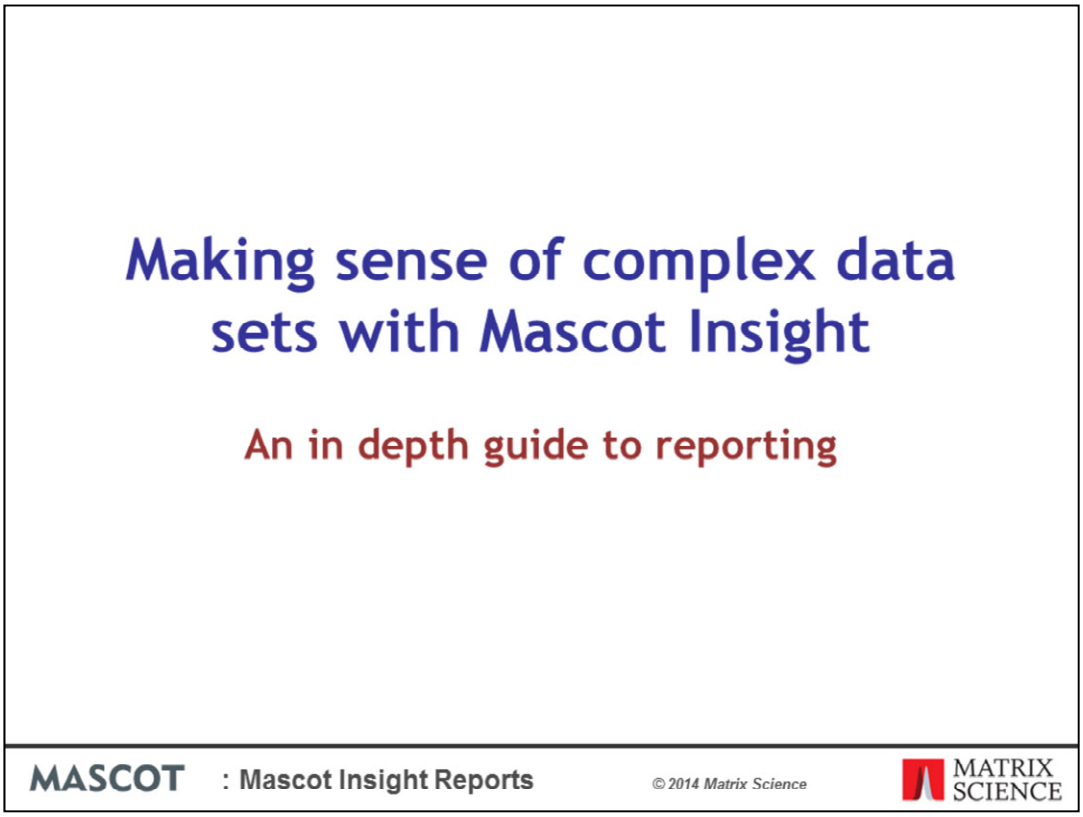
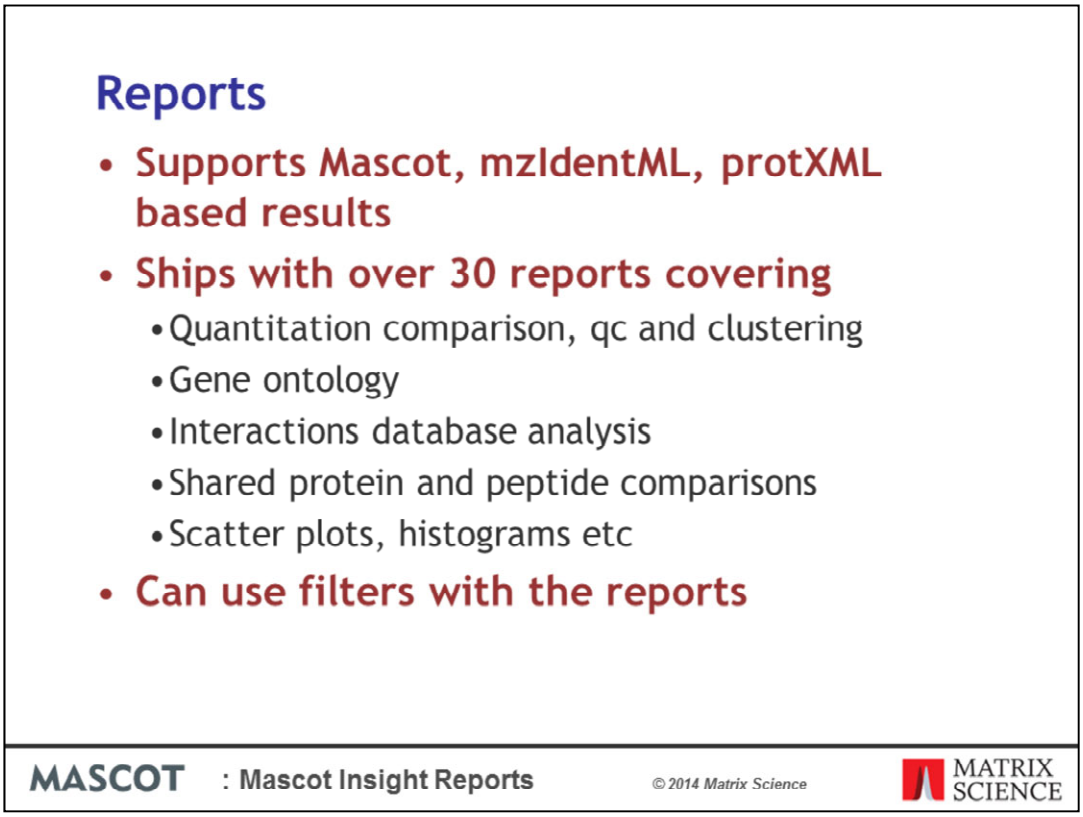
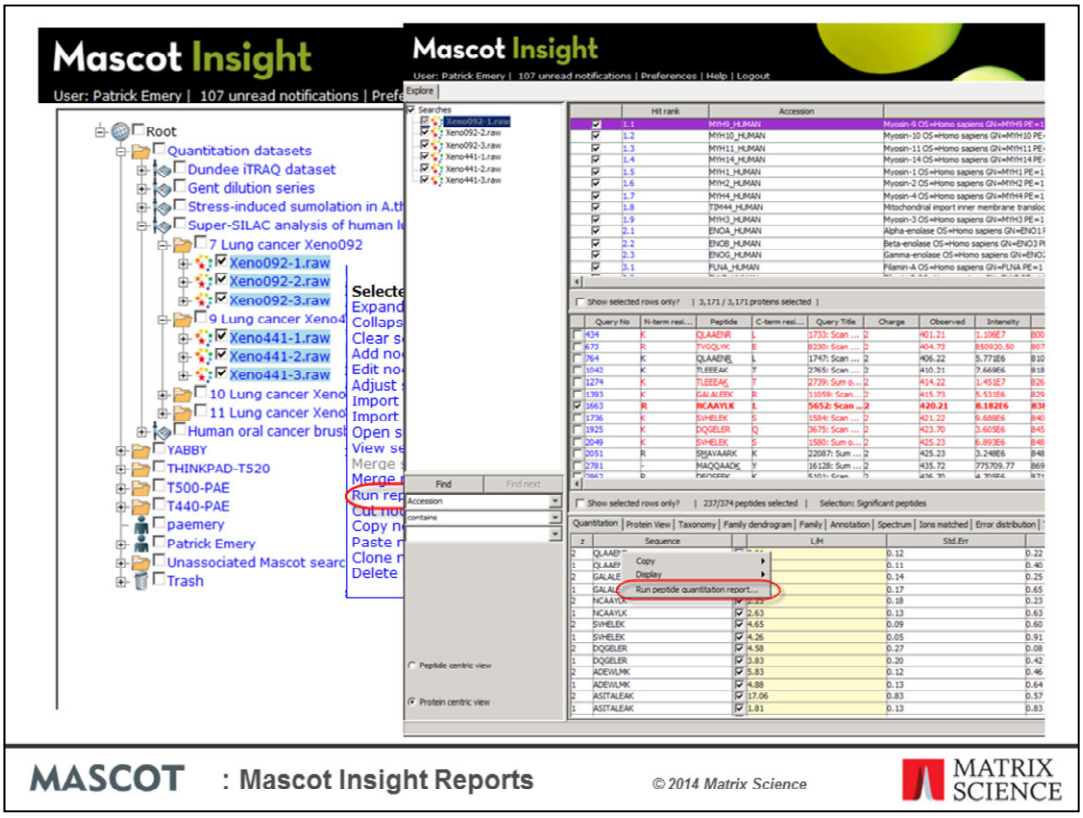


This presentation follows on from the talk presented by Patrick Emery at the beginning of this workshop but I will be focusing on the data analysis and reporting functions available in Mascot Insight.

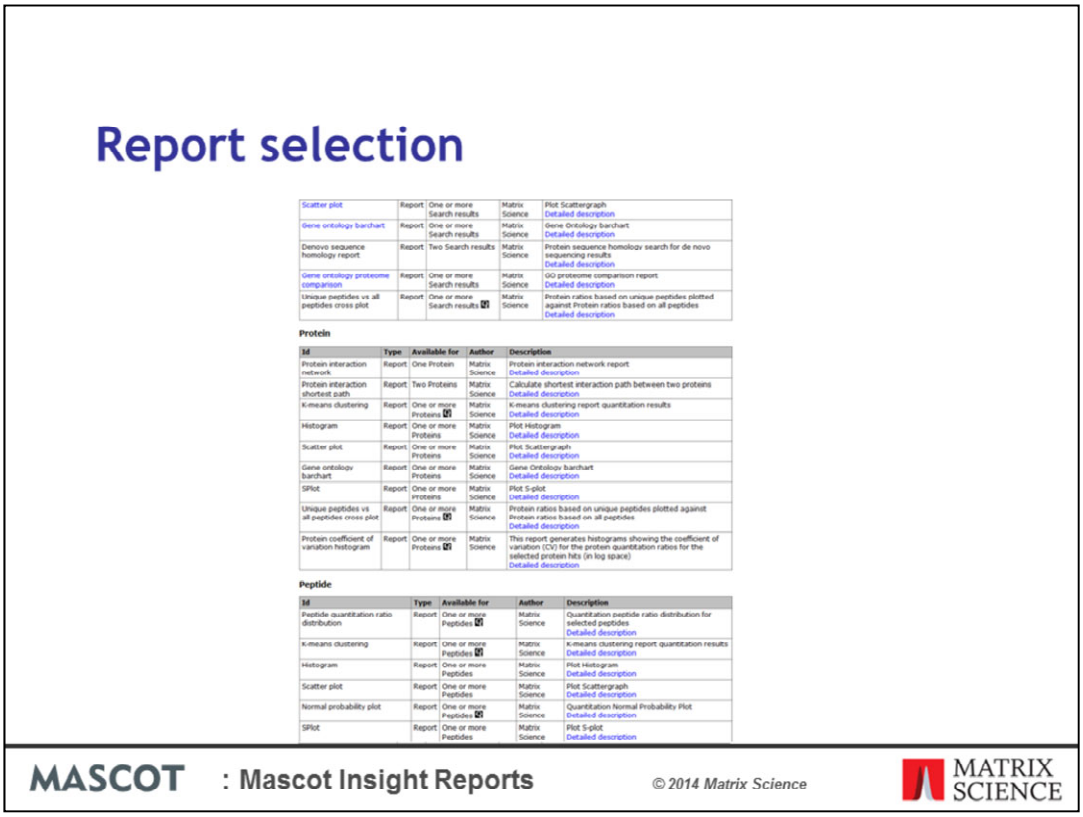


As mentioned previously Mascot Insight Can support Mascot, mzIdentML, protXML based results so these reports are not just relevant to Mascot Server results.

Insight ships with over 30 reports designed specifically for proteomics data. These cover a wide range of areas such as result comparison, quantitation and quantitation clustering/grouping, Gene Ontology analysis, Interactions database analysis and general graphing reports such as scatter plots. It is possible to stack some of the filters and reports and use the results as inputs in to later analysis. For example filtering out the proteins identified in the contaminates database used in the initial search so that the reporting only works with the matches to the target database.

