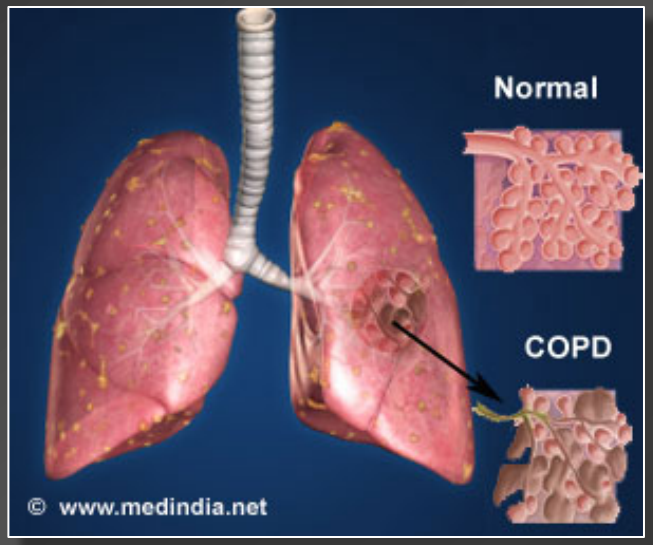
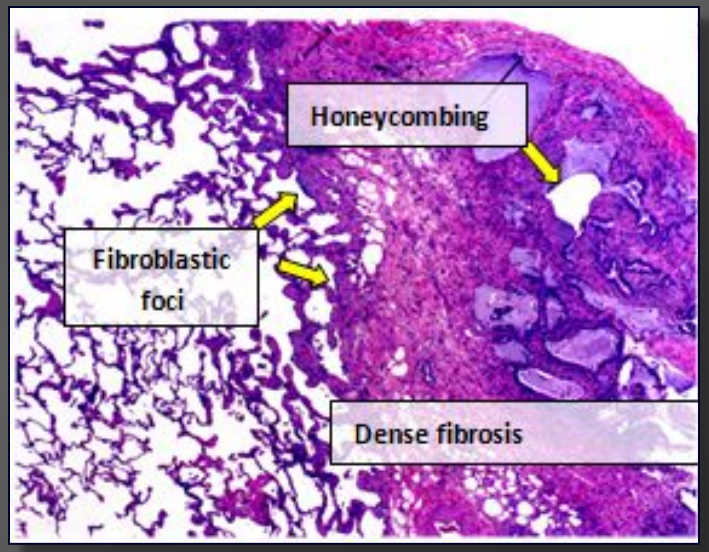
Causal modeling in the lung

Combining multiple data types to enhance clinical diagnosis





Takis Benos

Department of Computational & Systems Biology

University of Pittsburgh, SOM

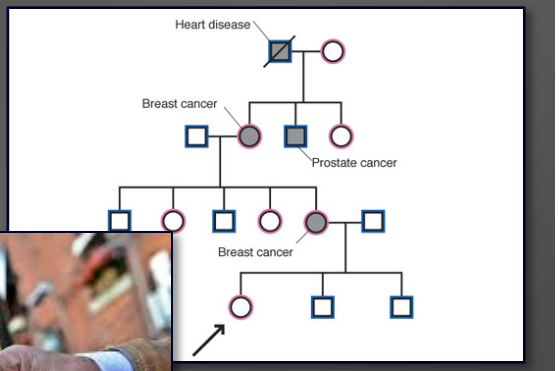
CCM Workshop, Pittsburgh, PA

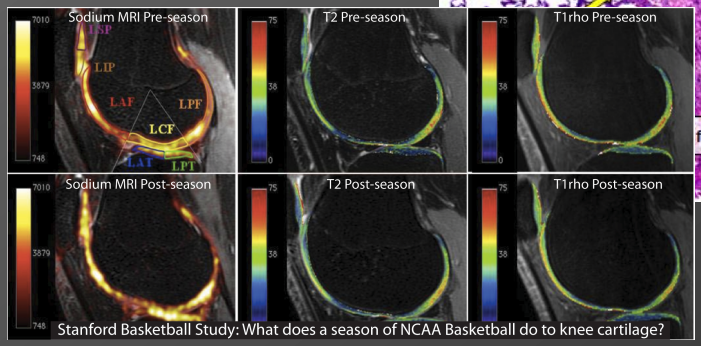
June 10, 2015

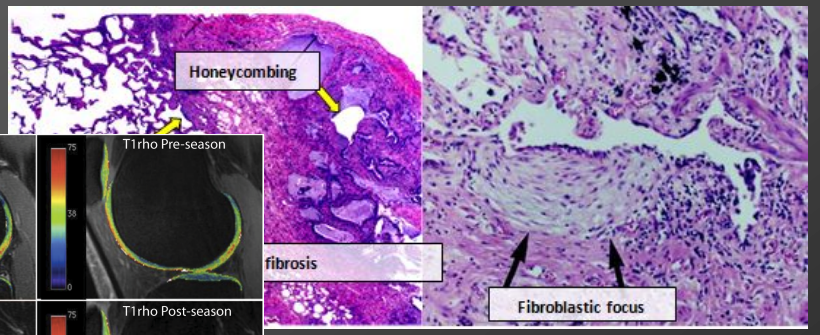
What kind of data do we collect from patients?

