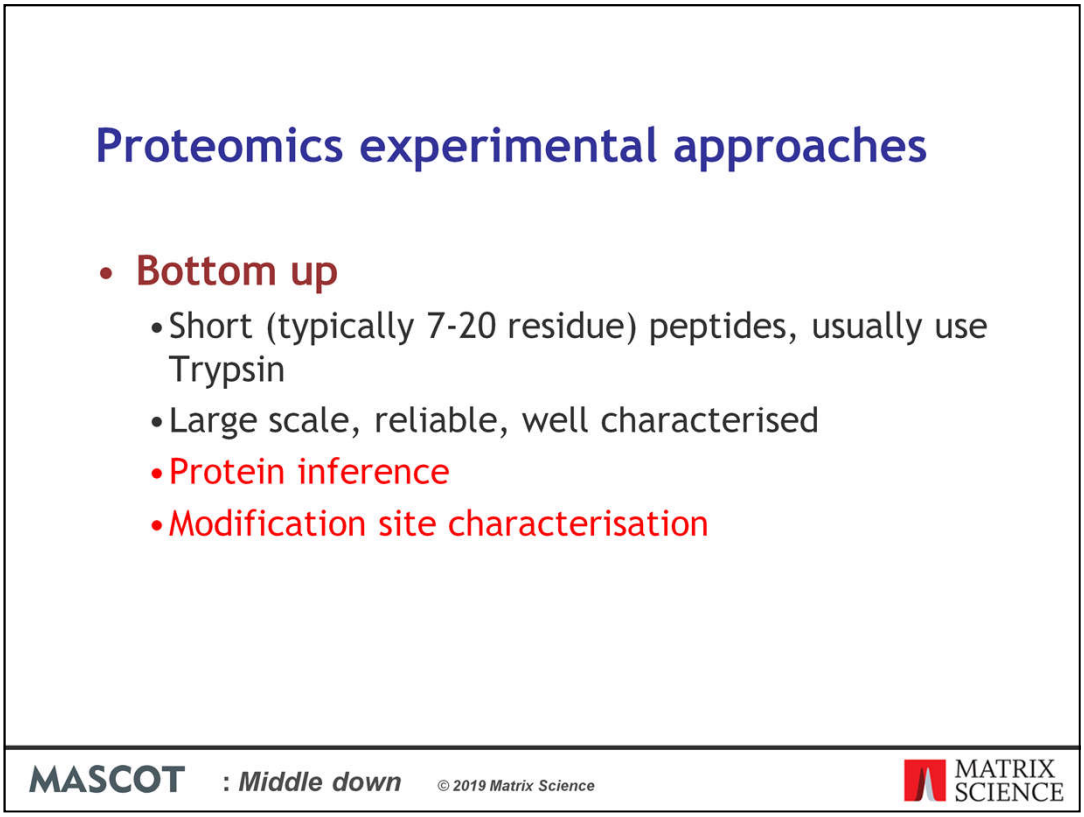
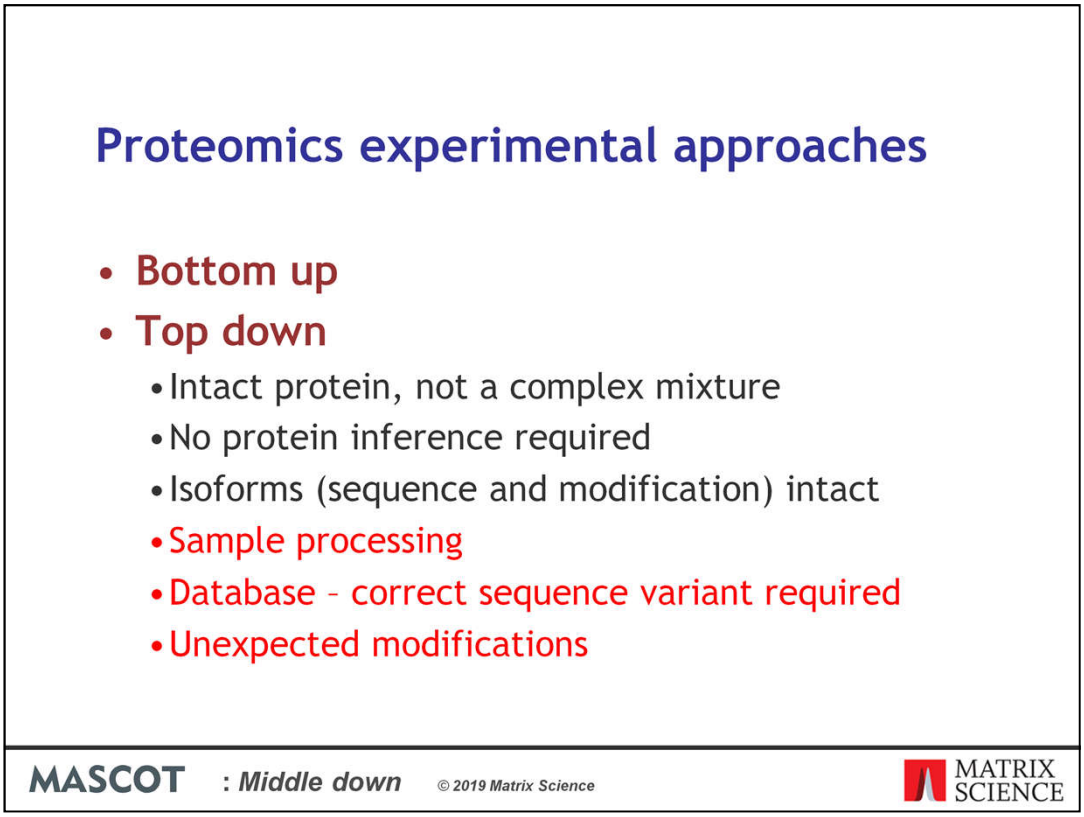
Trypsin isn't the only protease

Processing and searching middle down datasets with Mascot Distiller and Mascot Server



The most commonly used approach in proteomics is the bottom up experiment, where you're generally digesting a complex mixture with typically trypsin, to generate generally relatively short peptides. This is proven technology, tryptic peptides generally behave nicely in the MS and it works well for large scale experiments and is generally reliable – if you're trying to characterise what is present in a complex mixture, this is the most standard approach.

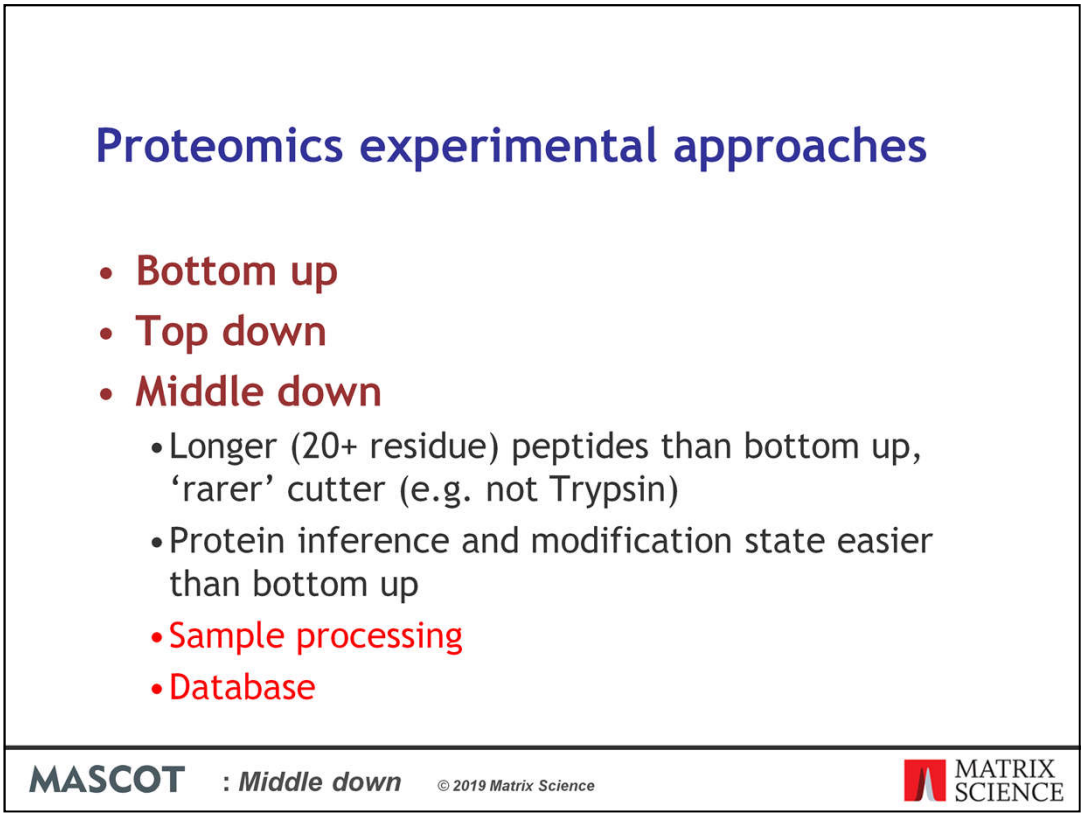
The downsides of the approach are that protein inference is tricky – you're trying to reconstruct which proteins are present in your mixture from the identified peptides, and there will almost always be multiple possibilities. The other issue is that because you're generally working on short peptides, you can't easily do detailed modification site characterisation – e.g. patterns of post translational modification across multiple sites on the intact protein.



At the other end of the spectrum is the top down approach. Here, you're taking the intact protein and carrying out MS/MS analysis directly on it. The big advantage of this is that you're maintaining the intact protein, so you can maintain your isoform information (both sequence and modification states).

Downsides – it isn't for complex mixtures, and your sample preparation and MS are generally more complex than for a shotgun bottom up experiment. When you carry out the database search, you need to have the correct sequence variant available and you need to have a good idea of what modifications to select for the search, or you won't get any matches back. Whereas with a bottom up experiment, you may miss some peptides, but you should still be able to get some information back.

