INSTITUTE FOR Systems Biology

Revolutionizing science. Enhancing life.

Spectrum Library Building and Searching

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Henry Lam

Strategies for Improving Reliability
in Protein/Peptide Identification Workshop

2007-11-15

Outline

INSTITUTE FOR Systems Biology
Revolutionizing science. Enhancing life.

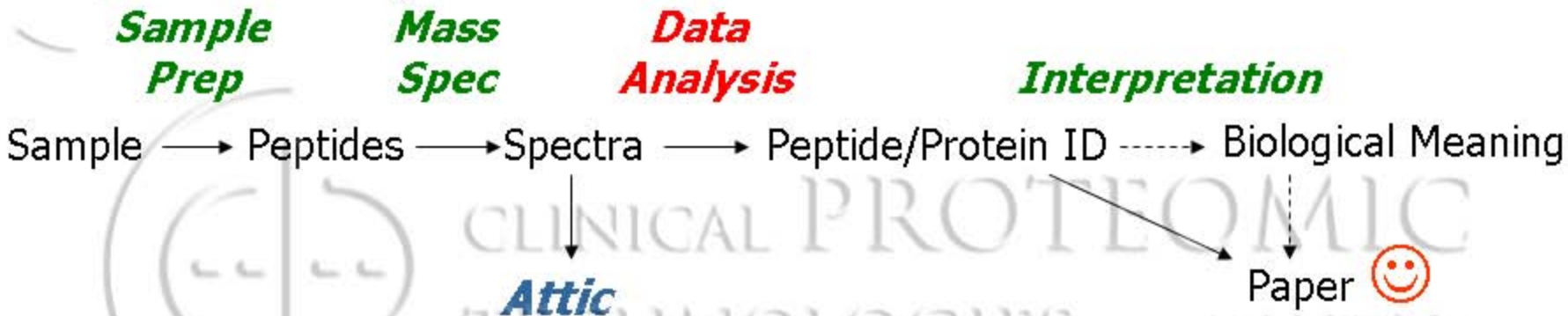
- Motivation
- Spectrum library searching
- Spectrum library building
- Challenges for the future

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Memoryless Workflow

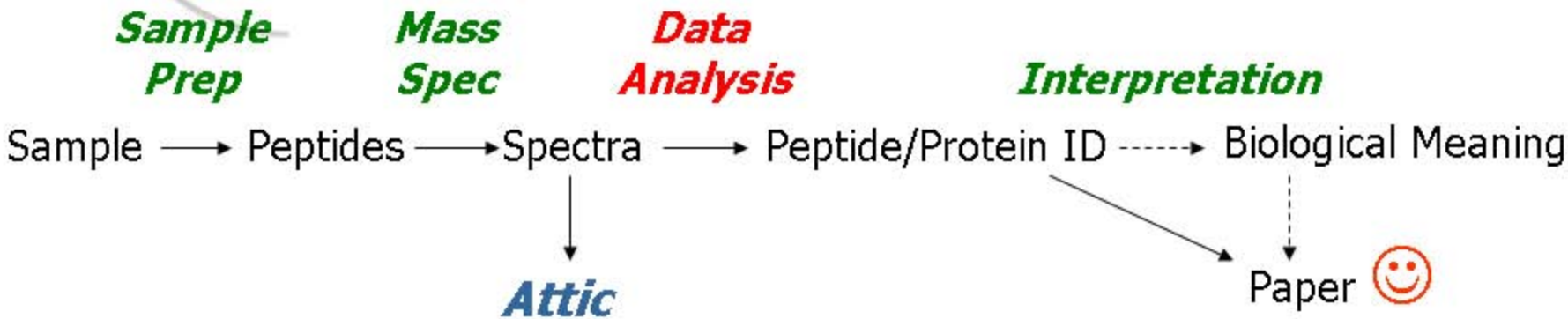
INSTITUTE FOR Systems Biology
Revolutionizing science. Enhancing life.

Experiment #1 by Joe:



CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Experiment #2 by Jane months later:



Learning from the Past

INSTITUTE FOR Systems Biology
Revolutionizing science. Enhancing life.

Experiment #1 by Joe

Sample Prep ***Mass Spec*** ***Data Analysis*** ***Interpretation***

Sample → Peptides → Spectra → Peptide/Protein ID ⇢ Biological Meaning

Peptide/Protein ID → Paper

Biological Meaning ⇢ Paper

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Spectrum libraries !!

