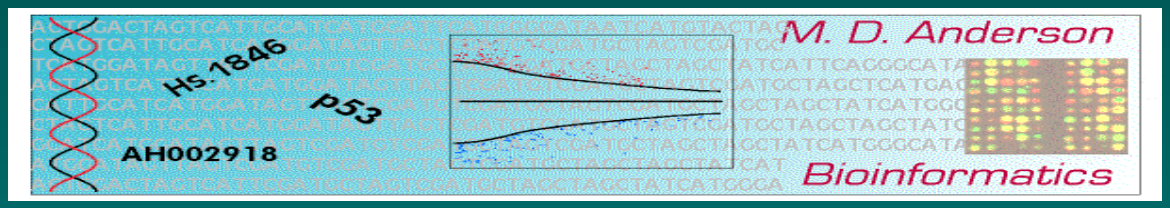
The Importance of Reproducibility in High-Throughput Biology: Case Studies in Forensic Bioinformatics

Keith A. Baggerly
Bioinformatics and Computational Biology
UT M. D. Anderson Cancer Center
kabagg@mdanderson.org

UNMC COPH Grand Rounds, Feb 15, 2012



Why is Reproducibility Important in H-T-B?

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 Cell Lines to Predict Sensitivity

ture.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 Chan[4], Robyn Sayer[4], Janiel Cragun[4], Hope Cottrill[4], Michael J Kelley[2], Rebecca Petersen[5], David Harpole[5], Jeffrey Marks[5], Andrew Berchuck[1,6], Geoffrey S Ginsburg[1,2], Phillip Febbo[1–3], Johnathan Lancaster[4] & Joseph R Nevins[1–3]

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.

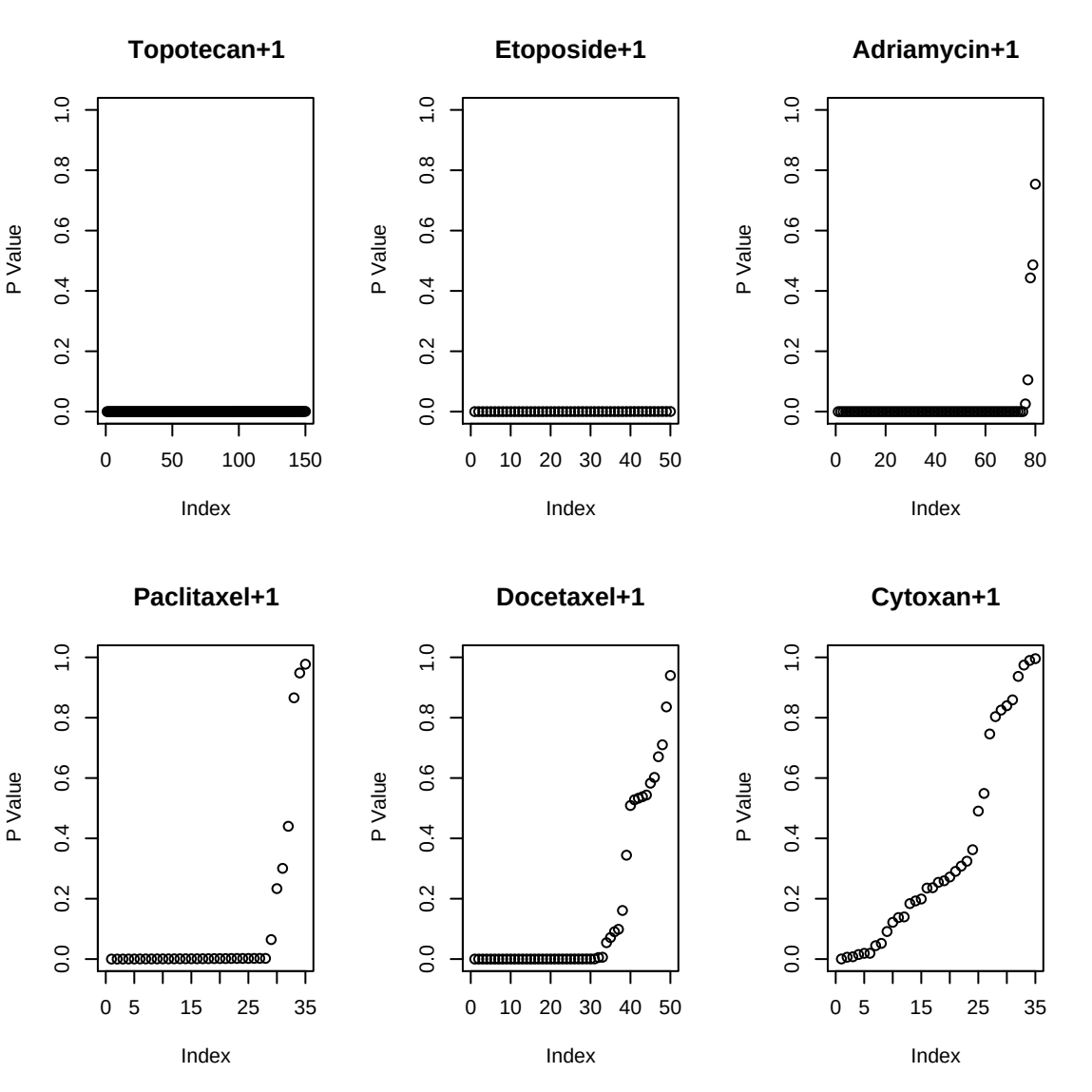
They provide examples using 7 commonly used agents.