If you're working on a well characterised organism, then database preparation is much less of a stumbling block than it used to be because the uniprot proteomes make it very easy to download known isoforms of a protein.

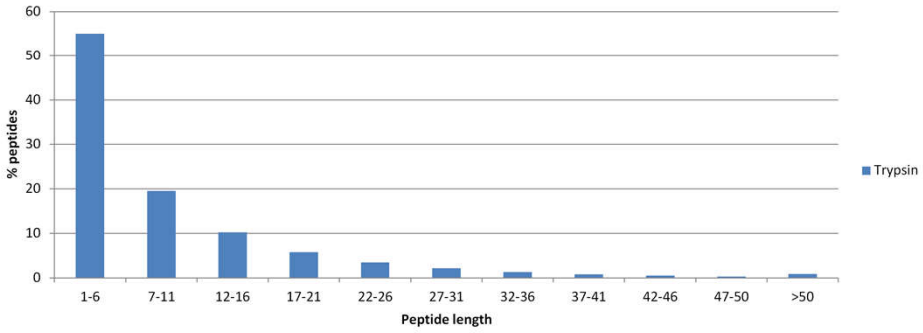


Middle down is an approach which tries to marry some of the advantages of the bottom up and top down methods. You use a cleavage agent which cuts proteins on a less frequently observed site than trypsin, to produce longer peptides – typically 20 or more residues. This means that protein inference and modification state characterisation are easier and more powerful than they are with bottom up approaches, and the experiment is generally simpler to carry out than a top down experiment (and better behaved in the MS) allowing for greater throughput and coverage.

Sample processing is generally trickier than bottom up (though easier than top down), and you do need to make sure you've got your database and search conditions set up correctly – you're more likely to completely miss a match due to a missing isoform that you are with bottom up. If you're characterising multiple peptides from the same protein, you obviously don't have the same potential power of modification state characterisation as you would do when analysing the intact protein.

Cleavage agents

Agent	Mean pep. length*	Mean pep. length >= 7
Trypsin	8 (~4 million)	16 (~1.8 million)

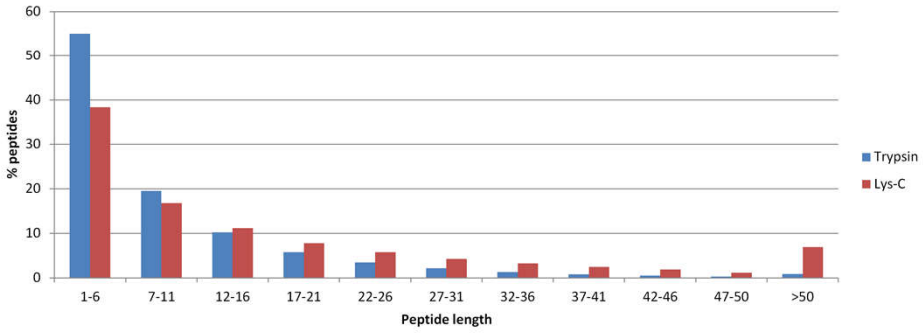


* In-silico digestion of Uniprot human sequences, no missed cleavages

In bottom up experiments, Trypsin is typically used as the protease to digest the sample. If we carry out a in-silico limit digest of protein sequences in the Uniprot human proteome, then the mean peptide length generated by Trypsin is 8 residues. Because getting a statistically significant match from a short peptide sequence is difficult, by default Mascot only looks for peptides which are 7 residues or longer. If we exclude peptides shorter than 7 residues from the calculation, the mean peptide length is now 16, but you've excluded approximately 55% of the theoretical peptides – fortunately, the loss of sequence coverage is nothing like as high as that at ~18.5%

Cleavage agents

Agent	Mean pep. length*	Mean pep. length >= 7
Trypsin	8 (~4 million)	16 (~1.8 million)
Lys-C	17 (~2.1 million)	26 (~1.3 million)



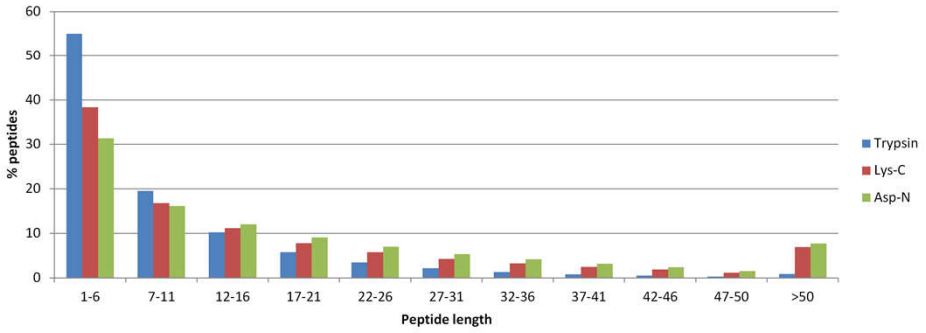
* In-silico digestion of Uniprot human sequences, no missed cleavages

Now lets take a look at some of the commonly used cleavage agents for middle down experiments. Lys-C produces a mean peptide length of 17, with a mean length of 26 if we're only looking at peptides of 7 residues or longer, and we're only losing ~38% of the theoretical peptides as too short when we exclude the short peptides, with 7% loss of sequence coverage.

As you can see from the graph, we are shifting the distribution of peptide sequence length upwards compared with Trypsin.

Cleavage agents

Agent	Mean pep. length*	Mean pep. length >= 7
Trypsin	8 (~4 million)	16 (~1.8 million)
Lys-C	17 (~2.1 million)	26 (~1.3 million)
Asp-N	19 (~1.9 million)	26 (~1.3 million)

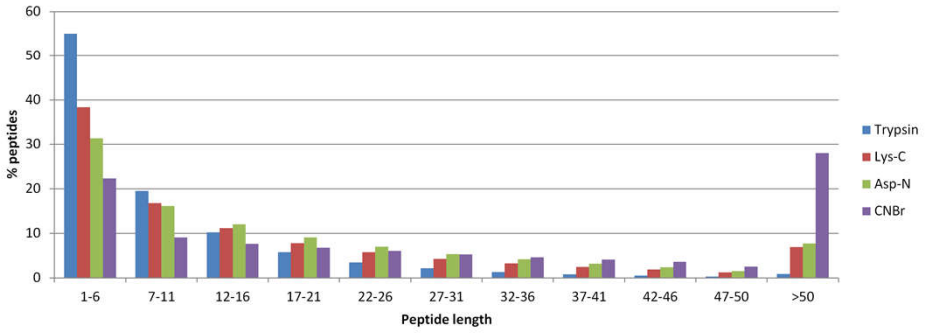


* In-silico digestion of Uniprot human sequences, no missed cleavages

Another commonly used protease in middle down experiments is Asp-N. The mean peptide length from this is 19 residues, and 26 again if we exclude the short peptides, this time only losing 32% of the theoretical peptides and 5% loss of sequence coverage. We're also shifting the peptide length distribution further upwards.

Cleavage agents

Agent	Mean pep. length*	Mean pep. length >= 7
Trypsin	8 (~4 million)	16 (~1.8 million)
Lys-C	17 (~2.1 million)	26 (~1.3 million)
Asp-N	19 (~1.9 million)	26 (~1.3 million)
CNBr	41 (~850k)	52 (~660k)



* In-silico digestion of Uniprot human sequences, no missed cleavages

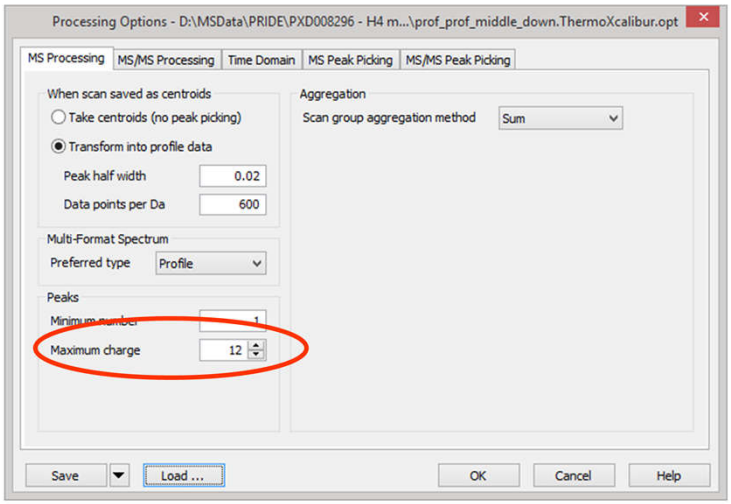
MASCOT : Middle down MATRIX SCIENCE

Sometimes chemical cleavage agents are used rather than a protease. CNBr gives us a mean peptide length now of 41, and 52 respectively – in fact nearly 30% of the theoretical peptides have a length greater than 50 residues. Excluding short peptides only excludes ~22% of our theoretical peptides, representing just 1.3% sequence coverage loss.

One practical consideration when handling either intact proteins or long peptides is the mass range of your instrument. A 26 residue peptide would typically have a mass in the range of ~2.9KDa, which would put even doubly charged data outside the detection range of many instruments. Therefore, with both Top and Middle down experiments, you do typically need to obtain higher charge state precursors than you would for a bottom up experiment.

