


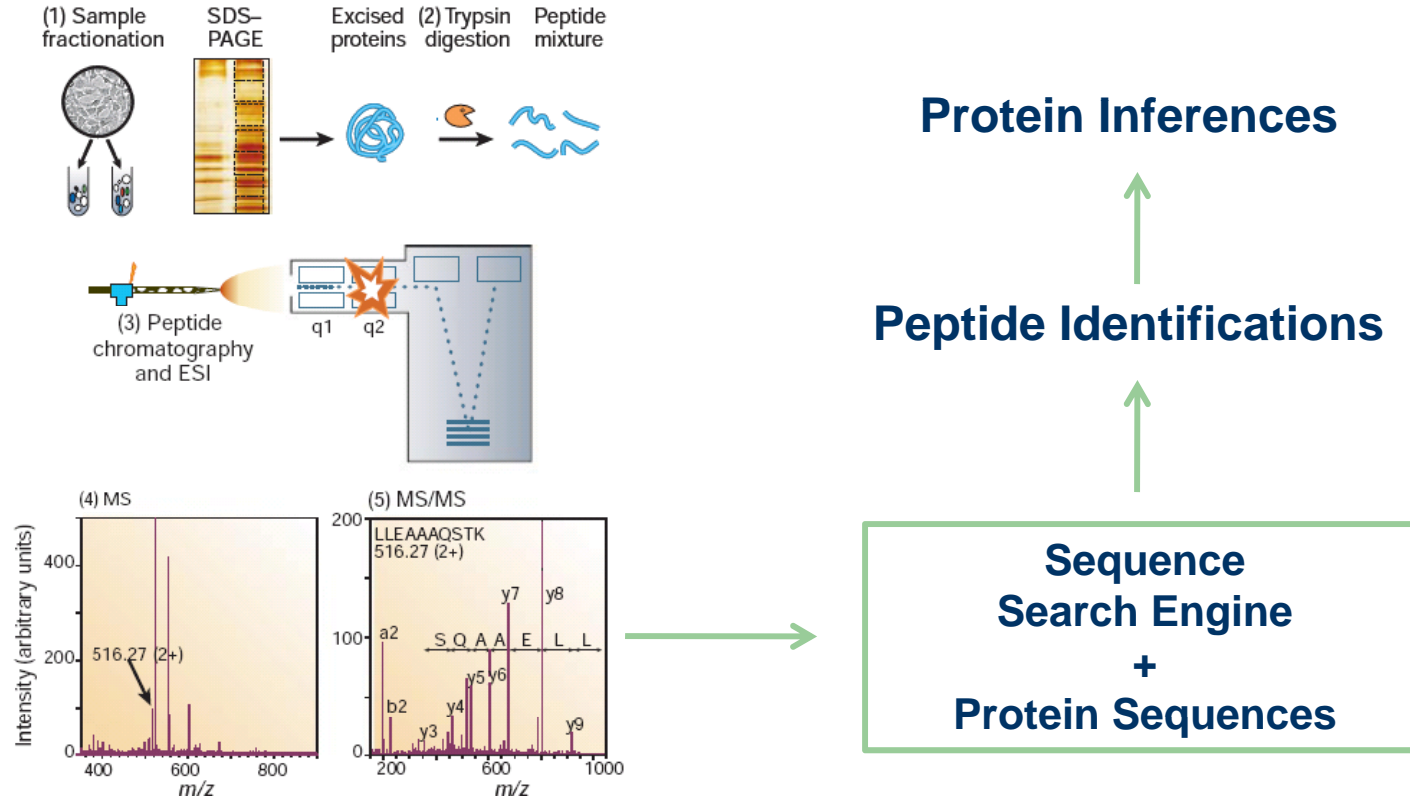
Comparison of Results from Theoretical Sequence Search Engines and Peptide Mass Spectral Libraries for Selected Biological Samples

Jeri Roth, Paul Rudnick, Qian Dong, Yuri Mirokhin,
Dmitrii Tchekhovskoi, Niksa Blonder and
Stephen E. Stein

Chemical and Biochemical Reference Data Division
National Institute of Standards and Technology

