



netherlands
proteomics
centre

RockerBox

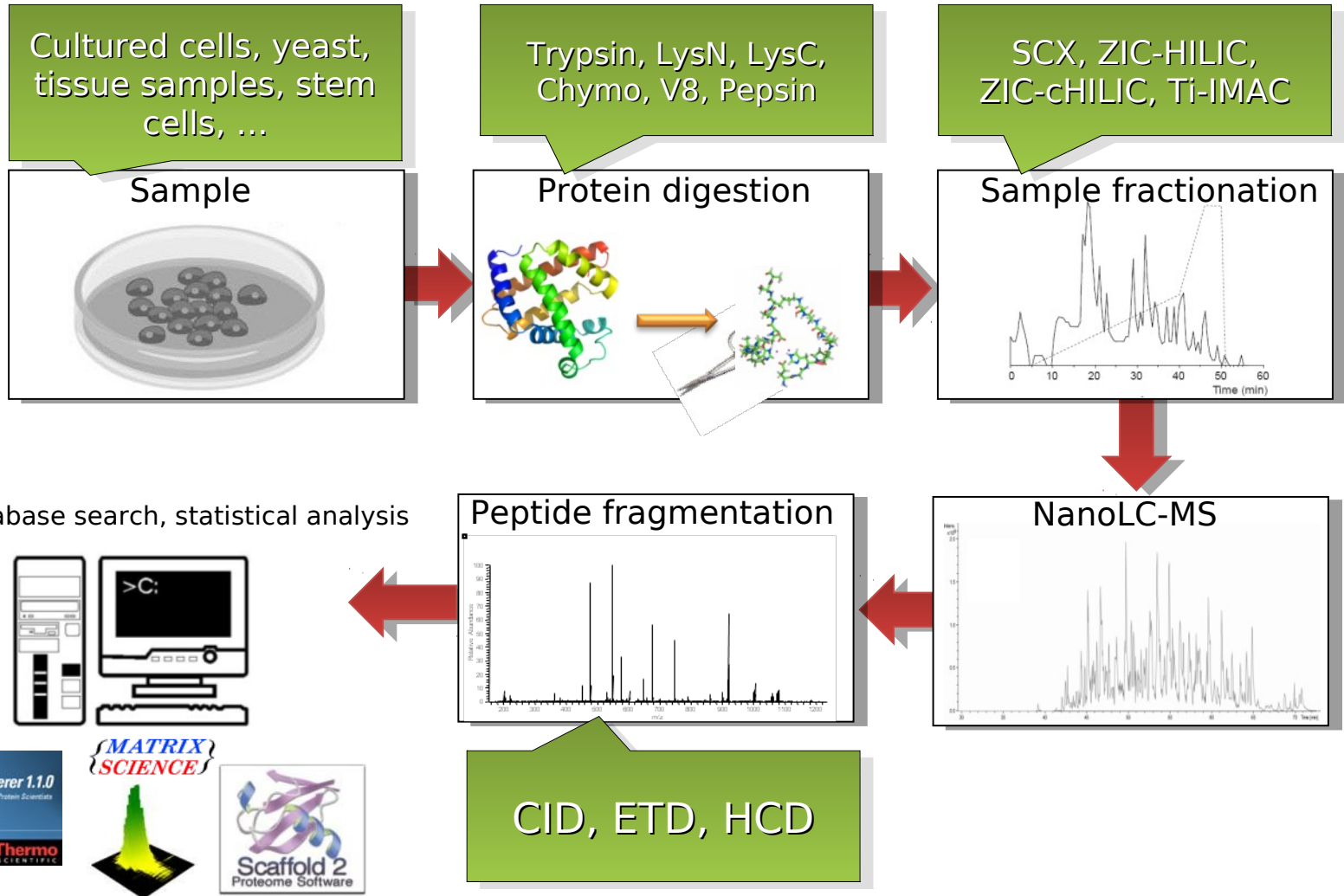
*Filtering massive Mascot
search results at the .dat
level*

Challenges

- “Big” experiments
- High amount of data
- Large raw and .dat files (> 2GB)

- How to handle our results??
 - The ‘2.2’ peptide summary could not be made by Mascot
 - MSQuant couldn’t load the result files

Proteomics workflows

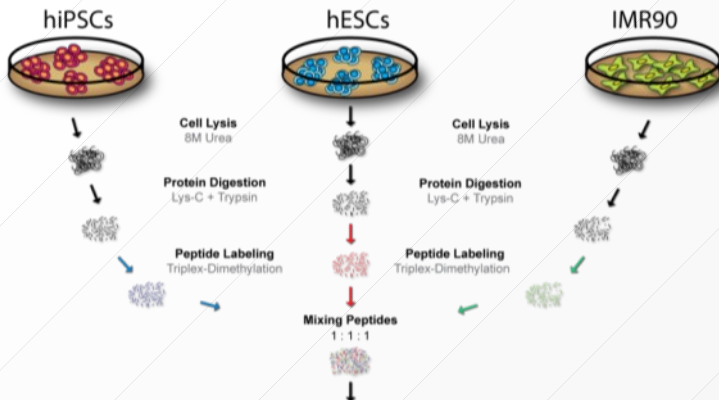


Boxes

- 'proteomics' mass spectrometers
 - 3 Orbitrap (Thermo)
 - 2 Orbitrap Velos(Thermo)
 - Quadropole TOFs (Agilent, Waters and AB Sciex)
 - 2 Triple Quad (AB Sciex, Thermo)
- 70 Terabytes of stored data
- Software:
 - Preprocessing scripts: in-house, MaxQuant, Proteome Discoverer, Scaffold, MSQuant...
 - Mascot 2.3 (Linux)

Large MS experiment

Biological Replica 1



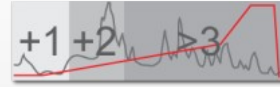
"Big SCX"



40 LC-MS/MS (CID)
14 LC-MS/MS (ETD)

PTM ids

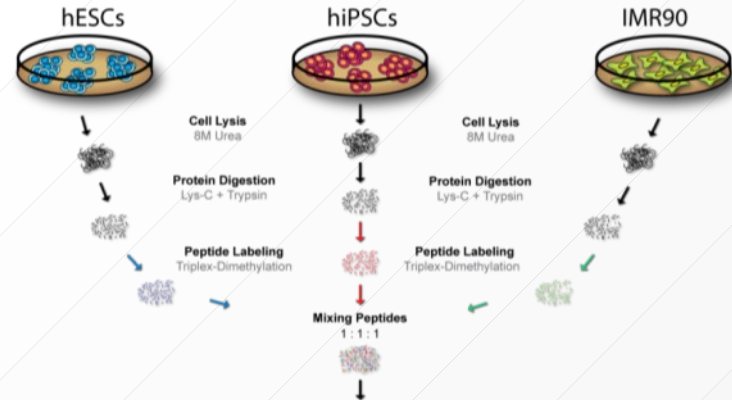
"Small SCX"



40 LC-MS/MS (CID)
14 LC-MS/MS (ETD)

Protein ids

Biological Replica 2



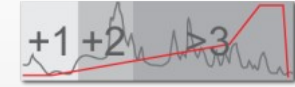
"Big SCX"



40 LC-MS/MS (CID)
14 LC-MS/MS (ETD)

PTM ids

"Small SCX"



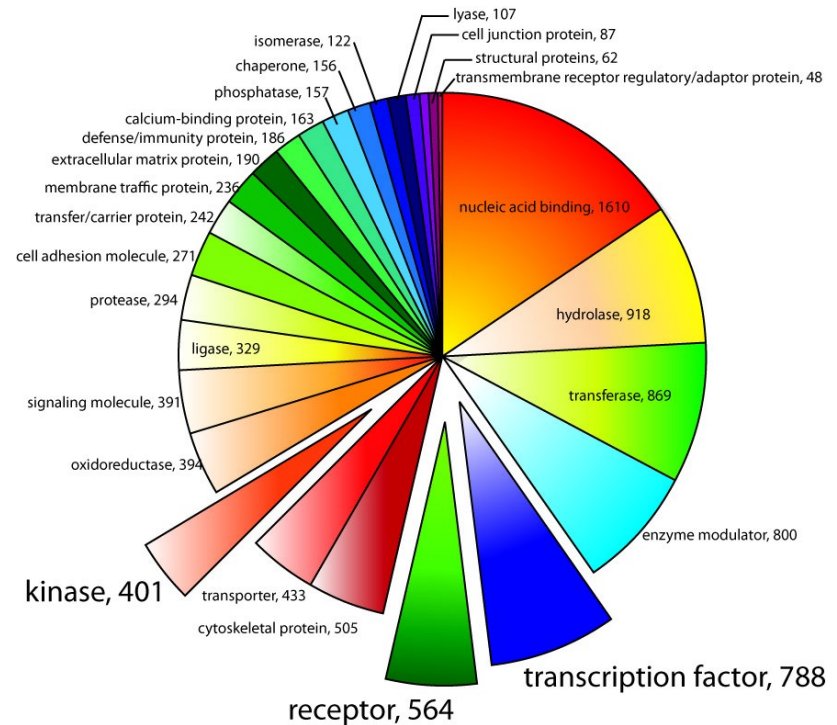
40 LC-MS/MS (CID)
14 LC-MS/MS (ETD)

Protein ids

Experimental design and MS results

- Biological replica (label swap)
- 196 LC-MS/MS (3 h gradient)
- LTQ-Orbitrap: CID/ETD
- 2,440,583 MS/MS spectra collected
- 568,054 PSMs (FDR=1.02%)
- 68,172 unique peptides
- **10,683 unique proteins**

130 GB raw files
12 GB .dat files



ROCKERBOX

Meeting the challenges

What is RockerBox?

