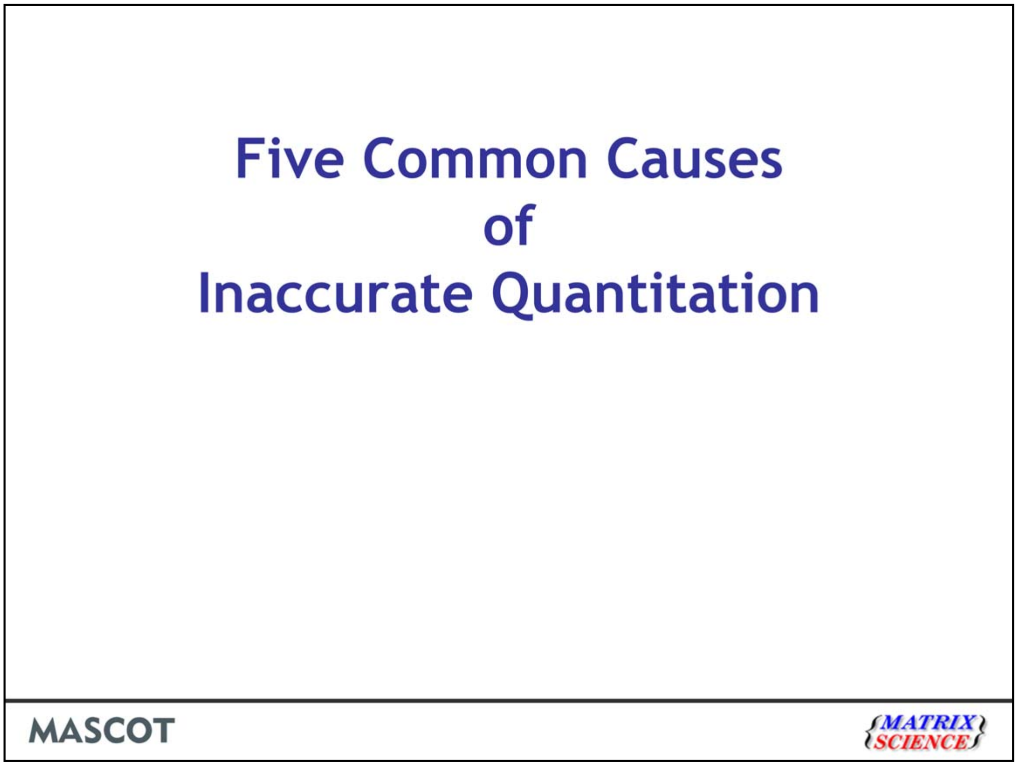


Five Common Causes of Inaccurate Quantitation

MASCOT

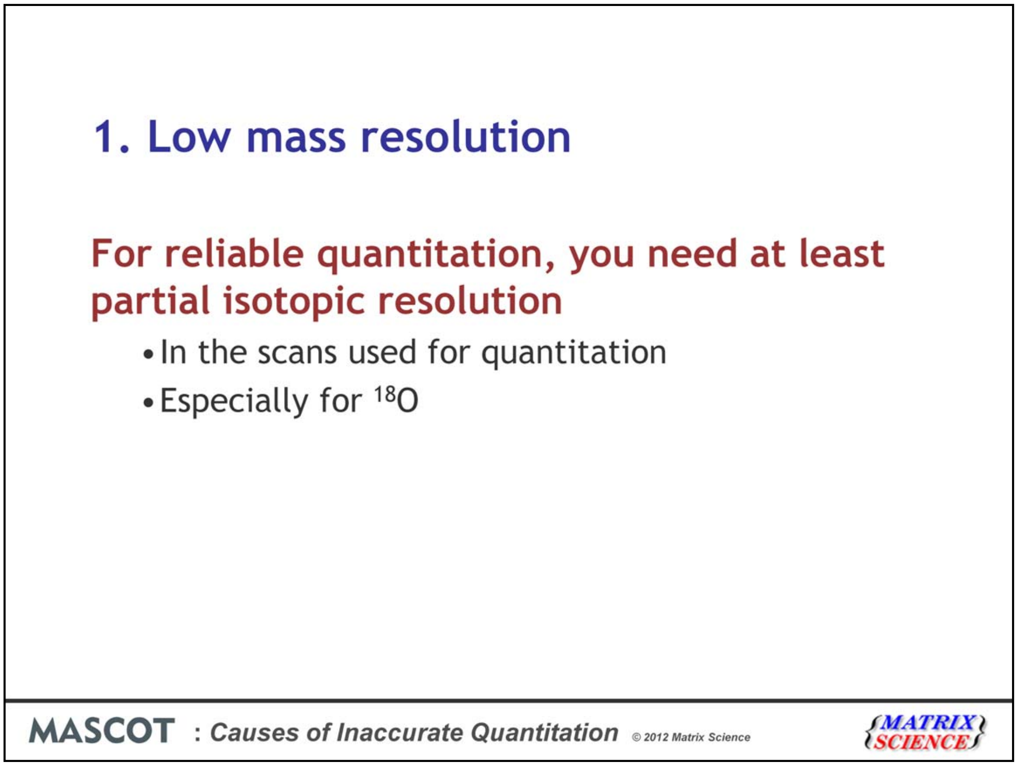
MATRIX SCIENCE

I must make it clear that I am only considering data processing problems

MASCOT : *Causes of Inaccurate Quantitation* © 2012 Matrix Science

MATRIX SCIENCE

I hardly dare think about all the things that can go wrong at the bench, or when acquiring the mass spec. data



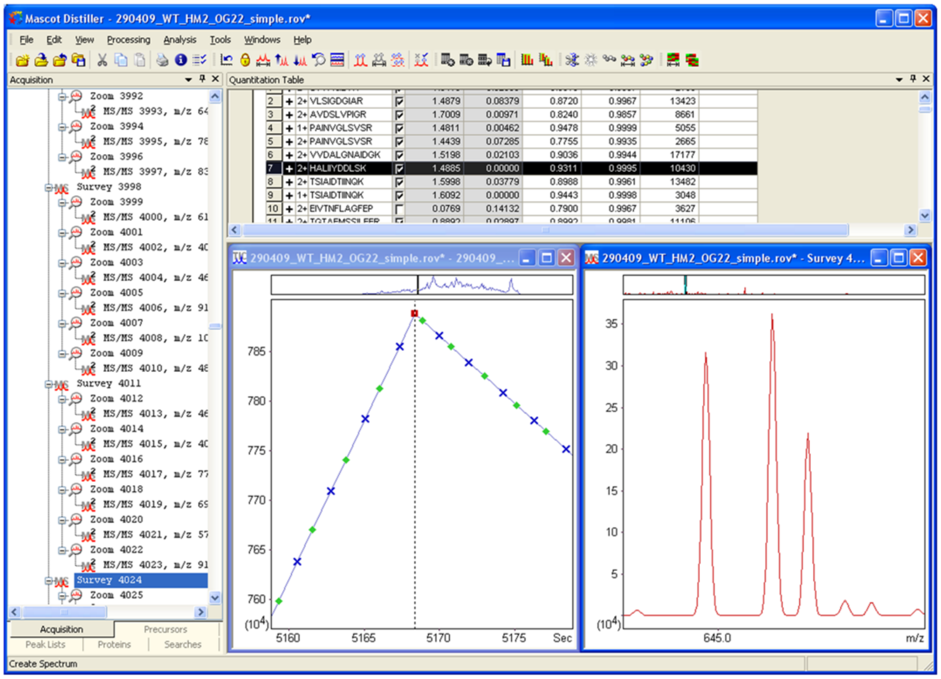
1. Low mass resolution

For reliable quantitation, you need at least partial isotopic resolution

- In the scans used for quantitation
- Especially for ^{18}O

MASCOT : *Causes of Inaccurate Quantitation* © 2012 Matrix Science

I think most people would agree that you need some degree of isotopic resolution to get reliable quantitation. If you are using a ‘classic’ ion trap, 3+ and even 2+ peaks may not show any isotopic resolution

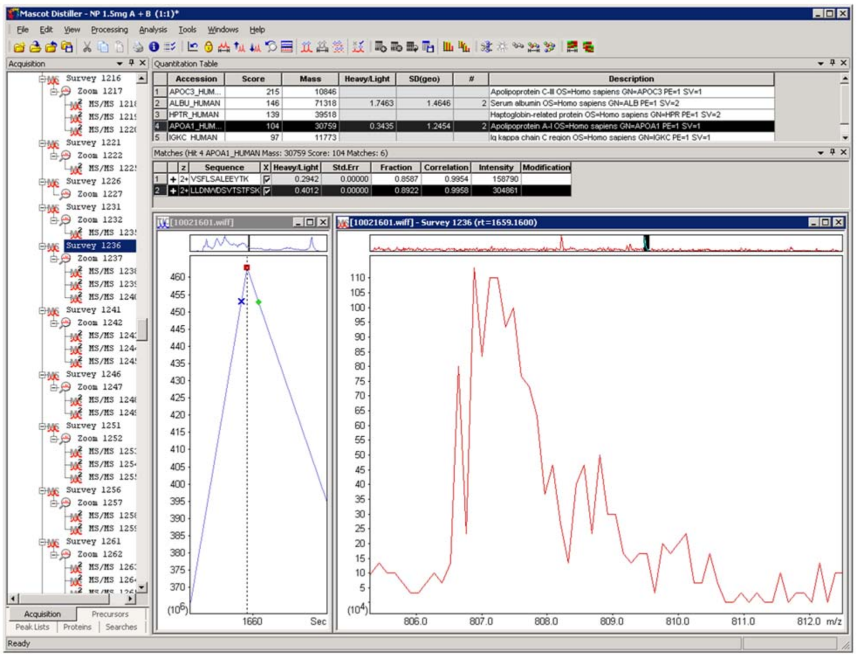


```
Zoom 3992
  MS/MS 3993, m/z 64
Zoom 3994
  MS/MS 3995, m/z 78
Zoom 3996
  MS/MS 3997, m/z 83
Survey 3998
Zoom 3999
  MS/MS 4000, m/z 61
Zoom 4001
  MS/MS 4002, m/z 40
Zoom 4003
  MS/MS 4004, m/z 46
Zoom 4005
  MS/MS 4006, m/z 91
Zoom 4007
  MS/MS 4008, m/z 10
Zoom 4009
  MS/MS 4010, m/z 48
Survey 4011
Zoom 4012
  MS/MS 4013, m/z 46
Zoom 4014
  MS/MS 4015, m/z 40
Zoom 4016
  MS/MS 4017, m/z 77
Zoom 4018
  MS/MS 4019, m/z 69
Zoom 4020
  MS/MS 4021, m/z 57
Zoom 4022
  MS/MS 4023, m/z 91
Survey 4024
Zoom 4025
```

2	+	2+	VLSIGDGIAR	☑	1.4879	0.08379	0.8720	0.9967	13423
3	+	2+	AVDSLVPIGR	☑	1.7009	0.00971	0.8240	0.9857	8661
4	+	1+	PAINVGLSVSR	☑	1.4811	0.00462	0.9478	0.9999	5055
5	+	2+	PAINVGLSVSR	☑	1.4439	0.07285	0.7755	0.9935	2665
6	+	2+	VVDALGNAIDGK	☑	1.5198	0.02103	0.9036	0.9944	17177
7	+	2+	HALIIYDDLSK	☑	1.4885	0.00000	0.9311	0.9995	10430
8	+	2+	TSIAIDTIINQK	☑	1.5998	0.03779	0.8988	0.9961	13482
9	+	1+	TSIAIDTIINQK	☑	1.6092	0.00000	0.9443	0.9998	3048
10	+	2+	EIVTNFLAGFEP	☐	0.0769	0.14132	0.7900	0.9967	3627
11	+	2+	TGTAFMSSILEER	☑	0.8892	0.02897	0.8992	0.9981	11106

MASCOT : *Causes of Inaccurate Quantitation* © 2012 Matrix Science

This is some 18O data from a ‘classic’ trap. This particular peptide gets a strong match. At first glance, the resolution in the precursor region of a survey scan looks pretty good. But, if you look more carefully, these peaks are not from resolved isotope distributions. For a start, there aren’t enough peaks. The data have been saved to the raw file as centroids, not profile, as is common practice with traps.