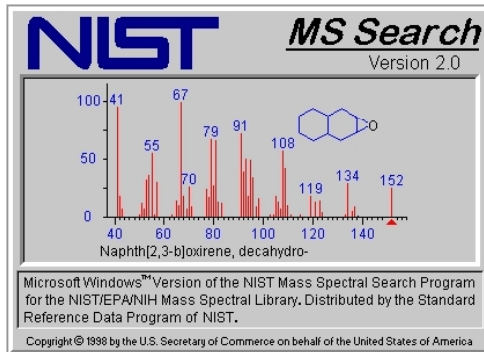


Shotgun Proteomics Workflow



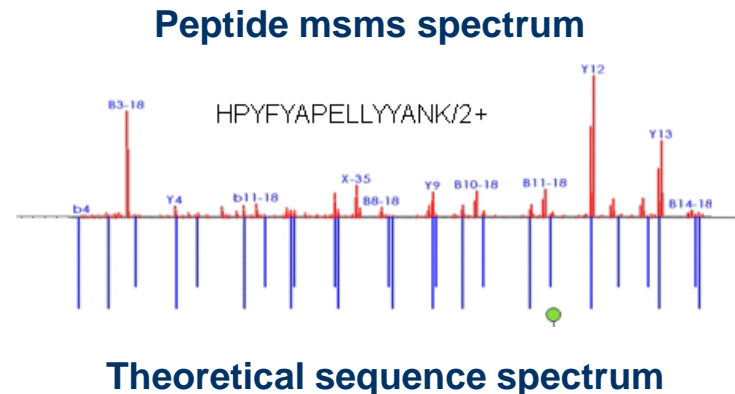
(from Aebersold & Mann, Nature 2003 422(6928):198-207)

Library Search vs Database Search



- Established method small molecules EI-MS
- Faster – smaller search space
- More accurate - includes ion intensities

- Can't find previously unidentified peptides
- Requires extensive data collection

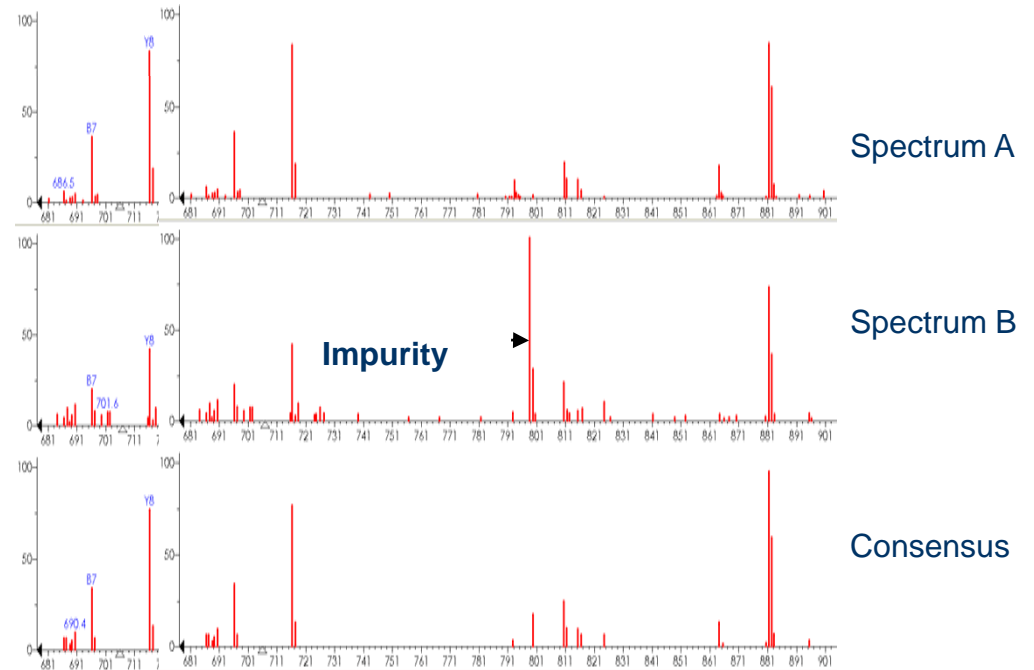


Overview of Library Creation

- Acquire raw proteomics data files from diverse sources – internal and external
- Convert to mgf format, organize and annotate files
- Identify peptides with several available sequence search engines
- Annotate the peaks and apply quality filters
- For each peptide ion, create a ‘consensus spectrum’ from replicate spectra
- Create library and perform quality testing

