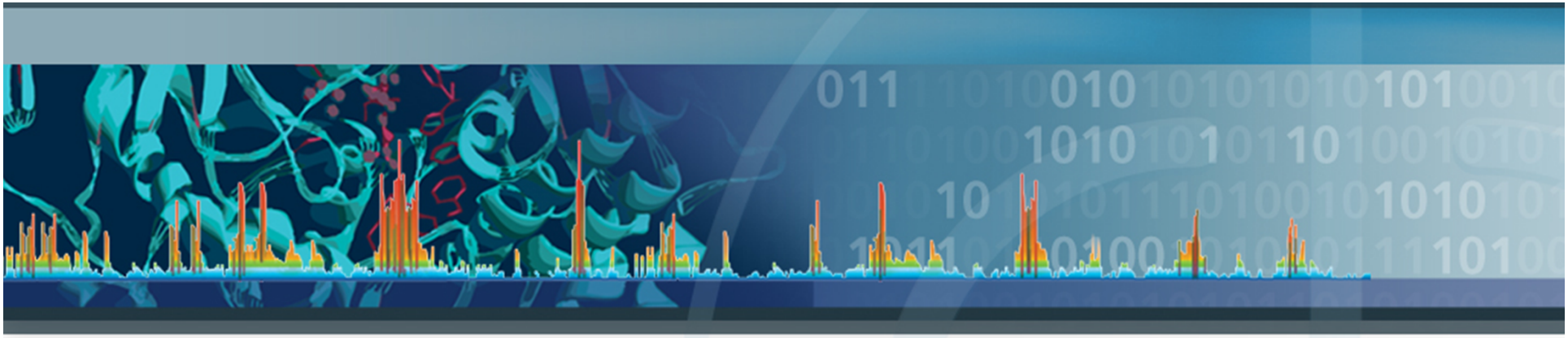


Challenges of Clinical Proteomics and Path Forward Defined by the CPTACs

Steven A. Carr, Broad Institute of MIT and Harvard
and
Daniel Liebler, Vanderbilt University School of Medicine

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

DEPARTMENT OF HEALTH & HUMAN SERVICES · USA

NATIONAL INSTITUTES OF HEALTH

NATIONAL CANCER INSTITUTE

What were the big questions in clinical proteomics at launch of program (2006)?

- If “signal” exists, can discovery proteomic methods detect it?
  - At what abundance levels can proteins be reliably identified?
  - How reproducible are discovery platforms* (intra-lab, inter-lab)?
  - Are estimates of differential protein expression in complex proteomes reproducible between labs and platforms?
  - How can “depth” of detection be improved?

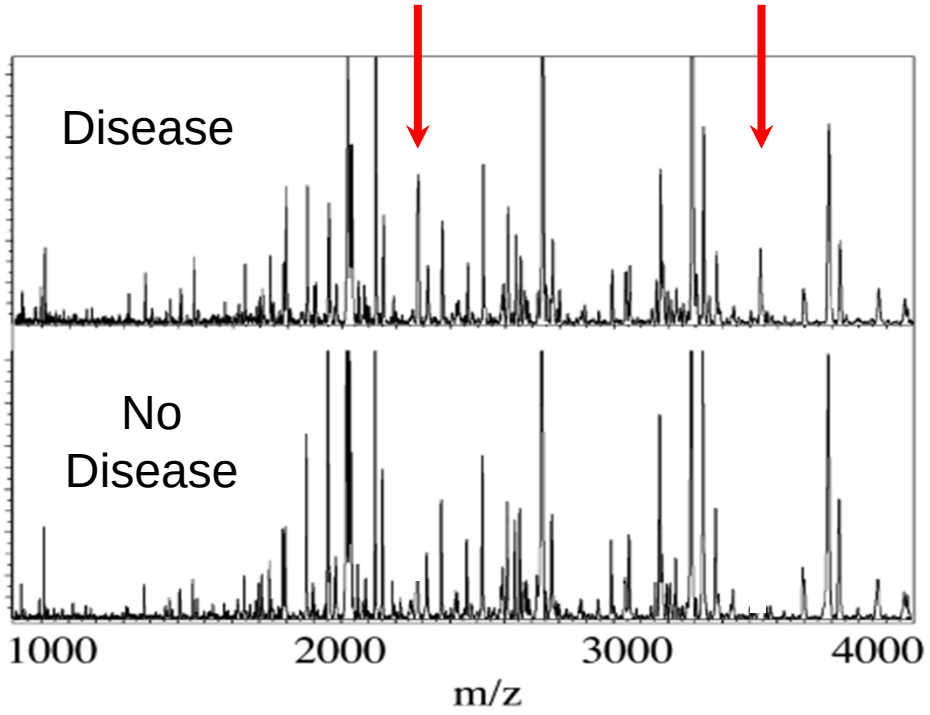
** “Platform” includes sample processing and fractionation methods as well as the LCMS system.*

What were the big questions in clinical proteomics at launch of program (2006)?

- What samples are best suited for proteomics-based discovery?
  - Plasma?
  - Tissue?
  - Other?

- What is the process for moving from “Discovery” to clinically useful marker(s)?
  - How do we credential and prioritize candidates from Discover?
  - What would a pipeline look like?

Pattern-based biomarker discovery in blood: what we knew



- **Collect spectra across large number of samples (minimal processing)**
- **Determine difference signals with machine learning (training)**
- **Identify discriminant portions of signatures (build models)**
- **Apply to test samples (other collection sites)**

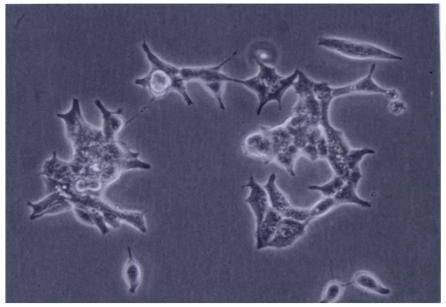
Pattern-based biomarker discovery in blood: what we knew

Pattern Methods:

- only the most abundant proteins detected (sensitivity)
- discriminating pattern often not disease specific (specificity)
- limited ability to derive identity (cannot migrate to other assay)
- Poor study designs used in many published studies

"Unbiased" discovery proteomics in blood: what we knew

Biological Samples (state a, b, c…)



Protein Mixtures

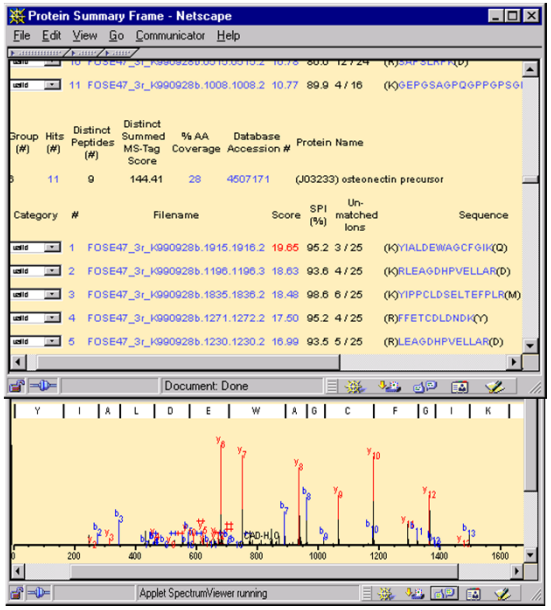
Digest to Peptides

LC/MS/MS

Separate Peptides

Analyze Peptides
- MW
- sequence

Data Analysis



Identify Peptides

Protein i.d. and Relative Abundance

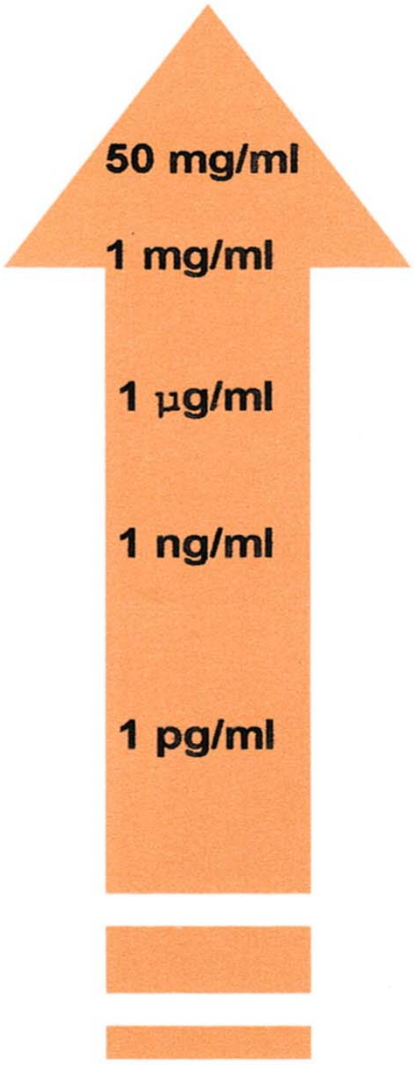
"Unbiased" discovery proteomics in blood: what we knew

