CLINICAL PROTEOMIC
TECHNOLOGIES FOR CANCER

# Spectral Libraries:

## Productivity Enhancers for Cancer Proteomics

Christopher R. Kinsinger Ph.D.
National Cancer Institute
Clinical Proteomic Technologies for Cancer

**NCI Board of Scientific Advisors**
*June 22, 2009*

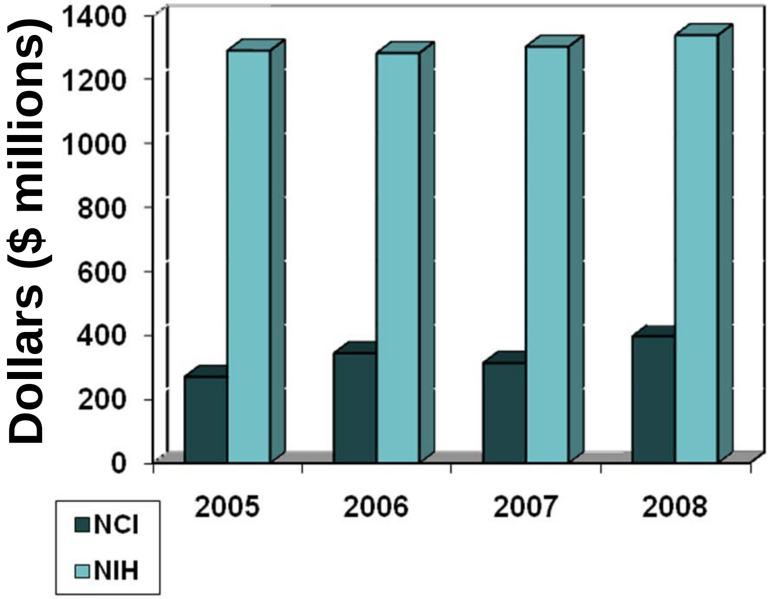
# Today’s presentation

## Outline:

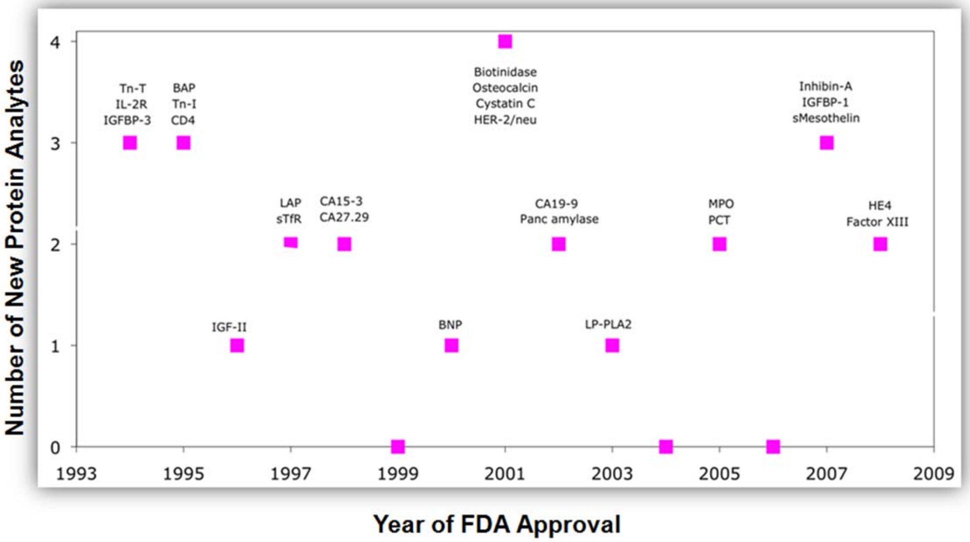
- State of cancer proteomics
- Critical role of Spectral Libraries
  - In discovery proteomics
  - In targeted proteomics
- Proposed concept
- Questions

