

Biomarker Discovery: Have We Learned How to Do This Yet?

ABRF Roundtable Discussion

7 February, 2005

Chair: Steve Carr, Broad Institute of Harvard
and MIT, Cambridge, MA

Speakers

- Leigh Anderson Plasma Proteome Institute, Washington, DC
- Kevin Coombes University of Texas, MD Anderson Cancer Center, Houston, TX
- Kojo Elenitoba-Johnson University of Utah School of Medicine, Salt Lake, UT
- Michael Gillette Broad Institute of Harvard and MIT, Cambridge, MA; MGH and HMS, Boston, MA
- Gregory J. Opiteck Caprion Pharmaceuticals, Inc., Montreal, QC, Canada
- Sushmita Mimi Roy SurroMed Inc., Menlo Park, CA

Topics

- I: Sample Type, Collection, Annotation, & Storage
- II: Sample Number and Processing
- III: Analytical Approaches & Pattern v. Identity
- IV: Biomarker Verification & Validation
- V: Data Standards & Infrastructure

I: Sample Type, Collection, Annotation, & Storage

- What are the requirements for samples for BM discovery?
- What human samples should be collected, and how should they be used?
- What are the respective roles of routinely accessible biofluids, “proximal” fluids, and solid tissues?
- What if any roles do cell lines, primary tissue culture systems, or animal models play in biomarker discovery?
- Must human sample collection be prospective, or can existing repositories be used?
- What are the essential features of a prospective sample collection protocol designed for biomarker discovery?
- How does biological variability need to be controlled, characterized, or understood in biomarker studies?
- How does one find sources of samples?

II: Sample Number, Annotation, & Processing

- What number of samples is required for initial candidate discovery? How is this determined?
- Is abundant protein removal necessary for biomarker discovery? What works? What is lost?
- What are useful methods for sample processing and fractionation for MS-based biomarker discovery?
- What is the expected abundance of useful potential biomarkers? Can they be routinely discovered with currently available methods?

III: Analytical Approaches & Pattern v. Identity

- What are the advantages and inherent problems of patterns-based methods?
- How is relative quantitation achieved in pattern methods? How does one manage added dimensions of chromatography in pattern analysis?
- Are patterns sufficient / acceptable for use as markers by clinical groups and regulatory agencies?
- Can one combine approaches or move to identity from a pattern-based approach?
- Do we need panels of biomarkers? How large? In what way combined?
- Is an “unbiased” strategy necessary for biomarker discovery?

IV: Biomarker Verification & Validation

- How should candidate biomarkers be prioritized? How combined with other markers and other information?
- When and how should a biomarker discovery effort move from candidate discovery to verification?
- How many samples are required for biomarker validation? How is that number determined?
- Is there a role for antibodies? At what point should these be incorporated, if available, or generated, if not available?

V: Data Standards & Infrastructure

- Is the proteomics data analysis infrastructure ready for the job of biomarker discovery?
- What data standards should obtain?
- How well must one understand currently available software and analysis tools to generate quality data and interpret them meaningfully?
- Is open access to primary proteomic data obtained in biomarker discovery efforts desirable? What information, formats, and standards are required?

I: Sample Type, Collection, Annotation, & Storage

- What is the role of “proximal” fluids in biomarker discovery?
- What if any roles do animal models play in biomarker discovery for human disease?

An Ideal Disease Biomarker

- Distinguishes healthy from diseased individuals with a high degree of accuracy
- Present during early stages of disease to allow effective therapeutic intervention
- Measurable in a readily accessible body fluid
- Leads to the development of a test that ultimately impacts mortality

An Ideal Disease Biomarker

- Distinguishes healthy from diseased individuals with a high degree of accuracy
- Present during early stages of disease to allow effective therapeutic intervention
- Measurable in a readily accessible body fluid
- Leads to the development of a test that ultimately impacts mortality

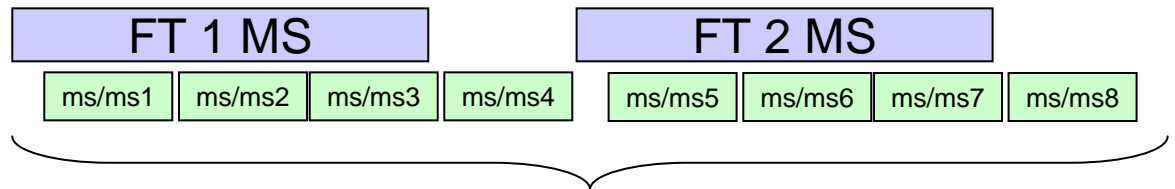
But HOW can we find these?
And why bother with model systems
and “off-target” sample materials?

Higher Data Quality: LTQ-FT Hybrid Instrument



- high mass accuracy: ≤ 2 ppm
- high resolution: 100,000 - 500,000
- Fast MS/MS data acquisition: 10x LCQ
- high sensitivity of linear ion trap for ms^n
- dynamic range in ppt range;
 - use biochemical separations at protein and peptide level to comp.