Identity-based methods:

- Only ca. 1000 proteins id'ed with high confidence in blood*
- Abundant proteins have very high and reproducible representation in the data, but
- Proteins at < 1ug/mL are poorly represented
  - Detection of lower abundance proteins requires fractionation and depletion, but precise relationship to platform (methods; instruments) remains unclear
- Low or only stochastic representation of proteins of interest (ng/mL and lower range).

*** States et al. Nat. Biotech. (2006) 24: 333**

Is blood the best sample for biomarker discovery?



Albumin
IgG, IgA
Haptoglobulin

IgE

CEA (10 ng/ml)
CK-MB (3 ng/ml)
PSA (1 ng/ml)

Troponin T (10 pg/ml)
FT3
TSH
NT-pro-BNP

Blood is <u>the</u> most complex proteome

- **Dynamic range of proteins in blood ~ 10^{11}**
- **Dynamic range MS ≤ 10^{3}**

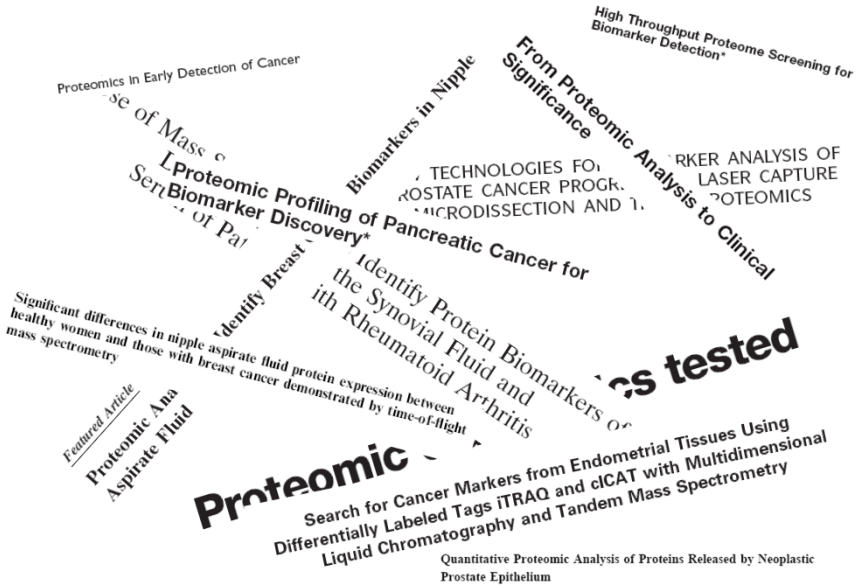
Are tissues and proximal fluids better samples for proteomics discovery?

Why tumor tissue and biofluids proximal to tumor?

- Potential biomarkers likely to exist in the tumor
- May be differentially cleaved/secreted/shed from tumor into surroundings
- Fluids close to “disease volume” expected to be markedly enriched in candidate markers

  100x to 1000x vs. plasma have been observed

The process for moving from “Discovery” to clinic: what we knew



- Lots of proteins being proposed as biomarkers but extremely few introduced into clinical practice
- Methods for **validation** of biomarkers well established but very expensive and slow; reserved for exceptionally promising candidates
- Credentialing constrained by chance availability of reliable antibody assays
    - **Can mass spectrometry provide the missing bridge?**

What were the systems and technical barriers?

- Inadequate supply of high quality biospecimens and clinical data
- Absence of coordinated systematic effort by expert proteomics labs focused on biomarker discovery
- Multiple ad hoc data analysis methods
- Insufficient tools for data capture and knowledge creation
- Enormous diversity, range, and dynamic nature of proteins to be measured
- Difficulty in measuring large number of features simultaneously

What were the systems and technical barriers?

Discovery leads to candidates, not biomarkers

- Extensive fraction. required to detect lower level proteins
- Low analysis throughput (e.g. 2 samples / 3 weeks)
- Data has high dimensionality (>>100 differences)
- Stochastic, incomplete sampling by MS system
- Recipe for high false discovery rate

Candidates must be confirmed and quantified in blood

- Need robust method with sufficient throughput to reliably credential large numbers of candidates in blood

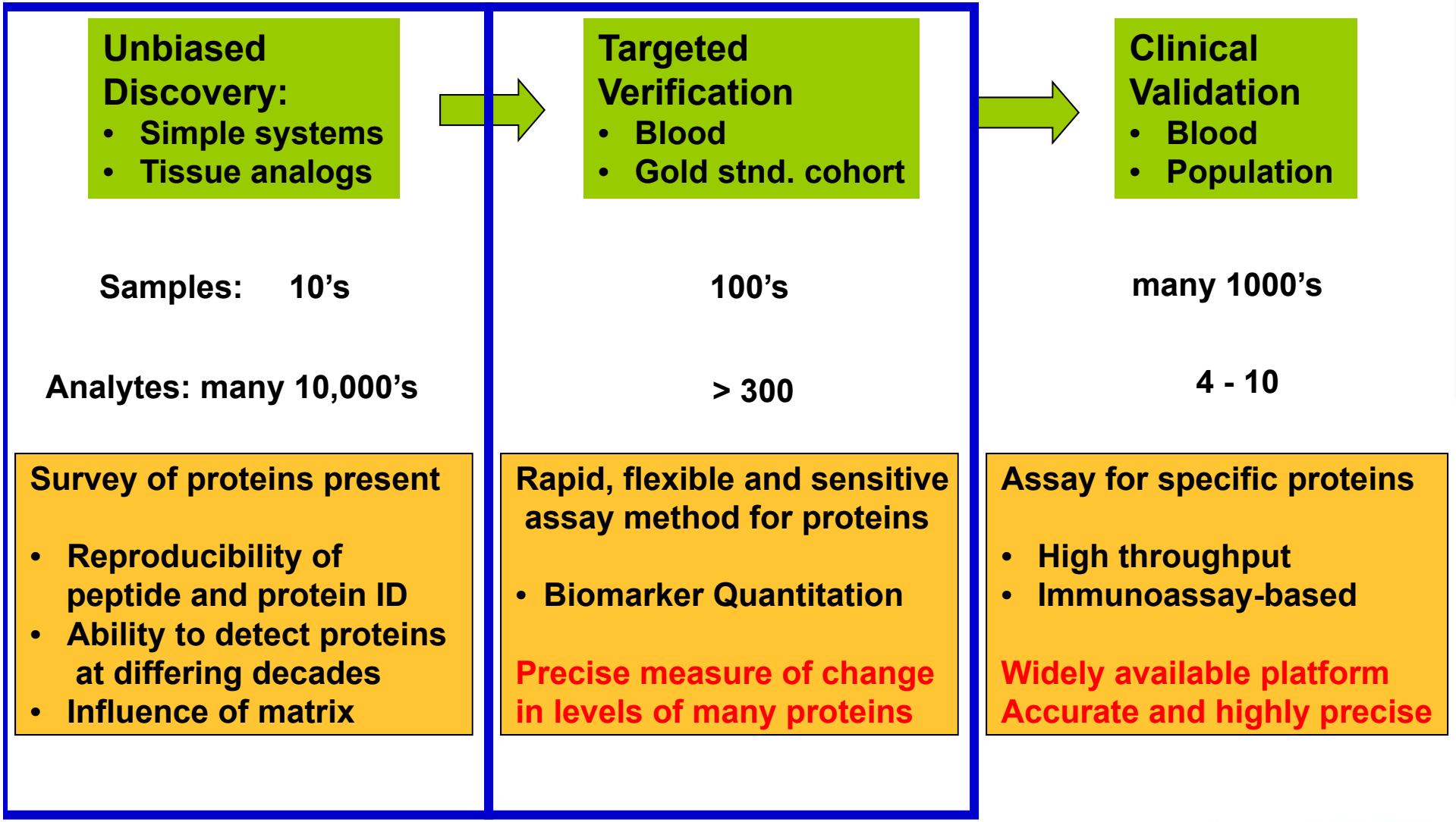
CPTAC's Response to Clinical Challenges in Proteomics

