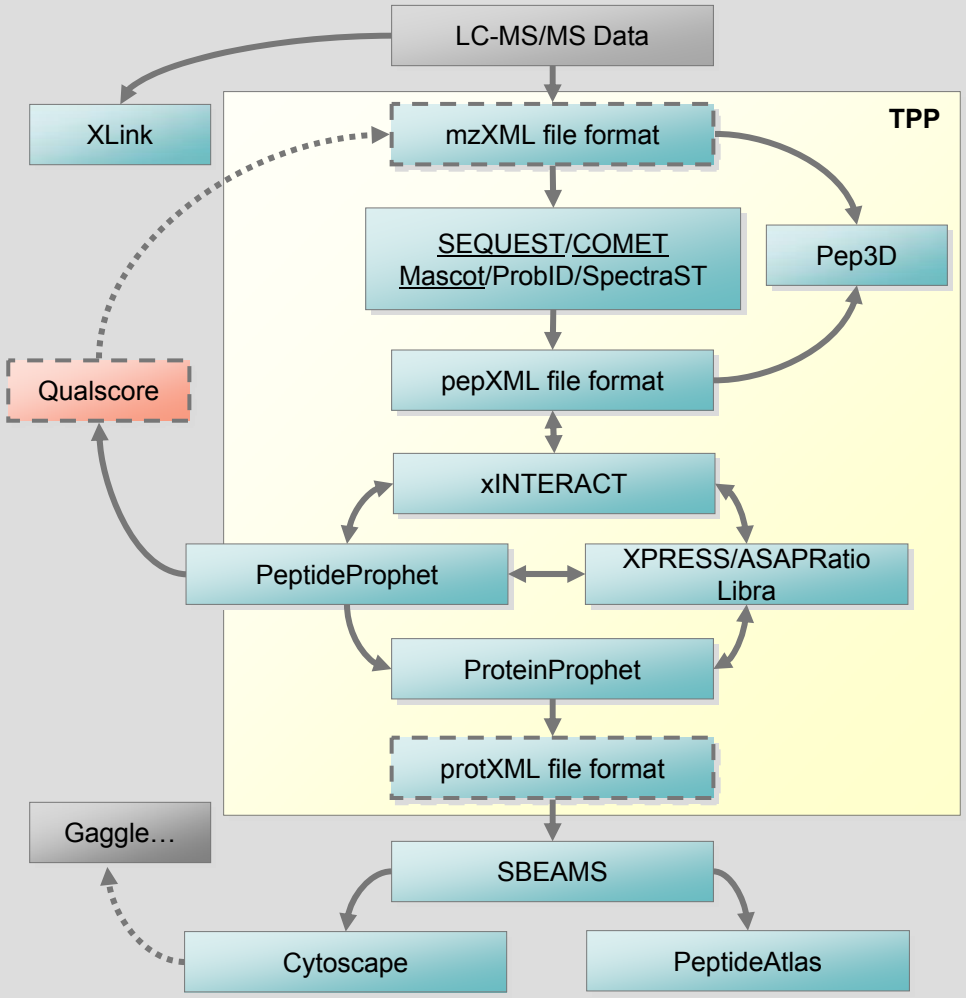
QualScore

Day 2

James Eddes

INSTITUTE FOR Systems Biology

QualScore



Nesvizhskii *et. al*. Mol Cell Proteomics. 2006 Apr;5(4):652-70. Epub 2005 Dec 12

Fraction of Spectra Left Unassigned in a Typical Search

Typical MS/MS search

- SEQUEST, IPI database
- semi-constrained (tryptic on one end)
- Met + 16
- +/- 3 Da, average mass

Average numbers (mix of ICAT/non-ICAT experiments):

- 10-15% of all spectra assigned peptide with high confidence
- 20-25 % of all high quality spectra are not assigned

What are these spectra?

