Data Extraction and Analysis for LC-MS Based Proteomics

Instructors

Jake Jaffe[1], Deep Jaitly[2], and Matt Monroe[2]

Co-Organizers

Josh Adkins[2] and Gordon Anderson[2]

[1] The Broad Institute, Cambridge, MA 02142

[2] Pacific Northwest National Laboratory, Richland, WA 99354

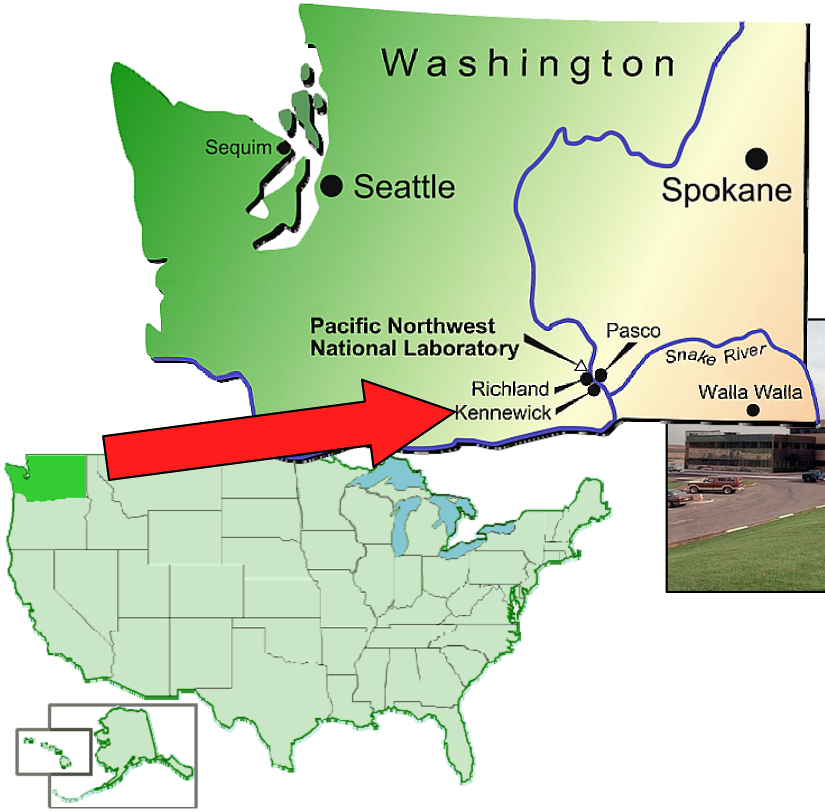
Course Outline

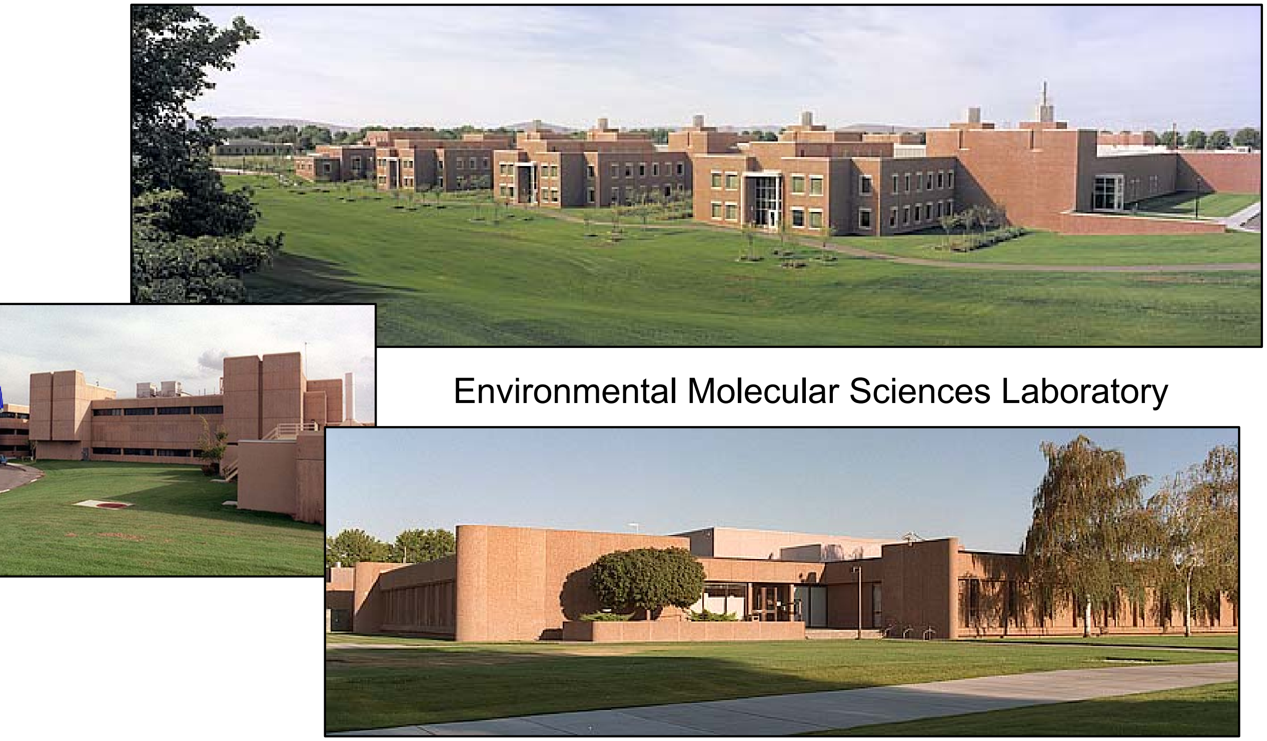
- Introduction (Adkins)
    - Goals
    - Our Historical Perspective
    - Why Use an LC-MS Approach
    - Data and Tools Availability
- Part I: Overview of Label-Free Quantitative Proteomics (Jaffe)
- Part II: Feature discovery in LC-MS datasets (Monroe and Jaitly)
- Part III: PEPPeR, GenePattern and Real-world examples (Jaffe)
- Break
- AMT tag Pipeline Demo (general)
- 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



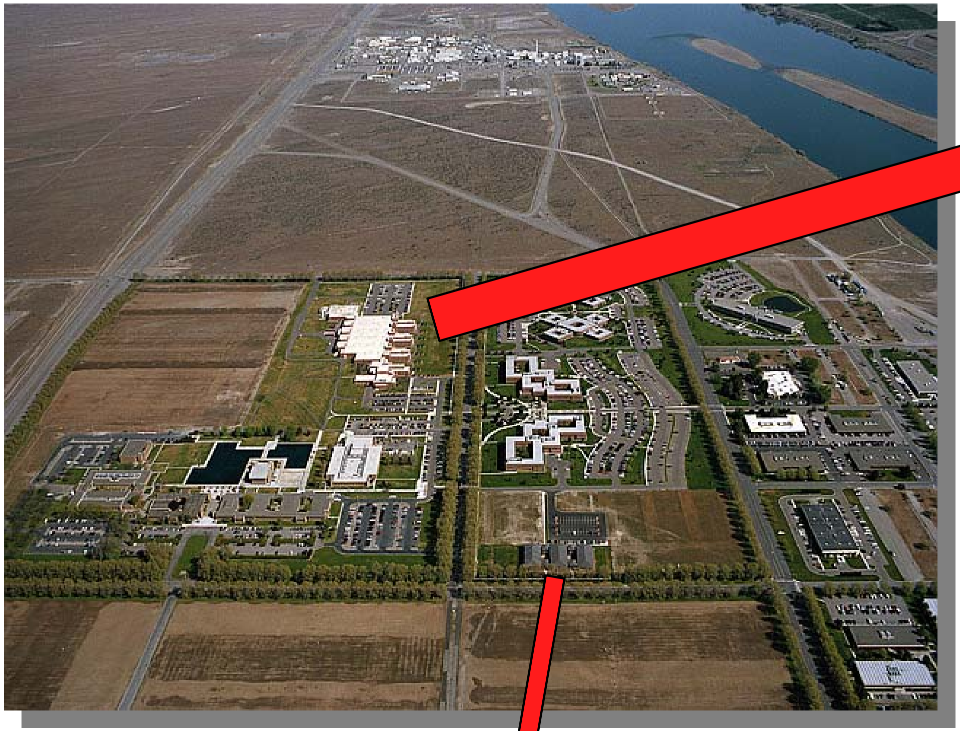


Environmental Molecular Sciences Laboratory

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**

BROAD INSTITUTE

“Realizing the promise of the genome project for human health”

A collaboration among MIT, Harvard, and affiliated teaching hospitals

Programs

- Cellular Circuits
- Medical Genetics
- Chemical Biology
- Cancer Research

Initiatives

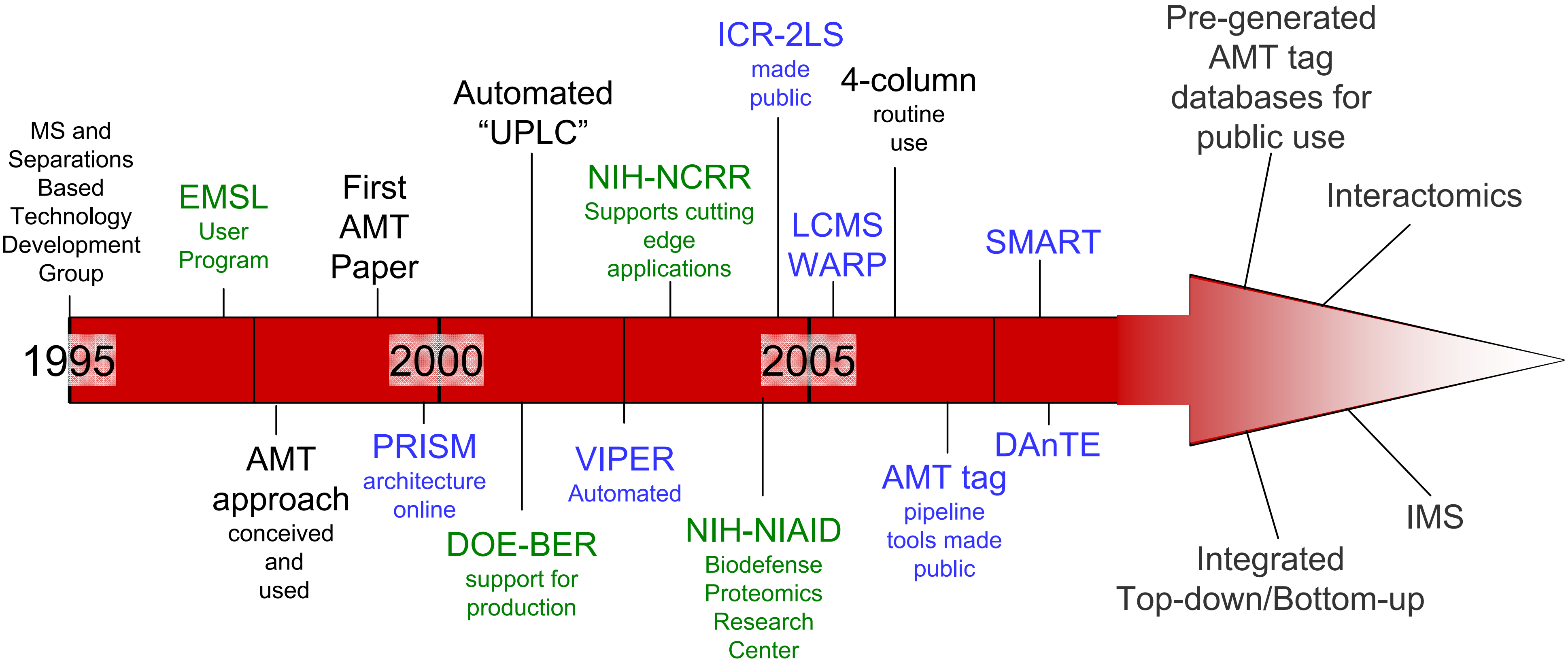
- Metabolic Disease
- Infectious Disease
- Psychiatric Disease
- Inflammatory Disease

Platforms

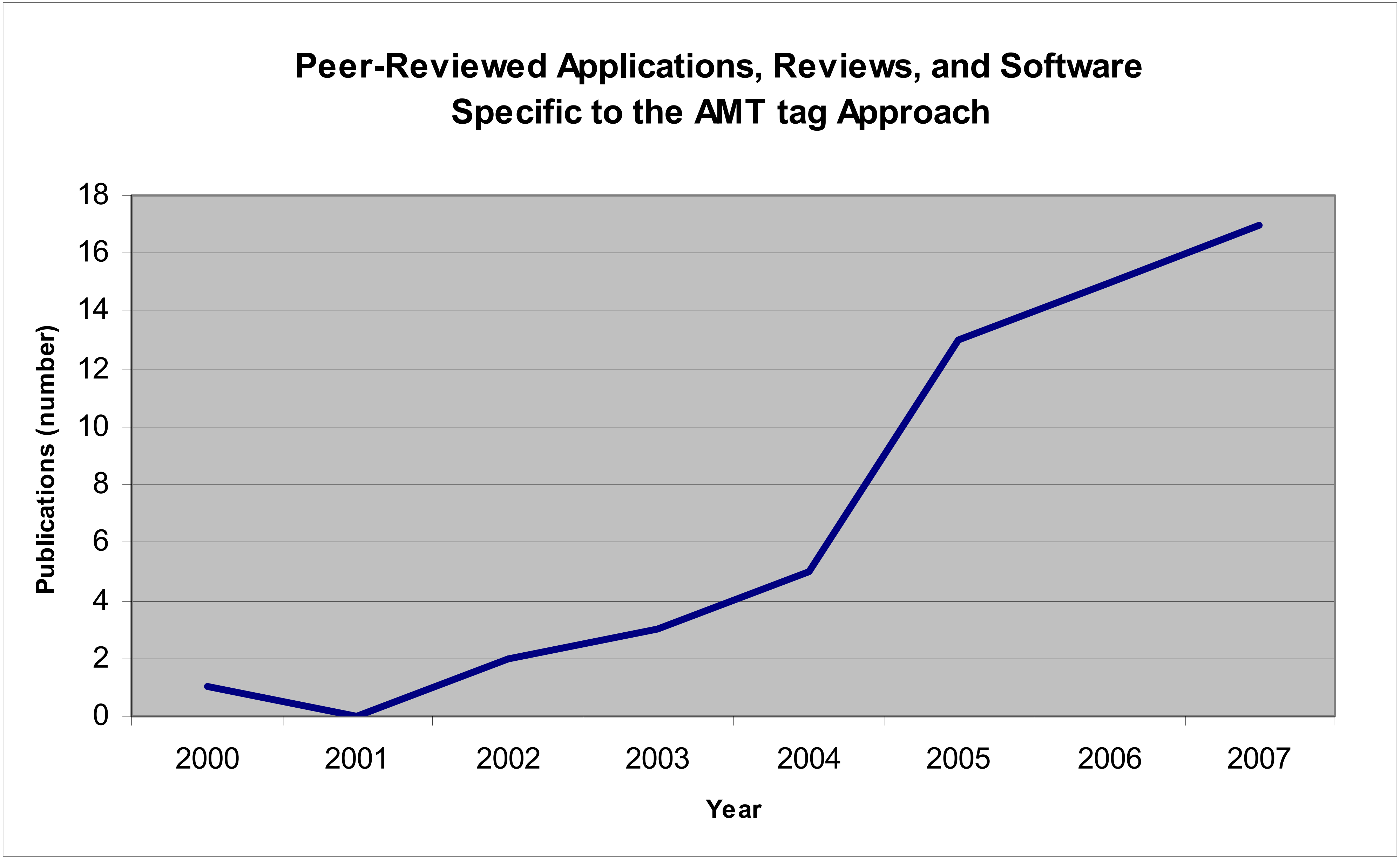
- Sequencing
- Genotyping
- Chemical Synthesis and Screening
- Proteomics and Metabolite Profiling
- Image Analysis

- **Scientific mission:** Create **comprehensive, broadly available tools** for genomic medicine; pioneer **applications** toward disease understanding and treatment
- **Organizational mission:** Enable **collaborative projects** not readily done in individual labs; empower scientists through access to tools and approaches

History/Evolution of PNNL Proteomics


MS and Separations Based Technology Development Group
EMSL User Program
First AMT Paper
Automated "UPLC"
NIH-NCRR Supports cutting edge applications
ICR-2LS made public
LCMS WARP
4-column routine use
SMART
Pre-generated AMT tag databases for public use
Interactomics
1995
2000
2005
AMT approach conceived and used
PRISM architecture online
DOE-BER support for production
VIPER Automated
NIH-NIAID Biodefense Proteomics Research Center
AMT tag pipeline tools made public
DAnTE
Integrated Top-down/Bottom-up
IMS

Key point: early access and experience with higher resolution LC and MS with ~1 ppm mass accuracy



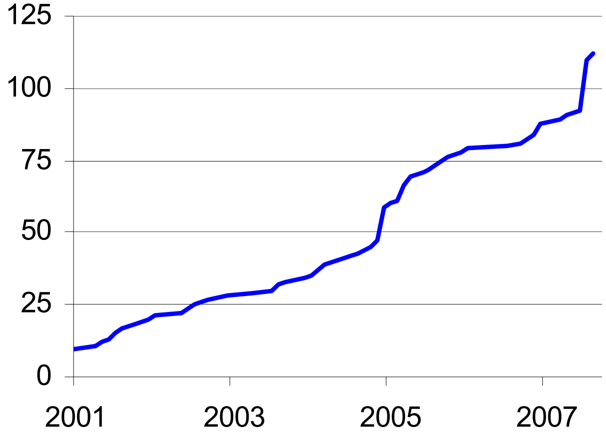
Peer-Reviewed Applications, Reviews, and Software Specific to the AMT tag Approach

Note: excludes non-AMT tag applications papers and excludes broader technology development papers

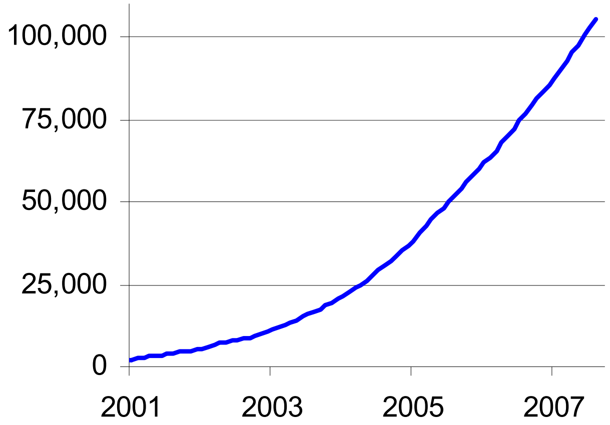
PRISM Data Trends

Organisms	115
Prepared Samples	>50,000
LC-MS(/MS) Analyses	>105,000
Automated Software Analyses	>277,000
Data Files	115 TB
Data in SQL Server databases	1 TB

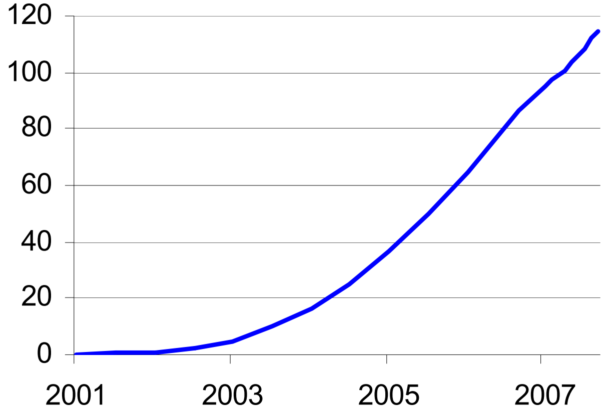
Organisms studied



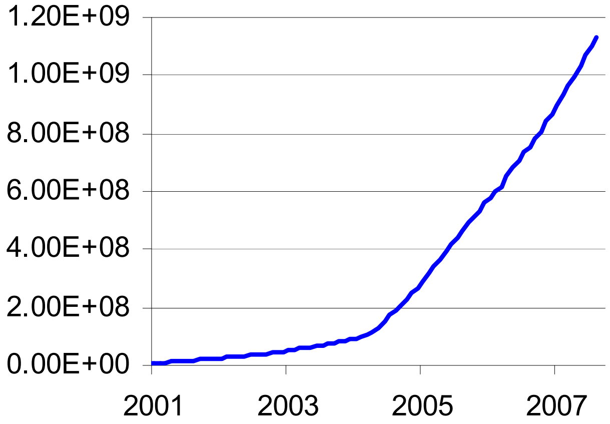
Datasets acquired (instrumental analyses)



TB data stored in PRISM

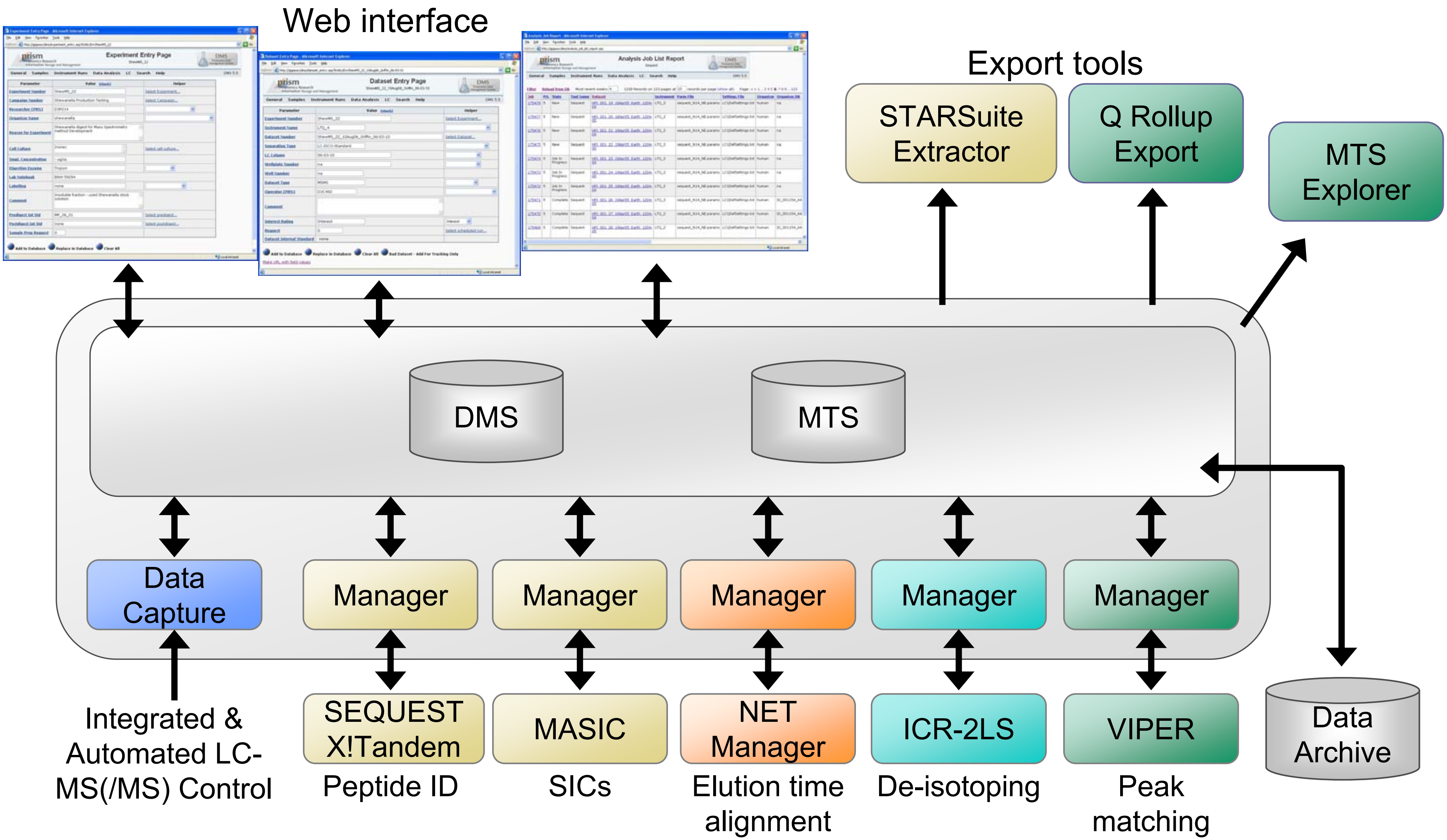


Over 1 billion mass spectra acquired



Proteomics Informatics Architecture

modular and loosely coupled for flexibility

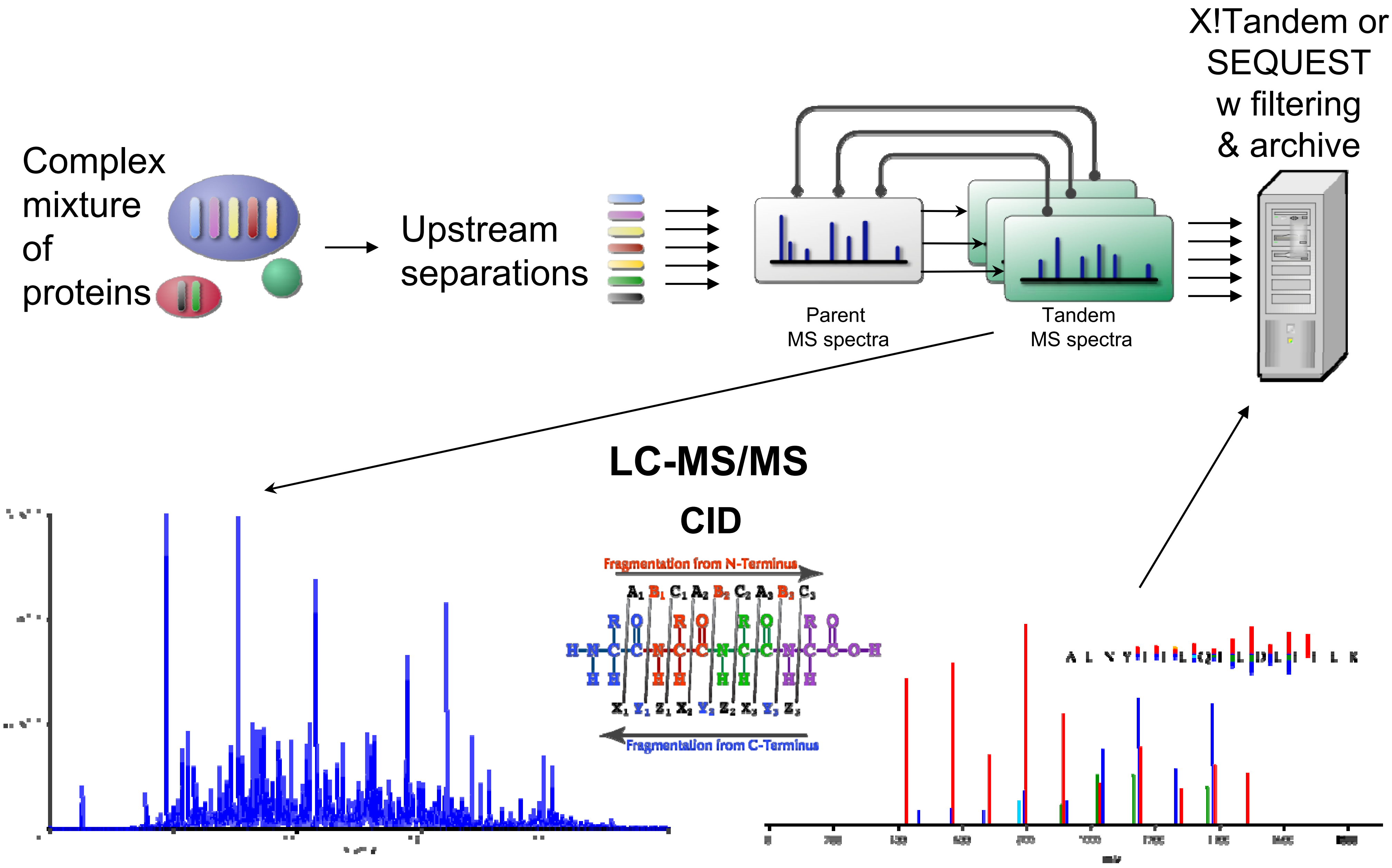


PRISM: G.R. Kiebel et. al. *Proteomics* **2006**, *6*, 1783-1790.

Motivations for LC-MS Based 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
  - Blocking and randomization needs

Shotgun or MuDPIT Proteomics



Complex mixture of proteins

Upstream separations

Parent MS spectra

Tandem MS spectra

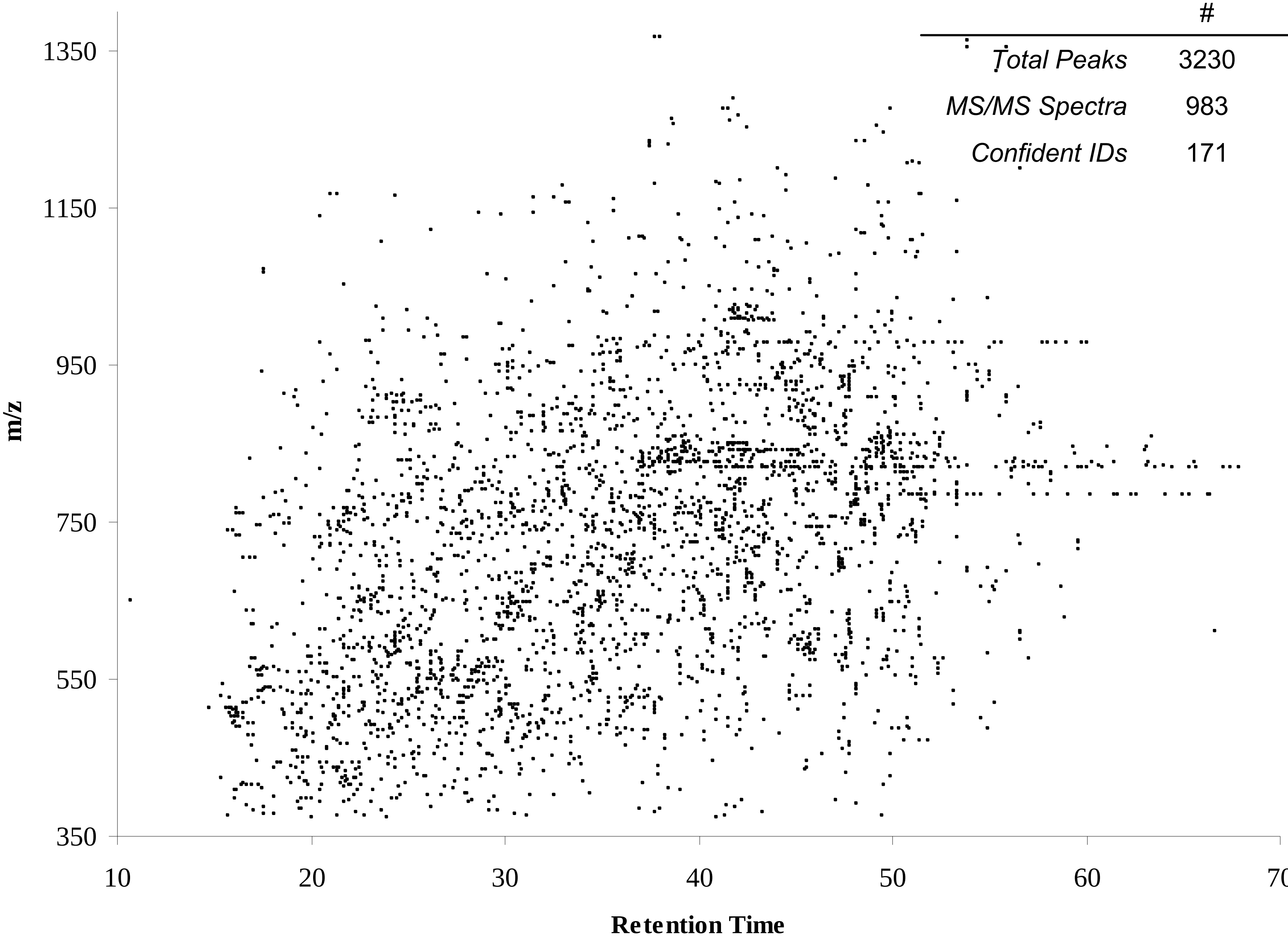
X!Tandem or SEQUEST w filtering & archive

LC-MS/MS

CID

Fragmentation from N-Terminus
A1 B1 C1 A2 B2 C2 A3 B3 C3
H-N-C-C-N-C-C-N-C-C-N-C-C-O-H
R O R O R O R O
H H H H H H H H
X1 Y1 Z1 X2 Y2 Z2 X3 Y3 Z3
Fragmentation from C-Terminus
m/z

LCMS Information Gauntlet

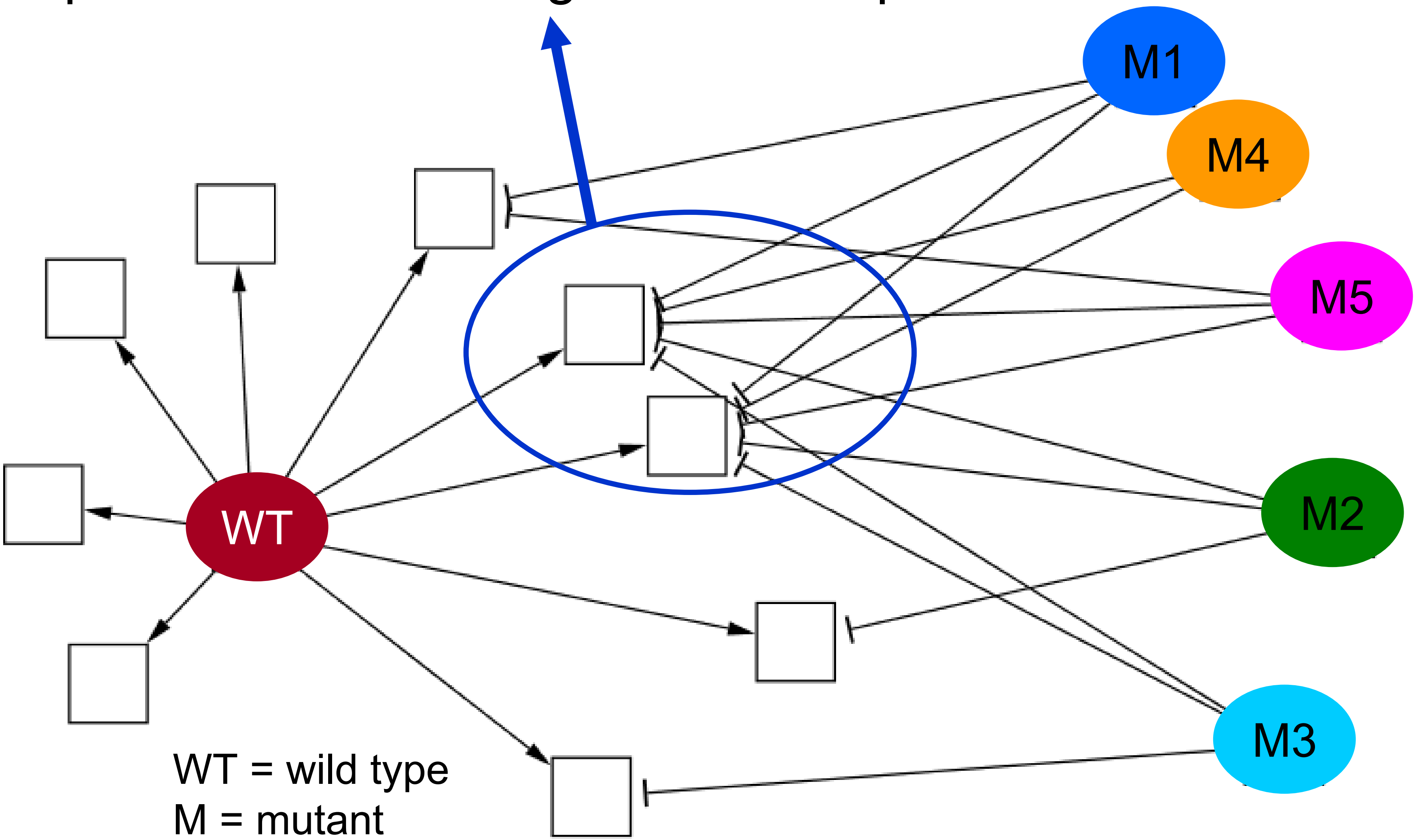


	#	%
Total Peaks	3230	--
MS/MS Spectra	983	30%
Confident IDs	171	5%

Courtesy Jake Jaffe

An example need for increased throughput
Analysis of Regulatory Mutants

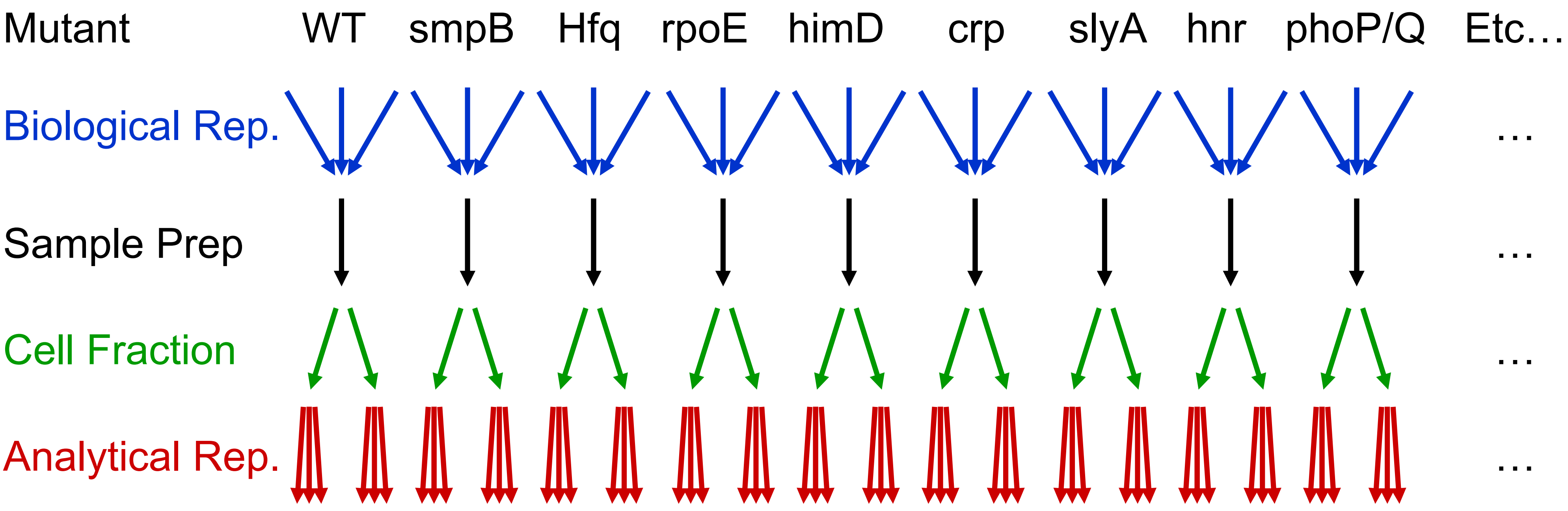
Hypothesis: Knock-out regulatory proteins involved in pathogenesis and the commonly regulated proteins represent the best targets for therapeutics



WT = wild type
M = mutant

Understanding Biological Regulation of *Salmonella*: **Demonstrates the need/use for increased throughput**

Replicate analysis to account for
natural biological and normal analytical variation



Mutant WT smpB Hfq rpoE himD crp slyA hnr phoP/Q Etc…

Biological Rep. …

Sample Prep …

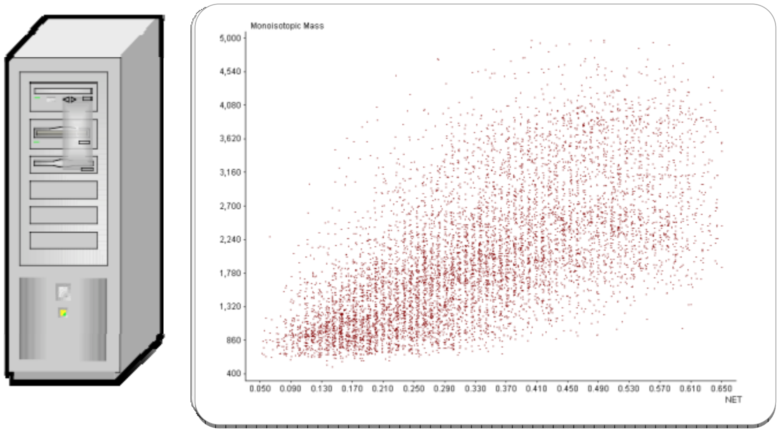
Cell Fraction …

Analytical Rep. …

X4 Contrasting Conditions

1080 analyses for 15 mutants
using biological pooling 360 analyses

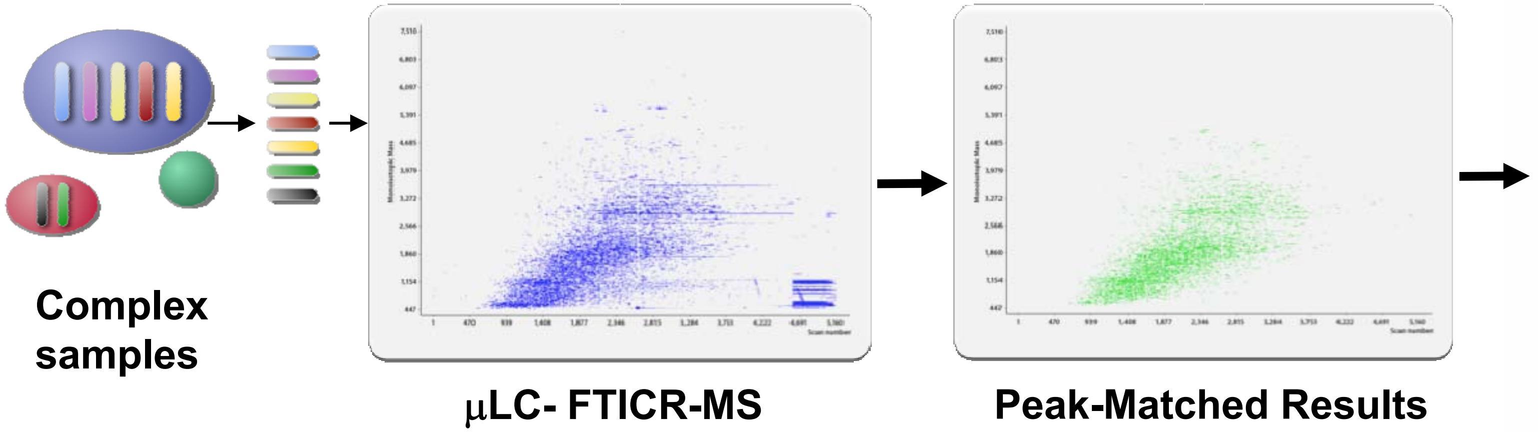
Accurate Mass and Time Tag Approach



SEQUEST and/or X!Tandem results

- Filtering
- Calculate exact mass
- Normalize observed rlution time

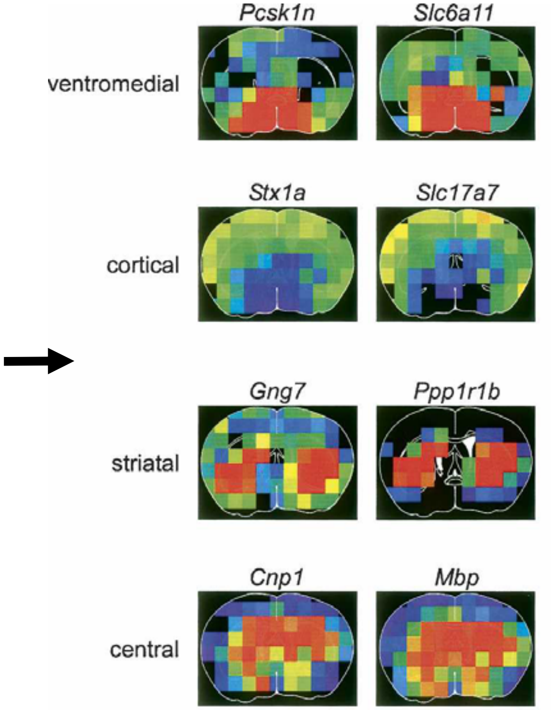
High-throughput LC-FTICR-MS Analysis (AMT) tag



Complex samples

μLC- FTICR-MS

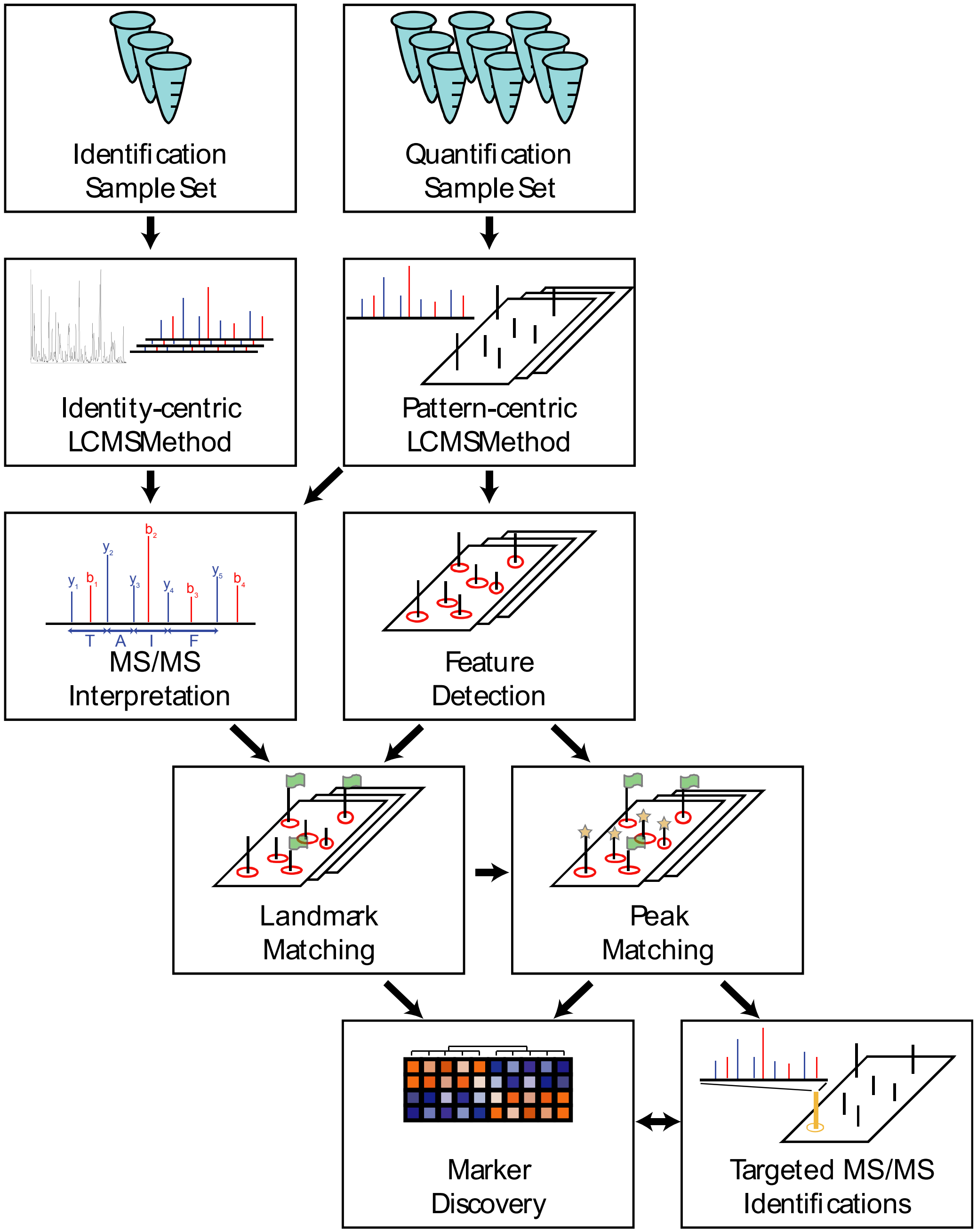
Peak-Matched Results



Compare abundances across samples

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

PEPPeR Pipeline



Identification Sample Set

Quantification Sample Set

Identity-centric LCMSMethod

Pattern-centric LCMSMethod

MS/MS Interpretation

Feature Detection

Landmark Matching

Peak Matching

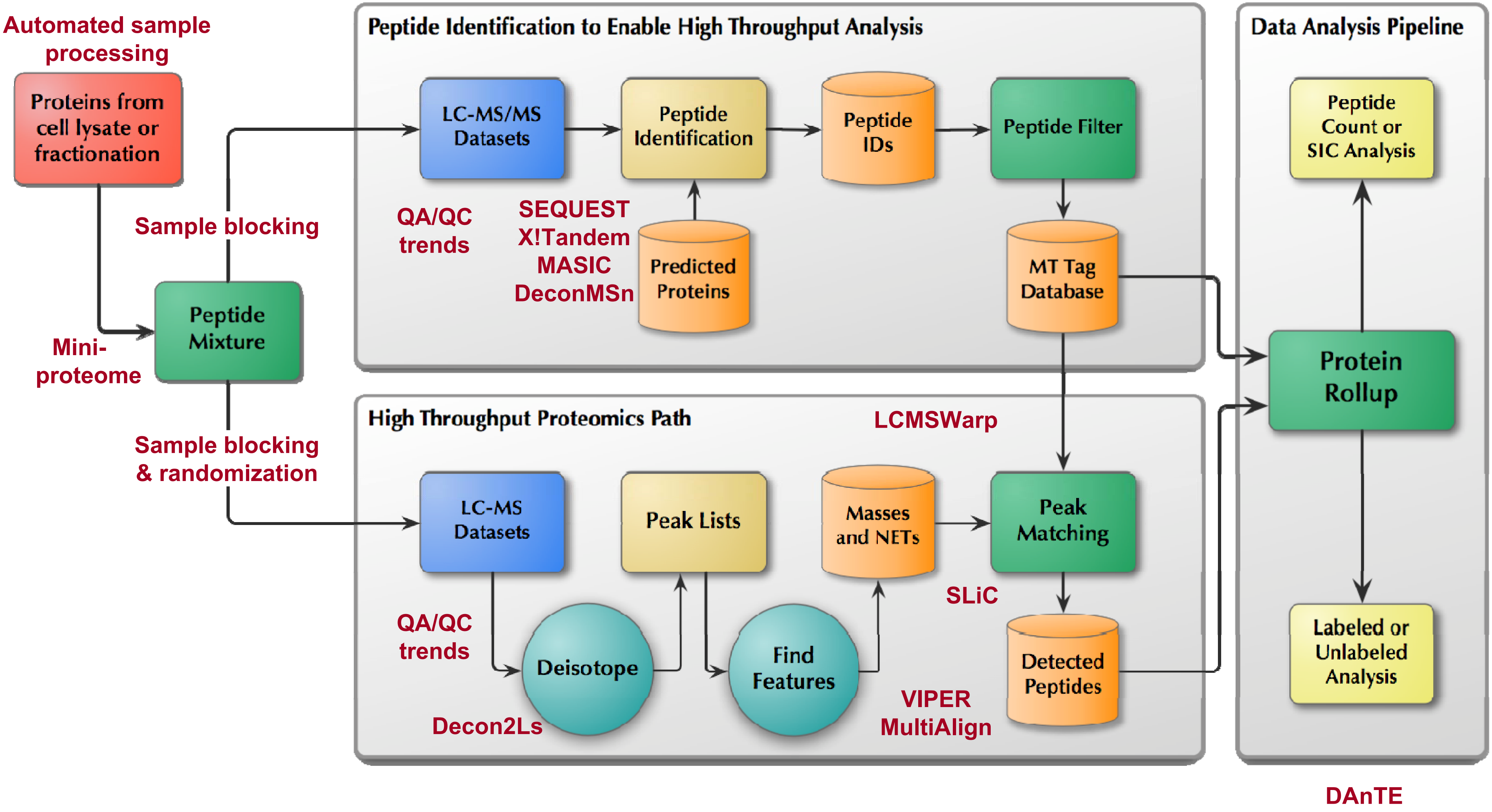
Marker Discovery

Targeted MS/MS Identifications

New Concerns with Larger Comparisons

- 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

Accurate Mass and Time (AMT) Tag Data Processing Pipeline



Automated sample processing

Proteins from cell lysate or fractionation

Sample blocking

Mini-proteome

Peptide Mixture

Sample blocking & randomization

Peptide Identification to Enable High Throughput Analysis

LC-MS/MS Datasets

Peptide Identification

Peptide IDs

Peptide Filter

QA/QC trends

SEQUEST
X!Tandem
MASIC
DeconMSn

Predicted Proteins

MT Tag Database

High Throughput Proteomics Path

LCMSWarp

LC-MS Datasets

Peak Lists

Masses and NETs

Peak Matching

QA/QC trends

Deisotope

Find Features

SLiC

Detected Peptides

Decon2Ls

VIPER
MultiAlign

Data Analysis Pipeline

Peptide Count or SIC Analysis

Protein Rollup

Labeled or Unlabeled Analysis

DAnTE

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

  J.N. Adkins, et. al. *Mol. Cell. Proteomics*. **2006**, *5* (8), 1450-1461.

- Analysis of purified viral particles of Monkeypox and Vaccinia viruses

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

- 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.

- Jake Jaffe will expand on a couple of examples such as primary tissue example; quantities too small for labeling

Course Related Software & Data

AMT tag Pipeline Software

PROTEOMICS
National Center for Research Resources

http://ncrr.pnl.gov

PEPPeR, software within GenePattern

GenePattern
A platform for integrative genomics

http://www.broad.mit.edu/cancer/software/genepattern/

PNNL's LCMS-based data repository

Biological MS Data and
Software Distribution Center

http://omics.pnl.gov

Currently in open beta-testing
>1 Terabyte available
More coming soon!

Salmonella typhimurium data resource

Resource Center for
Biodefense Proteomics Research

http://www.proteomicsresource.org

Other 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)

Example Data for the AMT tag Pipeline Demo

- *Salmonella typhimurium*, LC-MS/MS
  - Grown in LB (Luria-Bertani) up to log phase
  - Soluble portion of cell lysis
  - “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-MS
  - Different sample, grown and prepared in the same conditions
- LC-FTICR-MS analysis (11T FTICR)
  - Non-linear alignment and peak matching to the database
- DAnTE data
  - Similar experiment with new growth condition

Course Outline

- Introduction (Adkins)
- Part I: Overview of Label-Free Quantitative Proteomics (Jaffe)
  - When and why to use label free quantitative proteomics
  - Overview of the generic ‘label free’ pattern-based approach with guidelines
  - Discussion of alternate pipelines
- Part II: Feature discovery in LC-MS datasets (Monroe and Jaitly)
- Part III: PEPPeR, GenePattern and Real-world examples (Jaffe)
- Break
- AMT tag Pipeline Demo (general)
- Panel Discussion
  - Questions
  - Future Directions

Part I: An Overview of Label-free quantitative proteomics

Jacob D. Jaffe

The Broad Institute of Harvard and MIT

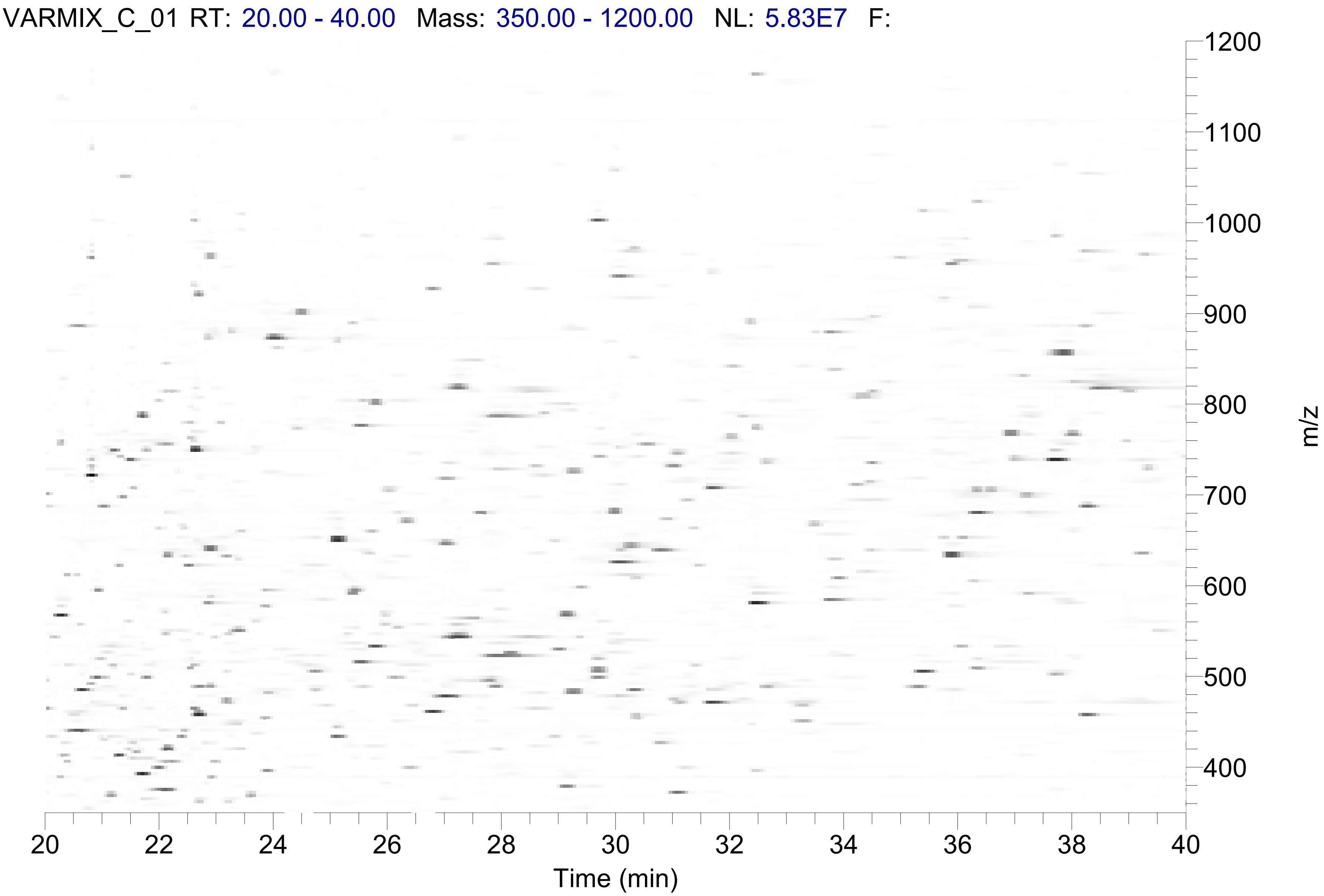
Proteomics Platform

Section Outline

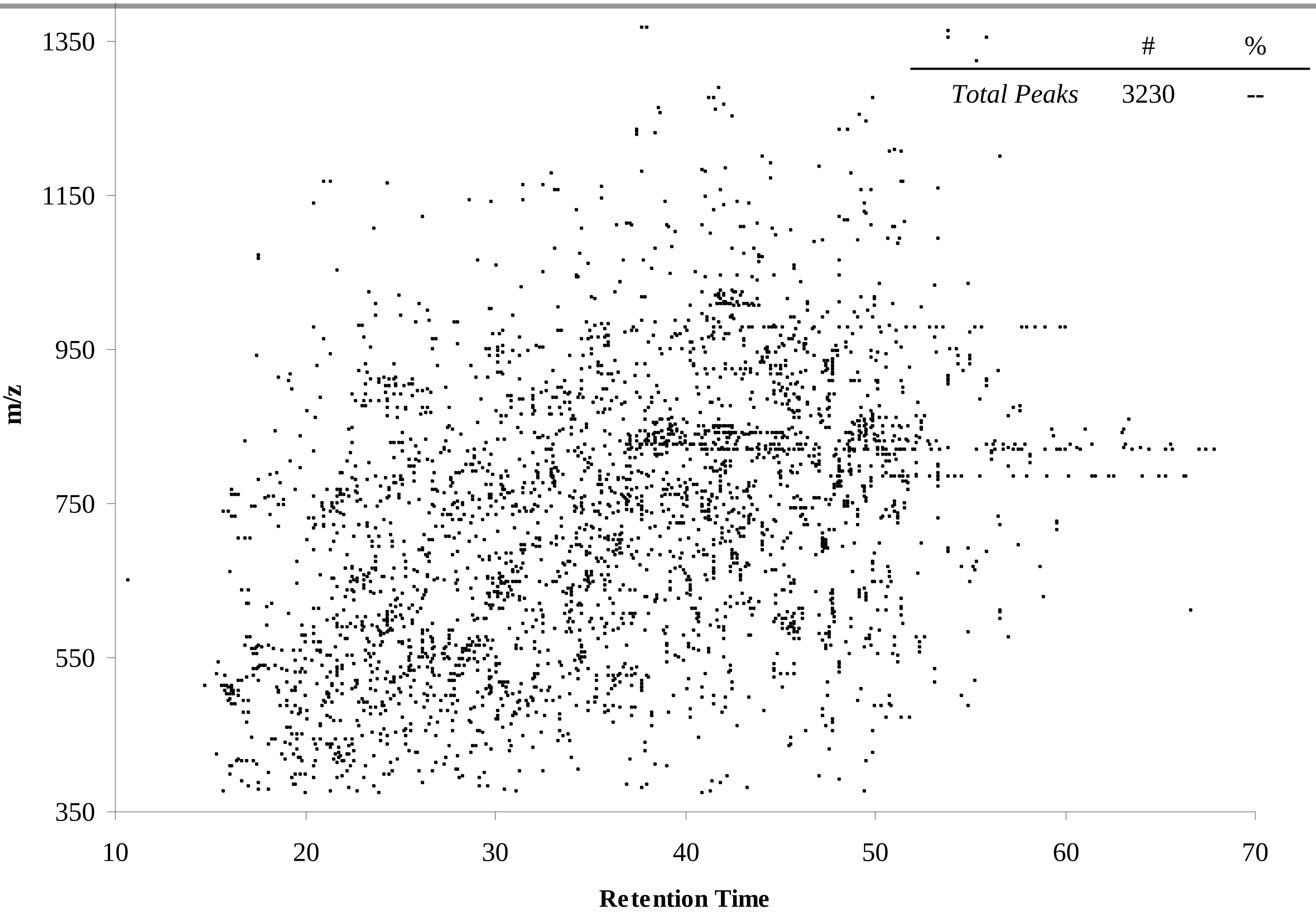
- When and why to use label free quantitative proteomics

- Overview of the generic ‘label free’ pattern-based approach with guidelines

A picture is worth 1000 parameters…

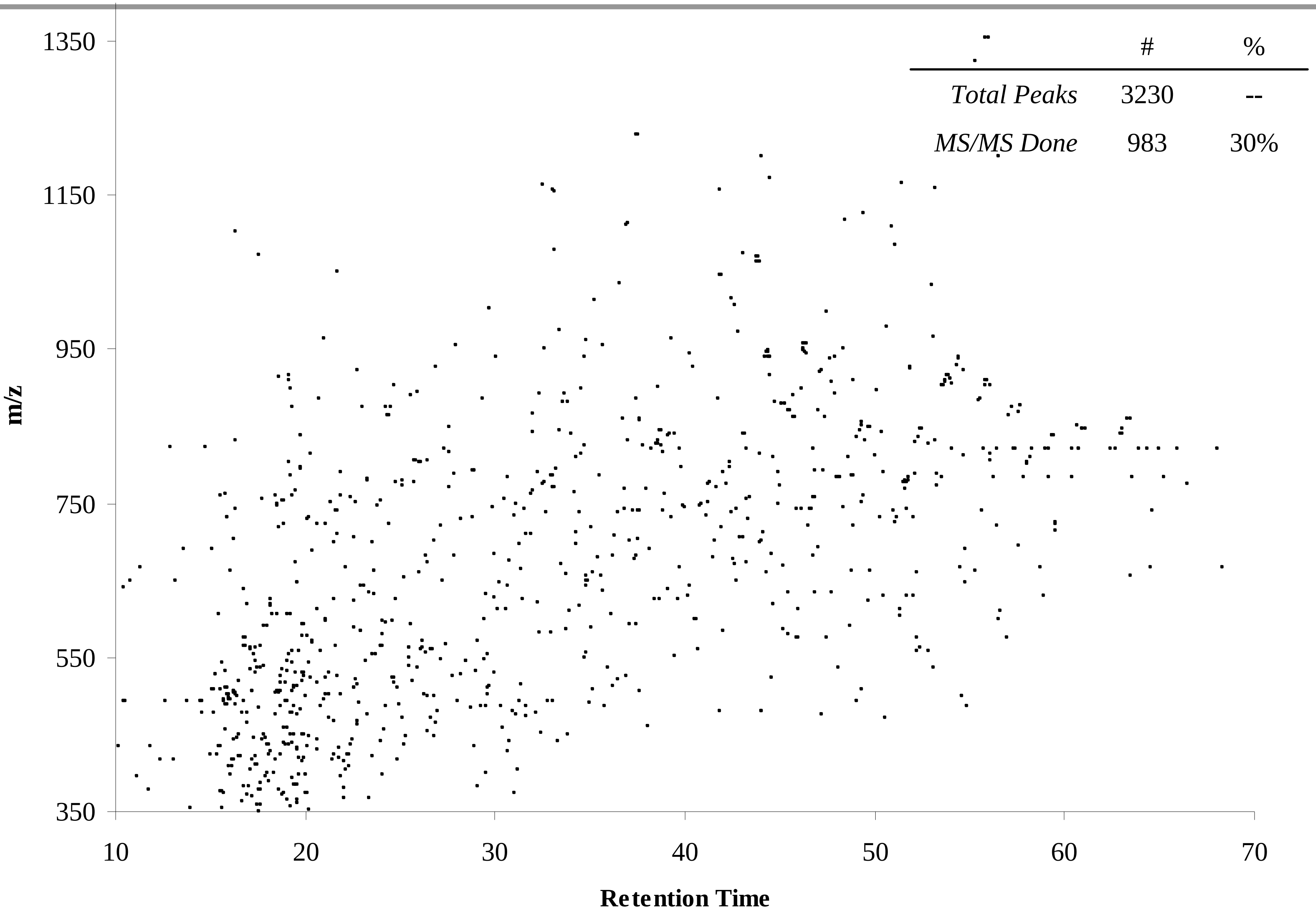


LCMS Information Funnel – Total Peaks



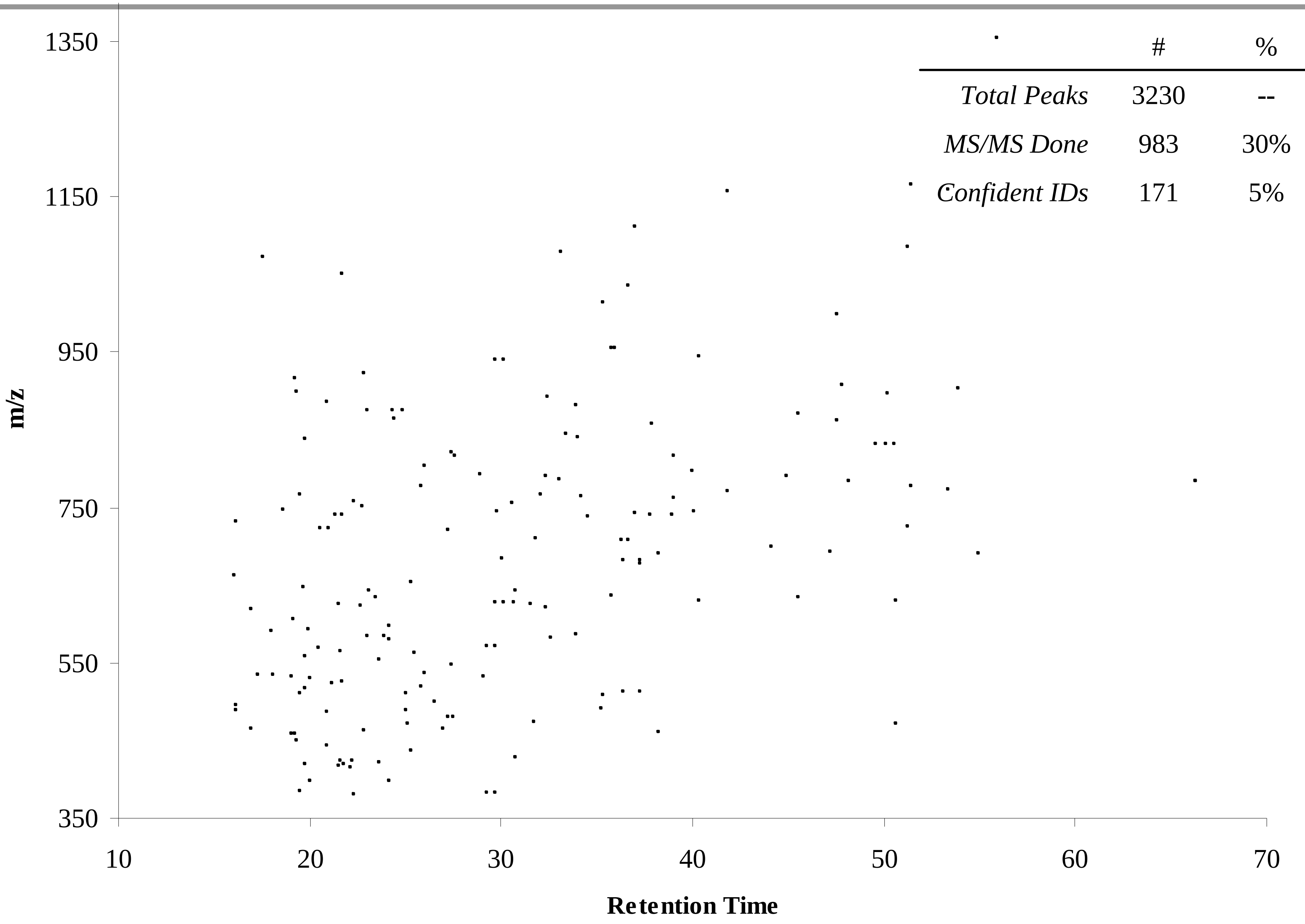
	#	%
Total Peaks	3230	--

LCMS Information Funnel – MS/MS Sampling



	#	%
Total Peaks	3230	--
MS/MS Done	983	30%

LCMS Information Funnel – MS/MS Identified



	#	%
Total Peaks	3230	--
MS/MS Done	983	30%
Confident IDs	171	5%

Definition of Label-free Quantitative Proteomics

- Use of raw mass spectral signal intensity (peaks) as a surrogate for the abundance of a peptide and/or protein

- Signal intensity from the same analyte is compared across multiple experimental conditions as the basis for quantitation

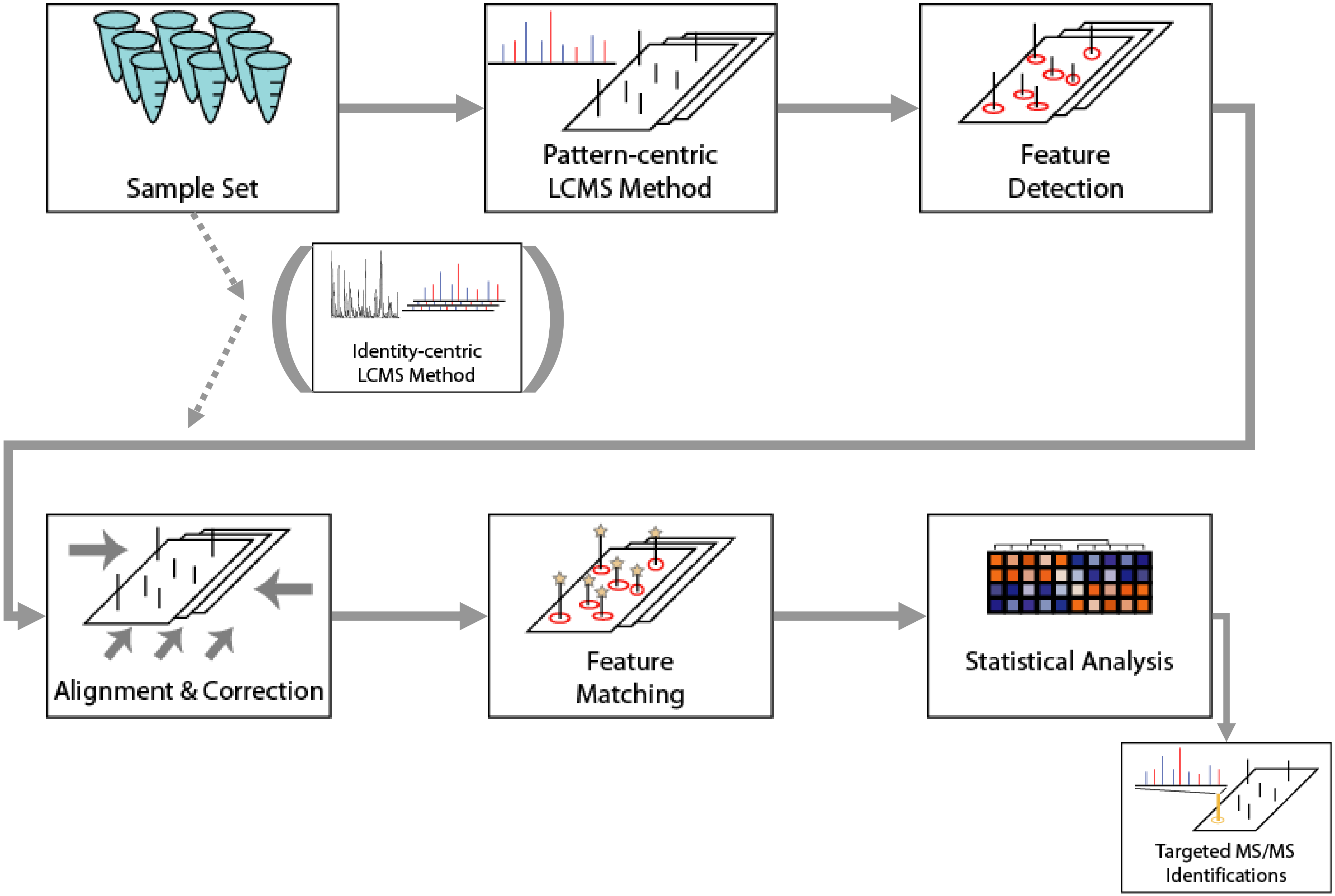
- When coupled to LC, peaks have dimensions in retention time and well as *m/z* and intensity

- Careful experimental processing and computational methods are required to extract quantitative information in label-free proteomics

Motivations for Label-free Quantitative Proteomics

- Microarray envy
  - Well-defined experiment, well-defined tools
- Differential detection and *quantification* of proteins
  - Biomarker discovery and pattern recognition
  - Biological insight into the real actors in the cell: proteins
  - Time course analysis
- MS^2 independent but friendly
  - SILAC and iTRAQ (labeling methods) require MS^2 ID for entry
  - Comprehensive! Quantify all the spots! Even the faint ones!
- Minimal sample workup
  - Primary tissue OK
  - No artifacts from labeling efficiency

The Generic 'Label-free' Workflow



Sample Set

Pattern-centric LCMS Method

Feature Detection

Identity-centric LCMS Method

Alignment & Correction

Feature Matching

Statistical Analysis

Targeted MS/MS Identifications

Best Practices: Getting Started

- Team approach
  - LCMS expert experimentalist
  - Computer scientist/programmer
  - Statistician
- Planning
  - Statistical power of study (consult statistician)
  - Identification of reliable sample sources
  - Instrument / Computational / Storage infrastructure
- Execution
  - Patience
  - Consistency

Best Practices: Samples

- BEST POSSIBLE **SAMPLES** AND **CONTROLS**
  - Relevant to disease or study target
  - Proximal to the source of differential markers
  - Consistent in composition
  - Controls appropriately matched (same subject if possible)
  - Enriched in likely differential markers
  - GARBAGE IN, GARBAGE OUT
- Sample processing pipeline TESTED and CONSISTENT
  - Abundant protein depletion (serum proteomics)?
  - Fractionation required?
  - Measure yields – are they consistent?
  - CLEAN!!!
- Collect more than you need – outlier removal!

