

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

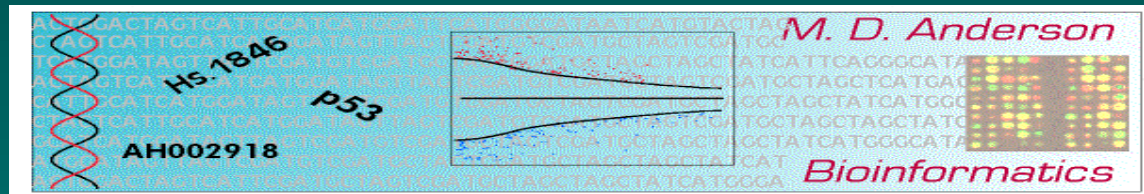
Keith A. Baggerly

Bioinformatics and Computational Biology

UT M. D. Anderson Cancer Center

kabagg@mdanderson.org

CSE, May 3, 2011



Why is Reproducibility Important in H-T-S?

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

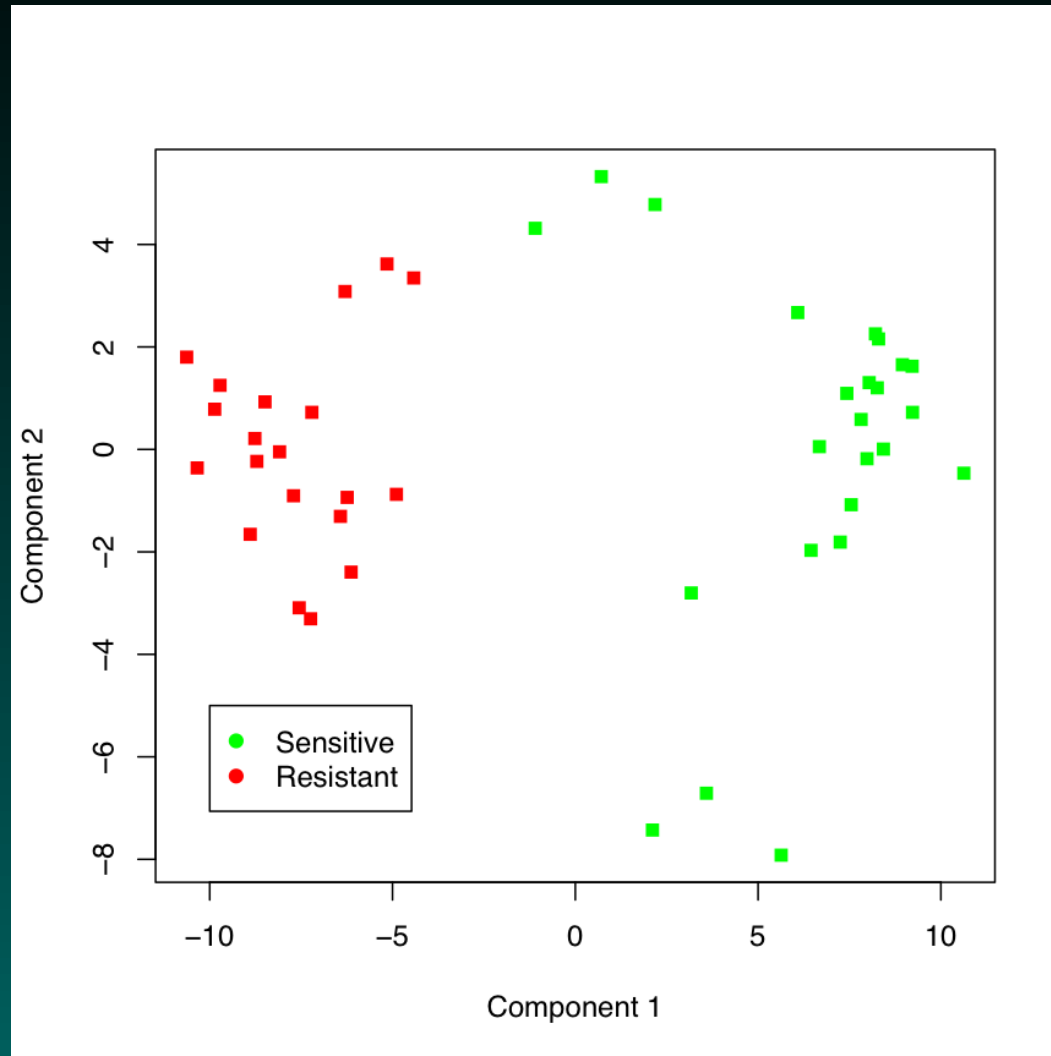
Potti et al (2006), Nature Medicine, 12:1294-1300.

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.

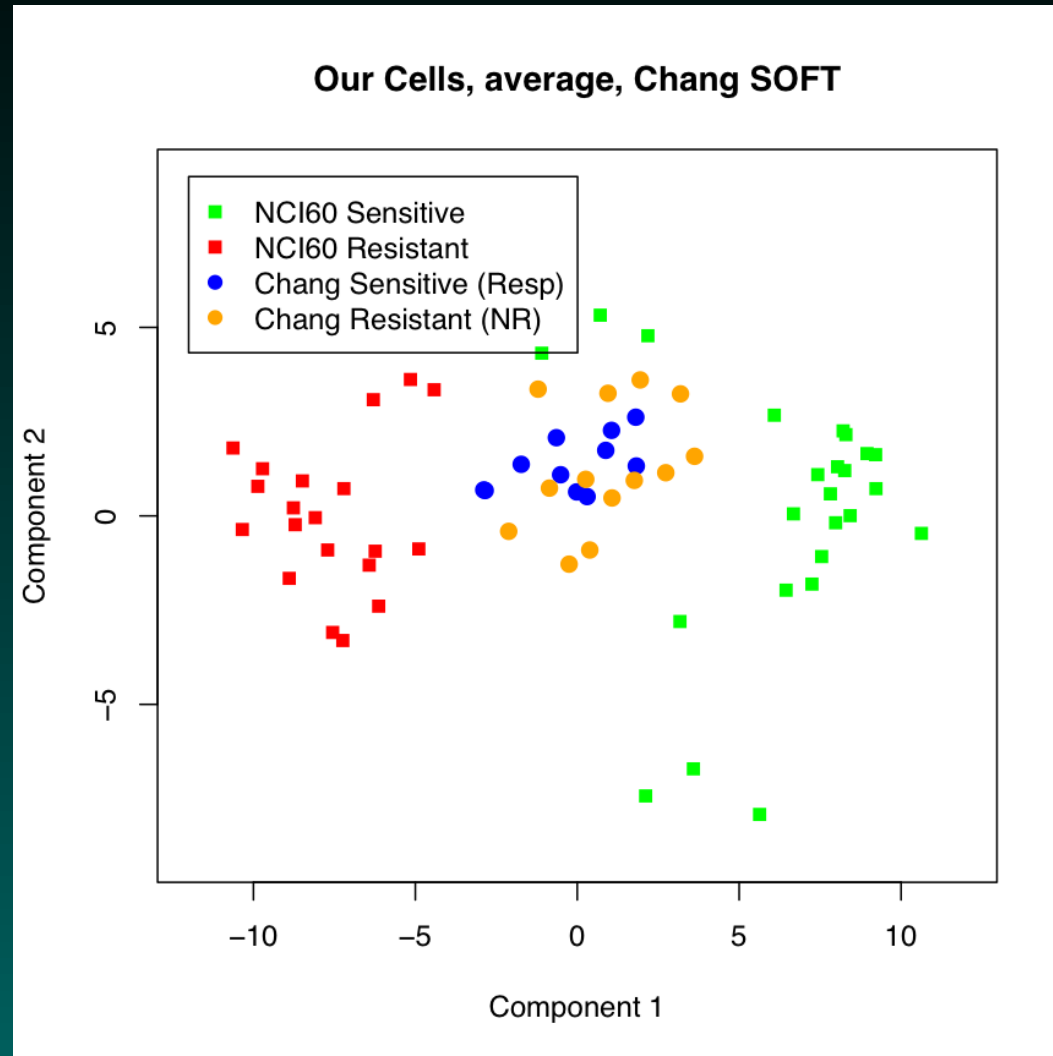
This got people at MDA very excited.

