



Reproducible Research Case Study

Identifying Harmful Constituents in Particulate Matter Air Pollution

Roger D. Peng, Associate Professor of Biostatistics
Johns Hopkins Bloomberg School of Public Health

What Causes PM to be Toxic?

- PM is composed of many different chemical elements
- Some components of PM may be more harmful than others
- Some sources of PM may be more dangerous than others
- Identifying harmful chemical constituents may lead us to strategies for controlling sources of PM

NMMAAPS

- The National Morbidity, Mortality, and Air Pollution Study (NMMAAPS) was a national study of the short-term health effects of ambient air pollution
- Focused primarily on particulate matter (PM_{10}) and ozone (O_3)
- Health outcomes included mortality from all causes and hospitalizations for cardiovascular and respiratory diseases
- Key publications
 - <http://www.ncbi.nlm.nih.gov/pubmed/11098531>
 - <http://www.ncbi.nlm.nih.gov/pubmed/11354823>
- Funded by the [Health Effects Institute](#)
 - Roger Peng currently serves on the Health Effects Institute Health Review Committee

NMMAAPS and Reproducibility

- Data made available at the Internet-based Health and Air Pollution Surveillance System (<http://www.ihapss.jhsph.edu>)
- Research results and software also available at iHAPSS
- Many studies (over 67 published) have been conducted based on the public data <http://www.ncbi.nlm.nih.gov/pubmed/22475833>
- Has served as an important test bed for methodological development

What Causes Particulate Matter to be Toxic?

Research

Cardiovascular Effects of Nickel in Ambient Air

Morton Lippmann,^{1*} Kazuhiko Ito,¹ Jing-Shiang Hwang,² Polina Maciejczyk,¹ and Lung-Chi Chen^{1*}

¹New York University School of Medicine, Nelson Institute of Environmental Medicine, Tuxedo, New York, USA; ²Institute of Chemistry, Academia Sinica, Taipei, Taiwan

<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1665439/>

- Lippmann *et al.* found strong evidence that Ni modified the short-term effect of PM_{10} across 60 US communities
- No other PM chemical constituent seemed to have the same modifying effect
- Too simple to be true?

A Reanalysis of the Lippmann *et al.* Study

Research

Does the Effect of PM₁₀ on Mortality Depend on PM Nickel and Vanadium Content? A Reanalysis of the NMMAPS Data

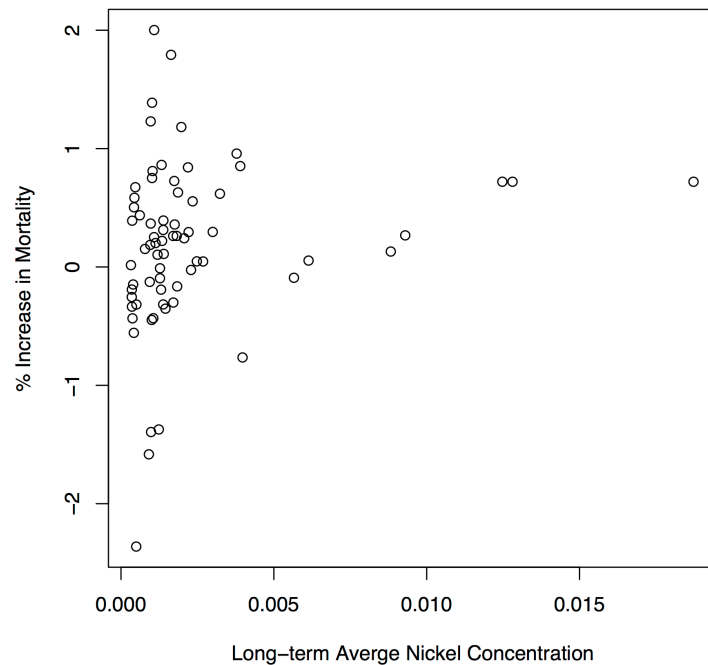
Francesca Dominici,¹ Roger D. Peng,¹ Keita Ebisu,² Scott L. Zeger,¹ Jonathan M. Samet,³ and Michelle L. Bell²

¹Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA; ²School of Forestry and Environmental Studies, Yale University, New Haven, Connecticut, USA; ³Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA

<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2137127/>

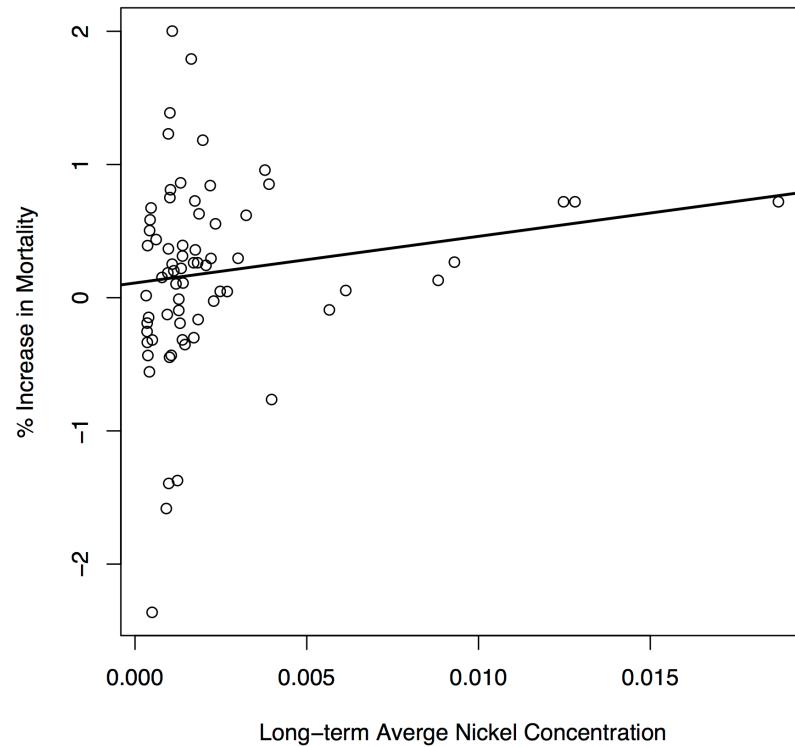
- Reexamine the data from NMMAPS and link with PM chemical constituent data
- Are the findings sensitive to levels of Nickel in New York City?

Does Nickel Make PM Toxic?



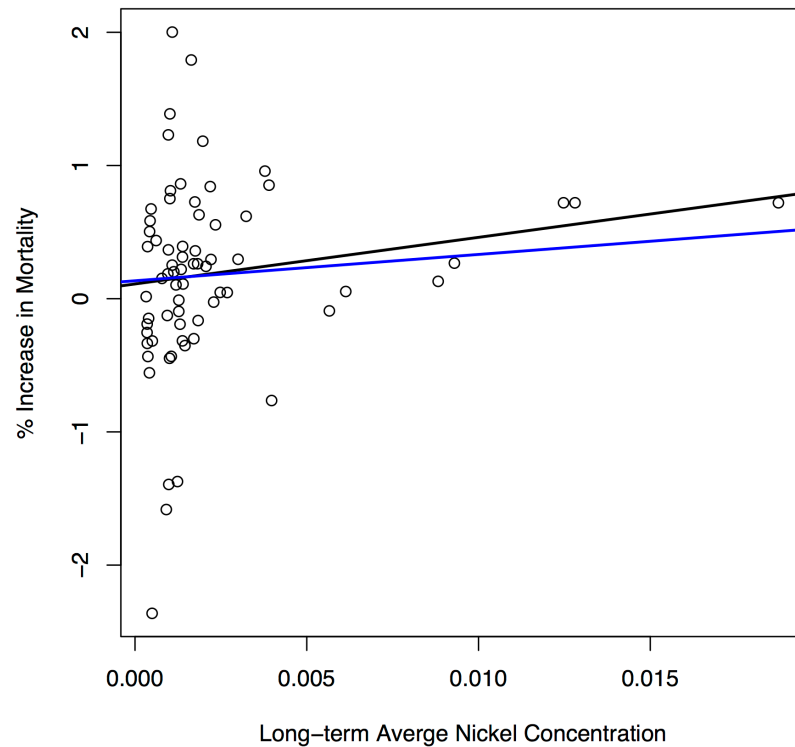
- Long-term average nickel concentrations appear correlated with PM risk
- There appear to be some outliers on the right-hand side (New York City)

Does Nickel Make PM Toxic?



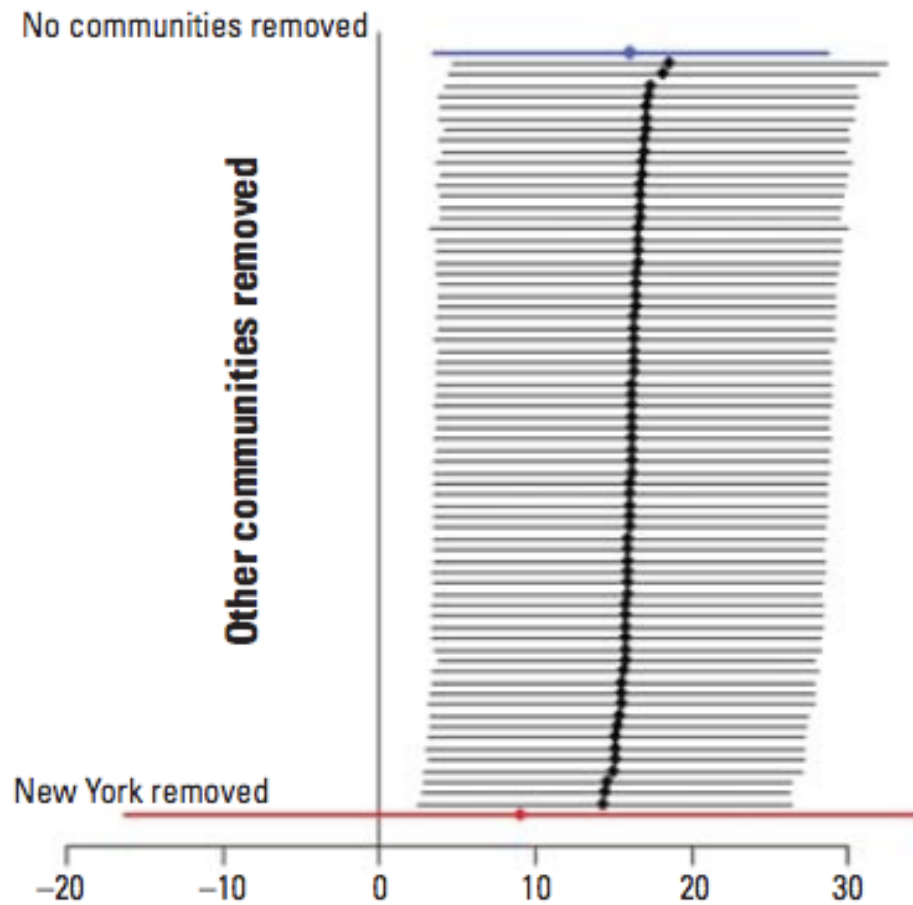
- Regression line statistically significant ($p < 0.01$)

Does Nickel Make PM Toxic?



- Adjusted regression line (blue) no longer statistically significant ($p < 0.31$)

Does Nickel Make PM Toxic?



What Have We Learned?

- New York does have very high levels of nickel and vanadium, much higher than any other US community
- There is evidence of a positive relationship between Ni concentrations and PM_{10} risk
- The strength of this relationship is highly sensitive to the observations from New York City
- Most of the information in the data is derived from just 3 observations

Lessons Learned

- Reproducibility of NMMAAPS allowed for a secondary analysis (and linking with PM chemical constituent data) investigating a novel hypothesis (Lippmann *et al.*)
- Reproducibility also allowed for a critique of that new analysis and some additional new analysis (Dominici *et al.*)
- Original hypothesis not necessarily invalidated, but evidence not as strong as originally suggested (more work should be done)
- Reproducibility allows for the scientific discussion to occur in a timely and informed manner
- This is how science works

The Importance of Reproducibility in High-Throughput Biology: Case Studies in Forensic Bioinformatics

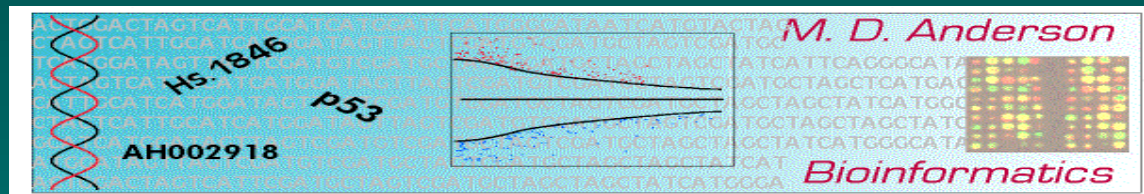
Keith A. Baggerly

Bioinformatics and Computational Biology

UT M. D. Anderson Cancer Center

kabagg@mdanderson.org

Cambridge, 4 September 2010



Why is RR So Important in H-TB?