This got people at MDA very excited.

Their Gene List and Ours

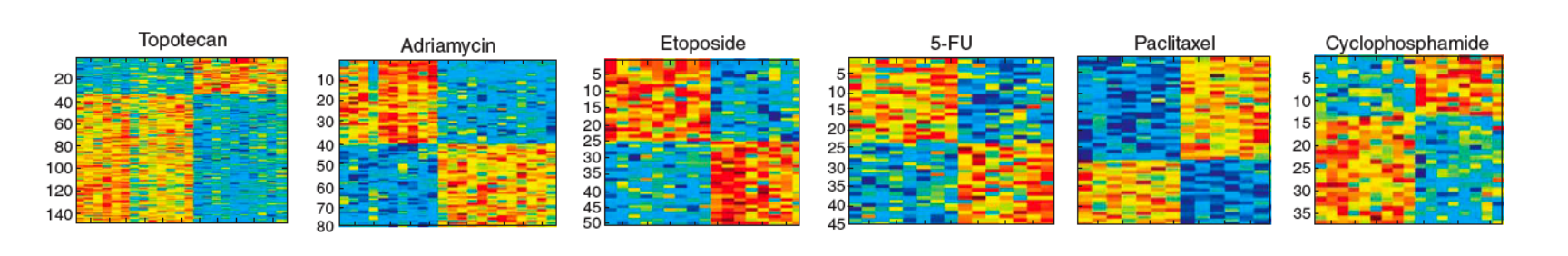
```
> temp <- cbind(
    sort(rownames(pottiUpdated)[fuRows]),
    sort(rownames(pottiUpdated)[
         fuTQNorm@p.values <= fuCut]);
> colnames(temp) <- c("Theirs", "Ours");
> temp
     Theirs          Ours
...
[3,] "1881_at"       "1882_g_at"
[4,] "31321_at"      "31322_at"
[5,] "31725_s_at"    "31726_at"
[6,] "32307_r_at"    "32308_r_at"
...
```

Offset P-Values: Other Drugs

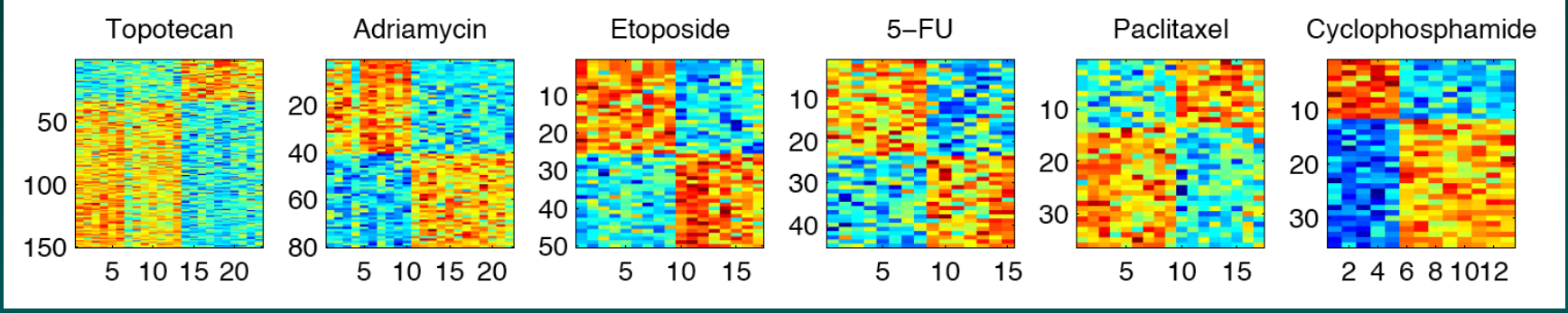


Heatmaps Match Exactly for Most Drugs

From the paper:



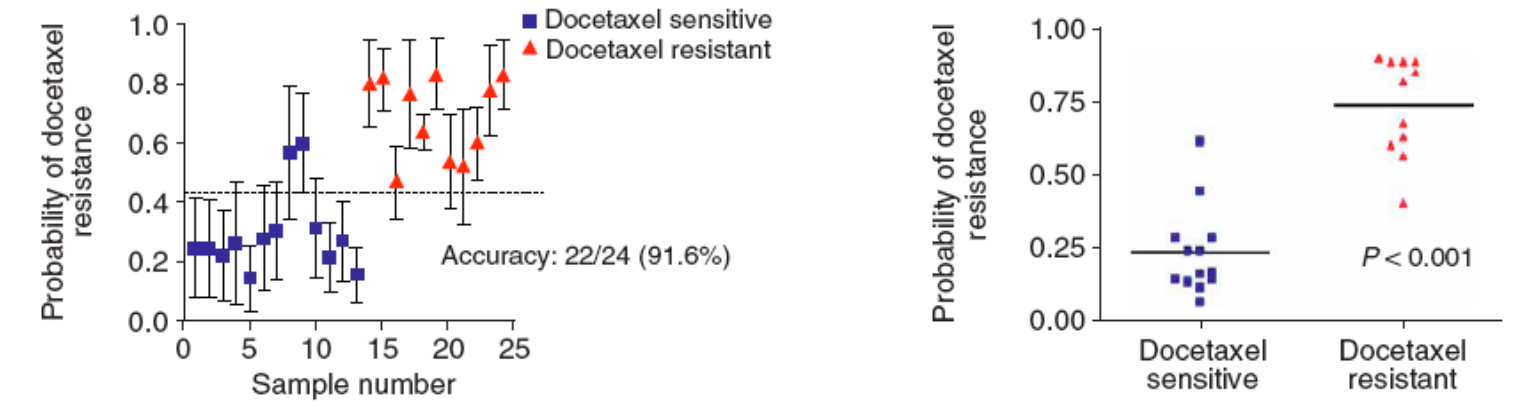
From the software:



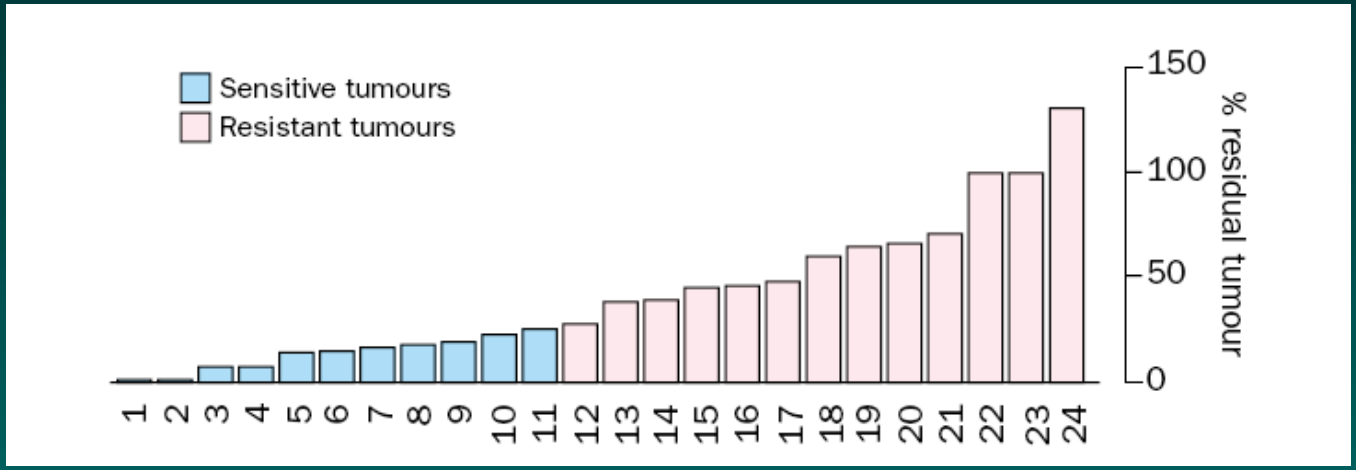
We match heatmaps but not gene lists?

Their software also gives *predictions*.

