Society for Immunotherapy of Cancer

How to Learn Cause-Effect Relations from Observational Clinical Data

SITC Cancer Immunotherapy Winter School

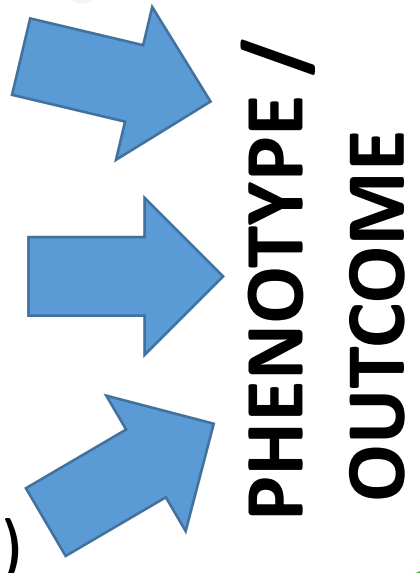
Houston, TX – January 2020

Takis Benos, PhD
Professor and Vice Chair
Department of Computational and Systems Biology
University of Pittsburgh School of Medicine

E-mail: benos@pitt.edu
Lab URL: http://www.benoslab.pitt.edu

Types of patient data currently collected

- Demographics, family history, patient's history
- Medical tests (lab tests, PFTs, vitals, etc)
- Clinical image data (H&E stained tissues, CT scans, etc)
- Omics data (gene expression, methylation, SNP/CNV, etc)



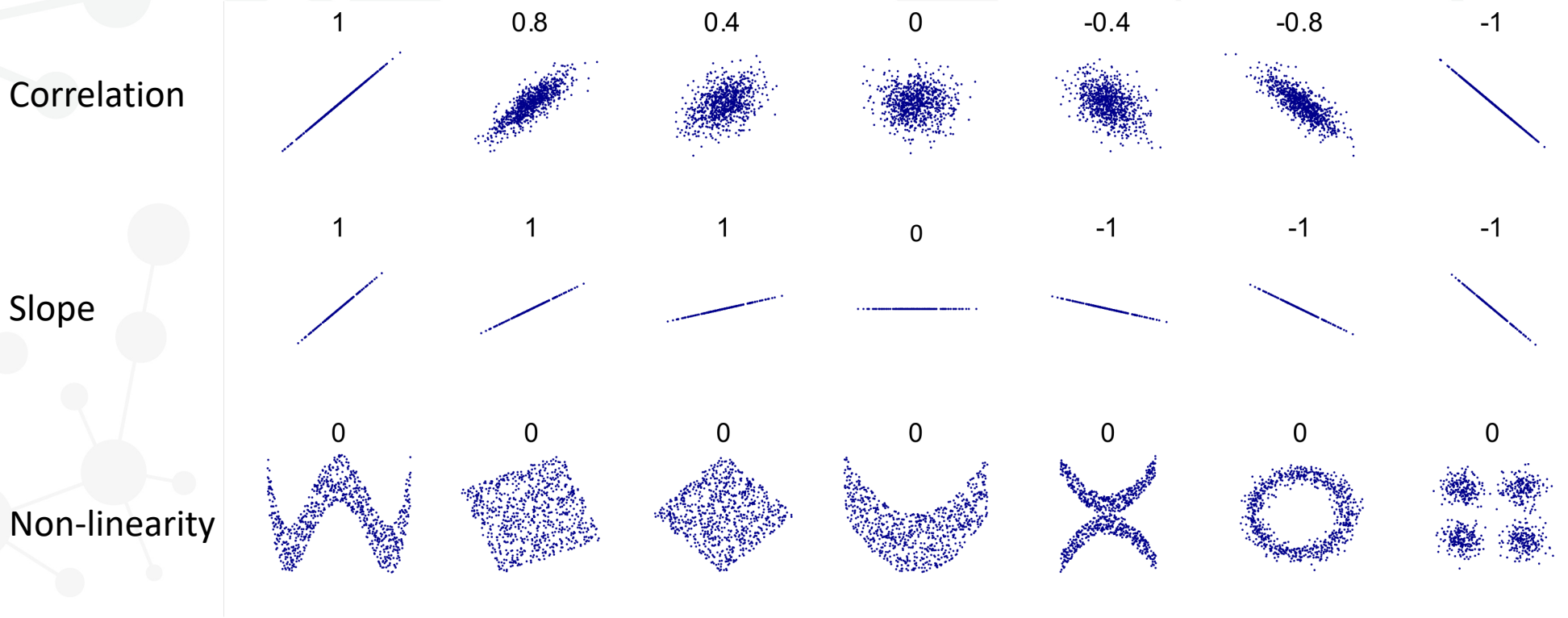
Society for Immunotherapy of Cancer

Data integration: how?

- **Correlations**
  - One variable at a time compared to other variables or outcome
- **Regressions**
  - One at a time or multiple variables compared to outcome
- **Graphical models**
  - All variables compared to all to identify direct relations
- **Other Machine learning methods**

Society for Immunotherapy of Cancer

Correlation: What does it mean?



DenisBoigelot, original uploader was Imagecreator

sitc
Society for Immunotherapy of Cancer

Types of correlation coefficients

- Pearson correlation coefficient

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

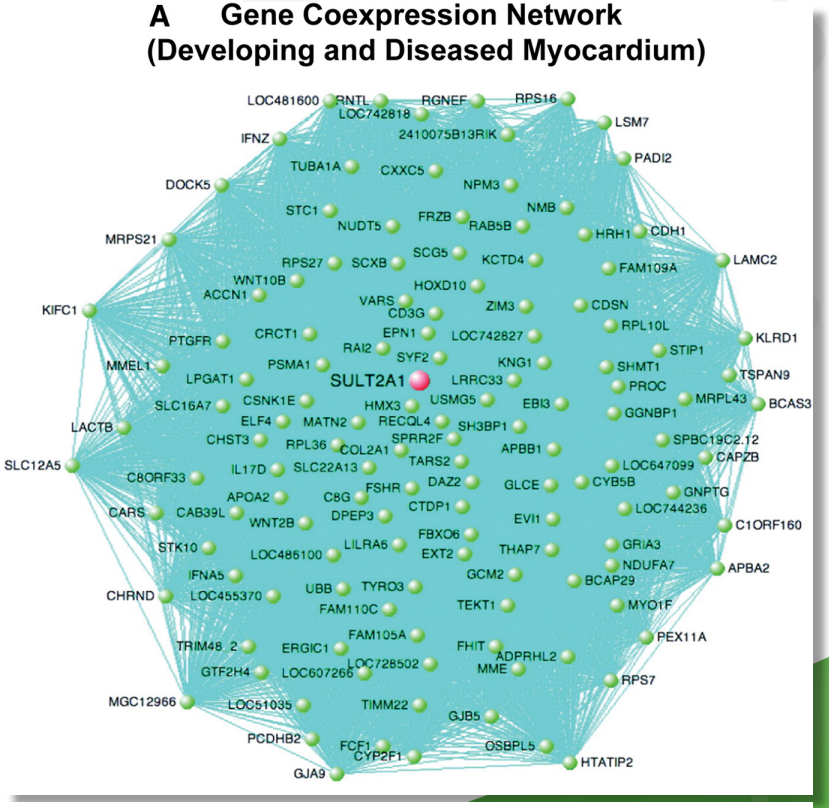
- Rank correlation coefficient
  - Spearman's rank: Pearson CC on the ranks of the values $r_s = 1 - \frac{6\sum D^2}{n(n^2-1)}$
  - Kendall's rank: $\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{n(n-1)/2}$

$$\tau = \frac{2}{n(n-1)} \sum_{i<j} \operatorname{sgn}(x_i - x_j)\operatorname{sgn}(y_i - y_j)$$

sitc
Society for Immunotherapy of Cancer

Correlation-based methods

- They are simple and thus very attractive
- They tend to overestimate the number of true connections
  - So we need to use prior or expert information to find testable hypotheses



Chan and Loscalzo, 2012, Circulation Research

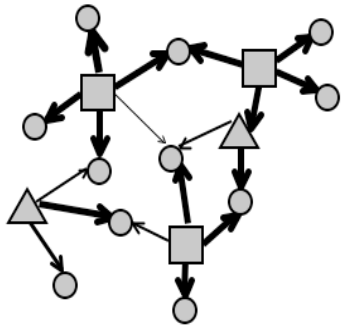
Society for Immunotherapy of Cancer

mirConnX: correlations with priors for miRNA:mRNA networks

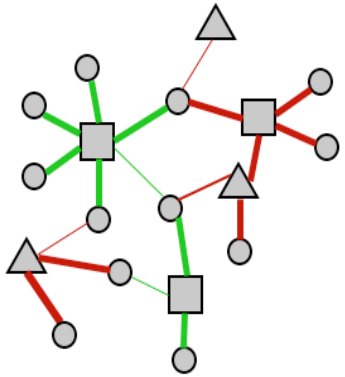
Grace Huang
PhD

miRConnX

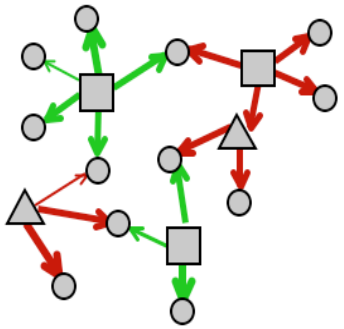
Static network

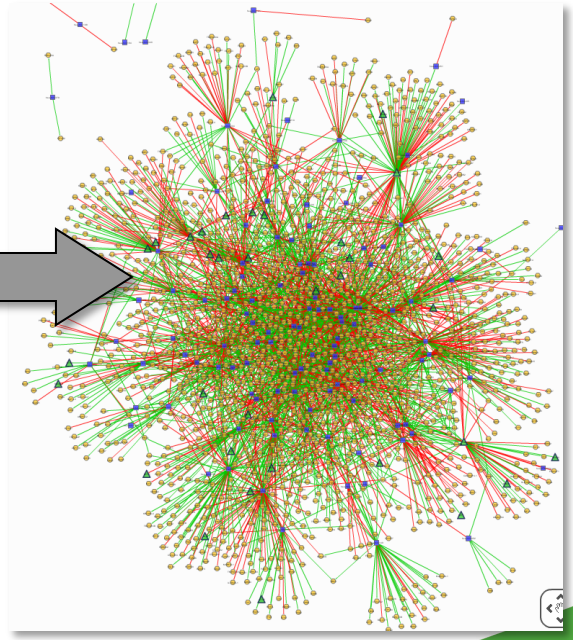


Dynamic network



Integration function





Transcription factor
microRNA
Other gene

Huang, Athanassiou, Benos, 2011, *Nucl Acids Res*

sitc
Society for Immunotherapy of Cancer

Discovery of important network module in Idiopathic Pulmonary Fibrosis (IPF)

