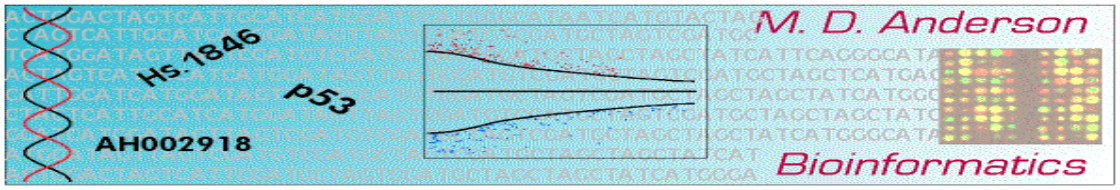
Why is Replicability Hard?

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

SISBID, July 19, 2016



Expanding our Brief

We’re mostly focusing on improving reproducibility.

The real problem is poor replicability.

We’re giving you tools to address the former because we have these on hand.

To help you address the latter, we can provide some awareness of what the major problems are, so you can keep these in mind and either avoid them or mitigate their impact.

So, what do we know about replicability?

Lack of Replicability is a Big Problem

"Why Most Published Research Findings are False"
Ioannidis (2005), PLoS Med, 2(8):e124

Numbers of studies not replicating in 4 surveys:

Begley and Ellis (2012), Nature, 483:531-3.
47/53, 89%.

Prinz et al (2011), Nat Rev Drug Discovery, 10:712-3.
52/67, 78%.

Glasziou et al (2008), BMJ, 336:1472-4.
41/80, 51%.

Vasilevsky et al (2013), PeerJ, 1:e148
128-9/238, 54%.

So What Should We Do?

Acknowledge the issue:

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

Present some partial fixes:

Better data sharing, code access
Grant RFAs

Ask for input:

NAS Meeting, Feb 26-7, 2015, videos

Is Replication Worse than Before?

Yes and no.

I don't think people have gotten markedly sloppier or worse over time.

We're just now beginning to recognize how much clutter exists in "the literature", awareness has increased.

"Big Data" does change some things.

Expanding dataset sizes can make some types of common errors harder to spot - errors can get "lost".

This may be addressable with better quality control.

Is Better Reproducibility Enough?

No. (It's a prerequisite for improvement.)

There are too many "experimenter degrees of freedom".

So, what can we do?

- Clarify terminology
- Specify strength of findings in terms of likely replicability
- Avoid stupidity (incorrect analyses)
- Be aware of sizes of assay variability
- Prespecify sanity checks and at least some hypotheses

What Problem Are We Addressing?

We defined “reproducibility” and “replicability” on day 1, but our definitions are not “consensus”.

Many journals (and fields) reverse our phrasing.

Reproducibility and replicability are often used interchangeably, mirroring common linguistic usage.

There are even further strata, depending on whether *you* can reproduce (or replicate) your results, or *someone else* can reproduce (or replicate) them.

We need to know which concept we’re discussing.
– Steve Goodman

P-Values $\neq$ P(Will Replicate); Part 1

Boos and Stefanski (2011), *Am Statistician*, 65(4):213-21.

“P-Value Precision and Reproducibility”

P-values are random variables, and as such have distributions with definable center and spread.

Variation of p-values is rarely considered.

“only the magnitude of $-\log_{10}(\mathrm{p}-\mathrm{value})$ is reliably determined”

“the probability of nonreplication of published studies with p-values in the range 0.005 to 0.05 is roughly 0.33.”

P-Values $\neq$ P(Will Replicate); Part 2

Johnson (2013), *PNAS*, 110(48):19313-7.

"Revised Standards for Statistical Evidence"

We should really be using Bayes Factors

Using uniformly most powerful Bayesian tests (UMPBTs) "provides a direct connection between significance levels, P-values, and Bayes factors"

"these results suggest that between 17% and 25% of marginally significant scientific findings are false"

More realistically, "These analyses suggest that the range 17-25%" is an *underestimate*.

P-Values $\neq$ P(Will Replicate); Part 3

Boos (Frequentist):

"to have the estimate of $P(p_{new} \leq 0.05)$ at least 90%, we need $p_{obs} \leq 0.001$"

Johnson (Bayesian):

"Make 0.005 the default level of significance"

Very different approaches, but comparable bottom lines.

We're using the wrong cutoffs if replication is the goal.

Be suspicious of marginal results!

Was the Analysis Done Correctly?

Dupuy and Simon (2007), *JNCI*, 99:147-57.

Was the right type of test used?

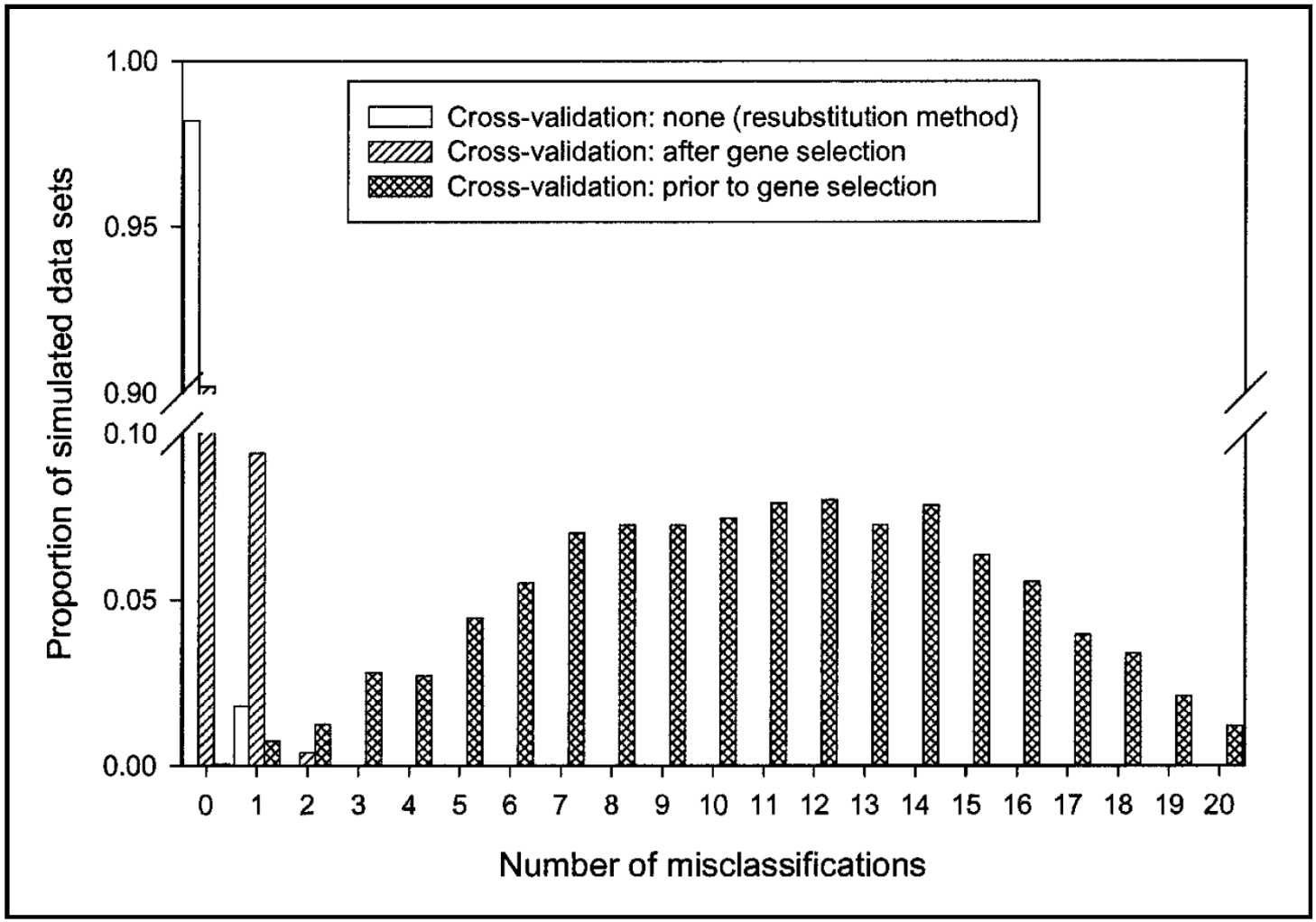
Were significance levels adjusted for multiple testing?

If there were training sets and test sets for developing classification rules, was the rule properly locked down?