Mascot Distiller - NP 1.5mg A + B (1:1)*
File Edit View Processing Analysis Tools Windows Help
Acquisition
Survey 1216
Zoom 1217
Survey 1221
Survey 1226
Survey 1231
Survey 1236
Survey 1241
Survey 1246
Survey 1251
Survey 1256
Survey 1261
Acquisition Precursors
Peak Lists Proteins Searches
Ready
[10021601.wiff]
1660 Sec
(10^6)
[10021601.wiff] - Survey 1236 (rt=1659.1600)
(10^4)
806.0 807.0 808.0 809.0 810.0 811.0 812.0 m/z

Quantitation Table

	Accession	Score	Mass	Heavy/Light	SD(geo)	#	Description
1	APOC3_HUM...	215	10846				Apolipoprotein C-III OS=Homo sapiens GN=APOC3 PE=1 SV=1
2	ALBU_HUMAN	146	71318	1.7463	1.4646	2	Serum albumin OS=Homo sapiens GN=ALB PE=1 SV=2
3	HPTR_HUMAN	139	39518				Haptoglobin-related protein OS=Homo sapiens GN=HPR PE=1 SV=2
4	APOA1_HUM...	104	30759	0.3435	1.2454	2	Apolipoprotein A-I OS=Homo sapiens GN=APOA1 PE=1 SV=1
5	IGKC_HUMAN	97	11773				Ig kappa chain C region OS=Homo sapiens GN=IGKC PE=1 SV=1

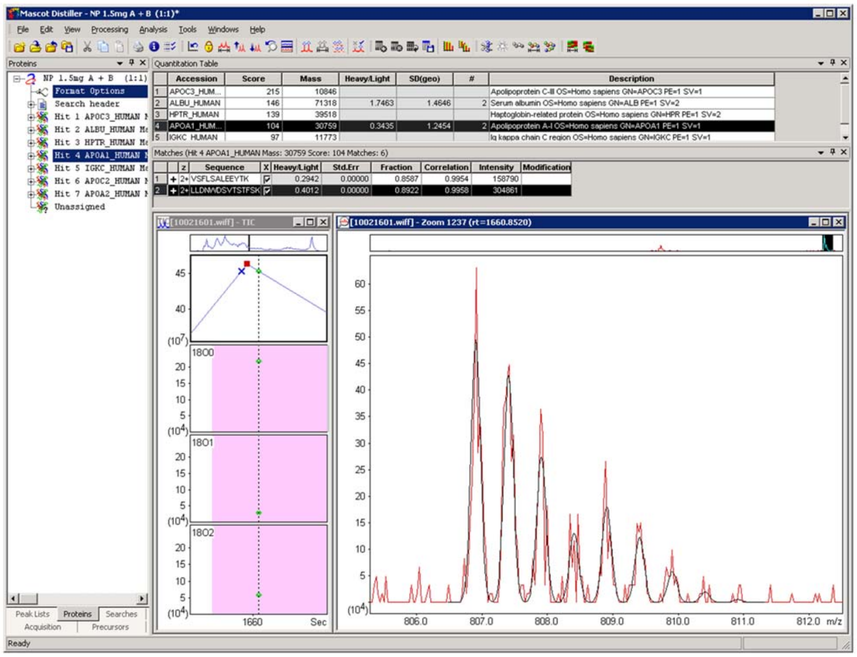
Matches (Hit 4 APOA1_HUMAN Mass: 30759 Score: 104 Matches: 6)

		z	Sequence	X	Heavy/Light	Std.Err	Fraction	Correlation	Intensity	Modification
1	+	2+	VSFLSALEEYTK	☑	0.2942	0.00000	0.8587	0.9954	158790	
2	+	2+	LLDNWDSVTSTFSK	☑	0.4012	0.00000	0.8922	0.9958	304861	

MASCOT : *Causes of Inaccurate Quantitation* © 2012 Matrix Science

MATRIX SCIENCE

This is a different file where the survey scans have been saved as profile data. Now we can see the true picture. When an unresolved distribution such as this is centroided, it gets broken up into peaks in an arbitrary way. Trying to use such survey scans for any type of quantitation would be difficult, whether saved as profile data or centroids. For 18O labelling, the situation is hopeless because the separation between heavy and light is only 4 Da, and it is essential to deconvolute the distributions.

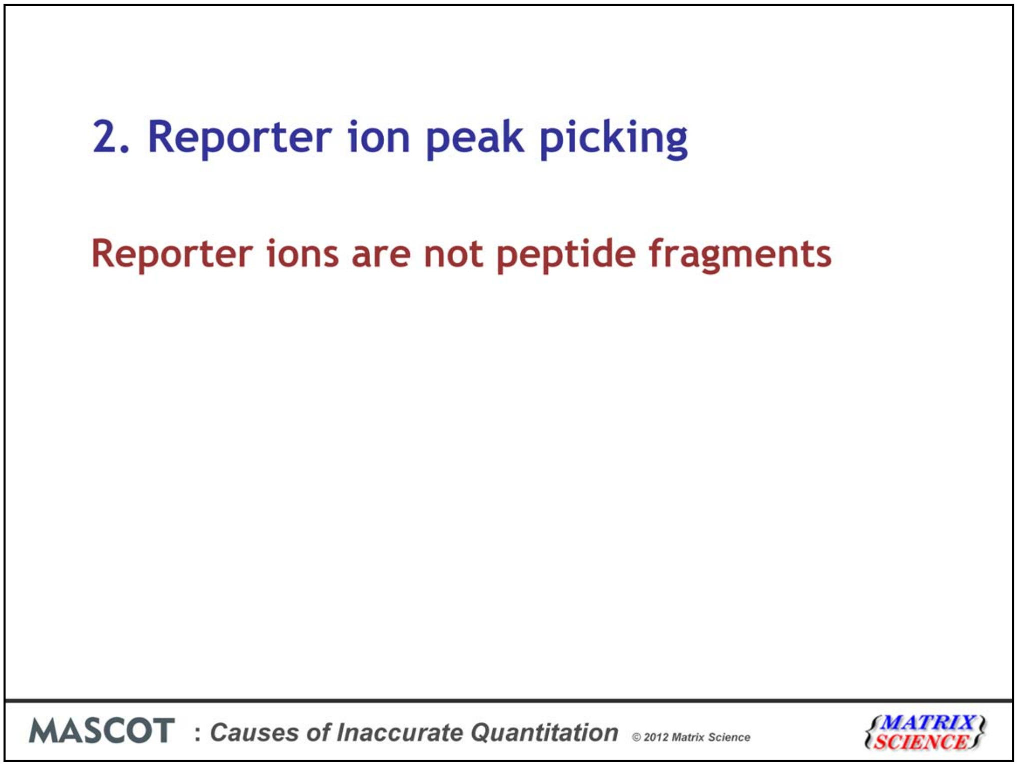


	Accession	Score	Mass	Heavy/Light	SD(geo)	#	Description
1	APOC3_HUM...	215	10846				Apolipoprotein C-III OS=Homo sapiens GN=APOC3 PE=1 SV=1
2	ALBU_HUMAN	146	71318	1.7463	1.4646	2	Serum albumin OS=Homo sapiens GN=ALB PE=1 SV=2
3	HPTR_HUMAN	139	39518				Haptoglobin-related protein OS=Homo sapiens GN=HPR PE=1 SV=2
4	APOA1_HUM...	104	30759	0.3435	1.2454	2	Apolipoprotein A-I OS=Homo sapiens GN=APOA1 PE=1 SV=1
5	IGKC_HUMAN	97	11773				Ig kappa chain C region OS=Homo sapiens GN=IGKC PE=1 SV=1

Matches (Hit 4 APOA1_HUMAN Mass: 30759 Score: 104 Matches: 6)

		z	Sequence	X	Heavy/Light	Std.Err	Fraction	Correlation	Intensity	Modification
1	+	2+	VSFLSALEEYTK	☑	0.2942	0.00000	0.8587	0.9954	158790	
2	+	2+	LLDNWDSVTSTFSK	☑	0.4012	0.00000	0.8922	0.9958	304861	

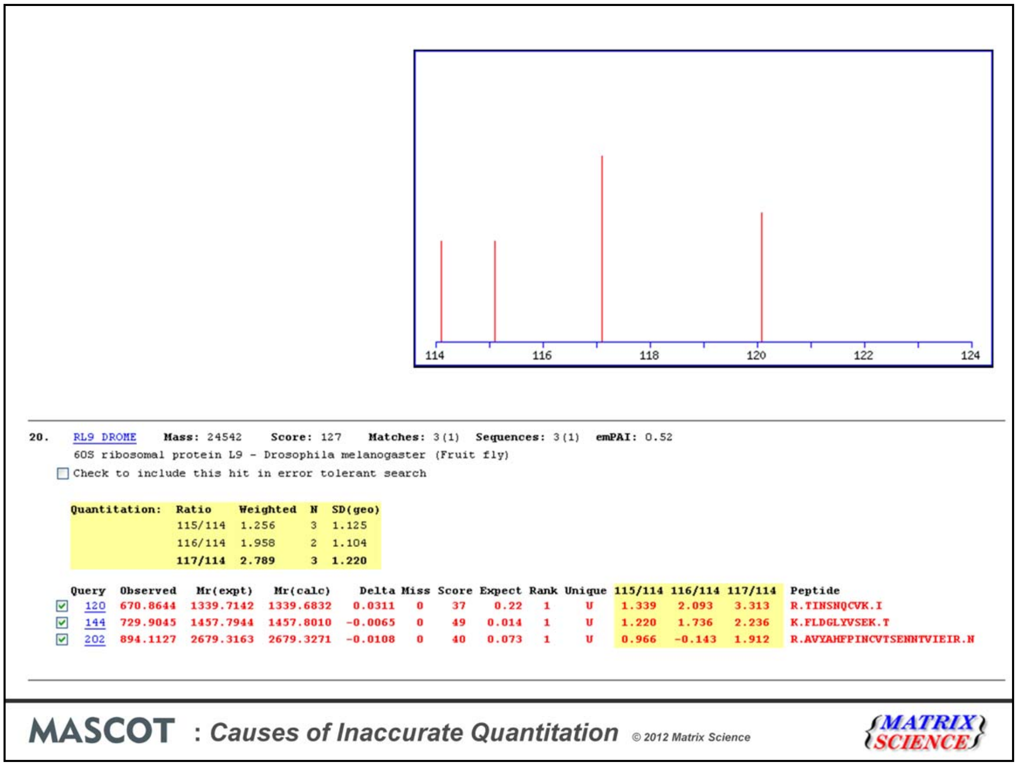
If we look at an adjacent zoom scan, we can see what the isotope pattern should look like. Signal to noise is still not great, but deconvolution becomes possible when the peaks are fully resolved. So, with zoom scans, even though you might only have a single scan for each precursor, you can get reasonable results from a standard trap.