Our intuition about what “makes sense” is very poor in high dimensions. To use “genomic signatures” as biomarkers, we need to know they’ve been assembled correctly.

Without documentation, we may need to employ *forensic bioinformatics* to infer what was done to obtain the results.

Let’s examine some case studies involving an important clinical problem: *can we predict how a given patient will respond to available chemotherapeutics?*

Using the NCI60 to Predict Sensitivity

ature.com/naturemedicine

Genomic signatures to guide the use of chemotherapeutics

Anil Potti^{1,2}, Holly K Dressman^{1,3}, Andrea Bild^{1,3}, Richard F Riedel^{1,2}, Gina Chan⁴, Robyn Sayer⁴,
Janiel Cragun⁴, Hope Cottrill⁴, Michael J Kelley², Rebecca Petersen⁵, David Harpole⁵, Jeffrey Marks⁵,
Andrew Berchuck^{1,6}, Geoffrey S Ginsburg^{1,2}, Phillip Febbo¹⁻³, Johnathan Lancaster⁴ &
Joseph R Nevins¹⁻³

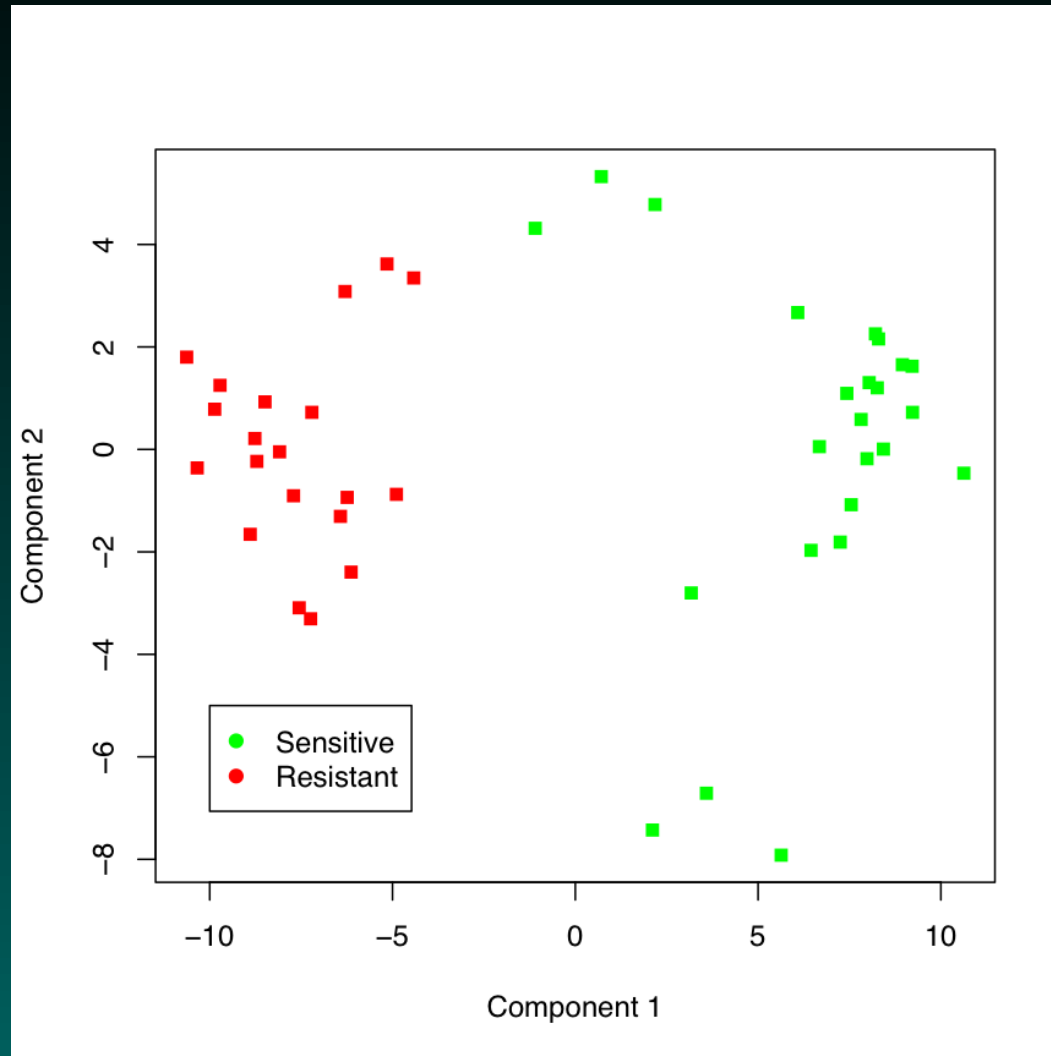
Potti et al (2006), Nature Medicine, 12:1294-1300.

The main conclusion is that we can use microarray data from cell lines (the NCI60) to define drug response “signatures”, which can be used to predict whether patients will respond.

They provide examples using 7 commonly used agents.

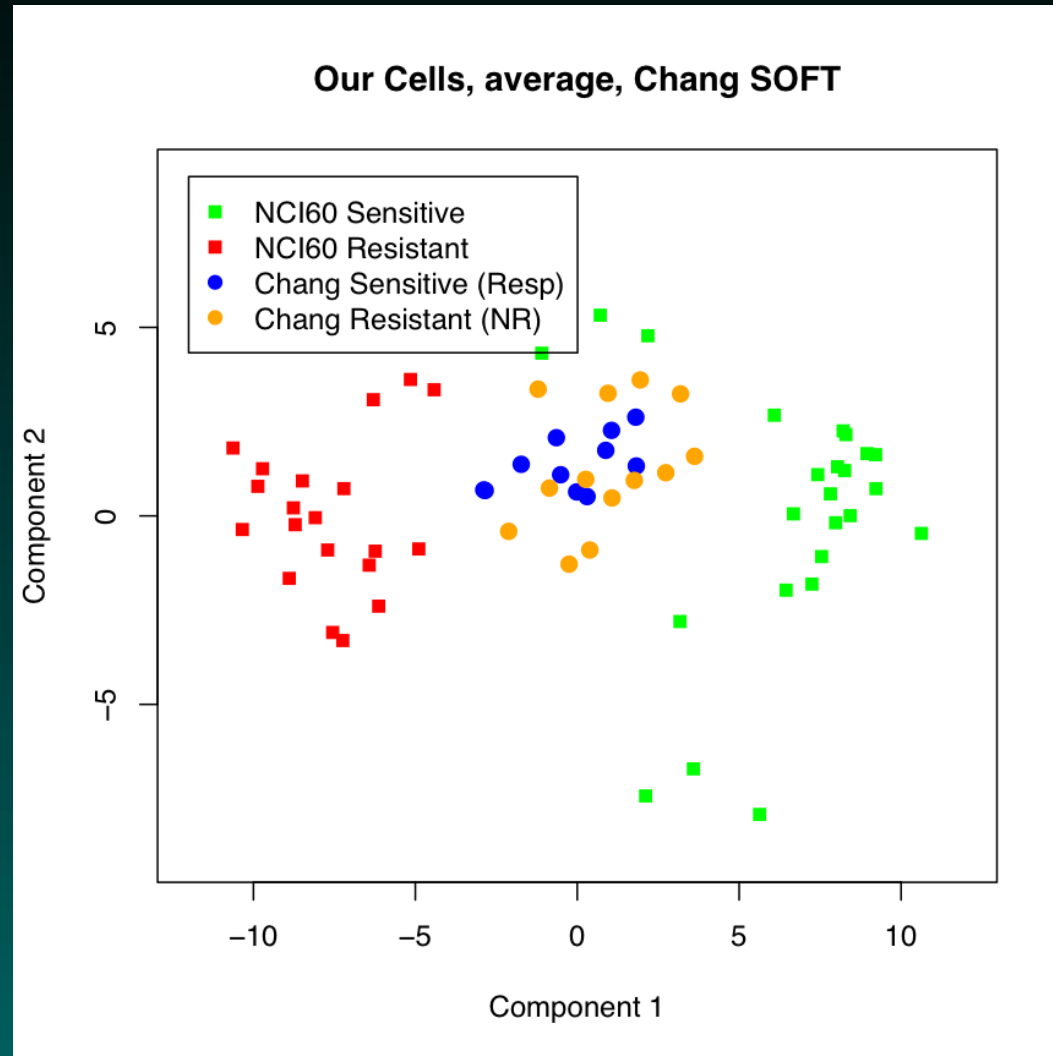
This got people at MDA very excited.

Fit Training Data



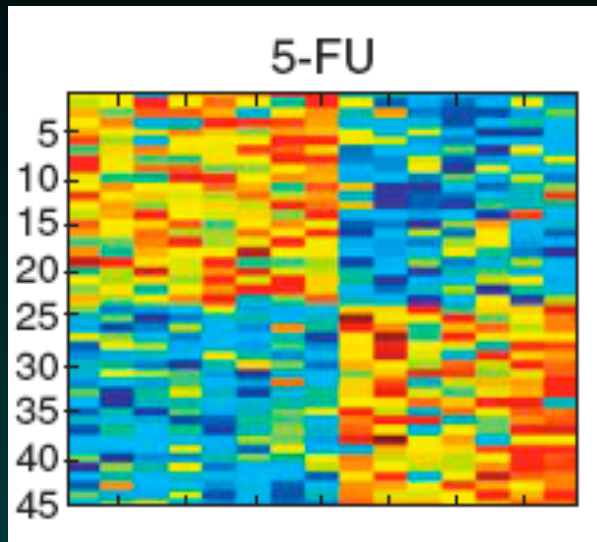
We want the test data to split like this...

Fit Testing Data



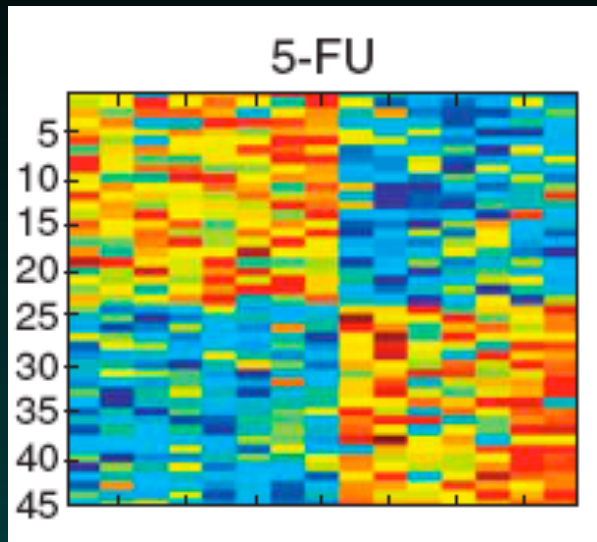
But it *doesn't*. Did we do something wrong?

