Proteome Informatics
Research Group

iPRG 2011:

A Study on the Identification of Electron Transfer Dissociation (ETD) Mass Spectra

ABRF 2011, San Antonio, TX

2/20/11

Proteome Informatics
Research Group

INTRODUCTION:

CID AND ETD FROM 30K FEET

2

Proteome Informatics Research Group

CID and ETD – differences

Collision Induced Dissociation (CID) relies on a series of bimolecular events (collisions) to provide the peptide precursor with sufficient energy to fragment (*ergodic* process). CID typically causes backbone fragmentation. y and b ions are by far the most prevalent fragment types.

Electron Transfer Dissociation (ETD) relies on the transfer of a single electron to a peptide precursor. This transfer likely creates a radical that very quickly decays into ion fragments (a *non-ergodic* process). Like CID, ETD typically causes backbone fragmentation, but mostly resulting in c and z ions.

3

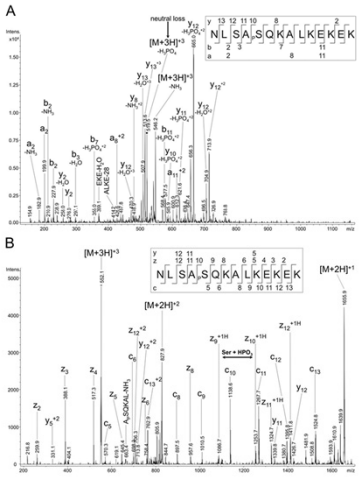
CID and ETD spectra - example

Proteome Informatics Research Group

NanoLC-ESI-MS/MS analysis of the CcO subunit IV.

CID

ETD



Helling S et al. Mol Cell Proteomics 2008;7:1714-1724

4

Proteome Informatics
Research Group

iPRG 2011 STUDY:

CONCEPT

5

Proteome Informatics Research Group

Study Goals

- Primary: Evaluate the ability of participants to identify ETD spectra
- Secondary: Find out why result sets might differ between participants
- Tertiary: Produce a benchmark dataset, along with a spectral library and an analysis resource

6

Proteome Informatics Research Group

Study Design

- Use a common, rich dataset
- Use a common sequence database
- Allow participants to use the bioinformatic tools and methods of their choosing
- Use a common reporting template
- Report results at an estimated 1% FDR (at the spectrum level)
- Ignore modification localization
- Ignore protein inference

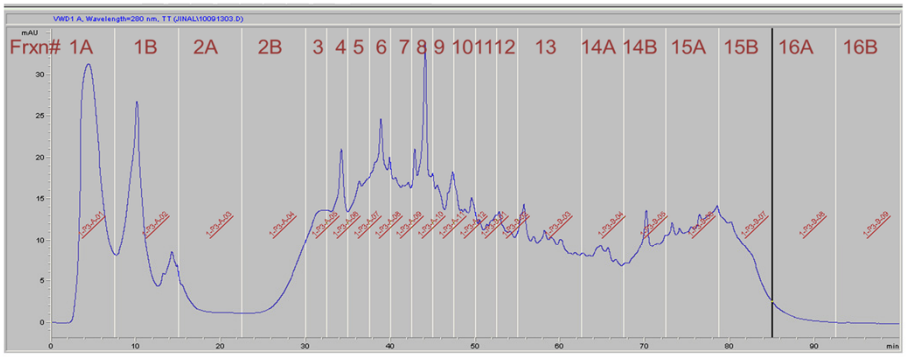
7

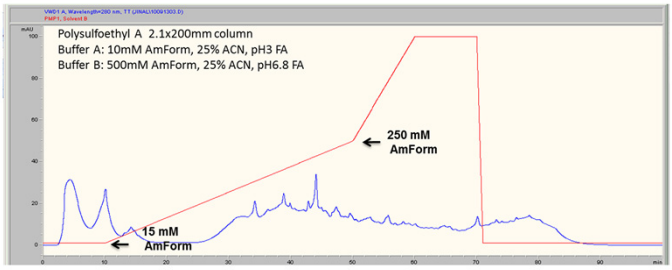
Proteome Informatics Research Group

The sample

NIST yeast lysate (six vials of RM8323), 228µg protein, LysC digest separated on SCX column

A & B pooled after concentrating





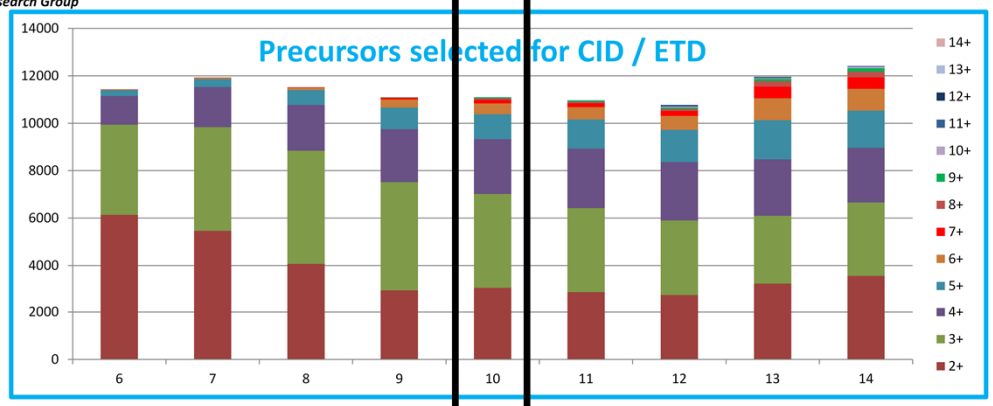
Sample prep by Robert Chalkley, UCSF

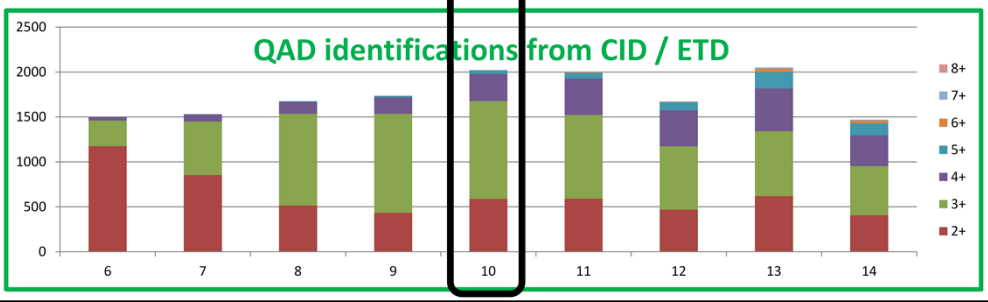
SCX by Jinal Patel, The Broad Institute

8

Choosing a fraction for the study

Proteome Informatics Research Group





9

Proteome Informatics Research Group

Study Materials (i)

- 1 LTQ-Orbitrap XL dataset (eq. 1 RAW file)
  - RAW, mzML, mzXML, MGF, dta – conversions by ProteoWizard 2.1.2051
- 1 fasta file (UniProtKB/SwissProt *S. cerevisiae* from Sept. 2010)
- 1 spectral library in SpectraST format (contributed by Henry Lam)
- 1 template (Excel)
- 1 on-line survey (Survey Monkey)

10

Study Materials (ii) – *additional data*

Proteome Informatics Research Group

Not public yet

Used to create library

Instrument:
XL – Orbitrap XL,
V – Velos Orbitrap

Fragmentation:
DT – Decision tree CID or ETD
DT – Decision tree HCD or ETD
H+E – HCD, ETD on each precursor
C+E – CID, ETD on each precursor
E – ETD only

