ThermoFisher SCIENTIFIC

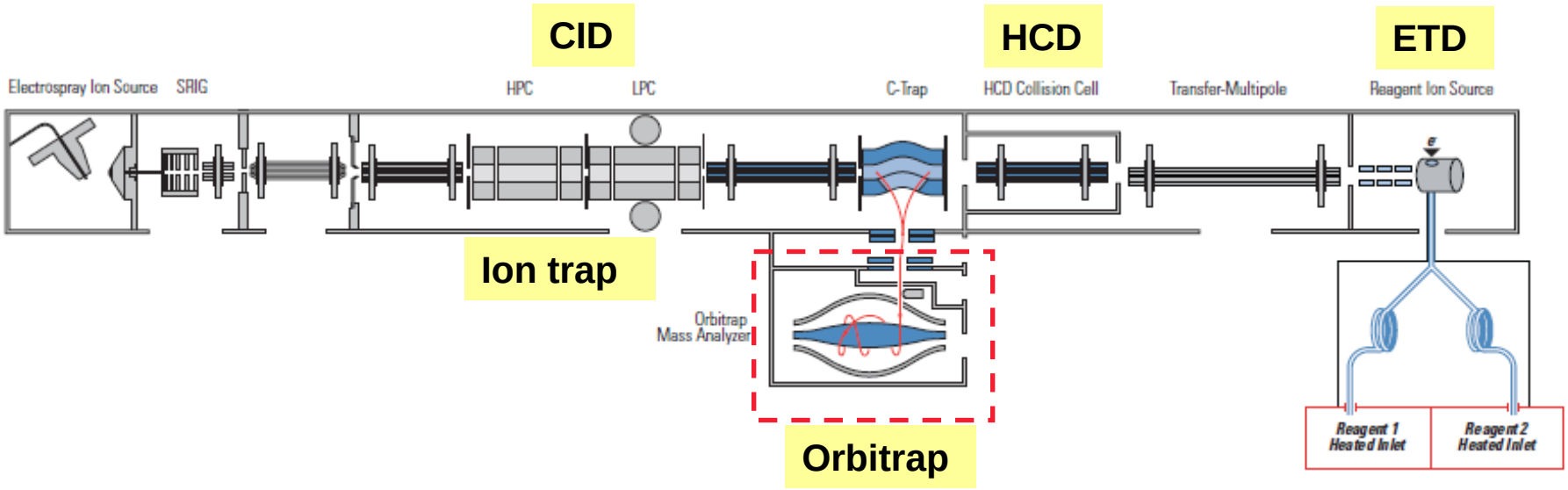
Proteome Discoverer 1.3 Software: Enhanced Tools For Protein Identification

David Horn

Proteomics Software Strategic Marketing Manager

September 22, 2011

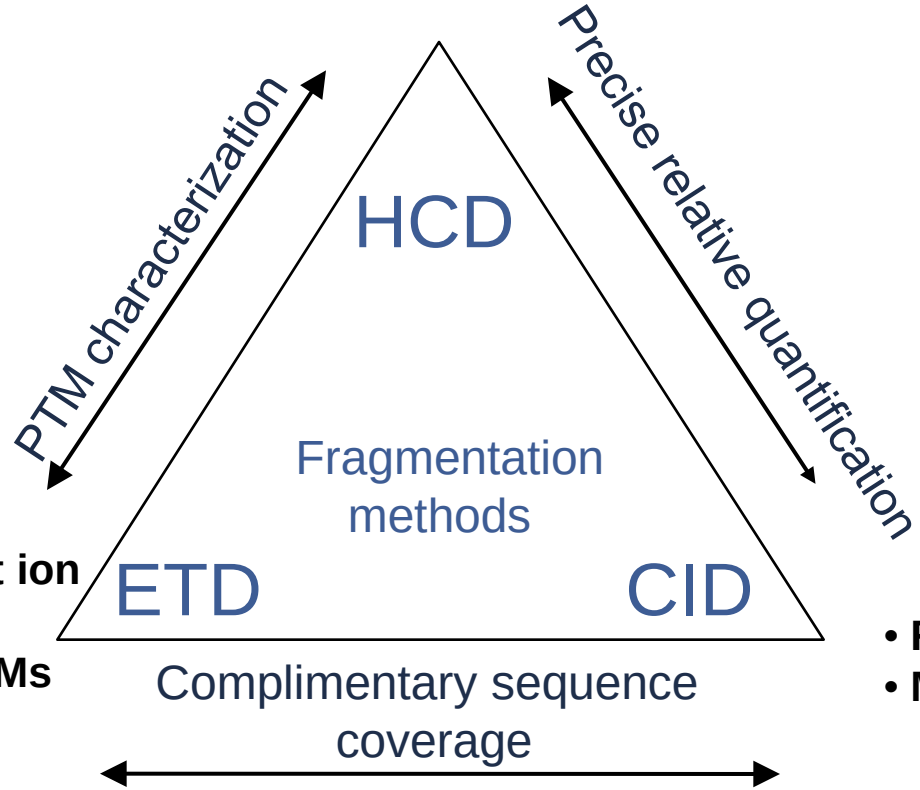
Hybrid Orbitrap Mass Spectrometers



- Full flexibility of fragmentation techniques (CID, HCD, ETD)
- Up to 240,000 resolving power for MS, low-ppm mass tolerance
- Highest sensitivity (i.e. ion trap with SRIG: 5-10X vs. LTQ)
- Top down (Intact protein analysis), Bottom up (Protein ID), PTMs (especially those requiring MS^n such as glycomics/glycoproteomics), Comprehensive Quan (SILAC, label free, targeted peptide quan with HRAM)

Multiple Fragmentation Methods Provide Complementary Information

- **High mass accuracy**
- **High energy fragmentation**
- **High intensity at low m/z**
- **Diagnostic ions (immonium & reporter ions)**



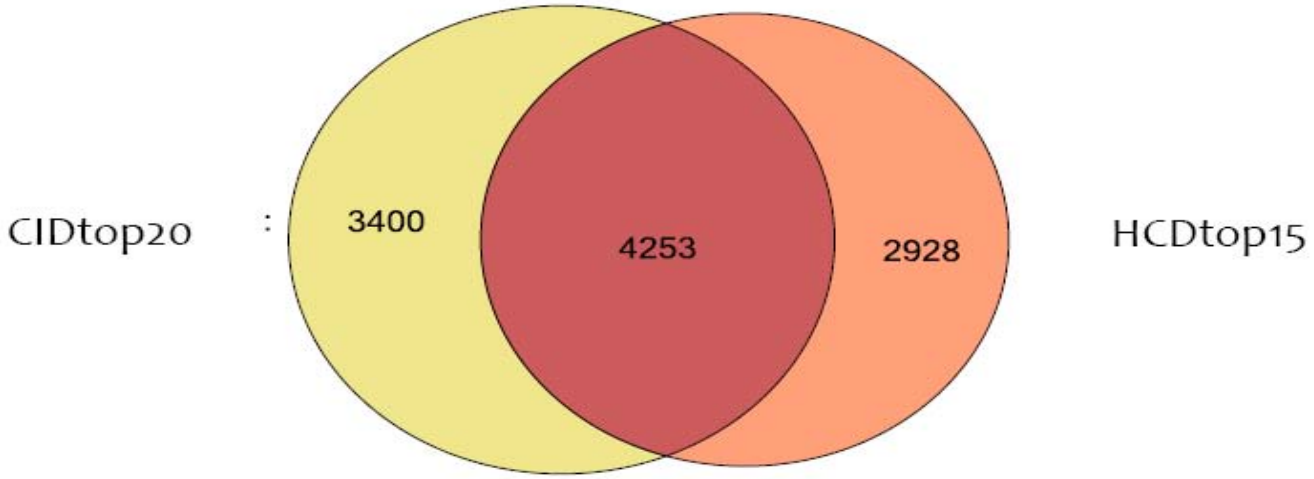
- **Complete fragment ion coverage**
- **Preservation of PTMs**

- **Robust**
- **Most versatile**

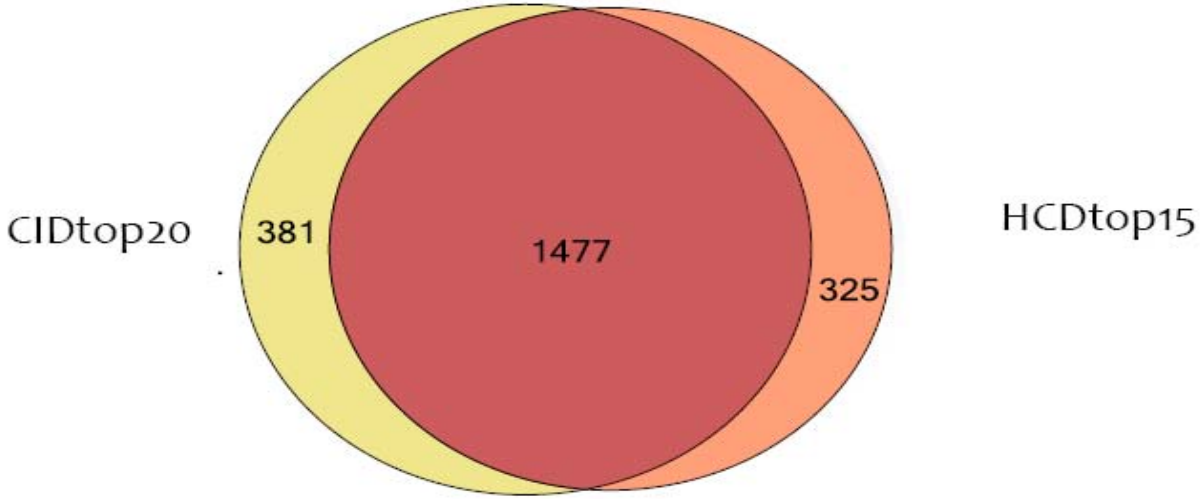
Benefits of multiple fragmentation methods

Multiple modes of fragmentation= x 1.5 more unique peptides

Peptides

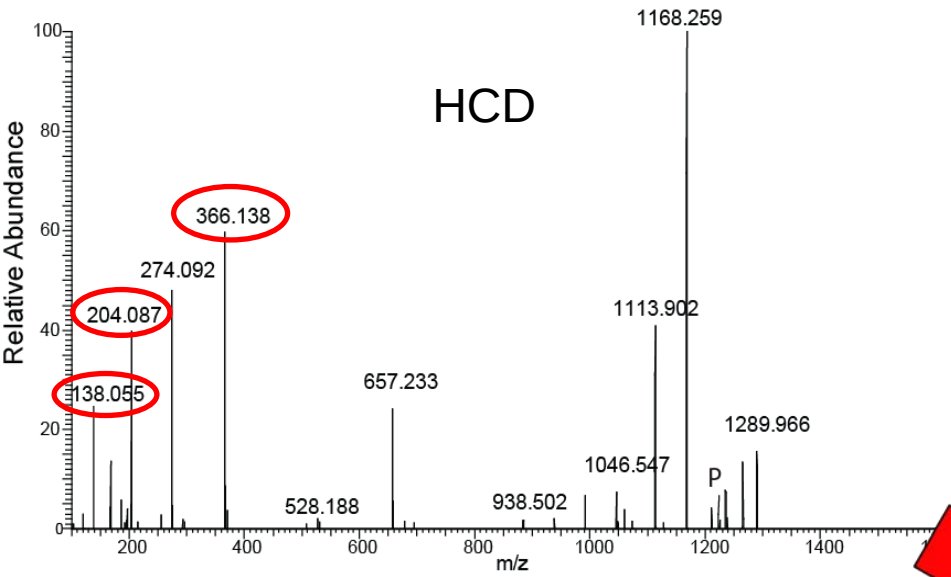


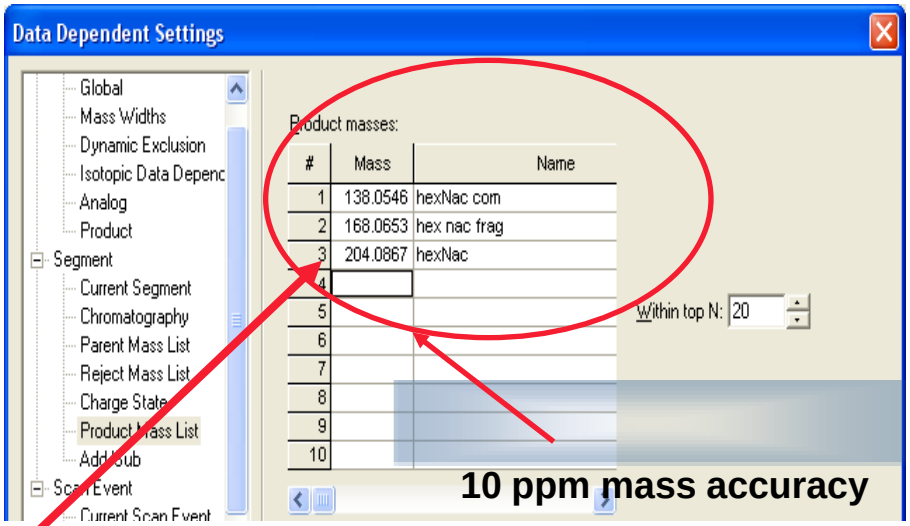
Proteins



Sample: 200 ng HeLa lysate; tryptic digest; 90 min gradient.

