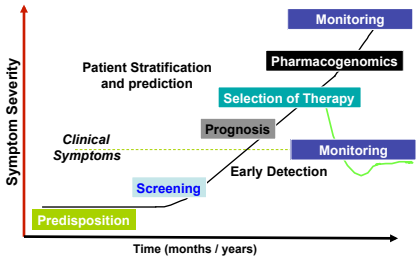
BROAD INSTITUTE

Proteomics-based Biomarker Discovery: Mirage or Emerging Reality?

Steven A. Carr
Broad Institute of MIT and Harvard

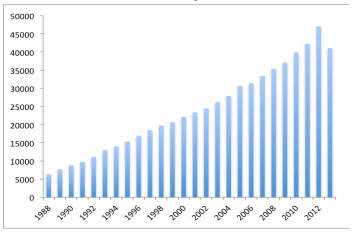
Biomarkers have tremendous clinical utility
Investment in new candidates has been vast…



Period	**Decades**
Number of tests	**> 10 million/yr**
Instruments	**Ca. 100,000 machines in hospitals, labs**
Accuracy	**CV ~5-10% worldwide at ≥100pg/ml**

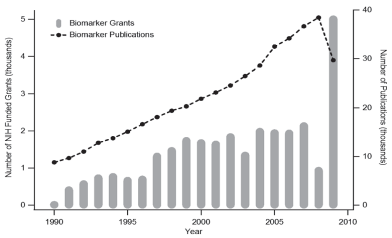
Cardiac damage	TnI, CK-MB, Mb, MPO, BNP
Cancer	PSA, CA-125, Her-2
Inflammation	CRP, SAA, cytokines, RF
Liver Damage	ALT, ALP, AST, GGT (enzyme assays)
Coagulation	AT-III, proteins C&S, fibrinogen, VWF
Allergy	IgE against various antigens
Infectious disease	HIV-1, Hepatitis BsAg

"Biomarker" Publications in Pubmed by Year



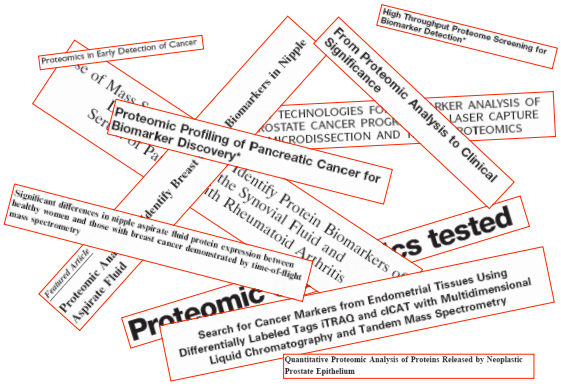
41,086 papers in 2013
> 370,000 in the past decade

NIH Biomarker Related Grants by Year



Pharmaceutical Intelligence 2014

…but the track record for translation of discoveries made using proteomics into clinical assays has been poor



=

Tower of Babel

- **Many proteins proposed as biomarkers but very few introduced into clinical practice (≤ 1/yr)**
- **Demonstrated successes for proteomics: 0**

Why? Can we do better?

Historical barriers to progress in proteomics-based BMD

- Absence of coordinated teams that included biostatisticians, clinicians and proteomics specialists has led to poor study design
- Few expert proteomics labs willing/able to focus on clinical sample analysis
  - perceived difficulties, long time horizons, poor reputation of field
  - Has resulted in many studies describing readily detectable, abundant proteins with no specific disease association
- MS-platforms inadequate for the task
  - Difficulty in repeatedly and precisely measuring large number of peptides/ proteins over >10^8 concentration range
  - Low number of patient samples used in Discovery – high FDR
- Multiple ad hoc, statistically indefensible data analysis methods used
- Need for methods to quantify large numbers of peptides/proteins from Discovery in 100's of patient samples
  - Must be robust, quantitative, highly multiplexed, sensitive, specific

There have been remarkable improvements in the practice and technologies of Proteomics over the past few years

- Appropriate study design
- Robust sample processing methods
- Quantitative, multiplexed labeling of peptides
- Data acquired with fast and sensitive high performance LC-MS/MS technology
- Statistically rigorous data analysis

Unprecedented definition of proteins in cells and tissues

- 10K – 12K distinct proteins
- Precise and reproducible

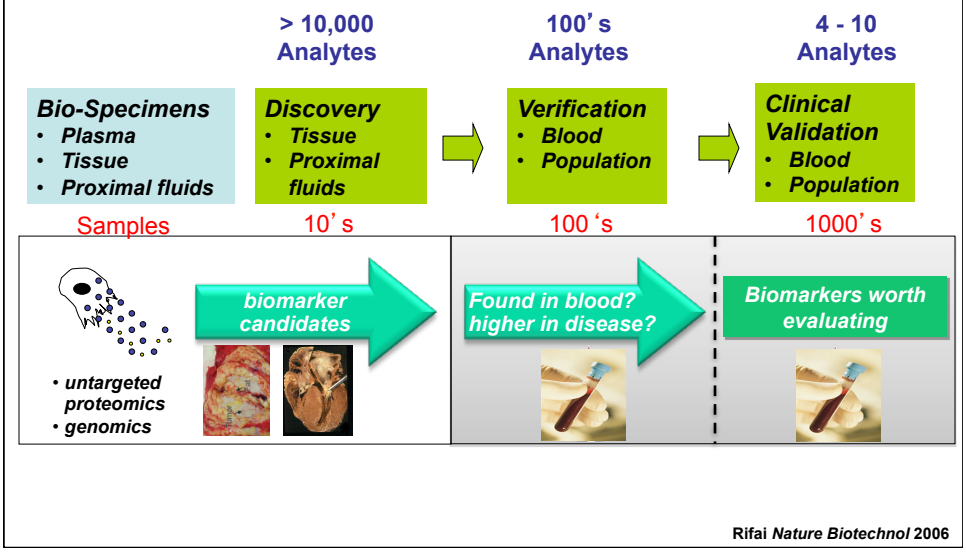
Deep and broad PTM coverage

- >30K phosphosites
- >20K ubiquitinated peps
- >8K acetylation sites

- **The number of proteins observed in tissues now begins to approximate the expressed proteome**
- **PTM analysis provide window into function and pathogenesis not accessible by genomic methods**

