Data Extraction and Analysis for LC-MS Based Proteomics

Instructors

Gordon Anderson, Charles Ansong, Matthew Monroe, and Ashoka Polpitiya

Pacific Northwest National Laboratory, Richland, WA 99354

Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by Battelle *Since 1965*

Course Outline

- Part I: Introduction and Overview of Label-Free Quantitative Proteomics (Anderson)
  - Goals
  - Data and Tools Availability
  - Quantitative Proteomics: Historical Perspective
- Part II: Feature Discovery in LC-MS Datasets (Monroe and Polpitiya)
- Break
- Part III: Biological Application of the AMT tag Approach (Ansong)
- AMT tag Analysis Software Demo
- Panel Discussion
  - Questions
  - Future Directions

Course Goals

- Understand the reasons for developing and applying an LC-MS-based approach to proteomics
- Discuss considerations of experimental design for larger scale experiments
- Develop a sense of the source of information, its relative complexity and the algorithms required to make use of this approach
- See (and participate) in a demonstration of the critical tools applied to “real” data
- Learn where to get more information

Pacific Northwest National Laboratory

Environmental Molecular Sciences Laboratory

Washington Wine Country

Pacific Northwest National Laboratory and EMSL

PNNL performs basic and applied research to deliver energy, environmental, and national security solutions for our nation.

W.R. Wiley Environmental Molecular Sciences Laboratory

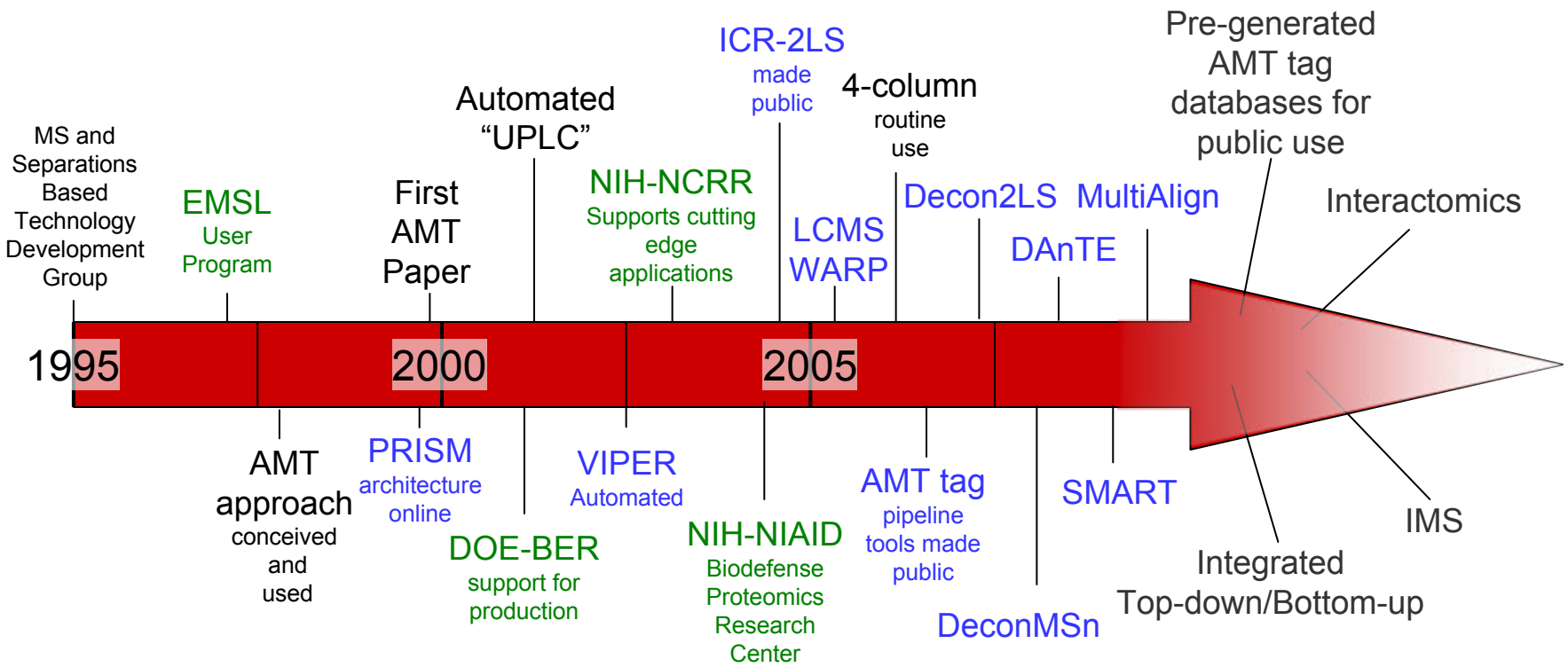
The Guest House at PNNL for EMSL Users

EMSL Mission

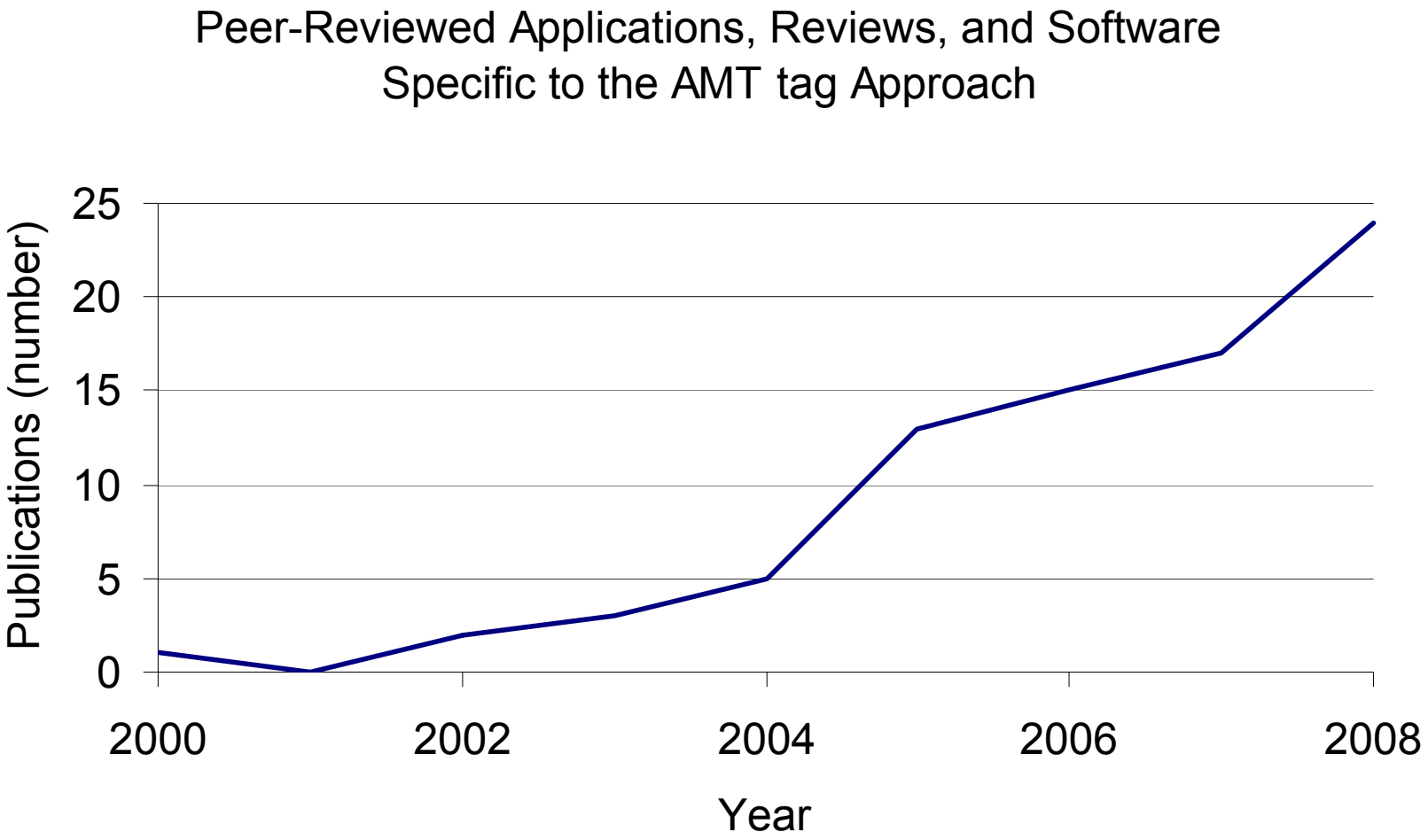
The W.R. Wiley Environmental Molecular Sciences Laboratory (EMSL), **a national scientific user facility** at Pacific Northwest National Laboratory, provides integrated experimental and computational resources for discovery and technological innovation in the environmental molecular sciences to support the needs of DOE and the nation.

To find out more and request access to the resource: **www.emsl.pnl.gov**

History/Evolution of PNNL Proteomics



AMT tag Approach Publication Trends

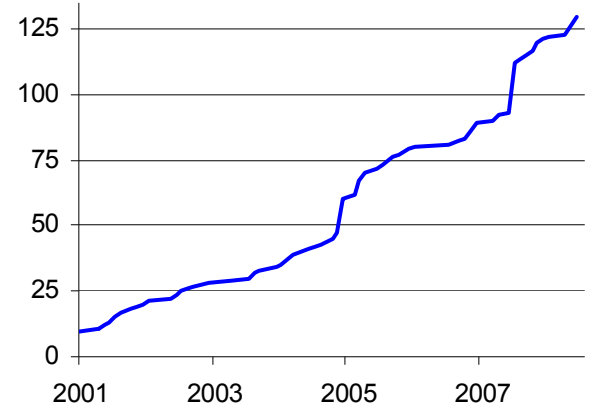


Publications from PNNL and collaborators. Excludes non-AMT tag applications papers and excludes broader technology development papers

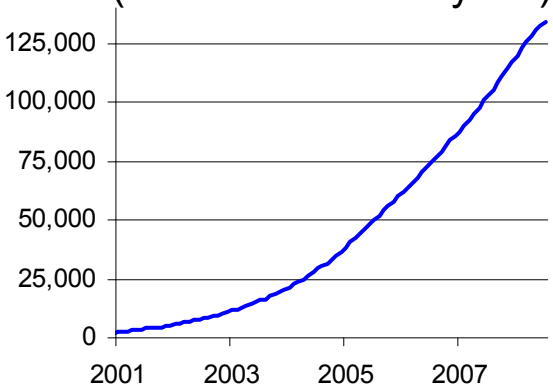
PRISM Data Trends

Organisms	145
Prepared Samples	>60,000
LC-MS(/MS) Analyses	>134,000
Automated Software Analyses	>350,000
Data Files	139 TB
Data in SQL Server databases	1.4 TB

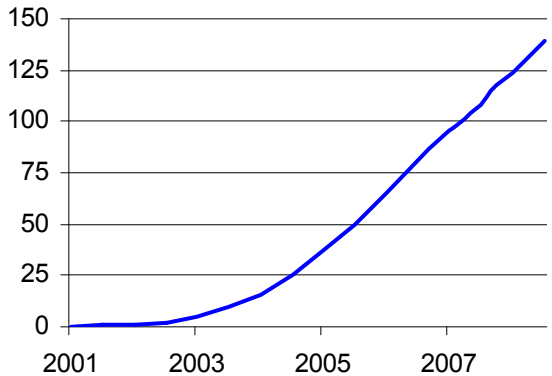
Organisms studied



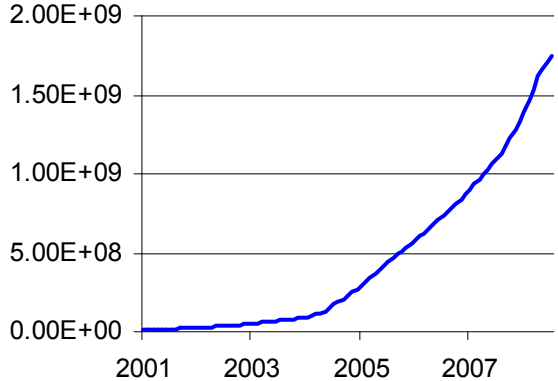
Datasets acquired (instrumental analyses)



TB data stored in PRISM



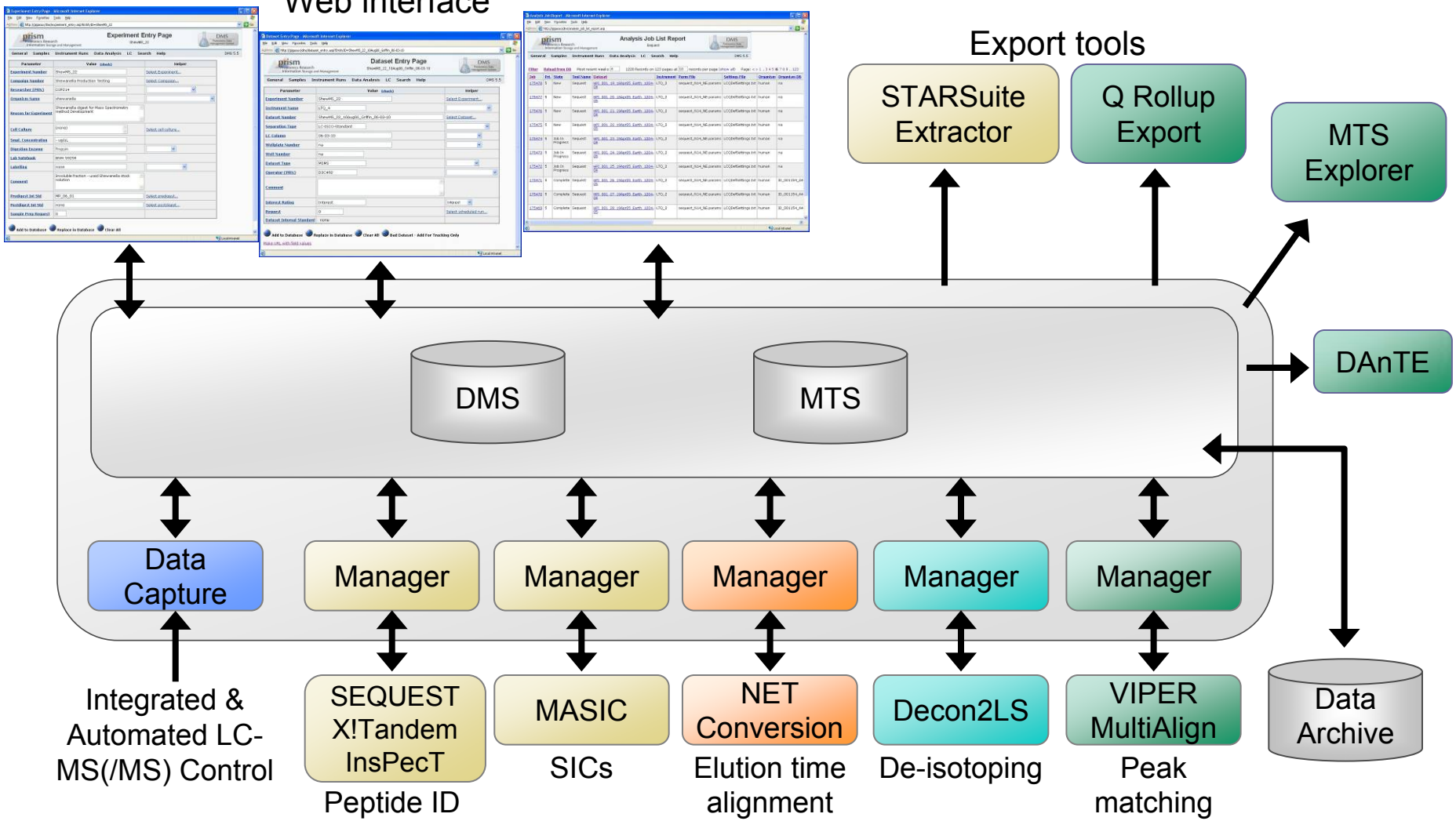
Over 1.5 billion mass spectra acquired



PRISM: Proteomics Research Information Storage and Management

Proteomics Informatics Architecture

modular and loosely coupled for flexibility

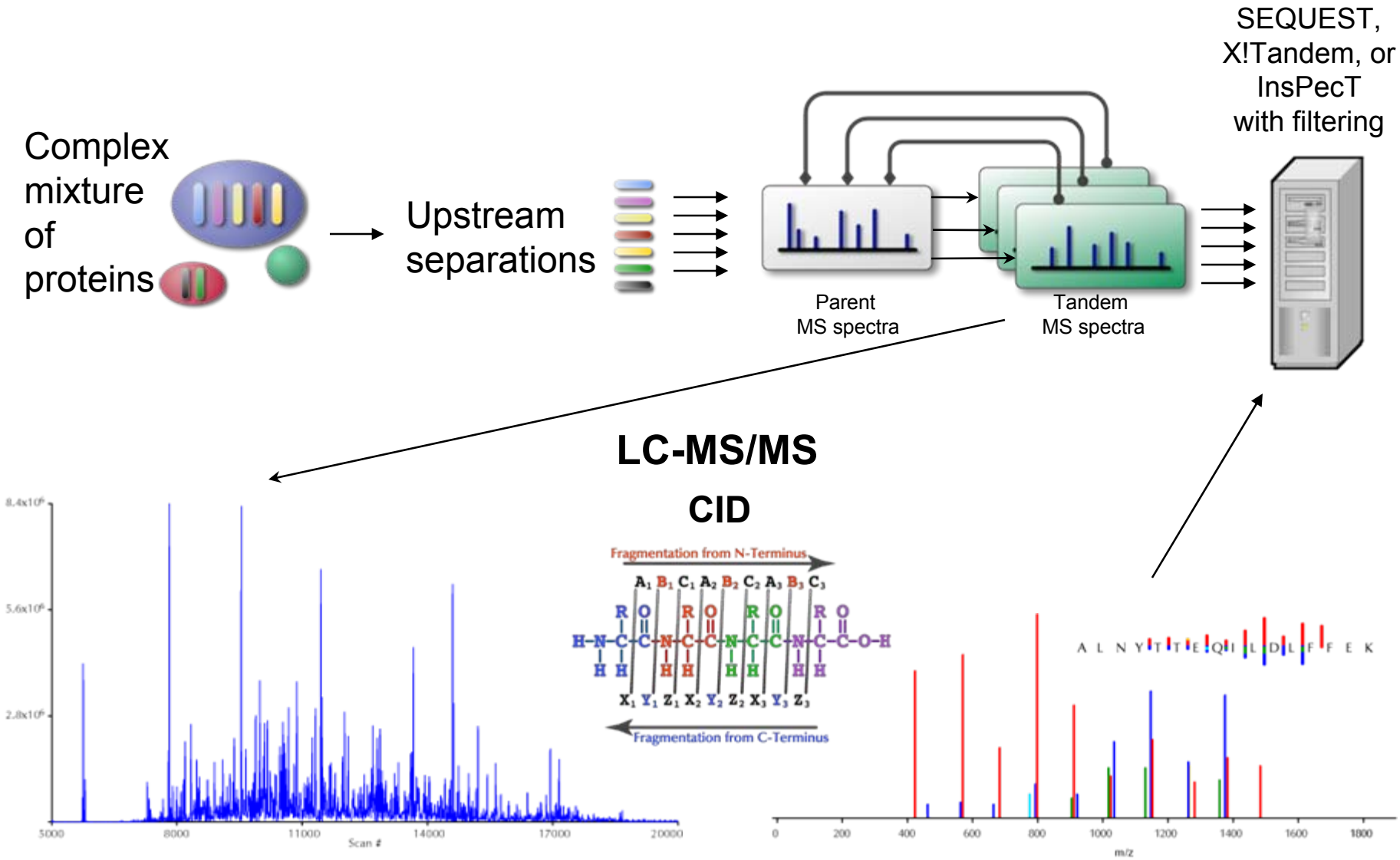


PRISM: G.R. Kiebel et. al. *Proteomics* **2006**, *6*, 1783-1790.

Motivations for Label-free LC-MS Proteomics

- Throughput, sensitivity, and sampling efficiency
  - Compared to LC-MS/MS based approaches
- Shortcomings with chemical/labeling methods
  - Multiple species need to be sampled for each “peptide”
  - Potentially more sample preparation steps or increased cost
  - Multiple analyses still required for statistical assessment
- New challenges for experimental design
  - Statistical blocking and sample order randomization
    - Helps to minimize the effects of systematic bias

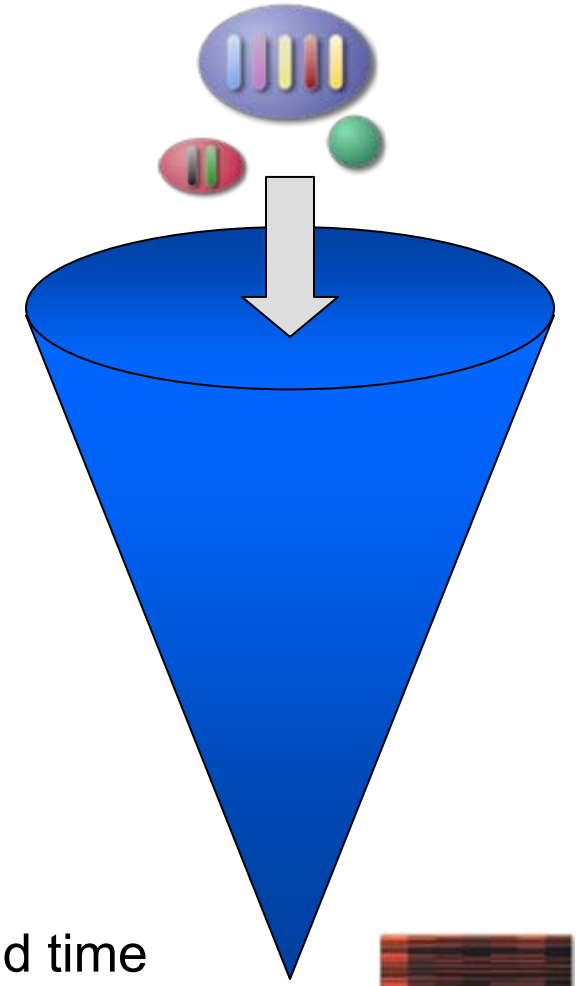
Shotgun or MuDPIT Proteomics



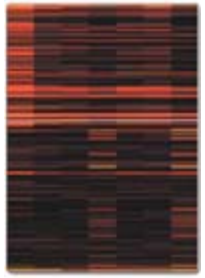
M. P. Washburn, D. Wolters, and J. R. Yates. *Nature Biotechnology* **2001**, *19 (3)*, 242-247.

LC-MS Information Funnel

- Biological sample analyzed by LC-MS
- Ions detected at instrument
- Deisotoped features
  - Charge and monoisotopic mass determined
- LC-MS features
  - Features observed in adjacent spectra with a defined chromatographic peak shape
- Identified features
  - LC-MS features that match a peptide in an AMT tag database
  - LC-MS features that are observed in multiple, related datasets at roughly the same mass and time

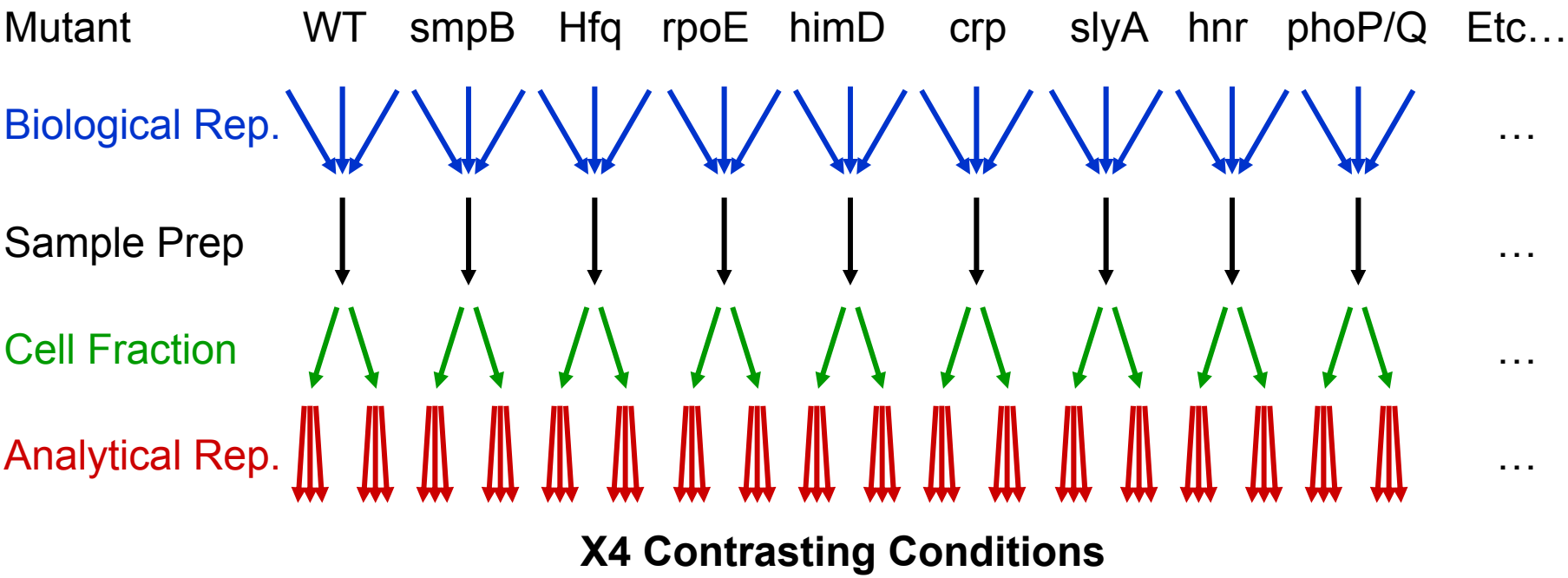


Biological knowledge



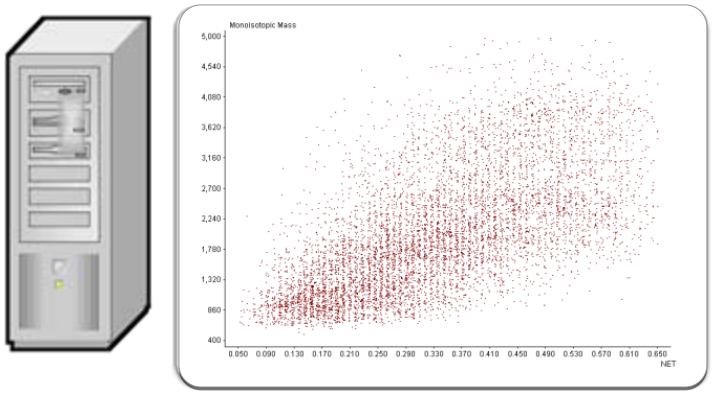
The Need/Use for Increased Throughput

- Replicate analysis to account for natural biological and normal analytical variation



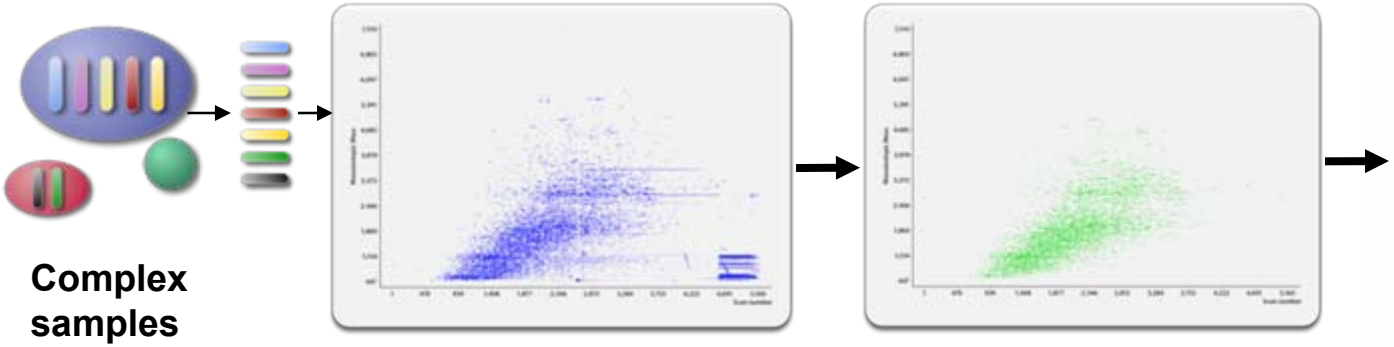
1080 analyses for 15 mutants
using biological pooling 360 analyses

Accurate Mass and Time Tag Approach



SEQUEST, X!Tandem, or InsPecT results
- Filtering
- Calculate exact mass
- Normalize observed elution time

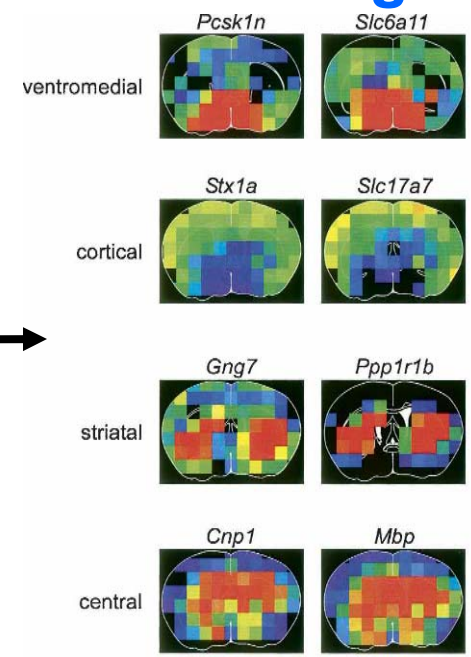
High-throughput LC-FTICR-MS Analysis with AMT tags



Complex samples

µLC- FTICR-MS

Peak-Matched Results



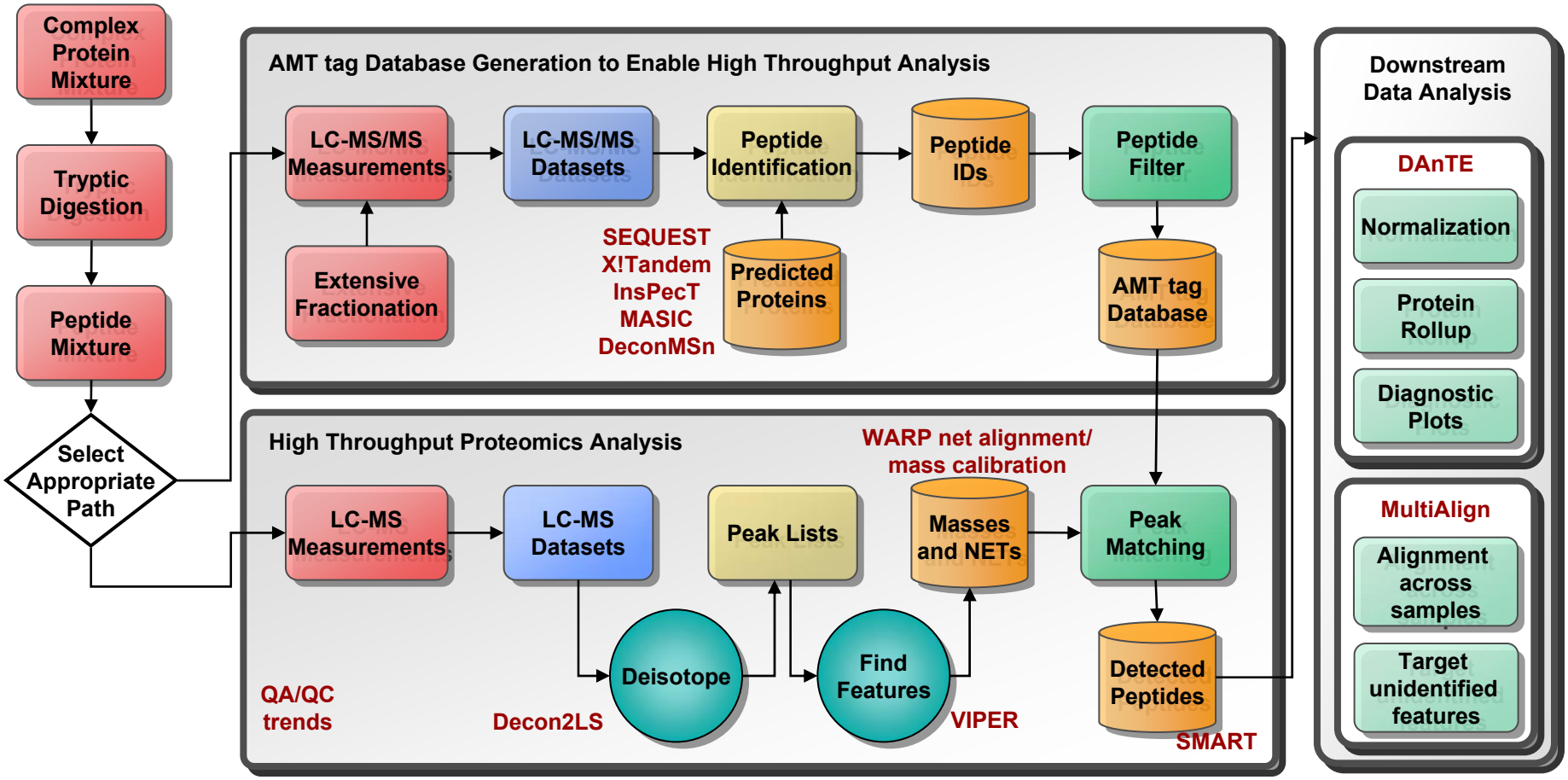
Compare abundances across samples

Example: V.A. Petyuk, et al. *Genome Research* **2007**, *17* (3), 328-336.

Considerations for Large Scale Studies
The need for blocking and randomization

- Column effects (PNNL operates 4 column systems)
  - Elution time variability, potential for carryover, and stationary phase life span
- Electrospray emitters
  - Alignment, wear, clogging, etc.
- Mass Spectrometer
  - Calibration, detector response, tuning, etc.
- Samples
  - Oxidation, degradation, and other chemical modifications
- QA/QC to assess system performance

Accurate Mass and Time (AMT) Tag Data Processing Pipeline



J.S. Zimmer et. al. *Mass. Spectrom. Rev.* **2006**, *25* (3), 450-482.

Recent Examples of Successful Applications using LC-MS Proteomics Approaches

- NIAID: *Salmonella* infecting host cells; small sample quantities → whole proteome coverage

  L. Shi, J.N. Adkins, et. al., *J. of Biological Chem.* **2006**, *281*, 29131-29140.

- Analysis of "Voxels" from mouse brains to reveal protein abundance patterns in brain structures

  V.A. Petyuk, et al. *Genome Research* **2007**, *17* (3), 328-336.

- The Mammary Epithelial Cell Secretome and Its Regulation by Signal Transduction Pathways

  J.K. Jacobs, et. al. *J. Proteome Res.* **2008**, *7* (2), 558-569.

- Analysis of purified viral particles of Monkeypox and Vaccinia viruses

  N.P. Manes, et. al. *J. Proteome Res.* **2008**, *7* (3), 960-968.

Course Related Software & Data

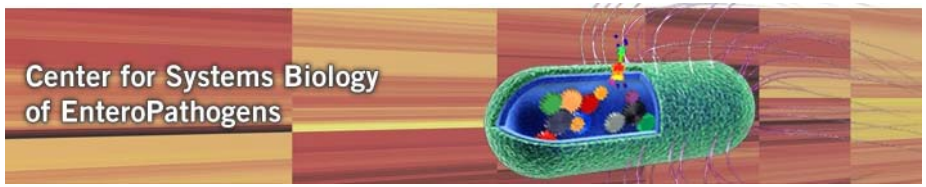
PNNL's Data and Software Distribution Website

http://omics.pnl.gov

PNNL's NCRR website

http://ncrr.pnl.gov

Salmonella Typhimurium data resource



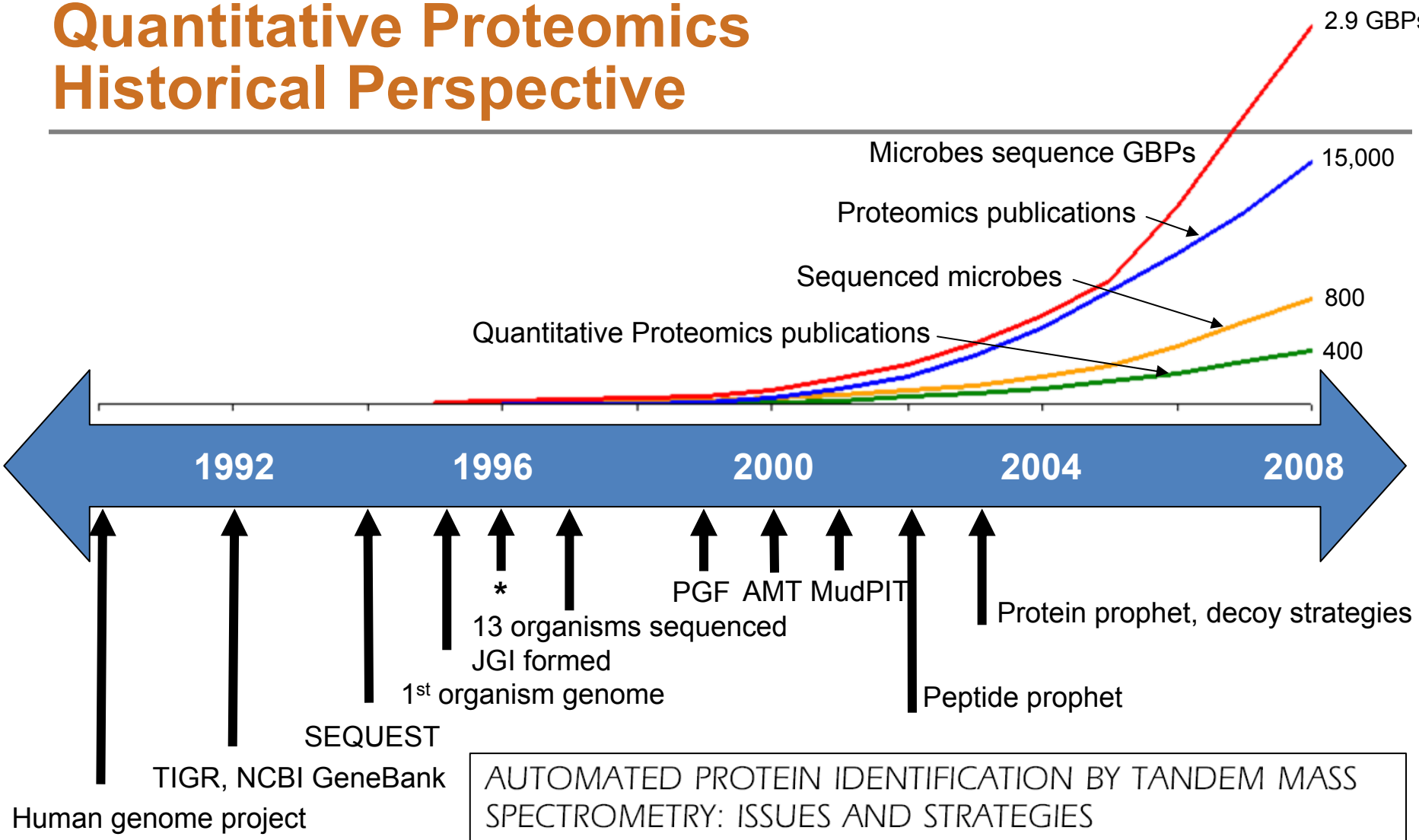
http://www.sysbep.org/

http://www.proteomicsresource.org

Selected Software Resources

- http://www.ms-utils.org/ (Magnus Palmblad)
- http://open-ms.sourceforge.net/index.php (European consortium)
- http://tools.proteomecenter.org/SpecArray.php (ISB)
- http://fiehnlab.ucdavis.edu/staff/kind/Metabolomics/Peak_Alignment/ (Tobias Kind with Oliver Fiehn)
- http://www.proteomecommons.org/tools.jsp (Phil Andrews and Jayson Falkner)
- http://www.broad.mit.edu/cancer/software/genepattern/ (Broad Institute)
- https://proteomics.fhcrc.org/CPAS/ (FHCRC)
- http://proteowizard.sourceforge.net/

Quantitative Proteomics Historical Perspective



* Separations with accurate mass MS, 1996

AUTOMATED PROTEIN IDENTIFICATION BY TANDEM MASS SPECTROMETRY: ISSUES AND STRATEGIES

Patricia Hernandez,[1]* Markus Müller,[3] and Ron D. Appel[1,2]
[1]*Swiss Institute of Bioinformatics, Geneva, Switzerland*
[2]*University of Geneva and Geneva University Hospital, Geneva, Switzerland*
[3]*Institute of Molecular Systems Biology, Swiss Federal Institute of Technology, Zürich, Switzerland*

Received 30 November 2004; received (revised) 29 March 2005; accepted 6 April 2005

Published online 11 November 2005 in Wiley InterScience (www.interscience.wiley.com) DOI 10.1002/mas.20068

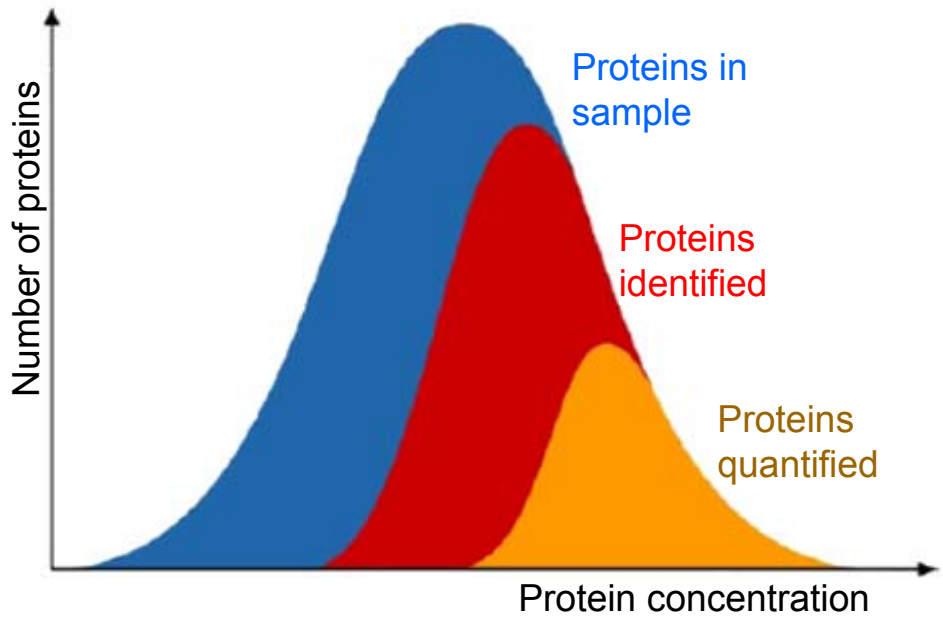
Proteomics Workflow

Cells / Tissue

Sample Processing → Instrument Analysis (LC-MS, LC-MS/MS) → Feature Extraction Identification → Quantitative Analysis

Purification
Fractionation
Protein extraction
Digestion
Labeling
Spiking

TOF
Ion Traps
Q-TOF
TOF-TOF
FTICR
Orbitrap



M. Bantscheff, M. Schirle, G. Sweetman, J. Rick, and B. Kuster, "Quantitative mass spectrometry in proteomics: a critical review," *Anal. Bioanal. Chem.* **2007**, *389 (4)*, 1017-1031.

Identification Strategies

- Proteomics
  - MS (Peptide Mass Fingerprinting or PMF)
    - Low complexity mixtures
  - MS/MS (Peptide Fragment Fingerprinting or PFF)
    - Comprehensive tool set available
  - Accurate Mass and Time (AMT) tag approach
    - Requires database of peptide IDs and LC elution times
    - High throughput
  - Validation
    - Peptide ID confidence
    - Peptide to protein assignment
    - Protein identification confidence
- Metabolomics
  - Identification tools less mature
    - Accurate mass can be used to determine molecular formula
    - Structural determination
      - Manual analysis of MS/MS spectra
      - NMR analysis

Pros and Cons of PMF/PFF Strategies

Computational method	Data	Pro/cons
Database dependent PMF search	MS	Accurate Database dependent Fast interpretation Few modifications can be included Only for simple protein mixtures
Database dependent PFF search	MS/MS	Accurate Complex protein mixtures Database dependent Moderate amount of modifications can be included Moderate search speed
Sequence tag + database search	MS/MS	Moderate accuracy[a)] Complex protein mixtures Database dependent Many modifications can be included Fast interpretation
De novo sequencing	MS/MS	Less accurate[a)] Complex protein mixtures Database independent Few modifications can be included Fast interpretation
Spectral library search	MS/MS	Highly accurate Complex protein mixtures Database dependent Few modifications Fast interpretation

a) Likely to become more accurate with future instrument and algorithmic improvements.

R. Matthiesen, "Methods, algorithms and tools in computational proteomics: a practical point of view," *Proteomics* **2007**, *7 (16),* 2815-2832.

Quantitation Strategies

- Proteomics
  - Label based (Relative / Absolute)
    - Metabolic labeling
    - Chemical labeling
    - Enzymatic labeling
  - Label free (Relative / Absolute)
  - Peptide to protein “rollup”
    - Degenerate peptide problem
    - Normalization methods
- Metabolomics
  - Primarily label free approaches
  - Does not suffer from the “rollup” challenge

Quantitation Strategies

	Application	Accuracy (process)	Quantitative proteome coverage	Linear dynamic range[a]
Metabolic protein labeling	Complex biochemical workflows Comparison of 2–3 states Cell culture systems only	+++	++	1–2 logs
Chemical protein labeling (MS)	Medium to complex biochemical workflows Comparison of 2–3 states	+++	++	1–2 logs
Chemical peptide labeling (MS)	Medium complexity biochemical workflows Comparison of 2–3 states	++	++	2 logs
Chemical peptide labeling (MS/MS)	Medium complexity biochemical workflows Comparison of 2–8 states	++	++	2 logs
Enzymatic labeling (MS)	Medium complexity biochemical workflows Comparison of 2 states	++	++	1–2 logs
Spiked peptides	Medium complexity biochemical workflows Targeted analysis of few proteins	++	+	2 logs
Label free (ion intensity)	Simple biochemical workflows Whole proteome analysis Comparison of multiple states	+	+++	2–3 logs
Label free (spectrum counting)	Simple biochemical workflows Whole proteome analysis Comparison of multiple states	+	+++	2–3 logs

[a] In MRM mode, dynamic range may be extended to 4–5 logs [65]

M. Bantscheff, M. Schirle, G. Sweetman, J. Rick, and B. Kuster, "Quantitative mass spectrometry in proteomics: a critical review," *Anal. Bioanal. Chem.* **2007**, *389 (4)*, 1017-1031.

Quantitation Strategies

Target	Name of method or reagent	Isotopes
Metabolic stable-isotope labeling		
None	^{15}N-labeling (^{15}N-ammonium salt)	^{15}N
	Stable isotope labeling by amino acids in cell culture (SILAC)	D, ^{13}C, ^{15}N
	Culture-derived isotope tags (CDIT)	D, ^{13}C, ^{15}N
	Bioorthogonal noncanonical amino acid tagging (BONCAT)	No isotope
Isotope tagging by chemical reaction		
Sulfhydryl	Isotope-coded affinity tagging (ICAT)	D
	Cleavable ICAT	^{13}C
	Catch-and-release (CAR)	^{13}C
	Acrylamide	D
	Isotope-coded reduction off of a chromatographic support (ICROC)	D
	2-vinyl-pyridine	D
	N-t-butyliodoacetamide	D
	Iodoacetanilide	D
	HysTag	D
	Solid-phase ICAT	D
	Visible isotope-coded affinity tags (VICAT)	^{13}C, ^{14}C and ^{15}N
	Acid-labile isotope-coded extractants (ALICE)	D
	Solid phase mass tagging	^{13}C
Amines	Tandem mass tag (TMT)	D
	Succinic anhydride	D
	N-acetoxysuccinamide	D
	N-acetoxysuccinamide: In-gel Stable-Isotope Labeling (ISIL)	D
	Acetic anahydride	D
	Proprionic anhydride	D
	Nicotinoyloxy succinimide (Nic-NHS)	D
	Isotope-coded protein labeling (ICPL,Nic-NHS)	D
	Phenyl isocyanate	D or ^{13}C
	Isotope-coded n-terminal sulfonation (ICens) 4-sulphophenyl isothiocyanate (SPITC)	^{13}C
	Sulfo-NHS-SS-biotin and 13C,D3-methyl iodide	^{13}C and D
	Formaldehyde	D
	Isobaric tag for realtive and absolute quantification (iTRAQ)	^{13}C, ^{15}N and ^{18}O
	Benzoic acid labeling (BA part of ANIBAL)	^{13}C
Lysines	Guanidination (O-methyl-isourea) mass-coded abundance tagging (MCAT)	No isotope
	Guanidination (O-methyl-isourea)	^{13}C and ^{15}N
	Quantitation using enhanced sequence tags (QUEST)	No isotope
	2-Methoxy-4,5-1H-imidazole	D
N-terminus protein	Differentially isotope-coded N-terminal protein sulphonation (SPITC)	^{13}C
N-terminus peptide	N-terminal stable-isotope labelling of tryptic peptides (pentafluorophenyl-4-anilino-4-oxobutanoate)	D or ^{13}C
Carboxyl	Methyl esterification	D
	Ethyl esterification	D
	C-terminal isotope-coded tagging using sulfanilic acid (SA)	^{13}C
	Aniline labeling (ANI part of ANIBAL)	^{13}C
Indole	2-nitrobenzenesulfenyl chloride (NBSCI)	^{13}C

Target	Name of method or reagent	Isotopes
Stable-isotope incorporation via enzyme reaction		
C-terminus peptide	Proteolytic ^{18}O-labeling ($H_2{}^{18}O$)	^{18}O
	Quantitative cysteinyl-peptide enrichment technology (QCET)	^{18}O
Absolute quantification		
None	Absolute quantification (AQUA)	D, ^{13}C, ^{15}N
	Multiplexed absolute quantification (QCAT)	D, ^{13}C, ^{15}N
	Multiplexed absolute quantification using concatenated signature (QconCAT)	D, ^{13}C, ^{15}N
	Stable Isotope Standards and Capture by Anti-Peptide Antibodies (SISCAPA)	D, ^{13}C, ^{15}N
Label-free quantification		
None	XIC-based quantification	No isotope
	Spectrum sampling (SpS)	No isotope
	Protein abundance index (PAI)	No isotope
	Exponentially modified protein abundance index (emPAI)	No isotope
	Probabilistic peptide scores (PMSS)	No isotope

A. Panchaud, M. Affolter, P. Moreillon, and M. Kussmann, "Experimental and computational approaches to quantitative proteomics: status quo and outlook," *J. Proteomics* **2008**, *71 (1)*, 19-33.

Validation

- Measurement validation
  - Peptide/Protein Identification
    - Confidence algorithms
    - Statistical models
  - Quantitation
    - Less mature than identification confidence
- Functional validation
  - Western blots
  - Gene knockout
  - Protein assays
  - Protein chemistry
- However, all measure something different

Active Software Development to Address Challenges

- Large array of available tools
  - No universal analysis workflow
- Tool functional categories
  - Peptide ID
  - Identification confidence
    - SMART, epic (PNNL active research)
  - Quantitation
    - A. Polpitiya et al., "DAnTE: a statistical tool for quantitative analysis of -omics data," *Bioinformatics* **2008**, *24 (13),* 1556-1558.
  - Data management / metadata capture
  - Workflow automation

MS identification	Validation	Quantitation	Storage
MASCOT [21, 35, 36][a]	MAE [37][a]	VEMS [38, 39][b]	CPAS [40, 41][b]
ProFound [42][b]	AMASS [43][a]	MSinspect [44][b]	PRIDE [45–47][b]
VEMS [38, 39, 48][b]	SALSA [49][a]	PEPPeR [50][b]	dbVEMS [38, 39][b]
Aldente [51][b]	Qscore [52][b]	OpenMS [53][b]	Proteios [54, 55][b]
MS-Fit[b]	PeptideProphet [30][b]	Mzmine [56, 57][b]	GPMDB [58][b]
PeptIdent[b]	Scope [59][c]	SpecArray [60][b]	**Raw data formats**
MS/MS (Direct Database)	EPIR [61][c]	Peplist [60][b]	MzXML [62][b]
VEMS [38, 39, 48][b]	Spider [63][b]	PEAKSQ (BSI)[c]	mzData [64][b]
MASCOT [21, 35, 36][a]	SILVER [65][b]	MSquant (http://msquant.sourceforge.net/)[b]	**Result formats**
Phenyx [31, 34](GENEBIO)[a]		MSight [66][b]	AnalysisXML [46][b]
SEQUEST (Thermo Finnigan)[c]		RelEx [22][b]	ProtXML [67][b]
X!tandem [68][b]		ASAPratio [69][b]	pepXML [67][b]
ProbId [70][b]		**2D-gels**	**Pipelines**
PopITAM [71][b]		Flicker [72][b]	TPP (tools.proteomecenter.org/TPP.php)[b]
OMSSA [73][b]		Melanie (www.gehealthcare.com) [74–76][a]	ProteinScape™ (www.proteinscape.com)[c]
P-mod [77][a]		PDQuest (www.bio-rad.com)[c]	Scaffold (www.proteomesoftware.com)[c]
PLGS (Waters)[c]		DeCyder (www.gehealthcare.com)[c]	TOPP [78][b]
Paragon (ABI)[c,d]		Delta2D (www.decodon.com)[c]	VEMS (http://personal.cicbiogune.es/rmatthiesen/)[c]
Spectral library search		Progenesis (www.nonlinear.com)[c]	PLGS (www.waters.com)[c]
X ! Hunter [58][b]		Proteomweaver (www.definiens.com/www.bio-rad.com)[c]	
Proteotypic Peptide search			
X! P3 [79][b]			
MS/MS (Tag database)			
GutenTag [80][a]			
InsPecT [81][b]			
Popitam [71][b]			
MS/MS (*De novo*)			
Lutefisk [82, 83][b]			
PepHMM [32][b]			
Sherenga [84][c]			
PepNovo [33][b]			
Peaks (BSI)[c,e]			

Community Development

a) Semi-commercial or must contact author
b) Freely available on the internet
c) Commercial or not available
d) Applied Biosystems
e) Bioinformatics Solutions

R. Matthiesen, "Methods, algorithms and tools in computational proteomics: a practical point of view," *Proteomics* **2007**, *7 (16),* 2815-2832.

Software Platforms for Label-free Quantitation

	PNNL Pipeline	PEPPeR	msInspect	SuperHirn	CRAWDAD
Lab	PNNL	Broad Institute	FHCRC	IMSB (Swiss)	Univ. Wash.
Feature Picker	Decon2LS/Viper	Mapquant (or any other)	msInspect	SuperHirn	CRAWDAD
Method	Spectrum de-isotoping then clustering	Image Analysis then de-isotoping	Wavelet decomposition then de-isotoping	Spectrum de-isotoping then merging	m/z channel binning
RT Alignment	Normalization, then linear or LCMSWARP	Relative, then linear, or LOESS (exp)	Iterative non-linear transformation	LOESS modeling	Dynamic time warping
m/z recalibration	Yes (dynamic)	Yes (quadratic)	No	No	No
Assignment of IDs to features	AMT database, normalized elution times	AMT database, relative elution order (Landmarks)	AMT database through user interaction	Yes, but not well documented at present	Yes, for differences only if they exist
Statistical Evaluation of assignment	Mass shift decoy and/or Bayesian Statistics	Bayesian Statistics	No	No	No
Unidentified Feature Recognition	Stored in database for later analysis	Data-dependent tolerance-based clustering	User specified tolerance-based clustering	Tolerance-based merging, heuristics	Difference mapping only
Runs on	Windows with GUI	Web-based (Linux or Windows install bases)	Java with GUI	Linux	Linux/Windows

Confident Quantitative Results

- Credible results require
  - Rigorous statistical models
  - Validation
    - Measurements
    - Functions
  - Full disclosure of procedures and methods
  - Dissemination
    - Data
    - Custom analysis software tools
- Data standards and release policies are critical
  - HUPO Proteomics Standards Initiative: http://www.psidev.info/
  - L. Martens and H. Hermjakob, "Proteomics data validation: why all must provide data," *Mol. Biosyst.* **2007**, *3 (8),* 518-522.

References

- A. Panchaud, M. Affolter, P. Moreillon, and M. Kussmann, "Experimental and computational approaches to quantitative proteomics: status quo and outlook," *J. Proteomics* **2008**, *71 (1)*, 19-33.
- A. Honda, Y. Suzuki, and K. Suzuki, "Review of molecular modification techniques for improved detection of biomolecules by mass spectrometry," *Anal. Chim. Acta.* **2008**, *623 (1)*, 1-10.
- T.O. Metz, J.S. Page, E.S. Baker, K.Q. Tang, J. Ding, Y.F. Shen, and R.D. Smith, "High-resolution separations and improved ion production and transmission in metabolomics," *Trac-Trends in Analytical Chemistry* **2008**, *27 (3)*, 205-214.
- L. Martens and H. Hermjakob, "Proteomics data validation: why all must provide data," *Mol. Biosyst*. **2007**, *3 (8)*, 518-522.
- B.J. Webb-Robertson and W.R. Cannon, "Current trends in computational inference from mass spectrometry-based proteomics," *Brief Bioinform.* **2007**, *8 (5)*, 304-317.
- T.O. Metz, Q. Zhang, J.S. Page, Y. Shen, S.J. Callister, J.M. Jacobs, and R.D. Smith, "The future of liquid chromatography-mass spectrometry (LC-MS) in //metabolic profiling and metabolomic studies for biomarker discovery," *Biomark. Med*. **2007,** *1 (1),* 159-185.

References

- R. Matthiesen, "Methods, algorithms and tools in computational proteomics: a practical point of view," *Proteomics* **2007**, *7 (16)*, 2815-2832.
- M. Bantscheff, M. Schirle, G. Sweetman, J. Rick, and B. Kuster, "Quantitative mass spectrometry in proteomics: a critical review," *Anal. Bioanal. Chem.* **2007**, *389 (4)*, 1017-1031.
- W. Urfer, M. Grzegorczyk, and K. Jung, "Statistics for proteomics: a review of tools for analyzing experimental data," *Proteomics* **2006**, *6 Suppl 2*, 48-55.
- P. Hernandez, M. Muller, and R.D. Appel, "Automated protein identification by tandem mass spectrometry: issues and strategies," *Mass Spectrom. Rev.* **2006**, *25 (2)*, 235-54.
- J. Peng, J.E. Elias, C.C. Thoreen, L.J. Licklider, and S.P. Gygi, "Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome," *J. Proteome Res.* **2003**, *2 (1)*, 43-50.
- S.A. Gerber, J. Rush, O. Stemman, M.W. Kirschner, and S.P. Gygi, "Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS," *Proc. Natl. Acad. Sci. USA* **2003**, *100 (12)*, 6940-5.
- R. Aebersold and M. Mann, "Mass spectrometry-based proteomics," *Nature.* **2003**, *422 (6928)*, 198-207.

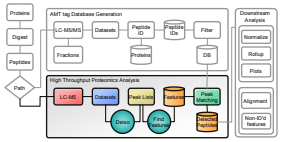
References

- R.D. Smith, G.A. Anderson, M.S. Lipton, L. Pasa-Tolic, Y. Shen, T.P. Conrads, T.D. Veenstra, and H.R. Udseth, "An accurate mass tag strategy for quantitative and high-throughput proteome measurements," *Proteomics* **2002**, *2 (5)*, 513-23.
- R.D. Smith, L. Pasa-Tolic, M.S. Lipton, P.K. Jensen, G.A. Anderson, Y. Shen, T.P. Conrads, H.R. Udseth, R. Harkewicz, M.E. Belov, C. Masselon, and T.D. Veenstra, "Rapid quantitative measurements of proteomes by Fourier transform ion cyclotron resonance mass spectrometry," *Electrophoresis* **2001,** *22 (9)*, 1652-68.
- T.P. Conrads, K. Alving, T.D. Veenstra, M.E. Belov, G.A. Anderson, D.J. Anderson, M.S. Lipton, L. Pasa-Tolic, H.R. Udseth, W.B. Chrisler, B.D. Thrall, and R.D. Smith, "Quantitative analysis of bacterial and mammalian proteomes using a combination of cysteine affinity tags and ^{15}N-metabolic labeling," *Anal. Chem.* **2001**, *73 (9)*, 2132-9.
- T.P. Conrads, G.A. Anderson, T.D. Veenstra, L. Pasa-Tolic, and R.D. Smith, "Utility of accurate mass tags for proteome-wide protein identification," *Anal. Chem.* **2000**, *72 (14)*, 3349-54.
- J.K. Nicholson, J.C. Lindon, and E. Holmes, "Metabonomics: understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data," *Xenobiotica* **1999**, *29 (11)*, 1181-1189.