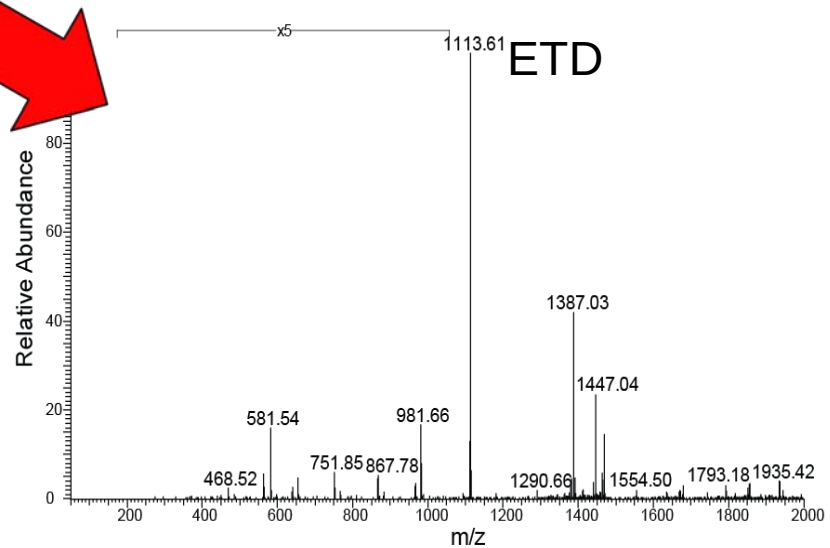
Example: HCD-triggered ETD for Identification of Glycopeptides





Oxonium Ions=Product Ions

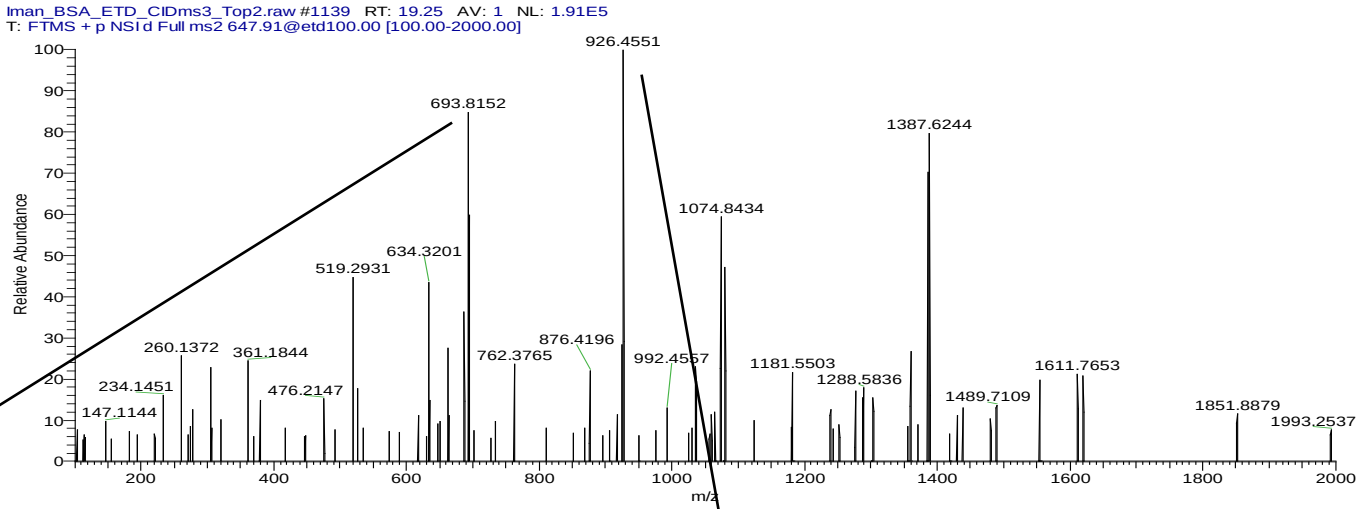
Hex	**m/z 163.06**
HexNAc	**m/z 204.087**
HexNAc fragment ions	**m/z 138.055**
	m/z 186.076
Hex-HexNAc	**m/z 366.138**



ThermoFisher SCIENTIFIC

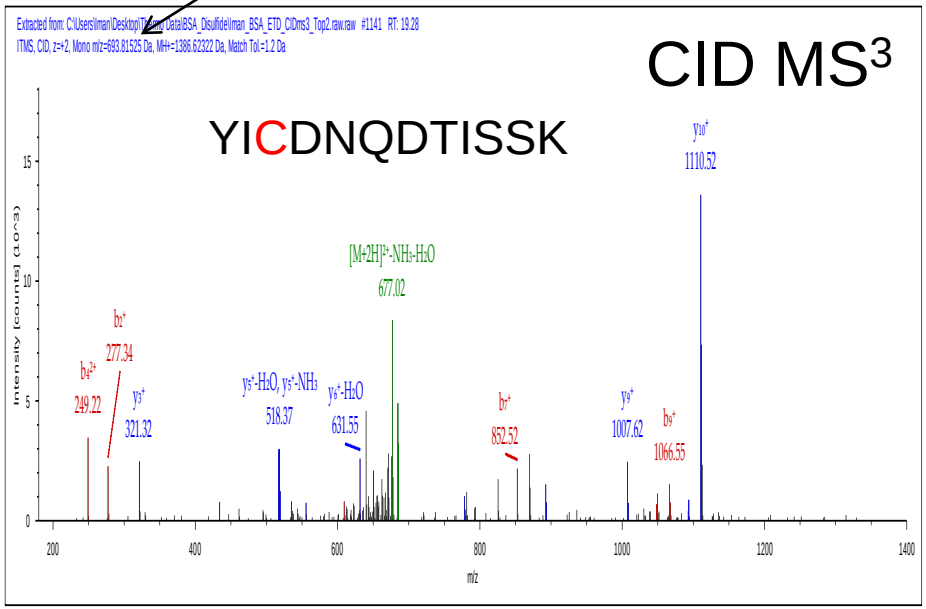
ETD triggered Top 2 MS^3 CID of disulfide-linked peptides

ETD MS^2



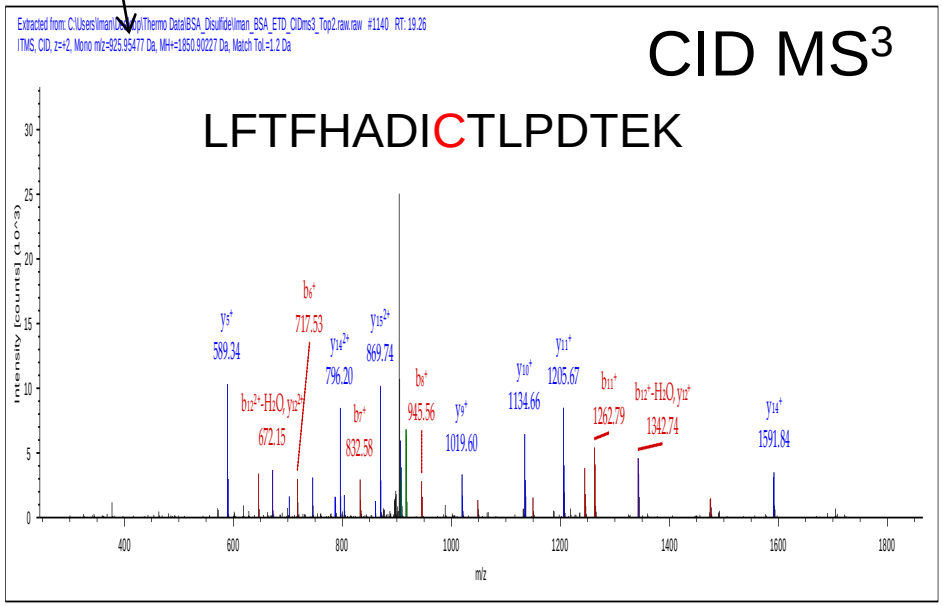
CID MS^3

YICDNQDTISSK



CID MS^3

LFTFHADICTLPDTEK



ThermoFisher SCIENTIFIC

Analysis of Complex Proteomics Datasets Containing Mixed MS/MS modes

- Flexibility, throughput, and sensitivity of Orbitrap and ion trap systems produce large, highly complex datasets that most database search software packages are unable to utilize effectively
- Some examples:
  - Data-dependent decision tree (ETD on larger peptides, CID or HCD on smaller) to increase the number of ID's in a run
  - HCD-triggered ETD for labile PTM detection (e.g. glycans)
  - MS^3 of phosphopeptides
  - ETD MS/MS-triggered CID MS^3 for disulfide mapping
  - Many different quantification experiments, including SILAC, TMT, dimethyl labeling, spectral counting, precursor ion label-free quantification

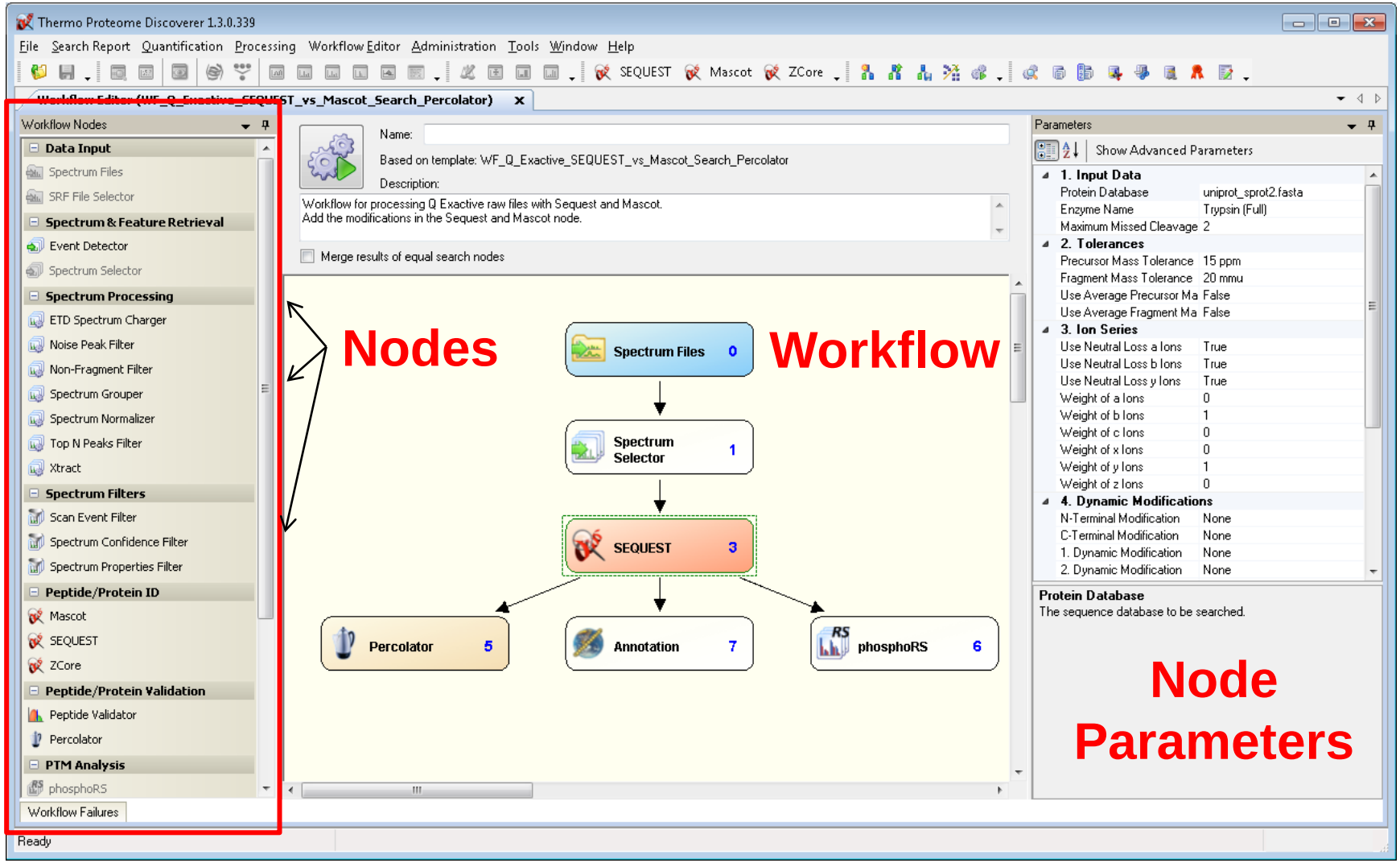
Solution: **Proteome Discoverer 1.3**

Proteome Discoverer 1.3

- **Workflow-based system for proteomics "deep sequencing"**

- **Key Features**
  - Workflow editor enables highly flexible searches
  - Easily searches complex experiments produced by hybrid Orbitrap systems (e.g. data dependent decision tree, HCD-triggered ETD for glycan analysis, ETD->CID MS^3 disulfide mapping, etc.)
  - SEQUEST and Mascot
  - Labeled quantification (SILAC, TMT)
  - Automated calculation of false discovery rate (Percolator) – New
  - Biological annotation node and automated upload to ProteinCenter – New
  - Phosphorylation site localization – New

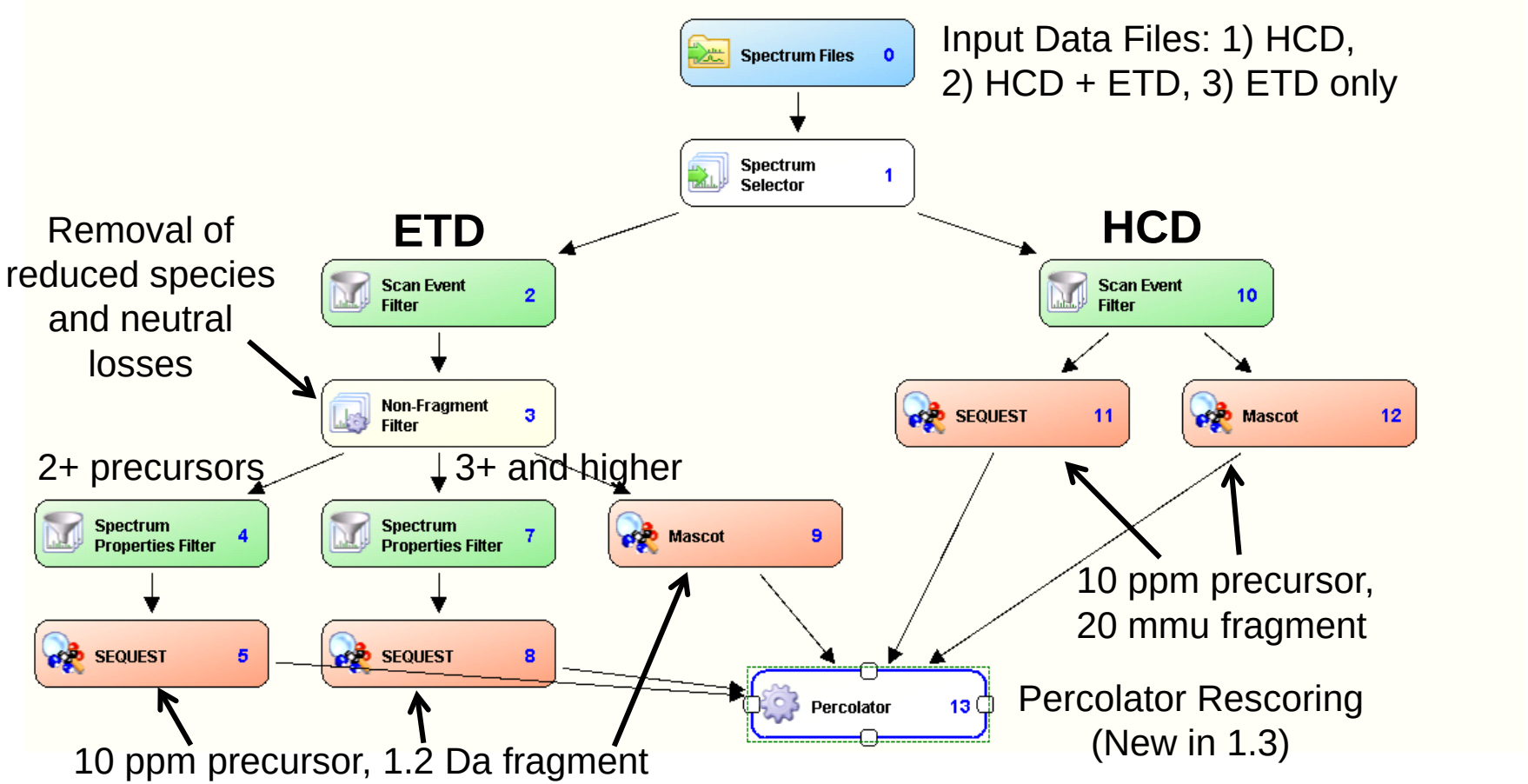
Proteome Discoverer Workflow Editor



ThermoFisher SCIENTIFIC

HCD-triggered ETD of O-GlcNAc Proteins

- Proteome Discoverer: 5 different database searches (3 SEQUEST, 2 Mascot), ETD-specific peak processing, Percolator post-processing



