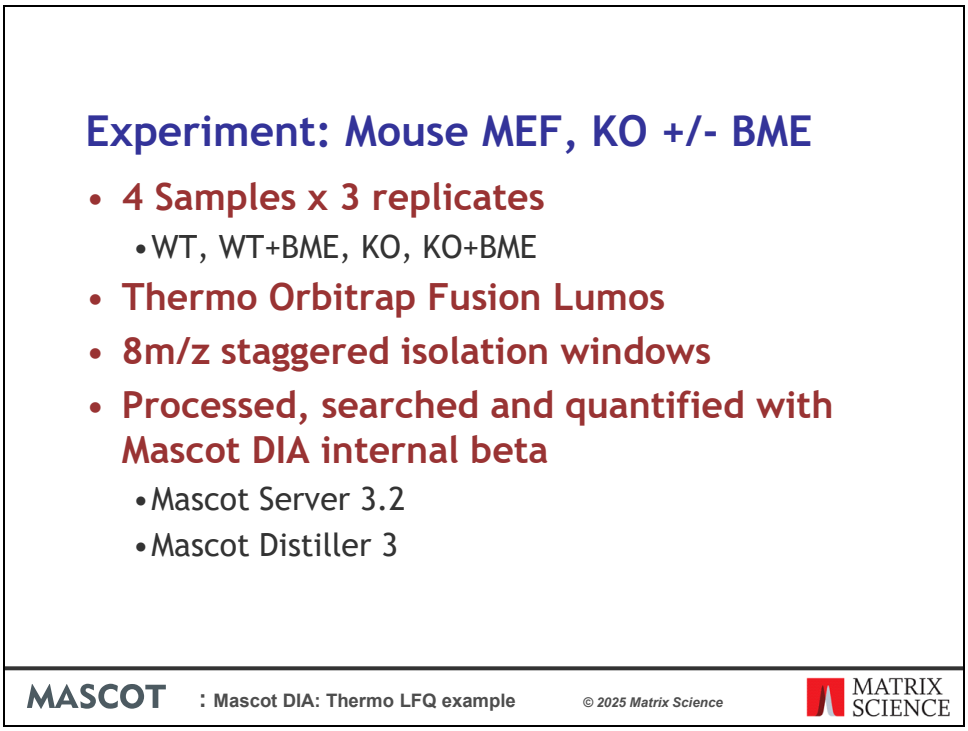
Mascot DIA: Thermo LFQ example

Patrick Emery, Matrix Science

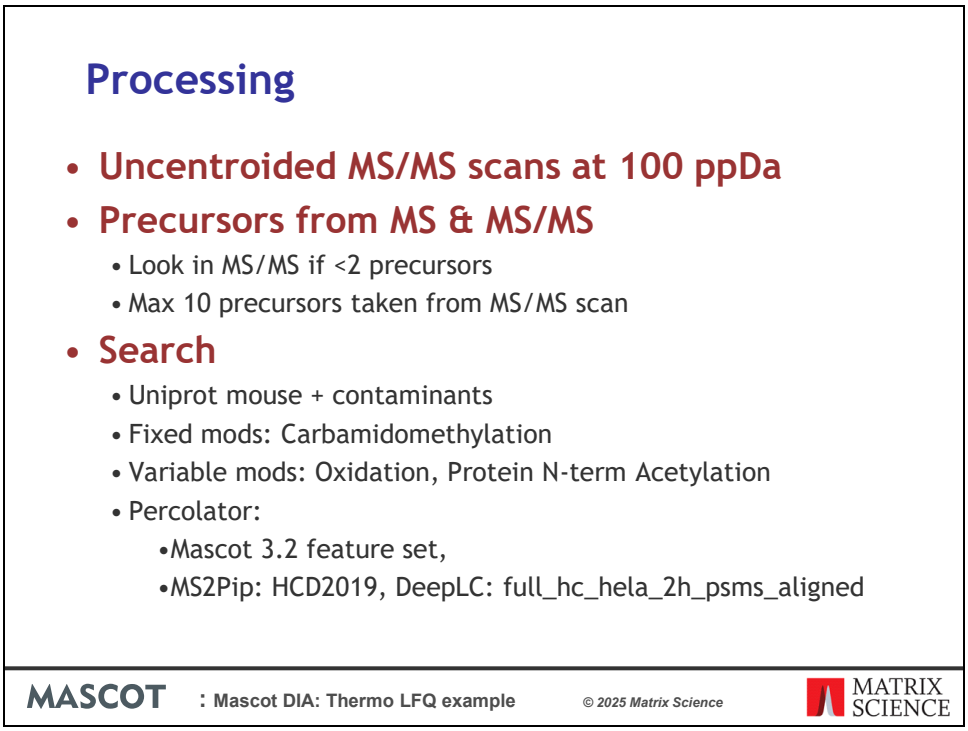
In this presentation I'll take you through an example of processing a Thermo Orbitrap label free quantitation dataset with Mascot DIA