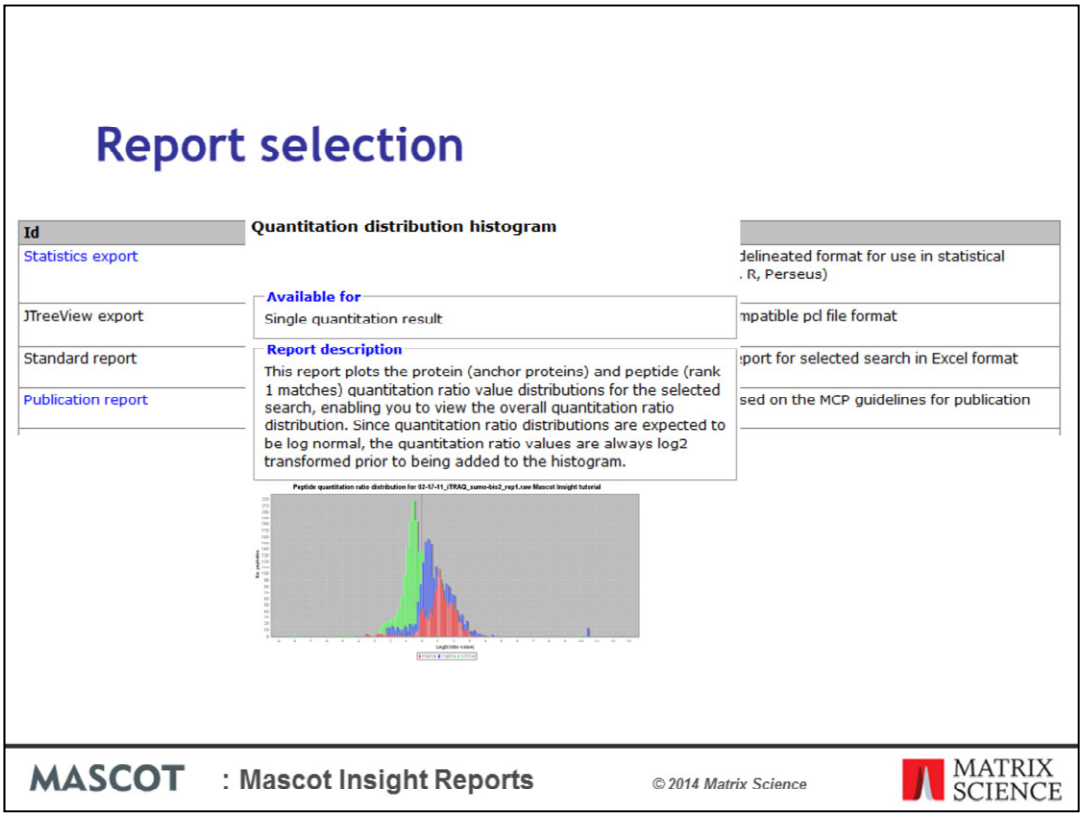


You can initiate reports from the data tree or within MIRA at the Search, Protein, Peptide or quantitation level.

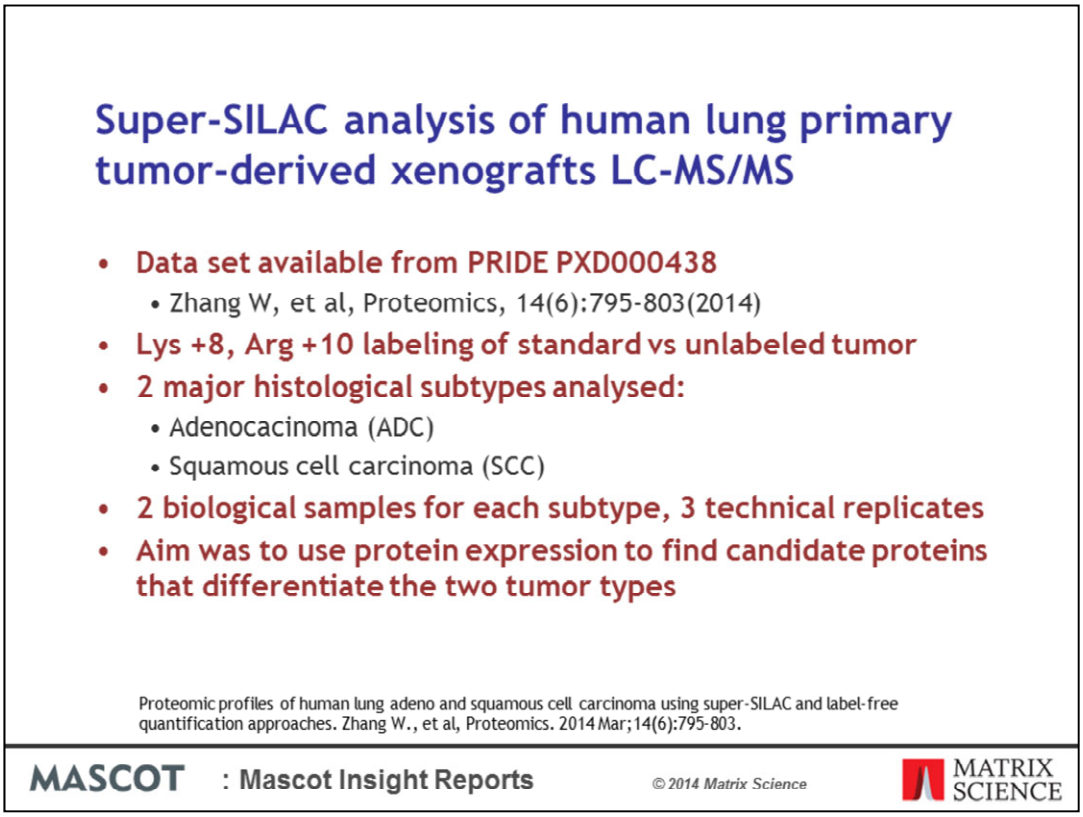


The report selection window appears. Only the reports that are suitable for the data selected are active in the list, shown in blue.

Quantitation reports are marked with a Q symbol.



There is a one line description for each report and a more detailed description a click away.



The data set that I am going to use for most of the examples is available at the EBI PRIDE repository.

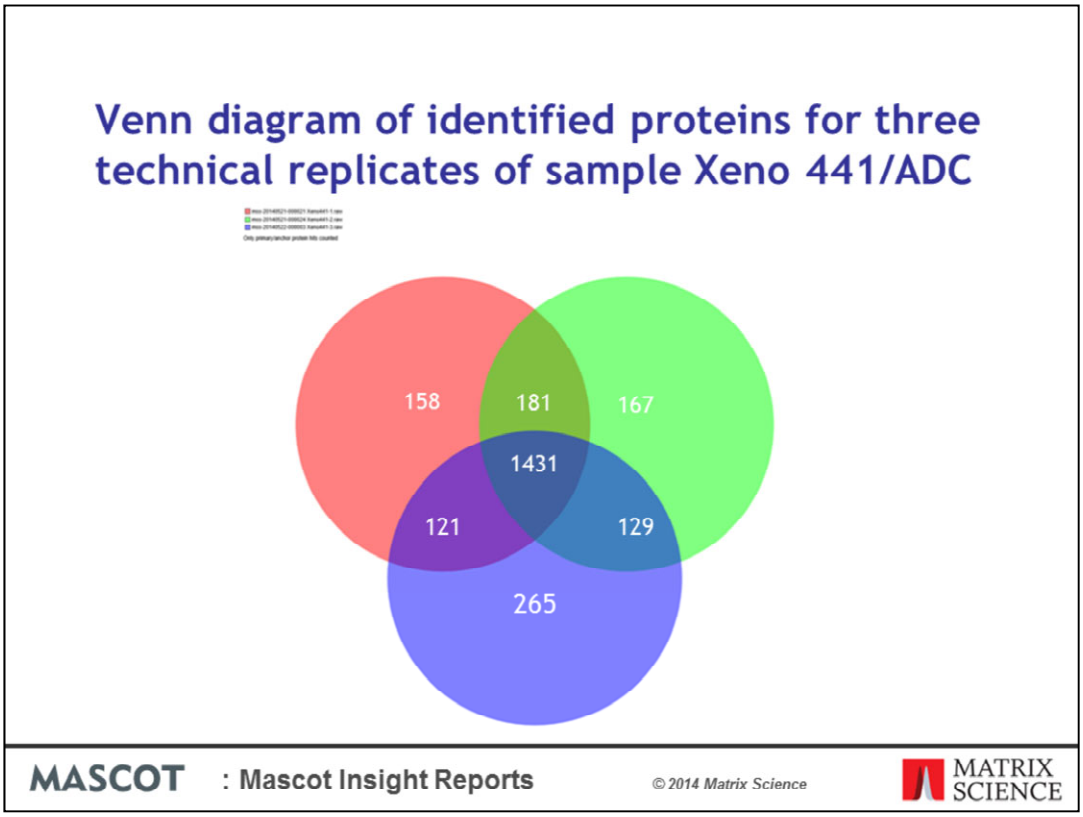
It is a Super-SILAC experiment using a lung cell culture labelled with Lys +8 and Arg +10 as a standard and comparing it to unlabeled human lung primary tumor-derived xenografts and analysed by LCMS/MS.

2 major histological subtypes were analysed, ADC and SCC with 2 biological samples for each subtype and 3 technical replicates

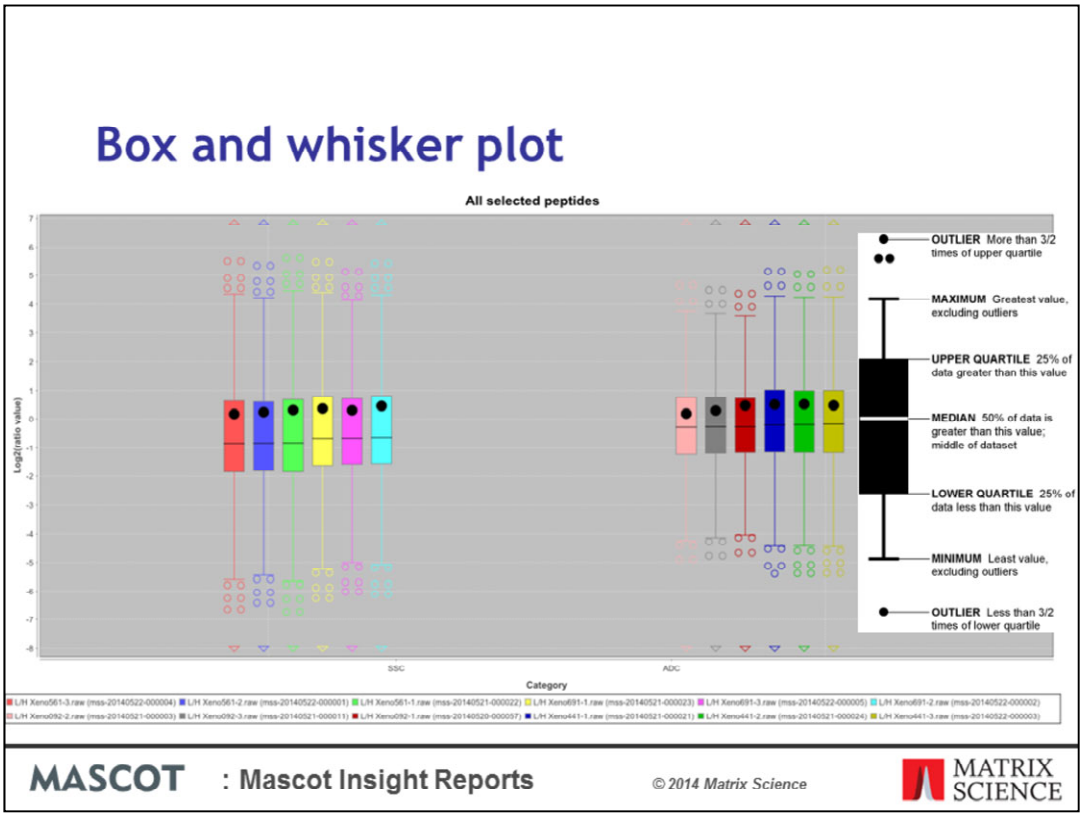
for each sample for a total of 12 MS/MS datasets. The data was analysed using the same search settings described in the publication.

The goal of the experiment was to compare the two lung tumor types and to use protein expression to differentiate them.

Reference: Proteomics. 2014 Mar;14(6):795-803. doi: 10.1002/pmic.201300382, Proteomic profiles of human lung adeno and squamous cell carcinoma using super-SILAC and label-free quantification approaches, Zhang W1, Wei Y, Ignatchenko V, Li L, Sakashita S, Pham NA, Taylor P, Tsao MS, Kislinger T, Moran MF.



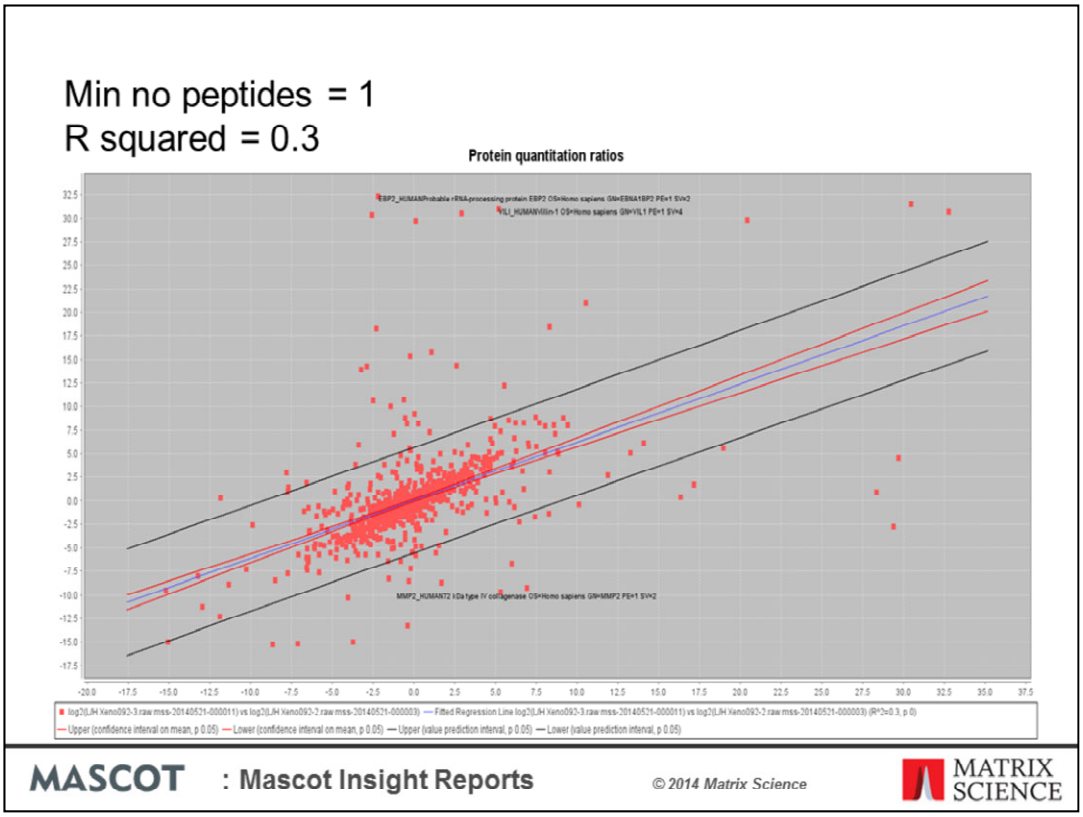
First I took a look at the technical replicates for each biological sample to see how consistent they are. These are the three replicates for the ADC sample Xeno 441. As you can see there is good overlap between the different analysis and each individual run identifies no more than 15% unique proteins.



