Extracting Meaningful Biomarkers from Gene Expression Data

Michael Bittner & Edward R. Dougherty

Translational Genomics Research Institute

Topics

- What is “meaningful”?
- Are our mathematical approaches appropriate given biological behaviors?
- What are our models?
- What differences arise in expectations given the models?
- Which model does this project use?

Meaning: A validated scientific theory

- *Mathematical model* consisting of variables and relations between the variables.
- *Operational definitions* relating the variables to observable (and measurable) phenomena.
- *Experimental design* to test predictions made by the model.
- A scientific theory is validated to the extent that predictions derived from it agree with experimental observations.

Translational Genomics Research Institute

Meaning: Non-Considerations

- **Richard Feynman**: "It is whether or not the theory gives predictions that agree with experiment. It is not a question of whether a theory is philosophically delightful, or easy to understand, or perfectly reasonable from the point of view of common sense. The theory of quantum electrodynamics describes Nature as absurd from the point of view of common sense. And it agrees fully with experiment. So I hope you can accept Nature as She is – absurd."

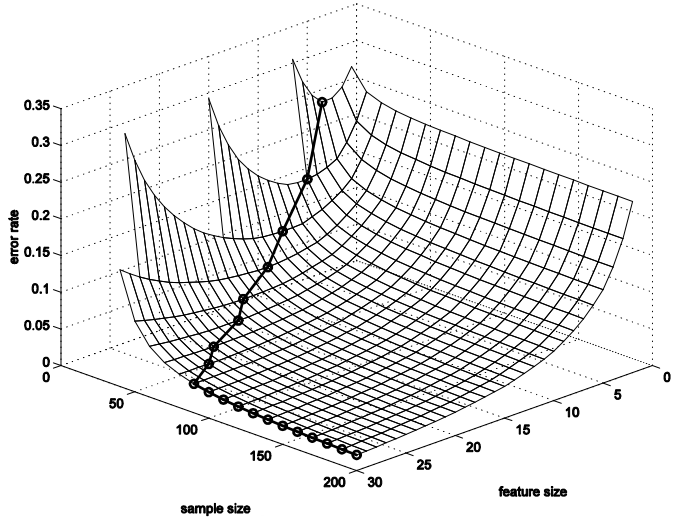
Mathematical Approaches: Statistics

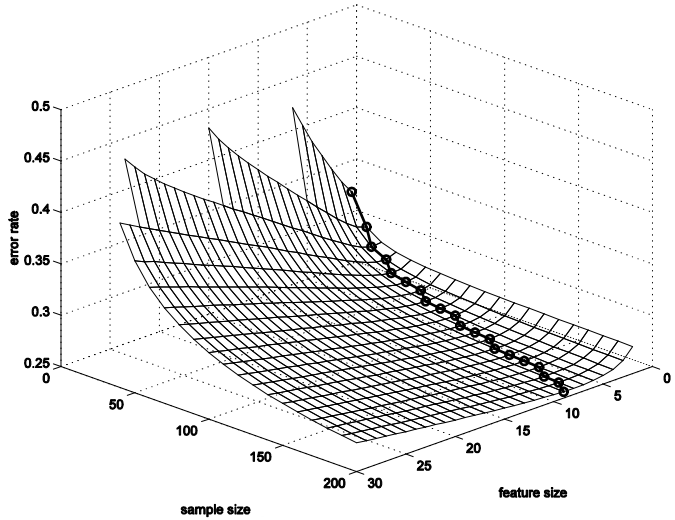
- Statistical requirements –
  - Randomness
  - Independence
- Biological Requirements
  - Orderly
  - Interdependent

Translational Genomics Research Institute

Peaking Phenomenon

- Optimal number of features depends on sample size, classification rule and feature-label distribution.
- Top: LDA, linear model, slightly correlated features.
- Bottom: LDA, linear model, highly correlated features.