Sample Prep ***Mass Spec*** ***Data Analysis*** ***Interpretation***

Sample → Peptides → Spectra → Peptide/Protein ID ⇢ Biological Meaning

Peptide/Protein ID → Paper

Biological Meaning ⇢ Paper

Experiment #2 by Jane months later

Spectrum Libraries

- Simply, collections of identified spectra
  - Rely on sequence searching to make initial identifications
- Enables future identifications by *spectral matching* (similar spectra => same ID)
  - Pros: Extremely fast, very high sensitivity and specificity
  - Cons: Can only find things previously seen

CLINICAL PROTEOMIC TECHNOLOGIES for CANCER

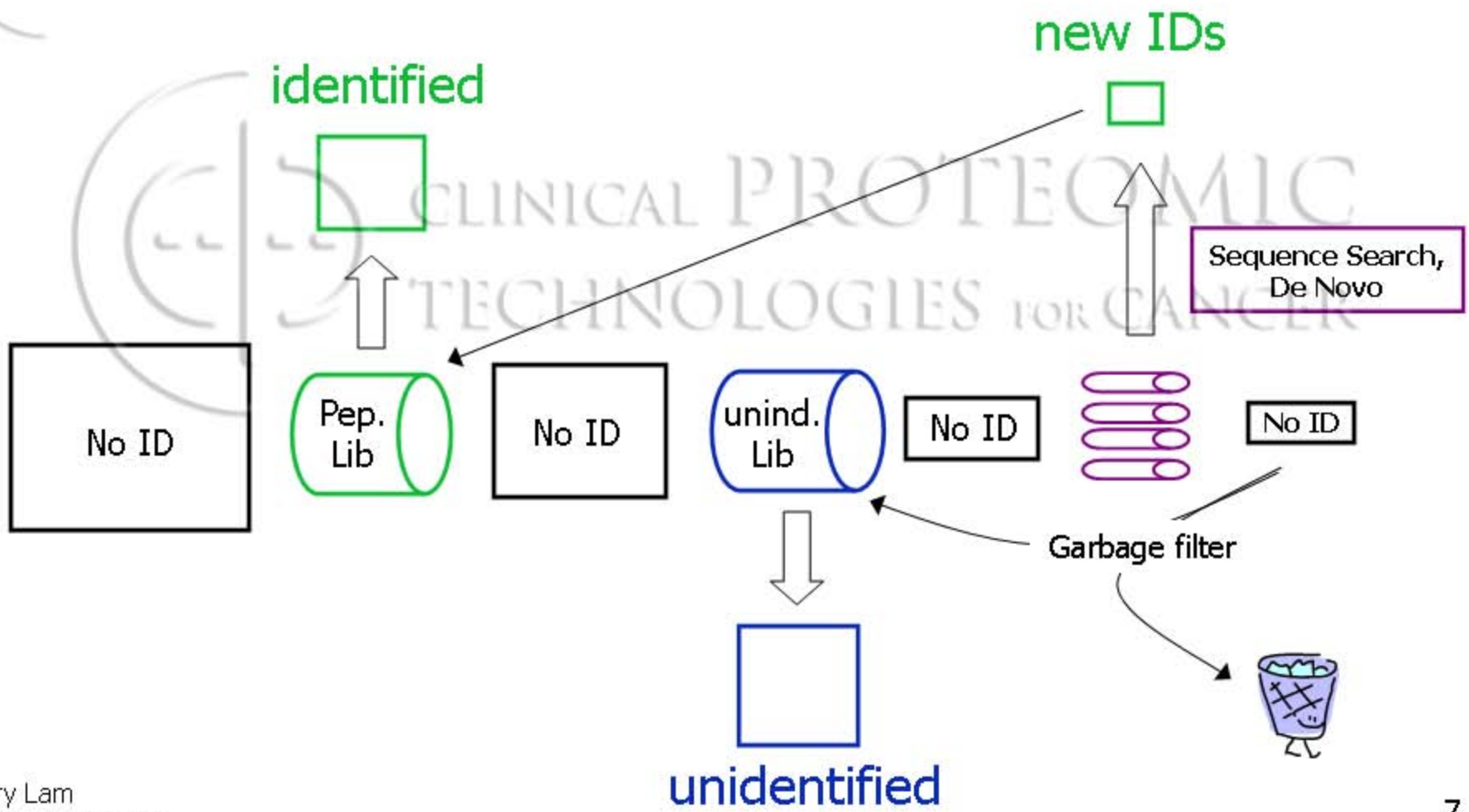
Spectrum Libraries

INSTITUTE FOR Systems Biology
Revolutionizing science. Enhancing life.

- Incomplete library coverage, but…
  - Limiting factor is MS technology, not libraries
  - More interesting biological questions are answered with targeted approaches
  - Can be integrated with sequence searching to improve pipeline

Spectrum Libraries

- "Peptide ID Pipeline of the future" – Steve Stein, NIST



CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Spectral searching

INSTITUTE FOR Systems Biology
Revolutionizing science. Enhancing life.

- NIST MS (NIST, Stein et al, JASMS 1994)
- LIBQUEST (Yates, Yates et al, Anal Chem. 1998)
- X!Hunter (GPM, Craig et al, JPR 2006)
- Bibliospec (MacCoss, Frewen et al, Anal Chem 2006)
- SpectraST (Aebersold, Lam et al, Proteomics 2007)

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Spectral searching

INSTITUTE FOR Systems Biology
Revolutionizing science. Enhancing life.

