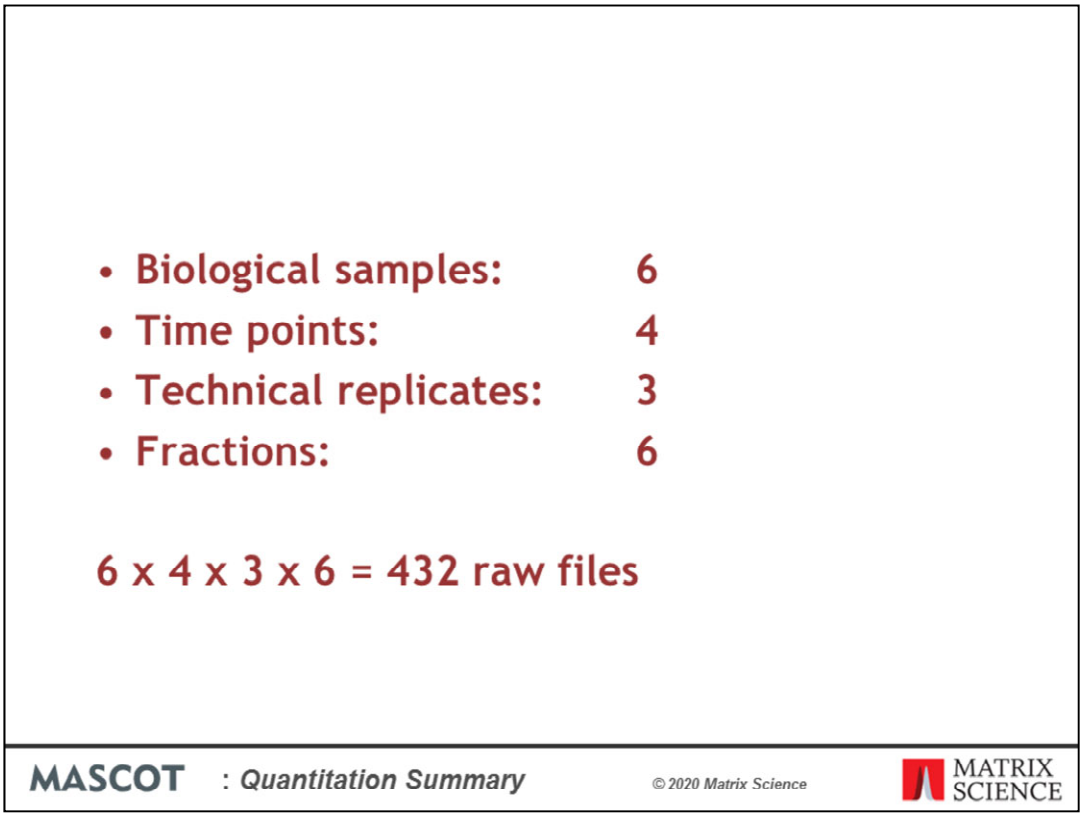
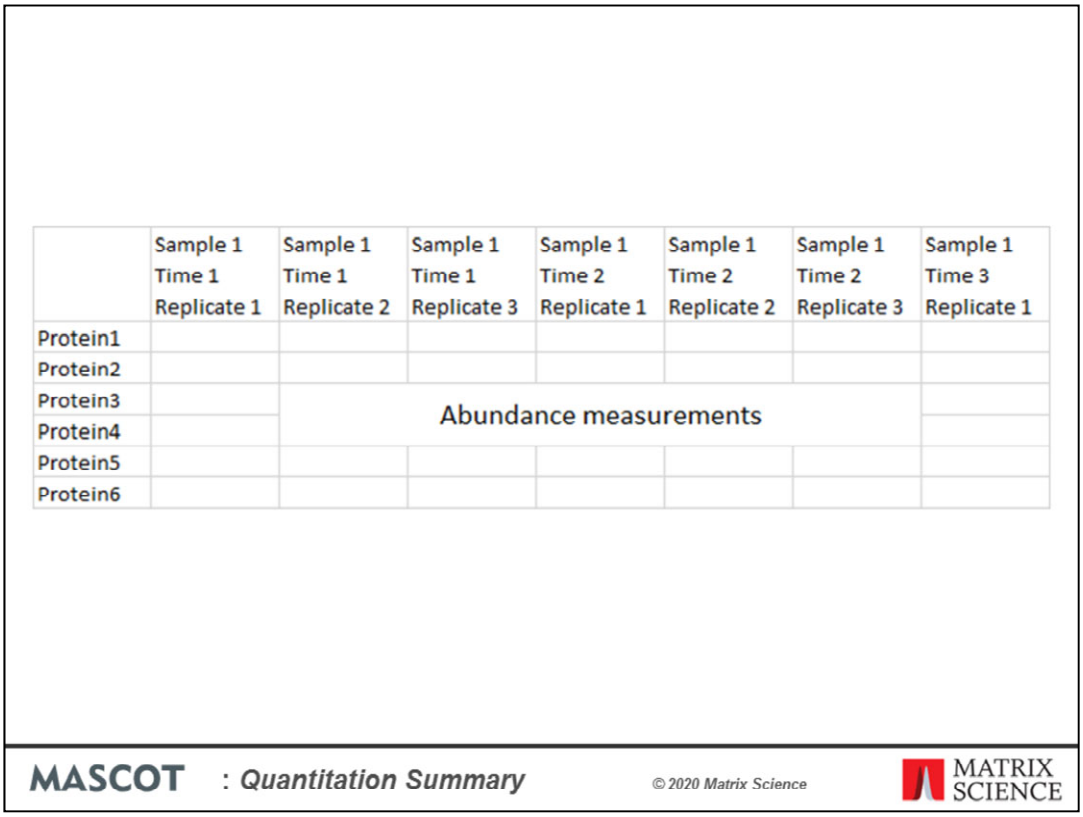
Quantitation Summary: exporting protein expression data for complex experiments

John Cottrell
Matrix Science



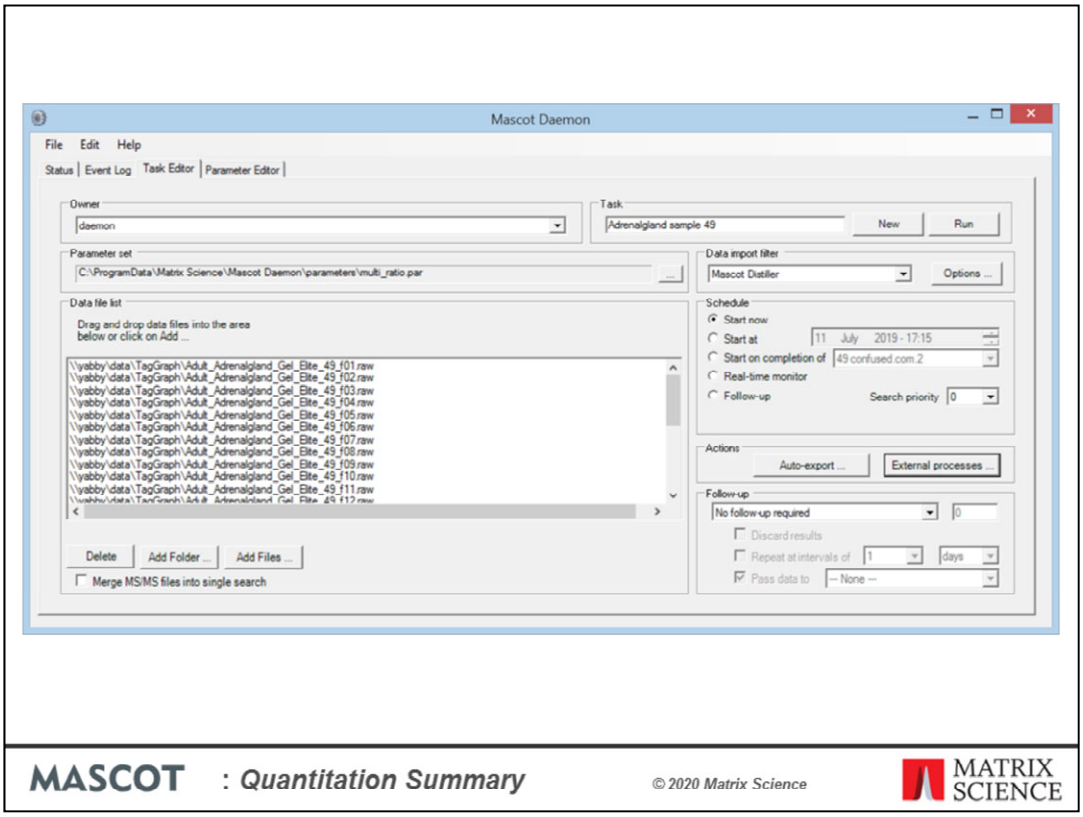
Studies that use mass spectrometry-based quantitation often contain very large numbers of individual analyses: samples from different sources or treatments or time points, possibly fractionated, with replicates and so forth.

Even a simple study can generate a large number of files. For example, imagine a label-free study of 6 biological samples, 3 control and 3 treated, each of which has been analysed at 4 time points in 3 technical replicates and each replicate has been separated into 6 fractions prior to analysis. This would result in 432 raw files.



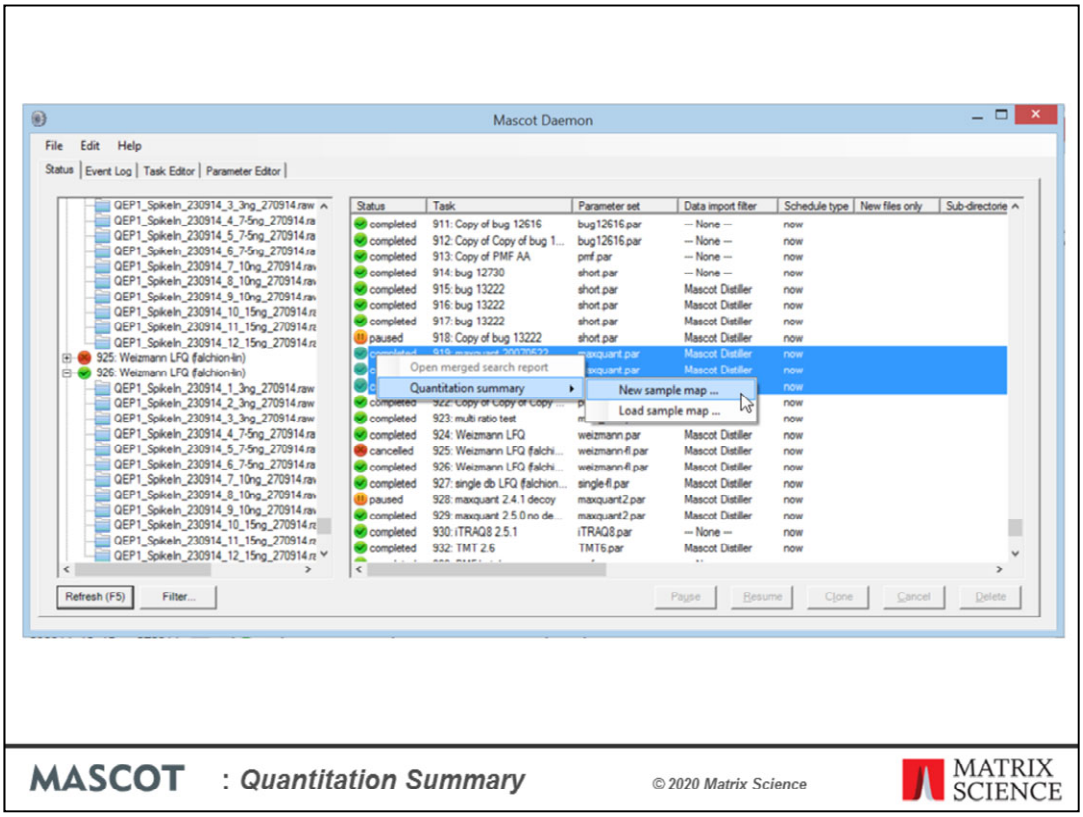
Using statistical methods to extract meaningful information, and report it as charts and tables is a complex task that requires custom scripting in a language such as R or specialised software such as Perseus. These take their input in spreadsheet form, called a Quantitation Summary, where the rows correspond to proteins and the columns contain expression data for the various samples in the form of abundances or ratios of abundances.

