Spectrum Mill - Overview

Karl Clauser

Broad Institute of MIT and Harvard

August 2021

SM Core Capabilities

Spectrum Mill provides modules for complete analysis of proteomic LC-MS/MS data, including:

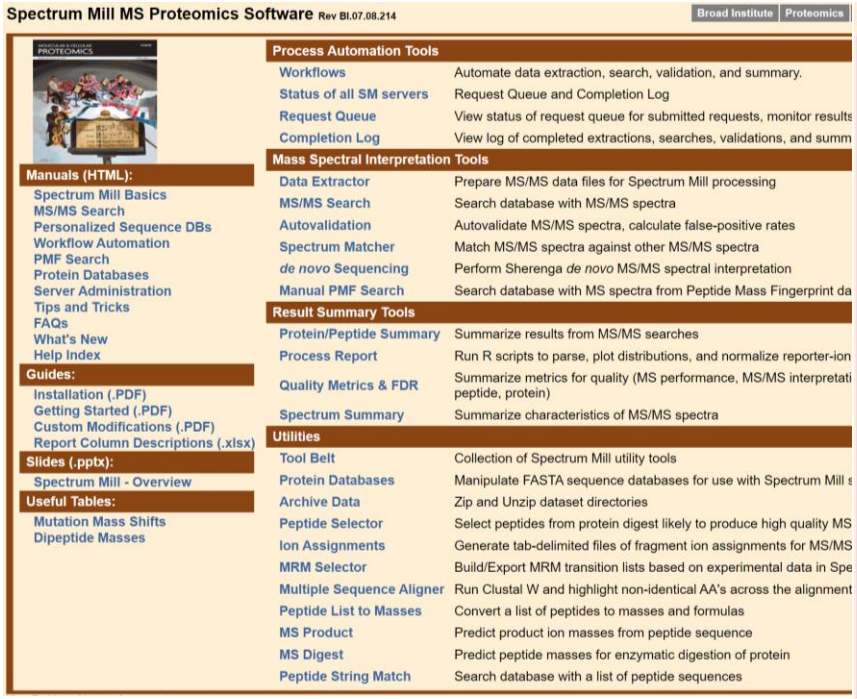
- Spectral pre-processing
  - directly from instrument vendor format files: Thermo Fisher, Agilent
- Identification of peptide spectrum matches - PSMs by database search.
  - PSM scoring capitalizes on low ppm product ion mass accuracies with fragment-ion type models optimized for orbitrap HCD, qToF-CID, ion trap CID, and ETD / ETHCD dissociation methods
- PTM-site localization
  - for phosphorylation, acetylation, and ubiquitylation with multi-sample site-level consolidation that handles localization ambiguity
- TMT-10, 11, 16, 18 / iTRAQ 4 reporter ion based quantitation
  - with PSM-level isotope correction and precursor ion purity filtering, and sample-level normalization
- Protein-level assembly of PSMs
  - with attention to isoform-specific and species-specific peptides
- Quality metric calculation/reporting
  - for MS acquisition, chromatography, spectral interpretation and FDR, digestion efficiency, PTM enrichment, and label incorporation
- Automation on an SM server through a service request manager
  - maximizes use of all CPUs, queues and processes tasks for multiple users via workflows with stored parameters

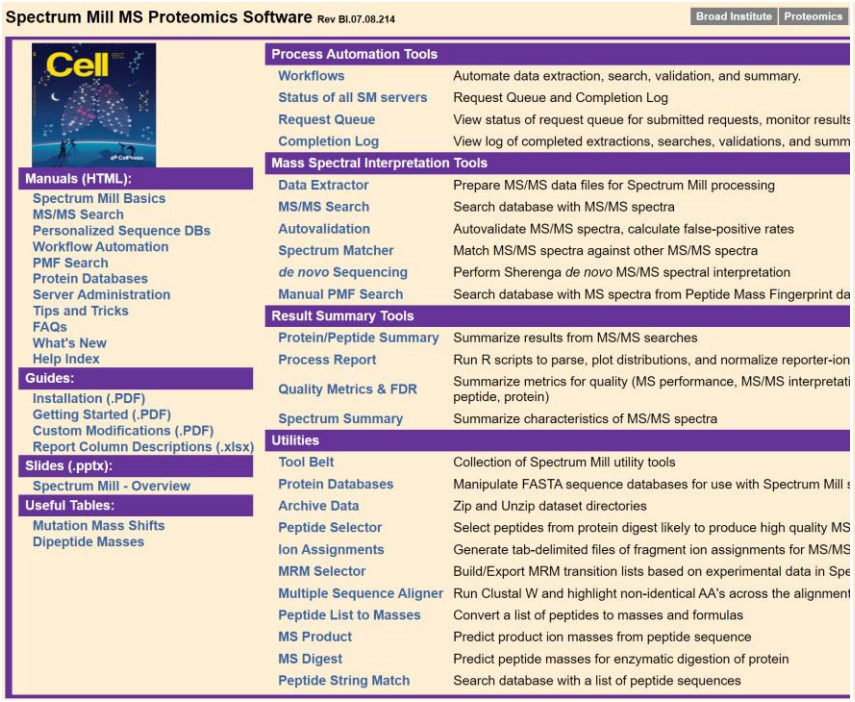
Topics Covered

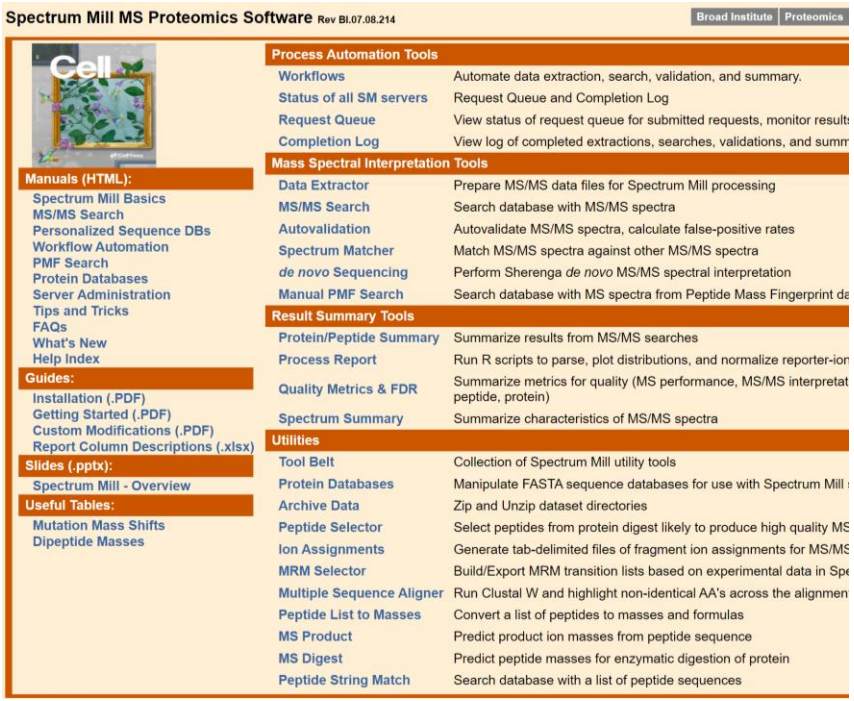
- Spectrum Mill architecture – program, files, workflow
- Configuring Extraction, MS/MS search, modifications
- PSM Scoring
- Phosphosite Localization
- Methods to control false discovery rates (FDR) of MS/MS search results
- Collapsing Peptide Spectrum Matches (PSM's) for Quantitation
  - Protein, Phosphosite levels
- Autovalidation

  False Discovery Rates (FDR) at the PSM, peptide, and protein levels.
- Quantitation features
  - Reporter Ion Correction Factors
- Automation – SRM, workflows
- Quality Metrics
- Process Report
- Future Directions

SM servers at Broad Institute







Cibola

24 CPUs 2.0 GHz
56 GB RAM

Manzano

24 CPUs 2.0 GHz
56 GB RAM

Eldorado

36 CPUs 2.0 GHz
24 GB RAM

Jurkat QC

HLA - Immunopeptidomics

Status of all SM servers at Broad Institute

Spectrum Mill - Request Queue cibola

Request Queue | Completion Log | Help

There are 4 requests in the queue. Available memory: 34.0 Gb of 51.5 Gb

Remove

	#	Task Id	Status	Task Type	Data Directory
☐	1	210723085635.25437	Queued	Autovalidation	CPTAC3/LUAD/Proteome/25CPTAC_LUAD_Proteome_
☐	2	210723085636.25438	Queued	Autovalidation	CPTAC3/LUAD/Proteome/25CPTAC_LUAD_Proteome_
☐	3	210723085638.25439	Queued	Autovalidation	CPTAC3/LUAD/Proteome/25CPTAC_LUAD_Proteome_
☐	4	210723085634.25409P Monitor	Running (535:1924)	MS/MS Search	CPTAC3/LUAD/Proteome/25CPTAC_LUAD_Proteome_

Spectrum Mill - Request Queue manzano

Request Queue | Completion Log | Help

There are 7 requests in the queue. Available memory: 26.5 Gb of 51.5 Gb

Remove

	#	Task Id	Status	Task Type	Data Directory
☐	1	210722213857.18237	Queued	Autovalidation	CPTAC3/PDAC/Phosphoproteome/10CPTAC_PDA_
☐	2	210722213859.18238	Queued	Autovalidation	CPTAC3/PDAC/Phosphoproteome/10CPTAC_PDA_
☐	3	210722213856.18235P Monitor	Running (986:998)	MS/MS Search	CPTAC3/PDAC/Phosphoproteome/10CPTAC_PDA_
☐	4	210723084219.19241	Queued	Autovalidation	CPTAC3/LUAD/Proteome/01CPTAC_LUAD_Proteon
☐	5	210723084220.19242	Queued	Autovalidation	CPTAC3/LUAD/Proteome/01CPTAC_LUAD_Proteon
☐	6	210723084221.19243	Queued	Autovalidation	CPTAC3/LUAD/Proteome/01CPTAC_LUAD_Proteon

Spectrum Mill - Request Queue eldorado

Request Queue | Completion Log | Help

There are 3 requests in the queue. Available memory: 18.4 Gb of 25.8 Gb

Remove

	#	Task Id	Status	Task Type	Data Directory
☐	1	210723091357.9	Queued	Autovalidation	HLA/External_Data/PXD01349/PXD019643_humanatlas/DN
☐	2	210723091358.10	Queued	P/P Summary	HLA/External_Data/PXD01349/PXD019643_humanatlas/DN
☐	3	210723091355.3P Monitor	Running (100:1493)	MS/MS Search	HLA/External_Data/PXD01349/PXD019643_humanatlas/DN

Spectrum Mill - Request Queue shiva

Request Queue | Completion Log | Help

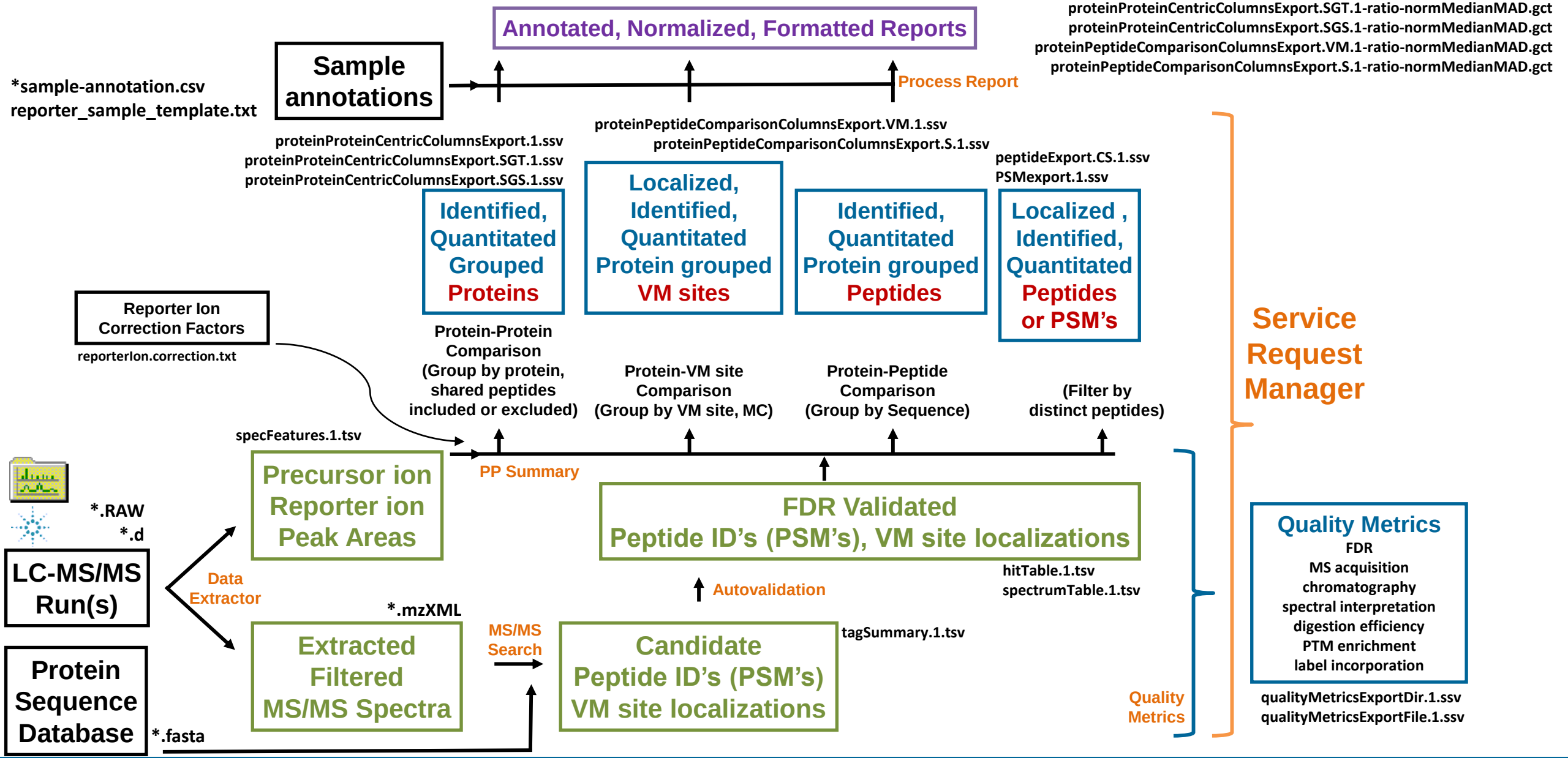
There are 3 requests in the queue. Available memory: 66.8 Gb of 90.2 Gb

Remove

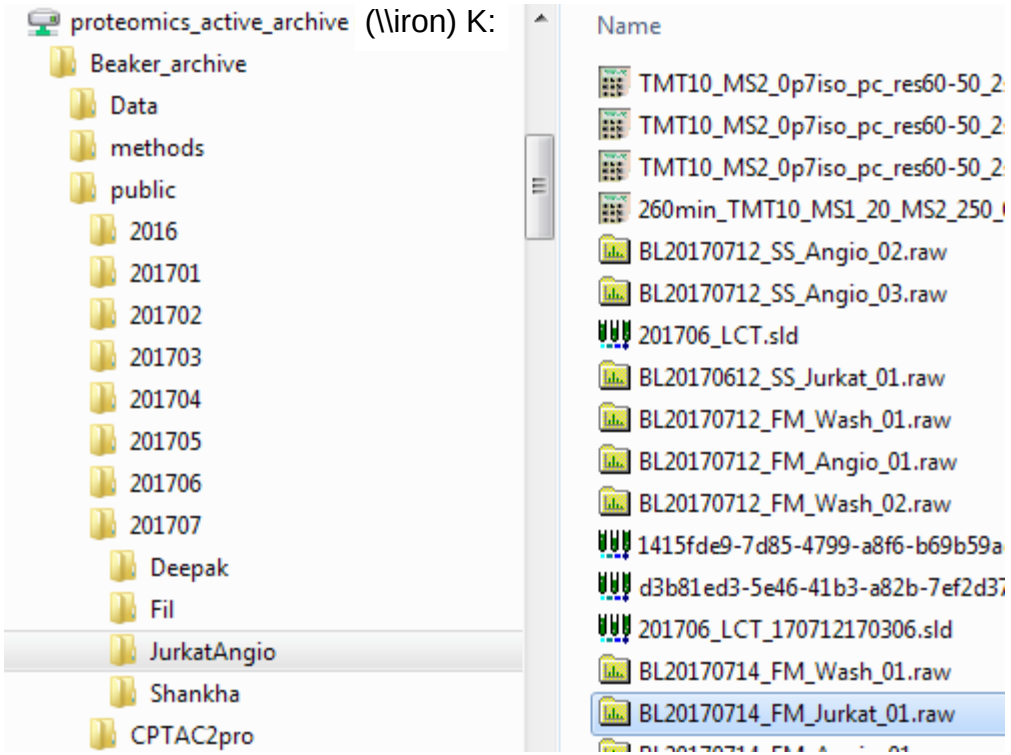
	#	Task Id	Status	Task Type	Data Directory
☐	1	210723082331.9793	Queued	Autovalidation	CPTAC3pancancer/PDAC/Phosphoproteome/16CPTAC_F
☐	2	210723082332.9794	Queued	Autovalidation	CPTAC3pancancer/PDAC/Phosphoproteome/16CPTAC_F
☐	3	210723082330.9791P Monitor	Running (384:998)	MS/MS Search	CPTAC3pancancer/PDAC/Phosphoproteome/16CPTAC_F

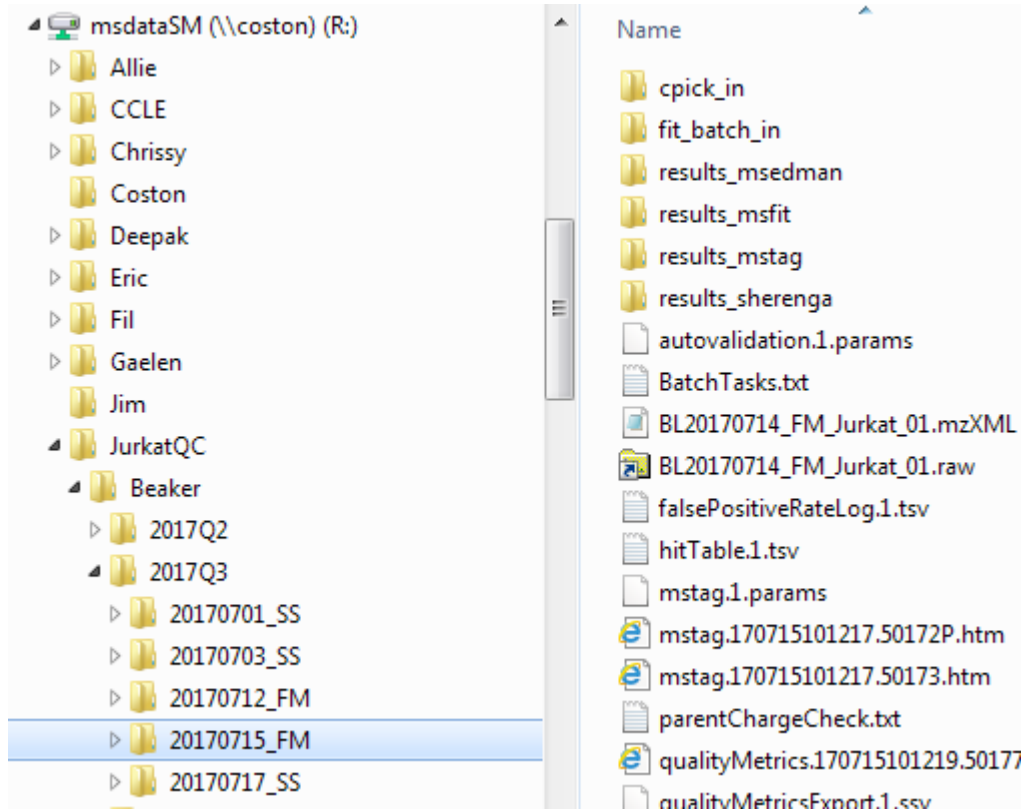
108 CPUs Grinding out PSMs

Spectrum Mill Workflow



Use Windows Explorer to Move data files to the SM server





<u>At Broad Institute</u>

- Nightly, data files backed up from instrument to archive server.
- Following extraction, raw data files automatically deleted from SM server, replaced by shortcut.

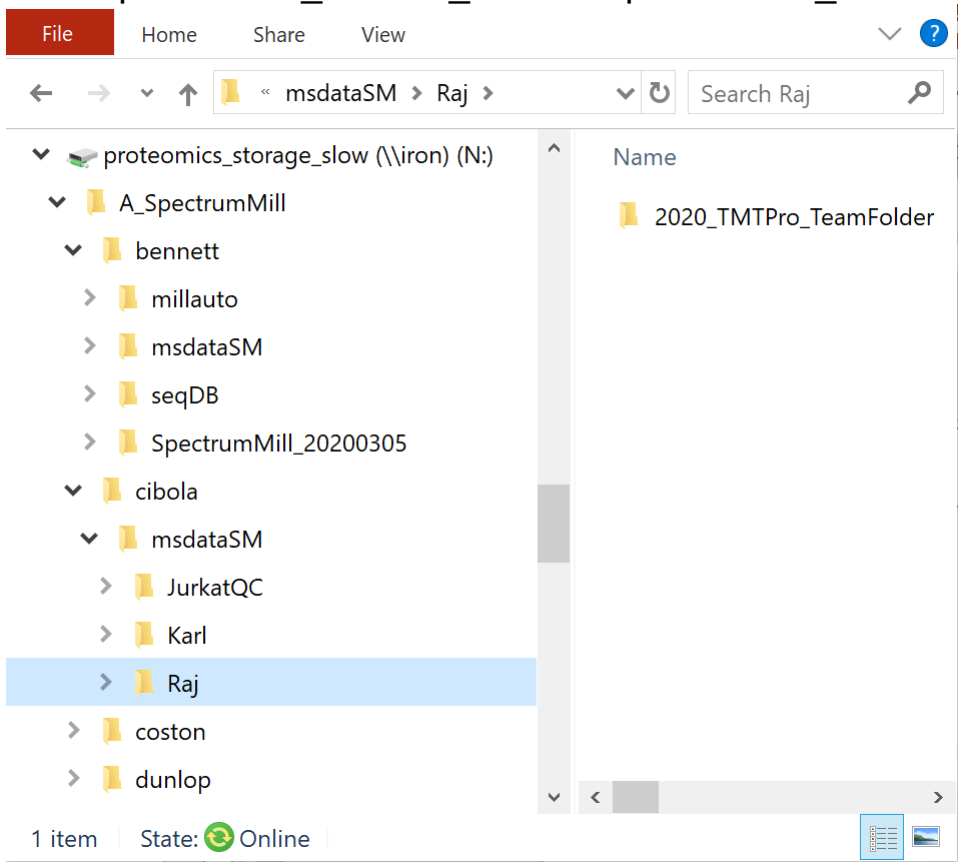
Maintaining Available SM Server Disk Space for Active Data Analysis

Idle Datasets

Might Ask,
Tell via Slack

Migrate to:
\\iron/proteomics_storage_slow/A_SpectrumMill

Older (read-only access)
\\iron/proteomics_active_archive/SpectrumMill_archive
\\iron/proteomics_inactive_archive/SpectrumMill_archive

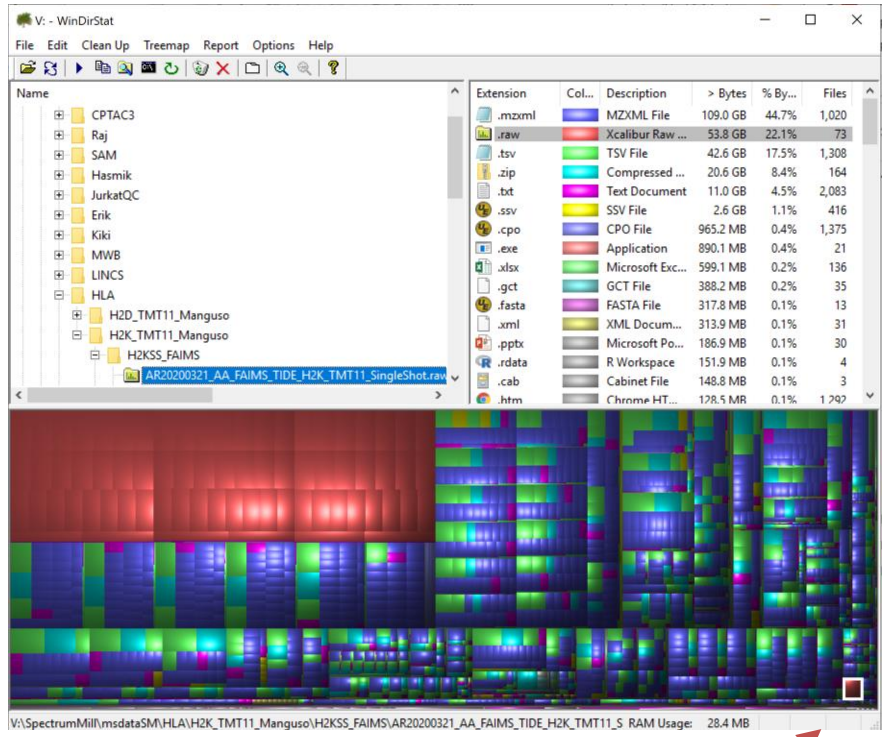


Stale .RAW files

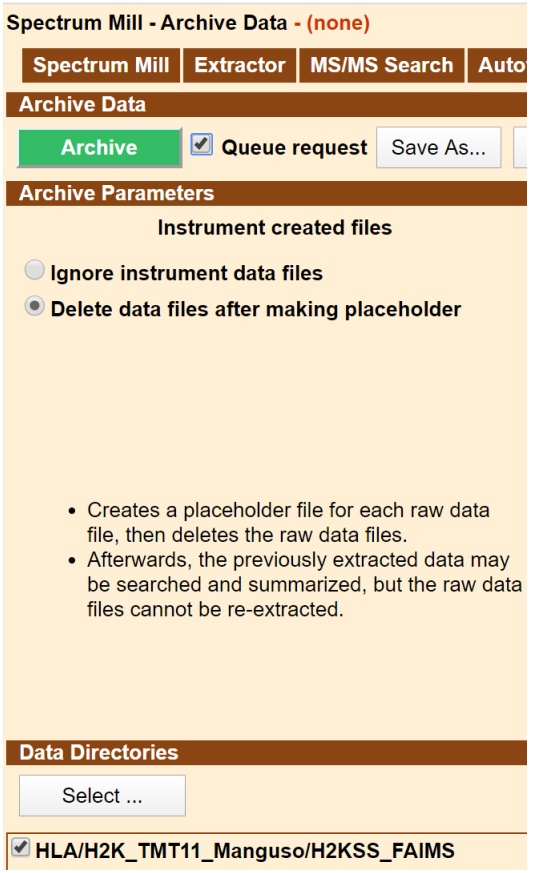
Don't ask,
Don't tell,
Just do it

Delete &leave behind a placeholder
User can tell what was once there

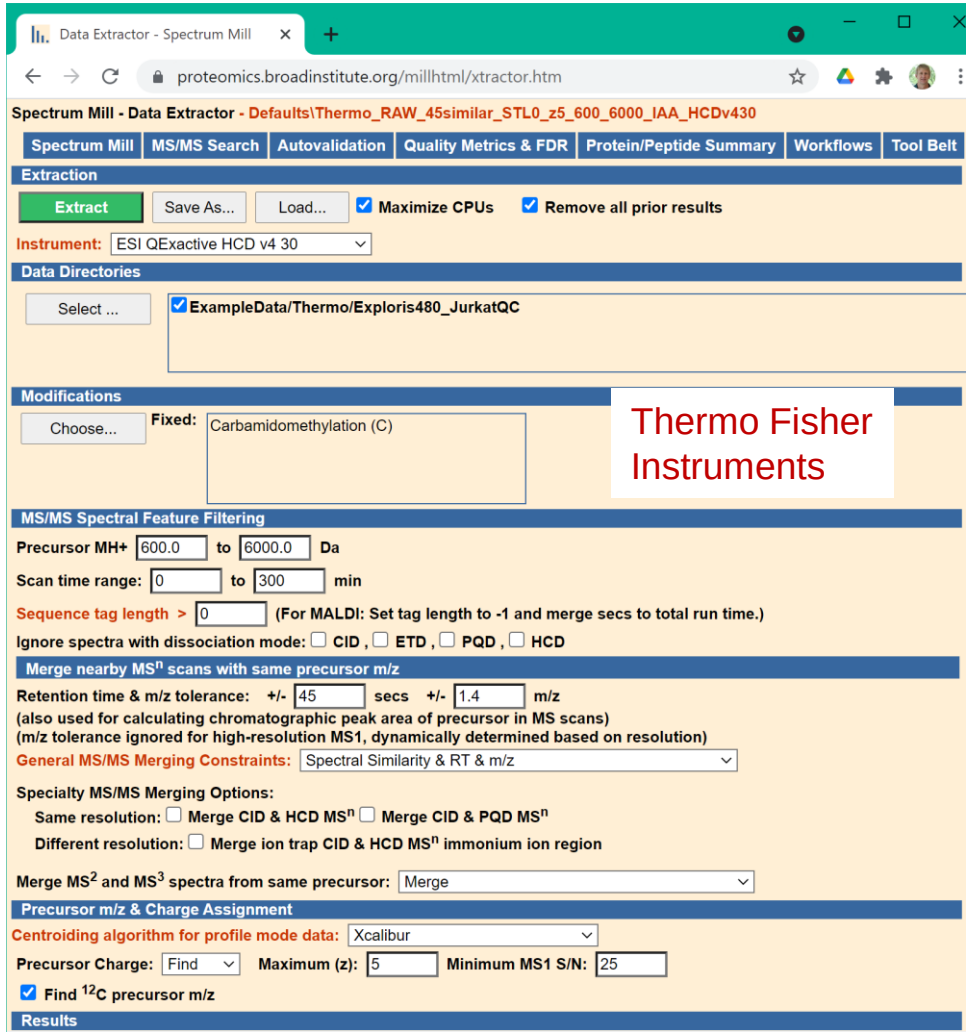
fresh

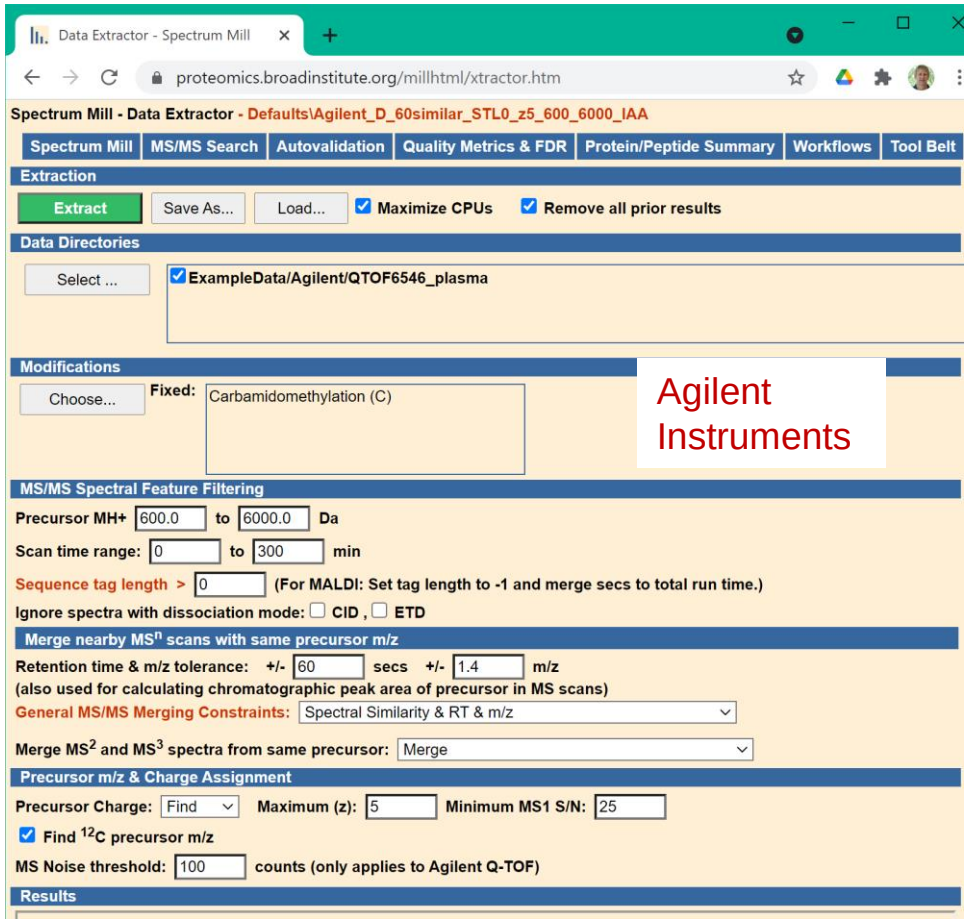


stale



Configuring an Extraction



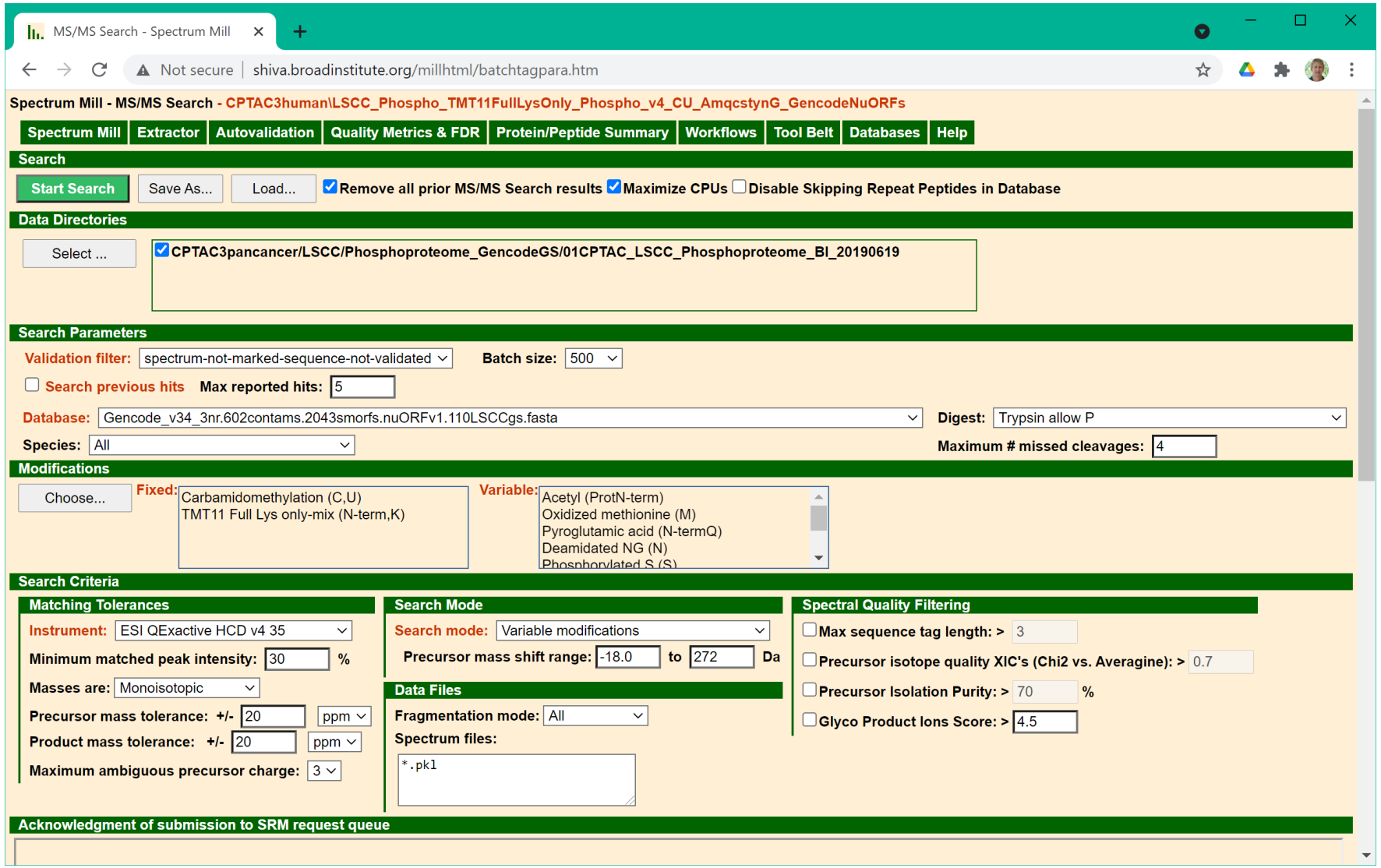


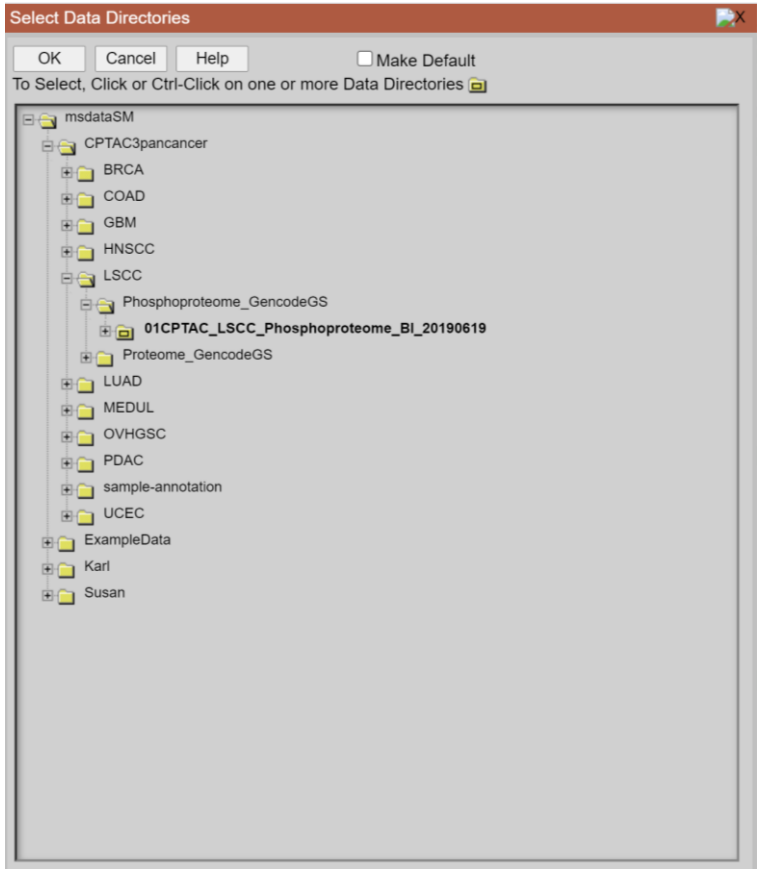
- Convert to .mzXML
- Filter out low quality MS/MS
- Extract reporter ion intensities

- Calculate precursor ion XIC's
- Correct precursor monisotopic m/z assignments

- Merge replicate MS/MS from same chromatographic peak

Configuring an MS/MS Search





Fixed, Mix and Variable Modifications

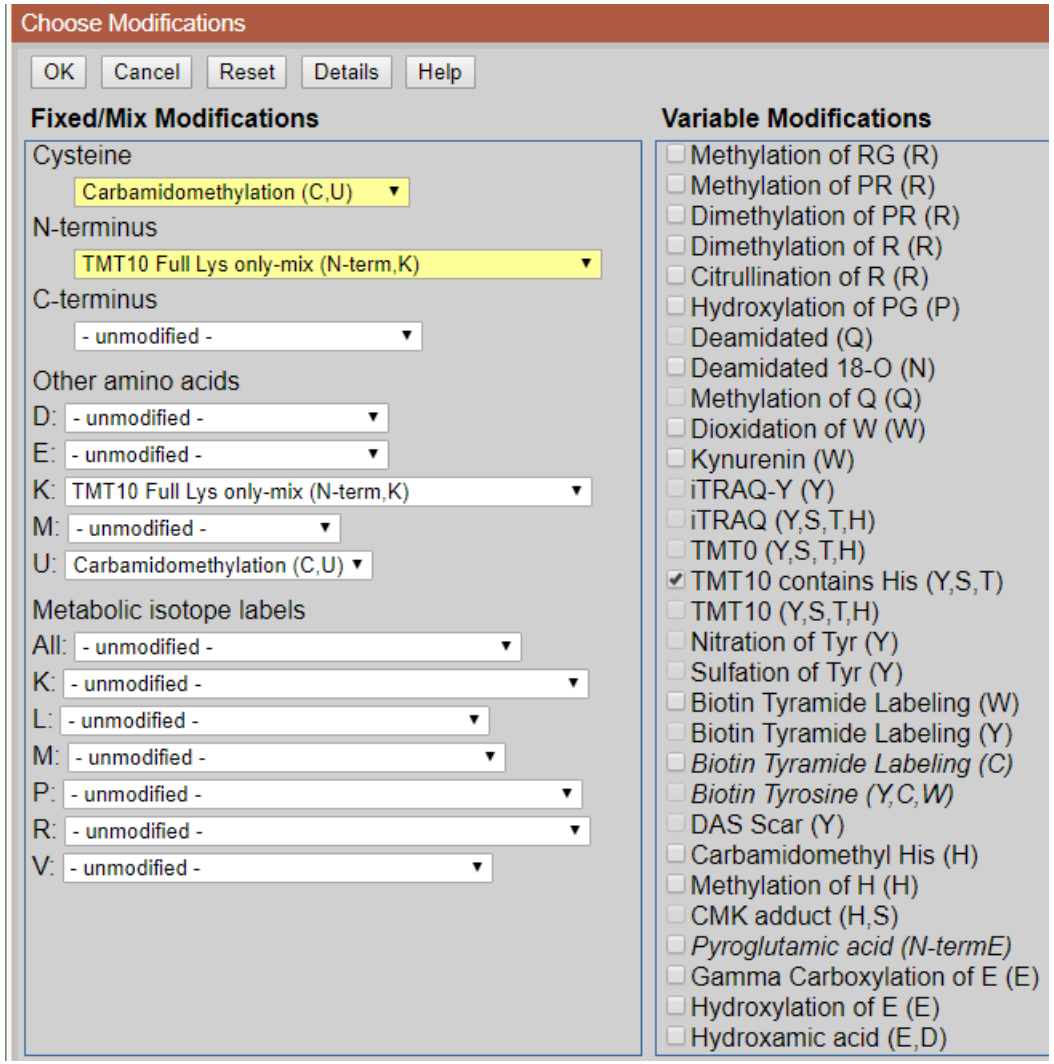
Fixed

Redefine the wild type as

Mix

TMT 10 full Lys only mix
Search in 2 cycles
Cycle 1: all K , and Nterm TMT10 labeled
Cycle 2: all K TMT10 labeled, Nterm unlabeled

TMT 10 partial mix
Search in 4 cycles
Cycle 1: unlabeled
Cycle 2: all K , and Nterm TMT10 labeled
Cycle 3: all K , and Nterm unlabeled
Cycle 4: all K unlabeled, Nterm TMT10 labeled



Variable

Allow
2 possibilities
for an AA.
Allow both
in 1 spectrum
if more than one
location/AA.