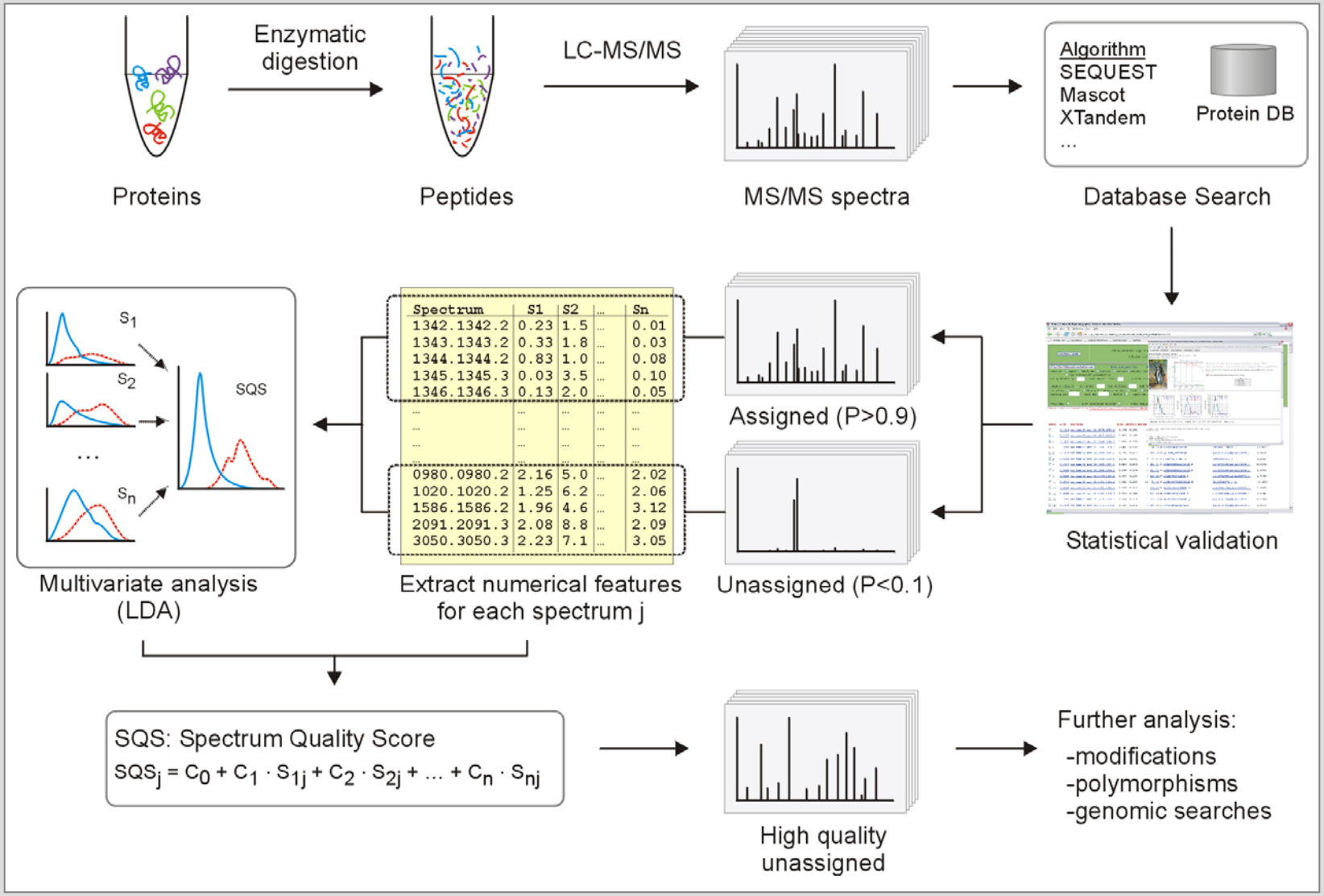
Biologically interesting peptides or boring modifications?

Why unassigned?

Possible causes of failure to assign peptide:

- Imperfect scoring scheme
- Constrained search (PTM, not tryptic etc.)
- Incorrect mass/ charge state
- Low spectrum quality / contaminant ion
- Correct sequence may not be in the database searched (e.g., SNP)
- Novel sequence (splice variants)

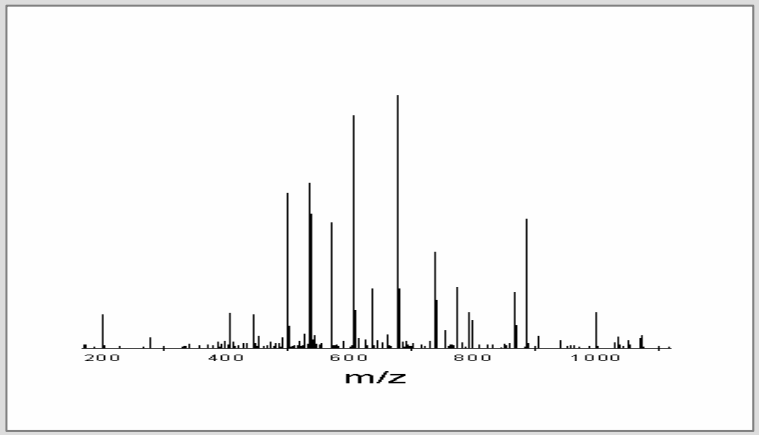
Finding and Mining High Quality Unassigned Spectra



Spectrum Features

- General features
  - number of peaks in the spectrum, distribution of peak intensities, etc.
- Sequence tags
  - length of the longest sequence tag extracted using *de novo* type algorithm.
- Complementary peaks
  - number of *y*/*b* fragment pairs summing up to the precursor ion mass.
- Neutral losses
  - loss of water etc.