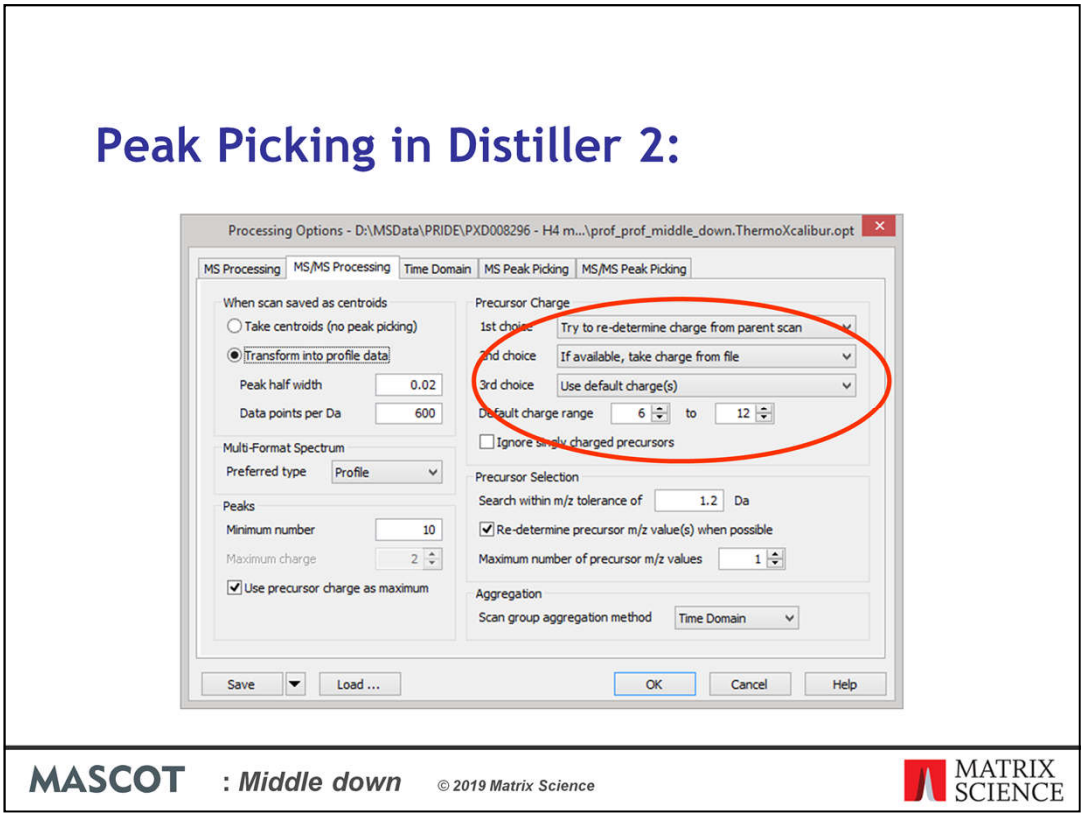
Another potential issue with some of the very long peptides would be exceeding the 16KDa precursor limit in Mascot. For example, approximately 4% of theoretical peptides from the CNBr digest are longer than 150 residues. These could exceed the precursor limit and require a Top Down licence in order to be searched.

Peak Picking in Distiller 1:



MASCOT : Middle down MATRIX SCIENCE

For a middle down experiment you will therefore normally be obtaining higher precursor charge states than for a bottom up dataset, but lower than those required for a top down experiment. To account for this in Mascot Distiller, you may want to take a look at your data to see what precursor charge state range is present and adjust the MS Processing maximum charge value. This value should be chosen carefully, however, as the processing time required increases proportionately to this value. Also, you should only look for higher charge states if the resolution of your instrument is sufficiently high to allow for this – charge state can only be reliably determined if there is some resolution between the isotope peaks.



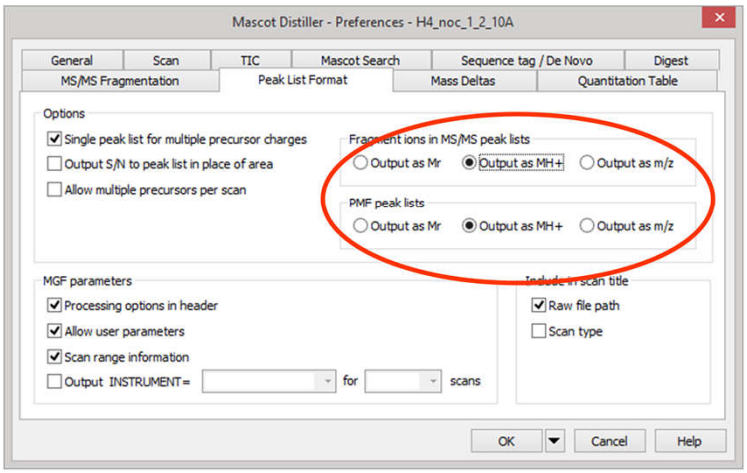
The other place in the processing options you need to take precursor charge into account is on the MS/MS processing tab. Here, you have a number of choices for determining the precursor charge of the MS/MS spectra, which are taken in order, falling through to a lower choice if the preceding method fails.

Here, my first choice is to determine the charge from the parent scan – whether this can work or not depends on the settings on the MS Processing tab we were looking at in the previous slide, and if there is sufficient resolution in the required region of the parent survey scan.

If that fails, the second choice is to try and take the charge state from the source raw file.

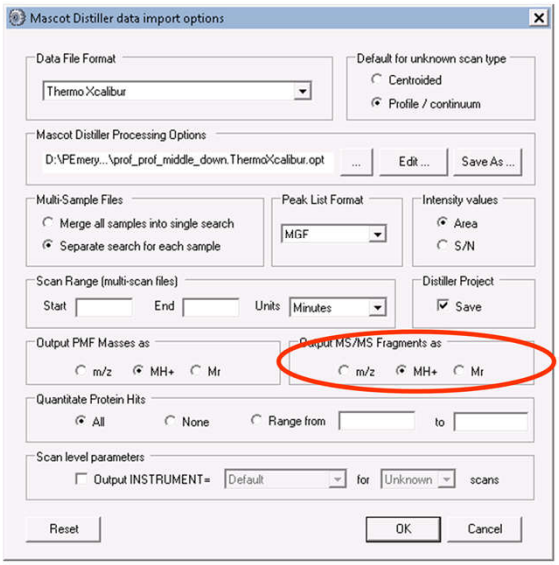
Finally, if the charge state is not available from the file, we'll output a range of possible charge states for the peak list.

Peak List Format 1:

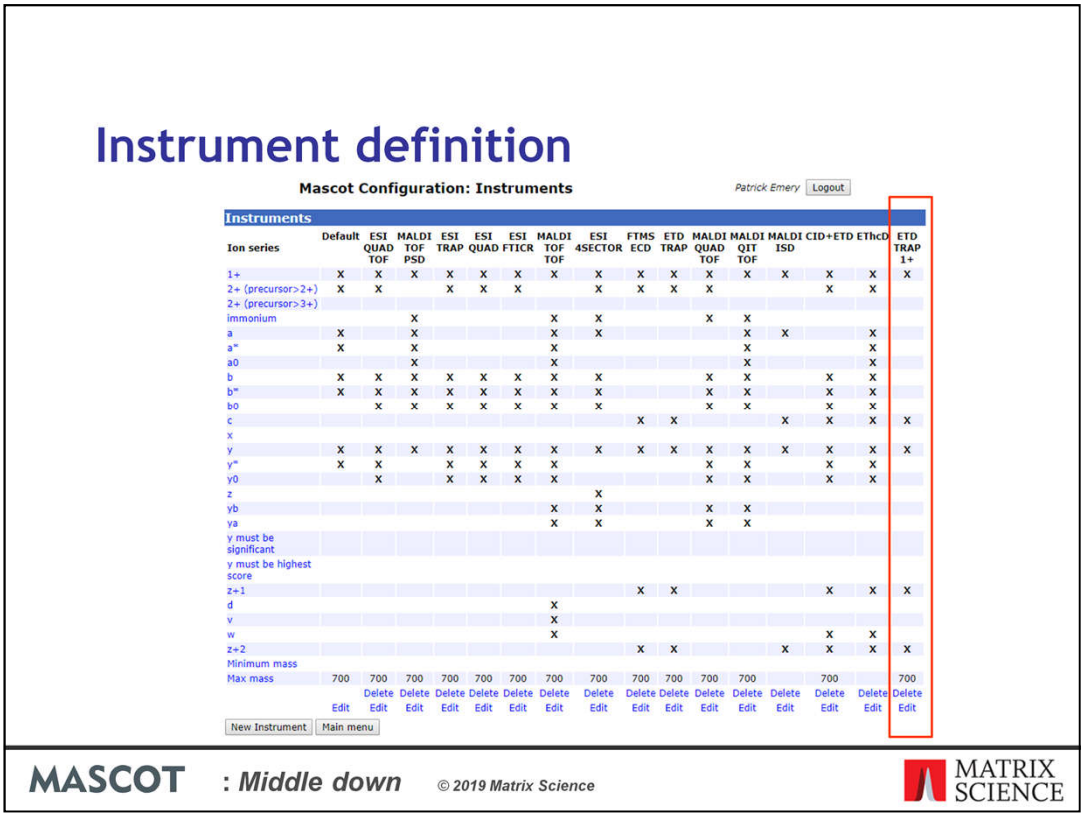


Mascot only matches 1+ and 2+ MS/MS series. With the higher precursor charge states we see in middle down datasets, it's likely you'll have fragment ions present at much higher charge states that this. Therefore, it's important to select the option to de-charge the MS/MS peak lists output from Distiller to MH+ values. You'll find this under the Tools->Preferences dialog on the Peak List Format tab in Distiller.