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

Symptoms / Characteristics

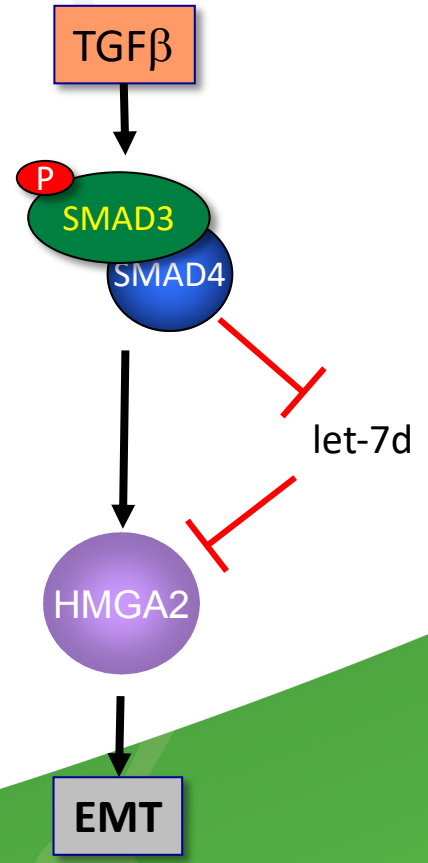
- Age: >50 yrs
- Cough w/o mucus
- Progressive dyspnea
- Characteristic “velcro-like” breathing
- Disfigurement of fingertips (clubbing of the digits)

sitc
Society for Immunotherapy of Cancer

Building a correlation regulatory network for TGF-β response

INGREDIENTS (DATA)

- mRNA and miRNA expression data
- RNA pol II ChIP assay for miRNA promoter identification
- SMAD3 ChIP assay for signaling cascade identification
- TGF-β stimulus



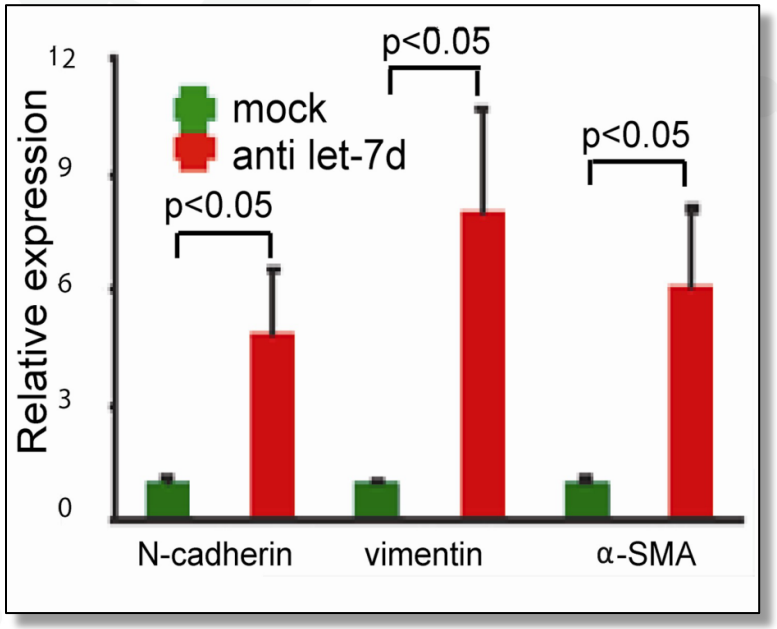
RESULT

- Identification of let-7d as a key molecule in a FFL involving SMAD and HMGA2

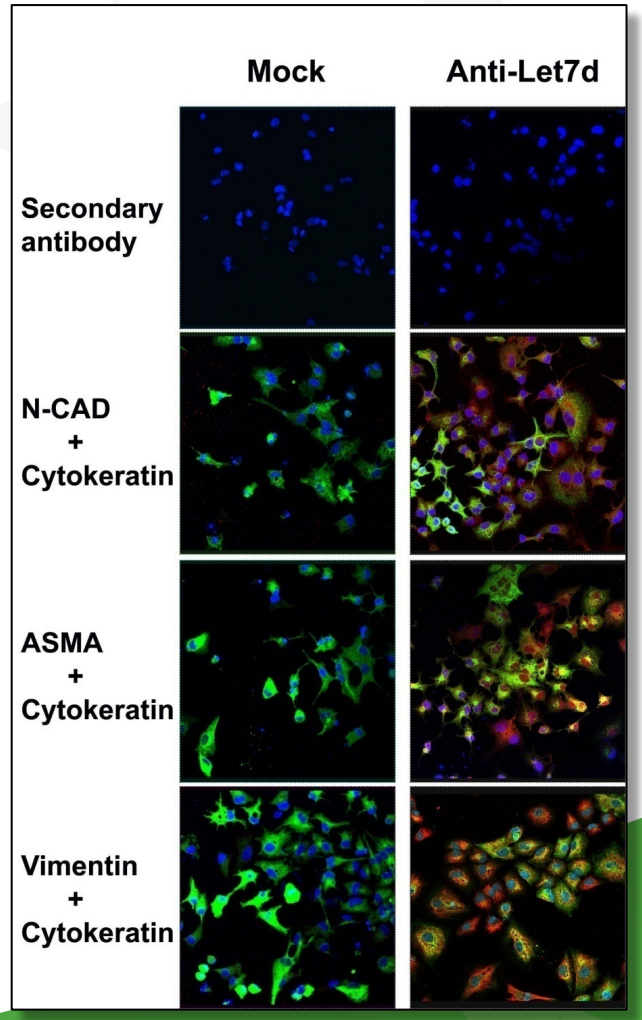
sitc
Society for Immunotherapy of Cancer

let-7d inhibition leads to EMT in cells

Naftali Kaminski MD



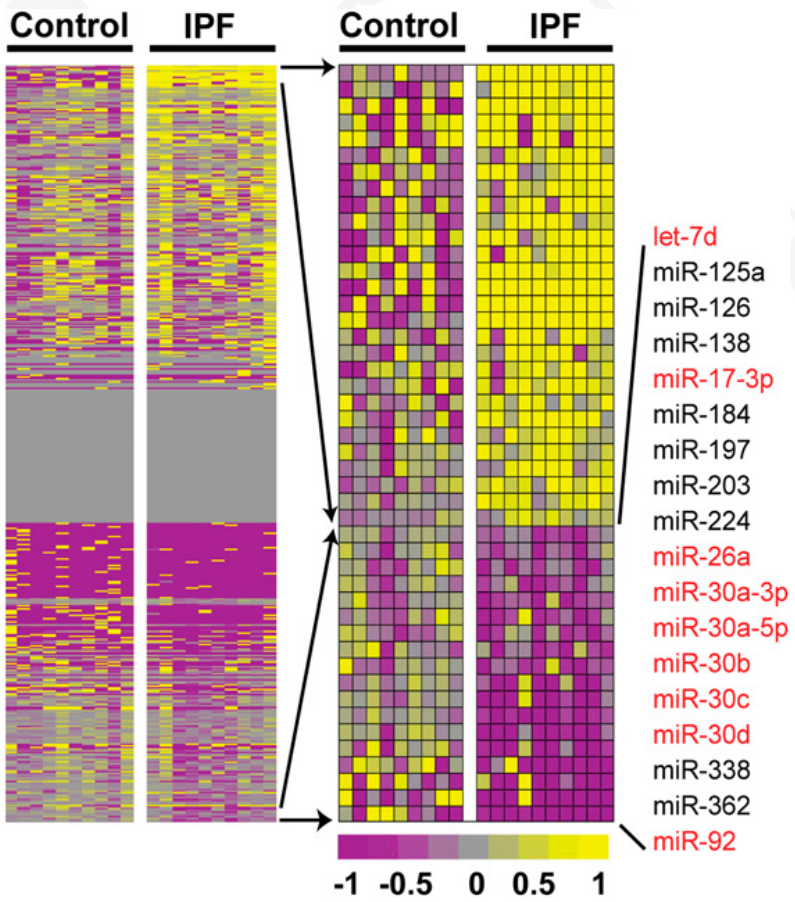
mRNA



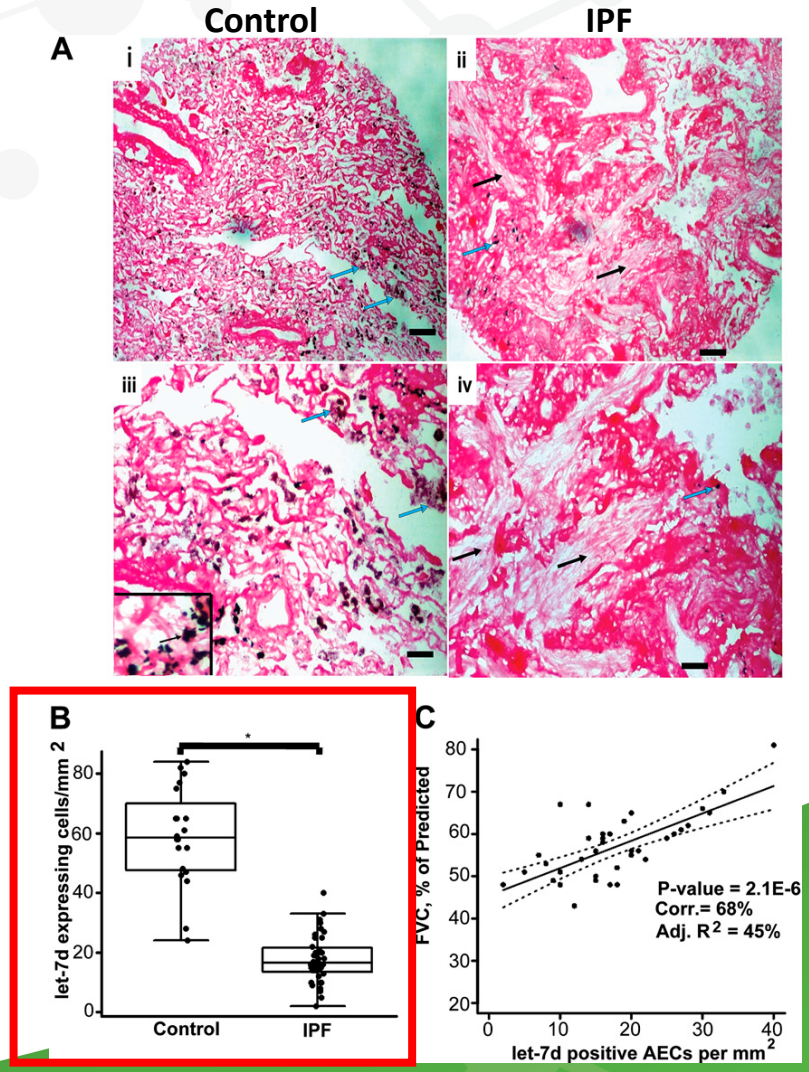
Pandit, Corcoran, …, Benos, Kaminski, 2010, *Am J Resp Critic Care Med*