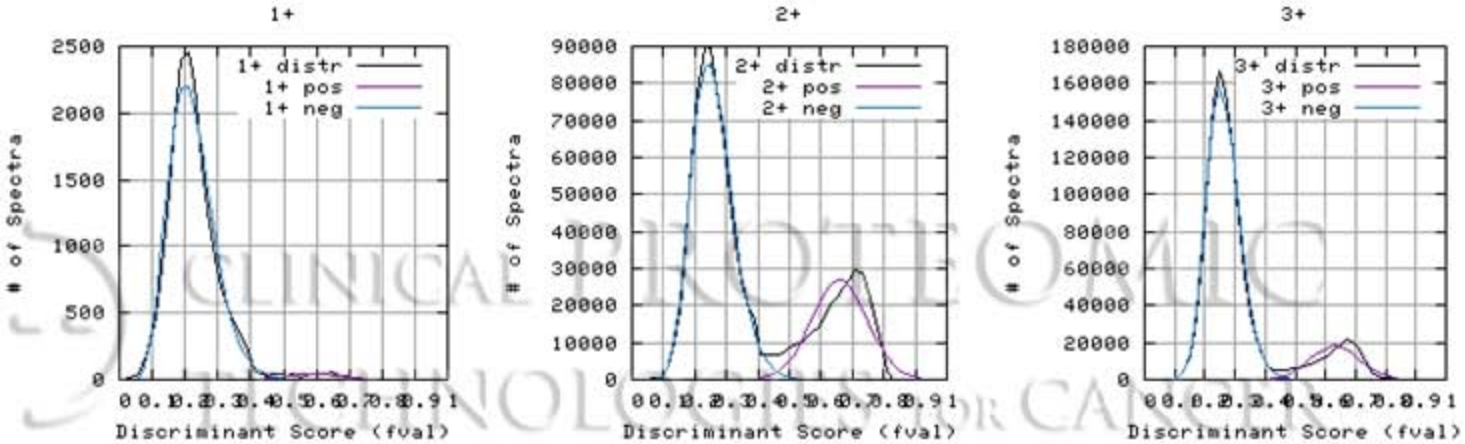
- Human plasma dataset (Novartis-GeneProt)
  - 7,000 runs, 2.4 million MS/MS spectra
  - SEQUEST search
    - Semitryptic, 3 Da precursor mass window
    - Against human IPI
    - 350,000 p>0.99 IDs (PeptideProphet)
    - Weeks on a 80-CPU cluster
  - SpectraST search
    - 3 Th precursor m/z window
    - Against NIST human spectrum library (2006)
    - 430,000 p>0.99 IDs (PeptideProphet)
    - Overnight on 1 CPU

CENTER FOR PROTEOMIC TECHNOLOGIES FOR CANCER

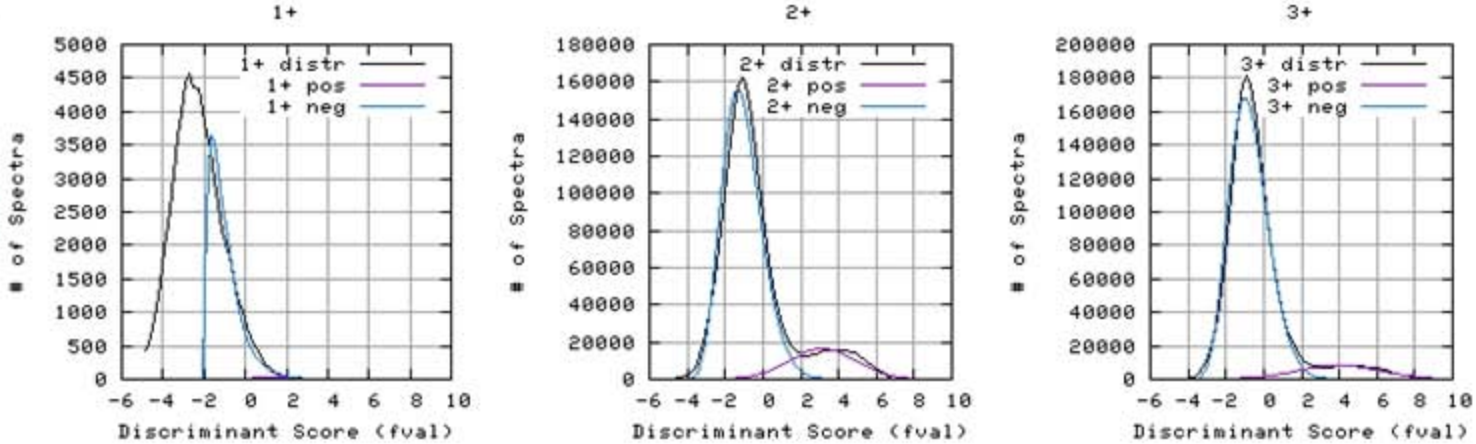
Score discrimination

INSTITUTE FOR Systems Biology
Revolutionizing science. Enhancing life.