In estimating out-of-sample prediction accuracy, was cross-validation employed correctly?

Dupuy & Simon found 21/42 papers surveyed flubbed at least one of the steps above.

Cross-Validation



Simon et al (2003), *JNCI*, 95(1):14-18.

Are the Stats Getting Better?

I'd like to think so.

Awareness of multiple testing and use of false discovery rates (FDRs) has certainly become more widespread.

That said, the FDA, NCI, and IOM were focusing on continued problems and locking down decision rules in 2012.

They're better, but not perfect.

Let's look at some examples of another stats/epi problem that's still with us...

A Proteomics Case Study: Feb 16 '02 Lancet

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 of the ovarian cancer cases.
- Correctly classified 46/50 of the normal cases.
- Correctly classified 16/16 of the benign disease 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

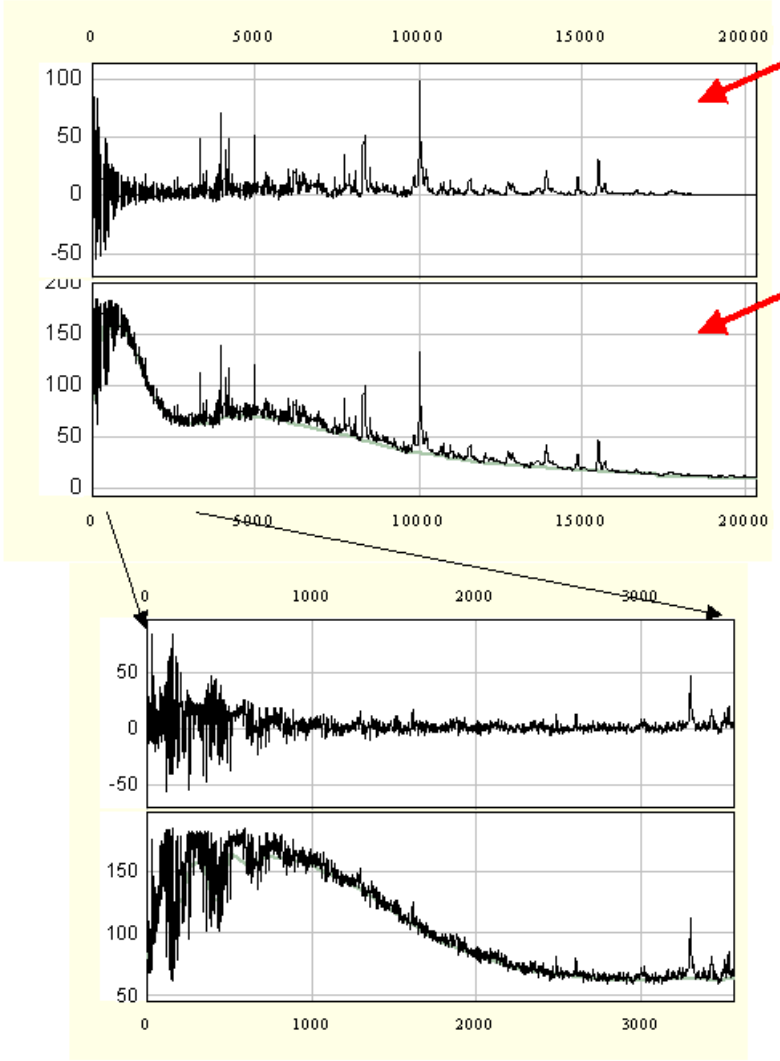
Data Set 2 – Followup: the same 216 samples, baseline subtracted, WCX2 chip