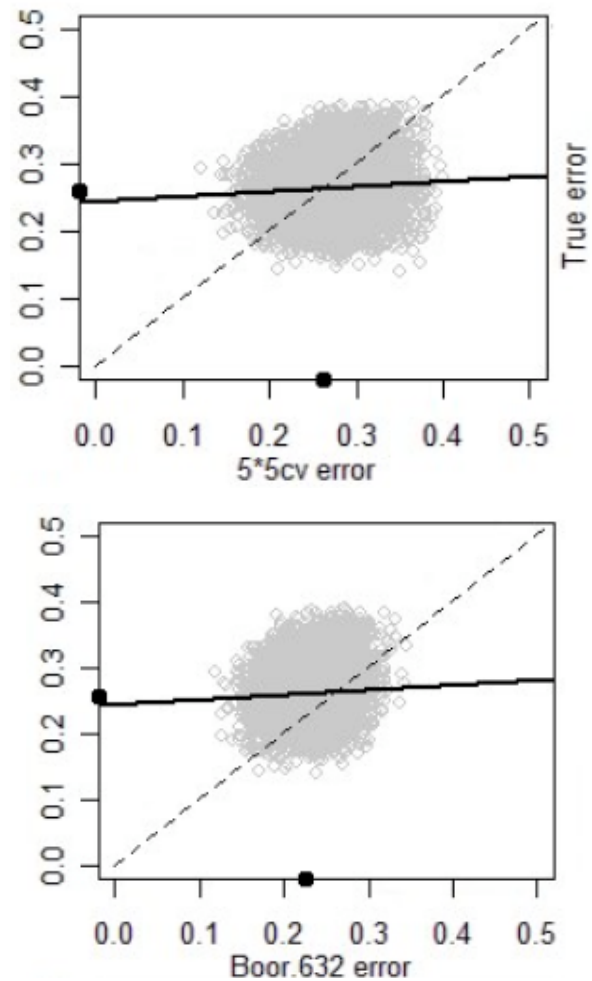
- Hua, J., Xiong, Z., Lowey, J., Suh, E., and E. R. Dougherty, "**Optimal Number of Features as a Function of Sample Size for Various Classification Rules**", *Bioinformatics*, **21**(8), 1509-1515, 2005.

Mathematical Approaches: Classification

- Error estimation requirements –
  - Feature Selection
  - Validation
- Typical biological study
  - Few samples per variable
  - Poor distributional estimate

Decorrelation of True and Estimated Errors

- In high dimensions, the key issue is decorrelation of the estimated and true errors, not estimator variance.
  - With or without feature selection.
- $Var_{dev}[\varepsilon_{est}] = \sigma^2_{est} + \sigma^2_{tru} - 2\rho\sigma_{est}\,\sigma_{tru}$
- Top: Cross-validation
  - 5 features out of 200, sample size 50
- Bottom: .632 bootstrap



- Hanczar, B., Hua, J., and E. R. Dougherty, "**Decorrelation of the True and Estimated Classifier Errors in High-dimensional Settings**" *EURASIP J. Bioinformatics and Systems Biology*, Article ID 38473, 12 pages, 2007.

Mathematical Approaches: Consequences

Nonvalidation of Reported Genetic Risk Factors for Acute Coronary Syndrome in a Large-Scale Replication Study

Thomas M. Morgan, MD
Harlan M. Krumholz, MD, MS
Richard P. Lifton, MD, PhD
John A. Spertus, MD, MPH

Conclusions Our null results provide no support for the hypothesis that any of the 85 genetic variants tested is a susceptibility factor for ACS. These results emphasize the need for robust replication of putative genetic risk factors before their introduction into clinical care.