<u>Overlabeled</u>
(TMT on S,T,Y
only when
Peptide
Contains His)
smaller search space

Database Search: General Considerations

- Choice of database
  - reference vs. personalized + contams + non Canonical ORFs + tagged bait protein
    - reference proteome (UniProt, RefSeq, Ensembl, Gencode)
  - decoy databases – estimation of false discovery rate

- Choice of enzyme
  - full vs. partial enzyme specificity, no enzyme for immunopeptidomics

- Choice of fixed and variable modifications
  - fixed modifications: target AA always considered as modified
  - variable modifications: target AA may or may not be modified
  - *expansion of search space!*

- Precursor ion (peptide) mass tolerance
  - depends on the measurement mass accuracy
  - lower tolerance decreases search space (fewer candidate masses)
  - higher tolerance helps FDR statistics, requires post-search tolerance filtering (3 std dev of mean)

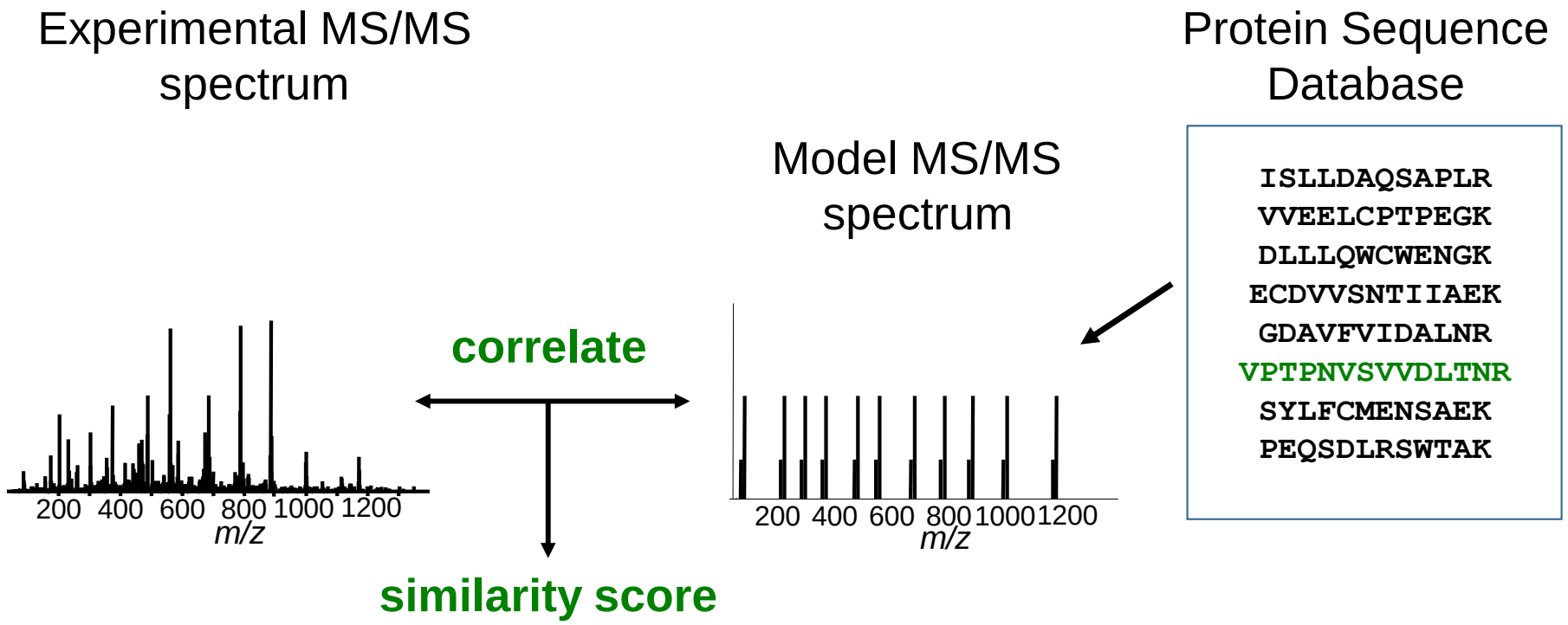
- MS/MS fragment ion mass tolerence
  - may not be the same as precursor mass tolerance (hybrid instruments)

Key Features of Spectrum Mill Database Searching with MS/MS Spectra

- Scoring specific to instrument types, dissociation modes
  - Ion types allowed and scoring weights
  - On board peak detection (de-isotoping, S/N thresholding, fragment z assignment).
- Product ion mass tolerance in ppm or Da.
- Spectra grouped in batches of similar precursor m/z, z for fast precursor mass filtering.
- Target-decoy FDR accomplished by on-the-fly peptide inner reversal
  - SAMPLER becomes SELPMAR
  - Search time is only ~1.5x as long as a target only database, instead of 2x for a concatenated forward/reverse database. 1x digestion of all proteins, 1x precursor mass filtering of all peptides, 2x MS/MS matching of every sequence (fwd & rev)
  - Search results do not comingle target and decoy hits.
    - For each PSM report top target hit, and delta Fwd – Rev score of top decoy hit
    - A false positive PSM had delta Fwd-Rev < 0. (Top Decoy hit has higher score)
- Multiple cycles for fixed/mix modifications or unknown charge, with combined single result.
- Variable modifications constrained by precursor mass shift, not by # of mods/peptide.
- Second pass search possible with leftover spectra from 1st pass.
- Homology modes available
  - Unassigned single mass gap
  - Allow single mutation per peptide

PSM Scoring

MS/MS Search Engines: look up answer in back of book



- Find best matching database peptide
- Can miss unanticipated modifications
  - post-translational, sample handling
- Determine peptide FDR by searching reversed DB

Algorithms: Mascot, Sequest, MaxQuant, Spectrum Mill, MSGF+, MSFragger, PEAKS, X-Tandem…

PSM scoring considerations

- Signal pre-processing
  - deisotoping, fragment charge assignment, noise removal

- Fragment ion types allowed and relative scoring weights
  - b/y, neutral loss (-H_2O, -NH_3, -H_3PO_4), internal, immonium
  - specific to instrument dissociation method (HCD, CID, ETD)

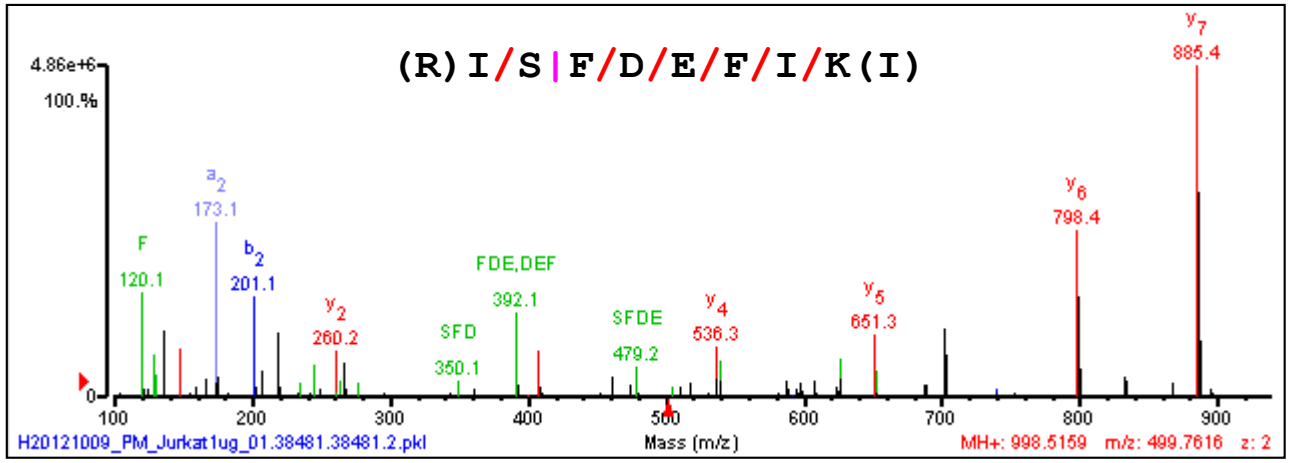
- Peak intensity
  - Contribute to score or just peak detection
  - Intensity used in model spectra (fragment ion type, mass range)

- Fragment mass tolerance units appropriate to mass analyzer
  - Da (ion trap, quadrupole), ppm (Orbitrap, ToF)

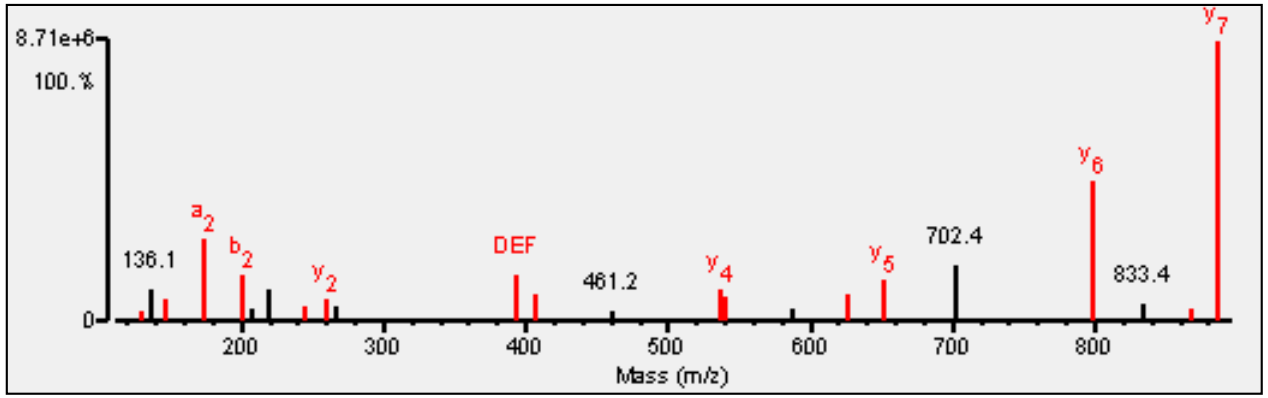
- Use of secondary scores
  - Gap between rank1 rank2 peptide
  - Sequence coverage of peptide
  - Unexplained peaks in spectrum
  - Ion series continuity

- Size of database may effect score

Spectrum Mill Scoring of MS/MS Interpretations



Peak Selection: De-Isotoping, S/N thresholding,
Parent - neutral removal, Charge assignment
Match to Database Candidate Sequences



Score
=
Assignment Bonus
(Ion Type Weighted)
+
Marker Ion Bonus
(Ion Type Weighted)
-
Non-assignment Penalty
(Intensity Weighted)

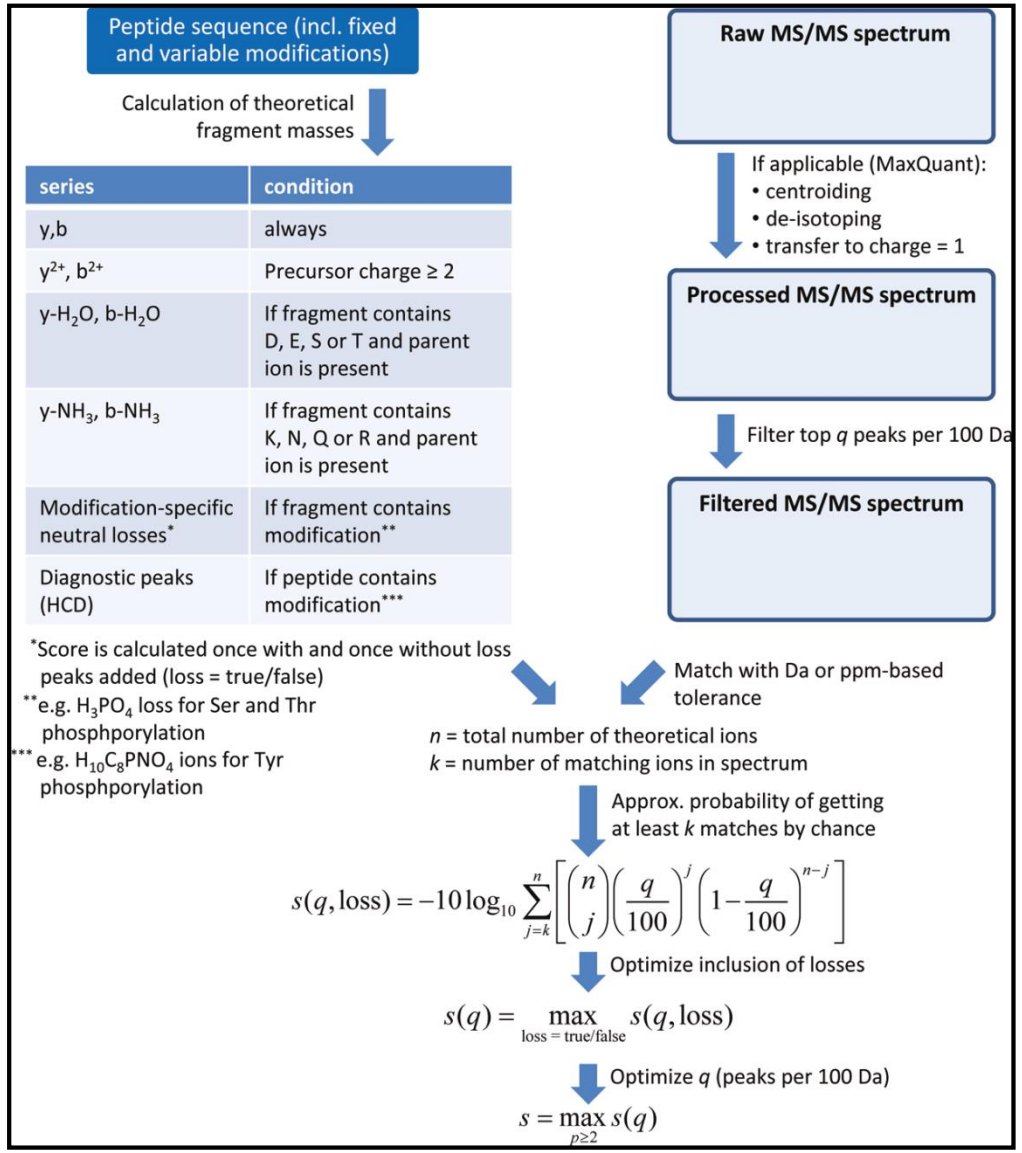
SPI (%)
Scored Peak Intensity

PIP(%)
Precursor Ion Purity

11.52
81.5%
80.3%

Fragment-ion (m/z)	120.080	129.102	130.086	136.075	147.112	173.128	201.122	207.112	219.112	245.075	260.195	267.115	392.143	407.263	461.163	536.304	539.210	587.366	626.242	651.332	702.390	798.400	833.430	867.416	885.430
Frac. Inten.(% of TIC)	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00
Rel. Inten.(% of BP)	0.53	0.22	3.83	11.45	8.14	29.49	16.93	4.54	11.12	5.41	8.07	5.92	16.63	9.57	3.61	11.65	9.01	4.52	10.00	15.16	20.21	50.16	6.43	4.73	100.00
Signal/Noise	17.8	6.4	2.7	11.1	7.4	30.8	17.0	3.5	10.7	4.4	7.4	5.0	16.7	9.0	2.5	45.2	33.6	14.0	38.0	60.6	82.6	213.7	22.4	14.9	431.7
Score	1.00	0.20	0.250	-0.115	1.000	0.250	1.000	-0.045	-0.111	0.500	1.000	-0.059	0.500	1.000	-0.036	1.000	0.500	-0.045	0.500	1.000	-0.202	1.000	-0.064	0.500	1.000
Ion-type	F	K	y_1-NH_3		y_1	a_2	b_2			DE	y_2		DEF	y_3		y_4	FDEF		SFDEF	y_5		y_8		y_7-H_2O	y_7
Delta ppm	-4.1	-5.0	-4.1		-6.2	-5.5	-5.0			-7.7	-7.9		-5.9	-5.6		-7.2	-6.7		-6.2	-4.0		-4.6		-9.6	-5.7

PSM Identification Scoring – Andromeda, MaxQuant



series	condition
y,b	always
y^{2+}, b^{2+}	Precursor charge ≥ 2
y-H_2O, b-H_2O	If fragment contains D, E, S or T and parent ion is present
y-NH_3, b-NH_3	If fragment contains K, N, Q or R and parent ion is present
Modification-specific neutral losses*	If fragment contains modification**
Diagnostic peaks (HCD)	If peptide contains modification***

*Score is calculated once with and once without loss peaks added (loss = true/false)
**e.g. H_3PO_4 loss for Ser and Thr phosphporylation
***e.g. $H_{10}C_8PNO_4$ ions for Tyr phosphporylation

$$s(q,\text{loss}) = -10\log_{10}\sum_{j=k}^{n}\left[\binom{n}{j}\left(\frac{q}{100}\right)^{j}\left(1-\frac{q}{100}\right)^{n-j}\right]$$

$$s(q) = \max_{\text{loss = true/false}} s(q,\text{loss})$$

$$s = \max_{p\geq 2} s(q)$$

Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen JV, Mann M, J. Proteome Res. 2011, 10, 1794–1805.

Peak selection assumptions

- Expects to use same # of peaks for each 100 Da region of spectrum
- Tall and short peak intensities equally diagnostic
- All regions of spectrum treated as equally likely
  - multiply charged fragments below precursor
  - some 100-300 m/z values not possible, di-pep AA combinations
- Tolerance in Da, not ppm

Scoring

- Scoring implies +/- 0.5 Da fragment tolerance
  - Matches separately constrained by user specified ppm or Da fragment tolerance
- Scoring implies all ion types have same value
- All ion types used for scoring

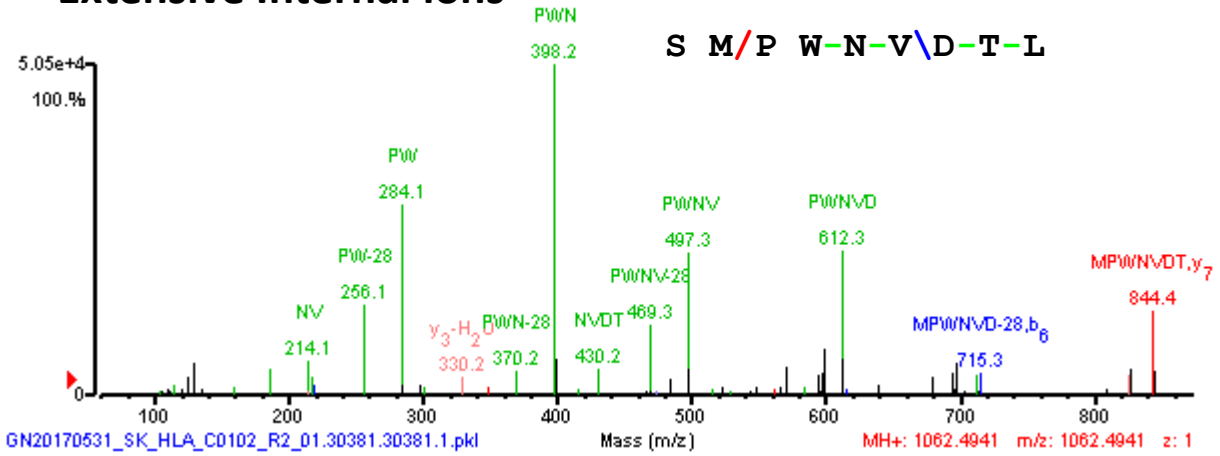
Final score optimizes

- # of highest intensity peaks used per 100 m/z
- Inclusion/exclusion of neutral loss ion types

Karl's take: Effective score, despite some awkward assumptions made to squeeze MS/MS interpretation into a binomial probability framework

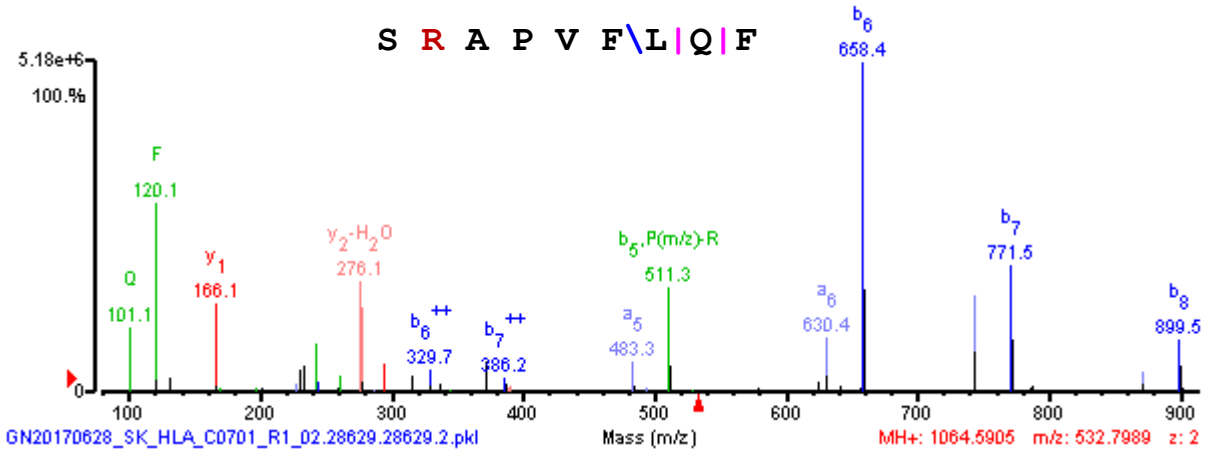
Example HCD MS/MS of range of scoring challenges

Extensive Internal ions



This peptide with no basic residues and a singly charged precursor ion fragments to yield mostly internal sequence ions.

Low Sequence Coverage



This peptide fragments to yield many different ion types, but they are derived from only 3 positions along the peptide backbone.

Key aspects of Spectrum Mill DB-search scoring revisions for HLA peptides

Added PSM quality threshold separate from primary score:

Minimum Backbone Cleavage Score (BCS) of 5

- Enforces uniformly higher minimum sequence coverage for each PSM, at least 4 or 5 residues of unambiguous sequence.
- Decreases false-positives associated with PSMs having fragmentation in a limited portion of the peptide that yield multiple ion types.
- BCS is a sequence coverage metric. The score is 1 or 0 for cleavage between adjacent AA's in the sequence (max score is peptide length-1). To receive a score, cleavage of the peptide backbone must be supported by presence of a primary ion type for HCD: b, y, or internal ion C-terminus (i.e. if the internal ion is for PWN then BCS is credited only for the backbone bond after the N).

BCS Example (top spectrum) **S M/P W-N-V\D-T-L**

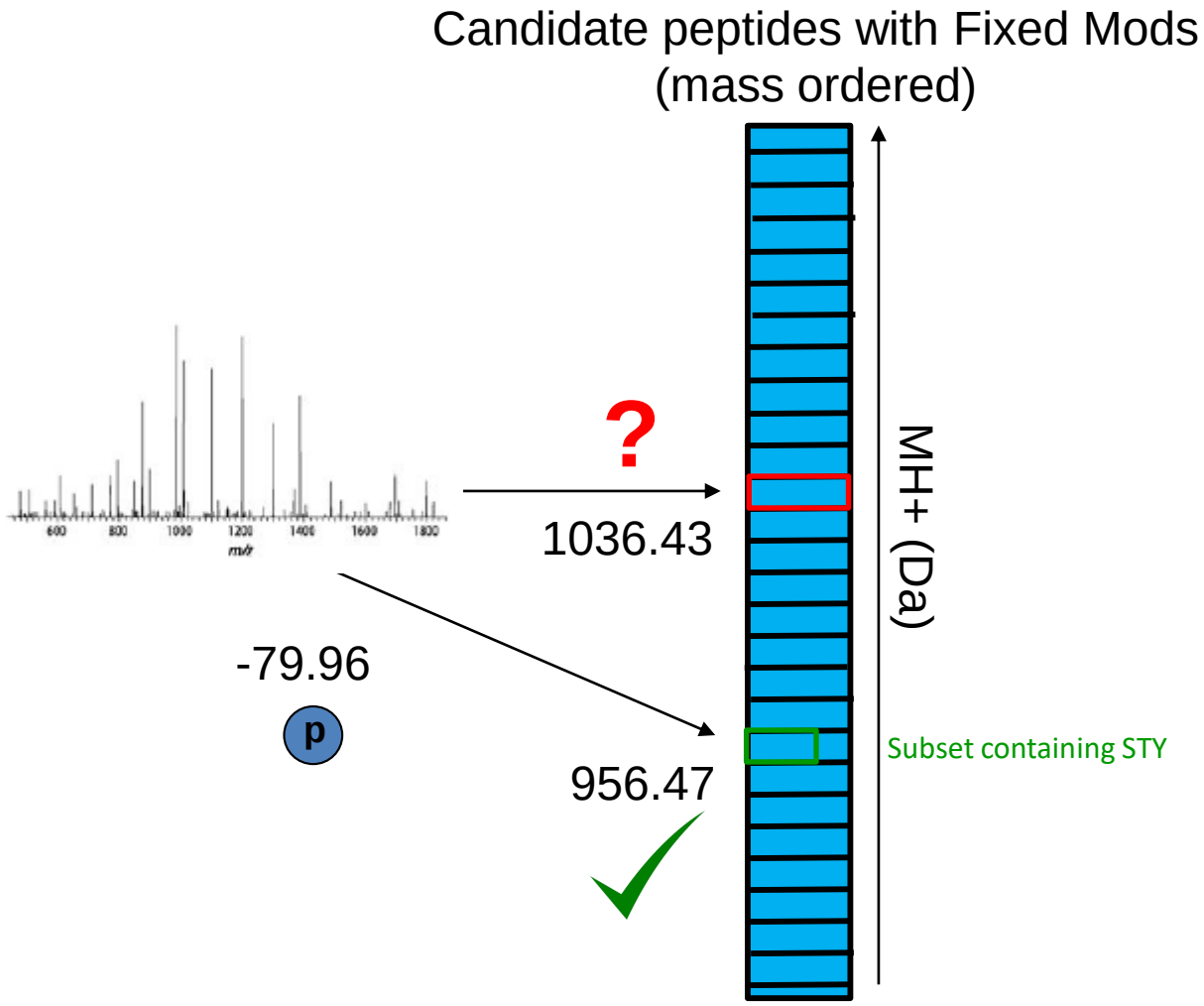
BCS: 6 **/ : y_7** **\ : b_6 ion** **-: internal ions (PW, PWN, PWNVD, PWNVDT)**

Enabled additional internal ion types: neutral loss of NH_3, H_2O, and CO.

- Minimal bonus score (10% of the score of b or y ions) for presence accompanying an internal b-ion (75% of the score of b or y-ions).
- Prevents intensity-based penalty score when the ion types are not enabled.

Search Space Expansion for Identification of Variably Modified Peptides

- Modifications alter the mass of the corresponding protein/peptide
- Have to be considered in database search
- Fixed modification: corresponding amino acid is always modified (e.g. carbamidomethylation of C)
- Variable modification: corresponding amino acid might or might not be modified (e.g. phosphorylation of S/T/Y)



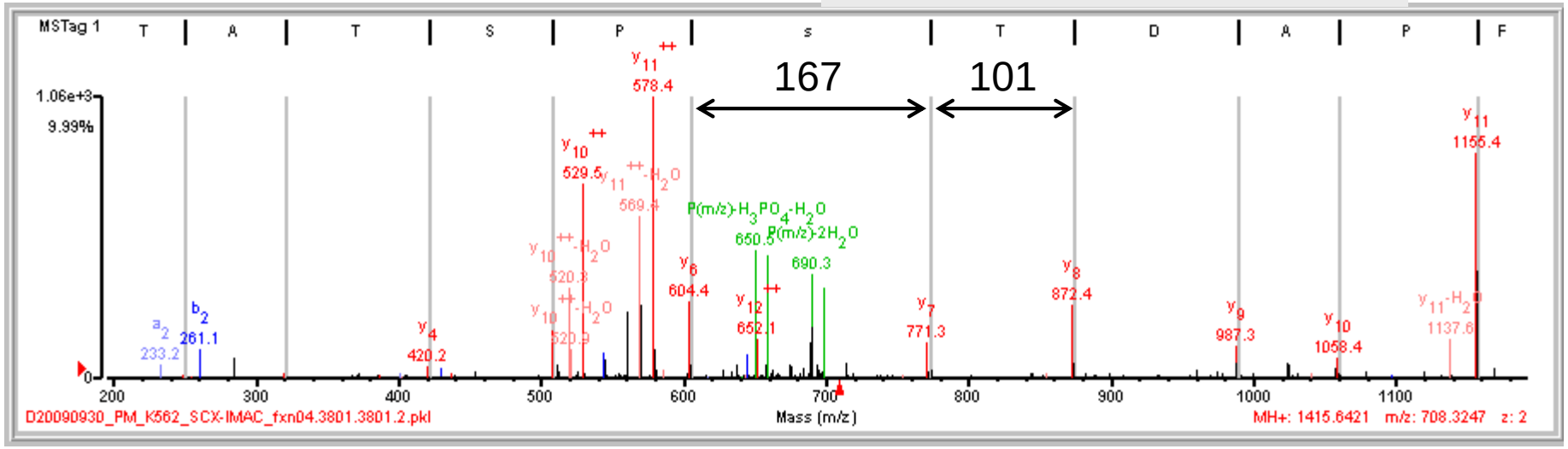
Variable mods require the search engine to evaluate candidate sequences with a precursor mass **offset** from the measured mass

Phosphosite Localization

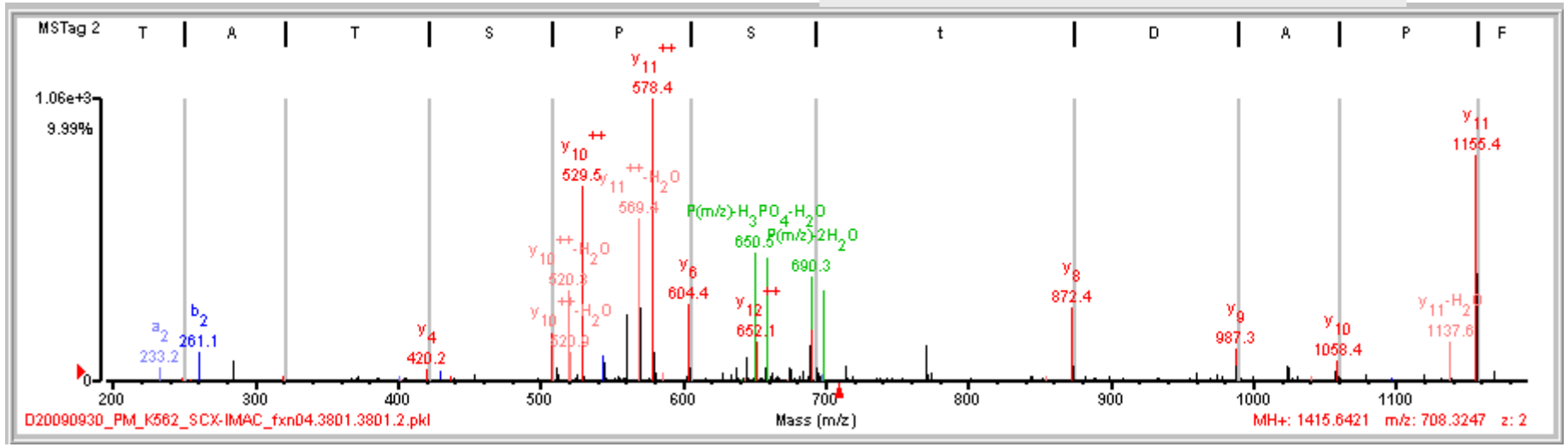
Localizing a Phosphorylation Site

same spectrum
2 different interpretations

L/F|P/A/D|T/s/P/S T A\T K



L/F|P/A/D|t S/P/S T A\T K

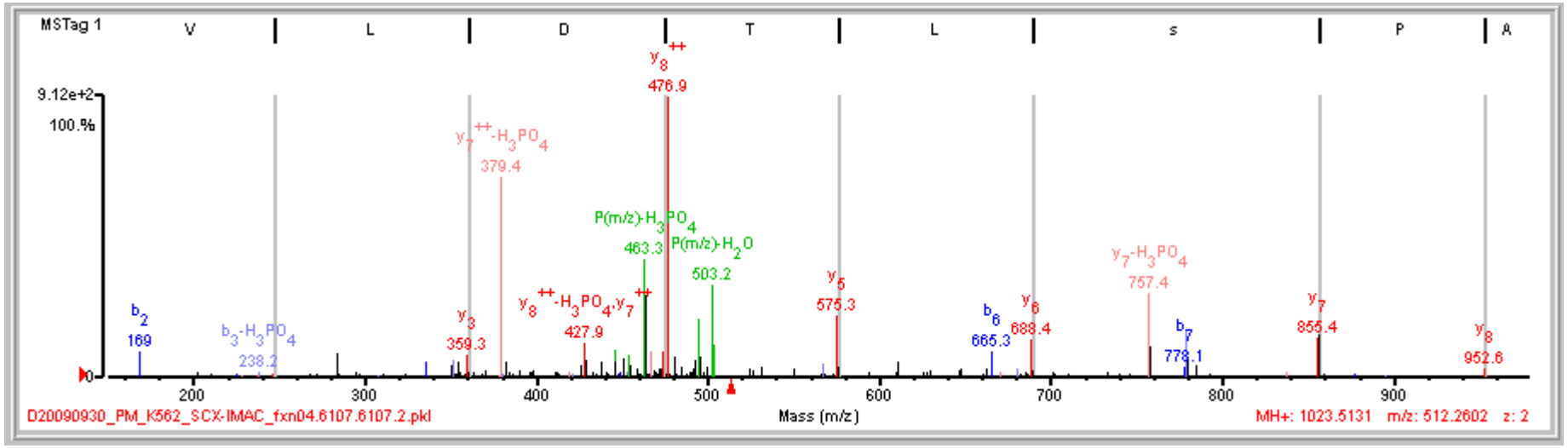


PTM Site Localization

Test all Locations, Examine Score Gaps

	Locations Tested	Conclusion
No possible ambiguity	AVsEEQQPALK	AVS(1.0)EEQQPALK # PO_4 sites = # S,T, or Y
Single Site	APsLTDLVK *	APS(0.99)LT(0.0)DLVK
	APSLtDLVK -	
	sSSAGPEGPQLDVPR *	S(0.50)S(0.50)S(0.0)AGPEGPQLDVPR
	SsSAGPEGPQLDVPR *	
	SSsAGPEGPQLDVPR -	
Multiple Sites	VTNDIsPEsSPGVGR *	VT(0.0)NDIS(0.99)PES(0.50)S(0.50)PGVGR
	VTNDIsPESsPGVGR *	
	VTNDISPEssPGVGR -	
	VtNDIsPESSPGVGR -	
	VtNDISPEsSPGVGR -	
	VtNDISPESsPGVGR -	

PTM Site Localization – Confident Localization

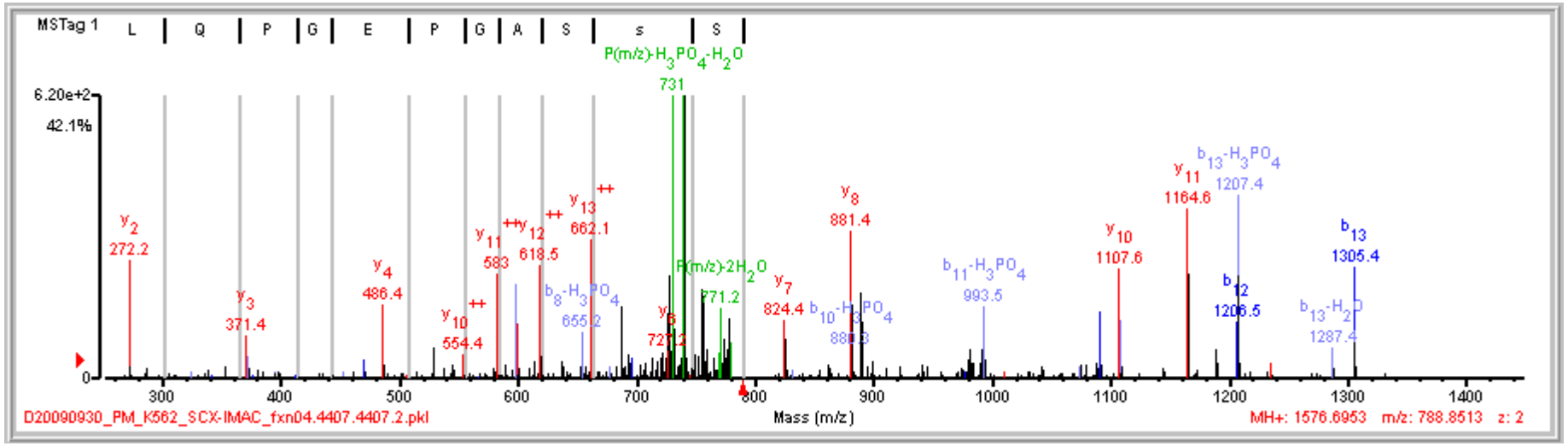


(K)A/P|s/L/T D|L\V K(S)

APS(0.99)LT(0.0)DLVK

y_6, y_7 ions provide confident localization to the Ser

PTM Site Localization – Ambiguous Localization

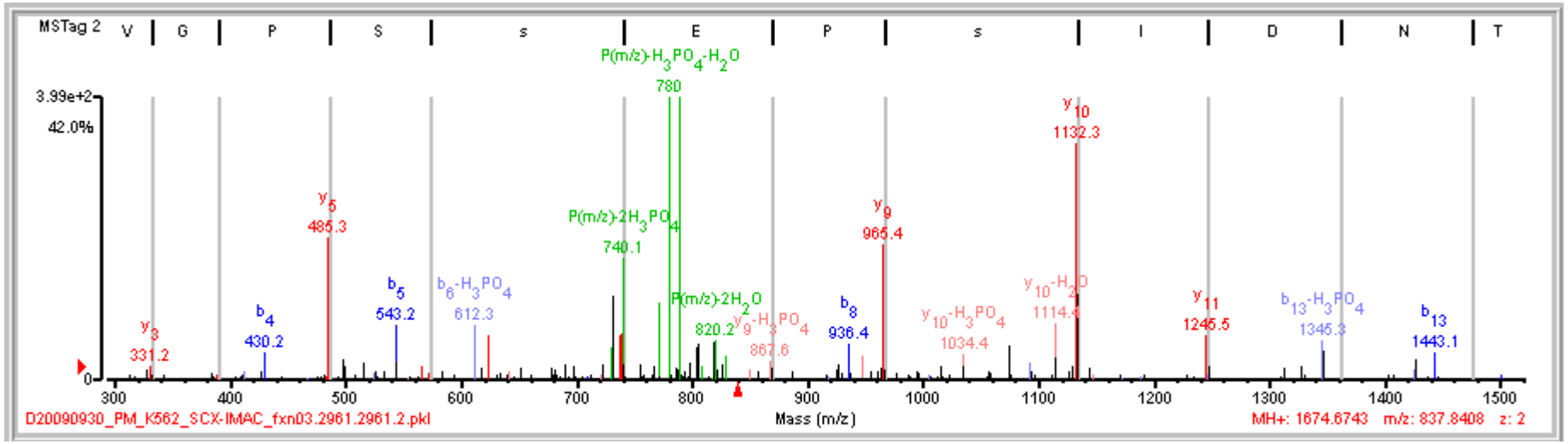


(R)S s/S/A/G/P E/G/P Q L|D|V|P R(E)

S(0.50)S(0.50)S(0.0)AGPEGPQLDVPR

y_{13}^{++} ion excludes localization to Ser 3

PTM Site Localization – Ambiguous Localization
2 sites: 1 confident, 1 ambiguous



(R)V T N D|I|s/P E|s S/P G V\G R(R)

VT(0.0)NDIS(0.99)PES(0.50)S(0.50)PGVGR

y_9, y_{10} ions provide confident localization to the Ser -6

y_9, y_{10} ions provide ambiguous localization to Ser-9, Ser-10

Spectrum Mill Variable Modification Localization Score

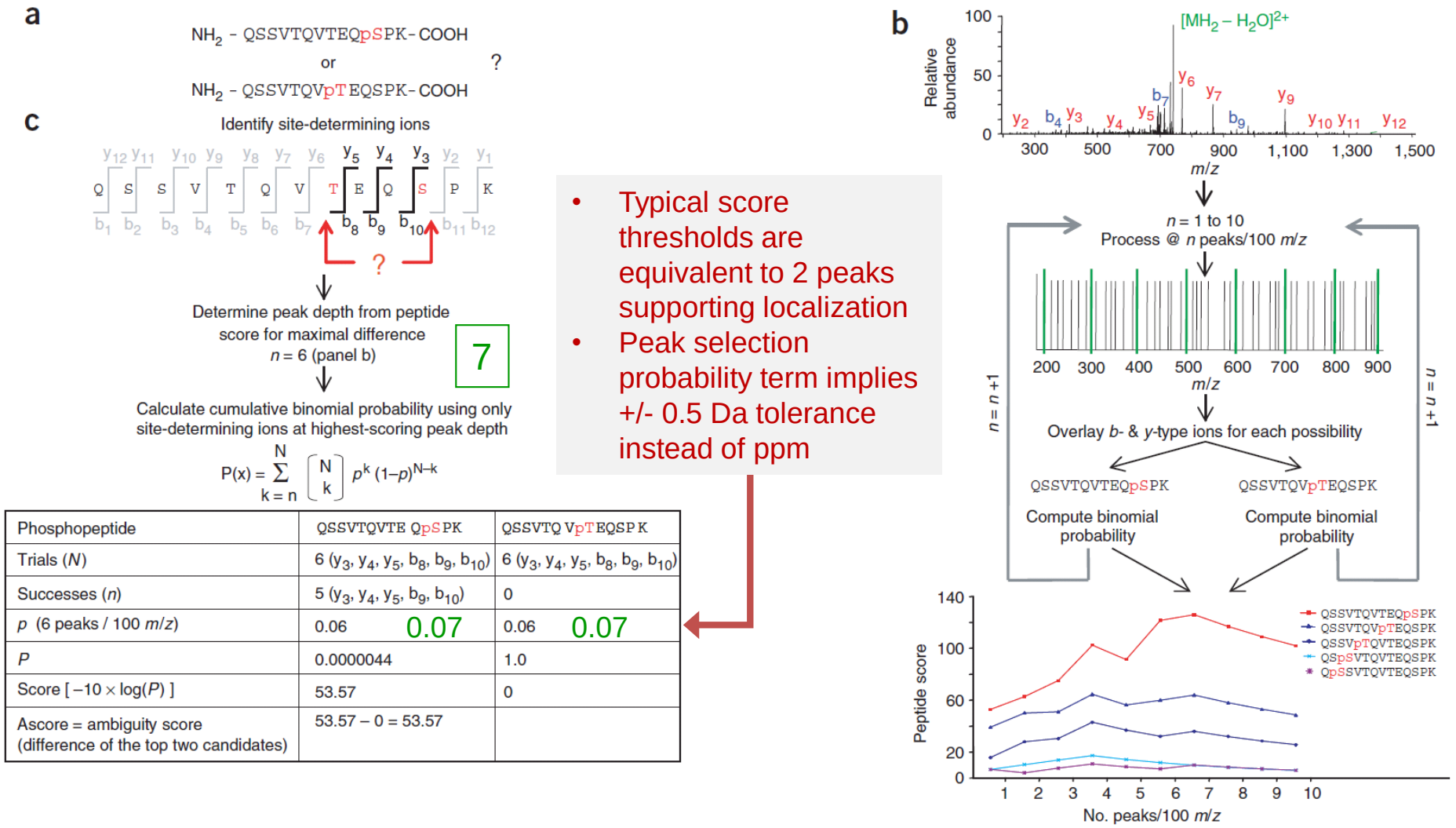
VML score = Difference in Score of same identified sequences with different variable modification localizations

VML score > 1.1 indicates confident localization

Why a threshold value of 1.1?