The dataset was shared with us by Professor Sue Weintraub of the **University of Texas Health Science Center at San Antonio**

It's a Mouse Embryonic Fibroblast sample, wild type and knock out cells grown in the presence or absence of beta-mercaptoethanol, each sample run with 3 replicates for a total of 12 raw files.

The isolation window strategy was the commonly used 8 Thomson staggered isolation windows approach. We reprocessed the data using the internal beta builds of Mascot DIA



The raw files were processed as follows:

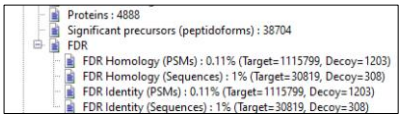
MS/MS scans had been captured as centroids, which were uncentroided in Mascot Distiller at 100 points per Da. For each MS/MS scan, candidate precursors were taken from the survey scan. If less than 2 precursors were identified in the survey scan then the software looked for additional precursors in the MS/MS scan using the complimentary ion pairs strategy outlined in the previous presentation. A maximum of 10 precursors were taken from the MS/MS scan, with precursors ranked by the number of times they were observed and the total intensity of the complimentary fragment ions

The peaklists were then searched on Mascot 3.2 with against the Uniprot mouse proteome and a contaminants database. Carbamidomethylation was selected as a fixed mod and Oxidation and Protein N-terminal acetylation as variable mods.

Results were refined using machine learning using the Mascot 3.2 percolator feature set and the following MS2Rescore models – HCD2019 for MS2Pip and full_hc_hela_2h for DeepLC

The fragment ions were decharged to MH+ to deconvolute the peaklists somewhat.