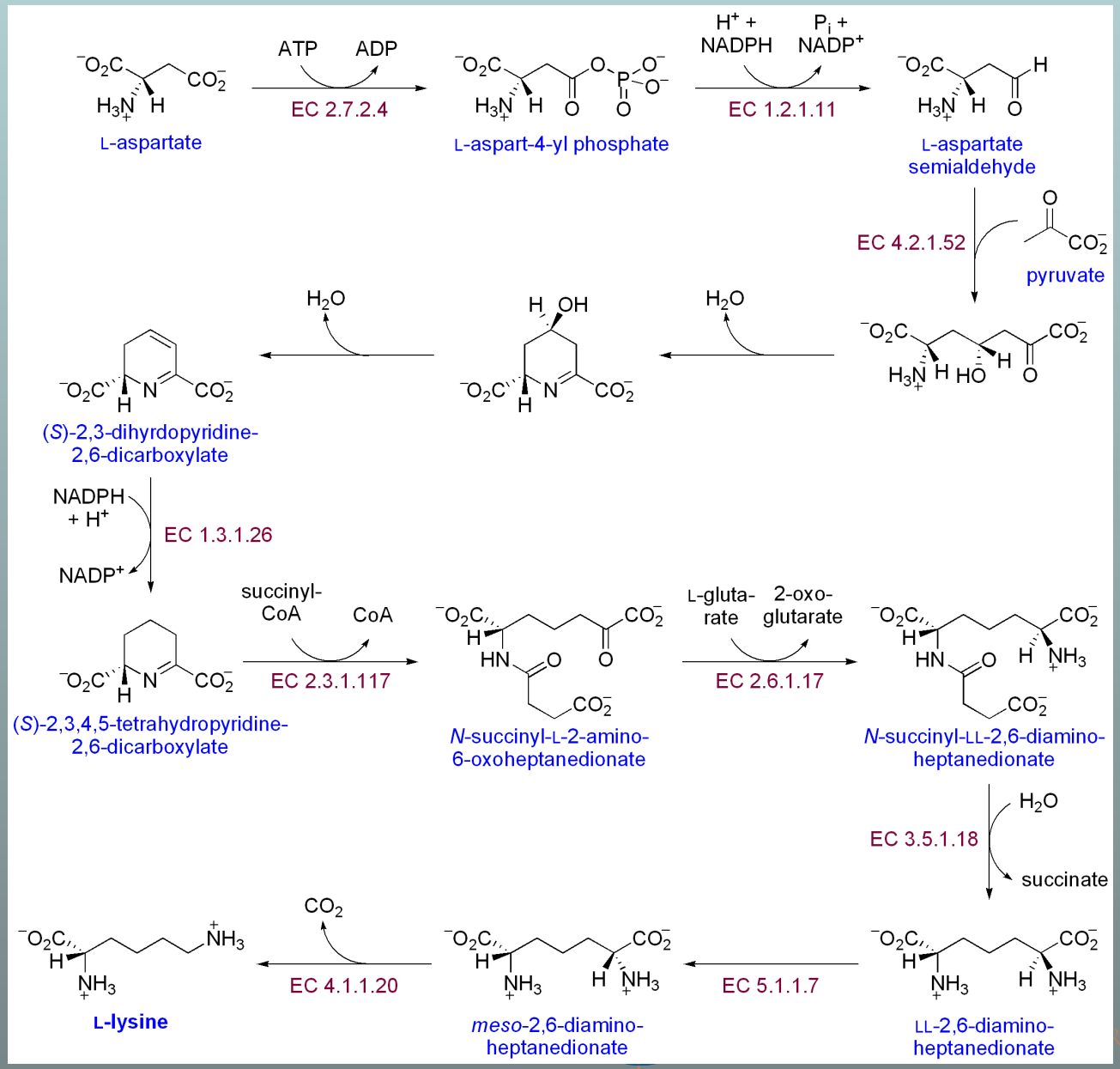
JAMA. 2007;297:1551-1561 www.jama.com

Models

- Newtonian
  - Single large influence
  - Background dependent
    - interaction measured has no consequence on overall system
- Relativistic
  - Multiple influences
  - Background independent
    - Interactions measured have direct influences on overall system

