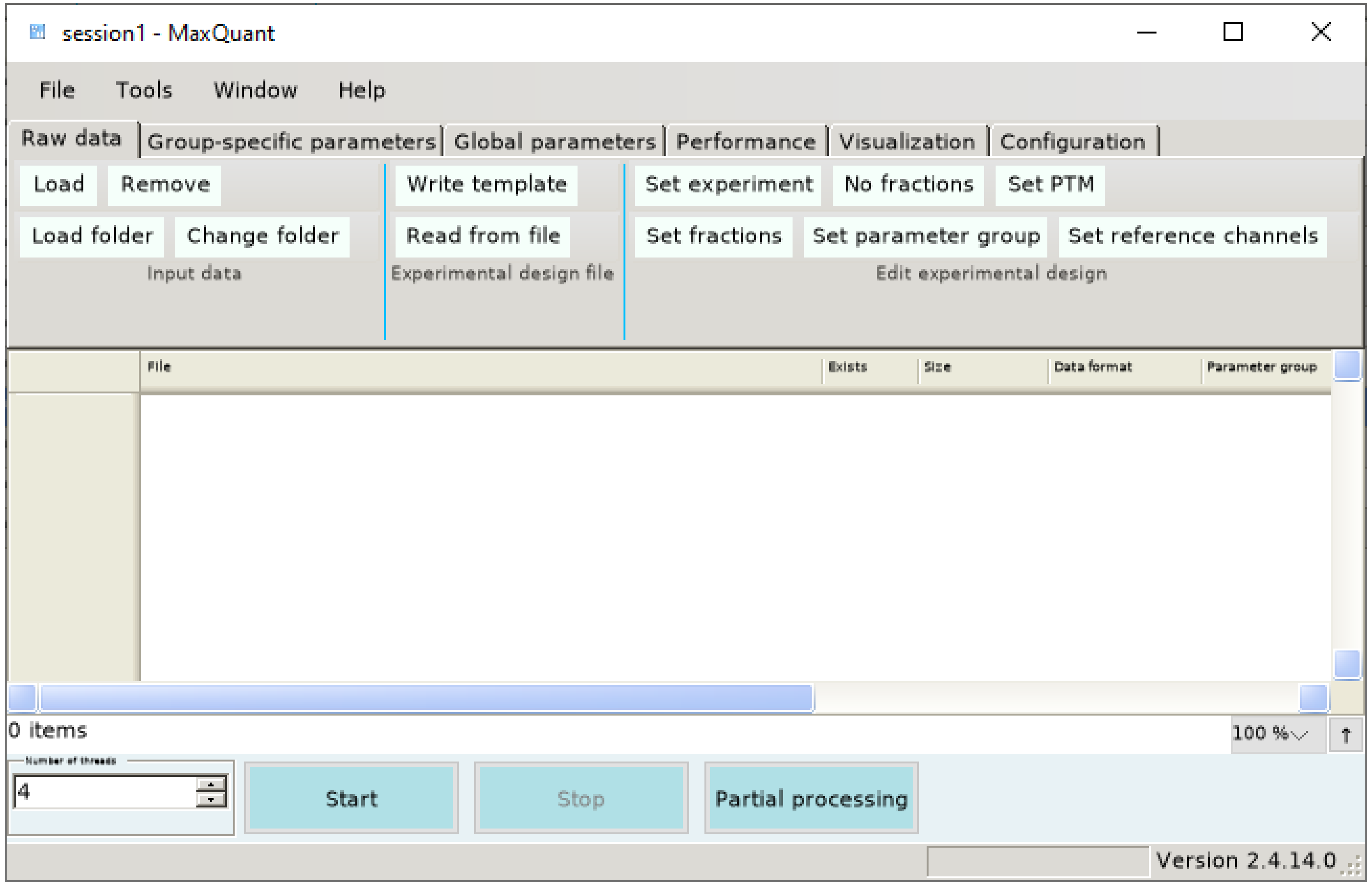
- Specific thresholds
  - Mass accuracy
  - LC time flexibility
  - Statistical thresholds

Normally either left at defaults, or set based on the machine you're using

Running MaxQuant (Label Free)

- Set Data
- Set Cores
- Set Search Sequences
- Set Quantitation
- Save Parameters

- Run search



0 items

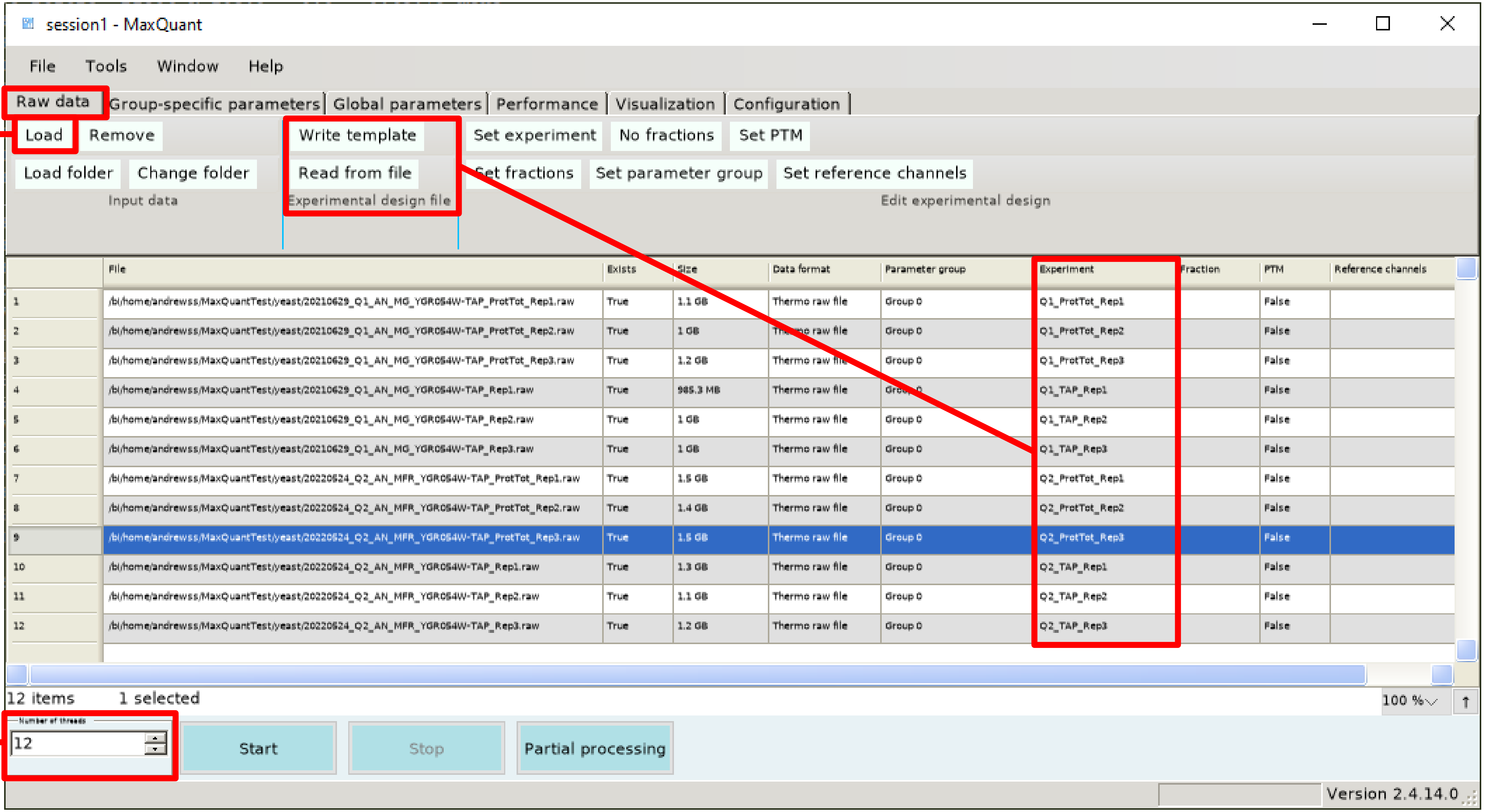
Number of threads: 4

Version 2.4.14.0

Load Raw Files

Select RAW files

Select Cores



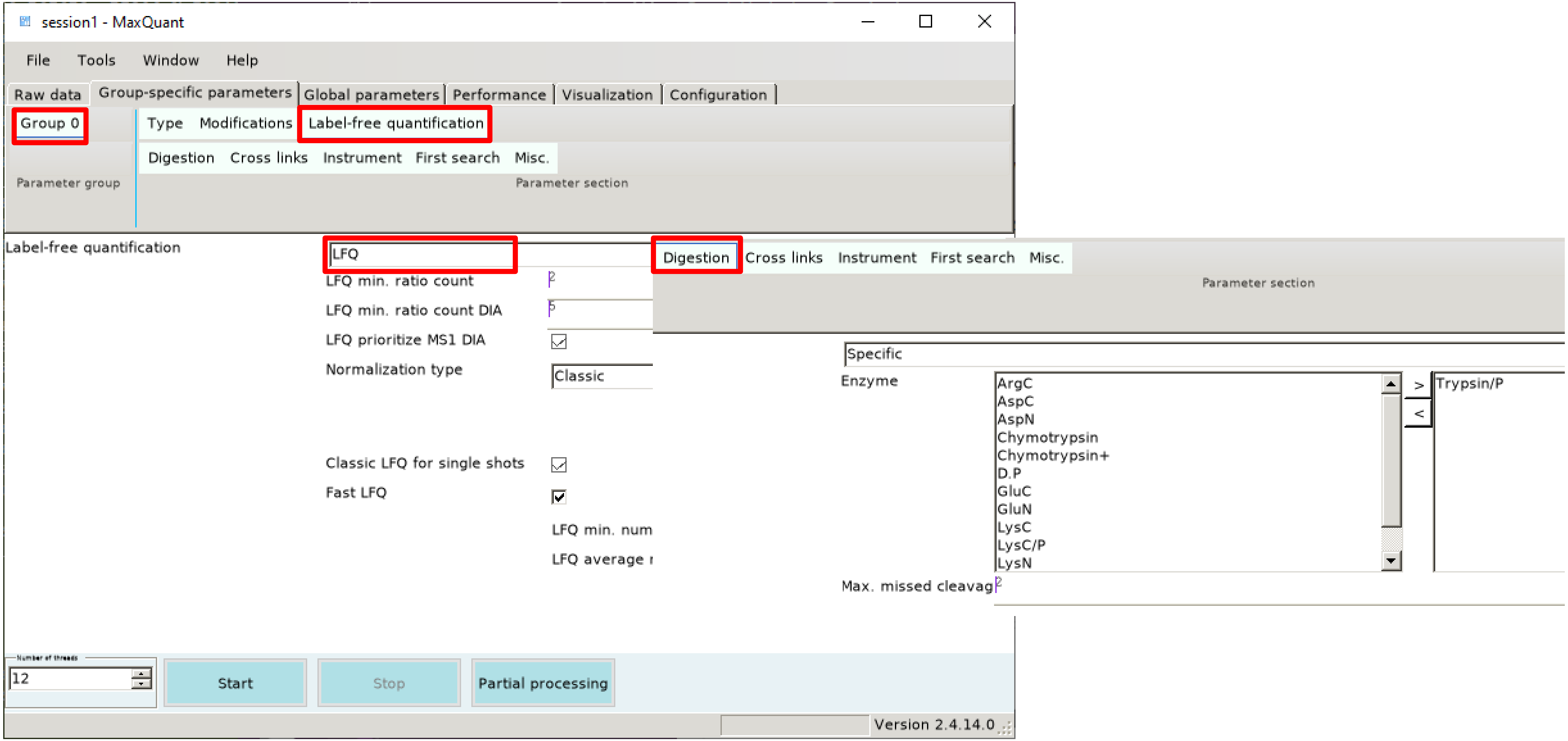
	File	Exists	Size	Data format	Parameter group	Experiment	Fraction	PTM	Reference channels
1	/bi/home/andrewss/MaxQuantTest/yeast/20210629_Q1_AN_MG_YGR054W-TAP_ProtTot_Rep1.raw	True	1.1 GB	Thermo raw file	Group 0	Q1_ProtTot_Rep1		False	
2	/bi/home/andrewss/MaxQuantTest/yeast/20210629_Q1_AN_MG_YGR054W-TAP_ProtTot_Rep2.raw	True	1 GB	Thermo raw file	Group 0	Q1_ProtTot_Rep2		False	
3	/bi/home/andrewss/MaxQuantTest/yeast/20210629_Q1_AN_MG_YGR054W-TAP_ProtTot_Rep3.raw	True	1.2 GB	Thermo raw file	Group 0	Q1_ProtTot_Rep3		False	
4	/bi/home/andrewss/MaxQuantTest/yeast/20210629_Q1_AN_MG_YGR054W-TAP_Rep1.raw	True	985.3 MB	Thermo raw file	Group 0	Q1_TAP_Rep1		False	
5	/bi/home/andrewss/MaxQuantTest/yeast/20210629_Q1_AN_MG_YGR054W-TAP_Rep2.raw	True	1 GB	Thermo raw file	Group 0	Q1_TAP_Rep2		False	
6	/bi/home/andrewss/MaxQuantTest/yeast/20210629_Q1_AN_MG_YGR054W-TAP_Rep3.raw	True	1 GB	Thermo raw file	Group 0	Q1_TAP_Rep3		False	
7	/bi/home/andrewss/MaxQuantTest/yeast/20220524_Q2_AN_MFR_YGR054W-TAP_ProtTot_Rep1.raw	True	1.5 GB	Thermo raw file	Group 0	Q2_ProtTot_Rep1		False	
8	/bi/home/andrewss/MaxQuantTest/yeast/20220524_Q2_AN_MFR_YGR054W-TAP_ProtTot_Rep2.raw	True	1.4 GB	Thermo raw file	Group 0	Q2_ProtTot_Rep2		False	
9	/bi/home/andrewss/MaxQuantTest/yeast/20220524_Q2_AN_MFR_YGR054W-TAP_ProtTot_Rep3.raw	True	1.5 GB	Thermo raw file	Group 0	Q2_ProtTot_Rep3		False	
10	/bi/home/andrewss/MaxQuantTest/yeast/20220524_Q2_AN_MFR_YGR054W-TAP_Rep1.raw	True	1.3 GB	Thermo raw file	Group 0	Q2_TAP_Rep1		False	
11	/bi/home/andrewss/MaxQuantTest/yeast/20220524_Q2_AN_MFR_YGR054W-TAP_Rep2.raw	True	1.1 GB	Thermo raw file	Group 0	Q2_TAP_Rep2		False	
12	/bi/home/andrewss/MaxQuantTest/yeast/20220524_Q2_AN_MFR_YGR054W-TAP_Rep3.raw	True	1.2 GB	Thermo raw file	Group 0	Q2_TAP_Rep3		False	

12 items 1 selected

Number of threads: 12

Version 2.4.14.0

Set Quantitation

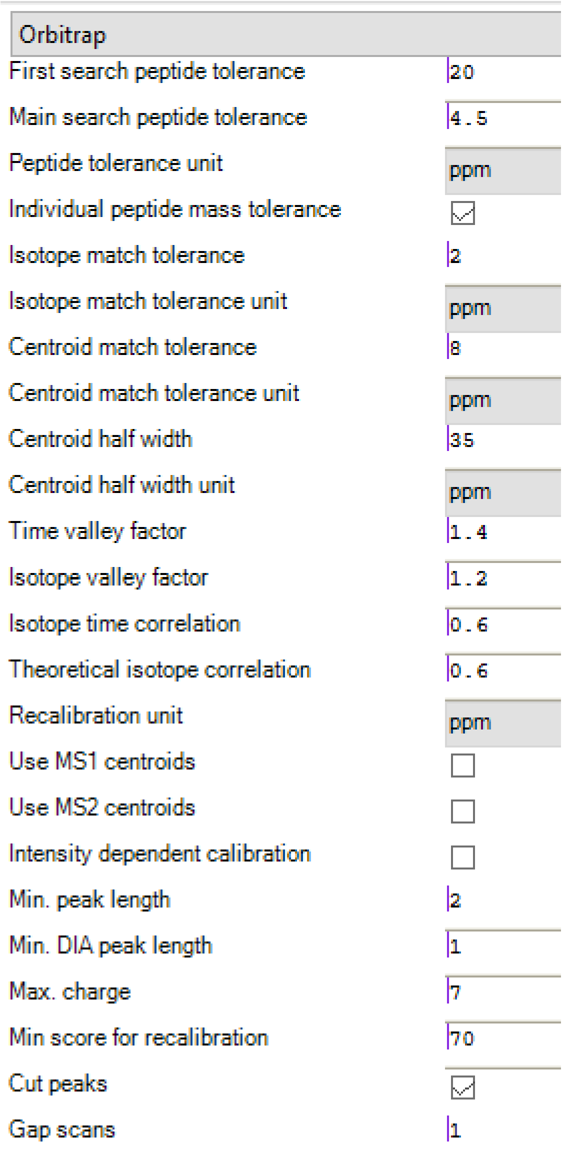


Label-free quantification: LFQ
LFQ min. ratio count: 2
LFQ min. ratio count DIA: 5
LFQ prioritize MS1 DIA: ☑
Normalization type: Classic
Classic LFQ for single shots: ☑
Fast LFQ: ☑
LFQ min. num
LFQ average r