Consensus Spectrum

- Rejects spurious peaks rather than averaging peak abundances.
- Outlier spectra are rejected by clustering.
- Minimum spectra required per consensus
current library - 2
2011 library - 3



R.AALPEDVNAPSGEAA.- /2+

NIST Libraries of Peptide Tandem Mass Spectra <http://peptide.nist.gov/>

| Library Name | Species | Num. Spectra | Theoretical Sequence Coverage |
|-----------------------|------------------------|--------------|-------------------------------|
| <i>C. elegans</i> | <i>C. elegans</i> | 98,151 | 11% |
| <i>D. radiodurans</i> | <i>D. radiodurans</i> | 11,913 | 11% |
| Drosophila | <i>D. melanogaster</i> | 113,877 | 11% |
| <i>E. coli</i> | <i>E. coli</i> | 58,967 | 25% |
| Human | <i>H. sapiens</i> | 345,489 | 21% |
| Human (qtof) | <i>H. sapiens</i> | 14,827 | 2% |
| Mouse | <i>M. musculus</i> | 156,495 | 12% |
| Rat | <i>R. norvegicus</i> | 58,705 | 3% |
| Yeast | <i>S. cerevisiae</i> | 90,507 | 22% |
| Yeast (qtof) | <i>S. cerevisiae</i> | 2,418 | 1% |
| Sigma UPS1 | protein standards mix | 3,542 | 89% |

Experimental Conditions

Human Colon

15 human colon samples from VUMC, digested with trypsin, alkylated with iodoacetamide, analyzed on LTQ Orbitrap, 138,803 ms² spectra

Sequence search engines – Mascot and OMSSA

Fasta – human from uniprot combined with reversed fasta

Search Parameters - trypsin, 2 missed cleavages, fixed mod – carbamidomethyl (C), variable mod – oxidation (M), precursor tolerance ± 1.6 Da, msms tolerance - ± 0.8 Da, charges +1-+3 and monoisotopic precursor.

Library search – MSPepsearch

Library was created from searches using the uniprot fasta above and filtered to include only tryptic peptides with charges +1-+3 and including only no modifications, carbamidomethyl (C) or oxidation (M)

Search Parameters – precursor tolerance ± 1.6 Da, msms tolerance ± 0.8 Da and monoisotopic precursor.

Estimating Error Levels Using Target Decoy Searches

Create a decoy database to be used together with real or target sequences when running a sequence search engine.

At a given score, assume all decoy peptide hits are false and that there are an equivalent number of false hits in the real sequence files.

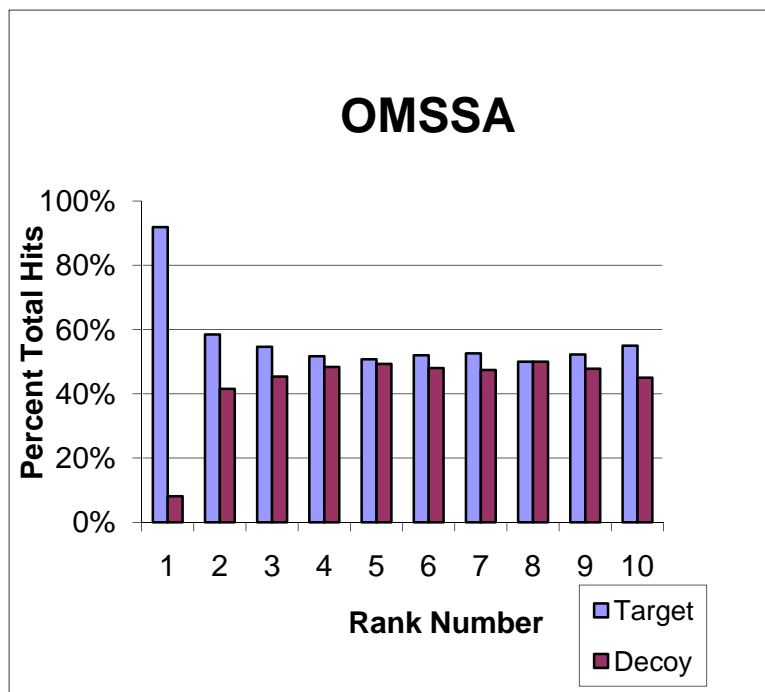
$$\text{False Discovery Rate (FDR)} \quad \frac{\text{FP}}{(\text{TP} + \text{FP})} \quad \frac{2 * \text{D}}{(\text{T} + \text{D})}$$

