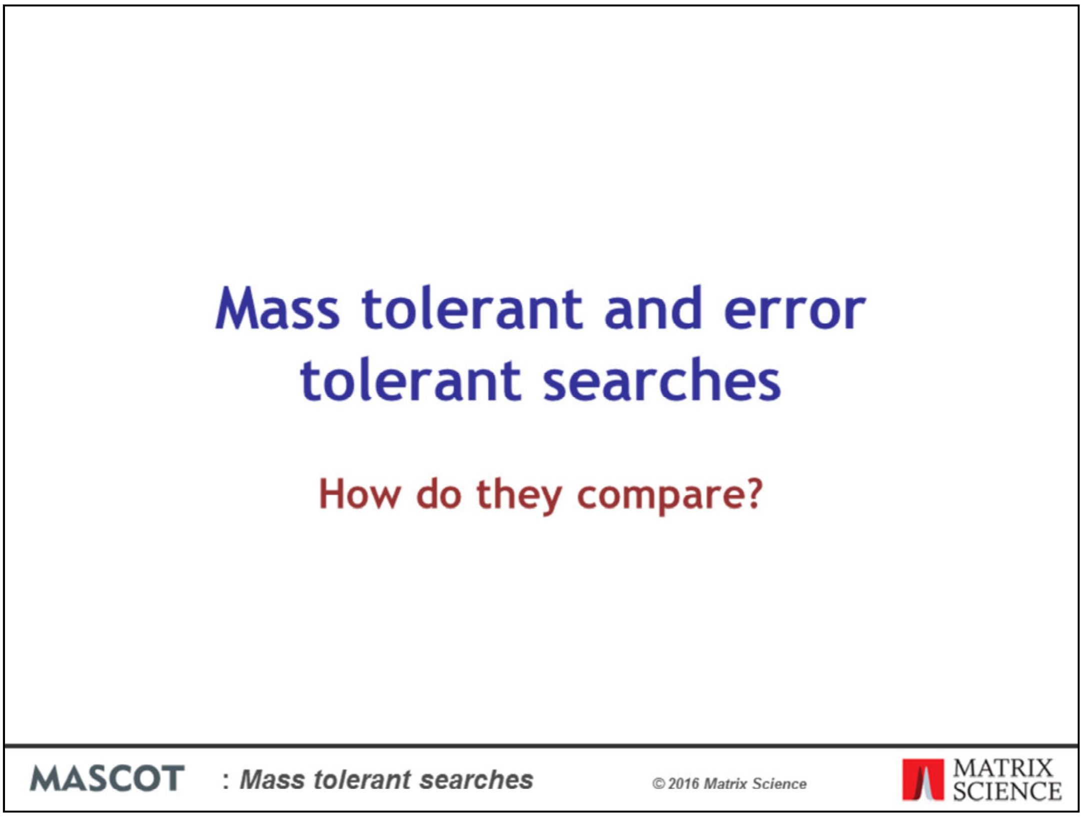
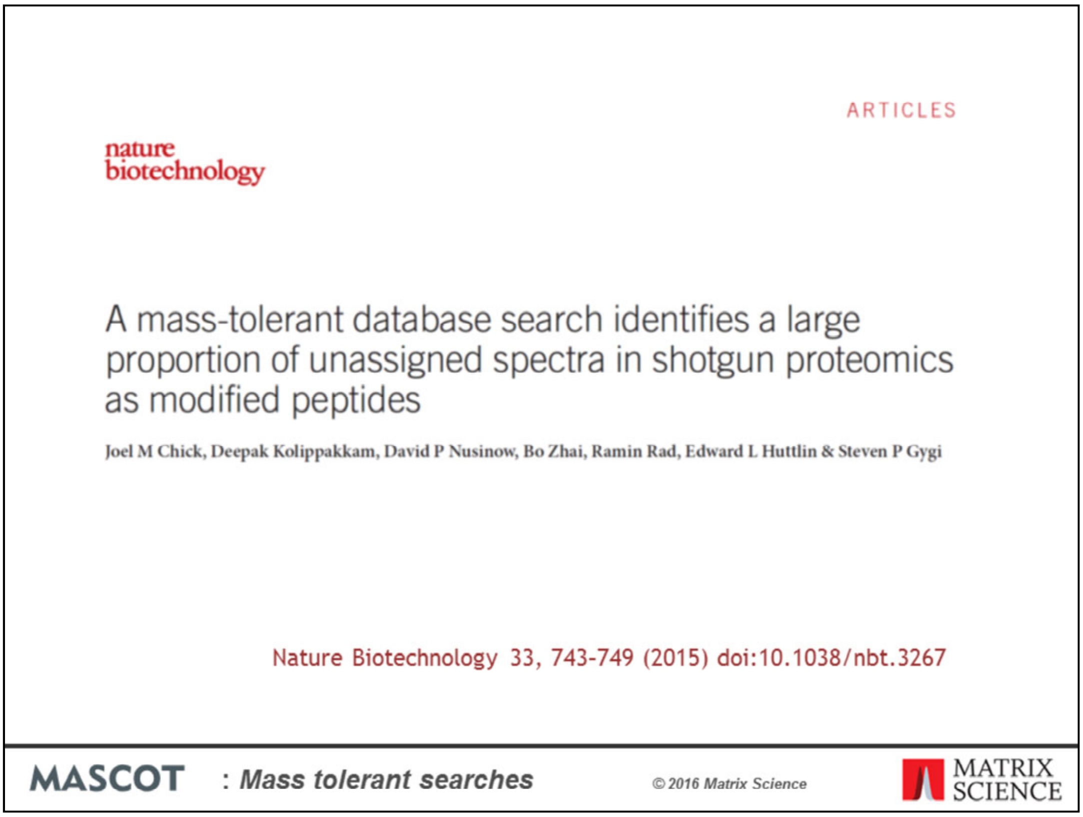


Mass tolerant and error tolerant searches

How do they compare?



ARTICLES

nature biotechnology

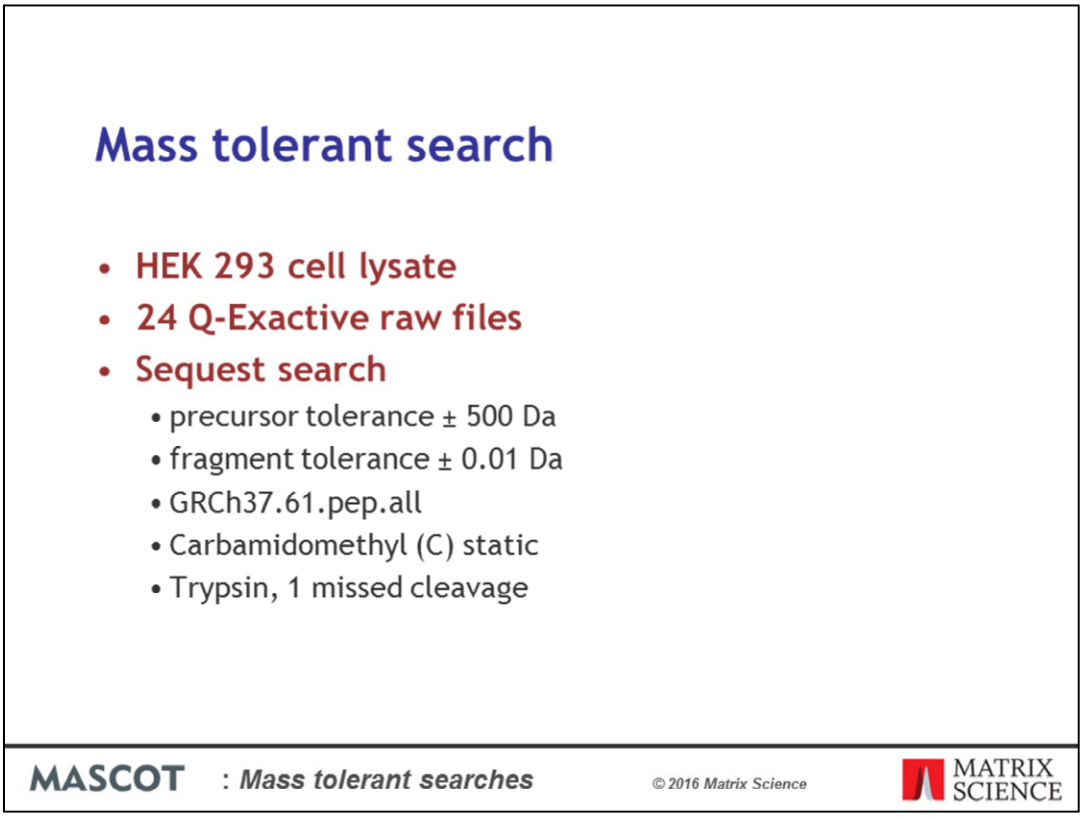
A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides

Joel M Chick, Deepak Kolippakkam, David P Nusinow, Bo Zhai, Ramin Rad, Edward L Huttlin & Steven P Gygi

Nature Biotechnology 33, 743-749 (2015) doi:10.1038/nbt.3267

MASCOT : *Mass tolerant searches* © 2016 Matrix Science MATRIX SCIENCE

Steven Gygi’s lab at Harvard Medical School published this paper in Nature Biotechnology last year. It describes the use of a very wide precursor mass tolerance, +/- 500 Da, to identify modified peptides in a Sequest search. How does this approach, which the authors also call an open search, compare with a "conventional" multi-pass search, such as the Mascot error tolerant search?

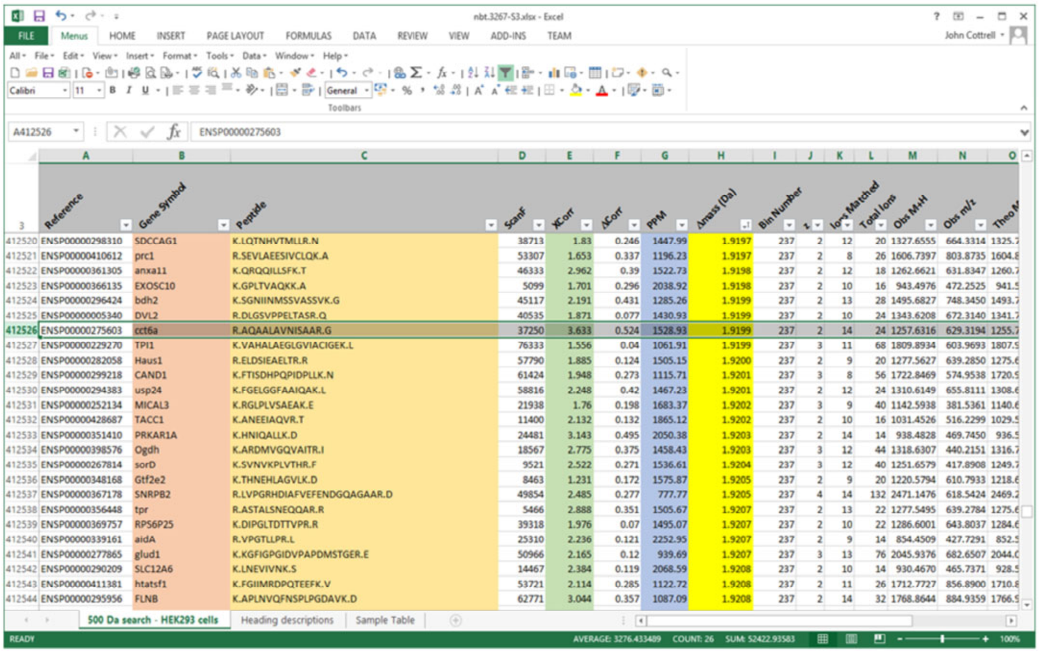


Mass tolerant search

- HEK 293 cell lysate
- 24 Q-Exactive raw files
- Sequest search
  - precursor tolerance ± 500 Da
  - fragment tolerance ± 0.01 Da
  - GRCh37.61.pep.all
  - Carbamidomethyl (C) static
  - Trypsin, 1 missed cleavage

MASCOT : *Mass tolerant searches* © 2016 Matrix Science MATRIX SCIENCE

The sample was a lysate of human embryonic kidney cells: 24 fractions analysed by Q-Exactive Orbitrap. Peak lists were searched against a human proteome database using Sequest. The only unusual aspect of the search was the 500 Da precursor tolerance.



