



*Proteome Informatics
Research Group*

ABRF iPRG2008 Study

Assessing the Quality and
Consistency of Protein Reporting
on a Common Data Set

ABRF2008, Salt Lake City, UT
February 12, 2008



*Proteome Informatics
Research Group*

The ABRF Proteome Informatics Research Group (iPRG)

Members

- Jayson A. Falkner – University of Michigan (*new member*)
- Jeffrey A. Kowalak – NIH (*E.B. Liaison*)
- William S. Lane - Harvard University
- Alexey I. Nesvizhskii - University of Michigan
- Brian C. Searle - Proteome Software (*incoming chair*)
- Sean L. Seymour - Applied Biosystems (*outgoing chair*)
- David L. Tabb - Vanderbilt University



*Proteome Informatics
Research Group*

iPRG2008 Study Description

- **Common starting point: MS dataset and FASTA DB**
 - All participants were given the same data set
 - The data set was available in both raw .wiff format and in several peak list formats (.mgf, .dta, mzXML, mzData)
 - A specific FASTA database was required
- **Task for respondents: Report a protein ID result as you would for a journal**
 - Given an Excel reporting template and minimal instructions, asked respondents to report proteins and peptides supporting them.



*Proteome Informatics
Research Group*

Study Goals

- **Primary Goal: Assess the current state of protein reporting.**
 - (1) Does the proteomics community still have a problem with reporting excessive numbers of proteins due to improper protein inference?
 - (2) Are people reporting accession ambiguity?
- **Secondary Goal: Assess similarity of reporting given a common MS analysis results (rather than a common sample)**
 - Expect results to be more similar than when acquisition is also a variable.
 - Read differences as variation in protein and peptide ID analysis only.
- **Secondary Goal: Develop a benchmark**
 - Develop a reference test for both software users and developers.



*Proteome Informatics
Research Group*

The Mass Spec Data Set

- A **mouse** liver differential expression experiment
 - Trypsin digestion
 - MMTS alkylation
 - iTRAQ® reagent labeled
- Separated into **13 cation exchange fractions**
- Analyzed on a 3200 QTRAP® system
- Total of **29 acquired files** since most fractions were analyzed with 1-2 rounds excluding prior precursors
- Total spectra in 29 files: **41,977 spectra**
- Converted raw .wiff files to .mgf, .dta, mzData, mzXML
 - Mass spec format inter-conversion is still problematic!!



*Proteome Informatics
Research Group*

The FASTA Database

- **Selected DB compiled by MGI – Mouse Genome Informatics**
 - Hosted by the Jackson Laboratory, Bar Harbor, Maine
 - ftp://ftp.informatics.jax.org/pub/sequence_dbs/seq_dbs.current/uniprotmusmgi.z
 - Version downloaded **Dec. 3, 2007** (53,826 protein sequences)
 - We added 74 contaminant proteins
 - Total sequences: **53,900**
- **Respondents could use:**
 - This database
 - A concatenated decoy DB (forward + reversed) we provided
 - Any variant database, as long as it was based on the provided DB.



*Proteome Informatics
Research Group*

Protein Inference Terminology

cluster – A cluster of proteins is a collection of homologous proteins that could cite some common MS/MS fragmentation evidence.

Accession 1	AAK PEPTIDE AAAKAK PEPTIDE BBBKAK PEPTIDE DDDKAA
Accession 2	AAAAK PEPTIDE AAAKAK PEPTIDE BBBKAAAK PEPTIDE EEEEK
Accession 3	AAAAAAAAAAAAAAAAAK PEPTI <u>D</u> EBBBK AAAAAK PEPTIDE FFFKA

- Common sequences not required – e.g. Ile/Leu difference
- Aggregation into a cluster is completely independent of how many of these isoforms can be detected in a given sample!

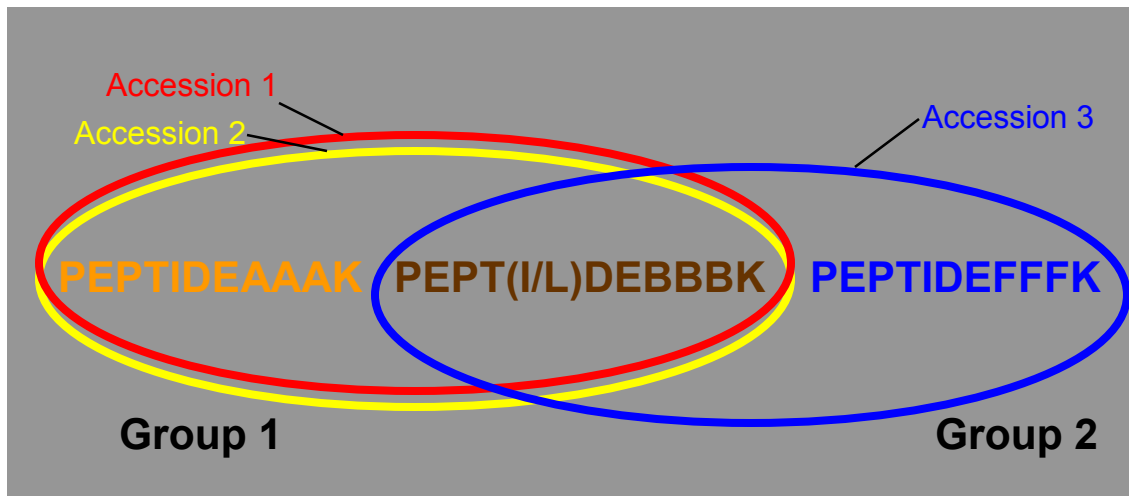


Proteome Informatics
Research Group

Protein Inference Terminology

group – One reported protein group implies the detection of a physical protein species. Because there may be ambiguity as to which accession best corresponds to this detected isoform, a group of accessions must be reported.

If we observe peptides A, B, and F in a set of peptide IDs ...



...we report a protein ID list:

<u>Protein N</u>	<u>Accession</u>	
1	Accession 1	} Group 1
1	Accession 2	
2	Accession 3	} Group 2

- Two isoforms can be detected in this cluster based on these IDs, thus two protein groups are reported in the results. Refer to this as a 'multi-detection' or 'multi-isoform' cluster.
- There is ambiguity in which accession is being detected in Group 1.



*Proteome Informatics
Research Group*

Determination of the 'Correct' Proteins

- **Each RG member analyzed the data independently with the tools of their choosing like any study respondent.**
- **All observed accessions were assigned to protein clusters using:**
 - Accessions listed in common groups in RG results
 - UniRef50 mapping
 - Proteins cited by shared peptides
- **Default assumption: the expected number of detectable isoforms per cluster is the maximal number supported by at least 3/6 RG results.**
- **Allowed manual inspection by RG members to argue for changes to the default assumption.**



iPRG Analysis Tools Used

*Proteome Informatics
Research Group*

	Peptide Identification	Protein Inference	Error Rate Assessment
Kowalak	Mascot	MassSieve	Peptide level 1% FDR, 2 distinct peptides (<i>concatenated decoy</i>)
Lane	SEQUEST	Proteomics Browser Suite	Protein & peptide 1% FDR, 2 distinct peptides (<i>concatenated decoy</i>)
Nesvizhskii	X!Tandem w/ k-score plug-in	Protein Prophet	Protein level ~1% FDR (P=0.9 →0.7% FDR) (<i>concatenated decoy</i>)
Searle	Mascot	Scaffold	Protein level 90% probability
Seymour	Paragon	Pro Group	Protein level 1% FDR (<i>concatenated decoy</i>)
Tabb	MyriMatch	IDPicker	Peptide 5% FDR, 2 distinct peptides

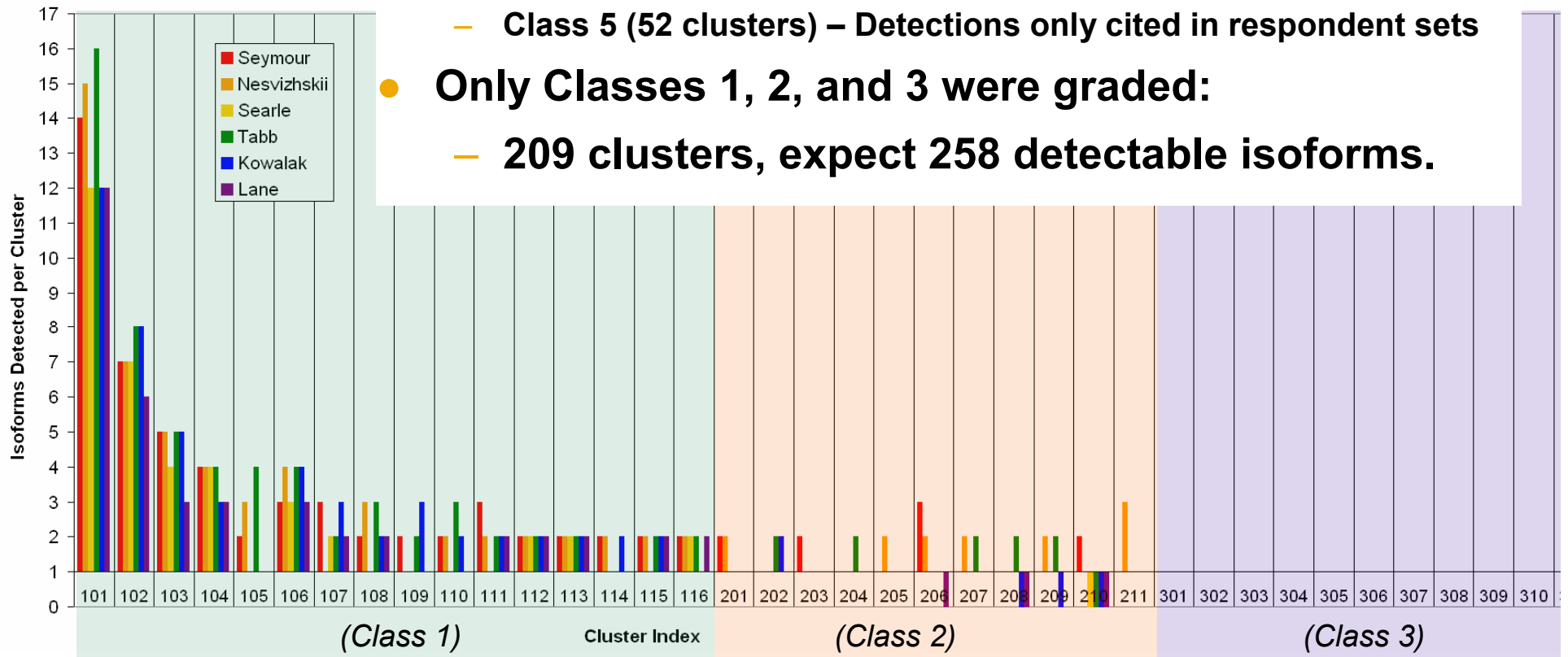
- **All tools were used automatically (no manual curation).**



The Resulting Annotation

Proteome Informatics
Research Group

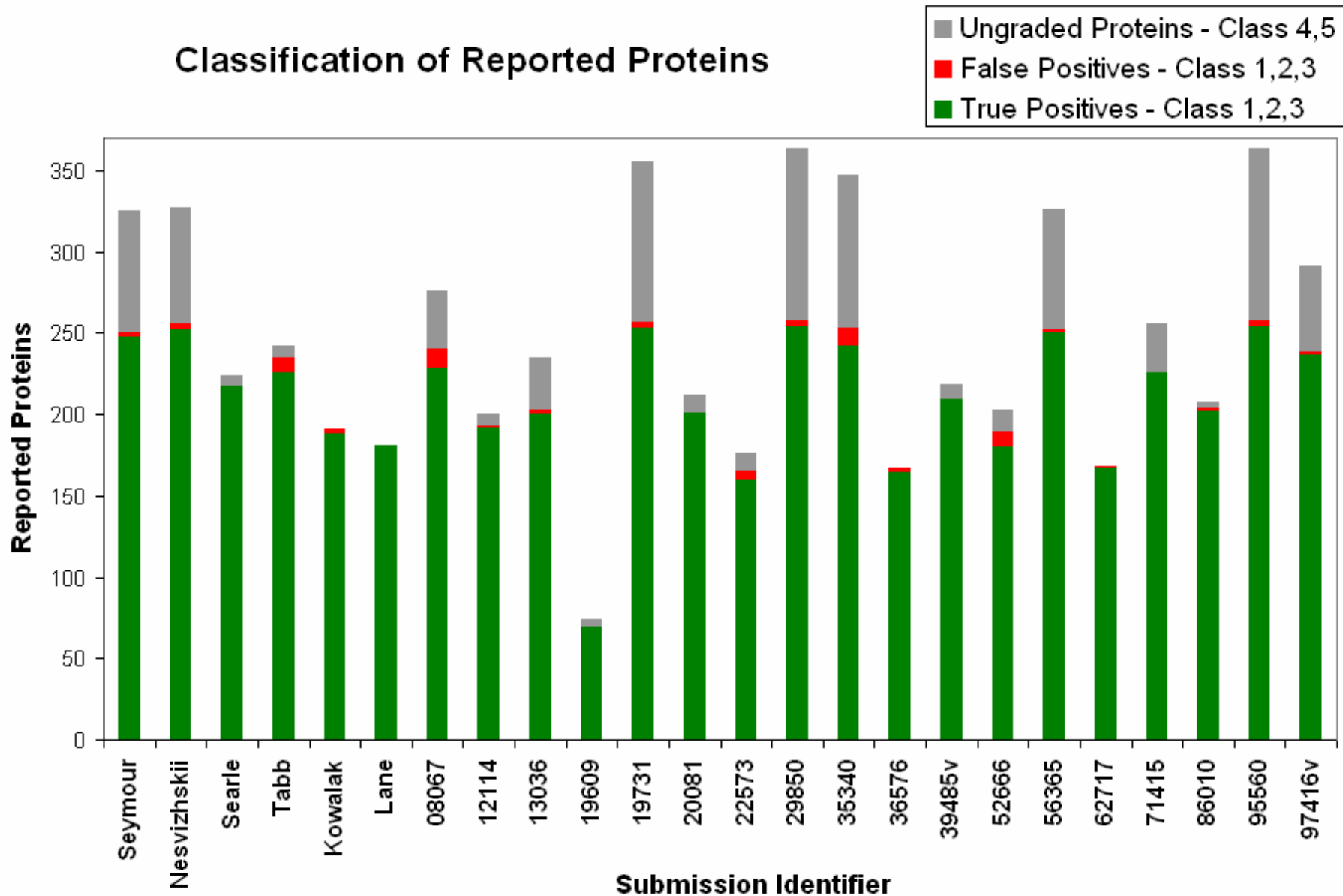
- **Protein clusters were grouped in classes:**
 - Class 1 (16 clusters) – RG consensus multi-detection clusters
 - Class 2 (11 clusters) – Debatable multi-detection clusters
 - Class 3 (182 clusters) – RG consensus single-detection clusters
 - Class 4 (154 clusters) – RG non-consensus detections
 - Class 5 (52 clusters) – Detections only cited in respondent sets
- **Only Classes 1, 2, and 3 were graded:**
 - 209 clusters, expect 258 detectable isoforms.





