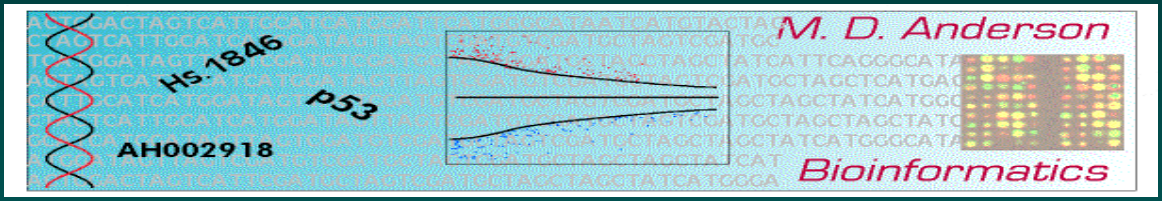
Why is Replicability Hard?

Keith A. Baggerly
Bioinformatics and Computational Biology
kabagg@gmail.com

FDA, Nov 16, 2018



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?

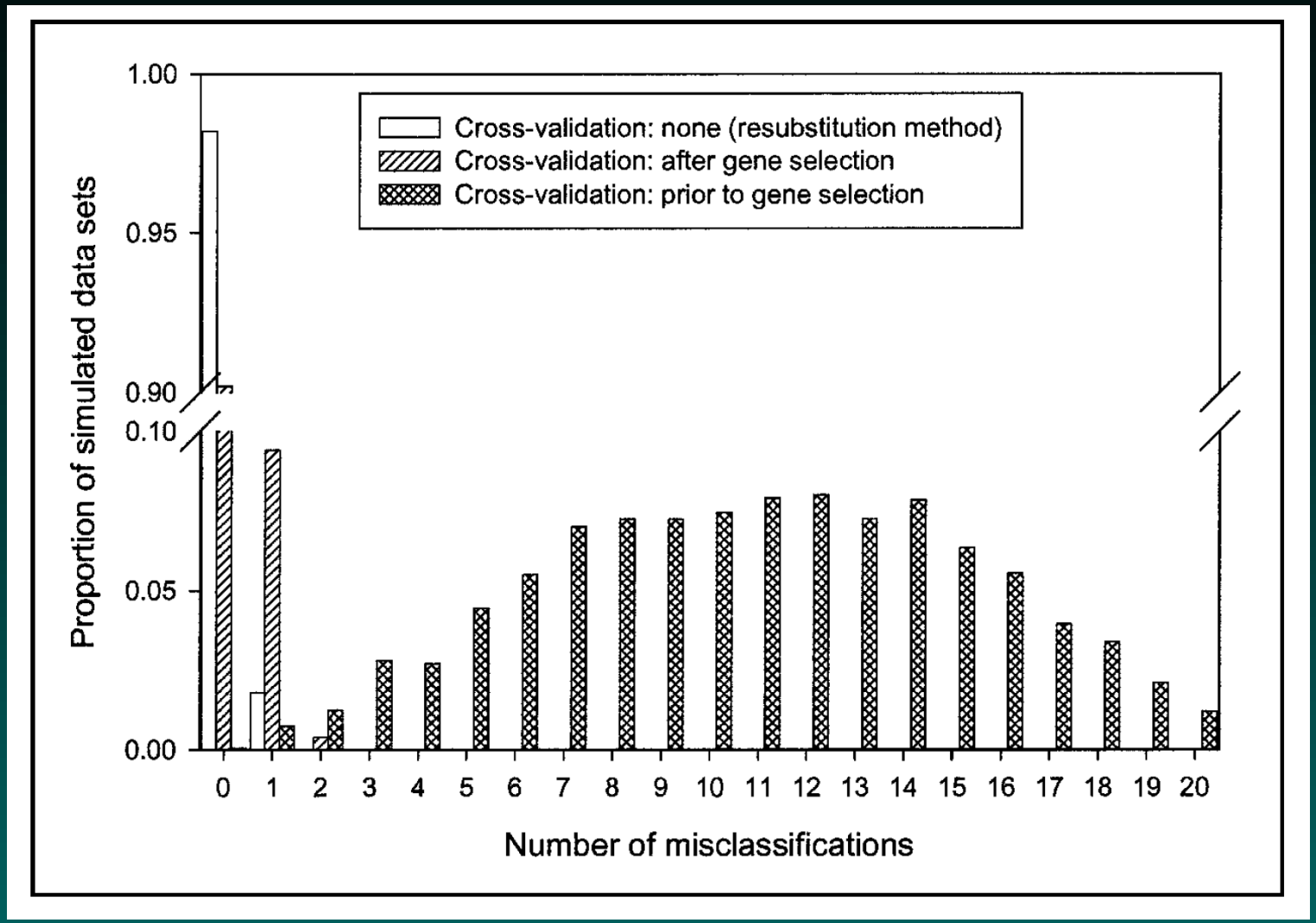
Were significance levels adjusted for multiple testing?

