Proteomics Data Analysis

Simon Andrews

V2025-3

Babraham Bioinformatics

Course Content

- Principles of Mass Spectrometry
- Types of Quantitative MS
- Processing MS Data
  - Running searches
  - Evaluating Quality Control
- Analysing MS Data in R
  - General proteomics with R
    - Data import
    - Quality control
    - Quantitation and normalisation
    - Imputation
    - Visualisation and exploration
    - Differential abundance
  - Automated analysis with MSstats

Related Courses

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

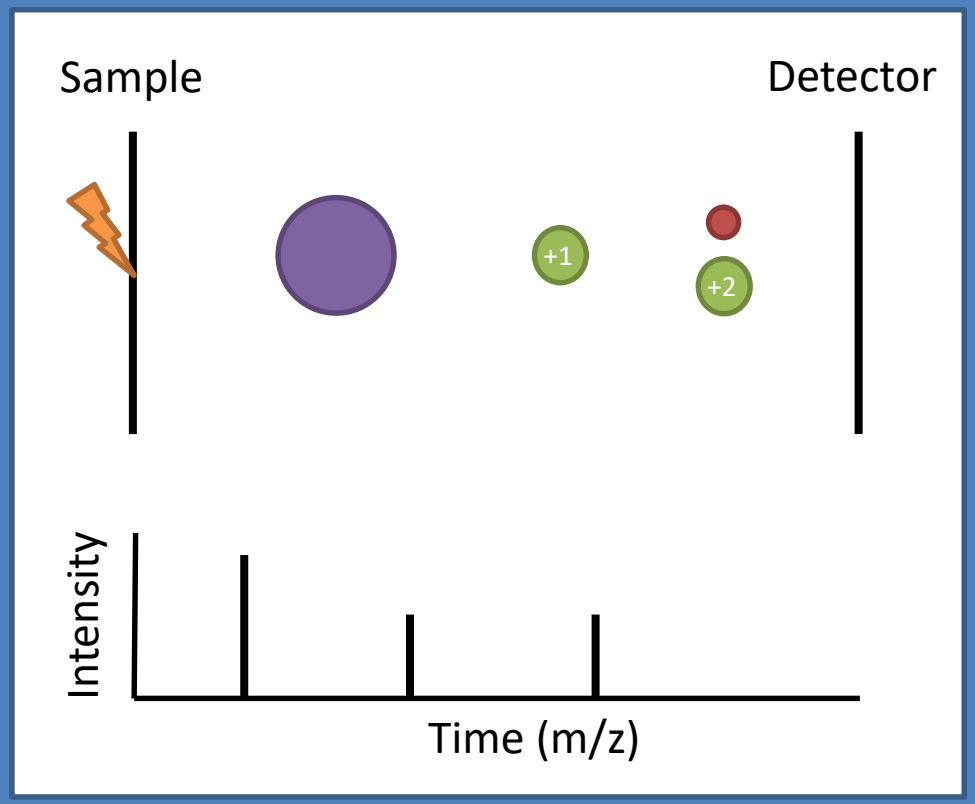
- Interpreting Gene Lists

Principles of Proteomics
Mass Spec

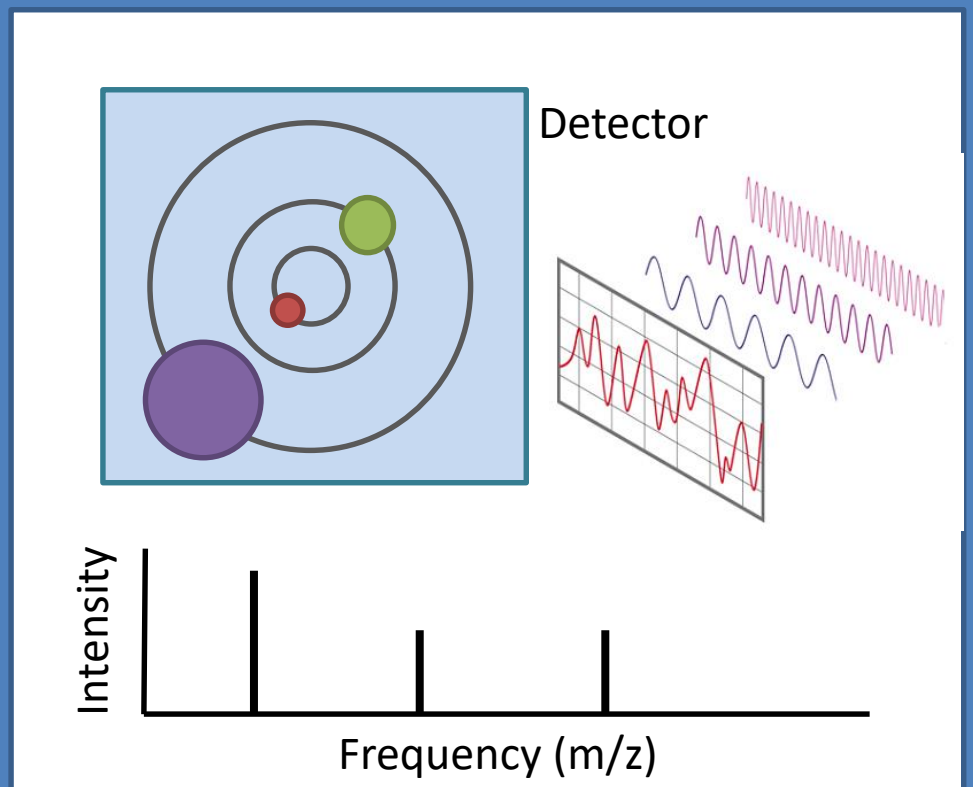
Babraham Bioinformatics

How mass spectrometers work

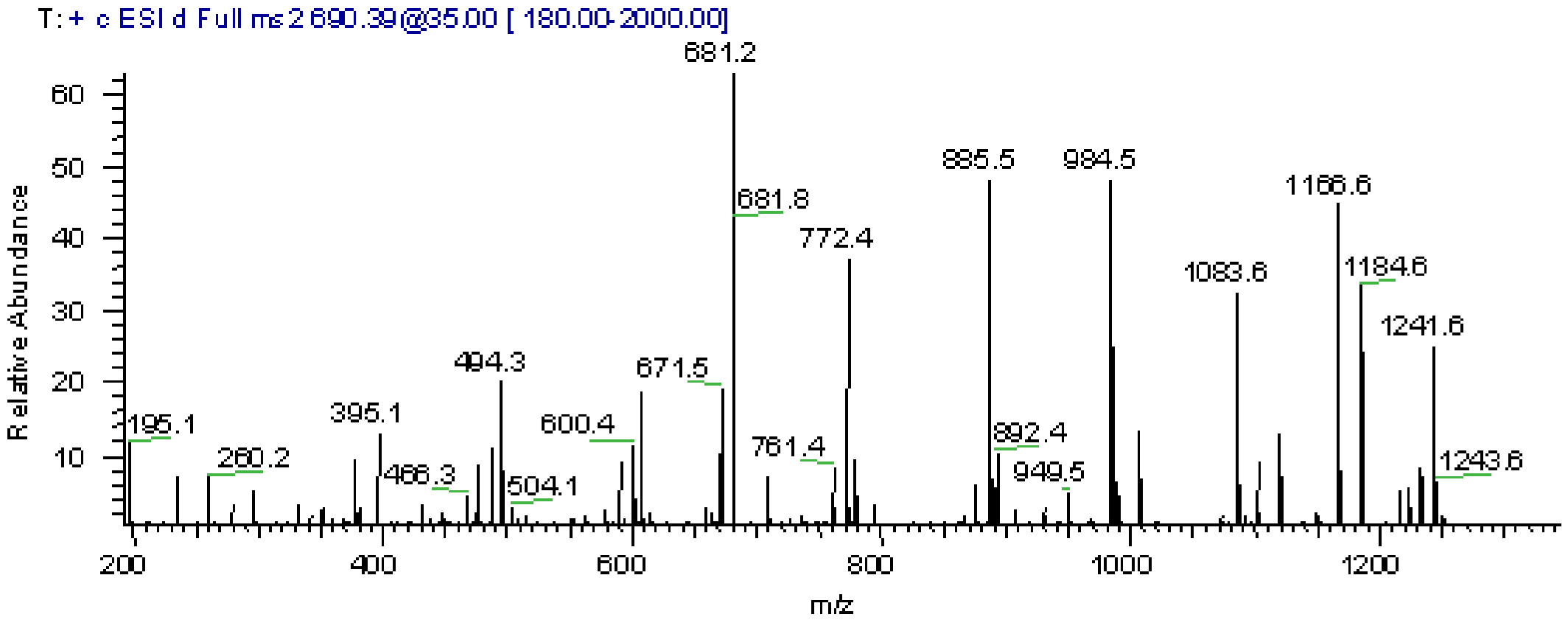
Time of Flight (TOF)



Fourier Transform Ion Cyclotron Resonance (FT-ICR)

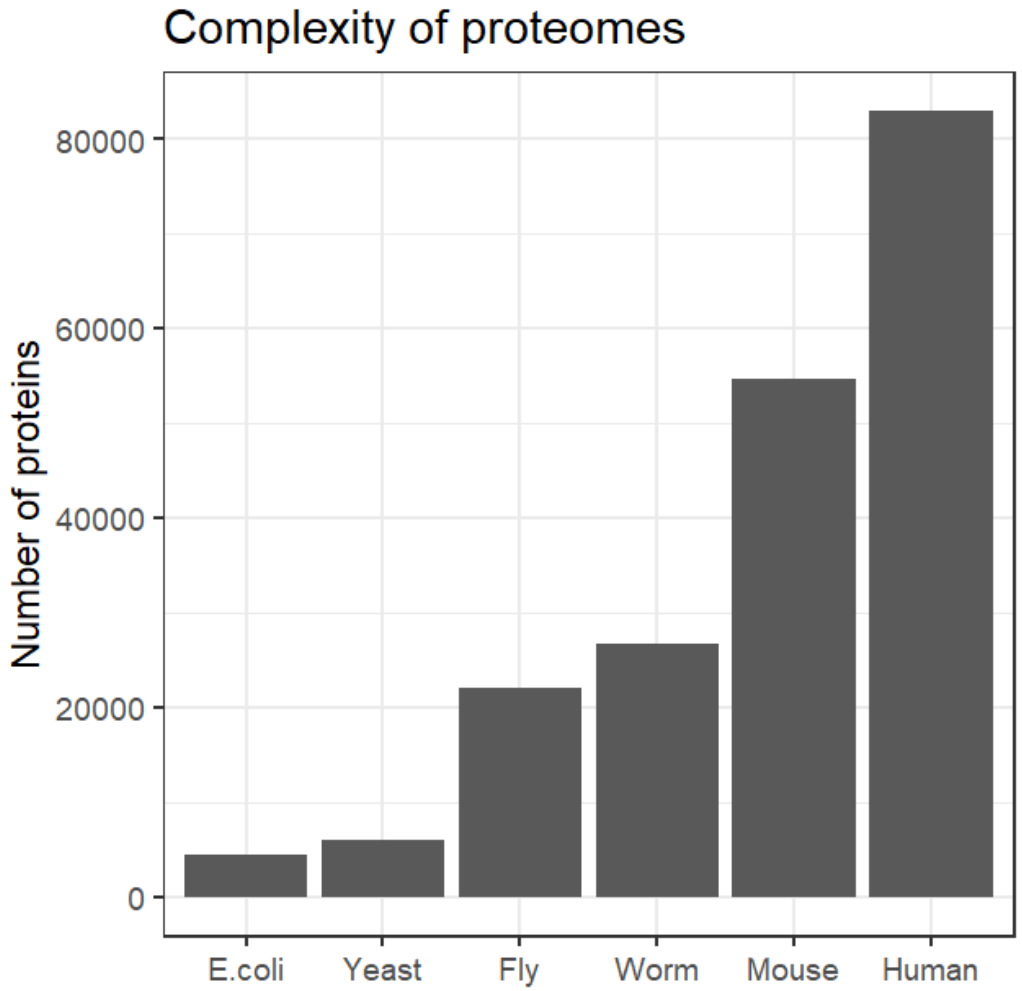


A typical mass spectrum



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

Measuring whole proteomes



- Whole proteins are so complex they are difficult to identify when processed whole
- Proteome samples are typically 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

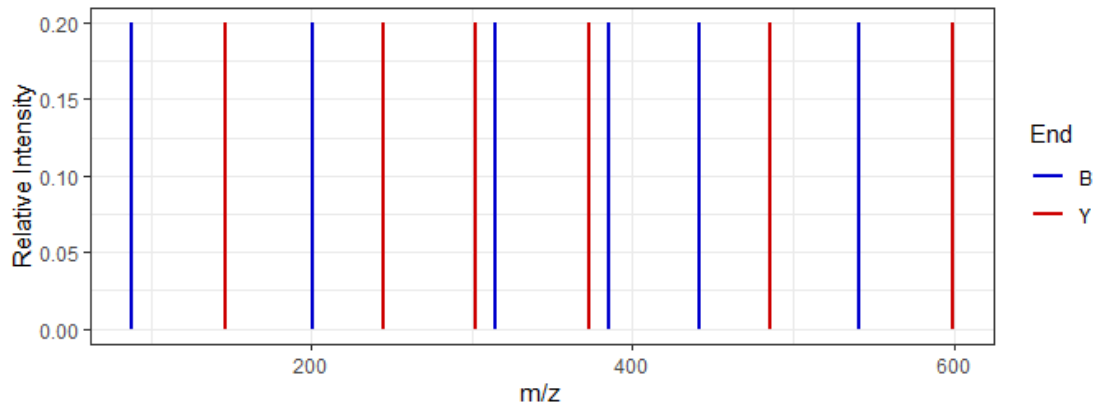
Just knowing a peptide's mass isn't enough to identify it

Tandem Mass Spectrometry

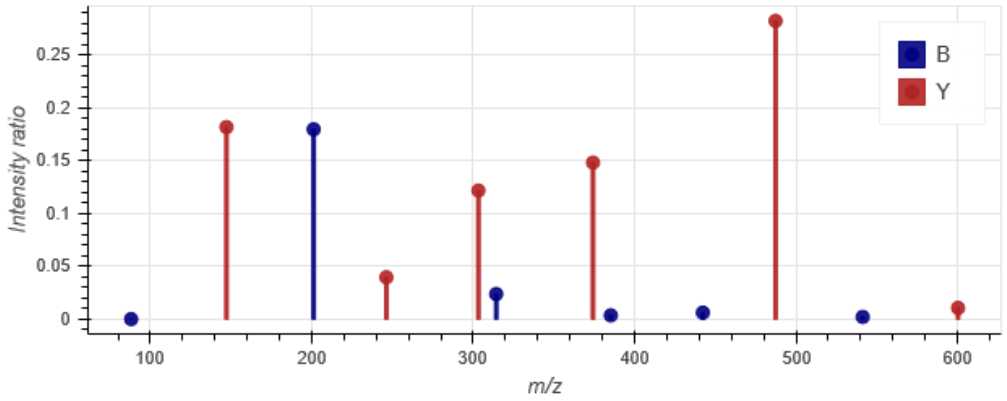
686Da	**SILAGVK**	
541Da	SILAGV K	147Da
442Da	SILAG VK	246Da
385Da	SILA GVK	303Da
314Da	SIL AGVK	374Da
201Da	SI LAGVK	487Da
88Da	S ILAGVK	600Da

Peptide MS2 Spectra

Theoretical



Observed

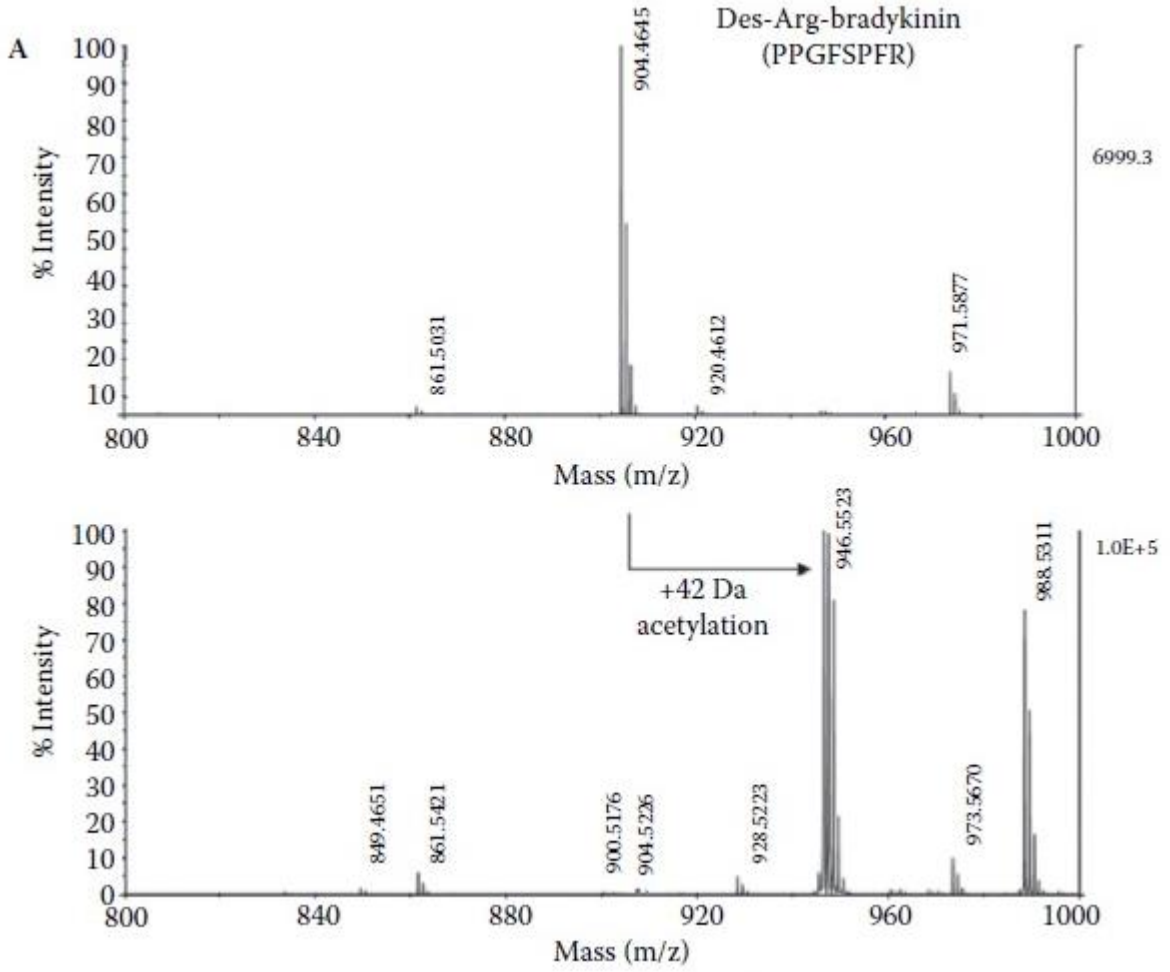


Searches are not performed by inferring sequence from spectra, but by scoring matches to predicted spectra

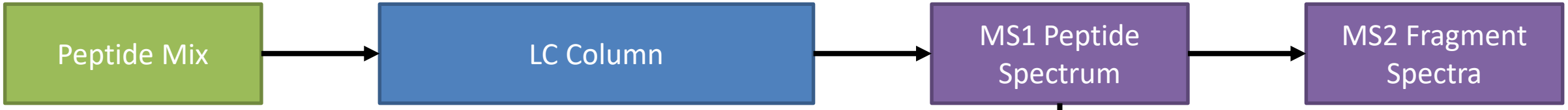
https://iomics.ugent.be/ms2pip/

Measuring Modifications

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



Problems with bottom up proteomics

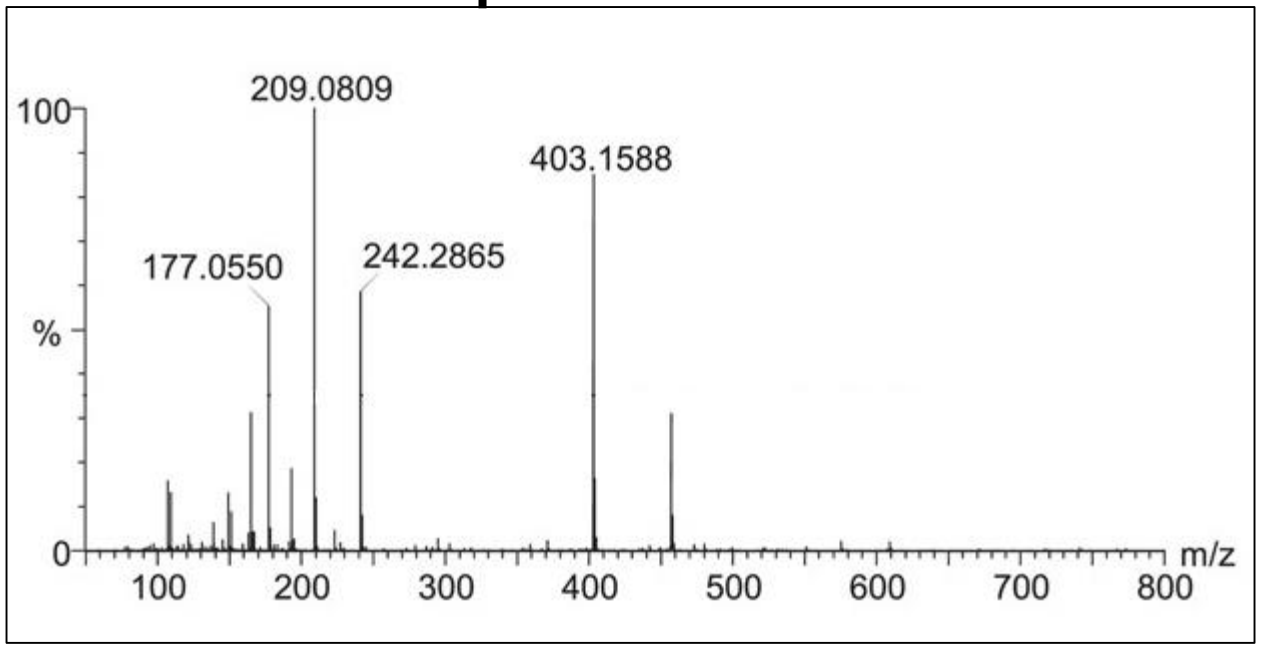


- Too many peaks for MS2
  - More LC Separation

    (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

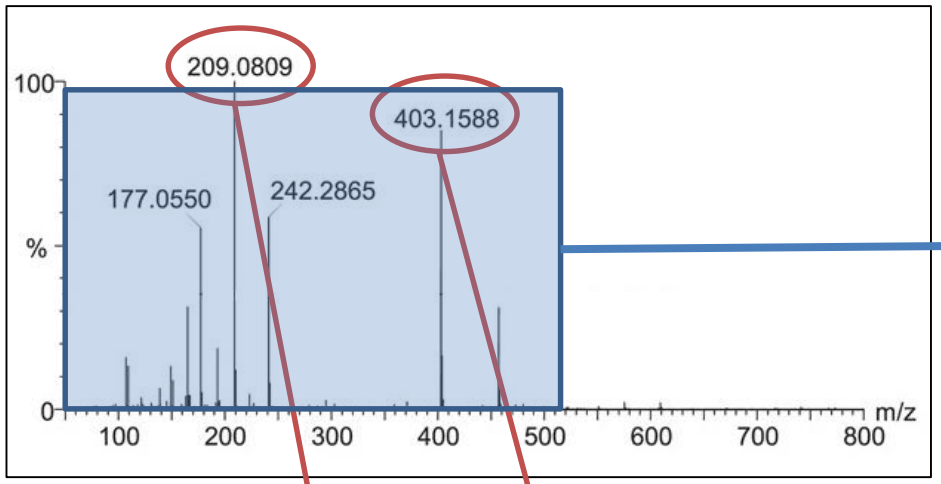
Data Independent Acquisition (DIA)

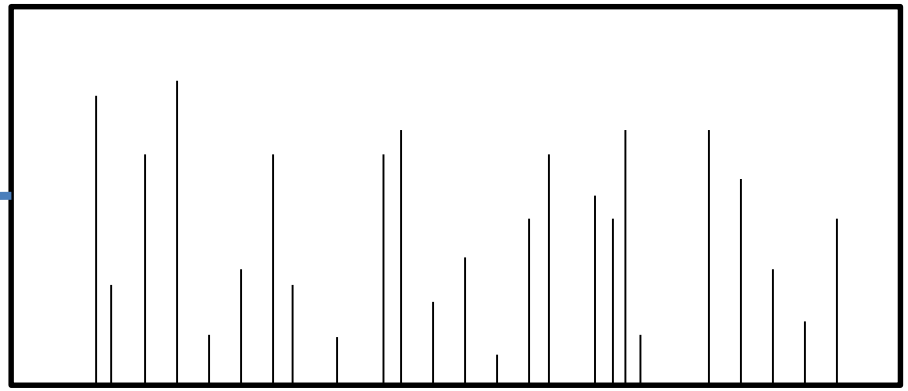
- Pick all peaks from MS1 (MZ range)
- Pass them simultaneously to MS2

- Mixed MS2 spectra
- More difficult spectrum matching

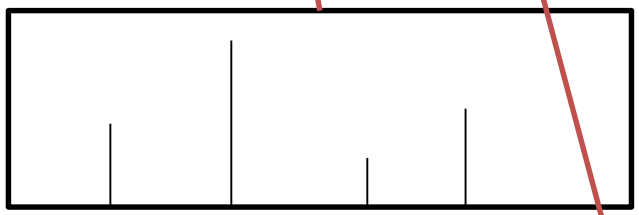
- Higher coverage
- More complete coverage

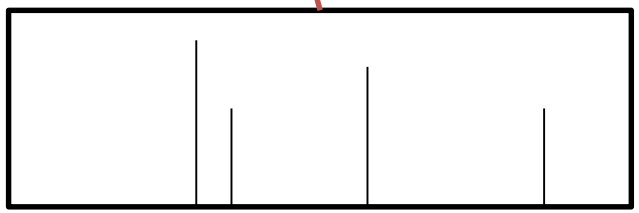
DIA vs DDA





DIA





DDA

Identifying Proteins from spectra

Babraham Bioinformatics

Database Searching

Spectronaut®
powered by Pulsar

PROTEOME DISCOVERER

DIA-NN

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

https://www.uniprot.org/proteomes

UP000005640

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

BUSCO

Single Duplicated Fragmented Missing [i]

n:13780 · primates_odb10

C:99.5% (S:37.8% D:61.7%) F:0% M:0.5%

cRAP protein sequences

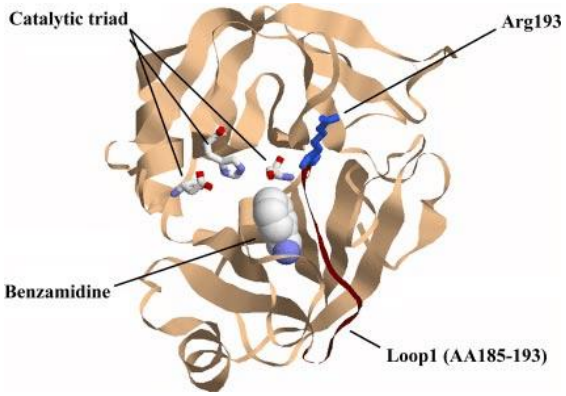
The **c**ommon **R**epository of **A**dventitious **P**roteins

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

MQ

N
N

DIA-NN

Take all proteins from your species of interest

UniProt

Generate Peptide Spectral Library

Shuffle Peptide Sequences

Generate Peptide Spectral Library

Search for peptide spectrum matches (PSMs)

Protein Libraries

REAL

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

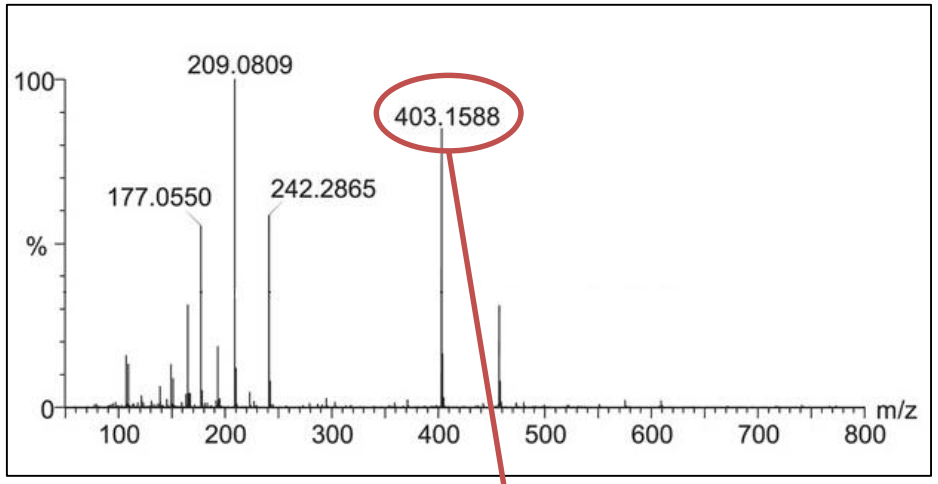
DECOY

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

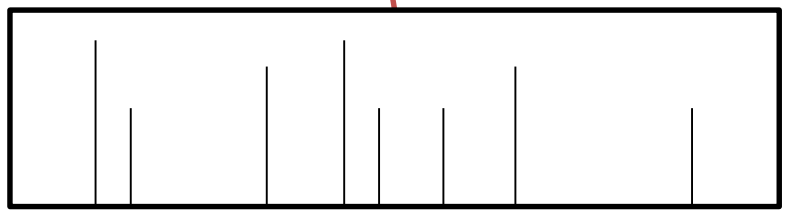
Decoy libraries can be reversed or shuffled

Peptide Spectrum Matches

MS1 Base Peak



MS2 Fragments



Find peptides with masses close to the parent peak

>P05067

MLPGLALLLAAWTARALEVPTDGNAGLLAEPQIAMFCGRLNMHMNV
QNGKWDSDP**SGTKTCIDTKEGIL**QYCQEVYPELQITNVVEANQPVTI
QNWCKRGRKQCKTHPHFVIPYRCLVGEFVSDALLVPDKCKFLHQERM
DVCETHLHW

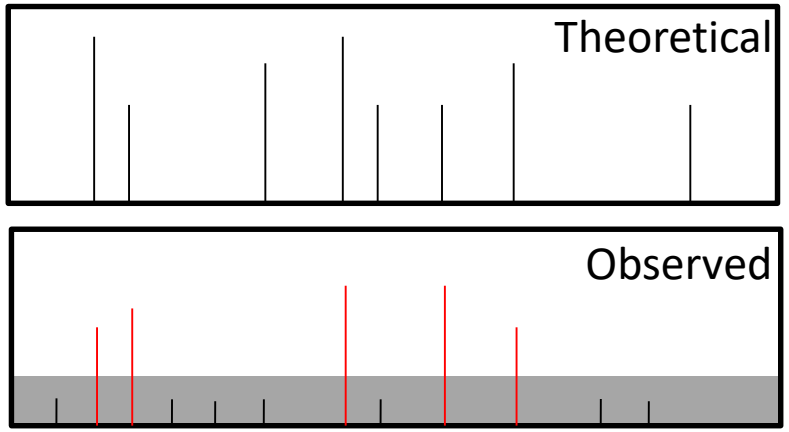
>P90210

MAVCGSAMSQSLLKTTQEPLARDPVKLPTTAASTPDAVDKYLETPGD
ENEHAHFQKAKERLEAKHRERMSQVMREWEEAERQAKNLPKADKKAV
IQHFQEKVESLEQEAANERQQLVET**HMARVEAMLNDRR**RLALENYIT
ALQAVPPRPRHVFNMLKKYVRAEQKDRQHTLKHFEH

Hundreds of candidates

Scoring a PSM match

Count Overlaps (Andromeda - MQ)



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

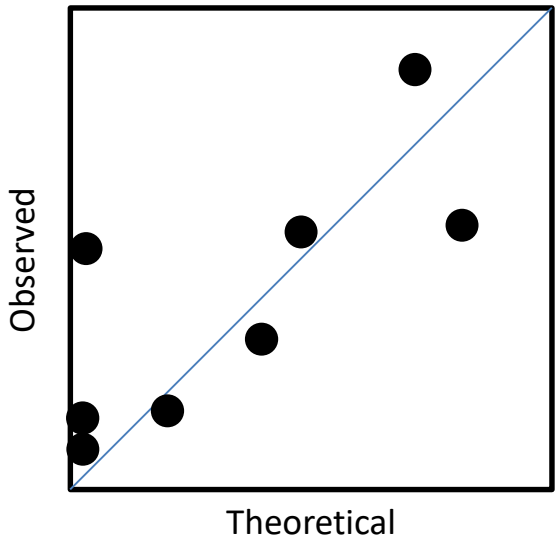
Optimize inclusion of losses

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

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

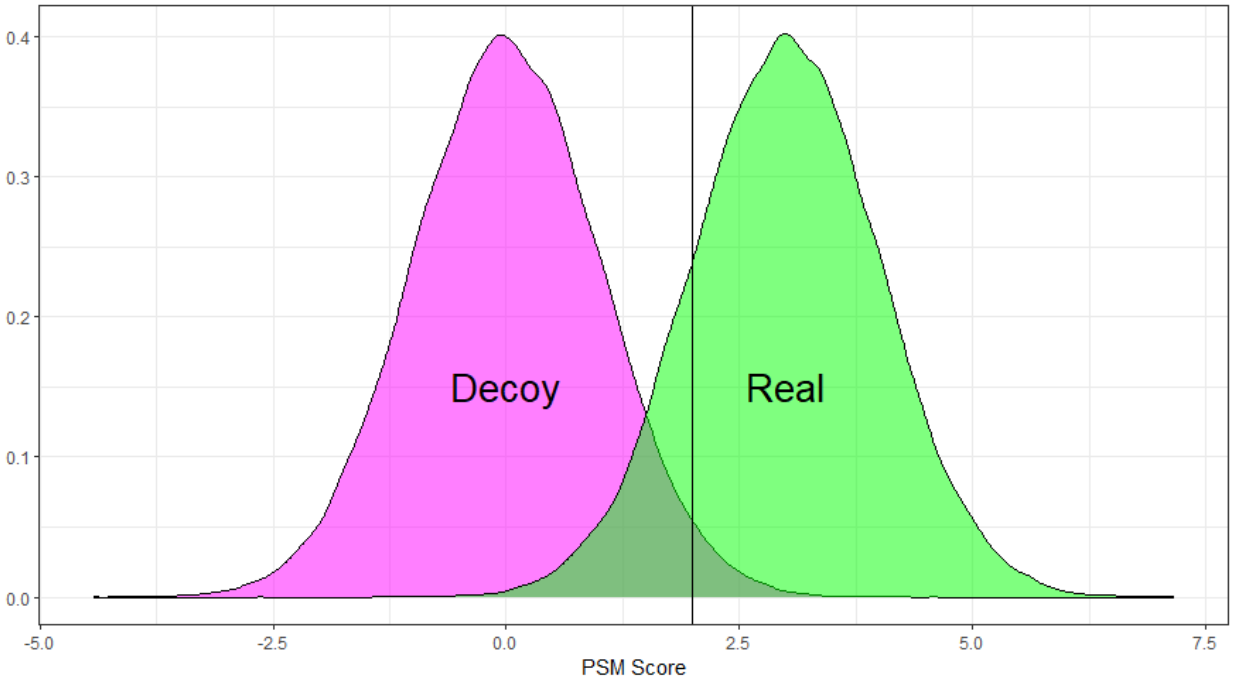
https://pubs.acs.org/doi/10.1021/pr101065j