Frac		Instrument	Fragmentation	MSMS Res/Acc	Spike?
9	D100914_yeast_SCX09_rak_ft8DT_pc_01.RAW	XL	DT (C or E)	LL	
	K100923_Yeast_SCX09_ft16DT_pcc_01.raw	V	DT (C or E)	LL	
	K100923_Yeast_SCX09_ff6f6HE_pcc_01.raw	V	H+E	HH	
	D100915_yeast_SCX09S_rak_ft8E_pc_02.RAW	XL	E	L	Y
	V20100923-23	V	C+E	LL	
10	D100914_yeast_SCX10_rak_ft8DT_pc_01.RAW	XL	DT (C or E)	LL	
	K100923_Yeast_SCX10_ft16DT_pcc_01.raw	V	DT (C or E)	LL	
	K100923_Yeast_SCX10_ff6f6HE_pcc_01.raw	V	H+E	HH	
	D100917_yeast_SCX10_rak_ft8E_pc_01.RAW	XL	E	L	
	D100930_yeast_SCX10S_rak_ft8E_pc_01.RAW	XL	E	L	Y
	V20100923-24	V	C+E	LL	
	V20100923-29	V	DT (C or E)	LL	
	V20100923-31	V	E	L	Y
	V20100923-32	V	DT (H or E)	HH	
11	D100914_yeast_SCX11_rak_ft8DT_pc_01.RAW	XL	DT (C or E)	LL	
	K100923_Yeast_SCX11_ft16DT_pcc_01.raw	V	DT (C or E)	LL	
	K100923_Yeast_SCX11_ff6f6HE_pcc_01.raw	V	H+E	HH	
	D100917_yeast_SCX11S_rak_ft8E_pc_02.RAW	XL	E	L	Y
	V20100923-25	V	C+E	LL	
	V20100923-30	V	DT (C or E)	LL	
	V20100923-33	V	DT (H or E)	HH	
12	D100914_yeast_SCX12_rak_ft8DT_pc_01.RAW	XL	DT (C or E)	LL	
	K100923_Yeast_SCX12_ft16DT_pcc_01.raw	V	DT (C or E)	LL	
	K100923_Yeast_SCX12_ff6f6HE_pcc_01.raw	V	H+E	HH	
	V20100923-26	V	C+E	LL	
	V20100923-27	V	E	LL	
S48	D100914_Sigma48_ft4t4_pcc_01.RAW	XL	C+E	LL	
	K100922_Sigma48_ft16DT_pcc_01.raw	V	DT (C or E)	LL	
	K100923_Sigma48_ff6f6HE_pcc_01.raw	V	H+E	HH	
	V20100923-28	V	E	L	

11

Proteome Informatics Research Group

Instructions to Participants

1. Retrieve and analyze the data file in the format of your choosing, with the method(s) of your choosing

2. Report the peptide to spectrum matches in the provided template

3. Fill out the survey

4. Attach a 1-2 page description of the methodology employed

12

Reporting Template (random example)

Proteome Informatics Research Group

ABRF iPRG 2011 Study Template: ETD Data Analysis

Instructions: Please fill in all fields required fields (marked with *). After deleting the example rows, create a new row for each *peptide spectrum match*. Indicate whether each match is better than a 1% FDR on the spectrum-level. Include identifications above and below threshold. Results should be sorted by 'Search Engine Score' from most to least confident. Additional instructions can be found above each field header. **Results should be emailed to 'anonymous.iprg2011@gmail.com' no later than Dec. 10, 2010.** Please make sure to fill out the REQUIRED survey (URL).

Identifiers should be unique scan numbers from data file. Retention times and spectrum indices (e.g., from the MGF file) are also acceptable if described in the 1-2 page methods report.	Measured precursor m/z as used by search engine (possibly after mono-isotopic peak corrections).	Precursor mass error in m/z or ppm. This value should be reported as the Precursor m/z - Theoretical precursor m/z.	Precursor charge reported by the search engine.	Use lowercase letters, a trailing symbol, a trailing delta mass value or a string in parentheses immediately following each residue containing a variable modification (see examples below). Localization ambiguity/certainty is not the focus of this study.	For each mod, list the position, amino acid residue and name of mod. For n-terminal modifications, use position=0 and residue=n-term. Mod localization is not the focus of this study. (not required)	Protein identifier(s) from Fasta file (see examples below). Use multiple values if peptide is found in multiple proteins Protein inference is not the focus of this study. (not required)	Peptide identification score reported by search engine and used for FDR calculation (e.g., E-value, p-value, probability, Mascot score, etc.)	'Y' indicates this match is BETTER than the confidence threshold. 'N' indicates the match is WORSE. Please report BOTH types of identifications in your ranked list.
Spectrum Identifier*	Precursor m/z*	Mass error*	Precursor Charge*	Peptide Sequence*	Modifications	Protein Accession(s)	Search Engine Score*	Better than 1% FDR threshold?*
Scan:1753	563.7818	-0.0004	2	mVGnRYLEK	1,M,oxidation;4,N,dea	Q12672	0.999999	Y
Scan:4669	842.4291	0.0004	2	FFGFTPEGVAERAQK		P23254	0.999999	Y
Scan:2156	673.3224	-0.0009	2	TSGYADRTAEFK		P22146	0.999999	Y
Scan:1571	414.8957	0.0007	3	nSTIKnHSLVK	1,N,deamidated;6,N,d	P41940	0.999999	Y
Scan:6017	838.5427	-0.0031	2	qGVLLPTRIKLLLTK	1,Gln->pyro-glu	P02365	0.999999	Y
Scan:4212	617.3125	0.0008	2	YKGnFTEFVK	4,N,deamidated	P16521	0.999999	Y
Scan:1587	658.3462	-0.0188	2	IIAENTNVAKDK		Q12447	0.999999	Y
Scan:7333	917.7254	-0.0013	4	GLVSDPAGSDALNVLKYFDYN\	28,C,carbidomethyl	P54839	0.999999	Y

13

Proteome Informatics
Research Group

iPRG 2011 STUDY:

PARTICIPATION

14

Soliciting Participants and Logistics

Proteome Informatics Research Group

Study advertised on the ABRF website and listserv, Molecular and Cellular Proteomics blogsite, ECD/ETD conference attendants, GenomeWeb and by direct invitation from iPRG members

Participant

1. Email participation request to 'iPRGxxxx@gmail.com'

2. Send official study letter with instructions

iPRG members

Questions / Answers

3. All further communication (e.g., questions, submission) through 'iPRGxxx.anonymous@gmail.com'

"Anonymizer"

15

Proteome Informatics Research Group

Participants (i) – *overall numbers*

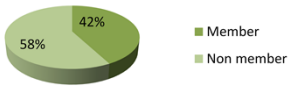
- 40 requests / 36 submissions (‘90% return’)
  - Some participants submitted two result sets
- 9 initialed iPRG member submissions (with appended ‘i’)
- 8 vendor submissions (identifiable by appended ‘v’)