SpectraST



SEQUEST



CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Lessons

- SpectraST identified 22% more spectra at same confidence cutoff
- Extra SpectraST identifications are mostly
  - Same IDs made by SEQUEST but at lower confidence
  - Repeated IDs of same peptide ions (from lower-quality spectra that SEQUEST failed to identify)
- SpectraST is rarely wrong with its confident IDs (~0.01% FDR)
- 99.5% of the missed IDs by SpectraST is not in the spectrum library

CANCER PROTEOMIC TECHNOLOGIES FOR CANCER

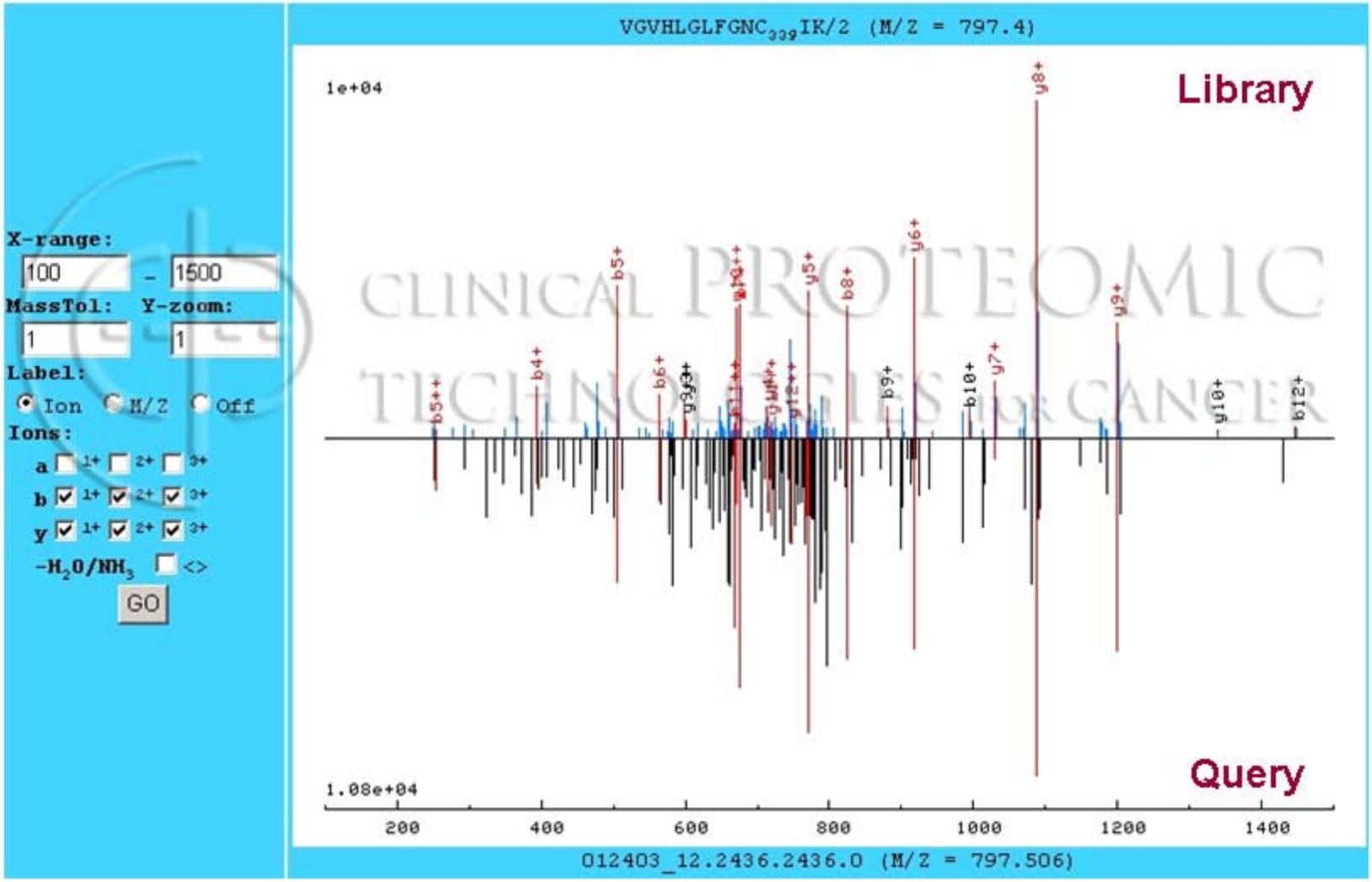
Why the improvement?

INSTITUTE FOR Systems Biology
Revolutionizing science. Enhancing life.

- Smaller search space
  - Only previously observed peptides are searched
- NIST library contains IDs of several sequence search engines
  - Implicitly combines multiple search algorithms
- More precise similarity scoring
  - Global similarity, not just presence of b- and y-ions

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Global similarity



Spectrum Library Building

INSTITUTE FOR Systems Biology
Revolutionizing science. Enhancing life.

- Centralized (NIST, PeptideAtlas, GPM)
  - Greater coverage
  - Higher quality
- Do-it-yourself (X!Hunter, SpectraST)
  - Specialized libraries
  - Proprietory data
  - Data organization and reusability

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Spectrum Library Building

INSTITUTE FOR Systems Biology
Revolutionizing science. Enhancing life.