# Current investment in proteomics

## NCI/NIH funding in proteomics

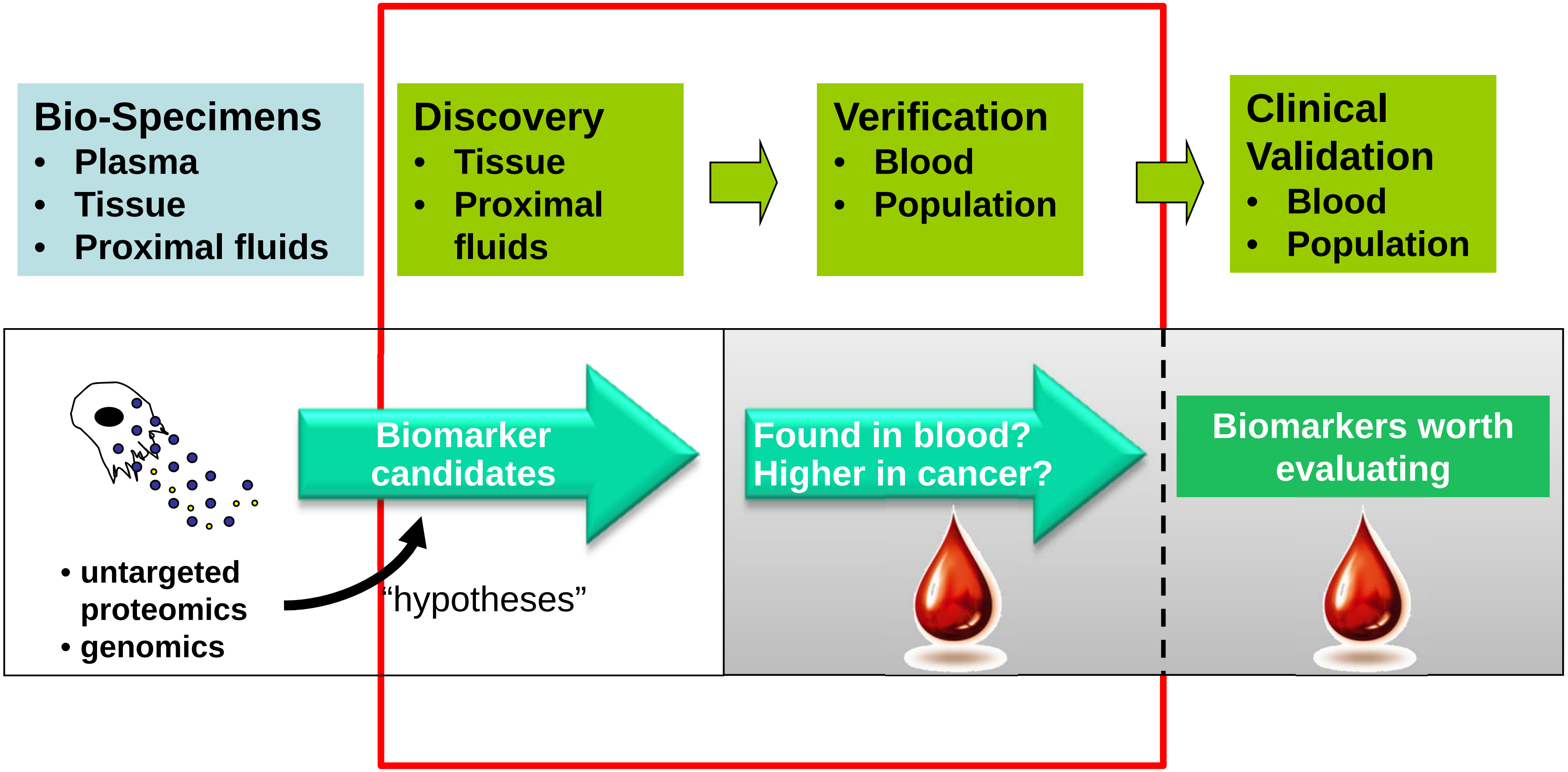


## Number of new FDA-approved protein analytes



**What if a small project could increase the efficiency of these dollars by 50%?**

# Enhancing discovery- and verification (targeted)-stage components in a biomarker development pipeline



**Bio-Specimens**
- Plasma
- Tissue
- Proximal fluids

**Discovery**
- Tissue
- Proximal fluids

**Verification**
- Blood
- Population

**Clinical Validation**
- Blood
- Population

- untargeted proteomics
- genomics

“hypotheses”

Biomarker candidates

Found in blood?
Higher in cancer?