QualScore



n spectral features

$x_1, x_2 \ldots x_n$

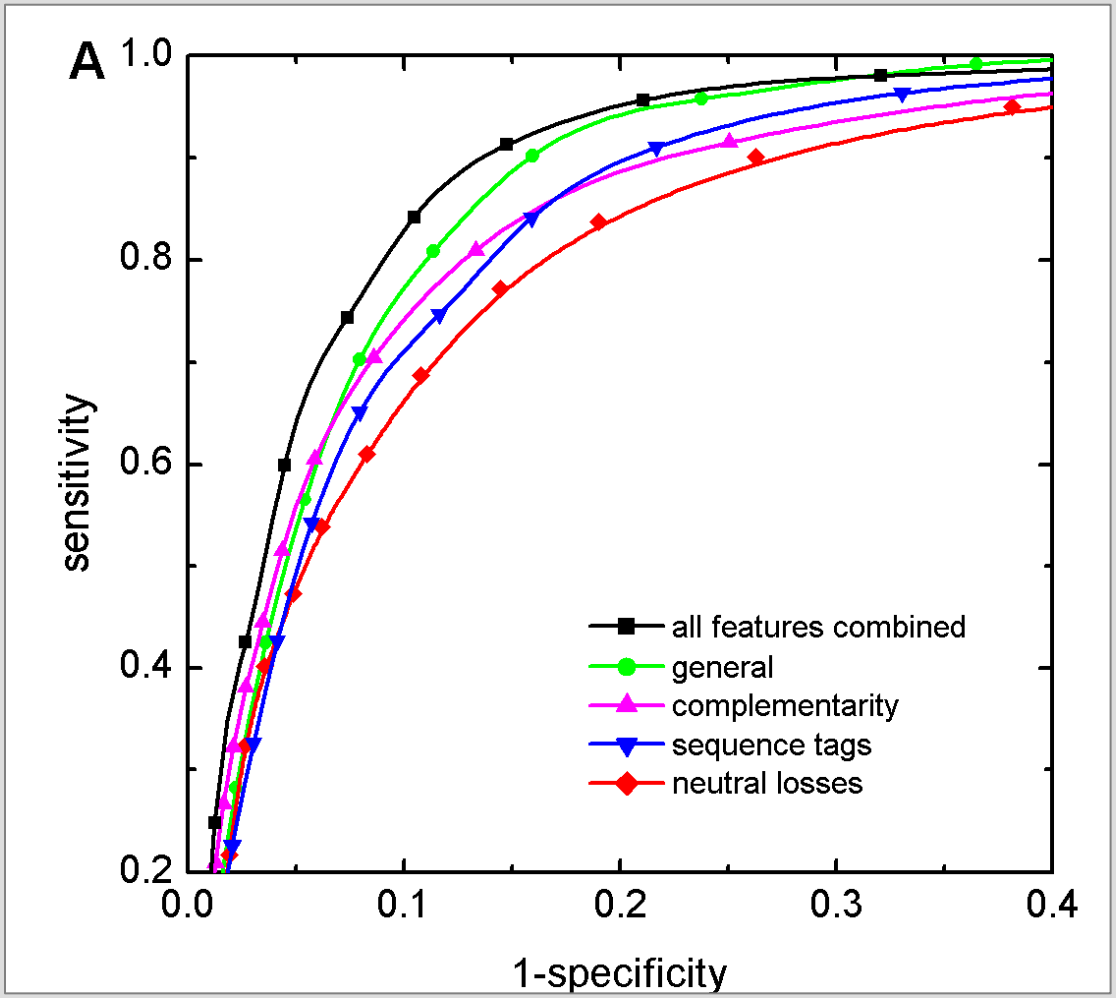
Composite score (QualScore):

$$QS = c_1x_1 + c_2x_2 + c_3x_3 + \ldots + c_nx_n$$

Linear discriminant function approach

- Coefficients c_i indicate statistical significance of each spectral feature x_i for spectral classification
- Dynamically trained (robustness)

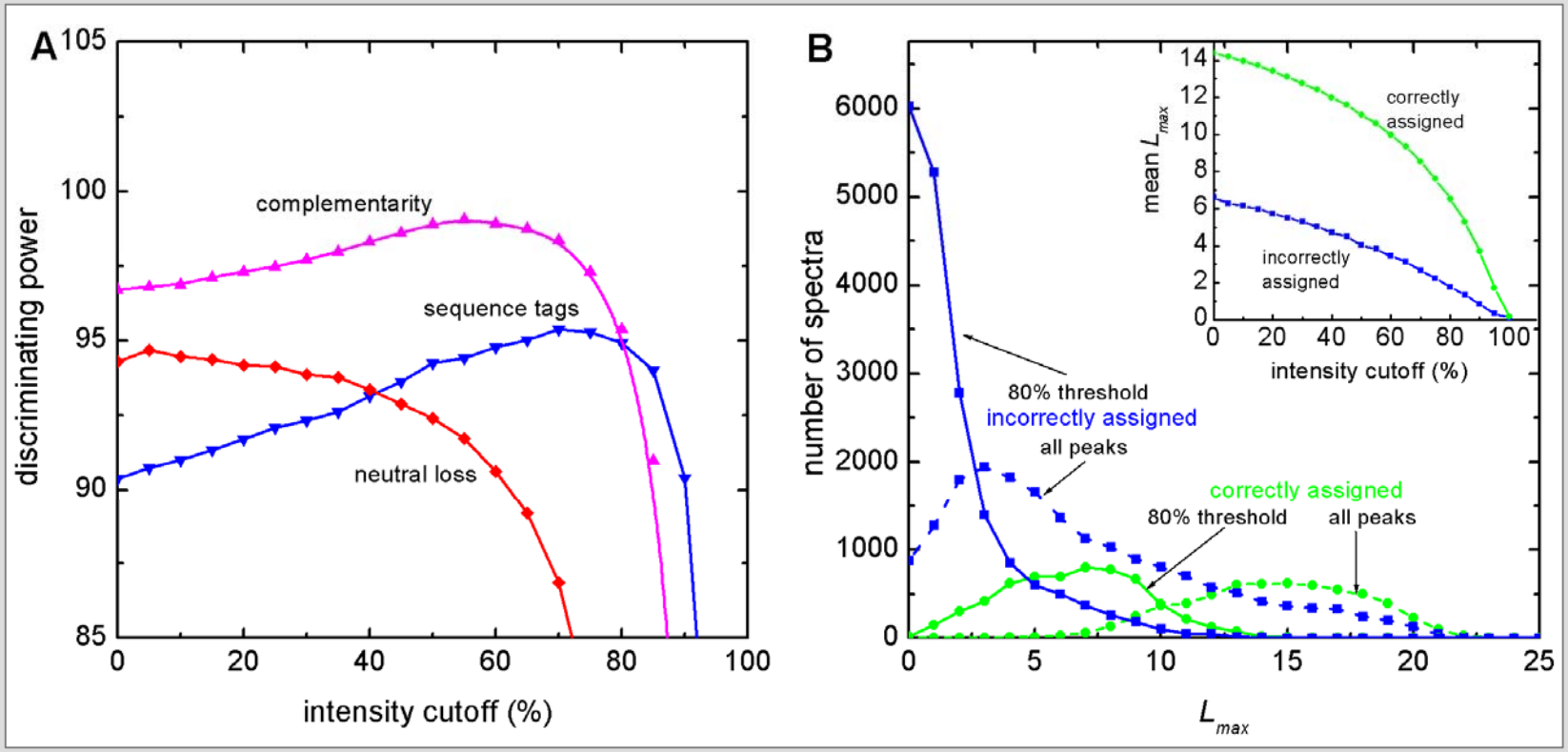
Statistical Significance of Spectrum Features