ThermoFisher SCIENTIFIC

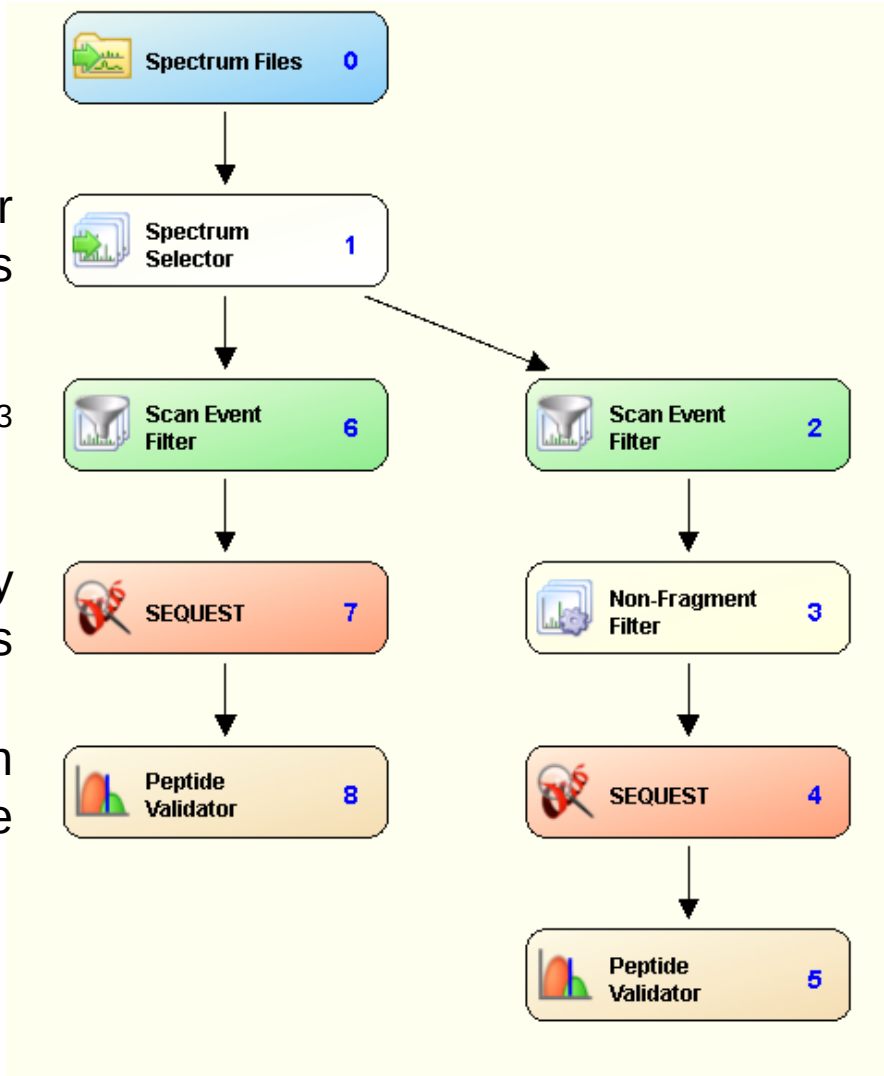
Workflow for ETD-triggered CID MS^3

MS^{n-1} precursor mass

CID MS^3

search with b, y ions

Assign confidence



ETD MS^2

Removal of charge reduced species

Search with c, y, z ions

Assign confidence

HCD-triggered ETD of O-GlcNAc Proteins

- Proteome Discoverer Results:

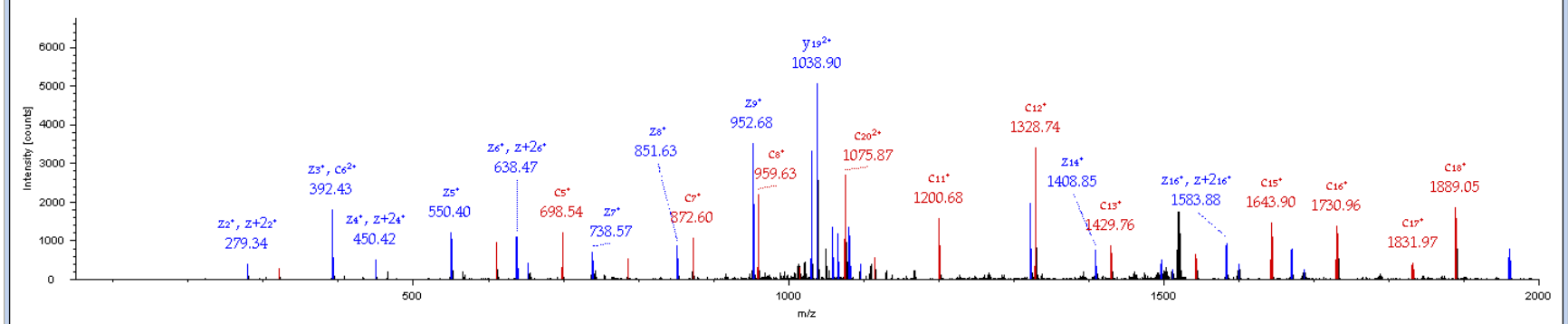
O-GlcNAc subset all mods no HexNAc HCD.msf

Proteins | Peptides | Search Input | Result Filters | Peptide Confidence | Search Summary

A3	A4	A8	A1	A1	Sequence	Protein Accessions	# Proteins	# Protein Groups	Activation Type	Modifications	IonScore	XCorr	Exp Value	Δ Score	Rank	Identity High	Homology Threshol	Char
										hexnac								
					fGVsSSSSGPSQTLTSTGNFK	P49790	1	1	ETD	N-Term(Carbami...	83		3.7E-008	1.00	1	39		
					FGVsSSSSGPSQTLTSTGNFK		1	0	ETD	S4(HexNAc)		6.67		1.00	1	38		
					FGVsSSSSGPSQTLTSTGNFK		1	0	ETD	S4(HexNAc)		6.79		1.00	1	38		
					FGVsSSSSGPSQTLTSTGNFK	P49790	1	1	ETD	S4(HexNAc)	116		1.7E-011	1.00	1	38		
					FGVsSSSSGPSQTLTSTGNFK	P49790	1	1	ETD	S4(HexNAc)	101		5.4E-010	1.00	1	38		
					GFDTSsSSSNSAASSSFK	P49790	1	1	ETD	S6(HexNAc)	74		1.4E-007	1.00	1	35		
					GFDTSSsSSNSAASSSFK		1	0	ETD	S7(HexNAc)		2.68		1.00	1	35		
					GFDTSSsSSNSAASSSFK		1	0	ETD	S7(HexNAc)		2.70		1.00	1	34		
					GFDTSSsSSNSAASSSFK	P49790	1	1	ETD	S7(HexNAc)	109		3.9E-011	1.00	1	34		
					GVSVsSSTTGLPDmTGSVYNK	Q5T6F2	2	1	ETD	S5(HexNAc), M1...		6.08		1.00	1	38		
					GVSVsSSTTGLPDmTGSVYNK	Q5T6F2	1	1	ETD	S5(HexNAc), M1...	66		2.0E-006	1.00	1	38		
					IETAVtsTPSASGQFSK		1	0	ETD	T6(HexNAc), S7(...		1.84		1.00	1			
					IETAVTsTPSASGQFSK		1	0	ETD	S7(HexNAc)		2.54		1.00	1			
					IETAVTsTPSASGQFSK		1	0	ETD	S7(HexNAc)		2.75		1.00	1			
					iGSSAPTttAANTSLMGIk	O75179	1	1	ETD	N-Term(Carbami...		3.77		1.00	1			

Fragment Match Spectrum

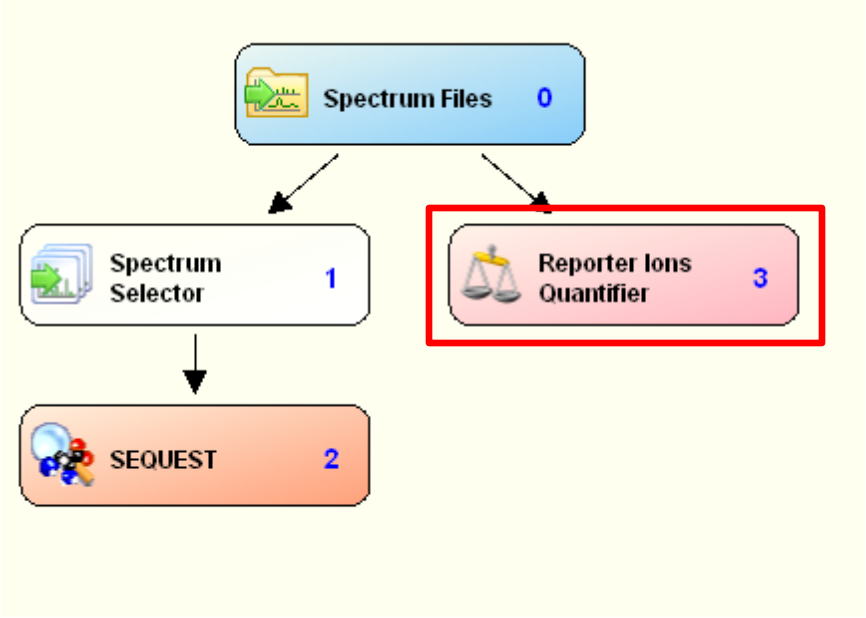
Extracted from: C:\Data\OGlcNAc\GlcNac_HCD_ETD.RAW #9442 RT: 146.01
ITMS, ETD, z=+3, Mono m/z=760.36108 Da, MH+=2279.06870 Da, Match Tol.=1.2 Da



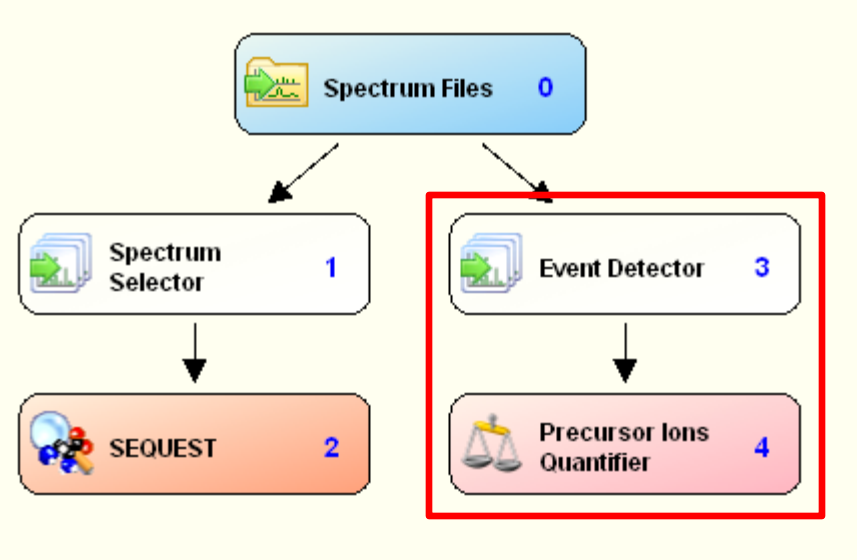
Ready | 389 Protein Group(s), 430/432 Merged Protein(s), 11983/11983 Peptide(s), 24998/24998 Search Input(s)

Expression Profiling with Proteome Discoverer

Reporter Ion Quantification (iTRAQ, TMT)



Precursor Ion Quantification (SILAC, ^{18}O, Dimethyl)



Expression ratios and variabilities for proteins and peptides:

Proteins | Peptides | Search Input | Result Filters | Peptide Confidence | Search Summary | Quantitation Summary

	127/126	127/126 Counts	127/126 Variability [%]	128/126	128/126 Counts	128/126 Variability [%]	129/126	129/126 Counts	129/126 Variability [%]	130/126	130/126 Counts
0	0.993	56;44	3.5	1.020	55;43	2.9	0.993	56;44	3.5	0.983	55;43
2	0.966	20;10	3.5	0.984	20;10	3.2	0.972	20;10	4.3	0.963	20;8
5	1.041	19;22	2.8	1.021	19;22	6.7	1.045	19;22	6.4	1.057	19;22
5	0.941	22;11	1.8	0.995	22;11	10.9	0.919	22;11	2.3	0.958	21;10
5	0.979	55;42	1.2	1.001	55;42	1.2	0.986	55;42	2.8	1.009	54;42
2											

Precursor Quan: Quan Result Display

Thermo Proteome Discoverer 1.2.0.208

File Search Report Quantification Processing Workflow Editor Administration Tools Window Help

SEQUEST Mascot ZCore

silac2noMT_AS4_pd1.2.0.208.msf*

Proteins | Peptides | Search Input | Result Filters | Peptide Confidence | Search Summary | Quantification Summary

Accession	Coverage	# PSMs	# Peptides	# AAs	MW [kDa]	calc. pI	Score	Area	Heavy/Light	Heavy/Light Count	Heavy/Light Variability [%]	
IPI:IPI00019502.3	32.35 %	94	48	1960	226.4	5.60	3617.37	3.489e7	0.301	52	15.6	Isoform 1 of Myosin
IPI:IPI00302592.2	34.79 %	95	60	2639	279.8	6.05	3047.40	2.807e7	0.266	46	14.5	Isoform 2 of Filamin
IPI:IPI00298994.6	24.12 %	58	40	2541	269.6	6.07	2235.23	4.359e6	0.394	30	18.6	Talin-1
IPI:IPI00382696.1	18.39 %	40	32	2578	275.5	5.81	1548.16	7.626e6	0.199	17	17.5	Isoform 2 of Filamin

Sequence	Modifications	IonScore	QuanResultID	Area	Quan Channel	Quan Info	Quan Usage	Heavy/Light	Light	Heavy
LVSPGSANETSSILVESVTR		111	1768	6.900e6	Light	Unique	Used	0.151	6.855e6	1.032e6
YmIGVTYGGDDIPLSPYR	M2(Oxidation)	94	1865	7.719e5	Light	Excluded by Method	Not Used		2.253e5	
GAGTGGLGLTVEGPdEAK	C15(Carbamido...	89	1268	3.913e6	Light	Excluded by Method	Not Used		1.785e6	3.084e5
SPFVVQVGEAcNPNAdR	C11(Carbamido...	88	1351	2.122e6	Light	Unique	Used	0.193	1.931e6	3.732e5

Peptide Ratio Distributions

IPI:IPI00382696.1: Isoform 2 of Filamin-B

Heavy/Light

R = -2.33 | 0.20
IQR = 0.29 | 0.04

log2 Fold Change

Quantification Spectrum

C:\data\SILAC 2plex - Arg10, Lys6\silac2noMT_AS4.RAW, RT=51.73-52.79 min
Event Spectrum: FTMS, Quantified Ion: z=+2, Mono m/z=1023.53943 Da, MH+=2046.07158 Da

1024.54336 z=2 QuanID=1768 Light cRT=52.31
1025.04466 z=2 QuanID=1768 Light cRT=52.32
1025.55018 z=2 Light cRT=52.28
1026.54381 cRT=52.69
1027.54801 cRT=52.64
1027.99249 cRT=51.92
1028.49968 cRT=51.87
1028.54367 z=2 QuanID=1768 Heavy cRT=52.23
1030.05023 z=2 QuanID=1768 Heavy cRT=52.32

Intensity [counts] (10^6)

m/z

Quan Channel Values

6855041

1032134

Area [counts · min] (10^6)

Light

Heavy

Quan Channels

Ready

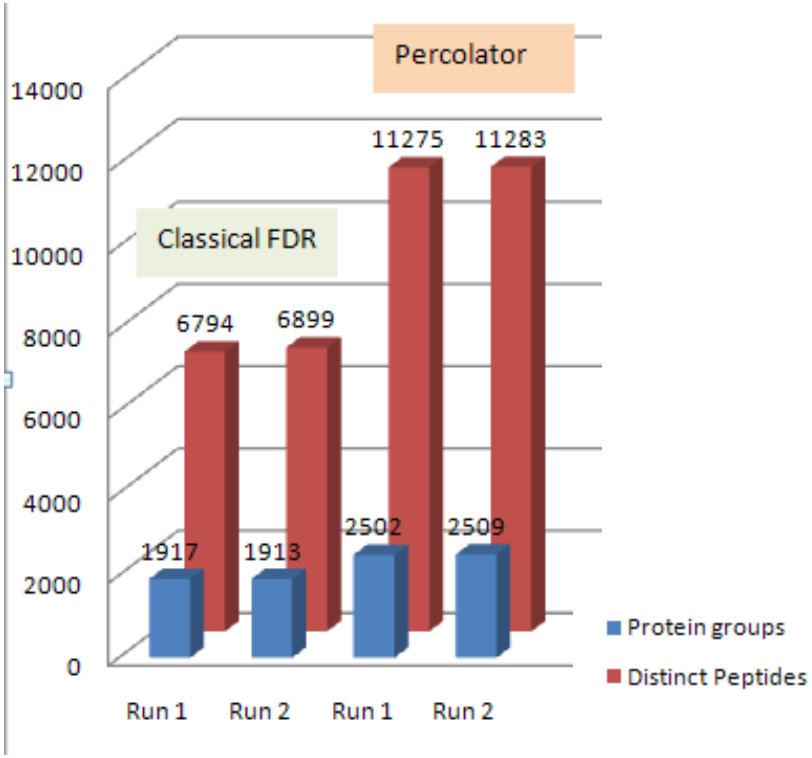
293 Protein Group(s), 1392/2099 Protein(s), 1595/7656 Peptide(s), 5955/5955 Search Input(s)

Deep data mining using Percolator

- Percolator uses >30 features of a peptide spectral match (PSM) to distinguish true positives from random matches
- Result: more peptides and proteins identified with high confidence (1% FDR)
  - CID, HCD: >30% increase in PSM's
  - ETD: Up to 80% increase in PSM's
- Publication
  - Käll *et al*, Nature Methods 4:923-925 (2007)

ASMS poster MP25

Breaking the 2000 proteins barrier in a standard LC run using a new benchtop Orbitrap instrument and multiple search engines.