Oct. 2006: CPTAC teams begin to develop detailed plans for technology assessment of proteomics

Agree a process pipeline to test; identify key problems and design studies to address

- Unbiased discovery
- Verification

Assessing performance of key process steps in the test biomarker pipeline



Adapted from Rifai, Gillette and Carr Nat. Biotech.2006

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

What were the systems and technical barriers and how did we propose to address them?

- Inadequate supply of high quality biospecimens and clinical data
  - Develop and employ common sample collection methods to insure high quality samples for clinical phase
- Absence of coordinated systematic effort by expert proteomics labs focused on biomarker discovery
  - Five expert proteomics labs with cutting-edge technology formed highly integrated and collaborative consortium committed to sharing samples, data, methods, knowledge
- Multiple ad hoc data analysis methods
  - MS data analyzed through common pipelines using common DB with defined criteria for confidence assignment
  - Produce highly qualified raw and processed data sets and make publicly available

What were the systems and technical barriers and how did we propose to address them?

- Insufficient supply and quality of test samples and reagents
  - Develop common samples and reagents for CPTACs and community
- Enormous diversity, range, and dynamic nature of proteins to be measured
  - Selectively enrich and fractionate
  - Employ and develop targeted MS approaches
- Difficulty in measuring large number of features simultaneously
  - Experts employing state-of-the-art high-performance MS
  - Develop novel methods to recover information from MS scans

What were the systems and technical barriers and how did we propose to address them?

- Need robust method with sufficient throughput to reliably credential large numbers of candidates in blood
  - Develop bridging technology from discovery in tissues, etc. to quantitative assay in blood
  - Build and assess performance of multiplexed MS-based targeted assays to screen for and quantify candidates in patient plasma

- Apply assay methods to breast cancer patient samples for verification
  - Each CPTAC prospectively collecting plasma

Some Key Questions Being Addressed by the CPTACs

- What is the representation of proteins present in a sample that are detected at each decade (1 ug/mL; 100 ng/mL, etc.) in an Unbiased Discovery experiment?
- How reproducible are various Discovery platforms to detect true differences between samples (what is the FDR)?
- How reproducible, accurate and sensitive are Verification platforms?
- Do discovery and verification platforms require different measurement endpoints, different specifications on the same endpoints or both?
- What is the impact of matrix complexity on Discovery and Verification?

Design Principles

Use of common "spikes" and common matrices key

- NIST-provided 20 protein standard
- Commercial equimolar 48 protein mixture
- > 150 Heavy-isotope labeled cancer-specific proteins (Argonne National Labs)

Three matrices chosen to mimic increasingly complex biological backgrounds encountered in proteomics

- Yeast: non-human; ca. 6000 ORFs;
- Cell lysates: analog of tissue which proteomics community is increasingly using for discovery work instead of blood
- Plasma: large pool created for verification studies

Some Metrics of Reproducibility and Sensitivity for Proteomics Platforms

Reproducibility of seeing a protein, as a function of the protein's concentration, measured by:

- Protein detection and coverage (e.g., MS identification, LOD)
- Protein quantitation (with or without internal standards: CV's, LOQ)

Reproducibility of observing a statistically significant difference in protein level between two samples (e.g., case and control)

- Number of differences observed (number of proteins)
- Identities of differences observed
- Magnitude of differences observed
- Ideally determined as a function of the protein's concentration

Working Groups Established

Design of Experiments

- Unbiased Discovery
- Verification Studies
  - heavy-labeled peptides for quantitation
  - Abs

Selection and Production of Matrices

- Yeast; Cell lines; Plasma

Post-translational Modifications

- PTM standards (in discussions with vendors)
- Experimental designs to test detection of glyco-, phosphoproteins

Working Groups Established

Data Analysis, Storage and Dissemination WG

- Repository for raw and processed data
- Pipeline for analysis of all data by a standard method
- Defined databases to search

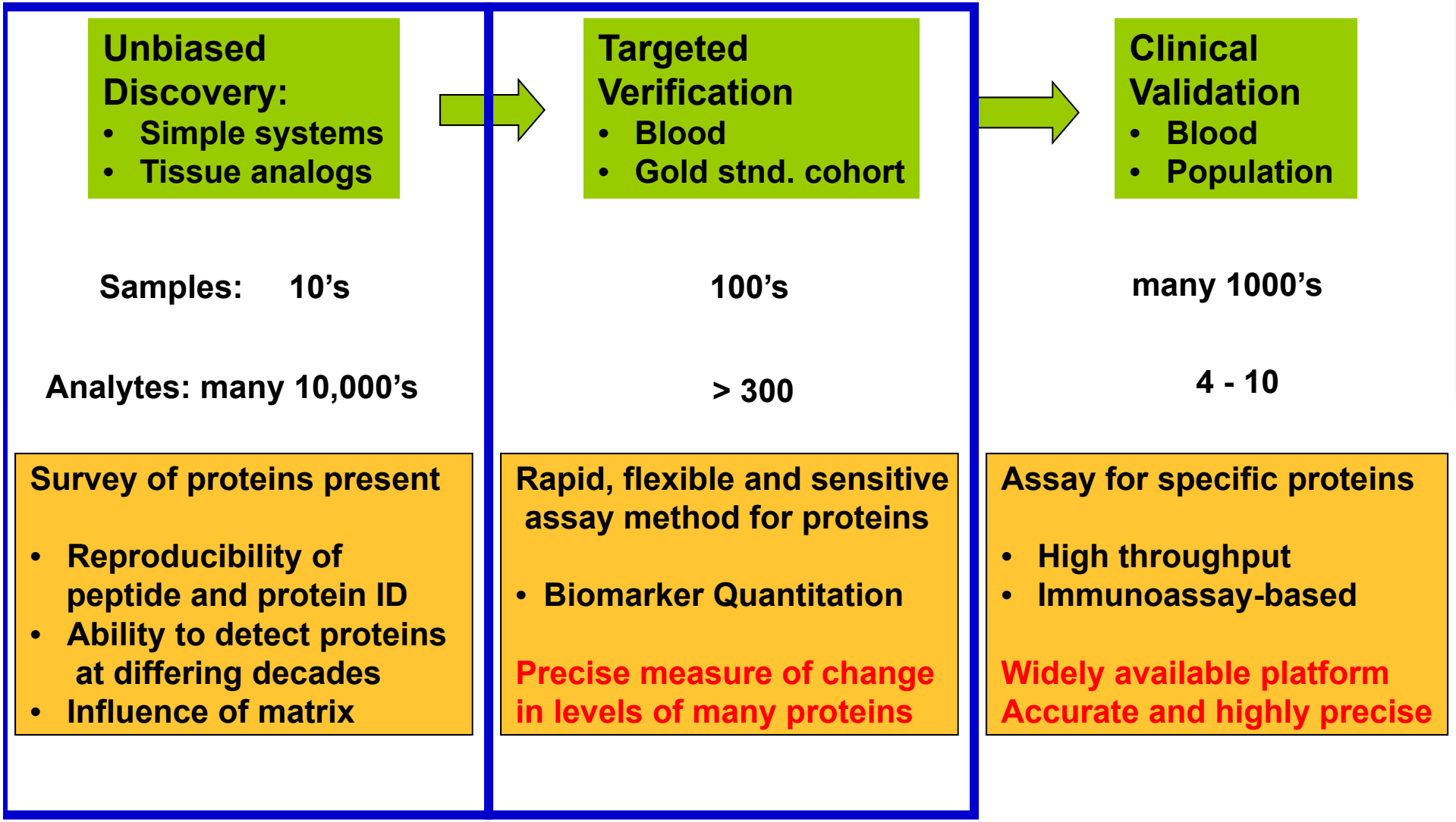
Biospecimens WG

- Collect, review and align sample collection protocols
- Develop mechanisms to maximize ability of groups to share samples
- Work toward use of standard collection protocol
- All CPTACs have clinical oncologists participating in their programs