16

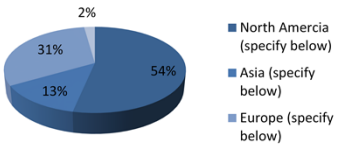
Participants (ii) - *demographics*

Proteome Informatics Research Group

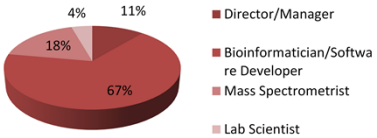
Membership



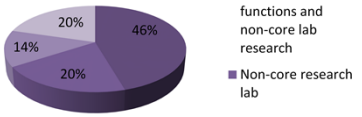
Location



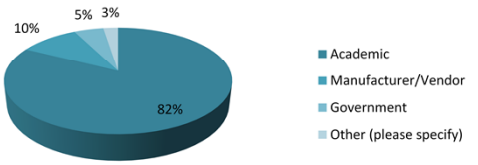
Primary Job Function



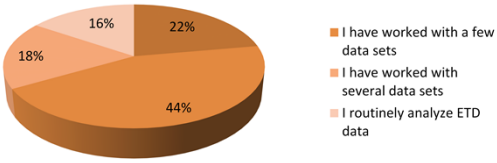
Resource Lab Status



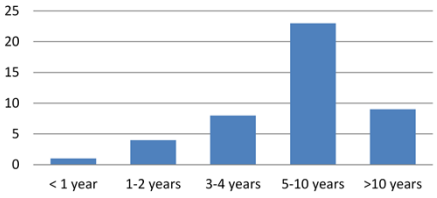
Type of Lab



ETD Data Interpretation Experience



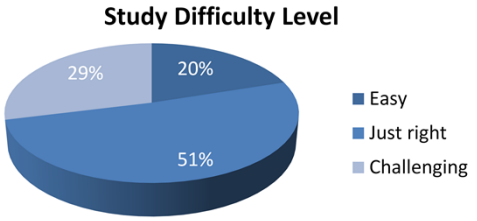
Years of Experience in Proteomics

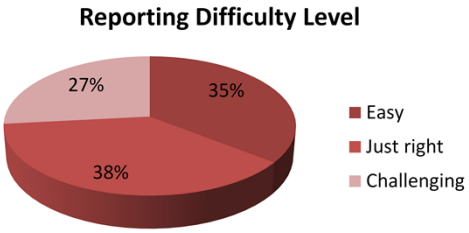


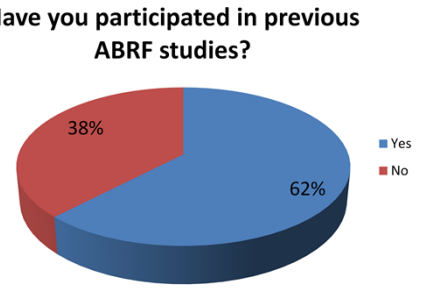
17

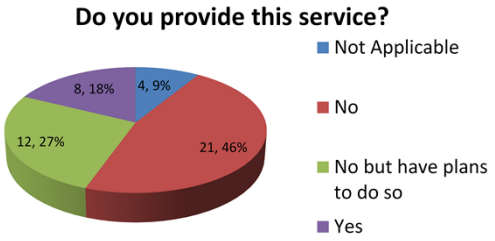
Proteome Informatics Research Group

Participants (iii) – *study opinions*





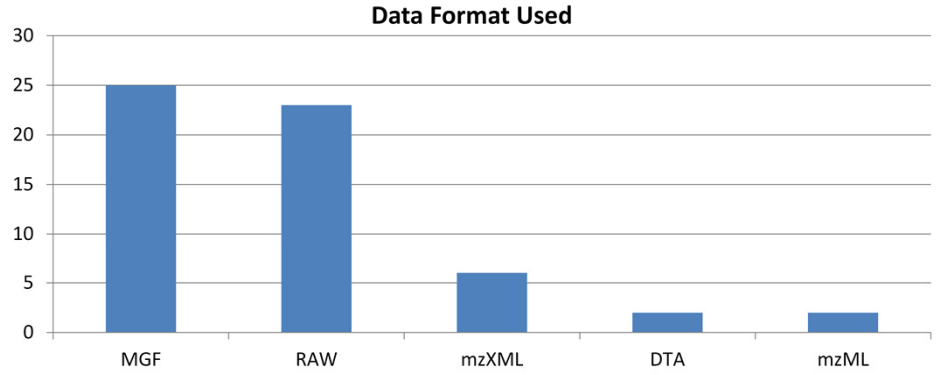




18

Participants (iv) – *methods (i)*

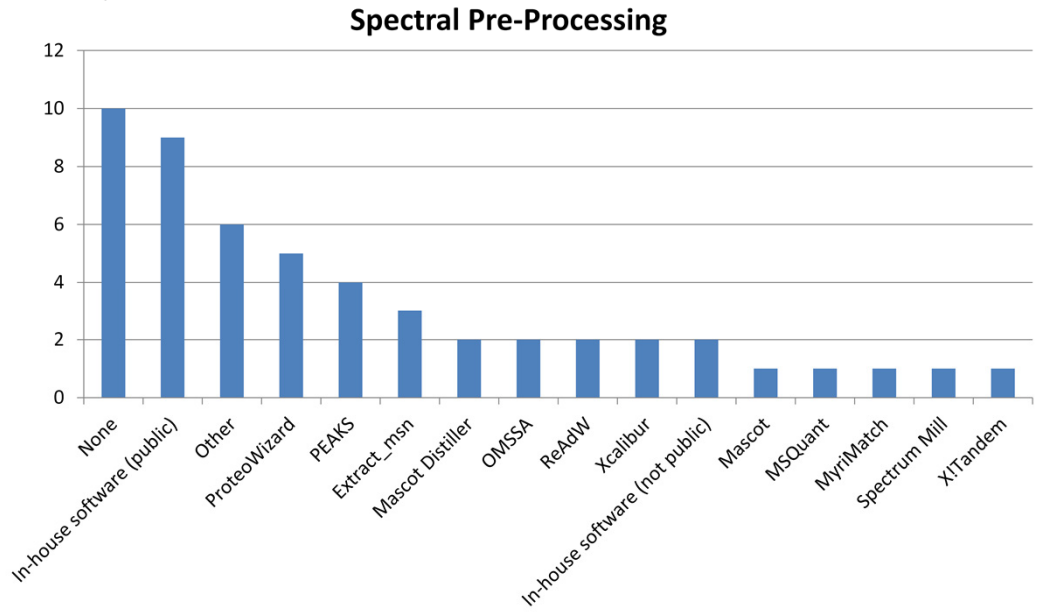
Proteome Informatics Research Group



19

Participants (v) – *methods (ii)*

Proteome Informatics Research Group



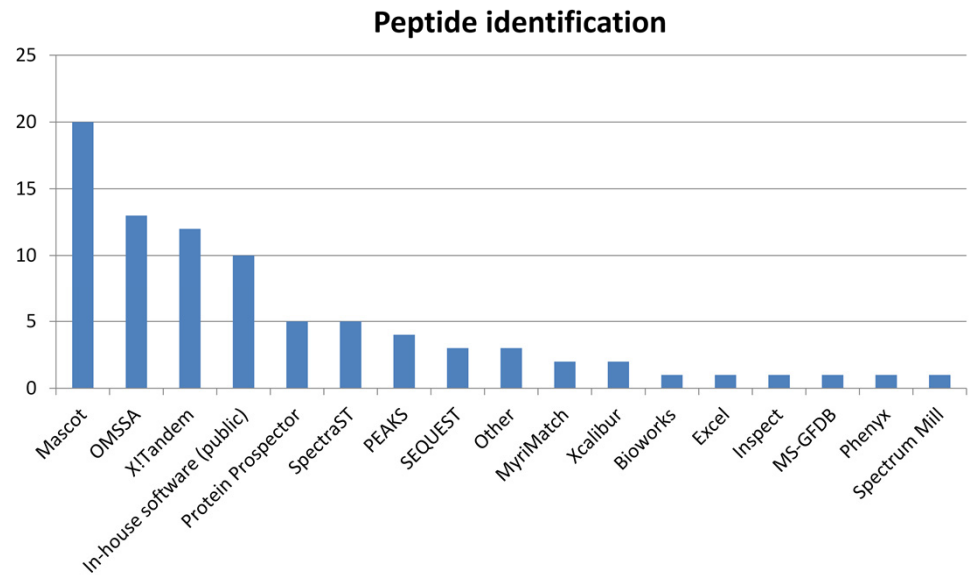
Other / in-house (public / non-public):

DTARefinery, DeconMSn, DTA Generator, Etdgenerator, RawExtractor, Hardklor, multiplierz, ReAdW, Byonic2

20

Proteome Informatics Research Group

Participants (vi) – *methods (iii)*

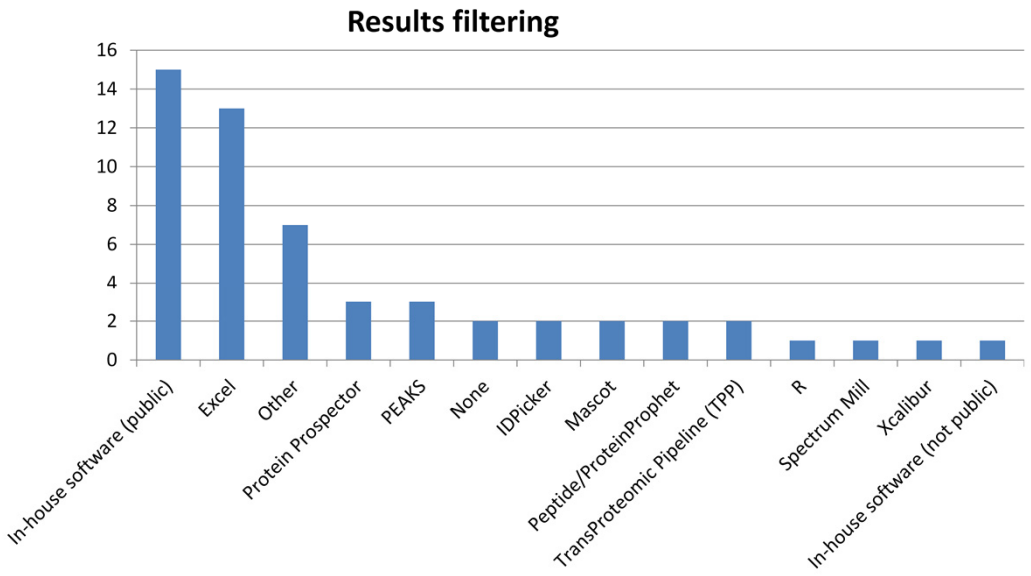


Other / in-house (public / non-public):
pFind, ByOnic, ProteinScape, MS_LIMS, PVIEW, PepArML, Byonic2, Proteome Discoverer