sitc
Society for Immunotherapy of Cancer

let-7d is downregulated in IPF patients

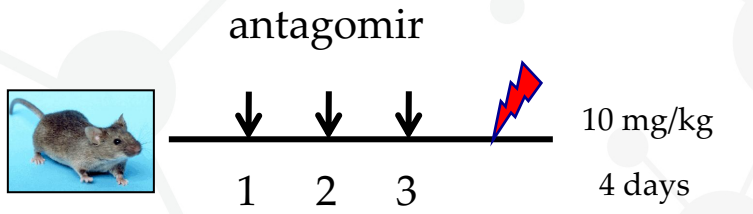


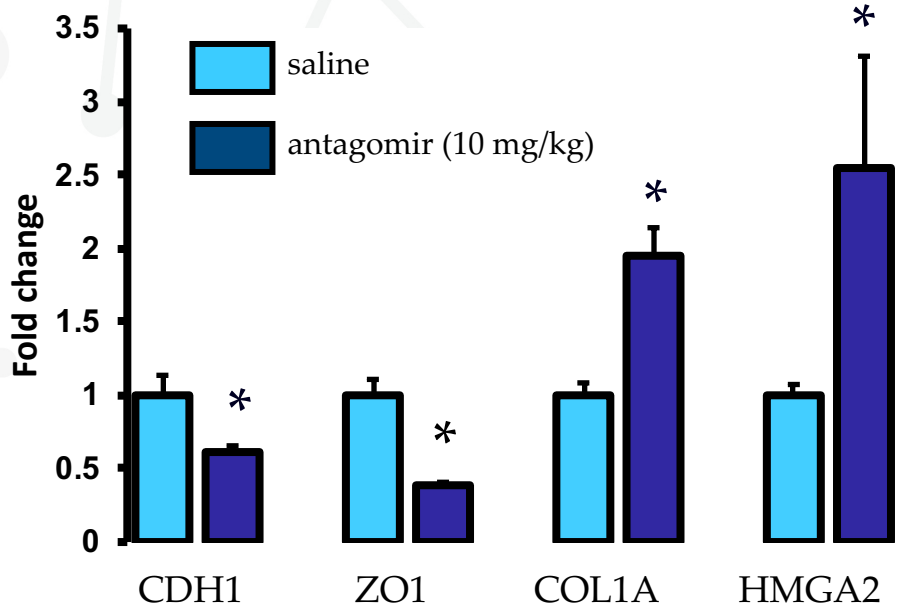
Pandit, Corcoran, …, Benos, Kaminski, 2010, *Am J Resp Critic Care Med*

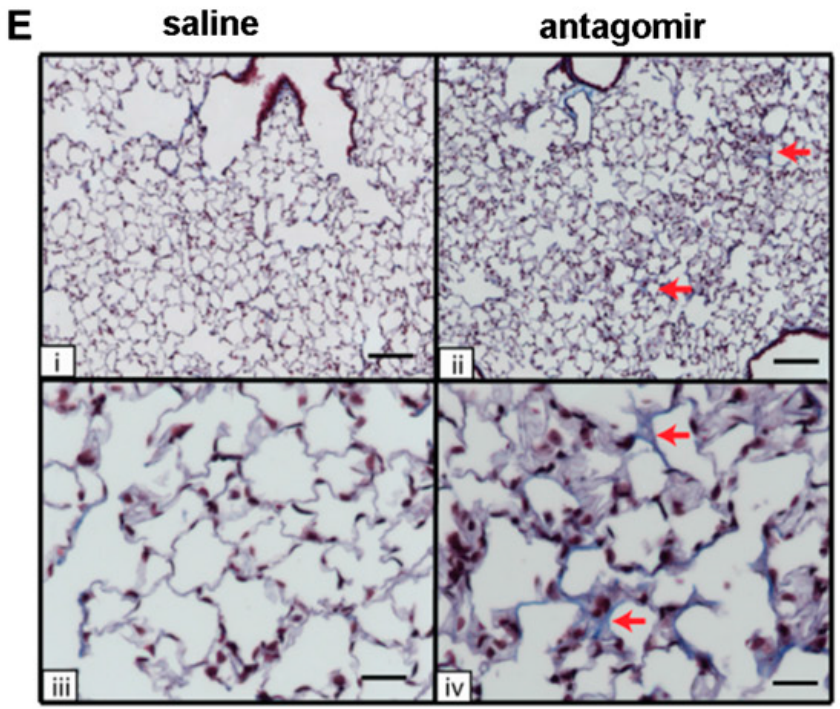


Society for Immunotherapy of Cancer

let-7d inhibition leads to lung fibrosis in mice







Collagen

Pandit, Corcoran, …, Benos, Kaminski, 2010, *Am J Resp Critic Care Med*

Society for Immunotherapy of Cancer

Correlations: what can and can not do

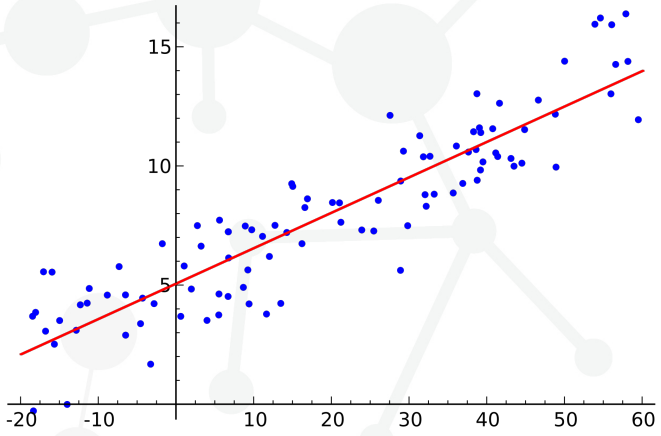
- They are easy to calculate and intuitive and can be very useful
- Provide all variables possibly related to our target variable

- Generate many ”false positive” edges
  - We need prior knowledge to generate testable hypotheses

- Correlation does not imply causation

Society for Immunotherapy of Cancer

Regression: a better way of modeling



$$\hat{Y} = \beta_0 + \sum_{i=1}^{N} \beta_i x_i + \varepsilon$$

Linear regression

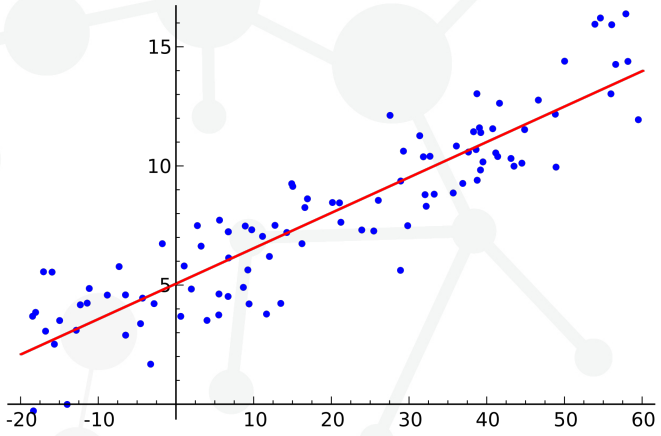
lasso

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \left(Y - \hat{Y} \right)^2 \qquad \sum_{j=1}^{p} |\beta_j| < \delta$$

$$\hat{Y} = \beta_0 + \sum_{i=1}^{N} \beta_i x_i + \beta_{age} x_{age} + \beta_{smk} x_{smk} + \cdots + \varepsilon$$

sitc
Society for Immunotherapy of Cancer

Regression: a better way of modeling

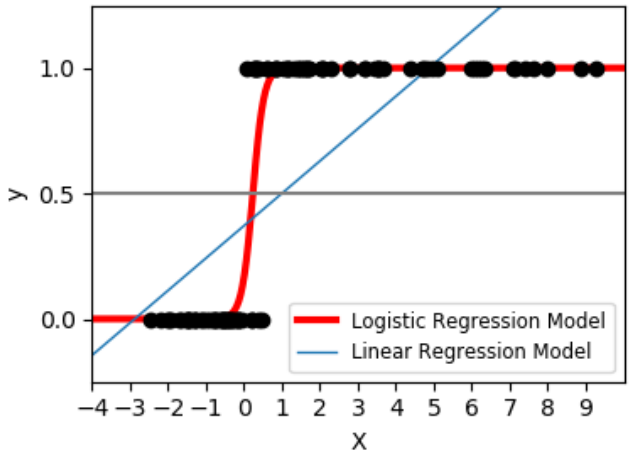


$$\hat{Y} = \beta_0 + \sum_{i=1}^{N} \beta_i x_i + \varepsilon$$ **Linear regression**

lasso

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \left(Y - \hat{Y} \right)^2 \qquad \sum_{j=1}^{p} |\beta_j| < \delta$$

$$\hat{Y} = \beta_0 + \sum_{i=1}^{N} \beta_i x_i + \beta_{age} x_{age} + \beta_{smk} x_{smk} + \cdots + \varepsilon$$



$$\ln \frac{p}{1-p} = \beta_0 + \beta_1 x_1$$ **Logistic regression**

$$P(y|x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

Society for Immunotherapy of Cancer

Regressions: applications

- QTL/eQTL analyses
  - Dependent: phenotype / Independent: SNPs
- Gene expression analyses
  - Mixed effects models to account for covariates
- Prediction / classification models

Society for Immunotherapy of Cancer

Regressions: what can and can not do

- They are intuitive and flexible
- Relatively fast to calculate
- Provide relative contributions of all predictors to the target variable

- In practice, it is not easy to implement interactive terms on predictors when number of predictors is large
  - This may result in misleading coefficients

Society for Immunotherapy of Cancer

Case study of lasso regression: optimal gastric cancer region removal

PNAS

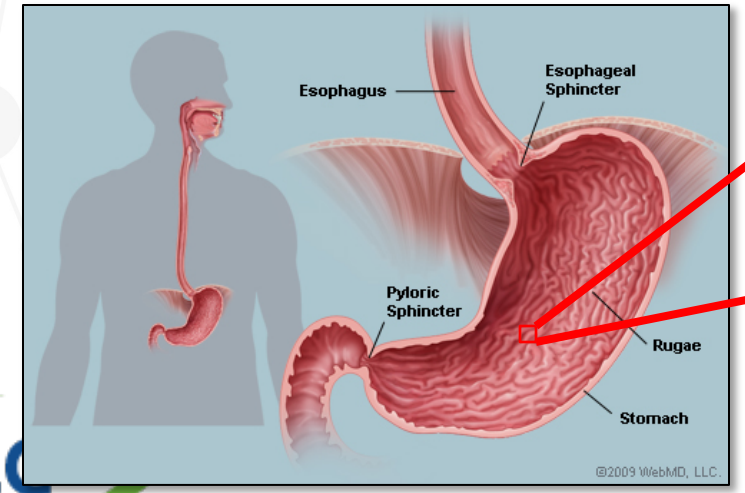
CrossMark click for updates

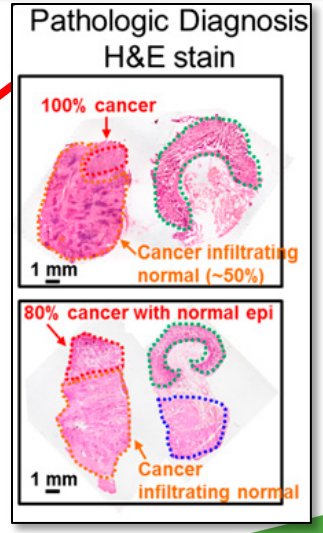
Molecular assessment of surgical-resection margins of gastric cancer by mass-spectrometric imaging