Overview of All Submissions

Classification of Reported Proteins

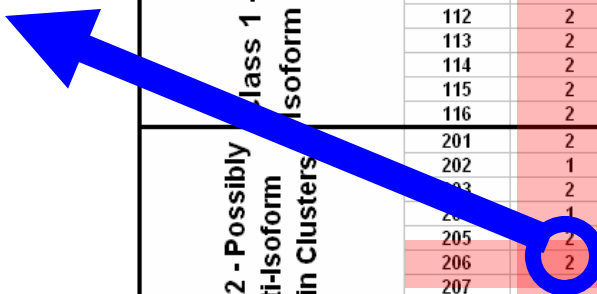




The Table...

Proteome Informatics
Research Group

RG expected number
of isoforms for
cluster 206 is two.



			29850	95560	19731	56365	Seymour	97416	71415	20081	12114	13036	Searle	62717	Kowalak	22573	36576	39485	52666	35340	Nesvizhskii	Tabb	08067	86010	Lane	19609	
Research Group/Vendor/Builder(non-vendor)/User			U	U	U	U	RG	V	U	U	U	U	RG	U	RG	U	U	V	U	B	RG	RG	?	B	RG	B	
Total Class 1,2,3 False Positives			4	4	4	2	2	1	0	0	1	3	0	1	3	5	3	0	9	11	4	9	12	2	0	0	
Total Class 1,2,3 False Negatives			4	4	5	8	10	21	32	57	66	58	41	91	70	98	94	49	78	16	6	32	30	56	77	189	
Protein Inference Tool			PG	PG	PG	PG	PG	PG	PG	PG	PG	Sc	Sc	Sc	MS	MS	MS	M	M	C	PP	ID	?	pC	PBS	SW	
Peptide ID Tool			P	P	P	P	P	P	P	P	P	M	M	M	M	M	M	M	M	B	C	My	?	pF	S	XI, Pe	
Total Reported Proteins			364	364	355	326	325	291	256	212	200	235	224	168	191	176	167	218	203	347	327	242	276	207	181	74	
Total Class 1,2,3 True Positives			254	254	253	250	248	237	226	201	192	200	217	167	188	160	164	209	180	242	252	226	228	202	181	69	
Cluster Class	Cluster Index	Expected # Detections																									
Class 1 - Consensus Multi-Isoform Protein Clusters	101	14	0	0	0	0	0	-1	-2	3	-2	-1	-2	-5	-2	-4	-2	-3	-1	1	1	2	1	-2	-2	6	
	102	7	0	0	0	0	0	-1	-1	-1	-1	-1	0	-1	1	-1	0	0	2	-1	0	1	2	0	-1	4	
	103	5	0	0	0	0	0	0	0	0	0	0	-1	-1	-1	0	-1	0	-2	0	0	0	0	0	-2	-3	
	104	4	0	0	0	0	0	0	0	-1	0	0	-1	0	-1	-1	-1	-1	-1	0	0	0	0	0	0	-1	-1
	105	3	-1	-1	-1	-1	-1	-2	-1	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	0	0	1	-1	-2	-2	-2	-2
	106	4	0	0	0	0	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	-2	-1	0	-1	0	0	0	0	3	-1	-2
	107	3	0	0	0	0	0	-1	-1	-1	-1	0	-1	0	-1	0	-1	0	-1	2	-1	-2	-1	0	0	-1	-2
	108	3	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	0	-1	0	0	0	-1	-1	-1
	109	2	1	1	1	1	0	0	0	-1	1	-1	-1	0	1	-1	-1	-1	1	2	-1	0	3	-1	-1	-1	-2
	110	2	0	0	0	0	0	-1	-1	-1	0	-1	-1	-1	0	-1	-1	-1	-1	-1	0	1	0	0	-1	-1	-2
	111	3	0	0	0	-1	0	-1	-1	-2	-2	-1	-2	-2	-1	-3	-1	-1	0	-1	-1	-1	-1	-1	-1	-1	-3
	112	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	0	0	0	0	0	0	0
	113	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-2
	114	2	0	0	0	0	0	0	0	-1	0	-1	-1	0	-1	1	-1	0	-1	0	-1	-1	-1	-1	-1	-1	-1
	115	2	0	0	0	0	0	-1	-1	-1	-1	0	-1	-1	0	-1	-1	0	0	0	0	0	0	0	0	0	-1
	116	2	0	0	0	0	0	0	0	-1	-1	0	0	-1	-1	-2	-2	0	-1	0	0	0	0	0	0	0	-1
Class 2 - Possibly Multi-Isoform Protein Clusters	201	2	0	0	0	0	0	0	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	-1	-1	-1	-1	-2	
	202	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	1	0	0	1	0	0	0	0	
	203	2	0	0	0	0	0	0	0	-1	0	-1	-1	-1	-1	-1	-1	-1	0	-1	-1	-1	-1	-1	-1	-1	
	204	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1	1	0	1	2	1	0
	205	2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	-1	-2	
	206	2	1	1	1	1	1	0	0	-1	0	0	-1	-1	-1	-1	3	-1	-1	0	1	0	-1	0	-1	-2	-1
	207	1	0	0	0	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	0	0	0	-1	-1	-2
	208	1	0	0	0	0	0	0	0	0	-1	-1	0	0	-1	-1	-1	-1	-1	-1	0	0	1	0	-1	-1	-1
	209	1	0	0	0	-1	0	0	0	0	0	0	0	1	-1	-1	-1	0	-1	1	1	1	0	0	1	0	-1
	210	1	1	1	1	0	1	1	0	-1	-1	0	-1	-1	-1	-1	-1	-1	-1	-1	0	0	-1	-1	-1	-1	-1
211	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	
	301	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
	302	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
	303	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
	304	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
	305	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
	306	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
	307	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
	308	1	0	0	-1	0	0	0	0	0	-1	1	0	0	0	2	0	0	0	-1	1	0	0	0	0	0	-1
	309	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	310	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	311	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	312	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	313	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	314	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0



The Table...

*Proteome Informatics
Research Group*

			29850	95560	19731	56365	Seymour	97416	71415	20081	12114	13036	Searle	62717	Kowalak	22573	36576	39485	52666	35340	Nesvizhskii	Tabb	08067	86010	Lane	19609	
Research Group/Vendor/Builder(non-vendor)/User			U	U	U	U	RG	V	U	U	U	U	RG	U	RG	U	U	V	U	B	RG	RG	?	B	RG	B	
Total Class 1,2,3 False Positives			4	4	4	2	2	1	0	0	1	3	0	1	3	5	3	0	9	11	4	9	12	2	0	0	
Total Class 1,2,3 False Negatives			4	4	5	8	10	21	32	57	66	58	41	91	70	98	94	49	78	16	6	32	30	56	77	189	
Protein Inference Tool			PG	PG	PG	PG	PG	PG	PG	PG	PG	Sc	Sc	Sc	MS	MS	MS	M	M	C	PP	ID	?	pC	PBS	SW	
Peptide ID Tool			P	P	P	P	P	P	P	P	P	M	M	M	M	M	M	M	M	B	C	My	?	pF	S	XI, Pe	
Total Reported Proteins			364	364	355	326	325	291	256	212	200	235	224	168	191	176	167	218	203	347	327	242	276	207	181	74	
Total Class 1,2,3 True Positives			254	254	253	250	248	237	226	201	192	200	217	167	188	160	164	209	180	242	252	226	228	202	181	69	
Cluster Class	Cluster Index	Expected # Detections																									
Class 1 - Consensus Multi-Isoform Protein Clusters	101	14	0	0	0	0	0	-1	-2	-3	-2	-1	-2	-5	-2	-4	-2	-3	-1	1	1	2	1	-2	-2	6	
	102	7	0	0	0	0	0	-1	-1	-1	-1	-1	0	-1	1	-1	0	0	2	-1	0	1	2	0	-1	4	
	103	5	0	0	0	0	0	0	0	0	0	-1	-1	-1	0	-1	0	-2	0	0	0	0	0	0	-2	-3	
	104	4	0	0	0	0	0	0	-1	0	0	-1	0	-1	-1	-1	-1	-1	-1	0	0	0	0	0	0	-1	
	105	3	-1	-1	-1	-1	-1	-2	-1	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	0	0	1	-1	-2	-2		
	106	4	0	0	0	0	-1	-1	-1	-1	-1	-1	-2	0	-2	-1	0	-1	0	0	0	0	0	3	-1	-2	
	107	3	0	0	0	0	0	-1	-1	-1	-1	0	-1	0	-1	0	-1	0	2	-1	-2	-1	0	0	-1	-2	
	108	3	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	-2	-1	-1	-1	-2	0	-1	0	0	0	-1	-1	-1	
	109	2	1	1	1	1	0	0	0	-1	1	-1	-1	0	1	-1	-1	-1	1	2	-1	0	3	-1	-1	-2	
	110	2	0	0	0	0	0	-1	-1	-1	0	-1	-1	-1	0	-1	-1	-1	-1	0	0	1	0	0	-1	-2	
	111	3	0	0	0	-1	0	-1	-1	-2	-1	-2	-2	-1	-3	-1	-1	0	-1	-1	-1	-1	-1	-1	-1	-3	
	112	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	0	0	0	0	0	0	
	113	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-2	
	114	2	0	0	0	0	0	0	0	0	-1	0	-1	-1	0	-1	1	-1	0	-1	0	-1	-1	-1	-1	-1	
	115	2	0	0	0	0	0	-1	-1	-1	-1	0	-1	-1	0	0	0	0	0	0	0	0	0	0	0	-1	
	116	2	0	0	0	0	0	0	0	-1	-1	0	0	-1	-1	-2	-2	0	-1	0	0	0	0	0	0	-1	
Class 2 - Possibly Multi-Isoform Protein Cluster	201	2	0	0	0	0	0	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	-1	-1	-1	-1	-2		
	202	1	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	0	1	0	0	0	0		
	203	1	0	0	0	0	0	0	0	-1	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1		
	204	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	1	0	1	0	0		
	205	2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	-1	-1	0	-1	-2		
	206	2	1	1	1	1	1	0	-1	0	-1	-1	-1	-1	3	-1	-1	0	1	0	-1	0	-1	-2	-1		
	207	2	0	0	0	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	0	0	-1	-1	-2		
	208	1	0	0	0	0	0	0	0	0	-1	-1	0	0	-1	-1	-1	-1	0	0	1	0	-1	-1	-1		
	209	1	0	0	0	-1	0	0	0	0	0	0	0	1	-1	-1	-1	0	-1	1	1	1	0	1	0		
	210	1	1	1	1	0	1	1	0	-1	-1	0	-1	-1	-1	-1	-1	-1	0	0	-1	-1	-1	-1	-1		
211	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0			
301	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0			
302	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0			
303	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0			
304	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0			
305	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0			
306	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0			
307	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0			
308	1	0	0	-1	0	0	0	0	0	-1	1	0	0	0	2	0	0	-1	1	0	0	0	0	-1			
309	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0			
310	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0			
311	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0			
312	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0			
313	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0			
314	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0			

Cell value of 1 means one too many detections (3) were reported for the cluster, meaning one false positive.



The Table...