All CPTAC groups represented and actively participating

- *Intra- and inter-WG teleconferences; face-to-face meetings*
- *Research studies involving all labs; analysis support from NIST*

Assessing performance of key process steps in the test biomarker pipeline



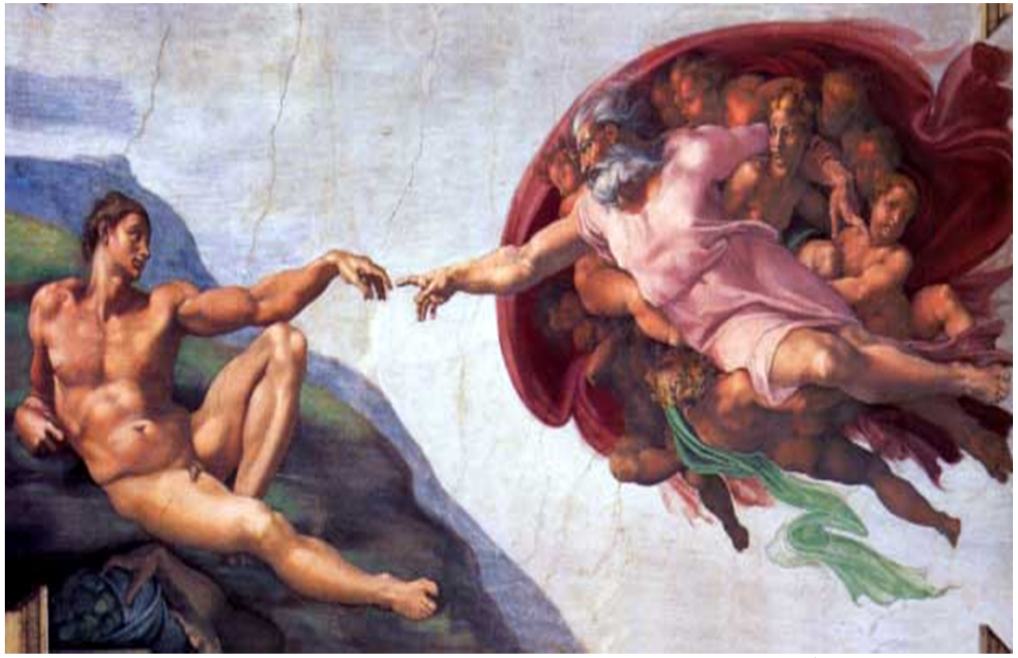
Adapted from Rifai, Gillette and Carr Nat. Biotech.2006

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

CPTAC Working Groups

- Unbiased discovery
- Targeted verification
- Data analysis, storage and dissemination
- Biospecimen collection
- Protein standards for verification
- Protein standards for PTM analysis
- Plasma standards
- Cell models

Shotgun proteome analysis

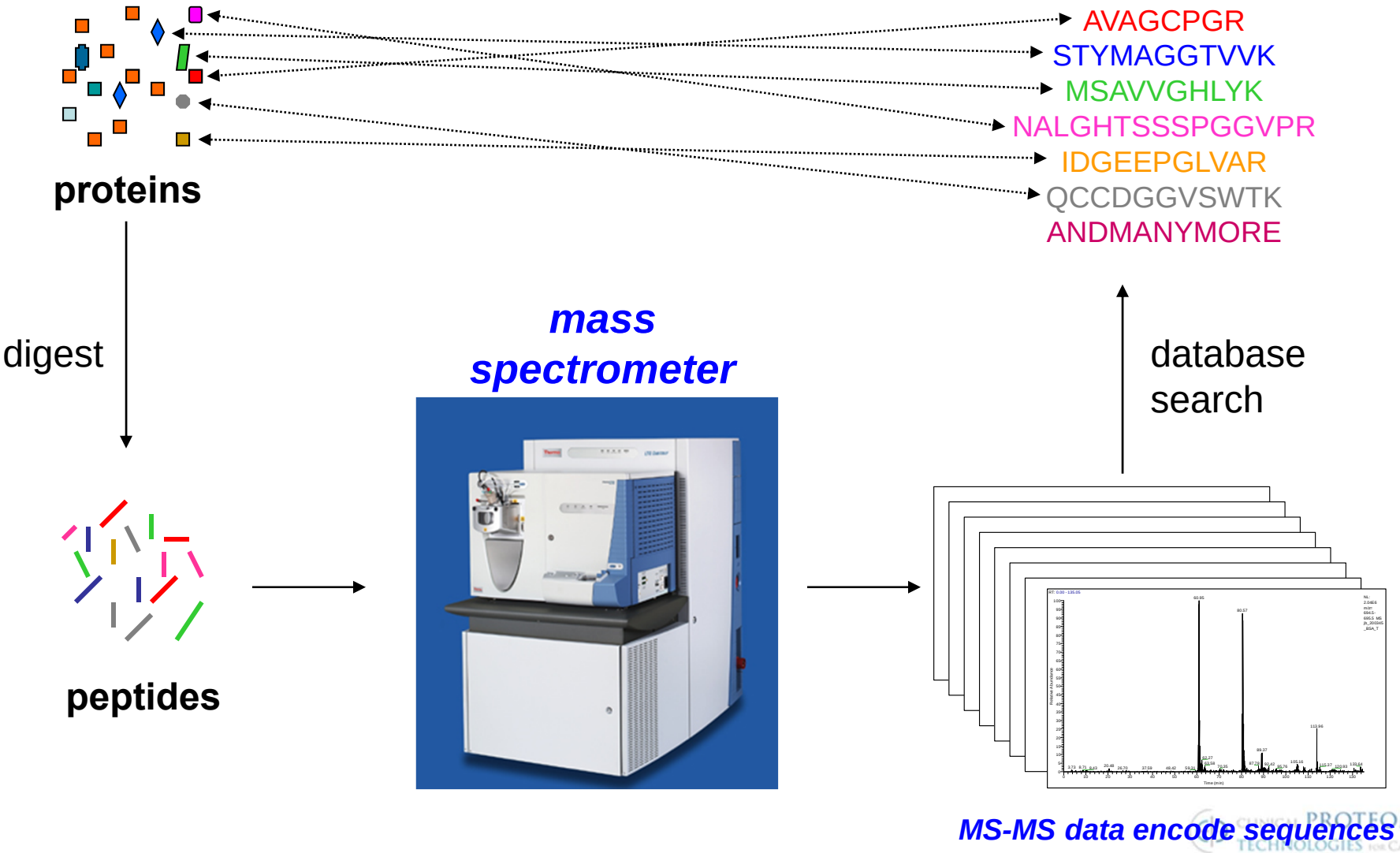


Intact Proteome

Digested Proteome

courtesy Stan Hefta, Bristol-Myers Squibb

Peptide sequences can be identified by tandem mass spectrometry (MS-MS)



Unbiased Discovery WG Goals

- Establish performance standards
  - standardized yeast proteome extract (NIST)
  - NCI-20 performance standard mix
- Develop SOPs for LC-MS-MS shotgun analyses
  - Joint development by NIST and CPTAC teams (1_07 - 10_07)
- Assess platform performance and develop QC metrics
  - detection limits, FDR, matrix effects, carryover, sources of intra- and inter-lab variation