- 1 implies that there is a distinguishing ion of b or y ion type
- 0.1 means that when unassigned, the peak is 10% the intensity of the base peak

Phosphosite Localization Scoring - Ascore



- Typical score thresholds are equivalent to 2 peaks supporting localization
- Peak selection probability term implies +/- 0.5 Da tolerance instead of ppm

Phosphopeptide	QSSVTQVTE QpSPK	QSSVTQ VpTEQSP K
Trials (*N*)	6 (y_3, y_4, y_5, b_8, b_9, b_{10})	6 (y_3, y_4, y_5, b_8, b_9, b_{10})
Successes (*n*)	5 (y_3, y_4, y_5, b_9, b_{10})	0
p (6 peaks / 100 *m/z*)	0.06 0.07	0.06 0.07
P	0.0000044	1.0
Score [$-10 \times \log(P)$]	53.57	0
Ascore = ambiguity score (difference of the top two candidates)	53.57 – 0 = 53.57	

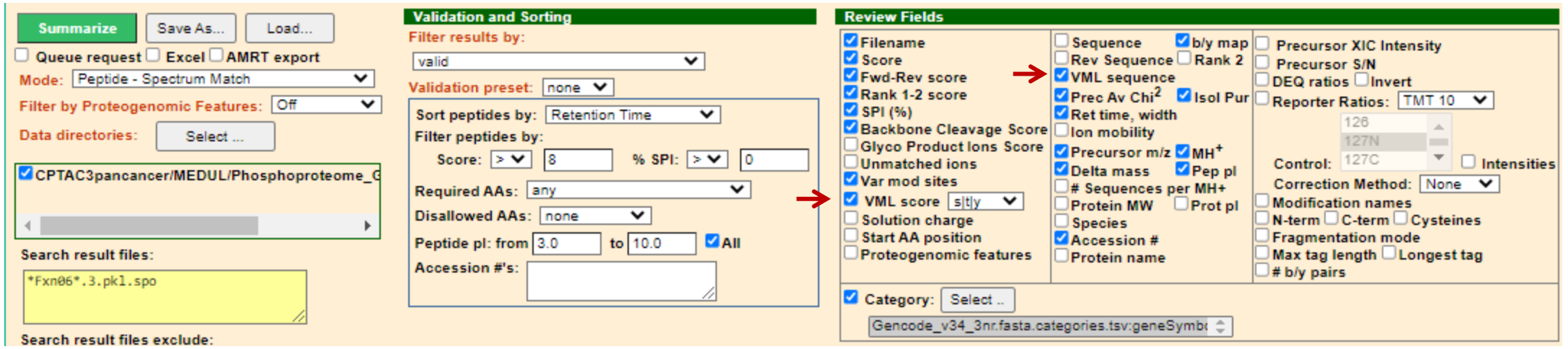
Beausoleil SA, Villen J, Gerber SA, Rush J, Gygi SP (2006) *Nat Biotechnol* 24:1285–1292.

http://ascore.med.harvard.edu/
Supports Sequest results only, Linux only

VML Score Features Reported

Filename	z	Score	Fwd-Rev Score	Rank1-Rank2 Score	SPI (%)	# Backbone Cleavages	Missed Cleavages	#KR	Score VML	# STY	# sty Sites	# Loc sty Sites	# Ambig sty Sites	Variable Sites	VM site Flanks	Sequence Map VML sequence
BL20160819_FM_Medullo_Phosphoproteome_Plex02_Fxn06.0005.0005.3	3	12.94	12.94	12.94	68.5	8	0	1	0.460	5	1	0	1	S99s	RRHSHSHsPMSTRRR	(R)H\S\H\S\H-s/P/M S T/R(R) HS(0.0)HS(0.50)HS(0.50)PMS(0.0)T(0.0)R
BL20160819_FM_Medullo_Phosphoproteome_Plex02_Fxn06.1390.1390.3	3	9.47	9.47	9.47	41.4	6	0	1	99	1	1	1	0	S623s	AHHAAGQsVRSGRLG	(K)A\H H\A\A\G Q s/V/R(S) AHHAAGQS(1.0)VR
BL20160819_FM_Medullo_Phosphoproteome_Plex02_Fxn06.3872.3872.3	3	11.65	11.65	11.65	63.1	7	1	2	2.135	2	1	1	0	S437s	VLTKSGRsAHQVARY	(K)S G\R\s/A/H/Q\V A/R(Y) S(0.0)GRS(0.99)AHQVAR
BL20160819_FM_Medullo_Phosphoproteome_Plex02_Fxn06.3908.4284.3	3	11.77	8.39	8.29	68.4	6	1	2	0.277	4	1	0	1	S409s	PPPKTRHsPTPQQSN	(K)T R H\s/P T/P/Q/Q S N/R(T) T(0.33)RHS(0.33)PT(0.33)PQQS(0.0)NR
BL20160819_FM_Medullo_Phosphoproteome_Plex02_Fxn06.4546.4546.3	3	9.98	6.78	4.81	71.6	6	1	2	1.214	4	1	1	0	S491s	EGKSSSRsPGPHPRS	(K)S S\S R\s/P G/P H/P/R(S) S(0.0)S(0.0)S(0.0)RS(0.99)PGPHPR
BL20160819_FM_Medullo_Phosphoproteome_Plex02_Fxn06.4548.4548.3	3	9.55	5.39	5.04	50.2	8	1	2	4.086	3	1	1	0	S61s	PASSHREsPRGSGGA	(R)G\P A\S\S\H R\E\s/P/R(G) GPAS(0.0)S(0.0)HRES(0.99)PR
BL20160819_FM_Medullo_Phosphoproteome_Plex02_Fxn06.4638.4638.3	3	9.93	9.05	8.54	68.0	6	2	3	0.751	5	1	0	1	S298s	SRRSTTKsPGPSRRS	(R)S T\T K-s/P/G/P S R/R(S) S(0.0)T(0.0)T(0.50)KS(0.50)PGPS(0.0)RR
BL20160819_FM_Medullo_Phosphoproteome_Plex02_Fxn06.4673.4673.3	3	8.74	7.18	7.35	67.5	5	0	2	1.945	2	1	1	0	S694s	PNKRHSPsPRPRAPQ	(R)H\S\P s/P R/P/R(A) HS(0.0)PS(0.99)PRPR
BL20160819_FM_Medullo_Phosphoproteome_Plex02_Fxn06.4881.4881.3	3	8.33	3.69	2.24	57.1	5	2	3	1.379	3	1	1	0	T310t	GTARRTGtPSDPRRR	(R)R T G\t/P S D\P-R/R(R) RT(0.0)GT(0.99)PS(0.0)DPRR
BL20160819_FM_Medullo_Phosphoproteome_Plex02_Fxn06.5095.5095.3	3	13.13	13.61	10.25	78.1	7	0	2	7.333	4	1	1	0	S544s	GSSAPEHsPRTSGLG	(R)S\P R\P G S S\A\P E\H s/P/R(T) S(0.0)PRPGS(0.0)S(0.0)APEHS(0.99)PR
BL20160819_FM_Medullo_Phosphoproteome_Plex02_Fxn06.5077.5077.3	3	11.46	7.21	6.63	69.7	7	0	2	99	1	1	1	0	S287s	LAQLLARsPPPPHRP	(R)s/P/P/P/P/H R/P/R(L) S(1.0)PPPPHRPR

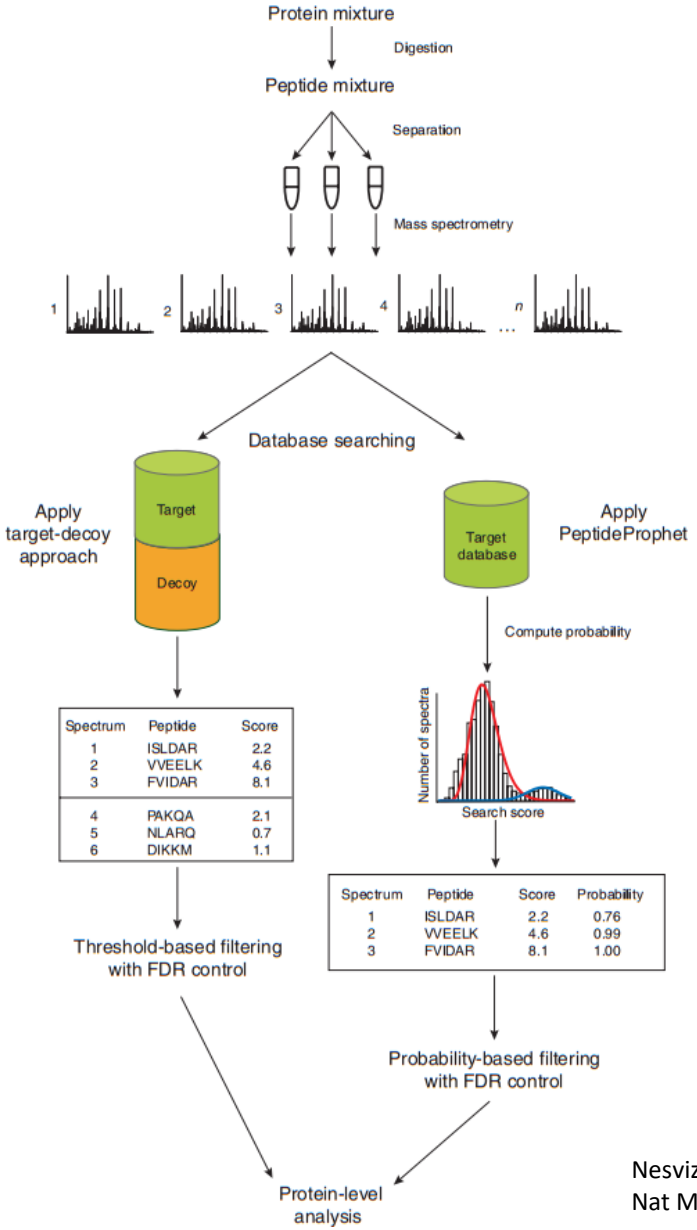
Site flanks +/- 7 AA in protein



Methods to control false discovery rates (FDR) of MS/MS search results

Estimating False Discovery Rate (FDR)

- Basic Approaches
  - Target-decoy approach
  - Bayesian posterior probability
- Goals
  - Accurate error estimation
  - Accurate sensitivity estimation
  - Comparable across labs and experiments



Nesvizhskii AI, Vitek O, Aebersold R. Nat Methods. 2007 Oct;4(10):787-97.

Creating a Target/Decoy Database

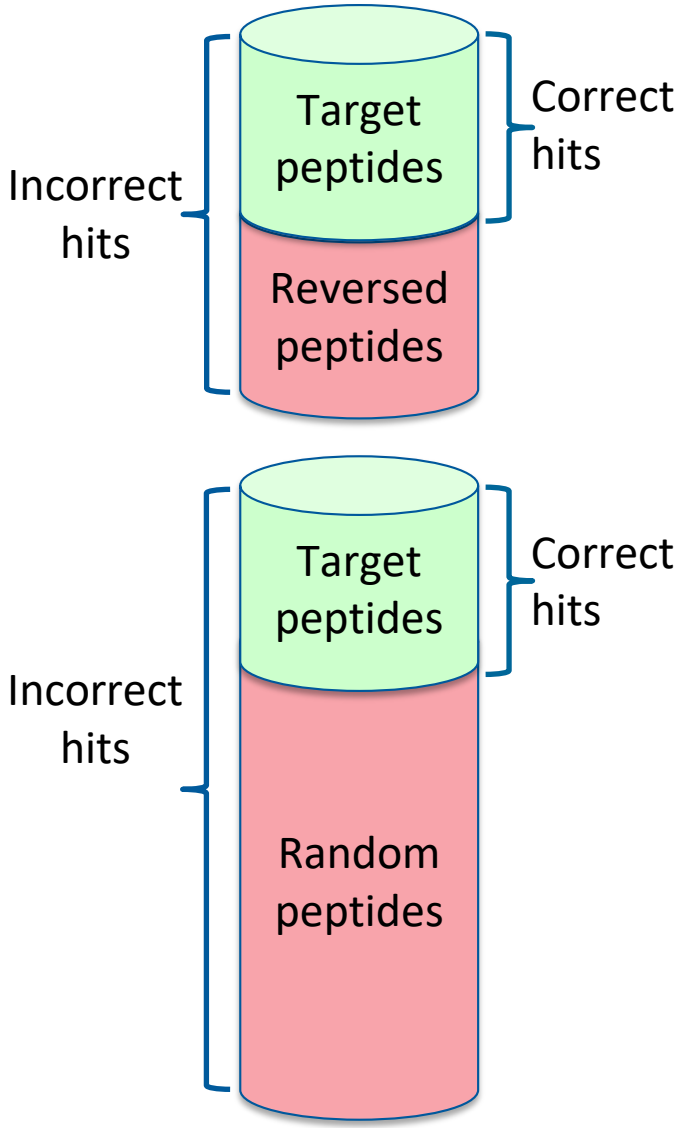
Reverse each protein sequence

- Amino acid composition conserved
- #, length of proteins conserved
- **# of peptides conserved**

Shuffle each protein sequence

- Amino acid composition conserved
- #, length of proteins conserved
- **# of peptides increases**
  - **Protein homology is lost**
  - **Yields more distinct peptides**

```
ProtA: ACDERFGHIK -> ACDER  FGHIK   3
ProtB: LMNPRFGHIK -> LMNPR  FGHIK
                   shuffled
RandA: CHERIGDAEK -> CHER  IGDAEK   5
RandB: PLKMHFRIGN -> PLK MHFR IGN
```



from James Eddes

SM

Reverse each peptide

SAMPLER becomes SELPMAR

Faster, skips digestion overhead

SM ignores peptides if decoy = target

Target-Decoy Thresholding – 1D Example

- Sort hits in descending score order
- Count target hits above chosen score
- Count decoy hits above chosen score
- For each score calculate FDR
  - FDR = (2 * FP) / (FP + TP)
- Filter list at score corresponding to chosen FDR

score

Rank	Xcorr	Sequence	Reference	Est. FP	Est TP	Est FN	Est TN	precision	sensitivity	accuracy
1	8.43	R.QLLLTADDRVNPCIGGVILFHETLYQK.A	IPI:IPI00549682.2	0	1	2152	456	1.000	0.000	0.175
2	8.15	R.SGPFGQIFRPDNFVFGQSGAGNNWAK.G	IPI:IPI00007752.1	0	2	2151	456	1.000	0.001	0.176
3	7.32	K.FEGEPATHTQPGVQLQSNTYDLQESNVR.L	IPI:IPI00216139.1	0	3	2150	456	1.000	0.001	0.176
4	7.30	K.GHYTEGAELVDSVLDVVR.K	IPI:IPI00007752.1	0	4	2149	456	1.000	0.002	0.176
5	7.12	K.DYPVVSIEDPFDQDDWGAWQK.F	IPI:IPI00465248.4	0	5	2148	456	1.000	0.002	0.177
6	7.08	K.EKLCYVALDFEQEM*ATAASSSSLEK.S	IPI:IPI00642997.1	0	6	2147	456	1.000	0.003	0.177
7	7.08	R.DLKPENLLIDQQGYIQVTDFGFAK.R	IPI:IPI00217960.1	0	7	2146	456	1.000	0.003	0.177
8	7.08	R.SHSTEPGLVLTLGQGDVGQLGLGENVM*ER.K	IPI:IPI00001661.2	0	8	2145	456	1.000	0.004	0.178
9	7.07	R.TTGIVMDSGDGVTHTVPIYEGYALPHAILR.L	IPI:IPI00642997.1	0	9	2144	456	1.000	0.004	0.178
10	6.99	K.KFDLGQDVIDFTGHALALYR.T	IPI:IPI00514395.1	0	10	2143	456	1.000	0.005	0.179
741	3.64	K.VILALGDYM*GATCHACIGGTNVR.N	IPI:IPI00328328.3	0	741	1412	456	1.000	0.344	0.459
742	3.64	K.EIDDSVLGQTGPYR.R	IPI:IPI00306516.1	0	742	1411	456	1.000	0.345	0.459
743	3.64	K.DVTNNVHYENYR.S	IPI:IPI00033025.7	0	743	1410	456	1.000	0.345	0.460
744	3.64	K.EHWNPAIVALVYNVLK.A	IPI:IPI00556528.1	0	744	1409	456	1.000	0.346	0.460
745	3.63	R.YISPDQLADLYK.G	IPI:IPI00013769.1	0	745	1408	456	1.000	0.346	0.460
746	3.63	K.HLREIDRLSLLEGRPEEDSDIEIYK.R	##IPI:IPI00479822.2	2	744	1409	454	0.997	0.346	0.459
747	3.63	K.IPEGLFDPSNVK.G	IPI:IPI00216622.1	2	745	1408	454	0.997	0.346	0.460
748	3.63	K.IDLGVHVDGFIANVAHTFVVDVAQGTQVTGR.K	IPI:IPI00299000.3	2	746	1407	454	0.997	0.346	0.460
749	3.63	R.KGETIFVERPLVAAQFLWNALYR.Y	IPI:IPI00013789.4	2	747	1406	454	0.997	0.347	0.460
750	3.62	K.SM*AASGNLGHTPFVDEL.-	IPI:IPI00305383.1	2	748	1405	454	0.997	0.347	0.461
1251	2.93	K.ELAEQLGLSTGEK.E	IPI:IPI00290460.3	14	1237	916	442	0.989	0.575	0.644
1252	2.92	R.IDAVNAETIR.E	IPI:IPI00289535.6	14	1238	915	442	0.989	0.575	0.644
1253	2.92	R.LFLLPQAHHIM*GTIGAIQSALAPPDSSTLLK.L	##IPI:IPI00444490.1	16	1237	916	440	0.987	0.575	0.643
1254	2.92	R.FLLSLPEHR.G	IPI:IPI00014151.3	16	1238	915	440	0.987	0.575	0.643
1255	2.92	R.SVDPDSPAEASGLR.A	IPI:IPI00003527.4	16	1239	914	440	0.987	0.575	0.644
1256	2.92	R.M*TDLLPNPAM*KGAPHAQASLR.R	##IPI:IPI00658084.1	18	1238	915	438	0.986	0.575	0.642
1257	2.92	R.VGQAVALR.A	IPI:IPI00012835.1	18	1239	914	438	0.986	0.575	0.643
1258	2.92	K.M*LLYTEVTR.Y	IPI:IPI00010154.3	18	1240	913	438	0.986	0.576	0.643
1259	2.92	R.LIVENLSSR.V	IPI:IPI00555857.1	18	1241	912	438	0.986	0.576	0.644
1260	2.91	K.WLNENAVEK.V	IPI:IPI00012007.5	18	1242	911	438	0.986	0.577	0.644
1801	2.27	R.FM*EQVIFK.Y	IPI:IPI00028091.1	76	1725	428	380	0.958	0.801	0.807
1802	2.27	K.EIRRVLEVLMEAM*K.A	##IPI:IPI00020915.1	78	1724	429	378	0.957	0.801	0.806
1803	2.26	R.M*VLAAAGGVEHQQLLDLAQK.H	IPI:IPI00013847.3	78	1725	428	378	0.957	0.801	0.806
1804	2.26	R.LEQDEYALR.S	IPI:IPI00022793.4	78	1726	427	378	0.957	0.802	0.806
1805	2.26	R.ALKIPQVRSTDM*IAAR.R	##IPI:IPI00457056.1	80	1725	428	376	0.956	0.801	0.805
1806	2.26	R.GGPNIITLADIVKDPVSR.T	IPI:IPI00016613.2	80	1726	427	376	0.956	0.802	0.806
1807	2.26	R.TQGVSTTDLVGR.M	IPI:IPI00015285.6	80	1727	426	376	0.956	0.802	0.806
1808	2.26	K.HFVLDECDK.M	IPI:IPI00644431.1	80	1728	425	376	0.956	0.803	0.806
1809	2.26	K.KGEEALFTTR.E	IPI:IPI00019004.1	80	1729	424	376	0.956	0.803	0.807
1810	2.26	R.ISPELSAEAM*GSGTSVVNR.R	##IPI:IPI00027003.1	82	1728	425	374	0.955	0.803	0.806
2101	1.94	R.CAHVARTYSIGR.S	IPI:IPI00396391.7	162	1939	214	294	0.923	0.901	0.856
2102	1.94	R.VKIERQEVSK.E	##IPI:IPI00028214.3	164	1938	215	292	0.922	0.900	0.855
2103	1.93	R.VSPQASVVLEHR.E	IPI:IPI00167300.3	164	1939	214	292	0.922	0.901	0.855
2104	1.93	K.FCLDNGAK.S	IPI:IPI00169383.2	164	1940	213	292	0.922	0.901	0.856
2105	1.93	R.VQAVERHLRSLEAEATR.K	##IPI:IPI00305267.2	166	1939	214	290	0.921	0.901	0.854
2106	1.93	K.EKEKASWSSLSMDEK.V	IPI:IPI00006579.1	166	1940	213	290	0.921	0.901	0.855
2107	1.93	R.YEEIVK.E	IPI:IPI00180730.1	166	1941	212	290	0.921	0.902	0.855
2108	1.93	K.LYENFISEFEHR.V	IPI:IPI00549672.2	166	1942	211	290	0.921	0.902	0.856
2109	1.93	K.ALAPTWEQLALGLEHSETVK.I	IPI:IPI00171438.2	166	1943	210	290	0.921	0.902	0.856
2110	1.93	R.CEPIAM*TVPR.K	IPI:IPI00010133.1	166	1944	209	290	0.921	0.903	0.856
2581	1.04	R.KSSLVTSK.L	IPI:IPI00644717.1	430	2151	2	26	0.833	0.999	0.834
2582	1.03	-.M*DWCCSM*PTK.G	IPI:IPI00457291.1	430	2152	1	26	0.833	1.000	0.835
2583	1.02	K.YVPPRR.V	IPI:IPI00304809.6	430	2153	0	26	0.834	1.000	0.835
2584	1.01	K.LDRHIIPPTRTHLFLLGK.L	##IPI:IPI00478412.2	432	2152	1	24	0.833	1.000	0.834
2585	1.01	K.NLSYDERSISIM*K.V	IPI:IPI00045496.1	432	2153	0	24	0.833	1.000	0.834
2586	0.99	R.GPGSLGRR.L	##IPI:IPI00552079.1	434	2152	1	22	0.832	1.000	0.833
2587	0.99	R.LSSQTRLPR.K	##IPI:IPI00655766.1	436	2151	2	20	0.831	0.999	0.832
2588	0.98	R.APQRLKR.A	##IPI:IPI00305305.3	438	2150	3	18	0.831	0.999	0.831
2589	0.97	R.LSVLQVRRAR.G	##IPI:IPI00456163.1	440	2149	4	16	0.830	0.998	0.830
2590	0.91	R.TIPLTFK.D	IPI:IPI00002841.1	440	2150	3	16	0.830	0.999	0.830

Decoy hits (reversed database)

$$FDR = \frac{n_{decoy}}{n_{decoy} + n_{forward}} \times 2$$

$$n_{reversed} = 13$$

$$n_{forward} = 2590 - 13 = 2577$$

$$FDR = \frac{13}{13 + 2577} \times 2$$

$$= 0.0100386 \approx 1\%$$

Target/Decoy FDR Tutorial
Elias & Gygi, *Nature Methods*, 4, 207-214, **2007**.

Target-Decoy Thresholding – 3D to 7D possible in SM

In SM to achieve FDR goal:

1. Filter - sequence length range (min, max)
2. Filter - min backbone cleavage score threshold is applied.
3. Filter - limit by modifications (i.e. phospho sty or not)
4. Filter - precursor mass error distributions calculated for each LC-MS/MS run, retain PSMs within 3 stddev of mean

3D 5. Optimize – thresholds for score, delta Rank1 – Rank2; separately for each precursor charge

   a. enforces the notion that a credible ID should have both a suitable absolute score and be separated from the rank2 match by a suitable margin

Example: for z=2 PSMs only

delta Rank1 – Rank2 (rows)

Score(columns)

Optimum: 15,010 PSMs @ FDR 1.0

score 4.4

dR1-R2 1.0

	8.0	7.9	7.8	7.7	7.6	7.5	7.4	7.3	7.2	7.1	7.0	6.9	6.8	6.7	6.6	6.5	6.4	6.3	6.2	6.1	6.0	5.9	5.8	5.7	5.6	5.5	5.4	5.3	5.2	5.1	5.0	4.9	4.8	4.7	4.6	4.5	4.4	4.3	4.2	4.1
1.1	0.53 12904	0.54 12975	0.57 13038	0.59 13118	0.64 13189	0.63 13266	0.63 13343	0.64 13415	0.64 13486	0.65 13551	0.66 13612	0.66 13682	0.65 13748	0.65 13812	0.68 13856	0.69 13917	0.70 13984	0.70 14034	0.71 14090	0.74 14142	0.75 14187	0.74 14237	0.74 14288	0.75 14346	0.75 14379	0.75 14418	0.75 14464	0.77 14514	0.78 14553	0.78 14582	0.78 14630	0.78 14665	0.80 14705	0.83 14741	0.83 14781	0.84 14809	0.85 14834	0.85 14858	0.85 14884	0.87 14908
1.0	0.63 13030	0.64 13104	0.67 13170	0.69 13253	0.74 13327	0.73 13404	0.73 13483	0.74 13560	0.73 13632	0.76 13700	0.77 13761	0.77 13835	0.76 13902	0.76 13966	0.79 14011	0.80 14072	0.81 14141	0.80 14192	0.81 14249	0.85 14304	0.86 14350	0.86 14401	0.86 14453	0.87 14512	0.87 14548	0.86 14587	0.86 14634	0.89 14686	0.91 14727	0.92 14757	0.92 14805	0.93 14841	0.95 14881	0.98 14917	0.98 14957	0.99 14985	1.00 15010	1.01 15035	1.01 15061	1.03 15088
0.9	0.66 13070	0.67 13147	0.70 13214	0.72 13298	0.76 13372	0.76 13450	0.75 13530	0.76 13607	0.76 13679	0.79 13749	0.80 13810	0.79 13884	0.79 13952	0.78 14017	0.81 14064	0.82 14127	0.83 14197	0.83 14248	0.84 14306	0.88 14361	0.89 14407	0.90 14460	0.90 14512	0.91 14573	0.90 14610	0.90 14649	0.90 14696	0.92 14749	0.95 14790	0.96 14820	0.96 14869	0.97 14906	0.99 14946	1.01 14983	1.01 15027	1.02 15055	1.03 15080	1.05 15105	1.04 15132	1.07 15160
0.8	0.72 13108	0.73 13185	0.75 13252	0.78 13336	0.82 13410	0.82 13488	0.81 13569	0.82 13646	0.82 13718	0.84 13789	0.85 13850	0.85 13924	0.84 13993	0.84 14058	0.86 14105	0.88 14168	0.90 14239	0.90 14291	0.91 14349	0.94 14405	0.95 14451	0.97 14506	0.96 14559	0.97 14620	0.97 14658	0.97 14697	0.96 14744	0.99 14798	1.02 14842	1.04 14873	1.03 14922	1.04 14960	1.07 15000	1.09 15038	1.10 15084	1.11 15113	1.12 15138	1.13 15163	1.13 15190	1.16 15219
0.7	0.73 13141	0.74 13218	0.77 13285	0.79 13370	0.85 13445	0.84 13525	0.85 13607	0.86 13684	0.86 13756	0.88 13828	0.89 13891	0.89 13966	0.88 14035	0.88 14100	0.90 14147	0.91 14211	0.95 14284	0.95 14336	0.96 14394	1.00 14450	1.01 14496	1.02 14552	1.01 14605	1.02 14666	1.03 14706	1.03 14745	1.03 14792	1.05 14847	1.09 14891	1.10 14922	1.10 14972	1.11 15010	1.13 15050	1.17 15089	1.18 15135	1.19 15164	1.22 15191	1.24 15216	1.23 15244	1.27 15274
0.6	0.79 13166	0.80 13243	0.83 13310	0.85 13395	0.91 13470	0.90 13551	0.91 13635	0.92 13712	0.91 13784	0.94 13858	0.95 13921	0.94 13997	0.94 14066	0.93 14131	0.96 14178	0.97 14242	1.01 14315	1.00 14367	1.01 14425	1.05 14481	1.06 14527	1.07 14583	1.07 14637	1.07 14699	1.09 14739	1.10 14780	1.09 14827	1.12 14882	1.15 14928	1.16 14959	1.16 15011	1.17 15049	1.19 15090	1.23 15129	1.24 15175	1.25 15204	1.29 15231	1.31 15258	1.31 15286	1.34 15316
0.5	0.85 13188	0.86 13267	0.88 13335	0.91 13421	0.96 13496	0.96 13579	0.97 13664	0.98 13741	0.97 13813	0.99 13887	1.00 13951	1.00 14027	0.99 14096	0.99 14161	1.01 14209	1.02 14273	1.06 14346	1.06 14398	1.07 14456	1.10 14512	1.13 14559	1.14 14615	1.13 14669	1.14 14731	1.15 14771	1.16 14812	1.16 14859	1.18 14914	1.22 14960	1.23 14992	1.22 15044	1.25 15084	1.27 15125	1.31 15164	1.31 15210	1.34 15240	1.38 15267	1.40 15294	1.41 15324	1.45 15355
0.4	0.91 13242	0.92 13321	0.94 13391	0.96 13479	1.02 13554	1.03 13638	1.03 13724	1.04 13802	1.04 13874	1.06 13949	1.07 14014	1.08 14092	1.10 14163	1.11 14230	1.13 14278	1.14 14343	1.18 14416	1.17 14469	1.18 14529	1.22 14586	1.24 14634	1.25 14691	1.25 14746	1.26 14809	1.27 14850	1.29 14893	1.31 14943	1.33 14998	1.37 15045	1.38 15078	1.39 15133	1.41 15174	1.43 15215	1.47 15254	1.49 15301	1.51 15333	1.55 15360	1.57 15389	1.58 15420	1.63 15453
0.3	0.95 13278	0.96 13357	0.98 13427	1.01 13515	1.06 13591	1.07 13675	1.08 13761	1.08 13840	1.08 13913	1.10 13988	1.11 14053	1.12 14131	1.14 14202	1.15 14271	1.17 14319	1.18 14385	1.22 14458	1.21 14511	1.22 14571	1.26 14628	1.29 14679	1.33 14738	1.32 14793	1.33 14856	1.34 14897	1.37 14940	1.39 14990	1.41 15045	1.44 15093	1.45 15126	1.46 15181	1.48 15222	1.51 15263	1.54 15303	1.56 15350	1.59 15382	1.62 15409	1.65 15439	1.63 15470	1.70 15503
0.2	0.99 13302	1.00 13381	1.03 13452	1.05 13540	1.10 13617	1.11 13701	1.12 13787	1.12 13867	1.12 13940	1.14 14015	1.16 14081	1.17 14161	1.19 14232	1.22 14302	1.24 14350	1.25 14416	1.28 14490	1.28 14543	1.29 14603	1.32 14660	1.36 14711	1.39 14770	1.39 14825	1.41 14889	1.42 14930	1.44 14973	1.46 15024	1.49 15079	1.52 15127	1.53 15160	1.54 15215	1.56 15256	1.58 15297	1.62 15337	1.64 15384	1.66 15416	1.70 15445	1.72 15475	1.73 15506	1.78 15539

(score, dR1-R2: with FDR > 1.0)

FDR
PSMs

HLA class I: B5703_20180828

Machine Learning to Combine Multiple Scores- Percolator

Not yet done with SM

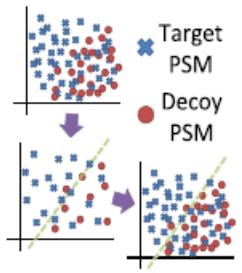
J. Am. Soc. Mass Spectrom. (2016) 27:1719–1727
DOI: 10.1007/s13361-016-1460-7

FOCUS: BIOINFORMATICS, SOFTWARE, AND MS-BASED "OMICS": RESEARCH ARTICLE

Fast and Accurate Protein False Discovery Rates on Large-Scale Proteomics Data Sets with Percolator 3.0

Matthew The,[1] Michael J. MacCoss,[2] William S. Noble,[2,3] Lukas Käll[1]

[1]Science for Life Laboratory, School of Biotechnology, KTH – Royal Institute of Technology, Box 1031, 17121, Solna, Sweden
[2]Department of Genome Sciences, School of Medicine, University of Washington, Seattle, WA 98195, USA
[3]Department of Computer Science and Engineering, University of Washington, Seattle, WA 98195, USA



Abstract. Percolator is a widely used software tool that increases yield in shotgun proteomics experiments and assigns reliable statistical confidence measures, such as *q* values and posterior error probabilities, to peptides and peptide-spectrum matches (PSMs) from such experiments. Percolator's processing speed has been sufficient for typical data sets consisting of hundreds of thousands of PSMs. With our new scalable approach, we can now also analyze millions of PSMs in a matter of minutes on a commodity computer. Furthermore, with the increasing awareness for the need for reliable statistics on the protein level, we compared several easy-to-understand protein inference methods and implemented the best-performing method—grouping proteins by their corresponding sets of theoretical peptides and then considering only the best-scoring peptide for each protein—in the Percolator package. We used Percolator 3.0 to analyze the data from a recent study of the draft human proteome containing 25 million spectra (PM:24870542). The source code and Ubuntu, Windows, MacOS, and Fedora binary packages are available from http://percolator.ms/ under an Apache 2.0 license.

SM scores/features to percolate

SM scores

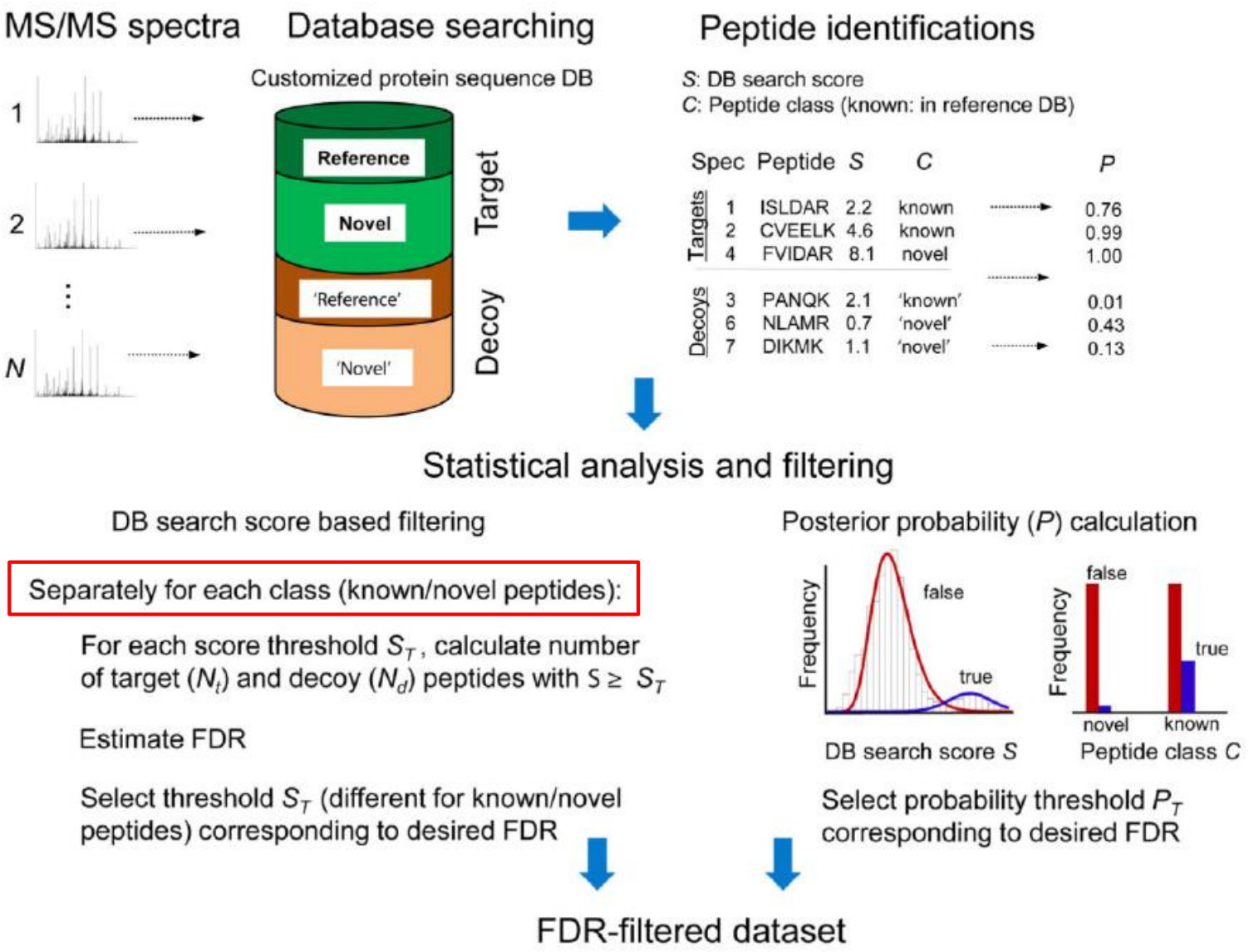
Score
- score (w/o penalty portion)
- penalty portion

Delta Rank1-Rank2 score
SPI – percent scored peak intensity
BCS – backbone cleavage score
Sequence coverage
Fragmentation category
sequences past parent filter (DB size dependency)
Parent mass error
Sequence length

SM Spectral Features

z
MH+
RT – retention time
Max sequence tag length
PIP – precursor ion purity
Precursor averagine Chi2

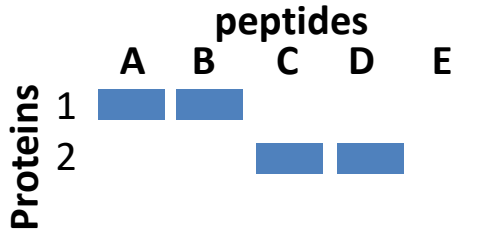
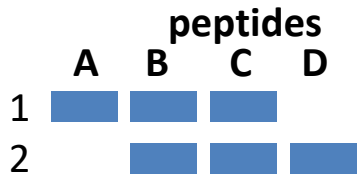
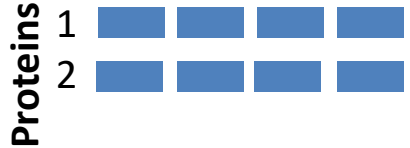
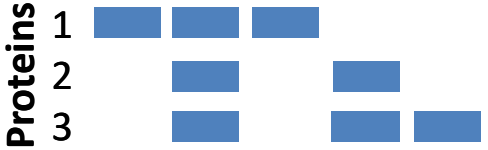
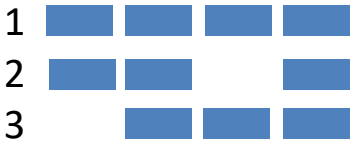
Calculate FDR separately for rare peptide subsets



Nesvizhskii AI. *Nat Methods*. 2014 11, 1114–1125.

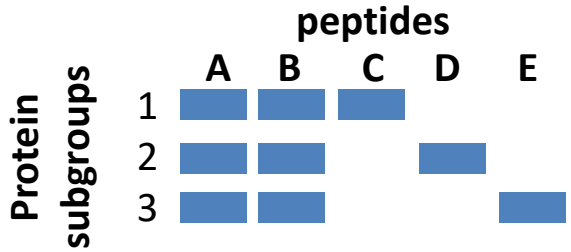
Collapsing Peptide Spectrum Matches (PSM's) for Quantitation at the Protein level and Phosphosite level

Protein Grouping and Shared peptides

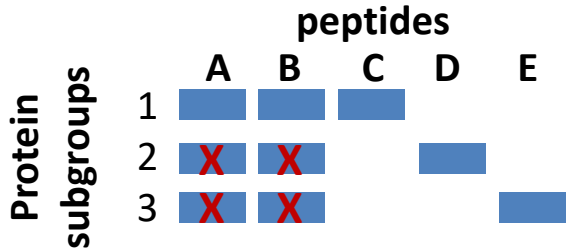
# groups	# subgroups				# groups	# subgroups
2	1, 1	2 Distinct proteins			1	2
1	1	2 Indistinguishable proteins			1	1
1	2	Subsumed Protein #2			1	1

Protein Inference Tutorial
Nesvizhskii, *Mol Cell Proteomics*, 4, 1419-1440, **2005**.

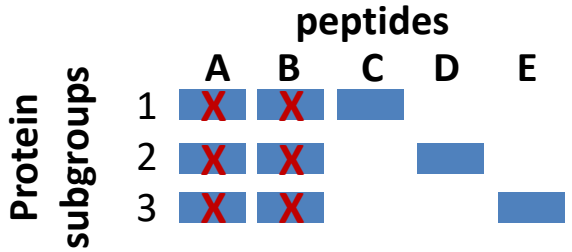
Shared Peptides and Protein Grouping



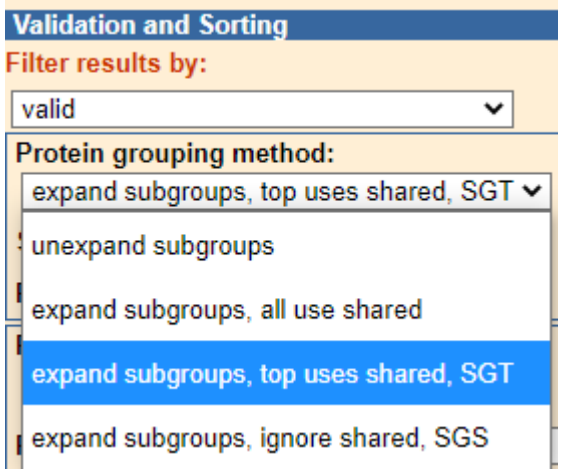
All subgroups used shared



Top subgroups uses shared
Subgroup Top (SGT)



Ignore shared
Subgroup Specific (SGS)



Ranking of subgroups is by protein score.
Protein score is Σ distinct peptide scores.