- How much spectral information to include in the spectra?
  - Top N peaks
  - Only "expected" peaks
  - Consistently present peaks
- What to do with multiple observations of the same peptide ion?
  - Just throw them all in
  - Pick the "best" replicate
  - Consensus

CPTAC PROTEOMIC TECHNOLOGIES FOR CANCER

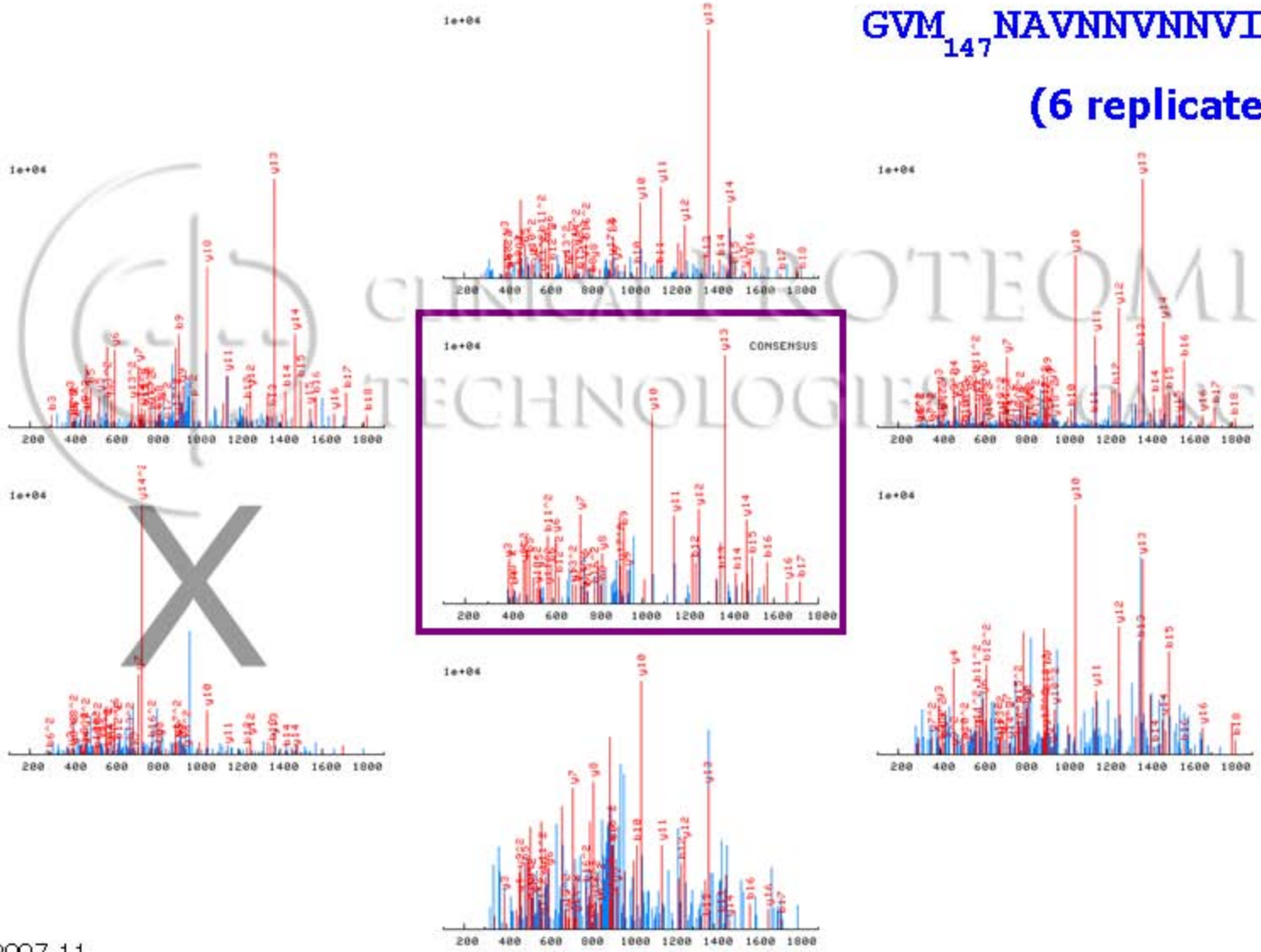
Consensus Spectrum

INSTITUTE FOR Systems Biology

Revolutionizing science. Enhancing life.