- Filtering .dat file peptide spectrum matches (PSMs)
- Charting of search results
- Combining .dat files^(new)
- Exporting text files with PSMs
- Cross-platform usability (Java)

van den Toorn HW, Muñoz J, Mohammed S, Raijmakers R, Heck AJ, van Breukelen B. **RockerBox: analysis and filtering of massive proteomics search results.** J Proteome Res. 2011 Mar 4;10(3):1420-4.

Input files

C:\Users\toom101.SOLISCOM\Desktop\F275563.dat

Messages

output system log

(ppm) over Retention time plot chart finished.
 [F274658.dat.db] Creating Mascot ions score over Mass delta (ppm) plot chart
 [F274658.dat.db] Mascot ions score over Mass delta (ppm) plot chart finished.
 [F275563.dat.db] Creating Mascot ions score over Mass delta (ppm) plot chart
 [F275563.dat.db] Mascot ions score over Mass delta (ppm) plot chart finished.
 [F275563.dat.db] Creating Mascot ions score over m/z value plot chart

Progress

Collecting decoy data for "Mascot ions score over m/z value plot" from F275563.dat.db 49%

memory 

File info

F275563.dat

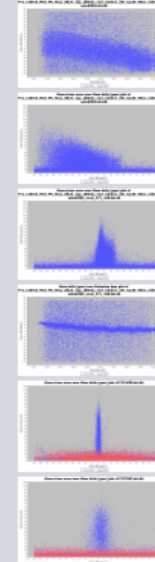
Parameters

**Proof-of-principle-masswindow
 HUEVC-Forward-Phospho-ETD-FT**

input file D:\Lars\Forward\ETD-FT\HUEVC-Forward-Phospho...
 queries: 18853
 user: henk (2)
 e-mail: h.w.p.vandentoom@uu.nl
instrument: FTMS-ECD
 pep. mass tol.: 10.0 ppm
 fragm. mass tol.: 0.05 Da
database: SwissProt
 398181 sequences (20407 after tax), 143572911 residues
 taxonomy: Homo sapiens (human)
 auto decoy: yes
enzyme: Trypsin
 max miscleavages: 2
 fixed mods: Carbamidomethyl (C)
 variable mods: Dimethyl (K) · Dimethyl (N-term) · Dimethyl:2H(4) (K) · Dimethyl:2H(4) (N-term) · Oxidation (M)