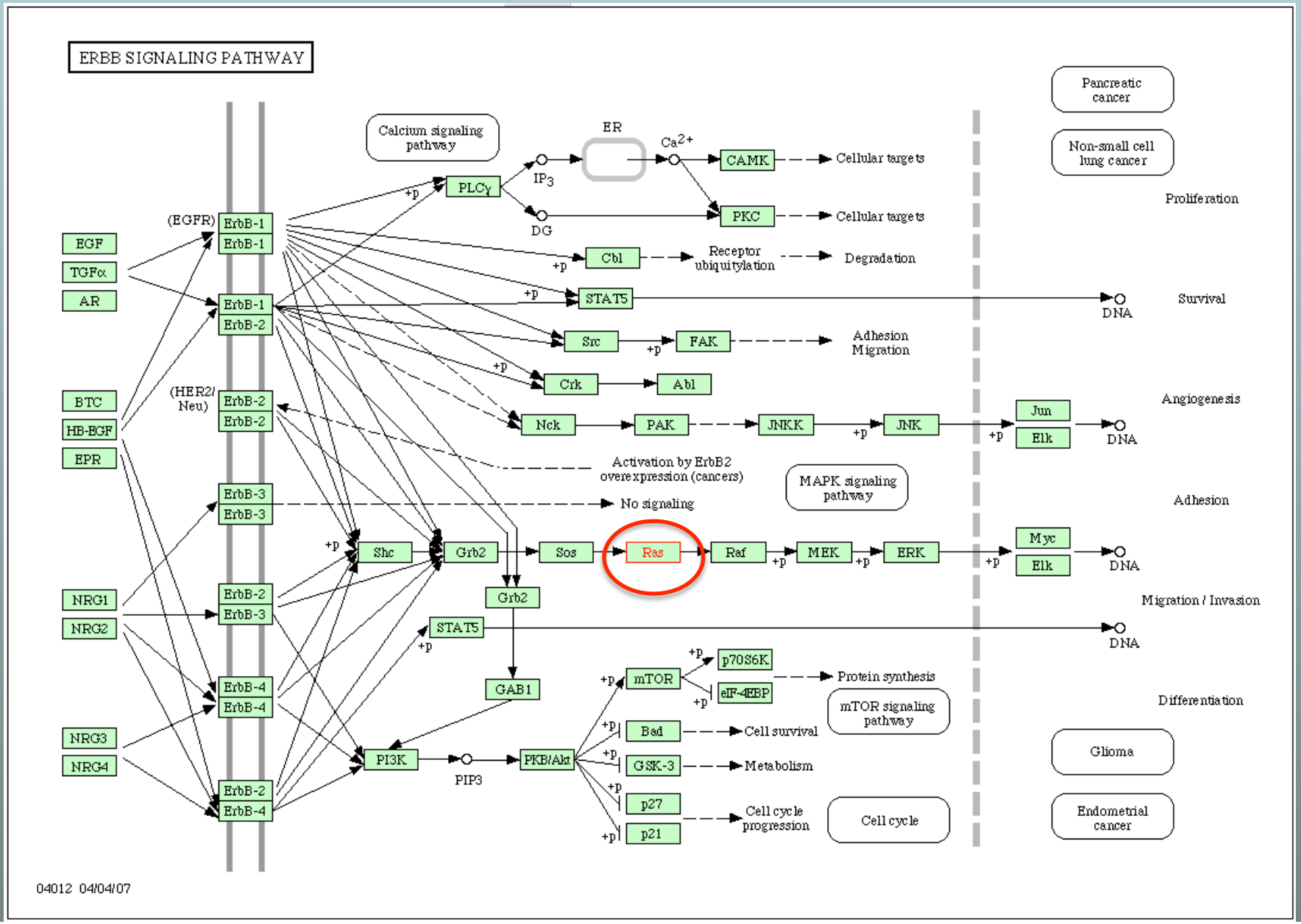
Translational Genomics Research Institute

Models: If And (Newtonian)



RESEARCH INSTITUTE

Models: If Or (Relativistic)