- Demographics, family history, patient's history
- Clinical tests
- Clinical image data
- OMICS

moz.com



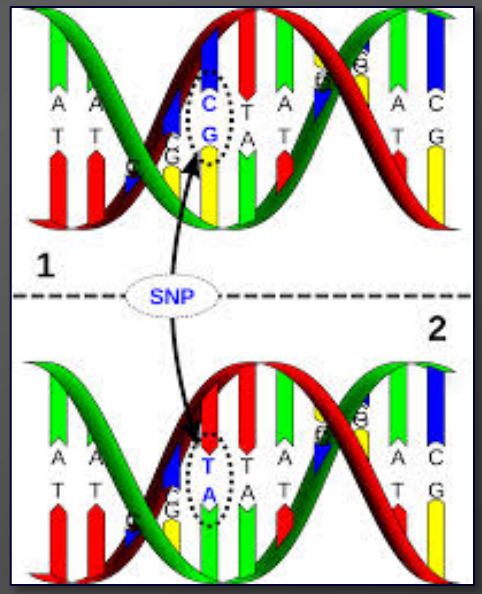




Other images: nih.gov

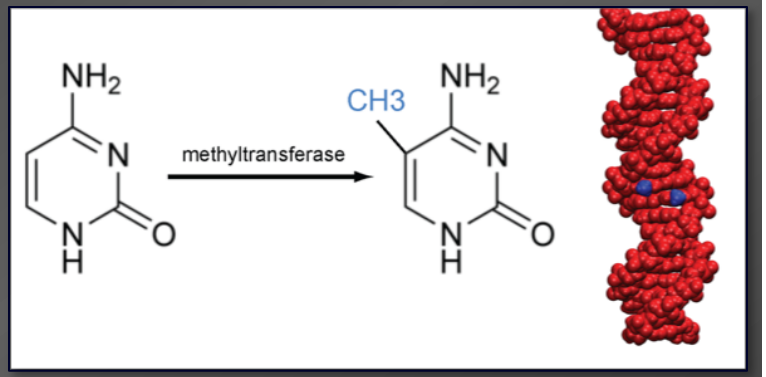
Single Nucleotide Polymorphisms (SNPs)

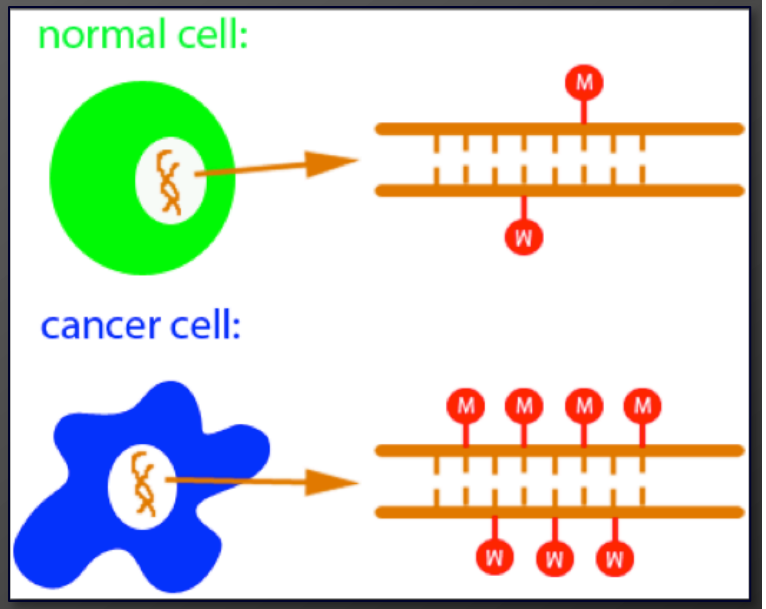
- Human Genome: ~6B bp (2 x 3B bp)
  - ~50M SNPs have been catalogued
- Each person: 3-5M SNPs
  - 150K SNPs are not catalogued
  - ~60 SNPs are specific to you
- Most SNPs have no functional consequences, but some are disruptive
  - ~100 SNPs disrupt a gene copy
  - ~20 SNPs disrupt both copies
- For every person we can sample 1-2 million SNPs



DNA Methylation

- [C | A] methylation is a covalent modification
- In mammals, 60-90% of the CpG are methylated
- DNA methylation pattern affects
  - Gene expression
  - Genome stability
  - X chromosome inactivation
- Our assay sampled ~2M methylation events