ROC curves

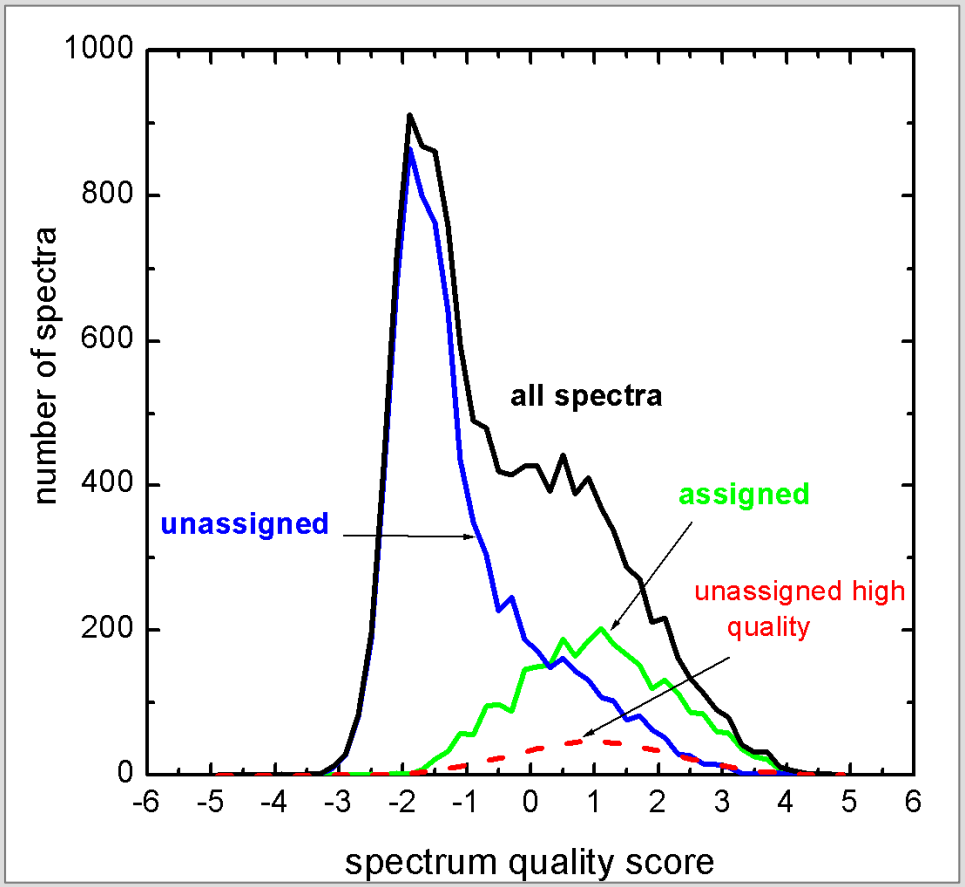
- Combining different classes of features improves performance of the classifier
- Individually: general spectrum features are best for filtering out bad quality spectra
- Complementary peaks are best for finding high quality spectra

Optimization of Spectrum Features



- Complementary ion pairs, sequence tags and neutral losses are computed using high intensity peaks only
- Optimal signal / noise threshold is different for each class