2. Reporter ion peak picking

Reporter ions are not peptide fragments

If you are using iTRAQ or TMT, it is very important to understand that the reporter ions are not peptide fragments. Make sure your peak picking software doesn’t try to apply some standard de-isotoping algorithm, designed for peptides. This can only distort the relative intensities of reporter ion peaks.



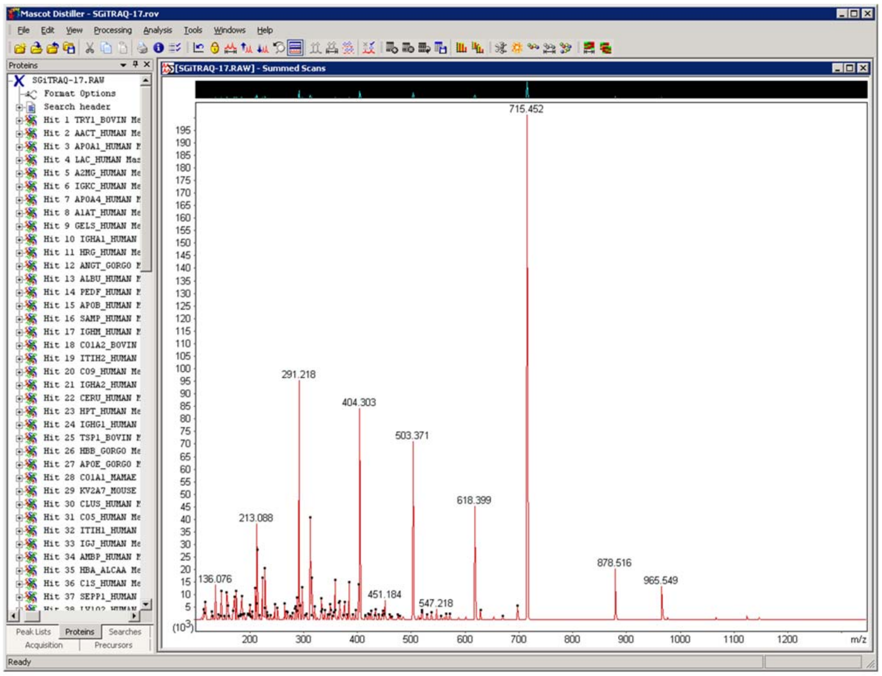
20. RL9_DROME **Mass:** 24542 **Score:** 127 **Matches:** 3(1) **Sequences:** 3(1) **emPAI:** 0.52
60S ribosomal protein L9 - Drosophila melanogaster (Fruit fly)
□ Check to include this hit in error tolerant search

Quantitation:	Ratio	Weighted	N	SD(geo)
	115/114	1.256	3	1.125
	116/114	1.958	2	1.104
	117/114	**2.789**	**3**	**1.220**

	Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Expect	Rank	Unique	115/114	116/114	117/114	Peptide
☑	120	670.8644	1339.7142	1339.6832	0.0311	0	37	0.22	1	U	1.339	2.093	3.313	R.TINSNQCVK.I
☑	144	729.9045	1457.7944	1457.8010	-0.0065	0	49	0.014	1	U	1.220	1.736	2.236	K.FLDGLYVSEK.T
☑	202	894.1127	2679.3163	2679.3271	-0.0108	0	40	0.073	1	U	0.966	-0.143	1.912	R.AVYAHFPINCVTSENNTVIEIR.N

A more serious problem is unreliable peak picking. If you look at a Mascot quantitation report for a reporter ions experiment, you may see large numbers of negative ratios. These are where the peak picking has missed a peak completely, giving a raw intensity of zero. The isotope correction then removes a little more intensity, and donates it to the adjacent peaks, so that the missing peak goes negative. Here, for example, the 116 has been missed.

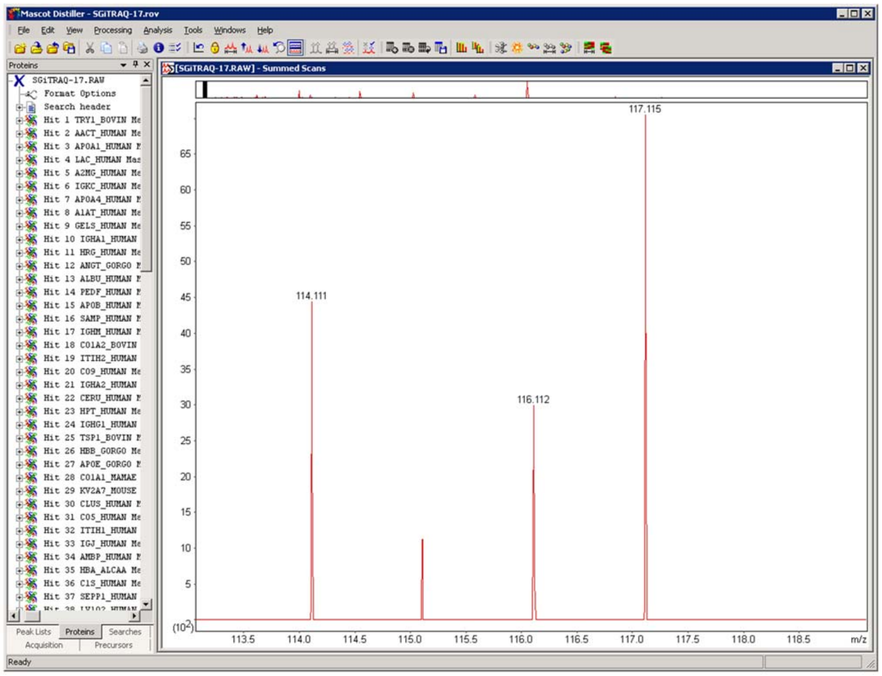
We decided not to suppress these negative ratios because they are a strong indicator that something is wrong with the peak picking. Usually, the problem is a setting that would be fine for sequence ions, such as ‘ignore peaks less than 1% of the base peak intensity’



```
SGiTRAQ-17.RAW
  Format Options
  Search header
  Hit 1 TRY1_BOVIN Me
  Hit 2 AACT_HUMAN Me
  Hit 3 APOA1_HUMAN M
  Hit 4 LAC_HUMAN Mas
  Hit 5 A2MG_HUMAN Me
  Hit 6 IGKC_HUMAN Me
  Hit 7 APOA4_HUMAN M
  Hit 8 A1AT_HUMAN Me
  Hit 9 GELS_HUMAN Me
  Hit 10 IGHA1_HUMAN
  Hit 11 HRG_HUMAN Me
  Hit 12 ANGT_GORGO M
  Hit 13 ALBU_HUMAN M
  Hit 14 PEDF_HUMAN M
  Hit 15 APOB_HUMAN M
  Hit 16 SAMP_HUMAN M
  Hit 17 IGHM_HUMAN M
  Hit 18 CO1A2_BOVIN
  Hit 19 ITIH2_HUMAN
  Hit 20 CO9_HUMAN Me
  Hit 21 IGHA2_HUMAN
  Hit 22 CERU_HUMAN M
  Hit 23 HPT_HUMAN Me
  Hit 24 IGHG1_HUMAN
  Hit 25 TSP1_BOVIN M
  Hit 26 HBB_GORGO Me
  Hit 27 APOE_GORGO M
  Hit 28 CO1A1_MAMAE
  Hit 29 KV2A7_MOUSE
  Hit 30 CLUS_HUMAN M
  Hit 31 CO5_HUMAN Me
  Hit 32 ITIH1_HUMAN
  Hit 33 IGJ_HUMAN Me
  Hit 34 AMBP_HUMAN M
  Hit 35 HBA_ALCAA Me
  Hit 36 C1S_HUMAN Me
  Hit 37 SEPP1_HUMAN
  Hit 38 LV102_HUMAN
```

MASCOT : *Causes of Inaccurate Quantitation* © 2012 Matrix Science MATRIX SCIENCE

Here, for example, the base peak has an intensity of approx 200,000 widgets. The reporter ions are relatively weak peaks, down at the bottom left

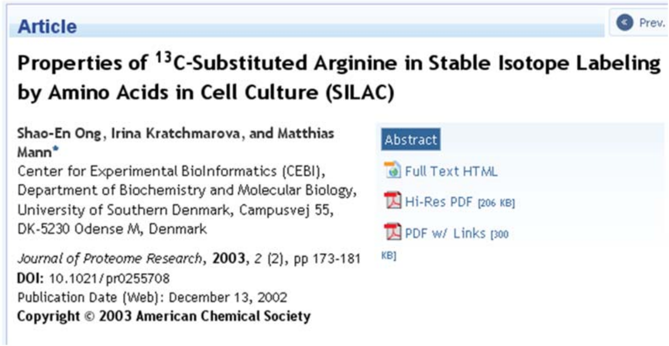


MASCOT : *Causes of Inaccurate Quantitation* © 2012 Matrix Science

MATRIX SCIENCE

If we zoom in, we can see that the 115 peak has been missed. 1% of the base peak is an intensity of 2000 and the 115 is below this. Not a problem for protein identification but a huge problem for iTRAQ quantitation. So, very important to ensure your peak picking settings are correct for these peaks

3. SILAC: Arg-Pro Conversion



Article

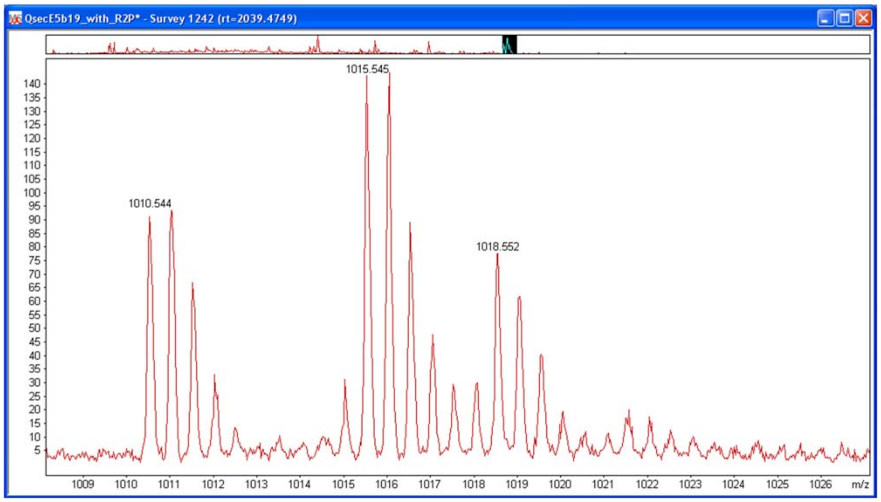
Properties of ^{13}C-Substituted Arginine in Stable Isotope Labeling by Amino Acids in Cell Culture (SILAC)

Shao-En Ong, Irina Kratchmarova, and Matthias Mann*
Center for Experimental BioInformatics (CEBI), Department of Biochemistry and Molecular Biology, University of Southern Denmark, Campusvej 55, DK-5230 Odense M, Denmark

Journal of Proteome Research, **2003**, *2* (2), pp 173-181
DOI: 10.1021/pr0255708
Publication Date (Web): December 13, 2002
Copyright © 2003 American Chemical Society

SILAC is extremely popular. Not everyone is aware of Arg-Pro conversion. Ong and colleagues reported how cells grown in media containing labelled arginine could yield peptides containing labelled proline. To obtain an accurate ratio, it becomes necessary to account for the label distributed across these additional peaks.