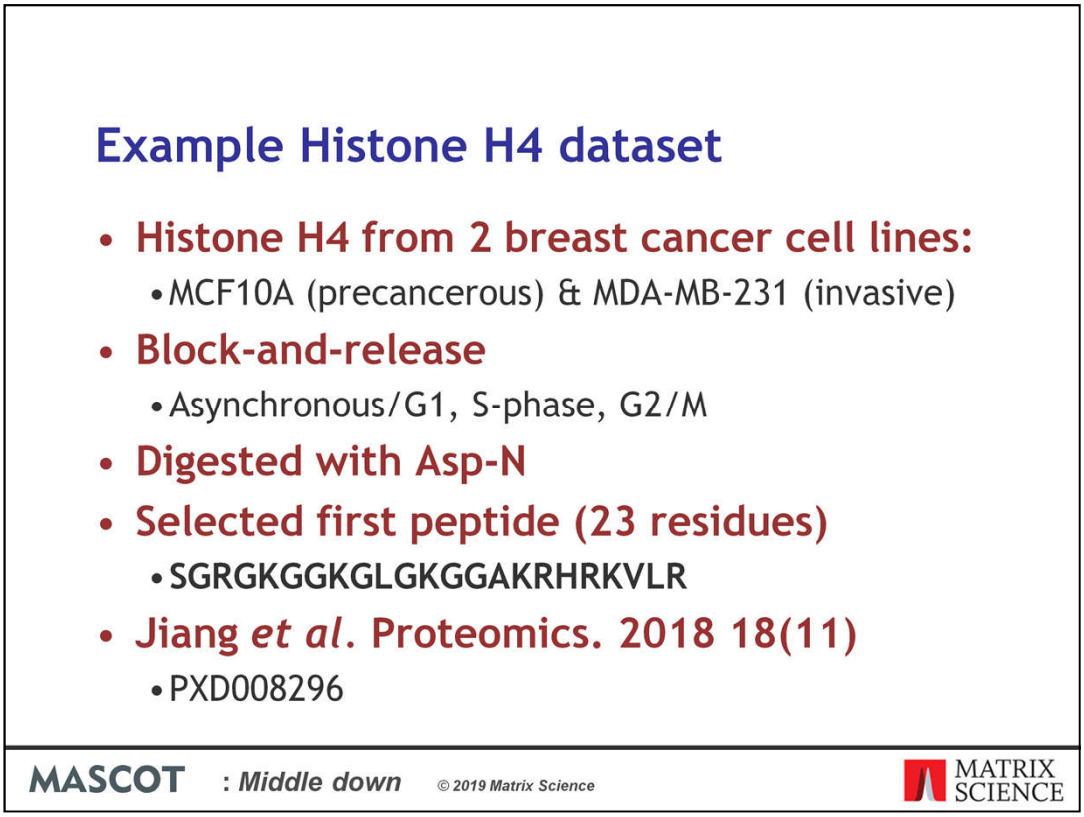
Peak List Format 2:



If you want to automate processing and searching using Mascot Daemon and the Distiller Daemon Toolkit, the equivalent setting is on the Mascot Distiller import options dialog.

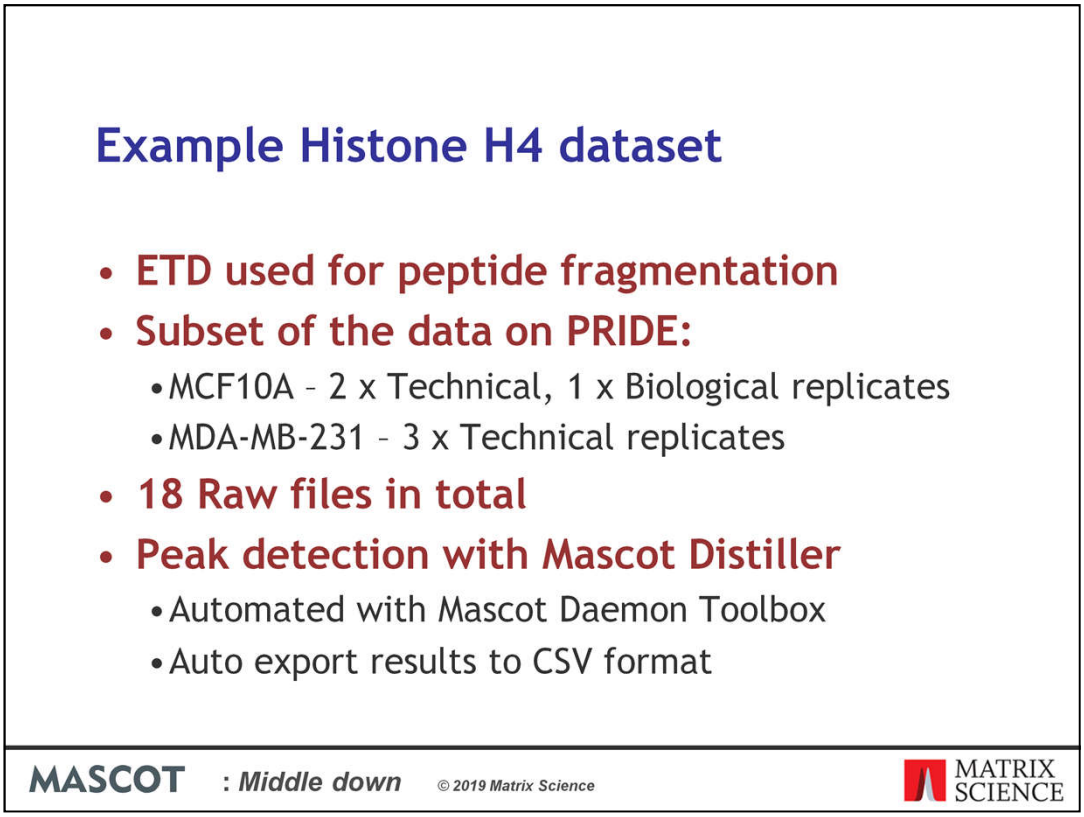


Of course, if you're doing this routinely, then you may also wish to add an Instrument definition to your Mascot server which only looks for 1+ ions series and use this for your searches. You can do this via the configuration editor on your Mascot server – here I've added a definition for the ETD-TRAP which only looks for 1+ ions.



A typical use of middle down proteomics is to look at regions of highly modified proteins such as Histones. As an example of how to use Mascot Distiller and Mascot Server to process and search middle down data, we're going to take a look at a middle down dataset comprised of Histone H4 from 2 cell lines, one pre-cancer cell line and one breast cancer cell line. This is a block and release experiment, so it starts with asynchronous cells (which will mostly be in G1 of the cell-cycle), blocked and release to synchronise them, with further samples taken when the cells were in S-phase and the G2/M phase.

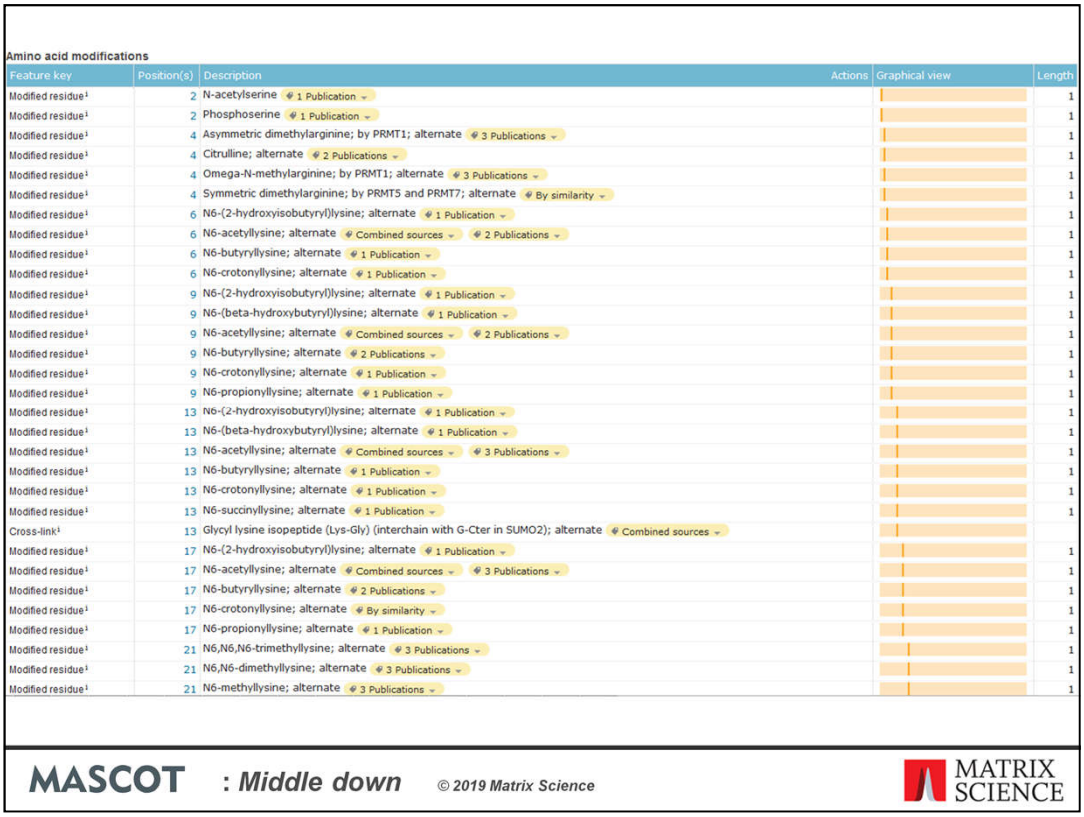
Digested with Asp-N and the first peptide from H4 was selected for MS/MS – that's a 23 residue lysine and arginine rich peptide peptide which has multiple possible cleavage sites for trypsin – in fact, you would be relying of incomplete digestion to get any matches to this from a trypsin digest.



You'll also often find CID used for peptide fragmentation in middle- and top- down experiments. However, ETD was used here, which also works well.

From the raw data available on PRIDE, I've taken 2 technical and 1 biological replicate for the MCF10A time line, and 3 technical replicates for the MDA-MB-231 cell line, for a total of 18 raw files across the 3 time points. This is excluding several technical and biological replicates which are incomplete in the PRIDE upload.

