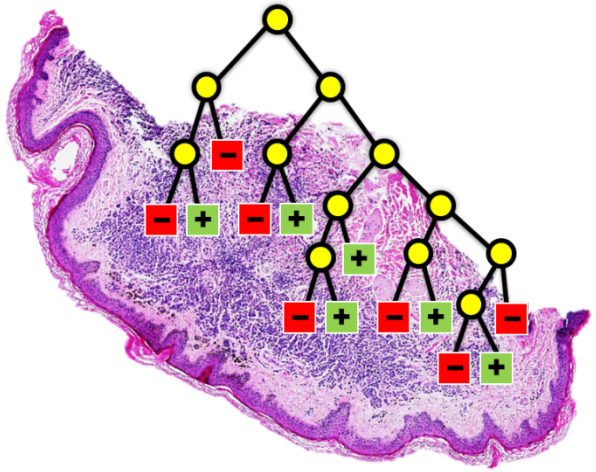
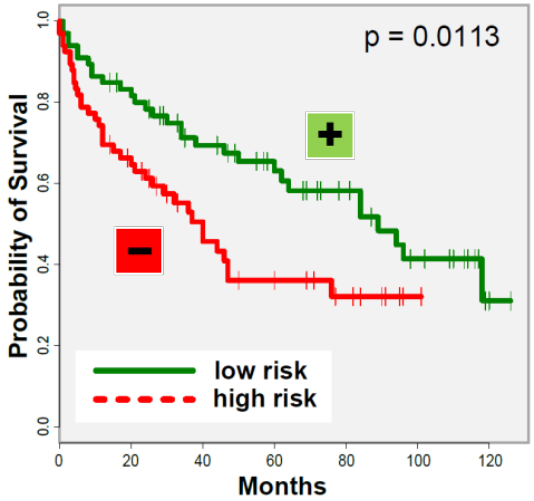
Thomas J. Fuchs
Dan Crichton (PI)

Reproducible Data Science in EDRN -

Exemplified with the PLCO Ovarian Phase III Validation Study





Previous EDRN Reproducibility Studies

- Lung Cancer
- Pancreas Cancer

VOLUME 26 · NUMBER 31 · NOVEMBER 1 2008

JOURNAL OF CLINICAL ONCOLOGY

ORIGINAL REPORT

Occurrence of Autoantibodies to Annexin I, 14-3-3 Theta and LAMR1 in Prediagnostic Lung Cancer Sera

Ji Qiu, Gina Choi, Lin Li, Hong Wang, Sharon J. Pitteri, Sandra R. Pereira-Faca, Alexei L. Krasnoselsky, Timothy W. Randolph, Gilbert S. Omenn, Cim Edelstein, Matt J. Barnett, Mark D. Thornquist, Gary E. Goodman, Dean E. Brenner, Ziding Feng, and Samir M. Hanash

From the Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA; and Department of Internal Medicine and Center for Computational Medicine and Biology, and Comprehensive Cancer Center, University of Michigan, Ann Arbor, MI.

Submitted January 16, 2008; accepted June 12, 2008; published online ahead of print at www.jco.org on September 15, 2008.

Supported by funding from the National Cancer Institute Early Detection Research Network program and the Canary Foundation.

Authors' disclosures of potential conflicts of interest and author contributions are found at the end of this article.

Corresponding author: Ji Qiu, PhD, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave North, M5-C800, Seattle, WA 98109; e-mail: jiqiu@fhcrc.org.

0732-183X/08/2631-5060/$20.00

DOI: 10.1200/JCO.2008.16.2388

A B S T R A C T

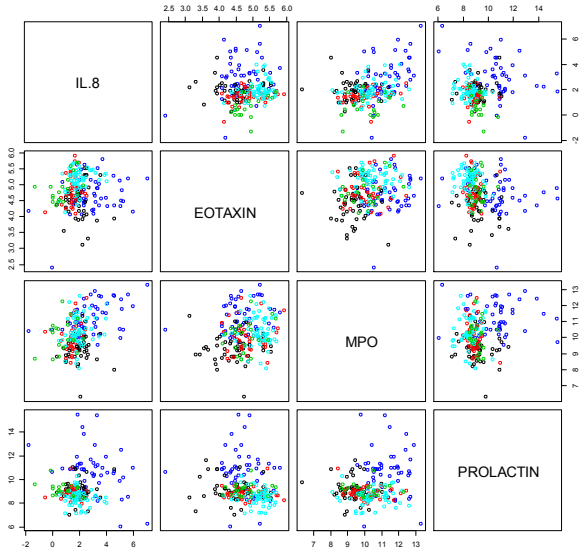
Purpose
We have implemented a high throughput platform for quantitative analysis of serum autoantibodies, which we have applied to lung cancer for discovery of novel antigens and for validation in prediagnostic sera of autoantibodies to antigens previously defined based on analysis of sera collected at the time of diagnosis.