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



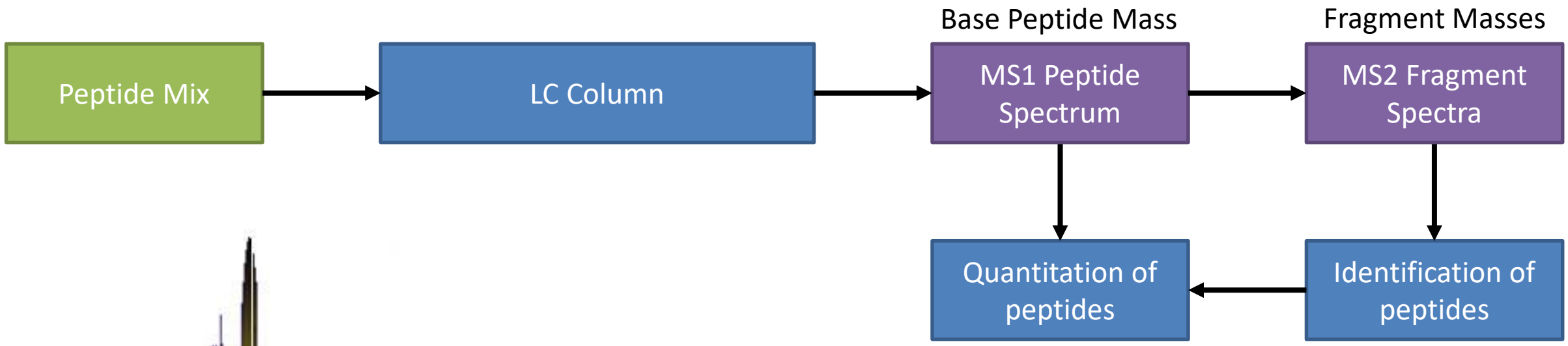
PEP Score
Probability of peptide being wrongly identified

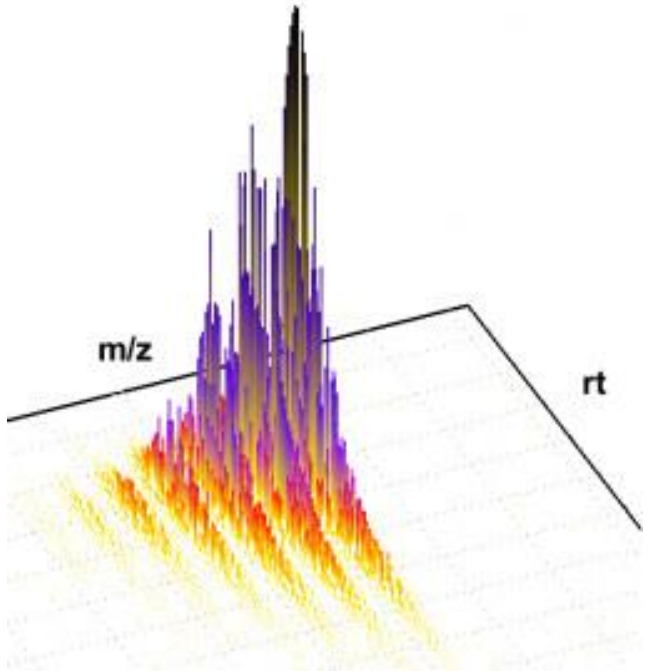
Q-value
Ratio of best Real hit to best Decoy hit

Quantitating Proteins

Babraham Bioinformatics

Label Free Quantitation





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

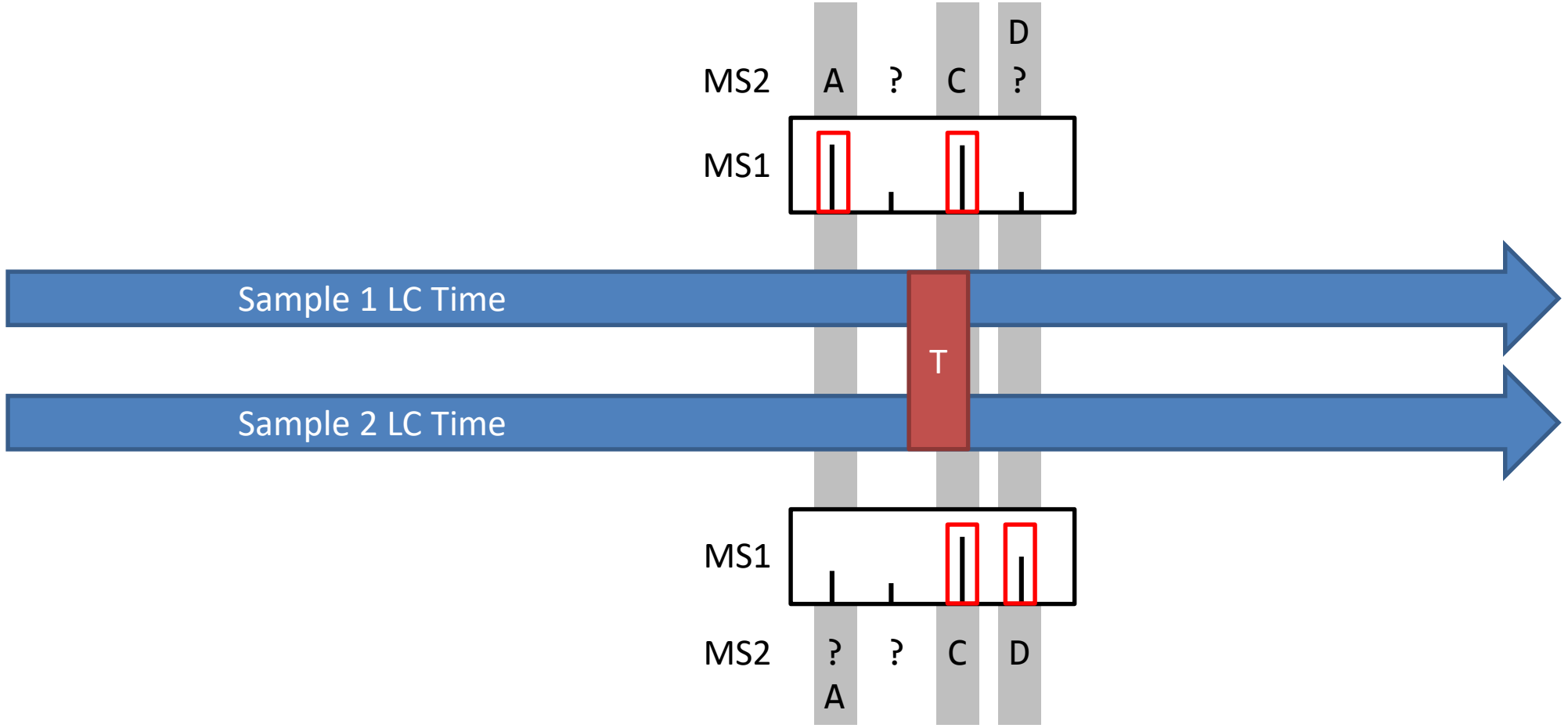
DOI:10.1186/1756-0381-2-4

Measuring multiple samples

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

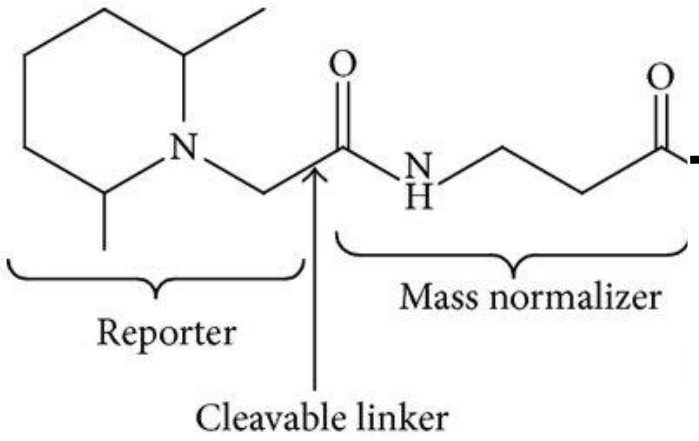
- How to measure consistently across samples?

Finding missing label free MS2 peaks



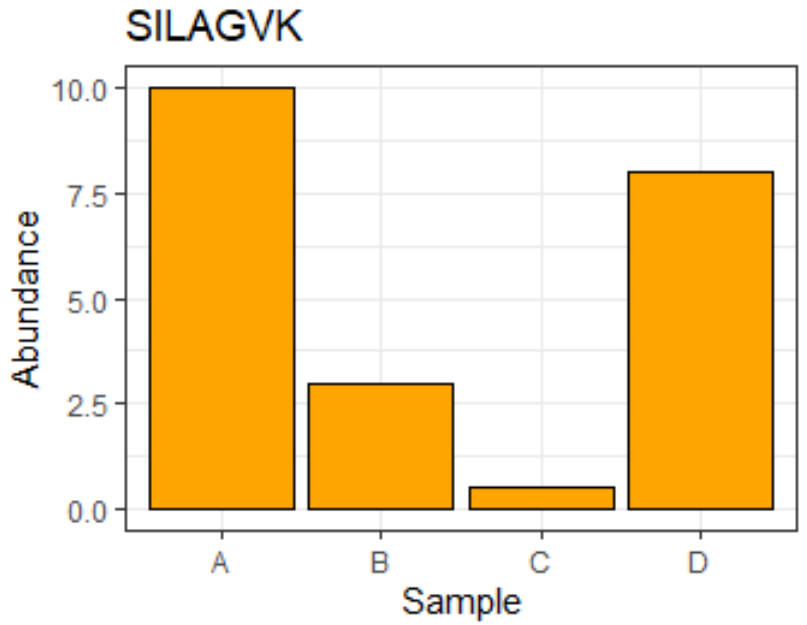
Matching MS1 base peaks based on LC time and M/Z allows more consistent data collection.

Tandem Mass Tagging (TMT)



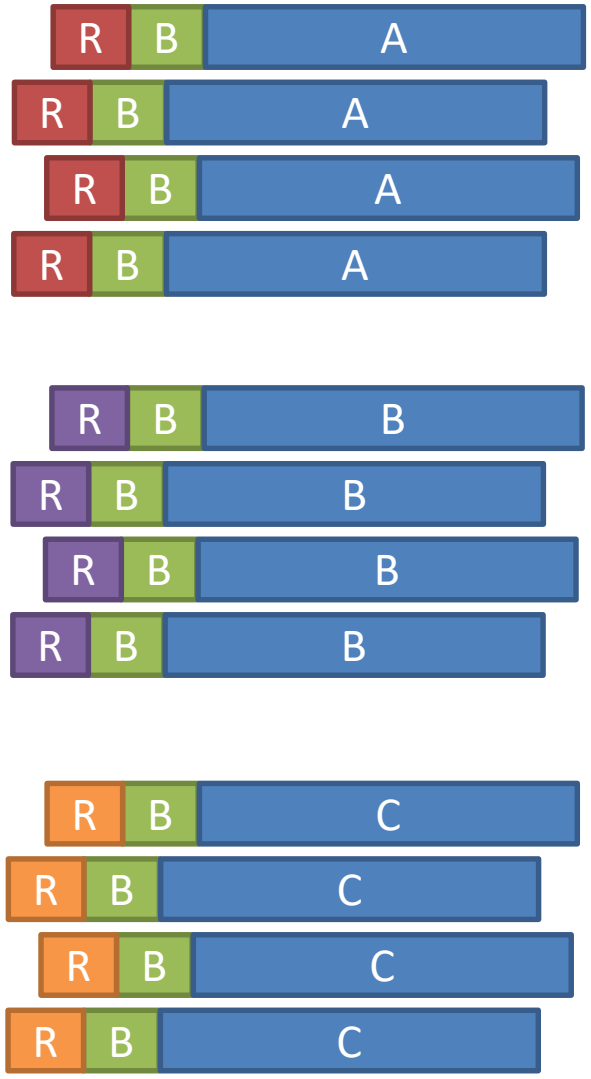
SILAGVR

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

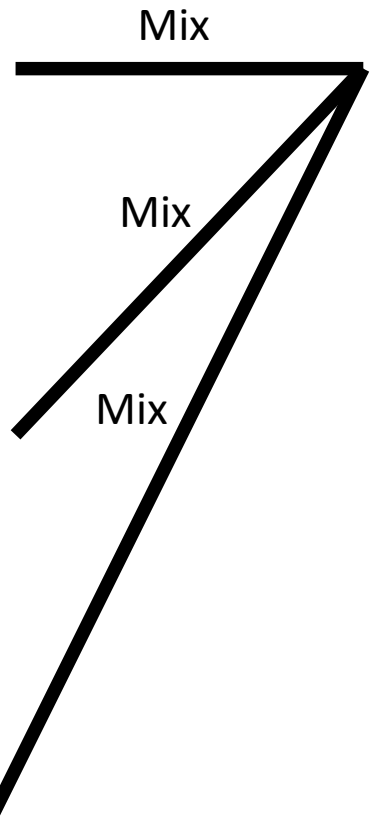


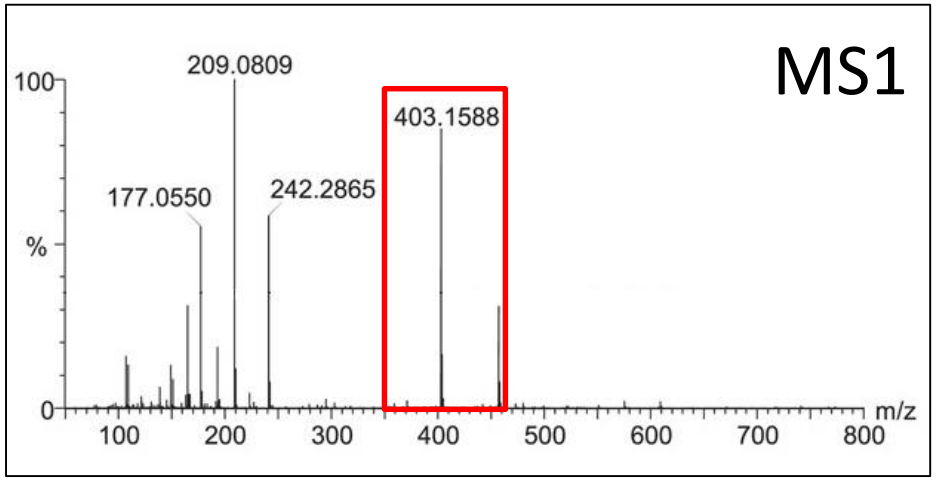
DOI:10.1155/2014/767812

Tandem Mass Tagging

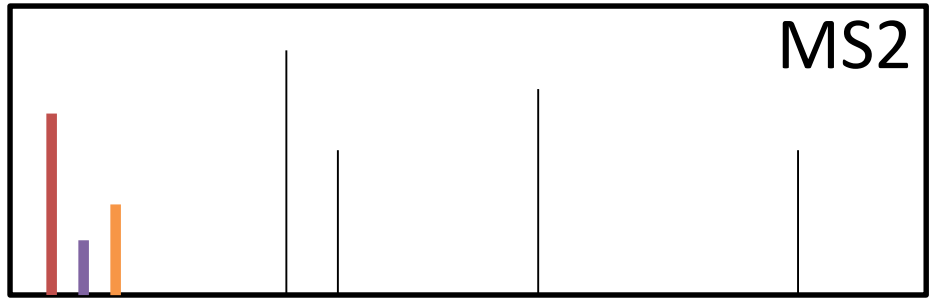


>15 reporters available





Peptides from all samples run together with a fixed mass shift

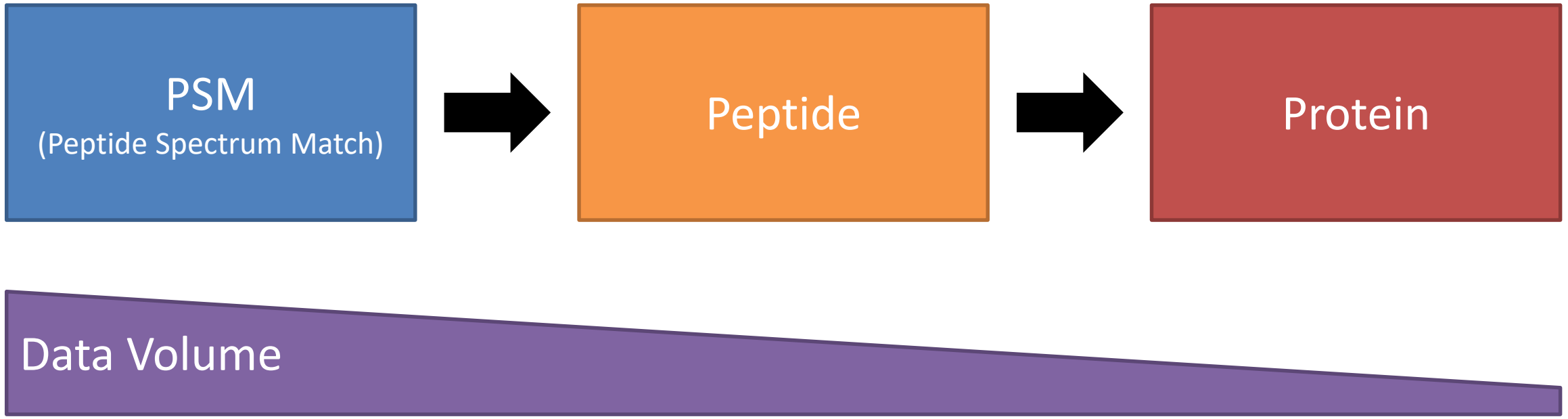


Reporters detach leaving a separately quantifiable signal

Moving from peptides to proteins

Babraham Bioinformatics

Levels of quantitation

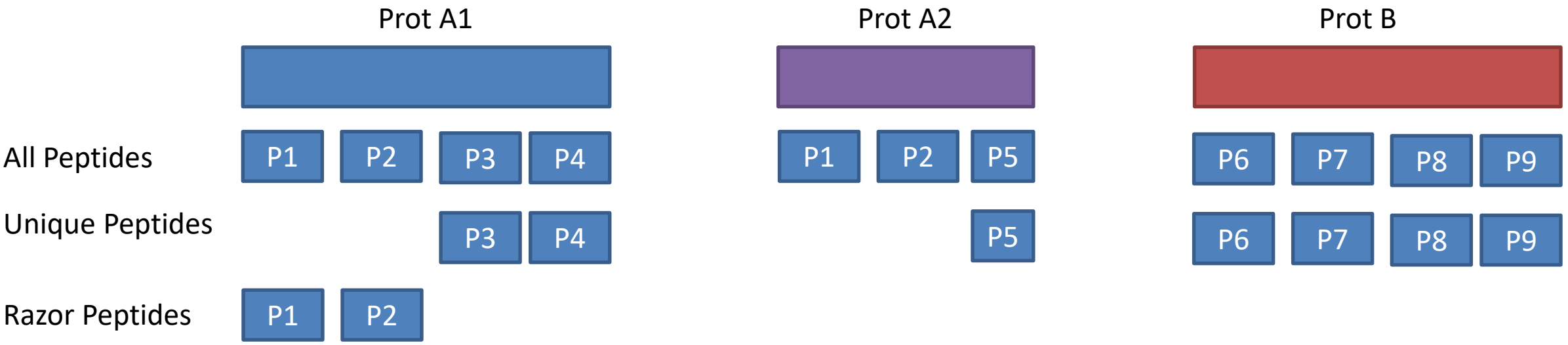


PSMs to Peptides

- 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



Assigning Razor Peptides

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

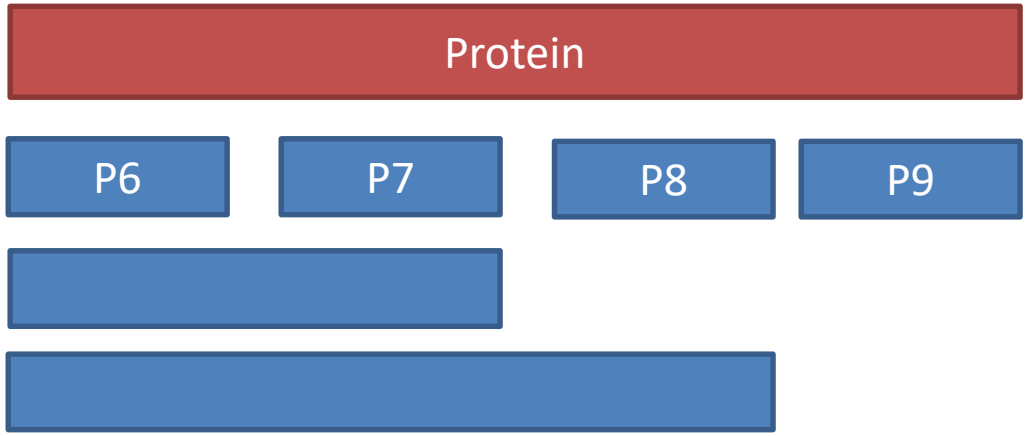
Quantitative value assignment

- 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

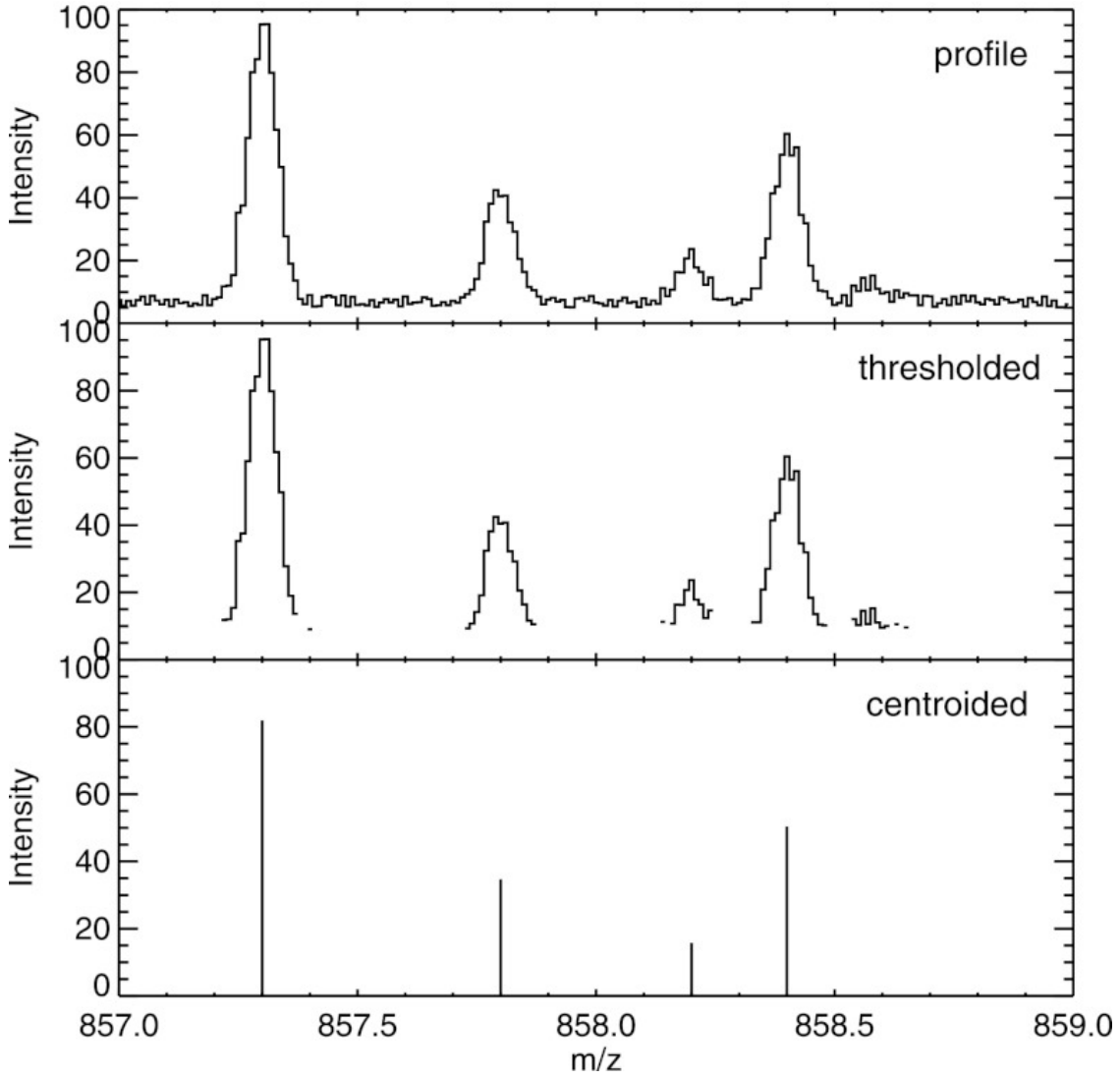
Instrument Provider	Extension	File type
Agilent	.D	instrument data format
Bruker	.BAF	instrument data format
Bruker	.FID	instrument data format
Bruker	.YEP	instrument data format
ABI/Sciex	.WIFF	QSTAR and QTRAP file format
ABI/Sciex	.t2d	4700 and 4800 file format
Thermo Xcalibur, Micromass (Waters), 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

ThermoFisher SCIENTIFIC

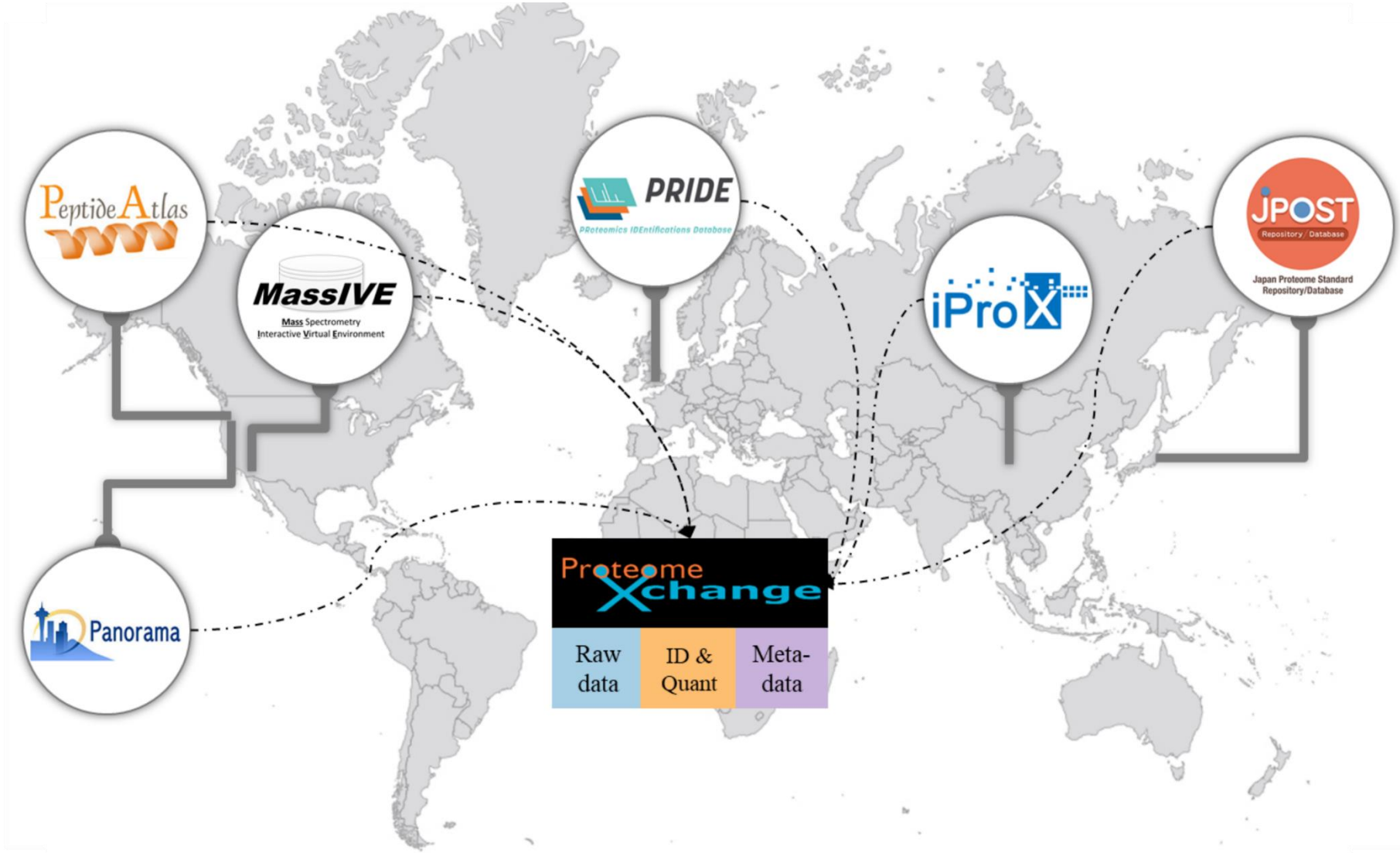
Most common format (>70% of PRIDE)

Information in RAW files

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



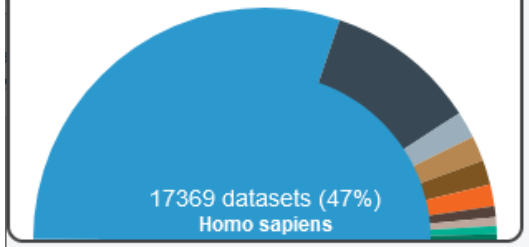
Data Repositories for Proteomics Mass Spec



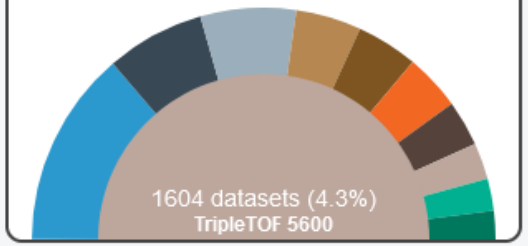
ProteomeXchange ProteomeCentral

Browse ProteomeXchange Datasets

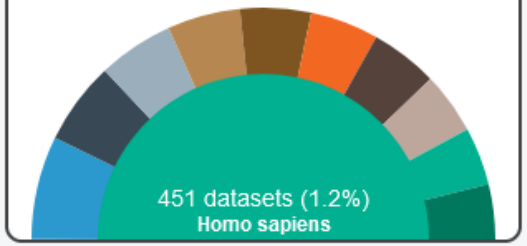
Top 10 Species



Top 10 Instruments



Top 10 Keywords



USI

Need to access individual spectra from a ProteomeXchange dataset?