Livia S. Eberlin[a], Robert J. Tibshirani[b], Jialing Zhang[a,c], Teri A. Longacre[d], Gerald J. Berry[d], David B. Bingham[d], Jeffrey A. Norton[e], Richard N. Zare[a,1], and George A. Poultsides[e]

[a]Department of Chemistry, Stanford University, Stanford, CA 94305-5080; [b]Departments of Health Research and Policy and of Statistics, Stanford University, Stanford, CA 94305-4065; [c]Department of Chemistry, Peking University, Beijing 100871, China; [d]Department of Pathology, Stanford University, Stanford, CA 94305-5324; and [e]Department of Surgery, Stanford University, Stanford, CA 94305-5641

Contributed by Richard N. Zare, January 7, 2014 (sent for review November 11, 2013)





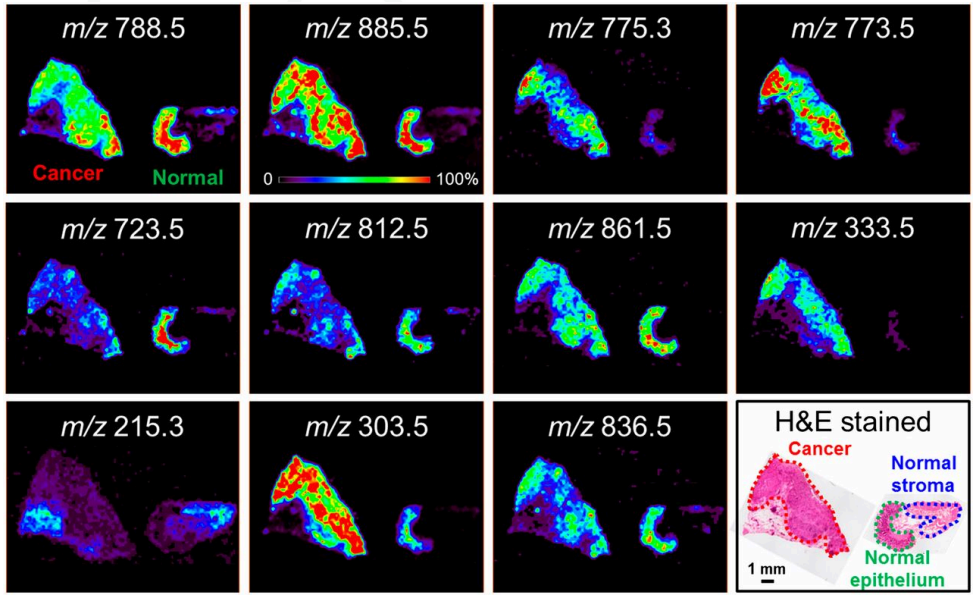
CANCER SURGERY ALGORITHM

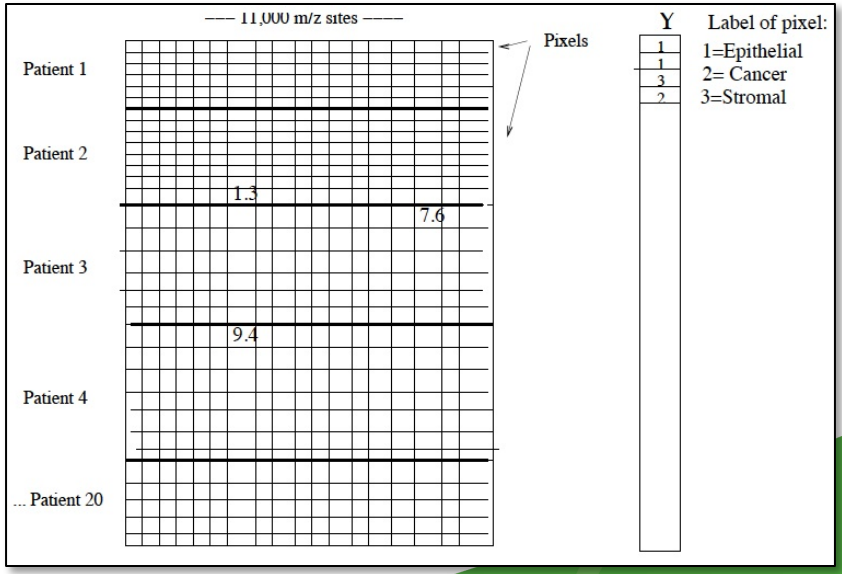
1. Surgeon removes tissue
2. Pathologist examines tissue (under the microscope)
3. If no margin is detected, GOTO #1
4. Repeat until margin is found

Society for Immunotherapy of Cancer

Technology to the rescue

- DESI (Desorption electrospray ionization)
  - An electrically charged “mist” is directed at the sample; surface ions are freed and enter the mass spec.





Society for Immunotherapy of Cancer

A modified lasso efficiently detects the tumor region

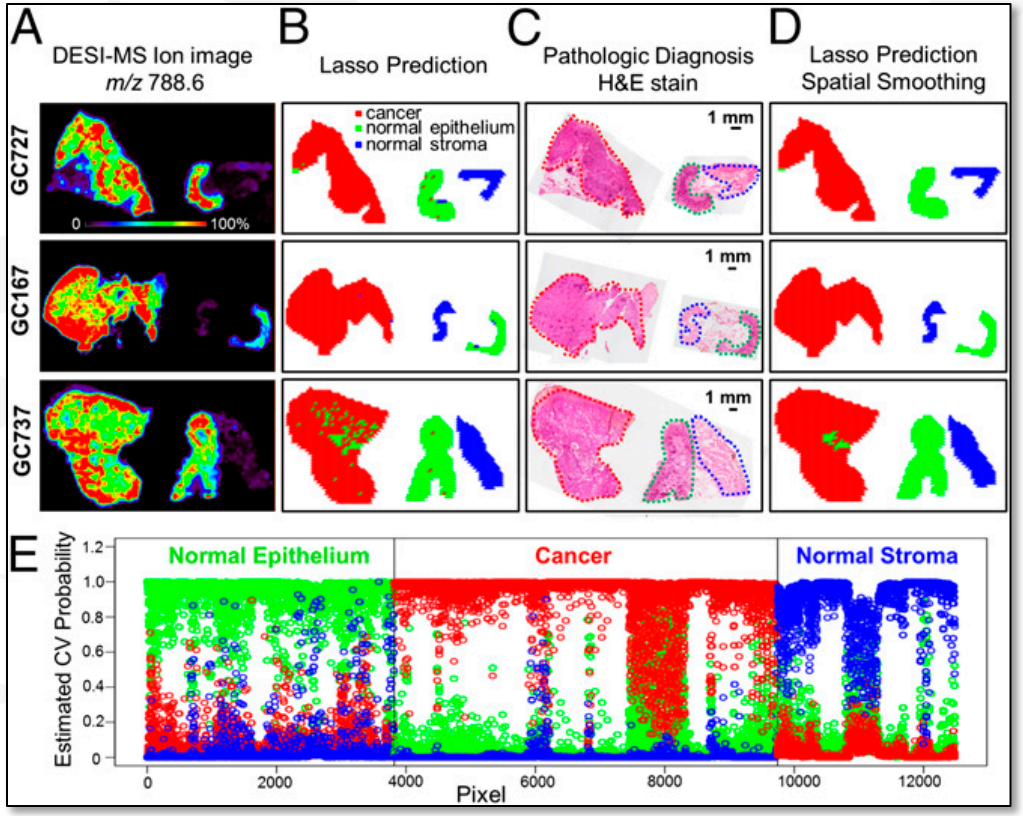


Table 1. Prediction results for the 12,480 pixels analyzed in the training set of samples, in comparison with pathological analysis

	Predicted			
Pathology*	Cancer	Epithelium	Stroma	Agreement, %
Cancer	5,809	114	2	98.0
Epithelium	134	3,566	118	93.4
Stroma	25	82	2,630	96.1
				Overall: 96.2

	Cancer	Normal	Agreement, %
Cancer	5,809	116	98.0
Normal	159	6,396	97.6
			Overall: 97.8

	Presmoothing Predicted			
Pathology	Cancer	Epithelium	Stroma	Agreement, %
Cancer	5,895	30	0	99.5
Epithelium	85	3,657	76	95.8
Stroma	4	34	2,699	98.6
				Overall: 98.2

	Cancer	Normal	Agreement, %
Cancer	5,895	30	99.5
Normal	89	6,466	98.6
			Overall: 99.0

*Pathologic analysis was performed on the same frozen tissue section used for DESI-MSI that was H&E stained after MSI analysis.

sitc
Society for Immunotherapy of Cancer

Probabilistic Graphical Models (PGMs)

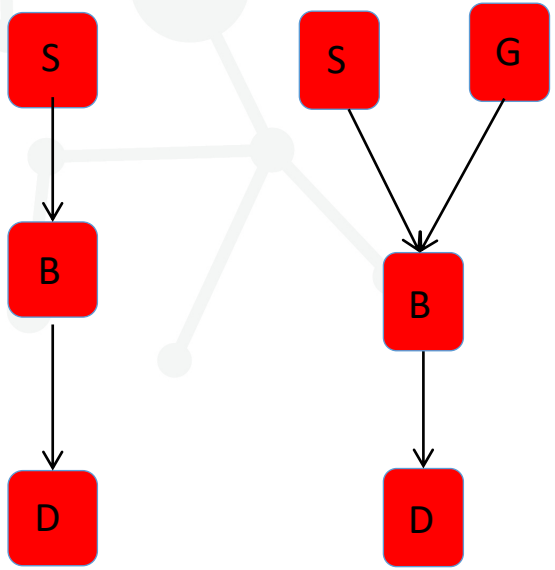
- A **graph** consists of a set of nodes (variables), some of which are connected through edges
  - Edge connections imply information transfer

- **Probabilistic graphical model** (PGM) is a model of the data in which a graph represents the conditional (in)dependencies between variables
  - PGMs can be either directed or undirected

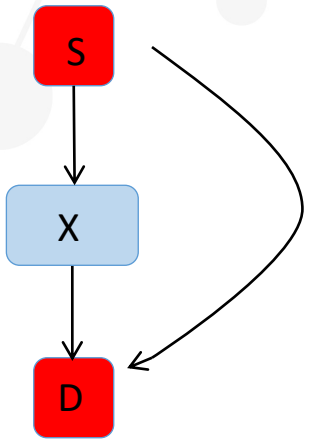
- **Causal graphs** are directed acyclic graphs (DAGs)

sitc
Society for Immunotherapy of Cancer

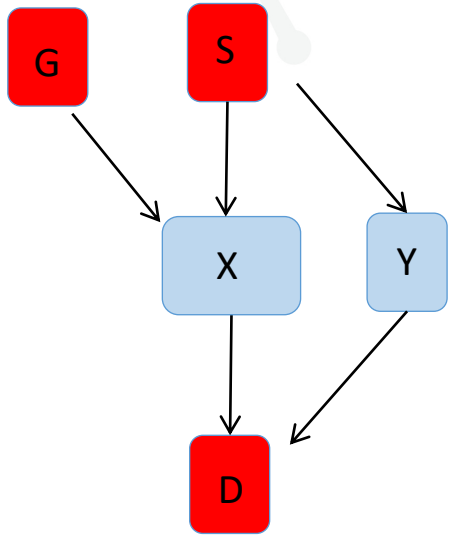
PGMs: Modeling causal dependencies



Ind(S,D | B)