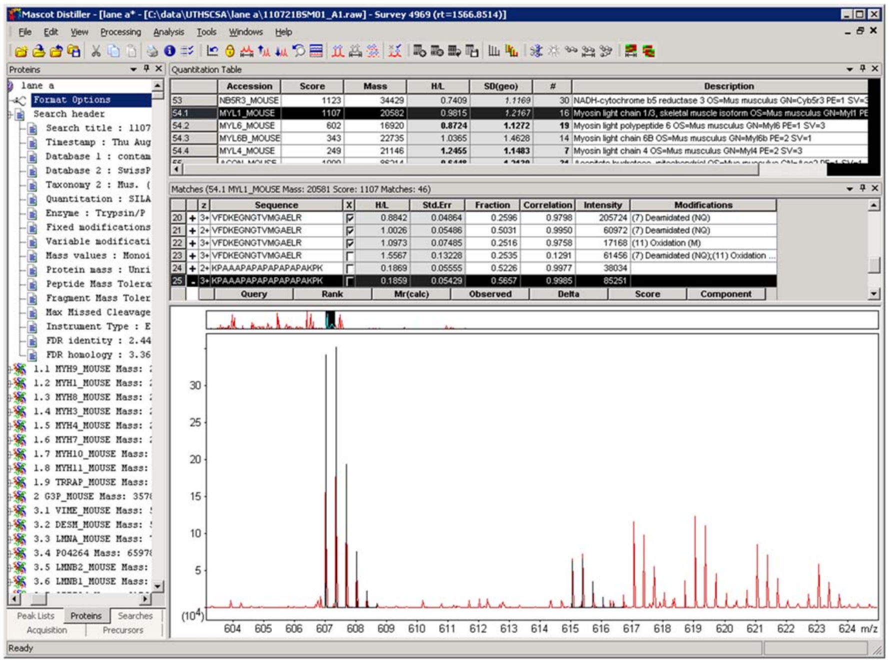
3. SILAC: Arg-Pro Conversion



MASCOT : *Causes of Inaccurate Quantitation* © 2012 Matrix Science

{MATRIX SCIENCE}

Here is an example for Arginine labelled with 13C(6)15N(4), +10. Some of the label has been incorporated as Proline The proline label is not identical to the arginine label. In this case, it is 13C(5)15N(1), +6. To get an accurate ratio, you need to sum the area of the two heavy distributions.



Mascot Distiller - [lane a* - [C:\data\UTHSCSA\lane a\110721BSM01_A1.raw] - Survey 4969 (rt=1566.8514)]

```
lane a
  Format Options
  Search header
    Search title : 1107
    Timestamp : Thu Aug
    Database 1 : contam
    Database 2 : SwissP
    Taxonomy 2 : Mus. (
    Quantitation : SILA
    Enzyme : Trypsin/P
    Fixed modifications
    Variable modificati
    Mass values : Monoi
    Protein mass : Unri
    Peptide Mass Tolera
    Fragment Mass Toler
    Max Missed Cleavage
    Instrument Type : E
    FDR identity : 2.44
    FDR homology : 3.36
  1.1 MYH9_MOUSE Mass:
  1.2 MYH1_MOUSE Mass:
  1.3 MYH8_MOUSE Mass:
  1.4 MYH3_MOUSE Mass:
  1.5 MYH4_MOUSE Mass:
  1.6 MYH7_MOUSE Mass:
  1.7 MYH10_MOUSE Mass:
  1.8 MYH11_MOUSE Mass:
  1.9 TRRAP_MOUSE Mass:
  2 G3P_MOUSE Mass: 357
  3.1 VIME_MOUSE Mass:
  3.2 DESM_MOUSE Mass:
  3.3 LMNA_MOUSE Mass:
  3.4 P04264 Mass: 6597
  3.5 LMNB2_MOUSE Mass:
  3.6 LMNB1_MOUSE Mass:
```

	Accession	Score	Mass	H/L	SD(geo)	#	Description
53	NB5R3_MOUSE	1123	34429	0.7409	*1.1169*	30	NADH-cytochrome b5 reductase 3 OS=Mus musculus GN=Cyb5r3 PE=1 SV=3
54.1	MYL1_MOUSE	1107	20582	0.9815	*1.2167*	16	Myosin light chain 1/3, skeletal muscle isoform OS=Mus musculus GN=Myl1 PE
54.2	MYL6_MOUSE	602	16920	**0.8724**	**1.1272**	**19**	Myosin light polypeptide 6 OS=Mus musculus GN=Myl6 PE=1 SV=3
54.3	MYL6B_MOUSE	343	22735	1.0365	1.4628	14	Myosin light chain 6B OS=Mus musculus GN=Myl6b PE=2 SV=1
54.4	MYL4_MOUSE	249	21146	**1.2455**	**1.1483**	**7**	Myosin light chain 4 OS=Mus musculus GN=Myl4 PE=2 SV=3

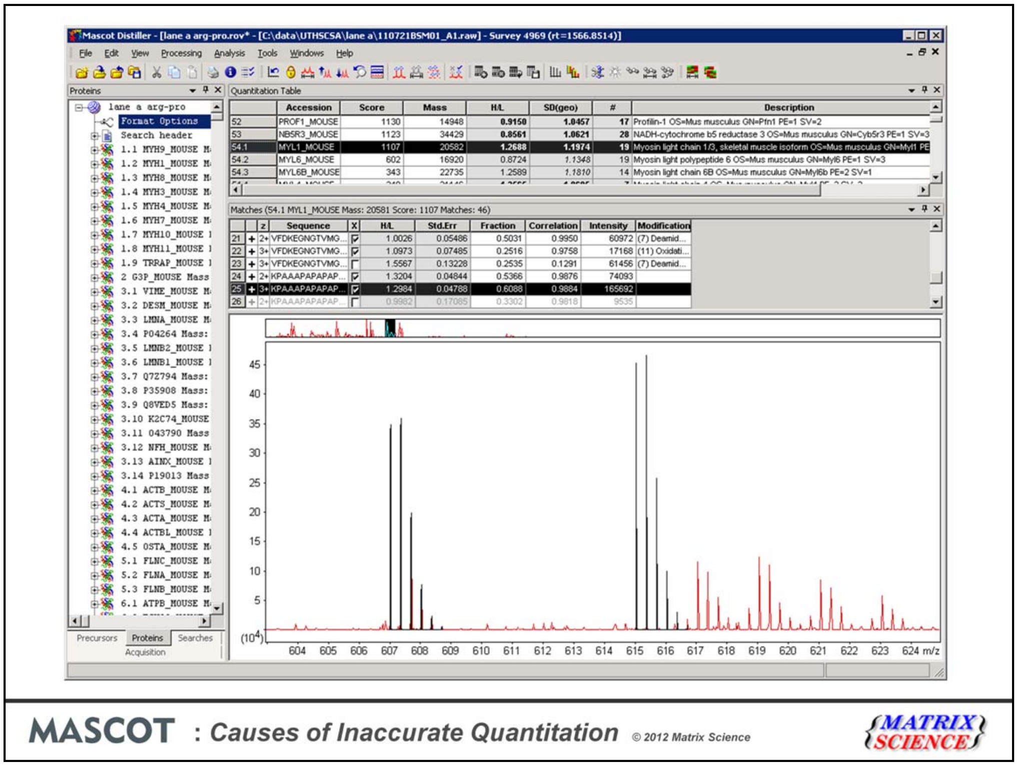
Matches (54.1 MYL1_MOUSE Mass: 20581 Score: 1107 Matches: 46)

		z	Sequence	X	H/L	Std.Err	Fraction	Correlation	Intensity	Modifications
20	+	3+	VFDKEGNGTVMGAELR	☑	0.8842	0.04864	0.2596	0.9798	205724	(7) Deamidated (NQ)
21	+	2+	VFDKEGNGTVMGAELR	☑	1.0026	0.05486	0.5031	0.9950	60972	(7) Deamidated (NQ)
22	+	3+	VFDKEGNGTVMGAELR	☑	1.0973	0.07485	0.2516	0.9758	17168	(11) Oxidation (M)
23	+	3+	VFDKEGNGTVMGAELR	☐	1.5567	0.13228	0.2535	0.1291	61456	(7) Deamidated (NQ);(11) Oxidation ...
24	+	2+	KPAAAPAPAPAPAPAPAKPK	☐	0.1869	0.05555	0.5226	0.9977	38034	
25	-	3+	KPAAAPAPAPAPAPAPAKPK	☐	0.1859	0.05429	0.5657	0.9985	85251	

Query	Rank	Mr(calc)	Observed	Delta	Score	Component

MASCOT : *Causes of Inaccurate Quantitation* © 2012 Matrix Science MATRIX SCIENCE

Not everyone sees this problem, and there are ways to minimise it. But, take a close look at your data from time to time. Here is a case where it is very strong and the ratios are seriously distorted. Without a correction, we only integrate the first heavy distribution, overlayed in black, and the ratio is 0.2 rather than 1



Mascot Distiller - [lane a arg-pro.rov* - [C:\data\UTHSCSA\lane a\110721BSM01_A1.raw] - Survey 4969 (rt=1566.8514)]

lane a arg-pro
Format Options
Search header
1.1 MYH9_MOUSE M
1.2 MYH1_MOUSE M
1.3 MYH8_MOUSE M
1.4 MYH3_MOUSE M
1.5 MYH4_MOUSE M
1.6 MYH7_MOUSE M
1.7 MYH10_MOUSE I
1.8 MYH11_MOUSE I
1.9 TRRAP_MOUSE I
2 G3P_MOUSE Mass
3.1 VIME_MOUSE M
3.2 DESM_MOUSE M
3.3 LMNA_MOUSE M
3.4 P04264 Mass:
3.5 LMNB2_MOUSE I
3.6 LMNB1_MOUSE I
3.7 Q7Z794 Mass:
3.8 P35908 Mass:
3.9 Q8VED5 Mass:
3.10 K2C74_MOUSE
3.11 O43790 Mass
3.12 NFH_MOUSE M
3.13 AINX_MOUSE I
3.14 P19013 Mass
4.1 ACTB_MOUSE M
4.2 ACTS_MOUSE M
4.3 ACTA_MOUSE M
4.4 ACTBL_MOUSE I
4.5 OSTA_MOUSE M
5.1 FLNC_MOUSE M
5.2 FLNA_MOUSE M
5.3 FLNB_MOUSE M
6.1 ATPB_MOUSE M