USI mzspec:

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

Identification Results

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

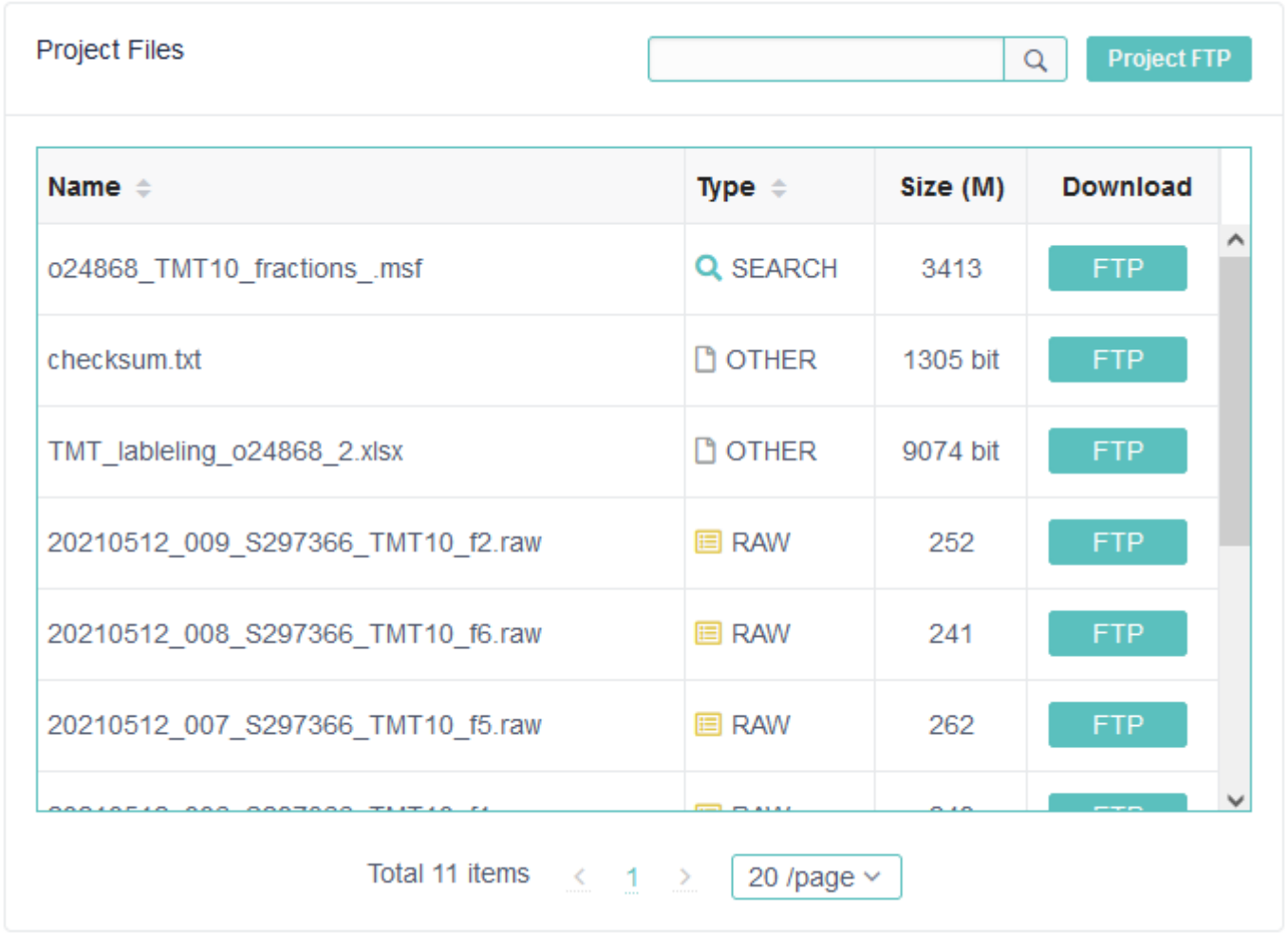
- Mass spec collection metrics
- Organism
- Modifications
- (Search method)

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



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?

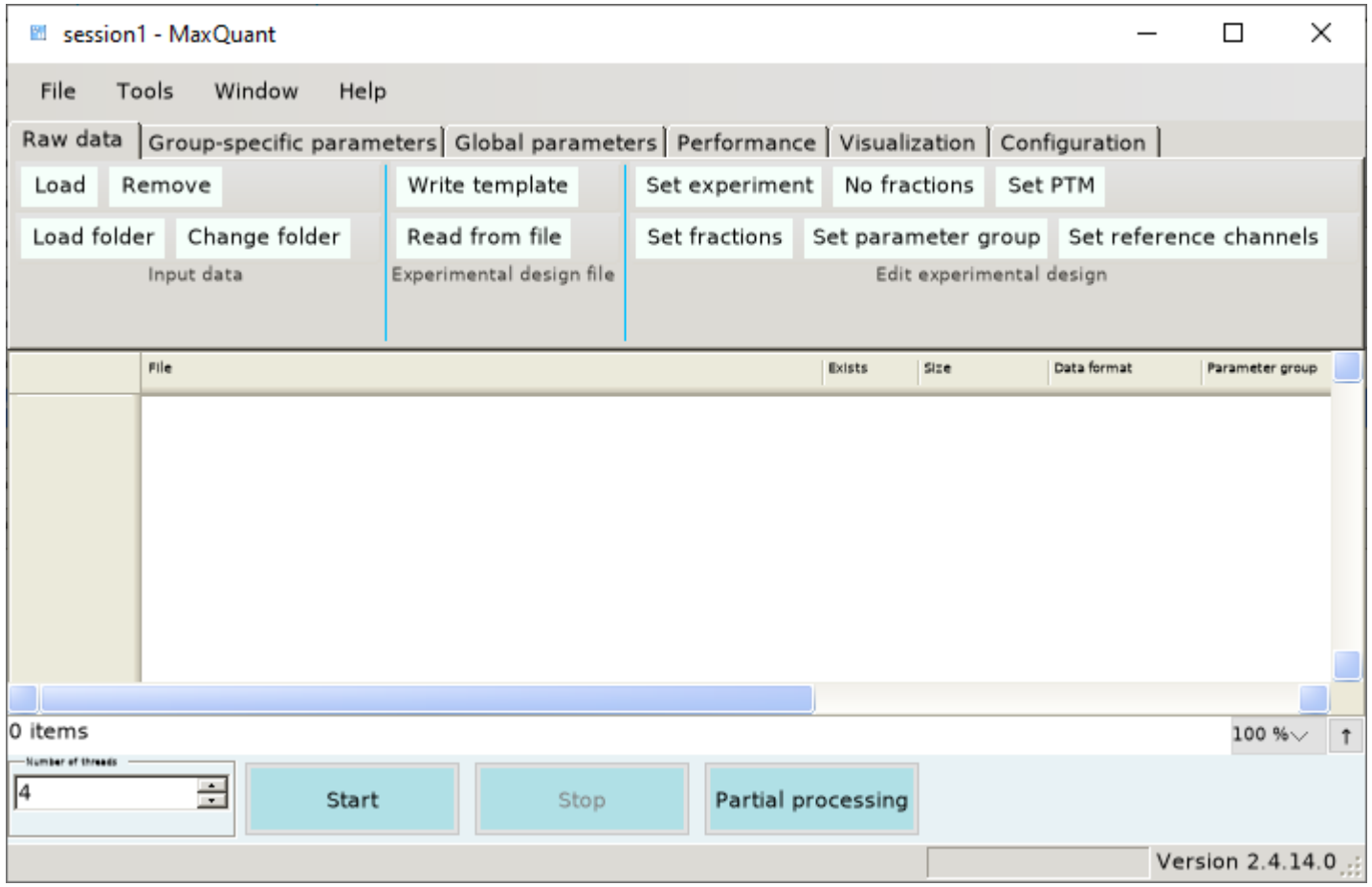
- Specific thresholds
  - Mass accuracy
  - LC time flexibility
  - Statistical thresholds

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

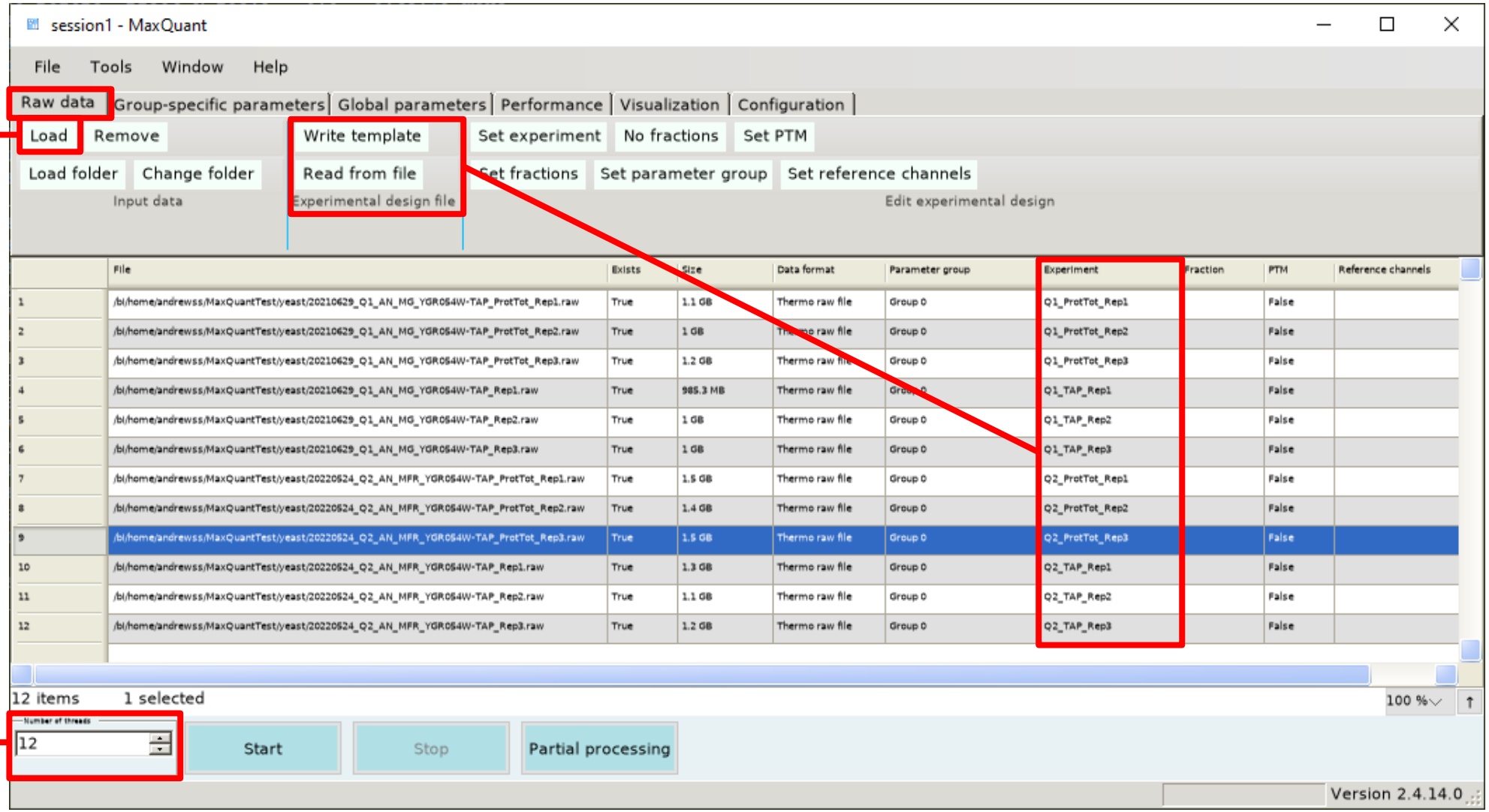
- Run search



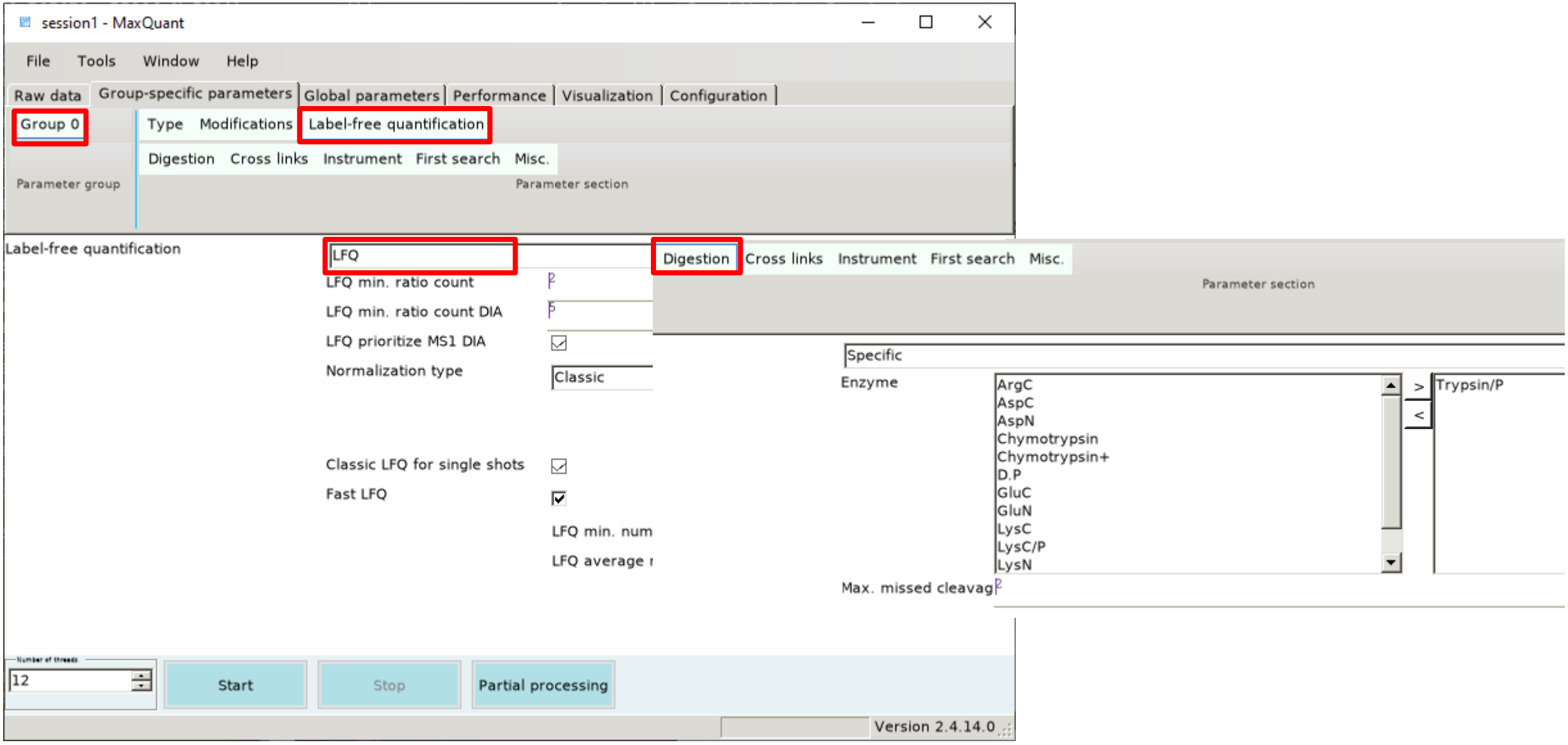
Load Raw Files

Select RAW files

Select Cores

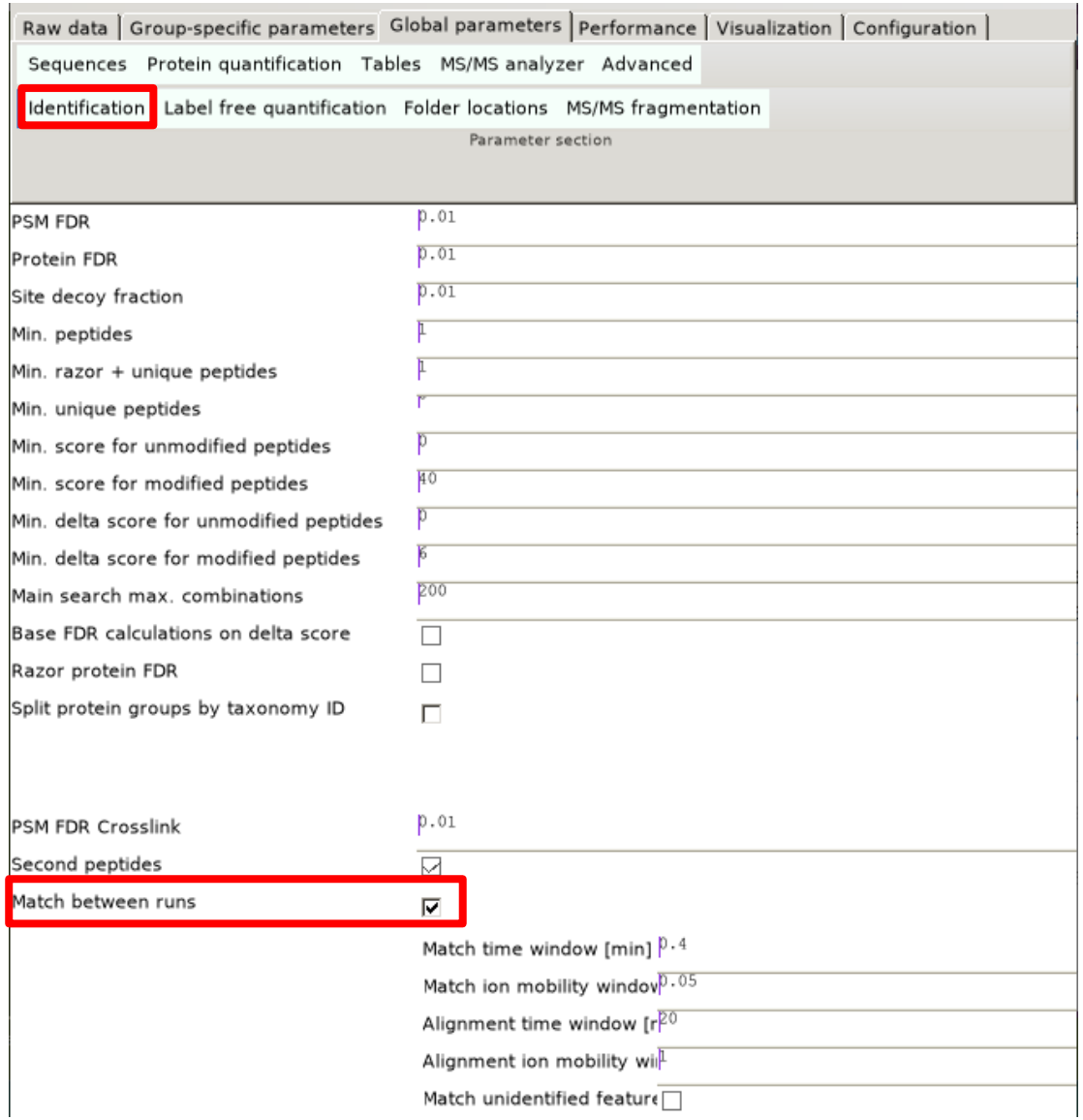


Set Quantitation

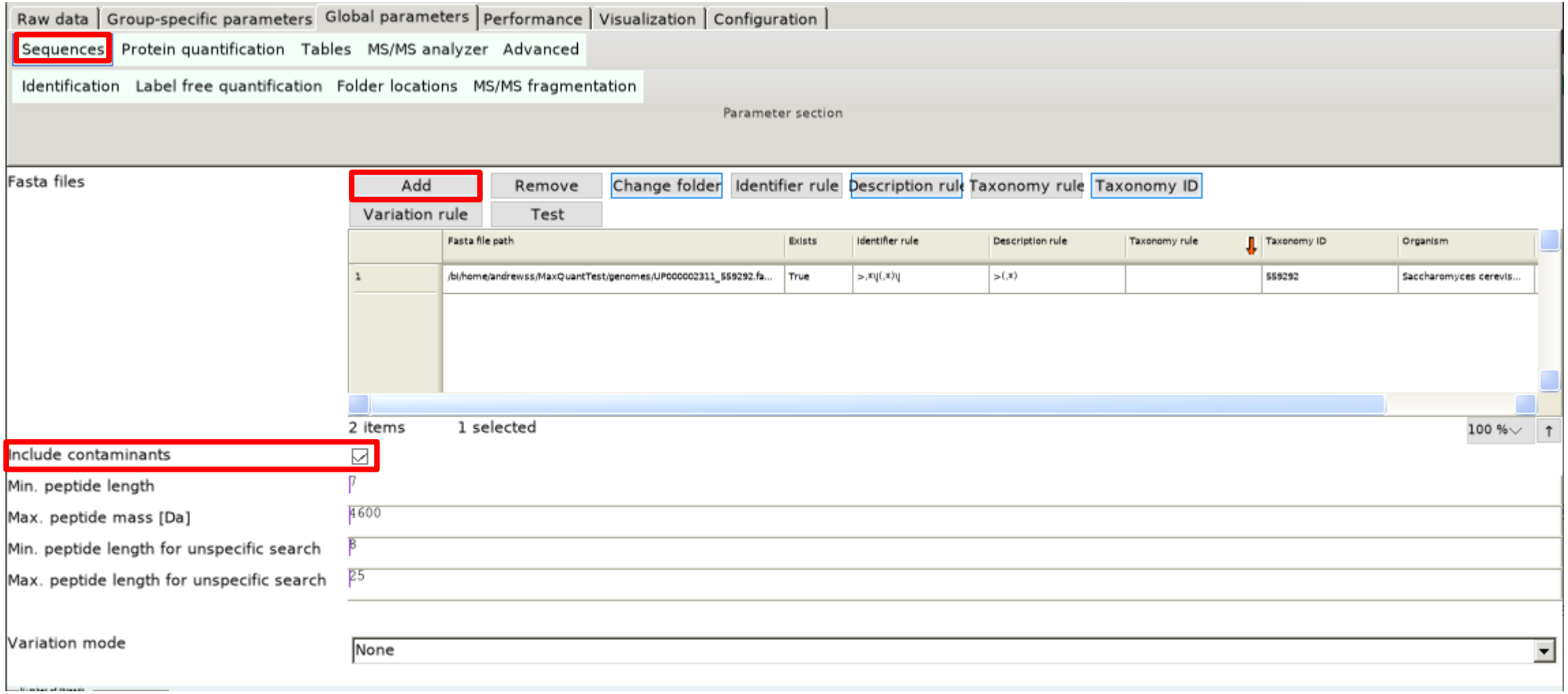


Identification Parameters

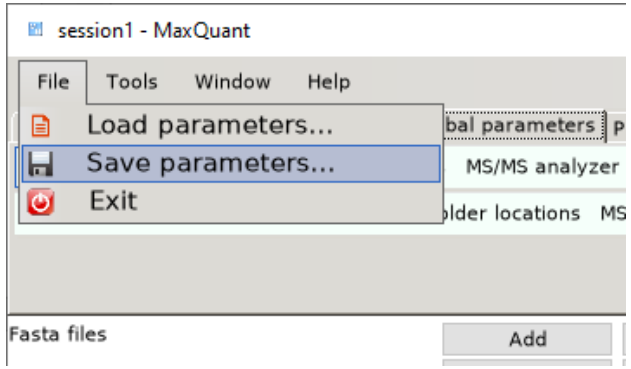
Orbitrap	
First search peptide tolerance	20
Main search peptide tolerance	4.5
Peptide tolerance unit	ppm
Individual peptide mass tolerance	☑
Isotope match tolerance	2
Isotope match tolerance unit	ppm
Centroid match tolerance	8
Centroid match tolerance unit	ppm
Centroid half width	35
Centroid half width 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