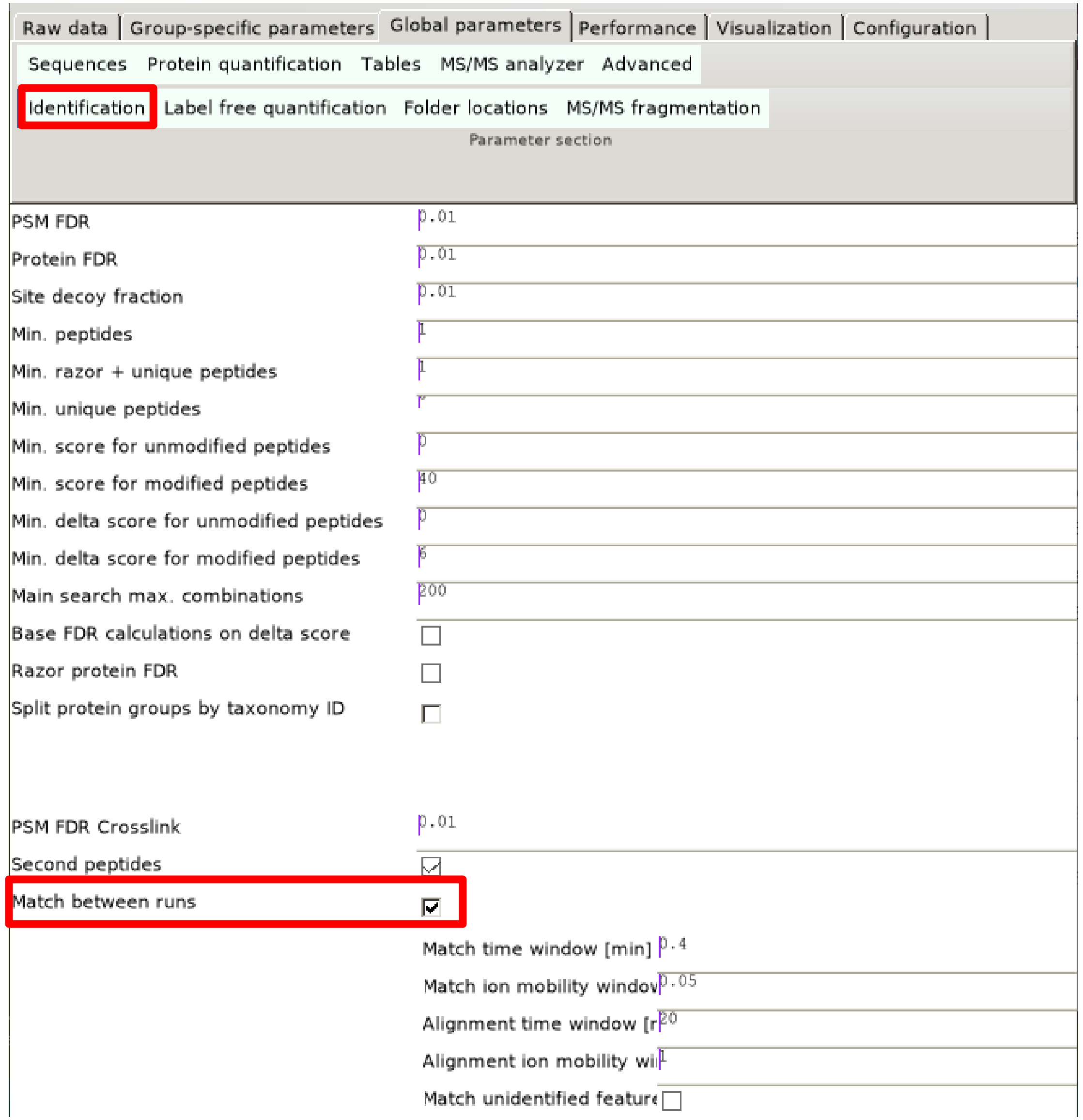
Specific
Enzyme: ArgC, AspC, AspN, Chymotrypsin, Chymotrypsin+, D.P, GluC, GluN, LysC, LysC/P, LysN
Selected: Trypsin/P
Max. missed cleavag: 2

Number of threads: 12

Identification Parameters

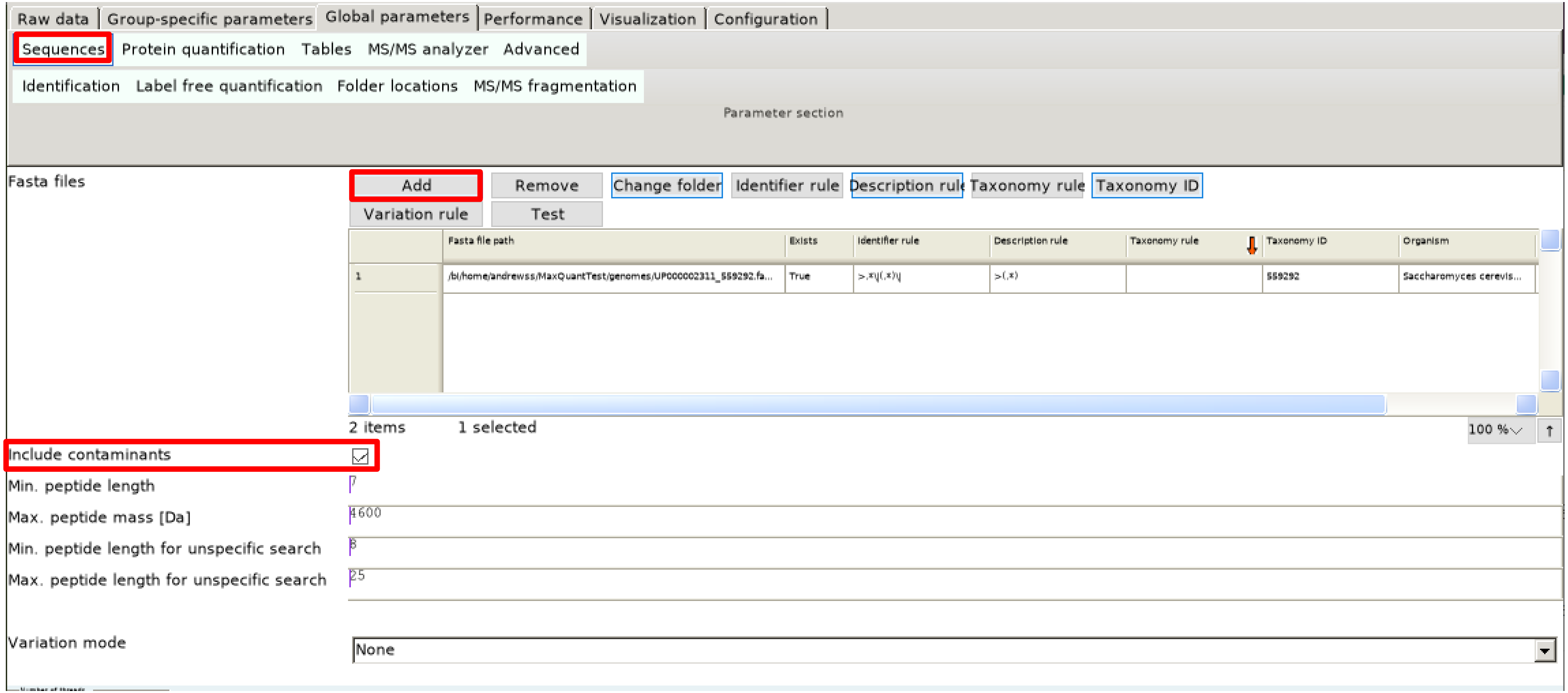


Orbitrap	
First search peptide tolerance	20
Main search peptide tolerance	4.5
Peptide tolerance unit	ppm
Individual peptide mass tolerance	☑
Isotope match tolerance	2
Isotope match tolerance unit	ppm
Centroid match tolerance	8
Centroid match tolerance unit	ppm
Centroid half width	35
Centroid half width unit	ppm
Time valley factor	1.4
Isotope valley factor	1.2
Isotope time correlation	0.6
Theoretical isotope correlation	0.6
Recalibration unit	ppm
Use MS1 centroids	☐
Use MS2 centroids	☐
Intensity dependent calibration	☐
Min. peak length	2
Min. DIA peak length	1
Max. charge	7
Min score for recalibration	70
Cut peaks	☑
Gap scans	1



PSM FDR	0.01
Protein FDR	0.01
Site decoy fraction	0.01
Min. peptides	1
Min. razor + unique peptides	1
Min. unique peptides	
Min. score for unmodified peptides	0
Min. score for modified peptides	40
Min. delta score for unmodified peptides	0
Min. delta score for modified peptides	6
Main search max. combinations	200
Base FDR calculations on delta score	☐
Razor protein FDR	☐
Split protein groups by taxonomy ID	☐
PSM FDR Crosslink	0.01
Second peptides	☑
Match between runs	☑
Match time window [min]	0.4
Match ion mobility window	0.05
Alignment time window [r	20
Alignment ion mobility wi	1
Match unidentified feature	☐

Search Sequences



Fasta files

	Fasta file path	Exists	Identifier rule	Description rule	Taxonomy rule	Taxonomy ID	Organism
1	/bl/home/andrewss/MaxQuantTest/genomes/UP000002311_559292.fa...	True	>.*|(.*)|	>(.*)		559292	Saccharomyces cerevis...

2 items 1 selected

Include contaminants ☑

Min. peptide length 7

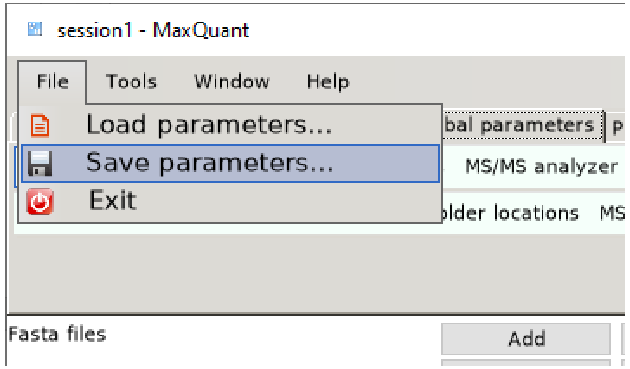
Max. peptide mass [Da] 4600

Min. peptide length for unspecific search 8

Max. peptide length for unspecific search 25

Variation mode None

Saving and Running



```
$ ls -l mqpar.xml
-rw-rw-r-- 1 andrewss bioinf 29631 Aug 20 10:09 mqpar.xml
```

```
maxquant_cmd mqpar.xml
```

```
ssub -o mqcmd.log --cores=12 --mem=20G maxquant_cmd mqpar.xml
```

Easier searches with `mqtemplate`

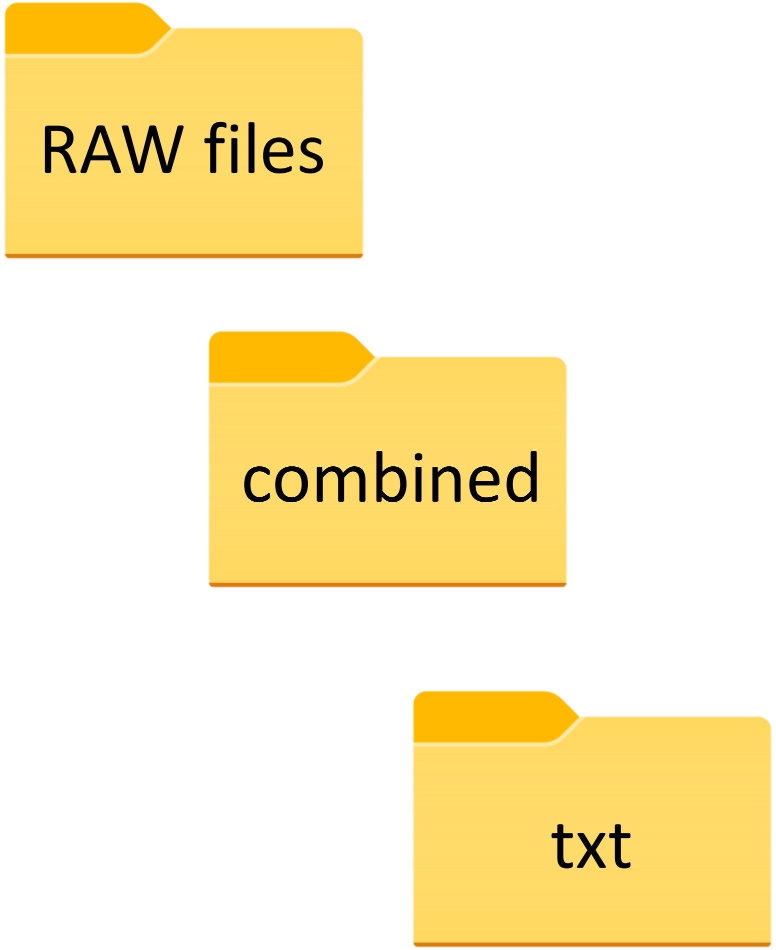
```
mqtemplate --template lfq --proteome mouse *raw
```

```
Proteome file is /bi/apps/mqtemplate/latest/proteomes/mouse_UP000000589_2024_08_23.fa
Template file is /bi/apps/mqtemplate/latest/templates/lfq.xml
Writing mqpar to /bi/home/andrewss/MaxQuantTest/example/mqpar.xml

Command to start searching:

ssub -o mqcmd.log --cores=12 --mem=24G maxquant_cmd mqpar.xml
```

Log File Whilst Running

```
Configuring
Assemble run info
Finish run info
Testing fasta files
Testing raw files

Feature detection
Deisotoping
MS/MS preparation
Calculating peak properties
Combining apl files for first search
Preparing searches
MS/MS first search

Read search results for recalibration
Mass recalibration
Calculating masses
MS/MS preparation for main search
Combining apl files for main search
MS/MS main search
Preparing combined folder
Correcting errors
Reading search engine results
Preparing reverse hits
Finish search engine results
Filter identifications (MS/MS)
Calculating PEP
Copying identifications
Applying FDR

Assembling second peptide MS/MS
Combining second peptide files
Second peptide search
Reading search engine results (SP)
Finish search engine results (SP)
Filtering identifications (SP)
Applying FDR (SP)
Re-quantification
Reporter quantification
Retention time alignment
Matching between runs 1
Matching between runs 2
Matching between runs 3
Matching between runs 4

Prepare protein assembly
Assembling proteins
Assembling unidentified peptides
Finish protein assembly
Updating identifications

Label-free preparation
Label-free normalization
Label-free quantification
Label-free collect
Estimating complexity
Prepare writing tables
Writing tables
Finish writing tables
```

Output Files



`evidence.txt`

All of the quantified data at PSM level

`summary.txt`

Overall summary metrics for the run

`proteinGroups.txt`

Details of the proteins which were joined

Quality Control of Search Results

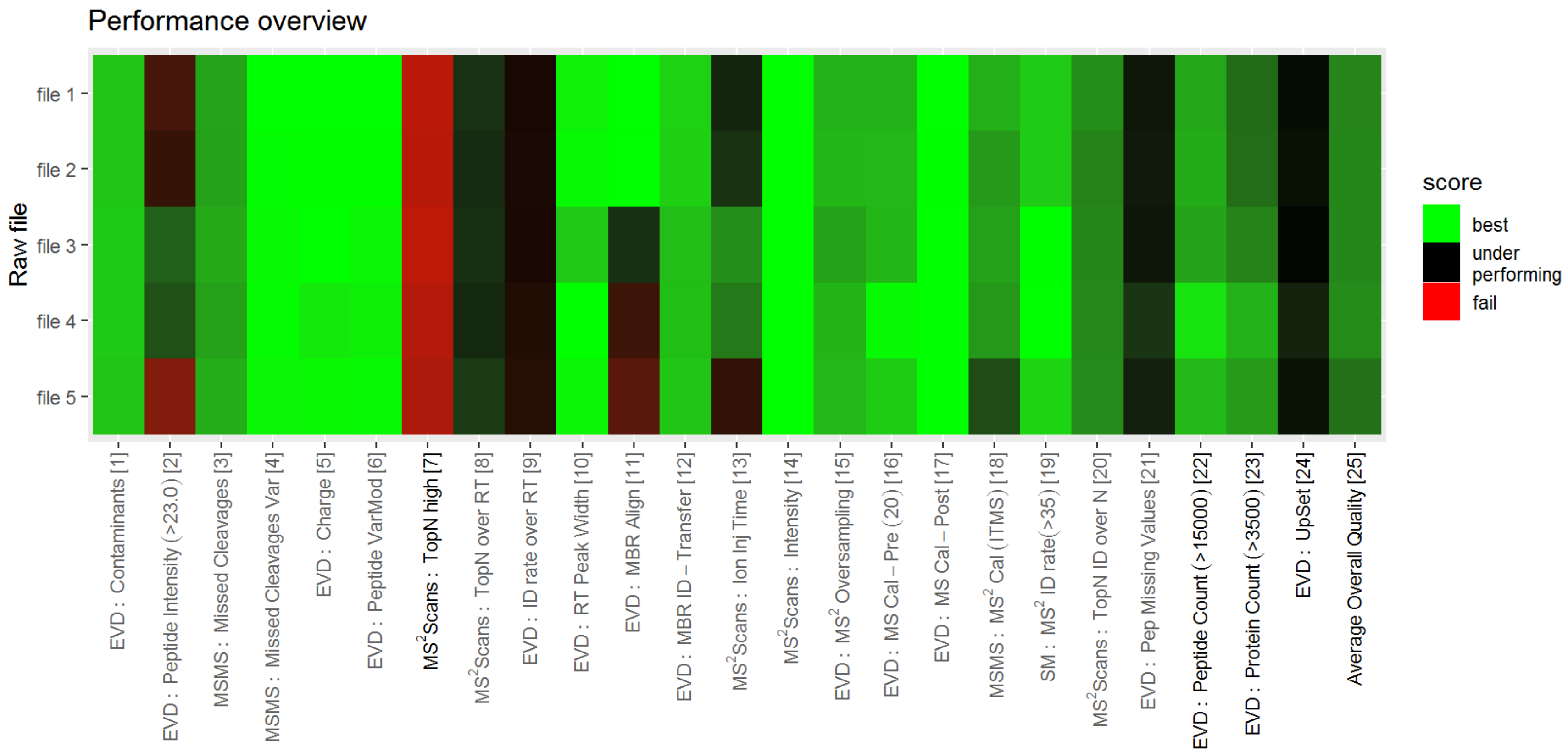
1. Problems during sample preparation
    - Digestion failed
    - Sample Contaminated
    - Low sample amount