21

Participants (vii) – *methods (iv)*

Proteome Informatics Research Group



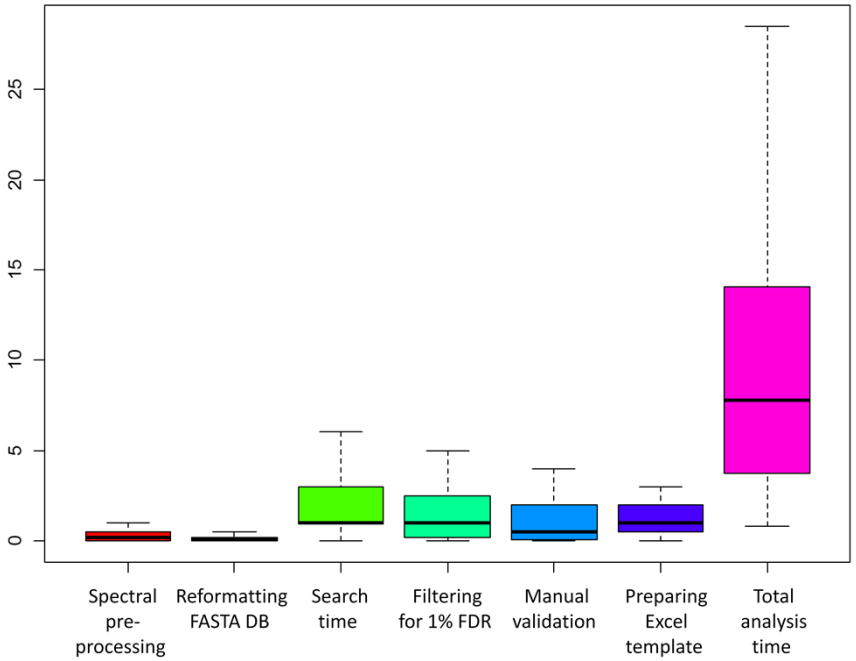
Other / in-house (public / non-public):

pBuild, ComByne, ProteinScape, Percolator, PVIEW, Epitomize, FDR Optimizer, MSblender, OmssaParser, MascotDatFile, multiplierz, ComputeFDR, Proteome discoverer

22

Proteome Informatics Research Group

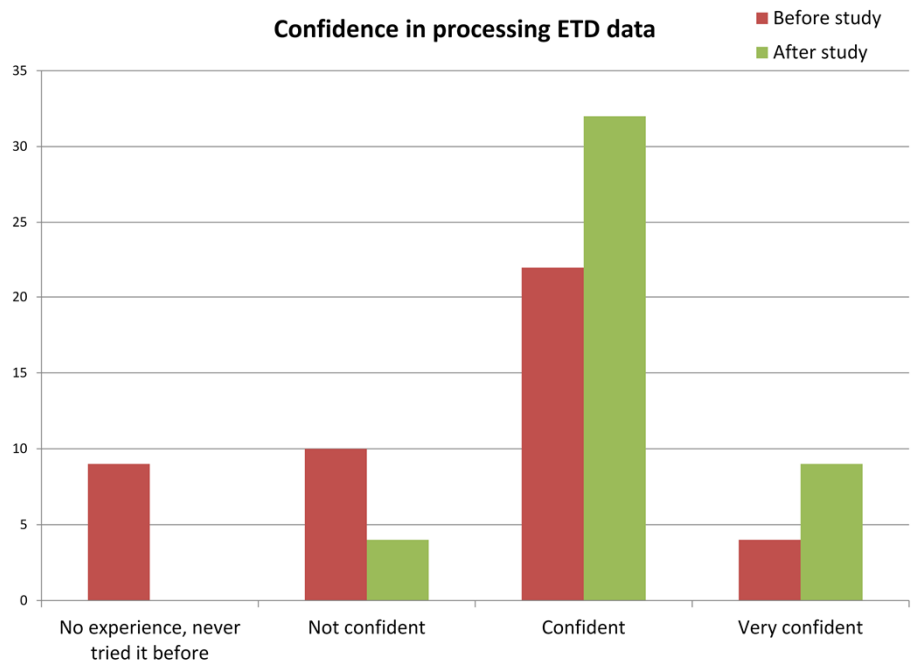
Participants (viii) – *time spent (hours)*



23

Proteome Informatics Research Group

Participants (ix) – *confidence*



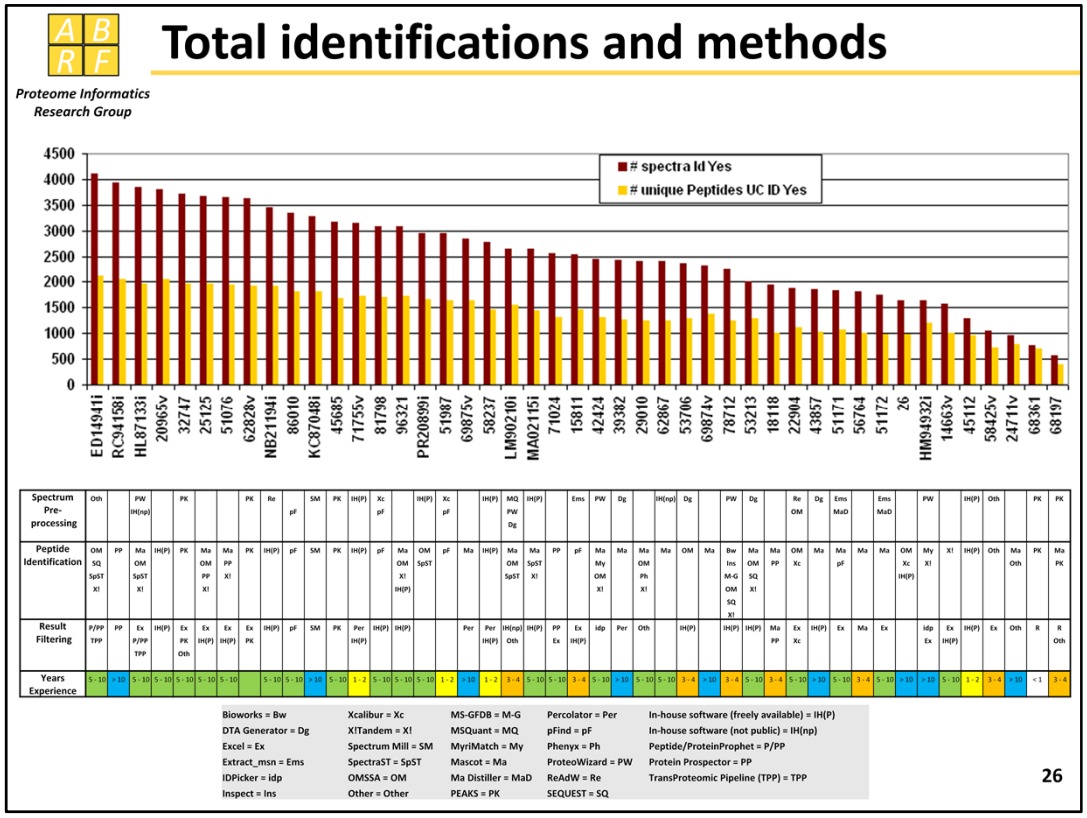
24

Proteome Informatics
Research Group

iPRG 2011 STUDY:

PRELIMINARY ANALYSIS

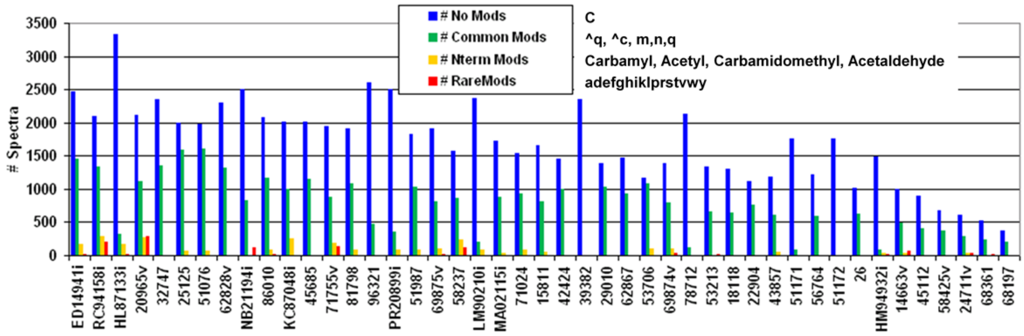
25



iPRG studies are not competitions. Leftmost is not meant to imply best, it just reflects the sorting criterion: total number of confident ids; this was chosen as a convenient means of sorting, and this sort is used throughout for consistency.