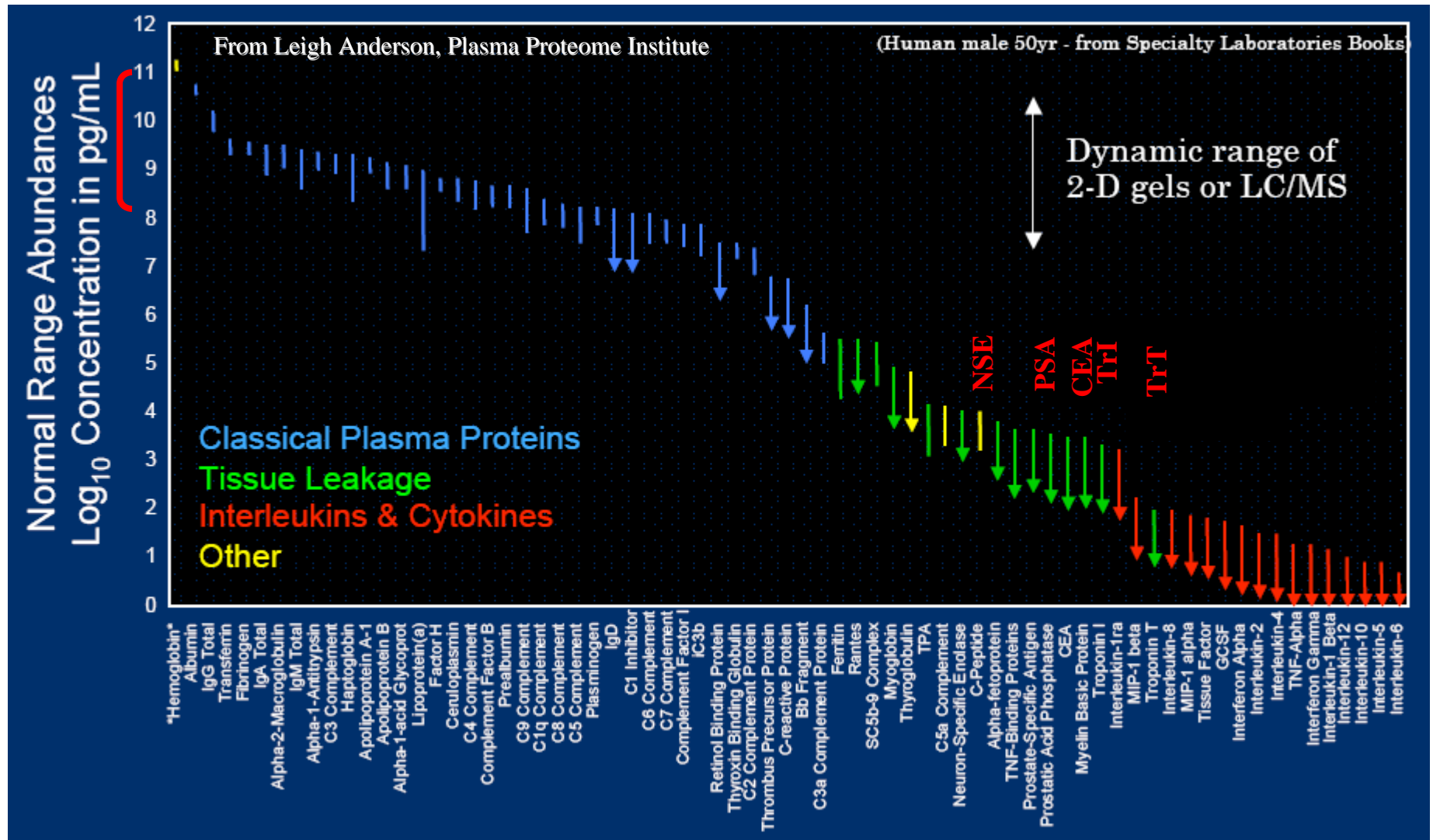
Current LTQ-FT Method Setup



DDA triggered from FT 1 MS survey scan

All within one segment, "continuous" FT cycling

Challenges of the Plasma Proteome

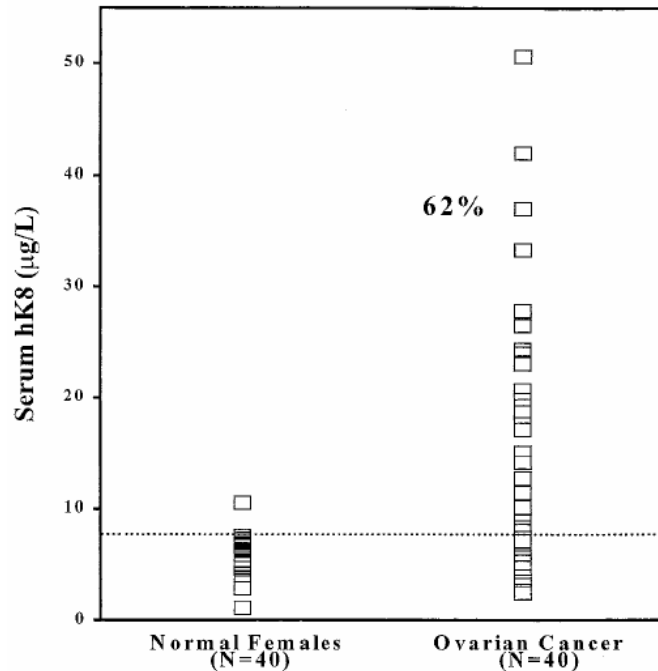


Proximal Fluids: Rationale

- Ultimate goal circulating diagnostic biomarkers in plasma
- Plasma biomarkers expected to be low abundance
 - Mechanism
 - differentially cleaved/secreted/shed proteins from small tumor/microenvironment volume
 - massively diluted into plasma
 - Markers with clinical currency low abundance
- Complexity & dynamic range of plasma daunting; **finding / validating candidates easier than primary discovery**
- **Fluids close to source expected to be markedly enriched in candidate markers**

Proximal Fluid Marker Enrichment: Concentration

CA125 mean level



CANCER RESEARCH 63, 2771-2774, June 1, 2003

| Histology | Serum | Ascites | Cyst Fluid |
|---------------------|-------|---------|------------|
| Serous CA | 696 | 18,560 | 44,850 |
| Endometrioid CA | 661 | 14,415 | 32,150 |
| Mucinous AdenoCA | 67 | 3,521 | 3931 |
| Undifferentiated CA | 861 | 3,909 | -- |
| Serous Cystadenoma | 7 | -- | 42,150 |
| Serous Cyst | 5 | -- | 6852 |
| Mucinous Adenoma | 11 | -- | 5692 |

CANCER November 1, 2002 / Volume 95 / Number 9

- HIP/PAP I pancreatic ductal carcinoma
- Discovery by MALDI in pancreatic juice, ELISA confirmation in serum
- Juice levels 1000x serum levels
- Juice cancer / control concentration difference 24x; serum 3x

CANCER RESEARCH 62, 1868-1875, March 15, 2002

Proximal Fluid Marker Enrichment: Fold Change

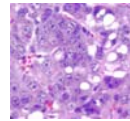
| | Benign (<i>n</i> = 30) | Malignant (<i>n</i> = 44) |
|---------------------------|-------------------------|----------------------------|
| <i>CEA (ng/ml)</i> | | |
| Serum | 1.41 ± 0.20 | 7.52 ± 2.50* |
| Pleural fluid | 1.80 ± 0.32 | 55.44 ± 21.8* |
| <i>CA 15-3 (U/ml)</i> | | |
| Serum | 15.52 ± 2.20 | 53.50 ± 8.70*** |
| Pleural fluid | 8.36 ± 0.71 | 49.01 ± 7.60*** |
| <i>CA 19-9 (U/ml)</i> | | |
| Serum | 7.70 ± 1.50 | 38.78 ± 11.33* |
| Pleural fluid | 3.60 ± 0.72 | 31.36 ± 9.23* |
| <i>CYFRA 21-1 (ng/ml)</i> | | |
| Serum | 3.77 ± 0.60 | 13.33 ± 1.70*** |
| Pleural fluid | 3.77 ± 0.55 | 32.30 ± 2.9*** |
| <i>NSE (µg/l)</i> | | |
| Serum | 9.60 ± 0.44 | 13.97 ± 1.30** |
| Pleural fluid | 5.50 ± 0.80 | 5.21 ± 0.42 |
| <i>TSA (mg/dl)</i> | | |
| Serum | 178.00 ± 5.10 | 138.87 ± 6.40*** |
| Pleural fluid | 67.18 ± 6.40 | 91.23 ± 5.30** |

* *P* < 0.05 as compared to benign effusions.
 ** *P* < 0.01 as compared to benign effusions.
 *** *P* < 0.001 as compared to benign effusions.