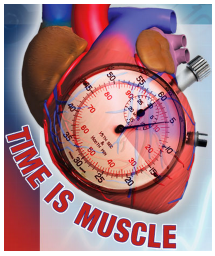
How well does this translate to biomarker discovery in biofluids?

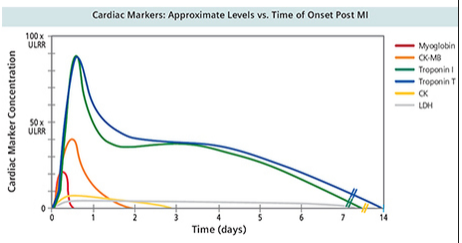
A functioning pipeline for biomarker development requires both *Discovery* and *Targeted* assay components



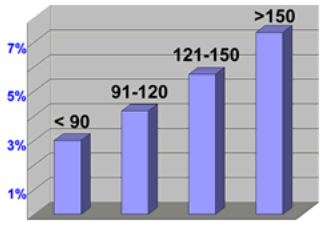
Rifai *Nature Biotechnol* 2006

Improved markers of early myocardial injury are needed

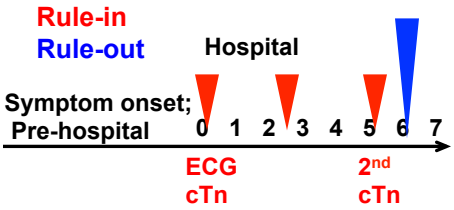




A minority of patients get angioplasty within the recommended 90 min

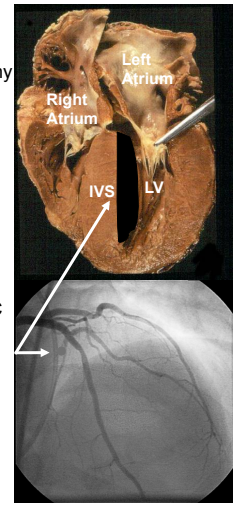


In-hospital mortality % by time (min.) from symptom onset to angioplasty



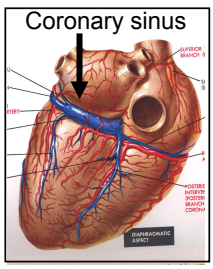
Plasma-based Discovery using a human model of myocardial injury (planned myocardial infarction)

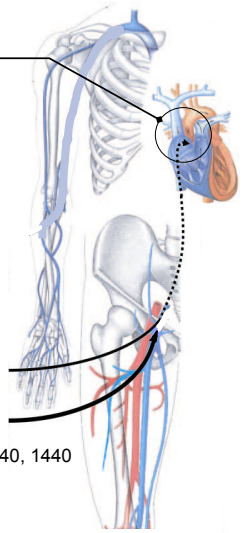
Hypertrophic Obstructive Cardiomyopathy (HOCM)