5-FU Heatmaps

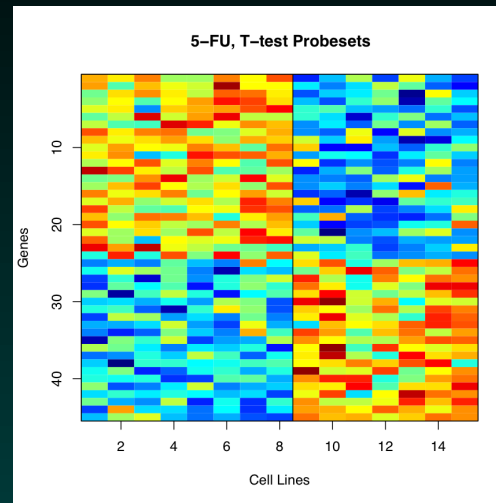


Nat Med Paper

5-FU Heatmaps

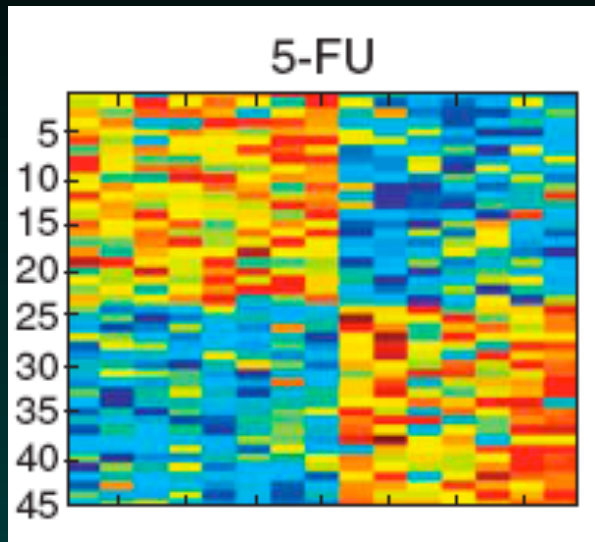


Nat Med Paper

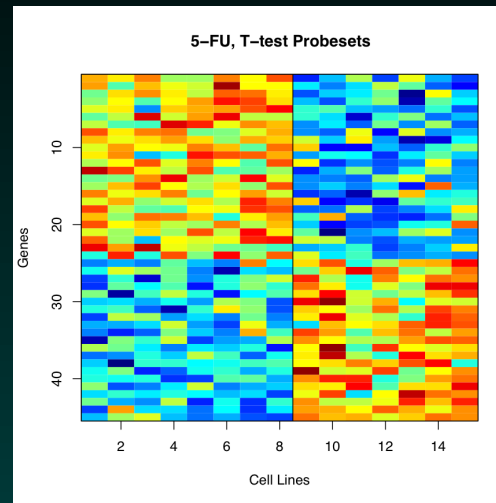


Our t-tests

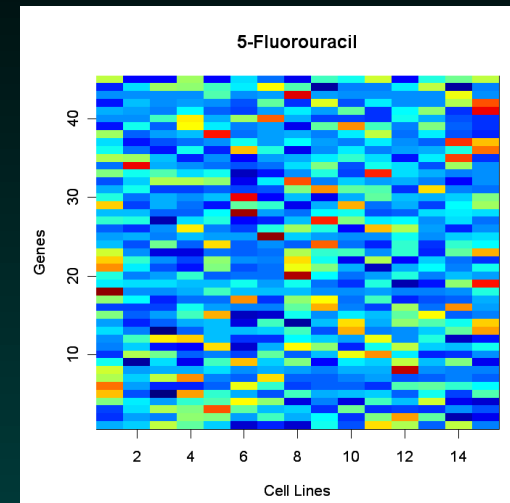
5-FU Heatmaps



Nat Med Paper



Our t-tests



Reported Genes

Their List and Ours

```
> temp <- cbind(  
  sort(rownames(pottiUpdated)[fuRows]),  
  sort(rownames(pottiUpdated)[  
    fuTQNorm@p.values <= fuCut]));  
> colnames(temp) <- c("Theirs", "Ours");  
> temp
```

Theirs

Ours

...

[3,] "1881_at" "1882_g_at"

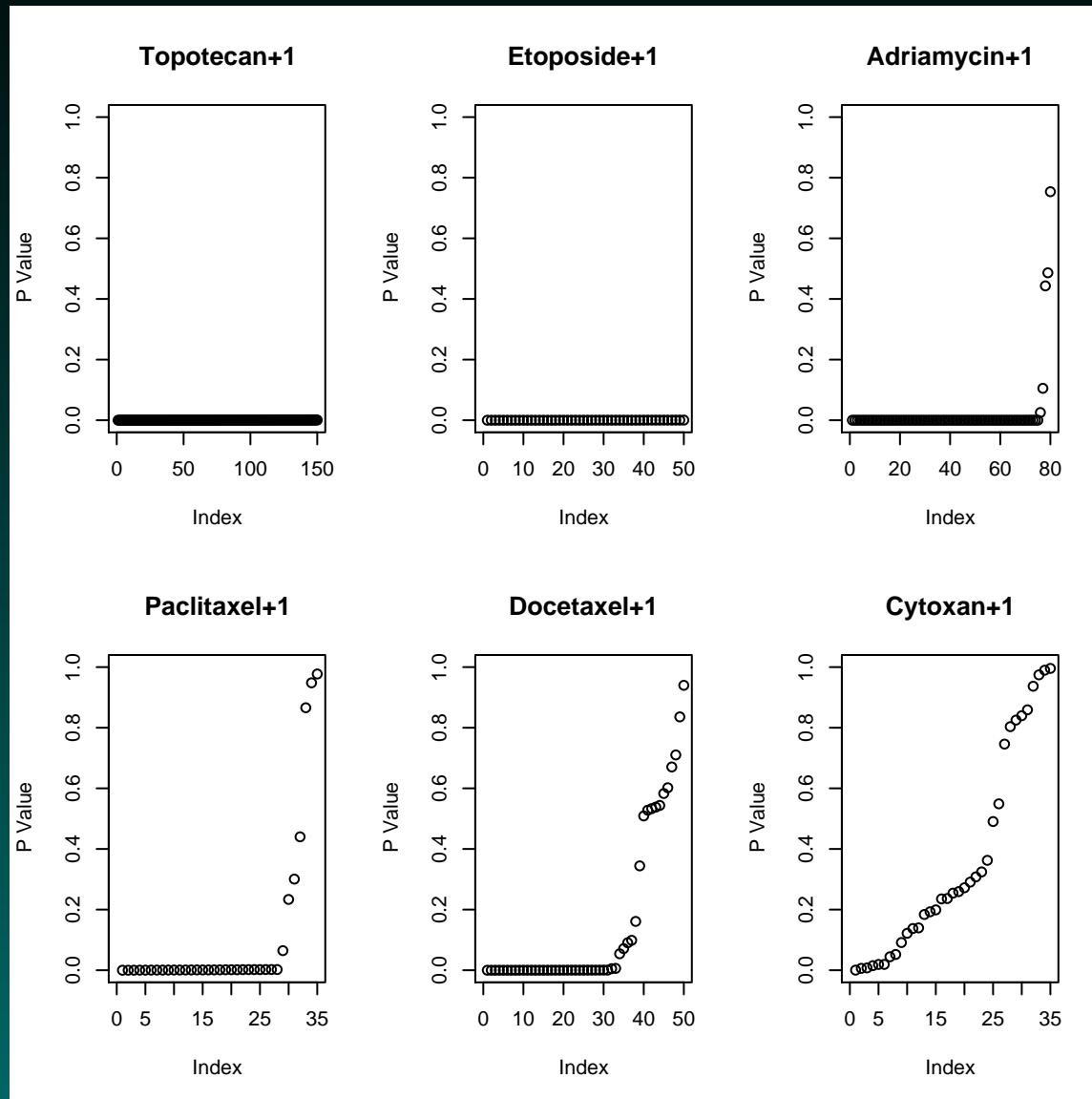
[4,] "31321_at" "31322_at"

[5,] "31725_s_at" "31726_at"

[6,] "32307_r_at" "32308_r_at"

...

Offset P-Values: Other Drugs



Using Their Software

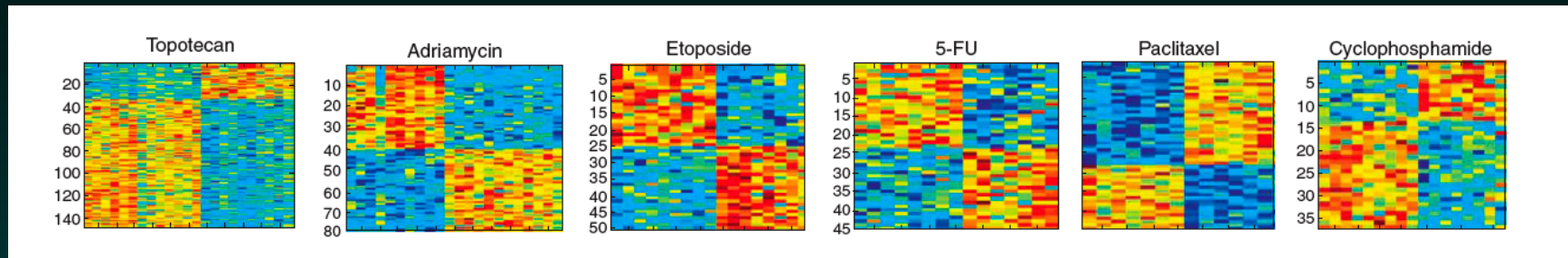
Their software requires two input files:

1. *a quantification matrix*, genes by samples, with a header giving classifications (0 = Resistant, 1 = Sensitive, 2 = Test)
2. *a list of probeset ids* in the same order as the quantification matrix. ***This list must not have a header row.***

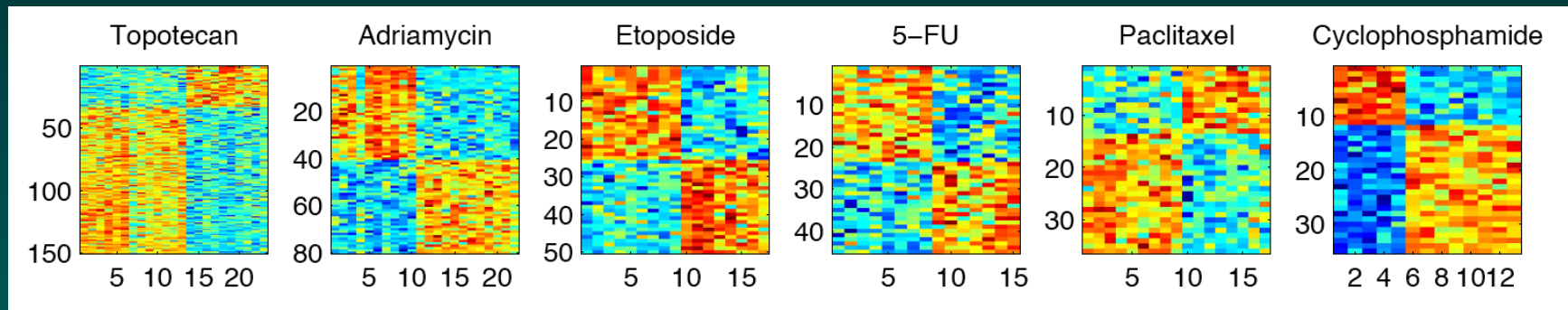
What do we get?

Heatmaps Match Exactly for Most Drugs!