Results



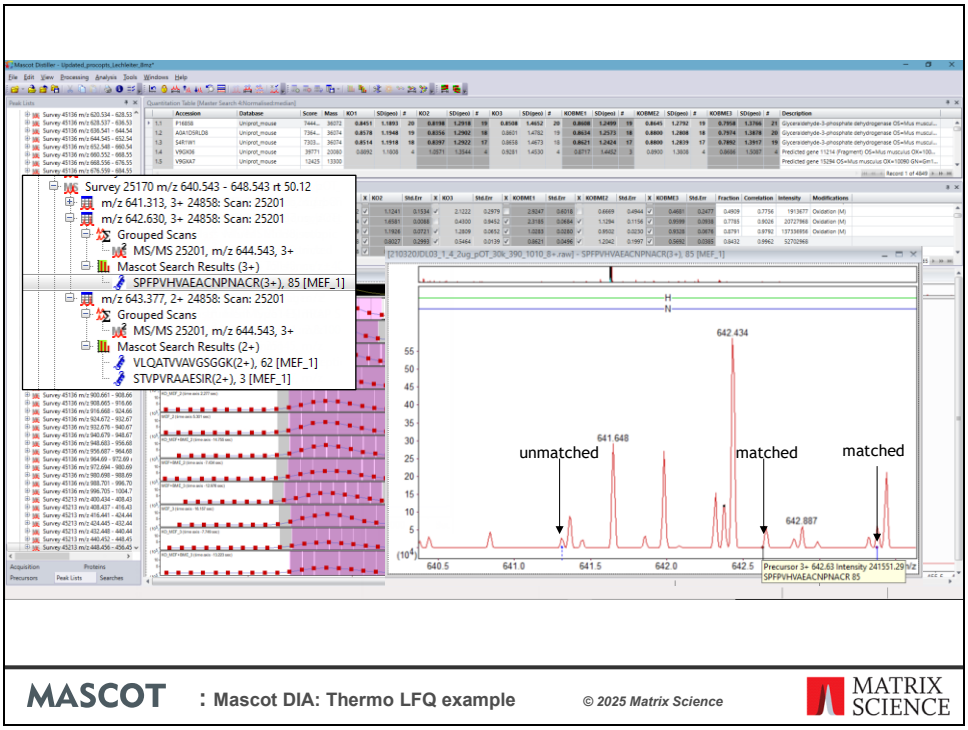
- **1% sequence FDR:**
  - 4888 proteins
  - 38,704 precursors/30,819 sequences/1,115,799 PSMs
  - 659,912/1,191,327 (55.4%) scans have 1 or more significant matches
- **Label free quantitation:**
  - Replicate (match between runs enabled)
  - KO/WT * 3, KO+BME/WT+BME * 3

MASCOT : Mascot DIA: Thermo LFQ example MATRIX SCIENCE

Processing and searching each file took about an hour on the system I used. Results were as follows:

Using a 1% peptide sequence false discovery rate filter, we found 4888 proteins from 38704 different precursors and 30819 peptide sequences with a total of ~1.1 million PSMs. This PSMs were distributed over ~660k scans (with scans having between 0 and 7 significant precursor matches), so just over 55% of the MS/MS scans in the dataset have 1 or more significant matches.

Results were imported into Mascot Distiller for label free quantitation using the “Replicate” method, which enables match between runs.