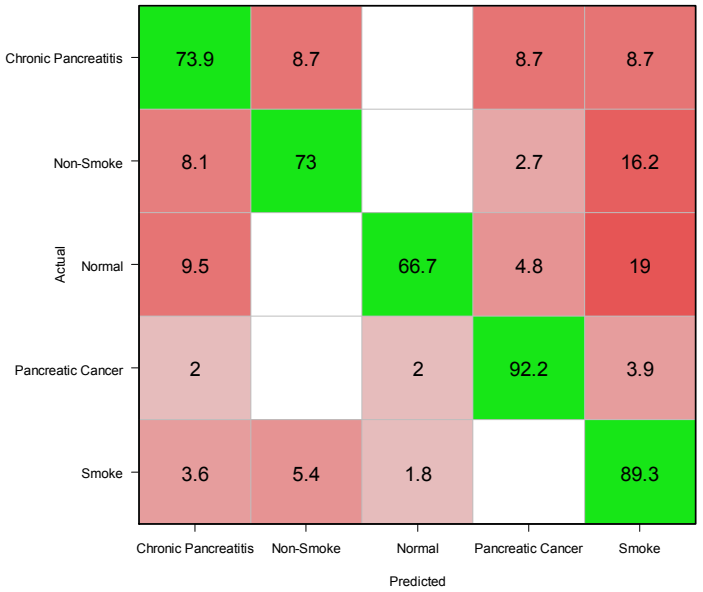
Materials and Methods
Proteins from human lung adenocarcinoma cell line A549 lysates were subjected to extensive fractionation. The resulting 1,824 fractions were spotted in duplicate on nitrocellulose-coated slides. The microarrays produced were used in a blinded validation study to determine whether annexin I, PGP9.5, and 14-3-3 theta antigens previously found to be targets of autoantibodies in newly diagnosed patients with lung cancer are associated with autoantibodies in sera collected at the presymptomatic stage and to determine whether additional antigens may be identified in prediagnostic sera. Individual sera collected from 85 patients within 1 year before a diagnosis of lung cancer and 85 matched controls from the Carotene and Retinol Efficacy Trial (CARET) cohort were hybridized to individual microarrays.

Results
We present evidence for the occurrence in lung cancer sera of autoantibodies to annexin I, 14-3-3 theta, and a novel lung cancer antigen, LAMR1, which precede onset of symptoms and diagnosis.

Conclusion
Our findings suggest potential utility of an approach to diagnosis of lung cancer before onset of symptoms that includes screening for autoantibodies to defined antigens.

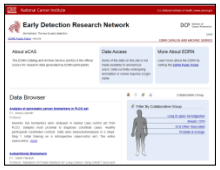
J Clin Oncol 26:5060-5066. © 2008 by American Society of Clinical Oncology





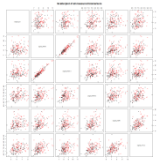
eCAS Statistical Analysis Pipeline

eCas raw data

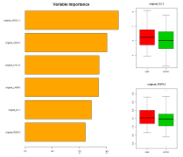


Data Preprocessing: Import into R → Cleaning and Preparation

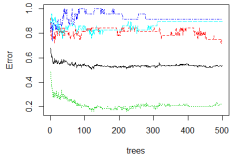
Data Inspection: Descriptive Statistics → Univariate Analysis → ...



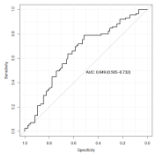
Statistical Modeling and Machine Learning: Random Forests → Regression → Bayesian Approaches → ...



Performance Analysis: Precision/ Recall → ROC + AUC → ...

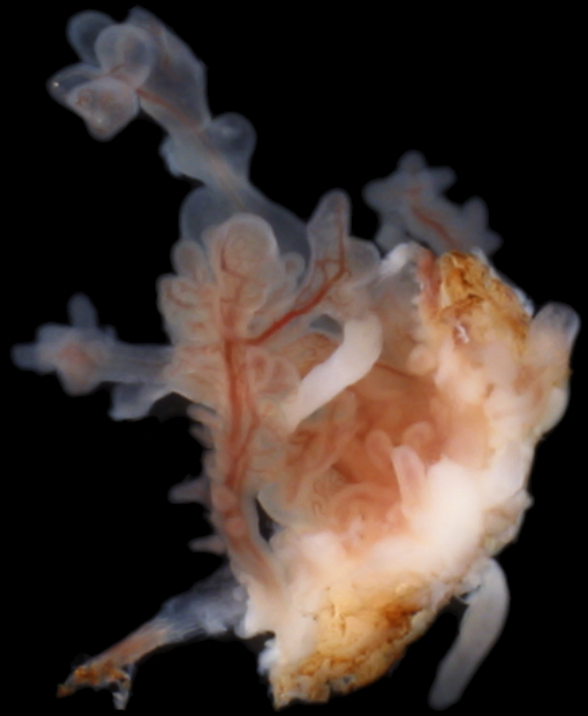


Reporting: Plotting → PDF Generation → ...