Quantitation Table

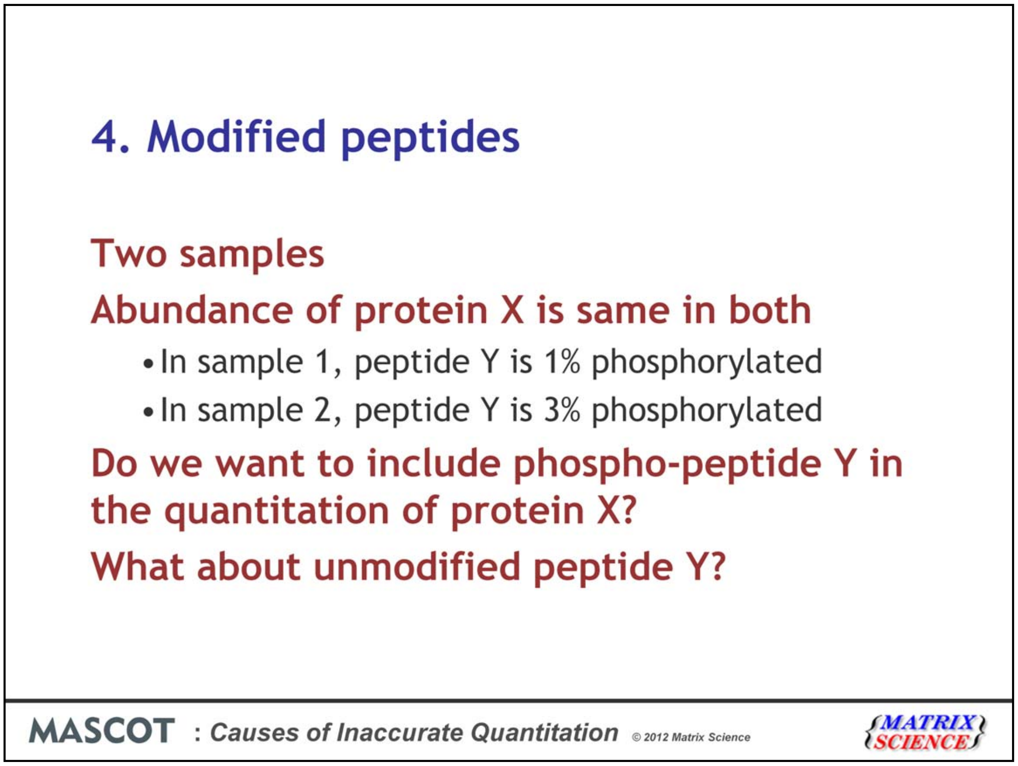
	Accession	Score	Mass	H/L	SD(geo)	#	Description
52	PROF1_MOUSE	1130	14948	**0.9150**	**1.0457**	**17**	Profilin-1 OS=Mus musculus GN=Pfn1 PE=1 SV=2
53	NB5R3_MOUSE	1123	34429	**0.8561**	**1.0621**	**28**	NADH-cytochrome b5 reductase 3 OS=Mus musculus GN=Cyb5r3 PE=1 SV=3
54.1	MYL1_MOUSE	1107	20582	**1.2688**	**1.1974**	**19**	Myosin light chain 1/3, skeletal muscle isoform OS=Mus musculus GN=Myl1 PE
54.2	MYL6_MOUSE	602	16920	0.8724	*1.1348*	19	Myosin light polypeptide 6 OS=Mus musculus GN=Myl6 PE=1 SV=3
54.3	MYL6B_MOUSE	343	22735	1.2589	*1.1810*	14	Myosin light chain 6B OS=Mus musculus GN=Myl6b PE=2 SV=1

Matches (54.1 MYL1_MOUSE Mass: 20581 Score: 1107 Matches: 46)

		z	Sequence	X	H/L	Std.Err	Fraction	Correlation	Intensity	Modification
21	+	2+	VFDKEGNGTVMG...	☑	1.0026	0.05486	0.5031	0.9950	60972	(7) Deamid...
22	+	3+	VFDKEGNGTVMG...	☑	1.0973	0.07485	0.2516	0.9758	17168	(11) Oxidati...
23	+	3+	VFDKEGNGTVMG...	□	1.5567	0.13228	0.2535	0.1291	61456	(7) Deamid...
24	+	2+	KPAAAPAPAPAP...	☑	1.3204	0.04844	0.5366	0.9876	74093	
25	+	3+	KPAAAPAPAPAP...	☑	1.2984	0.04788	0.6088	0.9884	165692	
26	+	2+	KPAAAPAPAPAP...	□	0.9982	0.17085	0.3302	0.9818	9535	

MASCOT : *Causes of Inaccurate Quantitation* © 2012 Matrix Science

With a correction, we sum all of the distributions and the ratio is closer to those of the non-proline containing peptides



4. Modified peptides

Two samples

Abundance of protein X is same in both

- In sample 1, peptide Y is 1% phosphorylated
- In sample 2, peptide Y is 3% phosphorylated

Do we want to include phospho-peptide Y in the quantitation of protein X?

What about unmodified peptide Y?

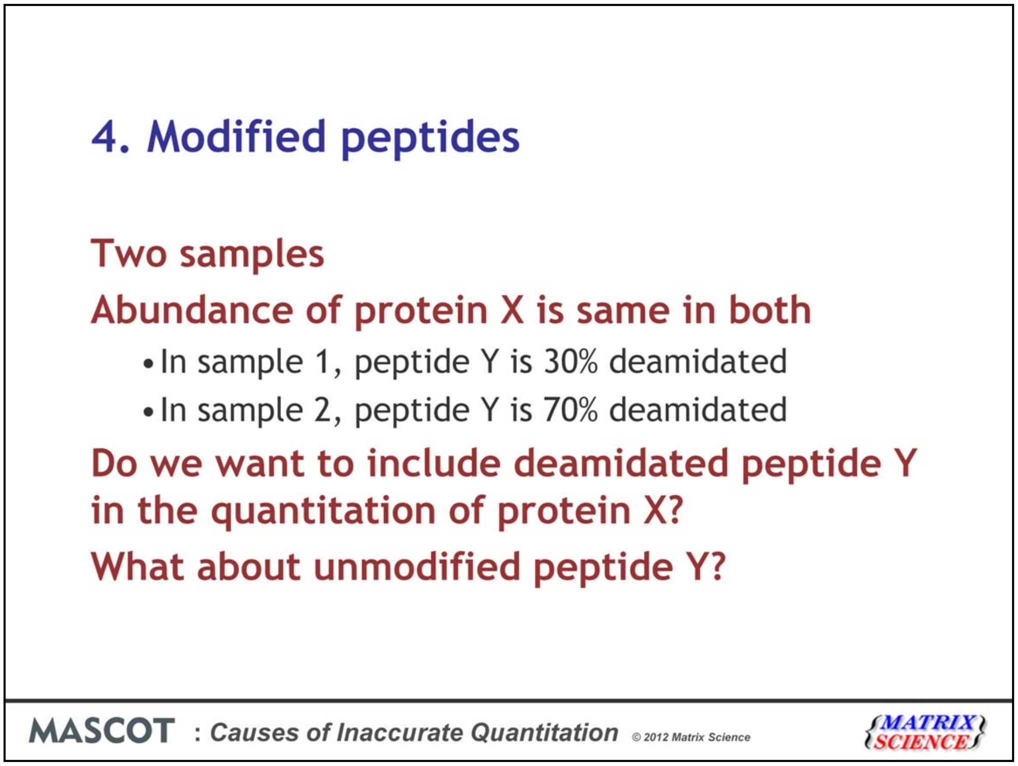
MASCOT : *Causes of Inaccurate Quantitation* © 2012 Matrix Science

MATRIX
SCIENCE

How should modified peptides be handled when we are interested in relative quantitation of proteins? I get the distinct impression that many people don’t give this a great deal of thought.

Consider this case. The abundance of the protein is the same in both samples but one of the peptides carries a low level of phosphorylation: 1% in one sample and 3% in the other. Clearly, we want to exclude this peptide because it will give us a ratio of 1:3 rather than 1:1.

Using the unmodified peptide is fine, because we’ll get a ratio of 99:97, which in most cases will be indistinguishable from 1:1



4. Modified peptides

Two samples

Abundance of protein X is same in both

- In sample 1, peptide Y is 30% deamidated
- In sample 2, peptide Y is 70% deamidated

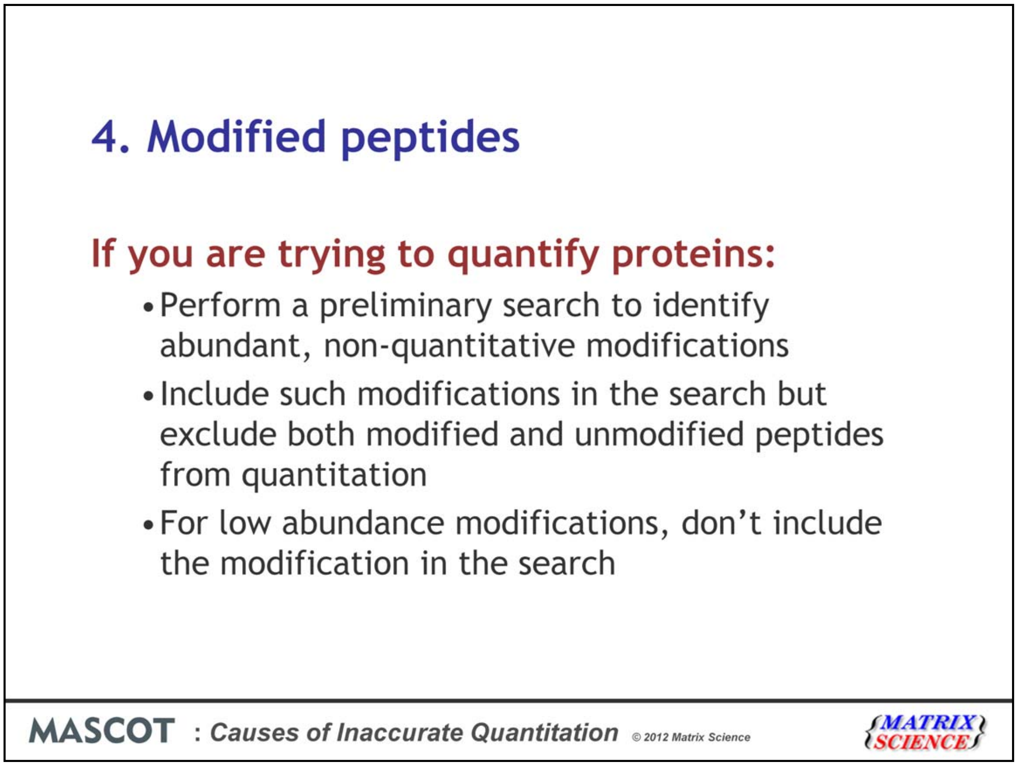
Do we want to include deamidated peptide Y in the quantitation of protein X?

What about unmodified peptide Y?