Characteristics of mRNA and microRNA expression

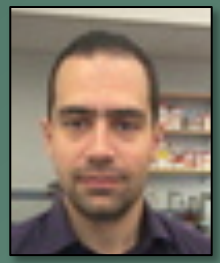
- mRNAs code for proteins (~20K genes in mammals)
- microRNAs (miRNAs) reduce the levels of their target mRNAs in the cytoplasm (~2-3K miRNAs known in mammals)

- (m/mi)RNA expression changes depending on cell type, disease status, developmental stage, etc. They can be:
  - Markers of disease status
    - Diagnostic markers
  - Markers of therapy outcome
    - Prognostic markers

Lung DBP: the team

THE (BIGGISH) DATA

Naftali Kaminski, MD



Argyris Tzouvelekis, MD

IMAGE ANALYSIS

Chakra Chennubhotla, PhD

Frank Schneider, MD

Akif Burak Tosun, Ph.D

CAUSAL MODELING

Takis Benos, PhD

Hyokyeong Lee, PhD

Lung diseases we are interested in

Chronic lung diseases – COPD (obstructive)

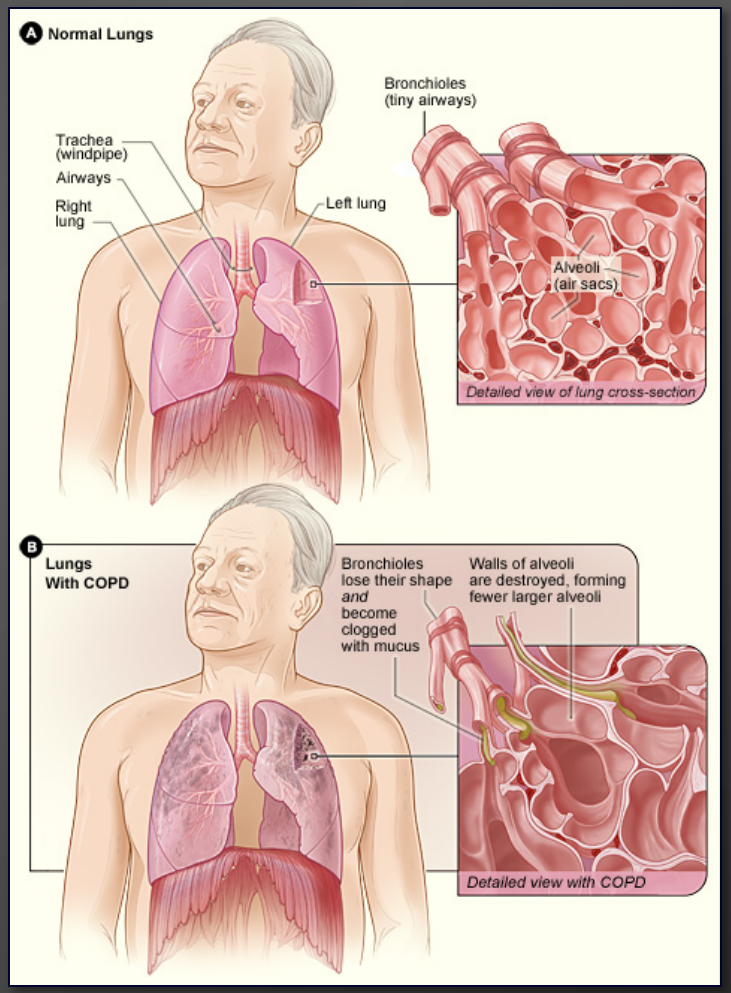
Physiological changes

- The airways and air sacs (alveoli) lose their elastic quality

- The walls between many of the air sacs are destroyed

- The walls of the airways become thick and inflamed

- The airways make more mucus, which can clog them



Chronic lung diseases – COPD (obstructive)

