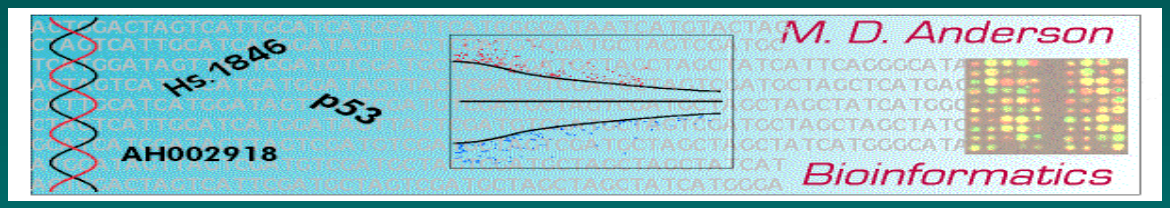
Reproducible Research with R and RStudio C01: Why

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

GCC, June 6, 2018



Why is Reproducibility Important in H-T B?

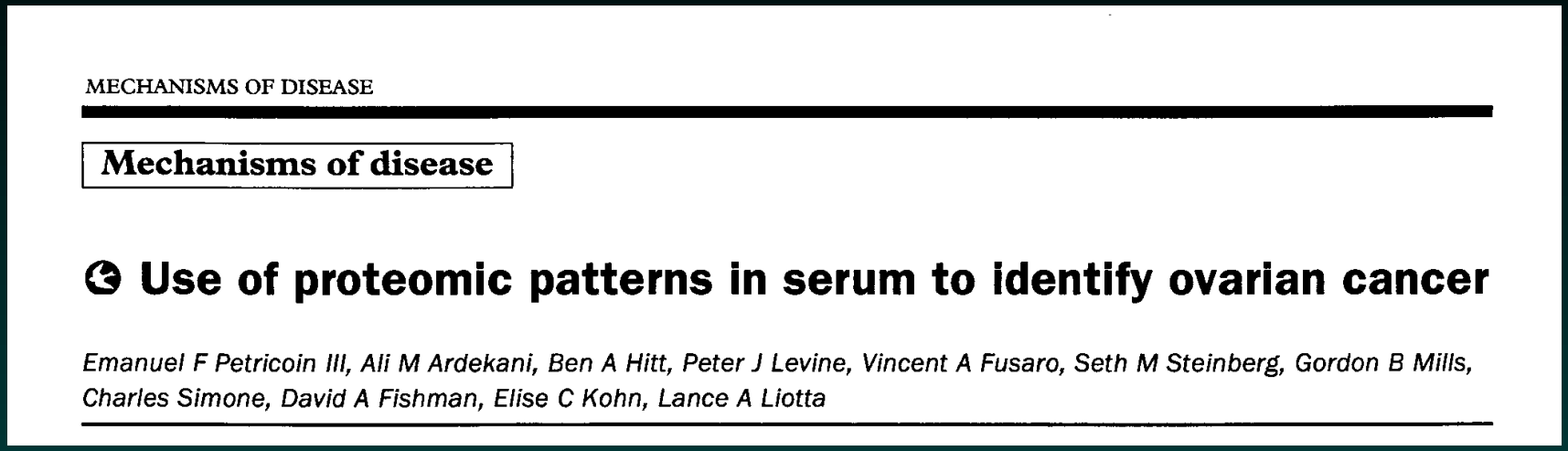
Our intuition about what “makes sense” is very poor in high-d.

To use “omics-based signatures” as biomarkers, we need to know they’ve been assembled correctly.

Without documentation, we may need to employ (lengthy!) *forensic bioinformatics* to infer what was done.

Let’s look at *some examples* in the context of diagnosis and treatment of cancer

Using Proteomics for Early Detection



MECHANISMS OF DISEASE

Mechanisms of disease

Use of proteomic patterns in serum to identify ovarian cancer

Emanuel F Petricoin III, Ali M Ardekani, Ben A Hitt, Peter J Levine, Vincent A Fusaro, Seth M Steinberg, Gordon B Mills, Charles Simone, David A Fishman, Elise C Kohn, Lance A Liotta

* 100 ovarian cancer patients
* 100 normal controls
* 16 patients with “benign disease”

Use 50 cancer and 50 normal spectra to train a classification method; test the algorithm on the remaining samples.

Their Results

* Correctly classified 50/50 ovarian cancer cases
* Correctly classified 46/50 normal cases
* Correctly classified 16/16 benign disease cases as “other”.