Biomarkers worth evaluating

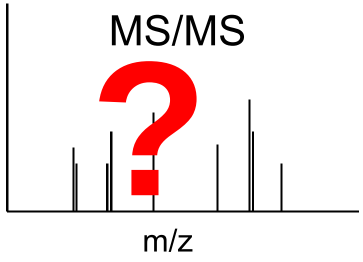
# Current state of peptide identification

**Proteomics experiment**

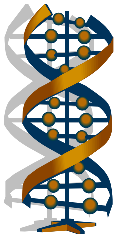
Protein mixture

Physical digest

Mass spec analysis



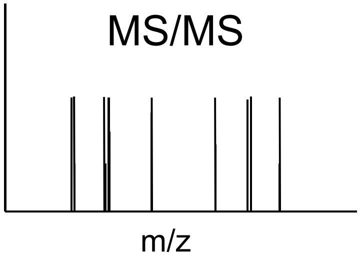
**Human genome sequence library**



Theoretical digest

(DNA → RNA → Proteins → Peptides

Fragmentation model



**Theoretical spectra**

**Match product ion spectra**

**Score**

# The Spectral Library solution

**What is a Spectral Library?**

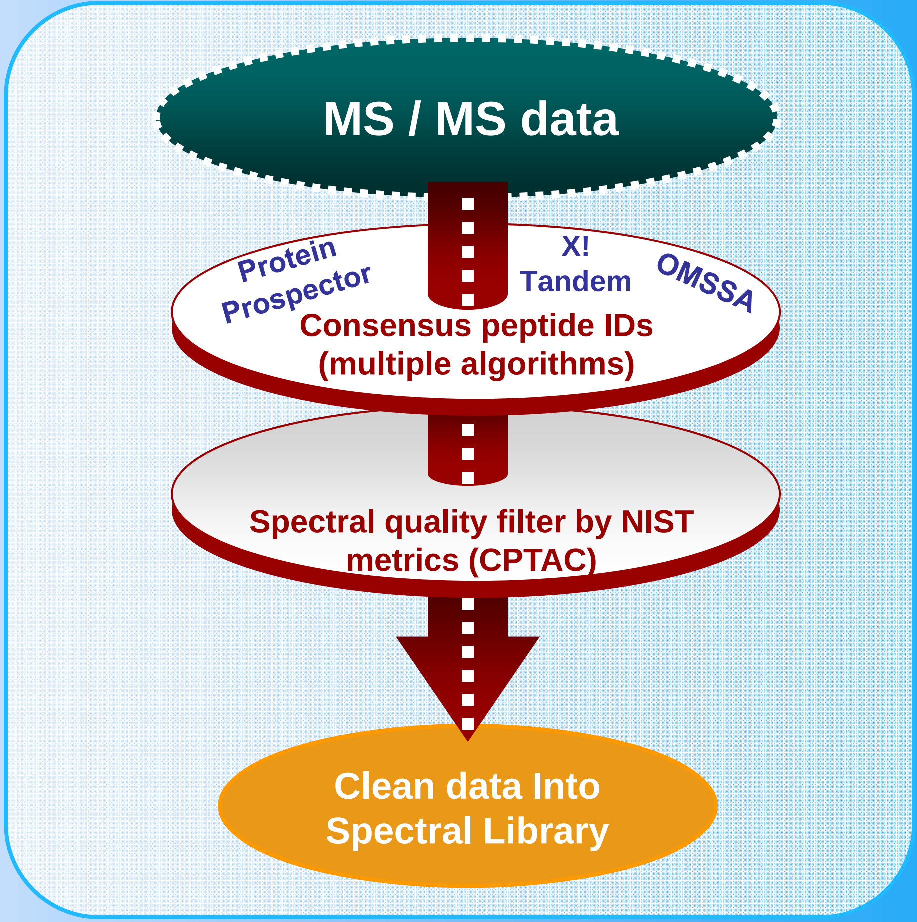
## THE GOLD STANDARD

- Used in chemical, drug, forensics industry
- Catalog of the highest quality MS spectra
- Proven method to identify unknown spectra
- **Maintained by NIST**

# The Spectral Library Solution

What is a Spectral Library?