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

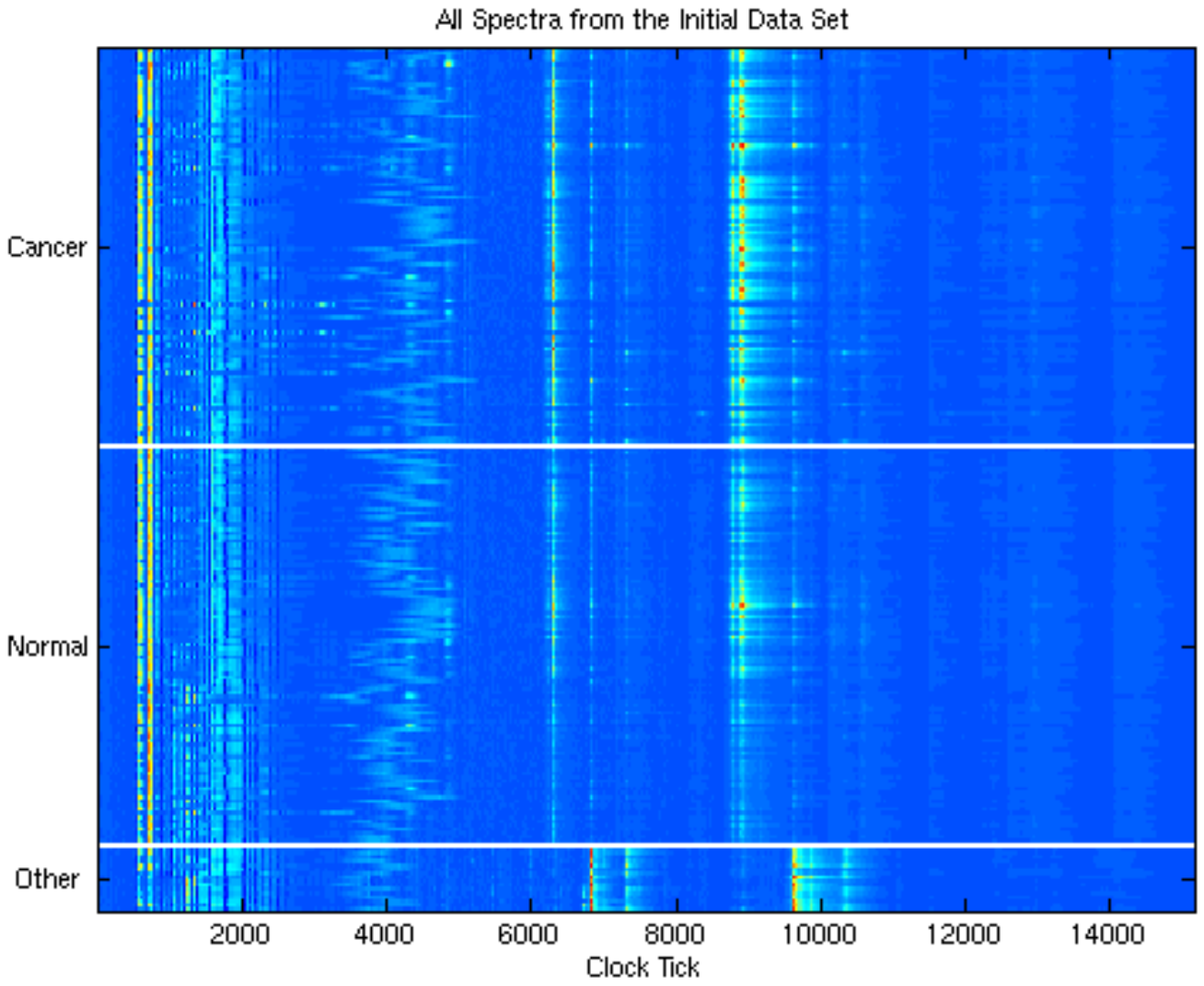
They analyzed this

Baseline Subtraction OFF

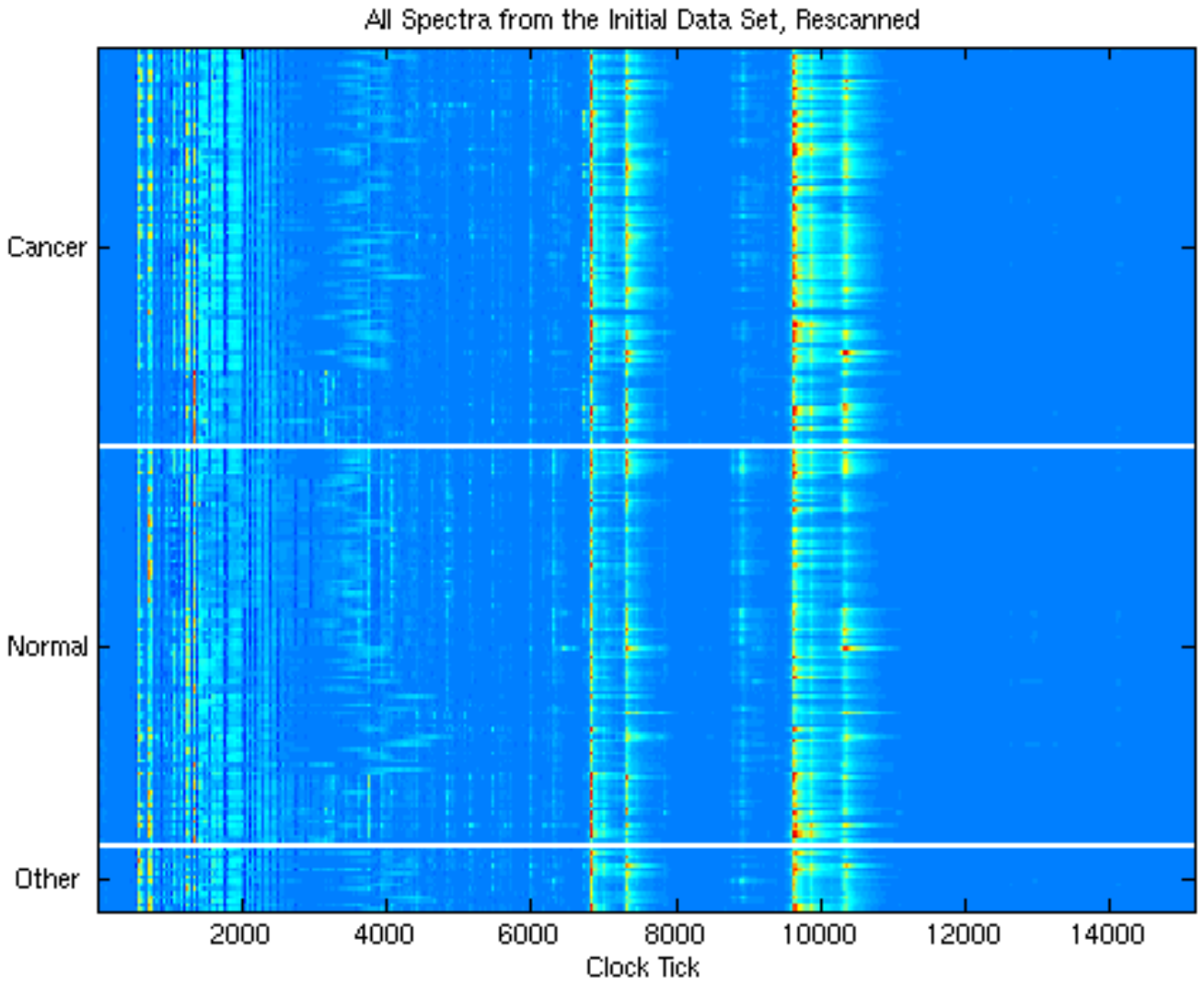
Zoomed in with Baseline Sub ON

Zoomed in with Baseline Sub OFF

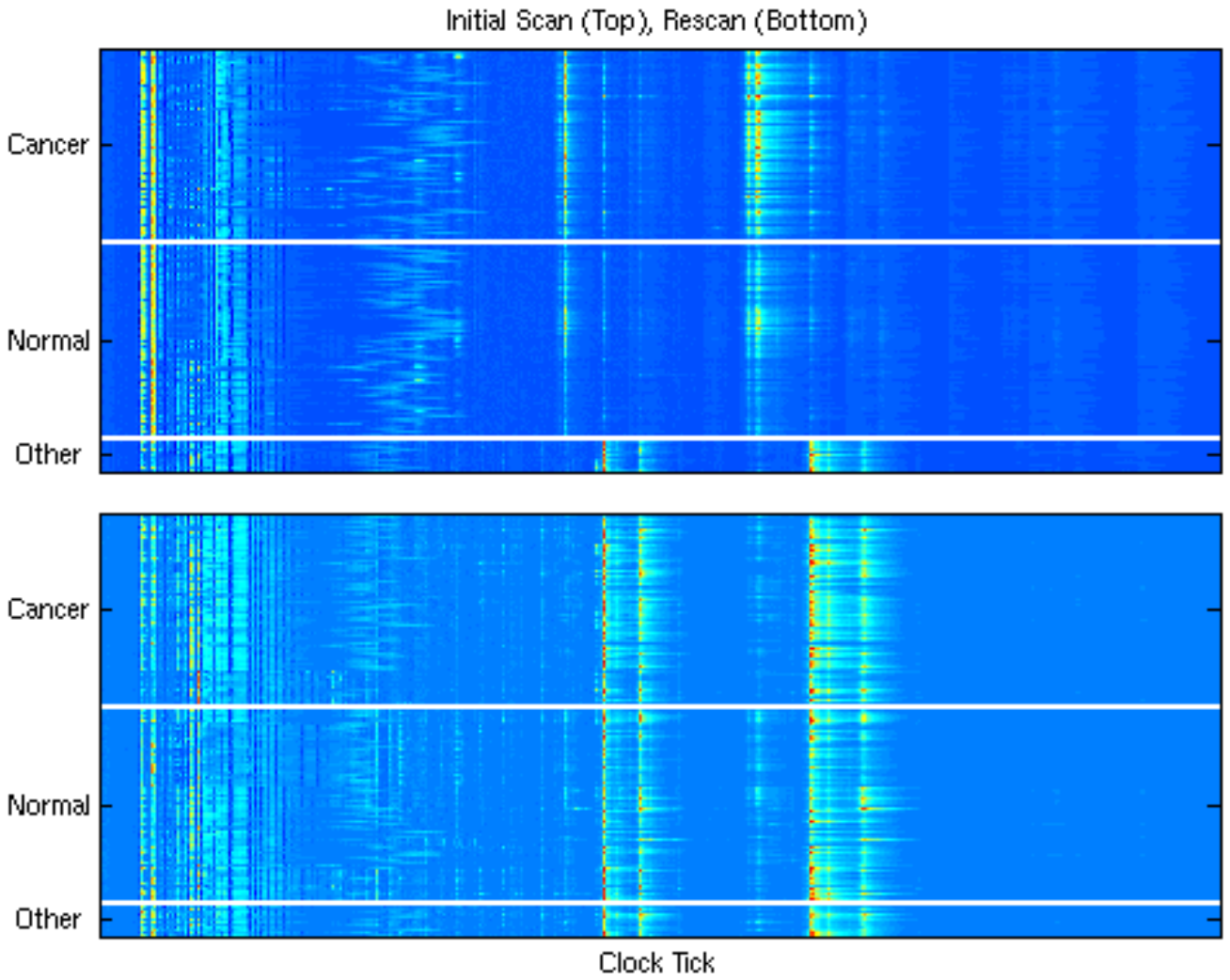
Some Structure is Visible in DS1



Or is it? Not in DS2



Processing Can Trump Biology (DS1 & DS2)



Are We Beating a Dead Horse?

Qstar data is higher resolution.