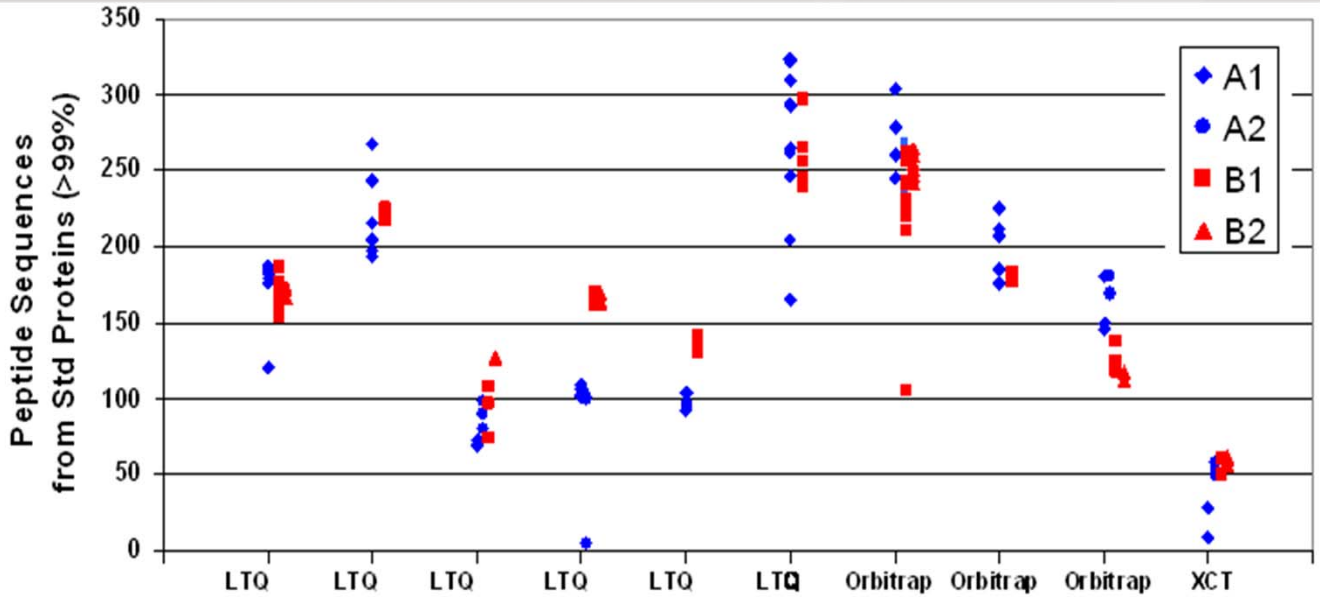
Unbiased Discovery WG Goals

- Assess detection efficiency for model biomarkers
  - yeast proteome extract spiked with 48 human protein mixture (Sigma) at concentrations equivalent to 10^2 to 10^5 copies/cell
- Implement standardized data analysis tools
  - peptide/protein ID pipeline (Vanderbilt)
  - data-sharing system (Tranche, Univ. Michigan)
- Significant inter-lab study publications

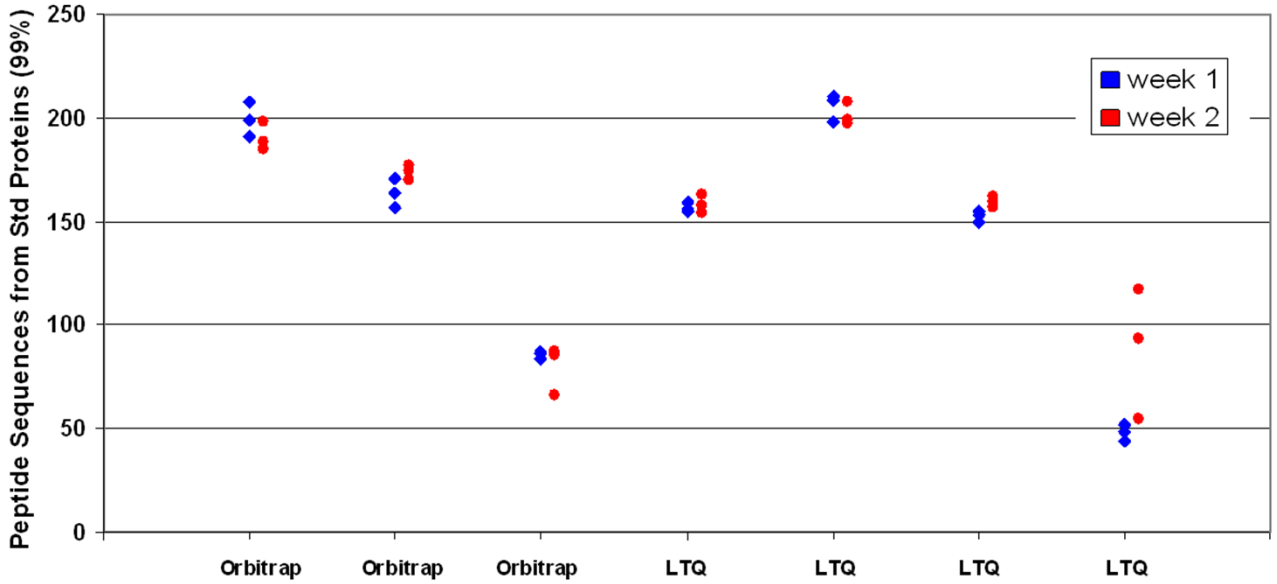
Unbiased Discovery Studies: status

Study	Design	Outcome
Study 1 **(November 2006)**	• **NCI-20 (20 human protein mix)** • **multiple instruments** • **no SOP**	• **high variability** - Between labs, instrument types, replicate analyses on same instrument
Study 2 **(February 2007)**	• **NCI20 mix** • **ion trap LC-MS instruments** • **initial SOP**	• **reduced variability** • **BUT, many variables not standardized**
Study 3 **(July 2007)**	• **yeast proteome** • **refined SOP**	• **further reduced variability** • **dynamic range estimates**
Study 5 **(November 2007)**	• **yeast proteome** • **finalized SOP**	• **lowest inter-lab CVs** • **yeast protein detection efficiency**
Study 6 **(February 2008)**	• **yeast proteome + 48 human protein spikes (concentration decades)** • **finalized SOP**	• **detection power across large concentration range in complex proteome**