Planned therapeutic MI by alcohol ablation

PLASMA as a proximal fluid



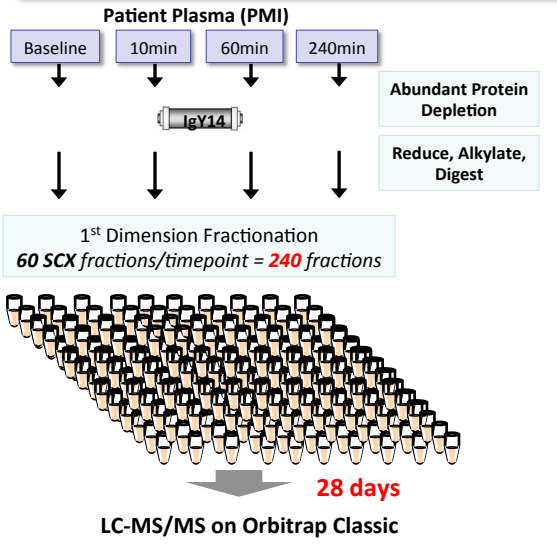


Coronary Sinus Samples
Time (min): Baseline, 10, 60

Femoral Vein Samples
Time (min): Baseline, 10, 60, 120, 240, 1440

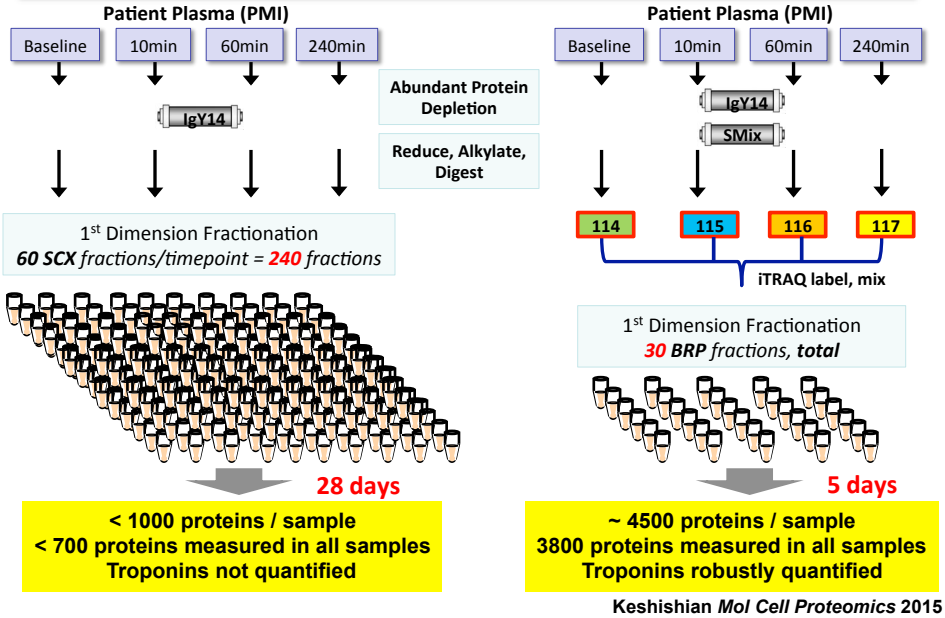
Collaboration with clinicians (Robert Gerszten, MGH), Clin Trials/Biostats experts (Marc Sabatine, BWH) and Proteomics Scientists in system where patients are their own controls

Original label-free approach was low throughput, yielded only modest depth of coverage with CVs of ca. 50%



Addona *Nature Biotechnol.* 2011

Optimized plasma processing has become at least 6X faster and 4X less expensive … and performs better



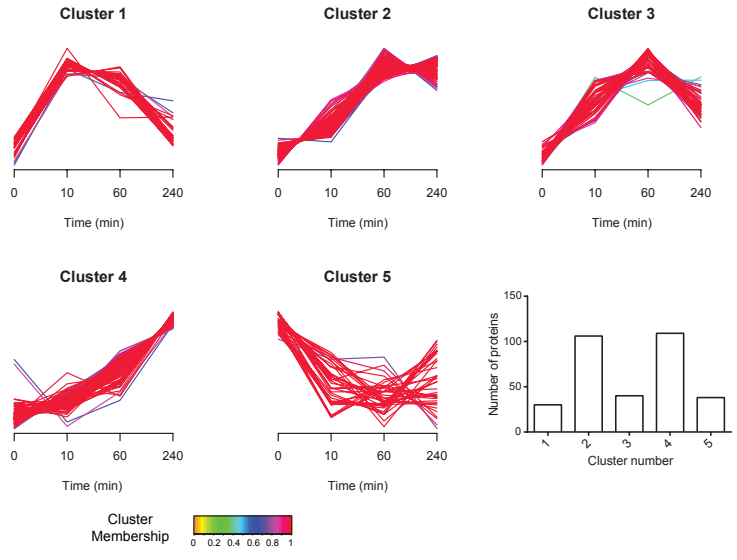
Keshishian *Mol Cell Proteomics* 2015

For the first time, we are seeing multiple peptides from the Troponins in the discovery plasma samples

		patient 1			patient 2			patient 3		
Protein Name	**sequence**	10min/BL	1hr/BL	4hr/BL	10min/BL	1hr/BL	4hr/BL	10min/BL	1hr/BL	4hr/BL
Troponin T, cardiac muscle	(K)DLNELQALIEAHFENR(K)	0.925	0.804	2.565	1.434	1.598	3.317			
Troponin T, cardiac muscle	(K)VLAIDHLNEDQLREK(A)	0.667	0.336	8.394				2.313	2.915	9.159
Troponin T, cardiac muscle	(K)ELWQSIYNLEAEKFDLQEK(F)	0.447	0.758	7.909						
Troponin T, cardiac muscle	(K)ELWQSIYNLEAEK(F)	0.997	1.099	5.074						
Troponin T, cardiac muscle	(K)YEINVLR(N)	0.737	0.544	4.653						
Troponin T, cardiac muscle	(R)KVLAIDHLNEDQLR(E)	0.612	1.004	2.011						
Troponin T, cardiac muscle	(K)IPDGERVDFDDIHRK(R)	1.039	1.012	4.887						
Troponin T, cardiac muscle	(K)FDLQEK(F)	0.998	0.742	5.813						
Troponin T, cardiac muscle	(K)EAEDGPMEESKPK(P)				4.955	12.097	16.693			
Troponin T, cardiac muscle	(K)DLNELQALIEAHFEnR(K)				2.01	3.273	6.255			
Troponin T, cardiac muscle	(K)VLAIDHLNEDQLR(E)								1.648	5.462
Troponin I, cardiac muscle	(R)CQPLELAGLGFAELQDLCR(Q)	1.019	0.749	9.528				0.721	1.345	2.503
Troponin I, cardiac muscle	(R)CQPLELAGLGFAELQDLCR(Q)	0.239	1.059	3.935						
Troponin I, cardiac muscle	(K)NITEIADLTQK(I)	0.349	0.639	8.649				1.296	1.982	7.403
Troponin I, cardiac muscle	(R)VDKVDEERYDIEAK(V)	0.634	0.874	3.569						
Troponin I, cardiac muscle	(R)EVGDWRK(N)	1.119	2.47	6.892						
Troponin I, cardiac muscle	(K)IFDLR(G)	0.606	0.592	6.617				4.438	4.87	14.579
Troponin I, cardiac muscle	(R)ISADAMMQALLGAR(A)	1.027	0.823	2.406						
Troponin I, cardiac muscle	(K)TLLLQIAK(Q)	0.574	0.361	5.574				1.88	2.366	6.093
Troponin C, slow skeletal and cardiac muscles	(K)NADGYIDLDELK(I)	1.443	2.725	27.039	2.062	1.877	2.209	5.056	5.884	8.78
Troponin C, slow skeletal and cardiac muscles	(K)AAVEQLTEEQKNEFK(A)	0.859	1.268	10.293				3.154	2.907	5.407
Troponin C, slow skeletal and cardiac muscles	(K)NADGYIDLDELK(I)	1.301	1.477	14.463	1.707	1.664	2.108	2.001	2.345	3.986
Troponin C, slow skeletal and cardiac muscles	(K)AAVEQLTEEQKNEFK(A)	1	0.959	5.006						
Troponin C, slow skeletal and cardiac muscles	(K)AAVEQLTEEQK(N)	0.894	0.916	4.077	4.695	3.992	12	1.193	1.517	2.446
Troponin C, slow skeletal and cardiac muscles	(K)GKSEEELSDLFR(M)	0.884	1.076	2.528				1.801	2.428	3.301
Troponin C, slow skeletal and cardiac muscles	(K)NADGYIDLDELK(I)	0.774	0.904	4.805						
Troponin C, slow skeletal and cardiac muscles	(K)AAFDIFVLGAEDGCISTK(E)	0.791	0.764	1.477				0.967	1.141	1.179
Troponin C, slow skeletal and cardiac muscles	(K)IMLQATGETITEDDIEELMK(D)				2.148	2.705	3.491			
Troponin C, slow skeletal and cardiac muscles	(R)IDYDEFLEFMK(G)				16.318	12.719	30.612	7.254	7.364	17.083