Next I generated box and whisker plot. For those not familiar with box and whisker plots here is a quick guide. The plot will give us global information for each ratio from each replicate, showing us things like the global mean and median ratios, the data spread, presence of outliers etc. The black circle is the average.

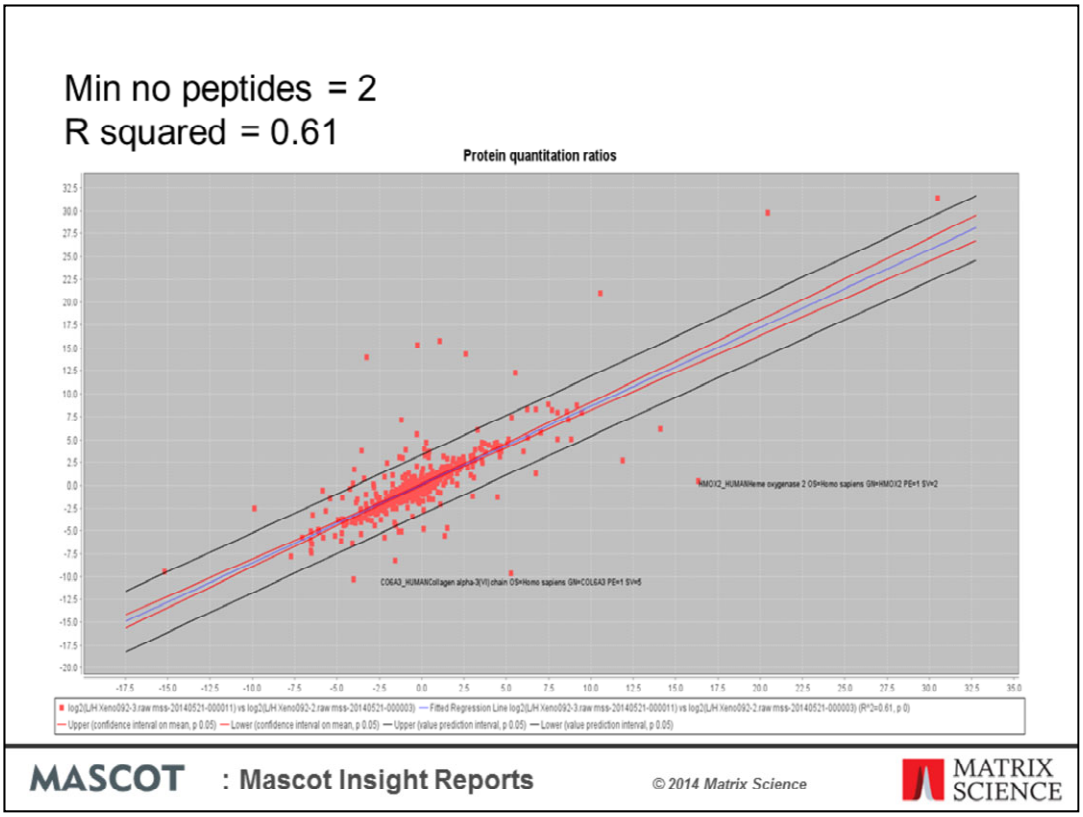
You can see the overall distributions of peptide ratios for each technical replicate, and also each tumour type, are pretty close to each other.

Once a report is generated, you can easily copy and paste it out of MIRA, or export it in a number of formats – including exporting the underlying data for the dataset as a CSV file, in case you want to get the data into another package such as Excel or R.

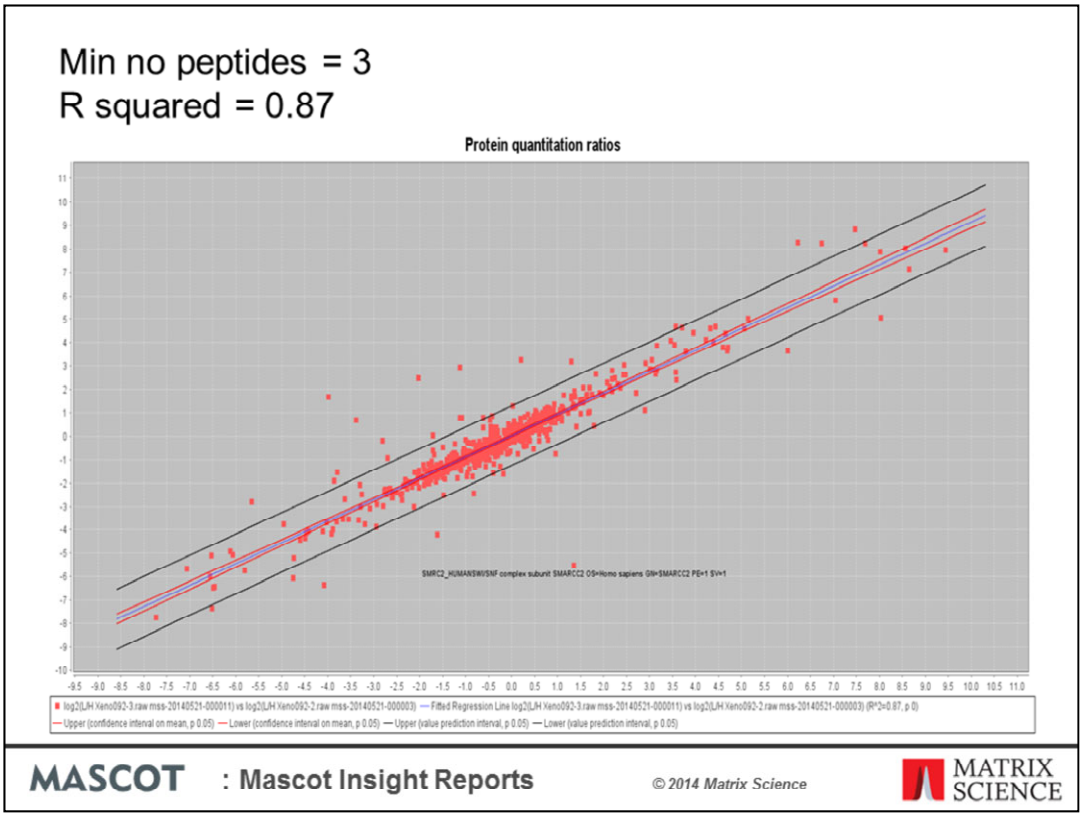


We can explore the quality of the data set further looking at the effect of the minimum number of quantitation values used per a protein on the ratio correlation. The question is how many peptide ratio measurements do I need for reliable quantitation results. For these examples I used two of Xeno 092 ADC replicates. Prior to creating the report I filtered out matches to the contaminants database.

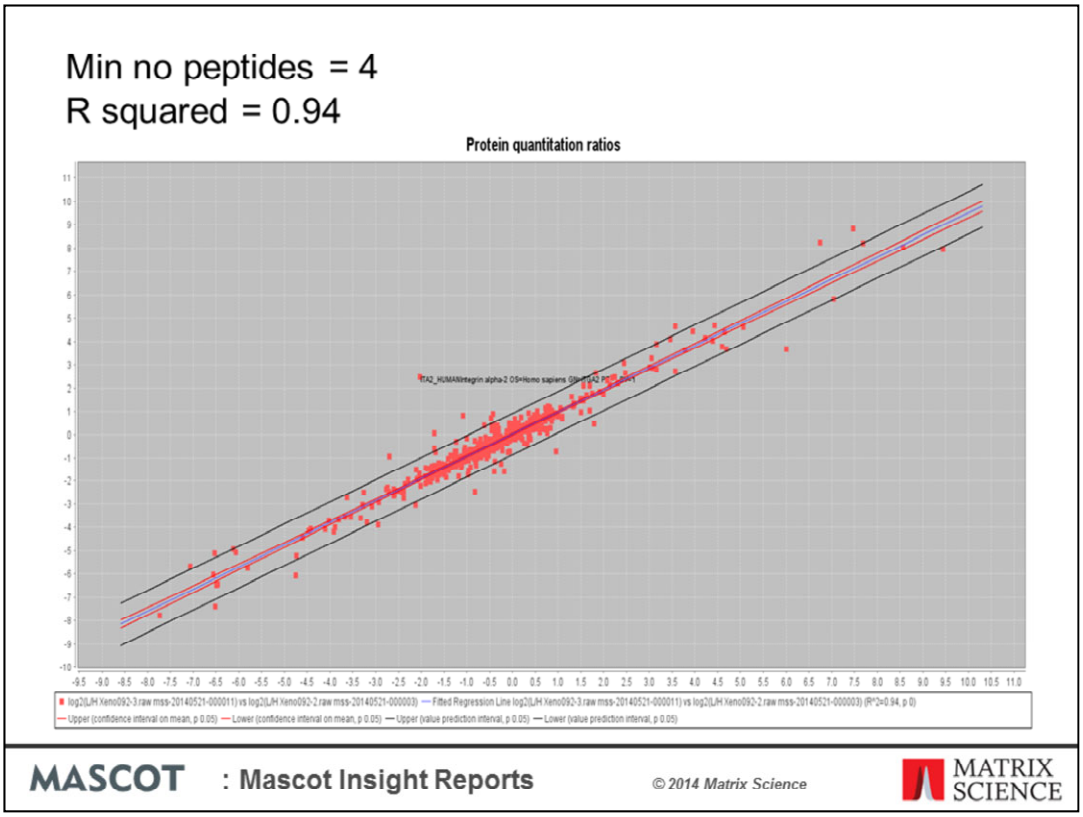
I then plotted the ratios from two replicates against each other using different numbers of minimum required peptide ratios per a protein (1-4), getting better and better correlations. Here is the plot for using 1 peptide ratio and it does not give me much confidence.



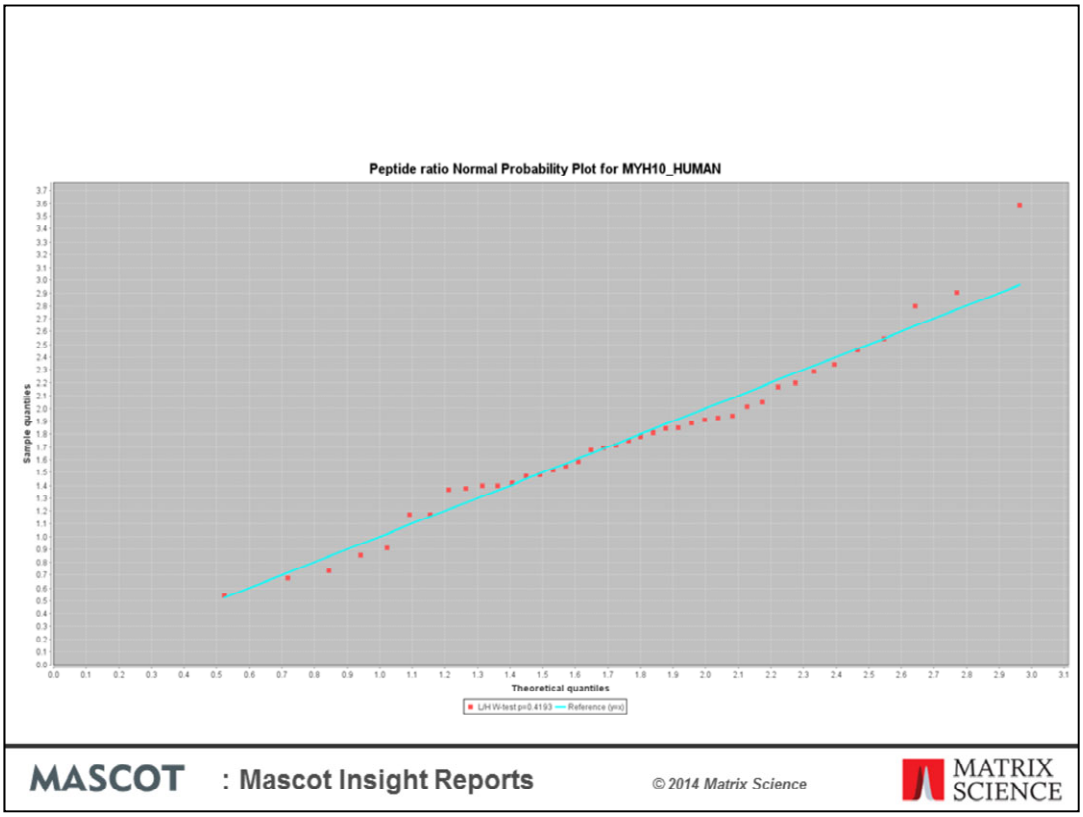
Using 2 peptides the protein ratio correlation starts to improve