Ideal Decoy Characteristics

Validating Reversed Fasta Decoy Human Colon Samples

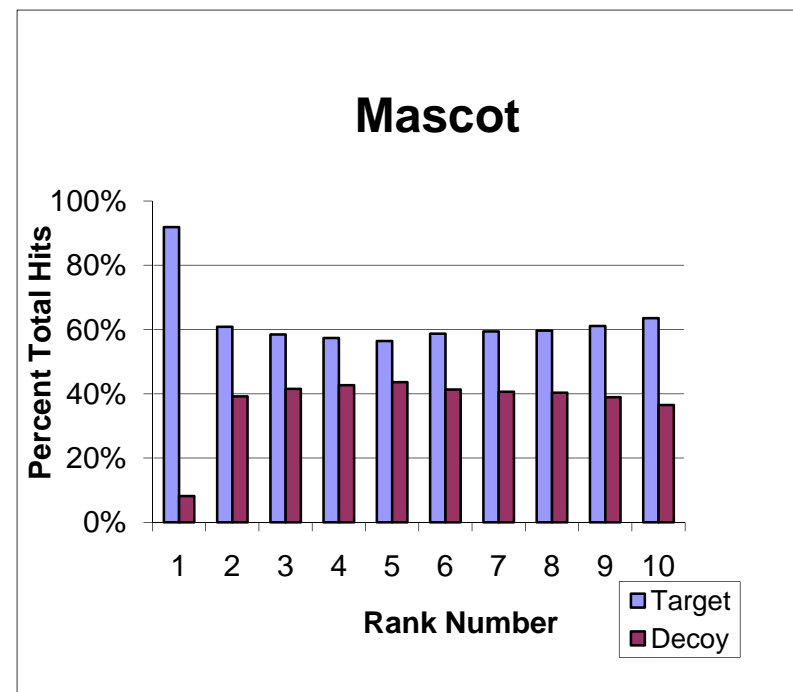
$$\frac{FP}{(TP + FP)}$$

$r = 1.1$

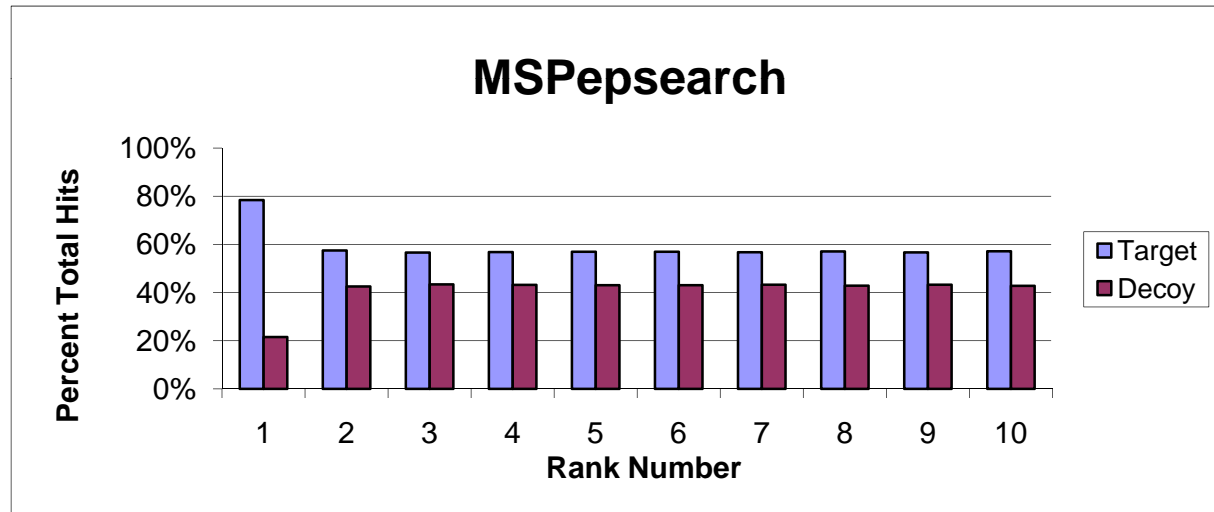


$$\frac{2 * r * D}{(T + r * D)}$$

$r = 1.46$



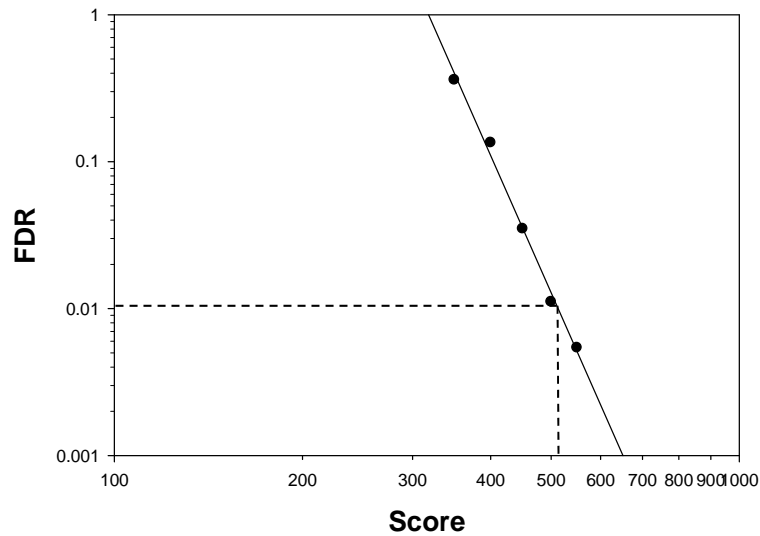
Validating Distracting Species Decoy Human Colon Samples



- Decoy created from filtered consensus spectra from *c elegans*, *drosophila*, *e-coli*, *myco smegmatis*, rat and yeast.
- Mean target/decoy ratio from ranks 3-10, $r = 1.32$.