Until the release of Mascot Server 2.7, we did not have a convenient way to create a Quantitation Summary from individual Mascot Server or Mascot Distiller result files. Mascot Daemon now includes this functionality. Searches run through Daemon that include label or label-free quantitation, including reporter methods such as iTRAQ and TMT, can be combined and annotated to create just such a Quantitation Summary.

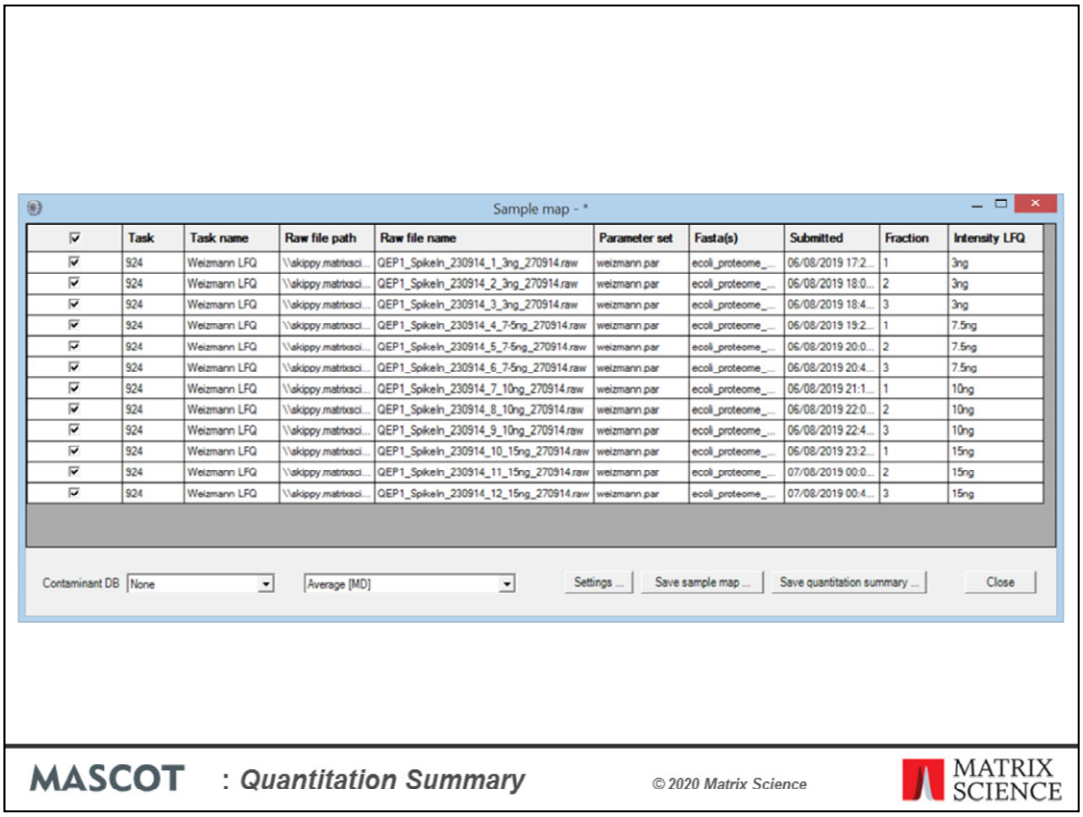


MASCOT : *Quantitation Summary* *© 2020 Matrix Science* MATRIX SCIENCE

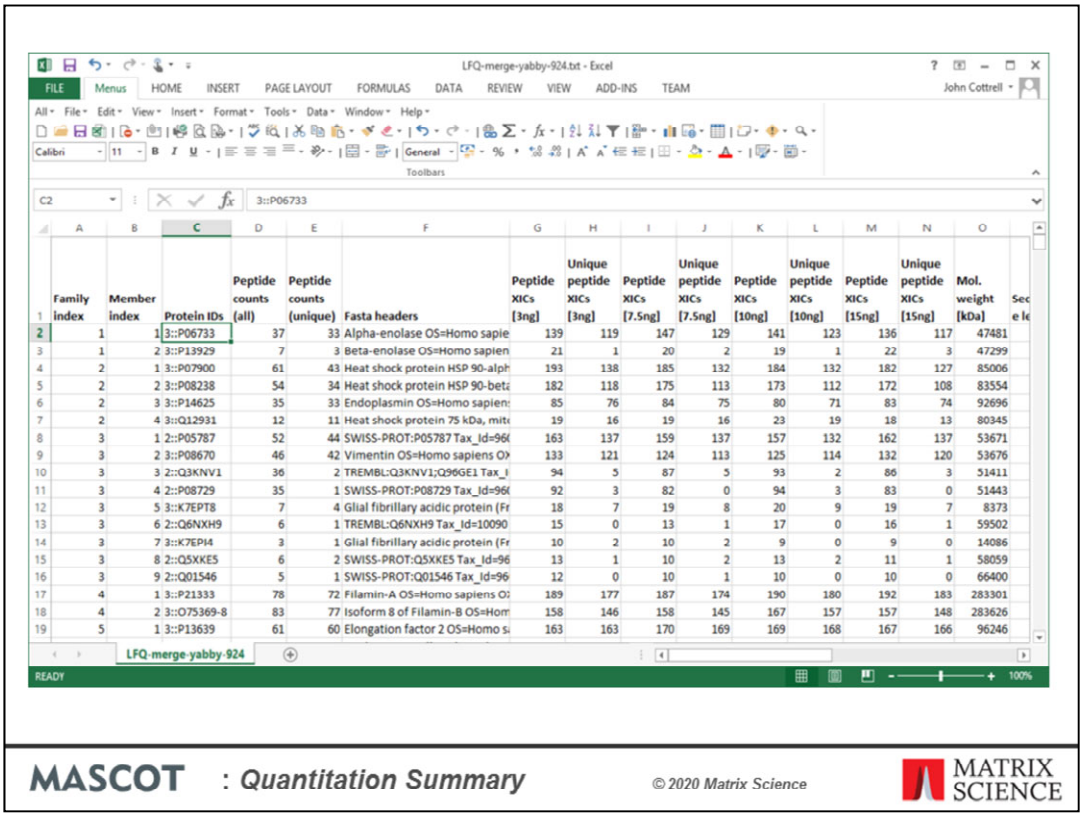
The steps are, first, use Daemon to submit the search and initiate quantitation. The analyses can be spread across any number of Daemon tasks, and can include existing results from earlier versions of Daemon. For reporter ion experiments, you can use Distiller for peak picking, but this is not a requirement. For MS1 quantitation methods, such as SILAC and label-free, the raw files must be peak picked and quantified by Mascot Distiller.



Second, you select the relevant tasks and create a Sample Map.

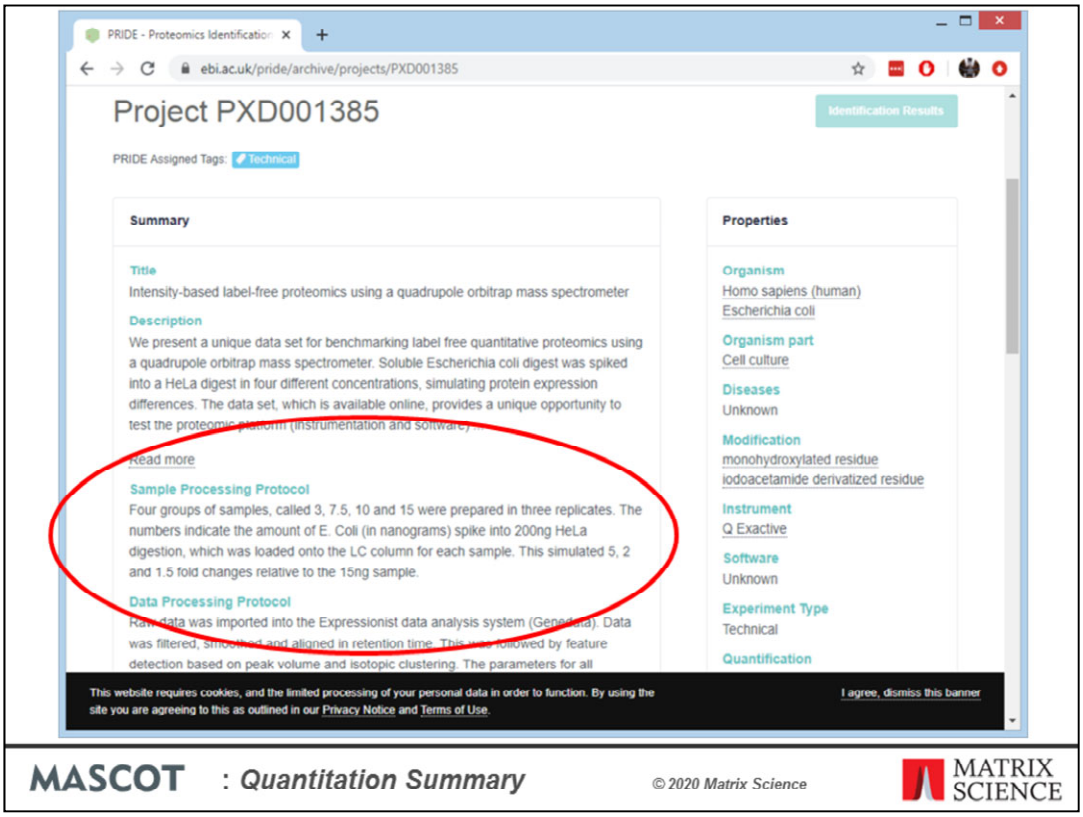


The Sample Map is used to annotate the list of result files with recognisable sample identifiers. In this case we just need to complete the two right-hand columns. Third, choose ‘Save quantitation summary’