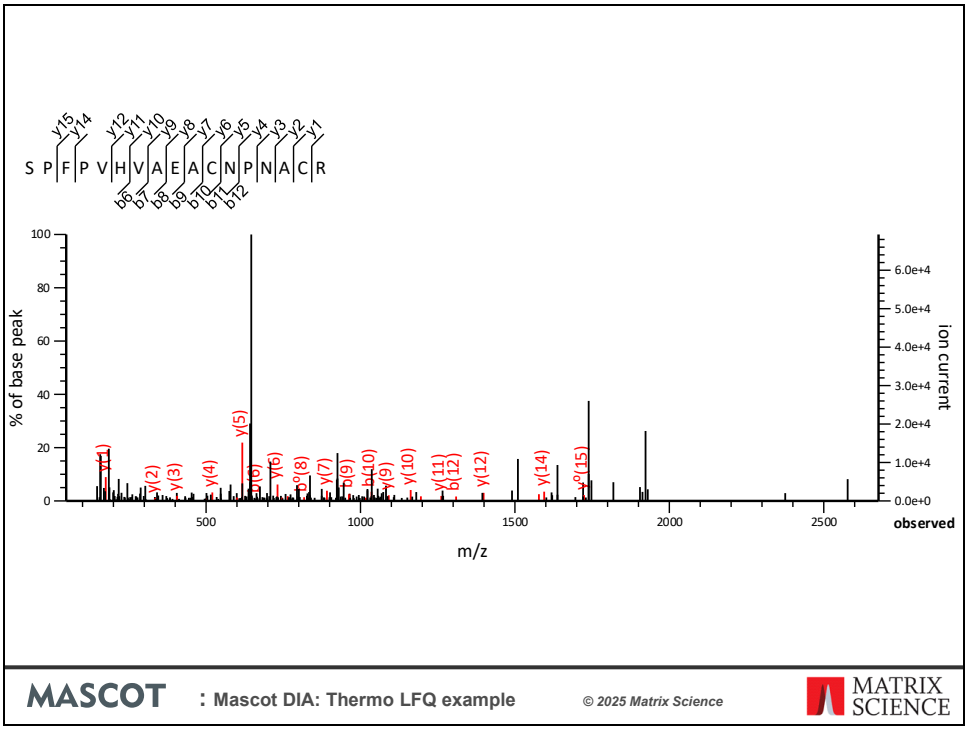


Once the data are processed, searched and quantified it doesn't look very different to a standard DDA dataset in Mascot Distiller.

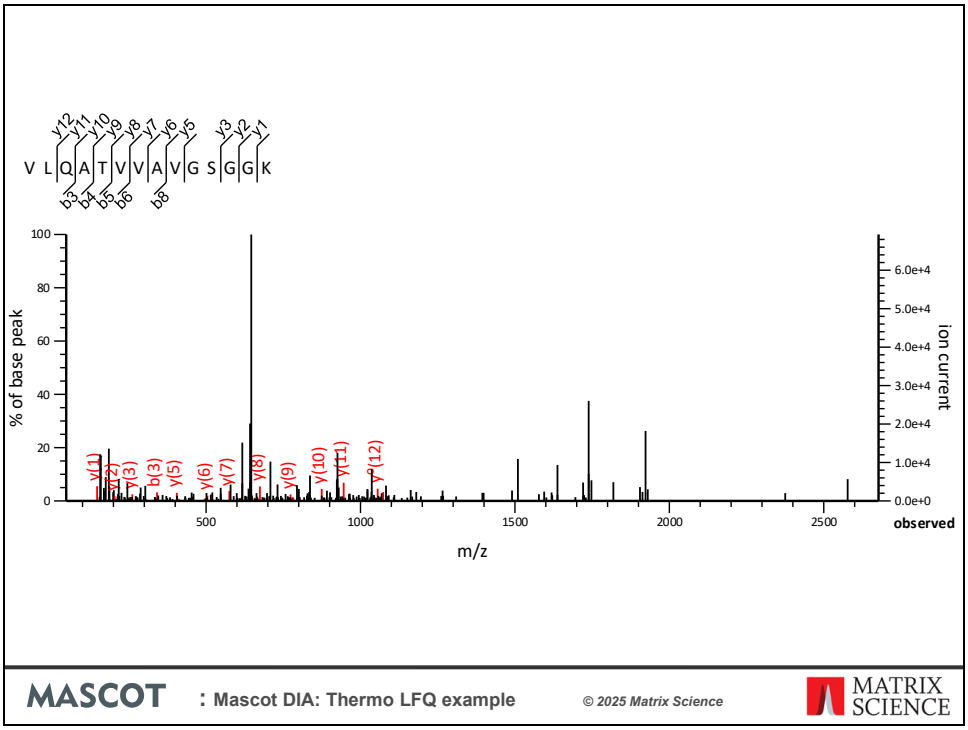
In the bottom left panel here we have the XICs for a quantified peptide – the grey bars are where we found significant matches to the peptide and as you can see, in this case we identified the peptide in all the samples and across the XIC peaks.

The peaklists tree is organised slightly differently, with the survey scan and isolation window grouping together the peaklists for the precursors assigned to that scan and window. In this case, 3 precursors were detected and we have significant matches to two of them.