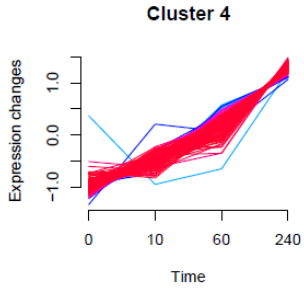
Keshishian *Mol Cell Proteomics* 2015

333 regulated proteins identified with a range of temporal behaviors



Keshishian *Mol Cell Proteomics* 2015

Cluster 4: Continuous risers
*Total of **99** proteins (Cont.)*

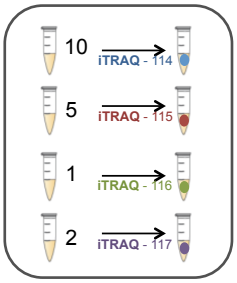


UV excision repair protein RAD23 homolog B
Puromycin-sensitive aminopeptidase
Methionine adenosyltransferase 2 subunit beta
Proteasome subunit alpha type-2
Proteasome subunit beta type-4
Heat shock 70 kDa protein 1A/1B
Pterin-4-alpha-carbinolamine dehydratase
Isoform 4 of FGGY carbohydrate kinase domain-containing protein
Citrate synthase, mitochondrial
cDNA FLJ59758, highly similar to S-methyl-5-thioadenosine phosphorylase (EC 2.4.2.28)
Proteasome subunit beta type-5
Proteasome subunit beta type-3
Glutathione S-transferase Mu 2
Proteasome subunit beta type-7
Isoform M1 of Pyruvate kinase isozymes M1/M2
Carbohydrate kinase domain-containing protein
Succinyl-CoA:3-ketoacid-coenzyme A transferase 1, mitochondrial
Xaa-Pro aminopeptidase 1
Proteasome subunit beta type-6
Pyruvate kinase isozymes M1/M2
5'-nucleotidase domain-containing protein 1
Farnesyltransferase, CAAX box, alpha, isoform CRA_a
Secernin-2
Glutathione S-transferase Mu 3
Microtubule-associated protein 4
Phospholipid hydroperoxide glutathione peroxidase, mitochondrial
Estradiol 17-beta-dehydrogenase 8
Stress-70 protein, mitochondrial
5'(3')-deoxyribonucleotidase, cytosolic type
Serine/threonine-protein phosphatase 5
UPF0364 protein C6orf211
Troponin C, slow skeletal and cardiac muscles
Troponin T, cardiac muscle
Troponin I, cardiac muscle
14-3-3 protein sigma
Cofilin-2
NEDD8
cDNA FLJ54408, highly similar to Heat shock 70 kDa protein 1
LIM domain-binding protein 3
GTP cyclohydrolase 1 feedback regulatory protein
BAG family molecular chaperone regulator 3
GSTT2 protein
Glucosamine 6-phosphate N-acetyltransferase
Beta-enolase
Mannose-6-phosphate isomerase
cDNA, FLJ93371, highly similar to Homo sapiens retinoid X receptor, gamma (RXRG), mRNA
Uncharacterized protein C7orf43
Probable aminopeptidase NPEPL1
Cytochrome P450 1B1

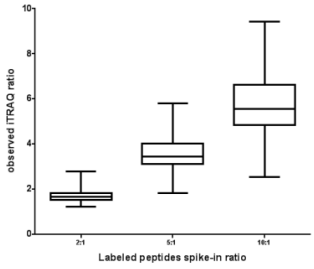
Keshishian *Mol Cell Proteomics* 2015

Ratios are compressed, but differential peptides/ proteins are still readily and reproducibly detected

Spike in 97 heavy peptides



Spike-in ratio	**2:1**	**5:1**	**10:1**
Observed average ratio	**1.67**	**3.44**	**5.54**
CV (%) of the ratio	14.0	22.4	25.7
Median CV (%) of 3 separate iTRAQ experiments	23.8	20.7	16.4



- Ratios are compressed 17 to 45% in plasma, similar to what we observe in cells and tissues (Mertins Mol Cell Proteomics 2011)
- Reproducibility of ratios observed ranged from 16 to 24%

Keshishian *Mol Cell Proteomics* 2015

Discovery defines a reduced set of "sentinel" marks that need to be repeatedly measured in a range perturbations

Perturbations:
- Disease
- Development
- Drug
- KO/KI

Analyte Valley of Death

Past: Westerns; Immunoassays

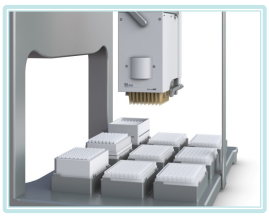
Not all proteins and PTMs of interest observed in all experiments

Precisely measure selected analytes in all experiments – no missing data!

