Standards for Proteomics Data Generated by LC-MS-MS

Ruedi Aebersold, Ph.D.
ETH Zurich, Switzerland and
Institute for Systems Biology
Seattle, Washington

Theses:

- Different requirements for data processing, dissemination and storage apply for mass spectrometry applied to the analysis of proteins and proteomes.
- Proteomics is a genomic science and needs to develop “genomics” data analysis/dissemination strategies

LC-MS/MS as a protein analysis tool

- Relatively low number of proteins analyzed per experiment
- Extensive (biological, manual) validation of data
- Projects centered in single group and focused on specific question
- Data stored in notebook or local computer
- Reports focused on the biological meaning of the data

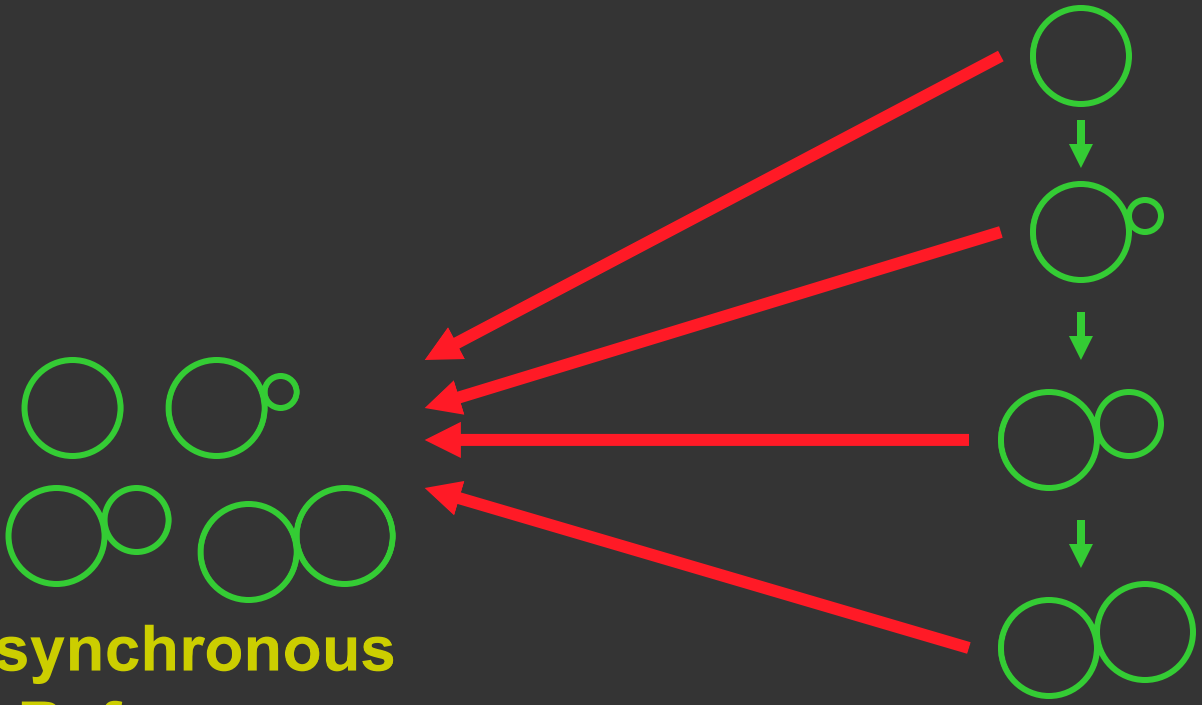
LC-MS/MS as a genomic technology

- Many – ideally, all – proteins in a proteome analyzed repeatedly
- Extensive and consistent biological or manual validation of all data impossible
- Value of information increases if data from multiple experiments/groups can be integrated and collectively mined
- Proteomics is a community effort
- Data are collected and organized in relational databases
- Whole data sets should be made accessible/published

Discussion Points

Many – ideally, all – proteins in a proteome analyzed repeatedly, generating large volumes of data

Synchronous Timepoint Samples Compared to Reference Sample



Asynchronous Reference Sample

Timepoint Samples from Yeast Cells Synchronously Transiting the Cell Cycle

Data Summary

	T0	T30	T60	T90	T120
T0	1648	1095	1184	1112	892
T30		1523	1055	1140	921
T60			1448	1051	871
T90				1713	960
T120					1229

- 2735/6562 proteins quantified across all timepoints (42%)
- 696 proteins quantified in every experiment
- 1513 proteins quantified in at least one timepoint
- 34,400 peptides quantified on average per timepoint
- >1 million mass spectra collected

Discussion Points

Many – ideally, all – proteins in a proteome analyzed repeatedly, generating large volumes of data

Current status:

- Large volumes of data are being generated to identify relatively small numbers of proteins
- Information from prior experiments is not used, making the process relatively inefficient

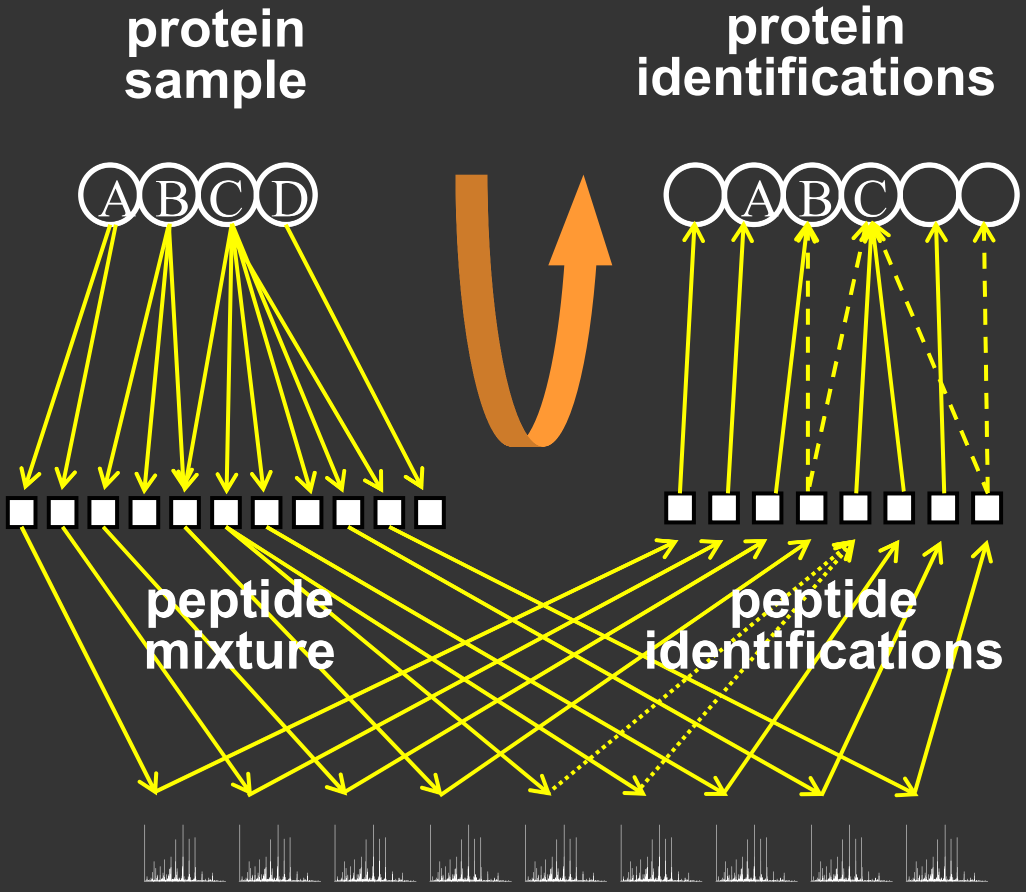
Recommendations:

- Improved strategies for more efficient data collection and analysis are required
- To develop those, access to data is essential

Discussion Points

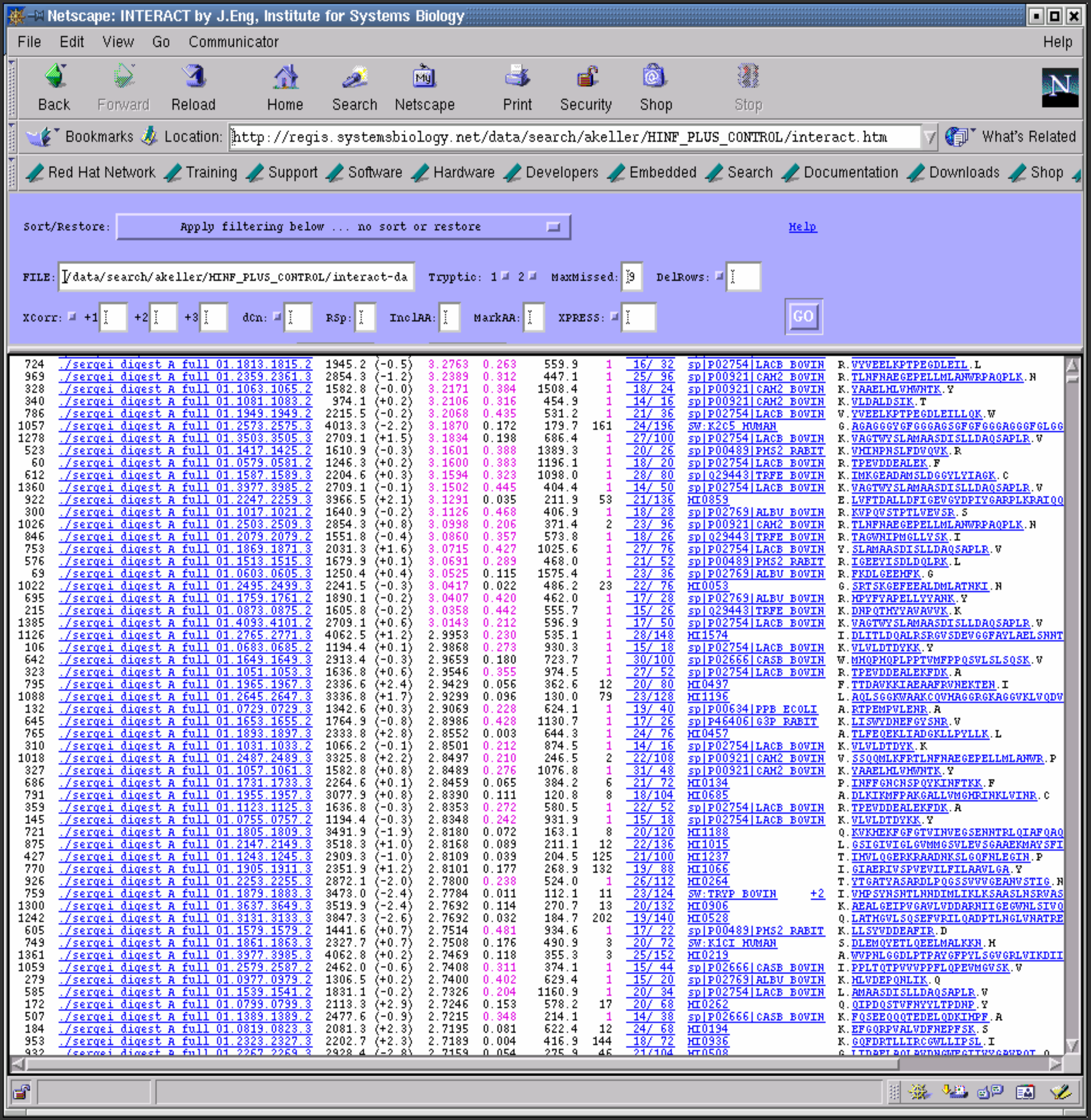
Extensive biological and/or manual validation of all data impossible