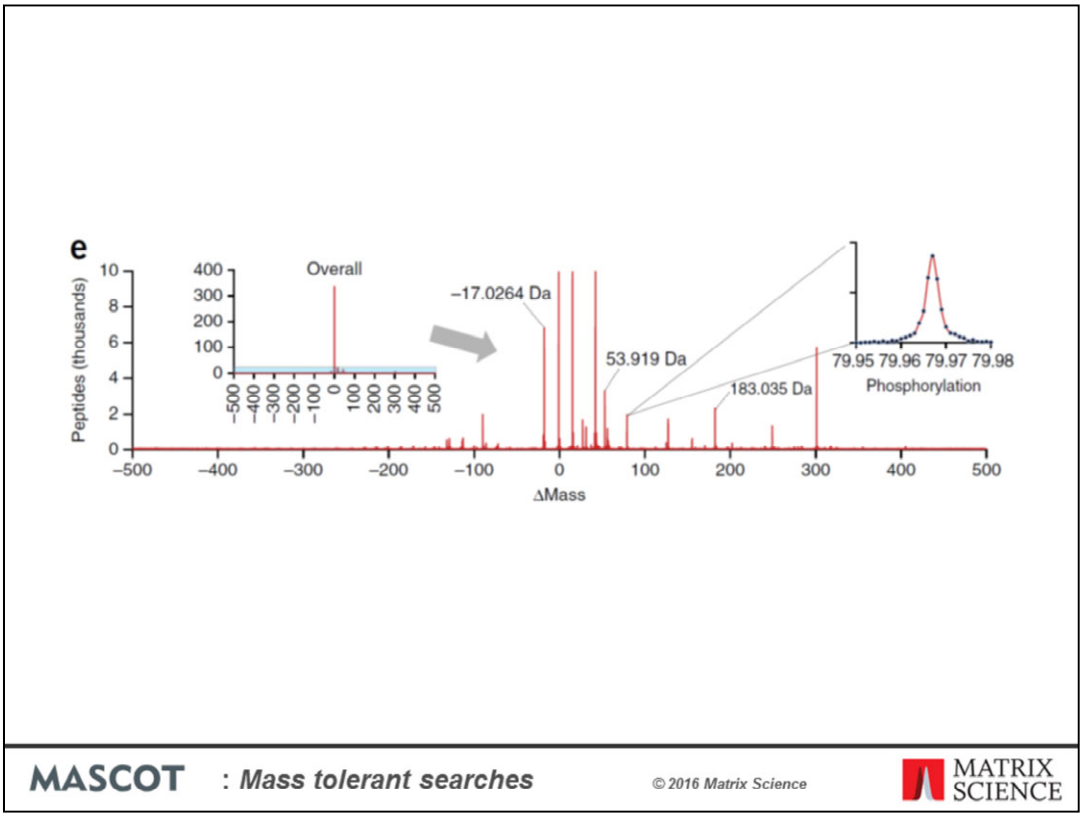
A412526 | ENSP00000275603

3	Reference	Gene Symbol	Peptide	ScanF	XCorr	ΔCorr	PPM	Δmass (Da)	Bin Number	z	Ions Matched	Total Ions	Obs M+H	Obs m/z	Theo M
412520	ENSP00000298310	SDCCAG1	K.LQTNHVTMLLR.N	38713	1.83	0.246	1447.99	1.9197	237	2	12	20	1327.6555	664.3314	1325.7
412521	ENSP00000410612	prc1	R.SEVLAEESIVCLQK.A	53307	1.653	0.337	1196.23	1.9197	237	2	8	26	1606.7397	803.8735	1604.8
412522	ENSP00000361305	anxa11	K.QRQQILLSFK.T	46333	2.962	0.39	1522.73	1.9198	237	2	12	18	1262.6621	631.8347	1260.7
412523	ENSP00000366135	EXOSC10	K.GPLTVAQKK.A	5099	1.701	0.296	2038.92	1.9198	237	2	10	16	943.4976	472.2525	941.5
412524	ENSP00000296424	bdh2	K.SGNIINMSSVASSVK.G	45117	2.191	0.431	1285.26	1.9199	237	2	13	28	1495.6827	748.3450	1493.7
412525	ENSP00000005340	DVL2	R.DLGSVPPELTASR.Q	40535	1.871	0.077	1430.93	1.9199	237	2	10	24	1343.6208	672.3140	1341.7
412526	ENSP00000275603	cct6a	R.AQAALAVNISAAR.G	37250	3.633	0.524	1528.93	1.9199	237	2	14	24	1257.6316	629.3194	1255.7
412527	ENSP00000229270	TPI1	K.VAHALAEGLGVIACIGEK.L	76333	1.556	0.04	1061.91	1.9199	237	3	11	68	1809.8934	603.9693	1807.9
412528	ENSP00000282058	Haus1	R.ELDSIEAELTR.R	57790	1.885	0.124	1505.15	1.9200	237	2	9	20	1277.5627	639.2850	1275.6
412529	ENSP00000299218	CAND1	K.FTISDHPQPIDPLLK.N	61424	1.948	0.273	1115.71	1.9201	237	3	8	56	1722.8469	574.9538	1720.9
412530	ENSP00000294383	usp24	K.FGELGGFAAIQAK.L	58816	2.248	0.42	1467.23	1.9201	237	2	12	24	1310.6149	655.8111	1308.6
412531	ENSP00000252134	MICAL3	K.RGLPLVSAEAK.E	21938	1.76	0.198	1683.37	1.9202	237	3	9	40	1142.5938	381.5361	1140.6
412532	ENSP00000428687	TACC1	K.ANEEIAQVR.T	11400	2.132	0.132	1865.12	1.9202	237	2	10	16	1031.4526	516.2299	1029.5
412533	ENSP00000351410	PRKAR1A	K.HNIQALLK.D	24481	3.143	0.495	2050.38	1.9203	237	2	14	14	938.4828	469.7450	936.5
412534	ENSP00000398576	Ogdh	K.ARDMVGQVAITR.I	18567	2.775	0.375	1458.43	1.9203	237	3	12	44	1318.6307	440.2151	1316.7
412535	ENSP00000267814	sorD	K.SVNVKPLVTHR.F	9521	2.522	0.271	1536.61	1.9204	237	3	12	40	1251.6579	417.8908	1249.7
412536	ENSP00000348168	Gtf2e2	K.THNEHLAGVLK.D	8463	1.231	0.172	1575.87	1.9205	237	2	9	20	1220.5794	610.7933	1218.6
412537	ENSP00000367178	SNRPB2	R.LVPGRHDIAFVEFENDGQAGAAR.D	49854	2.485	0.277	777.77	1.9205	237	4	14	132	2471.1476	618.5424	2469.2
412538	ENSP00000356448	tpr	R.ASTALSNEQQAR.R	5466	2.888	0.351	1505.67	1.9207	237	2	13	22	1277.5495	639.2784	1275.6
412539	ENSP00000369757	RPS6P25	K.DIPGLTDTTVPR.R	39318	1.976	0.07	1495.07	1.9207	237	2	10	22	1286.6001	643.8037	1284.6
412540	ENSP00000339161	aidA	R.VPGTLLPR.L	25310	2.236	0.121	2252.95	1.9207	237	2	9	14	854.4509	427.7291	852.5
412541	ENSP00000277865	glud1	K.KGFIGPGIDVPAPDMSTGER.E	50966	2.165	0.12	939.69	1.9207	237	3	13	76	2045.9376	682.6507	2044.0
412542	ENSP00000290209	SLC12A6	K.LNEVIVNK.S	14467	2.384	0.119	2068.59	1.9208	237	2	10	14	930.4670	465.7371	928.5
412543	ENSP00000411381	htatsf1	K.FGIIMRDPQTEEFK.V	53721	2.114	0.285	1122.72	1.9208	237	2	11	26	1712.7727	856.8900	1710.8
412544	ENSP00000295956	FLNB	K.APLNVQFNSPLPGDAVK.D	62771	3.044	0.357	1087.09	1.9208	237	2	14	32	1768.8644	884.9359	1766.9

500 Da search - HEK293 cells | Heading descriptions | Sample Table

AVERAGE: 3276.433489 COUNT: 26 SUM: 52422.93583 100%

What do the results of a mass tolerant search look like? Well, it's a long list of matches, just like a regular search, except some of them have substantial differences between the calculated and observed peptide mass.



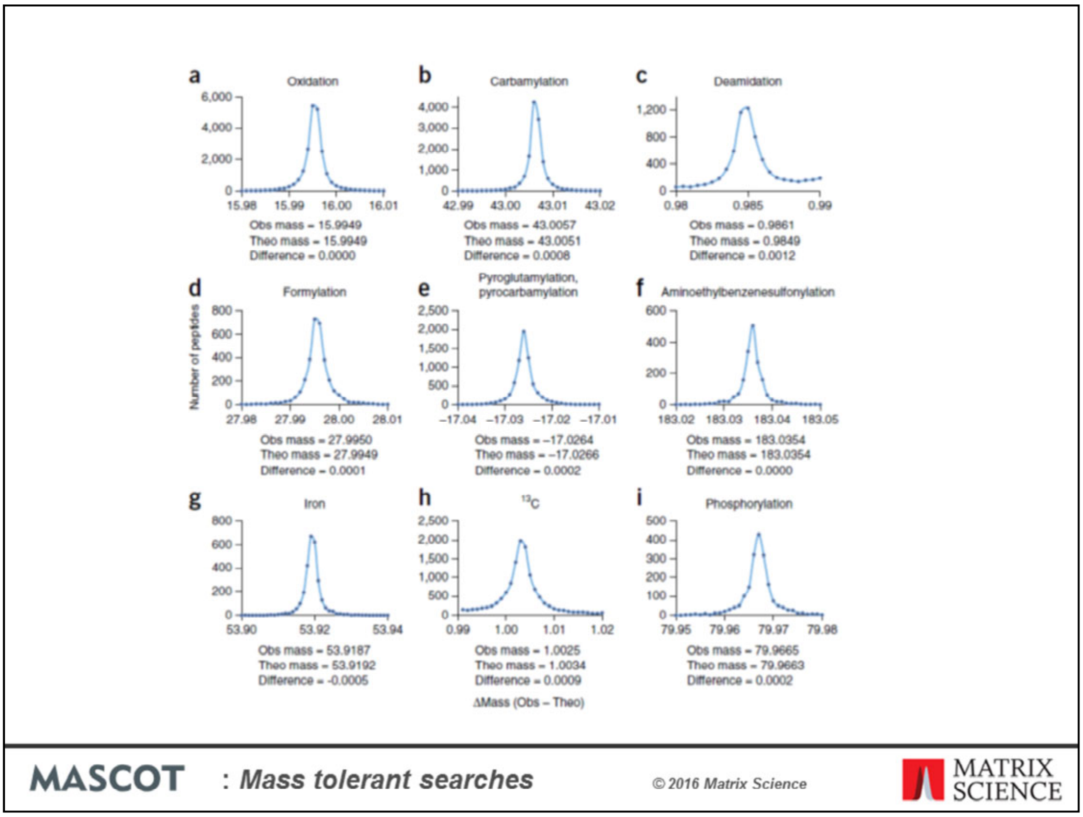
MASCOT : *Mass tolerant searches*

© 2016 Matrix Science

MATRIX SCIENCE

In the paper, this is summarized using a histogram of significant PSM count against mass difference. Most of the counts are for unmodified peptides, so the y axis has been expanded to show some of the more common delta masses, many of which correspond to the ‘usual suspects’ – ammonia loss, oxidation, acetylation, carbamylation, phosphorylation, etc.