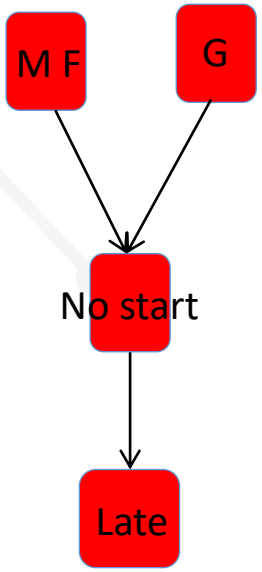
Dep(S,D | X)



Ind(S,D | {X, Y})

Society for Immunotherapy of Cancer

PGMs: Modeling orientations



Ind(MF, G | Ø)

Dep(MF, G | “No start”)

Ind(Late, G | “No start”)

Society for Immunotherapy of Cancer

Properties and Drawbacks of Graphical Models

- They can distinguish between direct and indirect effects
- They are *asymptotically* correct. 😄
- The output graph can be used to build predictive models

- They can incorporate prior information, if available ← Manatakis, Raghu and Benos, 2018
- The graph searches are relatively slow and heuristics are needed ← Raghu, Poon and Benos, 2018

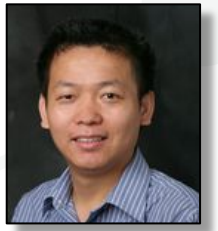
- They have some non-realistic assumptions (but they can be relaxed)
  - *There are no cycles in the graph*
  - All common causes are measured (no latent confounders) ← Raghu et al, 2017
  - Variables are either all continuous or all discrete ← Sedgewick et al, 2016
  - *All continuous variables should be normally distributed* ← Sedgewick et al, 2019

sitc
Society for Immunotherapy of Cancer

Vineet Raghu

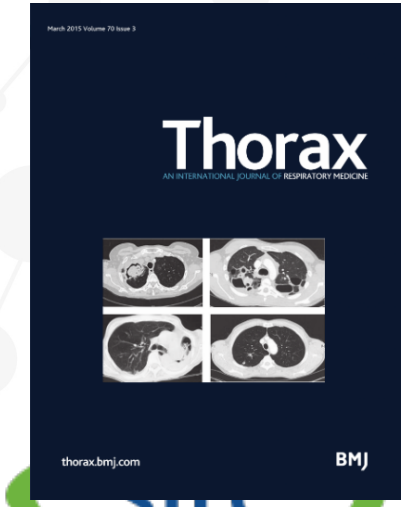
In collaboration with:

David Wilson MD



Jiantao Pu PhD

Factors determining malignancy of a lung nodule from low-dose CT scan and clinical data



Society for Immunotherapy of Cancer

Lung cancer

OPEN ACCESS

ORIGINAL ARTICLE

Feasibility of lung cancer prediction from low-dose CT scan and smoking factors using causal models

Vineet K Raghu,[1,2] Wei Zhao,[3,4] Jiantao Pu,[3] Joseph K Leader,[3] Renwei Wang,[5] James Herman,[6] Jian-Min Yuan,[5,7] Panayiotis V Benos,[1,2] David O Wilson[8]

* Corresponding author

Low dose CT scan screening reduces lung cancer mortality

The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 AUGUST 4, 2011 VOL. 365 NO. 5

Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening

The National Lung Screening Trial Research Team*

ABSTRACT

BACKGROUND

The aggressive and heterogeneous nature of lung cancer has thwarted efforts to reduce mortality from this cancer through the use of screening. The advent of low-dose helical computed tomography (CT) altered the landscape of lung-cancer screening, with studies indicating that low-dose CT detects many tumors at early stages. The National Lung Screening Trial (NLST) was conducted to determine whether screening with low-dose CT could reduce mortality from lung cancer.

The members of the writing team (who are listed in the Appendix) assume responsibility for the integrity of the article. Address reprint requests to Dr. Christine D. Berg at the Early Detection Research Group, Division of Cancer Prevention, National Cancer Institute, 6130 Executive Blvd., Suite 3112, Bethesda, MD 20892-7346, or at bergc@mail.nih.gov.

RESULTS

The rate of adherence to screening was more than 90%. The rate of positive screening tests was 24.2% with low-dose CT and 6.9% with radiography over all three rounds. A total of 96.4% of the positive screening results in the low-dose CT group and 94.5% in the radiography group were false positive results. The incidence of lung cancer was 645 cases per 100,000 person-years (1060 cancers) in the low-dose CT group, as compared with 572 cases per 100,000 person-years (941 cancers) in the radiography group (rate ratio, 1.13; 95% confidence interval [CI], 1.03 to 1.23). There were 247 deaths from lung cancer per 100,000 person-years in the low-dose CT group and 309 deaths per 100,000 person-years in the radiography group, representing a relative reduction in mortality from lung cancer with low-dose CT screening of 20.0% (95% CI, 6.8 to 26.7; P=0.004). The rate of death from any cause was reduced in the low-dose CT group, as compared with the radiography group, by 6.7% (95% CI, 1.2 to 13.6; P=0.02).

CONCLUSIONS

Screening with the use of low-dose CT reduces mortality from lung cancer. (Funded by the National Cancer Institute; National Lung Screening Trial ClinicalTrials.gov number, NCT00047385.)

- Follow-up CTs
- Unnecessary invasive biopsies
  - with potential serious complications
- Anxiety
- Increased healthcare costs

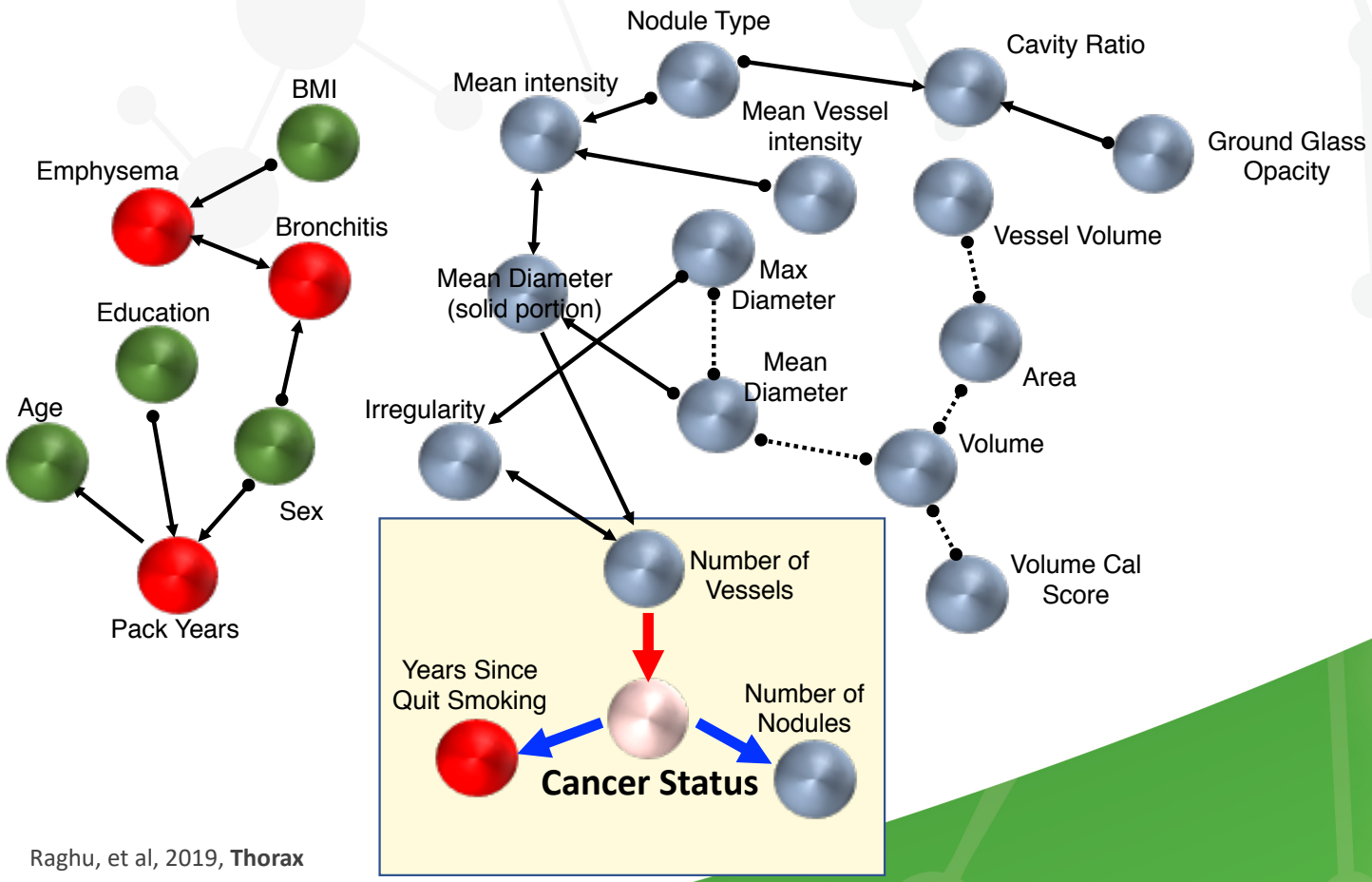
sitc
Society for Immunotherapy of Cancer

© Benos Lab / University of Pittsburgh 2014-2019

LCCM: Lung Cancer Causal Model

A. Training cohort	Lung cancer (n = 50)	Benign nod. (n = 42)	P value
Male, n (%)	25 (50)	28 (67)	0.162
Age (years), mean (SD)	63.6 (7.1)	65.2 (6.9)	0.261

B. Validation cohort	Lung cancer (n = 44)	Benign nod. (n = 82)	P value
Male, n (%)	23 (52)	48 (59)	0.626
Age, mean, years (SD)	65.23 (9.62)	66.93 (7.54)	0.313