Best Practices: Data Acquisition

- Resolution! Resolution! Resolution!
  - FTICR or Orbitrap recommended > 60,000 resolution
  - More ‘channels’
- Accuracy! Accuracy! Accuracy!
  - Calibrate mass often
  - Downstream recognition of “same” feature easier
  - Statistical confidence
- Consistency
  - LCMS methods and instrumentation
  - LC column and length
- Common Sense
  - MS^1 sampling rate -> chromatographic resolution
  - Tolerances and dynamic exclusion for MS^2 sampling
  - Carry over testing and sample randomization
  - SAVE THE SAMPLES!!!!!

Best Practices: Feature Picking

- Understand the method
  - No method is demonstrably ‘best’
  - Consult with expert help
  - All methods have parameters and tolerances that have to be tailored to your operating characteristics
  - There is no magic ‘black box’

- Patience
  - You will spend a long time collecting data; expect to spend at least as much time extracting and analyzing data
  - Budget time and resources to explore parameters on a subset of your data before doing feature picking *en masse*

Best Practices: Experiment Alignment

- Consistency in experimental execution
    - Makes life easier, less computational correction
- Pay attention to output of aligners
    - Methods may have metrics of alignment quality
    - Large corrections may signal outlier experiments
        - Consider discarding
- Intensity normalization
    - Total ion current (TIC)?
    - TIC of all features?
    - Subset of ‘housekeeping’ features?
    - Medians, means, etc?

Best Practices: Feature Assignment and Matching

- Assignment: annotation of an LCMS feature with a peptide identity (sequence)
  - Derived from external or embedded MS^2 data that has been searched against a database (i.e. Sequest, Mascot, etc)
  - AMT-based assignment (importance of mass accuracy)
  - Look for statistics!

- Matching: recognition that a feature is the same across multiple experiments irrespective of an identity assignment
  - Assignments can help
  - Accuracy and alignment are paramount
  - Take care with user-adjustable tolerances
  - Look for statistics!

Best Practices: Statistical Analysis

- Intensity normalization of features must be done prior to statistical analysis
  - Also address handling of missing values

- Understand what you are doing or seek assistance
  - Know your *p*-values from your *q*-values (and FDRs)

- Have a well-formulated statistical question
  - Most statistical tests are measured vs. the ‘null’ hypothesis
  - Decide in advance what levels of false discovery are acceptable
  - Significance level ≠ priority for follow-up

- There are many tools available
  - Some are more proteomics-amenable
    - Handling of intensity normalization
    - Handling of proteins as combinations of peptides

Best Practices: Following-up

- Targeted reinterrogation of samples for identification of ‘unidentified’ features

- Literature mining
  - Possible connections to your biological questions
  - Helps with prioritization

- Targeted assessment of interesting features in alternative matrices
  - I.e., discovered in tissue, but is it present in blood?
  - Methods other than mass spec, too!

Reference Chart of Label-free Platforms

	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

Part II: LC-MS Feature Discovery

- Introduction (Adkins)
- Part I: Overview of Label-Free Quantitative Proteomics (Jaffe)
- Part II: Feature discovery in LC-MS datasets (Monroe and Jaitly)
  - 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
- Part III: PEPPeR, GenePattern and Real-world examples (Jaffe)
- Break
- AMT tag Pipeline Demo (general)
- Panel Discussion
  - Questions
  - Future Directions

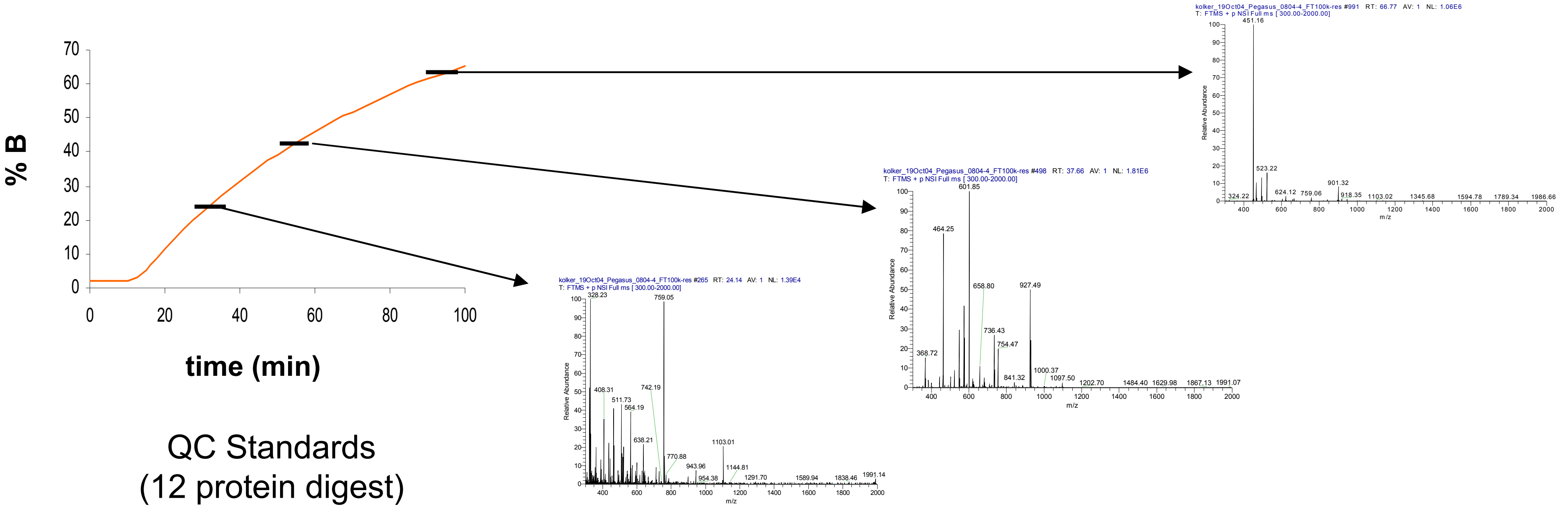
Part II: Feature discovery in LC-MS datasets

Navdeep Jaitly and Matthew E. Monroe

Pacific Northwest National Laboratory

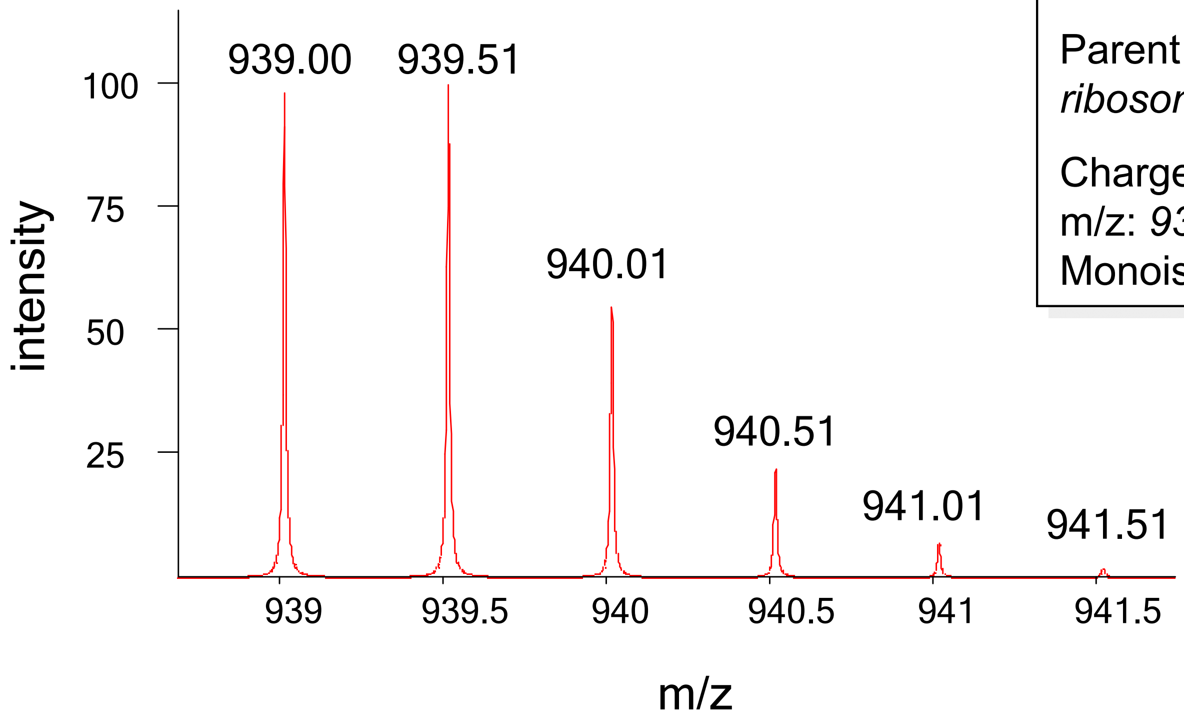
Structure of LC-MS Data

- 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 isotopic pattern in a mass spectrum
  - The pattern is dependent on the compound’s chemical composition, charge, and resolution of instrument



Theoretical Profile

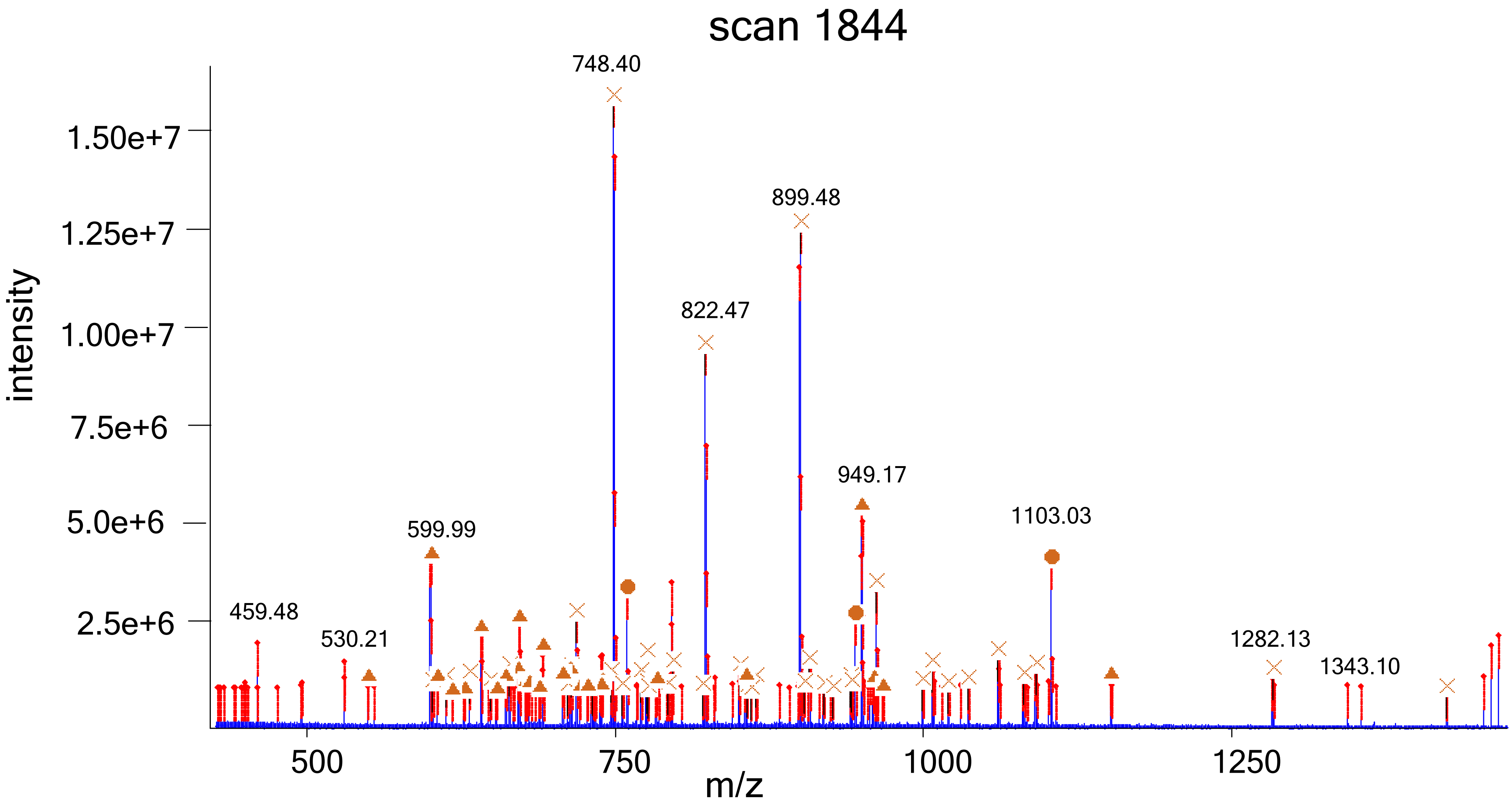
Peptide: *VKHPSEIVNVGDEINVK*

Parent Protein: *gi|16759851 30S ribosomal protein S1*

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

Structure of LC-MS Data

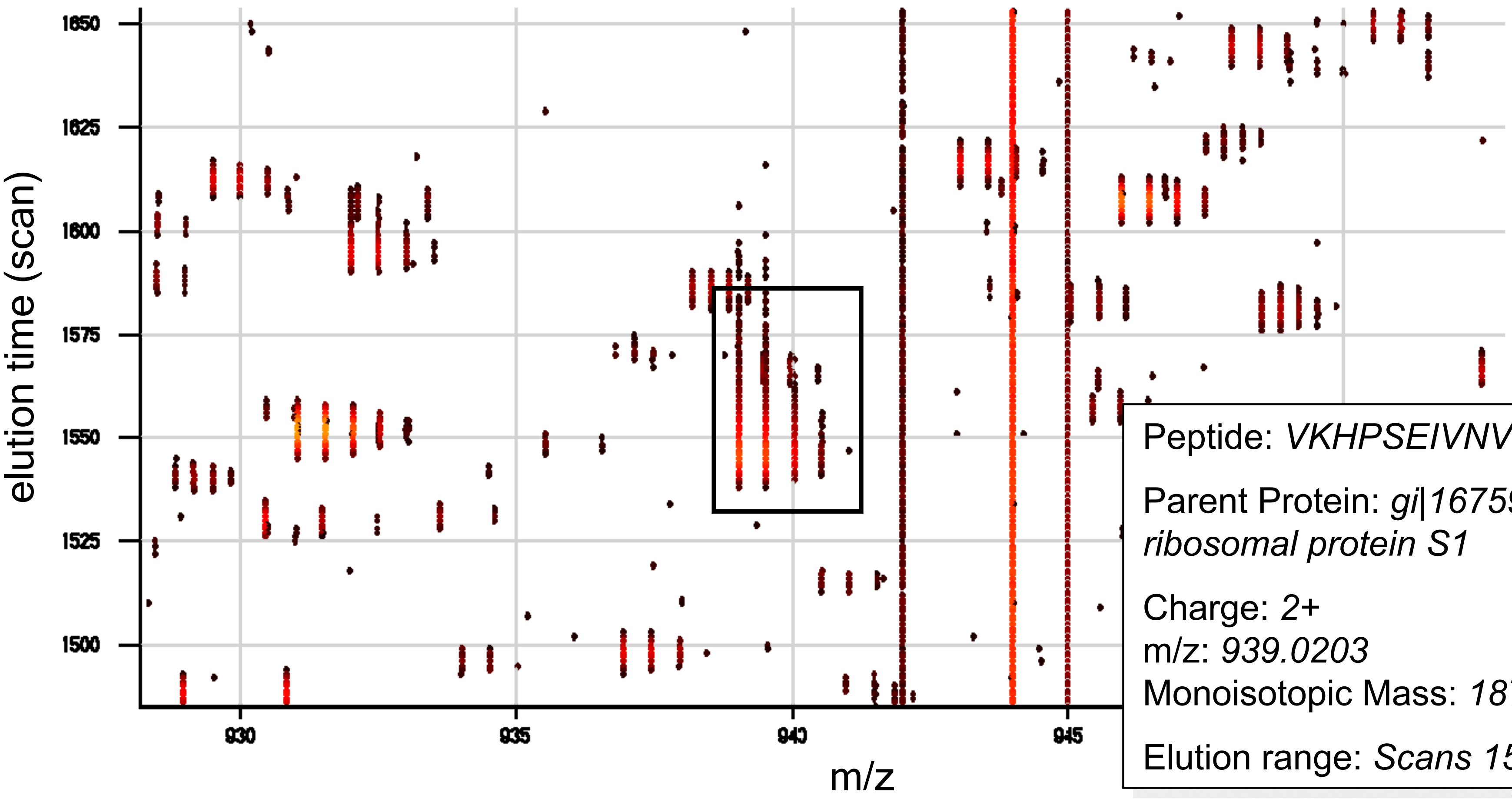
- A mass spectrum of a complex mixture contains overlaid distributions of several different compounds



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

Salmonella typhimurium dataset



Peptide: *VKHPSEIVNVGDEINVK*

Parent Protein: *gi|16759851 30S ribosomal protein S1*

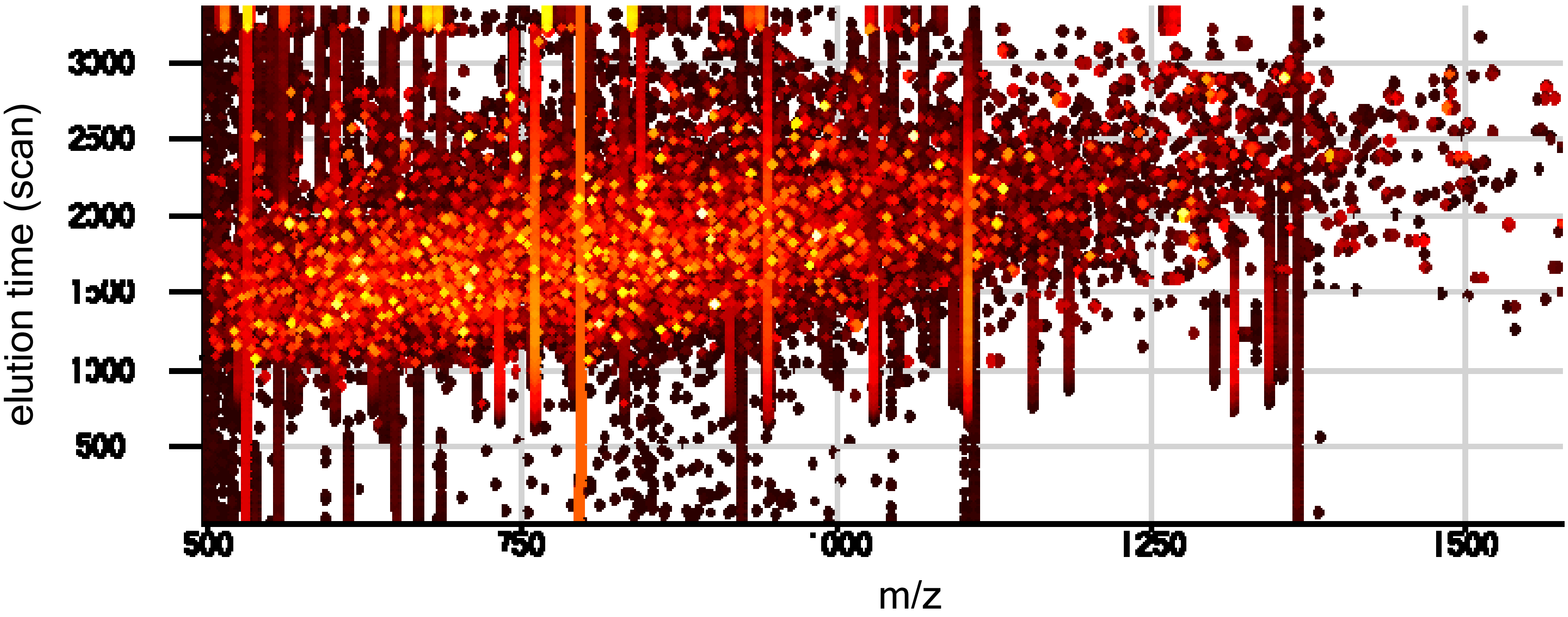
Charge: 2+
m/z: *939.0203*
Monoisotopic Mass: *1876.0054 Da*

Elution range: *Scans 1539 - 1593*

Feature Discovery in LC-MS data

- Goal: Infer *(mass, elution time, intensity)* of compounds that are present in data obtained from an LC-MS dataset
  - Compounds are termed LC-MS features since they are inferred from a three dimensional pattern, yet identity is unknown

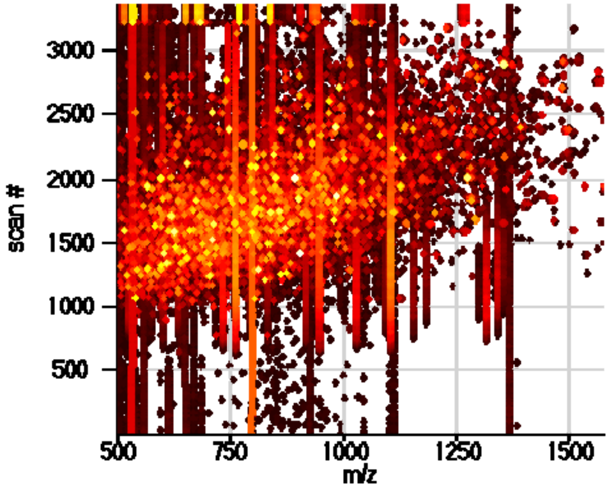
2D view of an LC-MS analysis of Salmonella typhimurium

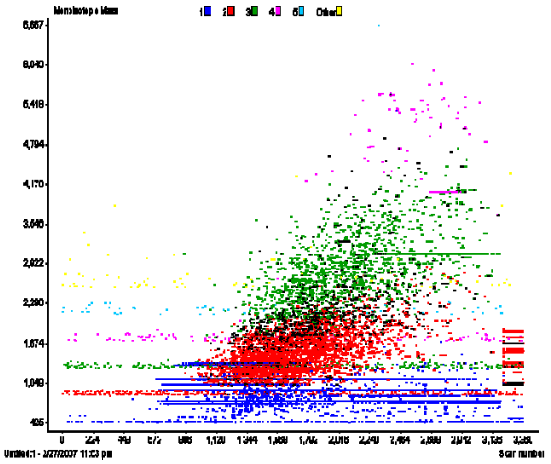


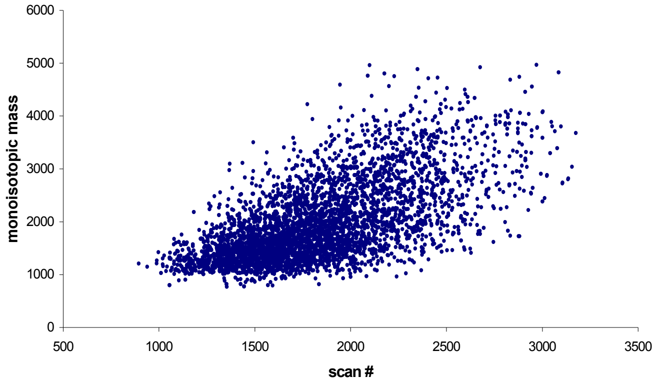
Feature Discovery in LC-MS data

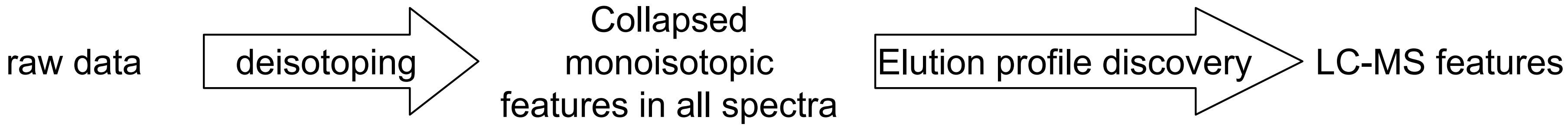
- Sequential process of finding features in each mass spectrum is followed by grouping of features over multiple spectra together

2D views of an LC-MS dataset in different stages of processing







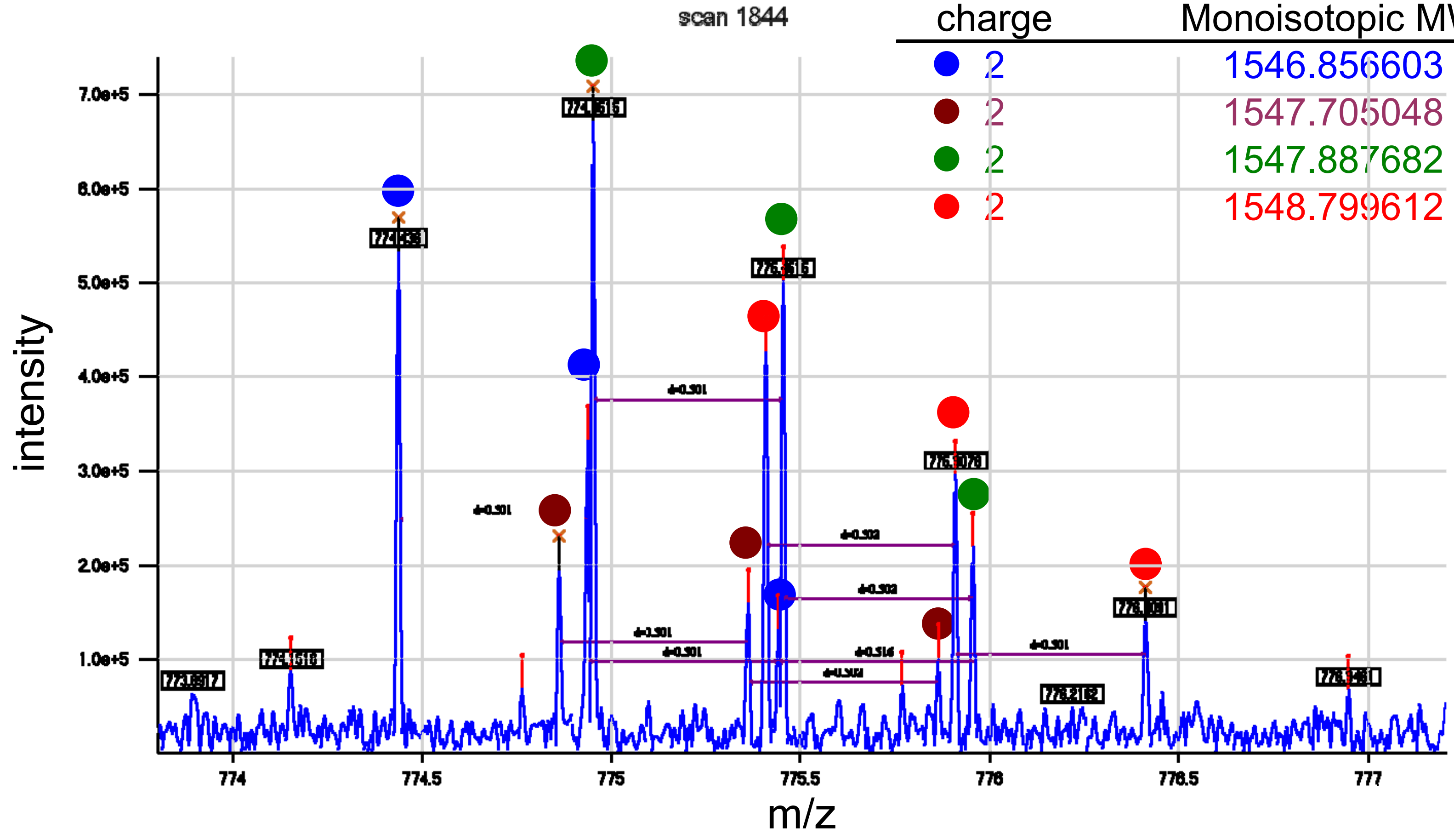


raw data → deisotoping → Collapsed monoisotopic features in all spectra → Elution profile discovery → LC-MS features

Feature discovery in individual spectra

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

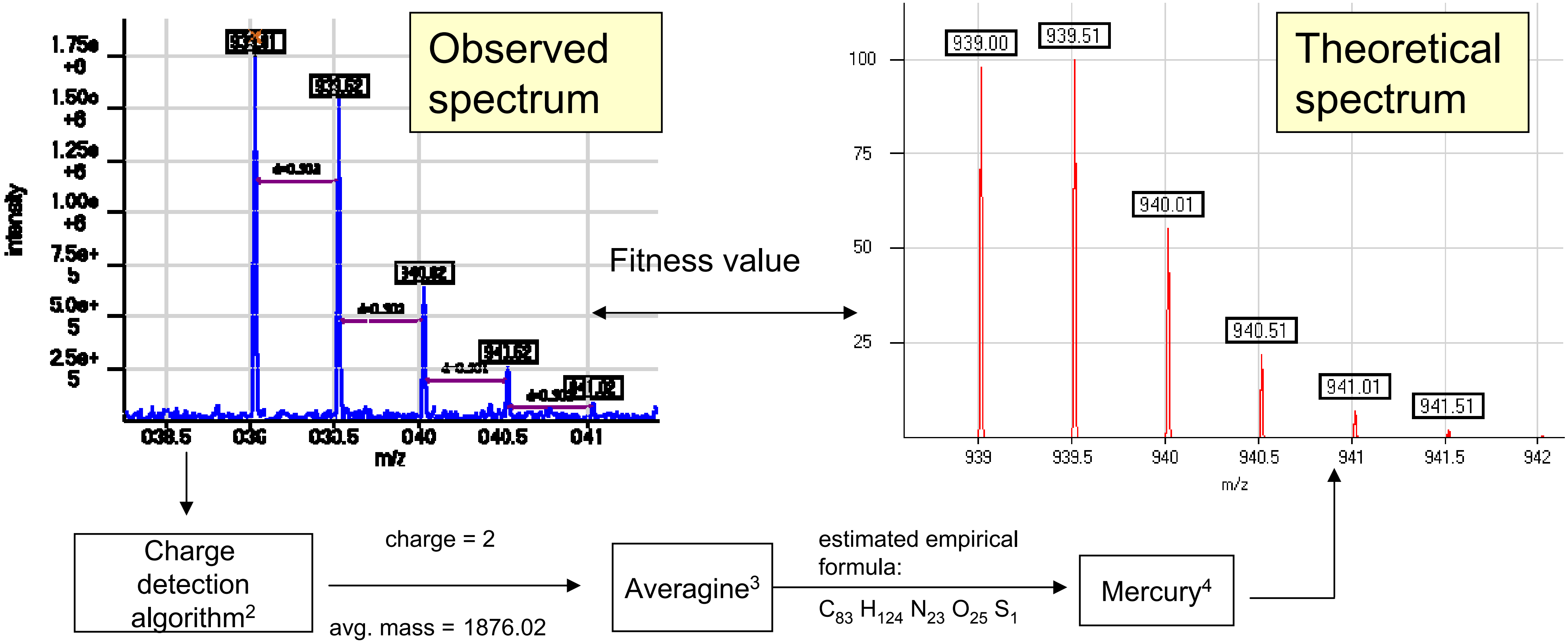
Deisotoping a mass spectrum of 4 overlapping species



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



Charge detection algorithm[2]

charge = 2

avg. mass = 1876.02

Averagine[3]

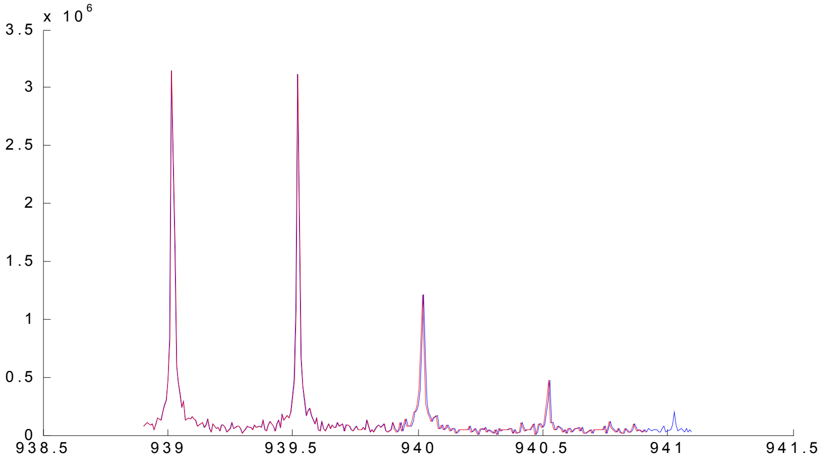
estimated empirical formula:

$C_{83} H_{124} N_{23} O_{25} S_1$

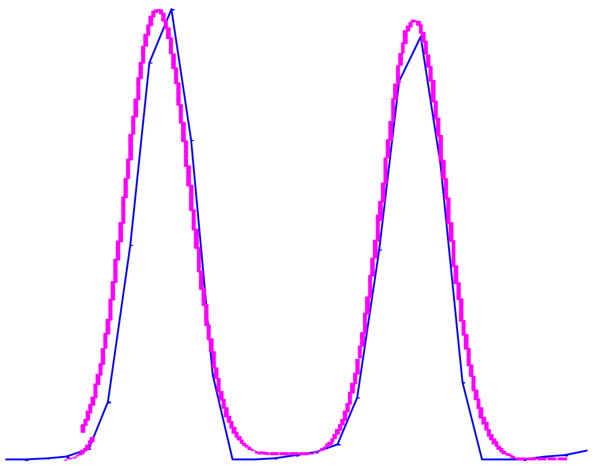
Mercury[4]

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



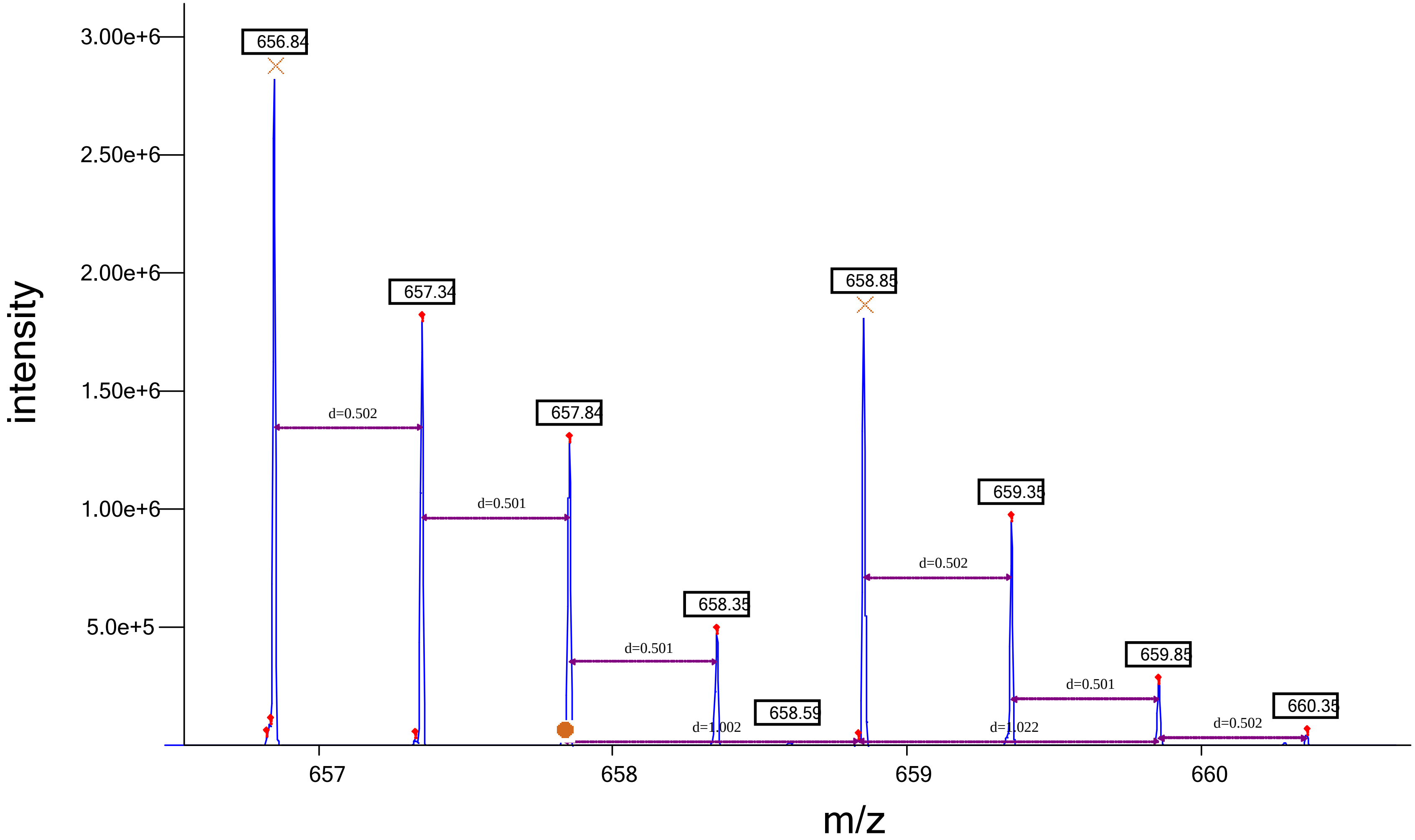
- Patterson (Autocorrelation) algorithm to detect charge of a peak in a complex spectrum
- Mercury algorithm used to guess an average empirical formula for a given mass
  - Averagine empirical formula of $C_{4.9384}$ $H_{7.7583}$ $N_{1.3577}$ $O_{1.4773}$ $S_{0.0417}$ → C_{83} H_{124} N_{23} O_{25} S 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/

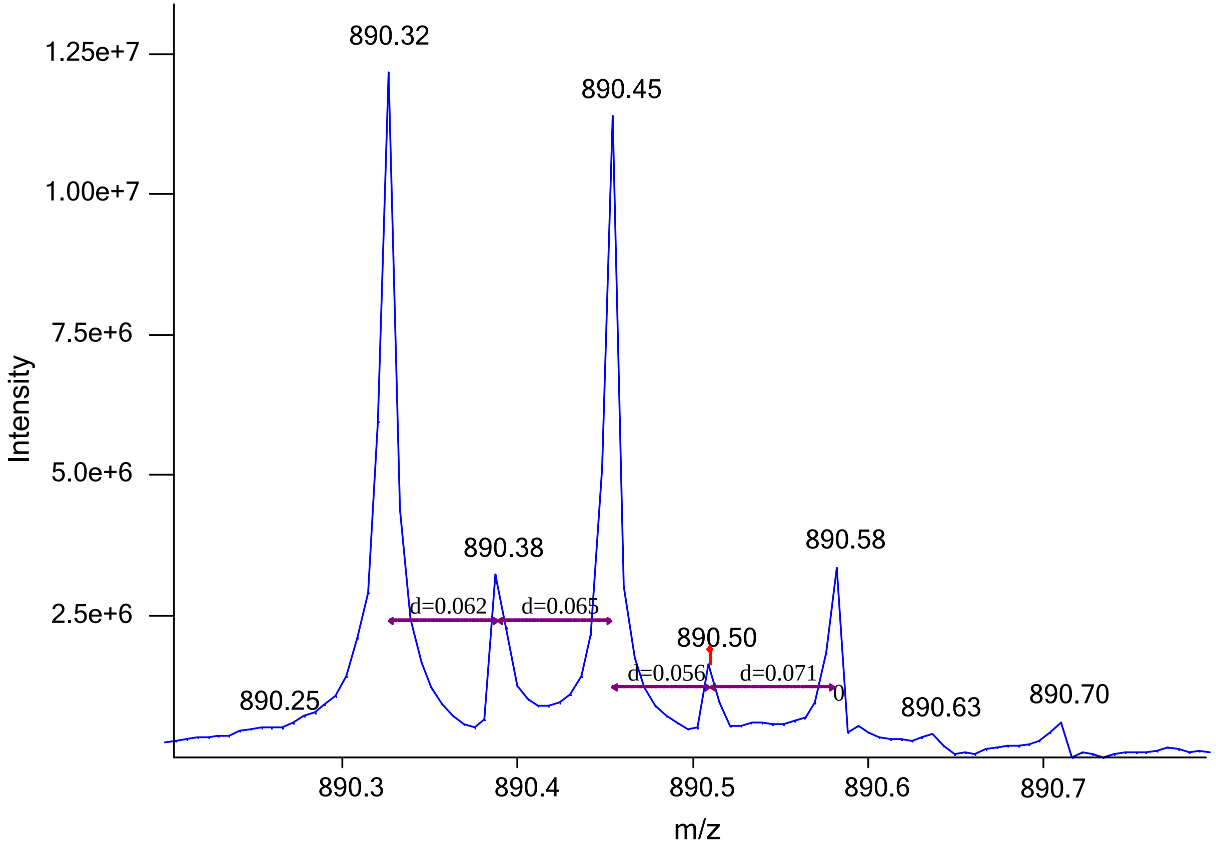
$^{16}O/^{18}O$ Mixtures

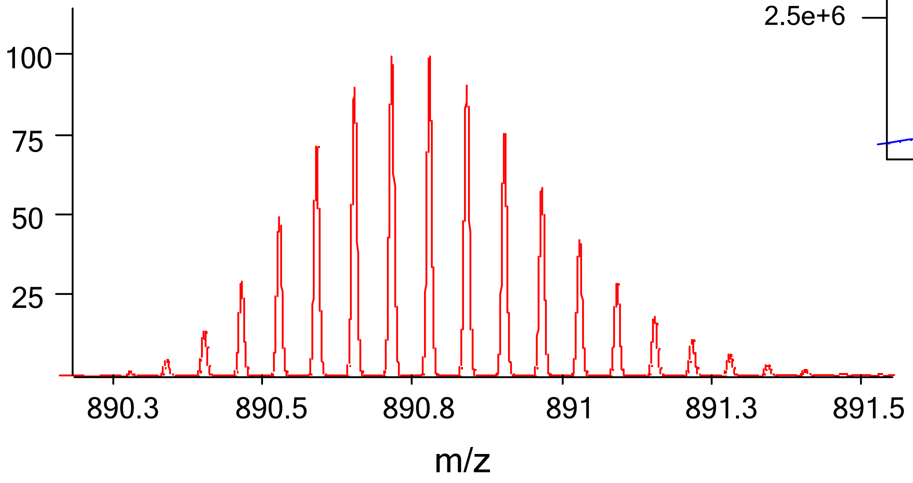
- Overlapping isotope patterns are separated by 4 Da
  - Creates challenges for deisotoping, particularly for charge states of 3+ or higher



Isotopic Composition

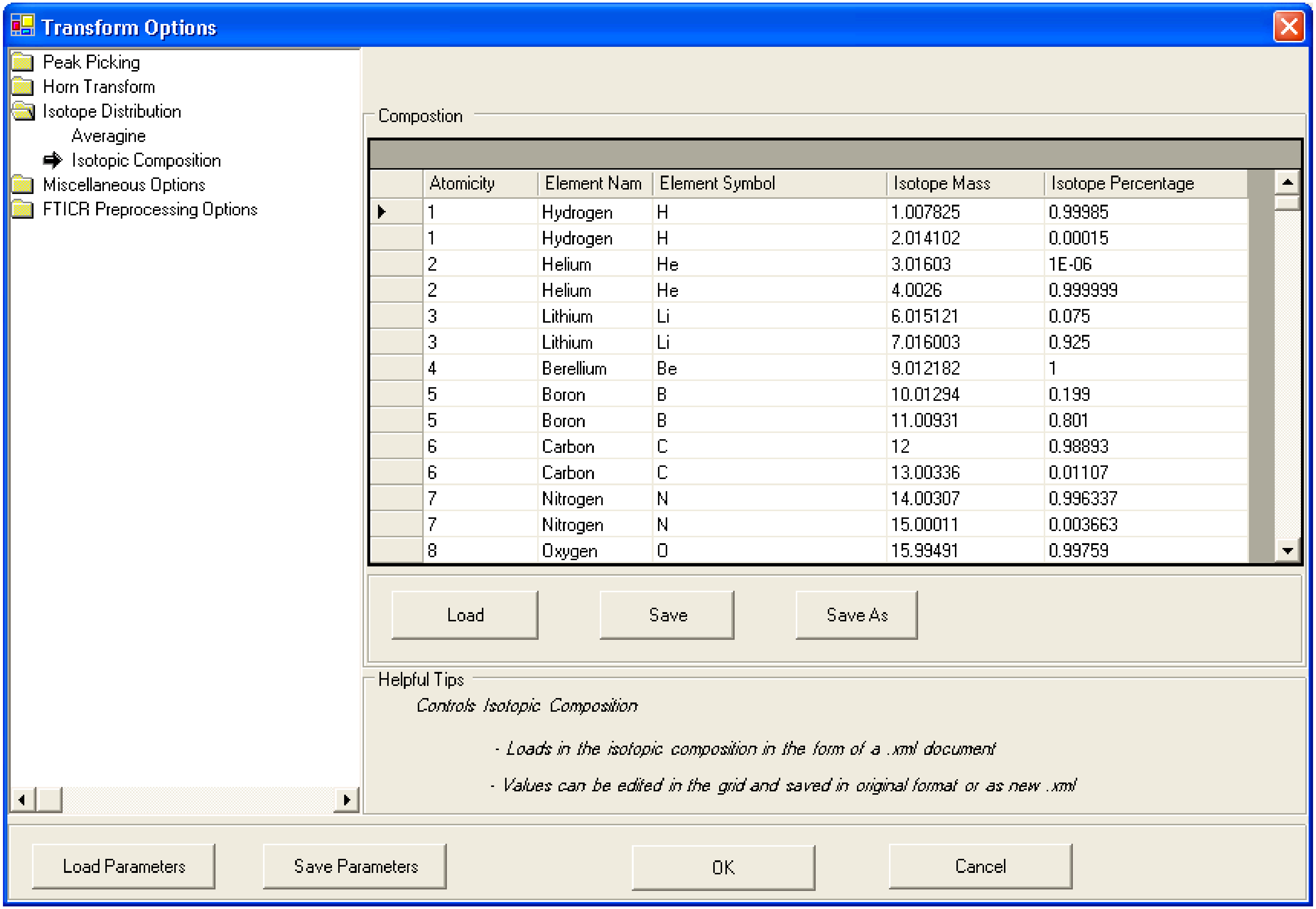
- Deviation from natural abundances
  - In ^{13}C, ^{15}N depleted media, isotopic composition of atoms is different from those found in nature
  - E.g., sulfur isotopes predominate the distribution at right
  - Constrast with an isotopic distribution of a peptide with similar mass and charge (16+),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





Isotopic Composition

- Decon2LS supports changing the isotope composition



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

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

Part II: LC-MS Feature Discovery

- Introduction (Adkins)
- Part I: Overview of Label-Free Quantitative Proteomics (Jaffe)
- Part II: Feature discovery in LC-MS datasets (Monroe and Jaitly)
  - ✓ 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
- Part III: PEPPeR, GenePattern and Real-world examples (Jaffe)
- Break
- AMT tag Pipeline Demo (general)
- Panel Discussion

Feature definition over elution time

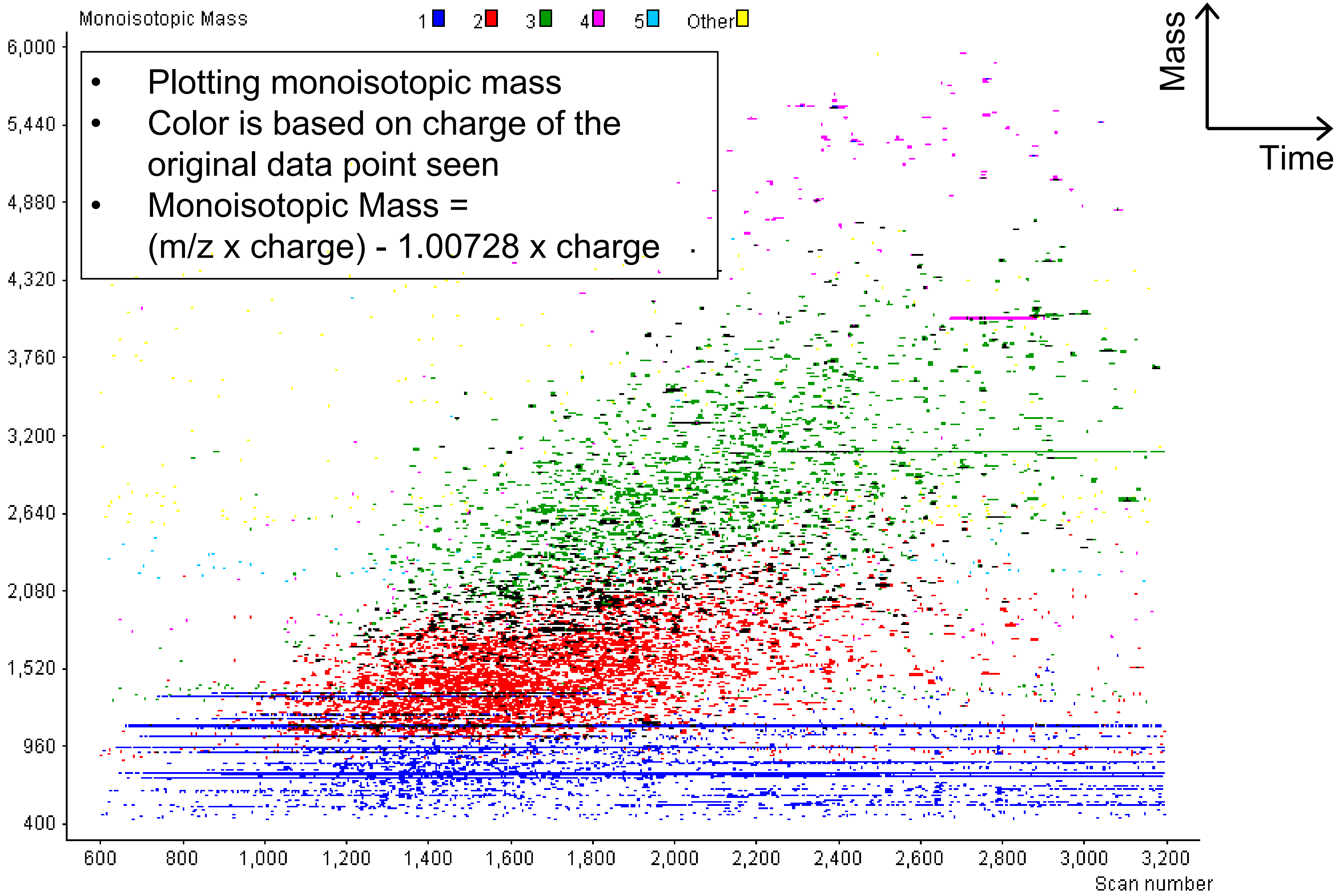
- Deisotoping collapses original data into data lists

						monoiso	most abu.		
scan num	charge	abundance	mz	fit	average mw	mw	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

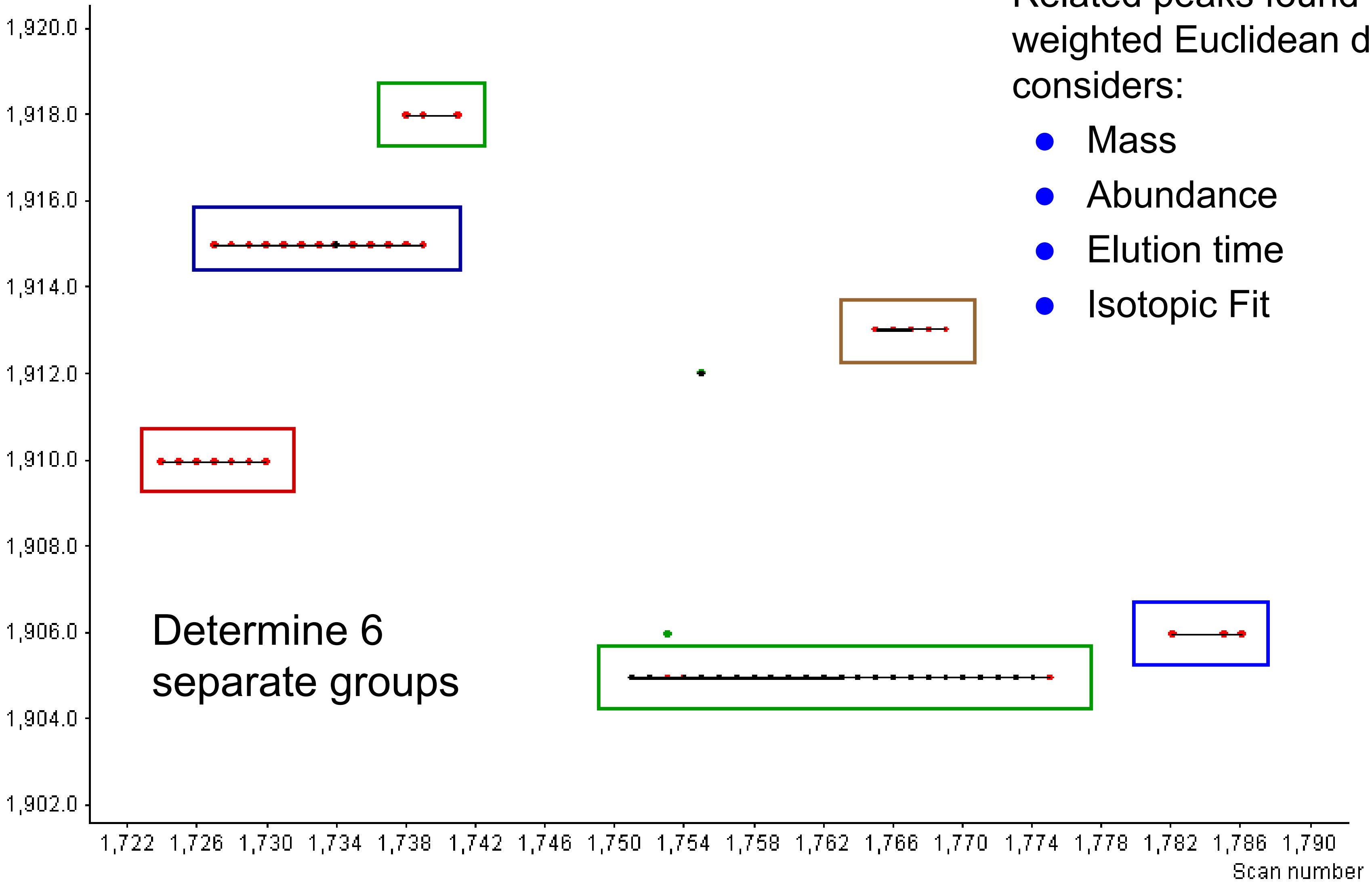
- Can visualize deisotoped data in two-dimensions



- 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



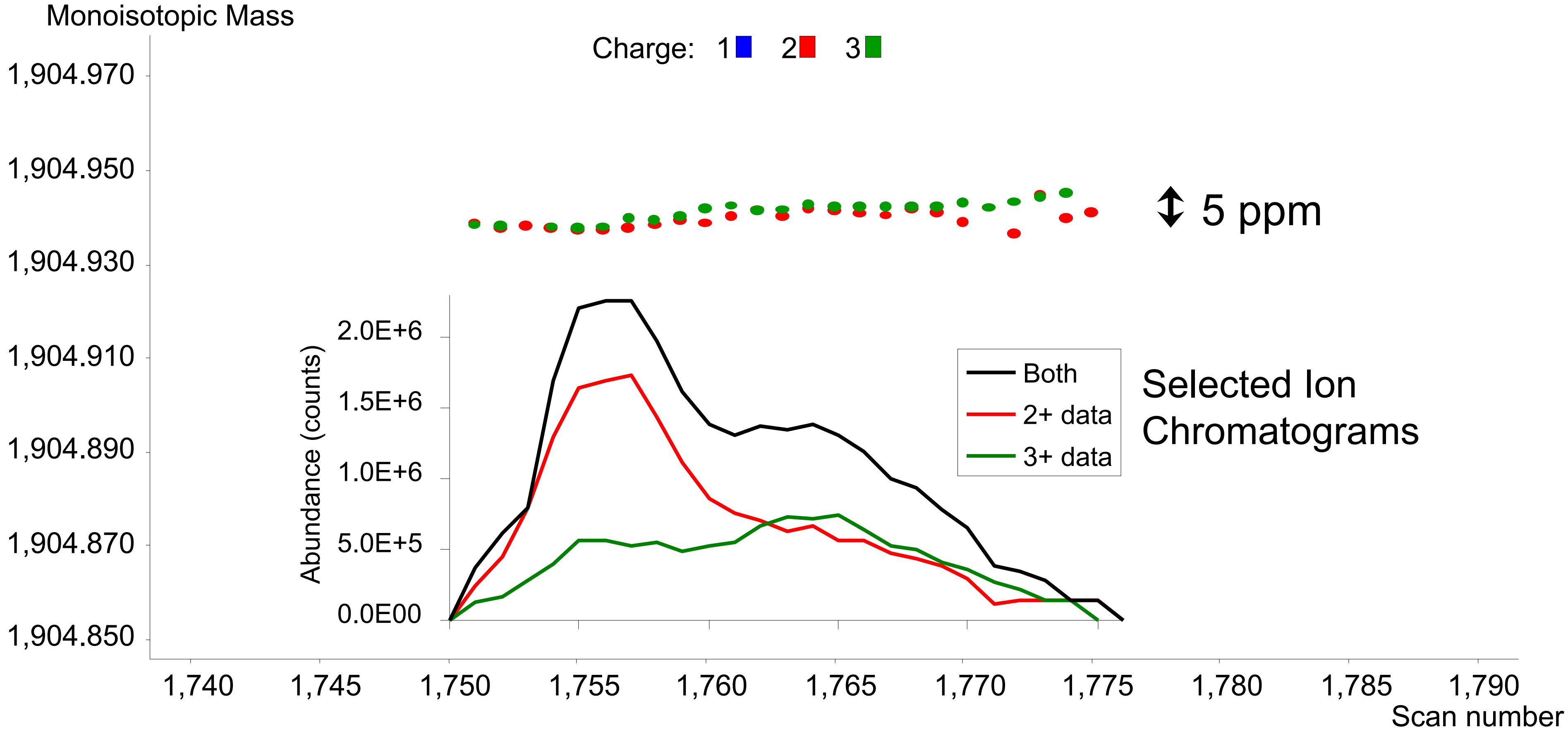
Determine 6 separate groups

Related peaks found using a weighted Euclidean distance; considers:

- Mass
- Abundance
- Elution time
- Isotopic Fit

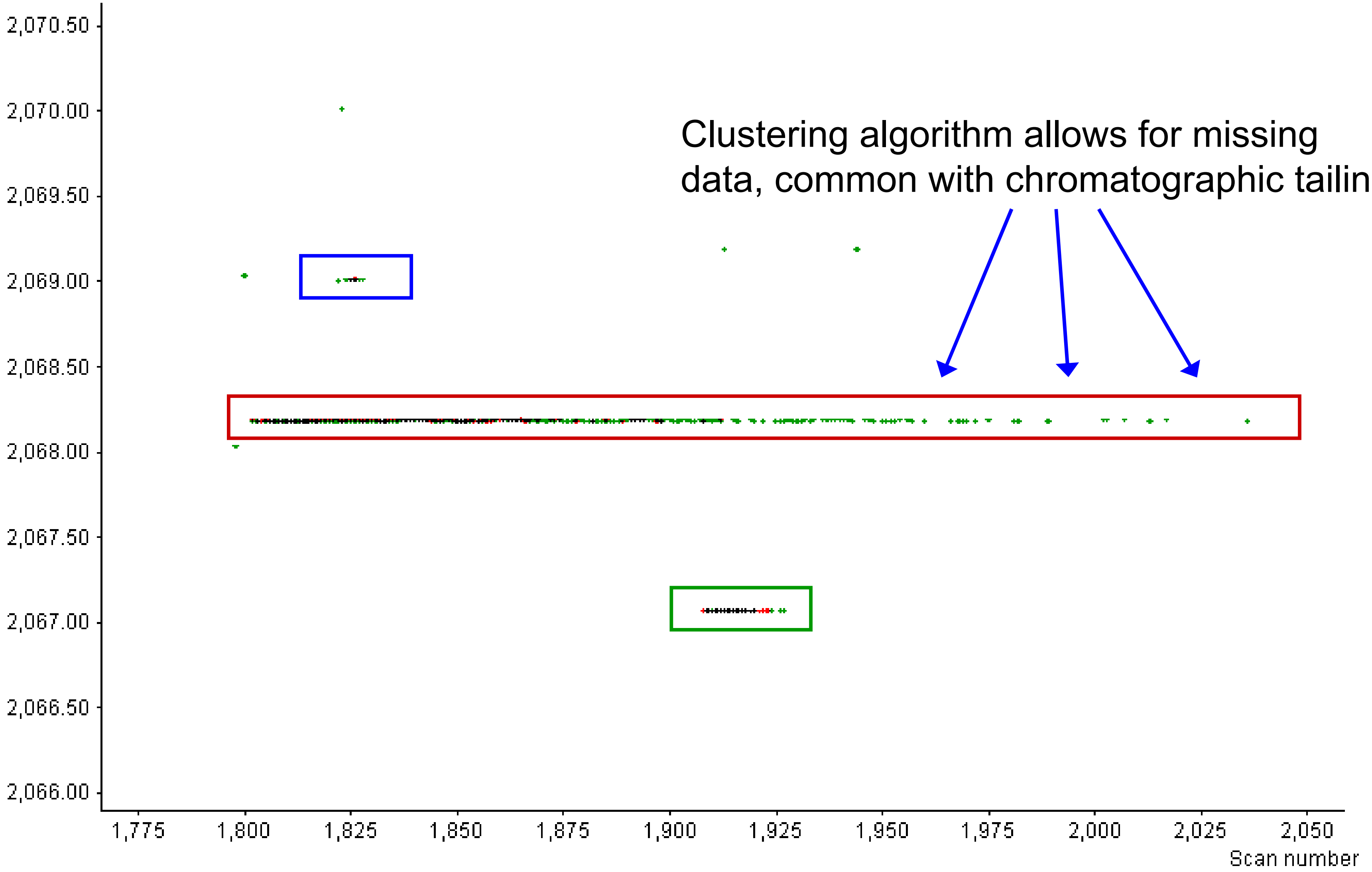
Feature definition over elution time

- Feature detail
  - Median Mass: 1904.9399 Da (more tolerant to outliers than average)
  - Elution Time: Scan 1757 (0.363 NET)
  - Abundance: 1.7x10^7 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

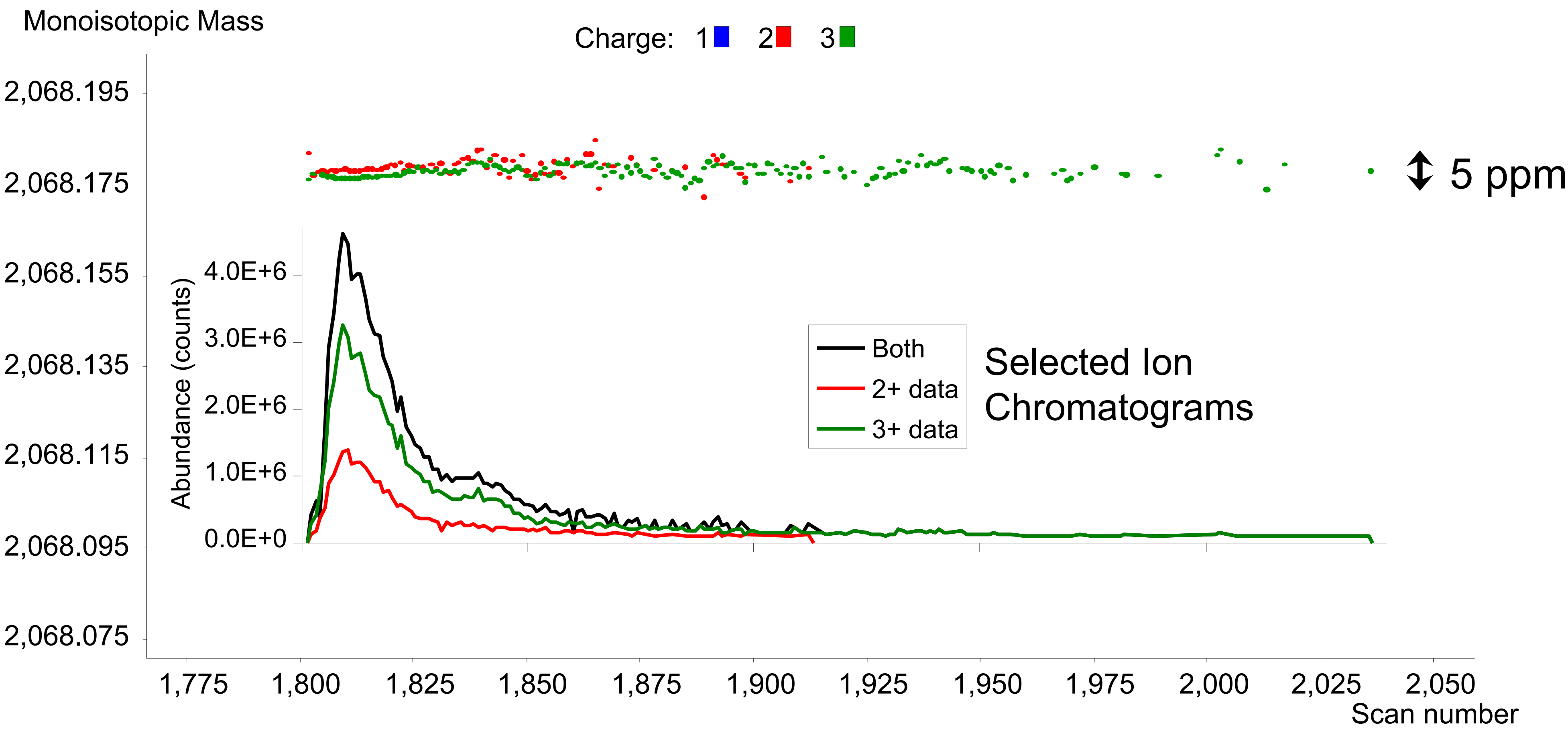
- Second example
  - LC-MS feature eluting over 7.5 minutes