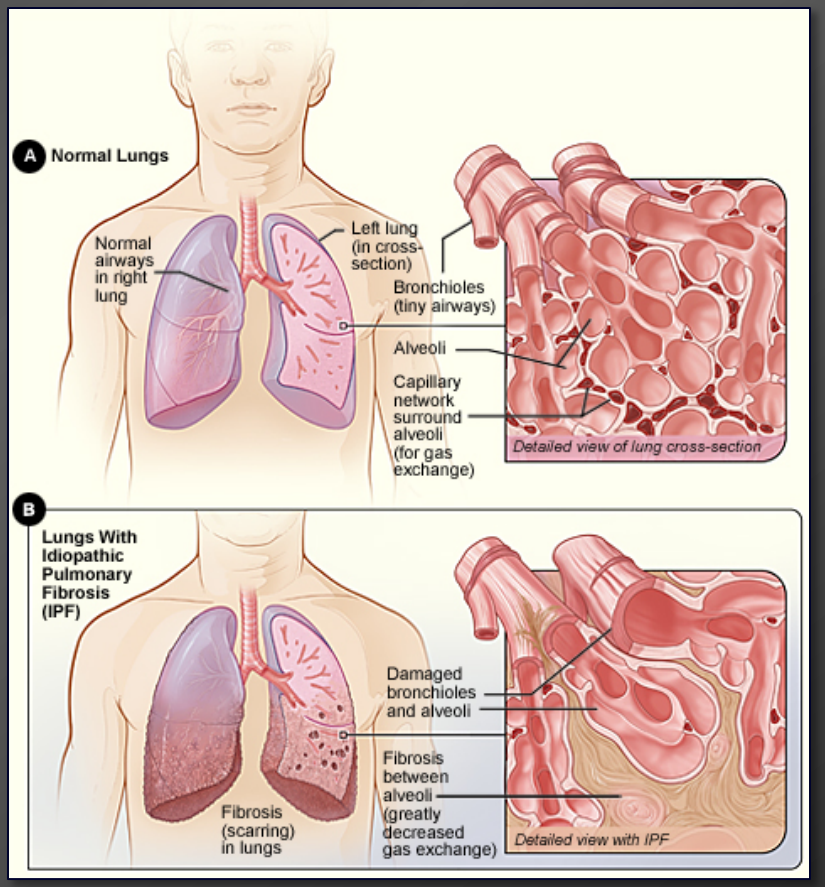
- **Symptoms**
  - Shortness of breath (dyspnea)
  - Cough with mucus
- **Risk factors**
  - Tobacco smoking
  - Air pollution
  - Occupational exposures
  - Genetics
- **Diagnosis** *via* spirometry
  - Forced Expiratory Volume in 1st sec (FEV_1)
  - Forced Vital Capacity (FVC)
    - FEV_1/FVC < 70% + other symptoms → COPD diagnosis
- **Treatment**
  - No known cure
- **Management**
  - Symptoms are treatable (e.g. bronchodilators, corticosteroids, etc)
  - Progression can be delayed by reduction of the risk factors

Chronic lung diseases – IPF (restrictive)

Physiological changes

- The tissue in the lungs becomes thick and stiff, or scarred over time (fibrotic tissue)

- This makes lungs unable to move oxygen to the bloodstream



Chronic lung diseases – IPF (restrictive)

- Symptoms
  - Age: >50 yrs
  - Cough w/o mucus
  - Progressive dyspnea
  - Characteristic "velcro-like" breathing
  - Disfigurement of fingertips (clubbing of the digits)

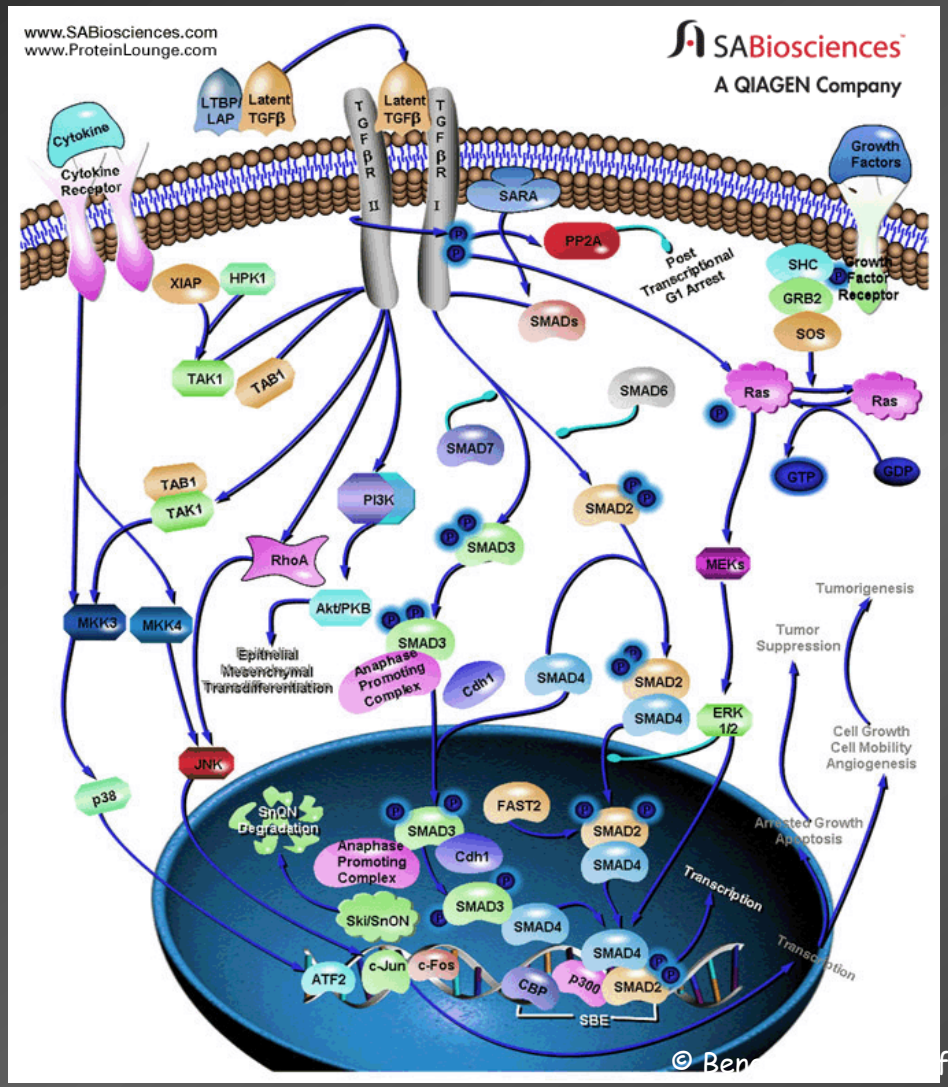
- Causes and Risk factors
  - "Idiopathic"
  - Tobacco smoking
  - Genetics
  - Environmental + occupational exposures (metal dust, wood dust, coal dust, silica, etc)