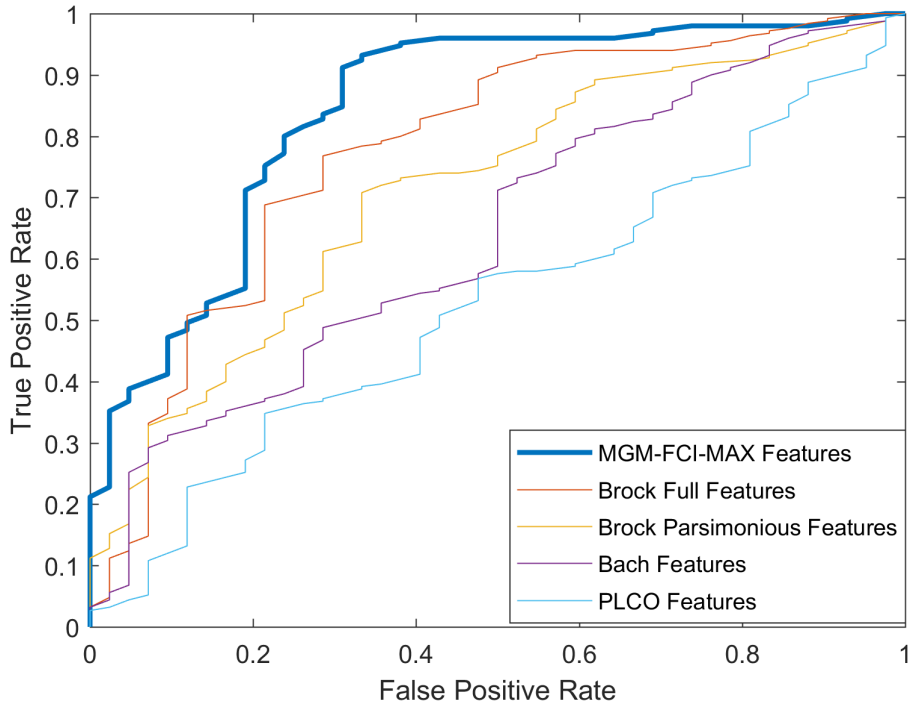


Raghu, et al, 2019, **Thorax**

Society for Immunotherapy of Cancer

© Benos Lab / Univ of Pittsburgh 2014-2019

LCCM outperforms existing lung cancer predictors (cross-validation)



Model	No. of Features	AUC (95% CI)	p-value	Features Used
MGM-FCI-MAX Features	3	0.882 (0.789, 0.975)	-	**Smoking:** Years Quit **Radiographic:** Nodule Count, Vessel Number
Brock Full Features	8	0.792 (0.699,0.885)	0.16	**Demographics:** Age, Sex, Family History Ca **Comorbidities:** Emphysema **Radiographic:** Nodule Size, Nodule Type, Nodule Location, Nodule Count
Brock Parsimonious Features	3	0.700 (0.607,0.793)	0.01	**Demographics**: Sex **Radiographic:** Nodule Location, Nodule Size
Bach Features	5	0.722 (0.629,0.815)	0.02	**Demographics**: Age, Sex **Smoking:** Cigarettes Per Day, Smoke Duration, Years Quit
PLCO Features	10	0.5613 (0.412,0.701)	<0.001	**Demographics**: BMI, Education, Family History Ca, Race **Comorbidities**: Ca History, COPD **Smoking**: Duration, Intensity, Smoking Status, Years Quit

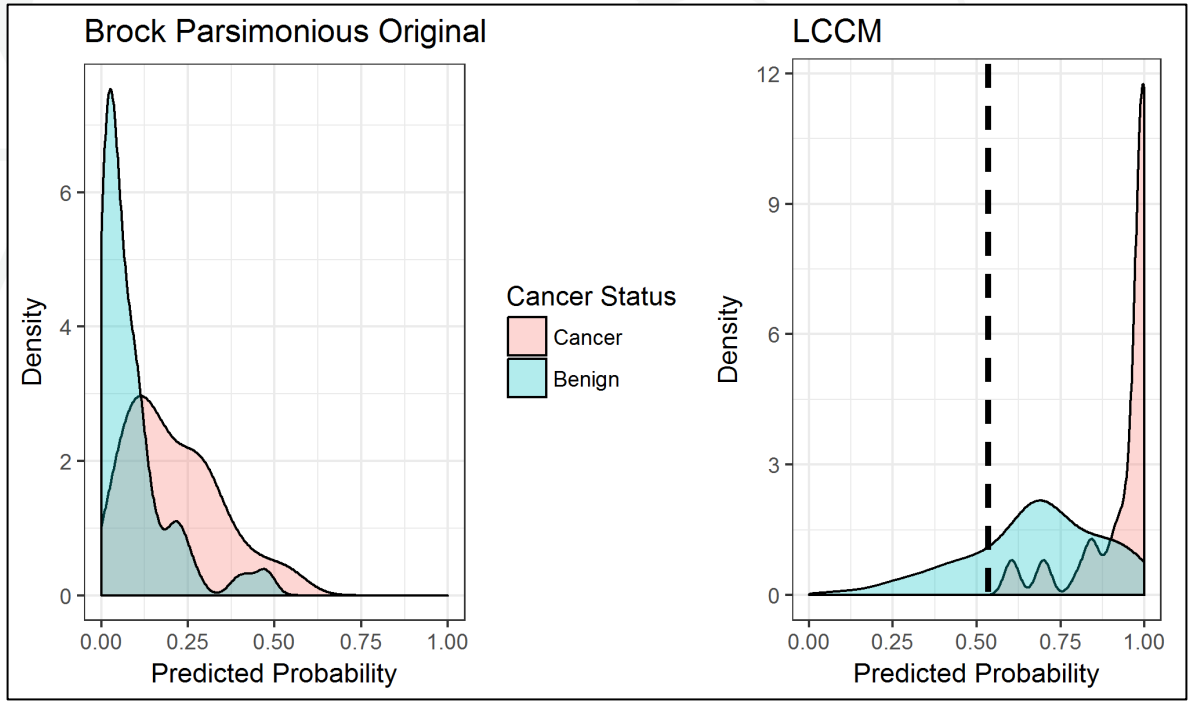
Predictors	Coefficient (95% CI)	*p*-value
Years since quit smoking	-0.178 (-0.349, -0.007)	0.041
Number of Vessels	0.238 (0.074, 0.510)	0.009
Number of Nodules	-0.203 (-0.325, -0.081)	0.001
Model Intercept	1.053	

Raghu, et al, 2019, Thorax

Society for Immunotherapy of Cancer

© 2019-2020 Benos Lab / University of Pittsburgh

LCCM can help reduce unnecessary follow up screenings



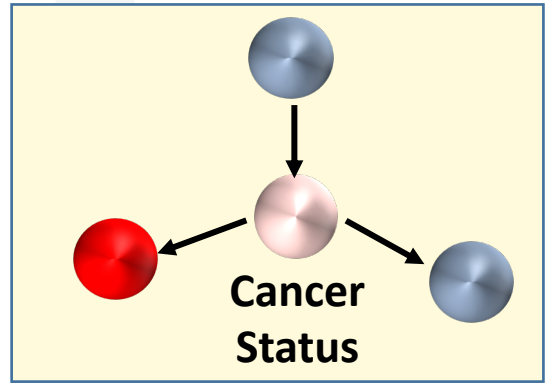
Raghu, et al, 2019, Thorax

Society for Immunotherapy of Cancer

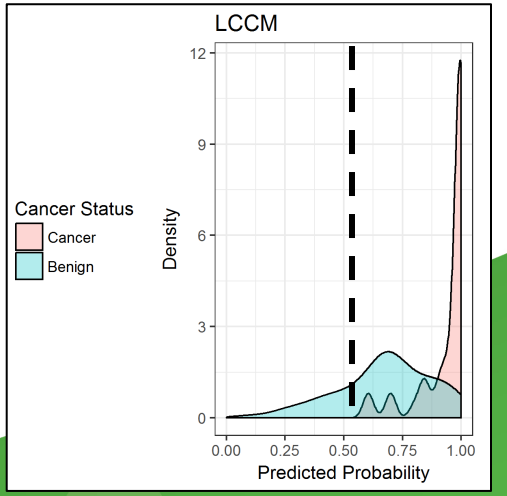
© 2019-2020 Benos Lab / University of Pittsburgh

What we learned from the LCCM study?

- Vasculature around a nodule and total number of nodules are important discriminants of nodule status



- LCCM in the future may help reduce unnecessary follow up screens for 28% of the benign nodule subjects



Society for Immunotherapy of Cancer

© 2019-2020 Benos Lab / University of Pittsburgh

AJ Sedgewick PhD

In collaboration with:

Hussein Tawbi MD

Disclosure:
US Patent Application No. 15/524,242, filed May 3, 2017

A SNP that predicts response to chemotherapy and suggests new combination therapy

SCIENTIFIC REPORTS

Article | OPEN ACCESS | Published: 01 March 2019

PARP1 rs1805407 Increases Sensitivity to PARP1 Inhibitors in Cancer Cells Suggesting an Improved Therapeutic Strategy

Irina Abecassis, Andrew J. Sedgewick, Marjorie Romkes, Shama Buch, Tomoko Nukui, Maria G. Kapetanaki, Andreas Vogt, John M. Kirkwood, Panayiotis V. Benos ✉ & Hussein Tawbi ✉

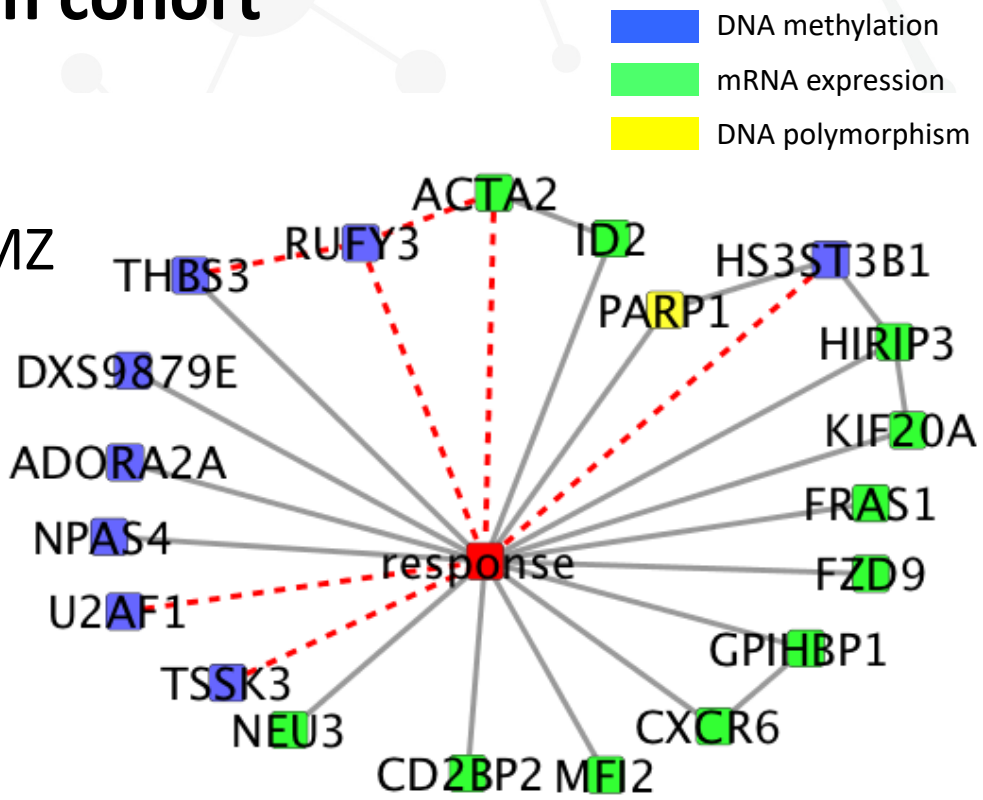
Scientific Reports **9**, Article number: 3309 (2019) | Download Citation ⬇