Protein Groups and Subgroups, Genes

luad-v1.0-proteome-ratio-norm-NArm.gct

#1.3																
12205	215	16	26													
id	id	id.description	geneSymbol	numColumnsProteinObserved	numSpectraProteinObserved	protein_mw	percentCoverage	numPepsUnique	scoreUnique	species	orfCategory	accession_number	accession_numbers	subgroupNum	entry_name	GeneSymbol
Sample.ID	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na
Sample	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na
Experiment	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na
Channel	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na
Type	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na
Aliquot	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na
Smoking.Status	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na
Stage	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na
Region.of.Origin	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na
Country.of.Origin	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na
Age	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na
Gender	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na
Ethnicity	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na
Height.cm	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na
Weight.kg	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na
BMI	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na
Cigarettes.per.Day	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na
Pack.Years.Smoked	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na
Smoking.History	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na
Secondhand.Smoke	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na
QC.status	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na
QC.status.1	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na
id	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na
normalization	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na
normalization.center	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na
normalization.scale	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na
NP_001317422.1	NP_001317	protein O	OSCP1	25	192	44,643	75.5	25	320.89	HOMO SAPIENS		NP_001317	NP_001317	4385.1	protein O	OSCP1
NP_001106273.1	NP_001106	sodium/c	SLC8A1	23	148	105,600	31.3	22	320.75	HOMO SAPIENS		NP_001106	NP_001106	4386.1	sodium/c	SLC8A1
NP_620156.1	NP_620156	aldose 1-	GALM	25	709	37,993	66	20	320.67	HOMO SAPIENS		NP_620156	NP_620156	4387.1	aldose 1-	GALM
NP_001446.1	NP_001446	forkhead	FOXO3	25	216	71,561	46.8	21	320.63	HOMO SAPIENS		NP_001446	NP_001446	4388.1	forkhead	FOXO3
NP_002006.2	NP_002006	forkhead	FOXO1	25	198	70,060	38.1	18	280.85	HOMO SAPIENS		NP_002006	NP_002006	4388.2	forkhead	FOXO1
NP_005929.2	NP_005929	forkhead	FOXO4	25	83	53,969	19.2	8	96.07	HOMO SAPIENS		NP_005929	NP_005929	4388.3	forkhead	FOXO4
NP_001278210.2	NP_001278	forkhead	FOXO6	24	79	57,486	18.4	8	92.14	HOMO SAPIENS		NP_001278	NP_001278	4388.4	forkhead	FOXO6
NP_057250.1	NP_057250	E3 SUMO-	PIAS1	25	159	72,805	46.3	23	320.61	HOMO SAPIENS		NP_057250	NP_057250	4389.1	E3 SUMO-	PIAS1
NP_006090.2	NP_006090	E3 SUMO-	PIAS3	20	85	69,043	35.8	18	191.87	HOMO SAPIENS		NP_006090	NP_006090	4389.2	E3 SUMO-	PIAS3
NP_001243764.1	NP_001243	queuine	QTRT2	25	258	49,116	60.8	19	320.61	HOMO SAPIENS		NP_001243	NP_001243	4390.1	queuine	QTRT2
NP_001243463.1	NP_001243	tricarboxy	SLC25A1	25	669	35,394	55.6	17	317.88	HOMO SAPIENS		NP_001243	NP_001243	4429.1	tricarboxy	SLC25A1
NP_001274316.1	NP_001274	tricarboxy	SLC25A1	25	500	23,438	66.3	13	251.51	HOMO SAPIENS		NP_001274	NP_001274	4429.2	tricarboxy	SLC25A1
NP_006811.2	NP_006811	interfero	IFI44L	24	206	52,006	63.9	23	317.83	HOMO SAPIENS		NP_006811	NP_006811	4430.1	interfero	IFI44L
NP_003013.1	NP_003013	SH3 doma	SH3BGRL	25	1131	12,774	91.2	16	317.77	HOMO SAPIENS		NP_003013	NP_003013	4431.1	SH3 doma	SH3BGRL
NP_116193.2	NP_116193	2-aminoe	ADO	25	355	30,093	78.5	21	317.74	HOMO SAPIENS		NP_116193	NP_116193	4432.1	2-aminoe	ADO

id	C3N.01799	C3N.01799.N	C3L.01890	C3L.01890.N	C3N.00572	C3N.00572.N	C3N.02423	C3N.02423.N	C3N.02729	C3N.02729.N	C3L.00263	C3L.00263.N	C3N.01410	C3N.01410.N	C3N.00578	C3N.00578.N	C3N.02587	C3N.02587.N	C3L.00893
Sample.ID	C3N.01	C3N.01	C3L.01	C3L.01	C3N.00	C3N.00	C3N.02	C3N.02	C3N.02	C3N.02	C3L.00	C3L.00	C3N.01	C3N.01	C3N.00	C3N.00	C3N.02	C3N.02	C3L.00
Sample	C3N-01	C3N-01	C3L-01	C3L-01	C3N-00	C3N-00	C3N-02	C3N-02	C3N-02	C3N-02	C3L-00	C3L-00	C3N-01	C3N-01	C3N-00	C3N-00	C3N-02	C3N-02	C3L-00
Experiment	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	3.00	3.00	3.00
Channel	#####	127N	127C	128N	128C	129N	129C	130N	#####	127N	127C	128N	128C	129N	129C	130N	#####	127N	127C
Type	Tumor	Norma	Tumor	Norma	Tumor	Norma	Tumor	Norma	Tumor	Norma	Tumor	Norma	Tumor	Norma	Tumor	Norma	Tumor	Norma	Tumor
Aliquot	CPT00	CPT00	CPT00	CPT00	CPT00	CPT00	CPT01	CPT01	CPT01	CPT01	CPT00	CPT00	CPT01	CPT01	CPT00	CPT00	CPT01	CPT01	CPT00
Smoking.Status	smoke	smoke	smoke	smoke	non-s	non-s	smoke	smoke	non-s	non-s	smoke	smoke	non-s	non-s	smoke	smoke	non-s	non-s	smoke
Stage	2A	2A	1A	1A	1A	1A					2B	2B			3A	3A			1B
Region.of.Origin	weste	weste	weste	weste	asian	asian	asian	asian	asian	asian	weste	weste	asian	asian	asian	asian	asian	asian	weste
Country.of.Origin	usa	usa	usa	usa	vietna	vietna	china	china	china	china	usa	usa	vietna	vietna	vietna	vietna	china	china	other
Age	81.00	81.00	58.00	58.00	48.00	48.00	70.00	70.00	45.00	45.00	79.00	79.00	64.00	64.00	48.00	48.00	59.00	59.00	71.00
Gender	male	male	femal	femal	femal	femal	male	male	male	male	male	male	male	male	male	male	femal	femal	male
Ethnicity	cauca	cauca	cauca	cauca	asian	asian	han	han	han	han	cauca	cauca	asian	asian	asian	asian	han	han	hispa
Height.cm	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####
Weight.kg	82.00	82.00	52.90	52.90	44.00	44.00	79.00	79.00	60.00	60.00	80.46	80.46	48.00	48.00	64.00	64.00	50.00	50.00	86.82
BMI	27.37	27.37	22.60	22.60	18.08	18.08	27.99	27.99	20.05	20.05	22.77	22.77	18.75	18.75	22.15	22.15	21.93	21.93	29.10
Cigarettes.per.Day	5.00	5.00	20.00	20.00	NA	NA	40.00	40.00	NA	NA	10.00	10.00	NA	NA	5.00	5.00	NA	NA	20.00
Pack.Years.Smoked	0.30	5.00	33.00	33.00	NA	NA	#####	#####	NA	NA	27.00	27.00	NA	NA	6.30	6.30	NA	NA	40.00
Smoking.History	Currer	Currer	Currer	Currer	Lifelo	Lifelo	Currer	Currer	Lifelo	Lifelo	Currer	Currer	Lifelo	Lifelo	Currer	Currer	Lifelo	Lifelo	Currer
Secondhand.Smoke		No or	Expos	Expos	No or	No or	Expos	Expos	Expos	Expos	Expos	Expos	No or	No or	No or	No or	Expos	Expos	Expos
QC.status	QC.pa	QC.pa	QC.pa	QC.pa	QC.pa	QC.pa	QC.pa	QC.pa	QC.pa	QC.pa	QC.pa	QC.pa	QC.pa	QC.pa	QC.pa	QC.pa	QC.pa	QC.pa	QC.pa
QC.status.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
id	C3N.01	C3N.01	C3L.01	C3L.01	C3N.00	C3N.00	C3N.02	C3N.02	C3N.02	C3N.02	C3L.00	C3L.00	C3N.01	C3N.01	C3N.00	C3N.00	C3N.02	C3N.02	C3L.00
normalization	2comp	2comp	2comp	2comp	2comp	2comp	2comp	2comp	2comp	2comp	2comp	2comp	2comp	2comp	2comp	2comp	2comp	2comp	2comp
normalization.center	0.60	-0.57	-0.11	-0.30	0.03	-0.27	0.28	-0.56	0.57	-0.37	0.32	-0.94	0.03	-0.73	0.17	-0.16	0.40	-0.27	0.20
normalization.scale	0.31	0.29	0.28	0.26	0.21	0.26	0.23	0.26	0.37	0.28	0.29	0.33	0.41	0.35	0.29	0.27	0.33	0.34	0.30
NP_001317422.1	-2.45	2.09	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1.03	2.13	3.05	0.96	1.87	0.13
NP_001106273.1	-1.27	-0.12	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.83	-0.96	1.33	0.19	0.27	0.02
NP_620156.1	-0.64	-0.53	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-0.55	0.74	1.47	0.44	0.72	4.18
NP_001446.1	0.19	1.17	-0.68	0.88	-0.22	1.34	-0.35	1.46	-1.19	1.61	0.90	0.86	-0.62	1.23	-0.25	0.57	-0.63	1.22	0.72
NP_002006.2	0.29	0.79	-0.45	0.21	0.15	2.48	-0.05	1.49	-0.14	2.61	0.30	0.26	-0.77	-0.04	-1.15	1.10	-1.59	2.18	-0.40
NP_005929.2	-0.41	0.19	-1.30	0.47	-1.02	2.75	-1.74	1.74	1.29	2.84	-0.89	-0.16	-1.24	-1.88	-1.59	2.43	-1.57	2.09	-0.25
NP_001278210.2	1.42	1.00	-0.97	0.53	0.71	1.37	-0.25	1.11	-0.90	2.78	1.88	0.26	-0.77	-1.96	-2.02	0.38	-1.19	1.93	0.21
NP_057250.1	0.33	0.64	0.50	0.87	0.08	0.89	-0.34	0.77	0.56	1.16	0.74	0.60	-1.13	1.59	-0.08	0.85	-0.23	0.76	0.67
NP_006090.2	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.00	#####	1.65
NP_001243764.1	0.12	-0.29	0.42	-0.45	0.44	-0.36	0.82	-0.35	0.59	-0.33	-0.16	0.27	0.34	0.11	0.41	-0.51	-0.23	-0.39	0.07
NP_001243463.1	0.87	-1.26	1.24	0.29	0.12	-1.44	-0.91	-1.49	3.86	-0.92	-2.61	-2.94	0.43	-0.23	-1.59	-1.15	-0.17	-0.98	-1.50
NP_001274316.1	0.95	-1.26	1.51	0.40	0.11	-1.33	-0.91	-0.91	4.73	-1.11	-2.61	-3.07	0.16	-0.23	-1.59	-1.47	0.02	-1.16	-1.72
NP_006811.2	-5.89	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-0.85	-2.45	0.08	0.29	0.19	-2.12
NP_003013.1	-0.17	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.46	1.16	1.64	1.21	1.20	0.50
NP_116193.2	-0.36	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.20	0.64	1.18	0.40	0.96	0.27

Protein group with multiple genes

Protein group with isoforms of 1 gene

Link from Group Number in Protein/Peptide Summary HTML Results Highlights Identified Peptides

Subgroup-specific Peptides

1

2

3

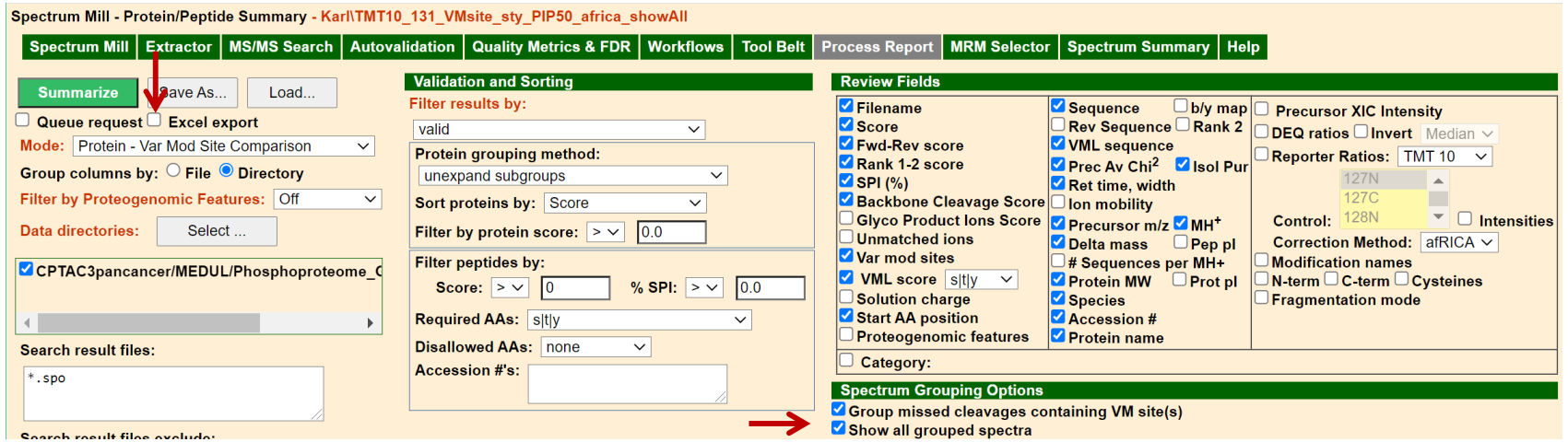
ID	Sub Group #	Length	Identical AA's	%ID	Species	Protein Name
54696354	1	327	327	100.0	Homo sapiens	protein phosphatase 1, catalytic subunit, beta isoform
46249376	1	327	326	99.7	Homo sapiens	protein phosphatase 1, catalytic subunit, beta isoform 1
45827798	2	286	254	88.8	Homo sapiens	protein phosphatase 1, catalytic subunit, alpha isoform 2
4506003	2	330	291	88.2	Homo sapiens	protein phosphatase 1, catalytic subunit, alpha isoform 1
56790945	2	341	290	85.0	Homo sapiens	protein phosphatase 1, catalytic subunit, alpha isoform 3
5668560	3	294	269	91.5	Homo sapiens	serine/threonine phosphatase 1 gamma
4506007	3	323	285	88.2	Homo sapiens	protein phosphatase 1, catalytic subunit, gamma isoform

```
54696354  (1)  MADG-ELNVDSLITRLLEVR-----------GCRPGKIVQMTEAEVRGLCIKSREIFLSQPIPLELEAPLKICGDIHGQYTDLLRLFEYGGFPPEANYLF
46249376  (1)  MADG-ELNVDSLITRLLEVR-----------GCRPGKIVQMTEAEVRGLCIKSREIFLSQPILLELEAPLKICGDIHGQYTDLLRLFEYGGFPPEANYLF
45827798  (2)  MSDSEKLNLDSIIGRLLEG-------------------------------------------------------------DIHGQYYDLLRLFEYGGFPPESNYLF
 4506003  (2)  MSDSEKLNLDSIIGRLLEVQ-----------GSRPGKNVQLTENEIRGLCLKSREIFLSQPILLELEAPLKICGDIHGQYYDLLRLFEYGGFPPESNYLF
56790945  (2)  MSDSEKLNLDSIIGRLLEGSRVLTPHCAPVQGSRPGKNVQLTENEIRGLCLKSREIFLSQPILLELEAPLKICGDIHGQYYDLLRLFEYGGFPPESNYLF
 5668560  (3)  MADLDKLNIDSIIQRLLEVR-----------GSKPGKNVQLQENEIRGLCLKSREIFLSQPILLELEAPLKICGDIHGQYYDLLRLFEYGGFPPESNYLF
 4506007  (3)  MADLDKLNIDSIIQRLLEVR-----------GSKPGKNVQLQENEIRGLCLKSREIFLSQPILLELEAPLKICGDIHGQYYDLLRLFEYGGFPPESNYLF

54696354  (1)  LGDYVDRGKQSLETICLLLAYKIKYPENFFLLRGNHECASINRIYGFYDECKRRFNIKLWKTFTDCFNCLPIAAIVDEKIFCCHGGLSPDLQSMEQIRRI
46249376  (1)  LGDYVDRGKQSLETICLLLAYKIKYPENFFLLRGNHECASINRIYGFYDECKRRFNIKLWKTFTDCFNCLPIAAIVDEKIFCCHGGLSPDLQSMEQIRRI
45827798  (2)  LGDYVDRGKQSLETICLLLAYKIKYPENFFLLRGNHECASINRIYGFYDECKRRYNIKLWKTFTDCFNCLPIAAIVDEKIFCCHGGLSPDLQSMEQIRRI
 4506003  (2)  LGDYVDRGKQSLETICLLLAYKIKYPENFFLLRGNHECASINRIYGFYDECKRRYNIKLWKTFTDCFNCLPIAAIVDEKIFCCHGGLSPDLQSMEQIRRI
56790945  (2)  LGDYVDRGKQSLETICLLLAYKIKYPENFFLLRGNHECASINRIYGFYDECKRRYNIKLWKTFTDCFNCLPIAAIVDEKIFCCHGGLSPDLQSMEQIRRI
 5668560  (3)  LGDYVDRGKQSLETICLLLAYKIKYPENFFLLRGNHECASINRIYGFYDECKRRYNIKLWKTFTDCFNCLPIAAIVDEKIFCCHGGLSPDLQSMEQIRRI
 4506007  (3)  LGDYVDRGKQSLETICLLLAYKIKYPENFFLLRGNHECASINRIYGFYDECKRRYNIKLWKTFTDCFNCLPIAAIVDEKIFCCHGGLSPDLQSMEQIRRI

54696354  (1)  MRPTDVPDTGLLCDLLWSDPDKDVQGWGENDRGVSFTFGADVVSKFLNRHDLDLICRAHQVVEDGYEFFAKRQLVTLFSAPNYCGEFDNAGGMMSVDETL
46249376  (1)  MRPTDVPDTGLLCDLLWSDPDKDVQGWGENDRGVSFTFGADVVSKFLNRHDLDLICRAHQVVEDGYEFFAKRQLVTLFSAPNYCGEFDNAGGMMSVDETL
45827798  (2)  MRPTDVPDQGLLCDLLWSDPDKDVQGWGENDRGVSFTFGAEVVAKFLHKHDLDLICRAHQVVEDGYEFFAKRQLVTLFSAPNYCGEFDNAGAMMSVDETL
 4506003  (2)  MRPTDVPDQGLLCDLLWSDPDKDVQGWGENDRGVSFTFGAEVVAKFLHKHDLDLICRAHQVVEDGYEFFAKRQLVTLFSAPNYCGEFDNAGAMMSVDETL
56790945  (2)  MRPTDVPDQGLLCDLLWSDPDKDVQGWGENDRGVSFTFGAEVVAKFLHKHDLDLICRAHQVVEDGYEFFAKRQLVTLFSAPNYCGEFDNAGAMMSVDETL
 5668560  (3)  MRPTDVPDQGLLCDLLWSDPDKDVLGWGENDRGVSFTFGAEVVAKFLHKHDLDLICRAHQVVEDGYEFFAKRQLVTLFSAPNYCGEFDNAGAMMSVDETL
 4506007  (3)  MRPTDVPDQGLLCDLLWSDPDKDVLGWGENDRGVSFTFGAEVVAKFLHKHDLDLICRAHQVVEDGYEFFAKRQLVTLFSAPNYCGEFDNAGAMMSVDETL

54696354  (1)  MCSFQILKPSEK-KAKY-QYGGLNS-GRPVTPPRT--ANPPKKR
46249376  (1)  MCSFQILKPSEK-KAKY-QYGGLNS-GRPVTPPRT--ANPPKKR
45827798  (2)  MCSFQILKPADKNKGKYGQFSGLNPGGRPITPPRN--SAKAKK-
 4506003  (2)  MCSFQILKPADKNKGKYGQFSGLNPGGRPITPPRN--SAKAKK-
56790945  (2)  MCSFQILKPADKNKGKYGQFSGLNPGGRPITPPRN--SAKAKK-
 5668560  (3)  MCSFQ-------------------------------------
 4506007  (3)  MCSFQILKPAEKKKPNA---------TRPVTPPRGMITKQAKK-
```

ClustalW aligns proteins in group SM colors identified peptides and labels subgroups

Collapsing PSM's to Phosphosite level – Show site-grouped PSMs - HTML

m/z	ppm	ID Score	Filename	VML score	#Loc sty sites	#Amb sty sites	iTRAQ 114/117	iTRAQ 115/117	iTRAQ 116/117	PIP (%)	Sequence	VML sequence	Gene name	PO4 Site
552.960	5.1	13.83	050.10050.3	0.00	0	1	0.96	0.08	1.27	76.7	RSStLSQLPGDK	RS(0.33)S(0.33)T(0.33)LS(0.0)QLPGDK	SPAG9	T581t
643.689	-0.1	10.84	7464.7464.3	0.00	0	1	0.93	0.06	1.45	94.5	KRSsTLSQLPGDK	KRS(0.25)S(0.25)T(0.25)LS(0.25)QLPGDK		
500.923	-1.4	8.06	558.11558.3	0.00	0	1	1.48	0.31	0.88	78.7	SsTLSQLPGDK	S(0.25)S(0.25)T(0.25)LS(0.25)QLPGDK		
750.882	0.8	7.23	174.11174.2	0.00	0	1	1.84	0.16	0.99	87.1	SStLSQLPGDK	S(0.0)S(0.50)T(0.50)LS(0.0)QLPGDK		
750.881	-0.2	7.09	572.11572.2	0.00	0	1	1.96	0.14	0.71	62.8	SStLSQLPGDK	S(0.0)S(0.50)T(0.50)LS(0.0)QLPGDK		
678.833	3.6	7.07	720.11720.2	1.04	0	1	0.99	0.10	0.96	82.9	SStLSQLPGDK	S(0.33)S(0.33)T(0.33)LS(0.0)QLPGDK		
551.943	0.1	10.81	410.20410.3	4.43	1	0	0.40	0.70	1.84	75.1	DGGsVVGASVFYK	DGGS(0.99)VVGAS(0.0)VFY(0.0)K	SPAG9	S691s
755.361	2.1	7.91	717.20717.2	7.91	1	0	1.03	0.89	1.60	94.0	DGGsVVGASVFYK	DGGS(0.99)VVGAS(0.0)VFY(0.0)K		
827.408	-2.0	7.73	423.20423.2	7.73	1	0	0.57	0.90	1.63	87.7	DGGsVVGASVFYK	DGGS(0.99)VVGAS(0.0)VFY(0.0)K		
721.043	0.7	16.86	757.16757.3	2.29	1	0	0.48	0.98	1.77	69.8	SAsQSSLDKLDQELK	S(0.0)AS(0.99)QS(0.0)S(0.0)LDKLDQELK	SPAG9	S716s
645.813	1.0	12.89	4977.4977.2	2.66	1	0	0.74	0.08	1.45	91.6	SAsQSSLDK	S(0.0)AS(0.99)QS(0.0)S(0.0)LDK		
541.035	2.5	12.36	777.16777.4	0.00	0	1	0.42	1.08	1.74	74.7	SASQsSLDKLDQELK	S(0.0)AS(0.50)QS(0.50)S(0.0)LDKLDQELK		
1081.057	-2.6	13.71	073.15073.2	1.21	1	0	0.44	0.76	1.56	87.9	SASQSsLDKLDQELK	S(0.0)AS(0.0)QS(0.0)S(0.99)LDKLDQELK	SPAG9	S719s
541.032	-1.7	8.73	034.15154.4	0.00	0	1	0.48	0.74	1.56	75.6	SASQSsLDKLDQELK	S(0.0)AS(0.0)QS(0.50)S(0.50)LDKLDQELK		
721.040	-2.5	7.37	039.15039.3	0.00	0	1	0.35	0.75	1.52	72.3	SASQSsLDKLDQELK	S(0.0)AS(0.0)QS(0.50)S(0.50)LDKLDQELK		
430.878	0.6	4.40	4593.4593.3	0.00	0	1	0.65	0.06	1.52	69.7	SASQSsLDK	S(0.0)AS(0.0)QS(0.50)S(0.50)LDK		



Combine repeat site-level observations with non-conflicting localizations

- Different precursor charge states
- Different sample-handling modifications
- Different trypsin missed cleavage states

Representative peptide: if confident localization, higher ID scores and longer peptide preferred, else best VML score.

Collapsing PSM's to Phosphosite level – Show site-grouped PSMs - SSV

PSMexportVMsitesPolished.1.ssv

number	filename	parent_charge	score	deltaForwardReverseScore	backbone_cleavage_score	missedCleavages	numKorR	scoreVML	numPossibleVMsites_STY	numActualVMSites_sty	numLocalizedVMsites_sty	numAmbiguousVMsites_sty	sequence	sequenceVML	accession_number
6442	05CPTAC_LUAD_	4	5.51	5.51	4	1	3	0	4	1	0	1	KKESsSmLATVK	KKES(0.33)S(0.33)S(0.33)M(1.0)LAT(0.0)VK	NP_001123917.1
6442	12CPTAC_LUAD_	3	5.27	3.72	5	1	2	0	4	1	0	1	KESsSmLATVK	KES(0.33)S(0.33)S(0.33)M(1.0)LAT(0.0)VK	NP_001123917.1
6442	09CPTAC_LUAD_	3	4.92	4.92	5	1	2	0	4	1	0	1	KESsSMLATVK	KES(0.33)S(0.33)S(0.33)MLAT(0.0)VK	NP_001123917.1
6442	04CPTAC_LUAD_	3	4.71	4.71	5	1	2	0	4	1	0	1	KESsSmLATVK	KES(0.33)S(0.33)S(0.33)M(1.0)LAT(0.0)VK	NP_001123917.1
6442	04CPTAC_LUAD_	2	4.27	4.27	6	0	1	0	4	1	0	1	ESSsmLATVK	ES(0.0)S(0.50)S(0.50)M(1.0)LAT(0.0)VK	NP_001123917.1
6443	20CPTAC_LUAD_	2	16.95	15.07	11	0	1	1.127	5	1	1	0	EEVSGsSAAVTENADSDR	EEVS(0.0)GS(0.99)S(0.0)AAVT(0.0)ENADS(0.0)DR	NP_001123917.1
6443	10CPTAC_LUAD_	2	9.55	9.55	9	0	1	0	5	1	0	1	EEVSGSsAAVTENADSDR	EEVS(0.33)GS(0.33)S(0.33)AAVT(0.0)ENADS(0.0)D	NP_001123917.1
6444	10CPTAC_LUAD_	4	21.91	21.91	17	1	2	0	9	1	0	1	EEVSGSSAAVTENADSDRI	EEVS(0.0)GS(0.0)S(0.0)AAVT(0.0)ENADS(0.0)DRIS	NP_001123917.1
6444	24CPTAC_LUAD_	4	18.74	22.07	16	1	2	0	9	1	0	1	EEVSGSSAAVTENADSDRI	EEVS(0.0)GS(0.0)S(0.0)AAVT(0.0)ENADS(0.0)DRIS	NP_001123917.1
6444	25CPTAC_LUAD_	4	12.42	12.42	12	1	2	0	9	1	0	1	EEVSGSSAAVTENADSDRI	EEVS(0.0)GS(0.0)S(0.0)AAVT(0.0)ENADS(0.0)DRIS	NP_001123917.1
6445	02CPTAC_LUAD_	3	12.26	12.26	13	0	1	0	6	1	0	1	LDSDFNIsSHSELENSSELK	LDS(0.0)DFNIS(0.50)S(0.50)HS(0.0)ELENS(0.0)S(0.	NP_001123917.1
6445	10CPTAC_LUAD_	3	10.67	10.67	11	0	1	0	6	1	0	1	LDSDFNIsSHSELENSSELK	LDS(0.0)DFNIS(0.33)S(0.33)HS(0.33)ELENS(0.0)S(	NP_001123917.1
6445	09CPTAC_LUAD_	3	9.17	9.17	11	0	1	0	6	1	0	1	LDSDFNISsHSELENSSELK	LDS(0.0)DFNIS(0.50)S(0.50)HS(0.0)ELENS(0.0)S(0.	NP_001123917.1
6445	05CPTAC_LUAD_	3	8.79	8.79	11	0	1	0	6	1	0	1	LDSDFNIsSHSELENSSELK	LDS(0.0)DFNIS(0.33)S(0.33)HS(0.33)ELENS(0.0)S(	NP_001123917.1
6445	14CPTAC_LUAD_	3	8.05	8.05	12	0	1	0	6	1	0	1	LDSDFNISsHSELENSSELK	LDS(0.0)DFNIS(0.50)S(0.50)HS(0.0)ELENS(0.0)S(0.	NP_001123917.1
6445	09CPTAC_LUAD_	3	7.46	7.46	9	0	1	0	6	1	0	1	LDSDFNIsSHSELENSSELK	LDS(0.0)DFNIS(0.33)S(0.33)HS(0.33)ELENS(0.0)S(	NP_001123917.1
6446	15CPTAC_LUAD_	3	11.45	8.48	6	0	1	5.573	3	1	1	0	sVHISTPEK	S(0.99)VHIS(0.0)T(0.0)PEK	NP_001123917.1
6446	10CPTAC_LUAD_	3	10.19	10.19	6	0	1	4.53	3	1	1	0	sVHISTPEK	S(0.99)VHIS(0.0)T(0.0)PEK	NP_001123917.1
6446	07CPTAC_LUAD_	3	6.34	6.34	6	0	1	5.149	3	1	1	0	sVHISTPEK	S(0.99)VHIS(0.0)T(0.0)PEK	NP_001123917.1
6447	09CPTAC_LUAD_	4	20.97	20.97	15	1	2	1.15	5	1	1	0	SVHIStPEKEPCAPLTIPSIR	S(0.0)VHIS(0.0)T(0.99)PEKEPCAPLT(0.0)IPS(0.0)IR	NP_001123917.1
6447	12CPTAC_LUAD_	4	20.31	18.94	15	1	2	1.135	5	1	1	0	SVHIStPEKEPCAPLTIPSIR	S(0.0)VHIS(0.0)T(0.99)PEKEPCAPLT(0.0)IPS(0.0)IR	NP_001123917.1
6447	03CPTAC_LUAD_	4	20.2	20.2	16	1	2	1.157	5	1	1	0	SVHIStPEKEPCAPLTIPSIR	S(0.0)VHIS(0.0)T(0.99)PEKEPCAPLT(0.0)IPS(0.0)IR	NP_001123917.1
6447	08CPTAC_LUAD_	4	19.99	19.99	14	1	2	1.177	5	1	1	0	SVHIStPEKEPCAPLTIPSIR	S(0.0)VHIS(0.0)T(0.99)PEKEPCAPLT(0.0)IPS(0.0)IR	NP_001123917.1
6447	10CPTAC_LUAD_	4	19.43	19.43	15	1	2	1.184	5	1	1	0	SVHIStPEKEPCAPLTIPSIR	S(0.0)VHIS(0.0)T(0.99)PEKEPCAPLT(0.0)IPS(0.0)IR	NP_001123917.1
6447	16CPTAC_LUAD_	4	19.02	19.02	15	1	2	1.221	5	1	1	0	SVHIStPEKEPCAPLTIPSIR	S(0.0)VHIS(0.0)T(0.99)PEKEPCAPLT(0.0)IPS(0.0)IR	NP_001123917.1
6447	14CPTAC_LUAD_	3	13.06	14.67	8	0	1	1.341	3	1	1	0	SVHIStPEK	S(0.0)VHIS(0.0)T(0.99)PEK	NP_001123917.1
6447	10CPTAC_LUAD_	3	12.01	12	8	0	1	1.253	3	1	1	0	SVHIStPEK	S(0.0)VHIS(0.0)T(0.99)PEK	NP_001123917.1
6447	17CPTAC_LUAD_	3	11.53	12.31	7	0	1	1.406	3	1	1	0	SVHIStPEK	S(0.0)VHIS(0.0)T(0.99)PEK	NP_001123917.1
6447	14CPTAC_LUAD_	4	20.65	20.65	14	1	2	0	5	1	0	1	SVHIStPEKEPCAPLTIPSIR	S(0.0)VHIS(0.50)T(0.50)PEKEPCAPLT(0.0)IPS(0.0)IF	NP_001123917.1
6447	04CPTAC_LUAD_	4	19.99	19.99	15	1	2	0	5	1	0	1	SVHIStPEKEPCAPLTIPSIR	S(0.0)VHIS(0.50)T(0.50)PEKEPCAPLT(0.0)IPS(0.0)IF	NP_001123917.1
6447	15CPTAC_LUAD_	4	17.04	14.84	14	1	2	0	5	1	0	1	SVHIStPEKEPCAPLTIPSIR	S(0.0)VHIS(0.50)T(0.50)PEKEPCAPLT(0.0)IPS(0.0)IF	NP_001123917.1
6447	13CPTAC_LUAD_	4	15.89	15.89	12	1	2	0	5	1	0	1	SVHIStPEKEPCAPLTIPSIR	S(0.0)VHIS(0.50)T(0.50)PEKEPCAPLT(0.0)IPS(0.0)IF	NP_001123917.1
6447	05CPTAC_LUAD_	4	15	15	13	1	2	0	5	1	0	1	SVHIsTPEKEPCAPLTIPSIR	S(0.0)VHIS(0.50)T(0.50)PEKEPCAPLT(0.0)IPS(0.0)IF	NP_001123917.1
6448	19CPTAC_LUAD_	4	26.17	24.25	15	0	1	11.191	6	1	1	0	DsFEMEEVQSTEGEAPHVF	DS(0.99)FEMEEVQS(0.0)T(0.0)EGEAPHVPAT(0.0)Y	NP_001123917.1
6448	19CPTAC_LUAD_	5	25.78	25.41	20	1	2	3.291	7	1	1	0	NIMTQQKDsFEMEEVQSTI	NIMT(0.0)QQKDS(0.99)FEMEEVQS(0.0)T(0.0)EGEA	NP_001123917.1
6448	15CPTAC_LUAD_	4	25.72	23.78	15	0	1	10.052	6	1	1	0	DsFEMEEVQSTEGEAPHVF	DS(0.99)FEMEEVQS(0.0)T(0.0)EGEAPHVPAT(0.0)Y	NP_001123917.1
6448	12CPTAC_LUAD_	4	25.6	25.49	18	1	2	2.259	7	1	1	0	NIMTQQKDsFEMEEVQSTI	NIMT(0.0)QQKDS(0.99)FEMEEVQS(0.0)T(0.0)EGEA	NP_001123917.1

Log file created by

Autovalidation
VM site Polishing

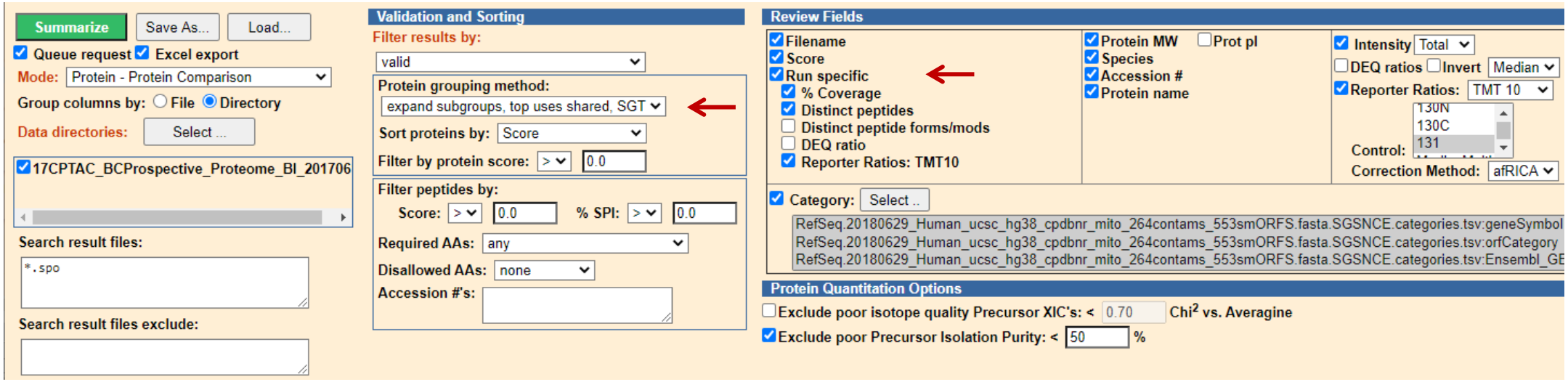
Combine repeat site-level observations with non-conflicting localizations

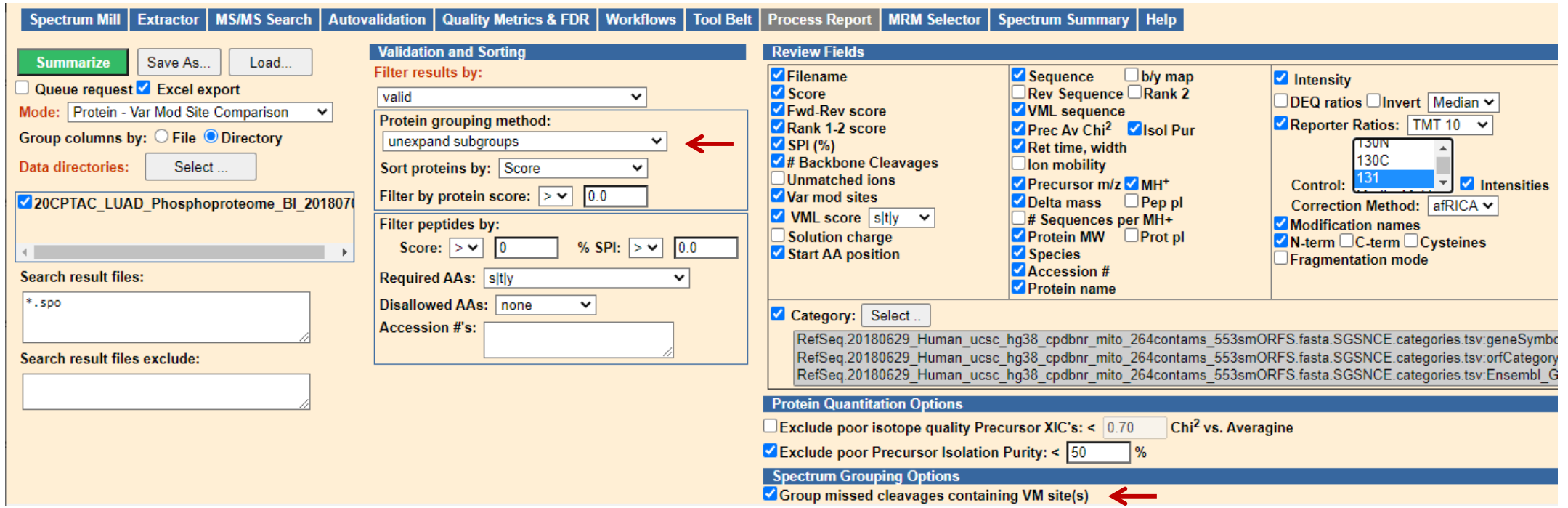
- Different precursor charge states
- Different sample-handling modifications
- Different trypsin missed cleavage states

Representative peptide: if confident localization, higher ID scores and longer peptide preferred, else best VML score.

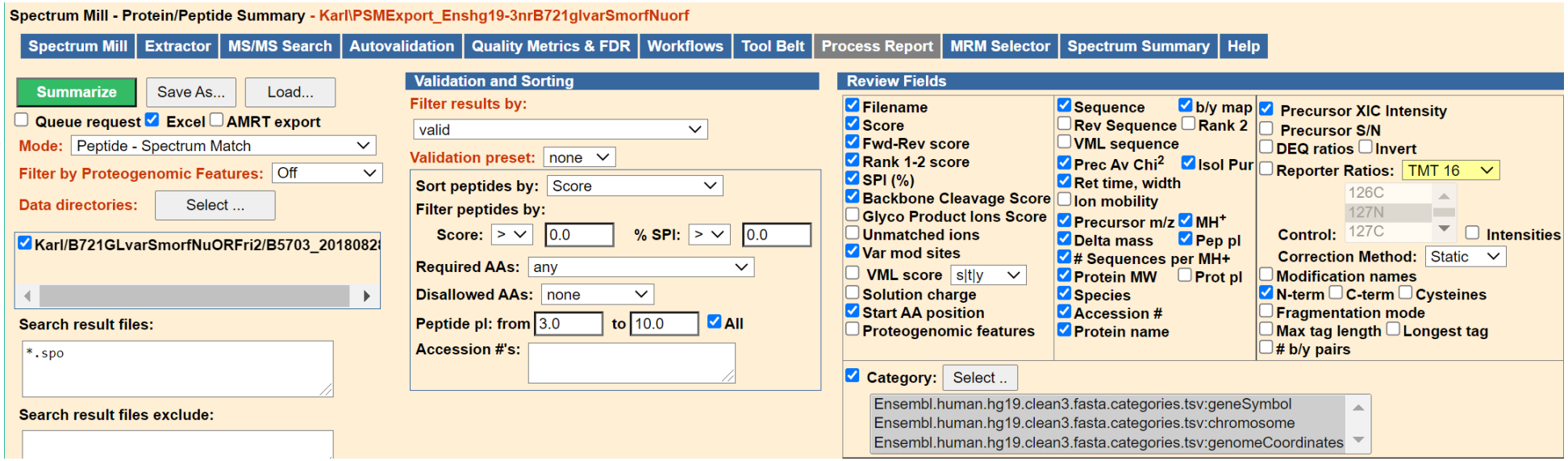
Sequence overlap complications:
Ambiguous localizations grouped with confident ones preferentially over other overlapping ambiguous localizations that are inconsistent with the confident localization.

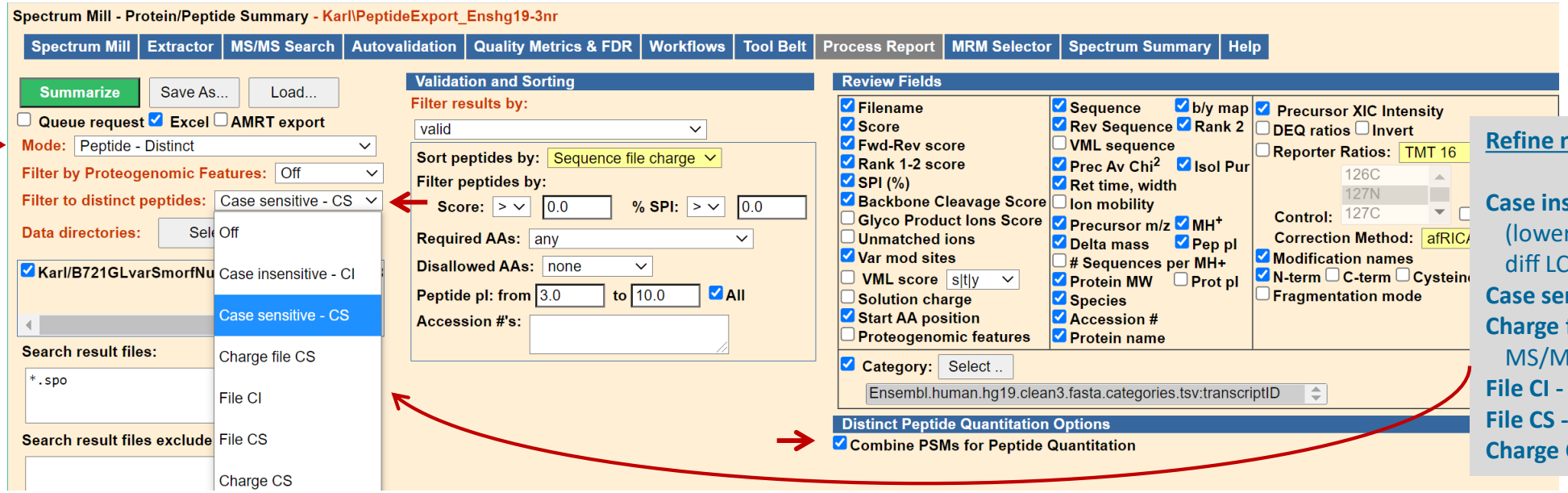
Protein / VM-site Summary Modes





PSM / Peptide Summary Modes





Refine notion of sameness for collapsing PSMs

Case insensitive CI - combine variable mods (lowercase AA's), no mods, diff precursor z, diff LC-MS/MS runs.

Case sensitive CS - separate mods & no mods

Charge file CS - separate diff precursor z, diff LC-MS/MS runs, diff mods.

File CI - separate diff LC-MS/MS runs

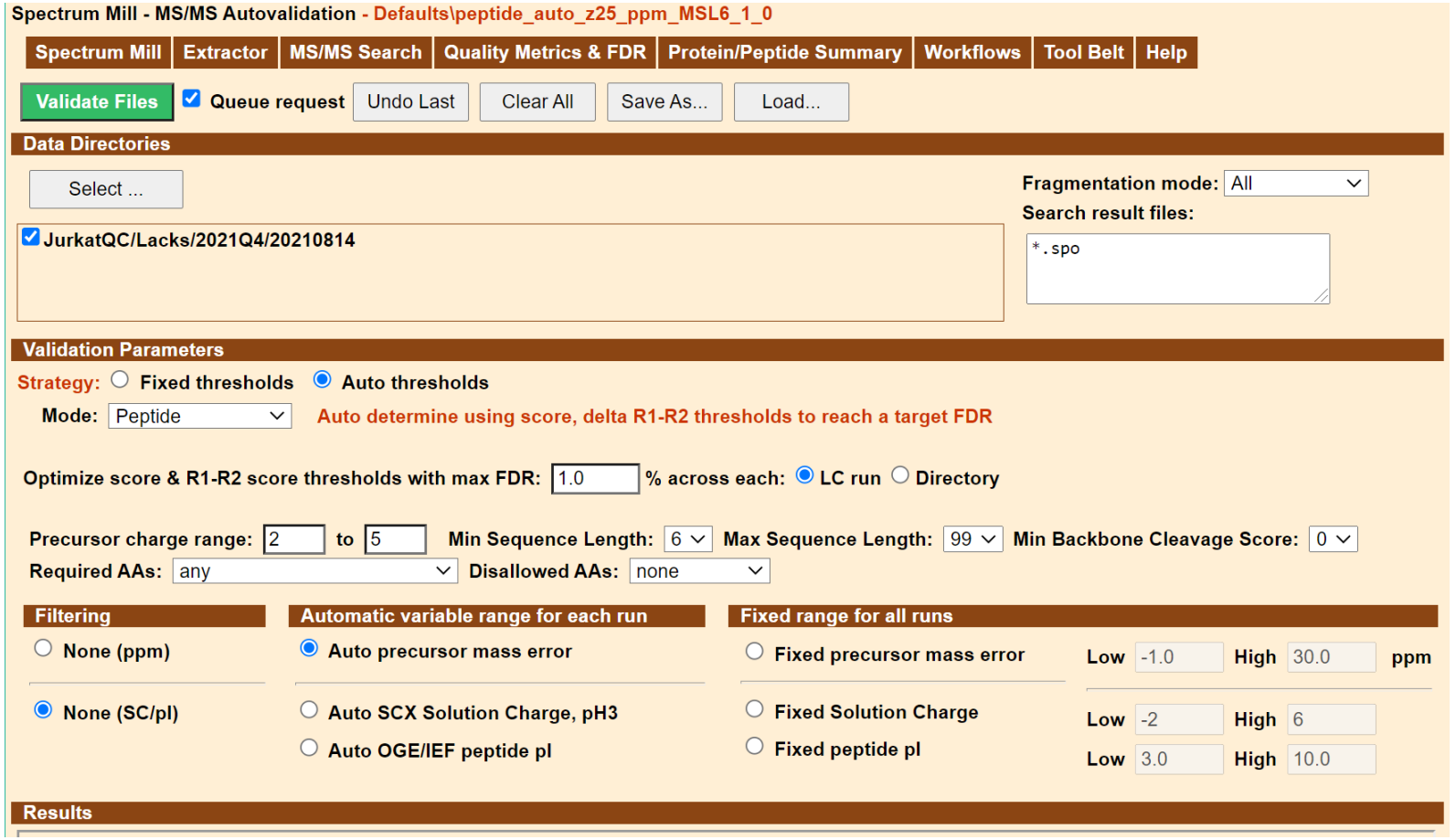
File CS - separate diff LC-MS/MS runs, diff mods

Charge CS - separate diff precursor z, diff mods

Autovalidation

Confident identification and False Discovery Rates (FDR) at the PSM, peptide, and protein levels.