Clustering algorithm allows for missing data, common with chromatographic tailing

Feature definition over elution time

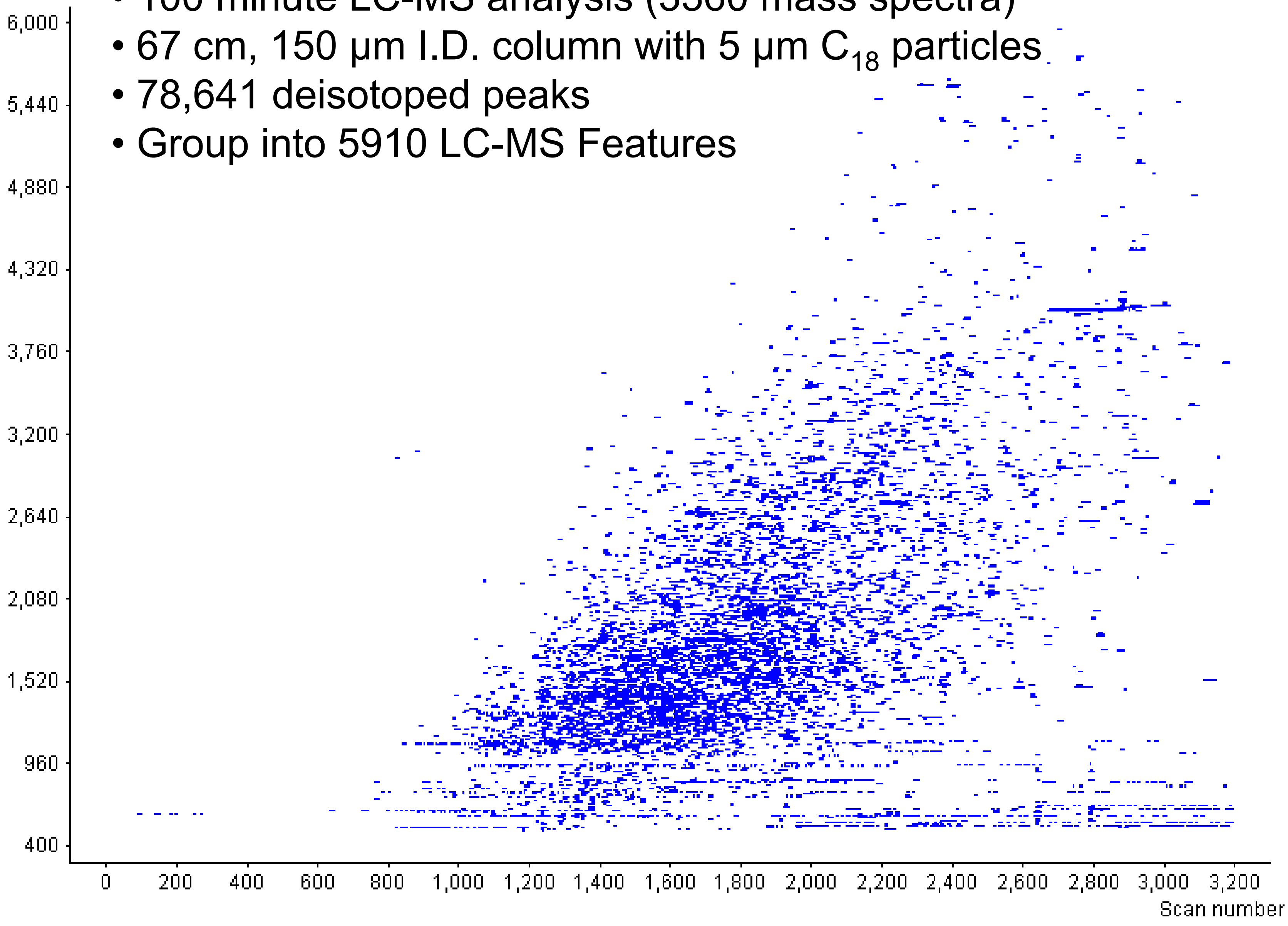
- Second example, feature detail
  - Median Mass: 2068.1781 Da
  - Elution Time: Scan 1809 (0.380 NET)
  - Abundance: 8.7×10^7 counts (area under 3+ SIC)
    - This example has primarily 3+ data; previous had even mix of 2+ and 3+ data



Selected Ion Chromatograms

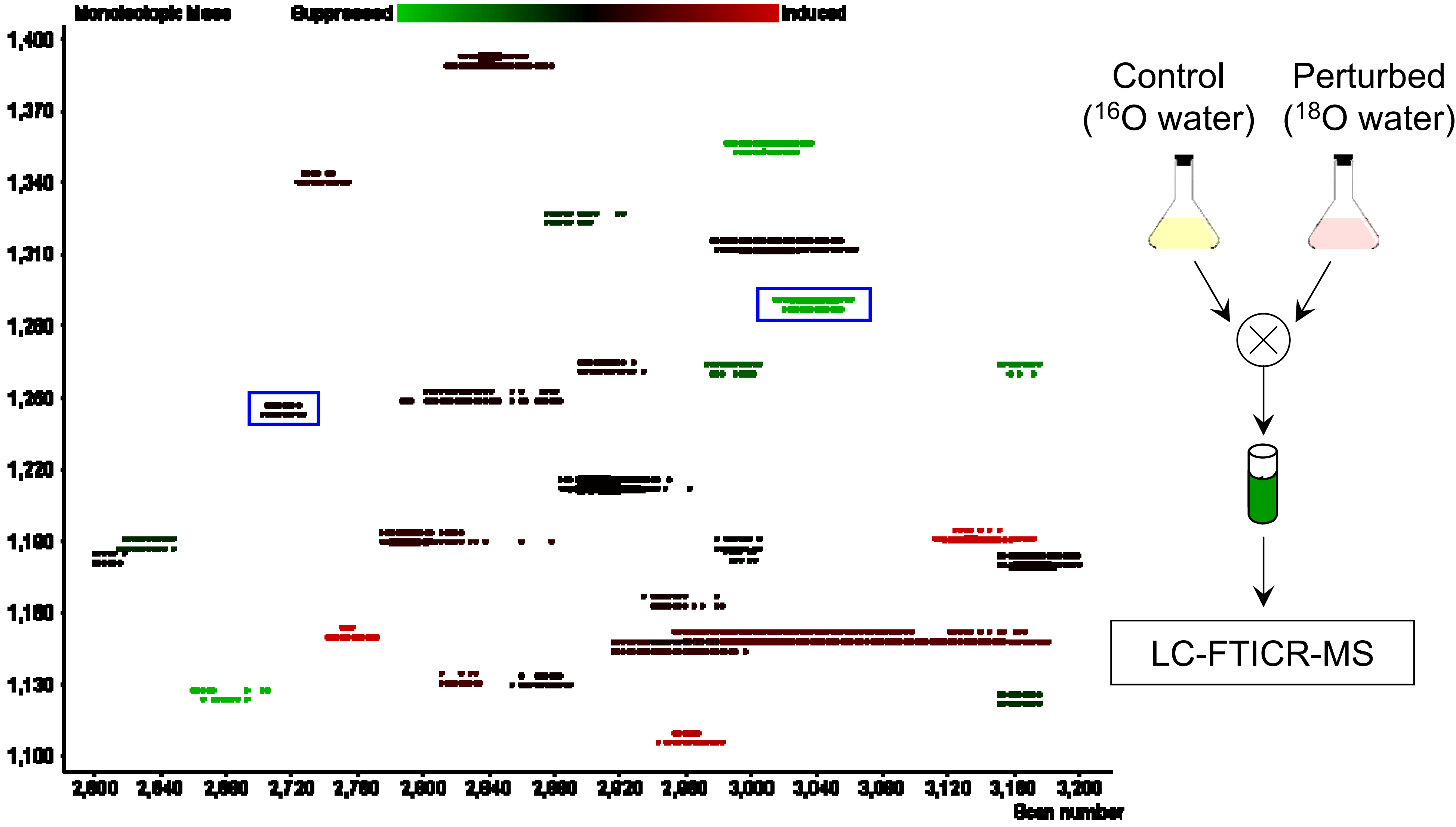
Feature definition over elution time

- 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



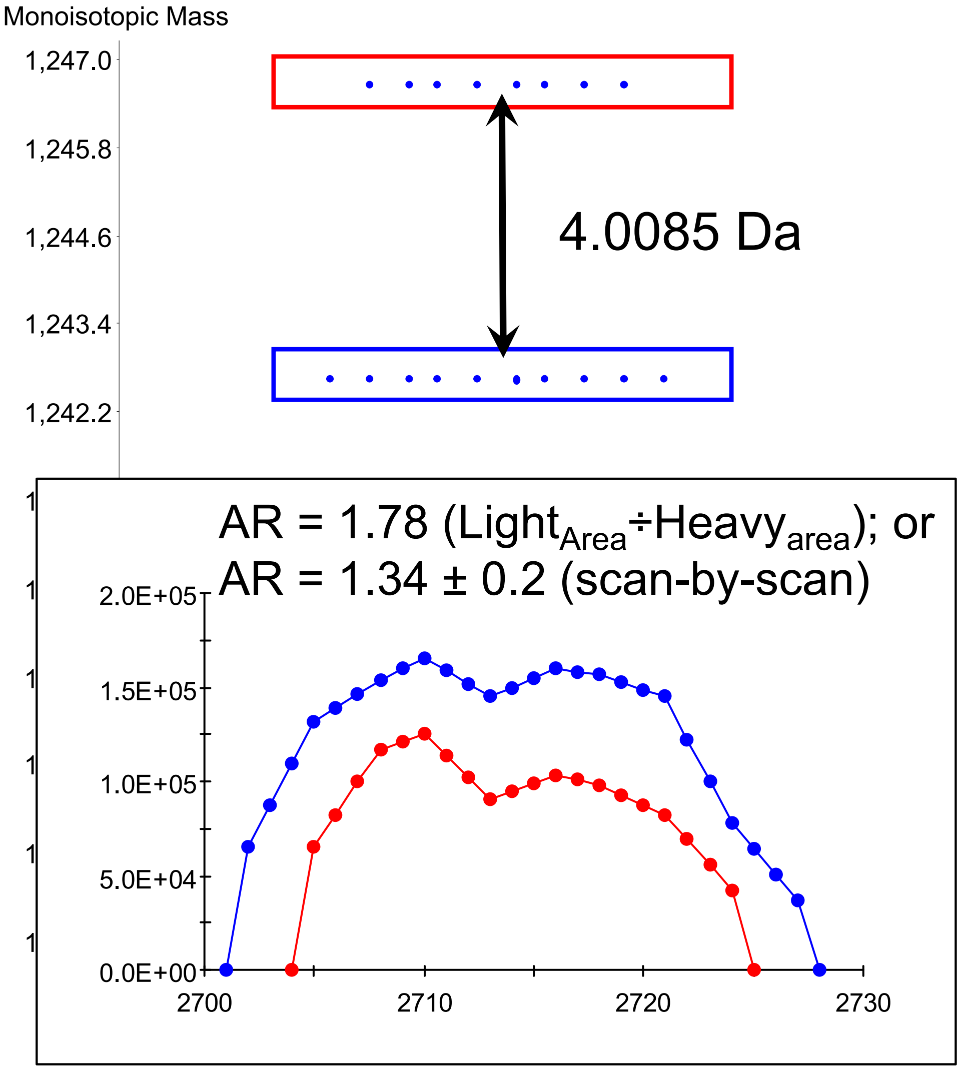
Control (^{16}O water)

Perturbed (^{18}O water)

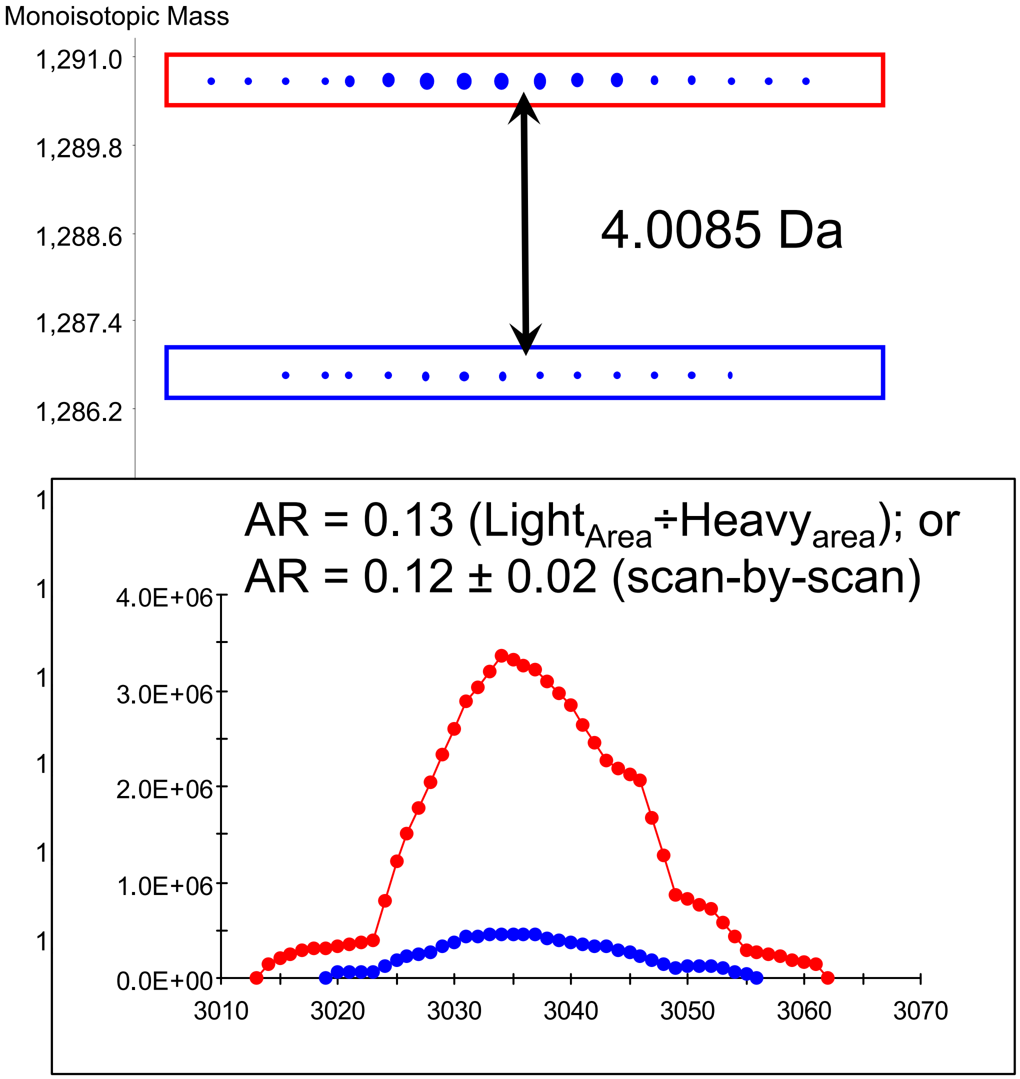
LC-FTICR-MS

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)



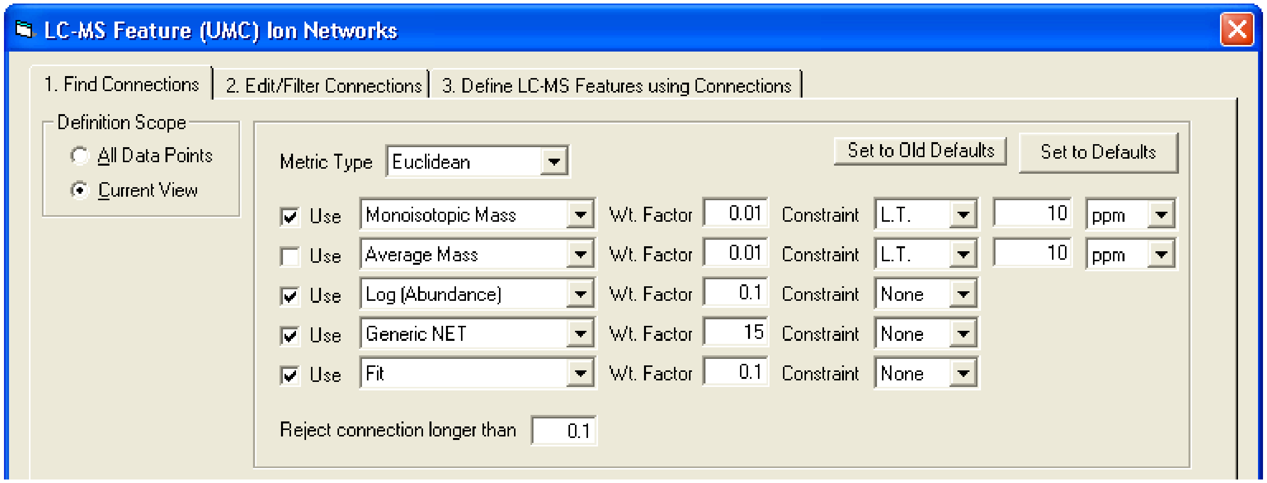
AR = 1.78 ($Light_{Area}$÷$Heavy_{area}$); or
AR = 1.34 ± 0.2 (scan-by-scan)



AR = 0.13 ($Light_{Area}$÷$Heavy_{area}$); or
AR = 0.12 ± 0.02 (scan-by-scan)

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

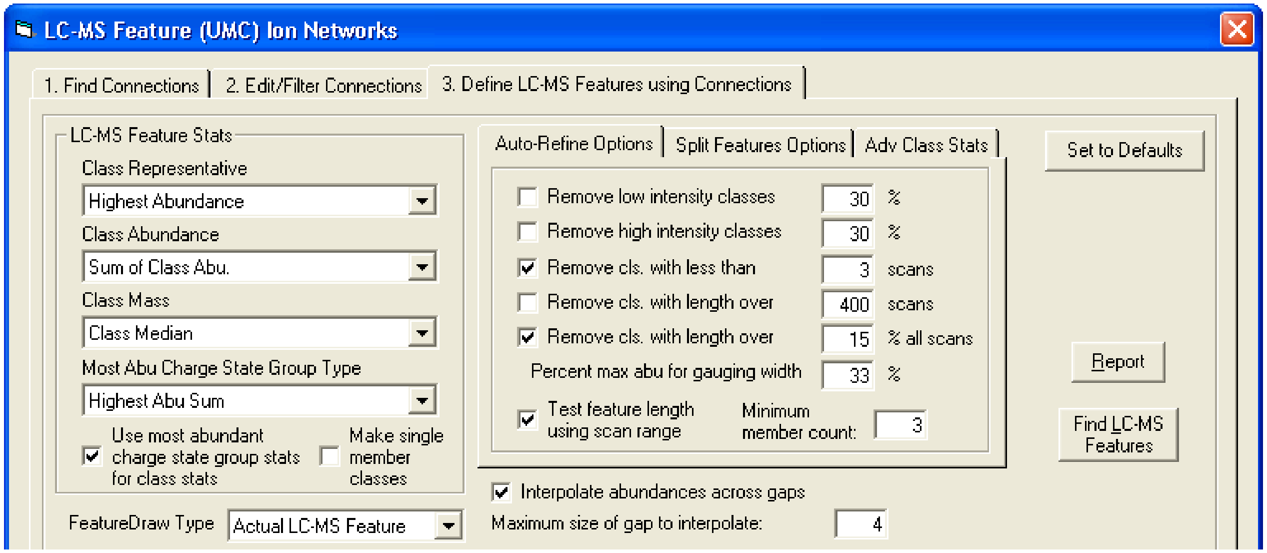
Definition Scope
○ All Data Points
● Current View

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

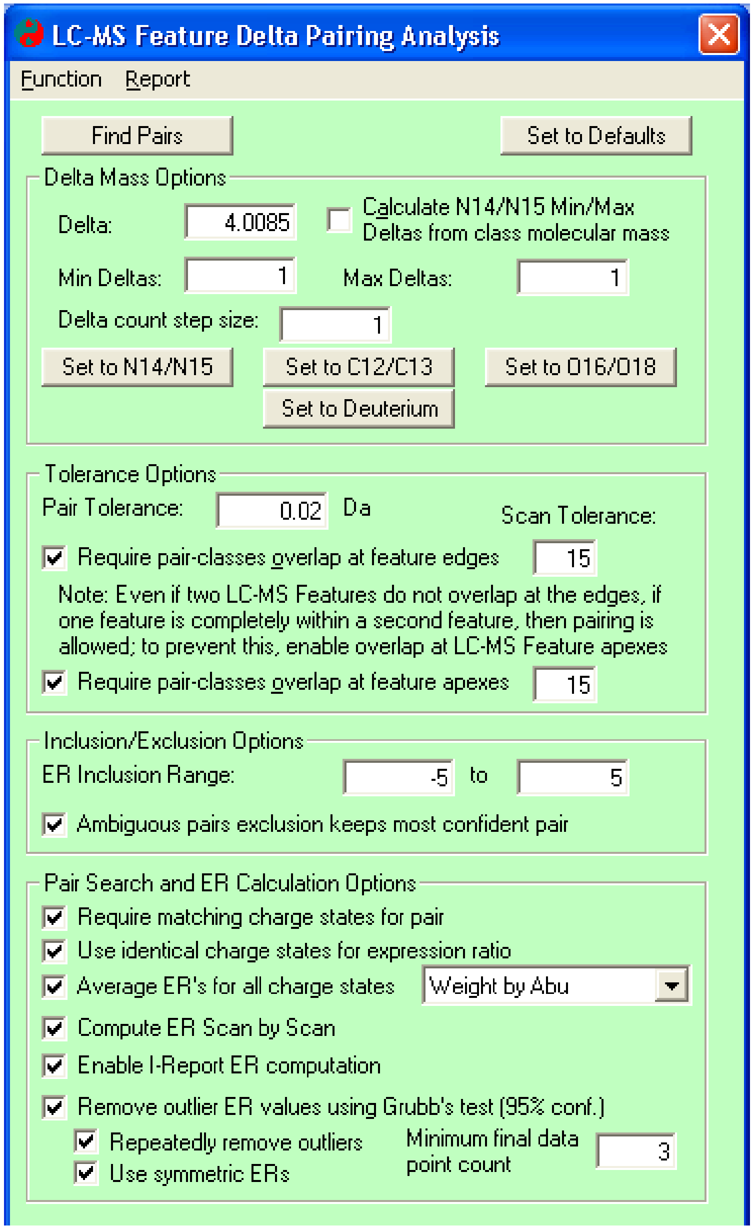
Set to Defaults

☐ 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

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 — Minimum final data point count 3
☑ Use symmetric ERs

Part II: LC-MS Feature Discovery

- Introduction (Adkins)
- Part I: Overview of Label-Free Quantitative Proteomics (Jaffe)
- Part II: Feature discovery in LC-MS datasets (Monroe and Jaitly)
  - ✓ 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
- Part III: PEPPeR, GenePattern and Real-world examples (Jaffe)
- Break
- AMT tag Pipeline Demo (general)
- Panel Discussion

Assembling an AMT tag DB

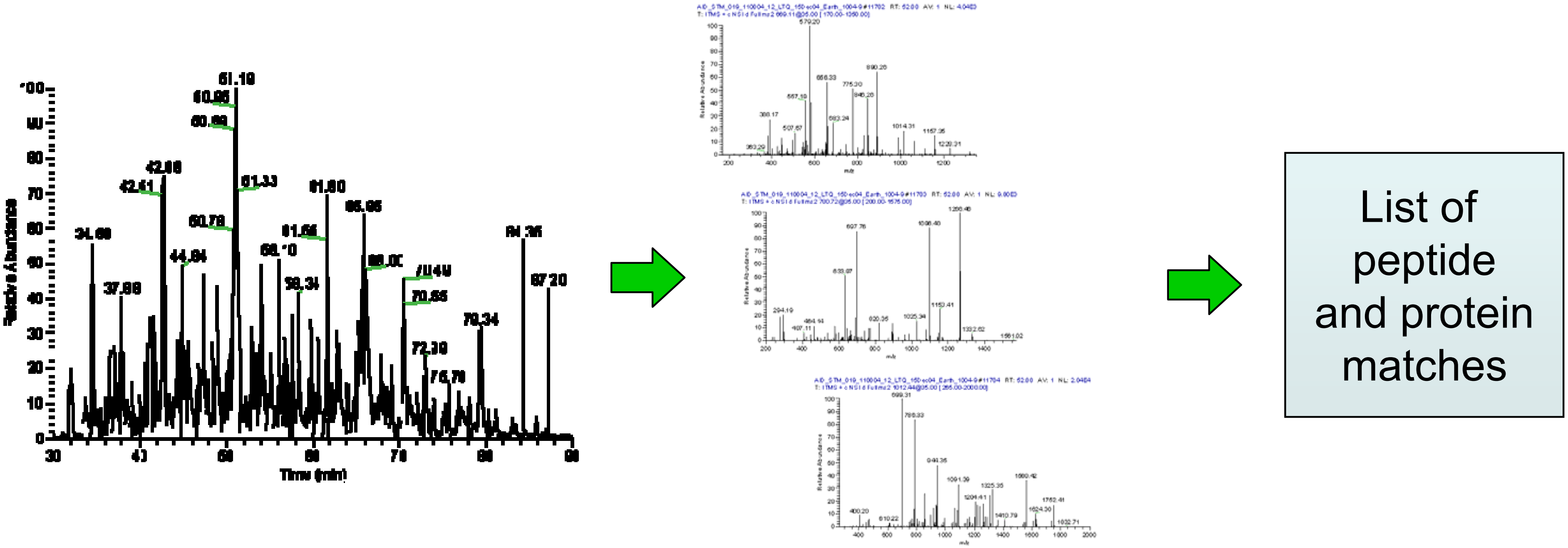
- Accurate Mass and Time (AMT) tag
  - Unique peptide sequence whose monoisotopic mass and normalized elution time are accurately known
  - AMT tags also track any modified residues in peptide
- AMT tag DB
  - Collection of AMT tags
- AMT tag approach articles
  - R.D. Smith et. al., *Proteomics* **2002**, *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.

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

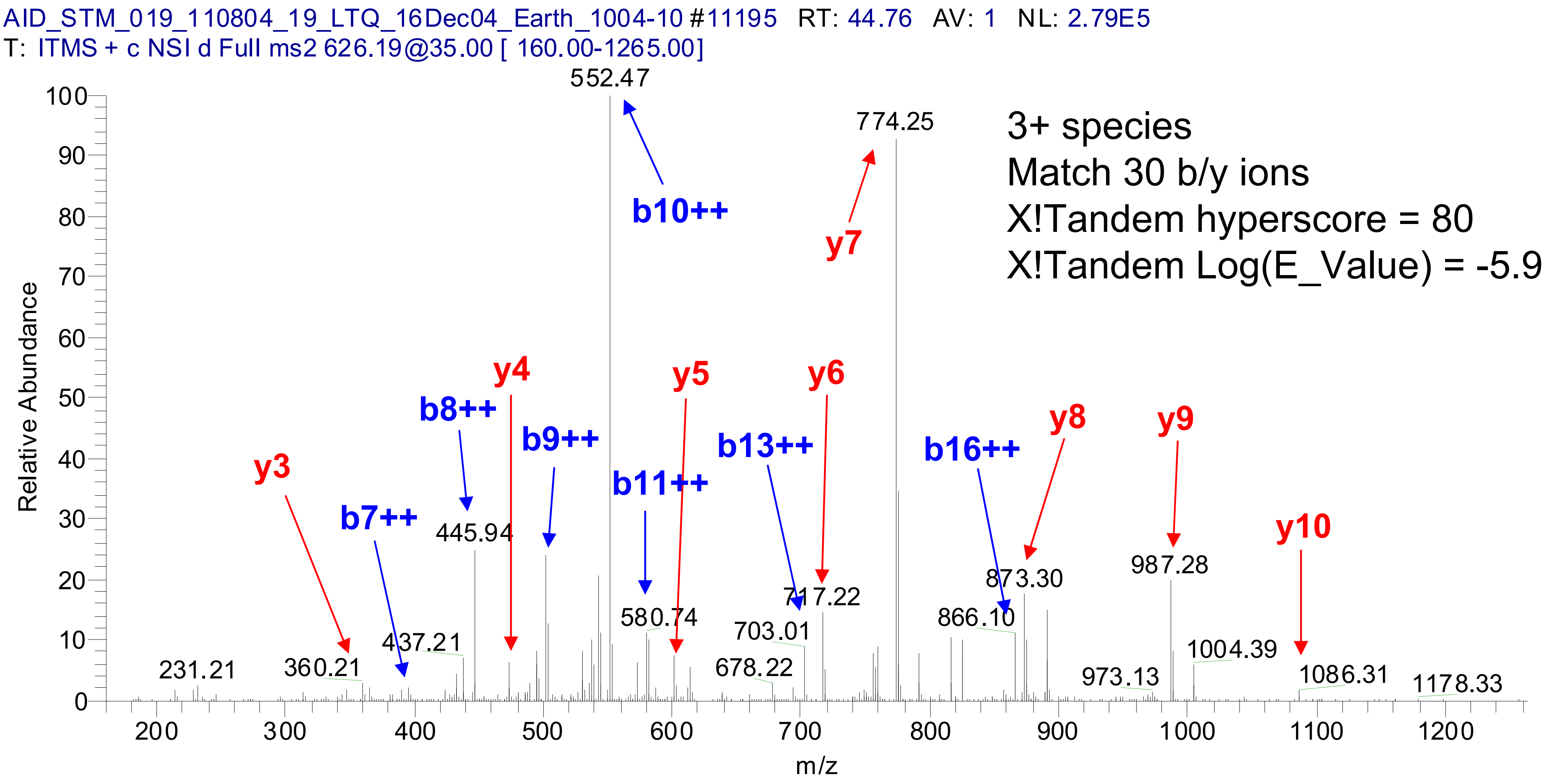
- Characterizing AMT tags
  - Analyze samples by LC-MS/MS
    - 10 minute to 180 minute LC separations
    - Obtain 1000's of MS/MS fragmentation spectra for each sample
  - Analyze spectra using SEQUEST, X!Tandem, etc.
    - SEQUEST: http://www.thermo.com/bioworks/
    - X!Tandem: http://www.thegpm.org/TANDEM/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



List of peptide and protein matches

Assembling an AMT tag DB

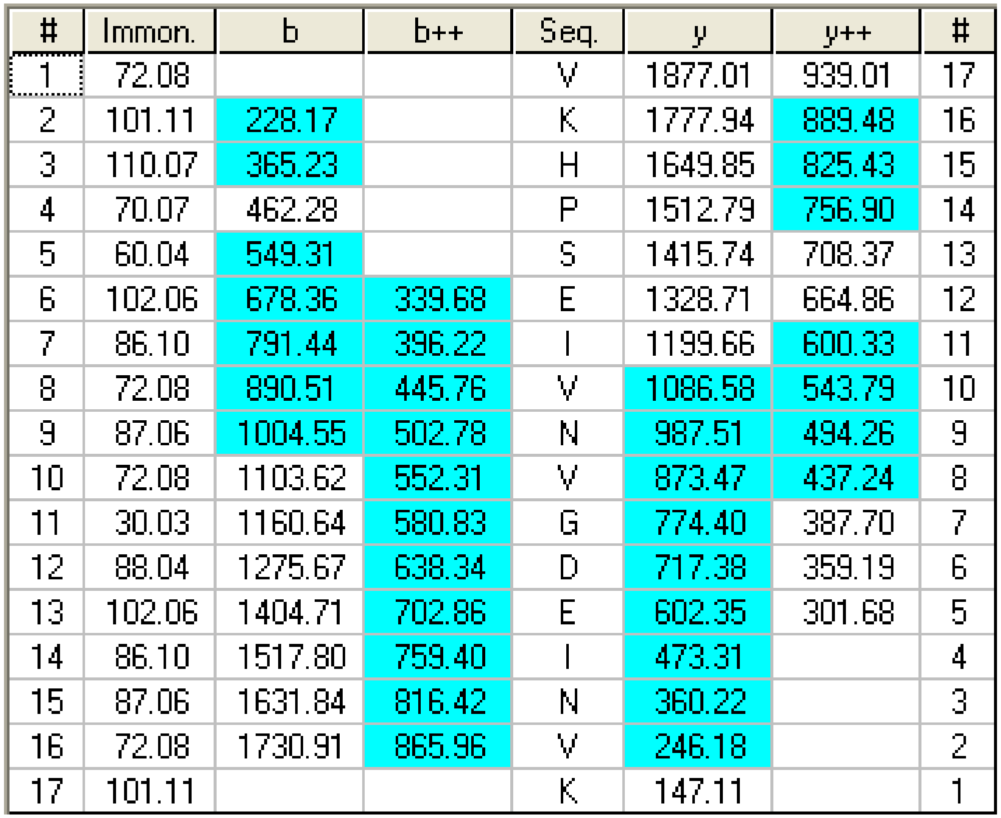
- AMT tag example
  - R.VKHPSEIVNVGDEINVK.V
  - Observed in scan 11195 of dataset #19 in an SCX fractionation series



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

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

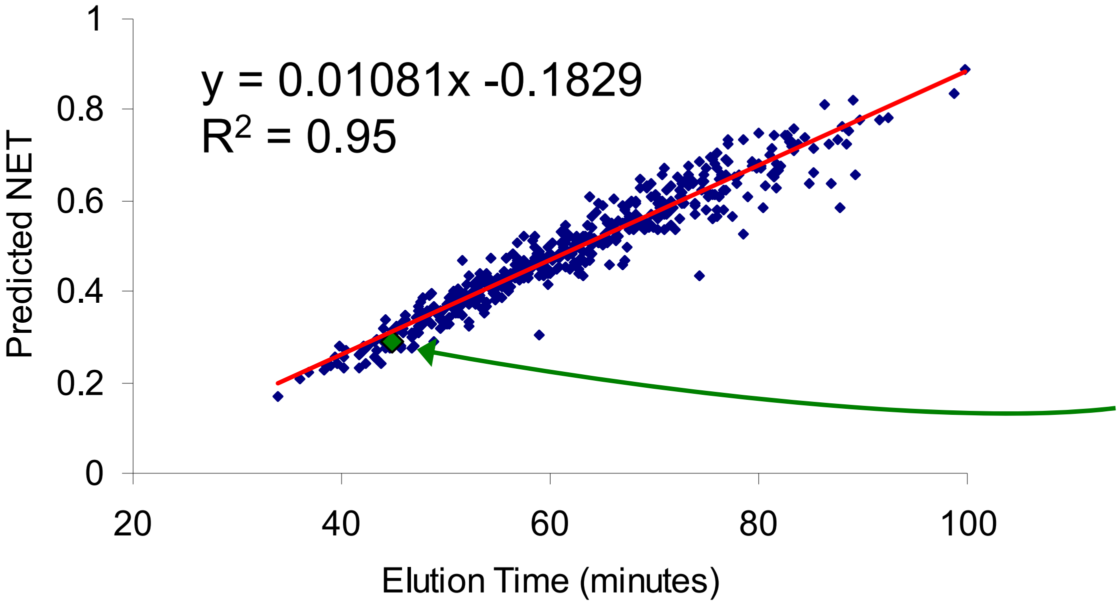
Peptide	X!Tandem Log (E_Value)	Elution Time	Predicted NET
R.AARPAKYSYVDENGETK.T	-6.1	33.958	0.167
R.LVHGEEGLVAAKR.I	-8.8	36.915	0.224
R.GIIKVGEEVEIVGIK.E	-8.2	53.003	0.415
K.RFNDDGPILFIHTGGAPALFAYHPHV.-	-7.3	62.583	0.519
K.KTGVLAQVQEALKGLDVR.E	-11.6	62.803	0.438
R.KVAAQIPNGSTLFIDIGTTPEAVAHALLGHSNLR.I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

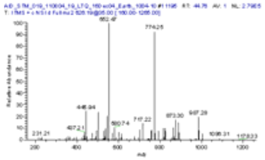
Example: 506 unique peptides used for alignment; Log(E_Value) ≤ -6



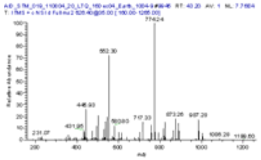
VKHPSEIVNVGDEINVK
Elution time: 44.923 minutes
Predicted NET: 0.292
Observed NET: 0.303

Assembling an AMT tag DB

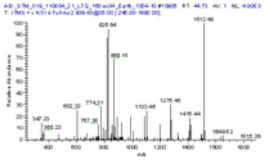
- AMT tag example
  - R.VKHPSEIVNVGDEINVK.V
  - Observed in 7 (of 25) LC-MS/MS datasets in the SCX fractionation series



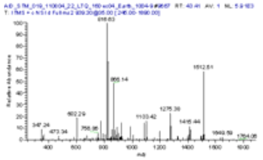
Analysis 1, scan 11195 → 3+, hyperscore 80, Obs. NET 0.303