2. Problems during Chromatography
    - Even amounts of data over time
    - Consistent rate between experiments

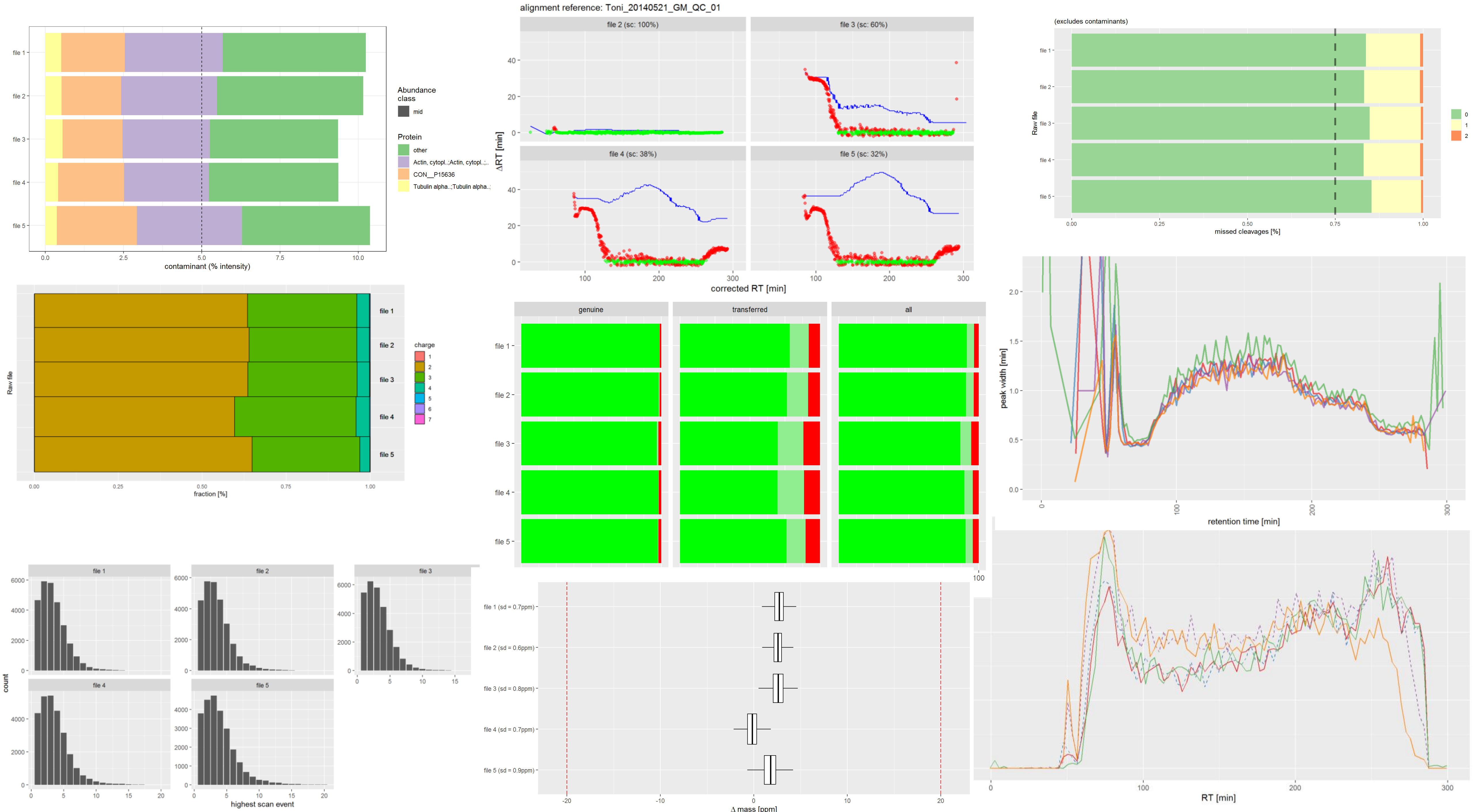
3. Problems with the Mass Spec
    - Poor mass accuracy
    - Poor matching to reference

PTXQC

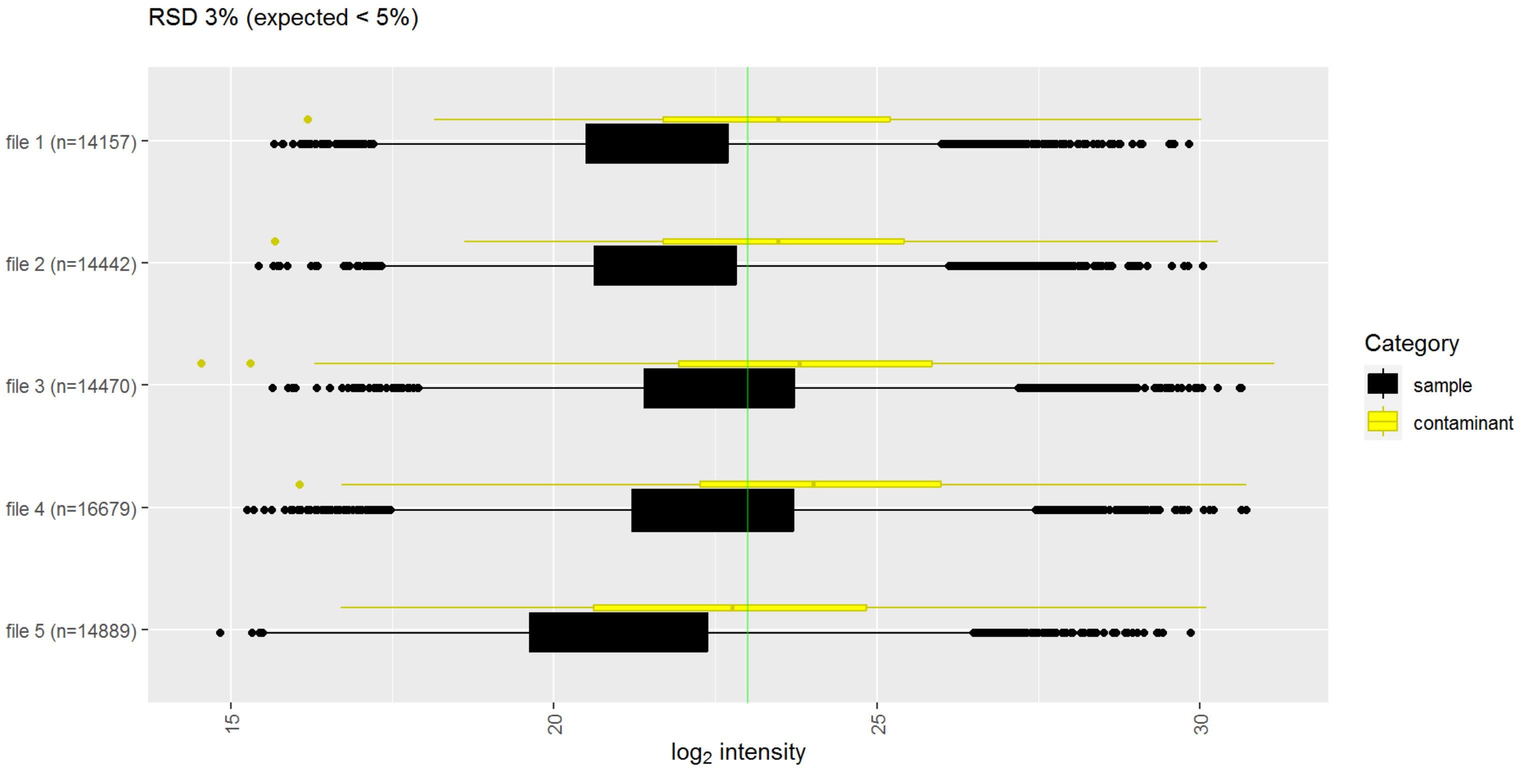
- R package – calculates a QC report from MQ or MzTab



https://cran.rstudio.com/web/packages/PTXQC/index.html

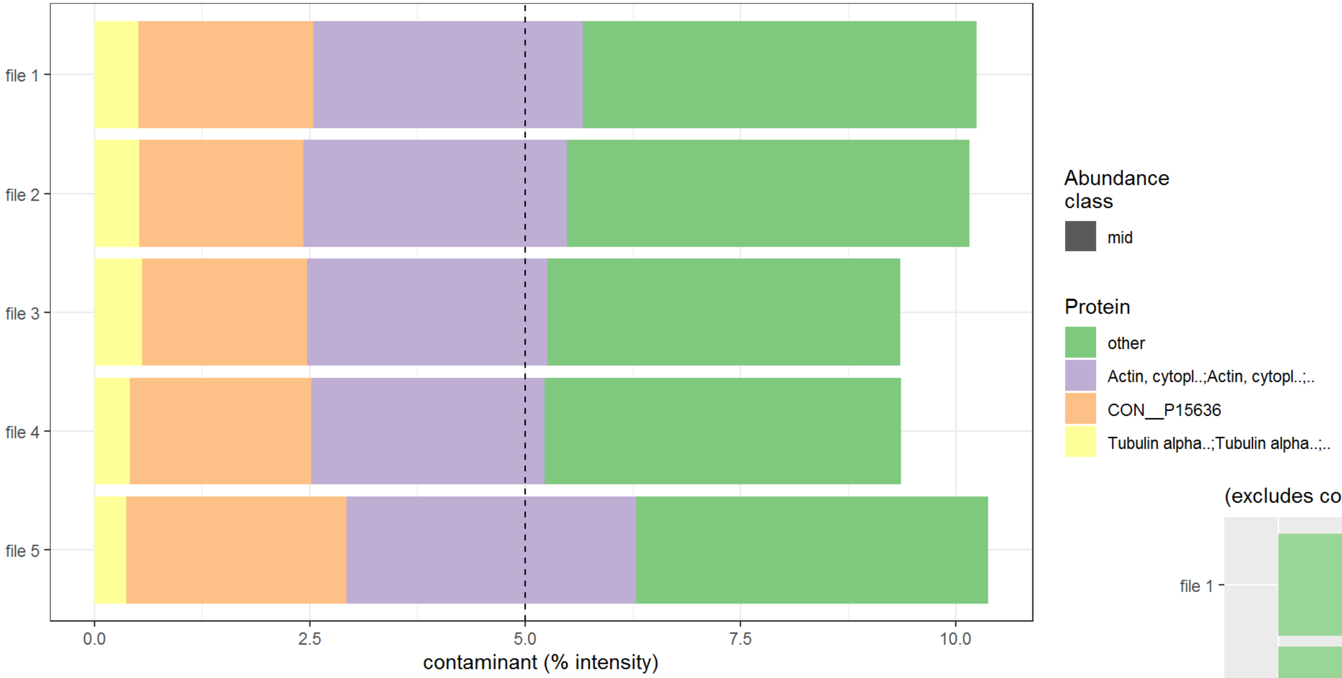

file 1
file 2
file 3
file 4
file 5
0.0
2.5
5.0
7.5
10.0
contaminant (% intensity)
Abundance class
mid
Protein
other
Actin, cytopl.;Actin, cytopl.;...
CON__P15636
Tubulin alpha..;Tubulin alpha..
alignment reference: Toni_20140521_GM_QC_01
file 2 (sc: 100%)
file 3 (sc: 60%)
file 4 (sc: 38%)
file 5 (sc: 32%)
ΔRT [min]
40
20
0
100
200
300
corrected RT [min]
(excludes contaminants)
Raw file
0.00
0.25
0.50
0.75
1.00
missed cleavages [%]
0
1
2
fraction [%]
charge
1
2
3
4
5
6
7
genuine
transferred
all
peak width [min]
2.0
1.5
1.0
0.5
0.0
retention time [min]
count
highest scan event
6000
4000
2000
0
5
10
15
20
file 1 (sd = 0.7ppm)
file 2 (sd = 0.6ppm)
file 3 (sd = 0.8ppm)
file 4 (sd = 0.7ppm)
file 5 (sd = 0.9ppm)
-20
-10
Δ mass [ppm]
RT [min]

Loading and Abundance



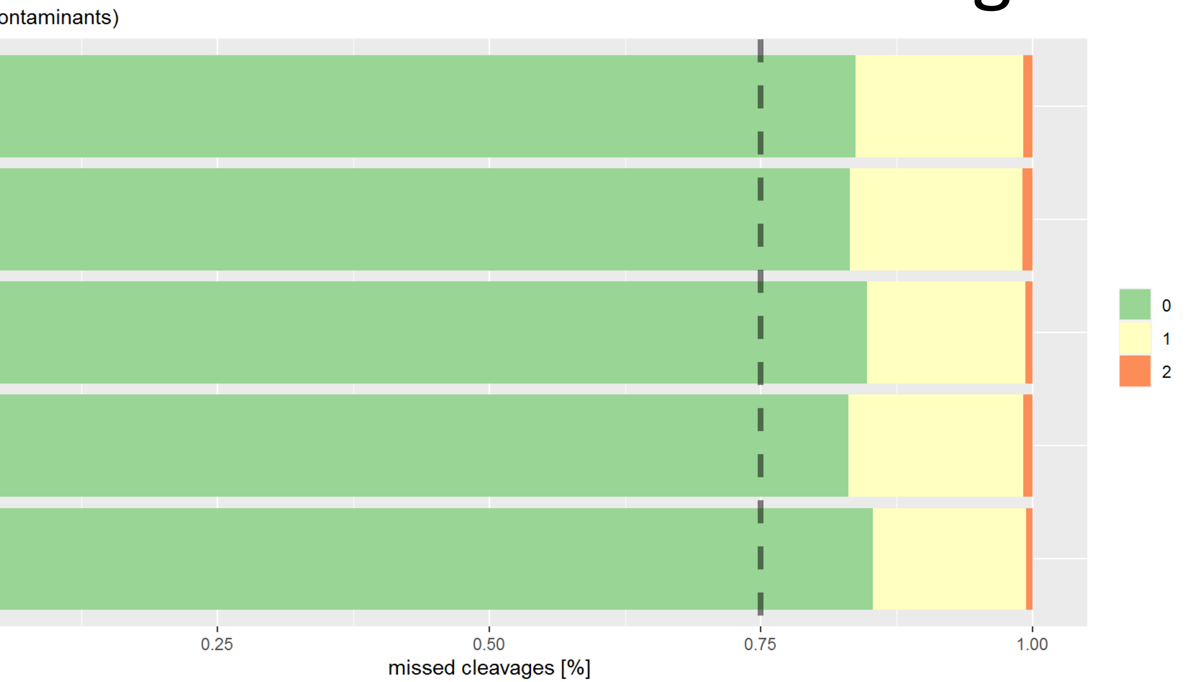
- Should be equal (ish)
- Lower is worse
  - Underloaded
  - Poor column
- RSD is reproducibility between files