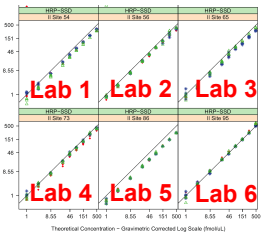
Desired assay properties:
- Highly specific
- Sensitive
- Highly precise
- Multiplexed
- Interference-free

Targeted MS (MRM, PRM) with labeled internal standards is specific, precise, reproducible, robust, and can be highly multiplexed

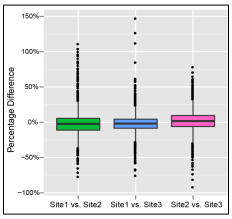
Automated Sample Processing



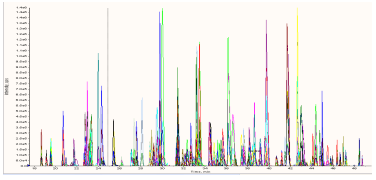
Precise and Reproducible



Robust



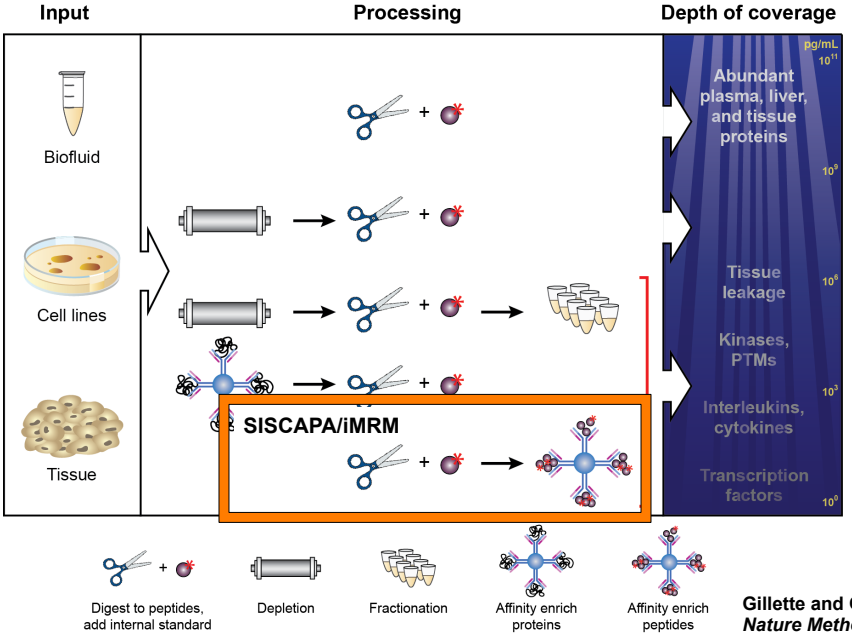
High Multiplex and Information Content



400-plex MRM assay; single 3h run

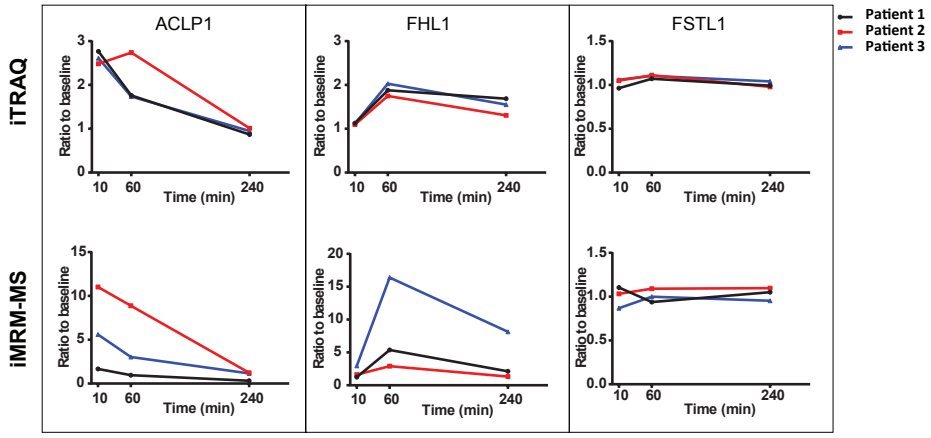
Numerous, well documented Studies
- Addona (2009) *Nature Biotech*
- Whiteaker (2011) *Mol Cell Proteomics*
- Addona (2011) *Nature Biotech*
- Kuhn (2011) *Mol Cell Proteomics*
- Hüttenhain (2012) *Sci Transl Med*
- Kennedy (2013) *Nature Methods*
- Keshishian (2014) *Mol Cell Proteomics*

Enrichment methods increase limits of detection



Gillette and Carr *Nature Methods*, 2013

iTRAQ-based quantification of peptides generally correlated well with iMRM results