Studies 1 and 2: NCI-20



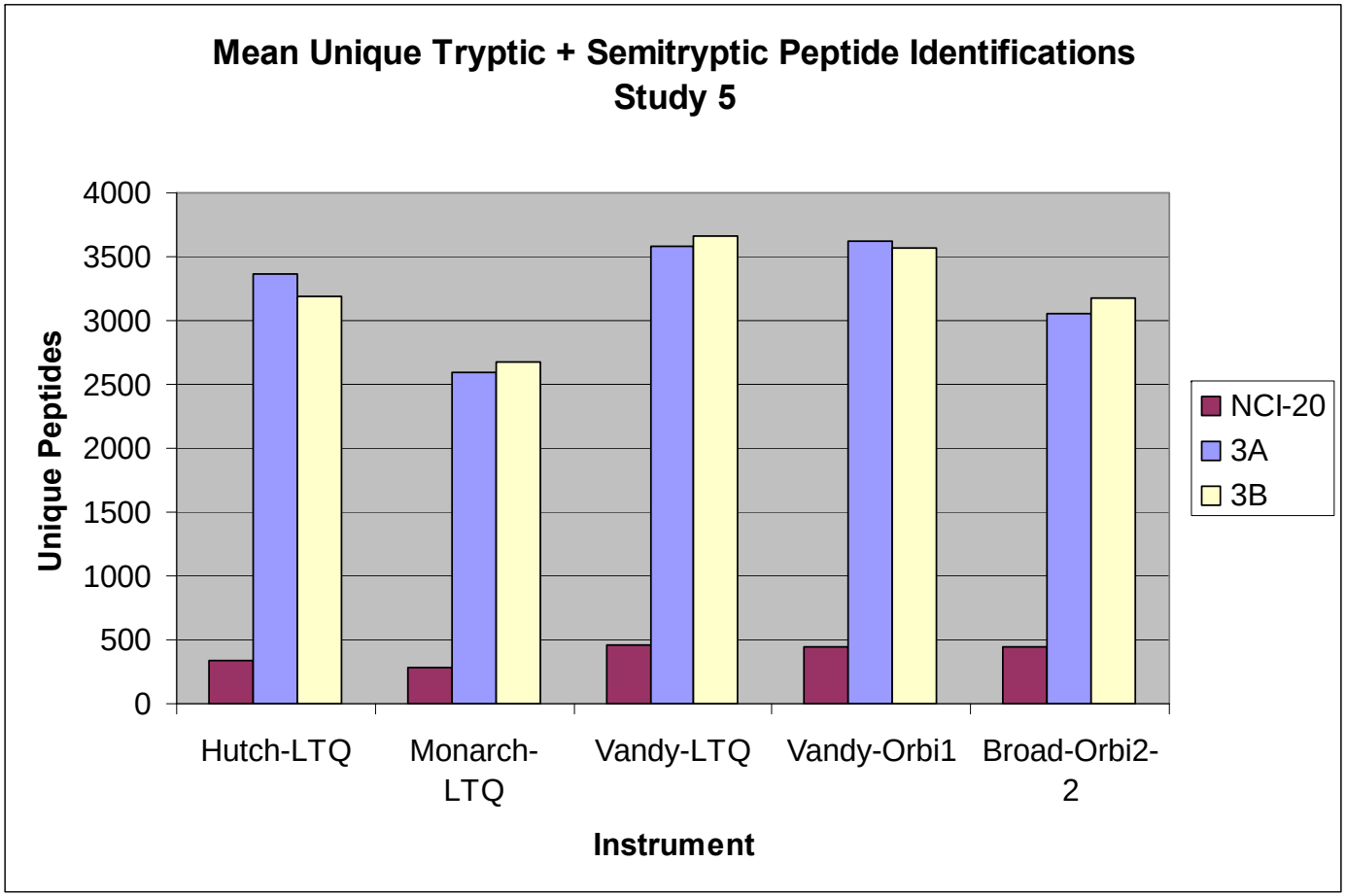
Study 1: No SOP



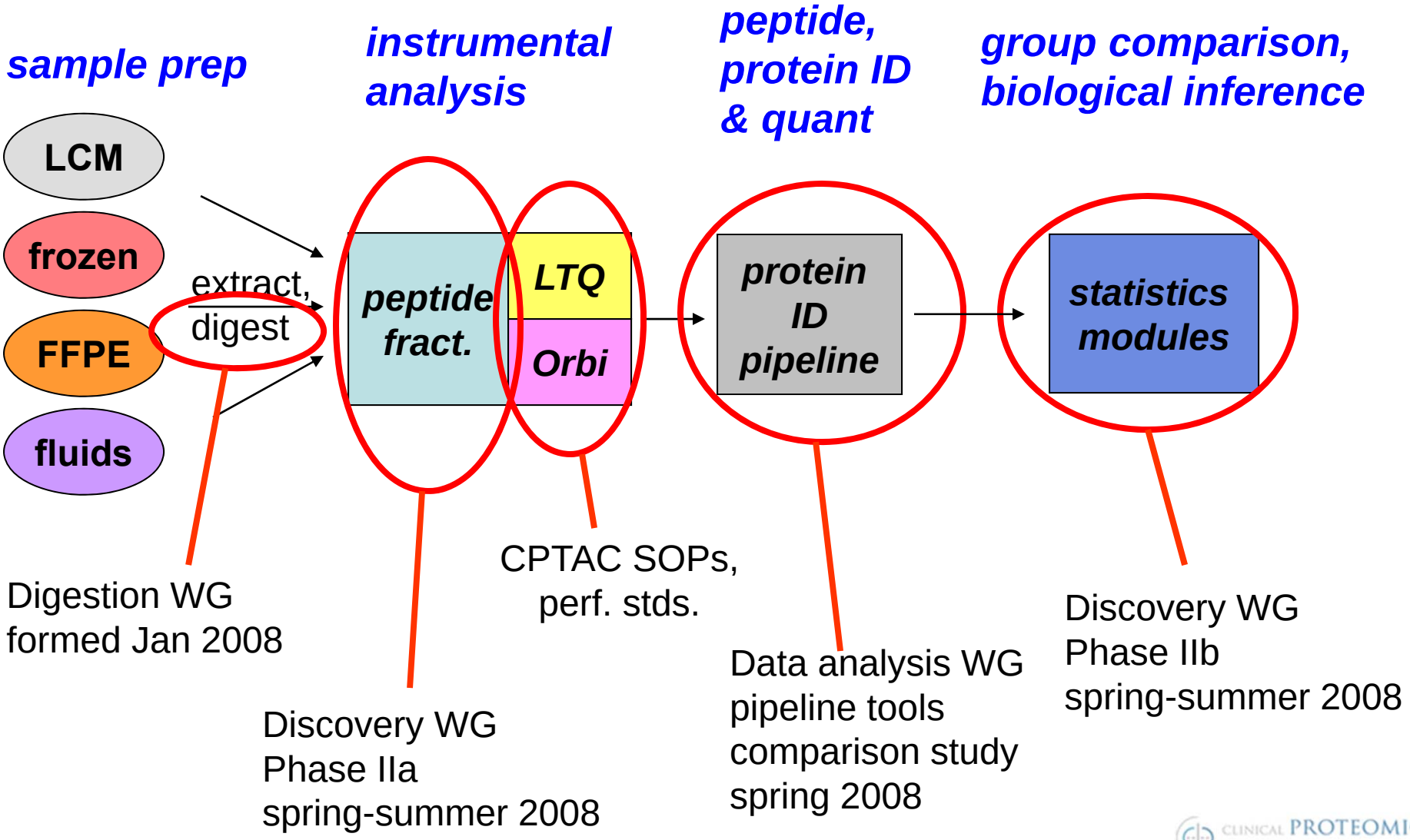
Study 2: Initial SOP

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

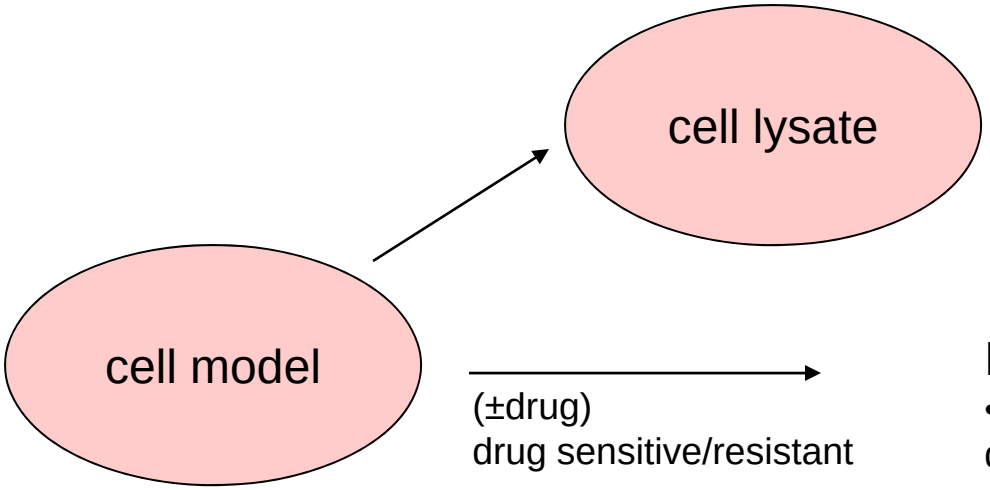
Study 5: Yeast proteome extract; finalized SOP