ThermoFisher SCIENTIFIC

Percolator in Proteome Discoverer

Thermo Proteome Discoverer 1.3.0.339

File Search Report Quantification Processing Workflow Editor Administration Tools Window Help

SILAC3.msf | Workflow Editor

Proteins | Peptides | Search Input | Result Filters | Peptide Confidence | Search Summary | Quantification Summary

	A2	Sequence	# PSMs	# Proteins	# Protein Groups	Protein Group Accessions	Modifications	q-Value	PEP	IonScore	Exp Value	ΔCn
4911		RTsLSYLNK	2	1	1	YBL060W	S3(Phospho)	0.00598	0.105	22	7.1E-002	0.0000
4912		EcADEMktTPk	1	1	1	YPR018W	C2(Carbamidomethyl); K7(	0.00598	0.105	13	3.5E-001	0.0000
4913		ksSPATkVPSkPDr	4	1	1	YGL207W	K1(Lys8); S2(Phospho); K7	0.00605	0.106	17	2.3E-001	0.0000
4914		ALDAsNAIDR	3	1	1	YJR068W	S5(Phospho)	0.00607	0.106	20	7.8E-002	0.0000
4915		kkDAsQEESLI	2	1	1	YBR127C	K1(Lys8); K2(Lys8); S5(Ph.	0.00609	0.107	13	5.5E-001	0.0000
4916		tcrHFISVILNSRGLETGK	2	1	1	YBR214W	T1(Phospho); C2(Carb	0.00609	0.107	13	1.4E+000	0.0000
4917		MsPVLTtPkr	2	1	1	YIL106W	S2(Phospho); T7(Phospho)	0.00616	0.107	20	1.3E-001	0.0000
4918		FLDkLGLsR					K4(Lys8); S9(Phospho)	0.00629	0.11	15	3.1E-001	0.0000
4919		KKDAsQEEsLI					S5(Phospho); S9(Phospho)	0.00636	0.111	25	2.6E-002	0.0000
4920		NSTPSDASSTKNtDHIV					T13(Phospho)	0.00641	0.112	12	1.0E+000	0.0000
4921		HLNTItLTk					T6(Phospho); K9(Lys8)	0.00643	0.112	18	1.8E-001	0.0000
4922		TkPAEEksAEPEVk					K2(Lys8); K7(Lys8); S8(Ph.	0.00643	0.112	11	1.1E+000	0.0000
4923		FSNGGASsR					S8(Phospho)	0.00647	0.113	14	8.9E-002	0.0000
4924		LLDYFk	1	1	1	YLR355C	K6(Lys8)	0.0065	0.114	16	1.5E-001	0.0000
4925		RLsGIMR	1	1	1	YBR156C	S3(Phospho)	0.00661	0.114	28	1.1E-002	0.0000

q-values and PEPs for peptide groups

Peptide Group Members

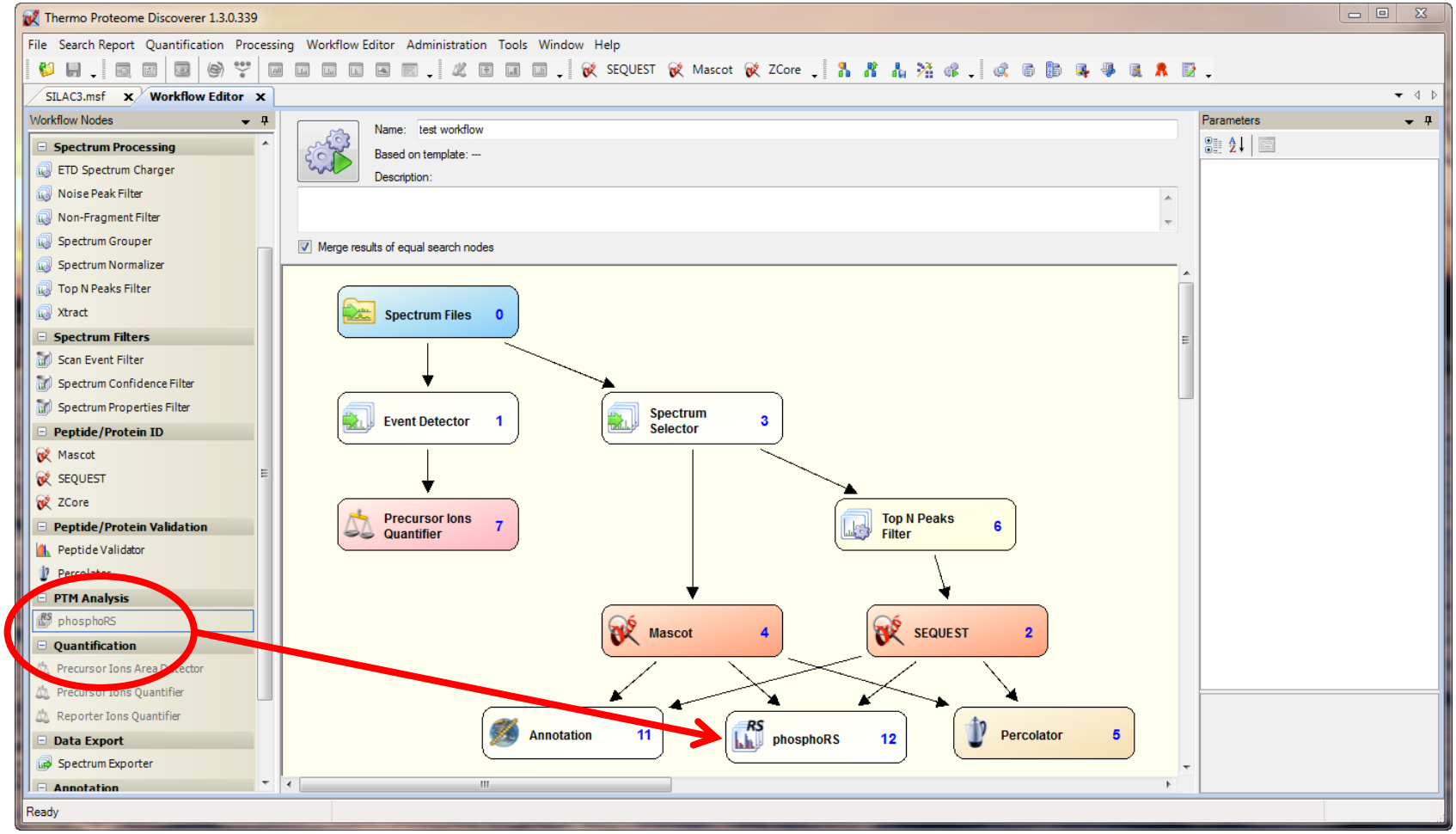
	Sequence	PSM Ambiguity	Protein Group Accessions	Modifications	Rank	q-Value	PEP	IonScore	ΔScore	ΔCn	Exp Value	Search Eng
1	ksSPATkVPSkPDr	Unambiguous	YGL207W	K1(Lys8); S2(Phospho); K7...	1	0.00605	0.106	17	0.2353	0.0000	2.3E-001	
2	kSsPATkVPSkPDr	Unambiguous	YGL207W	K1(Lys8); S3(Phospho); K7...	1	0.0357	0.572	15	1.0000	0.0000	3.8E-001	
3	ksSPATkVPSkPDr	Unconsidered	YGL207W	K1(Lys8); S2(Phospho); K7...	1	0.0641	0.843	15	1.0000	0.0000	3.8E-001	
4	kSsPATkVPSkPDr	Unconsidered	YGL207W	K1(Lys8); S3(Phospho); K7...	2			13		0.2353	5.7E-001	

q-values and PEPs for PSMs

Ready (s), 5758/5758 Peptide(s), 42326/42326 PSM(s), 70574/70574 Search Input(s)

phospho*RS* node in Proteome Discoverer

- Developed in conjunction with Karl Mechtler's group at University of Vienna for determination of phosphorylation site confidence



ThermoFisher SCIENTIFIC

phospho*RS* in Proteome Discoverer

Thermo Proteome Discoverer 1.3.0.339

File Search Report Quantification Processing Workflow Editor Administration Tools Window Help

SEQUEST Mascot ZCore

SILAC3.msf | Workflow Editor

Proteins | Peptides | Search Input | Result Filters | Peptide Confidence | Search Summary | Quantification Summary

	A2	Sequence	# PSMs	# Proteins	# Protein Groups	Protein Group Accessions	Modifications	pRS Score	pRS Probability	pRS Site Probabilities
4911		RTsLSYLNK	2	1	1	YBL060W	S3(Phospho)	8	4.0 %	T(2): 4.0; S(3): 4.0; S(5): 46.0; Y(6): 46.0
4912		EcADEMktTPk	1	1	1	YPR018W	C2(Carbamidomethyl); K7(...	20	84.2 %	T(8): 84.2; T(9): 15.8
4913		ksSPATkVPSkPDr	4	1	1	YGL207W	K1(Lys8); S2(Phospho); K7...	24	32.3 %	S(2): 32.3; S(3): 32.3; T(6): 32.3; S(10):...
4914		ALDAsNAIDR	3	1	1	YJR068W	S5(Phospho)	39	100.0 %	S(5): 100.0
4915		kkDAsQEESLI	2	1	1	YBR127C	K1(Lys8); K2(Lys8); S5(Ph...	40	98.1 %	S(5): 98.1; S(9): 1.9
4916		tcrHFISVILNSRGLETGK	2	1	1	YBR214W	T1(Phospho); C2(Carbam...	54	94.2 %	T(1): 94.2; S(7): 5.8; S(12): 0.0; T(17): 0.0
4917		MsPVLTtPkr	2	1	1	YIL106W	S2(Phospho); T7(Phospho)...	73	50.0 %	S(2): 100.0; T(6): 50.0; T(7): 50.0
4918		FLDkLGLsR	1				K4(Lys8); S8(Phospho)	46	100.0 %	S(8): 100.0
4919		KKDAsQEEsLI	1				S5(Phospho); S9(Phospho)	80	100.0 %	S(5): 100.0; S(9): 100.0
4920		NSTPSDASSTKNtDHIV	2				T13(Phospho)	23	8.7 %	S(2): 0.1; T(3): 0.3; S(5): 1.3; S(8): 8.7;...
4921		HLNTItLTk	3				T6(Phospho); K9(Lys8)	34	10.3 %	T(4): 89.5; T(6): 10.3; T(8): 0.3
4922		TkPAEEksAEPEVk	1				K2(Lys8); K7(Lys8); S8(Ph...	26	88.4 %	T(1): 11.6; S(8): 88.4
4923		FSNGGASsR	1				S8(Phospho)	36	89.3 %	S(2): 0.1; S(7): 10.6; S(8): 89.3
4924		LLDYFk	1	1	1	YLR355C	K6(Lys8)			
4925		RLsGIMR	1	1	1	YBR156C	S3(Phospho)	60	100.0 %	S(3): 100.0
4926		AINGLtMLK	1	0	0		T6(Phospho)	77	100.0 %	T(6): 100.0
4927		EASksPISSFVNDYr	3	1	1	YDR096W	K4(Lys8); S5(Phospho); R1...	21	16.2 %	S(3): 16.2; S(5): 16.2; S(8): 62.4; S(9): 4...

pRS scores for peptide groups

Peptide Group Members

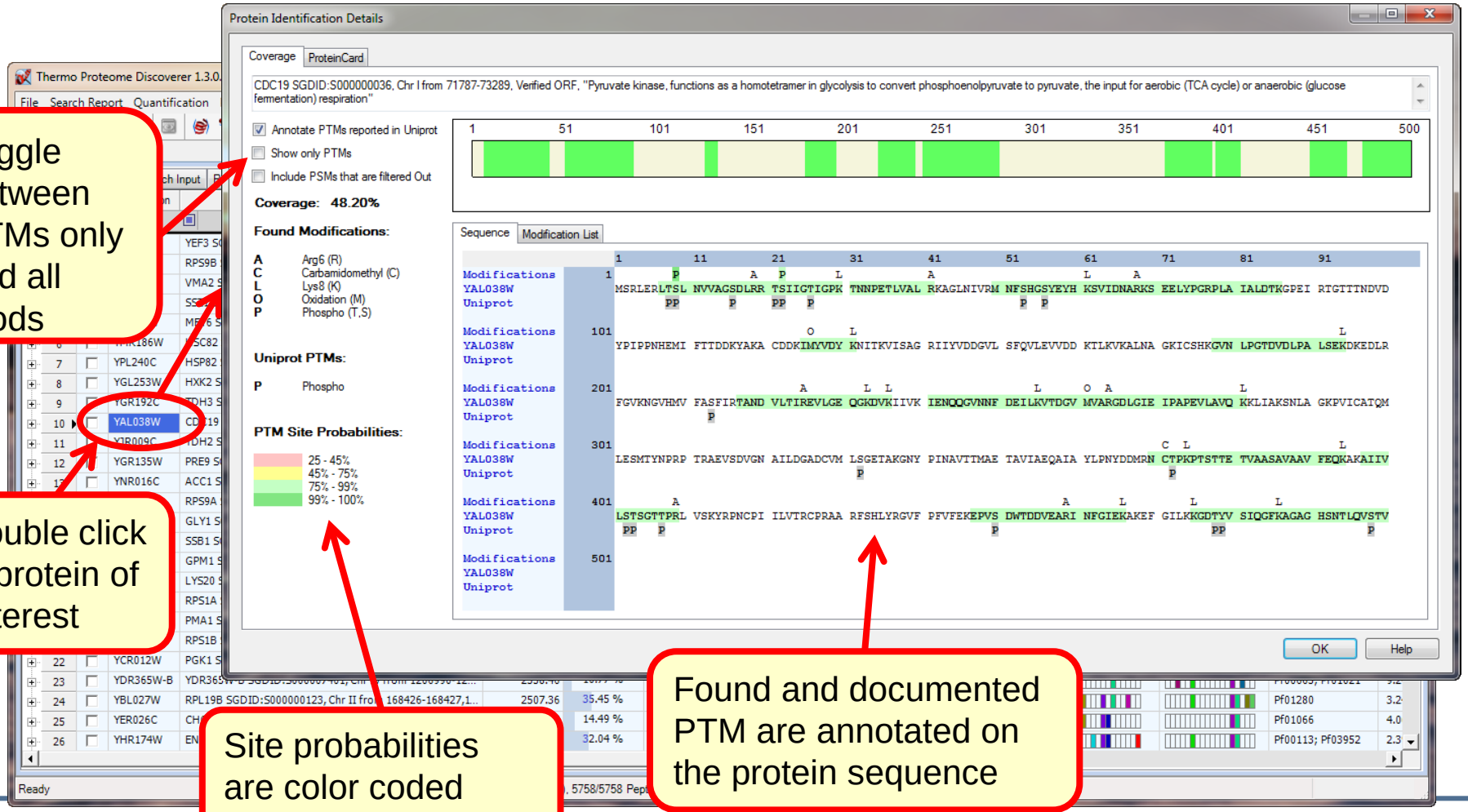
		Sequence	PSM Ambiguity	Protein Group Accessions	Modifications	Rank	pRS Score	pRS Probability	pRS Site Probabilities	q-Value
1		ksSPATkVPSkPDr	Unambiguous	YGL207W	K1(Lys8); S2(Phospho); K7...	1	24	32.3 %	S(2): 32.3; S(3): 32.3; T(6): 32.3; S(10): 3.1	0.00605
2		kSsPATkVPSkPDr	Unambiguous	YGL207W	K1(Lys8); S3(Phospho); K7...	1	14	31.8 %	S(2): 31.8; S(3): 31.8; T(6): 31.8; S(10): 4.6	0.0357
3		ksSPATkVPSkPDr	Unconsidered	YGL207W	K1(Lys8); S2(Phospho); K7...	1	14	31.8 %	S(2): 31.8; S(3): 31.8; T(6): 31.8; S(10): 4.6	0.0641
4		kSsPATkVPSkPDr	Unconsi...		S3(Phospho); K7...	2	24	32.3 %	S(2): 32.3; S(3): 32.3; T(6): 32.3; S(10): 3.1	