The search doesn’t say anything about the site of the modification, which has to be determined after the search using a separate algorithm. In the paper, A-Score was used for this.

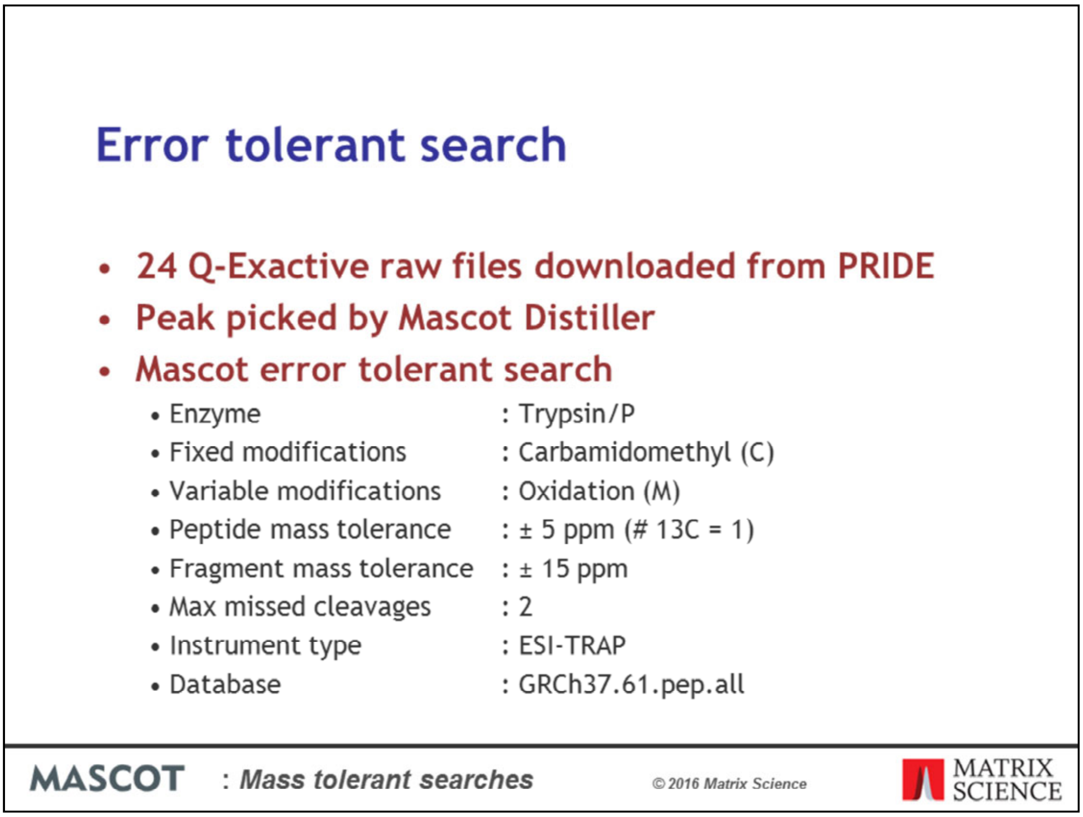


MASCOT : *Mass tolerant searches*

© 2016 Matrix Science

MATRIX SCIENCE

The authors used Gaussian fit analysis to divide the matches into 523 delta mass bins. This is Figure 2 from the publication, showing narrow, symmetric distributions in the delta mass distributions for selected modifications. Half widths are typically 0.005



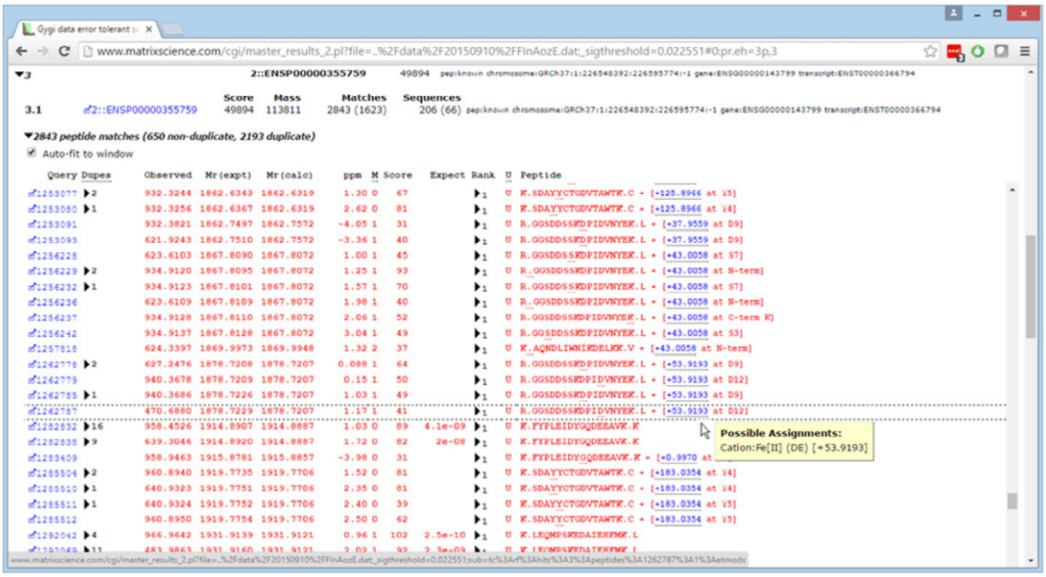
Error tolerant search

- 24 Q-Exactive raw files downloaded from PRIDE
- Peak picked by Mascot Distiller
- Mascot error tolerant search

• Enzyme	: Trypsin/P
• Fixed modifications	: Carbamidomethyl (C)
• Variable modifications	: Oxidation (M)
• Peptide mass tolerance	: ± 5 ppm (# 13C = 1)
• Fragment mass tolerance	: ± 15 ppm
• Max missed cleavages	: 2
• Instrument type	: ESI-TRAP
• Database	: GRCh37.61.pep.all

MASCOT : *Mass tolerant searches* © 2016 Matrix Science MATRIX SCIENCE

To make a comparison with the Mascot error tolerant search, we downloaded the raw files from PRIDE and processed them into peak lists using Mascot Distiller. The same database was searched using very standard settings. Target/decoy was used to set the false discovery rate for PSMs for the first pass search to 1%



www.matrixscience.com/cgi/master_results_2.pl?file=..%2Fdata%2F20150910%2FFInAozE.dat;_sigthreshold=0.022551#0:pr.eh=3p,3

▼3 2::ENSP00000355759 49894 pep:known chromosome:GRCh37:1:226548392:226595774:-1 gene:ENSG00000143799 transcript:ENST00000366794

		Score	Mass	Matches	Sequences	
3.1	2::ENSP00000355759	49894	113811	2843 (1623)	206 (66)	pep:known chromosome:GRCh37:1:226548392:226595774:-1 gene:ENSG00000143799 transcript:ENST00000366794

▼2843 peptide matches (650 non-duplicate, 2193 duplicate)

☑ Auto-fit to window

Query	Dupes	Observed	Mr(expt)	Mr(calc)	ppm	M	Score	Expect	Rank	U	Peptide
1253077	▶2	932.3244	1862.6343	1862.6319	1.30	0	67		▶1	U	K.SDAYYCTGDVTAWTK.C + [+125.8966 at Y5]
1253080	▶1	932.3256	1862.6367	1862.6319	2.62	0	81		▶1	U	K.SDAYYCTGDVTAWTK.C + [+125.8966 at Y4]
1253091		932.3821	1862.7497	1862.7572	-4.05	1	31		▶1	U	R.GGSDDSSKDPIDVNYEK.L + [+37.9559 at D9]
1253093		621.9243	1862.7510	1862.7572	-3.36	1	40		▶1	U	R.GGSDDSSKDPIDVNYEK.L + [+37.9559 at D9]
1256228		623.6103	1867.8090	1867.8072	1.00	1	45		▶1	U	R.GGSDDSSKDPIDVNYEK.L + [+43.0058 at S7]
1256229	▶2	934.9120	1867.8095	1867.8072	1.25	1	93		▶1	U	R.GGSDDSSKDPIDVNYEK.L + [+43.0058 at N-term]
1256232	▶1	934.9123	1867.8101	1867.8072	1.57	1	70		▶1	U	R.GGSDDSSKDPIDVNYEK.L + [+43.0058 at S7]
1256236		623.6109	1867.8109	1867.8072	1.98	1	40		▶1	U	R.GGSDDSSKDPIDVNYEK.L + [+43.0058 at N-term]
1256237		934.9128	1867.8110	1867.8072	2.06	1	52		▶1	U	R.GGSDDSSKDPIDVNYEK.L + [+43.0058 at C-term K]
1256242		934.9137	1867.8128	1867.8072	3.04	1	49		▶1	U	R.GGSDDSSKDPIDVNYEK.L + [+43.0058 at S3]
1257818		624.3397	1869.9973	1869.9948	1.32	2	37		▶1	U	K.AQNDLIWNIKDELKK.V + [+43.0058 at N-term]
1262778	▶2	627.2476	1878.7208	1878.7207	0.088	1	64		▶1	U	R.GGSDDSSKDPIDVNYEK.L + [+53.9193 at D9]
1262779		940.3678	1878.7209	1878.7207	0.15	1	50		▶1	U	R.GGSDDSSKDPIDVNYEK.L + [+53.9193 at D12]
1262785	▶1	940.3686	1878.7226	1878.7207	1.03	1	49		▶1	U	R.GGSDDSSKDPIDVNYEK.L + [+53.9193 at D9]
1262787		470.6880	1878.7229	1878.7207	1.17	1	41		▶1	U	R.GGSDDSSKDPIDVNYEK.L + [+53.9193 at D12]
1282832	▶16	958.4526	1914.8907	1914.8887	1.03	0	89	4.1e-09	▶1	U	K.FYPLEIDYGQDEEAVK.K
1282838	▶9	639.3046	1914.8920	1914.8887	1.72	0	82	2e-08	▶1	U	K.FYPLEIDYGQDEEAVK.K
1283409		958.9463	1915.8781	1915.8857	-3.98	0	31		▶1	U	K.FYPLEIDYGQDEEAVK.K + [+0.9970 at
1285504	▶2	960.8940	1919.7735	1919.7706	1.52	0	81		▶1	U	K.SDAYYCTGDVTAWTK.C + [+183.0354 at Y4]
1285510	▶1	640.9323	1919.7751	1919.7706	2.35	0	81		▶1	U	K.SDAYYCTGDVTAWTK.C + [+183.0354 at Y4]
1285511	▶1	640.9324	1919.7752	1919.7706	2.40	0	39		▶1	U	K.SDAYYCTGDVTAWTK.C + [+183.0354 at Y5]
1285512		960.8950	1919.7754	1919.7706	2.50	0	62		▶1	U	K.SDAYYCTGDVTAWTK.C + [+183.0354 at Y5]
1292042	▶4	966.9642	1931.9139	1931.9121	0.96	1	102	2.5e-10	▶1	U	K.LEQMPSKEDAIEHFMK.L
1292049	▶11	483.9863	1931.9160	1931.9121	2.02	1	92	2.3e-09	▶1	U	K.LEQMPSKEDAIEHFMK.L