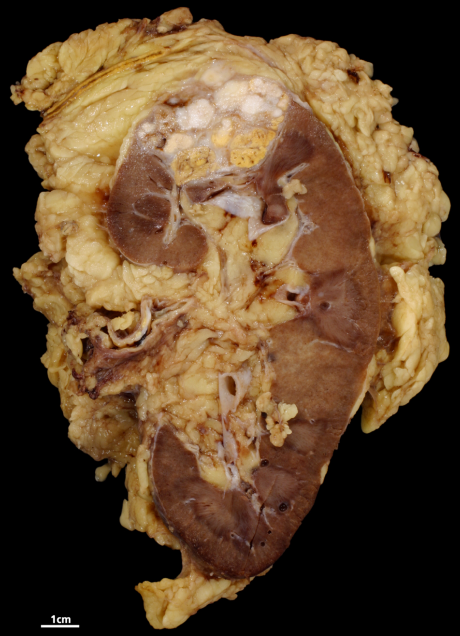


Going Beyond Reproducibility

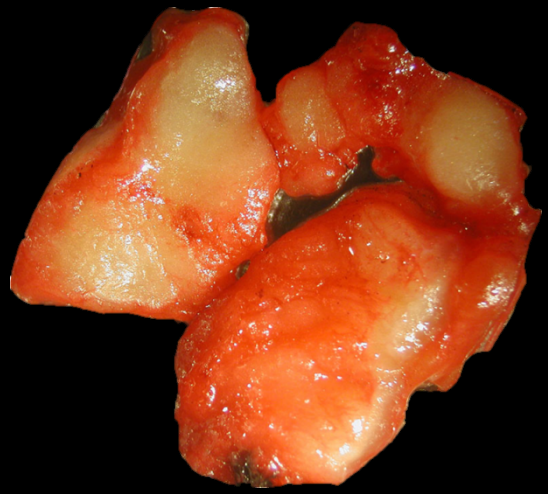
Can we improve on published models based on EDRN data?



Bladder



Kidney



Lymphnodes

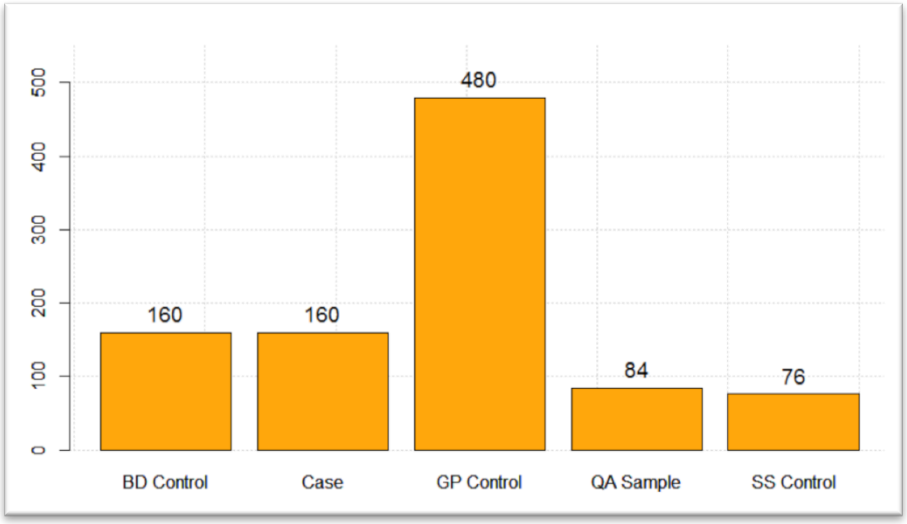
Reproducibility in EDRN – Example: Early Detection of Ovarian Cancer