Contributory Inputs

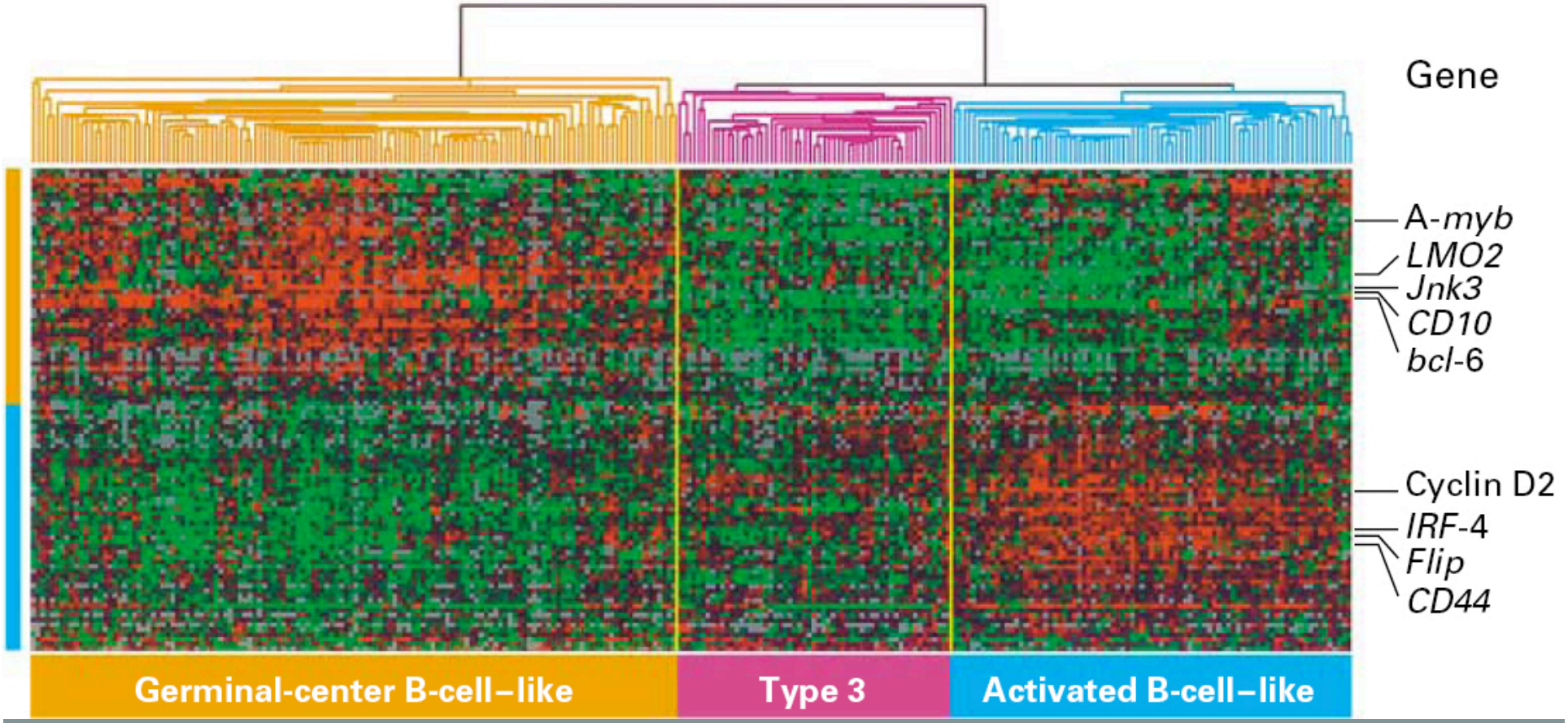
Association between Microdeletion and Microduplication at 16p11.2 and Autism

Lauren A. Weiss, Ph.D., Yiping Shen, Ph.D., Joshua M. Korn, B.S., Dan E. Arking, Ph.D., David T. Miller, M.D., Ph.D., Ragnheidur Fossdal, B.Sc., Evald Saemundsen, B.A., Hreinn Stefansson, Ph.D., Manuel A.R. Ferreira, Ph.D., Todd Green, B.S., Orah S. Platt, M.D., Douglas M. Ruderfer, M.S., Christopher A. Walsh, M.D., Ph.D., David Altshuler, M.D., Ph.D., Aravinda Chakravarti, Ph.D., Rudolph E. Tanzi, Ph.D., Kari Stefansson, M.D., Ph.D., Susan L. Santangelo, Sc.D., James F. Gusella, Ph.D., Pamela Sklar, M.D., Ph.D., Bai-Lin Wu, M.Med., Ph.D., and Mark J. Daly, Ph.D., for the Autism Consortium.

N Engl J Med 2008;358.