Digestion and Contaminants

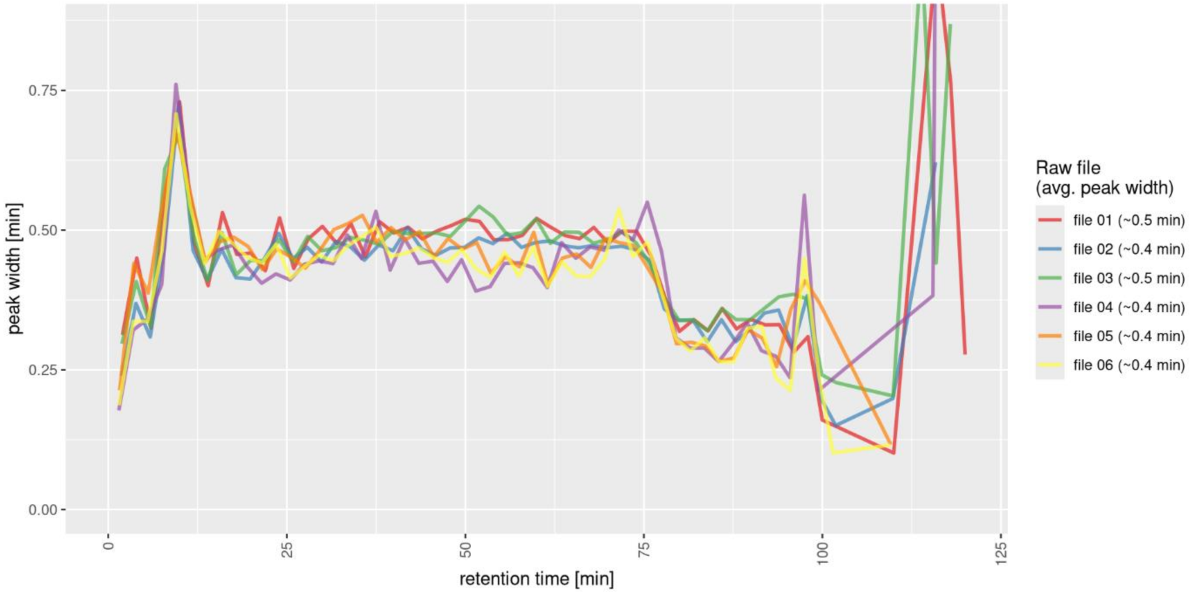


Contaminant Abundance

Missed Cleavages

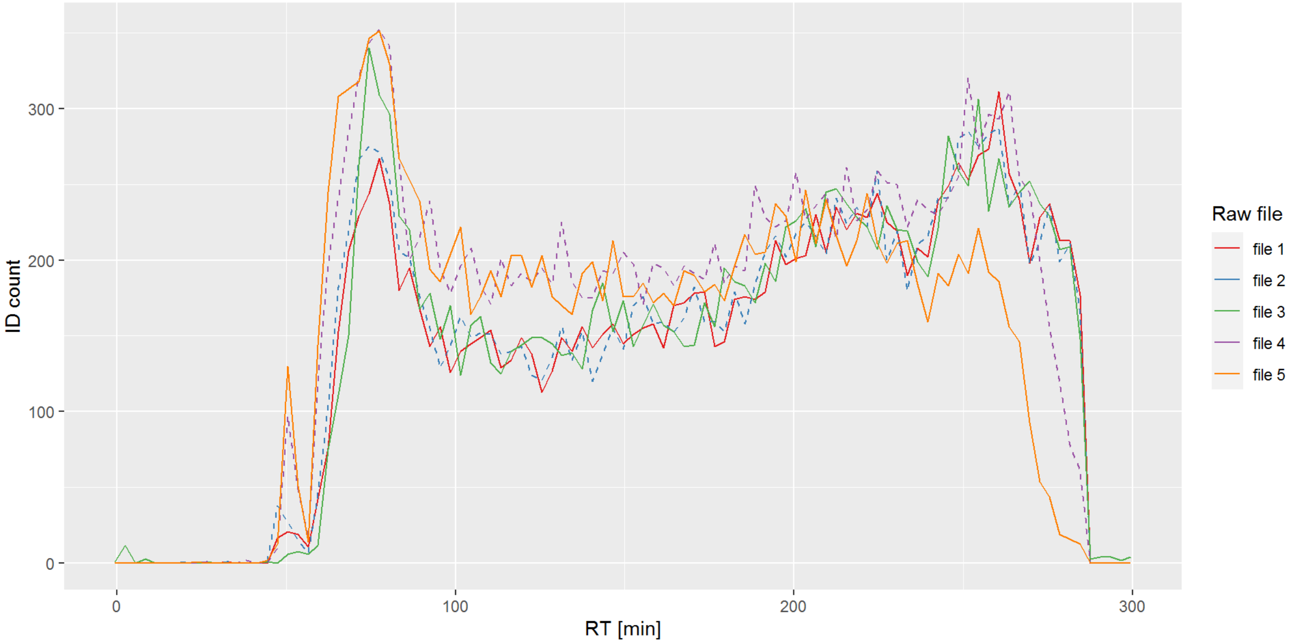


Chromatography Consistency

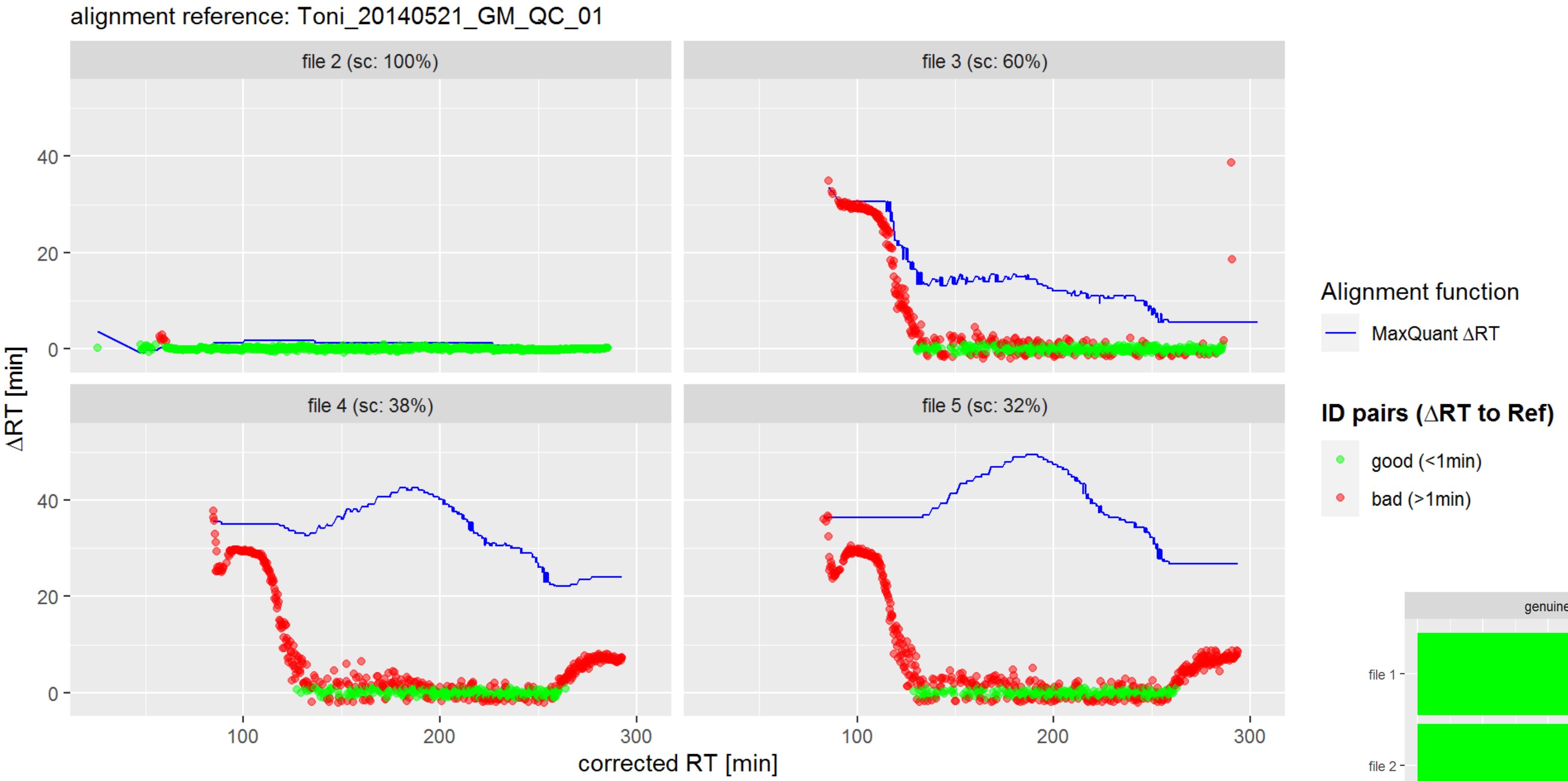


MS1 Peak Width

Consistent peptides over time

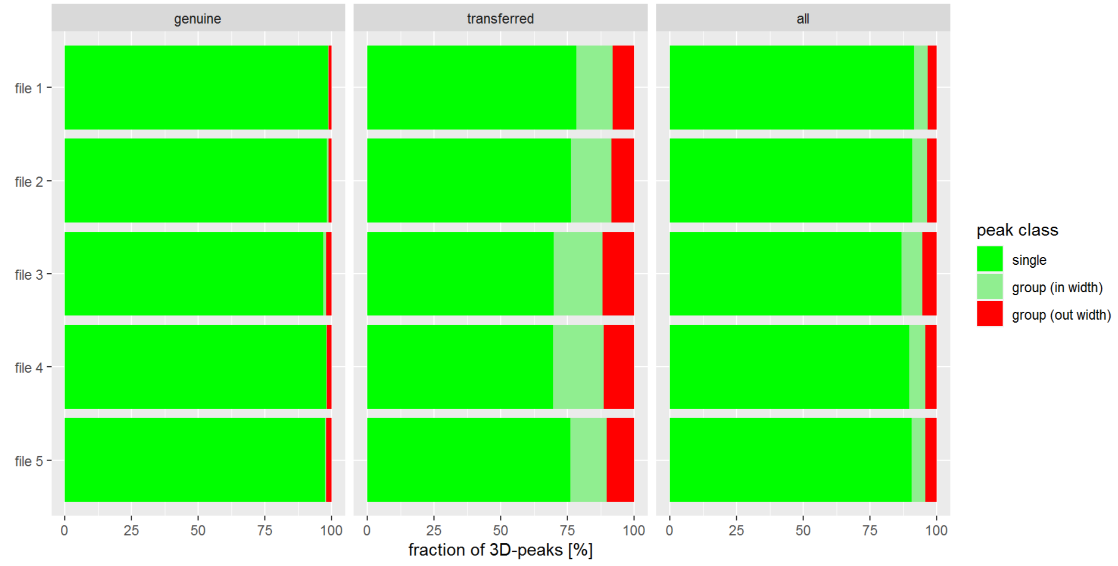


Match Between Runs

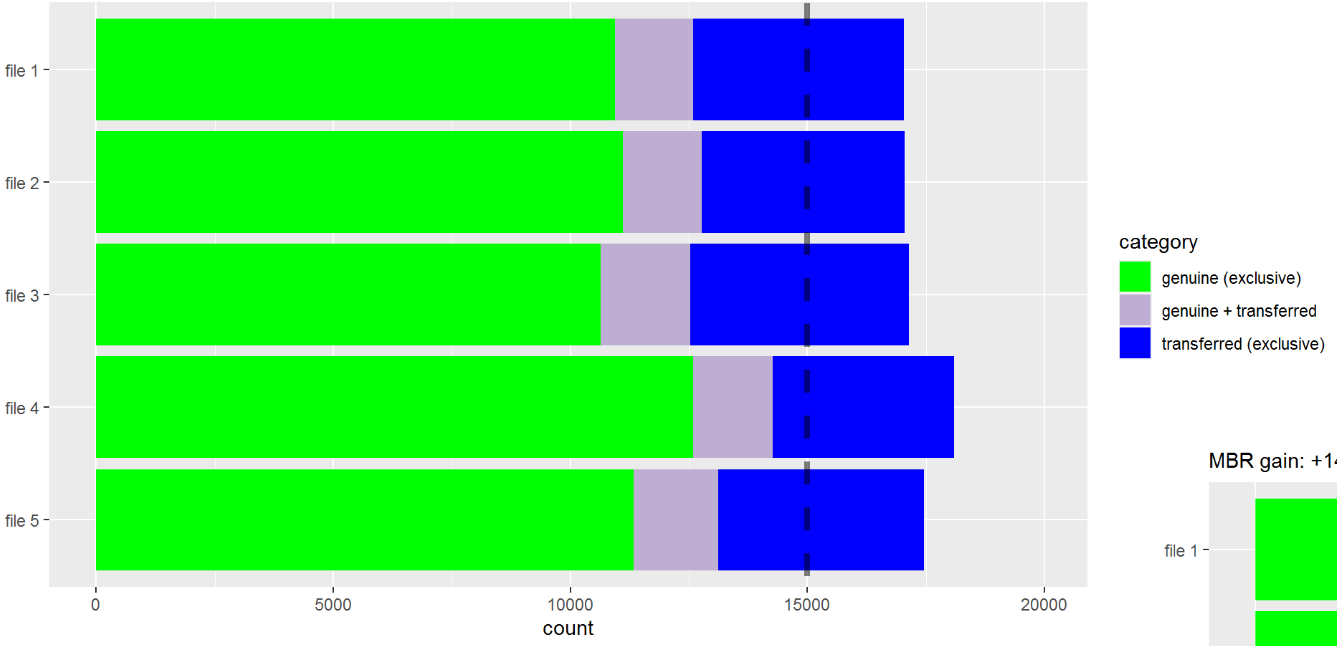


Were the retention times sufficiently close to allow base peak matching?

Were transferred peaks correct?

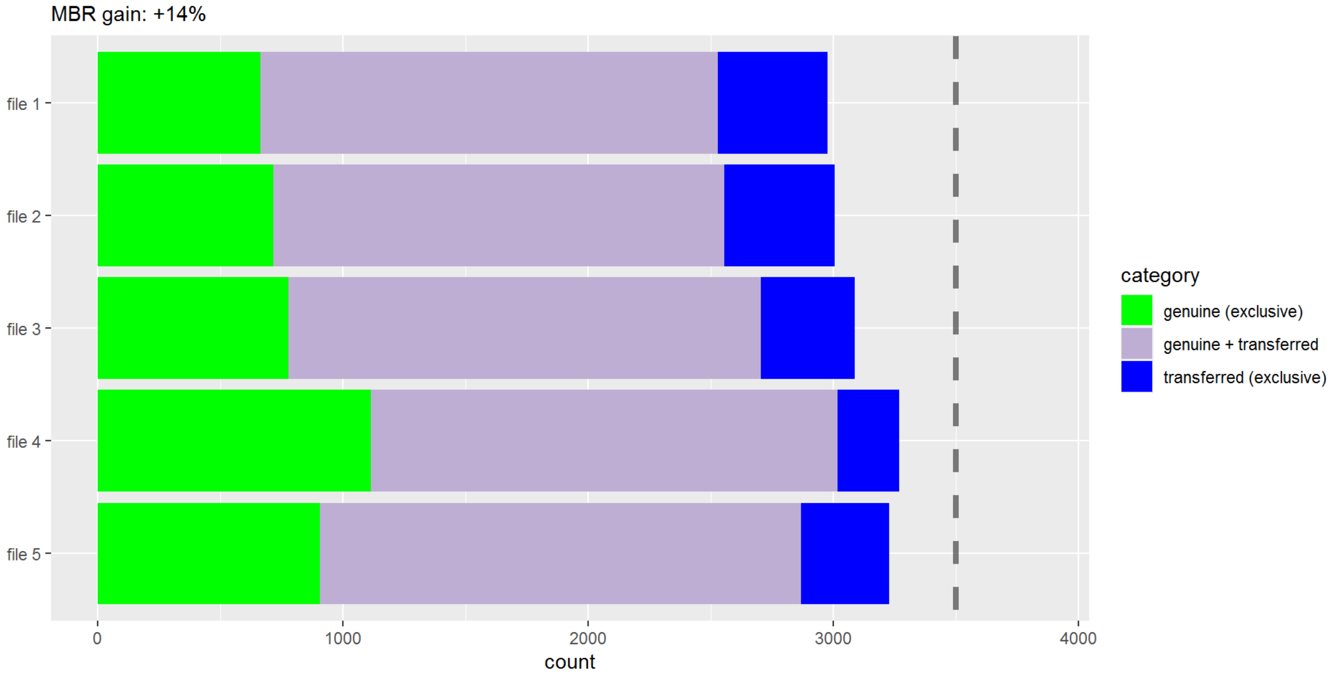


Peptide Identification

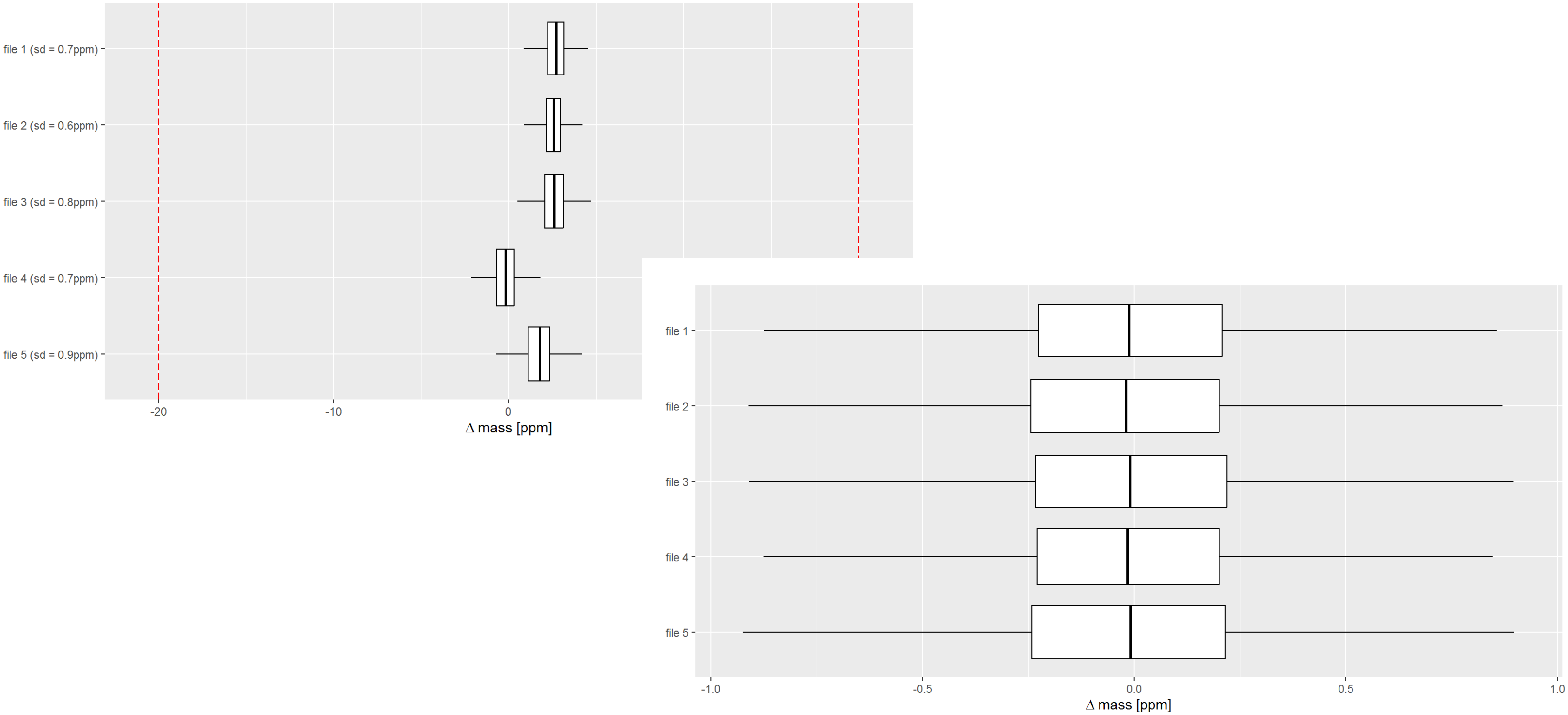


Peptide Level

Protein Level



Mass Accuracy



Exercise

Looking at QC Reports

Analysing Mass Spec Data with R

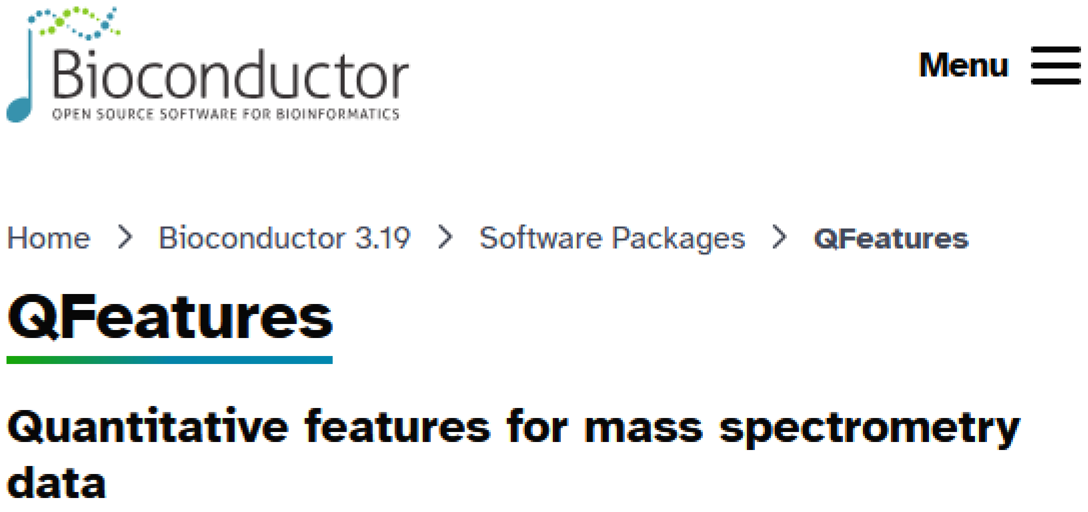
Babraham Bioinformatics

Bioconductor Package Environments

MSstats

Protein Significance Analysis in DDA, SRM and DIA for Label-free or Label-based Proteomics Experiments