Part II: LC-MS Feature Discovery

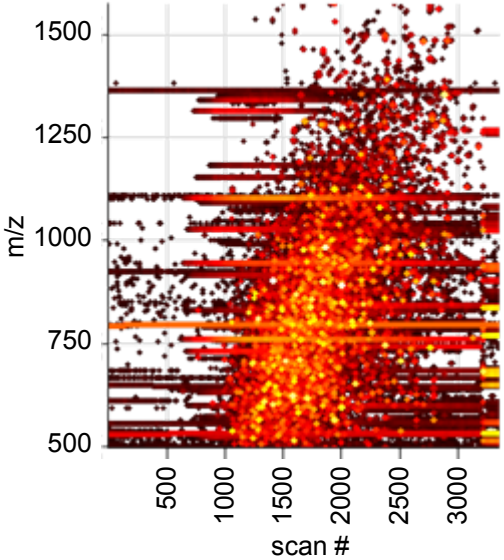
- Part I: Introduction and Overview of Label-Free Quantitative Proteomics (Anderson)
- Part II: Feature Discovery in LC-MS Datasets (Monroe and Polpitiya)
  - Structure of LC-MS Data
  - Feature discovery in individual spectra (deisotoping)
  - Feature definition over elution time
  - Identifying LC-MS Features using an AMT tag DB
  - Extending the AMT tag approach for feature based analyses
  - Estimating confidence of identified LC-MS features
  - Downstream quantitative analysis with DAnTE
- Break
- Part III: Biological Application of the AMT tag Approach (Ansong)
- AMT tag Analysis Software Demo
- Panel Discussion

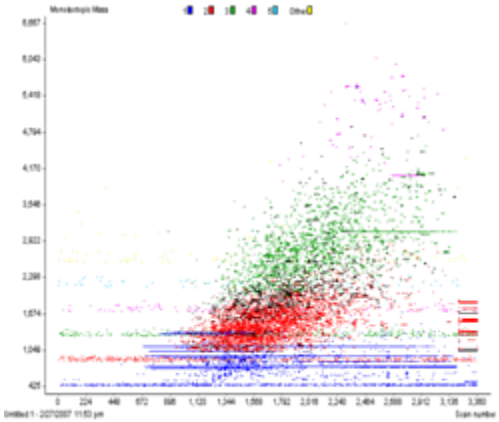
Feature Discovery in LC-MS Datasets

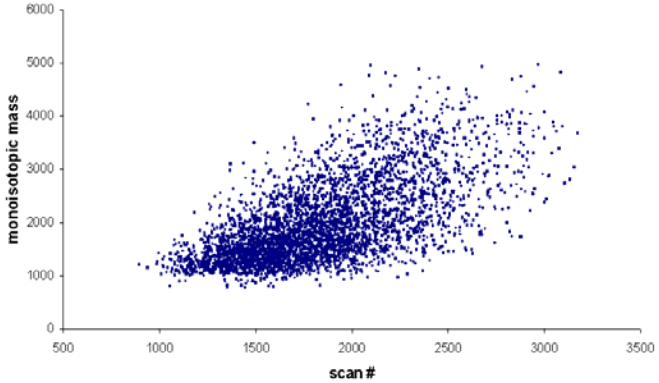


- Two-dimensional views of an LC-MS dataset in different stages of data processing
  - Several stages of processing are required to extract biological knowledge from raw LC-MS data

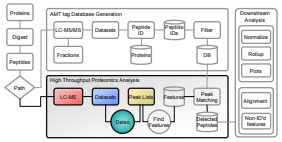
Raw Data → Deisotoping → Monoisotopic Features in Each Mass Spectrum → Elution profile discovery → LC-MS Features Characterized



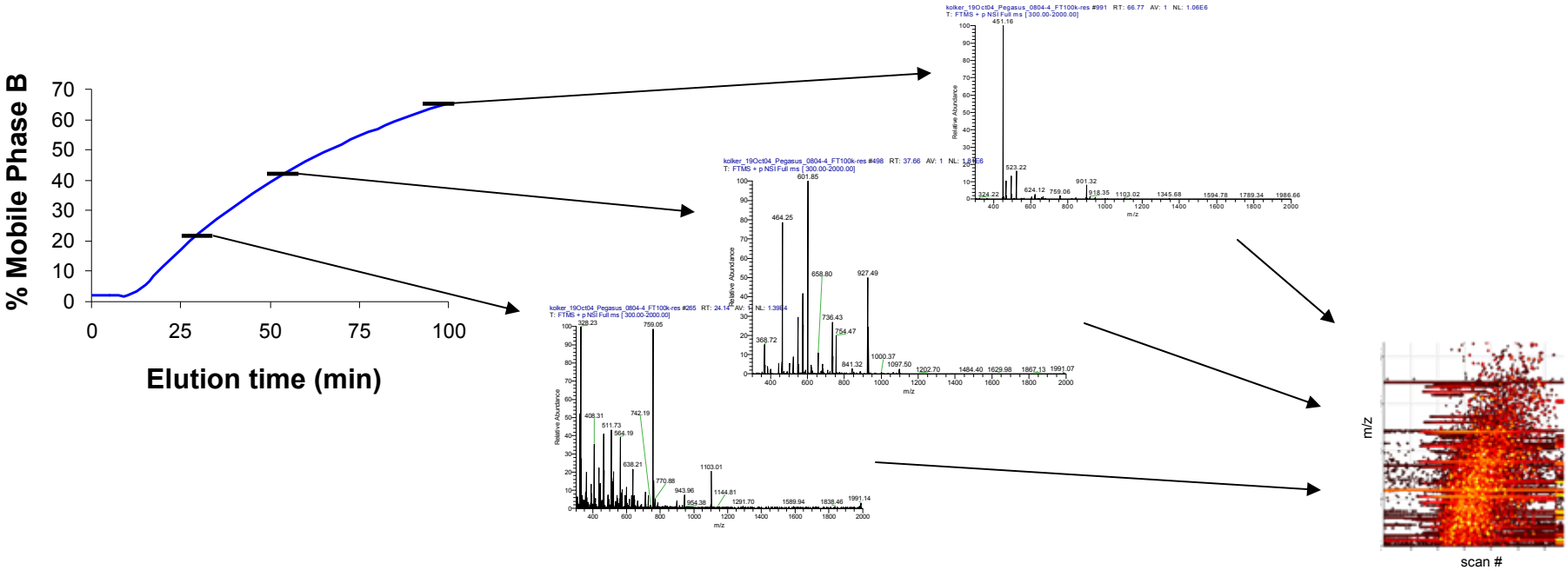




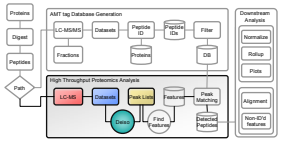
Structure of LC-MS Data



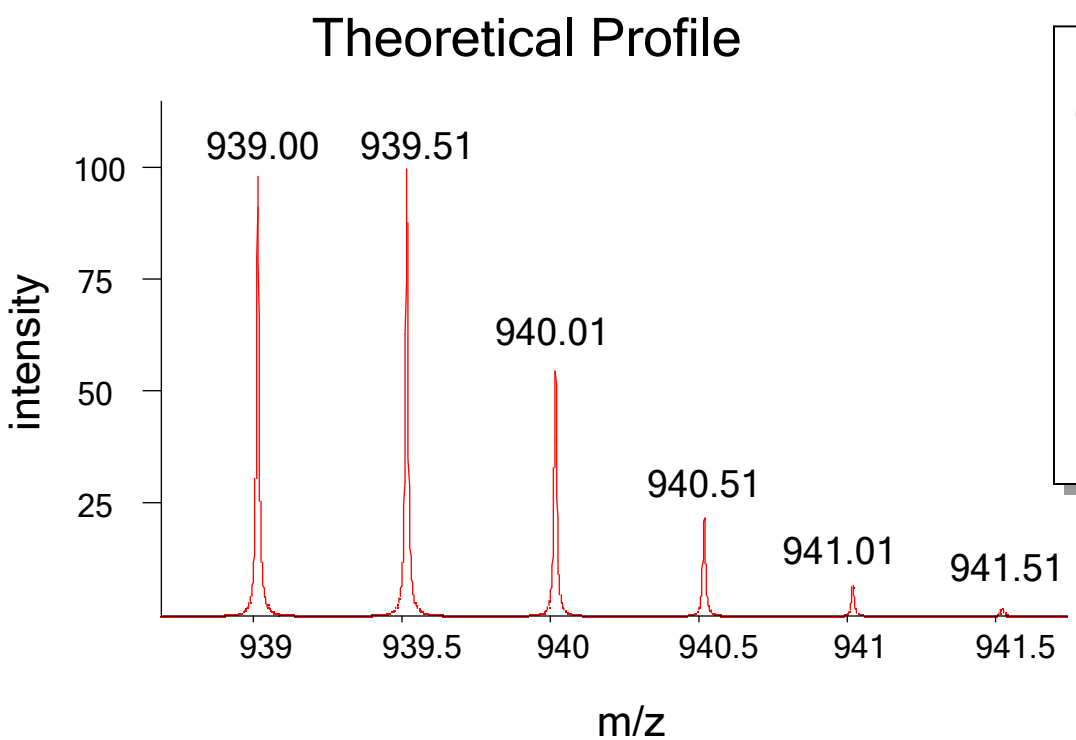
- Mass spectra capture the changing composition of peptides eluting from a chromatographic column
  - Complex peptide mixture on a column is separated by liquid chromatography over a period of time
  - Changing composition of the mobile phase causes different peptides to elute at different times
  - The components eluting from a column are sampled continuously by sequential mass spectra



Structure of LC-MS Data



- Each compound is observed as an <u>isotopic pattern</u> in a mass spectrum
  - The pattern is dependent on the compound's chemical composition, charge, and resolution of instrument



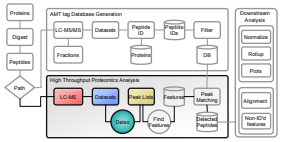
m/z: 939.0203
Charge: 2+
Monoisotopic Mass: 1876.0054 Da

Elution range: Scans 1539 - 1593

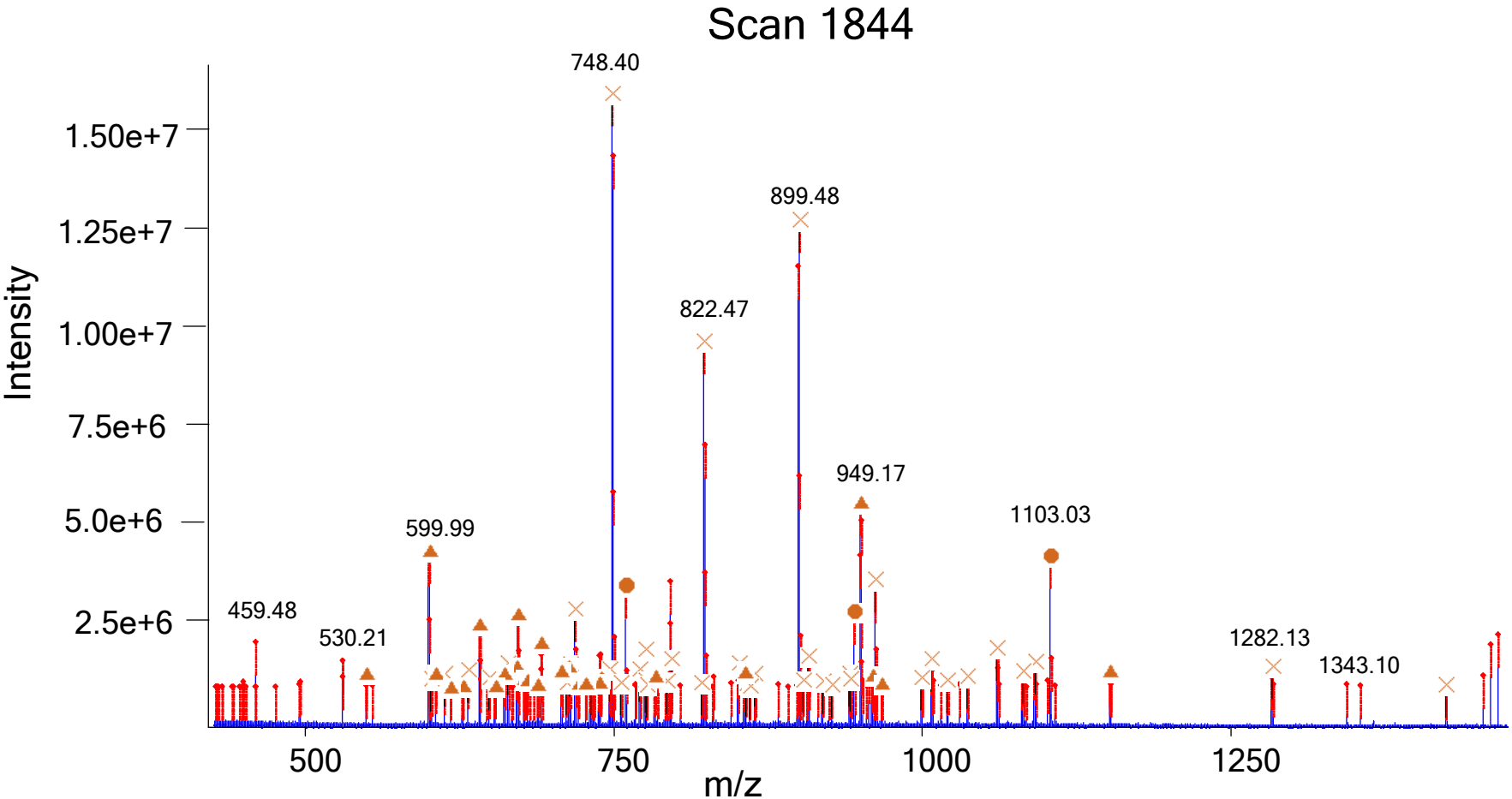
Peptide: VKHPSEIVNVGDEINVK

Parent Protein: gi|16759851 30S ribosomal protein S1

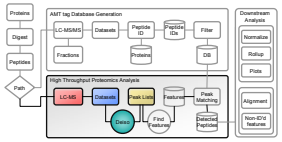
Structure of LC-MS Data



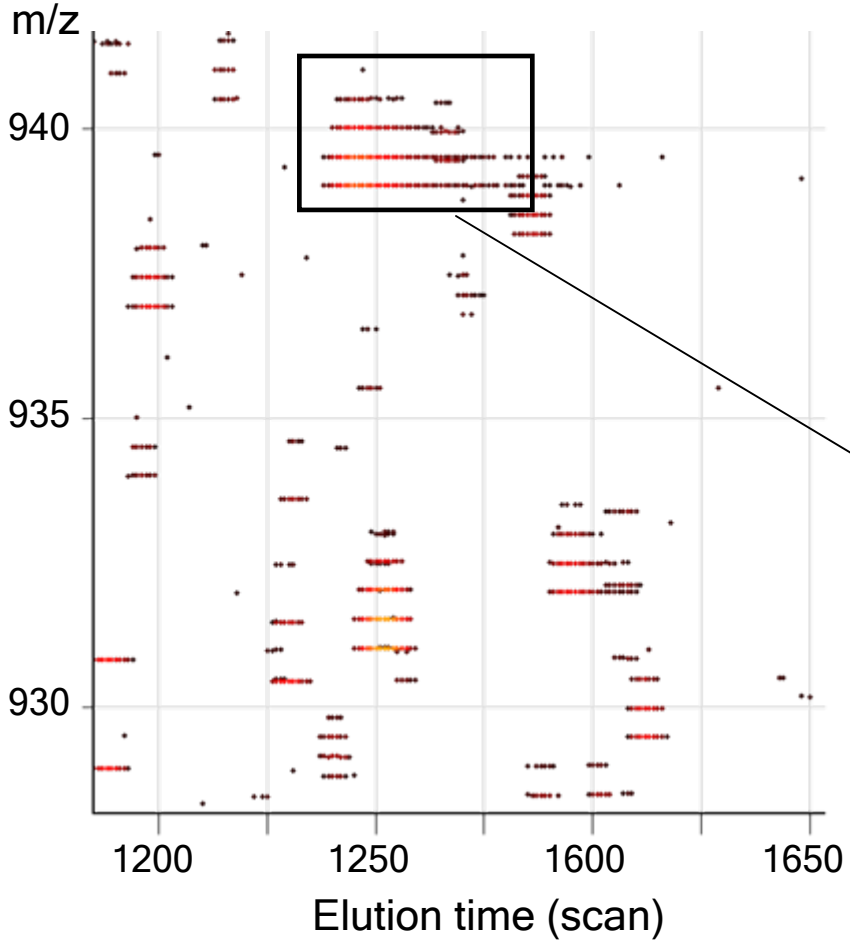
- A mass spectrum of a complex mixture contains overlaid distributions of several different compounds



Structure of LC-MS Data



- With LC as the first dimension, each compound is observed over multiple spectra, showing a three-dimensional pattern of m/z, elution time and abundance



- Goal: Infer mass, elution time, and intensity of compounds that are present in the LC-MS dataset
  - Compounds are termed LC-MS features since they are inferred from a 3D pattern, yet identity is unknown

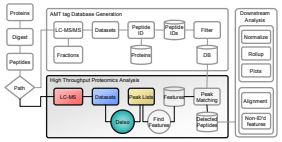
m/z: 939.0203
Charge: 2+
Monoisotopic Mass: 1876.0054 Da

Elution range: Scans 1539 - 1593

Peptide: VKHPSEIVNVGDEINVK

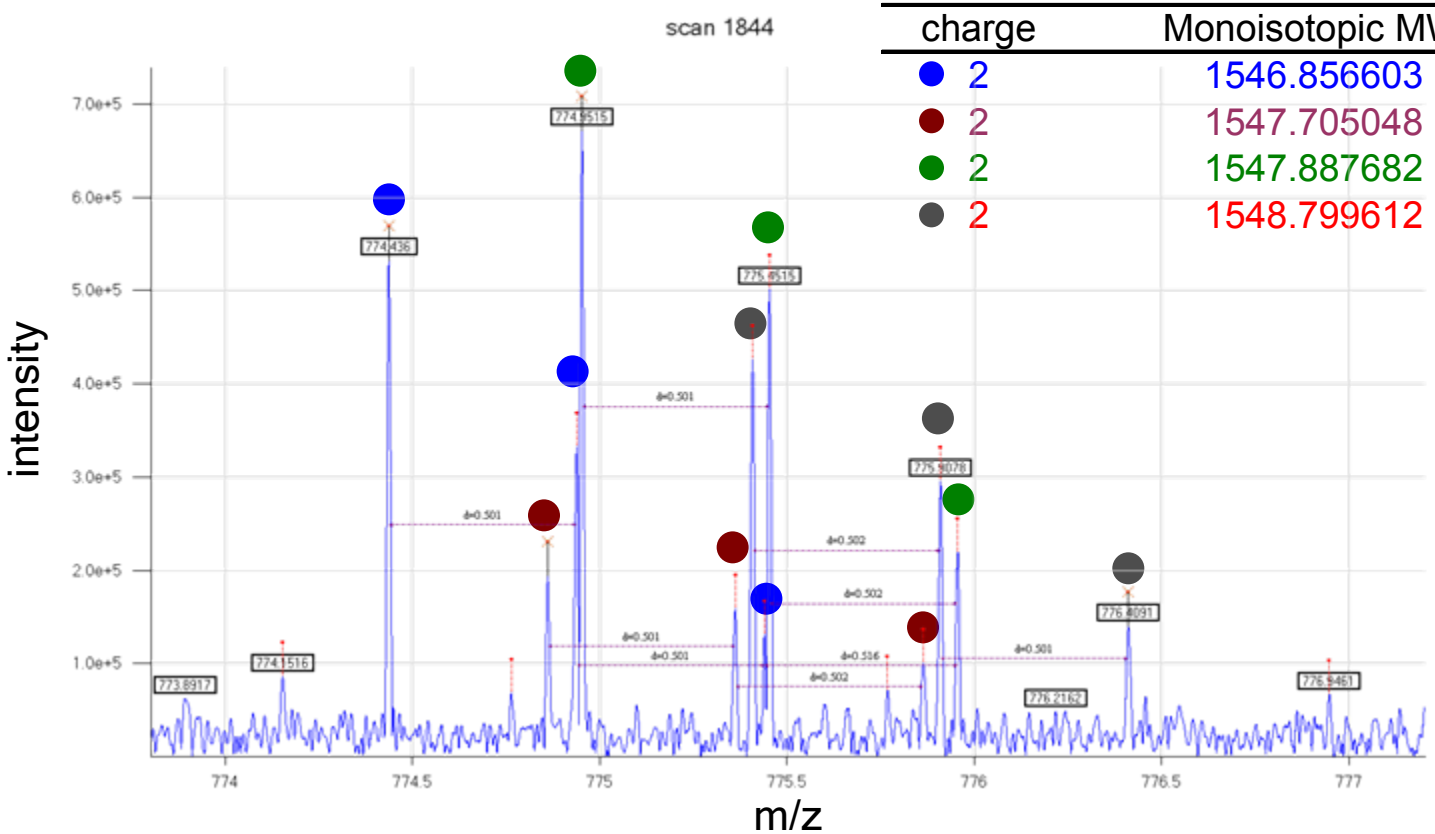
Parent Protein: gi|16759851 30S ribosomal protein S1

Feature Discovery in Individual Spectra



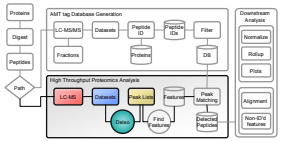
- Deisotoping
  - Process of converting a mass spectrum (m/z and intensity) into a list of species (mass, abundance, and charge)

Deisotoping a mass spectrum of 4 overlapping species

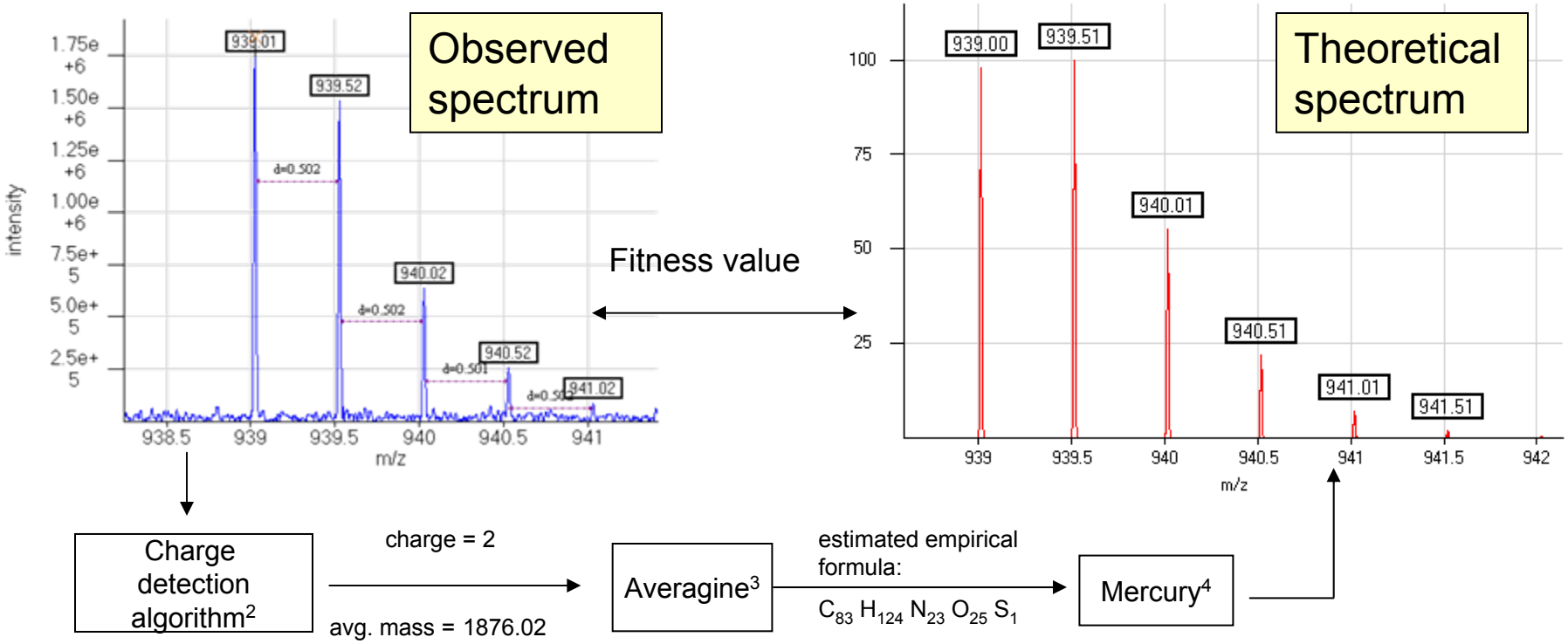


charge	Monoisotopic MW	abundance
2	1546.856603	533467
2	1547.705048	194607
2	1547.887682	671947
2	1548.799612	426939

Deisotoping an Isotopic Distribution

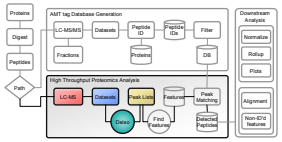


- Decon2LS deisotoping algorithm compares theoretical isotopic patterns with observed patterns

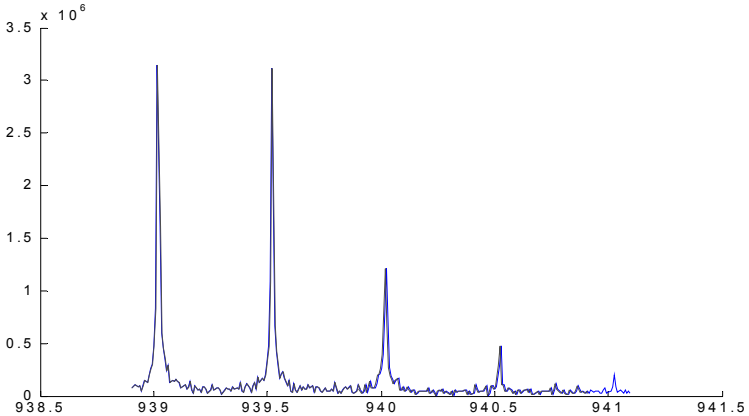


1. Horn, D.M., Zubarev, R.A., McLafferty, F.W. Automated Reduction and Interpretation of High Resolution Electrospray Mass Spectra of Large Molecules. *J. Am. Soc. Mass Spectrom.* **2000**, *11*, 320-332.
2. Senko, M. W.; Beu, S. C.; McLafferty, F. W. Automated assignment of charge states from resolved isotopic peaks for multiplycharged ions. *J. Am. Soc. Mass Spectrom.* **1995,** *6*, 52–56.
3. Senko, M. W.; Beu, S. C.; McLafferty, F. W. Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *J. Am. Soc. Mass Spectrom.* **1995,** *6*, 229–233.
4. Rockwood, A. L.; Van Orden, S. L.; Smith, R. D. Rapid Calculation of Isotope Distributions. *Anal. Chem.* **1995,** *67,* 2699–2704.

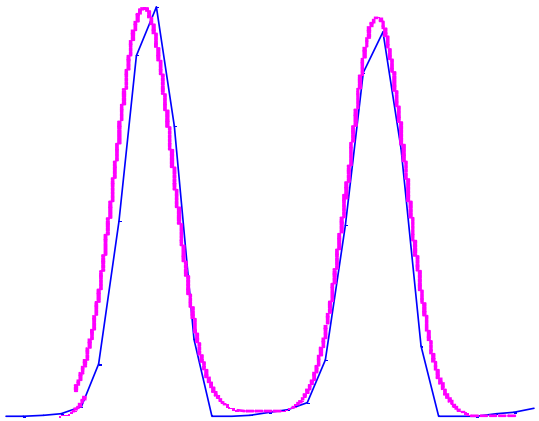
Deisotoping an Isotopic Distribution



- Patterson (Autocorrelation) algorithm to detect charge of a peak in a complex spectrum

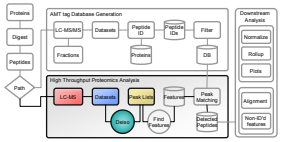


- Mercury algorithm used to guess an average empirical formula for a given mass
  - Averagine empirical formula of $C_{4.9384}\ H_{7.7583}\ N_{1.3577}\ O_{1.4773}\ S_{0.0417}$ → $C_{83}\ H_{124}\ N_{23}\ O_{25}\ S$ for 1876.02 Da
- Fitness (fit) functions to quantitate quality of match between theoretical and observed profiles

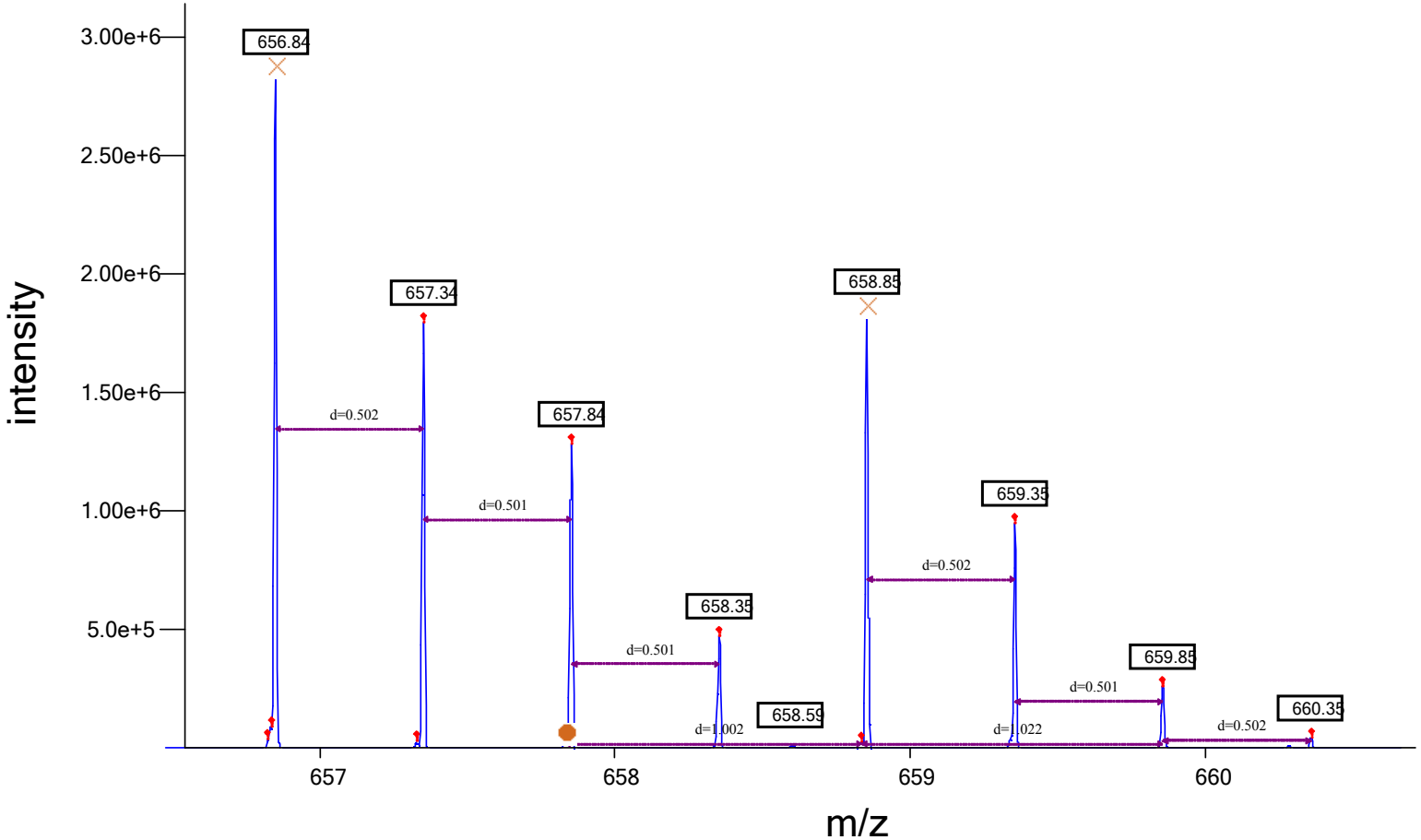


- For additional details, see the slides presented at 2007 US HUPO, available at http://ncrr.pnl.gov/training/workshops/

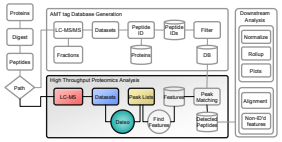
$^{16}O/^{18}O$ Mixtures



- Overlapping isotope patterns are separated by 4 Da
  - Creates challenges for deisotoping, particularly for charge states of 3+ or higher

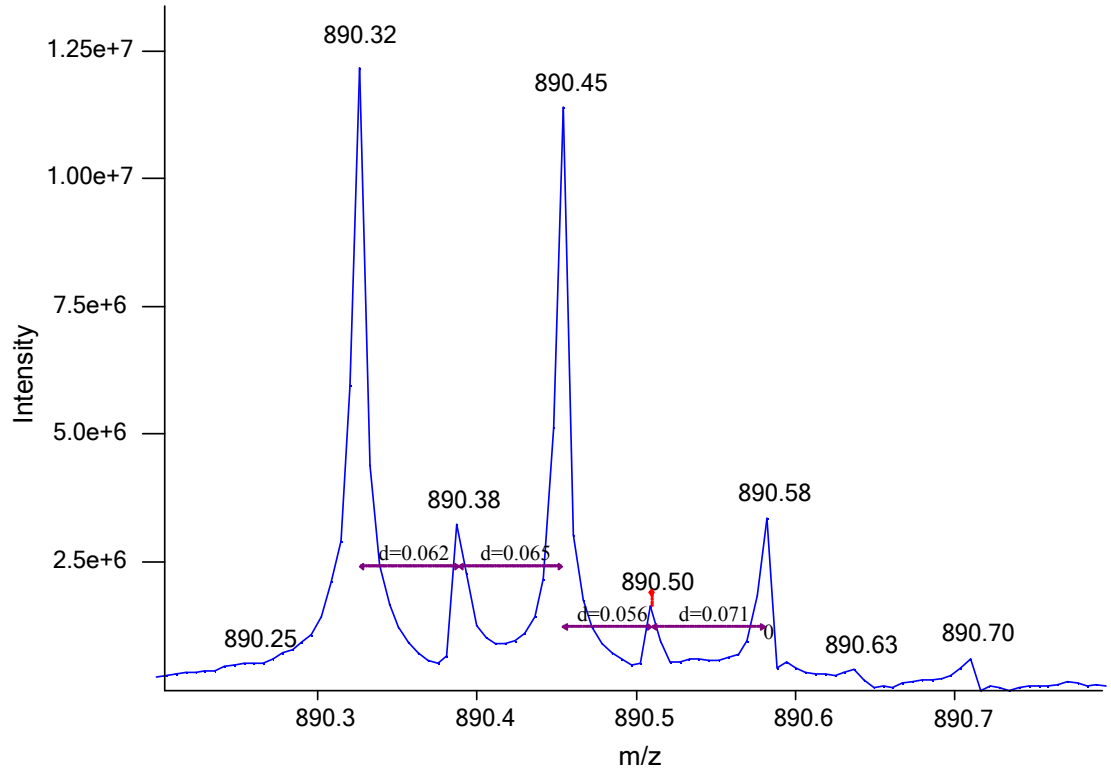


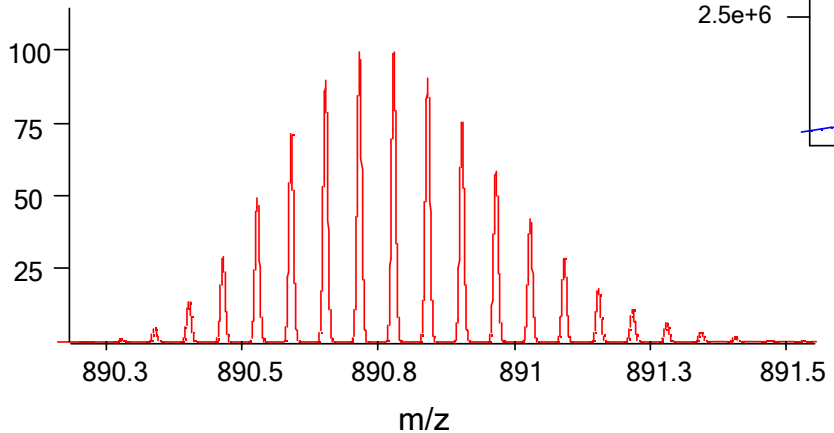
Isotopic Composition



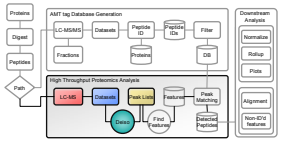
- Deviation from natural abundances
  - In ^{13}C, ^{15}N depleted media, isotopic composition of atoms is different from those found in nature
  - E.g., sulfur isotopes predominate the distribution at right
  - Contrast with an isotopic distribution of a peptide with similar mass and charge (16+), but a natural atomic distribution (below)

16+ sulfur containing peptide





Isotopic Composition



- Decon2LS supports changing the isotope composition

Transform Options

- Peak Picking
- Horn Transform
- Isotope Distribution
  - Averagine
  - Isotopic Composition
- Miscellaneous Options
- FTICR Preprocessing Options

Compostion

Atomicity	Element Nam	Element Symbol	Isotope Mass	Isotope Percentage
1	Hydrogen	H	1.007825	0.99985
1	Hydrogen	H	2.014102	0.00015
2	Helium	He	3.01603	1E-06
2	Helium	He	4.0026	0.999999
3	Lithium	Li	6.015121	0.075
3	Lithium	Li	7.016003	0.925
4	Berellium	Be	9.012182	1
5	Boron	B	10.01294	0.199
5	Boron	B	11.00931	0.801
6	Carbon	C	12	0.98893
6	Carbon	C	13.00336	0.01107
7	Nitrogen	N	14.00307	0.996337
7	Nitrogen	N	15.00011	0.003663
8	Oxygen	O	15.99491	0.99759

Load | Save | Save As

Helpful Tips

Controls Isotopic Composition

- Loads in the isotopic composition in the form of a .xml document

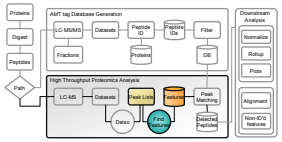
- Values can be edited in the grid and saved in original format or as new .xml

Load Parameters | Save Parameters | OK | Cancel

Part II: LC-MS Feature Discovery

- Part I: Introduction and Overview of Label-Free Quantitative Proteomics (Anderson)
- Part II: Feature Discovery in LC-MS Datasets (Monroe and Polpitiya)
  - ✓ Structure of LC-MS Data
  - ✓ Feature discovery in individual spectra (deisotoping)
  - Feature definition over elution time
  - Identifying LC-MS Features using an AMT tag DB
  - Extending the AMT tag approach for feature based analyses
  - Estimating confidence of identified LC-MS features
  - Downstream quantitative analysis with DAnTE
- Break
- Part III: Biological Application of the AMT tag Approach (Ansong)
- AMT tag Analysis Software Demo
- Panel Discussion

Feature Definition over Elution Time

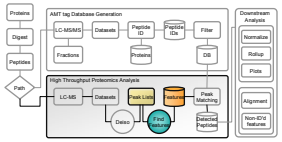


- Deisotoping collapses original data into data lists

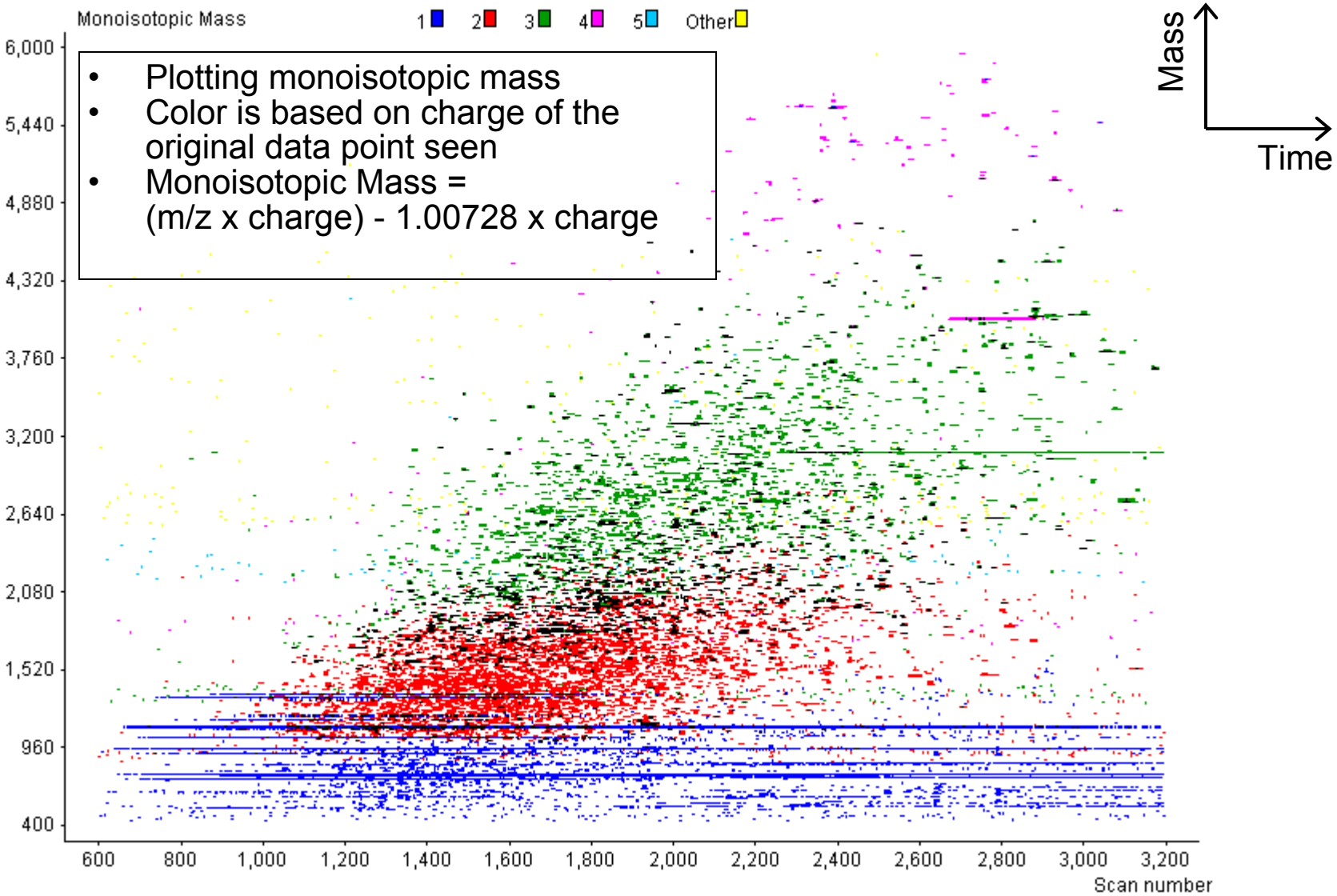
scan num	charge	abundance	mz	fit	average mw	monoiso mw	most abu. mw	fwhm	signal noise
1500	1	2772933	759.0649	0.0716	758.5222	758.0576	758.0576	0.0106	718.83
1500	1	2614913	1103.033	0.1111	1102.698	1102.026	1102.026	0.0222	74.04
1500	1	2422829	864.4919	0.0156	864.0073	863.4846	863.4846	0.0137	74.75
1500	2	2297822	563.3253	0.012	1125.322	1124.636	1124.636	0.006	77.94
1500	1	1213607	943.9815	0.1025	943.5518	942.9742	942.9742	0.0165	120.36
1500	3	988761	675.0246	0.02	2023.375	2022.052	2023.0549	0.0086	79.22
1500	2	734070	688.392	0.0384	1375.694	1374.77	1374.7695	0.009	92.09
1500	2	663954	642.3243	0.0253	1283.417	1282.634	1282.6341	0.0076	109.01
1500	1	661477	730.1117	0.024	729.5461	729.1045	729.1045	0.0096	39.06
1500	2	630657	689.3645	0.0446	1377.64	1376.715	1376.7145	0.0088	57.52
1500	2	569896	591.8343	0.0198	1182.379	1181.654	1181.6541	0.0065	111.2
1500	2	503993	757.8854	0.0706	1513.762	1512.753	1512.7533	0.0105	80.4
1500	2	451007	936.9389	0.0296	1873.091	1871.863	1872.8662	0.0156	46.74

- Goal: Given series of deisotoped mass spectra, group related data across elution time
  - Look for repeated monoisotopic mass values in sequential spectra, allowing for missing data
  - Can also look for expected chromatographic peak shape

Feature Definition over Elution Time

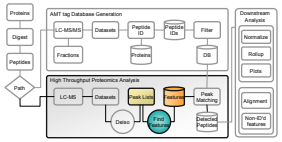


- Can visualize deisotoped data in two-dimensions



- Plotting monoisotopic mass
- Color is based on charge of the original data point seen
- Monoisotopic Mass = (m/z x charge) - 1.00728 x charge

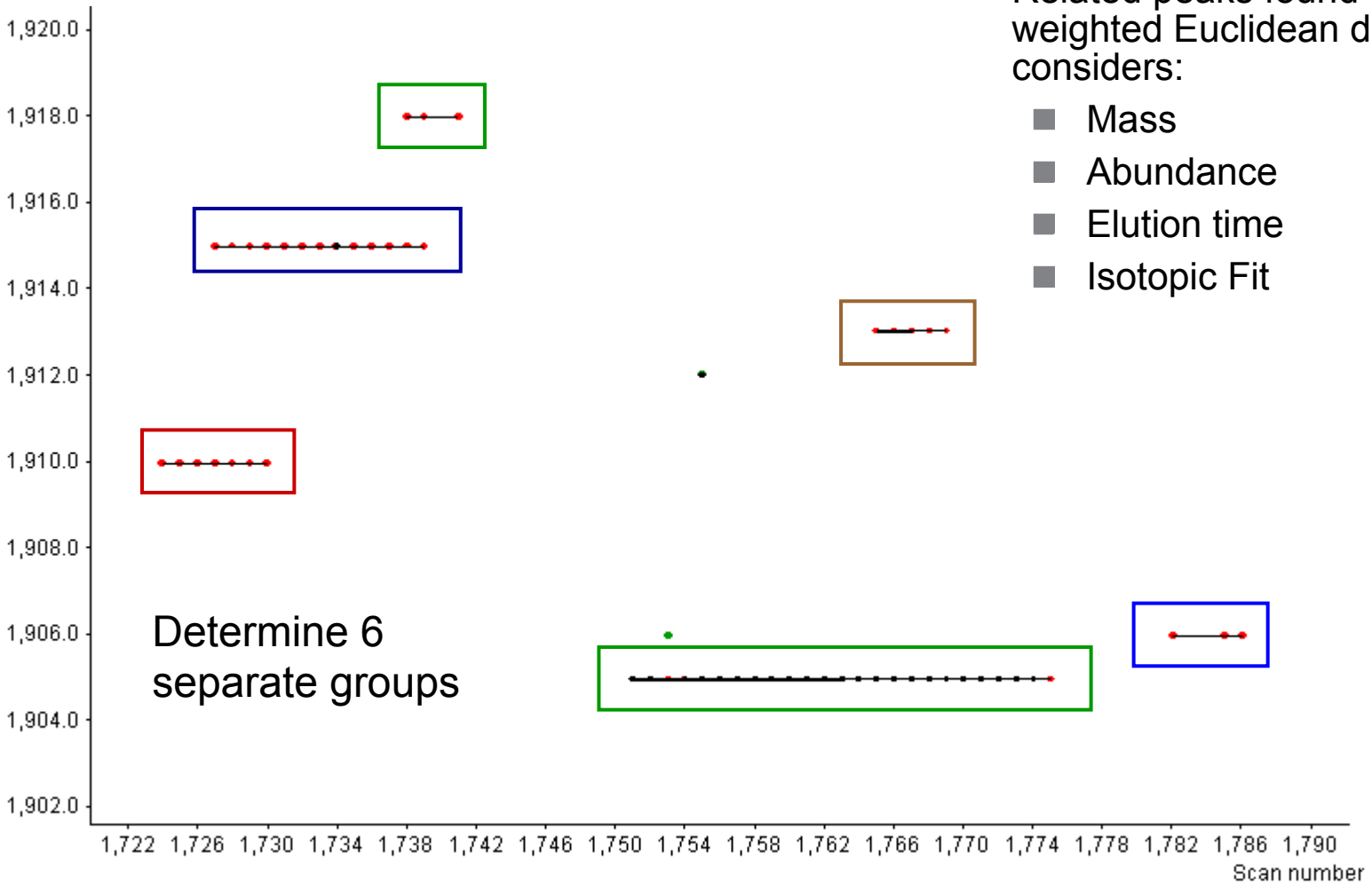
Feature Definition over Elution Time



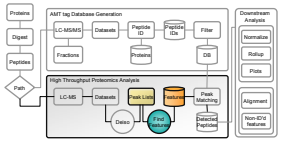
- Zoom-in view of species
  - Same species in multiple spectra need to be grouped together

Related peaks found using a weighted Euclidean distance; considers:

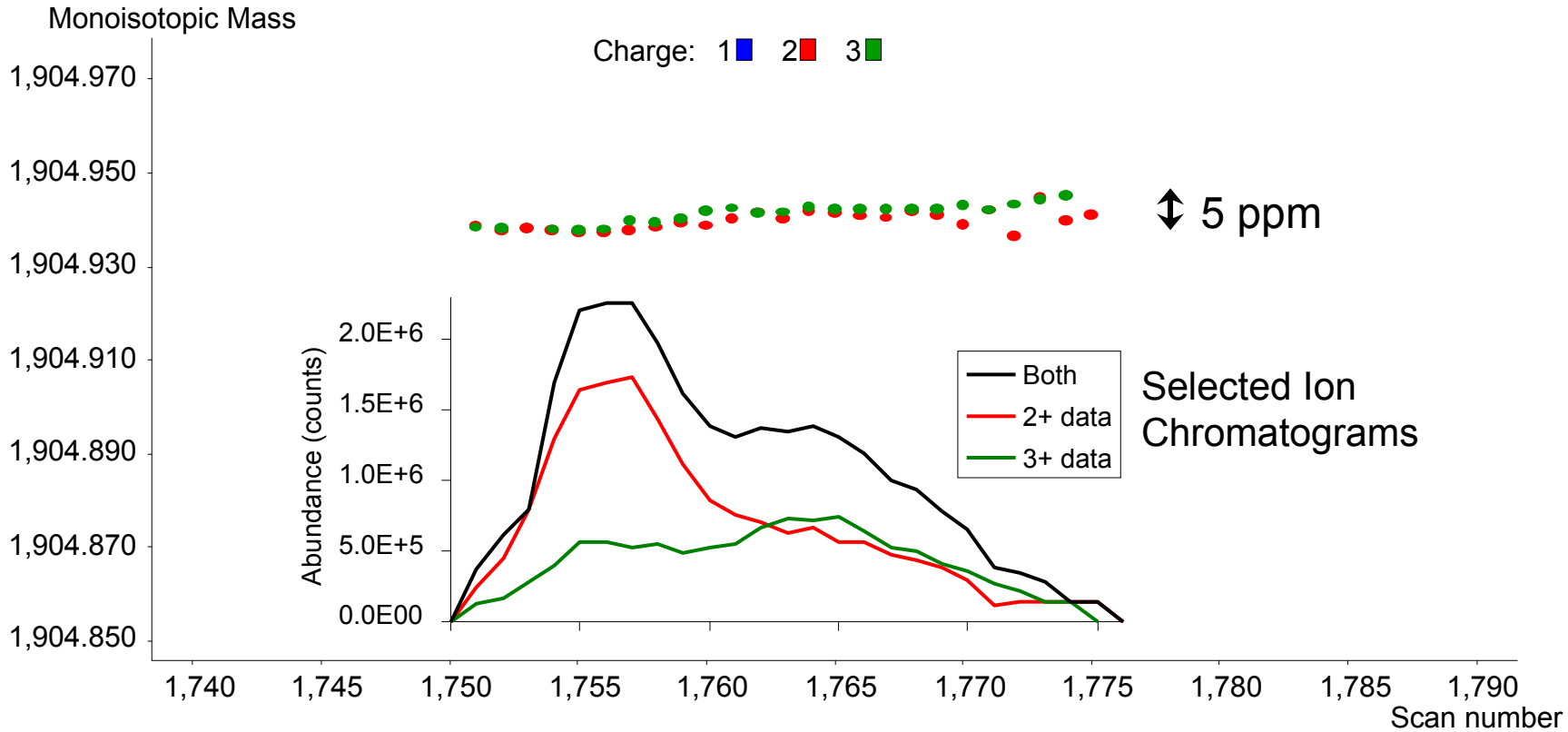
- Mass
- Abundance
- Elution time
- Isotopic Fit



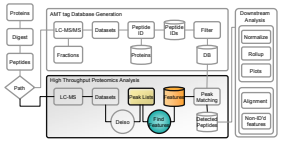
Feature Definition over Elution Time



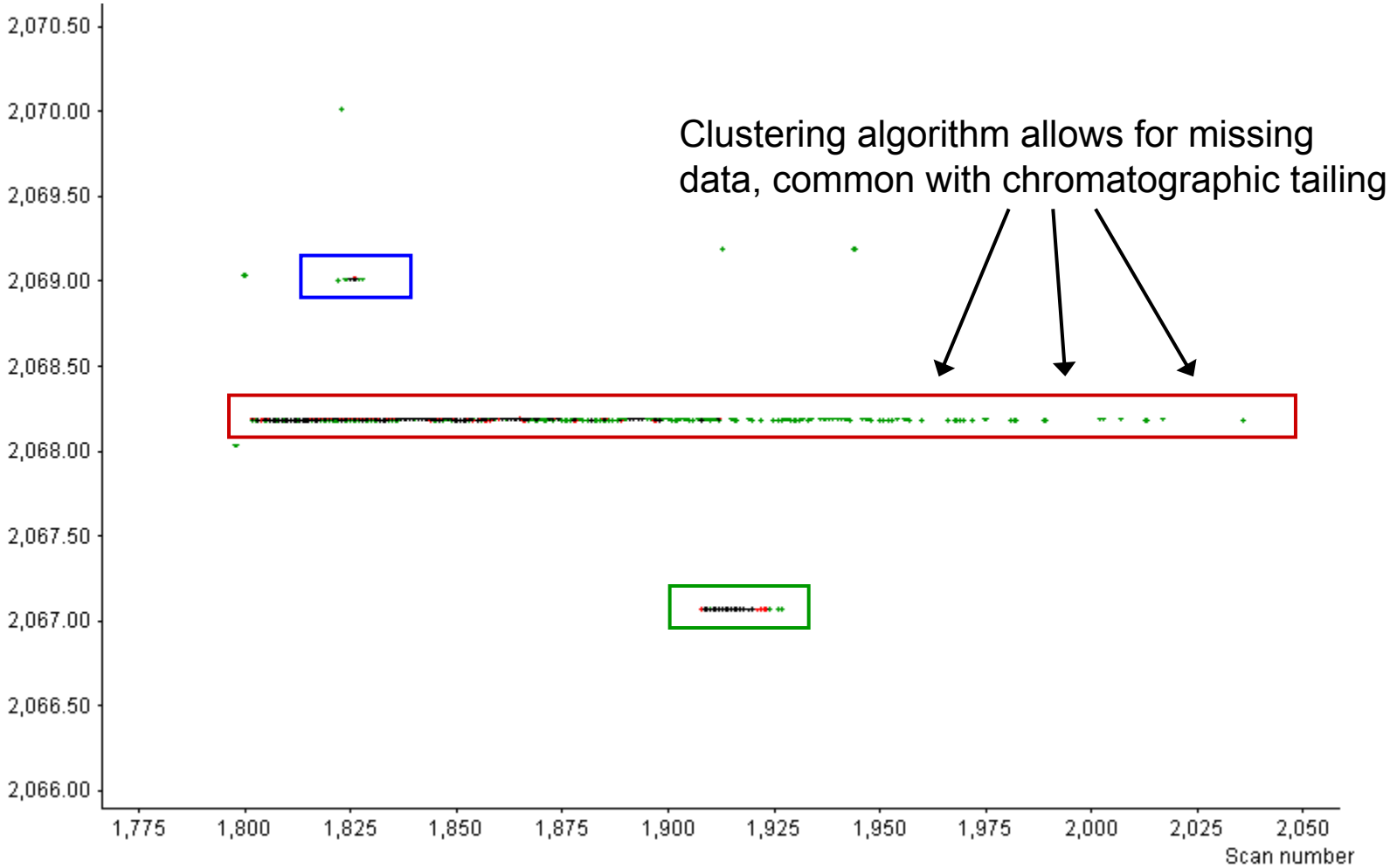
- Feature detail
  - Median Mass: 1904.9399 Da (more tolerant to outliers than average)
  - Elution Time: Scan 1757 (0.363 NET, aka normalized elution time)
  - Abundance: 1.7×10^7 counts (area under 2+ SIC)
    - See both 2+ and 3+ data
    - Stats typically come from the most abundant charge state



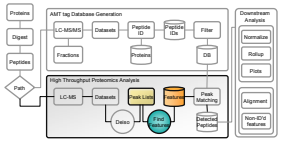
Feature Definition over Elution Time



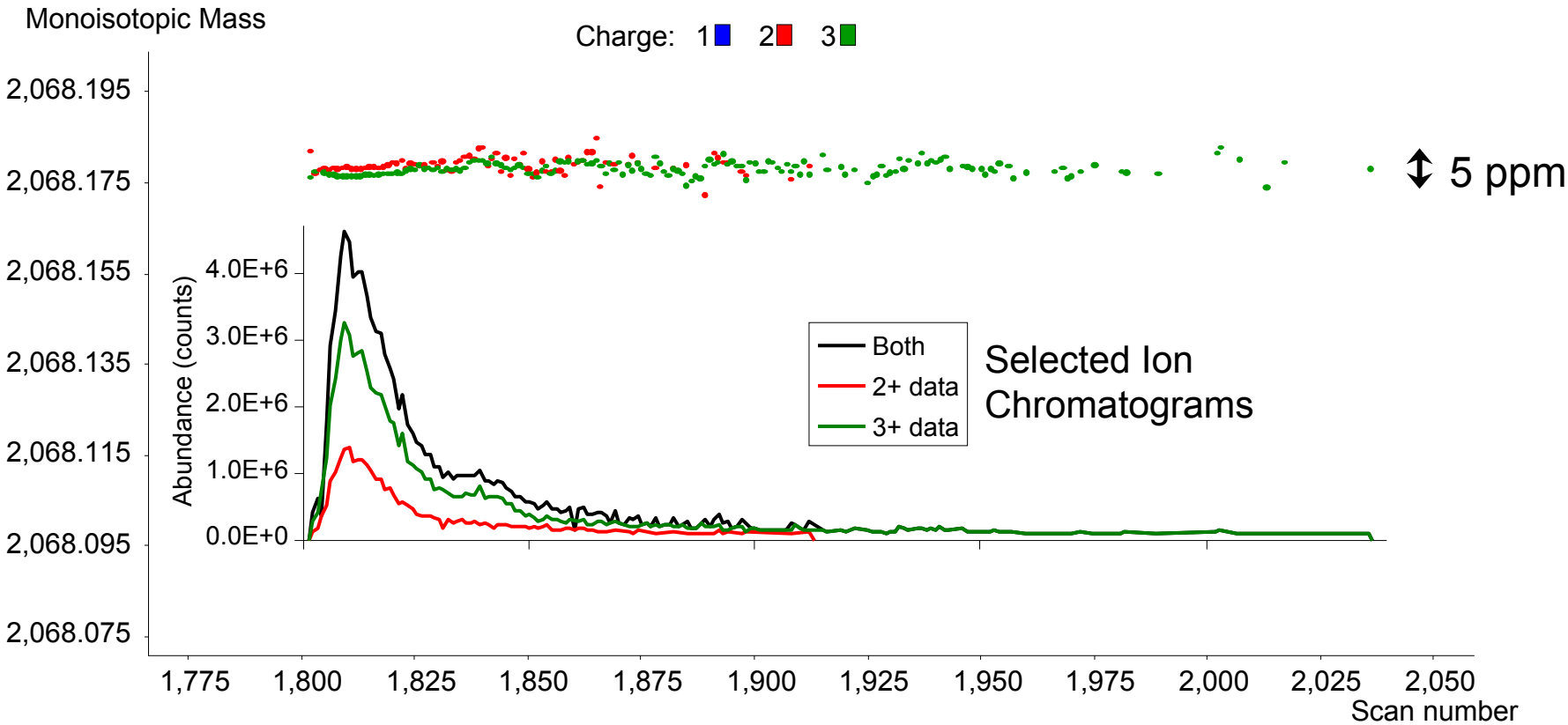
- Second example
  - LC-MS feature eluting over 7.5 minutes



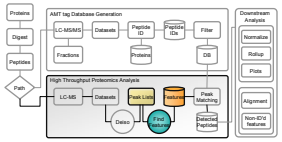
Feature Definition over Elution Time



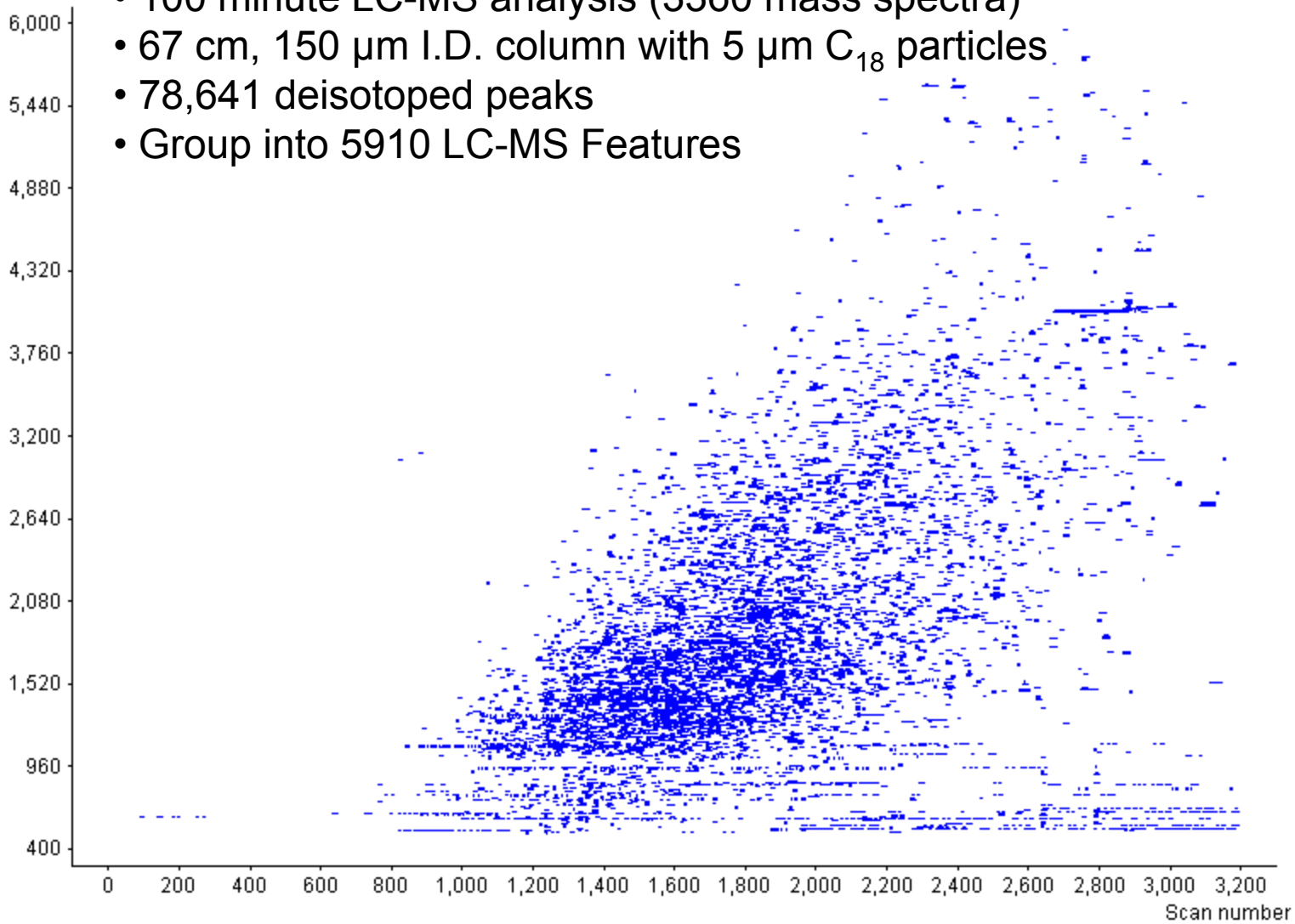
- Second example, feature detail
  - Median Mass: 2068.1781 Da
  - Elution Time: Scan 1809 (0.380 NET)
  - Abundance: 8.7×10^7 counts (area under 3+ SIC)
    - See both 2+ and 3+ data, though 3+ is more prevalent



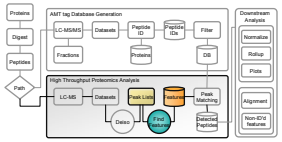
Feature Definition over Elution Time



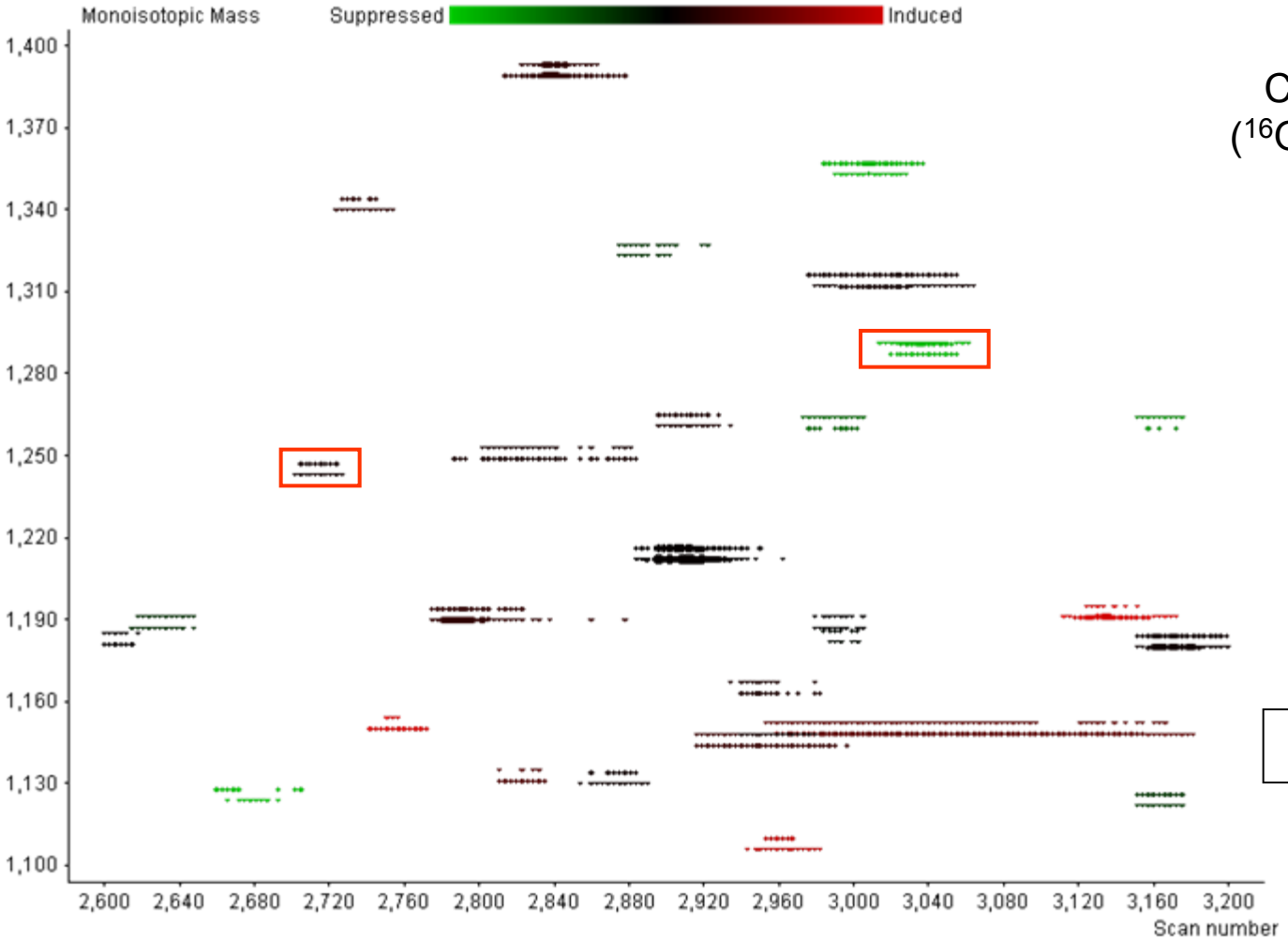
- Example: *S.* Typhimurium dataset on 11T FTICR
  - 100 minute LC-MS analysis (3360 mass spectra)
  - 67 cm, 150 µm I.D. column with 5 µm C_{18} particles
  - 78,641 deisotoped peaks
  - Group into 5910 LC-MS Features

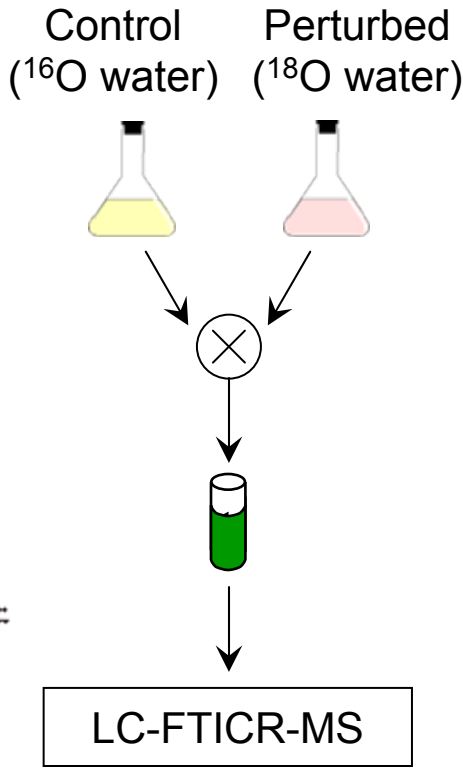


Isotopic Pairs Processing

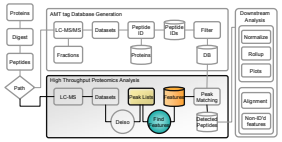


- Paired features typically have identical sequences, with and without an isotopic label
  - e.g. $^{16}O/^{18}O$ pairs have 4 Da spacing due to two ^{18}O atoms

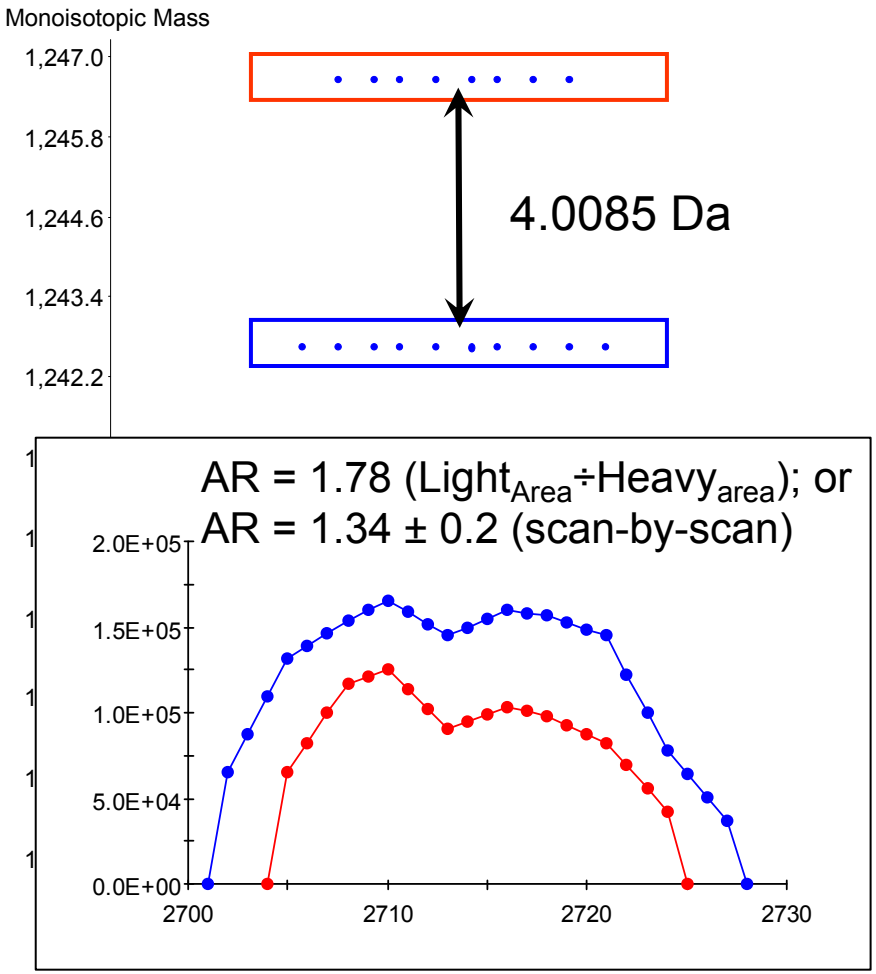


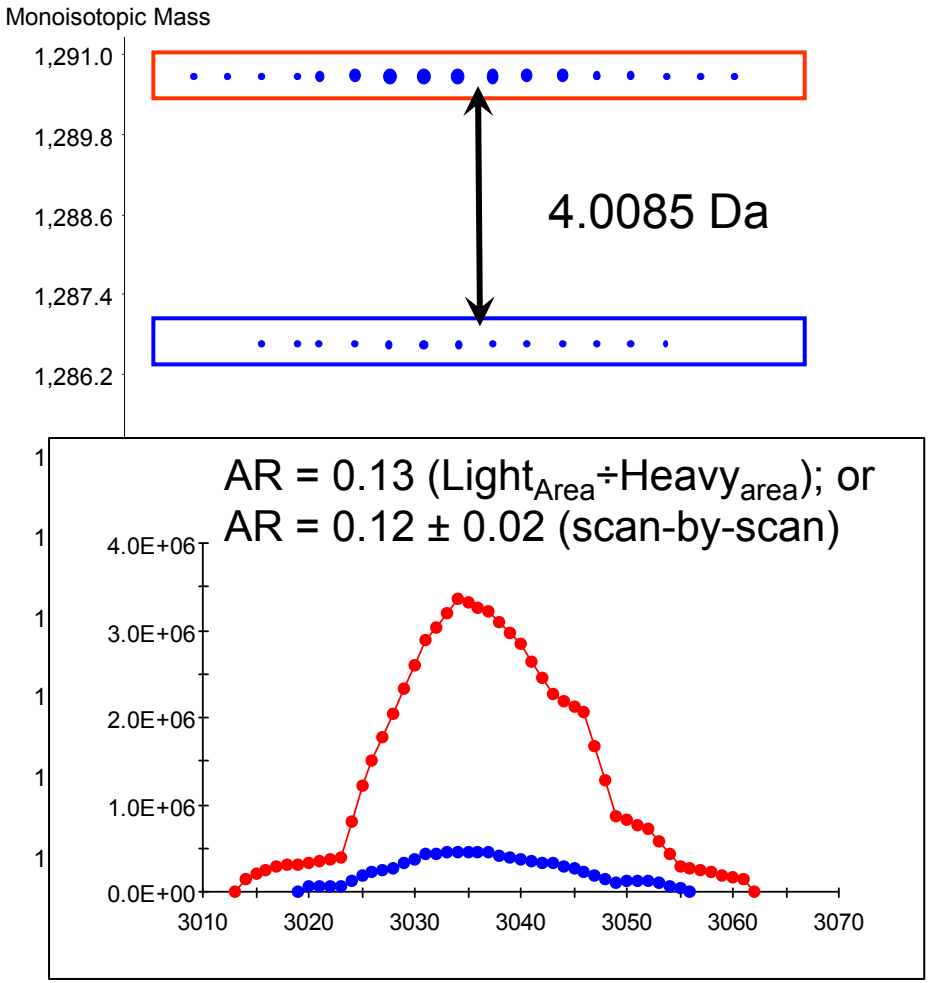


Isotopic Pairs Processing

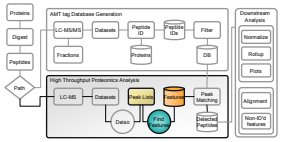


- Paired feature example: $^{16}O/^{18}O$ data
  - Compute AR using ratio of areas, or
  - Compute AR scan-by-scan, then average AR values (members must co-elute)





Feature Definition over Elution Time



- Numerous options in VIPER for clustering data to form LC-MS features and for finding paired features

LC-MS Feature (UMC) Ion Networks

1. Find Connections | 2. Edit/Filter Connections | 3. Define LC-MS Features using Connections

Definition Scope
- All Data Points
- Current View

Metric Type: Euclidean — Set to Old Defaults — Set to Defaults

☑ Use	Monoisotopic Mass	Wt. Factor	0.01	Constraint	L.T.	10	ppm
☐ Use	Average Mass	Wt. Factor	0.01	Constraint	L.T.	10	ppm
☑ Use	Log (Abundance)	Wt. Factor	0.1	Constraint	None		
☑ Use	Generic NET	Wt. Factor	15	Constraint	None		
☑ Use	Fit	Wt. Factor	0.1	Constraint	None		

Reject connection longer than: 0.1

LC-MS Feature (UMC) Ion Networks

1. Find Connections | 2. Edit/Filter Connections | 3. Define LC-MS Features using Connections

LC-MS Feature Stats
- Class Representative: Highest Abundance
- Class Abundance: Sum of Class Abu.
- Class Mass: Class Median
- Most Abu Charge State Group Type: Highest Abu Sum
- ☑ Use most abundant charge state group stats for class stats
- ☐ Make single member classes

FeatureDraw Type: Actual LC-MS Feature

Auto-Refine Options | Split Features Options | Adv Class Stats

- ☐ Remove low intensity classes 30 %
- ☐ Remove high intensity classes 30 %
- ☑ Remove cls. with less than 3 scans
- ☐ Remove cls. with length over 400 scans
- ☑ Remove cls. with length over 15 % all scans
- Percent max abu for gauging width 33 %
- ☑ Test feature length using scan range — Minimum member count: 3

☑ Interpolate abundances across gaps
Maximum size of gap to interpolate: 4

Set to Defaults — Report — Find LC-MS Features

LC-MS Feature Delta Pairing Analysis

Function Report

Find Pairs — Set to Defaults

Delta Mass Options
- Delta: 4.0085
- ☐ Calculate N14/N15 Min/Max Deltas from class molecular mass
- Min Deltas: 1 — Max Deltas: 1
- Delta count step size: 1
- Set to N14/N15 — Set to C12/C13 — Set to O16/O18 — Set to Deuterium

Tolerance Options
- Pair Tolerance: 0.02 Da — Scan Tolerance:
- ☑ Require pair-classes overlap at feature edges 15

Note: Even if two LC-MS Features do not overlap at the edges, if one feature is completely within a second feature, then pairing is allowed; to prevent this, enable overlap at LC-MS Feature apexes

- ☑ Require pair-classes overlap at feature apexes 15

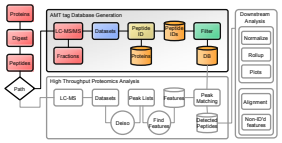
Inclusion/Exclusion Options
- ER Inclusion Range: -5 to 5
- ☑ Ambiguous pairs exclusion keeps most confident pair

Pair Search and ER Calculation Options
- ☑ Require matching charge states for pair
- ☑ Use identical charge states for expression ratio
- ☑ Average ER's for all charge states — Weight by Abu
- ☑ Compute ER Scan by Scan
- ☑ Enable I-Report ER computation
- ☑ Remove outlier ER values using Grubb's test (95% conf.)
  - ☑ Repeatedly remove outliers
  - ☑ Use symmetric ERs
  - Minimum final data point count: 3

Part II: LC-MS Feature Discovery

- Part I: Introduction and Overview of Label-Free Quantitative Proteomics (Anderson)
- Part II: Feature Discovery in LC-MS Datasets (Monroe and Polpitiya)
  - ✓ Structure of LC-MS Data
  - ✓ Feature discovery in individual spectra (deisotoping)
  - ✓ Feature definition over elution time
  - Identifying LC-MS Features using an AMT tag DB
  - Extending the AMT tag approach for feature based analyses
  - Estimating confidence of identified LC-MS features
  - Downstream quantitative analysis with DAnTE
- Break
- Part III: Biological Application of the AMT tag Approach (Ansong)
- AMT tag Analysis Software Demo
- Panel Discussion

Assembling an AMT tag DB



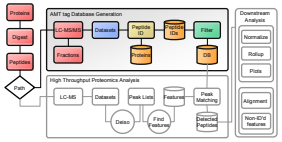
- Accurate Mass and Time (AMT) tag
  - Unique peptide sequence whose monoisotopic mass and normalized elution time are accurately known
  - AMT tags also track any modified residues in peptide
- AMT tag DB
  - Collection of AMT tags
- AMT tag approach articles
  - R.D. Smith et. al., *Proteomics* **2002**, *2*, 513-523.
  - J.S. Zimmer, M.E. Monroe et. al., *Mass Spec. Reviews* **2006**, *25*, 450-482.
  - L. Shi, J.N. Adkins, et. al., *J. of Biological Chem.* **2006**, *281*, 29131-29140.