MASCOT : *Causes of Inaccurate Quantitation* © 2012 Matrix Science

{MATRIX}
{SCIENCE}

What about a peptide that is more extensively modified? Again, the abundance of the protein is the same in both samples. In one sample, a hypothetical peptide is 30% deamidated, in the other 70% deamidated. We want to exclude this peptide because it will give us a ratio of 3:7 rather than 1:1. Unlike the previous case, the unmodified peptide is no better, giving a ratio of 7:3



4. Modified peptides

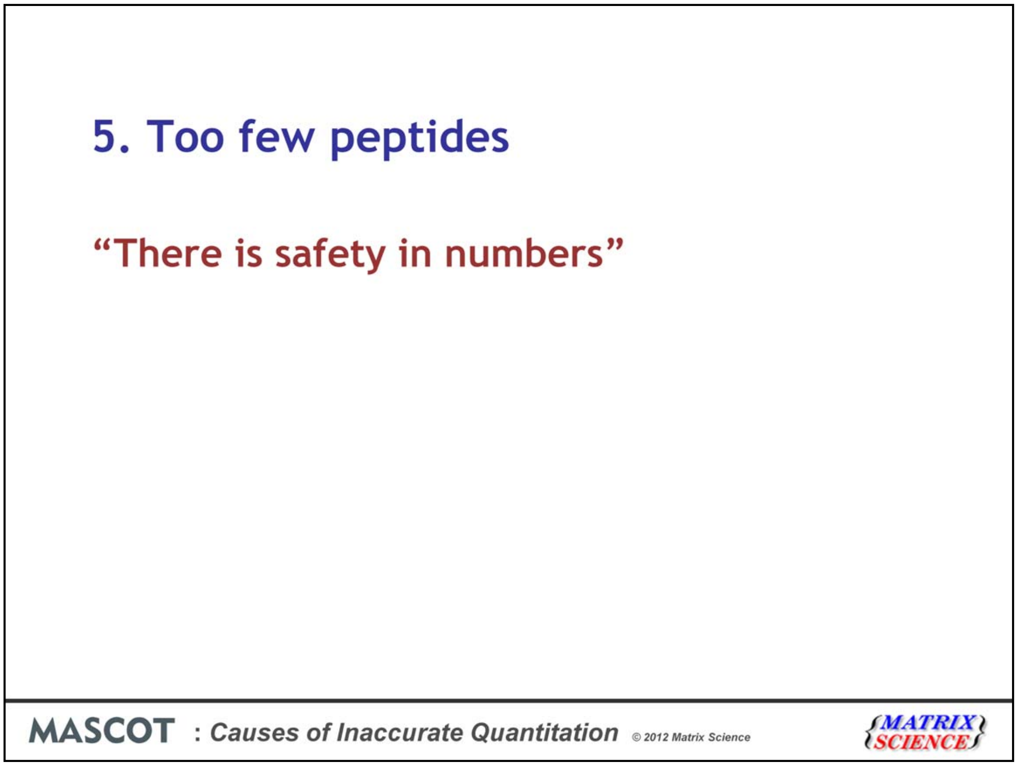
If you are trying to quantify proteins:

- Perform a preliminary search to identify abundant, non-quantitative modifications
- Include such modifications in the search but exclude both modified and unmodified peptides from quantitation
- For low abundance modifications, don't include the modification in the search

MASCOT : *Causes of Inaccurate Quantitation* © 2012 Matrix Science

{MATRIX}
{SCIENCE}

So, for relative quantitation of proteins…

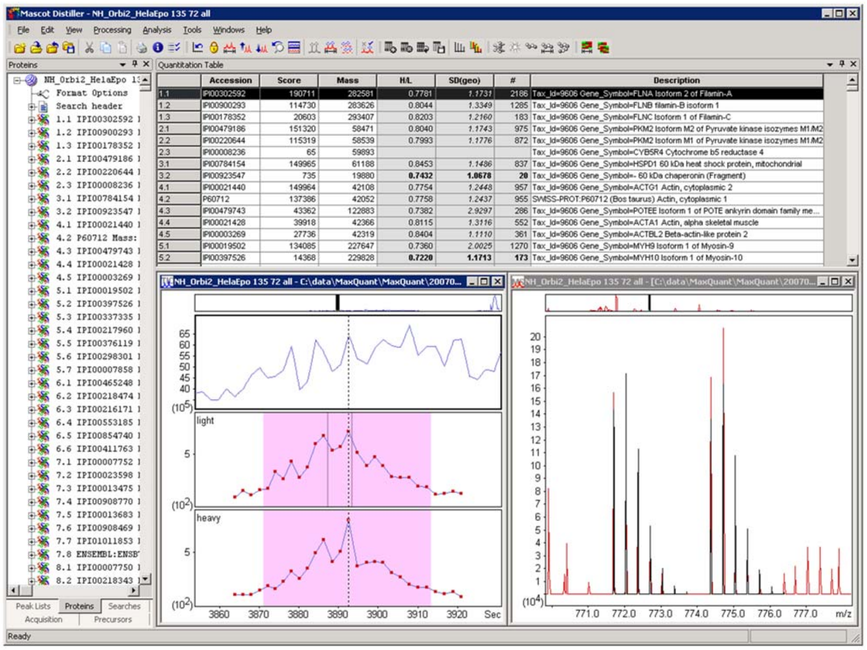


5. Too few peptides

"There is safety in numbers"

Finally, I suggest the main cause of inaccurate quantitation in a discovery experiment is having insufficient data. Peptide abundance is a surrogate for protein abundance. We assume, or rather hope, that the two are closely coupled. This only becomes a safe assumption when you look at a good sized population of peptides, and eliminate the outliers.

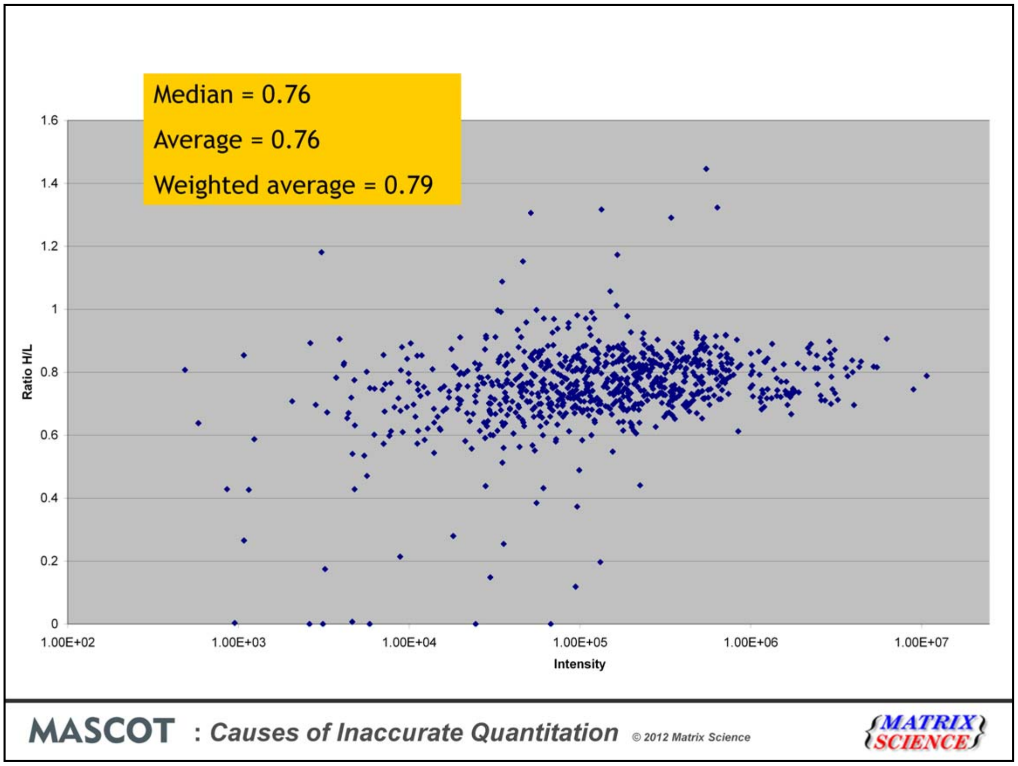
One of the main reasons for peptide abundance being different from protein abundance was just discussed: modified peptides. For post-digest labelling, another factor might be the enzyme digest conditions.


Mascot Distiller - NH_Orbi2_HelaEpo 135 72 all
File Edit View Processing Analysis Tools Windows Help
Proteins
Quantitation Table
NH_Orbi2_HelaEpo 135 72 all - C:\data\MaxQuant\MaxQuant\20070...
NH_Orbi2_HelaEpo 135 72 all - [C:\data\MaxQuant\MaxQuant\20070...
65
60
55
50
45
40
(10⁵)
light
5
(10²)
heavy
5
(10²)
3860
3870
3880
3890
3900
3910
3920
Sec
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
(10⁴)
771.0
772.0
773.0
774.0
775.0
776.0
777.0
m/z
Peak Lists
Proteins
Searches
Acquisition
Precursors
Ready

NH_Orbi2_HelaEpo 13
Format Options
Search header
1.1 IPI00302592 I
1.2 IPI00900293 I
1.3 IPI00178352 I
2.1 IPI00479186 I
2.2 IPI00220644 I
2.3 IPI00008236 I
3.1 IPI00784154 I
3.2 IPI00923547 I
4.1 IPI00021440 I
4.2 P60712 Mass:
4.3 IPI00479743 I
4.4 IPI00021428 I
4.5 IPI00003269 I
5.1 IPI00019502 I
5.2 IPI00397526 I
5.3 IPI00337335 I
5.4 IPI00217960 I
5.5 IPI00376119 I
5.6 IPI00298301 I
5.7 IPI00007858 I
6.1 IPI00465248 I
6.2 IPI00218474 I
6.3 IPI00216171 I
6.4 IPI00553185 I
6.5 IPI00854740 I
6.6 IPI00411763 I
7.1 IPI00007752 I
7.2 IPI00023598 I
7.3 IPI00013475 I
7.4 IPI00908770 I
7.5 IPI00013683 I
7.6 IPI00908469 I
7.7 IPI01011853 I
7.8 ENSEMBL:ENSB'
8.1 IPI00007750 I
8.2 IPI00218343 I

	Accession	Score	Mass	H/L	SD(geo)	#	Description
1.1	IPI00302592	190711	282581	0.7781	*1.1731*	2186	Tax_Id=9606 Gene_Symbol=FLNA Isoform 2 of Filamin-A
1.2	IPI00900293	114730	283626	0.8044	*1.3349*	1285	Tax_Id=9606 Gene_Symbol=FLNB filamin-B isoform 1
1.3	IPI00178352	20603	293407	0.8203	*1.2160*	183	Tax_Id=9606 Gene_Symbol=FLNC Isoform 1 of Filamin-C
2.1	IPI00479186	151320	58471	0.8040	*1.1743*	975	Tax_Id=9606 Gene_Symbol=PKM2 Isoform M2 of Pyruvate kinase isozymes M1/M2
2.2	IPI00220644	115319	58539	0.7993	*1.1776*	872	Tax_Id=9606 Gene_Symbol=PKM2 Isoform M1 of Pyruvate kinase isozymes M1/M2
2.3	IPI00008236	65	59893				Tax_Id=9606 Gene_Symbol=CYB5R4 Cytochrome b5 reductase 4
3.1	IPI00784154	149965	61188	0.8453	*1.1486*	837	Tax_Id=9606 Gene_Symbol=HSPD1 60 kDa heat shock protein, mitochondrial
3.2	IPI00923547	735	19880	**0.7432**	**1.0678**	**20**	Tax_Id=9606 Gene_Symbol=- 60 kDa chaperonin (Fragment)
4.1	IPI00021440	149964	42108	0.7754	*1.2448*	957	Tax_Id=9606 Gene_Symbol=ACTG1 Actin, cytoplasmic 2
4.2	P60712	137386	42052	0.7758	*1.2437*	955	SWISS-PROT:P60712 (Bos taurus) Actin, cytoplasmic 1
4.3	IPI00479743	43362	122883	0.7382	*2.9297*	286	Tax_Id=9606 Gene_Symbol=POTEE Isoform 1 of POTE ankyrin domain family me...
4.4	IPI00021428	39918	42366	0.8115	*1.3116*	552	Tax_Id=9606 Gene_Symbol=ACTA1 Actin, alpha skeletal muscle
4.5	IPI00003269	27736	42319	0.8404	*1.1110*	361	Tax_Id=9606 Gene_Symbol=ACTBL2 Beta-actin-like protein 2
5.1	IPI00019502	134085	227647	0.7360	*2.0025*	1270	Tax_Id=9606 Gene_Symbol=MYH9 Isoform 1 of Myosin-9
5.2	IPI00397526	14368	229828	**0.7220**	**1.1713**	**173**	Tax_Id=9606 Gene_Symbol=MYH10 Isoform 1 of Myosin-10

Here’s a very nice SILAC data set, containing over 4000 proteins in the minimal list and some of these have over a thousand peptide matches. High mass resolution and accuracy