pRS scores for PSMs

Ready ... s), 5758/5758 Peptide(s), 42326/42326 PSM(s), 70574/70574 Search Input(s)

Visualization of Found and Known PTMs

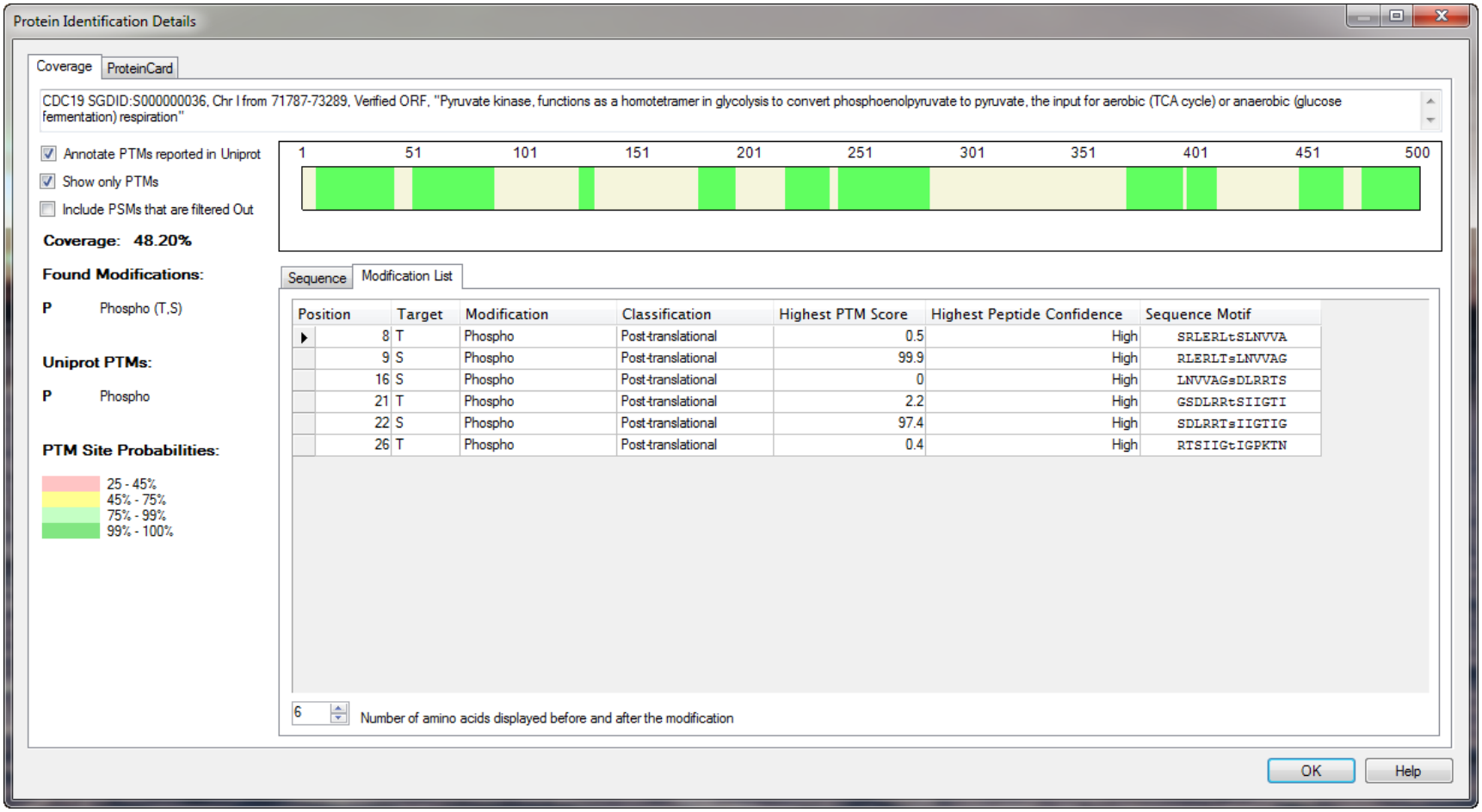
- The Protein ID Details view now presents a comprehensive overview of all modifications found in the MS/MS data of each protein



ThermoFisher SCIENTIFIC

Visualization of Identified PTMs

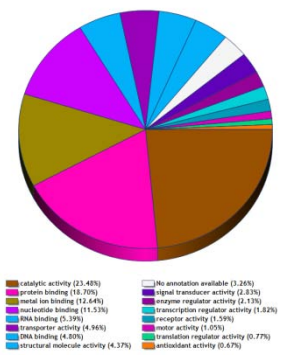
- Find detailed information about the found PTMs in the Modification List



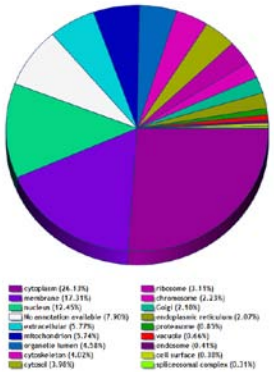
Providing biological context using the Annotation node

- 3 year support subscription includes annotation node:
  - Automatically queries hosted ProteinCenter server in Denmark for GO, protein family (Pfam), PTM, and ProteinCard annotations
    - All can be filtered
    - Displayed in the PD Viewer
  - Previous results can be annotated or re-annotated

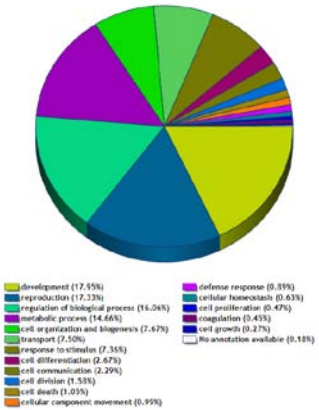
Molecular Function

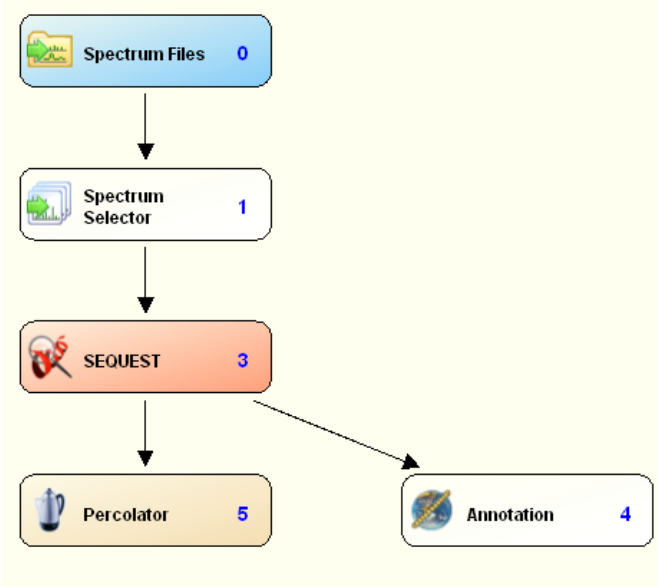


Cellular Component



Biological Process

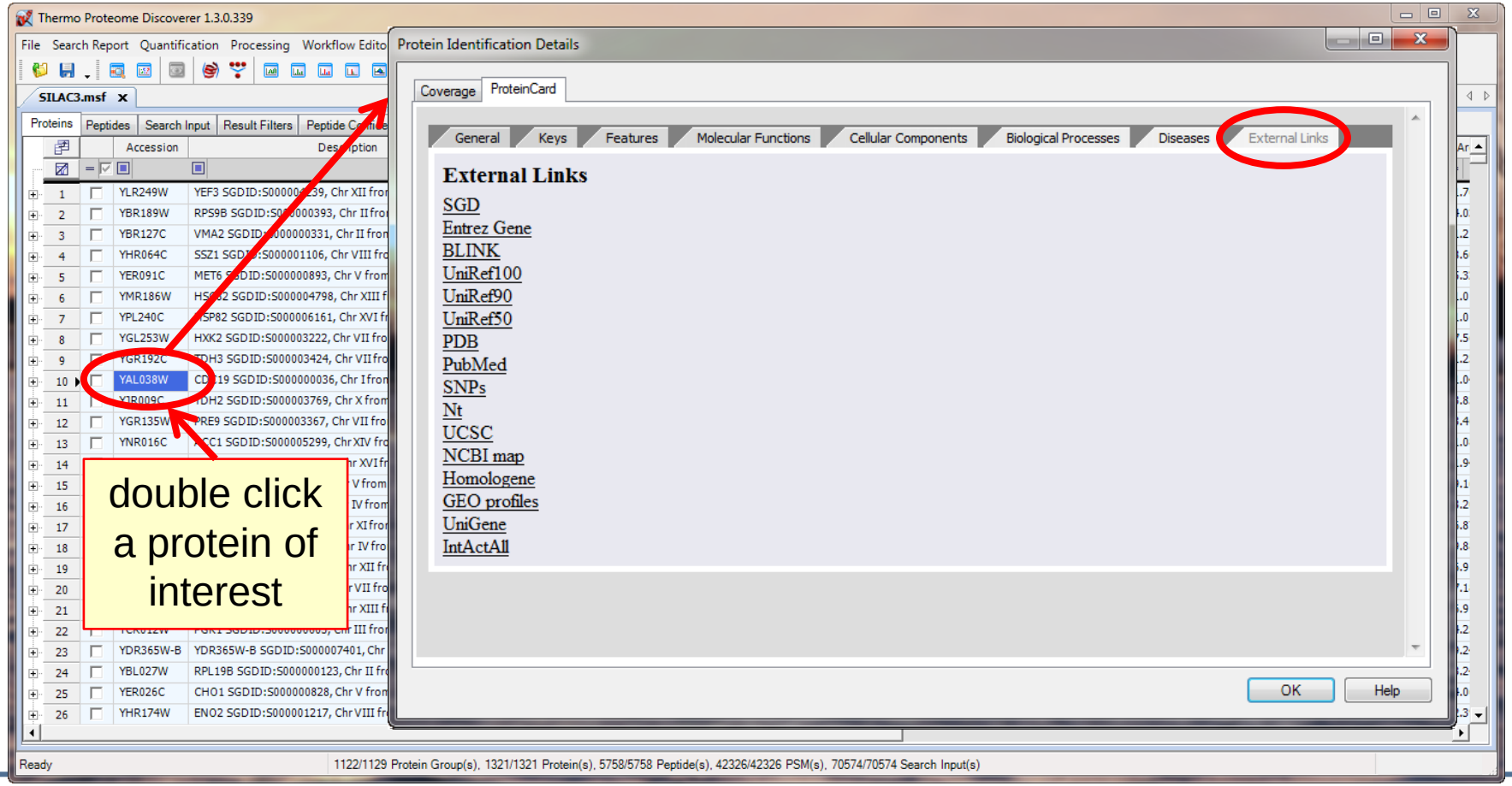




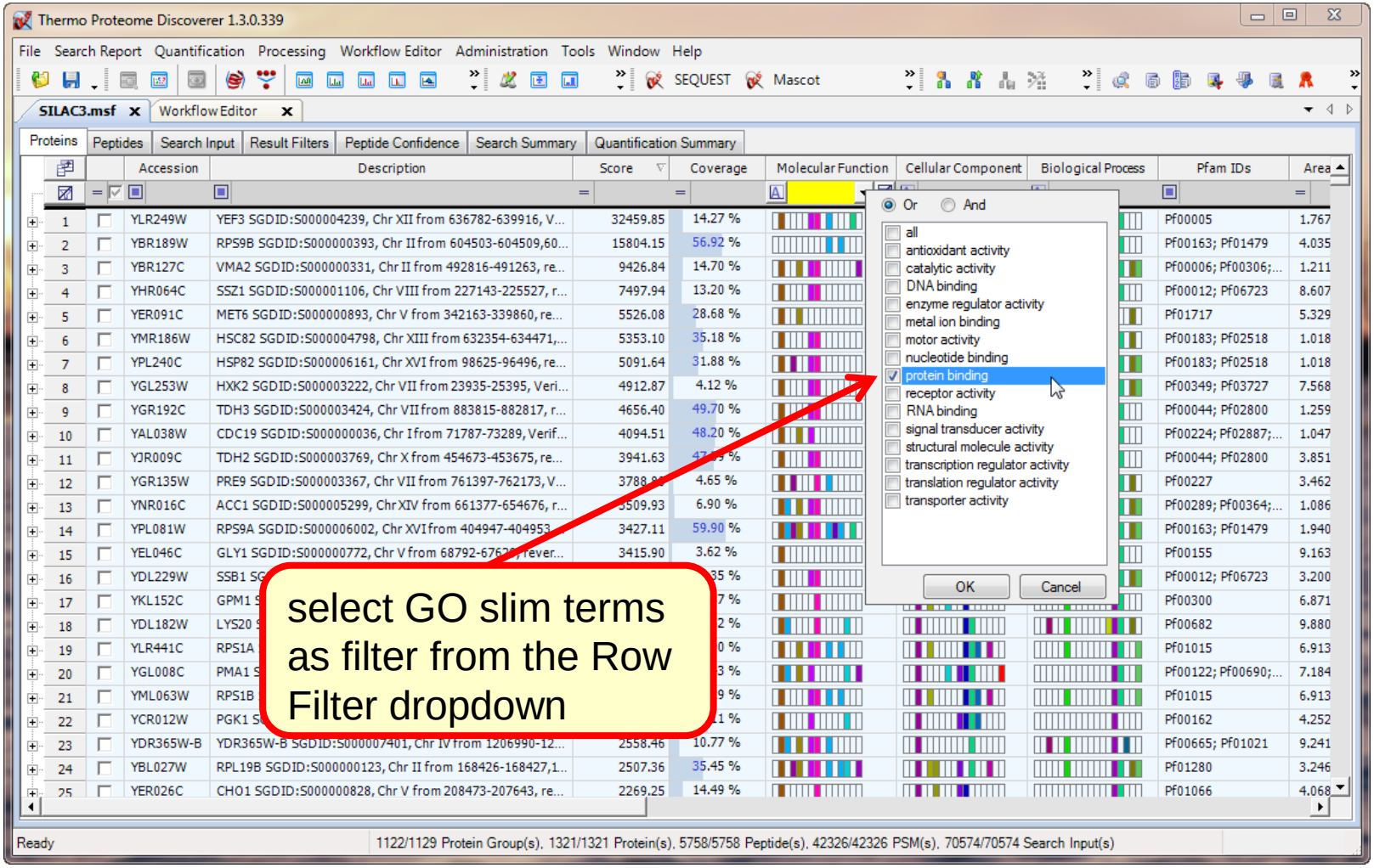
ASMS Poster ThP22 401
"Integration of a central protein repository into a standard data processing application for mining proteomics data"