More info

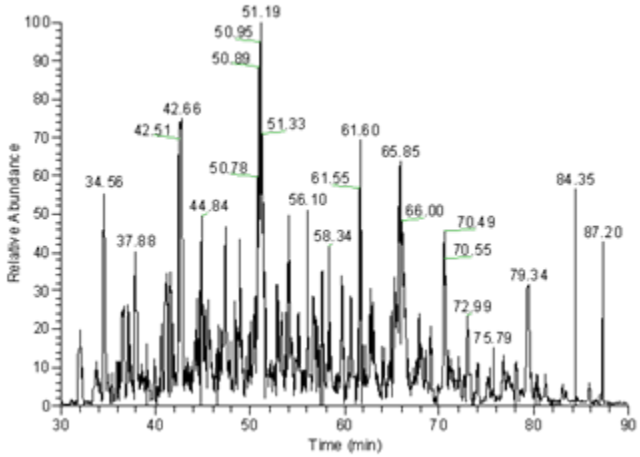
Assembling an AMT tag DB

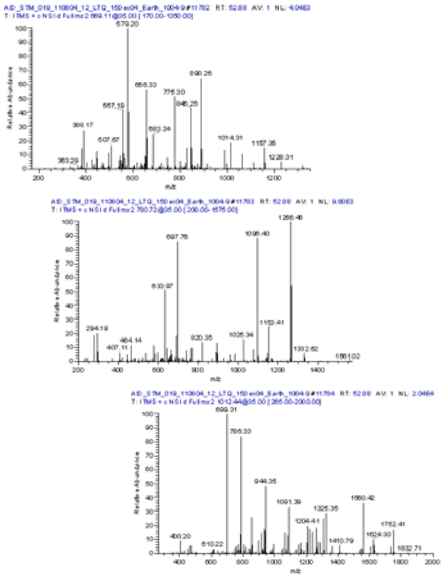
- What can we use an AMT tag DB for?
  - Query LC-MS/MS data to answer questions
    - How many distinct peptides were observed passing filter criteria?
    - Which peptides were observed most often by LC-MS/MS?
    - How many proteins had 2 or more partially or fully tryptic peptides?
  - Correlate LC-MS features to the AMT tags
    - Analyze multiple, related samples by LC-MS using a high mass accuracy mass spectrometer
      - e.g. Time course study, 5 data points with 3 points per sample
    - Characterize the LC-MS features
      - Deisotope to obtain monoisotopic mass and charge
      - Cluster in time dimension to obtain abundance information
    - Match to AMT tags to identify peptides
      - Align in mass and time dimensions
      - Match mass and time of LC-MS features to mass and time of AMT tags

Assembling an AMT tag DB



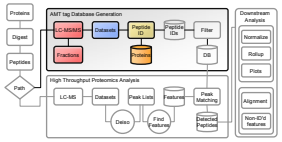
- Characterizing AMT tags
  - Analyze samples by LC-MS/MS
    - 10 minute to 180 minute LC separations
    - Obtain 1000's of MS/MS fragmentation spectra for each sample
  - Analyze spectra using SEQUEST, X!Tandem, InsPecT, etc.
    - SEQUEST: http://www.thermo.com/bioworks/
    - X!Tandem: http://www.thegpm.org/TANDEM/
    - InsPecT: http://bix.ucsd.edu/Software/Inspect.html
  - Collate results



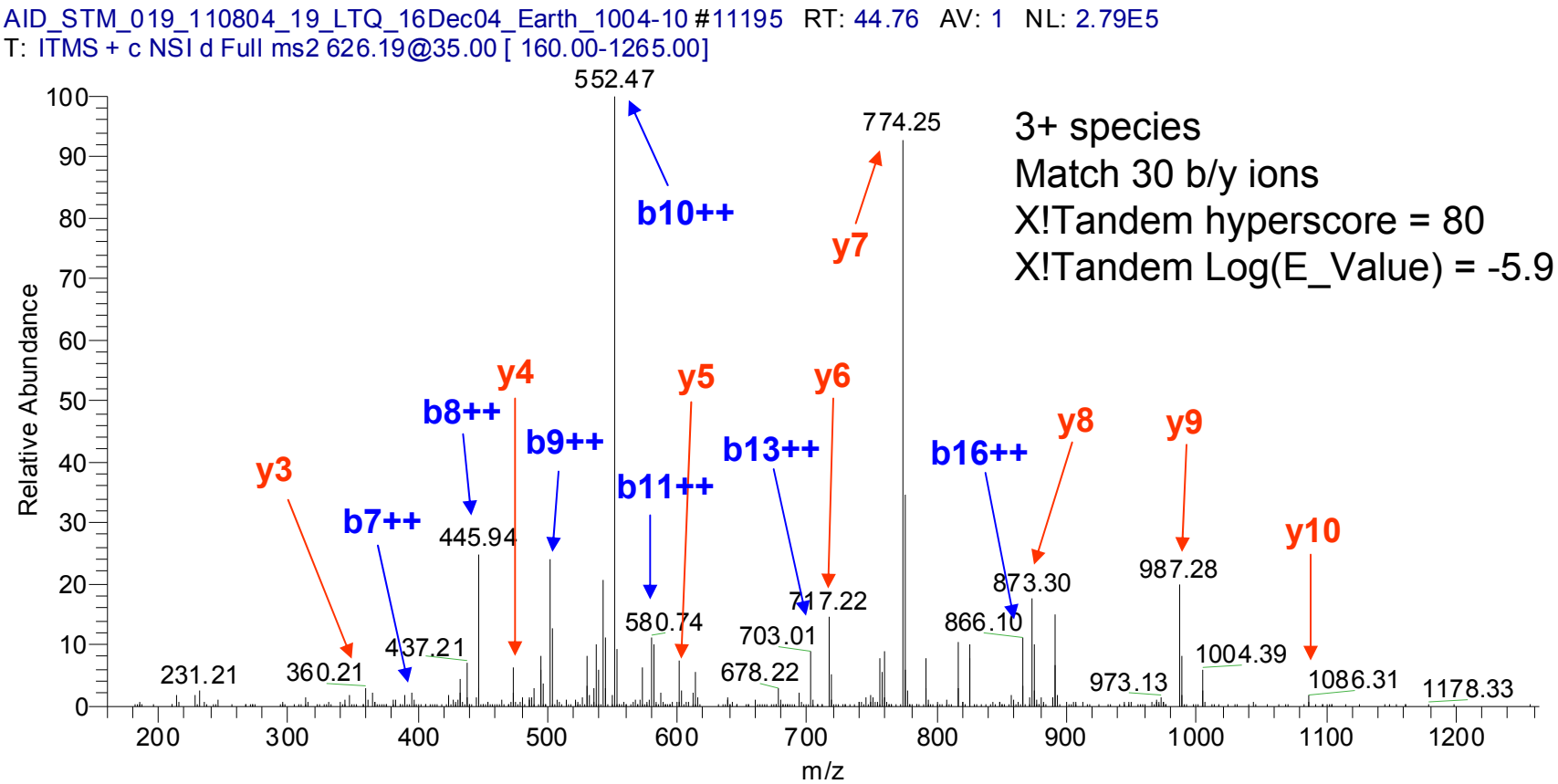


List of peptide and protein matches

Assembling an AMT tag DB

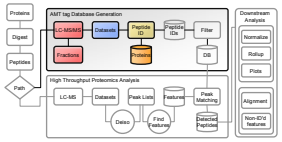


- AMT tag example
  - R.VKHPSEIVNVGDEINVK.V
  - Observed in scan 11195 of dataset #19 in an SCX fractionation series



More info

Assembling an AMT tag DB

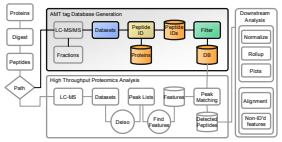


- AMT tag example
  - R.VKHPSEIVNVGDEINVK.V
  - Observed in scan 11195 of dataset #19 in an SCX fractionation series

#	Immon.	b	b++	Seq.	y	y++	#
1	72.08			V	1877.01	939.01	17
2	101.11	228.17		K	1777.94	889.48	16
3	110.07	365.23		H	1649.85	825.43	15
4	70.07	462.28		P	1512.79	756.90	14
5	60.04	549.31		S	1415.74	708.37	13
6	102.06	678.36	339.68	E	1328.71	664.86	12
7	86.10	791.44	396.22	I	1199.66	600.33	11
8	72.08	890.51	445.76	V	1086.58	543.79	10
9	87.06	1004.55	502.78	N	987.51	494.26	9
10	72.08	1103.62	552.31	V	873.47	437.24	8
11	30.03	1160.64	580.83	G	774.40	387.70	7
12	88.04	1275.67	638.34	D	717.38	359.19	6
13	102.06	1404.71	702.86	E	602.35	301.68	5
14	86.10	1517.80	759.40	I	473.31		4
15	87.06	1631.84	816.42	N	360.22		3
16	72.08	1730.91	865.96	V	246.18		2
17	101.11			K	147.11		1

3+ species
Match 30 b/y ions
X!Tandem hyperscore = 80
X!Tandem Log(E_Value) = -5.9

Assembling an AMT tag DB

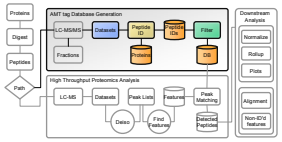


- Align related datasets using elution times of observed peptides
  - One option: utilize NET prediction algorithm to create theoretical dataset to align against
    - NET prediction uses position and ordering of amino acid residues to predict normalized elution time

Peptide	X!Tandem Log (E_Value)	Elution Time	Predicted NET
R.AARPAKYSYVDENGETK.T	-6.1	33.958	0.167
R.LVHGEEGLVAAKR.I	-8.8	36.915	0.224
R.GIIKVGEEVEIVGIK.E	-8.2	53.003	0.415
K.RFNDDGPILFIHTGGAPALFAYHPHV.-	-7.3	62.583	0.519
K.KTGVLAQVQEALKGLDVR.E	-11.6	62.803	0.438
R.KVAAQIPNGSTLFIDIGTTPEAVAHALLGHSNLR.I	-8.9	73.961	0.589
R.TFAISPGHMNQLRAESIPEAVIAGASALVLTSYLVR.C	-6.5	88.043	0.764

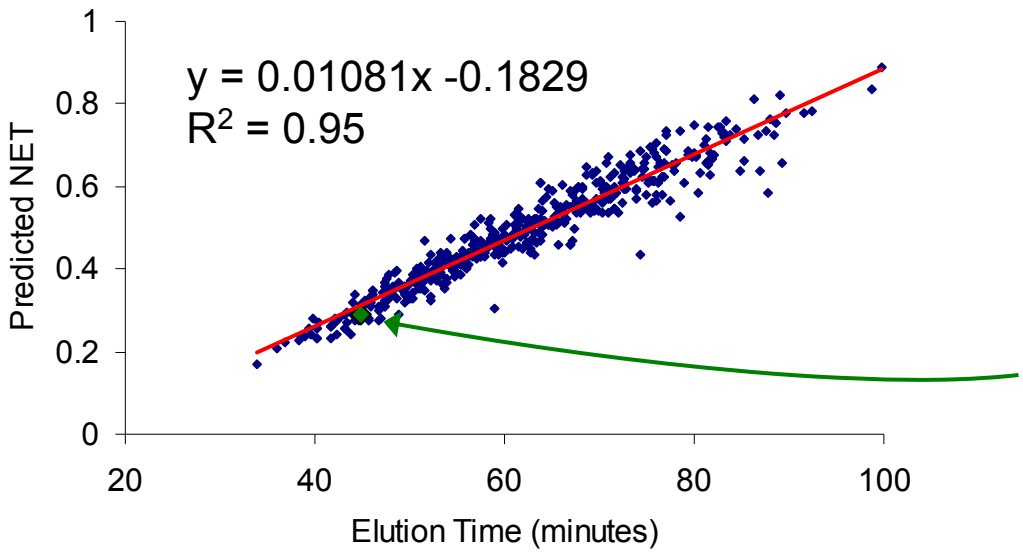
K. Petritis, L.J. Kangas, P.L. Ferguson, et al., *Analytical Chemistry* **2003**, *75*, 1039-1048.
K. Petritis, L.J. Kangas, B. Yan, et al., *Analytical Chemistry* **2006**, *78*, 5026-5039.

Assembling an AMT tag DB



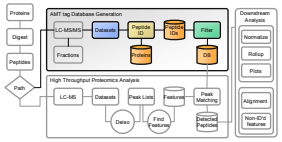
- Align related datasets using elution times of observed peptides
  - One option: utilize NET prediction algorithm to create theoretical dataset to align against
    - NET prediction uses position and ordering of amino acid residues to predict normalized elution time
  - Alignment yields NET values based on observed elution times
    - Observed NET = Slope×(Observed Elution Time) + Intercept

Example: 506 unique peptides used for alignment; Log(E_Value) ≤ -6

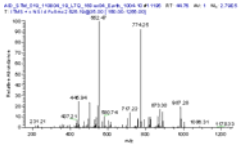


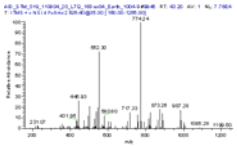
VKHPSEIVNVGDEINVK
Elution time: 44.923 minutes
Predicted NET: 0.292
Observed NET: 0.303

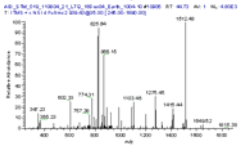
Assembling an AMT tag DB

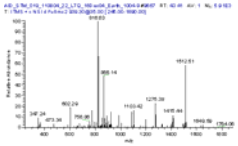


- AMT tag example
  - R.VKHPSEIVNVGDEINVK.V
  - Observed in 7 (of 25) LC-MS/MS datasets in the SCX fractionation series

Analysis 1, scan 11195  → 3+, hyperscore 80, Obs. NET 0.303

Analysis 2, scan 9945  → 3+, hyperscore 69, Obs. NET 0.298

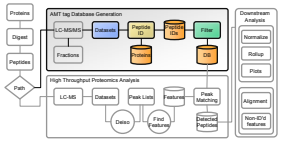
Analysis 3, scan 10905  → 2+, hyperscore 74, Obs. NET 0.301

Analysis 4, scan 9667  → 2+, hyperscore 77, Obs. NET 0.302

⋮ ⋮

Compute monoisotopic mass: 1876.0053 Da
Average Normalized Elution Time: 0.3021 (StDev 0.0021)

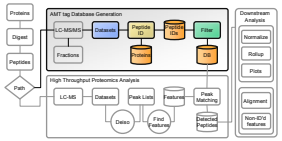
Assembling an AMT tag DB



- Mass and Time Tag Database
  - Repository for AMT tags
    - Mass, elution time, modified residues, and supporting information for each AMT tag
  - Allows samples of unknown composition to be matched quickly and efficiently, without needing to perform tandem MS
  - Assembled by analyzing a control set of samples, cataloging each peptide identification until subsequent analyses no longer provide new identifications

MT Tag ID	Peptide	LC-MS/MS Obs. Count	Calculated Monoisotopic Mass	Average Observed NET	Observed NET StDev
1662039	MTGRELKPHDR	1	1338.6826	0.143	0.000
17683899	SSALNTLTNQK	3	1175.6146	0.235	0.005
36609588	HRDLLGATNP…TLR	5	1960.0602	0.379	0.002
36715875	WVKVDGWDN…FER	11	2590.2815	0.459	0.011
36843675	MYGHLKGEVA…QER	8	2533.2304	0.557	0.005

Assembling an AMT tag DB



- Mini AMT tag DB
  - Database constructed from a relatively small number of datasets
  - e.g. 25 SCX fractionation samples from *S.* Typhimurium, each analyzed by LC-MS/MS and then by X!Tandem
  - Protein database: S_typhimurium_LT2_2004-09-19
    - 4550 proteins and 1.4 million residues

```
>STM1834 putative YebN family transport protein (yebN) {Salmonella typhimurium LT2}
MFAGGSDVFNGYPGQDVVMHFTATVLLAFGMSMDAFAASIGKGATLHKPKFSEALRTGLI
FGAVETLTPLIGWGLGILASKFVLEWNHWIAFVLLIFLGGRMIIEGIRGGSDEDETPLRR
HSFWLLVTTAIATSLDAMAVGVGLAFLQVNIIATALAIGCATLIMSTLGMMIGRFIGPML
GKRAEILGGVVLIGIGVQILWTHFHG
>STM1835 23S rRNA m1G745 methyltransferase (rrmA) {Salmonella typhimurium LT2}
MSFTCPLCHQPLTQINNSVICPQRHQFDVAKEGYINLLPVQHKRSRDPGDSAEMMQARRA
FLDAGHYQPLRDAVINLLRERLDQSATAILDIGCGEGYYTHAFAEALPGVTTFGLDVAKT
AIKAAAKRYSQVKFCVASSHRLPFADASMDAVIRIYAPCKAQELARVVKPGGWVVTATPG
PHHLMELKGLIYDEVRLHAPYTEQLDGFTLQQSTRLAYHMQLTAEAAVALLQMTPFAWRA
RPDVWEQLAASAGLSCQTDFNLHLWQRNR
```

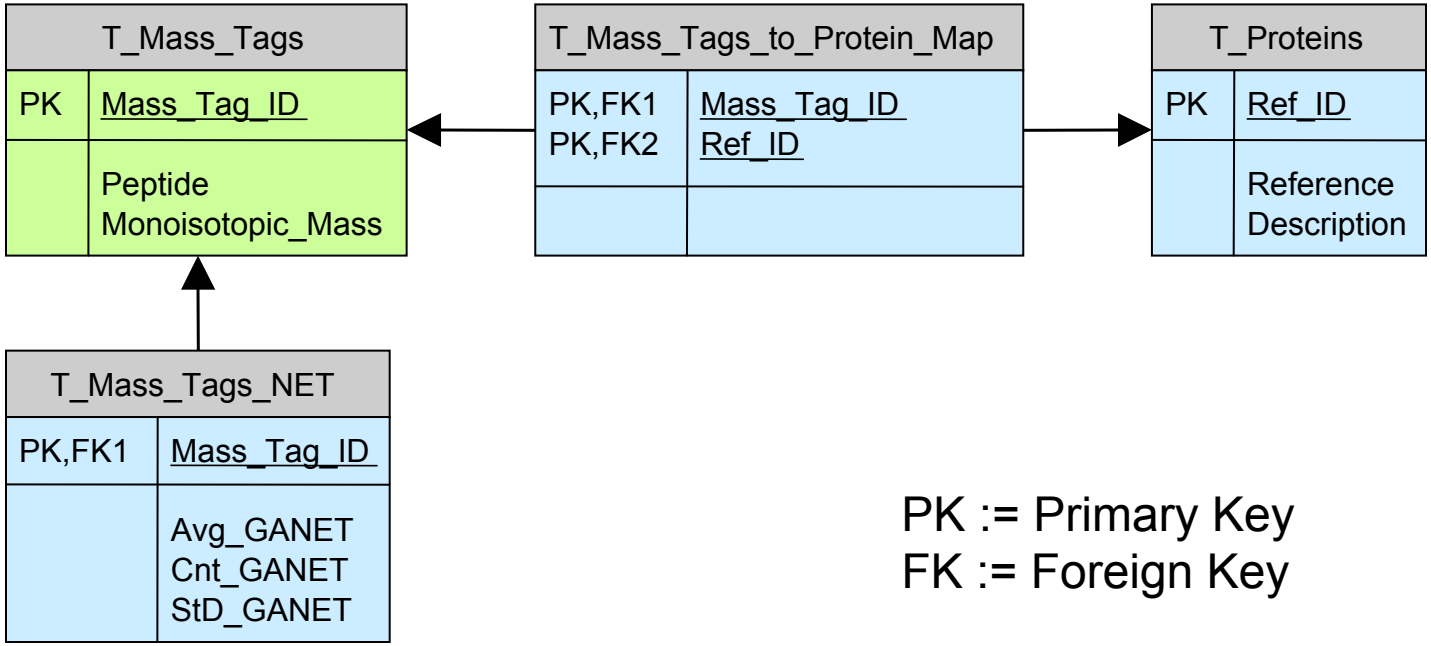
More info

Assembling an AMT tag DB

- Database Relationships
  - Minimum information required:
    - Single table with mass and normalized elution time (NET)

T_Mass_Tags	
PK	Mass_Tag_ID
	Peptide Monoisotopic_Mass NET

  - Expanded schema:

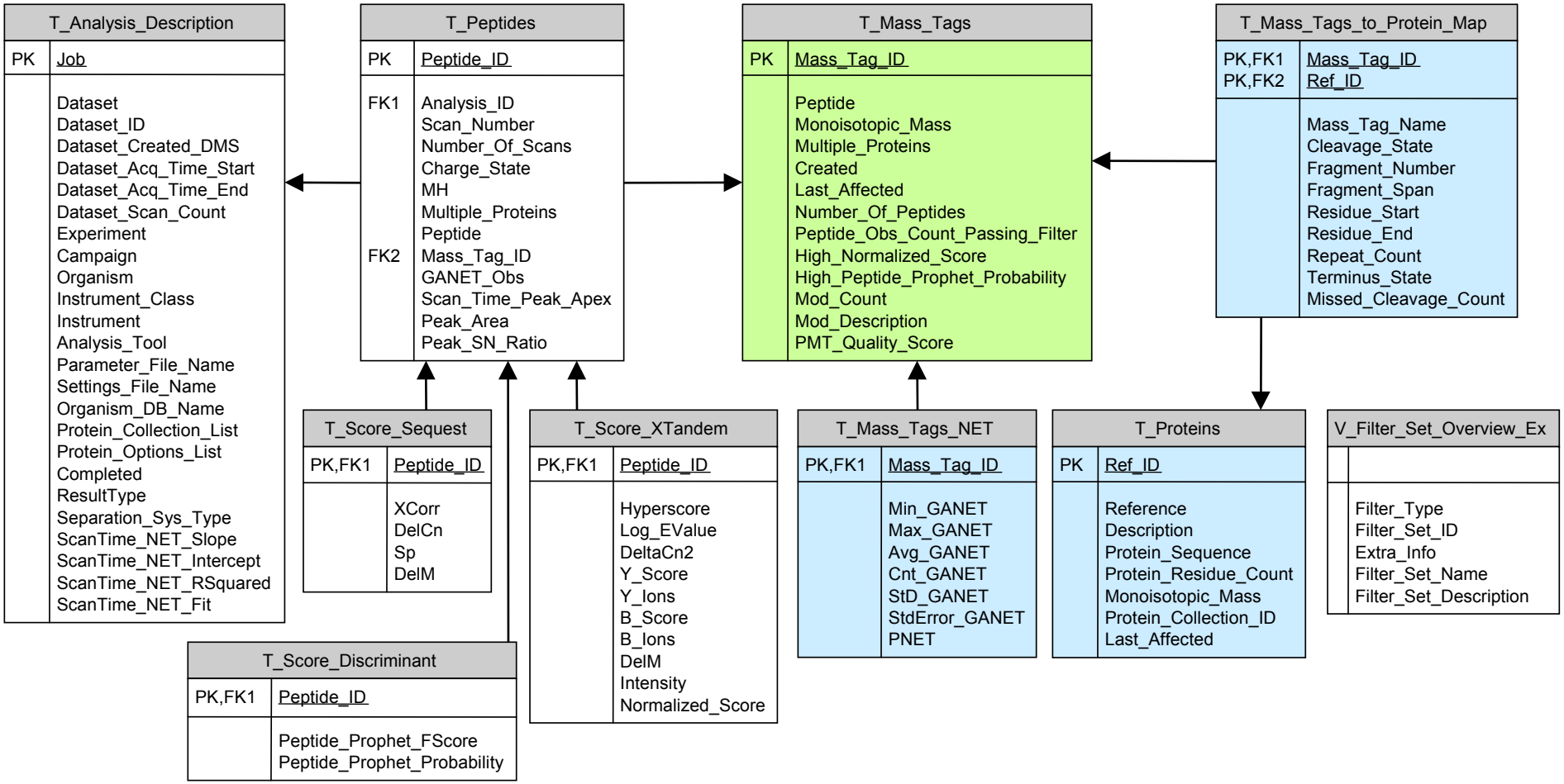


PK := Primary Key
FK := Foreign Key

More info

Assembling an AMT tag DB

- Microsoft Access DB Relationships
  - Full schema to track individual peptide observations



More info

Assembling an AMT tag DB

- Example data

T_Mass_Tags

Mass_Tag_ID	Peptide	Monoisotopic_Mass
24847	VKHPSEIVNVGDEINVK	1876.00533

T_Mass_Tags_NET

Mass_Tag_ID	Avg_GANET	Cnt_GANET	StD_GANET
24847	0.3021	7	2.11E-03

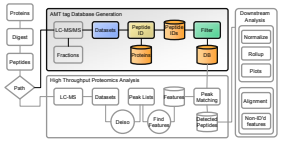
T_Peptides

Peptide_ID	Peptide	Mass Tag ID	Job	Scan Number	Charge State
53428	R.VKHPSEIVNVGDEINVK.V	24847	206386	11195	3
57461	R.VKHPSEIVNVGDEINVK.V	24847	206387	9945	3
61511	R.VKHPSEIVNVGDEINVK.V	24847	206388	10905	2
65386	R.VKHPSEIVNVGDEINVK.V	24847	206389	9667	2
69081	R.VKHPSEIVNVGDEINVK.V	24847	206390	9118	2
72556	R.VKHPSEIVNVGDEINVK.V	24847	206391	9159	2
76263	R.VKHPSEIVNVGDEINVK.V	24847	206392	9421	2

T_Score_XTandem

Peptide_ID	Hyperscore	Log(E_Value)
53428	80.2	-5.89
57461	69.2	-4.92
61511	74	-12.85
65386	77.2	-12.80
69081	69	-12.82
72556	78	-13.77
76263	60.3	-11.27

Assembling an AMT tag DB



- Processing steps

Thermo-Finnigan LTQ .Raw files

↓

MS/MS spectra files

↓

Peptide ID Results

↓

Tab delimited text files

↓

Summarized result files

↓

Microsoft Access DB

Convert to .Dta files or single _Dta.txt file using DeconMSn.exe. DeconMSn is similar to Thermo's Extract_MSn but has better support for data from LTQ-Orbitrap or LTQ-FT instruments.

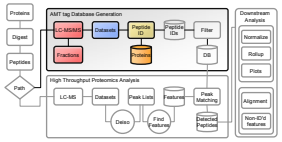
Process _Dta.txt file with X!Tandem or .Dta files with SEQUEST. Use the Peptide File Extractor to convert SEQUEST .Out files to Synopsis (_Syn.txt) files.

Convert X!Tandem .XML output files or SEQUEST _Syn.txt file to tab-delimited files using the Peptide Hit Results Processor (PHRP) application.

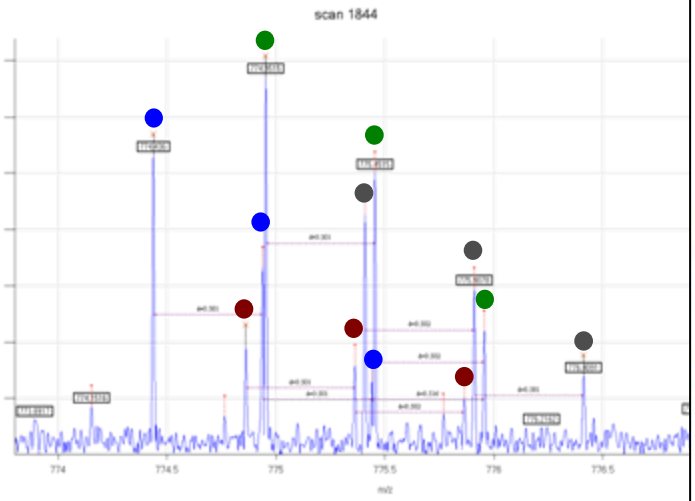
Align datasets using the MTDB Creator application

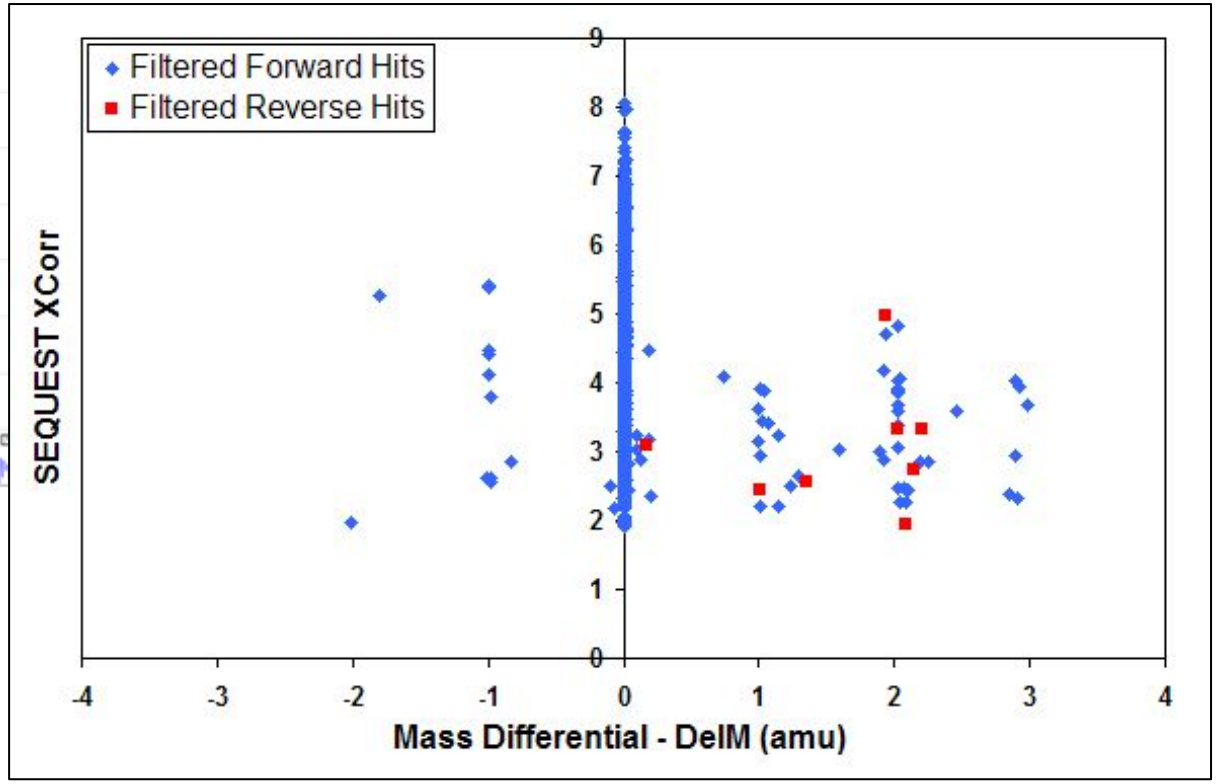
Load into database using MTDB Creator

DeconMSn



- Determines the monoisotopic mass and charge state of each parent ion chosen for fragmentation on a hybrid LC-MS/MS instrument using Decon2LS algorithms
- Replacement for the Extract_MSn.exe tool provided with SEQUEST and Bioworks

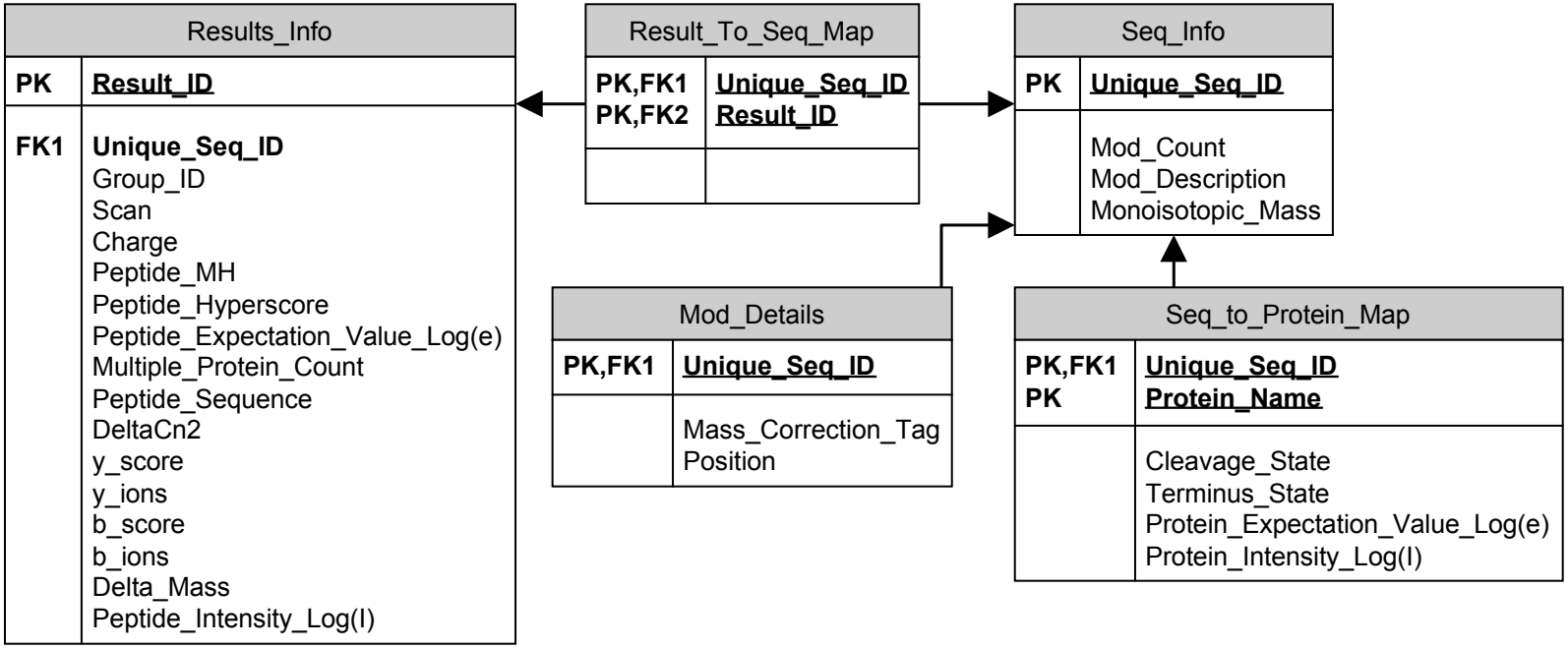




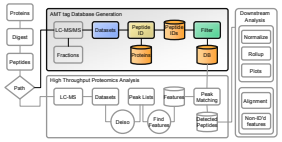
More info

Assembling an AMT tag DB

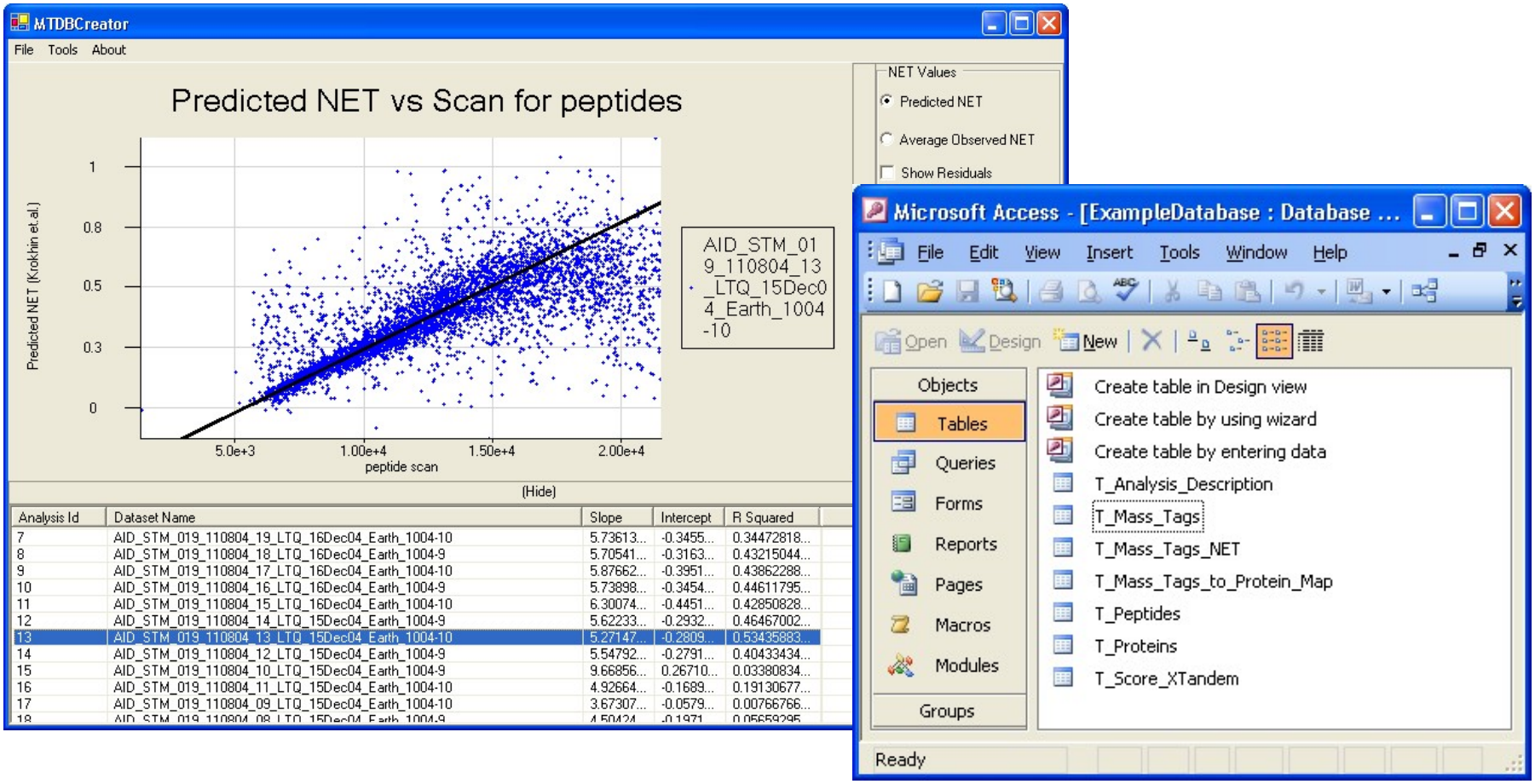
- Peptide Hit Results Processor (PHRP) relationships



MTDB Creator

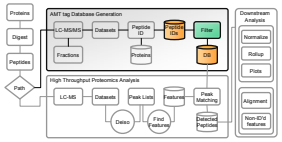


- MTDB Creator application
  - Allows external researchers to align multiple LC-MS/MS analyses and create a standalone AMT tag database

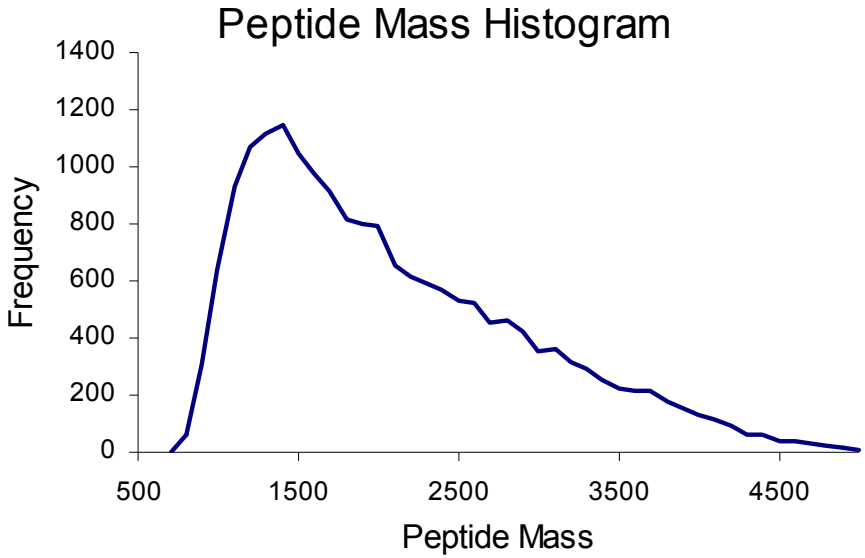


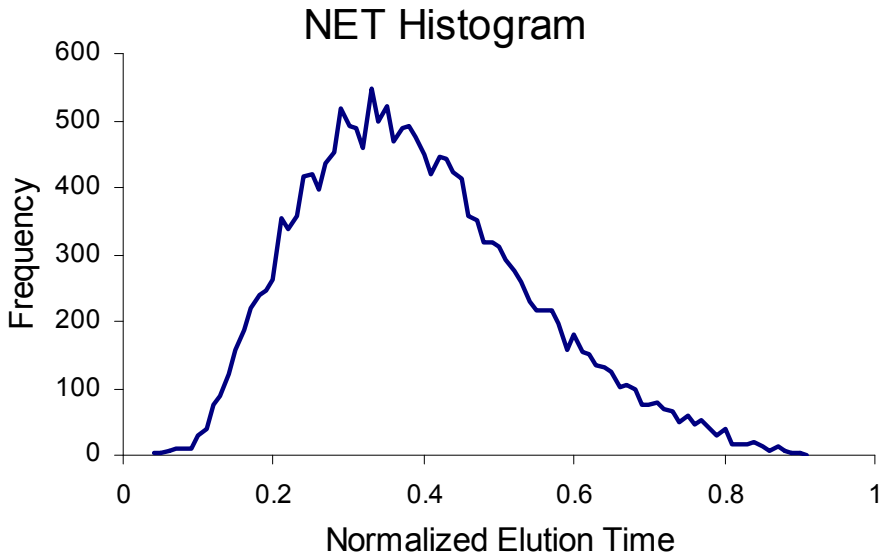
More info

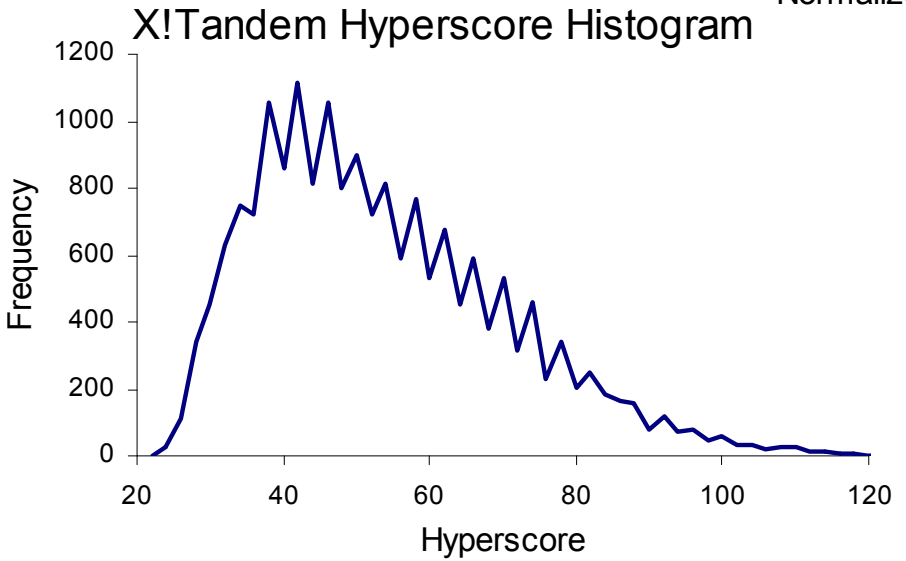
Assembling an AMT tag DB



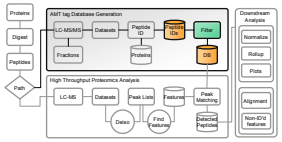
- Database histograms – filtered on Log(E_Value) ≤ -2



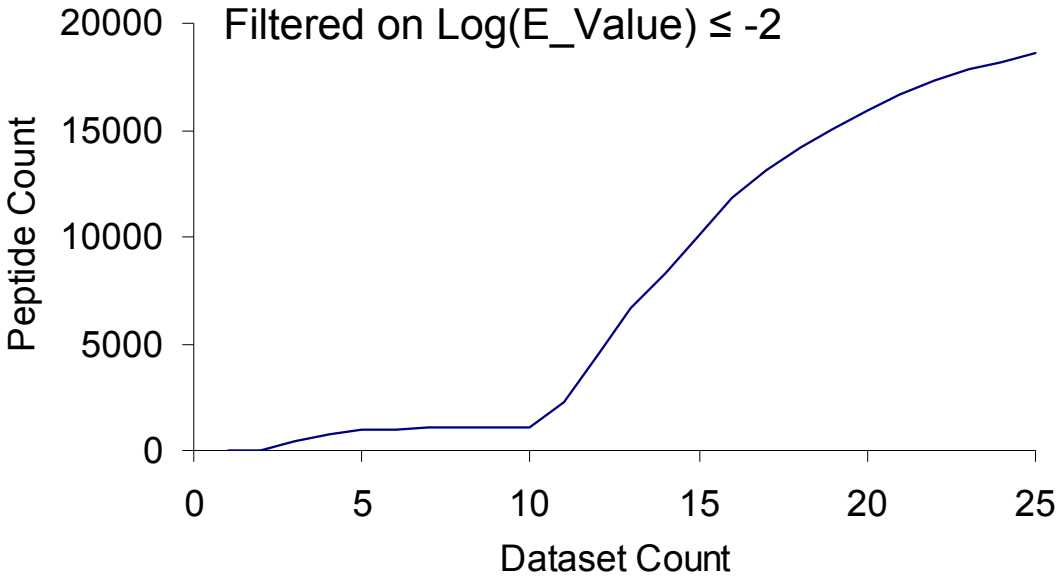




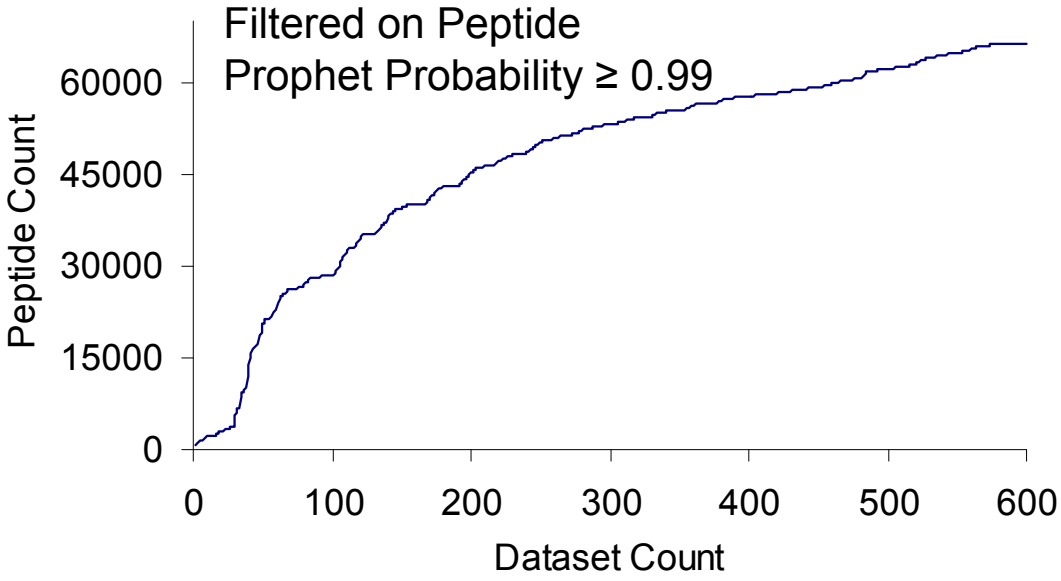
AMT tag DB Growth Trend



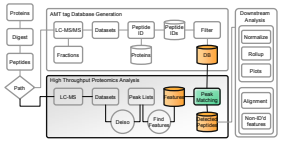
- Trend for Mini AMT tag DB
  - 25 SCX fractionation datasets of a single growth condition



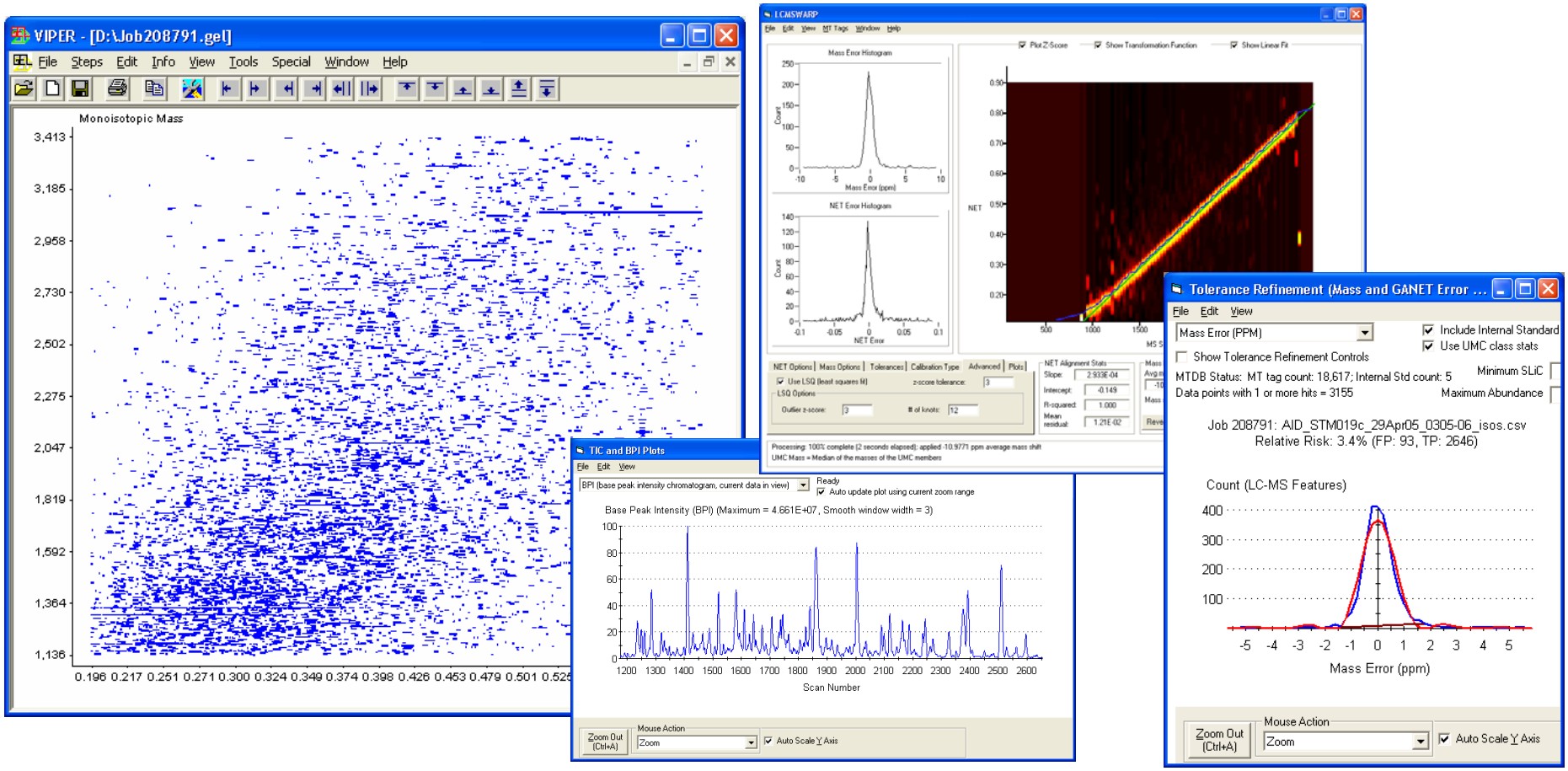
- Trend for Mature AMT tag DB
  - 521 different samples from several different growth conditions
  - Slope of curve decreases as more datasets are added and as fewer new peptides are seen



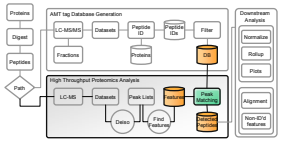
Identifying LC-MS Features



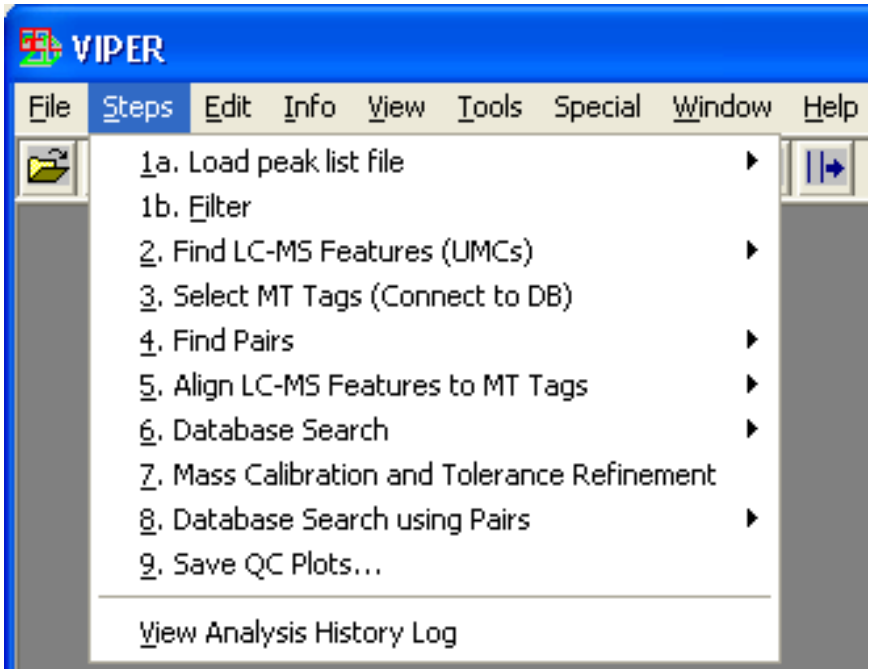
- VIPER software
  - Visualize and find features in LC-MS data
  - Match features to peptides (AMT tags)
  - Graphical User Interface and automated analysis mode



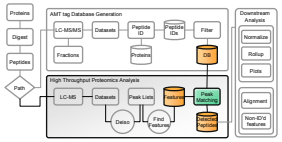
Identifying LC-MS Features



- Peak Matching Steps
  - ✓ Load LC-MS peak lists from Decon2LS
  - ✓ Filter data
  - ✓ Feature definition over elution time
  - Select AMT tags to match against
  - Optionally, find paired features (e.g. $^{16}O/^{18}O$ pairs)
  - Align LC-MS features to AMT tags using LCMSWarp
  - Broad AMT tag DB search
  - Search tolerance refinement
  - Final AMT tag DB search
  - Report results



Identifying LC-MS Features



- AMT tag database selection

Select/Modify Database Connection

MT_S_typhimurium_X347, 19273 MT tags
MT_Human_P255, 94865 MT tags
MT_BSA_P171, 6068 MT tags
MT_Human_P308, 1849 MT tags
MT_Human_P255, 20110 MT tags
MT_Shewanella_P196, 72213 MT tags
MT_C_Elegans_P237, 24103 MT tags
MT_Cyanothece_P290, 47239 MT tags

Details for the selected connection in the list at left

Database Name: MT_S_typhimurium_X347

Count of selected MT Tags in selected DB: 19273

Internal Standard Explicit:

Minimum XCorr: 1

Minimum discriminant: 0

Minimum peptide prophet: 0

Minimum PMT quality: 1

Avg Obs NET - from DB

Confirmed Only

Limit to MT tags from Dataset for Job

Sort by Most Recent

Link to Selected DB

Link to DB Not Listed Above

Break Current DB Link

Cancel

Ok

Details for the currently connected database

Database Name: MT_S_typhimurium_X347

Count of selected MT Tags in current DB: 18618

Select MT Tags

Internal Standard Explicit:

Minimum normalized XCorr: 1

Minimum discriminant score: 0

Minimum peptide prophet prob: 0

Minimum PMT quality score: 1

Avg Obs NET - from DB

Confirmed Only

Limit to MT tags from Dataset for Job

Database info for the current gel file

Override Job Info

Job number: 208791

MD_Type: 1

Save Job Info Changes

Source Filename: AID_STM019c_29Apr05_0305-06_isos.csv

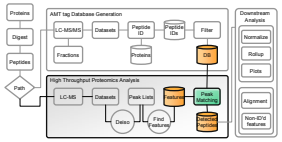
Path to Legacy DB (Access DB with MT Tags)

D:\Typhimurium_AMTs.mdb

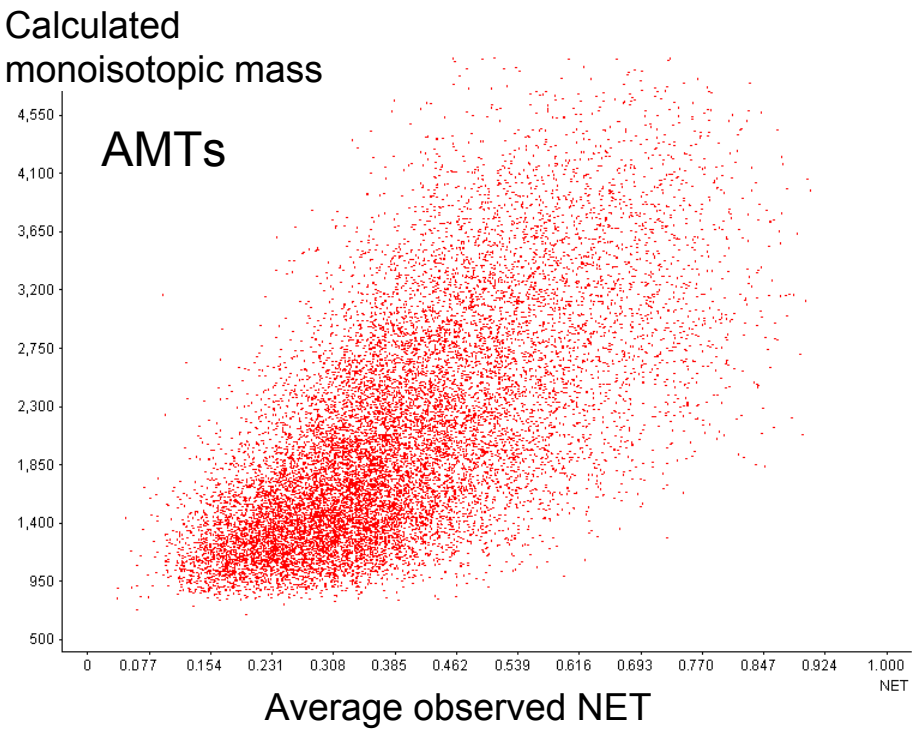
Browse

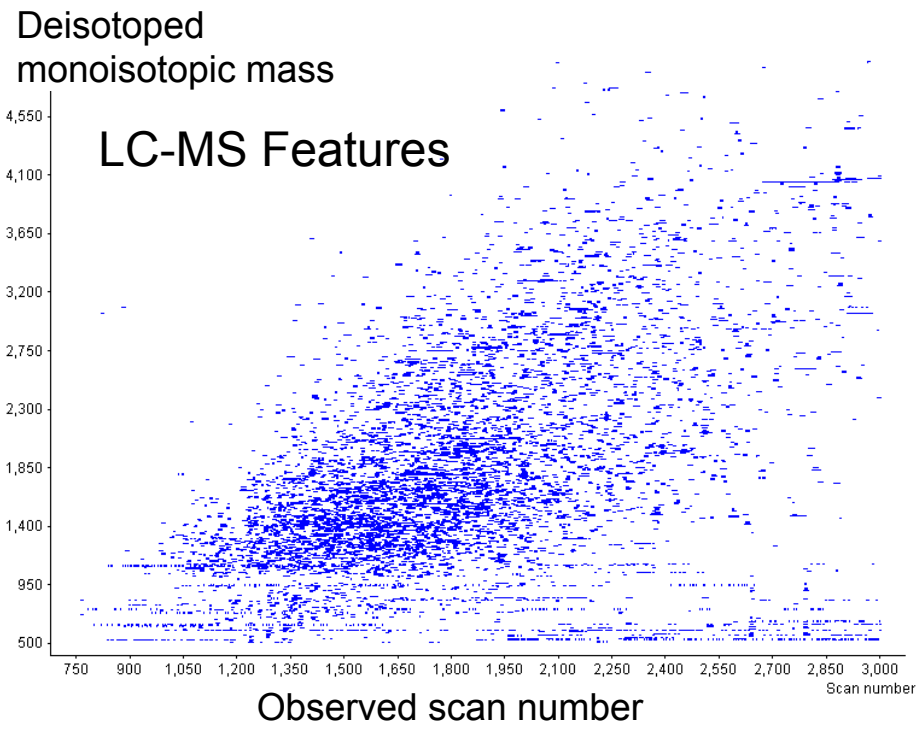
Connect to mass tag system (MTS) if inside PNNL or use standalone Microsoft Access DB

Alignment using LCMSWarp

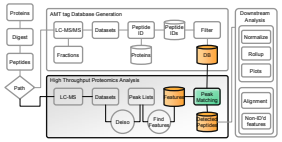


- Align scan number (i.e. elution time) of features to NETs of peptides in given AMT tag database
  - Match mass and NET of AMT tags to mass and scan number of MS features
  - Use LCMSWarp algorithm to find optimal alignment to give the most matches

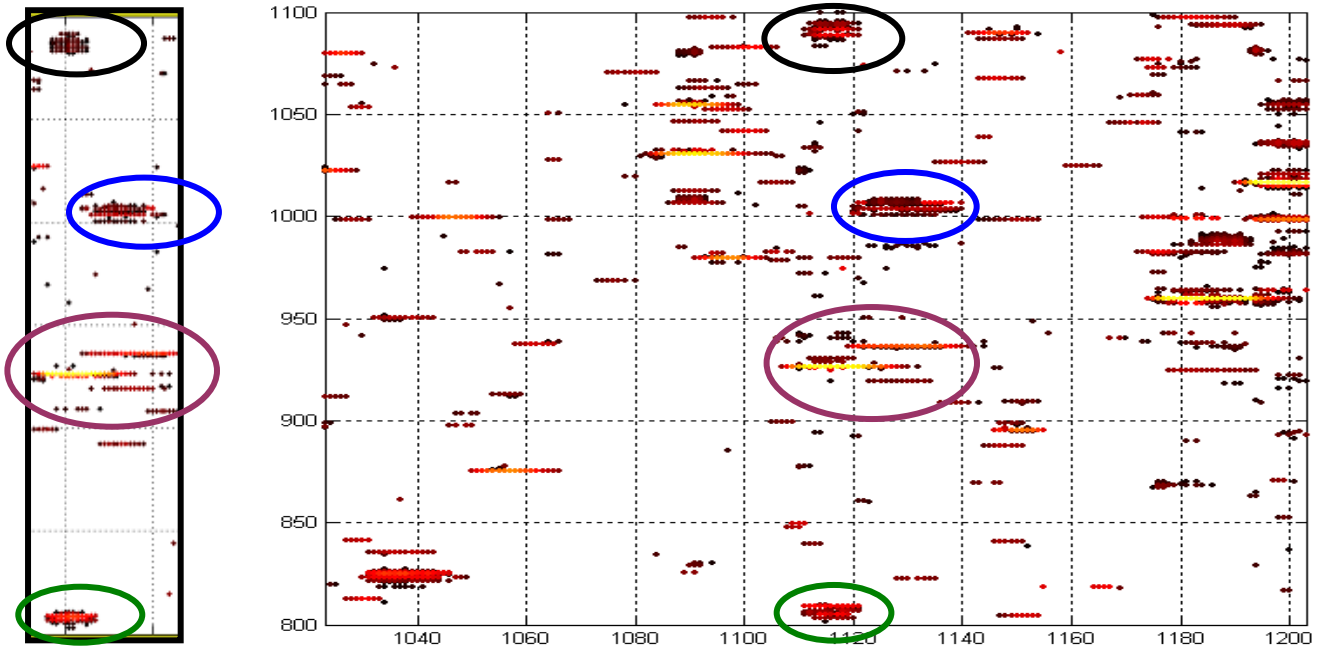


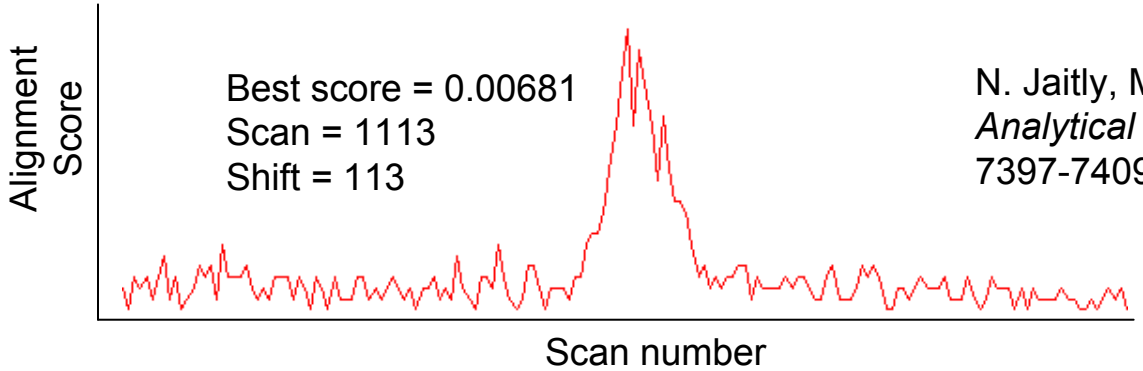


Alignment using LCMSWarp



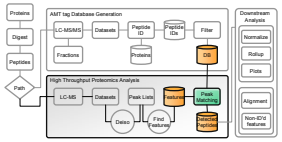
- LCMSWarp computes a similarity score from conserved local mass and retention time patterns



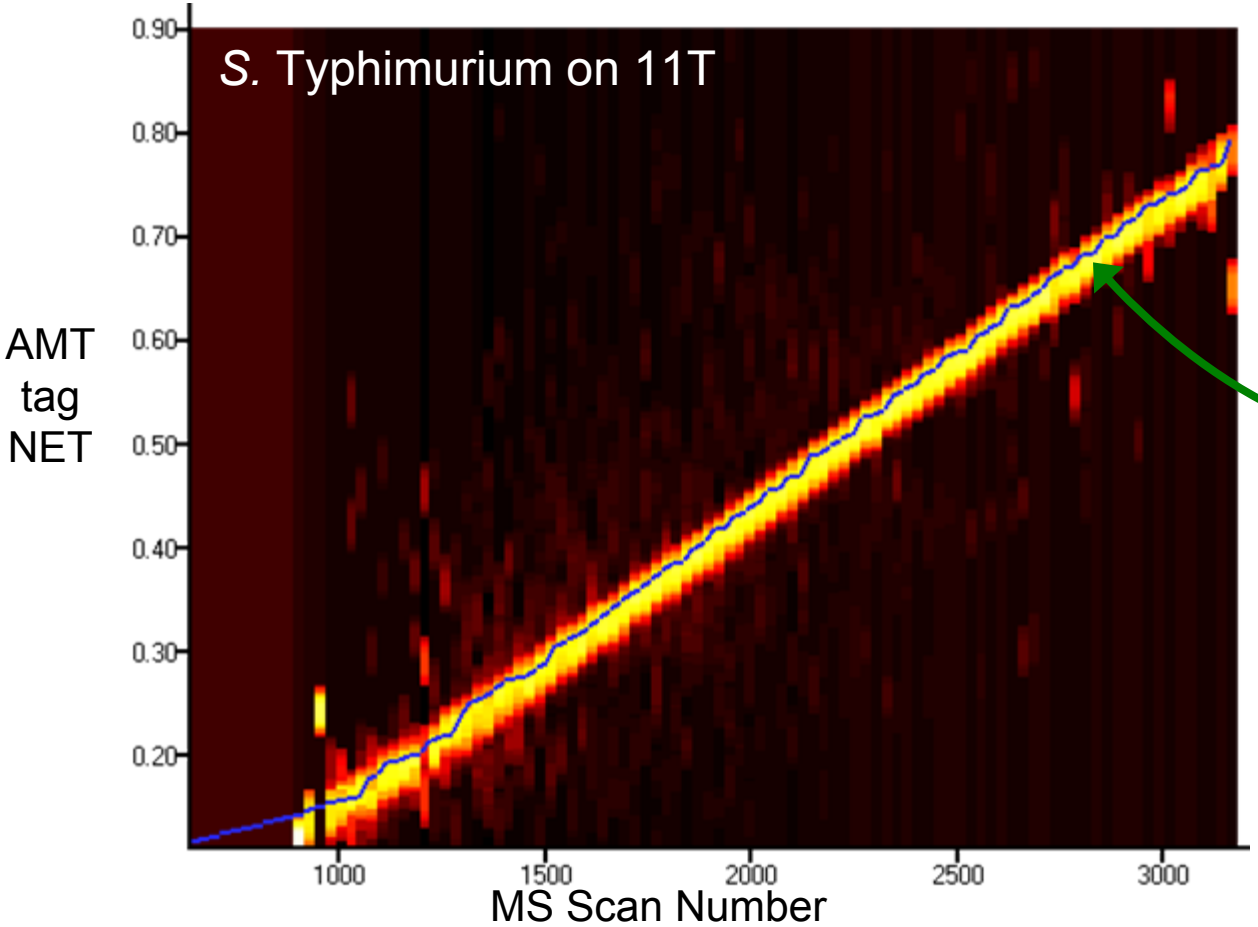


N. Jaitly, M.E. Monroe et. al., *Analytical Chemistry* **2006**, *78*, 7397-7409.

Alignment using LCMSWarp



- Similarity scores between LC-MS features and AMT tags are used to generate a score graph of similarity
- Best alignment is found using a dynamic programming algorithm that determines the transformation function with maximum likelihood

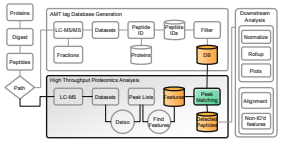


Heatmap of similarity score between LC-MS features and AMT tags (z-score representation)

Alignment Function

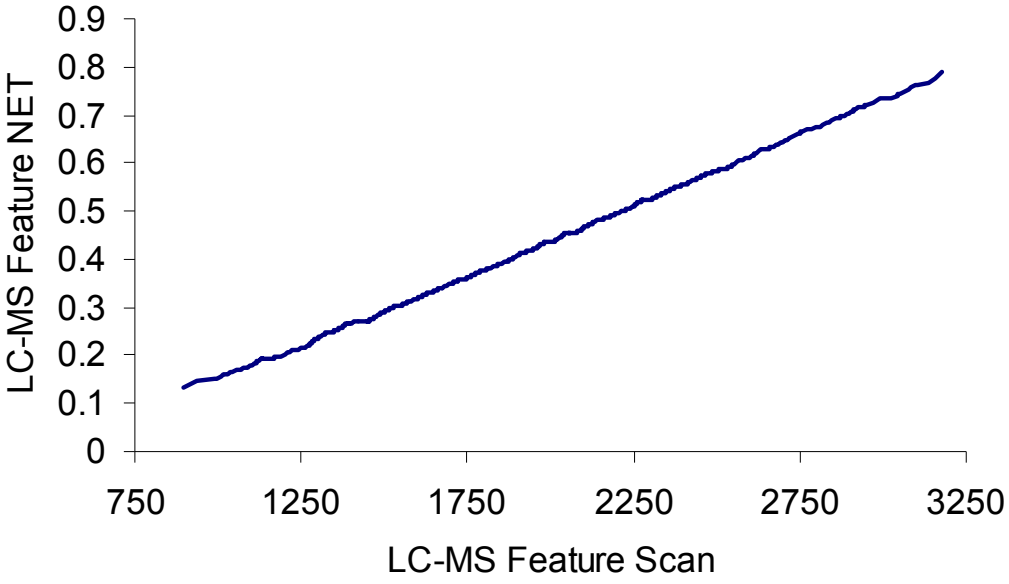
N. Jaitly, M.E. Monroe et. al., *Analytical Chemistry* **2006**, *78*, 7397-7409.