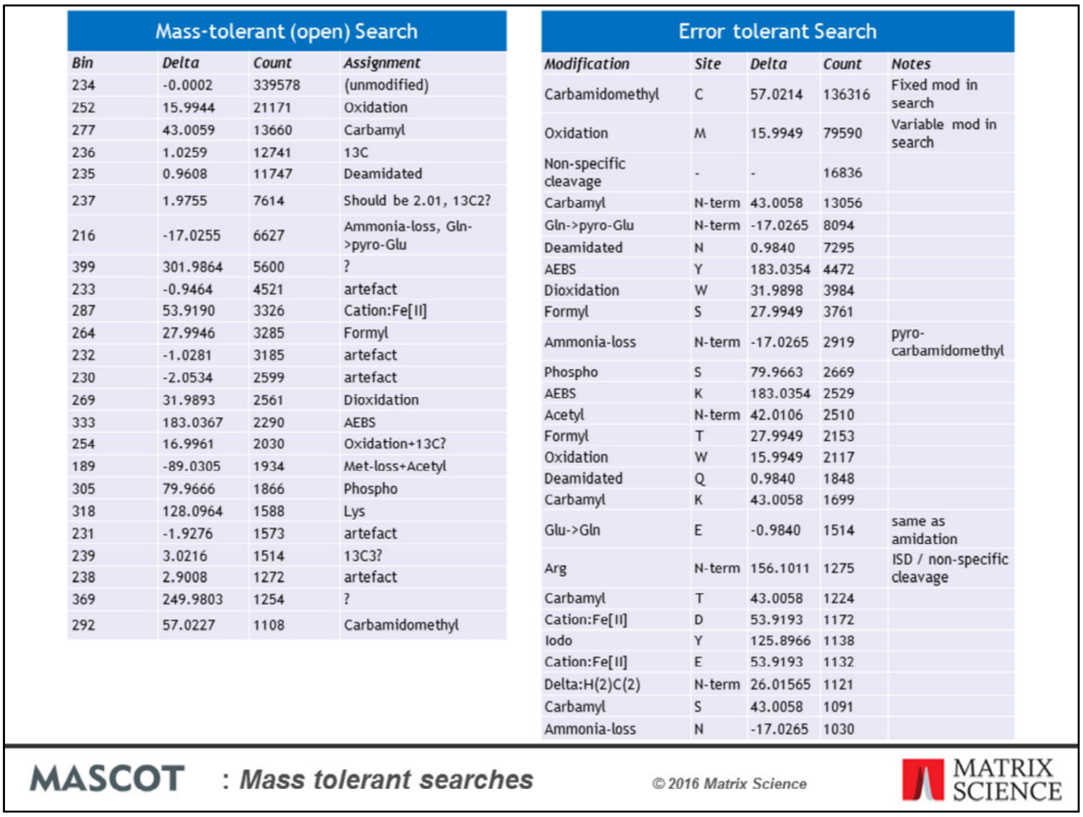
Possible Assignments:
Cation:Fe[II] (DE) [+53.9193]

www.matrixscience.com/cgi/master_results_2.pl?file=..%2Fdata%2F20150910%2FFInAozE.dat;_sigthreshold=0.022551;sub=tc%3Arf%3Ahits%3A3%3Apeptides%3A1262787%3A1%3Aetmods

In the automatic error tolerant search, every protein containing one or more significant matches from the first pass search is selected for a second pass search, which uses a much wider search space: all the modifications in the Unimod database, non-specific cleavage at one peptide terminus, and all possible single amino acid substitutions. For each peptide, these possibilities are tested serially. That is, we don't look for two unsuspected modifications on the same peptide, or an unsuspected modification plus a SNP, etc.

In the result report, these additional matches are displayed with a mass delta and a tooltip showing the modifications or SNPs that fit to the delta within the specified mass tolerance. For accurate data like this, where the precursor tolerance is 5ppm, there is usually just one possibility.

We can't claim that the way the matches are reported is infallible. Sometimes, the exact site of the modification will be uncertain. Other times, the error tolerant match has a score that is only slightly higher than an unmodified peptide, and we might prefer to take the simpler explanation. But, in general, for high accuracy data, the displayed modifications represent a reasonable interpretation.



Mass-tolerant (open) Search			
Bin	*Delta*	*Count*	*Assignment*
234	-0.0002	339578	(unmodified)
252	15.9944	21171	Oxidation
277	43.0059	13660	Carbamyl
236	1.0259	12741	13C
235	0.9608	11747	Deamidated
237	1.9755	7614	Should be 2.01, 13C2?
216	-17.0255	6627	Ammonia-loss, Gln->pyro-Glu
399	301.9864	5600	?
233	-0.9464	4521	artefact
287	53.9190	3326	Cation:Fe[II]
264	27.9946	3285	Formyl
232	-1.0281	3185	artefact
230	-2.0534	2599	artefact
269	31.9893	2561	Dioxidation
333	183.0367	2290	AEBS
254	16.9961	2030	Oxidation+13C?
189	-89.0305	1934	Met-loss+Acetyl
305	79.9666	1866	Phospho
318	128.0964	1588	Lys
231	-1.9276	1573	artefact
239	3.0216	1514	13C3?
238	2.9008	1272	artefact
369	249.9803	1254	?
292	57.0227	1108	Carbamidomethyl

Error tolerant Search				
Modification	*Site*	*Delta*	*Count*	*Notes*
Carbamidomethyl	C	57.0214	136316	Fixed mod in search
Oxidation	M	15.9949	79590	Variable mod in search
Non-specific cleavage	-	-	16836	
Carbamyl	N-term	43.0058	13056	
Gln->pyro-Glu	N-term	-17.0265	8094	
Deamidated	N	0.9840	7295	
AEBS	Y	183.0354	4472	
Dioxidation	W	31.9898	3984	
Formyl	S	27.9949	3761	
Ammonia-loss	N-term	-17.0265	2919	pyro-carbamidomethyl
Phospho	S	79.9663	2669	
AEBS	K	183.0354	2529	
Acetyl	N-term	42.0106	2510	
Formyl	T	27.9949	2153	
Oxidation	W	15.9949	2117	
Deamidated	Q	0.9840	1848	
Carbamyl	K	43.0058	1699	
Glu->Gln	E	-0.9840	1514	same as amidation
Arg	N-term	156.1011	1275	ISD / non-specific cleavage
Carbamyl	T	43.0058	1224	
Cation:Fe[II]	D	53.9193	1172	
Iodo	Y	125.8966	1138	
Cation:Fe[II]	E	53.9193	1132	
Delta:H(2)C(2)	N-term	26.01565	1121	
Carbamyl	S	43.0058	1091	
Ammonia-loss	N	-17.0265	1030	

MASCOT : *Mass tolerant searches*

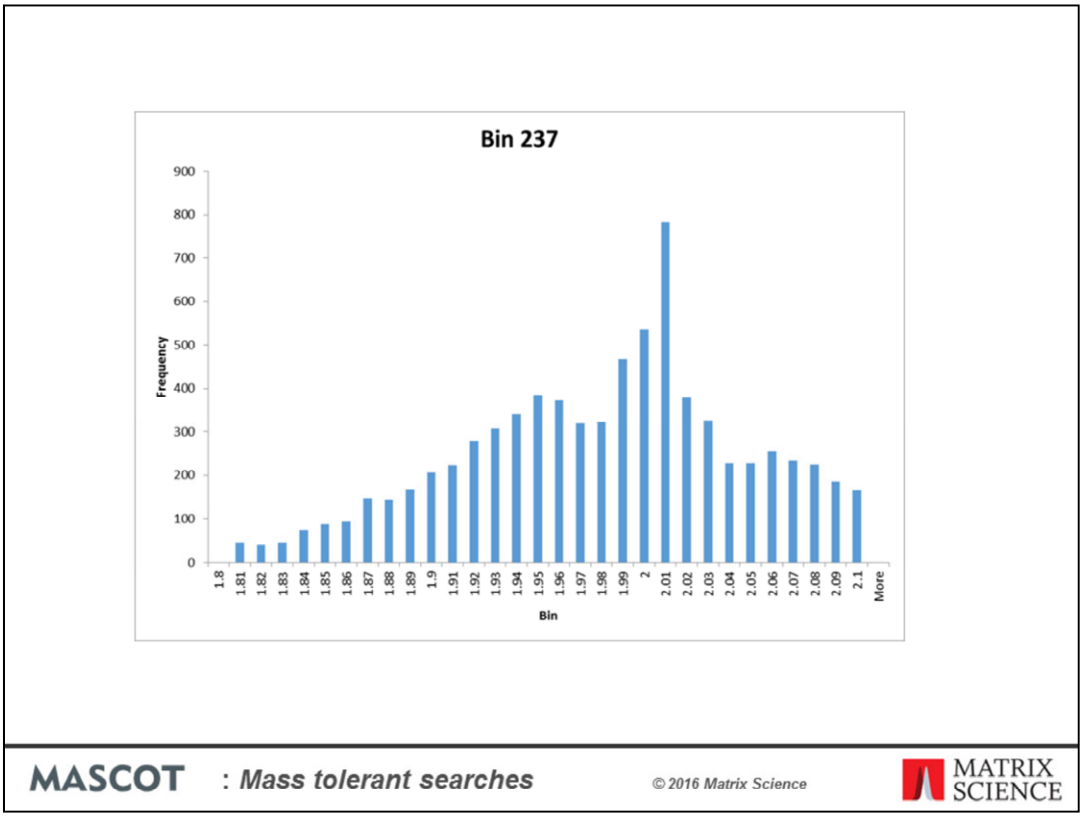
© 2016 Matrix Science

MATRIX SCIENCE

These tables list the most abundant matches from the two types of search with an arbitrary cut-off of 1000 instances. There are some differences in the way the results are reported that are not important. For example, the table for the error tolerant search includes the fixed and variable modifications while the table for the mass tolerant search includes unmodified peptides and 13C matches.