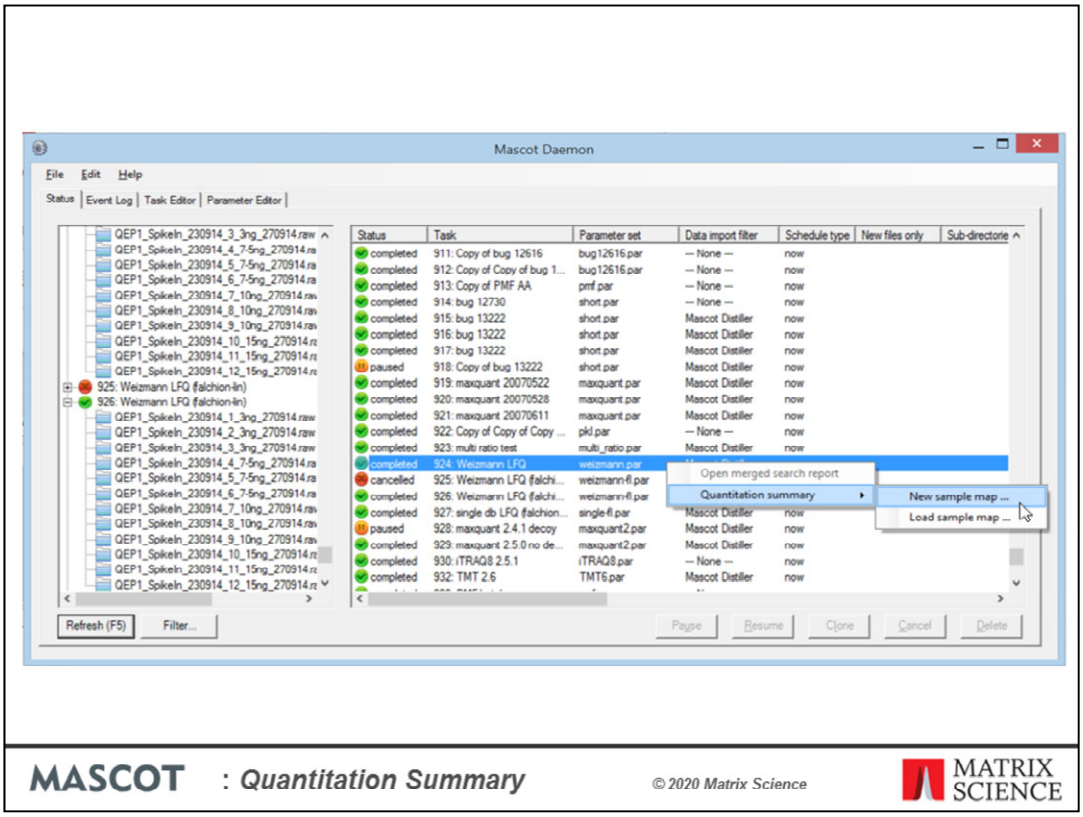


	Family index	Member index	Protein IDs	Peptide counts (all)	Peptide counts (unique)	Fasta headers	Peptide XICs [3ng]	Unique peptide XICs [3ng]	Peptide XICs [7.5ng]	Unique peptide XICs [7.5ng]	Peptide XICs [10ng]	Unique peptide XICs [10ng]	Peptide XICs [15ng]	Unique peptide XICs [15ng]	Mol. weight [kDa]
2	1	1	3::P06733	37	33	Alpha-enolase OS=Homo sapie	139	119	147	129	141	123	136	117	47481
3	1	2	3::P13929	7	3	Beta-enolase OS=Homo sapien	21	1	20	2	19	1	22	3	47299
4	2	1	3::P07900	61	43	Heat shock protein HSP 90-alph	193	138	185	132	184	132	182	127	85006
5	2	2	3::P08238	54	34	Heat shock protein HSP 90-beta	182	118	175	113	173	112	172	108	83554
6	2	3	3::P14625	35	33	Endoplasmin OS=Homo sapien	85	76	84	75	80	71	83	74	92696
7	2	4	3::Q12931	12	11	Heat shock protein 75 kDa, mit	19	16	19	16	23	19	18	13	80345
8	3	1	2::P05787	52	44	SWISS-PROT:P05787 Tax_Id=96	163	137	159	137	157	132	162	137	53671
9	3	2	3::P08670	46	42	Vimentin OS=Homo sapiens O	133	121	124	113	125	114	132	120	53676
10	3	3	2::Q3KNV1	36	2	TREMBL:Q3KNV1;Q96GE1 Tax_I	94	5	87	5	93	2	86	3	51411
11	3	4	2::P08729	35	1	SWISS-PROT:P08729 Tax_Id=96	92	3	82	0	94	3	83	0	51443
12	3	5	3::K7EPT8	7	4	Glial fibrillary acidic protein (Fr	18	7	19	8	20	9	19	7	8373
13	3	6	2::Q6NXH9	6	1	TREMBL:Q6NXH9 Tax_Id=10090	15	0	13	1	17	0	16	1	59502
14	3	7	3::K7EPI4	3	1	Glial fibrillary acidic protein (Fr	10	2	10	2	9	0	9	0	14086
15	3	8	2::Q5XKE5	6	2	SWISS-PROT:Q5XKE5 Tax_Id=96	13	1	10	2	13	2	11	1	58059
16	3	9	2::Q01546	5	1	SWISS-PROT:Q01546 Tax_Id=96	12	0	10	1	10	0	10	0	66400
17	4	1	3::P21333	78	72	Filamin-A OS=Homo sapiens O	189	177	187	174	190	180	192	183	283301
18	4	2	3::O75369-8	83	77	Isoform 8 of Filamin-B OS=Hom	158	146	158	145	167	157	157	148	283626
19	5	1	3::P13639	61	60	Elongation factor 2 OS=Homo s	163	163	170	169	169	168	167	166	96246

And Daemon creates and saves the Quantitation Summary, a tab delimited text file which is the input file for statistical analysis. The best way to explain how this works in detail is with a couple of examples

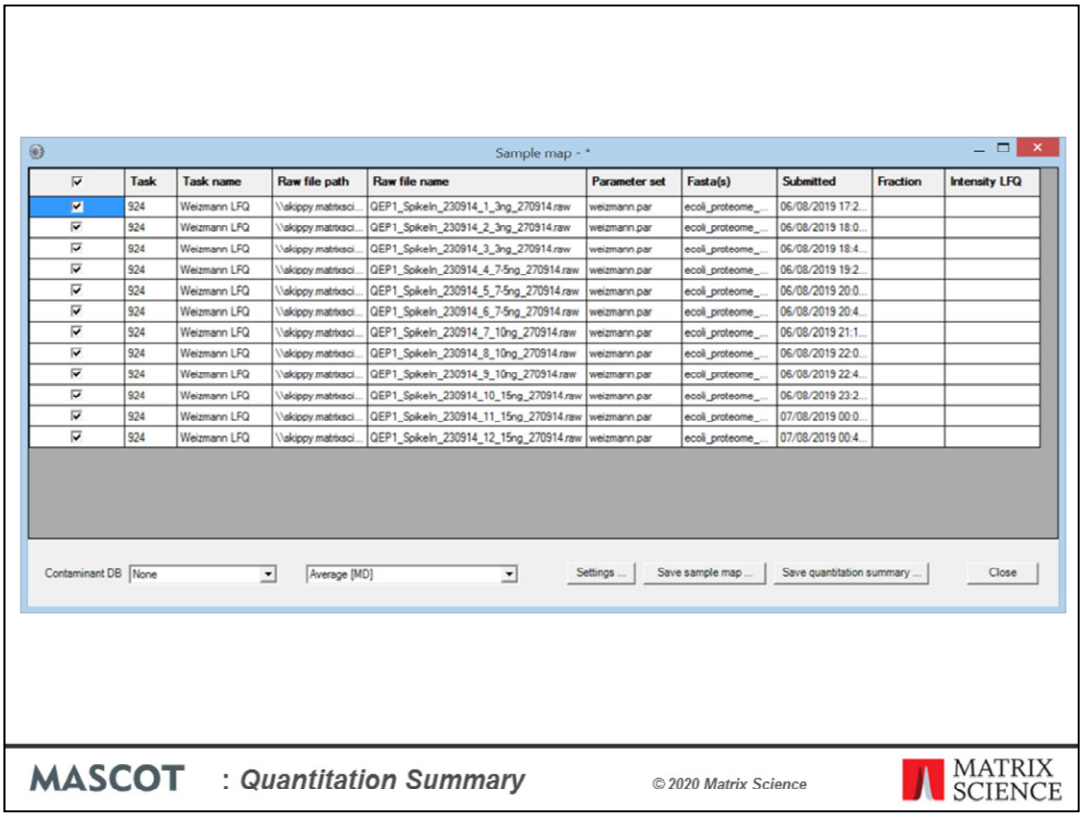


The first example is a label-free experiment. A set of 12 raw files was downloaded from PRIDE project PXD001385. According to the project description: "Four groups of samples, called 3, 7.5, 10 and 15 were prepared in three replicates. The numbers indicate the amount of E. Coli (in nanograms) spiked into 200ng HeLa digestion, which was loaded onto the LC column for each sample. This simulated 5, 2 and 1.5 fold changes relative to the 15ng sample."

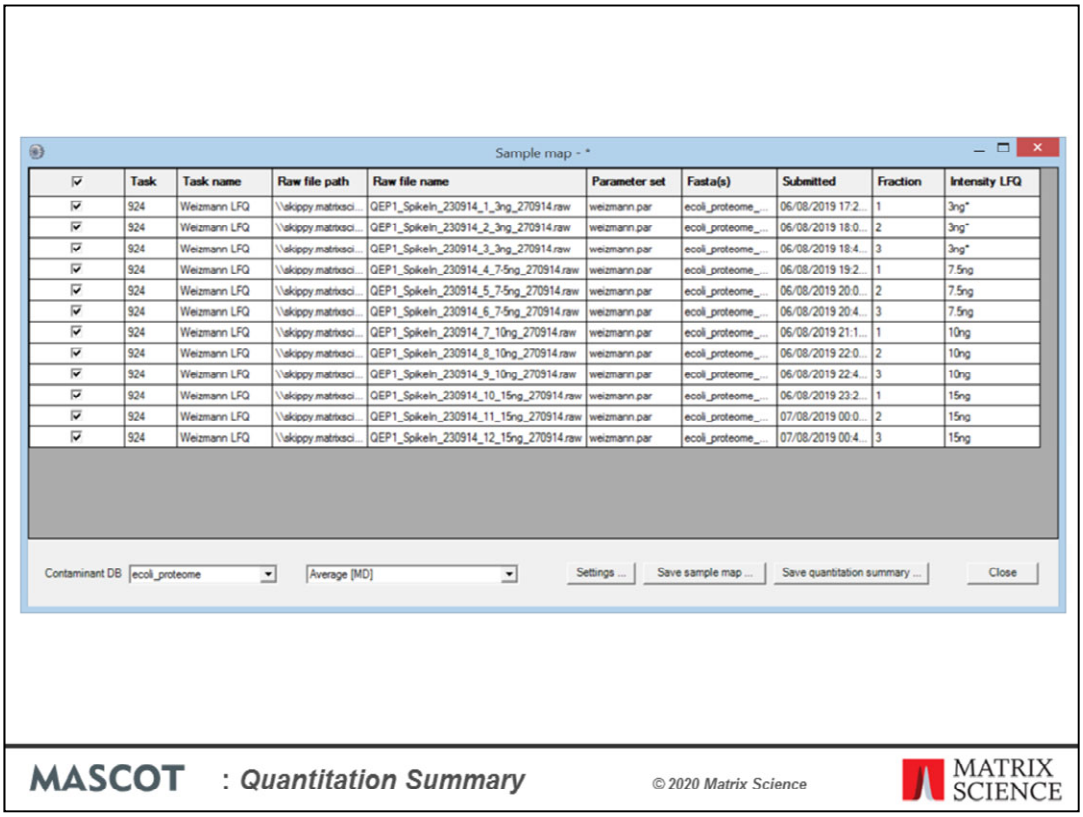


The files were processed in a single Mascot Daemon task, using Mascot Distiller for peak picking and quantitation. The data were searched against human and E. coli proteomes plus a contaminants database using typical search settings for Q Exactive data. The quantitation method was 'Average [MD]', which is label-free MS1 quantitation for individual files.