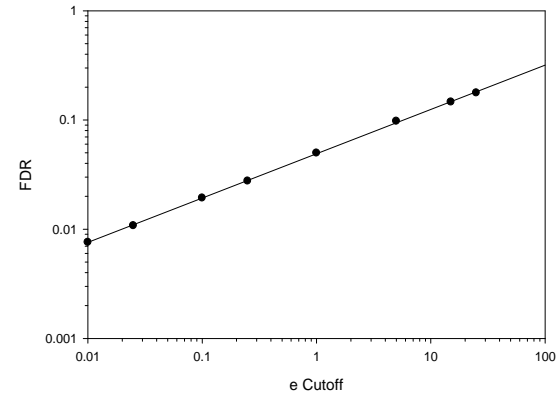
Score vs FDR Human Colon Samples

MSPepsearch

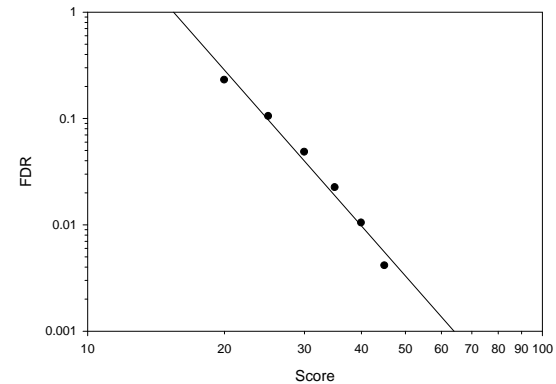


Score = 513 at FDR .01

OMSSA

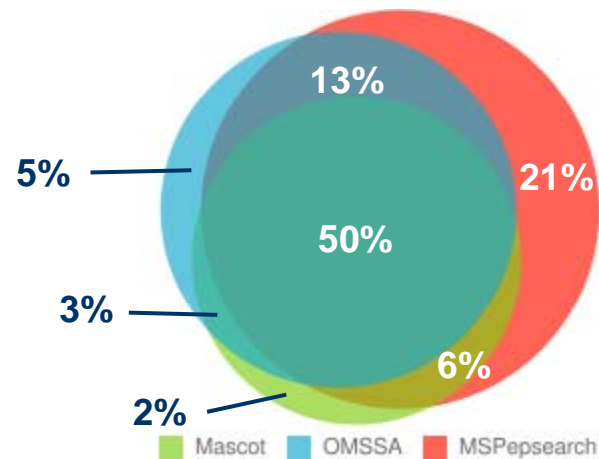


Mascot



Comparison of Results Human Colon Samples

| | MSPepsearch | Mascot | OMSSA |
|----------------------------|-------------|------------|-----------|
| Total Peptides | 39,420 | 23,038 | 28,364 |
| Unique Peptides | 12,356 | 8,424 | 9,854 |
| Overlap MSPepsearch | | 7,696 | 8,691 |
| Overlap Mascot | 7,696 | | 7,320 |
| Overlap All Three | 6,892 | | |
| Run Time | 27 min | 3 h 24 min | 11 h 24 m |



Experimental Conditions

Rat Liver Extracts

15 rat liver membrane samples from Stowers Institute for Medical Research via Tranche, digested with lysC and trypsin, alkylated with iodoacetamide, analyzed on LTQ.

Sequence search engines – Mascot and OMSSA

Fasta – rat from v3.23 IPI combined with reversed fasta

Search Parameters - trypsin, 2 missed cleavages, fixed mod – carbamidomethyl (C), variable mod – oxidation (M), precursor tolerance ± 1.6 Da, msms tolerance - ± 0.8 Da, charges +1-+3 and monoisotopic precursor.

Library search – MSPepsearch

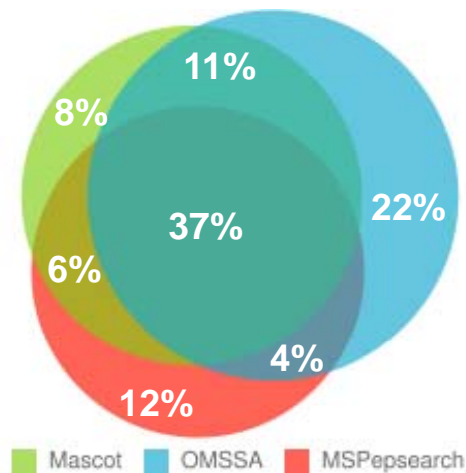
Library was created from searches using the IPI fasta above and filtered to include only tryptic peptides with charges +1-+3 and including only no modifications, carbamidomethyl (C) or oxidation (M).

