

Why you should care about statistics

Jeff Leek

@jtleek

Intersection of three disciplines

biology

statistics

computer science

u

A Venn diagram with three overlapping circles. The top-left circle is green and labeled 'Biology'. The top-right circle is blue and labeled 'Statistics'. The bottom circle is purple and labeled 'Computer Science'. The central area where all three circles overlap is labeled 'Genomic Data Science'. A handwritten orange note 'why is vital.?' is positioned above the 'Statistics' circle.

Biology

Statistics

why is vital.?

Genomic Data
Science

Computer Science

Seems like an exciting result!

Genomic signatures to guide the use of chemotherapeutics

personalize therapies.






Anil Potti^{1,2}, Holly K Dressman^{1,3}, Andrea Bild^{1,3}, Richard F Riedel^{1,2}, Gina Chan⁴, Robyn Sayer⁴, Daniel Cragun⁴, Hope Cottrill⁴, Michael J Kelley², Rebecca Petersen⁵, David Harpole⁵, Jeffrey Marks⁵, Andrew Berchuck^{1,6}, Geoffrey S Ginsburg^{1,2}, Phillip Febbo^{1,2,3}, Johnathan Lancaster⁴ & Joseph R Nevins^{1,2,3}

Using *in vitro* drug sensitivity data coupled with Affymetrix microarray data, we developed gene expression signatures that predict sensitivity to individual chemotherapeutic drugs. Each signature was validated with response data from an independent set of cell line studies. We further show that many of these signatures can accurately predict clinical response in individuals treated with these drugs. Notably, signatures developed to predict response to individual agents, when combined, could also predict response to multidrug regimens. Finally, we integrated the chemotherapy response signatures with signatures of oncogenic pathway deregulation to identify new therapeutic strategies that make use of all available drugs. The development of gene expression profiles that can predict response to

ARTICLE LINKS

- [Supplementary info](#)

ARTICLE TOOLS

-  [Send to a friend](#)
-  [Export citation](#)
-  [Export references](#)
-  [Rights and permissions](#)
-  [Order commercial reprints](#)

SEARCH PUBMED FOR

- [Anil Potti](#)
- [Holly K Dressman](#)
- [Andrea Bild](#)
- [Richard F Riedel](#)

image credits: Chan
<http://en.wikipedia.org/wiki/Protein>
http://en.wikipedia.org/wiki/Genetic_code

Major problems in the analysis

DERIVING CHEMOSENSITIVITY FROM CELL LINES: FORENSIC BIOINFORMATICS AND REPRODUCIBLE RESEARCH IN HIGH-THROUGHPUT BIOLOGY

BY KEITH A. BAGGERLY* AND KEVIN R. COOMBES†

U.T. M.D. Anderson Cancer Center

High-throughput biological assays such as microarrays let us ask very detailed questions about how diseases operate, and promise to let us personalize therapy. Data processing, however, is often not described well enough to allow for exact reproduction of the results, leading to exercises in “forensic bioinformatics” where aspects of raw data and reported results are used to infer what methods must have been employed. Unfortunately, poor documentation can shift from an inconvenience to an active danger when it obscures not just methods but errors. In this report, we examine several related papers purporting to use microarray-based signatures of drug sensitivity derived from cell lines to predict patient response. Patients in clinical trials are currently being allocated to treatment arms on the basis of these results. However, we show in five case studies that the results incorporate several simple errors that may be putting patients at risk. One theme that emerges is that the most common errors are simple (e.g., row or column offsets); conversely, it is our experience that the most simple errors are common. We then discuss steps we are taking to avoid such errors in our own investigations.

reproduce the
analysis
variable to }
reproduce }

An ongoing saga

NORTH CAROLINA

DURHAM COUNTY

DURHAM COUNTY

FILED

SEP 7 2011

IN THE GENERAL COURT OF
JUSTICE

SUPERIOR COURT DIVISION

1 CVS 4121

Richard Aiken, Jean K. Carroll, Executrix of the Estate of Harold G. Carroll, Jean K. Carroll, Individually, Peggy Cox, as Administratrix of the Estate of Paul F. Cox, Peggy Cox, Individually, Helene L. Fligel, Jason Gannon, as Personal Representative of the Estate of Jennifer L. Gannon, John Haddock, as Executor of the Estate of Karen Heath, Walter Jacobs, as Executor of the Estate of Juliet J. Jacobs, Walter Jacobs, Individually, Polly Johnson, as Executor of the Estate of Malcom W. Johnson, and Polly Johnson, Individually,
Plaintiffs

vs.