From the **paper**:

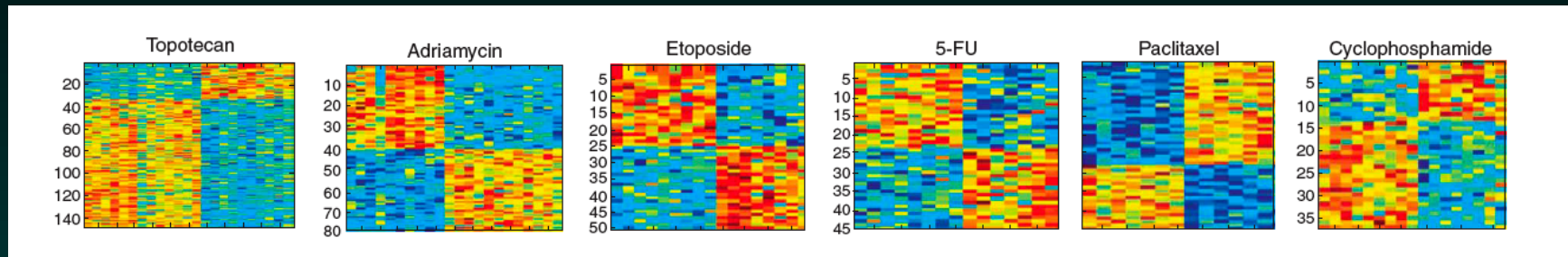


From the **software**:

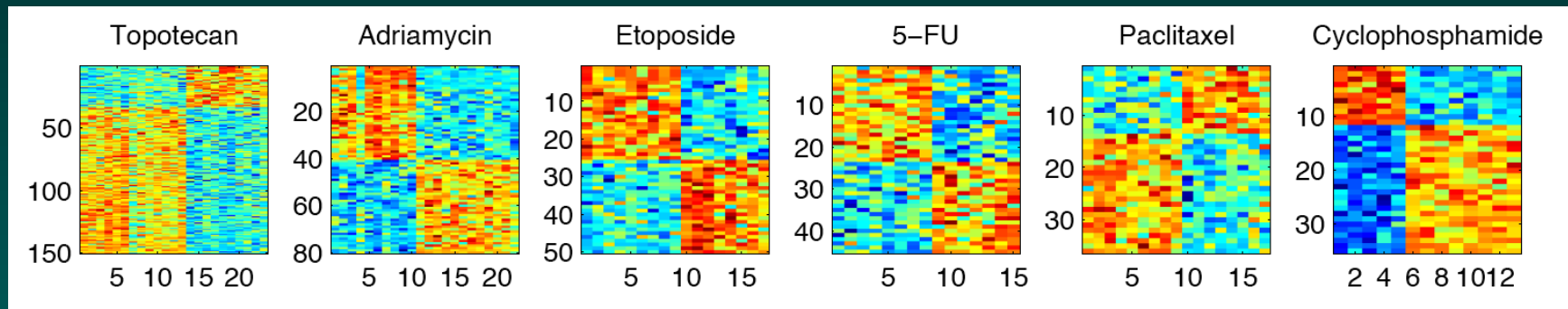


Heatmaps Match Exactly for Most Drugs!

From the **paper**:

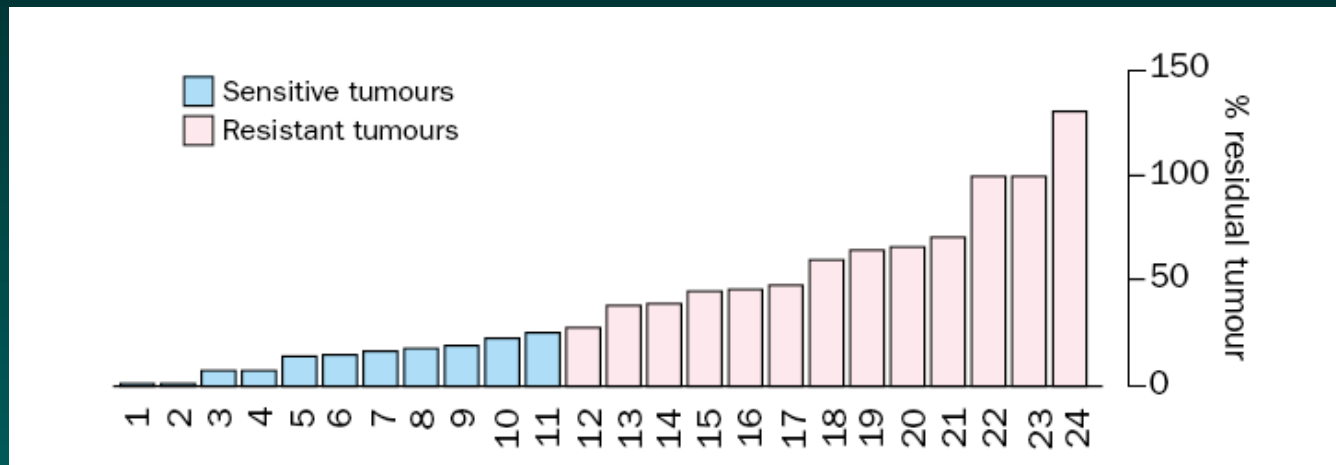
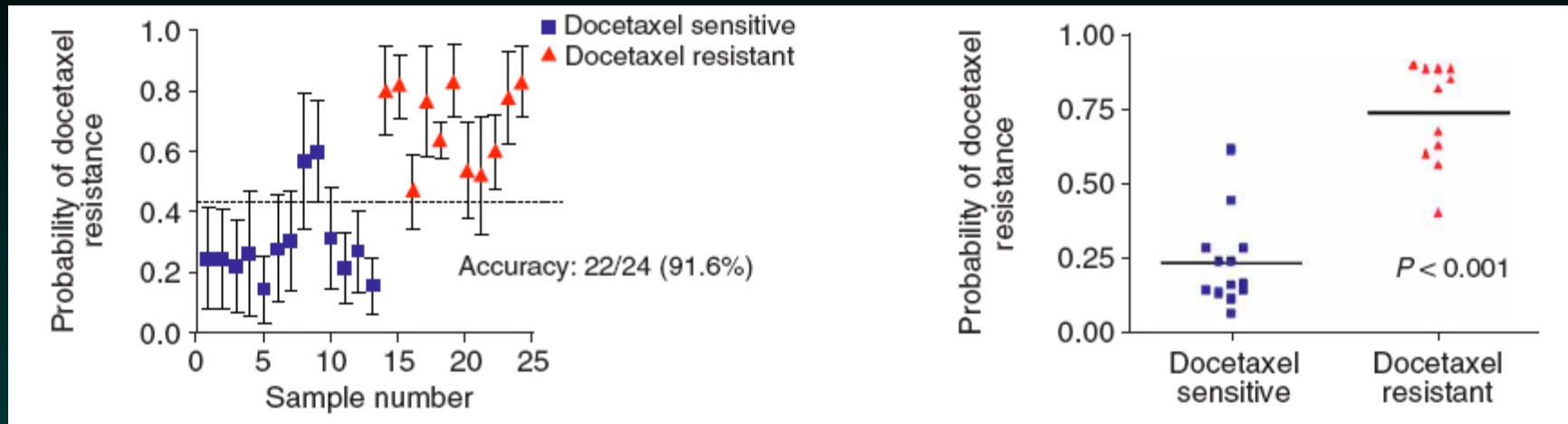


From the **software**:

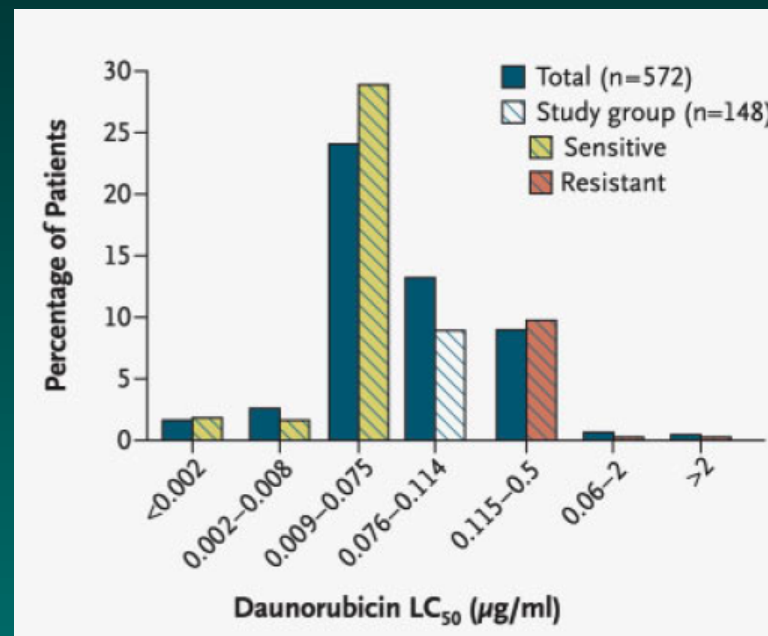
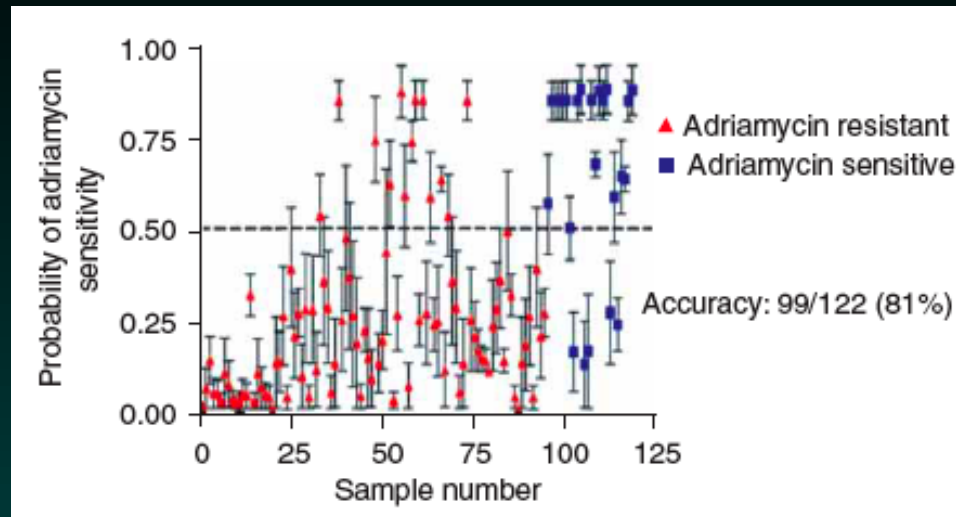


We match heatmaps but not gene lists? We'll come back to this, because their software also gives *predictions*.

Predicting Docetaxel (Chang 03)



Predicting Adriamycin (Holleman 04)



There Were Other Genes...

The 50-gene list for docetaxel has 19 “outliers”.

The initial paper on the test data (Chang et al) gave a list of 92 genes that separated responders from nonresponders.

Entries 7-20 in Chang et al's list comprise 14/19 outliers.

The others: ERCC1, ERCC4, ERBB2, BCL2L11, TUBA3.
These are the genes named to explain the biology.

RR Theme: Don't Take My Word For It!

Read the paper! Coombes, Wang & Baggerly, Nat Med, Nov 6, 2007, 13:1276-7, author reply 1277-8.

Try it yourselves! All of the raw data, documentation*, and code* is available from our web site (*and from Nat Med):

[http://bioinformatics.mdanderson.org/
Supplements/ReproRsch-Chemo](http://bioinformatics.mdanderson.org/Supplements/ReproRsch-Chemo).

Potti/Nevins Reply (Nat Med 13:1277-8)

Labels for Adria are correct – details on their web page.

They've gotten the approach to work again. (Twice!)

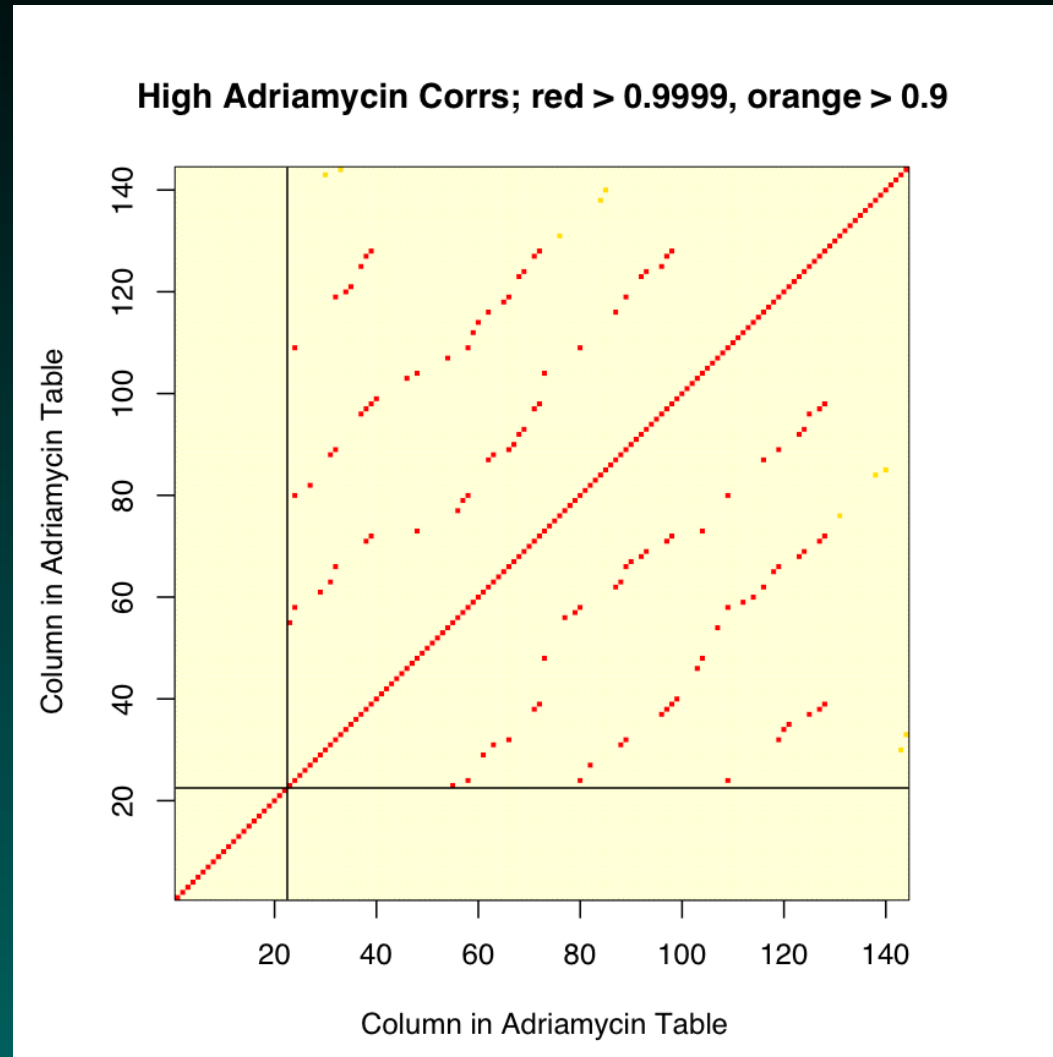
Pharmacogenomic Strategies Provide a Rational Approach to the Treatment of Cisplatin-Resistant Patients With Advanced Cancer