This is highlighted on the MS/MS scan showing the highlighted peptide match – the markers below the scan trace showing the m/z values for the precursors. The one with a dotted line was not matched while the two with solid lines did get significant matches in the search. Hover over them for a tool tip showing the precursor and match details



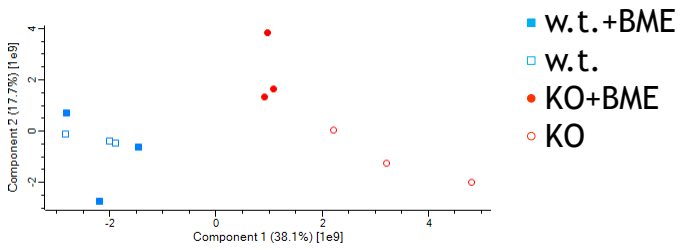
This is the match to the 3+ precursor from the scan



And this is to the matched 2+ precursor – as you can see for these two peptides the matches are using completely different fragment ions

Quantitation Results: PCA analysis

Protein intensity values exported to Perseus



MASCOT : Mascot DIA: Thermo LFQ example MATRIX SCIENCE

To confirm the quantitation had behaved as expected, I exported the normalised protein intensities for each sample into Perseus from the Max Planck institute and ran a principle component analysis.

As you can see, the WT and KO samples separate quite clearly on component 1. In the knock out samples, the samples grown in the presence or absence of beta mercaptoethanol also separate on components 1 and 2 while there doesn't seem to be so much difference in the WT samples.

There are indeed some proteins showing significant fold changes in the knock out samples between the presence and absence of bme, but, as you'd expect to find, not as many as between the wt and knock out samples.

For sake of simplicity, we'll concentrate on the wild type to knock out in the absence of bme from now on.

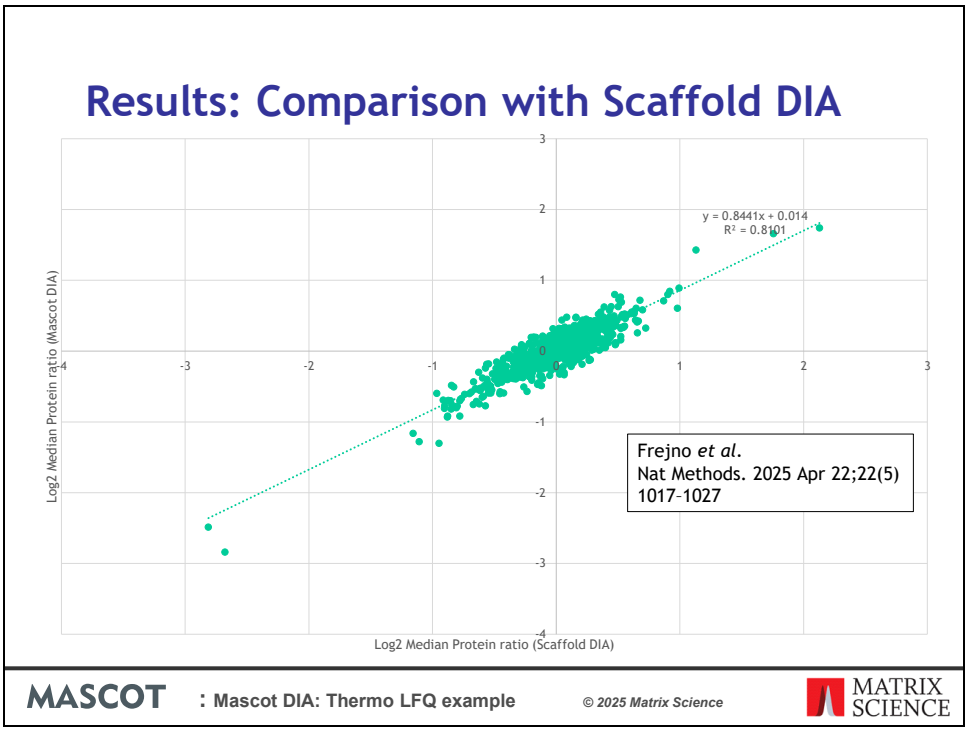
Quantitation Results: Comparison

- **Arbitrarily selected ~1000 proteins identified for all 3 KO/WT replicates in both Mascot DIA and Scaffold DIA**
  - There are far more shared proteins but the accession and grouping is different between the two packages
- **Calculated median protein ratio from the three KO/WT replicates in each case**

The next step was to compare the quantitation results from Mascot DIA with the original analysis carried out by Professor Weintraub using Scaffold DIA.