Protein Identification by MS/MS



MS/MS spectra

Output from search algorithm


Netscape: INTERACT by J.Eng, Institute for Systems Biology
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location: What's Related
Red Hat Network Training Support Software Hardware Developers Embedded Search Documentation Downloads Shop
GO

Location: http://regis.systemsbiology.net/data/search/akeller/HINF_PLUS_CONTROL/interact.htm

Sort/Restore: Apply filtering below ... no sort or restore   Help

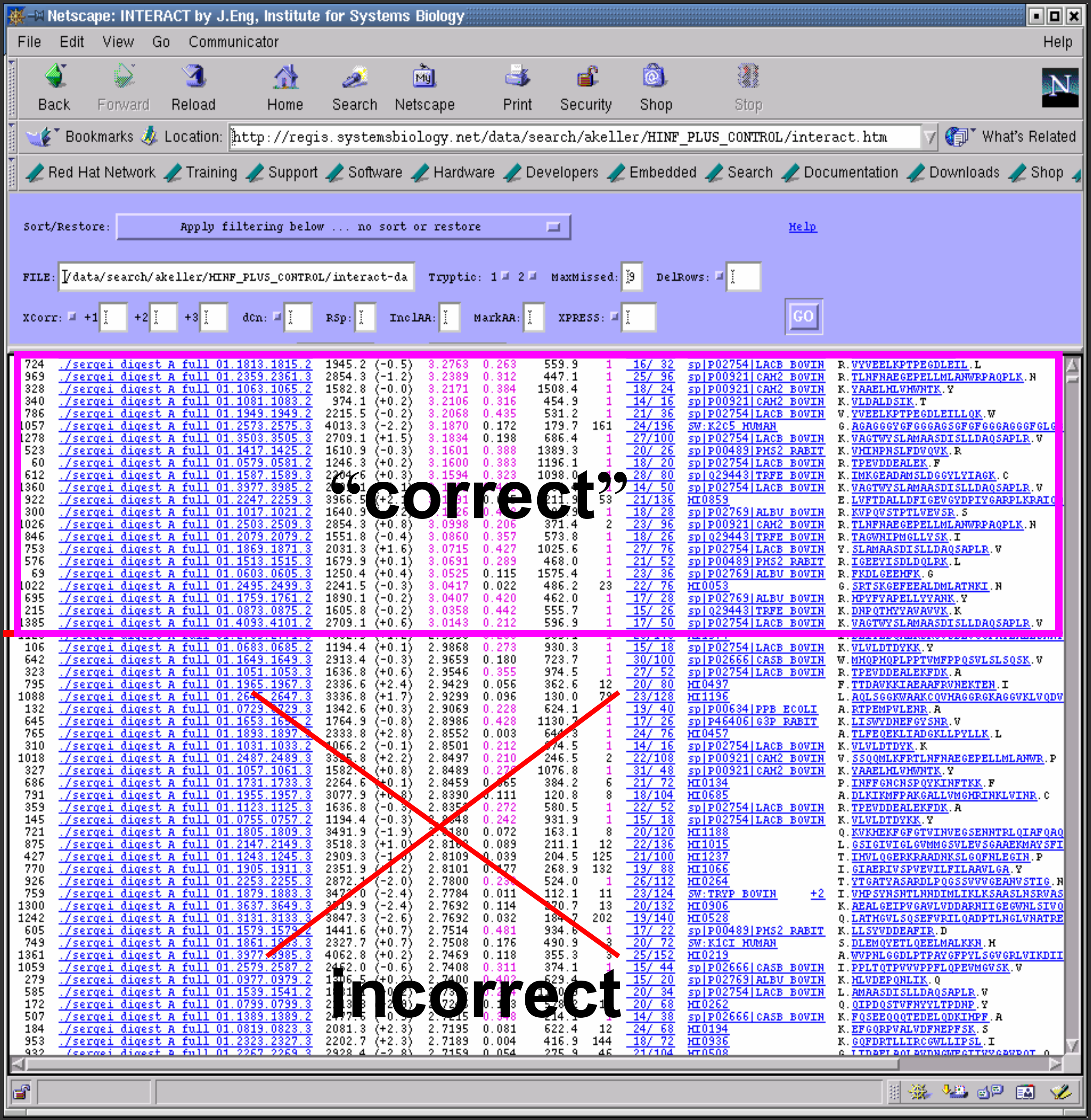
FILE: /data/search/akeller/HINF_PLUS_CONTROL/interact-da   Tryptic: 1 2   MaxMissed: 9   DelRows:

XCorr: +1 +2 +3   dCn: RSp: InclAA: MarkAA: XPRESS:

724	./sergei digest A full 01.1813.1815.2	1945.2	(-0.5)	3.2763	0.263	559.9	1	16/ 32	sp|P02754|LACB BOVIN	R.VYVEELKPTPEGDLEIL.L
969	./sergei digest A full 01.2359.2361.3	2854.3	(-1.2)	3.2389	0.312	447.1	1	25/ 96	sp|P00921|CAH2 BOVIN	R.TLNFNAEGEPELLMLANWRPAQPLK.N
328	./sergei digest A full 01.1063.1065.2	1582.8	(-0.0)	3.2171	0.384	1508.4	1	18/ 24	sp|P00921|CAH2 BOVIN	K.YAAELHLVHWNTK.Y
340	./sergei digest A full 01.1081.1083.2	974.1	(+0.2)	3.2106	0.316	454.9	1	14/ 16	sp|P00921|CAH2 BOVIN	K.VLDALDSIK.T
786	./sergei digest A full 01.1949.1949.2	2215.5	(-0.2)	3.2068	0.435	531.2	1	21/ 36	sp|P02754|LACB BOVIN	V.YVEELKPTPEGDLEILLQK.W
1057	./sergei digest A full 01.2573.2575.3	4013.3	(-2.2)	3.1870	0.172	179.7	161	24/196	SW:K2C5 HUMAN	G.AGAGGGYGFGGGAGSGFGFGGGAGGGFGLGG
1278	./sergei digest A full 01.3503.3505.3	2709.1	(+1.5)	3.1834	0.198	686.4	1	27/100	sp|P02754|LACB BOVIN	K.VAGTWYSLAMAASDISLLDAQSAPLR.V
523	./sergei digest A full 01.1417.1425.2	1610.9	(-0.3)	3.1601	0.388	1389.3	1	20/ 26	sp|P00489|PHS2 RABIT	K.VHINPNSLFDVQVK.R
60	./sergei digest A full 01.0579.0581.2	1246.3	(+0.2)	3.1600	0.383	1196.1	1	18/ 20	sp|P02754|LACB BOVIN	R.TPEVDDEALEK.F
612	./sergei digest A full 01.1587.1589.3	2204.6	(+0.3)	3.1594	0.323	1098.0	1	28/ 80	sp|Q29443|TRFE BOVIN	K.IMKGEADAMSLDGGYLYIAGK.C
1360	./sergei digest A full 01.3977.3985.2	2709.1	(-0.1)	3.1502	0.445	404.4	1	14/ 50	sp|P02754|LACB BOVIN	K.VAGTWYSLAMAASDISLLDAQSAPLR.V
922	./sergei digest A full 01.2247.2259.3	3966.5	(+2.1)	3.1291	0.035	211.9	53	21/136	HI0859	E.LVFTDALLDFIGEVGYDPIYGARPLKRAIQQ
300	./sergei digest A full 01.1017.1021.2	1640.9	(-0.2)	3.1126	0.468	406.9	1	18/ 28	sp|P02769|ALBU BOVIN	R.KVPQVSTPTLVEVSR.S
1026	./sergei digest A full 01.2503.2509.3	2854.3	(+0.8)	3.0998	0.206	371.4	2	23/ 96	sp|P00921|CAH2 BOVIN	R.TLNFNAEGEPELLMLANWRPAQPLK.N
846	./sergei digest A full 01.2079.2079.2	1551.8	(-0.4)	3.0860	0.357	573.8	1	18/ 26	sp|Q29443|TRFE BOVIN	R.TAGWNIPMGLLYSK.I
753	./sergei digest A full 01.1869.1871.3	2031.3	(+1.6)	3.0715	0.427	1025.6	1	27/ 76	sp|P02754|LACB BOVIN	Y.SLAMAASDISLLDAQSAPLR.V
576	./sergei digest A full 01.1513.1515.3	1679.9	(+0.1)	3.0691	0.289	468.0	1	21/ 52	sp|P00489|PHS2 RABIT	R.IGEEYISDLDQLRK.L
69	./sergei digest A full 01.0603.0605.3	1250.4	(+0.4)	3.0525	0.115	1575.4	1	23/ 36	sp|P02769|ALBU BOVIN	R.FKDLGEEHFK.G
1022	./sergei digest A full 01.2495.2499.3	2241.5	(-0.3)	3.0417	0.022	486.2	23	22/ 76	HI0053	G.SRTSKGEFEEALDMLATNKI.N
695	./sergei digest A full 01.1759.1761.2	1890.1	(-0.2)	3.0407	0.420	462.0	1	17/ 28	sp|P02769|ALBU BOVIN	R.HPYFYAPELLYYANK.Y
215	./sergei digest A full 01.0873.0875.2	1605.8	(-0.2)	3.0358	0.442	555.7	1	15/ 26	sp|Q29443|TRFE BOVIN	K.DNPQTHYYAVAVVK.K
1385	./sergei digest A full 01.4093.4101.2	2709.1	(+0.6)	3.0143	0.212	596.9	1	17/ 50	sp|P02754|LACB BOVIN	K.VAGTWYSLAMAASDISLLDAQSAPLR.V
1126	./sergei digest A full 01.2765.2771.3	4062.5	(+1.2)	2.9953	0.230	535.1	1	28/148	HI1574	I.DLITLDQALRSRGVSDEVGGFAYLAELSNNT
106	./sergei digest A full 01.0683.0685.2	1194.4	(+0.1)	2.9868	0.273	930.3	1	15/ 18	sp|P02754|LACB BOVIN	K.VLVLDTDYKK.Y
642	./sergei digest A full 01.1649.1649.3	2913.4	(-0.3)	2.9659	0.180	723.7	1	30/100	sp|P02666|CASB BOVIN	W.MHQPHQPLPPTVMFPPQSVLSLSQSK.V
323	./sergei digest A full 01.1051.1053.3	1636.8	(+0.6)	2.9546	0.355	974.5	1	27/ 52	sp|P02754|LACB BOVIN	R.TPEVDDEALEKFDK.A
795	./sergei digest A full 01.1965.1967.3	2336.6	(+2.4)	2.9429	0.056	362.6	12	20/ 80	HI0497	F.TTDAVKKIAEAAFRVNEKTEN.I
1088	./sergei digest A full 01.2645.2647.3	3336.8	(+1.7)	2.9299	0.096	130.0	79	23/128	HI1196	L.AQLSGGKWAAKCQVHAGGRGKAGGVKLVQDV
132	./sergei digest A full 01.0729.0729.3	1342.6	(+0.3)	2.9069	0.228	624.1	1	19/ 40	sp|P00634|PPB ECOLI	A.RTPEMPVLENR.A
645	./sergei digest A full 01.1653.1655.2	1764.9	(-0.8)	2.8986	0.428	1130.7	1	17/ 26	sp|P46406|G3P RABIT	K.LISWYDNEFGYSNR.V
765	./sergei digest A full 01.1893.1897.3	2333.8	(+2.8)	2.8552	0.003	644.3	1	24/ 76	HI0457	A.TLFEQEKLIADGKLLPYLLK.L
310	./sergei digest A full 01.1031.1033.2	1066.2	(-0.1)	2.8501	0.212	874.5	1	14/ 16	sp|P02754|LACB BOVIN	K.VLVLDTDYK.K
1018	./sergei digest A full 01.2487.2489.3	3325.8	(+2.2)	2.8497	0.210	246.5	2	22/108	sp|P00921|CAH2 BOVIN	V.SSQQMLKFRTLNFNAEGEPELLMLANWR.P
327	./sergei digest A full 01.1057.1061.3	1582.8	(+0.8)	2.8489	0.276	1076.8	1	31/ 48	sp|P00921|CAH2 BOVIN	K.YAAELHLVHWNTK.Y
686	./sergei digest A full 01.1731.1733.3	2264.6	(+0.1)	2.8459	0.065	384.2	6	21/ 72	HI0134	P.INFFGNCNSPQYKINFTKK.F
791	./sergei digest A full 01.1955.1957.3	3077.9	(+0.8)	2.8390	0.111	120.8	8	18/104	HI0685	A.DLKIKMFPAKGALLVMGHRINKLVINR.C
359	./sergei digest A full 01.1123.1125.3	1636.8	(-0.3)	2.8353	0.272	580.5	1	22/ 52	sp|P02754|LACB BOVIN	R.TPEVDDEALEKFDK.A
145	./sergei digest A full 01.0755.0757.2	1194.4	(-0.3)	2.8348	0.242	931.9	1	15/ 18	sp|P02754|LACB BOVIN	K.VLVLDTDYKK.Y
721	./sergei digest A full 01.1805.1809.3	3491.9	(-1.9)	2.8180	0.072	163.1	8	20/120	HI1188	Q.KVKHEKFGFGTVINVEGSENNTRLQIAFQAQ
875	./sergei digest A full 01.2147.2149.3	3518.3	(+1.0)	2.8168	0.089	211.1	12	22/136	HI1015	L.GSIGIVIGLGVMMGSVLEVSGAAEKMAYSFI
427	./sergei digest A full 01.1243.1245.3	2909.3	(-1.0)	2.8109	0.039	204.5	125	21/100	HI1237	T.IHVLQGERKRAADNKSLGQFNLEGIN.P
770	./sergei digest A full 01.1905.1911.3	2351.9	(+1.2)	2.8101	0.177	268.9	132	19/ 88	HI1066	I.GIAERIVSPVEVILFILAAVLGA.Y
926	./sergei digest A full 01.2253.2255.3	2872.1	(-2.0)	2.7800	0.238	524.0	1	26/112	HI0264	T.YTGATYASARDLPQGSSVVVGEANVSTIG.N
759	./sergei digest A full 01.1879.1883.3	3473.0	(-2.4)	2.7784	0.011	112.1	11	23/124	SW:TRYP BOVIN +2	I.VHPSYNSNTLNNDIMLIKLKSAASLNSRVAS
1300	./sergei digest A full 01.3637.3649.3	3519.9	(-2.4)	2.7692	0.114	270.7	13	20/132	HI0906	K.AEALGEIPVGAVLVDDARNIIGEGWNLSIVQ
1242	./sergei digest A full 01.3131.3133.3	3847.3	(-2.6)	2.7692	0.032	184.7	202	19/140	HI0528	Q.LATHGVLSQSEFVRILQADPTLNGLVNATRE
605	./sergei digest A full 01.1579.1579.2	1441.6	(+0.7)	2.7514	0.481	934.6	1	17/ 22	sp|P00489|PHS2 RABIT	K.LLSYVDDEAFIR.D
749	./sergei digest A full 01.1861.1863.3	2327.7	(+0.7)	2.7508	0.176	490.9	3	20/ 72	SW:K1CI HUMAN	S.DLEMQYETLQEELMALKKN.H
1361	./sergei digest A full 01.3977.3985.3	4062.8	(+0.2)	2.7469	0.118	355.3	3	25/152	HI0219	A.WVPNLGGDLPTPAYGFPYLSGVGRLVIKDII
1059	./sergei digest A full 01.2579.2587.2	2462.0	(-0.6)	2.7408	0.311	374.1	1	15/ 44	sp|P02666|CASB BOVIN	I.PPLTQTPVVVPPFLQPEVMGVSK.V
279	./sergei digest A full 01.0977.0979.2	1306.5	(+0.2)	2.7400	0.402	629.4	1	15/ 20	sp|P02769|ALBU BOVIN	K.HLVDEPQNLIK.Q
585	./sergei digest A full 01.1539.1541.2	1831.1	(-0.2)	2.7326	0.204	1160.9	1	20/ 34	sp|P02754|LACB BOVIN	L.AMAASDISLLDAQSAPLR.V
172	./sergei digest A full 01.0799.0799.3	2113.3	(+2.9)	2.7246	0.153	578.2	17	20/ 68	HI0262	Q.QIPDQSTVFNYYLTPDNP.Y
507	./sergei digest A full 01.1389.1389.2	2477.6	(-0.9)	2.7215	0.348	214.1	1	14/ 38	sp|P02666|CASB BOVIN	K.FQSEEQQQTEDELQDKIHPF.A
184	./sergei digest A full 01.0819.0823.3	2081.3	(+2.3)	2.7195	0.081	622.4	12	24/ 68	HI0194	K.EFGQRPVALVDFNEPFSK.S
953	./sergei digest A full 01.2323.2327.3	2202.7	(+2.3)	2.7189	0.004	416.9	144	18/ 72	HI0936	K.GQFDRTLLIRCGWLLIPSL.I
932	./sergei digest A full 01.2267.2269.3	2928.4	(-2.8)	2.7159	0.054	275.9	46	21/104	HI0508	G.LTDAELAQLAVDNGWEGTTVYGAVRQT.Q

sort by search score

Threshold Model



Location: http://regis.systemsbiology.net/data/search/akeller/HINF_PLUS_CONTROL/interact.htm

Sort/Restore: Apply filtering below ... no sort or restore

FILE: /data/search/akeller/HINF_PLUS_CONTROL/interact-da Tryptic: 1 2 MaxMissed: 9 DelRows:

XCorr: +1 +2 +3 dCn: RSp: InclAA: MarkAA: XPRESS:

724	./sergei digest A full 01.1813.1815.2	1945.2	(-0.5)	3.2763	0.263	559.9	1	16/ 32	sp|P02754|LACB BOVIN	R.VYVEELKPTPEGDLEIL.L
969	./sergei digest A full 01.2359.2361.3	2854.3	(-1.2)	3.2389	0.312	447.1	1	25/ 96	sp|P00921|CAH2 BOVIN	R.TLNFNAEGEPELLMLANWRPAQPLK.N
328	./sergei digest A full 01.1063.1065.2	1582.8	(-0.0)	3.2171	0.384	1508.4	1	18/ 24	sp|P00921|CAH2 BOVIN	K.YAAELHLVHWNTK.Y
340	./sergei digest A full 01.1081.1083.2	974.1	(+0.2)	3.2106	0.316	454.9	1	14/ 16	sp|P00921|CAH2 BOVIN	K.VLDALDSIK.T
786	./sergei digest A full 01.1949.1949.2	2215.5	(-0.2)	3.2068	0.435	531.2	1	21/ 36	sp|P02754|LACB BOVIN	V.YVEELKPTPEGDLEILLQK.W
1057	./sergei digest A full 01.2573.2575.3	4013.3	(-2.2)	3.1870	0.172	179.7	161	24/196	SW:K2C5 HUMAN	G.AGAGGGYGFGGGAGSGFGFGGGAGGGFGLG
1278	./sergei digest A full 01.3503.3505.3	2709.1	(+1.5)	3.1834	0.198	686.4	1	27/100	sp|P02754|LACB BOVIN	K.VAGTWYSLAMAASDISLLDAQSAPLR.V
523	./sergei digest A full 01.1417.1425.2	1610.9	(-0.3)	3.1601	0.388	1389.3	1	20/ 26	sp|P00489|PHS2 RABIT	K.VHINPNSLFDVQVK.R
60	./sergei digest A full 01.0579.0581.2	1246.3	(+0.2)	3.1600	0.383	1196.1	1	18/ 20	sp|P02754|LACB BOVIN	R.TPEVDDEALEK.F
612	./sergei digest A full 01.1587.1589.3	2204.6	(+0.3)	3.1594	0.323	1098.0	1	28/ 80	sp|Q29443|TRFE BOVIN	K.IMKGEADAMSLDGGYLYIAGK.C
1360	./sergei digest A full 01.3977.3985.2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1	14/ 50	sp|P02754|LACB BOVIN	K.VAGTWYSLAMAASDISLLDAQSAPLR.V
922	./sergei digest A full 01.2247.2259.3	3966.5	[illegible]	[illegible]	[illegible]	211.9	53	21/136	HI0859	E.LVFTDALLDFIGEVGYDPIYGARPLKRAIQ
300	./sergei digest A full 01.1017.1021.2	1640.9	[illegible]	[illegible]	[illegible]	[illegible]	1	18/ 28	sp|P02769|ALBU BOVIN	R.KVPQVSTPTLVEVSR.S
1026	./sergei digest A full 01.2503.2509.3	2854.3	(+0.8)	3.0998	0.206	371.4	2	23/ 96	sp|P00921|CAH2 BOVIN	R.TLNFNAEGEPELLMLANWRPAQPLK.N
846	./sergei digest A full 01.2079.2079.2	1551.8	(-0.4)	3.0860	0.357	573.8	1	18/ 26	sp|Q29443|TRFE BOVIN	R.TAGWNIPMGLLYSK.I
753	./sergei digest A full 01.1869.1871.3	2031.3	(+1.6)	3.0715	0.427	1025.6	1	27/ 76	sp|P02754|LACB BOVIN	Y.SLAMAASDISLLDAQSAPLR.V
576	./sergei digest A full 01.1513.1515.3	1679.9	(+0.1)	3.0691	0.289	468.0	1	21/ 52	sp|P00489|PHS2 RABIT	R.IGEEYISDLDQLRK.L
69	./sergei digest A full 01.0603.0605.3	1250.4	(+0.4)	3.0525	0.115	1575.4	1	23/ 36	sp|P02769|ALBU BOVIN	R.FKDLGEEHFK.G
1022	./sergei digest A full 01.2495.2499.3	2241.5	(-0.3)	3.0417	0.022	486.2	23	22/ 76	HI0053	G.SRTSKGEFEEALDMLATNKI.N
695	./sergei digest A full 01.1759.1761.2	1890.1	(-0.2)	3.0407	0.420	462.0	1	17/ 28	sp|P02769|ALBU BOVIN	R.HPYFYAPELLYYANK.Y
215	./sergei digest A full 01.0873.0875.2	1605.8	(-0.2)	3.0358	0.442	555.7	1	15/ 26	sp|Q29443|TRFE BOVIN	K.DNPQTHYYAVAVVK.K
1385	./sergei digest A full 01.4093.4101.2	2709.1	(+0.6)	3.0143	0.212	596.9	1	17/ 50	sp|P02754|LACB BOVIN	K.VAGTWYSLAMAASDISLLDAQSAPLR.V
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1	[illegible]	[illegible]	[illegible]
106	./sergei digest A full 01.0683.0685.2	1194.4	(+0.1)	2.9868	0.273	930.3	1	15/ 18	sp|P02754|LACB BOVIN	K.VLVLDTDYKK.Y
642	./sergei digest A full 01.1649.1649.3	2913.4	(-0.3)	2.9659	0.180	723.7	1	30/100	sp|P02666|CASB BOVIN	W.MHQPHQPLPPTVMFPPQSVLSLSQSK.V
323	./sergei digest A full 01.1051.1053.3	1636.8	(+0.6)	2.9546	0.355	974.5	1	27/ 52	sp|P02754|LACB BOVIN	R.TPEVDDEALEKFDK.A
795	./sergei digest A full 01.1965.1967.3	2336.6	(+2.4)	2.9429	0.056	362.6	12	20/ 80	HI0497	F.TTDAVKKIAEAAFRVNEKTEN.I
1088	./sergei digest A full 01.2645.2647.3	3336.8	(+1.7)	2.9299	0.096	130.0	79	23/128	HI1196	L.AQLSGGKWAAKCQVHAGGRGKAGGVKLVQDV
132	./sergei digest A full 01.0729.0729.3	1342.6	(+0.3)	2.9069	0.228	624.1	1	19/ 40	sp|P00634|PPB ECOLI	A.RTPEMPVLENR.A
645	./sergei digest A full 01.1653.1655.2	1764.9	(-0.8)	2.8986	0.428	1130.7	1	17/ 26	sp|P46406|G3P RABIT	K.LISWYDNEFGYSNR.V
765	./sergei digest A full 01.1893.1897.3	2333.8	(+2.8)	2.8552	0.003	644.3	1	24/ 76	HI0457	A.TLFEQEKLIADGKLLPYLLK.L
310	./sergei digest A full 01.1031.1033.2	1066.2	(-0.1)	2.8501	0.212	874.5	1	14/ 16	sp|P02754|LACB BOVIN	K.VLVLDTDYK.K
1018	./sergei digest A full 01.2487.2489.3	3325.8	(+2.2)	2.8497	0.210	246.5	2	22/108	sp|P00921|CAH2 BOVIN	V.SSQQMLKFRTLNFNAEGEPELLMLANWR.P
327	./sergei digest A full 01.1057.1061.3	1582.8	(+0.8)	2.8489	0.278	1076.8	1	31/ 48	sp|P00921|CAH2 BOVIN	K.YAAELHLVHWNTK.Y
686	./sergei digest A full 01.1731.1733.3	2264.6	(+0.1)	2.8459	0.065	384.2	6	21/ 72	HI0134	P.INFFGNCNSPQYKINFTKK.F
791	./sergei digest A full 01.1955.1957.3	3077.9	(+0.8)	2.8390	0.111	120.8	8	18/104	HI0685	A.DLKIKMFPAKGALLVMGHRINKLVINR.C
359	./sergei digest A full 01.1123.1125.3	1636.8	(-0.3)	2.8353	0.272	580.5	1	22/ 52	sp|P02754|LACB BOVIN	R.TPEVDDEALEKFDK.A
145	./sergei digest A full 01.0755.0757.2	1194.4	(-0.3)	2.8348	0.242	931.9	1	15/ 18	sp|P02754|LACB BOVIN	K.VLVLDTDYKK.Y
721	./sergei digest A full 01.1805.1809.3	3491.9	(-1.9)	2.8180	0.072	163.1	8	20/120	HI1188	Q.KVKHEKFGFGTVINVEGSENNTRLQIAFQAQ
875	./sergei digest A full 01.2147.2149.3	3518.3	(+1.0)	2.8168	0.089	211.1	12	22/136	HI1015	L.GSIGIVIGLGVMMGSVLEVSGAAEKMAYSFI
427	./sergei digest A full 01.1243.1245.3	2909.3	(-1.0)	2.8109	0.039	204.5	125	21/100	HI1237	T.IHVLQGERKRAADNKSLGQFNLEGIN.P
770	./sergei digest A full 01.1905.1911.3	2351.9	(-1.2)	2.8101	0.177	268.9	132	19/ 88	HI1066	I.GIAERIVSPVEVILFILAAVLGA.Y
926	./sergei digest A full 01.2253.2255.3	2872.1	(-2.0)	2.7800	0.230	524.0	1	26/112	HI0264	T.YTGATYASARDLPQGSSVVVGEANVSTIG.N
759	./sergei digest A full 01.1879.1883.3	3472.0	(-2.4)	2.7784	0.011	112.1	11	23/124	SW:TRYP BOVIN +2	I.VHPSYNSNTLNNDIMLIKLKSAASLNSRVAS
1300	./sergei digest A full 01.3637.3649.3	3519.9	(-2.4)	2.7692	0.114	270.7	13	20/132	HI0906	K.AEALGEIPVGAVLVDDARNIIGEGWNLSIVQ
1242	./sergei digest A full 01.3131.3133.3	3847.3	(-2.6)	2.7692	0.032	184.7	202	19/140	HI0528	Q.LATHGVLSQSEFVRILQADPTLNGLVNATRE
605	./sergei digest A full 01.1579.1579.2	1441.6	(+0.7)	2.7514	0.481	934.6	1	17/ 22	sp|P00489|PHS2 RABIT	K.LLSYVDDEAFIR.D
749	./sergei digest A full 01.1861.1863.3	2327.7	(+0.7)	2.7508	0.176	490.9	3	20/ 72	SW:K1CI HUMAN	S.DLEMQYETLQEELMALKKN.H
1361	./sergei digest A full 01.3977.3985.3	4062.8	(+0.2)	2.7469	0.118	355.3	3	25/152	HI0219	A.WVPNLGGDLPTPAYGFPYLSGVGRLVIKDII
1059	./sergei digest A full 01.2579.2587.2	2462.0	(-0.6)	2.7408	0.311	374.1	1	15/ 44	sp|P02666|CASB BOVIN	I.PPLTQTPVVVPPFLQPEVMGVSK.V
279	./sergei digest A full 01.0977.0979.2	1306.5	(+0.2)	2.7400	0.402	629.4	1	15/ 20	sp|P02769|ALBU BOVIN	K.HLVDEPQNLIK.Q
585	./sergei digest A full 01.1539.1541.2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1	20/ 34	sp|P02754|LACB BOVIN	L.AMAASDISLLDAQSAPLR.V
172	./sergei digest A full 01.0799.0799.3	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1	20/ 68	HI0262	Q.QIPDQSTVFNYYLTPDNP.Y
507	./sergei digest A full 01.1389.1389.2	2477.5	(+0.3)	2.7215	0.348	214.1	1	14/ 38	sp|P02666|CASB BOVIN	K.FQSEEQQQTEDELQDKIHPF.A
184	./sergei digest A full 01.0819.0823.3	2081.3	(+2.3)	2.7195	0.081	622.4	12	24/ 68	HI0194	K.EFGQRPVALVDFNEPFSK.S
953	./sergei digest A full 01.2323.2327.3	2202.7	(+2.3)	2.7189	0.004	416.9	144	18/ 72	HI0936	K.GQFDRTLLIRCGWLLIPSL.I
932	./sergei digest A full 01.2267.2269.3	2928.4	(-2.8)	2.7159	0.054	275.9	46	21/104	HI0508	G.LIDAELAQLAVDNGWEGIIVYGAVRQT.Q