$GVM_{147}NAVNNVNNVIAAAFK/2$

(6 replicates)

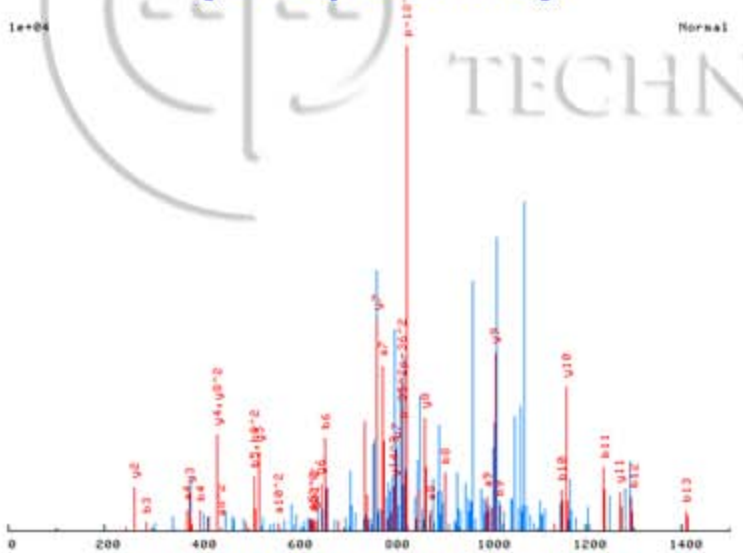


CLINICAL PROTEOMIC TECHNOLOGIES CENTER

Quality control

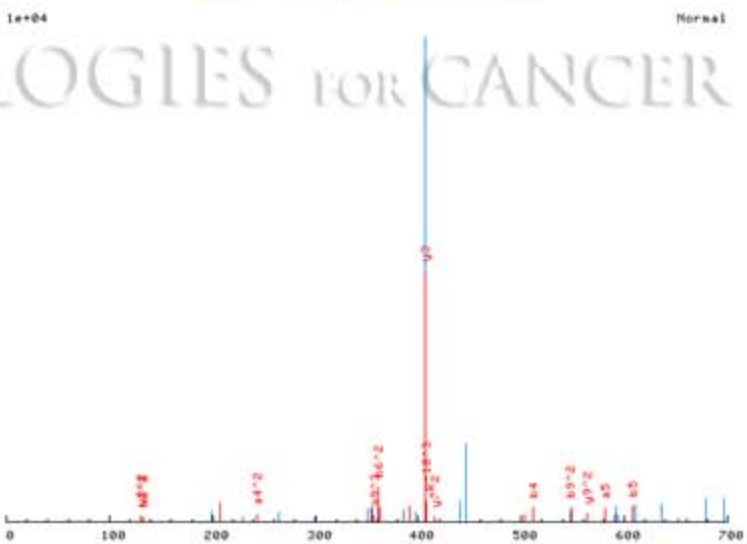
- Problem: There are still occasional false positives and terribly noisy spectra

AVDLLFFTDESGDSR/2
(2 replicates)



Possibly correctly identified, but impure spectrum

DFFTPNLFLK/3
(1 replicate)



False-positive impure spectrum

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

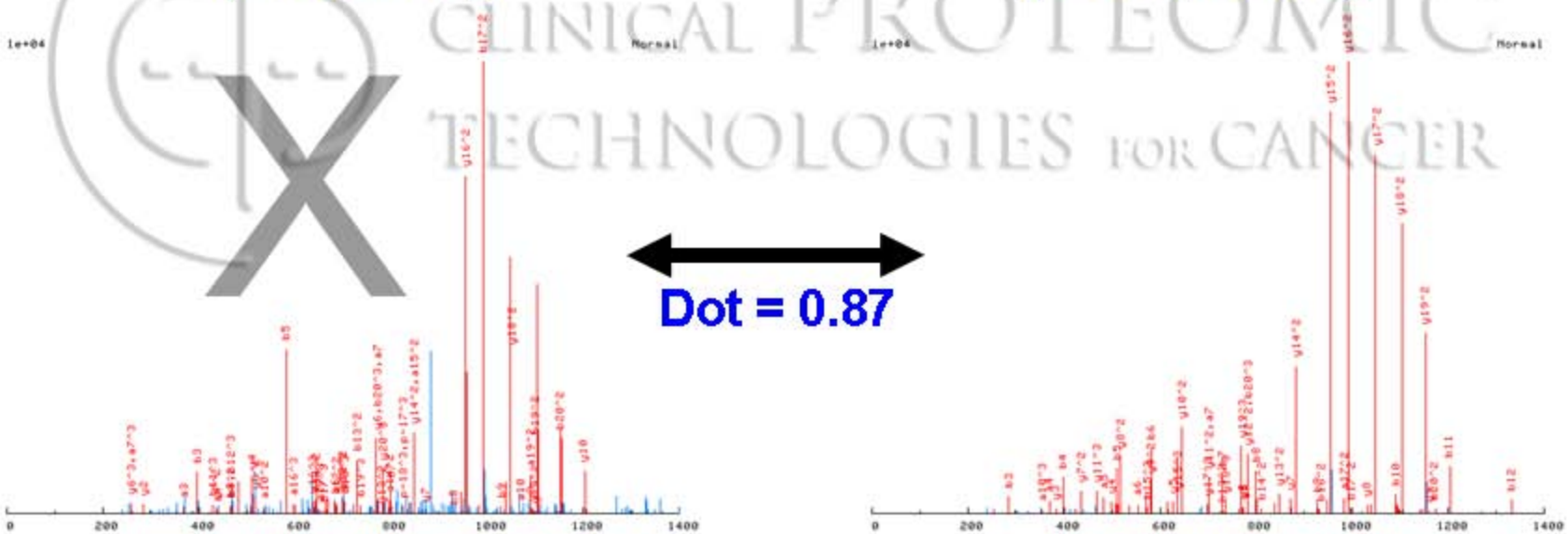
Quality control

INSTITUTE FOR Systems Biology
Revolutionizing science. Enhancing life.