This is still work in progress - because of the differences in accession selection same/subset differences between the two analyses, this is a little tricky, so we arbitrarily selected around 1000 protein accessions which matched between Mascot and Scaffold where we had quantitation results for all 3 replicates.

Then for each group we calculated the median protein ratio and median absolute deviation was calculated for KO/WT.

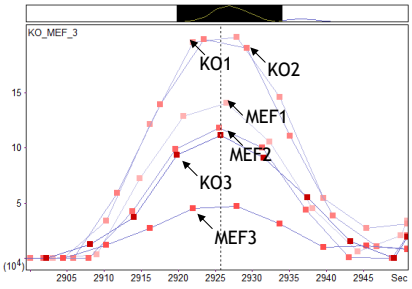


If we plot the ratios for these proteins and get Excel to calculate the line of best fit, there's a very strong correlation between the two software packages result with an R-squared on 0.81

We're not the first to find a good correlation between precursor and MS/MS based quantitation. Frejno et al. also found a strong correlation, with a very similar R-squared value, between MS1 based quantitation using Minora in Thermo PD and the MS/MS based quantitation in Chimerys on a different dataset.

Modifications: Error Tolerant Search

Modification	Delta	Type	Site	Total matches
Oxidation	15.994915	variable	M	5370
Carbamidomethyl	57.021464	fixed	C	4615
Acetyl	42.010565	variable	Protein N-term	2083
Non-specific cleavage		ET	-	614
Formyl	27.994922	ET	N-term	357
Trioxidation	47.984744	ET	C	341
Deamidated	0.984016	ET	N	320
Gln->pyro-Glu	-17.026532	ET	N-term	169
Deamidated	0.98401			
Phospho	79.9663			



MASCOT : Mascot DIA: Thermo LFQ example MATRIX SCIENCE

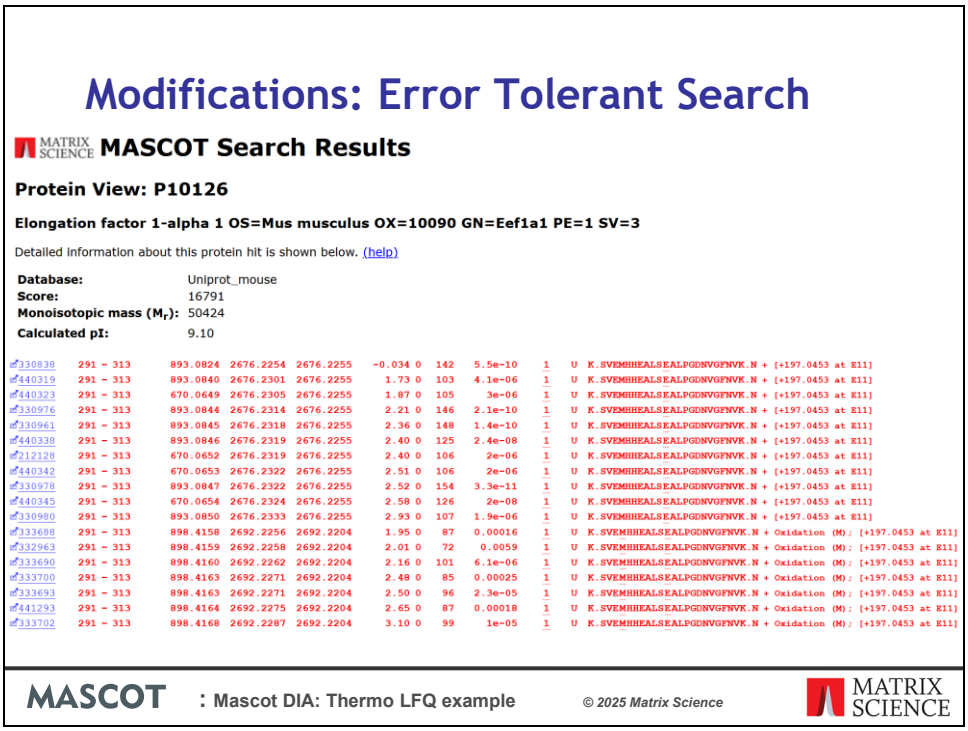
So, we have a good correlation between our results and those from Scaffold DIA, but while the results have good specificity the spectrum centric approach does have lower sensitivity than a peptide centric approach – because to get a match to a given peptide sequence we need a higher level of direct evidence in the peaklists.