- Streamlined workflow
  - Data Import
  - Data Aggregation and Normalisation
  - Differential abundance testing
- Little flexibility or control

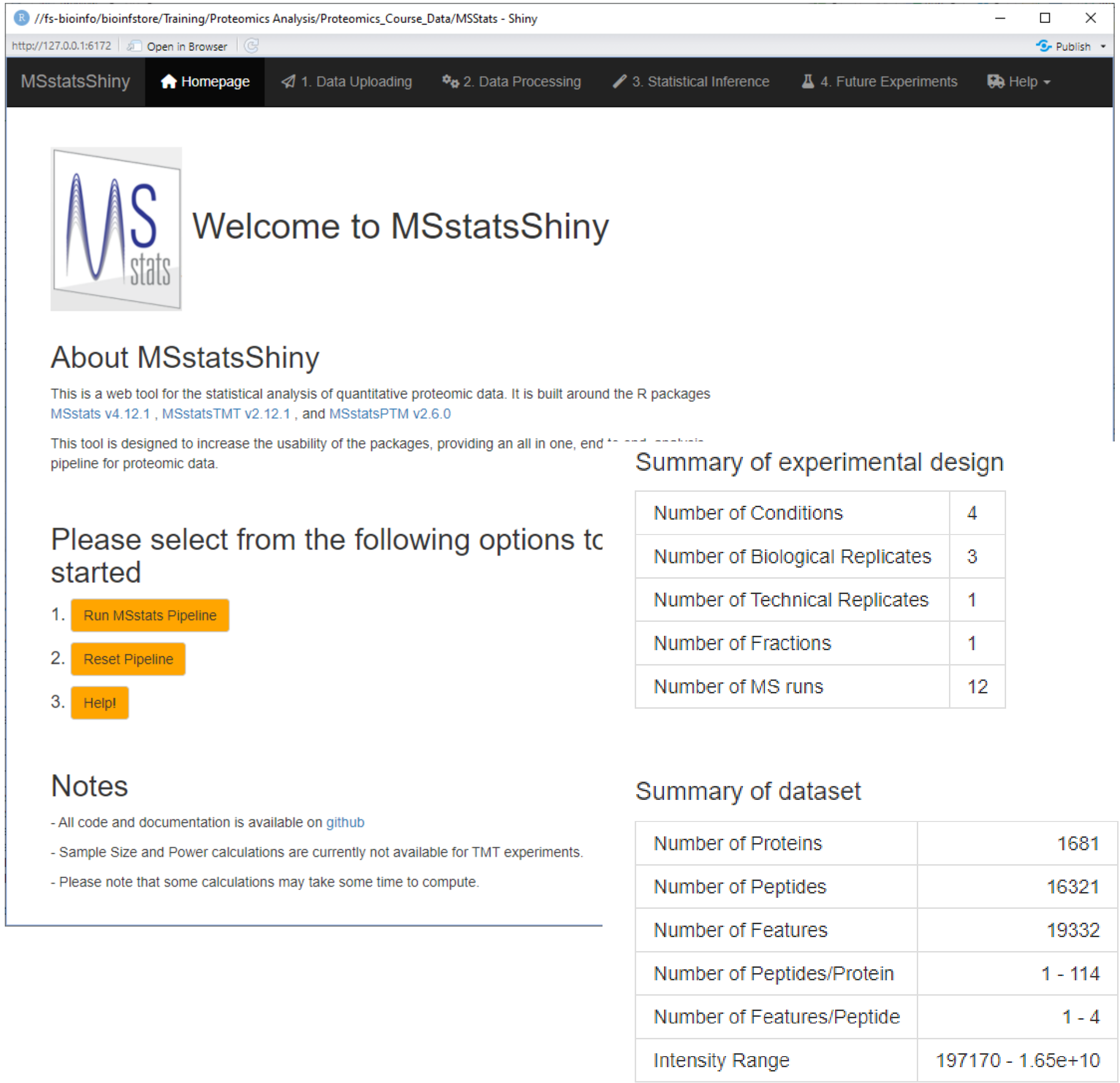


QFeatures

Quantitative features for mass spectrometry data

- Manual workflow
  - More user input in each step
  - More flexibility and options
  - Links externally for statistics

MSStats Shiny



Welcome to MSstatsShiny

About MSstatsShiny

This is a web tool for the statistical analysis of quantitative proteomic data. It is built around the R packages MSstats v4.12.1 , MSstatsTMT v2.12.1 , and MSstatsPTM v2.6.0

This tool is designed to increase the usability of the packages, providing an all in one, end
pipeline for proteomic data.

Please select from the following options tc started

1. Run MSstats Pipeline
2. Reset Pipeline
3. Help!

Notes

- All code and documentation is available on github

- Sample Size and Power calculations are currently not available for TMT experiments.

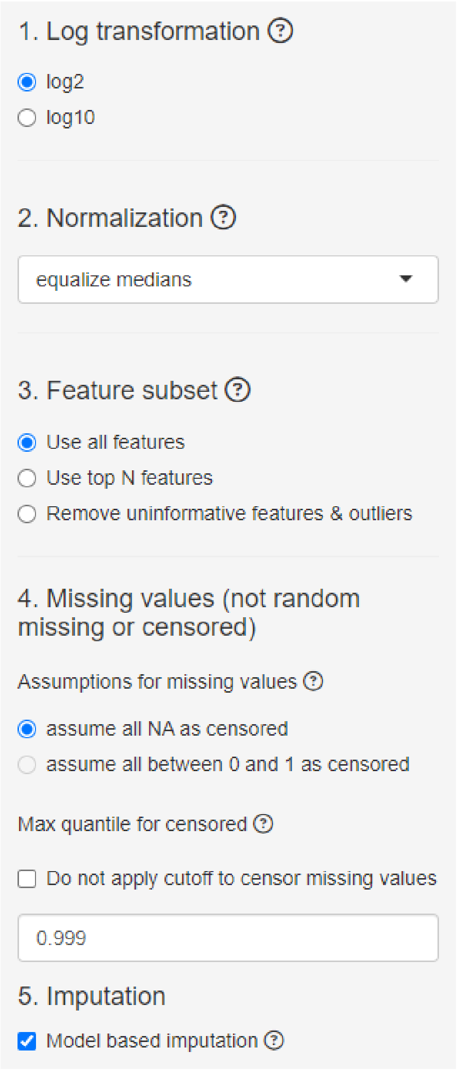
- Please note that some calculations may take some time to compute.

Summary of experimental design

Number of Conditions	4
Number of Biological Replicates	3
Number of Technical Replicates	1
Number of Fractions	1
Number of MS runs	12

Summary of dataset

Number of Proteins	1681
Number of Peptides	16321
Number of Features	19332
Number of Peptides/Protein	1 - 114
Number of Features/Peptide	1 - 4
Intensity Range	197170 - 1.65e+10



1. Log transformation ?

● log2
○ log10

2. Normalization ?

equalize medians ▾

3. Feature subset ?

● Use all features
○ Use top N features
○ Remove uninformative features & outliers

4. Missing values (not random missing or censored)

Assumptions for missing values ?

● assume all NA as censored
○ assume all between 0 and 1 as censored

Max quantile for censored ?

☐ Do not apply cutoff to censor missing values

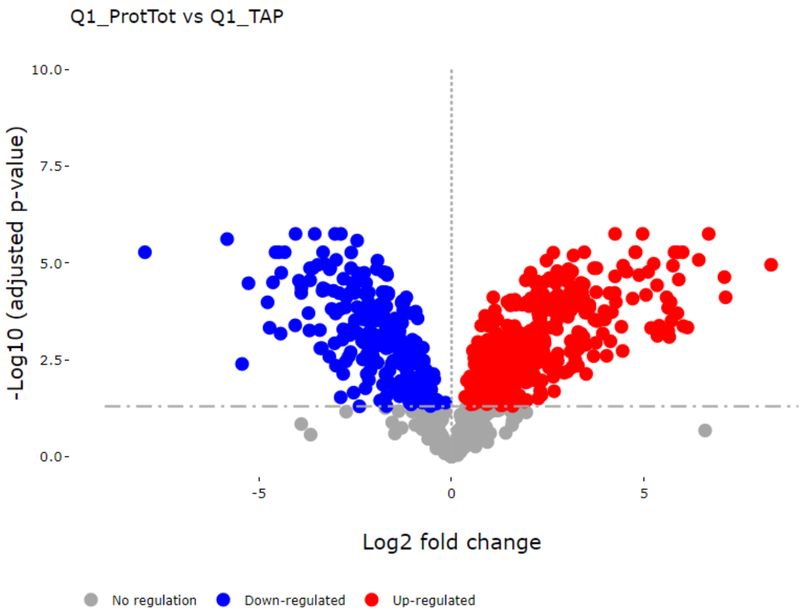
0.999

5. Imputation

☑ Model based imputation ?

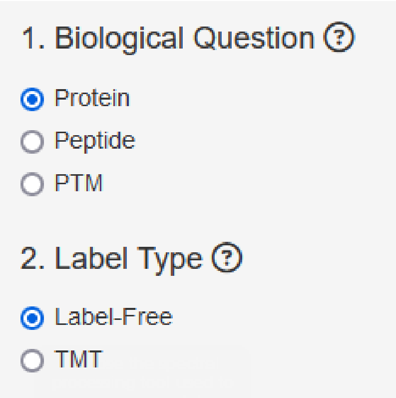
1. Define comparisons - contrast matrix ?

● All possible pairwise comparisons
○ Compare all against one
○ Create custom pairwise comparisons
○ Create custom non-pairwise comparisons



MSStats Shiny Workflow

- Define Experiment
  - Protein vs Peptide vs PTM
  - Mass Spec experiment type
- Load Data
  - Different imports from different programs

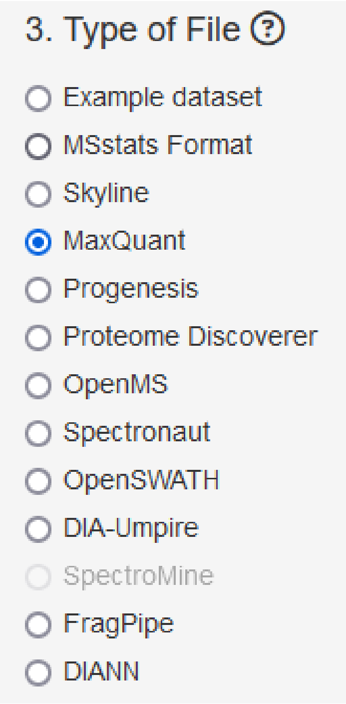


1. Biological Question

● Protein
○ Peptide
○ PTM

2. Label Type

● Label-Free
○ TMT



3. Type of File

○ Example dataset
○ MSstats Format
○ Skyline
● MaxQuant
○ Progenesis
○ Proteome Discoverer
○ OpenMS
○ Spectronaut
○ OpenSWATH
○ DIA-Umpire
○ SpectroMine
○ FragPipe
○ DIANN

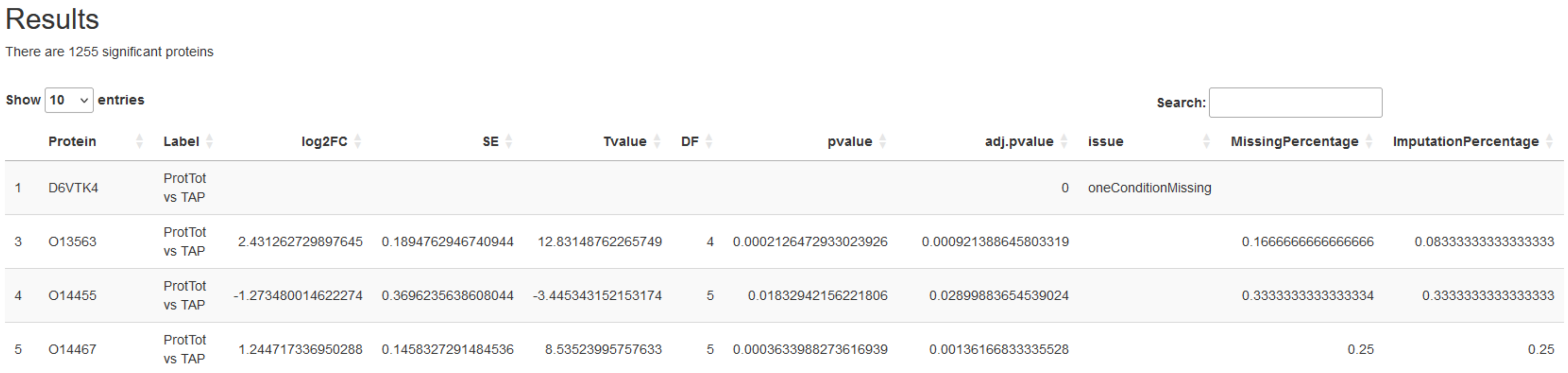
MSStats Shiny Workflow

- Protein Level Summarisation
  - Log Base
  - Normalisation
  - Filtering
  - Imputation
  - Summarisation

- Visualisation of individual proteins
  - Not very useful initially

MSStats Shiny Workflow

- Statistical analysis
  - Define comparison



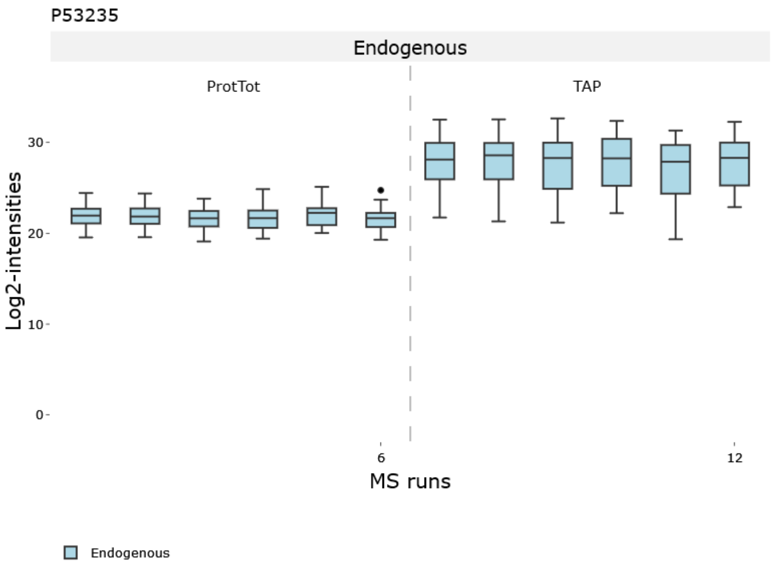
Results

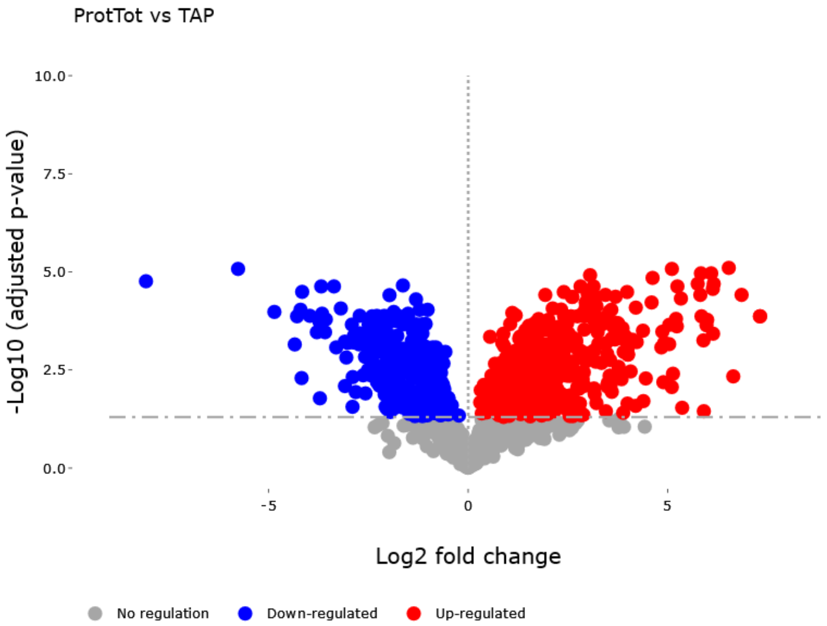
There are 1255 significant proteins

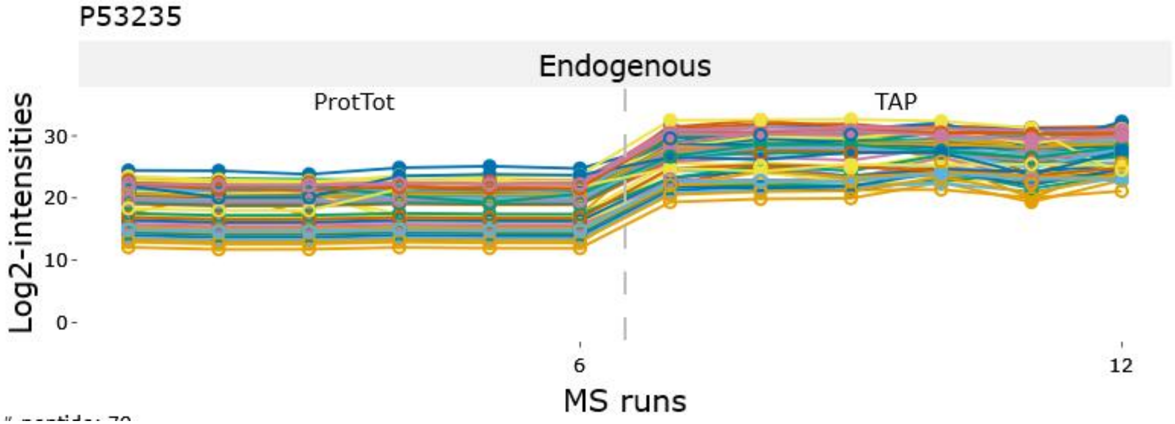
	Protein	Label	log2FC	SE	Tvalue	DF	pvalue	adj.pvalue	issue	MissingPercentage	ImputationPercentage
1	D6VTK4	ProtTot vs TAP						0	oneConditionMissing		
3	O13563	ProtTot vs TAP	2.431262729897645	0.1894762946740944	12.83148762265749	4	0.0002126472933023926	0.000921388645803319		0.1666666666666666	0.0833333333333333
4	O14455	ProtTot vs TAP	-1.273480014622274	0.3696235638608044	-3.445343152153174	5	0.01832942156221806	0.02899883654539024		0.3333333333333334	0.3333333333333333
5	O14467	ProtTot vs TAP	1.244717336950288	0.1458327291484536	8.53523995757633	5	0.0003633988273616939	0.00136166833335528		0.25	0.25