Proteome Informatics
Research Group

			29850	95560	19731	56365	Seymour	97416	71415	20081	12114	13036	Searle	62717	Kowalak	22573	36576	39485	52666	35340	Nesvizhskii	Tabb	08067	86010	Lane	19609	
Research Group/Vendor/Builder(non-vendor)/User			U	U	U	U	RG	V	U	U	U	U	RG	U	RG	U	U	V	U	B	RG	RG	?	B	RG	B	
Total Class 1,2,3 False Positives			4	4	4	2	2	1	0	0	1	3	0	1	3	5	3	0	9	11	4	9	12	2	0	0	
Total Class 1,2,3 False Negatives			4	4	5	8	10	21	32	57	66	58	41	91	70	98	94	49	78	16	6	32	30	56	77	189	
Protein Inference Tool			PG	PG	PG	PG	PG	PG	PG	PG	PG	Sc	Sc	Sc	MS	MS	MS	M	M	C	PP	ID	?	pC	PBS	SW	
Peptide ID Tool			P	P	P	P	P	P	P	P	P	M	M	M	M	M	M	M	M	B	C	My	?	pF	S	XI, Pe	
Total Reported Proteins			364	364	355	326	325	291	256	212	200	235	224	168	191	176	167	218	203	347	327	242	276	207	181	74	
Total Class 1,2,3 True Positives			254	254	253	250	248	237	226	201	192	200	217	167	188	160	164	209	180	242	252	226	228	202	181	69	
Cluster Class	Cluster Index	Expected # Detections																									
Class 1 - Consensus Multi-Isoform Protein Clusters	101	14	0	0	0	0	0	-1	-2	-3	-2	-1	-2	-5	-2	-4	-2	-3	-1	1	1	2	1	-2	-2	6	
	102	7	0	0	0	0	0	-1	-1	-1	-1	-1	0	-1	1	-1	0	0	2	-1	0	1	2	0	-1	4	
	103	5	0	0	0	0	0	0	0	0	0	-1	-1	-1	0	-1	0	-2	0	0	0	0	0	0	-2	3	
	104	4	0	0	0	0	0	0	-1	0	0	-1	0	-1	-1	-1	-1	-1	-1	0	0	0	0	0	-1	-1	
	105	3	-1	-1	-1	-1	-1	-2	-1	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	0	0	1	-1	-2	-2	-2	
	106	4	0	0	0	0	-1	-1	-1	-1	-1	-1	-1	-2	0	-2	-1	0	-1	0	0	0	0	3	-1	-2	
	107	3	0	0	0	0	0	-1	-1	-1	-1	0	-1	0	-1	0	-1	0	1	2	-1	-2	-1	0	0	-1	-2
	108	3	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	-2	-1	-1	-1	-2	0	-1	0	0	0	-1	-1	-1	
	109	2	1	1	1	1	0	0	0	-1	1	-1	-1	0	1	-1	-1	-1	1	2	-1	0	3	-1	-1	-2	
	110	2	0	0	0	0	0	-1	-1	-1	0	-1	-1	-1	0	-1	-1	-1	-1	-1	0	1	0	0	-1	-2	
	111	3	0	0	0	-1	0	-1	-1	-2	-1	-2	-2	-1	-2	-1	-3	-1	-1	0	-1	-1	-1	-1	-1	-3	
	112	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	0	0	0	0	0	0	0
	113	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-2
	114	2	0	0	0	0	0	0	0	-1	0	-1	-1	0	-1	1	-1	0	-1	0	-1	-1	-1	-1	-1	-1	-1
115	2	0	0	0	0	0	-1	-1	-1	-1	0	-1	-1	0	0	0	0	0	0	0	0	0	0	0	0	-1	
116	2	0	0	0	0	0	0	0	-1	-1	0	0	-1	-1	-2	-2	0	-1	0	0	0	0	0	0	0	-1	
Class 2 - Possibly Multi-Isoform Protein Clusters	201	2	0	0	0	0	0	0	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	-1	-1	-1	-1	-2	
	202	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	1	0	0	1	0	0	0	0	
	203	2	0	0	0	0	0	0	0	-1	0	-1	-1	-1	-1	-1	-1	-1	0	-1	-1	-1	-1	-1	-1	-1	
	204	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	1	1	0	1	2	1	0	0	
	205	2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	-1	-2	
	206	2	1	1	1	1	1	0	-1	0	0	-1	-1	-1	-1	3	-1	-1	0	1	0	-1	0	-1	-2	-1	
	207	2	0	0	0	0	0	0	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	0	0	-1	-1	-2	
209	1	0	0	0	0	0	0	0	0	-1	-1	0	0	-1	-1	-1	0	-1	0	0	1	0	-1	-1	-1		
210	1	1	1	1	0	1	1	0	-1	-1	0	-1	-1	-1	-1	-1	-1	-1	0	0	-1	-1	-1	-1	-1		
211	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0		
301	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
302	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
303	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
304	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	
305	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
306	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	
307	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
308	1	0	0	-1	0	0	0	0	0	-1	1	0	0	0	2	0	0	0	-1	1	0	0	0	0	0	-1	
309	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
310	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
311	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
312	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
313	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	
314	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

Cell value of 0 means the expected number of detections (2) were reported for the cluster.



The Table...

*Proteome Informatics
Research Group*

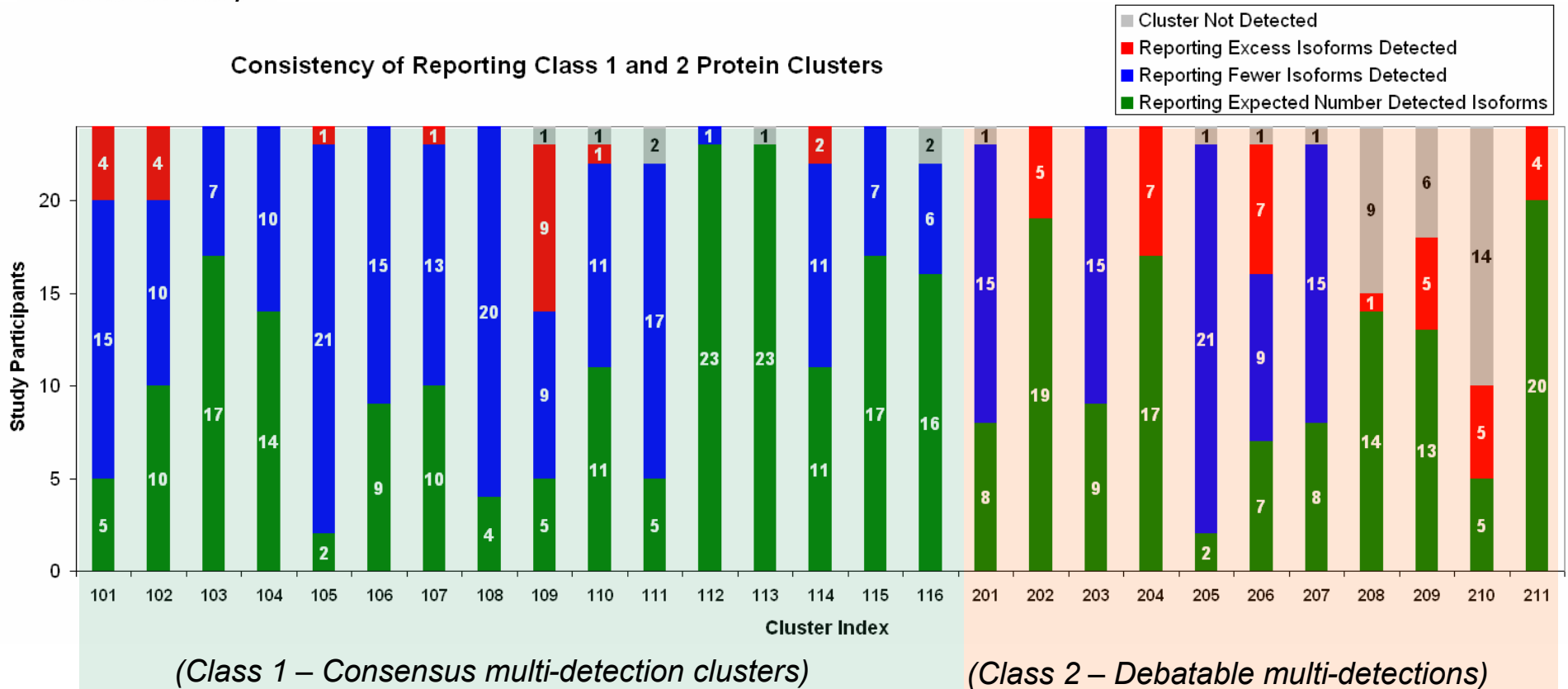
			29850	95560	19731	56365	Seymour	97416	71415	20081	12114	13036	Searle	62717	Kowalak	22573	36576	39485	52666	35340	Nesvizhskii	Tabb	08067	86010	Lane	19609	
Research Group/Vendor/Builder(non-vendor)/User			U	U	U	U	RG	V	U	U	U	U	RG	U	RG	U	U	V	U	B	RG	RG	?	B	RG	B	
Total Class 1,2,3 False Positives			4	4	4	2	2	1	0	0	1	3	0	1	3	5	3	0	9	11	4	9	12	2	0	0	
Total Class 1,2,3 False Negatives			4	4	5	8	10	21	32	57	66	58	41	91	70	98	94	49	78	16	6	32	30	56	77	189	
Protein Inference Tool			PG	PG	PG	PG	PG	PG	PG	PG	Sc	Sc	Sc	MS	MS	MS	M	M	C	PP	ID	?	pC	PBS	SW		
Peptide ID Tool			P	P	P	P	P	P	P	P	M	M	M	M	M	M	M	M	B	C	My	?	pF	S	XI, Pe		
Total Reported Proteins			364	364	355	326	325	291	256	212	200	235	224	168	191	176	167	218	203	347	327	242	276	207	181	74	
Total Class 1,2,3 True Positives			254	254	253	250	248	237	226	201	192	200	217	167	188	160	164	209	180	242	252	226	228	202	181	69	
Cluster Class	Cluster Index	Expected # Detections																									
Class 1 - Consensus Multi-Isoform Protein Clusters	101	14	0	0	0	0	0	-1	-2	3	-2	-1	-2	-5	-2	-4	-2	-3	-1	1	1	2	1	-2	-2	6	
	102	7	0	0	0	0	0	-1	-1	-1	-1	-1	0	-1	1	-1	0	0	2	-1	0	1	2	0	-1	4	
	103	5	0	0	0	0	0	0	0	0	0	-1	-1	-1	0	-1	0	-2	0	0	0	0	0	0	-2	-3	
	104	4	0	0	0	0	0	0	-1	0	0	-1	0	-1	-1	-1	-1	-1	-1	0	0	0	0	0	-1	-1	
	105	3	-1	-1	-1	-1	-1	-2	-1	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	0	0	1	-1	-2	-2	-2	
	106	4	0	0	0	0	-1	-1	-1	-1	-1	-1	-1	-2	0	-2	-1	0	-1	0	0	0	0	3	-1	-2	
	107	3	0	0	0	0	0	-1	-1	-1	-1	-1	0	-1	0	-1	0	-1	2	-1	-2	-1	0	0	-1	-2	
	108	3	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	-2	-1	-1	-1	-2	0	-1	0	0	0	-1	-1	-1	
	109	2	1	1	1	1	0	0	0	-1	1	-1	-1	0	1	-1	-1	-1	1	2	-1	0	3	-1	-1	-2	
	110	2	0	0	0	0	0	-1	-1	-1	0	-1	-1	-1	0	-1	-1	-1	-1	0	1	0	0	-1	-1	-2	
	111	3	0	0	0	-1	0	-1	-1	-2	-2	-1	-2	-2	-1	-3	-1	-1	0	-1	-1	-1	-1	-1	-1	-3	
	112	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	0	0	0	0	0	0	0	
	113	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-2	
	114	2	0	0	0	0	0	0	0	-1	0	-1	-1	0	-1	1	-1	0	-1	0	-1	-1	-1	-1	-1	-1	
	115	2	0	0	0	0	0	-1	-1	-1	-1	-1	-1	-1	0	-1	-1	0	0	0	0	0	0	0	0	-1	
	116	2	0	0	0	0	0	0	0	-1	-1	0	0	-1	-1	-2	-2	0	-1	0	0	0	0	0	0	-1	
Class 2 - Possibly Multi-Isoform Protein Clusters	201	2	0	0	0	0	0	0	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	-1	-1	-1	-1	-2	
	202	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	1	0	0	1	0	0	0	0	
	203	2	0	0	0	0	0	0	0	-1	0	-1	-1	-1	-1	-1	-1	-1	0	-1	-1	-1	-1	-1	-1	-1	
	204	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	1	0	1	1	0	1	2	1	0
	205	2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	-1	-1	0	-1	-2
	206	2	1	1	1	1	1	0	-1	0	-1	-1	-1	-1	3	-1	-1	0	1	0	-1	0	-1	-2	-1	-1	
	207	2	0	0	0	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	0	0	0	-1	-1	-2	
	208	1	0	0	0	0	0	0	0	0	-1	-1	0	0	-1	-1	-1	-1	-1	0	0	1	0	-1	-1	-1	
	209	1	0	0	0	0	0	0	0	0	0	0	0	1	-1	-1	-1	0	-1	1	1	1	0	1	0	-1	
	210	1	0	0	0	0	0	0	0	0	-1	-1	0	-1	-1	-1	-1	-1	-1	0	0	-1	-1	-1	-1	-1	
	211	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	
Class 3 - Single Isoform Protein Clusters	301	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
	302	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
	303	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
	304	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	0	0	0	0	0	-1	
	305	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	
	306	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	
	307	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
	308	1	0	0	-1	0	0	0	0	0	-1	1	0	0	0	2	0	0	-1	1	0	0	0	0	0	-1	
	309	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
	310	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
	311	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
	312	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
	313	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	-1	
314	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		

Cell value of -1 means 1 fewer than the maximal allowed detections, thus one false negative.