Systematic evaluation of discovery platforms: next steps



Cell models for marker discovery and verification



ANL
^{15}N proteins

Discovery platform
•detection efficiency in mammalian cell matrix

(±drug)
drug sensitive/resistant

Discovery platform
•consistent ID of protein differences between labs

•relevant phenotype
  •(±drug)
  •drug sensitive/resistant
•standardized culture conditions

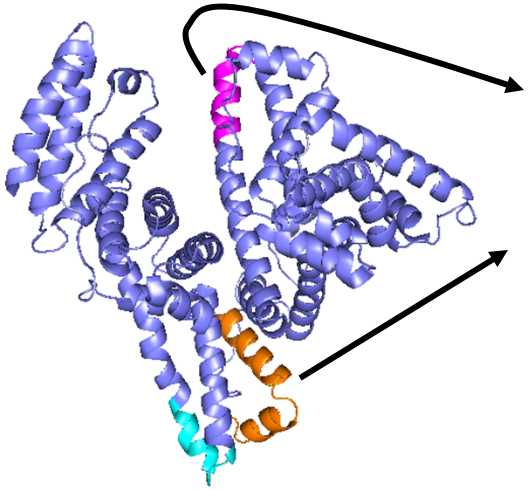
(±drug)
drug sensitive/resistant

Phosphoproteomics (e.g., pTyr intermediates related to EGFR signaling)
•standards and quality metrics for quantitative analyses in cells and tissues

Selection and production of protein standards: current status

- **Argonne National Laboratories to produce 500-1000 expressed, ^{15}N-labeled proteins for use as standards for targeted verification studies.**
- 1261 candidate cancer-related proteins on initial list
- 200-250 proteins to be expressed in mg quantities by 2009
- As of January 2008, 40 soluble proteins delivered in mg quantities; additional 110 in purification
- Critical reagents to spiking studies in plasma and cell model systems for verification and discovery studies

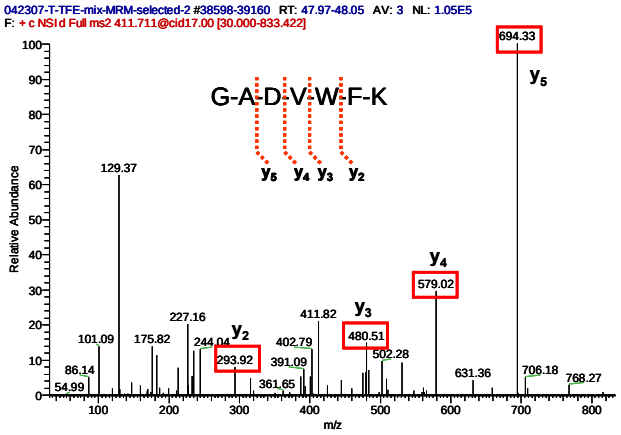
LC-MRM-MS for targeted protein quantitation

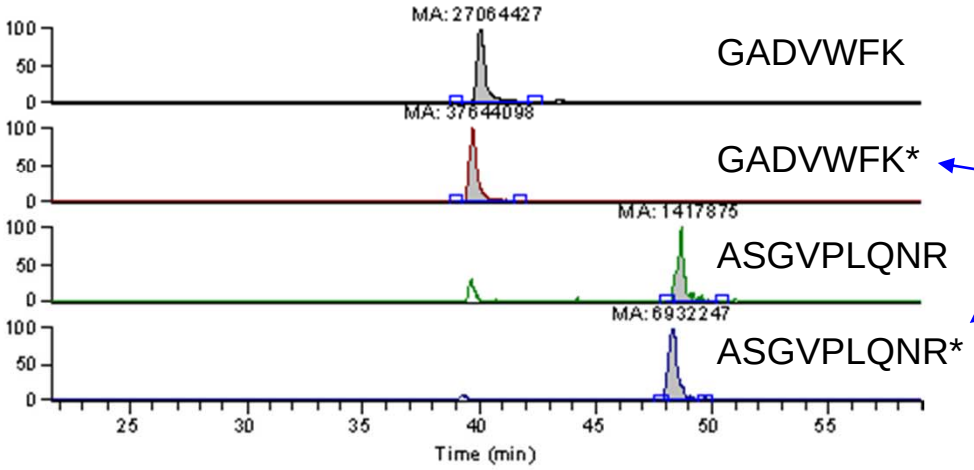


ASGVPLQNR

GADVWFK

select peptide-specific transitions





stable isotope internal standards

Absolute quantitation based on peak areas for target peptide and labeled standards

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Process for candidate verification in plasma by targeted MS

Candidate Protein Biomarkers

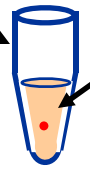
"Signature peptides" for candidate biomarkers
- Observed
- Defined *in silico*

Synthesize

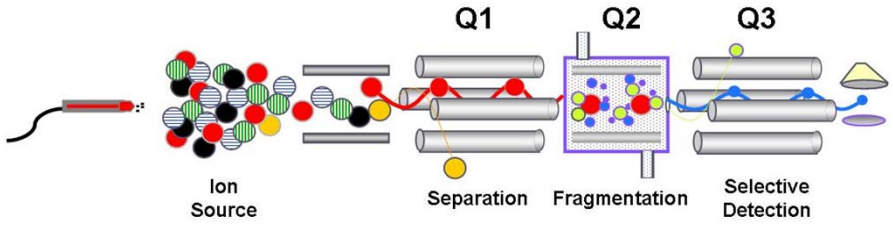
^{13}C-labeled & unlabeled signature peptides (characterized by LC-MS/MS)

Spike synthetic ^{13}C-labeled signature peptides

Endogenous ^{12}C signature peptides



+/- immunoaffinity peptide enrichment (SISCAPA)



Monitor signature peptide ratios
^{13}C peptide / ^{12}C peptide

Adapted from Rifai, Gillette and Carr Nat. Biotech. 2006

Targeted Verification WG Goals

- Establish a performance standard for LC-MS-MS-MRM system performance
  - Standardized human plasma containing 7 human proteins, provided by NIST
- Develop SOPs for comparison studies involving all CPTAC teams
  - Study SOP developed by NIST and CPTAC teams
- Generate dataset to develop metrics of LOD, LLOQ, accuracy (std) and precision (%CV) for an MRM based assay
- Identify and resolve sources of variation through inter-group studies and evolution of SOP
- Assess assay precision across all CPTAC sites
- Significant inter-lab study publications

Standard proteins for initial targeted verification experiment

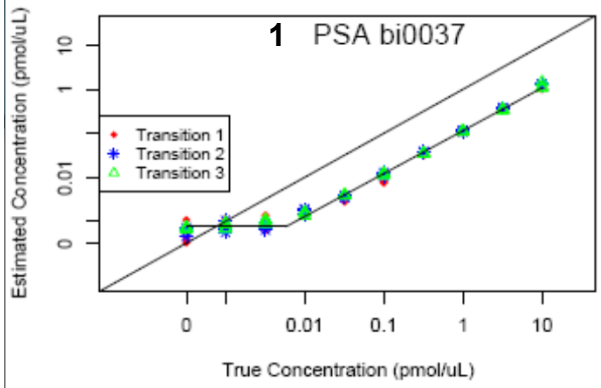
- Seven target proteins, five not found in human plasma and two found at low concentrations in human plasma, will be measured

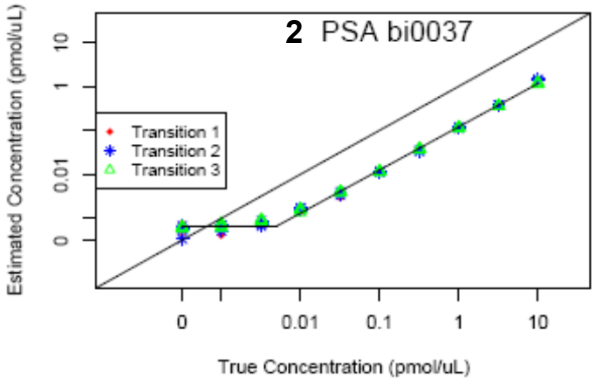
Target proteins and their ^{12}C signature peptides.