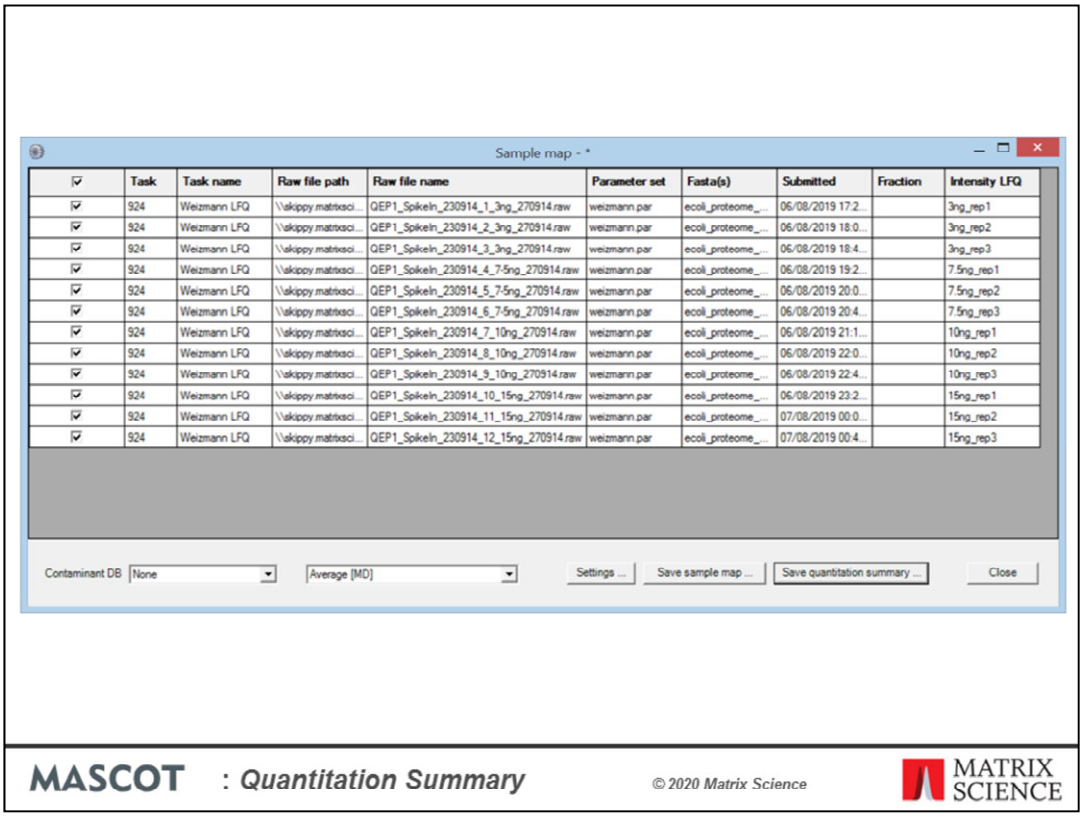
Once processing was complete, the task was selected in the list view on the Mascot Daemon status tab. Right clicking the selection invoked a context menu, from which Quantitation Summary; New sample map ... was chosen.



We've tried to minimise the amount of typing required to create a Sample Map. As in Excel, columns can be sorted so that a repeating value can be pasted to a range of cells. In this case, sorting on raw file name (by clicking the column header) is all that is required. For more complex data sets, sorting on file path or task name or time of submission may help organise the files in a useful way. This is particularly important when samples have been separated into large numbers of fractions. You don't want to have to type in every fraction number. Just sort appropriately, select the cell range in the fraction column, right click and choose Fill with integer series.

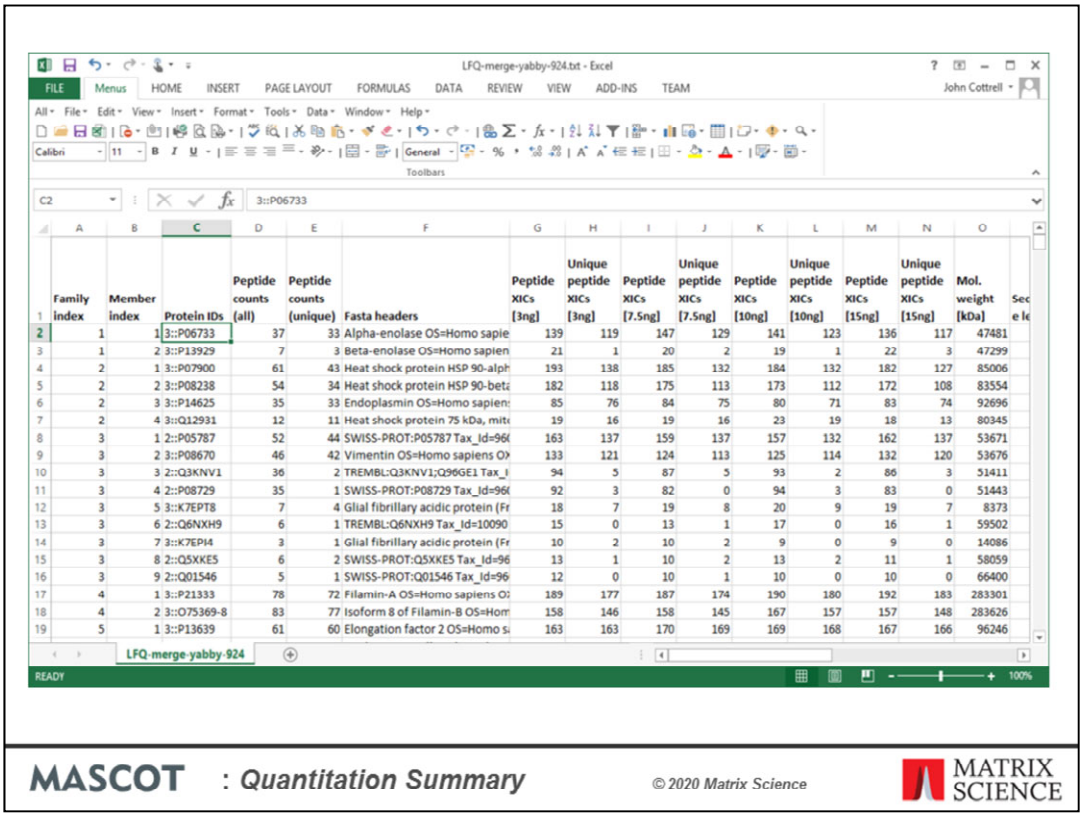


If a sample is not fractionated, the fraction cells can be left empty, unless you wish to merge replicates by treating them as fractions. In this example, there are two useful ways to fill in the two columns on the right that are used to identify the samples. Like this, to merge replicates as if they were fractions. An asterisk indicates the reference or control sample, and the Quantitation Summary will include ratios relative to this sample.



Or, like this, to create separate columns in the Quantitation Summary for each replicate; useful if you want statistics for variation across replicates. Sample identifiers can be anything you like as long as the combination of identifier and fraction number for each file is unique.

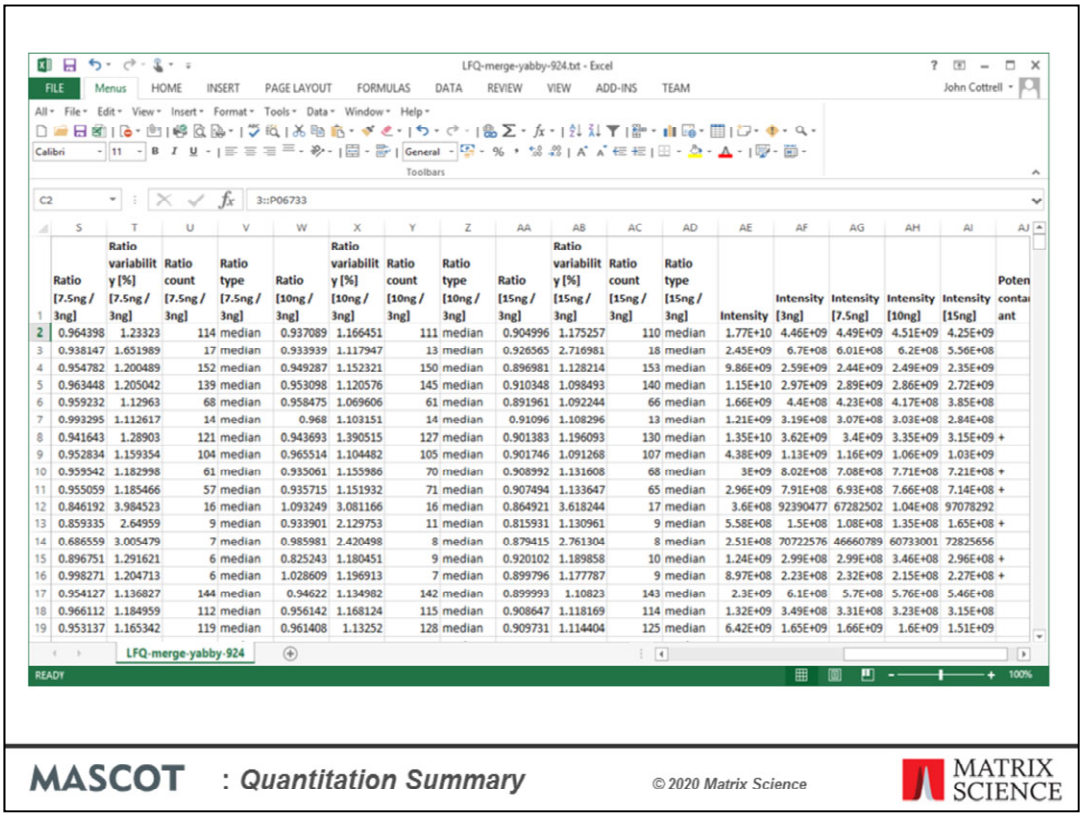
The Sample Map can be saved to a disk file, even if not complete, and reloaded as required. When Save quantitation summary ... is chosen, some validation is performed.



If all is present and correct, a progress dialog is displayed, because calculations can take some time for large numbers of files. The stages are

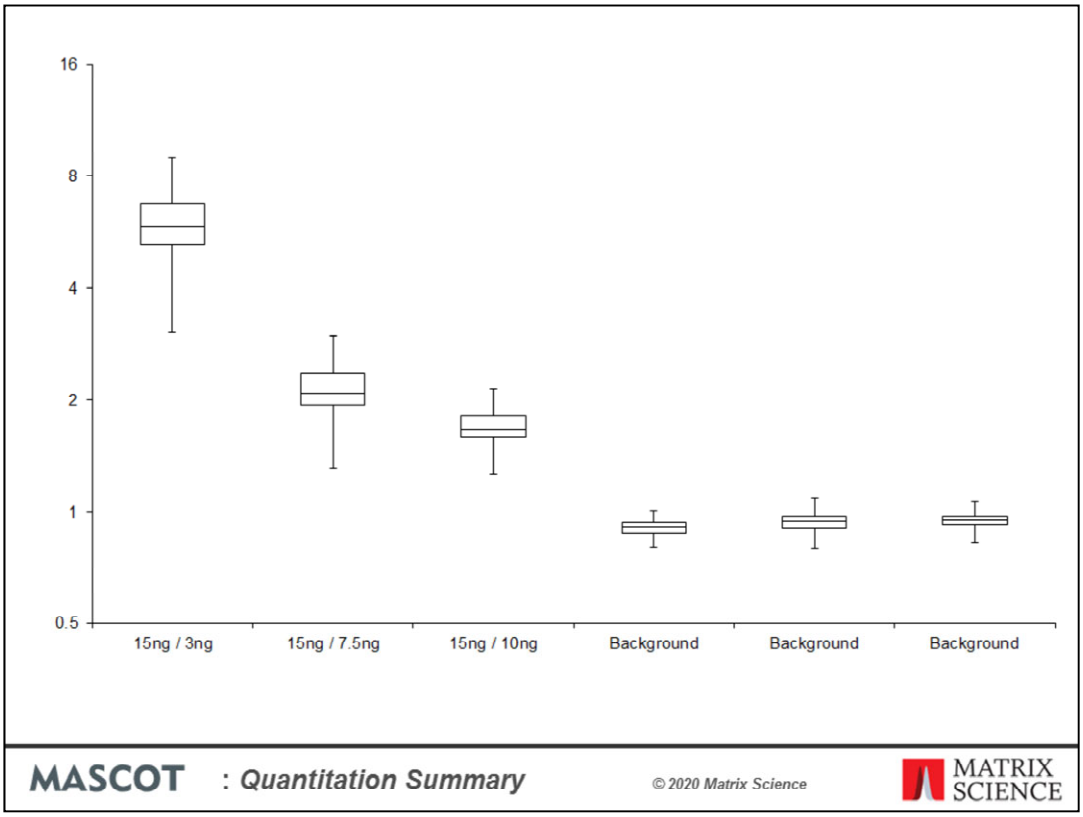
- Create a merged report of all search results
- Export the master list of proteins
- For each file, export the expression data for all peptides
- Assign the peptide data to proteins according to the master list
- Calculate protein abundances and ratios as required, including outlier detection
- Write everything to a disk file in TSV format