Data at
http://home.ccr.cancer.gov/ncifdaproteomics/
(used to be at http://clinicalproteomics.steem.com)

Large sample sizes, using serum

The Data Sets

3 data sets on ovarian cancer

Data Set 1 – The initial experiment. 216 samples, baseline subtracted, H4 chip

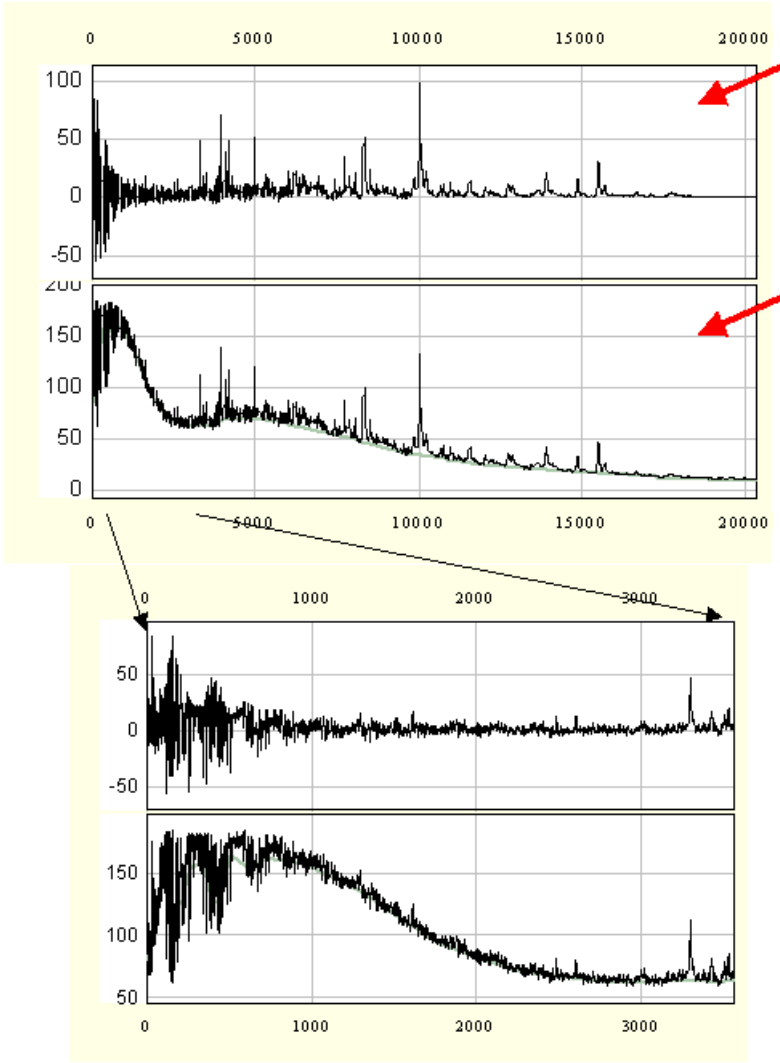
Data Set 2 – Followup: the same 216 samples, baseline subtracted, WCX2 chip

Data Set 3 – New experiment: 162 cancers, 91 normals, baseline NOT subtracted, WCX2 chip

A set of 5-7 separating peaks is supplied for each data set.

We tried to
(a) replicate their results, and
(b) check consistency of the proteins found.

We Can't Replicate their Results (DS1 & DS2)