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...

Big Data Can Help w Experimental Design

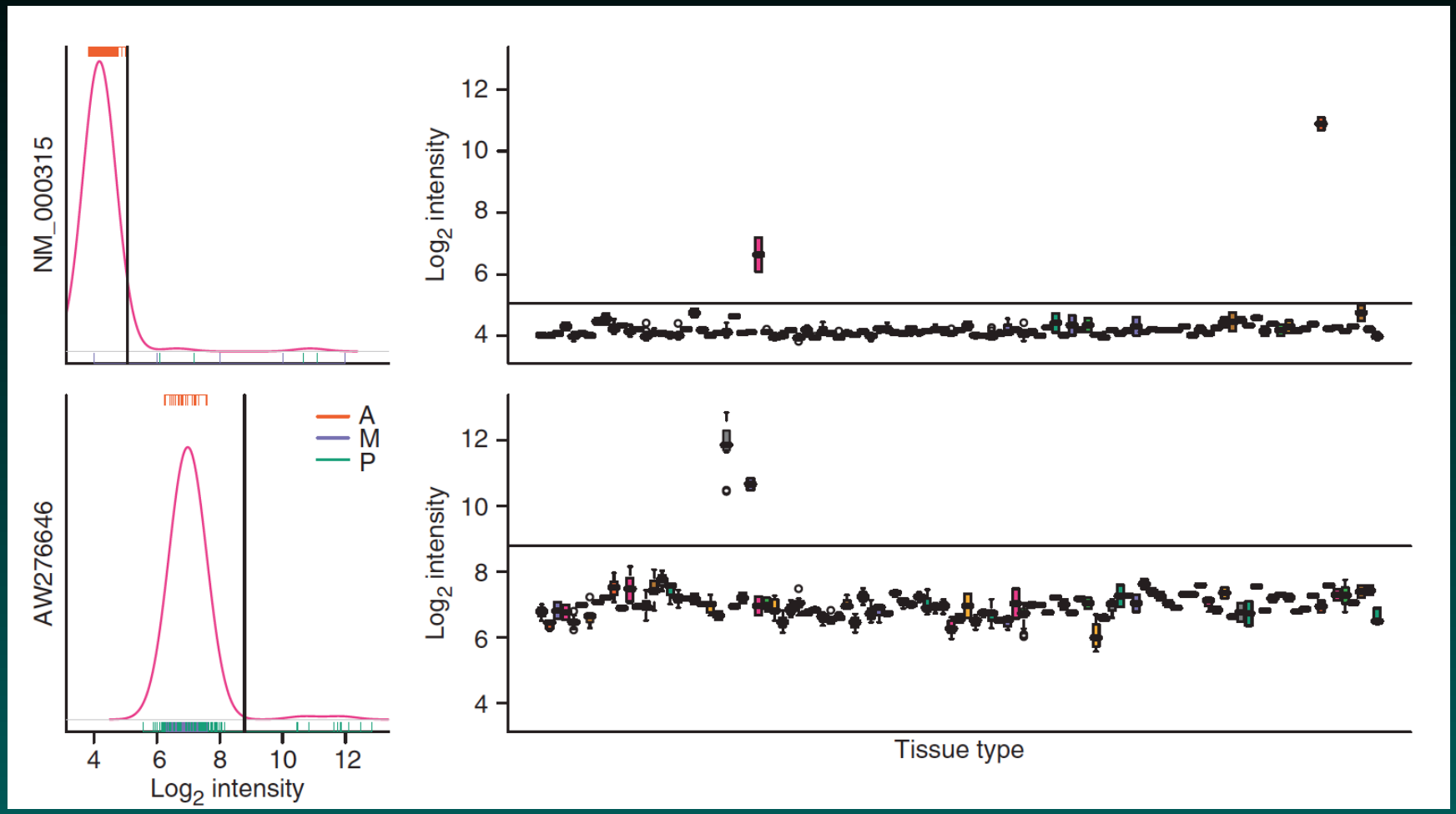
The proteomic studies were undermined by instrument drift within a single lab.