Search Sequences



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
```

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

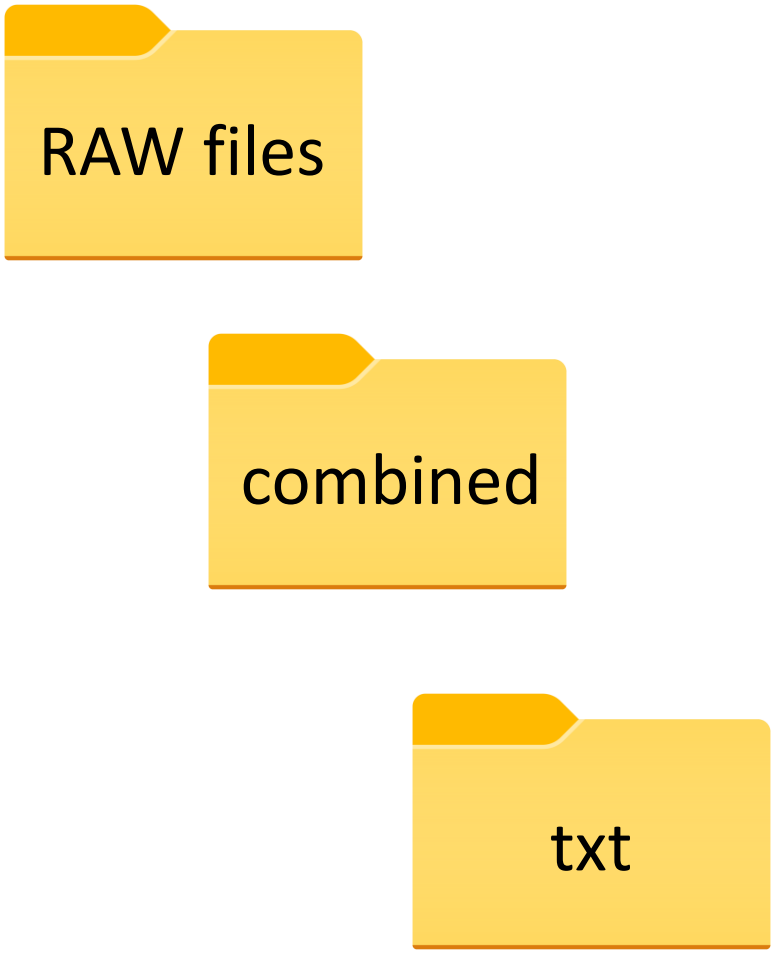
Read search results for recalibration
Mass recalibration
Calculating masses
MS/MS preparation for main search
Combining apl files for main search
MS/MS main search
Preparing combined folder
Correcting errors
Reading search engine results
Preparing reverse hits
Finish search engine results
Filter identifications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

Analysising Proteomics Data in R

Babraham Bioinformatics

Packages for Proteomics Analysis

MSstats

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

DEP

This is the **released** version of DEP; for the devel version, see DEP.

Differential Enrichment analysis of Proteomics data

QFeatures

Quantitative features for mass spectrometry data

PlexedPiper 0.4.2 Reference Articles Changelog

PlexedPiper

R package used at PNNL for processing isobaric labeling (e.g. TMT) proteomics data.

tidyproteomics 1.8.5 Overview Reference Articles Updates

tidyproteomics

An R package for the tidy-ing, post processing and analysis of quantitative proteomic data.

promor

Proteomics Data Analysis and Modeling Tools

CRAN 0.2.1 | downloads 278/month | downloads 7591 | R-CMD-check no status | test-coverage no status | License LGPL v2.1

- `promor` is a user-friendly, comprehensive R package that combines proteomics data analysis with machine learning-based modeling.

Main Steps in Analysis

- Load Data
  - Standard and Non-Standard Data formats
- Quality Control
  - PSM or Protein Level
- Quantitation / Normalisation
  - Also including missing value imputation
- Visualisation Exploration
  - Most important part!
- Statistical Analysis
  - Several Options depending upon design

Packages can make some of this easier, but they're all different and they get harder if you don't have standard input file formats.

None of the steps is hard once you know normal R/Tidyverse

	A	B	C	D	E	F	G	H	I	J
1		zygote_2_2	zygote_2_3	zygote_2_4	zygote_2_5	zygote_2_6	X2cell_3_4	2cell_3_5	2cell_3_6	2cell_3_7
2	Oas1h	139.4004	156.76555	160.084	156.19365	146.5641	1.88814085	0	0	21.05470179
3	Padi6	605.372	643.235	609.112	587.944	541.948	54.1708	42.8056	12.9178	36.9947
4	Pou2f1	9.428005	10.703015	9.4506	8.84173	10.01029	0.337351415	1.030093135	0	0
5	Pramef12	309.832	309.904	306.795	258.961	290.311	37.1347	26.6482	0	9.96366
6	Hsd3b7	69.77035	74.8341	75.2975	78.22125	62.61445	4.91305	0	0	9.99877637
7	Gmds	76.3921	76.3686	69.8795	63.6616	71.6125	0.297736	0	0	6.098
8	Nup107	47.0003	43.4688	37.2812	43.1323	47.1606	6.1709	0.12103	0	0
9	Mbl2	74.2078	87.3931	79.5226	67.3607	76.1681	15.3012	0	0	0.216471
10	Il31ra	28.5872	24.1273	28.1343	24.8542	29.8911	0	0	0	5.94977
11	Nuf2	110.794	119.863	134.625	121.496	105.549	29.2015	15.8391	0	7.63174
12	Dusp18	30.4356	24.3397	26.1815	26.0259	26.309	0	2.68527	6.65296	2.23609
13	Mat2b	477.3425	636.363	566.1895	538.308	560.617	39.875255	57.6356	63.32809	143.816
14	Nap1l4	73.77159333	71.08190543	62.73506047	63.81268333	58.75180916	0.3627249	5.058942599	8.165962767	2.986598137
15	Gtf2b	705.447	635.075	712.322	749.458	714.933	129.82	107.88	7.54549	198.708
16	Oas1e	945.771	1132.25	964.196	1041.23	1077.06	114.35	88.3087	81.8124	107.596
17	Zfp873	43.7166	39.1737	44.2615	43.8095	49.1028	0	0	0.233276	0
18	Ehf	30.852	34.2135	36.9557	39.1932	28.3825	0	0	7.6353	0
19	Stk3	35.0277	40.0571	32.4067	43.2943	37.2432	0	0.118031	0	5.10624

- Only gene name
- Pre-normalised values
- Missing = 0
- Conditions in sample names
- No other metrics

Metastasis

Endocytosis: a pivotal pathway for regulating metastasis

Imran Khan & Patricia S. Steeg

British Journal of Cancer **124**, 66–75 (2021) | Cite this article

	A	B	C	D	E	F
1	Genes	C:\Rawflles\Ref1101\Ref_1101-S01_1%.raw	C:\Rawflles\Ref1101\Ref_1101-S02_1%.raw	C:\Rawflles\Ref1101\Ref_1101-S03_1%.raw	C:\Rawflles\Ref1101\Ref_1101-S04_1%.raw	C:\Rawflles\Ref1101\Ref_1101-S05_1%.raw
2	2O16	1.88971e+07	1.49584e+07	1.2133e+07	1.20657e+07	1.09934e+07
3	3A339	188964				
4	3E324	1.03254e+07	6.6926e+06	5.91882e+06	1.03017e+07	1.08287e+07
5	4D656	4.73998e+08	1.34041e+08	1.4164e+08	3.05603e+08	2.30786e+08
6	5N226	4.7673e+07	4.03076e+07	3.71278e+07	4.92336e+07	3.06307e+07
7	AC3.5	7.56901e+07	1.07305e+08	9.52607e+07	1.32892e+08	8.439e+07
8	B0001.2	6.5548e+06	5.87955e+06	4.541e+06	4.62002e+06	6.90749e+06
9	B0001.4	7.58631e+07	9.58213e+07	8.52268e+07	1.02977e+08	1.0044e+08
10	B0001.7	589647			391112	
11	B0001.8	139444			69656	132116
12	B0024.11	2.66194e+07	2.7866e+07	2.45648e+07	2.1951e+07	2.29323e+07
13	B0024.13	7.51997e+06	6.98251e+06	6.81422e+06	7.66024e+06	1.34451e+07
14	B0024.4		2.18389e+07	3.05439e+07	6.68489e+07	1.81743e+08
15	B0035.13		2.24117e+07	1.27015e+07	2.31195e+07	
16	B0035.15	8.6501e+06	1.68235e+06	721716	3.87669e+06	4.94482e+06
17	B0035.3	5.22354e+07	7.27228e+07	6.43076e+07	8.27617e+07	7.63989e+07
18	B0041.8	4.02538e+06	5.00025e+06	4.66612e+06	6.55977e+06	4.99386e+06

- Gene IDs – no names
- Raw Intensity Values
- Missing = blank (NA in R)
- Only meaningless file names for samples
- Metadata supplied separately
- No search metrics

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Genes	C:\Rawfiles\Ref1059\Ref_1059_03.raw mT HOM - b_1	C:\Rawfiles\Ref1059\Ref_1059_07.raw mT HOM - b_2	C:\Rawfiles\Ref1059\Ref_1059_10.raw mT HOM - b_3	C:\Rawfiles\Ref1059\Ref_1059_14.raw mT HOM - b_4	C:\Rawfiles\Ref1059\Ref_1059_01.raw DELTA HET_1	C:\Rawfiles\Ref1059\Ref_1059_05.raw DELTA HET_2	C:\Rawfiles\Ref1059\Ref_1059_09.raw DELTA HET_3	C:\Rawfiles\Ref1059\Ref_1059_13.raw DELTA HET_4	C:\Rawfiles\Ref1059\Ref_1059_02.raw mT HET_1	C:\Rawfiles\Ref1059\Ref_1059_06.raw mT HET_2	C:\Rawfiles\Ref1059\Ref_1059_12.raw mT HET_3	C:\Rawfiles\Ref1059\Ref_1059_16.raw mT HET_4	C:\Rawfiles\Ref1059\Ref_1059_04.raw mT HOM +b_1	C:\Rawfiles\Ref1059\Ref_1059_08.raw mT HOM +b_2	C:\Rawfiles\Ref1059\Ref_1059_11.raw mT HOM +b_3	C:\Rawfiles\Ref1059\Ref_1059_15.raw mT HOM +b_4
2	(Bos	633769	667158	510547	452928	2.13E+06	2.12E+06	1.27E+06	1.51E+06	2.28E+06	2.18E+06	1.39E+06	1.43E+06	2.61E+06	2.15E+06	1.30E+06	1.37E+06
3	(S.avidinii)		275504	554503	3.77E+06	1.93E+06											
4	AVD	2.44E+09	2.28E+09	2.50E+09	2.57E+09	2.43E+09	2.78E+09	2.40E+09	2.45E+09	2.22E+09	2.13E+09	2.43E+09	2.54E+09	2.26E+09	2.15E+09	1.77E+09	1.90E+09
5	Aaas	1.43E+06	1.36E+06	1.22E+06	1.13E+06	7.77E+06	6.59E+06	6.03E+06	6.12E+06	3.49E+06	3.10E+06	3.42E+06	2.85E+06	5.26E+06	4.41E+06	3.75E+06	3.42E+06
6	Aacs	3.34E+06	3.46E+06	3.23E+06	3.25E+06	1.15E+07	9.26E+06	6.29E+06	7.66E+06	1.17E+07	1.03E+07	8.97E+06	9.96E+06	1.50E+07	1.48E+07	1.04E+07	1.03E+07
7	Aagab									669863			539313	1.82E+06	1.31E+06	976314	1.07E+06
8	Aak1	1.19E+07	1.12E+07	1.14E+07	1.14E+07	2.83E+07	2.91E+07	2.30E+07	2.40E+07	3.56E+07	3.05E+07	3.00E+07	3.18E+07	5.21E+07	5.11E+07	4.39E+07	4.45E+07
9	Aar2	165302				630393	377120	367766	399686	532196	472607	418727	409495	1.13E+06	967882	408205	569772
10	Aars1	6.09E+06	5.92E+06	8.47E+06	7.11E+06	1.20E+07	1.32E+07	1.25E+07	1.34E+07	1.18E+07	1.23E+07	1.61E+07	1.31E+07	1.94E+07	1.75E+07	1.87E+07	1.89E+07
11	Aarsd1	192173		469549		853193					743884	1.12E+06	958322				854443
12	Aasdhppt	651215	592978	447370	512395	2.06E+06	1.49E+06	1.01E+06	1.17E+06	1.88E+06	1.36E+06	1.45E+06	1.44E+06	2.44E+06	2.25E+06	1.37E+06	1.42E+06
13	Abca1					2.62E+06		1.44E+06	1.63E+06			1.92E+06	1.92E+06	3.19E+06	3.26E+06	2.96E+06	2.88E+06
14	Abca2					625328	516950	386555	460955								
15	Abca3					226413	298681		202501	234477	246548						

- Gene names plus other conventions
- Raw Intensity Values
- Missing = blank (NA in R)
- Filenames plus conditions as headers
- No search metrics

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Master Protein	Annotated Sequence	Positions in M	Modifications in Master Protei	Master Protein Descriptions	Modifications	# PSMs		Theo. M	# Prote	# Prote	C	Abundar	Abunda	Abunda	Abunda	Abunda
2	Q15046	[K].EICNAYTELNDPMR.[Q]	Q15046 [494-507]	Q15046 1xC6-CysPAT [C496(100)]	Lysine--tRNA ligase [OS=Homo sapiens]	1xOxidation [M13]; 1xTMTp	2	0	2210.019	1	1	##	-1.1	0.15	0.41	-0.26	0.29
3	Q05209	[K].IADGVNEINTENMVSSIE	Q05209 [309-339]	Q05209 1xPhospho [S332(100)]	Tyrosine-protein phosphatase non-receptor t	1xOxidation [M13]; 2xTMTp	1	1	4397.252	1	1	##					
4	P18031	[R].SGTFCLADTCLLLMDK.[R	P18031 [222-237]	P18031 2xC6-CysPAT [C226(100);	Tyrosine-protein phosphatase non-receptor t	1xOxidation [M14]; 1xTMTp	4	0	2797.389	1	1	##	-3.31	0.71	1.17	-0.44	-0.28
5	Q15046	[K].LVGEFLEVTCINPTFICDH	Q15046 [447-474]	Q15046 2xC6-CysPAT [C456(100);	Lysine--tRNA ligase [OS=Homo sapiens]	1xOxidation [M23]; 1xTMTp	4	0	4182.139	1	1	##	-4.19	0.36	0.72	-0.71	0.66
6	Q06124	[K].YSLADQTSGDQSPLPPCT	Q06124 [547-573]	Q06124 2xC6-CysPAT [C563(100);	Tyrosine-protein phosphatase non-receptor t	1xOxidation [M26]; 1xTMTp	11	0	3624.64	1	1	##	-1.55	0	0.11	-0.03	0.56
7	Q06124	[K].YSLADQTSGDQSPLPPCT	Q06124 [547-573]	Q06124 2xC6-CysPAT [C563(100);	Tyrosine-protein phosphatase non-receptor t	1xOxidation [M26]; 1xTMTp	1	0	3704.606	1	1	##	-0.85	0.19	0.72	0.01	0.16
8	P42684	[R].SNSTSSMSSGLPEQDR.[	P42684 [781-796]	P42684 1xPhospho [S783(99.6)]	Tyrosine-protein kinase ABL2 [OS=Homo sapi	1xOxidation [M7]; 1xTMTpr	1	0	2082.897	1	1	##	-1.24	-0.07	0.59	0.15	0.46
9	Q06124	[R].VYENVGLMQQQK.[S]	Q06124 [579-590]	Q06124 1xPhospho [Y580(100)]	Tyrosine-protein phosphatase non-receptor t	1xOxidation [M8]; 1xTMTpr	5	0	2141.096	1	1	##	-2.13	0.89	2.07	2.22	4.33
10	Q15046	[K].KEICNAYTELNDPMR.[Q	Q15046 [493-507]	Q15046 1xC6-CysPAT [C496(100)]	Lysine--tRNA ligase [OS=Homo sapiens]	1xTMTpro [K1]; 1xTMTpro [N	2	1	2626.327	1	1	##	-5.02	0.46	0.66	-6.64	0.91
11	Q06124	[K].GVDCDIDVPK.[T]	Q06124 [483-492]	Q06124 1xC6-CysPAT [C486(100)]	Tyrosine-protein phosphatase non-receptor t	1xTMTpro [K10]; 1xTMTpro [	1	0	1889.994	1	1	##	-4.43	0.2	0.43	-6.64	-1.03

- Uniprot IDs as annotation
- TMT Log Ratios as measurements
- Missing = blank (NA in R)
- Condition names part of sample name
- Many search metrics

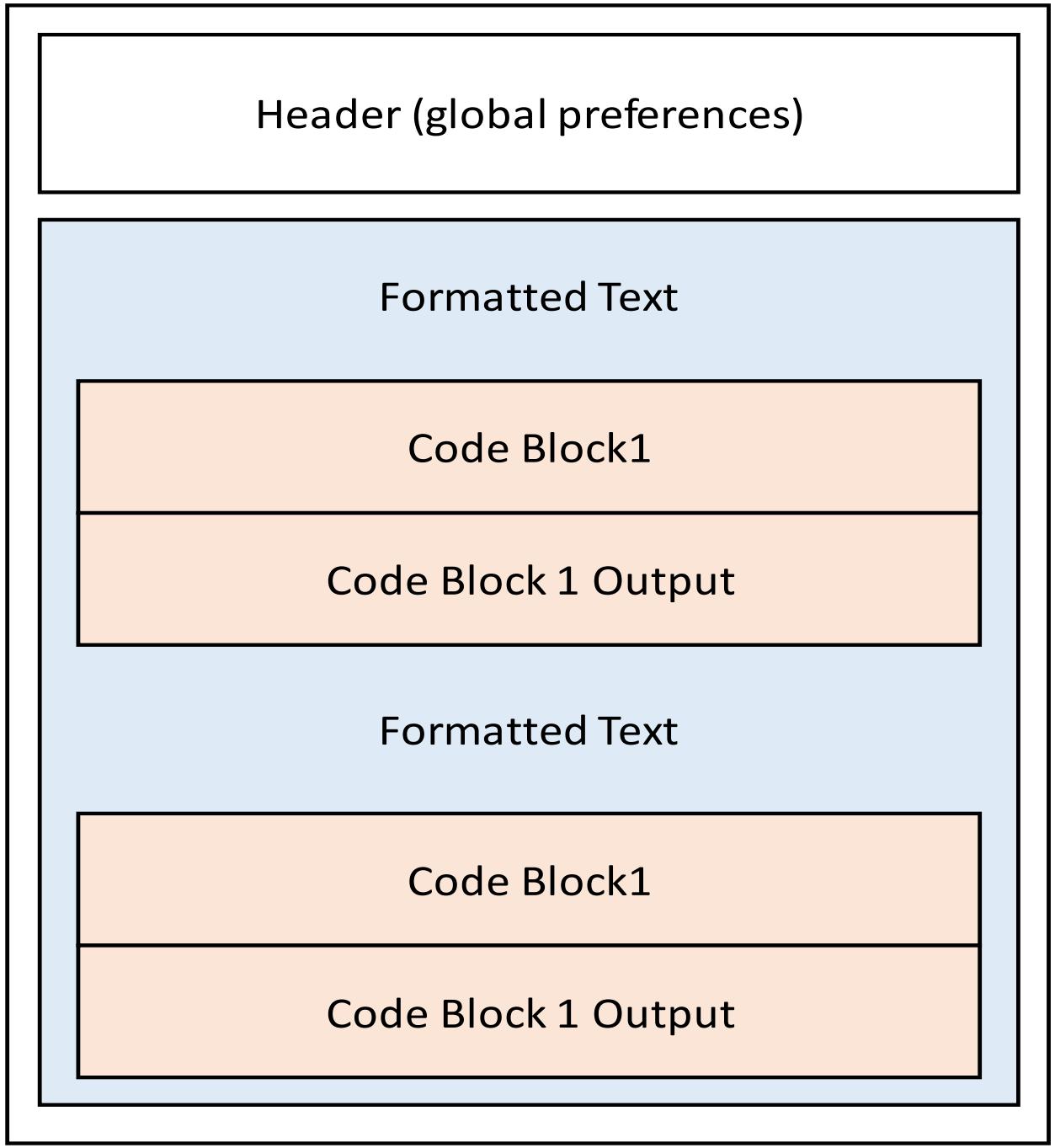
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1					Raw abundance				Normalized by median				As some Protein.Groups have mulitple protein IDs (highlighted), the first accession number was used when adding annotation (Gene Name, Keywords, etc)					
2	Contaminate proteins (bovine, keratins etc) are removed				Nuclear	Wash	Cytosol & Membrane	WCL	Nuclear	Wash	Cytosol & Membrane	WCL						
3	Protein.Group	Protein.Ids	Protein.Names	Genes	Ref_1061-1_DIA	Ref_1061-2_DIA	Ref_1061-3_DIA	Ref_1061-4_DIA	S1	S2	S3	S4	Accession	Gene name	Keywords	KEGG name	GOCC slim name	GOBP slim name
4	P02301;P68433;P	P02301;P68433;P842	H31_MOUSE;H32_MC	H3-3b;H3-5	2507490000	71446700	10375200	463926000	2507490000	27738687	26232189	952956466	P02301	H3-5	Acetylation;ADP-ribosylation;Chromosome;Citrullinati		cellular anatomical entity;	cellular component assemb
5	P0C0S6;Q3THW5	C0HKE1;C0HKE2;C0H	H2AV_MOUSE;H2AZ_	H2az1;H2a:	580423000	53580200	10139400	445798000	580423000	20802142	25636003	915719504	P0C0S6	H2az1	3D-structure;Acetylation;Ch	Systemic lupus erythemato	cellular anatomical entity;	biological regulation;cellul
6	P63260	P60710;P63260	ACTG_MOUSE	Actg1	416784000	577956000	300235000	426631000	416784000	224387421	759100680	876348318	P63260	Actg1	Acetylation;ATP-binding;Cy	Adherens junction;Arrhythr	apical part of cell;catalytic	actin cytoskeleton organiza
7	P62806	P62806	H4_MOUSE	H4c16	389175000	103582000	6571740	363546000	389175000	40214995	16615692	746764595	P62806	H4c1	3D-structure;Acetylation;Ch	Systemic lupus erythemato	cellular anatomical entity;	cellular component assemb
8	P43274	P43274	H14_MOUSE	H1-4	283031000	16459900	1438680	79413400	283031000	6390442	3637494	163124104	P43274	H1-4	Acetylation;ADP-ribosylation;Chromosome;Citrullinati		cellular anatomical entity;	biological regulation;cellul
9	P10853;P10854;Q	P10853;P10854;P706	H2B1B_MOUSE;H2B1	H2bc12;H2	223027000	27557700	1579900	79904400	223027000	10699087	3994548	164132673	P10853	H2bc7	Acetylation;ADP-ribosylati	Systemic lupus erythemato	cellular anatomical entity;	cellular component organiz
10	P43276	P43276	H15_MOUSE	H1-5	132770000	5975360	429377	27772500	132770000	2319892	1085618	57047856	P43276	H1-5	Acetylation;Chromosome;Citrullination;DNA-binding;H		cellular anatomical entity;	anatomical structure devel
11	P43277	P43277	H13_MOUSE	H1-3	124033000				124033000				P43277	H1-3	Acetylation;Chromosome;Citrullination;DNA-binding;H		cellular anatomical entity;	biological regulation;cellul
12	P14069	P14069	S10A6_MOUSE	S100a6	118717000		76144700	95578300	118717000		192520837	196328636	P14069	S100a6	3D-structure;Acetylation;Calcium;Cellmembrane;Cytop		cell projection;cellular anatomical entity;cytoplasm;cy
13	P48962	P48962;P51881;Q3V	ADT1_MOUSE	Slc25a4	106304000	44146800	20189400	55441300	106304000	17139690	51045972	113882700	P48962	Slc25a4	Acetylation;Antiport;Direct	Calcium signaling pathway	cellular anatomical entity;	biological regulation;cellul
14	Q61937	Q61937	NPM_MOUSE	Npm1	97173800	12937600	2680990	21185200	97173800	5022934	6778495	43516796	Q61937	Npm1	3D-structure;Acetylation;ADP-ribosylation;Chaperone;(		cellular anatomical entity;	biological regulation;biosy
15	P48036	P48036	ANXA5_MOUSE	Anxa5	94675700	16500400	9475930	27884700	94675700	6406166	23958516	57278327	P48036	Anxa5	Acetylation;Annexin;Bloodcoagulation;Calcium;Calciu		cell junction;cell projectior	biological regulation;cellul
16	P10107	P10107	ANXA1_MOUSE	Anxa1	94335000	26713000	6793360	22960000	94335000	10371138	17176026	47162436	P10107	Anxa1	Acetylation;Adaptiveimmunity;Annexin;Calcium;Calciu		cell projection;cell surface	actin cytoskeleton organiza
17	P10922	P10922	H10_MOUSE	H1-0	84643500			13583900	84643500			27902867	P10922	H1-0	Acetylation;ADP-ribosylation;Chromosome;Citrullinati		cellular anatomical entity;	biological regulation;cellul
18	P62983	P0CG49;P0CG50;P62	RS27A_MOUSE	Rps27a	82119600	56972500	35915400	43024700	82119600	22119179	90806883	88377599	P62983	Rps27a	3D-structure;Acetylation;AD	Ribosome	cell junction;cellular anato	amide metabolic process;b
19	P09405	P09405	NUCL_MOUSE	Ncl	79273700	25167200	6768200	25537400	79273700	9770991	17112413	52456707	P09405	Ncl	Acetylation;Cytoplasm;Dire	Pathogenic Escherichia col	cell cortex;cell surface;cell	anatomical structure forma

- Complex headers
- Uniprot IDs and Names
- Raw and Normalised Linear Intensity Values
- Missing = blank (NA in R)
- Real condition names
- No search metrics

R, Rstudio, Tidyverse, Notebooks

Notebook Structure

- Single overall text document, split into sections
  - Header (mostly preferences)
  - Body
    - Commentary (default)
    - R Code
    - Output (graphical and text)



Code


```
---
title: "Example Notebook"
output:
  html_document:
    df_print: paged
    toc: true
    toc_float: true
---

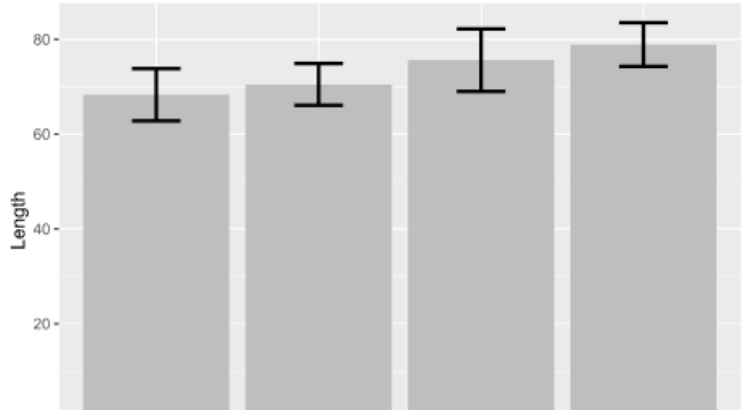
Introduction
============

This is an example of a notebook to show how they work.

```{r message=FALSE}
library(tidyverse)
```

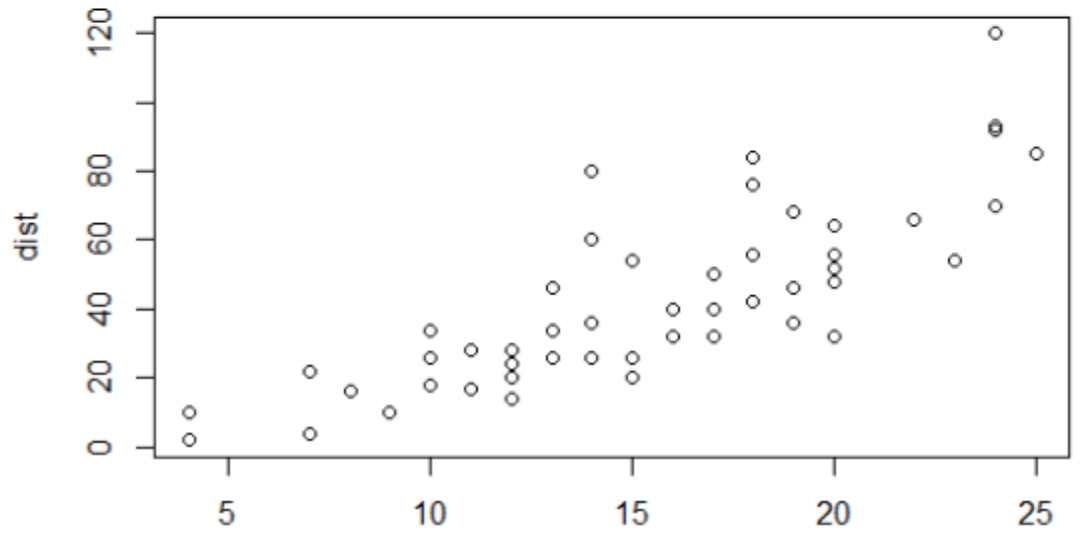
Processing
==========

Read the data

```{r message=FALSE}
read_tsv("small_file.txt") -> small
head(small)
```
```

Sample <chr>	Length <dbl>	Category <chr>
x_1	45	A
x_2	82	B
x_3	81	C
x_4	56	D
x_5	96	A

Summarise

We're going to calculate the mean of the lengths per category

```
small %>%
  group_by(Category) %>%
  summarise(
    count=n(),
    length=mean(Length)
  )
```

Category <chr>	count <int>	length <dbl>
A	10	68.3
B	10	70.5
C	10	75.6
D	10	78.9

4 rows

Plot

```
small %>%
  ggplot(aes(x=Category, y=Length)) +
  geom_bar(stat="summary", fun="mean", fill="grey")+
  stat_summary(geom="errorbar", width=0.3, size=1, fun.data=mean_se)
```



Output

Notebook sections

Header

Commentary

Code

```
---
title: "R Notebook"
output: html_notebook
---

This is an [R Markdown](http://rmarkdown.rstudio.com) Notebook. When you execute code within the notebook, the results appear beneath the code.

Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Ctrl+Shift+Enter*.

```{r}
plot(cars)
```
```



Sections are marked by special quotes

```
---
```
`---` for header

```
```{r}
```
```` ```{r} ```` for R code

Default for unquoted text is commentary

Basic Structures in R

`myfunc(x, value=y)`	Runs `myfunc` using data `x` and `y`
`myfunc(x,y) -> saveme`	Saves the output of `myfunc`
`funca() |> funcb()`	Passes data from `funca` to `funcb`

Loading in Data

```
read_delim("evidence.txt", name_repair = "minimal") -> data
```

```
 [1] "Sequence"                          "Length"                          "Modifications"
 [4] "Modified sequence"                 "Oxidation (M) Probabilities"     "Oxidation (M) Score Diffs"
 [7] "Acetyl (Protein N-term)"           "Oxidation (M)"                   "Missed cleavages"
[10] "Proteins"                          "Leading proteins"                "Leading razor protein"
[13] "Gene names"                        "Protein names"                   "Type"
[16] "Raw file"                          "Experiment"                      "MS/MS m/z"
[19] "Charge"                            "m/z"                             "Mass"
[22] "Uncalibrated - Calibrated m/z [ppm]" "Uncalibrated - Calibrated m/z [Da]" "Mass error [ppm]"
[25] "Mass error [Da]"                   "Uncalibrated mass error [ppm]"   "Uncalibrated mass error [Da]"
[28] "Max intensity m/z 0"               "Retention time"                  "Retention length"
[31] "Calibrated retention time"         "Calibrated retention time start" "Calibrated retention time finish"
[34] "Retention time calibration"        "Match time difference"           "Match m/z difference"
[37] "Match q-value"                     "Match score"                     "Number of data points"
[40] "Number of scans"                   "Number of isotopic peaks"        "PIF"
[43] "Fraction of total spectrum"        "Base peak fraction"              "PEP"
[46] "MS/MS count"                       "MS/MS scan number"               "MS/MS scan numbers"
[49] "MS3 scan numbers"                  "Score"                           "Delta score"
[52] "Combinatorics"                     "Intensity"                       "Reverse"
[55] "Potential contaminant"             "id"                              "Protein group IDs"
[58] "Peptide ID"                        "Mod. peptide ID"                 "MS/MS IDs"
[61] "Best MS/MS"                        "Oxidation (M) site IDs"          "Taxonomy IDs"
```

Understanding your starting data

- Measurement level
  - PSM
  - Peptide
  - Protein
- Protein ID / Name / Description
- Sample Name
  - Experiment Metadata
- Measured Intensity
  - Linear or Log
  - Raw or Normalised
- Match Type
  - Direct or Matched Peak?
  - Reversed match?
  - Contamination match?
- Match Score (PEP, Q-Value etc)
- Retention time on LC
- Errors
  - Mass Error
  - Retention Error

Adding in Experimental Design

- Need to annotate conditions
  - Biological condition
  - Multiple Columns for more complex designs
- Might be part of Experiment name or imported externally

Experiment <chr>
Q2_ProtTot_Rep1
Q2_ProtTot_Rep2
Q2_ProtTot_Rep3
Q2_TAP_Rep2
Q2_TAP_Rep3
Q1_TAP_Rep3
Q2_TAP_Rep1
Q1_TAP_Rep1
Q1_TAP_Rep2
Q1_ProtTot_Rep1

```
data |>
  mutate(
    condition = str_replace(Experiment,"_Rep.*","")
  ) -> data
```

Adding in Experimental Data

Experiment <chr>
Q2_ProtTot_Rep1
Q2_ProtTot_Rep2
Q2_ProtTot_Rep3
Q2_TAP_Rep2
Q2_TAP_Rep3
Q1_TAP_Rep3
Q2_TAP_Rep1
Q1_TAP_Rep1
Q1_TAP_Rep2
Q1_ProtTot_Rep1

	A	B	C	D
1	Experiment	Batch	Condition	Replicate
2	Q2_ProtTot_Rep1	Q2	Total	1
3	Q2_ProtTot_Rep2	Q2	Total	2
4	Q2_ProtTot_Rep3	Q2	Total	3
5	Q1_ProtTot_Rep1	Q1	Total	1
6	Q1_ProtTot_Rep2	Q1	Total	2
7	Q1_ProtTot_Rep3	Q1	Total	3
8	Q2_TAP_Rep1	Q2	Pulldown	1
9	Q2_TAP_Rep2	Q2	Pulldown	2
10	Q2_TAP_Rep3	Q2	Pulldown	3
11	Q1_TAP_Rep1	Q1	Pulldown	1
12	Q1_TAP_Rep2	Q1	Pulldown	2
13	Q1_TAP_Rep3	Q1	Pulldown	3

```
data |>
  left_join(metadata) -> data
```

Quality Control of PSM Search Results

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

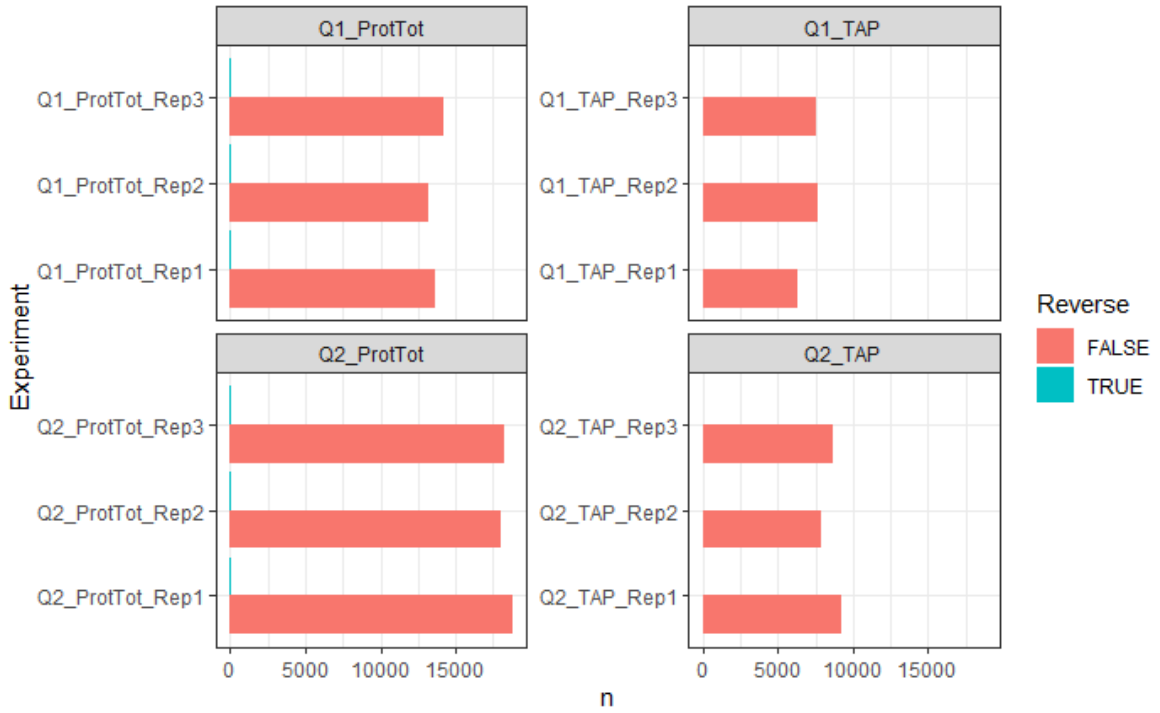
2. Problems during Chromatography
   - Even amounts of data over time
   - Consistent rate between experiments

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

Are we getting decoy hits reported?

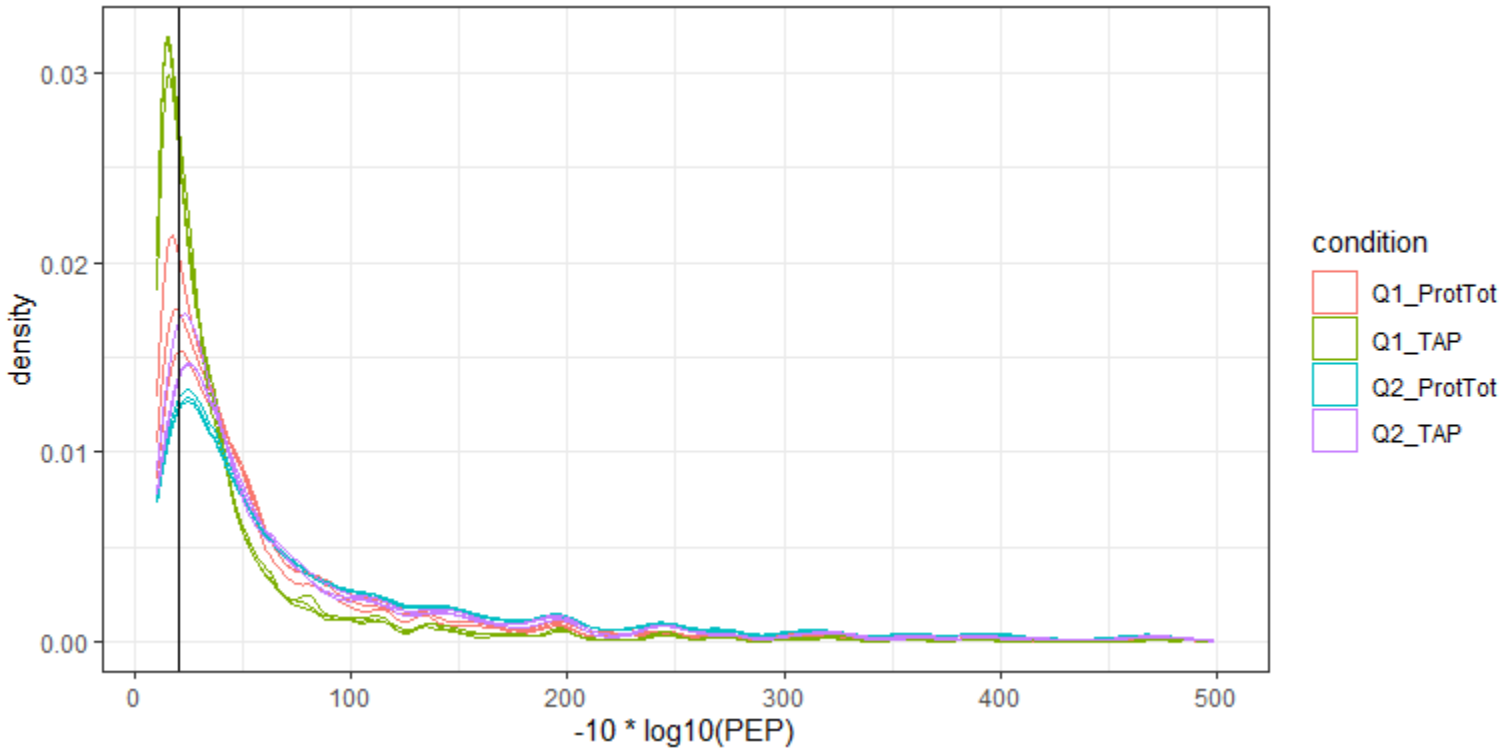
```
data |>
  group_by(
    Experiment,condition,Reverse
  ) |>
  count() |>
  mutate(Reverse = !is.na(Reverse)) |>
  ggplot(
    aes(x=Experiment, y=n, fill=Reverse)
  ) +
  geom_col(position="dodge") +
  coord_flip() +
  facet_wrap(
    vars(condition),
    scale="free_y“
  )


  data |>
    filter(is.na(Reverse)) -> data
```

Sequence <chr>	Experiment <chr>	Reverse <chr>
AAALNWLMIEGK	Q2_ProtTot_Rep2	+
AAALNWLMIEGK	Q1_ProtTot_Rep2	+
AAALNWLMIEGK	Q1_ProtTot_Rep3	+
AAALNWLMIEGK	Q2_ProtTot_Rep1	+
AAAALAGGK	Q2_ProtTot_Rep1	NA
AAAALAGGK	Q2_ProtTot_Rep2	NA
AAAALAGGK	Q2_ProtTot_Rep3	NA
AAAALAGGK	Q2_TAP_Rep2	NA



PEP Probability Distribution

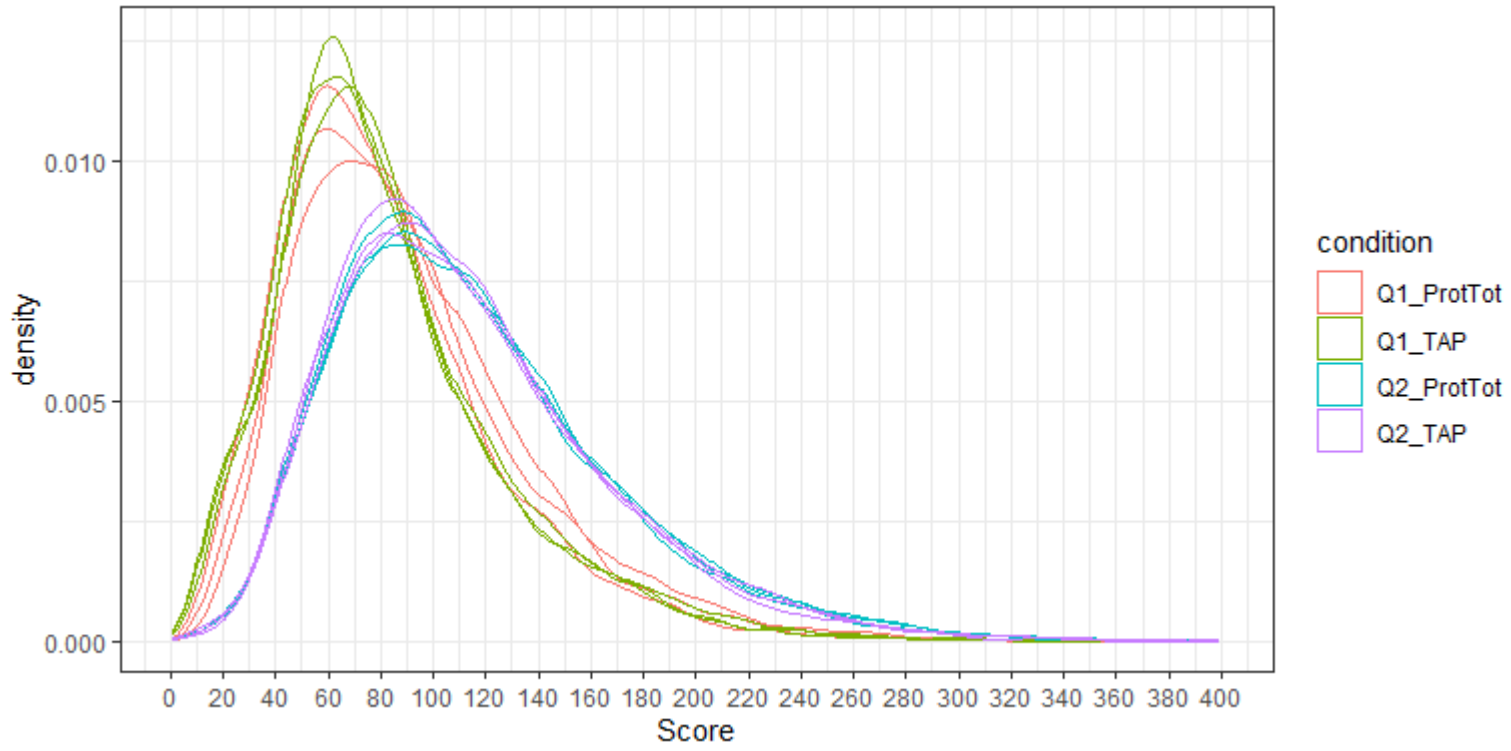
```
data |>
  ggplot(
    aes(
      x=-10*log10(PEP),
      group=Experiment,
      colour=condition
    )
  ) +
  geom_density() +
  geom_vline(xintercept = 20)
```



Phred Score = `-10 * log10(p)`

Match Score Distribution