This is the Quantitation Summary for the label-free data when we choose to merge replicates. Most columns are self-explanatory

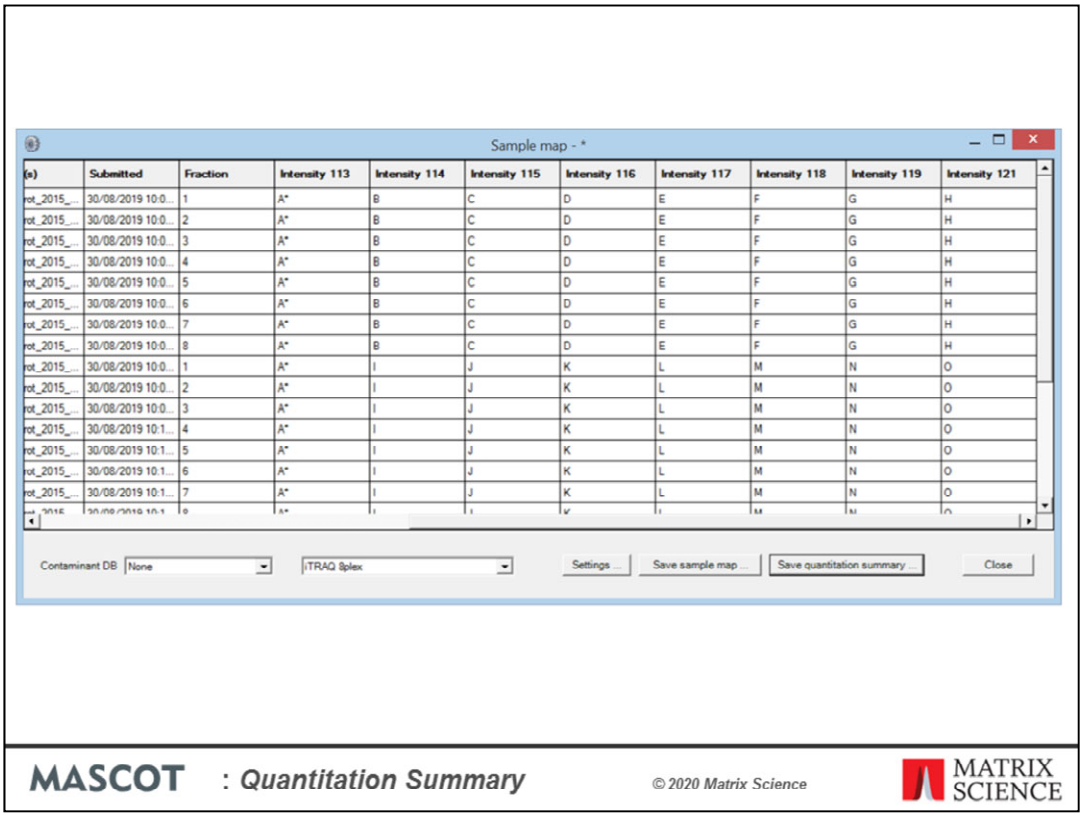


	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ
1	Ratio [7.5ng / 3ng]	Ratio variability [%] [7.5ng / 3ng]	Ratio count [7.5ng / 3ng]	Ratio type [7.5ng / 3ng]	Ratio [10ng / 3ng]	Ratio variability [%] [10ng / 3ng]	Ratio count [10ng / 3ng]	Ratio type [10ng / 3ng]	Ratio [15ng / 3ng]	Ratio variability [%] [15ng / 3ng]	Ratio count [15ng / 3ng]	Ratio type [15ng / 3ng]	Intensity	Intensity [3ng]	Intensity [7.5ng]	Intensity [10ng]	Intensity [15ng]	Potential contaminant
2	0.964398	1.23323	114	median	0.937089	1.166451	111	median	0.904996	1.175257	110	median	1.77E+10	4.46E+09	4.49E+09	4.51E+09	4.25E+09	
3	0.938147	1.651989	17	median	0.933939	1.117947	13	median	0.926565	2.716981	18	median	2.45E+09	6.7E+08	6.01E+08	6.2E+08	5.56E+08	
4	0.954782	1.200489	152	median	0.949287	1.152321	150	median	0.896981	1.128214	153	median	9.86E+09	2.59E+09	2.44E+09	2.49E+09	2.35E+09	
5	0.963448	1.205042	139	median	0.953098	1.120576	145	median	0.910348	1.098493	140	median	1.15E+10	2.97E+09	2.89E+09	2.86E+09	2.72E+09	
6	0.959232	1.12963	68	median	0.958475	1.069606	61	median	0.891961	1.092244	66	median	1.66E+09	4.4E+08	4.23E+08	4.17E+08	3.85E+08	
7	0.993295	1.112617	14	median	0.968	1.103151	14	median	0.91096	1.108296	13	median	1.21E+09	3.19E+08	3.07E+08	3.03E+08	2.84E+08	
8	0.941643	1.28903	121	median	0.943693	1.390515	127	median	0.901383	1.196093	130	median	1.35E+10	3.62E+09	3.4E+09	3.35E+09	3.15E+09	+
9	0.952834	1.159354	104	median	0.965514	1.104482	105	median	0.901746	1.091268	107	median	4.38E+09	1.13E+09	1.16E+09	1.06E+09	1.03E+09	
10	0.959542	1.182998	61	median	0.935061	1.155986	70	median	0.908992	1.131608	68	median	3E+09	8.02E+08	7.08E+08	7.71E+08	7.21E+08	+
11	0.955059	1.185466	57	median	0.935715	1.151932	71	median	0.907494	1.133647	65	median	2.96E+09	7.91E+08	6.93E+08	7.66E+08	7.14E+08	+
12	0.846192	3.984523	16	median	1.093249	3.081166	16	median	0.864921	3.618244	17	median	3.6E+08	92390477	67282502	1.04E+08	97078292	
13	0.859335	2.64959	9	median	0.933901	2.129753	11	median	0.815931	1.130961	9	median	5.58E+08	1.5E+08	1.08E+08	1.35E+08	1.65E+08	+
14	0.686559	3.005479	7	median	0.985981	2.420498	8	median	0.879415	2.761304	8	median	2.51E+08	70722576	46660789	60733001	72825656	
15	0.896751	1.291621	6	median	0.825243	1.180451	9	median	0.920102	1.189858	10	median	1.24E+09	2.99E+08	2.99E+08	3.46E+08	2.96E+08	+
16	0.998271	1.204713	6	median	1.028609	1.196913	7	median	0.899796	1.177787	9	median	8.97E+08	2.23E+08	2.32E+08	2.15E+08	2.27E+08	+
17	0.954127	1.136827	144	median	0.94622	1.134982	142	median	0.899993	1.10823	143	median	2.3E+09	6.1E+08	5.7E+08	5.76E+08	5.46E+08	
18	0.966112	1.184959	112	median	0.956142	1.168124	115	median	0.908647	1.118169	114	median	1.32E+09	3.49E+08	3.31E+08	3.23E+08	3.15E+08	
19	0.953137	1.165342	119	median	0.961408	1.13252	128	median	0.909731	1.114404	125	median	6.42E+09	1.65E+09	1.66E+09	1.6E+09	1.51E+09	

Over to the right are columns containing median ratios and total intensity values for each protein. One of the strengths of the Quantitation Summary is that it uses the same rigorous protein inference as the Mascot Protein Family Summary report.



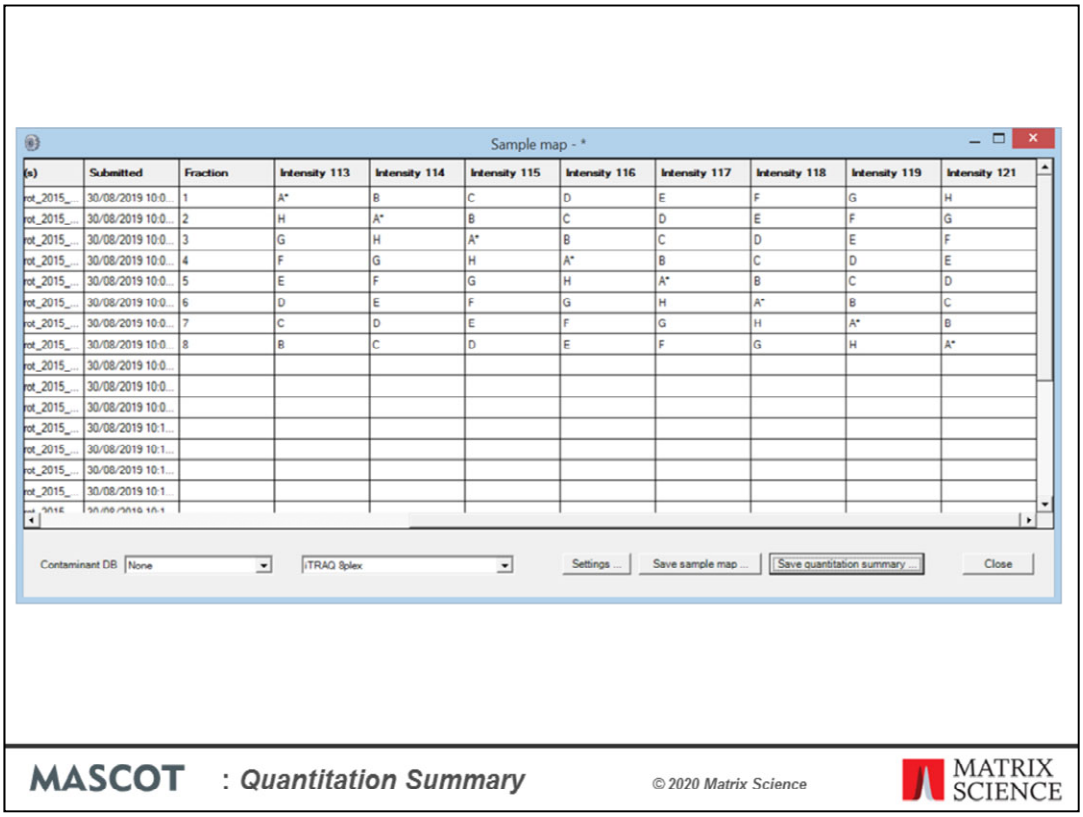
You may be able to get the report you want direct from Excel. For example, this box and whisker plot of the data was produced in Excel.