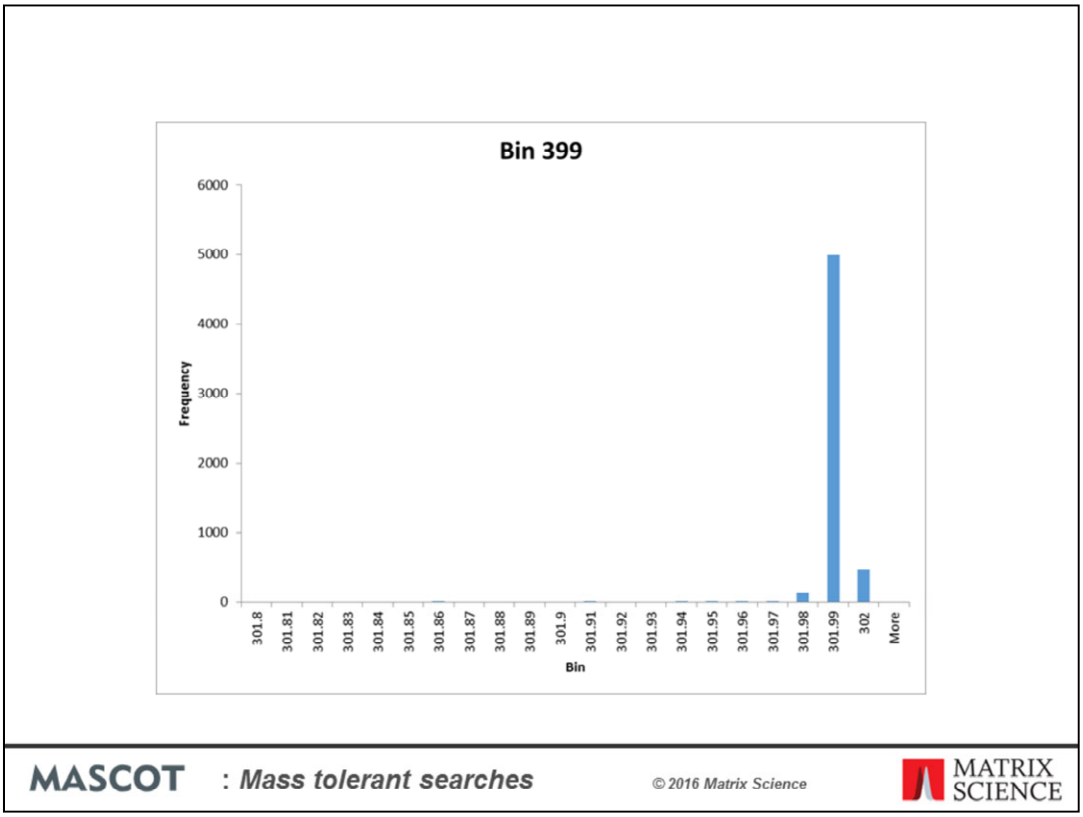
For the mass tolerant search, the counts are independent of specificity. For example, the carbamyl count includes carbamylation of N-term, S, T, and any other sites that are susceptible to this modification. The error tolerant search reports separate counts for each specificity, although this isn't always going to be meaningful. When alternative sites are close together or when the spectrum is noisy, there may be little difference in score between two alternatives. And, of course, if there is a choice of modifications within the precursor mass tolerance, the very identity of the mod may be uncertain, although such cases will be rare for this particular search because the tolerance was 5ppm.

The really important point is that several of the most abundant modifications from the mass tolerant search have a question mark against them or are labelled artefact. Let's look at the first three of these: bins 237, 399, and 233



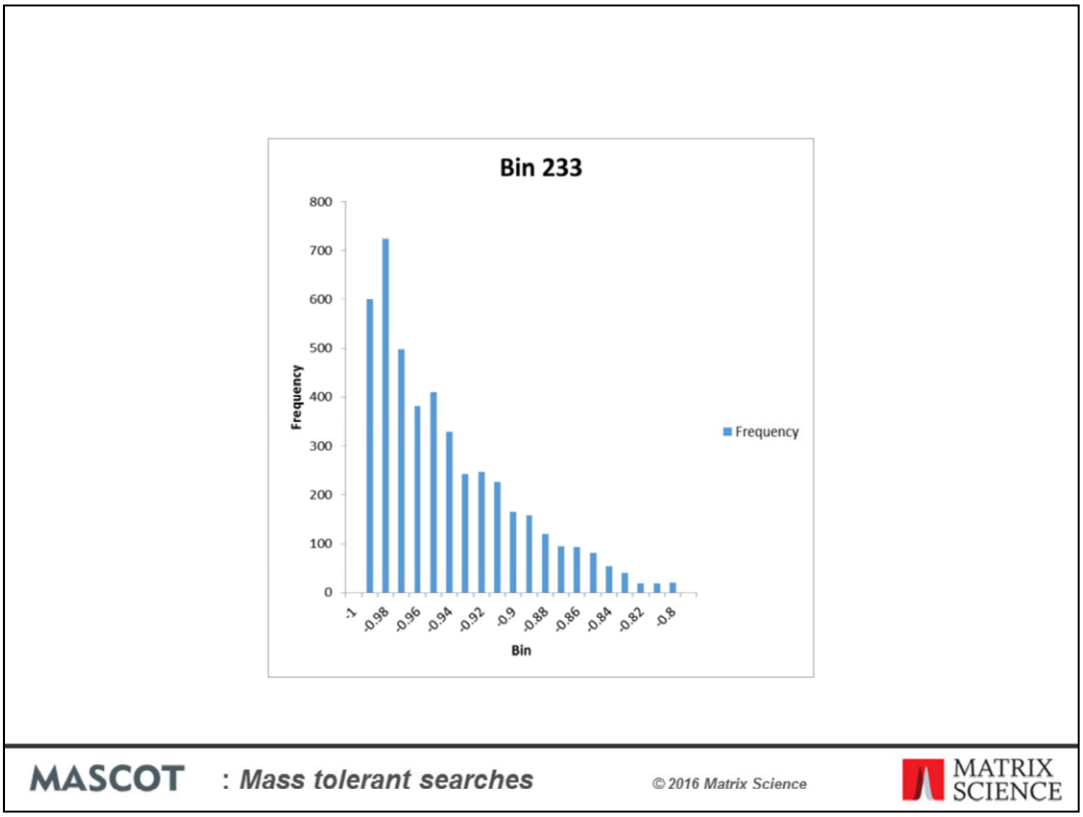
MASCOT : *Mass tolerant searches* © 2016 Matrix Science MATRIX SCIENCE

This is what the distribution looks like for bin 237. The paper reports this bin as a mass of 1.9755, which is the mean, but I suspect the spike at 2.01 is a better representative value, and can be assigned as 13C2. But what are all the other matches? It is essentially a continuum. The paper claims a peptide FDR of 0.12%, with just 625 modified peptides in total. If this is correct, and these matches are real, then each of these bins requires a different elemental composition, which is very hard to believe.



MASCOT : *Mass tolerant searches* © 2016 Matrix Science MATRIX SCIENCE

Bin 399 is a nice, clean peak. The problem is trying to figure out an assignment. It could be a combination of mods, so one approach is to look at combinations of the other high abundance modifications, but I haven't been able to come up with an assignment. The paper says nothing about this peak, even though it is the 7th most abundant modification. Does anyone have any ideas? Even with good mass accuracy, there are many possible elemental compositions for a mass of 302, and I haven't found any standard utility for listing possible formulae that includes negative counts for some elements, as may be required for a delta.

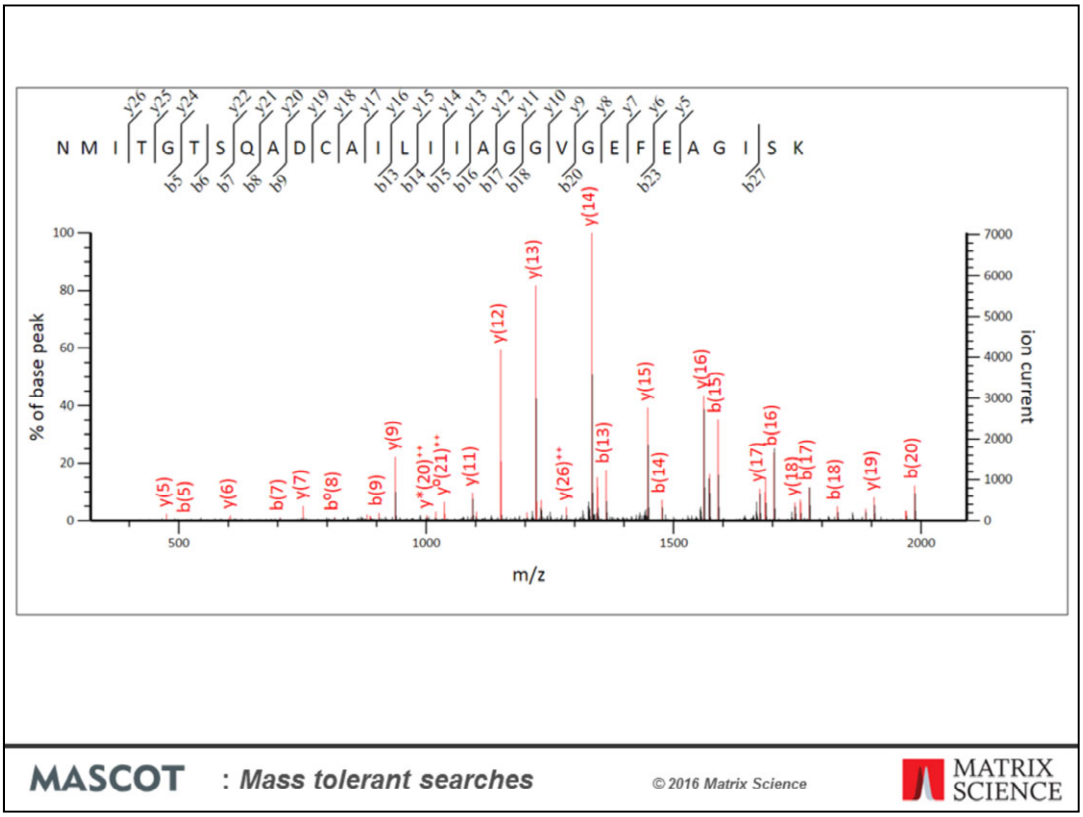


MASCOT : *Mass tolerant searches*

© 2016 Matrix Science

Bin 233 is another continuum. In the paper, it is assigned a mass of -0.95, but maybe -0.98 would be a better choice. As with the earlier example, even if you can come up with a composition for one or two of these channels, what are all the rest? These are not low level features, hidden in the grass.

The paper doesn't say anything about this issue, but we can make a good guess as to the likely cause if we consider exactly how a modification is found in a mass tolerant search



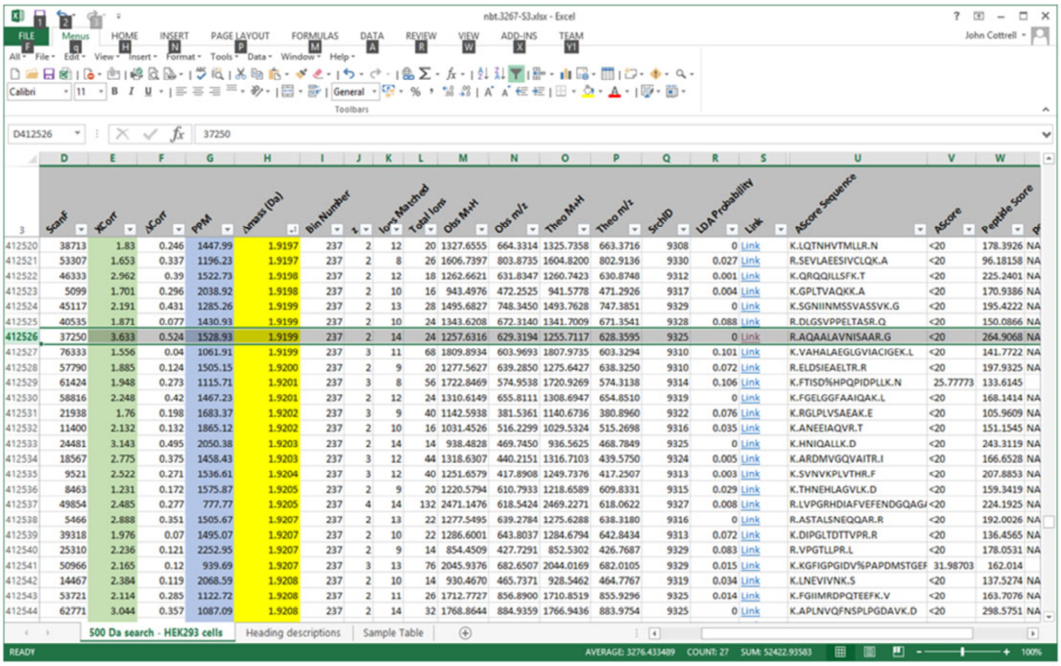
MASCOT : *Mass tolerant searches*

© 2016 Matrix Science

MATRIX SCIENCE

The calculated fragment masses used to test for a match are always those for the unmodified peptide. If you have a nice spectrum like this, with a good balance of b and y ions, and there is an unsuspected modification somewhere in the middle, this will take out roughly half the fragment matches. That is, the match is only based on those fragments that do not include the modified residue. If the modification was at or near a terminus, it would take out one complete series. For a modification on the amino terminus, you lose all the b ion matches and for a modification on the carboxy terminus, you lose all the y ion matches. If you have a good balance of b and y ions, this is much the same as having a modification in the middle – you lose half your matches - but if you only have one series it will give a bias. For example, if you only have y ions, then the closer the modification is to the C-terminus, the less likely you are to get any kind of match.