How is a Spectral Library built?


MS / MS data

Protein Prospector

X! Tandem

OMSSA

Consensus peptide IDs (multiple algorithms)

Spectral quality filter by NIST metrics (CPTAC)

Clean data Into Spectral Library

# The Spectral Library Solution

**What is a Spectral Library?**

**How is a Spectral Library built?**

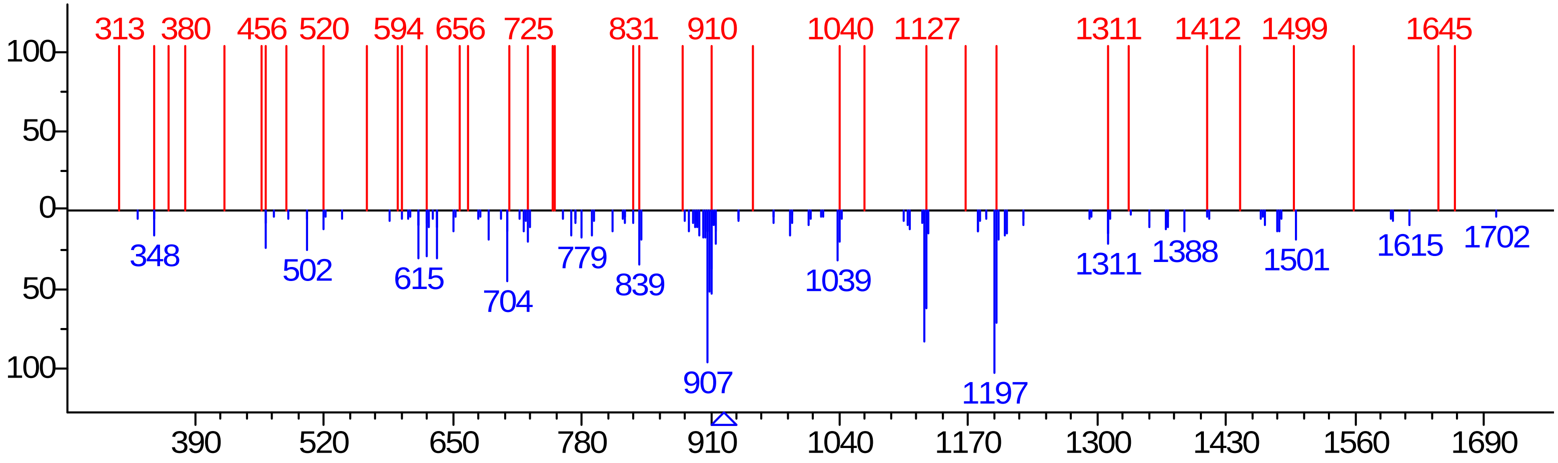
**What are the advantages?**

- Adds a $2^{nd}$ dimension of search
  - Intensity of peaks
  - Peptide sequence
- Improves speed and reliability in peptide ID
- Compilation of all observable peptides