ProteinCard for Identified Proteins

- ProteinCard can be displayed for every identified protein whose accession is tracked in ProteinCenter



Filtering identified proteins using GO and Pfam Annotations



ThermoFisher SCIENTIFIC

Annotated PTM's from UniProt

Coverage | ProteinCard

TDH2 SGDID:S000003769, Chr X from 454673-453675, reverse complement, Verified ORF, "Glyceraldehyde-3-phosphate dehydrogenase, isozyme 2, involved in glycolysis and gluconeogenesis; tetramer that catalyzes the reaction of glyceraldehyde-3-phosphate to 1,3 bis-phosphoglycerate; detected in the cytoplasm and cell wall"

- [x] Annotate PTMs reported in Uniprot
- [x] Show only PTMs
- [ ] Include PSMs That Are Filtered Out

Coverage: 47.59%

1 51 101 151 201 251 301 332

Found Modifications:

P Phospho (Y,T,S)

Uniprot PTMs:

P Phospho

PTM Site Probabilities:

25 - 45%
45% - 75%
75% - 99%
99% - 100%

Sequence | Modification List

```
                    1         11        21        31        41        51        61        71        81        91
Modifications     1
YJR009C             MVRVAINGFG RIGRLVMRIA LQRKNVEVVA LNDPFISNDY SAYMFKYDST HGRYAGEVSH DDKHIIVDGH KIATFQERDP ANLPWASLNI DIAIDSTGVF
Uniprot                                                           PP    P    P

Modifications   101                                   P         P  PP                                  P         P
YJR009C             KELDTAQKHI DAGAKKVVIT APSSTAPMFV MGVNEEKYTS DLKIVSNASC TTNCLAPLAK VINDAFGIEE GLMTTVHSMT ATQKTVDGPS HKDWRGGRTA
Uniprot                                                           P

Modifications   201 P     PPP                 P
YJR009C             SGNIIPSSTG AAKAVGKVLP ELQGKLTGMA FRVPTVDVSV VDLTVKLNKE TTYDEIKKVV KAAAEGKLKG VLGYTEDAVV SSDFLGDSNS SIFDAAAGIQ
Uniprot             P

Modifications   301
YJR009C             LSPKFVKLVS WYDNEYGYST RVVDLVEHVA KA
Uniprot                     P         PP
```

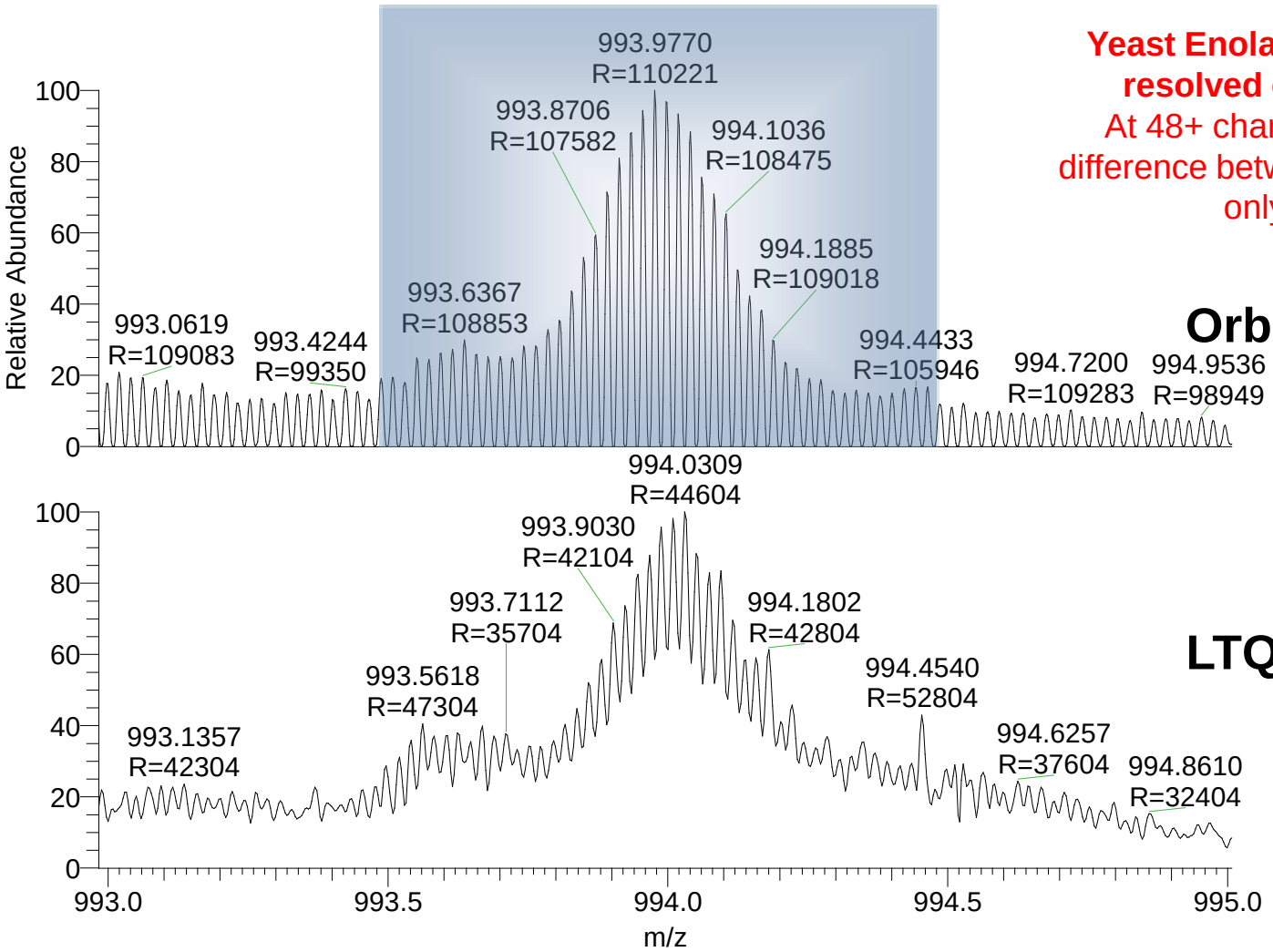
Known phosphorylation sites from UniProt annotated in protein sequence

Conclusions

- Proteome Discoverer 1.3 has the most comprehensive set of tools for analysis of proteomics data from Thermo mass spectrometers
- The new nodes developed in collaboration with 3rd parties demonstrate the flexibility of the Proteome Discoverer node-based workflow engine and hint at the expansion in the number of tools available for our customers
- Integration with ProteinCenter annotation adds biological inference to SEQUEST and Mascot search results

Supplementary Slides

Intact protein analysis example: Yeast Enolase (48 kDa)



Yeast Enolase, 48+ is baseline resolved on Orbitrap Elite

At 48+ charge state, the mass difference between isotopic peaks is only 0.021 Th!

Orbitrap Elite

LTQ Orbitrap Velos

Resolution 240000@400

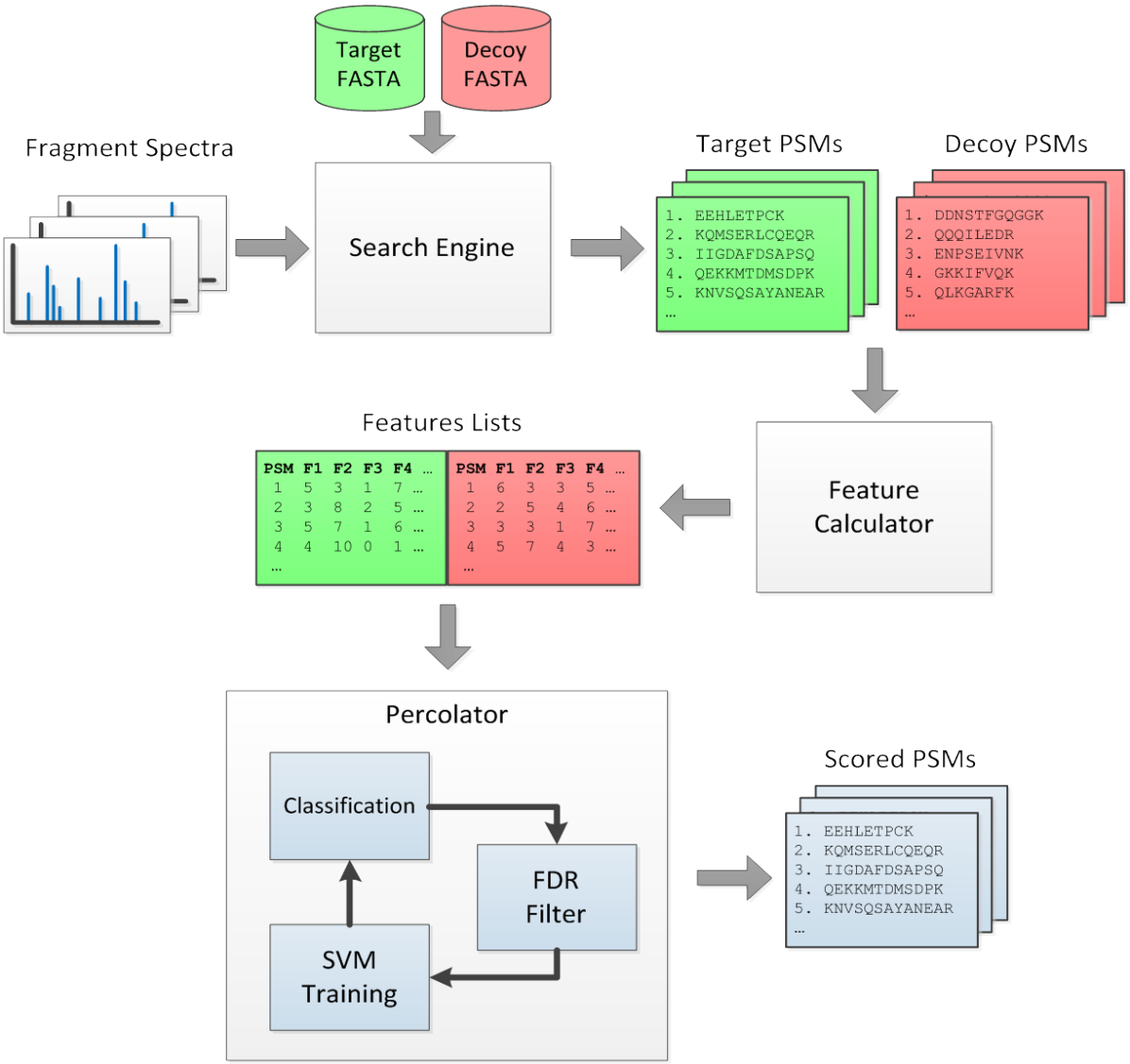
Improving protein analysis in Orbitrap mass spectrometry
Eugen Damoc; Eduard Denisov; Oliver Lange; Thomas Moehring; Alexander Makarov. **ASMS 2011, MP560**

Percolator

- The Percolator algorithm (http://per-colator.com) uses a set of features related to the quality of the peptide-spectrum matches (PSMs) with a semi-supervised method to train a machine learning algorithm called a support vector machine (SVM) to discriminate between correct and incorrect matches
- Does not require any expert-driven or subjective decisions, thereby eliminating any artificial biases
- The learnt classifier is specifically adapted and unique for each data set, thus, adapting to variations in data quality, protocols and instrumentation
- Percolator improves the sensitivity of existing database search algorithms at a constant false discovery rate
- Furthermore, Percolator assigns a statistically meaningful q-value to each PSM, as well as the probability of the individual PSM being incorrect

L. Käll, et al. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat Methods,* **2007***, 4*, 923-925

Percolator



q-values and Posterior Error Probabilities

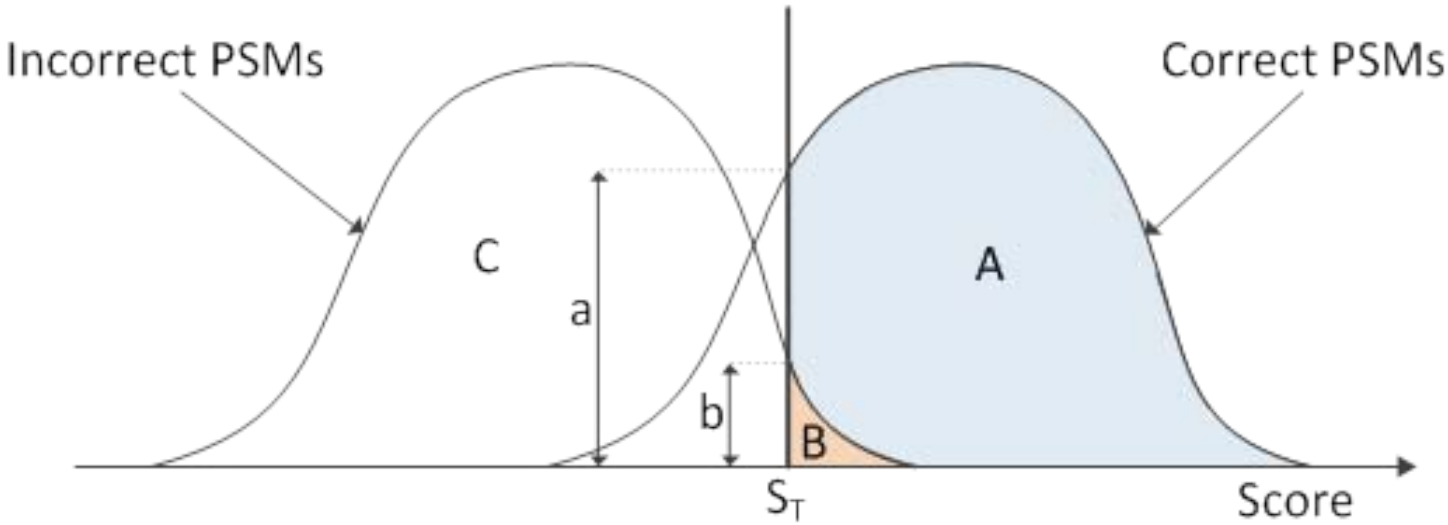
q-value

- the minimal FDR at which the identification is deemed correct
- Although associated with a single PSM, the q-value depends upon the data set in which the PSM occurs
- A q-value of 0.01 for peptide EAMRQPK matching spectrum s means that, if we try all possible FDR thresholds, then 1% is the minimal FDR at which this PSM will appear in the output list

Posterior Error Probability (PEP)

- quite simply, the probability that the observed PSM is incorrect.
- The PEP can be thought of as a local version of the FDR (“local FDR”). Whereas the FDR measures the error rate associated with a collection of PSMs, the PEP measures the probability of error for a single PSM.
- if the PEP of PSM (EAMRPK,s) is 5%, there is a 95% chance that peptide EAMRPK was in the spectrometer when the spectrum s was generated

q-values and Posterior Error Probabilities



$$FDR = \frac{A}{A + B}$$

$$PEP = \frac{a}{a+b}$$

L. Käll, et al., Posterior error probabilities and false discovery rates: two sides of the same coin., *J Proteome Res,* **2008***, 7*, 40-44

q-value or PEP: Which one is better?