Translational Genomics Research Institute

Analysis - If Or Agglomeration

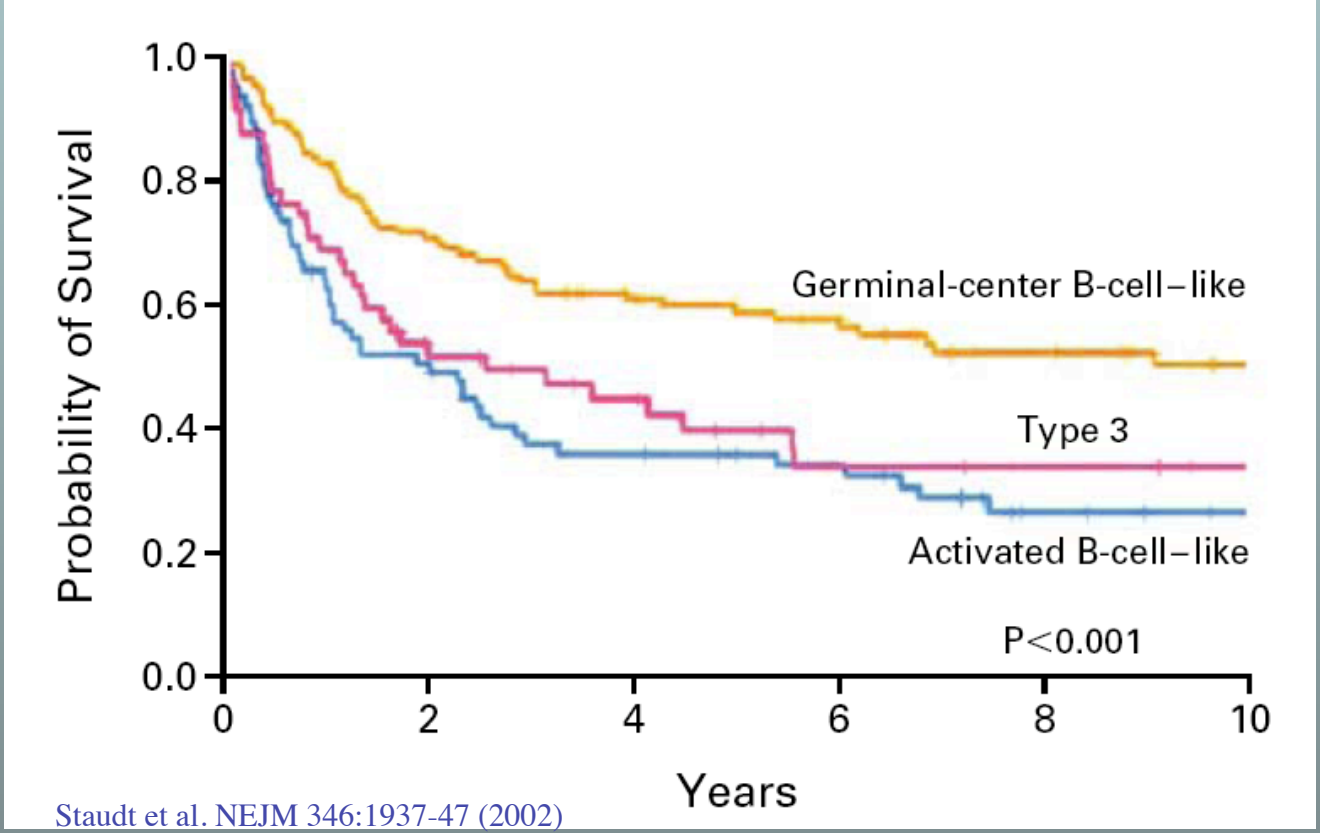


Staudt et al. NEJM 346:1937-47 (2002)