# How spectral libraries work

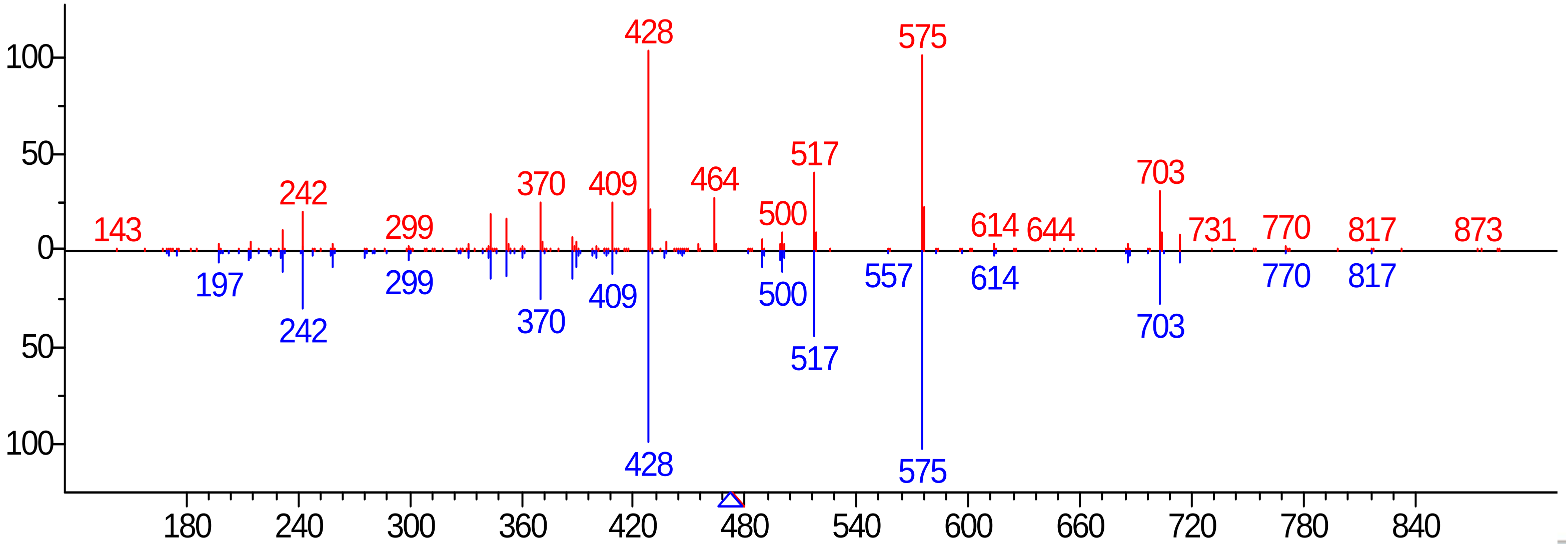
Theoretical spectrum

Physical spectrum



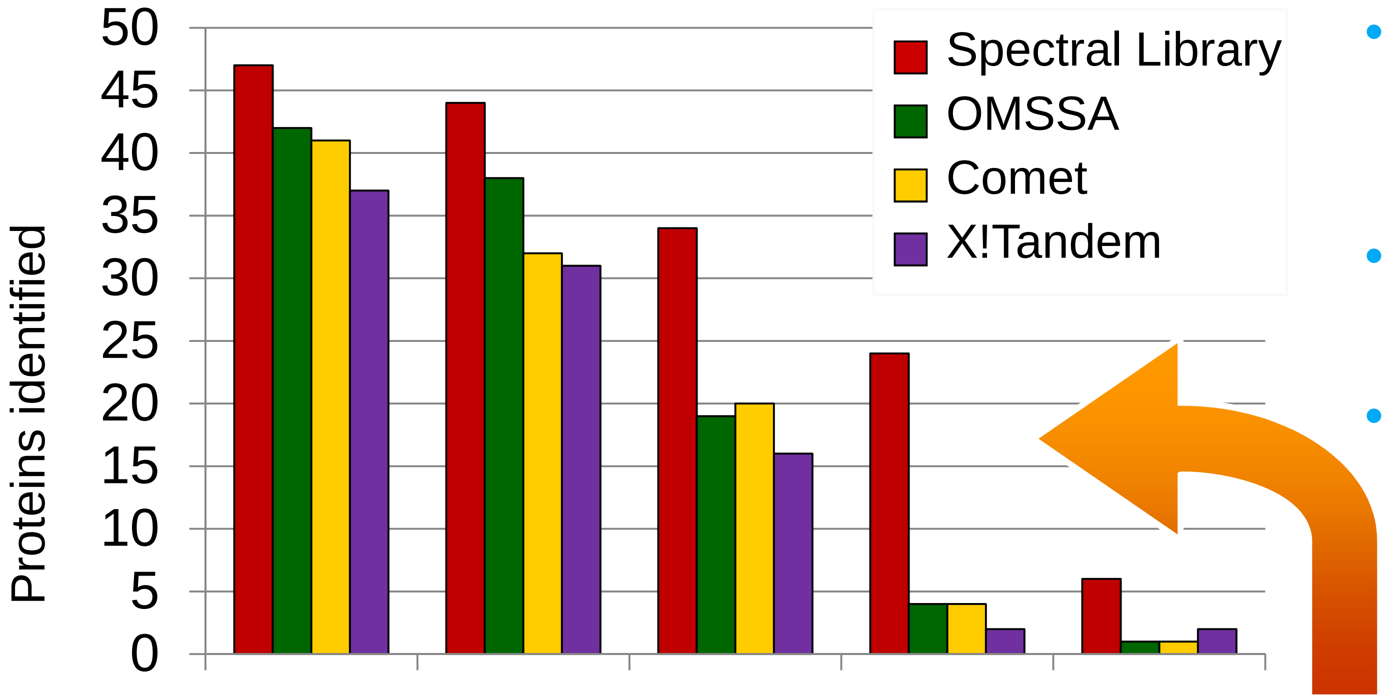
Library spectrum

Physical spectrum



# Enhancing discovery-stage: recombinant proteins

## *Enhancing