They posted this

Baseline Subtraction ON

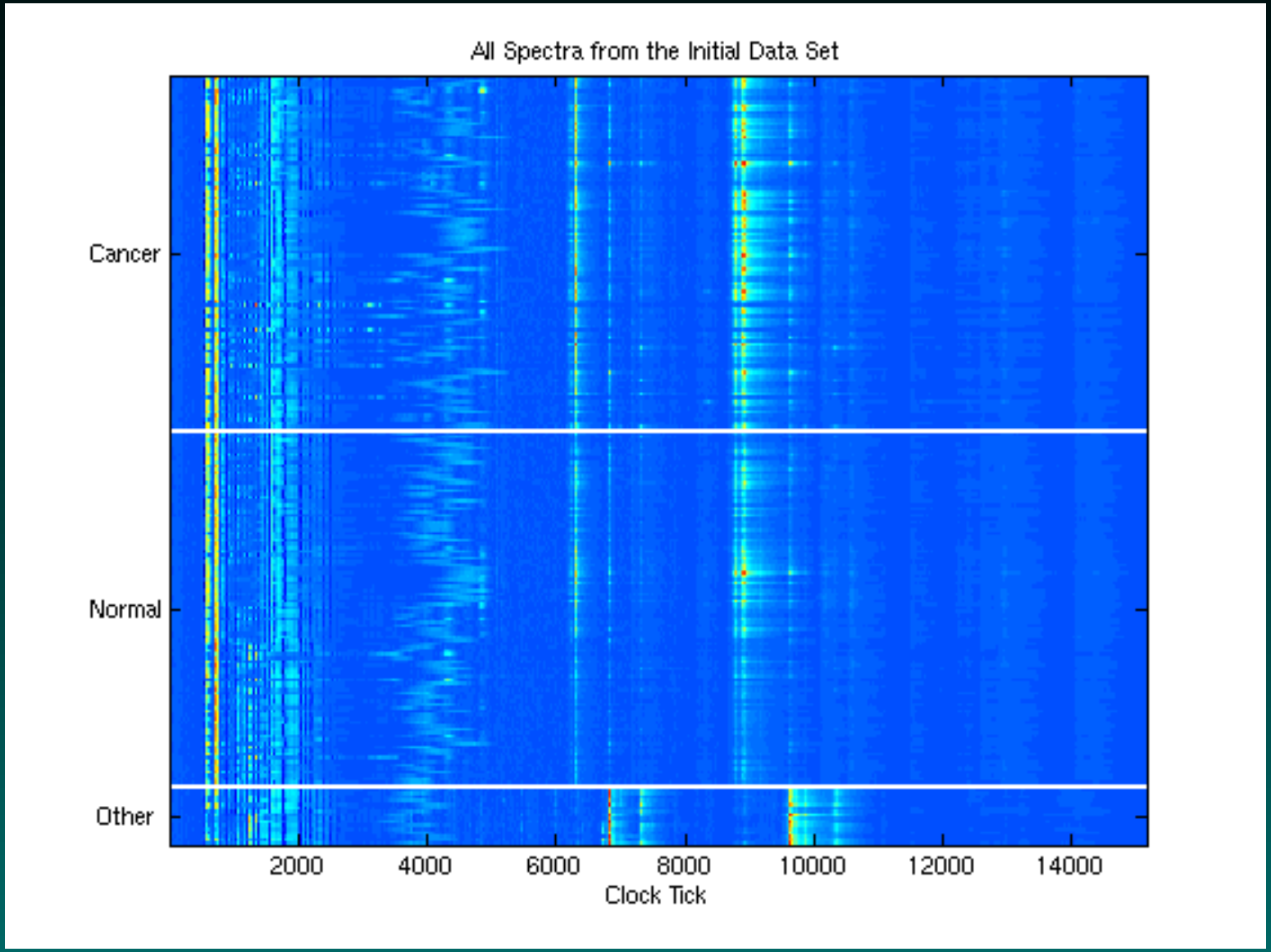
They analyzed this

Baseline Subtraction OFF

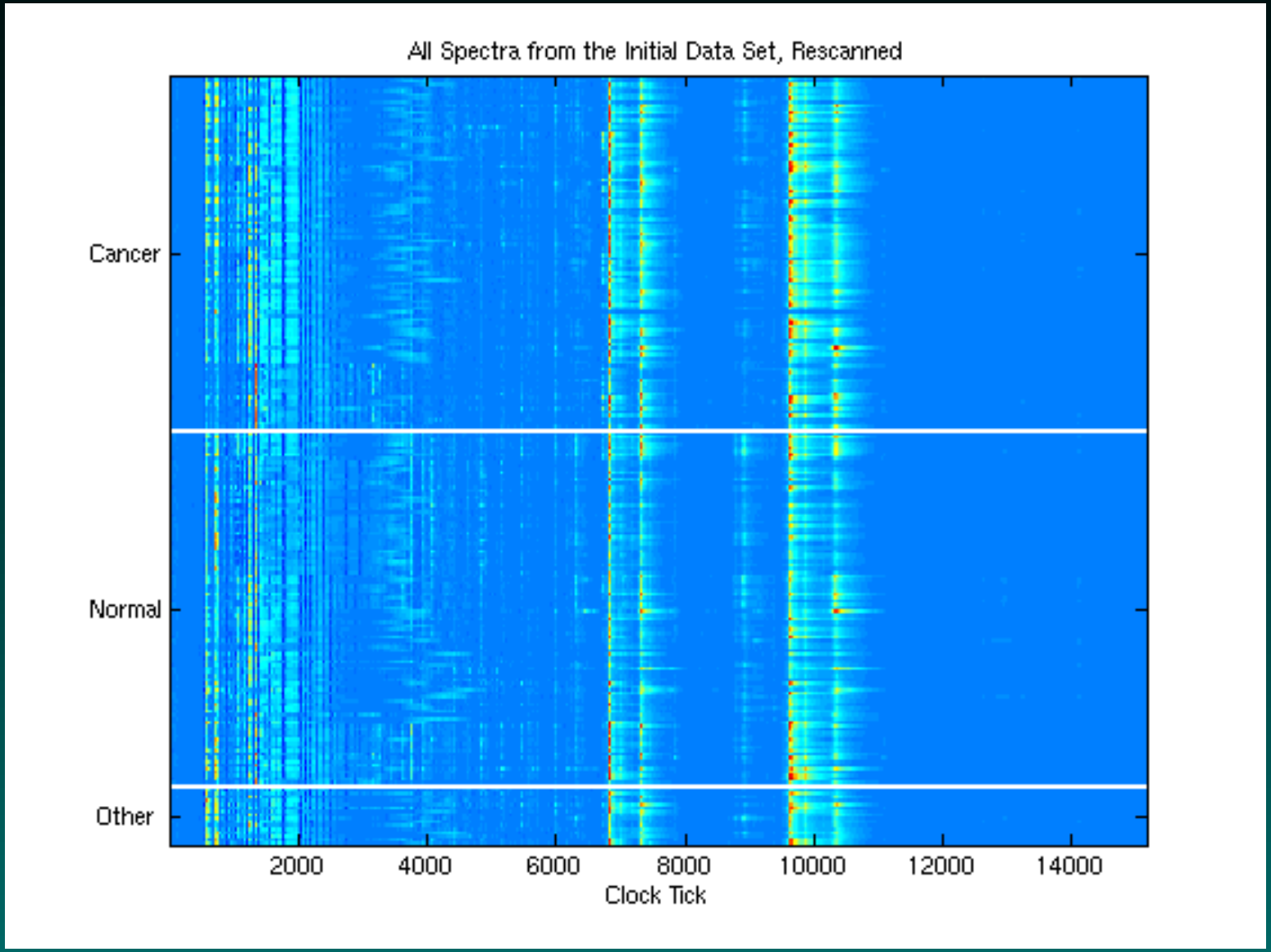
Zoomed in with Baseline Sub ON

Zoomed in with Baseline Sub OFF

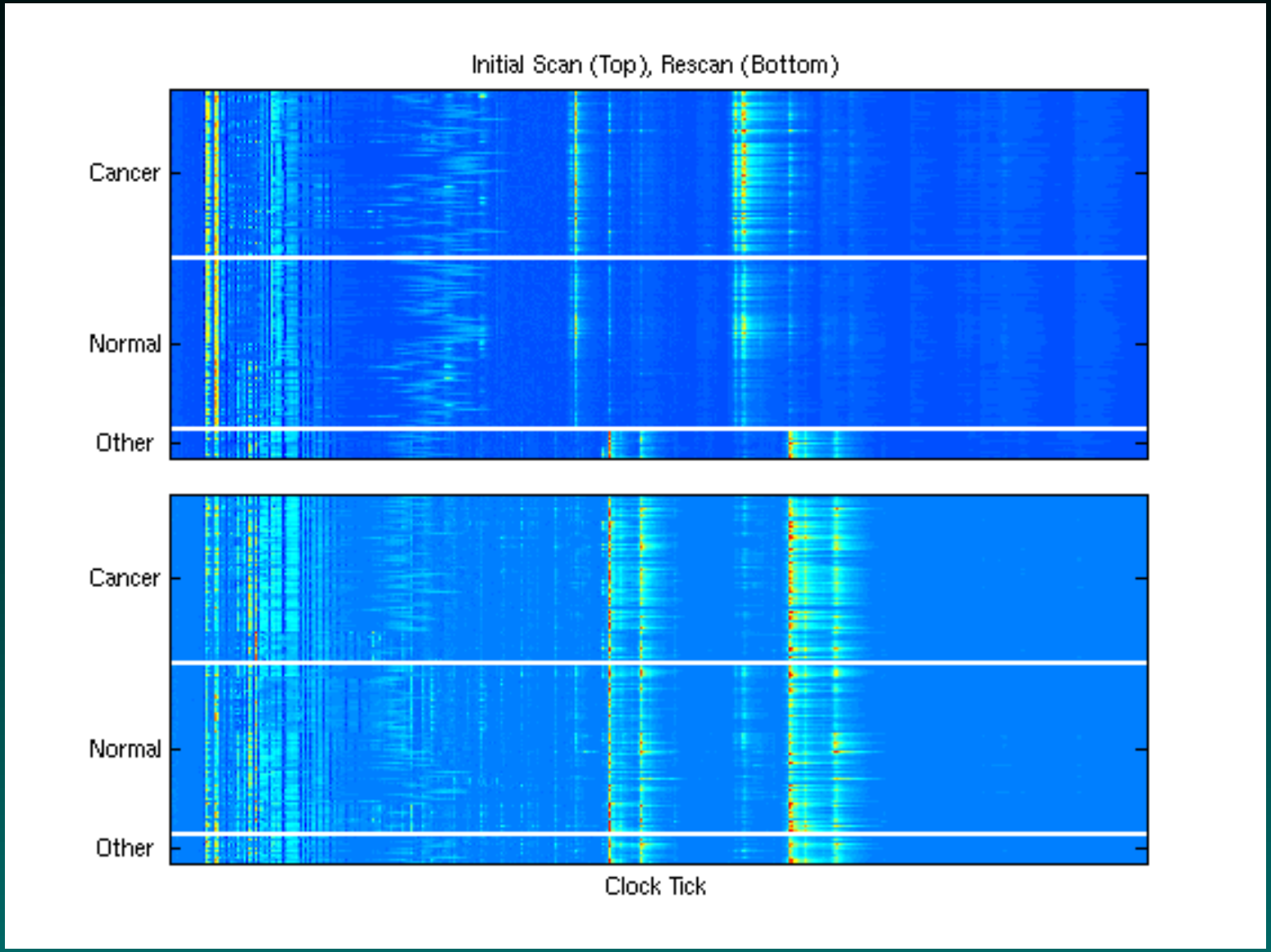
Some Structure is Visible in DS1



Or is it? Not in DS2



Processing Can Trump Biology (DS1 & DS2)



Meanwhile...

January 2004: Correlogic, Quest Diagnostics, Lab Corp announce plans to offer a “home brew” test called **OvaCheck**.

Samples would be sent in by clinicians for diagnosis.

Estimated market: 8-10 million women.
Estimated cost: $100-200 per test.

A Timeline

2004:

* Jan 29: Critiques available online
* Feb 3: New York Times coverage
* Feb 7: Statement from SGO
* Feb 18: FDA letter to Correlogic
* Mar 2: FDA letters to Quest, Lab Corp
* July: FDA rules OvaCheck is subject to pre-market review as a device

2006:

* FDA releases draft guidance on IVDMIAs
* NCI Clinical Proteomic Technologies for Cancer (CPTAC)

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, 12:1294-300.