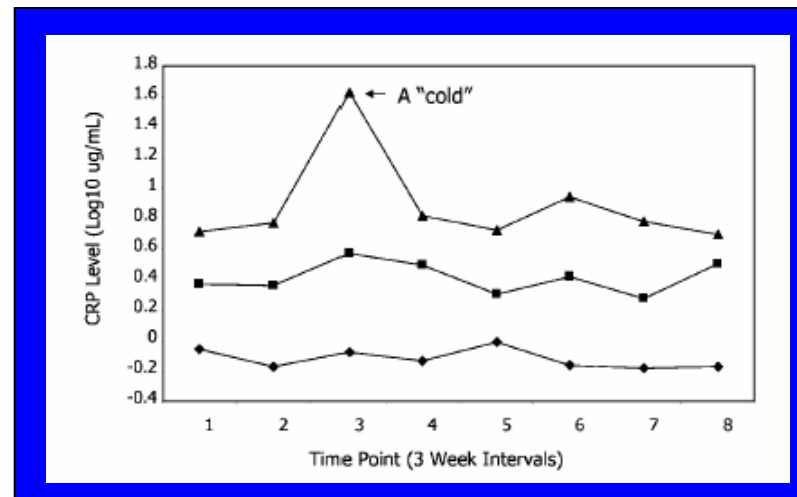
Lung Cancer 31 (2001) 9–16

Challenges of the Plasma Proteome

- Plasma proteome highly dynamic; “Noisy”



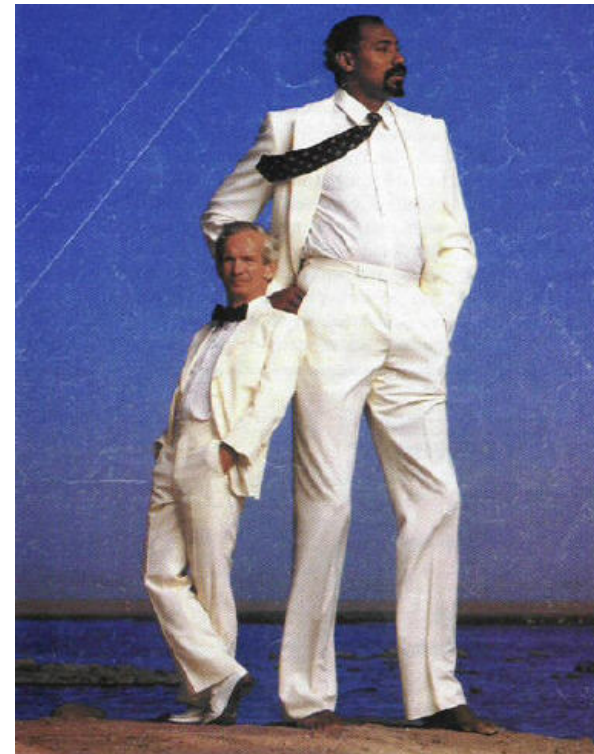
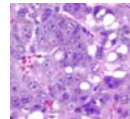
- Inter-individual variation in plasma proteins may obscure disease-specific intra-individual changes



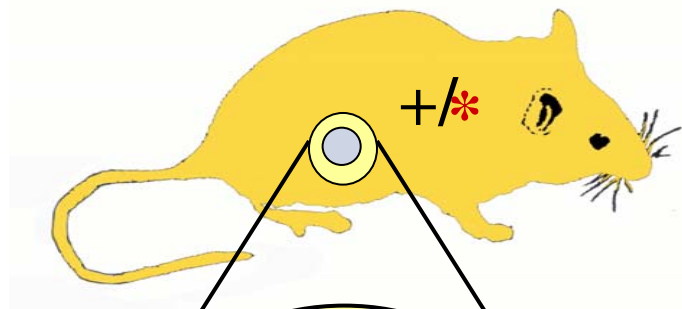
Anderson & Anderson Mol. Cell. Proteomics 1: 845-867, 2002

Challenges of the Plasma Proteome

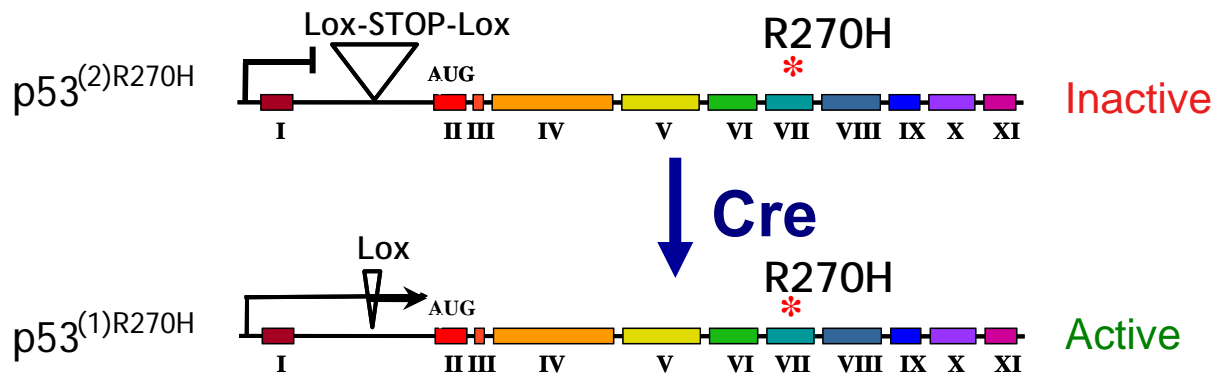
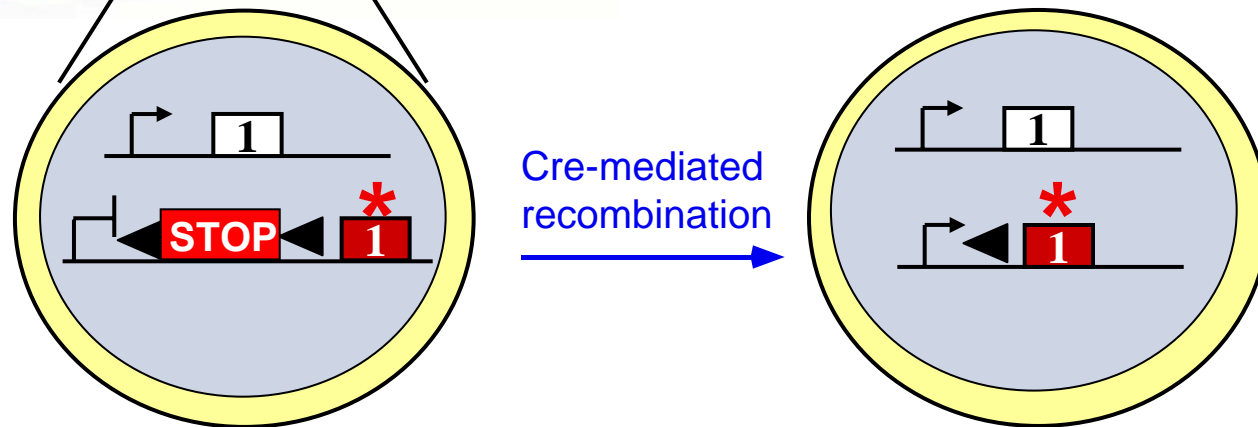
- Plasma proteome highly dynamic; “Noisy”