Search Parameters – precursor tolerance ± 1.6 Da, msms tolerance ± 0.8 Da and monoisotopic precursor.

Decoy library was created from filtered yeast spectra.

Comparison of Results Rat Liver Samples

| | MSPepsearch | Mascot | OMSSA |
|----------------------------|----------------|-----------|----------|
| Total Peptides | 19,848 (5,751) | 13,256 | 14,828 |
| Unique Peptides | 2,839 (909) | 2,965 | 3,521 |
| Overlap MSPepsearch | | 2,060 | 1,940 |
| Overlap Mascot | 2,060 | | 2,273 |
| Overlap All Three | 1,746 | | |
| Run Time | 1 h 6 m | 15 h 43 m | 9 h 10 m |



Experimental Conditions

Spiked Yeast

Sigma48 (equimolar standards of 48 proteins) were spiked into yeast digest at 5 concentration levels in addition to a yeast blank and direct standards as part Clinical Proteomics Technologies for Cancer (CPTAC) study 6. Samples were digested, alkylated with iodoacetamide and analyzed in triplicate on an LTQ-Orbitrap.

Sequence search engines – Mascot and OMSSA

Fasta – yeast from SGD plus Sigma48 combined with reversed fasta

Search Parameters - trypsin, 2 missed cleavages, fixed mod –carbamidomethyl (C), variable mod – oxidation (M), precursor tolerance ± 1.6 Da, msms tolerance - ± 0.8 Da, charges +1-+3 and monoisotopic precursor.