```
data |>
  ggplot(
    aes(
      x=Score,
      group=Experiment,
      colour=condition
    )
  ) +
  geom_density()
```



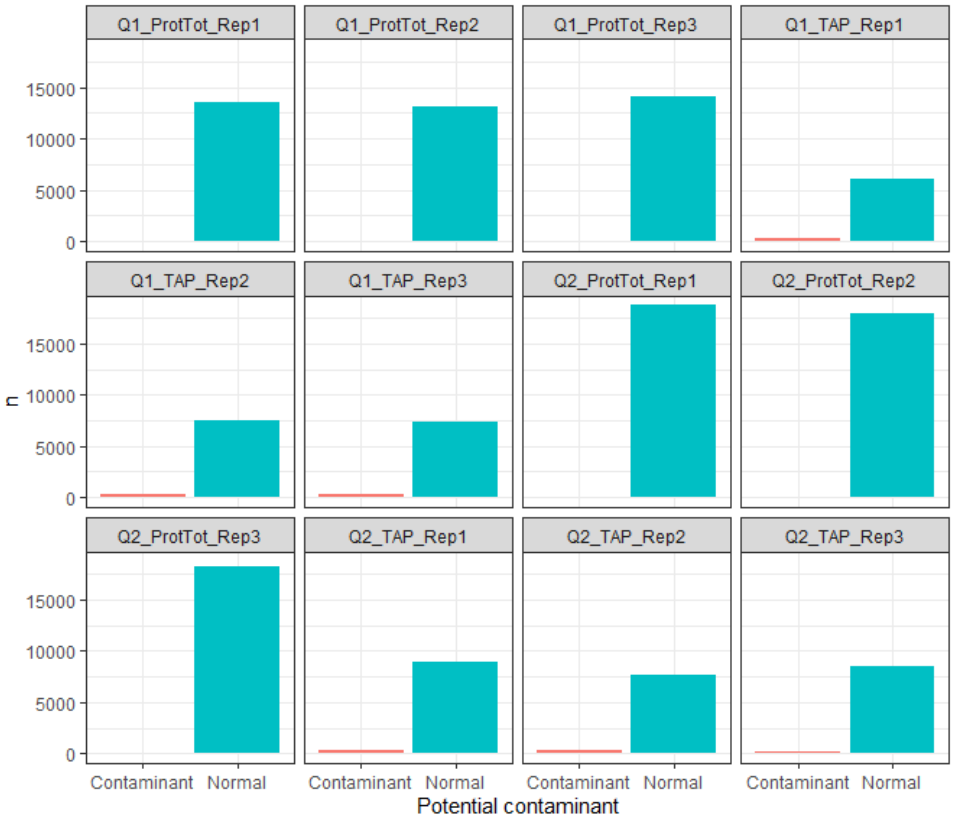
Incomplete Digestion

```
data |>
  ggplot(
    aes(
      x=`Missed cleavages`,
      fill=condition
    )
  ) +
  geom_bar() +
  coord_flip() +
  facet_wrap(vars(Experiment))
```

Total PSMs and Contaminants

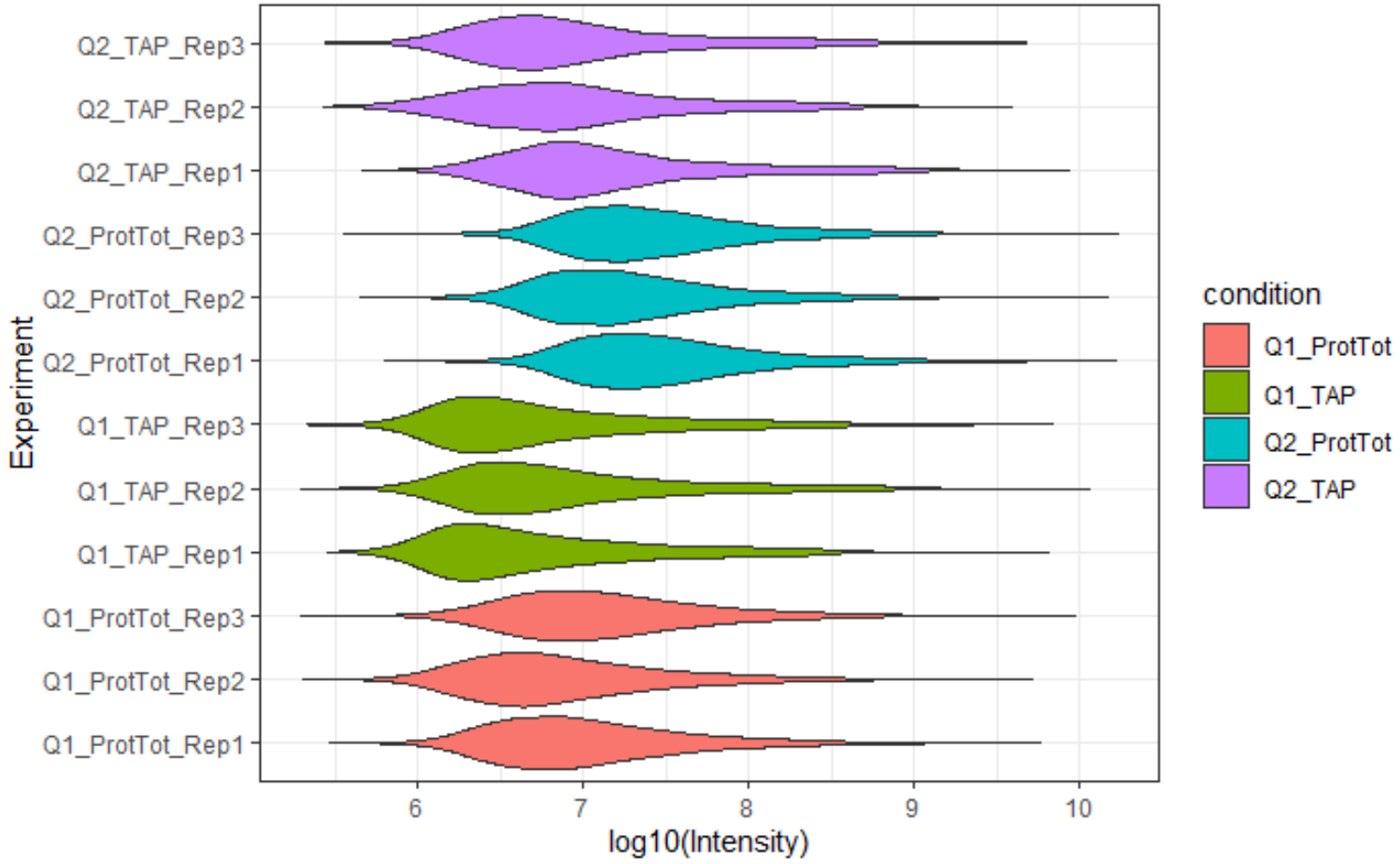
```
data |>
  mutate(
    `Potential contaminant` = if_else(
      is.na(`Potential contaminant`),
      "Normal",
      "Contaminant“
  )) |>
  group_by(
    Experiment,condition,`Potential contaminant`
  ) |>
  count() |>
  ungroup() |>
  ggplot(
    aes(
      x=`Potential contaminant`,
      y=n,
      fill=`Potential contaminant`
    )
  ) +
  geom_col(show.legend = FALSE) +
  facet_wrap(vars(Experiment))
```

Sequence <chr>	Potential contaminant <chr>	Length <dbl>	Modifications <chr>
ADLEMQIENLK	+	11	Oxidation (M)
ADLEMQIESLK	+	11	Oxidation (M)
ADLEMQIESLK	+	11	Oxidation (M)
ADLEMQIESLK	+	11	Oxidation (M)
AAAALAGGK	NA	9	Unmodified
AAAALAGGK	NA	9	Unmodified
AAAALAGGK	NA	9	Unmodified
AAAALAGGK	NA	9	Unmodified



Intensity Distributions

```
data |>
  ggplot(
    aes(
      x=Experiment,
      y=log10(Intensity),
      fill=condition
    )
  ) +
  geom_violin() +
  coord_flip()
```

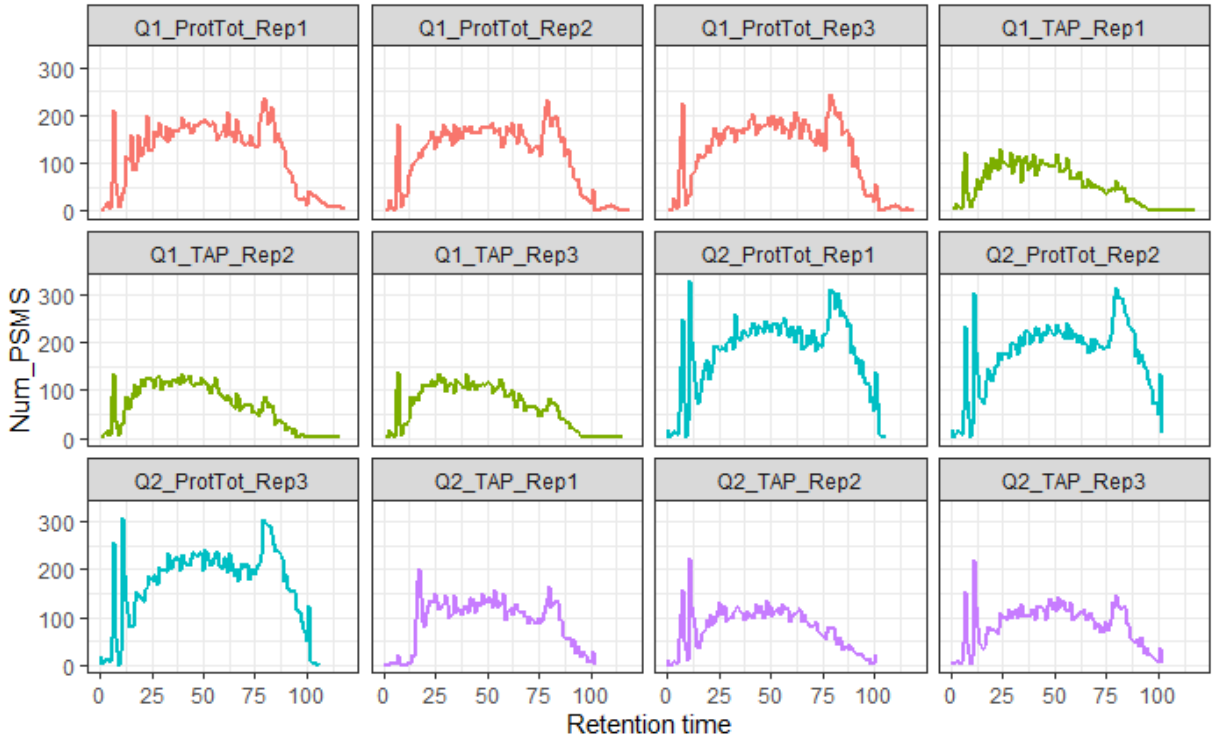


Check distribution shapes – check observed amounts are similar (or understand if not)

PSMs over LC Time

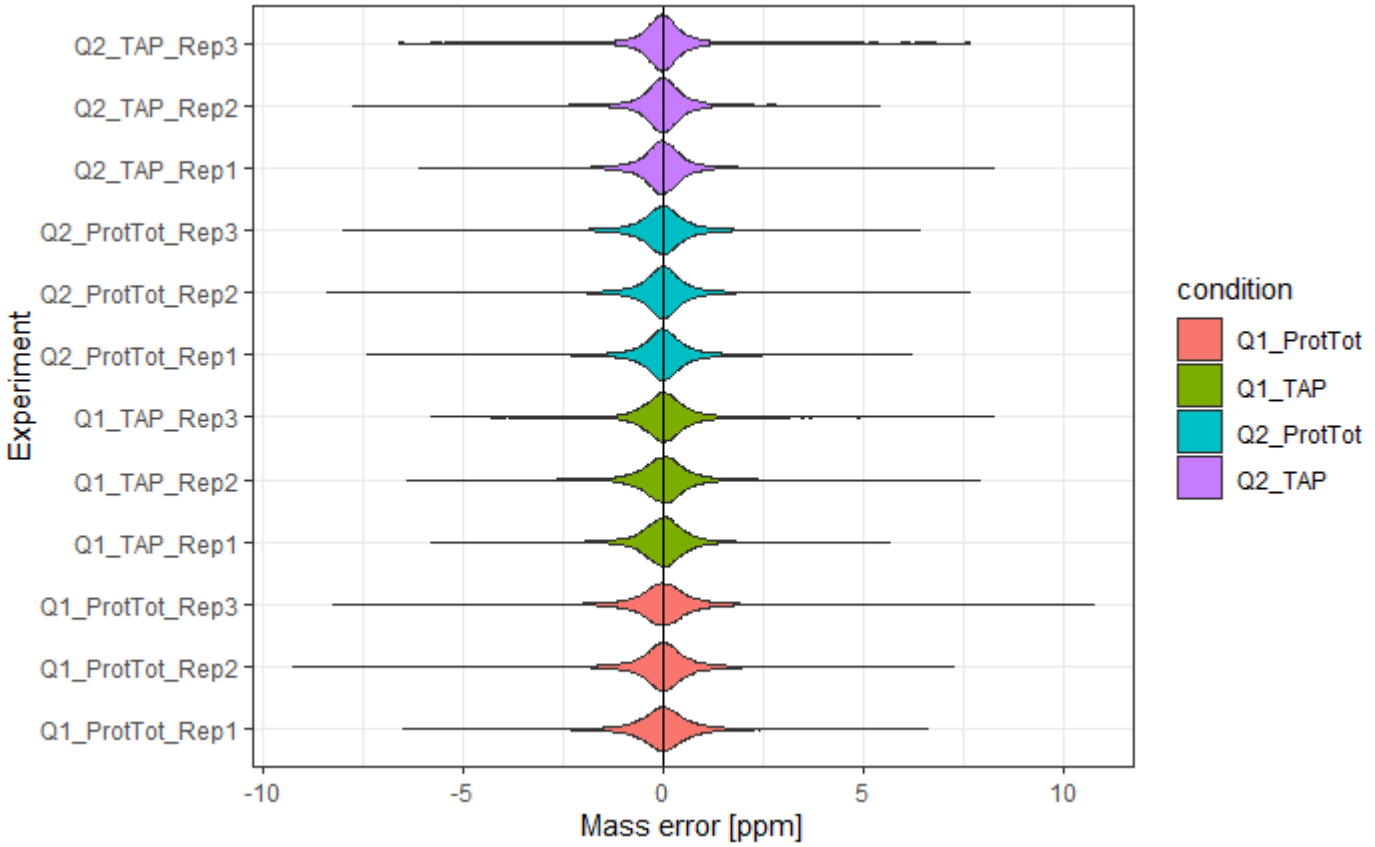
```
data |>
  mutate(
    `Retention time` = as.integer(`Retention time`)
  ) |>
  group_by(
    `Retention time`, Experiment, condition
  ) |>
  count(name="Num_PSMS") |>
  ggplot(
    aes(
      x=`Retention time`,
      y=Num_PSMS,
      colour=condition
    )
  ) +
  geom_line(show.legend = FALSE, linewidth=1) +
  facet_wrap(vars(Experiment))
```

Sequence <chr>	Retention time <dbl>	Intensity <dbl>
AAAALAGGK	8.7458	41189000
AAAALAGGK	8.7372	32882000
AAAALAGGK	8.7182	33508000
AAAALAGGK	8.7135	32531000
AAAALAGGK	9.0948	16542000
AAAALAGGKK	6.7033	39996000



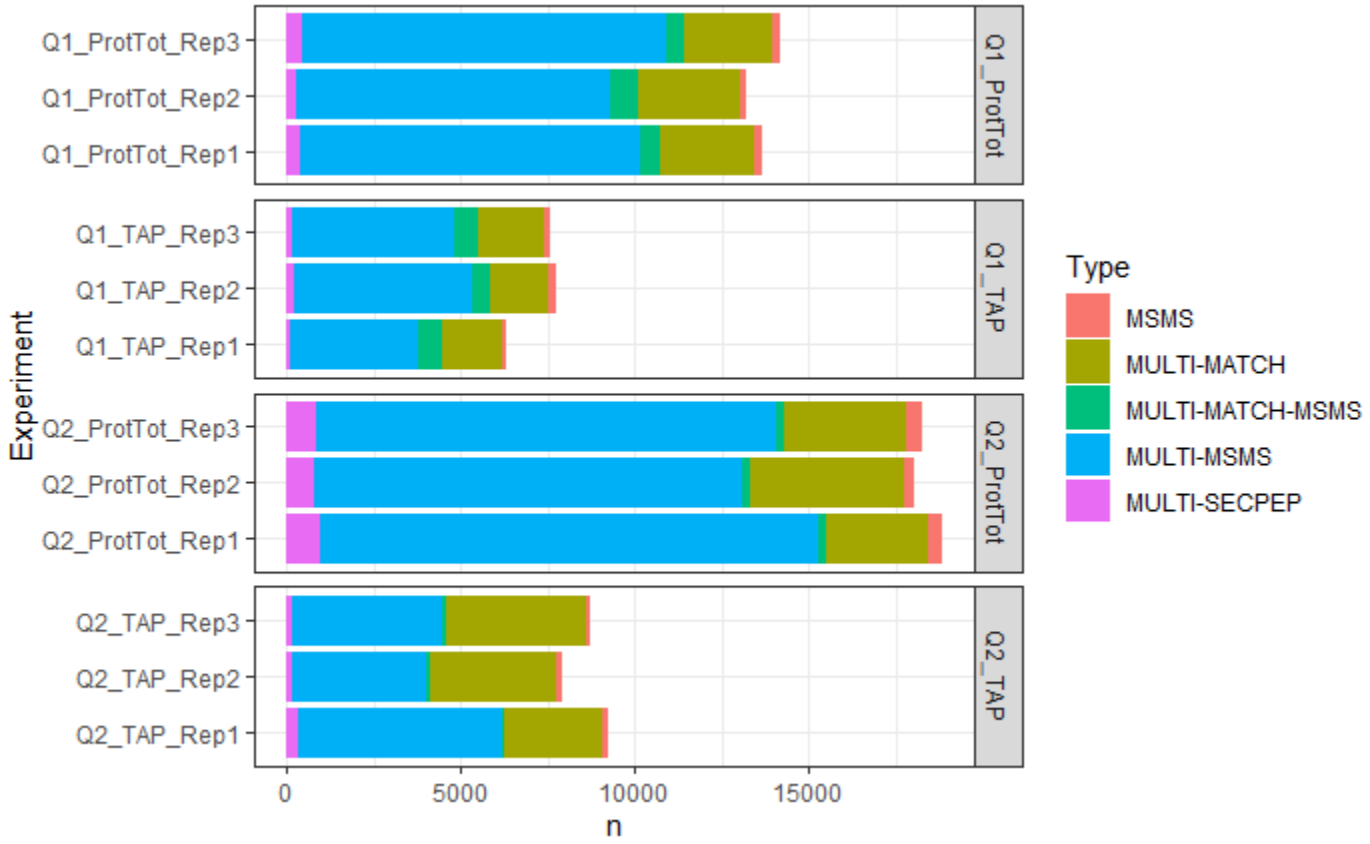
Mass Accuracy

```
data |>
  ggplot(
    aes(
      x=Experiment,
      y=`Mass error [ppm]`,
      fill=condition
    )
  ) +
  geom_violin() +
  geom_hline(yintercept = 0) +
  coord_flip()
```



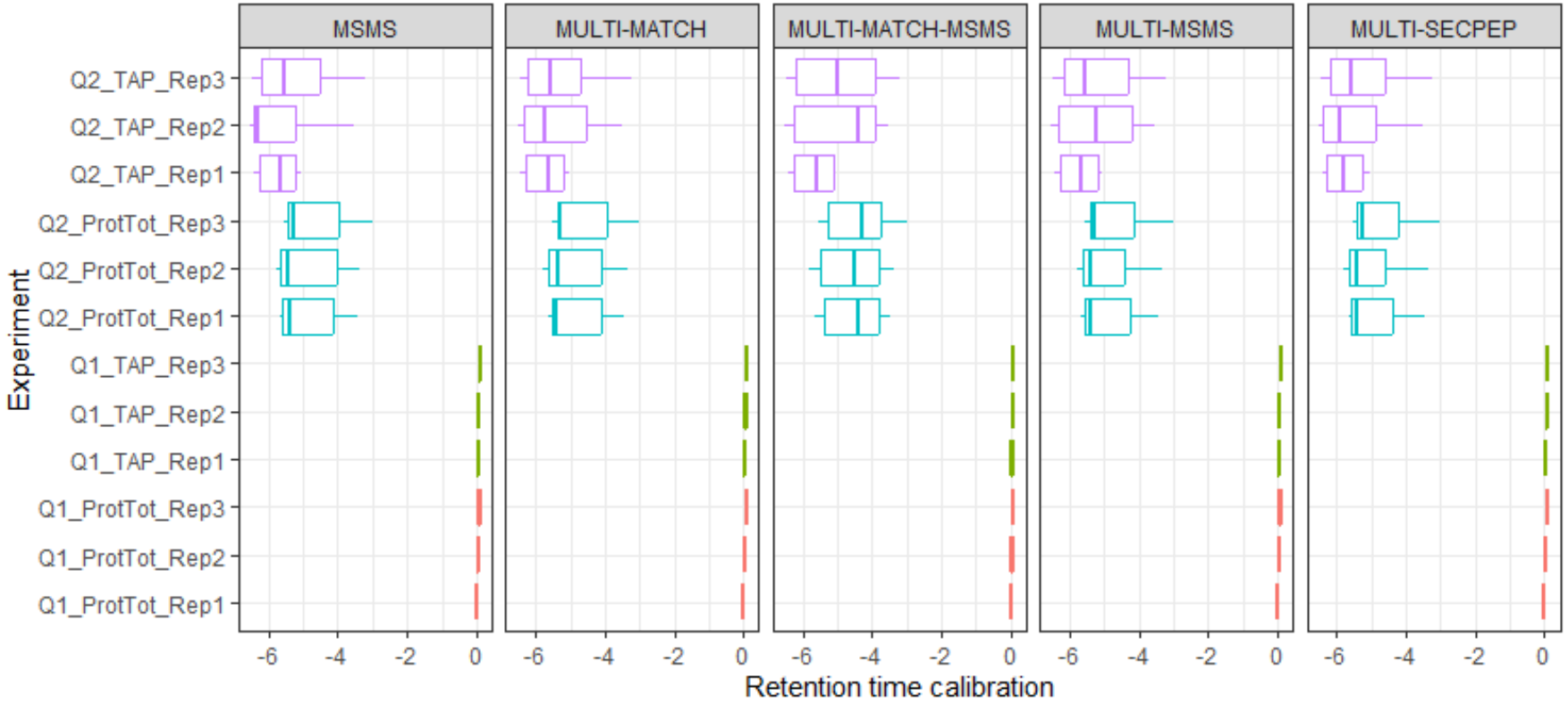
Measured vs Inferred PSMs

```
data |>
  group_by(Experiment, condition, Type) |>
  count() |>
  ggplot(
    aes(
      x=Experiment,
      y=n,
      fill=Type
    )
  ) +
  geom_col() +
  coord_flip() +
  facet_grid(
    rows=vars(condition),
    scale="free_y“
  )
```



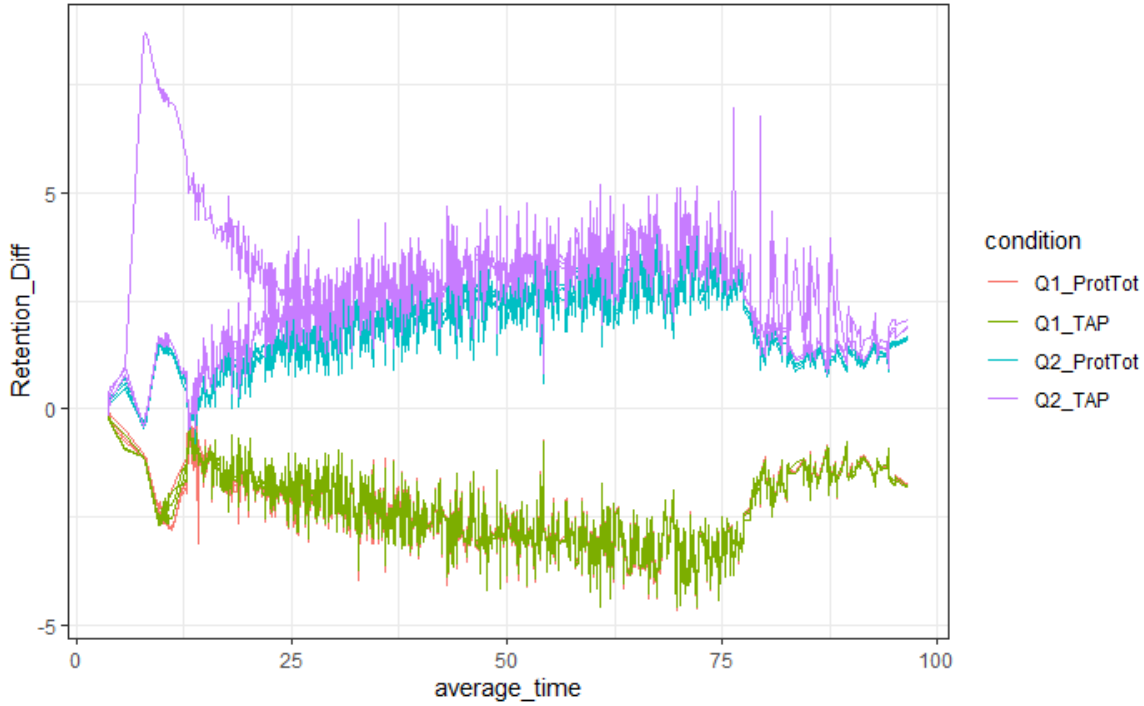
Retention Time Matching

```
data |>
  ggplot(
    aes(
      x=Experiment,
      y=`Retention time calibration`,
      colour=condition
    )
  ) +
  geom_boxplot() +
  coord_flip() +
  facet_grid(
    cols=vars(Type)
  )
```



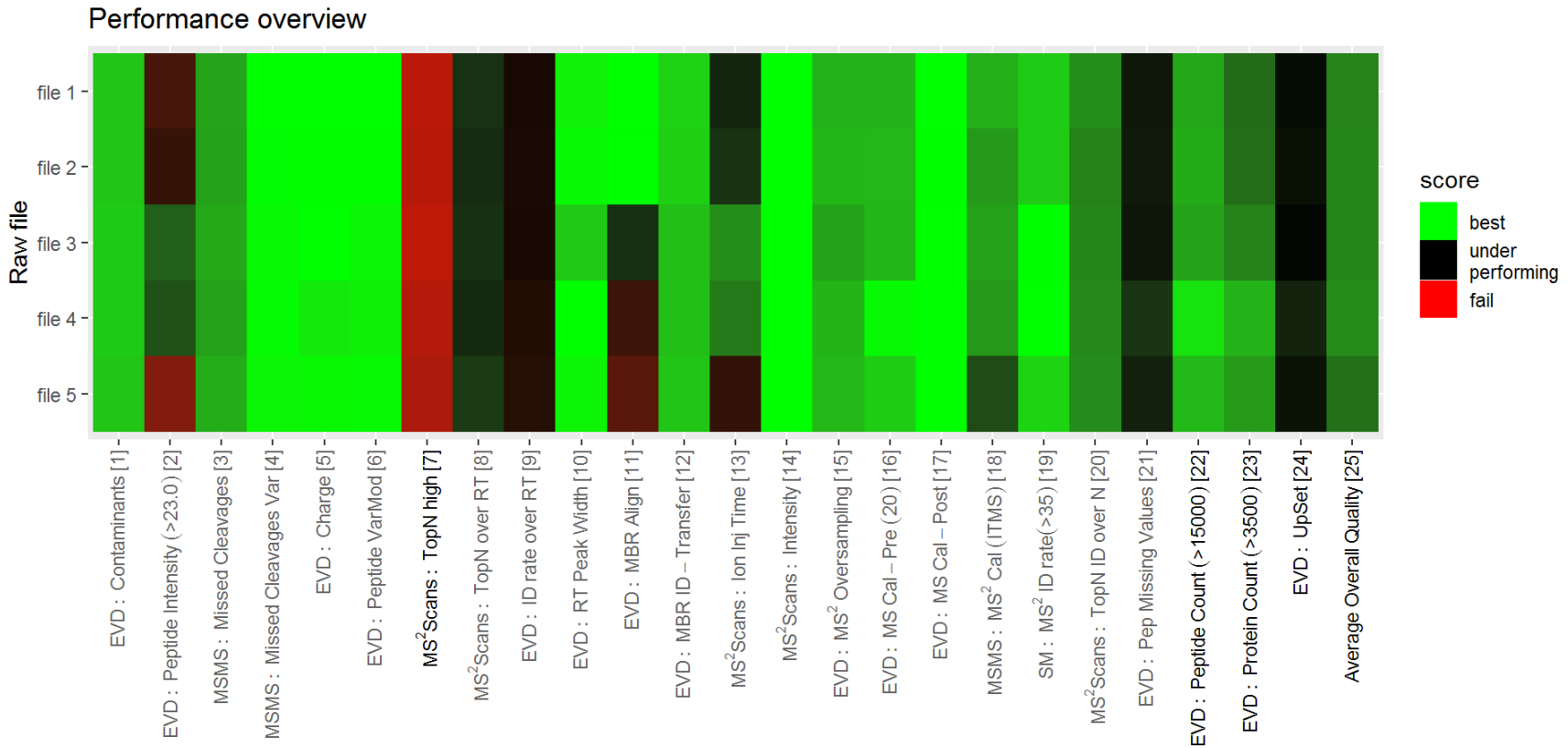
Retention Time Matching