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).

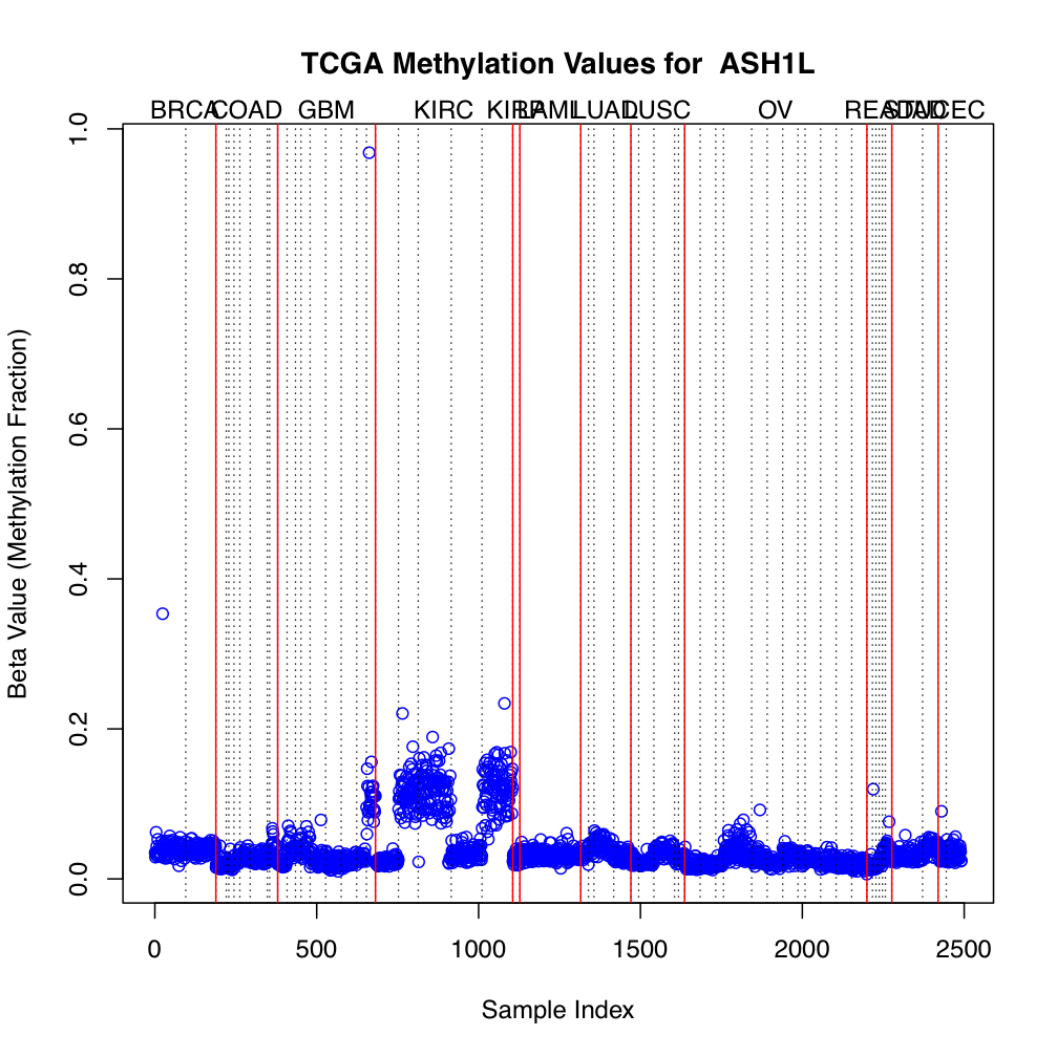
Big Data can help
by incorporating results from multiple studies