MSStats Shiny Workflow

- Visualisation
  - Volcano plot
  - Expression plots
- Data Export







Exercise

Running MSstats Shiny

MStats Manual

Load Search Data

Describe Experiment

Build Raw Data

Normalise

Statistical Comparison

Report

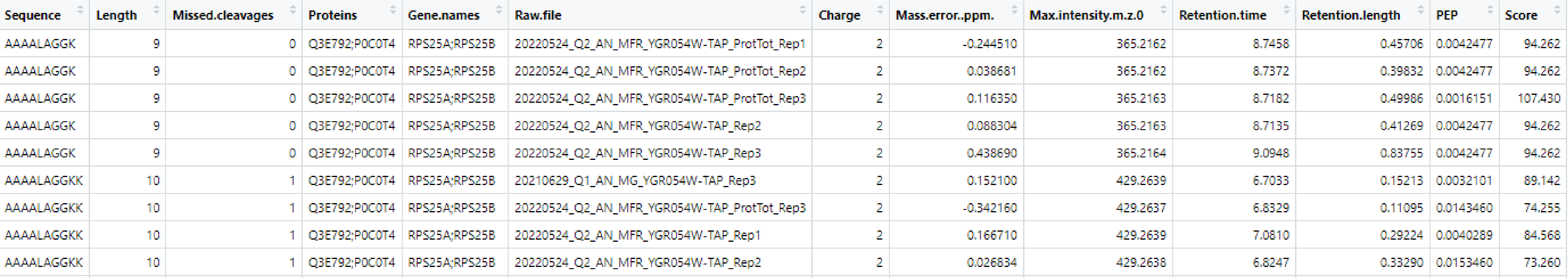
Loading Data

- MaxQuant
  - `evidence.txt`
  - `proteinGroups.txt`

- Spectronaut
  - `output_spectronaut.csv`

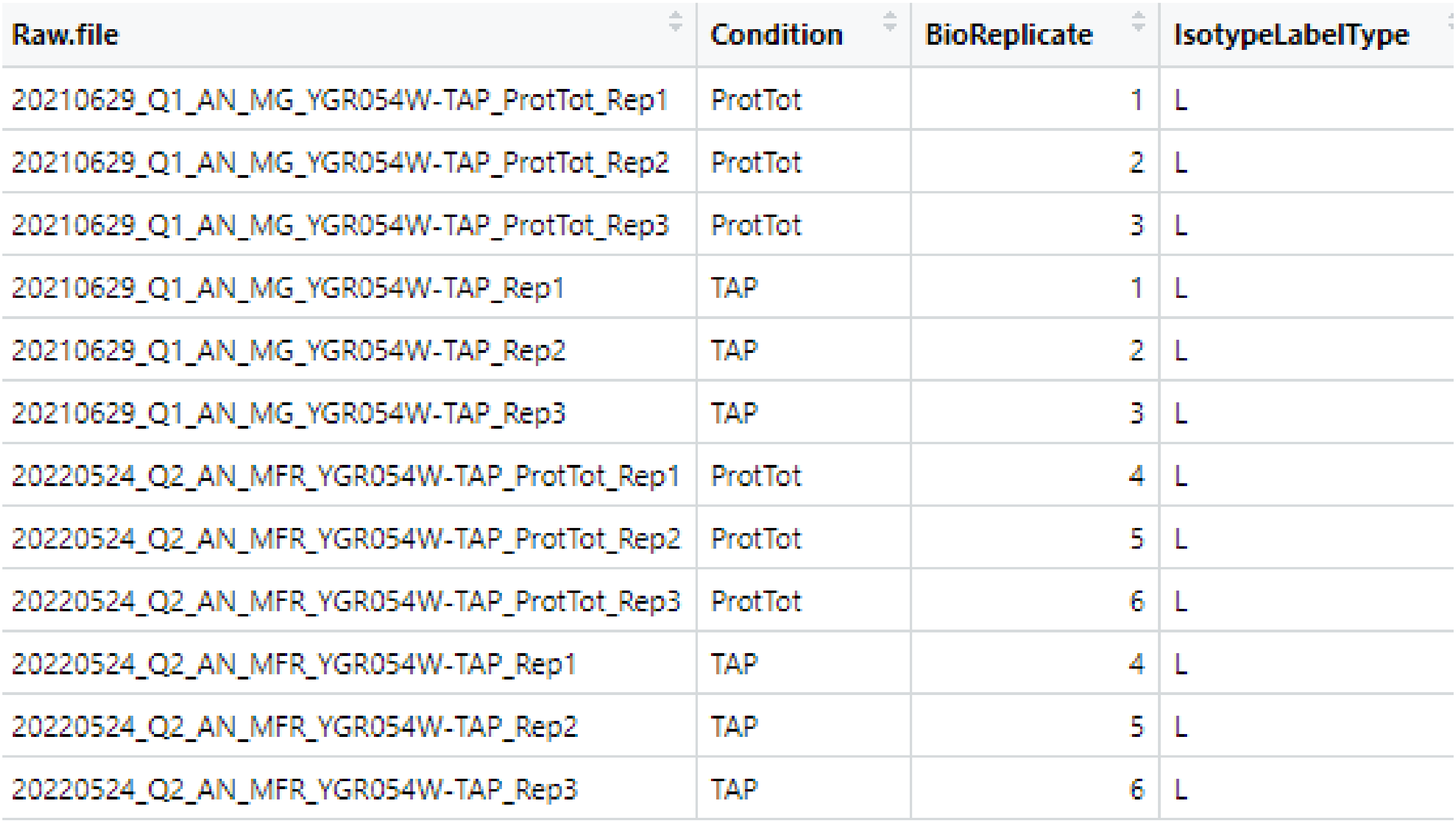
- ProteomeDiscoverer
  - `PSM result file`

Raw PSM Data



Sequence	Length	Missed.cleavages	Proteins	Gene.names	Raw.file	Charge	Mass.error..ppm.	Max.intensity.m.z.0	Retention.time	Retention.length	PEP	Score
AAAALAGGK	9	0	Q3E792;P0C0T4	RPS25A;RPS25B	20220524_Q2_AN_MFR_YGR054W-TAP_ProtTot_Rep1	2	-0.244510	365.2162	8.7458	0.45706	0.0042477	94.262
AAAALAGGK	9	0	Q3E792;P0C0T4	RPS25A;RPS25B	20220524_Q2_AN_MFR_YGR054W-TAP_ProtTot_Rep2	2	0.038681	365.2162	8.7372	0.39832	0.0042477	94.262
AAAALAGGK	9	0	Q3E792;P0C0T4	RPS25A;RPS25B	20220524_Q2_AN_MFR_YGR054W-TAP_ProtTot_Rep3	2	0.116350	365.2163	8.7182	0.49986	0.0016151	107.430
AAAALAGGK	9	0	Q3E792;P0C0T4	RPS25A;RPS25B	20220524_Q2_AN_MFR_YGR054W-TAP_Rep2	2	0.088304	365.2163	8.7135	0.41269	0.0042477	94.262
AAAALAGGK	9	0	Q3E792;P0C0T4	RPS25A;RPS25B	20220524_Q2_AN_MFR_YGR054W-TAP_Rep3	2	0.438690	365.2164	9.0948	0.83755	0.0042477	94.262
AAAALAGGKK	10	1	Q3E792;P0C0T4	RPS25A;RPS25B	20210629_Q1_AN_MG_YGR054W-TAP_Rep3	2	0.152100	429.2639	6.7033	0.15213	0.0032101	89.142
AAAALAGGKK	10	1	Q3E792;P0C0T4	RPS25A;RPS25B	20220524_Q2_AN_MFR_YGR054W-TAP_ProtTot_Rep3	2	-0.342160	429.2637	6.8329	0.11095	0.0143460	74.255
AAAALAGGKK	10	1	Q3E792;P0C0T4	RPS25A;RPS25B	20220524_Q2_AN_MFR_YGR054W-TAP_Rep1	2	0.166710	429.2639	7.0810	0.29224	0.0040289	84.568
AAAALAGGKK	10	1	Q3E792;P0C0T4	RPS25A;RPS25B	20220524_Q2_AN_MFR_YGR054W-TAP_Rep2	2	0.026834	429.2638	6.8247	0.33290	0.0153460	73.260

Building Annotation File

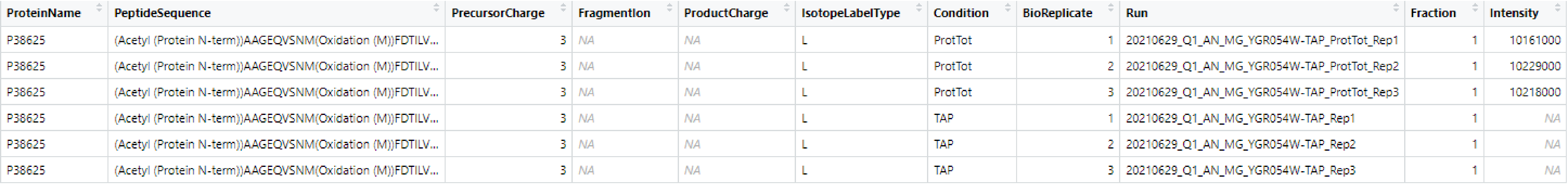


Raw.file	Condition	BioReplicate	IsotypeLabelType
20210629_Q1_AN_MG_YGR054W-TAP_ProtTot_Rep1	ProtTot	1	L
20210629_Q1_AN_MG_YGR054W-TAP_ProtTot_Rep2	ProtTot	2	L
20210629_Q1_AN_MG_YGR054W-TAP_ProtTot_Rep3	ProtTot	3	L
20210629_Q1_AN_MG_YGR054W-TAP_Rep1	TAP	1	L
20210629_Q1_AN_MG_YGR054W-TAP_Rep2	TAP	2	L
20210629_Q1_AN_MG_YGR054W-TAP_Rep3	TAP	3	L
20220524_Q2_AN_MFR_YGR054W-TAP_ProtTot_Rep1	ProtTot	4	L
20220524_Q2_AN_MFR_YGR054W-TAP_ProtTot_Rep2	ProtTot	5	L
20220524_Q2_AN_MFR_YGR054W-TAP_ProtTot_Rep3	ProtTot	6	L
20220524_Q2_AN_MFR_YGR054W-TAP_Rep1	TAP	4	L
20220524_Q2_AN_MFR_YGR054W-TAP_Rep2	TAP	5	L
20220524_Q2_AN_MFR_YGR054W-TAP_Rep3	TAP	6	L

Creating raw data object

```
MaxQtoMSstatsFormat(
  evidence = evidence,
  annotation = annotation,
  proteinGroups = protein_groups
) -> raw_data
```

- Removes contaminants
- Removes reverse (decoy) matches
- Removes proteins with 1 or 2 measures across all samples

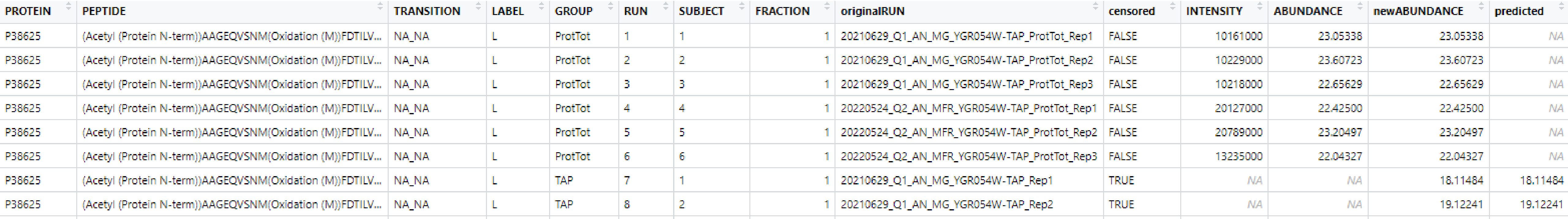


ProteinName	PeptideSequence	PrecursorCharge	FragmentIon	ProductCharge	IsotopeLabelType	Condition	BioReplicate	Run	Fraction	Intensity
P38625	(Acetyl (Protein N-term))AAGEQVSNM(Oxidation (M))FDTILV...	3	NA	NA	L	ProtTot	1	20210629_Q1_AN_MG_YGR054W-TAP_ProtTot_Rep1	1	10161000
P38625	(Acetyl (Protein N-term))AAGEQVSNM(Oxidation (M))FDTILV...	3	NA	NA	L	ProtTot	2	20210629_Q1_AN_MG_YGR054W-TAP_ProtTot_Rep2	1	10229000
P38625	(Acetyl (Protein N-term))AAGEQVSNM(Oxidation (M))FDTILV...	3	NA	NA	L	ProtTot	3	20210629_Q1_AN_MG_YGR054W-TAP_ProtTot_Rep3	1	10218000
P38625	(Acetyl (Protein N-term))AAGEQVSNM(Oxidation (M))FDTILV...	3	NA	NA	L	TAP	1	20210629_Q1_AN_MG_YGR054W-TAP_Rep1	1	NA
P38625	(Acetyl (Protein N-term))AAGEQVSNM(Oxidation (M))FDTILV...	3	NA	NA	L	TAP	2	20210629_Q1_AN_MG_YGR054W-TAP_Rep2	1	NA
P38625	(Acetyl (Protein N-term))AAGEQVSNM(Oxidation (M))FDTILV...	3	NA	NA	L	TAP	3	20210629_Q1_AN_MG_YGR054W-TAP_Rep3	1	NA

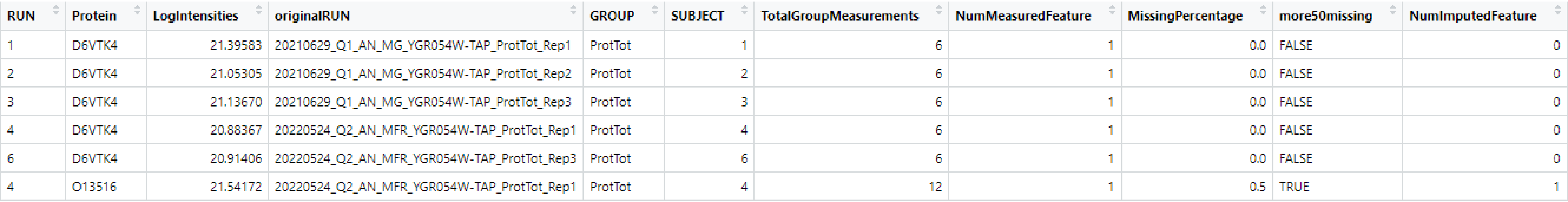
Quantitating

```
dataProcess(
  raw_data
) -> quantified_data
```