Autovalidation – Peptide/PSM level

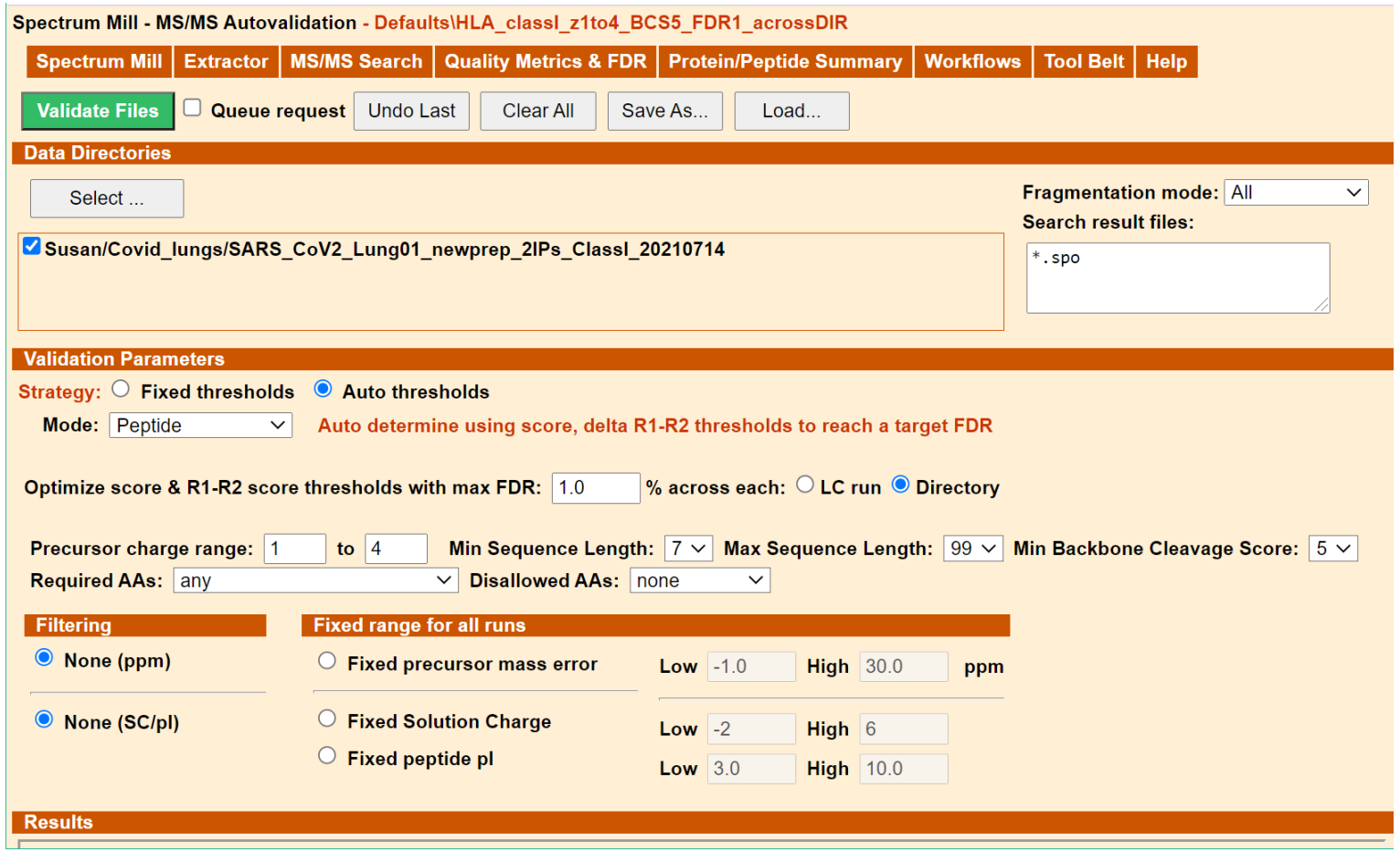


Separate Score thresholds for each:
- LC-MS/MS run
- Precursor charge

Separate precursor mass error range for each LC-MS/MS run
- Median +/- 3 std dev

Min Sequence length filter
- Short peptides often contribute to multiple proteins, inclusion may skew protein quantitation.
- Extractor MH+ lower limit: 750 for iTRAQ excludes short Arg peps
- Autovalidation MSL filter excludes short Lys & Arg peps

Autovalidation – Peptide/PSM level – immunopeptidomics extras



Separate Score thresholds for each:
- Precursor charge

Optimize thresholds by directory

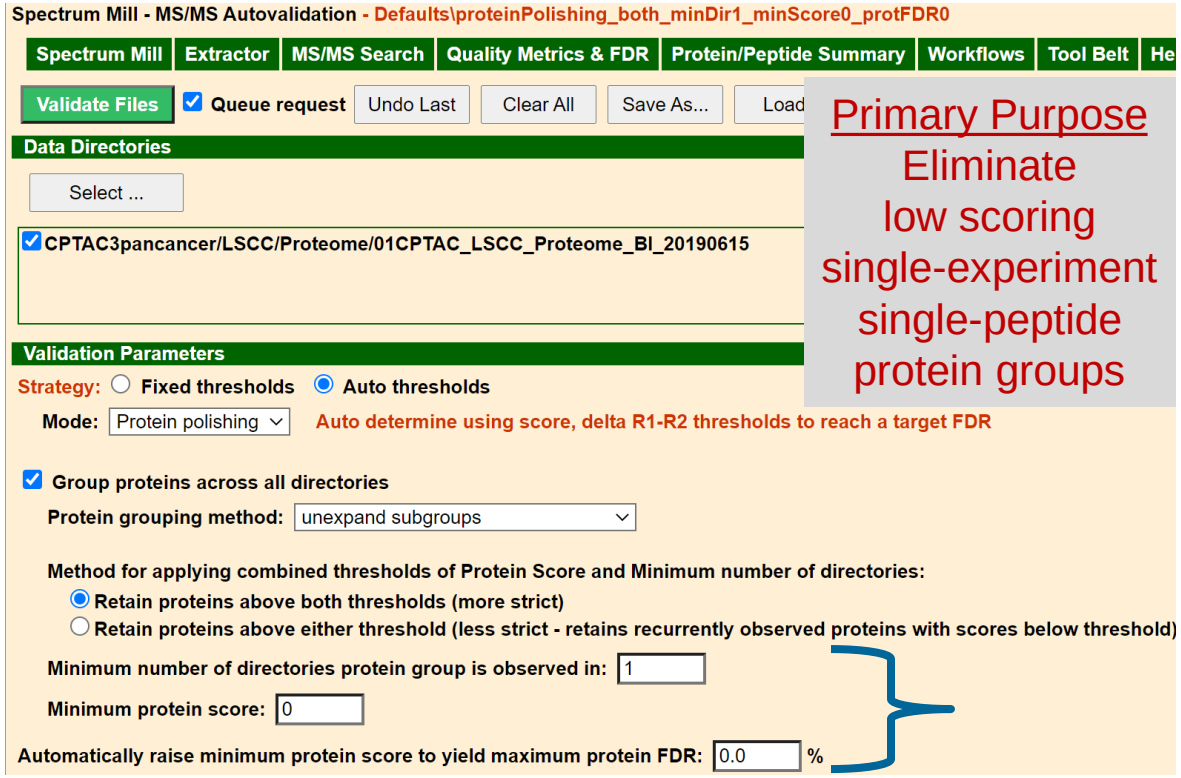
Useful for:
- Low abundance samples with <1,000 PSMs/LC run
- Low frequency matches in complex samples
  - PTM's in unenriched samples
  - Proteogenomic variants
  - High charge states (>4+)

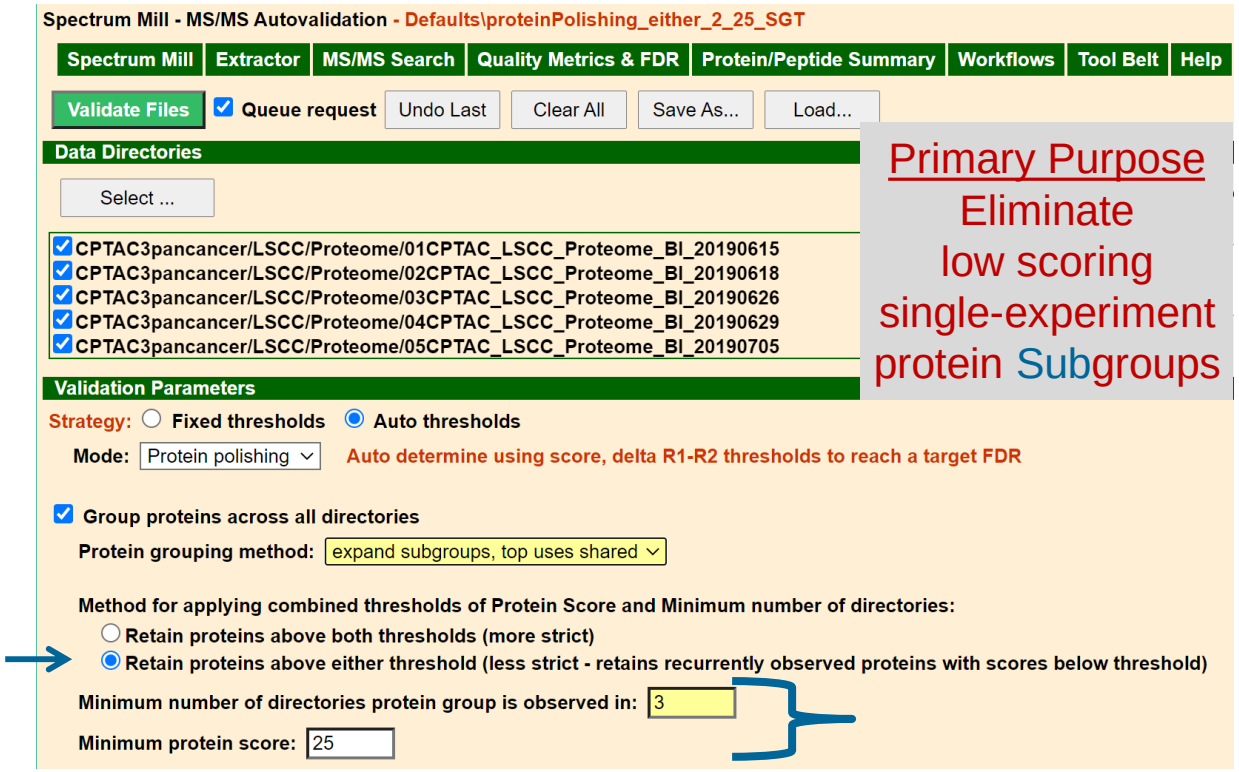
Backbone cleavage score

Enforces uniformly higher minimum sequence coverage for each PSM, at least 4 or 5 residues of unambiguous sequence.

Built for immunopeptidomics.

Autovalidation - Protein Polishing SGT





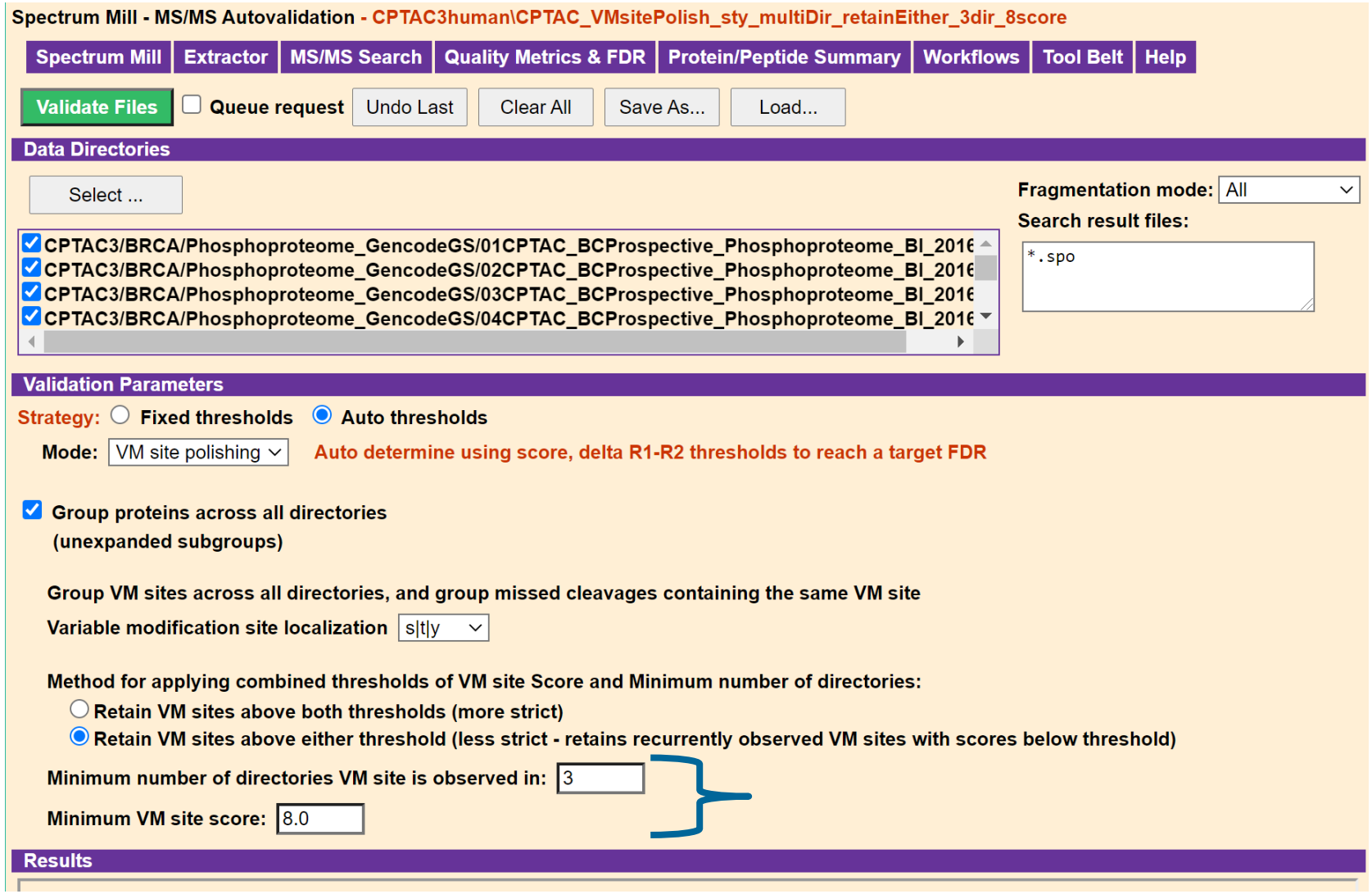
Assembles protein groups from the autovalidated PSM's, determines the maximum protein level score of a protein group that consists entirely of distinct peptides estimated to be false-positive identifications (PSM's with negative delta forward-reverse scores). Then PSM's are unvalidated if they contribute to protein groups that **BOTH**:

a) have protein scores at or below the larger of
   i. the minimum protein score
   ii. the max false-positive protein score
b) are observed in less than the minimum # of directories

From all directories, assembles protein subgroups from the autovalidated PSM's. Then PSM's are unvalidated if they contribute to protein subgroups that **EITHER**:

a) have protein subgroup-specific scores at or below the minimum protein score
   OR
b) are observed in less than the minimum # of directories

Autovalidation – VM site Polishing



Primary Purpose
Eliminate low-scoring, single-experiment phosphosites

From all directories, assembles VM sites from the autovalidated PSM's. Then PSM's are unvalidated if they contribute to VM sites that EITHER:

a) have VM site scores at or below the minimum VM site score
   OR
b) are observed in less than the minimum # of directories

Quantitation – reporter ion ratios (TMT, iTRAQ)

Key Features of Spectrum Mill Quantitation

- MS/MS reporter ion ratio based: iTRAQ, TMT
  - Median of all PSM ratios for each Protein, VM-site
    - Not sum of all PSM reporter ion intensities, then ratio for Protein, VM-site
  - Exclude ratios with Precursor Ion Purity, PIP < 50%
  - Exclude ½ of false positive ID's: Delta Fwd-Rev < 0
  - Exclude peptides with no label

- MS precursor ion ratio based: SILAC
  - Only 1 member of H/L pair H/M/L triplet needs to be triggered for MS/MS
  - Median of all H/L pair H/M/L triplet ratios for each Protein, VM-site
  - Exclude ratios from poor precursor ion isotope cluster shape

- Label – free
  - Only uses peptides identified by MS/MS
  - Sums up all precursor ion peak area for all PSM's for a protein

- Peptides shared between protein subgroups
  - Use shared in each subgroup
  - Use only unshared: Subgroup specific

Providing Reporter Ion Correction Factors

Spectrum Mill - Tool Belt Utilities

Spectrum Mill | Extractor | MS/MS Search | Autovalidation | Quality Metrics & FDR | Protein/Peptide S

Select a Tool:

- Stop process
- Create saved results file
- Create MS/MS search summary file
- View parameter table
- List modifications details
- Create Reporter Ion correction factors
- Apply Reporter Ion correction factors
- File collector
- Export PepXML
- Convert spectra

Create Reporter Ion correction factors

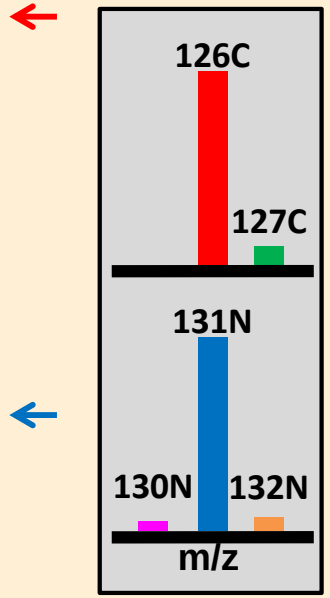
(iTRAQ 8-plex corrections do not vary by batch lot)

Create | TMT 16-plex

Batch: VA299611

Mass	% of -2	% of -1	% of +1	% of +2
126C	0.0	0.0	9.31	0.32
127N	0.0	0.0	8.98	0.32
127C	0.0	0.93	8.63	0.27
128N	0.0	0.95	8.38	0.26
128C	0.0	1.47	6.91	0.15
129N	0.0	1.46	6.86	0.15
129C	0.51	2.74	6.15	0.11
130N	0.49	2.76	5.98	0.11
130C	0.04	3.10	4.82	0.06
131N	0.04	3.09	4.75	0.06
131C	0.08	3.81	3.29	0.03
132N	0.04	2.84	3.51	0.02
132C	0.10	4.14	1.80	0.0
133N	0.36	3.64	1.94	0.0
133C	0.88	4.70	1.01	0.0
134N	0.4	4.92	1.05	0.0

Enter the ^{13}C values from the Certificate of Analysis



Updates file:
\\spectrumMill\msparams_mill\reporterIon.corrections.txt

Spectrum Mill - Tool Belt Utilities

Spectrum Mill | Extractor | MS/MS Search | Autovalidation | Quality Metrics & FDR | Protein/Peptid

Select a Tool:

- Stop process
- Create saved results file
- Create MS/MS search summary file
- View parameter table
- List modifications details
- Create Reporter Ion correction factors
- Apply Reporter Ion correction factors
- File collector
- Export PepXML
- Convert spectra

Apply Reporter Ion correction factors

Apply

Batch: TMT16_VA299611

Data Directory

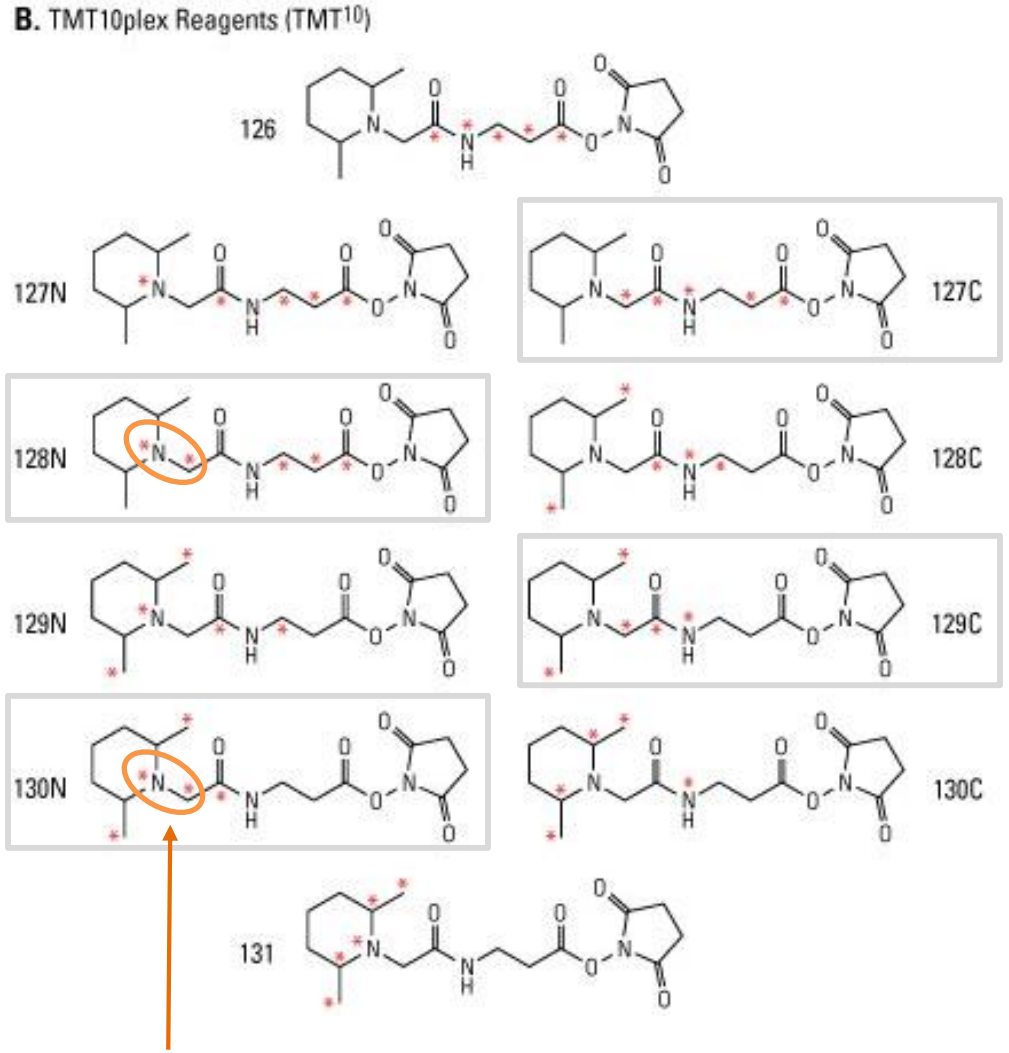
Select ... Hasmik/HIV_RAPMS_TMT16/HIV_RAPMS_TMT16_fxns

Batch: TMT16_VA299611				
126C	0.00	0.00	9.31	0.32
127N	0.00	0.00	8.98	0.32
127C	0.00	0.93	8.63	0.27
128N	0.00	0.95	8.38	0.26
128C	0.00	1.47	6.91	0.15
129N	0.00	1.46	6.86	0.15
129C	0.51	2.74	6.15	0.11
130N	0.49	2.76	5.98	0.11
130C	0.04	3.10	4.82	0.06
131N	0.04	3.09	4.75	0.06
131C	0.08	3.81	3.29	0.03
132N	0.04	2.84	3.51	0.02
132C	0.10	4.14	1.80	0.00
133N	0.36	3.64	1.94	0.00
133C	0.88	4.70	1.01	0.00
134N	0.40	4.92	1.05	0.00

Used in SM modules:
Quality Metrics
Protein/Peptide Summary

Creates file:
\\spectrumMill\msdataSM\Hasmik\HIV_RAPMS_TMT16\HIV_RAPMS_TMT16_All reporterIon.correction.txt

Source of Isotope Impurities



Probably most difficult combination to synthesize with high incorporation
>1% signal for 127C (in 128N) and for 129C (in 130N).

- +1 impurities originate from natural ^{13}C levels in the unlabeled C positions
  - C-reagents have only C-to-C +1.003 impurities
  - N-reagents have only N-to-N +1.003 impurities
- -1 impurities originate from incomplete incorporation of ^{15}N or ^{13}C in the synthetic precursors
  - C-reagents may have only C-to-C -1.003 impurities
  - N-reagents may have both:
    - N-to-C -0.997 impurities (^{15}N-^{14}N)
    - N-to-N -1.003 impurities (^{13}C-^{12}C)

TMT6/TMT10 Reporters			
126	+0		
127N	+1		^{15}N
127C	+1	^{13}C	
128N	+2	^{13}C	^{15}N
128C	+2	$^{13}C_2$	
129N	+3	$^{13}C_2$	^{15}N
129C	+3	$^{13}C_3$	
130N	+4	$^{13}C_3$	^{15}N
130C	+4	$^{13}C_4$	
131	+5	$^{13}C_4$	^{15}N

afRICA Reporter Ion Intensity Correction Method in SM

Shadforth IP, Dunkley TPJ, Lilley KS, and Bessant C
BMC Genomics 2005, 6:145 doi:10.1186/1471-2164-6-145

BMC Genomics

BioMed Central

Software

Open Access

i-Tracker: For quantitative proteomics using iTRAQ™

Ian P Shadforth*[1], Tom PJ Dunkley[2], Kathryn S Lilley[2] and Conrad Bessant[1]

Address: [1]Department of Analytical Science and Informatics, Cranfield University at Silsoe, Silsoe, Bedfordshire, UK and [2]Cambridge Centre for Proteomics, Biochemistry Department, Cambridge University, Cambridgeshire, UK

Email: Ian P Shadforth* - i.p.shadforth.s01@cranfield.ac.uk; Tom PJ Dunkley - tpjd2@cam.ac.uk; Kathryn S Lilley - k.s.lilley@bioc.cam.ac.uk; Conrad Bessant - c.bessant@cranfield.ac.uk

* Corresponding author

Purity correction

Each batch of iTRAQ reagents supplied by ABI is labelled with sixteen purity values indicating the percentages of each reporter ion that have masses differing by -2, -1, +1 and +2 Da from the nominal reporter ion mass due to isotopic variants. This information can be used to correct the peak areas calculated for each reporter ion to account for the losses to, and gains from, other reporter ions. Losses to ion peaks not in the reporter ion range are also accounted for in this method.

The simultaneous equations needed to solve this problem are fairly complicated, but can be framed such that Cramer's rule may be applied. A detailed explanation of

a,b,c,d,e,f,g,h,i,j,k,l,m,n,o,p are the sixteen purity correction values (as percentages) in the order:

114.1 – 2 Da, 115.1 – 2 Da, 116.1 – 2 Da, 117.1 – 2 Da, 114.1 – 1 Da, etc...

w,x,y,z represent the percentage of each peak expected to be present at the mass of the reporter ion associated with that peak. Here, w is for 114.1, x for 115.1 etc.:

$$w = (100 - (a + e + i + m))$$

$$x = (100 - (b + f + j + n))$$

$$y = (100 - (c + g + k + o))$$

$$z = (100 - (d + h + l + p))$$

The area (A_r) of each reporter ion peak ($_r$), as calculated above, can now be written in terms of the true areas of peaks (T_r):

$$\begin{aligned} A_{114.1} &= (w*T_{114.1}) + (f*T_{115.1}) + (c*T_{116.1}) \\ A_{115.1} &= (i*T_{114.1}) + (x*T_{115.1}) + (g*T_{116.1}) + (d*T_{117.1}) \\ A_{116.1} &= (m*T_{114.1}) + (j*T_{115.1}) + (y*T_{116.1}) + (h*T_{117.1}) \\ A_{117.1} &= (n*T_{115.1}) + (k*T_{116.1}) + (z*T_{117.1}) \end{aligned}$$

The task is now to calculate each T_r according to these equations.

4 equations and 4 unknown
-> Solve for unknowns
using **Cramer's rule**

The determinant of the matrix of coefficients can be found:

$$|C| = \begin{vmatrix} w, f, c, 0 \\ i, x, g, d \\ m, j, y, h \\ 0, n, k, z \end{vmatrix}$$

Algorithm for Reporter Ion Correction and Adjustment (*afRICA*)

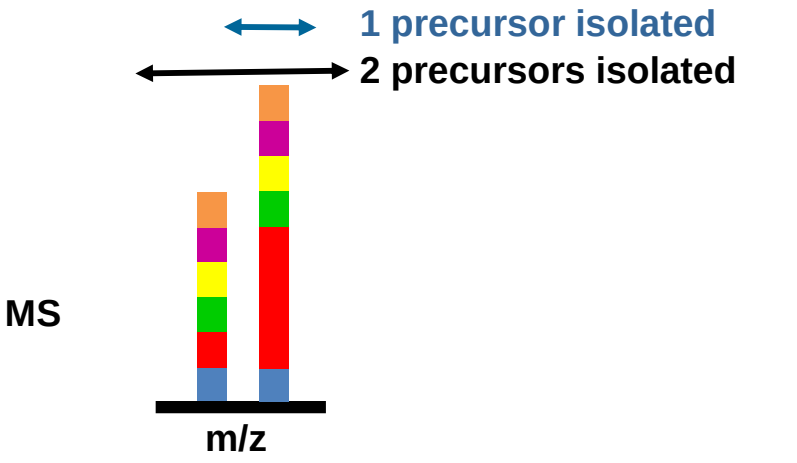
afRICA extends the i-TRACKER algorithm from iTRAQ to TMT

- 4x4 = 16 correction factors for iTRAQ4
- 6x4 = 24 correction factors for TMT6
- 10x4 = 40 correction factors for TMT10

Matrix of Coefficients

- iTRAQ4 4X4
- TMT6 6x6
- TMT10 5x5 (N), 5x5 (C)
- TMT11 5x5 (N), 6x6 (C)
- TMT16 8x8 (N), 8x8 (C)
- TMT18 9x9 (N), 9x9 (C)

Interference and Ratio Compression



Minimizing Interference

Data generation

- Use narrow MS1 precursor isolation
  - 0.7 m/z instead of 2.5

Data analysis

- Exclude MS/MS with low precursor ion purity (PIP) i.e. <50%
  - PIP = $\frac{\text{100\% x intensity precursor 1}}{\text{intensity all isolation window}}$

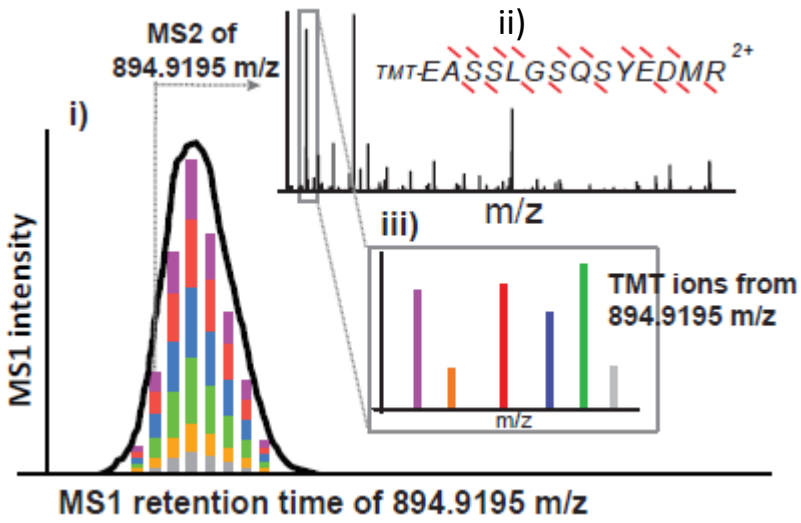
For 2 peptides, relative precursor intensities not proportional to reporter ion intensities

Factors:

- Precursor charge
- MS/MS Collision energy
- # TMT labels (Lys vs Arg)

De-multiplexing iTRAQ and TMT Experiments

Method is used when instead of quantifying across samples one seeks to quantify along proteins ↔ ↕



$$\text{Peptide}_{126}\ \text{abundance} = \frac{\text{MS2 reporter}_{126}\ \text{abundance}}{\sum_{k=126}^{131} \text{MS2 reporter}k\ \text{abundance}}\ x\ \text{MS1 precursor abundance}$$

$$\text{Protein}_{126}\ \text{abundance} = \sum_{j=0}^{n} \text{Peptide}_{126}\ \text{abundance}$$

n is # peptides in the protein

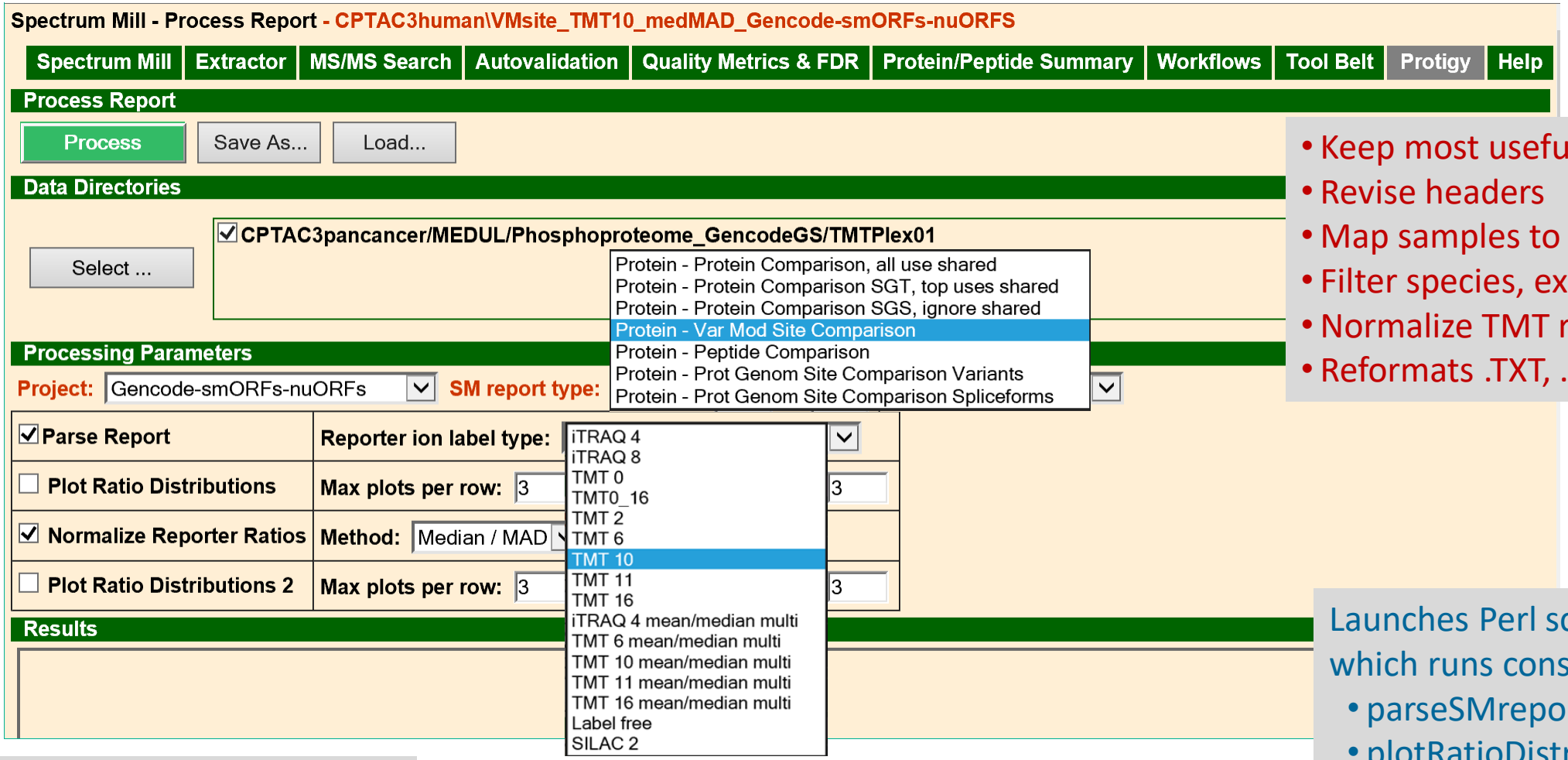
Repeat for each reporter ion

- 126
- 127
- 128
- 129
- 130
- 131

Not yet done directly in SM – done manually using values exported in SM reports, aka fractional intensity method

Process Report

Process Report



- Keep most useful columns from *.ssv
- Revise headers
- Map samples to reporter ions
- Filter species, exclude contaminants
- Normalize TMT ratios
- Reformats .TXT, .GCT

Launches Perl script processReport.pl which runs consecutive R scripts

- parseSMreport.r
- plotRatioDistributions.r
- normalizeReporterRatios.r
- plotRatioDistributions.r

Project-specific features in R scripts

Generates GCT format, if *-sample-annotation.csv present

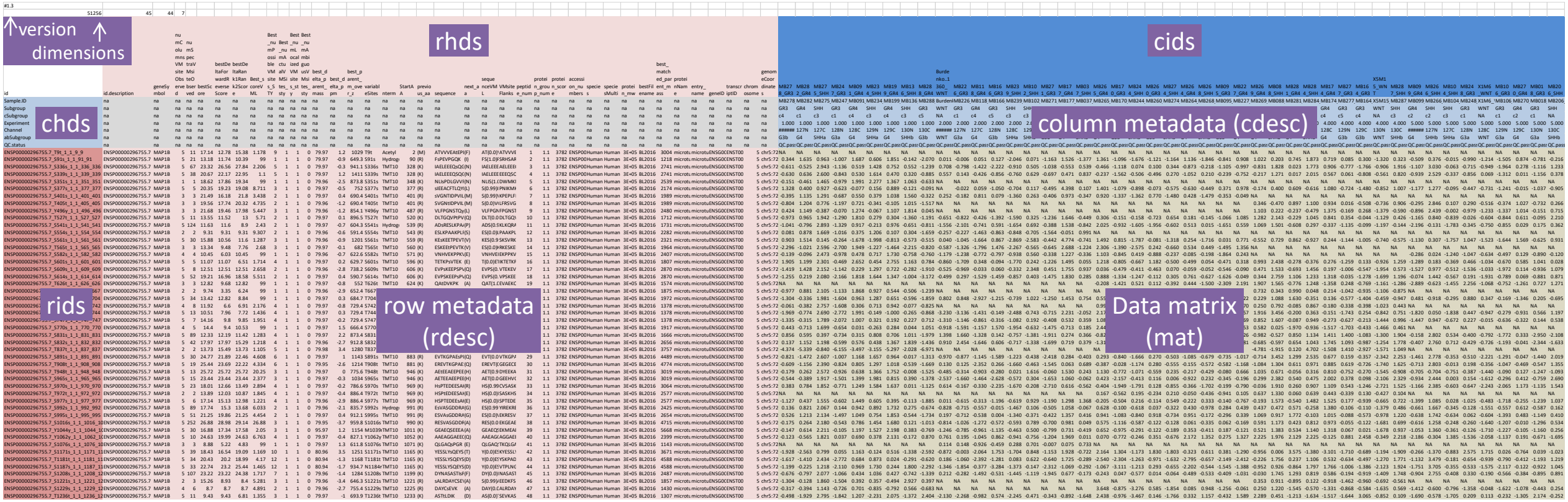
Parsed Report – Normalized, Filtered, Formatted - GCT

https://clue.io/connectopedia/gct_format

GCT (Gene Cluster Text) is a tab-delimited text file format convenient for analysis of matrix datasets which allows storing alongside the data from the experiment,

- column metadata about samples
- row metadata about proteins or VM sites

- Keep most useful columns from *.ssv
- Revise headers, map samples to reporter ions *-sample-annotation.csv
- Filter species, exclude contaminants
- Normalize TMT ratios



Create vertical layout file: *-sample-annotation.csv

Create a sample-annotation.csv file (vertical layout)

The requirements described below are, by design, harmonized between tools developed in the Broad Institute Proteomics Platform group including: Spectrum Mill and the downstream tools **Protigy** and **Panoply**.

In order for the parseSMreport script to correlate each reporter ion label with a sample name and metadata to produce a **GCT v1.3 format** output file you must provide a **comma-delimited .csv file** in the SM directory that contains the report that is parsed. While the filename may be prefixed at the user's discretion, the text file must be named with the suffix

sample-annotation.csv

And be organized like the example below:

The following considerations apply:

- Required columns are only: **Sample.ID**, **Experiment**, and **Channel**.
- Though not required for SM, the column: **Type** (Tumor, NAT, etc) is required downstream for Panoply use.
- Note that the sample-annotation file **must not contain a row for the sample(s) used as a control ion (denominator in ratios)**. However, users are encouraged to include the isobaric label name and control ion channel in the filename of the sample-annotation file.
- The headers of metadata columns need to all be R-compatible. No spaces, no control characters like (), and can not begin with a number.
- Replicates: Sample.IDs must be unique and cannot have duplicates. If replicate samples are present, they must have unique Sample.IDs, but can be identified as replicates by using an additional column (named "Participant" for downstream Panoply use) with identical ids.
- The column **Experiment** (integer values) **must be numbered in the exact same order** as the data directories appear in the report (or Sample.ID and metadata will be mis-applied) , which is the same order as they were selected when creating the report in P/P Summary. Note that the SM data directory need not, but could be a column in the sample-annotation file.
- The column **Channel** should include the reporter ion label used in SM reports. Please note the following nuances: For TMT6, 10 SM reports: 126, 131, while for TMT11, 16, 18 SM reports: 126C, 131N
  . The use of and N and C on every channel was implemented for clarity, and I regret not implementing the convention for the earliest reagents, in deference to the manufacturer's nomenclature, though considered it at the time. Now to ensure backwards compatibility the difference is what it is.

Sorting of the sample-annotation.csv file:

Sort as you wish. It has been expected that when some users prepare a sample-annotation file they might frequently end up with rows sorted in a different order than the data columns in the SM report to which they will be mated, for reasons like sorted by Sample.IDs, or the Channel column sorted alphabetically instead of by mass. So to reduce excess "quick questions" back to the developer, the script has been written to tolerate differences in sort. However, keep in mind that **the output .GCT file will maintain the order of the columns in the input .ssv file**. It is expected that the metadata columns in the sample-annotation file which become rows in the .GCT file will enable facile re-sorting of the data columns during downstream analyses.