Human Raft Dataset



SEQUEST ('ISB typical')
PeptideProphet

↓

Create training dataset:
unassigned: P < 0.1
assigned: P > 0.9

↓

Dynamically train Classifier

↓

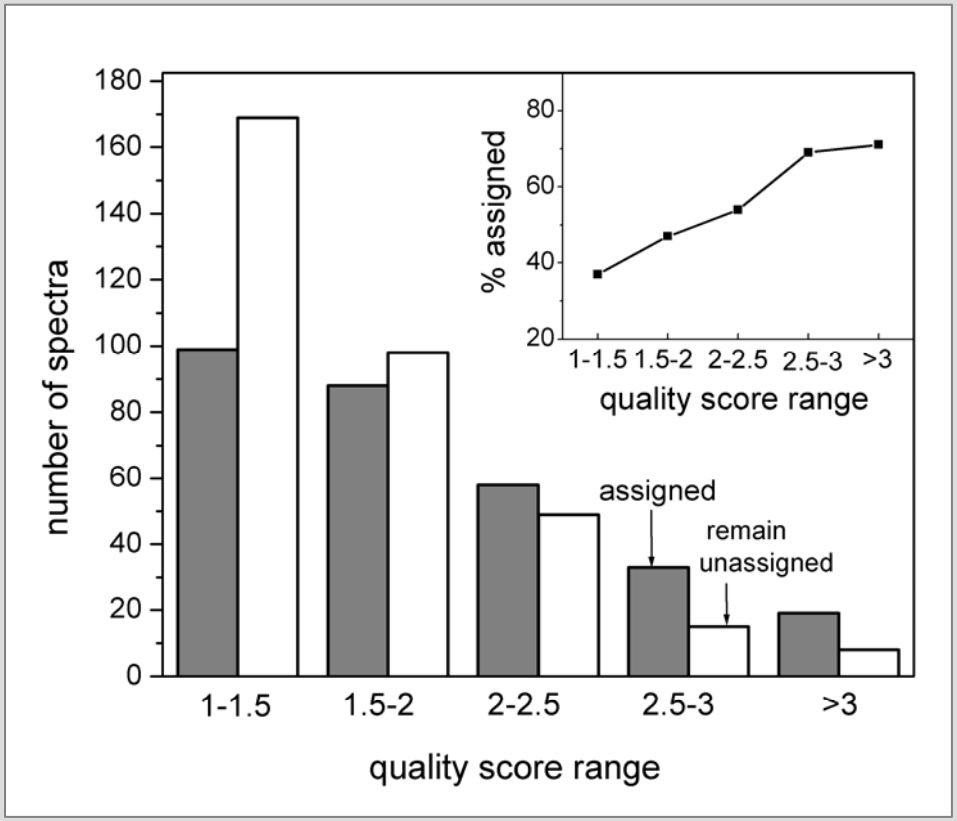
Apply to all spectra in the dataset; compute quality score for each spectrum

Potentially ~25% gain in the number of high confidence id's

Reanalysis of Unassigned High Quality Spectra

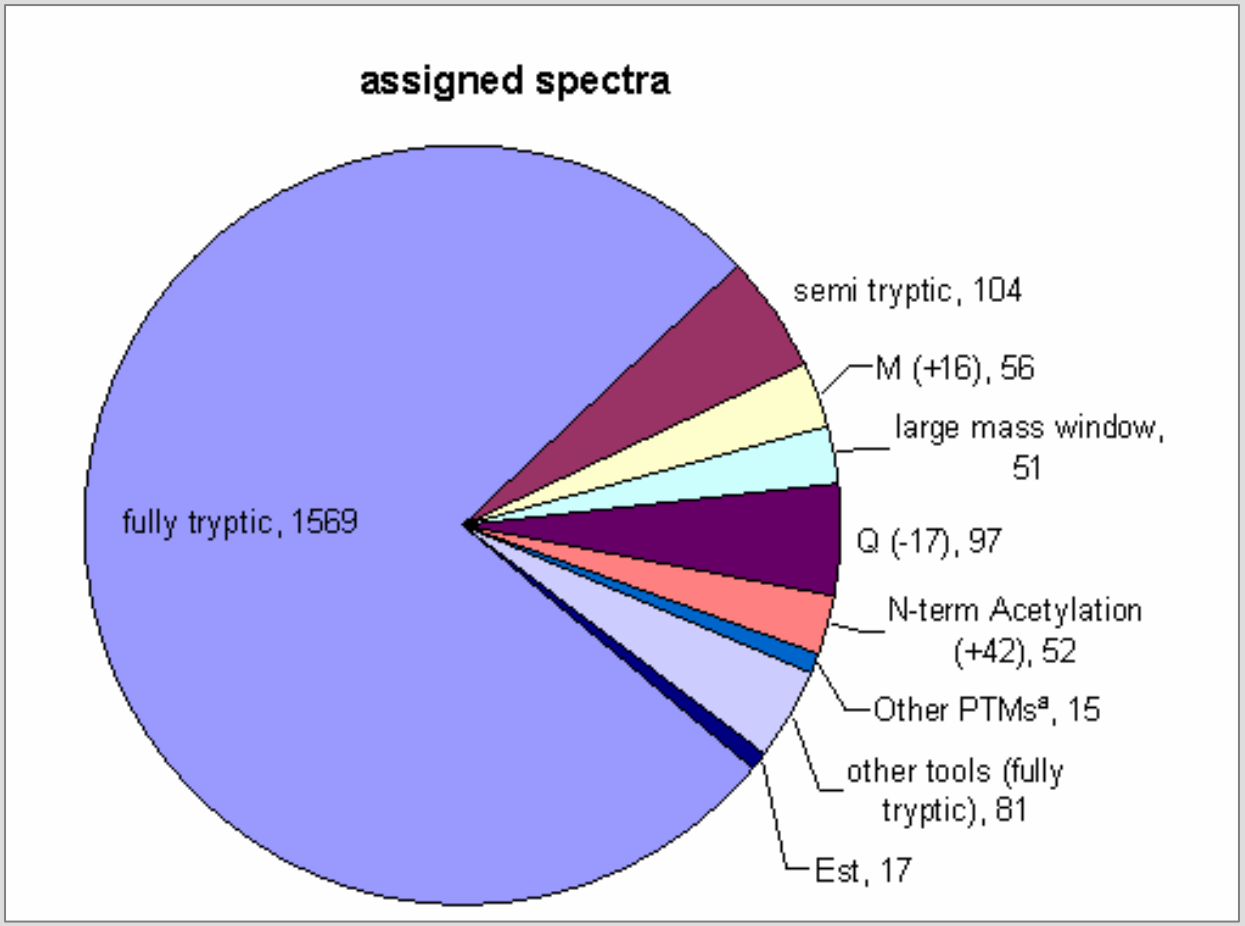
- Large mass tolerance search
  - SEQUEST, semi-tryptic, 4Da mass tolerance (previously 3Da).
- Q -17 search
  - SEQUEST, semi-tryptic, 3Da mass tolerance, allowing for conversion of glutamic residues to pyroglutamic acid (loss of 17 Da) as a variable PTM.
- Mascot search
  - Mascot, tryptic peptides only, 2 missed cleavages or less, 3Da mass tolerance, allowing for N-terminal acetylation as a variable PTM.
- Miscellaneous searches
  - XTandem with more than one type of PTM per peptide
  - SEQUEST and Mascot allowing for PTMs not specified in the previous searches (e.g., conversion of N-terminal glutamic acid residues to pyroglutamic acid, phosporylation, acetylation, guanidation, and etc.).
- EST database search