"correct"

incorrect

sort by search score

threshold

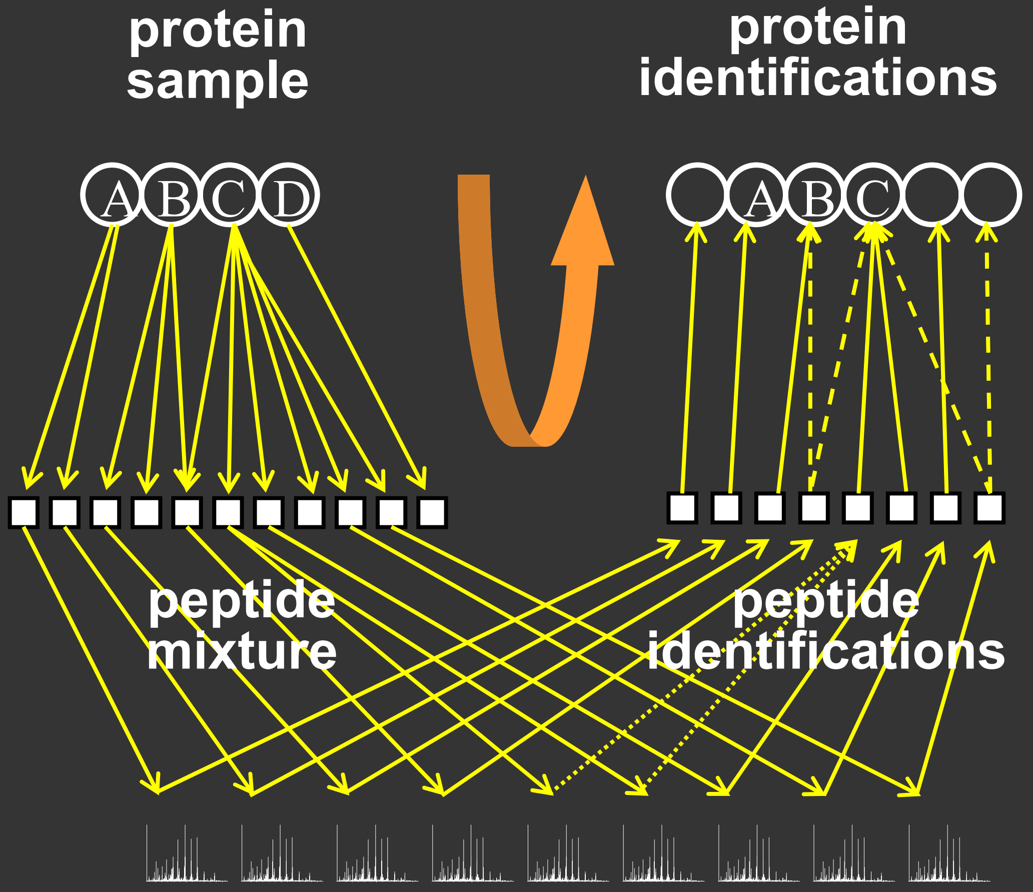
SEQUEST:
Xcorr > 2.0
ΔC_n > 0.1

MASCOT:
Score > 47

Difficulty Interpreting Protein Identifications based on MS/MS

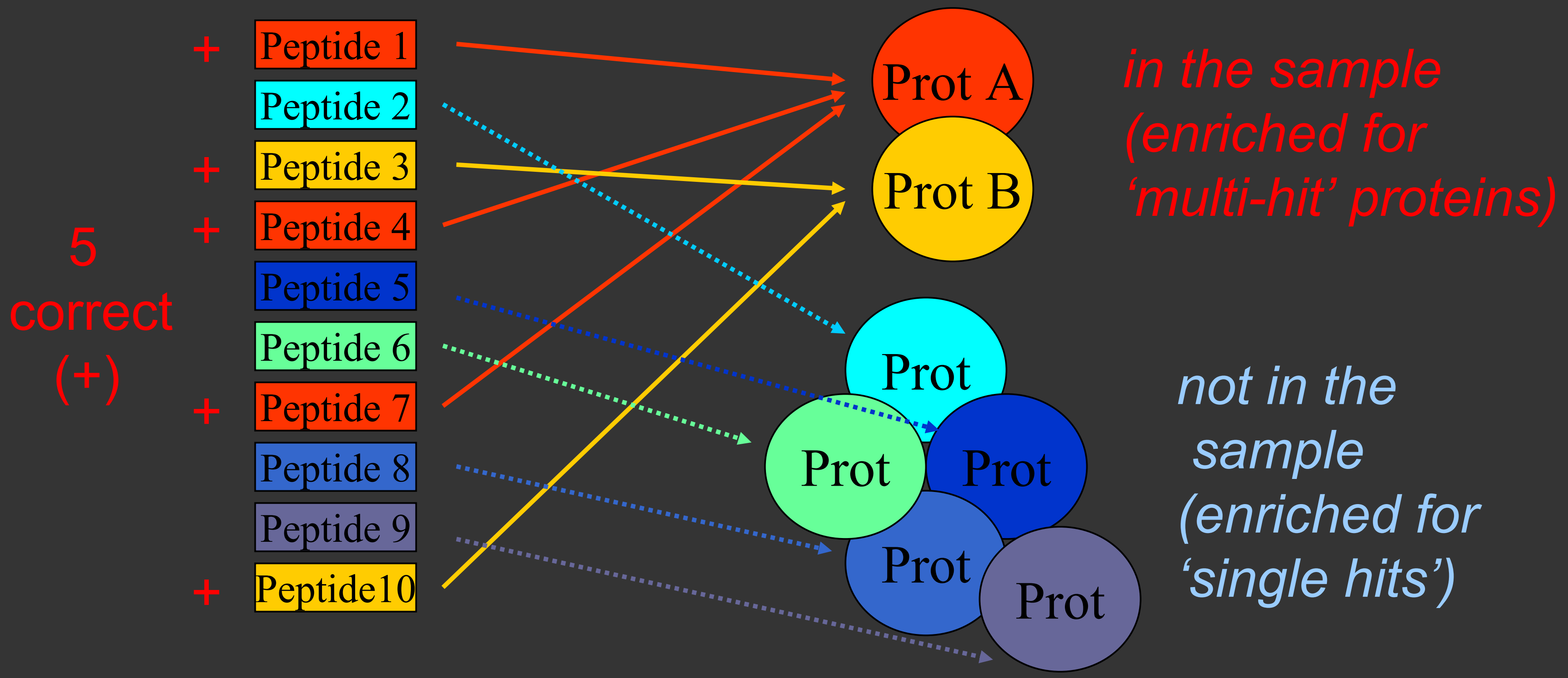
- Different search score thresholds used to filter data
- Unknown and variable false positive error rates
- No reliable measures of confidence