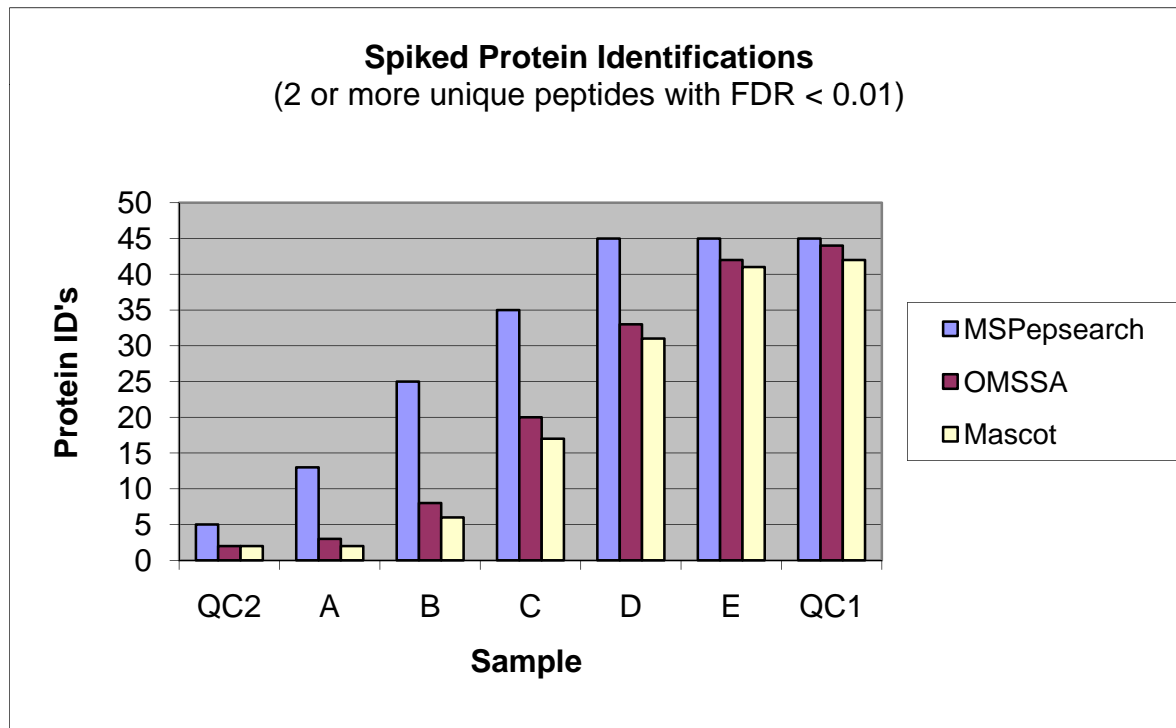
Library search – MSPepsearch

Used concatenation of yeast and Sigma48 libraries filtered to include only tryptic peptides with charges +1-+3 and including only no modifications, carbamidomethyl (C) or oxidation (M).

Search Parameters – precursor tolerance ± 1.6 Da, msms tolerance ± 0.8 Da and monoisotopic precursor.