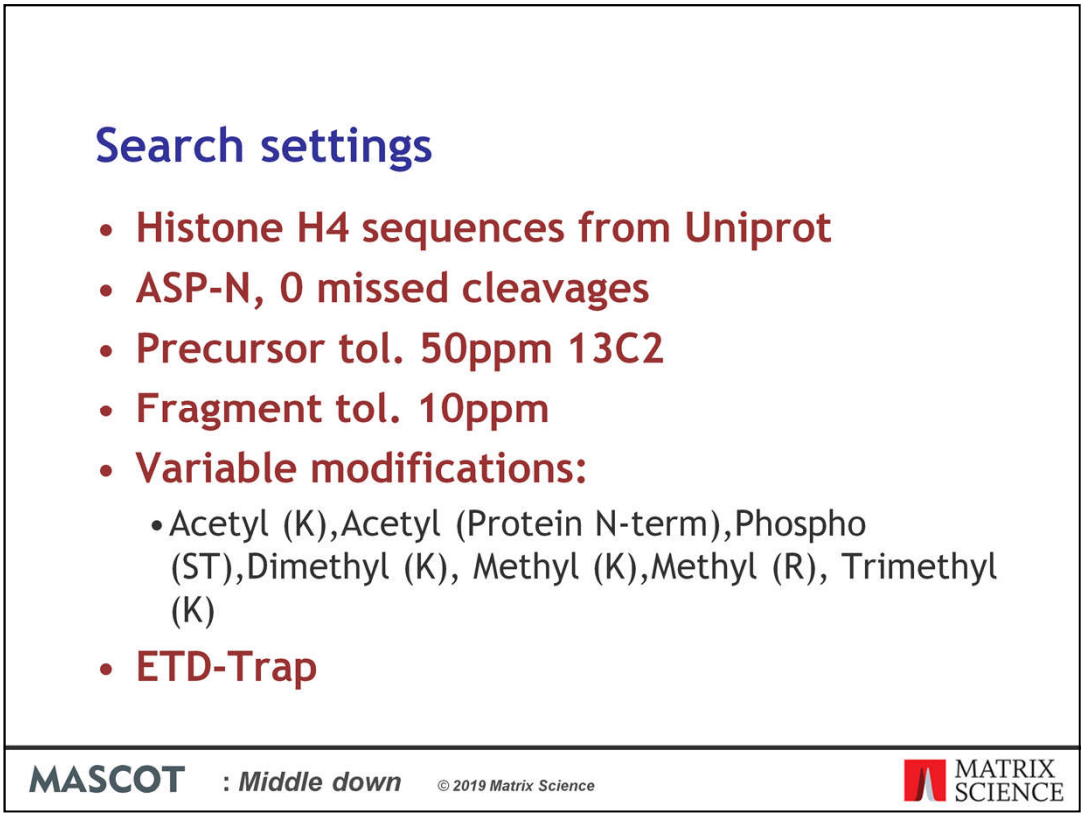
Peak detection was carried out with Mascot Distiller, automated using the Mascot Daemon Toolbox. I also used the auto-export feature in Daemon to export the search results to the CSV format.



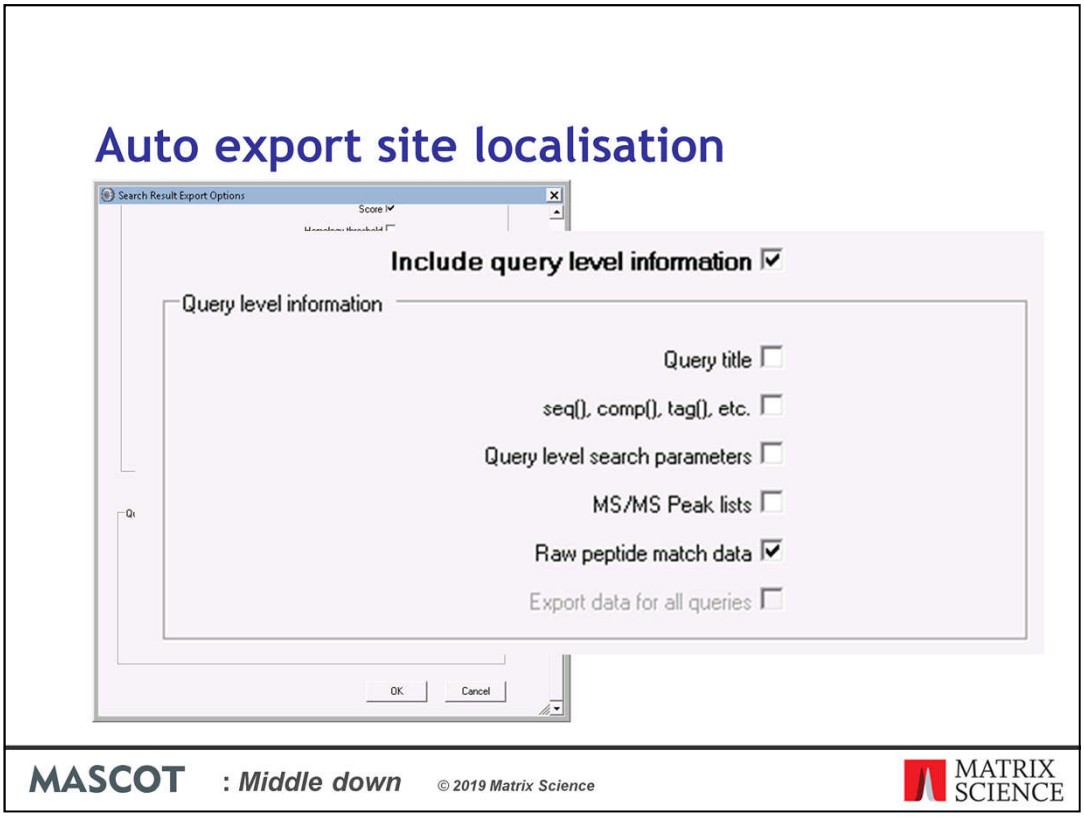
Amino acid modifications

Feature key	Position(s)	Description	Actions	Graphical view	Length
Modified residue[1]	2	N-acetylserine	1 Publication		1
Modified residue[1]	2	Phosphoserine	1 Publication		1
Modified residue[1]	4	Asymmetric dimethylarginine; by PRMT1; alternate	3 Publications		1
Modified residue[1]	4	Citrulline; alternate	2 Publications		1
Modified residue[1]	4	Omega-N-methylarginine; by PRMT1; alternate	3 Publications		1
Modified residue[1]	4	Symmetric dimethylarginine; by PRMT5 and PRMT7; alternate	By similarity		1
Modified residue[1]	6	N6-(2-hydroxyisobutyryl)lysine; alternate	1 Publication		1
Modified residue[1]	6	N6-acetyllysine; alternate	Combined sources, 2 Publications		1
Modified residue[1]	6	N6-butyryllysine; alternate	1 Publication		1
Modified residue[1]	6	N6-crotonyllysine; alternate	1 Publication		1
Modified residue[1]	9	N6-(2-hydroxyisobutyryl)lysine; alternate	1 Publication		1
Modified residue[1]	9	N6-(beta-hydroxybutyryl)lysine; alternate	1 Publication		1
Modified residue[1]	9	N6-acetyllysine; alternate	Combined sources, 2 Publications		1
Modified residue[1]	9	N6-butyryllysine; alternate	2 Publications		1
Modified residue[1]	9	N6-crotonyllysine; alternate	1 Publication		1
Modified residue[1]	9	N6-propionyllysine; alternate	1 Publication		1
Modified residue[1]	13	N6-(2-hydroxyisobutyryl)lysine; alternate	1 Publication		1
Modified residue[1]	13	N6-(beta-hydroxybutyryl)lysine; alternate	1 Publication		1
Modified residue[1]	13	N6-acetyllysine; alternate	Combined sources, 3 Publications		1
Modified residue[1]	13	N6-butyryllysine; alternate	1 Publication		1
Modified residue[1]	13	N6-crotonyllysine; alternate	1 Publication		1
Modified residue[1]	13	N6-succinyllysine; alternate	1 Publication		1
Cross-link[1]	13	Glycyl lysine isopeptide (Lys-Gly) (interchain with G-Cter in SUMO2); alternate	Combined sources		
Modified residue[1]	17	N6-(2-hydroxyisobutyryl)lysine; alternate	1 Publication		1
Modified residue[1]	17	N6-acetyllysine; alternate	Combined sources, 3 Publications		1
Modified residue[1]	17	N6-butyryllysine; alternate	2 Publications		1
Modified residue[1]	17	N6-crotonyllysine; alternate	By similarity		1
Modified residue[1]	17	N6-propionyllysine; alternate	1 Publication		1
Modified residue[1]	21	N6,N6,N6-trimethyllysine; alternate	3 Publications		1
Modified residue[1]	21	N6,N6-dimethyllysine; alternate	3 Publications		1
Modified residue[1]	21	N6-methyllysine; alternate	3 Publications		1