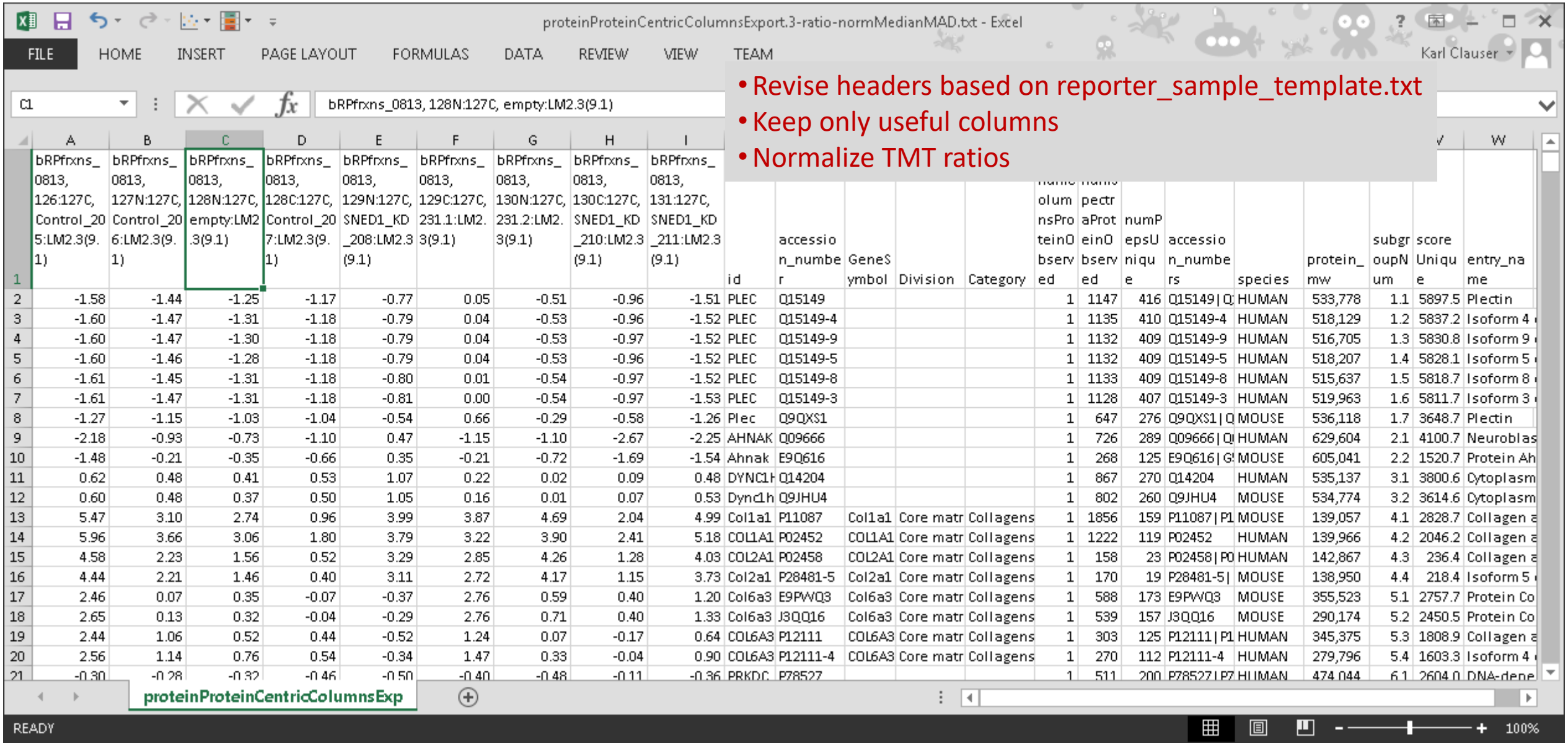
Example sample-annotation.csv file

Download a template suited to your labeling reagent: **TMT6 TMT10 TMT11**

- Manually create file, place in SM data directory
- Automatically read by script: parseSMreport.r

Sample.ID	Participant	Experiment	Channel	Type	Aliquot	Smoking.History	Stage	Country.of.Origin	Age	Gender	Ethnicity	Cigarettes.per.Day	Pack.Years.Smoked	Secondhand.Smoke	QC.status
C3L.02665	C3L-02665	1	126C	Tumor	CPT0192540004	Smoking history not avail	IIB	bulgaria	51	male	caucasian			Exposure t	QC.pass
C3L.02665.N	C3L-02665	1	127N	NAT	CPT0192590004	Smoking history not available		bulgaria	51	male	caucasian			Exposure t	QC.pass
C3L.01663	C3L-01663	1	127C	Tumor	CPT0091610003	Smoking history not avail	IIIA	usa	73	female	caucasian			Exposure t	QC.pass
C3L.01663.N	C3L-01663	1	128N	NAT	CPT0091620003	Smoking history not available		usa	73	female	caucasian			Exposure t	QC.pass
C3N.02575	C3N-02575	1	128C	Tumor	CPT0147680004	Current reformed smoker	IA	china	69	male	han	40	100	Exposure t	QC.pass
C3N.02575.N	C3N-02575	1	129N	NAT	CPT0147700003	Current reformed smoker within		china	69	male	han	40	100	Exposure t	QC.pass
C3L.02546	C3L-02546	1	129C	Tumor	CPT0112650003	Current reformed smoker,	IB	usa	74	male	caucasian	20	40	Exposure t	QC.pass
C3L.02546.N	C3L-02546	1	130N	NAT	CPT0112660003	Current reformed smoker, more		usa	74	male	caucasian	20	40	Exposure t	QC.pass
C3L.00965	C3L-00965	1	130C	Tumor	CPT0101130003	Current reformed smoker	IIA	russia	63	male	white(cauc	20	42	Yes	QC.pass
LUAD.Global.CR..pool.1.	LUAD Globa	1	131N		LUAD Global CR (pool 1)										QC.fail
C3L.02963	C3L-02963	2	126C	Tumor	CPT0197090003	Current smoker: Includes	I	usa	71	male	caucasian	20	56	Exposure t	QC.pass
C3L.02963.N	C3L-02963	2	127N	NAT	CPT0197100003	Current smoker: Includes daily a		usa	71	male	caucasian	20	56	Exposure t	QC.pass
C3N.04162	C3N-04162	2	127C	Tumor	CPT0222730003	Lifelong non-smoker: Less	IIA	ukraine	46	male	slavic			Exposure t	QC.pass
C3N.04162.N	C3N-04162	2	128N	NAT	CPT0222740003	Lifelong non-smoker: Less than 1		ukraine	46	male	slavic			Exposure t	QC.pass
C3L.02646	C3L-02646	2	128C	Tumor	CPT0202530004	Current smoker: Includes	IIIA	bulgaria	74	male	caucasian	50	5	Exposure t	QC.pass
C3L.02646.N	C3L-02646	2	129N	NAT	CPT0202580004	Current smoker: Includes daily a		bulgaria	74	male	caucasian	50	5	Exposure t	QC.pass
C3N.02285	C3N-02285	2	129C	Tumor	CPT0126740003	Current reformed smoker	IIB	poland	72	female	caucasian	15	35.3	Exposure t	QC.pass
C3N.02285.N	C3N-02285	2	130N	NAT	CPT0126760003	Current reformed smoker within		poland	72	female	caucasian	15	35.3	Exposure t	QC.pass
C3N.03875	C3N-03875	2	130C	Tumor	CPT0237410004	Current reformed smoker	IB	china	68	male	han	20		Exposure t	QC.pass
C3N.03875.N	C3N-03875	2	131N	NAT	CPT0237430003	Current reformed smoker within		china	68	male	han	20		Exposure t	QC.pass
C3N.03424	C3N-03424	3	126C	Tumor	CPT0245780003	Current reformed smoker,	IA	poland	74	male	caucasian	Unknown		Exposure t	QC.pass
C3N.03424.N	C3N-03424	3	127N	NAT	CPT0245790003	Current reformed smoker, more		poland	74	male	caucasian	Unknown		Exposure t	QC.pass
C3N.01017	C3N-01017	3	127C	Tumor	CPT0069890003	Current smoker: Includes	IIB	vietnam	56	male	asian	5	5	No or mini	QC.pass
C3N.01017.N	C3N-01017	3	128N	NAT	CPT0069900003	Current smoker: Includes daily a		vietnam	56	male	asian	5	5	No or mini	QC.pass
C3L.00081	C3L-00081	3	128C	Tumor	CPT0001020003	Lifelong non-smoker: Less	IIA	usa	61	female	caucasian			Exposure t	QC.pass
C3L.00081.N	C3L-00081	3	129N	NAT	CPT0001030003	Lifelong non-smoker: Less than 1		usa	61	female	caucasian			Exposure t	QC.pass
C3L.02649	C3L-02649	3	129C	Tumor	CPT0204800003	Current reformed smoker	IIIA	bulgaria	62	male	caucasian	40		Yes	QC.pass
C3L.02649.N	C3L-02649	3	130N	NAT	CPT0204820003	Current reformed smoker within		bulgaria	62	male	caucasian	40		Yes	QC.pass
C3L.01838	C3L-01838	3	130C	Tumor	CPT0091820003	Current reformed smoker,	IB	usa	70	male	caucasian	20	33	Exposure t	QC.pass
LSCC.Tumor.ONLY.CR	LSCC Tumor	3	131N		LSCC Tumor ONLY CR										QC.fail
C3N.02252	C3N-02252	4	126C	Tumor	CPT0128890003			other	68	female					QC.pass
C3N.02252.N	C3N-02252	4	127N	NAT	CPT0128900003			other	68	female					QC.pass
C3L.03963	C3L-03963	4	127C	Tumor	CPT0227400003	Current smoker: Includes	IIB	bulgaria	73	male	caucasian	20	60	Yes	QC.pass
C3L.03963.N	C3L-03963	4	128N	NAT	CPT0227420003	Current smoker: Includes daily a		bulgaria	73	male	caucasian	20	60	Yes	QC.pass
C3N.03072	C3N-03072	4	128C	Tumor	CPT0209020004	Current reformed smoker	IIIA	china	62	male	han	20		Exposure t	QC.pass
C3N.03072.N	C3N-03072	4	129N	NAT	CPT0209040003	Current reformed smoker within		china	62	male	han	20		Exposure t	QC.pass
C3L.02969	C3L-02969	4	129C	Tumor	CPT0217420003	Current reformed smoker	I	usa	76	male	caucasian	20	45	Exposure t	QC.pass
C3L.02969.N	C3L-02969	4	130N	NAT	CPT0217490003	Current reformed smoker within		usa	76	male	caucasian	20	45	Exposure t	QC.pass
C3N.03662	C3N-03662	4	130C	Tumor	CPT0219360003	Lifelong non-smoker: Less	IIA	ukraine	64	male	slavic			Exposure t	QC.pass
C3N.03662.N	C3N-03662	4	131N	NAT	CPT0219390003	Lifelong non-smoker: Less than 1		ukraine	64	male	slavic			Exposure t	QC.pass
C3N.03051	C3N-03051	5	126C	Tumor	CPT0208770003	Current smoker: Includes	IIIA	china	65	male	han	20		Exposure t	QC.pass

Parsed Report - Normalized, Formatted - TXT



	A	B	C	D	E	F	G	H	I														
1	bRPfrxns_0813, 126:127C, Control_205:LM2.3(9.1)	bRPfrxns_0813, 127N:127C, Control_206:LM2.3(9.1)	bRPfrxns_0813, 128N:127C, empty:LM2.3(9.1)	bRPfrxns_0813, 128C:127C, Control_207:LM2.3(9.1)	bRPfrxns_0813, 129N:127C, SNED1_KD_208:LM2.3(9.1)	bRPfrxns_0813, 129C:127C, 231.1:LM2.3(9.1)	bRPfrxns_0813, 130N:127C, 231.2:LM2.3(9.1)	bRPfrxns_0813, 130C:127C, SNED1_KD_210:LM2.3(9.1)	bRPfrxns_0813, 131:127C, SNED1_KD_211:LM2.3(9.1)	id	accession_number	GeneSymbol	Division	Category	numberOfColumnsProteinObserved	numberOfSpectraProteinObserved	numPepsUnique	accession_numbers	species	protein_mw	subgroupNum	scoreUnique	entry_name
2	-1.58	-1.44	-1.25	-1.17	-0.77	0.05	-0.51	-0.96	-1.51	PLEC	Q15149				1	1147	416	Q15149|Q	HUMAN	533,778	1.1	5897.5	Plectin
3	-1.60	-1.47	-1.31	-1.18	-0.79	0.04	-0.53	-0.96	-1.52	PLEC	Q15149-4				1	1135	410	Q15149-4	HUMAN	518,129	1.2	5837.2	Isoform 4
4	-1.60	-1.47	-1.30	-1.18	-0.79	0.04	-0.53	-0.97	-1.52	PLEC	Q15149-9				1	1132	409	Q15149-9	HUMAN	516,705	1.3	5830.8	Isoform 9
5	-1.60	-1.46	-1.28	-1.18	-0.79	0.04	-0.53	-0.96	-1.52	PLEC	Q15149-5				1	1132	409	Q15149-5	HUMAN	518,207	1.4	5828.1	Isoform 5
6	-1.61	-1.45	-1.31	-1.18	-0.80	0.01	-0.54	-0.97	-1.52	PLEC	Q15149-8				1	1133	409	Q15149-8	HUMAN	515,637	1.5	5818.7	Isoform 8
7	-1.61	-1.47	-1.31	-1.18	-0.81	0.00	-0.54	-0.97	-1.53	PLEC	Q15149-3				1	1128	407	Q15149-3	HUMAN	519,963	1.6	5811.7	Isoform 3
8	-1.27	-1.15	-1.03	-1.04	-0.54	0.66	-0.29	-0.58	-1.26	Plec	Q9QXS1				1	647	276	Q9QXS1|Q	MOUSE	536,118	1.7	3648.7	Plectin
9	-2.18	-0.93	-0.73	-1.10	0.47	-1.15	-1.10	-2.67	-2.25	AHNAK	Q09666				1	726	289	Q09666|Q	HUMAN	629,604	2.1	4100.7	Neuroblas
10	-1.48	-0.21	-0.35	-0.66	0.35	-0.21	-0.72	-1.69	-1.54	Ahnak	E9Q616				1	268	125	E9Q616|G	MOUSE	605,041	2.2	1520.7	Protein Ah
11	0.62	0.48	0.41	0.53	1.07	0.22	0.02	0.09	0.48	DYNC1H	Q14204				1	867	270	Q14204	HUMAN	535,137	3.1	3800.6	Cytoplasm
12	0.60	0.48	0.37	0.50	1.05	0.16	0.01	0.07	0.53	Dync1h	Q9JHU4				1	802	260	Q9JHU4	MOUSE	534,774	3.2	3614.6	Cytoplasm
13	5.47	3.10	2.74	0.96	3.99	3.87	4.69	2.04	4.99	Col1a1	P11087	Col1a1	Core matr	Collagens	1	1856	159	P11087|P1	MOUSE	139,057	4.1	2828.7	Collagen a
14	5.96	3.66	3.06	1.80	3.79	3.22	3.90	2.41	5.18	COL1A1	P02452	COL1A1	Core matr	Collagens	1	1222	119	P02452	HUMAN	139,966	4.2	2046.2	Collagen a
15	4.58	2.23	1.56	0.52	3.29	2.85	4.26	1.28	4.03	COL2A1	P02458	COL2A1	Core matr	Collagens	1	158	23	P02458|P0	HUMAN	142,867	4.3	236.4	Collagen a
16	4.44	2.21	1.46	0.40	3.11	2.72	4.17	1.15	3.73	Col2a1	P28481-5	Col2a1	Core matr	Collagens	1	170	19	P28481-5|	MOUSE	138,950	4.4	218.4	Isoform 5
17	2.46	0.07	0.35	-0.07	-0.37	2.76	0.59	0.40	1.20	Col6a3	E9PWQ3	Col6a3	Core matr	Collagens	1	588	173	E9PWQ3	MOUSE	355,523	5.1	2757.7	Protein Co
18	2.65	0.13	0.32	-0.04	-0.29	2.76	0.71	0.40	1.33	Col6a3	J3QQ16	Col6a3	Core matr	Collagens	1	539	157	J3QQ16	MOUSE	290,174	5.2	2450.5	Protein Co
19	2.44	1.06	0.52	0.44	-0.52	1.24	0.07	-0.17	0.64	COL6A3	P12111	COL6A3	Core matr	Collagens	1	303	125	P12111|P1	HUMAN	345,375	5.3	1808.9	Collagen a
20	2.56	1.14	0.76	0.54	-0.34	1.47	0.33	-0.04	0.90	COL6A3	P12111-4	COL6A3	Core matr	Collagens	1	270	112	P12111-4	HUMAN	279,796	5.4	1603.3	Isoform 4
21	-0.30	-0.28	-0.32	-0.46	-0.50	-0.40	-0.48	-0.11	-0.36	PRKDC	P78527				1	511	200	P78527|P7	HUMAN	474,044	6.1	2604.0	DNA-depe

Auxiliary Files

reporter_sample_template.txt

- Manually create file
  - ✓ Sort columns into order desired in output
  - ✓ Place in SM data directory
- Automatically read by script: parseSMreport.r

	A	B	C	D	E	F	G	H	I	J
1	129C	130N	126	127N	128C	128N	127C	129N	130C	131
2	231.1	231.2	Control_205	Control_206	Control_207	empty	LM2.3(9.1)	SNED1_KD_208	SNED1_KD_210	SNED1_KD_211
3	bRPfrxns_0813	bRPfrxns_0813	bRPfrxns_0813	bRPfrxns_0813	bRPfrxns_0813	bRPfrxns_0813	bRPfrxns_0813	bRPfrxns_0813	bRPfrxns_0813	bRPfrxns_0813

*norm-stats.txt

This file is automatically generated by the script: normalizeReporterRatios.r

proteinProteinCentricColumnsExport.3-ratio-norm-stats.txt - Excel

FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW TEAM Karl Clauser

H14

	A	B	C	D	E	F
1	norm.params.sample	norm.params.mean	norm.params.median	norm.params.sd	norm.params.mad	
2	bRPfrxns_0813, 126:127C, Control_205:LM2.3(9.1)	-0.187	-0.233	0.616	0.385	
3	bRPfrxns_0813, 127N:127C, Control_206:LM2.3(9.1)	0.135	0.090	0.637	0.420	
4	bRPfrxns_0813, 128N:127C, empty:LM2.3(9.1)	-4.958	-4.947	0.711	0.471	
5	bRPfrxns_0813, 128C:127C, Control_207:LM2.3(9.1)	-0.040	-0.077	0.554	0.387	
6	bRPfrxns_0813, 129N:127C, SNED1_KD_208:LM2.3(9.1)	-0.488	-0.528	0.729	0.513	
7	bRPfrxns_0813, 129C:127C, 231.1:LM2.3(9.1)	-0.043	-0.140	0.562	0.350	
8	bRPfrxns_0813, 130N:127C, 231.2:LM2.3(9.1)	-0.825	-0.912	0.800	0.593	
9	bRPfrxns_0813, 130C:127C, SNED1_KD_210:LM2.3(9.1)	0.090	0.053	0.553	0.353	
10	bRPfrxns_0813, 131:127C, SNED1_KD_211:LM2.3(9.1)	-0.186	-0.219	0.628	0.359	
11						

proteinProteinCentricColumnsExp

READY 100%

Create horizontal layout file: reporter_sample_template.txt

Create a reporter_sample_template.txt file (horizontal layout)

When the parseSMreport script is run to process a P/P Summary report it can update the column headers in the report to include meta information, like more specific sample names attached to each reporter ion. These names will also be propagated through to the ratio distribution plots, and reports following normalization.

In order to provide the script with information on a sample name for each reporter ion you must create a **tab-delimited text file** in the SM directory that contains the initial report that is parsed. The text file must be named:
reporter_sample_template.txt
And be organized like the following example:
This example is for an experiment involving 18 samples that were run in 2 separate TMT-10 plexes, with 9 samples and a common control (mix of all 18).

1. line 1: reporter ion. The parser looks only for a substring in each cells text for the reporter ion (126, 127N, etc). The A and B are optional, and added only for user convenience.
2. line 2: sample name, free text.
3. line 3: directory name, directory the data was present in when the report was generated. This is necessary to match up the name and reporter ion, when more than 1 directory contributed to the report.

Special considerations;

- Do not use a colon character : in any cells of the report, because colon will later be inserted by the parser to denote the numerator : denominator involved in a ratio.
- The examples have the columns in mass order. However, you can put the columns in any order. The parsed report will come out with the same column order as the reporter_sample_template.txt.
- When saving reporter_sample_template.txt in Excel
  Save as type:
  Text (Tab delimited) (*.txt)
  not
  Unicode Text (*.txt)
  You can diagnose problems caused by saving as Unicode Text by the funky characters in the script output pane in the browser in the reporterIons field.
  numerator: 126 denominator: 127C reporterIons: ÿþ1

A126	A127N	A127C	A128N	A128C	A129N	A129C	A130N	A130C	A131	B126	B127N	B127C	B128N	B128C	B129N
MD-8214C long survival, no chemo	MD-8226C long survival, unknown chemo	NP-8932N	PC-8592T	PI-8592N2	PI-8762N	PD-8832T long survival, palliative chemo	PD-8715 short survival, adj chemo	WD-8996C long survival, adj chemo	CC-all18	MD-7800T, long survival, unknown chemo	MD-1044C short survival, adj chemo	NP-8926N0308	PC-1254T	PI-8926C	PD-8216C long survival, adj chemo
TMT10A_bRP	TMT10A_bRP	TMT10A_bRP	TMT10A_bR	TMT10A_bRP	TMT10A_bRP	TMT10A_bRP	TMT10A_bRP	TMT10A_bRP	TMT10A_bRP	TMT10B_bRP	TMT10B_bRP	TMT10B_bRP	TMT10B_bRP	TMT10B_bRP	TMT10B_bR

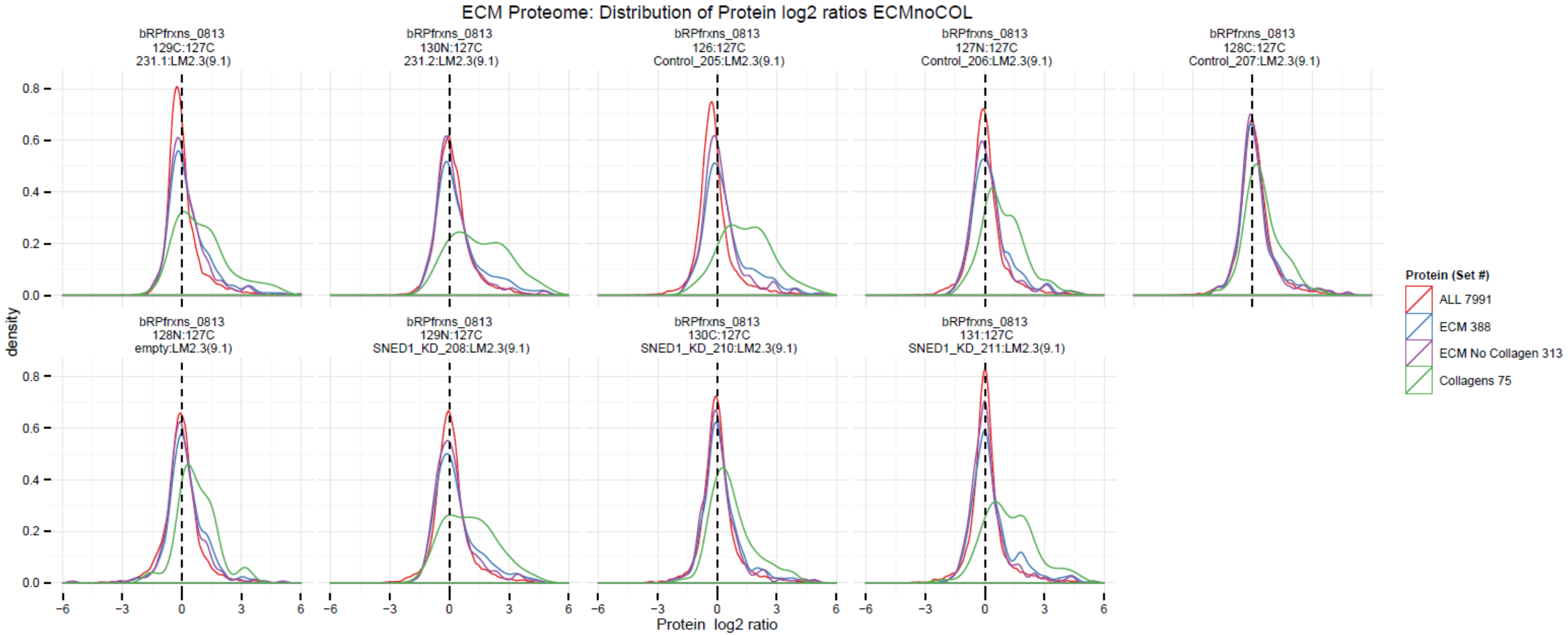
MeanMulti

If the control ion was designated as MeanMulti, when generating the SM report, this leads to the mean intensity of multiple reporter ions being used as the denominator in ratios for each of the TMT10 mass labels. Consequently, there will be 1 additional ratio for each directory in the report. In the corresponding reporter_sample_template, indicate which ions were used as the denominator by joining the ions with a dot character (and keep them in alphabetic order, to match the column headers in the .ssv file). The dot delimiter enables the parser to prevent the parser from expecting this channel to also be used as a numerator.

126	127N	127C	128N	128C	129N	129C	130N	130C	131	126.127N.128C
Control_205	Control_206	LM2.3(9.1)	empty	Control_207	SNED1_KD_208	231.1	231.2	SNED1_KD_210	SNED1_KD_211	Control_3_mean
bRPfrxns_0813	bRPfrxns_0813	bRPfrxns_0813	bRPfrxns_0813	bRPfrxns_0813	bRPfrxns_0813	bRPfrxns_0813	bRPfrxns_0813	bRPfrxns_0813	bRPfrxns_0813	bRPfrxns_0813

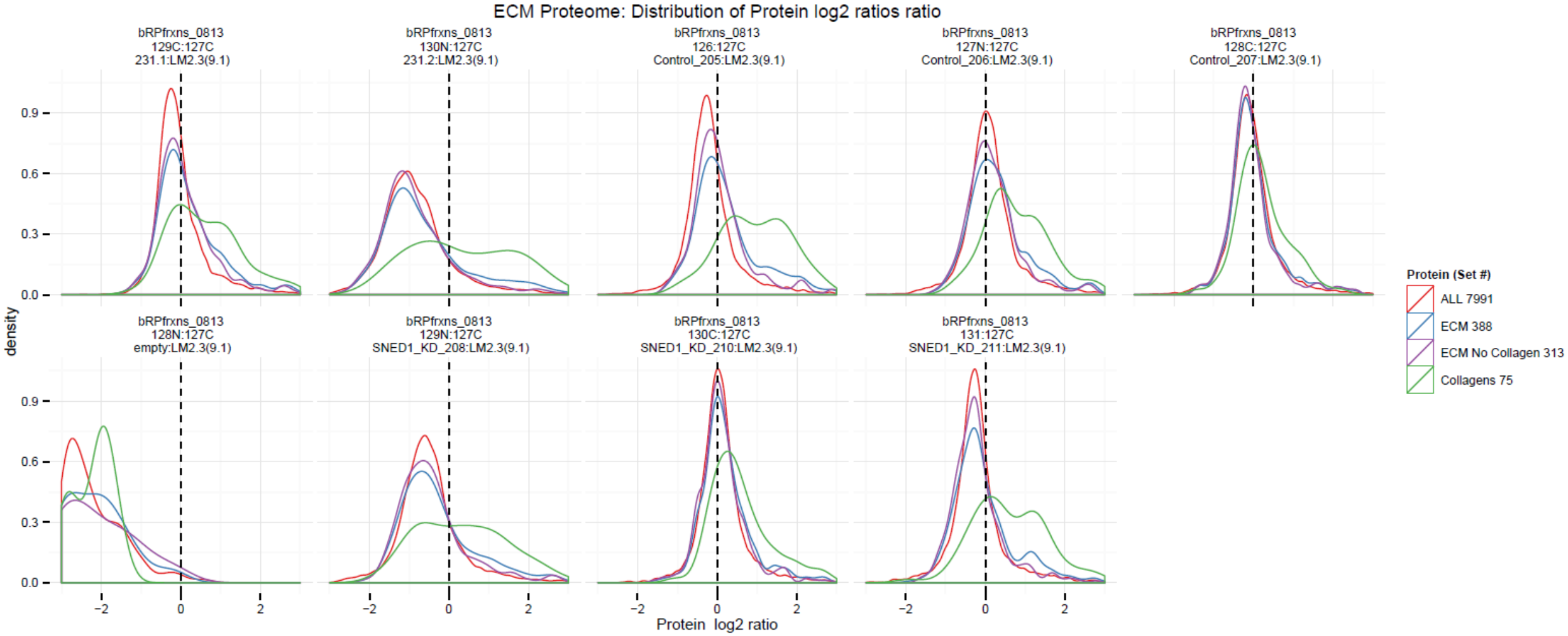
TMT ratio Distributions – Normalized by median/SD of ECM(no Coll) - SGS

proteinProteinCentricColumnsExport.SGS.10-ratio-normmedianSD-ECMnoCOL-ECM-Protein-distributions4EnC.pdf



TMT ratio Distributions – Unnormalized LM2 denom - SGS

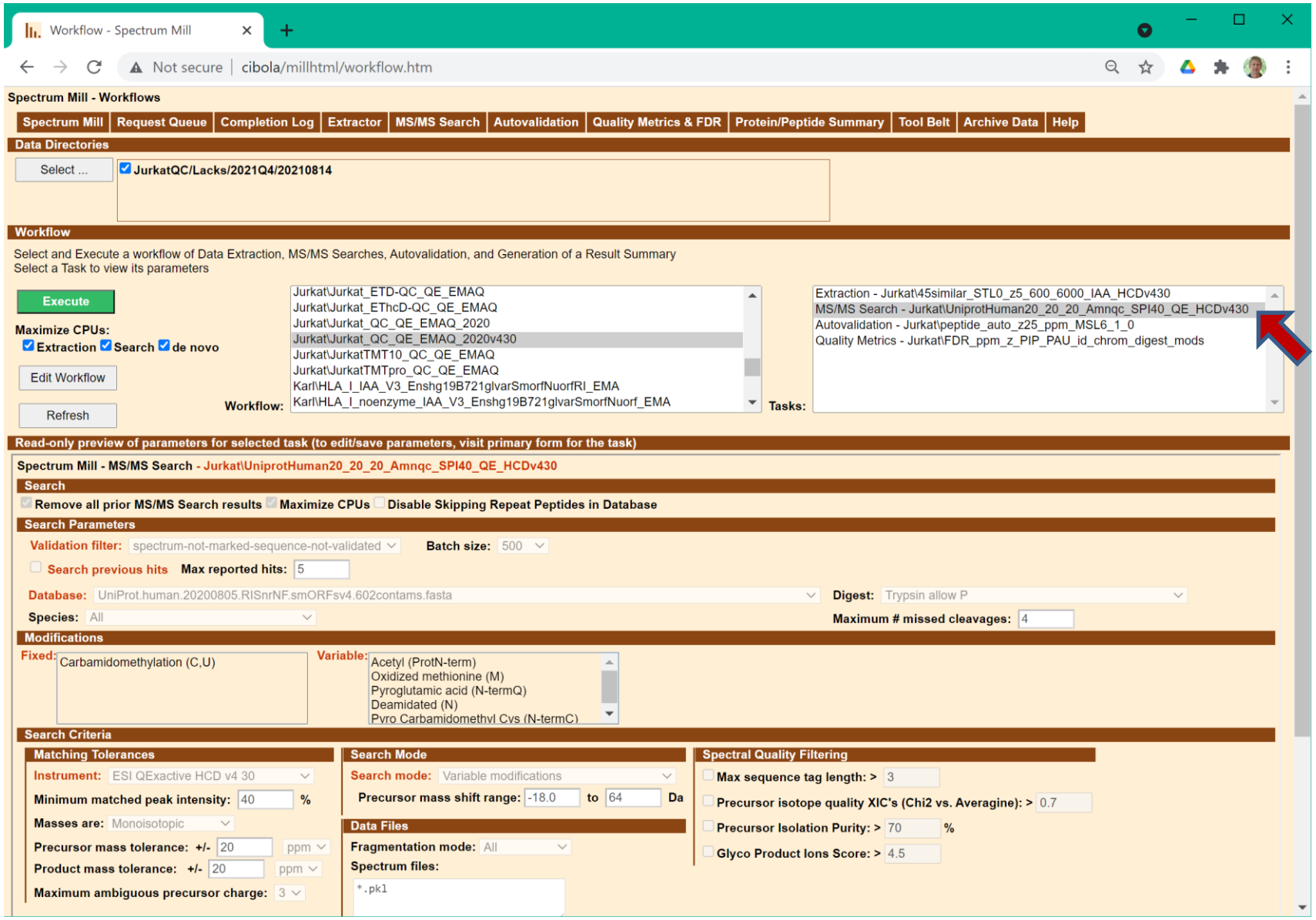
proteinProteinCentricColumnsExport.SGS.10-ratio-ECM-Protein-distributions4EnC.pdf



Automation

Service Request Manager (SRM)
and Workflows

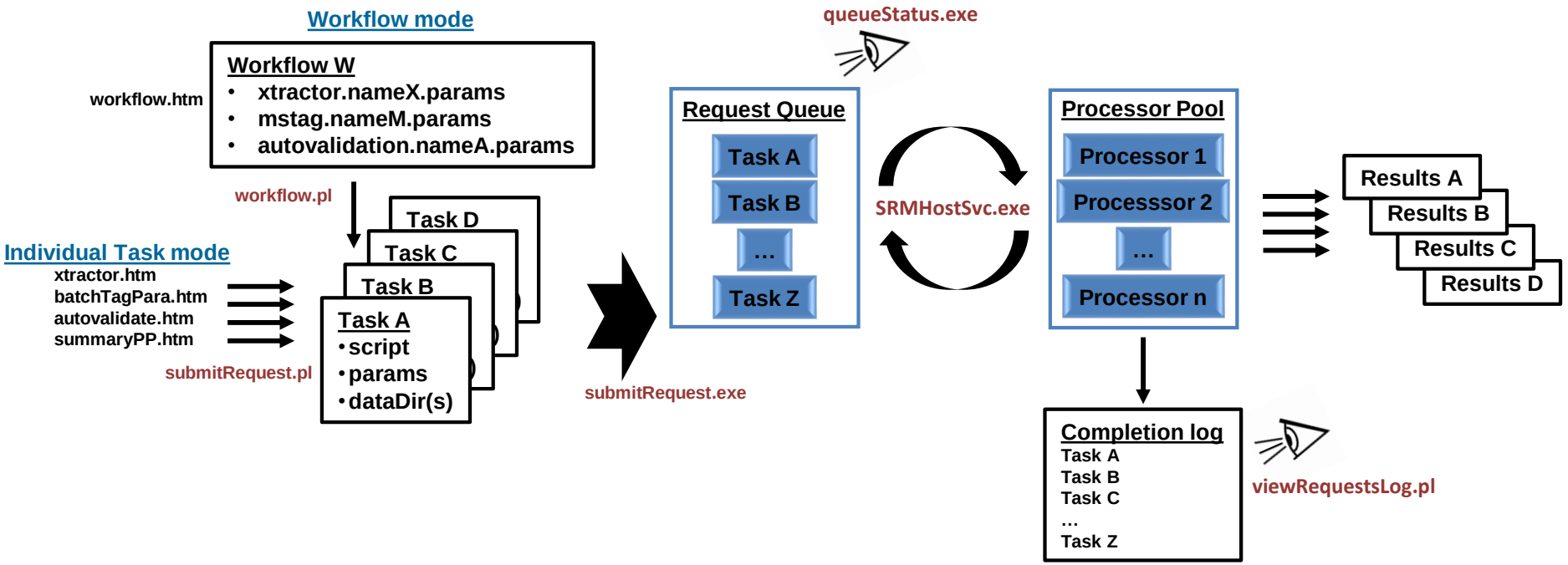
Simple Workflow – Broad Institute QC sample



List of tasks in workflow

Task parameter preview in normal layout (inactive)

Mechanisms for Controlling Processes - Service Request Manager (SRM)



SRM coordinates task execution on the queue to maximize CPU usage while maintaining workflow order dependencies

Tasks in each user's workflow execute in serial

runXtractor.pl → ***.mzXML, specFeatures.1.tsv**

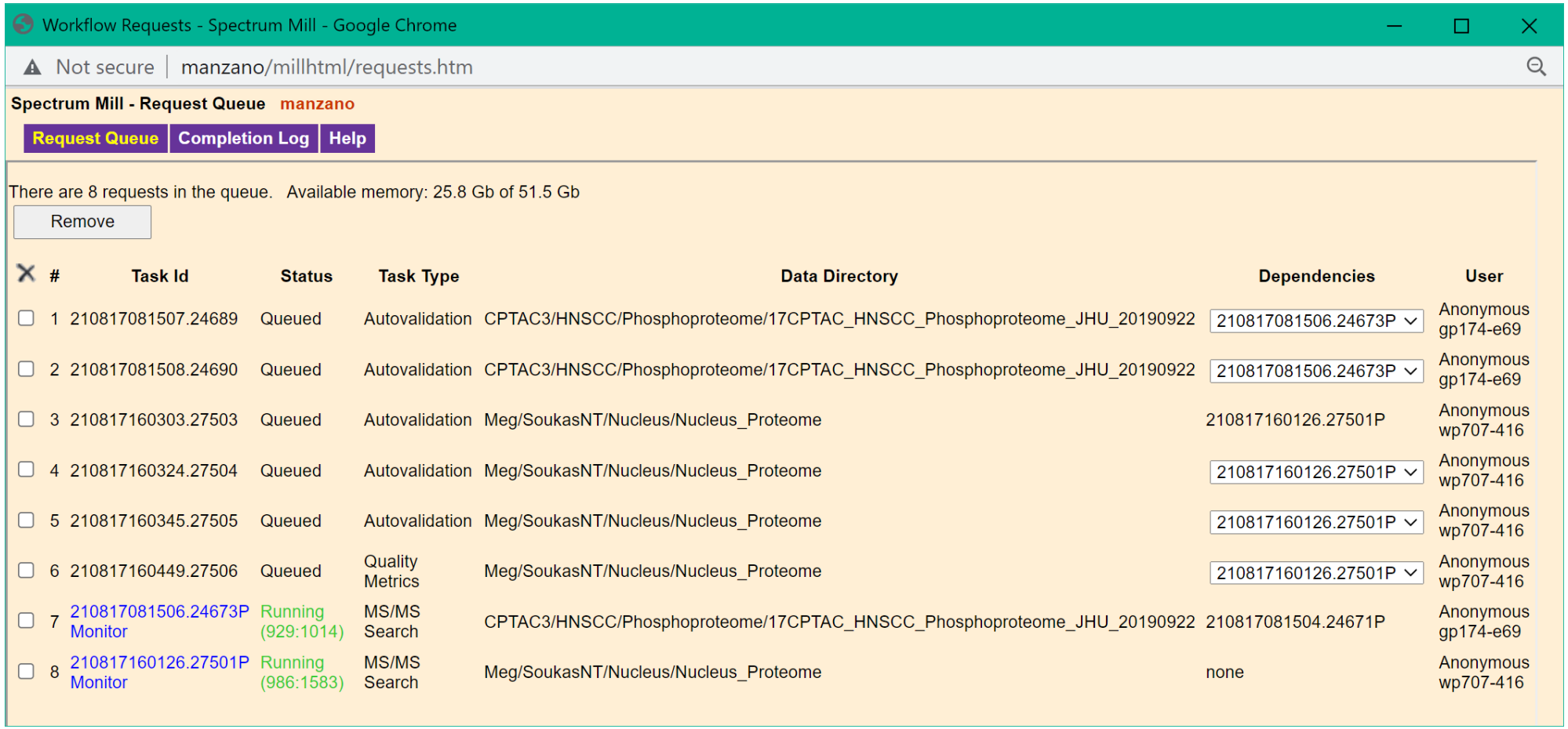
batchTagPara.pl → **spo.zip, tagSummary.1.tsv**

validateTable.pl → **hitTable.1.tsv, spectrumTable.1.tsv**

summaryPP.pl → **reportMode.1.ssv**

eyes peeking, not rocket ships blasting off

SRM Request Queue Status

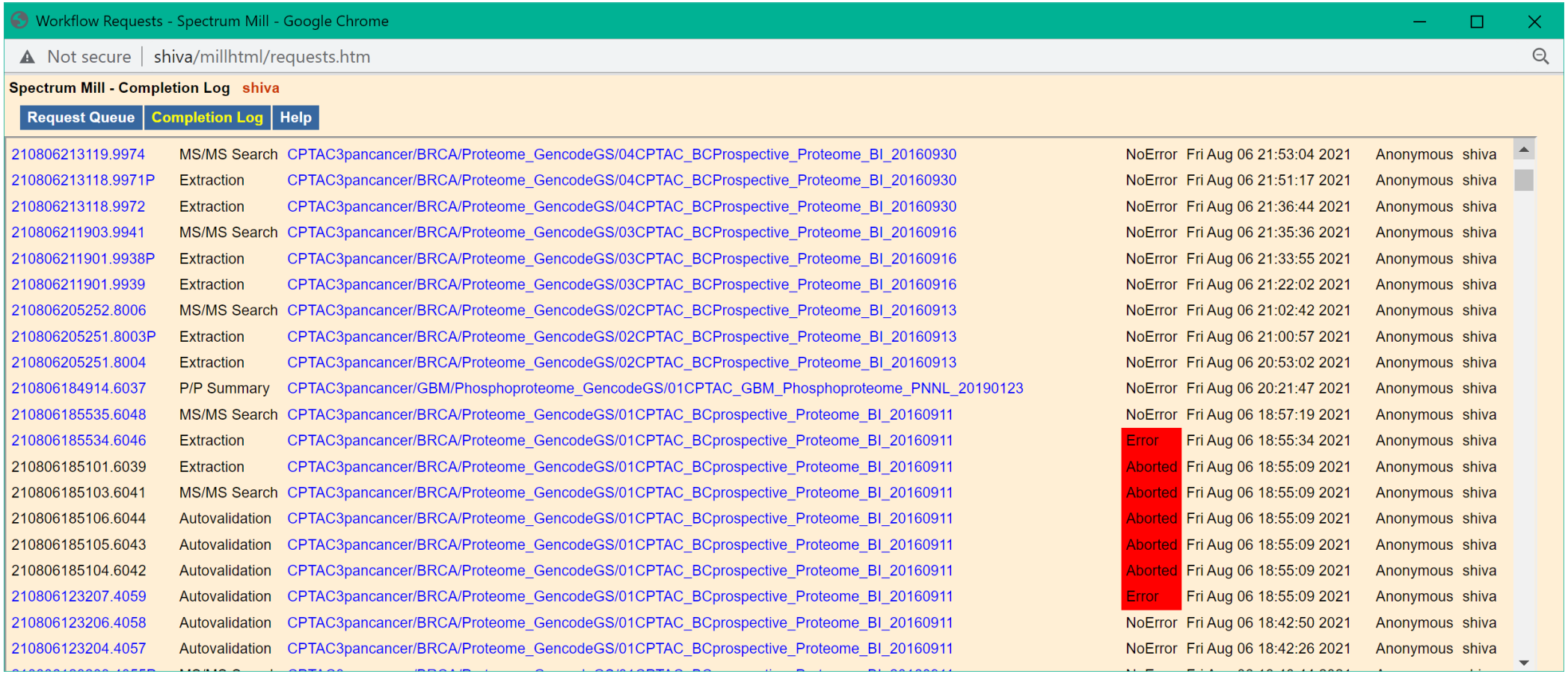


✕	#	Task Id	Status	Task Type	Data Directory	Dependencies	User
☐	1	210817081507.24689	Queued	Autovalidation	CPTAC3/HNSCC/Phosphoproteome/17CPTAC_HNSCC_Phosphoproteome_JHU_20190922	210817081506.24673P	Anonymous gp174-e69
☐	2	210817081508.24690	Queued	Autovalidation	CPTAC3/HNSCC/Phosphoproteome/17CPTAC_HNSCC_Phosphoproteome_JHU_20190922	210817081506.24673P	Anonymous gp174-e69
☐	3	210817160303.27503	Queued	Autovalidation	Meg/SoukasNT/Nucleus/Nucleus_Proteome	210817160126.27501P	Anonymous wp707-416
☐	4	210817160324.27504	Queued	Autovalidation	Meg/SoukasNT/Nucleus/Nucleus_Proteome	210817160126.27501P	Anonymous wp707-416
☐	5	210817160345.27505	Queued	Autovalidation	Meg/SoukasNT/Nucleus/Nucleus_Proteome	210817160126.27501P	Anonymous wp707-416
☐	6	210817160449.27506	Queued	Quality Metrics	Meg/SoukasNT/Nucleus/Nucleus_Proteome	210817160126.27501P	Anonymous wp707-416
☐	7	210817081506.24673P Monitor	Running (929:1014)	MS/MS Search	CPTAC3/HNSCC/Phosphoproteome/17CPTAC_HNSCC_Phosphoproteome_JHU_20190922	210817081504.24671P	Anonymous gp174-e69
☐	8	210817160126.27501P Monitor	Running (986:1583)	MS/MS Search	Meg/SoukasNT/Nucleus/Nucleus_Proteome	none	Anonymous wp707-416