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

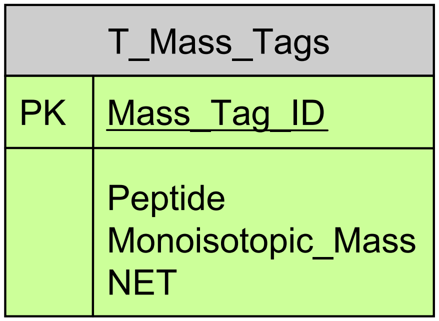
Assembling an AMT tag DB

- 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
```

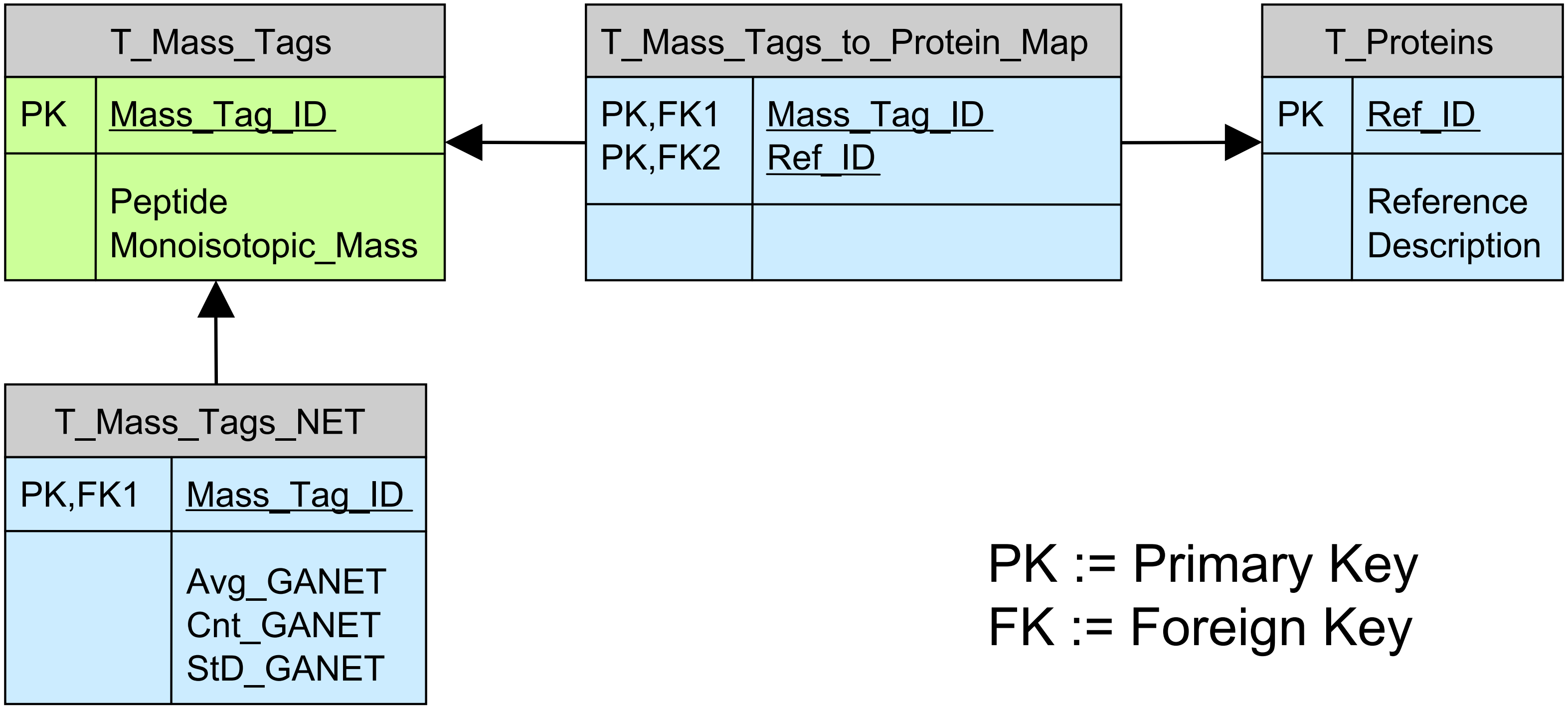
Assembling an AMT tag DB

- Database Relationships
  - Minimum information required:
    - Single table with Mass and NET



T_Mass_Tags	
PK	Mass_Tag_ID
	Peptide Monoisotopic_Mass NET

  - Expanded schema:



T_Mass_Tags	
PK	Mass_Tag_ID
	Peptide Monoisotopic_Mass

T_Mass_Tags_to_Protein_Map	
PK,FK1 PK,FK2	Mass_Tag_ID Ref_ID

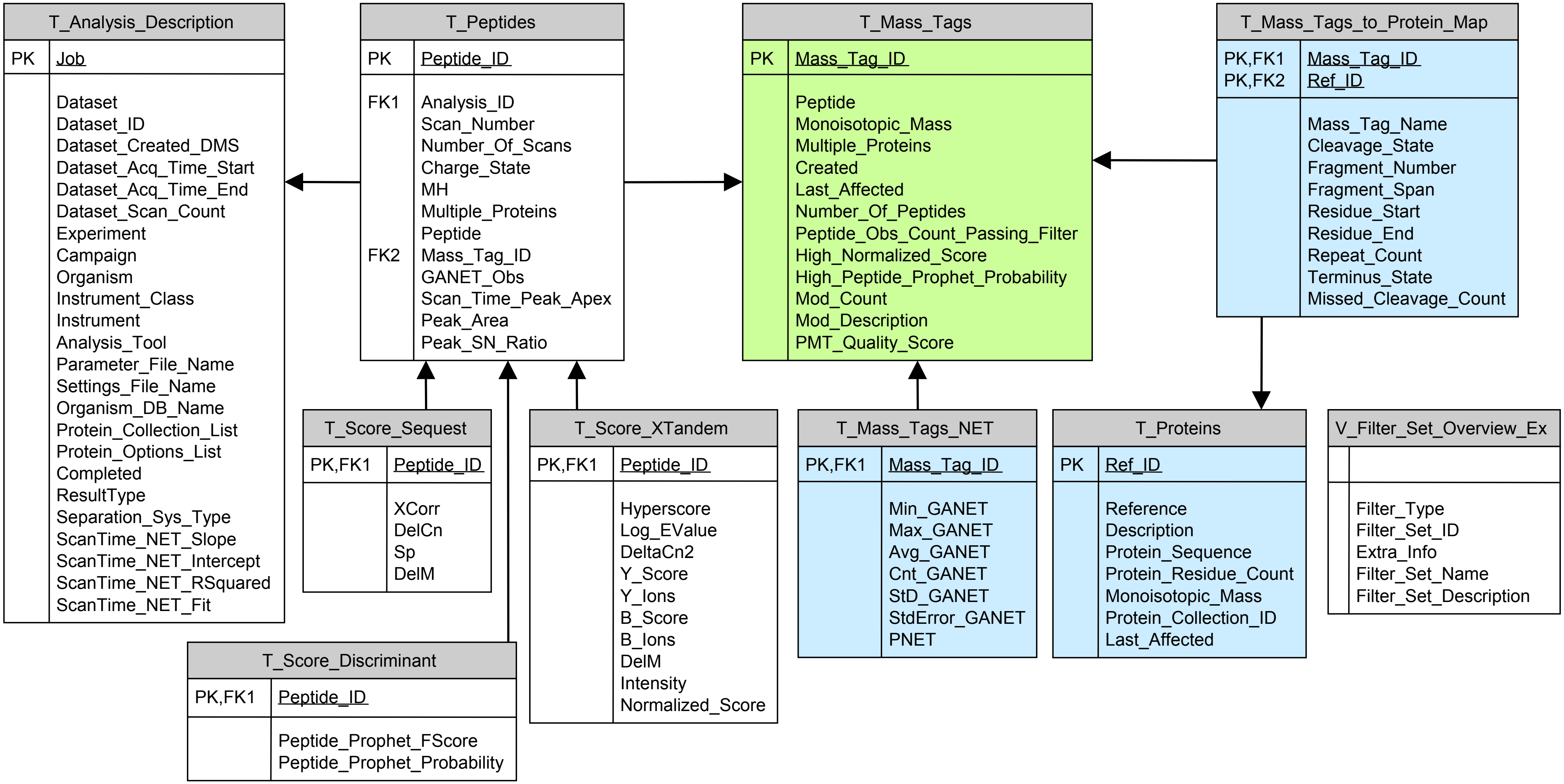
T_Proteins	
PK	Ref_ID
	Reference Description

T_Mass_Tags_NET	
PK,FK1	Mass_Tag_ID
	Avg_GANET Cnt_GANET StD_GANET

PK := Primary Key
FK := Foreign Key

Assembling an AMT tag DB

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



	T_Analysis_Description
PK	Job
	Dataset Dataset_ID Dataset_Created_DMS Dataset_Acq_Time_Start Dataset_Acq_Time_End Dataset_Scan_Count Experiment Campaign Organism Instrument_Class Instrument Analysis_Tool Parameter_File_Name Settings_File_Name Organism_DB_Name Protein_Collection_List Protein_Options_List Completed ResultType Separation_Sys_Type ScanTime_NET_Slope ScanTime_NET_Intercept ScanTime_NET_RSquared ScanTime_NET_Fit

	T_Peptides
PK	Peptide_ID
FK1	Analysis_ID
	Scan_Number Number_Of_Scans Charge_State MH Multiple_Proteins Peptide
FK2	Mass_Tag_ID
	GANET_Obs Scan_Time_Peak_Apex Peak_Area Peak_SN_Ratio

	T_Mass_Tags
PK	Mass_Tag_ID
	Peptide Monoisotopic_Mass Multiple_Proteins Created Last_Affected Number_Of_Peptides Peptide_Obs_Count_Passing_Filter High_Normalized_Score High_Peptide_Prophet_Probability Mod_Count Mod_Description PMT_Quality_Score

	T_Mass_Tags_to_Protein_Map
PK,FK1	Mass_Tag_ID
PK,FK2	Ref_ID
	Mass_Tag_Name Cleavage_State Fragment_Number Fragment_Span Residue_Start Residue_End Repeat_Count Terminus_State Missed_Cleavage_Count

	T_Score_Sequest
PK,FK1	Peptide_ID
	XCorr DelCn Sp DelM

	T_Score_XTandem
PK,FK1	Peptide_ID
	Hyperscore Log_EValue DeltaCn2 Y_Score Y_Ions B_Score B_Ions DelM Intensity Normalized_Score

	T_Mass_Tags_NET
PK,FK1	Mass_Tag_ID
	Min_GANET Max_GANET Avg_GANET Cnt_GANET StD_GANET StdError_GANET PNET

	T_Proteins
PK	Ref_ID
	Reference Description Protein_Sequence Protein_Residue_Count Monoisotopic_Mass Protein_Collection_ID Last_Affected

	V_Filter_Set_Overview_Ex
	Filter_Type Filter_Set_ID Extra_Info Filter_Set_Name Filter_Set_Description

	T_Score_Discriminant
PK,FK1	Peptide_ID
	Peptide_Prophet_FScore Peptide_Prophet_Probability

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

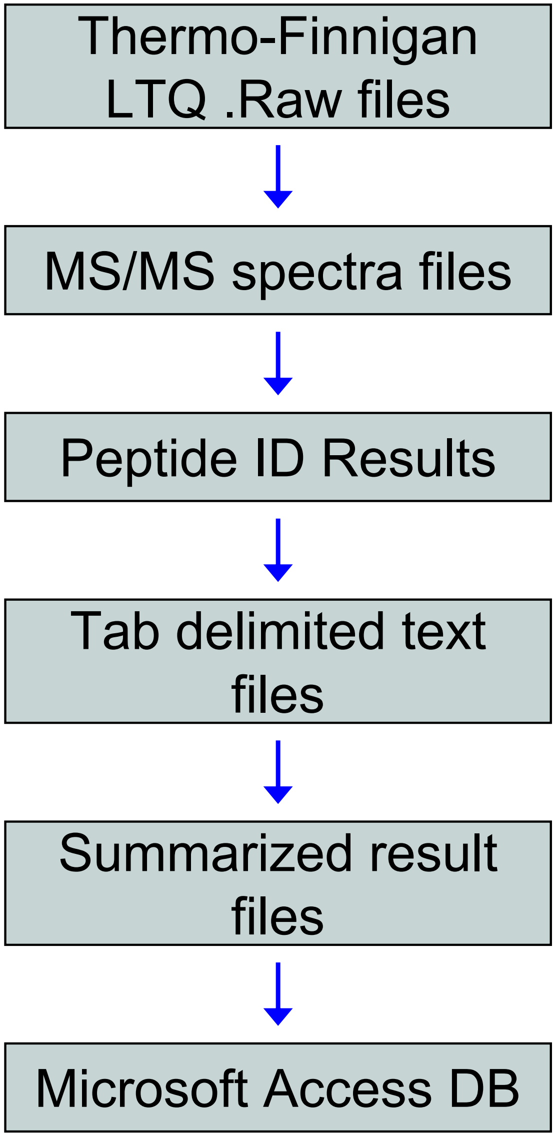
Peptide_ID	Peptide	Mass Tag ID	Job	Scan Number	Charge State
53428	R.VKHPSEIVNVGDEINVK.V	24847	206386	11195	3
57461	R.VKHPSEIVNVGDEINVK.V	24847	206387	9945	3
61511	R.VKHPSEIVNVGDEINVK.V	24847	206388	10905	2
65386	R.VKHPSEIVNVGDEINVK.V	24847	206389	9667	2
69081	R.VKHPSEIVNVGDEINVK.V	24847	206390	9118	2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

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.

MS/MS spectra files

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.

Peptide ID Results

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

Tab delimited text files

Align datasets using the MTDB Creator application

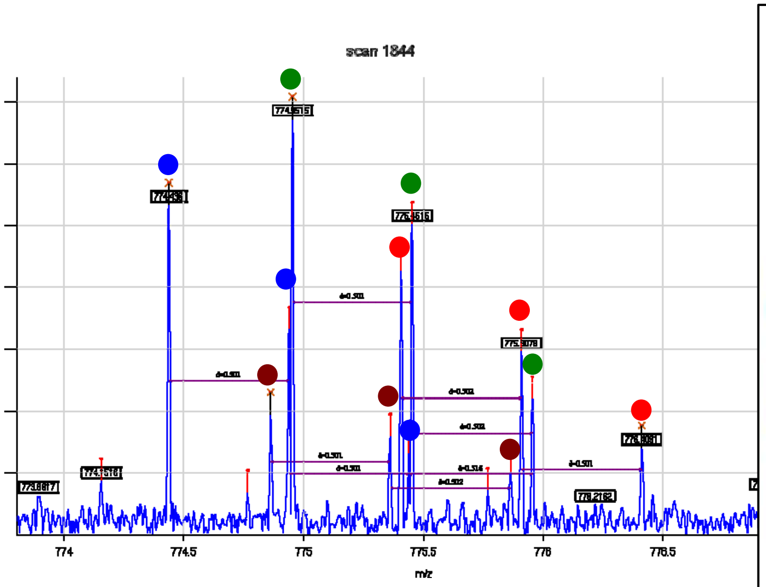
Summarized result files

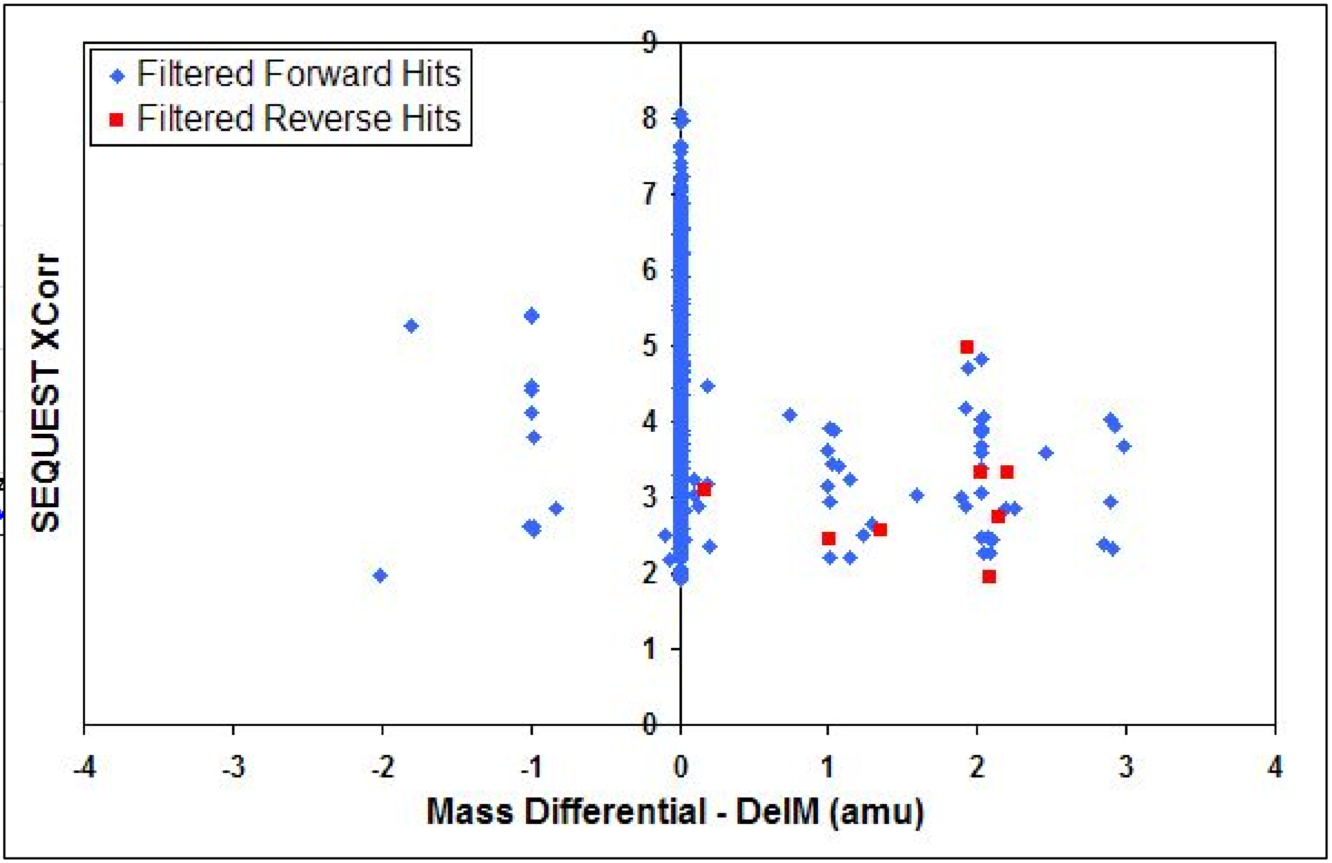
Load into database using MTDB Creator

Microsoft Access DB

DeconMSn

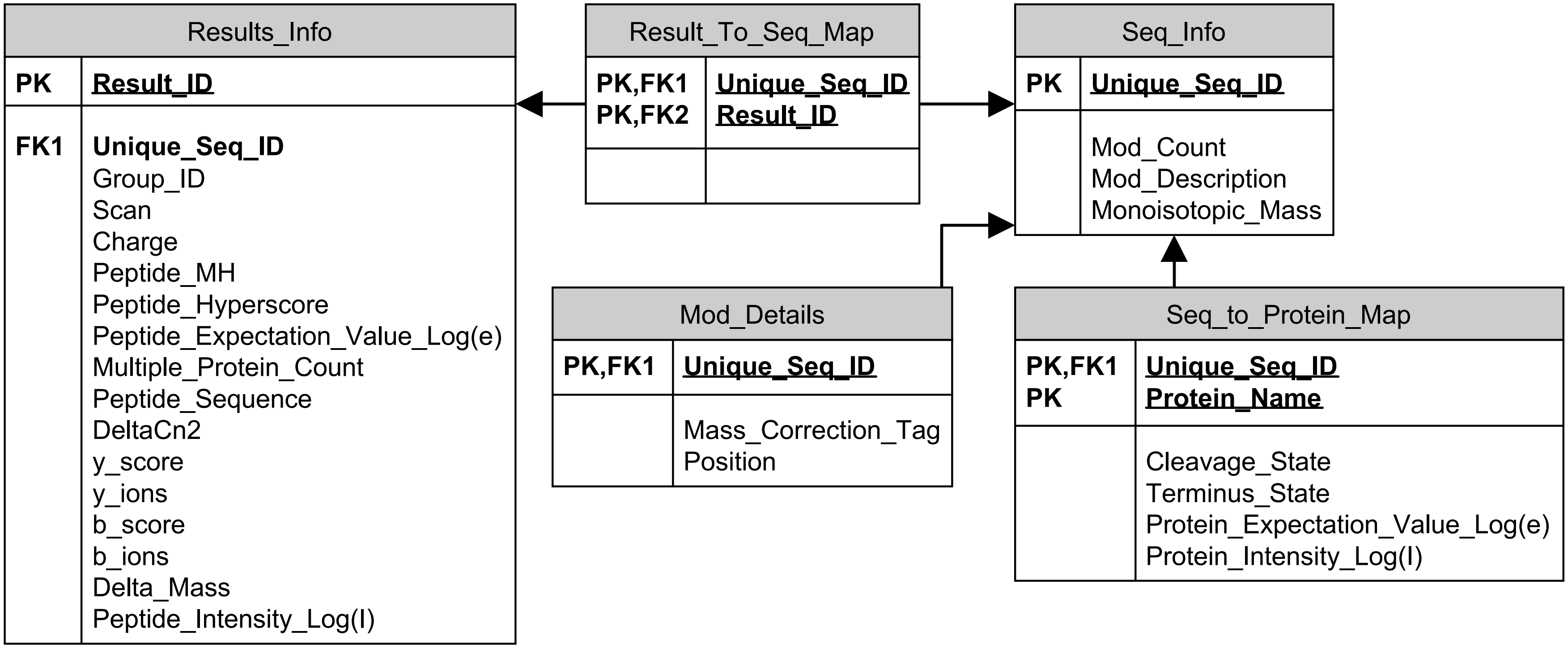
- Determines the monoisotopic mass and charge state of each parent ion chosen for fragmentation on a hybrid LC-MS/MS instrument using Decon2LS algorithms
- Replacement for the Extract_MSn.exe tool provided with SEQUEST and Bioworks





Assembling an AMT tag DB

- Peptide Hit Results Processor (PHRP) relationships



Results_Info	
PK	**Result_ID**
FK1	**Unique_Seq_ID** Group_ID Scan Charge Peptide_MH Peptide_Hyperscore Peptide_Expectation_Value_Log(e) Multiple_Protein_Count Peptide_Sequence DeltaCn2 y_score y_ions b_score b_ions Delta_Mass Peptide_Intensity_Log(I)

Result_To_Seq_Map	
PK,FK1 **PK,FK2**	**Unique_Seq_ID** **Result_ID**

Seq_Info	
PK	**Unique_Seq_ID**
	Mod_Count Mod_Description Monoisotopic_Mass

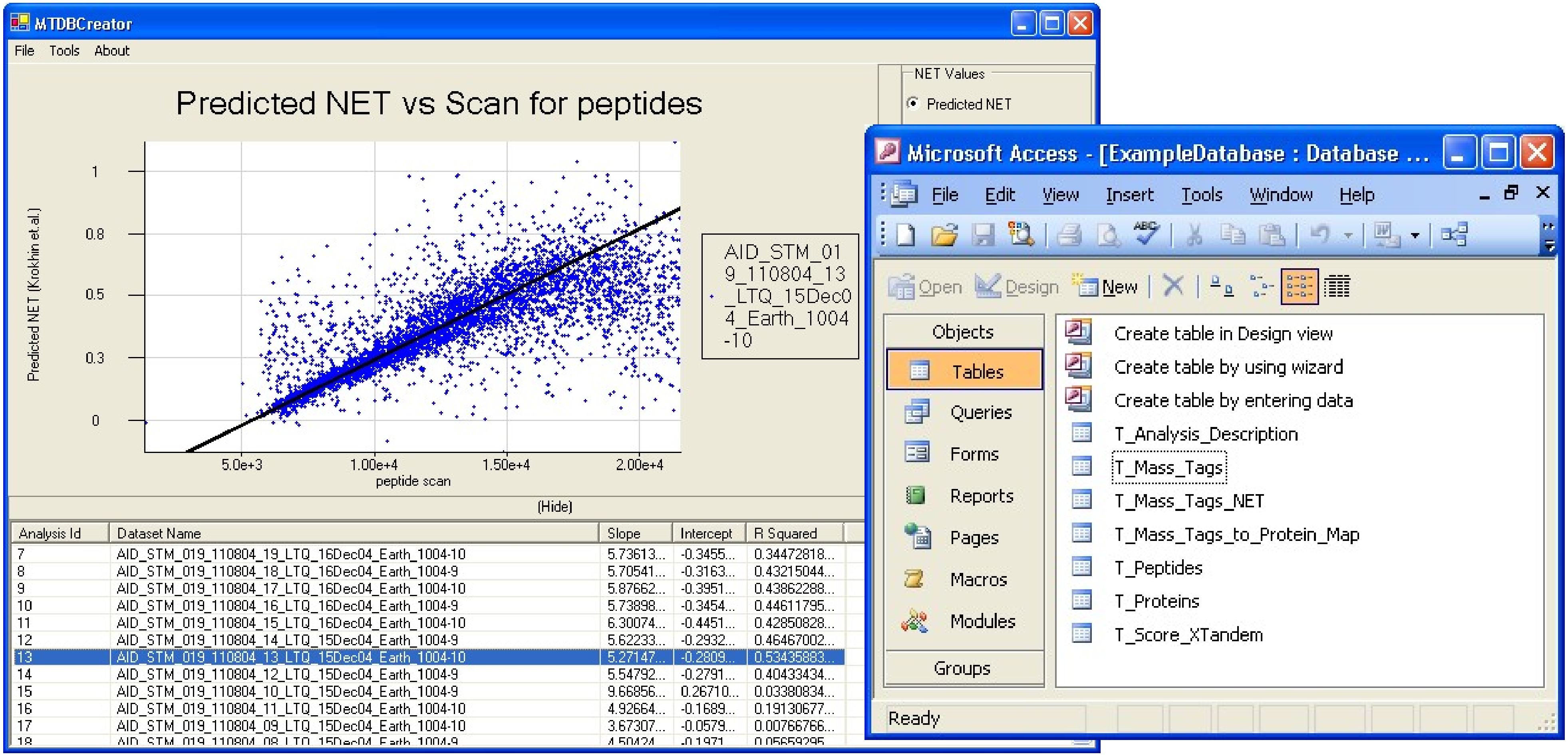
Mod_Details	
PK,FK1	**Unique_Seq_ID**
	Mass_Correction_Tag Position

Seq_to_Protein_Map	
PK,FK1 **PK**	**Unique_Seq_ID** **Protein_Name**

	Cleavage_State
	Terminus_State
	Protein_Expectation_Value_Log(e)
	Protein_Intensity_Log(I)

MTDB Creator

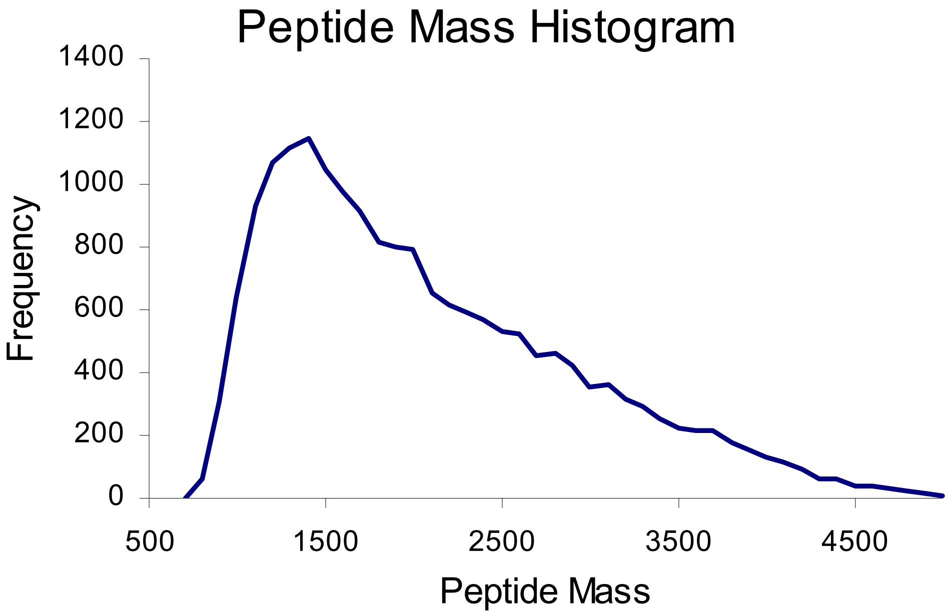
- MTDB Creator application
  - Allows external researchers to align multiple LC-MS/MS analyses, run PeptideProphet (for SEQUEST data) and create a standalone AMT tag database

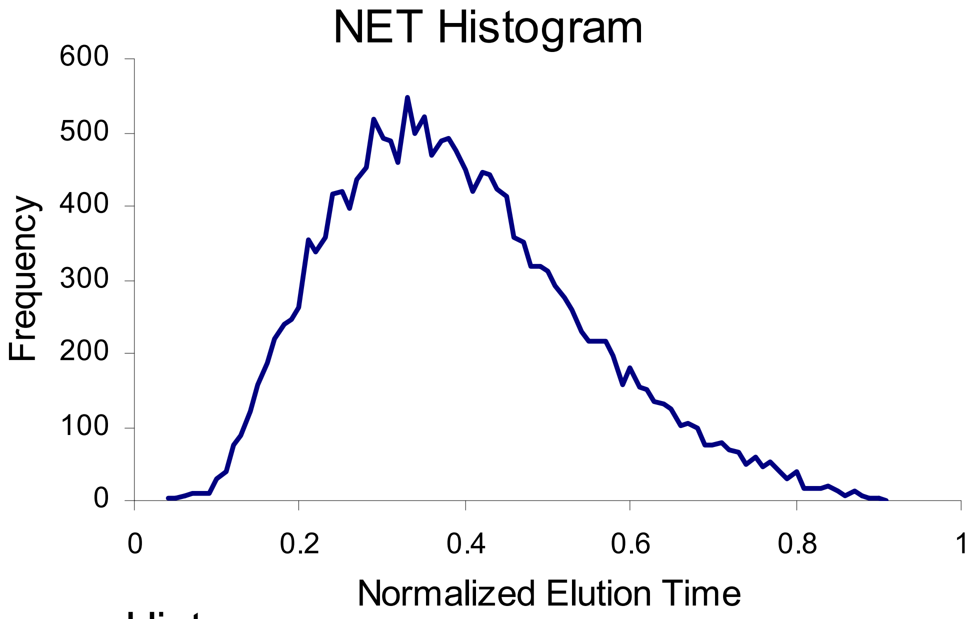

MTDBCreator
File Tools About
Predicted NET vs Scan for peptides
Predicted NET (Krokhin et.al.)
1
0.8
0.5
0.3
0
5.0e+3
1.00e+4
1.50e+4
2.00e+4
peptide scan
AID_STM_019_110804_13_LTQ_15Dec04_Earth_1004-10
NET Values
Predicted NET
(Hide)
Microsoft Access - [ExampleDatabase : Database ...
File Edit View Insert Tools Window Help
Open Design New
Objects
Tables
Queries
Forms
Reports
Pages
Macros
Modules
Groups
Create table in Design view
Create table by using wizard
Create table by entering data
T_Analysis_Description
T_Mass_Tags
T_Mass_Tags_NET
T_Mass_Tags_to_Protein_Map
T_Peptides
T_Proteins
T_Score_XTandem
Ready

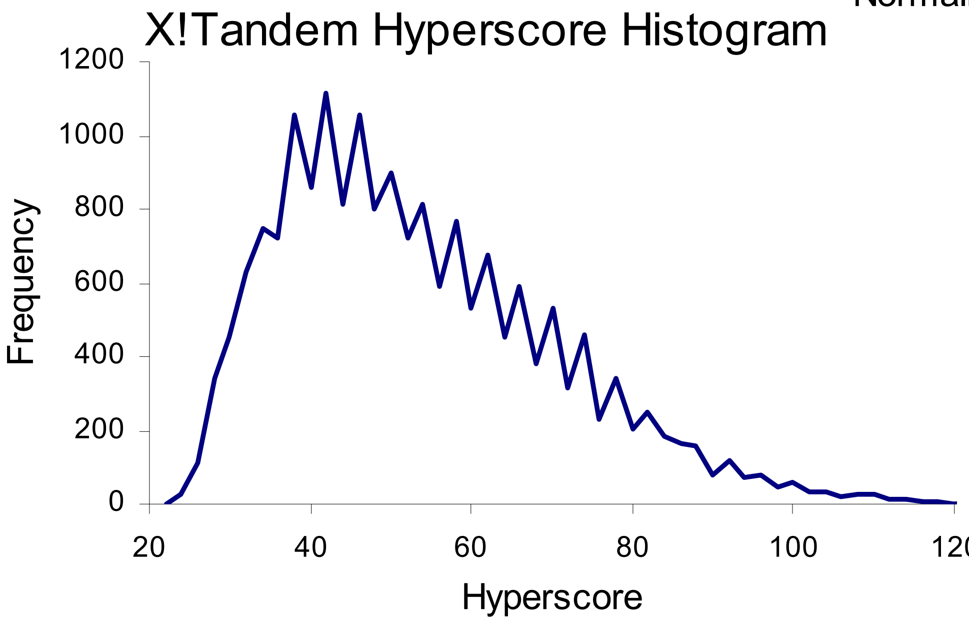
Analysis Id	Dataset Name	Slope	Intercept	R Squared
7	AID_STM_019_110804_19_LTQ_16Dec04_Earth_1004-10	5.73613...	-0.3455...	0.34472818...
8	AID_STM_019_110804_18_LTQ_16Dec04_Earth_1004-9	5.70541...	-0.3163...	0.43215044...
9	AID_STM_019_110804_17_LTQ_16Dec04_Earth_1004-10	5.87662...	-0.3951...	0.43862288...
10	AID_STM_019_110804_16_LTQ_16Dec04_Earth_1004-9	5.73898...	-0.3454...	0.44611795...
11	AID_STM_019_110804_15_LTQ_16Dec04_Earth_1004-10	6.30074...	-0.4451...	0.42850828...
12	AID_STM_019_110804_14_LTQ_15Dec04_Earth_1004-9	5.62233...	-0.2932...	0.46467002...
13	AID_STM_019_110804_13_LTQ_15Dec04_Earth_1004-10	5.27147...	-0.2809...	0.53435883...
14	AID_STM_019_110804_12_LTQ_15Dec04_Earth_1004-9	5.54792...	-0.2791...	0.40433434...
15	AID_STM_019_110804_10_LTQ_15Dec04_Earth_1004-9	9.66856...	0.26710...	0.03380834...
16	AID_STM_019_110804_11_LTQ_15Dec04_Earth_1004-10	4.92664...	-0.1689...	0.19130677...
17	AID_STM_019_110804_09_LTQ_15Dec04_Earth_1004-10	3.67307...	-0.0579...	0.00766766...
18	AID_STM_019_110804_08_LTQ_15Dec04_Earth_1004-9	4.50424	-0.1971	0.05659295

Assembling an AMT tag DB

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

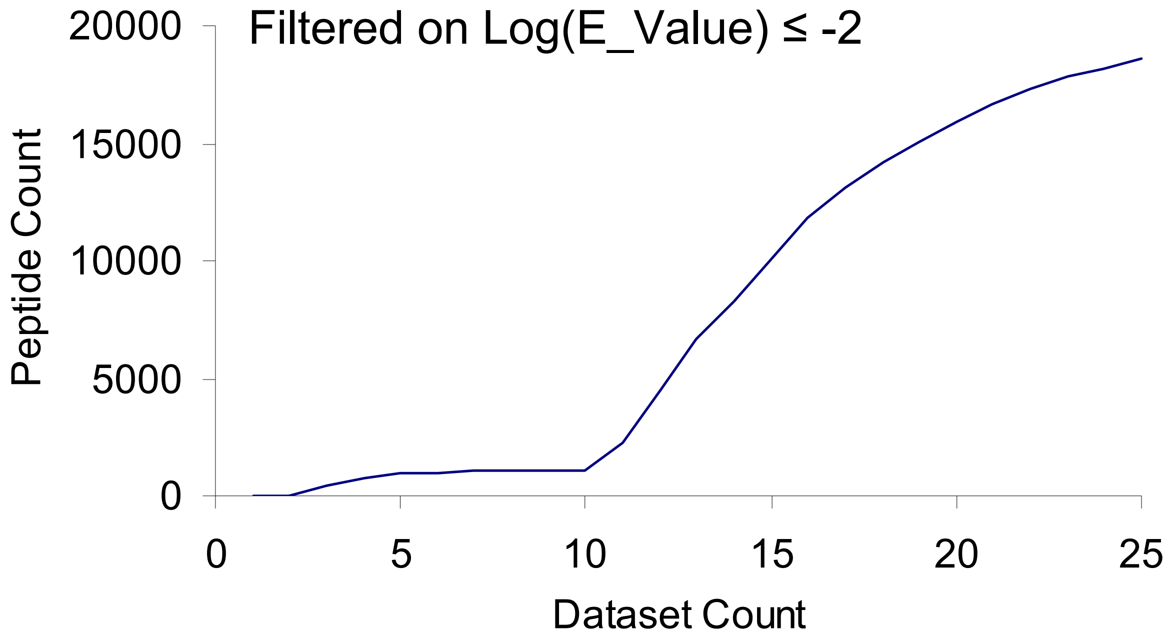






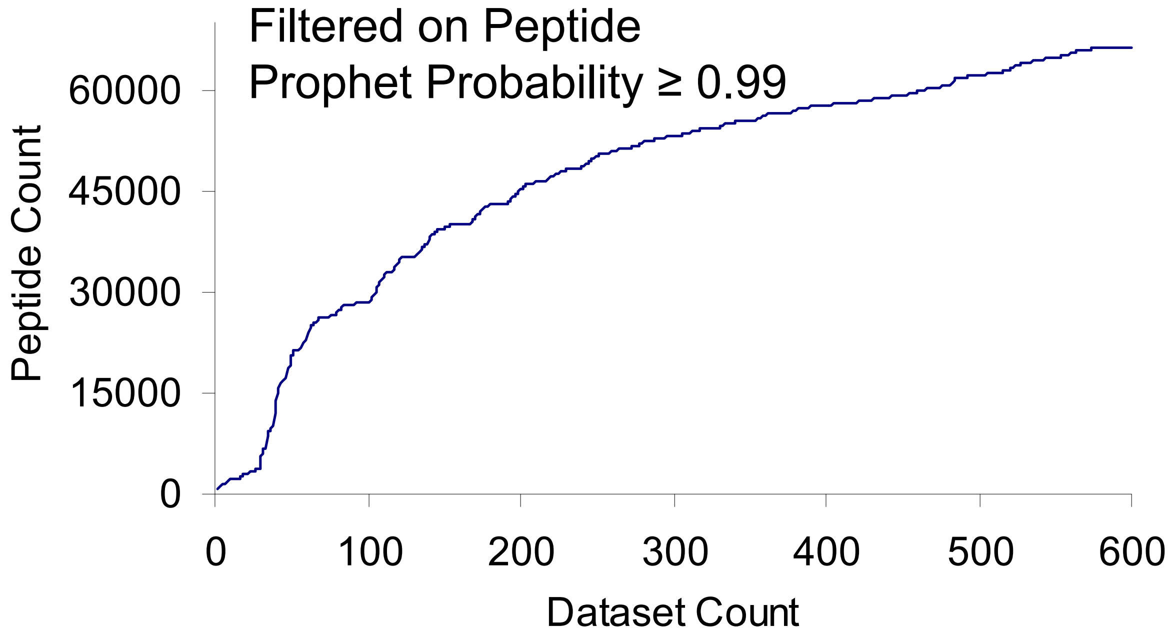
AMT Tag DB Growth Trend

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



Filtered on Log(E_Value) ≤ -2

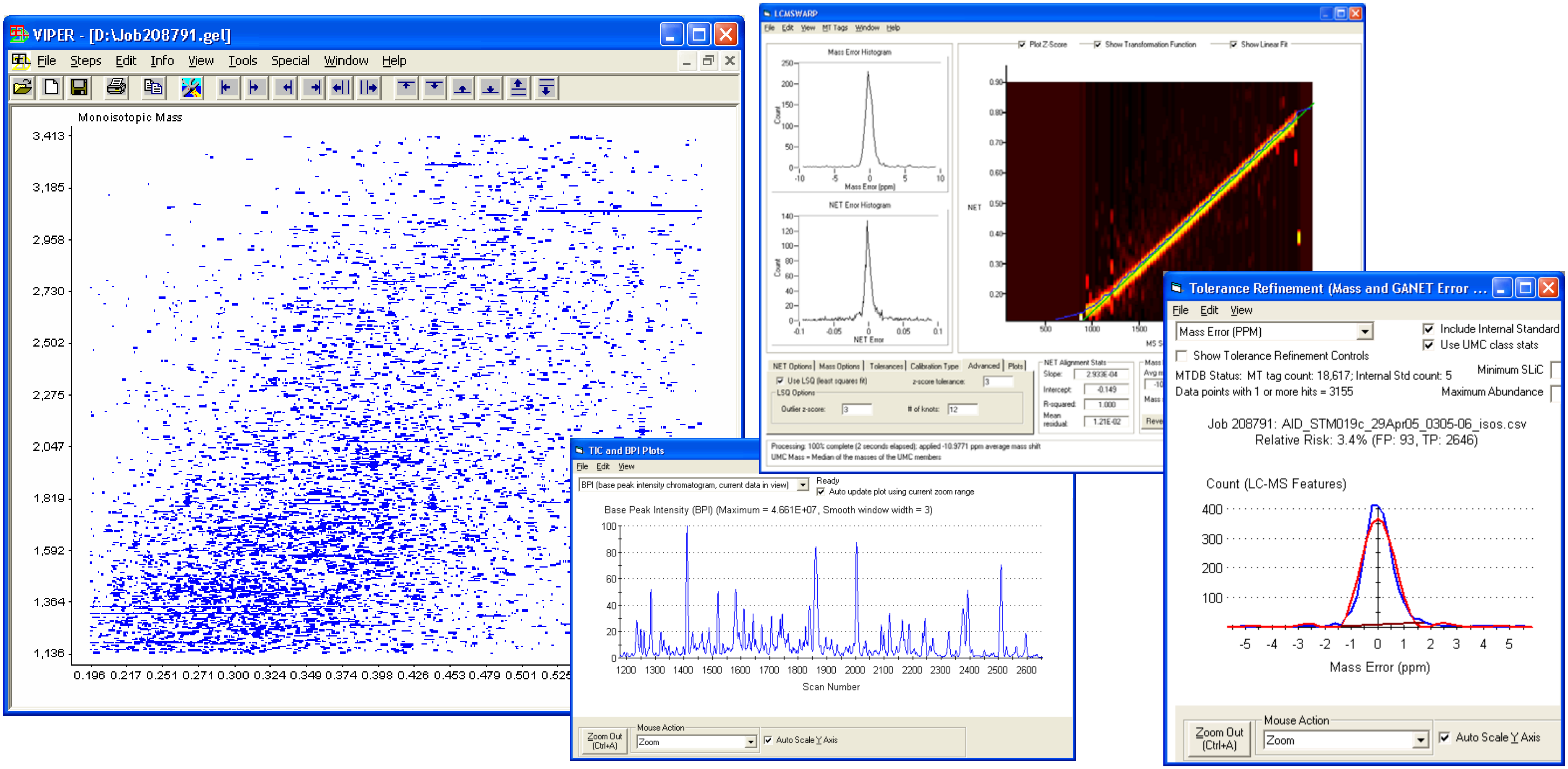
- Trend for Mature AMT tag DB
  - 521 different samples from ~25 different conditions
  - Slope of curve decreases as more datasets are added and as fewer new peptides are seen



Filtered on Peptide Prophet Probability ≥ 0.99

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



☑ Use LSQ (least squares fit)　z-score tolerance: 3

LSQ Options: Outlier z-score: 3　# of knots: 12

NET Alignment Stats: Slope: 2.933E-04　Intercept: -0.149　R-squared: 1.000　Mean residual: 1.21E-02

Processing: 100% complete (2 seconds elapsed); applied -10.9771 ppm average mass shift
UMC Mass = Median of the masses of the UMC members

BPI (base peak intensity chromatogram, current data in view)　Ready

☑ Auto update plot using current zoom range

Base Peak Intensity (BPI) (Maximum = 4.661E+07, Smooth window width = 3)

Mouse Action: Zoom　☑ Auto Scale Y Axis

Mass Error (PPM)　☑ Include Internal Standard　☑ Use UMC class stats

□ Show Tolerance Refinement Controls

MTDB Status: MT tag count: 18,617; Internal Std count: 5

Data points with 1 or more hits = 3155

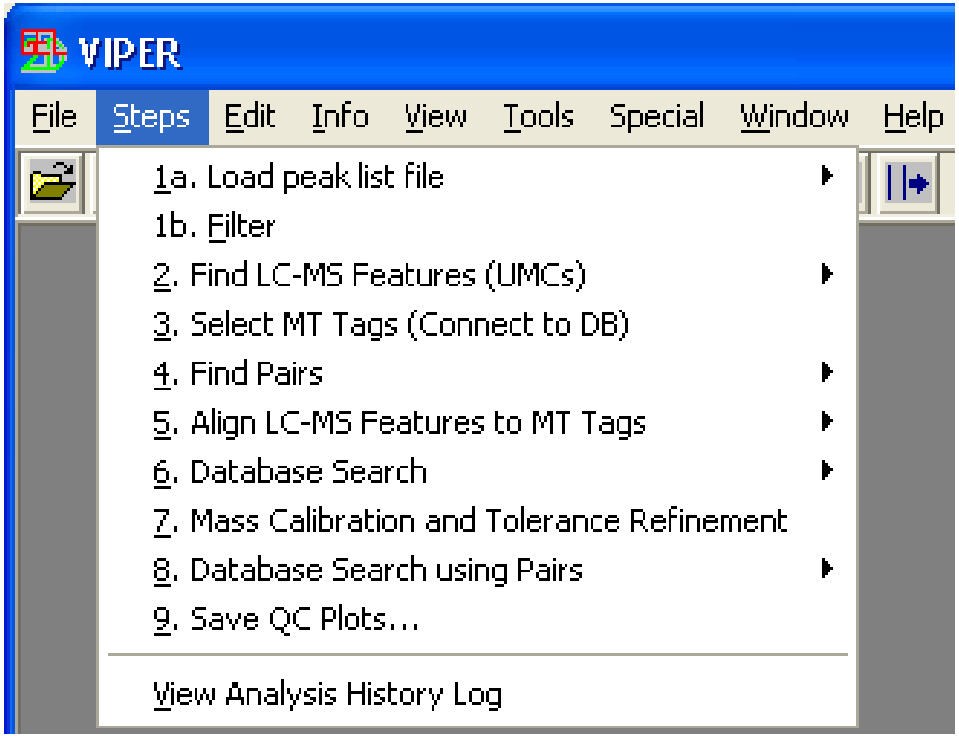
Minimum SLiC　Maximum Abundance

Job 208791: AID_STM019c_29Apr05_0305-06_isos.csv
Relative Risk: 3.4% (FP: 93, TP: 2646)

Mouse Action: Zoom　☑ Auto Scale Y Axis

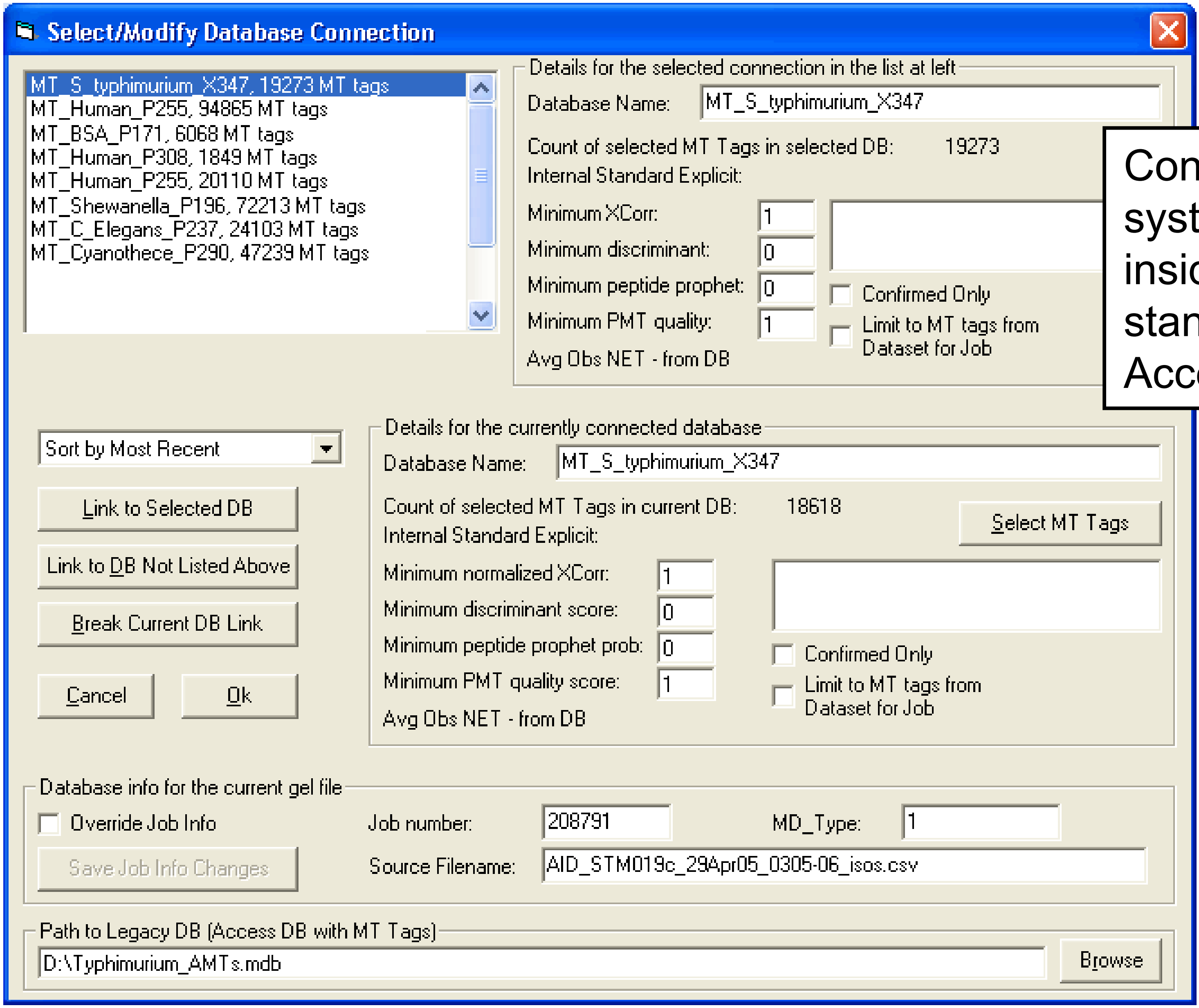
Identifying LC-MS Features

- 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


VIPER
File
Steps
Edit
Info
View
Tools
Special
Window
Help
1a. Load peak list file
1b. Filter
2. Find LC-MS Features (UMCs)
3. Select MT Tags (Connect to DB)
4. Find Pairs
5. Align LC-MS Features to MT Tags
6. Database Search
7. Mass Calibration and Tolerance Refinement
8. Database Search using Pairs
9. Save QC Plots…
View Analysis History Log

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

Details for the currently connected database

Database Name: MT_S_typhimurium_X347

Count of selected MT Tags in current DB: 18618

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

Source Filename: AID_STM019c_29Apr05_0305-06_isos.csv

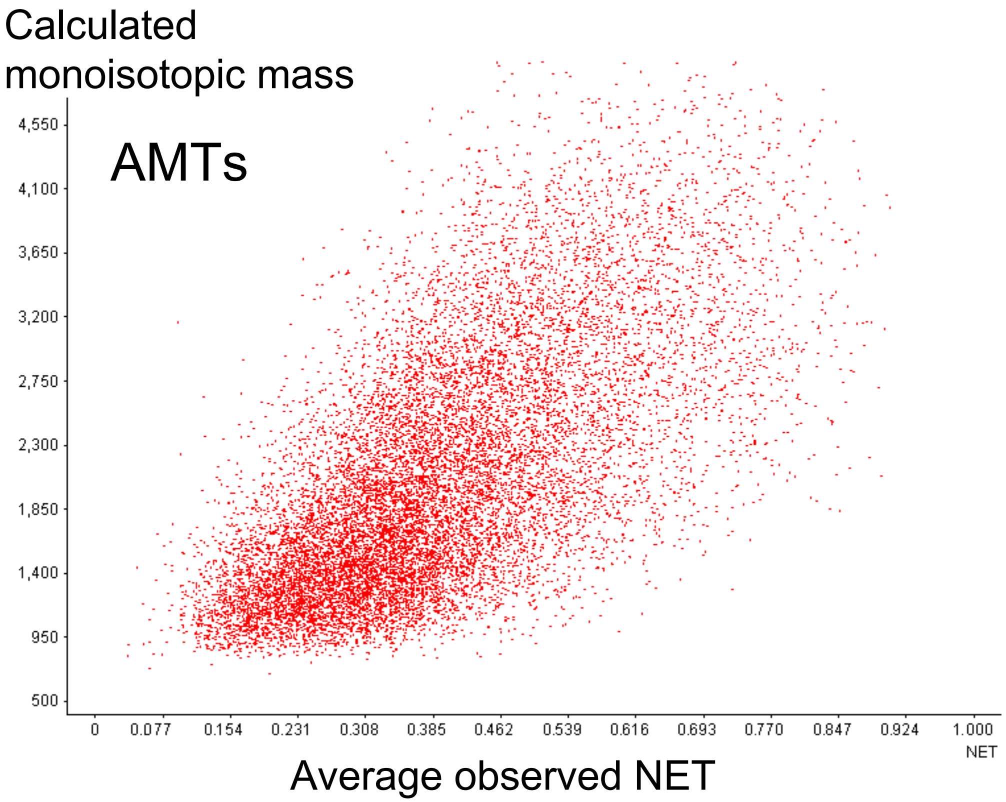
Path to Legacy DB (Access DB with MT Tags)

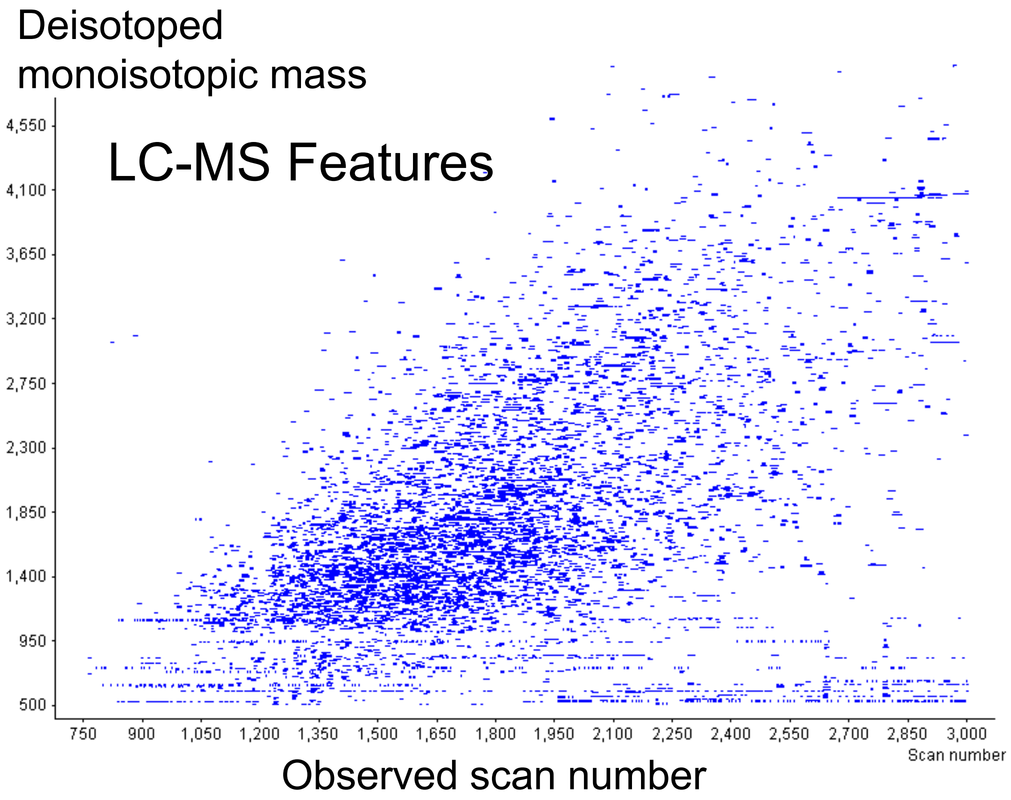
D:\Typhimurium_AMTs.mdb

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

Alignment using LCMSWarp

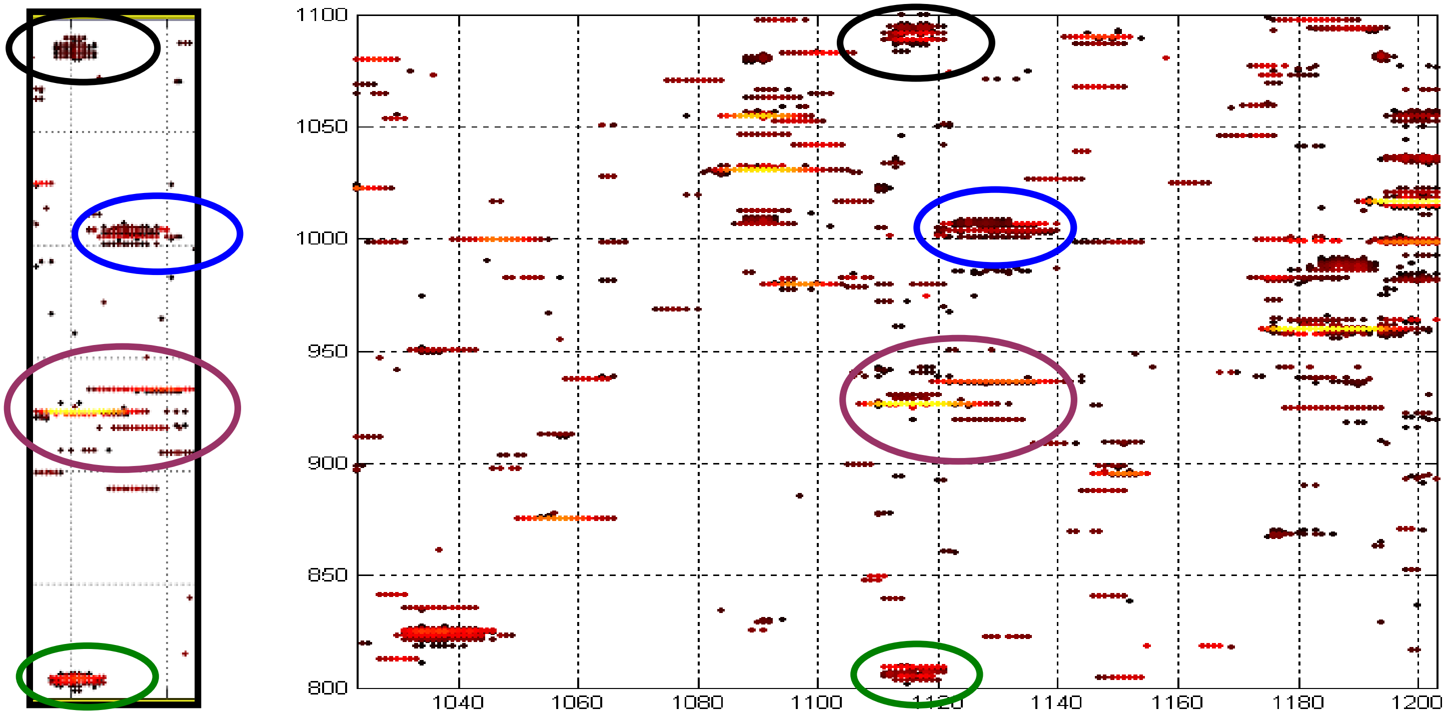
- Align scan number (i.e. elution time) of features to NETs of peptides in given AMT tag database
  - Match mass and NET of AMT tags to mass and scan number of MS features
  - Use LCMSWarp algorithm to find optimal alignment to give the most matches

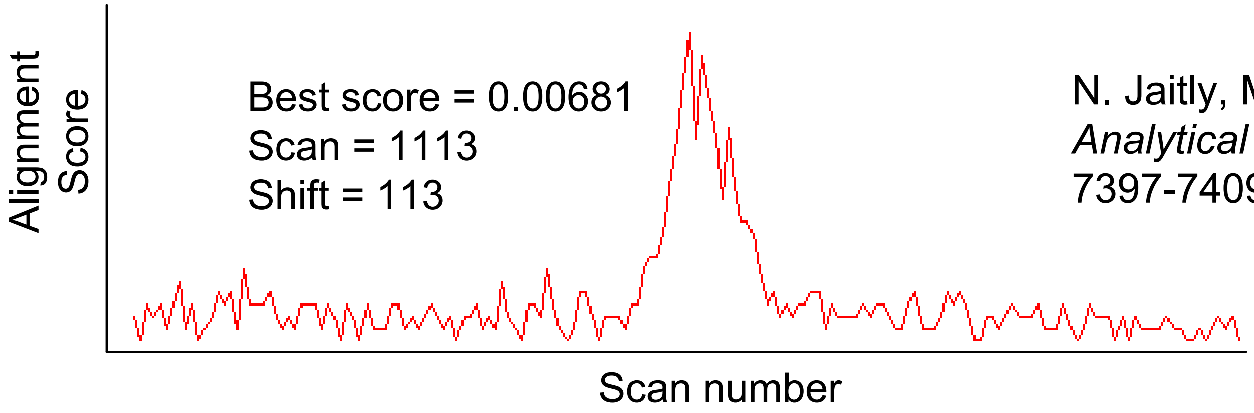




Alignment using LCMSWarp

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



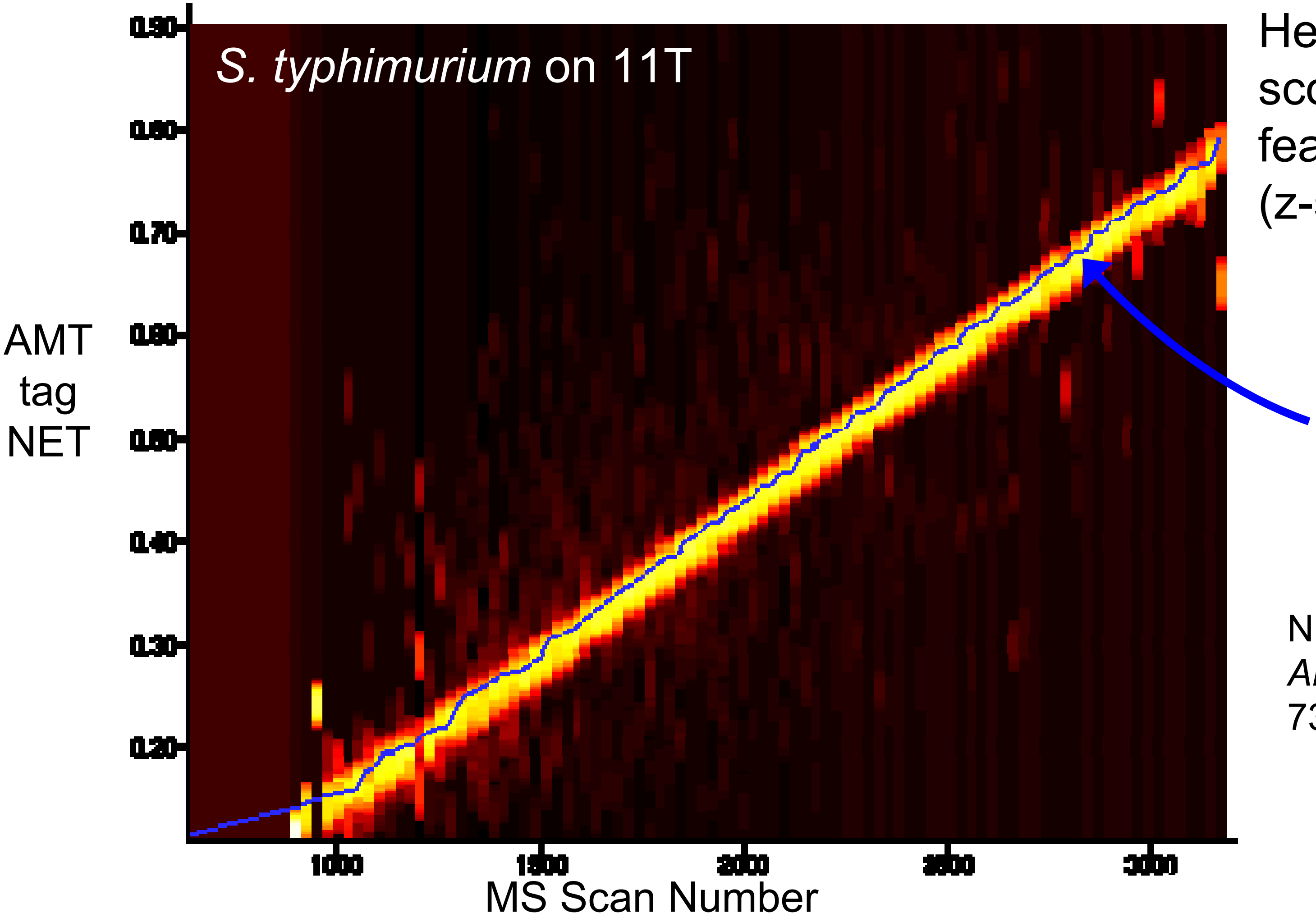


Best score = 0.00681
Scan = 1113
Shift = 113

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)

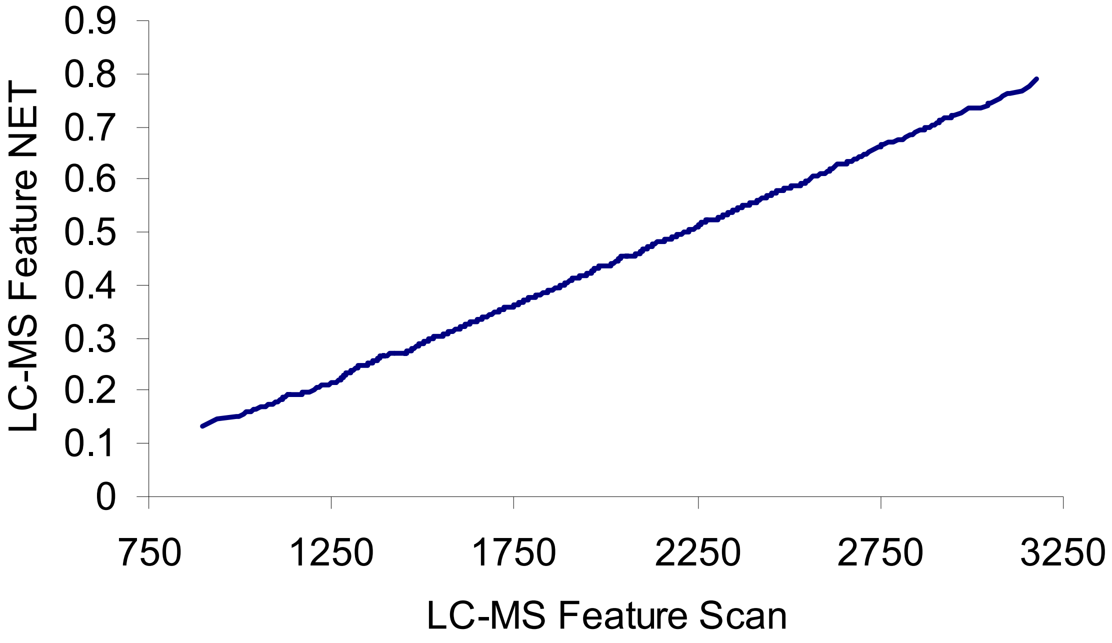
Alignment Function

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

Alignment using LCMSWarp

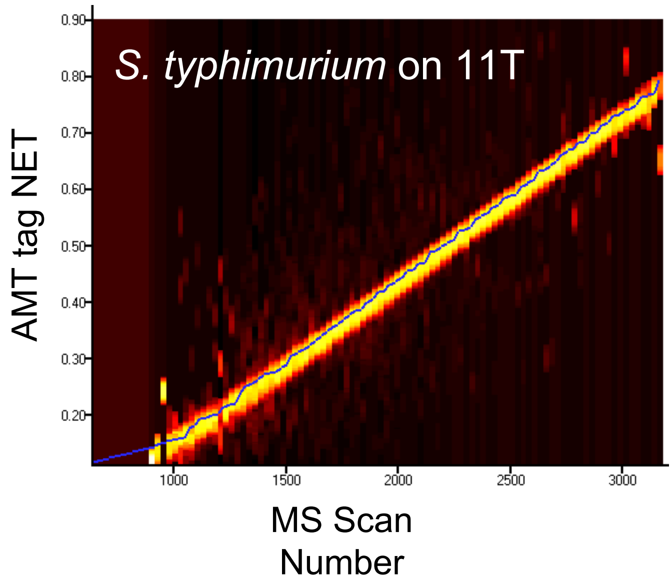
- Transformation function is used to convert from scan number to NET
  - Features centered at same scan number get the same obs. NET value
  - When matching LC-MS features to AMTs, we will search +/- a NET tolerance, which effectively allows for LC-MS features to shift around a little in elution time

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

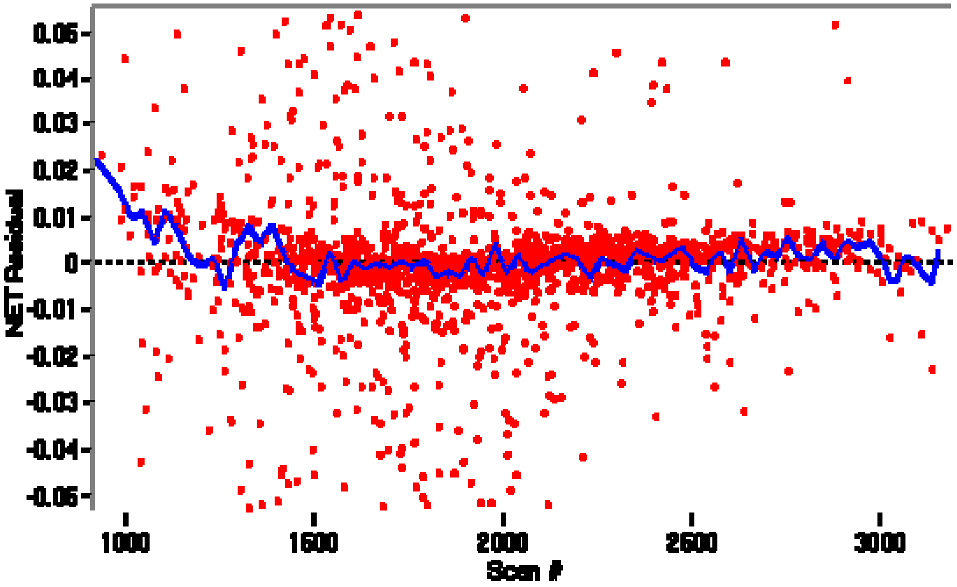


Alignment using LCMSWarp