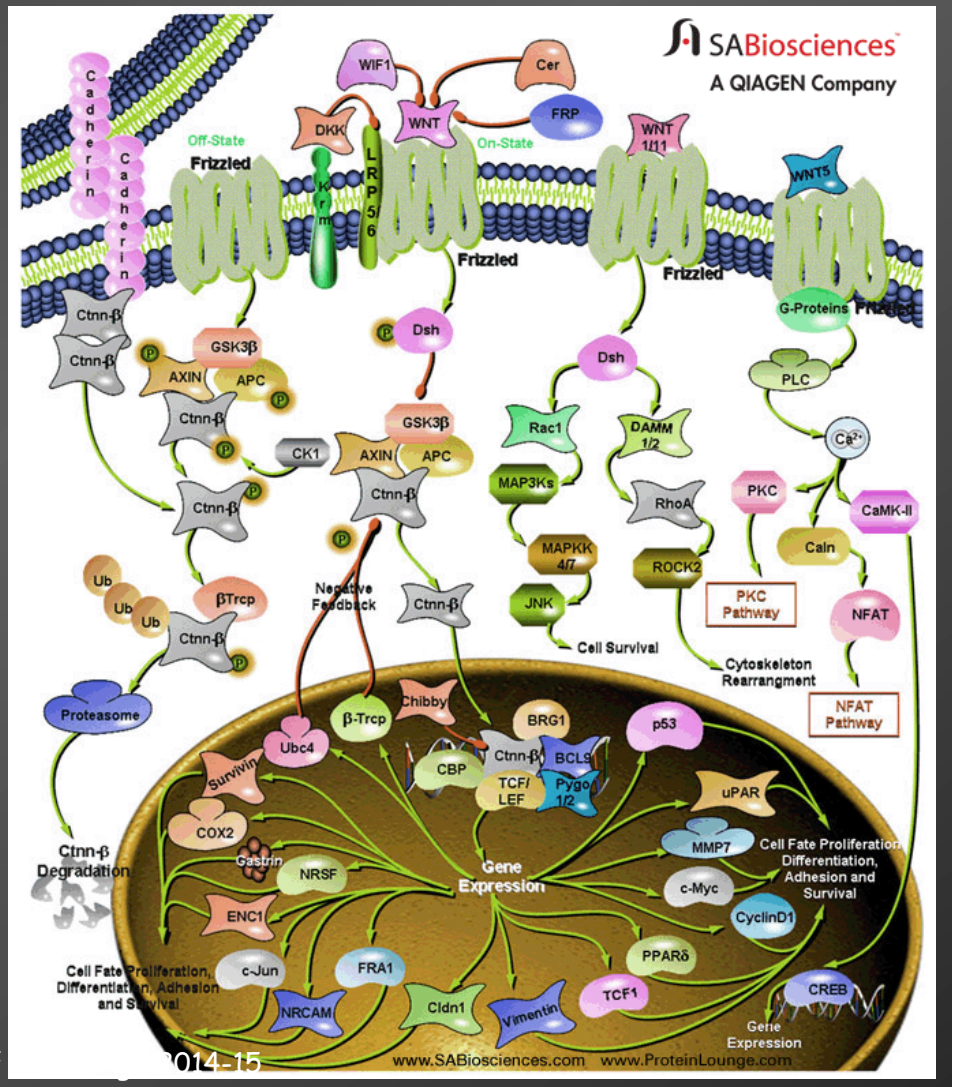
Prognosis: not good

2-5 years following diagnosis
5-yr survival: 20-40%

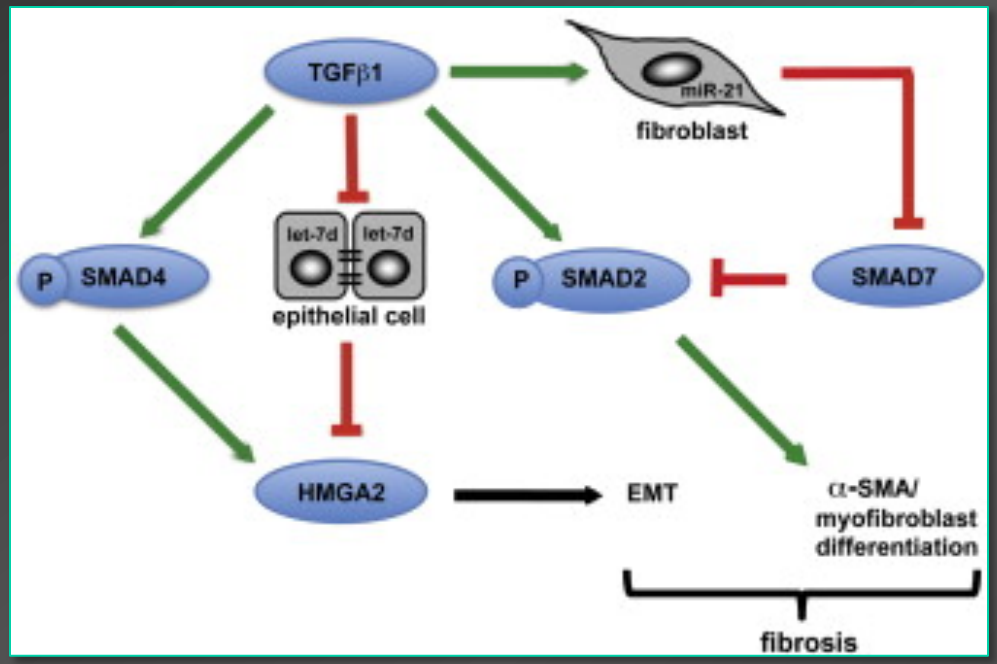
Pathways involved in IPF

TGF-β1 and Wnt signaling pathways





Transcriptional and post-transcriptional pathways