Alignment using LCMSWarp

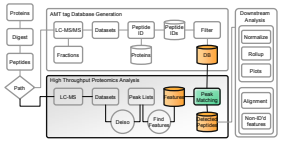


- Transformation function is used to convert from scan number to NET
  - Features centered at same scan number get the same obs. NET value
  - When matching LC-MS features to AMTs, we will search +/- a NET tolerance, which effectively allows for LC-MS features to shift around a little in elution time

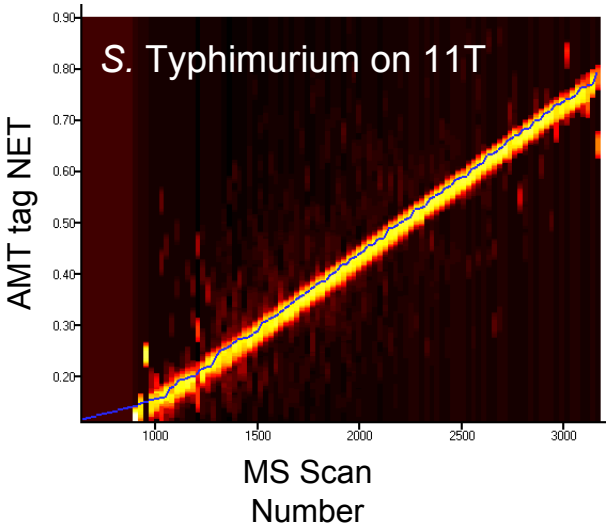
LC-MS Feature Scan	Matching AMT tag NET	LC-MS Feature NET
1011	0.1519	0.1569
1019	0.1626	0.1589
1019	0.1507	0.1589
1021	0.1653	0.1594
1027	0.1509	0.1609
1037	0.1519	0.1633
1042	0.183	0.1645
1055	0.1652	0.1677
1056	0.1862	0.1679
1056	0.1697	0.1679
1056	0.1682	0.1679



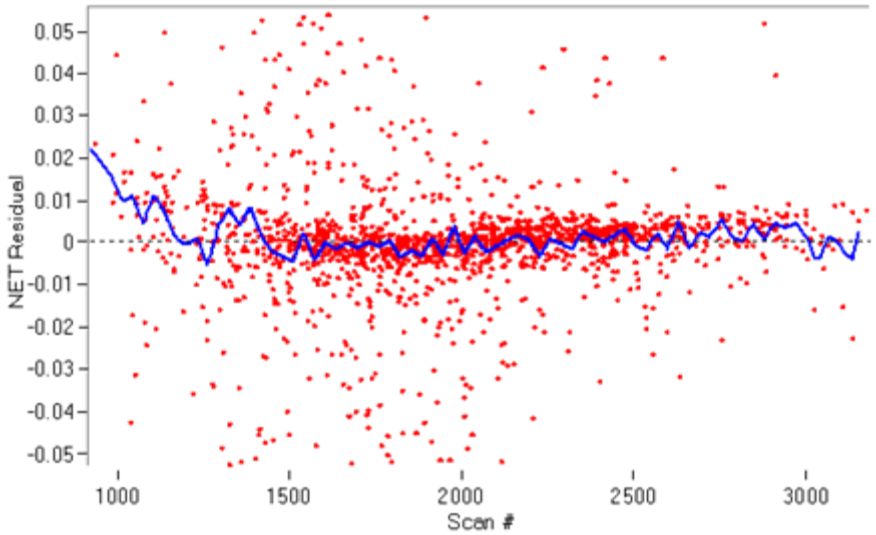
Alignment using LCMSWarp



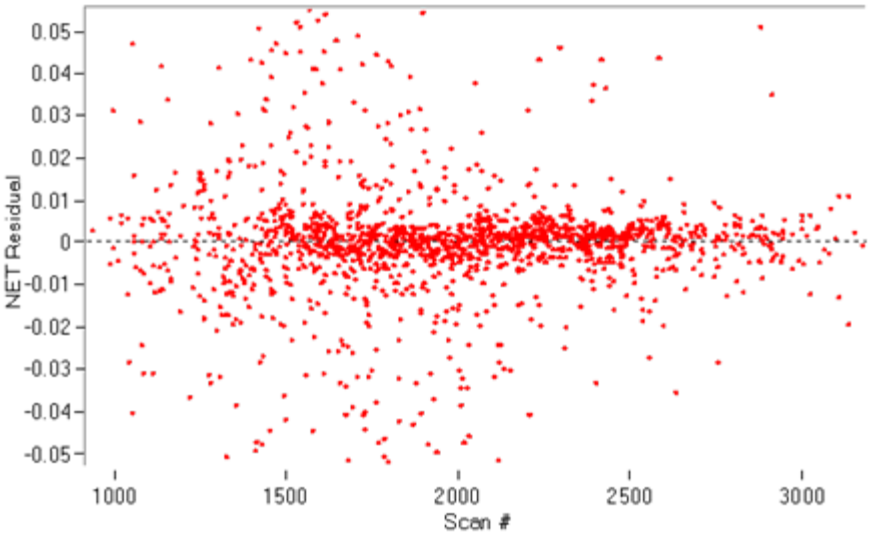
- NET Residual Plots
  - Difference between NET of LC-MS feature and NET of matching AMT tag
    - Indicates quality of alignment between features and AMT tags
  - This data shows nearly linear alignment between features and AMTs, but the algorithm can easily account for non-linear trends



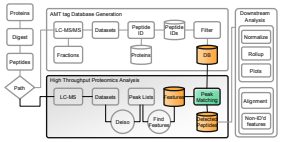
NET Residuals if a linear mapping is used



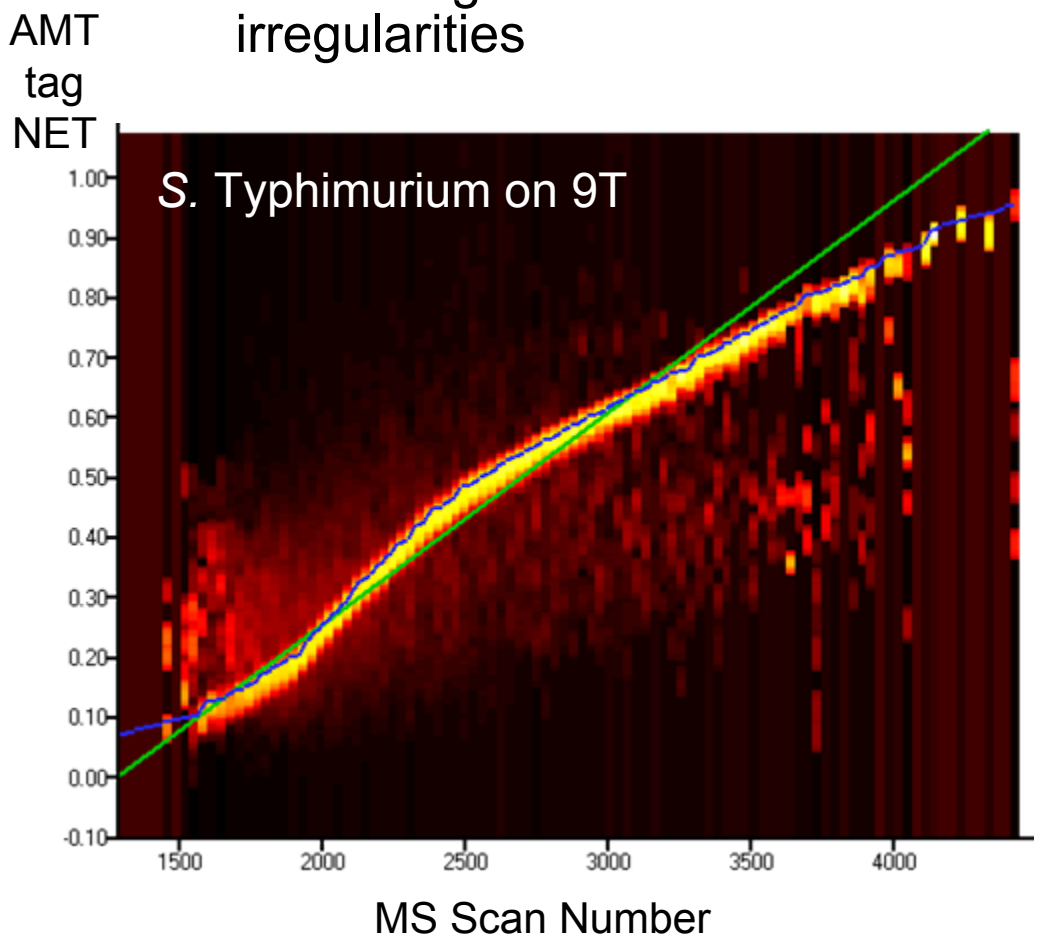
NET Residuals after LCMSWarp



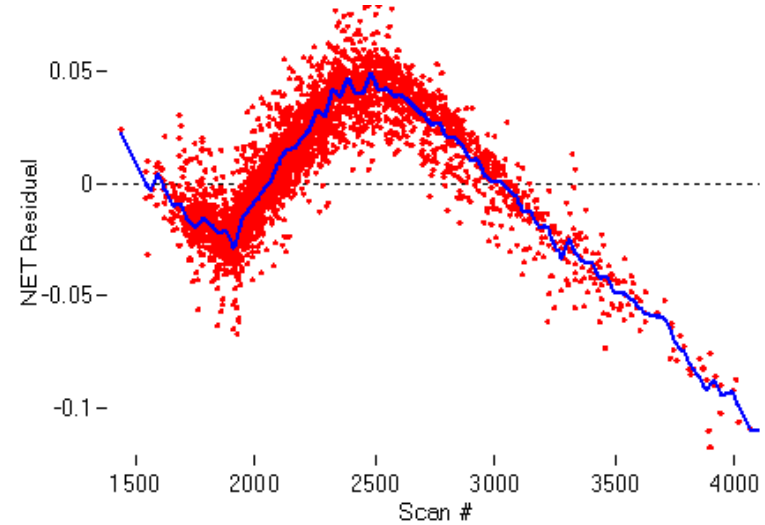
Alignment using LCMSWarp



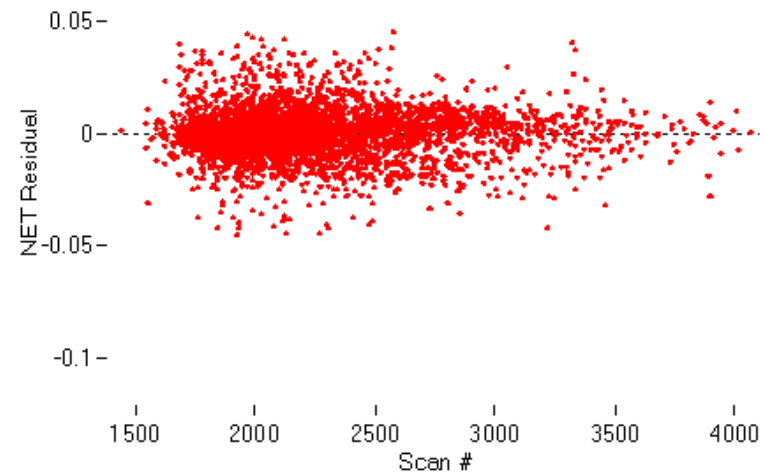
- Non-linear alignment example #1
  - Identical LC separation system, but having column flow irregularities



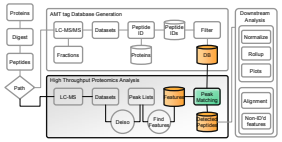
NET Residuals if a linear mapping is used



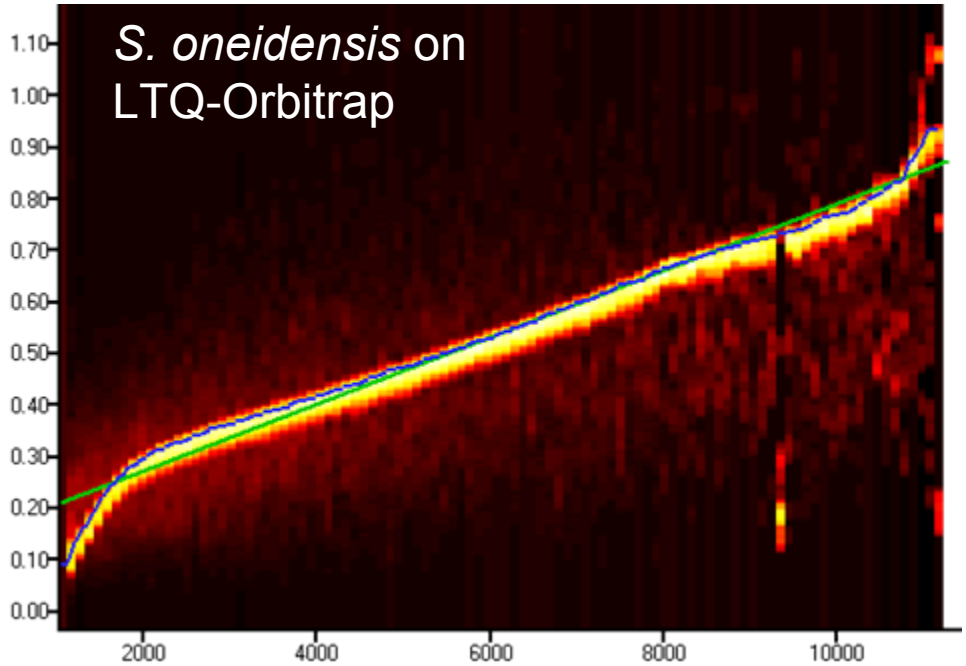
NET Residuals after LCMSWarp



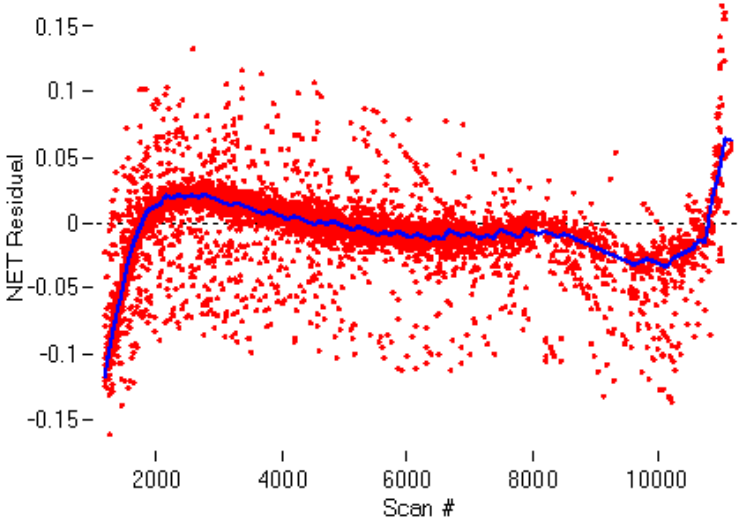
Alignment using LCMSWarp



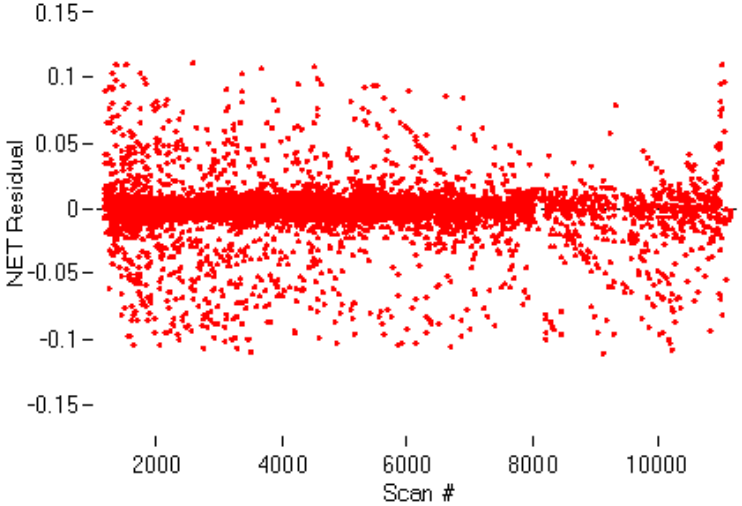
- Non-linear alignment example #2
  - AMT tag DB from C_{18} LC-MS/MS analyses using ISCO-based LC (**exponential dilution gradient**)
  - LC-MS analysis used C_{18} LC-MS via Agilent **linear gradient** pump



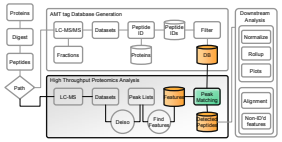
NET Residuals if a linear mapping is used



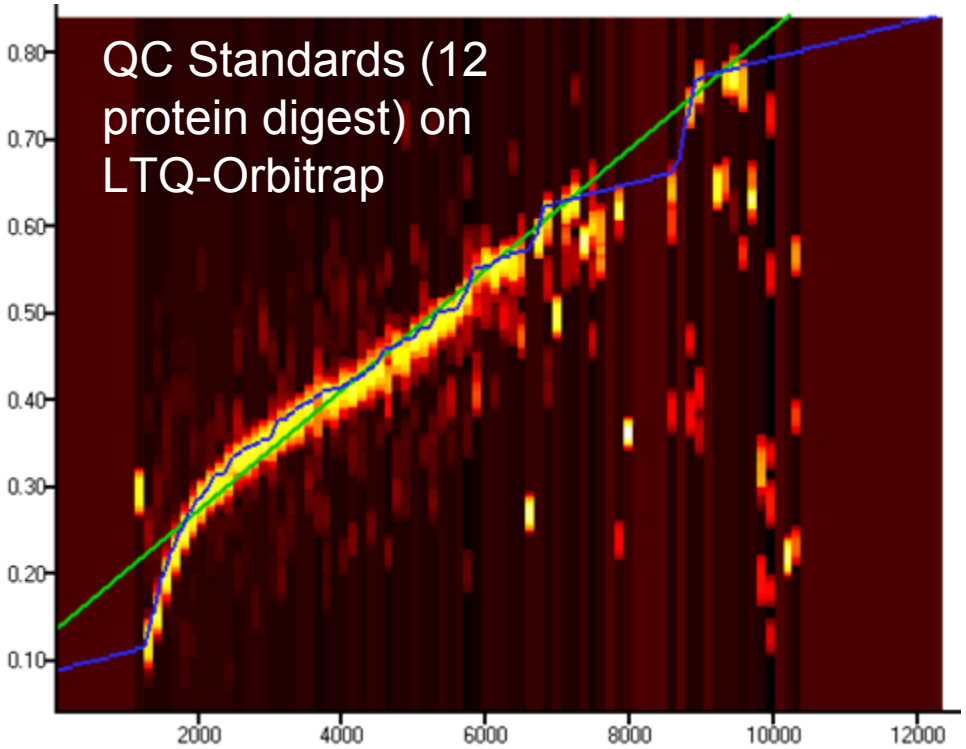
NET Residuals after LCMSWarp



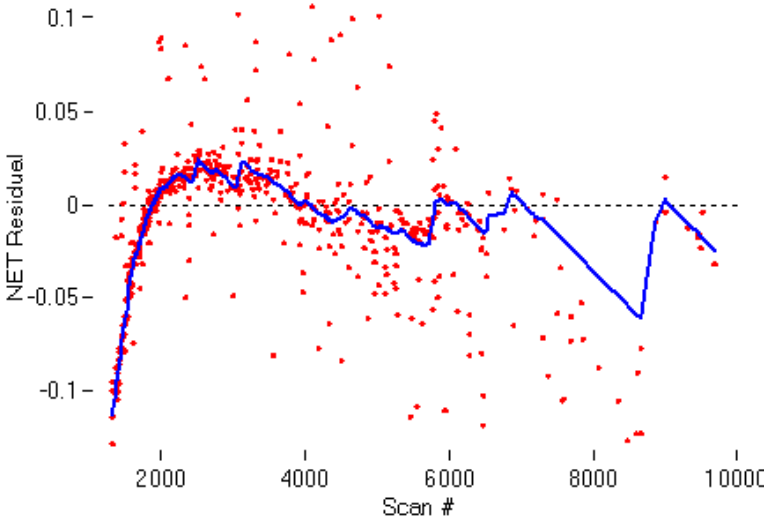
Alignment using LCMSWarp



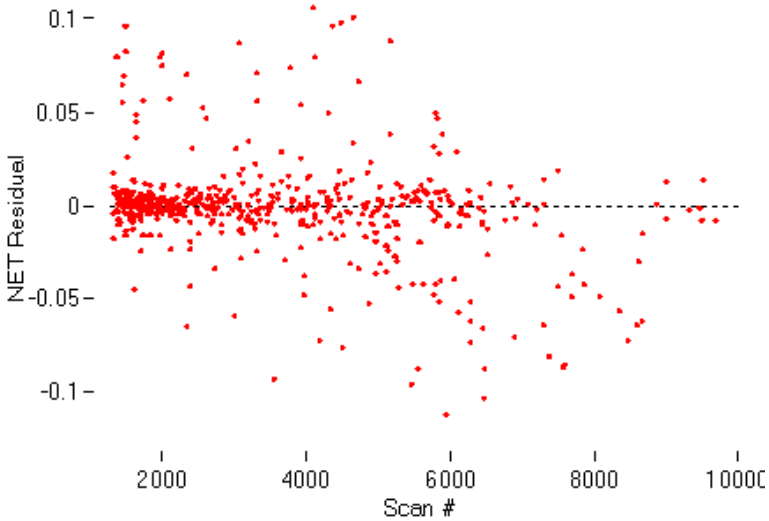
- Non-linear alignment example #3
  - AMT tag DB from C_{18} LC-MS/MS analyses using ISCO-based LC
  - LC-MS analysis used C_{18} LC-MS via Agilent linear gradient pump



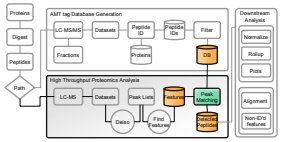
NET Residuals if a linear mapping is used



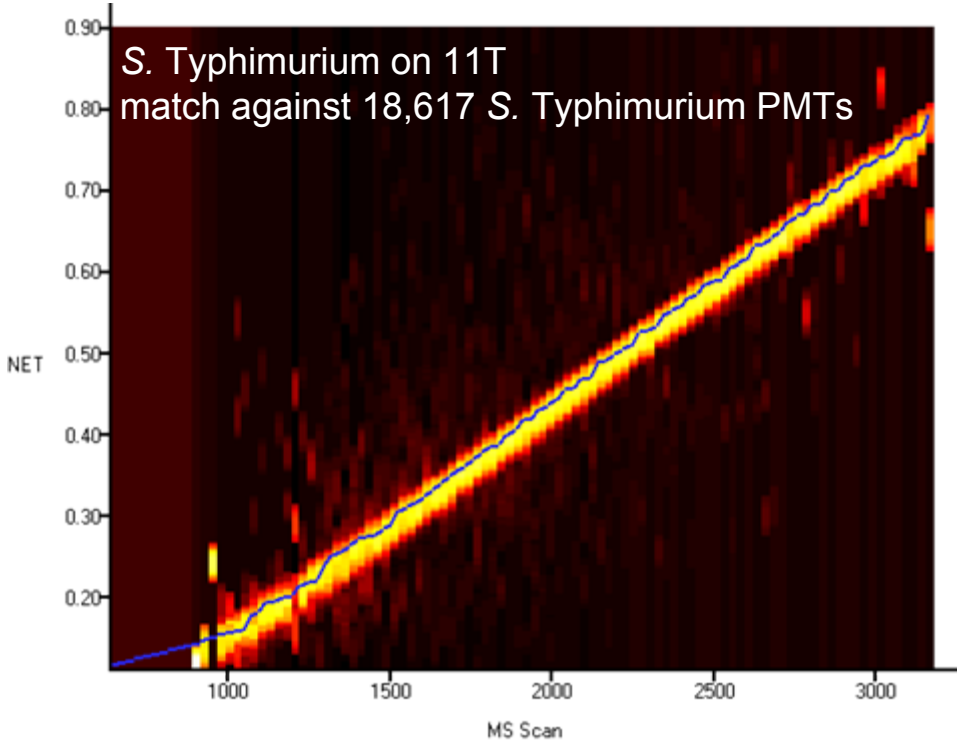
NET Residuals after LCMSWarp

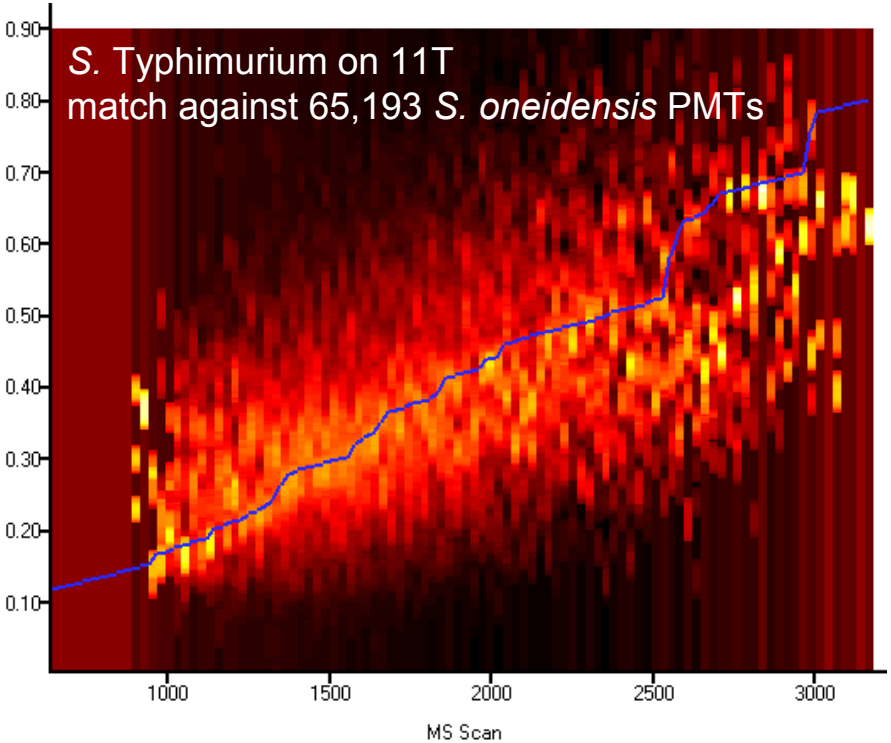


Alignment using LCMSWarp

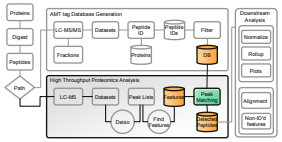


- LCMSWarp Features
  - Fast and robust
    - Previous method used least-squares regression, iterating through a large range of guesses (slow and often gave poor alignment)
  - Requires that a reasonable number of LC-MS features match the AMT tag DB



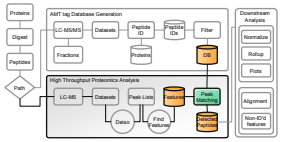


Alignment using LCMSWarp



- In addition to aligning data in time, we can also recalibrate the masses of the LC-MS features
  - Possible because mass and time values are available for both LC-MS features and AMT tags
- Two options for mass re-calibration
  - Bulk linear correction
  - Piece-wise correction via LCMSWarp
- Visualize mass differences using mass error histogram or mass residual plot

Mass Error Histogram

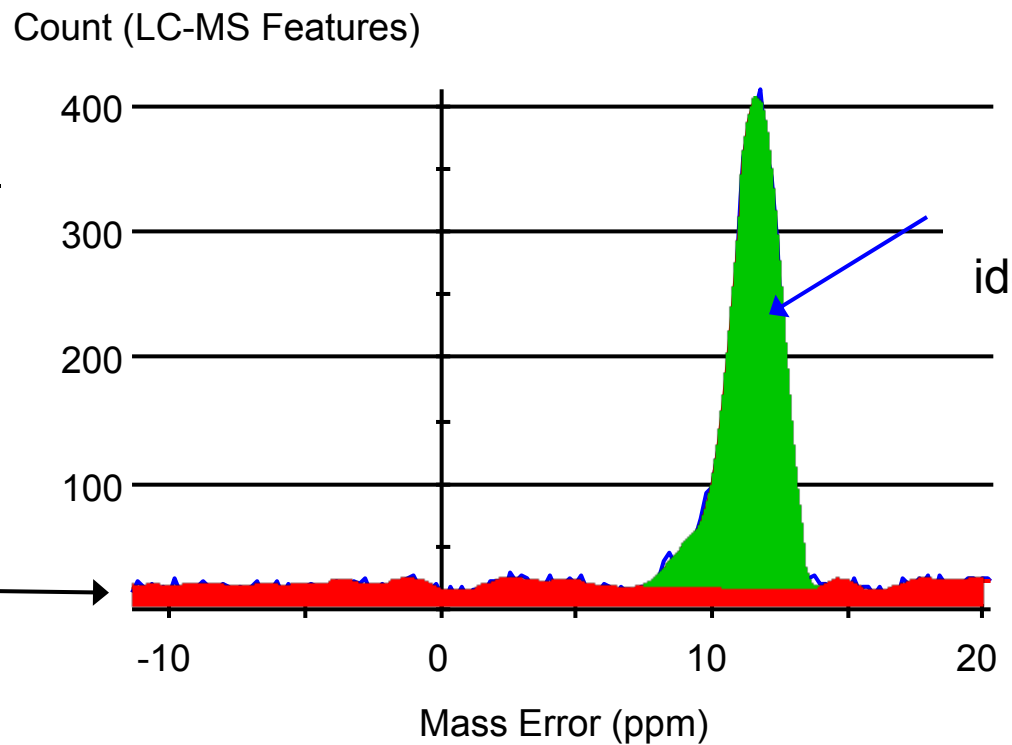


- List of binned mass error values
  - Difference between feature's mass and matching AMT tag's mass
  - Bin values to generate a histogram
  - Typically observe background false positive level

LC-MS Feature Mass (Da)	AMT tag Mass (Da)	Delta Mass (Da)	Mass Error (ppm)
1570.9005	1570.883	0.01745	11.1
1571.74325	1571.726	0.01770	11.3
1571.8498	1571.831	0.01912	12.2
1571.9107	1571.892	0.01848	11.8
1573.8381	1573.832	0.00569	3.6

<u>Match Tolerances</u>

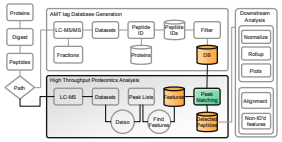
Mass: ±25 ppm
NET: ±0.05 NET



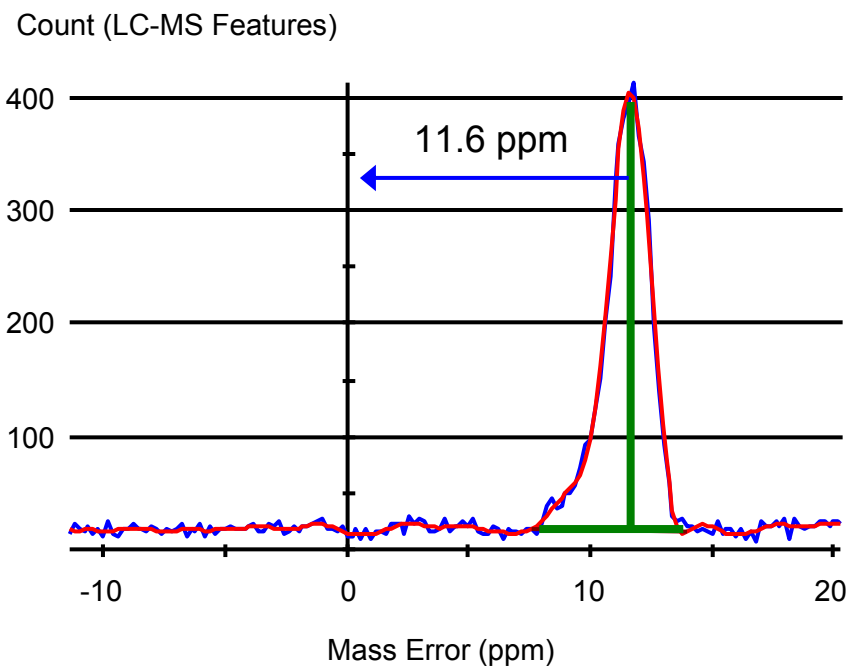
Likely true positive identifications

Likely false positive identifications

Mass Calibration



- Option 1: Bulk linear correction
  - Use location of peak in mass error histogram to adjust masses of all features
  - Shift by ppm mass; absolute shift amount increases as monoisotopic mass increases



Peak Center of mass: 11.6 ppm
Peak Width: 2 ppm at 60% of max
Peak Height: 404 counts/bin
Noise level: 19 counts/bin

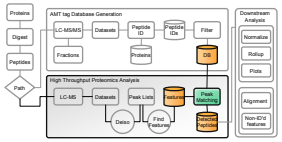
Shift all masses -11.6 ppm:

$$\Delta_{mass} = -11.6\text{ppm} \times \frac{mass_{old}}{1\text{x}10^6\ \text{ppm/Da}}$$

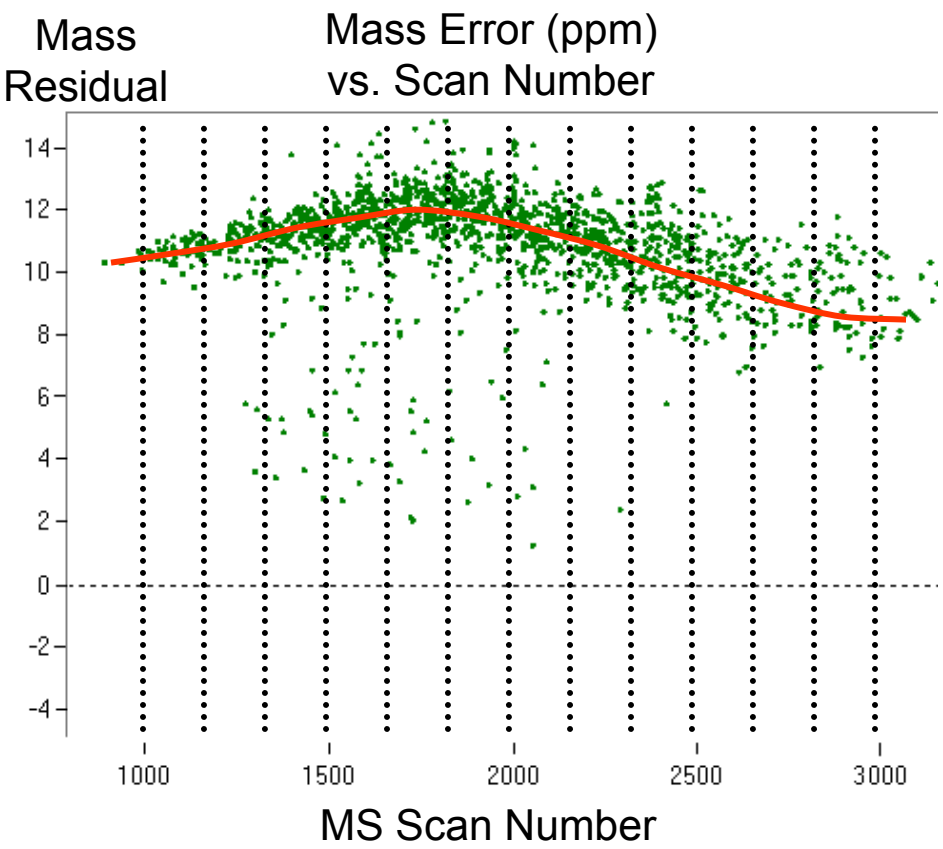
For 1+ feature at 1570.9005 Da,
Δ_{mass} = -0.0182 Da

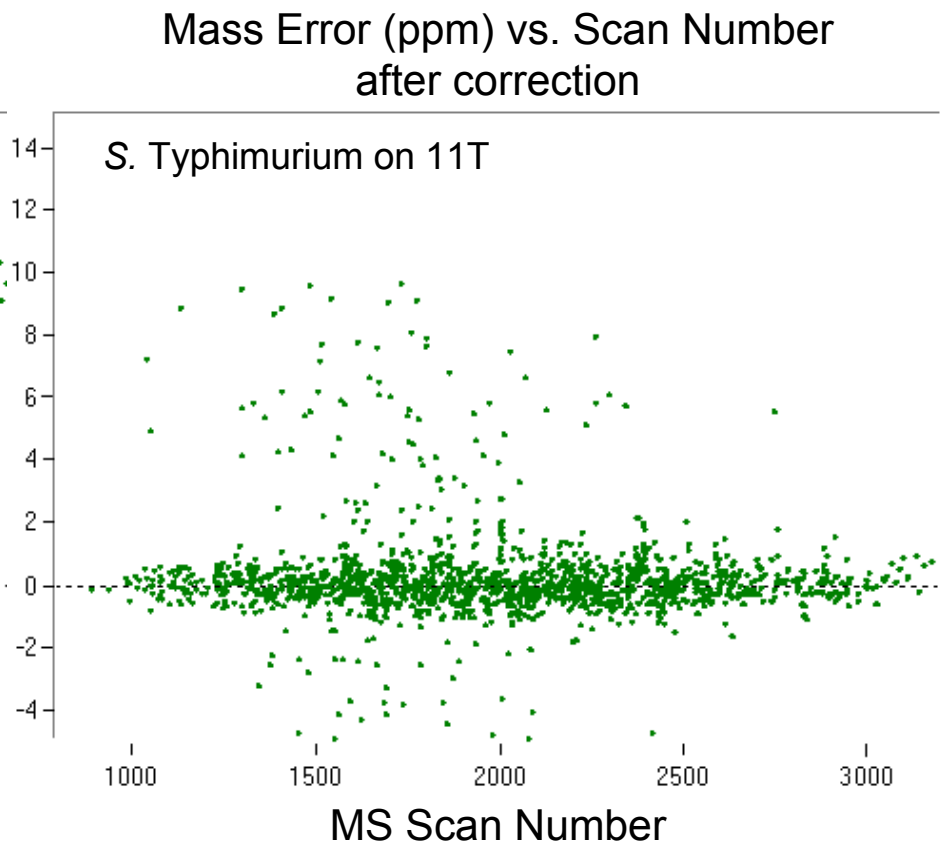
For 3+ feature at 2919.4658 Da,
Δ_{mass} = -0.0339 Da

Mass Calibration

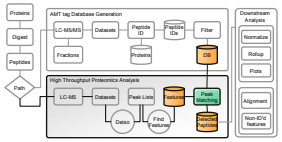


- Option 2: Piece-wise correction via LCMSWarp
  - Use smoothing splines to determine a smooth calibration curve which is a function of scan number

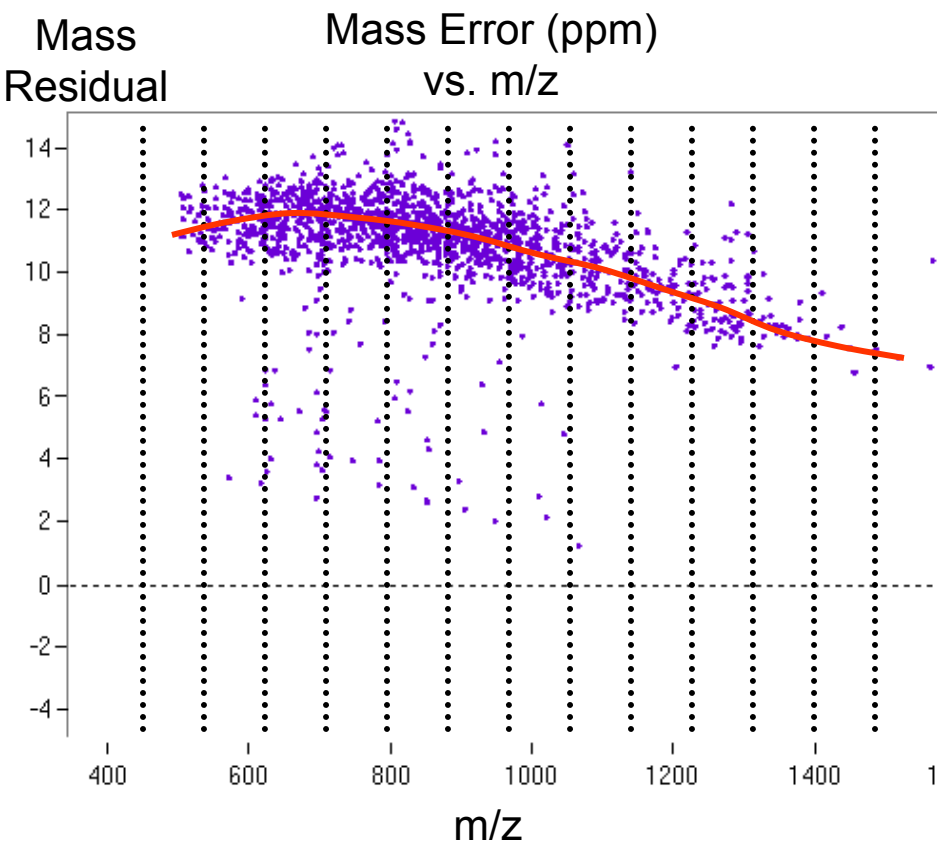


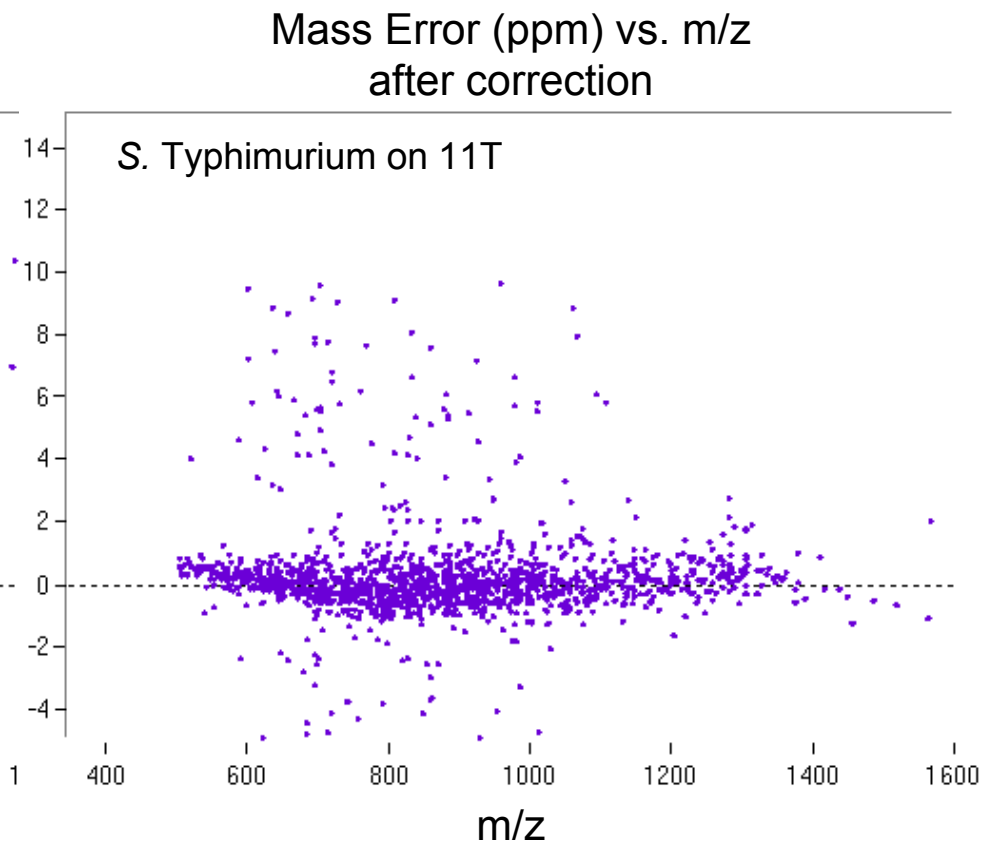


Mass Calibration

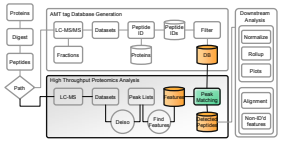


- Option 2: Piece-wise correction via LCMSWarp
  - Use a smoothing spline calibration which is a function of m/z
  - LCMSWarp utilizes a hybrid correction based on both mass error vs. time and mass error vs. m/z

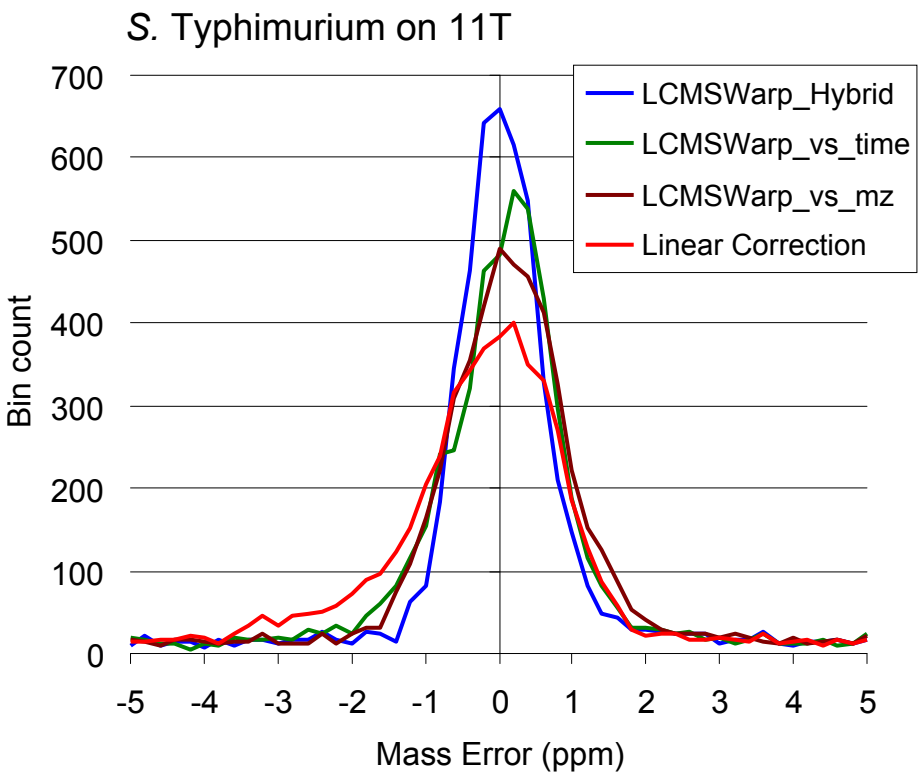


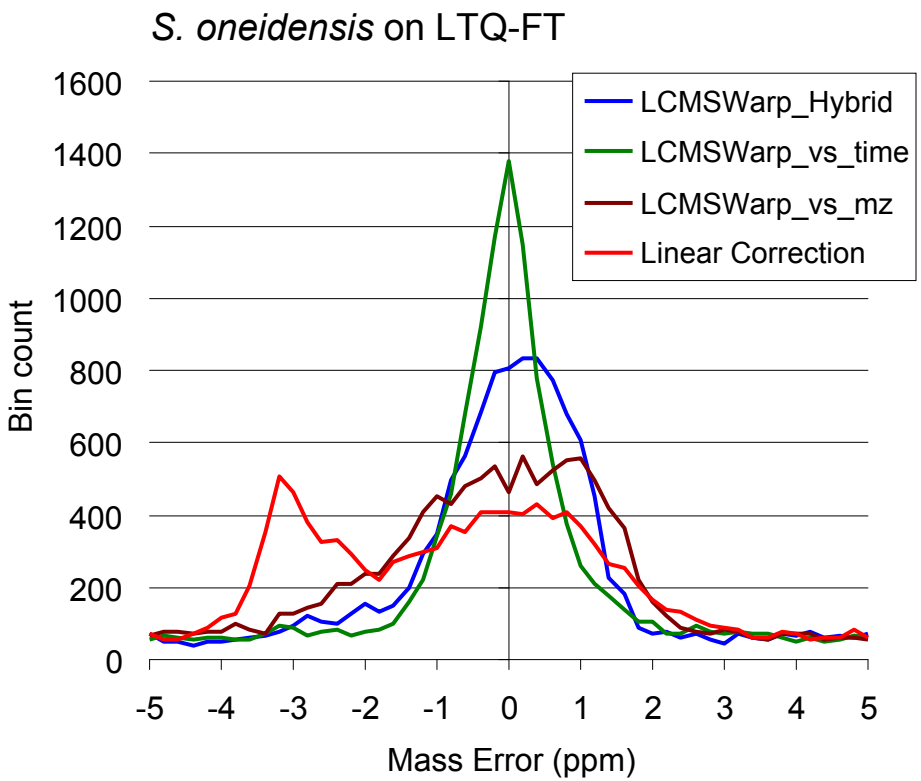


Mass Calibration

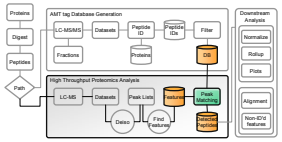


- Comparison of the three methods
  - Mass error histogram gets taller, narrower, and more symmetric
    - Linear → Mass error vs. m/z → Mass error vs. time → Hybrid
  - Not all datasets show the same trends, but Hybrid mass recalibration is generally superior



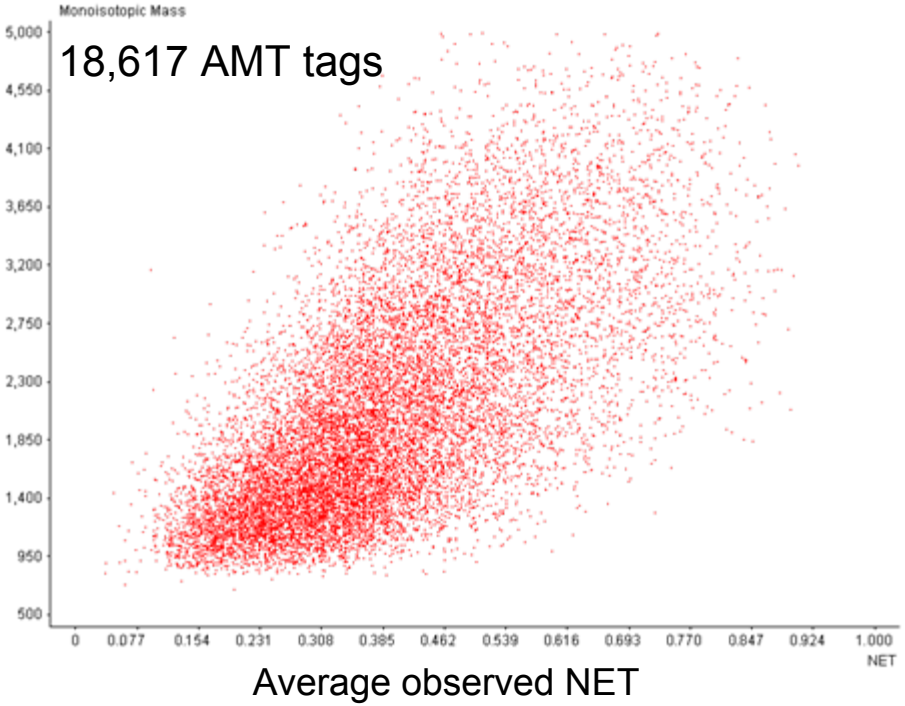


Identifying LC-MS Features

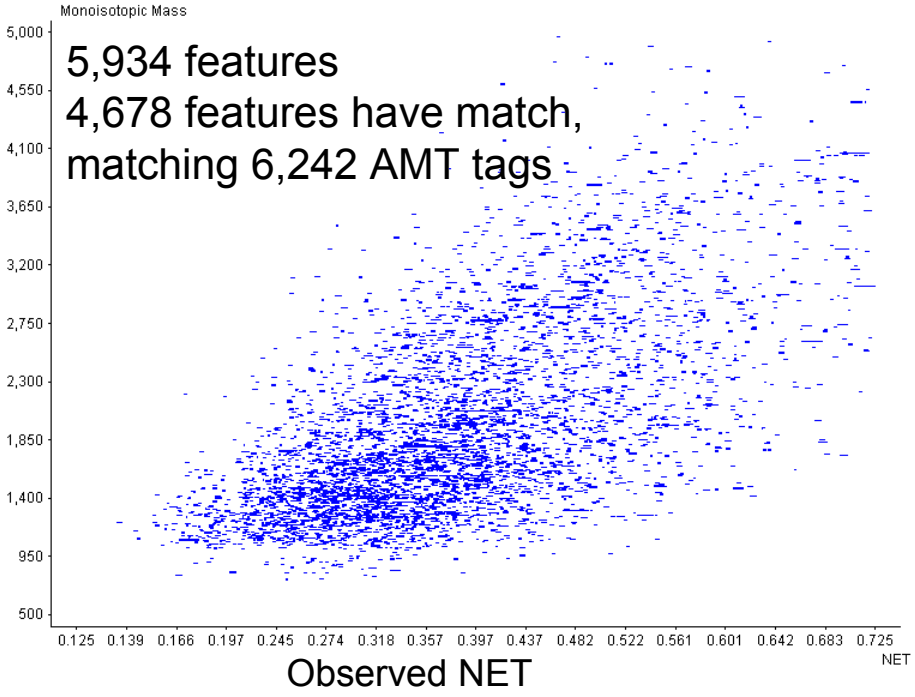


- Match features to LC-MS/MS IDs
- *S.* Typhimurium DB, from 25 LC-MS/MS analyses
  - 18,617 AMT tags, all fully or partially tryptic
  - Look for AMT tags within a broad mass range, e.g., ±25 ppm and ±0.05 NET of each feature

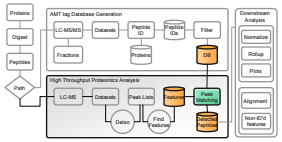
S. Typhimurium AMT tag Database



S. Typhimurium on 11T FTICR



Search Tolerance Refinement



- Can use mass error and NET error histograms to determine optimal search tolerances

Tolerance Refinement (Mass and GANET Error Plots)

File Edit View

Mass Calibration Refinement | Tolerance Refinement

Mass Tolerance Refinement
Minimum Tol. 0.75 ppm
Maximum Tol. 15 ppm
Adjustment multiplier 1

NET Tolerance Refinement
Minimum Tol. 0.0075 NET
Maximum Tol. 0.2 NET
Adjustment multiplier 1

Expectation Maximization
Use min or max tol if out of range

Expectation Maximization Options
Mass error tol. estimate 6 ppm
NETerror tol. estimate 0.05
Data to exclude from extremes 10 %
Use single data point errors for Mass
Use single data point errors for NET

Start Mass Tol Refinement
Start NET Tol Refinement

Criteria To Use Peak For Refinement
Minimum Height 10 counts/bin
Pct of Max for Finding Width 60
Minimum S/N for Low Abu 2.5

UMC Mass Statistics
Count: 5934 UMC's
Median Width: 1.9 ppm
Maximum Width: 17.751 ppm
Median StDev: 0.614 ppm
Maximum StDev: 7.372 ppm

Relative Risk Statistics
Mass: 7.7% (FP: 356, TP: 4245)
NET: 25% (FP: 1615, TP: 4844)

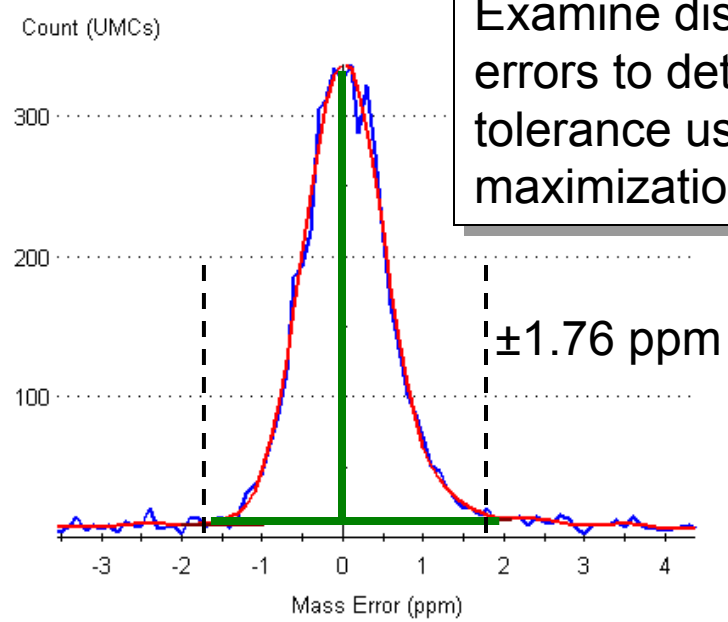
Current DB Search Tolerances
DB Mass Tolerance: 25 ppm
DB NET Tolerance: 0.05 NET

Mass Calibration Plot Status
Peak Center: 0 ppm
Peak Width: 1.1 ppm
Peak Height: 338 counts/bin
Noise: 8 counts
S/N: 43

NET Calibration Plot Status
Peak Center: 0.001 NET
Peak Width: 0.01 NET
Peak Height: 377 counts/bin
Noise: 37 counts
S/N: 10

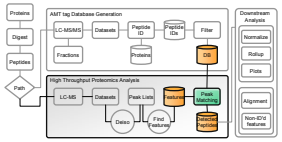
Mass Error (PPM)
Include Internal Standard matches
Use UMC class stats
Show Tolerance Refinement Controls
MTDB Status: MT tag count: 18,617; Internal Std count: 5
Minimum SLiC 0
Data points with 1 or more hits = 4763
Maximum Abundance 0

Job 208791: AID_STM019c_29Apr05_0305-06_isos.csv
Relative Risk: 7.7% (FP: 356, TP: 4245)

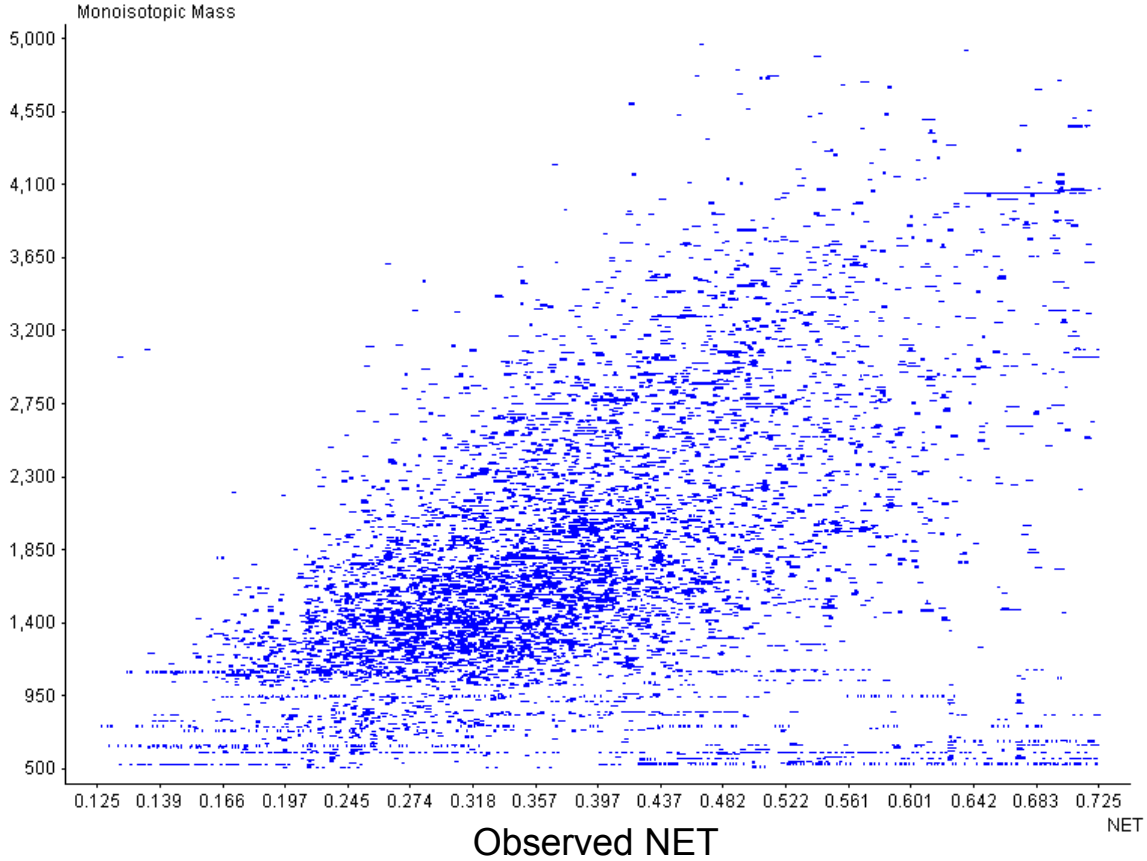


Examine distribution of errors to determine optimal tolerance using expectation maximization algorithm