```
data |>
  unite(PSM,Sequence,Modifications, Charge,sep=":") |>
  group_by(PSM) |>
  filter(n() == 12 & length(unique(Experiment))==12) |>
  group_by(PSM) |>
  mutate(
    average_time = mean(`Retention time`),
    Retention_Diff = `Retention time`-average_time
  )-> retention_times

retention_times |>
  ggplot(
    aes(
      x=average_time,
      y=Retention_Diff,
       group=Experiment,
       colour=condition
     )
  ) +  geom_line()
```

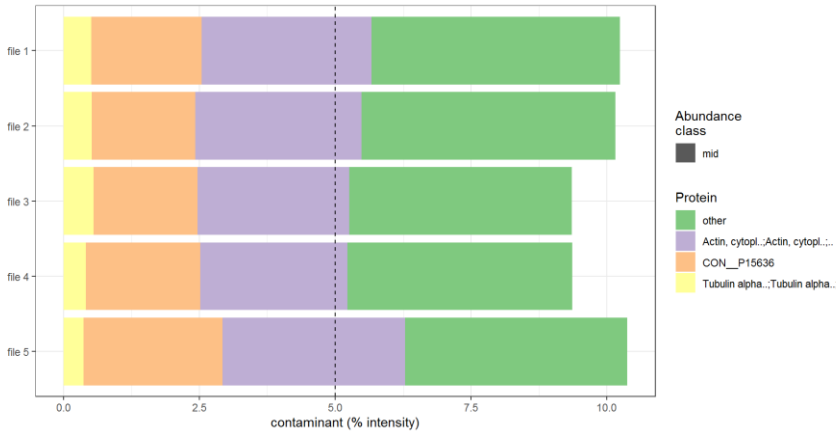


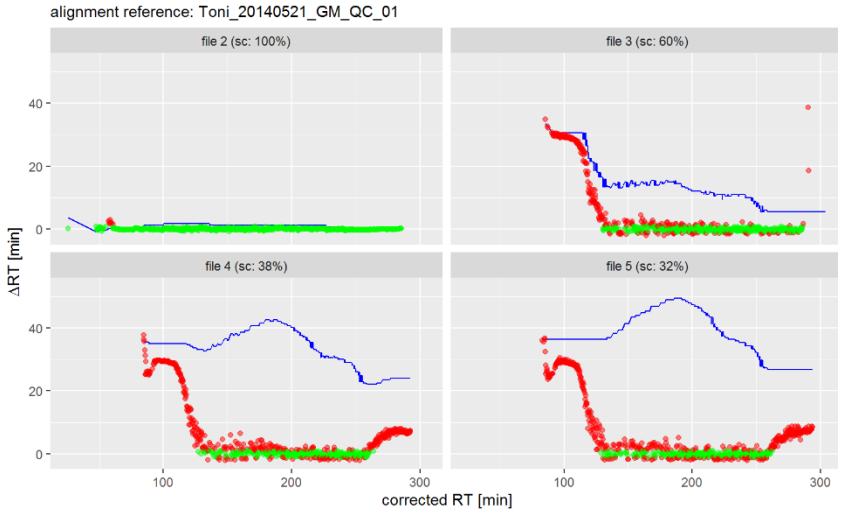
PTXQC

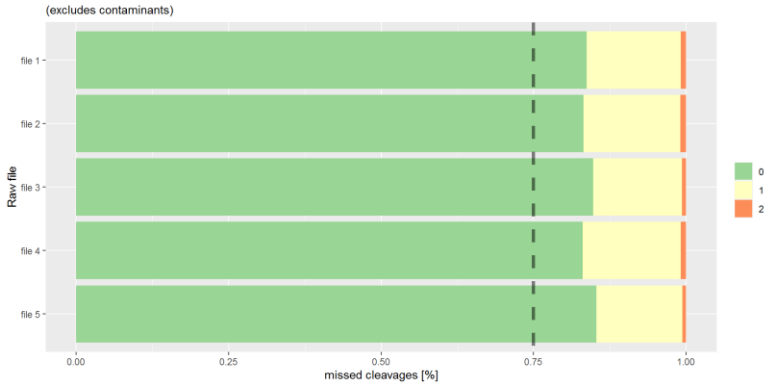
- R package – calculates a QC report from MQ or MzTab

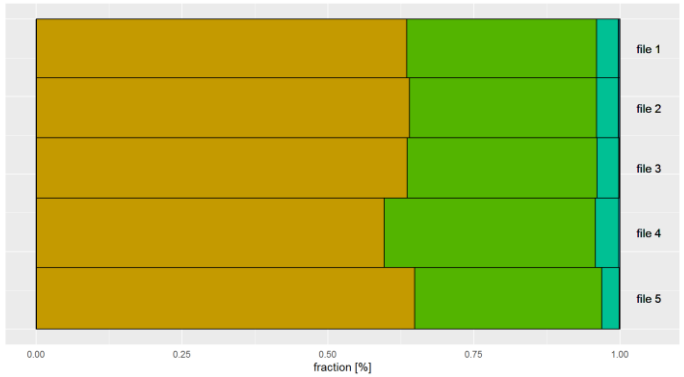


https://cran.rstudio.com/web/packages/PTXQC/index.html

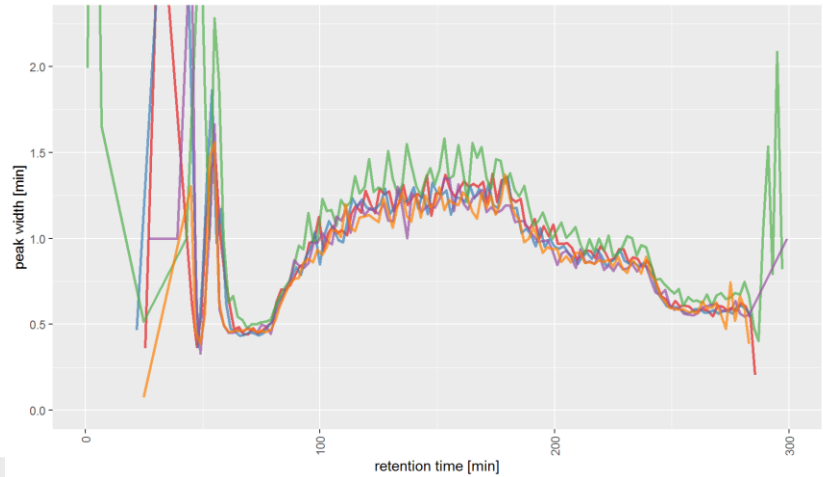

file 1
file 2
file 3
file 4
file 5
0.0
2.5
5.0
7.5
10.0
contaminant (% intensity)
Abundance class
mid
Protein
other
Actin, cytopl..;Actin, cytopl..;..
CON__P15636
Tubulin alpha..;Tubulin alpha..

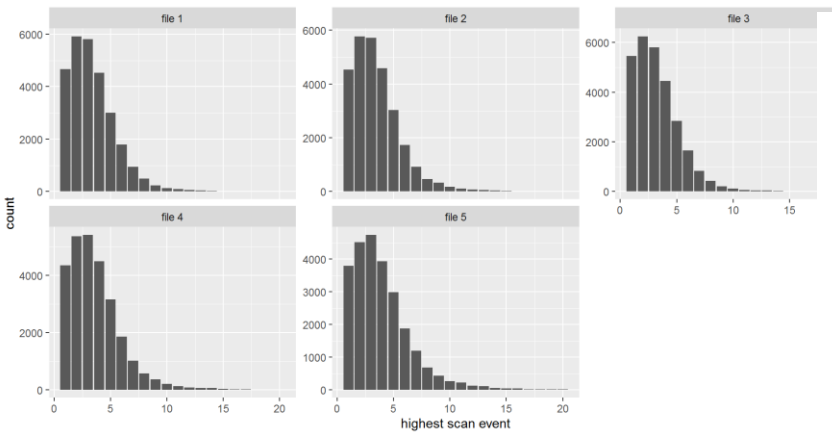

alignment reference: Toni_20140521_GM_QC_01
file 2 (sc: 100%)
file 3 (sc: 60%)
file 4 (sc: 38%)
file 5 (sc: 32%)
ΔRT [min]
corrected RT [min]

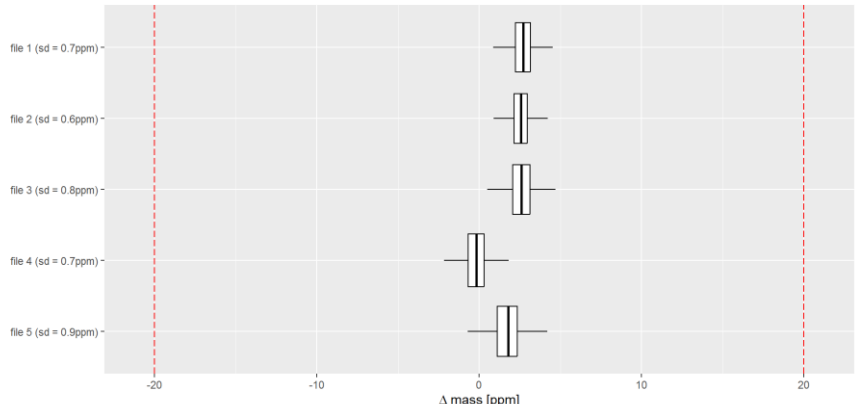

(excludes contaminants)
Raw file
file 1
file 2
file 3
file 4
file 5
0
1
2
missed cleavages [%]

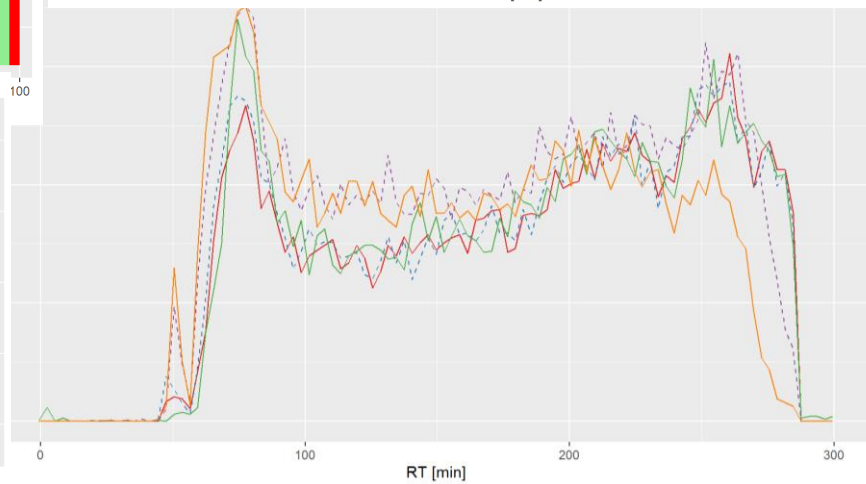

Raw file
file 1
file 2
file 3
file 4
file 5
charge
1
2
3
4
5
6
7
fraction [%]


genuine
transferred
all
file 1
file 2
file 3
file 4
file 5


peak width [min]
retention time [min]


file 1
file 2
file 3
file 4
file 5
count
highest scan event


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)
Δ mass [ppm]


RT [min]

Exercise

Quality Control of PSM data

Quantitation, Visualisation and Exploration

Babraham Bioinformatics

Loading Data (Protein Level)

```
read_delim("proteinGroups.txt", name_repair = "minimal") -> data
```

Pick out relevant Columns

```
data |>
  select(
    `Majority protein IDs`,
    `Gene names`,
    `Q-value`,
    `Reverse`,
    `Potential contaminant`,
    starts_with("LFQ Intensity ")
  ) -> data
```

Restructure and Add Condition information

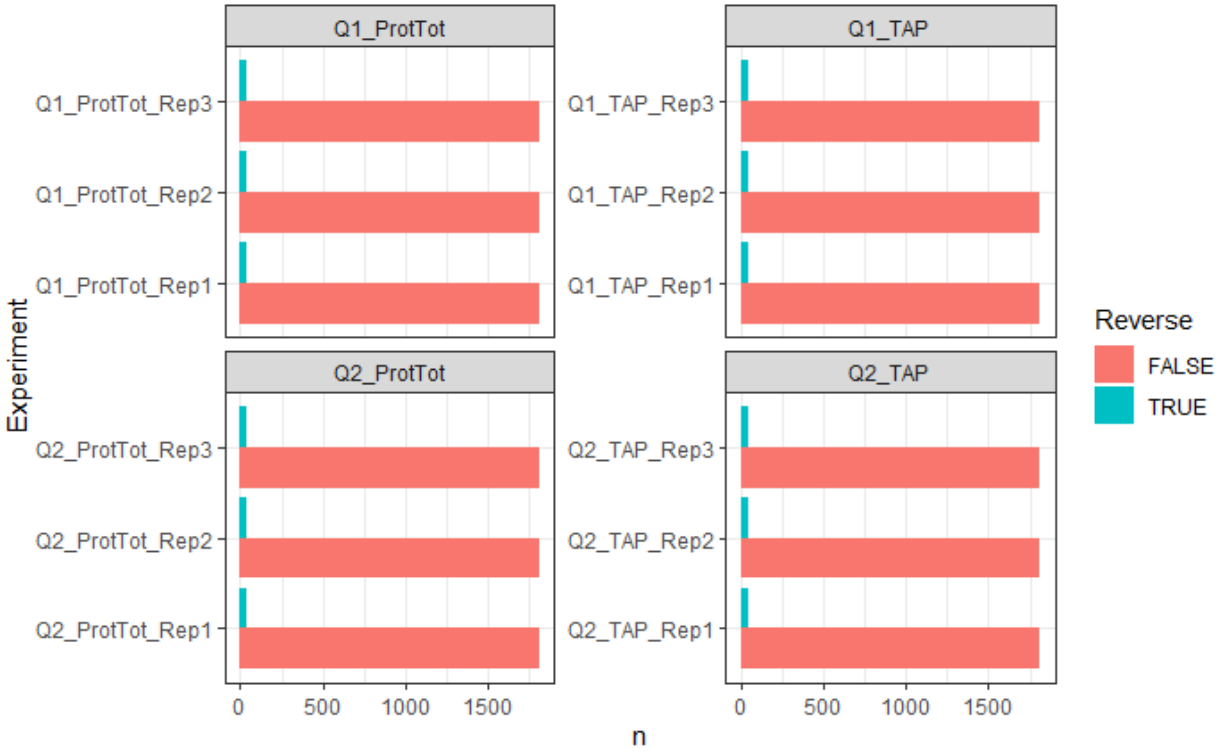
```
data |>
  pivot_longer(
    cols=starts_with("LFQ intensity"),
    names_to="Experiment",
    values_to="Intensity“
  ) |>
  mutate (
    Experiment = str_replace(Experiment, "LFQ intensity ","")
  ) |>
  mutate(
    Condition = str_replace(Experiment,"_Rep.*","")
  ) |>
  select(Experiment, Condition, everything()) -> data
```

Protein Level Data

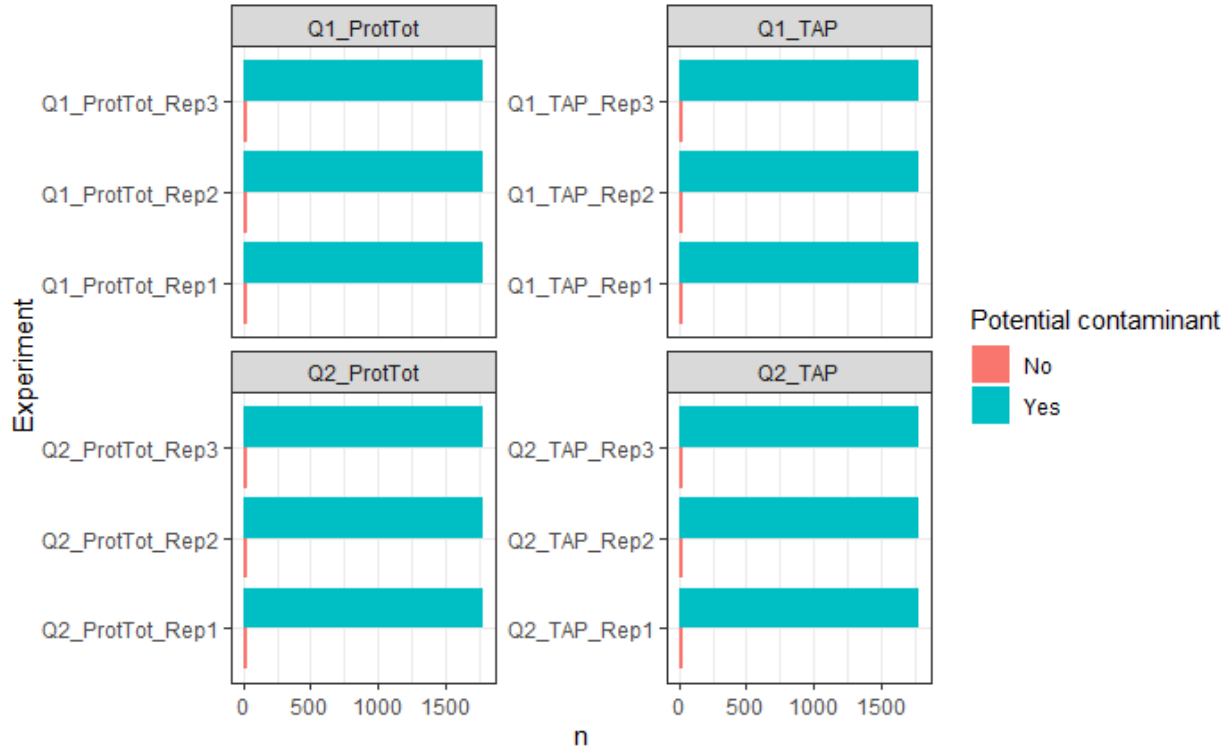
Experiment	Condition	Majority protein IDs	Gene names	Q-value	Reverse	Potential contaminant	Intensity
Q1_TAP_Rep1	Q1_TAP	P39077	CCT3	0.0000000	NA	NA	1.6059e+10
Q1_ProtTot_Rep3	Q1_ProtTot	P00925	ENO2	0.0000000	NA	NA	7.4979e+09
Q2_TAP_Rep1	Q2_TAP	P22336	RFA1	0.0000000	NA	NA	7.0751e+09
Q1_ProtTot_Rep1	Q1_ProtTot	P00925	ENO2	0.0000000	NA	NA	6.9097e+09
Q2_TAP_Rep1	Q2_TAP	P53235	YGR054W	0.0000000	NA	NA	6.6438e+09
Q1_ProtTot_Rep2	Q1_ProtTot	P00925	ENO2	0.0000000	NA	NA	6.5491e+09
Q2_TAP_Rep3	Q2_TAP	P53235	YGR054W	0.0000000	NA	NA	6.2282e+09
Q1_TAP_Rep3	Q1_TAP	P53235	YGR054W	0.0000000	NA	NA	6.2250e+09
Q1_TAP_Rep1	Q1_TAP	P53235	YGR054W	0.0000000	NA	NA	6.1224e+09
Q1_TAP_Rep3	Q1_TAP	CON__P00761	NA	0.0000000	NA	+	5.6815e+09
Q1_TAP_Rep2	Q1_TAP	P53235	YGR054W	0.0000000	NA	NA	5.5042e+09
Q1_ProtTot_Rep1	Q1_ProtTot	P00359	TDH3	0.0000000	NA	NA	5.4535e+09
Q2_ProtTot_Rep3	Q2_ProtTot	P00359	TDH3	0.0000000	NA	NA	5.4059e+09
Q2_ProtTot_Rep2	Q2_ProtTot	P39077	CCT3	0.0000000	NA	NA	5.3267e+09

Re-Check QC at Protein Level

Reverse Hits

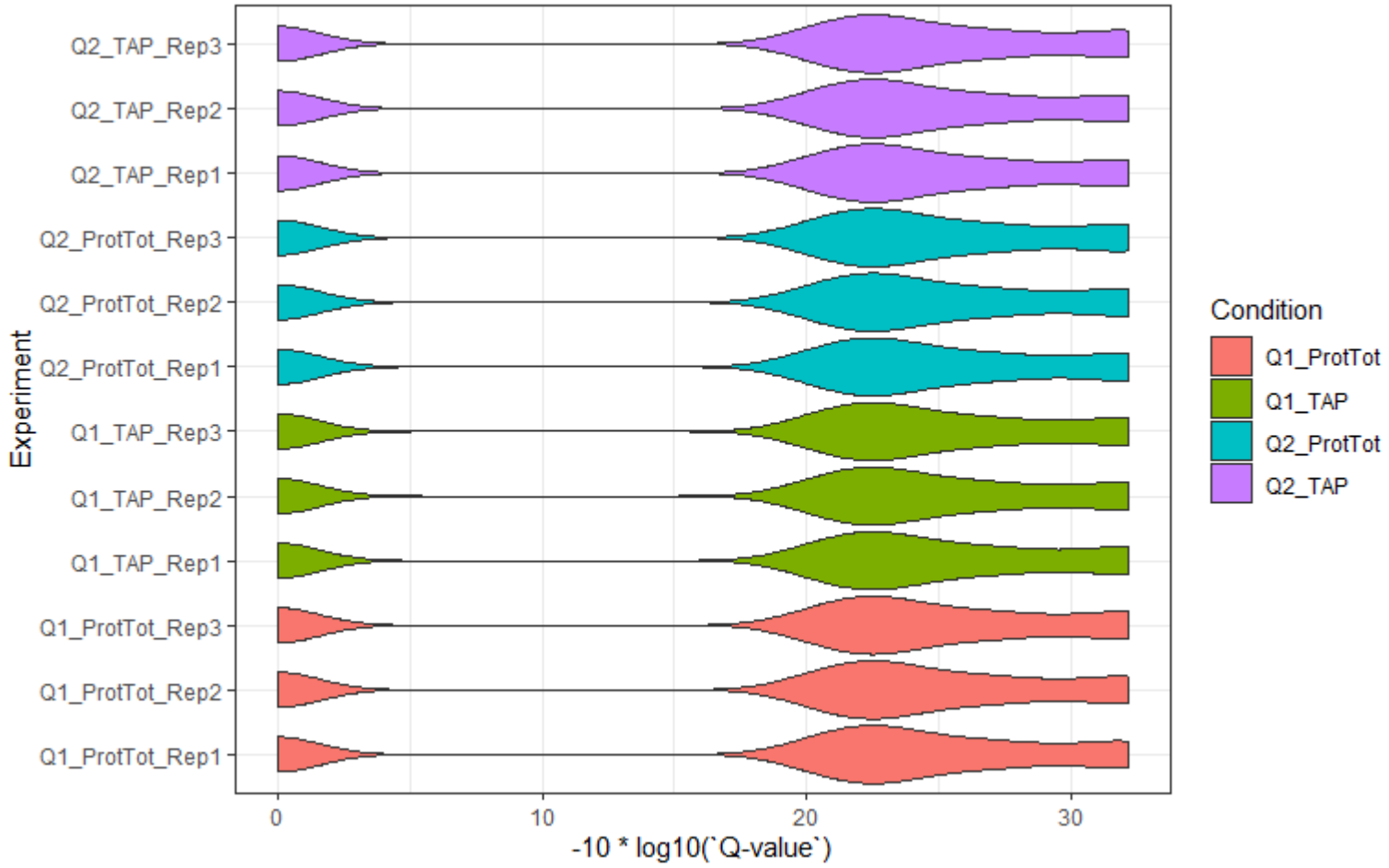


Contaminant Hits



Re-Check QC at Protein Level

Hit Qualities

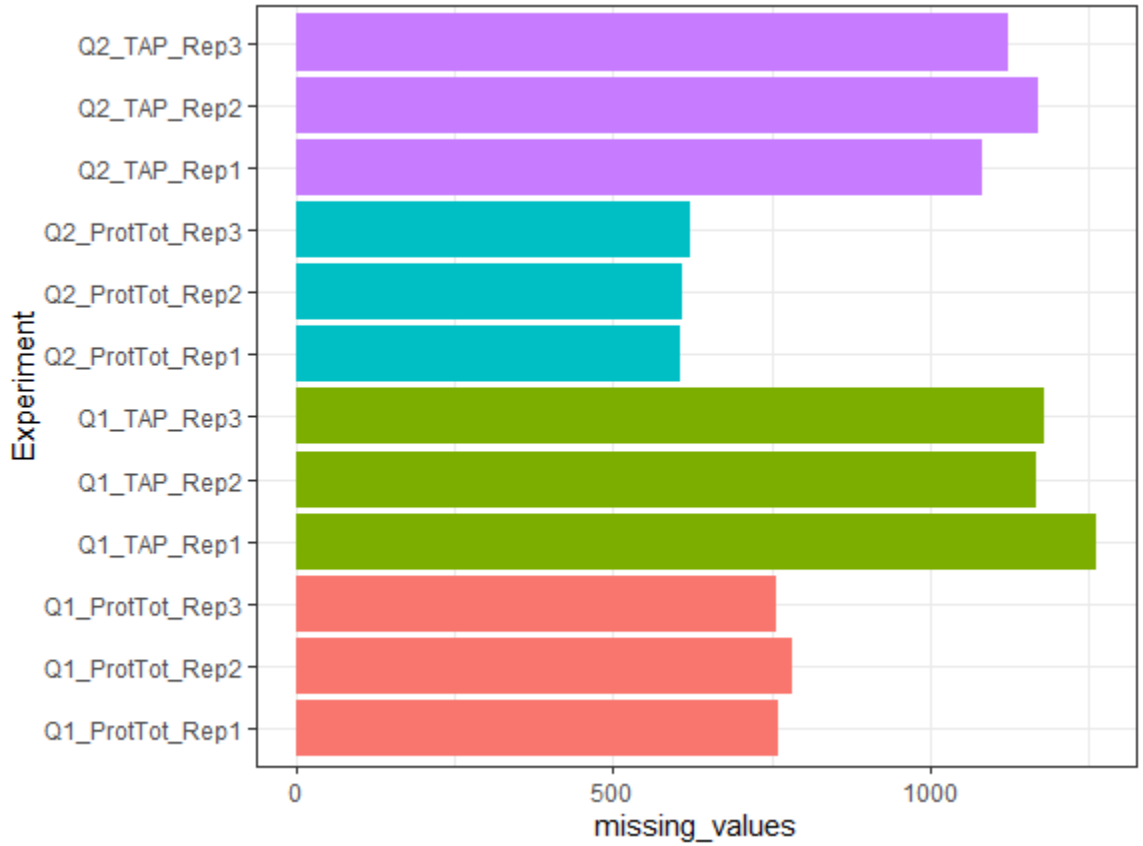


Missing Values

- Very likely in LFQ experiments
- Encoded differently in different file formats
  - Could be numerical zeroes
  - Could be missing (NA) values
  - Could just not be reported (not present in file)
- Different reasons
  - Missing at random (MAR) – technical failures
  - Missing not at random (MNAR) – low levels – missed observations

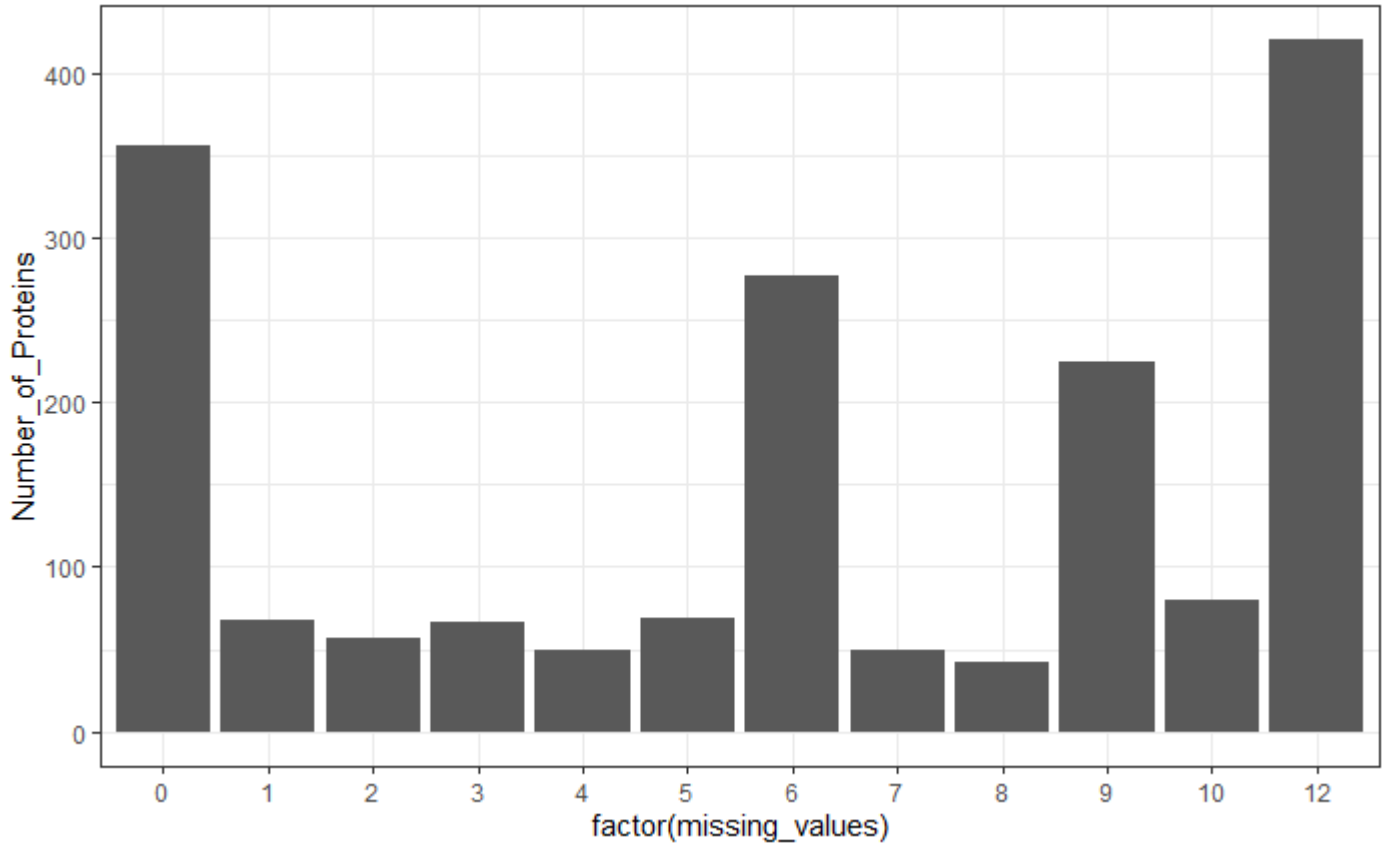
Missing Values

```
data |>
  group_by(Experiment,Condition) |>
  summarise(
    missing_values = sum(Intensity == 0)
  ) |>
  ungroup() |>
  ggplot(
    aes(
      x=Experiment,
      y=missing_values,
      fill=Condition
    )
  ) +  geom_col() +  coord_flip()
```



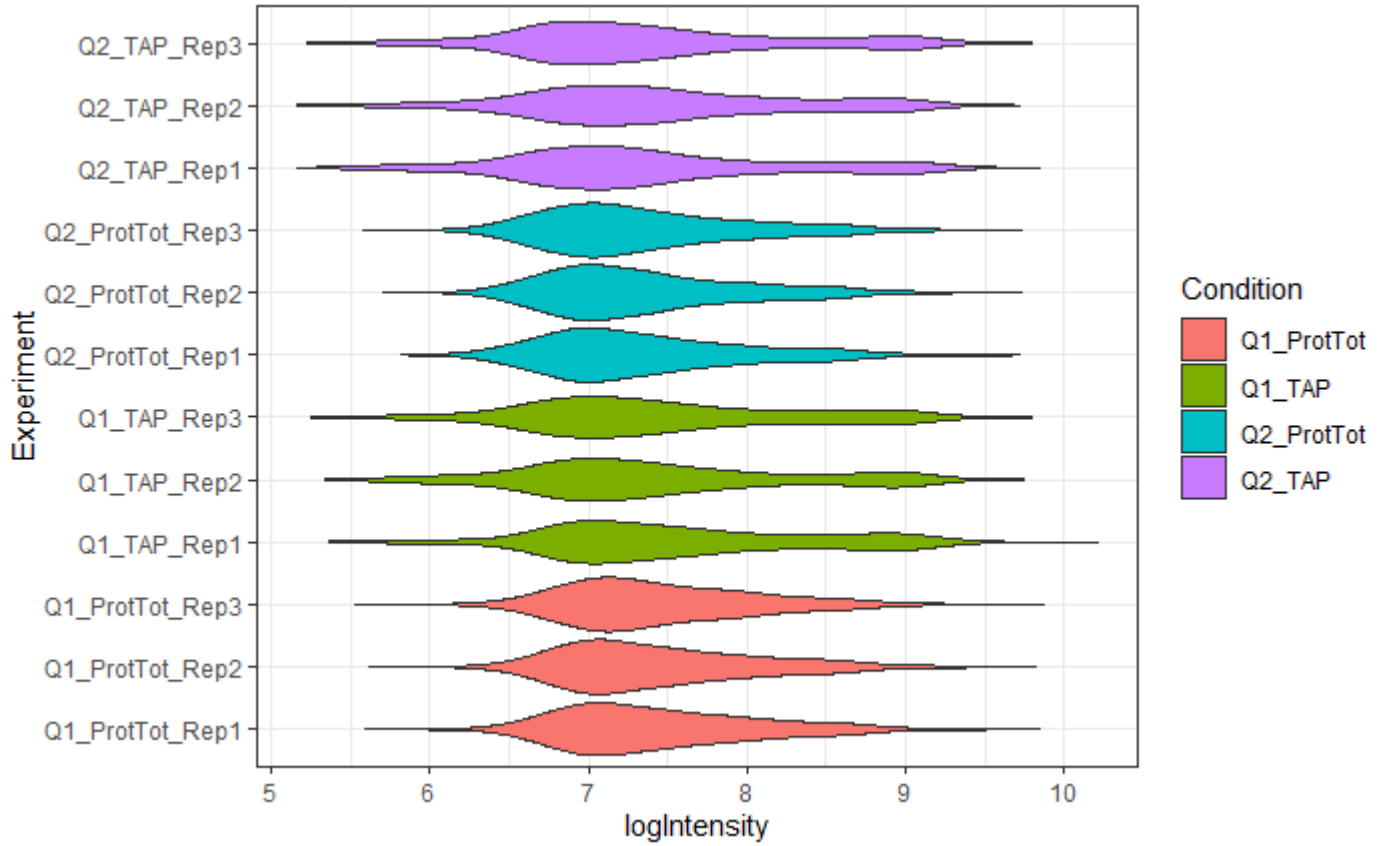
Missing Value Patterns Across Samples

```
data |>
  group_by(`Majority protein IDs`) |>
  summarise(
    missing_values = sum(Intensity == 0)
  ) |>
  ungroup() |>
  group_by(missing_values) |>
  count(name="Number_of_Proteins") |>
  ungroup() |>
  ggplot(
    aes(
      x=factor(missing_values),
      y=Number_of_Proteins
    )
  ) +  geom_col()
```



Intensity Distribution, Normalisation, Imputation