Predicting Docetaxel Response

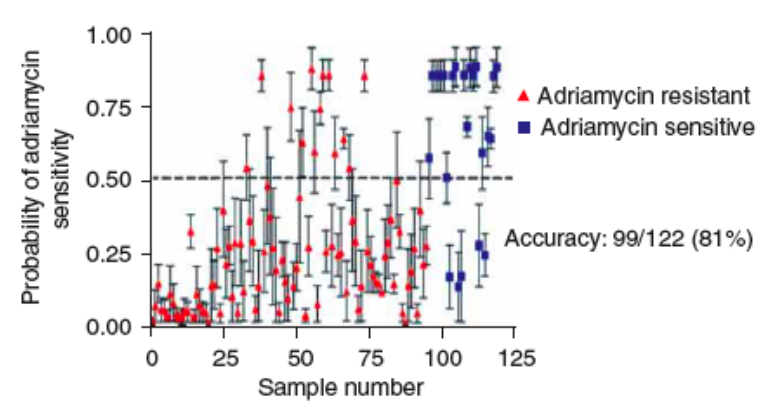


Potti et al, Nat Med 2006, 12:1294-300, Fig 1d

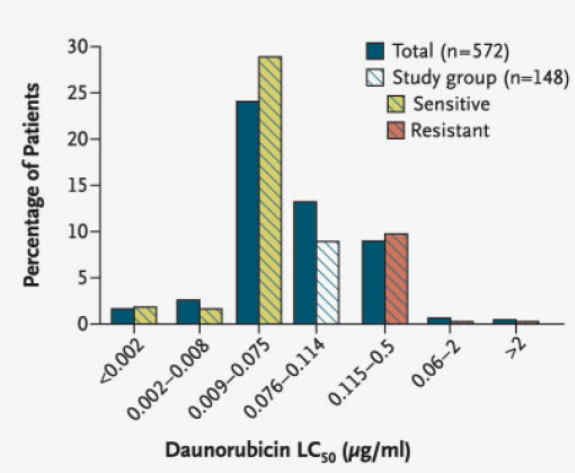


Chang et al, Lancet 2003, 362:362-9, Fig 2 top

Predicting Adriamycin Response

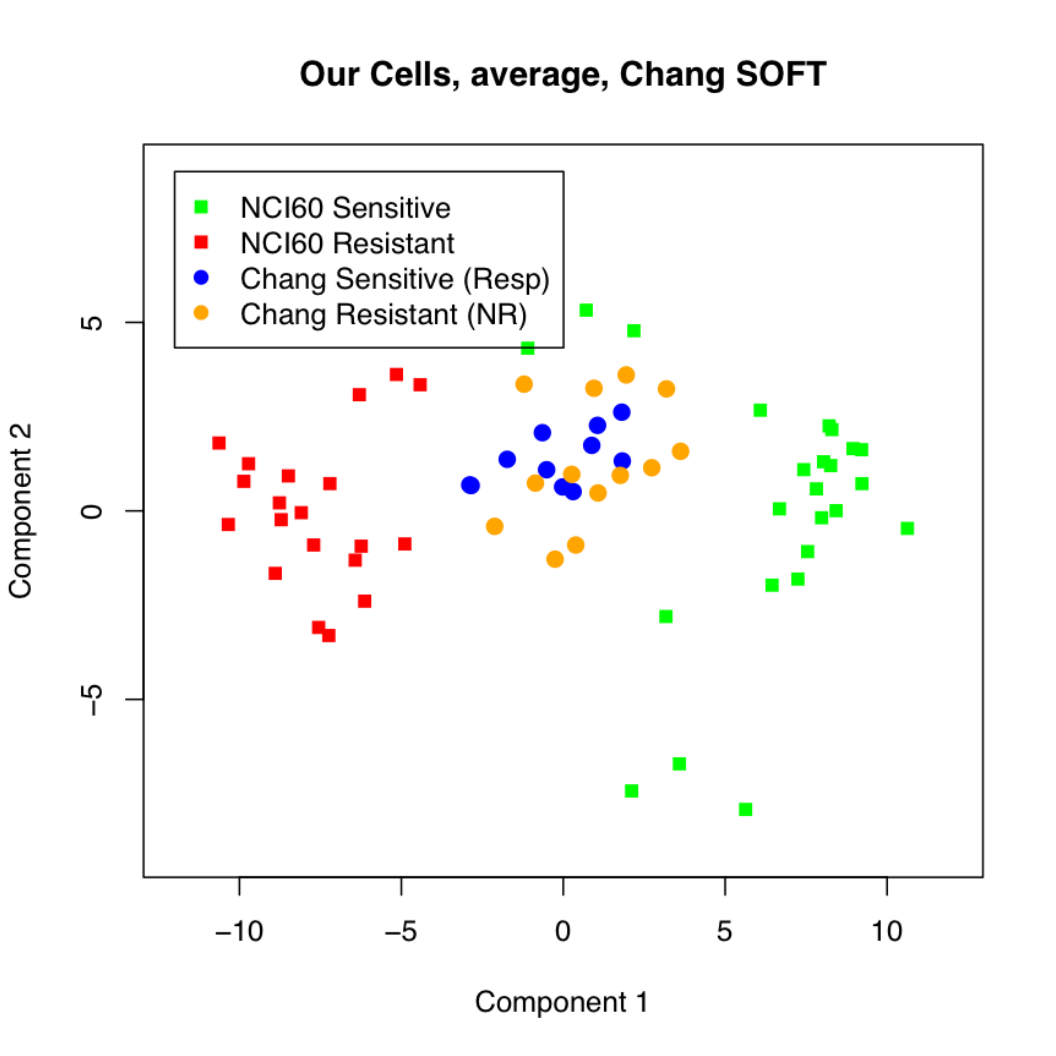


Potti et al, Nat Med 2006, 12:1294-300, Fig 2c



Holleman et al, NEJM 2004, 351:533-42, Fig 1

Trying it Ourselves



When we try it, *it doesn't work.*

Partial Timeline

2006:

* Nov 8: Our first questions to Potti and Nevins.
* Nov 21: Our first report describing errors.
* Nov-Dec: More reports/questions: Nov 27, Dec 4, 13, 27.

2007:

* Jan 24: We meet with Nevins at M.D. Anderson. We urge him to review the data.
* Feb-Apr: New data and code are posted. Some numbers change. We tell them we don't think it works.
* Apr 25: We send Potti and Nevins a draft for comment.
* May: We find problems with outliers. Potti and Nevins continue to insist it works, and want to "bring this to a close".