Fit Training Data



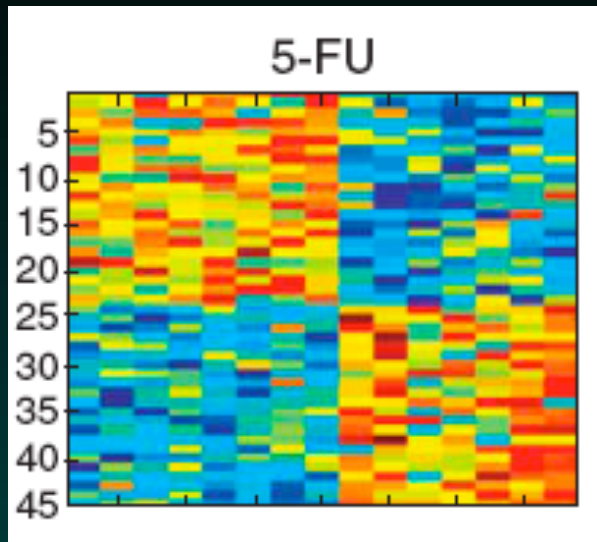
We want the test data to split like this...

Fit Testing Data



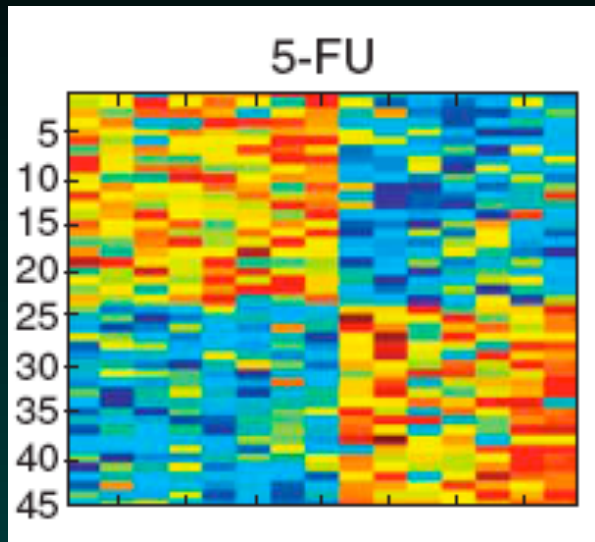
But it *doesn't*. Did we do something wrong?

5-FU Heatmaps

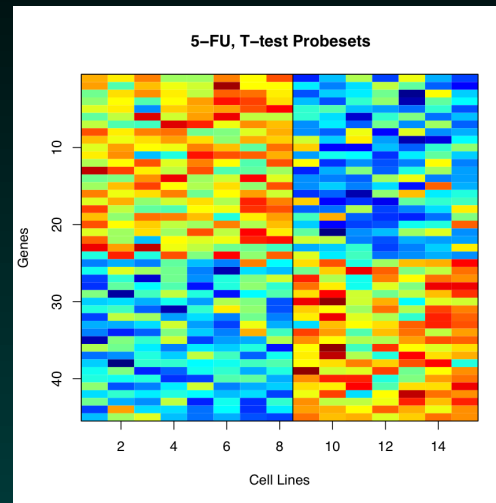


Nat Med Paper

5-FU Heatmaps

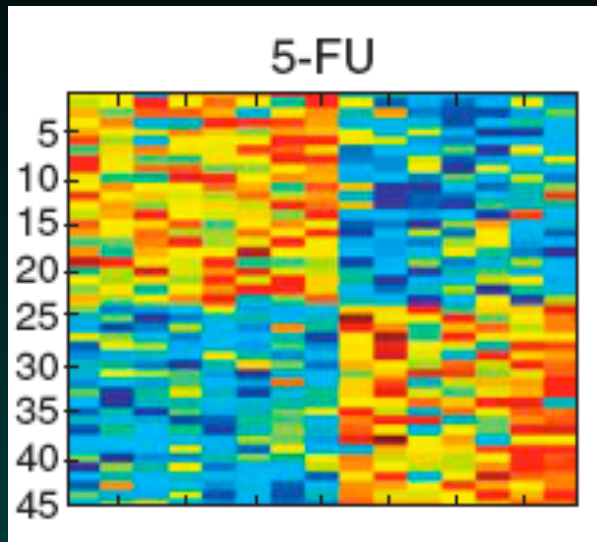


Nat Med Paper

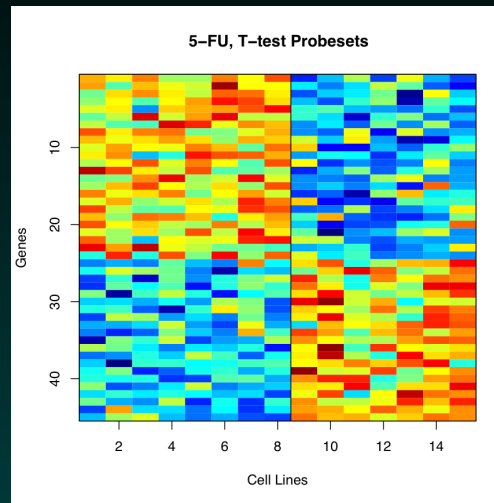


Our t-tests

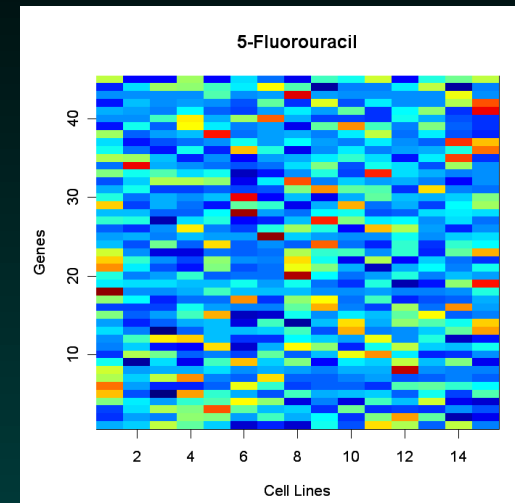
5-FU Heatmaps



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"