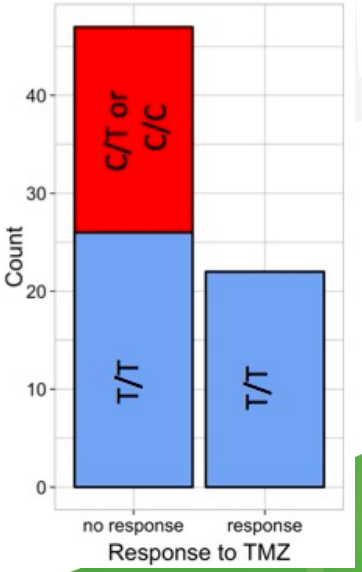
sitc
Society for Immunotherapy of Cancer

Identify cancer chemotherapy biomarkers

Hussein Tawbi MD

- **Metastatic melanoma Pittsburgh cohort**
- Subjects:
  - 69 subjects
  - Demographics and response to TMZ treatment
- **Data acquisition from tumor:**
  - Gene expression
  - miRNA expression
  - DNA methylation
  - SNP assay (selected SNPs)





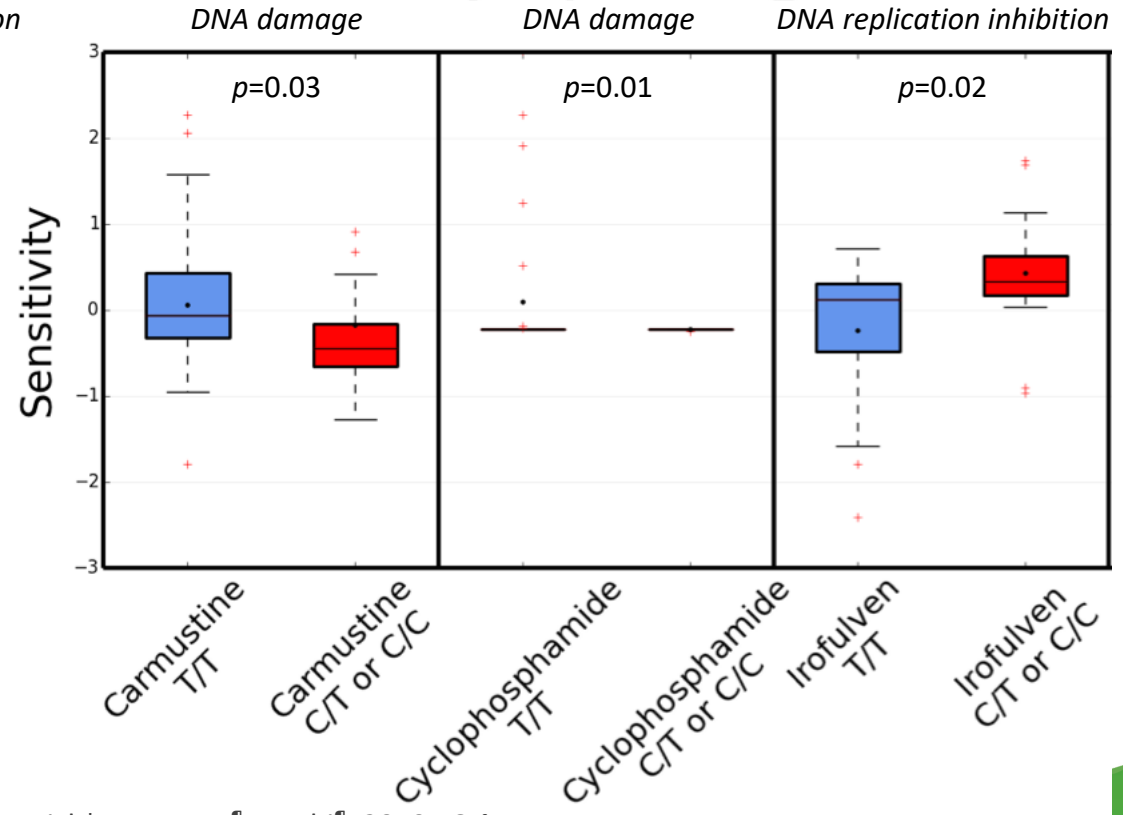
$p=10^{-5}$

Abecassis*, Sedgewick*, …, Benos¶, Tawbi¶, 2019, ***Sci Rep***,

Society for Immunotherapy of Cancer

Alkylating agents induce the strongest changes in drug sensitivity between carriers/non-carriers

AJ Sedgewick PhD



Abecassis*, Sedgewick*, …, Benos¶, Tawbi¶, 2019, ***Sci Rep***,

Society for Immunotherapy of Cancer

Hypothesis (testable)

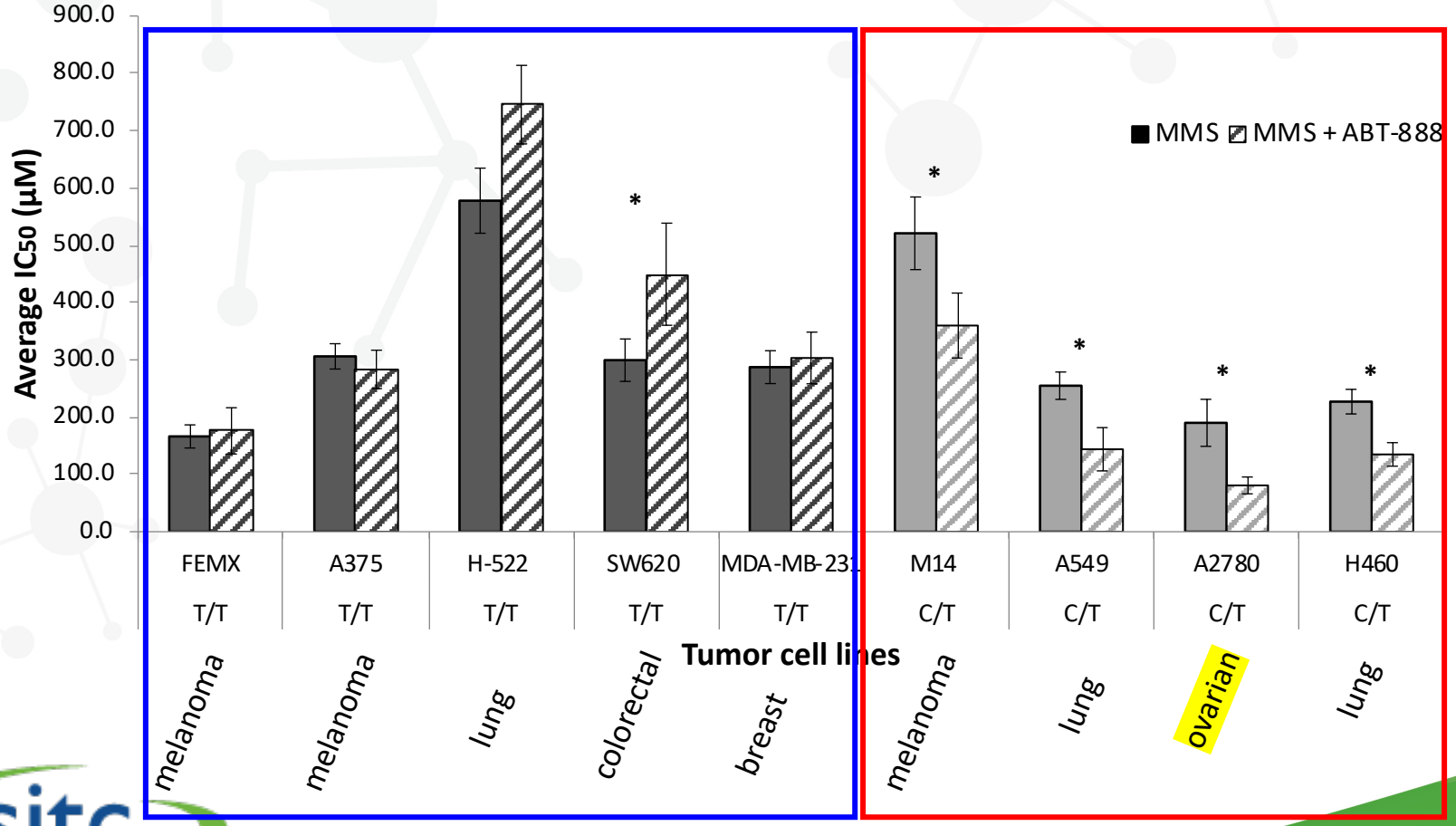
- The PARP1 SNP is directly related to improved DNA damage repair
  - Improved DNA damage repair ➔ worse response to chemotherapy

- Testing:

  Treat cells with PARP inhibitor (PARPi) ➔ do SNP cells require lower doses of alkylating agent than WT cells? (lower IC_{50})

Society for Immunotherapy of Cancer

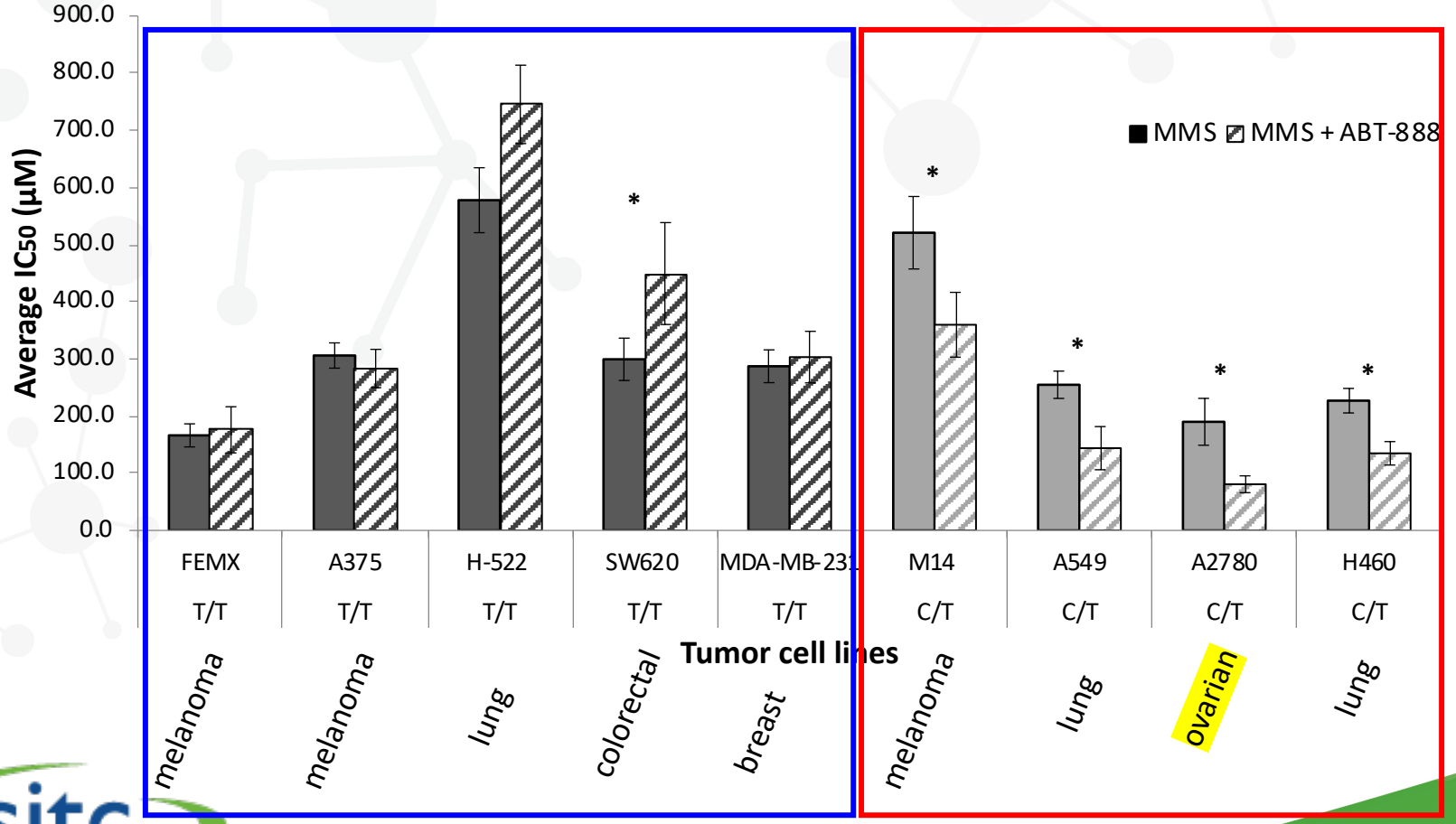
PARP-1 inhibition increases chemo efficiency to cell lines with the SNP