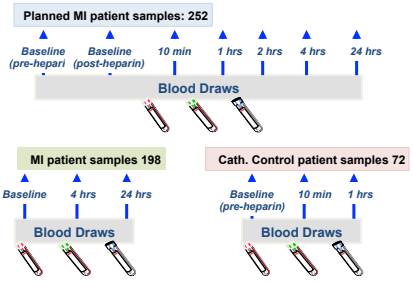


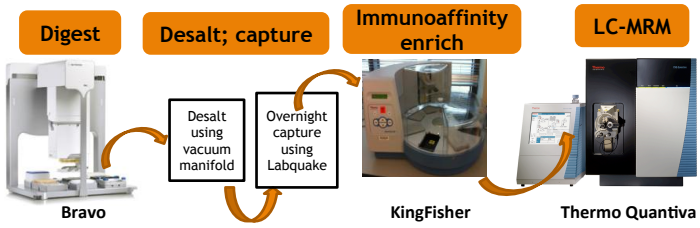
Keshishian *Mol Cell Proteomics* 2015

47-plex immunoMRM assay for CV disease biomarker candidates used to assay >650 patient samples in 3 months

Protein	Sequence	Protein	Sequence
MDH1	FVEGLPINDFSR	FSTL1	IQVDYDGHCK
HSPA8	[illegible]DPVEK	SCUBE2	GSVACECRPGFELAK
	[illegible]ETDTER	SPON1	AQWPAWQPLNVR
DPP3		FGL2	EEINVLHGR
		GLO1	GFGHIGIAVPDVYSACK
THSD[illegible]		GSTP1	ASCLYGQLPK
		[illegible]IPEPPS	VLGATLLPDLIQK
ESM1			IDYGVFAK
	ETCN[illegible]		[illegible]SLTFEPLTLVPIQTK
ANGPTL4	DCQELFQV[illegible]		[illegible]ALNAWR
	QSGLFEIQPQGSF[illegible]		[illegible]LDETLK
MIF	LLCGLLAER		[illegible]
FLT1	NVYTGEEILQK		
	YLAVPTSK	EZR	
TNNT2	VLAIDHLNEDQLR	CLIC1	
LAP3	LFEASIETGDR	PCNA	LV[illegible]
	GSPNANEPPLVFVGK	CTSL1	VFQEPL[illegible]
	TIQVDNTDAEGR	DKK3	DQDGEILLPR
TNNI3	NITEIADLTQK	VCAN	LGEPNYGAEIR
ACLP1	ILNPGEYR	VWF	EAPDLVLQR
	DTPVLSELPEPVVAR		
FHL1	AIVAGDQNVEYK		
	NPITGFGK		
MYL3	AAPAPAPPPEPERPK		
	ALGQNPTQAEVLR		
IL 33	TDPGVFIGVK		
	VLLSYYESQHPSNESGDGVDGK		

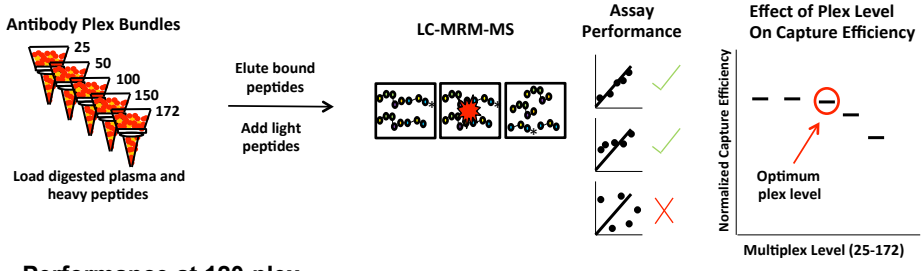
LOQ range in 47-plex iMRM assay: 3 -230ng/mL Median CV: 15%