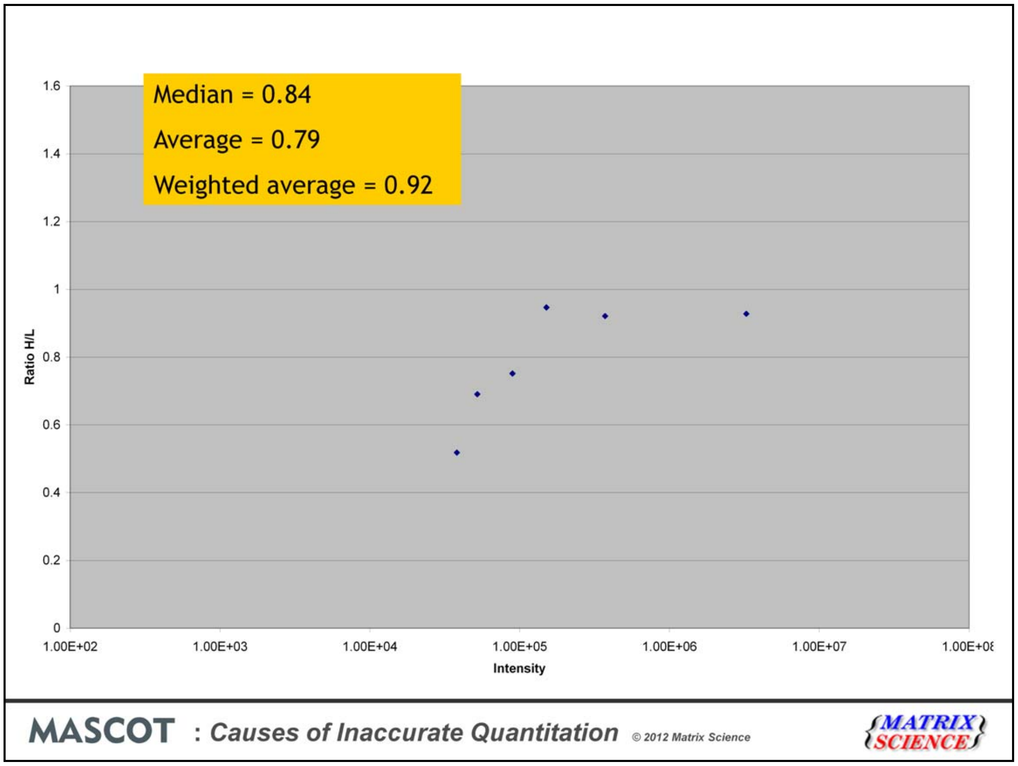
Median = 0.76

Average = 0.76

Weighted average = 0.79

MASCOT : *Causes of Inaccurate Quantitation* © 2012 Matrix Science

Here is a plot of 838 peptide ratios for a protein near the top of the list. The y axis should really be logarithmic, but a linear scale makes it easier to visualise the data. Possibly the extreme measurements are outliers caused by some failure in peak picking or chromatogram integration. Possibly they are peptides that have been misassigned. Possibly they are modified or processed in some way that makes them not representative of the abundance of the protein. In doesn’t really matter because, when you have this many measurements, you can see where the centre of gravity is. Somewhere just under 0.8, yes? It doesn’t matter whether you take the mean or the weighted mean or the median … we still get the same ratio.



Median = 0.84

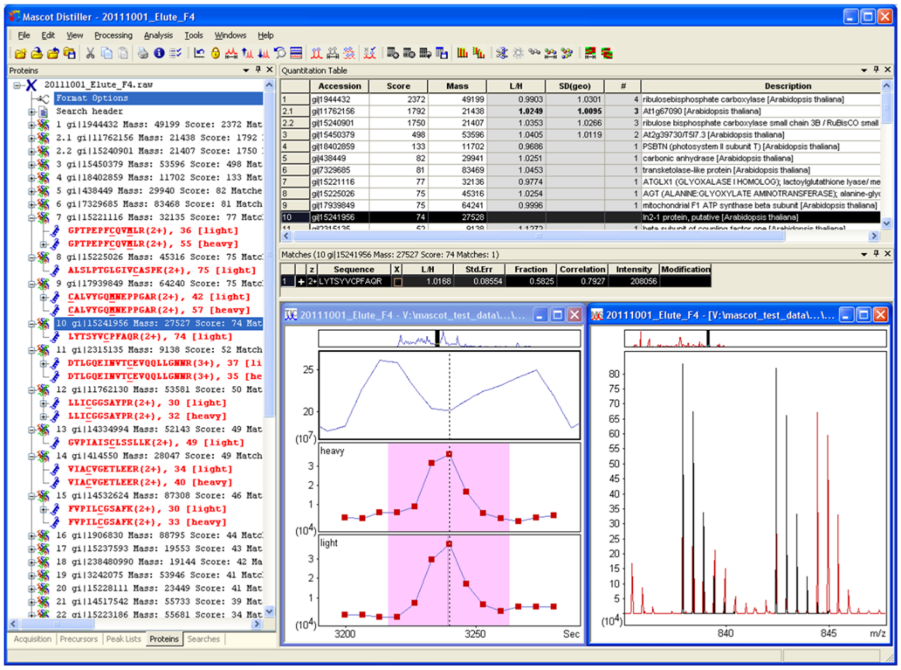
Average = 0.79

Weighted average = 0.92

MASCOT : *Causes of Inaccurate Quantitation* © 2012 Matrix Science

Go further down the list, to the low abundance proteins, where you have handful of measurements. You don’t need statistics to tell you that this is a less reliable measurement. For these 6 ratios, the average and weighted average are quite different. When people ask me which is the ‘best’ way to calculate the protein ratio from the peptide ratio, I’m tempted to reply that, if it makes much difference, you need more data.

The really dangerous situation is when these 6 peptide ratio measurements are all for the same peptide sequence, or maybe two sequences. Then it becomes a lottery.



Mascot Distiller - 20111001_Elute_F4
File Edit View Processing Analysis Tools Windows Help
Proteins
Quantitation Table
Acquisition Precursors Peak Lists Proteins Searches
20111001_Elute_F4 - V:\mascot_test_data\...\...
20111001_Elute_F4 - [V:\mascot_test_data\...\...
25
20
(10^7)
heavy
3
2
1
(10^4)
light
3
2
1
(10^4)
3200
3250
Sec
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
(10^4)
840
845
m/z

```
20111001_Elute_F4.raw
  Format Options
  Search header
  1 gi|1944432 Mass: 49199 Score: 2372 Mat
  2.1 gi|11762156 Mass: 21438 Score: 1792
  2.2 gi|15240901 Mass: 21407 Score: 1750
  3 gi|15450379 Mass: 53596 Score: 498 Mat
  4 gi|18402859 Mass: 11702 Score: 133 Mat
  5 gi|438449 Mass: 29940 Score: 82 Matche
  6 gi|7329685 Mass: 83468 Score: 81 Match
  7 gi|15221116 Mass: 32135 Score: 77 Matc
    GPTPEPFCQVMLR(2+), 36 [light]
    GPTPEPFCQVMLR(2+), 55 [heavy]
  8 gi|15225026 Mass: 45316 Score: 75 Matc
    ALSLPTGLGIVCASPK(2+), 75 [light]
  9 gi|17939849 Mass: 64240 Score: 75 Matc
    CALVYGQMNEPPGAR(2+), 42 [light]
    CALVYGQMNEPPGAR(2+), 57 [heavy]
  10 gi|15241956 Mass: 27527 Score: 74 Mat
    LYTSYVCPFAQR(2+), 74 [light]
  11 gi|2315135 Mass: 9138 Score: 52 Match
    DTLGQEINVTCEVQQLLGNNR(3+), 37 [li
    DTLGQEINVTCEVQQLLGNNR(3+), 35 [he
  12 gi|11762130 Mass: 53581 Score: 50 Mat
    LLICGGSAYPR(2+), 30 [light]
    LLICGGSAYPR(2+), 32 [heavy]
  13 gi|14334994 Mass: 52143 Score: 49 Mat
    GVPIAISCLSSLLK(2+), 49 [light]
  14 gi|414550 Mass: 28047 Score: 49 Match
    VIACVGETLEER(2+), 34 [light]
    VIACVGETLEER(2+), 40 [heavy]
  15 gi|14532624 Mass: 87308 Score: 46 Mat
    FVPILCGSAFK(2+), 30 [light]
    FVPILCGSAFK(2+), 33 [heavy]
  16 gi|1906830 Mass: 88795 Score: 44 Matc
  17 gi|15237593 Mass: 19553 Score: 43 Mat
  18 gi|238480990 Mass: 19144 Score: 42 Ma
  19 gi|3242075 Mass: 53946 Score: 41 Matc
  20 gi|15228111 Mass: 23449 Score: 41 Mat
  21 gi|14517542 Mass: 55733 Score: 39 Mat
  22 gi|15223186 Mass: 55681 Score: 34 Mat
```

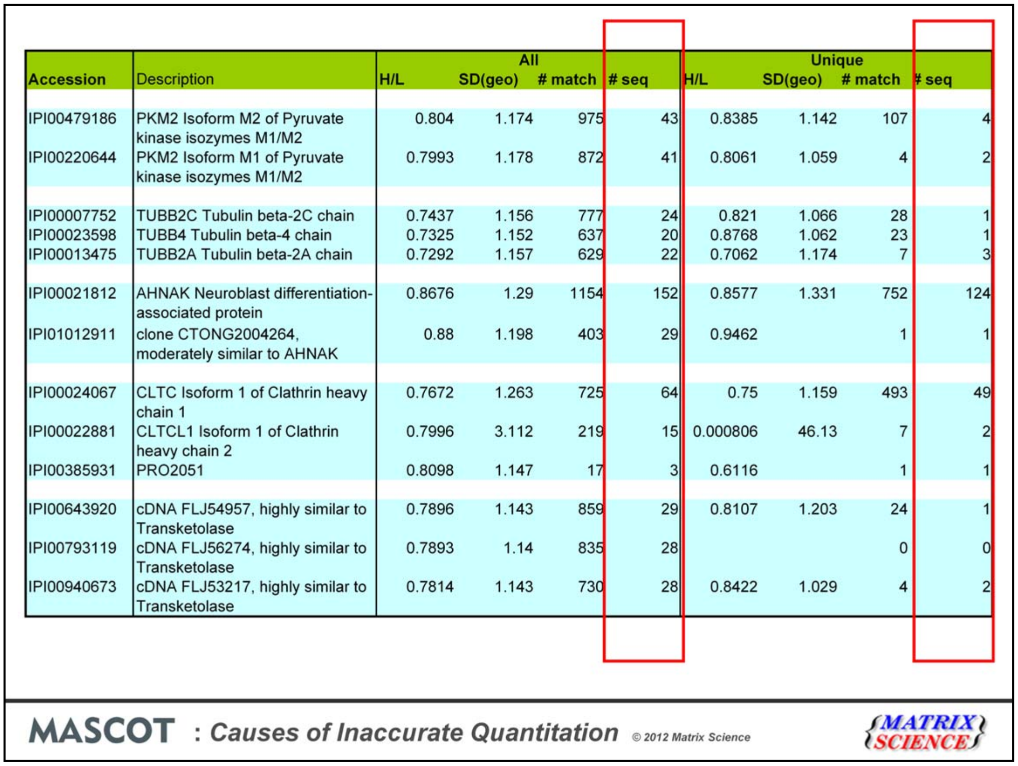
	Accession	Score	Mass	L/H	SD(geo)	#	Description
1	gi|1944432	2372	49199	0.9903	1.0301	4	ribulosebisphosphate carboxylase [Arabidopsis thaliana]
2.1	gi|11762156	1792	21438	**1.0249**	**1.0095**	**3**	At1g67090 [Arabidopsis thaliana]
2.2	gi|15240901	1750	21407	1.0353	1.0266	3	ribulose bisphosphate carboxylase small chain 3B / RuBisCO small
3	gi|15450379	498	53596	1.0405	1.0119	2	At2g39730/T5I7.3 [Arabidopsis thaliana]
4	gi|18402859	133	11702	0.9686		1	PSBTN (photosystem II subunit T) [Arabidopsis thaliana]
5	gi|438449	82	29941	1.0251		1	carbonic anhydrase [Arabidopsis thaliana]
6	gi|7329685	81	83469	1.0453		1	transketolase-like protein [Arabidopsis thaliana]
7	gi|15221116	77	32136	0.9774		1	ATGLX1 (GLYOXALASE I HOMOLOG); lactoylglutathione lyase/ me
8	gi|15225026	75	45316	1.0254		1	AGT (ALANINE:GLYOXYLATE AMINOTRANSFERASE); alanine-glyc
9	gi|17939849	75	64241	0.9996		1	mitochondrial F1 ATP synthase beta subunit [Arabidopsis thaliana]
10	gi|15241956	74	27528				ln2-1 protein, putative [Arabidopsis thaliana]
11	gi|2315135	52	9138	1.1272		1	beta subunit of coupling factor one [Arabidopsis thaliana]