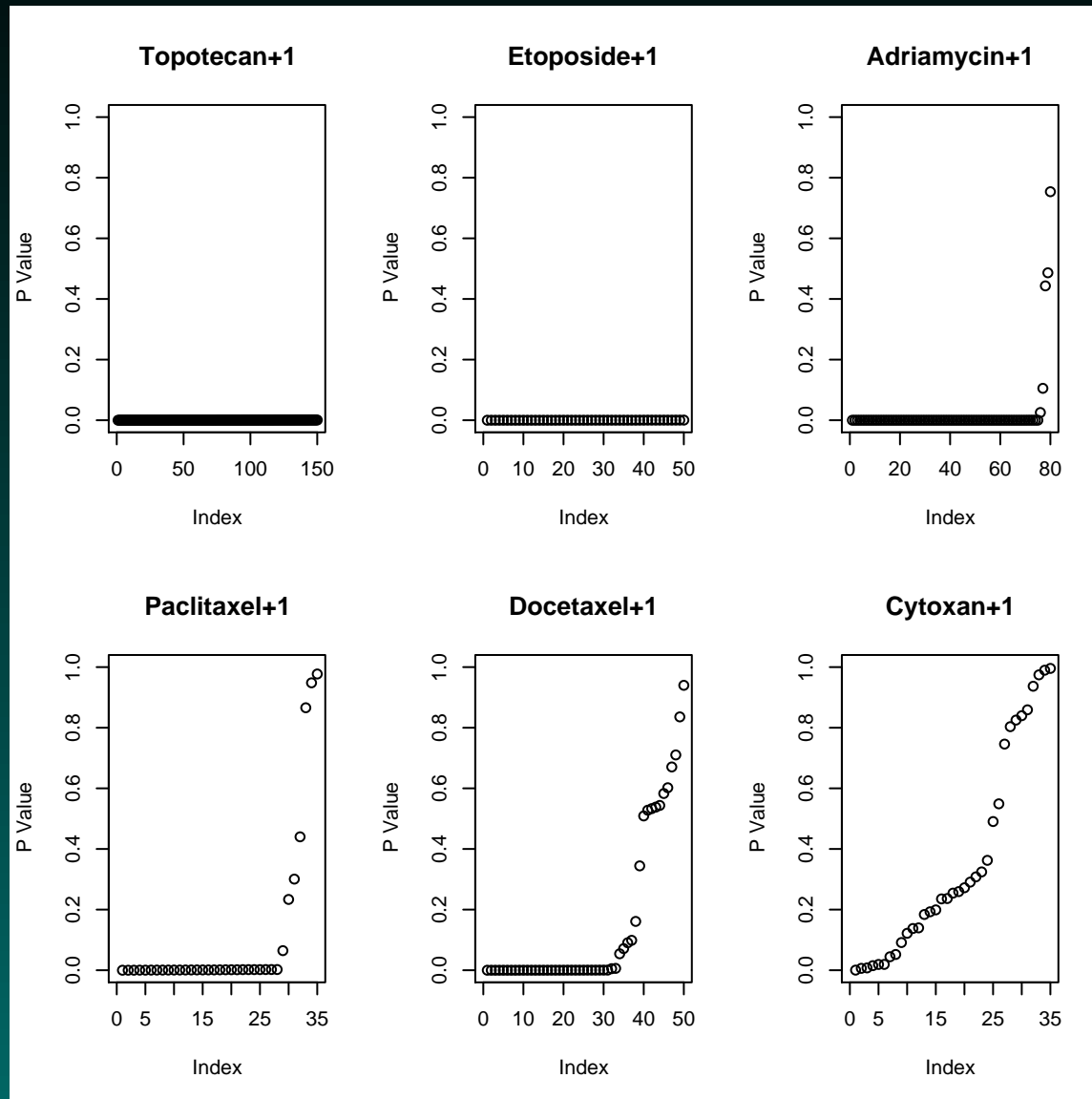
[4,] "31321_at" "31322_at"

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

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

...

Offset P-Values: Other Drugs



Using Their Software

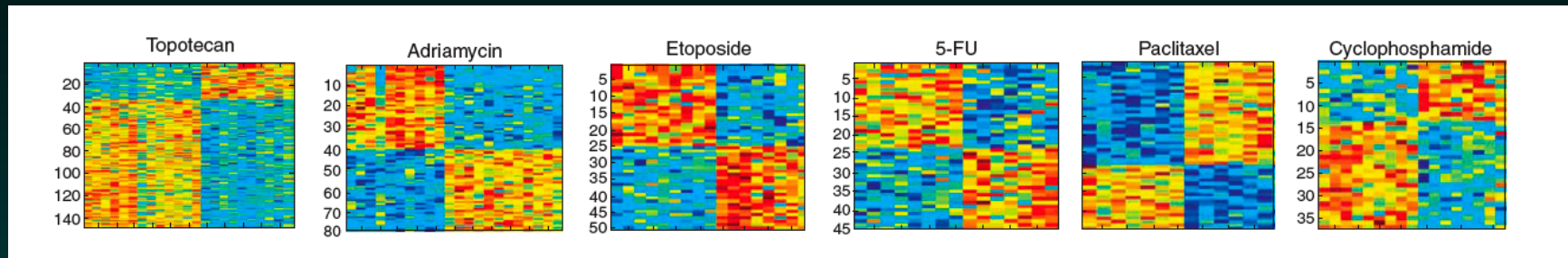
Their software requires two input files:

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.***

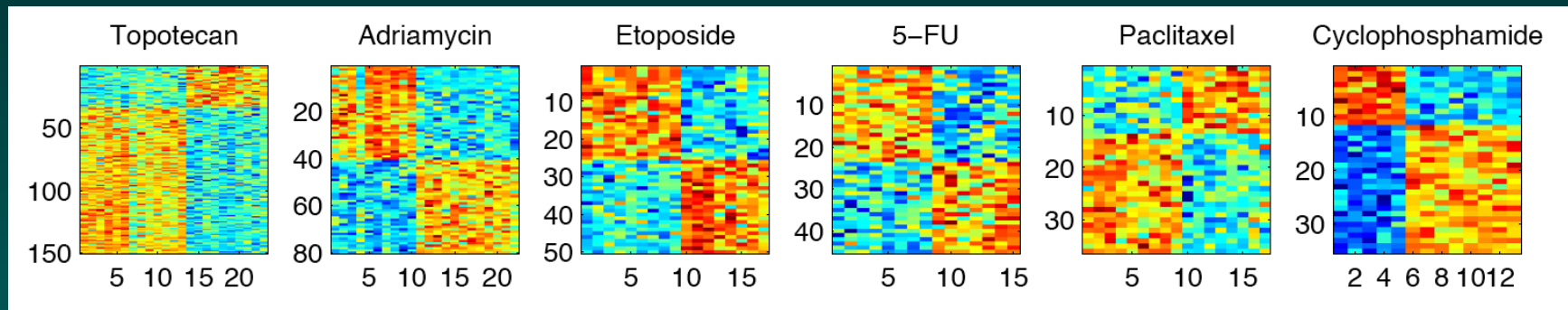
What do we get?

Heatmaps Match Exactly for Most Drugs!

From the **paper**:

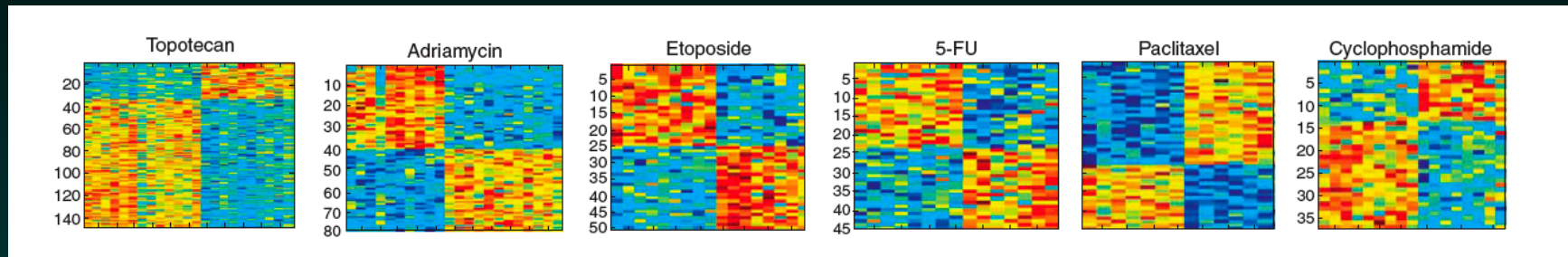


From the **software**:

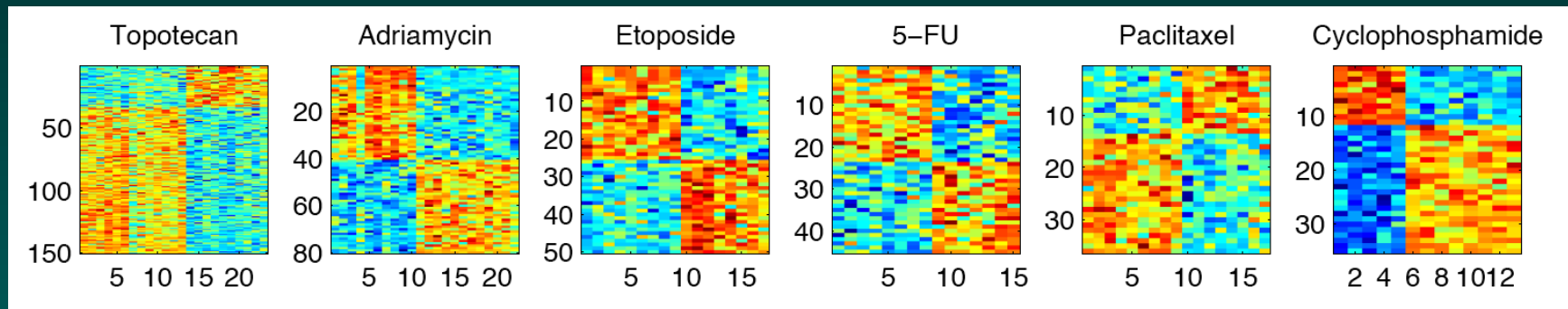


Heatmaps Match Exactly for Most Drugs!