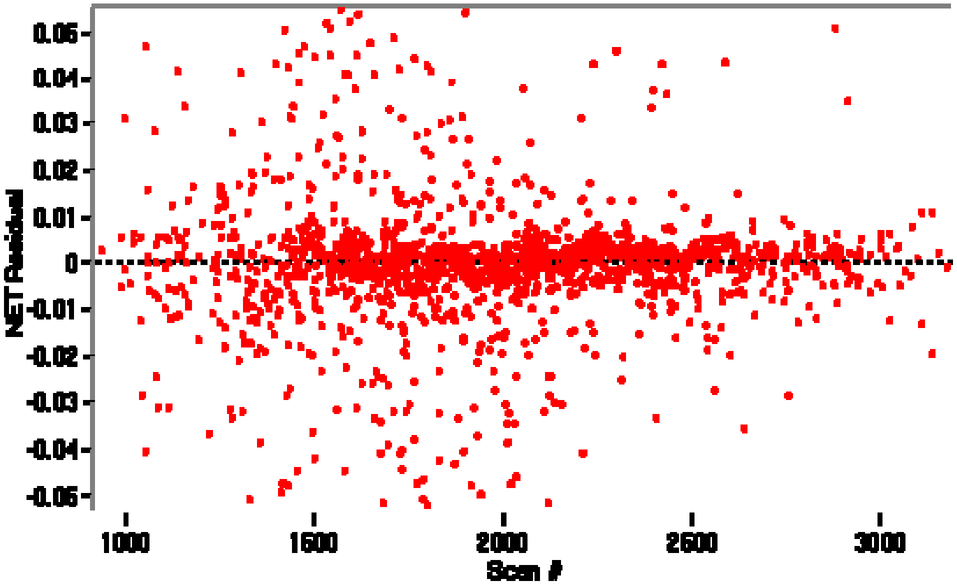
- NET Residual Plots
  - Difference between NET of LC-MS feature and NET of matching AMT tag
    - Indicates quality of alignment between features and AMT tags
  - This data shows nearly linear alignment between features and AMTs, but the algorithm can easily account for non-linear trends



NET Residuals if a linear mapping is used

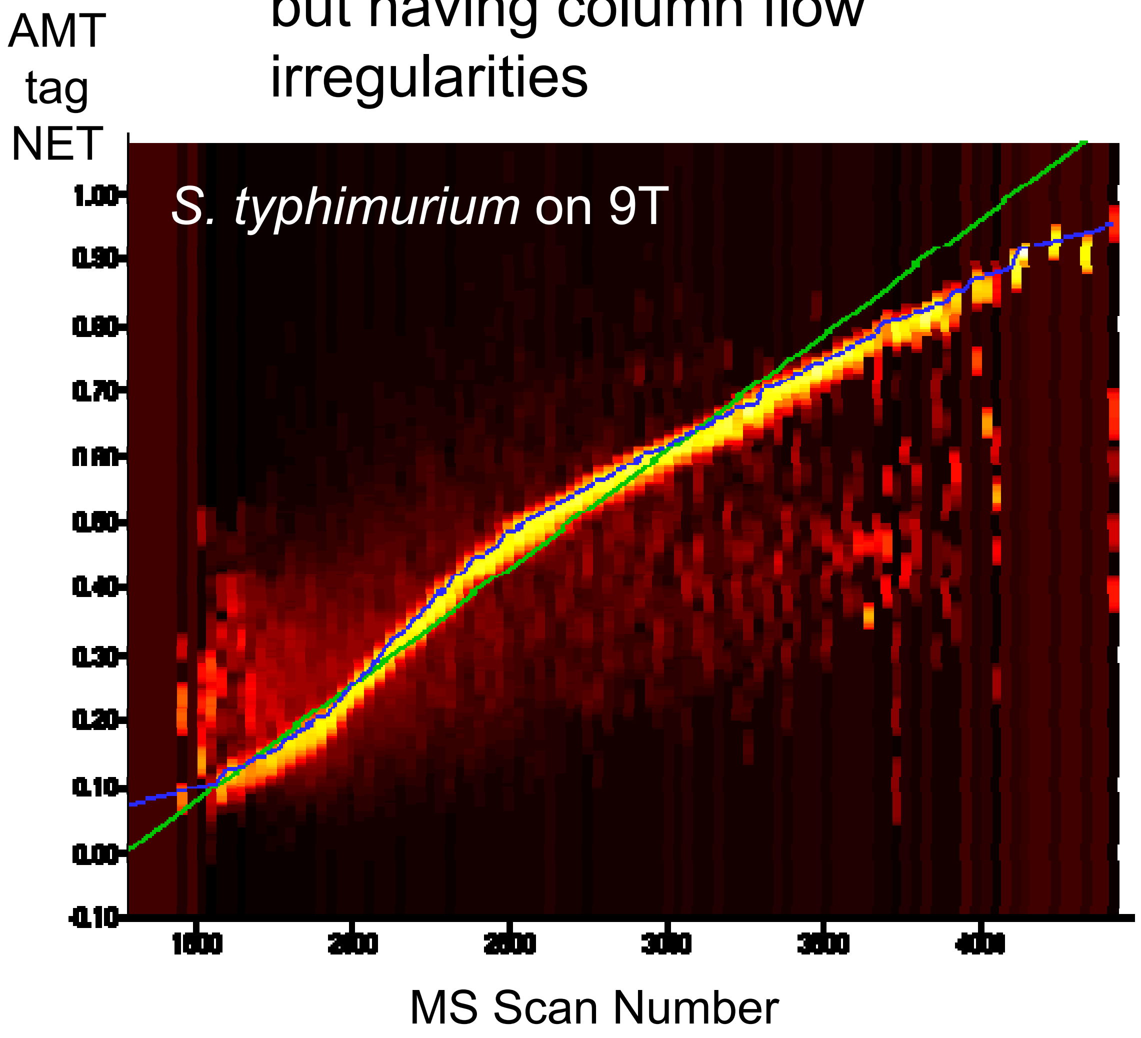


NET Residuals after LCMSWarp

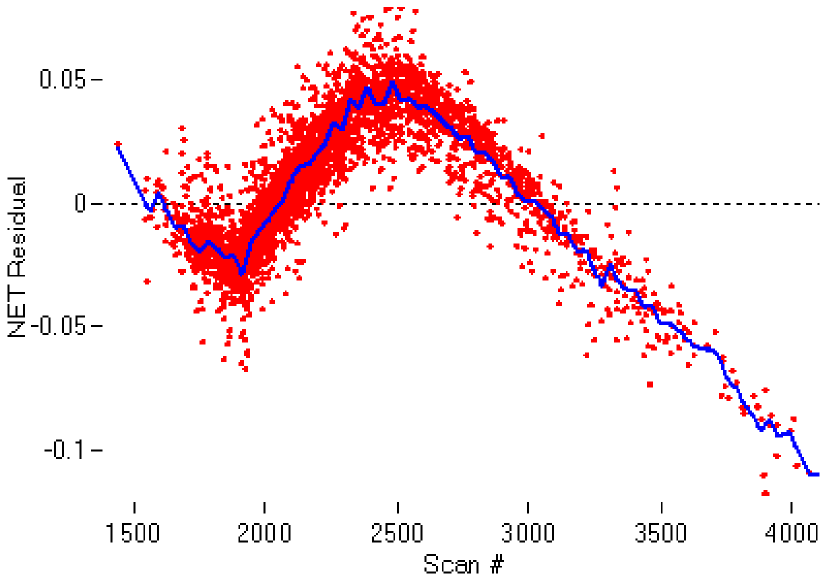


Alignment using LCMSWarp

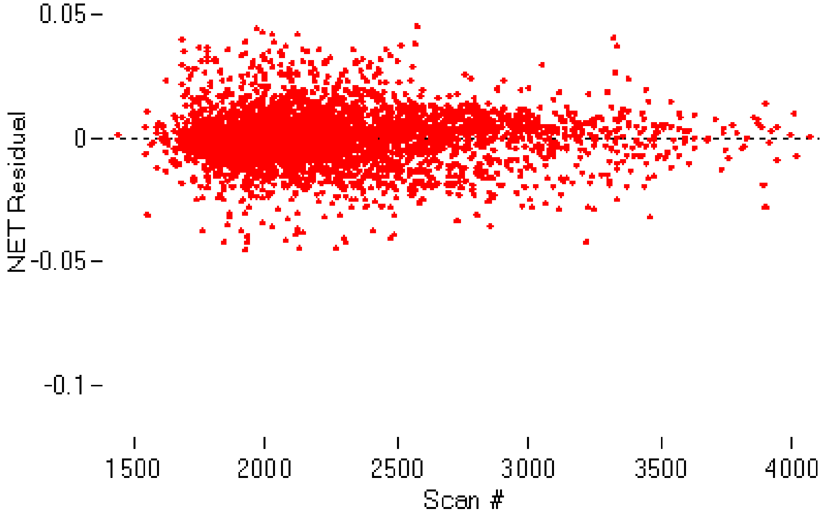
- Non-linear alignment example #1
  - Identical LC separation system, but having column flow irregularities



NET Residuals if a linear mapping is used

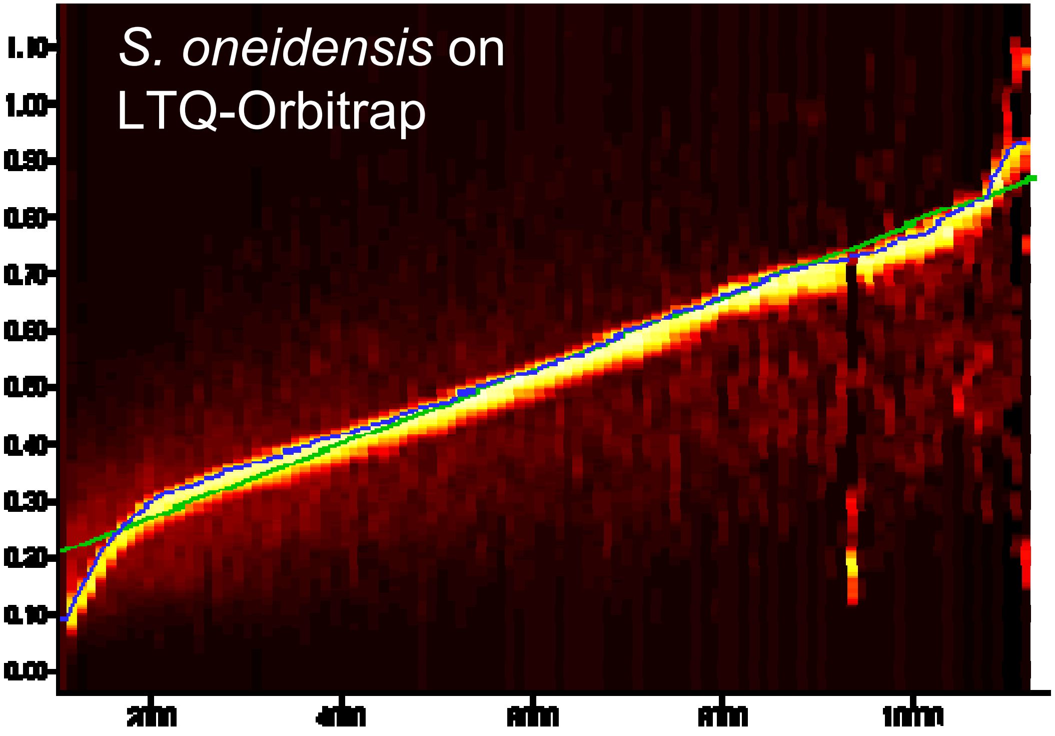


NET Residuals after LCMSWarp

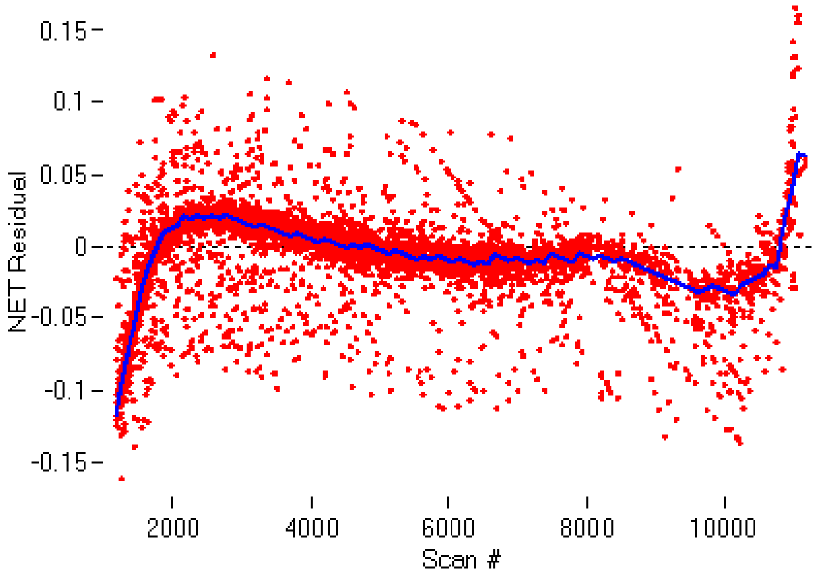


Alignment using LCMSWarp

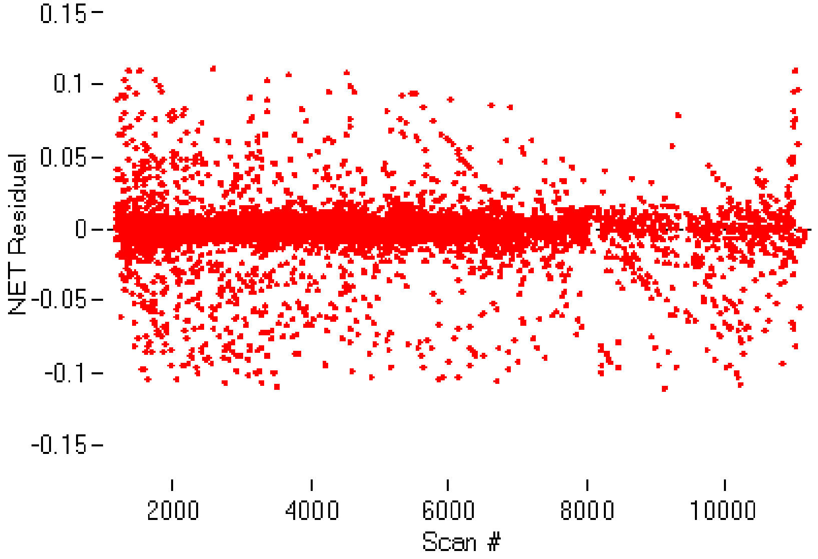
- Non-linear alignment example #2
  - AMT Tag DB from C_{18} LC-MS/MS analyses using ISCO-based LC (**exponential dilution gradient**)
  - LC-MS analysis used C_{18} LC-MS via Agilent **linear gradient** pump


S. oneidensis on LTQ-Orbitrap

NET Residuals if a linear mapping is used

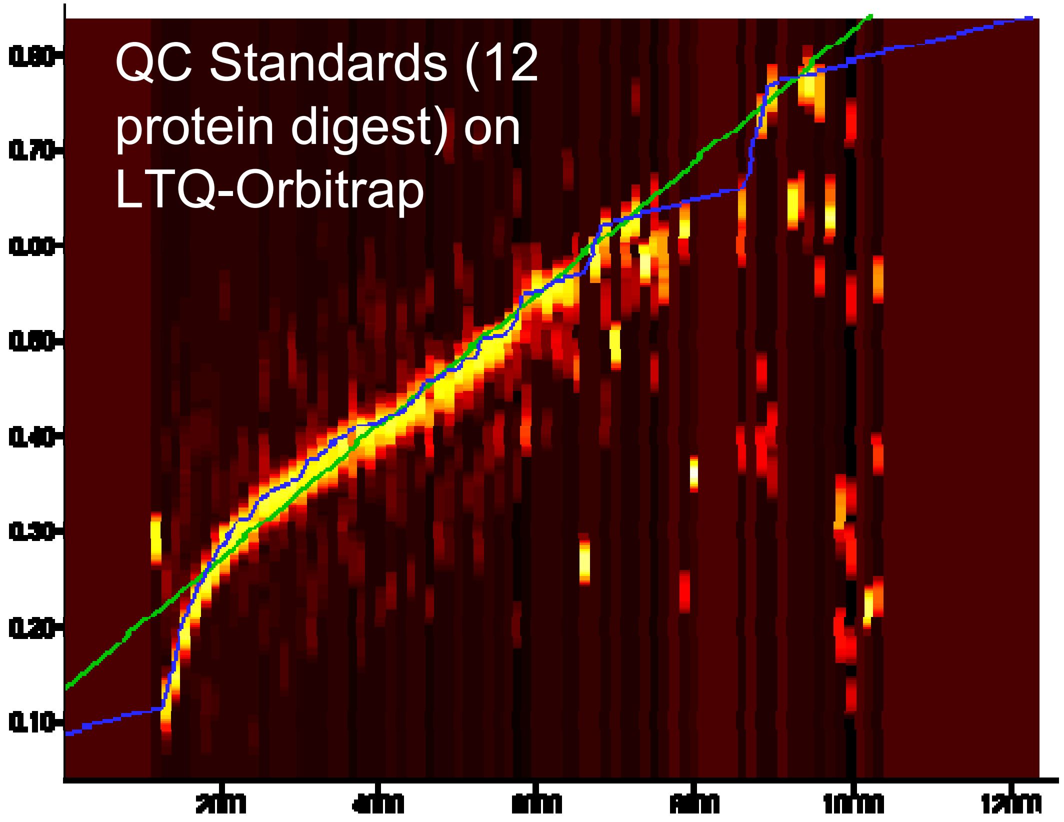


NET Residuals after LCMSWarp



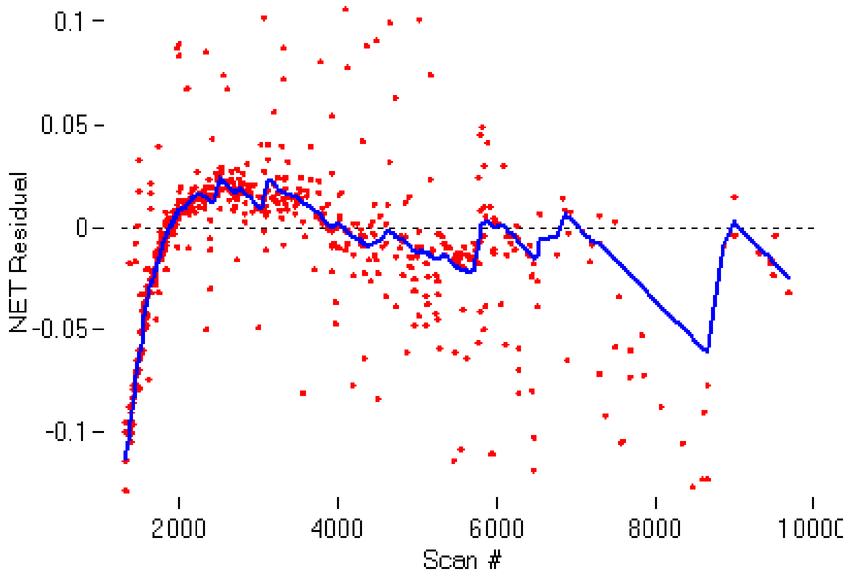
Alignment using LCMSWarp

- 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

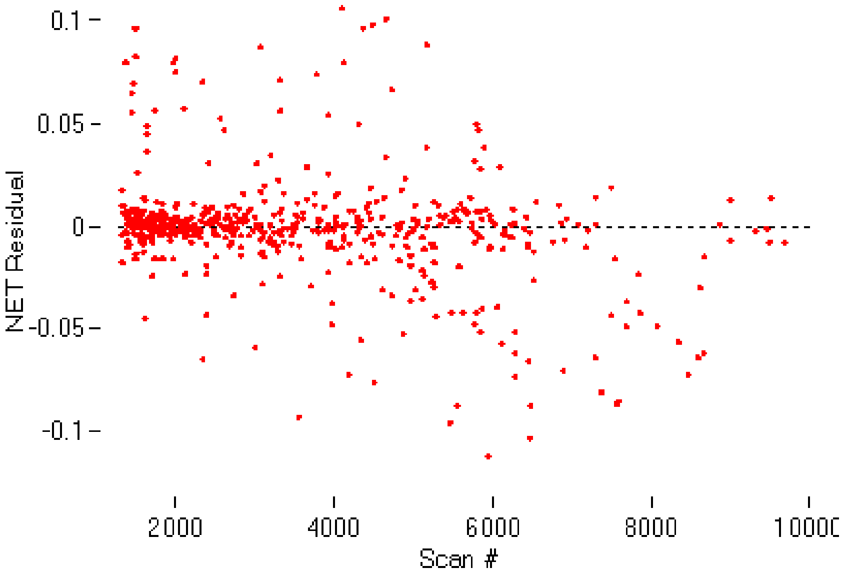


QC Standards (12 protein digest) on LTQ-Orbitrap

NET Residuals if a linear mapping is used

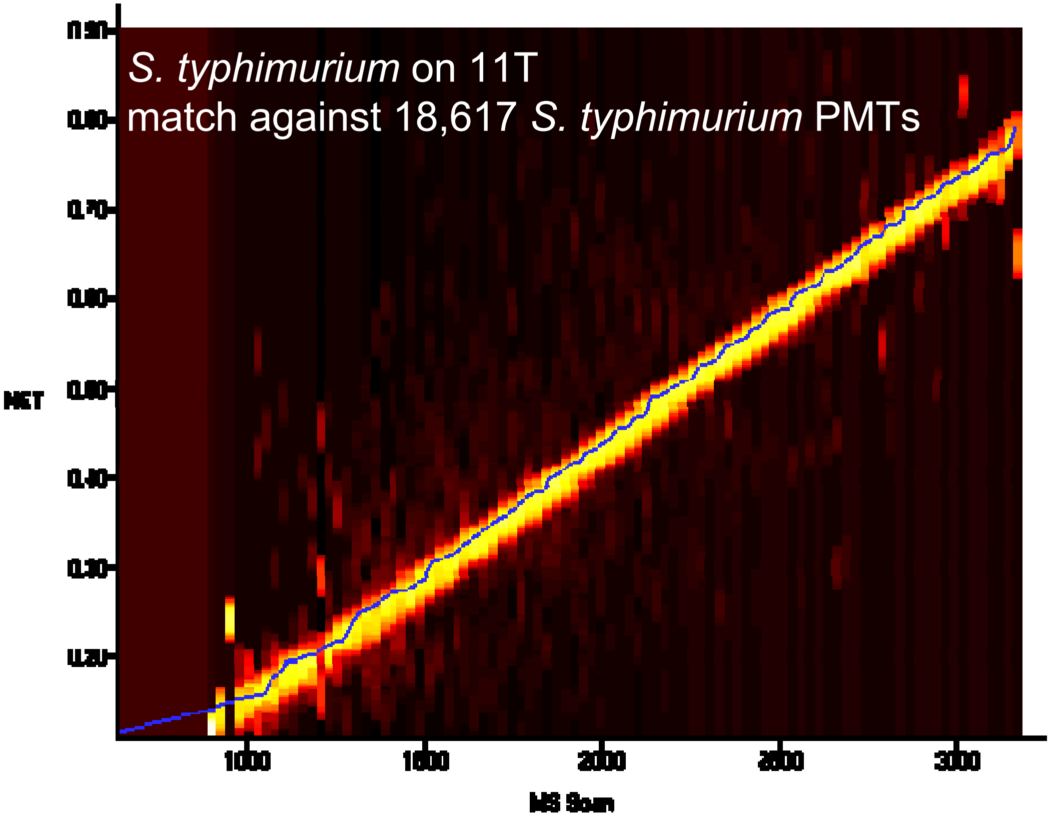


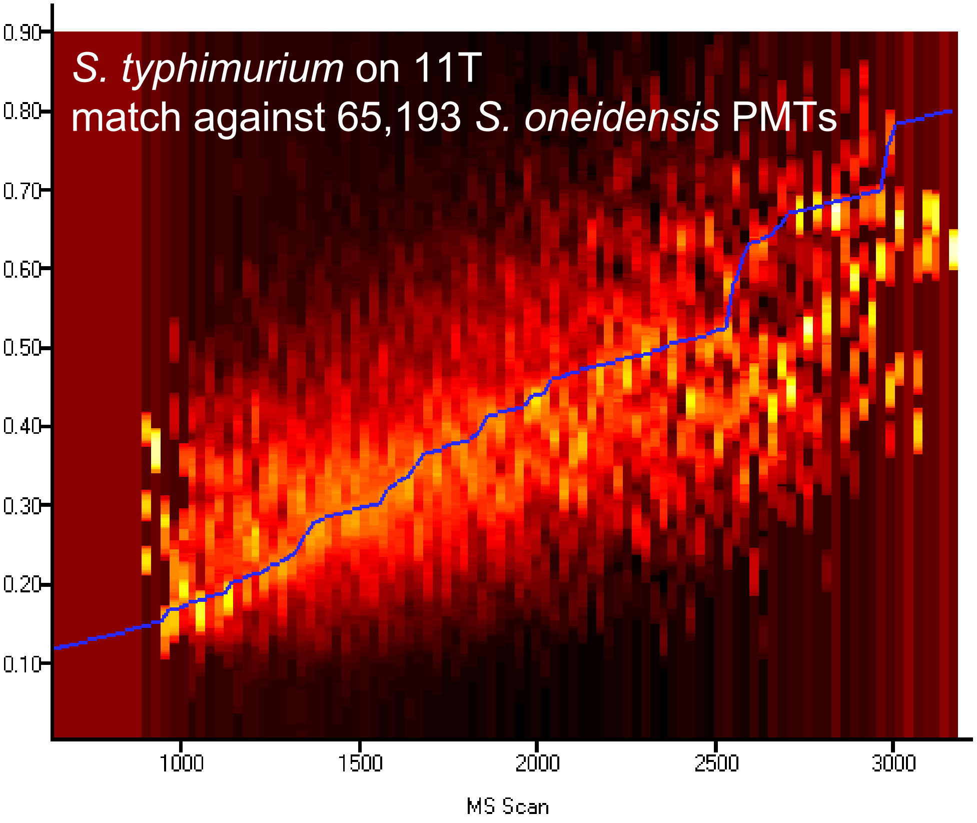
NET Residuals after LCMSWarp



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

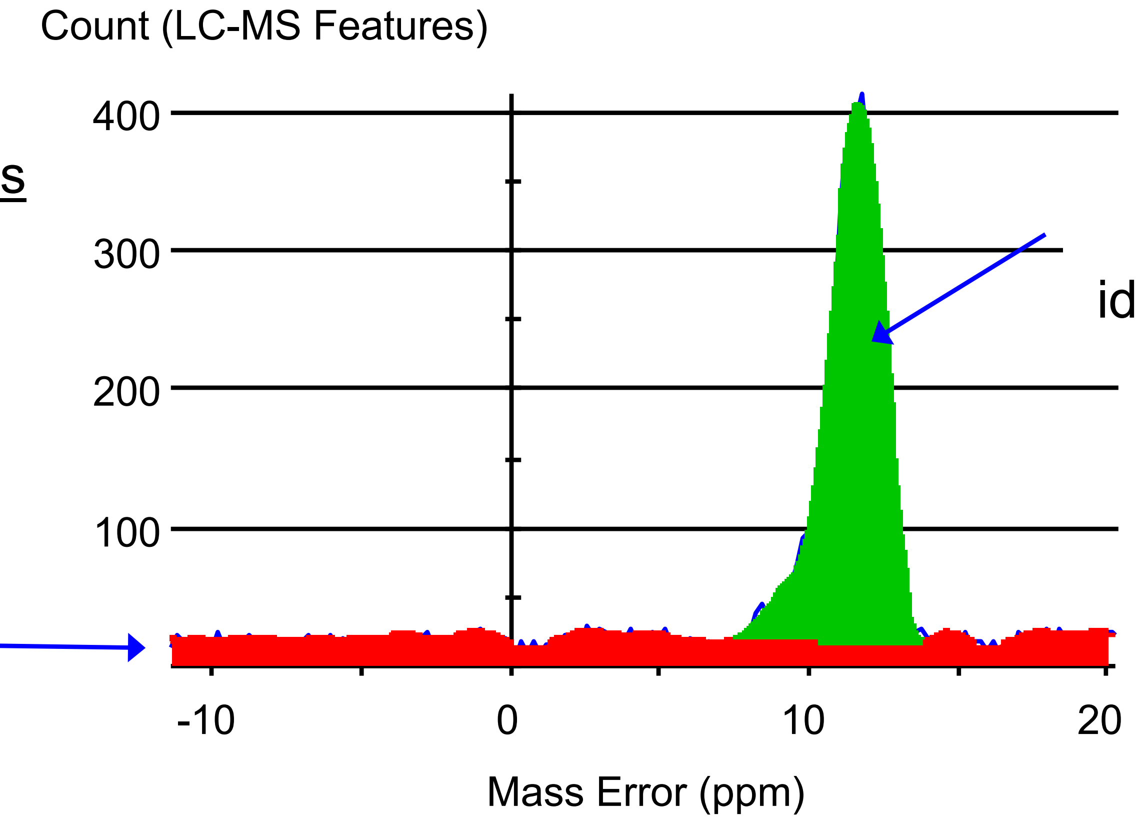
Mass Error Histogram

- 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

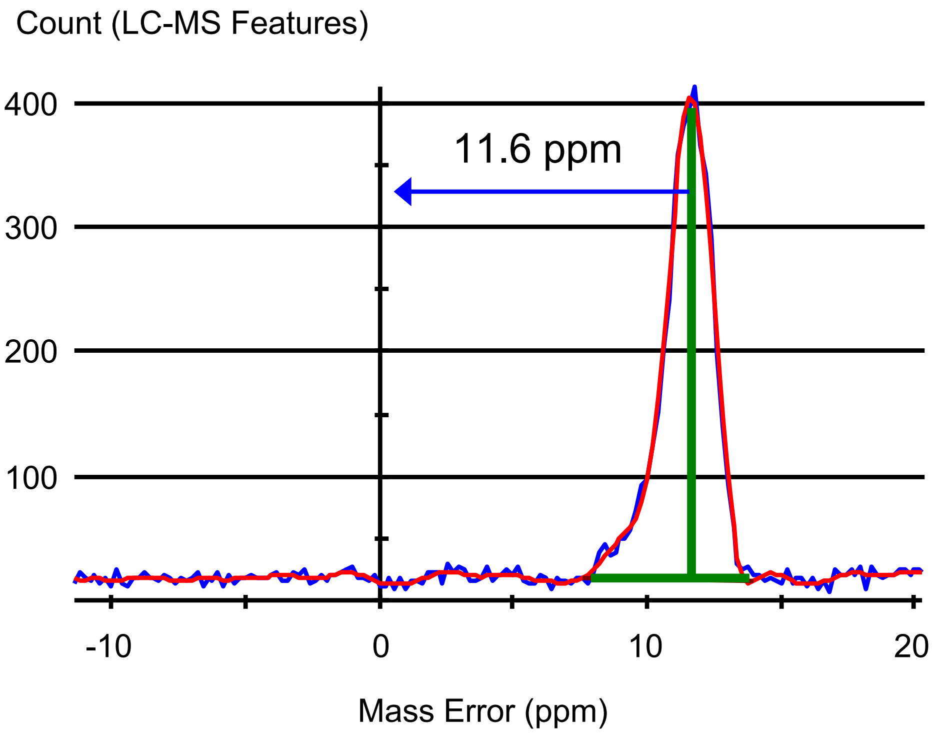


Likely true positive identifications

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:

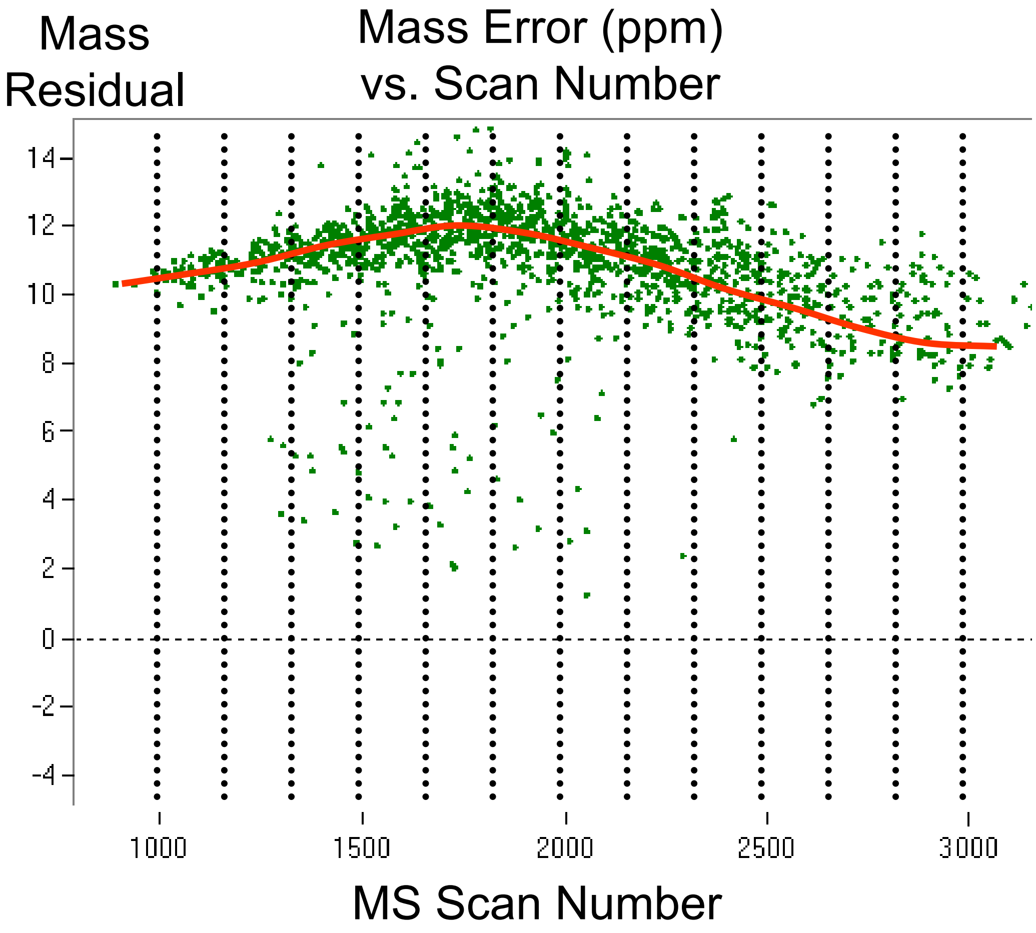
$$\Delta_{mass} = -11.6\text{ppm} \times \frac{\text{mass}_{old}}{1\text{x}10^{6}\ \text{ppm/Da}}$$

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

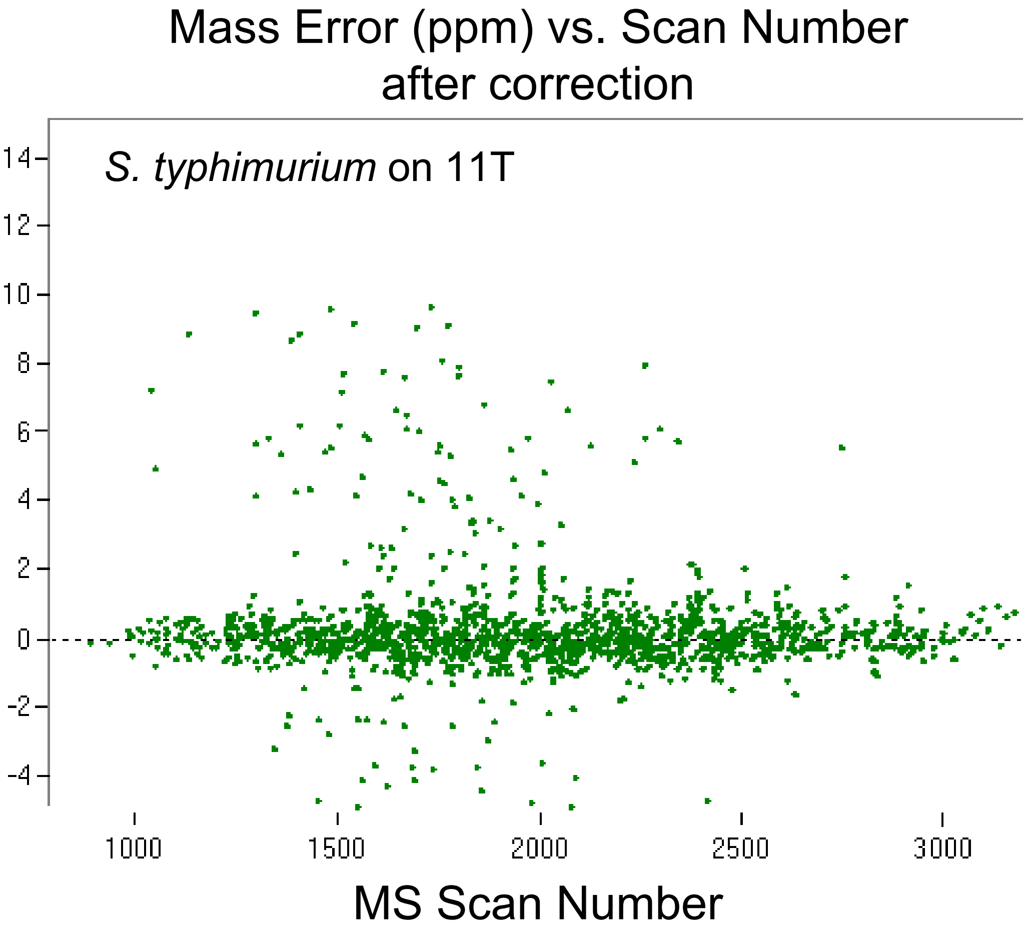
For 3+ feature at 2919.4658 Da,
Δ_{mass} = -0.0339 Da

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 Residual

Mass Error (ppm) vs. Scan Number

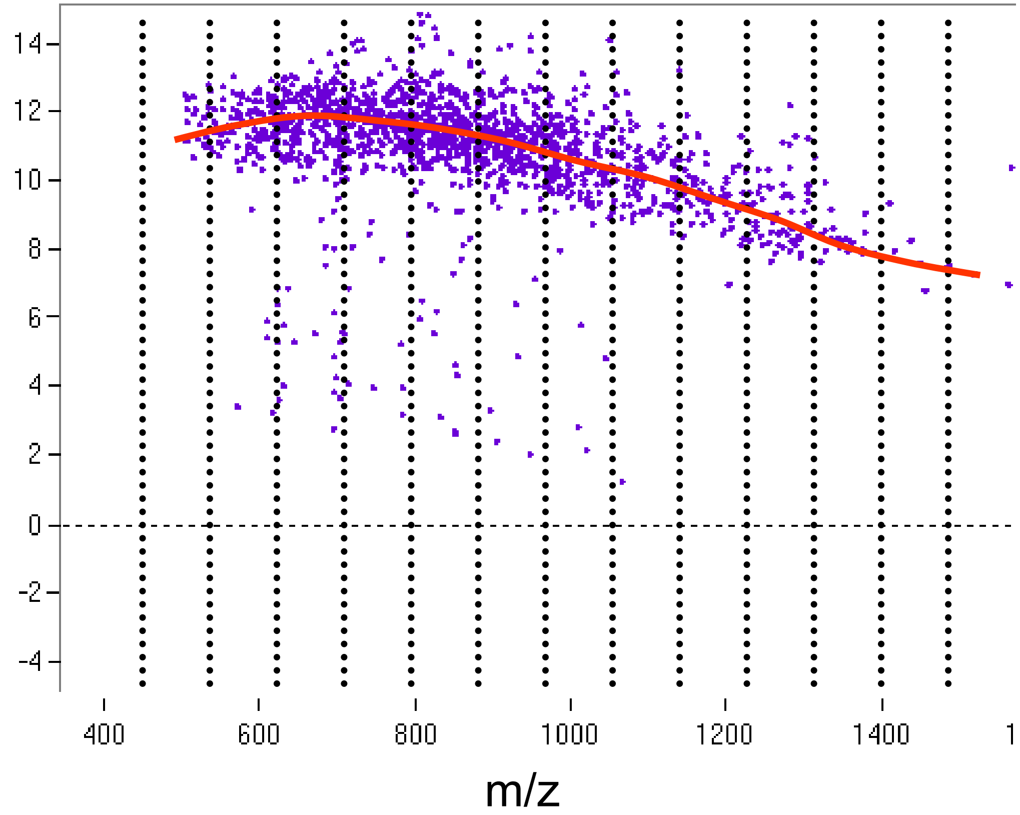

Mass Error (ppm) vs. Scan Number after correction

Mass Calibration

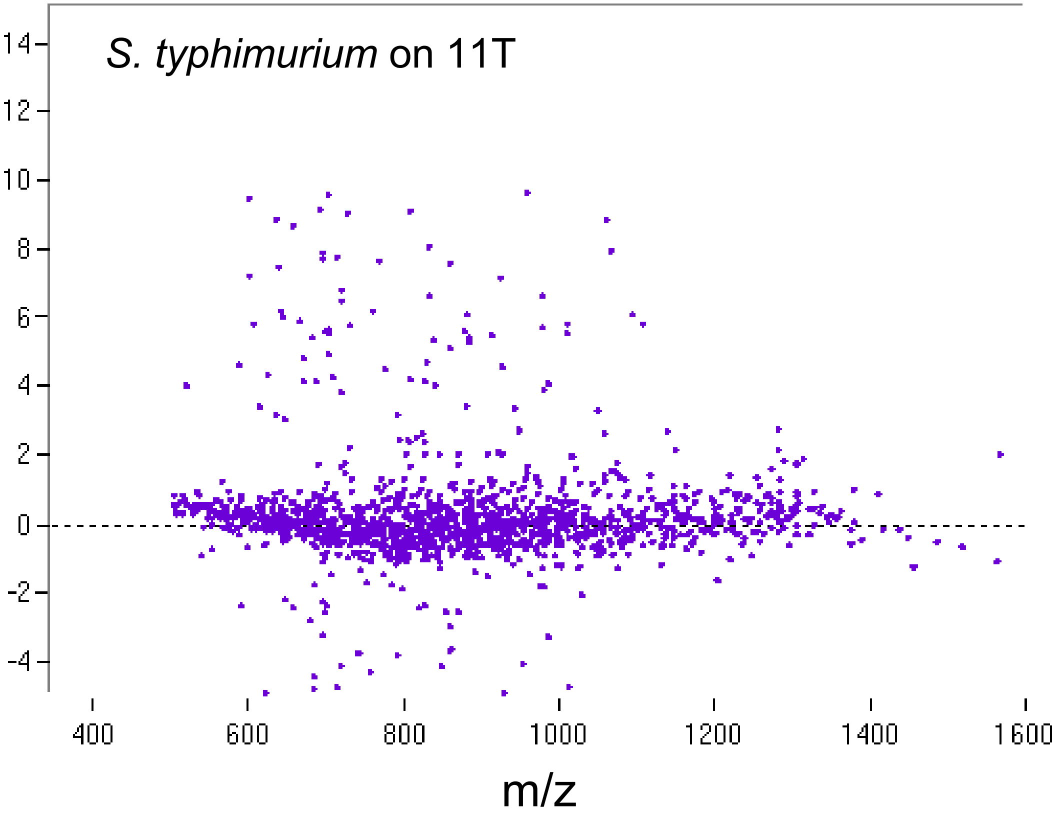
- Option 2: Piece-wise correction via LCMSWarp
  - Use a smoothing spline calibration which is a function of m/z
  - LCMSWarp utilizes a hybrid correction based on both mass error vs. time and mass error vs. m/z

Mass Residual

Mass Error (ppm) vs. m/z

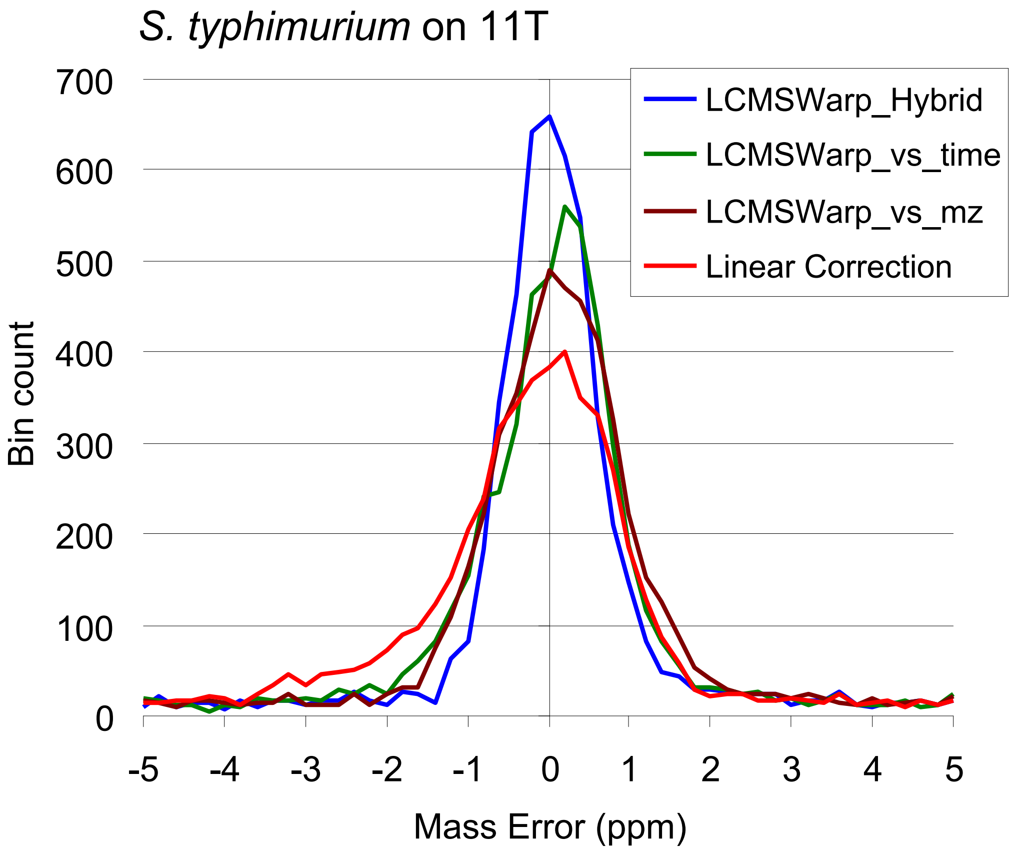


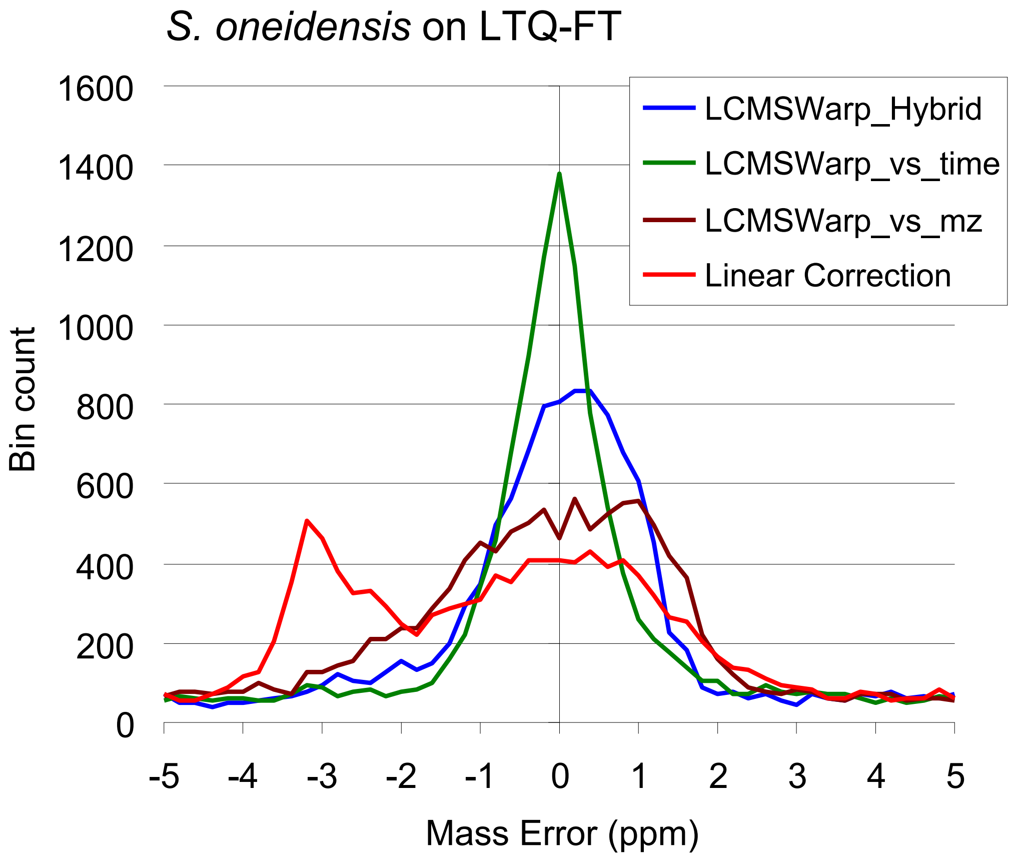
Mass Error (ppm) vs. m/z after correction



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

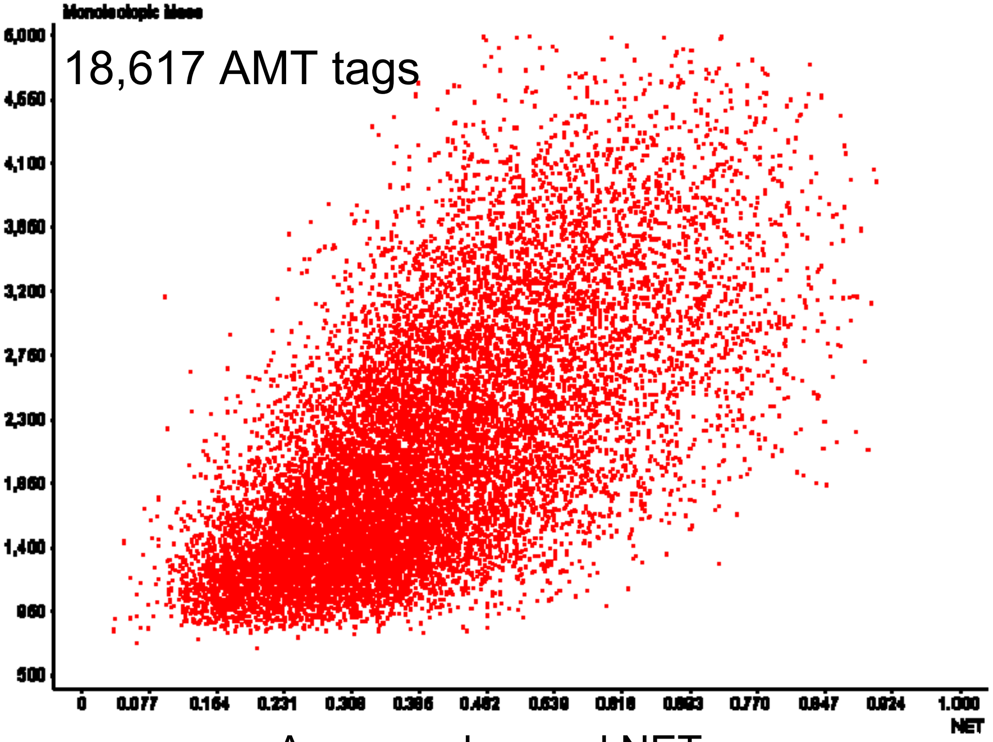




Identifying LC-MS Features

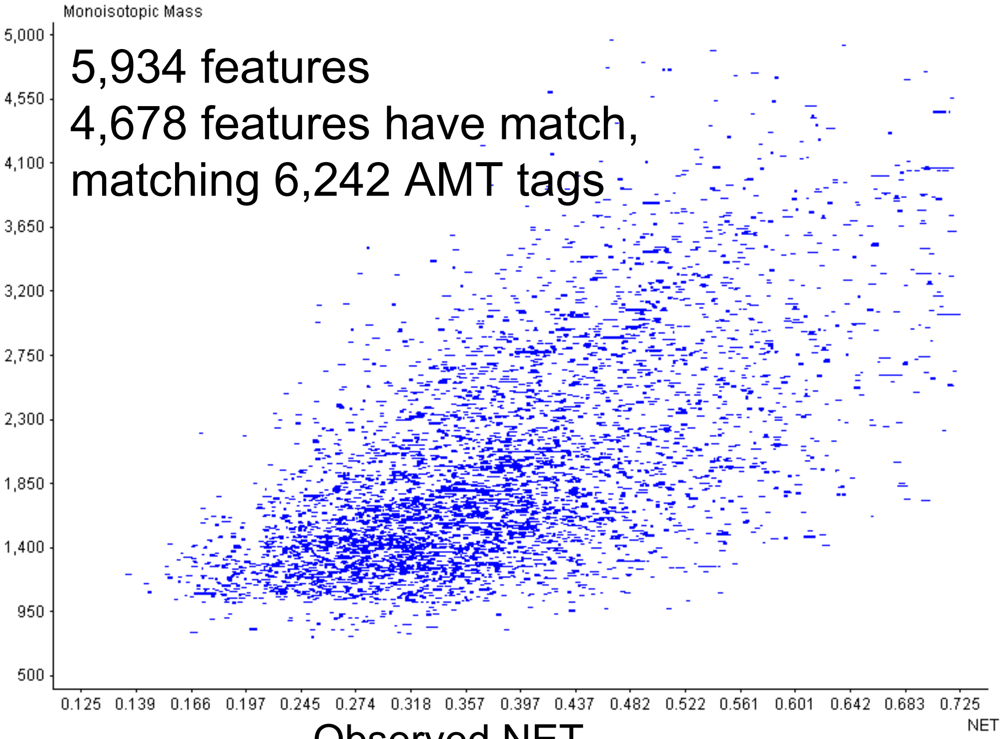
- Match Features to LC-MS/MS IDs
- *S. typhimurium* DB, from 25 LC-MS/MS analyses
  - 18,617 AMT tags, all fully or partially tryptic
  - Look for AMT tags within a broad mass range, e.g., ±25 ppm and ±0.05 NET of each feature

S. typhimurium AMT Tag Database



Average observed NET

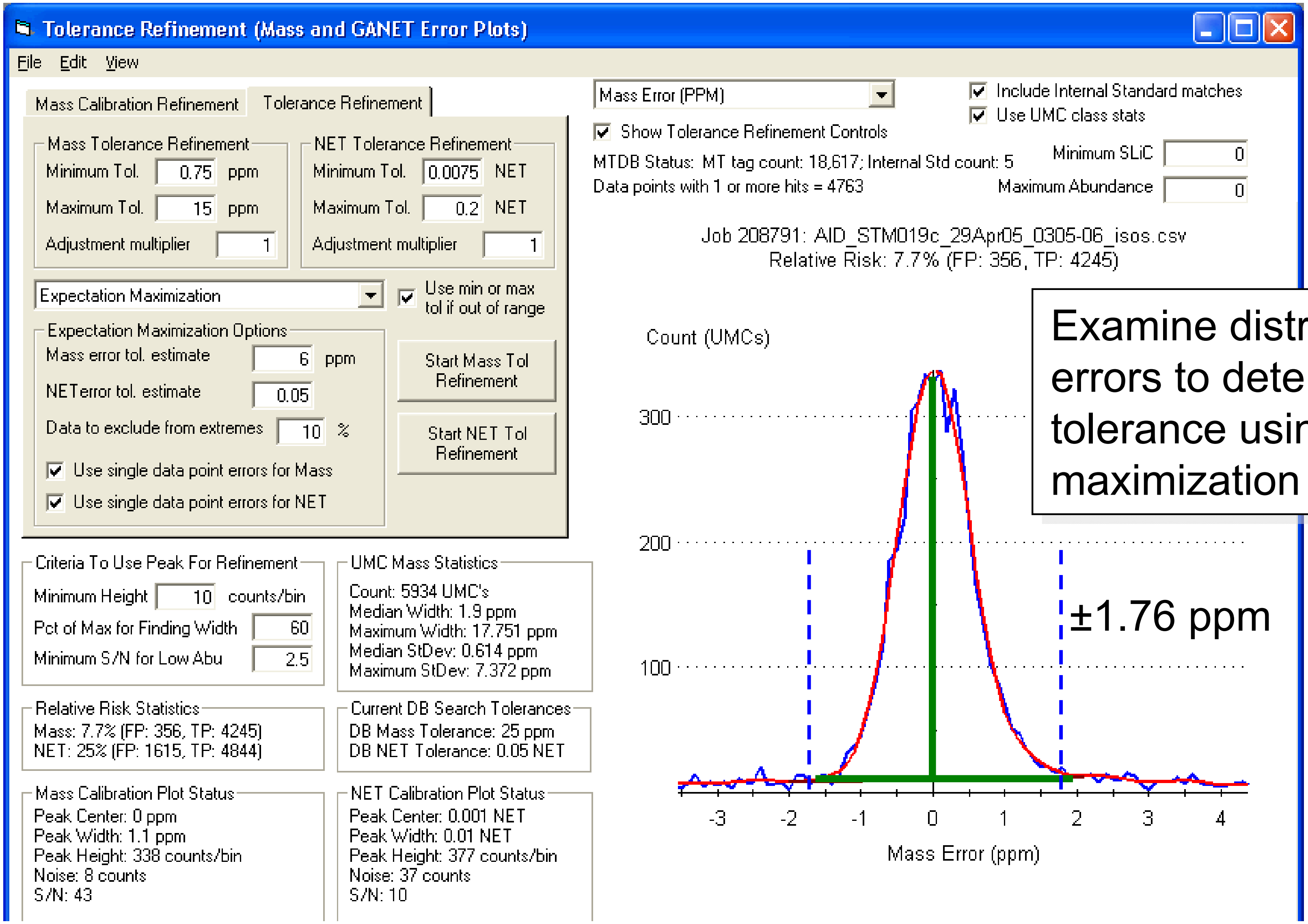
S. typhimurium on 11T FTICR



Observed NET

Search tolerance refinement

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



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

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
Data points with 1 or more hits = 4763
Minimum SLiC 0
Maximum Abundance 0

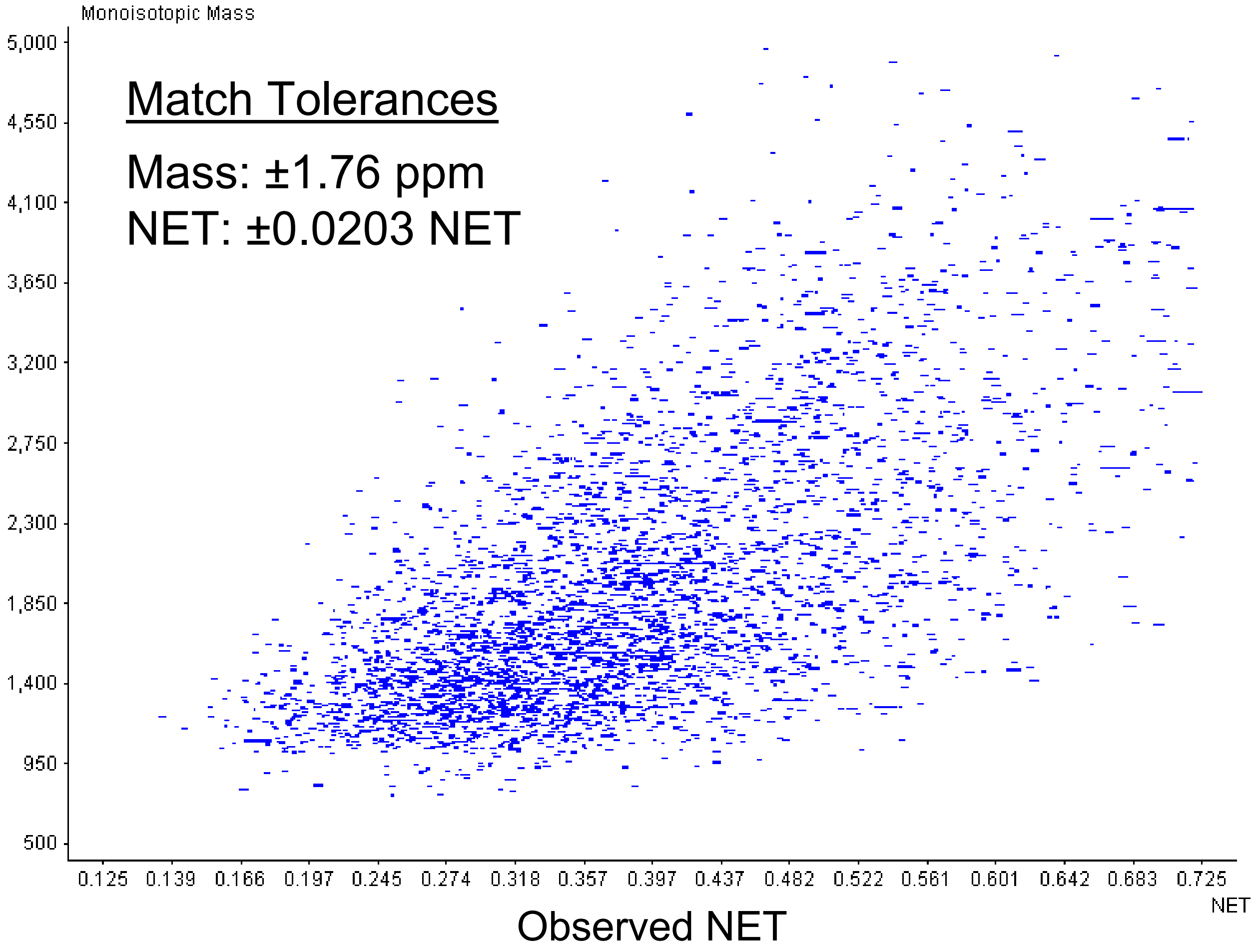
Job 208791: AID_STM019c_29Apr05_0305-06_isos.csv
Relative Risk: 7.7% (FP: 356, TP: 4245)

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

±1.76 ppm

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



Match Tolerances

Mass: ±1.76 ppm
NET: ±0.0203 NET

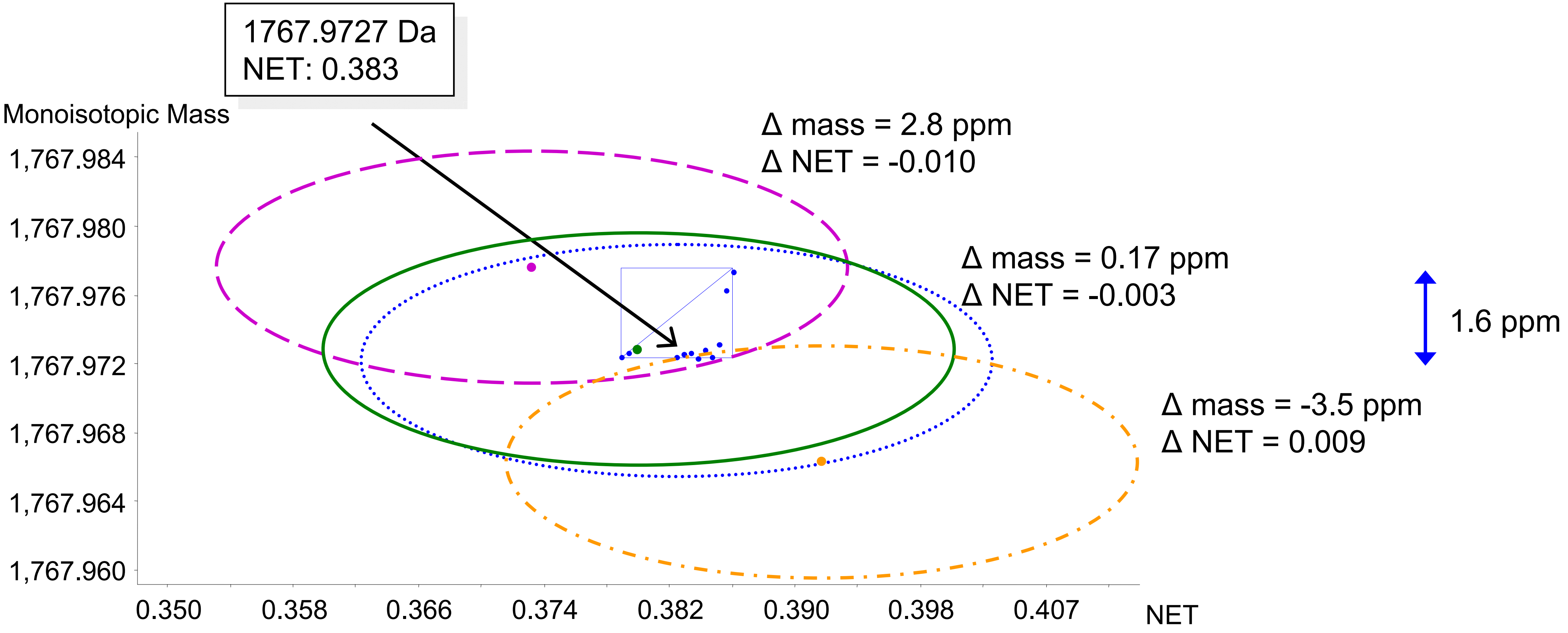
Identifying LC-MS Features

- 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



1767.9727 Da
NET: 0.383

Δ mass = 2.8 ppm
Δ NET = -0.010

Δ mass = 0.17 ppm
Δ NET = -0.003

Δ mass = -3.5 ppm
Δ NET = 0.009

1.6 ppm

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

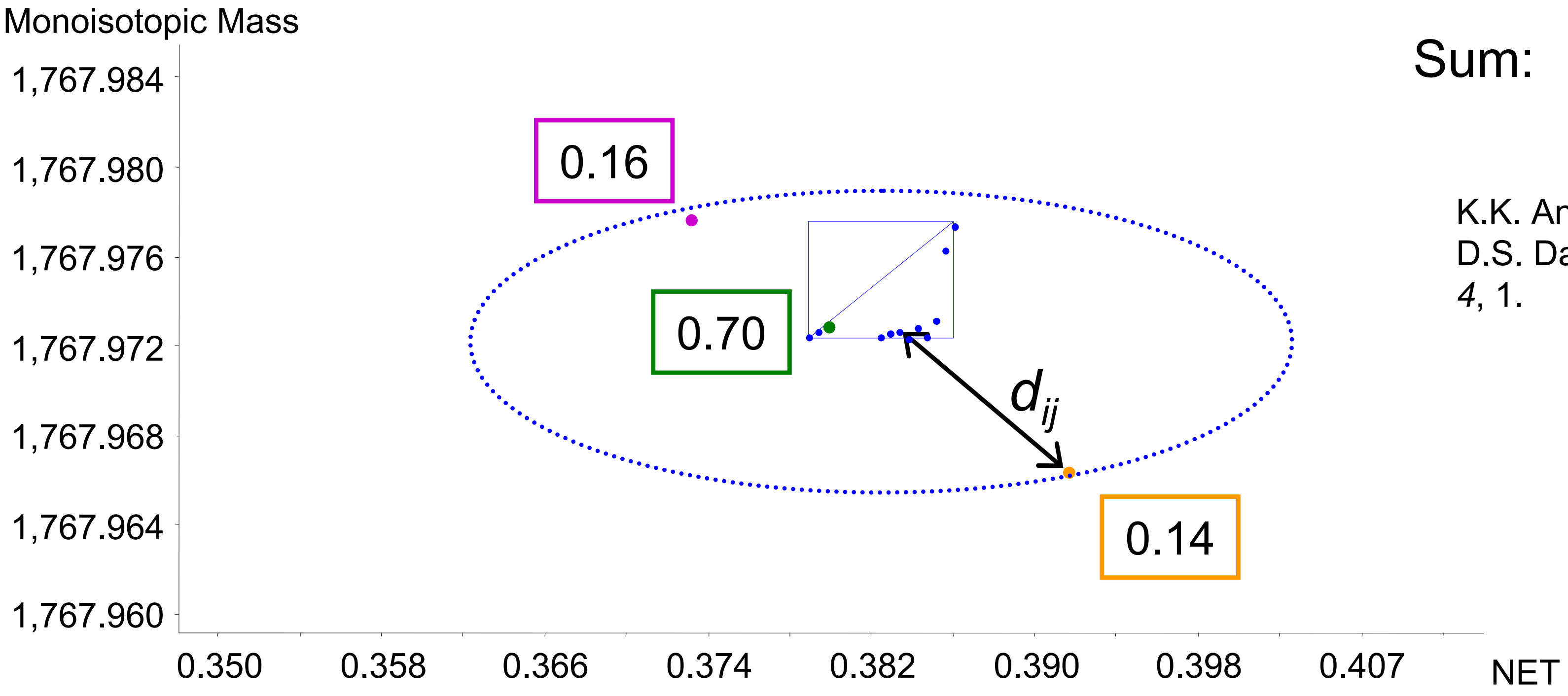
$$p_{ij} = \frac{(\sigma_{mj}\sigma_{tj})^{-1}\exp(-d_{ij}^2/2)}{\left(\sum_{k=1}^{N}(\sigma_{mk}\sigma_{tk})^{-1}\exp(-d_{ik}^2/2)\right)}$$

<u>Match Tolerances</u>

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	

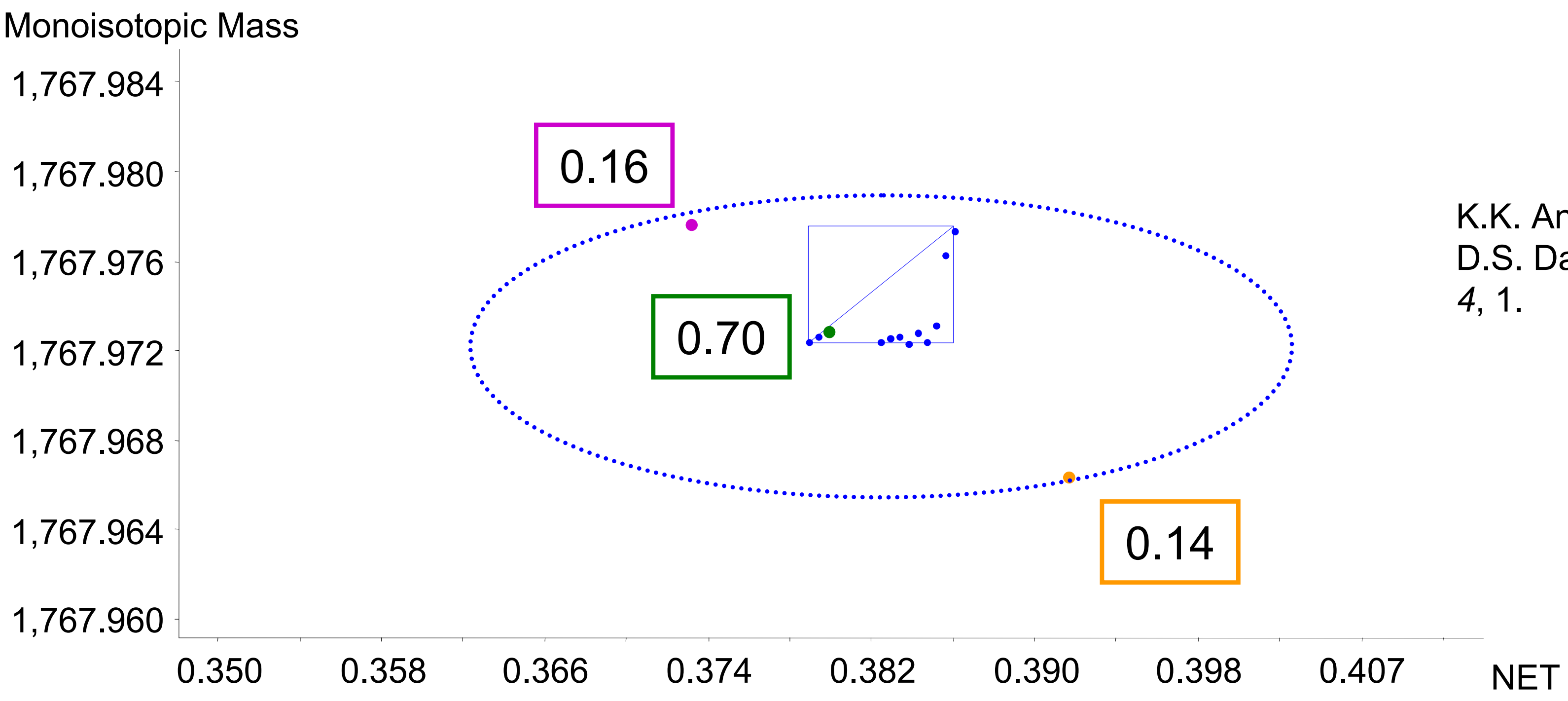


K.K. Anderson, M.E. Monroe, and D.S. Daly. *Proteome Science* **2006**, *4*, 1.

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

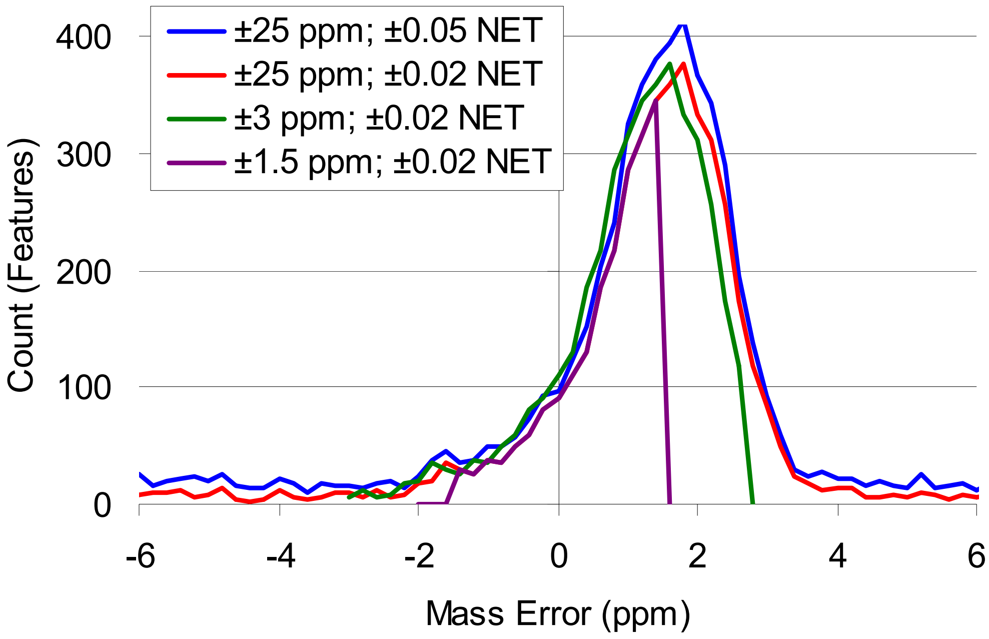


K.K. Anderson, M.E. Monroe, and D.S. Daly. *Proteome Science* **2006**, *4*, 1.

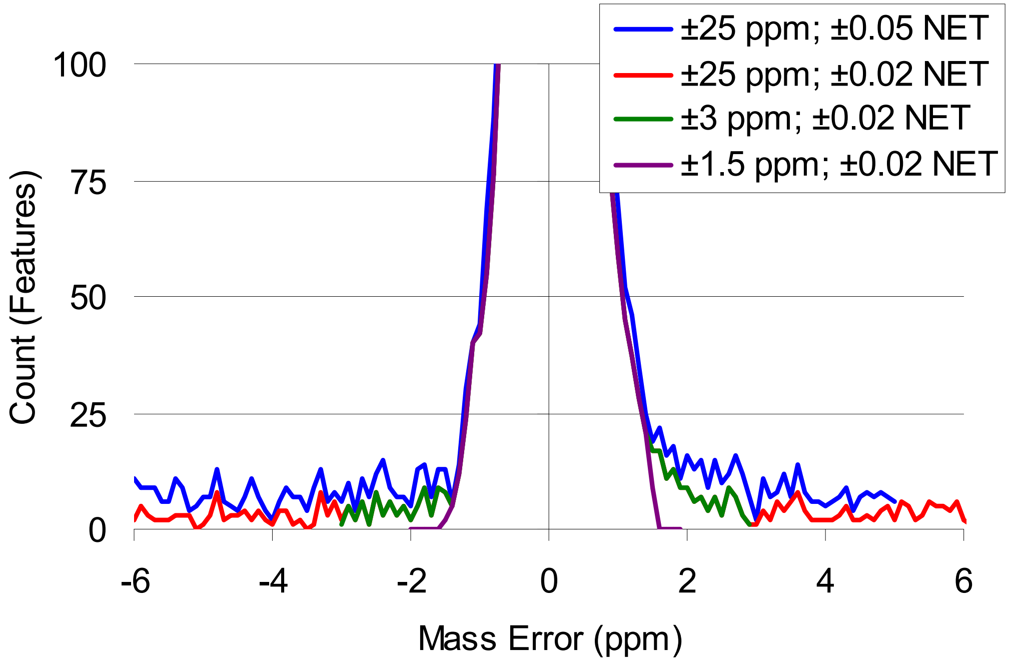
Search tolerance refinement

- 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

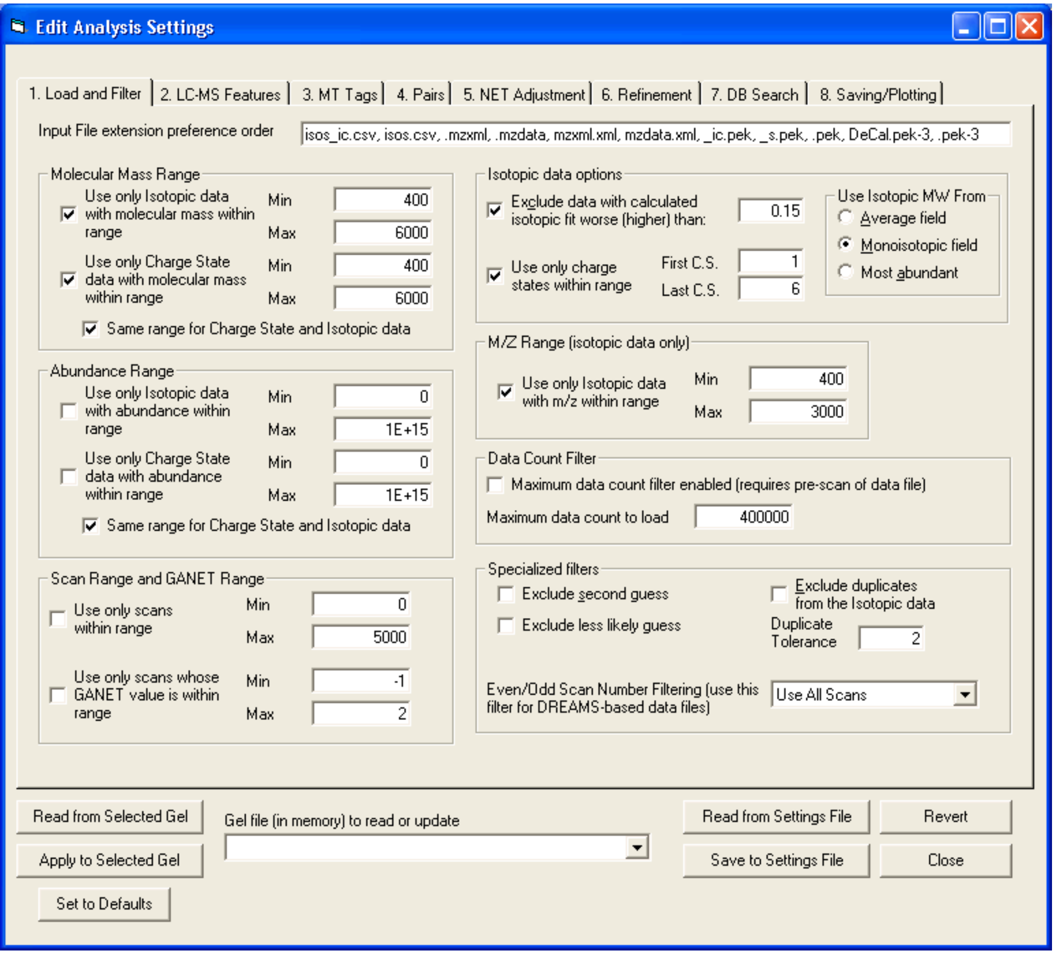


Mass error histograms with LCMSWarp mass correction



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

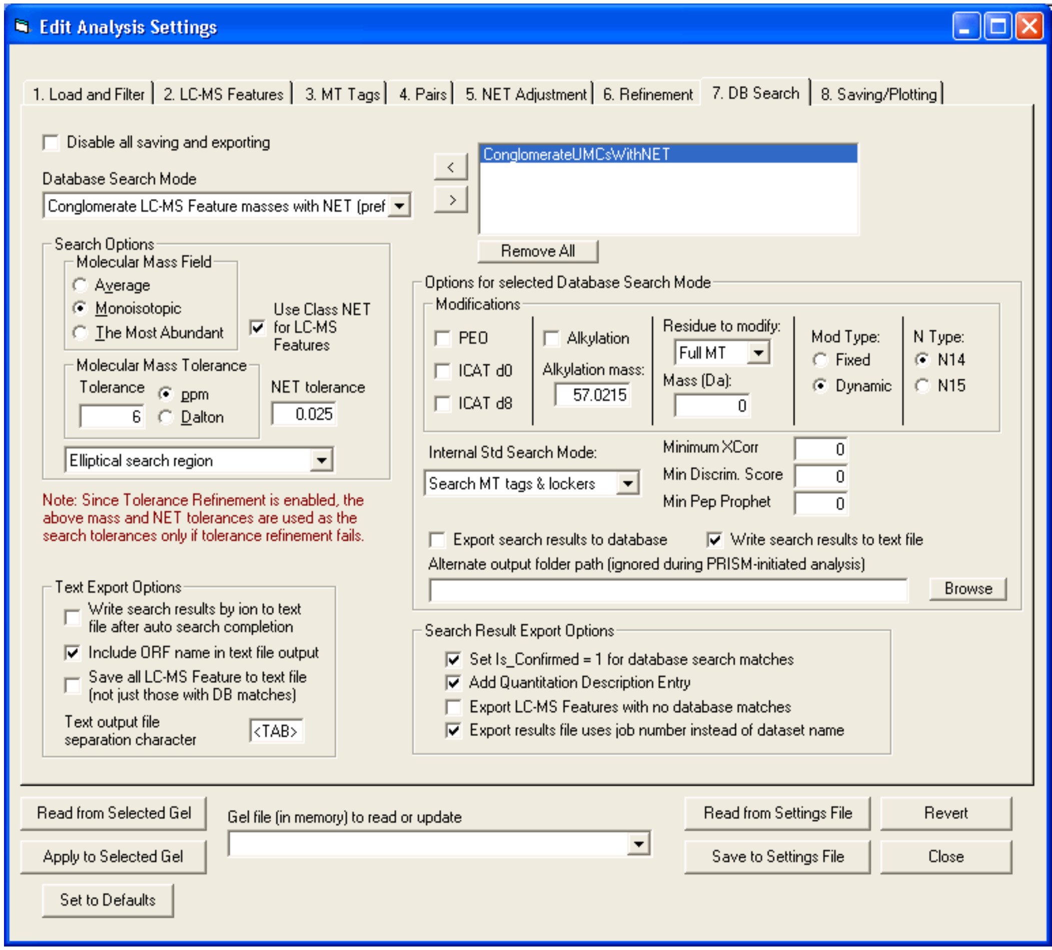
Isotopic data options
- ☑ Exclude data with calculated isotopic fit worse (higher) than: 0.15
- ☑ Use only charge states within range — First C.S. 1, Last C.S. 6
- Use Isotopic MW From: ○ Average field ● Monoisotopic field ○ Most abundant

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.

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

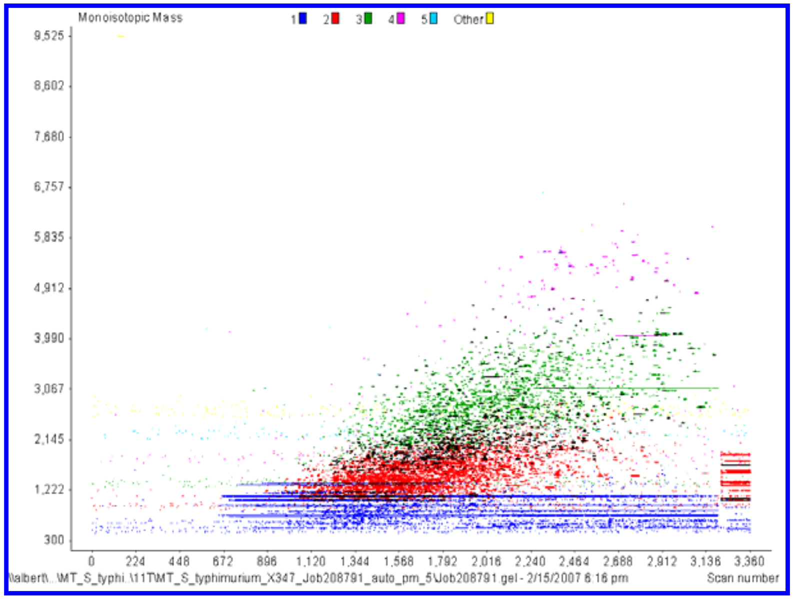
Text Export Options
- □ Write search results by ion to text file after auto search completion
- ☑ Include ORF name in text file output
- □ Save all LC-MS Feature to text file (not just those with DB matches)
- Text output file separation character: <TAB>

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

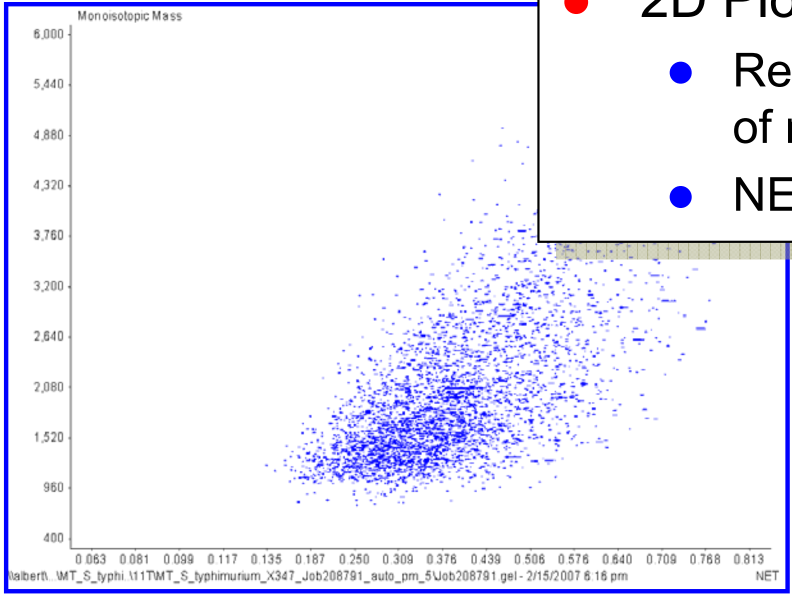
Read from Selected Gel | Apply to Selected Gel | Set to Defaults | Gel file (in memory) to read or update | Read from Settings File | Save to Settings File | Revert | Close

Peak Matching Results

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

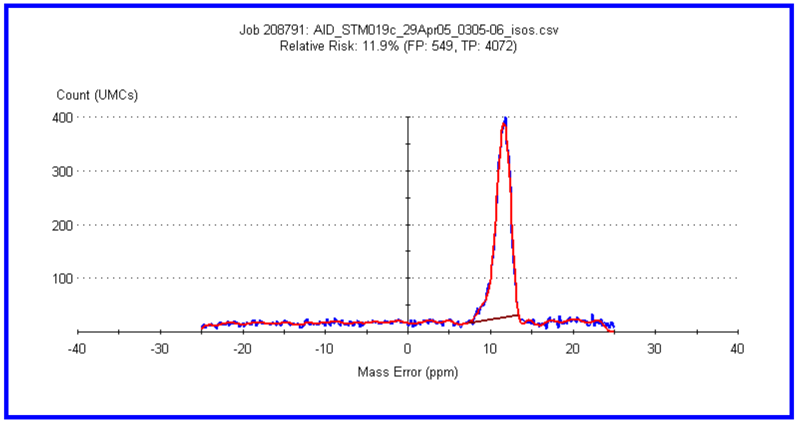


Data Searched

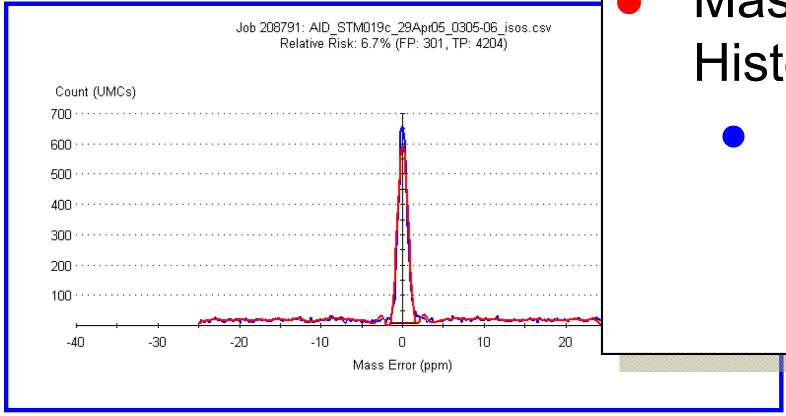


Data With Matches

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



Mass Errors Before Refinement

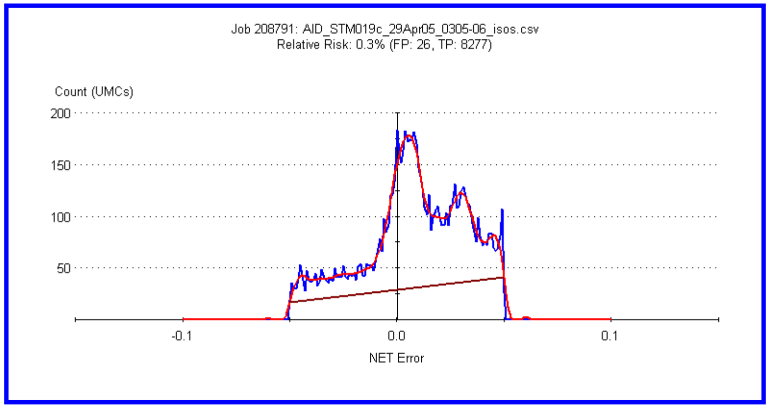


Mass Errors After Refinement

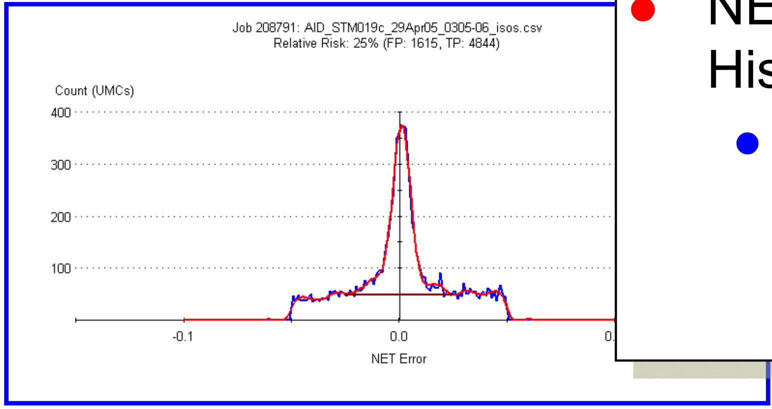
- Mass Error Histogram Metrics
  - Well defined, symmetric mass error peak centered at 0 ppm

Peak Matching Results

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

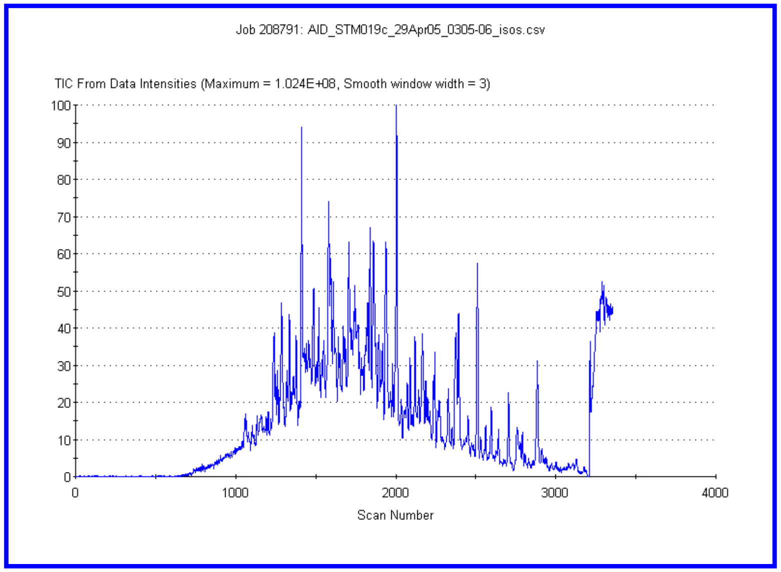


NET Errors Before Refinement

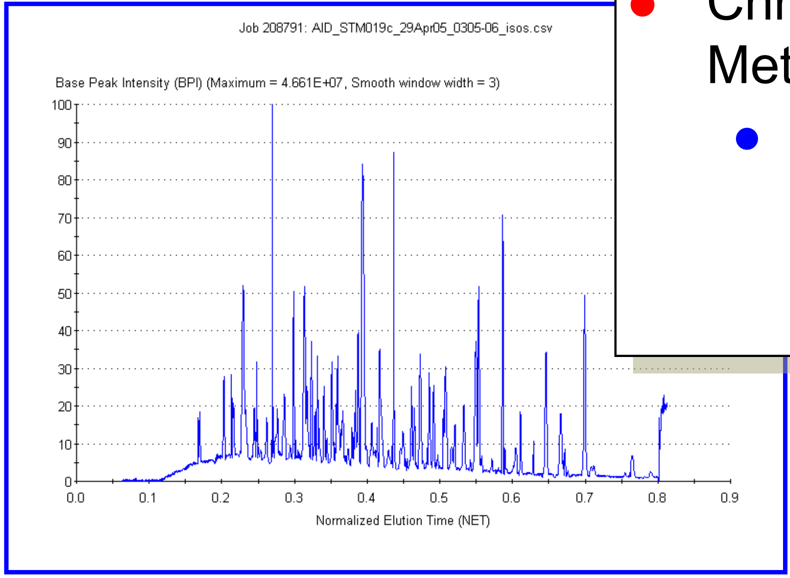


NET Errors After Refinement

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



Total Ion Chromatogram (TIC)

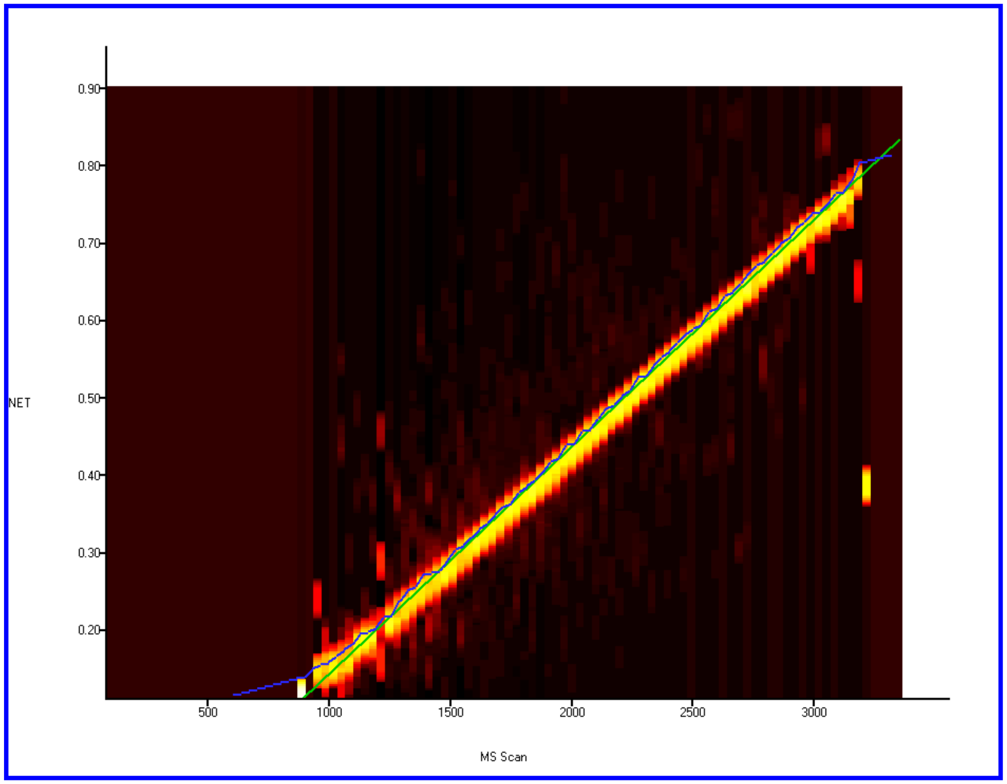


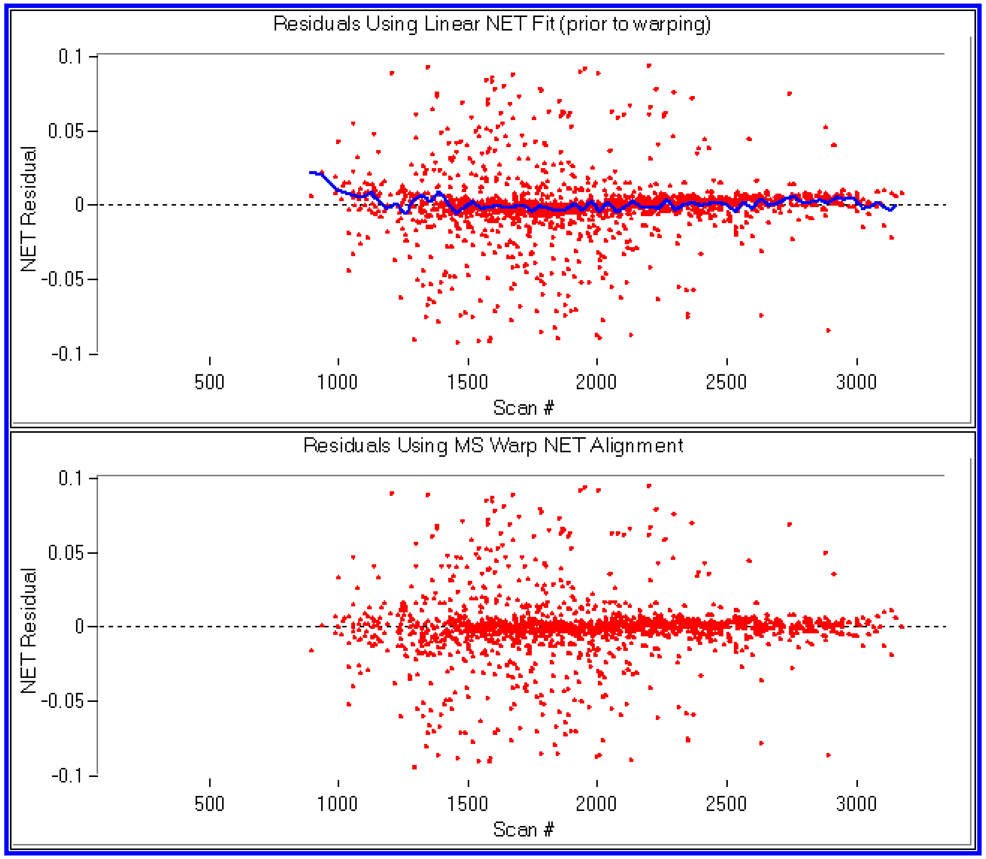
Base Peak Intensity (BPI) Chromatogram

- 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

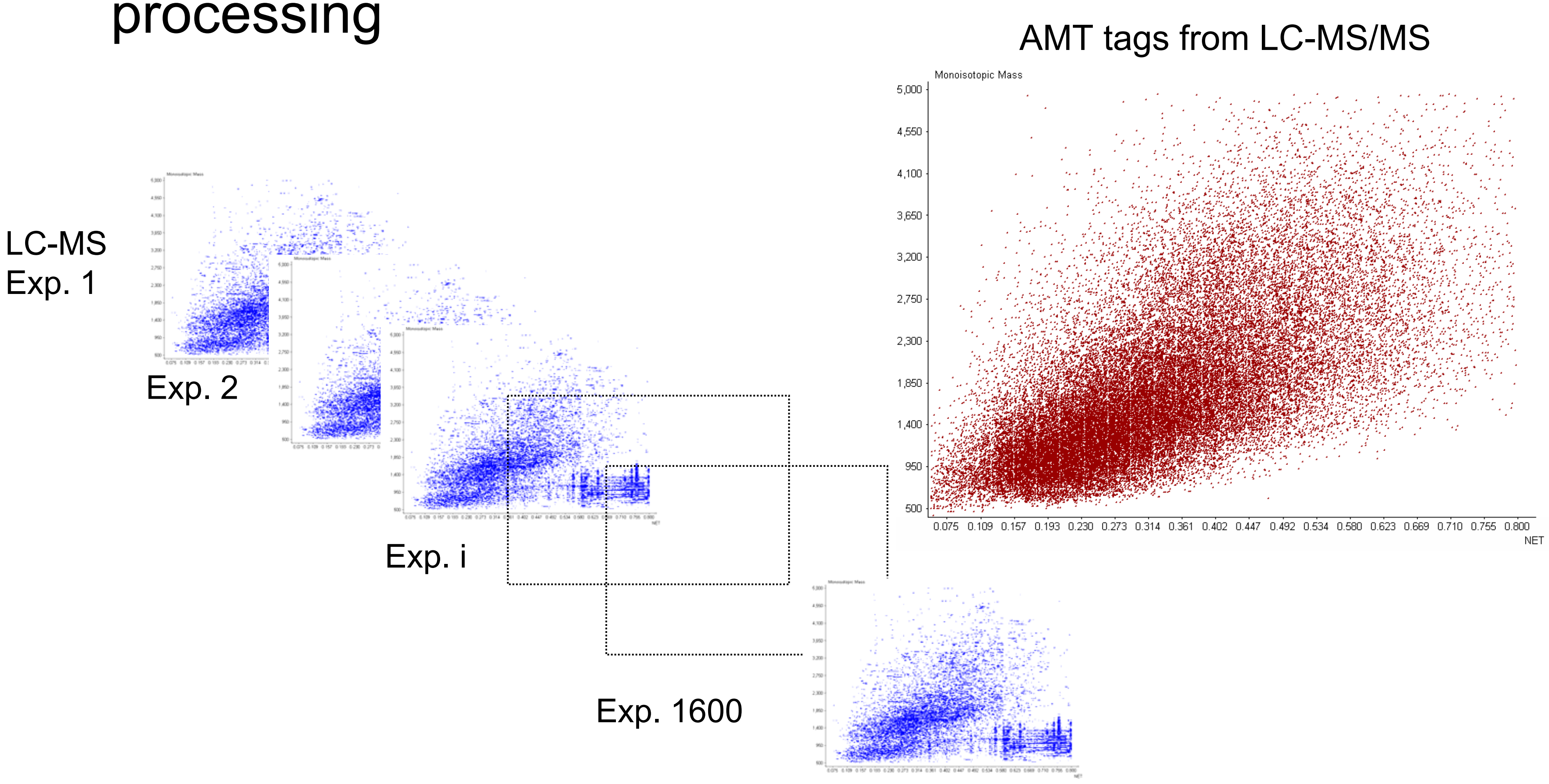
- NET Alignment Residual Metrics
  - Data after recalibration should be narrowly distributed around zero

Part II: LC-MS Feature Discovery

- Introduction (Adkins)
- Part I: Overview of Label-Free Quantitative Proteomics (Jaffe)
- Part II: Feature discovery in LC-MS datasets (Monroe and Jaitly)
  - ✓ 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
- Part III: PEPPeR, GenePattern and Real-world examples (Jaffe)
- Break
- AMT tag Pipeline Demo (general)
- Panel Discussion

Current AMT Tag Pipeline

- Individual LC-MS datasets are aligned to an AMT tag database independently
- Results are combined together after independent processing



AMT tags from LC-MS/MS

LC-MS
Exp. 1

Exp. 2

Exp. i

Exp. 1600

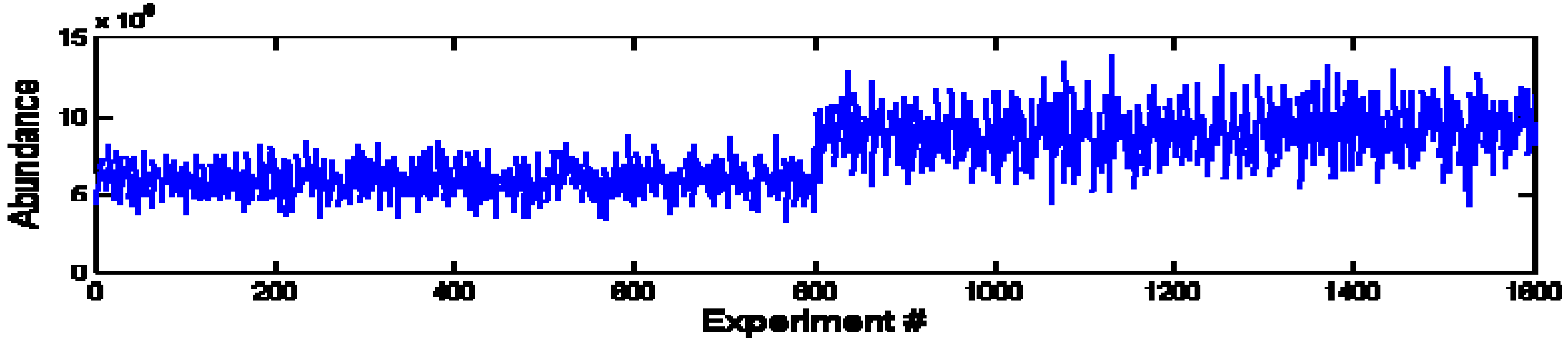
Current AMT Tag Pipeline

- For each peptide identified by peak matching, find the abundance of that peptide in all the peak matchings to create a profile

LC-MS				LC-MS/MS			
Experiment #	**Scan #**	**Mass**	**Abundance**	**Peptide**	**NET**	**Mass**	**ORFName**
1	2027	1063.56	3320000	TPHPALTEAK	0.18	1063.57	P006|BGAL_ECOLI
2	2300	1063.56	3524300	TPHPALTEAK	0.18	1063.57	P006|BGAL_ECOLI
3	-	-	-	-	-	-	-
1600	2400	1063.56	481000	TPHPALTEAK	0.18	1063.57	P006|BGAL_ECOLI

Collate Abundances

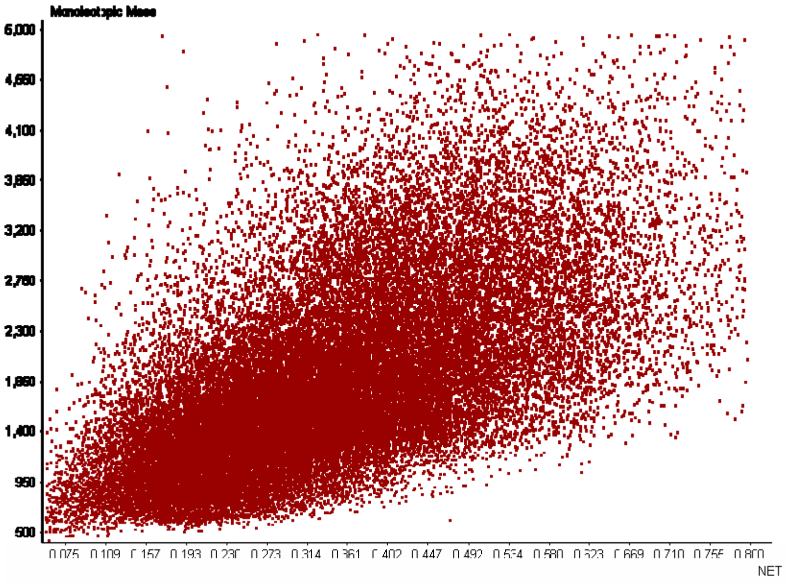
Peptide	NET	Mass	ORFName	Exp 1	Exp 2	Exp 3	Exp i	Exp 1600
TPHPALTEAK	0.18	1063.57	P006|BGAL_ECOLI	3320000	3524300	-	381000	381000



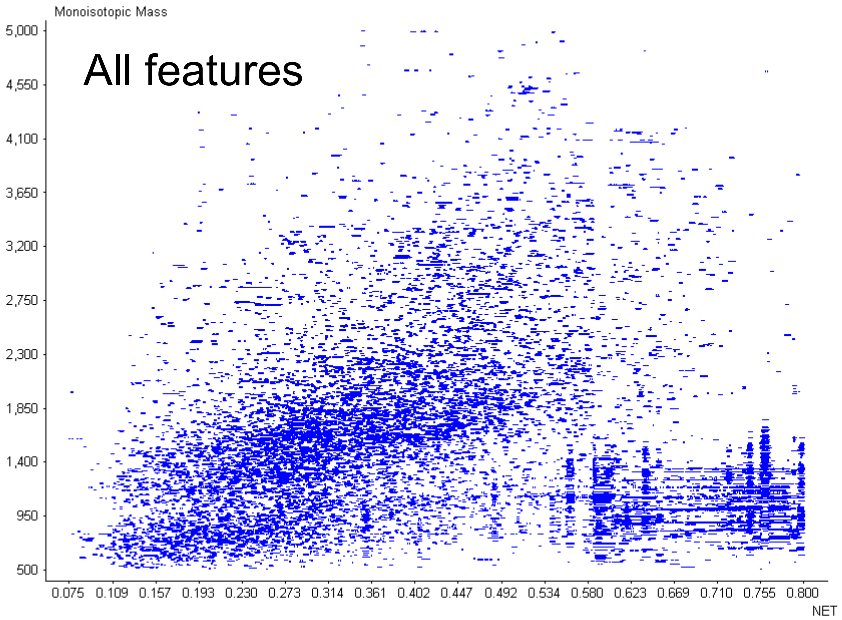
Current AMT Tag Pipeline

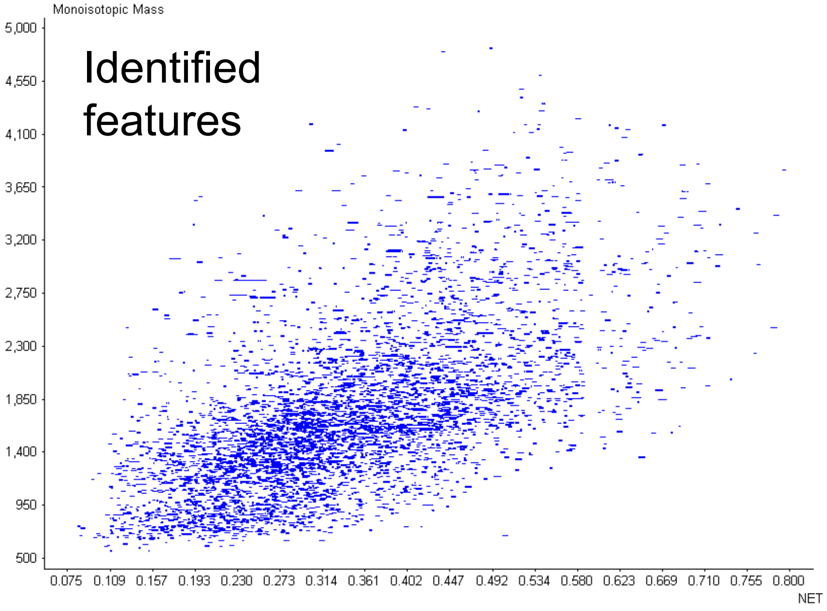
- LC-MS features without matches may represent useful information, but are effectively ignored

AMT tags from LC-MS/MS



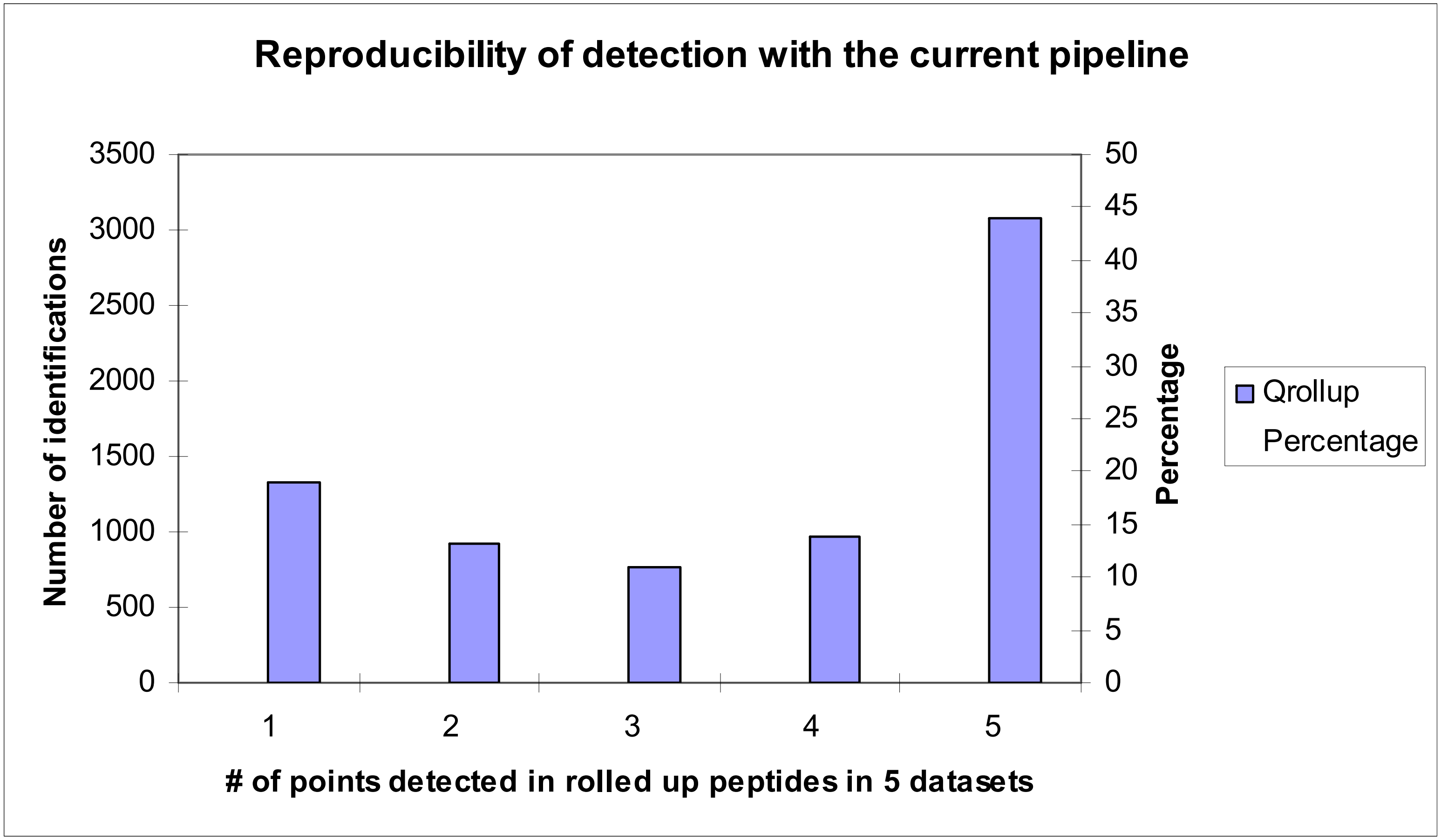
LC-FTICR-MS





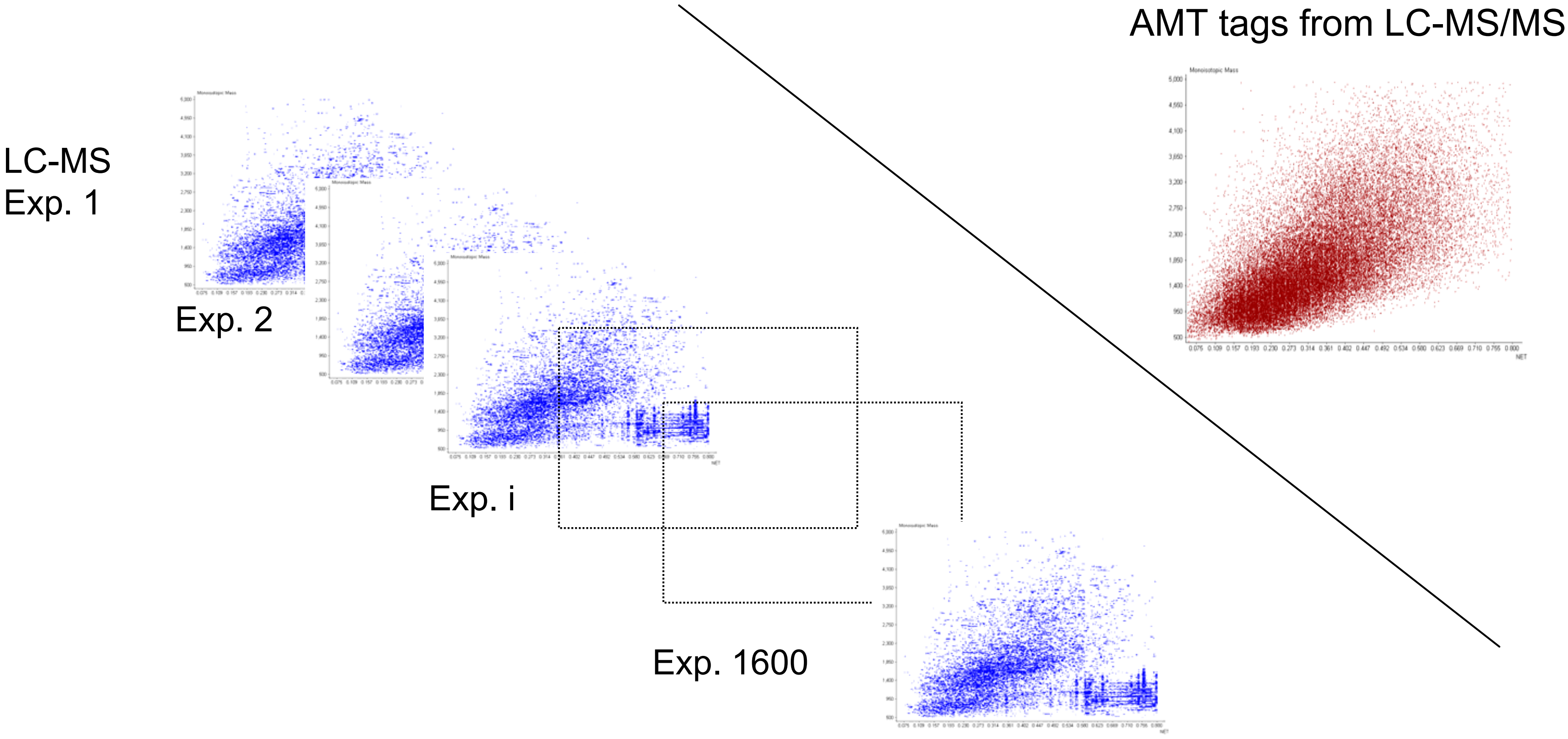
Other issues

- Independent processing of each dataset results in more missing data, because of the lack of statistics
- Lower abundance features suffer more, but are not the only casualties



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



AMT tags from LC-MS/MS

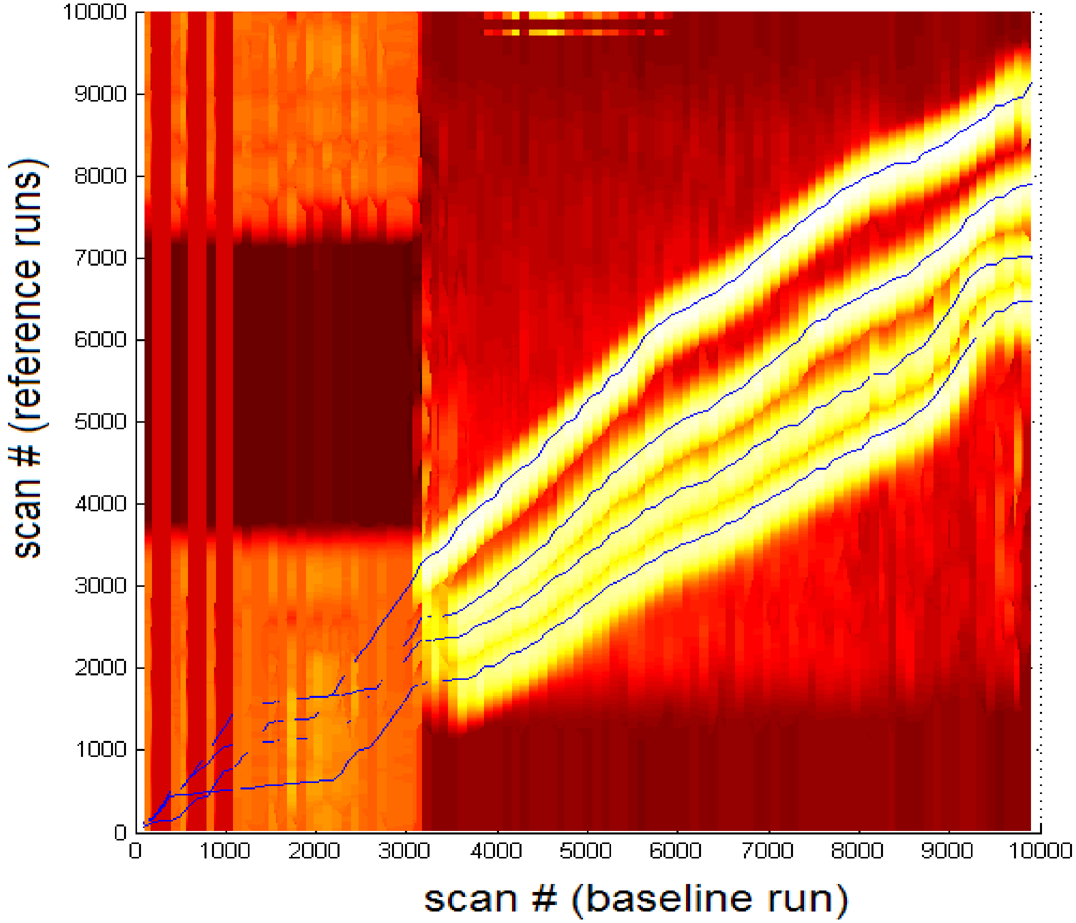
LC-MS
Exp. 1

Exp. 2

Exp. i

Exp. 1600

Align all datasets to common baseline



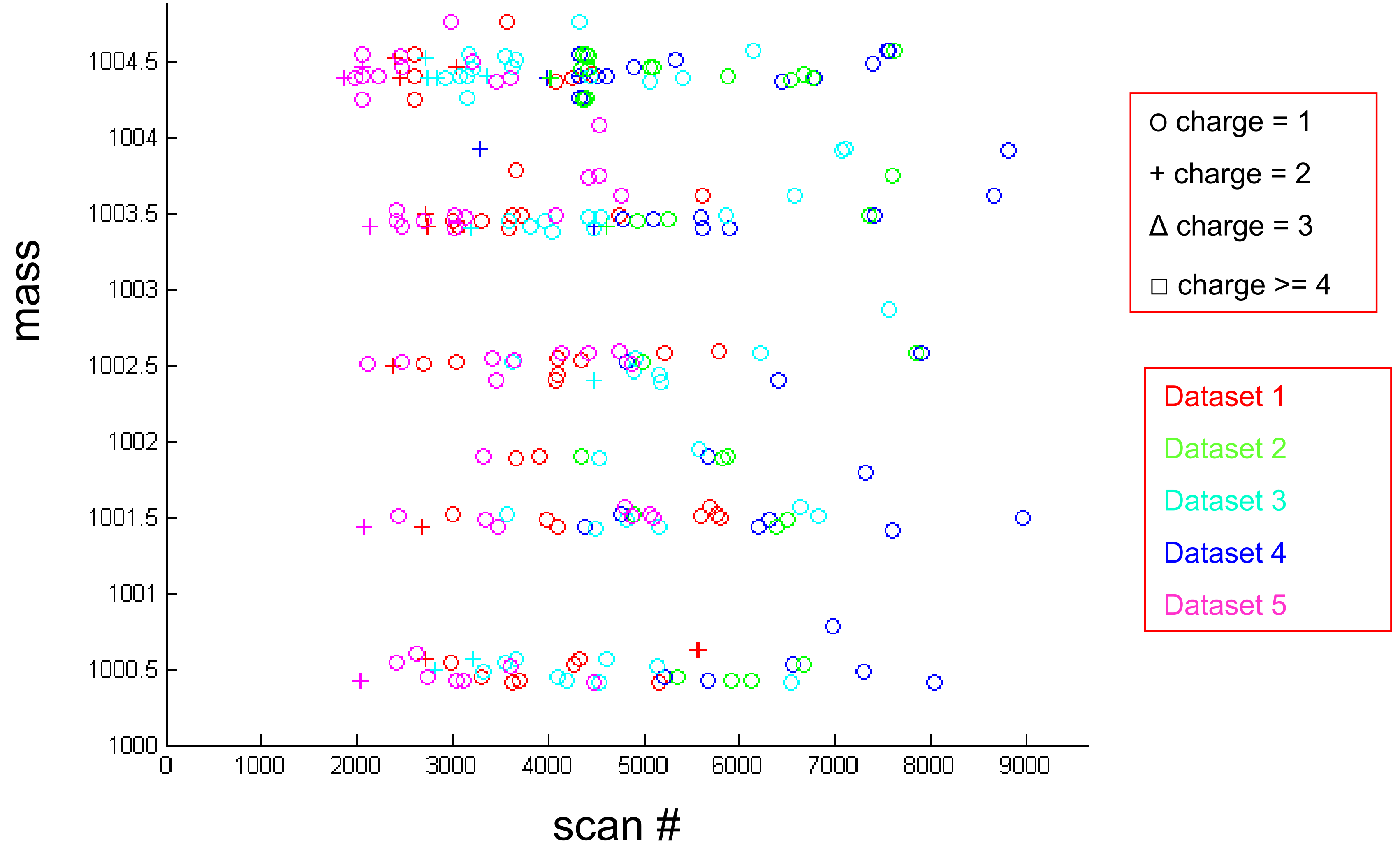
- Alignment Functions

Score plots for alignment of 4 datasets against arbitrary baseline run

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

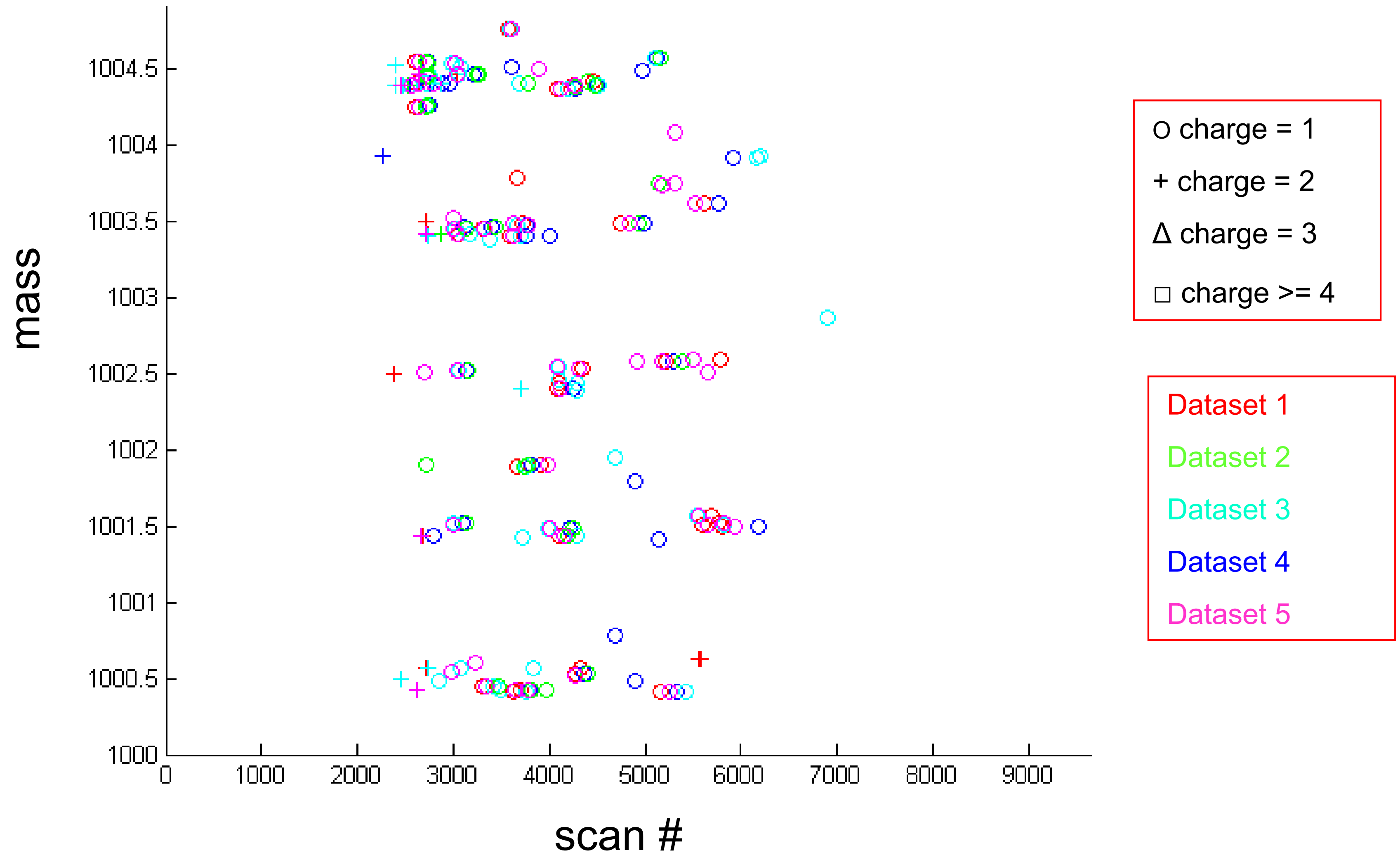
Alignment of Multiple LC-MS Datasets

- Obvious need for alignment before finding common features
  - Mass section of 5 LC-MS datasets before LC alignment



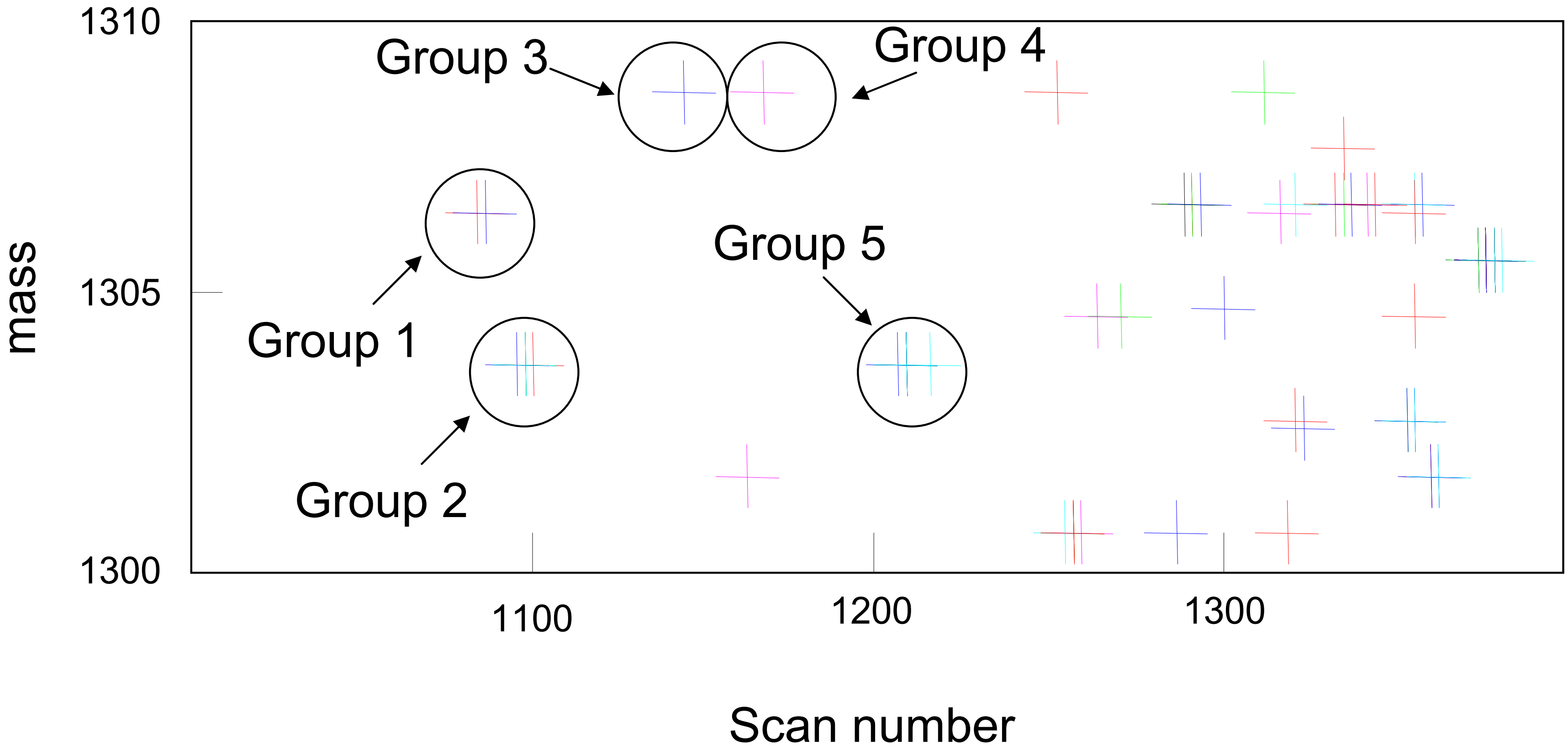
Alignment of Multiple LC-MS Datasets

- Obvious need for alignment before finding common features
  - Mass section of 5 LC-MS datasets after LC alignment



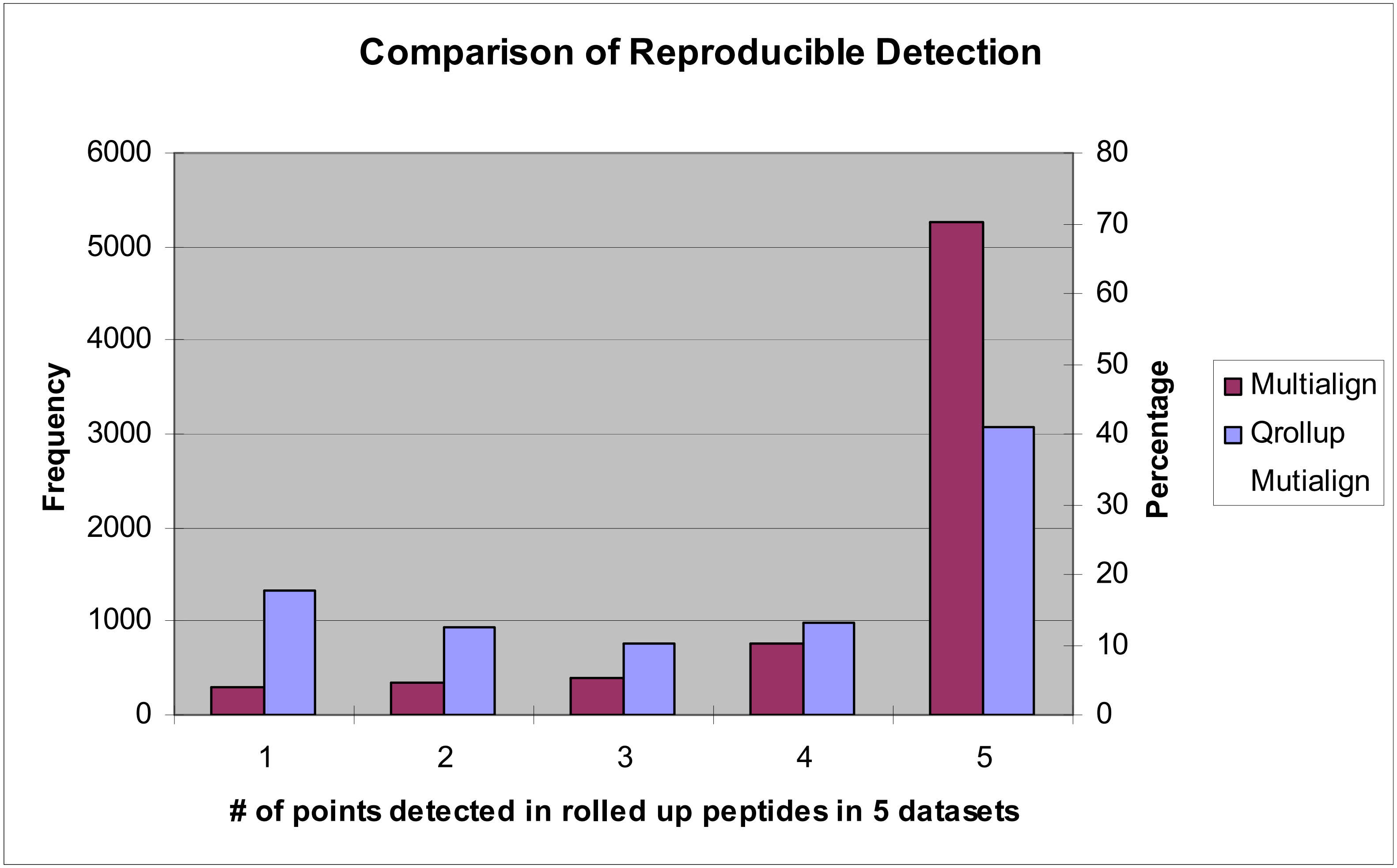
Clustering Features

- Create abundance profiles by finding similar features (using mass and retention time) across all LC-MS datasets, rather than analyzing each dataset separately and then collating results



Identifying Clustered Features

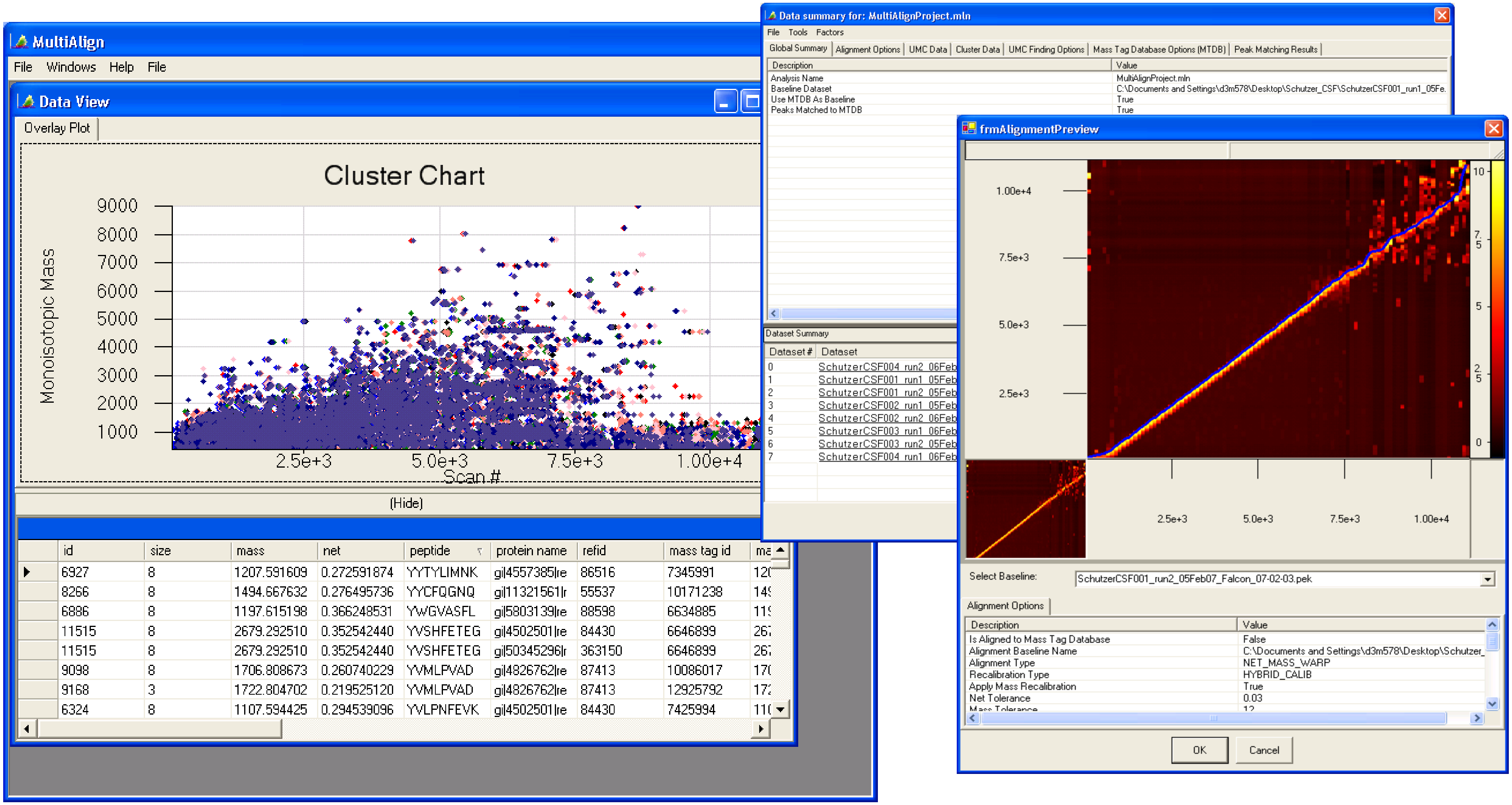