David S. Hsu, Bala S. Balakumaran, Chaitanya R. Acharya, Vanja Vlahovic, Kelli S. Walters, Katherine Garman, Carey Anders, Richard F. Riedel, Johnathan Lancaster, David Harpole, Holly K. Dressman, Joseph R. Nevins, Phillip G. Febbo, and Anil Potti

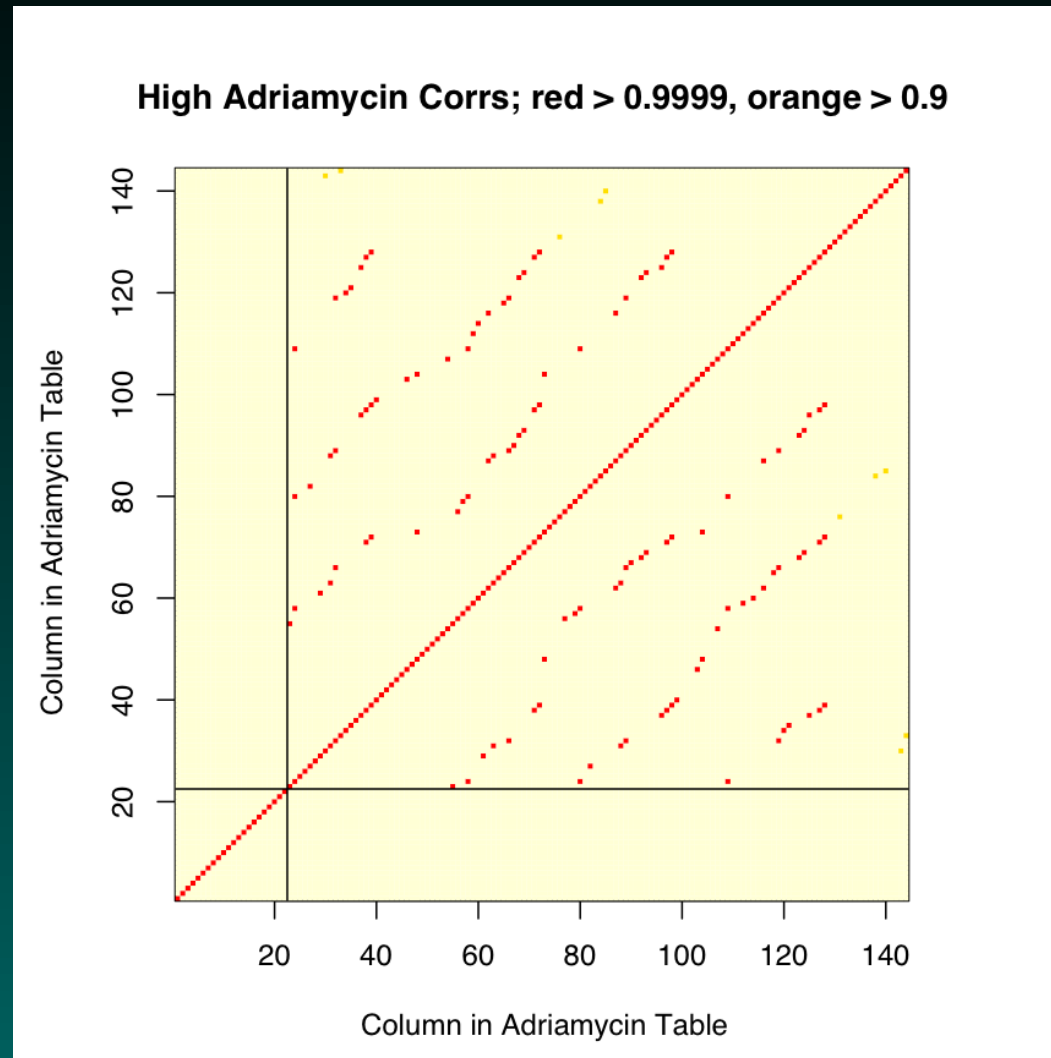
Validation of gene signatures that predict the response of breast cancer to neoadjuvant chemotherapy: a substudy of the EORTC 10994/BIG 00-01 clinical trial

Hervé Bonnefoi, Anil Potti, Mauro Delorenzi, Louis Mauriac, Mario Campone, Michèle Tubiana-Hulin, Thierry Petit, Philippe Rouanet, Jacek Jassem, Emmanuel Blot, Véronique Becette, Pierre Farmer, Sylvie André, Chaitanya R Acharya, Sayan Mukherjee, David Cameron, Jonas Bergh, Joseph R Nevins, Richard D Iggo

Adriamycin 0.9999+ Correlations (Reply)



Adriamycin 0.9999+ Correlations (Reply)



Redone in Aug 08, “using only the 95 unique samples”

The First 20 Files Now Named

Sample ID		Response	
1	GSM44303	RES	
2	GSM44304	RES	
3	GSM9653	RES	
4	GSM9653	RES	
5	GSM9654	RES	
6	GSM9655	RES	
7	GSM9656	RES	
8	GSM9657	RES	
9	GSM9658	SEN	
10	GSM9658	SEN	
11	GSM9694	RES	
12	GSM9695	RES	
13	GSM9696	RES	
14	GSM9698	RES	
15	GSM9699	SEN	
16	GSM9701	RES	
17	GSM9708	RES	
18	GSM9708	SEN	
19	GSM9709	RES	
20	GSM9711	RES	

15 duplicates; 6 inconsistent. (61R, 13S, 6B) vs (22,48,10).

Validation 1: Hsu et al

Pharmacogenomic Strategies Provide a Rational Approach to the Treatment of Cisplatin-Resistant Patients With Advanced Cancer

David S. Hsu, Bala S. Balakumaran, Chaitanya R. Acharya, Vanja Vlahovic, Kelli S. Walters, Katherine Garman, Carey Anders, Richard F. Riedel, Johnathan Lancaster, David Harpole, Holly K. Dressman, Joseph R. Nevins, Phillip G. Febbo, and Anil Potti

J Clin Oncol, Oct 1, 2007, 25:4350-7.

Same approach, using **Cisplatin** and **Pemetrexed**.

For cisplatin, U133A arrays were used for training. **ERCC1**, **ERCC4** and **DNA repair** genes are identified as “important”.

With some work, we matched the heatmaps. (Gene lists?)

The 4 We Can't Match (Reply)

203719_at, ERCC1,
210158_at, ERCC4,
228131_at, ERCC1, and
231971_at, FANCM (DNA Repair).

The last two probesets are special.

*These probesets aren't on the U133A arrays that were used.
They're on the U133B.*

Some Timeline Here...

Nat Med Nov 06*, Nov 07*, Aug 08. JCO Lung Oct 07*.
Lancet Oncology Breast Dec 07*. (* errors reported)

Some Timeline Here...

Nat Med Nov 06*, Nov 07*, Aug 08. JCO Lung Oct 07*.
Lancet Oncology Breast Dec 07*. (* errors reported)

May/June 2009: **we learn clinical trials had begun.**

2007: pemetrexed vs cisplatin, pem vs vinorelbine.

2008: docetaxel vs doxorubicin, topotecan vs dox (Moffitt).

Some Timeline Here...

Nat Med Nov 06*, Nov 07*, Aug 08. JCO Lung Oct 07*.
Lancet Oncology Breast Dec 07*. (* errors reported)

May/June 2009: **we learn clinical trials had begun.**

2007: pemetrexed vs cisplatin, pem vs vinorelbine.

2008: docetaxel vs doxorubicin, topotecan vs dox (Moffitt).


Sep 1. Paper submitted to *Annals of Applied Statistics*.

Sep 14. Paper online at *Annals of Applied Statistics*.

Sep-Oct: Story covered by *The Cancer Letter*, Duke starts internal investigation, suspends trials.

So, what happened next?

Jan 29, 2010

The logo for 'The Cancer Letter' features the word 'THE' in a small, white, sans-serif font on the left. To its right, the word 'CANCER' is written in a large, bold, white, sans-serif font. Below 'CANCER', the word 'LETTER' is written in a smaller, white, sans-serif font. The entire logo is set against a solid red rectangular background.

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

**Duke In Process To Restart Three Trials
Using Microarray Analysis Of Tumors**

By Paul Goldberg

Duke University said it is in the process of restarting three clinical trials using microarray analysis of patient tumors to predict their response to chemotherapy.

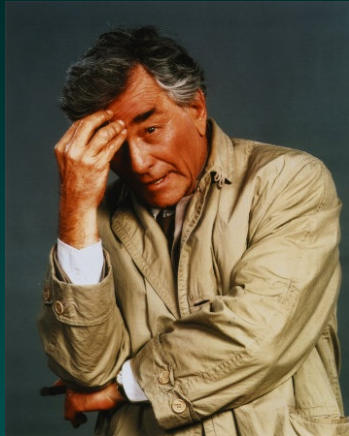
Their investigation's results *"strengthen ... confidence in this evolving approach to personalized cancer treatment."*

Why We're Unhappy...

“While the reviewers approved of our sharing the report with the NCI, *we consider it a confidential document*” (Duke). A *future paper* will explain the methods.

Why We're Unhappy...

“While the reviewers approved of our sharing the report with the NCI, *we consider it a confidential document*” (Duke). A *future paper* will explain the methods.