Conditional LSL-K-ras^{G12D} / P53^{R270H} Mouse

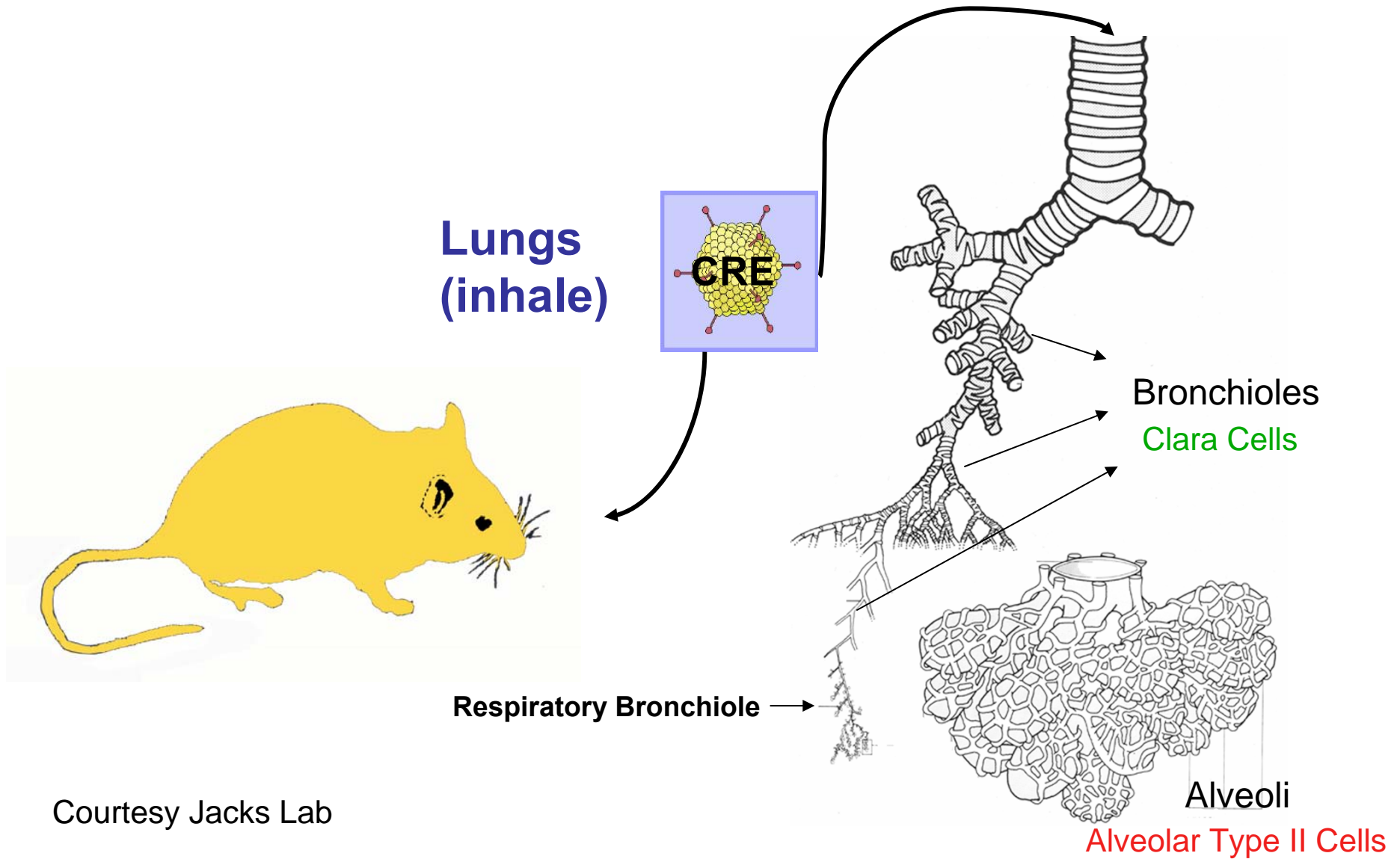


+ Wild-type K-ras or P53 allele
* Conditional K-ras or P53 allele



Tyler Jacks Lab

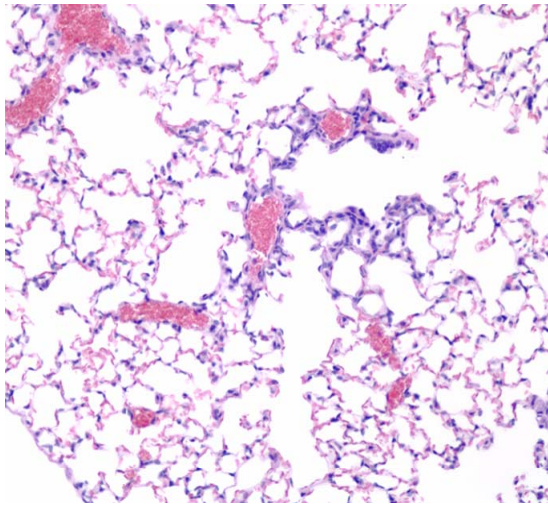
Adenoviral Delivery of Cre Recombinase to the Lungs of LSL-K-ras^{G12D} Mice



Courtesy Jacks Lab

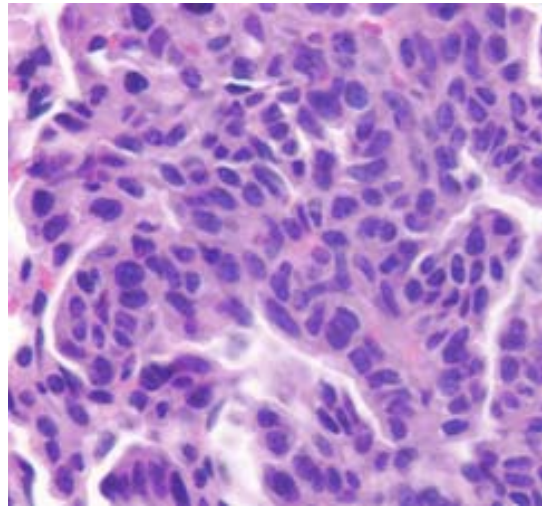
Stages of Tumor Progression in LSL-K-ras^{G12D} Mice