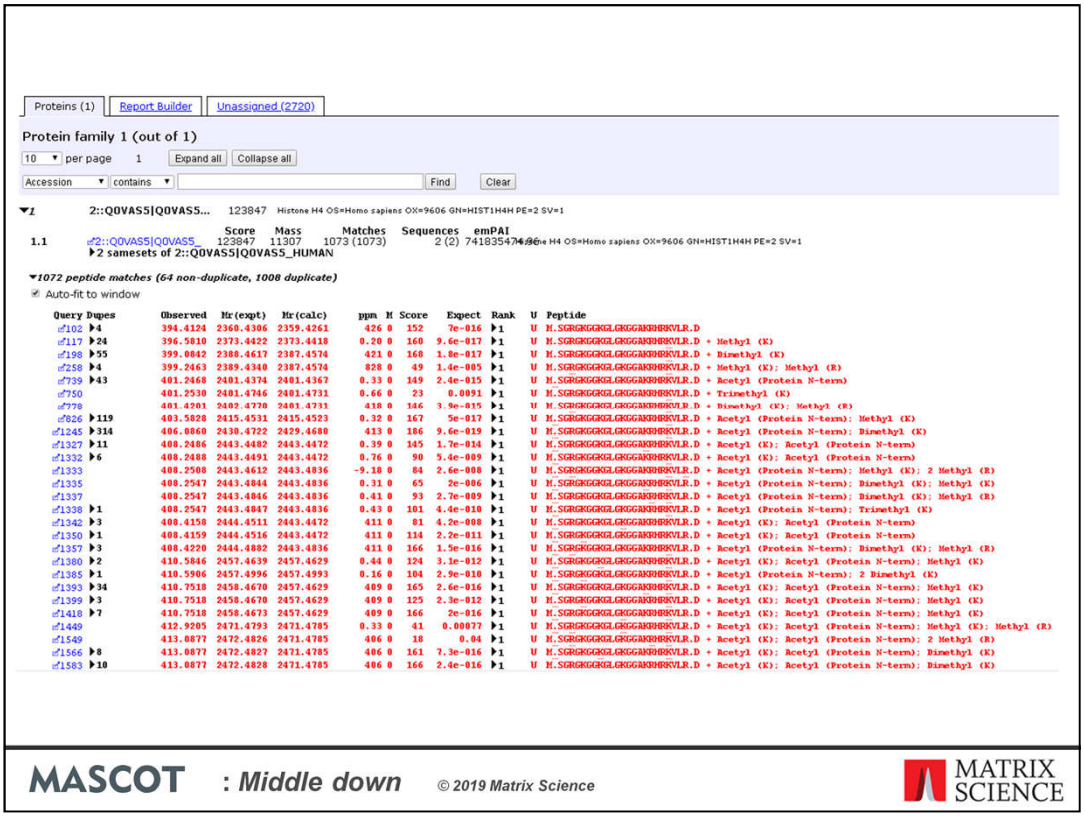
As this screenshot from Uniprot shows, this N-terminal stretch of Histone H4 can be highly modified. Acetylation and methylation are the most common modifications, but they're far from the only types observed.



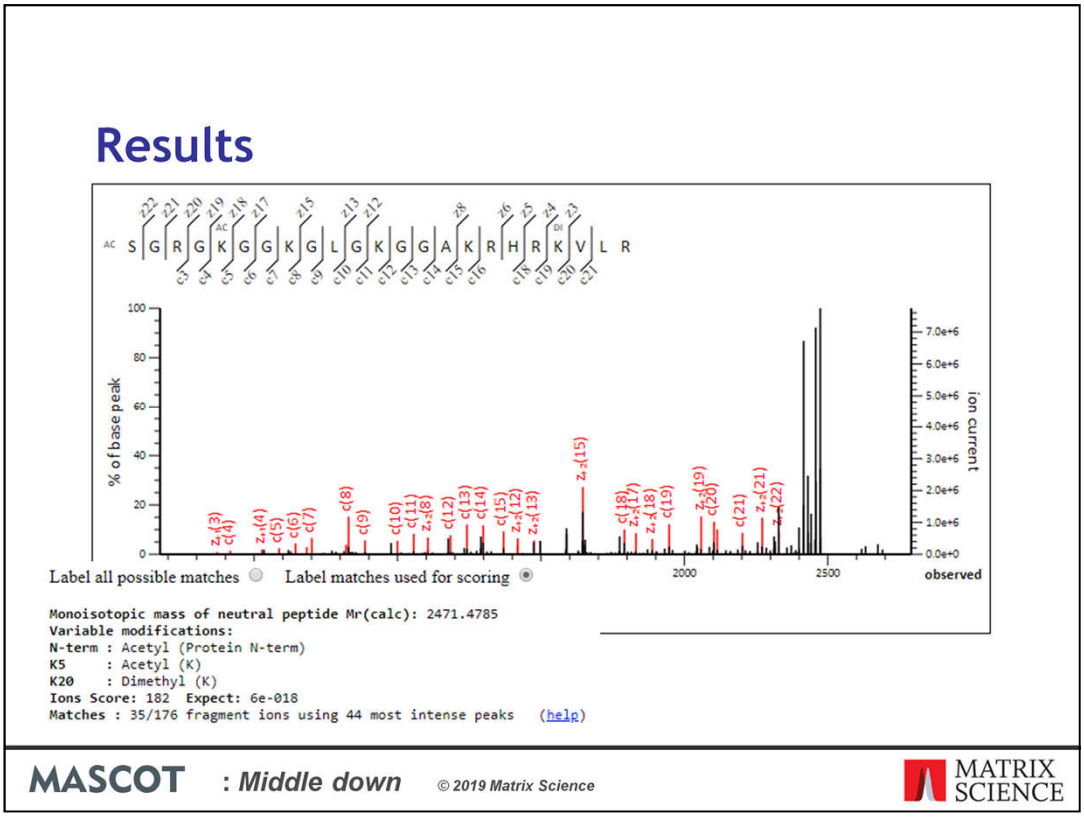
So here are our search settings – the generated peak lists were search against human Histone H4 sequences from Uniprot, using Asp-N as the enzyme with no missed cleavages. Other search settings were derived from the paper, except for the Precursor tolerance, where I've used a much tighter value, but allowed for 13C. We have a fragment tolerance of 10ppm and have selected a raft of lysine acetylation and methylation variable modifications. In addition, we'll look for Phospho Serine, and Argine methylation.



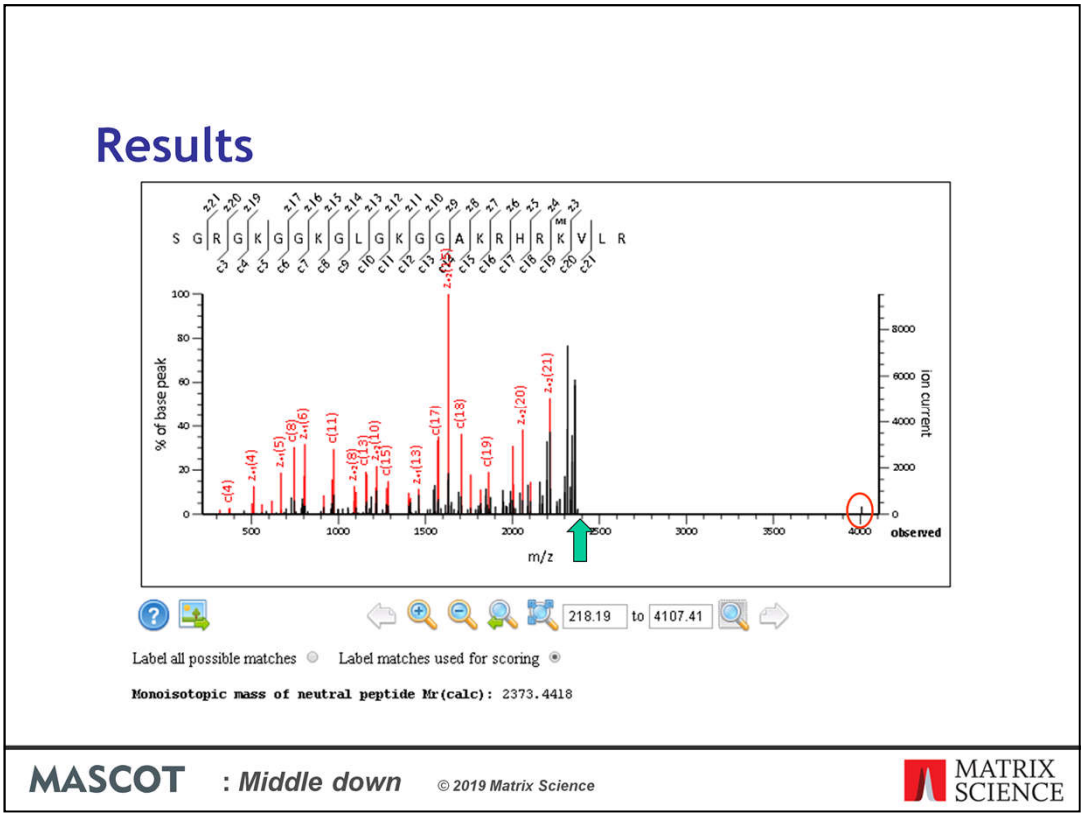
I used the 'Auto Export' option in Daemon 2.6 to automatically export the search results into the CSV format. Since I'm interested in modification patterns and site localisation, I also wanted that data including in the export. To do this, you need to check the 'Include query level information' checkbox and under that the 'Raw peptide match data' in the auto export dialog.