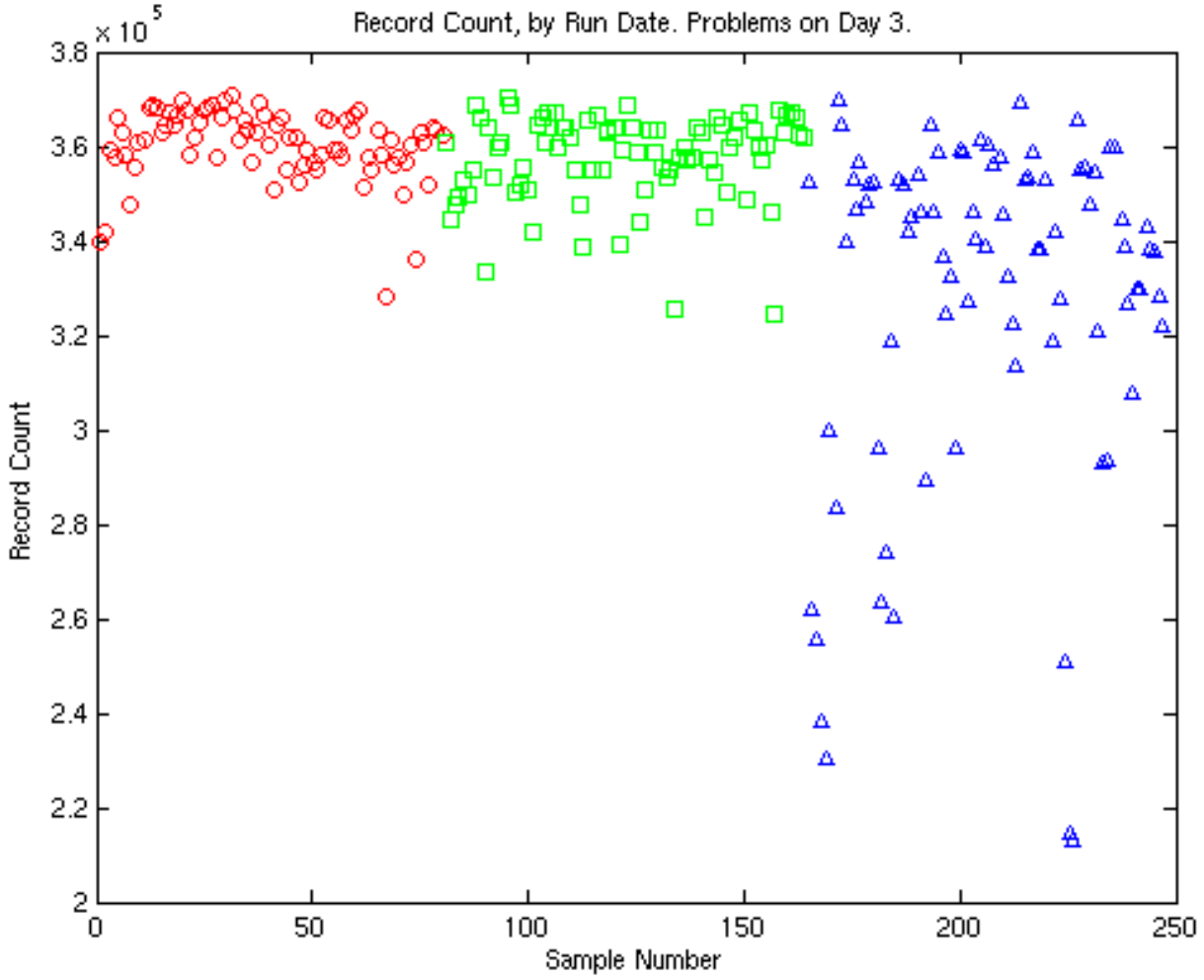
They've added some QA/QC steps to remove bad spectra.

Still using patterns.

Reported results are even better.

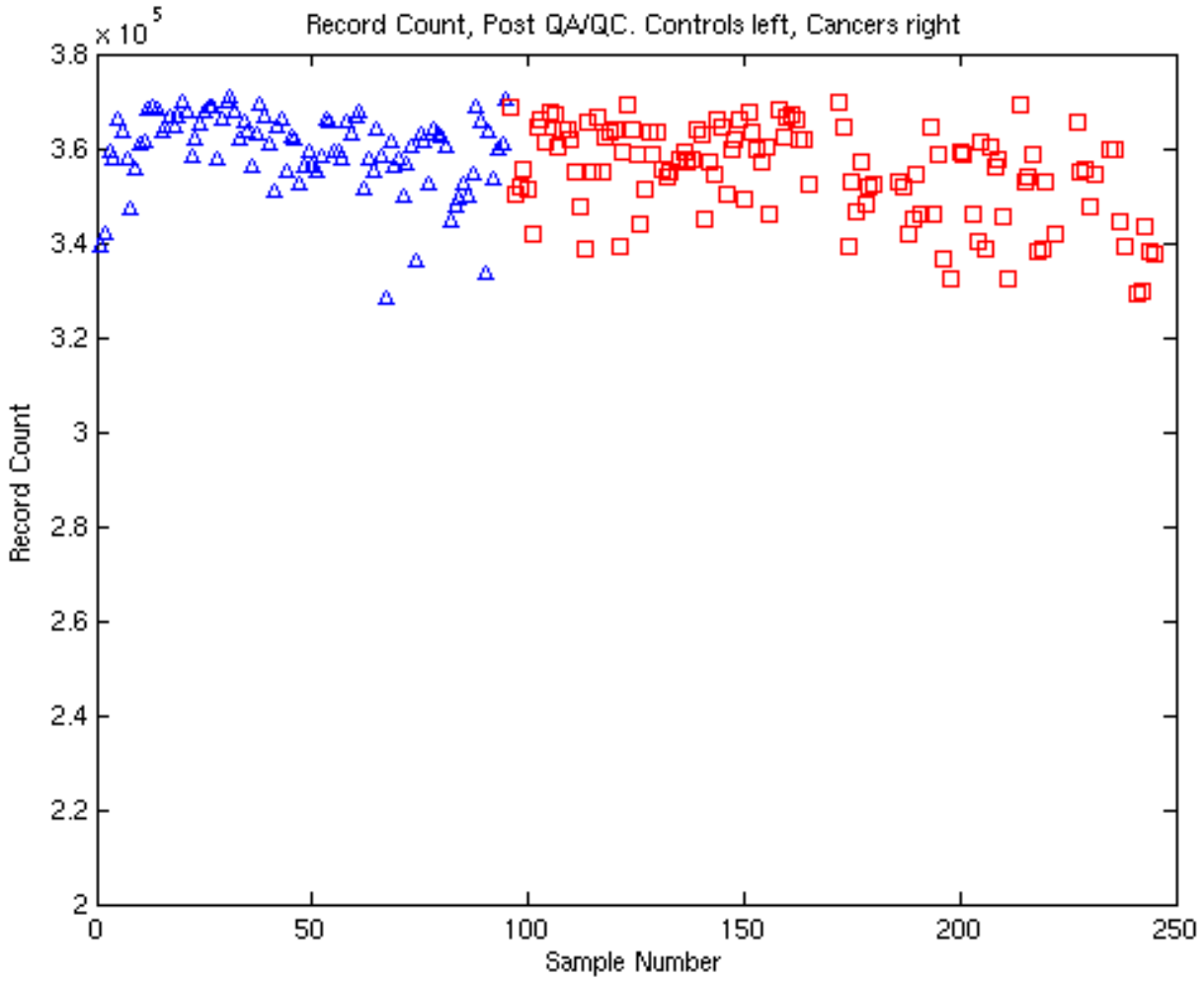
Endocrine-Related Cancer (Jul '04) – 100% sensitivity and specificity.