Decoy library was created from filtered mouse spectra

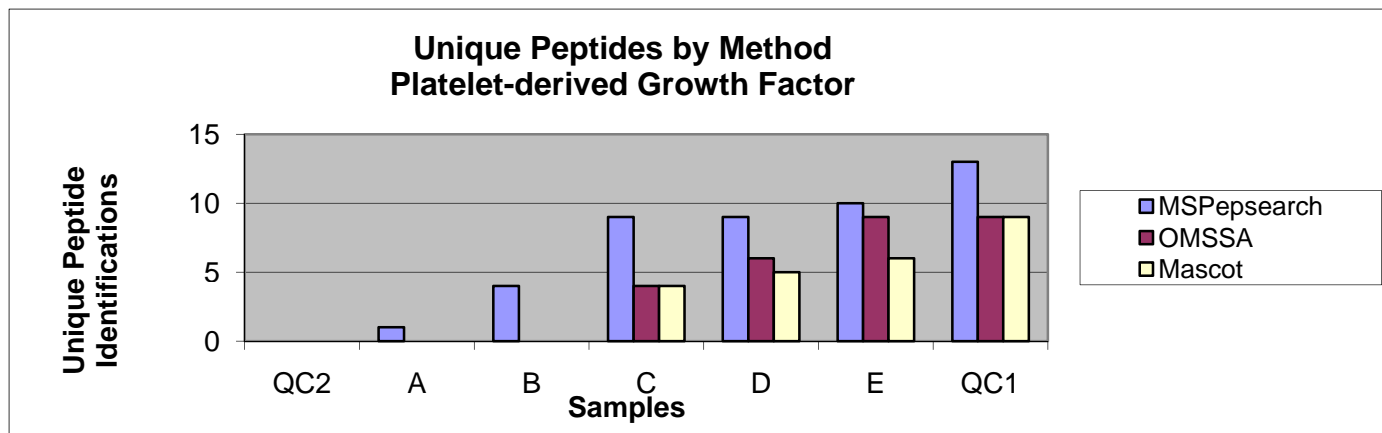
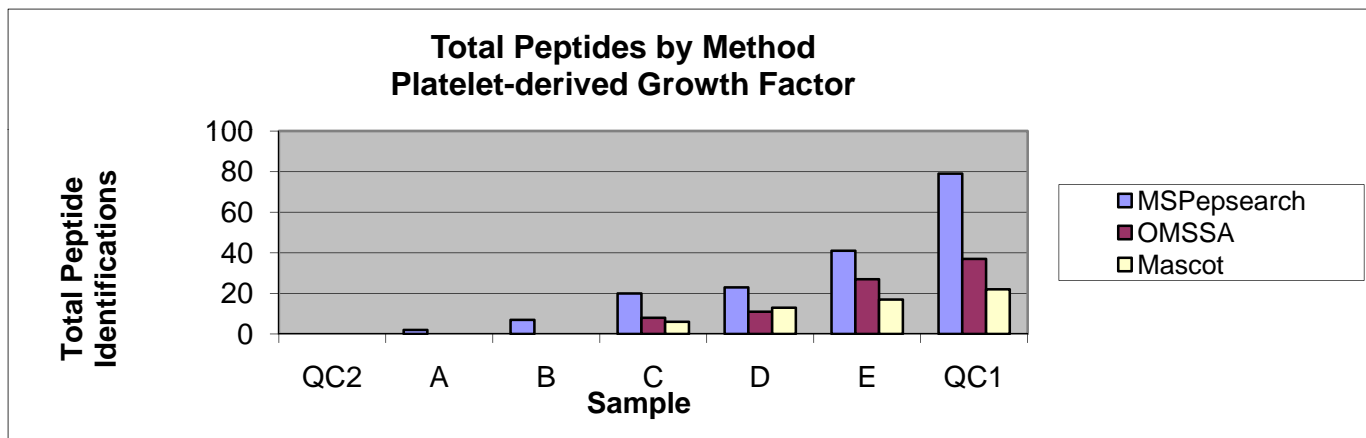
Comparison of Results Spiked Yeast Samples



**Sigma48 Protein
Concentrations
(fmol/ μ l)**

| | |
|-----|-------------|
| QC2 | yeast blank |
| A | 0.25 |
| B | 0.74 |
| C | 2.2 |
| D | 6.7 |
| E | 20 |
| QC1 | 20 |

Comparison of Results Peptides for Individual Protein



Summary

- Spectral library searching has many theoretical advantages over sequence searching
- Decoy libraries created from other species used with a simple target/decoy bias correction enable direct comparisons at FDR .01 with standard methods.
- Even at “low” coverage, library searching yielded equivalent or better results for real world samples.
- Library searching is an excellent addition to standard multiple search engine strategy.
- For known standards with high coverage, library searching is far more sensitive.

Thank You

- NIST collaborators
- Contributors of raw data



Numerous Individual Laboratories

- Future contributors – raw msms data

Contact Info

<http://peptide.nist.gov> jeri.roth@nist.gov