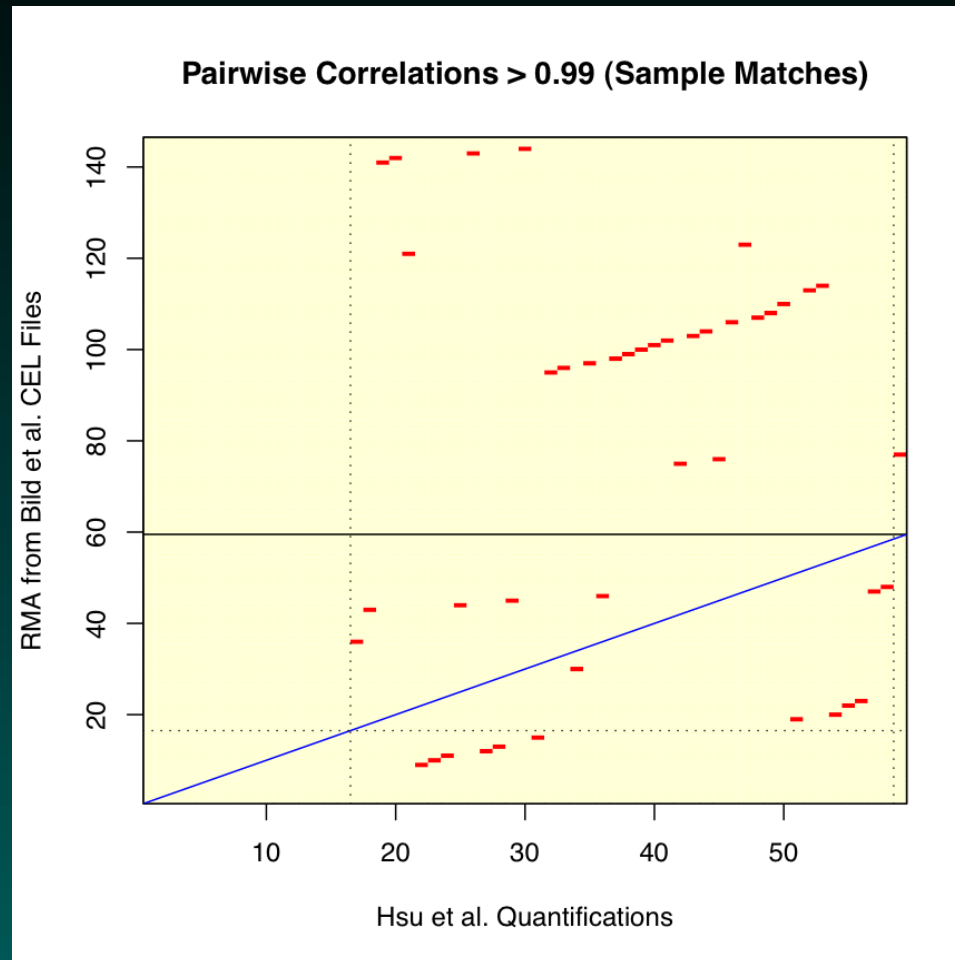


oh, there's just one more thing...

In mid-Nov (mid-investigation), the Duke team posted new data for cisplatin and pemetrexed (in trials since '07).

These included quantifications for 59 ovarian cancer test samples (from GSE3149) used for predictor validation.

We Tried Matching The Samples



We correlated the 59 vectors with all samples in GSE3149.
43 samples are mislabeled; 16 don't match at all.

FOI(L)A!

April 7: Paul Goldberg of *the Cancer Letter* requests “access to and copies of the report (and attendant data)” from the NCI under the Freedom of Information Act (FOIA).

May 3: redacted report supplied.

FOI(L)A!

April 7: Paul Goldberg of *the Cancer Letter* requests “access to and copies of the report (and attendant data)” from the NCI under the Freedom of Information Act (FOIA).

May 3: redacted report supplied.

“we were unable to identify a place where the statistical methods were described in sufficient detail to independently replicate the findings of the papers.” – **review panel**

The report makes no mention of the problems with cisplatin/pemetrexed that arose during the investigation.

May 14, 2010

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.

“We have asked [CALGB] to remove the Lung Metagene Score from the trial, because we were unable to confirm the score’s utility” – *Jeff Abrams, CTEP director*

(The NCI doesn’t directly sponsor the resumed trials.)

July 16, 2010

The image shows the front cover of 'The Cancer Letter'. The top half has a red background with the title 'THE CANCER LETTER' in white, bold, sans-serif capital letters. 'THE' is smaller and to the left of 'CANCER'. 'LETTER' is below 'CANCER'. The bottom half has a white background. It contains the mailing address 'PO Box 9905 Washington DC 20016 Telephone 202-362-1809' in a black sans-serif font. Below the address is a large, bold, black headline: 'Prominent Duke Scientist Claimed Prizes He Didn't Win, Including Rhodes Scholarship'. At the bottom of the white section, the author's name 'By Paul Goldberg' is written in a smaller, italicized black font.

THE CANCER LETTER

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

**Prominent Duke Scientist Claimed Prizes
He Didn't Win, Including Rhodes Scholarship**

By Paul Goldberg

July 19, 2010

“Duke administrators accomplished something monumental: **they triggered a public expression of outrage from biostatisticians.**”

A Baron, K Bandeen-Roche, D Berry, J Bryan,
V Carey, K Chaloner, M Delorenzi, B Efron,
R Elston, D Ghosh, J Goldberg, S Goodman,
F Harrell, S Hilsenbeck, W Huber, R Irizarry,
C Kendzioriski, M Kosorok, T Louis, JS Marron,
M Newton, M Ochs, G Parmigiani*, J Quackenbush,
G Rosner, I Ruczinski, Y Shyr*, S Skates,
TP Speed, JD Storey, Z Szallasi, R Tibshirani,
S Zeger

Req to Varmus, DoD, ORI, Duke: suspend trials.

Subsequent Events, and a Caveat

Duke announces trials resuspended

NPR blog, Science blog, Nature blog, NYT blog, article

Lancet Oncology issues Expression of Concern

Varmus & Duke request IOM Involvement

Questions raised about NEJM paper

JCO launches investigation

More awards found to be wrong, COI claims

<http://groups.google.com/group/reproducible-research>

Correspondence to Nature

Subsequent Events, and a Caveat

Duke announces trials resuspended

NPR blog, Science blog, Nature blog, NYT blog, article

Lancet Oncology issues Expression of Concern

Varmus & Duke request IOM Involvement

Questions raised about NEJM paper

JCO launches investigation

More awards found to be wrong, COI claims

<http://groups.google.com/group/reproducible-research>

Correspondence to Nature

We've seen problems like these before. CAMDA 2002.

Proteomics 2003-5. TCGA current. Others at MDA.

Some Observations

The most common mistakes are simple.

Confounding in the Experimental Design

Mixing up the sample labels

Mixing up the gene labels

Mixing up the group labels

(Most mixups involve simple switches or offsets)

This simplicity is often hidden.

Incomplete documentation

Unfortunately, we suspect

The most simple mistakes are common.

What Should the Norm Be?

For papers?

Things we look for:

1. Data (often mentioned, given MIAME)
2. Provenance
3. Code
4. Descriptions of Nonscriptable Steps
5. Descriptions of Planned Design, if Used.

For clinical trials?

Some Lessons

Is our own work reproducible?

Literate Programming. For the past two years, we have required reports to be prepared in *Sweave*.

Reusing Templates.

Report Structure.

Executive Summaries.

Appendices. Some things we want to know all the time: *SessionInfo*, *Saves*, and *File Location*.

The buzz phrase is *reproducible research*.

Some Acknowledgements

Kevin Coombes

Shannon Neeley, Jing Wang

David Ransohoff, Gordon Mills

Jane Fridlyand, Lajos Pusztai, Zoltan Szallasi

MDACC Ovarian SPORE, Lung SPORE, Breast SPORE

Now in the *Annals of Applied Statistics!* Baggerly and Coombes (2009), 3(4):1309-34.

[http://bioinformatics.mdanderson.org/
Supplements/ReproRsch-All](http://bioinformatics.mdanderson.org/Supplements/ReproRsch-All)

Validation 2: Bonnefoi et al

Validation of gene signatures that predict the response of breast cancer to neoadjuvant chemotherapy: a substudy of the EORTC 10994/BIG 00-01 clinical trial

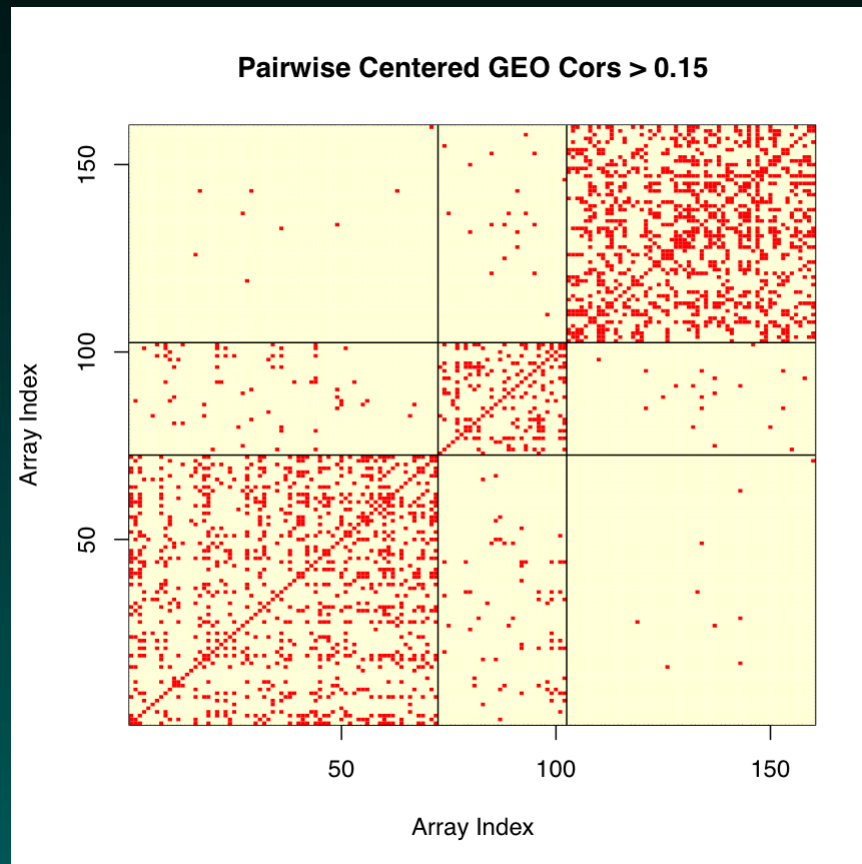
Hervé Bonnefoi, Anil Potti, Mauro Delorenzi, Louis Mauriac, Mario Camponé, Michèle Tubiana-Hulin, Thierry Petit, Philippe Rouanet, Jacek Jassem, Emmanuel Blot, Véronique Becette, Pierre Farmer, Sylvie André, Chaitanya R Acharya, Sayan Mukherjee, David Cameron, Jonas Bergh, Joseph R Nevins, Richard D Iggo

Lancet Oncology, Dec 2007, 8:1071-8. (early access Nov 14)

Similar approach, using signatures for Fluorouracil, Epirubicin Cyclophosphamide, and Taxotere to predict response to combination therapies: **FEC** and **TET**.

Potentially improves ER- response from 44% to 70%.

We Might Expect Some Differences...



High Sample Correlations
after Centering by Gene

Array Run Dates

How Are Results Combined?

Potti et al predict response to TFAC, Bonnefoi et al to TET and FEC. Let $P()$ indicate prob sensitive. The rules used are as follows.