- Align mass and elution time of clusters to AMT tag database, then identify clusters by matching to AMT tags



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



id	size	mass	net	peptide	protein name	refid	mass tag id	ma
6927	8	1207.591609	0.272591874	YYTYLIMNK	gi|4557385|re	86516	7345991	120
8266	8	1494.667632	0.276495736	YYCFQGNQ	gi|11321561|r	55537	10171238	14
6886	8	1197.615198	0.366248531	YWGVASFL	gi|5803139|re	88598	6634885	11
11515	8	2679.292510	0.352542440	YVSHFETEG	gi|4502501|re	84430	6646899	26
11515	8	2679.292510	0.352542440	YVSHFETEG	gi|50345296|r	363150	6646899	26
9098	8	1706.808673	0.260740229	YVMLPVAD	gi|4826762|re	87413	10086017	170
9168	3	1722.804702	0.219525120	YVMLPVAD	gi|4826762|re	87413	12925792	17
6324	8	1107.594425	0.294539096	YVLPNFEVK	gi|4502501|re	84430	7425994	110

Description	Value
Analysis Name	MultiAlignProject.mln
Baseline Dataset	C:\Documents and Settings\d3m578\Desktop\Schutzer_CSF\SchutzerCSF001_run1_05Fe.
Use MTDB As Baseline	True
Peaks Matched to MTDB	True

Dataset #	Dataset
0	SchutzerCSF004_run2_06Feb
1	SchutzerCSF001_run1_05Feb
2	SchutzerCSF001_run2_05Feb
3	SchutzerCSF002_run1_05Feb
4	SchutzerCSF002_run2_06Feb
5	SchutzerCSF003_run1_06Feb
6	SchutzerCSF003_run2_05Feb
7	SchutzerCSF004_run1_06Feb

Select Baseline: SchutzerCSF001_run2_05Feb07_Falcon_07-02-03.pek

Description	Value
Is Aligned to Mass Tag Database	False
Alignment Baseline Name	C:\Documents and Settings\d3m578\Desktop\Schutzer_
Alignment Type	NET_MASS_WARP
Recalibration Type	HYBRID_CALIB
Apply Mass Recalibration	True
Net Tolerance	0.03
Mass Tolerance	12

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.

LC-MS Feature Discovery

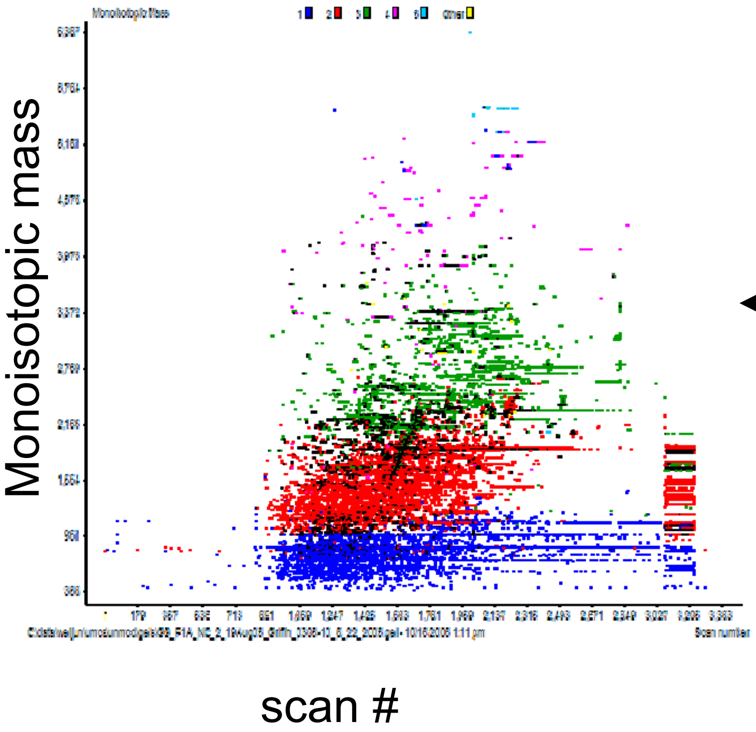
- Similar approaches and software tools: Low Res LC-MS
  - Signal maps software
    - A. Prakash et. al. *Mol. Cell Proteomics* **2006**, *5*, 423-432.
  - Informatics platform for global proteomic profiling using LC-MS
    - D. Radulovic, et al. *Mol. Cell. Proteomics* **2004**, *3*, 984-997.
  - Computational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

Part II: LC-MS Feature Discovery

- Introduction (Adkins)
- Part I: Overview of Label-Free Quantitative Proteomics (Jaffe)
- Part II: Feature discovery in LC-MS datasets (Monroe and Jaitly)
  - ✓ 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
- Part III: PEPPeR, GenePattern and Real-world examples (Jaffe)
- Break
- AMT tag Pipeline Demo (general)
- Panel Discussion

Developing Confidence Metrics

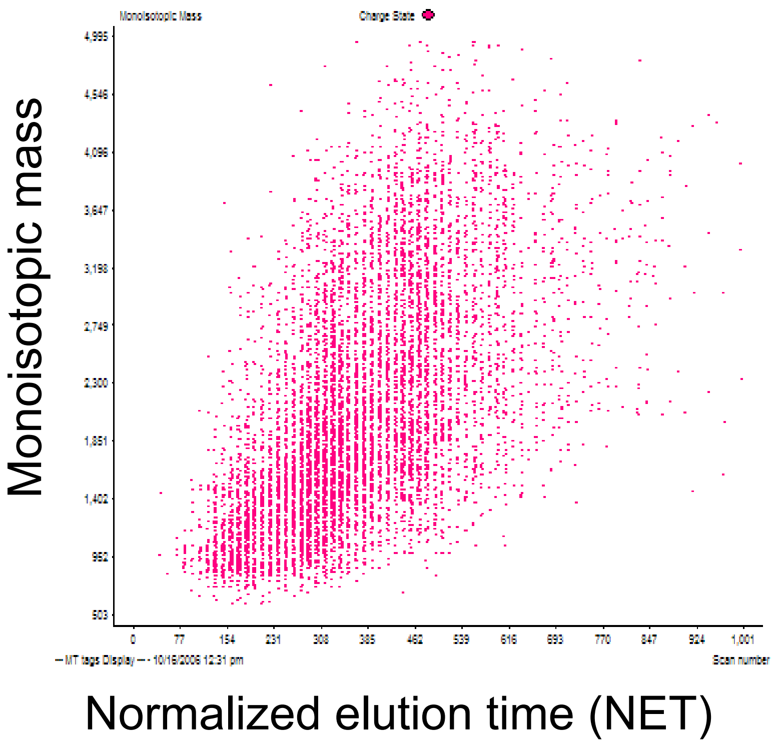
- LC-MS data is aligned against an AMT tag database
- Each LC-MS feature is matched to the closest AMT tag in mass and normalized LC elution time (NET) dimensions



LC-MS dataset

Alignment &

Peak Matching



AMT tag Database

How do we control the errors in the process?

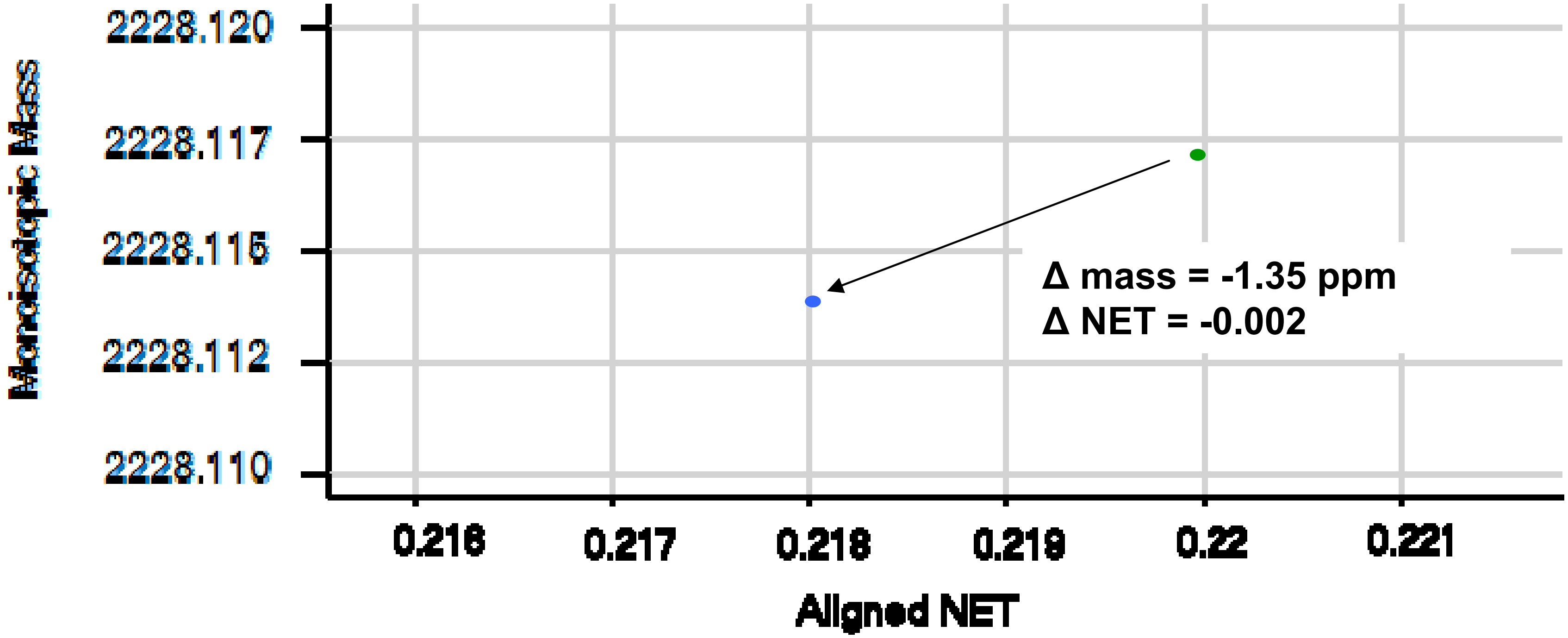
Controlling Rate of Random Matches

- Size of database and degree of noise in database affects the rate of random matches
  - Building more confident AMT tag database (e.g., using strict filtering) decreases background false positives
  - But, increases false negatives
- To date, *ad hoc* rules have been used
  - *Subjectively pleasing* threshold values selected for different parameters, such as mass error tolerance, LC NET error tolerance, etc.
  - False discovery rate (FDR) was estimated using decoy methods
  - Rules were accepted if results seemed satisfactory, otherwise parameters were re-optimized
  - But, chosen parameters may not result in optimal results

Metrics Associated with a Candidate Identification

- Each match between an LC-MS feature and a peptide AMT tag is described by a mass error and an LC NET error

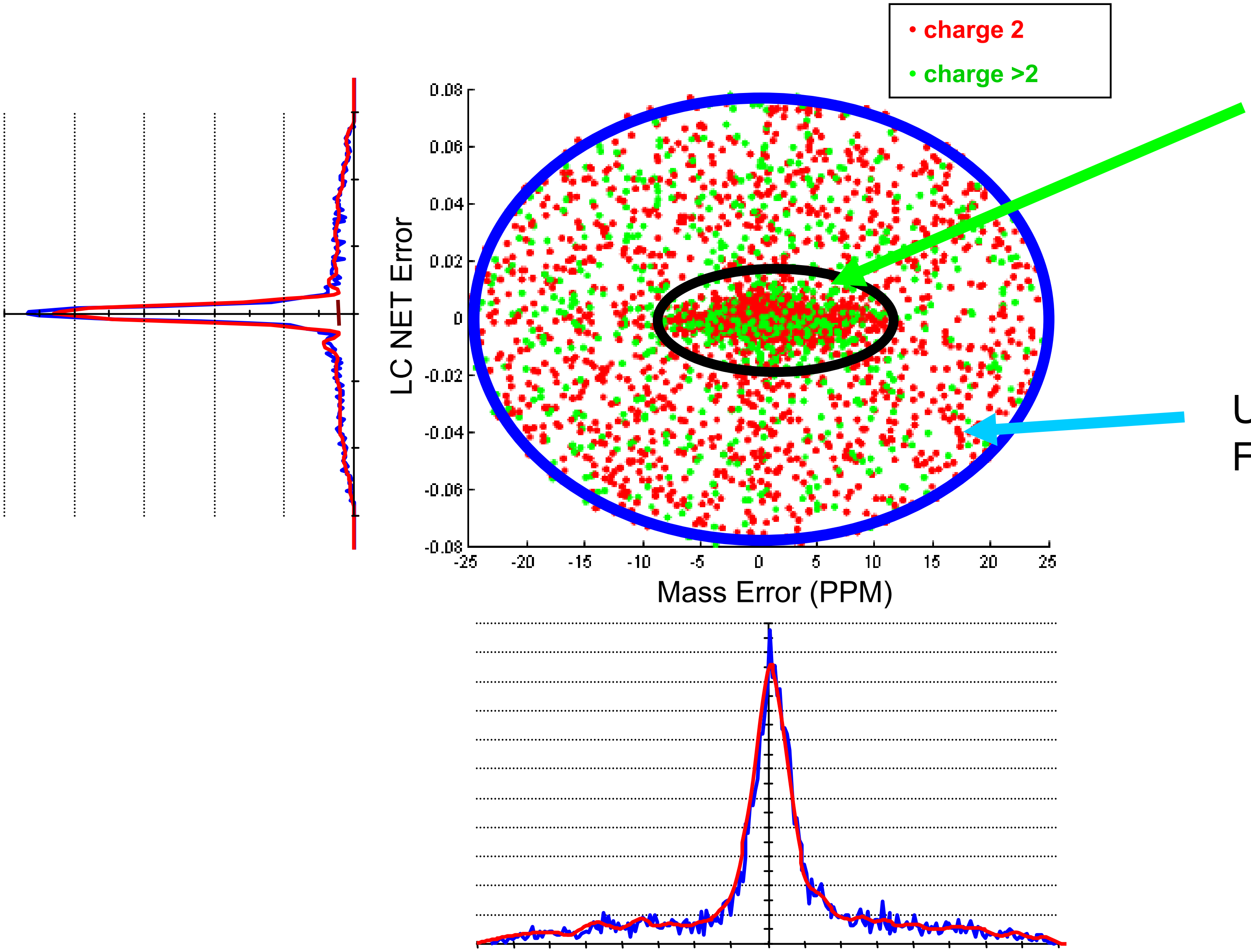
Mass	Scan	Aligned NET	Peptide	NET	Mass	ORFName
2228.114	1097	0.218	TETQEKNPLPSKETIEQEK	0.22	2228.117	Thymosin beta-4



Δ mass = -1.35 ppm
Δ NET = -0.002

Distribution of Peak Matches

- True and false matches resulting from peak matching display different mass and LC NET error distributions

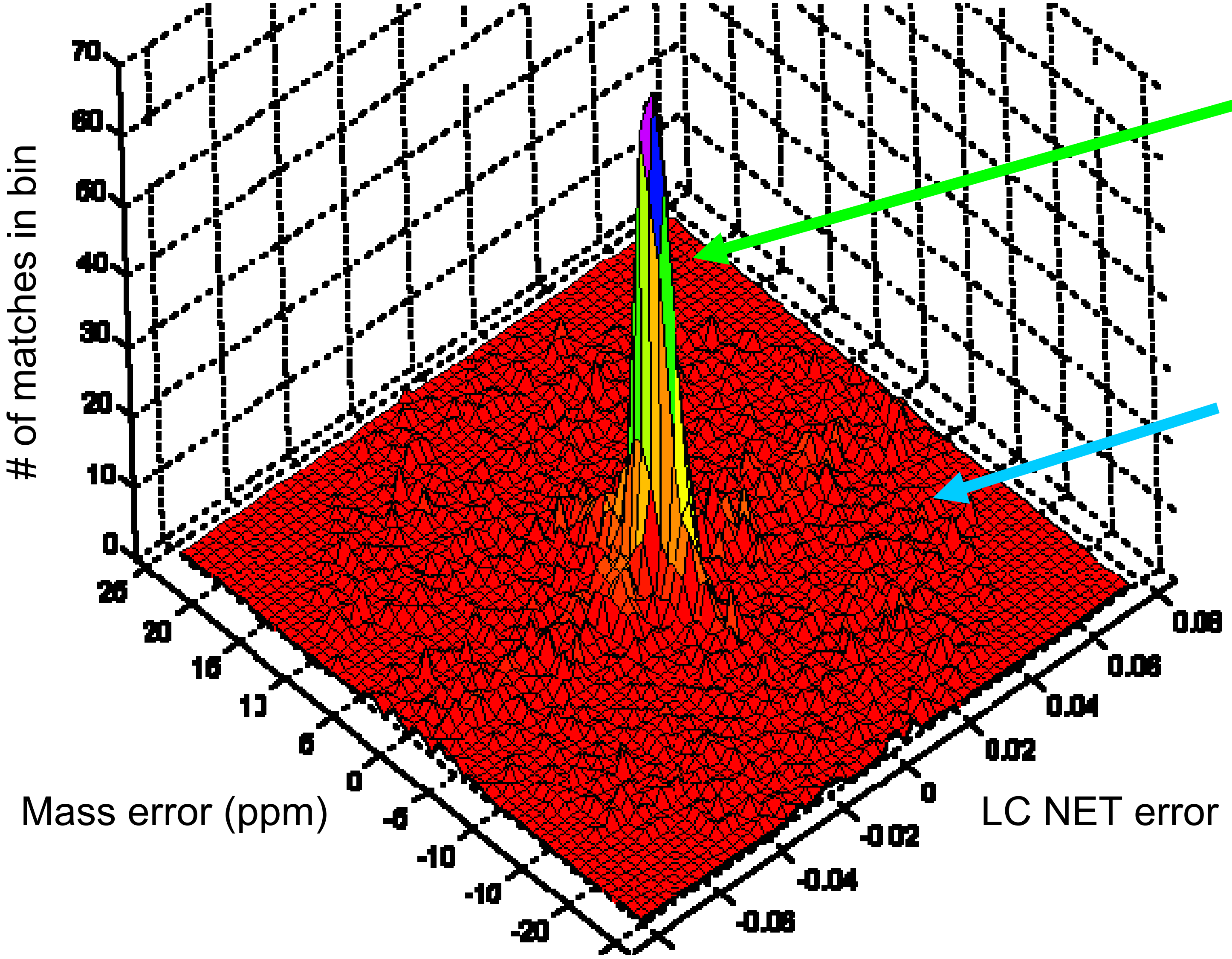


Centrally Distributed True Matches

Uniformly Distributed False Matches

Distribution of LC-MS Peak Matches

- Density plot of mass and LC NET error distributions is a sum of true and false components

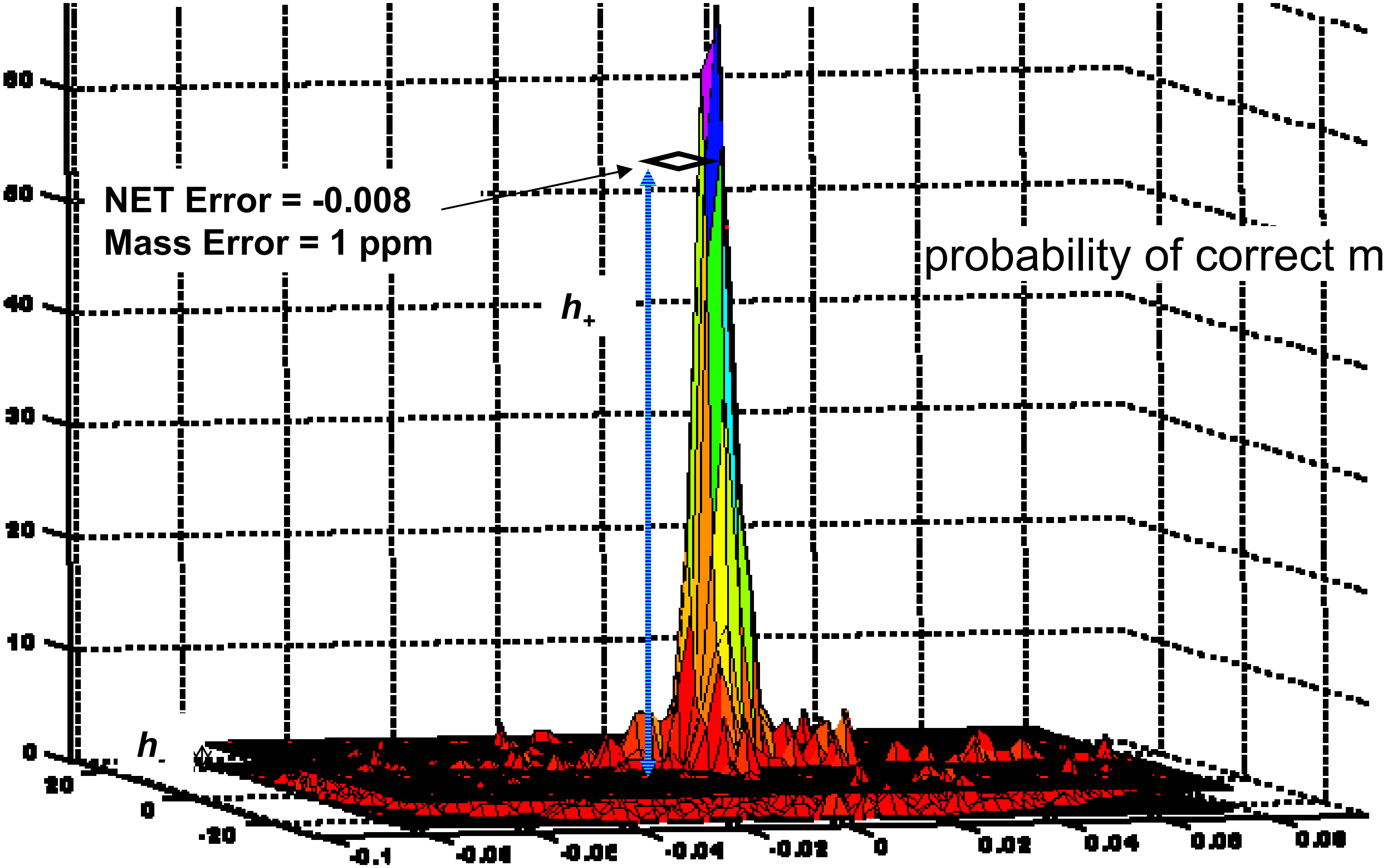


Centrally Distributed
True Matches

Uniformly Distributed
False Matches

Estimating the Probability a Match is Correct

- Approach: the probability that a peak match is correct can be estimated from where its mass and LC NET error values lie on the two-dimensional distribution



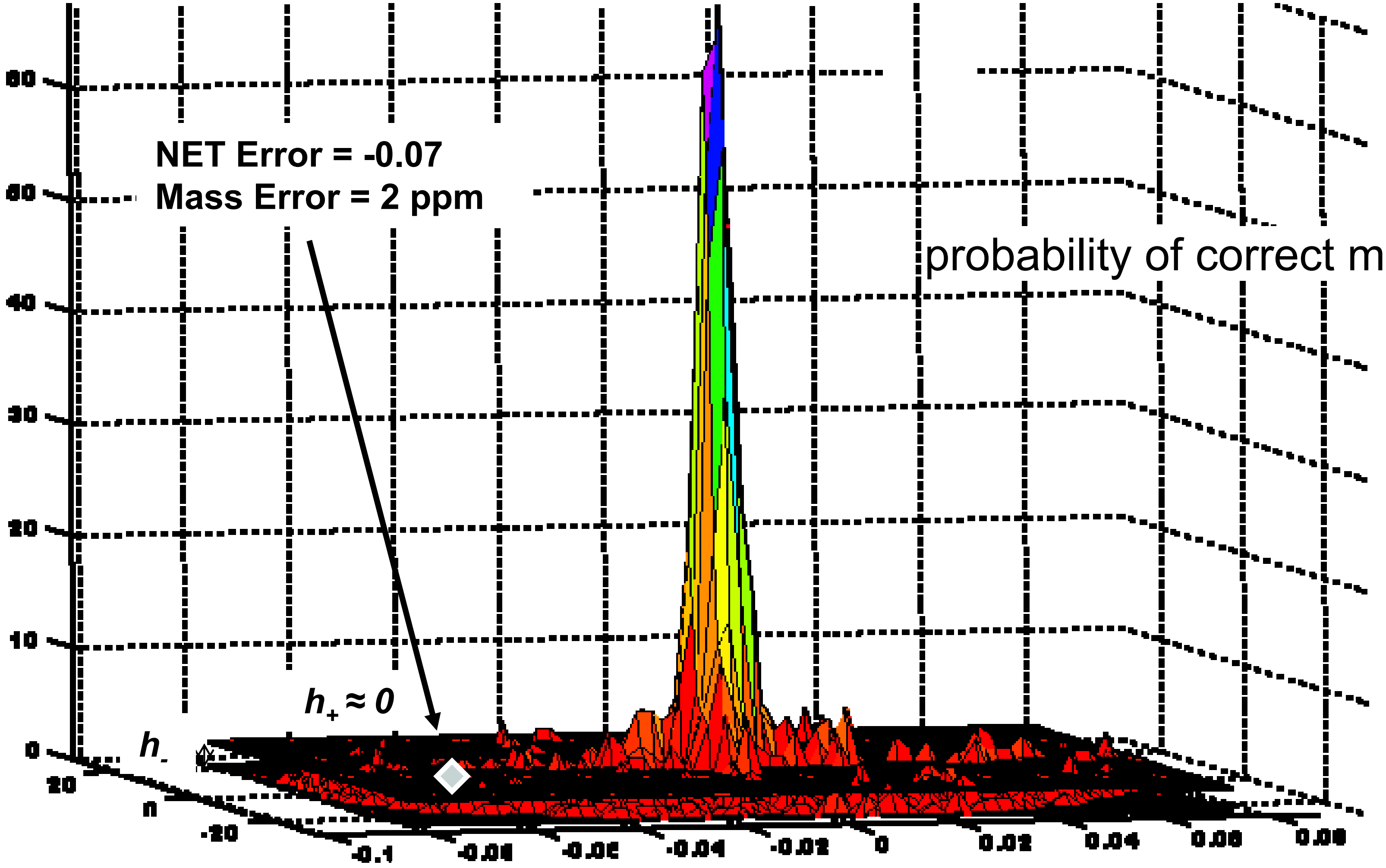
NET Error = -0.008
Mass Error = 1 ppm

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

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

Estimating the Probability a Match is Correct

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



NET Error = -0.07
Mass Error = 2 ppm

$h_+ \approx 0$

h_-

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

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

Optimizing the overall matching process

- General approach; calculate confidence in peak match based on:
  - Mass and LC NET errors
  - Instrumental performance for an analysis
    - Mass error precision
    - LC-NET precision
  - LC-MS/MS ID quality (e.g., SEQUEST XCorr or X!Tandem expectation values)
- Inter-related effects of different parameters on each other complicate simple choices:
  - Lower mass and LC NET errors should allow choice of lower scores
  - Higher scores should allow somewhat wider mass and LC NET tolerances

→ **For practical value we need a single metric that calculates and combines all these factors automatically**

- *S*tatistical *M*ethod for *A*ssignment of *R*elative *T*ruth (*SMART*) – More details to be presented at ASMS 2008 Bioinformatics oral session

Part II: LC-MS Feature Discovery

- Introduction (Adkins)
- Part I: Overview of Label-Free Quantitative Proteomics (Jaffe)
- Part II: Feature discovery in LC-MS datasets (Monroe and Jaitly)
  - ✓ 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
- Part III: PEPPeR, GenePattern and Real-world examples (Jaffe)
- Break
- AMT tag Pipeline Demo (general)
- Panel Discussion

Downstream Data Analysis

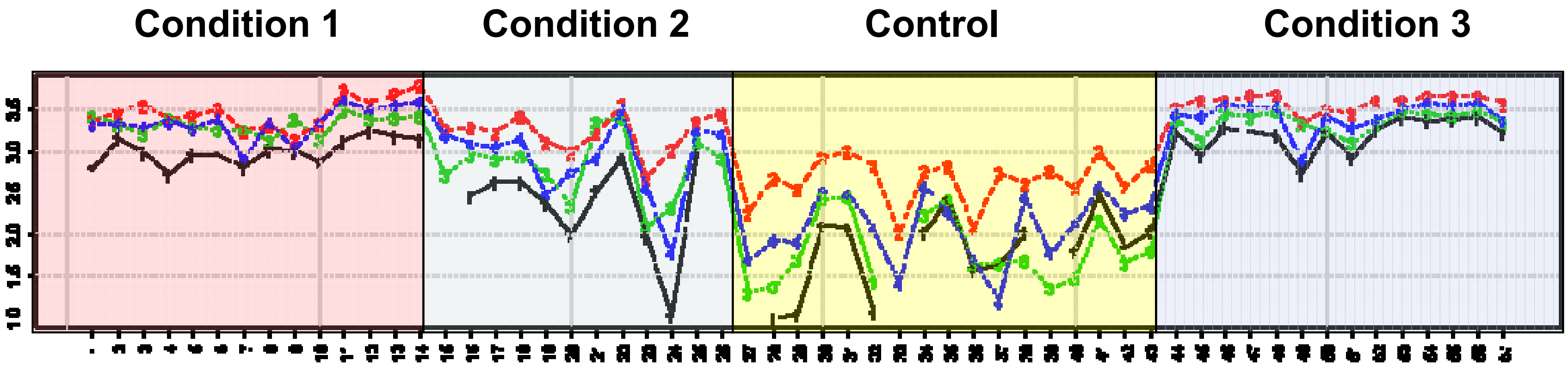
- Quantitative protein inference from peptide data
- Complications
  - Multiple, possibly inconsistent peptide measurements for same protein
  - Systematic abundance variation within and between conditions
    - How should we use information from blocking and randomization of experiments?
  - High rate 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

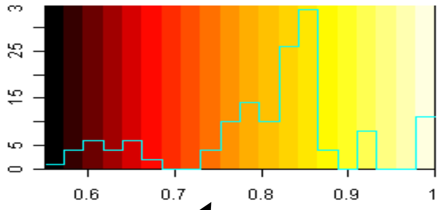
    1. SADLNVDSIISYWK
    2. LLLTSTGAGIIDVIK
    3. LIVGFPAYGHTFILSDPSK
    4. IPELSQSLDYIQVMTYDLHDPK

  - Plot peptide abundance across 57 datasets (for 4 conditions)



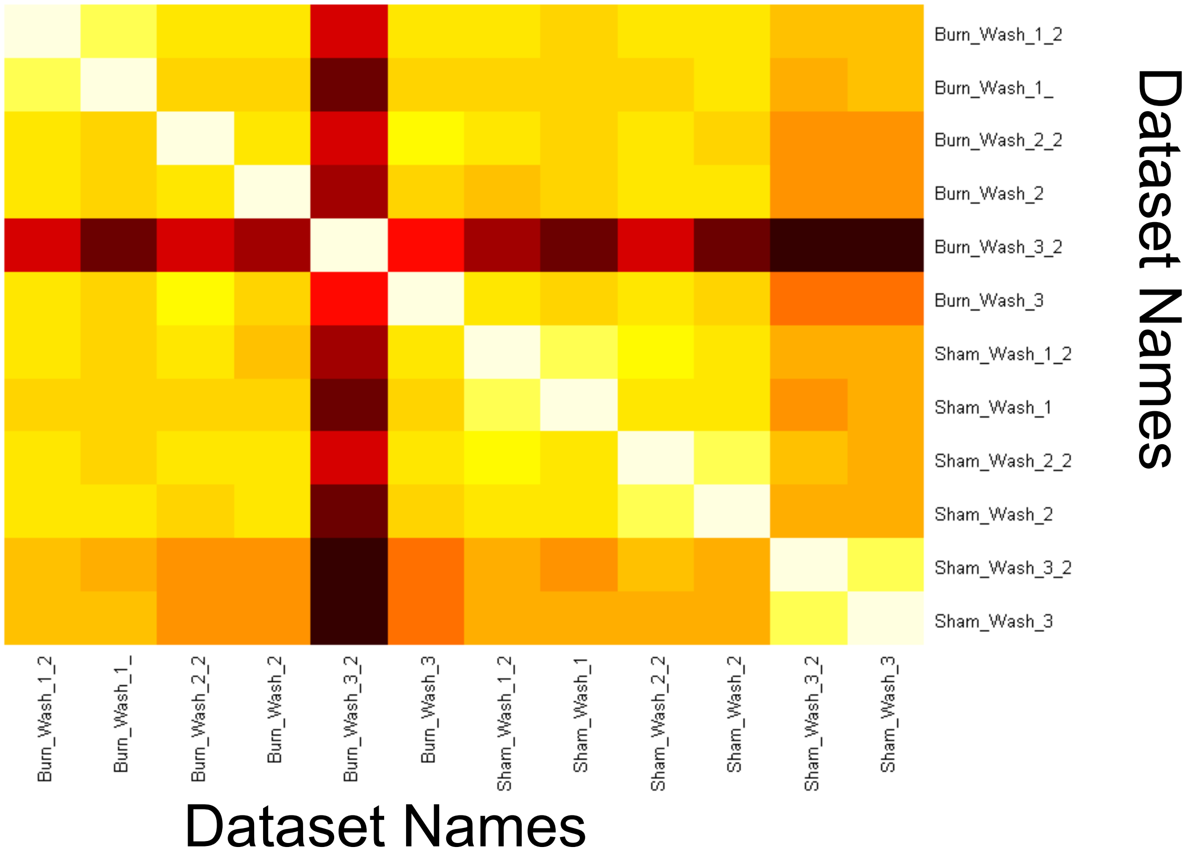
→ Outlier detection and normalization need to be performed before meaningful abundance information can be inferred

Outlier Detection

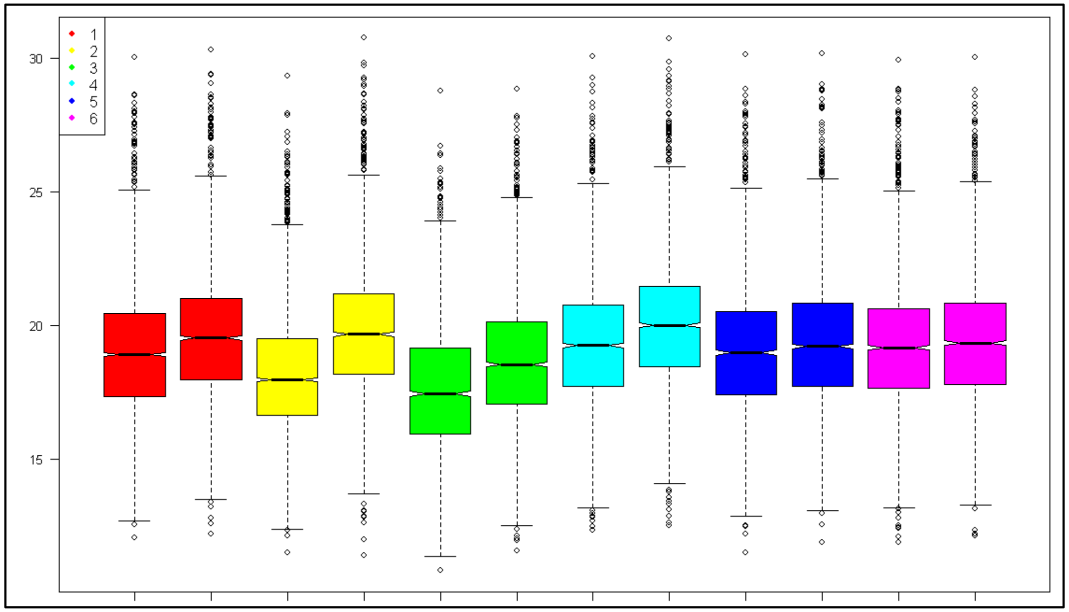


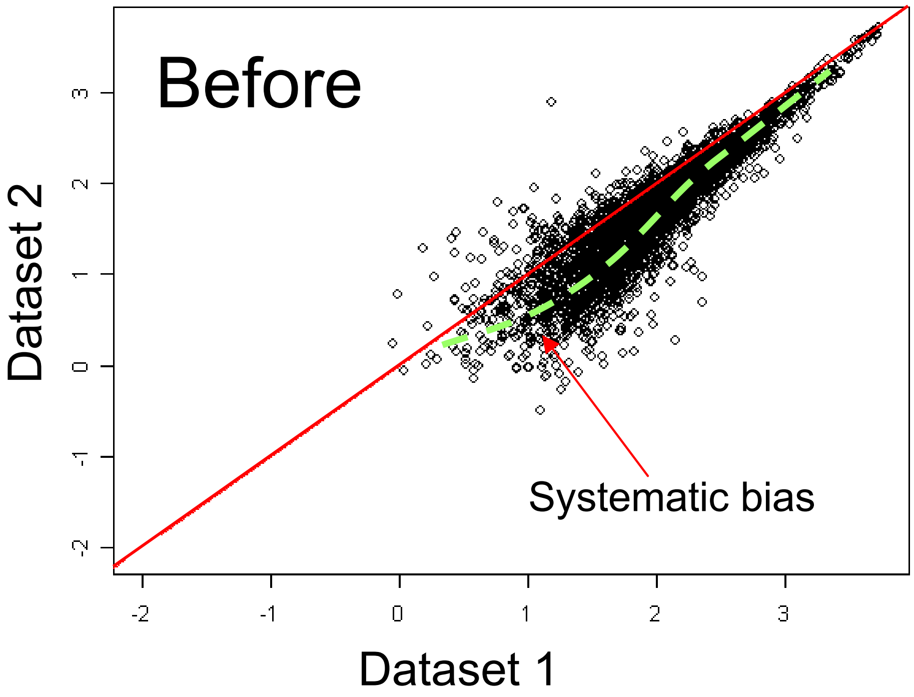
Color legend with overlaid histogram of correlation values

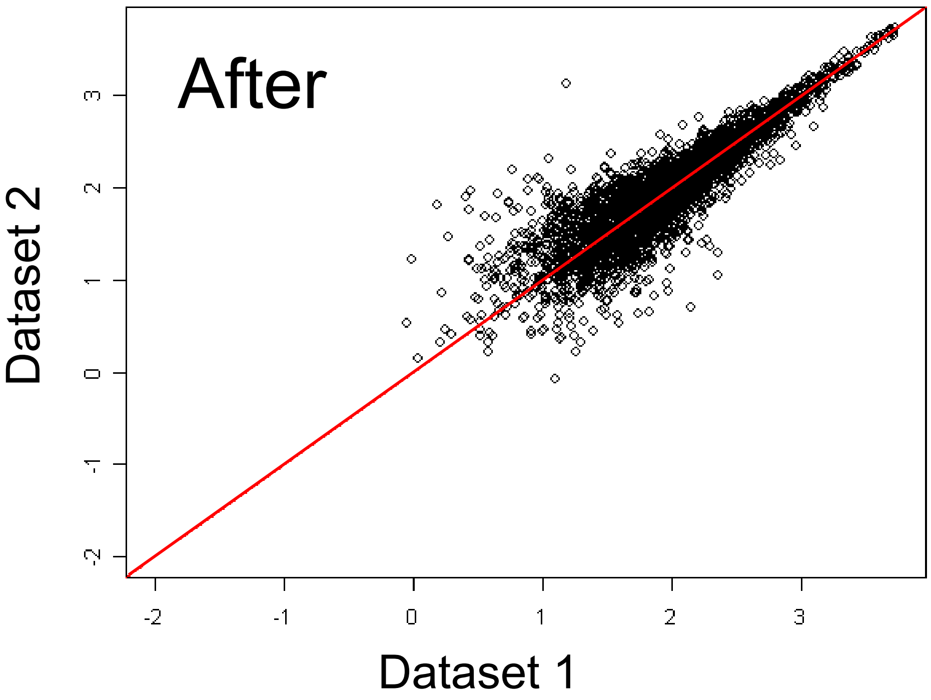
Outlier dataset



Normalization

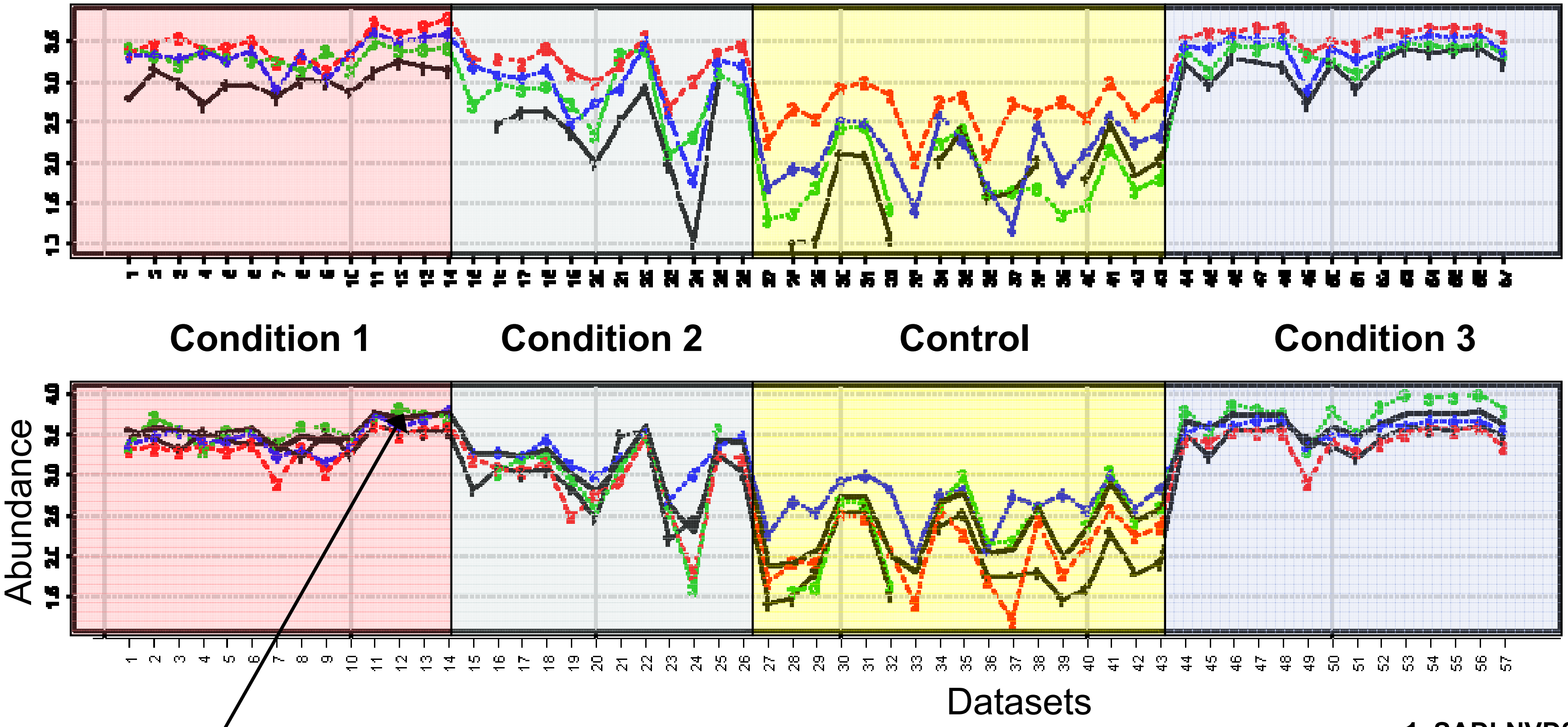






Infer Protein Abundances from Peptide Abundances

- 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)

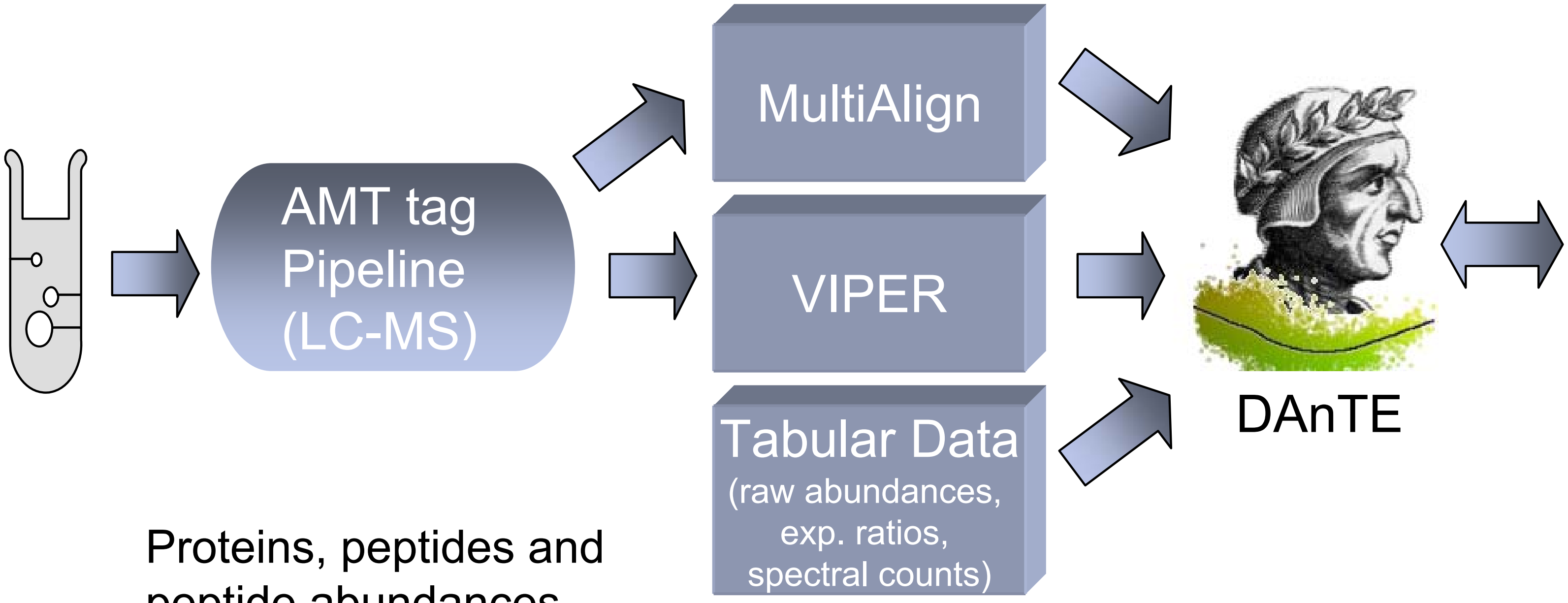
Scaled peptide abundances for this protein's 4 peptides

Median protein abundance (dark black line)

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

Data Analysis Tool Extension (DAnTE)

A software tool for downstream quantitative protein inference



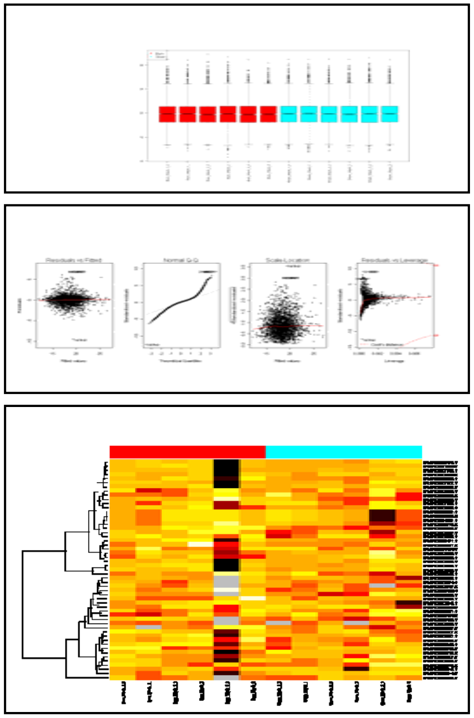
AMT tag Pipeline (LC-MS)

MultiAlign

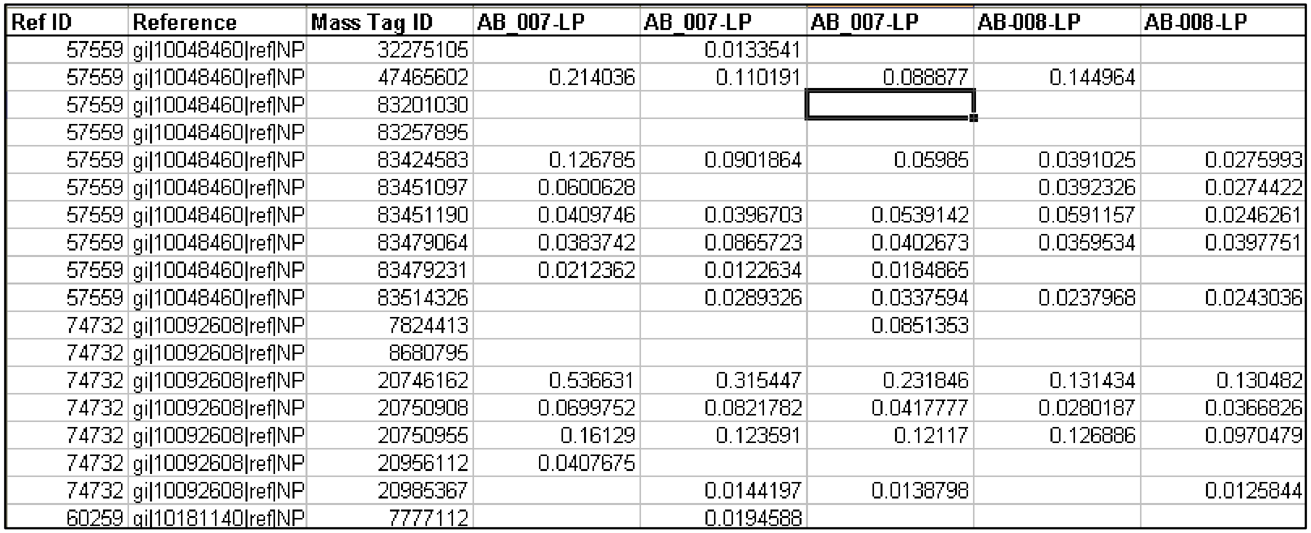
VIPER

Tabular Data (raw abundances, exp. ratios, spectral counts)

DAnTE



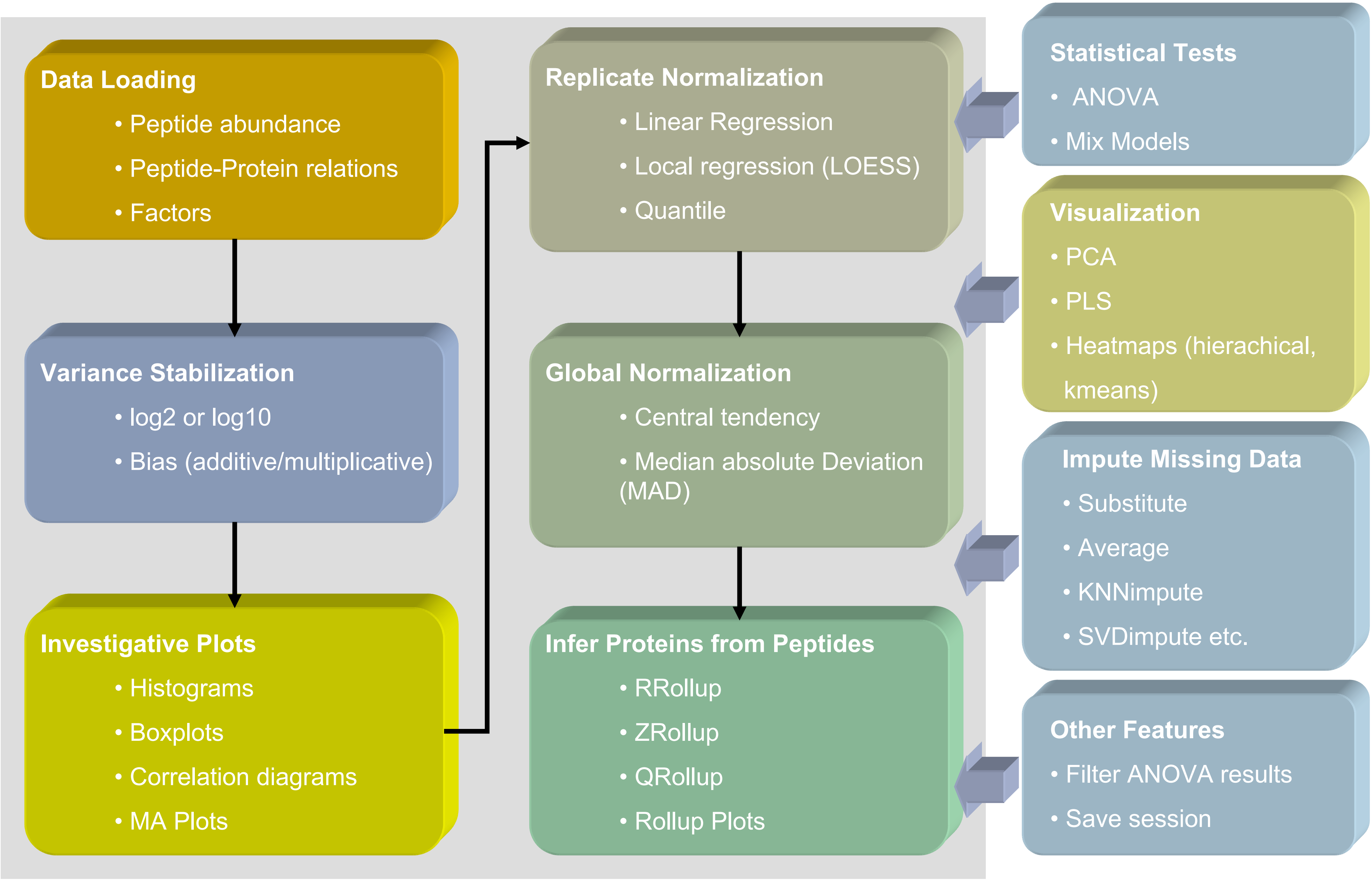
Proteins, peptides and peptide abundances



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			

Software by Ashoka Polpitiya

Interactive Analysis in DAnTE



Data Loading

- Peptide abundance
- Peptide-Protein relations
- Factors

Variance Stabilization

- log2 or log10
- Bias (additive/multiplicative)

Investigative Plots

- Histograms
- Boxplots
- Correlation diagrams
- MA Plots

Replicate Normalization

- Linear Regression
- Local regression (LOESS)
- Quantile

Global Normalization

- Central tendency
- Median absolute Deviation (MAD)

Infer Proteins from Peptides

- RRollup
- ZRollup
- QRollup
- Rollup Plots

Statistical Tests

- ANOVA
- Mix Models

Visualization

- PCA
- PLS
- Heatmaps (hierachical, kmeans)

Impute Missing Data

- Substitute
- Average
- KNNimpute
- SVDimpute etc.

Other Features

- Filter ANOVA results
- Save session

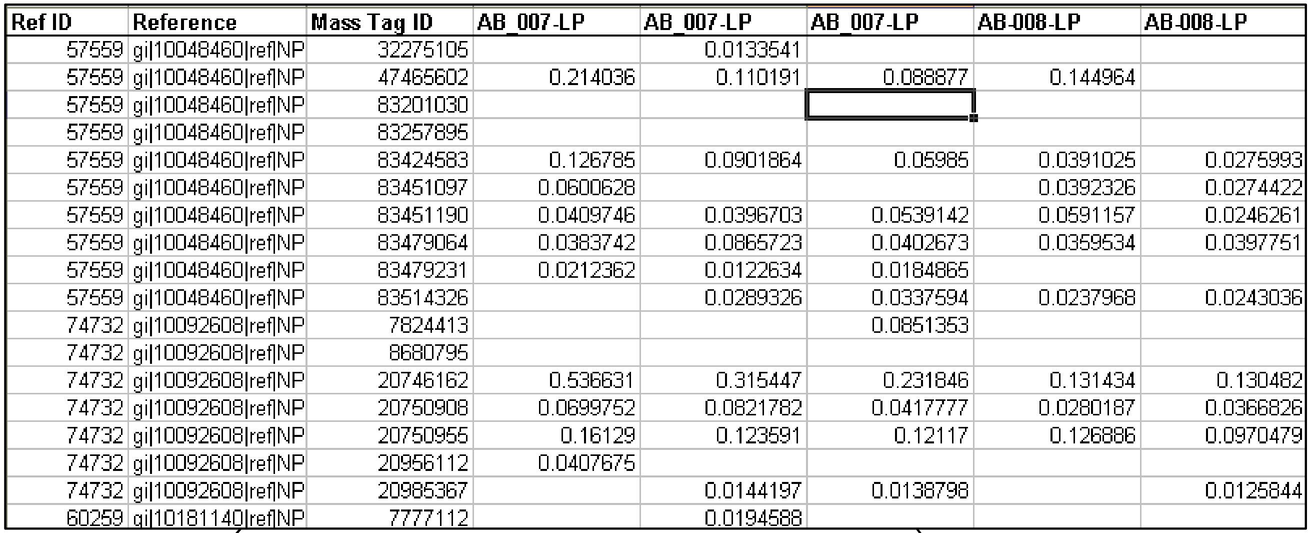
Outline of a Typical Analysis

- Load data
- Examine diagnostic plots
- Define factors
- Normalize
  - Within a Factor
    - Linear regression
    - LOESS (LOcal regrESSion)
    - Quantile
  - Across Factors
    - MAD
    - Central tendency
- Infer protein abundances from peptide abundances
  - RRollup, QRollup, and ZRollup
- ANOVA
- Save the results to a session file (.dnt)

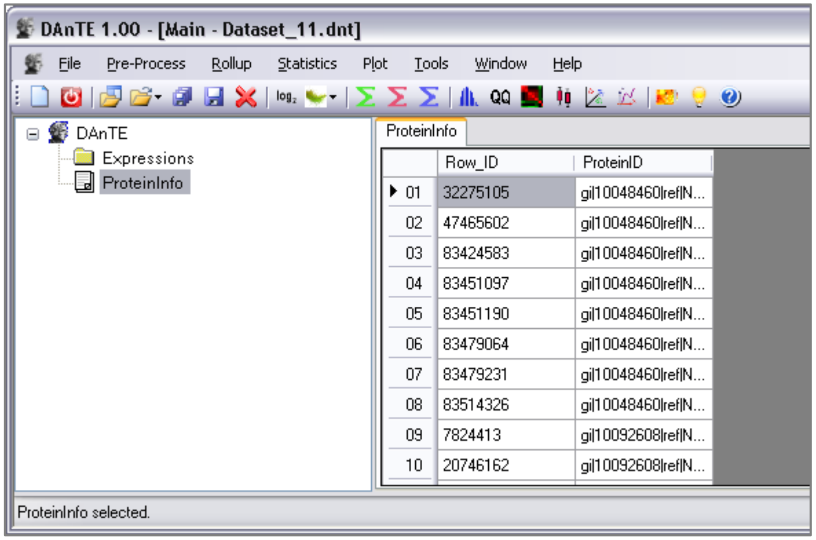
Load Data

Proteins, peptides, and peptide abundances

Tabular Data File:

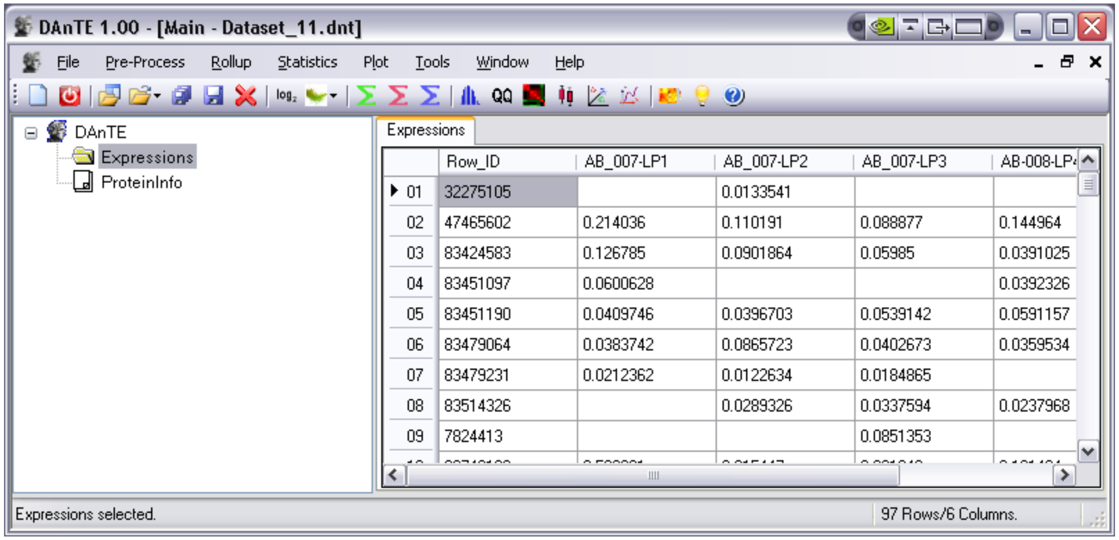


Ref ID	Reference	Mass Tag ID	AB_007-LP	AB_007-LP	AB_007-LP	AB-008-LP	AB-008-LP
57559	gi|10048460|ref|NP	32275105		0.0133541			
57559	gi|10048460|ref|NP	47465602	0.214036	0.110191	0.088877	0.144964	
57559	gi|10048460|ref|NP	83201030					
57559	gi|10048460|ref|NP	83257895					
57559	gi|10048460|ref|NP	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			



	Row_ID	ProteinID
01	32275105	gi|10048460|ref|N...
02	47465602	gi|10048460|ref|N...
03	83424583	gi|10048460|ref|N...
04	83451097	gi|10048460|ref|N...
05	83451190	gi|10048460|ref|N...
06	83479064	gi|10048460|ref|N...
07	83479231	gi|10048460|ref|N...
08	83514326	gi|10048460|ref|N...
09	7824413	gi|10092608|ref|N...
10	20746162	gi|10092608|ref|N...

ProteinInfo selected.

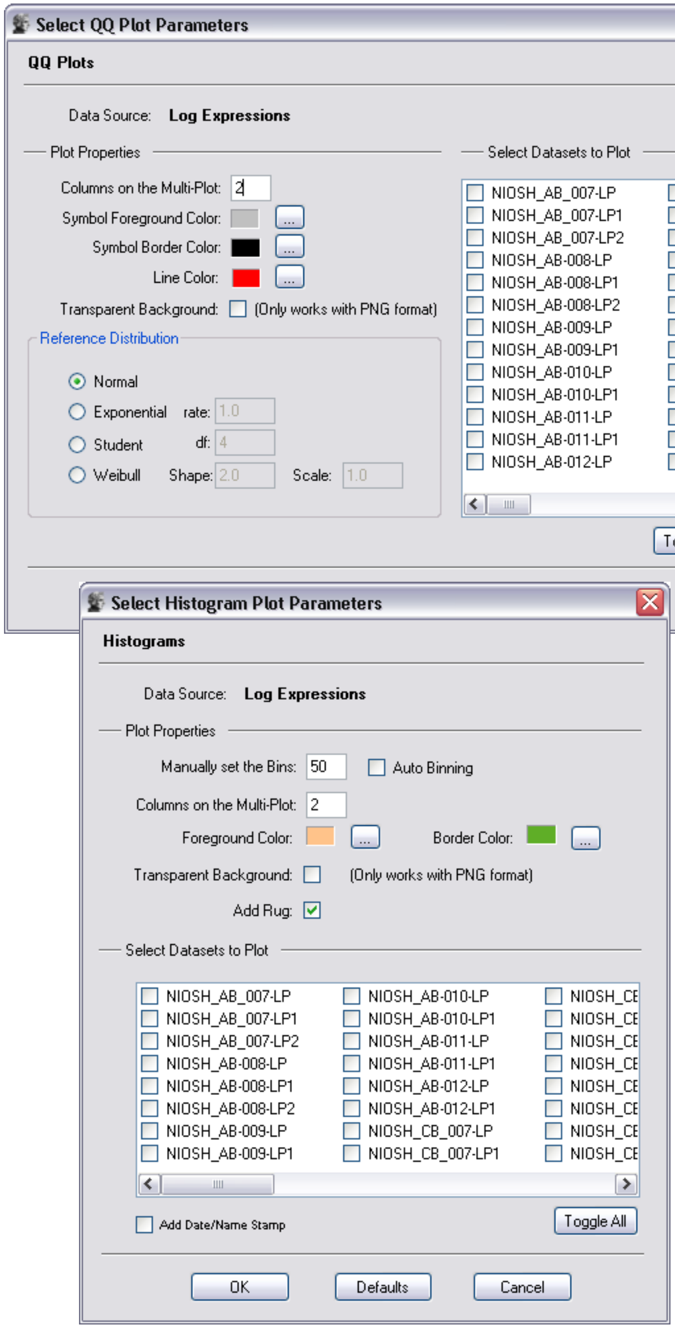


	Row_ID	AB_007-LP1	AB_007-LP2	AB_007-LP3	AB-008-LP
01	32275105		0.0133541		
02	47465602	0.214036	0.110191	0.088877	0.144964