As Ville mentioned though, that does also have advantages. Like a DDA search, we're not dependent on spectral libraries – either generated from DDA data or using one of the machine learning tools. You can search using any enzyme definition, using any modification against any FASTA database.

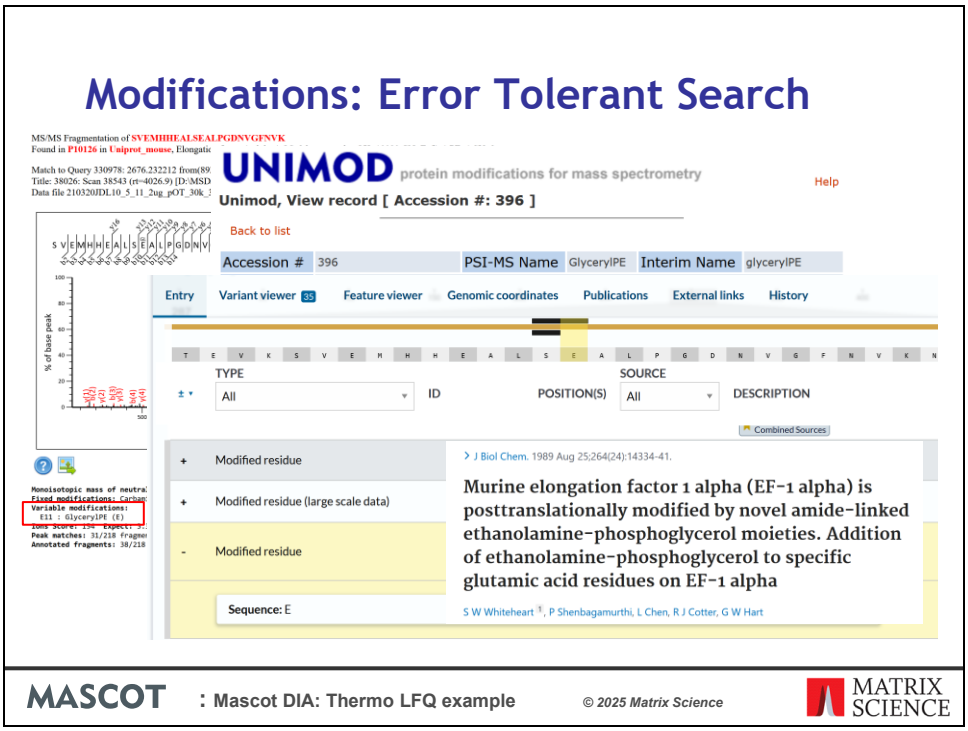
As an example of that, I carried out an error tolerant search on one of the wild type data files to look for unsuspected modifications, amino acid substitutions or non-specific cleavage products

If we take a look at the results from this, we can see the error tolerant pass found 614 semi-tryptic matches and then things like formylation which will have happened during sample preparation. Other commonly found modifications are deamidation, Gln->Pyro-Glu. A bit further down the table we have 80 phospho serines identified. I did then go back and repeat the analysis of the entire dataset with Phospho ST selected and here's an example of a doubly phosphorylated peptide which was originally identified by the error tolerant search and which is ~twice as abundant in the knock

out sample than the wild type



The error tolerant search is also an effective way to pick up less common PTMs and sequence variants. Here we have the matches to Elongation factor 1-alpha 1, and we have a whole series of matches identified with a +197.0453 at glutamic acid 301 – in fact we don't have the peptide sequence matched without the modification.