A Gene Expression Barcode



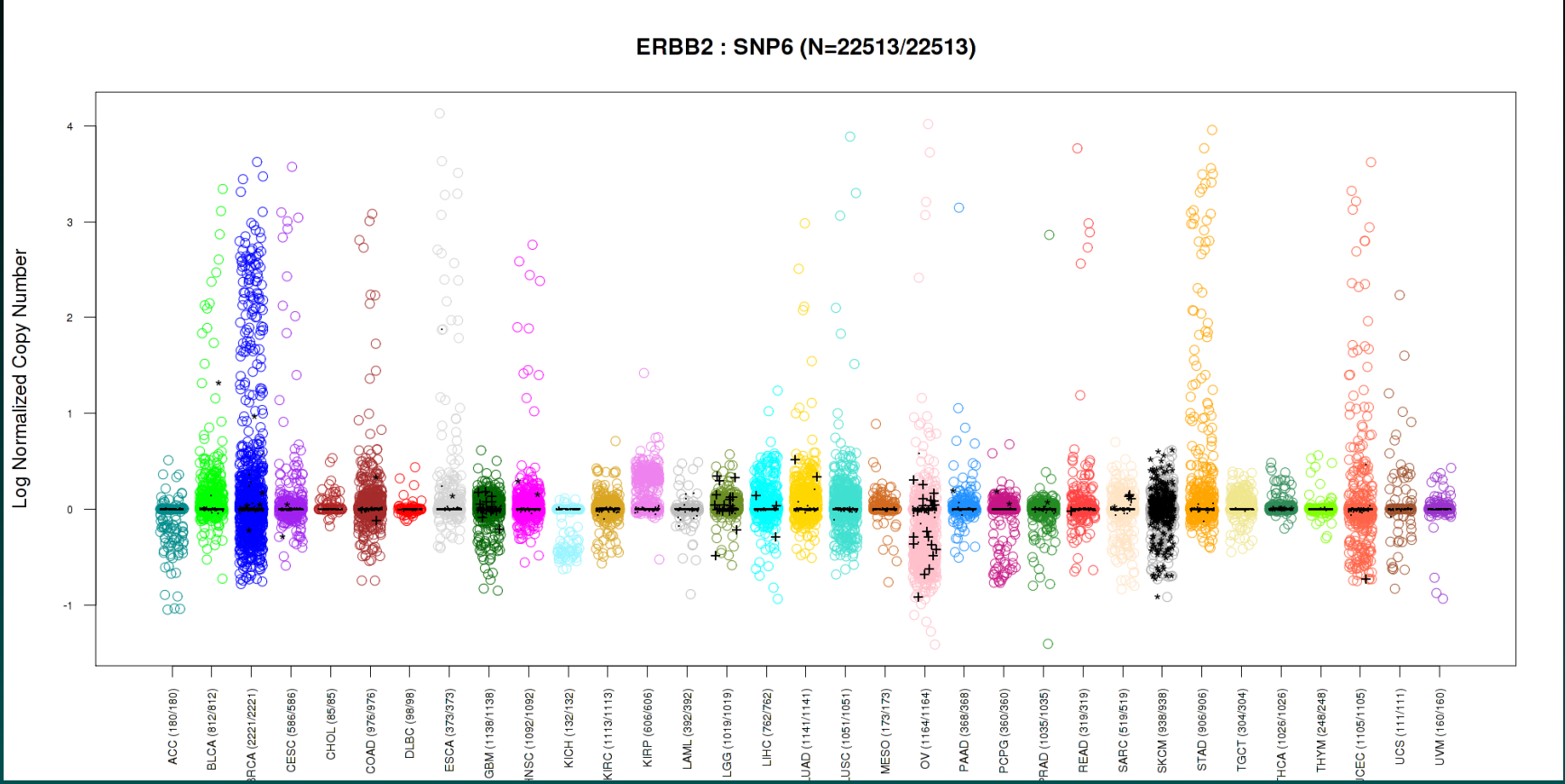
Zilliox and Irizarry (2007), *Nat Meth*, 4(11):911-3.