From the **paper**:

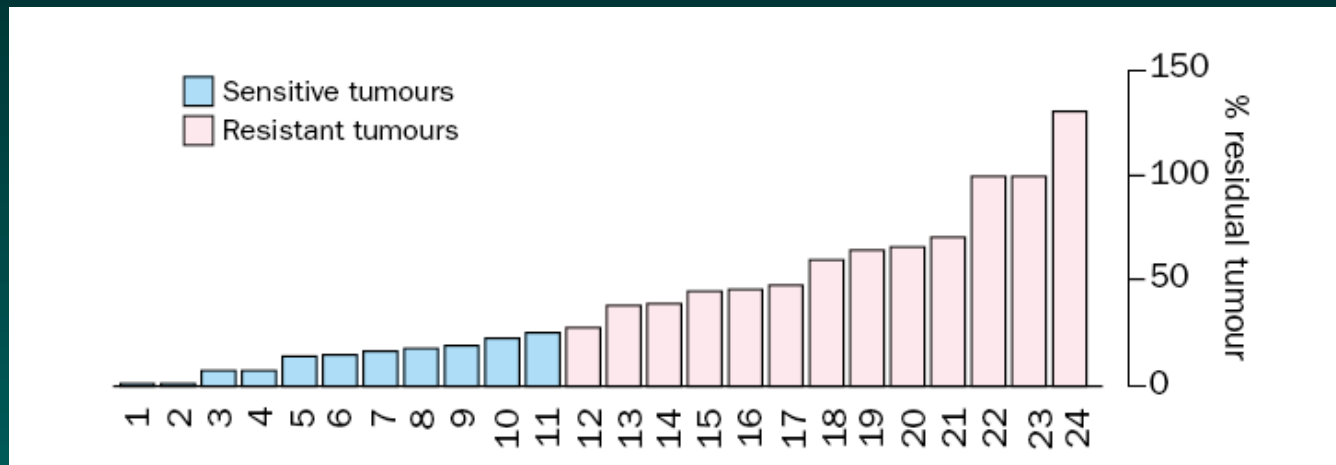
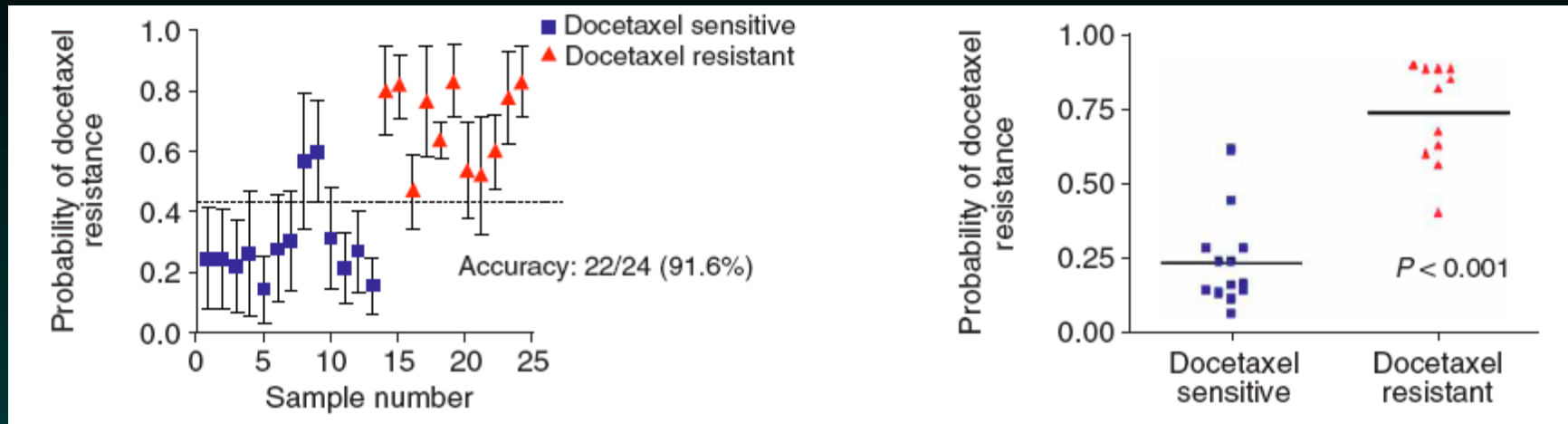


From the **software**:

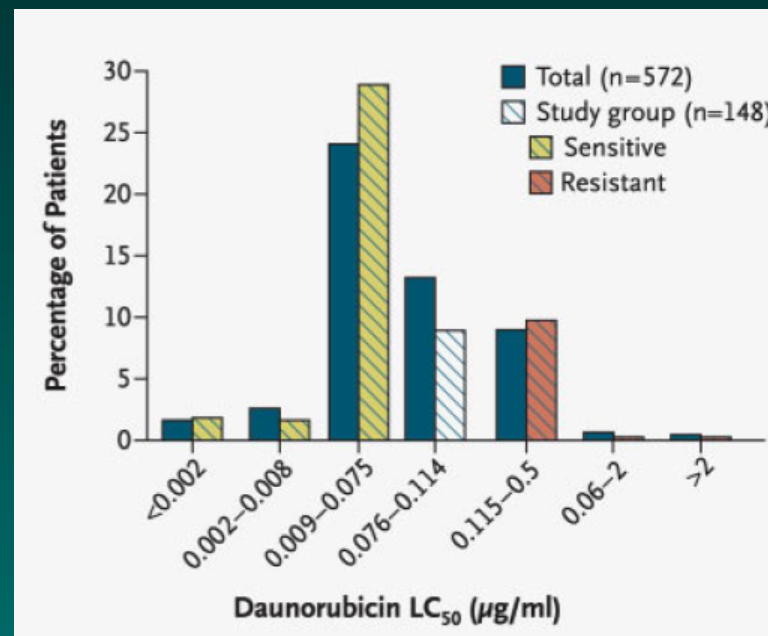
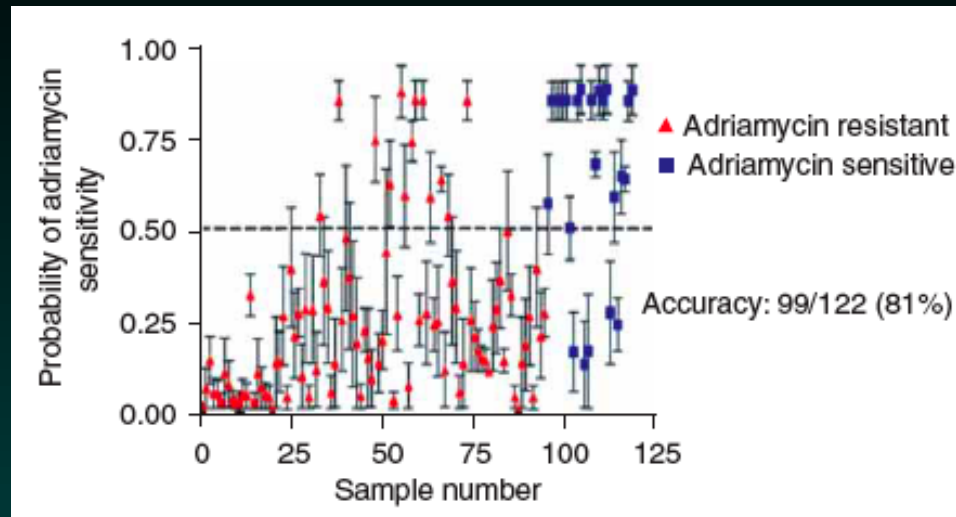


We match heatmaps but not gene lists? We'll come back to this, because their software also gives *predictions*.