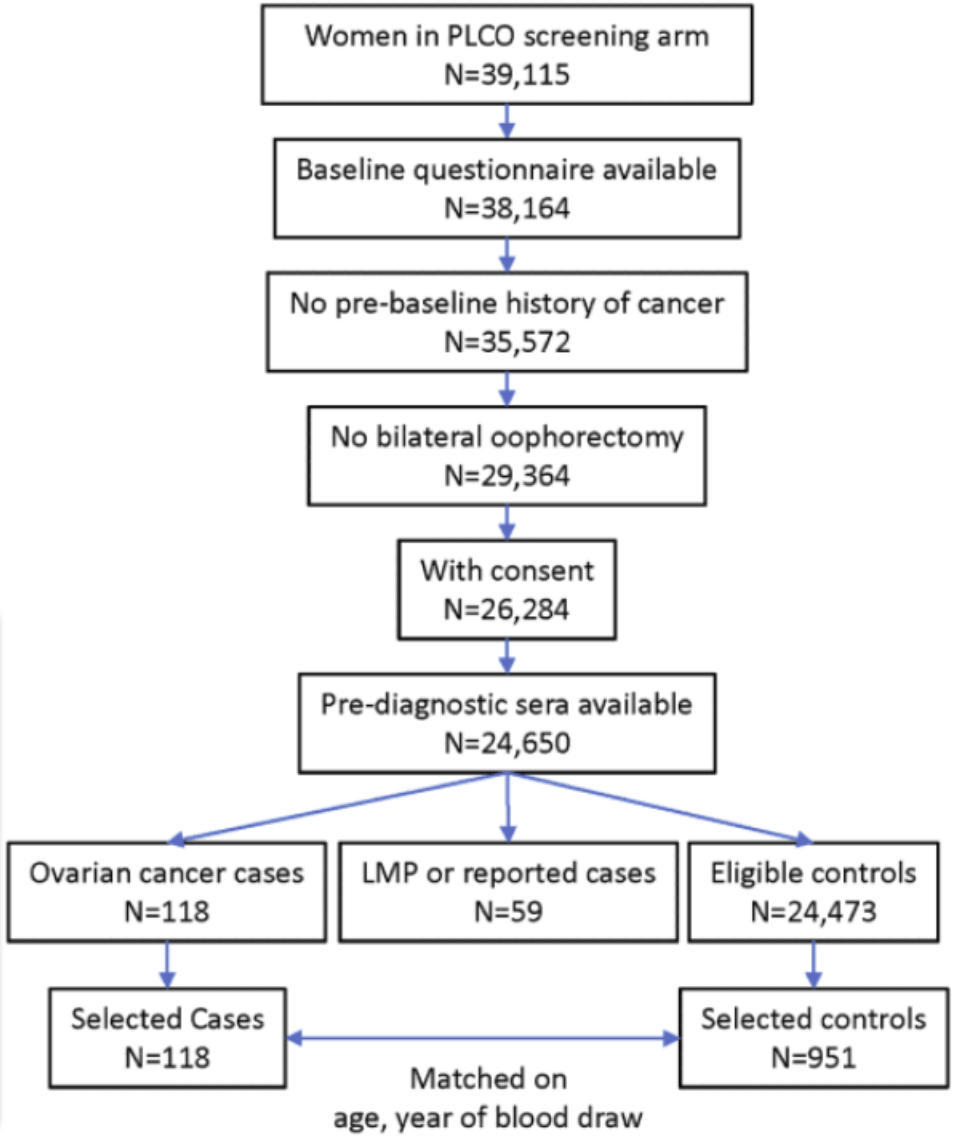
Published in final edited form as:
Cancer Prev Res (Phila). 2011 March ; 4(3): 375–383. doi:10.1158/1940-6207.CAPR-10-0193.

A Framework for Evaluating Biomarkers for Early Detection: Validation of Biomarker Panels for Ovarian Cancer

Claire S. Zhu[1], Paul F. Pinsky[1], Daniel W. Cramer[2], David F. Ransohoff[3], Patricia Hartge[4], Ruth M. Pfeiffer[4], Nicole Urban[5], Gil Mor[6], Robert C. Bast Jr.[7], Lee E. Moore[4], Anna E. Lokshin[8], Martin W. McIntosh[5], Steven J. Skates[9], Allison Vitonis[2], Zhen Zhang[10], David C. Ward[11], James T. Symanowski[12], Aleksey Lomakin[13], Eric T. Fung[14], Patrick M. Sluss[9], Nathalie Scholler[15], Karen H. Lu[7], Adele M. Marrangoni[8], Christos Patriotis[1], Sudhir Srivastava[1], Saundra S. Buys[16], and Christine D. Berg[1] for the PLCO Project Team

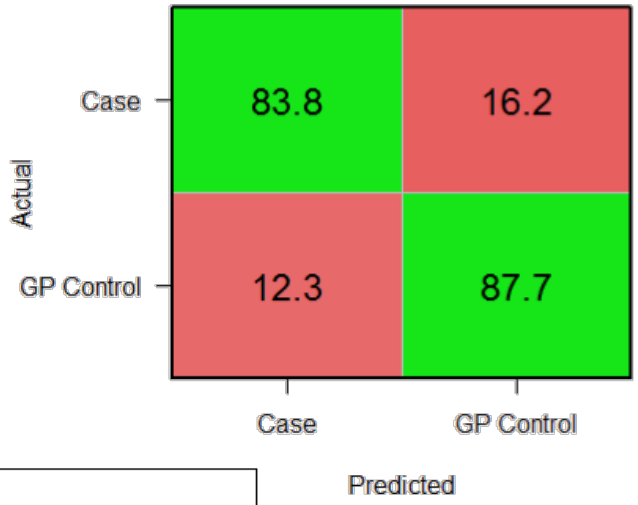
Phase III Validation of a Consensus Panel of Ovarian Cancer Biomarkers