- Log transforms and Normalises
- Summarises Proteins
- Imputes missing values



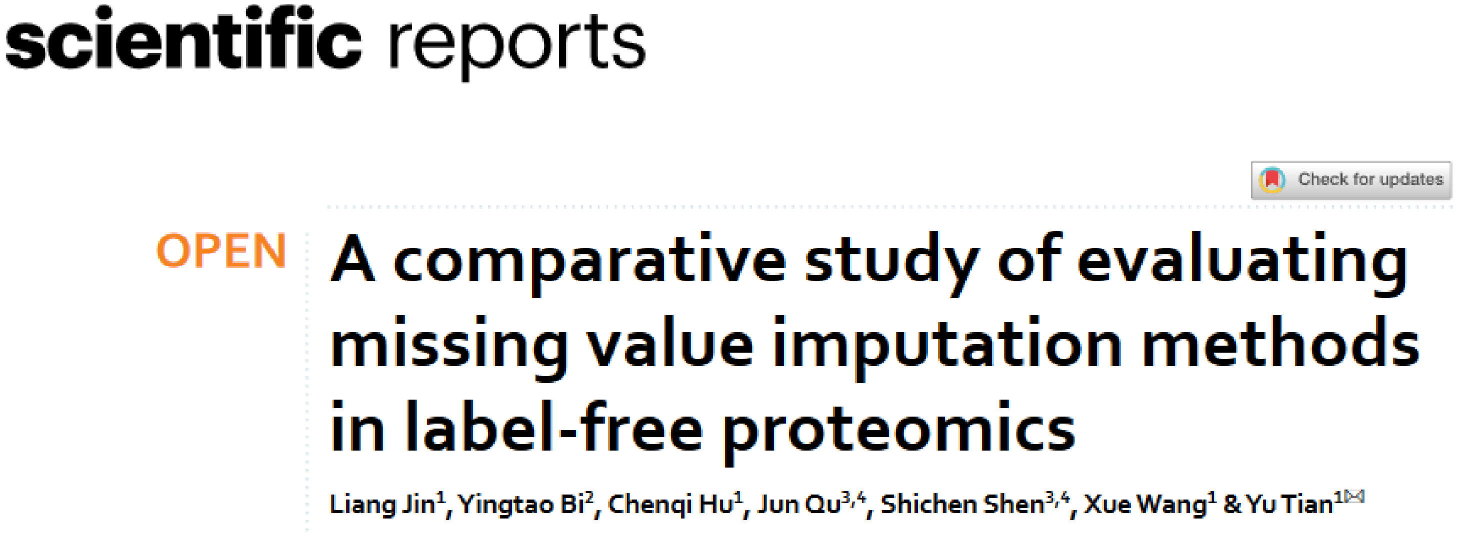
PROTEIN	PEPTIDE	TRANSITION	LABEL	GROUP	RUN	SUBJECT	FRACTION	originalRUN	censored	INTENSITY	ABUNDANCE	newABUNDANCE	predicted
P38625	(Acetyl (Protein N-term))AAGEQVSNM(Oxidation (M))FDTILV...	NA_NA	L	ProtTot	1	1	1	20210629_Q1_AN_MG_YGR054W-TAP_ProtTot_Rep1	FALSE	10161000	23.05338	23.05338	NA
P38625	(Acetyl (Protein N-term))AAGEQVSNM(Oxidation (M))FDTILV...	NA_NA	L	ProtTot	2	2	1	20210629_Q1_AN_MG_YGR054W-TAP_ProtTot_Rep2	FALSE	10229000	23.60723	23.60723	NA
P38625	(Acetyl (Protein N-term))AAGEQVSNM(Oxidation (M))FDTILV...	NA_NA	L	ProtTot	3	3	1	20210629_Q1_AN_MG_YGR054W-TAP_ProtTot_Rep3	FALSE	10218000	22.65629	22.65629	NA
P38625	(Acetyl (Protein N-term))AAGEQVSNM(Oxidation (M))FDTILV...	NA_NA	L	ProtTot	4	4	1	20220524_Q2_AN_MFR_YGR054W-TAP_ProtTot_Rep1	FALSE	20127000	22.42500	22.42500	NA
P38625	(Acetyl (Protein N-term))AAGEQVSNM(Oxidation (M))FDTILV...	NA_NA	L	ProtTot	5	5	1	20220524_Q2_AN_MFR_YGR054W-TAP_ProtTot_Rep2	FALSE	20789000	23.20497	23.20497	NA
P38625	(Acetyl (Protein N-term))AAGEQVSNM(Oxidation (M))FDTILV...	NA_NA	L	ProtTot	6	6	1	20220524_Q2_AN_MFR_YGR054W-TAP_ProtTot_Rep3	FALSE	13235000	22.04327	22.04327	NA
P38625	(Acetyl (Protein N-term))AAGEQVSNM(Oxidation (M))FDTILV...	NA_NA	L	TAP	7	1	1	20210629_Q1_AN_MG_YGR054W-TAP_Rep1	TRUE	NA	NA	18.11484	18.11484
P38625	(Acetyl (Protein N-term))AAGEQVSNM(Oxidation (M))FDTILV...	NA_NA	L	TAP	8	2	1	20210629_Q1_AN_MG_YGR054W-TAP_Rep2	TRUE	NA	NA	19.12241	19.12241



RUN	Protein	LogIntensities	originalRUN	GROUP	SUBJECT	TotalGroupMeasurements	NumMeasuredFeature	MissingPercentage	more50missing	NumImputedFeature
1	D6VTK4	21.39583	20210629_Q1_AN_MG_YGR054W-TAP_ProtTot_Rep1	ProtTot	1	6	1	0.0	FALSE	0
2	D6VTK4	21.05305	20210629_Q1_AN_MG_YGR054W-TAP_ProtTot_Rep2	ProtTot	2	6	1	0.0	FALSE	0
3	D6VTK4	21.13670	20210629_Q1_AN_MG_YGR054W-TAP_ProtTot_Rep3	ProtTot	3	6	1	0.0	FALSE	0
4	D6VTK4	20.88367	20220524_Q2_AN_MFR_YGR054W-TAP_ProtTot_Rep1	ProtTot	4	6	1	0.0	FALSE	0
6	D6VTK4	20.91406	20220524_Q2_AN_MFR_YGR054W-TAP_ProtTot_Rep3	ProtTot	6	6	1	0.0	FALSE	0
4	O13516	21.54172	20220524_Q2_AN_MFR_YGR054W-TAP_ProtTot_Rep1	ProtTot	4	12	1	0.5	TRUE	1

Imputation

- Can greatly expand the coverage of your data
- Restored values based on assumptions which may not be true
- Statistics doesn’t account for what is imputed



scientific reports

Check for updates

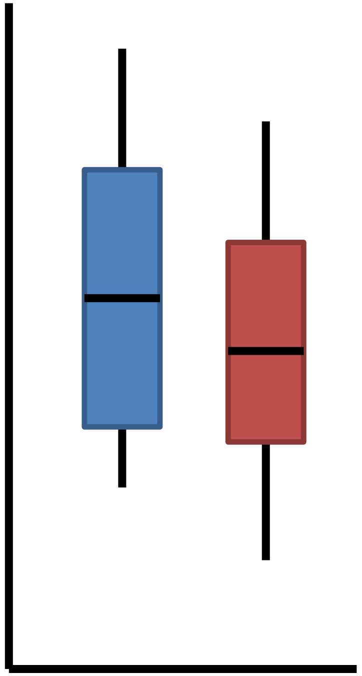
OPEN **A comparative study of evaluating missing value imputation methods in label-free proteomics**

Liang Jin[1], Yingtao Bi[2], Chenqi Hu[1], Jun Qu[3,4], Shichen Shen[3,4], Xue Wang[1] & Yu Tian[1]✉

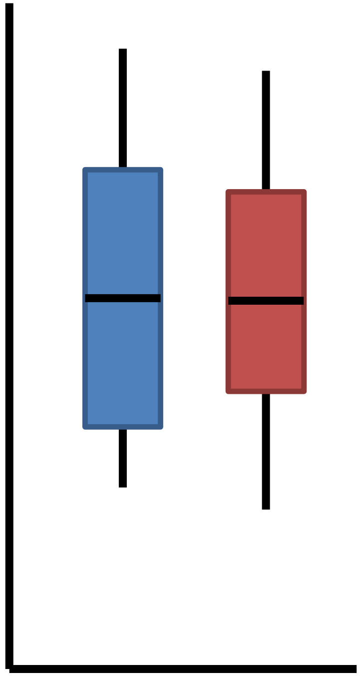
Many Methods

- Lowest observed value
- Random normal value
- Nearest neighbours
- Random forest

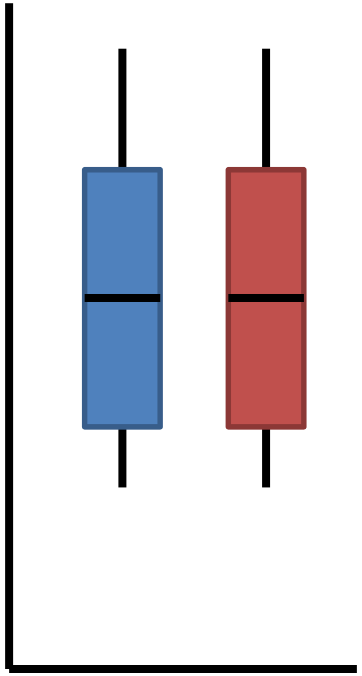
Normalisation



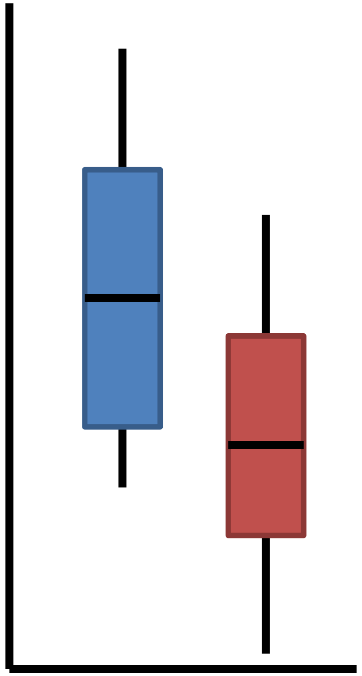
Original



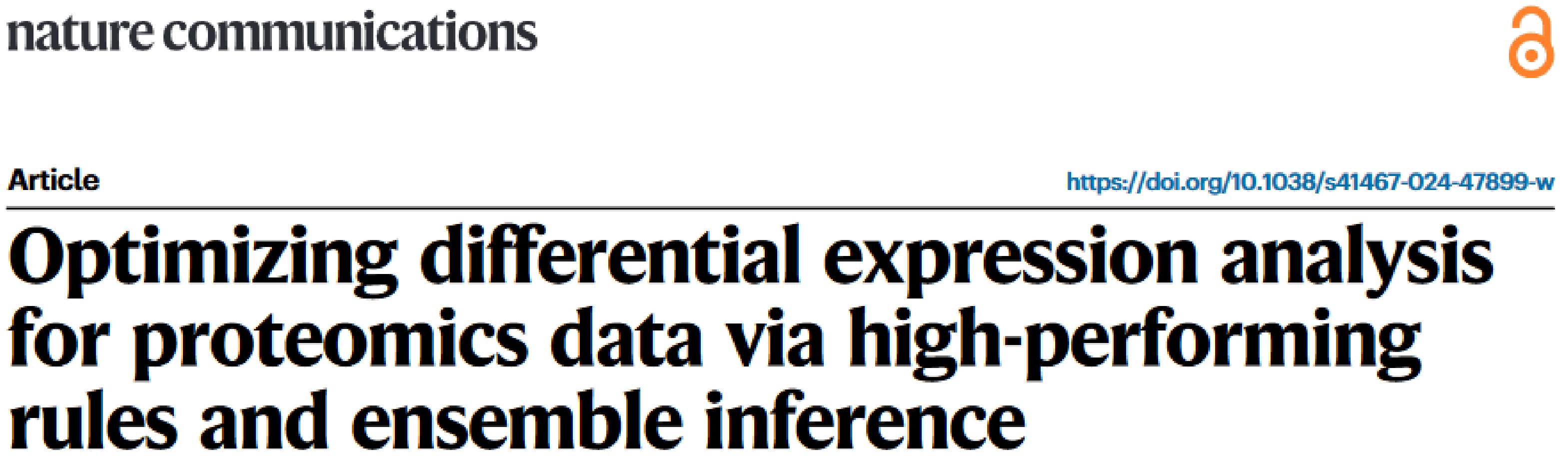
Equalise Medians



Quantile



Specific Genes

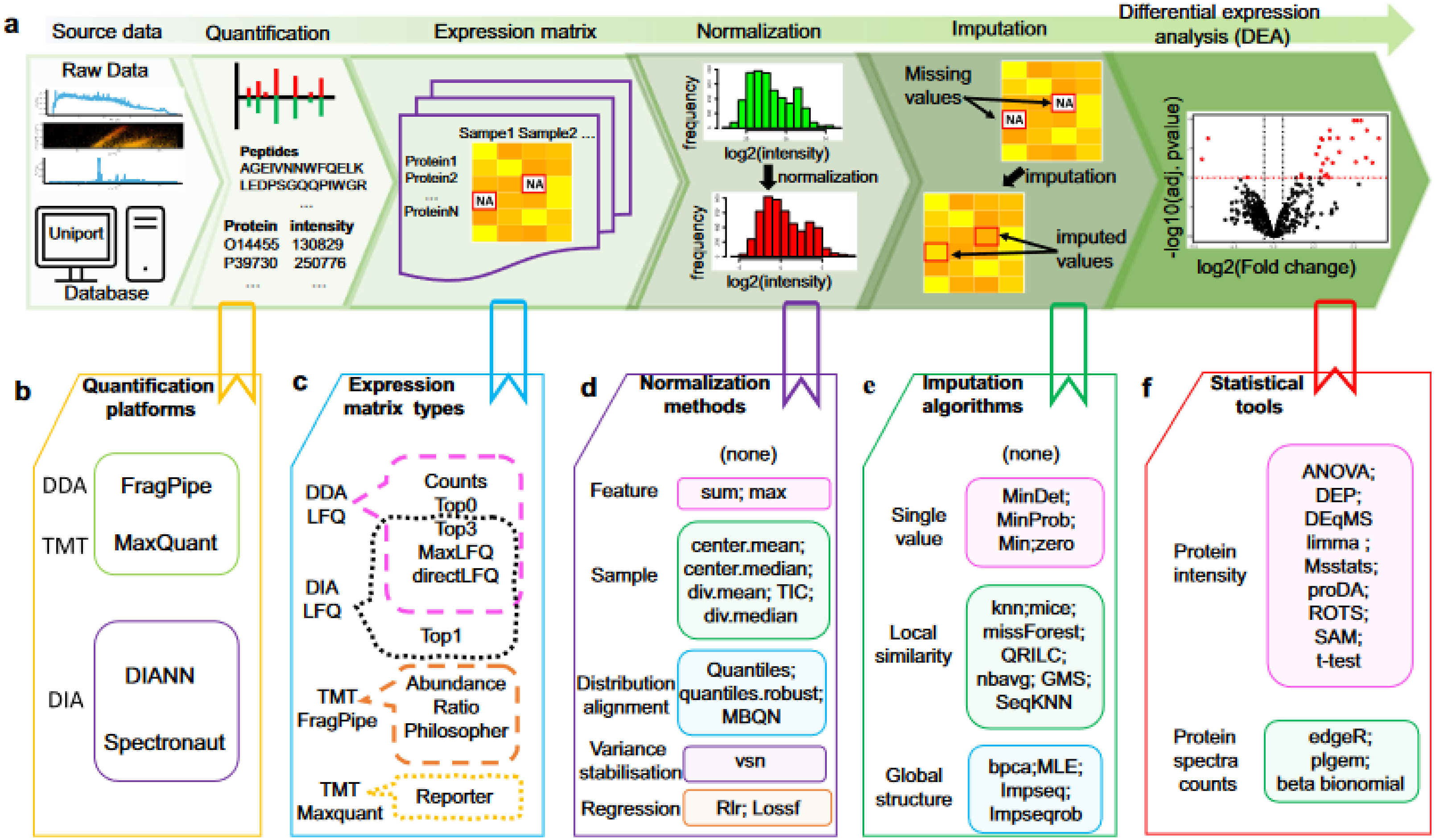


nature communications

Article https://doi.org/10.1038/s41467-024-47899-w

Optimizing differential expression analysis for proteomics data via high-performing rules and ensemble inference

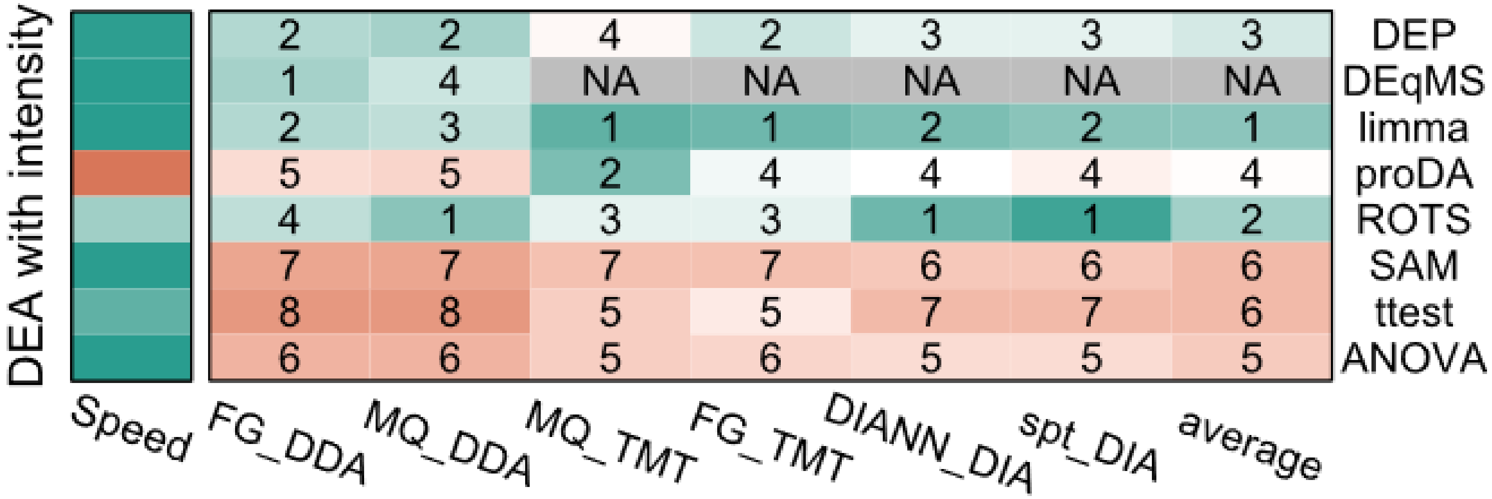
- Search Software
- Quantitation method
- Normalisation method
- Statistical test