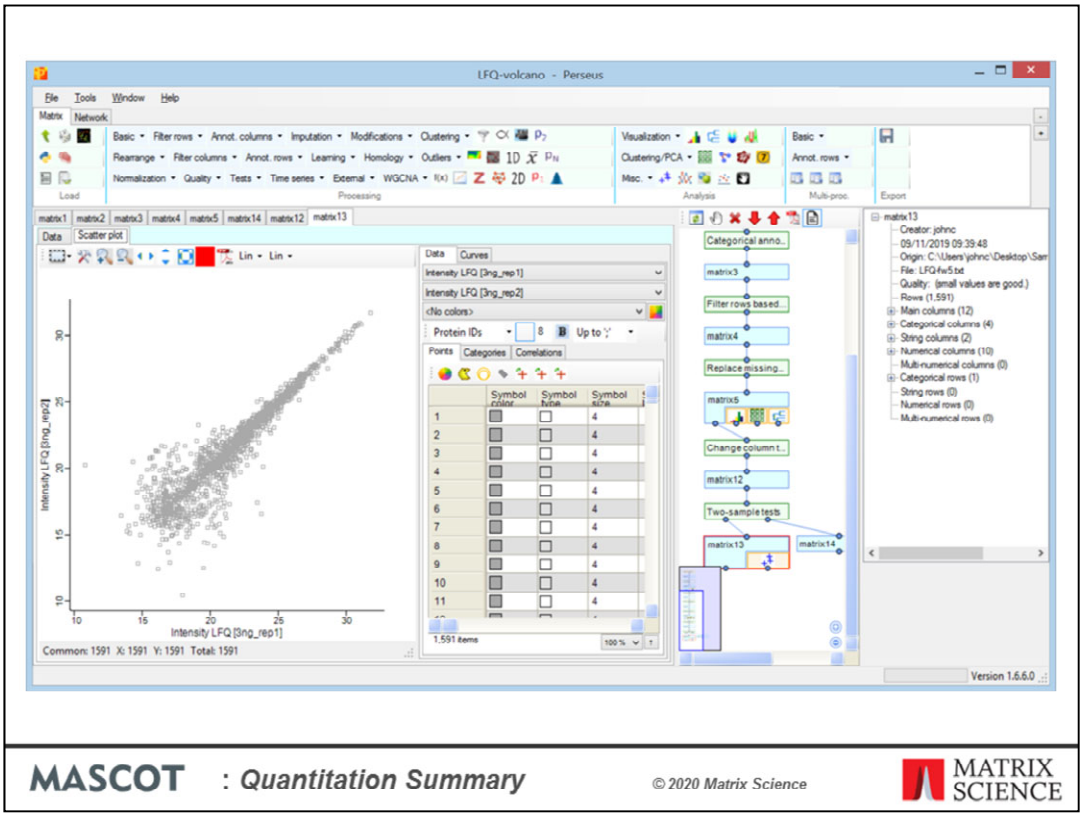
For a label-free experiment, there is a single column for the sample identifier. For experiments that use isotopic labels, there will be a column for each component specified in the quantitation method. If it was a typical SILAC experiment with two components, light for unlabelled and heavy for labelled, there would be two columns labelled *Intensity light* and *Intensity heavy*. An experiment that uses isobaric tags might have eight or more components.

This is a sample map for 8plex iTRAQ data. There are many ways of conducting such an study. This shows a case where there are 8 fractions for each sample, so the first 8 rows shows the same arrangement of samples, A to H. These fractions will be merged in the Quantitation summary, and A has an asterisk, so there will be columns for ratios to sample A as well as the total intensities for each channel. The second set of rows contains 7 new samples, plus reference sample A.

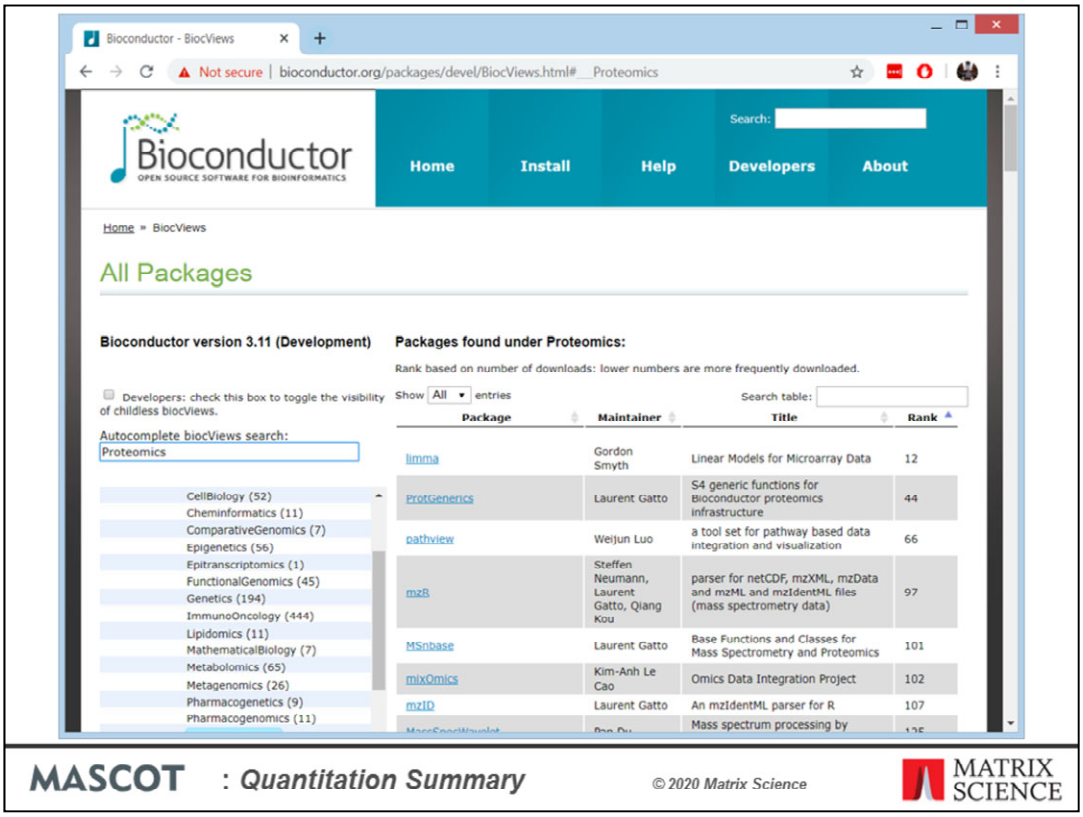
If the rows were replicates, and not fractions, then using the same channel for a sample across multiple replicates would be missing a trick.



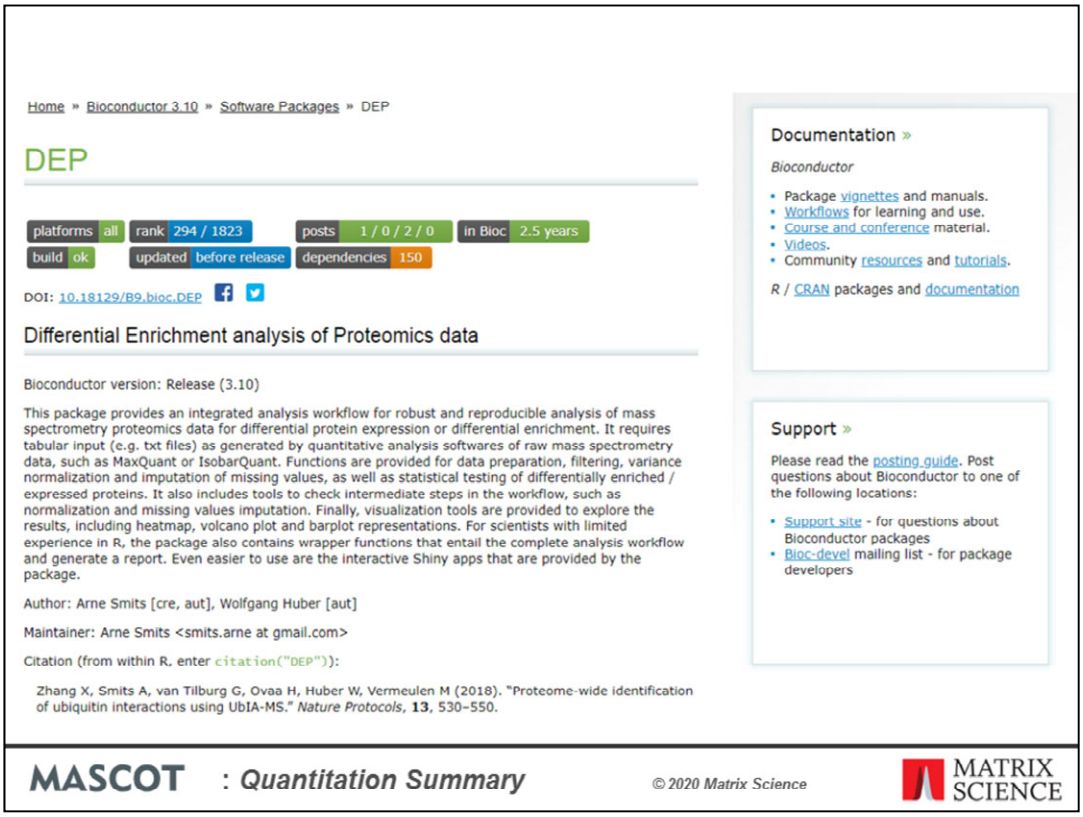
Better to rotate the labels, so as to reduce or eliminate systematic errors. Ideally, a so-called Latin Square, where each sample is rotated through all possible tags, as shown here for the first 8 rows. Rows are merged by sample identifier, so that the Quantitation Summary contains the correct ratio and intensity information.



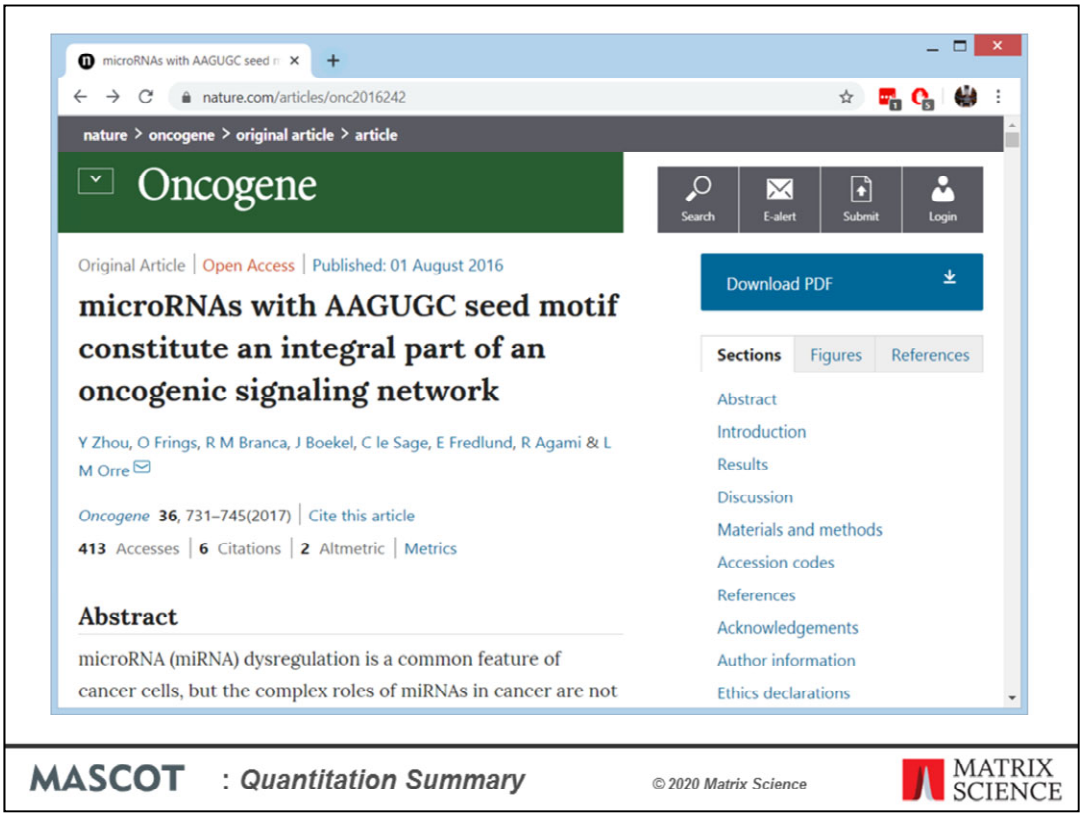
Having created a Quantitation Summary, what can you do with it? One option is to open it in Perseus, from the Max Planck Institute. This is a good choice if you prefer to manipulate the data using a spreadsheet type of approach