The Jabberbatch in TCGA



TCGA expression data in 2010

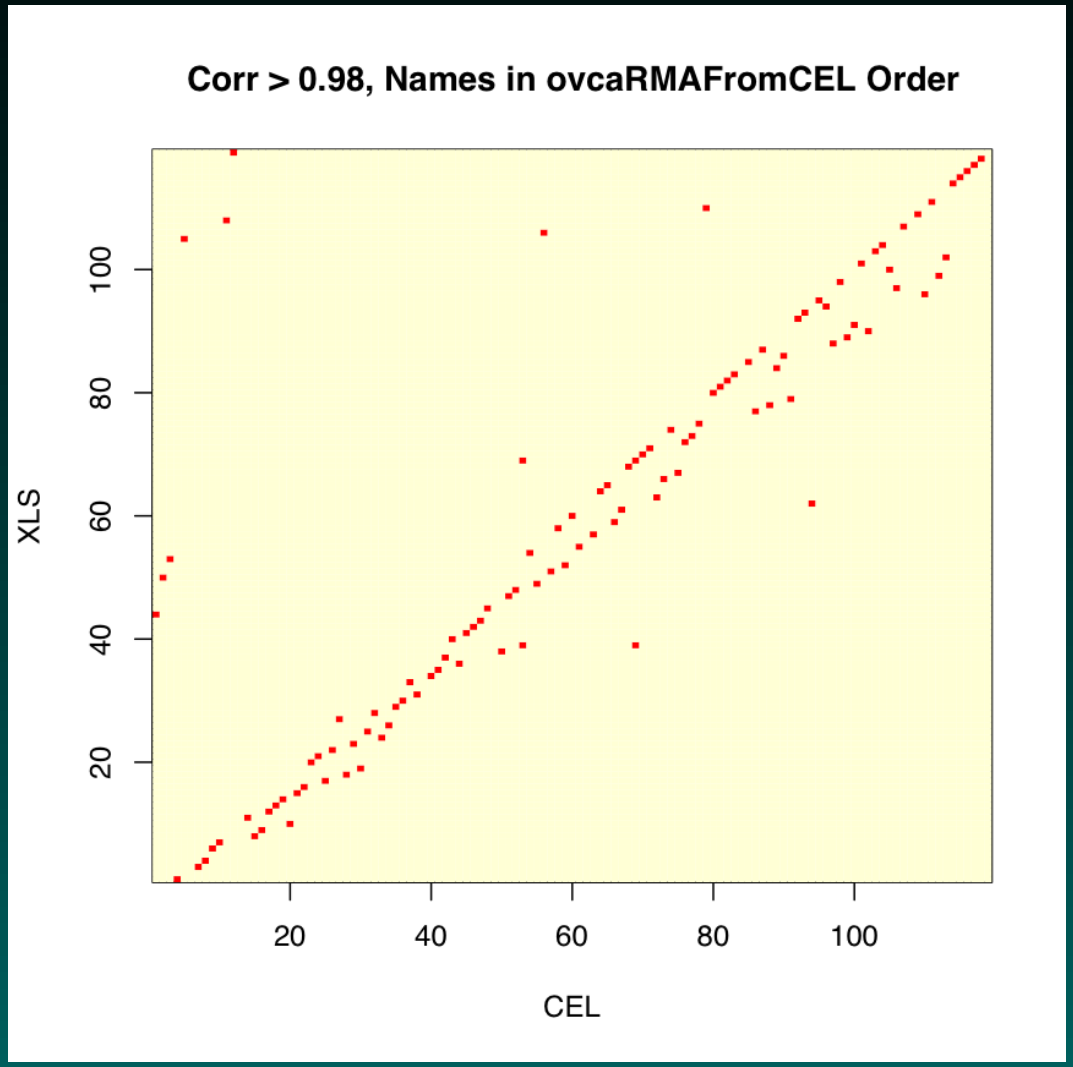
Can We Tell What's Big from a Big Study?