Protein Identification by MS/MS



MS/MS spectra

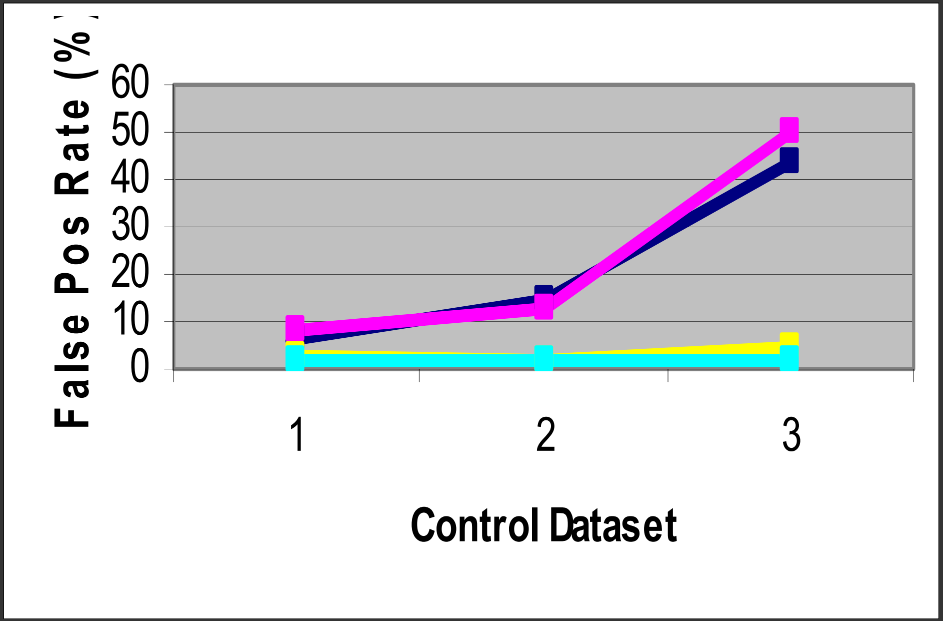
Amplification of False Positive Error Rate from Peptide to Protein Level



Peptide Level: 50% False Positives

Protein Level: 71% False Positives

Protein ID False Positive Rate: Control Dataset Examples



Data Filters:

- Publ. threshold model #1
- Publ. threshold model #2
- Statistical model ($p \geq 0.5$)
- Statistical model predicted

Control Datasets:

1 18 purified proteins vs. 18+Human (22 runs)

2 Halobacterium vs. Halo+Human (4 runs)

3 Halobacterium vs. Halo+Human (45 runs)

False Positive Error Rates among Single-hit Proteins

	Control Dataset		
Data Filter	*1*	*2*	*3*
Publ. Threshold model #1	11%	37%	67%
Publ. Threshold model#2	14%	32%	82%

Control Datasets:

1 18 purified proteins vs. 18+Human (22 runs)

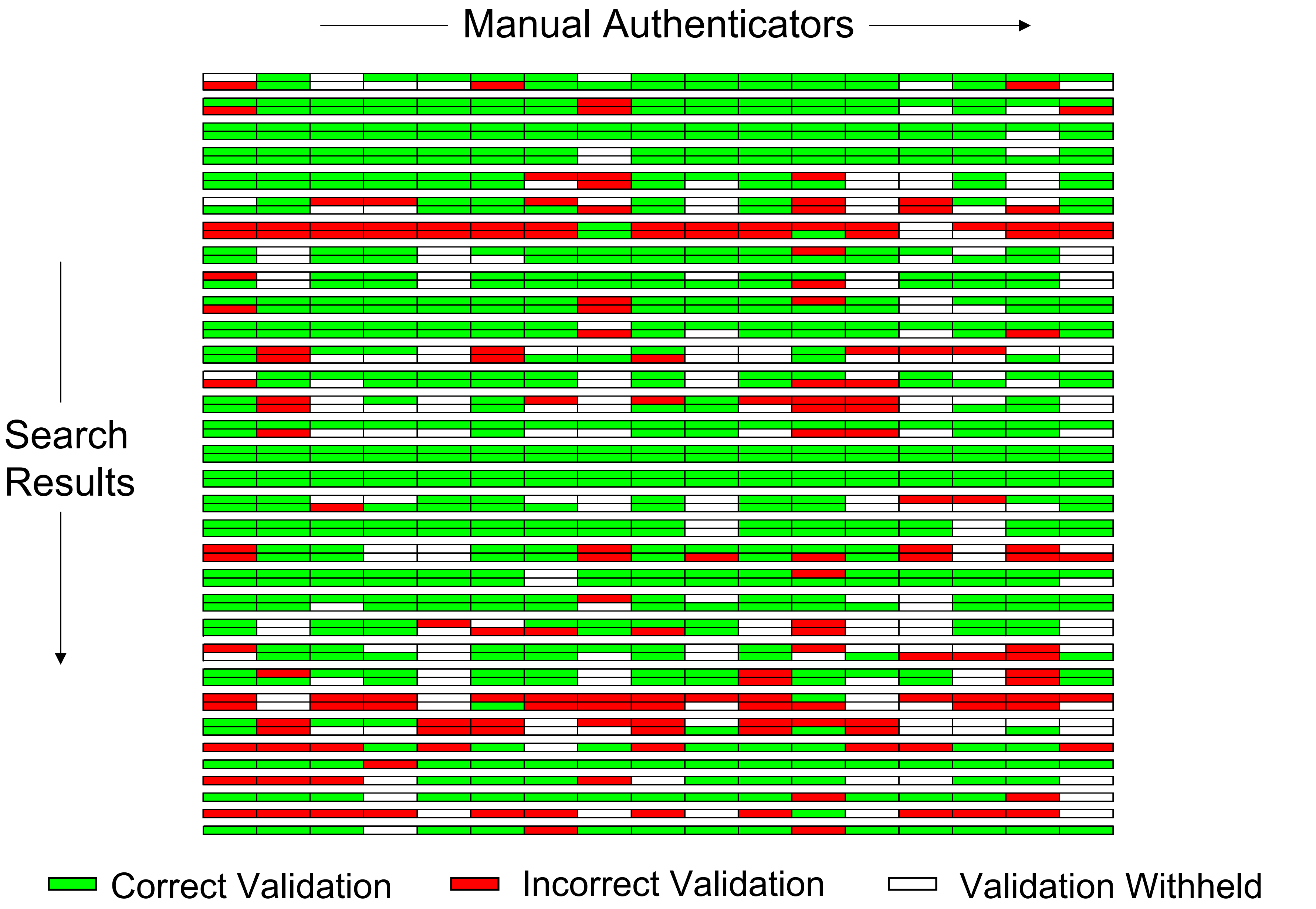
2 Halobacterium vs. Halo+Human (4 runs)

3 Halobacterium vs. Halo+Human (45 runs)

Serum Protein Identifications from Large-scale (~375 run) Experiment

Data Filter	*# ids*	*# non-single hits*	*# single-hits*
Publ. Threshold model#1	2257	359	1898
Publ. Threshold model #2	2742	441	2301
Statistical model, $p \geq 0.5$ *(predicted error rate: 7%)*	713	511	202

Consistency of Manual Validation of SEQUEST Search Results



Discussion Points

Extensive (biological, manual) validation of all data impossible

Current status:

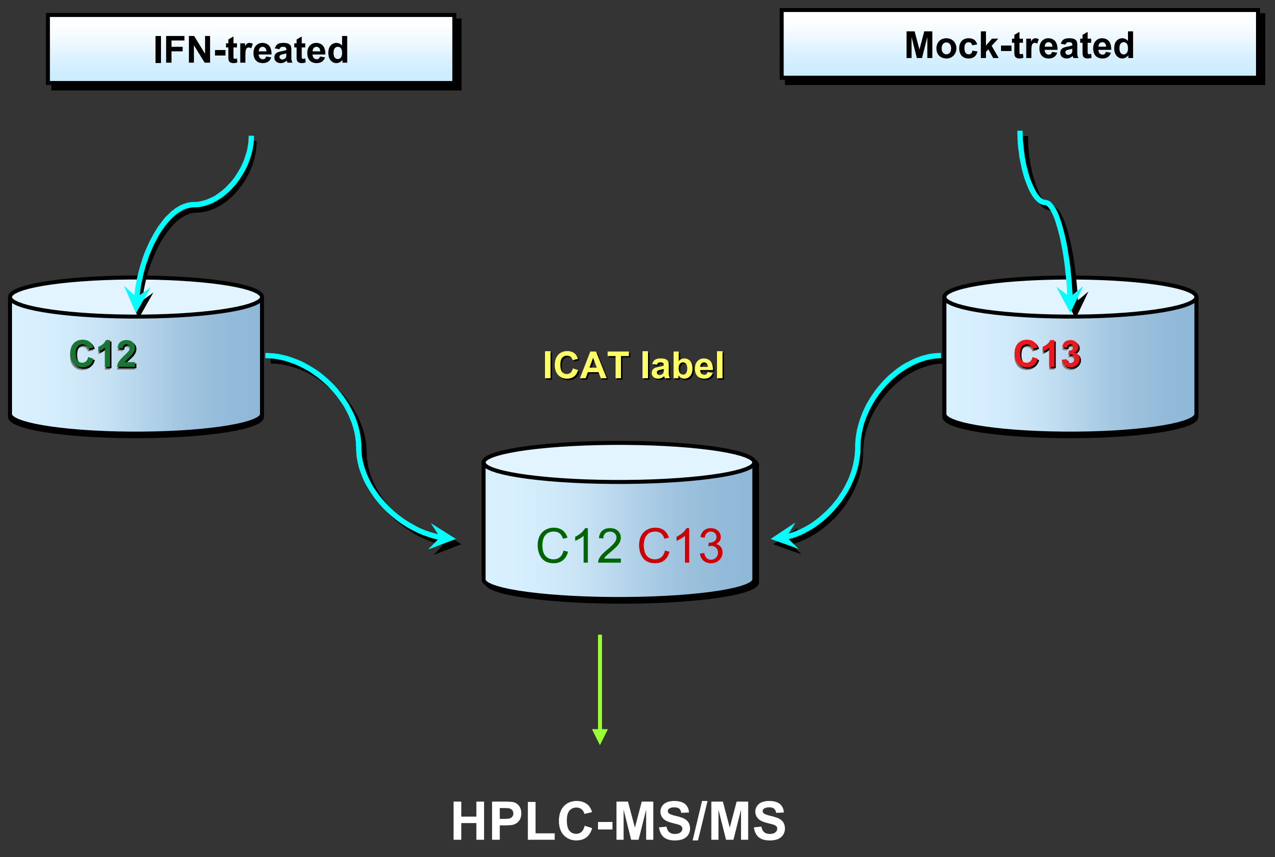
- Peptide and protein identifications are largely made based on threshold model
- Manual validation is often used as “gold standard”