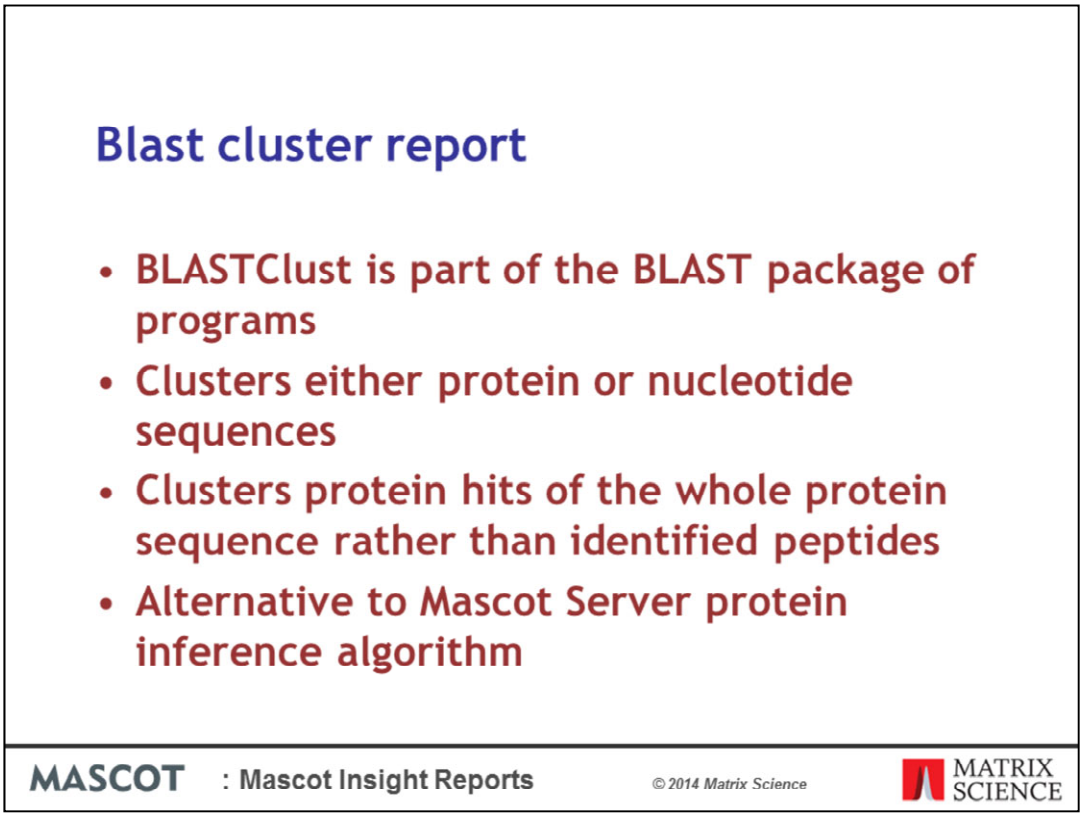
With three measurements - this is more like



Now we are there. This final plot shows that using 4 peptides gives a much higher ratio correlation between technical replicates than using just one measurement per a protein. This should be common sense but we still have queries from customers about obtaining as many protein quantitation ratios as possible even if they one have one or two peptide measurements.

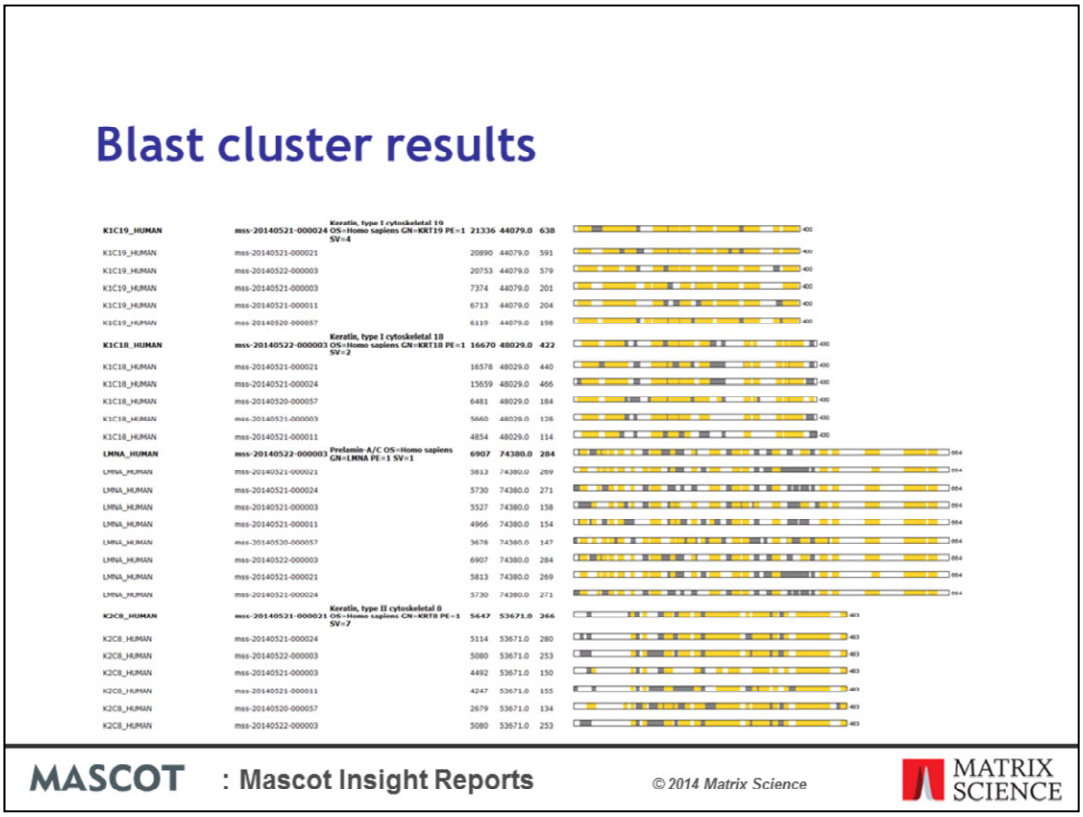


We would certainly be expecting Keratins to be differently expressed in a lung tumor. Here I show a Peptide ratio Normal probability plot for the Keratin, type I C 18 peptides. The peptide quantitation data deviated from the straight line of a theoretical normal distribution. This may be due to the large number of shared peptides between different keratin isoforms. Here is a plot for MYH10_HUMAN which is a lot closer to a straight line and of the quality we are looking for.

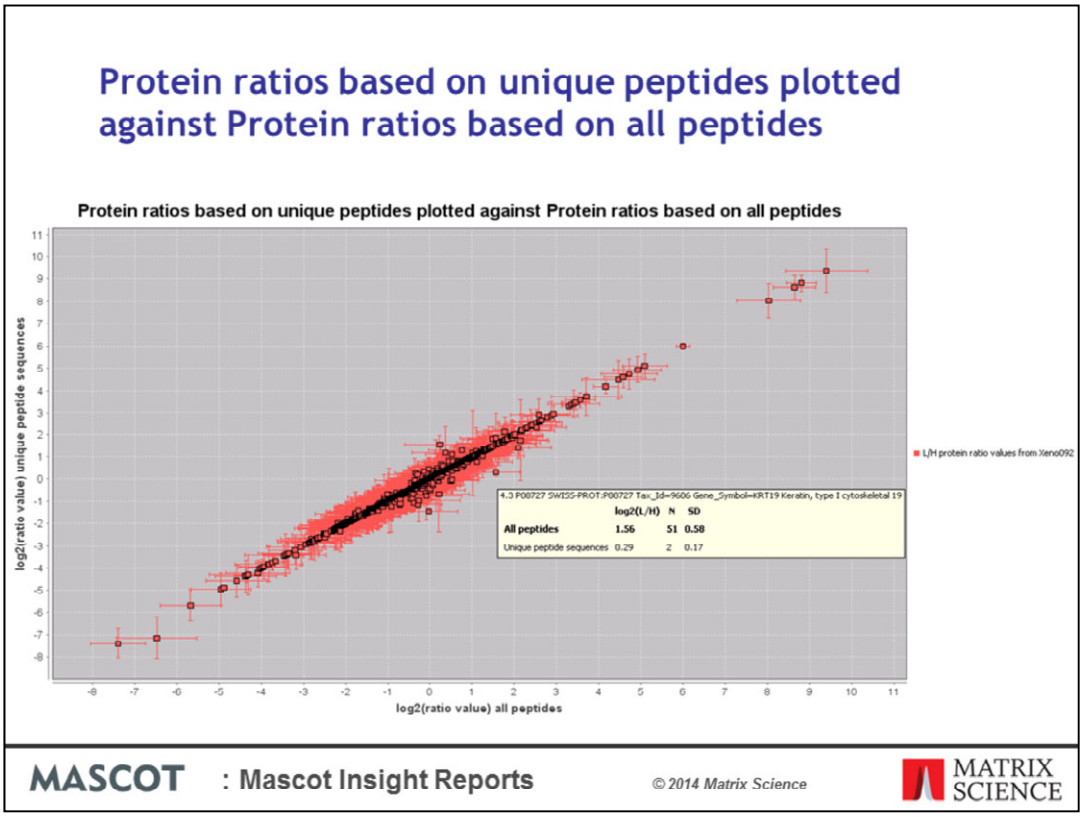


BLASTClust is a program within the standalone BLAST package used to cluster either protein or nucleotide sequences. The program begins with pairwise matches and places a sequence in a cluster if the sequence matches at least one sequence already in the cluster. In the case of proteins, the blastp algorithm is used to compute the pairwise matches.

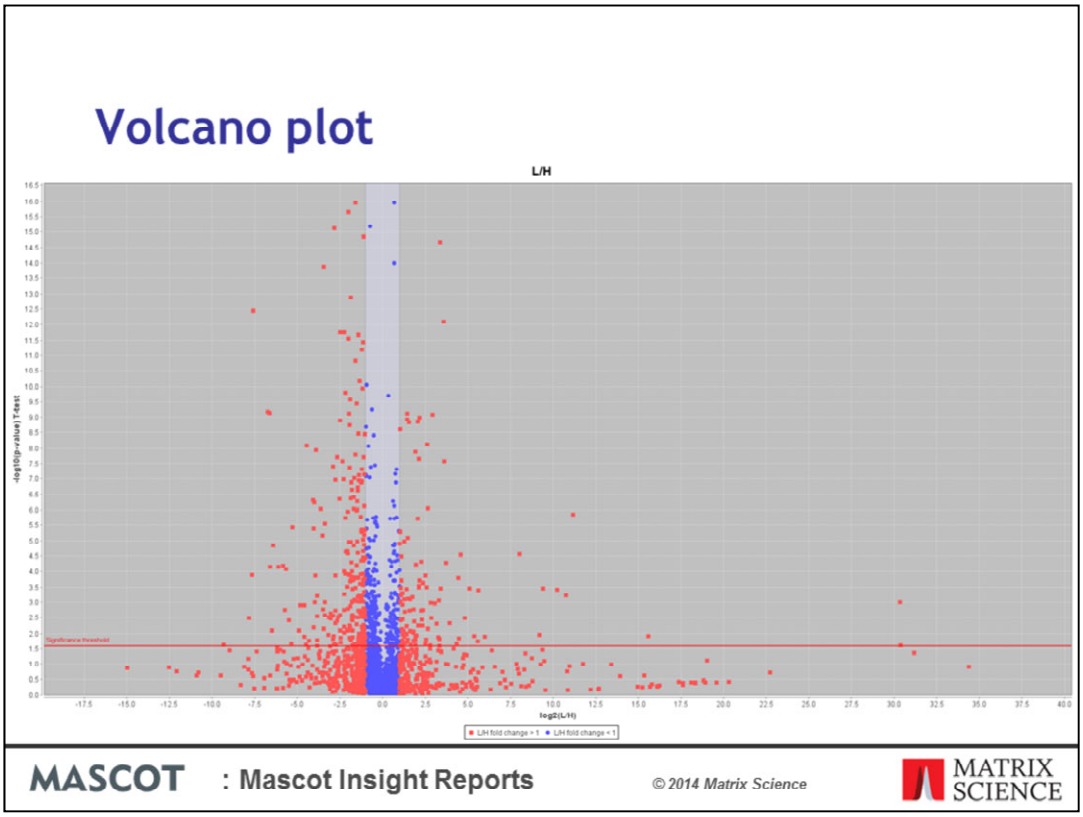
BLASTClust clusters protein hits on the whole protein sequence rather than just the identified peptides. This means that isoforms and protein families will be clustered together while proteins that only share one peptide sequence are unlikely to be grouped together. This makes it an alternative protein grouping algorithm to those used by Mascot Server.



The peptide sharing between Keratins become obvious once you use the BLAST cluster algorithm to group them. There were 77 keratin sequences in the first group.



One final quality plot that illustrates the problem with shared peptides. This is a plot of the Log2 transformed Protein ratios based on unique peptides plotted against Log2 transformed Protein ratios based on all peptides. Proteins with no shared peptides or no difference fall on the diagonal line. However some of the proteins are quite a way of the line with a wide standard deviation, as shown by the error bars. Moussing over the data point or clicking on them give the accession number and information on the protein in question. In this case Keratin type 1 C19 whose unique peptides barely changed with a non significant Log2 transformed ratio of 0.29 (close to 1.2), compared to its shared peptides with a significant up regulation of Log2 transformed ratio of 1.56 (close to 2.9).