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.

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*.

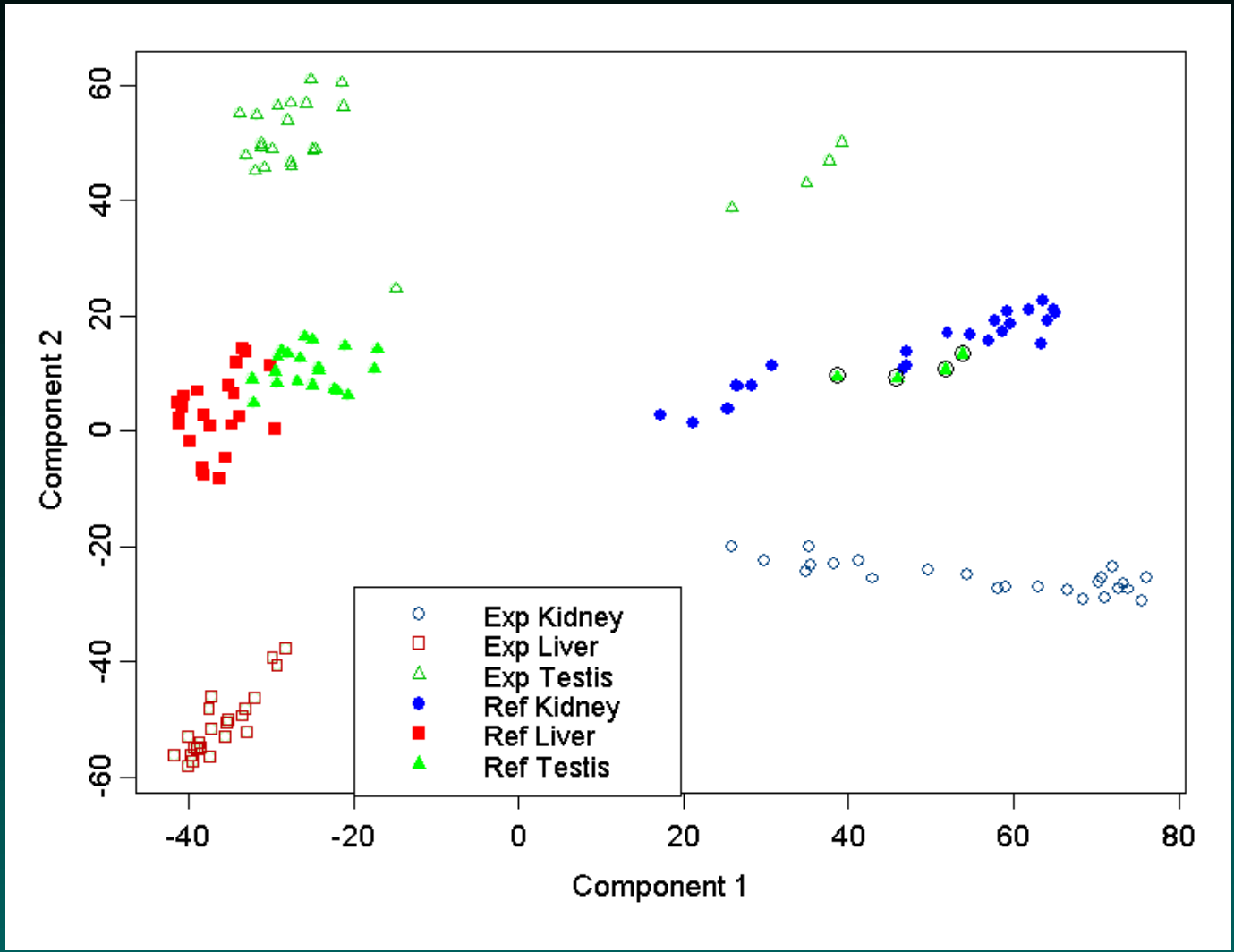
A Thought Experiment

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

The Proteomics Data Mining Competition

Let's try another competition!