Pandit et al. "MicroRNAs in idiopathic pulmonary fibrosis" Translational Research, 2011

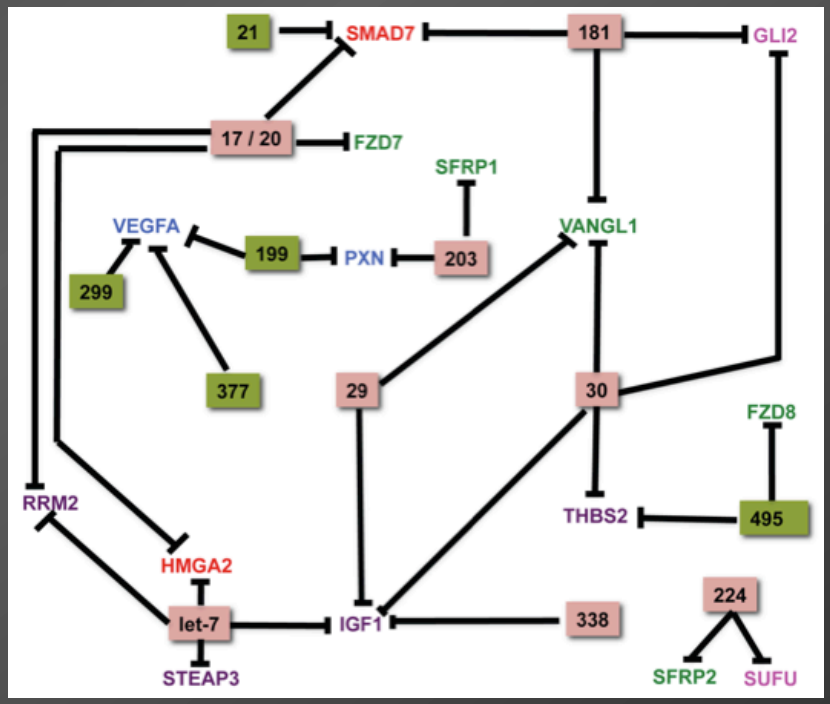


Image source: Naftali Kaminski, MD

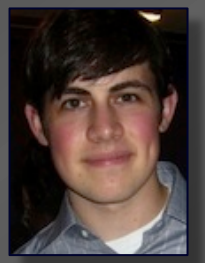
The Lung Genomics Research Consortium (LGRC)