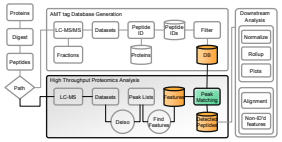
Identifying LC-MS Features



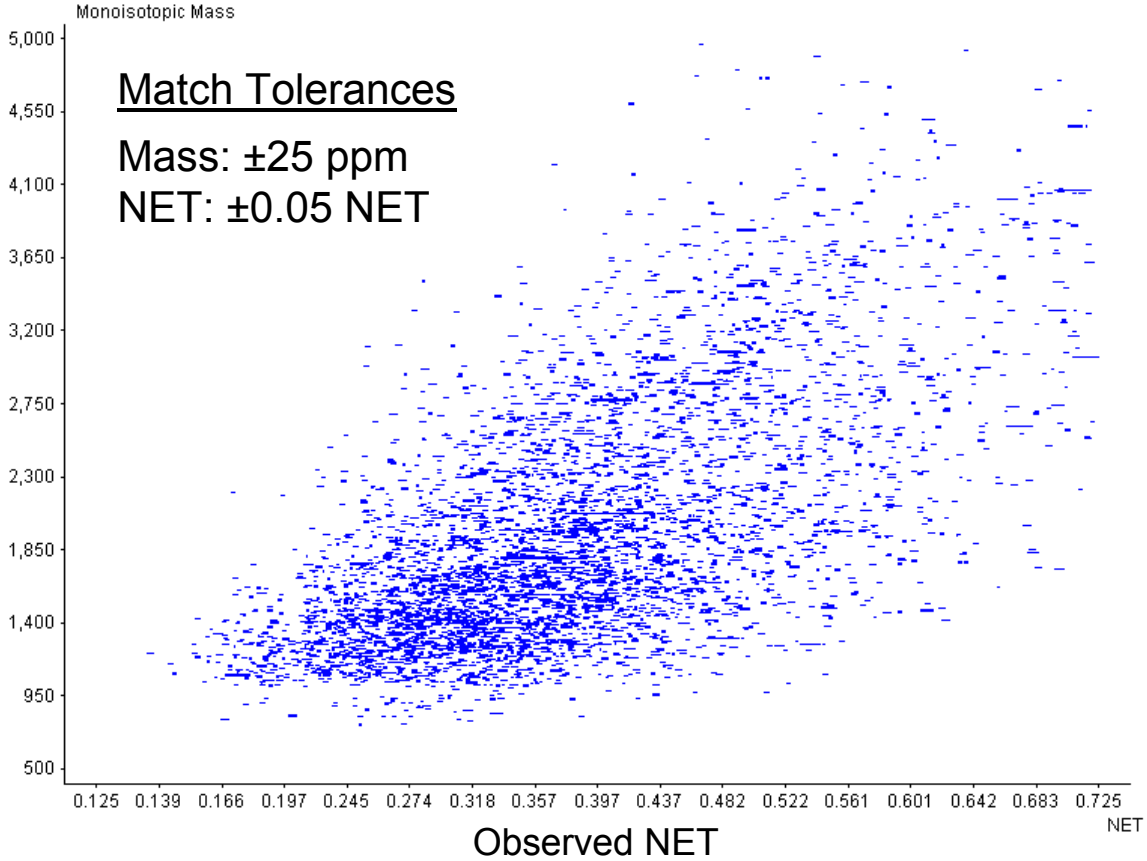
- Repeat search with final search tolerances
  - 5,934 features



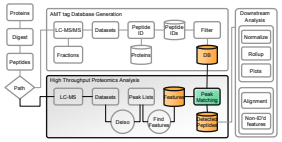
Identifying LC-MS Features



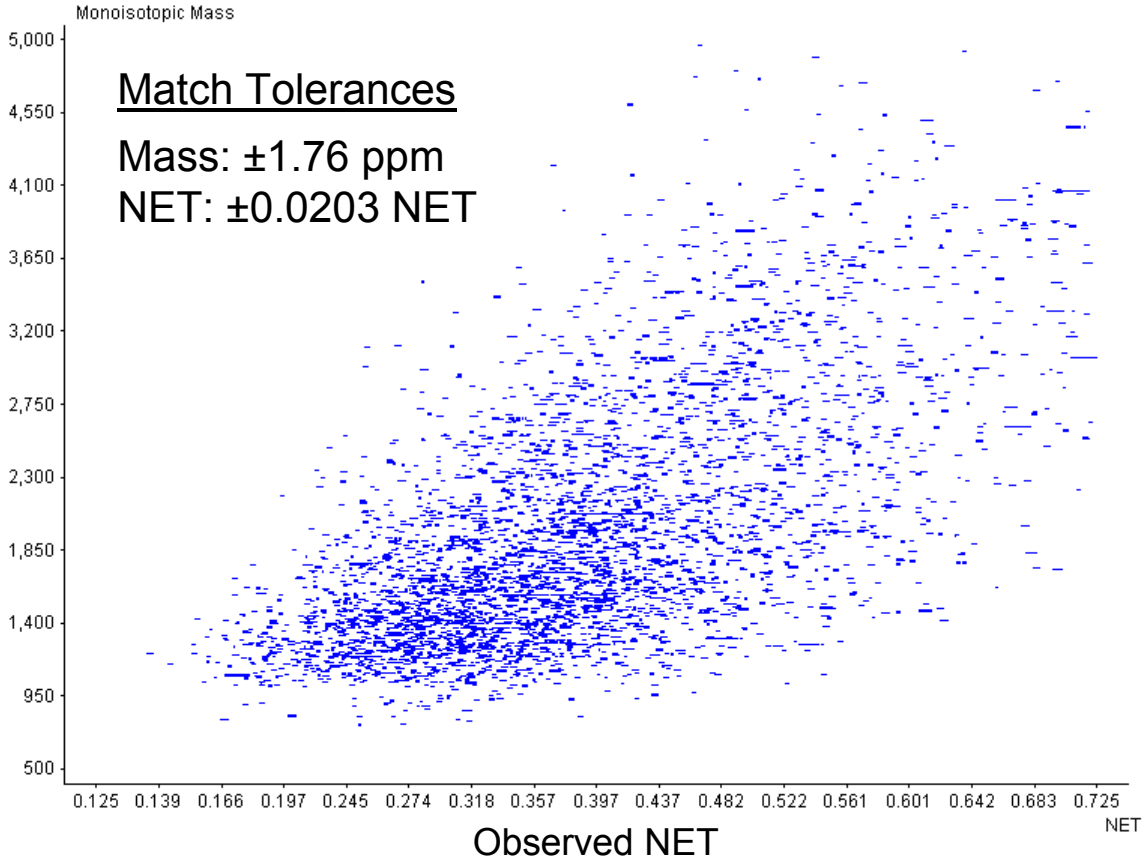
- Repeat search with final search tolerances
  - 5,934 features
  - 3,866 features with matches



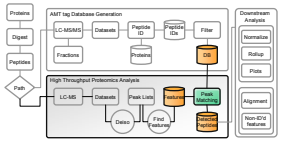
Identifying LC-MS Features



- Repeat search with final search tolerances
  - 5,934 features
  - 3,866 features with matches
  - 3,958 out of 18,617 AMT tags matched using ±1.76 ppm



Identifying LC-MS Features

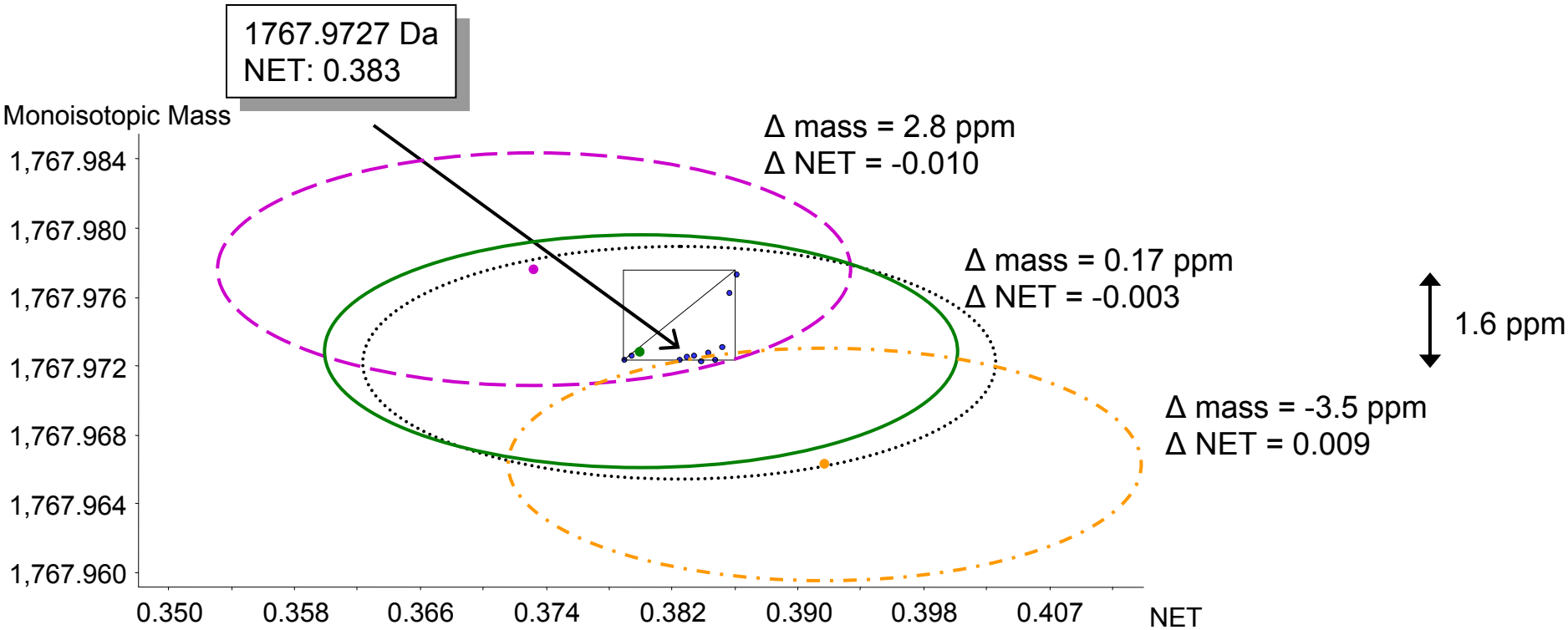


- Caveat: given feature can match more than one AMT tag
  - Need measure of ambiguity

Match Tolerances

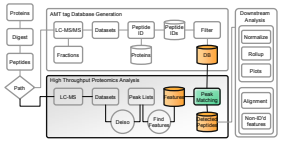
Mass: ±4 ppm
NET: ±0.02 NET

AMT Tag ID	Peptide	Mass (Da)	NET
35896216	T.RALMQLDEALRPSLR.S	1767.9777	0.373
105490	K.DLETIVGLQTDAPLKR.A	1767.9730	0.380
36259992	R.SIGIAPDVLICRGDRAI.P	1767.9664	0.392



More info

Identifying LC-MS Features



$$d_{ij}^2 = \frac{(m_i - \mu_{mj})^2}{\sigma_{mj}^2} + \frac{(t_i - \mu_{tj})^2}{\sigma_{tj}^2}$$

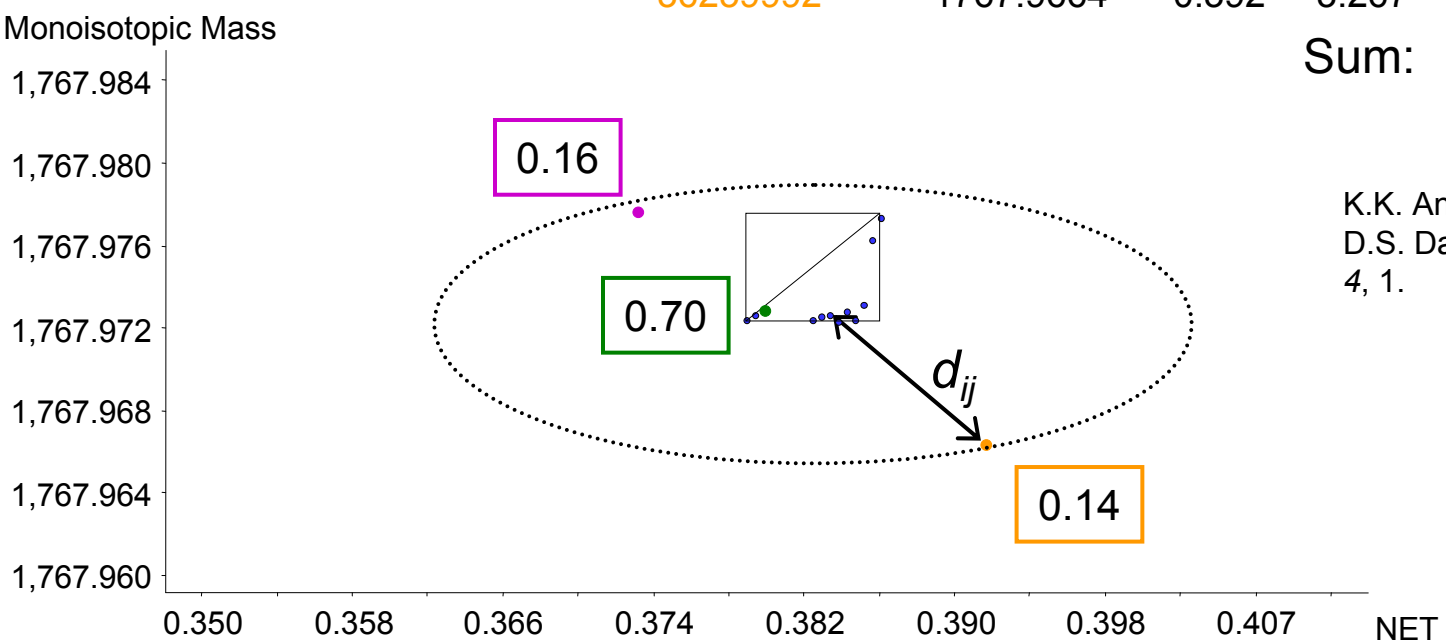
σ_{mj} = 4 ppm, σ_{tj} = 0.025

$$p_{ij} = \frac{(\sigma_{mj}\sigma_{tj})^{-1}\exp(-d_{ij}^2/2)}{\left(\sum_{k=1}^{N}(\sigma_{mk}\sigma_{tk})^{-1}\exp(-d_{ik}^2/2)\right)}$$

Match Tolerances

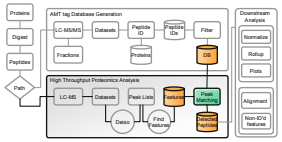
Mass: ±4 ppm
NET: ±0.02 NET

AMT Tag ID	Mass (Da)	NET	d_{ij}^2	Numerator	p_{ij}
35896216	1767.9777	0.373	3.012	6273.3	0.16
105490	1767.9730	0.380	0.090	27042.5	0.70
36259992	1767.9664	0.392	3.267	5521.4	0.14
			Sum:	38837.2	



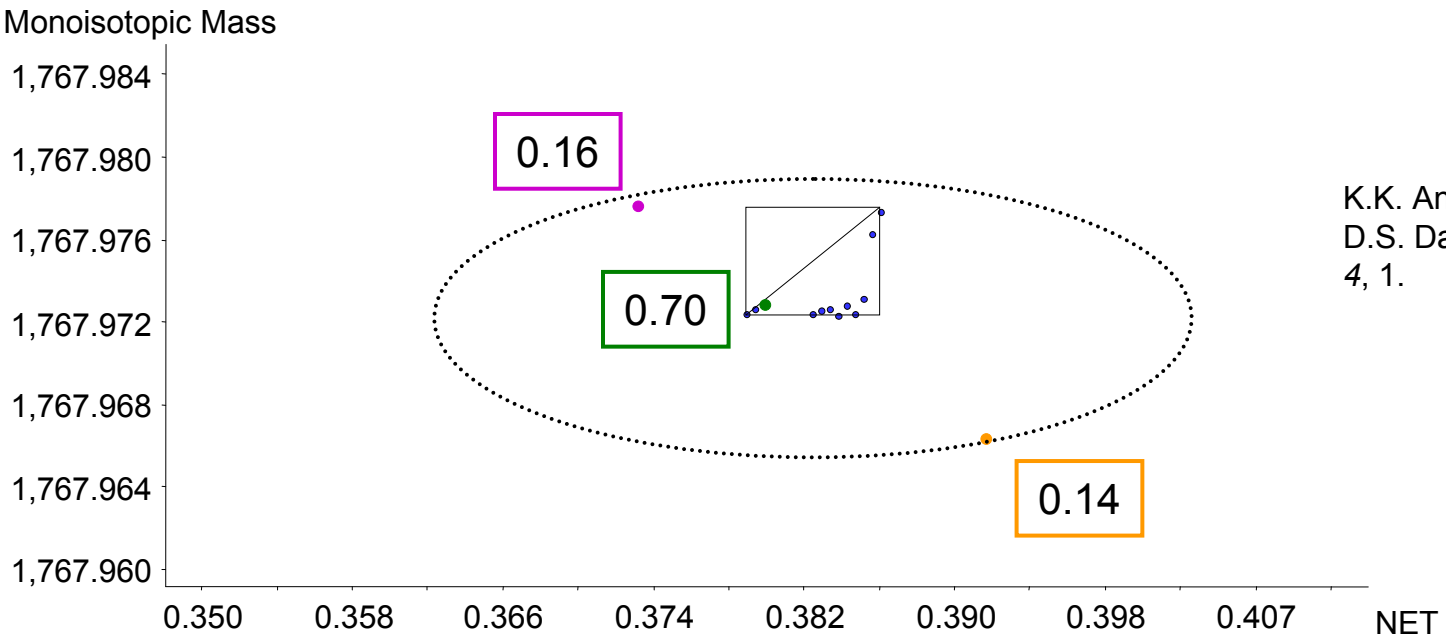
K.K. Anderson, M.E. Monroe, and D.S. Daly. *Proteome Science* **2006**, *4*, 1.

Identifying LC-MS Features



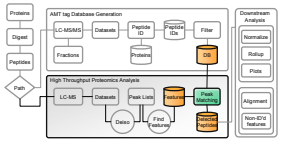
- VIPER reports a score that measures the uniqueness of each match

AMT Tag ID	Peptide	Mass (Da)	NET	SLiC Score	Average XCorr	Avg Disc Score
35896216	T.RALMQLDEALRPSLR.S	1767.9777	0.373	0.16	3.13	0.61
105490	K.DLETIVGLQTDAPLKR.A	1767.9730	0.380	0.70	3.68	0.97
36259992	R.SIGIAPDVLICRGDRAI.P	1767.9664	0.392	0.14	2.15	0.06



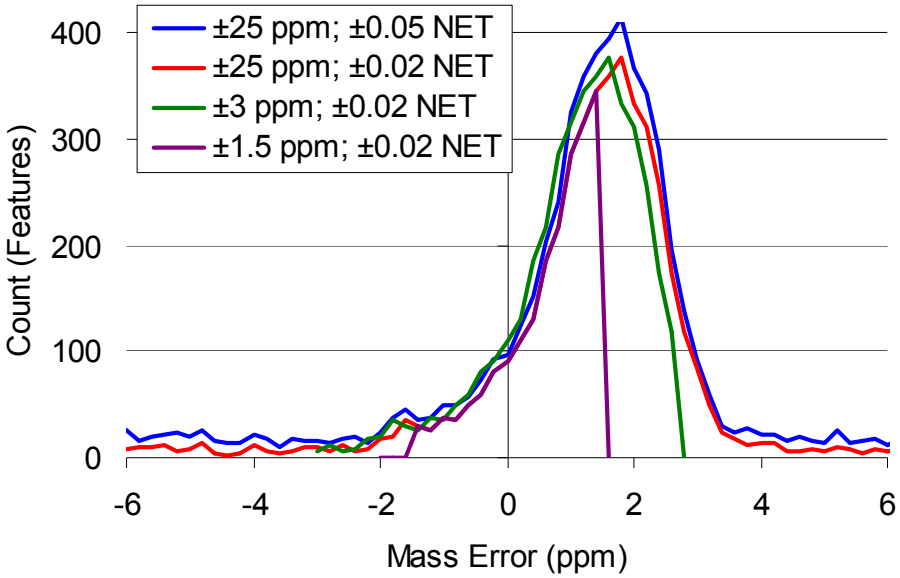
K.K. Anderson, M.E. Monroe, and D.S. Daly. *Proteome Science* **2006**, *4*, 1.

Search Tolerance Refinement

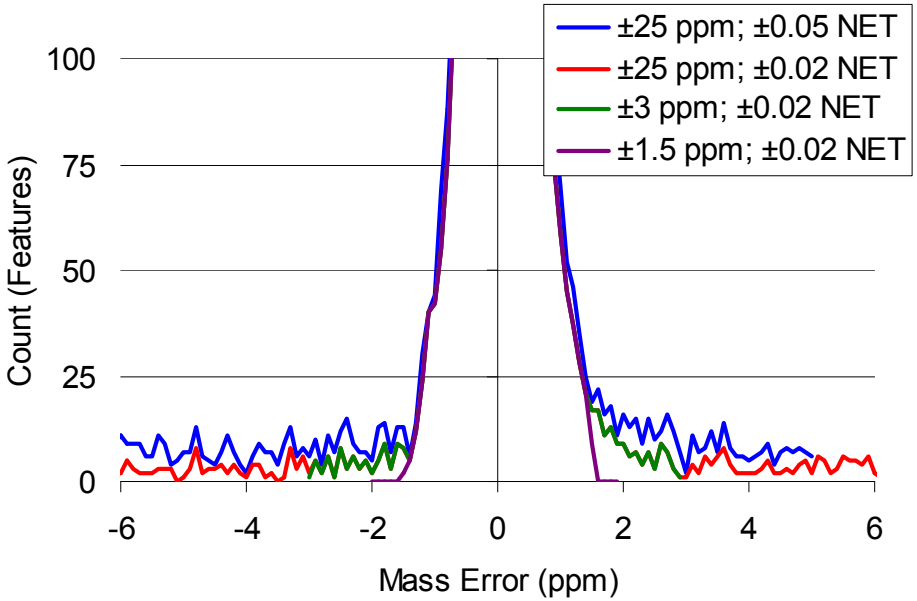


- Effect of search tolerances on Mass Error histogram
  - If mass error plot not centered at 0, then narrow mass windows exclude valid data
  - Decreasing mass and/or NET tolerance reduces background false positive level

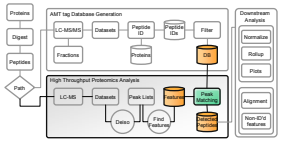
Mass error histograms with linear mass correction



Mass error histograms with LCMSWarp mass correction



Automated Peak Matching



- Automated processing using VIPER
  - Processing steps and parameters defined in .Ini file
    - Separate .Ini file for $^{14}N/^{15}N$ pairs and $^{16}O/^{18}O$ pairs

Edit Analysis Settings

1. Load and Filter | 2. LC-MS Features | 3. MT Tags | 4. Pairs | 5. NET Adjustment | 6. Refinement | 7. DB Search | 8. Saving/Plotting

Input File extension preference order: isos_ic.csv, isos.csv, .mzxml, .mzdata, mzxml.xml, mzdata.xml, _ic.pek, _s.pek, .pek, DeCal.pek-3, .pek-3

Molecular Mass Range
- Use only Isotopic data with molecular mass within range — Min 400, Max 6000
- Use only Charge State data with molecular mass within range — Min 400, Max 6000
- Same range for Charge State and Isotopic data

Abundance Range
- Use only Isotopic data with abundance within range — Min 0, Max 1E+15
- Use only Charge State data with abundance within range — Min 0, Max 1E+15
- Same range for Charge State and Isotopic data

Scan Range and GANET Range
- Use only scans within range — Min 0, Max 5000
- Use only scans whose GANET value is within range — Min -1, Max 2

Isotopic data options
- Exclude data with calculated isotopic fit worse (higher) than: 0.15
- Use only charge states within range — First C.S. 1, Last C.S. 6
- Use Isotopic MW From: Average field / Monoisotopic field / Most abundant

M/Z Range (isotopic data only)
- Use only Isotopic data with m/z within range — Min 400, Max 3000

Data Count Filter
- Maximum data count filter enabled (requires pre-scan of data file)
- Maximum data count to load: 400000

Specialized filters
- Exclude second guess
- Exclude less likely guess
- Exclude duplicates from the Isotopic data — Duplicate Tolerance 2
- Even/Odd Scan Number Filtering (use this filter for DREAMS-based data files): Use All Scans

Read from Selected Gel | Apply to Selected Gel | Set to Defaults | Gel file (in memory) to read or update | Read from Settings File | Save to Settings File | Revert | Close

Edit Analysis Settings

1. Load and Filter | 2. LC-MS Features | 3. MT Tags | 4. Pairs | 5. NET Adjustment | 6. Refinement | 7. DB Search | 8. Saving/Plotting

- Disable all saving and exporting
- Database Search Mode: Conglomerate LC-MS Feature masses with NET (pref
- ConglomerateUMCsWithNET — Remove All

Search Options
- Molecular Mass Field: Average / Monoisotopic / The Most Abundant
- Use Class NET for LC-MS Features
- Molecular Mass Tolerance: Tolerance 6 ppm / Dalton
- NET tolerance 0.025
- Elliptical search region

Note: Since Tolerance Refinement is enabled, the above mass and NET tolerances are used as the search tolerances only if tolerance refinement fails.

Text Export Options
- Write search results by ion to text file after auto search completion
- Include ORF name in text file output
- Save all LC-MS Feature to text file (not just those with DB matches)
- Text output file separation character: <TAB>

Options for selected Database Search Mode

Modifications
- PEO
- ICAT d0
- ICAT d8
- Alkylation — Alkylation mass: 57.0215
- Residue to modify: Full MT — Mass (Da): 0
- Mod Type: Fixed / Dynamic
- N Type: N14 / N15

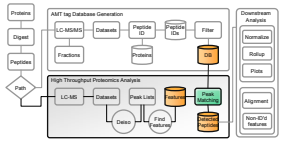
Internal Std Search Mode: Search MT tags & lockers
- Minimum XCorr 0
- Min Discrim. Score 0
- Min Pep Prophet 0

- Export search results to database
- Write search results to text file
- Alternate output folder path (ignored during PRISM-initiated analysis) — Browse

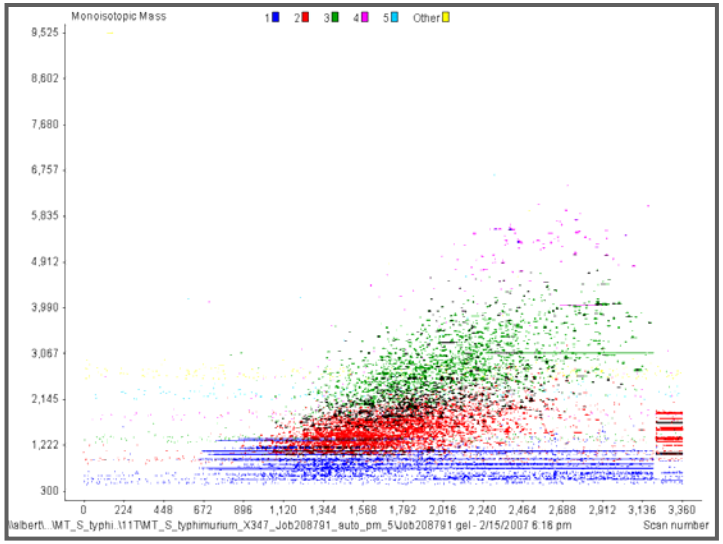
Search Result Export Options
- Set Is_Confirmed = 1 for database search matches
- Add Quantitation Description Entry
- Export LC-MS Features with no database matches
- Export results file uses job number instead of dataset name

Read from Selected Gel | Apply to Selected Gel | Set to Defaults | Gel file (in memory) to read or update | Read from Settings File | Save to Settings File | Revert | Close

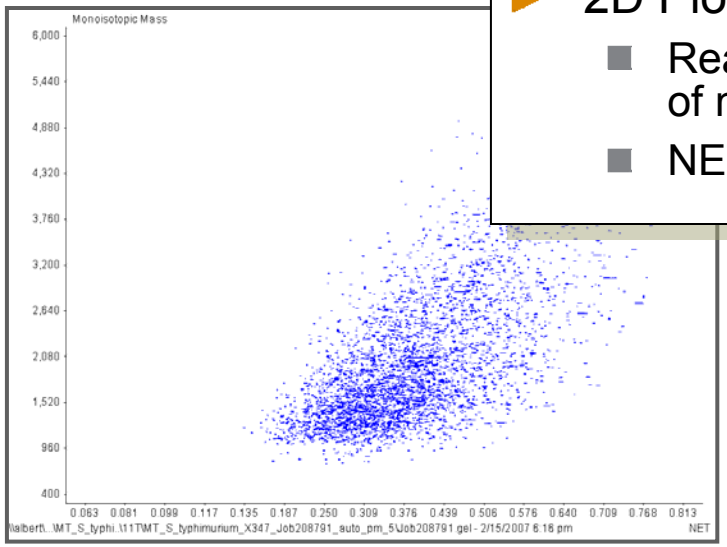
Peak Matching Results



- Browsable result folders for visual QC of each dataset
  - *S.* Typhimurium on 11T FTICR

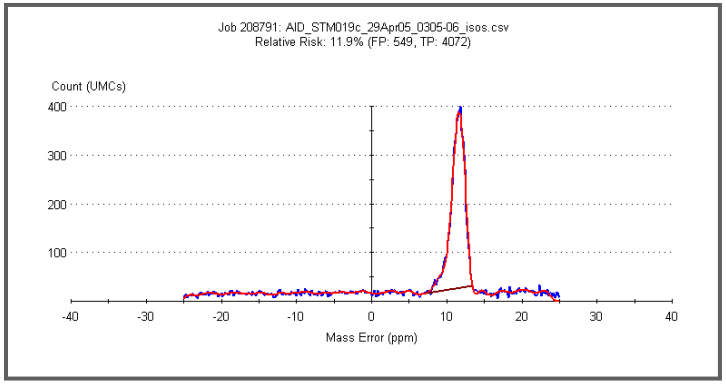


Data Searched

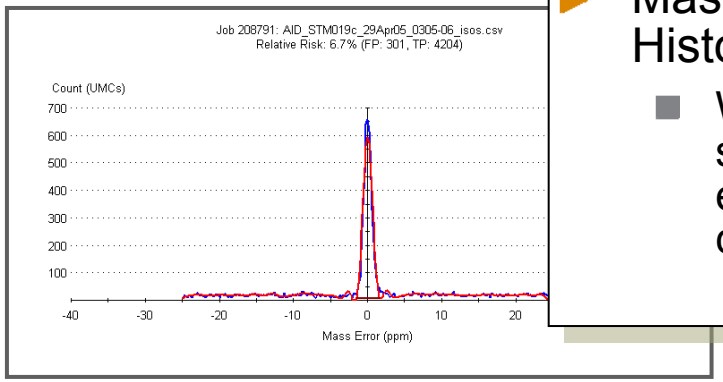


Data With Matches

- 2D Plot Metrics
  - Reasonable number of matches
  - NET range ≈ 0 to 1



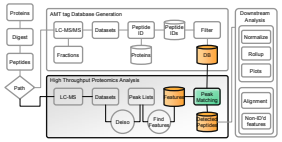
Mass Errors Before Refinement



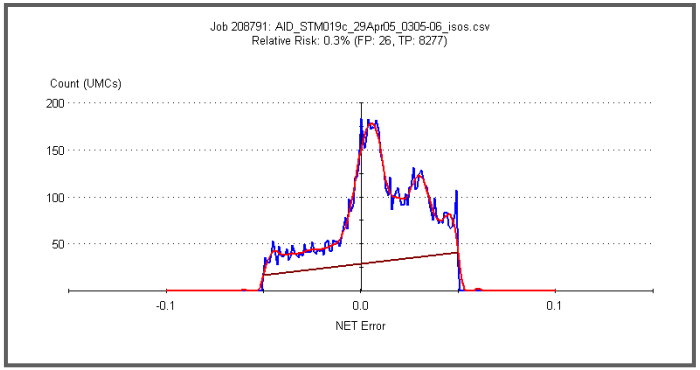
Mass Errors After Refinement

- Mass Error Histogram Metrics
  - Well defined, symmetric mass error peak centered at 0 ppm

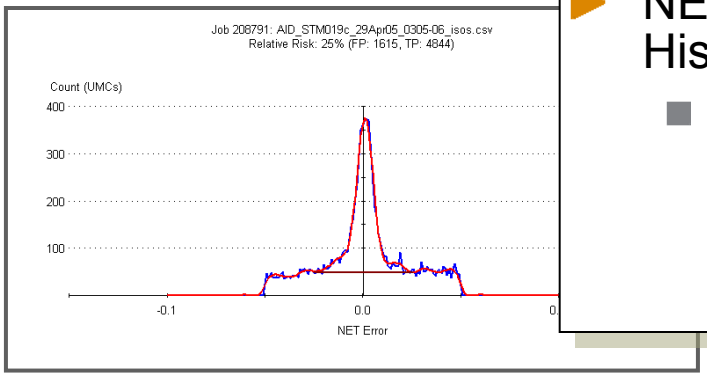
Peak Matching Results



- Browsable result folders for visual QC of each dataset
  - *S.* Typhimurium on 11T FTICR

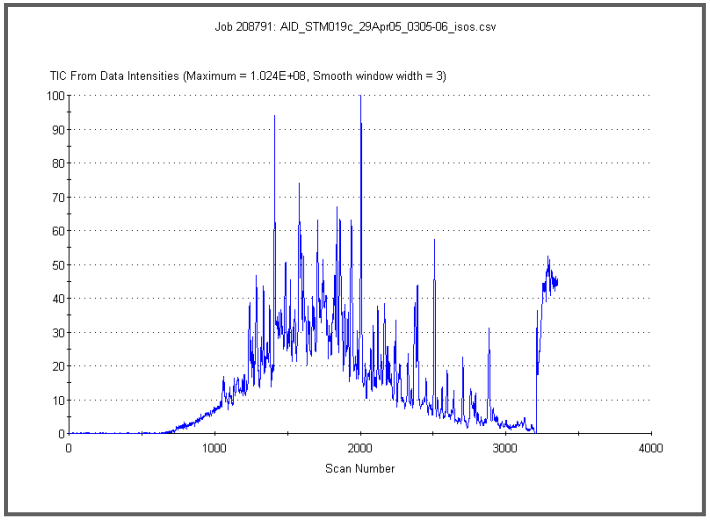


NET Errors Before Refinement

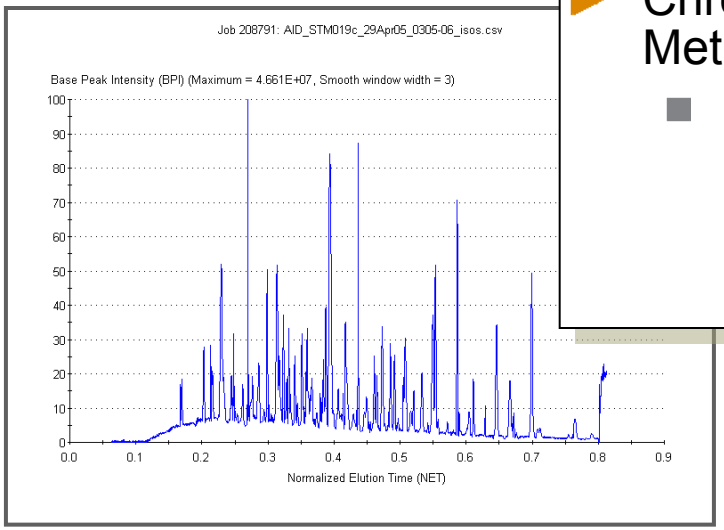


NET Errors After Refinement

- NET Error Histogram Metrics
  - Well defined, symmetric NET error peak centered at 0



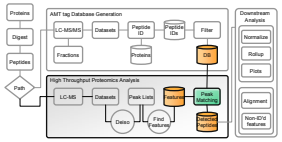
Total Ion Chromatogram (TIC)



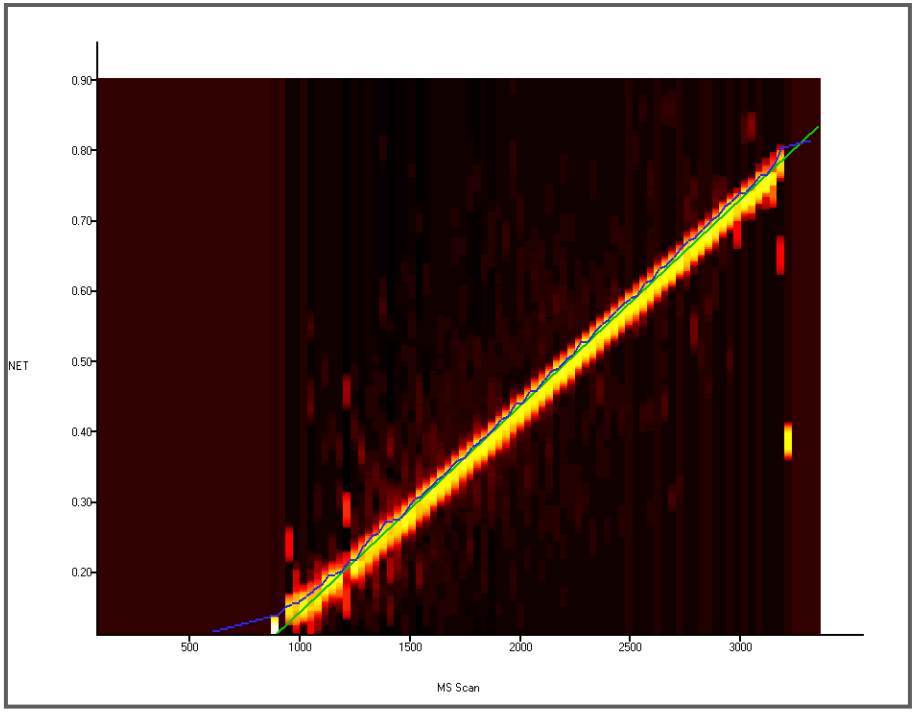
Base Peak Intensity (BPI) Chromatogram

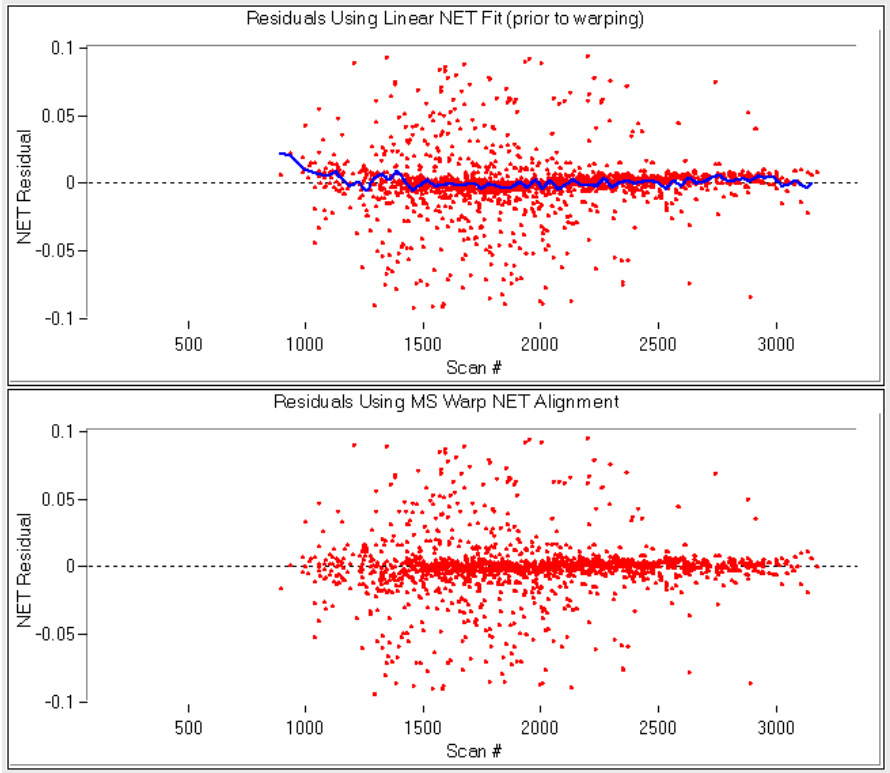
- Chromatogram Metrics
  - Narrow peaks evenly distributed throughout separation window

Peak Matching Results



- Browsable result folders for visual QC of each dataset
  - *S.* Typhimurium on 11T FTICR

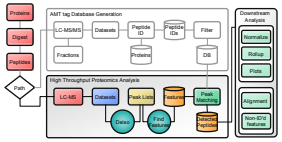




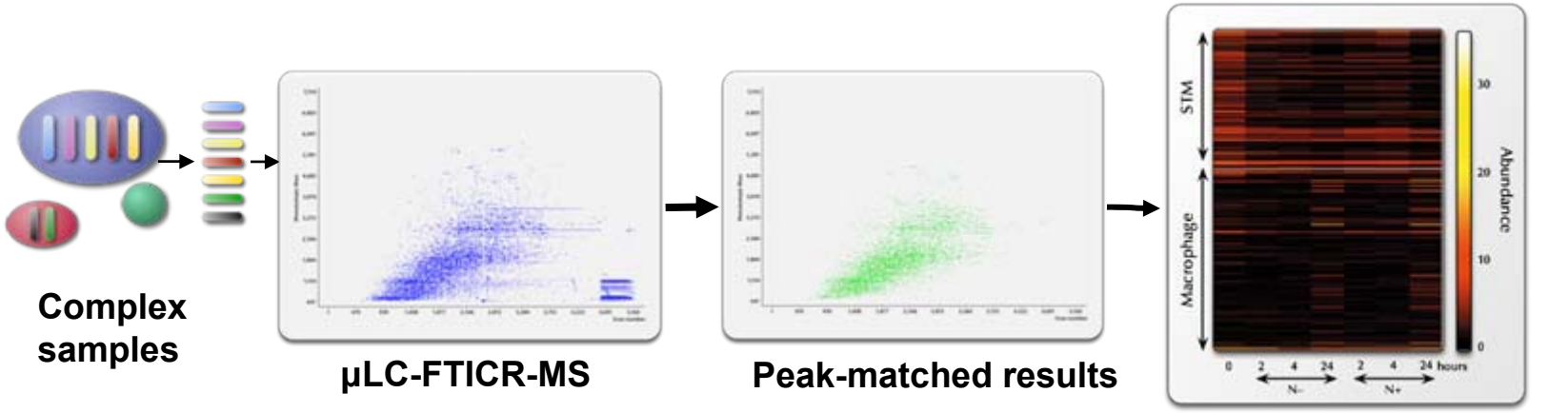
- NET Alignment Surface Metrics
  - Should show a smooth, bright yellow, diagonal line

- NET Alignment Residual Metrics
  - Data after recalibration should be narrowly distributed around zero

Peak Matching Results



- Peptide identification list
  - Peptides (and thus proteins) identified in each sample
  - Peptide abundance estimates (relative abundance)
  - Confidence metrics
- Next, need to compare peptide abundances (and/or protein abundances) between samples to reveal biological information

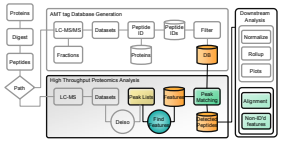


Compare abundances across multiple proteomes

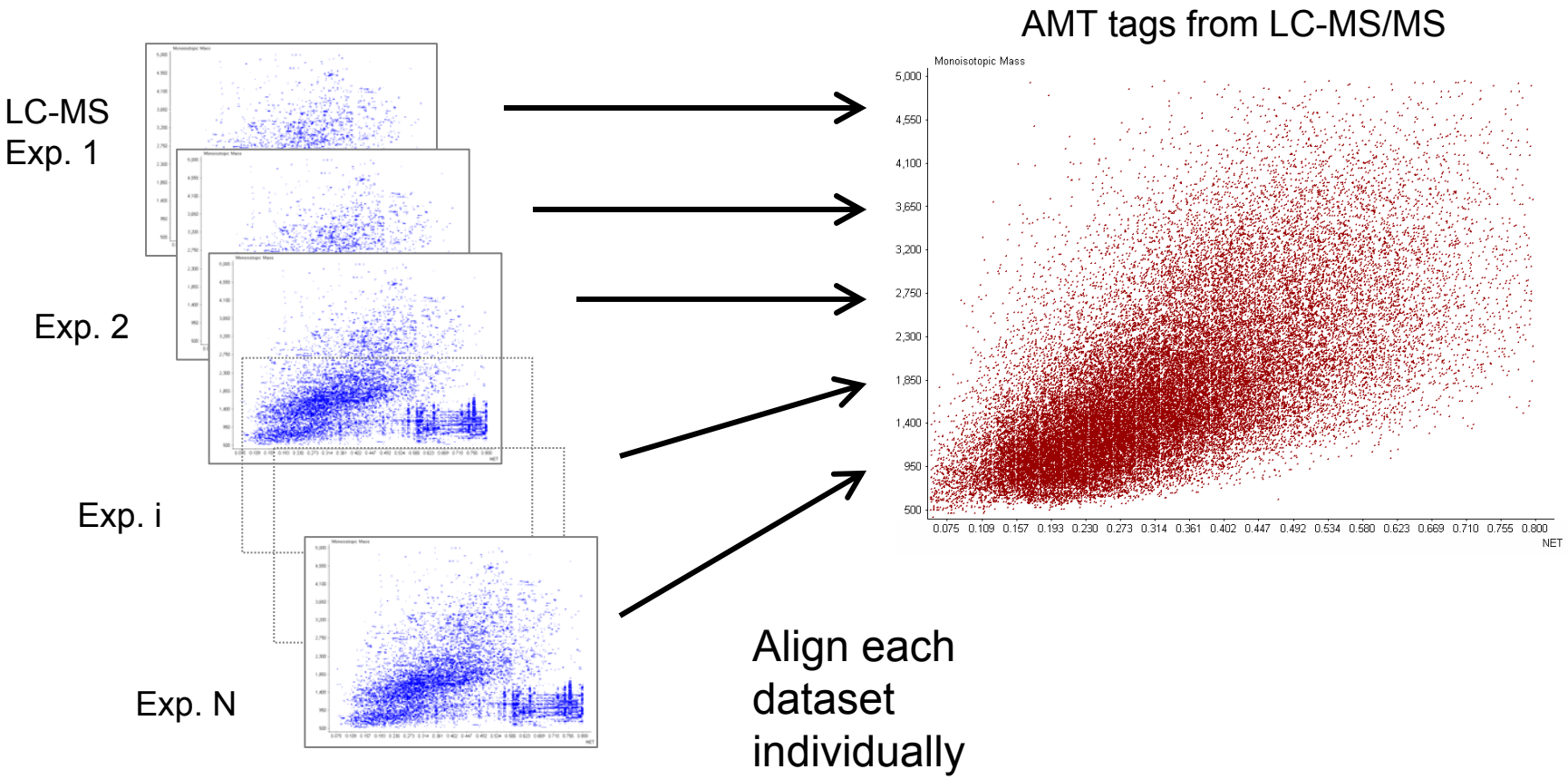
Part II: LC-MS Feature Discovery

- Part I: Introduction and Overview of Label-Free Quantitative Proteomics (Anderson)
- Part II: Feature Discovery in LC-MS Datasets (Monroe and Polpitiya)
  - ✓ Structure of LC-MS Data
  - ✓ Feature discovery in individual spectra (deisotoping)
  - ✓ Feature definition over elution time
  - ✓ Identifying LC-MS Features using an AMT tag DB
  - Extending the AMT tag approach for feature based analyses
  - Estimating confidence of identified LC-MS features
  - Downstream quantitative analysis with DAnTE
- Break
- Part III: Biological Application of the AMT tag Approach (Ansong)
- AMT tag Analysis Software Demo
- Panel Discussion

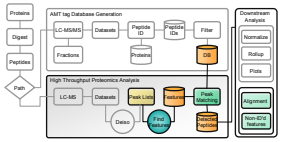
Independent Dataset Processing



- Individual LC-MS datasets are aligned to an AMT tag database sequentially
- Identified peptides are compared after independently processing each dataset

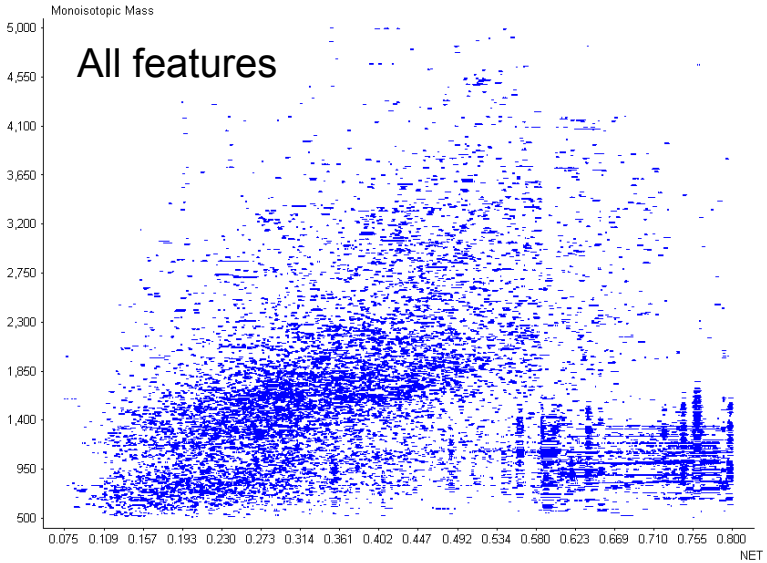


Independent Dataset Processing

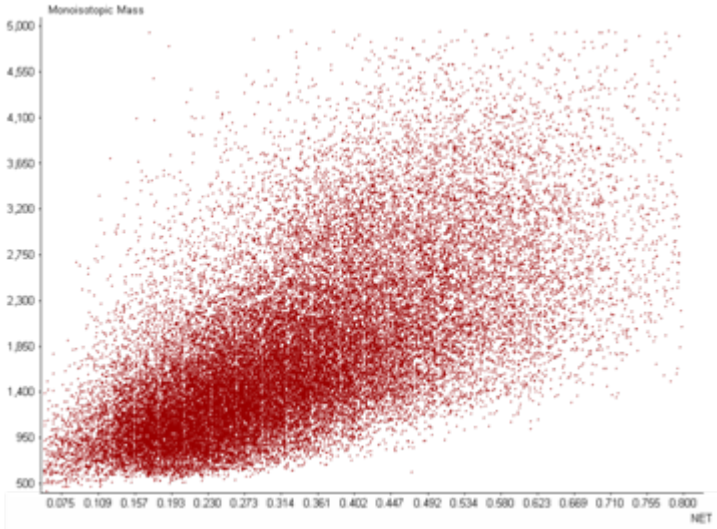


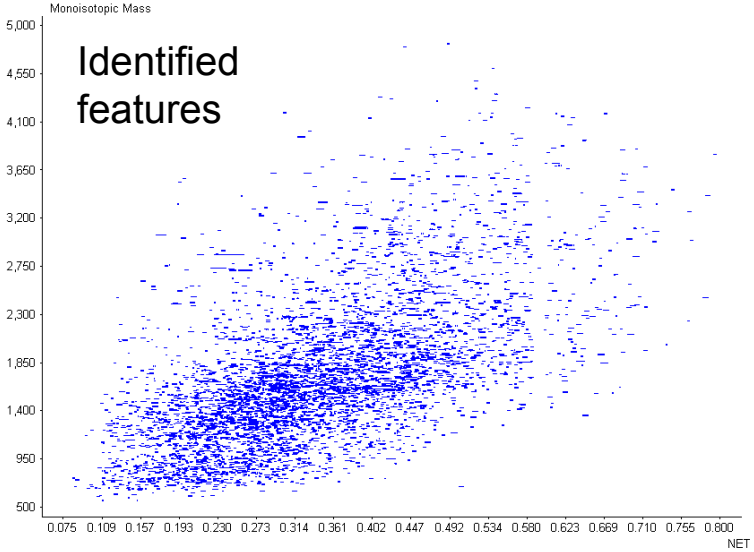
- LC-MS features without matches may represent useful information, but are effectively ignored

LC-FTICR-MS

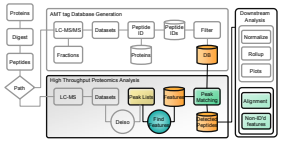


AMT tags from LC-MS/MS





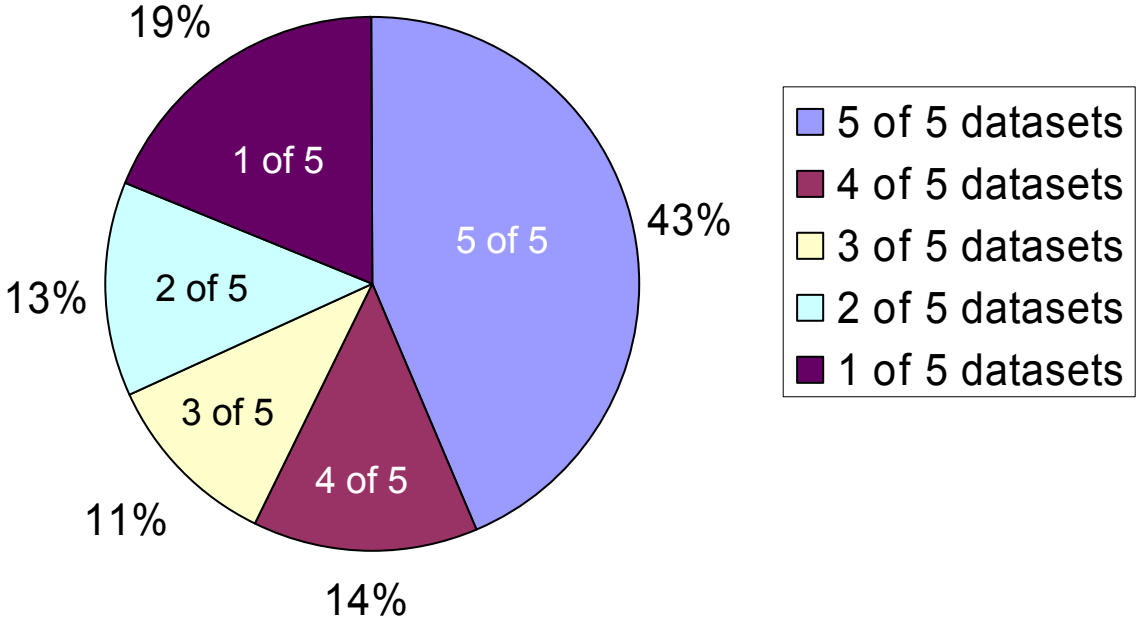
Independent Dataset Processing



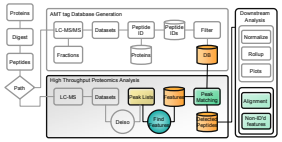
- Independent processing of each dataset results in more missing data
  - For example, if a low abundance peptide is not identified as an LC-MS feature in a given dataset, then that peptide identification is missing from that dataset
- Lower abundance features suffer more, but are not the only casualties

Peptide Detection Reproducibility in Replicate Datasets

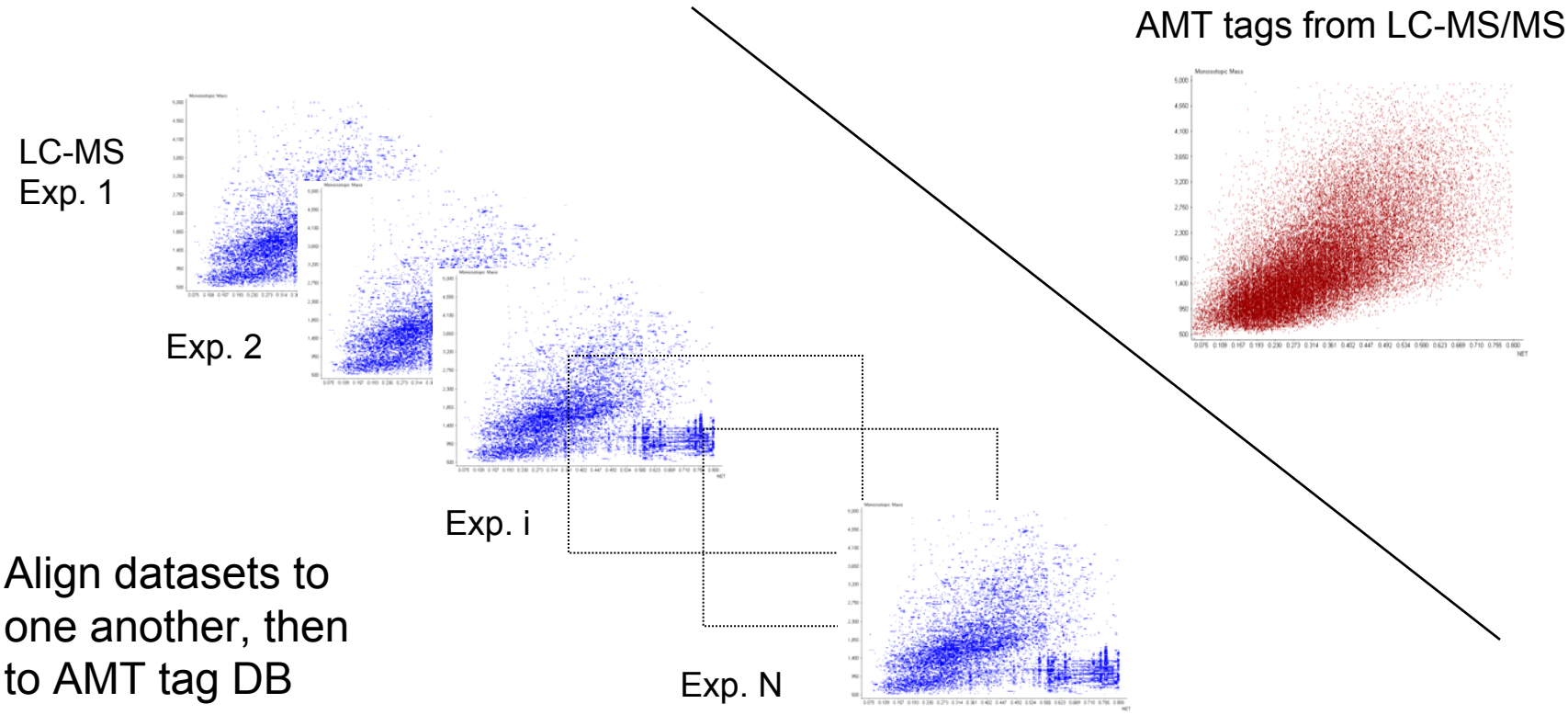
Frequency	Number of Peptides
Seen in all 5	3079
4 of 5	974
3 of 5	766
2 of 5	926
1 of 5	1332



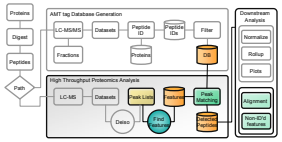
Extended AMT tag Method



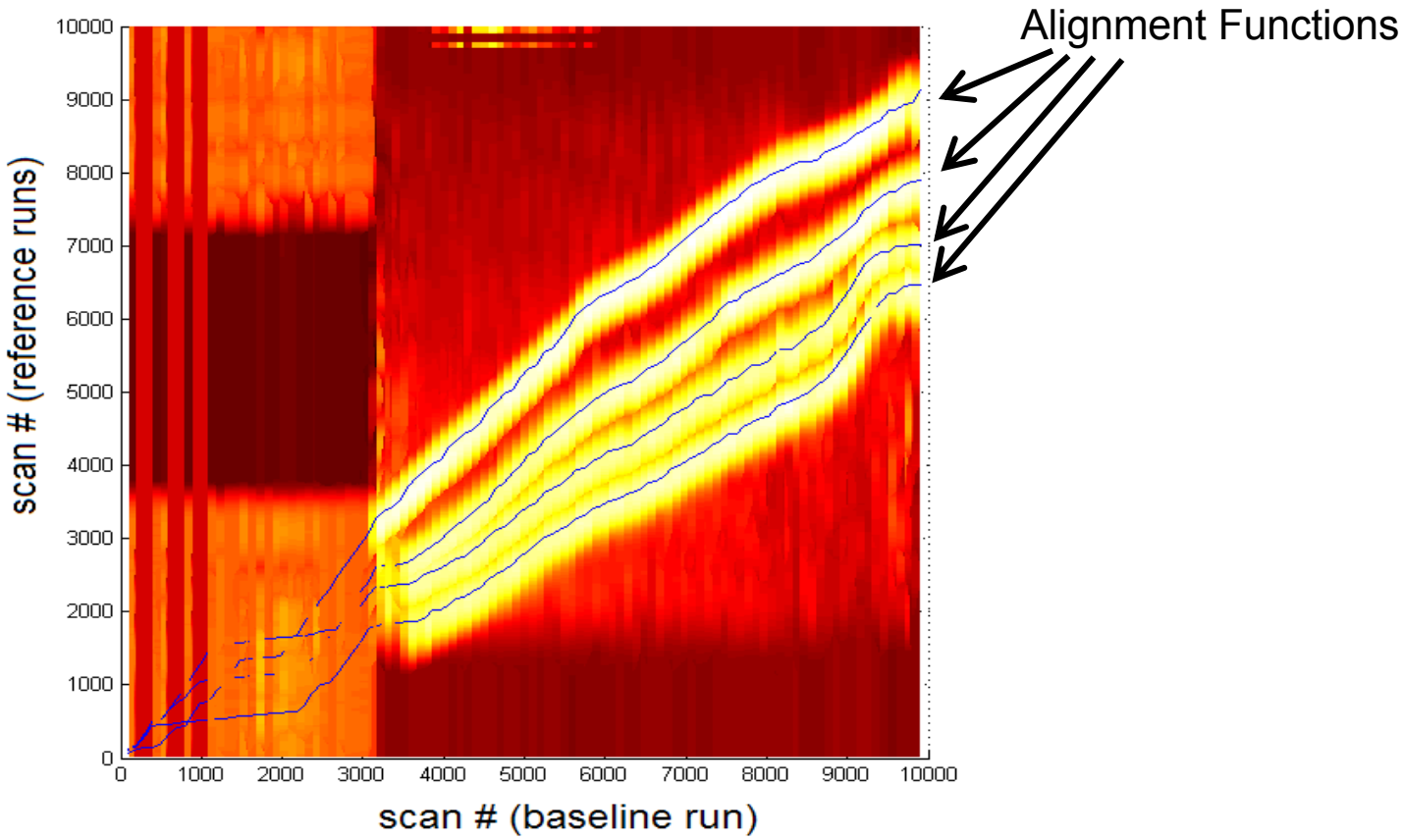
- Find common features based on mass and time patterns in all datasets first (with or without the AMT tag database)
- Align resulting groups of features to the AMT tag database using statistics from a larger number of features



Extended AMT tag Method



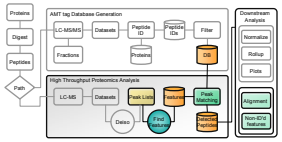
- Align all datasets to common baseline



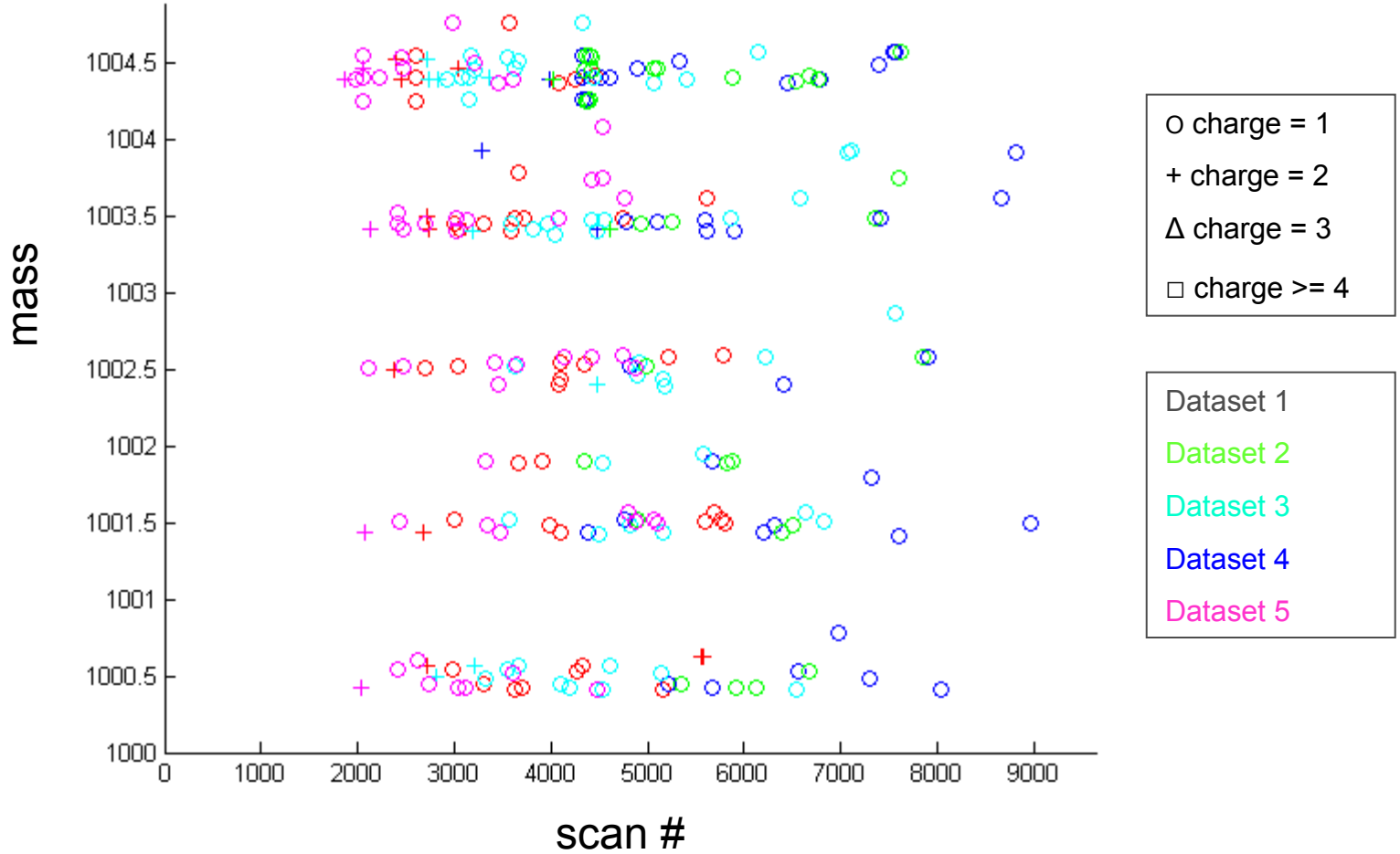
Score plots for alignment of 4 datasets against arbitrary baseline run

N. Jaitly, M.E. Monroe et. al., *Analytical Chemistry* **2006**, *78*, 7397-7409.

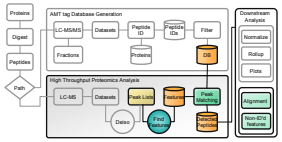
Extended AMT tag Method



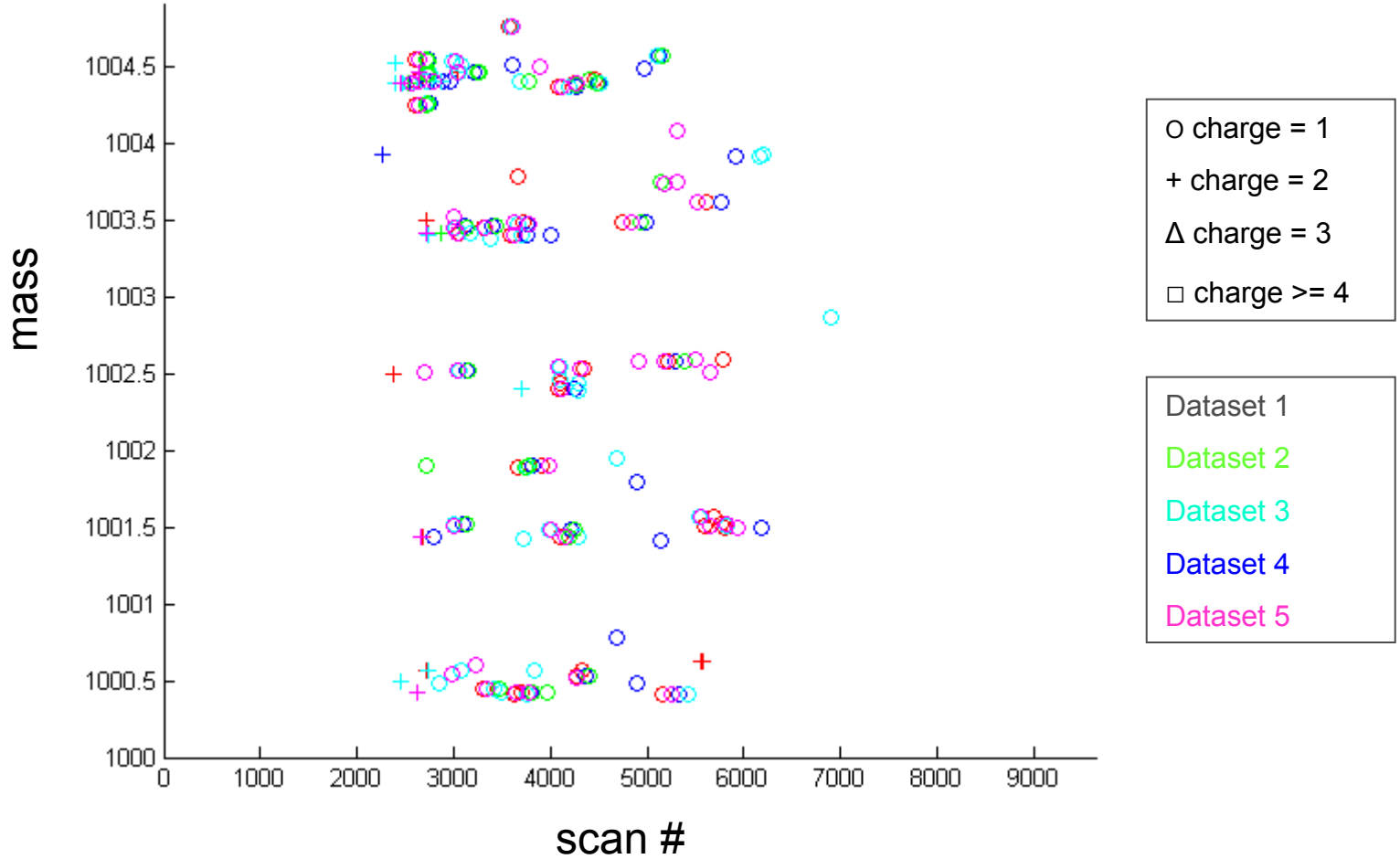
- Example alignment of 5 LC-MS datasets
  - Mass section before LC alignment