Consistency in Multi-Detection Clusters

*Proteome Informatics
Research Group*

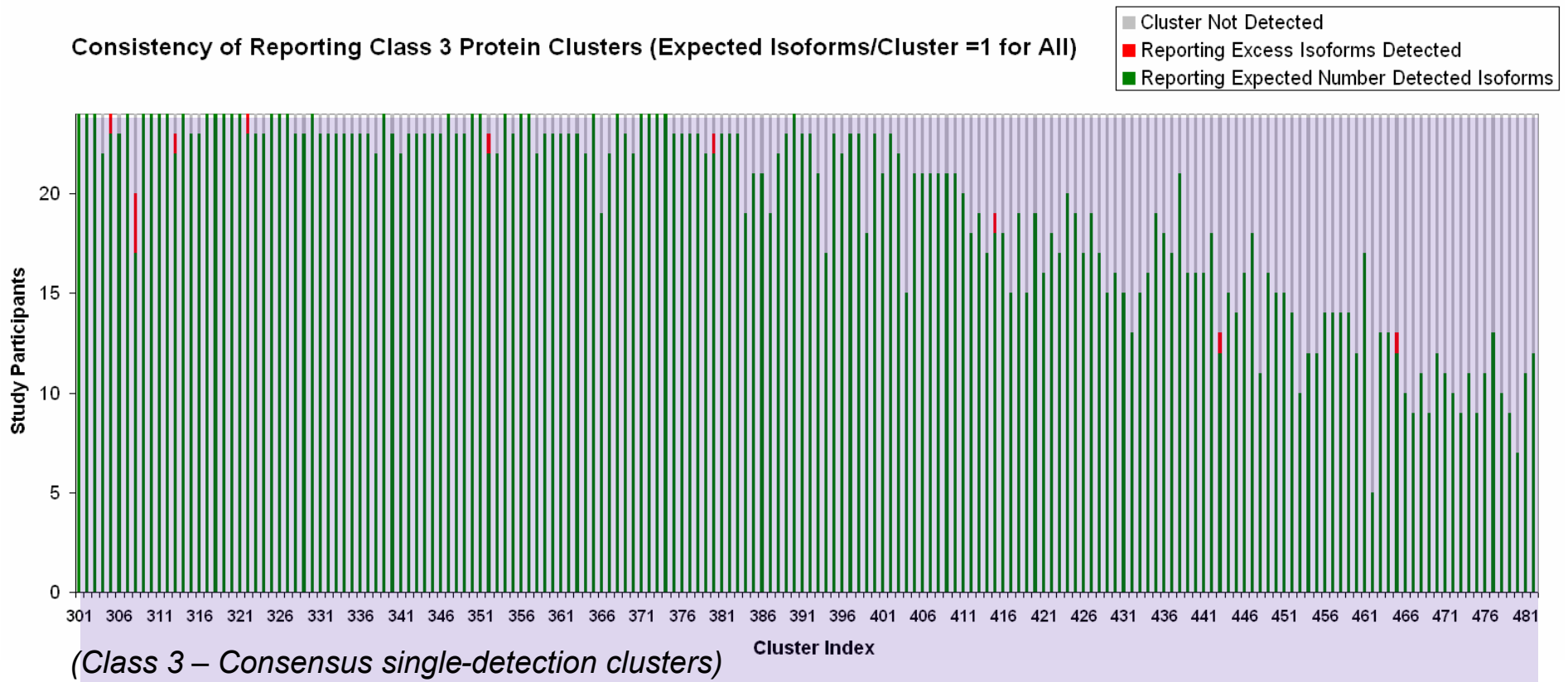


- More blue than red means more people under-reported than over-reported.



Proteome Informatics
Research Group

No Gross Over-Reporting due to Protein Inference Problems



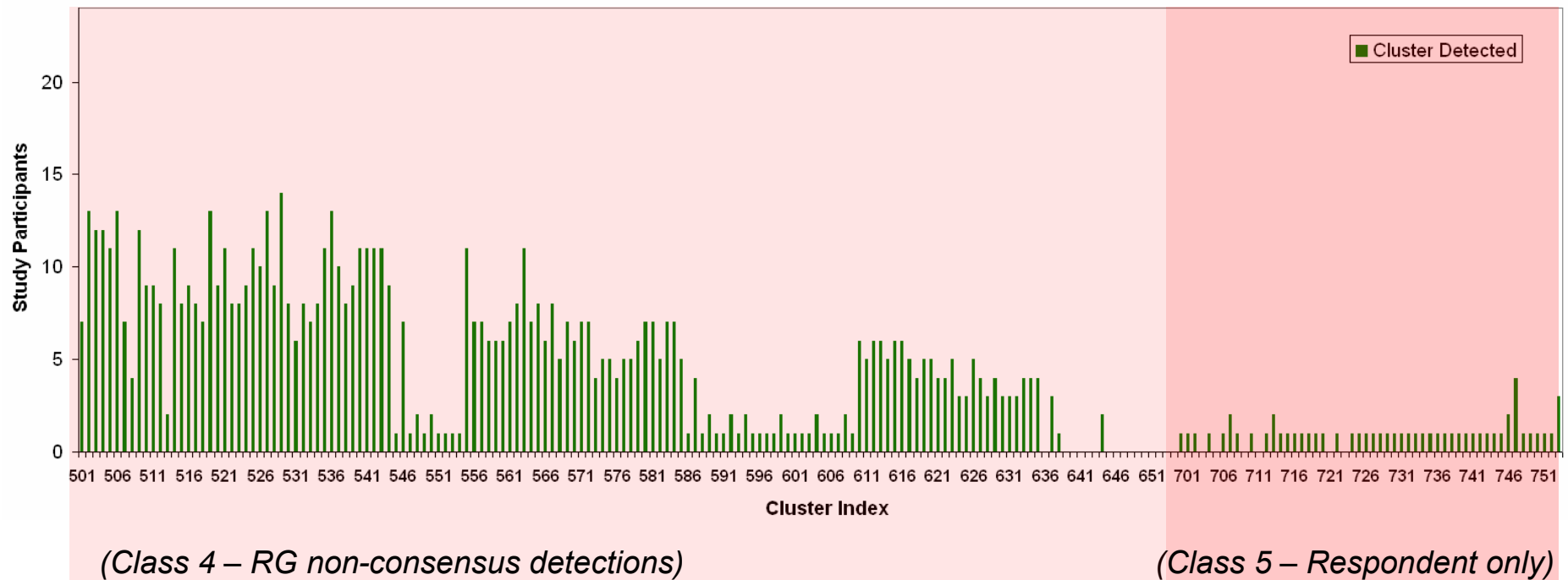
- These were effectively bait for reporting excessive isoforms.
- Very little excessive reporting due to inference errors was seen!



Clusters Not Graded were Less Consistent

*Proteome Informatics
Research Group*

Consistency of Reporting Class 4 and 5 Protein Clusters (Classes Not Graded, thus No Expected Value)



- **Protein clusters that were not graded (Class 4 and 5) were much less consistently detected by study participants.**



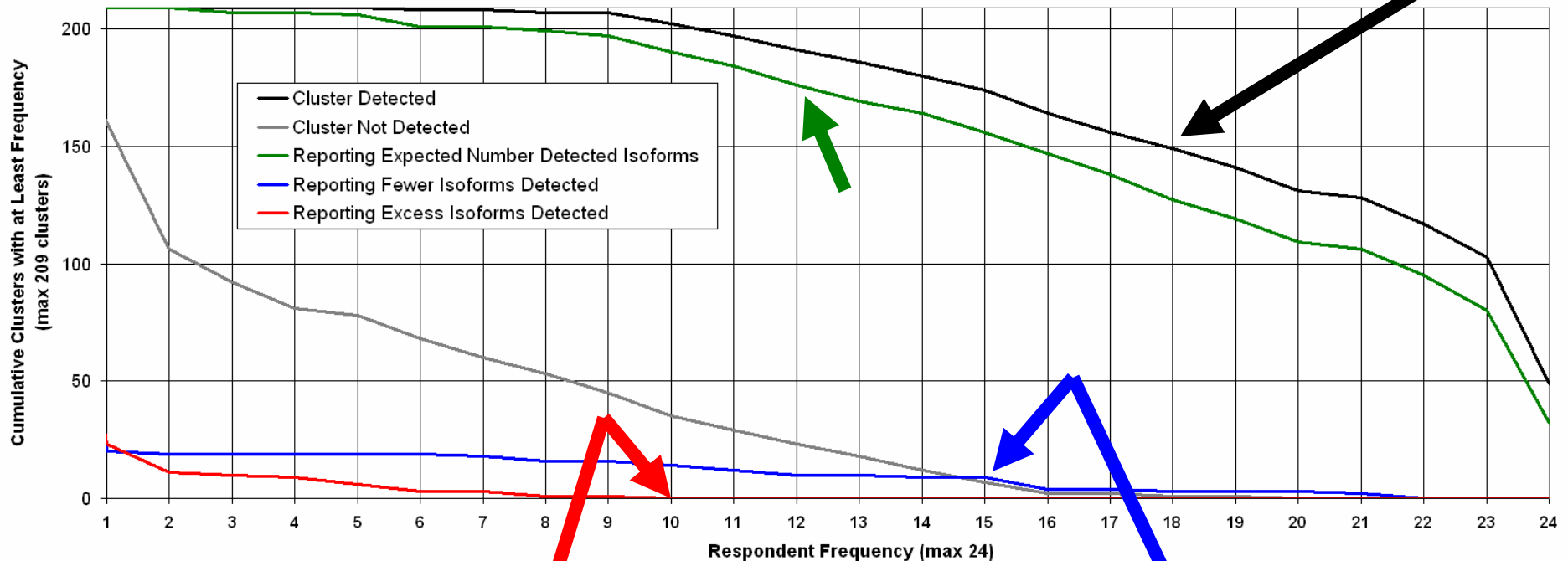
Consistency of Results

Proteome Informatics
Research Group

Example interpretation 4: For 176 of the 209 graded protein clusters (84%), at least half (12 or more) study participants reported the expected number of isoforms.

Example interpretation 1: 150 of the 209 graded protein clusters were detected by 18 or more study participants

Consistency of Reporting across All Participants - All Graded Protein Clusters (Class 1,2, and 3)



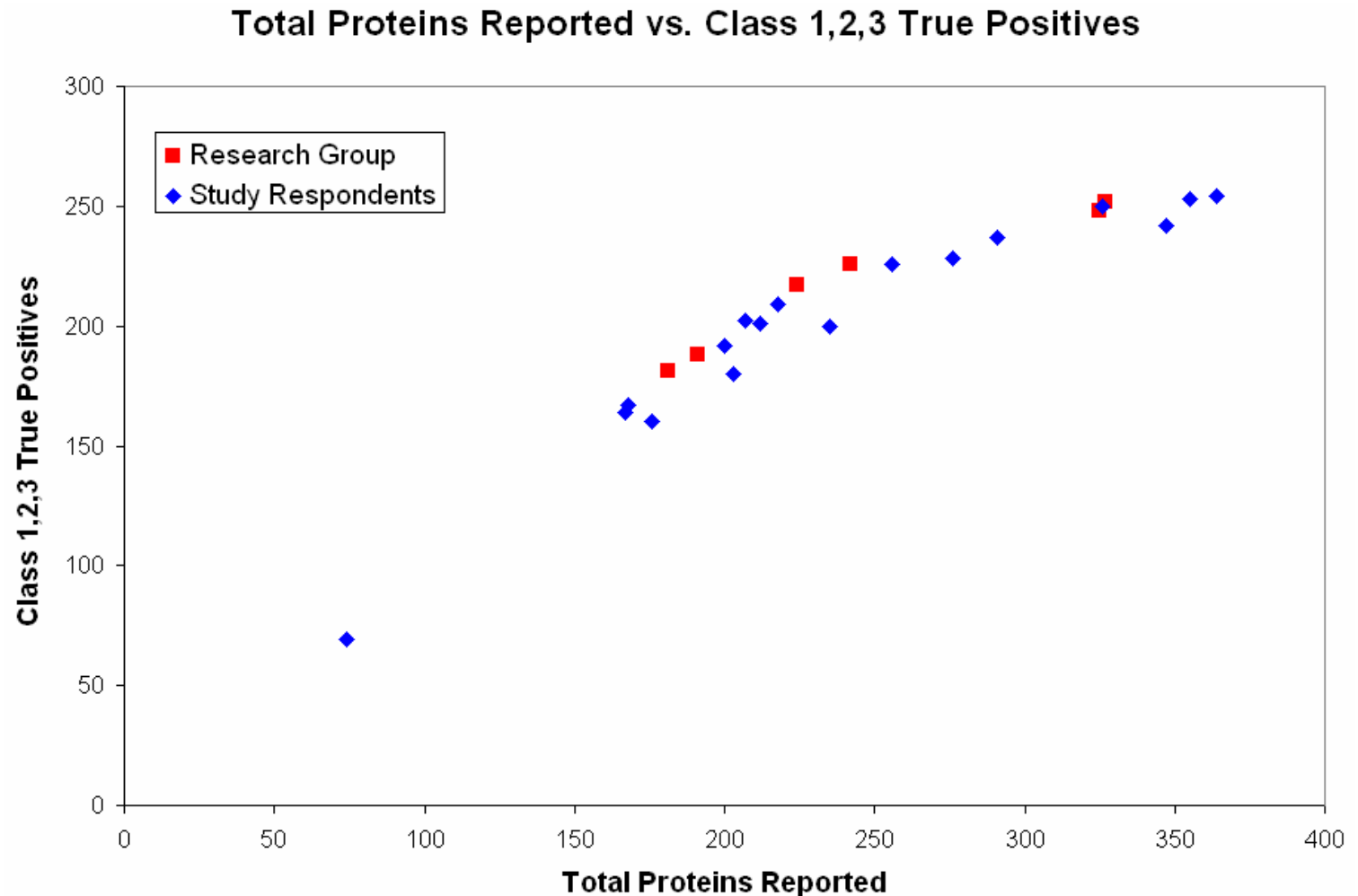
Example interpretation 3: There were no graded protein clusters where 10 or more study participants reported an excessive number of isoforms.