User A's workflow executes in parallel with user B's workflow

Any user can add/remove tasks from queue

SRM Completion Log

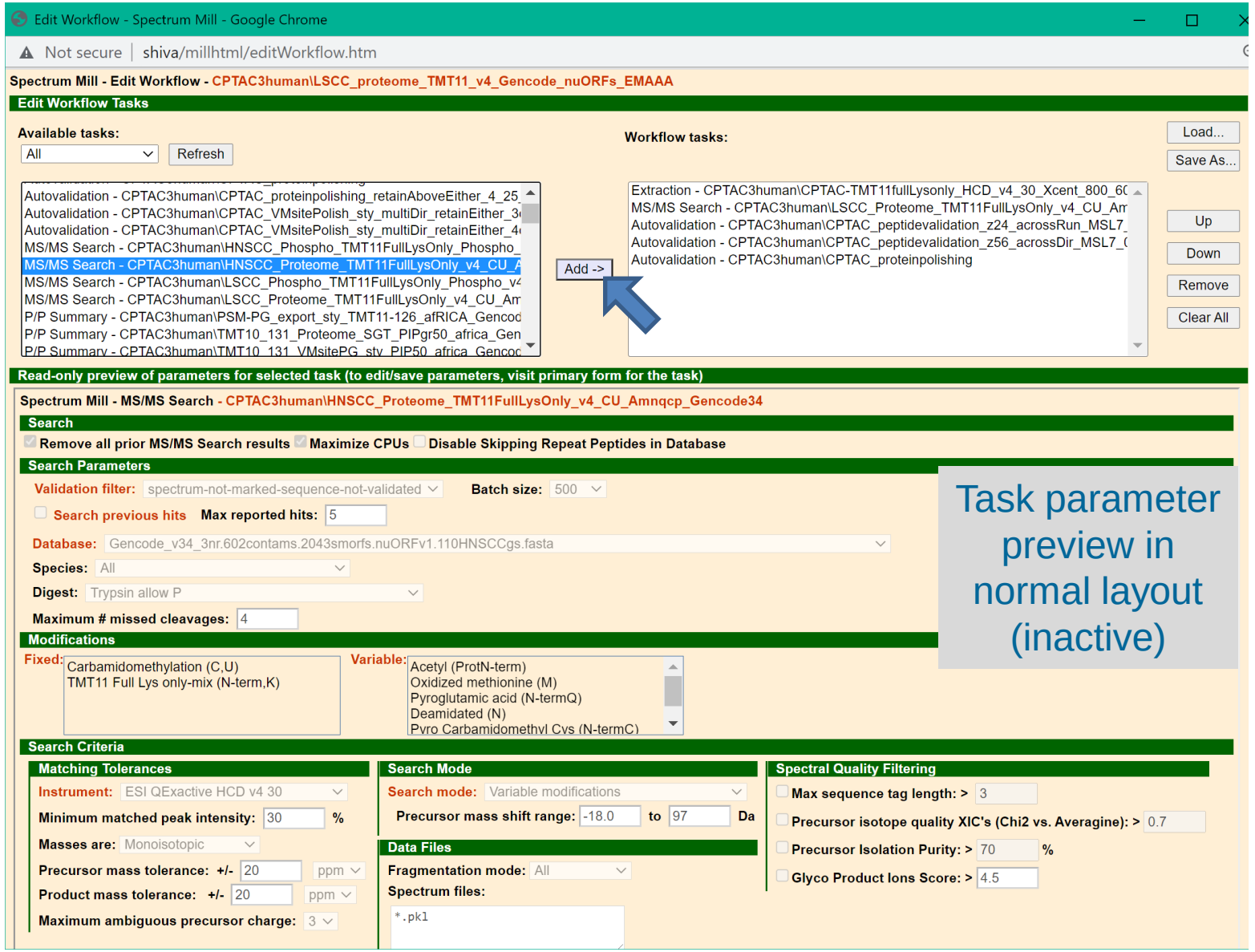


210806213119.9974	MS/MS Search	CPTAC3pancancer/BRCA/Proteome_GencodeGS/04CPTAC_BCProspective_Proteome_BI_20160930	NoError	Fri Aug 06 21:53:04 2021	Anonymous	shiva
210806213118.9971P	Extraction	CPTAC3pancancer/BRCA/Proteome_GencodeGS/04CPTAC_BCProspective_Proteome_BI_20160930	NoError	Fri Aug 06 21:51:17 2021	Anonymous	shiva
210806213118.9972	Extraction	CPTAC3pancancer/BRCA/Proteome_GencodeGS/04CPTAC_BCProspective_Proteome_BI_20160930	NoError	Fri Aug 06 21:36:44 2021	Anonymous	shiva
210806211903.9941	MS/MS Search	CPTAC3pancancer/BRCA/Proteome_GencodeGS/03CPTAC_BCProspective_Proteome_BI_20160916	NoError	Fri Aug 06 21:35:36 2021	Anonymous	shiva
210806211901.9938P	Extraction	CPTAC3pancancer/BRCA/Proteome_GencodeGS/03CPTAC_BCProspective_Proteome_BI_20160916	NoError	Fri Aug 06 21:33:55 2021	Anonymous	shiva
210806211901.9939	Extraction	CPTAC3pancancer/BRCA/Proteome_GencodeGS/03CPTAC_BCProspective_Proteome_BI_20160916	NoError	Fri Aug 06 21:22:02 2021	Anonymous	shiva
210806205252.8006	MS/MS Search	CPTAC3pancancer/BRCA/Proteome_GencodeGS/02CPTAC_BCProspective_Proteome_BI_20160913	NoError	Fri Aug 06 21:02:42 2021	Anonymous	shiva
210806205251.8003P	Extraction	CPTAC3pancancer/BRCA/Proteome_GencodeGS/02CPTAC_BCProspective_Proteome_BI_20160913	NoError	Fri Aug 06 21:00:57 2021	Anonymous	shiva
210806205251.8004	Extraction	CPTAC3pancancer/BRCA/Proteome_GencodeGS/02CPTAC_BCProspective_Proteome_BI_20160913	NoError	Fri Aug 06 20:53:02 2021	Anonymous	shiva
210806184914.6037	P/P Summary	CPTAC3pancancer/GBM/Phosphoproteome_GencodeGS/01CPTAC_GBM_Phosphoproteome_PNNL_20190123	NoError	Fri Aug 06 20:21:47 2021	Anonymous	shiva
210806185535.6048	MS/MS Search	CPTAC3pancancer/BRCA/Proteome_GencodeGS/01CPTAC_BCprospective_Proteome_BI_20160911	NoError	Fri Aug 06 18:57:19 2021	Anonymous	shiva
210806185534.6046	Extraction	CPTAC3pancancer/BRCA/Proteome_GencodeGS/01CPTAC_BCprospective_Proteome_BI_20160911	Error	Fri Aug 06 18:55:34 2021	Anonymous	shiva
210806185101.6039	Extraction	CPTAC3pancancer/BRCA/Proteome_GencodeGS/01CPTAC_BCprospective_Proteome_BI_20160911	Aborted	Fri Aug 06 18:55:09 2021	Anonymous	shiva
210806185103.6041	MS/MS Search	CPTAC3pancancer/BRCA/Proteome_GencodeGS/01CPTAC_BCprospective_Proteome_BI_20160911	Aborted	Fri Aug 06 18:55:09 2021	Anonymous	shiva
210806185106.6044	Autovalidation	CPTAC3pancancer/BRCA/Proteome_GencodeGS/01CPTAC_BCprospective_Proteome_BI_20160911	Aborted	Fri Aug 06 18:55:09 2021	Anonymous	shiva
210806185105.6043	Autovalidation	CPTAC3pancancer/BRCA/Proteome_GencodeGS/01CPTAC_BCprospective_Proteome_BI_20160911	Aborted	Fri Aug 06 18:55:09 2021	Anonymous	shiva
210806185104.6042	Autovalidation	CPTAC3pancancer/BRCA/Proteome_GencodeGS/01CPTAC_BCprospective_Proteome_BI_20160911	Aborted	Fri Aug 06 18:55:09 2021	Anonymous	shiva
210806123207.4059	Autovalidation	CPTAC3pancancer/BRCA/Proteome_GencodeGS/01CPTAC_BCprospective_Proteome_BI_20160911	Error	Fri Aug 06 18:55:09 2021	Anonymous	shiva
210806123206.4058	Autovalidation	CPTAC3pancancer/BRCA/Proteome_GencodeGS/01CPTAC_BCprospective_Proteome_BI_20160911	NoError	Fri Aug 06 18:42:50 2021	Anonymous	shiva
210806123204.4057	Autovalidation	CPTAC3pancancer/BRCA/Proteome_GencodeGS/01CPTAC_BCprospective_Proteome_BI_20160911	NoError	Fri Aug 06 18:42:26 2021	Anonymous	shiva

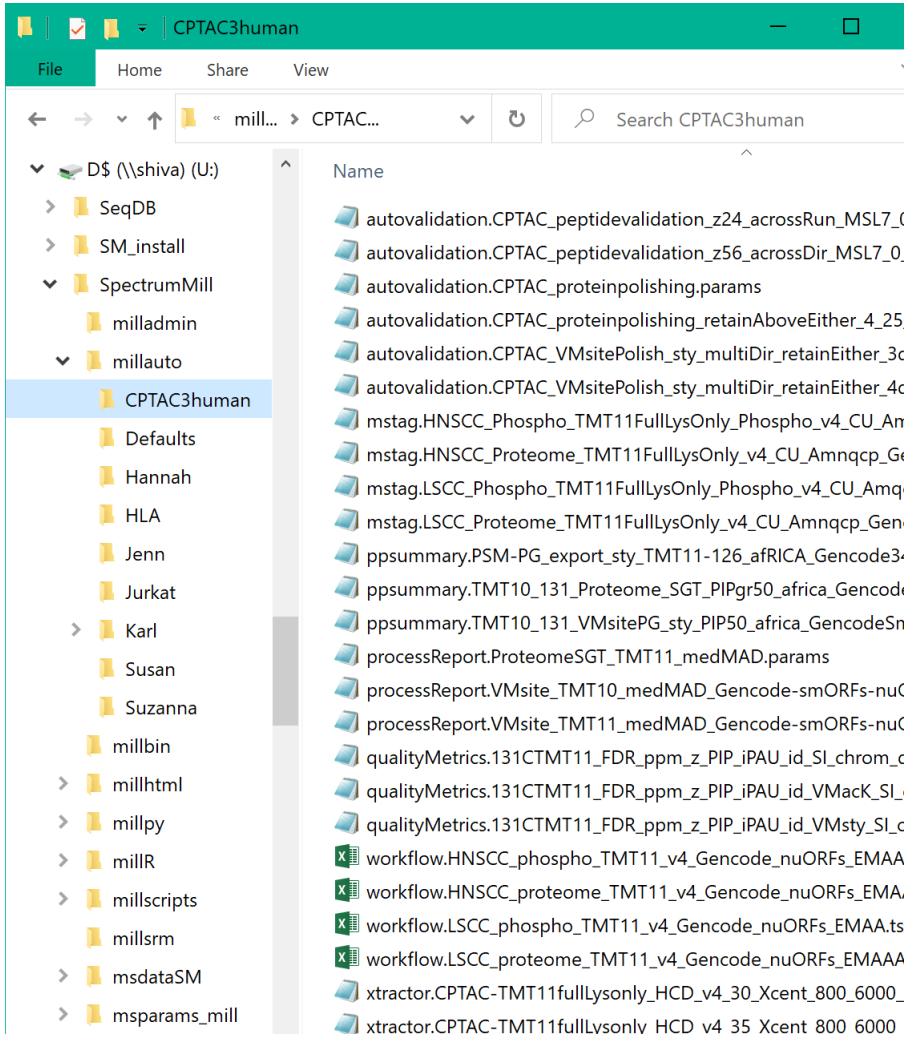
If a task fails, the SRM aborts all subsequent tasks in the workflow

Edit Workflow User Interface

Edit an existing workflow to replace a task



Task parameter preview in normal layout (inactive)



Workflow and Saved Parameter files

workflow.LSCC_phospho_TMT11_v4_Gencode_nuORFs_EMAA.tsv

requestScript	paramsFile
runXtractor.pl	CPTAC3human\xtractor.CPTAC-TMT11fullLysonly_HCD_v4_35_Xcent_800_6000_45sec_z6.params
batchTagPara.pl	CPTAC3human\mstag.LSCC_Phospho_TMT11FullLysOnly_Phospho_v4_CU_AmqcstynG_GencodeNuORFs.params
validateTable.pl	CPTAC3human\autovalidation.CPTAC_peptidevalidation_z24_acrossRun_MSL7_0_8_BCS3.params
validateTable.pl	CPTAC3human\autovalidation.CPTAC_peptidevalidation_z56_acrossDir_MSL7_0_4_BCS3.params

Workflow file lists each task and its parameter file name

Workflows and saved parameters files located at:
SpectrumMill/millauto

mstag.LSCC_Phospho_TMT11FullLysOnly_Phospho_v4_CU_AmqcstynG_GencodeNuORFs.params

```
msparams_dir=msparams_mill%2F
seqdb_dir=D%3A%5CSeqDB%5C
requestScript=batchTagPara.pl
input_program_name=mstag
removeResults=1
hide_html_links=1
validationState=spectrum-not-marked-sequence-not-validated
batchSize=500
max_reported_hits=5
database=Gencode_v34_3nr.602contams.2043smorfs.nuORFv1.
110LSCCgs.fasta
full_mw_range=1
full_pi_range=1
results_to_file=1
use_instrument_ion_types=1
species=All
enzyme=Trypsin%20allow%20P
missed_cleavages=4
fixedMods=carbamidomethylationCU
fixedMods=TMT11-Full-Lys
varMods=Acetyl
varMods=Oxidized-Methionine
varMods=pyroGlu
varMods=Deamidated-NG
varMods=Phosphorylated-S
varMods=Phosphorylated-T
varMods=Phosphorylated-Y
varMods=pyroCarbamidomethylCys
instrument_nameMSMS=3
instrument_name=ESI-QEXACTIVE-HCD-v4-35-20
minMatchedPercent=30
parent_mass_convert=monoisotopic
parent_mass_tolerance=20
tolerance_units_precursor=ppm
fragment_mass_tolerance=20
tolerance_units=ppm
max_ms_prod_charge=3
search_type=variable
enable_multiple_mods=1
user_min_parent_shift=-18.0
user_max_parent_shift=272
parent_shift_type=%2B%2F-
parent_shift=130.0
mutation_matrix_off=0
star_ions_off=1
homology_gap_mode=0
unknome=1
dissociationMethod=ALL
spectrumFilenames=%2A.pkl
contaminantProductIonsScoreFilter=4.5
=
```

Parameter file lists:
name=value
pairs
Perl-CGI style

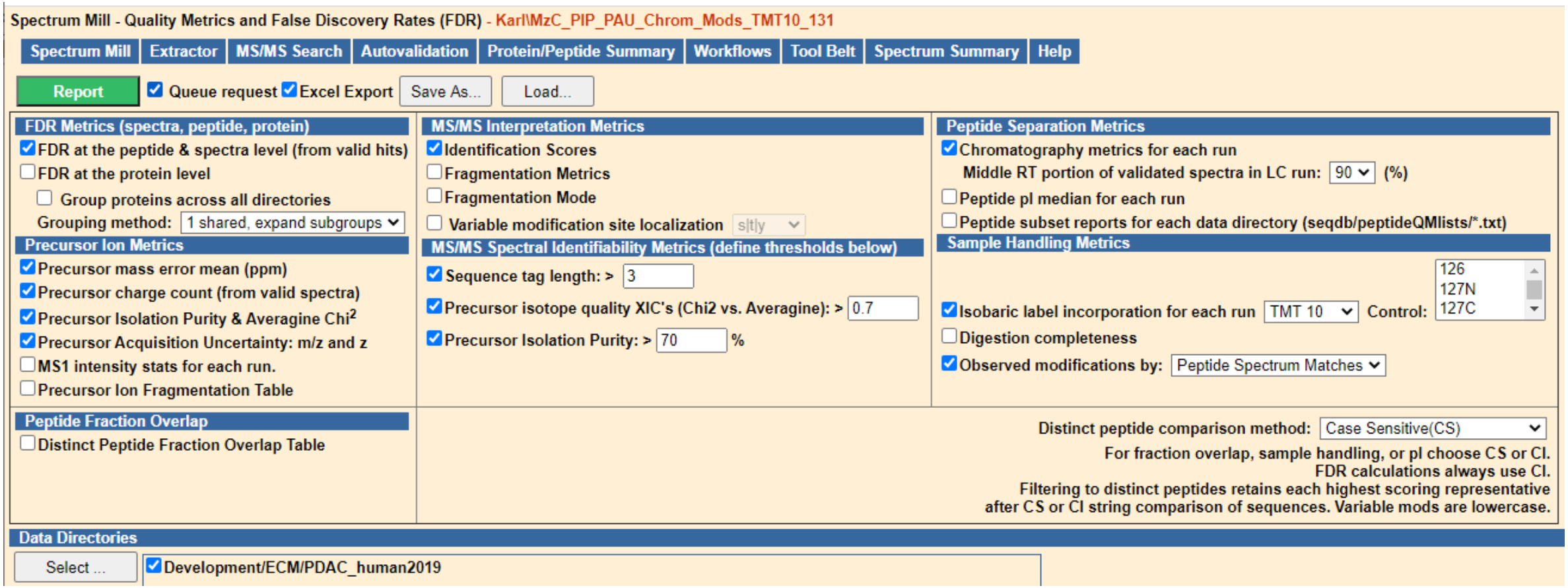
Same as .params files written in each SM data directory

Quality Metrics

Troubleshooting

- LC Gradient and column performance
- Measuring data acquisition strategy changes
- Mass calibration drift
- MS ion optics performance
- Digestion efficiency
- IMAC phospho enrichment
- TMT/iTRAQ labeling efficiency
- Cysteine reduction/alkylation failure

Quality Metrics & FDR



Concepts for some of these metrics were developed and described in: Rudnick PA, Clauser KR, Kilpatrick LE, et. al., "Performance metrics for liquid chromatography-tandem mas spectrometry systems in proteomics analyses", Mol Cell Proteomics. 2010 Feb;9(2):225-41 (http://www.ncbi.nlm.nih.gov/pubmed/19837981).

MS and Chromatography Metrics

Measure effect of changes in acquisition parameters and chromatography on newly installed QExactive Plus at Broad Institute

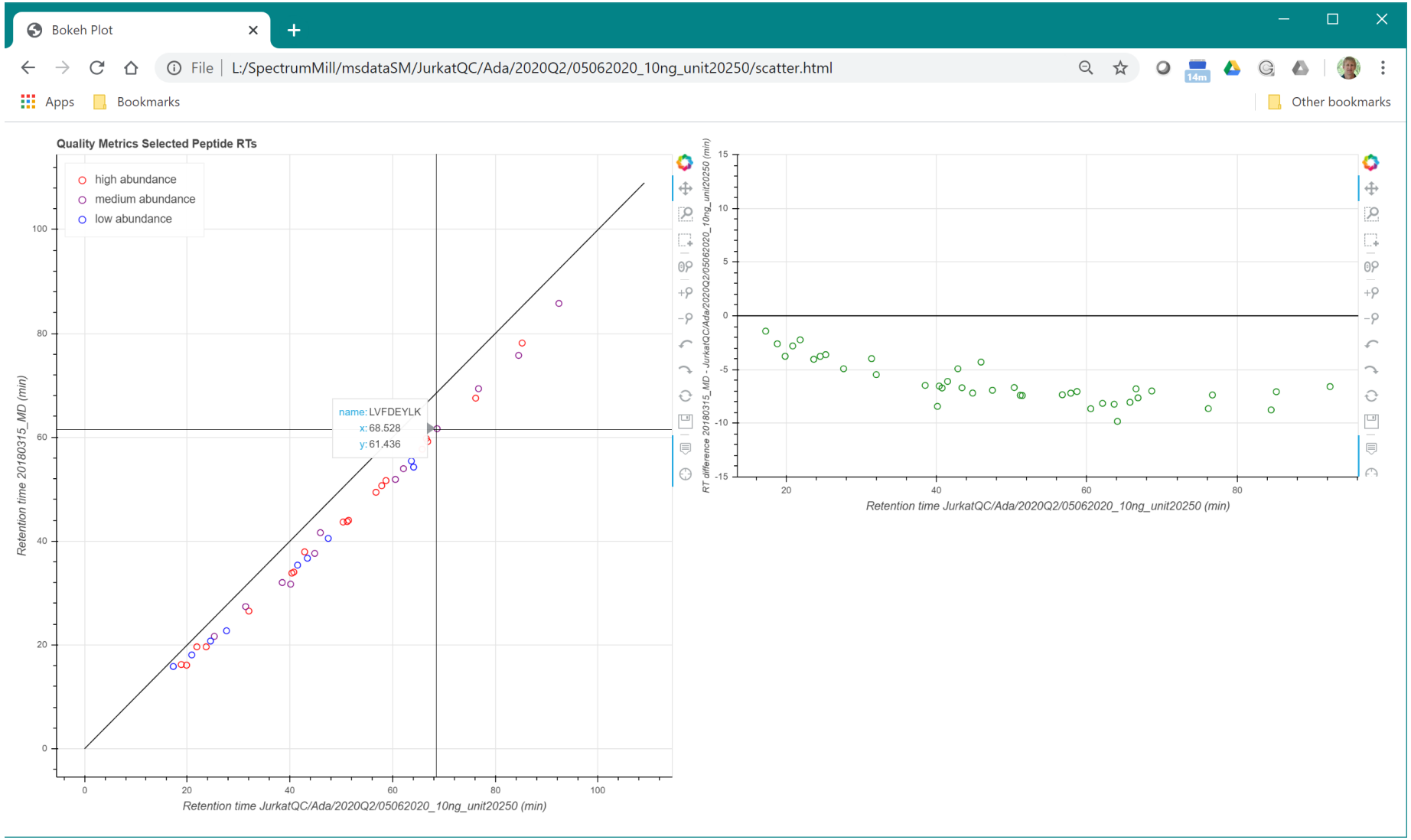
Comment	Pep Match	MS/MS spectra collected C	MS/MS spectra collected MI m/z = 0.0	MS/MS spectra filtered F	MS/MS spectra valid V	MS/MS spectra valid MI m/z = 0.0	Median ID Score	Mean Precursor Mass Error (ppm)	Start time mid 90 % run (min)	End time mid 90 % run (min)	Time span mid 90 % run (min)	Gradient Shape mid 90 % filtered spectra in run	Median MS1 peak width mid 90 % run (sec)	Total precursor XIC mid 90 % matched spectra in run	PIP bin 1 100-90 Spectra (%)	PIP bin 6 <50 Spectra (%)	Distinct Peps CS Total (#)	FDR Spectra (%)
	on	52,270	-	47,892	34,811	-	12.07	0.3	21	98	78	9997754	12	6.15E+13	22.1	9.7	27,929	0.72
Isolation width decreased 2.5 to 2.0	on	51,827	-	49,963	36,876	-	12.68	1.1	17	98	81	9987753	12	6.31E+13	29.7	6.4	29,598	0.73
LC-Klaus	on	49,015	-	47,140	32,487	-	11.73	-0.2	18	98	80	9999876	15	5.35E+13	26.1	9.4	26,176	0.78
	pref	60,705	19,121	57,497	38,127	8,198	11.79	2.2	21	100	79	9897654	11	4.88E+13	27.0	7.6	30,759	0.87
16 min longer gradient, isolation width decreased 2.0 to 1.6	pref	68,003	23,946	63,633	42,998	10,670	12.36	-0.4	20	115	94	8898754	13	6.13E+13	35.0	4.8	34,719	0.96

Pep Match (on/pref)
Precursor isotope cluster

Lengthened
Gradient

Narrowed
isolation width

Quality Metrics - Selected Peptide RTs vs gold standard run



Gold Standard

Current Experiment

Interactive graphics
Python script feeds
Bokeh
javascript library

Namrata Udeshi

SeqDB\peptideQMlists*.txt
Lists of readily observable human peptides in cell lysates.

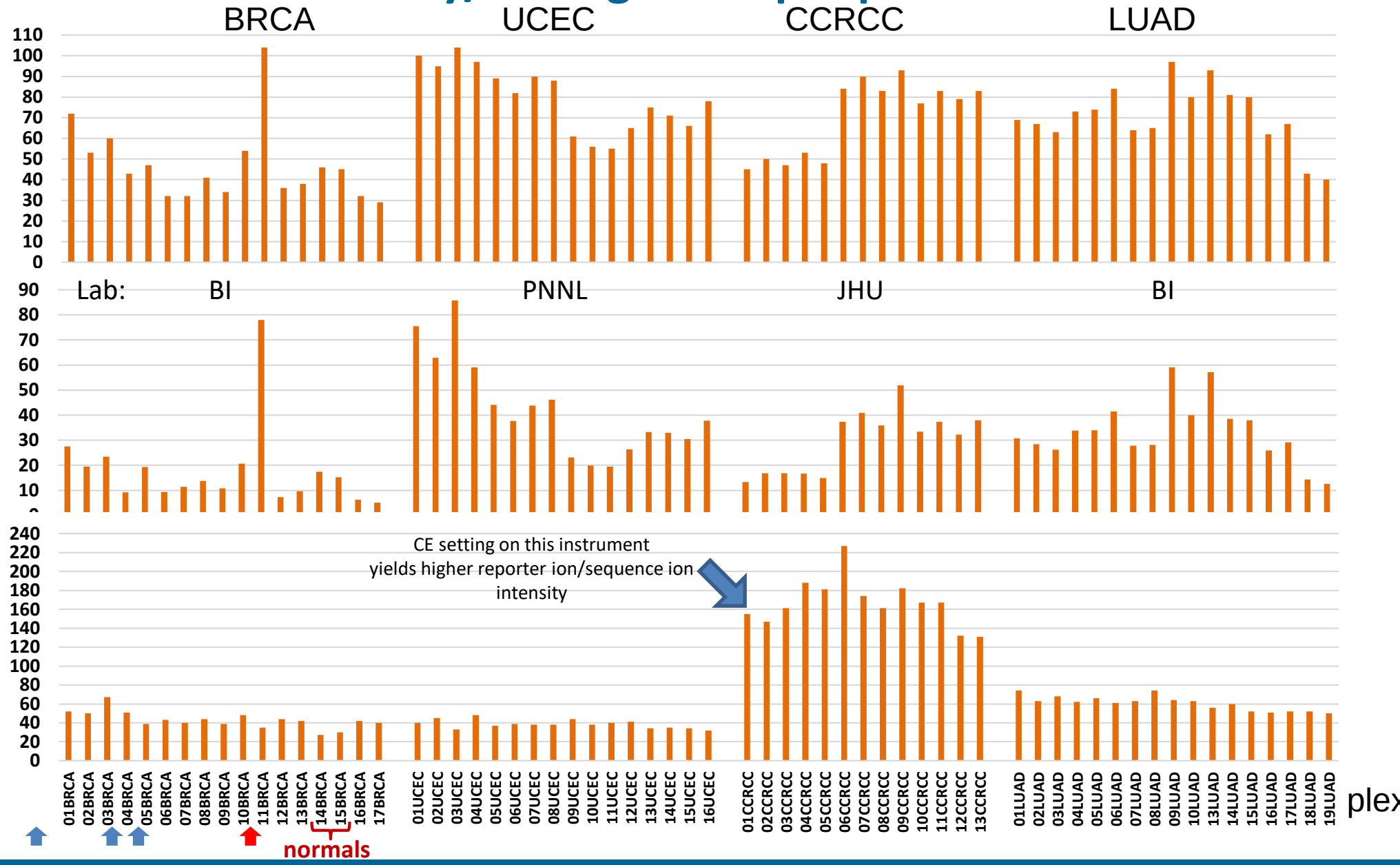
All are Thermo Fusion Lumos

Instrument Sensitivity/Tuning - Phosphoproteome

Median MS2 fill time mid 90% matched spectra (msec)

Spectra Reaching max MS2 fill time mid 90% matched spectra (%)

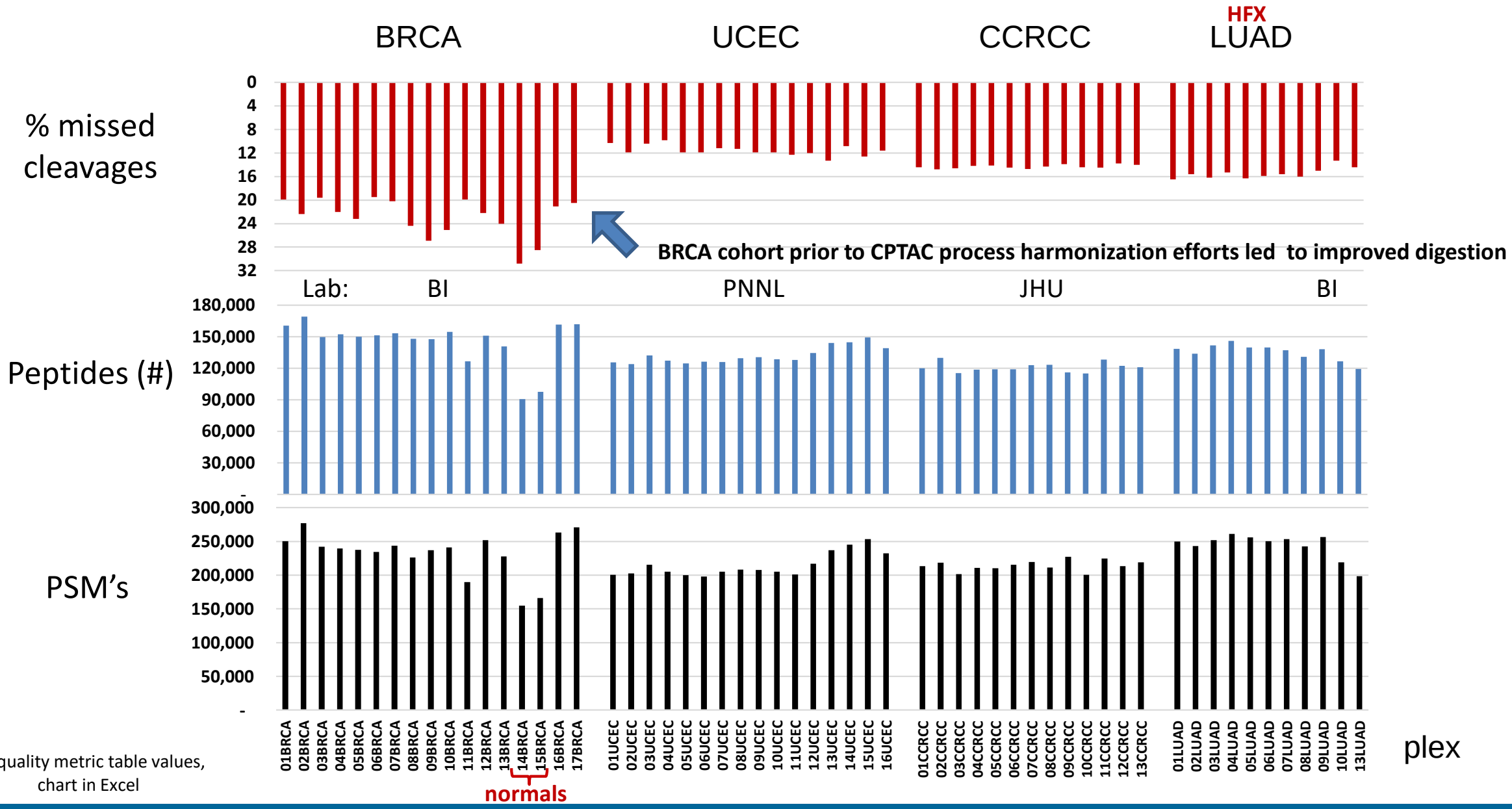
Median S/N All Reporters



Instrument Maintenance

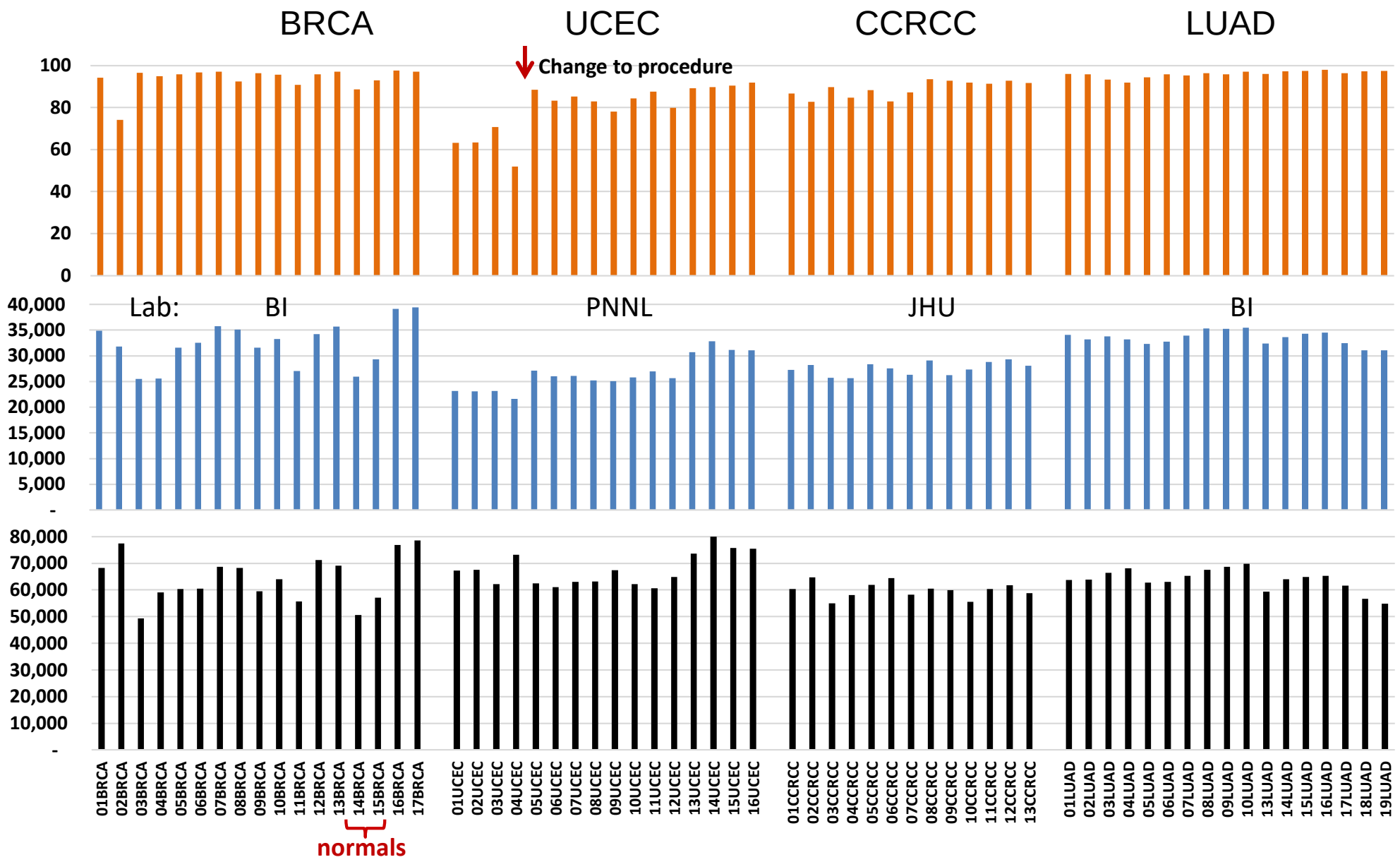
- Ion optics cleaning
- Replaced turbo & ion funnel

Tracking Digestion Efficiency



SM quality metric table values, chart in Excel

Tracking IMAC Phospho -enrichment



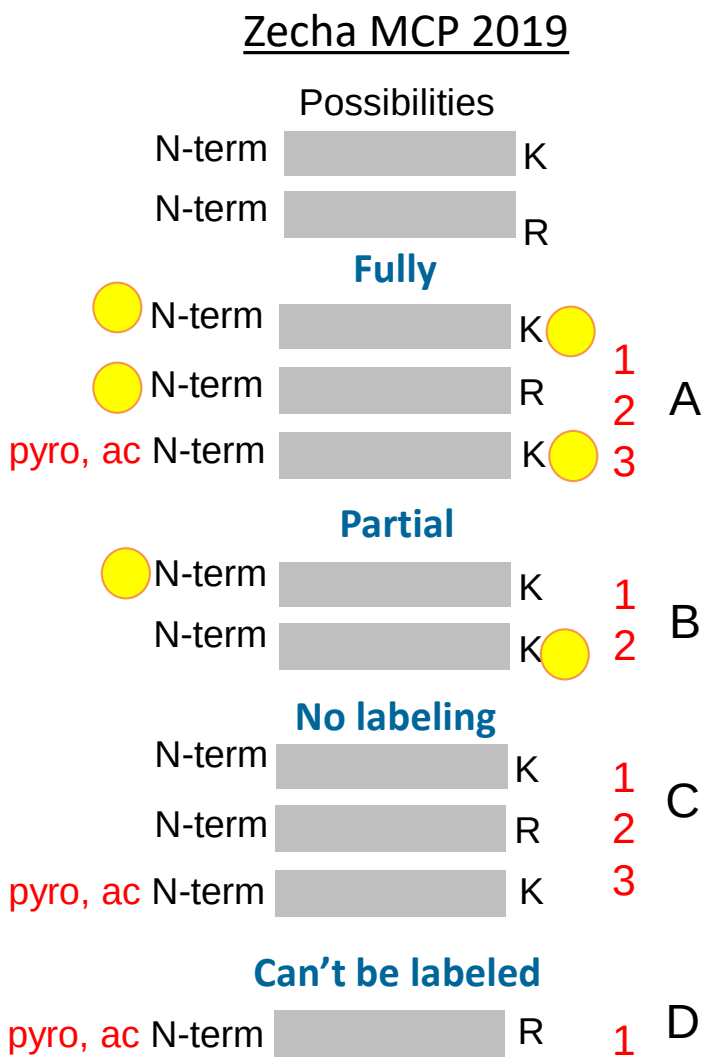
SM quality metric table values, chart in Excel

Phosphorylation Quality Metrics

Directory	Raw Files	MS/MS spectra valid V	s|t|y Sites spectra (%)	s|t|y Sites spectra (#)	s|t|y Sites Localized spectra (%)	s|t|y Sites Distinct Peptides (CI#)	s|t|y Sites Distinct Peptides (CI%)	s|t|y Sites FDR Spectra (%)	s|t|y Sites FDR Distinct Peptide (%)	s|t|y Sites Distinct Peptides (CI#)	FDR Spectra (%)	FDR Distinct Peptide (%)	Distinct Peptides (CI#)
ɔme2/01TCGA	13	77,152	94.5	72,930	56.4	32,179	92.4	0.47	0.92	32,179	0.51	0.98	34,830
ɔme2/02TCGA	13	82,578	90.0	74,288	55.7	31,843	85.9	0.51	1.00	31,843	0.52	0.98	37,072
ɔme2/03TCGA	13	91,894	77.8	71,460	57.0	31,193	75.6	0.48	0.95	31,193	0.50	0.96	41,256
ɔme2/04TCGA	13	98,744	72.6	71,690	57.9	31,921	68.1	0.57	1.12	31,921	0.51	0.96	46,887
ɔme2/05TCGA	13	90,028	83.4	75,068	57.6	31,748	78.2	0.49	1.05	31,748	0.50	0.98	40,620
ɔme2/06TCGA	13	96,310	53.2	51,264	58.7	22,933	51.4	0.64	1.18	22,933	0.51	0.93	44,613
ɔme2/07TCGA	13	83,351	73.0	60,858	58.4	26,668	68.9	0.54	1.03	26,668	0.47	0.86	38,695
ɔme2/08TCGA	13	102,722	66.7	68,521	55.1	30,355	66.3	0.65	1.23	30,355	0.49	0.92	45,811
ɔme2/09TCGA	13	87,016	67.1	58,360	55.5	27,513	65.1	0.62	1.19	27,513	0.51	0.91	42,250
ɔme2/10TCGA	13	88,406	78.5	69,368	55.7	29,632	74.7	0.58	1.15	29,632	0.52	1.00	39,650
ɔme2/11TCGA	13	60,593	92.5	56,029	49.4	25,075	90.0	0.52	0.95	25,075	0.50	0.89	27,864
ɔme2/12TCGA	13	63,858	94.1	60,117	50.5	26,540	92.5	0.45	0.81	26,540	0.45	0.80	28,689
ɔme2/13TCGA	13	81,810	85.2	69,701	49.5	28,186	79.7	0.51	0.93	28,186	0.49	0.86	35,384
ɔme2/14TCGA	13	68,580	90.7	62,196	48.5	26,335	87.6	0.49	0.92	26,335	0.49	0.91	30,077
ɔme2/15TCGA	13	80,970	95.0	76,933	57.2	31,203	92.8	0.47	0.97	31,203	0.50	1.02	33,629
ɔme2/16TCGA	13	83,601	93.1	77,874	56.2	32,908	90.0	0.46	0.91	32,908	0.49	0.91	36,573
ɔme2/17TCGA	13	71,216	94.5	67,305	56.7	29,337	91.9	0.48	0.95	29,337	0.49	0.93	31,907
ɔme2/18TCGA	13	71,290	93.9	66,930	53.0	27,688	91.6	0.47	0.91	27,688	0.49	0.93	30,234
ɔme2/19TCGA	13	85,670	92.5	79,252	55.9	35,012	90.2	0.5	0.98	35,012	0.51	0.97	38,814
ɔme2/20TCGA	13	85,860	94.7	81,314	55.8	35,404	93.1	0.48	0.96	35,404	0.51	1.00	38,046
ɔme2/21TCGA	13	90,324	96.6	87,277	55.3	35,215	95.1	0.47	0.98	35,215	0.50	1.04	37,043
ɔme2/22TCGA	13	83,748	97.5	81,684	56.0	34,594	96.3	0.46	0.97	34,594	0.49	1.01	35,935
ɔme2/23TCGA	13	84,628	94.6	80,032	55.2	31,800	92.1	0.43	0.94	31,800	0.46	0.98	34,537
ɔme2/24TCGA	13	75,699	95.7	72,416	55.4	28,790	93.8	0.46	0.97	28,790	0.48	0.99	30,697

- Track level of phospho enrichment across IMAC experiments
- Separate FDR calculation for phosphopeptides and all peptides

Determining if you should relabel your samples



SM terminology conflicts with Zecha 2019

Labeled – PSMs from A+B

Fully labeled – PSMs from A excluding A3

No label – PSMs from C and D

Partially labeled – PSMs from B

Completely labeled: 100- (Partially labeled + No label)

Metrics to determine if the samples should be relabeled

Underlabeled = B + C = Partially labeled + No label

Completely labeled = 100- (B + C) = 100 -(Partially labeled + No label)

SM Quality Metrics – TMT labeling

ProteomeDirectory	MS/MS spectra valid V	Isobaric Labeled Spectra (%)	Isobaric Fully Labeled Spectra (%)	Isobaric No Label Spectra (%)	Isobaric Partially Labeled Under-labeled both Spectra (%)	Isobaric Completely Labeled Spectra (%)	Isobaric Only Lys Label Spectra (%)	Isobaric Only Nterm Label Under-labeled Lys Spectra (%)	Isobaric Only Lys Label Under-labeled N-term Spectra (%)	Isobaric No Label Under-labeled both Spectra (%)	Isobaric Under-labeled total Spectra (%)	PSMs Containings|t|y (#)
Karl/Zecha_MCP_2019/ProteomeOverlabel/01Pr	305,976	97.2	93.6	2.8	2.6	94.6	3.1	0.5	2.1	0.3	2.9	34855
Karl/Zecha_MCP_2019/ProteomeOverlabel/Com	296,635	96.0	91.6	4.0	3.5	92.6	3.8	0.6	2.8	0.5	3.9	34391

Proteome	PSMs	F	P	U	O	DP	FDR
Mertins	305,976	94.6	2.6	0.3	11.4	163,322	0.6
Zecha	296,635	92.6	3.5	0.5	11.6	168,606	0.6

PhosphoDirectory	MS/MS spectra valid V	Isobaric Labeled Spectra (%)	Isobaric Fully Labeled Spectra (%)	Isobaric No Label Spectra (%)	Isobaric Partially Labeled Under-labeled both Spectra (%)	Isobaric Completely Labeled Spectra (%)	Isobaric Only Lys Label Spectra (%)	Isobaric Only Nterm Label Under-labeled Lys Spectra (%)	Isobaric Only Lys Label Under-labeled N-term Spectra (%)	Isobaric No Label Under-labeled both Spectra (%)	Isobaric Under-labeled total Spectra (%)
Karl/Zecha_MCP_2019/Phosphoproteome/01Ph	88,921	98.1	92.5	1.9	3.9	94.2	4.8	0.7	3.2	0.3	4.1
Karl/Zecha_MCP_2019/Phosphoproteome/Com	90,734	97.2	89.3	2.9	6.2	91.0	7.0	0.8	5.4	0.3	6.4
						P	F				U

Phospho	PSMs	F	P	U	O	DP	FDR
Mertins	88,921	94.2	3.9	0.3	-	42,124	0.9
Zecha	90,734	91.0	6.2	0.3	-	42,785	0.8

F	PSMs with all available amine groups are labeled. (includes fully labeled, blocked N-terms)
P	PSMs with a label, but also unlabeled amine on N-term or Lysine.
U	PSMs with no label, but have unlabeled amine on N-term or Lysine.
O	PSMs with a label on Ser,Thr, Tyr; in peptides that contain His

Fully labeled - old

PSMs with an N-terminal label and a label on Lys (if present).