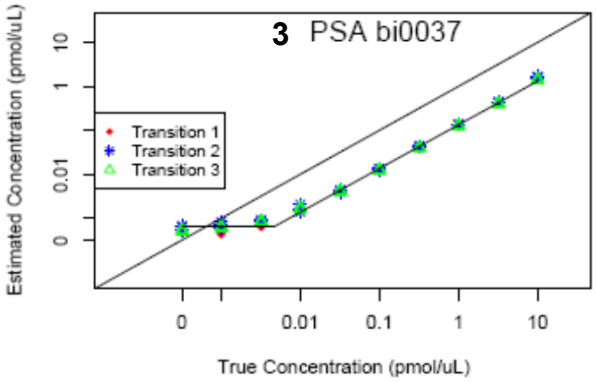
Protein	Species	Signature Peptide		
		Identifier	Sequence*	MH+ (mono)
prostate specific antigen (PSA)	human	bi0037	LSEPAELTDAVK	1272.67
		bi0161	IVGGWE<u>C</u>EK	1077.17
peroxidase	horse radish	bi0166	SSDLVALSGGHTFGK	1475.63
leptin	mouse	bi0167	INDISHTQSVSAK	1399.54
myelin basic protein (MBP)	bovine	bi0169	HGFLPR	725.86
		bi0170	YLASASTMDHAR	1322.47
myoglobin	horse	bi0171	LFTGHPETLEK	1271.45
aprotinin	bovine	bi0173	AGL<u>C</u>QTFVYGG<u>C</u>R	1488.61
C-reactive protein (CRP)	human	bi0231	ESDTSYVSLK	1128.54
		bi0202	GYSIFSYATK	1136.56
		ni0001	YEVQEVFTKPQLWP	1820.92

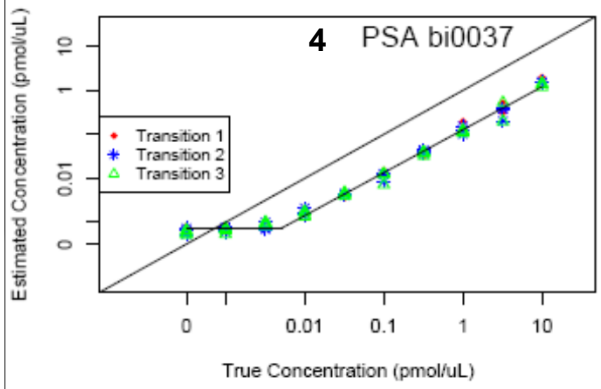
*cysteines have been carboxyamidomethylated

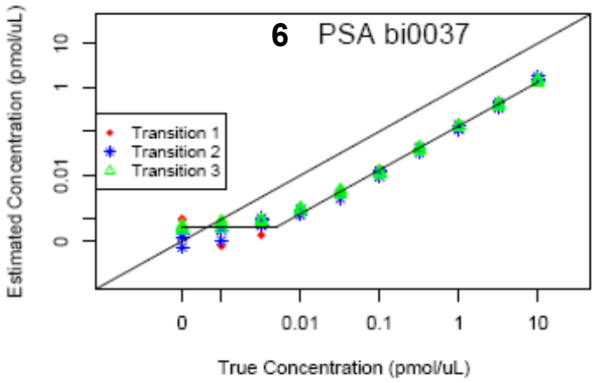
MRM Processed Data

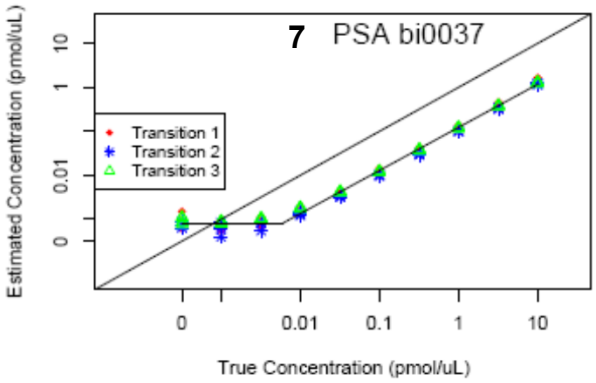


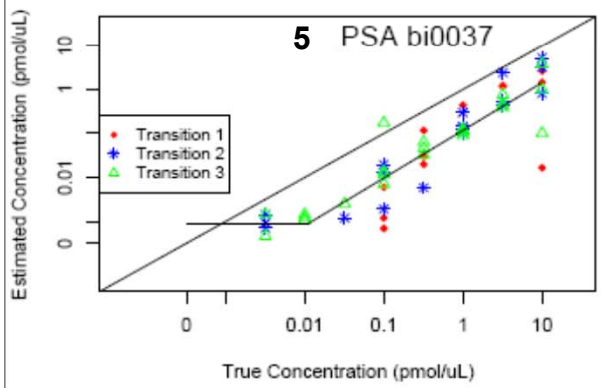












Expected Concentration = Area Ratio * 0.1

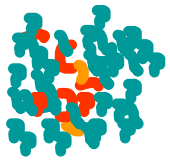
Area Ratio = Area Analyte / Area IS

Interlab verification studies: next steps

- Evaluate methods to increase MRM assay sensitivity to ng/mL range
  - Abundant protein depletion
  - Fractionation
- Determine degree to which assays can be multiplexed
- Test immunoaffinity enrichment as adjunct to MRM
  - Ab capture of protein
  - Ab capture of signature peptide (SISCAPA)
- Develop MRM assays for agreed list of breast cancer proteins
  - Using integrated genomics approach to prioritize (Candidate WG)
  - Define/predict best peptides for assay development
- Build reagent collection of peptides and Abs
  - 13C-labeled and unlabeled peptides
  - Anti-peptide Abs (rabbit polyclonal and monoclonal)

Targeted MS with Ab-capture for increased LOQ of Biomarkers from Plasma (SISCAPA[1])

Plasma-derived peptides



1. Add ^{13}C-labeled signature peptide

2. anti-peptide Ab capture

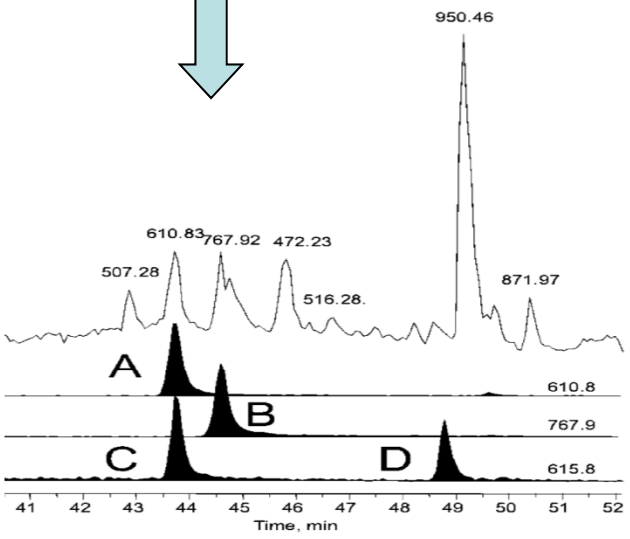
NFPSPVDAAFR

NFPSPVDAAFR

native (R = ^{12}C) and exogenous (R = ^{13}C) forms of peptide

Advantages over ELISA

- Only requires 1 Ab
- Ab need not recognize native protein
- Relaxed selectivity requirements for Ab
- Highly multiplexable



Enrich and decrease complexity: improves LOD and LOQ

1. Anderson et al. (2004) J. Prot. Res. 3: 235

PCC Membership

Program Coordinating Committee Members

- Steve Carr, Chair, Broad Institute of MIT and Harvard
- Susan Fisher, Co-Chair, UCSF
- Dan Liebler, Vanderbilt University
- Paul Tempst, MSKCC
- Fred Regnier, Indiana University
- Henry Rodriguez, NCI

Ad-hoc members

- Lee Hartwell, FHCRC
- Gordon Mills, M.D. Anderson Cancer Center
- Joe Gray, Lawrence Berkeley National Laboratory
- David Ransohoff, U. of North Carolina