Proteome Informatics Research Group

Modifications



Common	4	4	3	4	3	3	3	3	3	3	4	?	?	3	4	2
N-terminal	1	6		5		1	1			1	3	?	?	1		2
Rare	1	14		16	1	1	4		2	1		?	?	1	3	0

Common	?	3	3	3	?	3	3	3	?	?	3	4	3	3	?
N-terminal	?	2		1	?	4	1		?	?		1	2		?
Rare	?	1	1		?				?	?	1	1	1		?

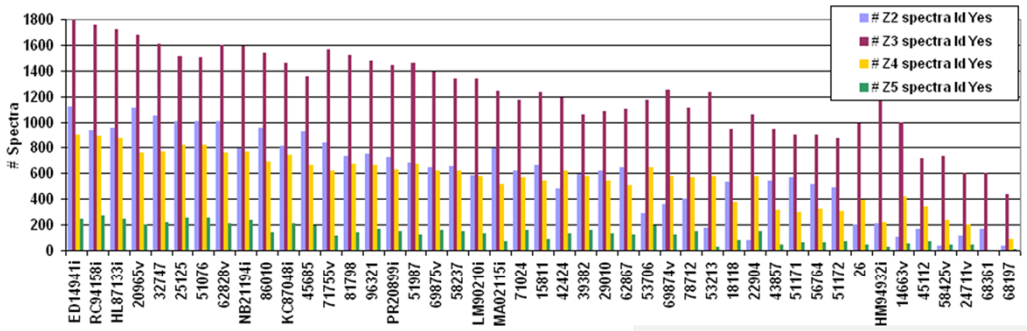
Common	3	4	3	3	3	2	3	2	2	?	2	2	3	3
N-terminal			1				1	2	1	?		?		
Rare					2		1		23	?	1	?	3	1

27

Questions marks and blanks are present when participants either did not clearly indicate which modifications were allowed for or did not include them in a sequence specific manner in their results.