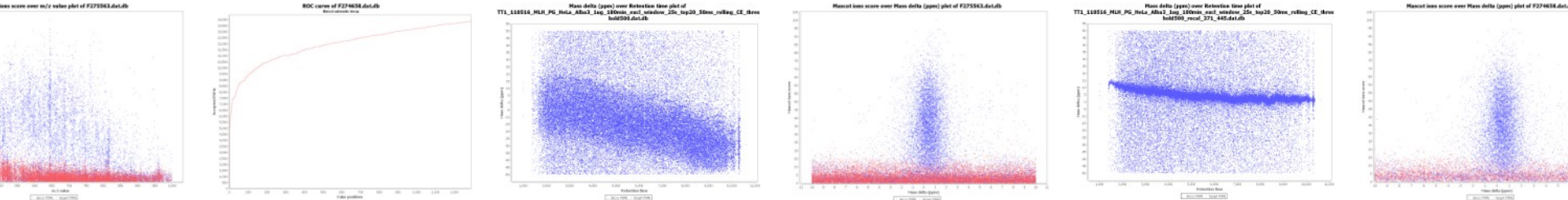
export .par file

Charts



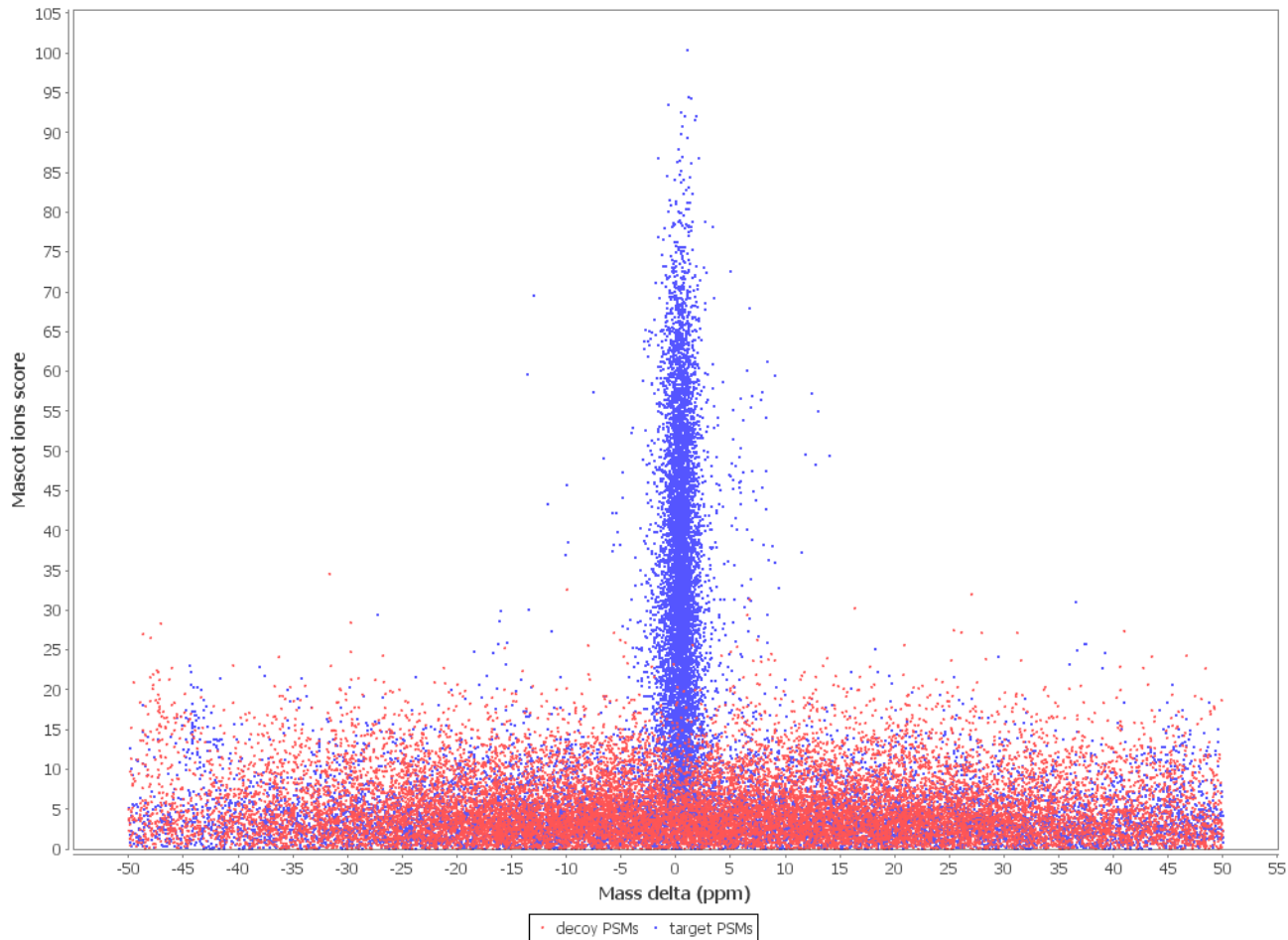
>>>

CHARTS



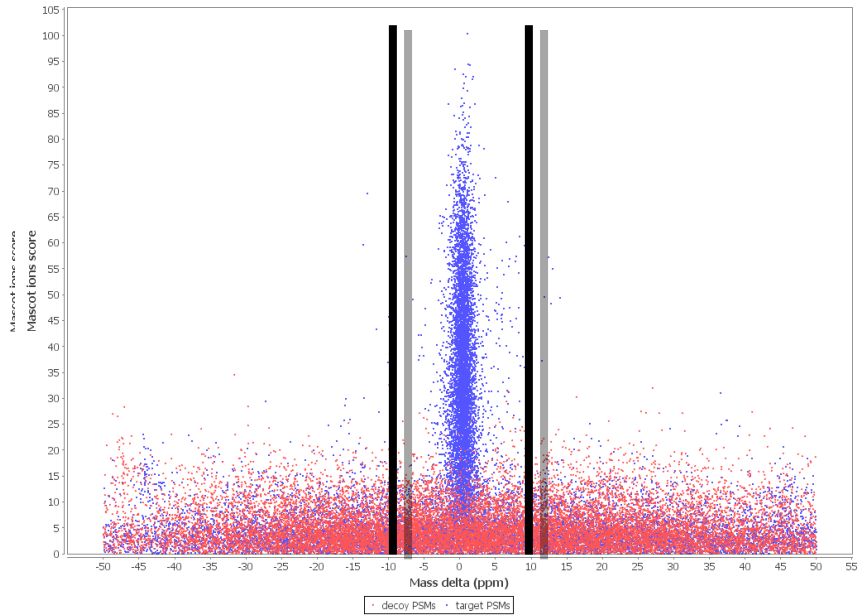
Example data set

Mascot ions score over Mass delta (ppm) plot of F274658.dat.db

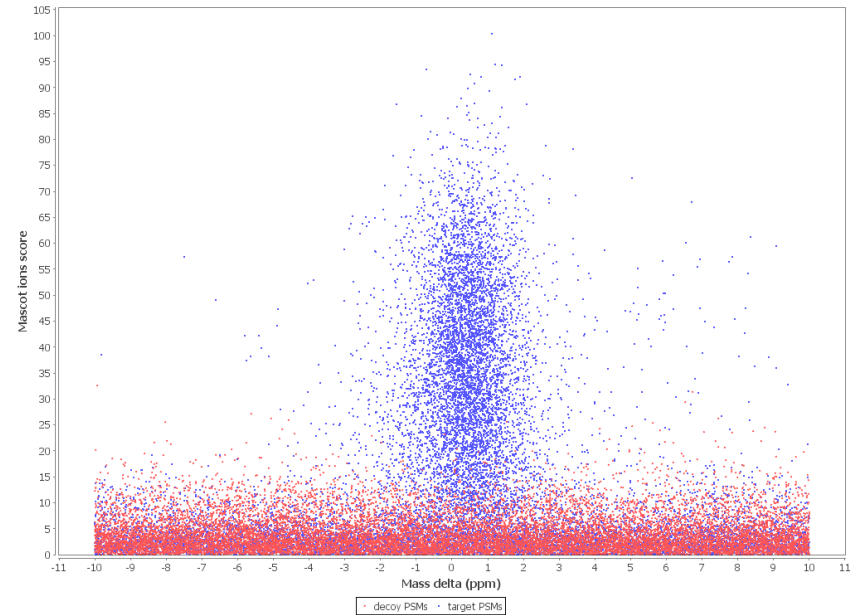


Wide search window

Mascot ions score over Mass delta (ppm) plot of F274658.dat.db

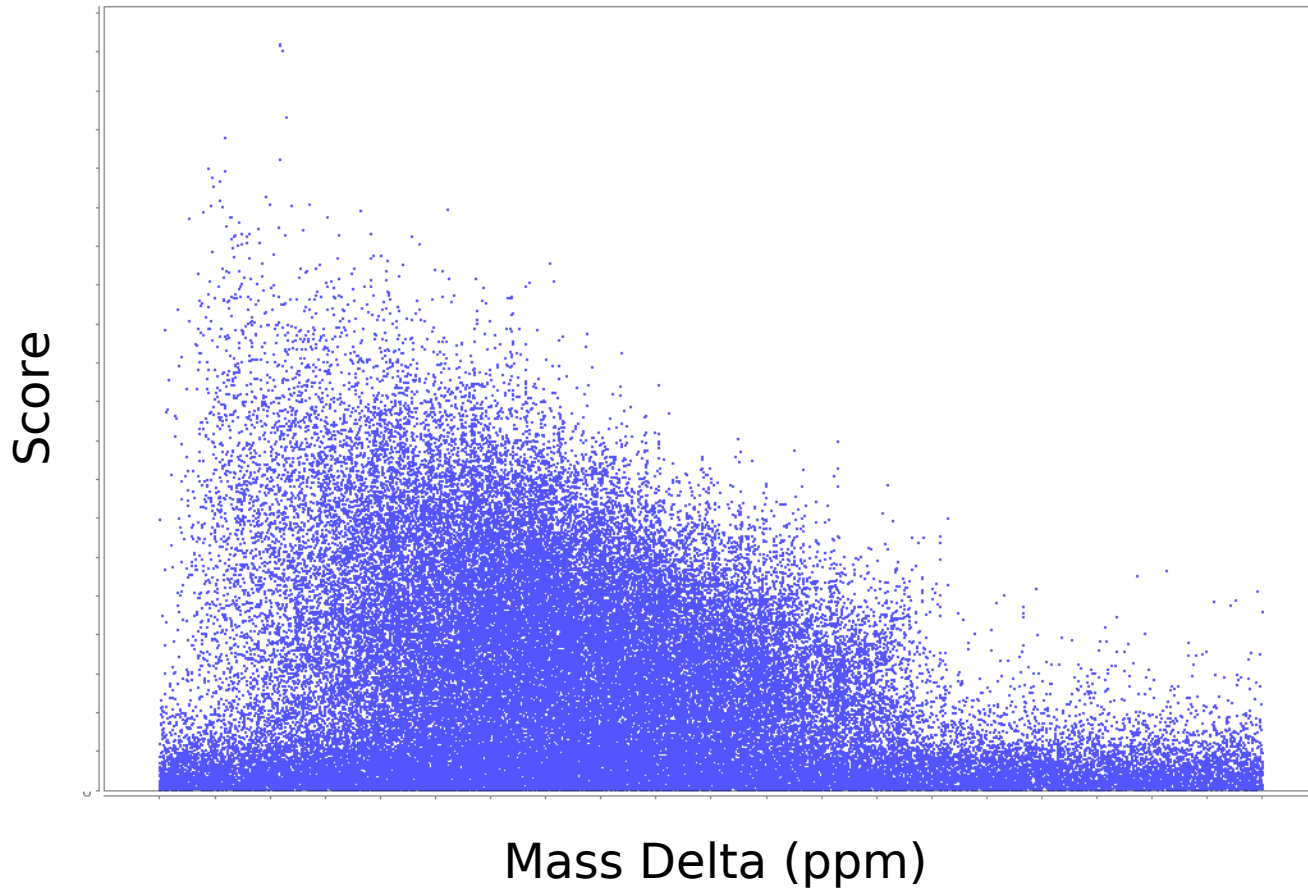


Mascot ions score over Mass delta (ppm) plot of F275563.dat.db

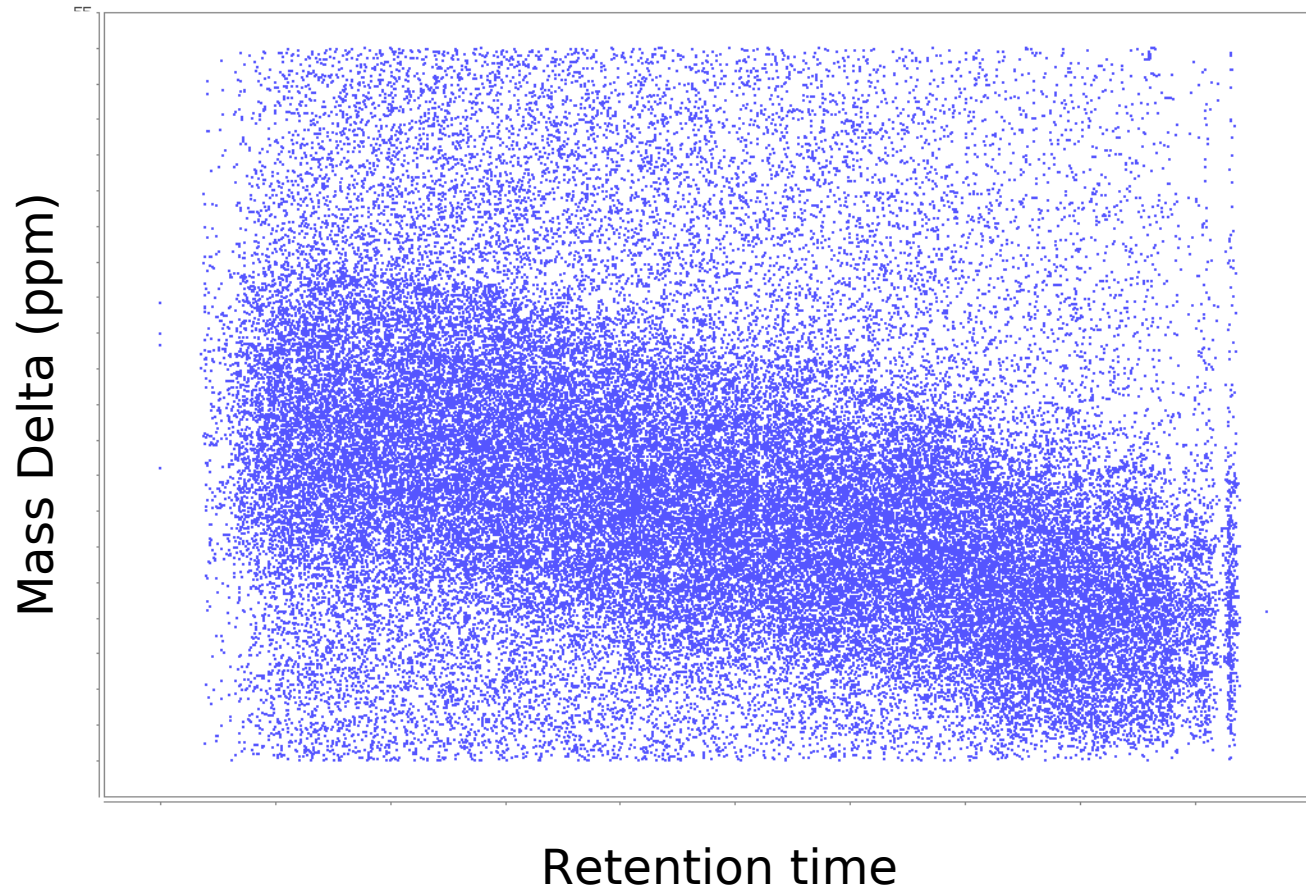


TOF: score vs. mass delta

Mascot ions score over Mass delta (ppm) plot of
TT1_110516_MLH_PG_HeLa_Alba3_1ug_180min_excl_window_25s_top20_50ms_rolling_CE_thres
hold500.dat.db