2 weeks



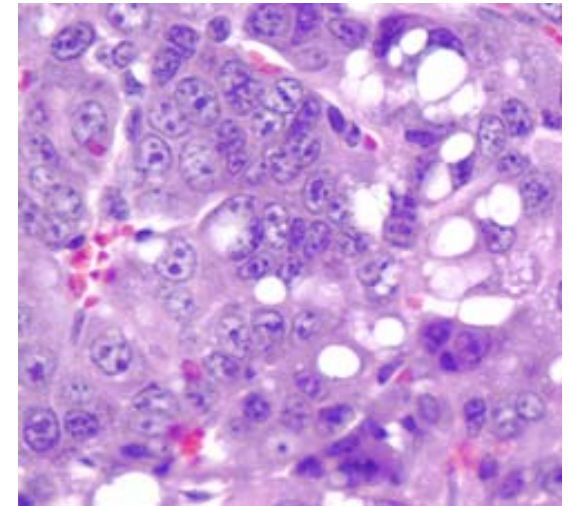
Atypical adenomatous hyperplasia

6 weeks



Papillary adenoma

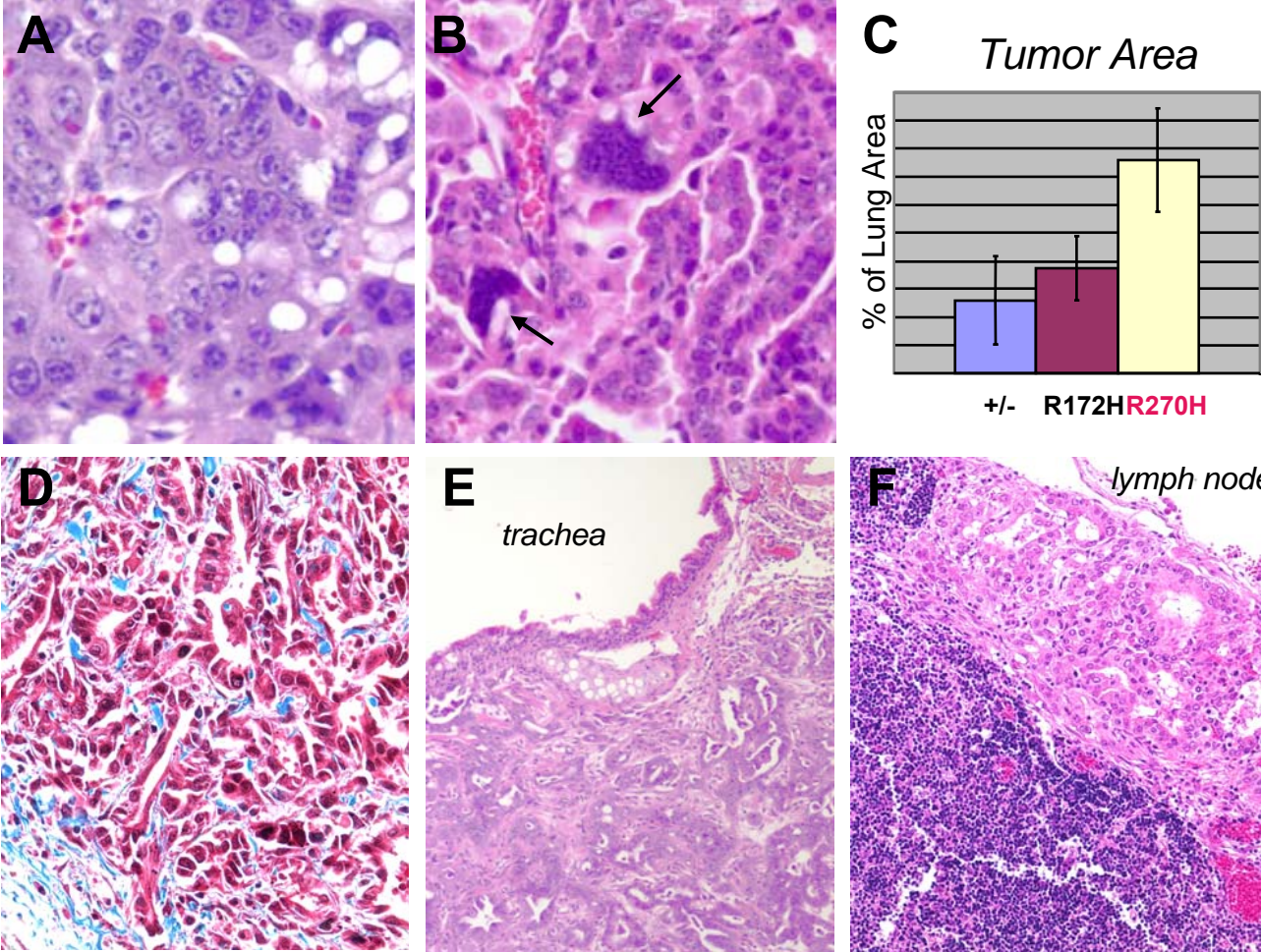
16 weeks



Adenocarcinoma

Courtesy Jacks Lab

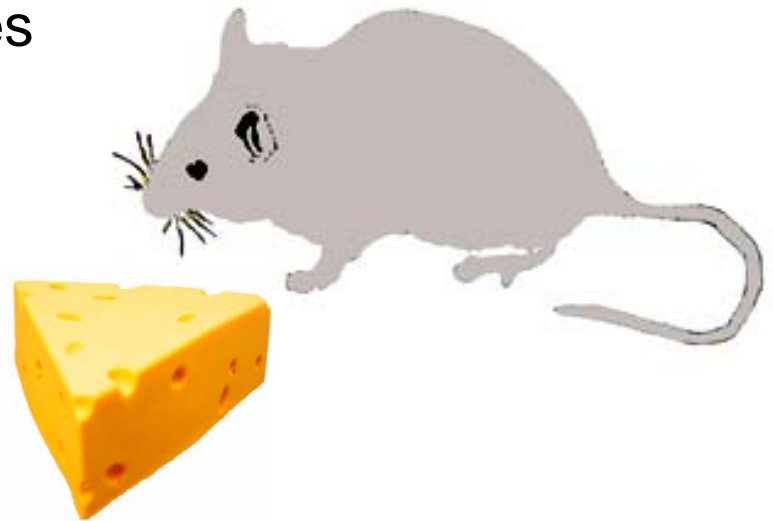
Combining K-ras and p53 Mutations: New and Improved Models of Human Lung Cancer



Courtesy
Jacks Lab

Advantages of Mouse Models for Biomarker Discovery

- Accurate models at both the genetic and histopathologic level
- Large numbers of animals readily obtained (also facilitates pooling)
- Inbred and uniform genetic background (129/sv)
- Controlled environmental exposures
- Uniform sample collection
- Timed tumor progression
- “Patients” as their own controls



III: Analytical Approaches & Pattern v. Identity

- Is there a role for pattern-based methods in proteomic biomarker discovery? How might pattern be used?

Pattern-based Diagnostics Venerable in Medicine

- Hippocrates: *Karkinos* based on crab-like appearance of growth pattern or vasculature
- Galen: *Rubor, dolor, calor, and tumor*, cornerstones of diagnosis of local inflammation
- Virchow: Microscopic pattern as standard for diagnosis

Pattern-based Diagnostics Contemporary in Medicine

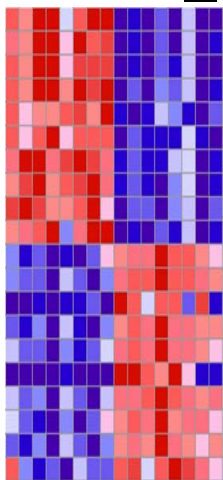
- Hippocrates: *Karkinos* based on crab-like appearance of growth pattern or vasculature
 - Galen: *Rubor, dolor, calor, and tumor*, cornerstones of diagnosis of local inflammation
 - Virchow: Microscopic pattern as standard for diagnosis
-
- Chest pressure, arm radiation, nausea, diaphoresis
 - Bilateral hilar lymph node enlargement in o/w clear CXR
 - Histopathology with standard stains
 - Recognition of and reliance upon a pattern typically precedes, and provides impetus for, deeper understanding

Pattern-based Diagnostics Contemporary in Medicine

- Cytogenetics may be diagnostic before genetic mechanisms understood
- DNA methylation may have clear disease association but unclear biological significance
- Gene expression analyses proffer disease signatures in high dimensionality expression space

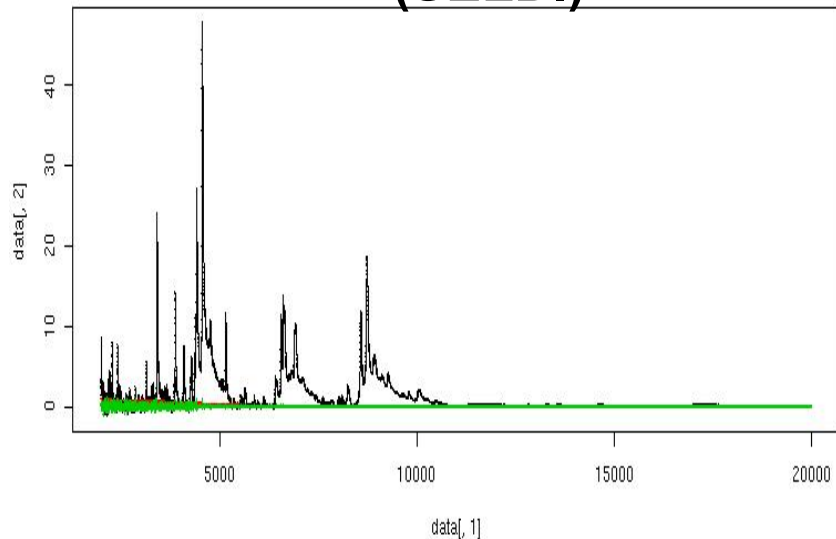
– Components have biologically interpretable labels, but these do not improve performance characteristics as biomarkers *per se*