Recommendations:

- Develop, validate and use statistical models that calculate accurate false positive and false negative error rates for peptide AND protein identifications
- Discourage manual validation of spectra as “gold standard”.
- Tools should be transparent and generally available

Discussion Points

- Value of information increases if data from multiple experiments/groups can be integrated and collectively mined
- Proteomics is a community effort
- Data are collected and organized in relational databases



IFN-treated

Mock-treated

C12

C13

ICAT label

C12 C13

HPLC-MS/MS

Wei Yan et al. MCP 2004

Name	Cellular pathway	Probability	ASAPRatio Mean	ASAPRatio Std.
DNAH11: dynein, axonemal, heavy polypeptide 11	moto protein complex	0.94	9999	-1
UBE2L6: ubiquitin-conjugating enzyme E2L 6	ubiquitination and protein degradation	0.57	9999	-1
IFIT1: interferon-induced protein with tetratricopeptide repeats 1	unknown and ESIs	0.48	9999	-1
GPR111: G protein-coupled receptor 111	G-protein coupled receptor and G-protein signaling	0.63	21.270	4.741
PASK PAS domain co				1.024
ADRM1: adhesion re				1.043
CSA_PPIasePEPTIDY				1.070
AHCY: S-adenosylhom				0.936
IFIT4: interferon-indu				0.794
FLJ32915: hypothetica				4.883
GNB1: guanine nucleo				0.133
G1P2: interferon, alp				0.661
MTP: microsomal trigly				0.751
PLCD1: phospholipase				0.116
CD7: CD7 antigen (p4				2.204
PRKR protein kinase, dependent				0.659
KIAA1276: KIAA1276				0.058
NUDT2: nudix (nucleos				0.224
CABC1: chaperone, A				1.659
ACACA: acetyl-Coenzy				0.259
KNS2: kinesin 2 60/70				0.335
LOC151636: rhysin 2	cytoskeleton and intracellular transport?	1	2.975	0.231
M96: likely ortholog of mouse metal response element binding transcription factor 2	transcription	0.98	2.923	0.390
ETFA: electron-transfer-flavoprotein, alpha polypeptide (glutaric aciduria II)	electron transfer	0.45	2.890	0.484
NMI: N-myc (and STAT) interactor	signaling pathway; transcriptio	0.57	2.875	0.138
GSA7: ubiquitin activating enzyme E1-like protein	ubiquitination and protei	0.98	2.844	0.663
MGC3207: hypothetical protein MGC3207		0.61	0.499	0.071
SPK: symplekin		1	0.496	0.029
KRT10: keratin 10 (epidermolytic hyperkeratosis; kerat plantaris)		0.97	0.495	0.055
SARDH: sarcosine dehydrogenase		0.98	0.484	0.008
TRA1: tumor rejection antigen (gp96) 1		1	0.452	0.165
GPS1: G protein pathway suppressor 1		0.98	0.455	0.138
SRRM2: serine/arginine repetitive matrix 2		0.82	0.434	0.224
KIAA0007: KIAA0007 protein		1	0.426	0.014
FACL4: fatty-acid-Coenzyme A ligase, long-chain 4		0.98	0.416	0.081
FXR2: fragile X mental retardation, autosomal homolog		0.95	0.391	0.074
TUBA6: tubulin alpha 6		1	0.383	0.165
CPSF4: cleavage and polyadenylation specific factor 4		0.96	0.378	0.154
MAPRE1: microtubule-associated protein, RP/EB fami		0.98	0.339	0.016
OAT: ornithine aminotransferase (gyrate atrophy)		0.98	0.331	0.018
PPGB: protective protein for beta-galactosidase (galact		1	0.323	0.084
WNT9A: wingless-type MMTV integration site family, memb		0.99	0.316	0.091
FASN: fatty acid synthase	lipid and fatty acid metabolism	0.99	0.304	0.100
Ig lambda chain C regions	immune response	0.98	0.265	0.110
G2AN: alpha glucosidase II alpha subunit	carbohydrate metabolism	1	0.198	0.033
Hypothetical protein FLJ21140	unknown	0.71	0.043	0.064
KRT6: keratin 6	cytoskeleton and intracellular transport	1	0.003	0.008
MIG-6: Gene 33/Mig-6	signaling pathway	0.99	0.000	-1.250
HIC1: hypermethylated in cancer 1	transcription suppression	0.94	0.000	-1.250

	S100	P100	P3	Sum	Unique ID
$P \geq 0.9$	523	270	671	1464	1113
$P \geq 0.4$	590	330	748	1668	1272

54 IFN-induced proteins (2-fold)

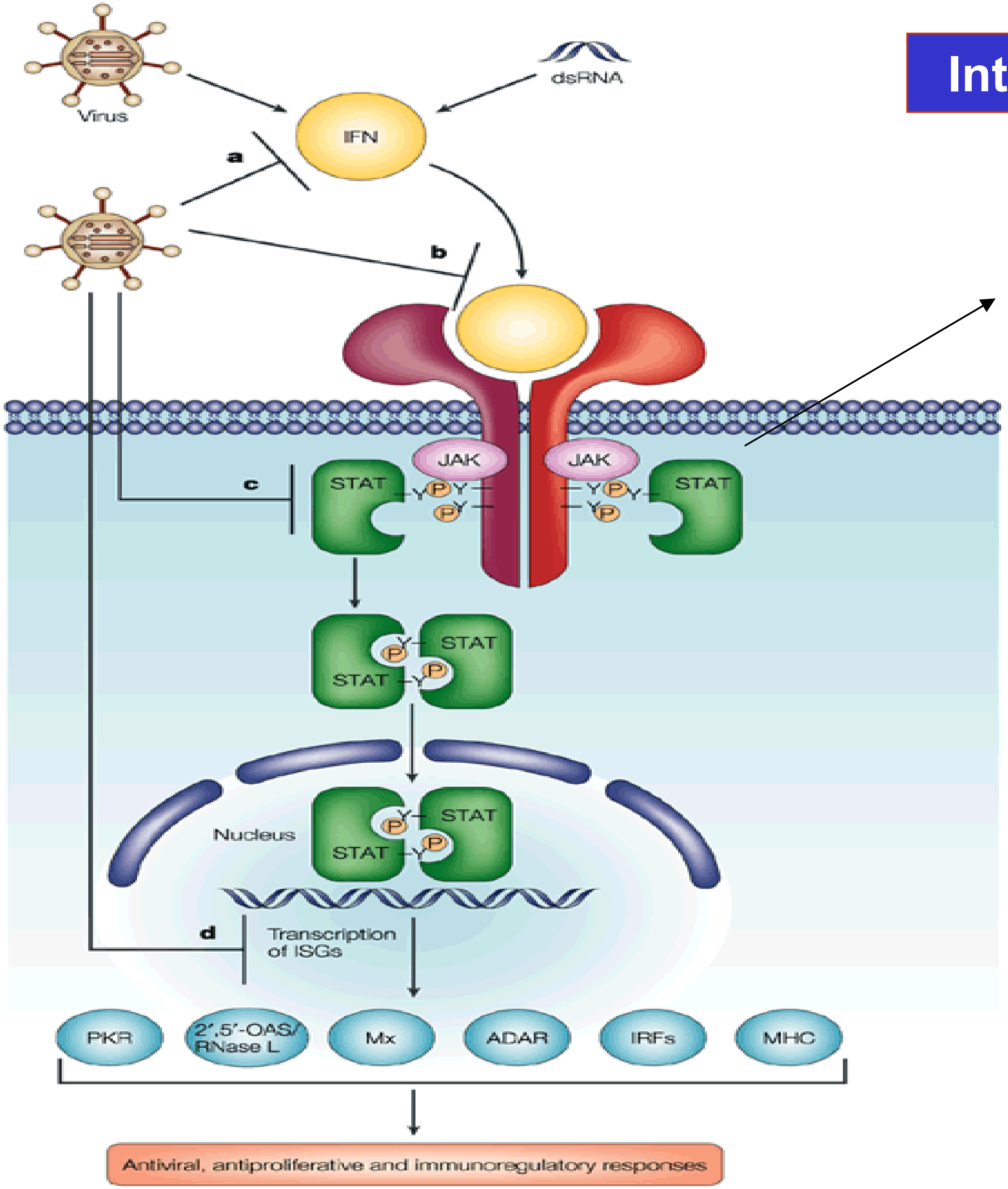
15 previously reported

39 novel

23 IFN-repressed proteins (0.5-fold)

Lots of data -what does it mean?