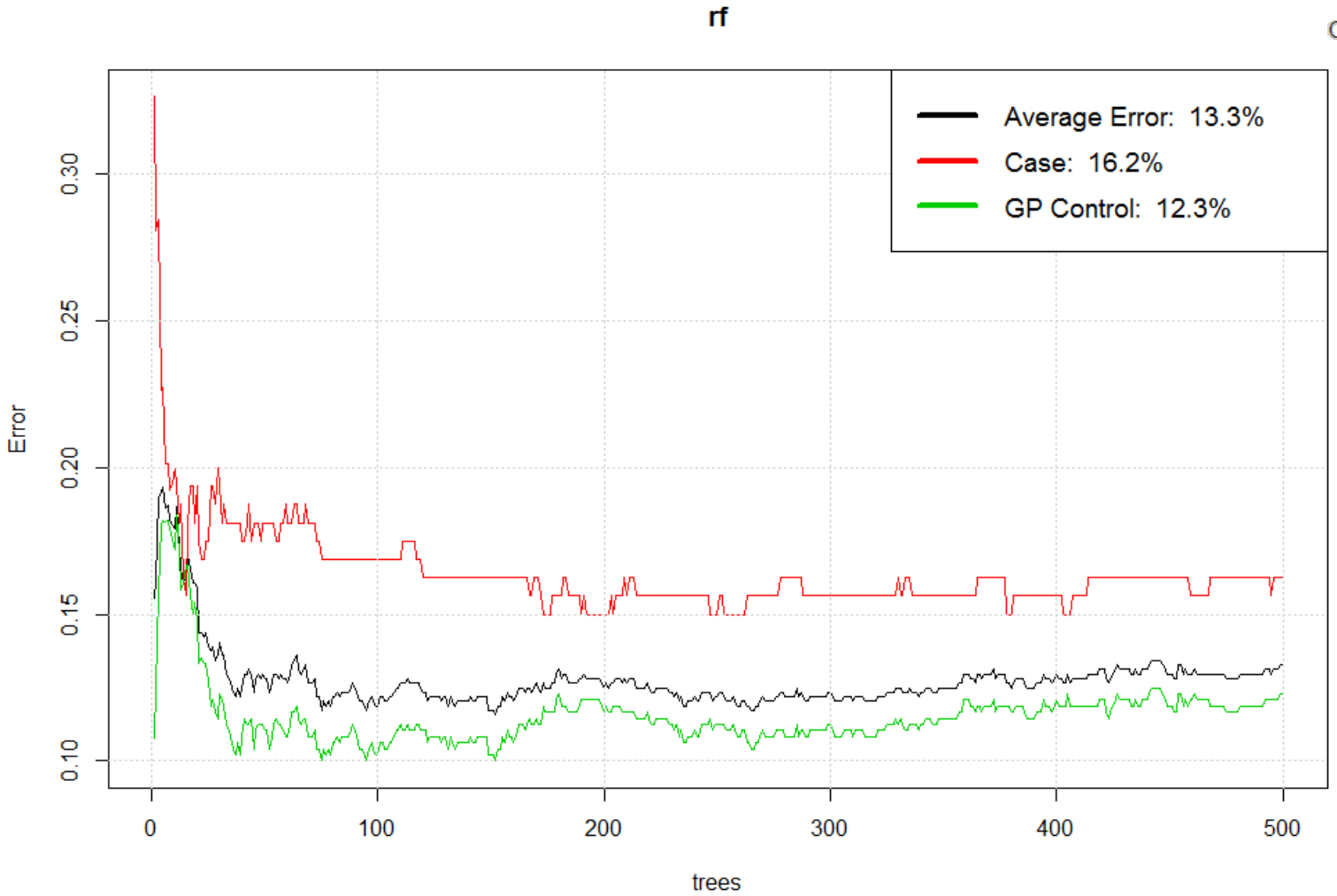




Classification Experiment

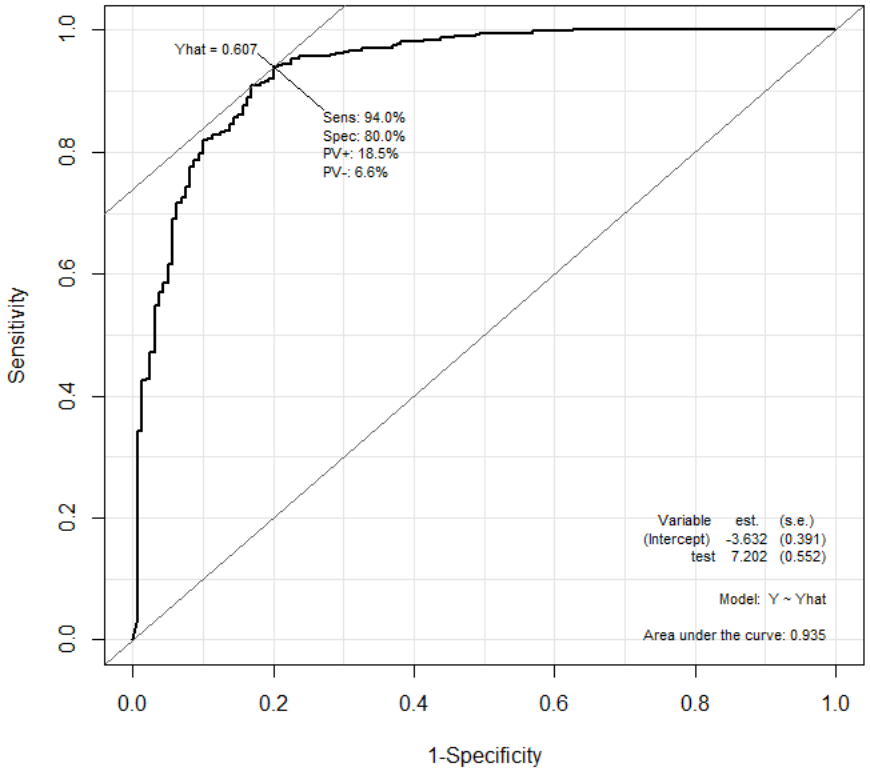
Model:
Random forest with 500 trees.
Distribution proportional class weights.
GINI index.