The critical weakness of the mass tolerant approach is that the mass of the modification comes solely from the difference between the calculated mass of the peptide and the observed mass of the precursor; the fragment masses play no part in determining the modification mass.



3	Scanf	XCorr	ΔCorr	PPM	Δmass (Da)	Bin Number	z	Ions Matched	Total Ions	Obs M+H	Obs m/z	Theo M+H	Theo m/z	SrchID	LDA Probability	Link	AScore Sequence	AScore	Peptide Score	P
412520	38713	1.83	0.246	1447.99	1.9197	237	2	12	20	1327.6555	664.3314	1325.7358	663.3716	9308	0	Link	K.LQTNHVTMLLR.N	<20	178.3926	NA
412521	53307	1.653	0.337	1196.23	1.9197	237	2	8	26	1606.7397	803.8735	1604.8200	802.9136	9330	0.027	Link	R.SEVLAEESIVCLQK.A	<20	96.18158	NA
412522	46333	2.962	0.39	1522.73	1.9198	237	2	12	18	1262.6621	631.8347	1260.7423	630.8748	9312	0.001	Link	K.QRQQILLSFK.T	<20	225.2401	NA
412523	5099	1.701	0.296	2038.92	1.9198	237	2	10	16	943.4976	472.2525	941.5778	471.2926	9317	0.004	Link	K.GPLTVAQKK.A	<20	170.9386	NA
412524	45117	2.191	0.431	1285.26	1.9199	237	2	13	28	1495.6827	748.3450	1493.7628	747.3851	9329	0	Link	K.SGNIINMSSVASSVK.G	<20	195.4222	NA
412525	40535	1.871	0.077	1430.93	1.9199	237	2	10	24	1343.6208	672.3140	1341.7009	671.3541	9328	0.088	Link	R.DLGSVPPELTASR.Q	<20	150.0866	NA
412526	37250	3.633	0.524	1528.93	1.9199	237	2	14	24	1257.6316	629.3194	1255.7117	628.3595	9325	0	Link	R.AQAALAVNISAAR.G	<20	264.9068	NA
412527	76333	1.556	0.04	1061.91	1.9199	237	3	11	68	1809.8934	603.9693	1807.9735	603.3294	9310	0.101	Link	K.VAHALAEGLGVIACIGEK.L	<20	141.7722	NA
412528	57790	1.885	0.124	1505.15	1.9200	237	2	9	20	1277.5627	639.2850	1275.6427	638.3250	9310	0.072	Link	R.ELDSIEAELTR.R	<20	197.9325	NA
412529	61424	1.948	0.273	1115.71	1.9201	237	3	8	56	1722.8469	574.9538	1720.9269	574.3138	9314	0.106	Link	K.FTISD%HPQPIDPLLK.N	25.77773	133.6145	
412530	58816	2.248	0.42	1467.23	1.9201	237	2	12	24	1310.6149	655.8111	1308.6947	654.8510	9319	0	Link	K.FGELGGFAAIQAK.L	<20	168.1414	NA
412531	21938	1.76	0.198	1683.37	1.9202	237	3	9	40	1142.5938	381.5361	1140.6736	380.8960	9322	0.076	Link	K.RGLPLVSAEAK.E	<20	105.9609	NA
412532	11400	2.132	0.132	1865.12	1.9202	237	2	10	16	1031.4526	516.2299	1029.5324	515.2698	9316	0.035	Link	K.ANEEIAQVR.T	<20	151.1545	NA
412533	24481	3.143	0.495	2050.38	1.9203	237	2	14	14	938.4828	469.7450	936.5625	468.7849	9325	0	Link	K.HNIQALLK.D	<20	243.3119	NA
412534	18567	2.775	0.375	1458.43	1.9203	237	3	12	44	1318.6307	440.2151	1316.7103	439.5750	9324	0.005	Link	K.ARDMVGQVAITR.I	<20	166.6528	NA
412535	9521	2.522	0.271	1536.61	1.9204	237	3	12	40	1251.6579	417.8908	1249.7376	417.2507	9313	0.003	Link	K.SVNVKPLVTHR.F	<20	207.8853	NA
412536	8463	1.231	0.172	1575.87	1.9205	237	2	9	20	1220.5794	610.7933	1218.6589	609.8331	9315	0.029	Link	K.THNEHLAGVLK.D	<20	159.3419	NA
412537	49854	2.485	0.277	777.77	1.9205	237	4	14	132	2471.1476	618.5424	2469.2271	618.0622	9327	0.008	Link	R.LVPGRHDIAFVEFENDGQAG/	<20	224.1925	NA
412538	5466	2.888	0.351	1505.67	1.9207	237	2	13	22	1277.5495	639.2784	1275.6288	638.3180	9316	0	Link	R.ASTALSNEQQAR.R	<20	192.0026	NA
412539	39318	1.976	0.07	1495.07	1.9207	237	2	10	22	1286.6001	643.8037	1284.6794	642.8434	9313	0.072	Link	K.DIPGLTDTTVPR.R	<20	136.4565	NA
412540	25310	2.236	0.121	2252.95	1.9207	237	2	9	14	854.4509	427.7291	852.5302	426.7687	9329	0.083	Link	R.VPGTLLPR.L	<20	178.0531	NA
412541	50966	2.165	0.12	939.69	1.9207	237	3	13	76	2045.9376	682.6507	2044.0169	682.0105	9329	0.015	Link	K.KGFIGPGIDV%PAPDMSTGEF	31.98703	162.014	
412542	14467	2.384	0.119	2068.59	1.9208	237	2	10	14	930.4670	465.7371	928.5462	464.7767	9319	0.034	Link	K.LNEVIVNK.S	<20	137.5274	NA
412543	53721	2.114	0.285	1122.72	1.9208	237	2	11	26	1712.7727	856.8900	1710.8519	855.9296	9325	0.014	Link	K.FGIIMRDPQTEEFK.V	<20	163.7076	NA
412544	62771	3.044	0.357	1087.09	1.9208	237	2	14	32	1768.8644	884.9359	1766.9436	883.9754	9325	0	Link	K.APLNVQFNSPLPGDAVK.D	<20	298.5751	NA

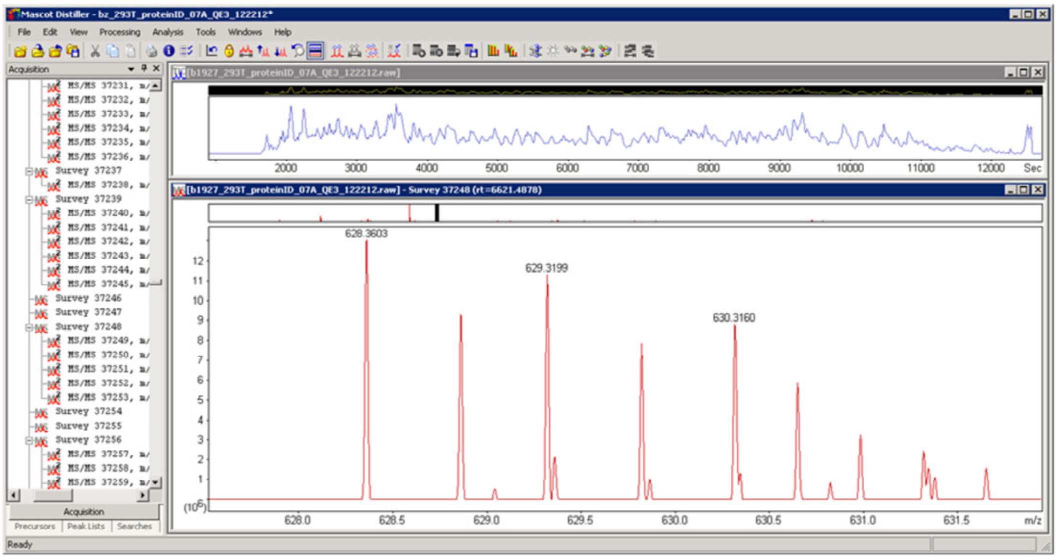
MASCOT : *Mass tolerant searches*

© 2016 Matrix Science

MATRIX SCIENCE

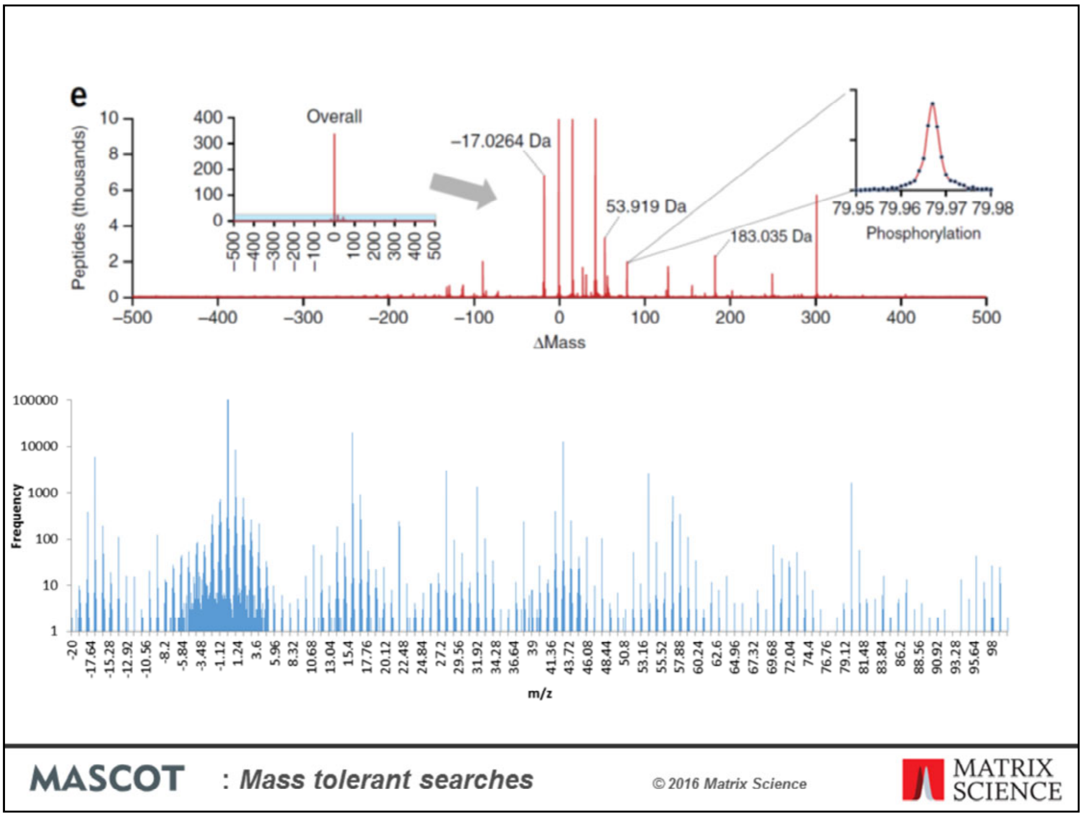
Let’s take one of the strong matches from the continuum of bin 237. Observed m/z 629.3194 and a match to AQAALAVNISAAR. The difference between the observed mass and the calculated mass is 1.92 Da, which doesn’t fit to anything in Unimod and is outside the ‘allowed’ range of mass defects for peptide-like molecules.