Percent of Previously Unassigned Spectra Assigned After All Additional Searchers

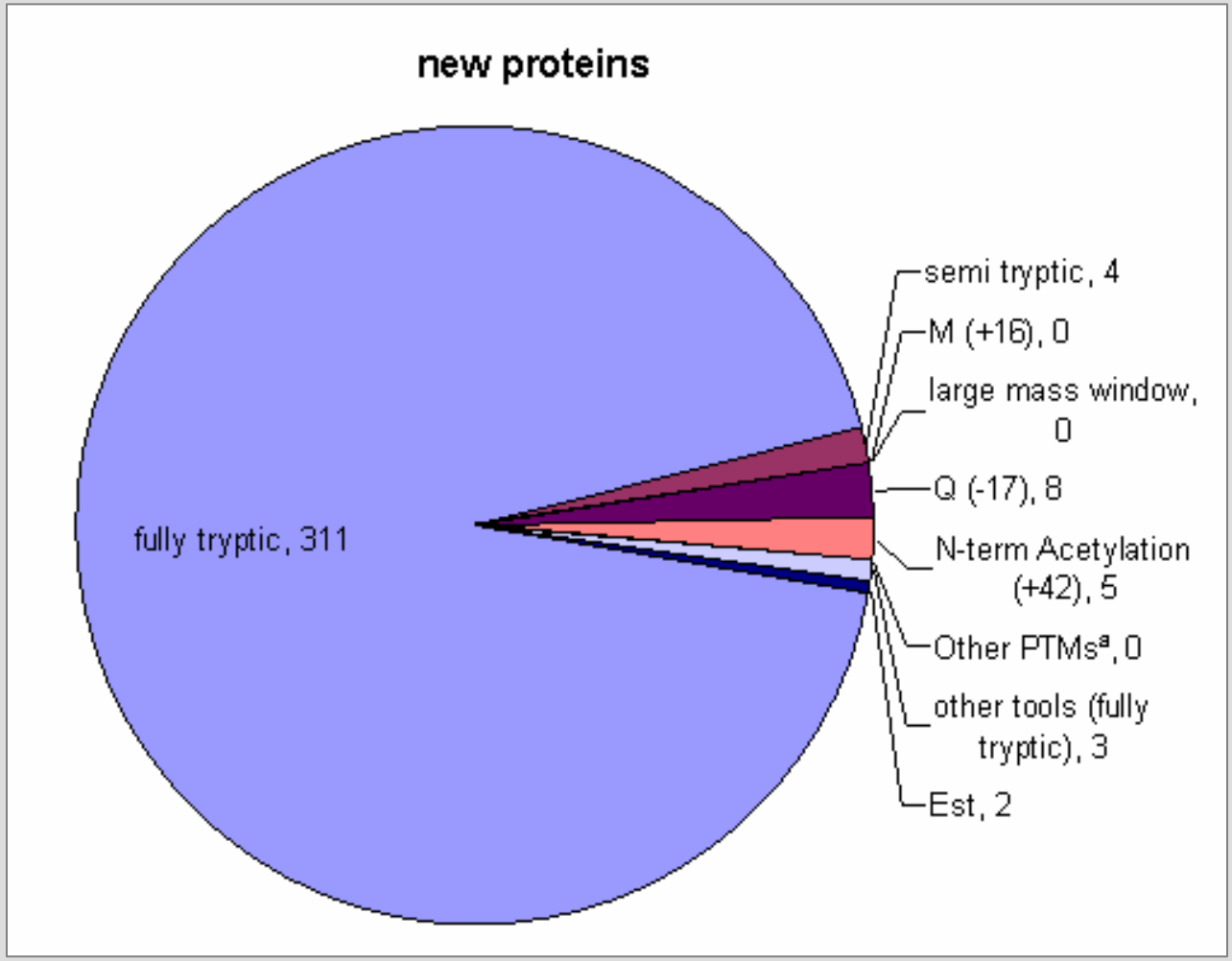


- The higher the spectrum quality, the more likely the spectrum is assigned with high confidence.
- Spectra of very high quality (QS>3) were unassigned in the initial search.
- More than 70% of these were eventually assigned