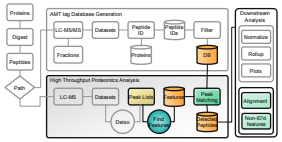
Extended AMT tag Method



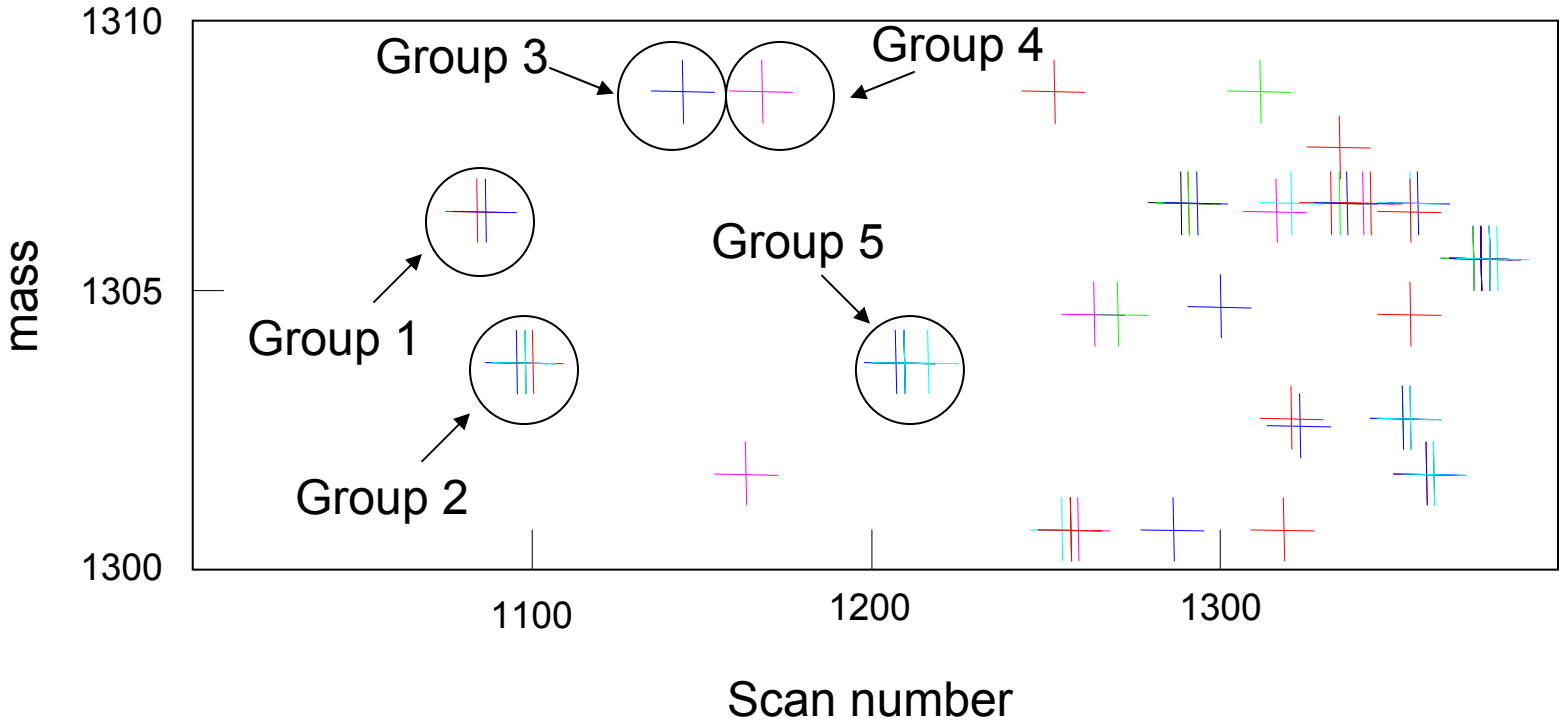
- Example alignment of 5 LC-MS datasets
  - Mass section after LC alignment



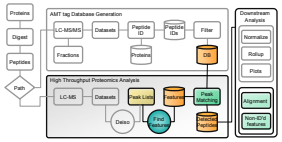
Extended AMT tag Method



- Cluster similar features (using mass and retention time) across all LC-MS datasets, rather than analyzing each dataset separately and then collating results
  - Can obtain abundance profiles for features seen in multiple datasets



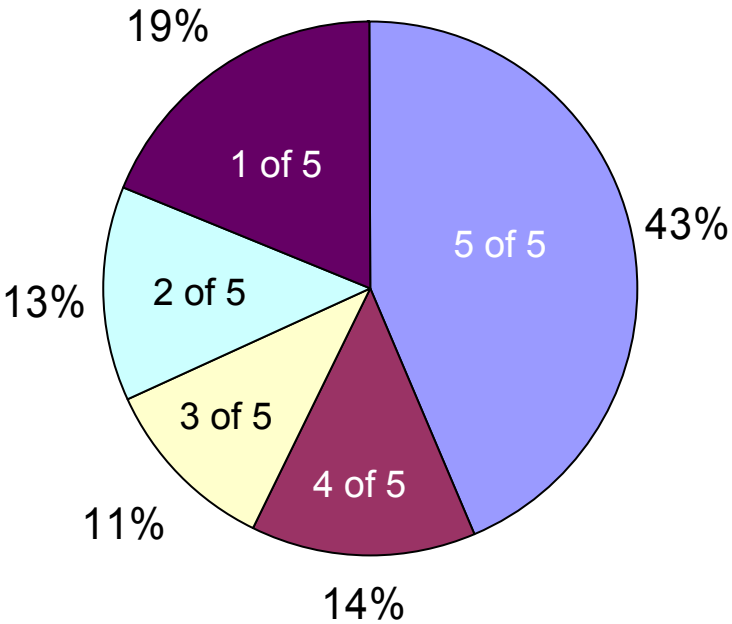
Extended AMT tag Method



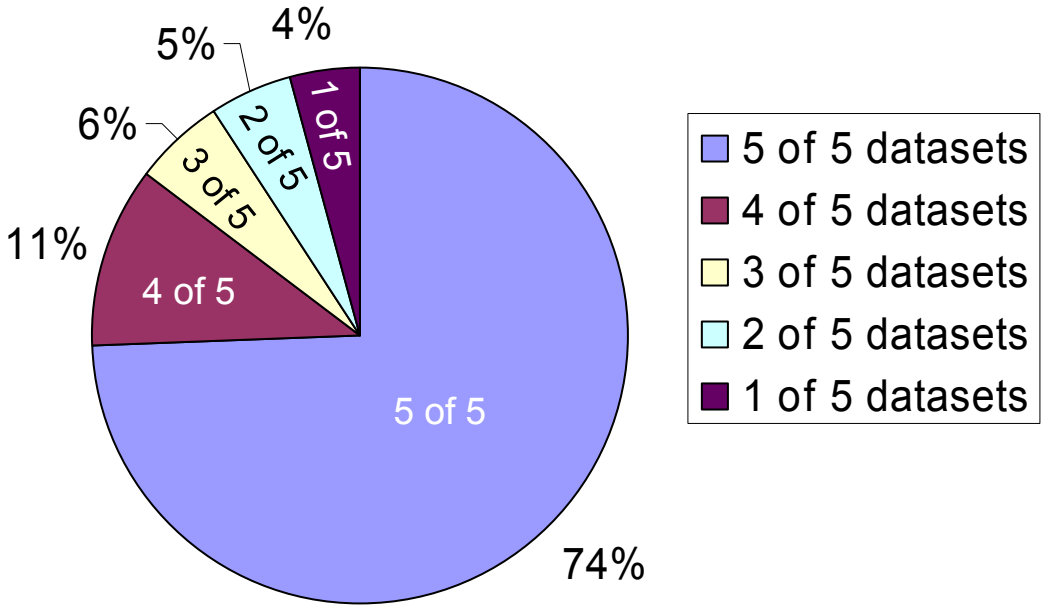
- Identify clustered features by aligning mass and elution time of clusters to AMT tags in database

Peptide Detection Reproducibility Comparison

Independent processing

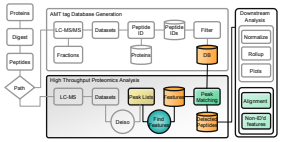


Cluster first, then match to AMT tags

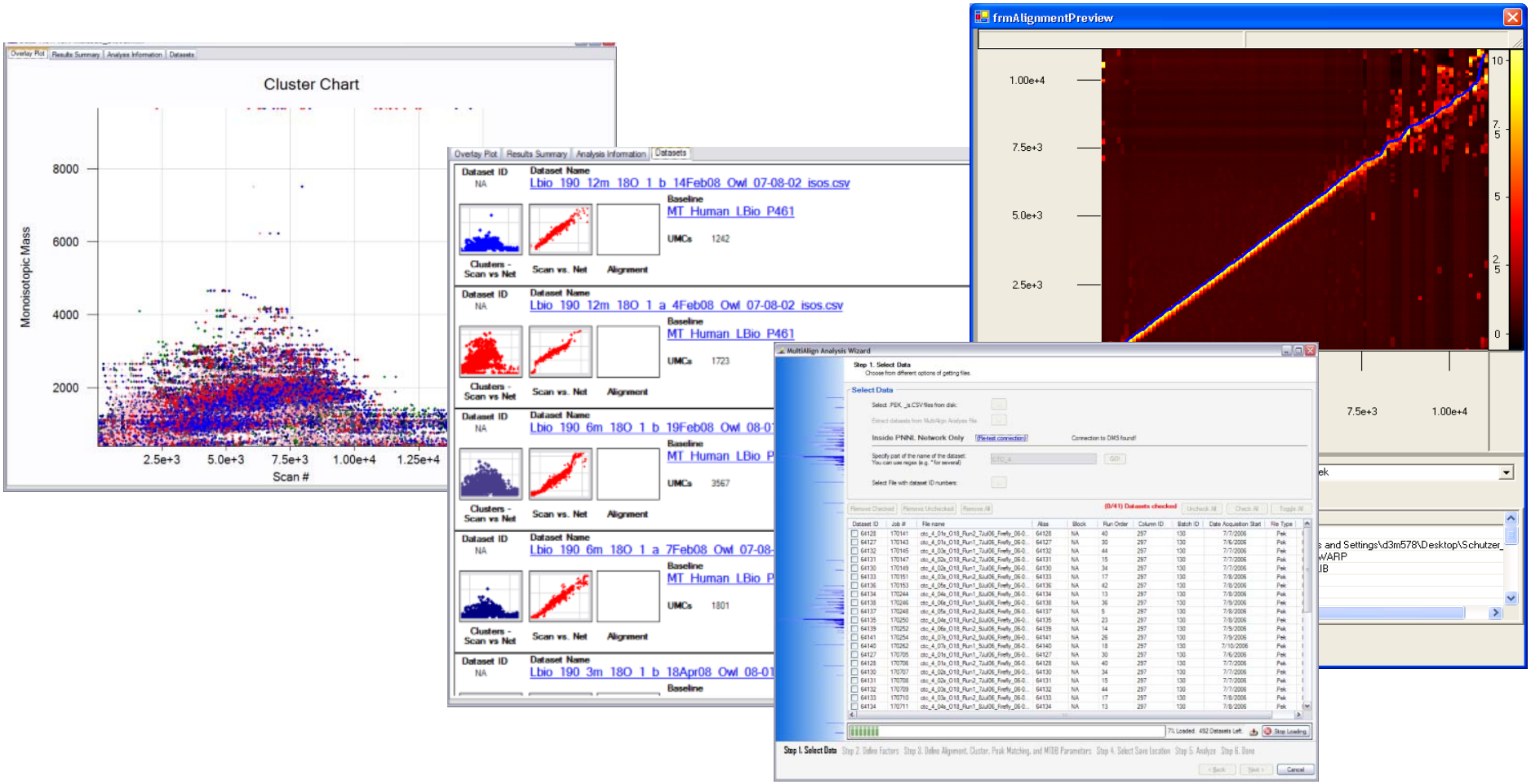


Fewer missing values observed with clustered feature approach

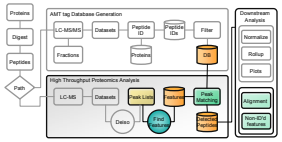
MultiAlign



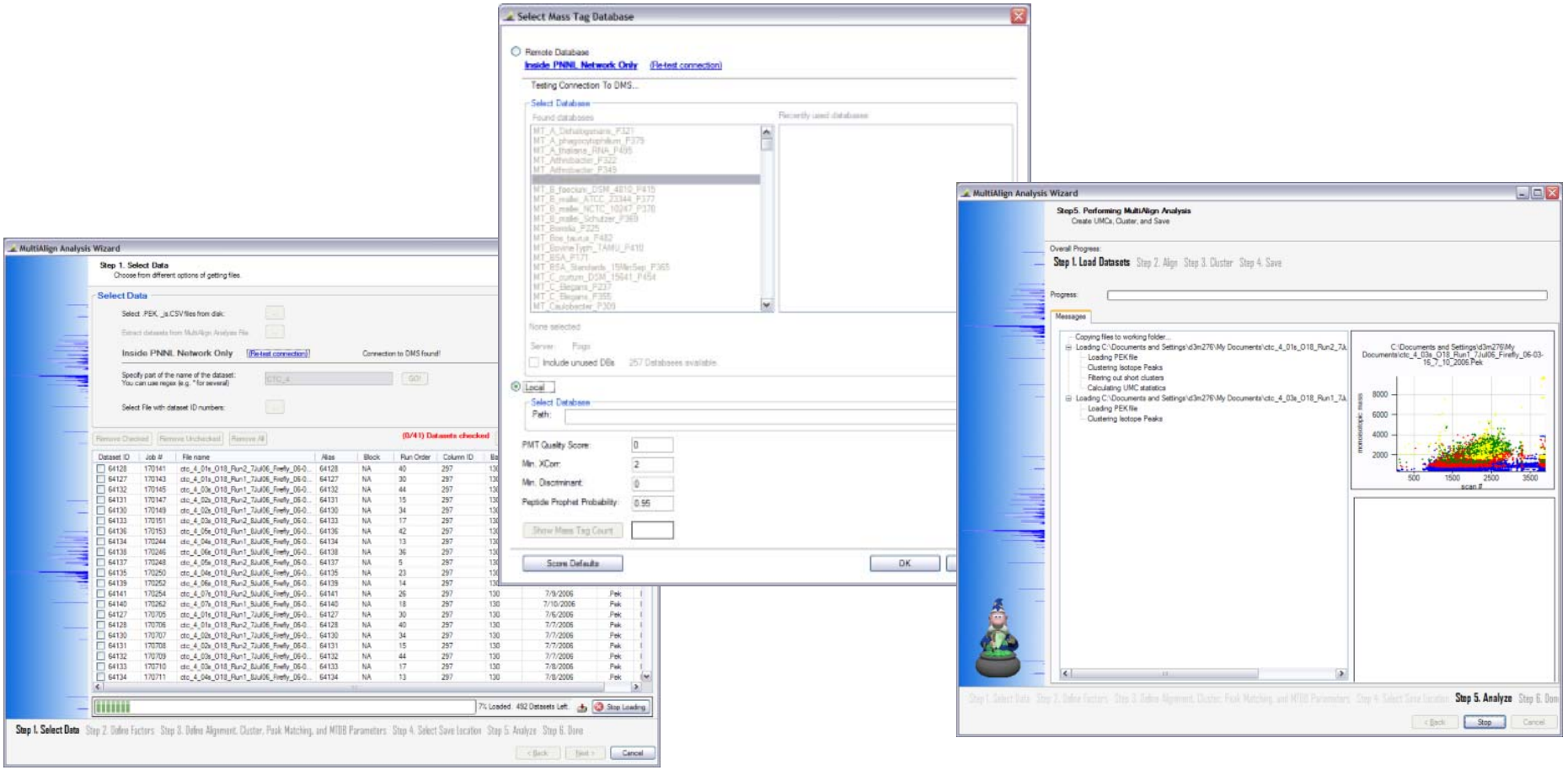
- Represents next version of the feature identification process
- Along with MTDB Creator it represents a standalone, redistributable version of the AMT tag process



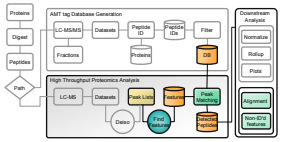
MultiAlign



- Analysis Wizard takes user through analysis setup
- Can perform single dataset alignment to AMT tag DB or multiple dataset alignment to baseline dataset or DB



MultiAlign



- Dataset summary pages and overview are available for visual QC

Overlay Plot | Results Summary | Analysis Information | Datasets

Dataset ID	Dataset Name	Baseline	UMCs
NA	Lbio_190_12m_18O_1_b_14Feb08_Owl_07-08-02_isos.csv	MT_Human_LBio_P461	1242
NA	Lbio_190_12m_18O_1_a_4Feb08_Owl_07-08-02_isos.csv	MT_Human_LBio_P461	1723
NA	Lbio_190_6m_18O_1_b_19Feb08_Owl_08-01-08_isos.csv	MT_Human_LBio_P461	3567
NA	Lbio_190_6m_18O_1_a_7Feb08_Owl_07-08-02_isos.csv	MT_Human_LBio_P461	1801
NA	Lbio_190_3m_18O_1_b_18Apr08_Owl_08-01-08_isos.csv	Baseline	

Clusters - Scan vs Net | Scan vs. Net | Alignment

Overlay Plot | Results Summary | Analysis Information | Datasets

UMC Finding Options | Alignment Options | Mass Tag Database Options (MTDB)

Description	Value
Is Aligned to Mass Tag Database	True
Alignment Baseline Name	MT_Human_LBio_P461
Alignment Type	NET_MASS_WARP
Recalibration Type	HYBRID_CALIB
Apply Mass Recalibration	True
Net Tolerance	0.03
Mass Tolerance	12
Mass Calibration LSQ Number Knots	12
Mass Calibration LSQZScore	2.5
Max Z-Score	3
Mass Calibration Max Jump	20
Mass Calibration Number of Mass Delta Bins	50
Mass Calibration Number of X Slices	12
Mass Calibration Window	25
Mass Calibration Use LSQ	False
Use Promiscuous Points	False
Max Promiscuity	3
Max Time Jump	
Contraction Factor	
Number of Time Sections	100

Copy Selected Items

More info

LC-MS Feature Discovery

- Similar approaches and software tools: High Res LC-MS
  - CRAWDAD
    - G.L. Finney et al. *Analytical Chemistry* **2008**, *80*, 961-971.
  - msInspect
    - M. Bellew et. al. *Bioinformatics* **2006**, *22*, 1902-1909.
  - PEPPeR
    - J. Jaffe et.al. *Mol. Cell. Proteomics* **2006**, *5*, 1927-1941.
  - SpecArray (Pep3D, mzXML2dat, PepList, PepMatch, PepArray)
    - X.-J. Li, et. al. *Mol Cell Proteomics* **2005**, *4*, 1328-1340.
  - SuperHIRN
    - L.N. Mueller et al. *Proteomics* **2007**, *7*, 3470-3480.
  - Surromed label-free quantitation software (MassView)
    - W. Wang et al. *Analytical Chemistry* **2003**, *75*, 4818-4826.
  - XCMS (for Metabolite profiling)
    - C.A. Smith et. al. *Analytical Chemistry* **2006**, *78*, 779-787.

More info

LC-MS Feature Discovery

- Similar approaches and software tools: Low Res LC-MS
  - Signal maps software
    - A. Prakash et. al. *Mol. Cell Proteomics* **2006**, *5*, 423-432.
  - Informatics platform for global proteomic profiling using LC-MS
    - D. Radulovic, et al. *Mol. Cell. Proteomics* **2004**, *3*, 984-997.
  - Computational Proteomics Analysis System (CPAS)
    - A. Rauch et. al. *J. Proteome Research* **2006,** *5*, 112-121.

Part II: LC-MS Feature Discovery

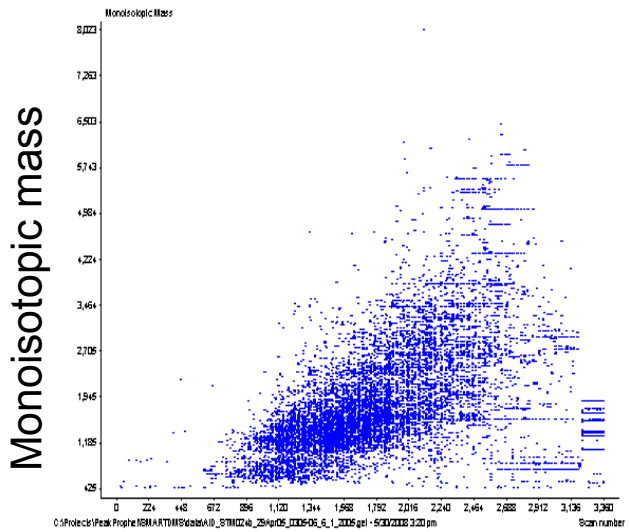
- Part I: Introduction and Overview of Label-Free Quantitative Proteomics (Anderson)
- Part II: Feature Discovery in LC-MS Datasets (Monroe and Polpitiya)
  - ✓ Structure of LC-MS Data
  - ✓ Feature discovery in individual spectra (deisotoping)
  - ✓ Feature definition over elution time
  - ✓ Identifying LC-MS Features using an AMT tag DB
  - ✓ Extending the AMT tag approach for feature based analyses
  - Estimating confidence of identified LC-MS features
  - Downstream quantitative analysis with DAnTE
- Break
- Part III: Biological Application of the AMT tag Approach (Ansong)
- AMT tag Analysis Software Demo
- Panel Discussion

More info

Accurate Mass and Time (AMT) tag pipeline

- Features in an LC-MS dataset are matched to a database of peptides previously identified by LC-MS/MS analyses using specified mass and elution time tolerances

LC-MS dataset

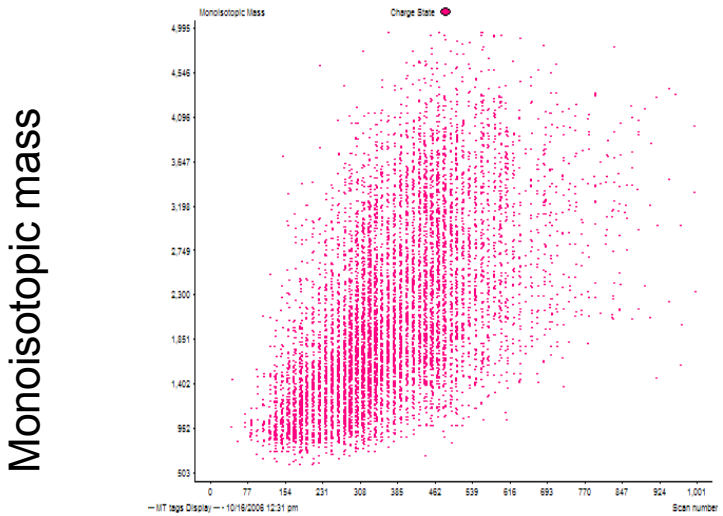


scan #

Alignment &

Matching

AMT tag Database



Normalized elution time (NET)

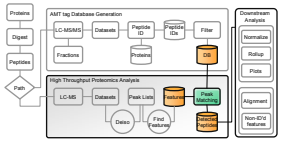
(monoisotopic mass, elution time, abundance)

(Peptide Sequence, monoisotopic mass, elution time)

|Δmass| < mass tolerance

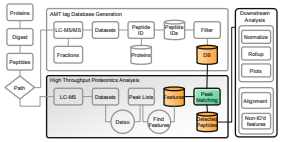
|ΔNET| < elution time tolerance

Current Method for Assessing and Controlling Rate of Random Matches

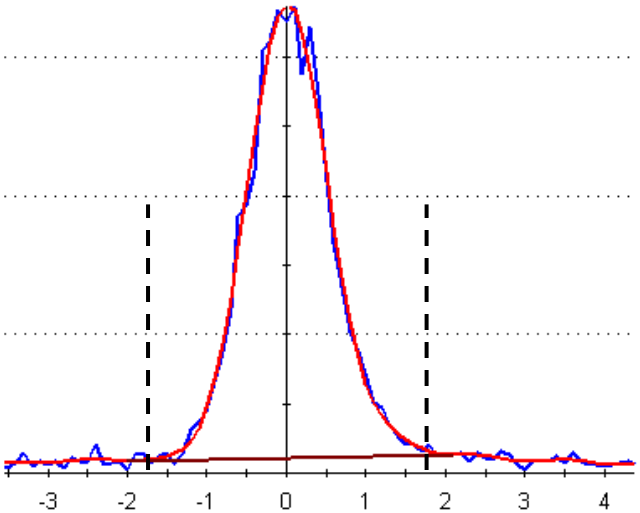


- Decoy database searching is often used to assess rate of random AMT tag matches
  - Shift database (or features) by some mass, and re-match to database
  - The number of matches reflects the random rate of errors
- Rate of error controlled by
  - Building more stringent LC-MS/MS databases (higher XCorr, hyperscore, Mascot score, Peptide Prophet score, etc)
  - Reducing mass and elution time tolerances
- Overall method consists of iteratively controlling and assessing error to choose “optimal” parameters

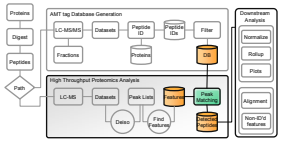
Challenges with Current Method



- Balancing false positives and false negatives is a tricky game
  - Building more confident LC-MS/MS database decreases background false positives, but increases false negatives
  - Reducing mass and elution time tolerances has a similar effect
  - Manually chosen parameters may look suitable, but in reality be sub-optimal
- Each identification is either accepted or rejected
  - In reality some identifications are better than others – higher MS/MS confidence, lower mass and elution time differences, etc.

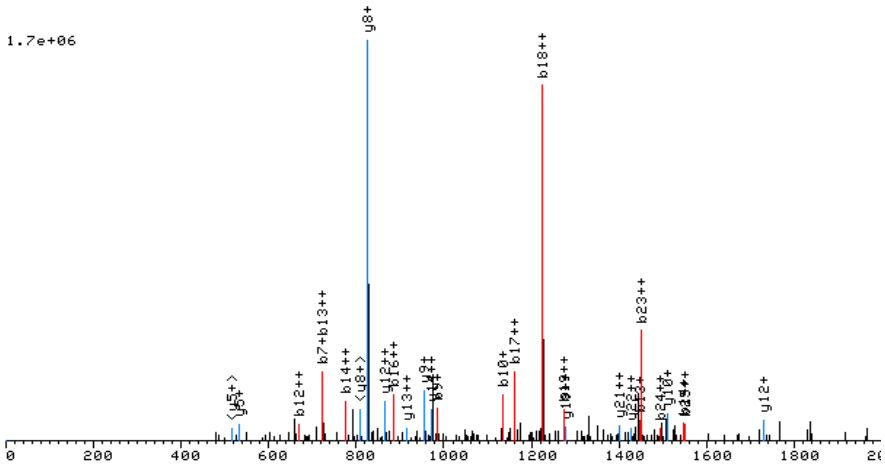


Statistical Method for MS/MS Identification Confidence



- Peptide Prophet – A Statistical Model to estimate probability that an MS/MS spectrum is correctly identified
  - Uses a Linear function (F-Score) of result metrics from SEQUEST to calculate an overall value representing the confidence of identifications

ARHGGEDYVFSLLTGYCEPPTGVSLR



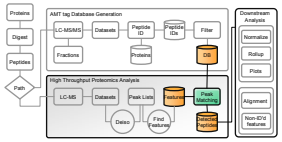
c_1	c_2	c_3	c_4
XCorr	DelCN	SpRank	d_M
4.015	0.325	1	0.14

Parameters Used

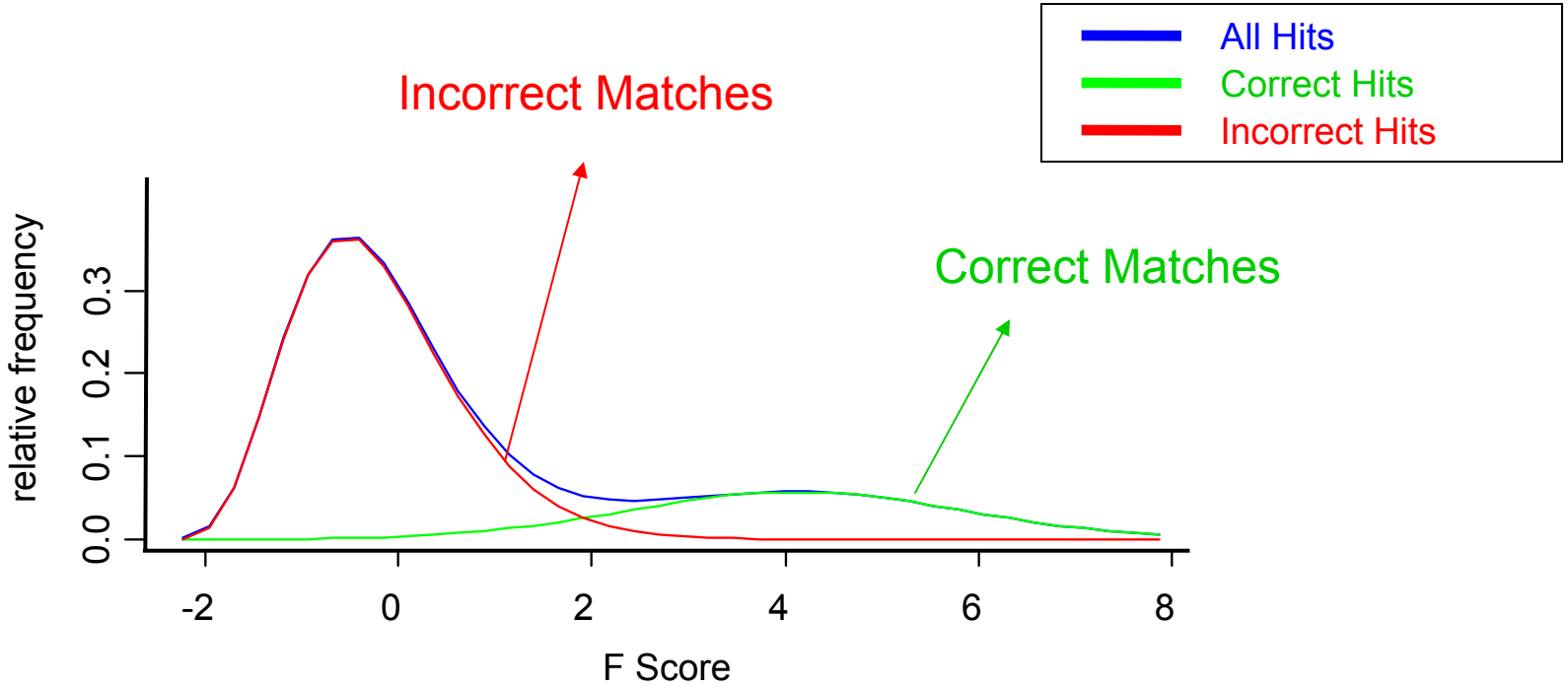
$$F(x_1, x_2, \ldots, x_s) = c_0 + \sum_{i=1}^{s} c_i x_i$$

A. Keller, A.I. Nesvizhskii, E. Kolker, R. Aebersold, "Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search," *Analytical Chemistry* **2002**, *74 (20)*, 5383-5392.

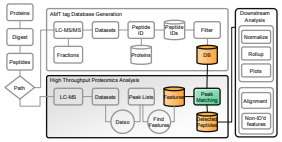
Peptide Prophet F-Score Distributions



- Overall F-Score distribution is bimodal, with distinct distributions for correct and incorrect matches
- Probability that an identification is correct can be computed from relative probabilities of coming from correct or incorrect distribution

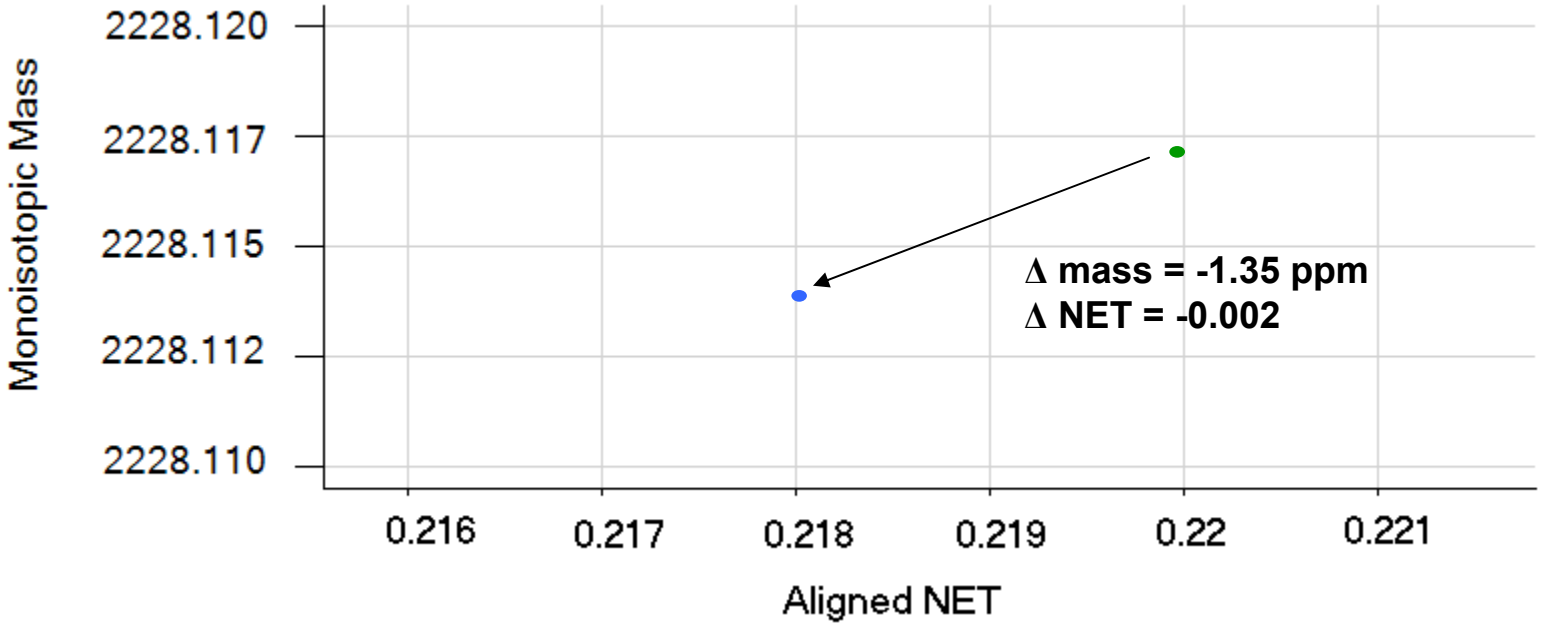


Metrics Associated with a Candidate Identification from AMT tag Pipeline

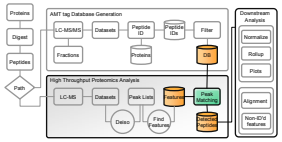


- Each match between an LC-MS feature and a peptide AMT tag is described by a mass error and an LC NET error, plus metrics related to the MS/MS spectra that identified to the AMT tag

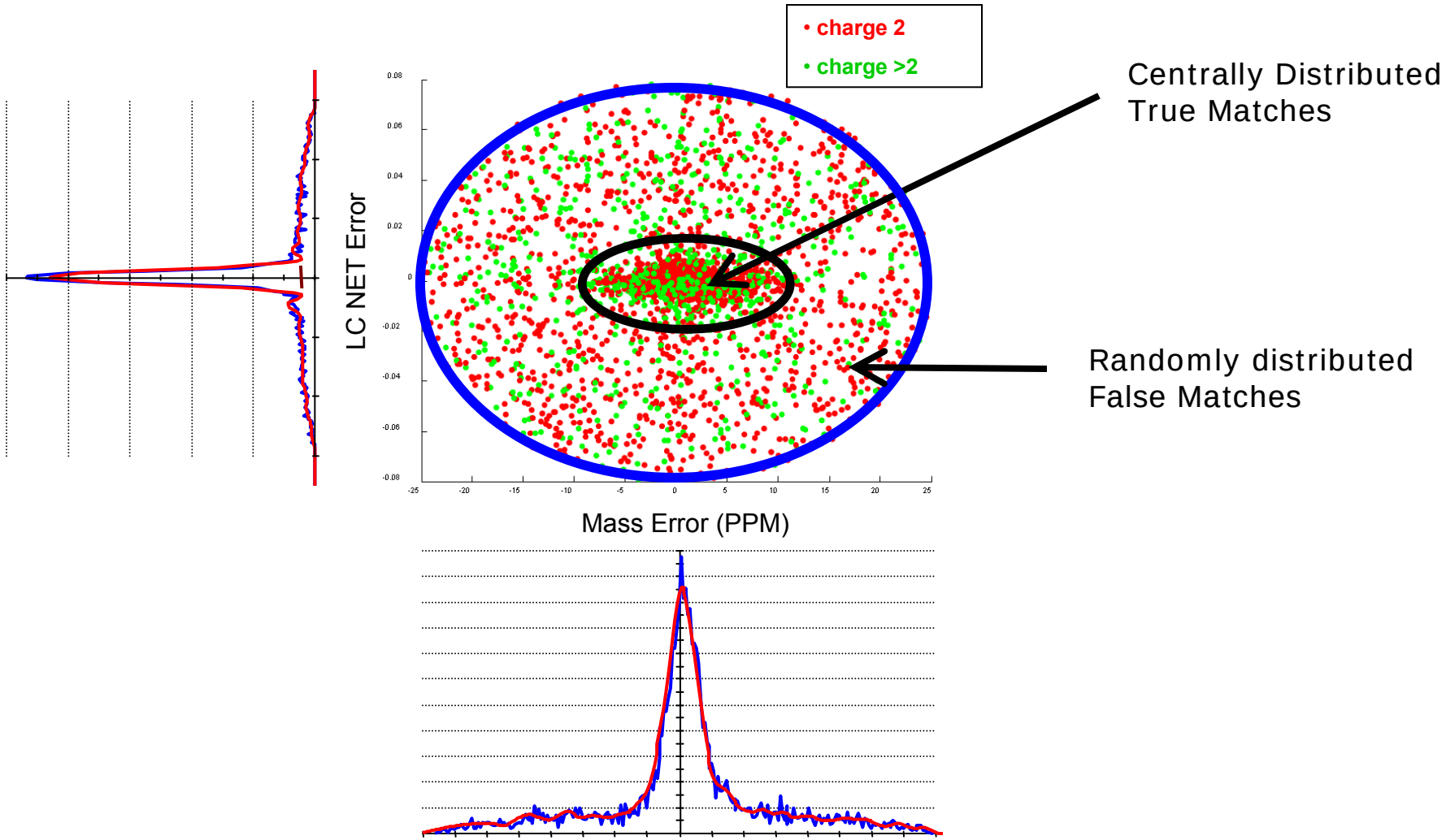
Mass	Scan	Aligned NET	Peptide	NET	Mass	Discriminant Score (Peptide Prophet)
2228.114	1097	0.218	TETQEKNPLPSKETIEQEK	0.220	2228.117	3.1



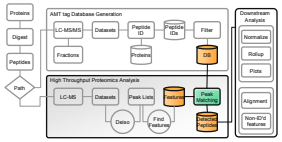
Distribution of Mass and LC Normalized Elution Time (NET) differences



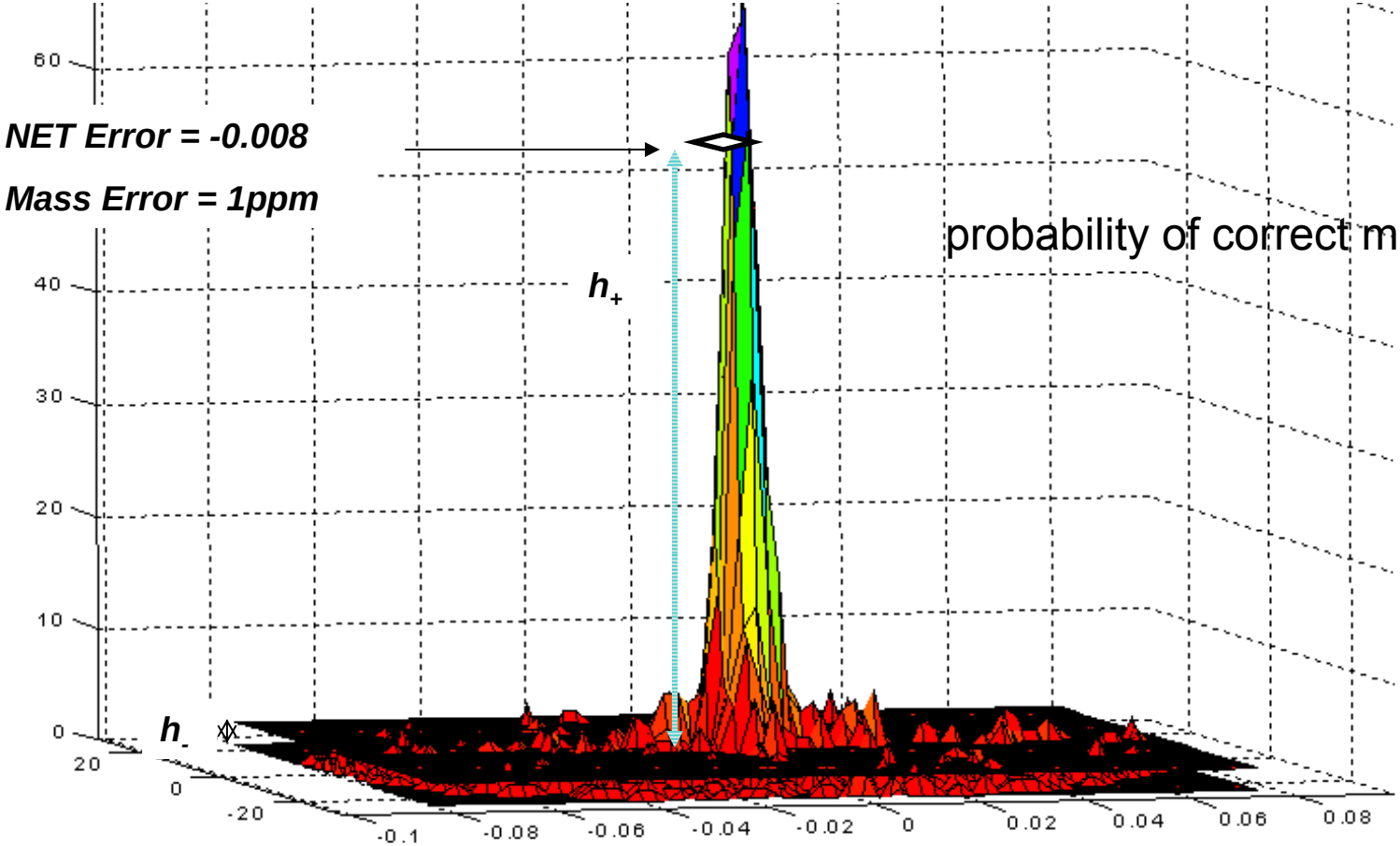
- True and false matches to LC-MS features show different mass and LC NET error distributions



Estimating the Probability a Match is Correct



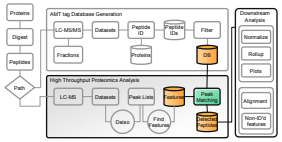
- The probability that a match is correct depends on where its mass and LC NET error value lies on the two-dimensional distribution



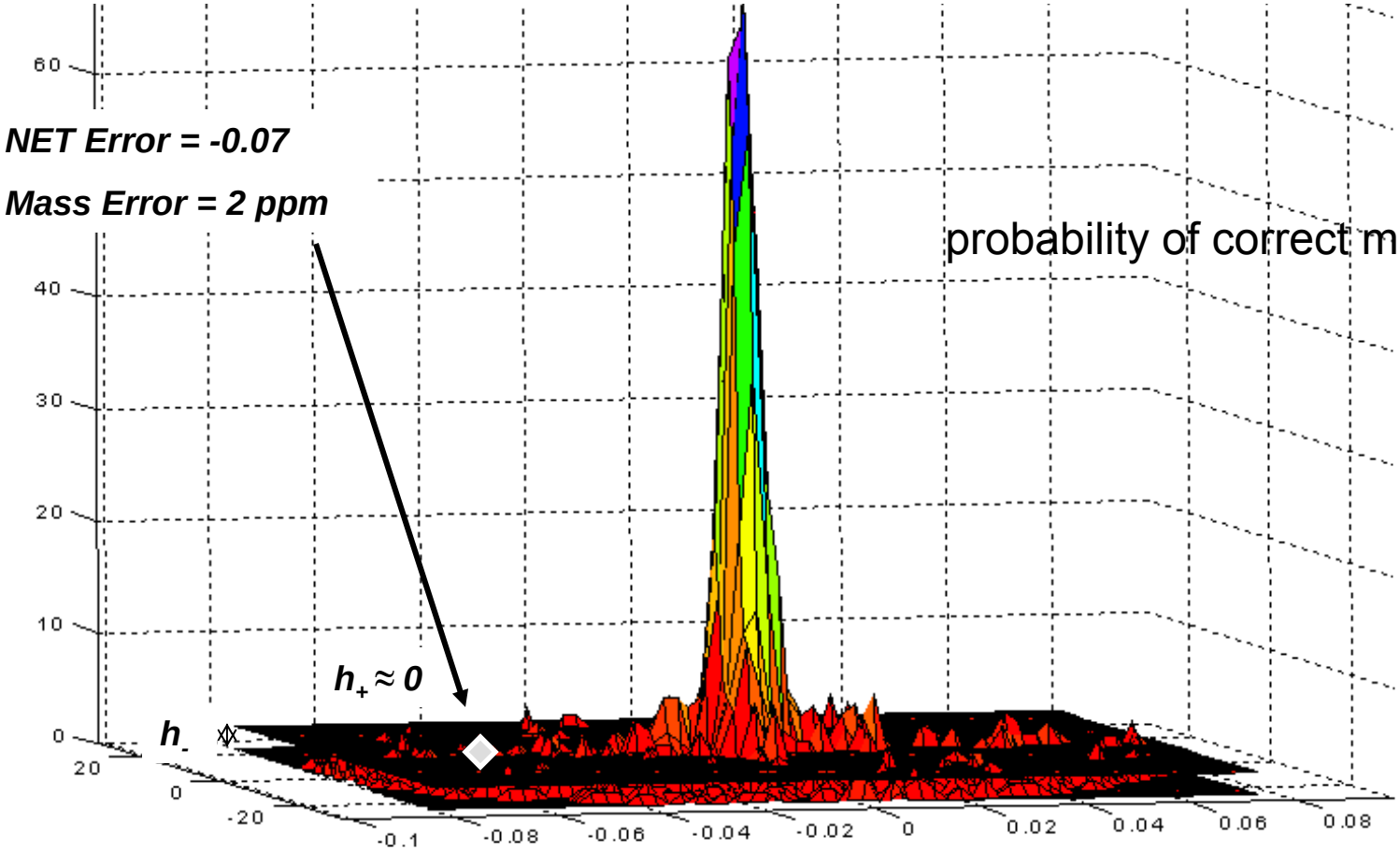
$$\text{probability of correct match} = \frac{52}{52 + 2}$$

$$\text{probability of correct match} = \frac{\text{height of central part}}{\text{height of central part} + \text{height of random part}} = \frac{h_+}{h_+ + h_-}$$

Estimating the Probability a Match is Correct



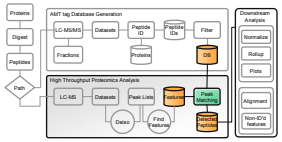
- The probability that a match is correct depends on where its mass and LC NET error value lies on the two-dimensional distribution



$$\text{probability of correct match} = \frac{0}{0+2}$$

$$\text{probability of correct match} = \frac{\text{height of central part}}{\text{height of central part} + \text{height of random part}} = \frac{h_+}{h_+ + h_-}$$

Statistical *M*ethod for *A*ssignment of *R*elative *T*ruth (*SMART*) Score

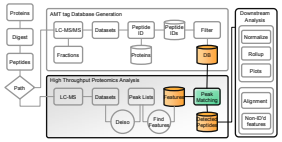


- A *SMART* score estimates the probability that an LC-MS feature has been identified by an AMT tag
  - Combines the mass and LC NET error between the LC-MS feature and the AMT tag with the probability that the MS/MS identification in the AMT tag DB is correct
  - Assumes independence of mass error, NET error and F-score

$$p(+_{match} \mid \delta m, \delta net, Fscore, peptide) = \frac{p(\delta m, \delta net \mid +_{match})\, p(Fscore \mid +_{peptide})\, p(+_{peptide})}{p(\delta m, \delta net \mid +_{match})\, p(Fscore \mid +_{peptide})\, p(+_{peptide}) + p(\delta m, \delta net \mid random_{match})\, p(Fscore \mid random_{match})\, p(random_{match})}$$

Algorithm developed by Navdeep Jaitly

Data Model and Model Fitting



- Data to Model:
  - Mass Errors, NET Errors, F-Score distribution
  - Expectation maximization algorithm used to find optimal parameters for the distributions

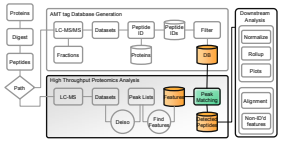
Error Type	Distribution for Model Fitting
Real Mass Errors	ppm errors distributed normally (truncated)
Random Mass Errors	Da Errors distributed normally (truncated)
Real NET Errors	Distributed normally
Random NET Errors	Uniform Background
Real F Scores	F-Scores normally distributed as a function of mass
Random F Scores	Gamma distribution

$$p(+_{match} \mid \delta m, \delta net, Fscore, peptide) = \frac{p(\delta m, \delta net \mid +_{match})\, p(Fscore \mid +_{peptide})\, p(+_{peptide})}{p(\delta m, \delta net \mid +_{match})\, p(Fscore \mid +_{peptide})\, p(+_{peptide}) + p(\delta m, \delta net \mid random_{match})\, p(Fscore \mid random_{match})\, p(random_{match})}$$

Algorithm developed by Navdeep Jaitly

More info

Data Model Example: Mass Error Distribution

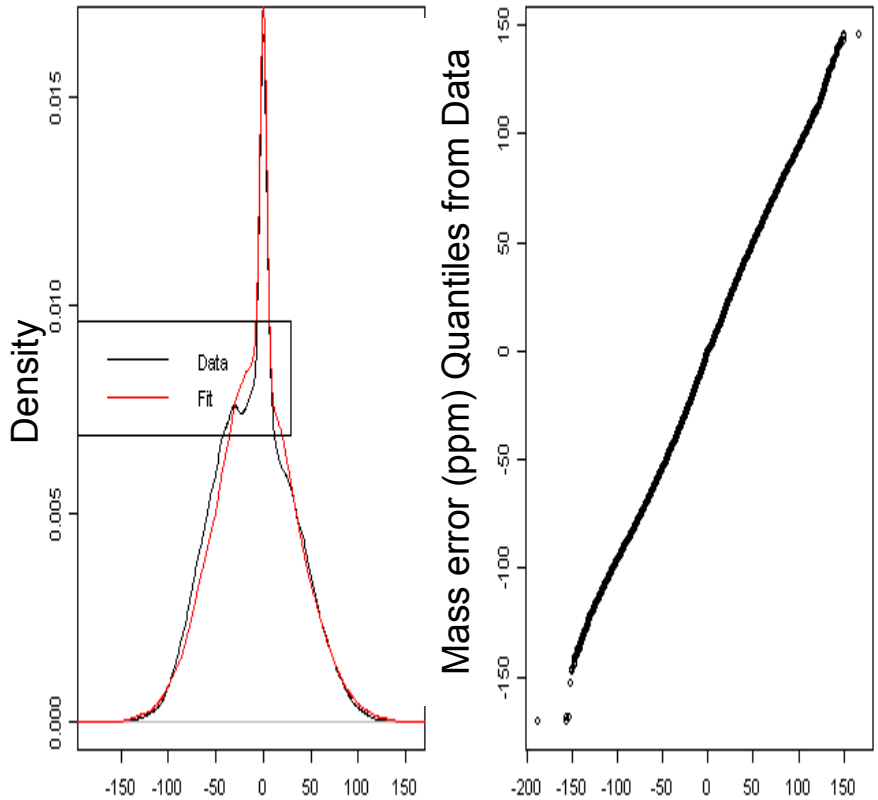


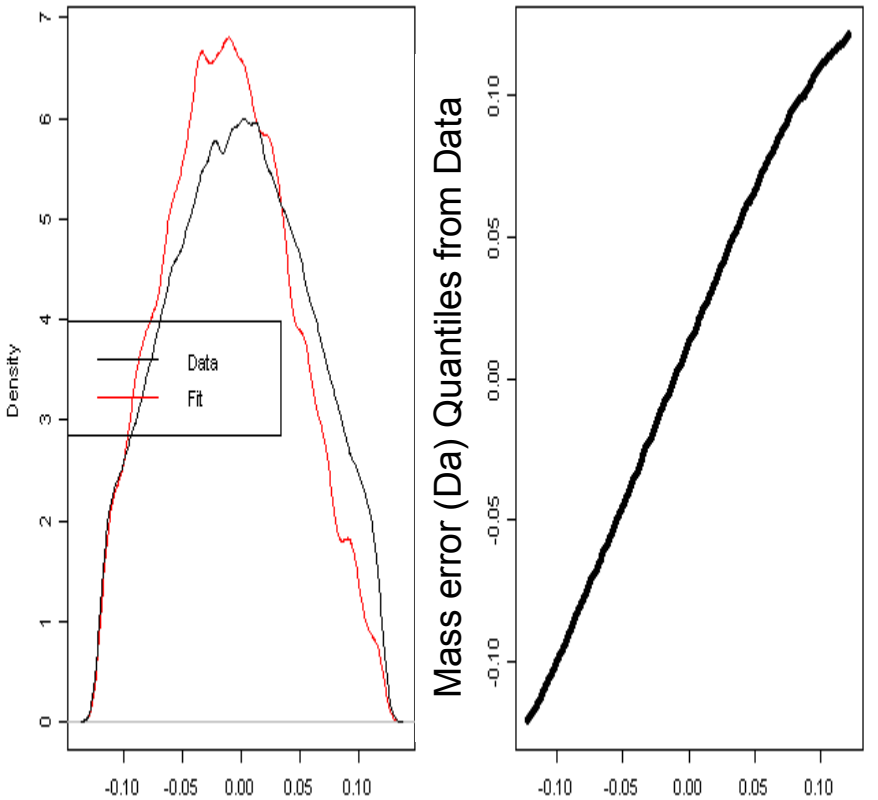
Regular Matches

(Real and Random Distributions)

Decoy (11 Da Shifted) Matches

(Random Distribution)



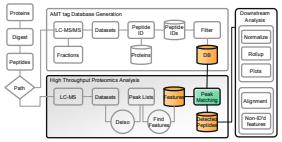


Salmonella Typhimurium protein extract sample

N. Jaitly, J.N. Adkins, and R.D. Smith, preliminary results.

More info

Data Model Example: NET Error Distribution

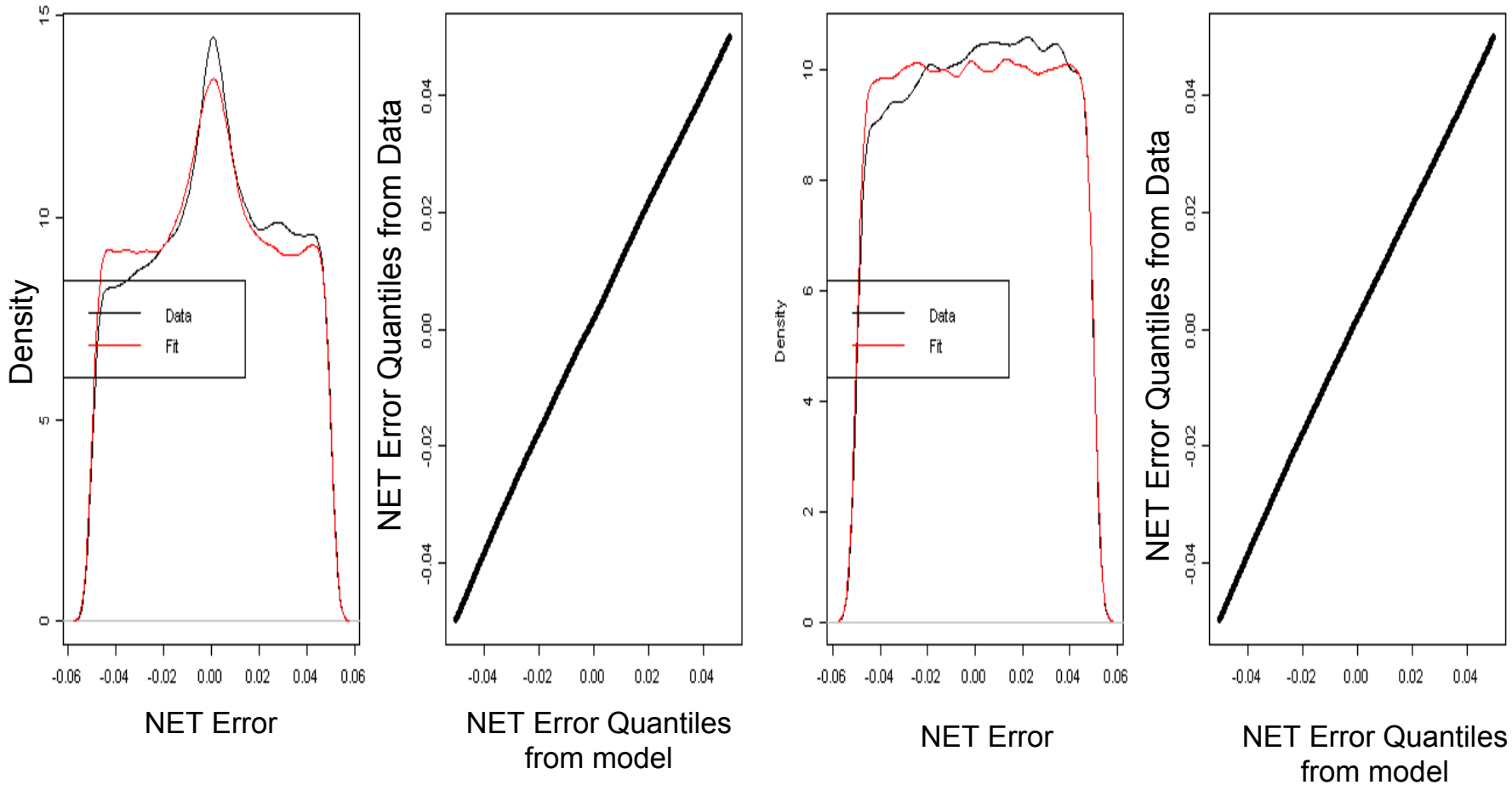


Regular Matches

(Real and Random Distributions)

Decoy (11 Da Shifted) Matches

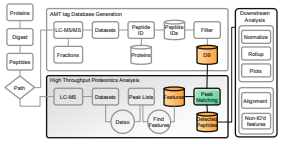
(Random Distribution)



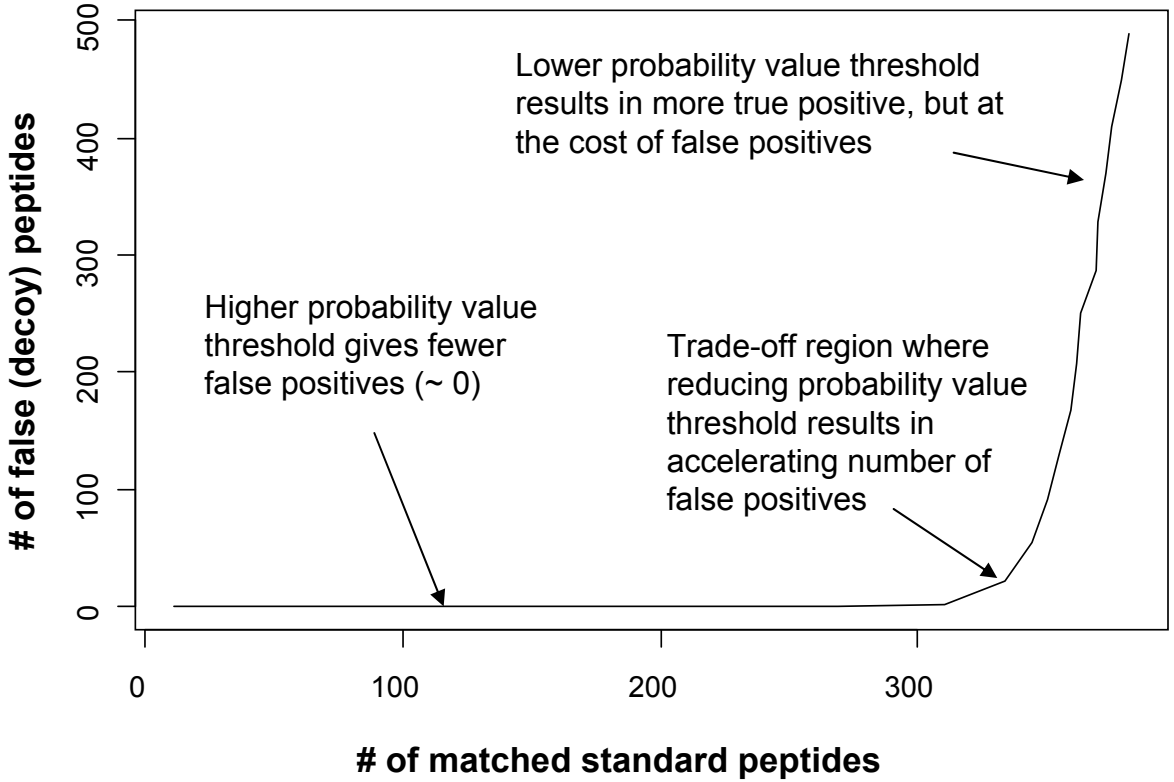
Salmonella Typhimurium protein extract sample

N. Jaitly, J.N. Adkins, and R.D. Smith, preliminary results.

Performance Curves



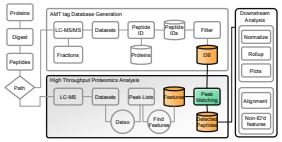
- Match LC-MS features to AMT tags and compute the probability that the identified feature is indeed the MS/MS derived AMT tag
- Accept matches whose computed probabilities are greater than a specified threshold
- Balance sensitivity and specificity by changing score thresholds



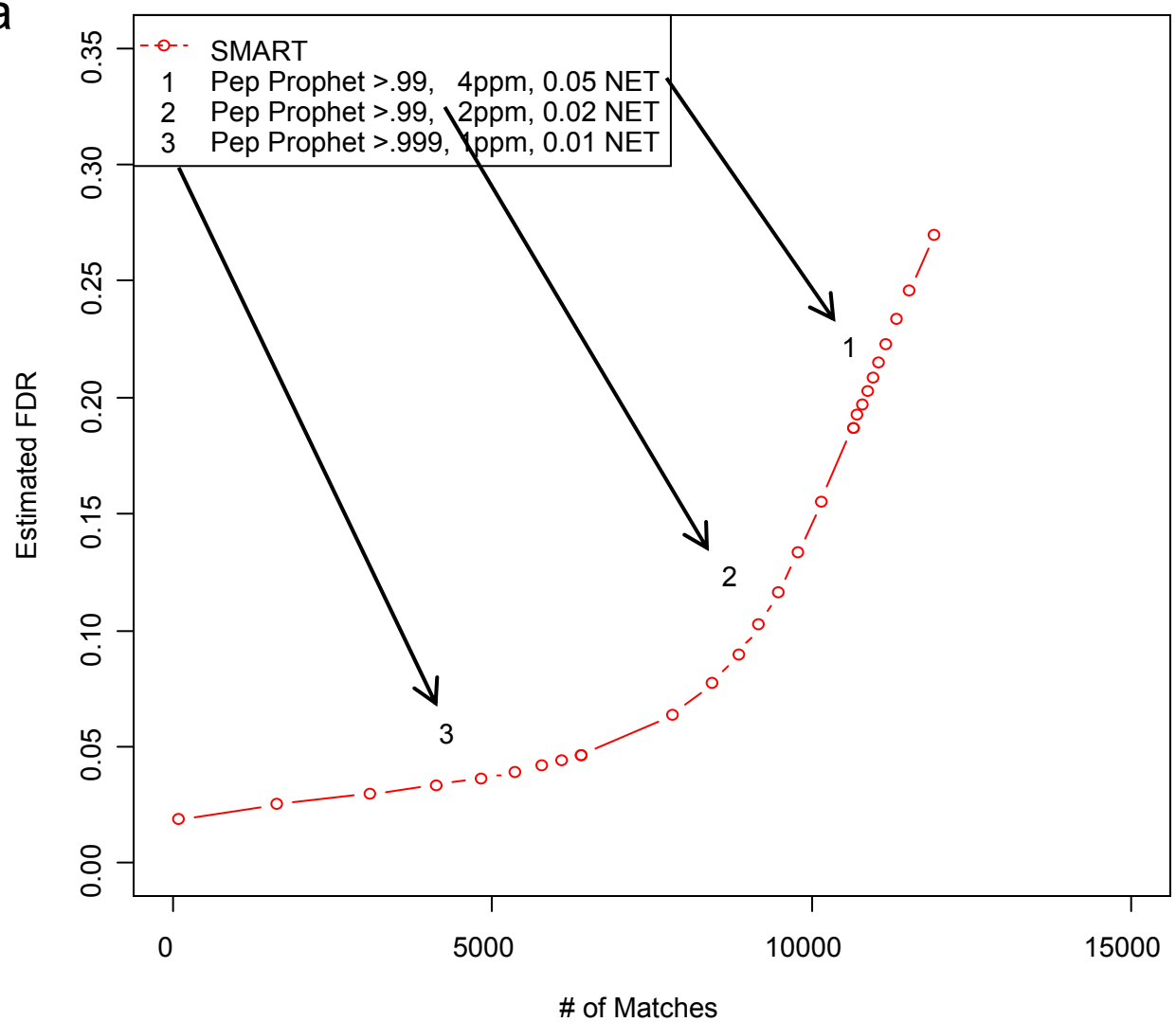
- QC sample matched to database with 8000 decoy proteins
- High number of true positives achieved with very few false positives

N. Jaitly, J.N. Adkins, and R.D. Smith, preliminary results.

Performance Curves

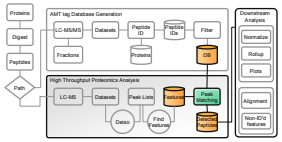


- *Salmonella* Typhimurium protein extract sample compared to typical criteria

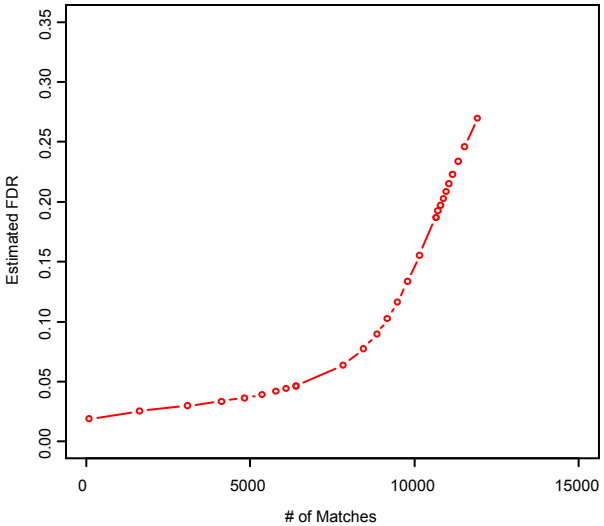


N. Jaitly, J.N. Adkins, and R.D. Smith, preliminary results.

Summary



- The Statistical Method for Assignment of Relative Truth (SMART) score is a model to estimate confidence of peak matching
  - SMART provides a measure to prioritize acceptable matches using one number, by defining a probability score combining disparate information
  - SMART allows calculation of FDR for identifications and estimates the tradeoff between false negatives and false positives
- Evaluation shows good correlation with observed number of correct answers