- *May be wrong (early generation arrays)*
- *May be effectively ambiguous (ESTs)*
- *Interpretation may be highly speculative, or utterly elusive*

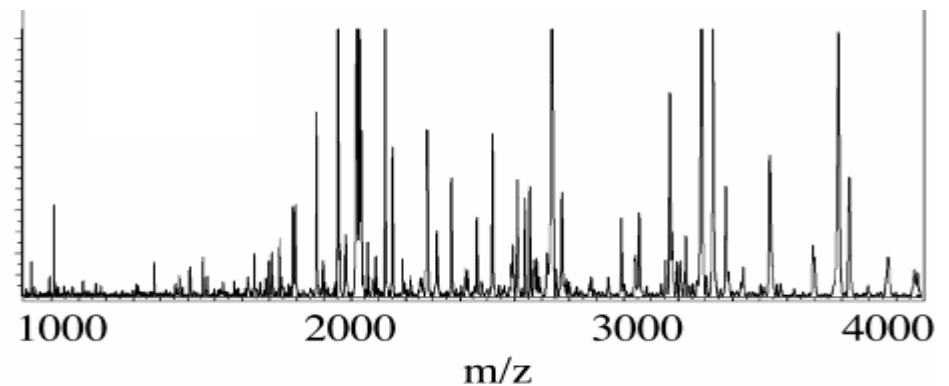


Pattern as biomarker

**Low performance MALDI-TOF MS
with on target sample capture
(SELDI)**



**High performance MALDI-TOF
MS with off-target sample capture
(Tempst, Anal. Chem. 2004)**

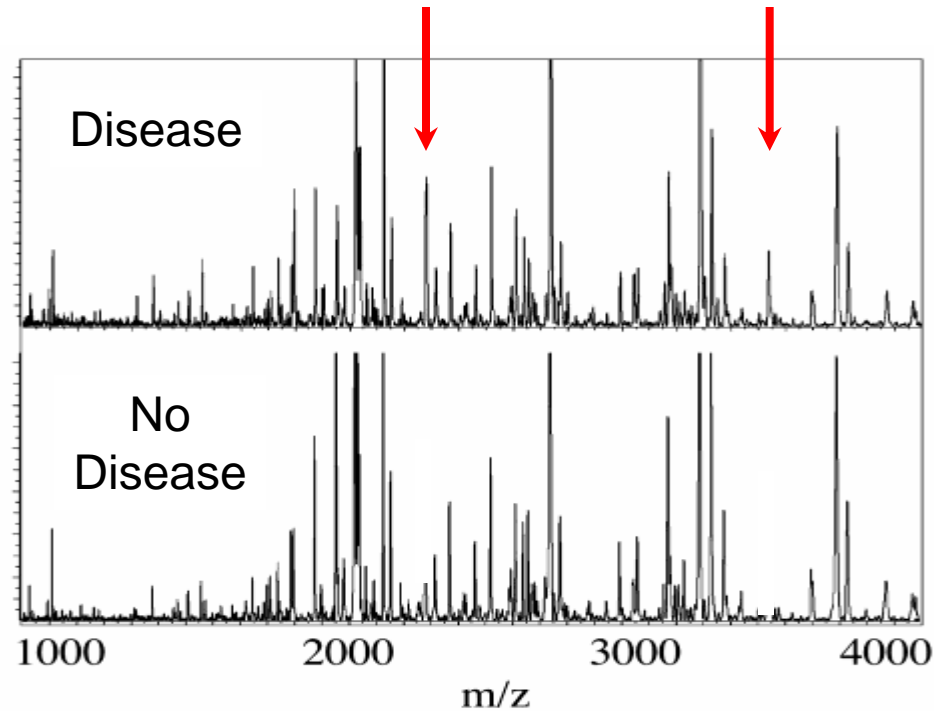


Distinguish *strategy* from *platform*

Utility of Identity

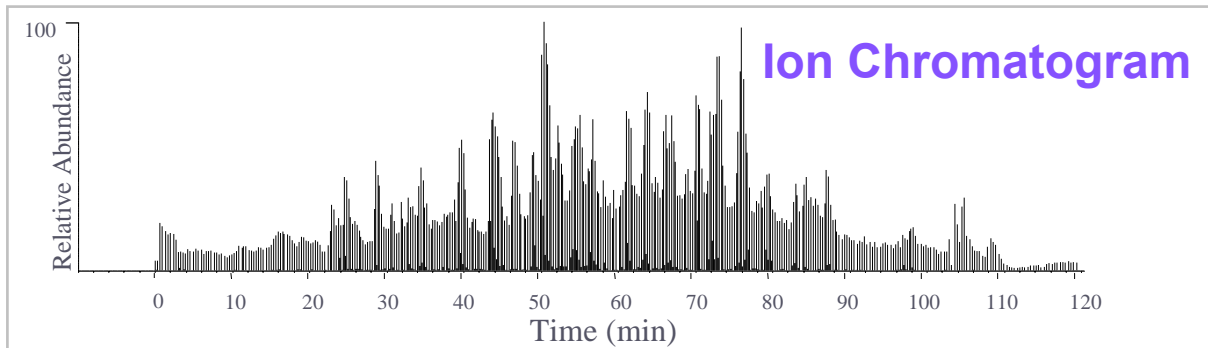
- May provide biological insight
 - Disease pathogenesis
 - Therapeutic targets
- May increase confidence in markers
 - If compelling association with disease biology
- May facilitate transfer of biomarkers to alternative diagnostic platform

Pattern *to guide* biomarker discovery

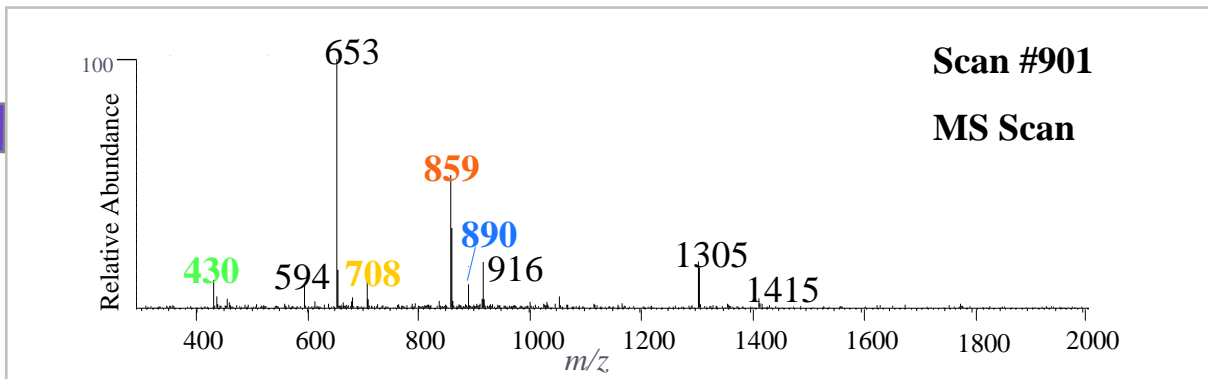


- Protein *or* peptide level
- Collect full MS spectra across large number of samples
- Determine difference signals with machine learning
- Identify discriminant portions of signatures