The LGRC resource

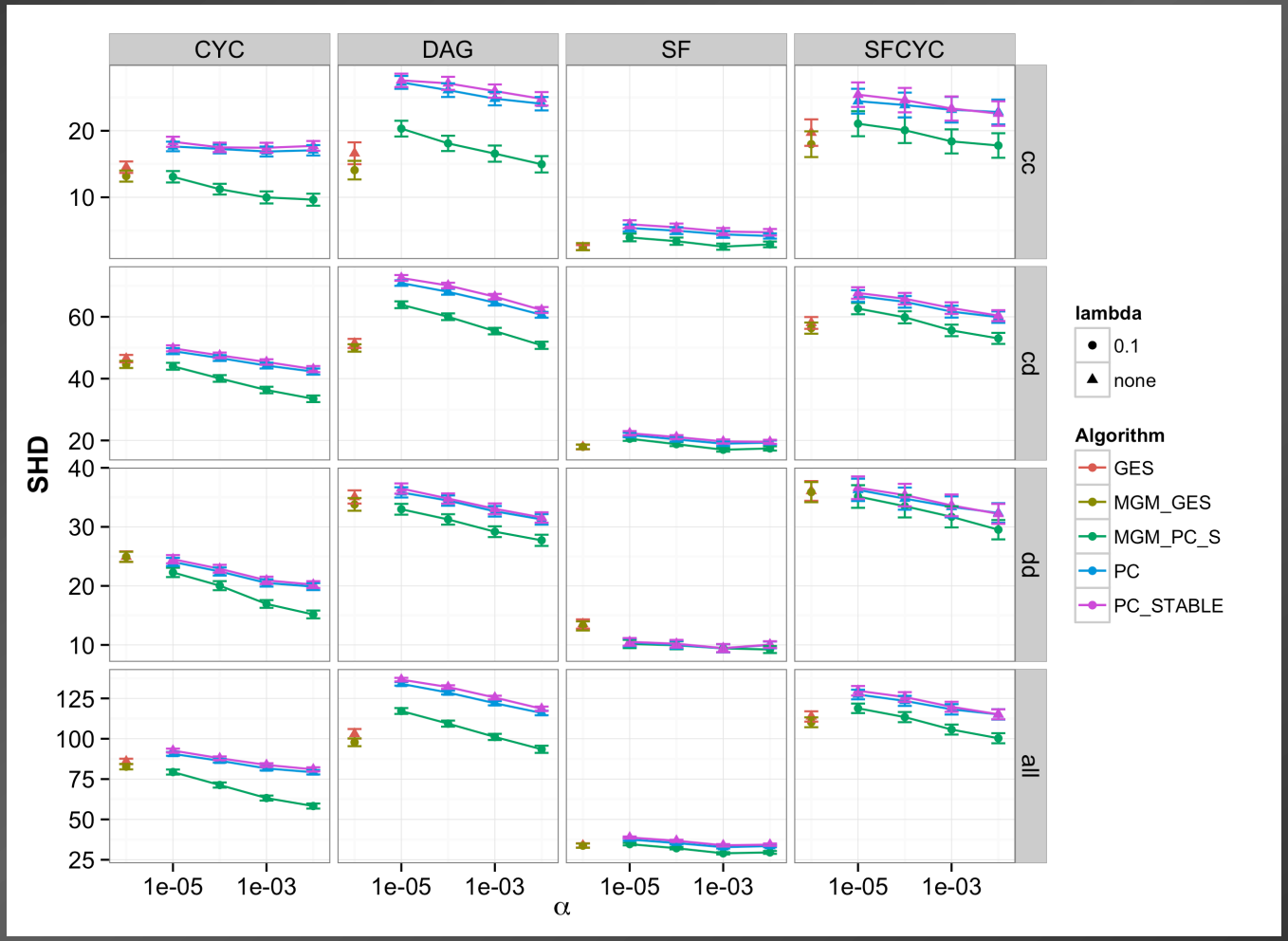
- LGRC resource includes a variety of omic, clinical and image data on chronic lung diseases such as
  - Chronic Obstructive Pulmonary Disease (COPD)
  - Interstitial lung disease (ILD), including Idiopathic Pulmonary Fibrosis (IPF)

- LGRC initial scope: use these data to…
  - Identify people at risk of developing COPD or ILD
  - Make an early diagnosis
  - Determine causes of disease
  - Provide personalized treatment

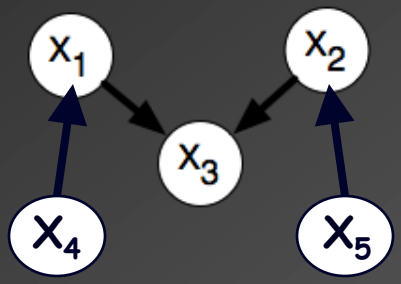
MGM-Learn methods perform better than GES, PC, PC_STABLE



AJ Sedgewick

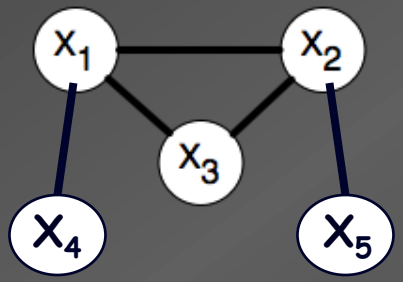


"Undirecting" colliders has surprises



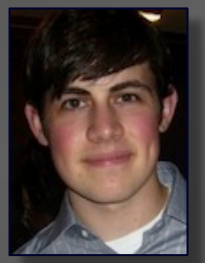
Collider: $x_1 \perp x_2 \mid \varnothing$ and $x_1 \not\perp x_2 \mid x_3$