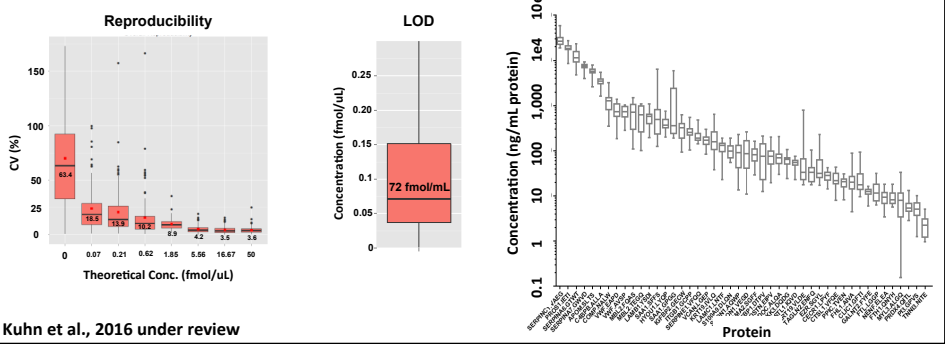


Semi-automated patient sample processing and analysis

We have extended immunoMRM to >150-plex

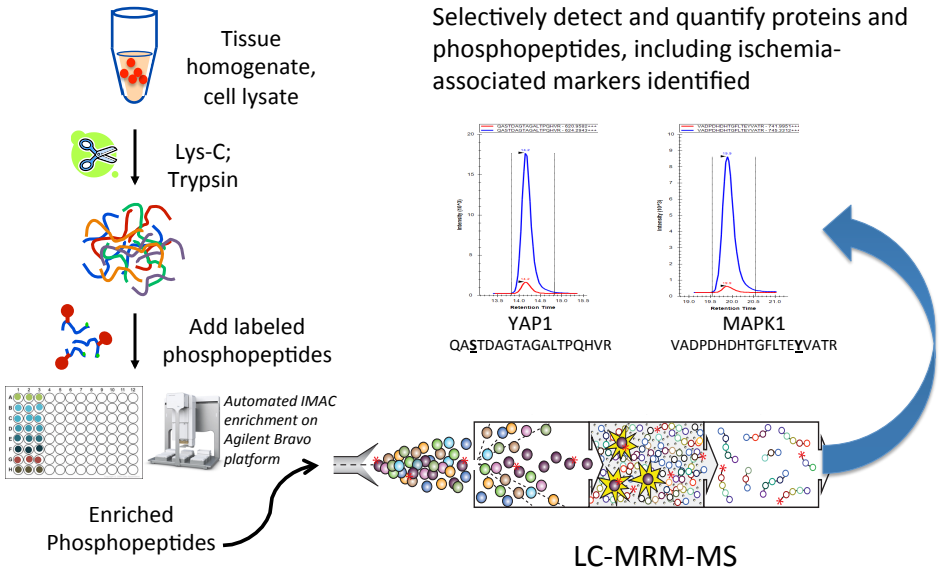


Performance at 120-plex

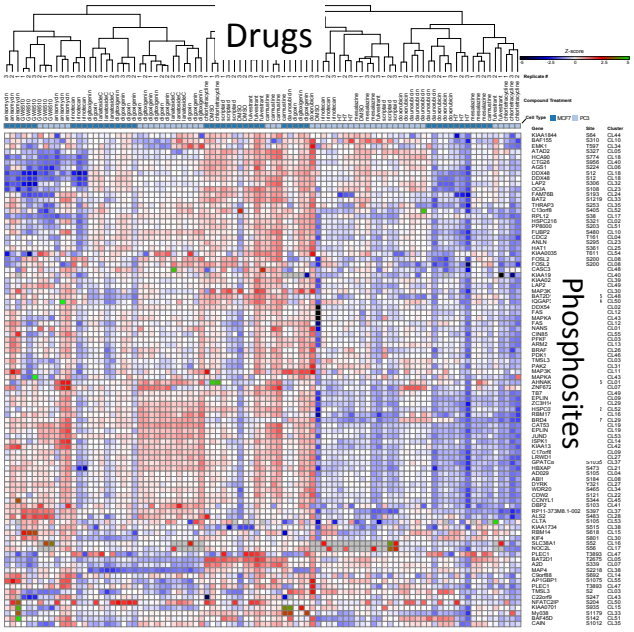


Kuhn et al., 2016 under review

Develop targeted-MS peptide and phosphopeptide marker panels



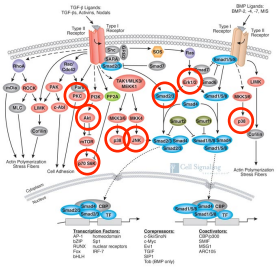
Automated, 100-plex targeted MS phosphopeptide assay



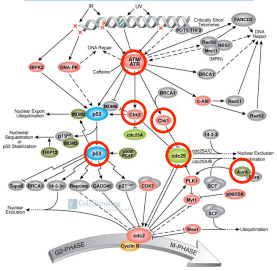
Abelin *Mol Cell Proteomics* 2016

Monitoring cell signaling by targeted MS of phosphopeptides

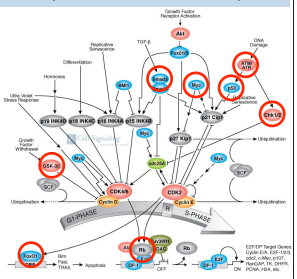
TGFB Signaling



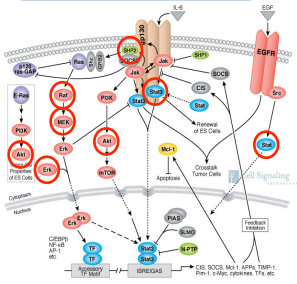
Cell Cycle Control: G2/M DNA Damage Checkpoint