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.

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](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!)

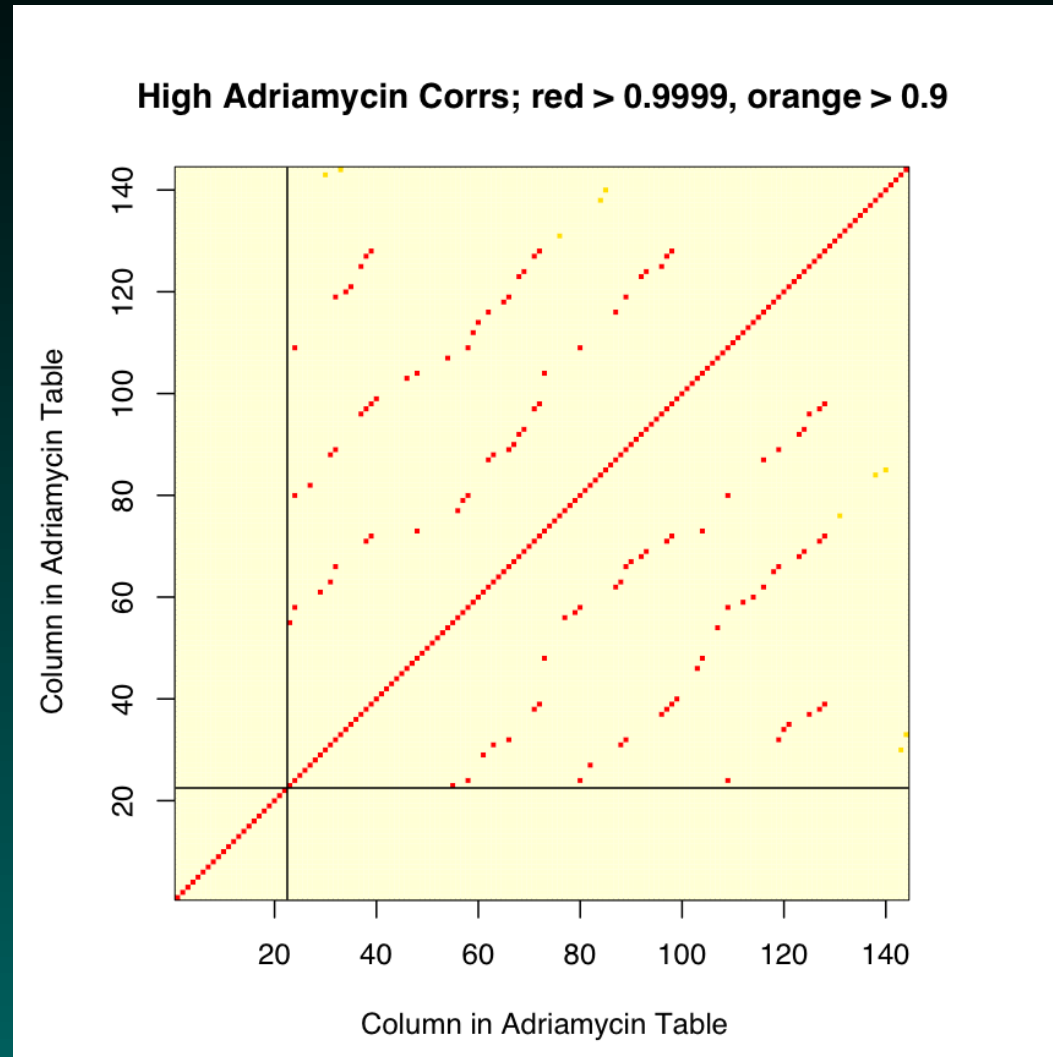
Pharmacogenomic Strategies Provide a Rational Approach to the Treatment of Cisplatin-Resistant Patients With Advanced Cancer

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

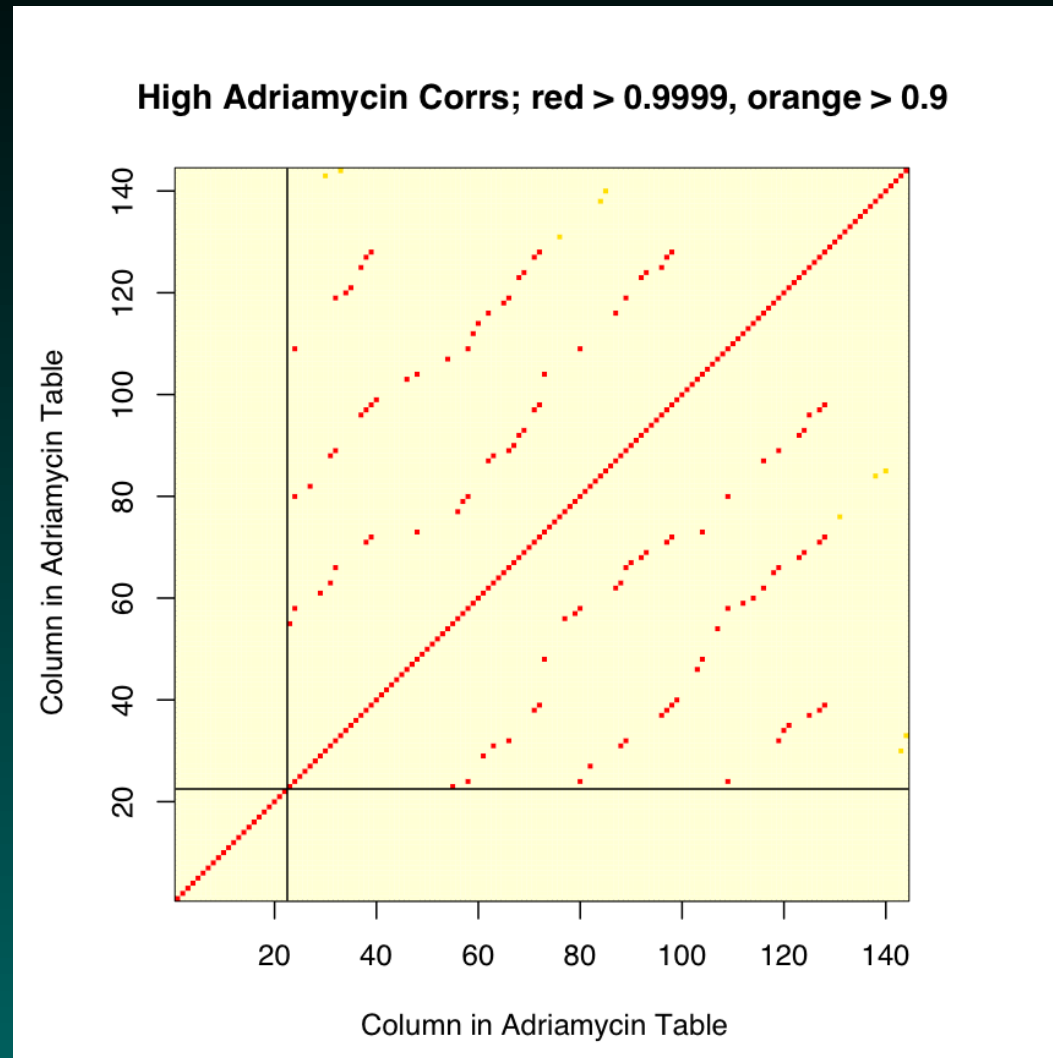
Validation of gene signatures that predict the response of breast cancer to neoadjuvant chemotherapy: a substudy of the EORTC 10994/BIG 00-01 clinical trial

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 Aug 08, “using ... 95 unique samples” (also wrong)

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.