Translational Genomics Research Institute

Analysis - If Or Agglomeration

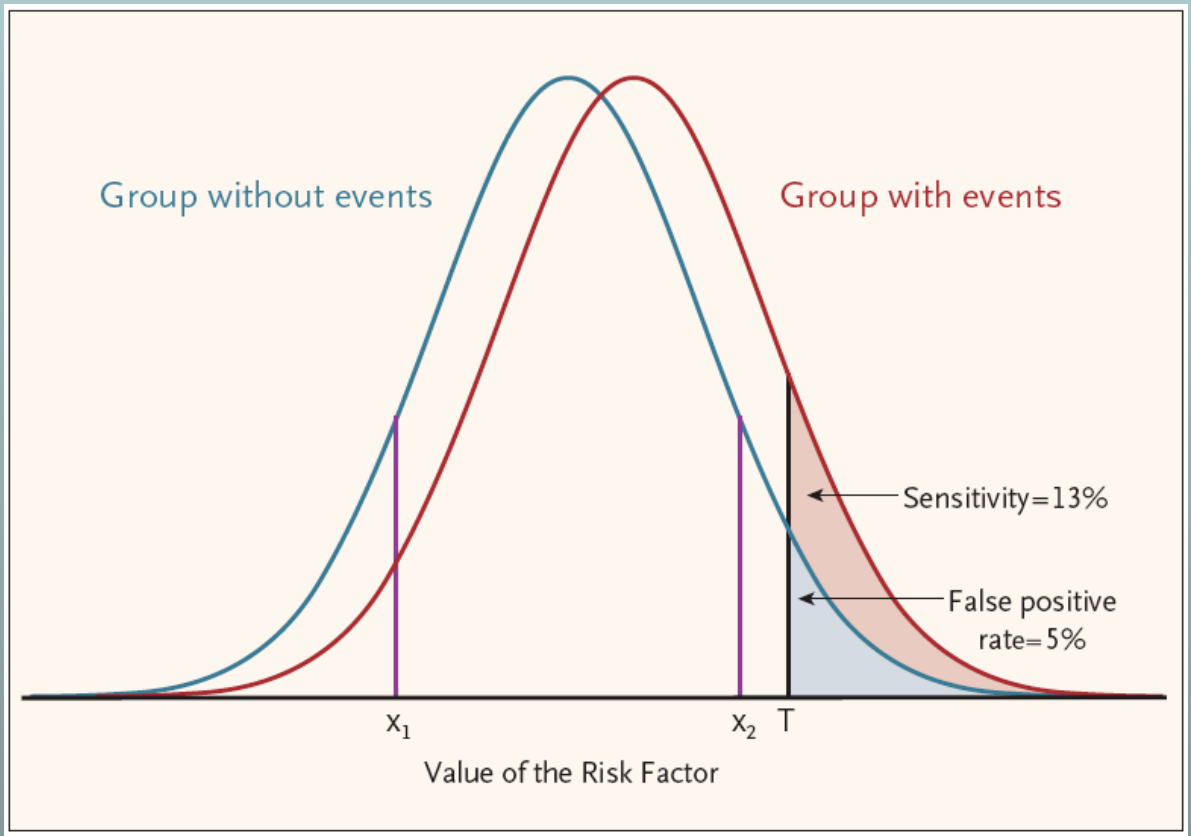
$$\log h(t) = f(h_0(t), \alpha + \beta_1 X_1 + \cdots + \beta_k X_k)$$



Staudt et al. NEJM 346:1937-47 (2002)

Translational Genomics Research Institute

Agglomeration vs Classification



STATISTICS AND MEDICINE

The Limitations of Risk Factors as Prognostic Tools

James H. Ware, Ph.D.

NEJM 355;25 (2006)

How does this play out?

OPEN ACCESS Freely available online

PLoS GENETICS

Huntingtin Interacting Proteins Are Genetic Modifiers of Neurodegeneration

Linda S. Kaltenbach[1], Eliana Romero[2], Robert R. Becklin[1], Rakesh Chettier[1], Russell Bell[1], Amit Phansalkar[1], Andrew Strand[3], Cameron Torcassi[4], Justin Savage[1], Anthony Hurlburt[1], Guang-Ho Cha[2], Lubna Ukani[2], Cindy Lou Chepanoske[1], Yuejun Zhen[1], Sudhir Sahasrabudhe[1], James Olson[3], Cornelia Kurschner[1], Lisa M. Ellerby[4], John M. Peltier[1], Juan Botas[2*], Robert E. Hughes[1,4*]