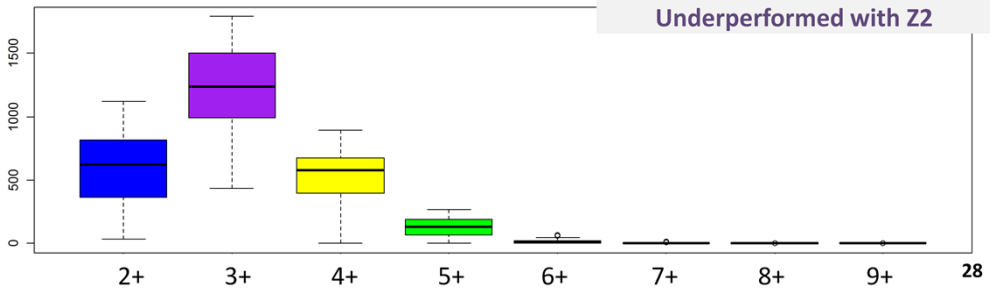
Charge state distributions

Proteome Informatics Research Group

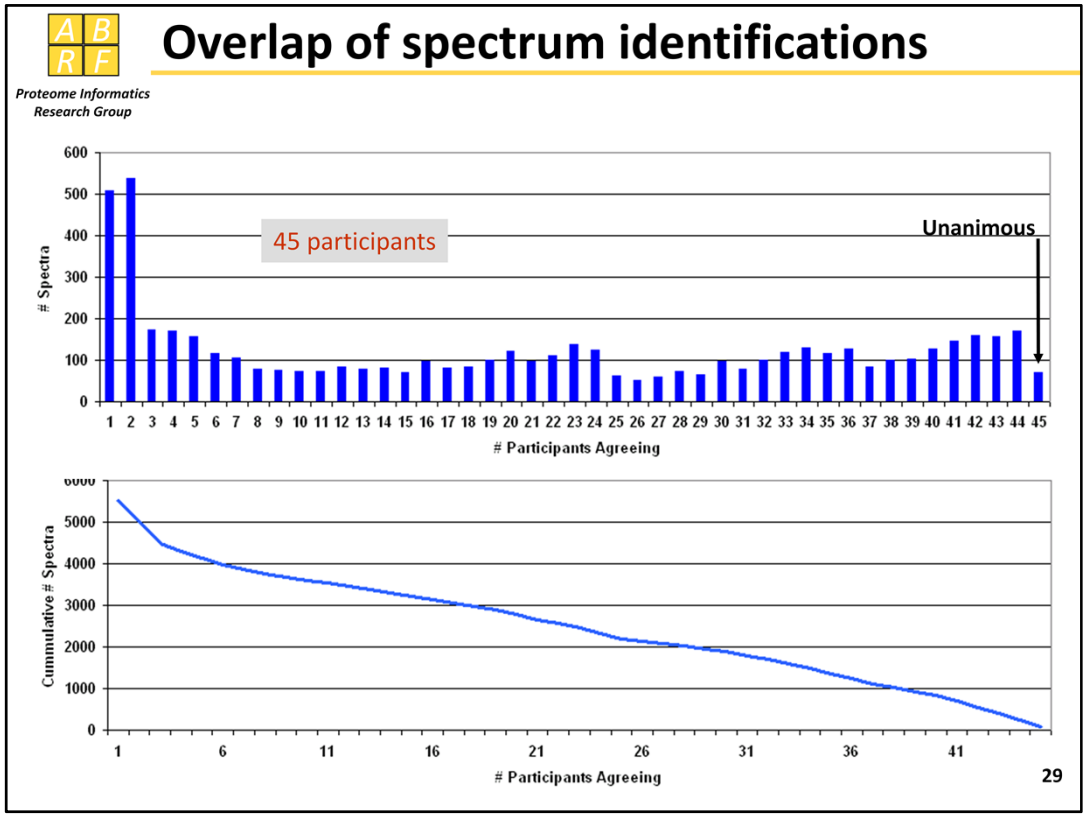


**** * ********

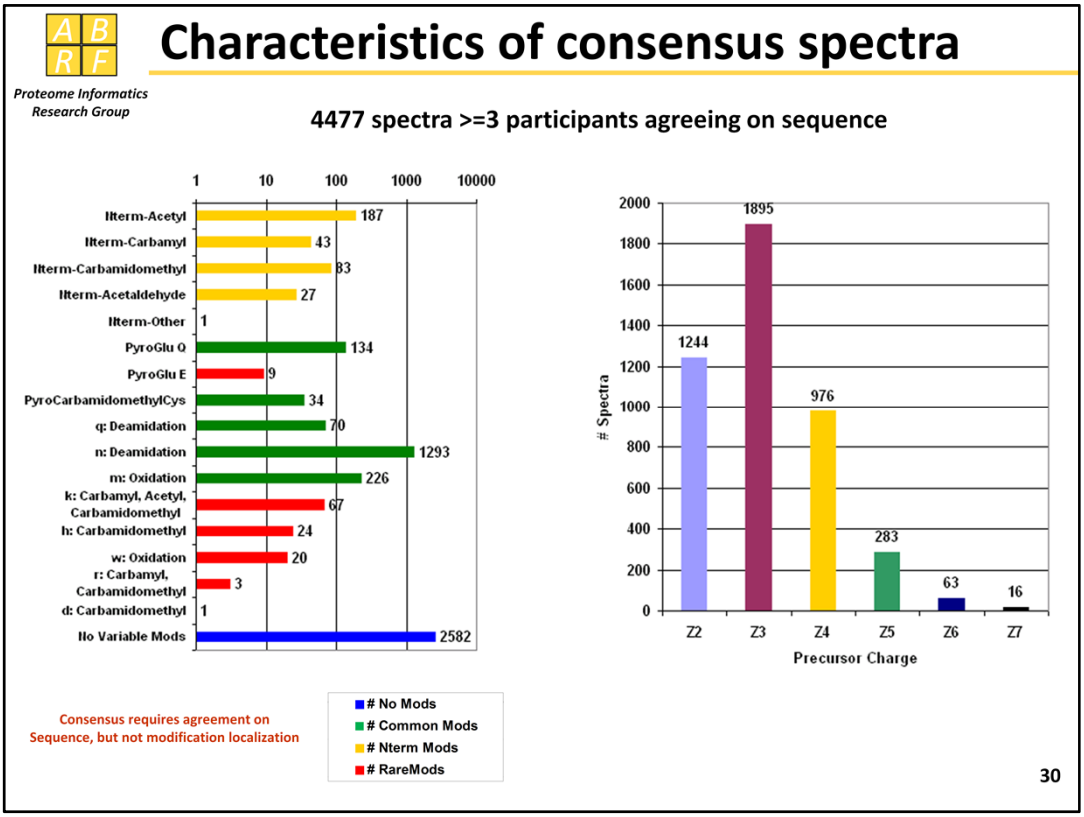
Underperformed with Z2



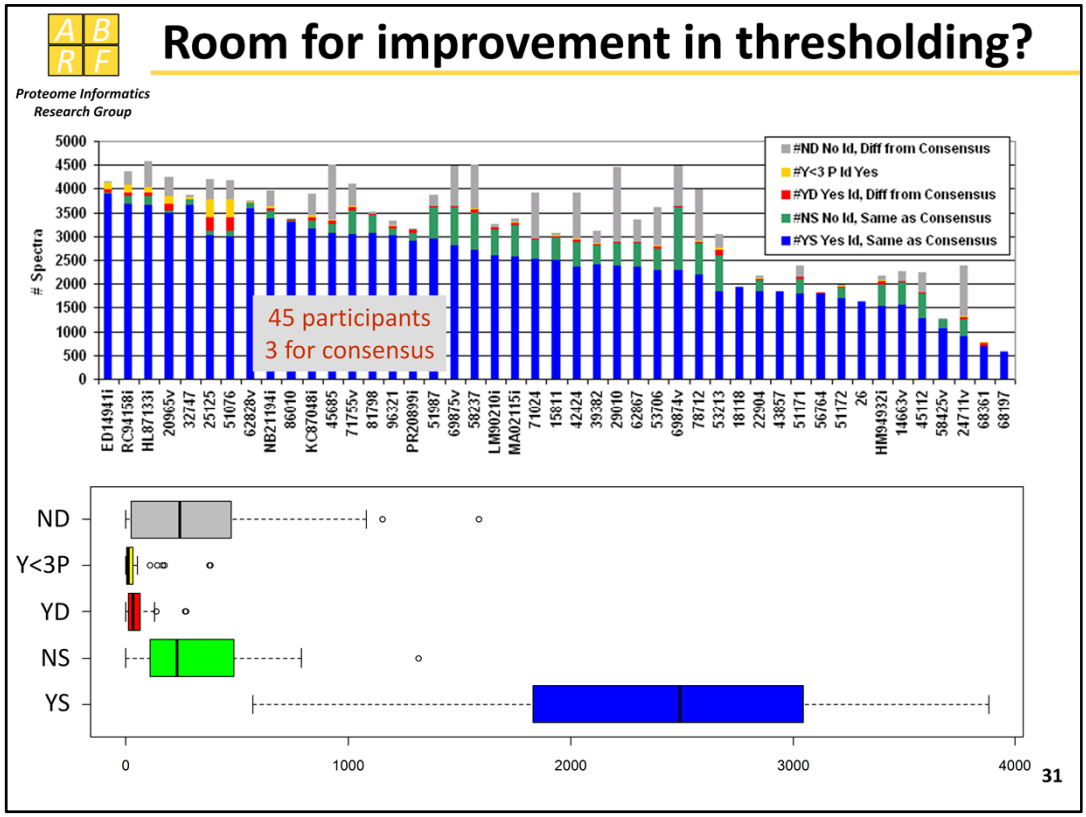
28



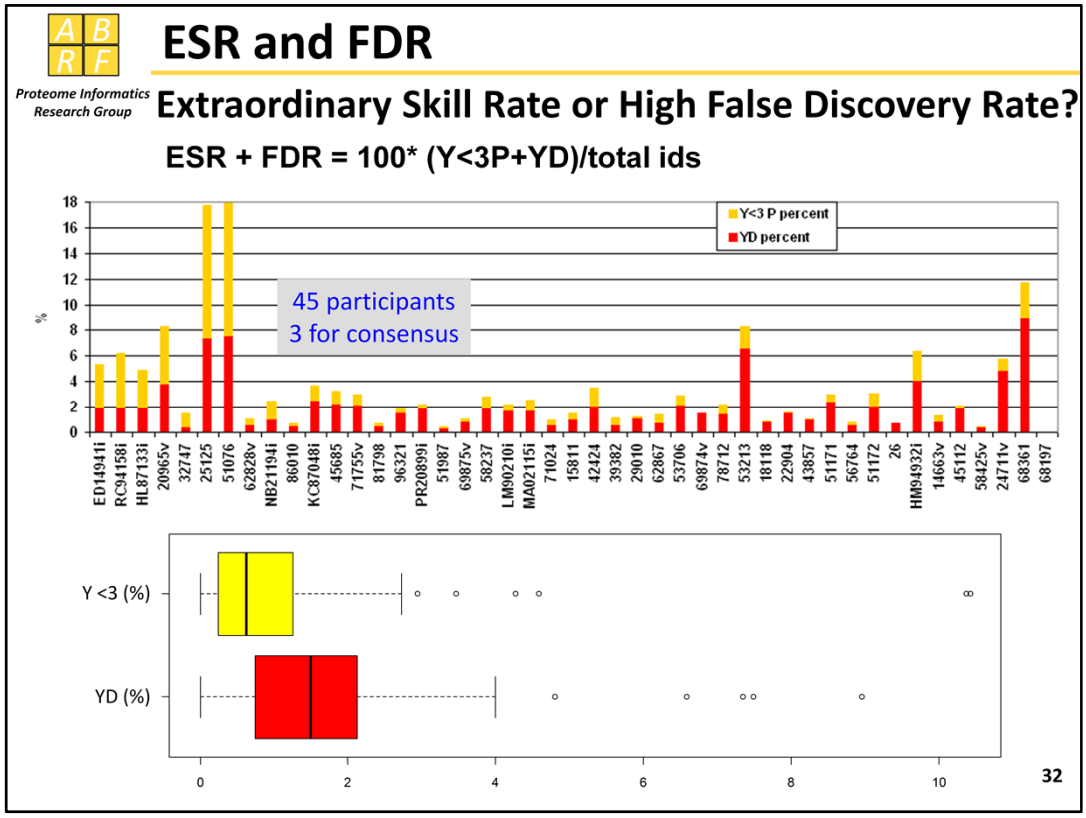
Since an inordinate number of spectra had only 1 or 2 participants agreeing, we selected 3 participants agreeing as the threshold for denoting consensus agreement. Consensus requires agreement on sequence, so do note that this still allows for disagreement on modification localization.