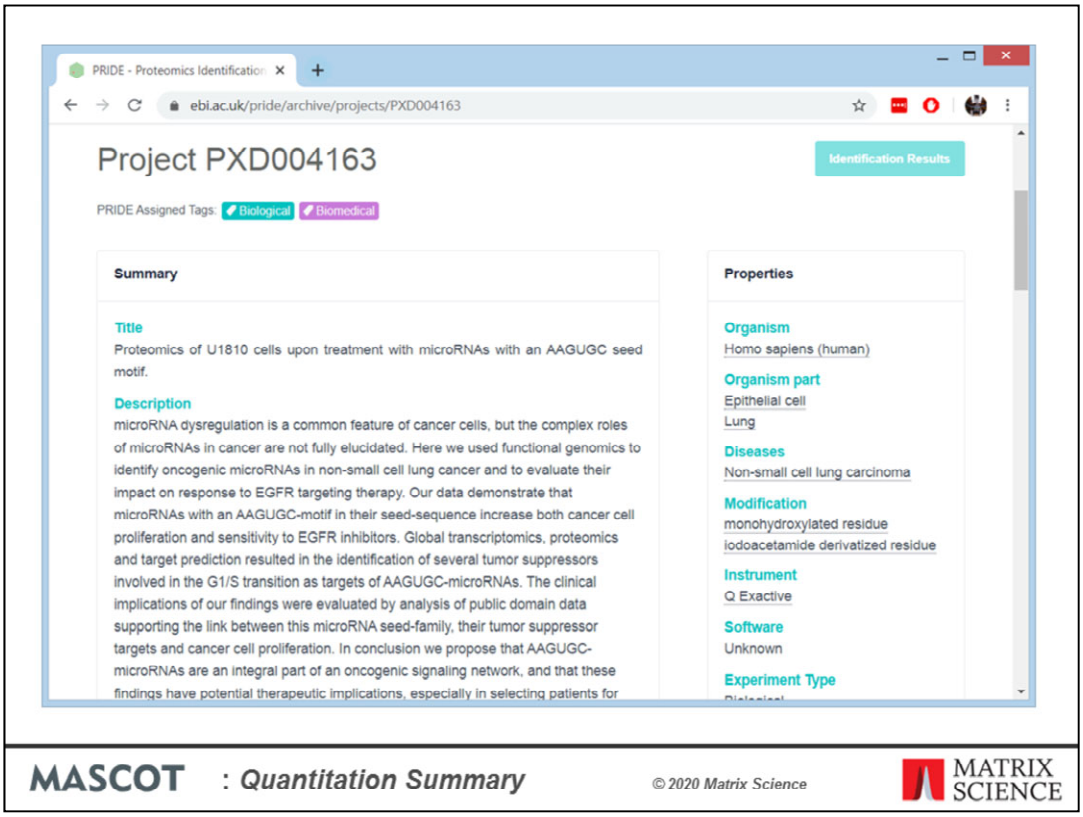
If you are willing to do a bit of scripting, the R language provides access to a huge range of statistical and graphical tools. Bioconductor is a collection of packages for genomic and proteomic applications. Currently, 135 packages are indexed under proteomics and 91 under mass spectrometry.



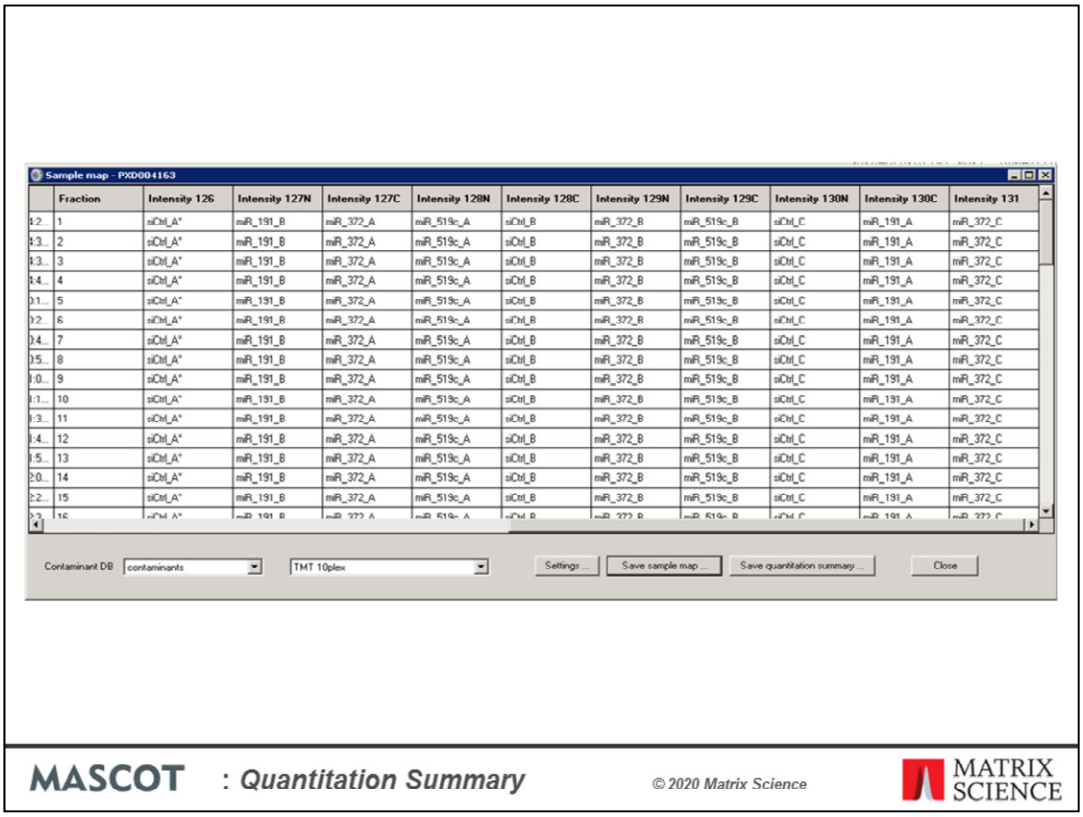
I'll use a package called DEP to illustrate the types of analysis that can be achieved with a few lines of scripting.



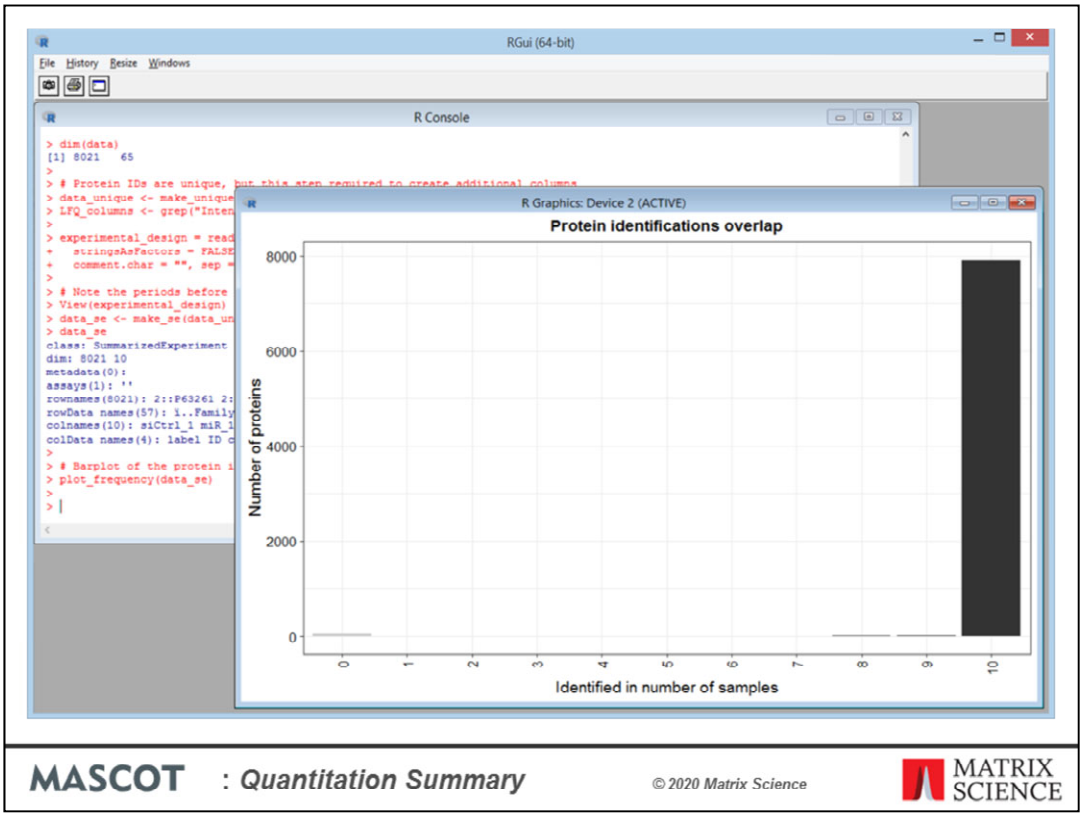
The data comes from this study to identify oncogenic microRNAs in non-small cell lung cancer. Quantitation used 10plex TMT



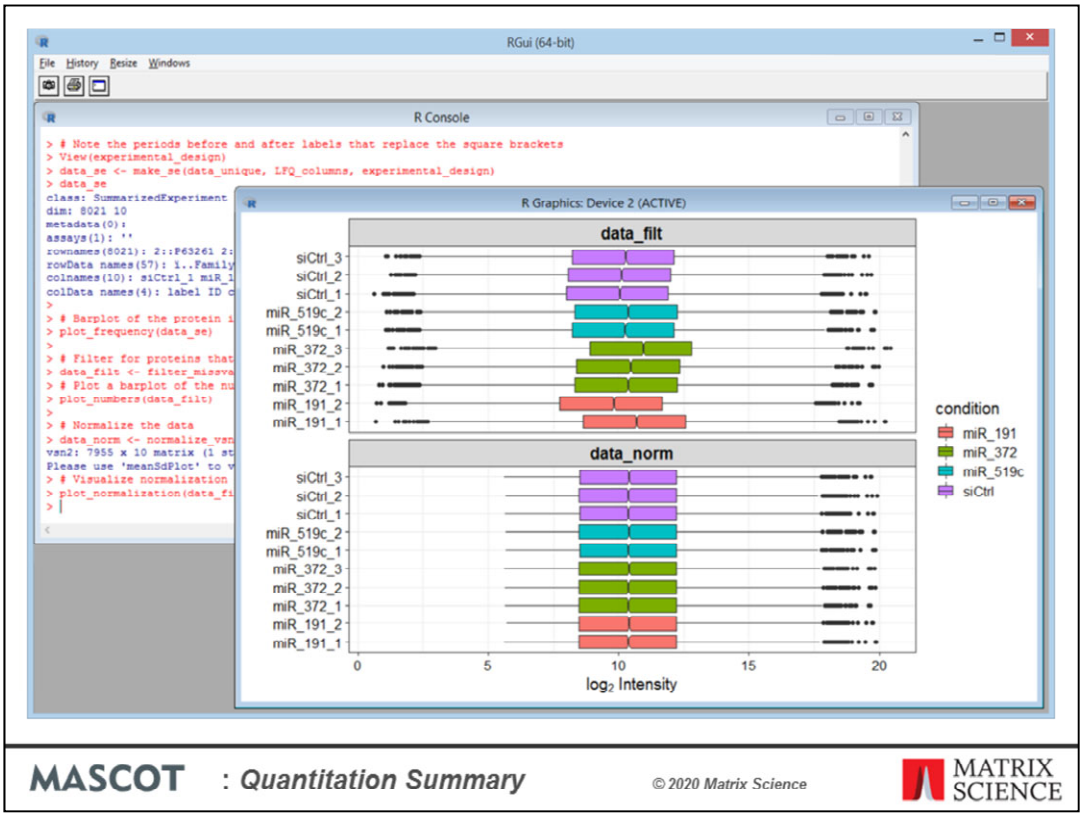
72 files downloaded from PRIDE project PXD004163 were processed and searched using Mascot Daemon, as described earlier.