Example interpretation 2: For only 9 of the 209 graded protein clusters did 15 or more study participants report fewer than the allowed number of isoforms.



*Proteome Informatics
Research Group*

The Effect of Protein List Size



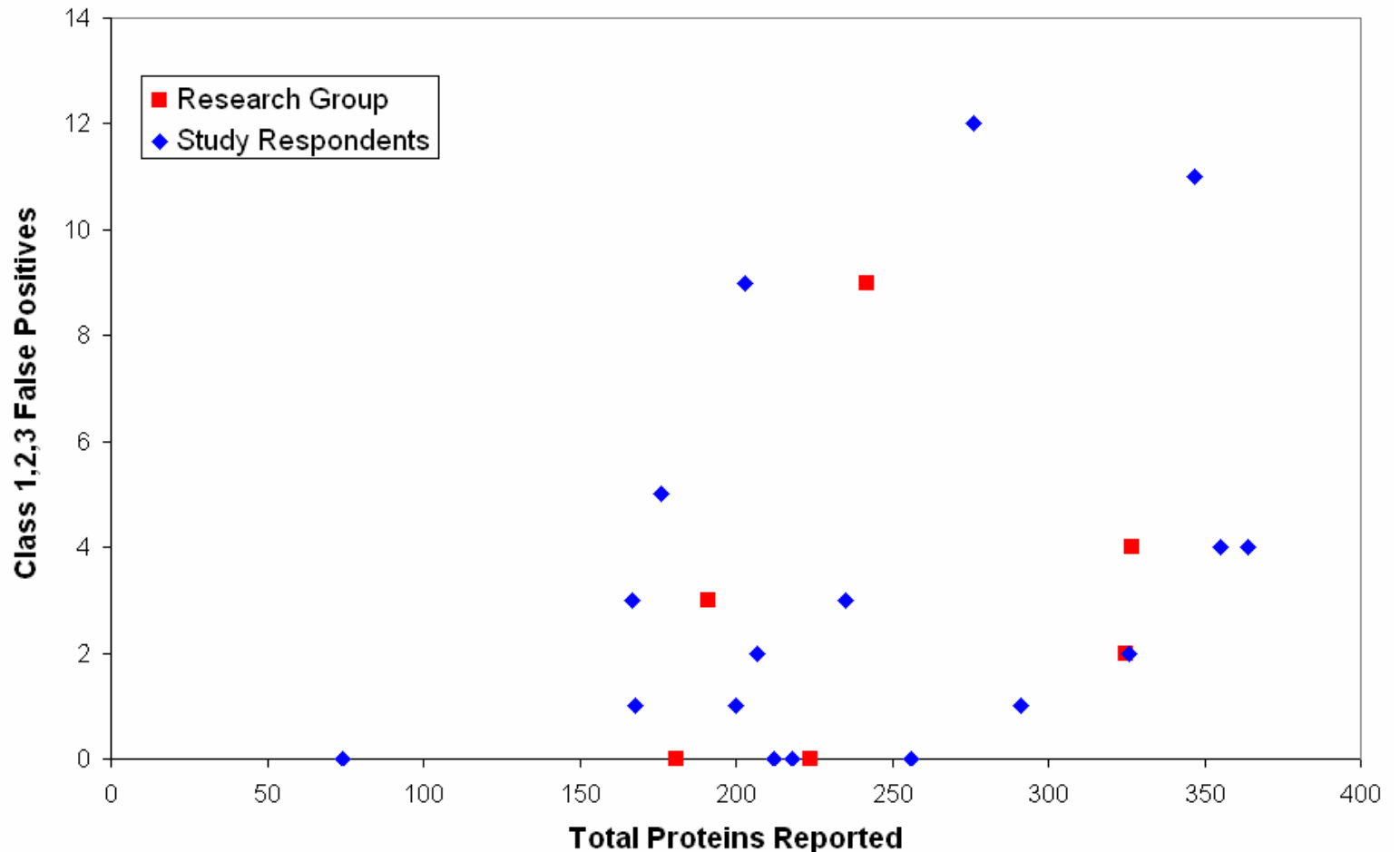
- **As expected, people reporting longer overall protein lists also got more true positives in the subset of clusters that were graded.**



*Proteome Informatics
Research Group*

The Effect of Protein List Size

Total Proteins Reported vs. Class 1,2,3 False Positives

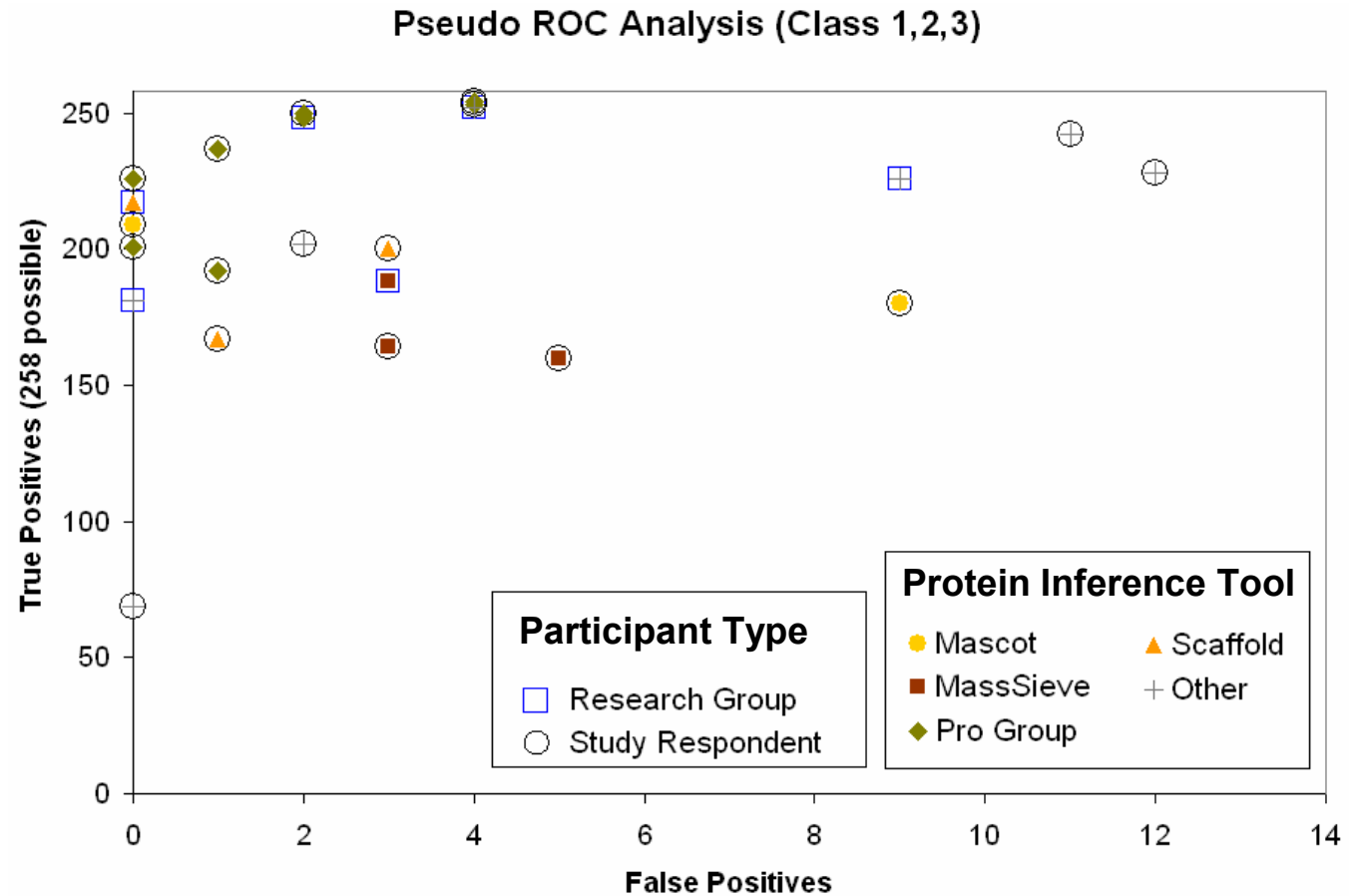


- **The same relationship does not hold with false positives because there are real differences in quality of results.**



Proteome Informatics
Research Group

False Positives vs. True Positives



- Closer to the top left is generally better (limit of 258 correct, 0 wrong).
- Even with the same software, there are clear differences in where people set thresholds.
- Some prefer virtually no errors while others will tolerate a few errors to get longer lists.
- There are also some real differences in quality of results.



*Proteome Informatics
Research Group*

Reporting of Accession Ambiguity

The Paris Guidelines:

“The apparent ambiguity in peptide assignment requires reporting of a protein group.”

“Authors should explain and be able to justify cases where a single protein from a protein group has been singled out or that more than one member of a protein group is present.”

The implication of this is that if a detected protein cannot be resolved to a single accession number, the ambiguity among several accessions should be reported.



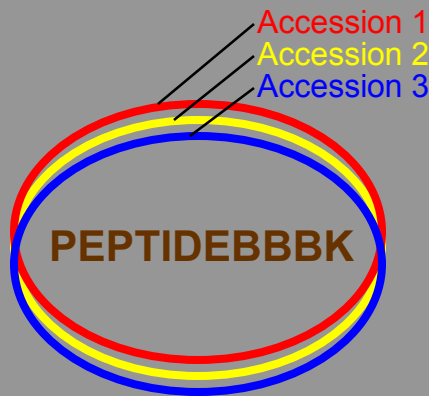
Why Report Accession Ambiguity?

Proteome Informatics
Research Group

**Protein
Cluster:**

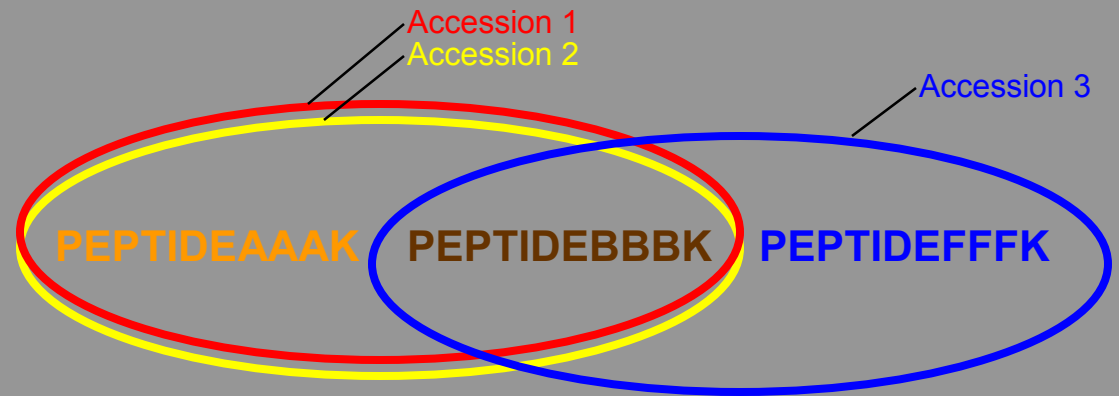
Accession 1	AAA	PEPTIDEAAAK	AAPEPTIDE	BBBKAA	PEPTIDE	DDDKAA
Accession 2	AAAAA	PEPTIDEAAAK	AAPEPTIDE	BBBKAAAA	PEPTIDE	EEEEK
Accession 3	AAAAAAAAAAAAAAAAAAAA	PEPTIDE	BBBK	AAAAAA	PEPTIDE	FFFFKA

If we observe only peptide B:



1 detected protein species
ambiguity among 3 accessions

If we observe peptides A, B, and F:



2 detected protein species
ambiguity among 2 accessions in one detection

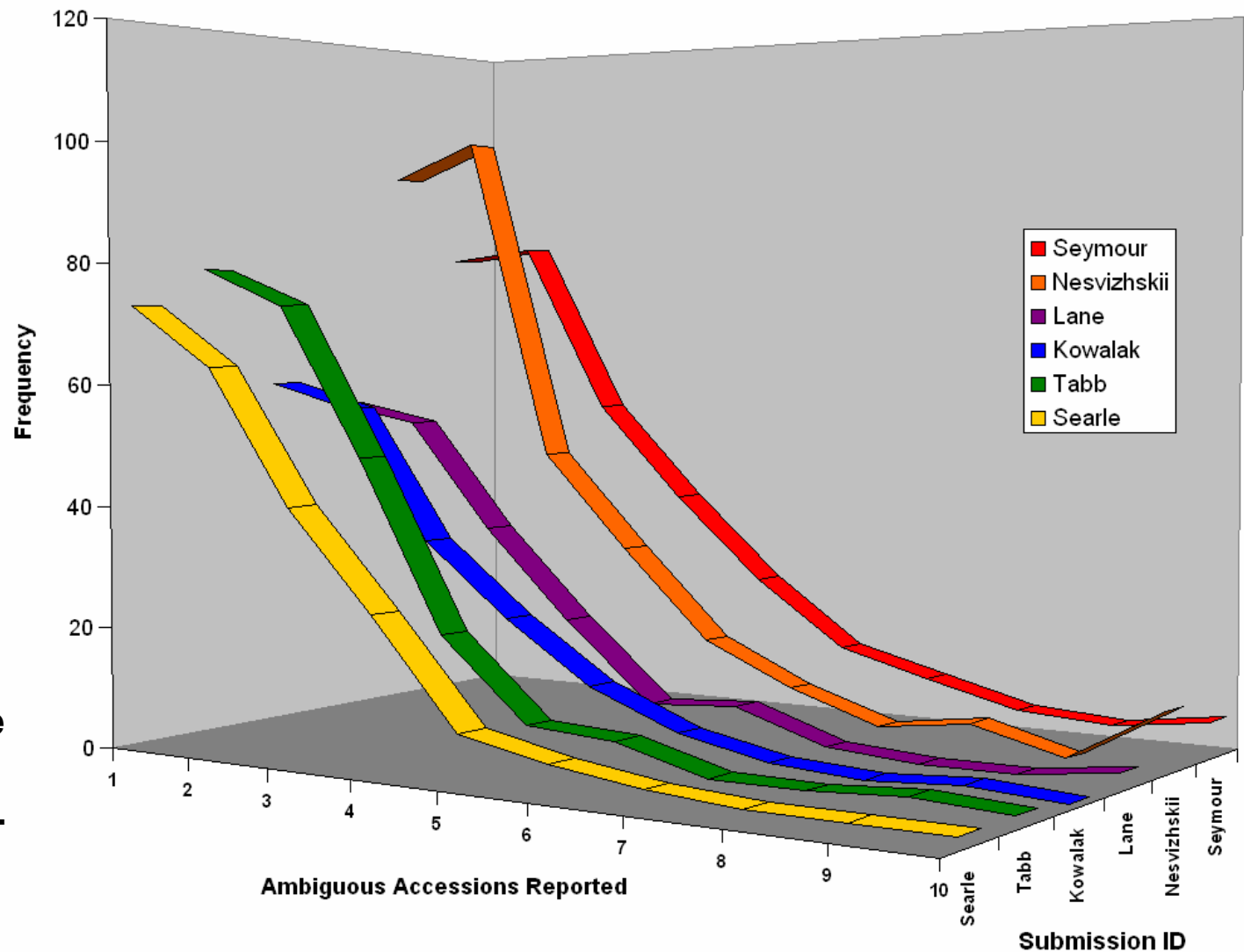
- 25 • What happens if we find peptide D in subsequent MS data acquisition?