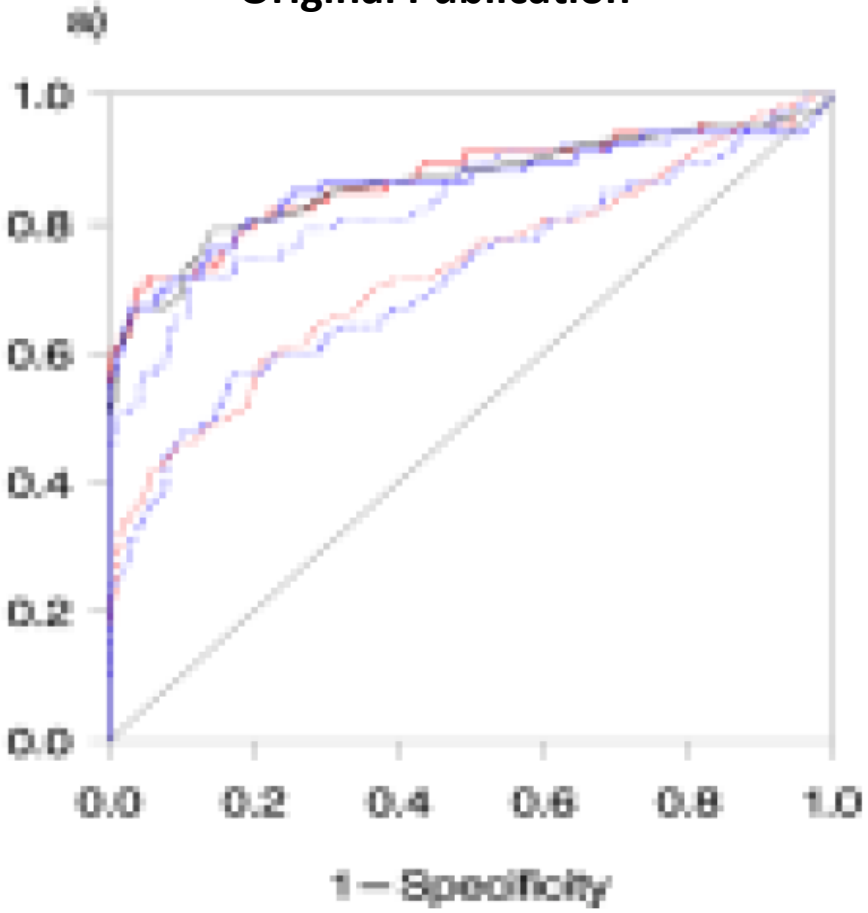


ROC curve

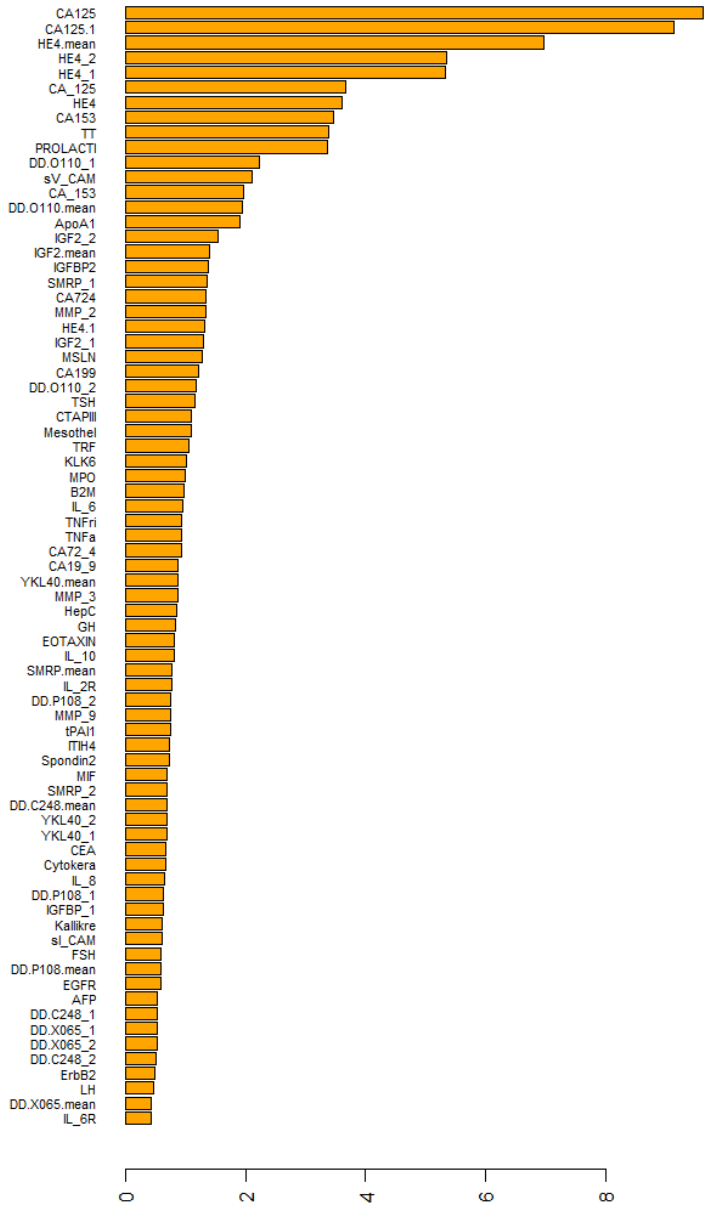
Our model: AUC = 0.935



Original Publication



Biomarker Importance Analysis