A Repro 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.

Prompted by this example, we imposed new rules for writing reports in our department.

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

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

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

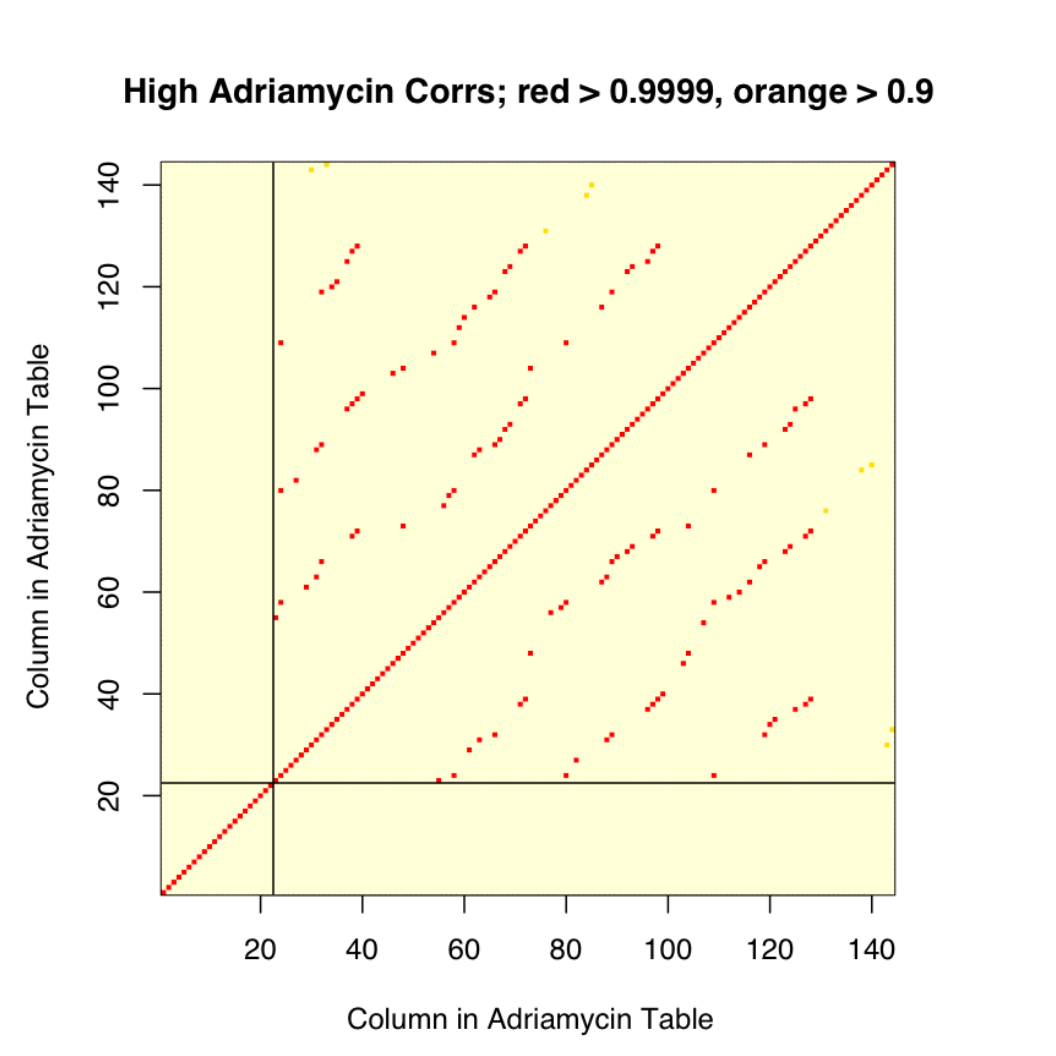
Pharmacogenomic Strategies Provide a Rational Approach to the Treatment of Cisplatin-Resistant Patients With Advanced Cancer

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)



Redone Aug 08, “using ... 95 unique samples”.

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).

Another problem –

The last two probesets aren't on the U133A arrays that were used. They're on the U133B.