identification of low abundant proteins*



- 48 human proteins spiked into yeast
- High-end MS platforms
- 500%+ increase in identification!

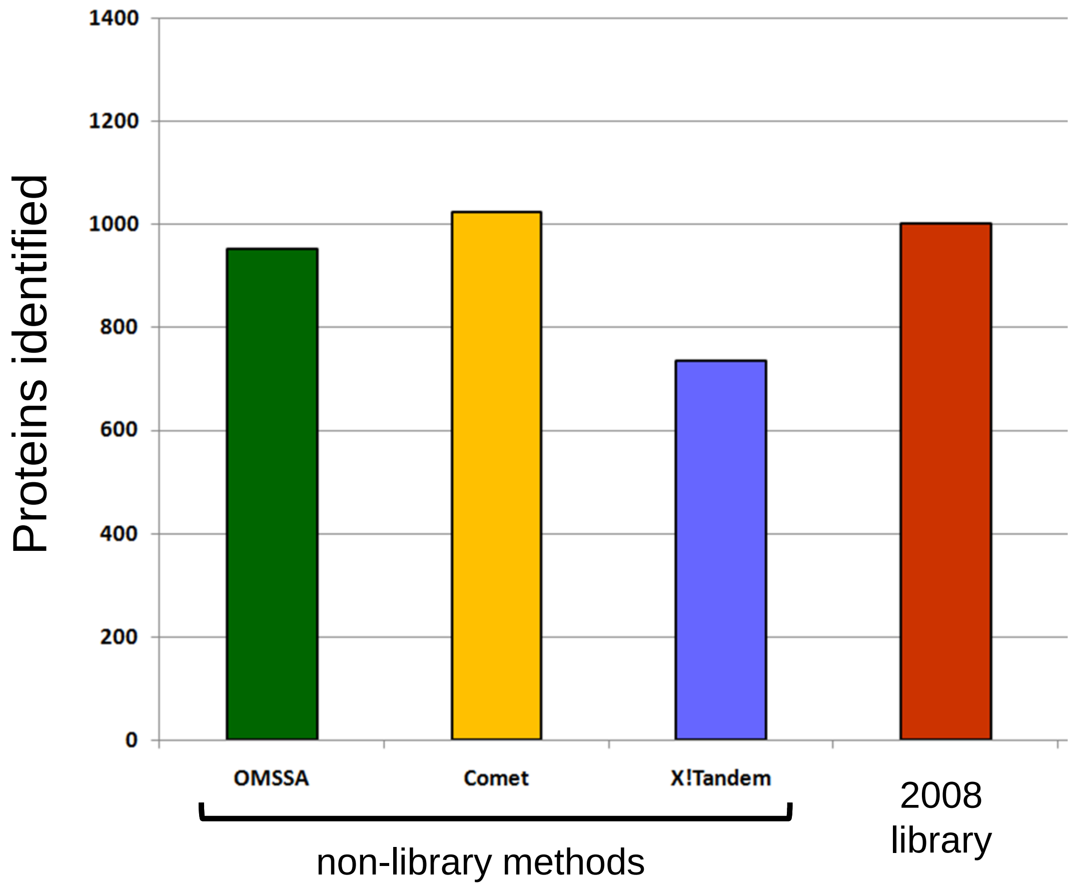
**Spectral library enables identification of low-abundant proteins**