Matches (10 gi|15241956 Mass: 27527 Score: 74 Matches: 1)

		z	Sequence	X	L/H	Std.Err	Fraction	Correlation	Intensity	Modification
1	+	2+	LYTSYVCPFAQR	□	1.0168	0.08554	0.5825	0.7927	208056	

For this reason, I'm not a fan of methods that focus on a small number of peptides for each protein. Methods such as ICAT or COFRADIC. The idea is to simplify the problem. But I feel they throw out the baby with the bathwater.

This is an ICAT example. The data quality is beautiful. High mass resolution. Clean and symmetrix XIC peaks. But, with only one peptide for most proteins, I simply don't feel confident that we are getting reliable protein quantitation.

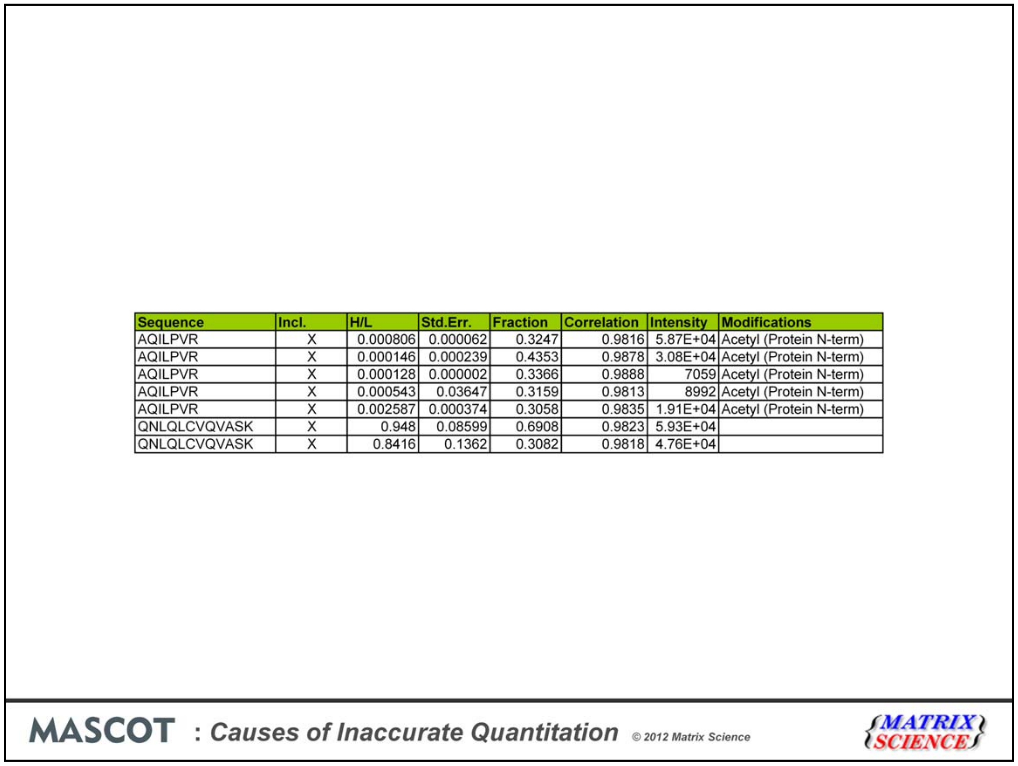


Accession	Description	All				Unique			
		H/L	SD(geo)	# match	# seq	H/L	SD(geo)	# match	# seq
IPI00479186	PKM2 Isoform M2 of Pyruvate kinase isozymes M1/M2	0.804	1.174	975	43	0.8385	1.142	107	4
IPI00220644	PKM2 Isoform M1 of Pyruvate kinase isozymes M1/M2	0.7993	1.178	872	41	0.8061	1.059	4	2
IPI00007752	TUBB2C Tubulin beta-2C chain	0.7437	1.156	777	24	0.821	1.066	28	1
IPI00023598	TUBB4 Tubulin beta-4 chain	0.7325	1.152	637	20	0.8768	1.062	23	1
IPI00013475	TUBB2A Tubulin beta-2A chain	0.7292	1.157	629	22	0.7062	1.174	7	3
IPI00021812	AHNAK Neuroblast differentiation-associated protein	0.8676	1.29	1154	152	0.8577	1.331	752	124
IPI01012911	clone CTONG2004264, moderately similar to AHNAK	0.88	1.198	403	29	0.9462		1	1
IPI00024067	CLTC Isoform 1 of Clathrin heavy chain 1	0.7672	1.263	725	64	0.75	1.159	493	49
IPI00022881	CLTCL1 Isoform 1 of Clathrin heavy chain 2	0.7996	3.112	219	15	0.000806	46.13	7	2
IPI00385931	PRO2051	0.8098	1.147	17	3	0.6116		1	1
IPI00643920	cDNA FLJ54957, highly similar to Transketolase	0.7896	1.143	859	29	0.8107	1.203	24	1
IPI00793119	cDNA FLJ56274, highly similar to Transketolase	0.7893	1.14	835	28			0	0
IPI00940673	cDNA FLJ53217, highly similar to Transketolase	0.7814	1.143	730	28	0.8422	1.029	4	2

MASCOT : *Causes of Inaccurate Quantitation* © 2012 Matrix Science

A related question is whether it’s a good idea to restrict quantitation to “unique peptide”. That is, peptides that are not shared with other proteins. Here are a few examples from a large SILAC data set. There is solid evidence for the presence of all of these proteins from high scoring, unique peptides. But, when we look at the peptides that are quantified, a very high proportion are shared between isoforms. For example, these tubulins. Each have some 6 or 7 hundred matches to twenty odd distinct sequences. But, almost all of these are shared. When eliminated, we end up with just 1 distinct sequence each for two of the tubulins. Too few for any kind of reliable measurement.

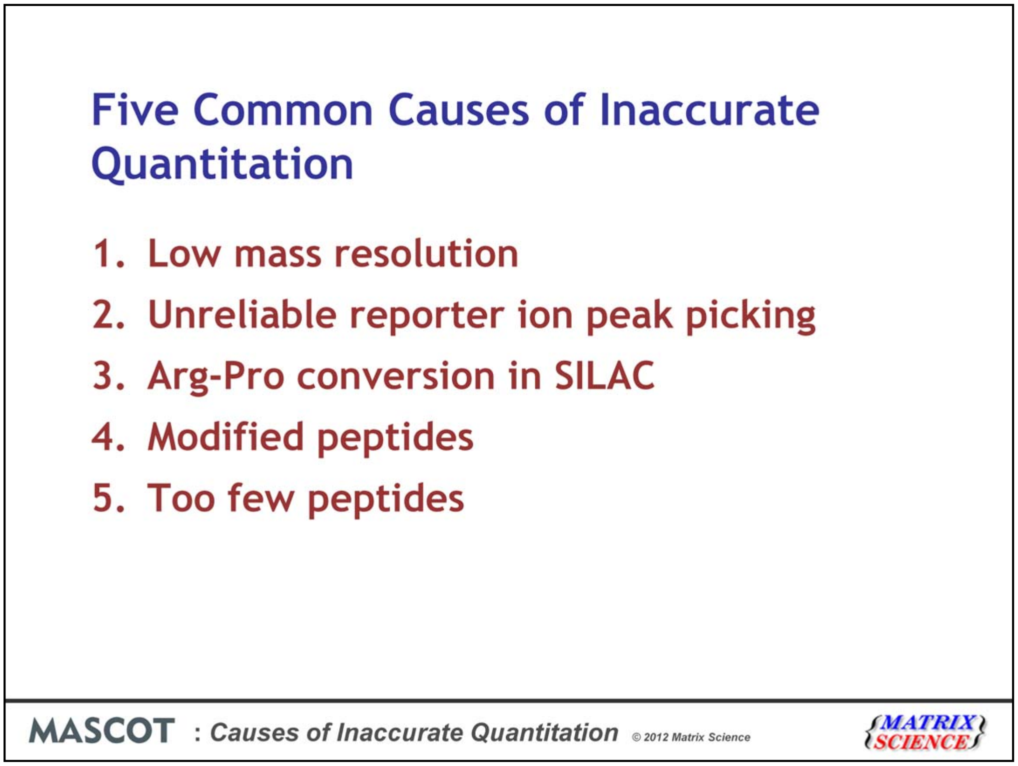
Does removal of the shared matches reveal any up or down regulation? You may think you see one here, the H/L for this Clathrin goes from 0.8 to near zero. However, note that it is down to only 7 matches to two distinct sequences. If we look at what these are



Sequence	Incl.	H/L	Std.Err.	Fraction	Correlation	Intensity	Modifications
AQILPVR	X	0.000806	0.000062	0.3247	0.9816	5.87E+04	Acetyl (Protein N-term)
AQILPVR	X	0.000146	0.000239	0.4353	0.9878	3.08E+04	Acetyl (Protein N-term)
AQILPVR	X	0.000128	0.000002	0.3366	0.9888	7059	Acetyl (Protein N-term)
AQILPVR	X	0.000543	0.03647	0.3159	0.9813	8992	Acetyl (Protein N-term)
AQILPVR	X	0.002587	0.000374	0.3058	0.9835	1.91E+04	Acetyl (Protein N-term)
QNLQLCVQVASK	X	0.948	0.08599	0.6908	0.9823	5.93E+04	
QNLQLCVQVASK	X	0.8416	0.1362	0.3082	0.9818	4.76E+04	

Matches to one of the sequences are both in the 0.8 ballpark. The other sequence is post-translationally modified, which makes it unreliable for quantitation of the protein.

It doesn’t always make sense to limit discovery quantitation to unique peptides. Maybe better to study cases where the variance of the measurements is larger than expected and see whether there is evidence for the peptides belonging to two populations



Five Common Causes of Inaccurate Quantitation

1. Low mass resolution
2. Unreliable reporter ion peak picking
3. Arg-Pro conversion in SILAC
4. Modified peptides
5. Too few peptides

MASCOT : *Causes of Inaccurate Quantitation* © 2012 Matrix Science

MATRIX SCIENCE

To summarise