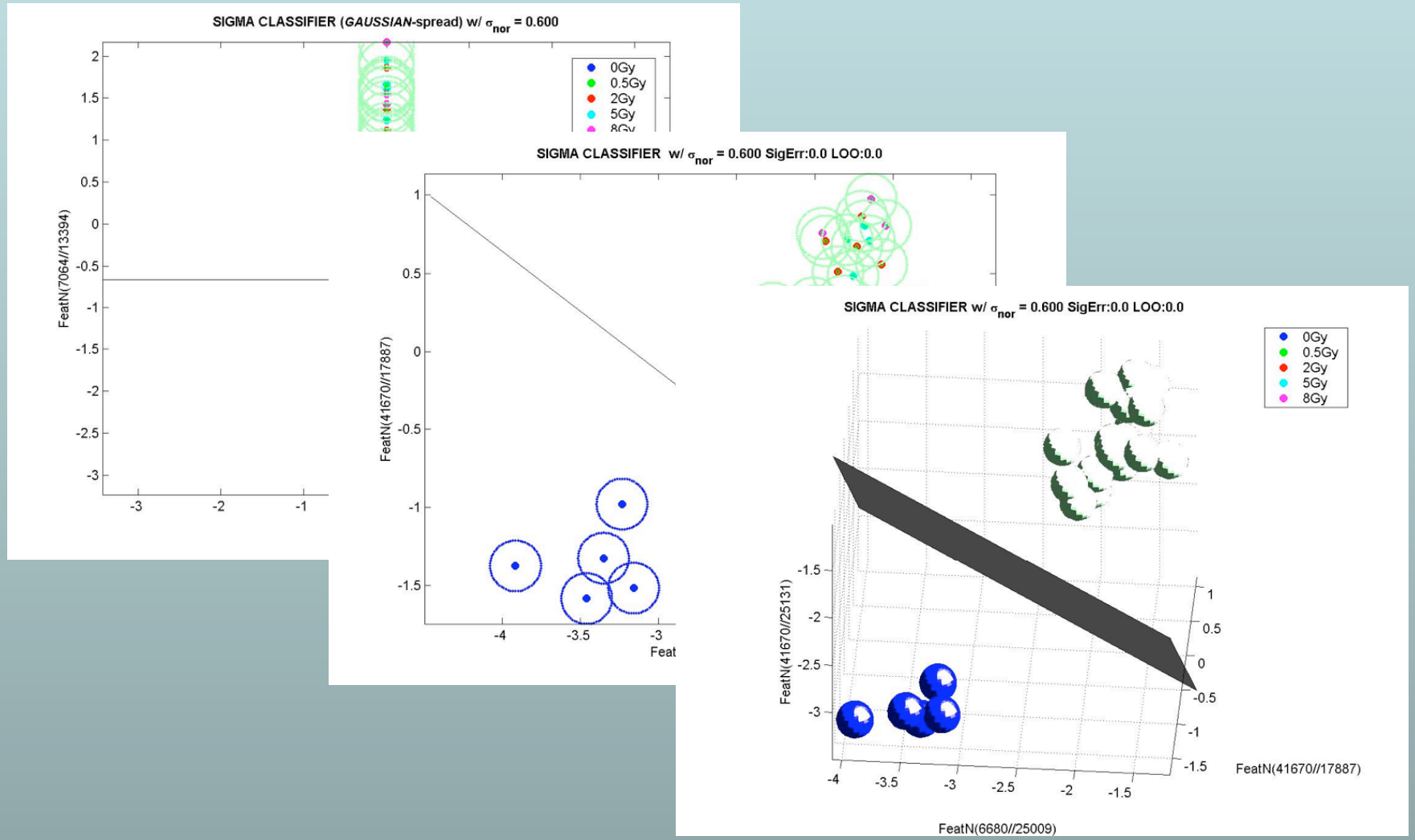
Efficient and Rapid Induction of a Chronic Myelogenous Leukemia-Like Myeloproliferative Disease in Mice Receiving P210 bcr/abl-Transduced Bone Marrow

By Warren S. Pear, Juli P. Miller, Lanwei Xu, John C. Pui, Benny Soffer, Robert C. Quackenbush, Ann Marie Pendergast, Roderick Bronson, Jon C. Aster, Martin L. Scott, and David Baltimore

Blood 92:3780-92 (1998)

Translational Genomics Research Institute

Dosimeter Model



TRANSLATIONAL GENOMICS RESEARCH INSTITUTE

Collaborators

Texas A&M
Edward Dougherty
Anirrudha Datta
Ulisses Braga Neto
Blaise Hanczar

U. Sao Paulo
Junior Barrera
Ronaldo Hashimoto

Tampere U.
Jakko Astola
Ioan Tabus

TGen
Seungchan Kim
Yoga Balagurunathan
Marcel Brun
Jianping Hua
Sima Chao
Edward Suh
James Lowey

NCI
Yidong Chen

Translational Genomics Research Institute