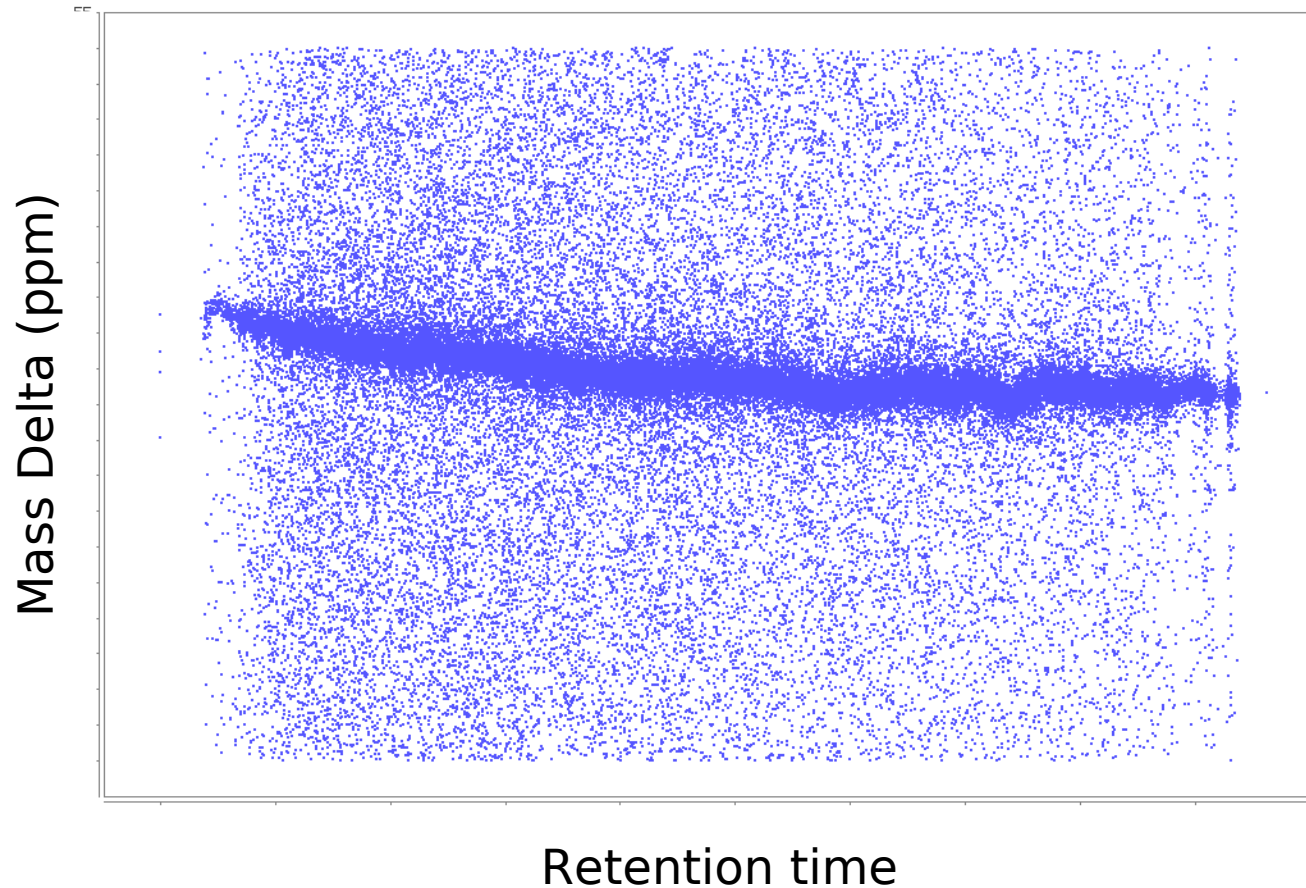


Mass delta (ppm) over Retention time plot of
TT1_110516_MLH_PG_HeLa_Alba3_1ug_180min_excl_window_25s_top20_50ms_rolling_CE_thres
hold500.dat.db



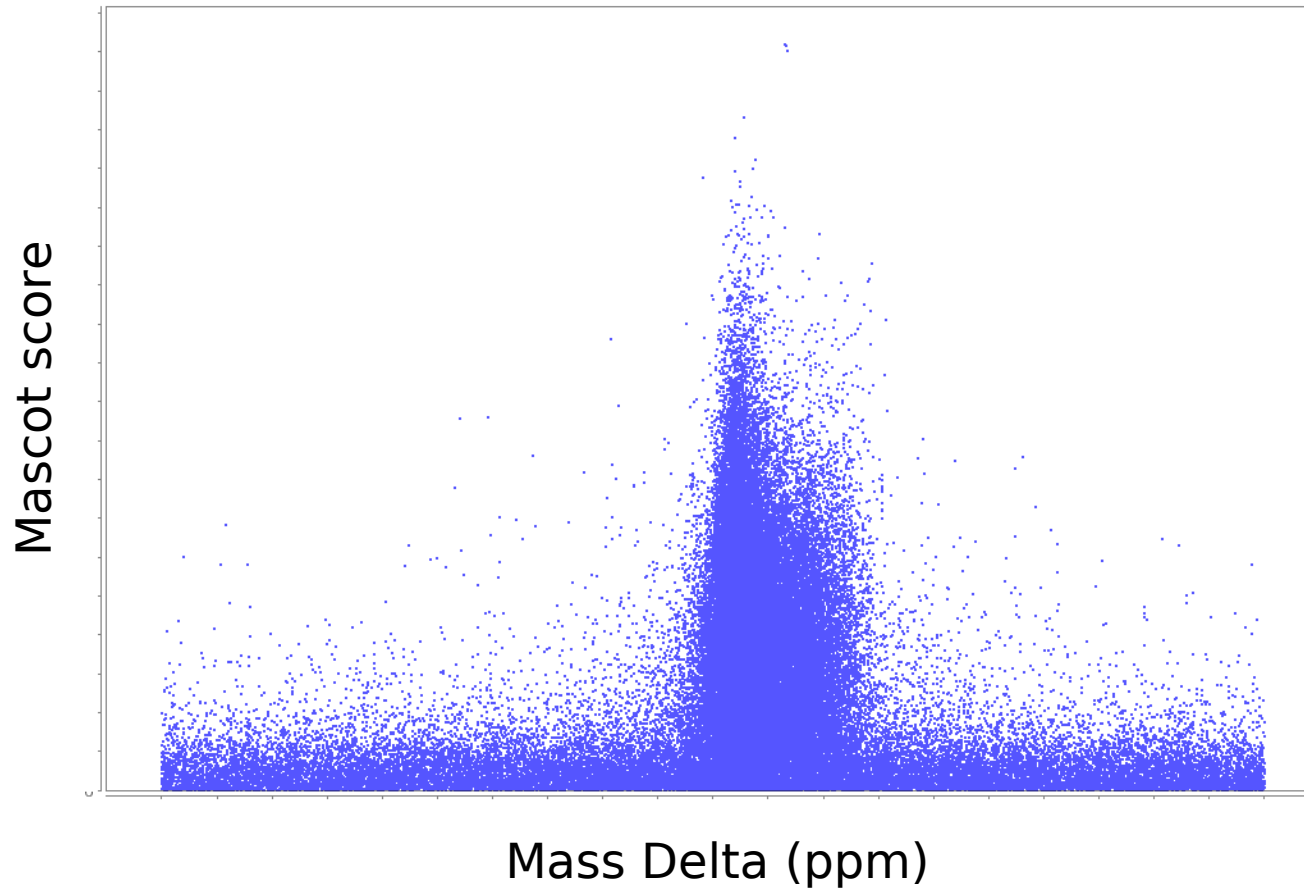
Mass-based calibration applied

Mass delta (ppm) over Retention time plot of
TT1_110516_MLH_PG_HeLa_Alba3_1ug_180min_excl_window_25s_top20_50ms_rolling_CE_thres
hold500_recal_371_445.dat.db