If we locate this scan in the original raw file and take a look at the precursor region of the survey scan …



This is what we find. The precursor with an m/z of 629.3199 is in the middle . But, we can see two other precursors, equally strong. Notice the difference between the first two: 0.96 m/z at charge 2+, corresponding to a mass difference of 1.92. It seems pretty clear that the mass tolerant search hasn't really discovered a modified peptide. The instrument was targeting 629.32 but the fragments in the MS/MS spectrum that gave the strongest match came from the precursor at 628.36. In effect, the precursor mass was 'wrong'. Since there is nothing to tie the fragment masses to the precursor mass, the error goes undetected, and a spurious modification is reported.

How often this happens is hard to say. The Gygi paper reports 185,000 modified peptides in the open search that were not found in the standard search, and it would be a mammoth task to make a forensic analysis of these. What we can say is that whenever there are overlapping precursors, there is a very real possibility of the 'wrong' mass being taken, causing the inference of a spurious modification.



MASCOT : *Mass tolerant searches*

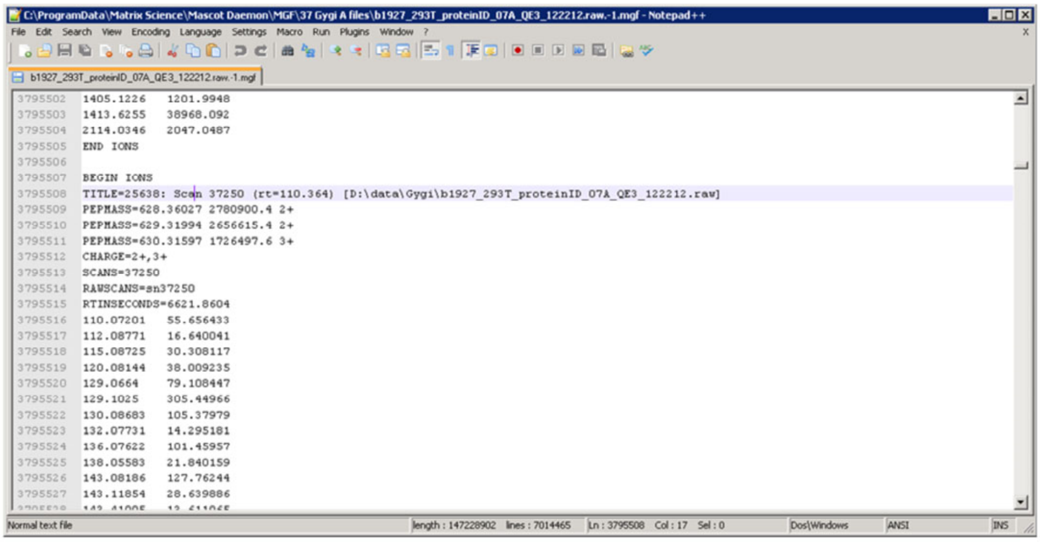
© 2016 Matrix Science

MATRIX SCIENCE

Going back to the summary histogram, here we have zoomed into the central range, -20 to +100 Da, and switched to a log scale for the counts. You can see that the region that has the highest level of ‘background’ is the region around 0 Da. This is what you would expect for overlapping distributions of the same charge. The width of the instrument selection window is user adjustable, but I believe 4 m/z units is typical, so we are likely to see false modifications of a few da at most on unmodified peptides. However, the artefact applies equally to modified peptides. You might believe you have discovered a peptide with a delta of 40 Da or 44 Da and find it is actually a modification of 42 Da from a different precursor.

(If the overlapping distributions have different charge states, then the spurious modification could be very large, but usually these cases will fall outside the mass range studied in the paper, +/- 500 Da.)

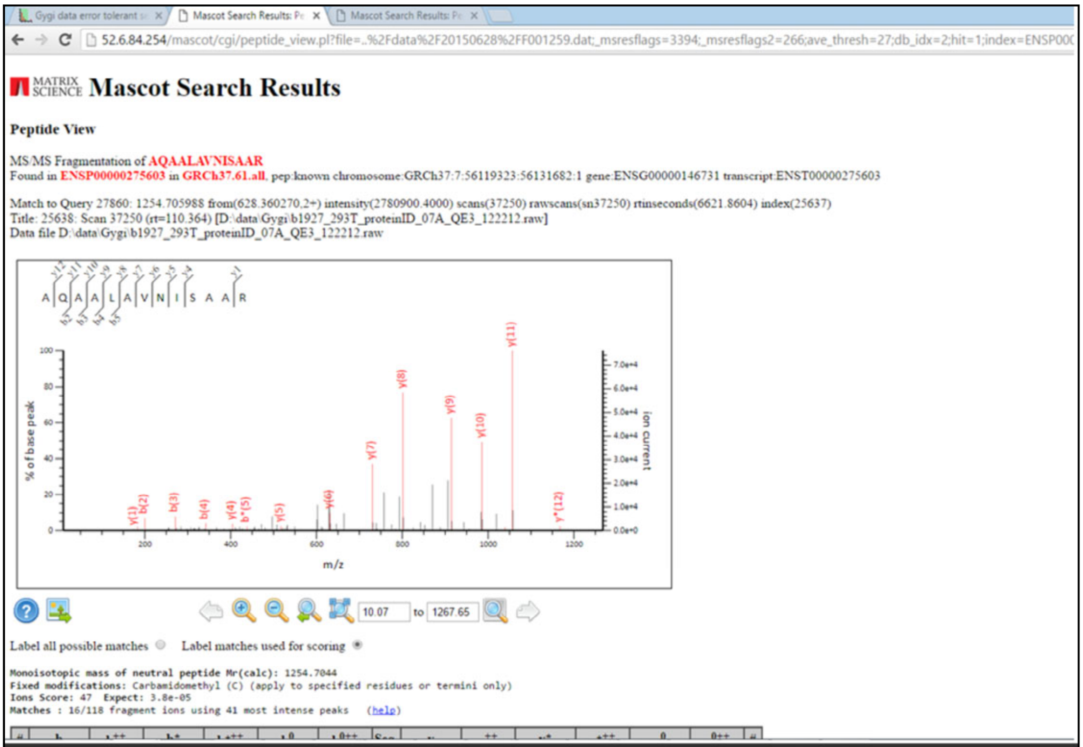
Note that such errors are outside the scope of the FDR as estimated by target/decoy. The delta mass plays no part in the scoring and a match is counted as true or false independently of whether the delta mass is true or false.


C:\ProgramData\Matrix Science\Mascot Daemon\MGF\37 Gygi A files\b1927_293T_proteinID_07A_QE3_122212.raw.-1.mgf - Notepad++
File Edit Search View Encoding Language Settings Macro Run Plugins Window ?
b1927_293T_proteinID_07A_QE3_122212.raw.-1.mgf
Normal text file
length : 147228902 lines : 7014465
Ln : 3795508 Col : 17 Sel : 0
Dos\Windows
ANSI
INS

```
1405.1226   1201.9948
1413.6255   38968.092
2114.0346   2047.0487
END IONS

BEGIN IONS
TITLE=25638: Scan 37250 (rt=110.364) [D:\data\Gygi\b1927_293T_proteinID_07A_QE3_122212.raw]
PEPMASS=628.36027 2780900.4 2+
PEPMASS=629.31994 2656615.4 2+
PEPMASS=630.31597 1726497.6 3+
CHARGE=2+,3+
SCANS=37250
RAWSCANS=sn37250
RTINSECONDS=6621.8604
110.07201   55.656433
112.08771   16.640041
115.08725   30.308117
120.08144   38.009235
129.0664    79.108447
129.1025    305.44966
130.08683   105.37979
132.07731   14.295181
136.07622   101.45957
138.05583   21.840159
143.08186   127.76244
143.11854   28.639886
```

The error tolerant search is much more constrained because it is looking for a fit to the modified peptide. A false modification will simply make the match worse. In this particular case, if the MS/MS spectrum was only associated with the central precursor m/z value of 629.3199, there would be no match. This is better than a false match, of course, but Distiller 2.5 and Mascot Server 2.5 introduced support for multiple precursor m/z values for a single MS/MS spectrum. This is what the Distiller peak list looks like.



Gygi data error tolerant s
Mascot Search Results: Pe
Mascot Search Results: Pe
52.6.84.254/mascot/cgi/peptide_view.pl?file=..%2Fdata%2F20150628%2FF001259.dat;_msresflags=3394;_msresflags2=266;ave_thresh=27;db_idx=2;hit=1;index=ENSP00
y12 y11 y10 y9 y8 y7 y6 y5 y4 y1
A Q A A L A V N I S A A R
b2 b3 b4 b5
% of base peak
ion current
m/z
y(1) b(2) b(3) b(4) y(4) b*(5) y(5) y(6) y(7) y(8) y(9) y(10) y(11) y*(12)
0 20 40 60 80 100
200 400 600 800 1000 1200
0.0e+0 1.0e+4 2.0e+4 3.0e+4 4.0e+4 5.0e+4 6.0e+4 7.0e+4
10.07 to 1267.65

Mascot Search Results

Peptide View

MS/MS Fragmentation of **AQAALAVNISAAR**
Found in **ENSP00000275603** in **GRCh37.61.all**, pep:known chromosome:GRCh37:7:56119323:56131682:1 gene:ENSG00000146731 transcript:ENST00000275603

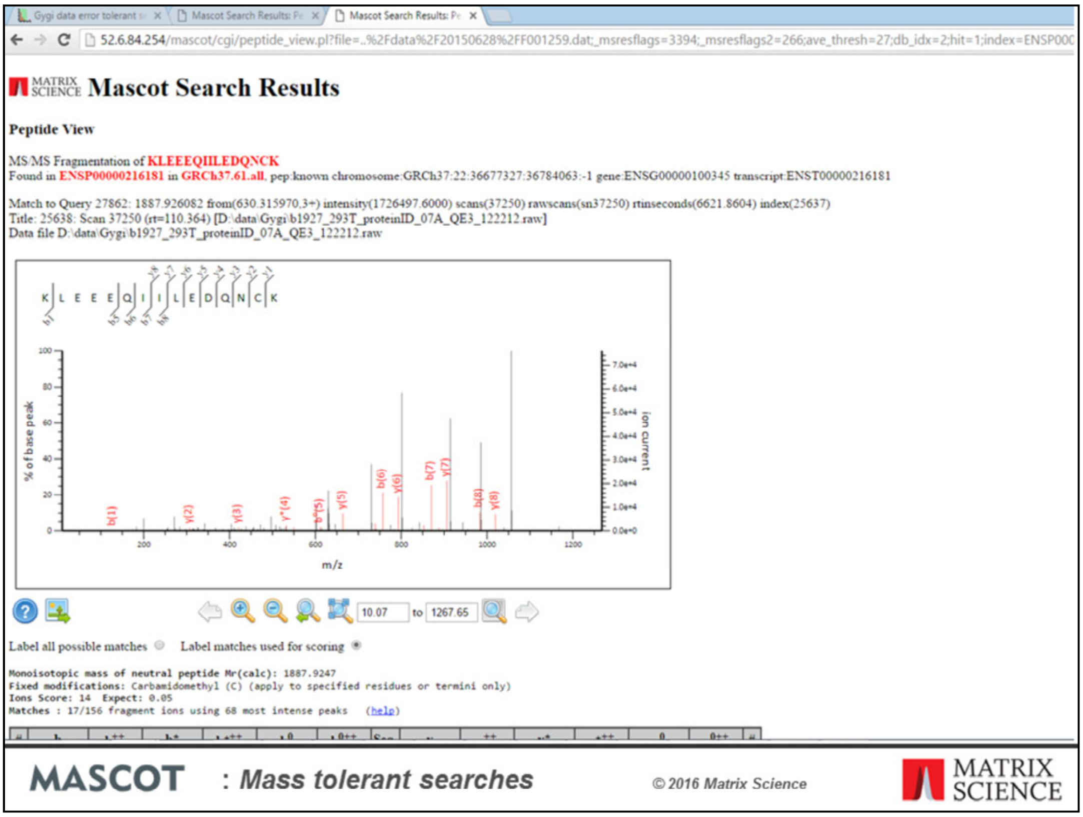
Match to Query 27860: 1254.705988 from(628.360270,2+) intensity(2780900.4000) scans(37250) rawscans(sn37250) rtinseconds(6621.8604) index(25637)
Title: 25638: Scan 37250 (rt=110.364) [D:\data\Gygi\b1927_293T_proteinID_07A_QE3_122212.raw]
Data file D:\data\Gygi\b1927_293T_proteinID_07A_QE3_122212.raw