What's Going On? Part I



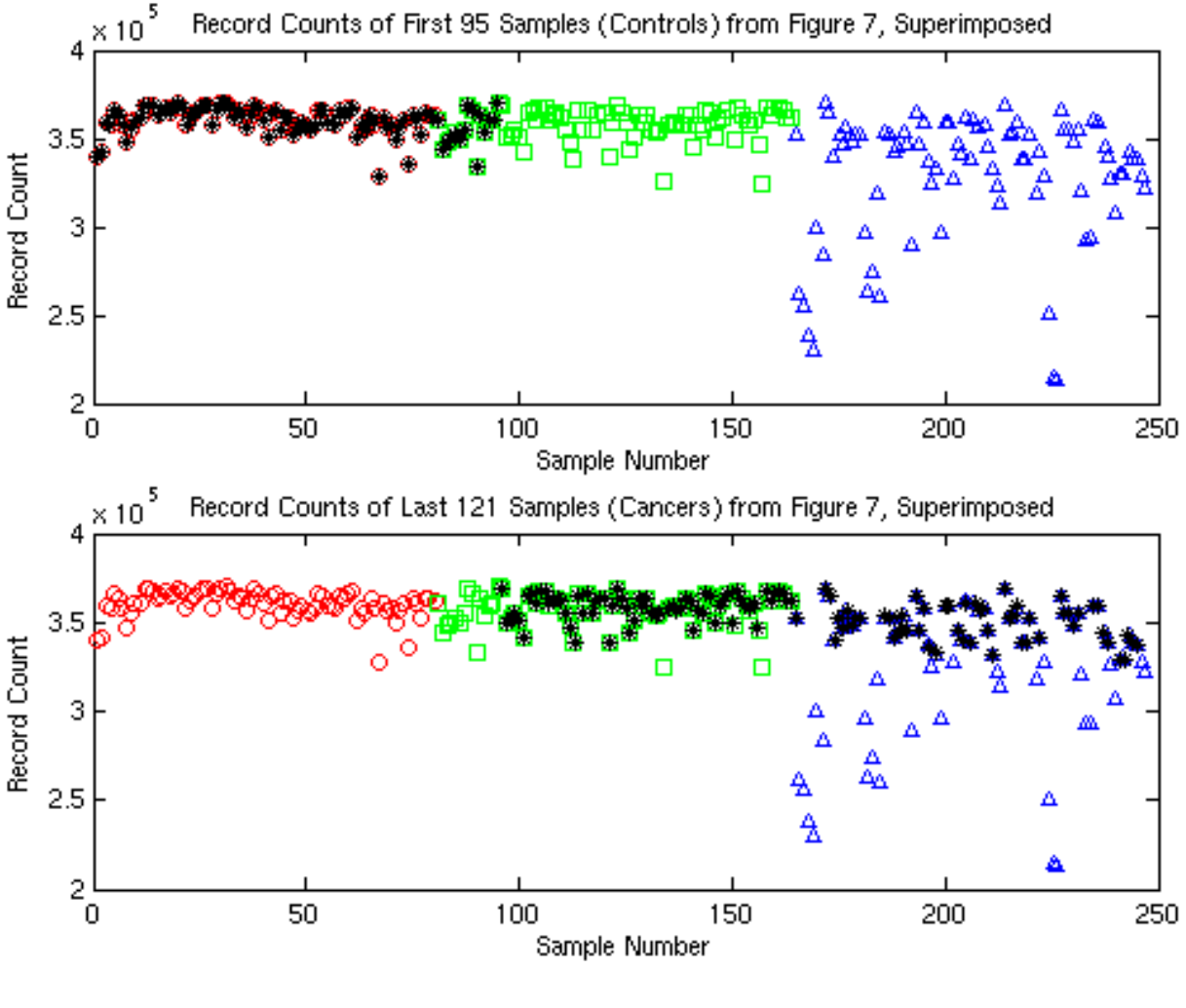
Conrads et al, ERC (Jul '04), Fig 6a

What's Going On? Part II



Conrads et al, ERC (Jul '04), Fig 7

What's Going On? Part III



Conrads et al, ERC (Jul '04), Fig 6a & 7

Meanwhile...

In January 2004, Correlogic, Quest Diagnostics and Lab Corp announced plans to offer a “home brew” test called **OvaCheck**: samples would be sent in by clinicians for diagnosis.

Estimated market: 8 to 10 million women. Estimated cost: 100-200 dollars/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)

Are Things Better Now?

New York Times, 2.3.04

New Cancer Test Stirs Hope and Concern

By ANDREW POLLACK

Jill Doimer's mother died in 2002 from ovarian cancer, detected too late to be effectively treated.

So Ms. Doimer is eagerly awaiting the introduction of a new test that holds the promise of detecting early-stage ovarian cancer far more accurately than any test available now, using only blood from a finger prick.

Not only does she plan to be tested, but an advocacy group she helped found, Ovarian Awareness of Kentucky, also intends to spread the word to women and doctors.

"If it's going to happen to me or anyone I know, I want it to be caught at an early stage," said Ms. Doimer, who lives in Louisville.

The new test, expected to be available in the next few months, could have a big effect on public health if it works as advertised. That is because when ovarian cancer is caught early, when it is treatable by surgery, more than 90 percent of women live five years or longer. But right now, about three-quarters of cases are detected after the cancer has advanced, and then only 35 percent of women survive five years.

The test is also the first to use a new technology that some believers say could revolutionize diagnostics. It looks not for a single telltale protein — like the prostate-specific antigen, or P.S.A., used to diagnose prostate cancer — but rather for a complex fingerprint formed by all the proteins in the blood. Similar tests are being developed for prostate, pancreatic, breast and other cancers. The technique may work for other diseases as well.

"I've been in cancer research for 40 years and I think it's the most important breakthrough in those years," said Dr.

Continued on Page 6

Cancer Test For Women Raises Hope, And Concern

By ANDREW POLLACK

A new blood test aimed at detecting ovarian cancer at an early, still treatable stage is stirring hopes among women and their physicians. But the Food and Drug Administration and some experts say the test has not been proved to work.

New York Times, Aug 26, 2008.

Are Things Better With *Arrays*?

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

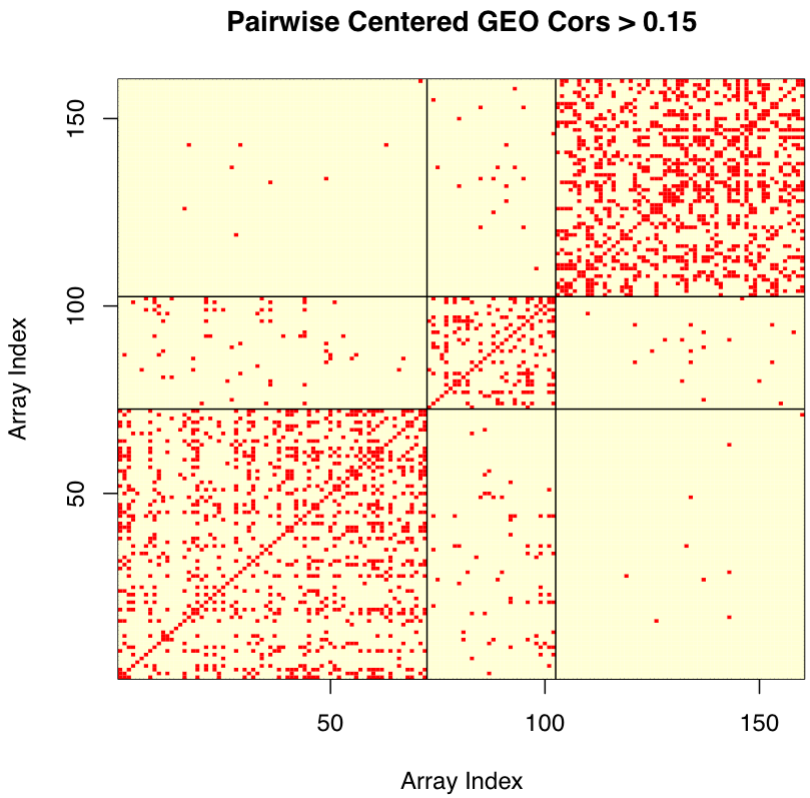
Hervé Bonnefoi, Anil Potti, Mauro Delorenzi, Louis Mauriac, Mario Campone, Michèle Tubiana-Hulin, Thierry Petit, Philippe Rouanet, Jacek Jassem, Emmanuel Blot, Véronique Becette, Pierre Farmer, Sylvie André, Chaitanya R Acharya, Sayan Mukherjee, David Cameron, Jonas Bergh, Joseph R Nevins, Richard D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

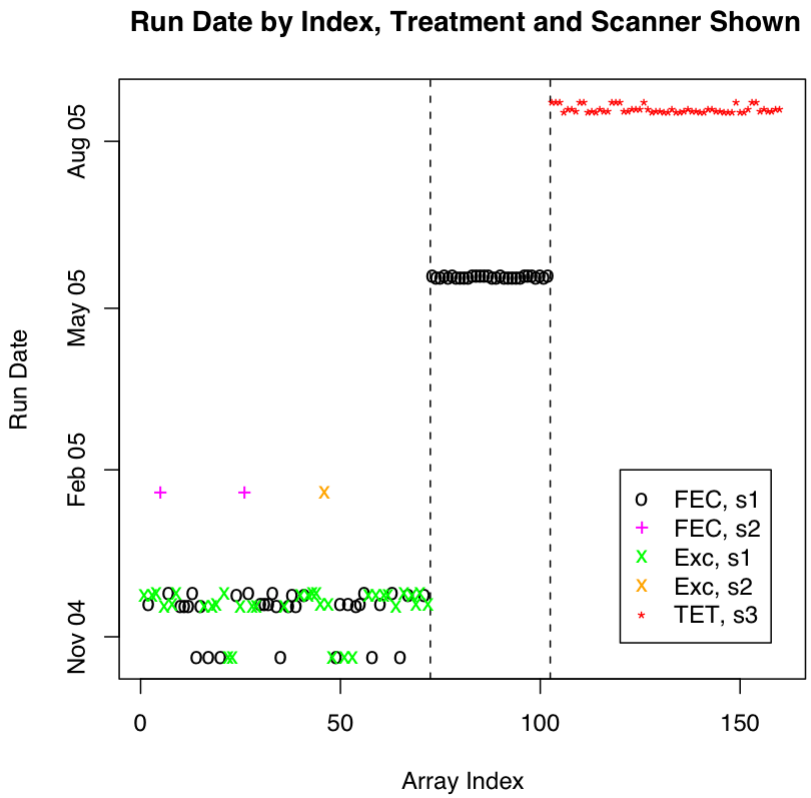
Lancet Oncology, Dec 2007, 8:1071-8. (early access Nov 14)

Similar approach, using signatures for Fluorouracil, Epirubcin (used Adriamycin), Cyclophosphamide, and Taxotere (Docetaxel) to predict response to one of two combination therapies: FEC and TET.