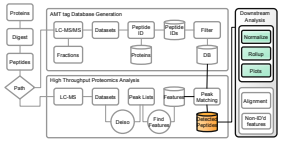
Part II: LC-MS Feature Discovery

- Part I: Introduction and Overview of Label-Free Quantitative Proteomics (Anderson)
- Part II: Feature Discovery in LC-MS Datasets (Monroe and Polpitiya)
  - ✓ Structure of LC-MS Data
  - ✓ Feature discovery in individual spectra (deisotoping)
  - ✓ Feature definition over elution time
  - ✓ Identifying LC-MS Features using an AMT tag DB
  - ✓ Extending the AMT tag approach for feature based analyses
  - ✓ Estimating confidence of identified LC-MS features
  - Downstream quantitative analysis with DAnTE
- Break
- Part III: Biological Application of the AMT tag Approach (Ansong)
- AMT tag Analysis Software Demo
- Panel Discussion

Downstream Data Analysis

- Quantitative protein inference from peptide data
- Complications
  - Multiple, possibly inconsistent peptide measurements for same protein
  - Systematic abundance variation within and between conditions
    - How should we use information from blocking and randomization of experiments?
  - High rate of missingness in peptide measurements
- Need to combine off the shelf statistical methods and novel solutions
  - Clustering
  - ANOVA
  - PCA

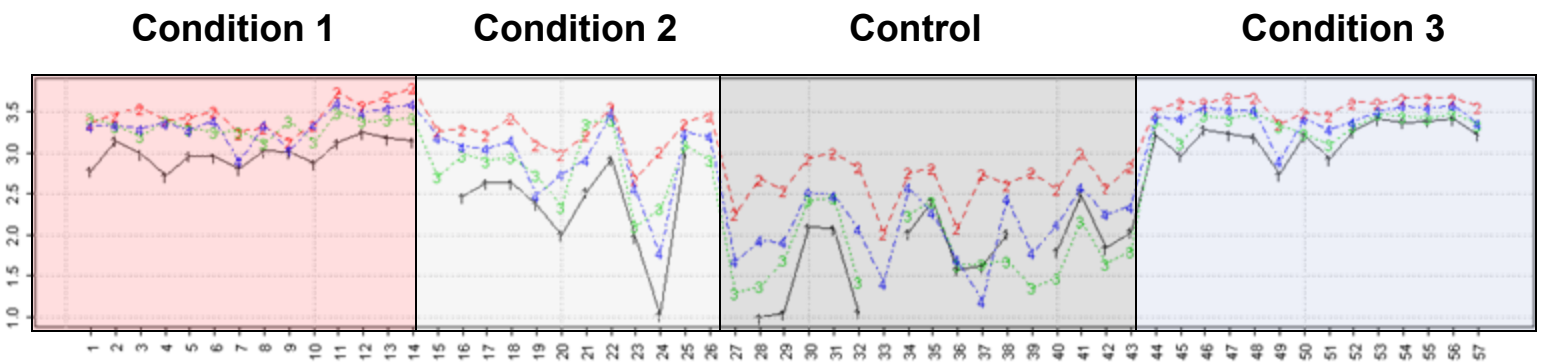
Infer Protein Abundances from Peptide Abundances



- Multiple peptides observed for each protein
  - For example, protein with 4 peptides

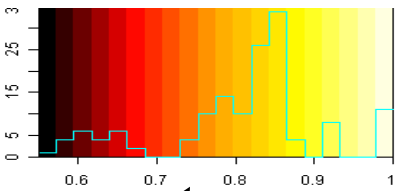
    1. SADLNVDSIISYWK
    2. LLLTSTGAGIIDVIK
    3. LIVGFPAYGHTFILSDPSK
    4. IPELSQSLDYIQVMTYDLHDPK

  - Plot peptide abundance across all datasets (for 4 conditions)



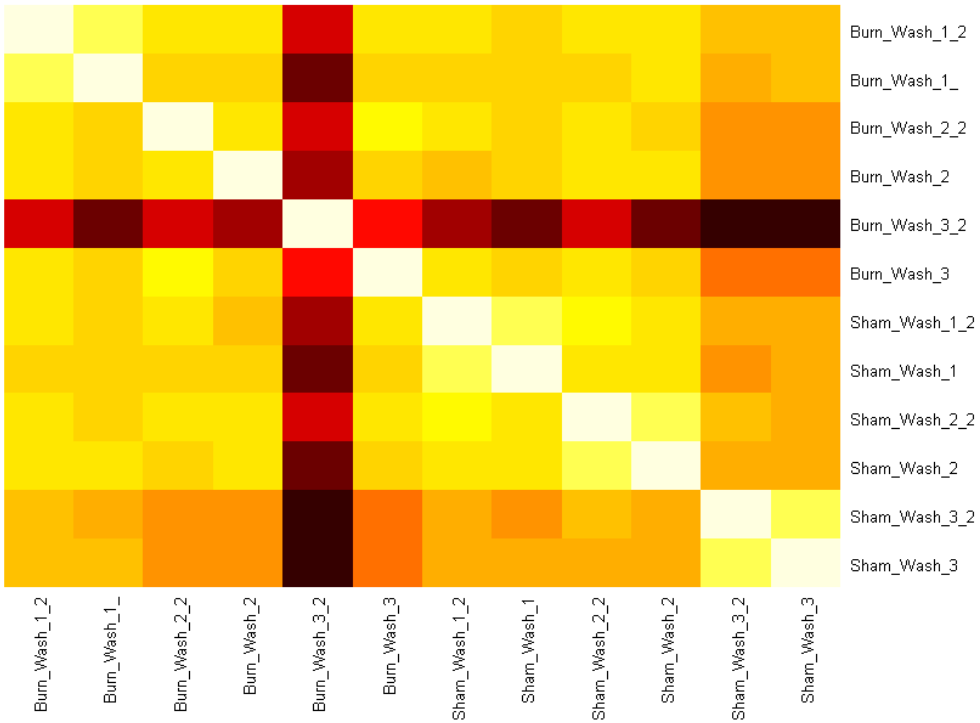
→ Outlier detection and normalization need to be performed before meaningful abundance information can be inferred

Outlier Detection

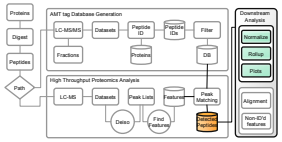


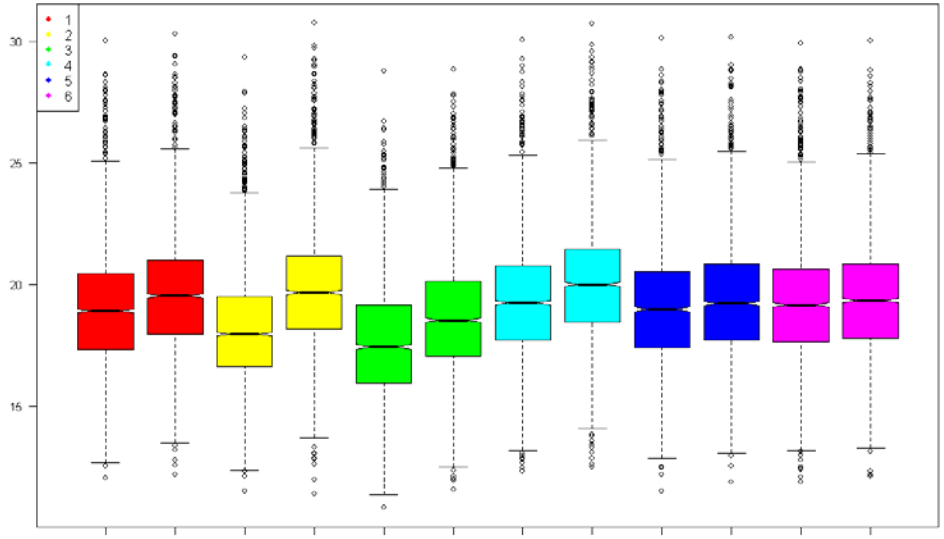
Color legend with overlaid histogram of correlation values

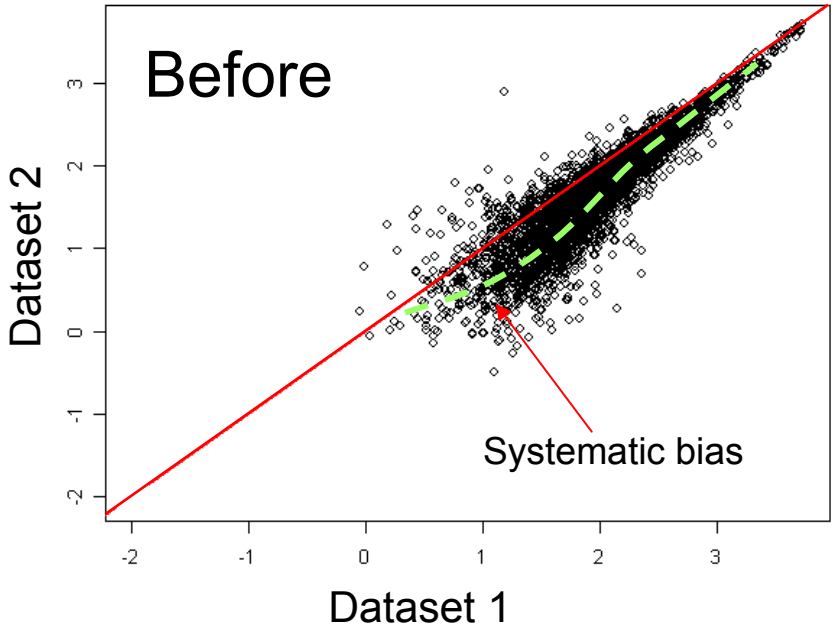
Outlier dataset

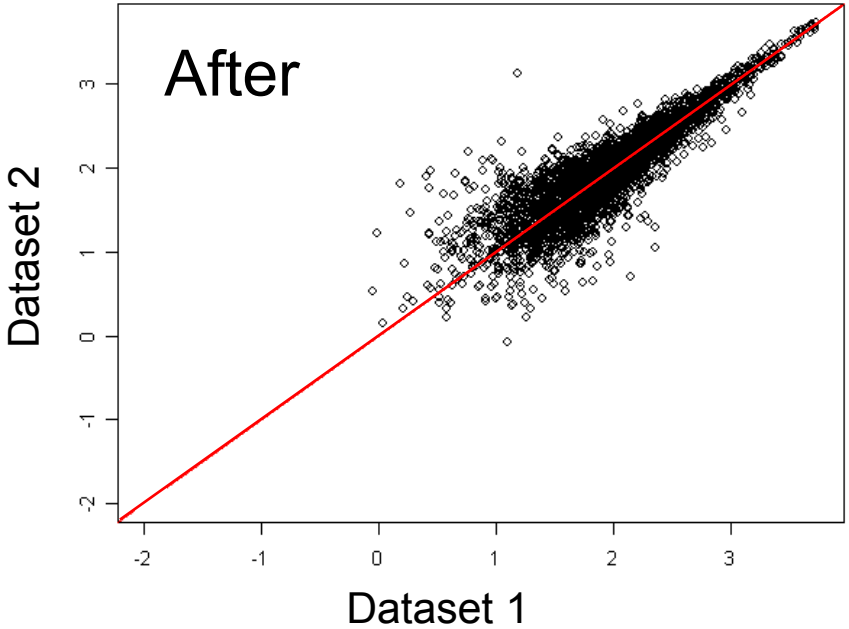


Normalization

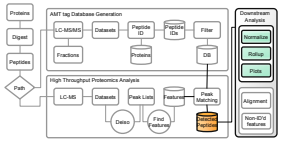




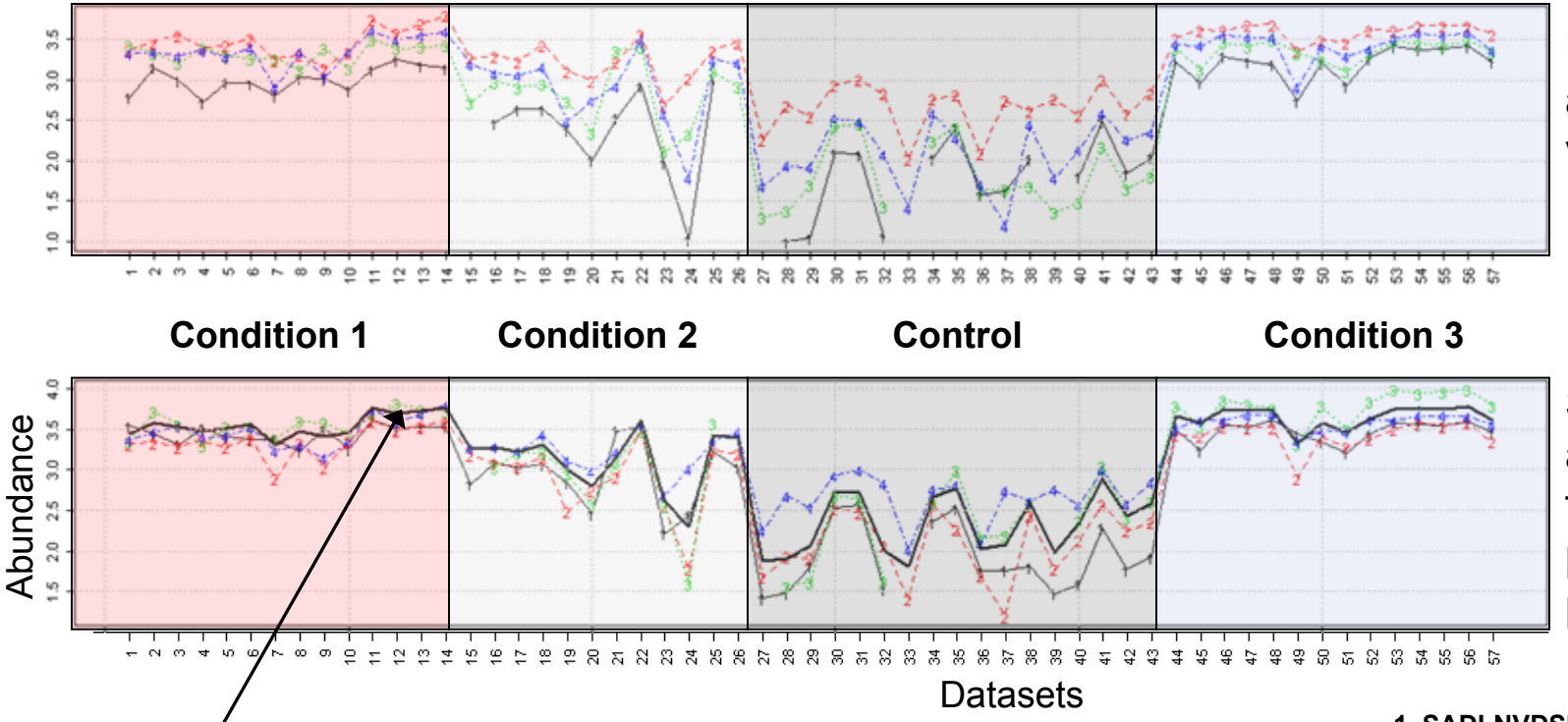




Infer Protein Abundances from Peptide Abundances



- Scale peptide abundances to an automatically chosen "optimal" reference peptide for each protein
- Estimate relative protein abundance using scaled peptides



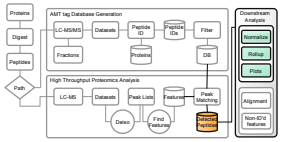
Raw peptide abundances vs. dataset (for 1 protein)

Scaled peptide abundances for this protein's 4 peptides

Median protein abundance (dark black line)

1. SADLNVDSIISYWK
2. LLLTSTGAGIIDVIK
3. LIVGFPAYGHTFILSDPSK
4. IPELSQSLDYIQVMTYDLHDPK

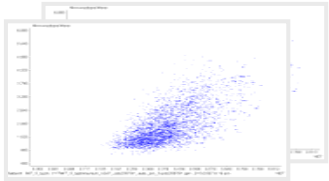
DAnTE: Data Analysis Tool Extension

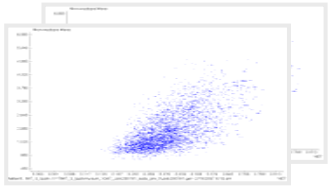


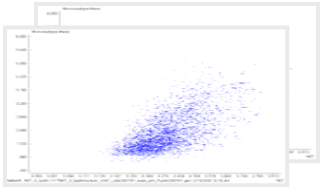
- Statistical tool designed to address analysis needs associated with quantitative bottom-up, proteomics data.

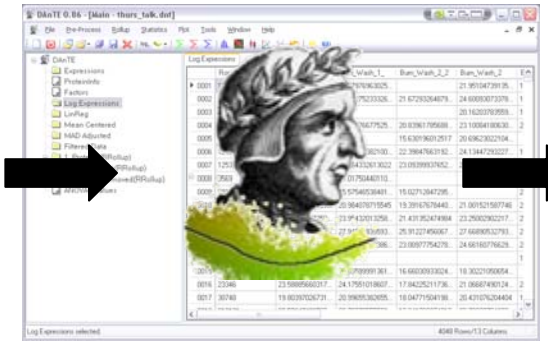
Capture experimental design through factors:

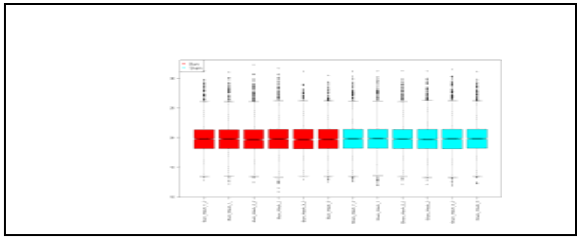
- Biological conditions
- Biological replicates
- Technical replicates

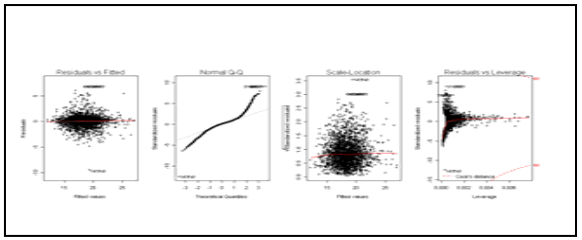


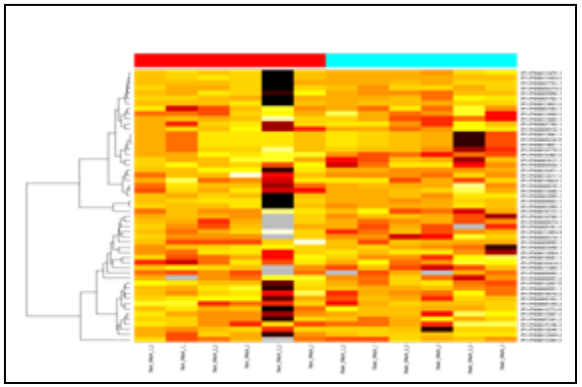




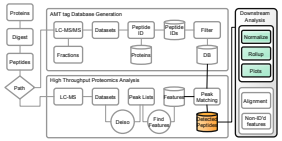






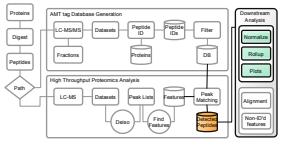


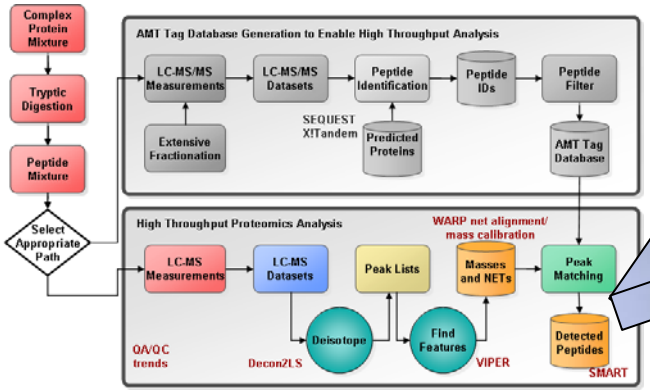
Downstream Data Analysis



- Challenge:
  - With thousands of peptides/proteins in hundreds of samples, how does one mine the data for <u>significant</u> trends of interest?
- General Analysis Steps in DAnTE:
  - Outliers: identify bad datasets
  - Normalization: remove any <u>systematic variations</u> due to instrument/sample processing etc.
  - “Rollup”: infer protein abundances from peptide abundances
  - ANOVA: look for statistically <u>significant</u> features
  - Cluster: explore <u>trends/patterns</u>

Where does DAnTE fit in?





AMT tag pipeline

Multialign

VIPER

Tabular Data
(raw abundances, exp. ratio values, spectral counts)

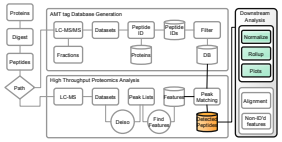
DAnTE

Biological Interpretations

Ref ID	Reference	Mass Tag ID	AB_007-LP	AB_007-LP	AB_007-LP	AB-008-LP	AB-008-LP
57559	gi|10048460|ref|NP	32275105		0.0133541			
57559	gi|10048460|ref|NP	47465602	0.214036	0.110191	0.088877	0.144964	
57559	gi|10048460|ref|NP	83201030					
57559	gi|10048460|ref|NP	83257895					
57559	gi|10048460|ref|NP	83424583	0.126785	0.0901864	0.05985	0.0391025	0.0275993
57559	gi|10048460|ref|NP	83451097	0.0600628			0.0392326	0.0274422
57559	gi|10048460|ref|NP	83451190	0.0409746	0.0396703	0.0539142	0.0591157	0.0246261
57559	gi|10048460|ref|NP	83479064	0.0383742	0.0865723	0.0402673	0.0359534	0.0397751
57559	gi|10048460|ref|NP	83479231	0.0212362	0.0122634	0.0184865		
57559	gi|10048460|ref|NP	83514326		0.0289326	0.0337594	0.0237968	0.0243036
74732	gi|10092608|ref|NP	7824413			0.0851353		
74732	gi|10092608|ref|NP	8680795					
74732	gi|10092608|ref|NP	20746162	0.536631	0.315447	0.231846	0.131434	0.130482
74732	gi|10092608|ref|NP	20750908	0.0699752	0.0821782	0.0417777	0.0280187	0.0366826
74732	gi|10092608|ref|NP	20750955	0.16129	0.123591	0.12117	0.126886	0.0970479
74732	gi|10092608|ref|NP	20956112	0.0407675				
74732	gi|10092608|ref|NP	20985367		0.0144197	0.0138798		0.0125844
60259	gi|10181140|ref|NP	7777112		0.0194588			

Proteins, peptides and peptide abundances

DAnTE supports a wide array of interactive analyses



Data Loading
- Peptide abundance
- Peptide-Protein relations
- Factors
- Spectral counts

Variance Stabilization
- log2 or log10
- Bias (additive/multiplicative)

Investigative Plots
- Histograms
- Boxplots
- Correlation diagrams
- MA Plots

Normalization (within a factor)
- Linear Regression
- Local regression (LOESS)
- Quantile

Normalization (across factors)
- Central tendency
- Median absolute Deviation (MAD)

Infer Proteins from Peptides
- RRollup
- ZRollup
- QRollup
- Rollup Plots

Statistical Tests
- ANOVA/Non-parametric
- Mix Models

Visualization
- PCA / PLS
- Heatmaps (hierachical, kmeans)

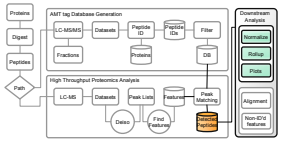
Impute Missing Data
- KNNimpute
- SVDimpute
- Other

Other Features
- Filter ANOVA results
- Save session

Easily extendable to plugin more modules

Statistical Environment

Example Dataset

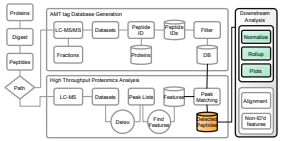


- Proteomics analysis of two *Salmonella* regulatory knock-out mutants, **Hfq** and **SmpB**

- Relevant factor is mutant type
  - Wild type (WT)
  - hfq knock-out mutant
  - smpB knock-out mutant

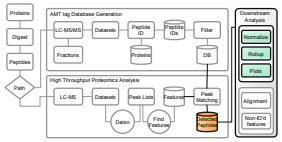
- Three biological replicates for each mutant

Outline of an Analysis



- Load data
- Log transform
- Diagnostic Plots
- Define factors
- Normalize
  - Within a Factor
    - Linear regression
    - LOESS (LOcal regrESSion) ←
    - Quantile
  - Across Factors
    - MAD
    - Central tendency ←
- Rollup
- ANOVA
- Save the results to a session file (.dnt file)

Load Data



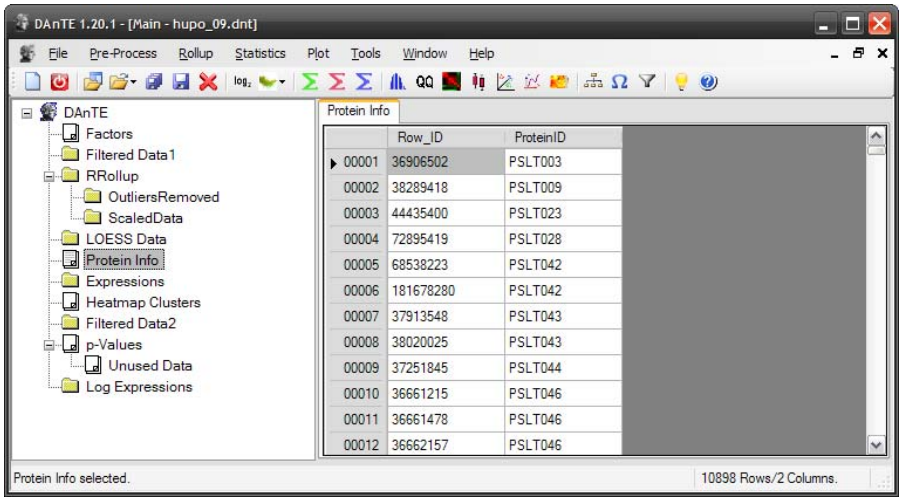
Experiments

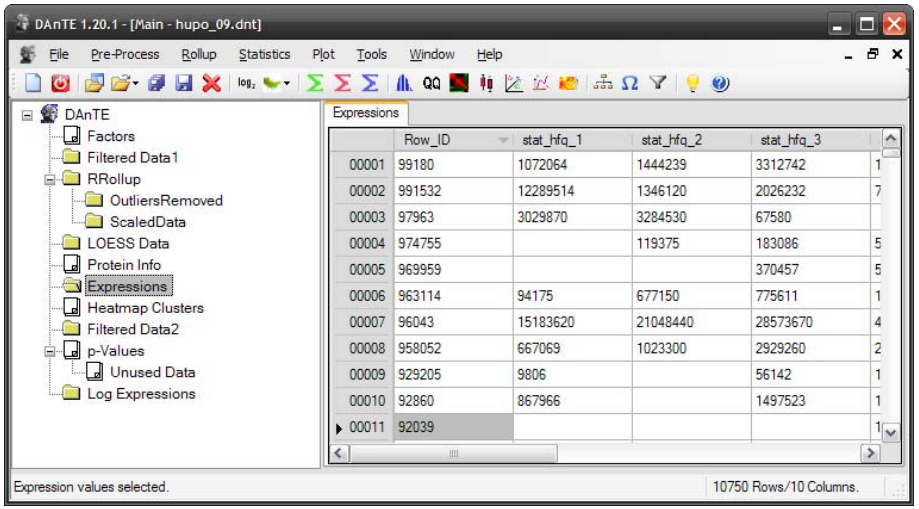
Multialign quantitation data

Reference	MassTagId	stat_hfq_1	stat_hfq_2	stat_hfq_3	stat_smpB_1	stat_smpB_2	stat_smpB_3	stat_wt_1	stat_wt_2	stat_wt_3
PSLT003	36906502						873217	809678	1222380	
PSLT009	38289418		19706	488302	398689		429441	444871	1143506	
PSLT023	44435400					633509	1715887	84851	46112	
PSLT028	72895419	419178							90115	418787
PSLT042	68538223	1882920	3768120							
PSLT042	181678280			7505360				5720460		4313170
PSLT043	37913548			175315				500771		388522
PSLT043	38020025				1674304		2266864			815086
PSLT044	37251845	735419	2995801	1141880						
PSLT046	36661215	1160162	477031	2088928						
PSLT046	36661478	823980	50765	2080703						
PSLT046	36662157	2028274	561722	3338537						
PSLT046	36683548	1157102	487085	677000						
PSLT046	37163770				882143	159208		56647	194201	

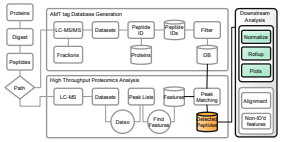
Protein-Peptide relationships

Peptide abundances

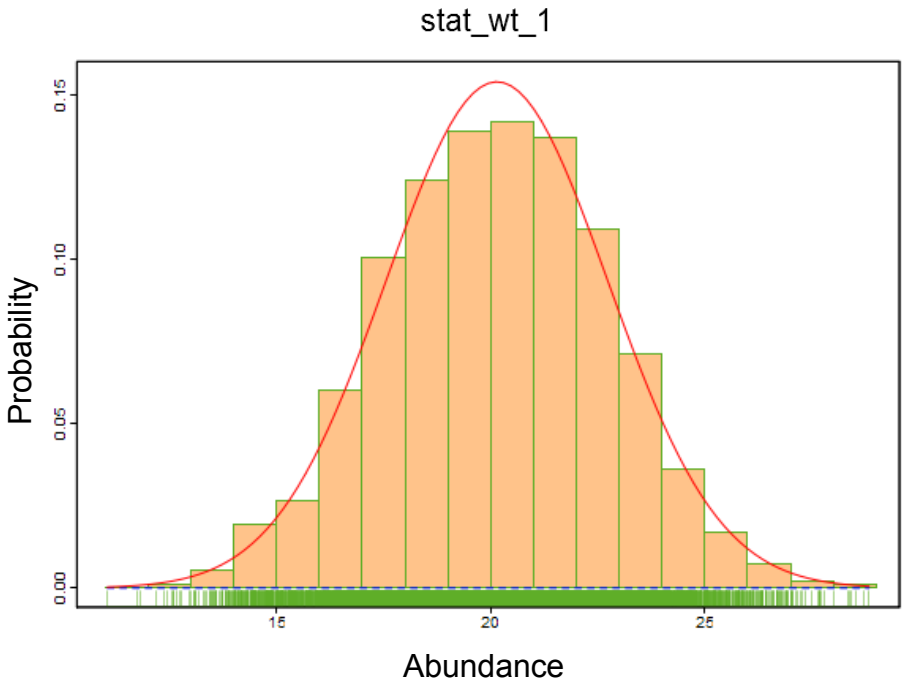


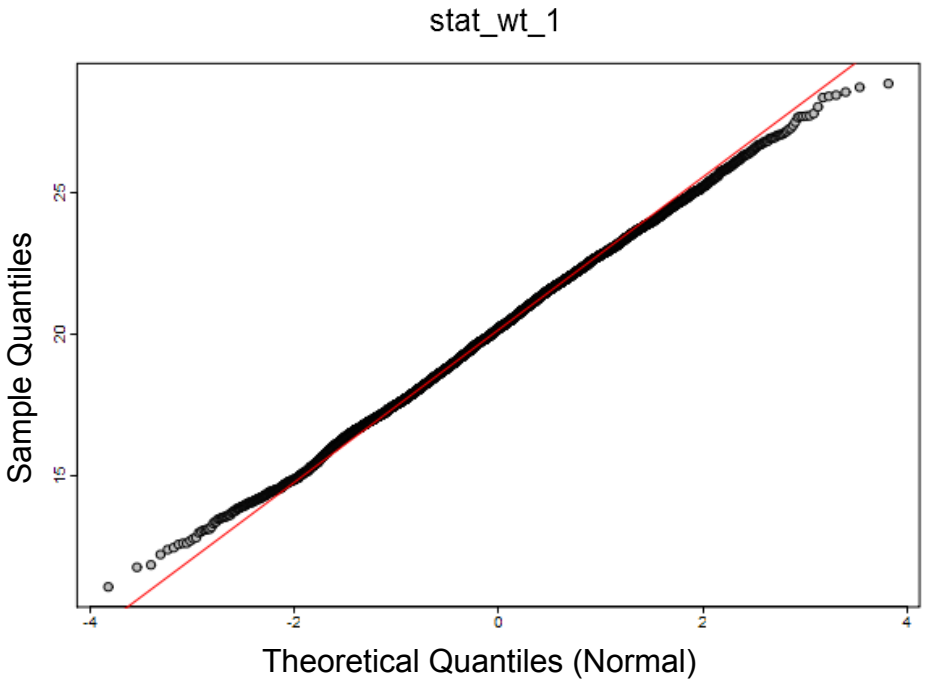


Diagnostic Plots: Normality

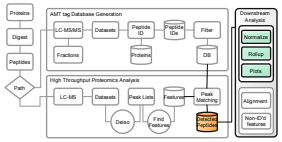


- Distribution of abundance values in a single dataset
  - Should see Normal distribution

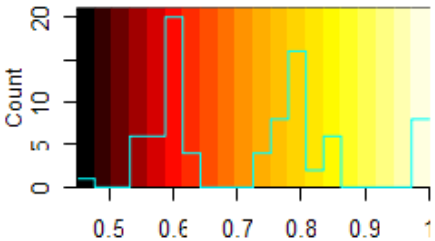




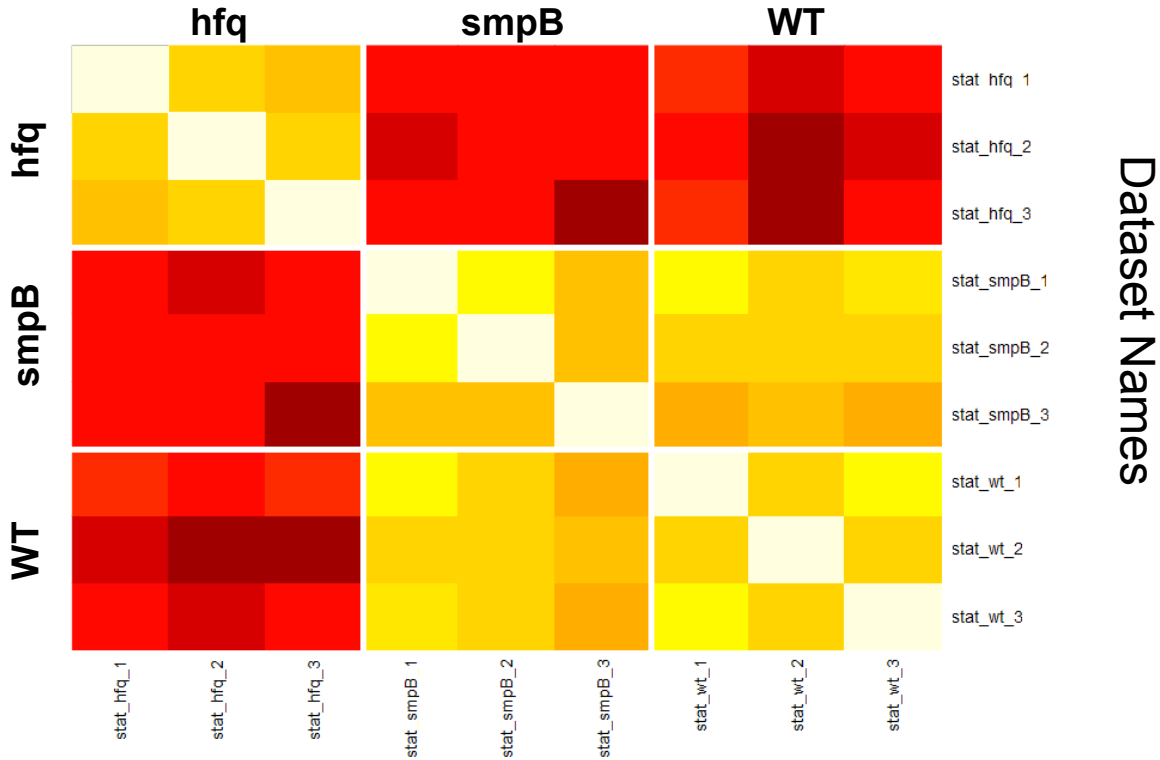
Diagnostic Plots: Correlations



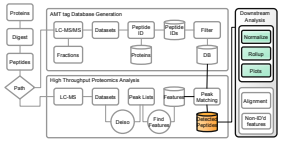
- Assess repeatability of replicate datasets
- Compare biological conditions

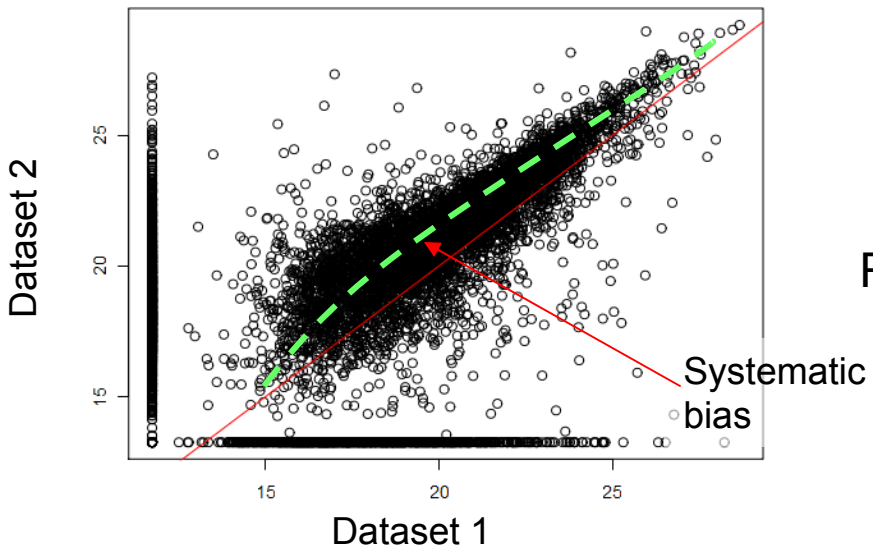


Color legend with overlaid histogram of correlation values

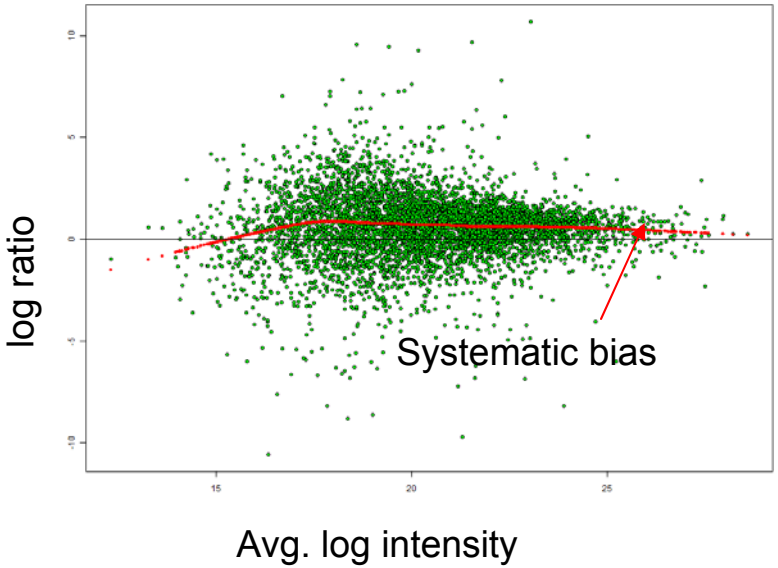


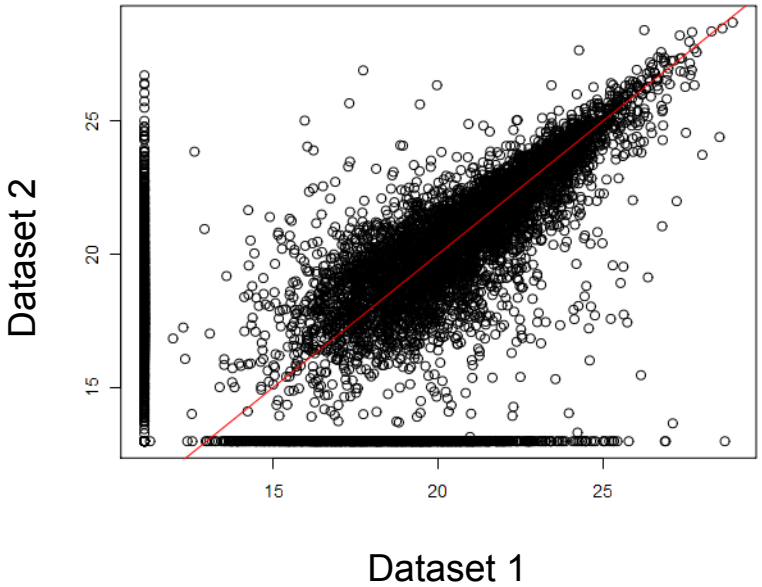
Normalization: LOESS



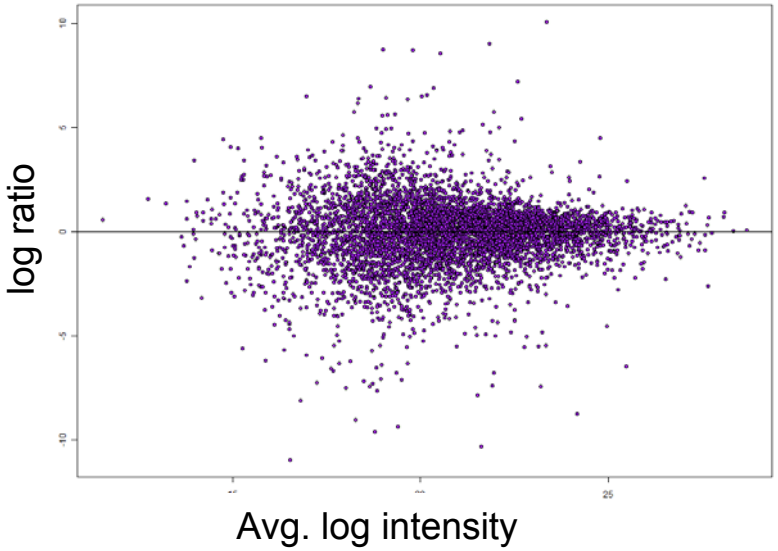


Raw

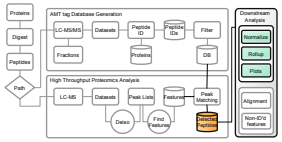




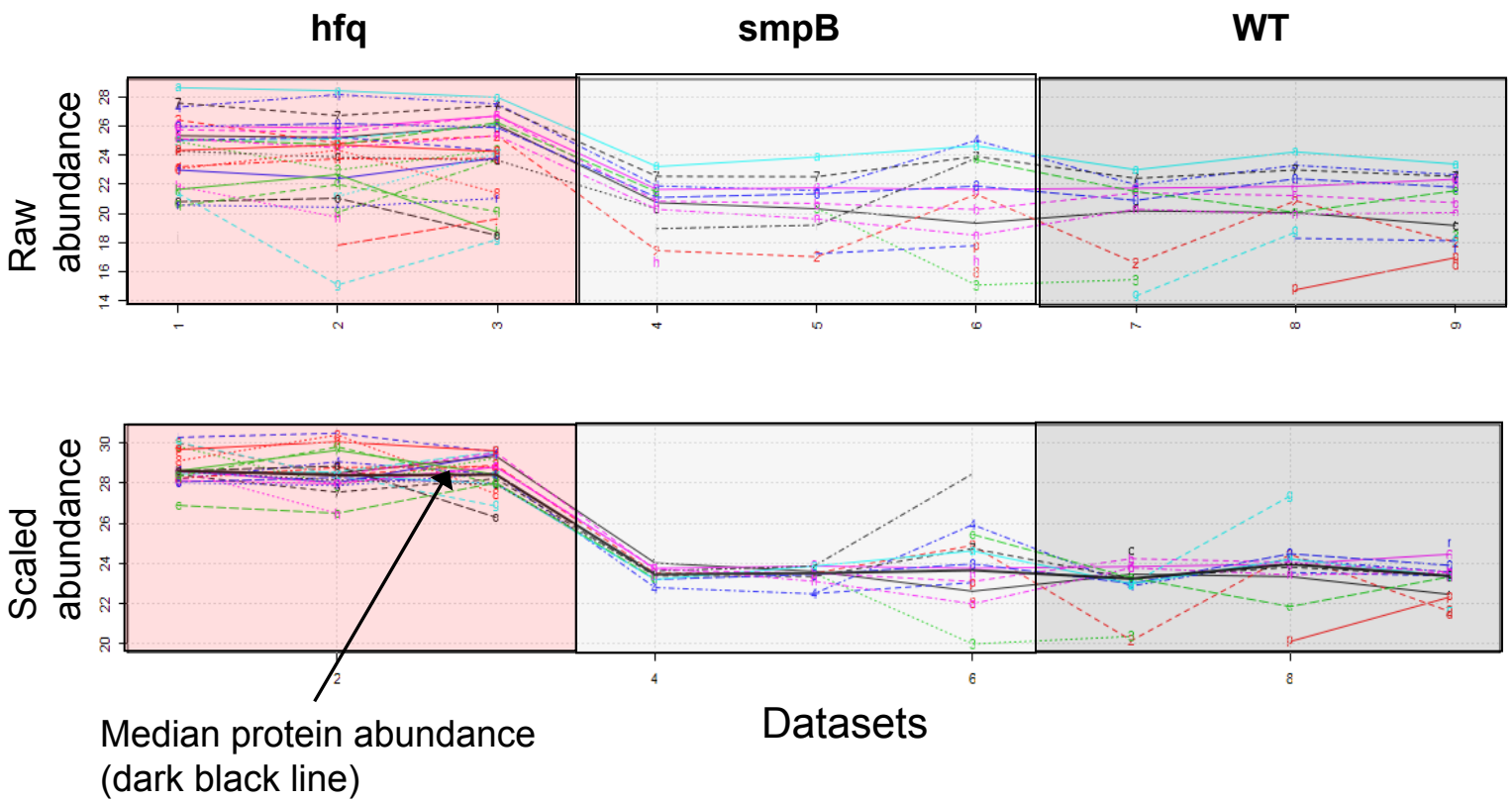
After LOESS



Inferring Protein Abundances from Peptide Abundances



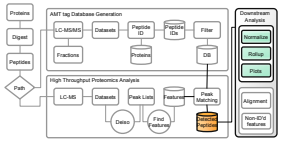
- Three algorithms for rolling up abundances
- Additional algorithms can be added as needed

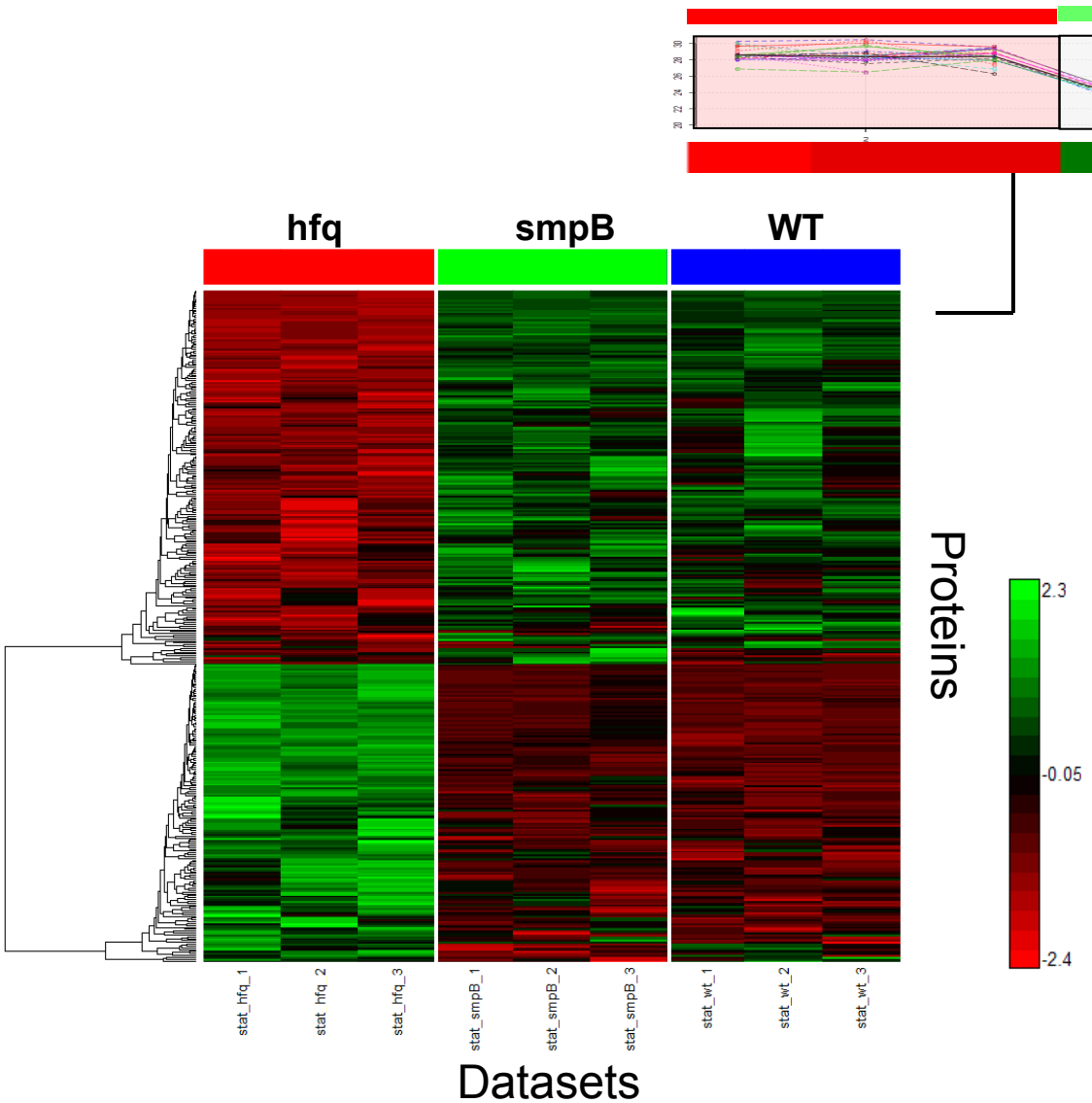


Raw peptide abundances vs. dataset (for 1 protein)

Scaled peptide abundances for this protein's 28 peptides

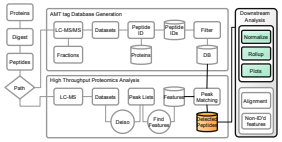
Proteins Showing Significant Changes





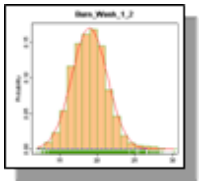
- ANOVA identified 361 proteins with significant abundance changes between the mutants (with an FDR of 5%)
- Hierarchical clustering groups related proteins

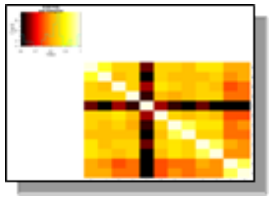
DAnTE Feature List

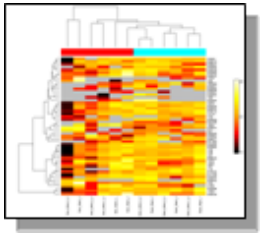


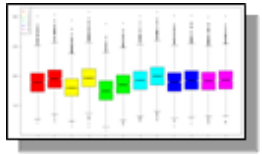
- Data loading with peptide-protein group information
- Spectral Count Analysis
- Log transform
- Factor Definitions
- Normalization
  - Linear Regression
  - Loess
  - Quantile normalization
  - Median Absolute Deviation (MAD) Adj.
  - Mean Centering
- Missing Value Imputation
  - Simple
    - mean/median of the sample
    - Substitute a constant
  - Advance
    - Row mean within a factor
    - kNN method
    - SVDimpute
- Save tables / factors / session

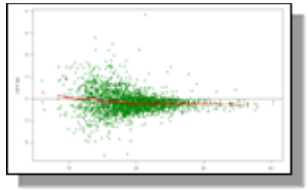
- Plots
  - Histograms
  - QQ plots
  - Boxplots
  - Correlation plots
  - MA plots
  - PCA/PLS plots
  - Protein rollup plots
  - Heatmaps
- Rolling up to Proteins
  - Reference peptide based scaling (RRollup)
  - Z-score averaging (ZRollup)
  - QRollup
- Statistics
  - ANOVA
    - Provisions for unbalanced data
    - Random effects (multi level) models (REML)
  - Normality test (Shapiro-Wilks)
  - Non-parametric methods (Wilcoxon, Kruskal-Walis tests)
  - Q-values
  - Filters

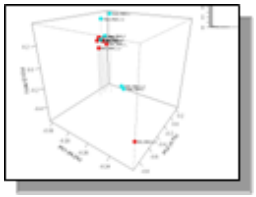




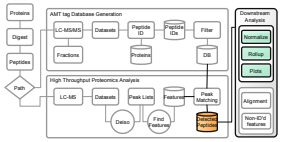






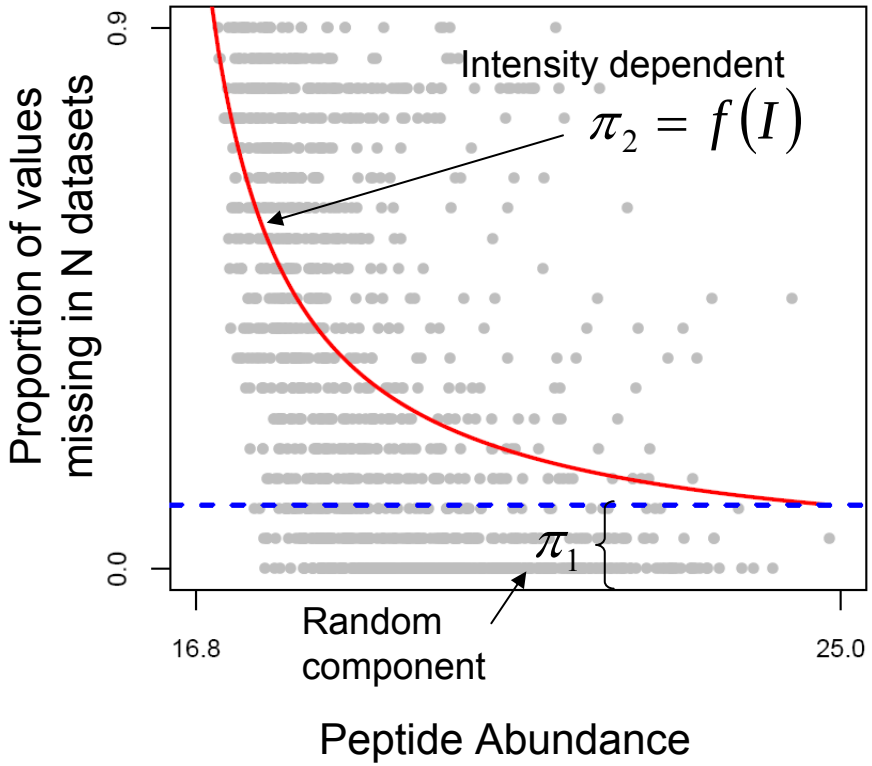


New Significance Test for Proteins with Missing Peptide Abundance Values

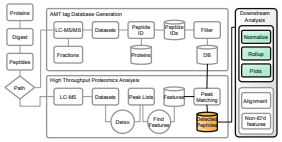


- Missing values in peptides can bias statistical tests for significance (ANOVA)
- Missingness is modeled as a mixture of two probability distributions
  - Too low to detect (left censored)
  - Completely at random
- Use a likelihood estimation to find the parameters of the mixture
- Perform a likelihood ratio test to compute $\Pr[P_1 > P_2]$

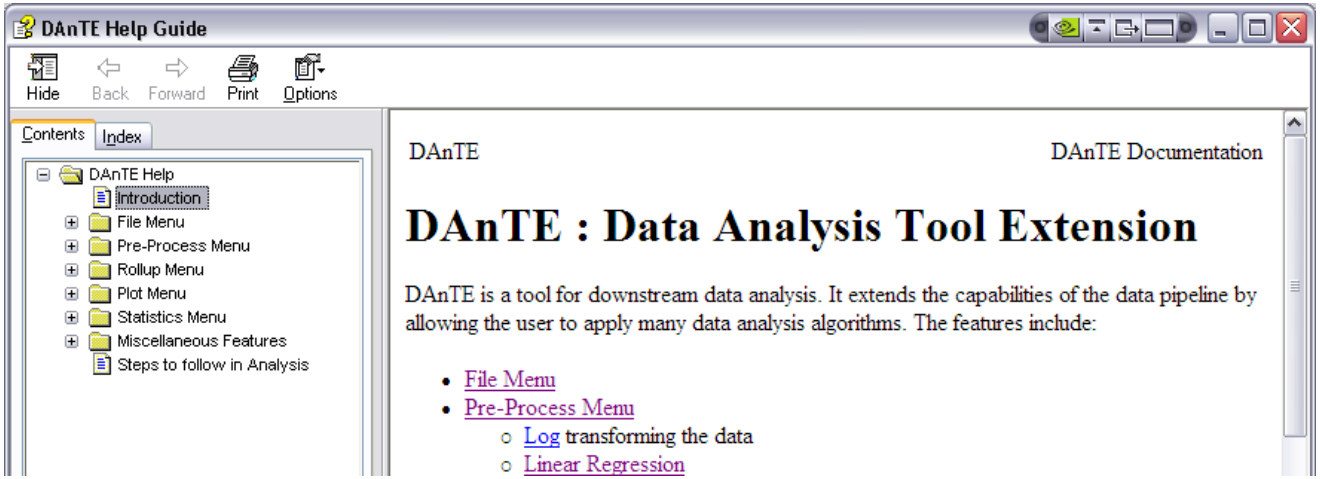
where P1 and P2 are proteins with peptides having missing values.



Further Details



- Help file



- Manuscript in *Bioinformatics*

Category: Data and Text Mining

DAnTE: a statistical tool for quantitative analysis of -omics data

Ashoka D. Polpitiya, Wei-Jun Qian, Navdeep Jaitly, Vladislav A. Petyuk, Joshua N. Adkins, David G. Camp II, Gordon A. Anderson and Richard D. Smith*

Pacific Northwest National Laboratory, Richland, WA 99352, USA.