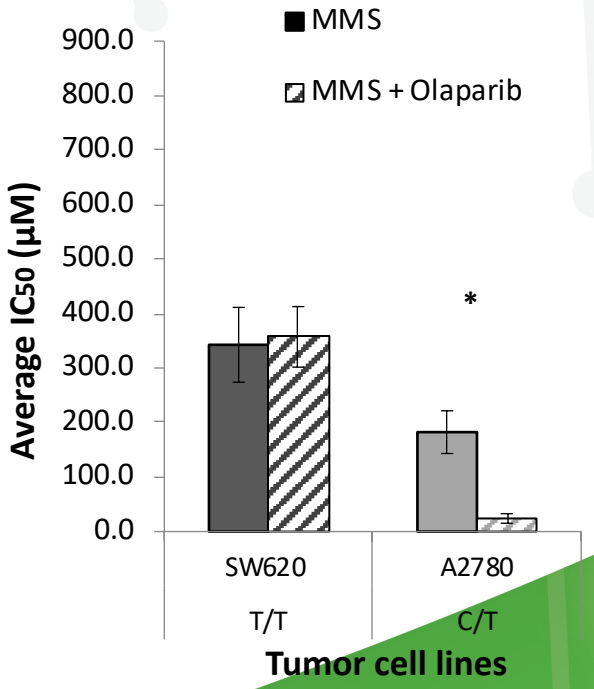


Abecassis*, Sedgewick*, …, Benos¶, Tawbi¶, 2019, ***Sci Rep***,

Society for Immunotherapy of Cancer

PARP-1 inhibition increases chemo efficiency to cell lines with the SNP

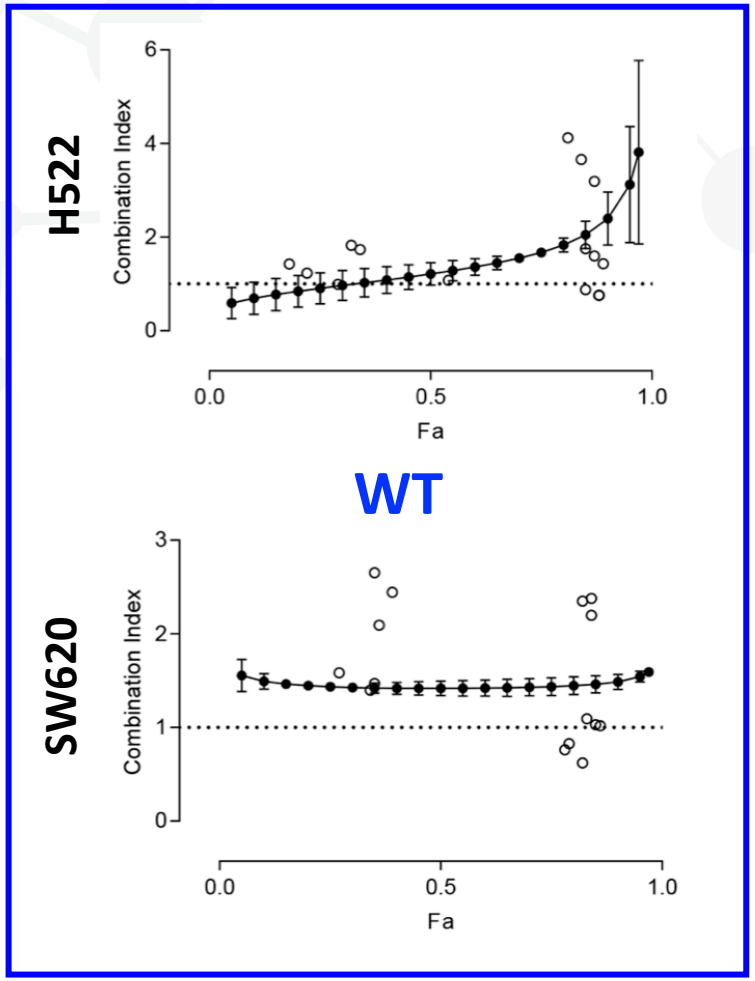


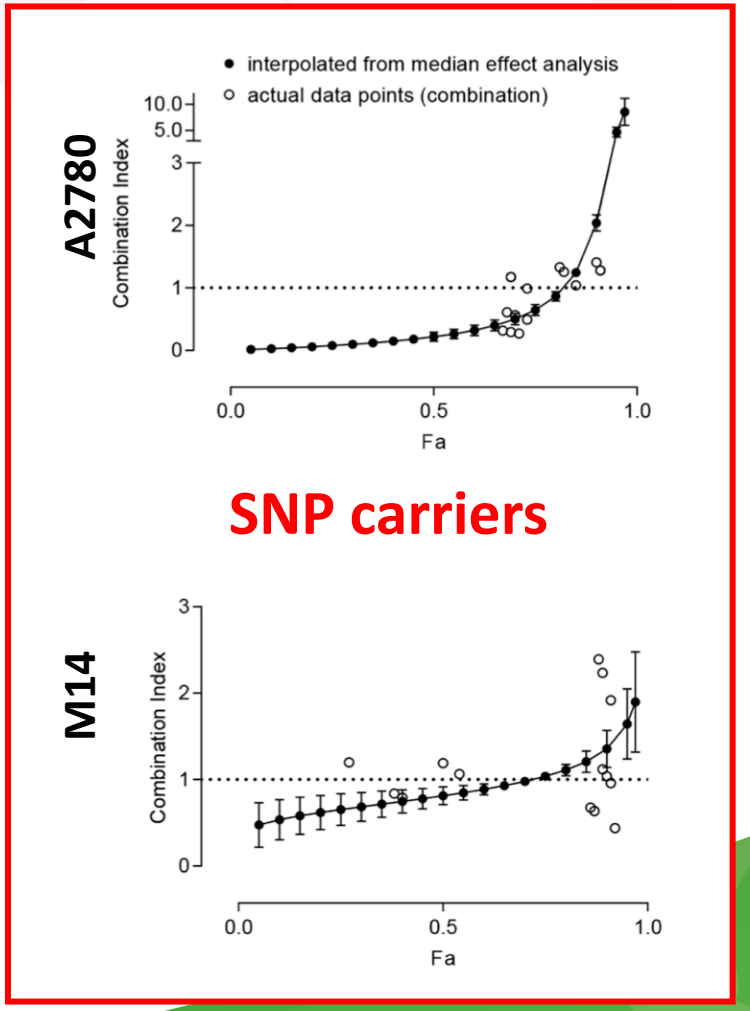


Abecassis*, Sedgewick*, …, Benos¶, Tawbi¶, 2019, ***Sci Rep***,

Society for Immunotherapy of Cancer

PARP-1 inhibition increases chemo efficiency to cell lines with the SNP





Abecassis*, Sedgewick*, …, Benos¶, Tawbi¶, 2019, ***Sci Rep***,

Society for Immunotherapy of Cancer

Summary of *Causal-MGM* applications

- ✓We identified blood biomarker proteins and comorbidities that are directly linked to longitudinal lung function decline in COPD patients (creatinine, TNF-α, GERD, etc)

- ✓We identified a PARP1 SNP that is a marker for no response to chemotherapy and we found evidence to suggest that the SNP carriers may benefit from combination therapy (chemo + PARP1 inhibitors)

sitc
Society for Immunotherapy of Cancer

Acknowledgements:
Some current collaborations

Causal modeling on mixed data (NLM R01)

Clark Glymour, PhD – Philosophy, CMU

Peter Spirtes, PhD – Philosophy, CMU

Joe Ramsey, PhD – Philosophy, CMU

Cloud interfaces (NHLBI U01)

Panos Chrysanthis, PhD – Computer Science, Pitt

COPD progression & subtyping (NHLBI U01)

Frank Sciurba, MD – Pulmonary Medicine, Pitt / UPMC

Biomarkers for cancer treatment response (NLM R01)

John M. Kirkwood MD –Medicine, Pitt / UPMC

Hussein Tawbi MD – MD Anderson

FUNDING

NHLBI, NLM, NCI, NHGRI (BD2K)

sitc

Society for Immunotherapy of Cancer

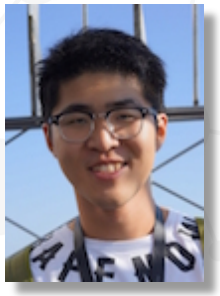
© 2019-2020 Benos Lab / University of Pittsburgh

Benos' laboratory

MD FELLOW

Craig Riley MD
(co-advised: Frank Sciurba)

ISB STUDENT



Feng Shan
(co-advised: Dario Vignalli)

sitc
Society for Immunotherapy of Cancer

Electronic contacts:

benos@pitt.edu

http://www.benoslab.pitt.edu

Questions???

Electronic contacts:

benos@pitt.edu

http://www.benoslab.pitt.edu

References (algorithms):

- Sedgewick et al, “Learning mixed graphical models with separate sparsity parameters and stability-based model selection”, 2016, ***BMC Bioinformatics***
- Sedgewick et al, “Mixed Graphical Models for Integrative Causal Analysis with Application to Chronic Lung Disease Diagnosis and Prognosis”, 2019, ***Bioinformatics***
- Raghu et al, “Comparison of Strategies for Scalable Causal Discovery of Latent Variable Models from Mixed Data”, 2019, ***Interntl J of Data Science and Analytics***
- Manatakis*, Raghu*, Benos, “piMGM: Incorporating Multi-Source Priors in Mixed Graphical Models for Learning Disease Networks”, 2018, ***Bioinformatics***
- Buschur, Chikina, Benos, “Causal network perturbations for instance-specific analysis of single cell and disease samples”, 2019, ***Bioinformatics***
- Raghu et al, “Evaluation of Causal Structure Learning Methods on Mixed Data Types”, 2018, ***Proc Mach Learn Res***