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

We Might Expect Some Differences...





High Sample Correlations Array Run Dates

See Leek et al, Nat Rev Gen 2010 for more examples.

Experimental Design is Important

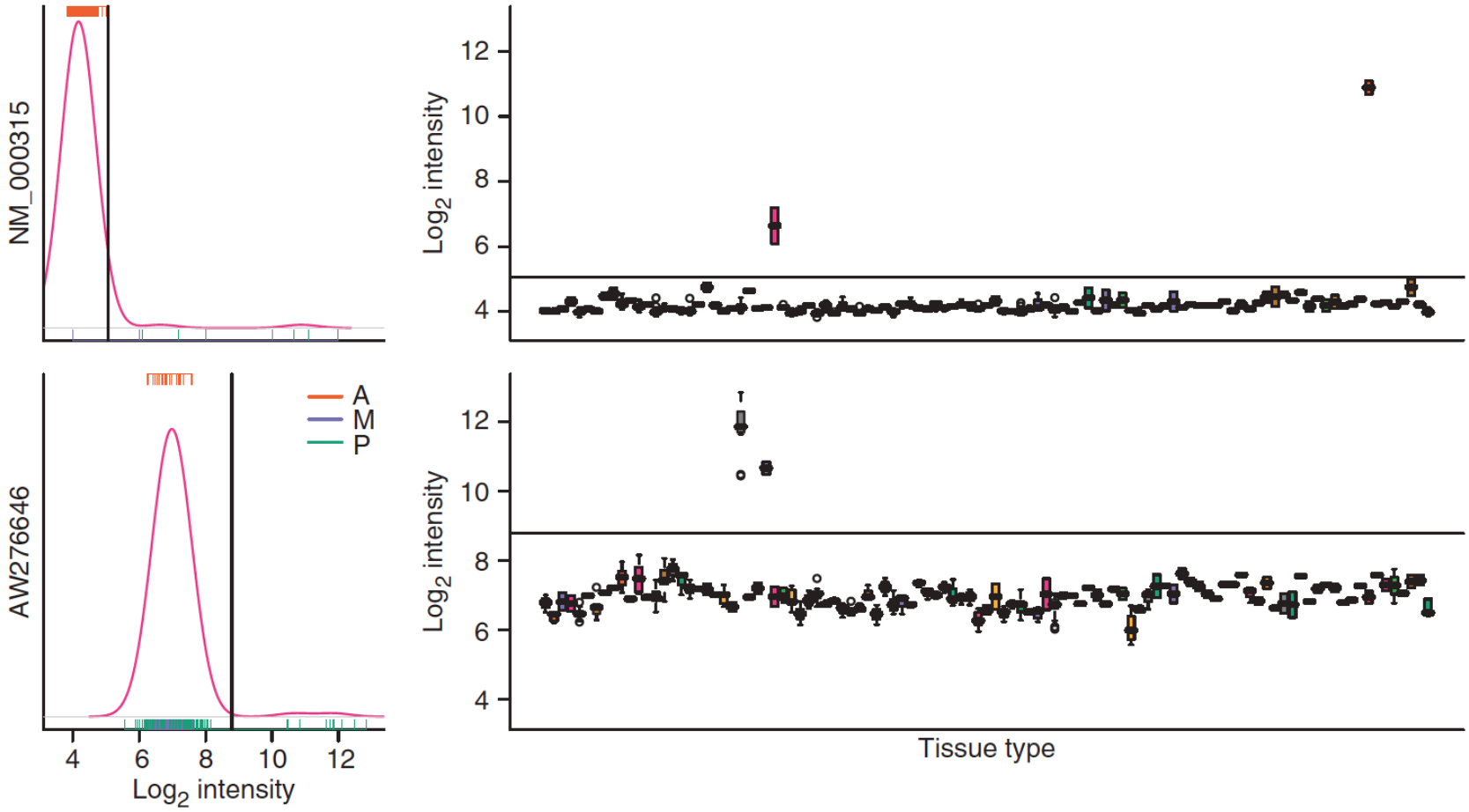
The proteomic studies were undermined by instrument drift within a single lab.

Effects *across labs* are often far larger (e.g., the array studies alluded to above)

Effects that are highly significant in one study but of modest magnitude may have problems surviving translation from lab to lab or assay to assay (use volcano plots).

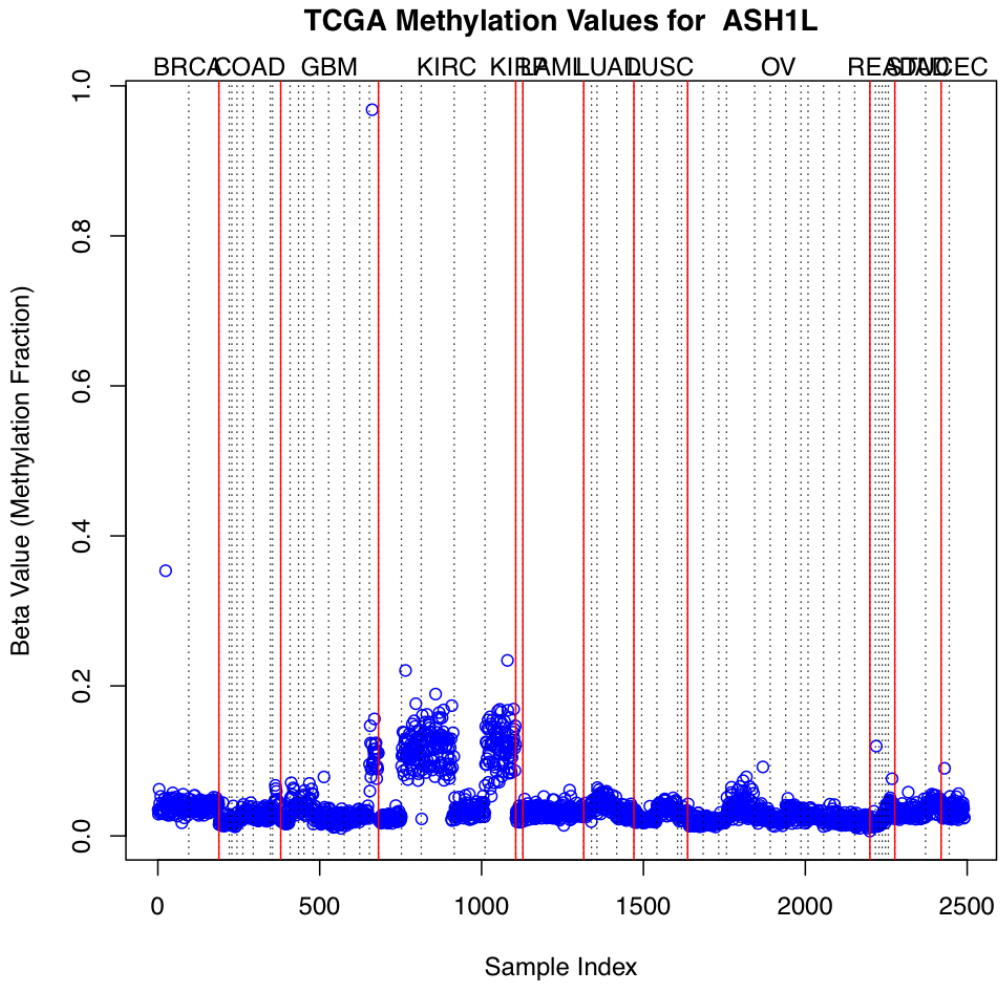
Big Data can help
by incorporating results from multiple studies