Bioinformatics 2008; doi: 10.1093/bioinformatics/btn217

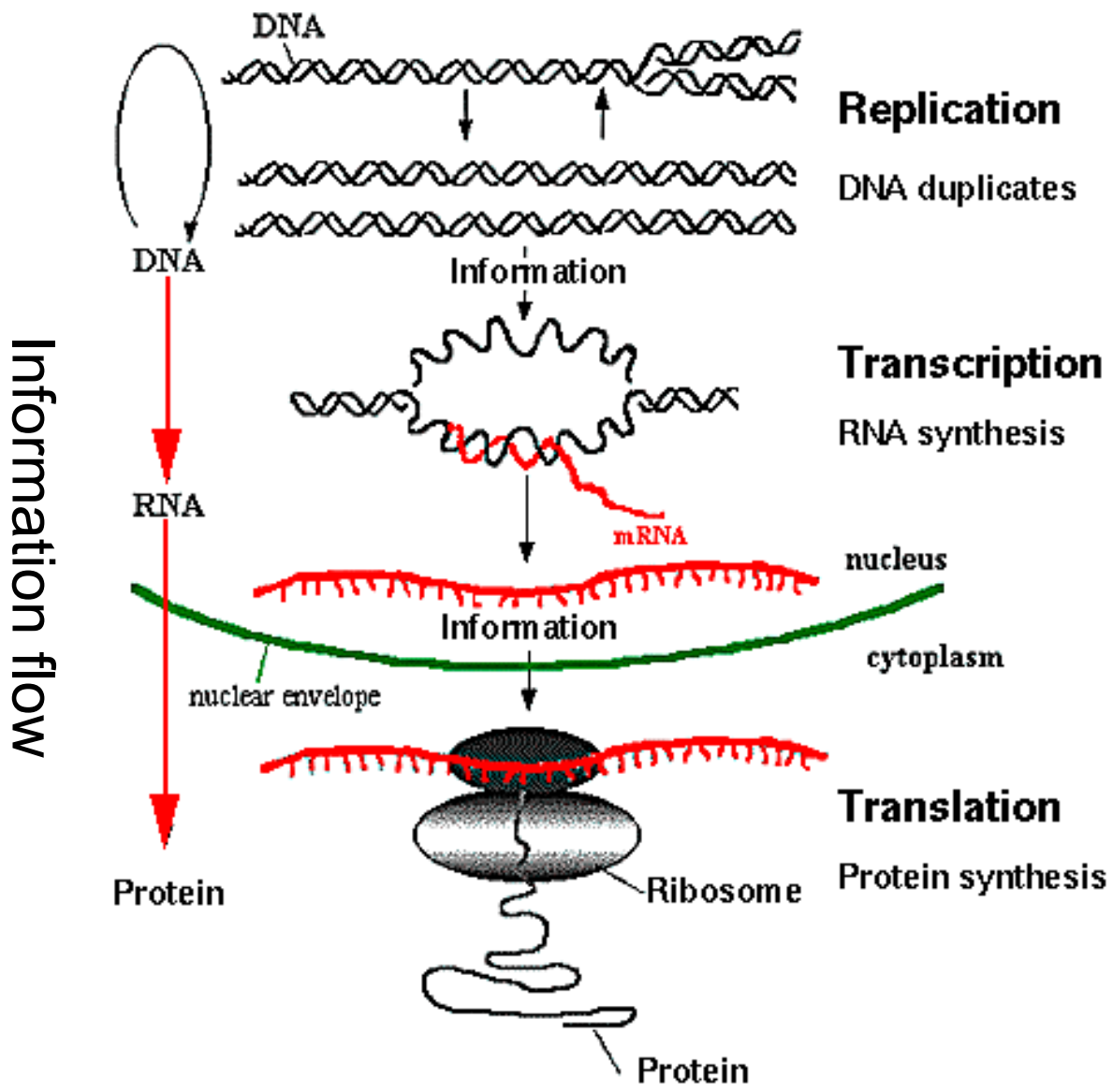
Part III: Biological Application

- Part I: Introduction and Overview of Label-Free Quantitative Proteomics (Anderson)
- Part II: Feature Discovery in LC-MS Datasets (Monroe and Polpitiya)
- Break
- Part III: Biological Application of the AMT tag Approach (Ansong)
  - *Salmonella* Typhimurium
- AMT tag Analysis Software Demo
- Panel Discussion

Biological Application of the AMT tag Approach

- *Salmonella* Typhimurium projects to be discussed:
- Global role of translational regulation in control of gene expression
- Identification of targets for therapeutic intervention

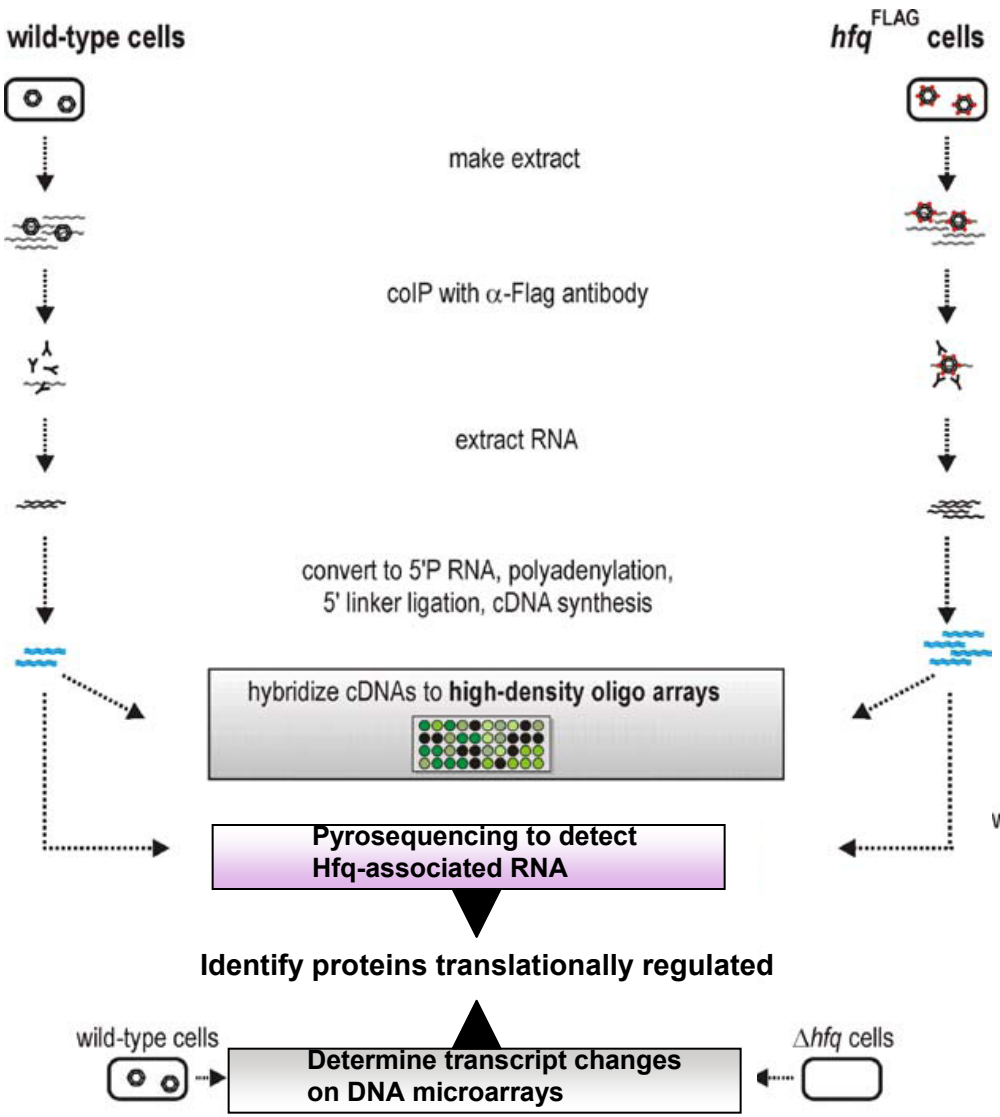
The “Central Dogma” of Biology



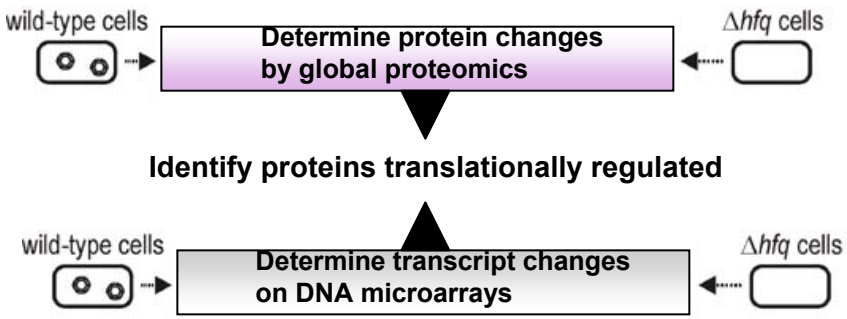
Post-transcriptional Regulation in Bacteria

- Little is known about how RNA binding proteins facilitate the global control of gene expression at the post-transcriptional level
- Bacterial regulation of translation is mediated by relatively few identified protein factors
  - Primarily those encoded by hfq and smpB
- Identifying proteins regulated by Hfq and SmpB is an important step in understanding the impact of post-transcriptional gene regulation in Enterobacteria

Post-transcriptional Regulation in Bacteria: Global identification of Hfq targets in *Salmonella*

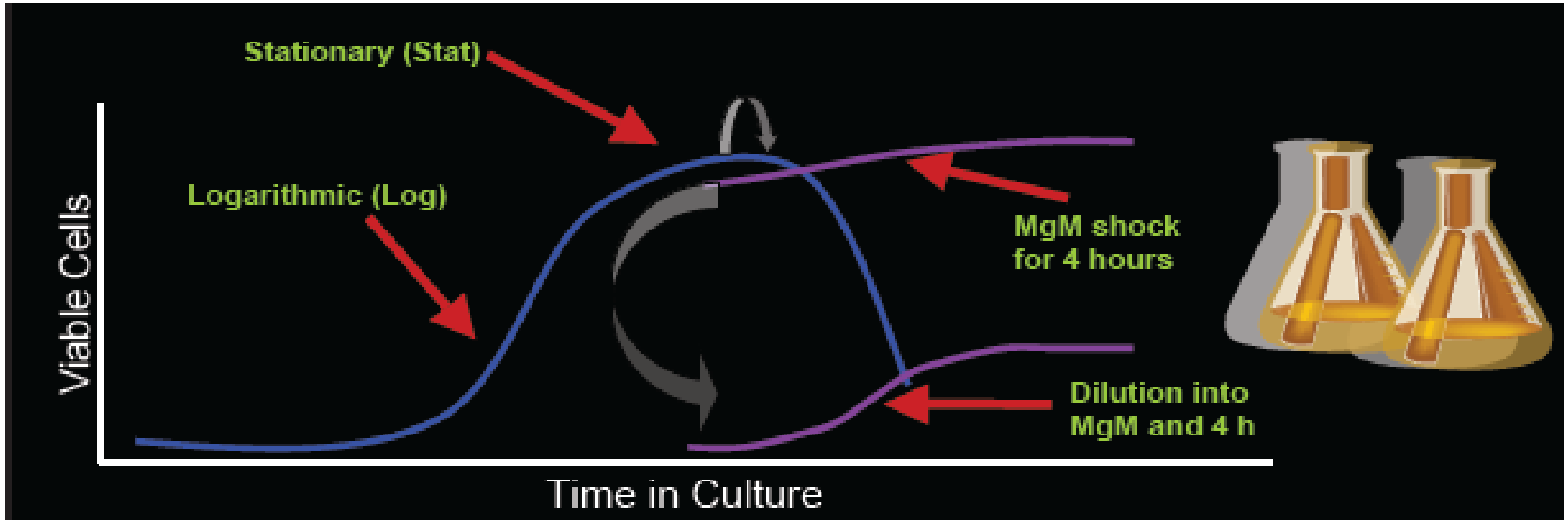


- Hfq strongly binds to RNA molecules
  - Binding is somewhat non-specific
- Possibility of spurious results introduced by tagging and/or co-IP



Adapted from Sittka et al.
PLoS Genetics 2008

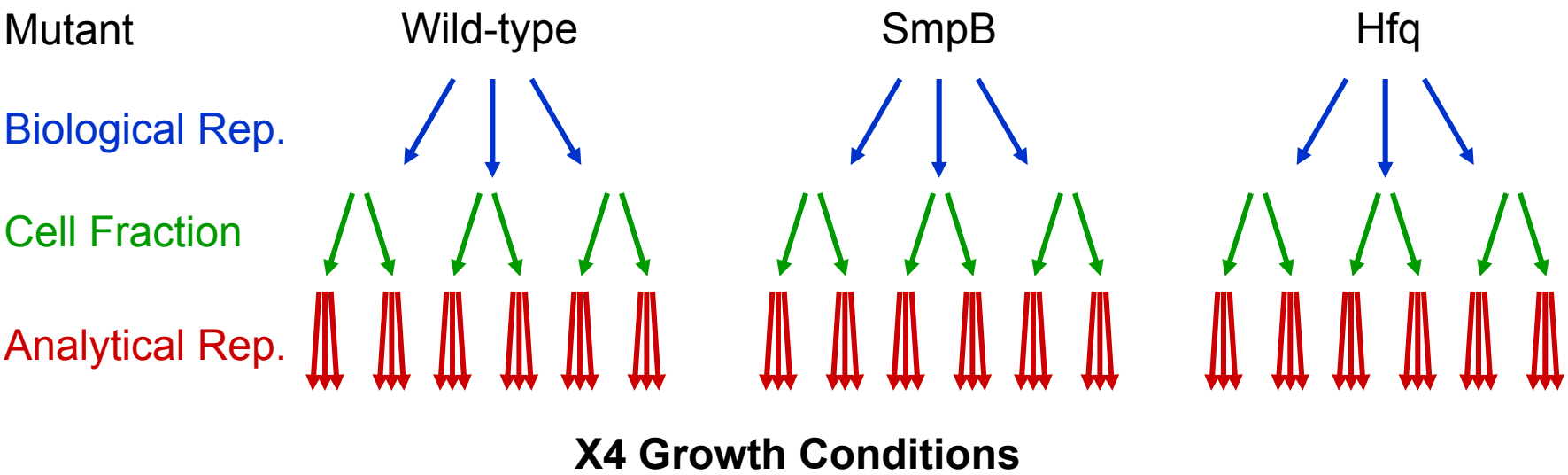
Salmonella Typhimurium Growth in Host-free Culture Conditions



- Laboratory conditions
  - Growth in Luria Bertani broth (LB) to log phase (Log)
  - Growth in LB to stationary phase (Stat)
- Intracellular environment mimic
  - Stat phase cells transferred to acidic minimal media for 4hr, then harvested (MgM Shock)
  - Stat phase cells diluted 1:100 into acidic minimal media growth for 4hr, then harvested (MgM Dilution)

Experimental Design: The Need/Use for Increased Throughput

- Replicate analysis to account for natural biological and normal analytical variation

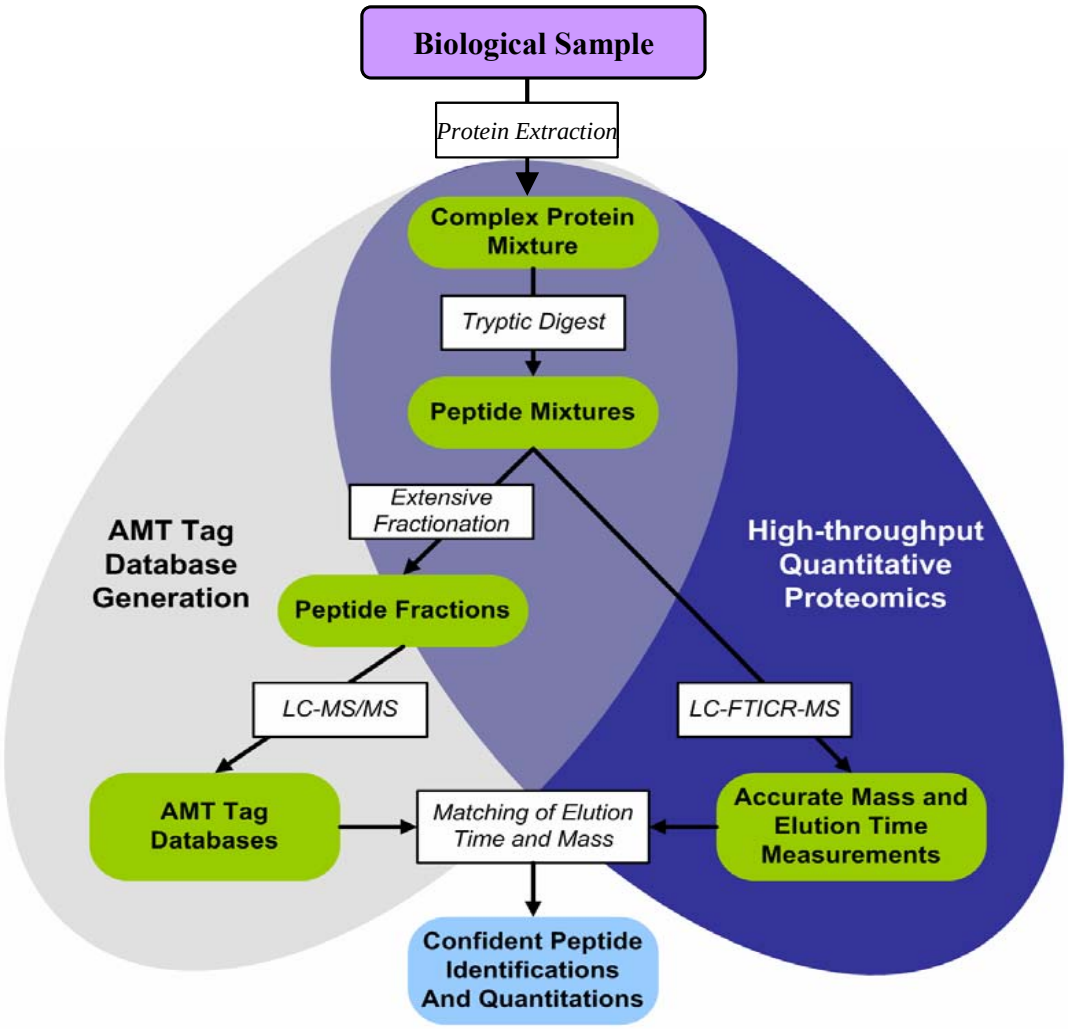


216 global proteome analyses for 2 mutants
LC-MS analysis order blocked and randomized

Ansong PLoS ONE 2009 – In press

Accurate Mass and Time Tag Approach

- AMT tag Database Generation
  - Complex samples fractionated, analyzed by LC-MS/MS, then analyzed using SEQUEST
- High-throughput Quantitative Proteomics
  - Peptides identified from LC-MS peaks by matching LC-MS features to AMT tags
  - Protein abundance value calculation via peptide rollup



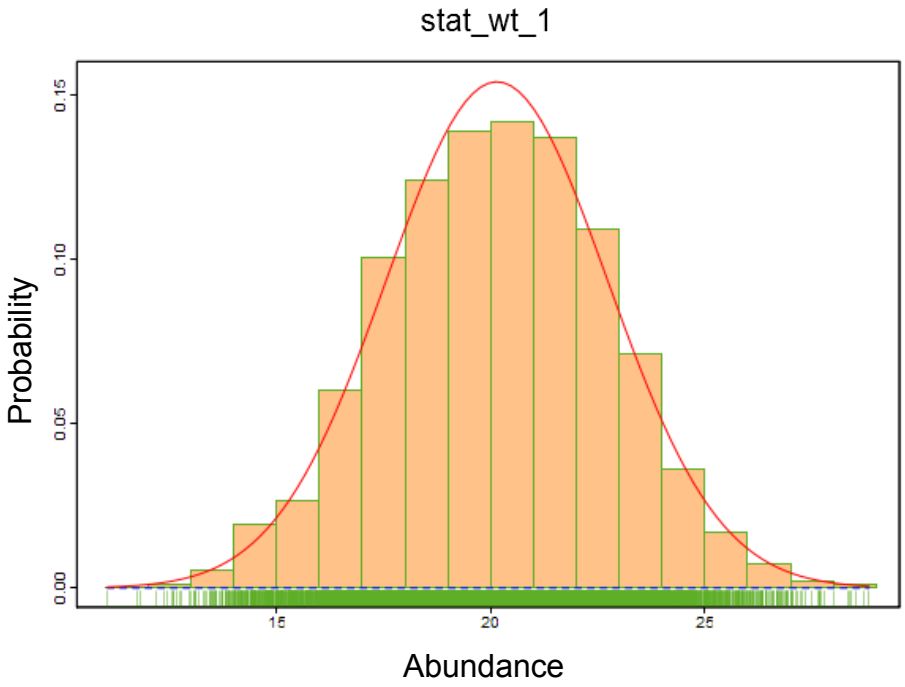
Adkins MCP 2006
– source of AMT tag Database

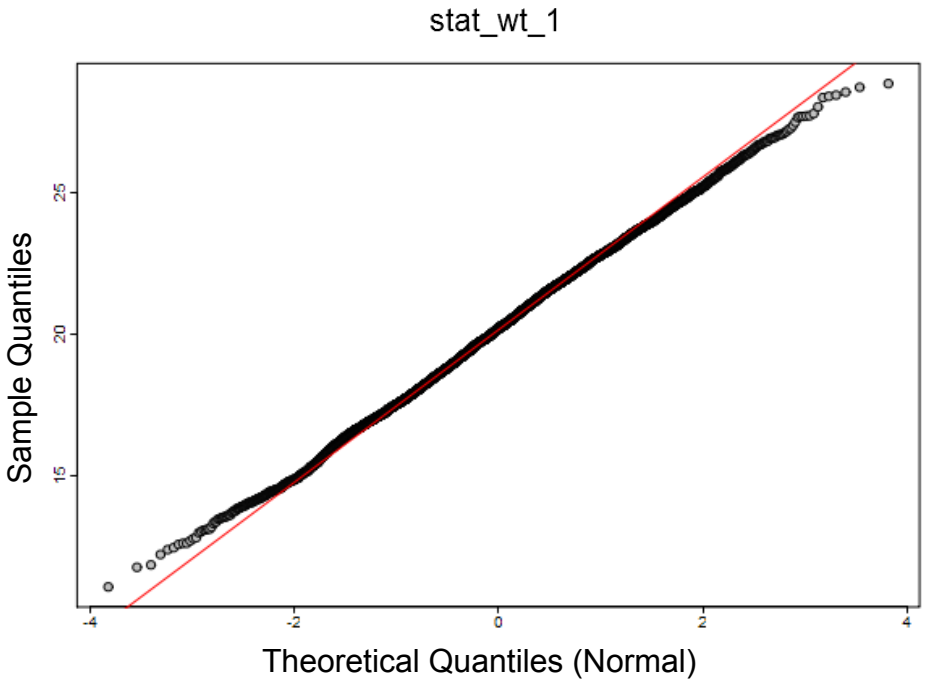
AMT tag Database Overview

- AMT tag database
  - 3 *Salmonella* Typhimurium growth conditions
  - Analyzed both unfractionated and SCX fractionated samples for each condition
  - 932 LC-MS/MS analyses
    - Primarly used two Thermo LTQ mass spectrometers
    - 65 days of instrument acquisition time
  - Data processed with MASIC and SEQUEST
  - 74,000 AMT tags pass filters
- LC-MS Data
  - Four growth conditions for three mutants
  - 216 LC-MS analyses over 16 days of instrument time on a Thermo LTQ-orbitrap mass spectrometer
    - Note: equivalent LC-MS/MS analyses would require 300 days
  - Data processed with Decon2LS and MultiAlign
  - Identified 27,753 AMT tags with ~1% FDR

Diagnostic Plots: Normality

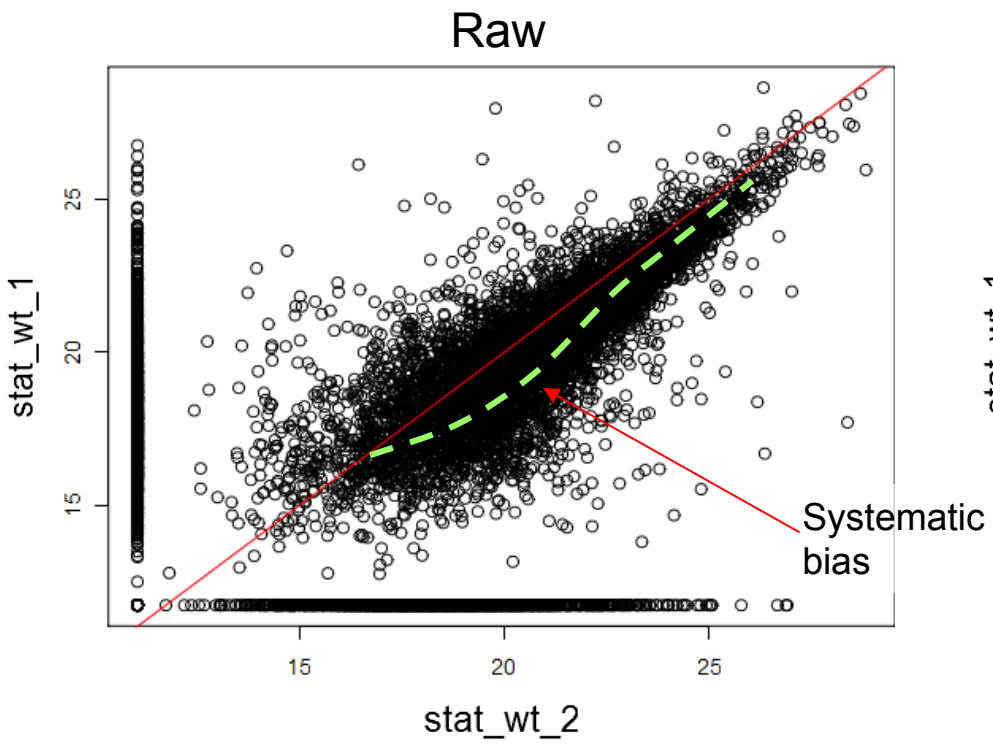
- Distribution of abundance values in a single dataset
  - Should see Normal distribution

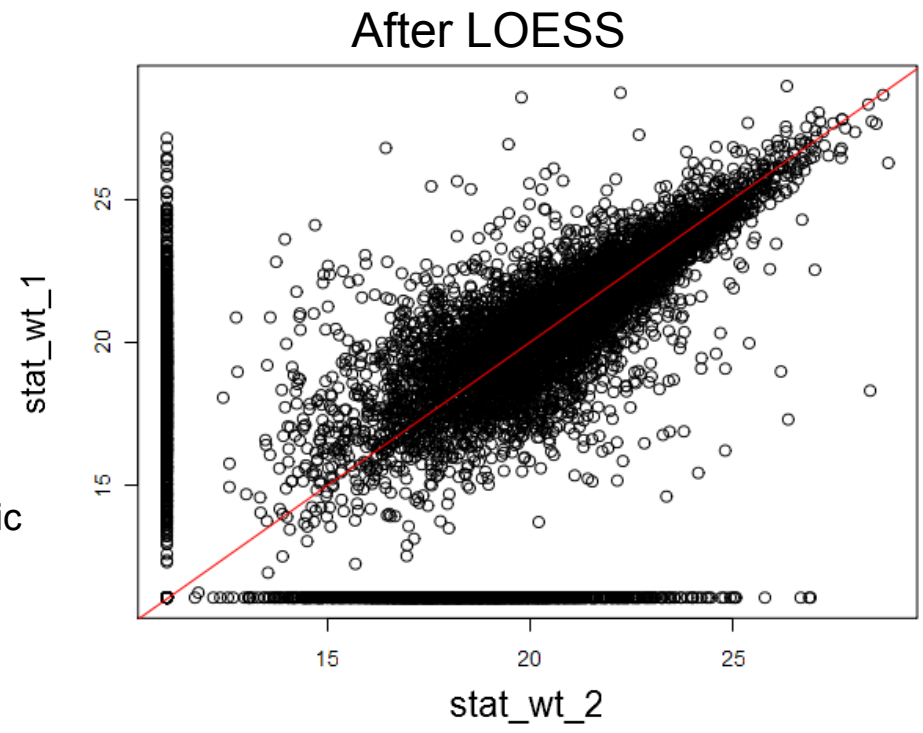




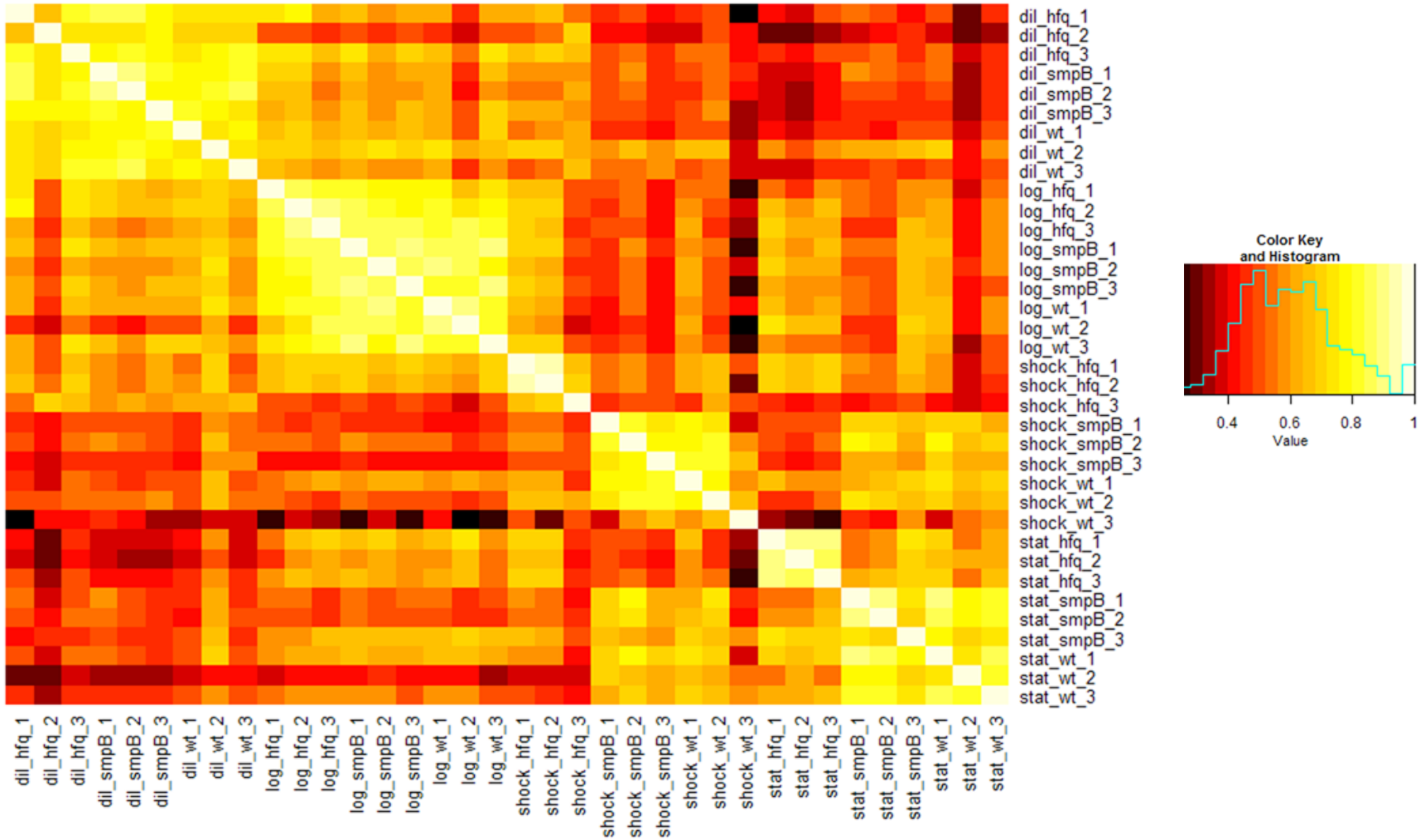
Normalization: LOESS

- Correlation plot (scatter style) of all *Salmonella* peptide identifications from one biological replicate of WT correlated to a second biological replicate of WT



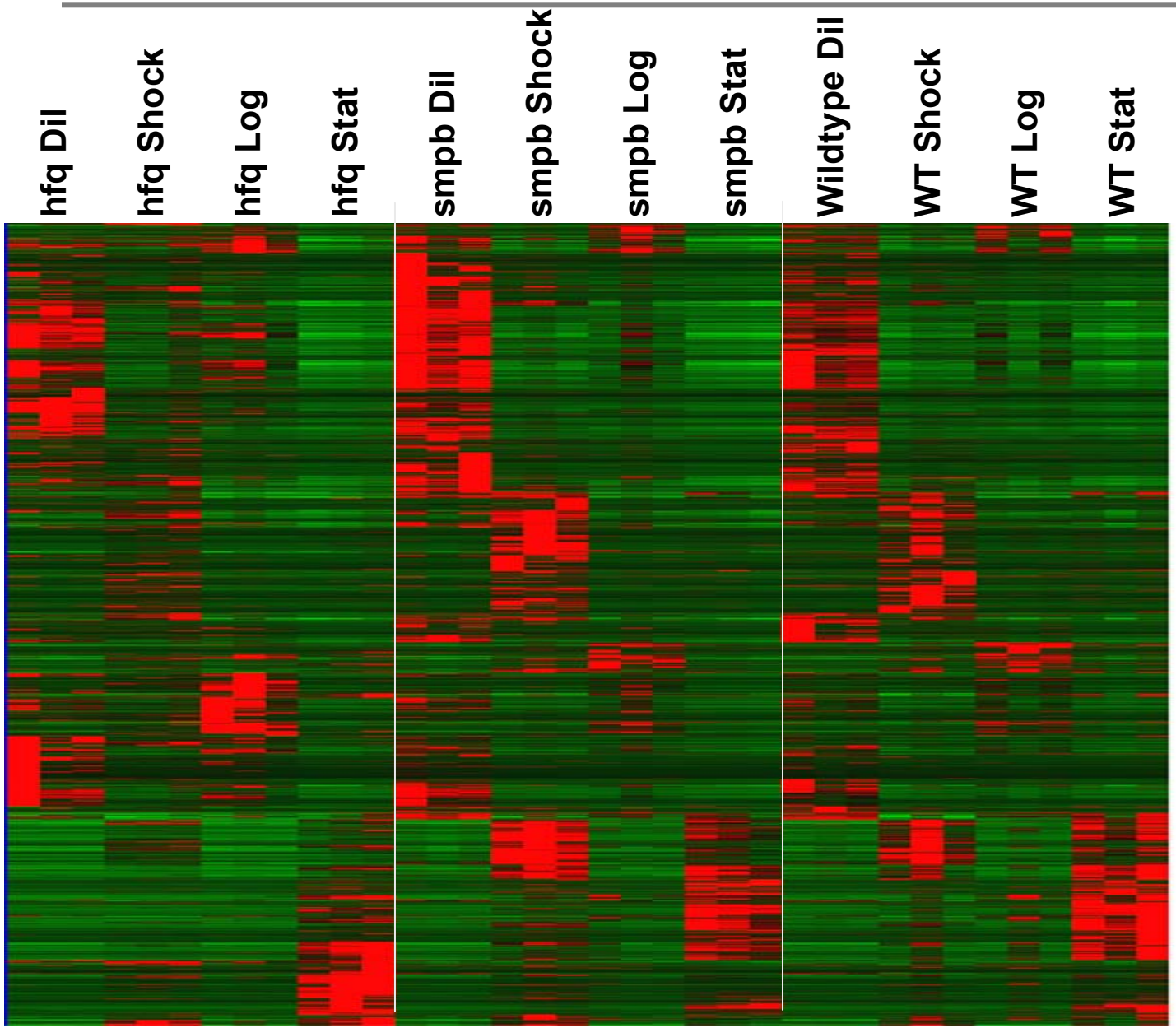


Replicate Reproducibility / Outlier Removal



- Pearson's pairwise correlation plot of all *Salmonella* peptide identifications from each triplicate analyses correlated to itself and every other analyses
- Each replicate analysis has a strong correlation to the other two replicate analyses for each sample, indicating reproducibility between replicate runs

Peptide Abundance Overview



~20,000 Peptides Clustered

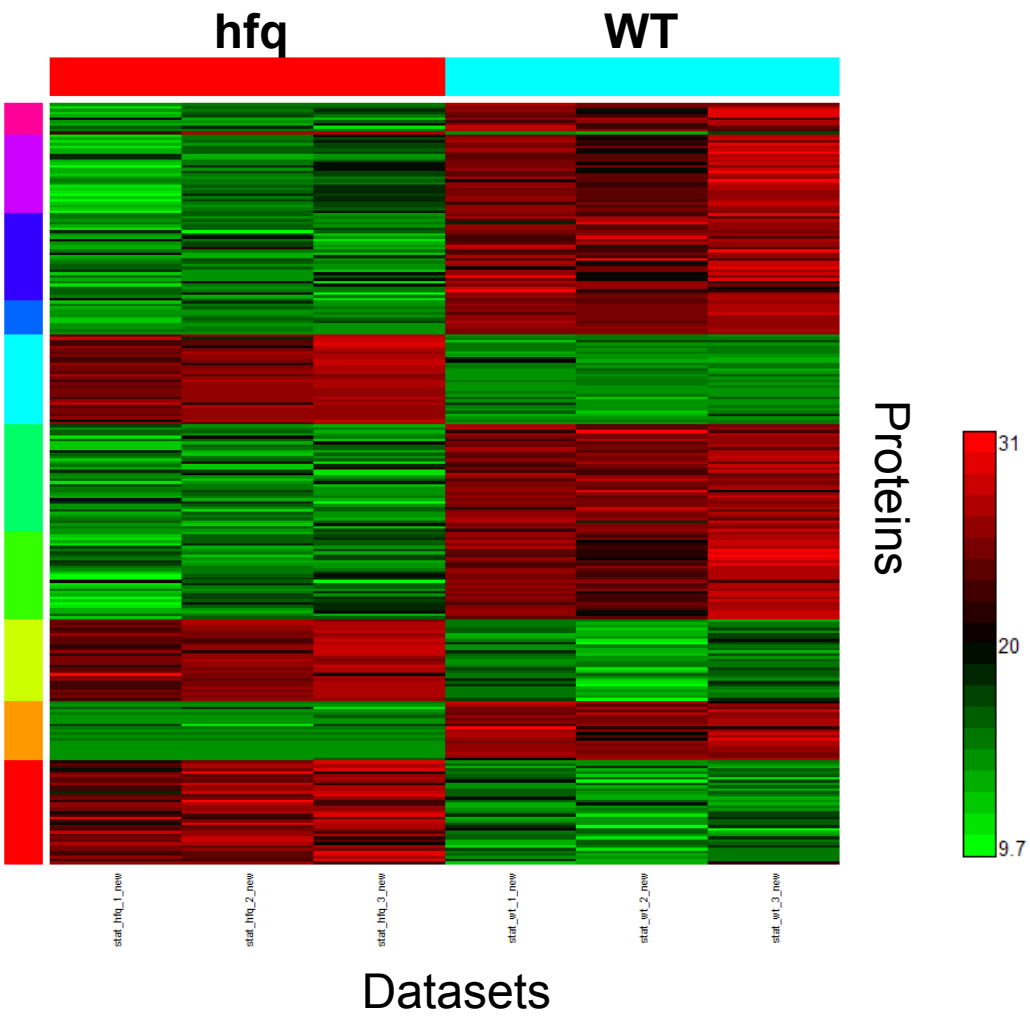
Represents ~1600 Proteins

~36% coverage of *Salmonella* proteome

Biological replicates have good agreement

Ansong PLoS ONE 2009 – In press

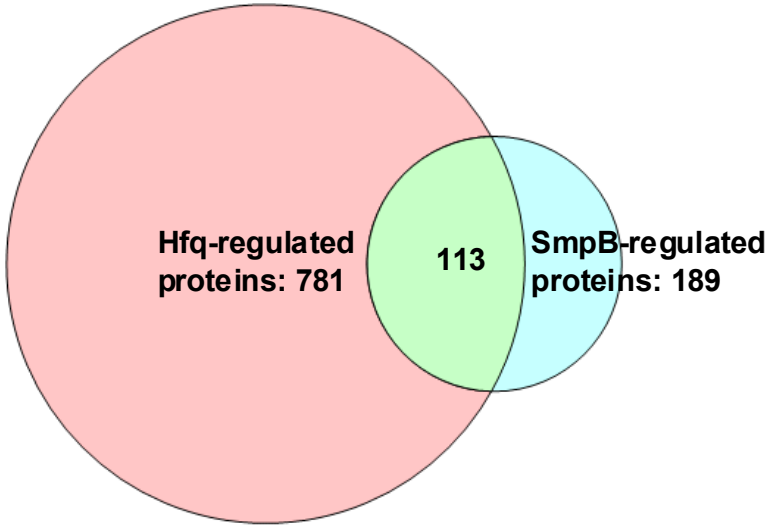
Subset of Proteins Showing Significant Changes

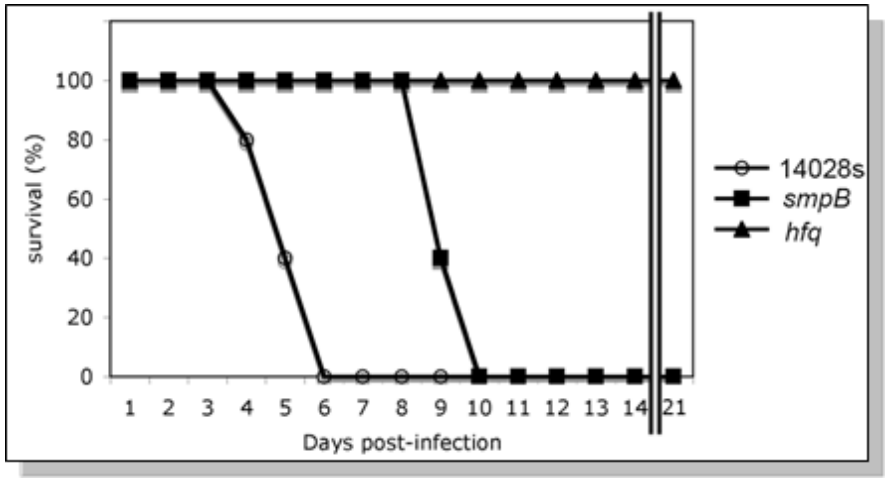


- *Salmonella* WT and hfq mutant grown in LB to Stat phase
- ANOVA between the two groups
  - 270 proteins show significant changes for a false discovery rate of 5%
- 10 groups via K-means clustering

Ansong PLoS ONE 2009 – In press

Functional Analysis





Functional Categories	Total of observed proteome	Hfq regulated	SmpB regulated	% Hfq regulated	% SmpB regulated
Amino acid biosynthesis	84	37	15	44.05	17.86
Biosynthesis of cofactors, prosthetic groups, and carriers	93	34	6	36.56	6.45
Cell envelope	131	58	10	44.27	7.63
Cellular processes	117	74	22	63.25	18.80
Central intermediary metabolism	86	42	17	48.84	19.77
DNA metabolism	56	27	7	48.21	12.50
Energy metabolism	250	149	24	59.60	9.60
Fatty acid and phospholipid metabolism	36	20	1	55.56	2.78
Hypothetical proteins	27	9	1	33.33	3.70
Mobile and extrachromosomal element functions	17	8	6	47.06	35.29
Protein fate	93	45	14	48.39	15.05
Protein synthesis	146	67	15	45.89	10.27
Purines, pyrimidines, nucleosides, and nucleotides	64	33	5	51.56	7.81
Regulatory functions	82	34	5	41.46	6.10
Signal transduction	7	2	0	28.57	0.00
Transcription	31	13	4	41.94	12.90
Transport and binding proteins	113	61	15	53.98	13.27
Unclassified	94	50	11	53.19	11.70
Unknown function	202	99	27	49.01	13.37

Ansong PLoS ONE 2009 – In press

Validating the Proteomics Dataset

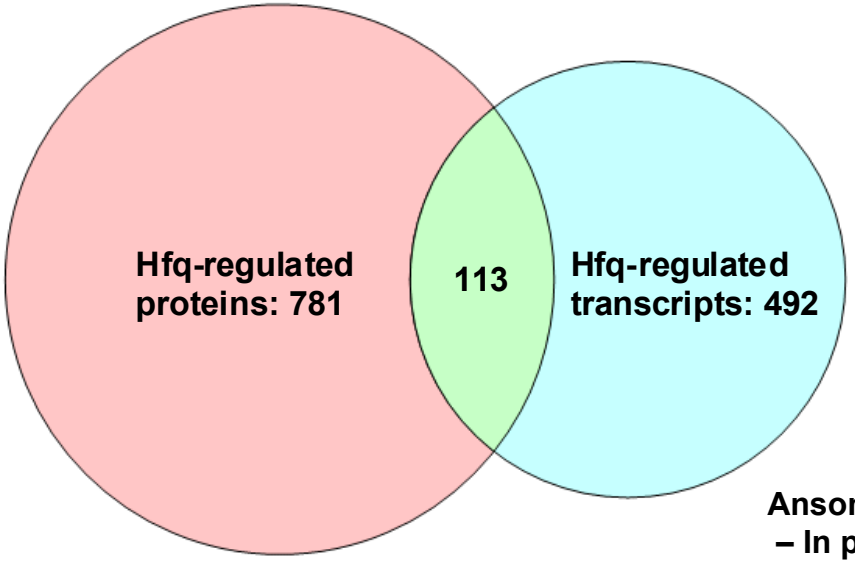
Name	Function	Reg.: Sittka et al.	Reg.: Figueroa-Bossi et al.	Reg.: This study
CarA	carbamoyl-phosphate synthetase, glutamine-hydrolysing small subunit	(-)		(-)
Upp	undecaprenyl pyrophosphate synthetase (di-trans,poly-cis-decaprenylcistrans	(-)		(-)
Dps	stress response DNA-binding protein; starvation induced resistance to H2O2;	(-)		(-)
FliC	flagellin, filament structural protein	(-)		(-)
LuxS	quorum sensing protein, produces autoinducer - acyl-homoserine lactone-sig	(-)		(-)
SipA	cell invasion protein	(-)		(+)
FadA	3-ketoacyl-CoA thiolase; (thiolase I, acetyl-CoA transferase), small (beta) sub	(-)		(-)
OsmY	hyperosmotically inducible periplasmic protein, RpoS-dependent stationary ph	(-)		(-)
RpsD	30S ribosomal subunit protein S4	(-)		(-)
SurA	peptidyl-prolyl cis-trans isomerase, survival protein	(+)		(+)
HtrA	periplasmic serine protease Do, heat shock protein	(+)		(+)
Tpx	thiol peroxidase	(+)		(+)
OppA	ABC superfamily, oligopeptide transport protein with chaperone properties	(+)		(+)
GlpQ	glycerophosphodiester phosphodiesterase, periplasmic	(+)		(+)
STM2494	putative inner membrane or exported	(+)		(+)
GreA	transcription elongation factor, cleaves 3' nucleotide of paused mRNA	(+)		(+)
FkpA	FKBP-type peptidyl-prolyl cis-trans isomerase (rotamase)	(+)		(+)
DppA	ABC superfamily, dipeptide transport protein	(+)		(+)
AphA	non-specific acid phosphatase/phosphotransferase, class B	(+)		(+)
GlnH	ABC superfamily (bind_prot), glutamine high-affinity transporter	(+)		(+)
MglB	ABC superfamily (peri_perm), galactose transport protein	(+)		(+)
GlpK	glycerol kinase	(+)		(+)
KatE	catalase hydroperoxidase HPII(III), RpoS dependent		(-)	(-)
PduD	Propanediol utilization: dehydratase, medium subunit		(+)	(+)
PduO	Propanediol utilization: B12 related		(+)	(+)
AphA	non-specific acid phosphatase/phosphotransferase, class B		(+)	(+)

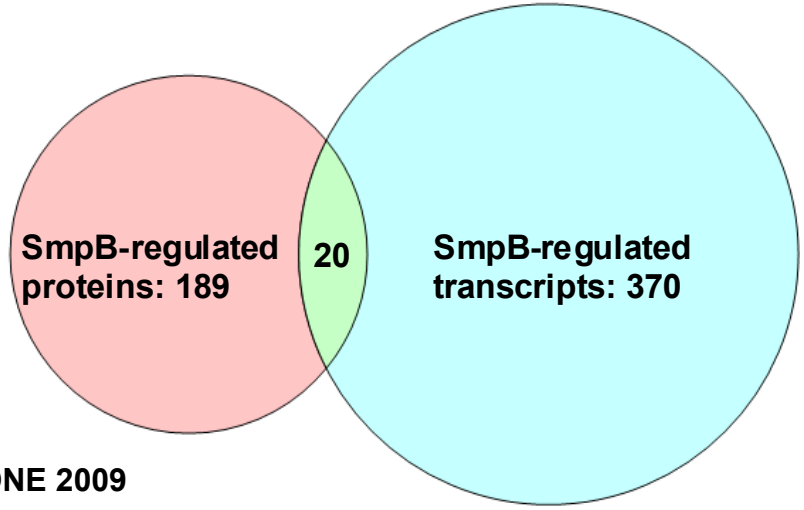
25/26 proteins are in general agreement with literature

Ansong PLoS ONE 2009 – In press

Alteration of Protein Expression Mediated Post-transcriptionally

- Integrating information from transcriptional analysis





Ansong PLoS ONE 2009 – In press

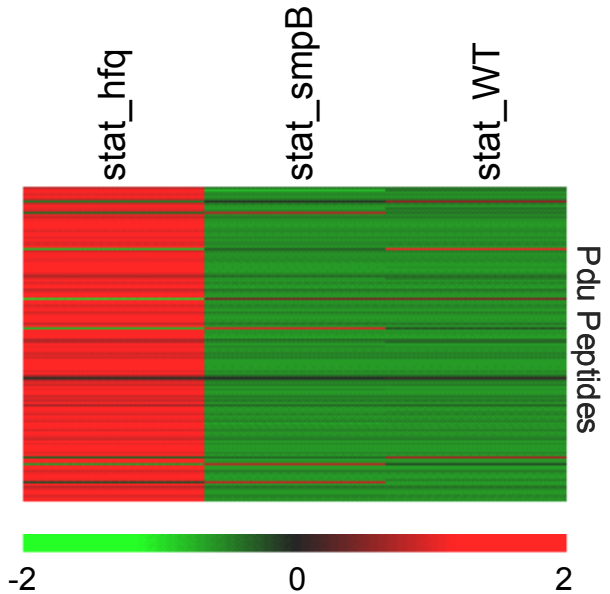
Published global transcriptional analyses of Hfq-regulated genes

Organism	% Hfq-regulated transcripts
Salmonella Typhimurium[1]	20 %
Pseudomonas aeruginosa[2]	15 %
E. coli[3]	6 %
Vibrio cholerae[4]	6%

- 1 – Sittka et al. 2008
- 2 – Sonnleitner et al. 2006
- 3 – Gusibert et al. 2007
- 4 – Ding et al. 2004

Validation of Translational Regulation in *Salmonella*: The propanediol utilization (pdu) operon

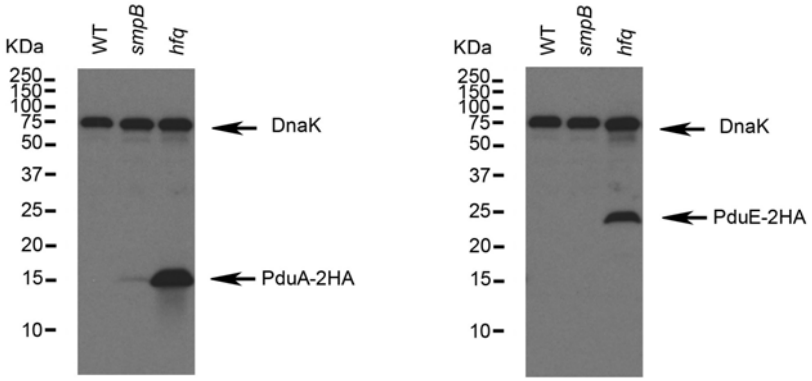
- LC-MS analysis of protein levels



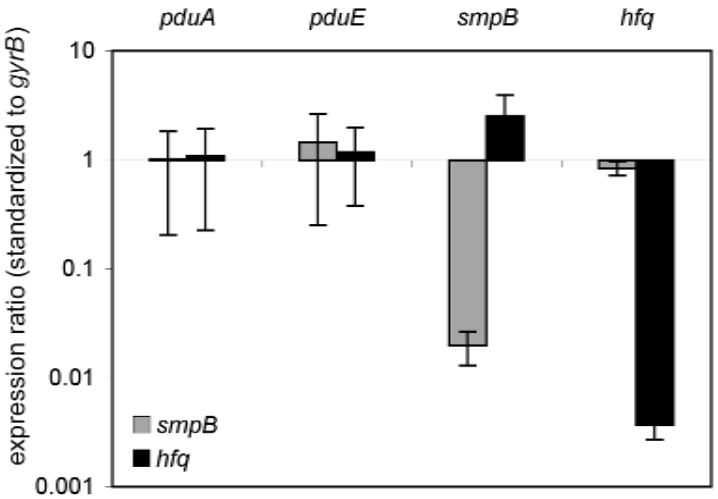
- pdu required for replication in macrophages
- Deletion of hfq or smpB reduces expression of Pdu proteins levels, but mRNA levels show no change
- Suggests role for translational regulation in *Salmonella* pathogenesis

Ansong PLoS ONE 2009 – In press

- Confirmatory western blot analysis



- qRT-PCR analysis of mRNA levels



Additional Benefit of the AMT tag Approach: Increased depth of coverage increases confidence metrics

LTQ-Orbitrap LC-MS/MS datasets, SEQUEST analysis results

Protein	Description	PeptideSequence	Unique Peptides	Peptide Count
PSLT046	putative carbonic anhydrase	K.IVGSMYHLTGGKVEFFEV.-	6	11
PSLT046	putative carbonic anhydrase	K.NVELTIENIRK.N	6	2
PSLT046	putative carbonic anhydrase	K.VVLVIGHTR.C	6	14
PSLT046	putative carbonic anhydrase	R.FRENRPAKHDYLAQK.R	6	3
PSLT046	putative carbonic anhydrase	R.KGSNYDFVDAVAR.K	6	1
PSLT046	putative carbonic anhydrase	R.VAGNISNR.D	6	7
			6	38

Roll up to protein level

Protein	Description	hfq_count	wt_count
PSLT046	putative carbonic anhydrase	38	1

AMT tag analysis of the same LTQ-Orbitrap datasets

Protein	Description	PeptideSequence	Unique Peptides	Peptide Count	hfq_abundance	wt_abundance
PSLT046	putative carbonic anhydrase	R.KGSNYDFVDAVAR.K	13	52	1242040.333	1060
PSLT046	putative carbonic anhydrase	R.KGSNYDFVDAVARK.N	13	12	985149.3333	1060
PSLT046	putative carbonic anhydrase	R.KNVELTIENIRK.N	13	37	1976177.667	1060
PSLT046	putative carbonic anhydrase	R.FRENRPAKHDYLAQK.R	13	6	773729	1060
PSLT046	putative carbonic anhydrase	R.APAEIVLDAGIGETFNSR.V	13	7	1060	125424
PSLT046	putative carbonic anhydrase	K.VVLVIGHTR.C	13	41	4955406.667	1060
PSLT046	putative carbonic anhydrase	R.KNVELTIENIR.K	13	103	1042892	221046
PSLT046	putative carbonic anhydrase	K.IVGSMYHLTGGKVEFFEV.-	13	43	1520394.333	1060
PSLT046	putative carbonic anhydrase	A.ASLSKEERDGMTPDAVIEHFK.Q	13	23	116143.5	1060
PSLT046	putative carbonic anhydrase	A.ASLSKEERDGMTPDAVIEHFKQGNLR.F	13	11	1019158.333	1060
PSLT046	putative carbonic anhydrase	R.NSIAGQYPAAVILSCIDSR.A	13	25	943607.6667	1109590
PSLT046	putative carbonic anhydrase	K.NSPVLKQLEDEKKIK.I	13	11	1237220.667	1060
			13	371		

Roll up to protein level

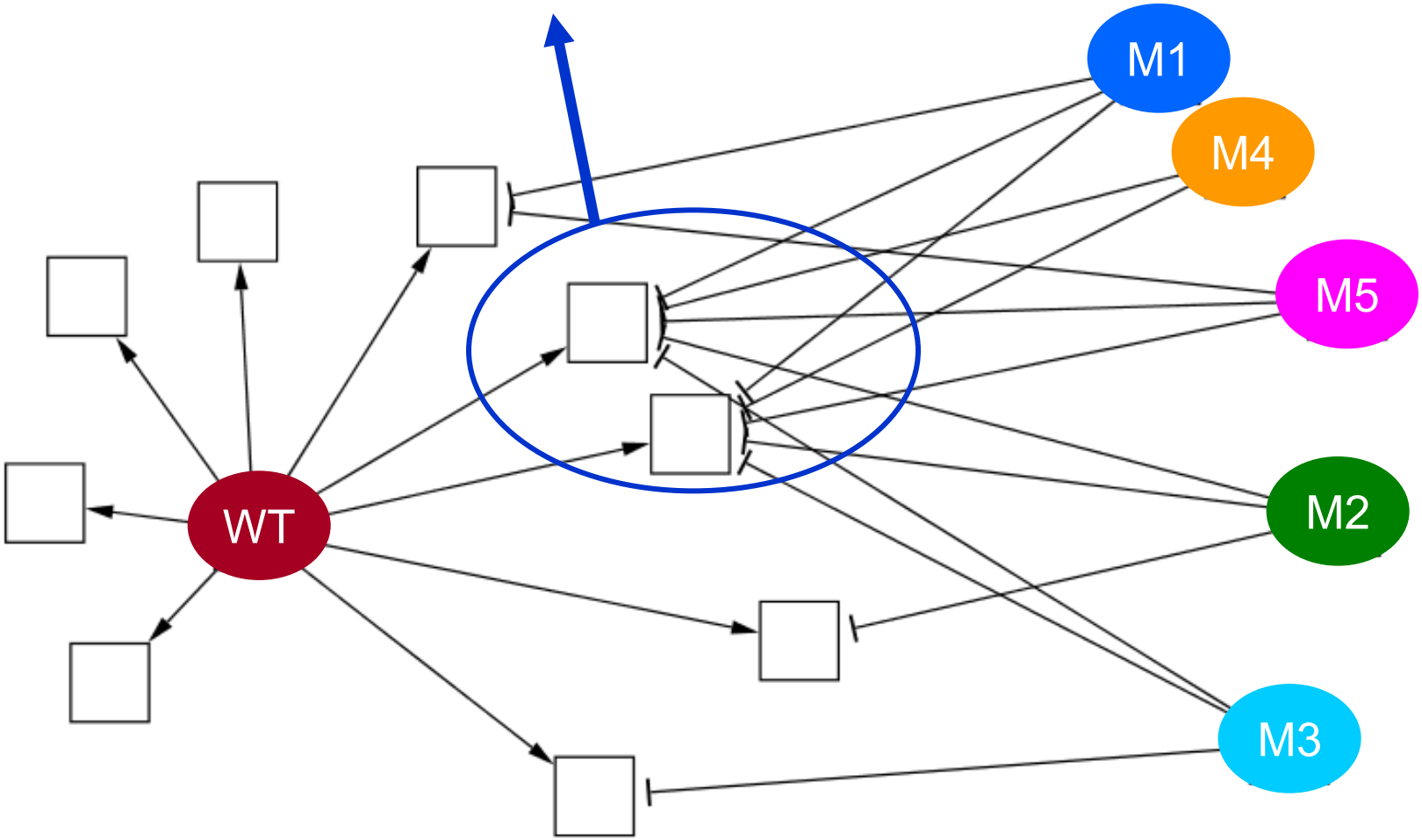
Protein	Description	hfq_abundance	wt_abundance	Fold Change
PSLT046	putative carbonic anhydrase	19.99	10.42	9.57

S. Typhimurium: Public health burden

- Bacterial pathogen with a broad host range
- Humans: causes salmonellosis (severe form of food poisoning)
  - Potentially life threatening to infants, elderly, and immuno-compromised
  - 40,000+ reported salmonellosis cases/year in the U.S.
  - Financial cost ~ $3 billion/year (Tauxe 1986, USDA 07/08)
- Isolates becoming resistant to frontline antibiotics (fluoroquinolone drugs, Lowry 2006)
- Imperative to identify new targets for therapeutic intervention

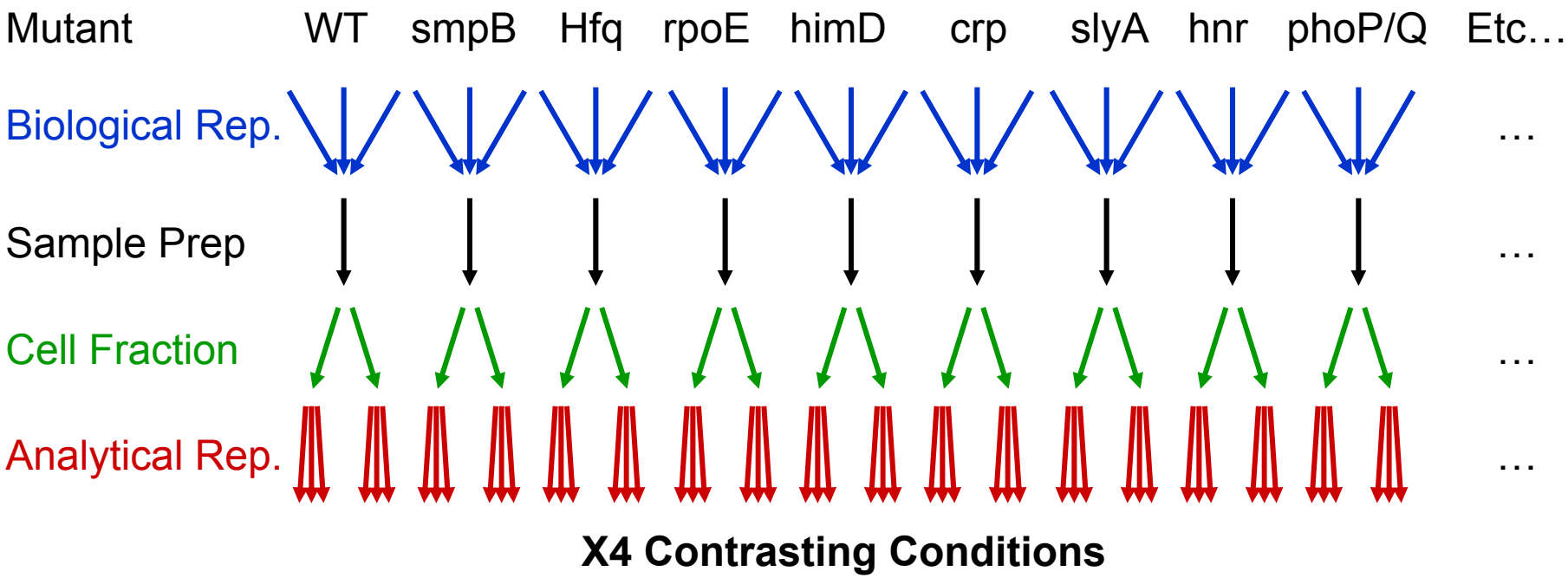
Understanding Regulation of Virulence in *Salmonella*

- Hypothesis: Knock-out regulatory proteins involved in pathogenesis and the commonly regulated proteins represent the best targets for therapeutics



Experimental Design: The Need/Use for Increased Throughput

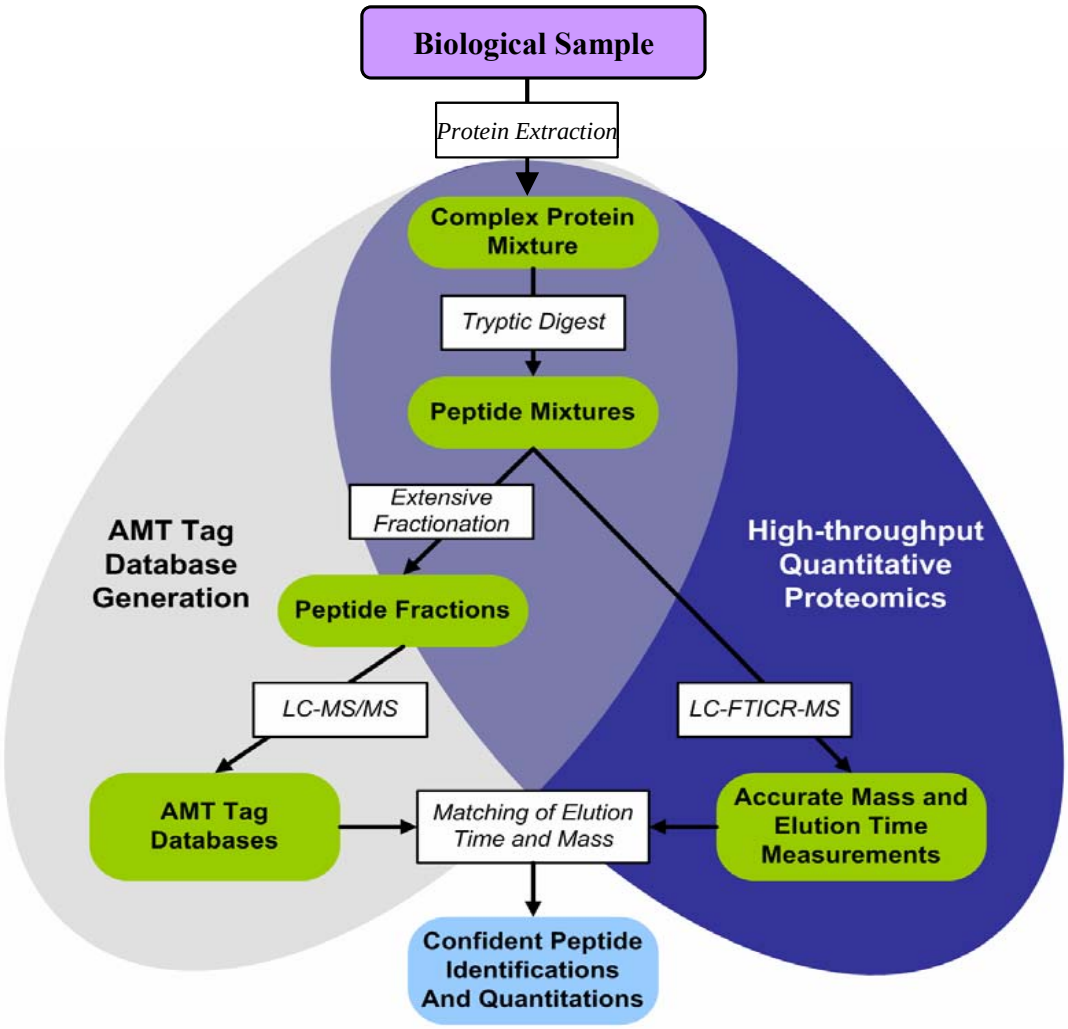
- Replicate analysis to account for natural biological and normal analytical variation



1080 analyses for 15 mutants
using biological pooling 360 analyses

Accurate Mass and Time Tag Approach

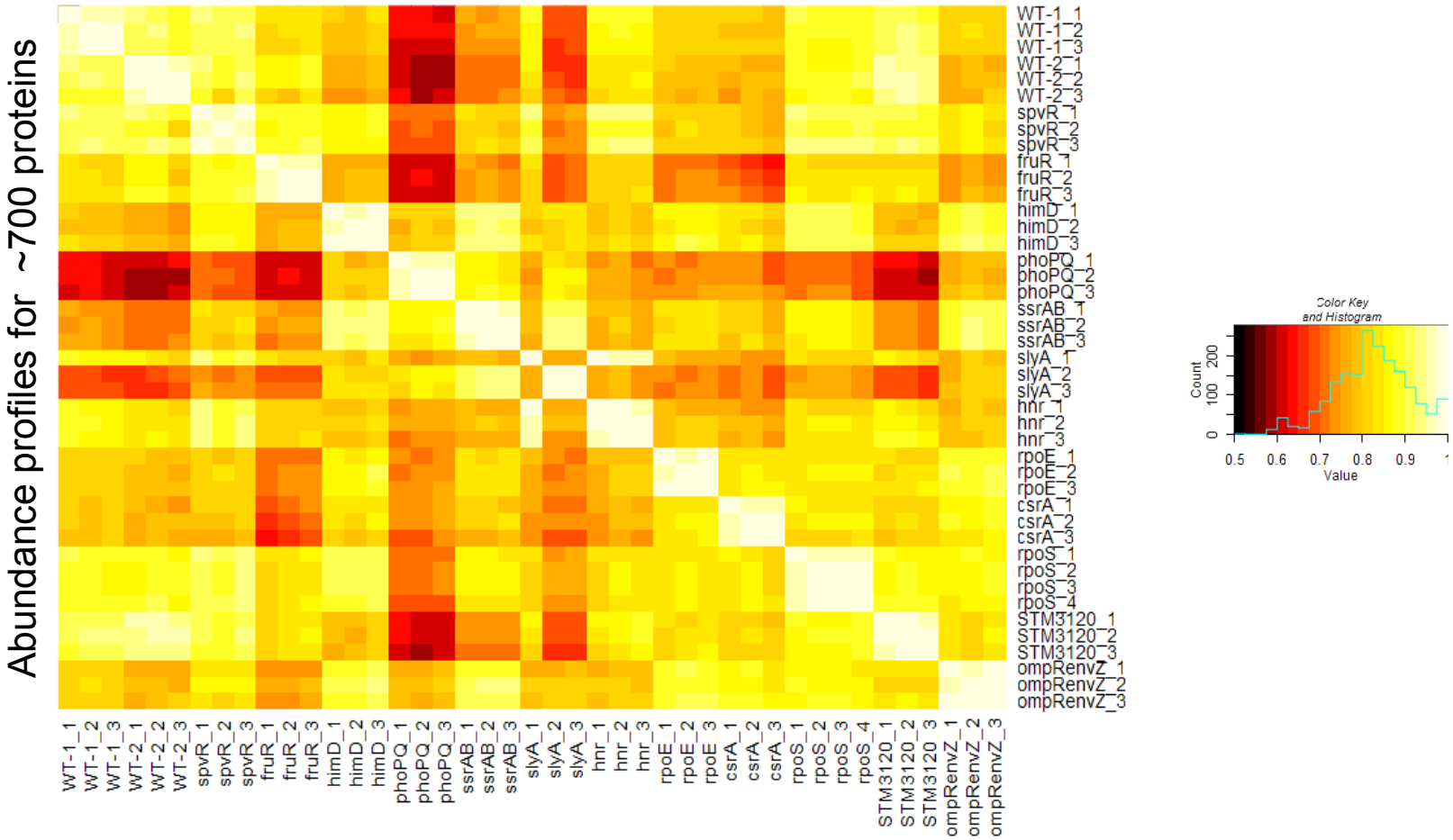
- AMT tag Database Generation
  - Complex samples fractionated, analyzed by LC-MS/MS, then analyzed using SEQUEST
- High-throughput Quantitative Proteomics
  - Peptides identified from LC-MS peaks by matching LC-MS features to AMT tags
  - Protein abundance value calculation



AMT tag Database Overview

- AMT tag database
  - 4 *Salmonella* Typhimurium growth conditions
  - Analyzed both unfractionated and SCX fractionated samples for each condition
  - 1349 LC-MS/MS analyses
    - Primarily used two Thermo LTQ's and one LTQ-Orbitrap
    - 77 days of instrument time
  - Data processed with MASIC and SEQUEST
  - 70,000 AMT tags pass filters
- LC-MS Data
  - Four growth conditions for three mutants
  - 409 LC-MS analyses over 10 days of instrument time on a Thermo LTQ-orbitrap mass spectrometer (30 minute separations)
    - Note: equivalent LC-MS/MS analyses would require 450 days
  - Data processed with Decon2LS and VIPER
  - Identified 16,367 AMT tags with ~1% FDR

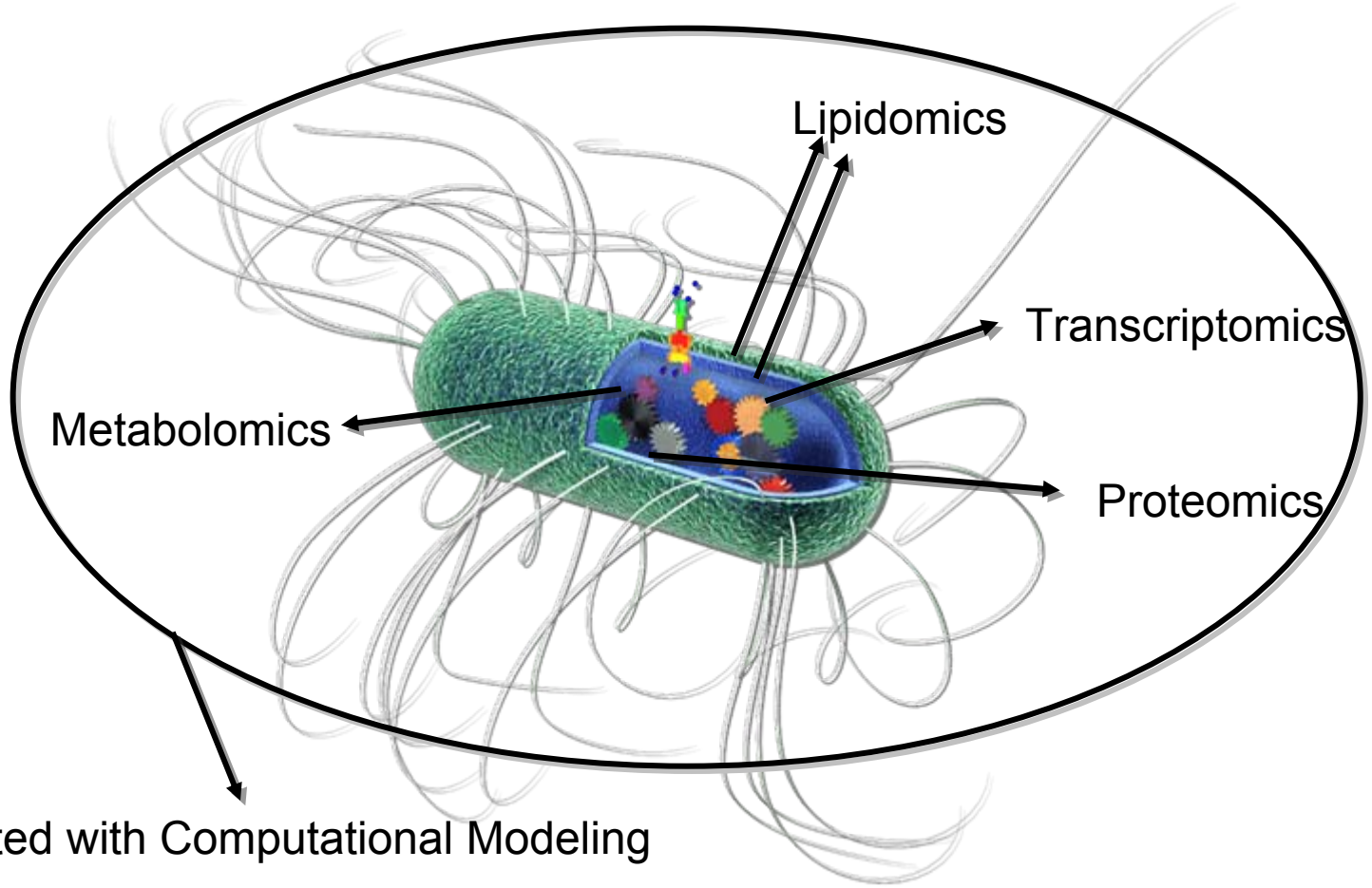
Replicate Reproducibility / Outlier Removal



- Pearson's pairwise correlation plot of all *Salmonella* peptide identifications from each triplicate analyses correlated to itself and every other analyses
- Each replicate analysis has a strong correlation to the other two replicate analyses for each sample, indicating reproducibility between replicate runs

Conclusion

- The AMT tag approach and the associated informatics pipeline enables Systems Biology experiments that would be impractical using shotgun proteomics and/or chemical/labeling methods



AMT tag Analysis Software Demo

- Part I: Introduction and Overview of Label-Free Quantitative Proteomics (Anderson)
- Part II: Feature Discovery in LC-MS Datasets (Monroe and Polpitiya)
- Break
- Part III: Biological Application of the AMT tag Approach (Ansong)
- AMT tag Analysis Software Demo
  - MTDB Creator
  - VIPER
  - DAnTE
  - See the supplied DVD for the software installers for the software that will be used in the live demo
- Panel Discussion

Example Data for the AMT tag Software Demo

- *Salmonella* Typhimurium, LC-MS/MS
  - Grown in LB (Luria-Bertani) up to log phase
  - “Mini AMT tag” database, composed of 25 SCX fractions analyzed by LC-MS/MS
  - Mass and time tag database composed from searches using X!Tandem (Log E_Value ≤ -2)
  - Linear alignment of datasets for AMT tag database
- LC-FTICR-MS analysis (LTQ-Orbitrap)
  - Knock-out mutant sample, grown and prepared in the same conditions
  - Non-linear alignment and peak matching to the database

AMT tag Software Demo

- Thanks to the many developers, beta testers, and users

Funding for Tool Development

- DOE Office of Biological and Environmental Research
  - http://ober-proteomics.pnl.gov/
- NIH
  - National Center for Research Resources
    - http://ncrr.pnl.gov/
  - National Institute of Allergy and Infectious Diseases
    - http://www.sysbep.org/
  - National Cancer Institute
  - National Institute of General Medical Sciences
  - National Institute of Diabetes & Digestive & Kidney Diseases

See the supplied DVD for the software installers for the software that will be used in the live demo

- **MTDB Creator**
  - \Software_Installers\MTDBCreator\MTDBCreatorInstall.msi
  - Data file: \Software_Installers\MTDBCreator\data\DatasetsDescription.txt
- **VIPER**
  - \Software_Installers\VIPER\VIPER_Installer.msi
  - \Software_Installers\VIPER\LCMSFeatureFinder (Install this after installing Viper).msi
  - Data file: \VIPER_Data\MT_S_typhimurium_X347\Job219616_LTQ-Orbitrap\Job219616.gel
- **DAnTE**
  - \Software_Installers\DAnTE\R_Installers\R-2.7.2-win32.exe
  - \Software_Installers\DAnTE\R_Installers\RSrv250_pl1.exe
  - \Software_Installers\DAnTE\DAnTE_Standalone_Installer\DAnTESetup.exe
  - Data file: \DAnTE_Data\USHUPO2009_DAnTE_data.dnt

- Note: Live demo will not be until the end of the course