Differential Abundance Statistics

limma **DEqMS** **ROTS** **proDA** **MSstats** **DEP**

- Quantitative analysis on normalised data
- T-test based (LIMMA)
- Mixture Models (MSstats, DEqMS)



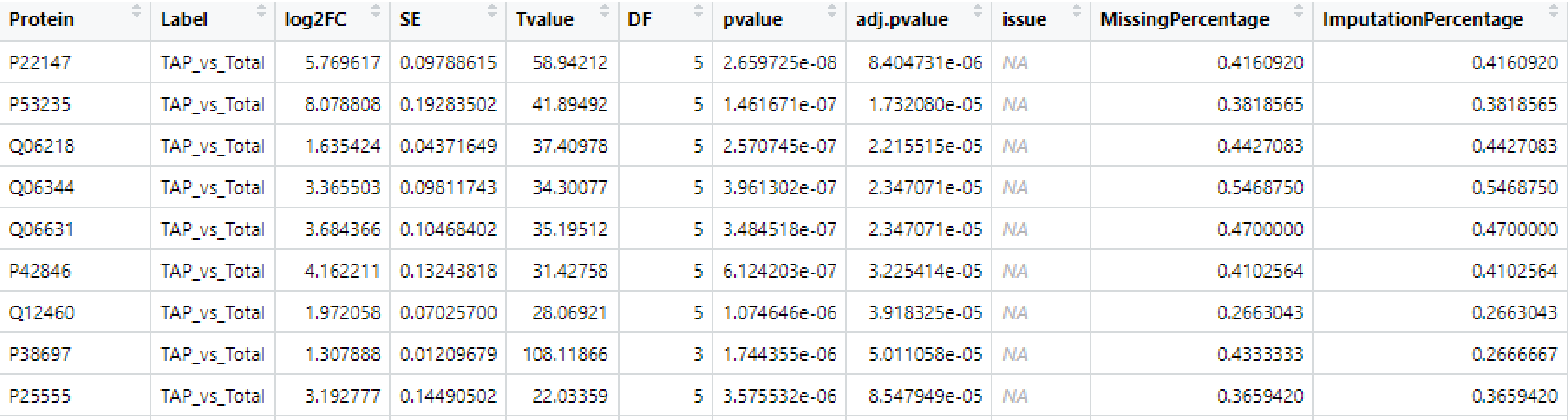
DEA with intensity

	Speed	FG_DDA	MQ_DDA	MQ_TMT	FG_TMT	DIANN_DIA	spt_DIA	average
DEP	🟩	2	2	4	2	3	3	3
DEqMS	🟩	1	4	NA	NA	NA	NA	NA
limma	🟩	2	3	1	1	2	2	1
proDA	🟥	5	5	2	4	4	4	4
ROTS	🟩	4	1	3	3	1	1	2
SAM	🟩	7	7	7	7	6	6	6
ttest	🟩	8	8	5	5	7	7	6
ANOVA	🟩	6	6	5	6	5	5	5

Running Differential Abundance

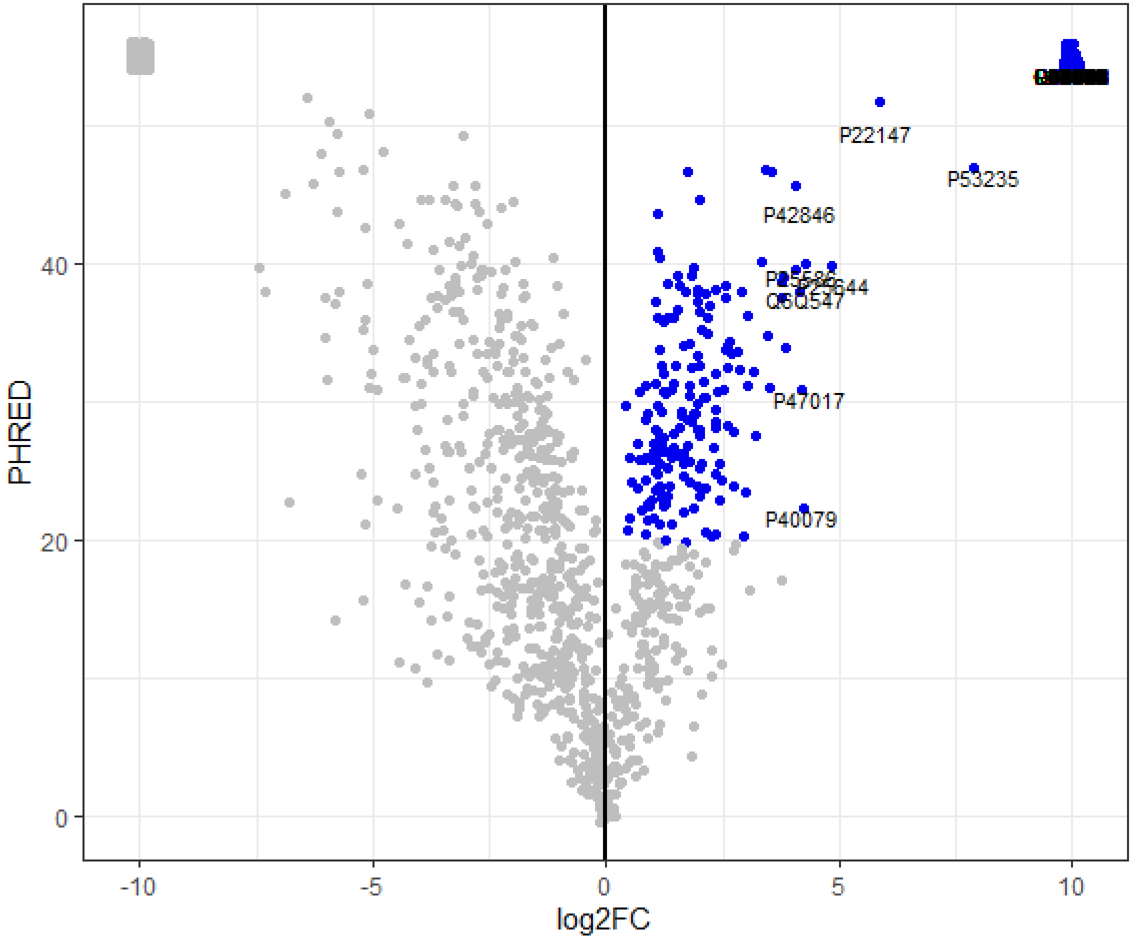
```
             ProtTot TAP
TAP_vs_Total      -1   1
```

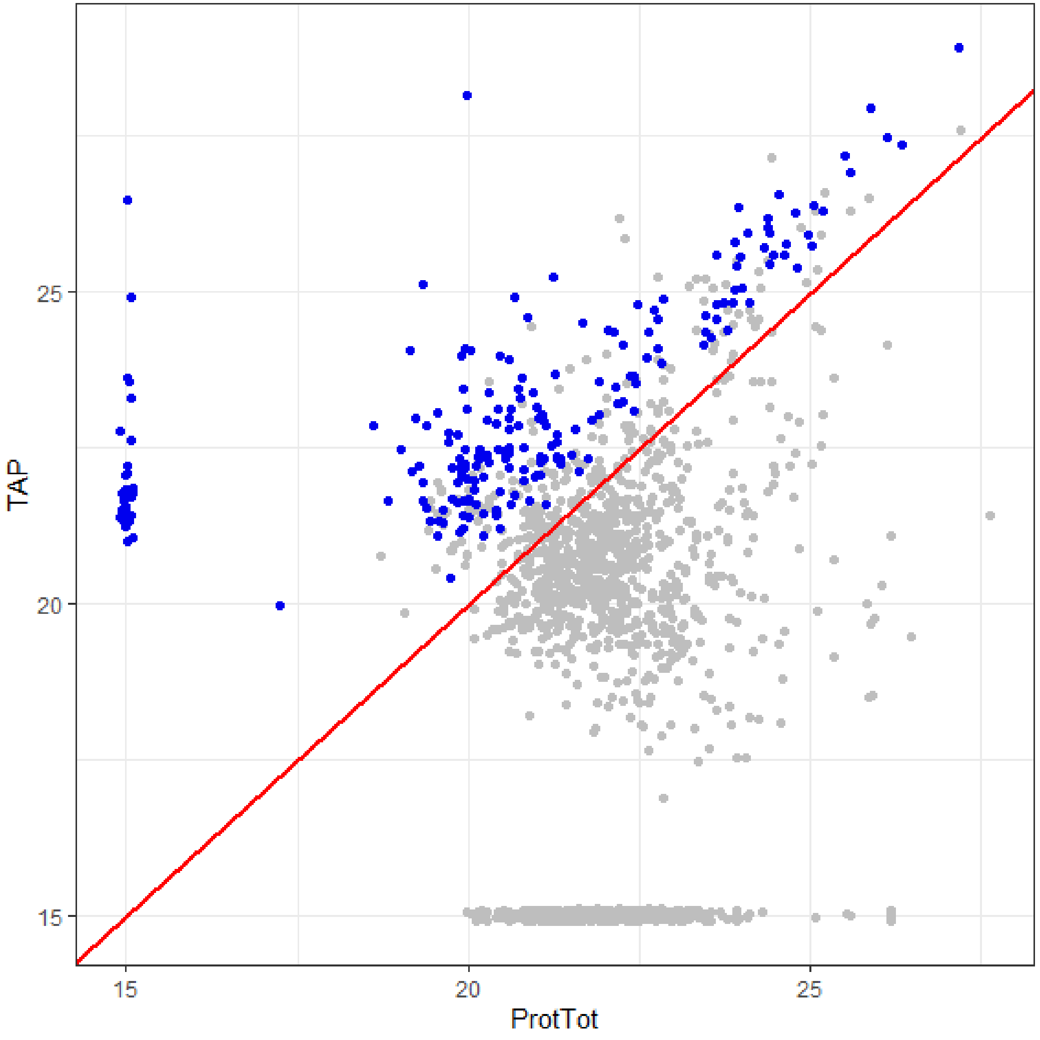
```
groupComparison(
  contrast.matrix = contrasts,
  data=quantified_data
) -> comparison_result
```



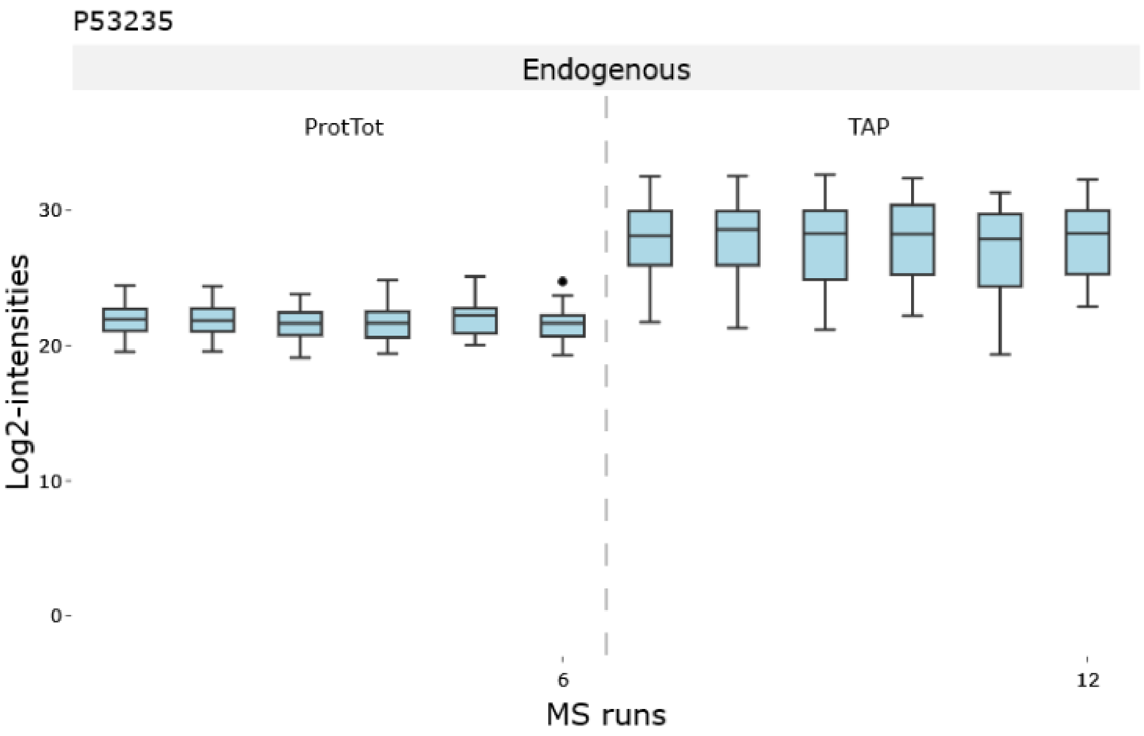
Protein	Label	log2FC	SE	Tvalue	DF	pvalue	adj.pvalue	issue	MissingPercentage	ImputationPercentage
P22147	TAP_vs_Total	5.769617	0.09788615	58.94212	5	2.659725e-08	8.404731e-06	NA	0.4160920	0.4160920
P53235	TAP_vs_Total	8.078808	0.19283502	41.89492	5	1.461671e-07	1.732080e-05	NA	0.3818565	0.3818565
Q06218	TAP_vs_Total	1.635424	0.04371649	37.40978	5	2.570745e-07	2.215515e-05	NA	0.4427083	0.4427083
Q06344	TAP_vs_Total	3.365503	0.09811743	34.30077	5	3.961302e-07	2.347071e-05	NA	0.5468750	0.5468750
Q06631	TAP_vs_Total	3.684366	0.10468402	35.19512	5	3.484518e-07	2.347071e-05	NA	0.4700000	0.4700000
P42846	TAP_vs_Total	4.162211	0.13243818	31.42758	5	6.124203e-07	3.225414e-05	NA	0.4102564	0.4102564
Q12460	TAP_vs_Total	1.972058	0.07025700	28.06921	5	1.074646e-06	3.918325e-05	NA	0.2663043	0.2663043
P38697	TAP_vs_Total	1.307888	0.01209679	108.11866	3	1.744355e-06	5.011058e-05	NA	0.4333333	0.2666667
P25555	TAP_vs_Total	3.192777	0.14490502	22.03359	5	3.575532e-06	8.547949e-05	NA	0.3659420	0.3659420

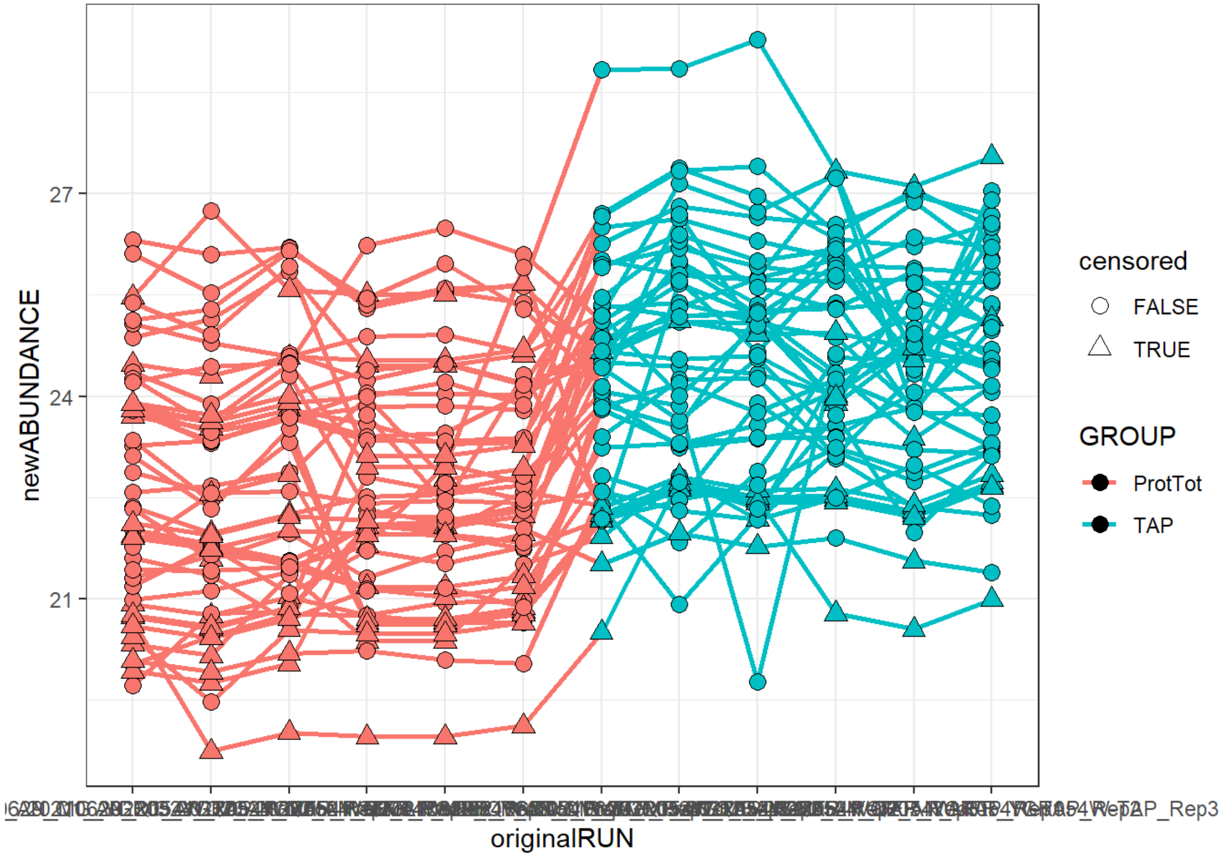
Plotting Hits





More Detailed Information





Exercise

Running MSstats in R