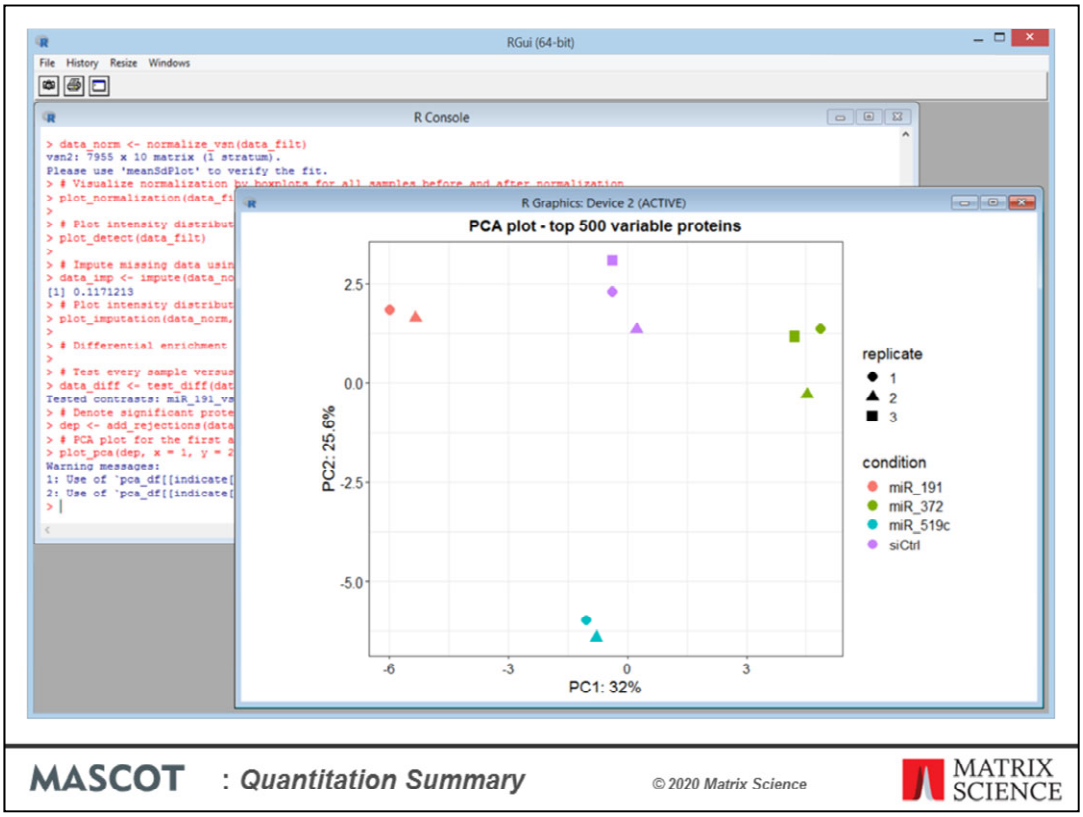
The Sample Map looks like this. 3 replicates of the control and one of the microRNA treatments, 2 replicates of the other two treatments. Peptide FDR was set to 1% by target/decoy.



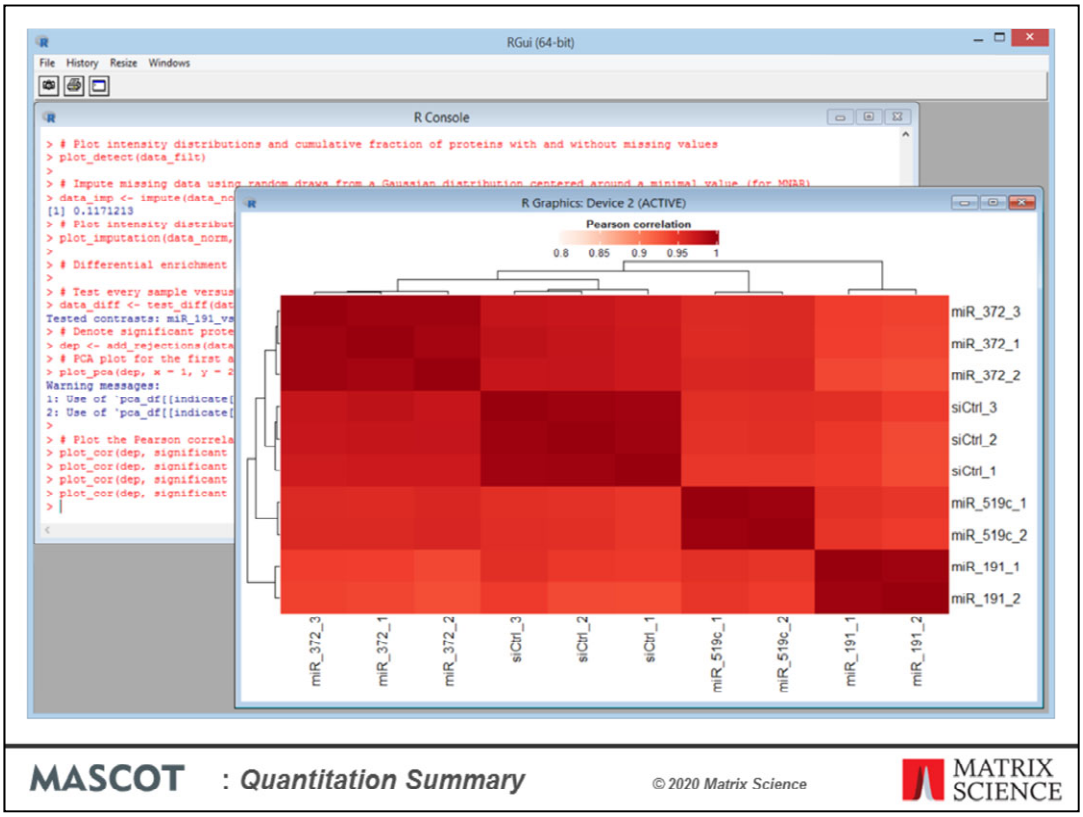
Using DEP, we can very easily create a number of informative charts. Some are for QC, such as this one, which shows we have data for almost all 8021 proteins across all 10 channels – very few missing values.



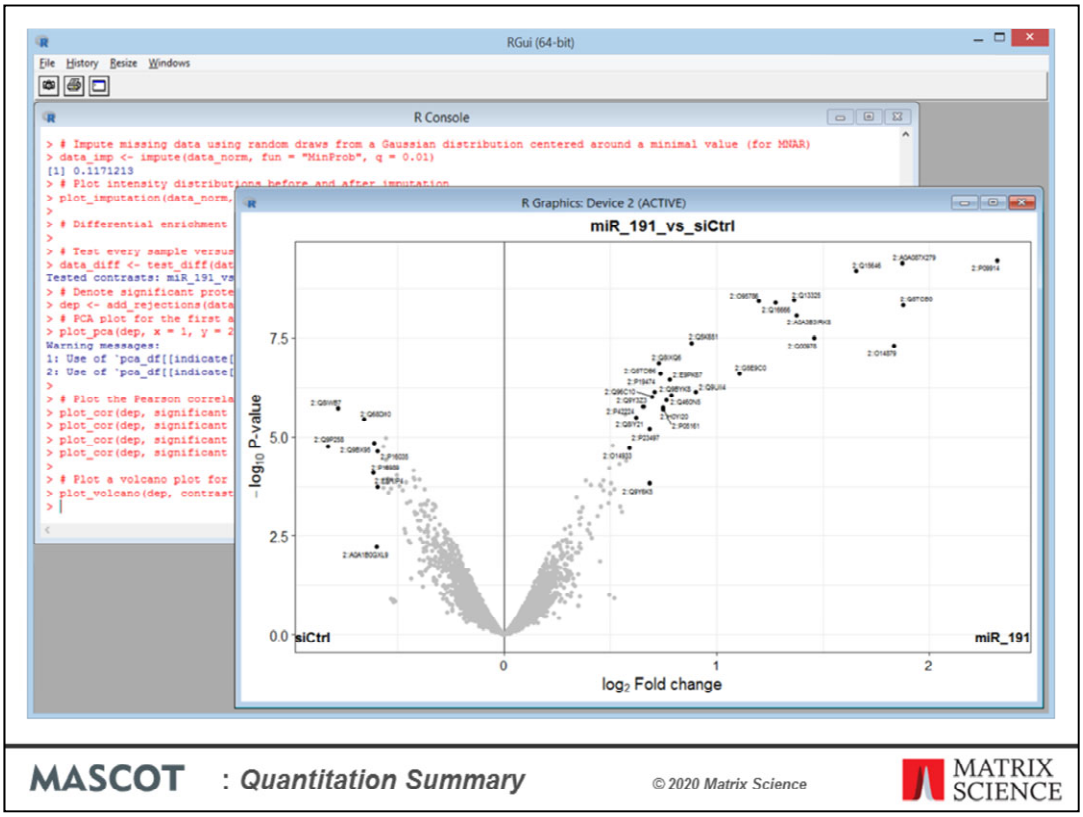
A box plot showing the intensities before and after normalisation



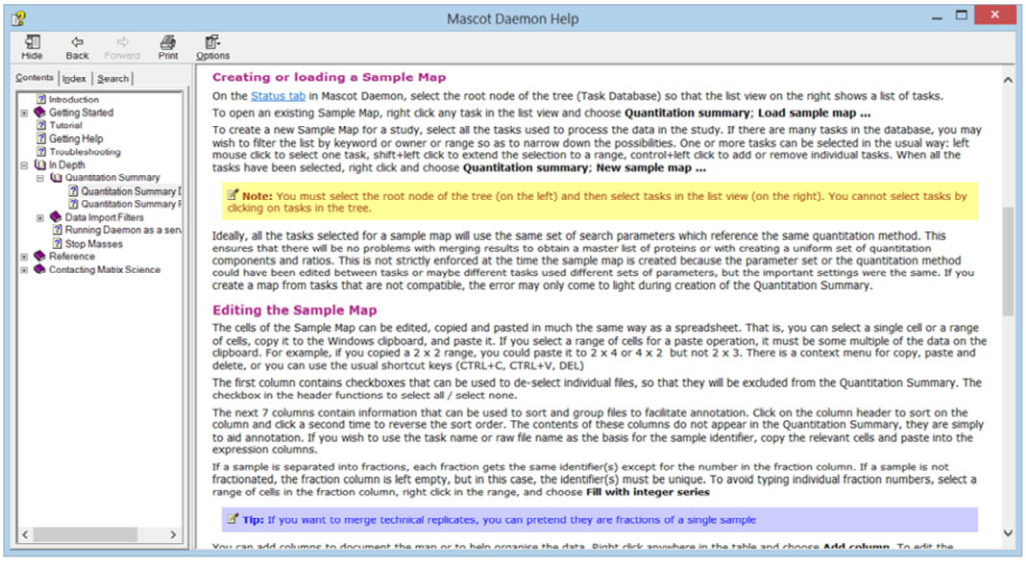
PCA shows the replicates cluster nicely



A heat map for sample to sample similarity



Finally, a volcano plot for fold changes between one treatment and the control. The outlier proteins are labelled with their identifiers.



As always, detailed help and reference material for the Sample Map and Quantitation Summary can be found in the Mascot Daemon help file