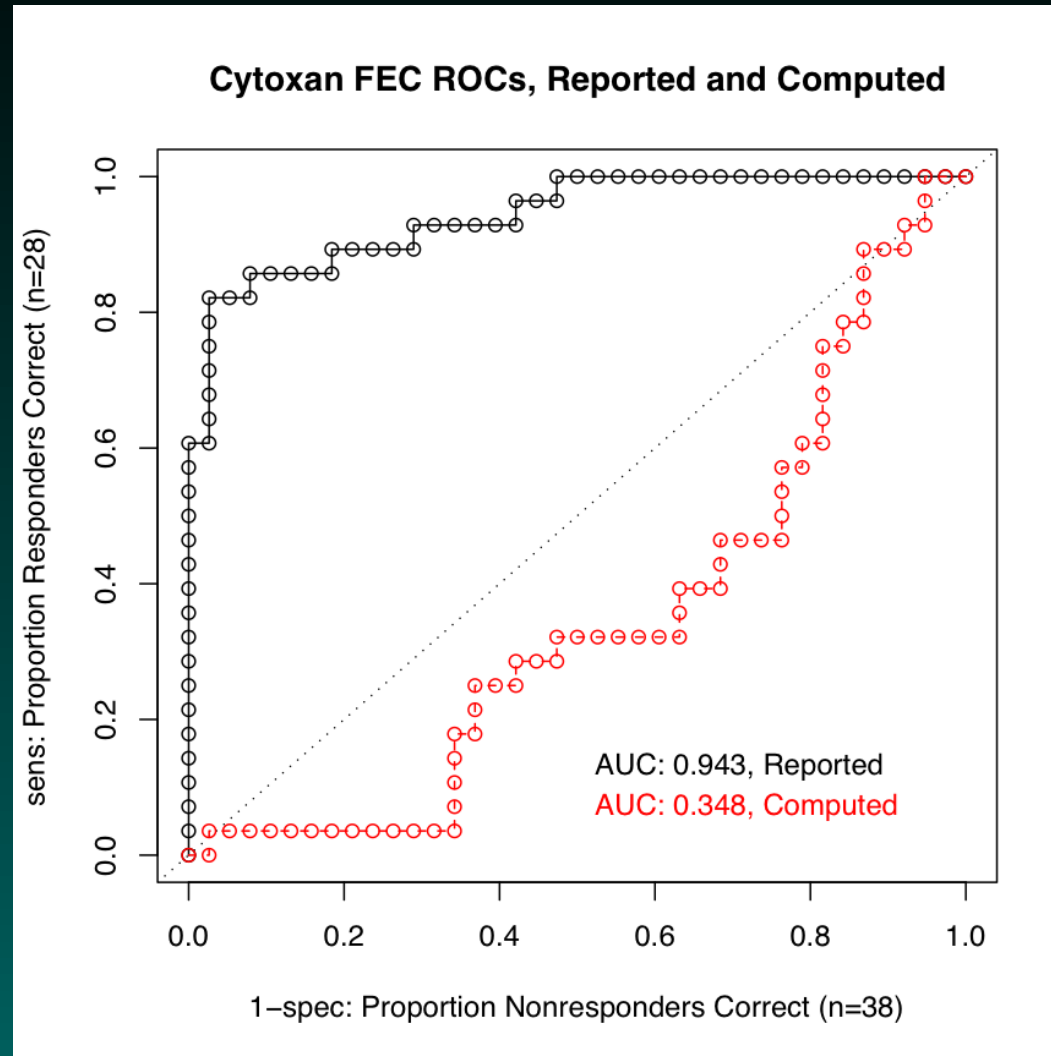
$$P(TFAC) = P(T) + P(F) + P(A) + P(C) - P(T)P(F)P(A)P(C).$$

$$P(ET) = \max[P(E), P(T)].$$

$$P(FEC) = \frac{5}{8}[P(F) + P(E) + P(C)] - \frac{1}{4}.$$

Each rule is different.

Predictions for Individual Drugs? (Reply)



Does cytoxan make sense?

What About Blinded Validation?

“Data was made available to us, blinded. All we got was the gene expression data. We ran the predictions and sent it back to the EORTC investigators” – *Joe Nevins, Oct 2.*

What About Blinded Validation?

“Data was made available to us, blinded. All we got was the gene expression data. We ran the predictions and sent it back to the EORTC investigators” – *Joe Nevins, Oct 2.*

Sample info supplied:

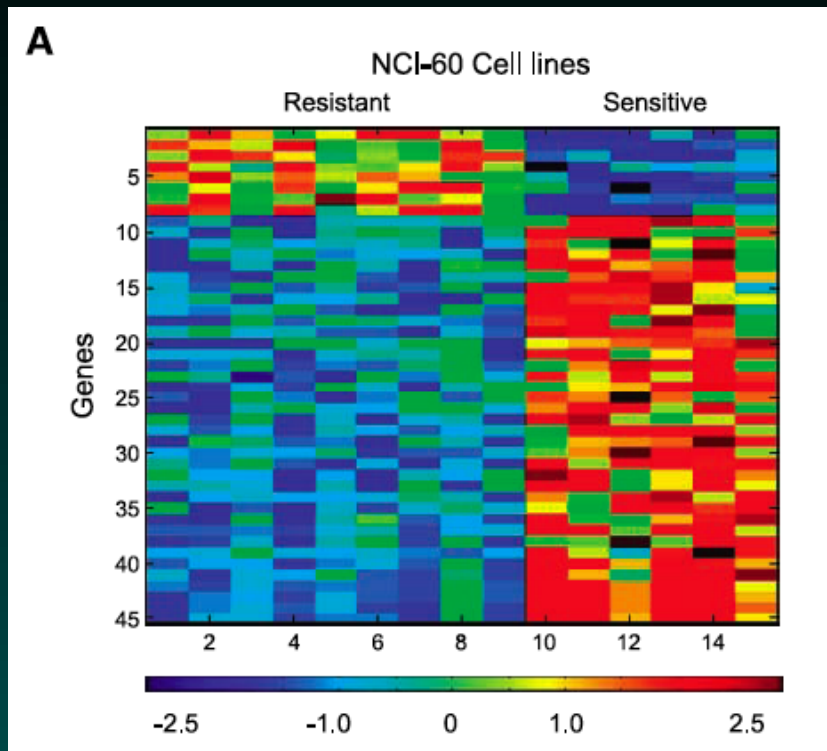
Arm, Composite label

A, npCR Ep P- T3 N1 HB01 ...

A, pCR Ep Pp T2 N1 HB04

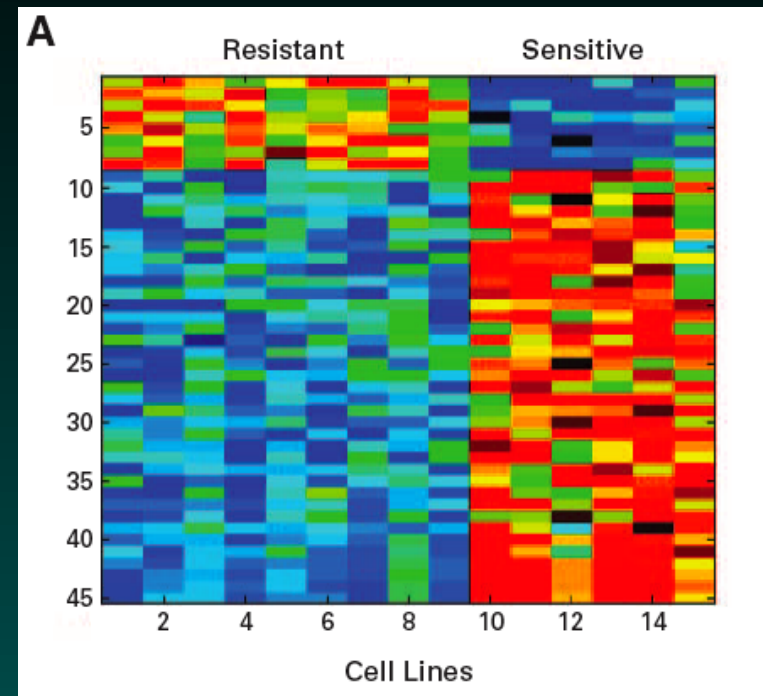
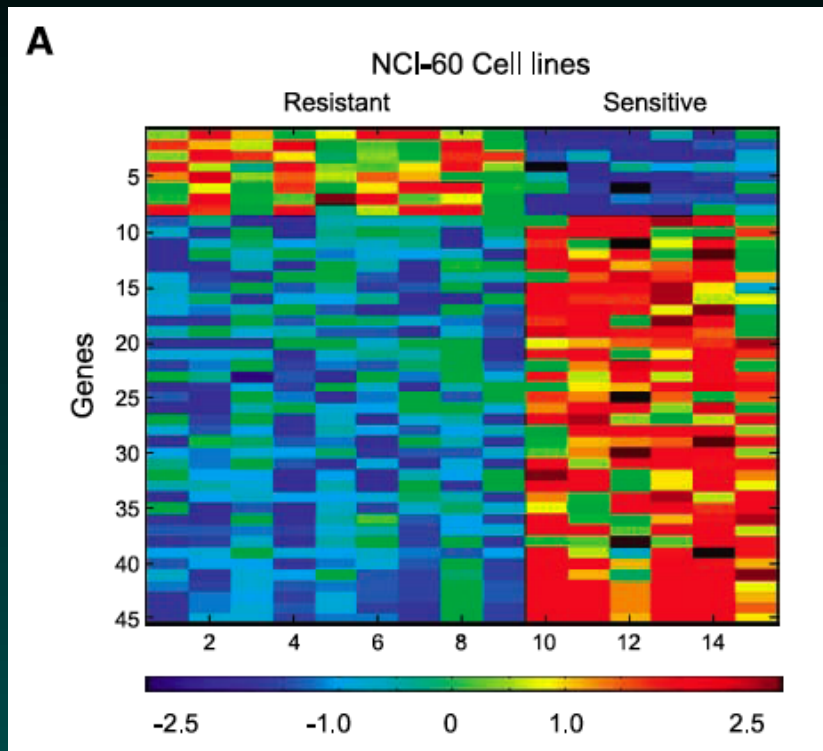
The data weren't blinded.

Temozolomide Heatmaps



Augustine et al., 2009, *Clin Can Res*, 15:502-10, Fig 4A.
Temozolomide, NCI-60.

Temozolomide Heatmaps



Augustine et al., 2009, *Clin Can Res*, 15:502-10, Fig 4A.
Temozolomide, NCI-60.

Hsu et al., 2007, *J Clin Oncol*, 25:4350-7, Fig 1A.
Cisplatin, Gyorffy cell lines.

Index

Title

Cell Line Story

1. Trying it Ourselves
2. Matching Features
3. Using Software/Making Predictions
4. The Reply
5. Adriamycin Followup
6. Hsu et al (Cisplatin)
7. Bonnefoi et al (Combination Therapy)
8. More Recent (Temozolomide)
9. Timeline, Trials, Cancer Letter
10. Trial Restart and Objections
11. FOIA
12. Final Lessons

What is the question?

By Jeff Leek* and Roger D. Peng

Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD 21205, USA.

*Corresponding author. E-mail: jleek@jhsph.edu, jtleek@gmail.com

Mistaking the type of question being considered is the most common error in data analysis.

Over the past 2 years, increased focus on statistical analysis brought on by the era of big data has pushed the issue of reproducibility out of the pages of academic journals and into the popular consciousness (1). Just weeks ago, a paper about the relationship between tissue-specific cancer incidence and stem cell divisions (2) was widely misreported because of misunderstandings about the primary statistical argument in the paper (3). Public pressure has contributed to the massive recent adoption of reproducible research, with corresponding improvements in reproducibility. But an analysis can be fully reproducible and still be wrong. Even the most spectacularly irreproducible analyses—like those underlying the ongoing lawsuits (4) over failed genomic signatures for chemotherapy assignment (5)—are ultimately reproducible (6). Once an analysis is reproducible, the key question we want to answer is, “Is this data analysis correct?” We have found that the most frequent failure in data analysis is mistaking the type of question being considered.

Any specific data analysis can be broadly classified into one of six types (see the figure). The least challenging of these is a descriptive data analysis, which seeks to summarize the measurements in a single data set without further interpretation. An example is the United States Census, which aims to describe how many people live in different parts of the United States, leaving the interpretation and use of these counts to Congress and the public.

An exploratory data analysis builds on a descriptive analysis by searching for discoveries, trends, correlations, or relationships between the measurements to generate ideas or hypotheses. The four-star planetary system Tatooine was discovered when amateur astronomers explored public astronomical data from the Kepler telescope (7). An exploratory analysis like this seeks to make discoveries, but can rarely confirm those discoveries. Follow-up studies and additional data were needed to confirm the existence of Tatooine (8).

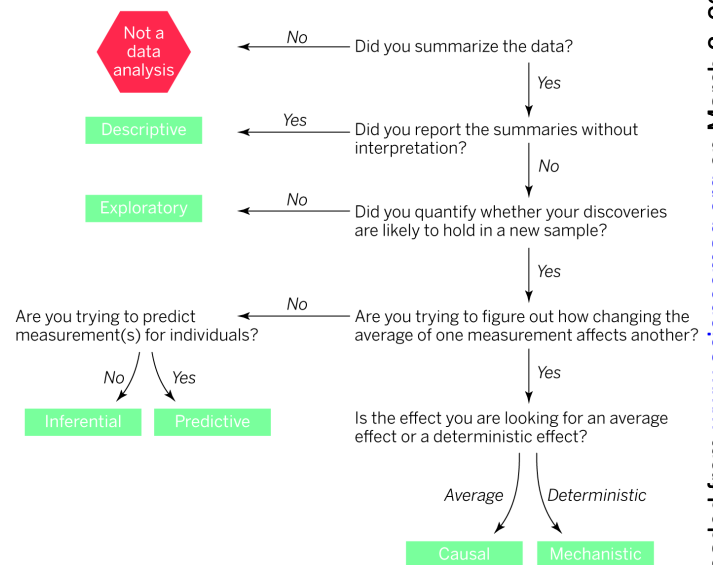
An inferential data analysis quantifies whether an observed pattern will likely hold beyond the data set in hand. This is the most common statistical analysis in the formal scientific literature. An example is a study of whether air pollution correlates with life expectancy at the state level in the United States (9). In nonrandomized experiments, it is usually only possible to determine the existence of a relationship between two measurements, but not the underlying

mechanism or the reason for it.