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

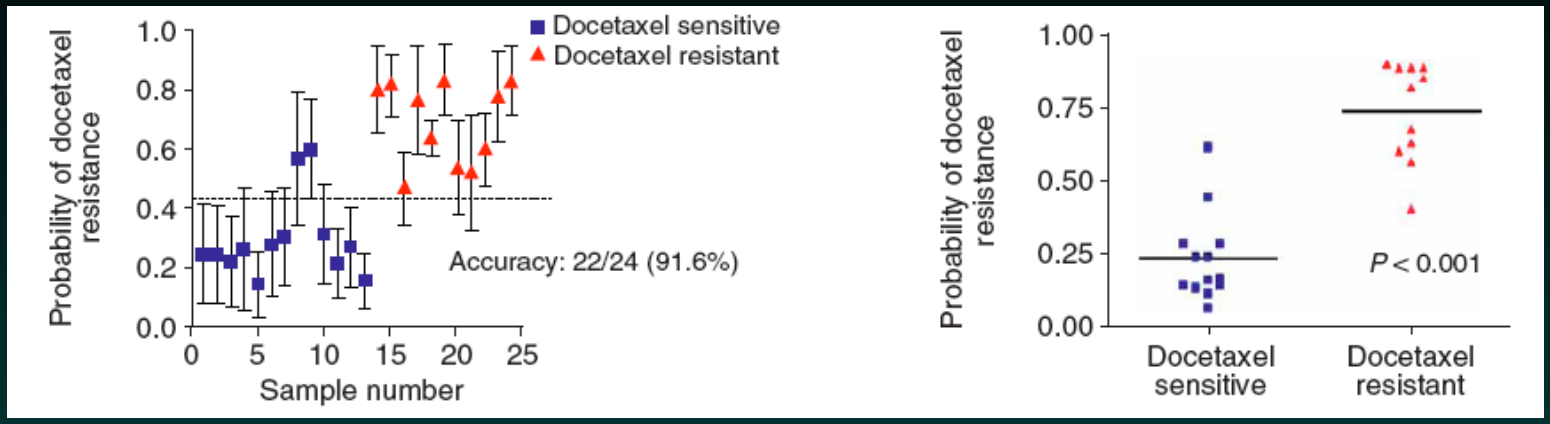
They provide examples using 7 commonly used agents.

This got people at MDA very excited.

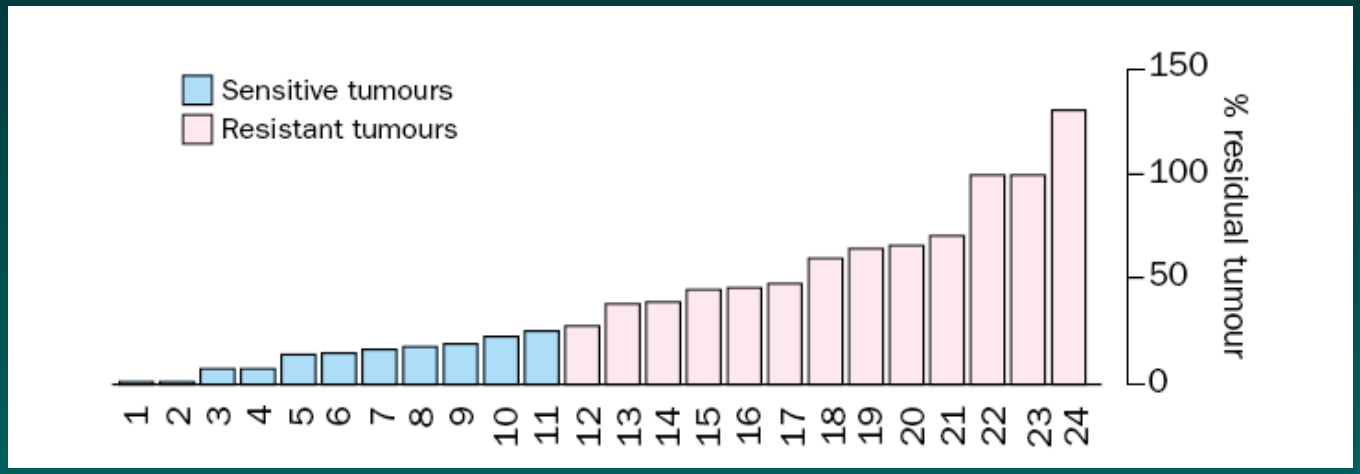
Their Gene List and Ours

```
> 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"
...
```