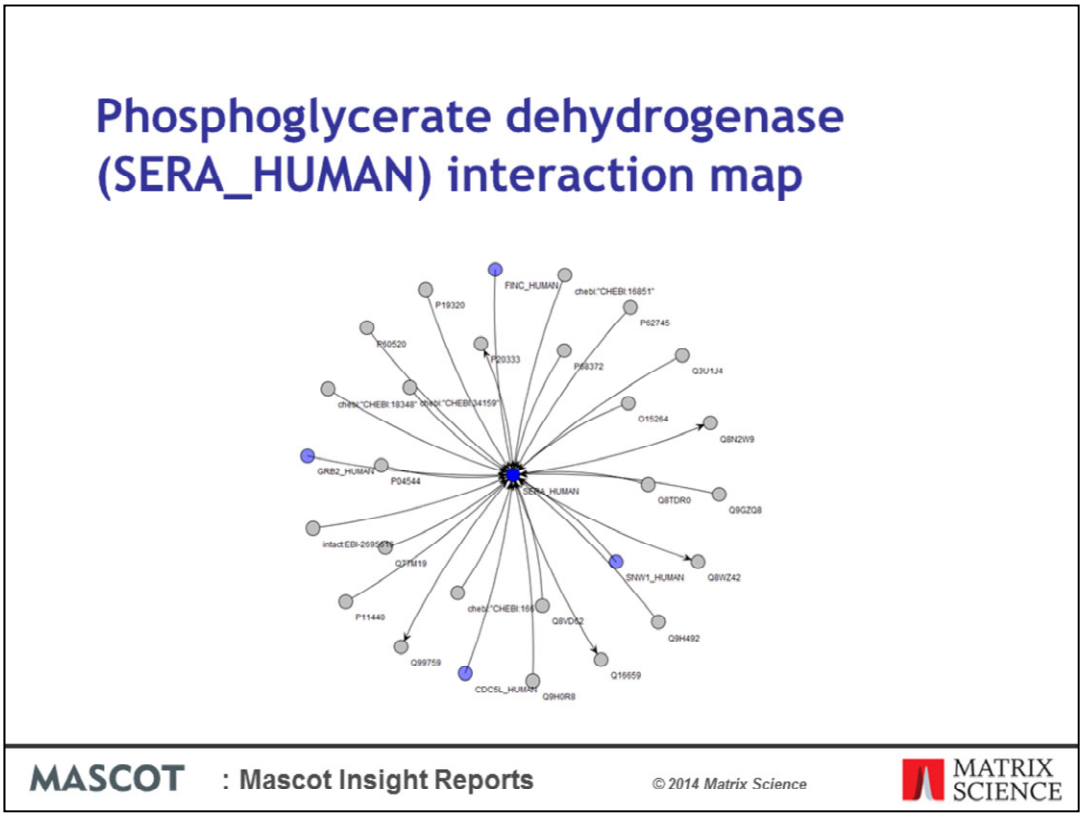


Now that I am satisfied that the overall quality of the data set is good I want to know which are the differently regulated proteins. This is a Volcano plot, generated for one of the ADC fractions, Xeno092-1. A Volcano plot combines the t-statistic for the protein quantitation ratio with the protein quantitation ratio fold change. Using the Volcano plot we can identify the proteins whose protein quantitation ratio value are above the significance threshold and their Protein ratio values after log2 transformation are greater than 1 or -1. These proteins are the significantly up or down regulated proteins of interest. The original paper used protein ratio values after log2 transformation greater than 2 or -2.

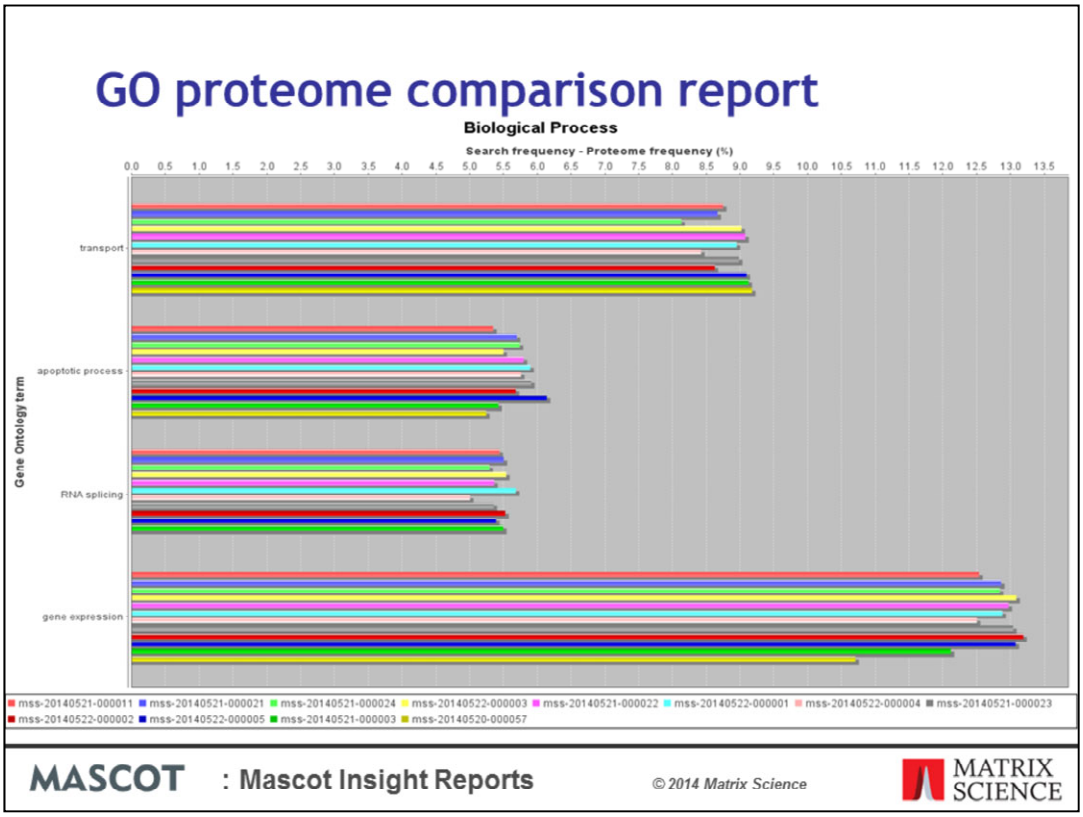
The top 40 up and down regulated proteins as identified from a volcano plot, minus the 4 Keratin proteins that were shown to have not normal peptide probability plots, were used for the Hierarchical clustering report. The Hierarchical clustering report uses unsupervised hierarchical clustering to group both the proteins and the quantitation ratios into clusters and then generates a heatmap of the results. Here we are looking at the heatmap generated form. Red is up and green is down regulated – the hierarchical clustering algorithm has clustered the matching ratios from each sample, and we can see from the heatmap that we can identify candidate proteins to differentiate the two types of tumour based on up and down regulated proteins.

In the original publication the authors go on to use the candidate signature proteins to differentiate an additional 12 tumour samples that were quantified by label-free spectral counting.

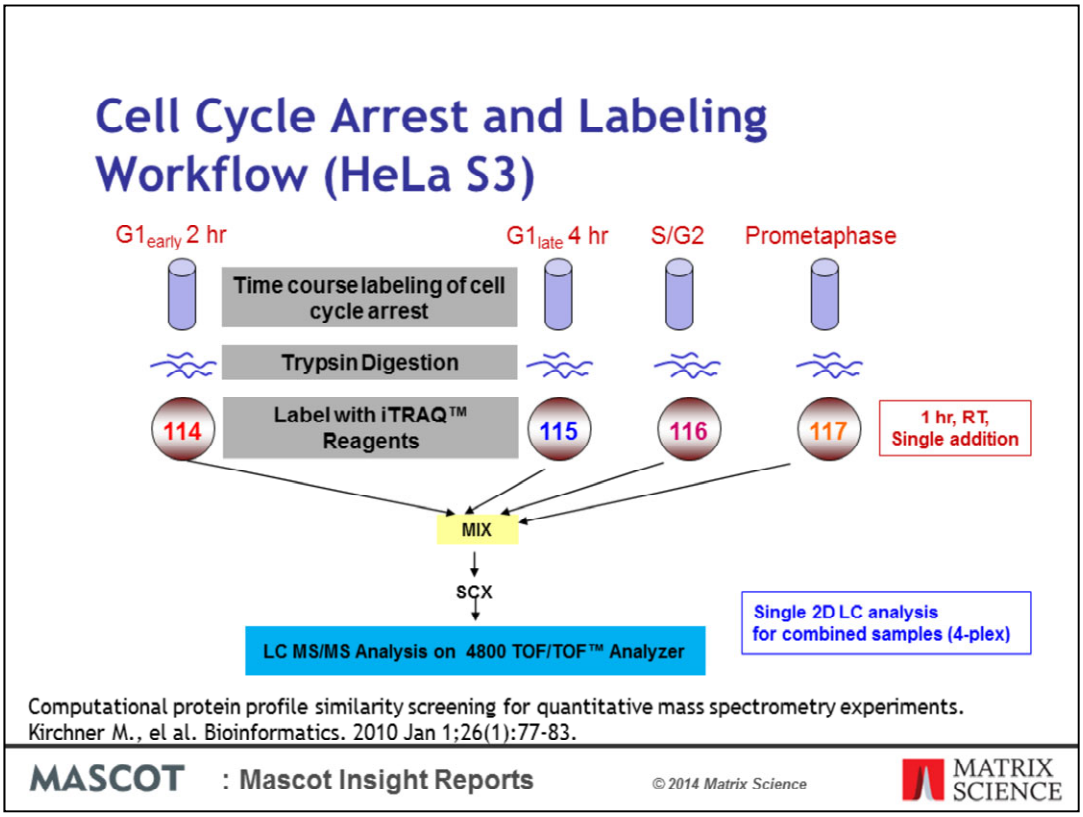
Other quantitation specific plots include a Bland-Altman plot, a QQ plot for peptide ratios matching a protein hit



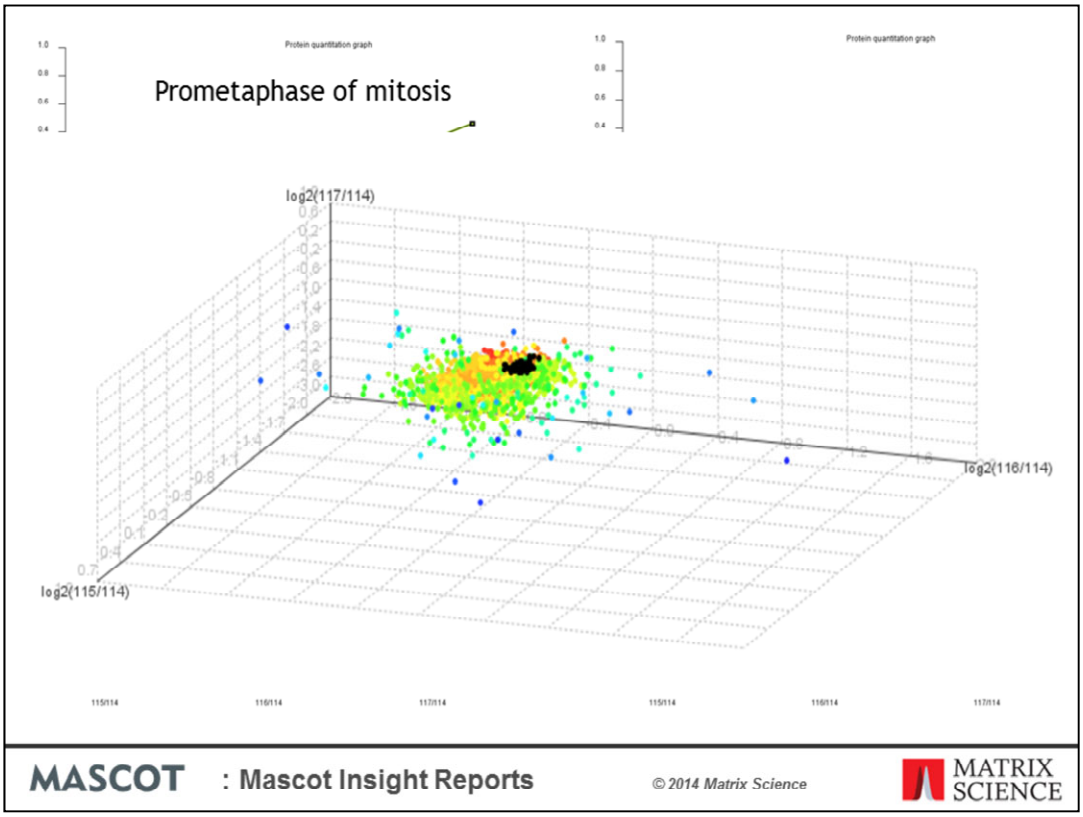
We can explore the candidate "signature" proteins identified in the clustering by running an interaction report on the accession numbers. Here we look at Phosphoglycerate dehydrogenase (SERA_HUMAN) which is know as a key enzyme involved in the re-routing of glycolytic carbon into serine biosynthesis. It was found more highly expressed in SCC, and is known to be upregulated in 70% of estrogen receptor negative breast cancers, and is required for tumorigenesis in in vivo breast cancer models. Proteins in blue have been identified in the data set.