Concentration (fmol/uL)

# Enhancing discovery-stage: clinical tissue

Colon tissue data

**Enriching library enriches discovery**

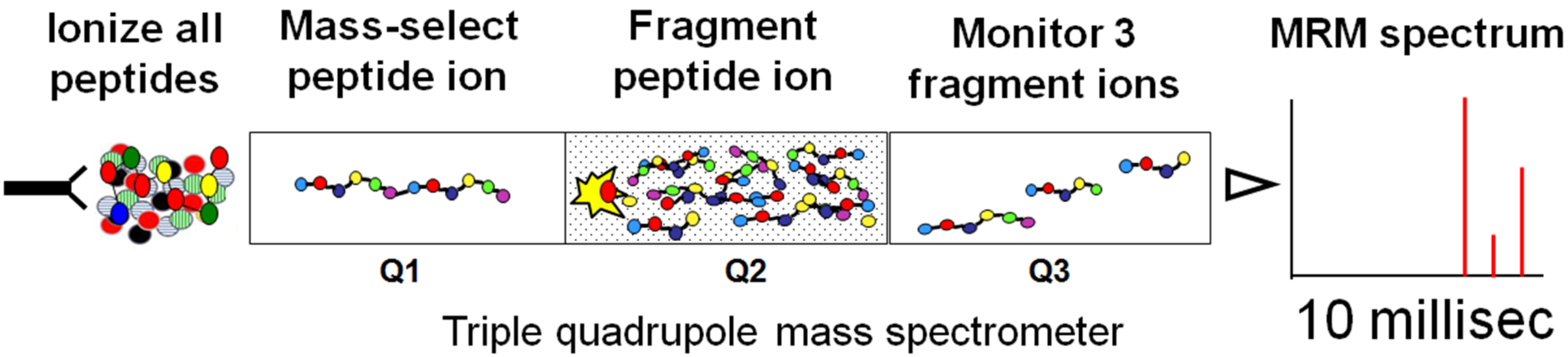


**Over 35% additional proteins identified when library contains one additional dataset**

# Spectral libraries for verification (targeted) stage proteomics

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

**MRM-MS** **(Multiple Reaction Monitoring)**



- **Quantitative mass spectrometry**

*From Steve Carr*

# Enhancing verification-stage

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

- **Select 3-5 target peptides**
  - Representative of parent protein
  - Detectable by mass spectrometer

**Provided by Spectral Library**

# Overview of proposed concept:
## Adding value to biomarker development

**Goal:** Develop public library that anchors proteomic analysis to the physical properties of a peptide through its MS/MS spectrum

### Human Spectral Library

- High quality human biological samples
  - Tissue
  - Recombinant proteins
- High quality peptide spectra
- Coordination among data generators, library developers and data integrators

### Advantages of Spectral Library

- Accelerates and improves ID of low abundant proteins for discovery
- Provides increasing registry of known peptides
- Becomes an index of assay design
- Creates a community-wide resource and shared interest to foster interactions among diverse research groups