Predicting Response: Docetaxel

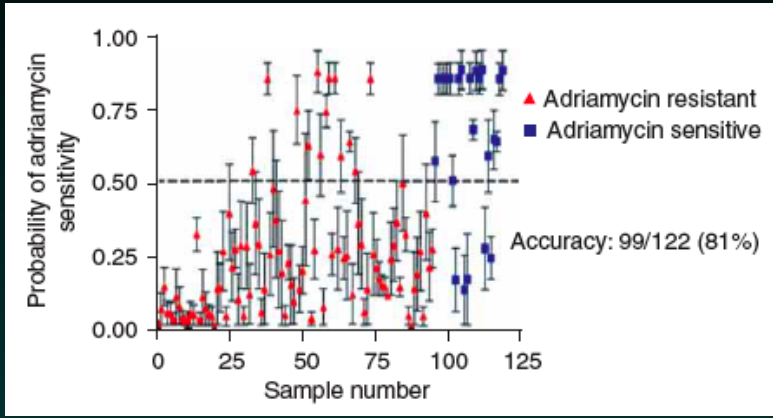


Potti et al (2006), Nature Medicine, 12:1294-300, Fig 1d

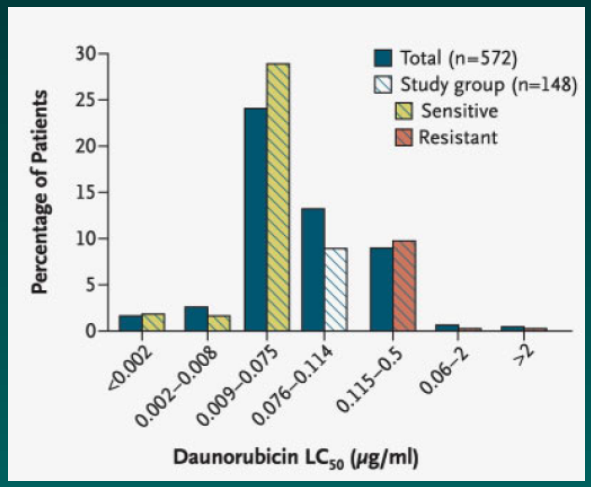


Chang et al, Lancet 2003, 362:362-9, Fig 2 top

Predicting Response: Adriamycin



Potti et al (2006), Nature Medicine, 12:1294-300, Fig 2c



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

Partial Timeline

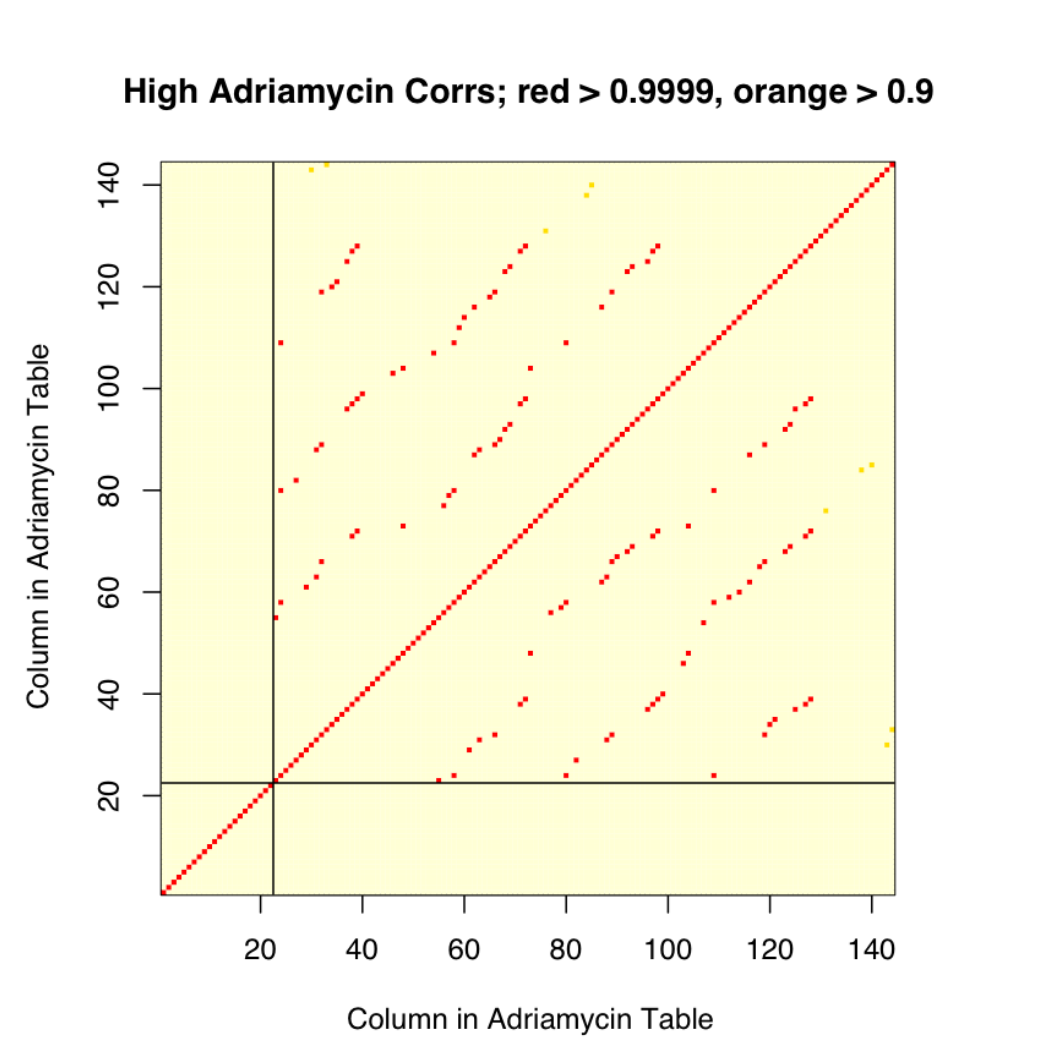
2006:

* 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".