A Gene Expression Barcode



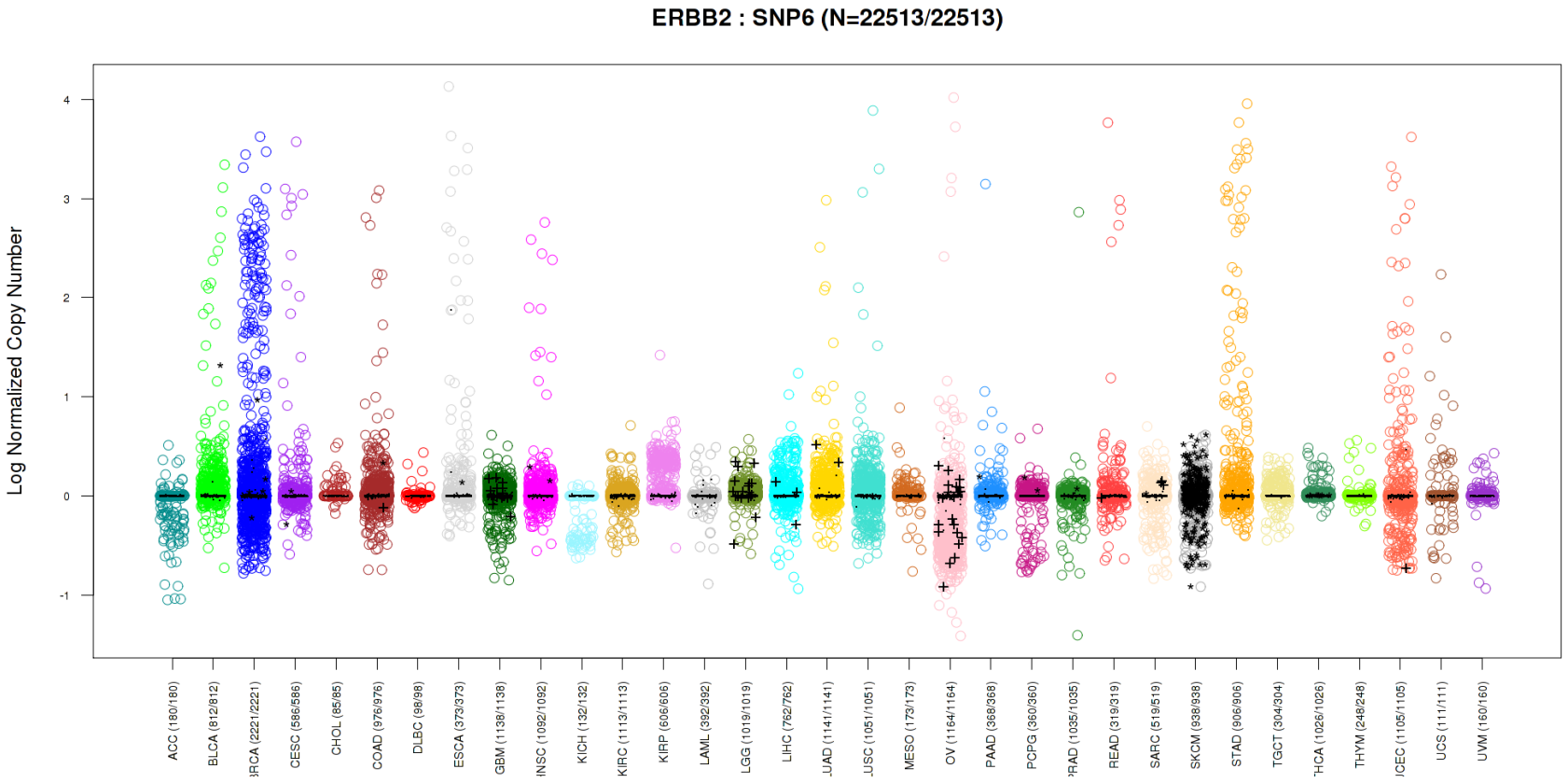
Zilliox and Irizarry (2007), *Nat Meth*, 4(11):911-3.

The Jabberbatch in TCGA



TCGA expression data in 2010

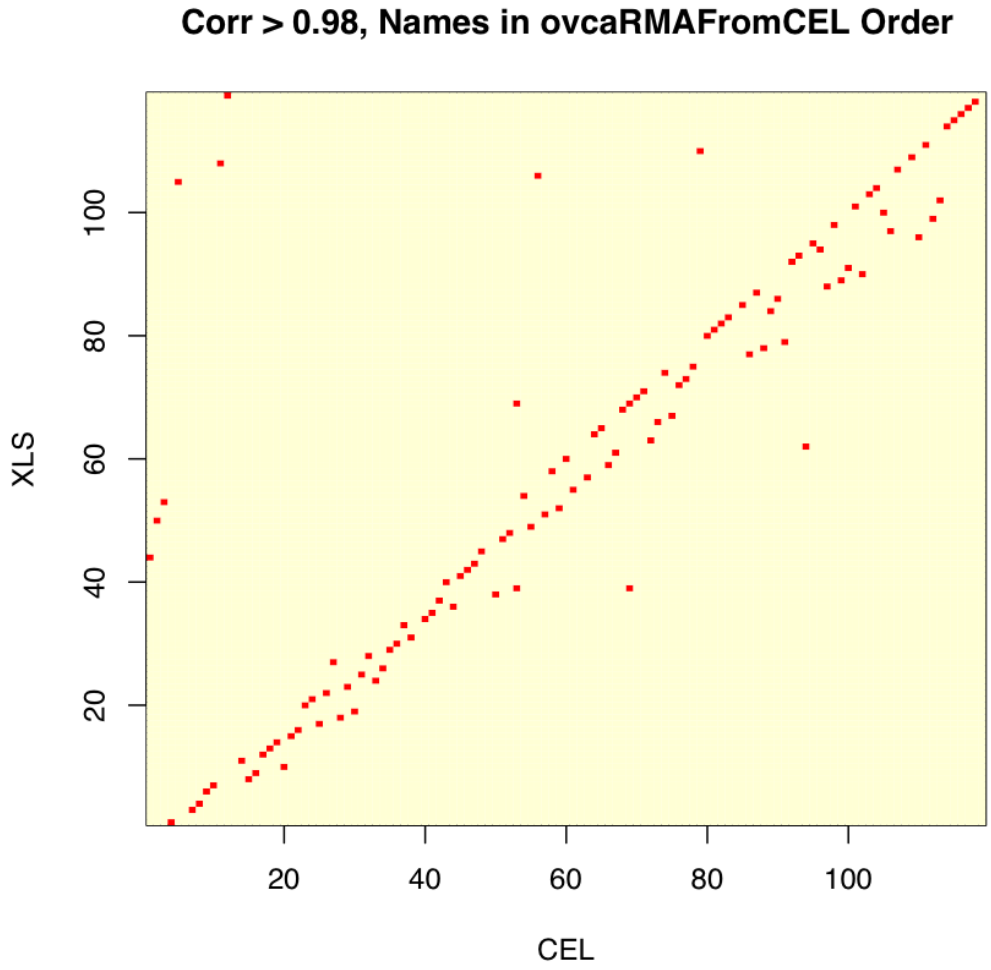
Can We Tell What's Big from a Big Study?



HER2 (ERBB2) copy number and Breast Cancer

Of course, all of this is predicated on another key assumption...

Is the Data What We Think it is?



Dressman et al, JCO, Feb 10, 2007.

In the Beginning was HeLa...