COMPLAINT
(JURY TRIAL DEMANDED)

*How the analysis was done?
Stat was done incorrect*

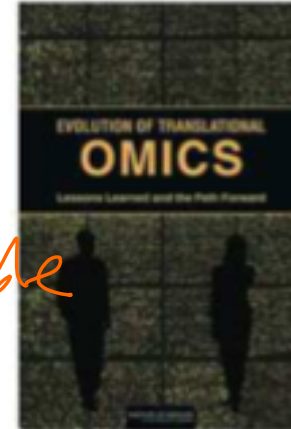
This situation spurred an IOM report

For more information visit www.iom.edu/translationalomics

Evolution of Translational Omics

Lessons Learned and the Path Forward

stat issues! how to set reproducible



Sequencing the human genome opened a new era in biomedical science. Researchers have begun to untangle the complex roles of biology and genetics in specific diseases, and now better understand why particular therapies do or do not work in individual patients. New technologies have made it feasible to measure an enormous number of molecules within a tissue or cell; for example, genomics investigates thousands of DNA sequences, and proteomics examines large numbers of proteins. Collectively, these technologies are referred to as *omics*.

Caring about statistics

stat creates major problem.

Background

Many groups, including our own, have proposed the use of DNA methylation profiles as biomarkers for various disease states. While much research has been done identifying DNA methylation signatures in cancer vs. normal etc., we still lack sufficient knowledge of the role that differential methylation plays during normal cellular differentiation and tissue specification. We also need thorough, genome level studies to determine the meaning of methylation of individual CpG dinucleotides in terms of gene expression.

Results

relative knowledge where stat falls.

In this study, we have used (insert statistical method here) to compile unique DNA methylation signatures from normal human heart, lung, and kidney using the Illumina Infinium 27 K methylation arrays and compared those to gene expression by RNA sequencing. We have identified unique signatures of global DNA methylation for human heart, kidney and liver, and showed that DNA methylation data can be used to correctly classify various tissues. It indicates that DNA methylation reflects tissue specificity and may play an important role in tissue differentiation. The integrative analysis of methylation and RNA-Seq data showed that gene methylation and its transcriptional levels were comprehensively correlated. The location of methylation markers in terms of distance to transcription start site and CpG island showed no effects on the regulation of gene expression by DNA methylation in normal tissues.

1993

FIELD

NOTEWORTHY APPLICATIONS

Artificial Intelligence

machine learning, natural language processing, vision, mathematical models of cognition and learning

Chemistry

chemical and biomolecular engineering

Computational Science

computational fluid mechanics, computational materials sciences

Earth and Planetary Science

climate modeling, seismology, geographic information systems

Marketing

online advertising, consumer behavior

Physical Sciences

astronomy, particle physics, geophysics, space sciences

Signal Processing

compressed sensing, inverse imaging

Statistics

no app areas.

Biology

genomics, proteomics, ecoinformatics, computational cell biology

Economics

macroeconomic policy, taxation, labor economics, microeconomics, finance, real estate

Engineering

sensor networks (traffic control, energy-efficient buildings, brain-machine interface)

Environmental Sciences

deforestation, climate change, impacts of pollution

Humanities

digital humanities, archeology, land use, cultural geography, cultural heritage

Law

privacy, security, forensics, drug/human/CBRNe trafficking, criminal justice, incarceration, judicial decision making, corporate law

Linguistics

historical linguistics, corpus linguistics, psycholinguistics,

2013