**Strengthens the first stages of the biomarker development pipeline**

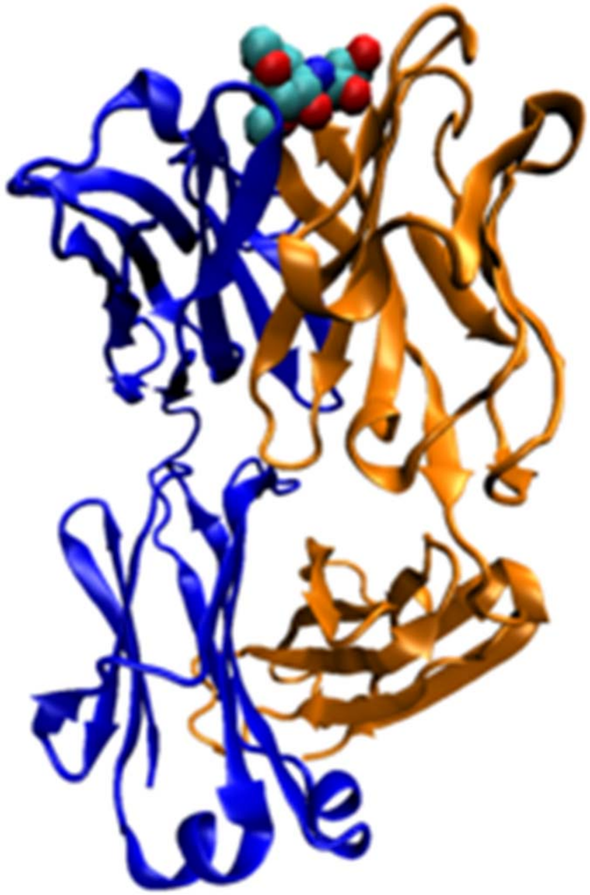
# Representation of cancer tissue in library is dismal

| Sample source | % of files in current library are from … |
|---|---|
| Colon | 2.31 |
| CSF | 1.07 |
| Kidney | 1.59 |
| Liver | 2.40 |
| Lymph | 6.15 |
| Plasma | 77.77 |
| Red Blood Cells | 2.95 |

- All other tissue types are <1% of current library
- Project will catalog proteins from 15 tissue types
- Increase total number of peptides in library by 50%

# Further expanding a library with non-native proteins

- Begin with tissue samples
- Identify key proteins missing from library (TCGA, SPOREs, ICBP, etc.)
- Fill gaps with recombinant proteins or synthetic peptides
- At least 70% of peptides are unmodified
- Complete coverage of protein
- Aids identification of high-priority, low-abundant proteins



# Evaluation criteria

- Increase number of peptide spectra in spectral library
- Increase number of proteins represented in library
- Provide a sustainable, caBIG-compatible data repository for proteomics data

# Components of concept:
## Leveraging NCI resources and partners

## NCI resources

1. **Sample source (tissue, protein/peptide production)**

2. **Protein analysis (data generators)**
   - Partners: NCI-F
   - Generate high quality spectra to more extensively represent all human protein sequences

3. **Data coordinating center**
   - Partner: CBIIT/caBIG®
   - Maintain CPAS database of experiments, peptides, proteins, and raw spectra; ensure quality of and completeness of annotation; leverage caBIG® data portal capabilities and Cancer Center network

## Leveraged activities

4. **Spectral Library development**
   - Partner: NIST
   - Receive peptide spectra from CBIIT and incorporate into human spectral library

5. **Data integration with other resources**
   - Partner: NCBI
   - Acquire data submitted to the NCI for incorporation into NIH Peptidome database

# Proposed Spectral Library (timeline & budget)

| Initiative title | FY10 | FY11 | FY12 | $ |
|---|---|---|---|---|
| 1) **Biospecimen tissue acquisition** (OBBR) | $666,667 | $433,333 | | **1.1 million** |
| 2) **Recombinant protein production** (RFP) | | $300,000 | $300,000 | **600,000** |
| 3) **Protein analysis** (Data generators) (RFP) | $800,000 | $800,000 | $800,000 | **2.4 million** |
| 4) **Data coordinating center** (CBIIT) | $233,333 | $233,333 | $233,333 | **700,000** |
| 5) **Spectral Library development** (NIST) | | | | **Leveraged activity** |
| 6) **Data integration with other NIH resources** (NCBI Peptidome) | | | | **Leveraged activity** |
| **Total:** | 1.7 million | 1.7 million | 1.3 million | 4.8 million |

# Summary

- Registry of high quality, assigned peptide spectra
- Enhance biomarker development in both the discovery and verification phases
- Augment existing spectral libraries by 50% with spectra from cancer-relevant proteins
- Spectral libraries will increase efficiency of NCI’s investment in proteomics