If we take a look at the peptide view that mass shift is suggested to be GlycerylPE. This is it's Unimod entry. GlycerylPE is a shorthand name for glycerylphosphorylethanolamine. If we take a look at the Uniprot entry for mouse Elongation factor 1-alpha 1, we can see that this is a known and published modification, but one which is unlikely to have been in the training data used by Prosit or similar tools.

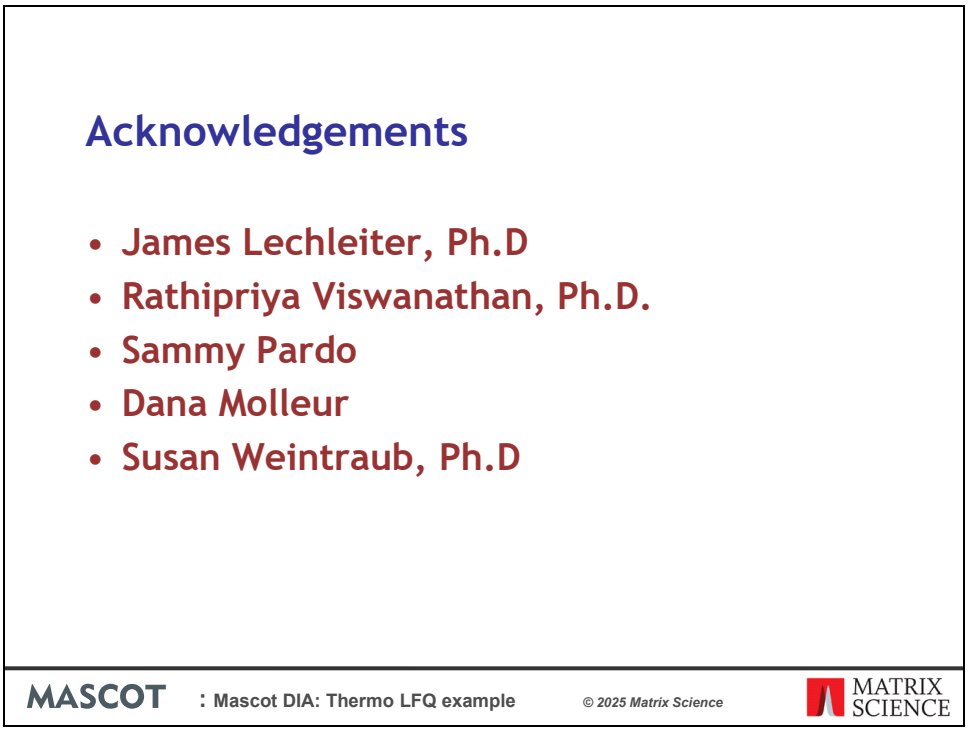
In summary

- 12 Thermo raw files processed and searched using Mascot DIA
- Results obtained are in agreement with the previous analysis
  - Including quantitation
- Error tolerant search identified
  - Semi specific cleavage
  - Processing artefacts
  - Additional PTMs

In summary, we reprocessed and search 12 Thermo raw files using Mascot DIA.

The results were in agreement with the previous analysis done in Scaffold DIA

A separate error tolerant search additionally identified semi-specific cleavage products, processing artefacts and additional unsuspected PTMs



And finally I'd like to thank

Professor James Lechleiter,

Dr Rathipriya Viswanathan

Sammy Pardo

Dana Molleur

Professor Susan Weintraub

From the **University of Texas Health Science Center at San Antonio** for sharing the data with us, and special thanks to Sue for the useful discussions we have about the data and experiment.