FFTAICDMVAWLGYTPYKVTY/3
(1 replicate)

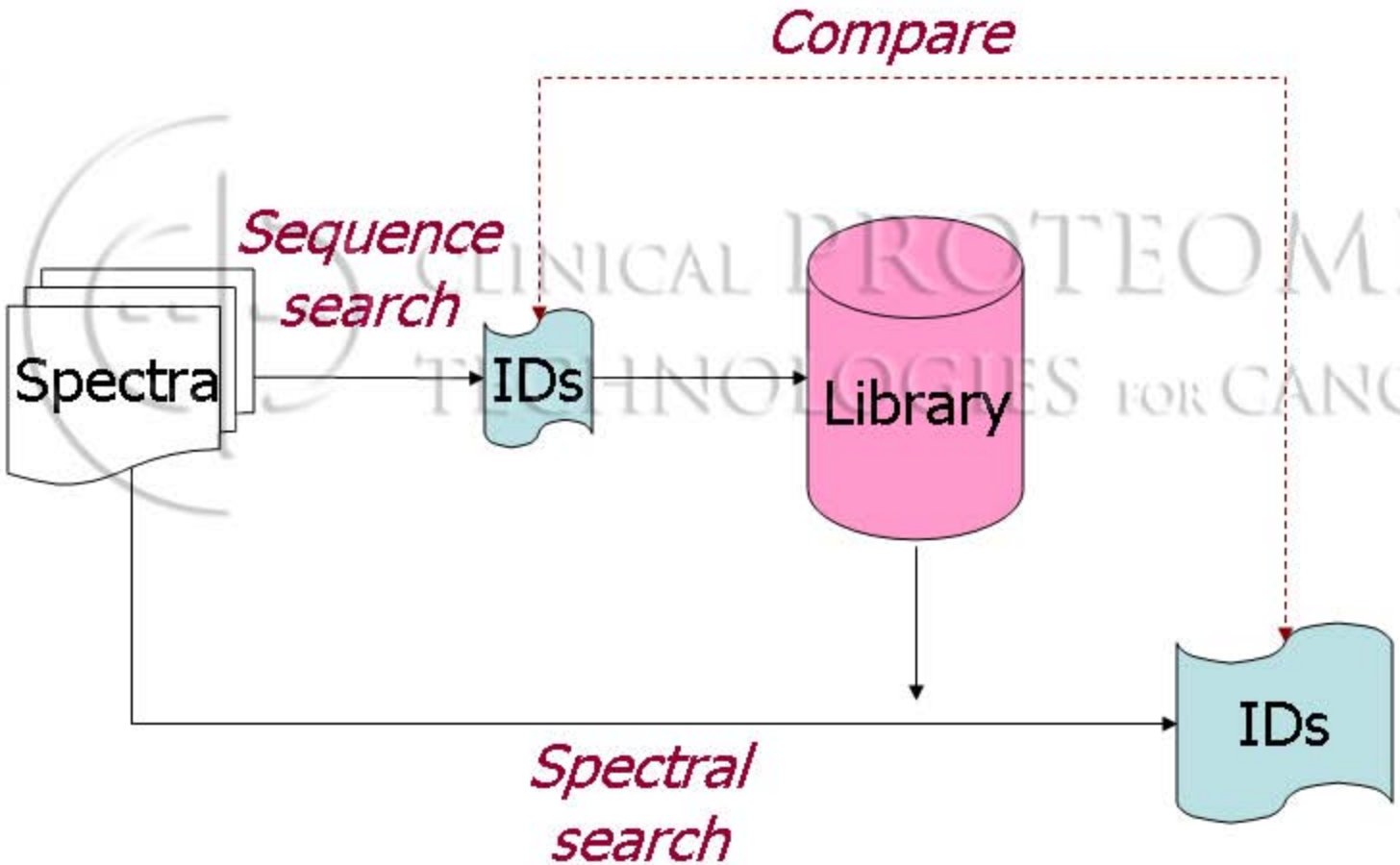
ALVLIAFAQYLQQC$_{160}$PFEDHVK/3
(100 replicates)