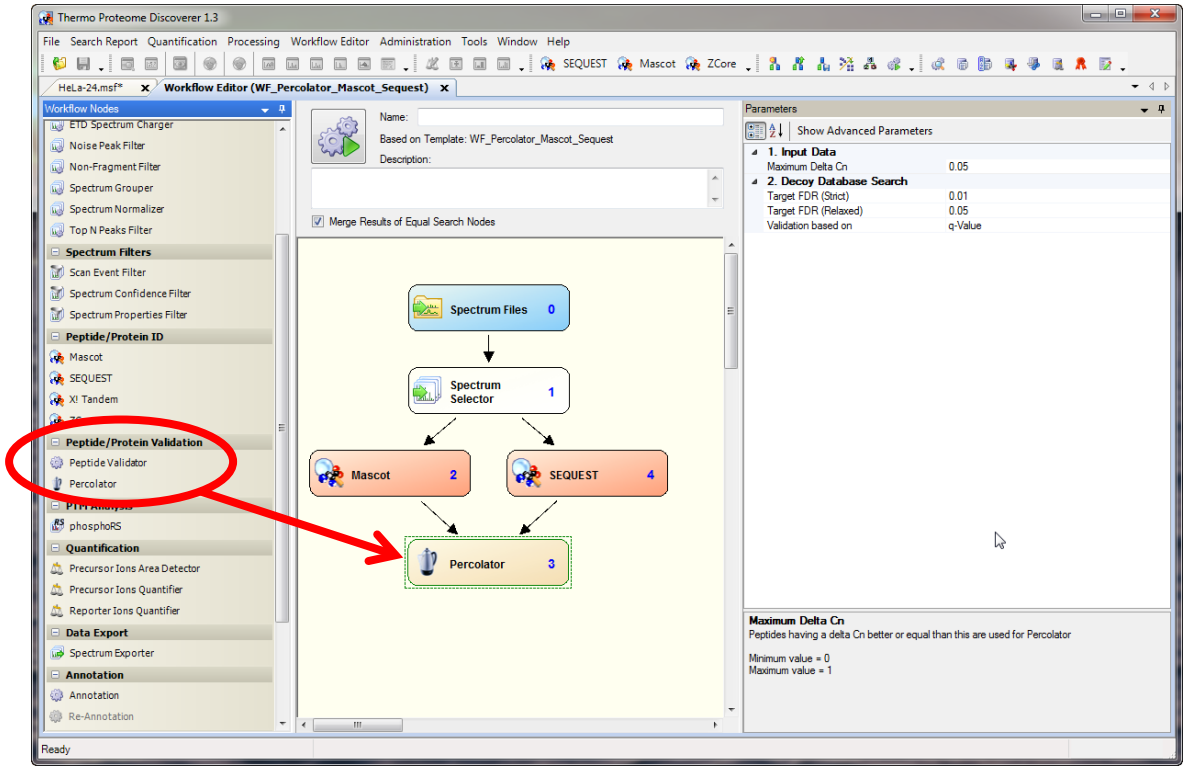
PEPs and q-values are complementary, and are useful in different situations:

- If you are interested in determining which proteins are expressed in a certain cell type under a certain set of conditions, or if your follow-up analysis will involve looking at groups of PSMs, for example, considering all proteins in a known pathway, evaluating enrichment with respect to GO categories, or performing experimental validation on a group of proteins, then the q-value is an appropriate measure.
- If the goal of your experiment instead is to determine the presence of a specific peptide or protein, then the PEP is more relevant. For example, imagine that you are interested in determining whether a certain protein is expressed in a certain cell type under a certain set of conditions. In this scenario you should examine the PEPs of your detected PSMs. Likewise, imagine that you have identified a large set of PSMs using a q-value threshold, and among them, you identify a single PSM that is intriguing. Before deciding to dedicate significant resources to investigating a single result, you should examine the PEP associated with that PSM. This is because, although the q-value associated with that PSM may be 0.01, the PEP is always greater than or equal to 0.01. In practice, the PEP values for PSMs near the q = 0.01 threshold are likely to be much larger than 1%.

L. Käll, et al., Posterior error probabilities and false discovery rates: two sides of the same coin., *J Proteome Res,* **2008***, 7*, 40-44

Percolator in Proteome Discoverer

- Available as a new node under “Peptide Validation”



<u>Note:</u> Percolator needs a sufficient number of PSMs from the target and the decoy search. If less than 200 target or decoy PSMs were identified, the percolation is rejected. The same is true if less than 20% decoy PSMs are available compared to the number of target matches.

Percolator in Proteome Discoverer

- Proteome Discoverer uses a set of more than 30 features describing the quality of a PSM

Search engine/scoring specific features	
Mascot Ions Score	Only used for Mascot searches
SEQUEST Xcorr	Only used for SEQUEST searches
SEQUEST	
X! Tandem	
Delta Cn	
Binomial Sc	

Peptide/Precursor related features	
% Isolation Interference	Fraction of ion current in the isolation window not attributed to the identified precursor
MH+ [Da]	
Delta Mass	
Delta Mass	
Absolute De	
Absolute De	
Peptide Len	
Is z=1	
Is z=2	
Is z=3	
Is z=4	
Is z=5	
Is z>5	

Fragment series related features	
Fragment Coverage Series A, B, C [%]	Coverage of the N-terminal fragment ion series. The coverage is separately calculated for each series used and the maximum coverage is used.
Fragment Coverage Series X, Y, Z [%]	Coverage of the C-terminal fragment ion series.
Log Matched	
Log Matched	
Longest Sequ	
Longest Sequ	

Fragment related features	
IQR Fragment Delta Mass [Da]	Inter-quartile range of the distribution of mass errors in Dalton of all fragments considered.
IQR Fragment Delta Mass [ppm]	Inter-quartile range of the distribution of mass errors in ppm of all fragments considered.
Mean Fragment Delta Mass [Da]	Arithmetic mean range of the distribution of
Mean Fragm	
Mean Absol	
Mean Absol	

Digestion related features	
# Missed Cleavages	Number of missed cleavages
FASTA related features	
Log Peptides Matched	Logarithm of the number of candidates in the precursor mass window
Spectrum related features	
Log Total Intensity	Logarithm of the total ion current of the fragment spectrum
Fraction Matched Intensity [%]	Fraction of the total ion current of the fragment spectrum that is matched by fragments of the PSM

Percolator in Proteome Discoverer

Thermo Proteome Discoverer 1.3.0.339

File Search Report Quantification Processing Workflow Editor Administration Tools Window Help

SEQUEST Mascot ZCore

SILAC3.msf | Workflow Editor

Proteins | Peptides | Search Input | Result Filters | Peptide Confidence | Search Summary | Quantification Summary

Percolator

Set Filter Type | Apply Filters | Apply FDRs

Processing Node Name	Workflow Name	Short Display Name
Mascot	SILAC3	A2

Modest Confidence Filter Settings

FDR Settings

Threshold	0.05
Validation based on	q-Value

High Confidence Filter Settings

FDR Settings

Threshold	0.01
Validation based on	q-Value

Threshold
Peptides having a threshold better than the threshold are valid.

Minimum value = 0.0
Maximum value = 1.0

Threshold
Peptides having a threshold better than the threshold are valid.

Minimum value = 0.0
Maximum value = 1.0

Ready | 1122/1129 Protein Group(s), 1321/1321 Protein(s), 5758/5758 Peptide(s), 42326/42326 PSM(s), 70574/70574 Search Input(s)

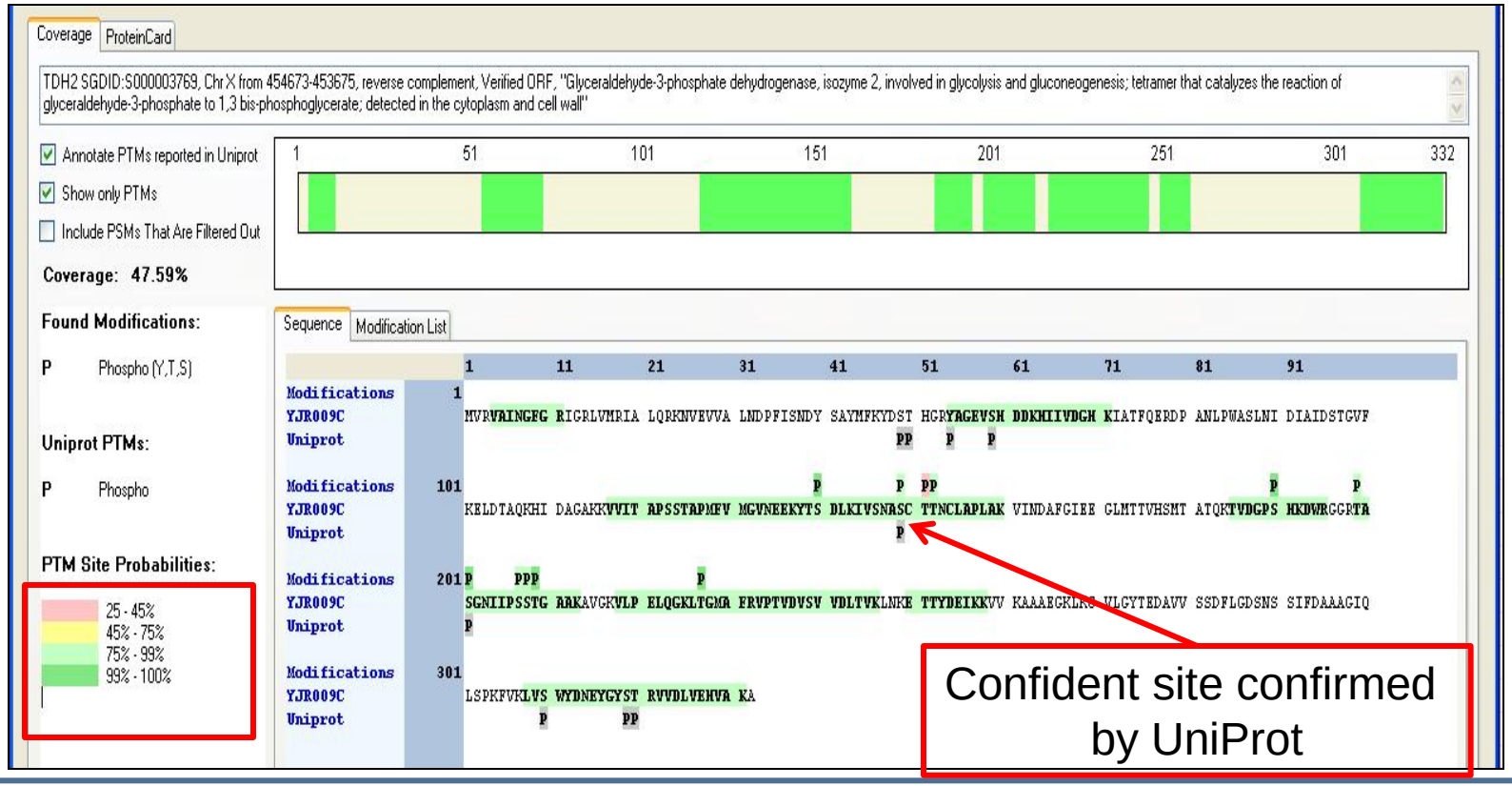
Define whether the confidence estimation is based on Percolator scores (default when Percolator was used) or classical FDR on score thresholds

Click “Apply Filters” after changes

Define the q-value (or PEP) thresholds to devide PSMs in high, medium, and low confident hits on the Peptide Confidence tab

Confident PTM analysis Using PhosphoRS

- Developed in collaboration with Karl Mechtler's lab at University of Vienna for phosphorylation site confidence determination
  - phosphoRS Score: is it phosphorylated?
  - site probability Score: confidence in site localization

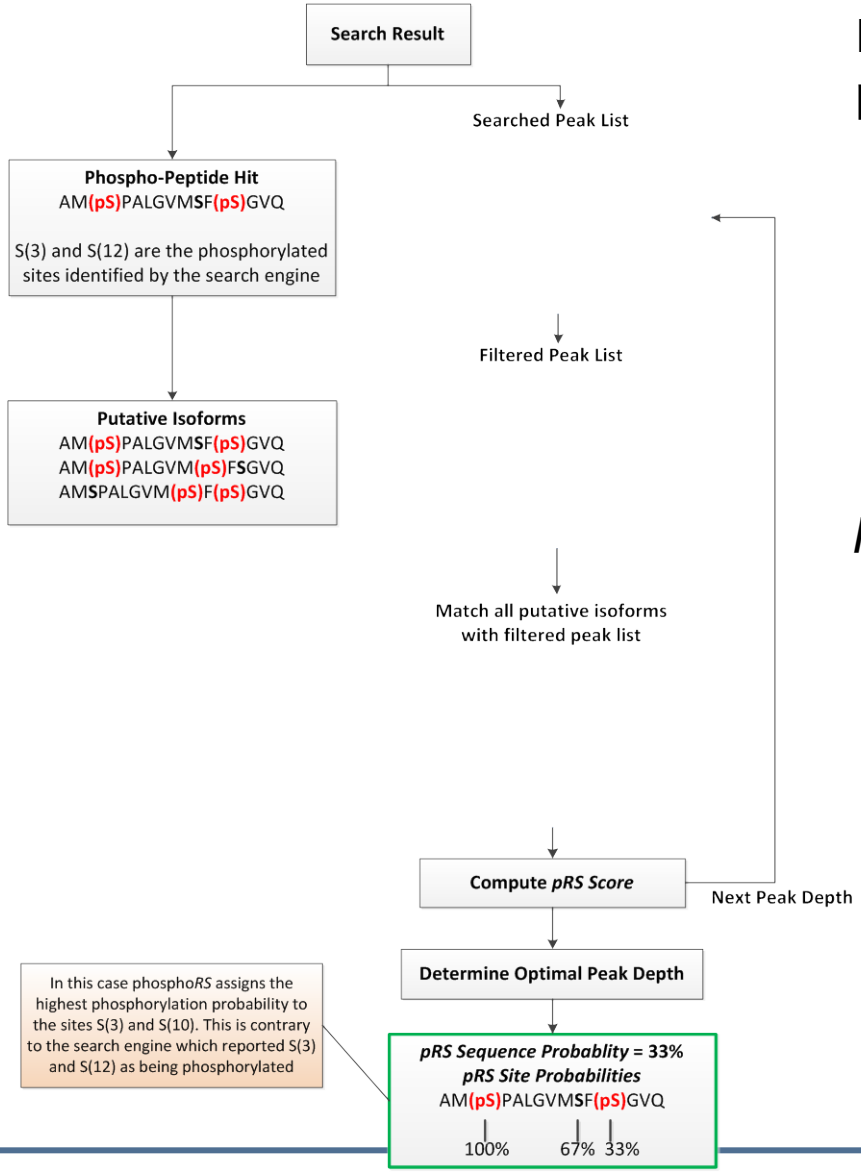


phospho*RS*

- automated calculation of individual site probabilities for each putative phospho-site
- works for all common fragmentation techniques (CID, ETD, and HCD) and all available database search engines
- validated and optimized by analysis of LC-MS/MS data of more than 150 synthetic phospho-peptides with precisely known phospho-sites