Adriamycin 0.9999+ Correlations



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

Why We Care

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

Why We Care

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:

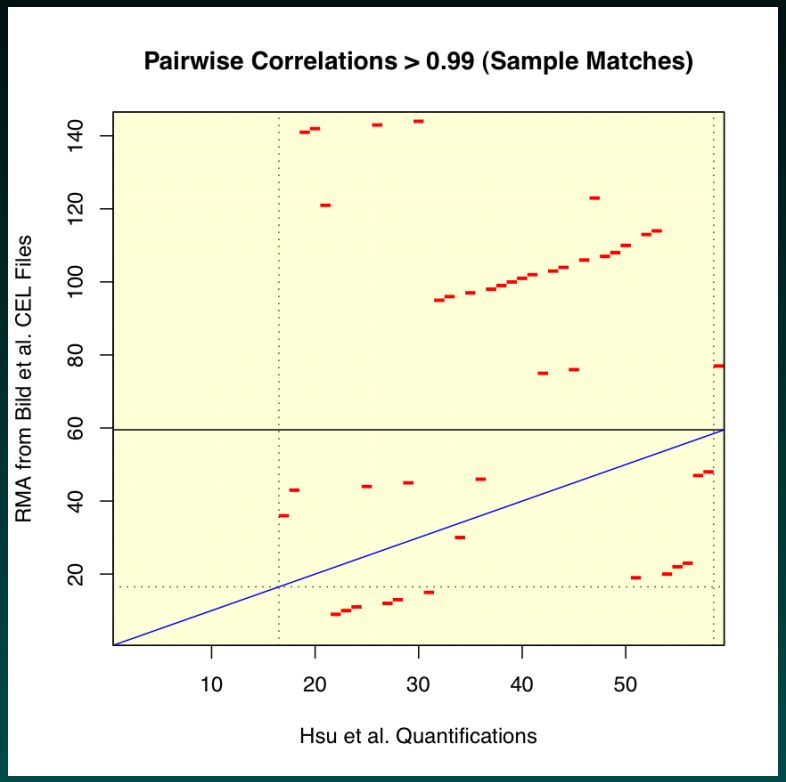
Story covered by *The Cancer Letter*; Oct 2, Oct 23.
NCI raises concerns with Duke's IRB behind the scenes.
Duke starts internal investigation, suspends trials.

New Data

Early-Nov ’09 (mid-investigation), the Duke team posted new data for cisplatin and pemetrexed (in lung trials since ’07).

These included quantifications for the 59 ovarian cancer test samples (from GSE3149, which has 153 samples) they used to validate their predictor.

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.

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

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.

This did give us one more option...

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.

This did give us one more option...

In May 2010, we obtained a copy of the reviewers' report from the NCI under FOIA (Cancer Letter, May 14).

In our assessment (and others'), it didn't justify restarting trials.

There was no mention of our Nov 2009 report.

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

Other Developments

117 patients were enrolled in the trials.
Sep, 2011: Patient lawsuits filed (11+ settlements).

Misconduct investigation (Jul 2010-Nov 2015).
10 retractions, 6+ “partial retractions”
FDA Review, Discussions with Duke IRB

Jul 8, 2011: Front Page, NY Times.

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

Mar 23, 2012: IOM Report Released.
http://www.iom.edu/Reports/2012/Evolution-of-Translational-Omics.aspx

Some Cautions/Observations

These cases are 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

This is not an Isolated Problem

Ioannidis et al. (2009), *Nat. Gen.*, **41**:149-55. Tested reproducibility of microarray papers. Could reproduce 2/18.