As a final analysis I have plotted a Gene Ontology proteome comparison report. Proteome frequencies are calculated using UniProt species proteomes.
The plot is search frequency - proteome frequency - so over represented GO terms are positive values, under represented would be negative. The report here has been filtered to show on differences >= 5%; so nothing highly under represented but we have a large over representation of proteins involved in gene expression, RNA splicing, apoptosis etc. not unexpected for a rapidly dividing cell line.

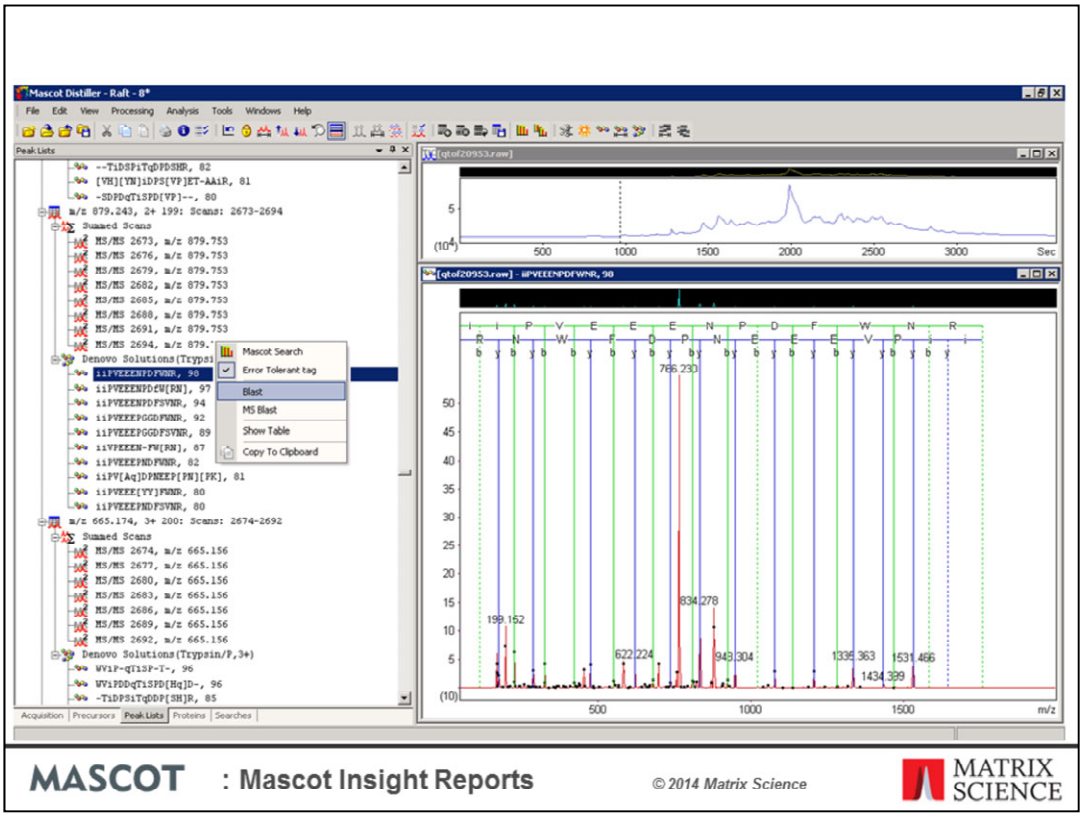


I want to show one other clustering plot, the K-means clustering report. The nature of the Super-SILAC data set, two tumors compared to a labelled standard was not a suitable data set for this type of analysis. Instead I have used a 4plex iTRAQ data set based on samples taken from 4 different points in the cell cycle. The data is from a collaboration between the labs of Darryl Pappin at Cold Spring Harbour Laboratory and Hanno Steen at the Children's Hospital, Boston.



The main difference between K-means clustering and hierarchical clustering is that K-means is a supervised clustering algorithm, which means that you have to specify the number of clusters the algorithm should go an create. Here we have four different clusters from an iTRAQ dataset, where each ratio represents a different points in the cell cycle, G1 early, G1 late, S/G2 and Prometaphase. The four different clusters are showing different behaviour across those points. For example the proteins clustered together in the top left plot increase in concentration during the Prometaphase of mitosis period of the cell cycle while those below it are at their highest concentration during the Synthesis/Gap 2 period.

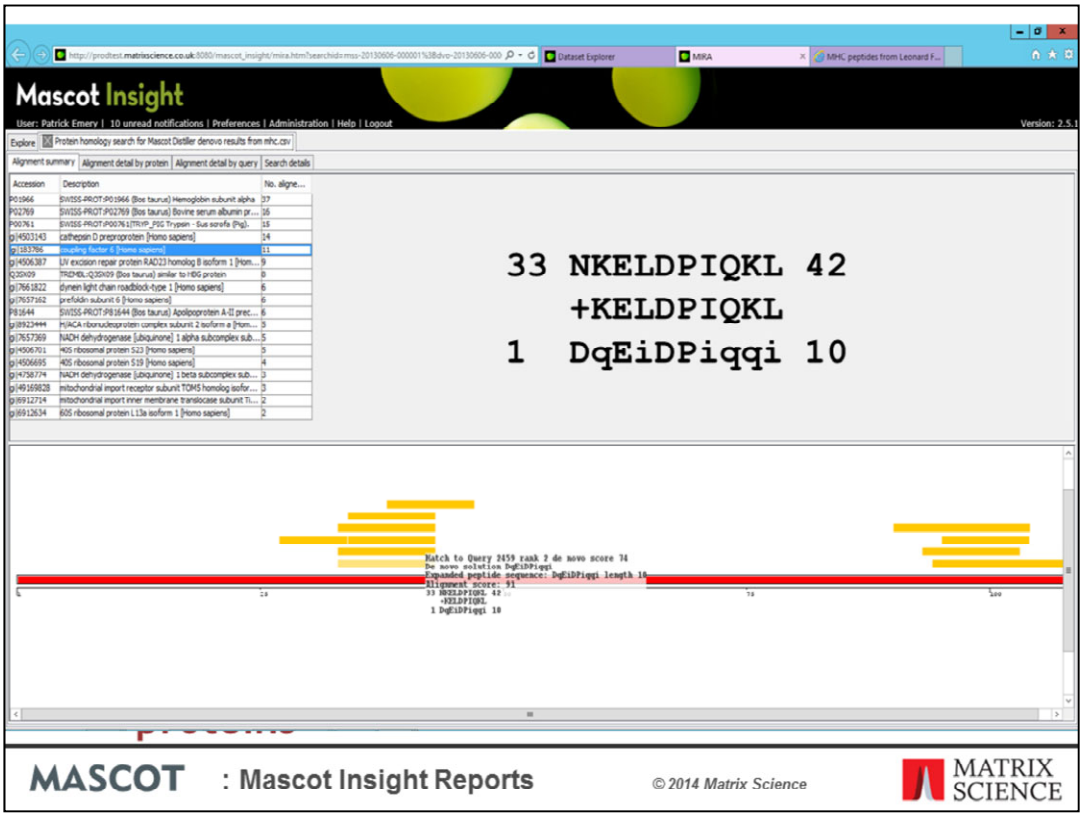
We can also plot a scatter graph of the clusters – here we have the data for that iTRAQ dataset plotted as a 3D scatter plot. The clusters generated are colour coded. You can rotate 3D scatter plots to get a better view of a particular cluster, hover over a data point and you get a tooltip giving you the protein name. You can also select individual clusters in order to highlight them. Here, I've selected the cluster which we're showing in the top left had graph.



This is a very different type of report from the previous ones in that it has nothing to do with quantitation. The Mascot Distiller search toolbox includes a powerful de novo sequencing algorithm. One of the most common reasons for carrying out a de novo sequence search is to try to find matches to MS/MS spectra which failed to get a match from a standard database search, which is what I've done here.

The data set is from a MHC (major histocompatibility complex) analysis using no enzyme and no modifications in the search settings. After the initial search of 1064 spectra there were a total of 290 matches and 774 unassigned spectra.

After the de novo search has completed we have a series of de novo sequence matches of various quality. One of the problems de novo sequencing is trying to verify the results – In this case, have we identified any additional matches to proteins we identified in the initial Mascot database search? The standard way try to answer this question would be to submit the high scoring de novo solutions for a sequence homology search, such as a BLAST or MS-BLAST search, or for an error tolerant sequence tag search. These approaches can be very time consuming as the process of submitting the searches is largely manual.

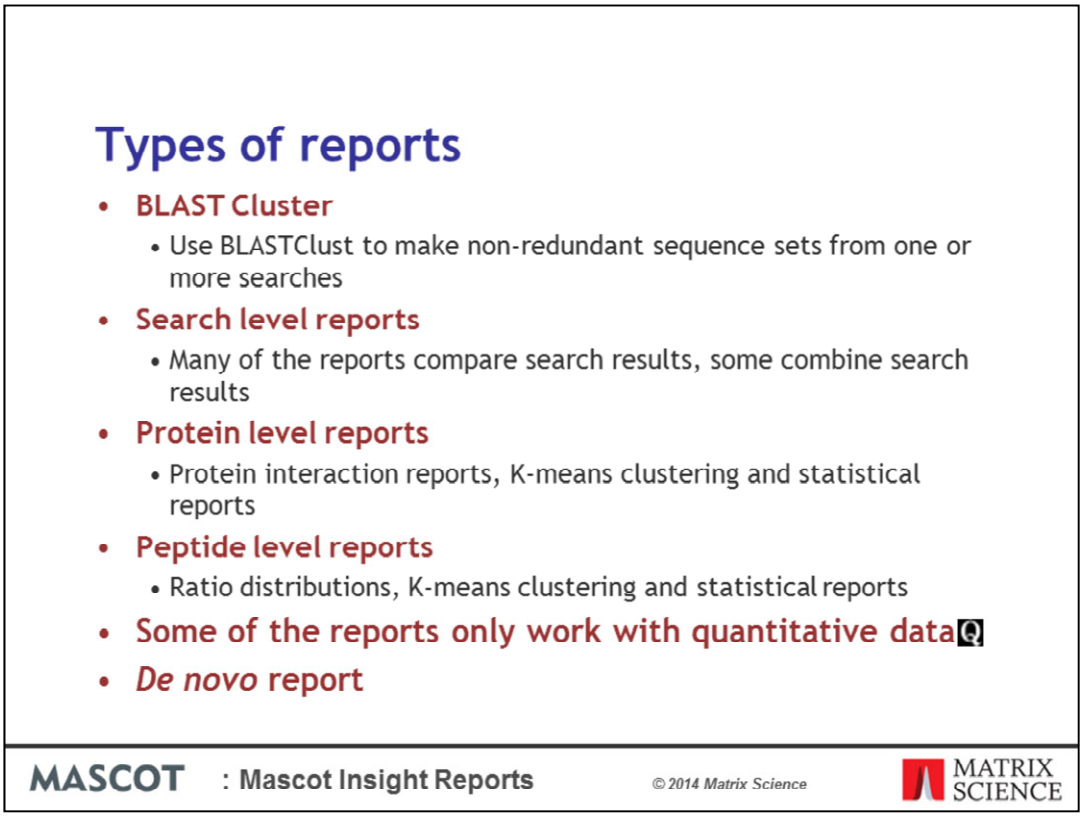


In Mascot Insight, we have implemented a report which allows you to use the protein hit sequences from a selected search result to carry out a BLAST like sequence homology search using de novo solutions, in order to try to find additional possible matches to spectra from your dataset, drilling down into the unassigned MS/MS spectra. The sequence homology search has been tailored to de novo solution data, and allows for Q->K and F->M* in the alignment without penalty for example.

Out of the de novo results we took the 156 spectra with 1 or more solutions and a *De novo* score >= 40, that is the good quality de novo matches. These 156 spectra expanded into 1082 *de novo* solutions which were tested against the 106 proteins identified from the original database search (does many more alignments than this because ambiguity in sequences is 'exploded'). The search results were 79 spectra with good matches to a previously identified protein.

(e.g. you could potentially ignore half of the of the spectra (79/156) with the best de-novo solutions when MHC hunting)

This allows us to see if we have any convincing de novo sequence matches to peptides from the protein hits from the initial Mascot database search – for example, this looks like a good match to a deamidated sequence also identified in the original search where the differences at the n-terminus prevented us from getting a sequence database match.

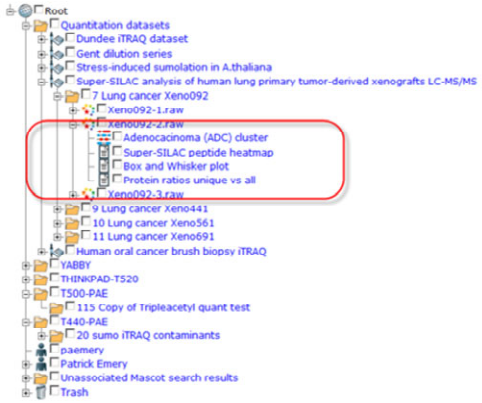


That's the end of the data analysis examples and I would like to quickly review the types of reports that Mascot Insight ships with.

And there is one special case of report a *De novo* report, which I have just shown, that works with the *De novo* results generated by Mascot Distiller.