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.*

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

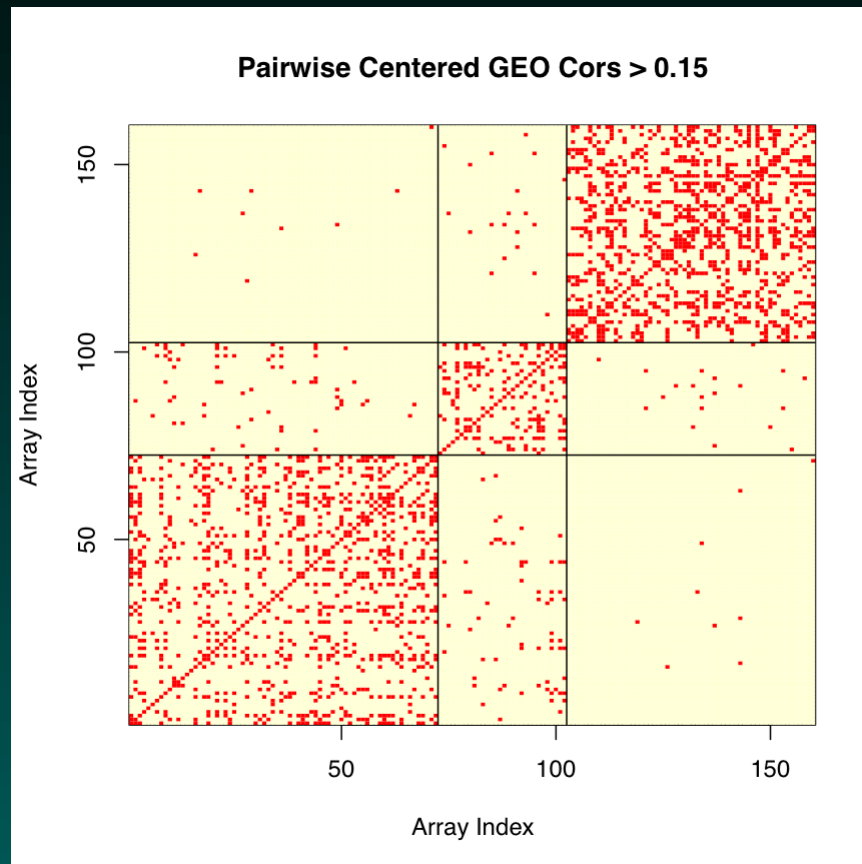
Hervé Bonnefoi, Anil Potti, Mauro Delorenzi, Louis Mauriac, Mario Camponé, Michèle Tubiana-Hulin, Thierry Petit, Philippe Rouanet, Jacek Jassem, Emmanuel Blot, Véronique Becette, Pierre Farmer, Sylvie André, Chaitanya R Acharya, Sayan Mukherjee, David Cameron, Jonas Bergh, Joseph R Nevins, Richard D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

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.

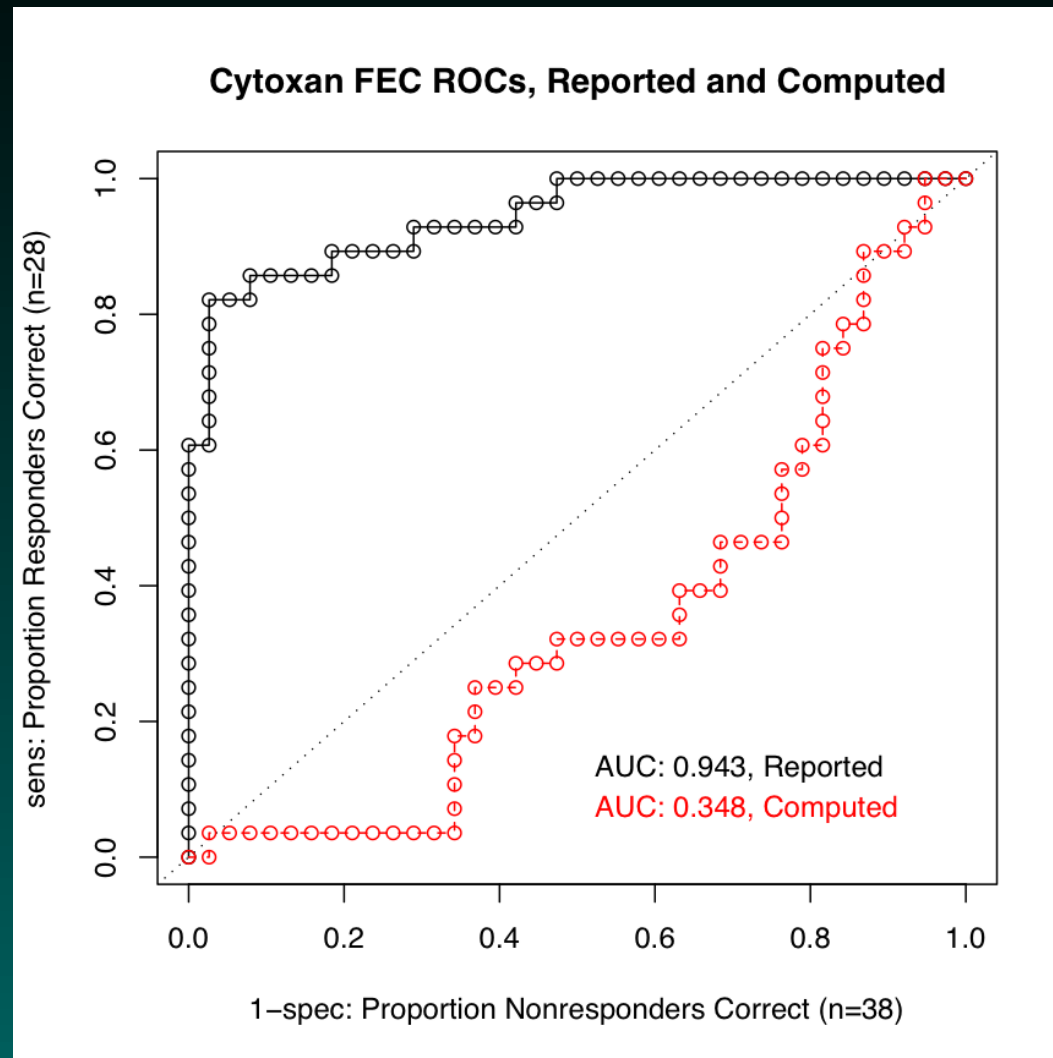
$$P(TFAC) = P(T) + P(F) + P(A) + P(C) - P(T)P(F)P(A)P(C).$$

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

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

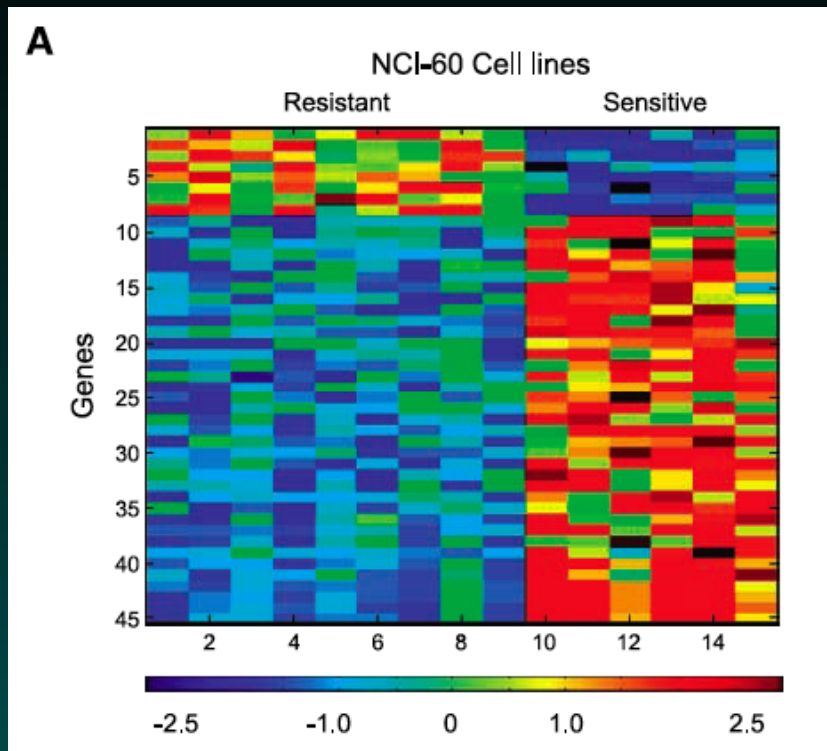
Each rule is different.