```
data |>
  ggplot(
    aes(
      x=Experiment,
      y=logIntensity,
      fill=Condition
    )
  ) +
  geom_violin() +
  coord_flip()
```



Summarisation will also add normalisation – should be more closely matched than PSMs

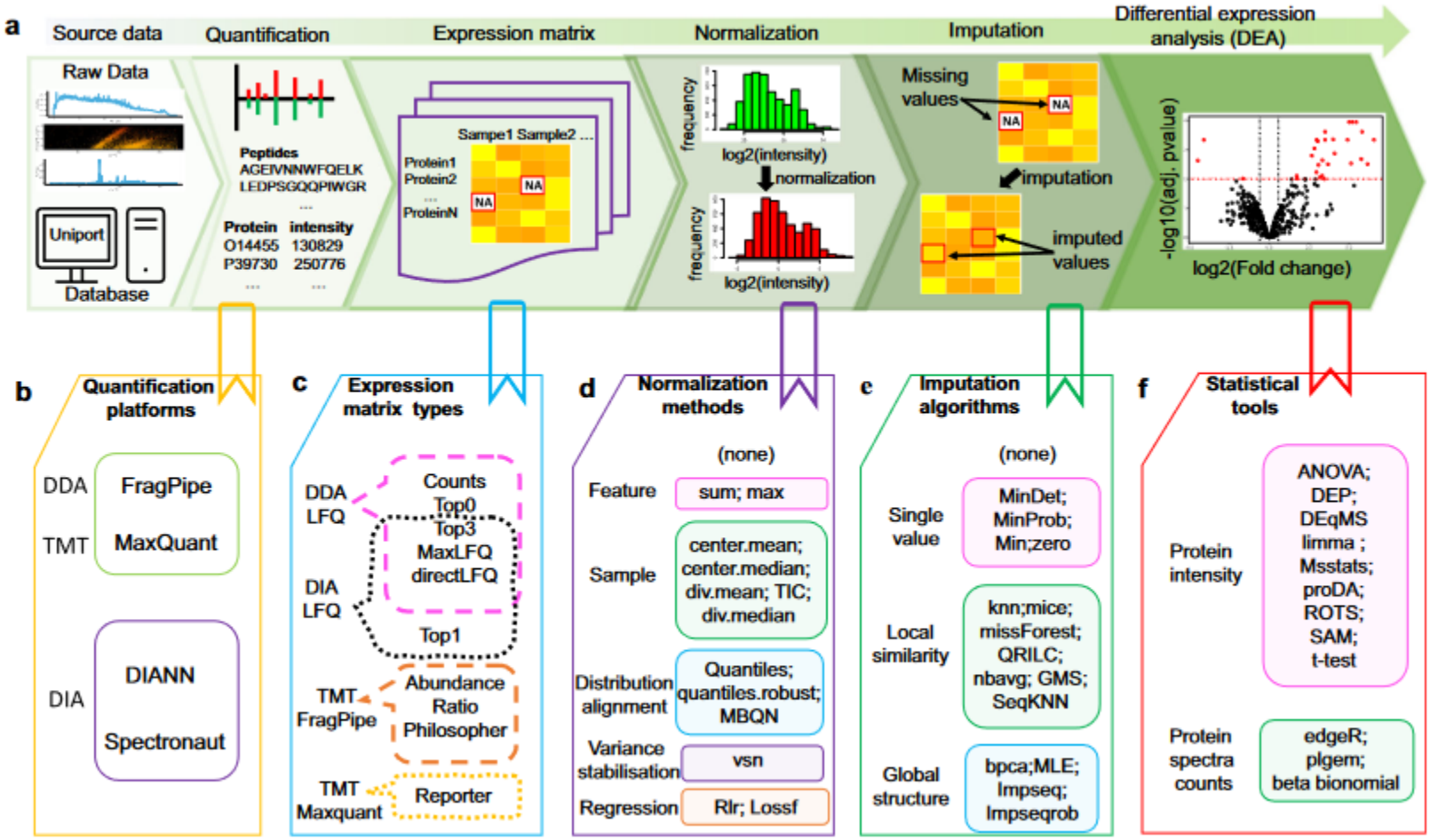
nature communications

Article

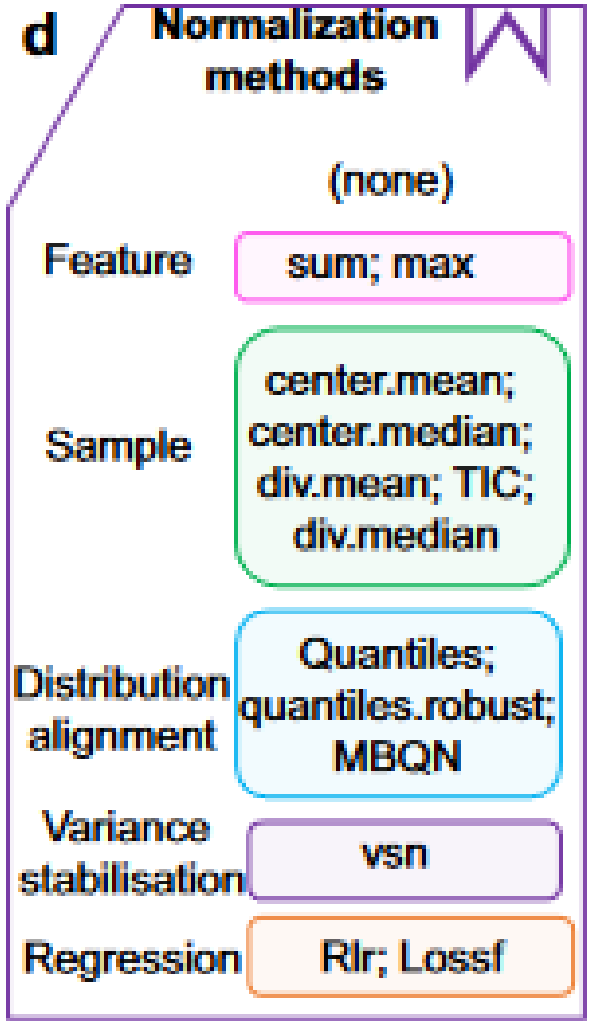
https://doi.org/10.1038/s41467-024-47899-w

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

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

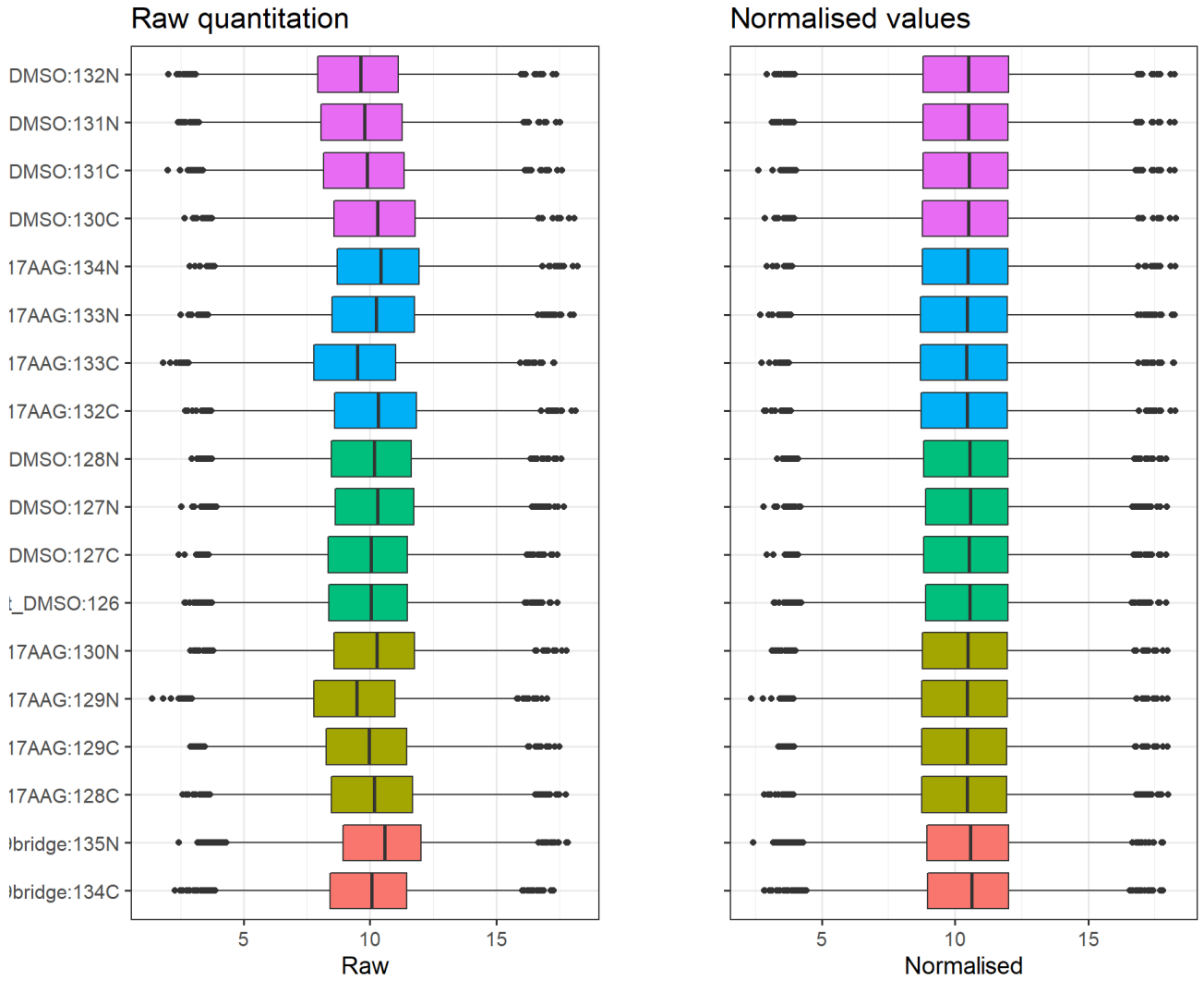


Normalisation



- Using original upstream values is often optimal
- Some methods susceptible to outliers
- Some methods break with skewed data
- Some methods aim to adjust variance

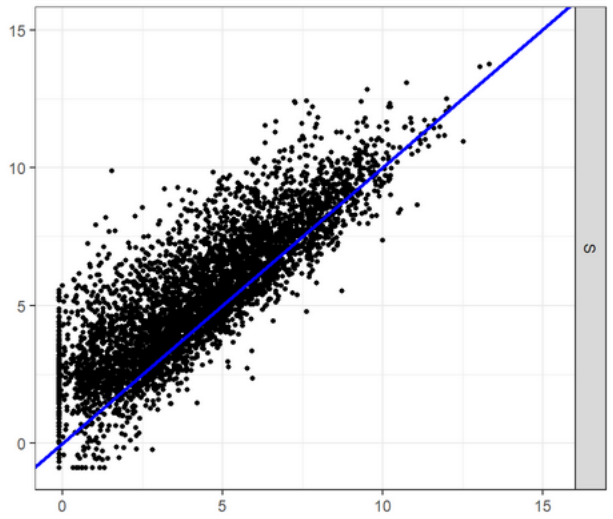
Median Centering

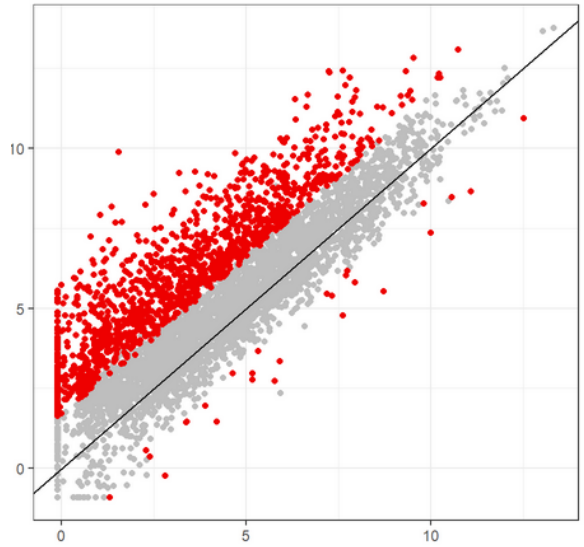



```
data |>
  group_by(Experiment) |>
  mutate(
    norm = logInt - median(logInt)
  ) -> data
```

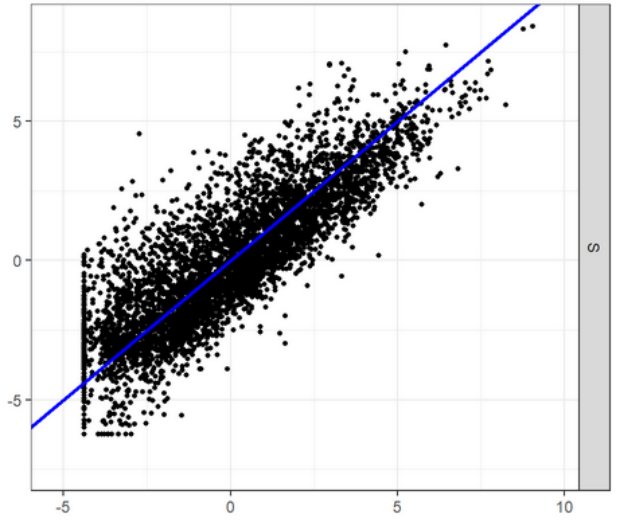
Normalisation Methods

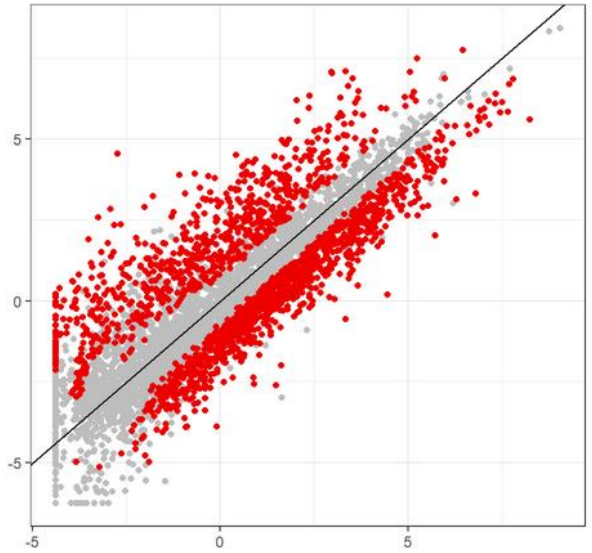
Original Quantitation



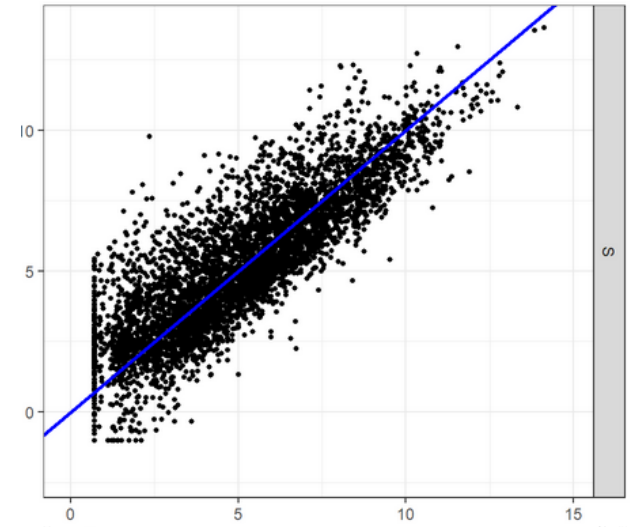


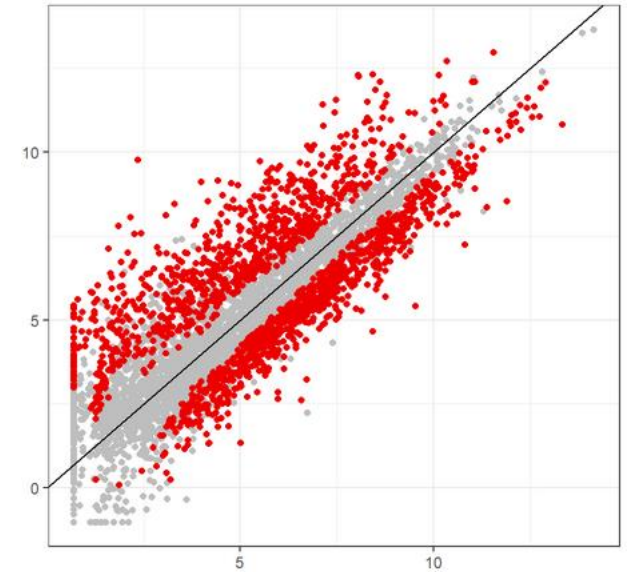
Global (Mean) Normalisation





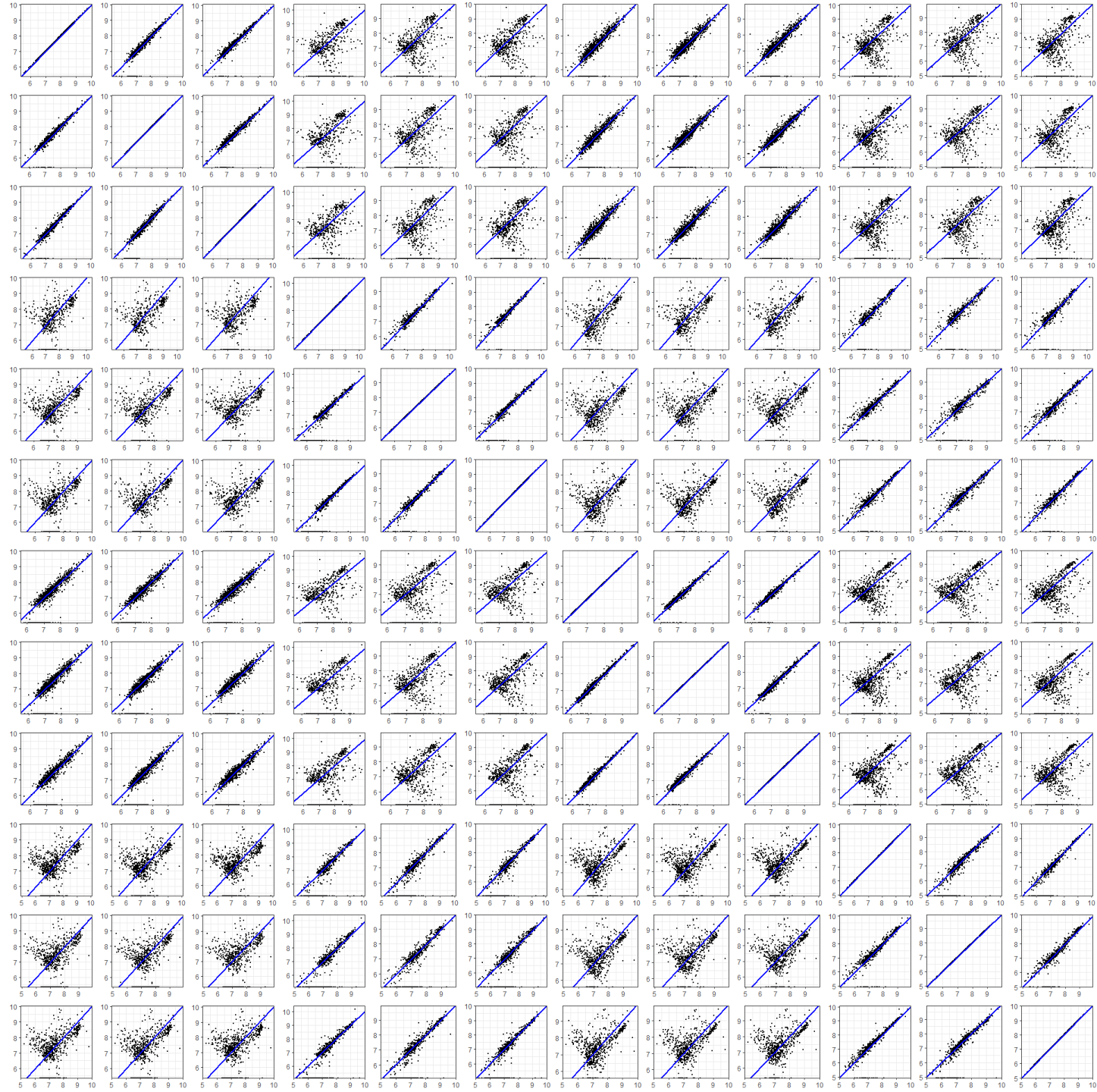
Size Factor (Median) Normalisation





All vs All

- Distribution plot is not always helpful
- Unusual designs (especially enrichment)
- Scatterplots are always more informative

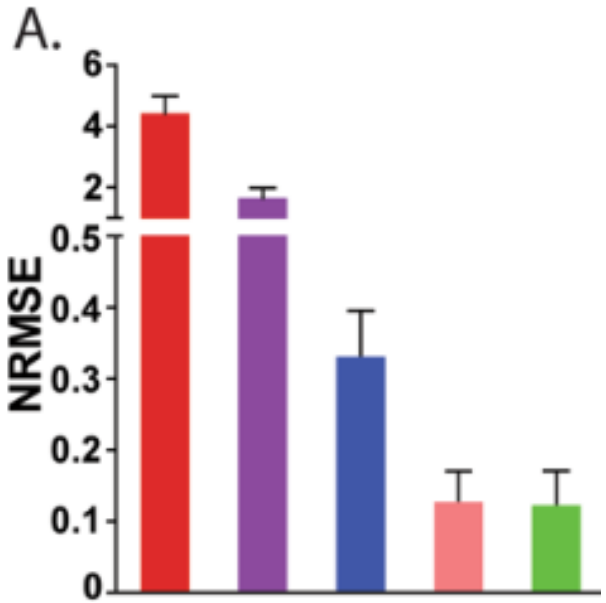


Imputation Methods

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]✉

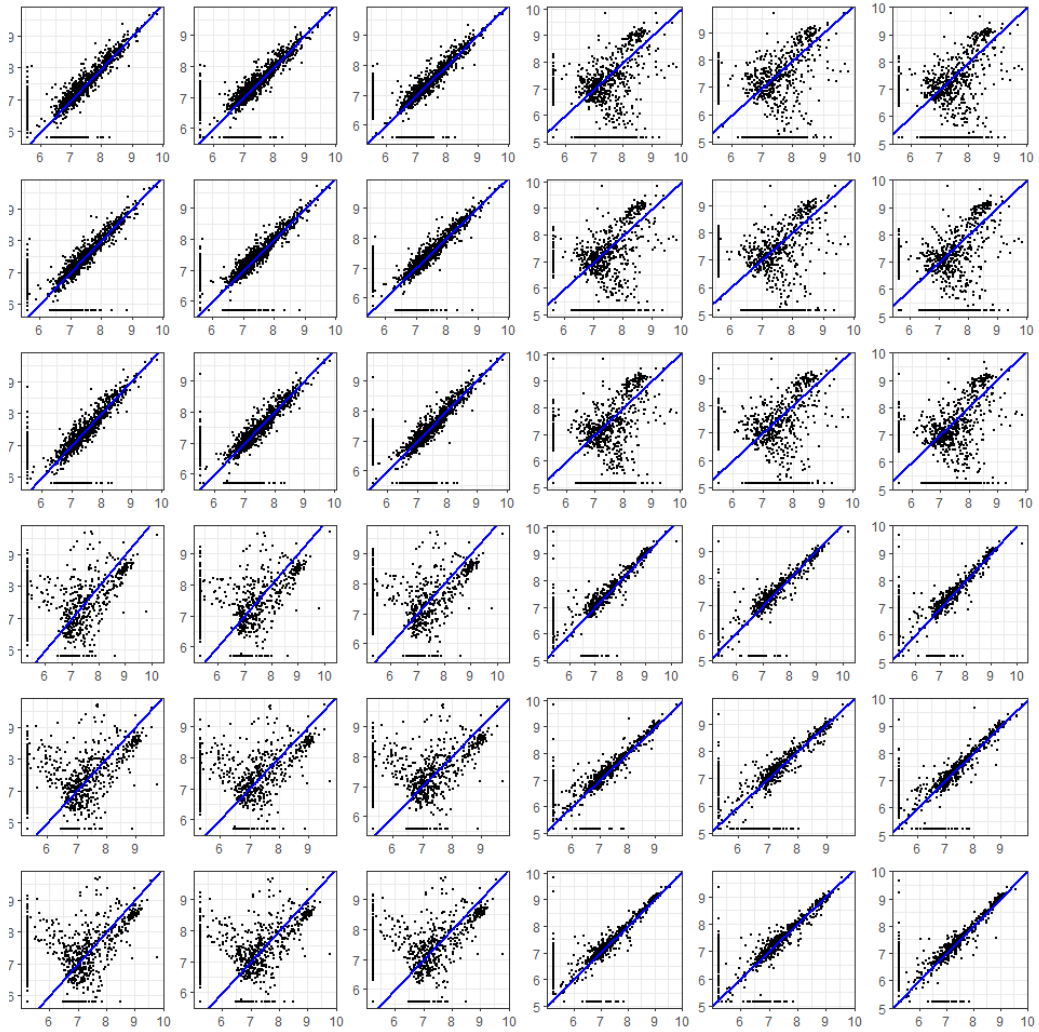


- Lowest observation
- Random Drawing
- K nearest neighbours
- Random forest

https://www.nature.com/articles/s41598-021-81279-4

Lowest Value Imputation

```
data |>
  group_by(Experiment) |>
  mutate(
    Intensity2 = replace(
      Intensity,
      Intensity==0,
      min(Intensity[Intensity != 0])
    )
  ) |>
  mutate(
    logIntensity = log10(Intensity2)
  ) -> data_simpleimpute
```



Random Forest Imputation

missForest

missForest is a nonparametric, mixed-type imputation method for basically any type of data. Here, we host the R-package "missForest" for the statistical software R.

The method is based on the publication Stekhoven and Bühlmann, 2012. The R package contains a vignette on how to use "missForest" in R including many helpful examples.

	Q1_ProtTot_Rep1 <dbl>	Q1_ProtTot_Rep2 <dbl>	Q1_ProtTot_Rep3 <dbl>	Q1_TAP_Rep1 <dbl>	Q1_TAP_Rep2 <dbl>	Q1_TAP_Rep3 <dbl>
O13563	19323000	21091000	17655000	*NA*	2089500	4420100
O14455	200660000	172160000	145660000	7.0774e+08	741060000	657080000
O14467	60694000	58253000	53757000	2.0570e+07	25402000	20104000
O43137	8525700	9944700	9821800	*NA*	*NA*	*NA*
P00330	2082600000	2132800000	2065800000	4.7175e+07	39745000	46975000
P00358	183300000	223790000	298120000	*NA*	*NA*	*NA*
P00359	5453500000	5119500000	5048400000	5.2562e+07	52270000	61381000
P00360	39415000	26623000	31036000	*NA*	*NA*	6527600
P00431	5703400	6166200	5406600	*NA*	*NA*	*NA*
P00445	257900000	240840000	224840000	*NA*	*NA*	*NA*

- Missing values NA (not zero)
- One column per sample (wide not long)
- Protein IDs as data frame rownames

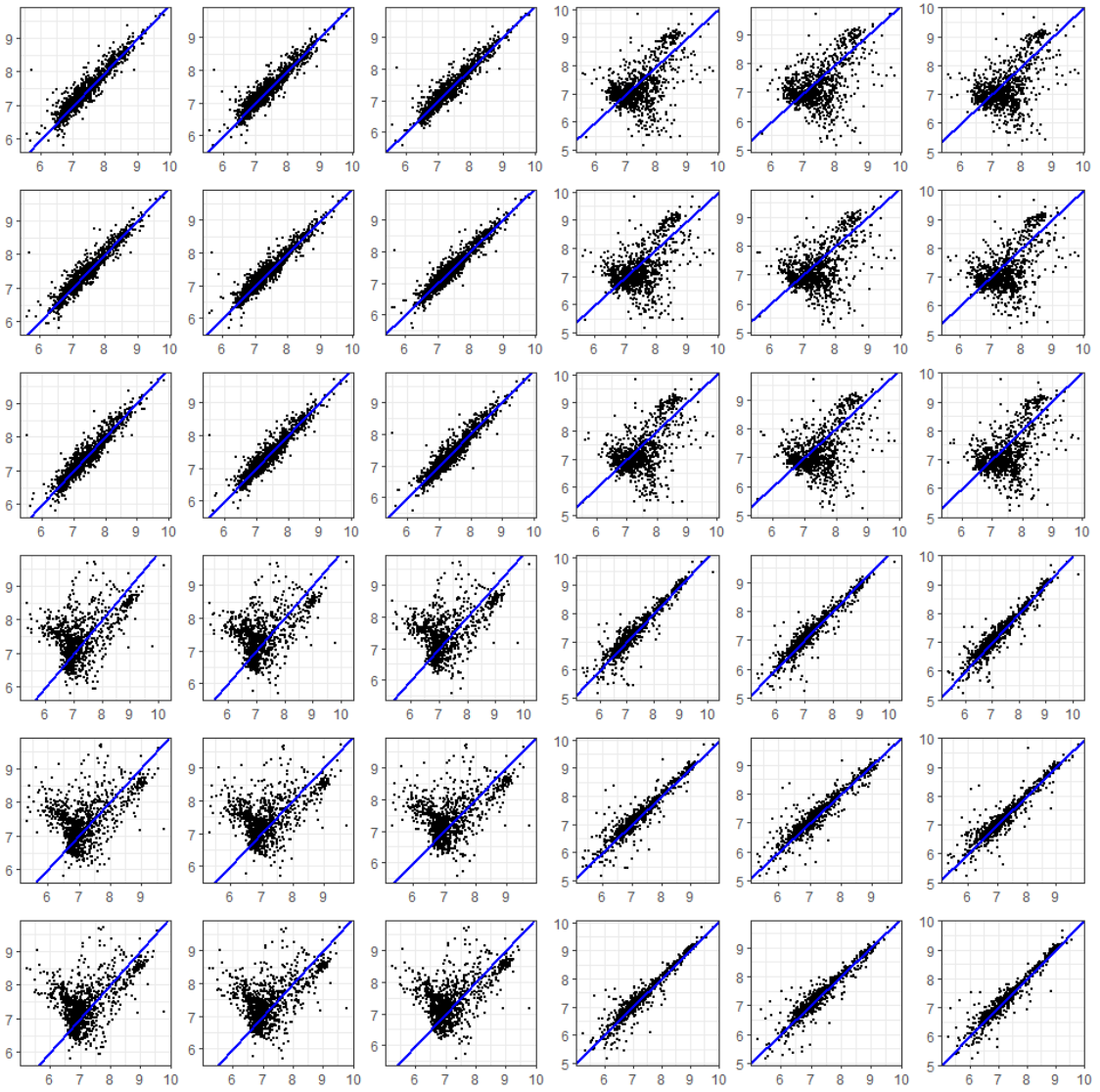
Random Forest Imputation

missForest

missForest is a nonparametric, mixed-type imputation method for basically any type of data. Here, we host the R-package "missForest" for the statistical software R.

The method is based on the publication Stekhoven and Bühlmann, 2012. The R package contains a vignette on how to use "missForest" in R including many helpful examples.

```
data |>
  select(`Majority protein IDs`,Experiment,Intensity) |>
  mutate(
    Intensity=replace(Intensity, Intensity==0, NA)
  ) |>
  pivot_wider(
    names_from=Experiment,
    values_from=Intensity
  ) |>
  column_to_rownames(var="Majority protein IDs") |>
  missForest()  -> data_forestimpute

data_forestimpute$ximp |>
  as_tibble(
    rownames="Majority protein IDs“
  ) -> data_forestimpute
```



Clustering Samples (PCA)

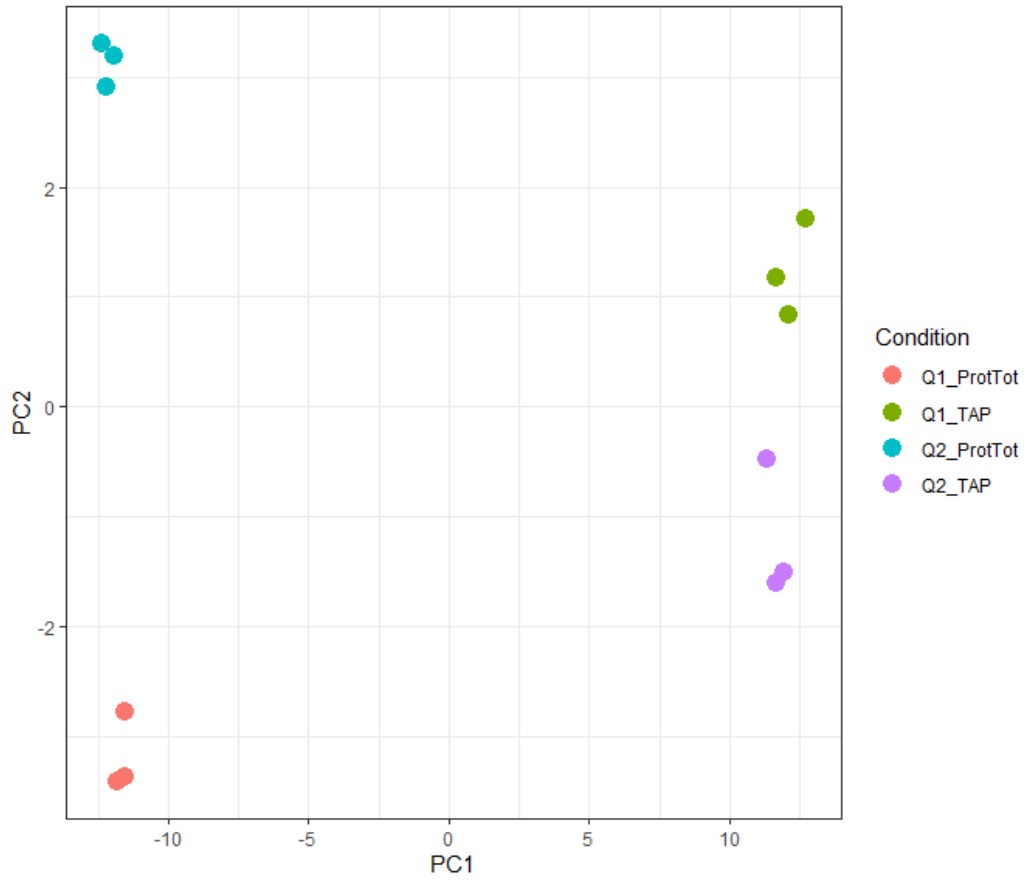
	O13563 <dbl>	O14455 <dbl>	O14467 <dbl>	O43137 <dbl>	P00330 <dbl>	P00358 <dbl>	P00359 <dbl>
Q1_ProtTot_Rep1	7.286075	8.302461	7.783146	6.930730	9.318606	8.263162	9.736675
Q1_ProtTot_Rep2	7.324097	8.235932	7.765318	6.997592	9.328950	8.349841	9.709228
Q1_ProtTot_Rep3	7.246868	8.163340	7.730435	6.992191	9.315088	8.474391	9.703154
Q1_TAP_Rep1	6.618976	8.849874	7.313234	6.796732	7.673712	6.807167	7.720672
Q1_TAP_Rep2	6.320042	8.869853	7.404868	6.759187	7.599283	6.669266	7.718253
Q1_TAP_Rep3	6.645432	8.817618	7.303282	6.760278	7.671867	6.750281	7.788034
Q2_ProtTot_Rep1	7.358981	8.374547	7.700808	6.783696	9.302937	8.854330	9.717121
Q2_ProtTot_Rep2	7.357344	8.301573	7.616108	6.829657	9.287511	8.624530	9.688438
Q2_ProtTot_Rep3	7.223080	8.336480	7.661576	6.766301	9.306725	8.813387	9.732868
Q2_TAP_Rep1	6.389343	8.843644	7.474041	6.766352	7.684361	6.913883	7.926995

- Rows are samples
  - (data frame rownames)
- Columns are Proteins
- No missing values
- Use R `prcomp` function

	PC1	PC2	PC3	PC4	PC5
Q1_ProtTot_Rep1	-11.58305	-3.5974748	0.8841948	0.005587525	-0.017716942
Q1_ProtTot_Rep2	-11.64434	-2.8864430	0.6160709	-0.123940570	-0.006679375
Q1_ProtTot_Rep3	-11.83566	-3.7090352	0.5243871	-0.140226891	0.114382490
Q1_TAP_Rep1	11.80773	0.6585281	2.5487219	2.521615018	-1.991443344
Q1_TAP_Rep2	12.88332	1.1566075	3.1540439	-2.069007829	2.432179534
Q1_TAP_Rep3	12.30649	0.6171419	2.0783029	-0.504767753	-1.160171452
Q2_ProtTot_Rep1	-12.49518	3.5022566	-0.5151248	0.078622603	-0.096536407
Q2_ProtTot_Rep2	-12.14816	3.3054506	-0.5858884	0.006436275	0.201530354
Q2_ProtTot_Rep3	-12.36482	3.1757716	-0.6446025	0.096724957	-0.088876603
Q2_TAP_Rep1	11.70348	-0.9714074	-3.1344744	-2.355042891	-1.803315189
Q2_TAP_Rep2	12.00338	-1.0577770	-2.4120315	2.573983116	2.052956895
Q2_TAP_Rep3	11.36679	-0.1936188	-2.5136000	-0.089983562	0.363690039

Clustering Samples (PCA)