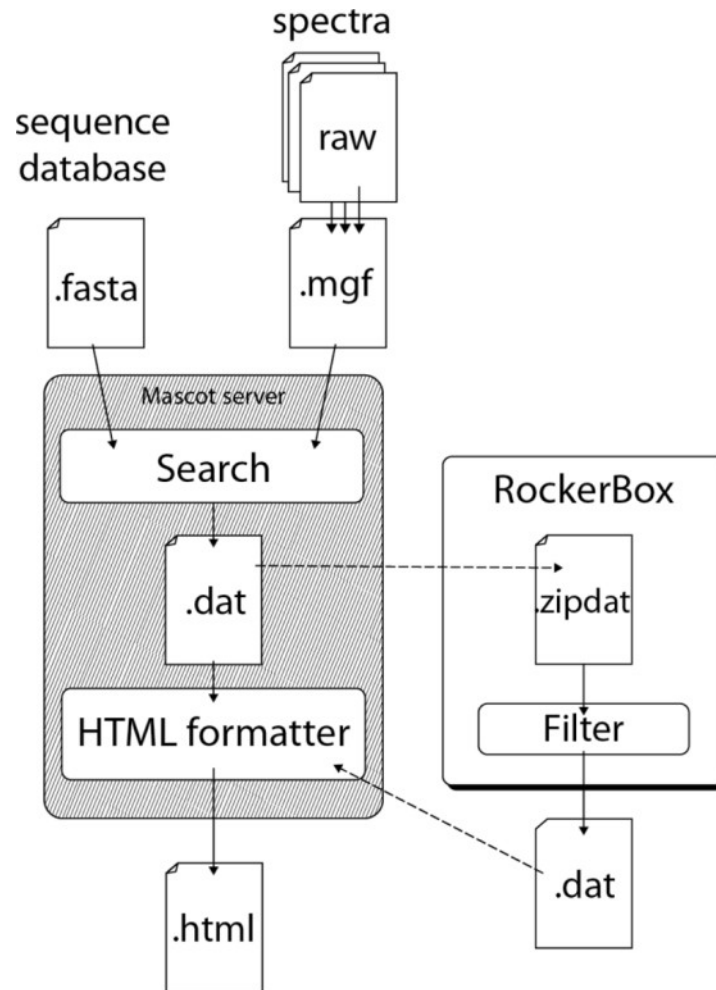


Mass-based calibration applied

Mascot ions score over Mass delta (ppm) plot of
TT1_110516_MLH_PG_HeLa_Alba3_1ug_180min_excl_window_25s_top20_50ms_rolling_CE_thres
hold500_recal_371_445.dat.db



Workflow



Removing PSMs?

- Many spectra are not matched to a correct peptide sequence
 - Low quality real spectra (signal/noise ratio)
 - Spectra from non-peptide origins
 - Mixed peptide spectra
 - Spectra from peptides not in the database
- These low quality matches are abundant
 - Typically around 50%
- Which PSMs really matter?

ROCKERBOX FILTERING METHODS

An overview

Manual filter: full control

Manual filter

Score cutoff

Mascot score cutoff

Modifications

Peptide contain:

Mass delta

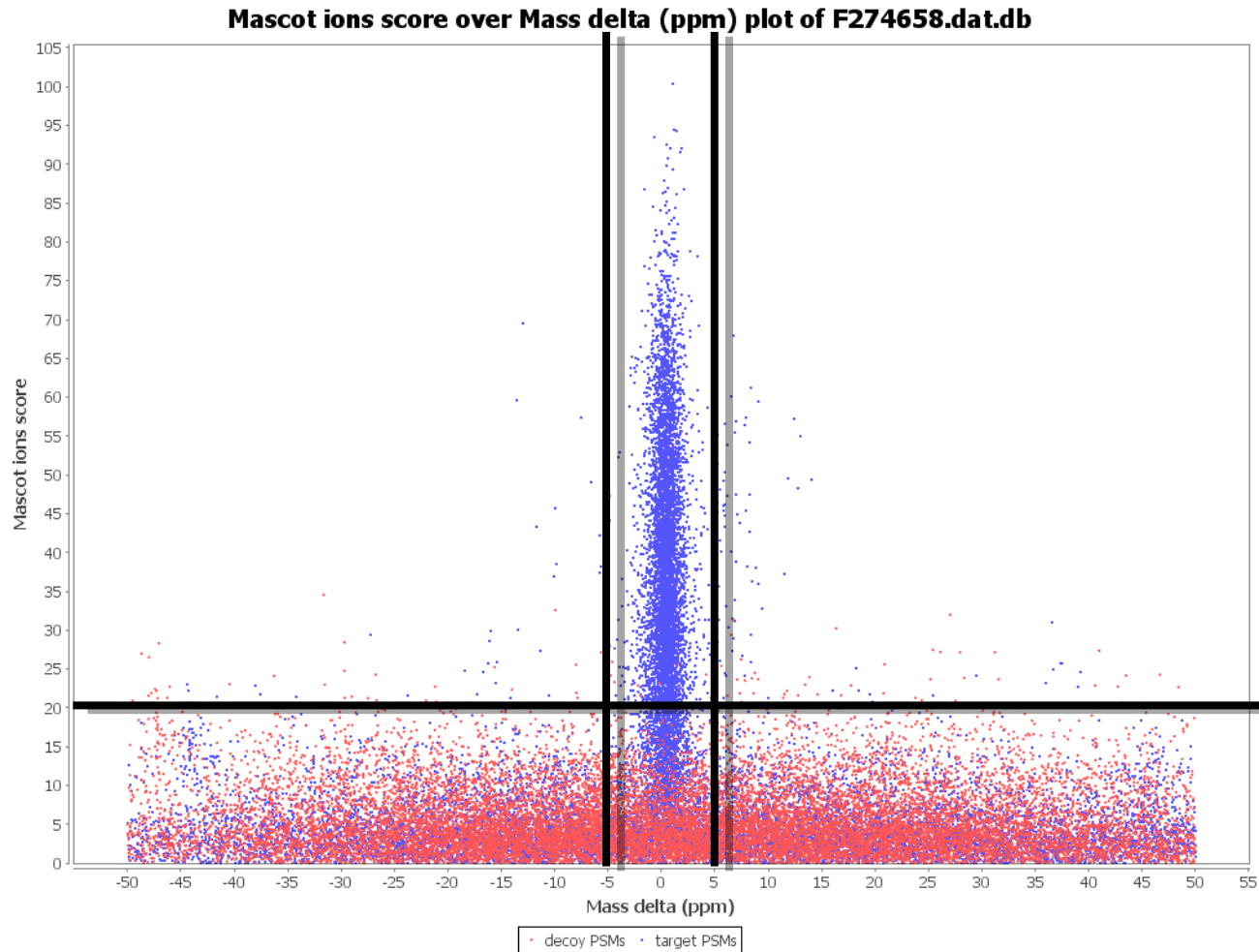
Mass delta should be between and ppm.

Only use highest ranking PSMs
 keep mascot automatic decoy

Export method

- Mascot score
- Modifications
- Mass delta

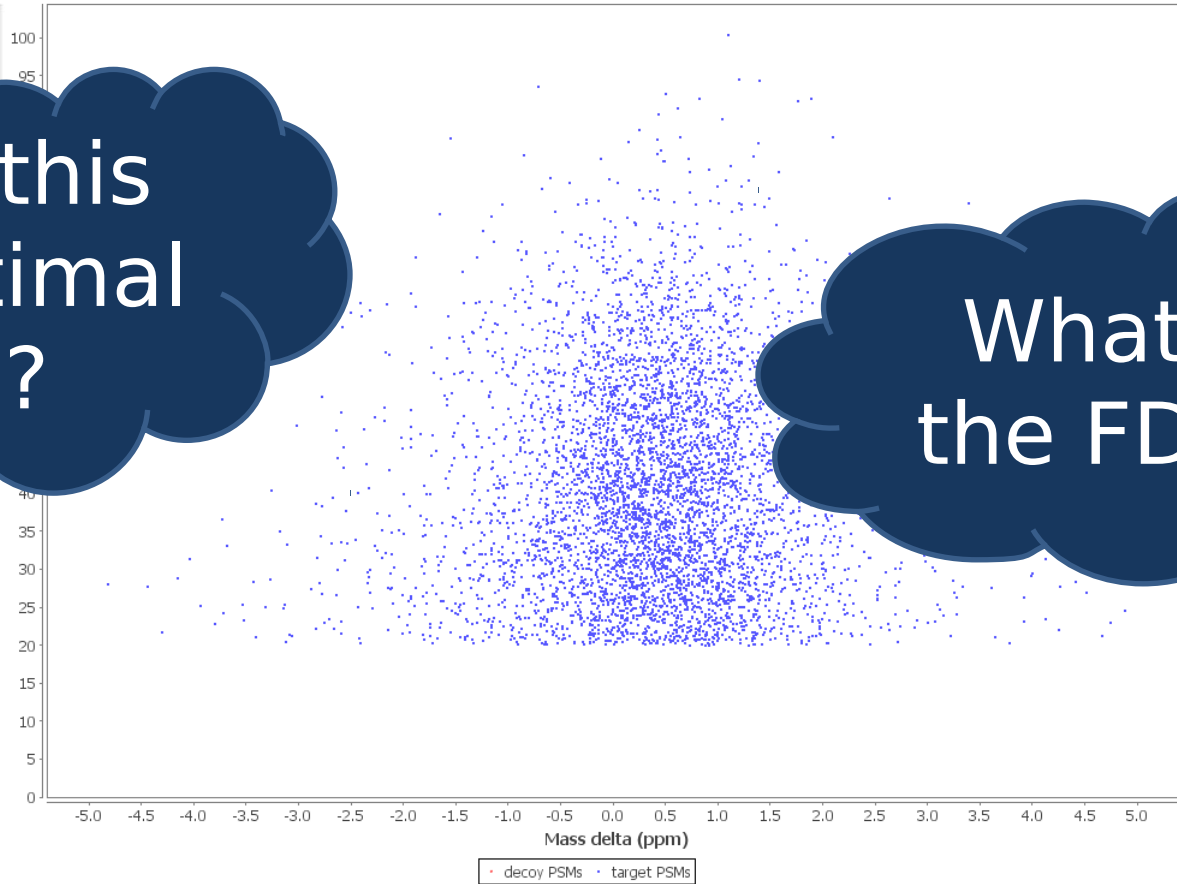
Manual filtering



Manual filter results

Mascot ions score over Mass delta (ppm) plot of F274658.dat.db

Score versus Mass Delta chart of F274658.dat.db



Is this optimal ?

