

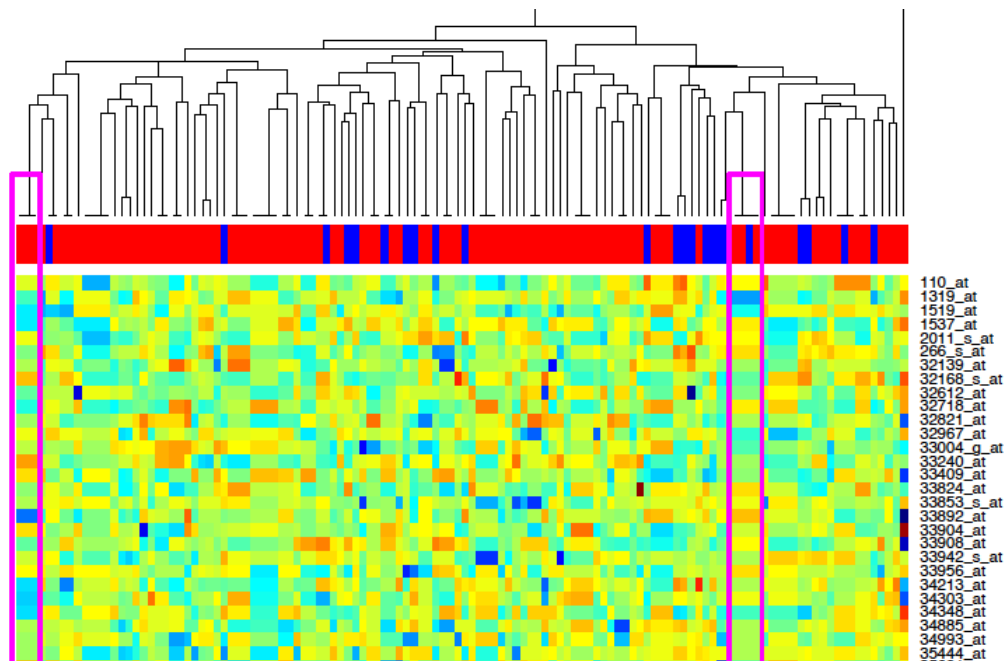
Reproducible research practice for genome-scale biology

**VJ Carey, Ph.D., Harvard University and
Bioconductor Foundation of N.A., Inc**

- discovering when things go wrong: forensic bioinformatics
- commercial pressures vs. transparency
- prospects for built-in forensic soundness

Reconstructing an analysis, finding a flaw

- columns are samples, rows are genes, colors indicate measured intensities of gene expression
- clumpy rows indicate accidental duplication of columns
- topmost red/blue column indicates sample labeling – some duplicated columns were labeled case in some instances, control in others



- a heartbreaking observation – and when you peek under the hood, more problems crop up

The catalog of error

Submitted to the Annals of Applied Statistics

arXiv: [math.PR/0000000](https://arxiv.org/abs/math.PR/0000000)

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 purport-

Some consequences



PO Box 9905 Washington DC 20016 Telephone 202-362-1809

NCI Raises New Questions About Duke Genomics Research, Cuts Assay From Trial

By Paul Goldberg

In a new setback to a controversial group of genomics researchers at Duke University, NCI officials eliminated a biomarker test from an ongoing phase III clinical trial.

The decision by the NCI Cancer Therapy Evaluation Program to remove the Lung Metagene Score assay from the trial conducted by the Cancer and Leukemia Group B challenges a Duke technology that has not previously attracted scrutiny.

The Duke group, headed by Joseph Nevins and Anil Potti, has made so many errors in their publications that the university suspended three clinical trials based on the group's technology. The trials were later restarted.

“We have asked [CALGB] to remove the Lung Metagene Score from the

Vol. 36 No. 18
May 14, 2010

© Copyright 2010 The Cancer Letter Inc.
All rights reserved. Price \$375 Per Year.
To subscribe, call 800-513-7042
or visit www.cancerletter.com.