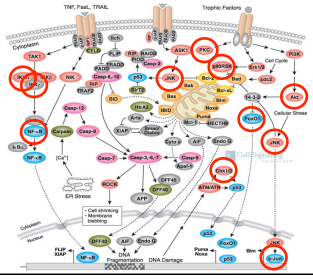
Cell Cycle Control: G1/S Checkpoint



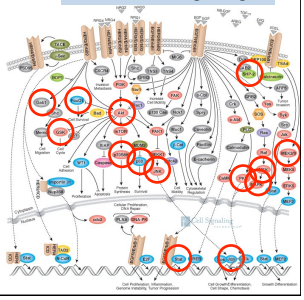
JAK / STAT Signaling



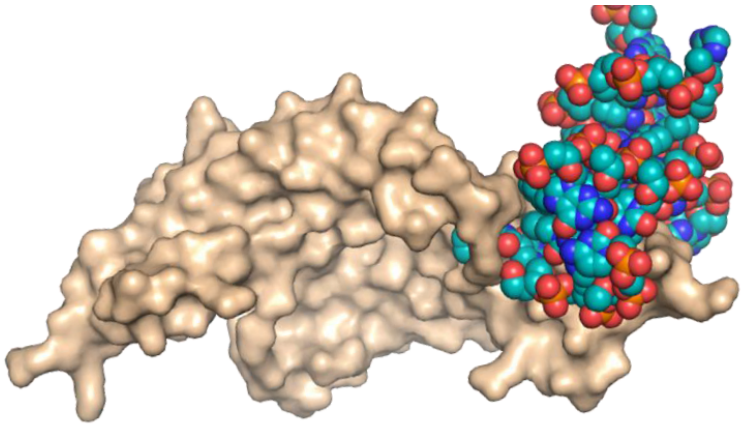
Regulation of Apoptosis



Erb B / HER Signaling

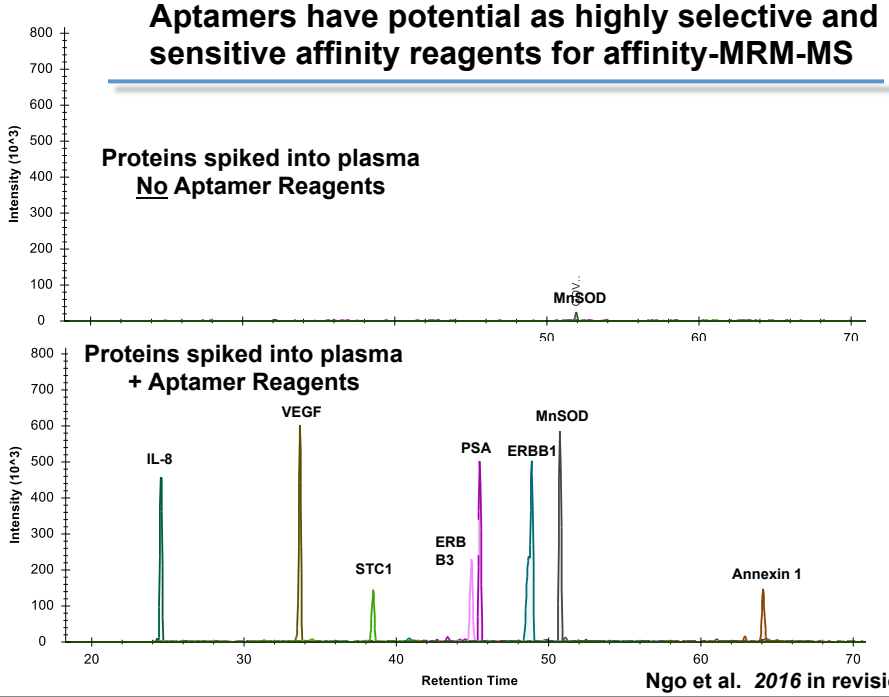


Evaluating Aptamers as affinity reagents for protein enrichment and quantification by targeted mass spectrometry



- Single-stranded DNA; 40 - 60 nucleotides in length
- Evolved using SELEX to obtain aptamers with very slow-off rate

Aptamers have potential as highly selective and sensitive affinity reagents for affinity-MRM-MS



Ngo et al. *2016* in revision

Availability of well validated MS-based assays for proteins and PTM's will help to alleviate the "reproducibility crisis"

NIH plans to enhance reproducibility

Francis S. Collins and **Lawrence A. Tabak** discuss initiatives that the US National Institutes of Health is exploring to restore the self-correcting nature of preclinical research.

"Efforts by the NIH alone will not be sufficient to effect real change in this unhealthy environment."

REPRODUCIBILITY OF RESEARCH FINDINGS

Preclinical research generates many secondary publications, even when results cannot be reproduced.

Journal impact factor	Number of articles	Mean number of citations of non-reproduced articles*	Mean number of citations of reproduced articles
>20	21	248 (range 3–800)	231 (range 82–519)
5–19	32	169 (range 6–1,909)	13 (range 3–24)

Results from ten-year retrospective analysis of experiments performed prospectively. The term 'non-reproduced' was assigned on the basis of findings not being sufficiently robust to drive a drug-development programme.
*Source of citations: Google Scholar, May 2011.

Establishing Best Practices for reproducible and reliable protein measurements and creation of community resources

Targeted Peptide Measurements in Biology and Medicine: Best Practices for Mass Spectrometry-based Assay Development Using a Fit-for-Purpose Approach*

Developed at an NIH-sponsored Workshop with participants from:

- Pharma, Clinical Labs, IVD companies, FDA, AACC, Biotech, Journals

Outcomes:

- Recommended criteria for 3 Tiers of assay validation
- Development of publication guidelines

CPTAC Assay Portal: using Tier 2 validation criteria for assay acceptance

https://assays.cancer.gov/

* Carr et al. *Mol Cell Proteomics* 2014

Conclusions

- Clinical proteomics begins with "Clinical" – invest in defining the question or need and finding the right samples
- Modern proteomic approaches and technologies when coherently integrated can yield new biological insights and novel, sufficiently credentialed biomarker candidates that merit real clinical evaluation
- New, targeted MS-based methods enable highly specific and sensitive quantitative measurement of proteins and their modifications in high multiplex
  - MRM-MS and accurate mass, high resolution variants of MRM (aka PRM) are becoming the new workhorse technologies
  - Broad availability of this resource will change paradigms for how experiments are planned and executed
  - With technological evolution, convergence of discovery and verification is likely

Proteomics Group, Broad Institute of MIT and Harvard

Acknowledgements

Broad Proteomics
- Sue Abbatiello
- Rushdy Ahmad
- Michael Burgess
- Karl Clauser
- Amanda Creech
- Lola Fagbami
- Mike Gillette
- Emily Hartmann
- Jake Jaffe
- Hasmik Keshishian
- Eric Kuhn
- D.R. Mani
- Philipp Mertins
- Jinal Patel
- Lindsay Pino
- Jana Qiao
- Monica Schenone
- Tanya Svink
- Namrata Udeshi
- Janice Williamson

University of Washington
- Michael MacCoss
- Brendan MacLean

FHCRC
- Amanda Paulovich
- Jeff Whiteaker
- Lei Zhao
- Regine Shoenherr
- Pei Wang

Mass. General Hospital
- Robert Gerszten
- Nir Hacohen
- Nicolas Chevrier

Brigham and Womens Hospital
- Marc Sabatine

Funding Agencies
Women's Cancer Research Fund, EIF
Susan G. Komen for the Cure
NIH: NCI and NHLBI
Bill and Melinda Gates Foundation