03	83424583	0.126785	0.0901864	0.05985	0.0391025
04	83451097	0.0600628			0.0392326
05	83451190	0.0409746	0.0396703	0.0539142	0.0591157
06	83479064	0.0383742	0.0865723	0.0402673	0.0359534
07	83479231	0.0212362	0.0122634	0.0184865	
08	83514326		0.0289326	0.0337594	0.0237968
09	7824413			0.0851353	

Expressions selected.

97 Rows/6 Columns.

Diagnostic Plots: Check Normality



Select QQ Plot Parameters

QQ Plots

Data Source: **Log Expressions**

Plot Properties

Columns on the Multi-Plot: 2

Symbol Foreground Color: ▨ ...

Symbol Border Color: ■ ...

Line Color: 🟥 ...

Transparent Background: □ (Only works with PNG format)

Reference Distribution

● Normal

○ Exponential rate: 1.0

○ Student df: 4

○ Weibull Shape: 2.0 Scale: 1.0

Select Datasets to Plot

□ NIOSH_AB_007-LP □ NIOSH
□ NIOSH_AB_007-LP1 □ NIOSH
□ NIOSH_AB_007-LP2 □ NIOSH
□ NIOSH_AB-008-LP □ NIOSH
□ NIOSH_AB-008-LP1 □ NIOSH
□ NIOSH_AB-008-LP2 □ NIOSH
□ NIOSH_AB-009-LP □ NIOSH
□ NIOSH_AB-009-LP1 □ NIOSH
□ NIOSH_AB-010-LP □ NIOSH
□ NIOSH_AB-010-LP1 □ NIOSH
□ NIOSH_AB-011-LP □ NIOSH
□ NIOSH_AB-011-LP1 □ NIOSH
□ NIOSH_AB-012-LP □ NIOSH

Toggle All

Select Histogram Plot Parameters

Histograms

Data Source: **Log Expressions**

Plot Properties

Manually set the Bins: 50 □ Auto Binning

Columns on the Multi-Plot: 2

Foreground Color: ▨ ... Border Color: 🟩 ...

Transparent Background: □ (Only works with PNG format)

Add Rug: ☑

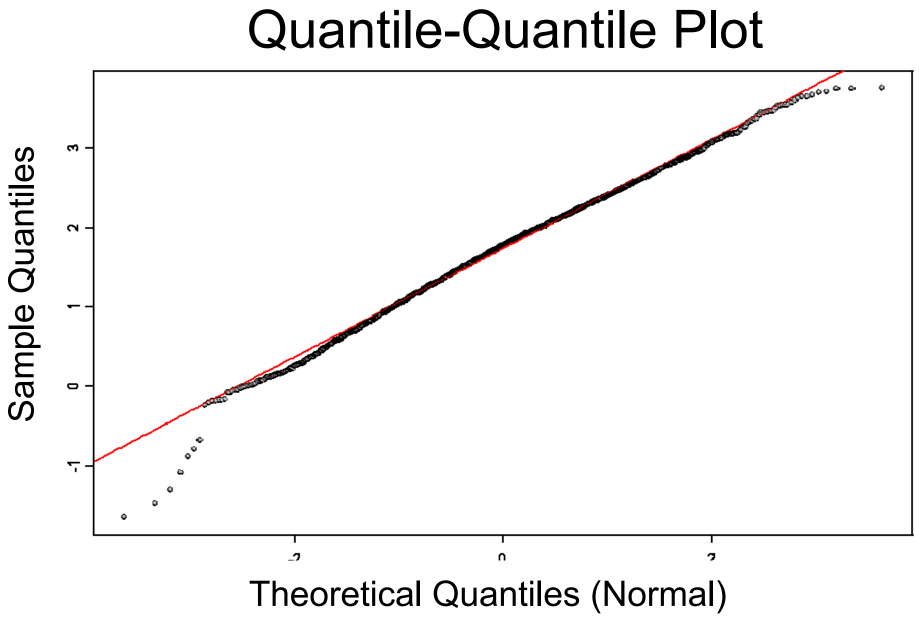
Select Datasets to Plot

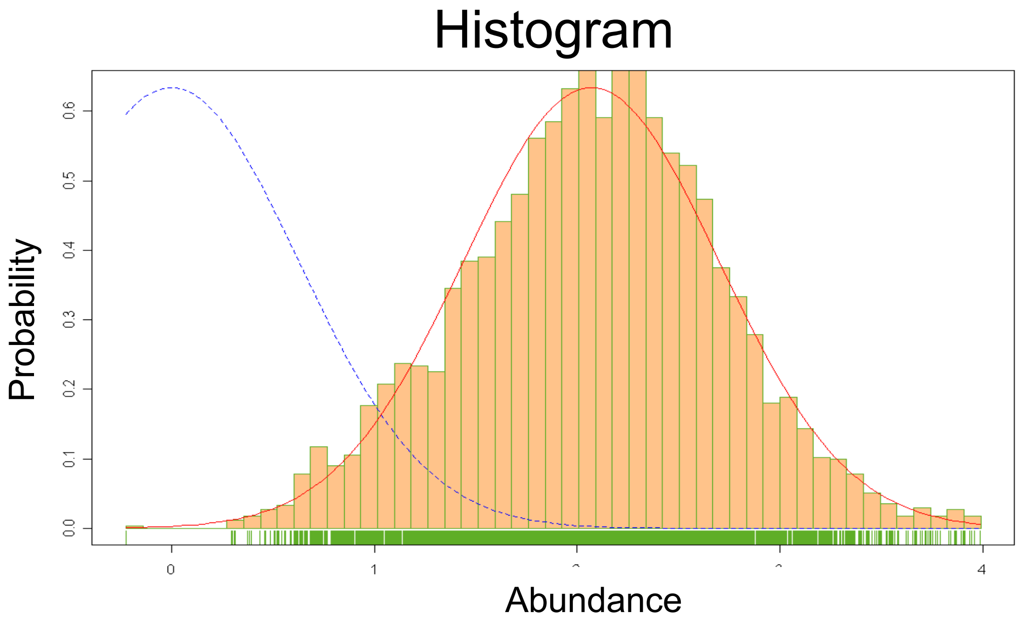
□ NIOSH_AB_007-LP □ NIOSH_AB-010-LP □ NIOSH_CE
□ NIOSH_AB_007-LP1 □ NIOSH_AB-010-LP1 □ NIOSH_CE
□ NIOSH_AB_007-LP2 □ NIOSH_AB-011-LP □ NIOSH_CE
□ NIOSH_AB-008-LP □ NIOSH_AB-011-LP1 □ NIOSH_CE
□ NIOSH_AB-008-LP1 □ NIOSH_AB-012-LP □ NIOSH_CE
□ NIOSH_AB-008-LP2 □ NIOSH_AB-012-LP1 □ NIOSH_CE
□ NIOSH_AB-009-LP □ NIOSH_CB_007-LP □ NIOSH_CE
□ NIOSH_AB-009-LP1 □ NIOSH_CB_007-LP1 □ NIOSH_CE

□ Add Date/Name Stamp

Toggle All

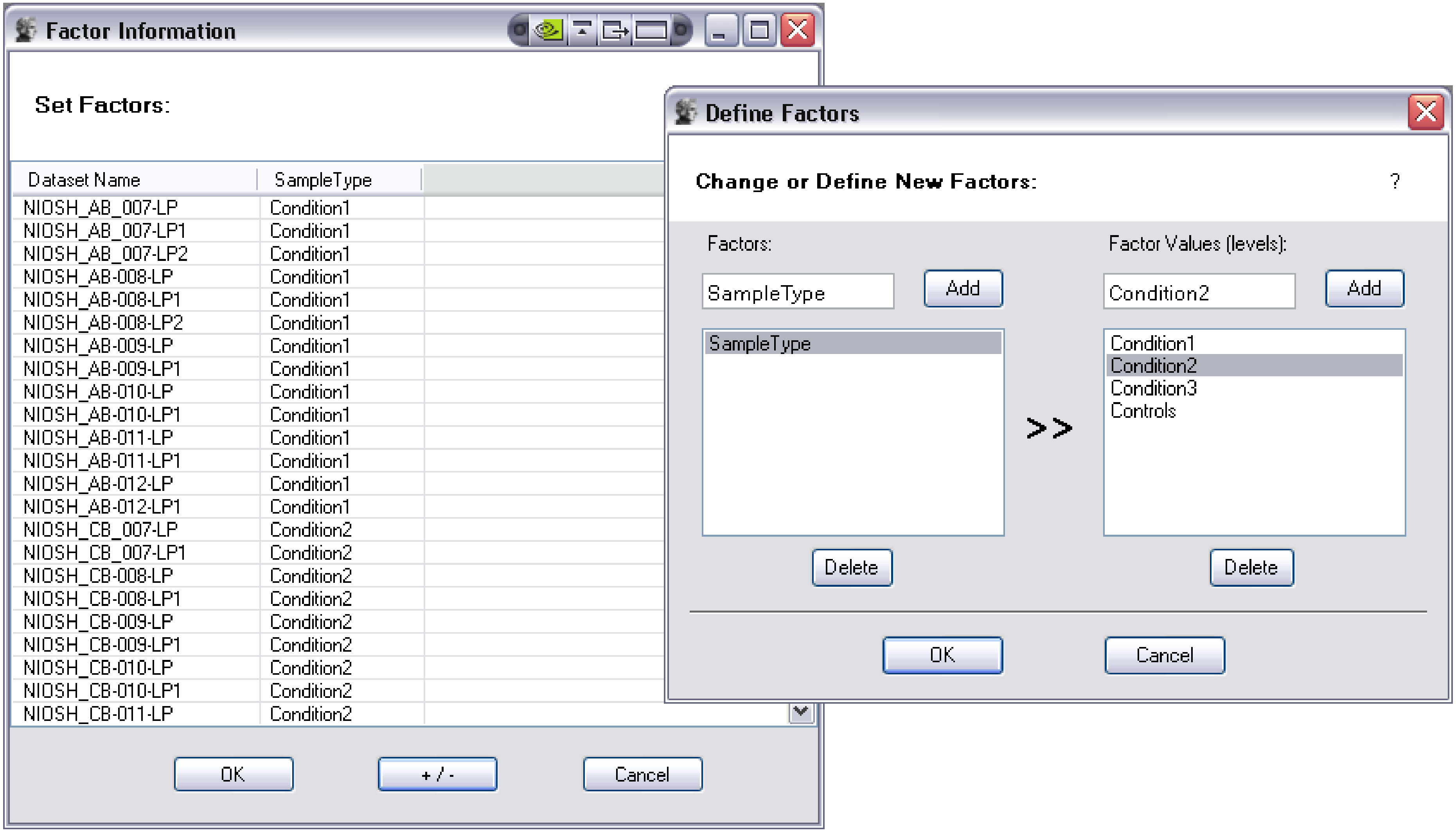
OK Defaults Cancel





Factors

- Capture experimental design through factors
  - For example, gender, sample type, technical replicate, and/or biological replicate



Dataset Name	SampleType
NIOSH_AB_007-LP	Condition1
NIOSH_AB_007-LP1	Condition1
NIOSH_AB_007-LP2	Condition1
NIOSH_AB-008-LP	Condition1
NIOSH_AB-008-LP1	Condition1
NIOSH_AB-008-LP2	Condition1
NIOSH_AB-009-LP	Condition1
NIOSH_AB-009-LP1	Condition1
NIOSH_AB-010-LP	Condition1
NIOSH_AB-010-LP1	Condition1
NIOSH_AB-011-LP	Condition1
NIOSH_AB-011-LP1	Condition1
NIOSH_AB-012-LP	Condition1
NIOSH_AB-012-LP1	Condition1
NIOSH_CB_007-LP	Condition2
NIOSH_CB_007-LP1	Condition2
NIOSH_CB-008-LP	Condition2
NIOSH_CB-008-LP1	Condition2
NIOSH_CB-009-LP	Condition2
NIOSH_CB-009-LP1	Condition2
NIOSH_CB-010-LP	Condition2
NIOSH_CB-010-LP1	Condition2
NIOSH_CB-011-LP	Condition2

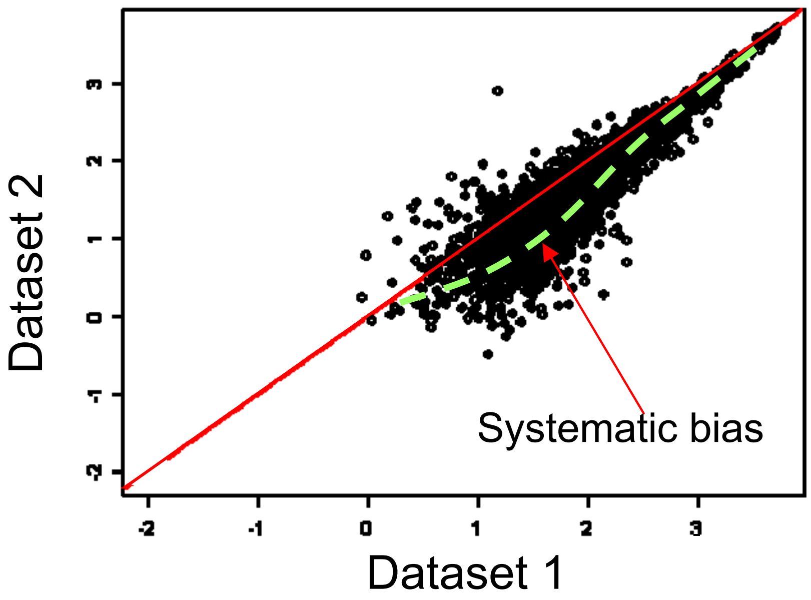
Factors: SampleType

SampleType

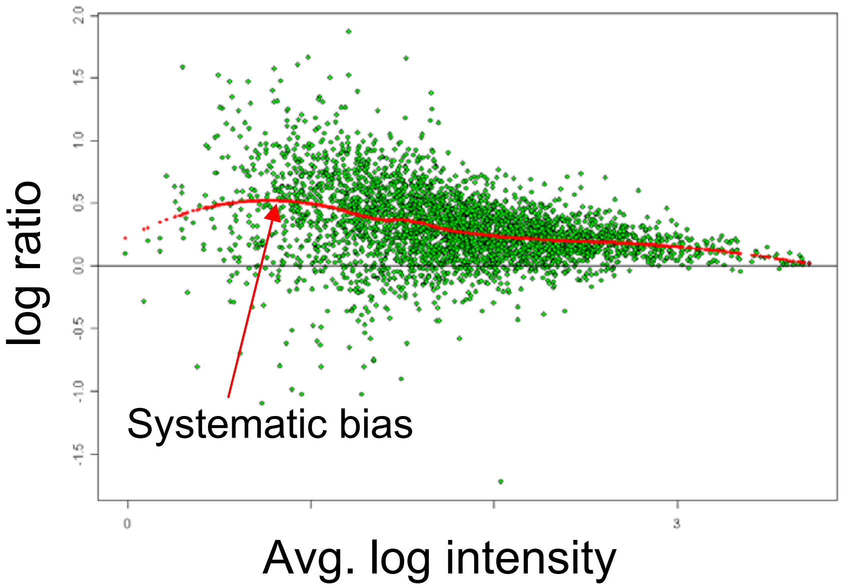
Factor Values (levels): Condition2

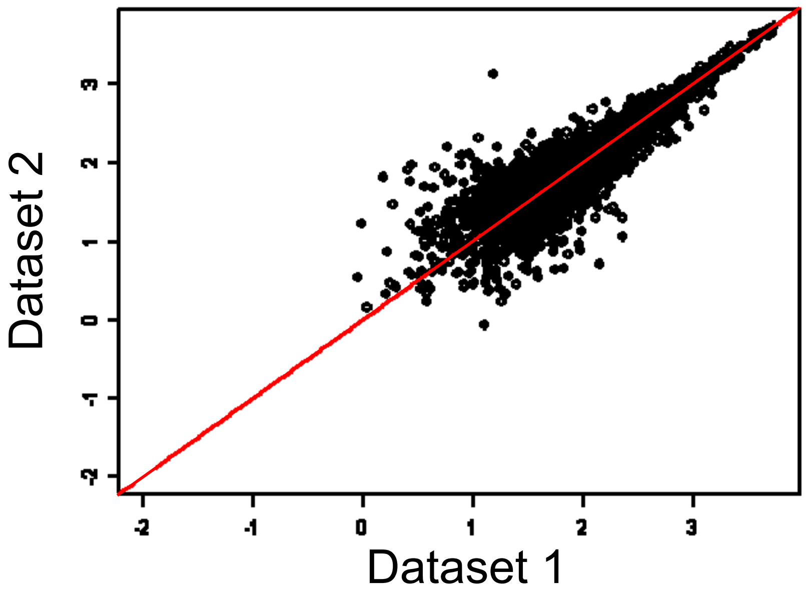
Condition1
Condition2
Condition3
Controls

Normalization: LOESS

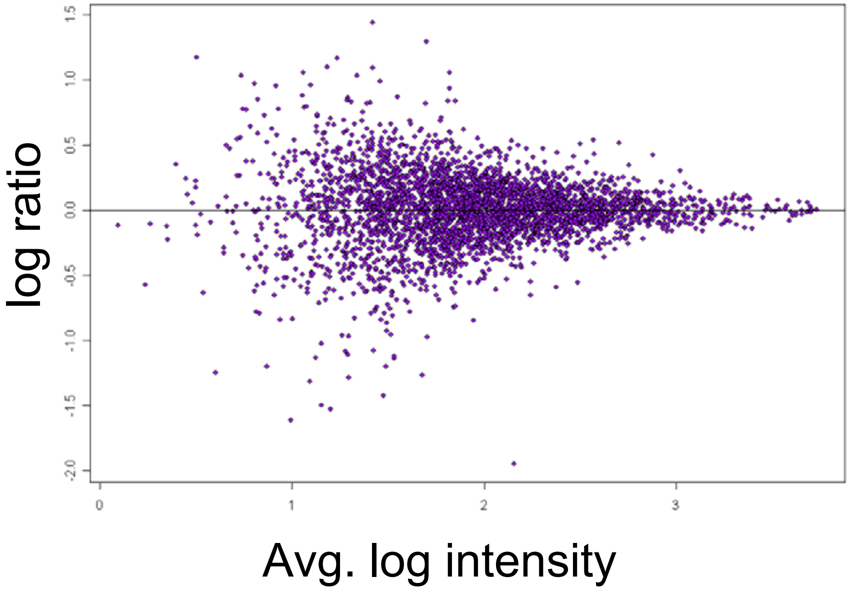


Raw



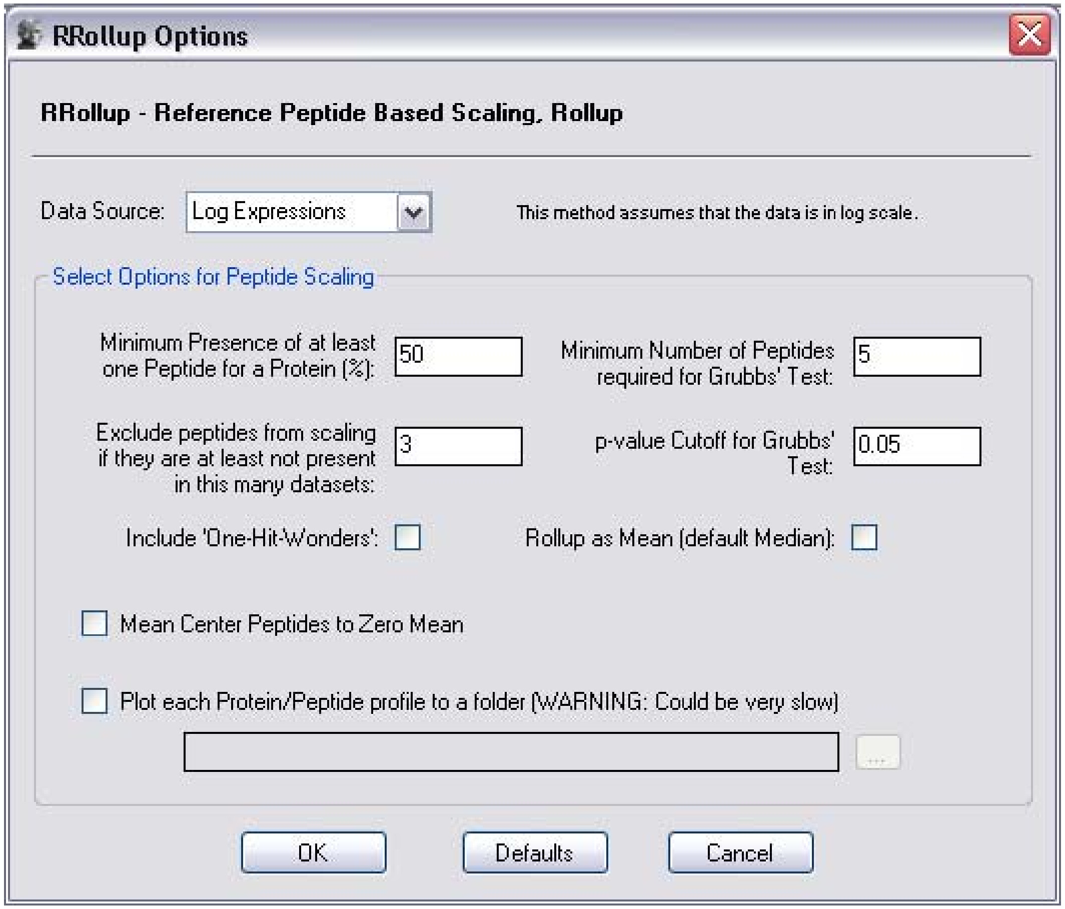


After LOESS



Protein Abundance Inference

- DAnTE currently has 3 different algorithms for rolling up peptide abundances to infer protein abundances
- Additional algorithms can be added as needed



RRollup Options

RRollup - Reference Peptide Based Scaling, Rollup

Data Source: Log Expressions

This method assumes that the data is in log scale.

Select Options for Peptide Scaling

Minimum Presence of at least one Peptide for a Protein (%): 50

Minimum Number of Peptides required for Grubbs' Test: 5

Exclude peptides from scaling if they are at least not present in this many datasets: 3

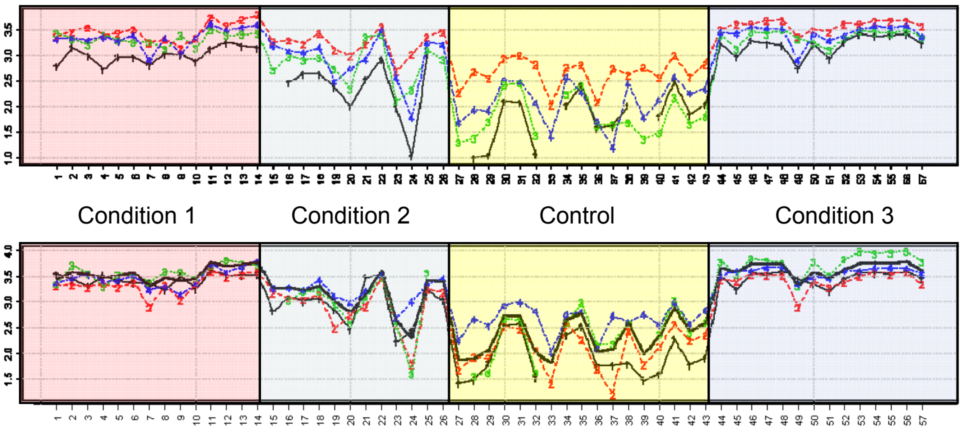
p-value Cutoff for Grubbs' Test: 0.05

Include 'One-Hit-Wonders': □

Rollup as Mean (default Median): □

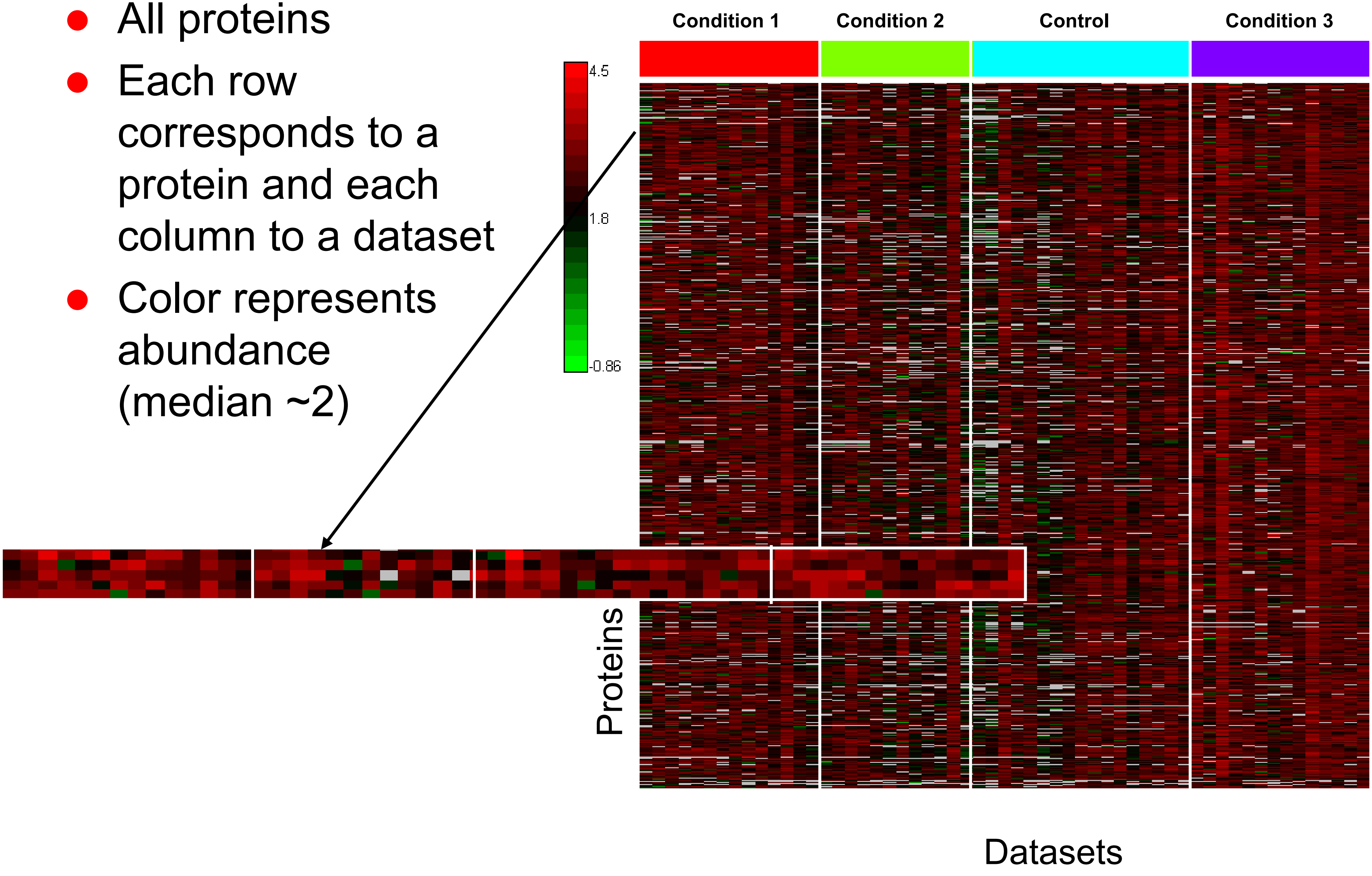
□ Mean Center Peptides to Zero Mean

□ Plot each Protein/Peptide profile to a folder (WARNING: Could be very slow)

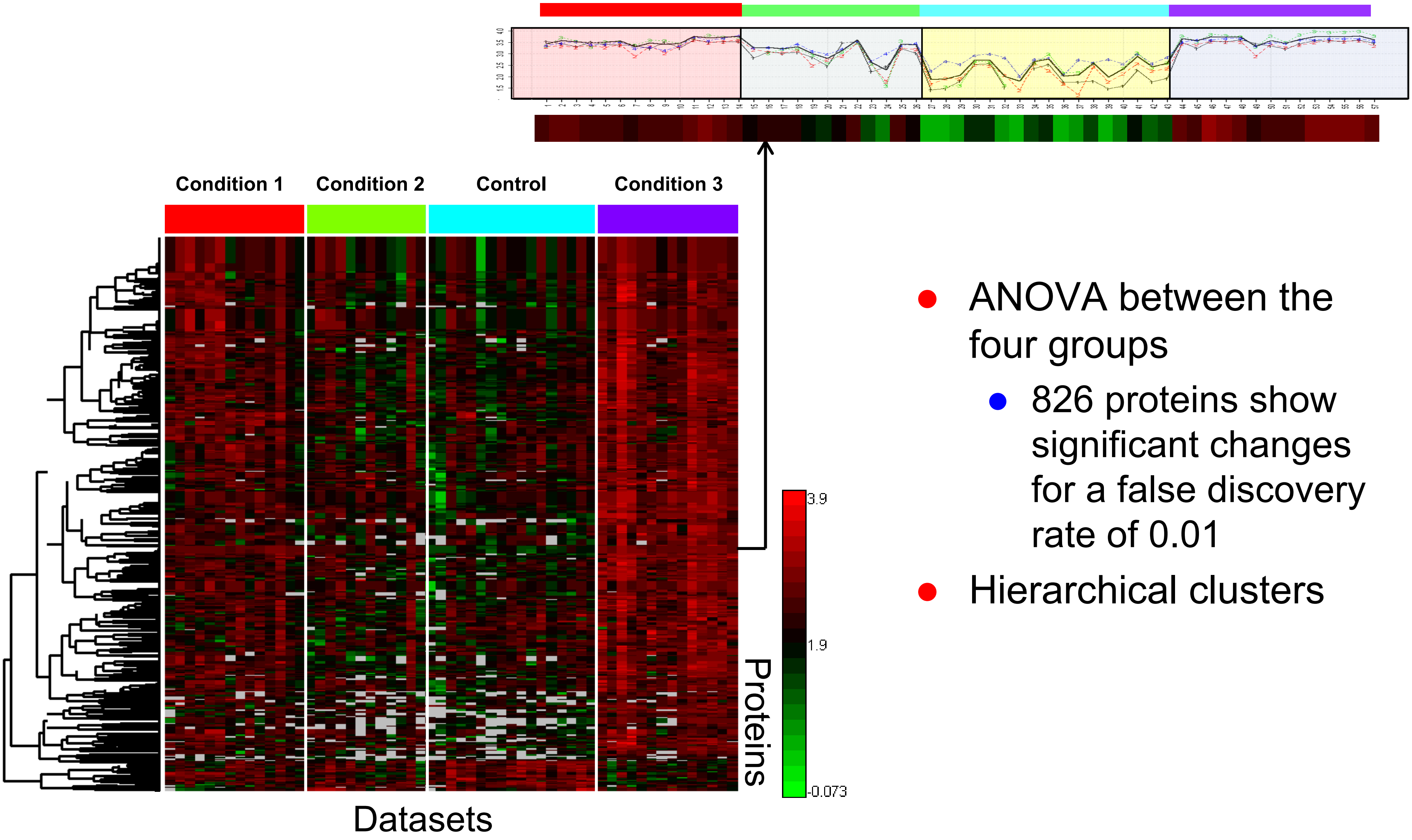


Protein Heatmap

- All proteins
- Each row corresponds to a protein and each column to a dataset
- Color represents abundance (median ~2)



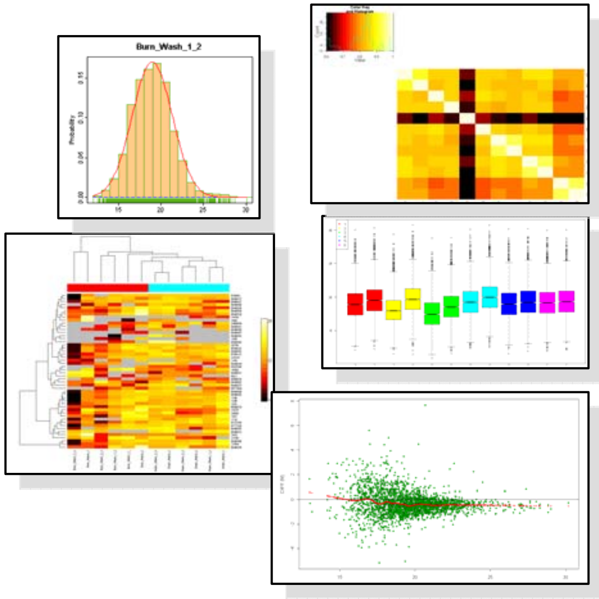
Significant Proteins



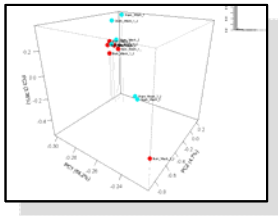
- ANOVA between the four groups
  - 826 proteins show significant changes for a false discovery rate of 0.01
- Hierarchical clusters

Complete DAnTE Feature List

- Data loading with peptide-protein group information
- 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

Course Outline

- Introduction (Adkins)
- Part I: Overview of Label-Free Quantitative Proteomics (Jaffe)
- Part II: Feature discovery in LC-MS datasets (Monroe and Jaitly)
- Part III: PEPPeR, GenePattern and Real-world examples (Jaffe)
  - PEPPeR: a self-contained web-based Biomarker Discovery pipeline
  - GenePattern: a suite of analysis and visualization tools that works with just about anything
- Break
- AMT tag Pipeline Demo (general)
- Panel Discussion
  - Questions
  - Future Directions

Part III: PEPPeR, GenePattern and Real-world examples

Jacob D. Jaffe

The Broad Institute of Harvard and MIT

Proteomics Platform

Section Outline

- PEPPeR: a self-contained web-based Biomarker Discovery pipeline
- GenePattern: a suite of analysis and visualization tools that works with just about anything
- Examples of use in the real world
  - Proof of principle by accidental discovery of markers
  - In-silico defractionation
  - Breast cancer biomarker discovery

PEPPeR:

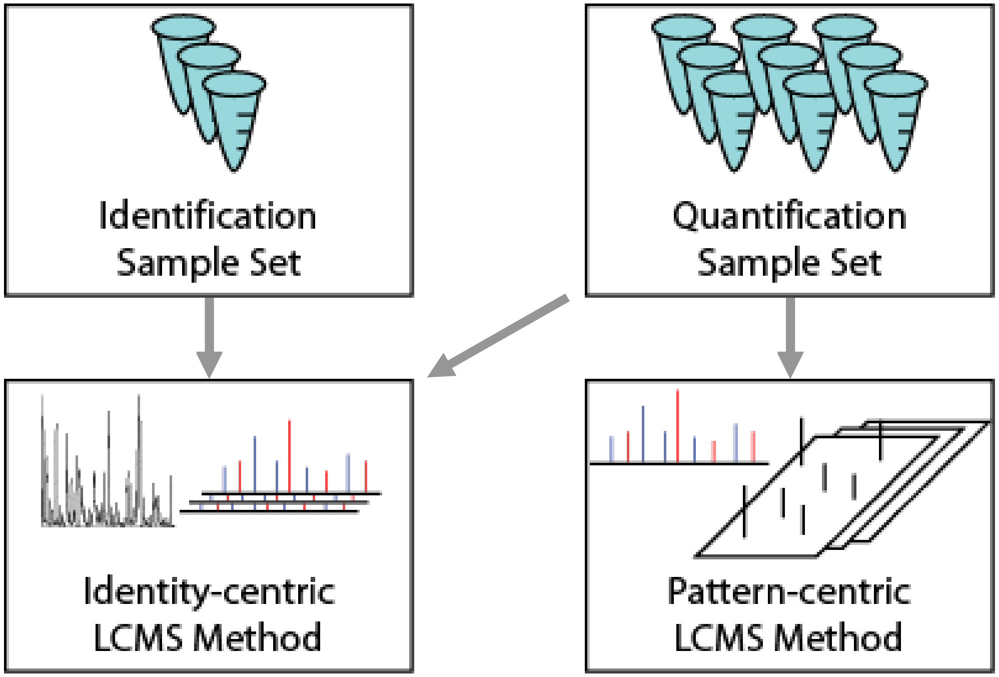
Platform for Experimental Proteomics Pattern Recognition

Jaffe JD, Mani DR, Leptos KC, Church GM, Gillette MA, Carr SA. PEPPeR, a Platform for Experimental Proteomic Pattern Recognition. *Mol Cell Proteomics*. 2006 Oct;5(10):1927-1941.

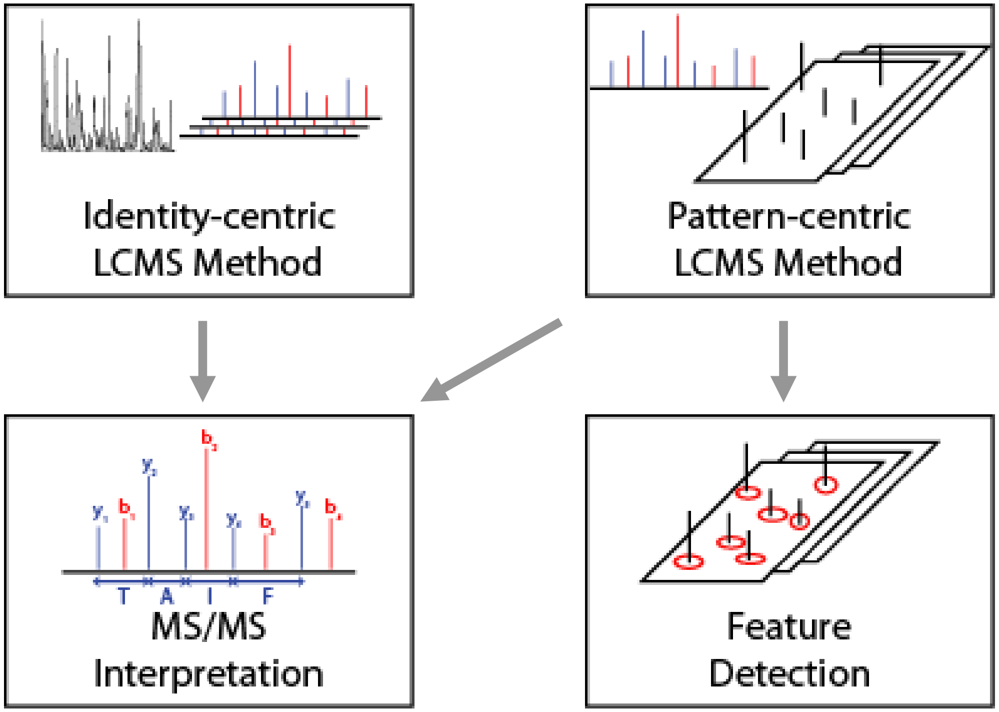
Multiple LCMS Experiments: Good with the Bad

- There is a lot of information in there
  - Peptide/protein IDs
  - Quantitative data
  - Statistical assessment
- The information may be noisy
  - Retention time drift
  - Instrument response noise
- Are there methods to leverage this information?
  - Without ‘perfect’ chromatography?
  - Without strict alignment?

PEPPeR Concepts – Samples and Data Acquisition



PEPPeR Concepts – Data Processing



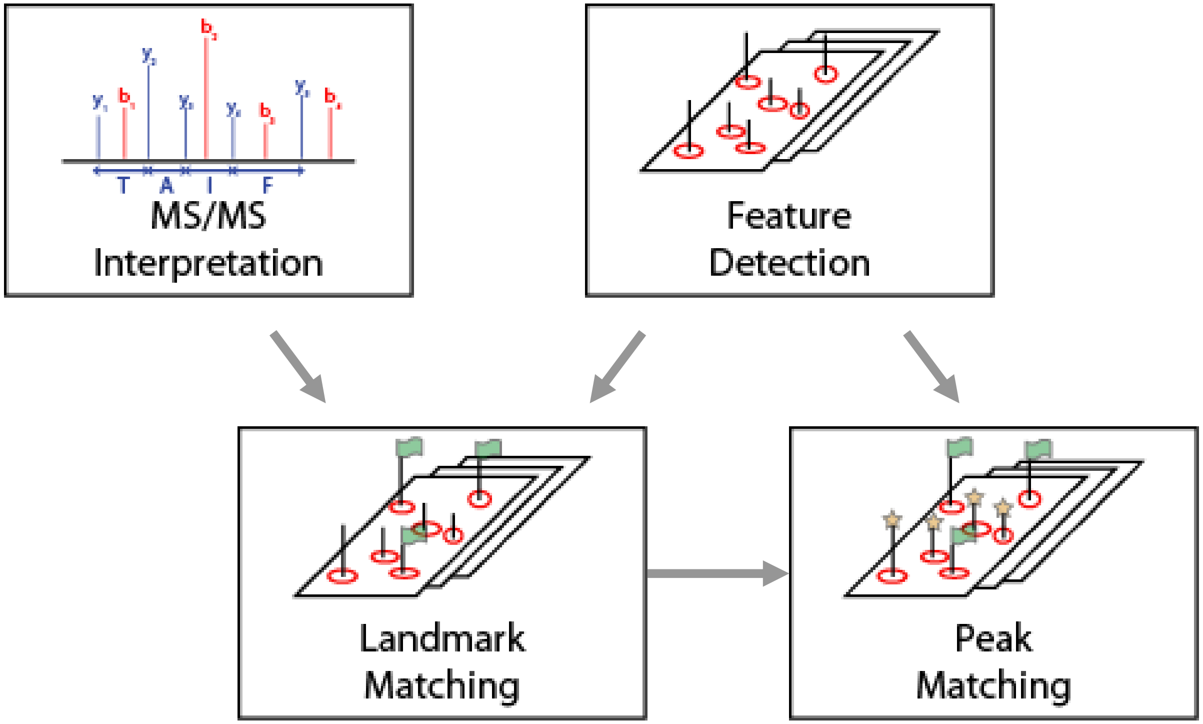
Identity-centric LCMS Method

Pattern-centric LCMS Method

MS/MS Interpretation

Feature Detection

PEPPeR Concepts – Processing Continued…



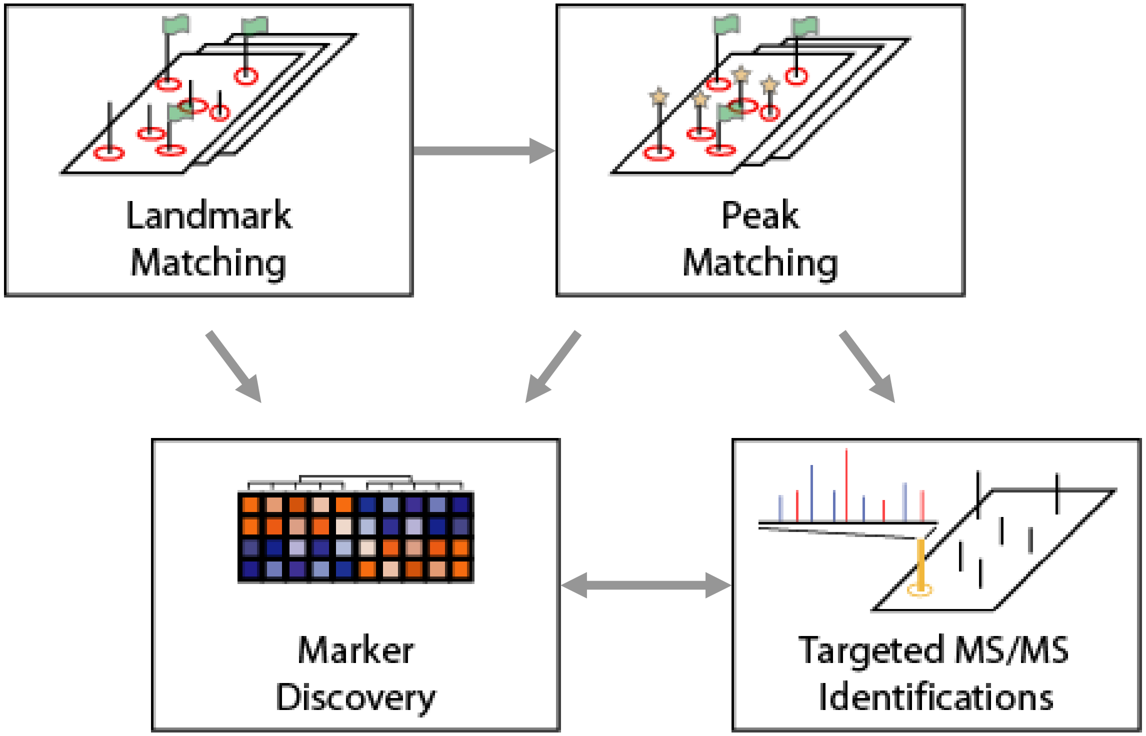
MS/MS Interpretation

Feature Detection

Landmark Matching

Peak Matching

PEPPeR Concepts – Analysis and Follow up



Landmark Matching

Peak Matching

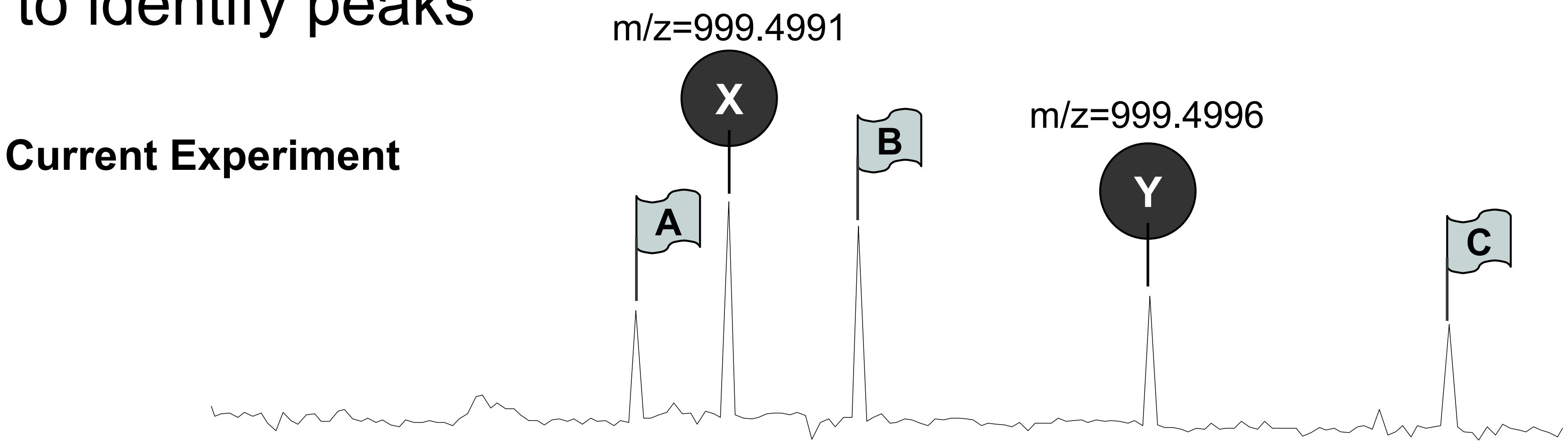
Marker Discovery

Targeted MS/MS Identifications

Landmark Matching: Identity Propagation

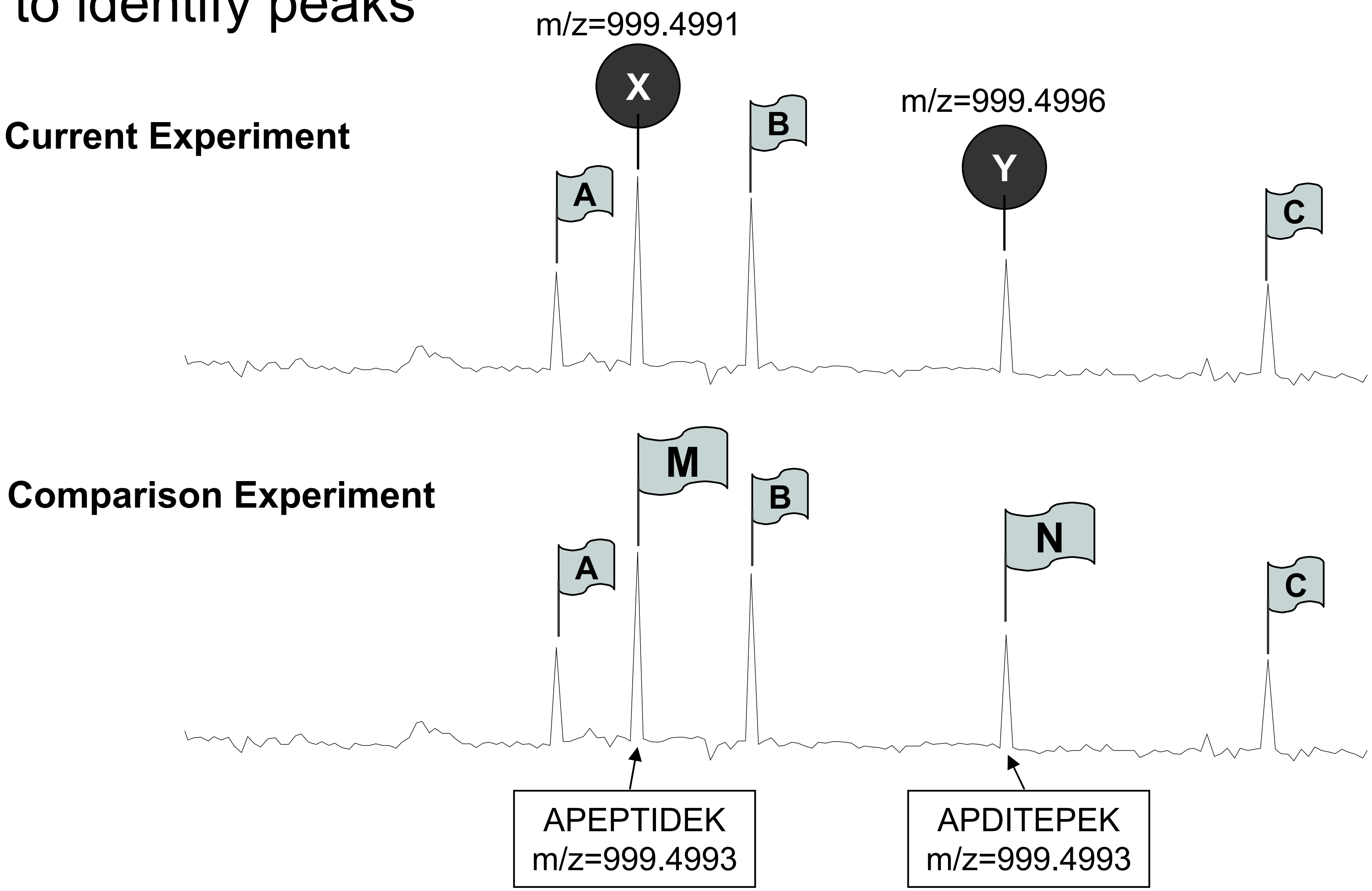
Landmark Matching

- Use accurate mass, relative retention order comparison to identify peaks



Landmark Matching: Identity Propagation

- Use accurate mass, relative retention order comparison to identify peaks



m/z=999.4991

m/z=999.4996

Current Experiment

Comparison Experiment

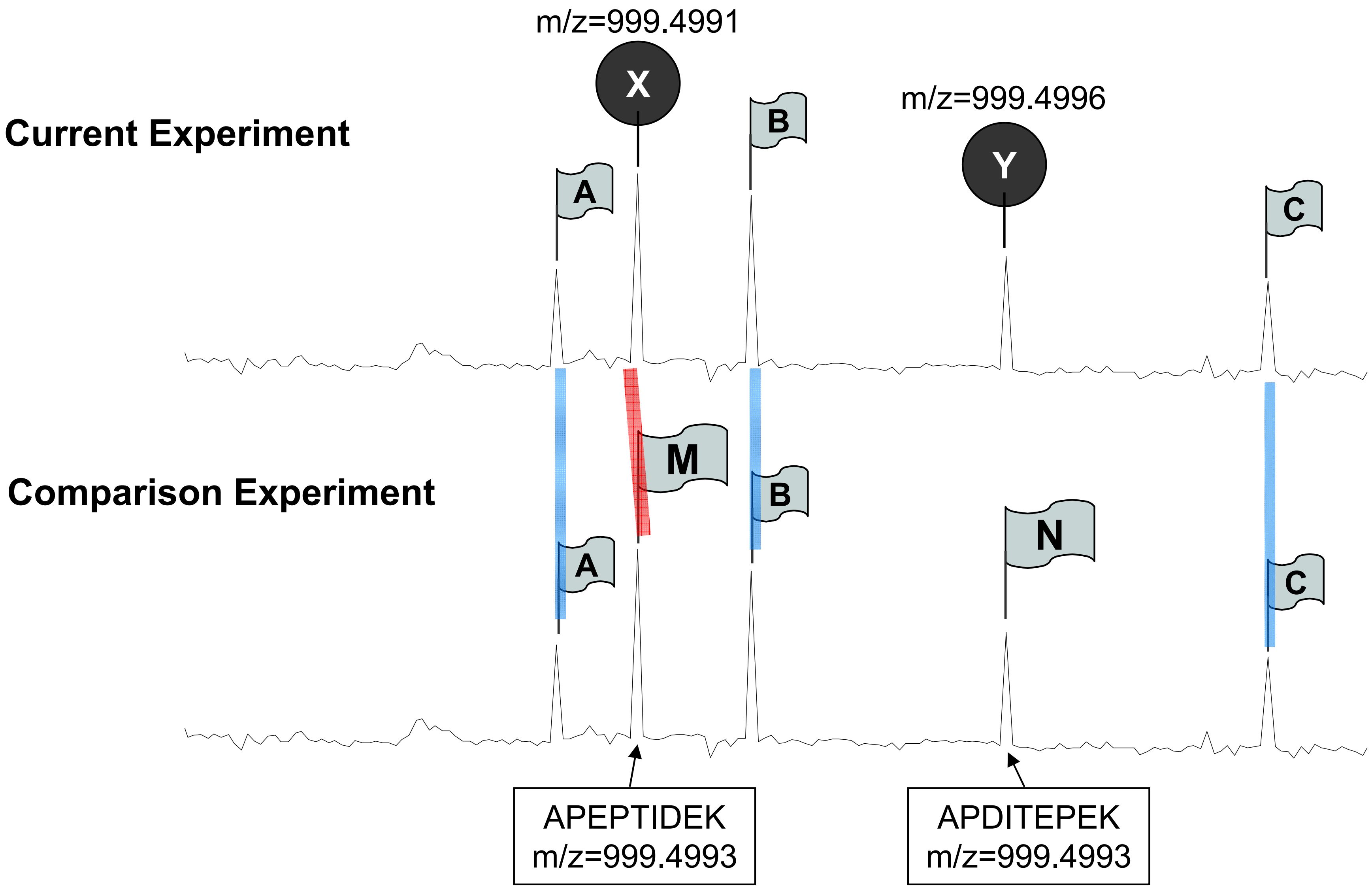
APEPTIDEK
m/z=999.4993

APDITEPEK
m/z=999.4993

Landmark Matching: Identity Propagation

Landmark Matching

Is **X** = **M** ?



Current Experiment

m/z=999.4991

m/z=999.4996

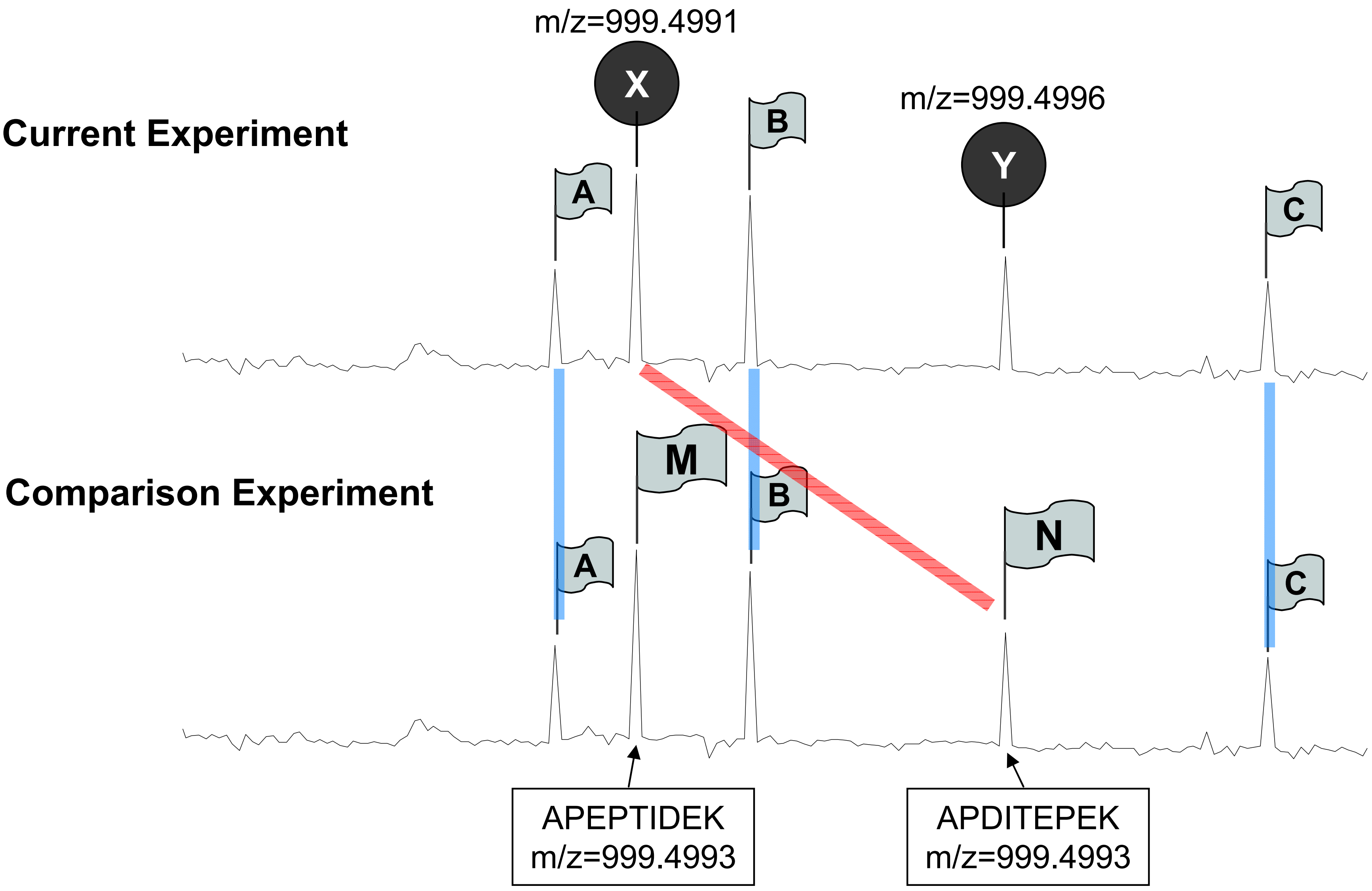
Comparison Experiment

APEPTIDEK
m/z=999.4993

APDITEPEK
m/z=999.4993

Landmark Matching: Identity Propagation

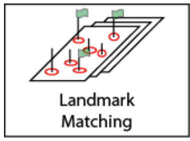
Is **X** = **N** ?



Nuts and bolts: How it works

- Match features to sequenced peptides in a single LCMS run
- Refine/recalibrate m/z tolerance
- Re-match features to sequenced peptides in a single LCMS run
- Now compare list of all features to Basis Set for mass, relative elution order matches given landmarks as reference points – ***propagation of identified features across multiple experiments***

Landmark Scoring and Confidence



$$S = \sum_{i=1}^{w} \left[\xi(\Lambda_{-i}, \Lambda_0) + \xi(\Lambda_0, \Lambda_i) \right]$$

$$\xi(m,n) = \begin{cases} 1 \text{ if } \tau(m) < \tau(n) \\ \text{if } \tau(m) > \tau(n) \begin{cases} 0.5 \text{ if } \tau(n) - \tau(m) < \delta \text{ and } \mu(m) + \sigma(m) > \mu(n) - \sigma(n) \\ -1 \text{ if } else \end{cases} \\ 0 \text{ if } else \end{cases}$$

$$P_{overall} = P_{m/z} P_{landmark}$$

$$P_{landmark} = P(landmark \mid m/z) = \frac{P(m/z \mid landmark) P(landmark)}{P(m/z \mid landmark) P(landmark) + (1 - P(m/z \mid landmark))(1 - P(landmark))}$$

Let:

Λ be a list of peptides observed in the comparison experiment ordered
by elution time. Here, elution time is defined by the centroid of all
MS/MS scans leading to the identification of the peptide.
Λ_0 is defined as the position of the putative assignment in Λ

μ(x) be the centroid of elution time of peptide ***x***
in the comparison experiment (in scans)

σ(x) be the standard deviation of elution time of peptide ***x***
in the comparison experiment (in scans)

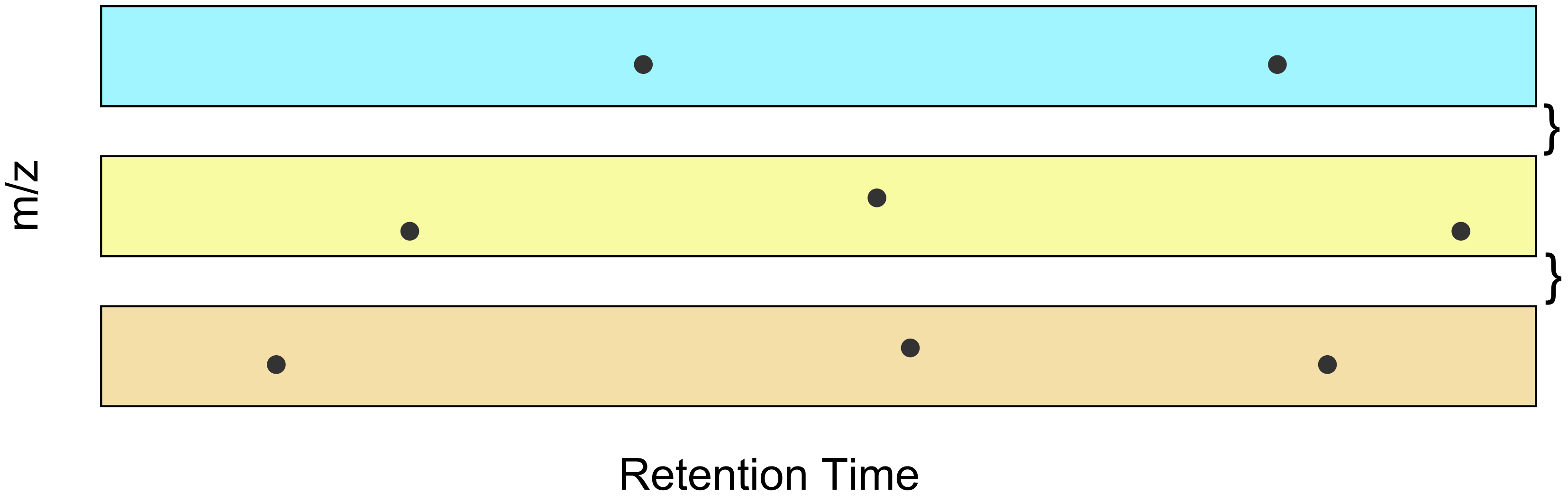
τ(x) be the centroid of elution time of peptide ***x***
in the current experiment (in seconds)

δ be the average retention time peak width, such that peptides eluting
within δ sec are considered to be co-eluting (typically δ = 30 s)

w the number of peptides to consider before and after
the putative assignment on the landmark list (typically ***w*** = 3)

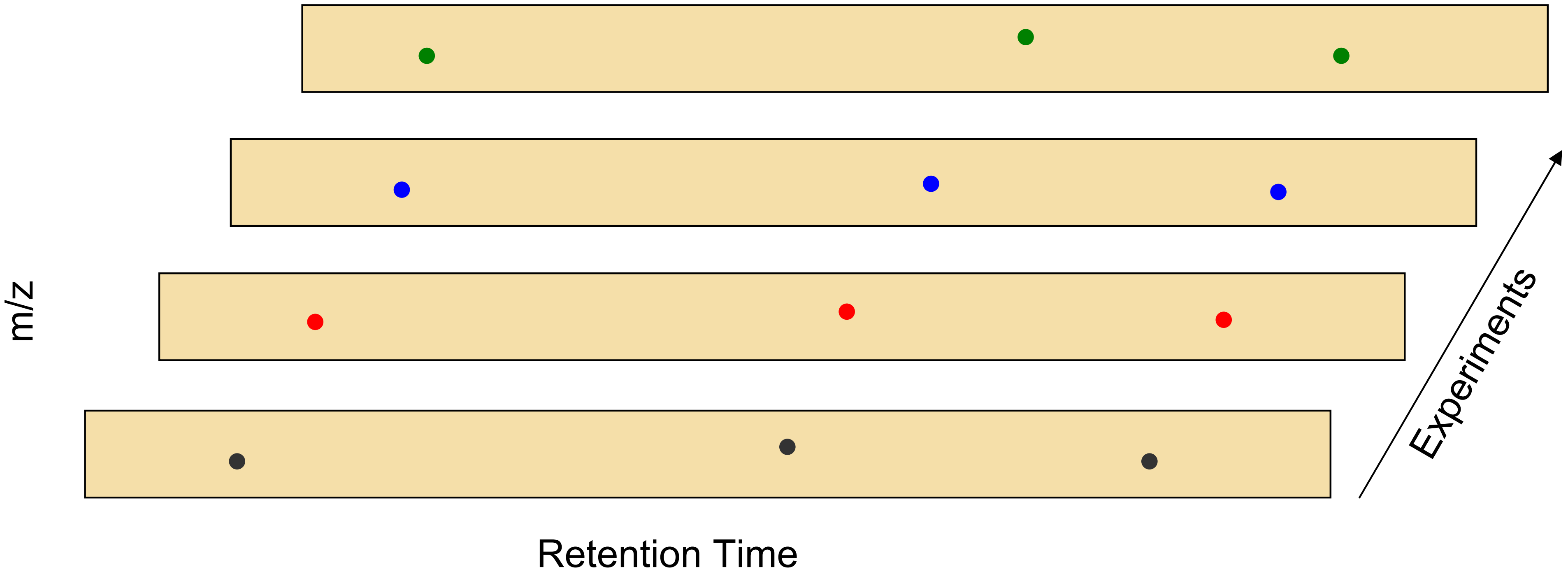
Peak Matching: Recognizing Identical Features

- Use landmarks to derive corrections and tolerances for clustering of features across LCMS experiments
    - Break down the problem to make it parallelizable



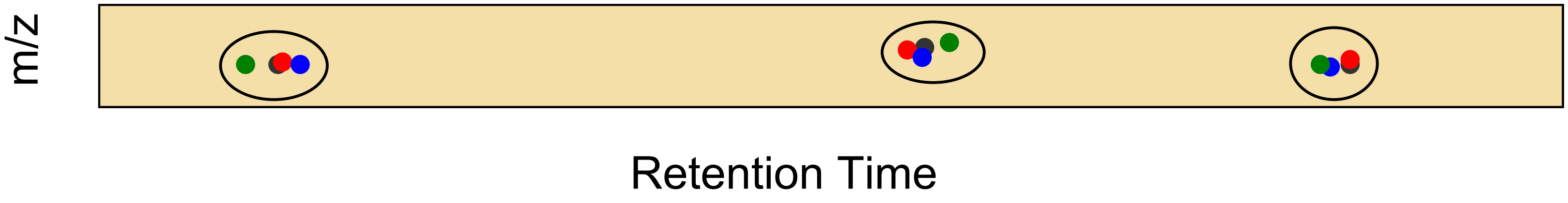
Peak Matching: Recognizing Identical Features

- Use landmarks to derive corrections and tolerances for clustering of features across LCMS experiments



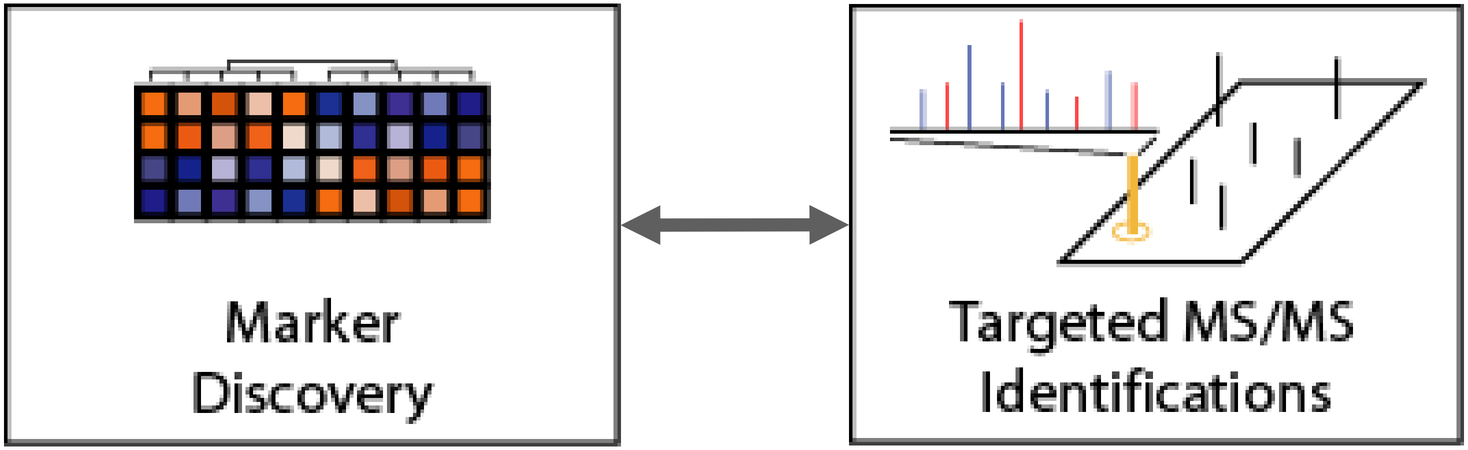
Peak Matching: Recognizing Identical Features

- Use landmarks to derive corrections and tolerances for clustering of features across LCMS experiments
  - Gaussian mixture model (GMM) with parameters determined by maximizing Likelihood ratio using Expectation Maximization (EM)