Reporting of Accession Ambiguity

*Proteome Informatics
Research Group*

Reporting of Accession Ambiguity - Research Group



- The Research Group protein lists ranged from 2.5 to 3.7 average accessions per protein detection.

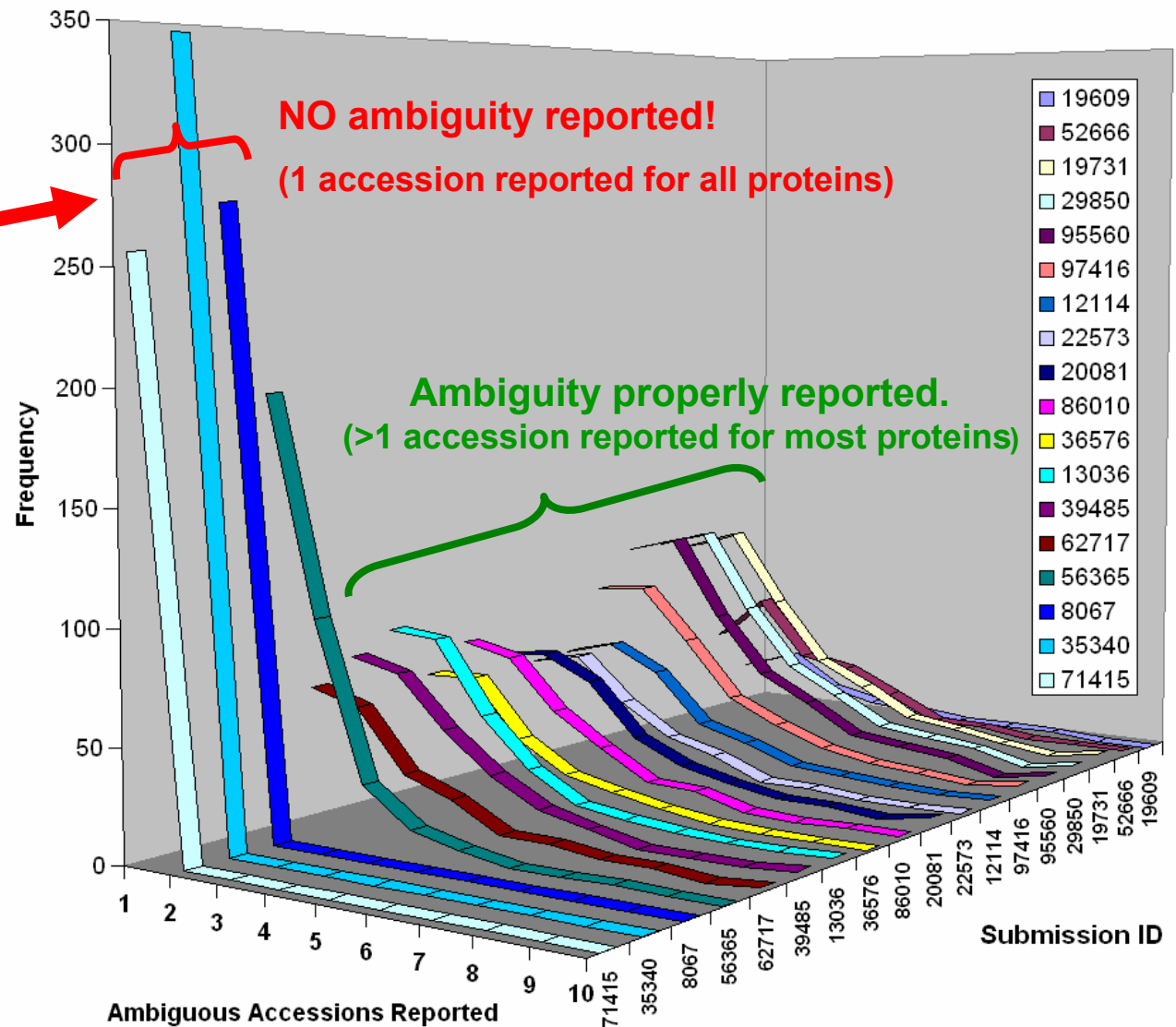


Reporting of Accession Ambiguity

Proteome Informatics
Research Group

- 3 respondents reported no accession ambiguity (NOT appropriate).
- Most respondents reported adequate accession ambiguity.
- A user's decision vs. software's output – both required for proper reporting.

Reporting of Accession Ambiguity - Study Respondents

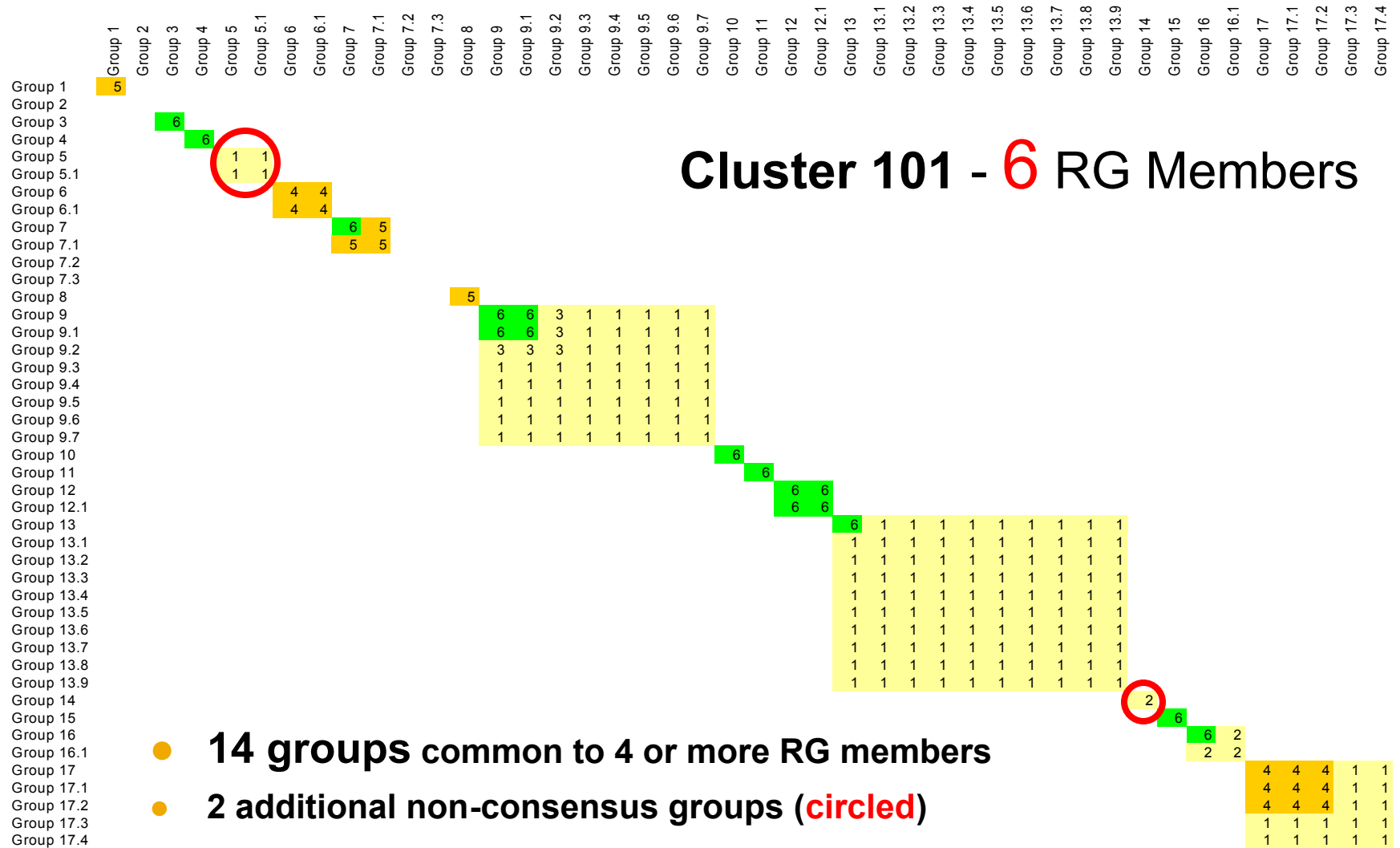




Consistency of Accession Grouping

*Proteome Informatics
Research Group*

- Sum across all 6 RG members to see consensus on groups:



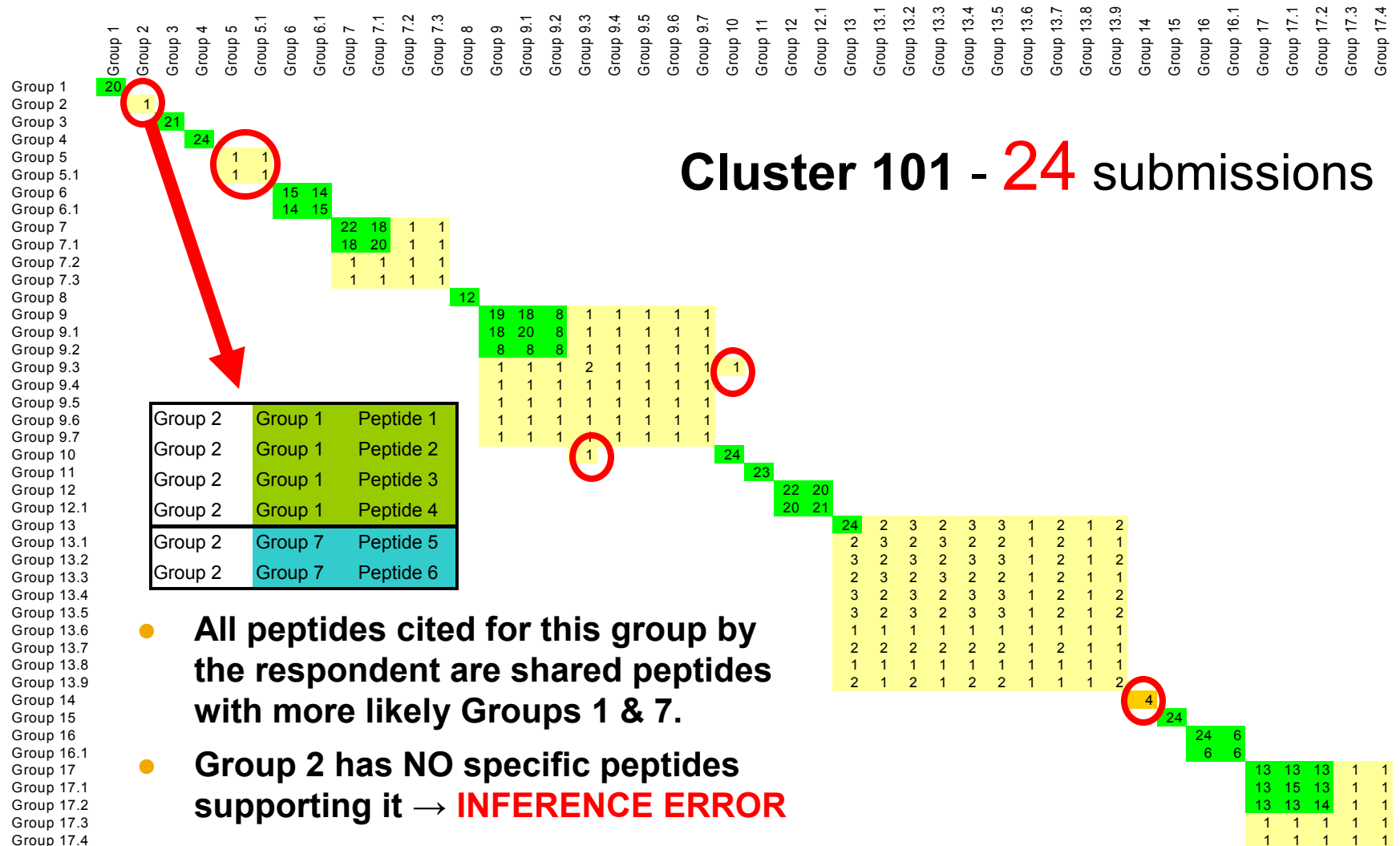
- 14 groups common to 4 or more RG members
- 2 additional non-consensus groups (circled)