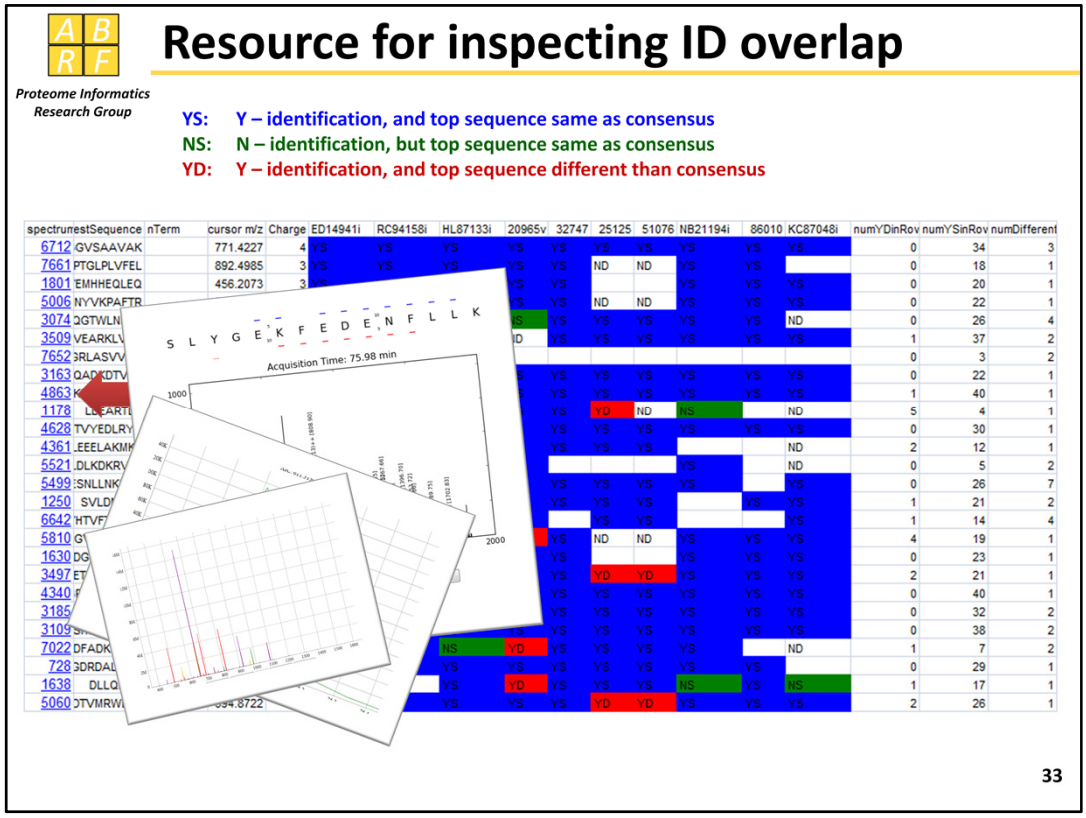
We would also liked to have categorized how many identified spectra were derived from enzymatic specificity of full, semi, or none. This information was not readily collected. However, for some participants 10-20% of confident identifications came from something other than full enzymatic specificity.



The green (NS) bars represent the room for improving confidence threshold setting. If one could improve the decision making about confidence in a peptide spectral match, without increasing FDR then the sum of the heights of the blue (YS) and green (NS) bars appears to be within reach for many participants to substantially improve their overall identification totals.



When particular identifications are reported by less than 3 participants it is difficult to tell whether that represents extraordinary skill or just another false positive. On the other hand, disagreement with the consensus (YD) is more likely to indicate a wrong answer. Consequently, the YD rate serves as a surrogate for the minimum FDR level. Note that many participants, especially those to the far left tend to have a YD rate >1%, which suggests they have underestimated their FDR level. The study was requested to be performed at 1% FDR.



Complete spreadsheet will be available for download at http://www.abrf.org/index.cfm/group.show/ProteomicsInformaticsResearchGroup.53.htm

Proteome Informatics
Research Group

iPRG 2011 STUDY:

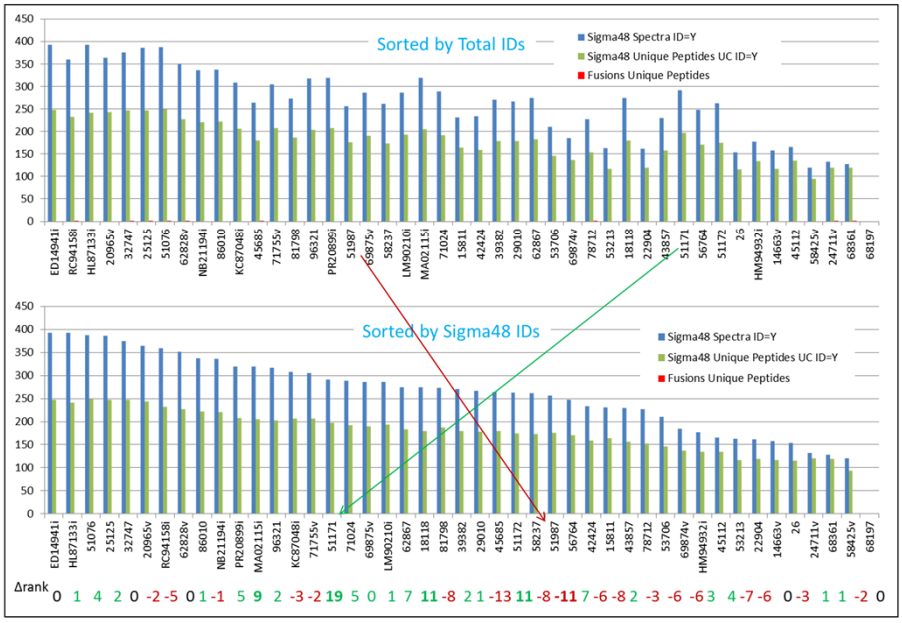
TWO SURPRISES AT THE END

34

Surprise Nº 1: Sigma-48 spike-in

Proteome Informatics Research Group

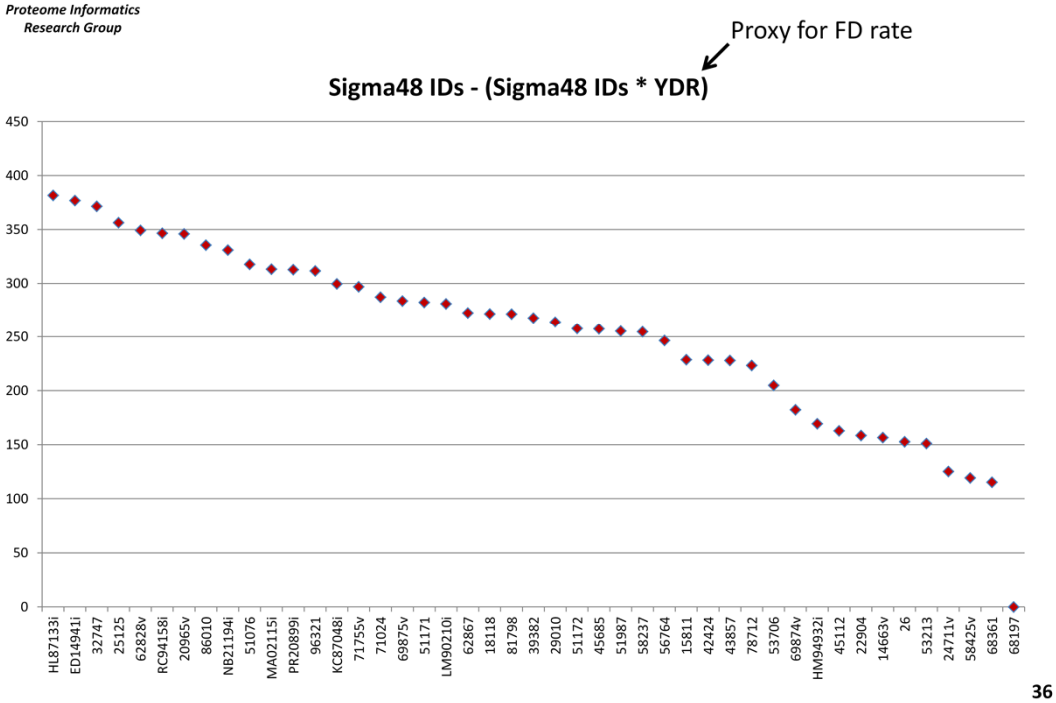
Sigma 48 digested separately and not subjected to SCX. Spike-in at a level to yield ~10% of ID peptides. ***Biases against non-iPRG DB users, and SCX prediction users***.



35

Sigma-48 as TP estimator

Proteome Informatics Research Group



36

Surprise Nº 2: Dr. Moreaudnick...