$A \not\perp B \mid Z \setminus \{A, B\}$



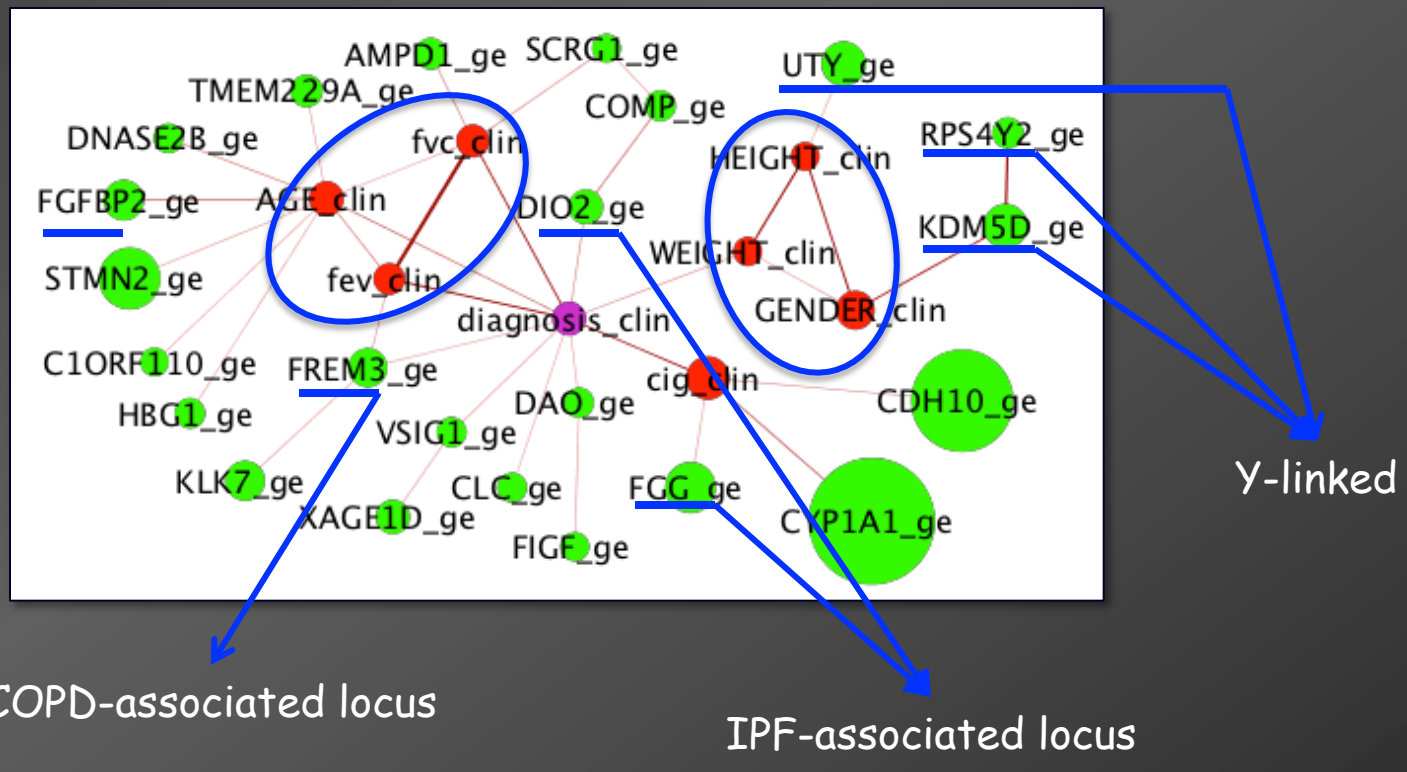
$x_1 \not\perp x_2 \mid x_3$

Identify associations between clinical features, omics and disease using graphical models (undirected)



AJ Sedgewick

We ran MGM on gene (mRNA) expression and few clinical variables



Summary

- MGM-Learn is a new causal structure learning algorithm that works on continuous and discrete variables
- MGM-Learn can be used to analyze multi-modal data of various types
- Applied to LGRC limited data it uncovers interesting associations between clinical and omics variables

When I was saying "we"…

MGM-Learn development

Andrew J Sedgewick

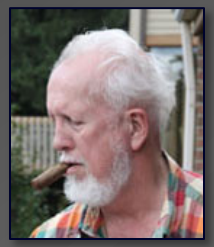
(CompBio PhD student)

MGM-Learn testing

Andrew J Sedgewick

Ivy Shi, Dept of Bioengineering, Univ of Pittsburgh

In collaboration with:



Clark Glymour, Dept Philosophy, CMU

Peter Spirtes, Dept Philosophy, CMU

Joe Ramsey, Dept Philosophy, CMU

Acknowledgements: Funding

NLM: R01 LM012087
NHLBI: R01 HL118536
NHLBI: U01 HL108642

NHGRI: U54 HG007934

Thank you!

Electronic contacts:
benos@pitt.edu
http://www.benoslab.pitt.edu

Questions???

Electronic contacts:

benos@pitt.edu

klea@pitt.edu

URL: http://www.benoslab.pitt.edu

Materials: http://www.benoslab.pitt.edu/ccd