An Example Protein Inference Error

Proteome Informatics
Research Group

- If we do this with ALL submissions, we can see some clear outliers:

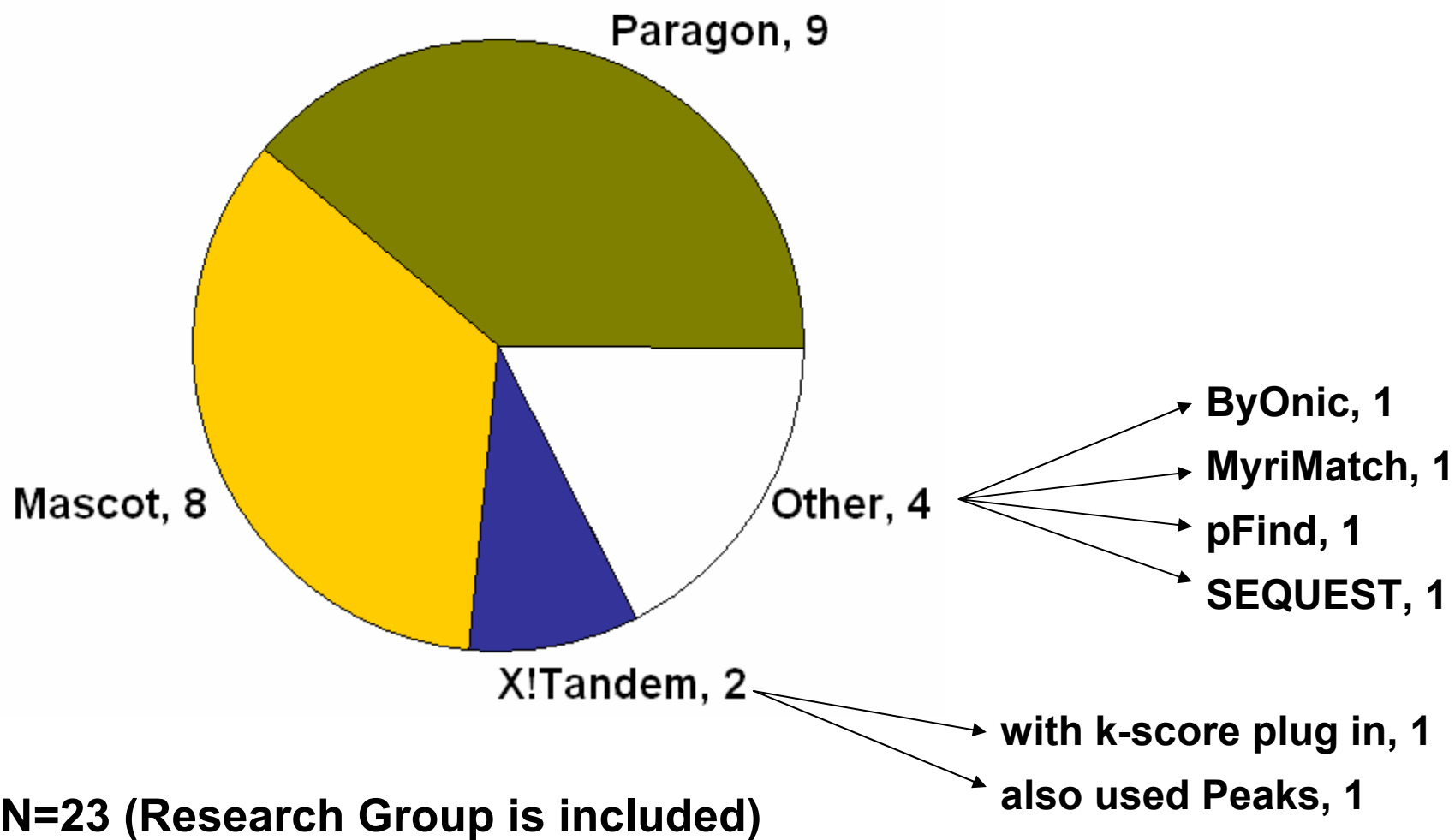


- All peptides cited for this group by the respondent are shared peptides with more likely Groups 1 & 7.
- Group 2 has NO specific peptides supporting it → **INFERENCE ERROR**



Survey: Peptide ID Software

Peptide Identification Tool Used

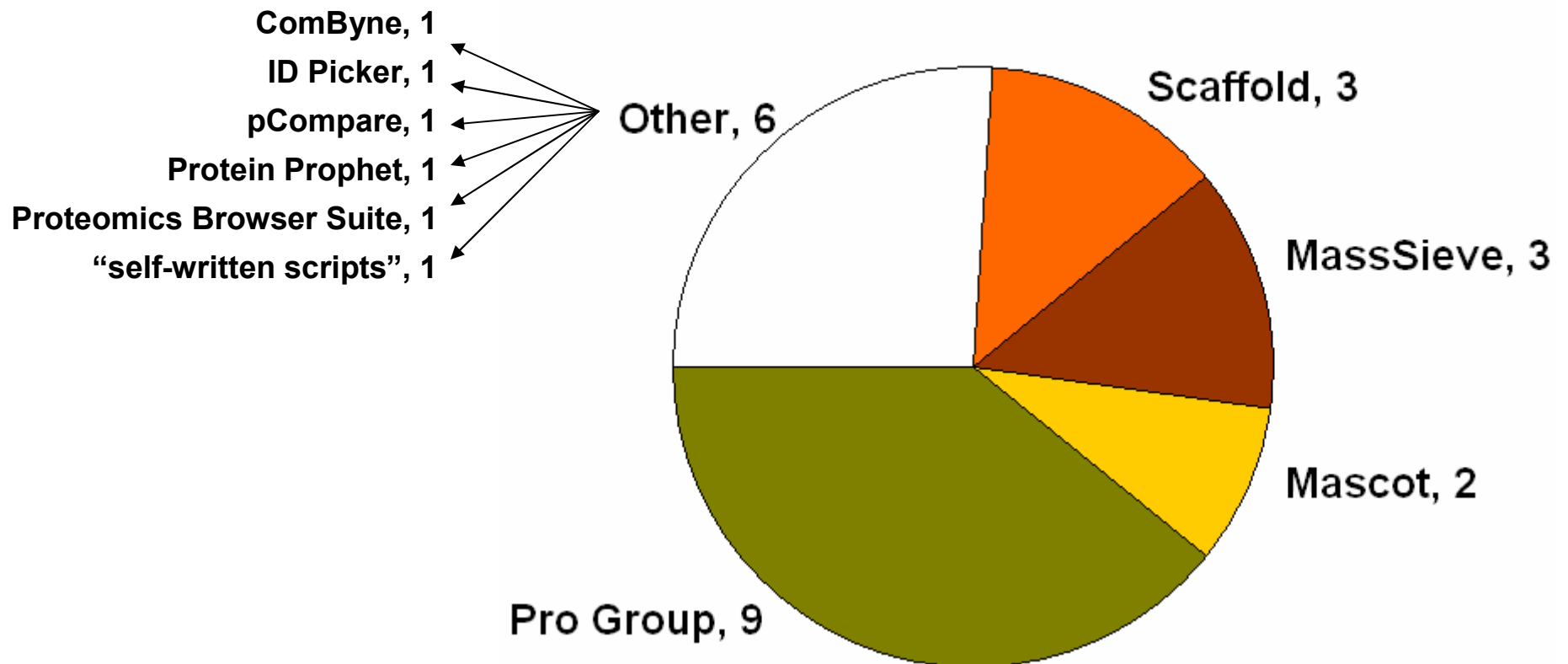




*Proteome Informatics
Research Group*

Survey: Protein Inference Software

Protein Inference Tool Used



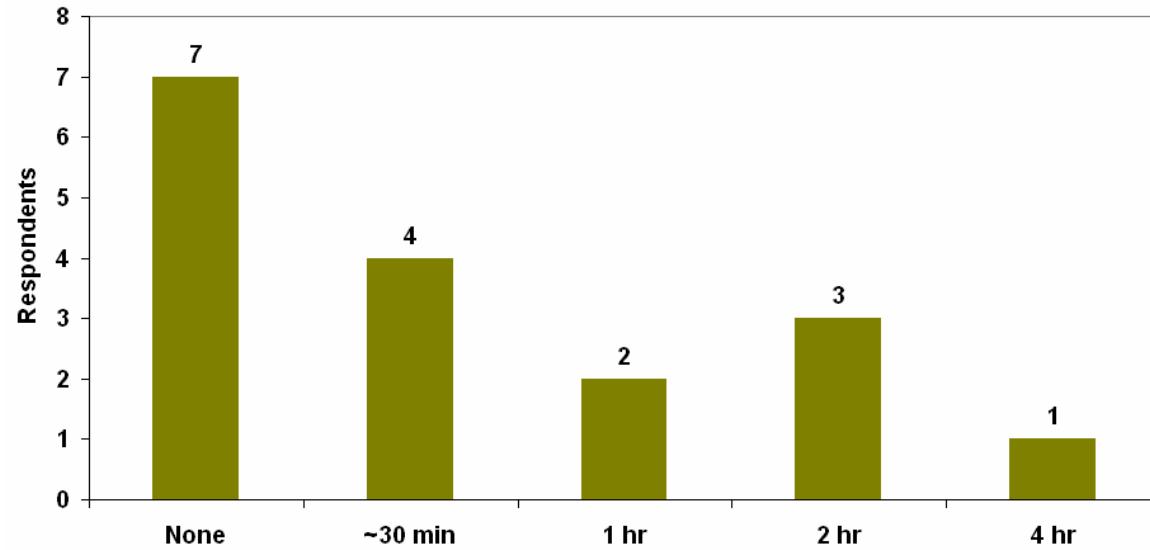
N=23 (Research Group is included)



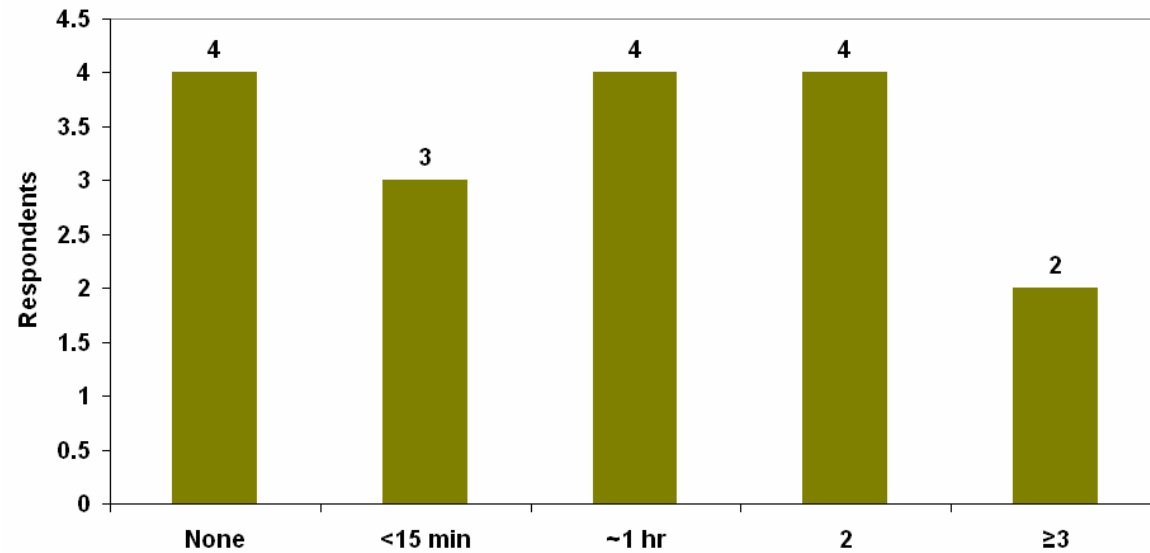
*Proteome Informatics
Research Group*

Survey: Time Inspecting IDs

Time Spent Inspecting Peptide Identifications



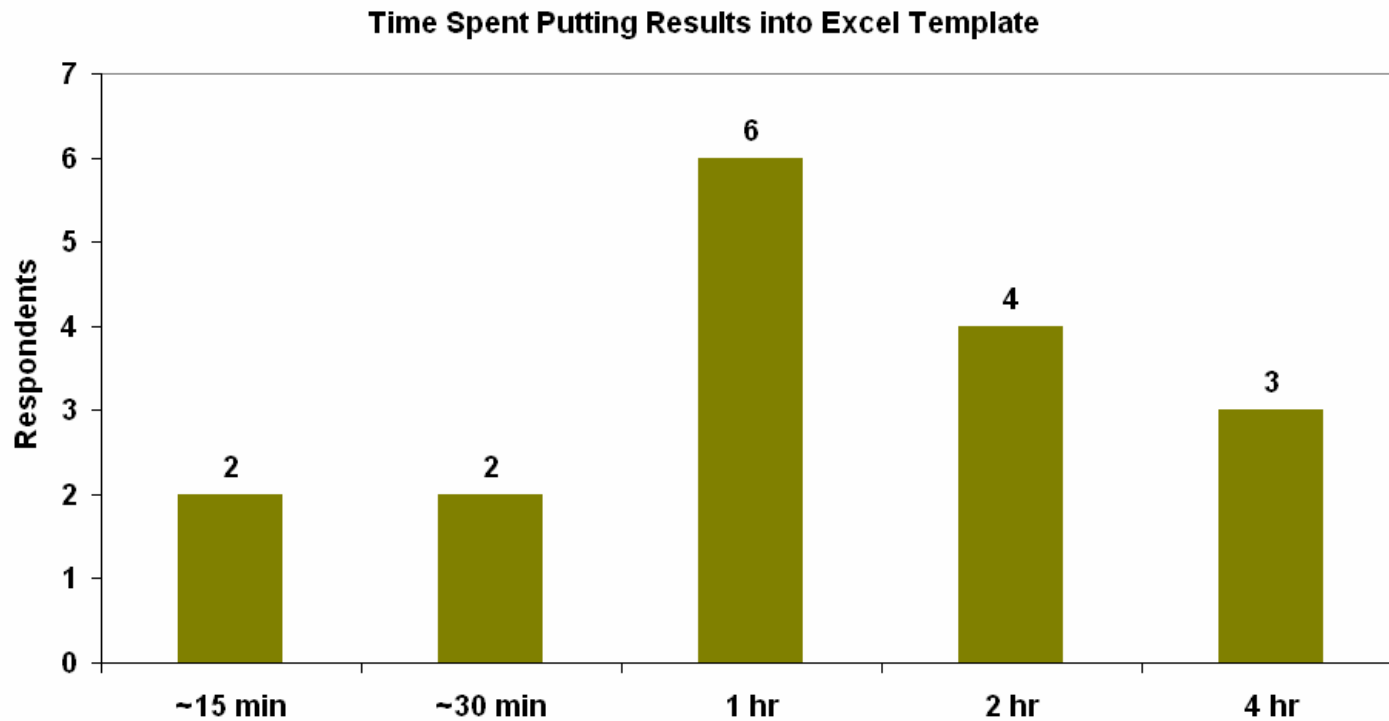
Time Spent Inspecting Protein Identifications