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

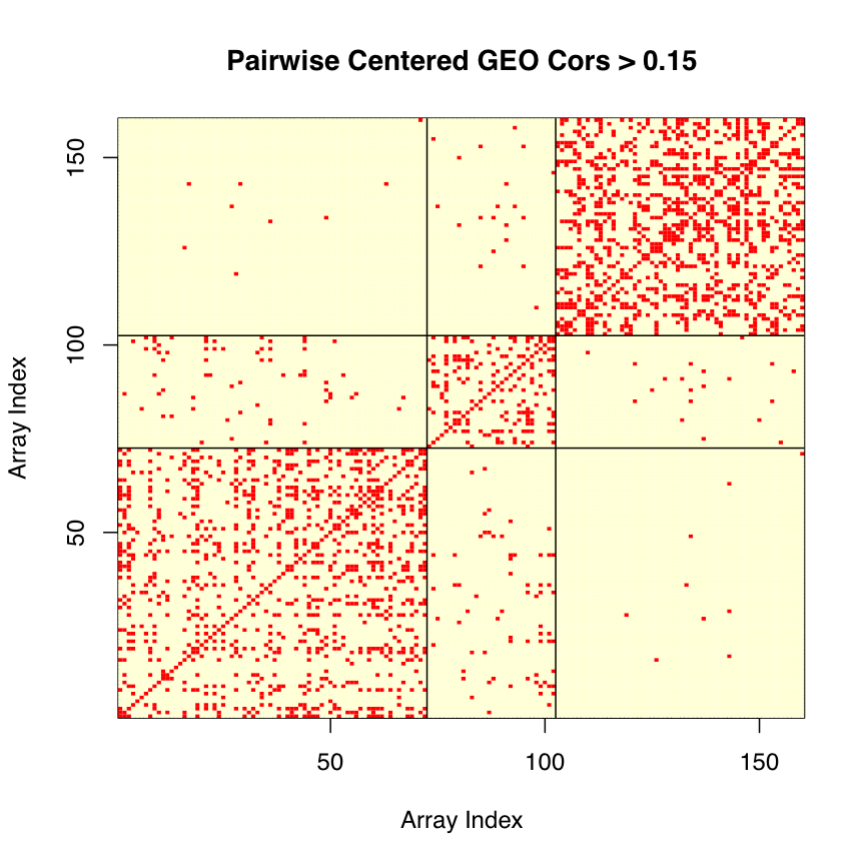
Hervé Bonnefoi, Anil Potti, Mauro Delorenzi, Louis Mauriac, Mario Campone, Michèle Tubiana-Hulin, Thierry Petit, Philippe Rouanet, Jacek Jassem, Emmanuel Blot, Véronique Becette, Pierre Farmer, Sylvie André, Chaitanya R Acharya, Sayan Mukherjee, David Cameron, Jonas Bergh, Joseph R Nevins, Richard D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

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

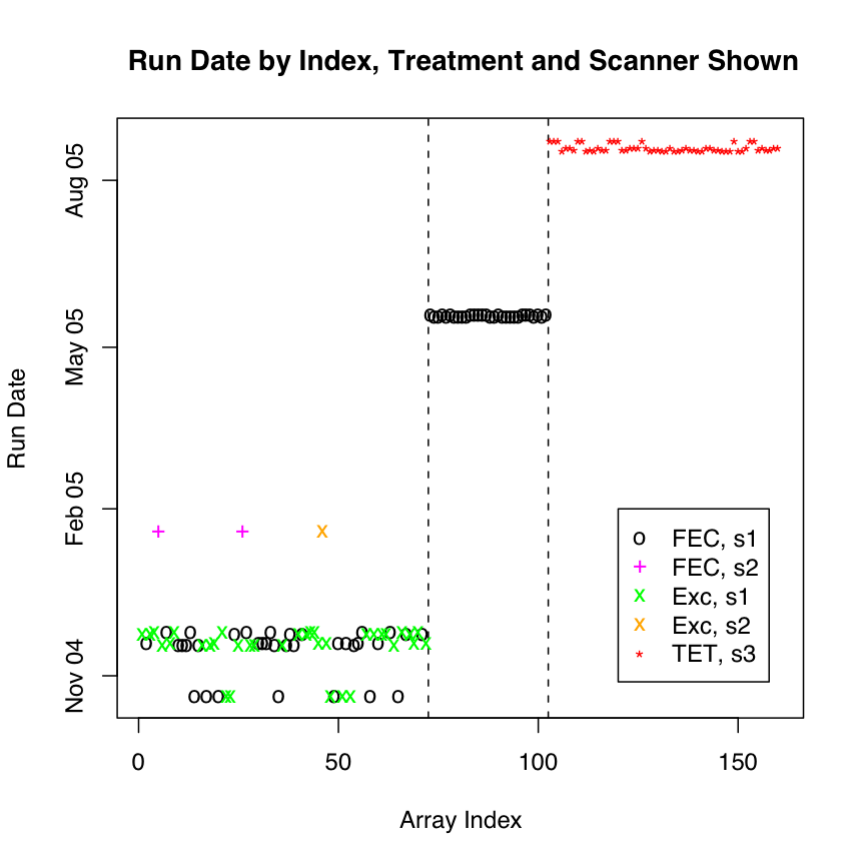
Similar approach, using signatures for Fluorouracil, Epirubicin, Cyclophosphamide, and Taxotere to predict response to combination therapies: FEC and TET.

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

Experimental Design Matters



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.

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.

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

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.

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

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

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.

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

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

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

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.