What's the FDR?

18853 → 5335

What's an FDR

- False Discovery Rate
- The FDR is the *proportion of matches in the result set, expected to be false*
 - Usually a percentage

FDR estimation methods

T_s = Accepted target (known) sequences

D_s = Accepted decoy (nonsense) sequences

$$\text{FDR}_s \approx D_s / (T_s + D_s)$$

Competitive

- Decoy and target sequences combined in one database
- A spectrum matches either a decoy or a target sequence

Non-competitive

- Search separate Decoy and Target databases
- A spectrum can match both decoy and target sequences

Automatic FDR based filtering

FDR based filtering

FDR based filter

Target FDR

Enter the target FDR value %

Split

Apply filter to separate mass spectrometry runs (combined input file)

Decoy

decoy method

Additional filters

Only use first ranking PSMs

Use mass window

massdelta

Precursor mass between and ppm.

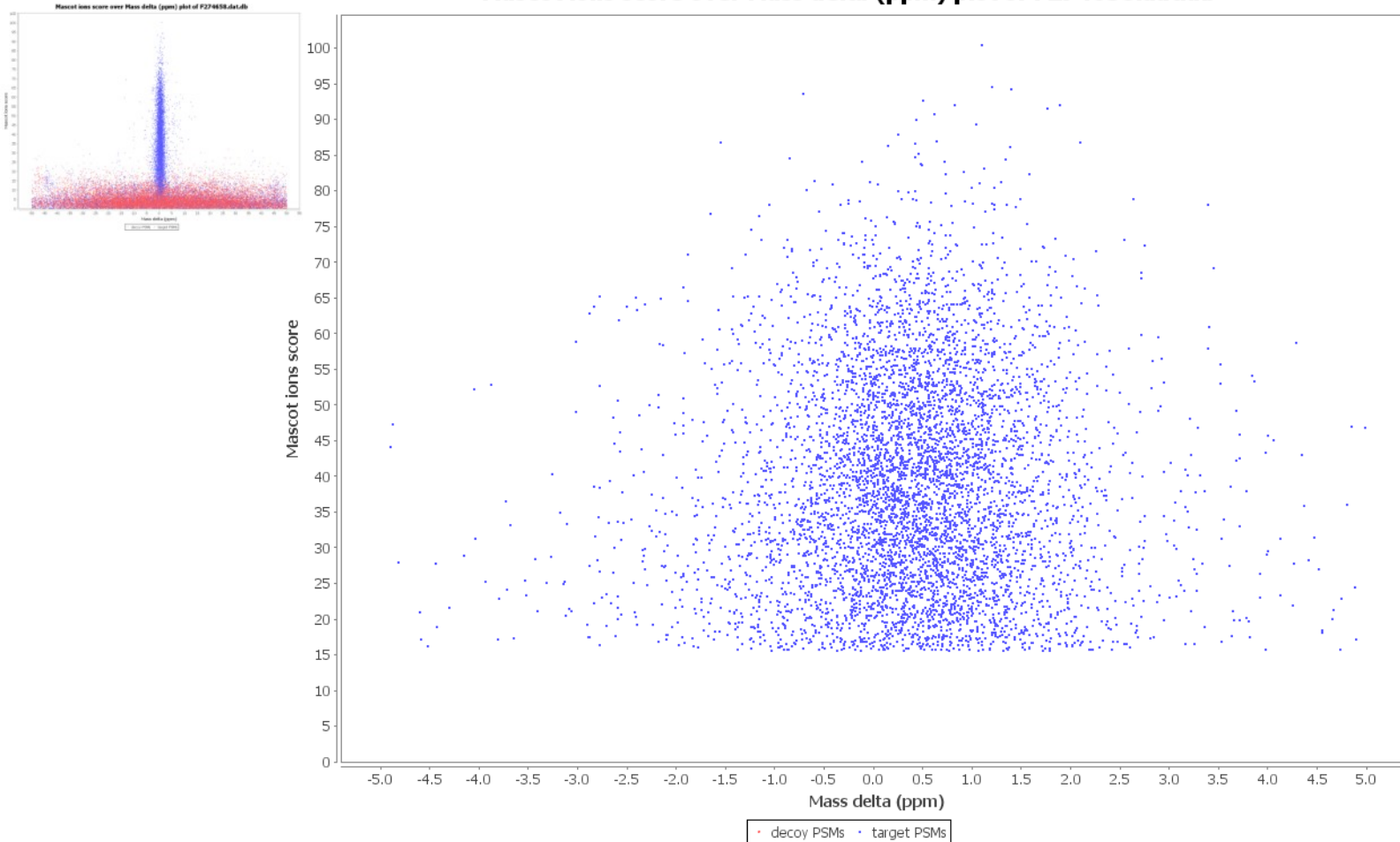
filter order (and logic)

Export method

- FDR guaranteed
- 50% Automatic
- Mass window
- Different decoy strategies
- Possibility to use on separate mass spec runs