Sharing reports

- **Export reports**
  - Variety of formats, including CSV, SVG, PNG and HTML.
  - Publication export (MCP guidelines)
  - Or saved to the dataset explorer tree
- **Direct access to Mascot Insight**
  - For interactive reports collaborators can log on to Insight and evaluate the results directly
  - No limits to the number of concurrent users
  - Control access to the reports and Insight features via built in security

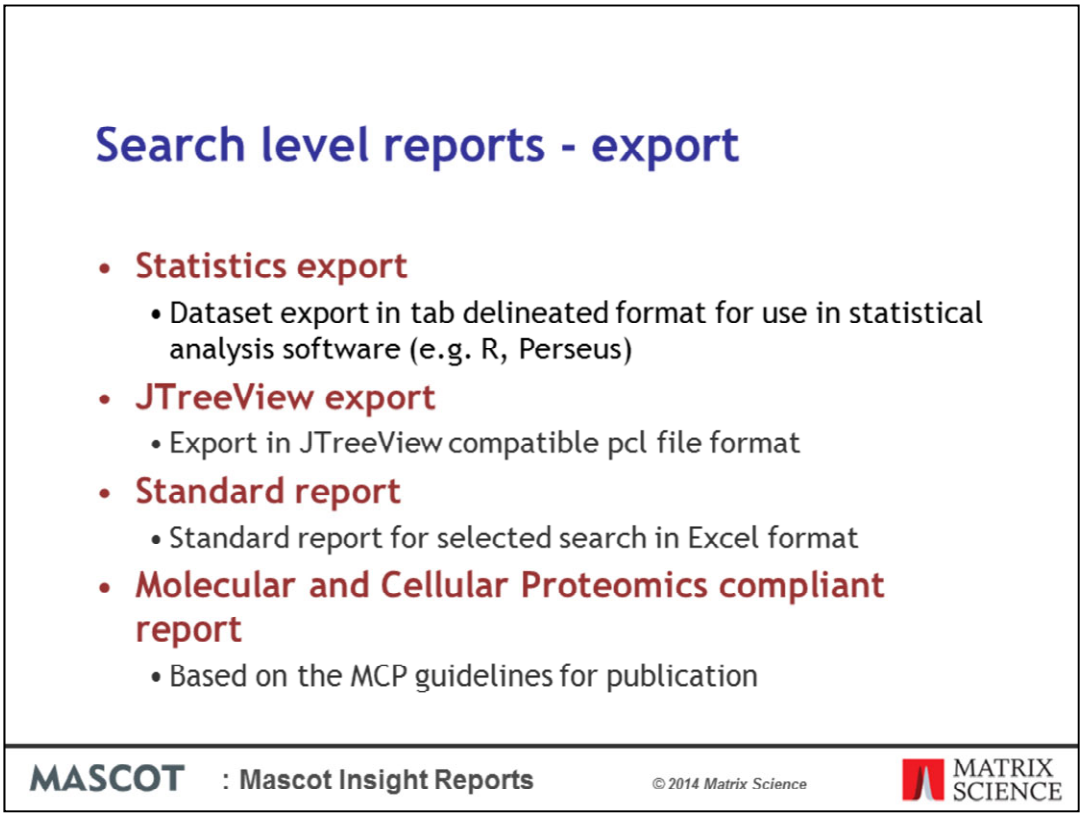


MASCOT : Mascot Insight Reports MATRIX SCIENCE

I have shown examples of many different kinds of reports and I want to make it clear that reports are not locked into Insight. Once a report is generated, they can be exported in a variety of formats, including CSV, SVG, PNG and HTML, or saved to the dataset explorer tree to be easily viewed at a later date.

Members of the lab or collaborators can be given access to the data to make their own analysis or view any saved reports. As mentioned in the previous talk there is no limit on the number of users for the software and that goes for concurrent users too.

Access to the results is controlled by Mascot Server security using the standard groups and uses configuration.



You can also export data out of Insight for use with other software.

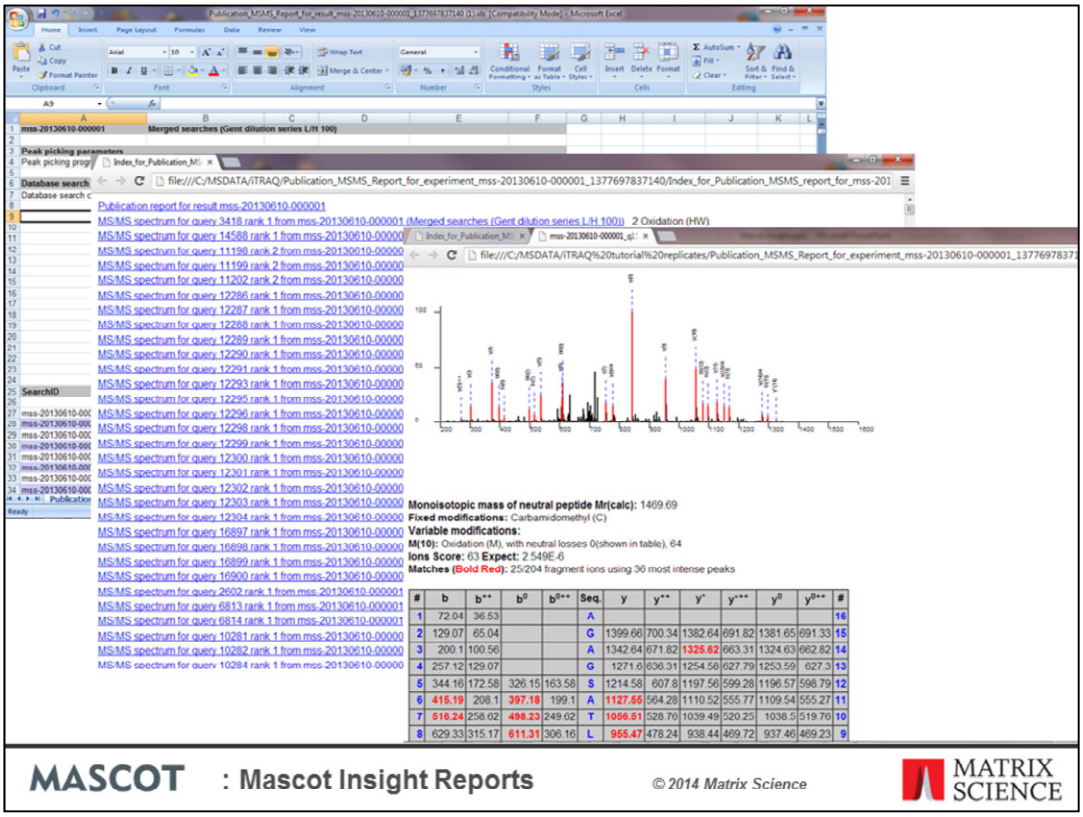
The different export reports work for one or more search results and format then export data for use by third party software, simple excel use or for publication compliant reports.

Files created by the Statistics export, which exports protein, peptide and quantitation data in a tab delineated format, can be used by third party software like the R statistics package or Perseus from the Matthias Mann group at MPI Munich.

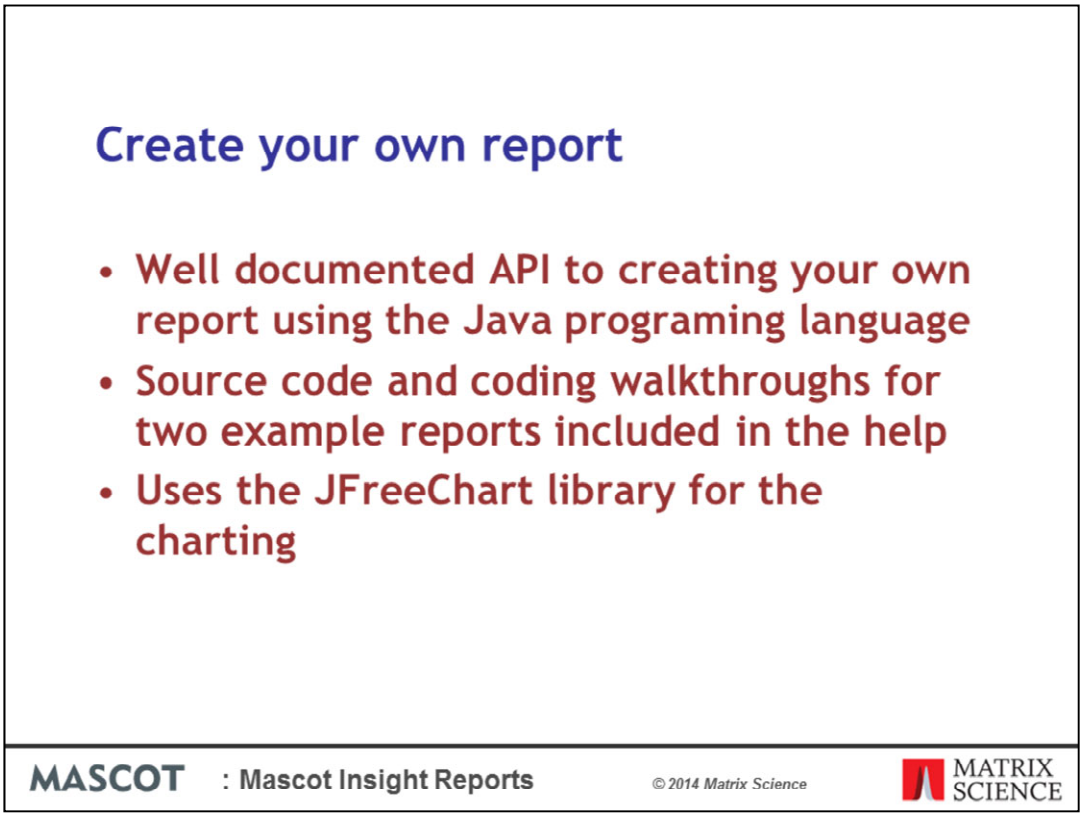
The Java TreeView format generates input files from quantified data read for use with the Java TreeView. Java TreeView is a program that visualizes hierarchical clustering of gene expression data, or in our case protein expression data.

The Standard report is equivalent to the a standard csv export from a Mascot Server web page report. It is only available for search results that have been merged together in Mascot Insight and for MzIdentML and ProtXML results. If the search result is available on the Mascot Server then this option will be disabled.

The MCP report meets the reporting guidelines from MCP for publication of data. The end result is a zip file that contains a report with all the protein, peptide and spectral information required by MCP.



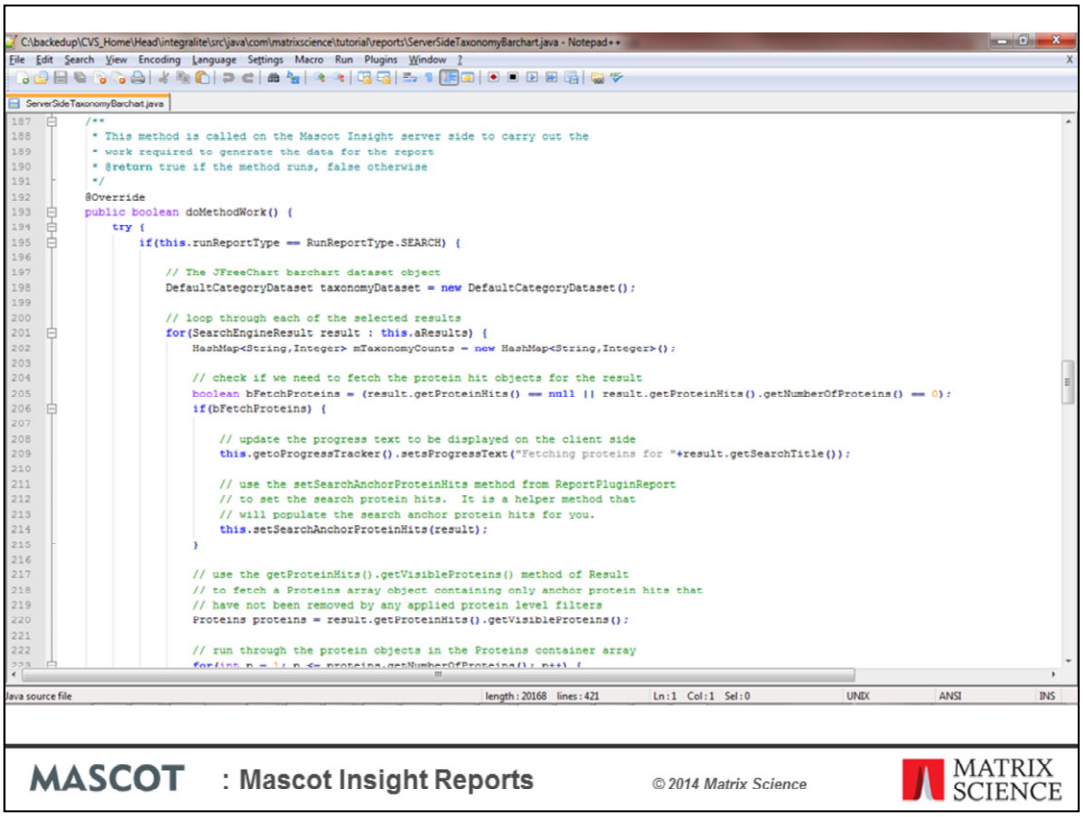
Here is an example of the files generated by the Publication export report, which exports data for selected search results in a report designed to meet a number of the criteria specified by the Molecular and Cellular Proteomics guidelines for publication. The Publication report generates a zip archive containing an index HTML page, with links to the main Excel report file and links to static HTML pages which contain the spectrum views for MS/MS peptide matches (or MS protein matches for PMF data). A separate archive is generated for each search result, or you can generate a single report for a merged dataset.