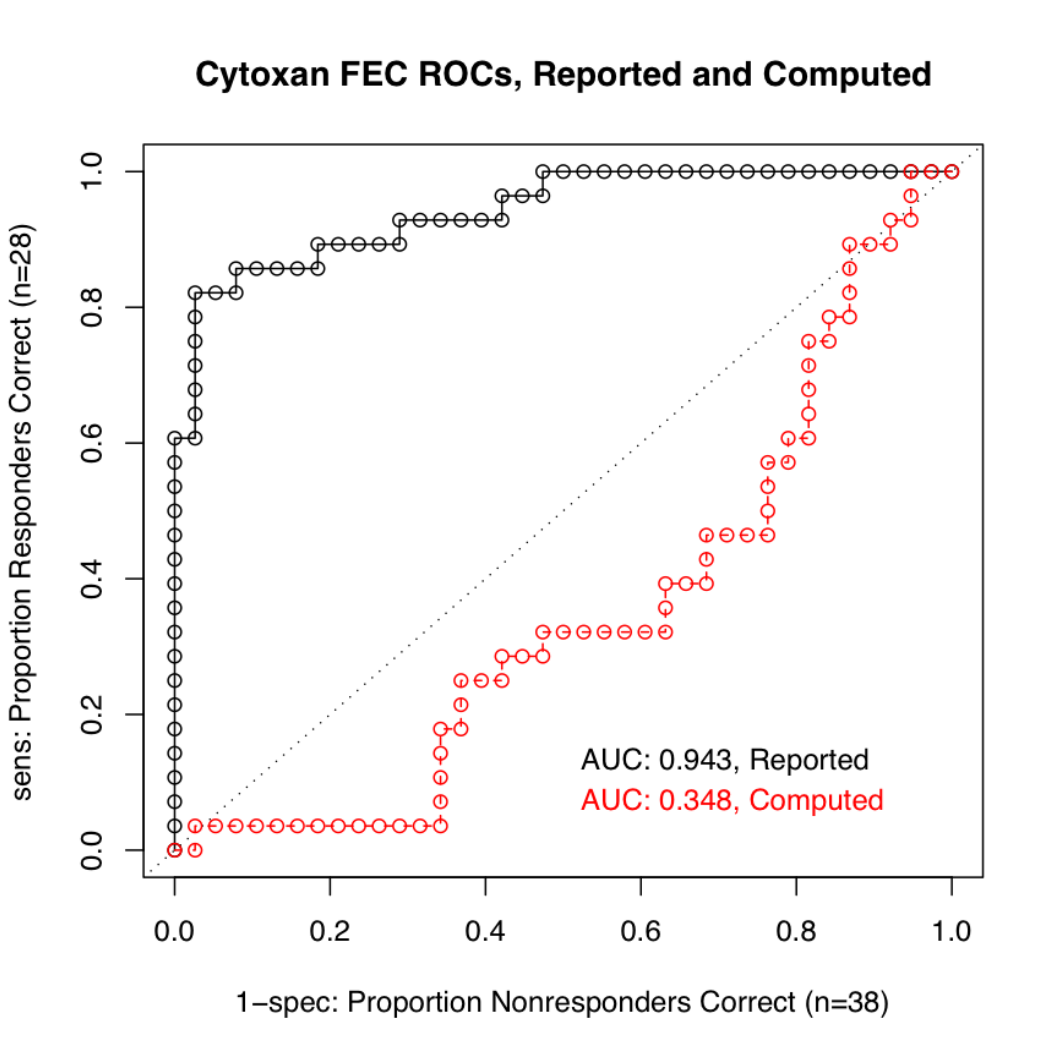
$$P(TFAC) = P(T)+P(F)+P(A)+P(C)-P(T)P(F)P(A)P(C).$$

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

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

Each rule is different.

Predictions for Individual Drugs?



Does cytoxan make sense?

The Story Takes A Darker Turn

Jun 2009: we learn clinical trials had begun.

2007: pemetrexed vs cisplatin, pem vs vinorelbine.

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

The Story Takes A Darker Turn

Jun 2009: we learn clinical trials had begun.
2007: pemetrexed vs cisplatin, pem vs vinorelbine.
2008: docetaxel vs doxorubicin, topotecan vs dox (Moffitt).

Sep 1, 2009: We submit a paper describing case studies to the *Annals of Applied Statistics*.

Sep 14, 2009: Paper accepted and available online at the *Annals of Applied Statistics*.

Sep-Oct 2009: Story covered by *The Cancer Letter*. NCI raises concerns with Duke's IRB behind the scenes. Duke starts internal investigation, suspends trials.

Are Our Objections Moot?

"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.*

Are Our Objections Moot?

"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.