Excludes blocked N-termini (acetyl or pyro) on peptides with labeled lysines.

Completely labeled - new

100 - No label - Partially labeled

Includes blocked N-termini (acetyl or pyro) on peptides with labeled lysines.

Too confusing!
should eliminate column: Fully

Tumor/NAT – Evaluate Relative Peptide Load Across Samples

LSCC

plex	126C_ 131C	127N_ 131C	127C_ 131C	128N_ 131C	128C_ 131C	129N_ 131C	129C_ 131C	130N_ 131C	130C_ 131C	131N_ 131C
1	1.38	0.97	1.07	0.89	1.09	1.03	0.97	0.99	1.25	0.94
2	1.14	0.99	1.35	1.07	1.11	0.84	0.94	1.02	1.25	0.94
3	1.17	1.14	1.12	0.92	1.04	0.98	1.05	1.00	0.95	1.04
4	1.16	1.18	1.09	0.99	1.07	0.97	1.05	1.00	1.03	1.05
5	1.13	1.17	1.05	0.86	0.97	1.00	0.90	0.99	1.05	1.07
6	1.17	0.87	1.07	0.85	1.05	0.87	1.10	0.93	1.22	0.89
7	1.10	1.04	1.14	1.12	1.19	0.96	1.01	1.00	1.08	1.10
8	1.00	0.97	0.91	0.91	1.00	0.86	0.96	0.89	0.97	0.92
9	1.17	1.10	1.24	0.99	1.02	0.99	1.08	1.02	1.08	0.94
10	0.95	1.02	1.15	1.01	0.94	1.05	1.05	1.02	1.01	0.96
11	1.17	0.91	1.13	1.04	1.06	1.09	1.10	0.82	1.07	0.95
12	1.09	1.05	0.96	0.98	1.09	1.03	0.98	1.00	1.10	1.03
13	1.08	1.01	1.13	1.04	1.19	1.09	1.12	1.02	1.12	1.09
14	1.13	1.00	1.14	0.97	0.99	0.93	0.92	0.90	0.92	0.85
15	1.10	1.09	1.05	0.95	1.09	1.02	1.15	0.99	1.32	1.06
16	1.07	0.94	1.02	0.95	1.14	0.96	1.08	0.92	1.03	0.90
17	1.21	1.10	1.09	0.98	1.04	1.00	1.07	0.93	0.93	0.93
18	1.31	1.08	1.27	1.10	1.09	0.94	1.15	0.89	1.15	0.97
19	1.14	1.07	1.15	1.03	1.05	1.01	1.05	1.01	1.04	0.84
20	1.07	0.99	1.06	0.97	1.21	0.96	1.09	0.99	1.07	0.91
21	1.17	1.03	1.16	1.13	1.10	1.00	1.07	0.97	1.08	1.09
22	1.16	1.04	1.20	0.81	1.09	1.01	1.25	0.98	1.08	0.98

LUAD

126C_ 131	127N_ 131	127C_ 131	128N_ 131	128C_ 131	129N_ 131	129C_ 131	130N_ 131	130C_ 131	plex
1.30	0.98	0.90	0.92	0.99	0.97	1.07	0.90	1.21	1
1.50	1.10	1.17	0.67	0.89	0.80	1.08	1.14	0.99	2
1.28	1.16	1.17	1.02	1.08	0.90	1.10	1.03	1.12	3
1.26	1.01	1.25	1.11	1.00	0.91	1.15	1.01	1.00	4
1.33	1.26	0.83	0.91	0.93	0.81	1.12	0.97	1.12	5
1.20	1.05	1.09	1.10	0.89	1.03	1.21	1.00	1.34	6
1.27	1.11	1.15	0.93	1.19	0.81	1.12	0.90	0.89	7
1.24	1.31	1.06	0.96	1.05	0.88	1.01	1.00	1.08	8
1.17	0.59	1.11	0.92	0.90	0.80	0.95	0.56	1.08	9
1.22	1.07	1.14	0.90	0.91	0.88	1.01	0.87	1.12	10
1.21	1.16	1.22	0.83	1.13	0.70	1.16	0.71	0.92	11
1.01	0.88	1.04	0.68	1.16	0.99	1.03	0.64	0.90	12
1.09	0.98	0.82	0.87	0.90	0.96	0.98	0.85	0.94	13
1.16	1.16	0.98	0.91	1.12	0.84	1.05	0.99	1.10	14
1.35	1.04	1.13	1.01	1.15	0.90	1.11	1.01	1.21	15
1.41	1.03	1.13	0.95	1.18	0.93	1.07	0.95	1.01	16
1.11	1.11	1.20	1.15	1.05	1.08	1.08	0.92	1.10	17
1.16	1.08	1.06	1.07	1.07	0.87	1.00	0.66	1.06	18
1.29	1.14	1.20	1.01	0.97	0.64	1.11	0.93	1.31	19
1.30	1.25	1.24	1.13	1.19	0.94	0.99	0.89	1.07	20
1.47	1.14	1.18	0.96	1.20	0.87	1.15	1.03	1.22	21
1.37	0.99	1.29	0.96	1.08	0.57	1.43	0.77	1.46	22
1.22	0.84	1.09	0.79	0.93	0.80	1.24	0.99	1.18	23
1.00	1.02	1.06	0.88	1.02	0.60	1.11	0.86	0.82	24
0.96	0.86	1.09	0.85	1.39	0.89	1.14	0.90	0.92	25

Proteome

LSCC tighter thresholds applied for mixing controls +/- 15% (LSCC) from CR instead of +/-25%(LUAD)

plex	126C_	127N_	127C_	128N_	128C_	129N_	129C_	130N_	130C_	131N_
1	1.83	0.63	1.27	0.72	1.11	0.83	1.08	0.90	1.64	0.88
2	0.89	0.48	1.58	0.96	1.26	0.49	1.19	0.91	1.75	0.92
3	1.13	0.99	1.00	0.47	1.15	1.02	1.02	0.48	0.94	1.18
4	1.37	1.03	0.91	0.71	1.28	0.81	1.13	0.72	1.10	0.87
5	1.25	0.97	0.68	0.55	1.08	0.97	0.63	0.73	1.08	0.95
6	0.94	0.60	1.17	0.46	1.22	0.80	1.15	0.86	1.41	0.76
7	1.32	0.73	1.22	1.16	1.41	0.70	1.13	0.84	1.16	1.23
8	0.80	0.79	0.75	0.82	1.10	0.69	0.83	0.55	0.90	0.82
9	1.13	1.12	1.83	0.90	1.11	0.80	1.33	0.87	1.48	0.88
10	1.32	0.95	1.20	0.79	1.04	0.94	1.17	0.98	1.18	0.76
11	1.25	0.57	1.34	0.82	1.08	1.10	1.20	0.48	1.02	0.82
12	0.91	0.73	1.15	0.81	0.96	0.72	1.07	0.83	1.34	1.17
13	1.11	0.79	1.43	0.99	1.39	1.13	1.31	0.92	1.23	1.21
14	1.33	0.77	1.32	0.84	1.23	0.91	1.22	0.85	0.99	0.87
15	1.07	0.81	0.94	0.78	1.28	0.90	1.28	0.80	1.70	1.16
16	0.97	0.78	0.96	0.77	1.24	0.76	1.53	0.77	1.26	0.86
17	1.34	1.12	1.29	0.82	1.24	0.89	1.48	0.80	0.96	0.81
18	1.73	0.95	0.90	1.12	1.07	0.80	1.22	0.55	1.53	0.78
19	0.87	0.65	1.37	0.97	1.21	0.91	1.03	0.78	1.26	0.80
20	1.15	0.87	0.85	0.72	1.43	0.64	1.14	0.79	1.31	0.76
21	1.33	0.77	0.97	1.06	1.32	0.89	1.17	0.61	0.78	1.16
22	1.17	0.91	1.47	0.45	1.09	0.91	1.65	0.83	1.21	0.88
	T	N	T	N	T	N	T	N	T	T/N

Phosphoproteome

Tumors tend to yield more phosphorylated material than NATs

126C_	127N_	127C_	128N_	128C_	129N_	129C_	130N_	130C_	plex
1.38	0.68	0.69	0.77	1.03	0.93	1.23	0.79	1.42	1
1.78	0.72	1.17	0.48	0.75	0.42	1.31	0.95	0.77	2
1.28	0.86	1.32	0.64	1.08	0.51	1.50	0.97	1.33	3
1.02	0.62	1.63	0.87	1.16	0.95	1.36	0.89	0.78	4
1.18	0.85	0.81	0.80	0.71	0.48	1.01	0.84	1.33	5
1.05	0.63	1.41	0.82	0.87	0.73	1.26	0.80	1.55	6
1.63	0.85	0.88	0.58	1.24	0.49	1.28	0.65	0.98	7
1.43	0.92	1.12	0.86	1.21	0.72	1.16	0.63	1.36	8
1.34	0.37	1.39	0.72	1.22	0.67	1.07	0.25	1.29	9
1.03	0.76	1.40	0.61	1.28	0.66	0.92	0.39	1.70	10
1.19	1.01	1.30	0.59	1.15	0.53	1.66	0.35	0.73	11
1.21	0.61	1.38	0.35	1.46	0.90	0.83	0.27	0.91	12
1.12	0.36	0.58	0.44	1.13	0.90	1.45	0.65	1.18	13
1.24	0.98	1.15	0.53	1.47	0.63	1.27	0.75	1.32	14
1.85	0.87	1.48	0.86	1.56	0.89	1.11	0.93	1.41	15
1.91	1.08	1.74	0.59	1.75	0.69	1.39	0.67	1.25	16
1.13	0.97	1.52	0.95	1.21	1.02	1.38	0.87	1.65	17
1.14	0.88	1.48	0.93	1.14	0.77	1.04	0.35	0.89	18
1.33	1.06	1.29	0.89	1.04	0.34	1.03	0.83	0.88	19
1.34	1.24	1.76	1.14	1.66	0.90	1.33	0.80	1.30	20
1.68	1.03	1.23	0.62	1.20	0.69	1.08	1.03	1.37	21
1.54	0.80	1.64	0.98	0.91	0.28	1.65	0.69	1.20	22
1.59	0.60	1.46	0.57	1.15	0.71	1.77	0.80	1.52	23
0.95	0.87	1.03	0.75	1.32	0.39	1.01	0.74	0.70	24
1.15	0.46	1.18	0.82	1.25	0.82	1.48	0.74	0.74	25
T	N	T	N	T	N	T	N	T/N	

SM quality metric – TMT channel balance

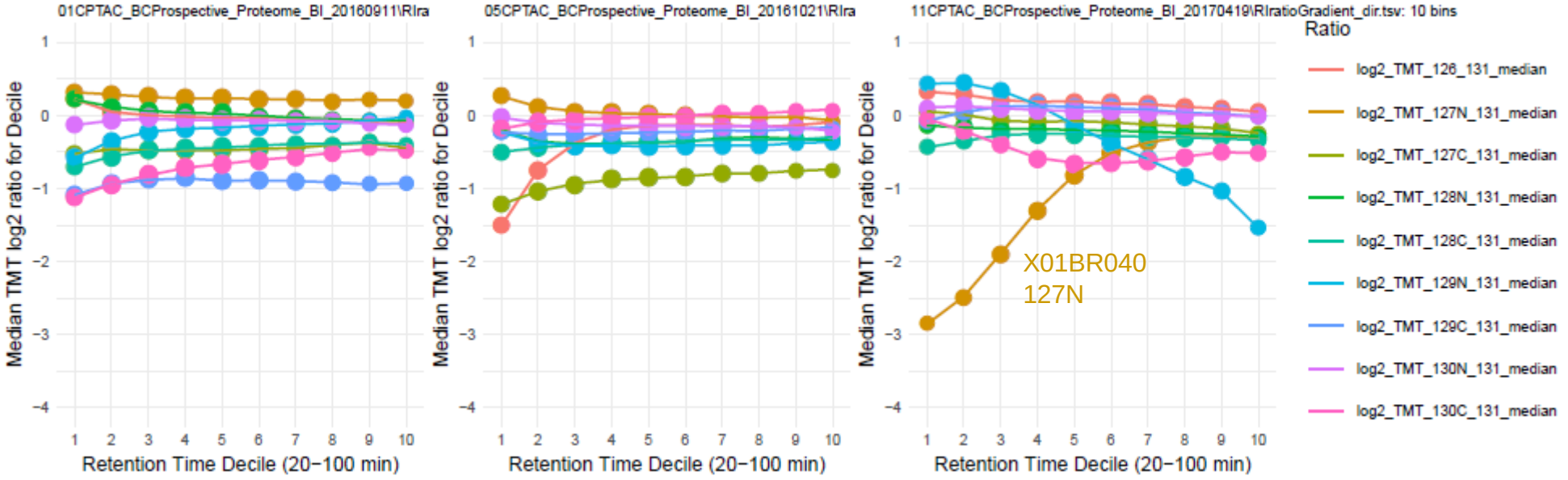
PSM level reporter ion ratio
(summed intensity of all PSMs
Each reporter ion/common ref)

SM quality metric table values, coloring done in Excel

SM Quality Metrics Plots – TMT Ratio vs Retention Time Decile

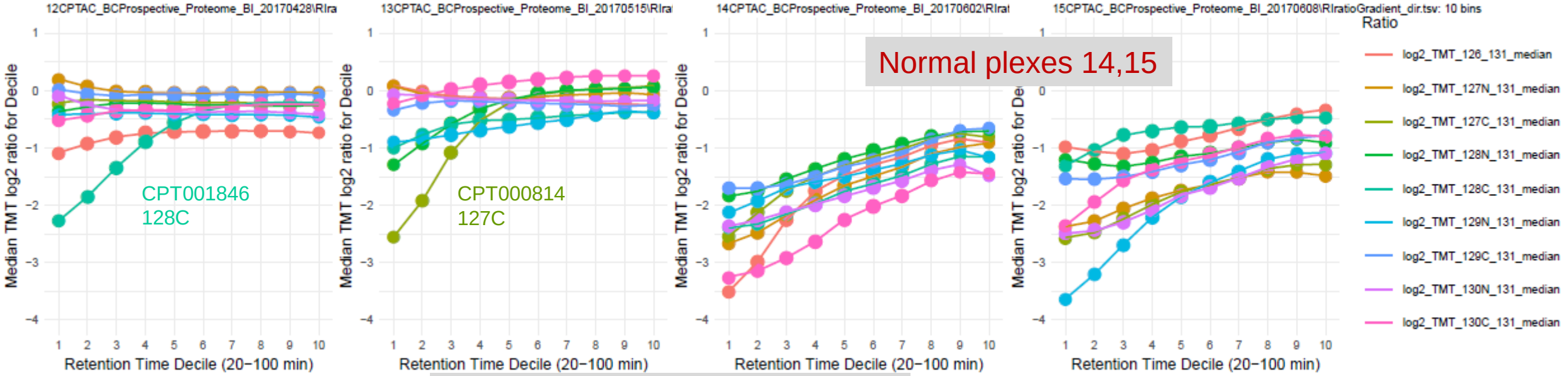
Proteome

Tumor plexes 1,5,11-13



All the LC-MS/MS runs were split up into retention time deciles. Ie from the RT range 20-100 minutes, 10 bins, each bin is 8 minutes wide. Then take the median ratio for each TMT channel, in each RT bin, from only the PSMs of identified MS/MS spectra. That median is across all the fractions for a plex.

Point area scaled to # PSMs/decile



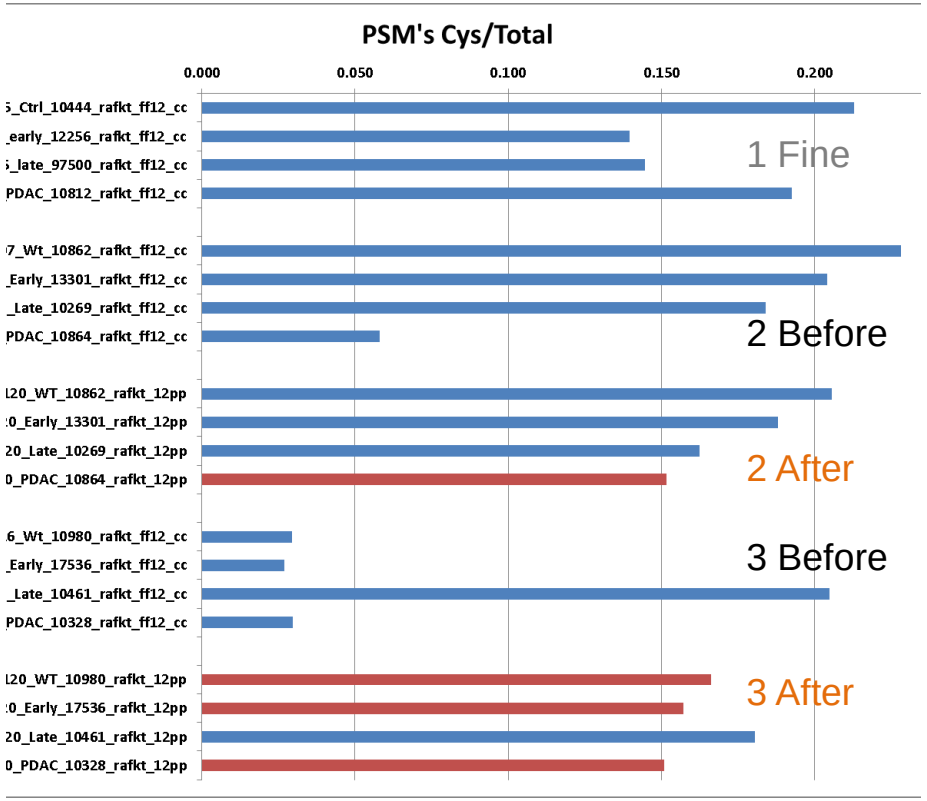
Deciles combined across all 25 bRP fractions

Plot made when QM run with isobaric label incorporation metrics on

Reduction/Alkylation Problem

PSMs Containing no Var Mods (#)	PSMs Containing m (#)	PSMs Containing c (#)	PSMs Containing p (#)	PSMs Containing C (#)	PSM's Cys/Total
5101	285	133	918	1527	0.213
3857	219	61	1051	778	0.140
4147	225	77	1107	866	0.145
5852	309	129	1178	1580	0.193
5204	522	223	1247	1855	0.228
4785	470	185	1072	1475	0.204
4161	287	157	1212	1170	0.184
6395	424	26	1371	533	0.058
7617	1199	303	1752	2431	0.206
7408	1078	259	1469	2084	0.188
6685	764	212	1524	1613	0.162
14035	946	113	1770	2770	0.152
1184	200	2	1144	77	0.029
1928	258	3	1004	91	0.027
4483	580	117	1381	1495	0.205
3684	447	5	1323	177	0.030
4017	434	22	1571	1061	0.166
5540	578	28	1434	1273	0.157
12714	1058	234	2129	3127	0.180
11774	936	70	1934	2376	0.151

Reduction/alkylation repeated due to low #'s of PSM's Containing Cys



SM quality metric table values, coloring & plotting done in Excel

Future Directions

- Process Report
  - ✓ Export GCT, vertical reporter-sample template
  - Export of reporter ion intensity, de-multiplex precursor intensity.
- Quality Metrics
  - More graphic display of metrics
  - Report FDR at VM-site level
  - Integrate Retention Time prediction
- Autovalidation
  - Integrate Percolator/ML for more sensitive use of SM scores

✓ Personalized sequence databases – routine use for SAAV's and spliceforms
  - ✓ Generalize features that are hard-coded for CPTAC
  - ✓ Automate conversion of .fa/.maf for neoantigen candidates
  - Support SAAV's that are common in the population
    - minor allele frequency > 0.5%
    - ICGC-TCGA pancancer shared mutations

- 2-10x faster searches for long peptides with many variable mods.
  - faster fragment ion pre-matching to more rapidly exclude poor matching sequences.
- Fully integrate Sherenga de novo MS/MS interpretation
  - improve accuracy of HLA peptide antigen identification
  - optimize FDR in large-scale projects.
- Spectrum Matcher - propagate identifications of related spectra
- Protein Quantitation
  - dynamic allocation of shared peptides, beyond subgroup-specific, subgroup-top
  - automate identification of isoform regulation (exon splicing, pre-propeptide processing)
  - store protein grouping for faster report re-generation
- Elastic processing capacity
  - Try Google Cloud-based VMs, evaluate access to and I/O time for extraction from Google Cloud Storage bucket.
    - Create SM server image for virtual-distribution
  - Dockerize key modules of SM to enable Terra integration and more facile processing of CPTAC data from multiple PCC's.

Load Instrument Gauges

Spectrum Mill - Quality Metrics Dashboard

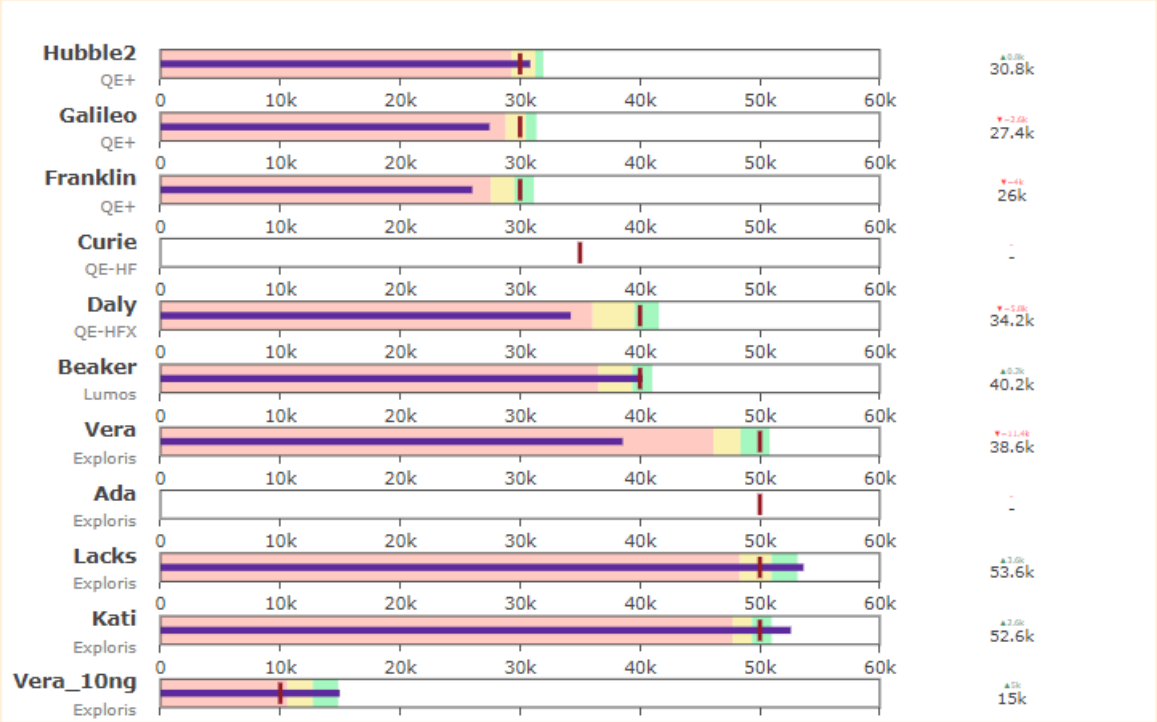
Spectrum Mill | Quality Metrics | Help - slides covering overall design

The following instruments are fully populated with metrics from start of Q4-2020 to present: Franklin, Galileo, Hubble2, Kati, Lacks, Beaker, Daly, Vera, Vera_10ng

Choose an option: Harvest Instrument QM Logs | Load Instrument Gauges | Graph Metrics

Instrument Jurkat QC Status

Gauges show latest total peptide count with respect to historical value ranges of 2nd quartile and 3rd quartile.



Bullet chart above produced by Plotly JavaScript library

Consider switching the underlying code to D3 instead of Plotly

Graph Metrics

Spectrum Mill - Quality Metrics Dashboard

Spectrum Mill | Quality Metrics | Help - slides covering overall design

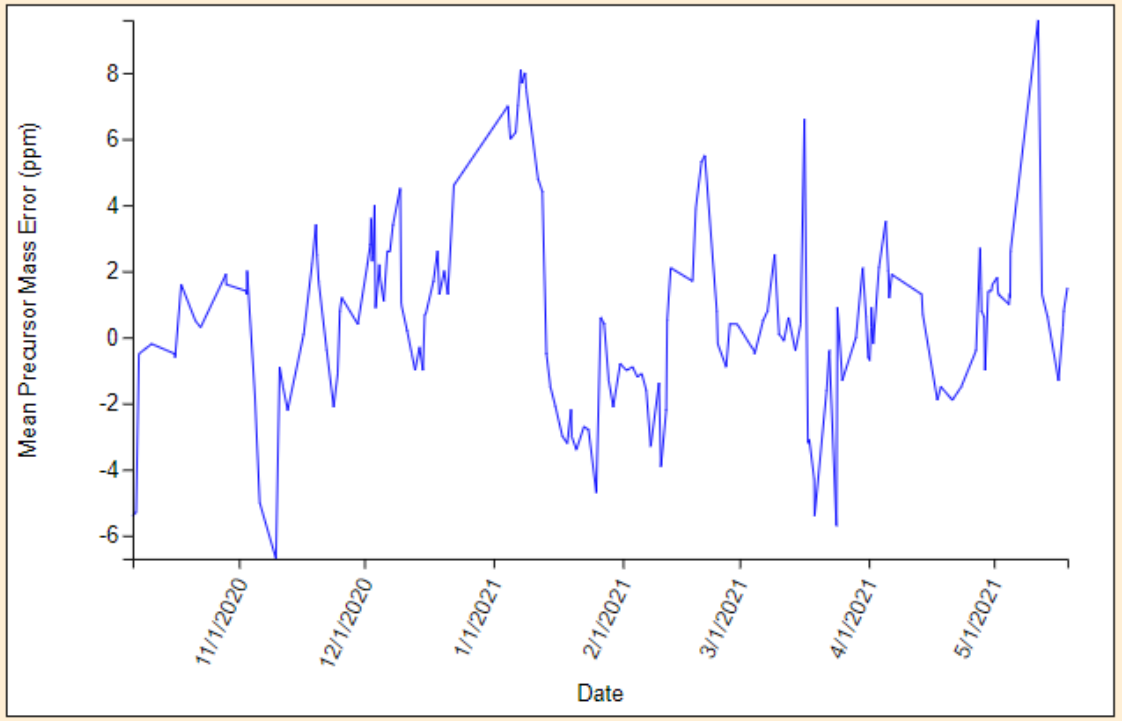
The following instruments are fully populated with metrics from start of Q4-2020 to present: Franklin, Galileo, Hubble2, Kati, Lacks, Beaker, Daly

Choose an option: Harvest Instrument QM Logs | Load Instrument Gauges | Graph Metrics

Line Charts

Click and drag to zoom. Double click to reset zoom.

Instrument: Lacks Quality Metric: Mean Precursor Mass Error (ppm)



Line chart produced by D3 JavaScript library using Support Vector Graphics (SVG)

Report to Plots - prototype

Not secure | manzano/millhtml/reportToPlots.htm

Apps Bookmarks Other bookmarks

Spectrum Mill - Report To Plots - HLA\PSM_z1-3_len8-11

Spectrum Mill | Extractor | MS/MS Search | Autovalidation | Quality Metrics & FDR | Protein/Peptide Summary | Workflows | Tool Belt | Protigy | Help

Report to Plots

Make Plots | Save As... | Load...

Data Directories

Select ...

☑ Karl/PatientsIP_snvIndel_nuORF_v3/Mel02_13240-002_20190419_10IPs

Peptide Filtering

Min Precursor Charge: 1 ▼ Max Precursor Charge: 3 ▼

Min Sequence Length: 8 ▼ Max Sequence Length: 11 ▼

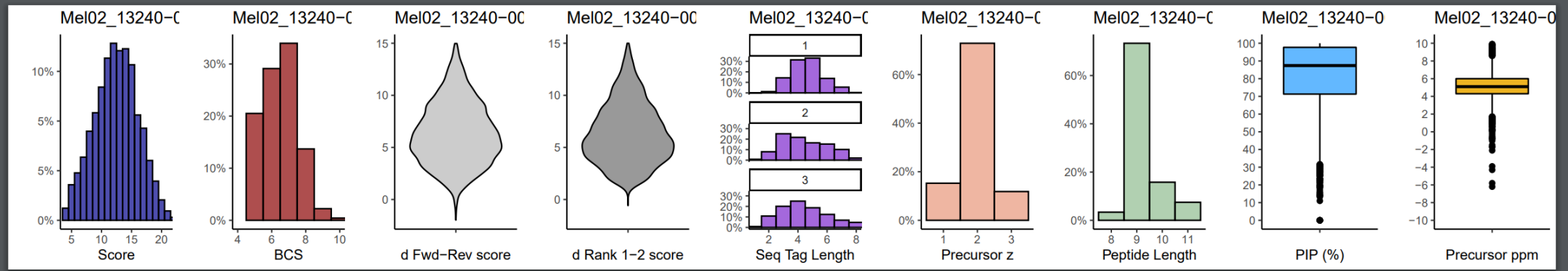
Plot Parameters

SM report type: PSM ▼

☑ Plot Score Metrics

Results

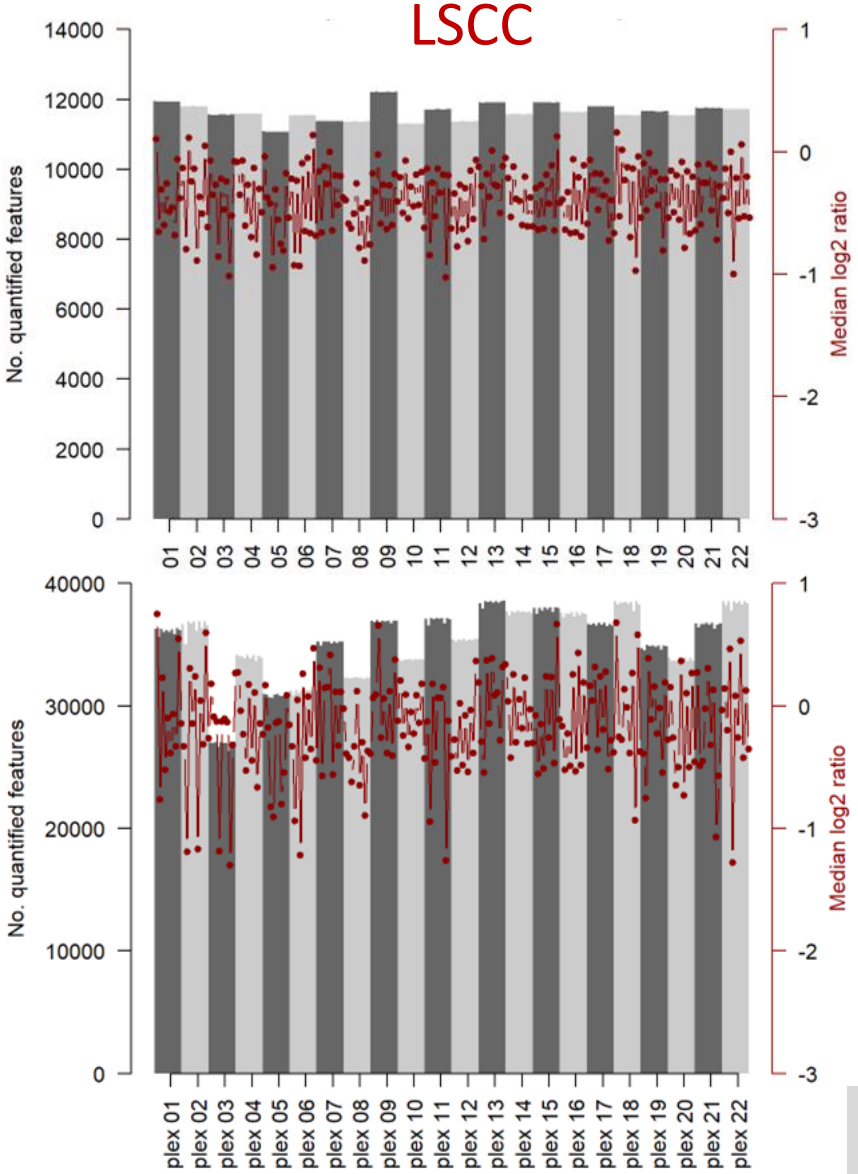
R Graphics Output 1 / 1



Similar Depth of Proteins & Phosphosites per TMT plex & channel

LSCC more uniform loading per TMT channel

LSCC



Proteins

Median # / plex

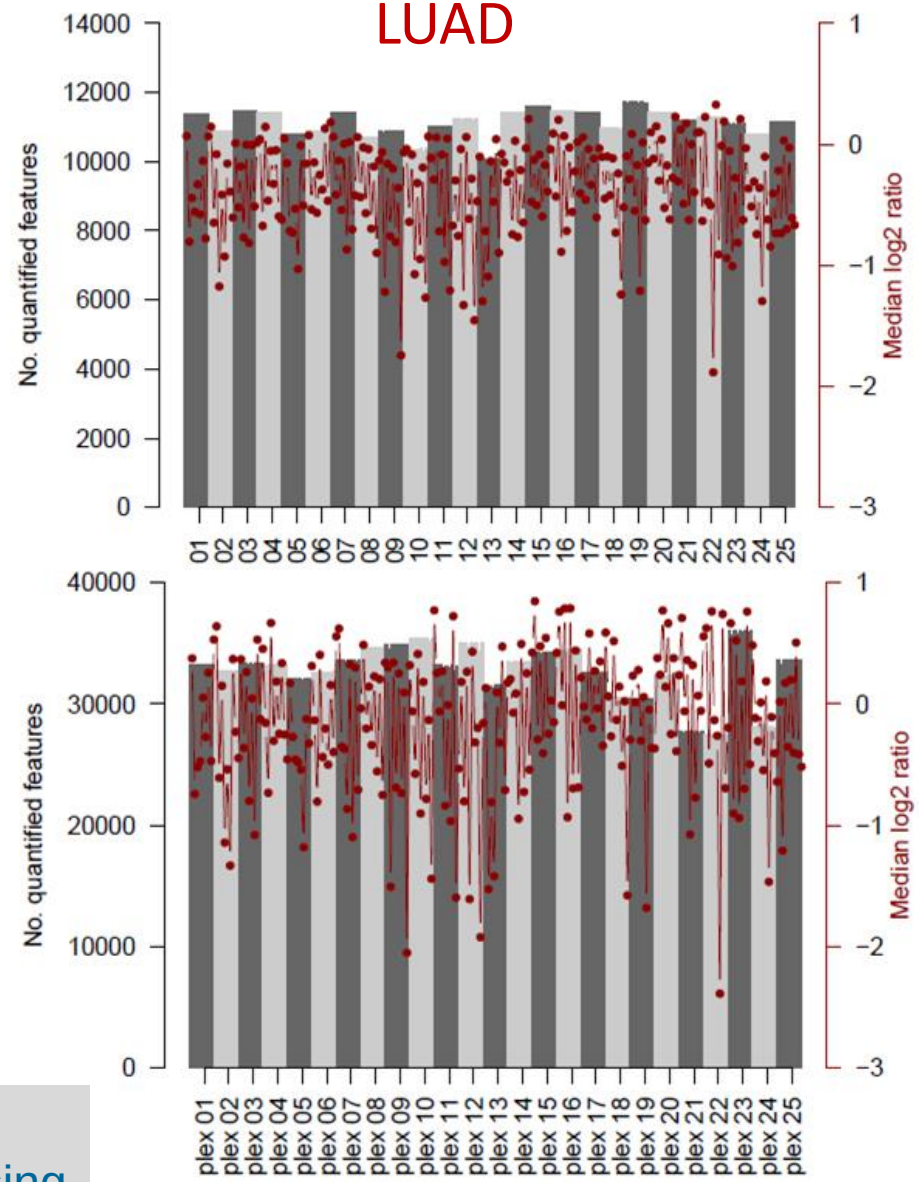
11,612

Phospho Sites

Median # / plex

36,282

LUAD



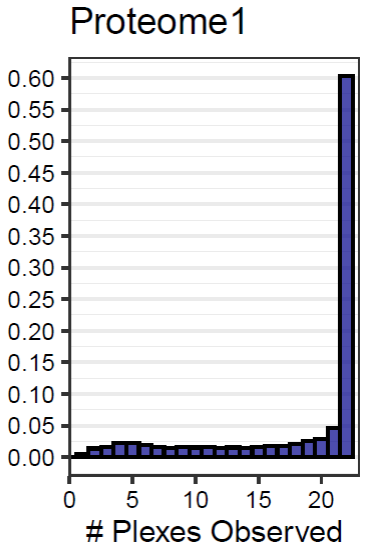
11,180

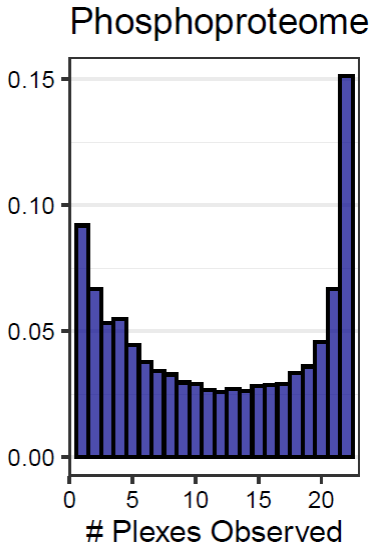
33,245

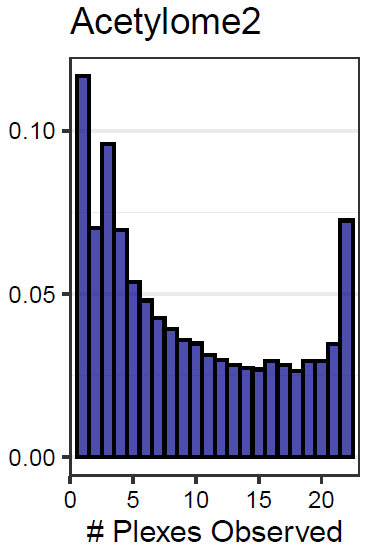
Generalize hardcoded CPTAC directory parsing

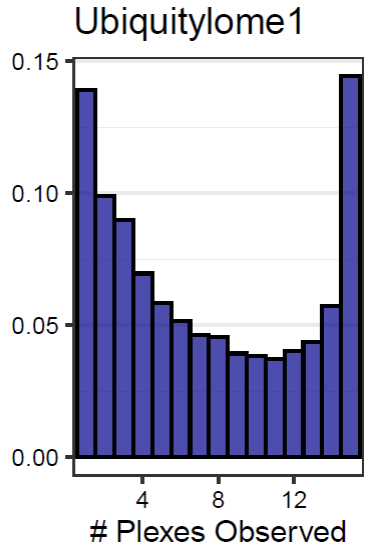
Missing Value Histograms - prototype

LSCC



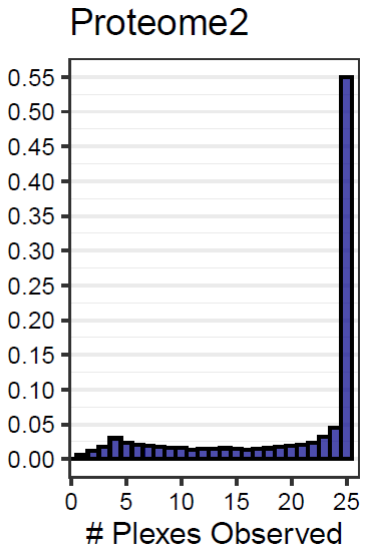


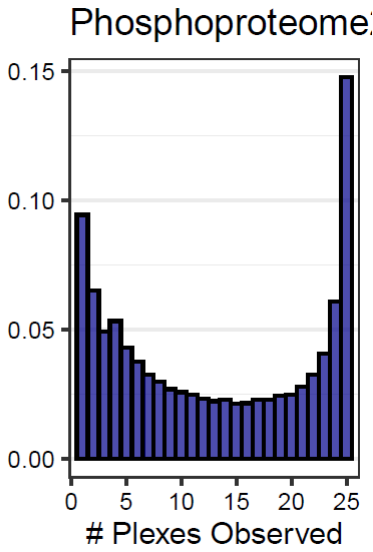


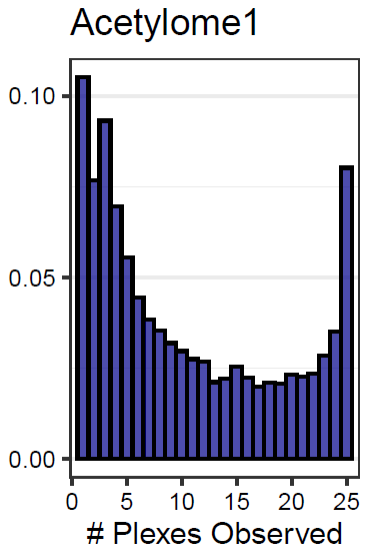


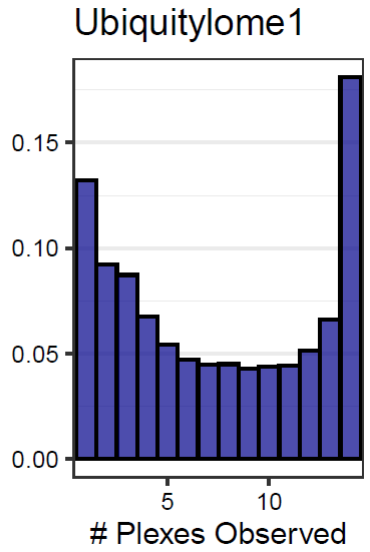
LSCC

LUAD









LUAD

Acknowledgments

Vlado Dancik
Broad Institute

Joe Roark
Agilent Technologies

- DR Mani
- Karsten Krug
- Jake Jaffe
- Namrata Udeshi