Interferon (IFN) Pathway



Katze et al (2002) 2: 675

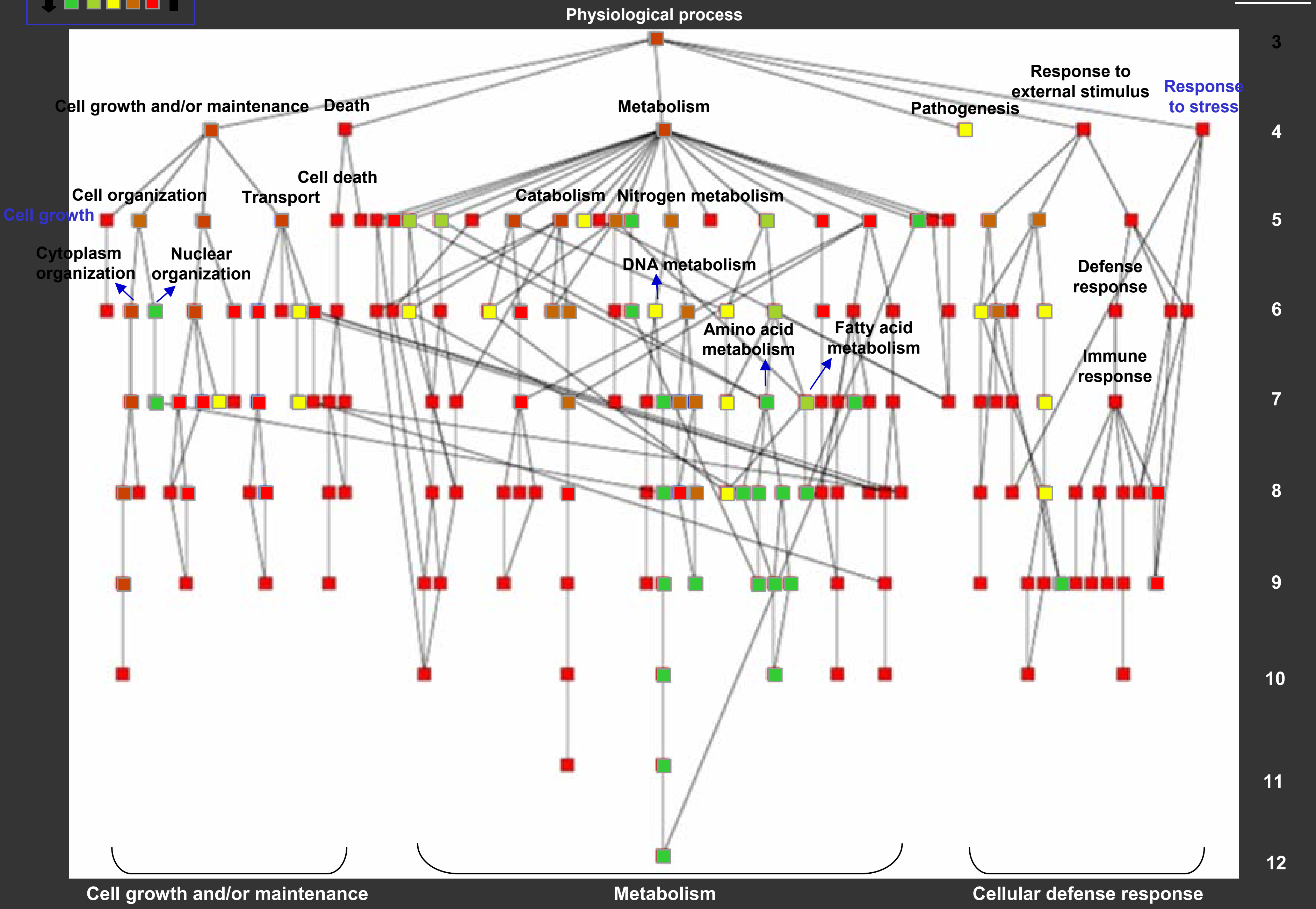
Nature Reviews | Immunology

2.215 ± 0.079

	IFN / Mock
PKR	3.963 ± 0.659
2',5'-OAS	2.460 ± 0.076
Mx	2.359 ± 0.149
ADAR	1.398 ± 0.118
IRFs	Not identified
MHC	
β-2-microglobulin (MHC I)	2.768 ± 0.583
IFI-30 (MHC II)	2.219 ± 0.183

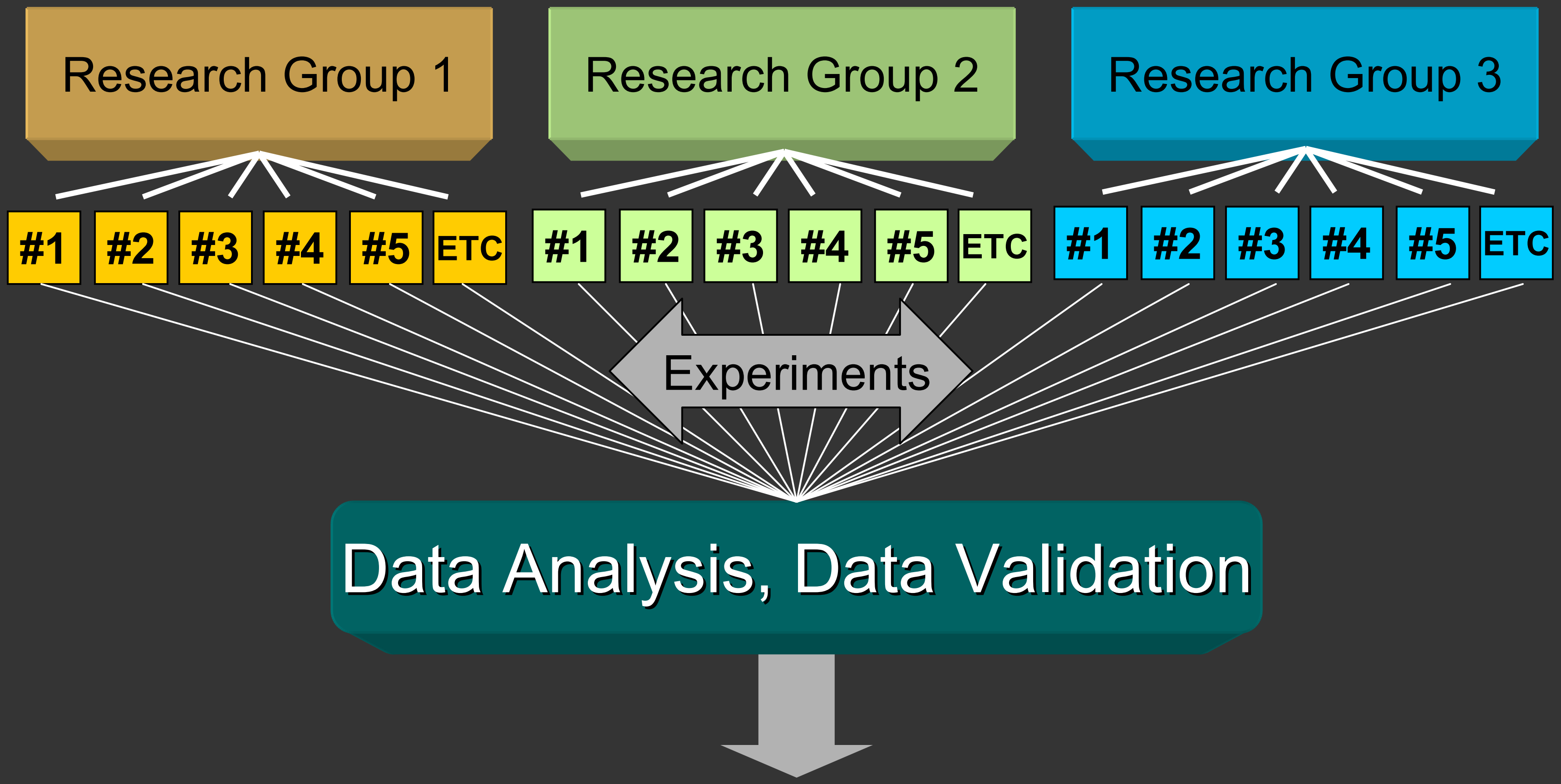
GO Analysis of Interferon regulated proteins

⬇ 🟩 🟩 🟨 🟧 🟥 ⬆



- Single proteomics datasets tend to rediscover the known
- New insights can be made from the comparison of many datasets

Data collection in proteomics



Research Group 1

#1 #2 #3 #4 #5 ETC

Research Group 2

#1 #2 #3 #4 #5 ETC

Research Group 3

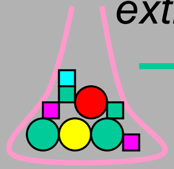
#1 #2 #3 #4 #5 ETC

Experiments

Data Analysis, Data Validation

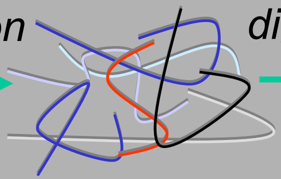
Relational Database

From Peptides to Genome Annotation



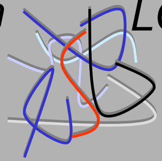
Sample

extraction



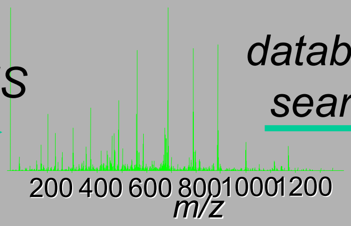
Proteins

digestion



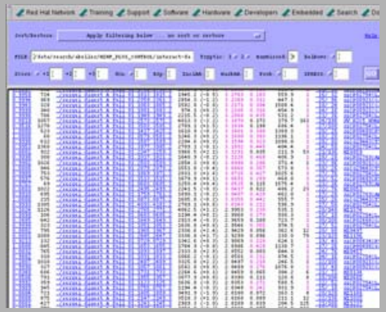
Peptides

LC-MS/MS



Mass Spectrum

database search



Peptides

statistical filtering

Spectrum	*Peptide*	*Probability*
Spectrum 1	LGEYGH	1.0
…	…	…
Spectrum N	EIQKKF	0.3

BLAST protein database

Map to genome

Peptide	*…*	*Chrom*	*Start_Coord*	*End_Coord*	*…*
PAp00007336	…	X	132217318	132217368	…
…	…	…	…	…	…

PeptideAtlas Database

visualization

Genome Browser

Tools: www.proteomecenter.org

Desiere et al , Genome Biology, (2004)

Discussion Points

- Value of information increases if data from multiple experiments/groups can be integrated and collectively mined
- Proteomics is a community effort
- Data are collected and organized in relational databases

Current status:

- Very little proteomics data publicly accessible
- Publications usually only show conclusions but not data

Recommendations:

- Develop and support infrastructure for data sharing and mining
- Make data access condition for publication

Summary

If proteomics is to truly operate as a discipline of the genomic sciences, data processing, management and dissemination strategies proven in other fields of genomics must be applied. These include:

- Statistical validation of large data sets
- Providing community access to all data (not just selected data points)
- Providing transparent tools for data processing to community