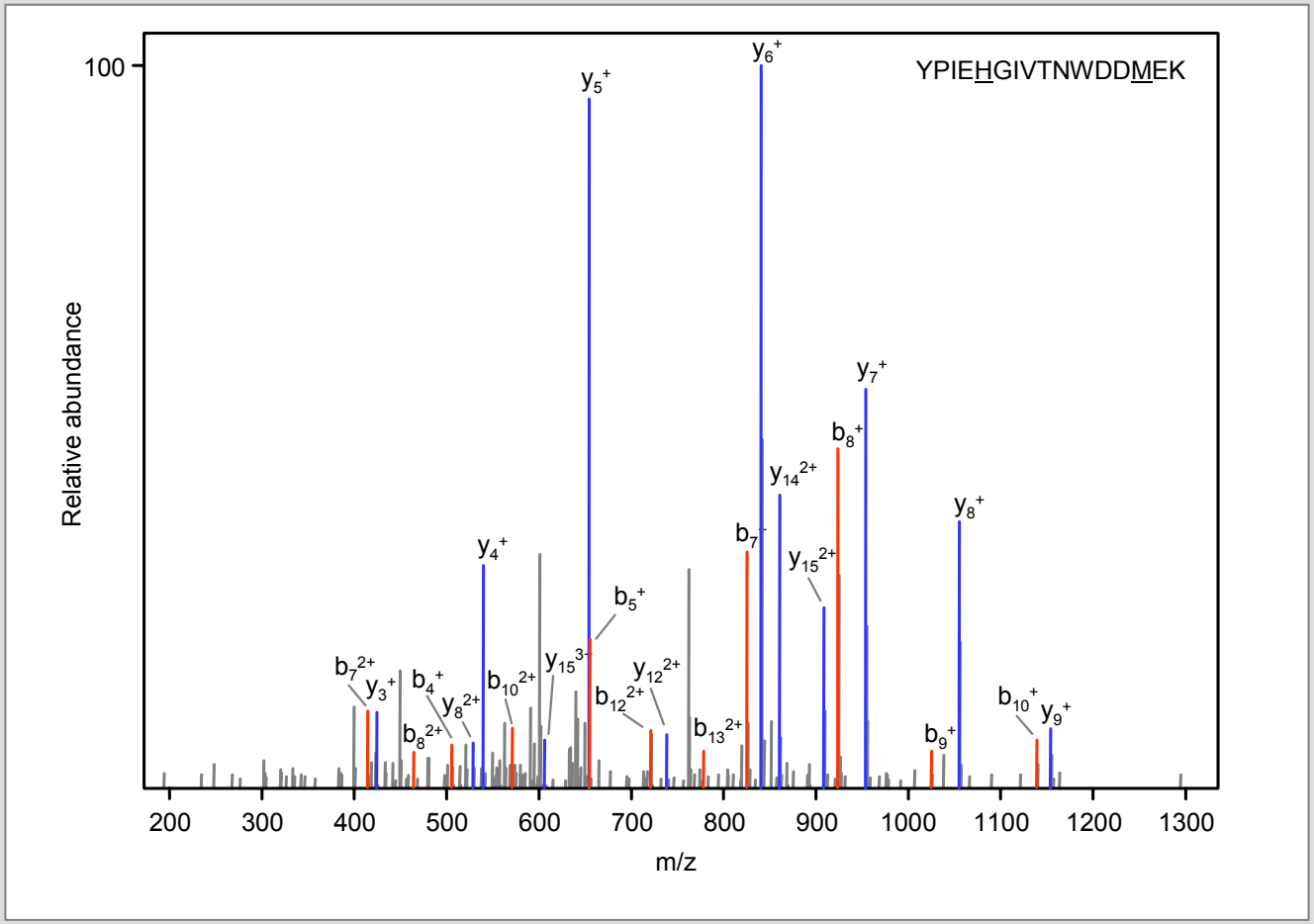
What Are Those Additional Identifications?



Do They Add Any New Proteins?

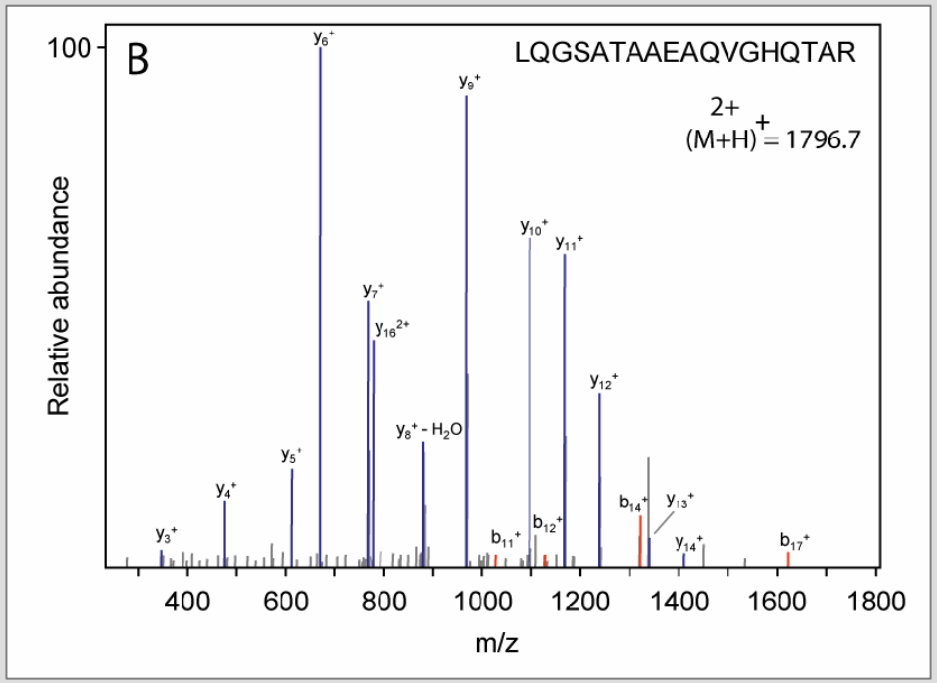


Any Biologically Interesting Peptides/Proteins?