"we would not be able to reproduce the reported probabilities with the information we have about how they were obtained." – *Mauro Delorenzi, Oct 23.*

Jan 29, 2010

THE CANCER LETTER

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."*

We Asked for the Data

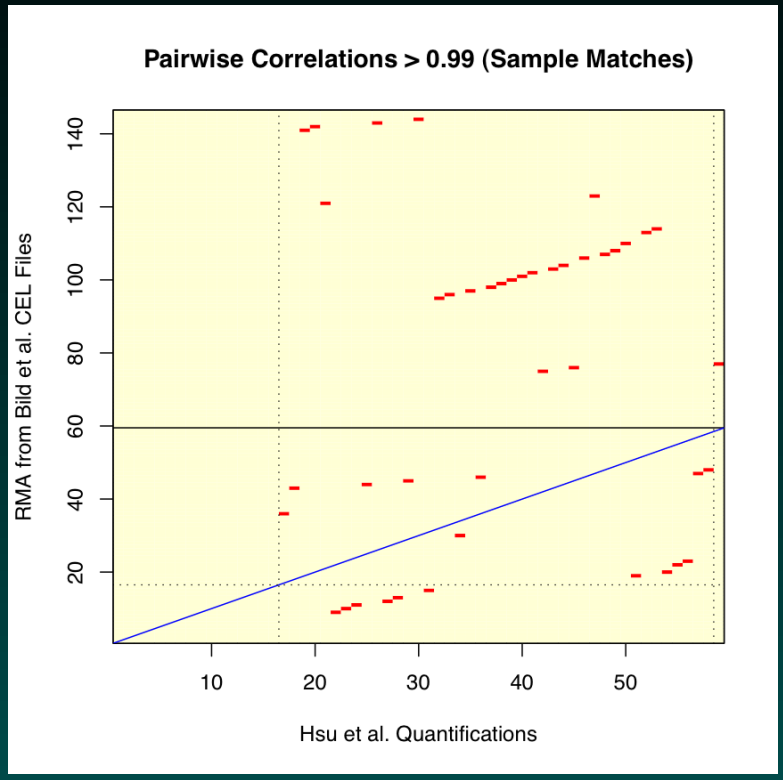
"While the reviewers approved of our sharing the report with the NCI, *we consider it a confidential document*" (Duke). A *future paper* will explain the methods.

There was, however, a major new development the restart announcement didn't mention.

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



43 samples are mislabeled.

16 samples don't match because the genes are mislabeled.

All of the validation data are wrong.

A Catalyzing Event: July 16, 2010

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

Jul 19/20: Letter to Varmus; Duke resuspends trials.
Oct 22/9: First call for paper retraction.
Nov 9: Duke terminates trials.
Nov 19: call for Nat Med retraction, Potti resigns

Dec 20, 2010: the NCI Speaks

Sep 2009: Our paper received. Similar problems noted with CALGB 30702 application. Concerns sent to Duke IRB.

Nov 2009-Mar 2010: Data underlying the Lung Metagene Score (LMS) used in CALGB 30506 reexamined. Signature found unjustified and unstable. LMS pulled from trial.

April 2010: NCI learns it is partially funding NCT00509366. Data, code immediately requested.

May 2010: Problems found with cisplatin, pemetrexed signatures.

June 29, 2010: Duke team visits NCI. NCI directs that search for data justifying trials be conducted.

The IOM Reviews

Dec 20, 2010: NCI, FDA Presentations.

Mar 30-1, 2011: Case Studies. Joe Nevins presents.
I present. Duke historical document supplied.
Details clarify what happened with our Nov 2009 report.

Jun 30, 2011: NCI Presentation.

Aug 22, 2011: Duke Institutional Response.

Nov 4, 2011: Moffitt trial in *The Cancer Letter*.

Links to MP3 audio, documents, our annotations:

http://bioinformatics.mdanderson.org/Supplements/ReproRsch-All/Modified/index.html

Other Developments

Sep, 2011: Patient lawsuits filed.

Dec, 2011: NC Medical Board: reprimand, some settlements.

Dec, 2011: SC Medical Board: Potti application letter.

Jan, 2012: Acharya et al JAMA, Dressman et al JCO papers retracted.

Wikipedia

Online reputation manager

Press coverage

Feb 12, 2012: 60 Minutes.
http://www.cbsnews.com/8301-18560_162-57376073/deception-at-duke/

The Story's Not Over Yet...

For now, we have a soapbox.

This is not an unmixed blessing.

What can we learn, and how are things moving forward?

Some Cautions/Observations/Lessons

This case is pathological.

But we've seen similar problems before.

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

What Should the Reproducibility Norm Be?

In our group we've prepared reports in *Sweave* since 2007.

For papers? (Baggerly + lots, *Nature*, Sep 22, 2010)

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?

Reasons for Hope

1. Our Own Experience
2. Duke's Trans Med Qual Framework (TMQF) Team
3. The NCI and Trials it Funds
4. The IOM, the FDA, and IDEs
5. Journals, Code and Data

Acknowledgements

Kevin Coombes

Shannon Neeley, Jing Wang
David Ransohoff, Gordon Mills
Jane Fridlyand, Lajos Pusztai, Zoltan Szallasi

M.D. Anderson Ovarian, Lung and Breast SPOREs

Baggerly and Coombes (2009), *Annals of Applied Statistics*, **3(4)**:1309-34. http://bioinformatics.mdanderson.org/Supplements/ReproRsch-All

For updates: http://bioinformatics.mdanderson.org/Supplements/ReproRsch-All/Modified.

Index