  - Number of clusters determined using Bayesian Information Criterion (BIC)
  - Coalesce clusters if M/Z and RT variation is within tolerance



Parameterized Peaks

Peak ID	*m/z*	*R.T.*	*z*	*Run 1*	*Run 2*	*Run 3*	*Run …*	*Identity*
1	490.3144	62.0	3	607.6	544.2	581.0	…	
2	743.3549	56.2	3	694.4	682.6	691.4	…	
3	999.4991	22.5	2	209.6	247.6	232.6	…	APEPTIDEK
4	396.7187	20.5	3	321.7	344.9	318.5	…	
5	934.6045	31.7	2	722.7	753.0	701.3	…	
6	678.1993	32.4	3	371.2	387.2	441.4	…	
7	999.4994	56.8	2	857.1	811.0	750.5	…	APDITEPEK
8	526.6502	46.0	3	183.6	169.0	155.2	…	
9	1105.3597	69.4	3	1130.1	1075.7	1075.1	…	
10	1292.0880	34.5	2	709.7	614.0	656.0	…	



Marker Discovery

Targeted MS/MS Identifications

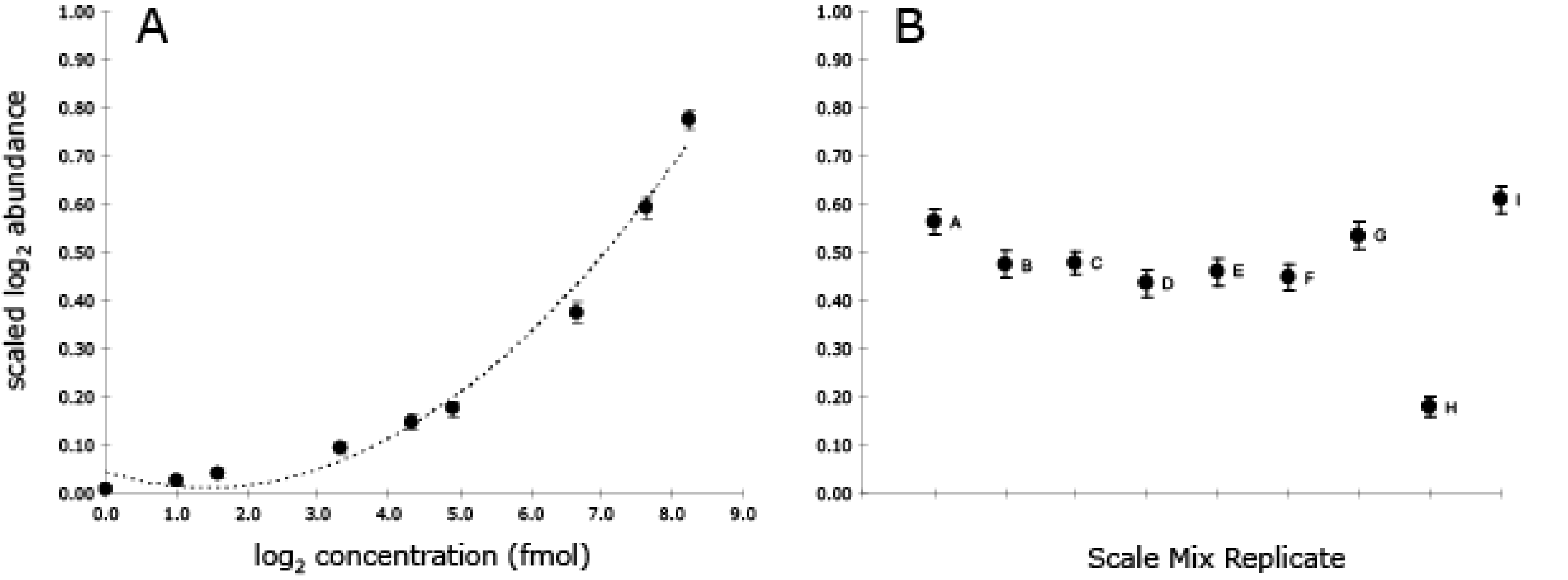
Calibration and Landmark Performance

Scale Mixture

	A	B	C	D	E	F	G	H	I
Aprotinin	1	2	3	10	20	30	100	200	300
Ribonuclease A	300	1	2	3	10	20	30	100	200
Myoglobin	200	300	1	2	3	10	20	30	100
beta-Lactoglobulin	100	200	300	1	2	3	10	20	30
alpha Casein	30	100	200	300	1	2	3	10	20
Carbonic anhydrase	20	30	100	200	300	1	2	3	10
Ovalbumin	10	20	30	100	200	300	1	2	3
Fibrinogen	3	10	20	30	100	200	300	1	2
BSA	2	3	10	20	30	100	200	300	1
Transferrin	100	100	100	100	100	100	100	100	100
Plasminogen	30	30	30	30	30	30	30	30	30
beta-Galactosidase	10	10	10	10	10	10	10	10	10

All concentrations in fmol/ul (nM)
Inject 1 ul x 5 replicates each

Peaks with IDs (avg. per run):
165 $\Rightarrow$ 281 +70%
False positive rate:
93% $p < 0.005$
100% $p < 0.05$
False negative rate:
~2%



Measurement of Ratios with Variability

Variability Mixture

	Person A		Person B		Person C		Person D		Person E	
	α	β	α	β	α	β	α	β	α	β
Aprotinin	100	5	100	5	100	5	100	5	100	5
Ribonuclease A	100	100	100	100	100	100	100	100	100	100
Myoglobin	100	100	100	100	100	100	100	100	100	100
beta-Lactoglobulin	50	1	50	1	50	1	50	1	50	1
alpha Casein	100	10	100	10	100	10	100	10	100	10
Carbonic anhydrase	100	100	100	100	100	100	100	100	100	100
Ovalbumin	5	10	5	10	5	10	5	10	5	10
Fibrinogen	25	25	25	25	25	25	25	25	25	25
BSA	200	200	200	200	200	200	200	200	200	200
Transferrin	10	5	10	5	10	5	10	5	10	5
Plasminogen	2.5	25	2.5	25	2.5	25	2.5	25	2.5	25
beta-Galactosidase	1	10	1	10	1	10	1	10	1	10

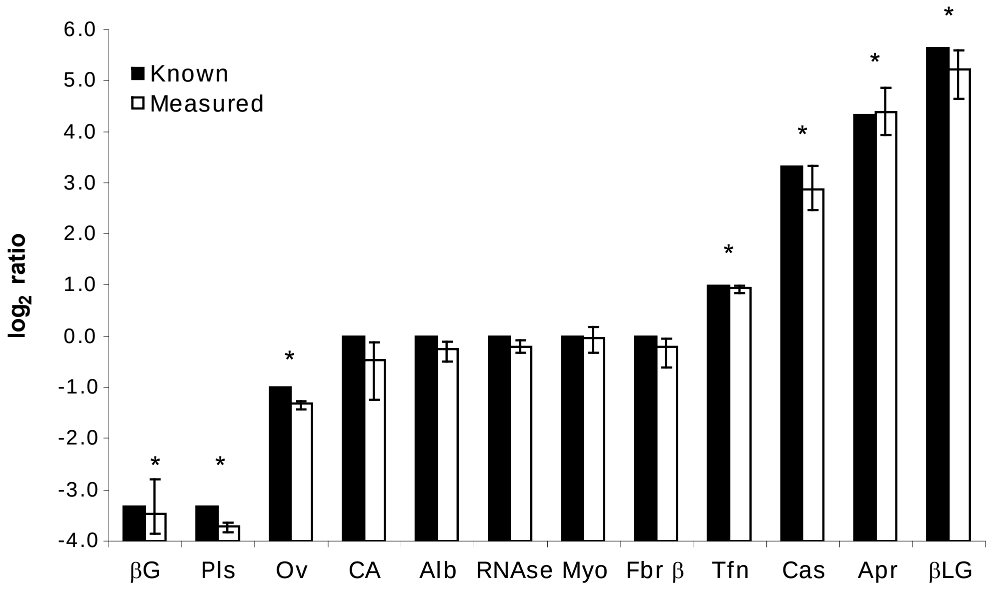
All concentrations in fmol/ul (nM)
Inject 1 ul x 5 replicates each

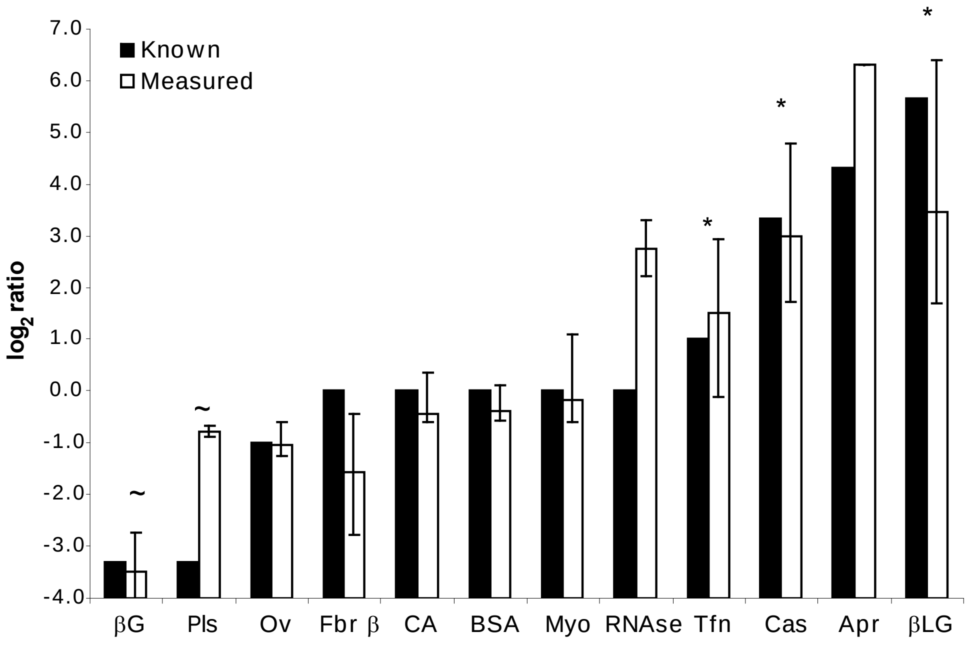
Complex Variability Mixture:

Mix α + Mitochondrial Protein from 2 wk. mouse liver

Mix β + Mitochondrial Protein from 6 wk. mouse liver

1 prep each sample, 6 injections each





PEPPeR and GenePattern

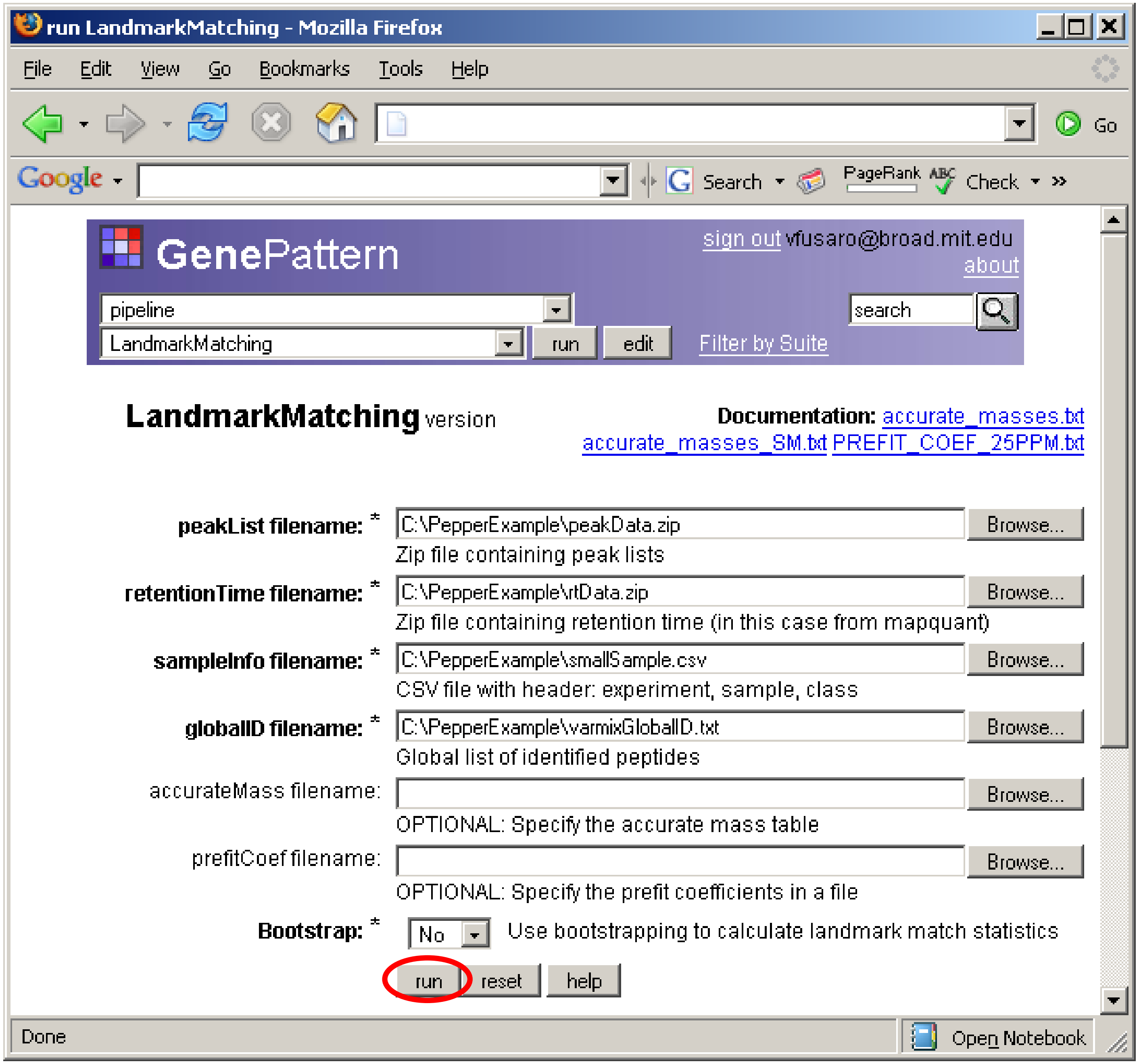
GenePattern

A platform for integrative genomics

- GenePattern is a suite of tools originally developed for microarray analysis
  - AIM: reproducible research through well-defined processing pipelines

- Many analysis modules available
  - PEPPeR: Landmark Matching and Peak Matching
  - Daisy-chainable into pipelines
  - Feed into statistical tools

PEPPeR in GenePattern



run LandmarkMatching - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Go
Google
Search
PageRank
Check
GenePattern
sign out vfusaro@broad.mit.edu
about
pipeline
search
Filter by Suite
run
edit
Done
Open Notebook

LandmarkMatching version

Documentation: accurate_masses.txt accurate_masses_SM.txt PREFIT_COEF_25PPM.txt

peakList filename: *	C:\PepperExample\peakData.zip Zip file containing peak lists	Browse...
retentionTime filename: *	C:\PepperExample\rtData.zip Zip file containing retention time (in this case from mapquant)	Browse...
sampleInfo filename: *	C:\PepperExample\smallSample.csv CSV file with header: experiment, sample, class	Browse...
globalID filename: *	C:\PepperExample\varmixGlobalID.txt Global list of identified peptides	Browse...
accurateMass filename:	 OPTIONAL: Specify the accurate mass table	Browse...
prefitCoef filename:	 OPTIONAL: Specify the prefit coefficients in a file	Browse...
Bootstrap: *	No — Use bootstrapping to calculate landmark match statistics	

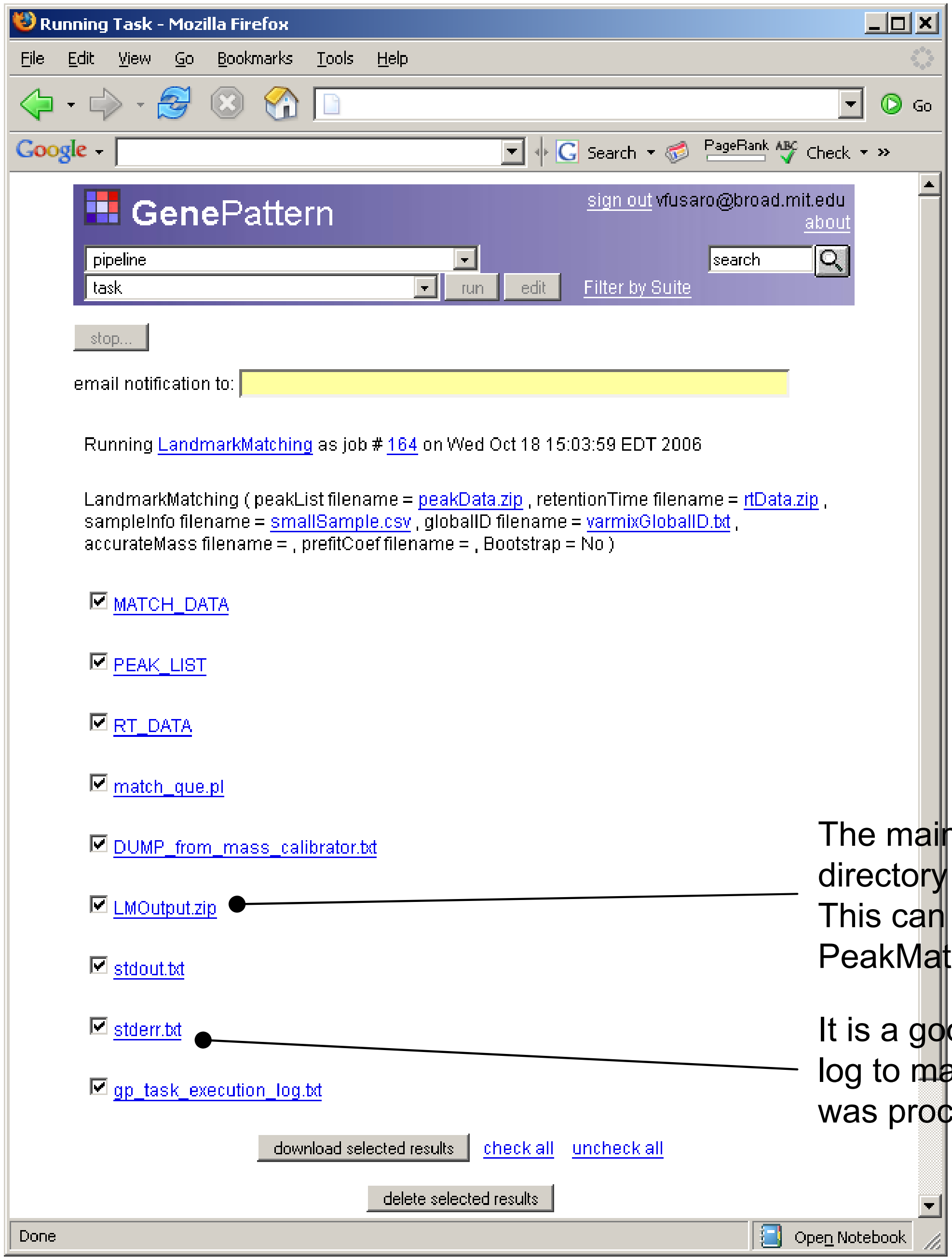
run reset help

Insert your favorite stuff here…

- Landmark Matching is platform agnostic
  - Need to get your data into a few simple flat-file formats and then zip them up together
  - Search engines i.e. SEQUEST, SpectrumMill, Mascot, etc.
  - Peak Pickers: MAPQUANT, msInspect, Decon2LS, etc.
  - Some helper apps can be found with the PEPPeR bundle on the GenePattern website

- All works via web-client interface
  - Just press go (but beware of this!)

Landmark Matching Output



vfusaro@broad.mit.edu

email notification to:

Running LandmarkMatching as job # 164 on Wed Oct 18 15:03:59 EDT 2006

LandmarkMatching (peakList filename = peakData.zip , retentionTime filename = rtData.zip , sampleInfo filename = smallSample.csv , globalID filename = varmixGlobalID.txt , accurateMass filename = , prefitCoef filename = , Bootstrap = No)

- ☑ MATCH_DATA
- ☑ PEAK_LIST
- ☑ RT_DATA
- ☑ match_que.pl
- ☑ DUMP_from_mass_calibrator.txt
- ☑ LMOutput.zip
- ☑ stdout.txt
- ☑ stderr.txt
- ☑ gp_task_execution_log.txt

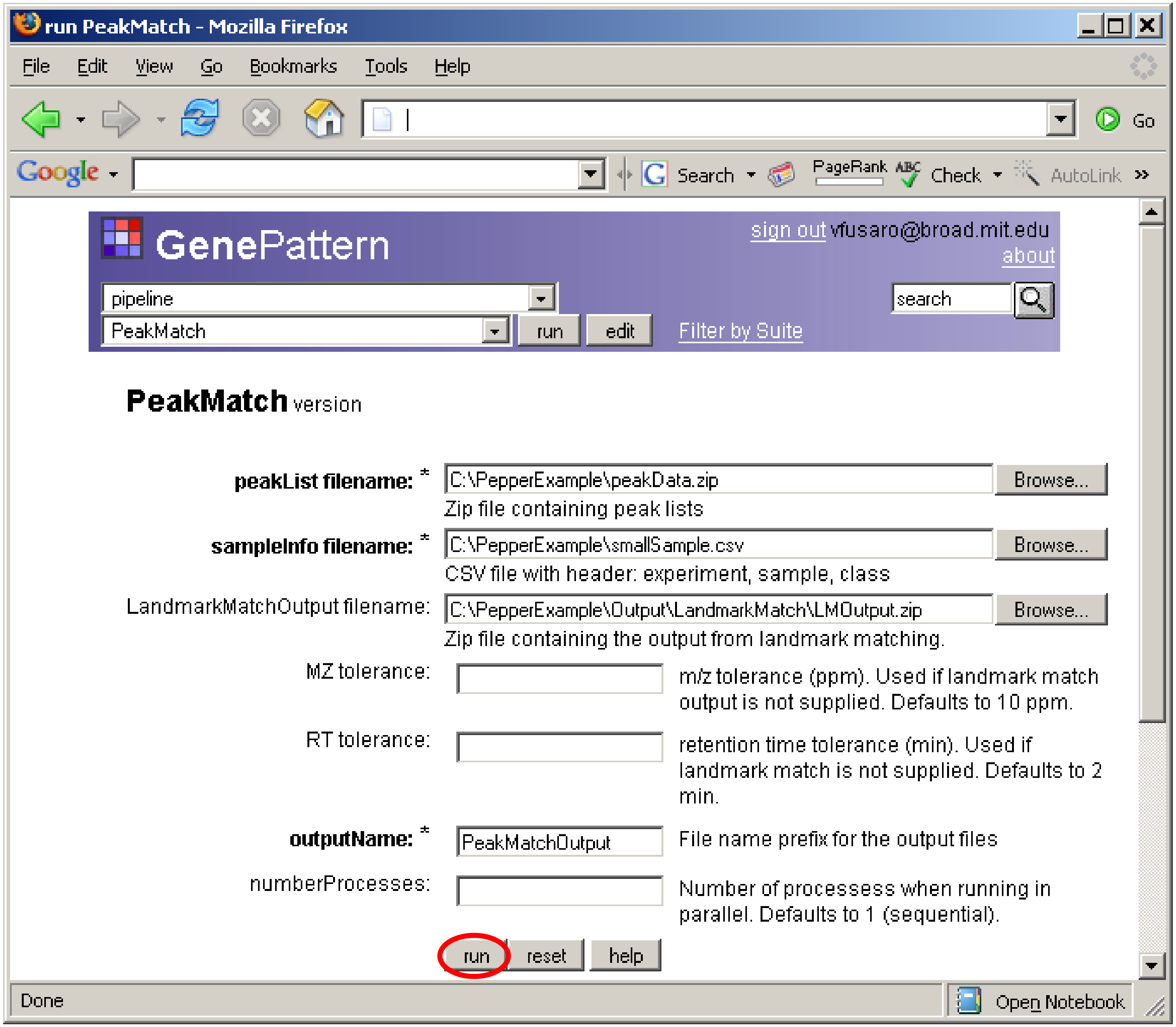
download selected results   check all   uncheck all

delete selected results

The main output is a zipped directory of all the processed files. This can be used as input into the PeakMatch module.

It is a good idea to check the error log to make sure that everything was processed correctly.

Peak Matching Interface



PeakMatch version

peakList filename: * C:\PepperExample\peakData.zip
Zip file containing peak lists

sampleInfo filename: * C:\PepperExample\smallSample.csv
CSV file with header: experiment, sample, class

LandmarkMatchOutput filename: C:\PepperExample\Output\LandmarkMatch\LMOutput.zip
Zip file containing the output from landmark matching.

MZ tolerance: m/z tolerance (ppm). Used if landmark match output is not supplied. Defaults to 10 ppm.

RT tolerance: retention time tolerance (min). Used if landmark match is not supplied. Defaults to 2 min.

outputName: * PeakMatchOutput File name prefix for the output files

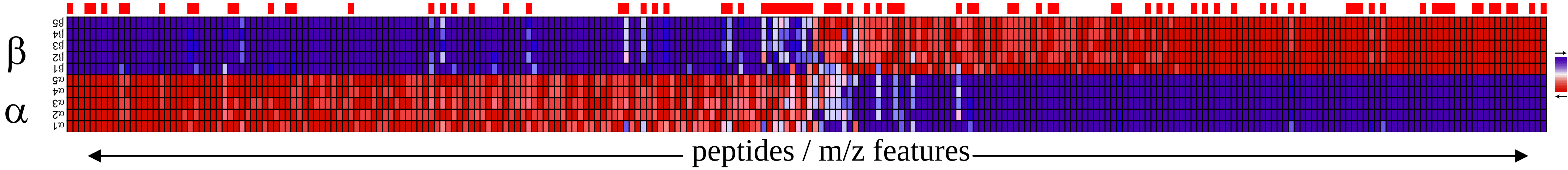
numberProcesses: Number of processess when running in parallel. Defaults to 1 (sequential).

GenePattern Downstream Tools

- **Differential analysis/marker selection**
  - Gene/Class neighbors
  - Comparative marker selection
  - Gene Set Enrichment Analysis
- **Class Prediction – supervised learning – with cross-validation**
  - Regression trees
  - K-nearest neighbors
  - Neural networks
  - Support Vector Machine
- **Class Discovery – unsupervised learning**
  - Hierarchical clustering
  - Self-organizing maps
  - Principal Component Analysis
- **Data Visualization**
  - Heat Maps, etc.

Note: Data analysis on subsequent slides done using GenePattern

Discovery of Novel Markers with PEPPeR



- Designed accurate mass ‘inclusion lists’ to hit these targets

- Confident IDs of previously identified peptides agree 100% of the time (59/59)
- 60 novel confident peptide IDs
  - 25 belong to proteins in the mix
    - 24/25 are changing
  - 35 are from proteins not designed to be in the mixture

gi Number	Species	Name
223424	*E. coli*	RNA polymerase β'
38491462	*E. coli*	GroEL
42144	*E. coli*	NusA
42818	*E. coli*	RNA polymerase β
42900	*E. coli*	Ribosomal protein S1
26249756	*E. coli*	Argininosuccinate synthase
8099322	*B. taurus*	κ-casein

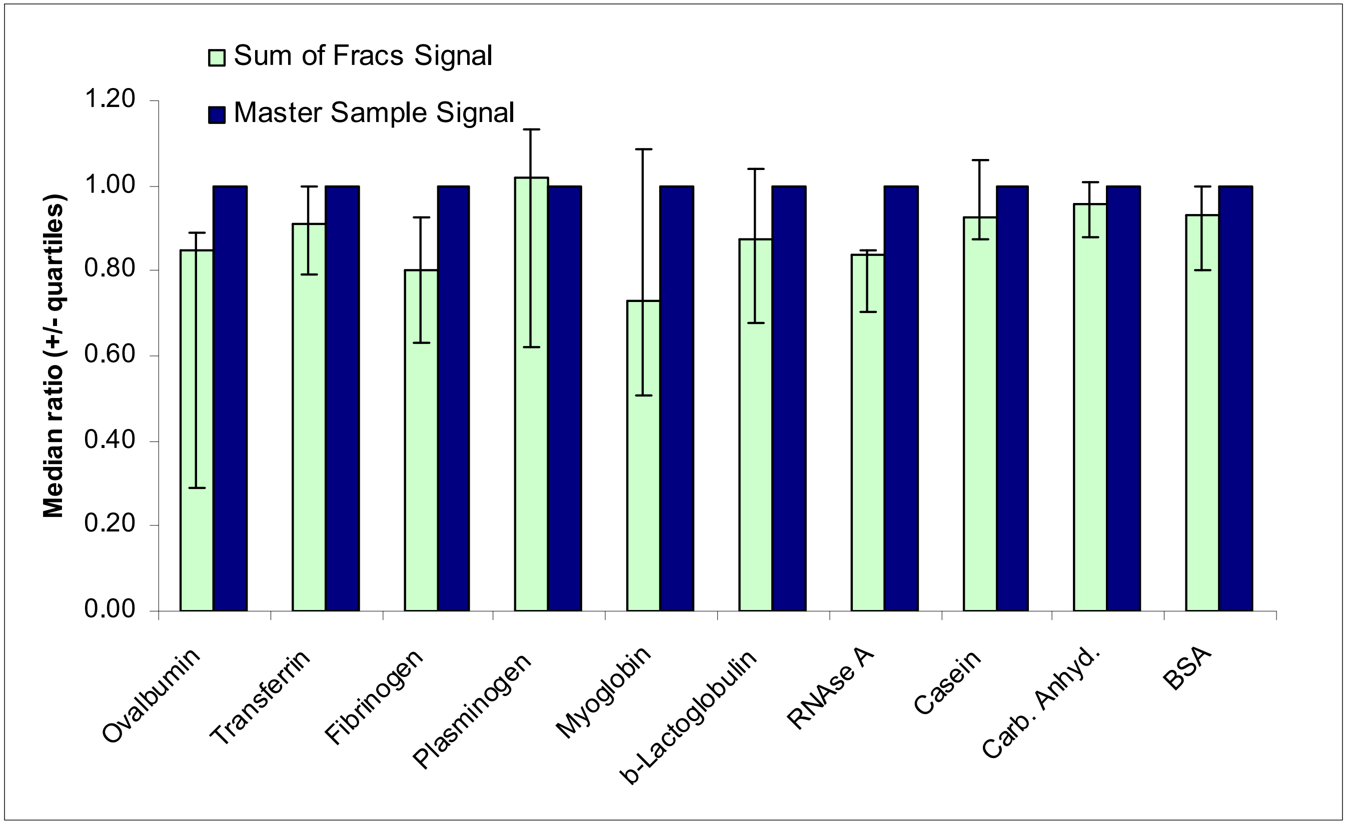
B-Galactosidase had 1:10 ratio!
Casein had 10:1 ratio!

In-silico defractionation of 2D-LC

- Wanted to mimic SCX fractionation scheme

	Frac. 1	Frac. 2	Frac. 3	Frac. 4	Frac. 5	Master
Ovalbumin	0	0	0	5	0	5
Transferrin	0	0	0	0	5	5
Fibrinogen	0	12.5	0	0	0	12.5
Plasminogen	5	5	5	5	5	25
Myoglobin	0	0	50	0	0	50
β-Lactoglobulin	0	0	12.5	25	12.5	50
RNAse A	50	10	0	0	0	60
Casein	5	50	5	0	0	60
Carb. Anhyd.	0	0	0	50	50	100
BSA	100	0	0	0	0	100

All values in fmol injected

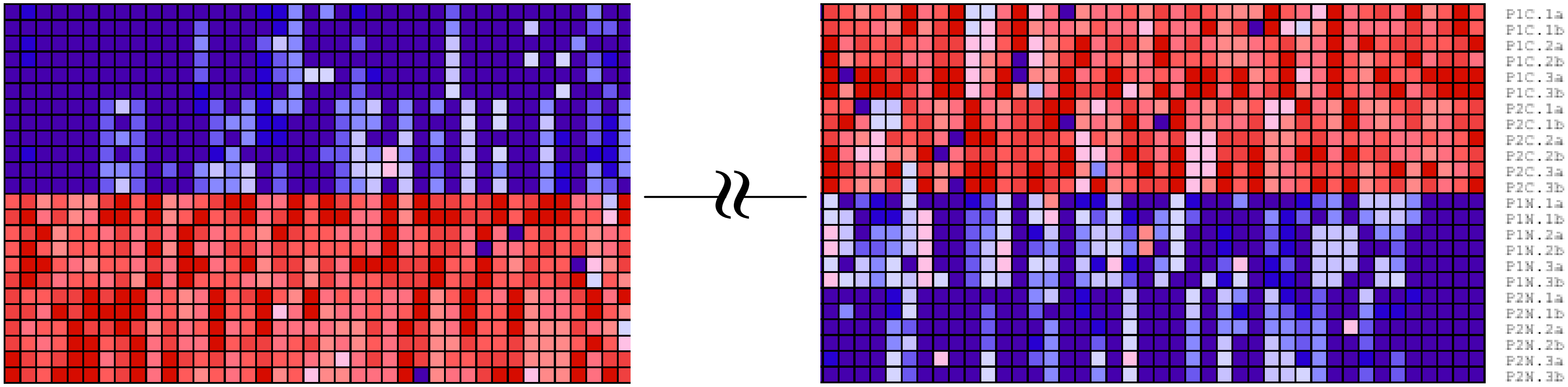


Breast Cancer Biomarker Discovery

- Sample source: nipple aspirate fluid (NAF) from malignancy affected breast
  - Unaffected contra-lateral breast used for control
  - Pools of several patient samples made <- low starting material
- Samples depleted of abundant proteins by affinity chromatography
- Separate ID-centric (fractionation) and Pattern Centric runs conducted for PEPPeR analysis
- Performed marker selection with allowed FDR of 5%

Breast Cancer Marker Selection

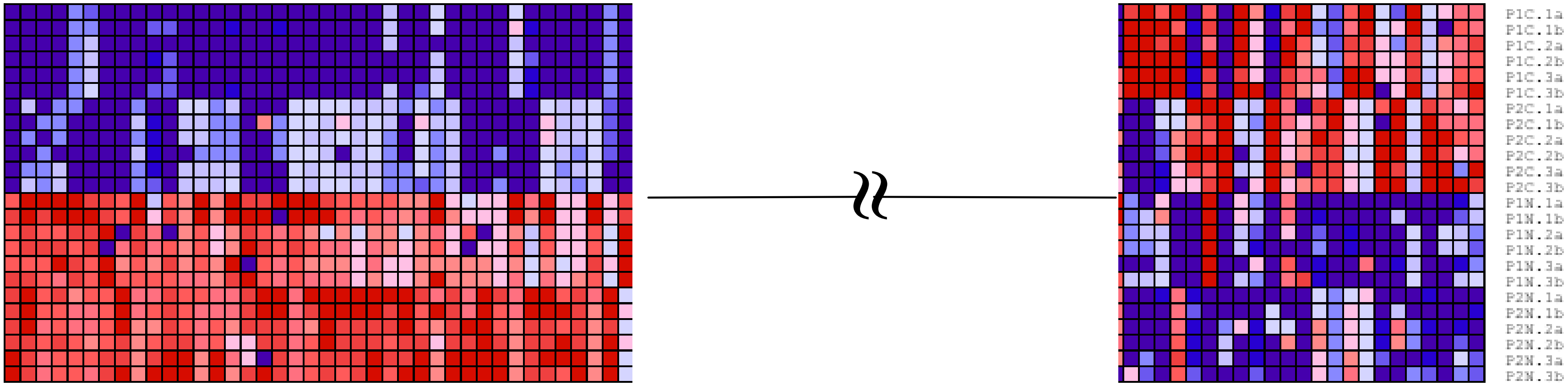
n=1520



Features down in cancer

Features up in cancer

n=264



Assigned peptides down in cancer

Assigned peptides up in cancer

Features vs. Assignments

- There's more out there than we can catalog
  - Low intensity features never trigger MS^2 in complex samples
  - Unidentified features may be better classifiers

- Direct follow-up easily achieved
  - We know exactly where and when to look
  - Targeted accurate mass methods can be employed

- Hopefully increase coverage and confidence in certain **proteins** as markers, rather than just peptides or features

Summary – what I hope you learned

- PEPPeR: Landmark Matching and Peak Matching
  - Keep track of all of those pesky peaks that you picked!

- GenePattern: A web-based tool to coordinate reproducible research

- An entrée into downstream discovery methods in an automated pipeline (more GenePattern)

- Some real world examples of its application

Acknowledgements

- Broad Proteomics:
  - Steve Carr
  - **D.R. Mani**
  - Vincent Fusaro
  - Mike Gillette

- Broad GenePattern Team:
  - **Michael Reich**
  - Josh Gould

- Church Lab, Harvard Medical School
  - George Church
  - **Kyriacos Leptos**

URLs:

- PEPPeR / GenePattern:
  - http://www.broad.mit.edu/cancer/software/genepattern/
  - http://www.broad.mit.edu/cancer/software/genepattern/desc/proteomics.html
- MAPQUANT:
  - http://arep.med.harvard.edu/MapQuant/

Live DEMO Time

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

Note: PNNL is always looking for good and knowledgeable informatics staff and post-docs (see us afterwards for more information, or visit http://jobs.pnl.gov/)

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://proteomicsresource.org/
    - National Cancer Institute
    - National Institute of General Medical Sciences
    - National Institute of Diabetes & Digestive & Kidney Diseases