Going beyond an inferential data analysis, which quantifies the relationships at population scale, a predictive data analysis uses a subset of measurements (the features) to predict another measurement (the outcome) on a single person or unit. Web sites like FiveThirtyEight.com use polling data to predict how people

will vote in an election. Predictive data analyses only show that you can predict one measurement from another; they do not necessarily explain why that choice of prediction works.

Data analysis flowchart



A causal data analysis seeks to find out what happens to one measurement on average if you make another measurement change. Such an analysis identifies both the magnitude and direction of relationships between variables on average. For example, decades of data show a clear causal relationship between smoking and cancer (10). If you smoke, it is certain that your risk of cancer will increase. The causal effect is real, but it affects your average risk.

Finally, a mechanistic data analysis seeks to show that changing one measurement always and exclusively leads to a specific, deterministic behavior in another. For example, data analysis has shown how wing design changes air flow over a wing, leading to decreased drag. Outside of engineering, mechanistic data analysis is extremely challenging and rarely achievable.

Mistakes in the type of data analysis and therefore the conclusions that can be drawn from data are made regularly. In the last 6 months, we have seen inferential analyses of the relationship between cellphones and brain cancer inter-

preted as causal (11) or the exploratory analysis of Google search terms related to flu outbreaks interpreted as a predictive analysis (12). The mistake is so common that it has been codified in standard phrases (see the table).

Common mistakes		
REAL QUESTION TYPE	PERCEIVED QUESTION TYPE	PHRASE DESCRIBING ERROR
Inferential	Causal	"correlation does not imply causation"
Exploratory	Inferential	"data dredging"
Exploratory	Predictive	"overfitting"
Descriptive	Inferential	"n of 1 analysis"

Determining which question is being asked can be even more complicated when multiple analyses are performed in the same study or on the same data set. A key danger is causal creep—for example, when a randomized trial is used to infer causation for a primary analysis and data from secondary analyses are given the same weight. To accurately represent a data analysis, each step in the analysis should be labeled according to its original intent.

Confusion between data analytic question types is central to the ongoing replication crisis, misconstrued press releases describing scientific results, and the controversial claim that most published research findings are false (13, 14). The solution is to ensure that data analytic education is a key component of research training. The most important step in that direction is to know the question.

REFERENCES AND NOTES

1. "How science goes wrong." *The Economist*, 19 October 2013; see www.economist.com/news/leaders/21588069-scientific-research-has-changed-world-now-it-needs-change-itself-how-science-goes-wrong.
2. C. Tomasetti, B. Vogelstein, *Science* **347**, 78 (2015). [Medline doi:10.1126/science.1260825](#)
3. See www.bbc.com/news/magazine-30786970.
4. Duke's Legal Stance: We Did No Harm, *The Cancer Letter Publications* (2015); see www.cancerletter.com/articles/20150123_2.
5. A. Potti et al., *Nat. Med.* **12**, 1294 (2006). [Medline doi:10.1038/nm1491](#)
6. K. A. Baggerly, K. R. Coombes, *Ann. Appl. Stat.* **3**, 1309 (2009). [doi:10.1214/09-AOAS291](#)
7. "Planet with four stars discovered by citizen astronomers," *Wired UK* (2012); see www.wired.co.uk/news/archive/2012-10/15/four-starred-planet.
8. M. E. Schwamb et al., <http://arxiv.org/abs/1210.3612> (2013).
9. A. W. Correia et al., *Epidemiology* **24**, 23 (2013). [Medline doi:10.1097/EDE.0b013e3182770237](#)
10. O. A. Panagiotou et al., *Cancer Res.* **74**, 2157 (2014).
11. E. Oster, Cellphones Do Not Give You Brain Cancer, *FiveThirtyEight* (2015); see <http://fivethirtyeight.com/features/cellphones-do-not-give-you-brain-cancer/>.
12. D. M. Lazer, R. Kennedy, G. King, A. Vespignani, The Parable of Google Flu: Traps in Big Data Analysis (2014); see <http://dash.harvard.edu/handle/1/12016836>.
13. L. R. Jager, J. T. Leek, *Biostatistics* **15**, 1 (2014). [Medline doi:10.1093/biostatistics/kxt007](#)
14. A. Gelman, K. O'Rourke, *Biostatistics* **15**, 18, discussion 39 (2014). [Medline doi:10.1093/biostatistics/kxt034](#)

Published online 26 February 2015
10.1126/science.aaa6146

Opinion: Reproducible research can still be wrong: Adopting a prevention approach

Jeffrey T. Leek^{a,1} and Roger D. Peng^b

^aAssociate Professor of Biostatistics and Oncology and ^bAssociate Professor of Biostatistics, Johns Hopkins University, Baltimore, MD

Reproducibility—the ability to recompute results—and replicability—the chances other experimenters will achieve a consistent result—are two foundational characteristics of successful scientific research. Consistent findings from independent investigators are the primary means by which scientific evidence accumulates for or against a hypothesis. Yet, of late, there has been a crisis of confidence among researchers worried about the rate at which studies are either reproducible or replicable. To maintain the integrity of science research and the public's trust in science, the scientific community must ensure reproducibility and replicability by engaging in a more preventative approach that greatly expands data analysis education and routinely uses software tools.

We define reproducibility as the ability to recompute data analytic results given an observed dataset and knowledge of the data analysis pipeline. The replicability of a study is the chance that an independent experiment targeting the same scientific question will produce a consistent result (1). Concerns among scientists about both have gained significant traction recently due in part to a statistical argument that suggested most published scientific results may be false positives (2). At the same time, there have

been some very public failings of reproducibility across a range of disciplines from cancer genomics (3) to economics (4), and the data for many publications have not been made publicly available, raising doubts about the quality of data analyses. Popular press articles have raised questions about the reproducibility of all scientific research (5), and the US Congress has convened hearings focused on the transparency of scientific research (6). The result is that much of the scientific enterprise has been called into question, putting funding and hard won scientific truths at risk.

From a computational perspective, there are three major components to a reproducible and replicable study: (i) the raw data from the experiment are available, (ii) the statistical code and documentation to reproduce the analysis are available, and (iii) a correct data analysis must be performed. Recent cultural shifts in genomics and other areas have had a positive impact on data and code availability. Journals are starting to require data availability as a condition for publication (7), and centralized databases such as the National Center for Biotechnology Information's Gene Expression Omnibus are being created for depositing data generated by publicly funded scientific experiments. New

computational tools such as knitr, iPython notebook, LONI, and Galaxy (8) have simplified the process of distributing reproducible data analyses.

Unfortunately, the mere reproducibility of computational results is insufficient to address the replication crisis because even a reproducible analysis can suffer from many problems—confounding from omitted variables, poor study design, missing data—that threaten the validity and useful interpretation of the results. Although improving the reproducibility of research may increase the rate at which flawed analyses are uncovered, as recent high-profile examples have demonstrated (4), it does not change the fact that problematic research is conducted in the first place.

The key question we want to answer when seeing the results of any scientific study is “Can I trust this data analysis?” If we think of problematic data analysis as a disease, reproducibility speeds diagnosis and treatment in the form of screening and rejection of poor data analyses by referees, editors, and other scientists in the community (Fig. 1).

This medication approach to research quality relies on peer reviewers and editors to make this diagnosis consistently—which is a tall order. Editors and peer reviewers at medical and scientific journals often lack the training and time to perform a proper evaluation of a data analysis. This problem is compounded by the fact that datasets and data analyses are becoming increasingly complex, the rate of submission to journals continues to increase (9), and the demands on statisticians to referee are increasing. These pressures have reduced the efficacy of peer review in identifying and correcting potential false discoveries in the medical literature. Crucially, the medication approach does not address the problem at its source.

We suggest that the replication crisis needs to be considered from the perspective of



Fig. 1. Peer review and editor evaluation help treat poor data analysis. Education and evidence-based data analysis can be thought of as preventative measures.

Author contributions: J.T.L. and R.D.P. wrote the paper.

¹To whom correspondence should be addressed. Email: jtleek@jhu.edu.

Any opinions, findings, conclusions, or recommendations expressed in this work are those of the authors and do not necessarily reflect the views of the National Academy of Sciences.

primary prevention. If we can prevent problematic data analyses from being conducted, we can substantially reduce the burden on the community of having to evaluate an increasingly heterogeneous and complex population of studies and research findings. The best way to prevent poor data analysis in the scientific literature is to (i) increase the number of trained data analysts in the scientific community and (ii) identify statistical software and tools that can be shown to improve reproducibility and replicability of studies.

How can we dramatically scale up data science education in the short term? One approach that we have taken is through massive online open courses (MOOCs). The Johns Hopkins Data Science Specialization (jhudatascience.org) is a sequence of nine courses covering the full spectrum of data science skills from formulating quantitative questions, to cleaning data, to statistical analysis and producing reproducible reports. Thus far, we have enrolled more than 1.5 million students in this Specialization. A complementary approach is crowd-sourced short courses such as Data and Software Carpentry (software-carpentry.org) that have addressed the extreme demand for data science knowledge on a smaller scale.

However, simply increasing data analytic literacy comes at a cost. Most scientists in these programs will receive basic to moderate training in data analysis, creating the potential for producing individuals with

enough skill to perform data analysis but without enough knowledge to prevent mistakes.

To improve the global robustness of scientific data analysis, we must couple education efforts with the identification of data analytic strategies that are most reproducible and replicable in the hands of basic or intermediate data analysts. Statisticians must bring to bear their history of developing rigorous methods to the area of data science.

A fundamental component of scaling up data science education is performing empirical studies to identify statistical methods, analysis protocols, and software that lead to increased replicability and reproducibility in the hands of users with basic knowledge. We call this approach evidence-based data analysis. Just as evidence-based medicine

applies the scientific method to the practice of medicine, evidence-based data analysis applies the scientific method to the practice of data analysis. Combining massive scale education with evidence-based data analysis can allow us to quickly test data analytic practices in a population most at risk for data analytic mistakes (10).

In much the same way that epidemiologist John Snow ended a London cholera epidemic by removing a pump handle to make contaminated water unavailable, we have an opportunity to attack the crisis of scientific reproducibility at its source. Dramatic increases in data science education, coupled with robust evidence-based data analysis practices, have the potential to prevent problems with reproducibility and replication before they can cause permanent damage to the credibility of science.

1 Peng RD (2011) Reproducible research in computational science. *Science (New York)* 6060:1226–1227.

2 Ioannidis JP (2005) Why most published research findings are false. *PLoS Med* 2(8):e124.

3 Baggerly KA, Coombes KR (2009) Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology. *Ann Appl Stat* 3(4):1309–1334.

4 Herndon T, Ash M, Pollin R (2014) Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. *Camb J Econ* 38(322):257–279.

5 Marcus G, Davis E (2014) Eight (no, nine!) problems with big data. Available at www.nytimes.com/2014/04/07/opinion/eight-no-nine-problems-with-big-data.html?_r=0. Accessed November 19, 2014.

6 Alberts B, Stodden V, Young S, Choudhury S (2013) Testimony on scientific integrity & transparency. Available at www.science.

house.gov/hearing/subcommittee-research-scientific-integrity-transparency. Accessed November 19, 2014.

7 Bloom T (2014) PLOS's new data policy: Part two. Available at blogs.plos.org/everyone/2014/02/24/plos-new-data-policy-public-access-data-2/. Accessed November 19, 2014.

8 Goecks J, Nekrutenko A, Taylor J; Galaxy Team (2010) Galaxy: A comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* 11(8):R86.

9 Jager LR, Leek JT (2014) An estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics* 15(1):1–12.

10 Fisher A, Anderson GB, Peng R, Leek J (2014) A randomized trial in a massive online open course shows people don't know what a statistically significant relationship looks like, but they can learn. *PeerJ* 2:e589.