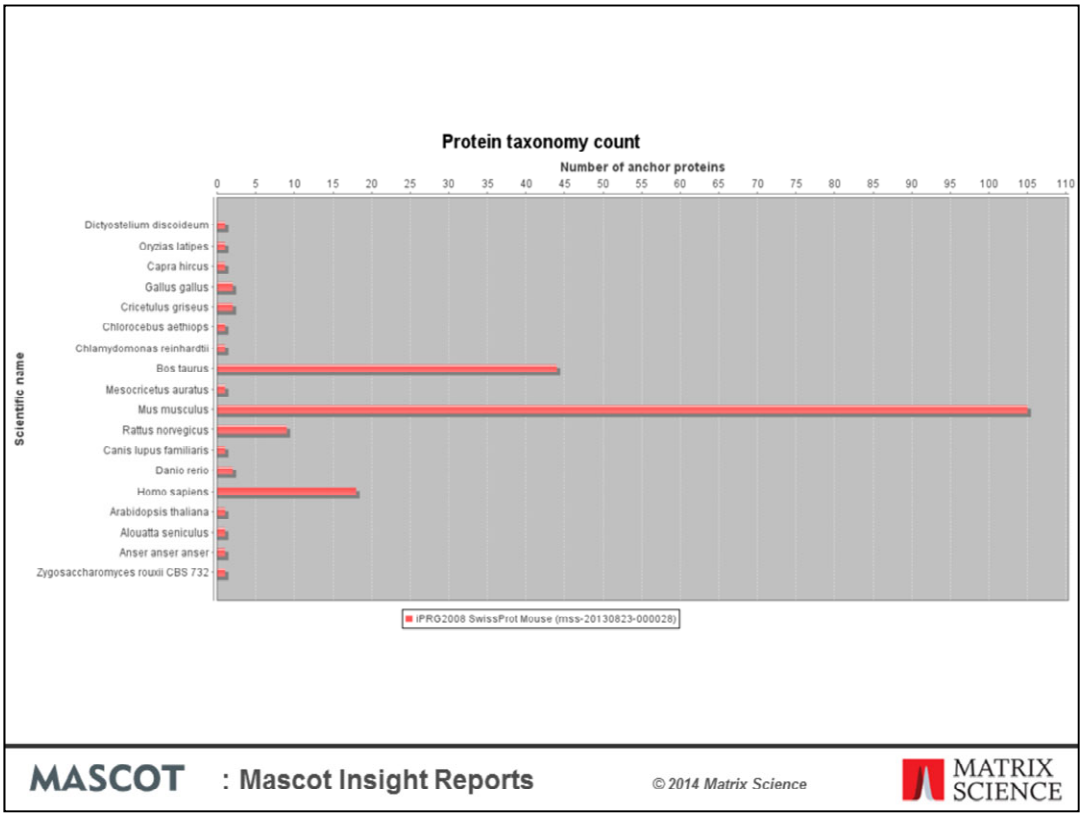
That brings me to the end of my examples of the reports currently included with Mascot Insight. In addition to the shipped reports, you can write your own – to do this you will need to have a working knowledge of the Java programming language.

We include a well documented API to creating your own report along with two documented examples to get you started.

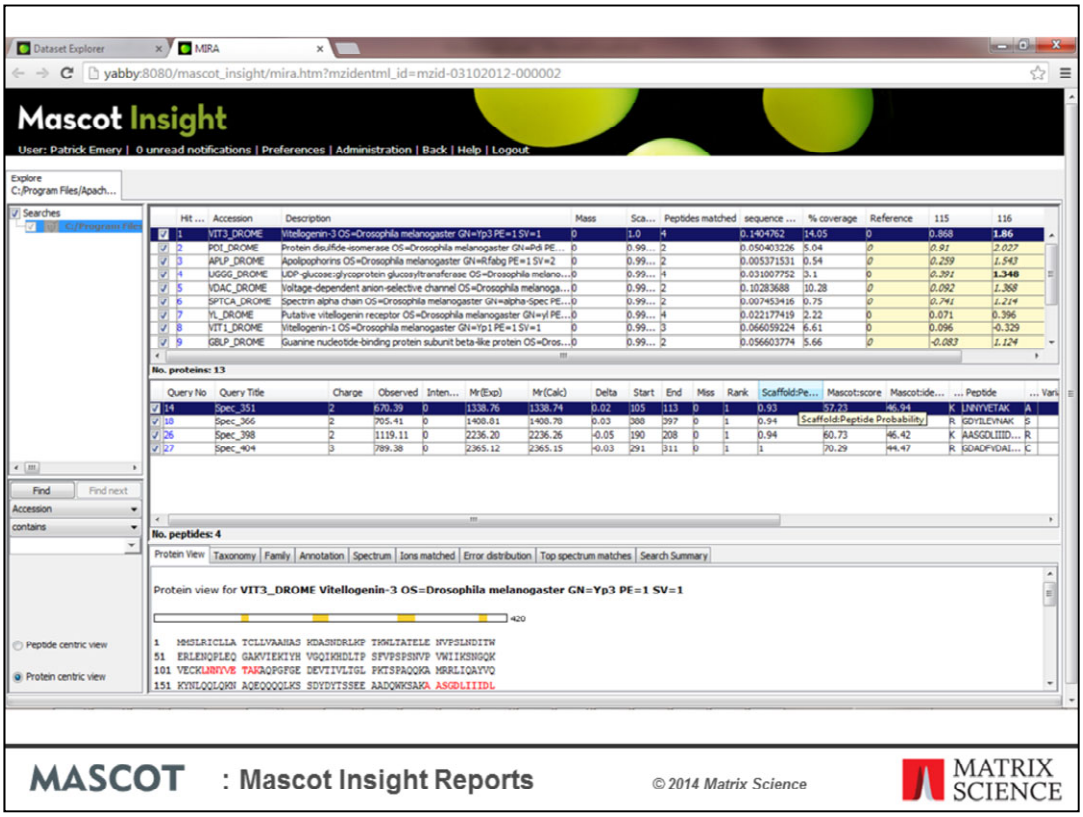
MI uses the popular JFreeChart library for the charting.



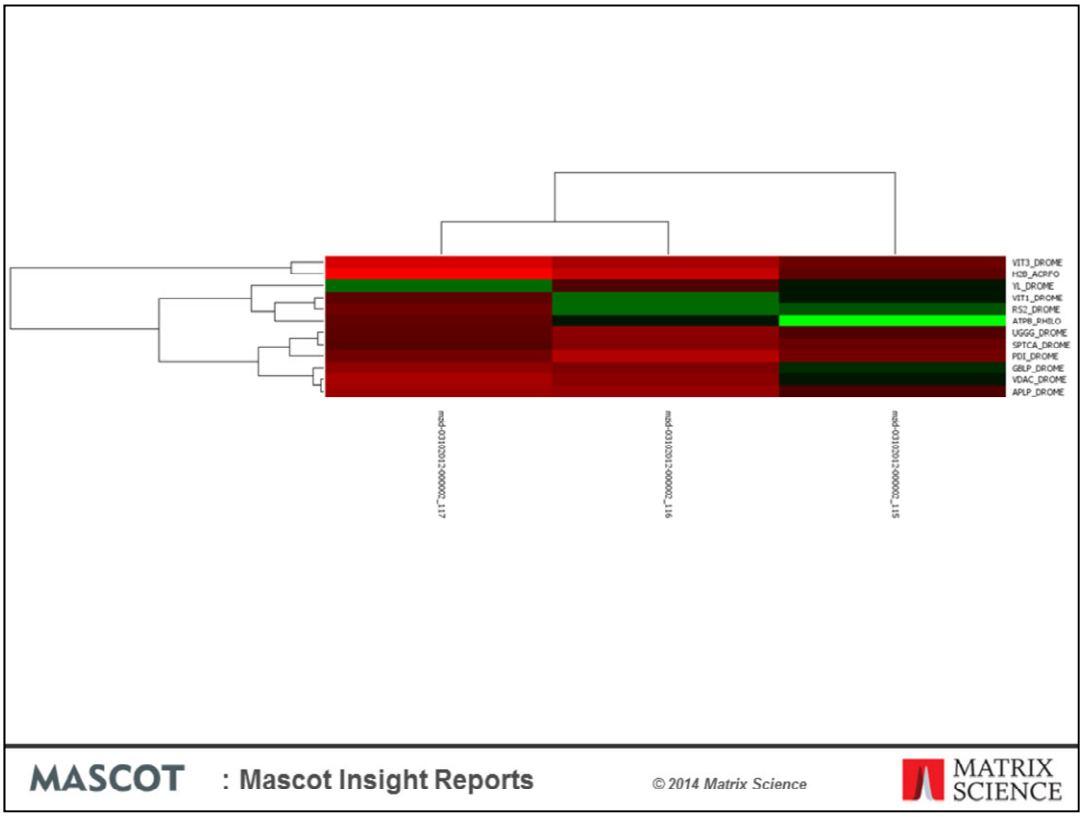
You can write both chart type reports for use in MIRA, and export reports to generate export files in any format you choose. The example shown here will generate a barchart of taxonomy ids for a selected search result. The API we're using here is a high level API, so that the report works without any modification for Mascot results, and for searches imported from protXML and mzIdentML datasets.



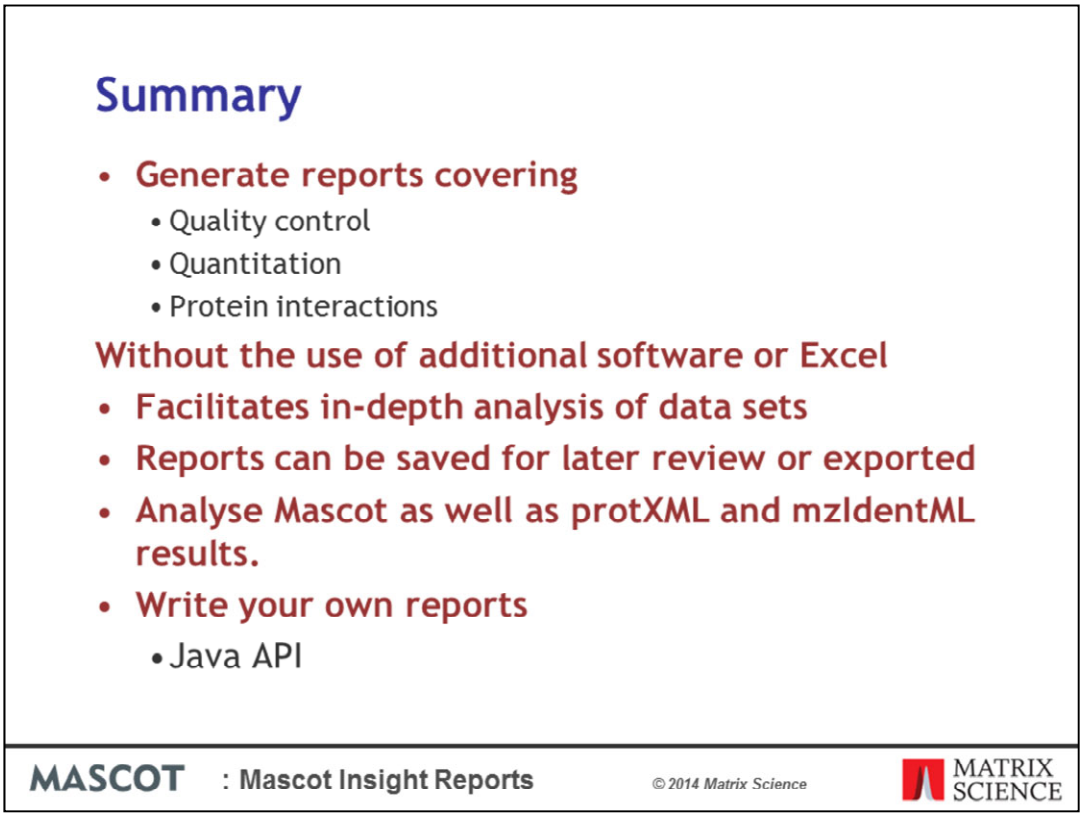
And this is the output from the report run against the iPRG2008 dataset searched against the whole of SwissProt, with the largest number of protein hits being to mouse as expected since this is a mouse dataset.



As I mentioned earlier in the presentation, you can import search result data from any source in either the protXML of mzIdentML standard formats. Here we are looking at an mzIdentML format result exported from Scaffold Q+, and we've also imported the Scaffold quantML quantitation result export file as well. We can use these results to generate reports, in the same way that we can do for Mascot search results.



As an example, here is the output from the hierarchical clustering report for this data set



So, in summary, Mascot Insight ships with a large number of reports, which cover areas such as protein and peptide level comparisons between data sets and quantitation based reports including reports for comparison, ratio clustering and quality control. This is all done in Insight without the use of additional software or Excel.

You can also look at gene ontology and Protein interactions including mapping the quantitation results on the onto the interaction diagram.

As I have shown using Mascot Insight allows you to carry out in-depth analysis of a data set to determine both the quality of the data and to discover the significantly regulated proteins.

The results of these reports can be saved with the project or copied to Word or Powerpoint documents or exported into popular graphics formats.

You can analyse not only Mascot Server results but also results from any software that exports protXML or mzIdentML results

And, if you have a specific report you want to be able to generate, you can write your own reports in Java using a common API across different result formats