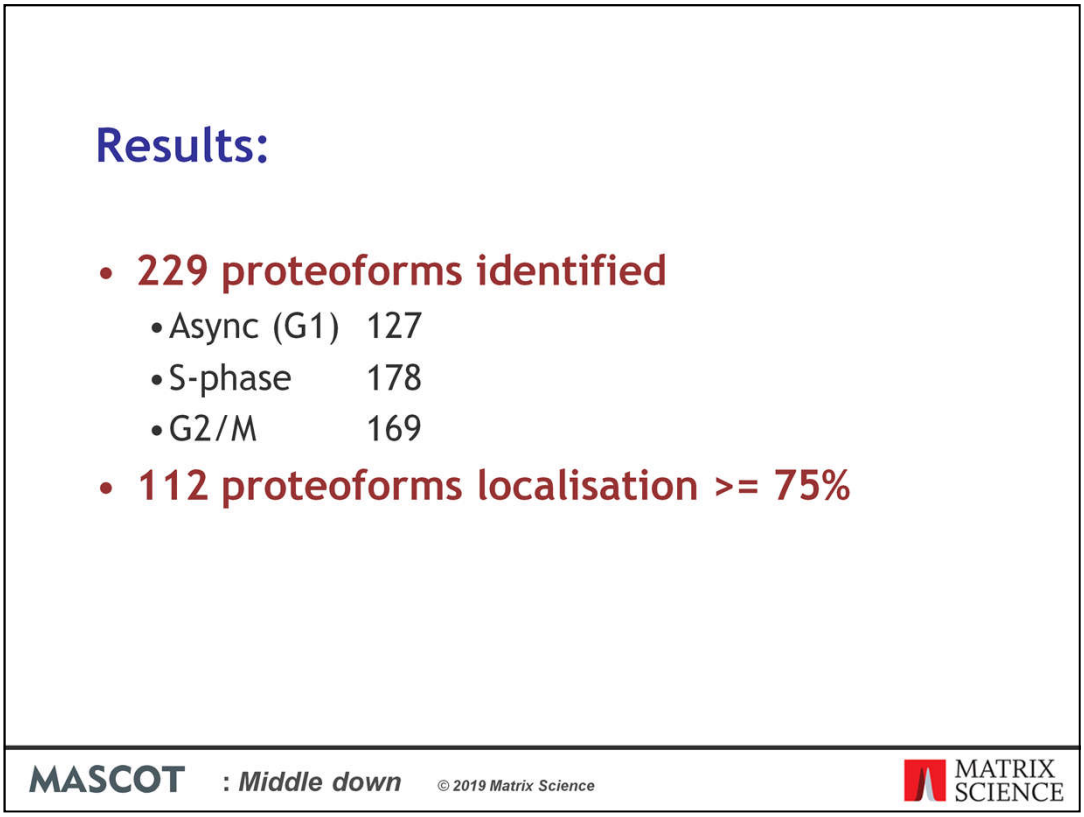
Here's an example result set from one of the G2/M phase raw files – as you can see, we're getting a lot of very high scoring matches to our target peptide, and we've got a large number of different modification patterns identified.



If we take a look at a fairly typical match – we're getting good quality matches from the data, the example shown here is by no means atypical. Long runs of c- and z- matches, and ion scores > 100 are not uncommon.

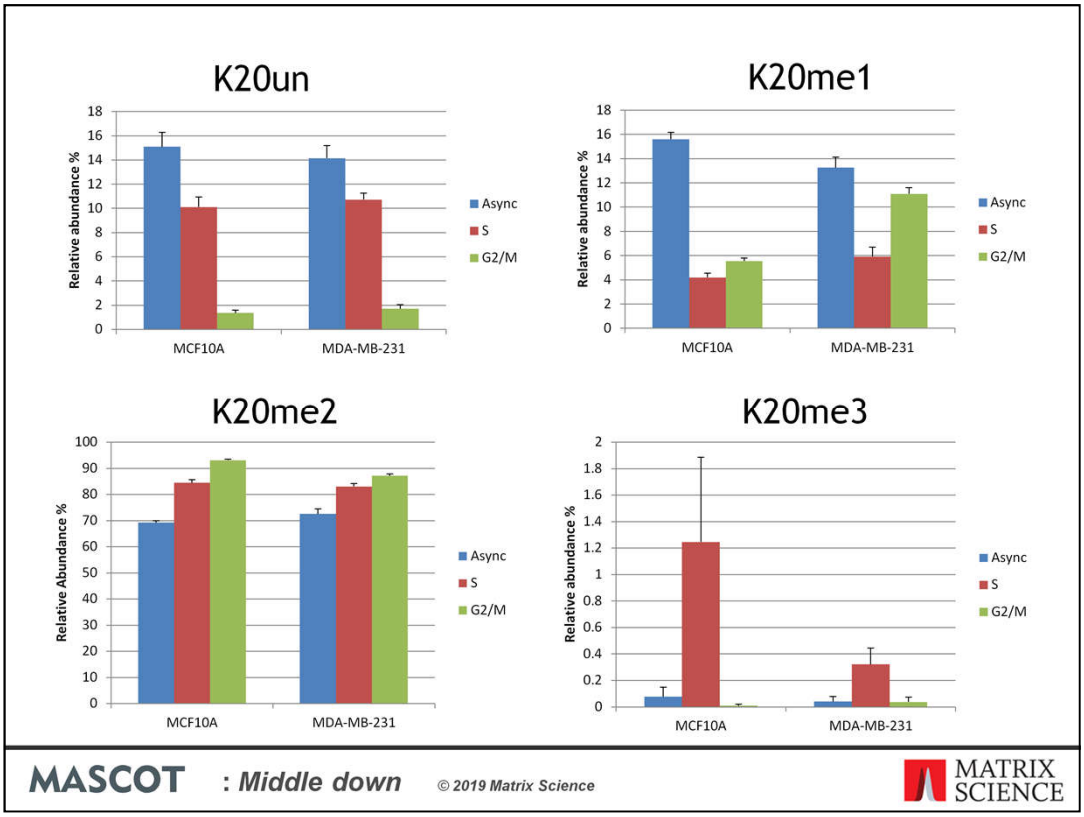


We're also seeing generally very reliable fragment charge state determination from Mascot Distiller. The Mr of this peptide is roughly 2.4KDa – we only have a single low intensity fragment ion in the peaklist with an m/z above this value, so the vast majority of identified peaks in this peaklist also had their correct charge state determined, and have therefore been correctly decharged.

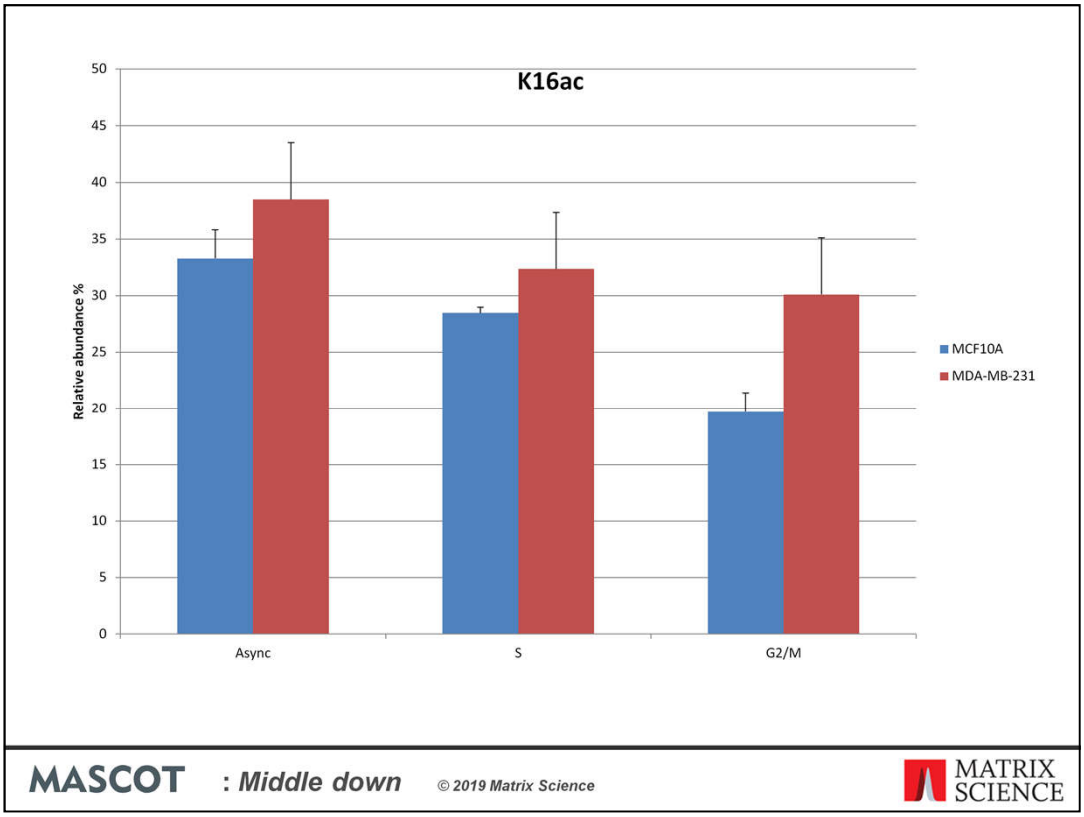


Across the two cell lines, we identified 229 proteoforms. We found 127 different possible proteoforms is the asynchronous datasets, increasing to 178 identified in S-phase and a slight decrease to 169 in G2/M phase.