Label all possible matches ○ Label matches used for scoring ●

```
Monoisotopic mass of neutral peptide Mr(calc): 1254.7044
Fixed modifications: Carbamidomethyl (C) (apply to specified residues or termini only)
Ions Score: 47  Expect: 3.8e-05
Matches : 16/118 fragment ions using 41 most intense peaks   (help)
```

MASCOT : *Mass tolerant searches* © 2016 Matrix Science MATRIX SCIENCE

The lowest m/z precursor gets the correct match to the unmodified peptide


Gygi data error tolerant s
Mascot Search Results: Pe
Mascot Search Results: Pe
52.6.84.254/mascot/cgi/peptide_view.pl?file=..%2Fdata%2F20150628%2FF001259.dat;_msresflags=3394;_msresflags2=266;ave_thresh=27;db_idx=2;hit=1;index=ENSP000
K L E E E Q I I L E D Q N C K
y8 y7 y6 y5 y4 y3 y2 y1
b1 b5 b6 b7 b8
% of base peak
ion current
m/z
b(1)
y(2)
y(3)
y*(4)
b*(5)
y(5)
b(6)
y(6)
b(7)
y(7)
b(8)
y(8)
10.07
to
1267.65

MATRIX SCIENCE **Mascot Search Results**

Peptide View

MS/MS Fragmentation of **KLEEEQIILEDQNCK**
Found in **ENSP00000216181** in **GRCh37.61.all**, pep:known chromosome:GRCh37:22:36677327:36784063:-1 gene:ENSG00000100345 transcript:ENST00000216181

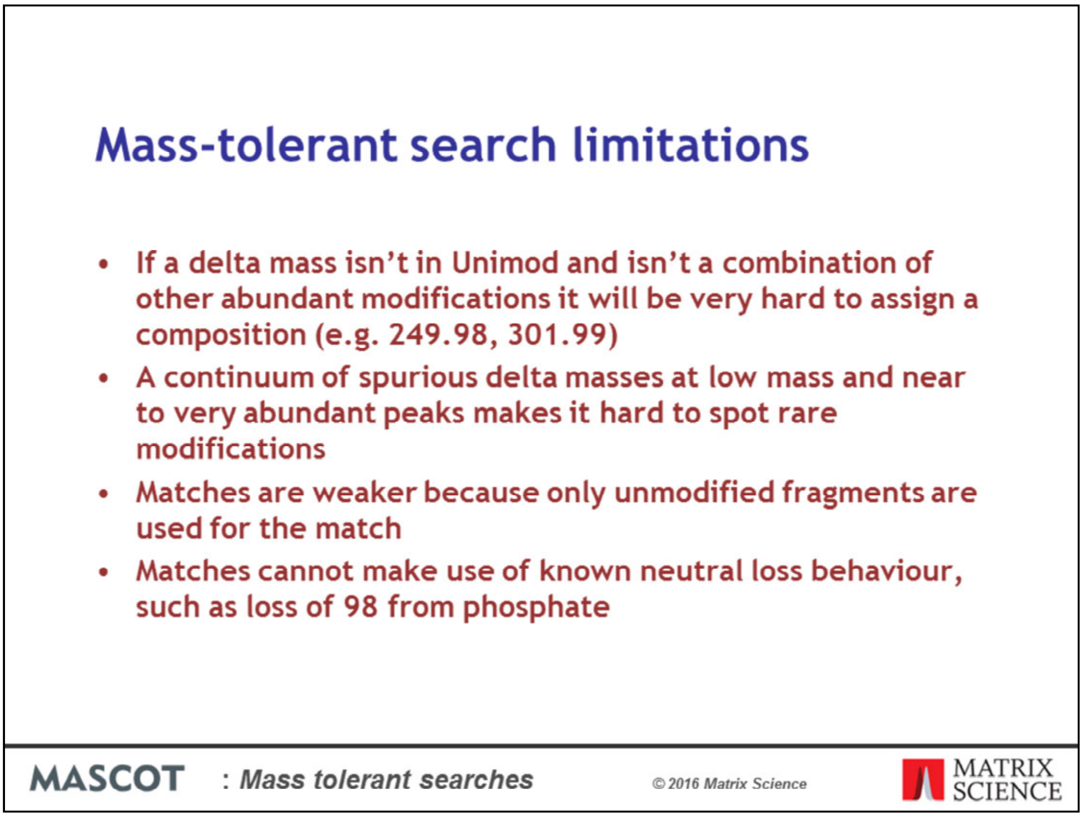
Match to Query 27862: 1887.926082 from(630.315970,3+) intensity(1726497.6000) scans(37250) rawscans(sn37250) rtinseconds(6621.8604) index(25637)
Title: 25638: Scan 37250 (rt=110.364) [D:\data\Gygi\b1927_293T_proteinID_07A_QE3_122212.raw]
Data file D:\data\Gygi\b1927_293T_proteinID_07A_QE3_122212.raw

Label all possible matches ○ Label matches used for scoring ●

```
Monoisotopic mass of neutral peptide Mr(calc): 1887.9247
Fixed modifications: Carbamidomethyl (C) (apply to specified residues or termini only)
Ions Score: 14  Expect: 0.05
Matches : 17/156 fragment ions using 68 most intense peaks   (help)
```

MASCOT : *Mass tolerant searches* © 2016 Matrix Science MATRIX SCIENCE

The higher m/z value, which has charge 3+, gives a match to a completely different peptide. If you compare the fragment matches, you'll see that this is a very nice example of a chimeric spectrum. These two precursors account for all the intense fragment peaks, so it isn't surprising that we don't get a match for the middle precursor



Mass-tolerant search limitations

- If a delta mass isn’t in Unimod and isn’t a combination of other abundant modifications it will be very hard to assign a composition (e.g. 249.98, 301.99)
- A continuum of spurious delta masses at low mass and near to very abundant peaks makes it hard to spot rare modifications
- Matches are weaker because only unmodified fragments are used for the match
- Matches cannot make use of known neutral loss behaviour, such as loss of 98 from phosphate

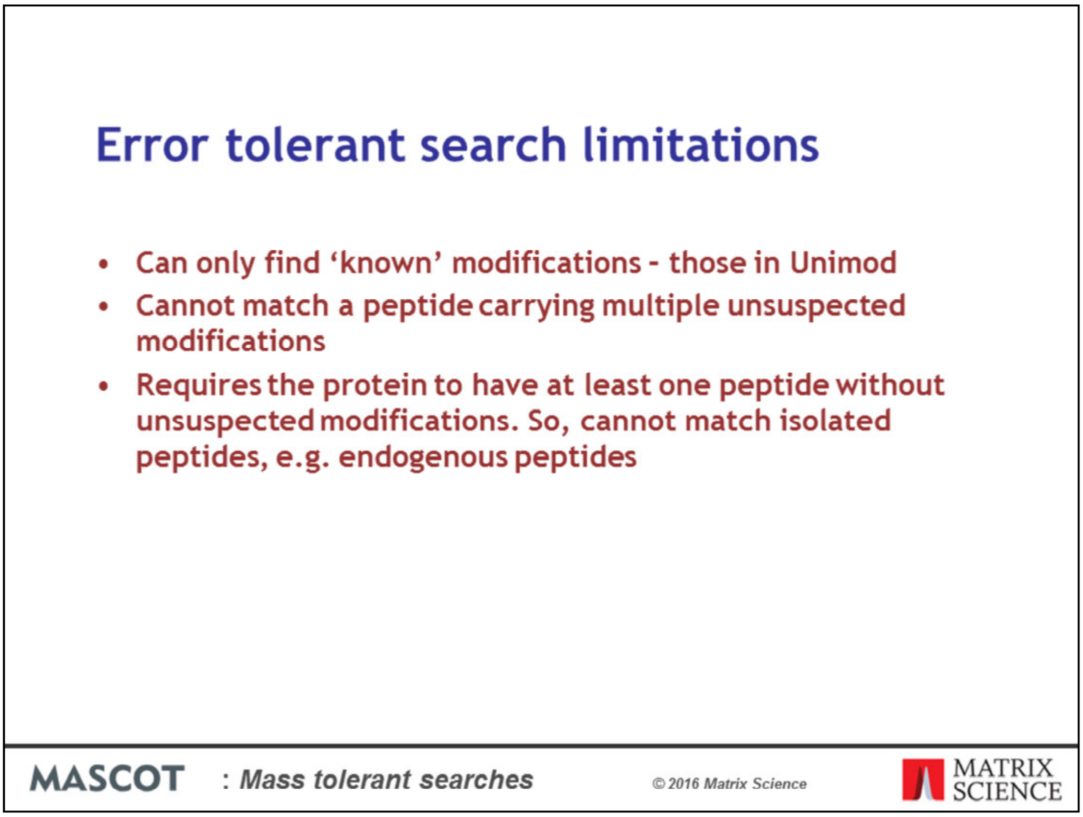
MASCOT : *Mass tolerant searches* © 2016 Matrix Science MATRIX SCIENCE

In summary, the mass tolerant search certainly shows spikes for delta masses corresponding to common modifications. But, if the mass isn’t in Unimod, it will be a challenge to figure out the chemical identity.

The site of modification has to be determined separately by using a calculation such as A-Score.

A background of spurious delta masses caused by taking the wrong precursor mass makes it difficult to identify low abundance modifications.

Matches to modified peptides are weaker than in a conventional search or an error tolerant search because only half the fragments are available for matching, on average. Similarly, the matching cannot take advantage of known neutral loss behaviour



Error tolerant search limitations

- Can only find 'known' modifications - those in Unimod
- Cannot match a peptide carrying multiple unsuspected modifications
- Requires the protein to have at least one peptide without unsuspected modifications. So, cannot match isolated peptides, e.g. endogenous peptides

MASCOT : *Mass tolerant searches* © 2016 Matrix Science MATRIX SCIENCE

One of the limitations in an error tolerant search are that it can only find 'known' modifications. As Unimod becomes more comprehensive, this becomes less of a concern.

For each peptide, it tests modifications serially, so it will not give a match to a peptide with multiple unsuspected modifications, such as might be found in a histone. In practice, the same limitation applies to the mass tolerant search; each modification takes out potential fragment matches, so having two or more makes getting a match very unlikely unless they are on adjacent residues.

I think the final limitation is the most serious. You can't use a two pass approach on endogenous peptides.

Maybe this is the most appropriate application for the mass tolerant search. If the data complexity is kept low, so that chimeric spectra are very rare, then the mass tolerant search may be an easier way to find modifications on endogenous peptides than an error tolerant sequence tag search