Reproducibility: changing the policies and culture of cell line authentication

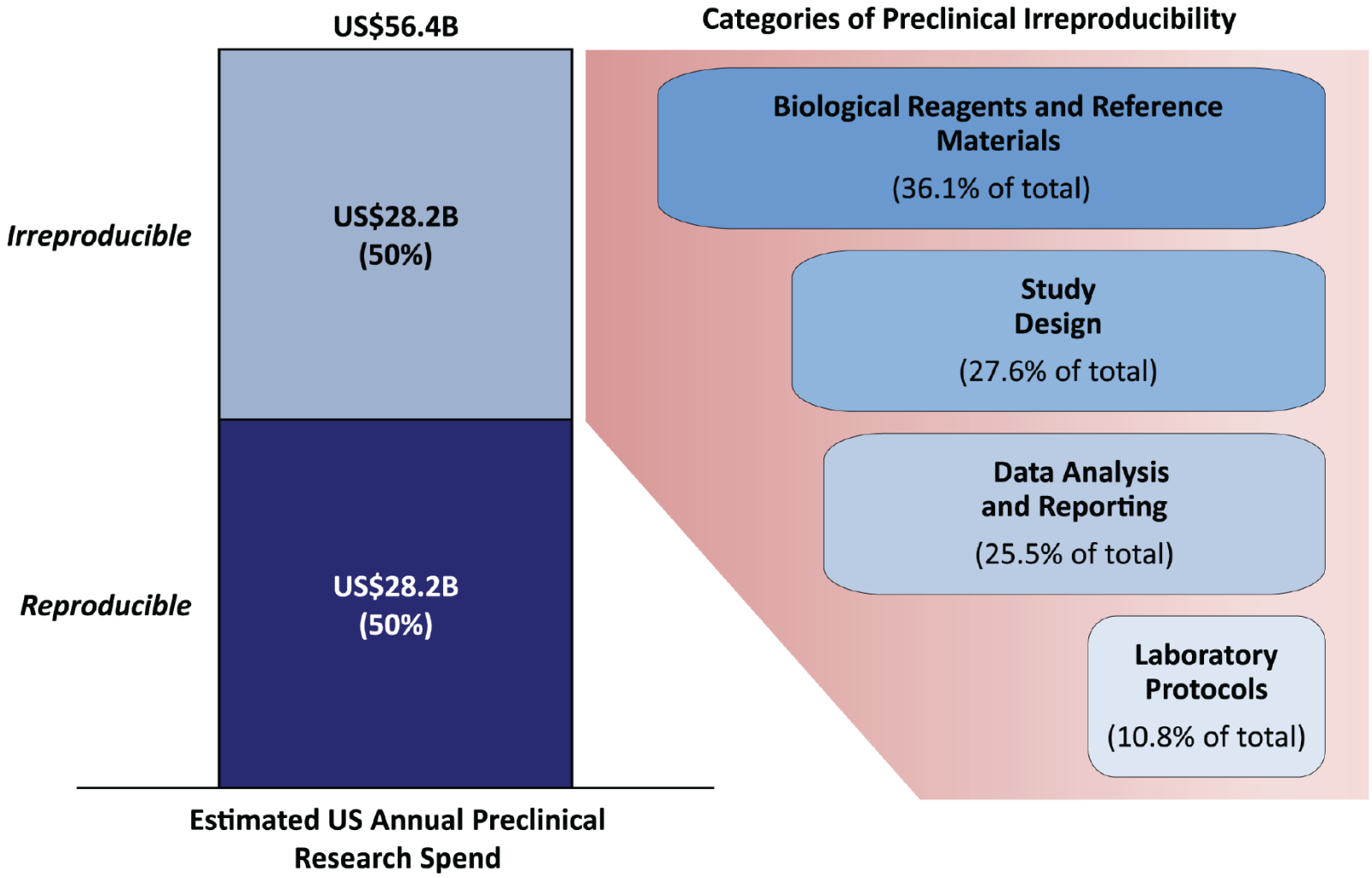
Leonard P Freedman[1], Mark C Gibson[1], Stephen P Ethier[2], Howard R Soule[3], Richard M Neve[4] & Yvonne A Reid[5]

Table 1 | Select reports of misidentified or cross-contaminated cell lines by major cell repositories

Cell type	Total number of lines	Number of false cell lines	Percentage of false cell lines	Ref.
Lymphoma, leukemia	550	82	15	39
Ovarian cancer	51	15	30	40
Adenoid cystic carcinoma	6	6	100	41
Thyroid cancer	40	17	43	42
Head, neck cancer	122	37	30	43
Esophageal adenocarcinoma	14	3	21	44
Total	783	160	20 (average)	

Freedman et al (2015), *Nat Meth*, 12(6):493-7.

Some Cost Breakdowns



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

Might Simple Tests Help?

In examining the data and experimental context, can we prespecify some things we should and shouldn't see?

In looking at panels of cell lines, the data should cluster by tissue type. (We know how to resolve cell line identity, if people would just check!)

The correlation tests above are also tests of this type.

For expression values of genes flagged as interesting *plot them by run date*.

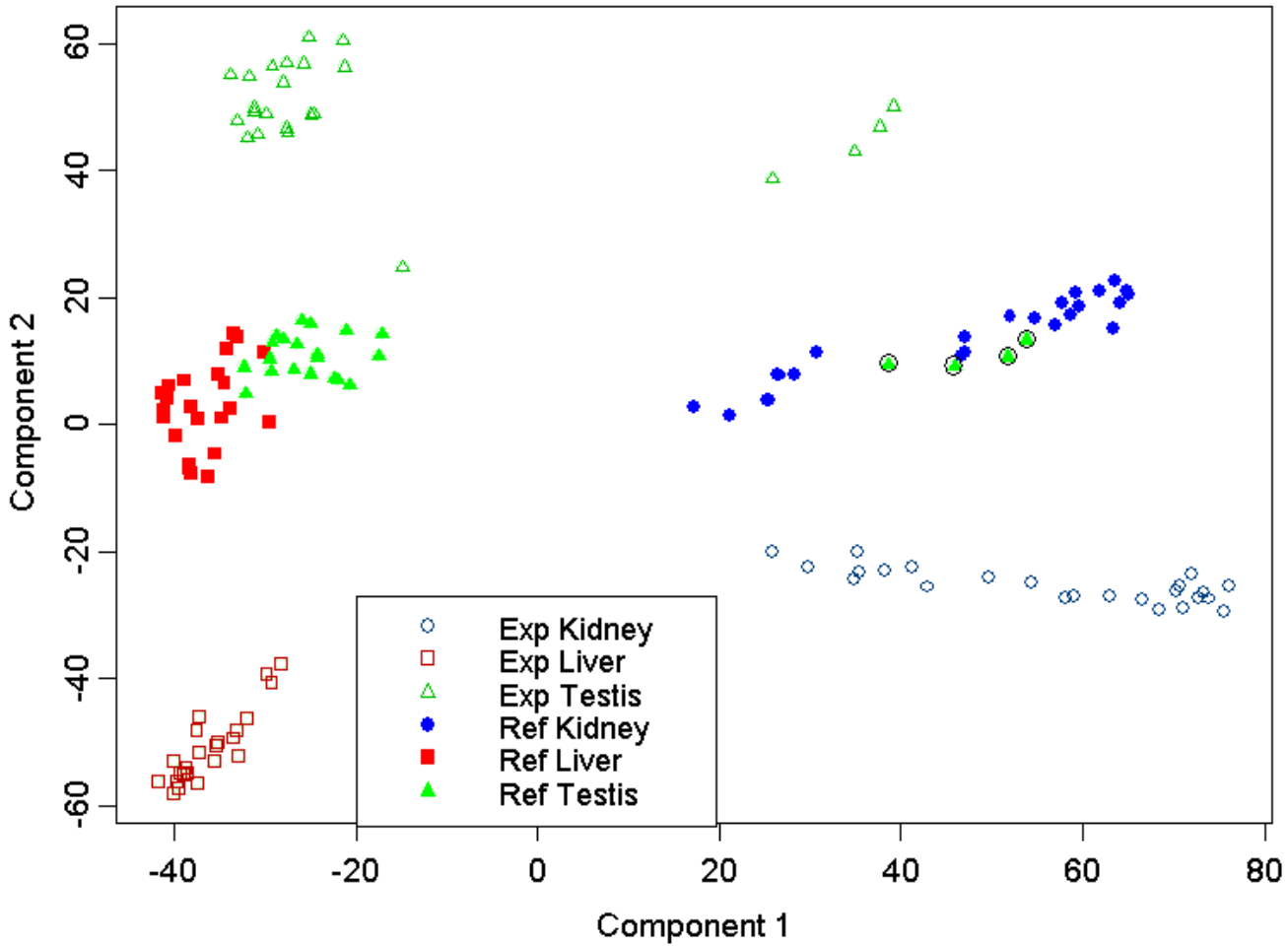
A Thought Experiment

Say you're measuring samples from
three organs, and a pool of all three
as a reference.

You run two color arrays with dye swaps, 24 samples per tissue * 2 labels * 3 tissues = 144 arrays.

If you summarize all the data by sample (there are 288), how many clusters would you expect?

When Bad Things Happen to Good Data



Stivers et al, CAMDA 2002

More Positive and Negative Controls

Prespecify

Before we analyze the data, can you write down 5-10 genes that should change, and directions they should change in?

Once you've written them down, give me *half* the list.

Try things that shouldn't work

Negative controls can often be generated by *label scrambling*.

Is the number of differences between two classes much bigger than the number we find when we randomly allocate samples to "group1" or "group2"?

Summary

Poor replication is a big problem

The biggest contributors are things we can avoid

Many problems can be detected by application of straightforward quality control tests

The best time to think of these questions is very early in the process!