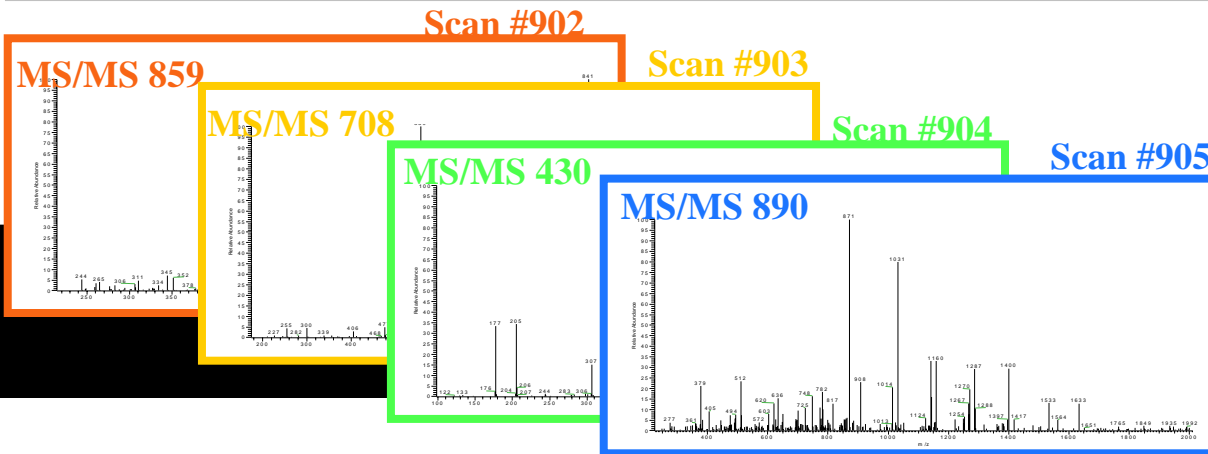
Pattern to enrich biomarker discovery



**Data
Dependent
Acquisition**

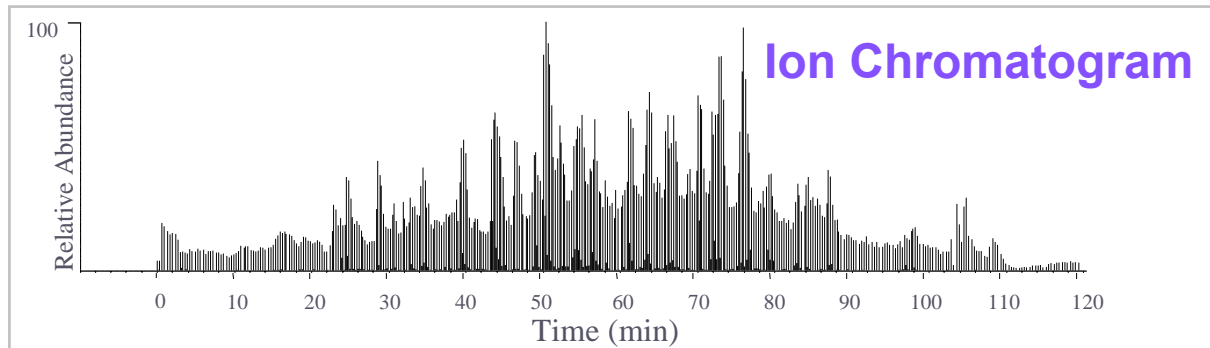


**Repeat
Process**

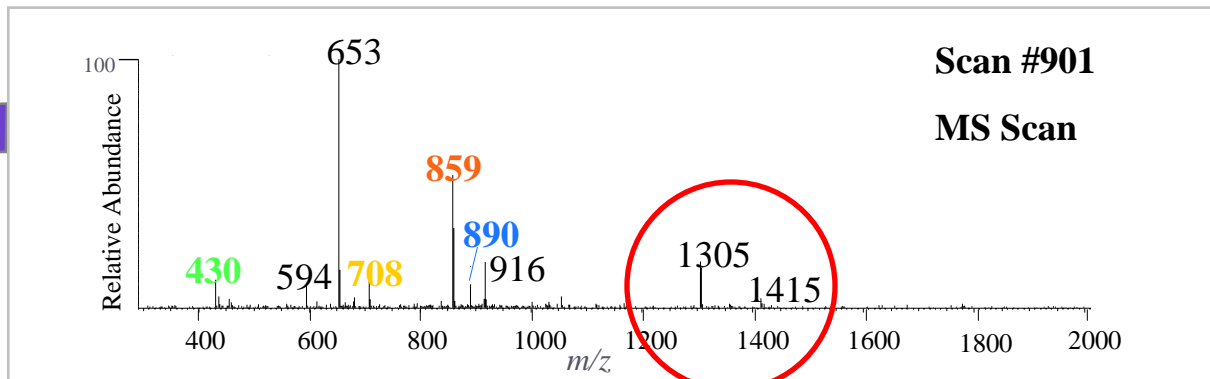


**1000's of
MS/MS Spectra**

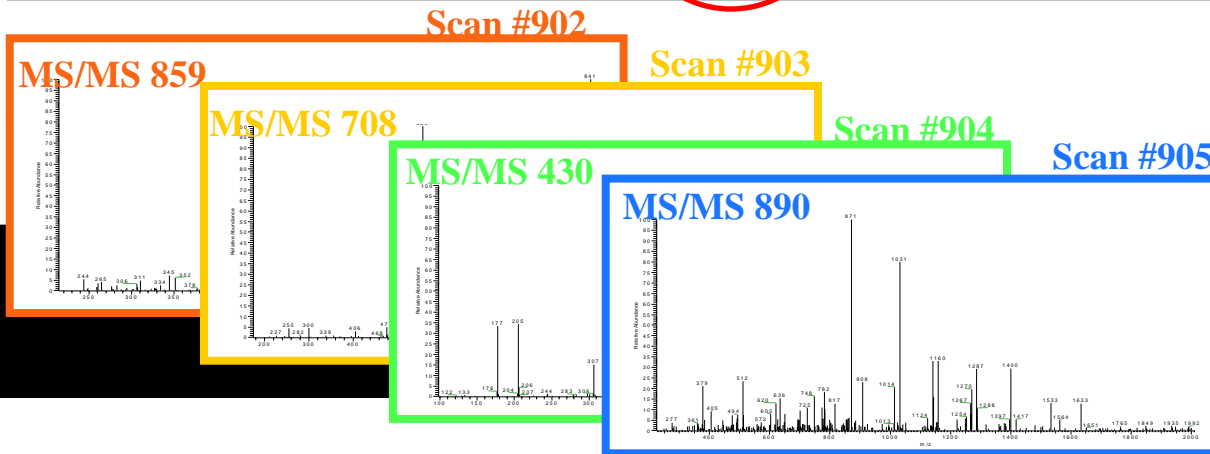
Pattern to enrich biomarker discovery



Data
Dependent
Acquisition



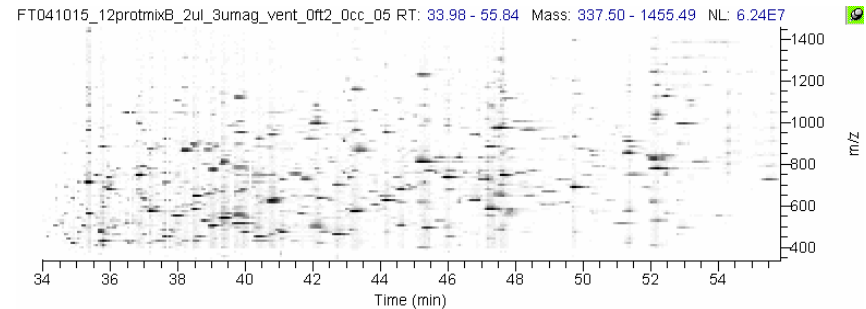
Repeat
Process



1000's of
MS/MS Spectra

Pattern Recognition Overview

1. Data preparation & preprocessing
 - De-noising
 - Centroiding/Peak detection
 - Signal-to-noise-based filtration
 - De-isotoping
2. Normalization
 - Across multiple LC-MS scans
3. Peak Matching
 - Locate and align peptides across multiple LC-MS scans
 - Methods:
 - Gaussian mixture modeling
 - Hierarchical clustering
 - Considerations:
 - Large RT drift and small mass imprecision
 - Use identified peptides to constrain matching
4. Data Table Assembly
 - Concatenation of multiple aligned fractions for a biological sample
5. Machine Learning



| Features | Sample ₁ | Sample ₂ | ... | Sample _n |
|---------------------------------------|---------------------|---------------------|-----|---------------------|
| (rt ₁ , m/z ₁) | 0.4 | 3.8 | | n/a |
| (rt ₂ , m/z ₂) | n/a | 10.0 | | 1.0 |
| ... | | | | |
| (rt _k , m/z _k) | 0.1 | n/a | | n/a |

Pattern Recognition Overview

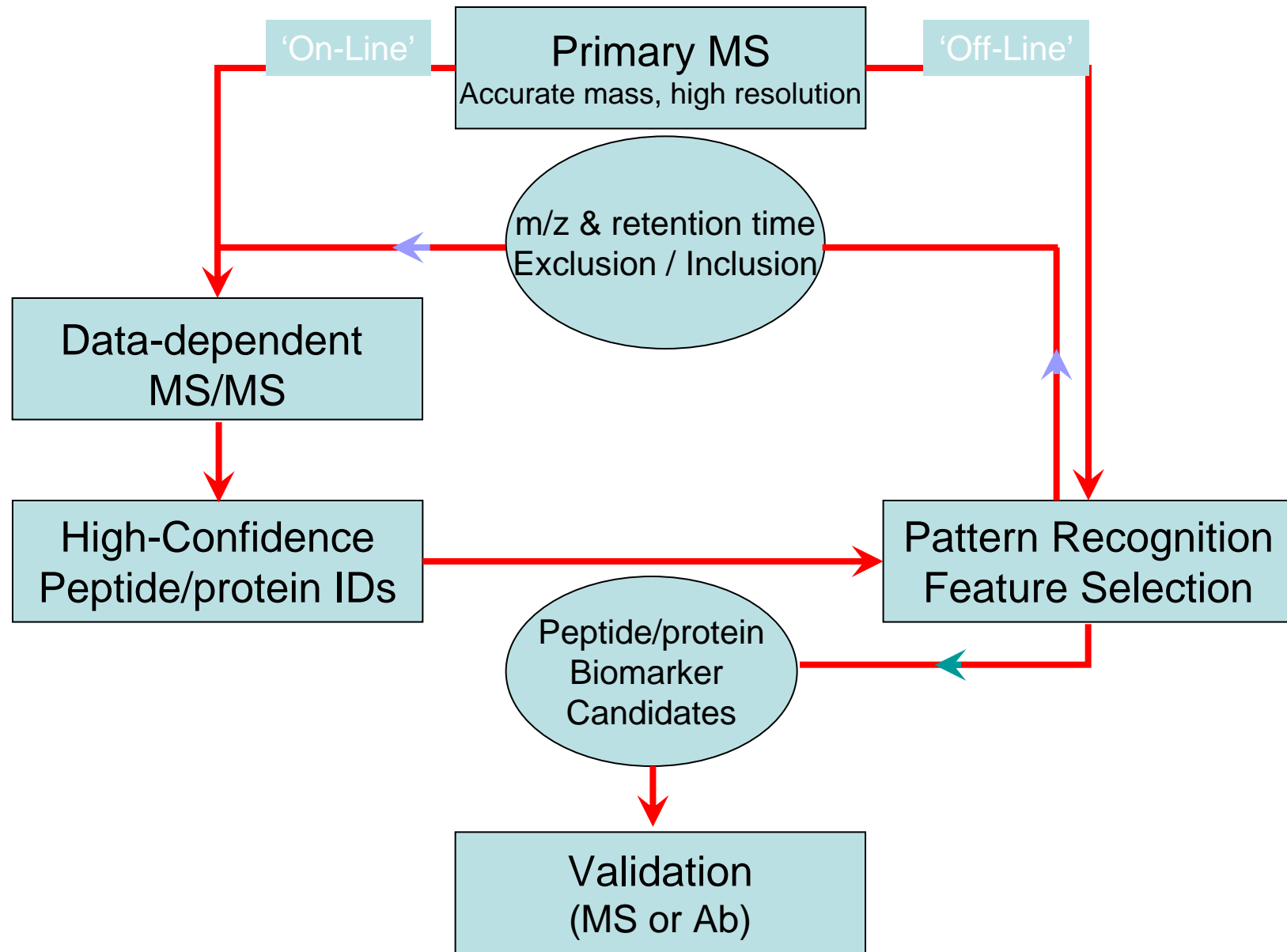
1. Data preparation & preprocessing
 - De-noising
 - Centroiding/Peak detection