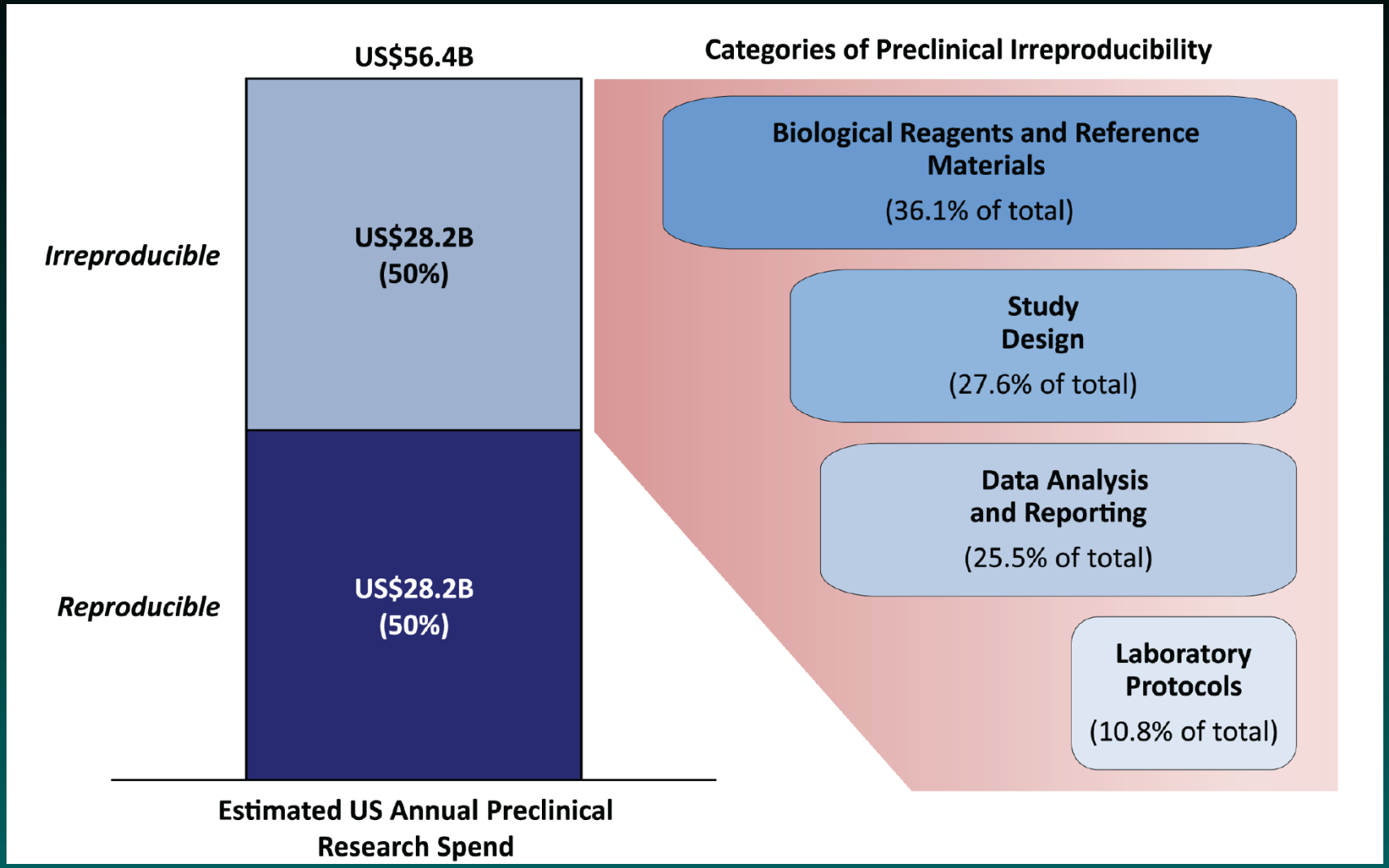
Begley and Ellis (2012), *Nature*, **483**:531-3. Amgen attempted replication of clinical "breakthroughs" prior to further study. Validated 6/53.

NCI focus meeting Sep 2012.

Collins and Tabak (2014), *Nature*, **505**:612-3.

SISBID RR Short Course July, 2015,2016,2017

Some Cost Breakdowns



Freedman et al (2015), PLoS Biology, 13(6):e1002165

What Have We and Others Suggested?

Exploiting a Teachable Moment...

Baggerly et al *Nature* (2010)

Give us your data, your code, your huddled masses

Records of data provenance

Checking existence as a task for journals and reviewers
(are there links? are they live?)

NCI Guidelines in *Nature* Oct 2013

Reasons for Hope

1. Our Own (Evolving!) Experience
2. Better tools (knitr, markdown, GitHub, the tidyverse)
3. Journals, Code and Data
4. The IOM, the FDA, and IDEs*
5. The NCI and Trials it Funds
6. OSTP, Congress, Science, Nature
7. NIH Rigor and Reproducibility Initiative

Some Places to Learn More

Karl Broman's Tools for RR Course

Roger Peng's Coursera course and notes (2013)

Christopher Gandrud's book (2e, 2015)

Yihui Xie's book (2e, 2015)

Hadley Wickham's R Packages book (2015)

NAS meeting, Feb 26-7, 2015

ENAR Webinar, Nov 20, 2015

SISBID Reproducible Research Short Course, July 2018

Some Reports

Baggerly, Morris and Coombes (2004), *Bioinformatics*, **20(5)**:777-785.
Baggerly, Edmonson, Morris and Coombes (2004), *Endocrine-Related Cancer*, **11**:583-584.
Baggerly, Morris, Edmonson and Coombes (2005), *J. Natl. Cancer Inst.*, **97**:307-309.

Coombes, Wang and Baggerly (2007), *Nat. Med.*, **13**:1276-7.
Baggerly and Coombes (2009), *Ann. App. Statist.*, **3(4)**:1309-34. http://bioinformatics.mdanderson.org/Supplements/ReproRsch-All
Baggerly and Coombes (2011), *Clin. Chem.*, **57(5)**:688-90.

More at http://bioinformatics.mdanderson.org.

Acknowledgments

Kevin Coombes

Yang Zhao, Ying Wang, Shelley Herbrich
Shannon Neeley, Jing Wang
David Ransohoff, Gordon Mills
Jane Fridlyand, Lajos Pusztai, Zoltan Szallasi

M.D. Anderson Ovarian, Lung and Breast SPOREs

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

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

Thanks!