Predictions for Individual Drugs? (Reply)



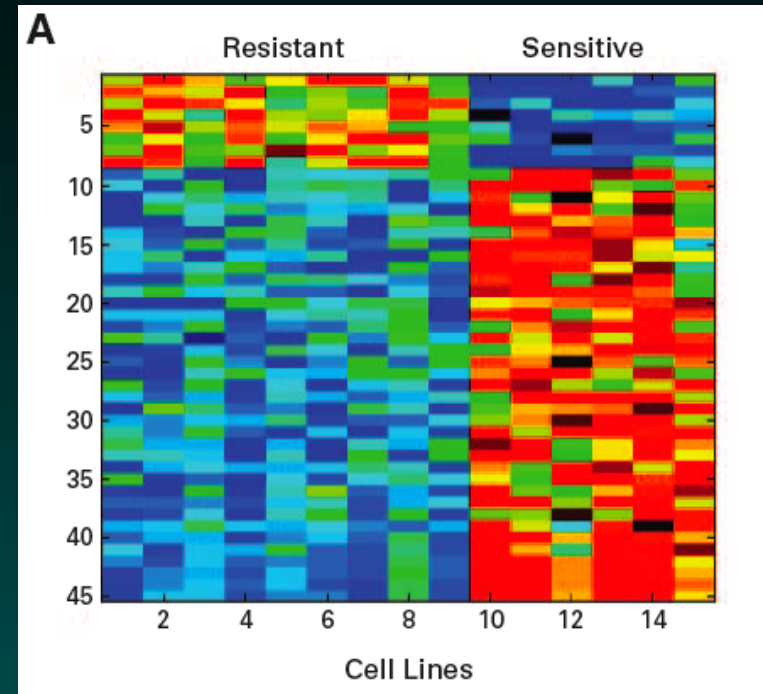
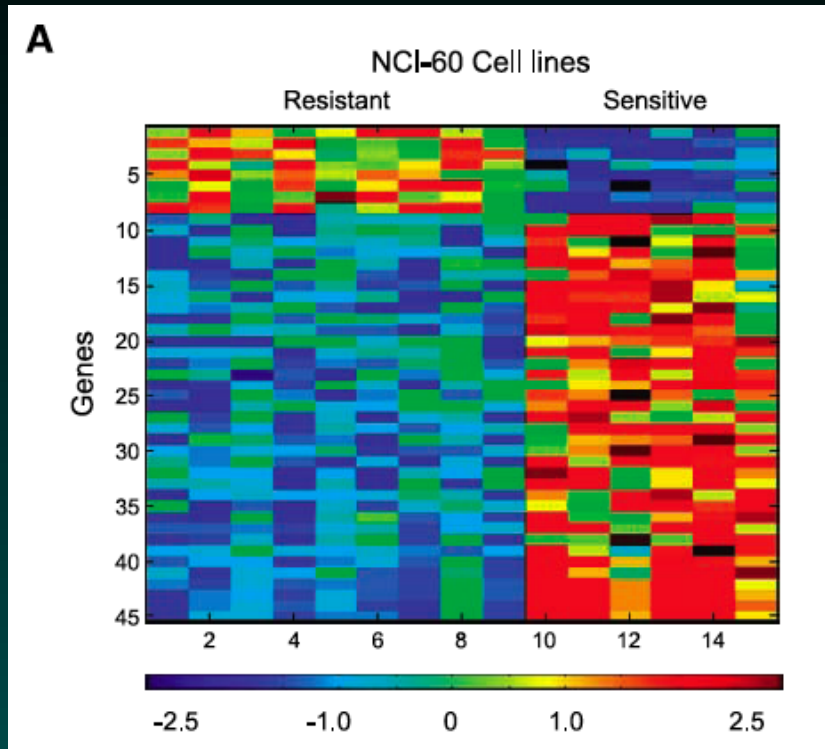
Does cytoxan make sense?

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.

Some Timeline Here...

Nat Med Nov 06*, Nov 07*, Aug 08. JCO Feb 07*, Oct 07*.
Lancet Oncology Dec 07*. PLoS One Apr 08. CCR Jan 09*.
(* errors reported)

Some Timeline Here...

Nat Med Nov 06*, Nov 07*, Aug 08. JCO Feb 07*, Oct 07*.
Lancet Oncology Dec 07*. PLoS One Apr 08. CCR Jan 09*.
(* 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 Feb 07*, Oct 07*.
Lancet Oncology Dec 07*. PLoS One Apr 08. CCR Jan 09*.
(* 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.

Jan 29, 2010

The logo for 'The Cancer Letter' is displayed on a red rectangular background. The word 'THE' is in a small, white, sans-serif font. 'CANCER' is in a large, bold, white, sans-serif font. 'LETTER' is in a medium-sized, white, sans-serif font.

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...

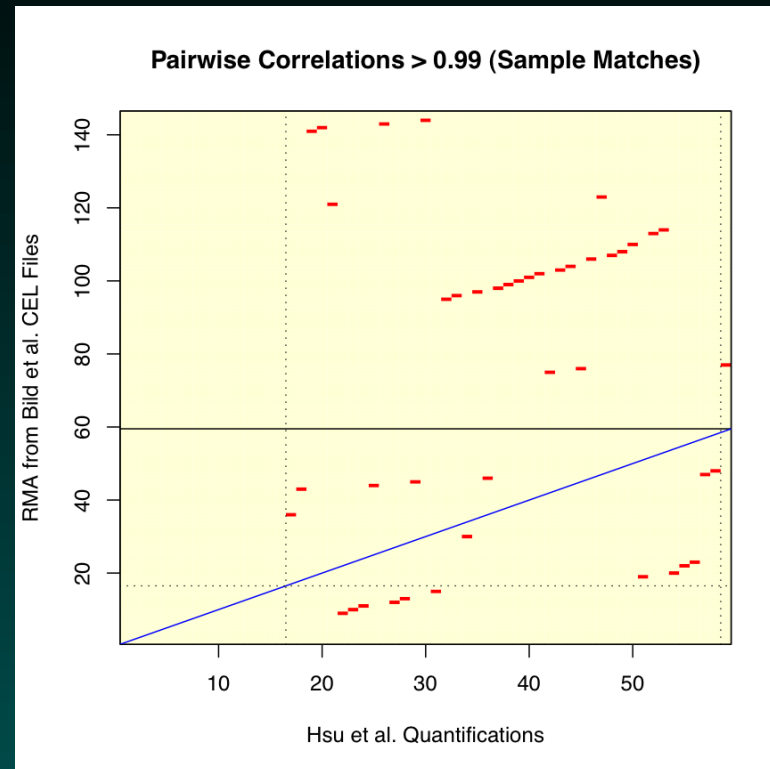
“While the reviewers approved of our sharing the report with the NCI, *we consider it a confidential document*” (Duke). A *future paper* will explain the methods.