 - Signal-to-noise-based filtration
 - De-isotoping
2. Normalization
 - Across multiple LC-MS scans
3. **Peak Matching**
 - Locate and align peptides across multiple LC-MS scans
 - Methods:
 - Gaussian mixture modeling
 - Hierarchical clustering
 - Considerations:
 - Large RT drift and small mass imprecision
 - Use identified peptides to constrain matching
4. Data Table Assembly
 - Concatenation of multiple aligned fractions for a biological sample
5. Machine Learning

- Locate and align peptides across multiple LC-MS scans
- Methods:
 - Gaussian mixture modeling
 - Hierarchical clustering
- Considerations:
 - Large RT drift and small mass imprecision
 - Use identified peptides to constrain matching

Machine Learning

- Feature selection
 - Features (rt,m/z,fraction) up- or down-regulated in phenotype
 - Methods:
 - T-test
 - Decision tree variable importance
 - Genetic algorithms
- Classification
 - Multivariate models for classifying samples into predefined phenotypes
 - Methods:
 - *k*-Nearest neighbor
 - Weighted voting
 - Decision trees
 - Genetic algorithms

Biomarker Discovery by Identity *and* Pattern



Acknowledgments

MIT Center for Cancer Research

Tyler Jacks
Alice Tsang Shaw

Alejandro Sweet-Cordero
Denise Crowley
Jim Dowdle
Erica Jackson
David Kirsch
JudiAnn Ramiscal

The Broad Institute of MIT and Harvard

Proteomics Platform

Steve Carr
Terri Addona
Betty Chang
Karl Clauser
Jake Jaffe
Hasmik Keshishian
D.R. Mani
Matt Sigakis
Veronica Saenz-Vash