Personalized Medicine:
The Cancer Letter
Obtains “Confidential”
Documents From Duke
IRB Investigation

... Page 3

Report Doesn't Quell
Concerns About Duke

Responding to charges of analytical and administrative error

- Duke researchers and administrators reinstated the suspended trials subsequent to an independent reevaluation of analyses, the report of which was submitted to NCI
- A redacted version of the report was obtained from NCI by Cancer Letter using FOIA
- The reviewers were kept anonymous and some of their findings, and the data/tools used to establish the reinstatement, were not released

Review of Genomic Predictors for Clinical Trials from Nevins, Potti, and Barry

December 22, 2009

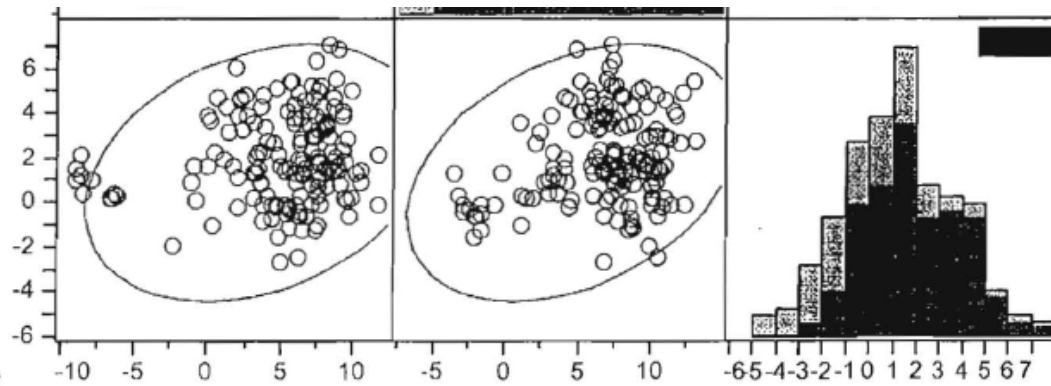
Summary

We were given two charges by Dr. Harrelson and the Duke IRB. The first was "Have the methodology errors originally communicated by the M.D. Anderson Cancer Center researchers, Baggerly and Coombes, been adequately addressed by the Duke researchers?" and the second "Do the methods as originally developed and as applied in the context of these trials remain valid?" We reviewed the responses provided, read the accumulated literature, reviewed the R code, attempted to replicate the methods proposed, and conducted a two-hour in person meeting with Drs. Potti, Nevins and Barry on December 16, 2009. For the first charge we think that the Duke investigators have,

Focusing on the breast cancer trial predictor of Adriamycin sensitivity, we attempted to independently replicate results using [REDACTED] software. We've made a few minor modifications in various steps, but don't expect these to be significant. In the official prediction algorithm for the clinical trial, 73 reference samples are used which were generated previously using the same protocol as in the trial. However, no sensitivity phenotypes are available for these samples, so in order to obtain some concrete results in validating performance of the predictive methodology, we used the 133 samples from MDA, as in [REDACTED]. Instead of [REDACTED] [REDACTED], we performed a single iteration of two-fold cross-validation, splitting the 133 samples randomly into two groups of 66 and 67. We first used the 66 as the reference sample and the 67 as a validation set, and then reversed their roles. We performed the following analysis steps, intended to parallel those in auto.MARCOM.TEST.R:

1. On Adria_U95.txt: [REDACTED]

2. On MDA133.rma.txt: [REDACTED]



matrix:

_Color



[Redacted text]

. This result emphasizes the importance of using a proper reference set in this kind of approach.

[Redacted text]

How to avoid this?

- A basic rebuttal of the investigators: the analyses were themselves correct, but disseminations were contaminated with errors after the fact
- There is some plausibility to this: creating web-accessible resources can introduce errors that are irrelevant to the basic analyses
- The plausibility and acceptability of this response is demolished by the scope of errors found after a first round of revisions, and by review of the evident misunderstandings of statistical analysis and reproducibility in the published dialogues (e.g., vehement retorts by Duke investigators in response to *J Clin Oncol.* 2008 Mar 1;26(7):1186-7; author reply 1187-8).
- MD Anderson suggestion: in the analysis of genome-scale data, common errors are simple; simple errors are common

Remedies in engineering and law

- Engineering
 - Work as if you are the independent investigator replicating the proposed analysis – build the web (more generally, disseminable) resources at an early stage and work only from them
 - When base resources are revised/reversioned, rerun the script
 - With highly accessible concurrent computing (e.g., multicore/SGE with R) massive but well-scripted workflows can be rerun and reevaluated on short notice
- Law
 - Redaction and high level of confidentiality are a response to risk of being ‘scooped’
 - If property rights over federally funded research results are so important, legal and administrative tools for securing them simultaneously with transparency must be developed
 - See work on RRS by Victoria Stodden (Columbia U. Stats) and colleagues