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER



Similar spectra having conflicting IDs

Evaluating Library Building Methods



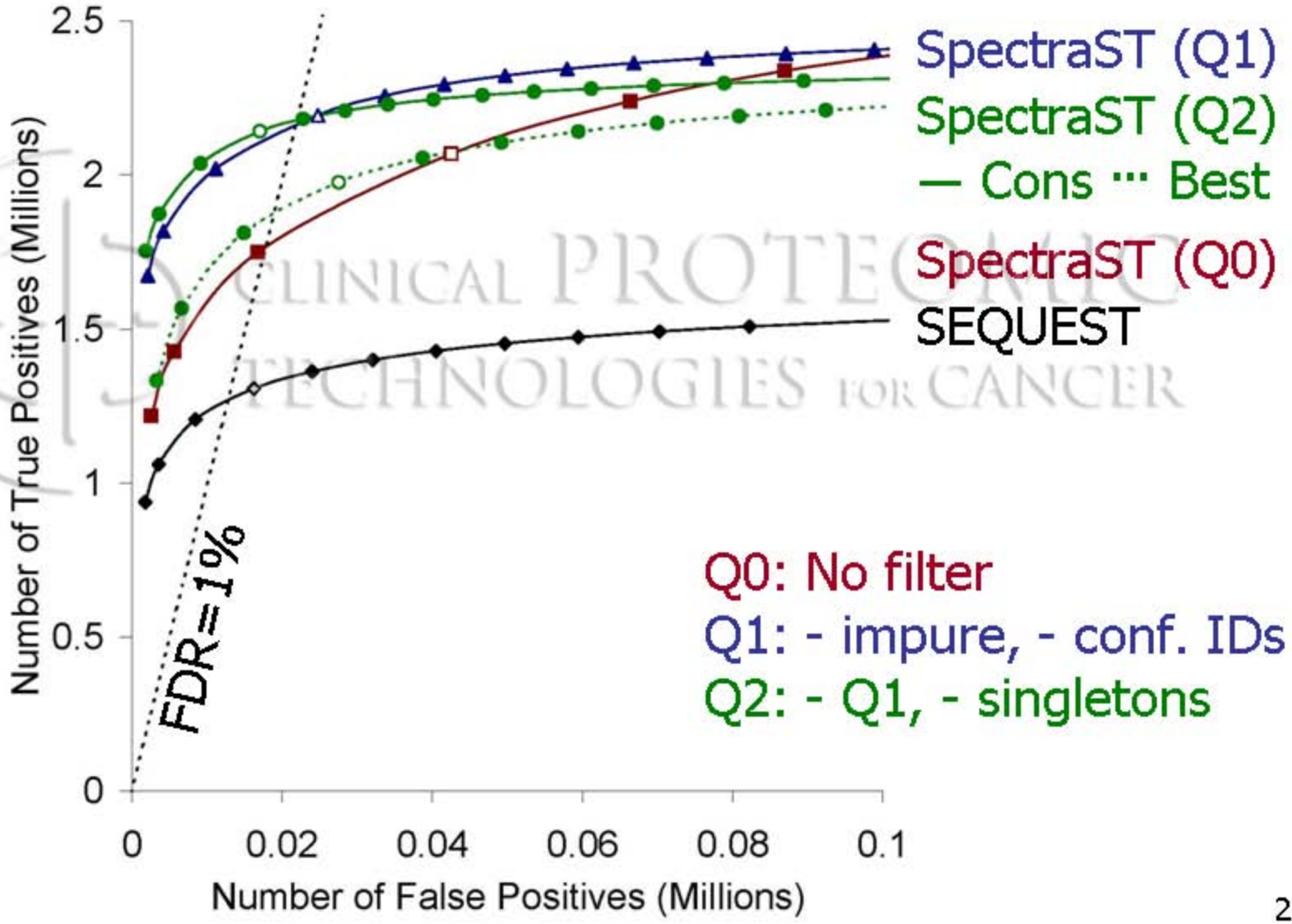
Evaluating Library Building Methods

INSTITUTE FOR Systems Biology
Revolutionizing science. Enhancing life.

- Use entire Human Plasma PeptideAtlas
  - 40 datasets, 1.3M identified spectra at p>0.9
  - Consensus vs best-replicate
  - Different quality levels
- Speed
  - Library building took about 2 days on 1 CPU

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Evaluating Library Building Methods



CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Lessons

- Up to 67% more confident IDs at same false discovery rates by SpectraST compared to SEQUEST
  - 1000x faster!
- Quality of library matters
  - Higher quality means better discrimination, more IDs at same confidence cutoff
  - Eventually trade off with lower coverage
- Consensus is much better than best-replicate

CLINICAL PROTEOMIC TECHNOLOGIES for CANCER

Challenges

- Algorithm optimization
  - Searching / scoring
  - Consensus building
  - Proper quality filters (coverage vs quality)
- Statistical validation
  - How do you determine accurate false discovery rates?
- Pipeline development
  - Integration with sequence searching

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Challenges

INSTITUTE FOR Systems Biology
Revolutionizing science. Enhancing life.

- Centralized library building
  - How to gather data from community efficiently?
  - How to ensure quality of data?
  - How finely should we specialize libraries?
    - Different instrument type? Different instrument setting (e.g. CE)?
    - Coverage vs quality

- General user friendliness!

CANCER PROTEOMIC TECHNOLOGIES FOR CANCER

Acknowledgements

INSTITUTE FOR Systems Biology

Revolutionizing science. Enhancing life.

Ruedi Aebersold

Steve Stein
(NIST)

Eric Deutsch

National **Heart**
Lung and Blood Institute