```
data_forestimpute |>
  select(
    `Majority protein IDs`,
    Experiment,logIntensity
  ) |>  pivot_wider(
    names_from=Experiment,
    values_from=logIntensity
  ) |>
  column_to_rownames("Majority protein IDs") |>
  t() |>
  prcomp() -> pca_result

pca_result$x |>
  as_tibble(rownames="Experiment") -> pca_result

pca_result |>
  left_join(data |> distinct(Experiment,Condition)) |>
  ggplot(aes(x=PC1, y=PC2, colour=Condition)) +
  geom_point(size=4)
```



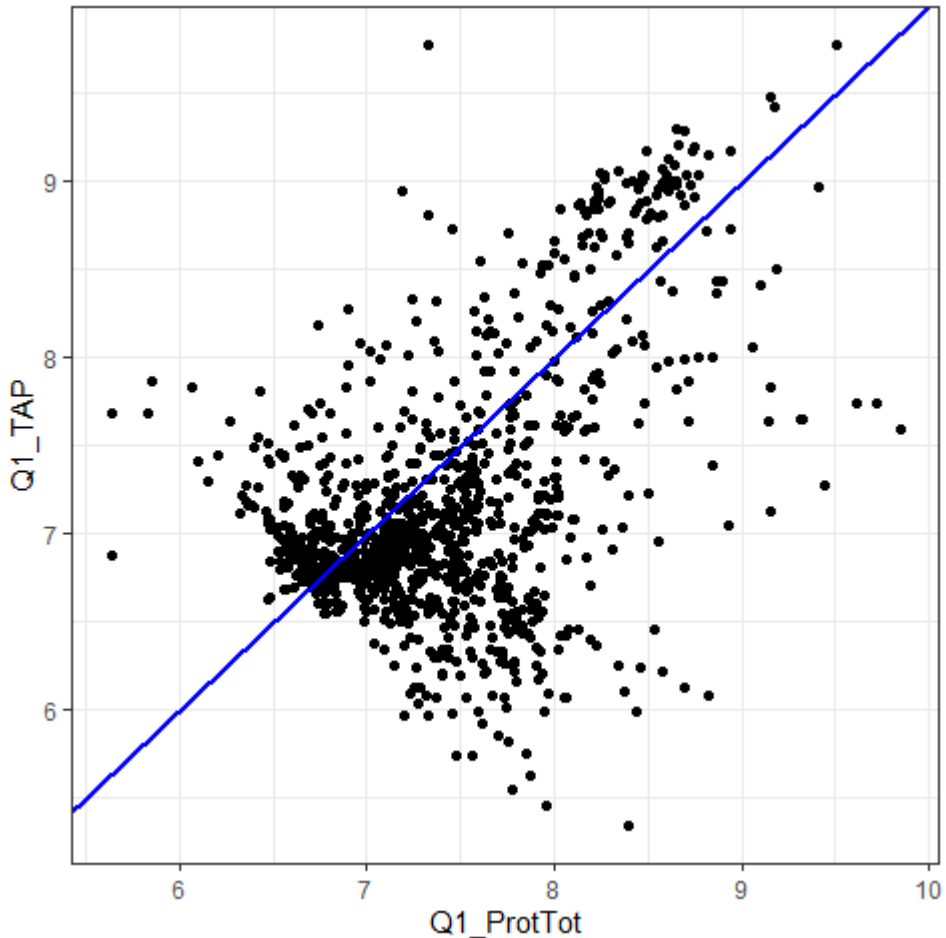
Summarising Conditions

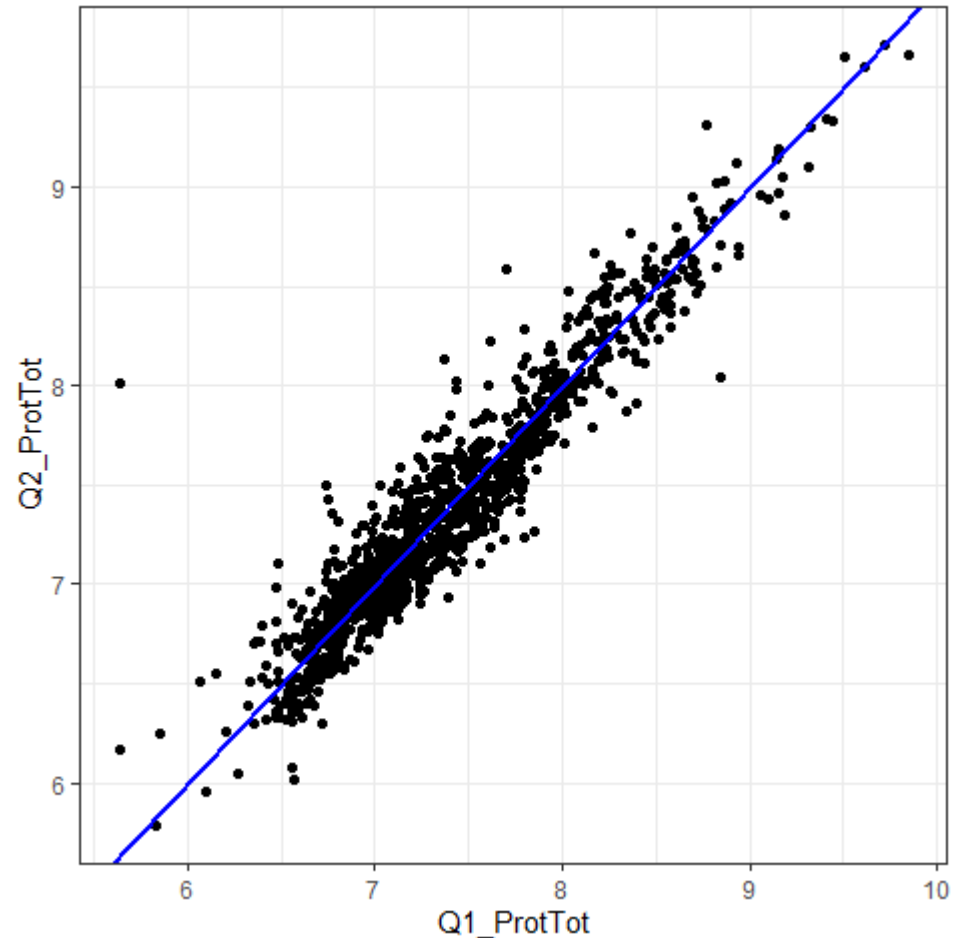
```
data_forestimpute |>
  group_by(
    `Majority protein IDs`,Condition
  ) |>
  summarise(
    logIntensity = mean(logIntensity)
  ) |>
  pivot_wider(
    names_from=Condition,
    values_from=logIntensity
  ) -> data_per_condition
```

- Mean of log transformed values
  - Not the same as log of original means
  - Should always work on log data

Majority protein IDs <chr>	Q1_ProtTot <dbl>	Q1_TAP <dbl>	Q2_ProtTot <dbl>	Q2_TAP <dbl>
O13563	7.285680	6.598893	7.313135	6.487595
O14455	8.233911	8.845782	8.337533	8.845681
O14467	7.759633	7.340462	7.659497	7.501618
O43137	6.973504	6.775351	6.793218	6.756045
P00330	9.320881	7.648287	9.299057	7.647194
P00358	8.362465	7.036090	8.764082	6.862700

Condition Scatterplots





Exercise

Visualising, Normalising, Exploring

Statistical Analysis

Babraham Bioinformatics

Statistics for Proteomics

- Intensity value is a continuous measurement
- Values are stable on a log scale

- Standard continuous statistics are applicable
  - T-Test, ANOVA, Linear Models

rstatix

Provides a simple and intuitive pipe-friendly framework, coherent with the 'tidyverse' design philosophy, for performing basic statistical tests, including t-test, Wilcoxon test, ANOVA, Kruskal-Wallis and correlation analyses.

- More specialised statistics can offer more power

DEqMS ROTS proDA MSstats DEP

Multiple Testing Correction

- Thousands of tests performed
  - One for each protein
- Raw p-values only valid if only one test performed
  - Some stats will correct automatically – need to check
- False positive rate will be high if many tests used

- P-value correction
  - Benjamini and Hochberg False Discovery Rate
  - `p.adjust` function in R

Testing over multiple conditions (ANOVA)

```
data |>
  group_by(`Majority protein IDs`) |>
  anova_test(
    logIntensity ~ Condition
  ) |>
  ungroup() -> anova_results

anova_results |>
  as_tibble() |>
  mutate(
    fdr = p.adjust(p, method="fdr")
  ) |>
  select(
    `Majority protein IDs`,DFn,DFd,fdr
  ) -> anova_results
```

- Highly powered
  - Often over-powered
- Doesn't say which conditions drive changes
  - Post-hoc tests / filtering

Majority protein IDs <chr>	DFn <dbl>	DFd <dbl>	fdr <dbl>
O13563	3	8	5.290739e-04
O14455	3	8	1.106394e-06
O14467	3	8	1.610413e-05
O43137	3	8	7.047890e-05
P00330	3	8	4.519810e-10
P00358	3	8	1.380839e-07

Simple T-Test

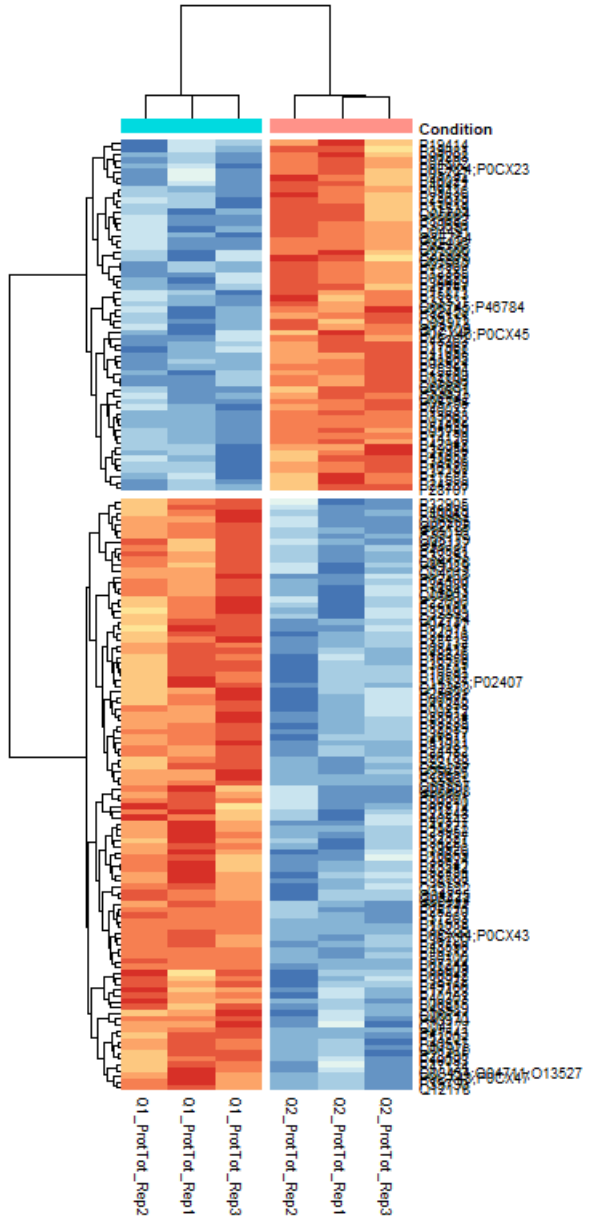
(usually 2 conditions)

```
data |>
  filter (str_detect(Experiment,"ProtTot")) |>
  group_by(`Majority protein IDs`) |>
  t_test(logIntensity ~ Condition) |>
  as_tibble() |>
  mutate(fdr = p.adjust(p, method="fdr")) |>
  mutate(
    significant = if_else(
      fdr < 0.05,
      "Significant",
      "Not Significant")
  ) |>
  select(
    `Majority protein IDs`,
    group1,group2,
    statistic, fdr,
    significant
  ) |>
  arrange(fdr) -> q1_vs_q2_t_test
```

Majority protein IDs <chr>	group1 <chr>	group2 <chr>	statistic <dbl>	fdr <dbl>	significant <chr>
P38235	Q1_ProtTot	Q2_ProtTot	42.59942	0.0172483	Significant
P13663	Q1_ProtTot	Q2_ProtTot	-21.08736	0.0180037	Significant
P14120	Q1_ProtTot	Q2_ProtTot	-21.01963	0.0180037	Significant
P37263	Q1_ProtTot	Q2_ProtTot	22.38147	0.0180037	Significant
P38009	Q1_ProtTot	Q2_ProtTot	28.11876	0.0180037	Significant
P43590	Q1_ProtTot	Q2_ProtTot	16.71045	0.0180037	Significant

Plotting Statistical Results (Heatmap)

```
q1_vs_q2_t_test |>
  filter(fdr<0.05) |>
  pull(`Majority protein IDs`) -> sig_proteins

data |>
  filter (
    str_detect(Experiment,"ProtTot")
  ) |>
  filter(
    `Majority protein IDs` %in% sig_proteins
  ) |>
  tidy_heatmap(
    rows=`Majority protein IDs`,
    columns=Experiment,
    values = logIntensity,
    annotation_col = Condition,
    scale="row",
    cluster_rows = TRUE
)
```



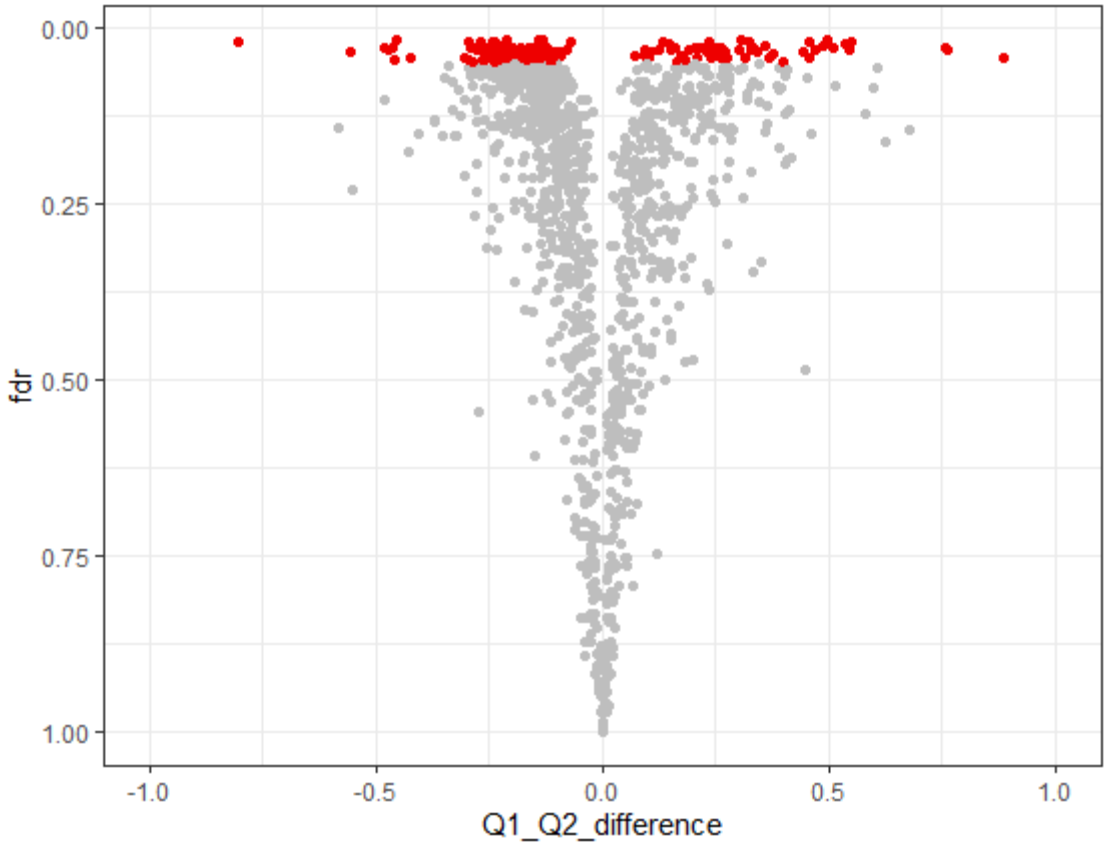
Plotting statistical results (Volcano Plot)

```
data |>
  filter (str_detect(Experiment,"ProtTot")) |>
  group_by(`Majority protein IDs`, Condition) |>
  summarise(logIntensity = mean(logIntensity)) |>
  ungroup() |>
  pivot_wider(
    names_from=Condition,
    values_from=logIntensity
  ) |>
  mutate(Q1_Q2_difference = Q2_ProtTot - Q1_ProtTot) |>
  select(`Majority protein IDs`,Q1_Q2_difference) |>
  right_join(t_test_results) ->t_test_results
```

Majority protein IDs <chr>	Q1_Q2_difference <dbl>	group1 <chr>	group2 <chr>	statistic <dbl>	fdr <dbl>	significant <chr>
O13563	0.02745538	Q1_ProtTot	Q2_ProtTot	-0.5464083	0.69585120	Not Significant
O14455	0.10362188	Q1_ProtTot	Q2_ProtTot	-2.2842027	0.19397384	Not Significant
O14467	-0.10013591	Q1_ProtTot	Q2_ProtTot	3.4580365	0.10246277	Not Significant
O43137	-0.18028588	Q1_ProtTot	Q2_ProtTot	6.3073772	0.03927156	Significant
P00330	-0.02182393	Q1_ProtTot	Q2_ProtTot	3.0312232	0.12073348	Not Significant
P00358	0.40161772	Q1_ProtTot	Q2_ProtTot	-4.2894526	0.06717984	Not Significant

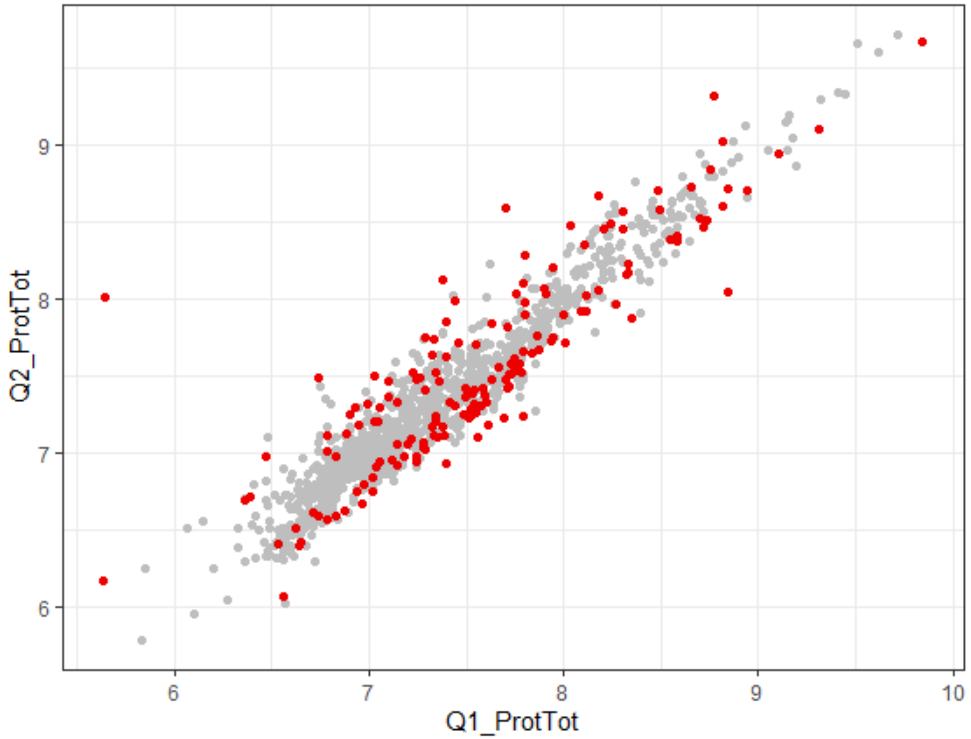
Plotting statistical results (Volcano Plot)

```
t_test_result |>
  ggplot(
    aes(
      x=Q1_Q2_difference,
      y=fdr,
      colour=significant
    )
  ) +
  geom_point() +
  scale_y_log10() +
  scale_y_reverse() +
  scale_colour_manual(values=c("grey","red2"))
```



Highlighting hits on a Scatterplot

```
data |>
  filter (str_detect(Experiment,"ProtTot")) |>
  group_by(`Majority protein IDs`, Condition) |>
  summarise(logIntensity = mean(logIntensity)) |>
  ungroup() |>
  pivot_wider(
    names_from=Condition,
    values_from=logIntensity
  ) |>
  left_join(
    q1_vs_q2_t_test |> select(`Majority protein IDs`,significant)
  ) |>
  arrange(significant) |>
  ggplot(aes(x=Q1_ProtTot, y=Q2_ProtTot, colour=significant)) +
  geom_point() +
  scale_colour_manual(values=c("grey","red2"))
```



LIMMA

- More powerful quantitative statistics
- Improved estimation of biological variance in poorly measured samples

- Relies on predicting variance across proteins
  - Magnitude of measurement
  - Number of measured peptides

DEqMS

- Should check whether the relationship exists in your data
  - Many people don’t!

Building LIMMA data

```
data |>
  filter(str_detect(Condition,"ProtTot")) |>
  select(
    `Majority protein IDs`,Experiment,logIntensity
  ) |>
  pivot_wider(
    names_from=Experiment,
    values_from=logIntensity
  ) |>
  column_to_rownames("Majority protein IDs") -> limma_data
```

	Q1_ProtTot_Rep1 <dbl>	Q1_ProtTot_Rep2 <dbl>	Q1_ProtTot_Rep3 <dbl>	Q2_ProtTot_Rep1 <dbl>	Q2_ProtTot_Rep2 <dbl>	Q2_ProtTot_Rep3 <dbl>
O13563	7.286075	7.324097	7.246868	7.358981	7.357344	7.223080
O14455	8.302461	8.235932	8.163340	8.374547	8.301573	8.336480
O14467	7.783146	7.765318	7.730435	7.700808	7.616108	7.661576
O43137	6.930730	6.997592	6.992191	6.783696	6.829657	6.766301
P00330	9.318606	9.328950	9.315088	9.302937	9.287511	9.306725
P00358	8.263162	8.349841	8.474391	8.854330	8.624530	8.813387

Building LIMMA Design

```
tibble(
  Experiment = colnames(limma_data),
  Control = 1,
  Q1_vs_Q2 = if_else(
    startsWith(Experiment,"Q1"),1,0
  )) |>
  column_to_rownames("Experiment") -> limma_design
```

	Control <dbl>	Q1_vs_Q2 <dbl>
Q1_ProtTot_Rep1	1	1
Q1_ProtTot_Rep2	1	1
Q1_ProtTot_Rep3	1	1
Q2_ProtTot_Rep1	1	0
Q2_ProtTot_Rep2	1	0
Q2_ProtTot_Rep3	1	0

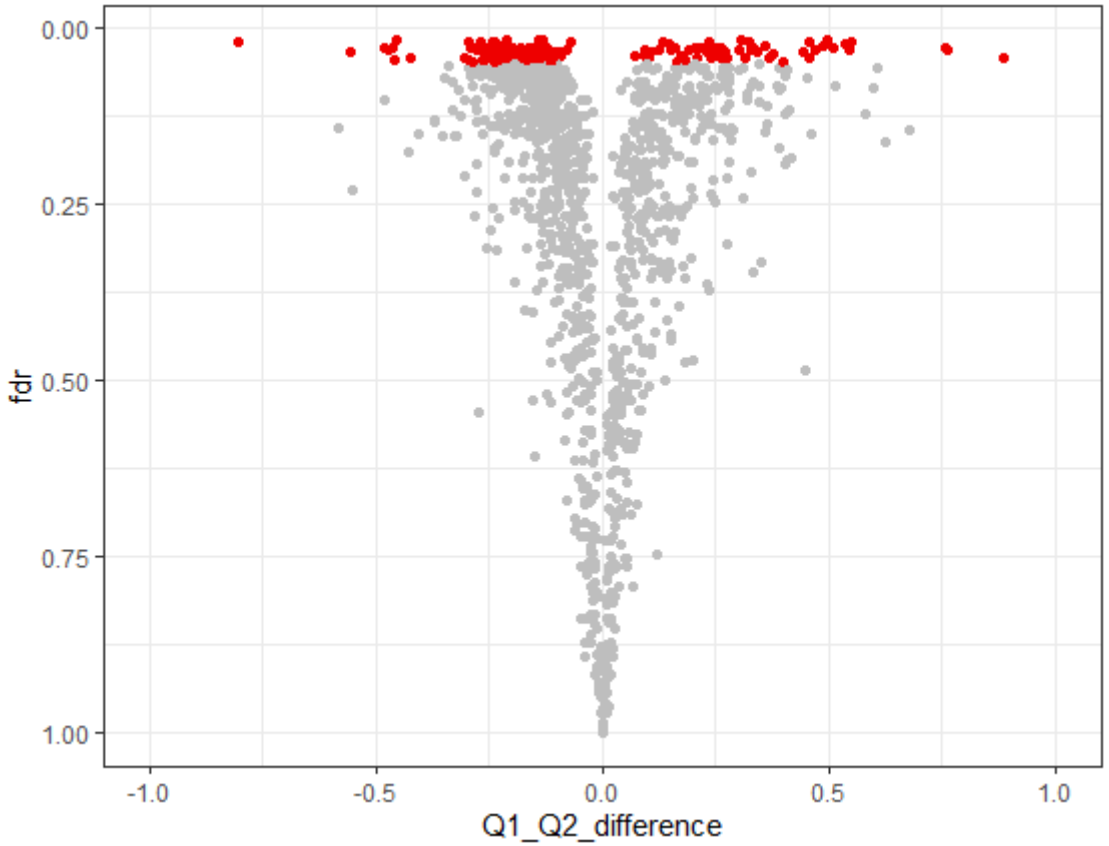
Running LIMMA

```
lmFit(limma_data,limma_design) -> limma_fit

eBayes(limma_fit) -> limma_fit

topTable(
  limma_fit,
  coef="Q1_vs_Q2",
  adjust="BH",
  number = nrow(limma_data)
) |>
  as_tibble(
    rownames="Majority Protein IDs“
  ) -> limma_results
```

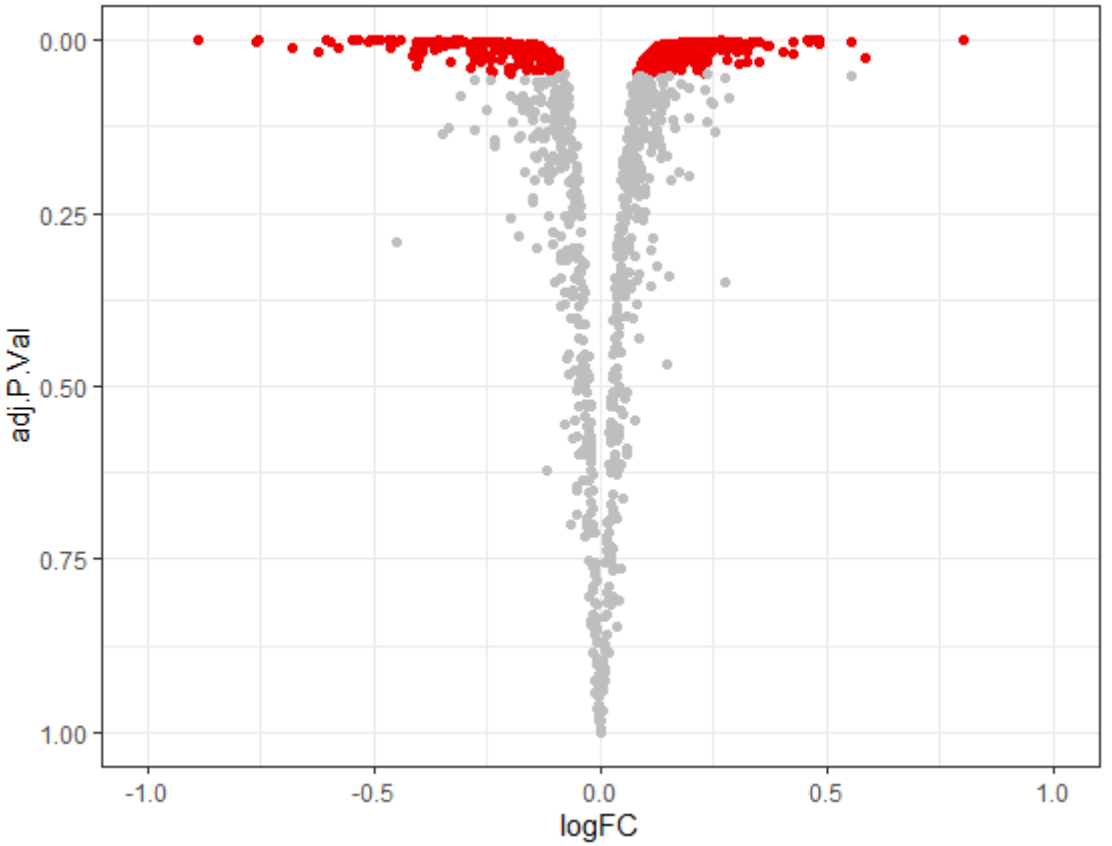
Majority Protein IDs <chr>	logFC <dbl>	AveExpr <dbl>	t <dbl>	P.Value <dbl>	adj.P.Val <dbl>	B <dbl>
P07262	-2.3716942	6.828316	-46.85688	4.275354e-10	5.382670e-07	12.788278
P54839	-0.7557947	7.750687	-24.85060	3.728122e-08	2.346853e-05	9.603240
P25294	-0.5496872	7.711969	-16.71981	5.920557e-07	1.760807e-04	7.013988
P05744	0.8029173	8.443816	16.32866	6.975407e-07	1.760807e-04	6.850459
P38009	0.4562778	7.332783	16.08098	7.753641e-07	1.760807e-04	6.744489
P14120	-0.4950243	8.421784	-15.89822	8.391457e-07	1.760807e-04	6.665051

Volcano Comparison

T-Test

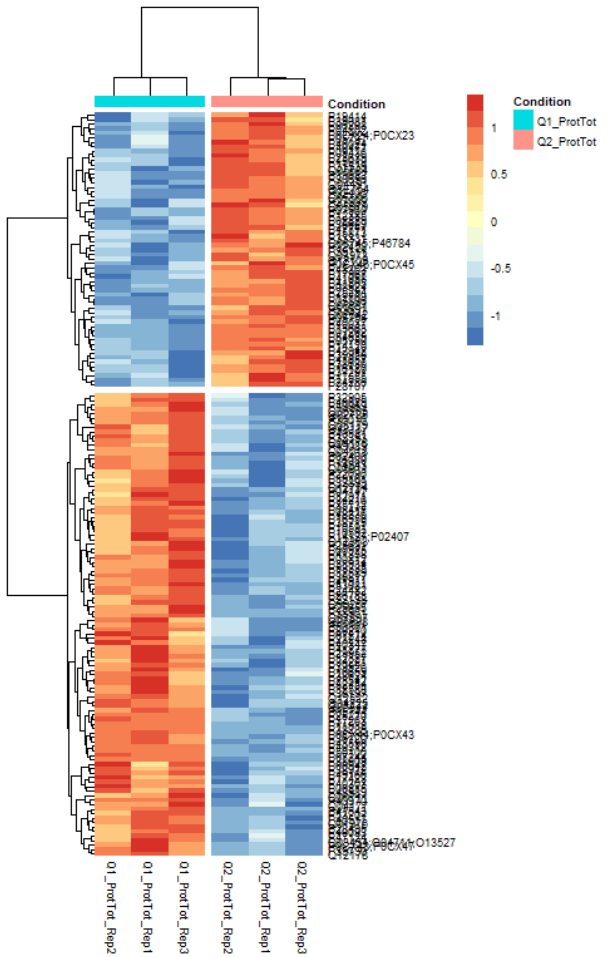


LIMMA

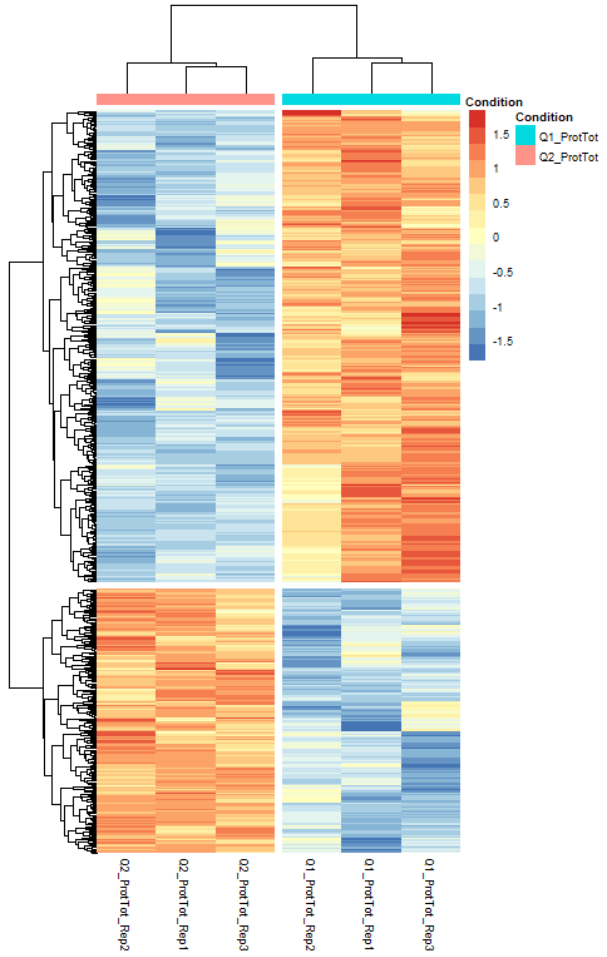


Heatmap Comparison

T-Test



LIMMA



Exercise

Statistical Analysis