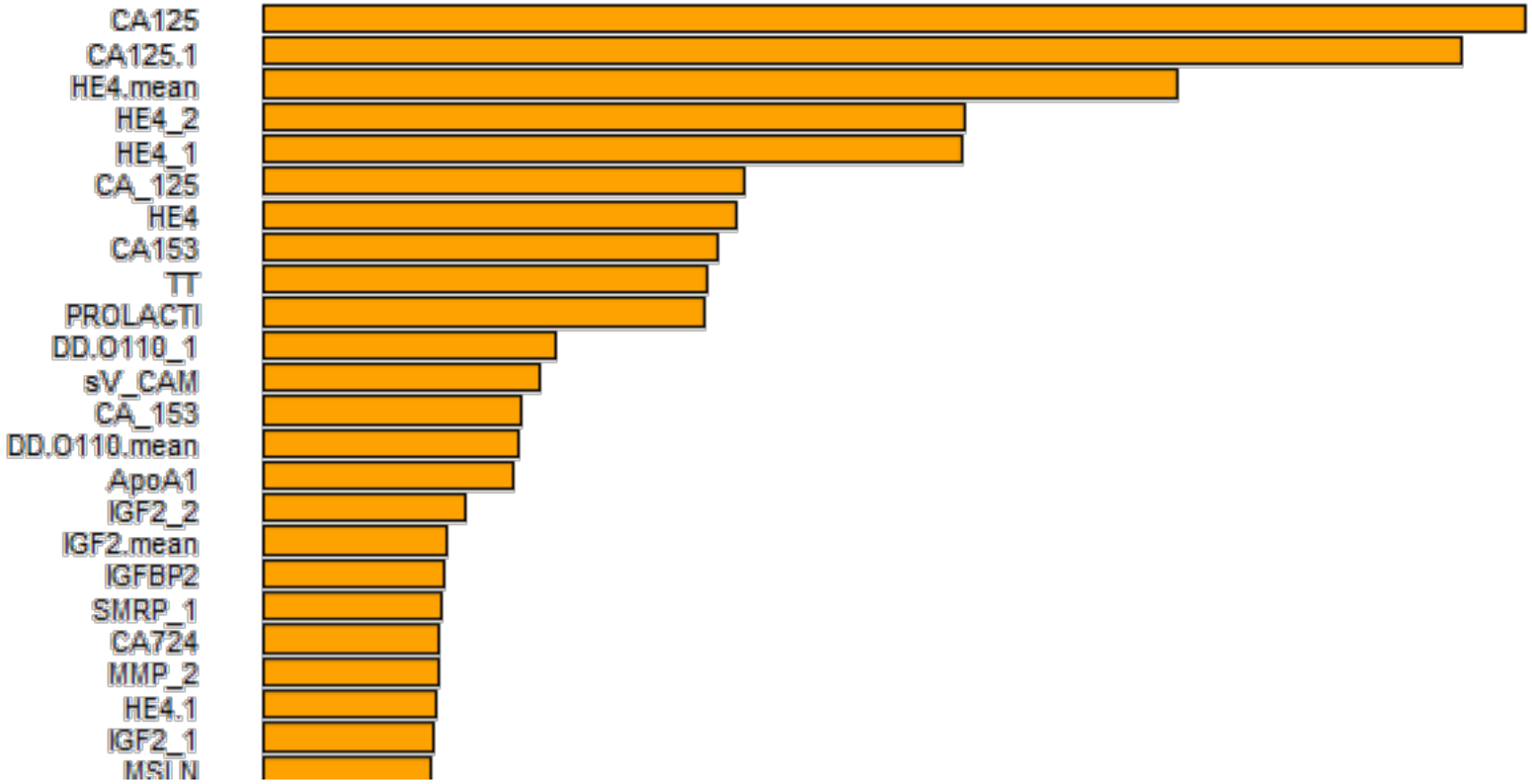


Model Comparison

FHCRC Model: CA125, HE4, HK11, B7-H4, DcR3, SMR, Spondin-2

Boston-NW Model: CA125, HE4, CA72.4, CA15.3, SMR

JPL Model:

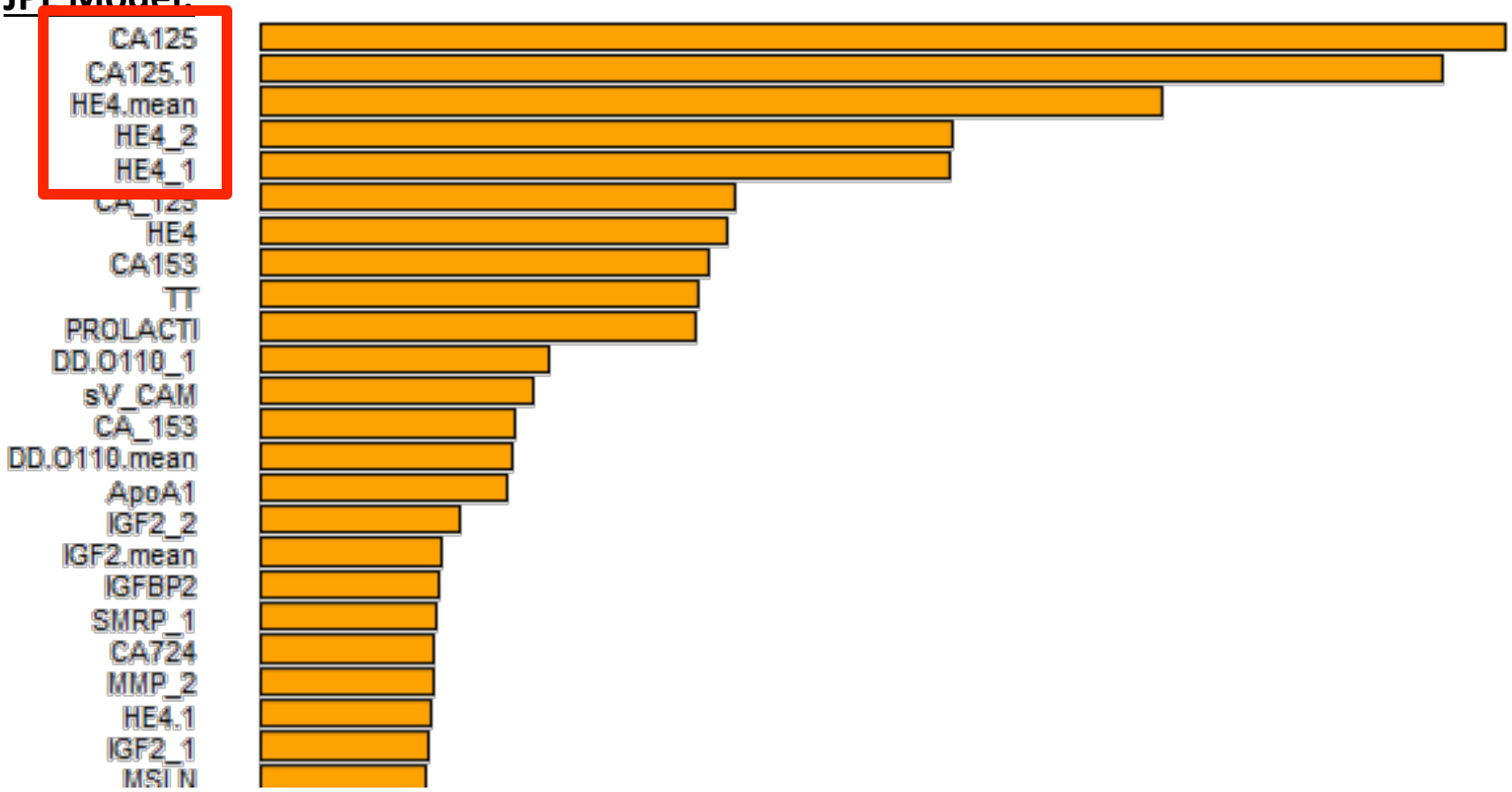


Model Comparison

FHCRC Model: CA125, HE4, HK11, B7-H4, DcR3, SMR, Spondin-2

Boston-NW Model: CA125, HE4, CA72.4, CA15.3, SMR

JPL Model:



Conclusion

- EDRN provides a solid **platform for reproducible research**
- Pipeline development at JPL will allow for a **plug-and-play interface** for running past but also novel statistical models
- **Non-Linearity makes a difference!**
  It is time to go beyond simple linear combinations for complex disease behavior

EDRN Informatics Center

Dan Crichton - PI

Maureen Colbert
(Dartmouth)

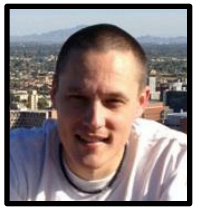
Mike Joyce

Heather Kincaid

Rishi Verma

Kristen Anton
(Dartmouth)

Andrew Hart



Sean Kelly

Chris Mattmann

Thank you for your attention!

Questions welcomed!

Thomas J. Fuchs
thomas.fuchs@jpl.nasa.gov