There was also a major new development that 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 don't match at all.

The first 16 don't match because the genes are mislabeled.

We reported this to Duke and to the NCI in mid-November.

All data was stripped from the websites within the week.

More Timeline

“While the reviewers approved of our sharing the report with the NCI...”

More Timeline

“While the reviewers approved of our sharing the report with the NCI...”

April, 2010. Review report sought from the NCI under the Freedom of Information Act (FOIA).

May, 2010. Redacted report supplied; gaps noted.

More Timeline

“While the reviewers approved of our sharing the report with the NCI...”

April, 2010. Review report sought from the NCI under the Freedom of Information Act (FOIA).

May, 2010. Redacted report supplied; gaps noted.

May, 2010. NCI and CALGB pull lung metagene signature from an ongoing phase III trial.

Duke trials continue.

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

Subsequent Events

July 19/20: letter to Varmus; Duke resuspends trials

July 30: Varmus & Duke request IOM Involvement

Oct 22/29: call to retract JCO paper

Nov 9: **Duke announces trials terminated**

Nov 19: call to retract Nat Med paper, Potti resigns

There have been at least three developments since.

These involve the **NCI**, the **FDA**, and **Duke**.

Dec 20, 2010: The IOM Meets, the NCI Speaks

At the first meeting of the IOM review panel, Lisa McShane outlined some NCI interactions with the Duke group.

An [MP3](#) is available from *the Cancer Letter*.

Our questions prompted the NCI to ask questions of its own. The NCI released about 550 pages of documents giving the details.

We posted [our annotation](#) of these documents and [an overall timeline](#) on Jan 14th.

We're going to look at one case they examined.

The Lung Metagene Score (LMS)

The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

A Genomic Strategy to Refine Prognosis in Early-Stage Non–Small-Cell Lung Cancer

Anil Potti, M.D., Sayan Mukherjee, Ph.D., Rebecca Petersen, M.D.,
Holly K. Dressman, Ph.D., Andrea Bild, Ph.D., Jason Koontz, M.D.,
Robert Kratzke, M.D., Mark A. Watson, M.D., Ph.D., Michael Kelley, M.D.,
Geoffrey S. Ginsburg, M.D., Ph.D., Mike West, Ph.D., David H. Harpole, Jr., M.D.,
and Joseph R. Nevins, Ph.D.

One of “the most significant advances on the front lines of cancer.” – Ozols et al., JCO, 25:146-62, 2007, ASCO survey of 2006.

The Background of CALGB 30506

When this trial was proposed to the NCI, they had questions about the amount of blinding, so they asked for one more test.

The Background of CALGB 30506

When this trial was proposed to the NCI, they had questions about the amount of blinding, so they asked for one more test.

The LMS failed.

It almost achieved significance going the wrong way.

The Background of CALGB 30506

When this trial was proposed to the NCI, they had questions about the amount of blinding, so they asked for one more test.

The LMS failed.

It almost achieved significance going the wrong way.

After post-hoc adjustments (adjusting for batches), the Duke group attained some success (with IB, not IA).

At CALGB's urging, this went forward, but the NCI only allowed the LMS to be used for stratification, not allocation.

After our Paper

The NCI asked to see the data and code associated with the post-hoc adjustment.

Using the code and data provided, the NCI couldn't reproduce the success of the post-hoc adjustment.

More, they noticed another problem.

They ran the algorithm and got predictions.

After our Paper

The NCI asked to see the data and code associated with the post-hoc adjustment.

Using the code and data provided, the NCI couldn't reproduce the success of the post-hoc adjustment.

More, they noticed another problem.

They ran the algorithm and got predictions.

They ran it again.

After our Paper

The NCI asked to see the data and code associated with the post-hoc adjustment.

Using the code and data provided, the NCI couldn't reproduce the success of the post-hoc adjustment.

More, they noticed another problem.

They ran the algorithm and got predictions.

They ran it again.

The predictions changed.

The Changes Weren't Subtle

Some scores changed from 5% to 95%.

From one run to the next, high/low classifications changed about 25% of the time.

The Changes Weren't Subtle

Some scores changed from 5% to 95%.

From one run to the next, high/low classifications changed about 25% of the time.

The NCI was unhappy. They had understood the rule to be “locked down”, so that predictions wouldn't change.

The Changes Weren't Subtle

Some scores changed from 5% to 95%.

From one run to the next, high/low classifications changed about 25% of the time.

The NCI was unhappy. They had understood the rule to be “locked down”, so that predictions wouldn't change.

The NCI publicly yanked the LMS from CALGB 30506 in May 2010.

The Changes Weren't Subtle

Some scores changed from 5% to 95%.

From one run to the next, high/low classifications changed about 25% of the time.

The NCI was unhappy. They had understood the rule to be “locked down”, so that predictions wouldn't change.

The NCI publicly yanked the LMS from CALGB 30506 in May 2010.

The NEJM paper was retracted March 2, 2011.

The FDA Visits

On Jan 28, 2011, *the Cancer Letter* reported that an FDA Audit team was visiting Duke to examine how the trials had been run.

A key question was whether Investigational Device Exemptions (IDEs) were obtained before the signatures were used to guide therapy. They weren't.

The FDA Visits

On Jan 28, 2011, *the Cancer Letter* reported that an FDA Audit team was visiting Duke to examine how the trials had been run.

A key question was whether Investigational Device Exemptions (IDEs) were obtained before the signatures were used to guide therapy. They weren't.

Some aspects of this issue were anticipated by the IOM committee (Baggerly and Coombes, *Clinical Chemistry*, 57(5):688-90).

Per the FDA, **genomic signatures are medical devices.**

Duke's Initial Investigation

One thing that had puzzled us since the trial restarts was how the external reviewers missed the new errors we reported.

Duke's Initial Investigation

One thing that had puzzled us since the trial restarts was how the external reviewers missed the new errors we reported.

Jan 11, 2011: *Nature* talks to Duke.

The Duke deans overseeing the investigation, in consultation with the acting head of Duke's IRB, decided not to forward our report to the reviewers.

This was done to avoid “biasing the review”.

The IOM Meets Again: Mar 30-31

Science: Ned Calogne, Rich Simon, Chuck Perou, Sumitha Mandrekar

Institutional Responsibility: Peter Pronovost (Hopkins)

Case Studies: Laura Van't Veer, Steve Shak, Joe Nevins

Responsible Parties: journal editors (Cathy DeAngelis, JAMA, Veronique Kiermer, Nature, Katrina Kelner, Science), authors and PIs (Stott Zeger), institutions (Albert Reece, U Md, Harold Paz, Penn State, Scott Zeger, Hopkins)

Forensics: Keith Baggerly

<http://bioinformatics.mdanderson.org/Supplements/ReproRsch-All/Modified/IOM/>

Some Cautions/Observations

We've seen problems like these 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

Unfortunately, we suspect

The most simple mistakes are common.

How Should We Pursue Reproducibility?

AAAS, Feb 19: All slides and an MP3 of the audio are available from

<http://www.stanford.edu/~vcs/AAAS2011/>

ENAR, Mar 23: Panel on ethics in biostatistics. Handouts are available from our Google Group.

What's Missing? *Nature*, Mar 23

AACR, Apr 5: Difficulties in moving biomarkers to the clinic.

CSE, May 3

NCI, Jun 23-4 What should the NCI be looking for in grants that it funds?

Check back...

What Should the Norm Be?

For our group? Since 2007, we have prepared reports in *Sweave*.

For papers? (Baggerly, *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?

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)

Index

Title
Cell Line Story
Trying it Ourselves
Matching Features
Using Software/Making Predictions
Outliers
The Reply
Adriamycin Followup
Hsu et al (Cisplatin)
Timeline, Trials, Cancer Letter
Trial Restart and Objections
Final Lessons