*Proteome Informatics
Research Group*

Survey: Time Formatting

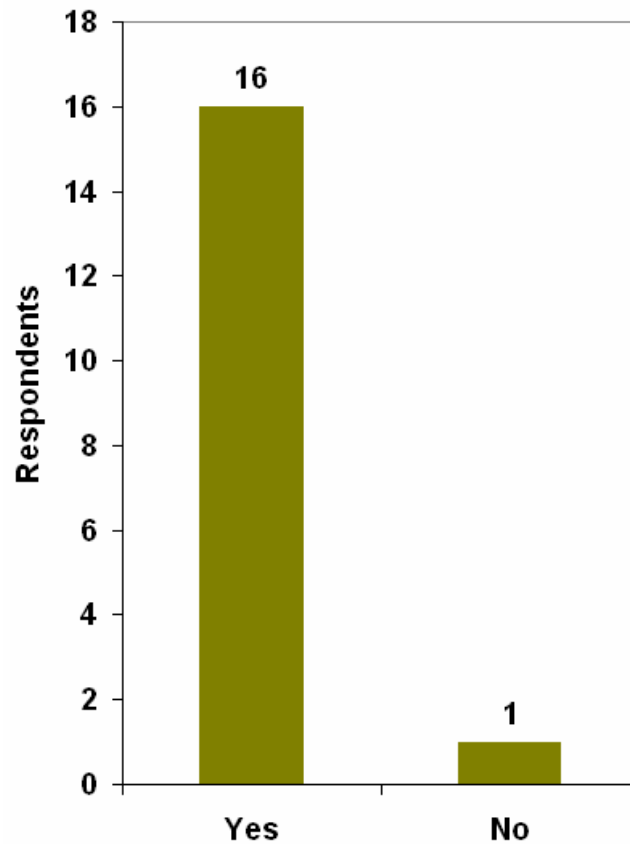




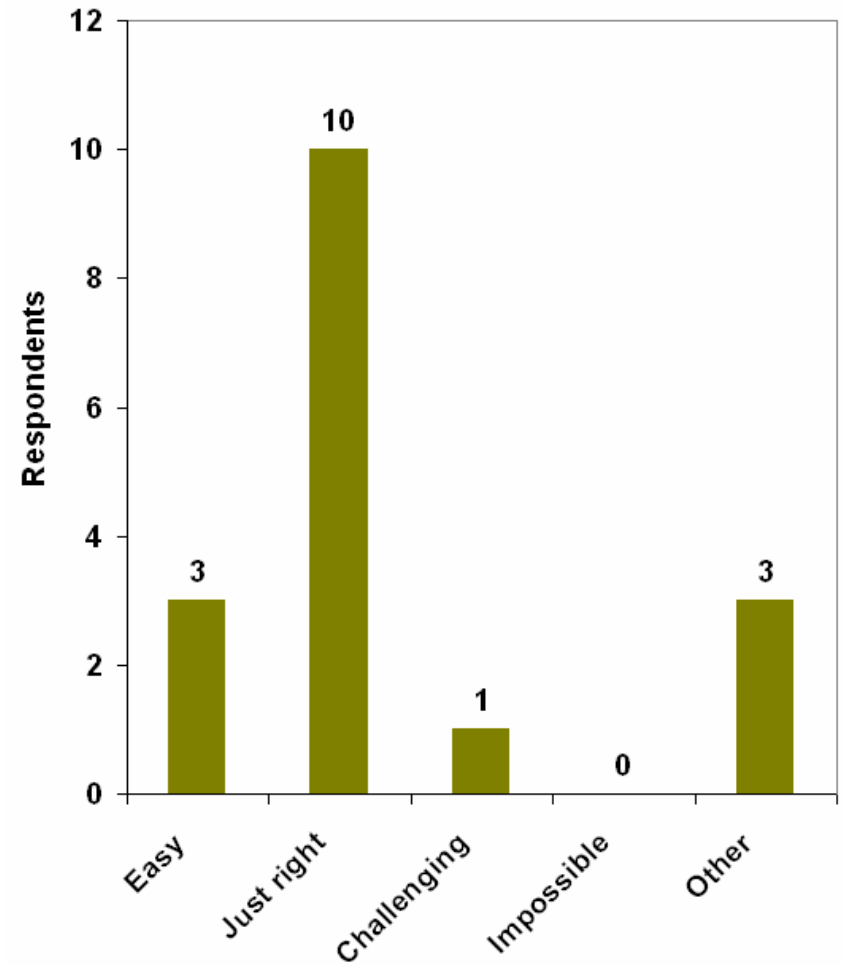
Survey: Opinions

*Proteome Informatics
Research Group*

Do you think this type of study is useful?



How would you rate this study's level of difficulty?

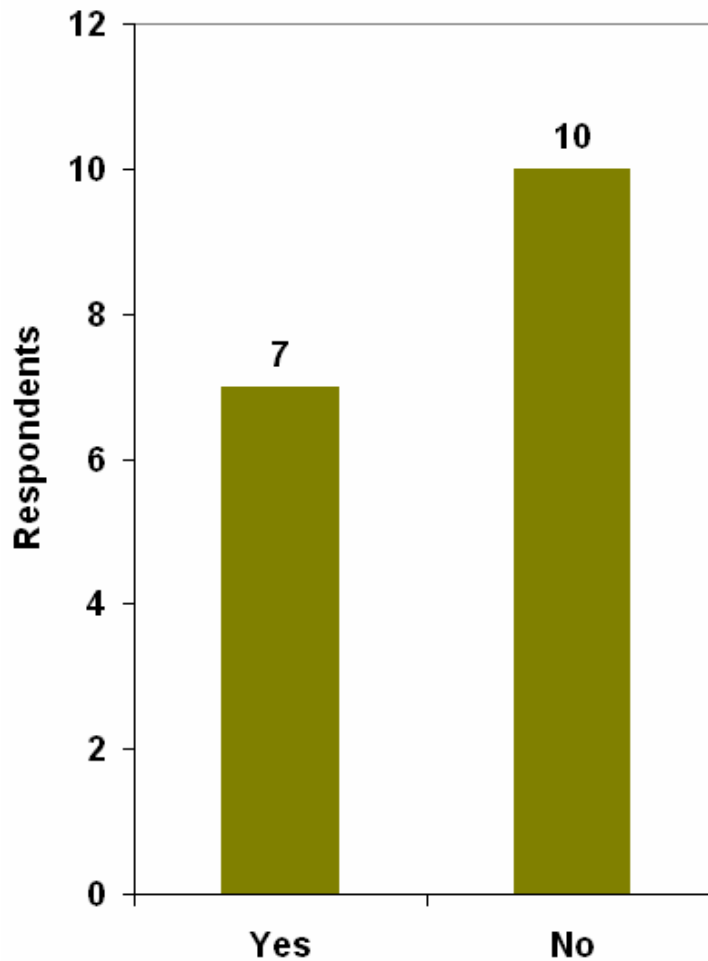




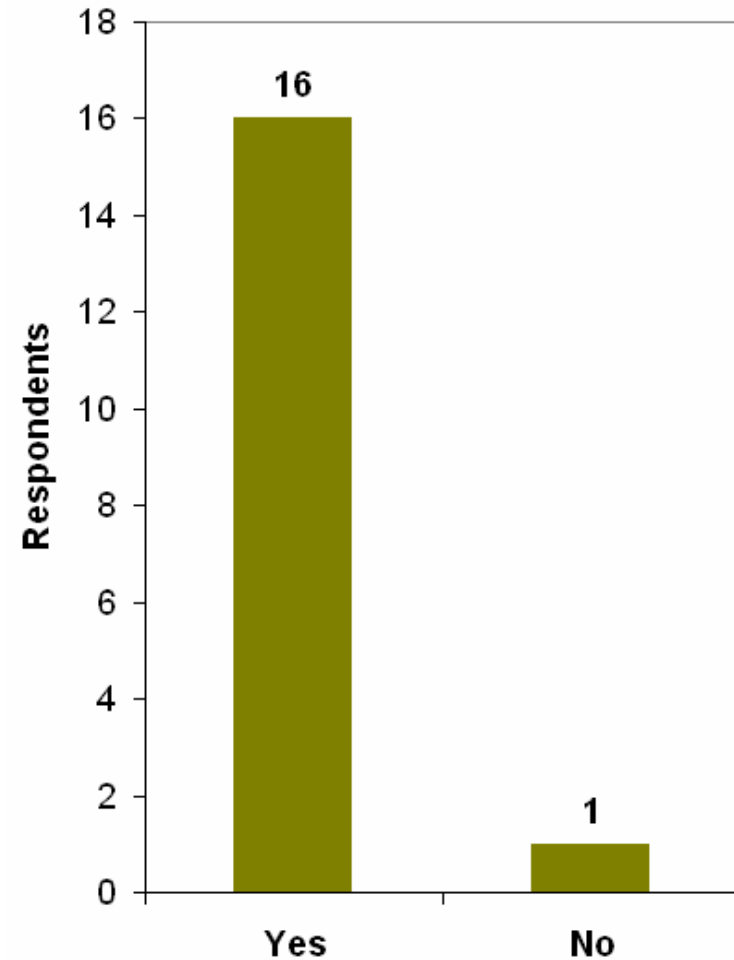
*Proteome Informatics
Research Group*

Survey: Opinions

Have you participated in previous ABRF studies?



Based on this study, would you consider participating in future ABRF studies?





*Proteome Informatics
Research Group*

Conclusions

- **Primary Goal: Assess the current state of protein reporting.**
 - (1) Does the proteomics community still have a problem with excessive numbers of proteins due to improper protein inference?
 - **Virtually all respondents used some type of protein inference software.**
 - **No gross inflation of protein numbers due to protein inference problems was seen in any respondent set.**
 - **Conclusions may not be representative of the whole proteomics community. The study sample is small and probably biased.**
 - **We ask that more groups submit an analysis of the data.**
 - (2) Are people reporting accession ambiguity as they should?
 - **Some respondent protein lists lacked proper reporting of ambiguity among multiple accessions for a detected protein.**
 - **Some of this is software user's decision to discard other accessions.**
 - **It is unknown if any of this results from inadequate software.**



*Proteome Informatics
Research Group*

Conclusions

- **Secondary Goal: Assess similarity of reporting given a common MS analysis results (rather than a common sample)**
 - Expect results to be more similar than when acquisition is also a variable.
 - Read differences as variation in protein and peptide ID analysis only.
- **The range of reported numbers of proteins still varies significantly, despite a common mass spectral data set starting point.**
- **The range persists even within people using common software.**
- **Personal choice of required stringency appears to be a major factor.**
- **Real differences in quality of results were also observed.**



*Proteome Informatics
Research Group*

Conclusions

- **Secondary Goal: Develop a benchmark**
 - Develop a reference test for both software users and developers.
- **Upon completion of this study, we intend to make publicly available:**
 - **The MS data set in all formats**
 - **The FASTA database**
 - **The protein consensus annotation**
 - **The Excel results template and grading tool.**
- **The study is currently still open! We need more labs submit their analysis of the data to improve the significance of the results.**
 - Email: iPRG2008@gmail.com if interested!



Acknowledgements

*Proteome Informatics
Research Group*

iPRG Members

- Jayson A. Falkner – University of Michigan (*new member*)
 - Jeffrey A. Kowalak - National Institutes of Health
 - William S. Lane - Harvard University
 - Alexey I. Nesvizhskii - University of Michigan
 - Brian C. Searle - Proteome Software (*incoming chair*)
 - Sean L. Seymour - Applied Biosystems (*outgoing chair*)
 - David L. Tabb - Vanderbilt University
-
- Renee Robinson - Anonymizer (Harvard University)

All iPRG2008 study respondents – Thank you!

Poster and slides available after ABRF at: www.abrf.org/iprg