Proteome Informatics Research Group

Previously unidentified yeast protein

Orig. yeast sequences

>sp|P47002|SSY5_YEAST SPS-sensor serine protease component SSY5 OS=Saccharomyces cerevisiae GN=SSY5 PE=1 SV=2
MVRFFGLNKKKNEEKENTDLPADNEQNAAETSSSNVSGNEERIDPNVDTNPENANNDDA
STTFGSSIQSSSIFSRGRMTYGTGASSSMATSEMRSHSSGHSGSSSKNLQGFKDVGKPL
RAVSFLSPVKEEESQDTQNTLDVSSSTSSTLATSENARENSFSRRSITLEYIHKSLSEL
EENLVDIMDDIHQDVISISKAVIEAIEYFKEFLPTTRDLSRISLEKSSSLRKINKIVL
HFLDNLLVSDAFSNSRSILLRRFYFFLFKLNLITDDDLSSGVLPCLSVFCIGSHCNLP
SMDKLGMILDELTKMDSSTISDQEGAFTAPILRGTPKSSILTIMFGLPNLQHEHYEMTK
VLYSLFPDVHMYCVKDYIKKAASAVGSIPSHTAATIDTIAPTKFQFSPPYAVSENPLELP
ISMSLSTETSAKITGTLGGYLFPQTGSDKKFSQFASCSFAITCAHVVLSEKQDYPNVMVP
SNVLQTSYKKVLTKESDRYPDGSVEKTAFLEEVQRIDQNLNWQKSNKFGQVVWGERAIVD
HRLSDFAIIKVNSSFKCQNTLGNGLKSFPDPTLRFQNLHVKRKIFKMKPGMKVFKIGAST
GYTSGELNSTKLVYWADGKLQSSEFVVASPTPLFASAGDSGAWILTKLEDRLGLLVGML
HSYDGEQRQFGLFTPIGDILERLHAVTKIQWDIDPQLDG
>sp|P38788|SSZ1_YEAST Ribosome-associated complex subunit SSZ1 OS=Saccharomyces cerevisiae GN=SSZ1 PE=1 SV=2
MSSPVIGITFGNTSSSIAYINPKNDVDVIANPDGERAIPSALSYVGEDEYHGGQALQQLI
RNPKNTIINFRDFIGLPFDKCDVSKCANGAPAVEVDGKVGFVISRGEGKEEKLTVDEVVS
RHLNRLKLAAEDYIGSAVKEAVLTVPTNFSEEQKTALKASAAKIGLQIVQFINEPSAALL
AHAEQFPFEKDVNVVVADFGGIRSDAAVIAVRNGIFTILATAHDLSLGGDNLDTELVEYF
ASEFQKKYQANPRKNARSLAKLKANSSITKKTLSNATSATISIDSLADGFDYHASINRMR
YELVANKVFAQFSSFVDSVIAKAELDPLDIDAVLLTGGVSFTPKLTTNLEYTLPESVEIL
GPQNKNASNNPNELAASGAALQARLISDYDADELAEALQPVIVNTPHLKKPIGLIGAKGE
FHPVLLAETSFPVQKKLTLKQAKGDFLIGVYEGDHHIEEKTLEPIPKEENAEEDDESEWS
DDEPEVVREKLYTLGTKLMELGIKNANGVEIIFNINKDGALRVTARDLKTGNAVKGEL

+

>sp|P01344|IGF2_HUMAN Insulin-like growth factor II OS=Homo sapiens GN=IGF2 PE=1 SV=1
MGIPMGKSMLVLLTFLAFASCCIAAYRPSETLCGGELVDTLQFVCGDRGFYFSRPASRVS
RRSRGIVEECCFRSCDLALLETYCATPAKSERDVSTPPTVLPDNFPRYPVGKFFQYDTWK
QSTQRLRRGLPALLRARRGHVLAKELEAFREAKRHRPLIALPTQDPAHGGAPPEMASNRK

Fusion junction (underlined)

Final fusion in FASTA

>sp|P47002|SSY5_YEAST SPS-sensor serine protease component SSY5 OS=Saccharomyces cerevisiae GN=SSY5 PE=1 SV=2
MGIPMGKSMLVLLTFLAFASCCIAAYRPSETLCGGELVDTLQFVCGDRGFYFSRPASRVS
RRSRGIVEECCFRSCDLALLETYCATPAKSERDVSTPPTVLPDNFPRYPVGKFFQYDTWK
QSTQRLRRGLPALLRARRGHVLAKELEAFREAKRHRPLIALPTQDPAHGGAPPEMASNRK
MSSPVIGITFGNTSSSIAYINPKNDVDVIANPDGERAIPSALSYVGEDEYHGGQALQQLI
RNPKNTIINFRDFIGLPFDKCDVSKCANGAPAVEVDGKVGFVISRGEGKEEKLTVDEVVS
RHLNRLKLAAEDYIGSAVKEAVLTVPTNFSEEQKTALKASAAKIGLQIVQFINEPSAALL
AHAEQFPFEKDVNVVVADFGGIRSDAAVIAVRNGIFTILATAHDLSLGGDNLDTELVEYF
ASEFQKKYQANPRKNARSLAKLKANSSITKKTLSNATSATISIDSLADGFDYHASINRMR
YELVANKVFAQFSSFVDSVIAKAELDPLDIDAVLLTGGVSFTPKLTTNLEYTLPESVEIL
GPQNKNASNNPNELAASGAALQARLISDYDADELAEALQPVIVNTPHLKKPIGLIGAKGE
FHPVLLAETSFPVQKKLTLKQAKGDFLIGVYEGDHHIEEKTLEPIPKEENAEEDDESEWS
DDEPEVVREKLYTLGTKLMELGIKNANGVEIIFNINKDGALRVTARDLKTGNAVKGEL

Next protein in FASTA

Sigma48 protein

37

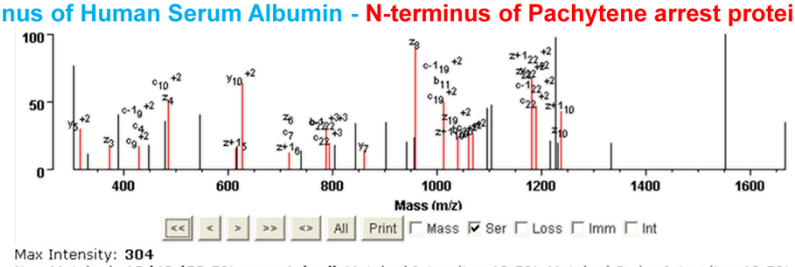
Identification of *fusion* peptides

Proteome Informatics Research Group

Five participants reported the peptide:

KLVAASQAALGLMNYLETQLNKK

C-terminus of Human Serum Albumin - N-terminus of Pachytene arrest protein SAE3

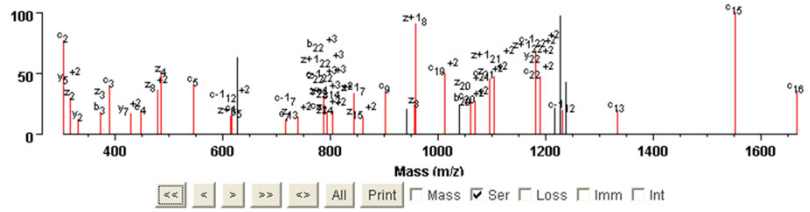


Max Intensity: **304**
Num Matched: **19/40 (52.5% unmatched)** Matched Intensity: **46.5%** Matched Series Intensity: **46.5%**

Consensus Answer:

rare mod

Acetyl-SRSGVAVADESLTAFNDLKLG*K(Carbamidomethyl)*K^{4+}



Max Intensity: **304**
Num Matched: **33/40 (17.5% unmatched)** Matched Intensity: **79.6%** Matched Series Intensity: **79.6%**

38

Proteome Informatics Research Group

Conclusions

- Study went well and had global participation, with quite a few first-timers joining in
- Earlier software problems with interpreting doubly-charged precursors have been largely cleared up
- Experience with software is probably a better measure of performance than the actual tool used
- People are generally over-optimistic about how reliable their results are (FDR underestimation)
- However, false negatives (NS) are generally much higher than false positives, so there is room for improvement there

39

Proteome Informatics Research Group

What did the participants think?

"I work in an environment without a group of peers doing proteomics. So having a study like this definitely gives me a chance to compare notes with others and benchmark my abilities. It also gives me a piece of evidence to prove my ability to people that are not in the field, such as users, administrators, advisory committee. I believe all core only facilities should participate in these studies."

100% of participants found the study useful

"The results of this study will be good to show how much room for improvement there is for the popular identification tools in ETD analysis. It's a good opportunity for lesser known and more open software to make a significant impact in the field."

Proteome Informatics Research Group

Thank you! Questions?

THANK YOU TO ALL STUDY PARTICIPANTS!

iPRG

~~Manor Askenazi~~
Nuno Bandeira
Robert Chalkley
Karl Clauser
Eric Deutsch
Henry Lam *(~~ad-hoc member~~)*
~~Paul Rudnick~~
Tom Neubert *(EB liaison)*
Hayes McDonald *(chair elect)*
~~Lennart Martens *(chair)*~~
John Cottrell *(member elect)*
Matt Chambers *(member elect)*
Ruixiang Sun *(member elect)*
Eugene Kapp *(member elect)*

Indispensable help came from:

Jinal Patel, The Broad Institute
Namrata Udeshi, The Broad Institute
Jeremy Carver, UCSD

41