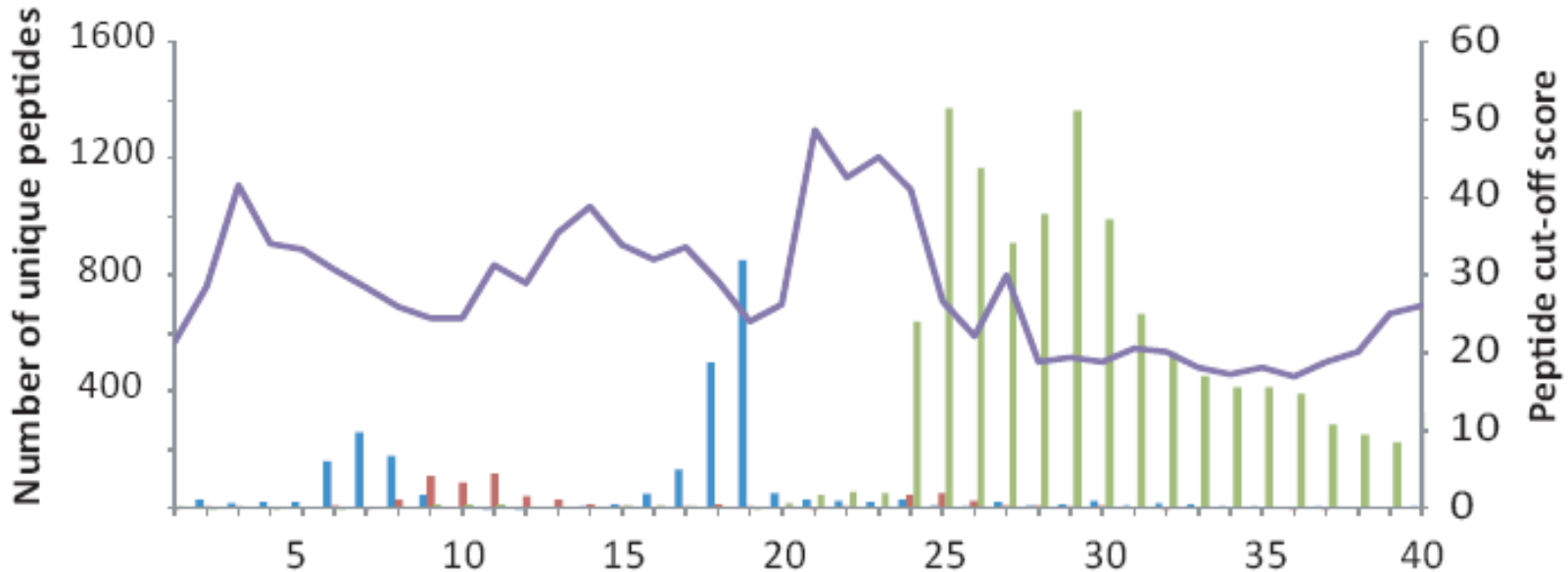
FDR based filtered file

Mascot ions score over Mass delta (ppm) plot of F274658.dat.db



cutoff: 15.6
18853 → 5880

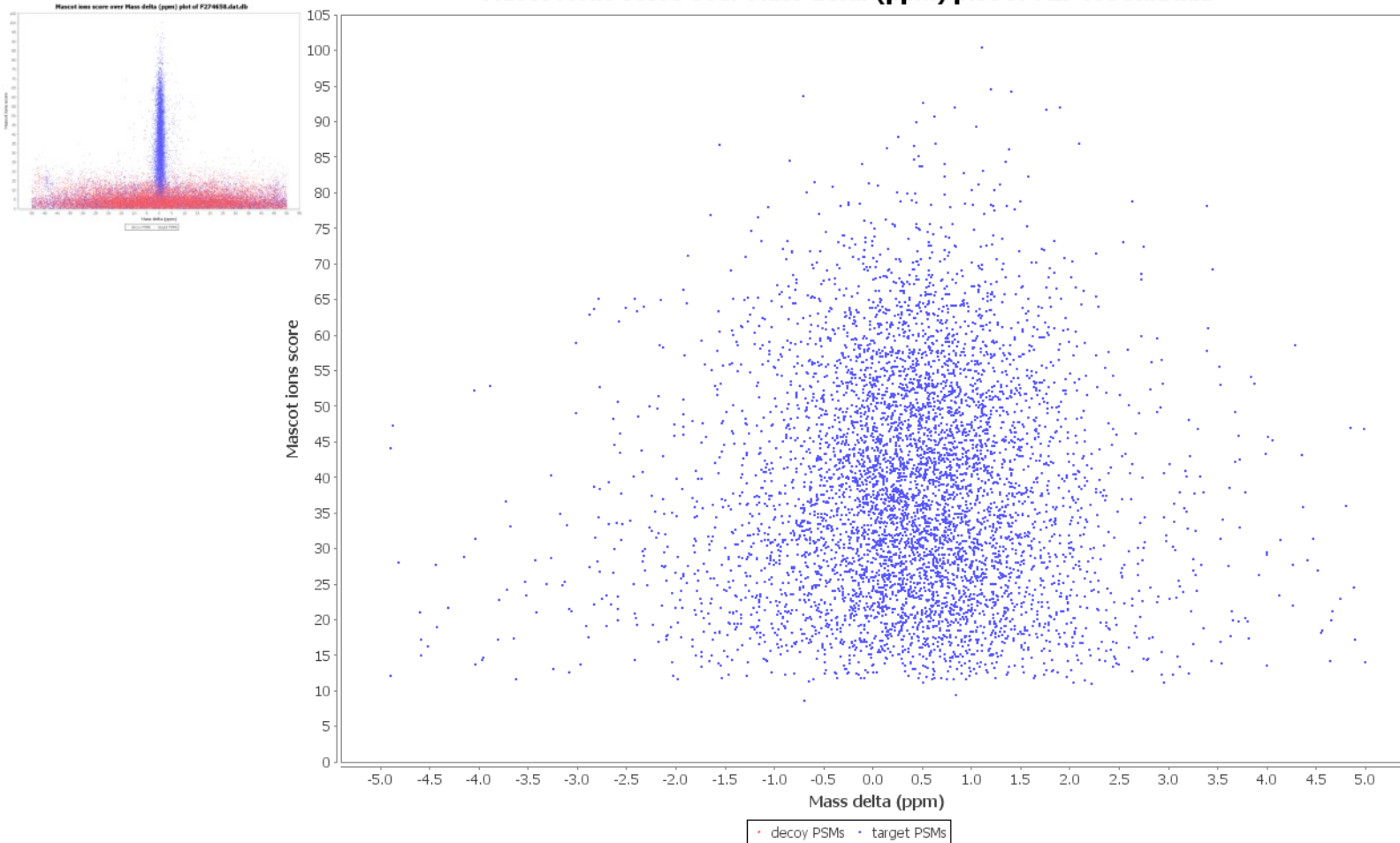
Fractions are not the same...



2 P N-Ac 1 P increasing # basic residue

FDR based filtered file

Mascot ions score over Mass delta (ppm) plot of F274658.dat.db

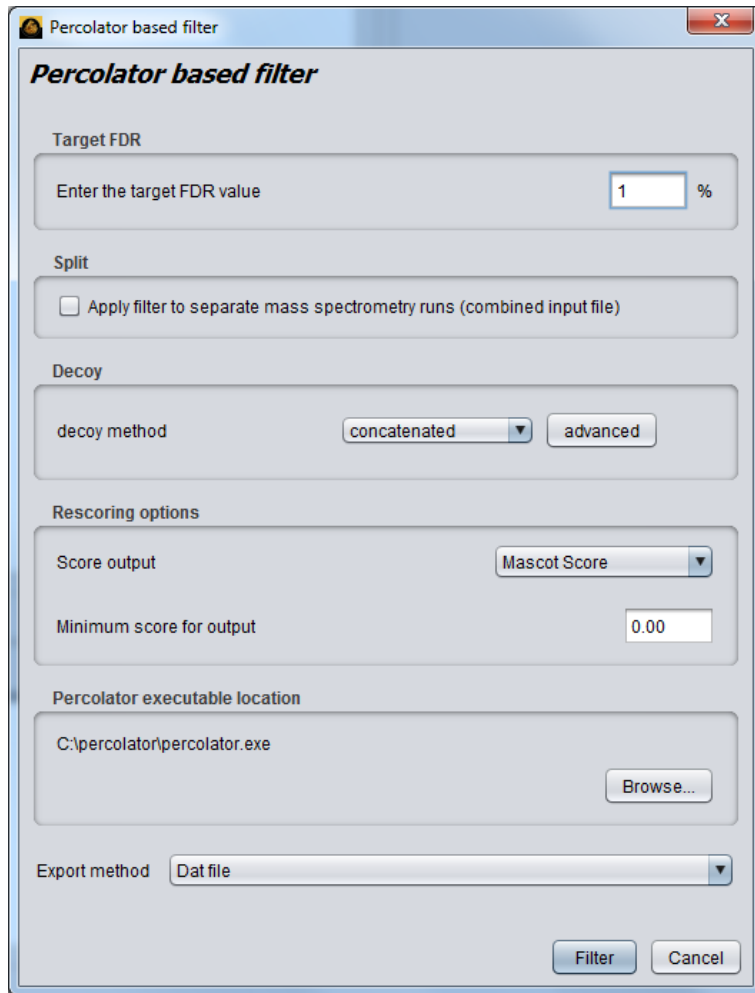


18853 → 5992

PSM properties

	Feature	Description
Score	id	Identifier. RockerBox uses the form *db*_querynumber_rank, in which *db* may be 'target' for Mascot automatic decoy real database, 'decoy' for Mascot automatic decoy scrambled database or 'combined' for a concatenated decoy strategy
	label	-1 if decoy, 1 if target PSM
	charge	Precursor charge
	score	Mascot score
	deltaScore	Difference between current rank score and 'next' rank score
Mass	mr	Measured precursor mass
	deltaM	Delta mass between precursor mass and matched peptide mass
	deltaMPpm	deltaM relative to matched peptide mass
	absDeltaM	Absolute value of deltaM
	absDeltaMPpm	Absolute value of deltaMPpm
	isoDeltaM	Delta mass allowing for 1, 2, 3 or 4 Dalton difference
	isoDeltaMPpm	isoDeltaM relative to matched peptide mass
Fragment matching	missedCleavages	Number of missed cleavages
	fragMassError*	RMS error of the MS2 spectrum to the theoretical spectrum
	totalIntensity*	Total intensity of the MS2 spectrum
	intMatchedTot*	Total intensity of matched MS2 peaks
	relIntMatchedTot*	intMatchedTot divided by totalIntensity
	fractionsMatched*	Fraction of all MS2 peaks matched
	peptide proteins	Peptide sequence The list of proteins from the search database that contain the peptide sequence

Using the Percolator algorithm



The screenshot shows a dialog box titled "Percolator based filter". It contains several sections for configuring the algorithm:

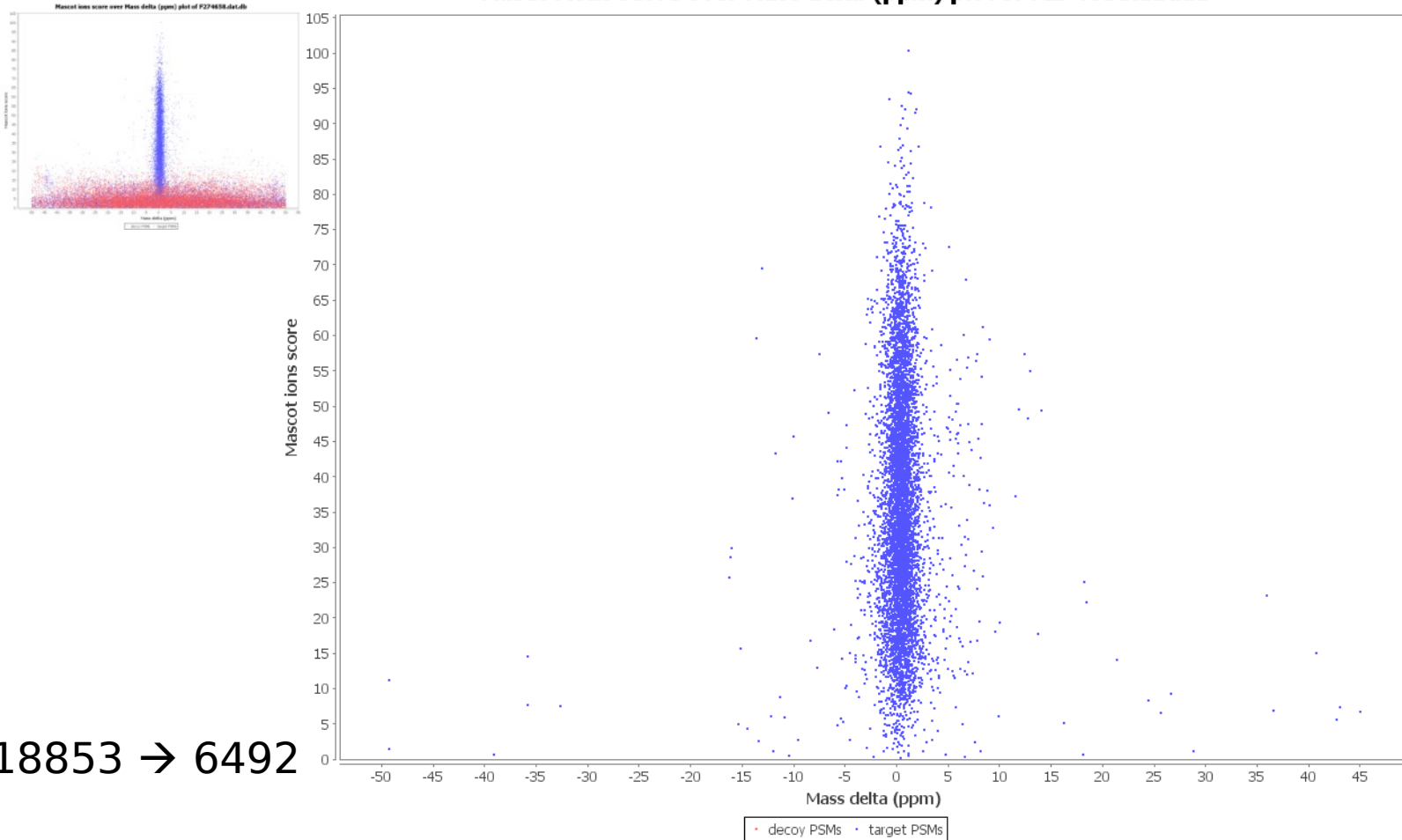
- Target FDR:** A text input field with "1" and a percentage sign.
- Split:** A checkbox labeled "Apply filter to separate mass spectrometry runs (combined input file)".
- Decoy:** A dropdown menu for "decoy method" set to "concatenated" and a button labeled "advanced".
- Rescoring options:** A dropdown menu for "Score output" set to "Mascot Score" and a text input field for "Minimum score for output" set to "0.00".
- Percolator executable location:** A text input field with "C:\percolator\percolator.exe" and a "Browse..." button.
- Export method:** A dropdown menu set to "Dat file".

At the bottom, there are "Filter" and "Cancel" buttons.

- Fully automatic
- Both decoy strategies
- Different score outputs
- Apply to separate spectrometry runs

Percolator filtered file

Mascot ions score over Mass delta (ppm) plot of F274658.dat.db



Overview of filtering methods

File	Size	Number of PSMs
Original .dat	95 MB	18853
Manually filtered	12 MB	5335
FDR 1%	11 MB	5880
FDR 1% per fraction	15 MB	5992
Percolator FDR 1%	15 MB	6671

PHOSPHORYLATION SITE COUNTS

Use case

Experiment

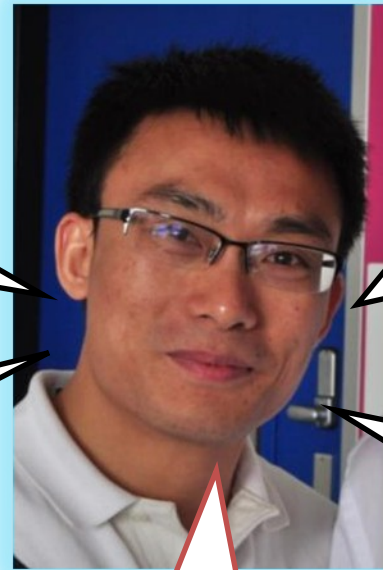
- Separate distinct peptide populations using SCX
- Enrich Phosphopeptide using Ti-IMAC
- Test in multiple fragmentation methods

Use case: phosphorylation counting

- Houjiang Zhou:

Which unique sites (protein level) do I have?

Maximize the number of phosphopeptides identified



which fragmentation method is best for phosphopeptides?
CID, ETD, HCD

Reliable data with known FDR

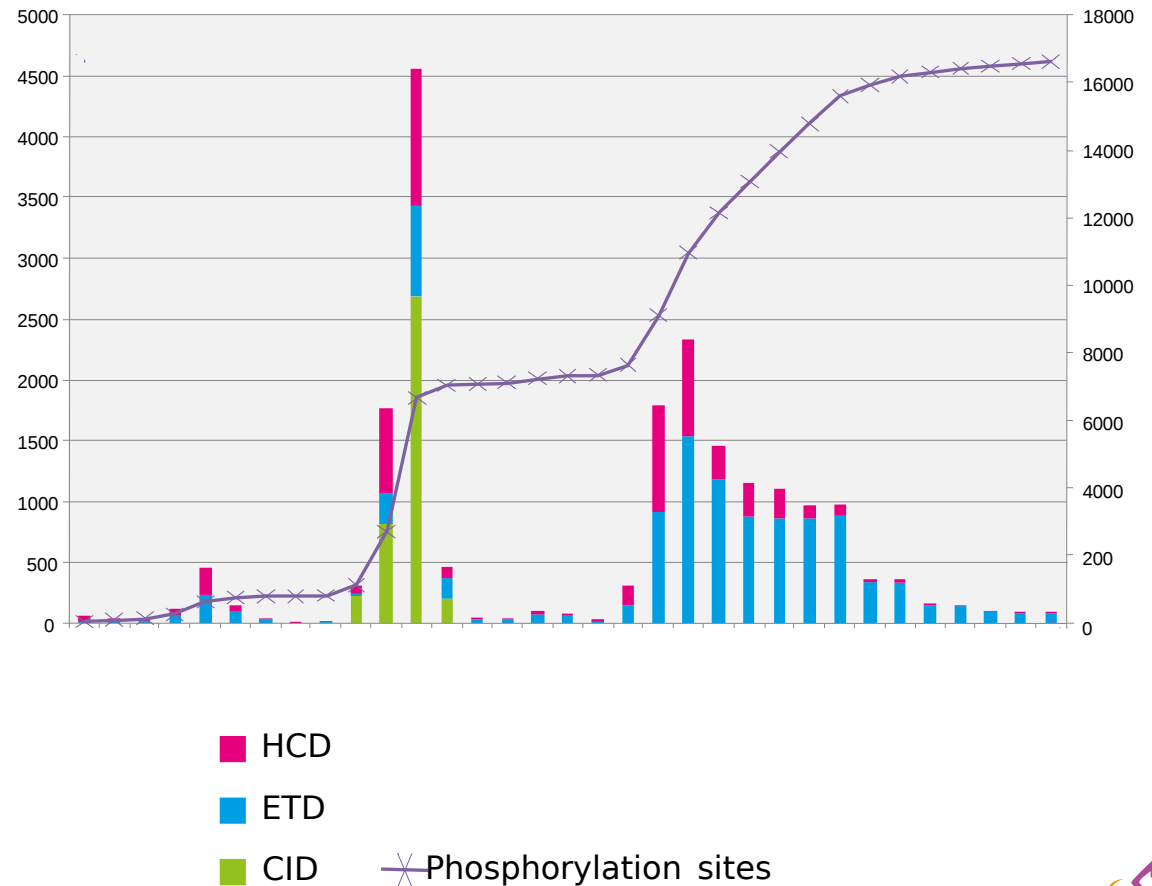
I need it tomorrow, the reviewer is waiting

Data analysis (WIP)

- Use RockerBox to filter the .dat file
 - using Percolator, FDR 1%, with a minimum Mascot score of 20
- Extract *PTM delta scores* using RockerBox csv export
- Count the phosphorylated peptides and phosphorylation sites on the proteins

Results

19,692 uni.phosphopeptides
16,624 uni.phosphosites
3862 phosphoproteins



Conclusions

- RockerBox helps to alleviate size problems
 - Complex research problems can be addressed more easily

Acknowledgements

{MATRIX}
{SCIENCE}

Lukas Käll

Javier Muñoz, Reinout Raijmakers,
Shabaz Mohammed

Bas van Breukelen, Albert J.R. Heck



Universiteit Utrecht

netherlands
proteomics
centre



nbic

netherlands
bioinformatics
centre

Thank you



<http://www.hecklab.nl>



netherlands
proteomics
centre



Availability

<http://trac.nbic.nl/rockerbox>