T. Taus et al., phosphoRS, a novel phospho-site localization tool, 59th ASMS 2011, WOA pm 3:30

phospho*RS*



Probability $P_{S,i}$ to match k_S or more peaks purely by chance:

$$P_{S,i}(X \geq k_S) = \sum_{k=k_S}^{n} \binom{n}{k} p^k (1-p)^{(n-k)}$$

n: number of theoretical fragment ions
k_S : number of matched fragment peaks
p: probability to match a fragment by chance

pRS-Scores:

$$pRS_{S,i} = -10 \cdot log_{10}\left(P_{S,i}(X \geq k_S)\right)$$

$$pRS\ Sequence\ Probability = \frac{P_{S,i_{optimal}}^{-1}}{\sum_S P_{S,i_{optimal}}^{-1}}$$

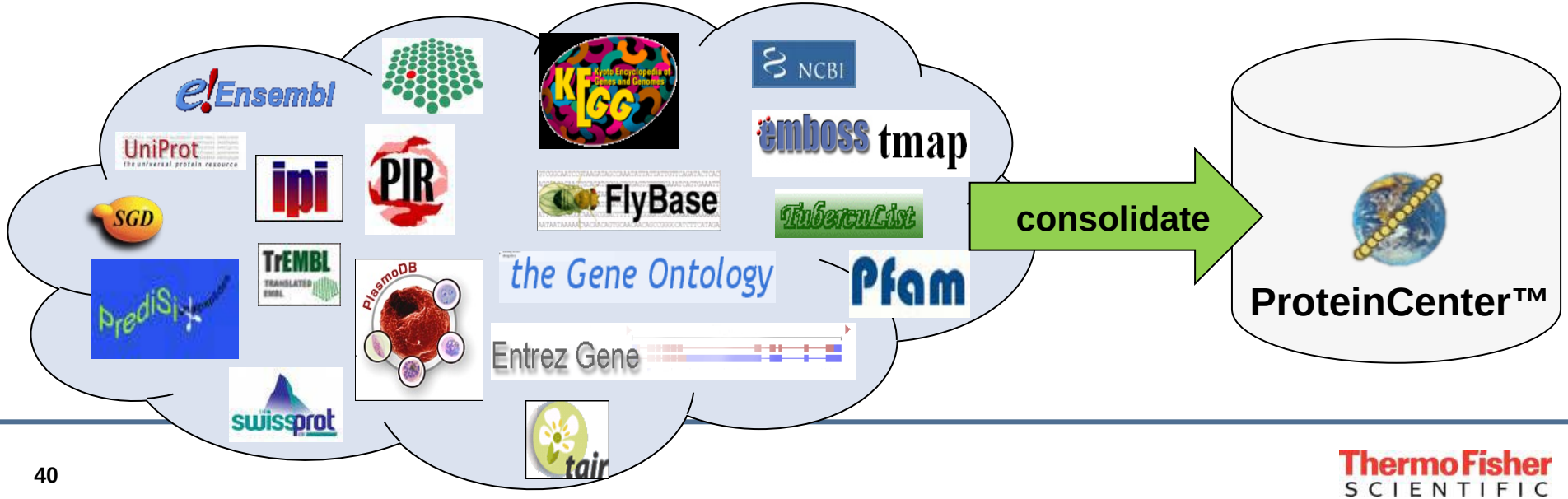
$$pRS\ Site\ Probability = \sum_{\substack{all\ isoforms \\ where\ this\ site \\ is\ phsophorylated}} pRS\ Seq.Prob._i$$

phospho*RS*

- pRS Score
  - This peptide score is based on the cumulative binomial probability that the observed match is a random event. The value of the pRS Score strongly depends on the data scored, but usually scores above 50 give good evidence for a good PSM.
- pRS Sequence Probability
  - This value estimates the probability (0-100%) that the respective isoform is correct.
- pRS Site Probabilites
  - For each phosphorylation site this is an estimation of the probability (0-100%) for the respective site being truly phosphorylated. pRS Site Probabilities above 75% are good evidence that the respective site is truly phosphorylated.

ProteinCenter™

- Protein-centric data warehouse specifically designed for interpretation of proteomics data
- Enables the comparison of data sets searched against different databases and different versions of databases
- > 16 million protein sequences from the major public protein databases distilled from 130 million accession codes from past and present versions
- The consolidated database is updated bi-weekly



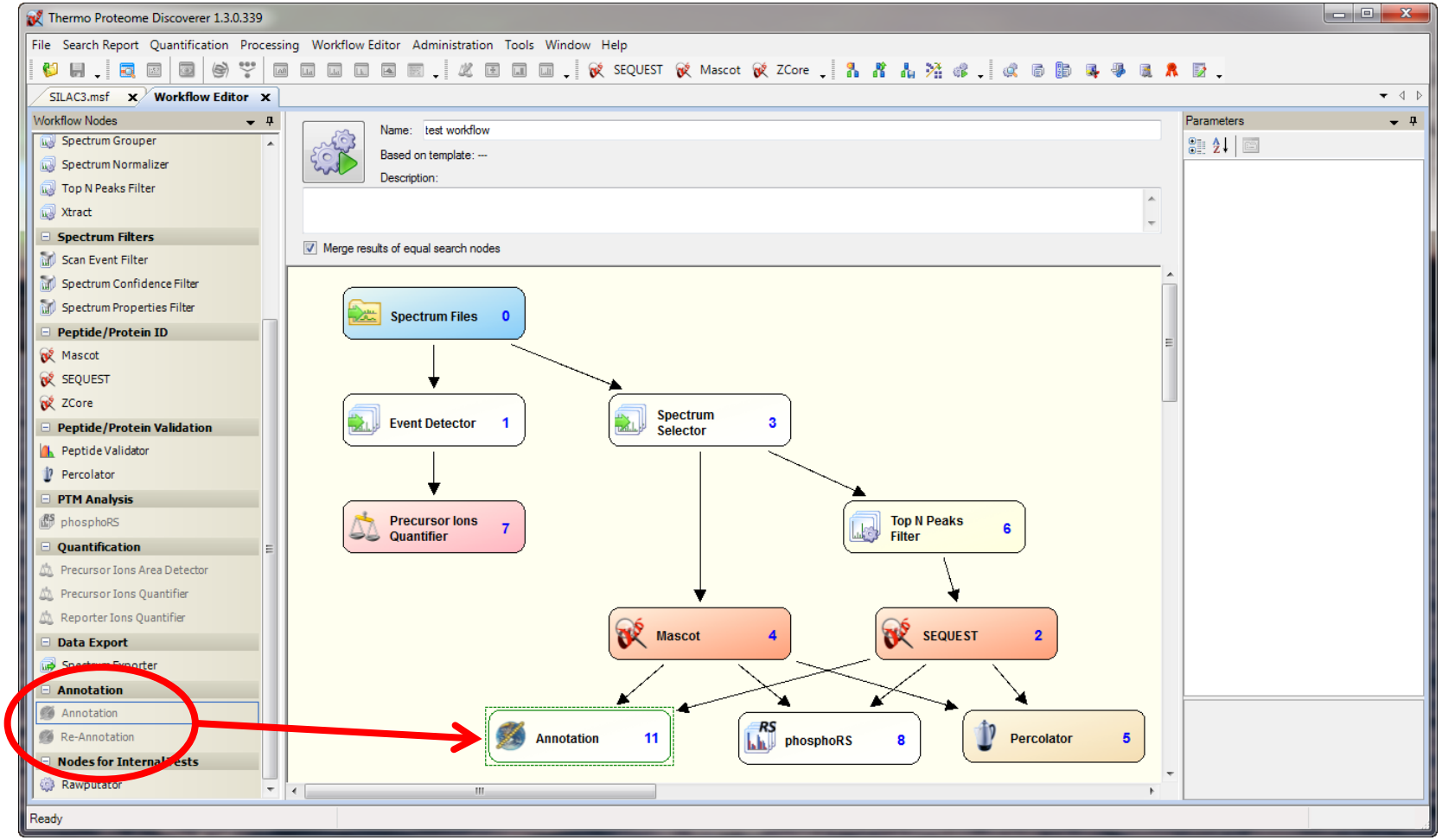
ThermoFisher SCIENTIFIC

Annotate Identified Proteins Using the ProteinCenter™ Repository

- Provides concise, precise and focused annotation for a given protein
- Use the Annotation node to directly retrieve protein annotations from the web service provided by ProteinCenter
- Currently the following annotations are retrieved:
  - Gene Ontology (GO, http://www.geneontology.org) and GO slim annotations
    - provides a widely used controlled vocabulary to describe the function, localization and process a protein is assigned to
    - approx. 35000 terms is organized in dependency hierarchies.
    - Subsets taken on a high level of this hierarchy (termed GO slims) are widely used to give an overview of the biological impact of a molecule
  - Protein family (Pfam, Welcome Trust Sanger Institute) annotations
  - Modifications documented in the Uniprot database

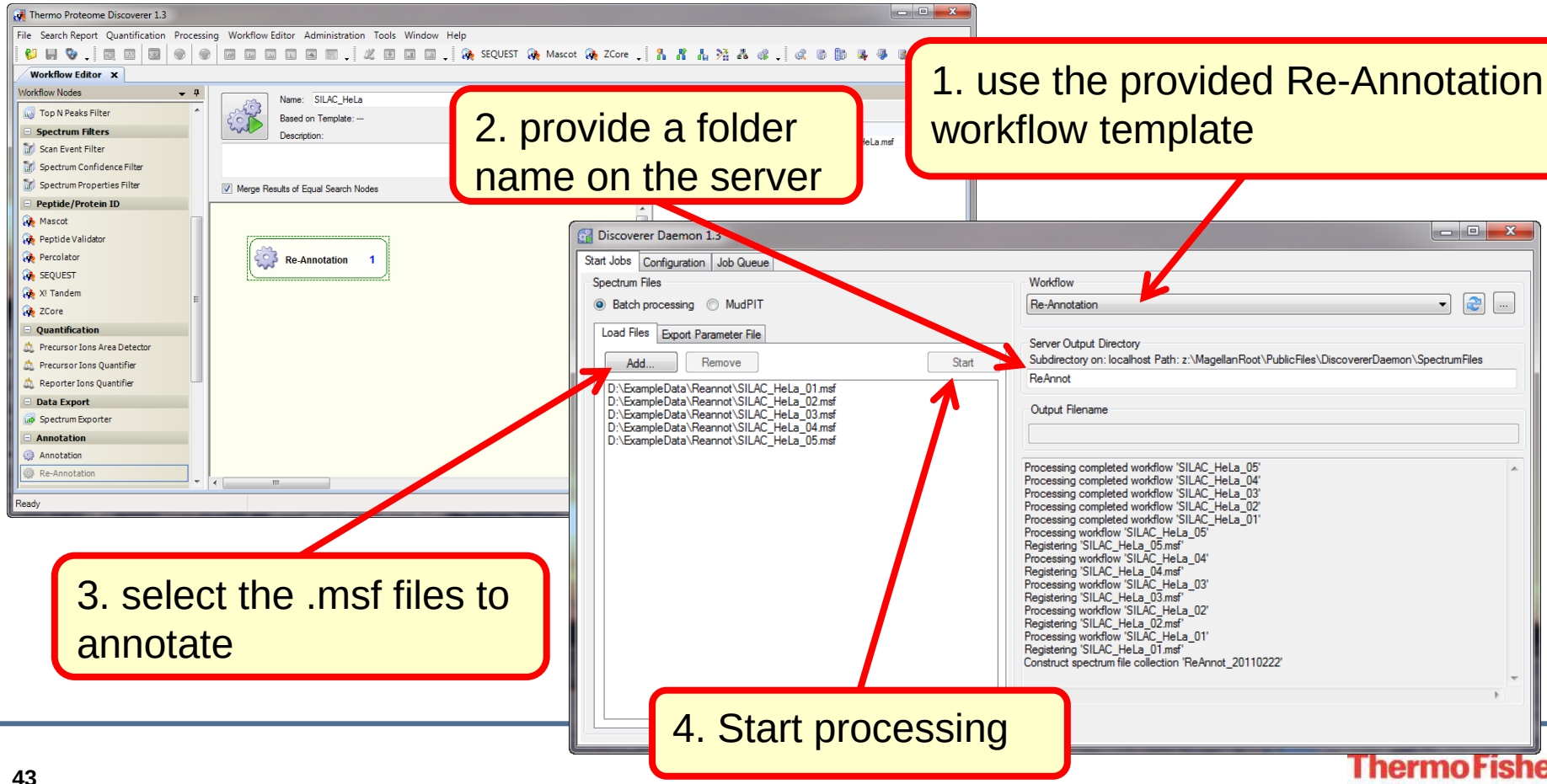
Use the Annotation Node in a Workflow

- The Annotation node automatically connects with every search node



(Re) Annotate Existing .msf Files

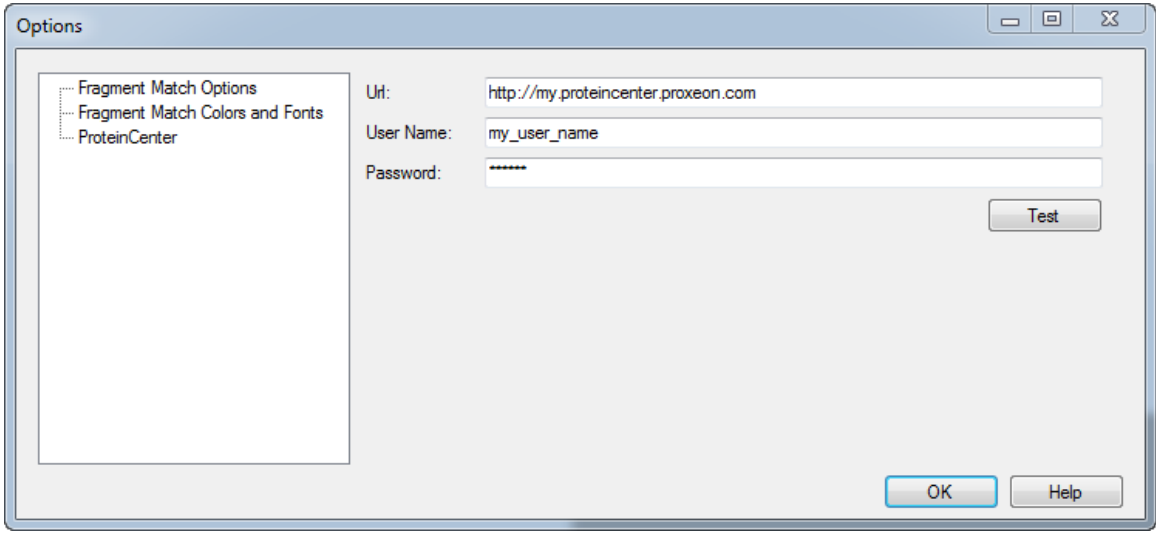
- Use the Re-Annotation node to annotate existing result files that do not yet contain annotations or update existing annotations
- Can be automated with the Discoverer Daemon



ThermoFisher SCIENTIFIC

Automatic Transfer of Results from Proteome Discoverer to ProteinCenter™

- You need a user account on a ProteinCenter™ server
- Configure your account settings under Tools > Options > ProteinCenter



- Export all or selected proteins with Tools > Export to ProteinCenter