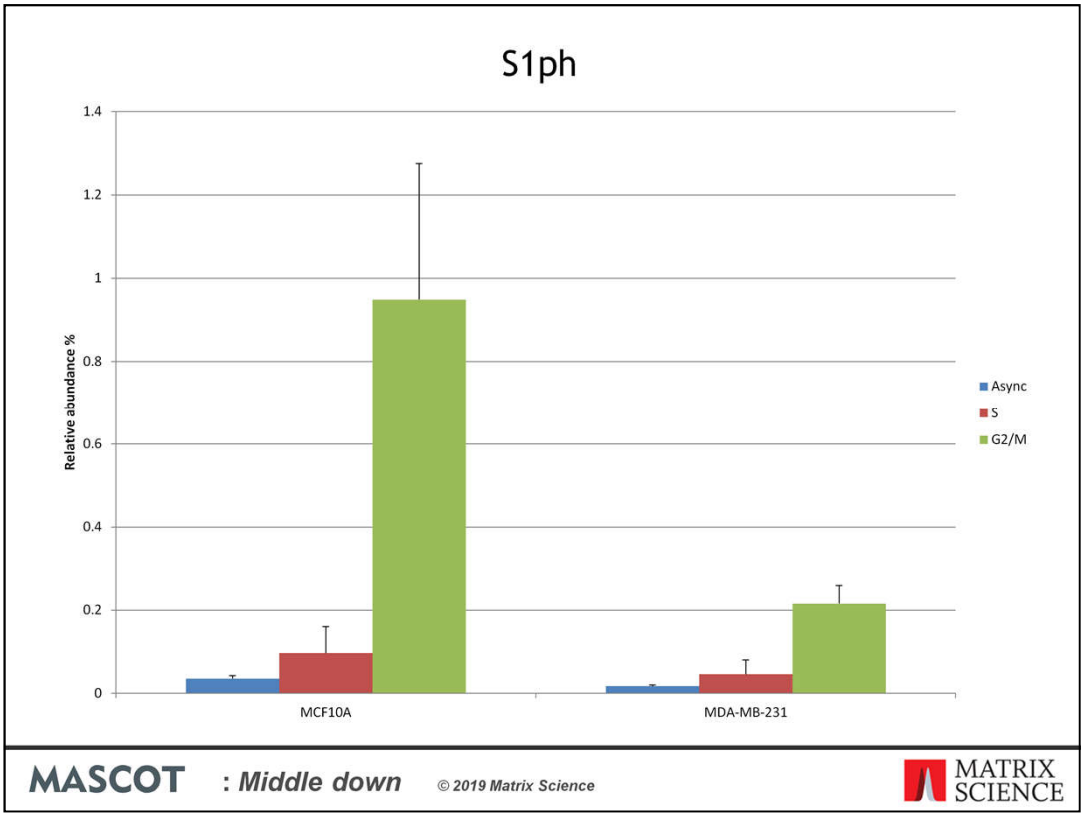
Of those, 112 have a localisation score from the Mascot delta score of 75% or better. Of the -forms which are excluded, some could represent mixtures of different modification patterns, whilst others don't have enough information in the MS/MS to be able to chose between two (or more) different arrangements. For the next step of analysis, peptides with a Mascot delta score site localisation of less than 75% were excluded from the analysis.



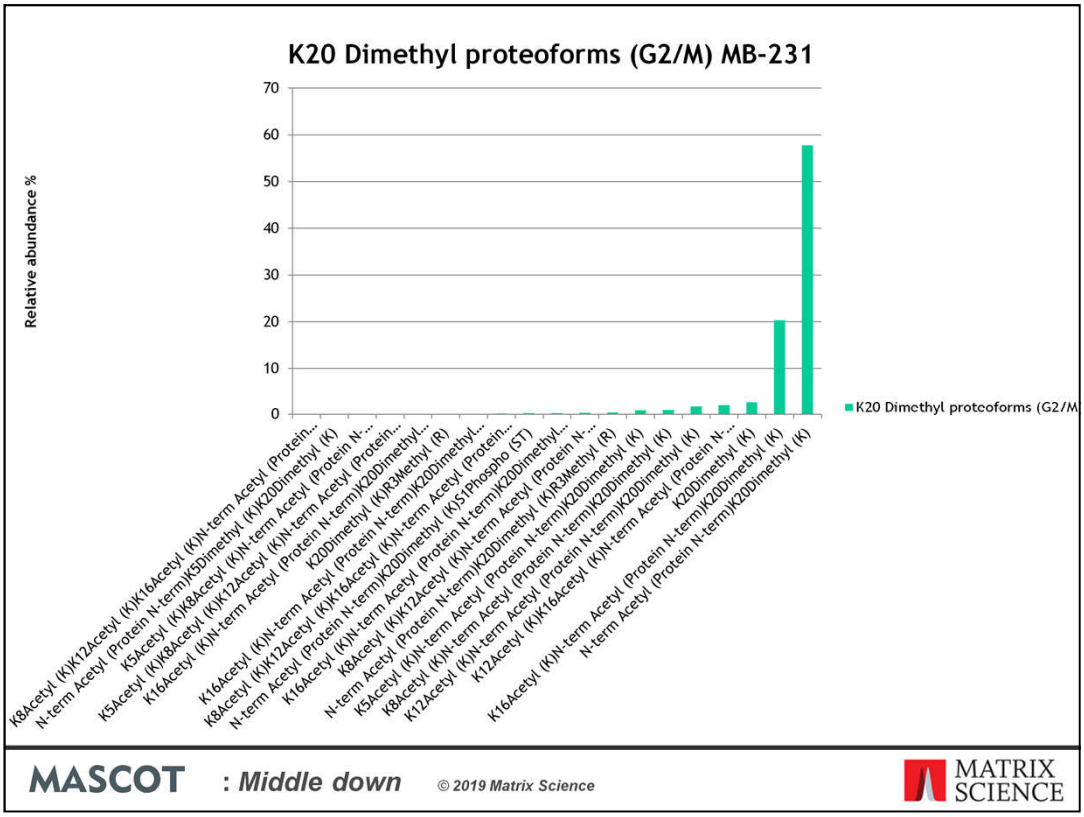
Here, we're looking at levels of methylation on Lysine 20 across the cell-cycle for the two cell lines. To generate this, we've collapsed all the K20 modified proteoforms together and calculated the relative abundance based on the precursor intensities of the matched MS/MS spectra. What we see is in line with the results presented by Jiang *et al.* – the levels of un-methylated K20 are similar in the two cell-lines and fall across the cell-cycle. While the general patterns of mono, di and tri-methylation are similar between the two cell lines across the cell-cycle, there are some obvious differences. Mono-methylation levels in G2/M in MDA-MB-231 appear to be significantly greater than in MCF10A – corresponding to a roughly equivalent reduction in di-methylation.



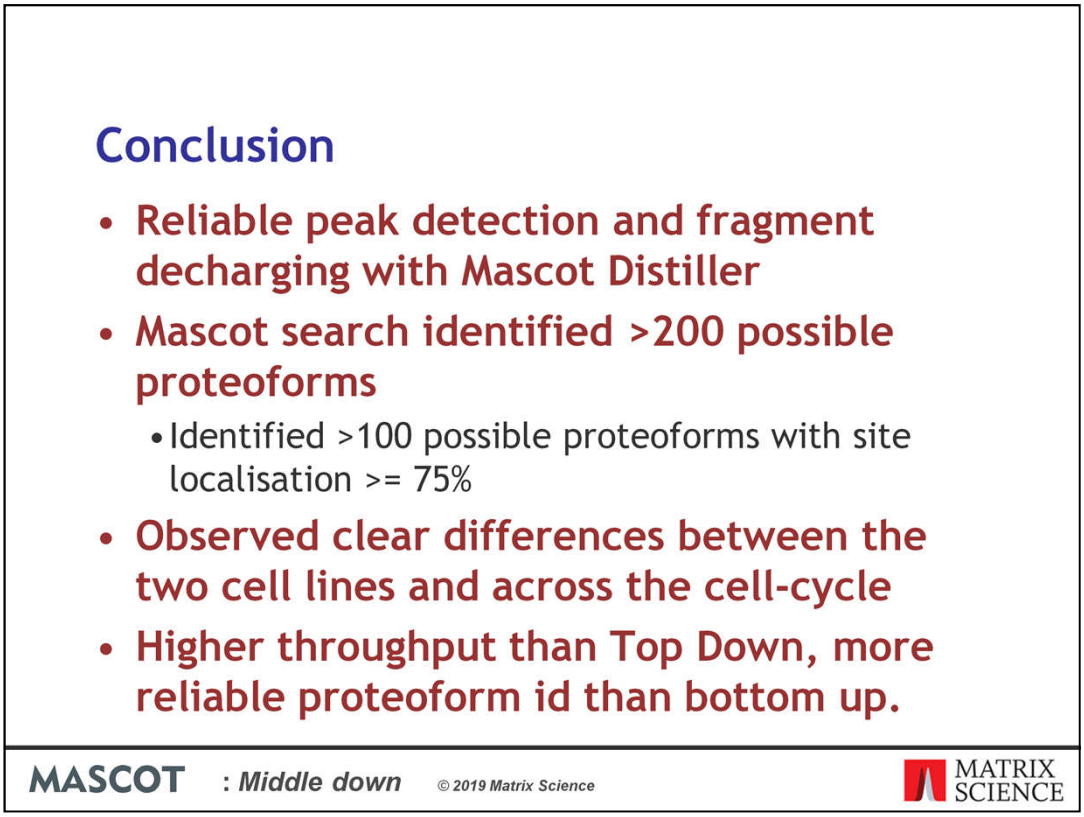
Here we're looking at the relative abundance of all the K16 acetylated proteoforms – there is a clear difference between the two cell lines, with MDA-MB-231 showing consistently higher levels of K16 acetylation than MCF10A



S1 Phosphorylation has been seen across a number of different organisms, and is associated with chromatin condensation during mitosis, and we do see a large increase in the abundance of S1ph at mitosis. This is most noticeable in the precancerous MCF10A cell-line. The increase observed in the MDA-MB-231 cancerous cell line is much less dramatic.



Of course, the data are much richer than that shown by simply collapsing all the proteoforms for a specific residue of interest together. Here, I've plotted the relative abundance of all the K20 Dimethyl containing proteoforms identified for one of the MB-231 result files. As you can see, the vast majority of the identified K20 Dimethyl peptides are found with protein N-terminal acetylation. The next most common is with the addition of K16 acetylation, and then the relative abundance of other proteoforms falls away rapidly.



In conclusion, we were able to carry out reliable peak detection and fragment ion decharging using Mascot Distiller, and to easily automate the peak-picking, search and export process using Mascot Daemon and the Distiller Mascot Daemon Toolkit.

We obtained good results from the Mascot searches and were able to identify more than 200 possible proteoforms from the data we reprocessed and searched. Of those, over 100 had site localisation scores of >= 75% and were then retained for further analysis.

We observed clear differences between the two cell lines and across the cell-cycle, in line with the published results.

The use of Middle down allowed for higher throughput of proteoform identification than would have been (easily) possible with Top Down approached, whilst also allowing for much more reliable proteoform identification than would have been possible with bottom up techniques. However, unlike a top down experiment, we can't easily get a measure for all proteoforms across the entire protein.