YPIEHGIVTNWDDMEK from Actin, cytoplasmic 1 protein (SW:P02570) containing methylated histidine at position 5

Searching Genomic Databases

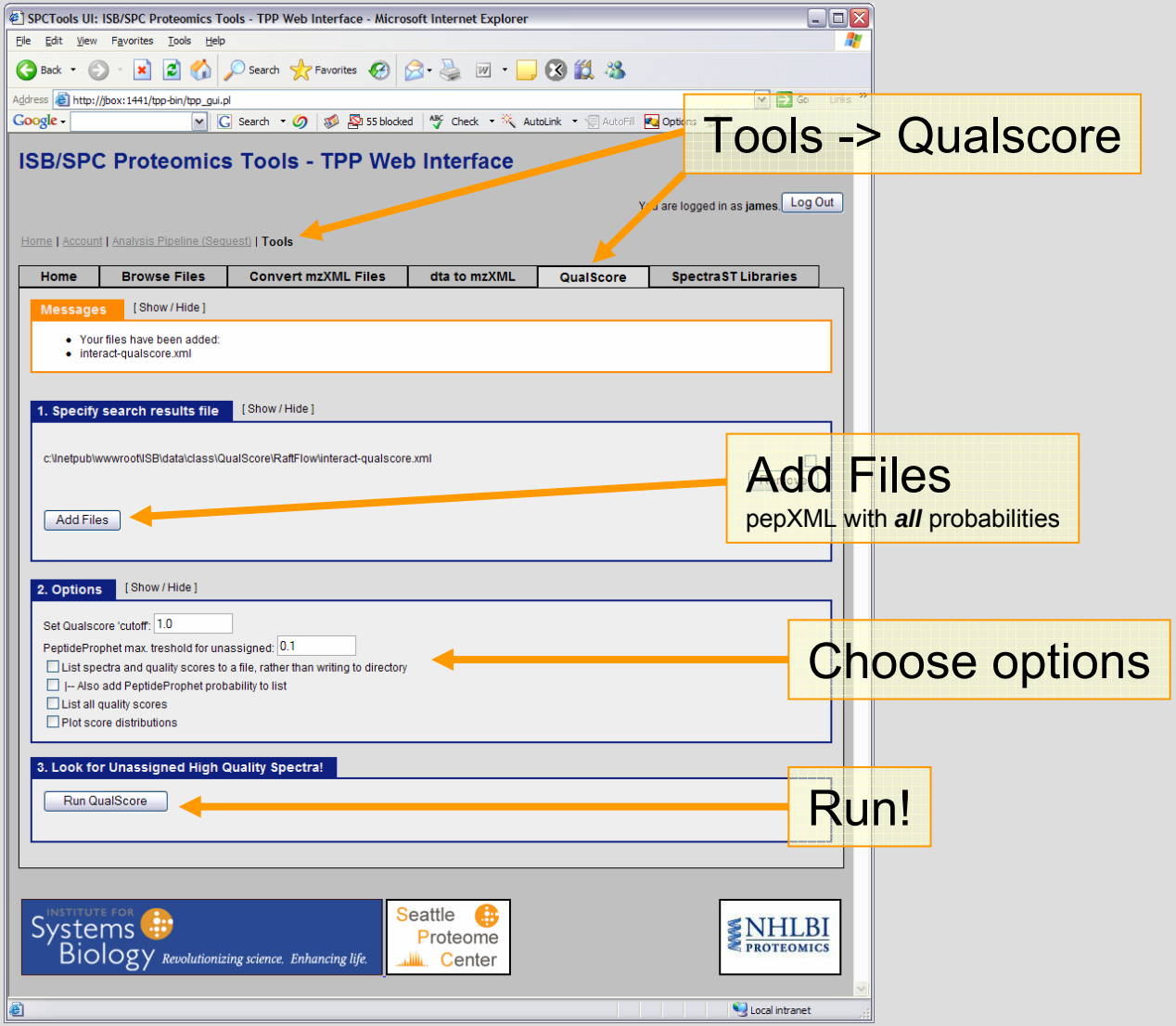


- Human lipid rafts
- Search against EST database

LQGSATAAEAQVGHQTAR (>10 EST sequences)

This intron-exon spanning peptide identifies a novel splice variant of the Lck-interacting transmembrane adaptor 1 protein (LIME1, NP_060276). LIME1 was shown to be a raft-associated protein in several recent studies.

Running QualScore



Acknowledgements

Alexey Nesvizhskii

Mathijs Vogelzang
Franz Roos
Jonas Grossmann
Sasha Baginsky

Nichole King

Ruedi Aebersold