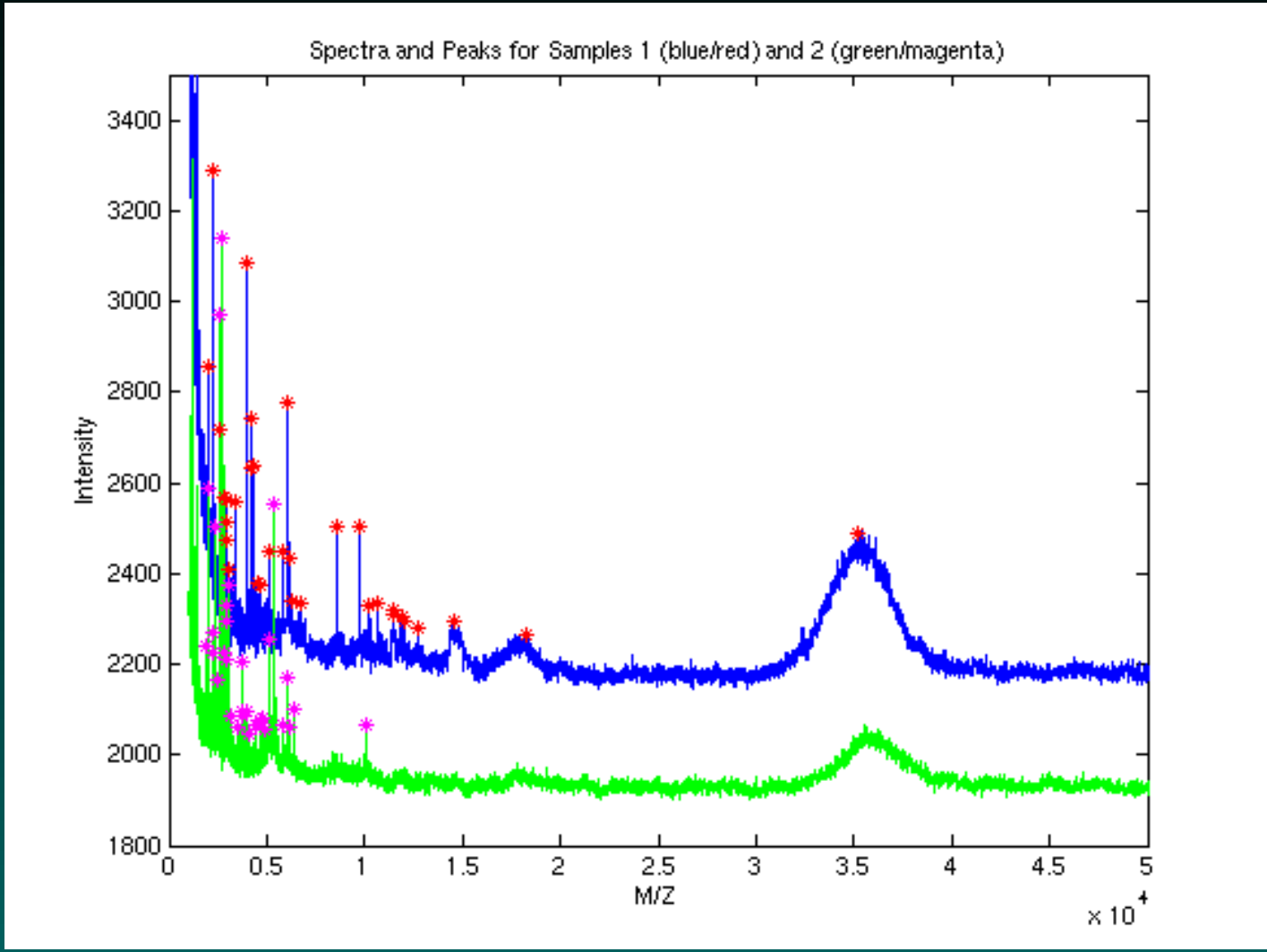
41 samples, 24 with disease*, 17 controls.

20 fractions (spectra) per sample.

Goal: distinguish the two groups.
We know this can be done due to the "zip effect".

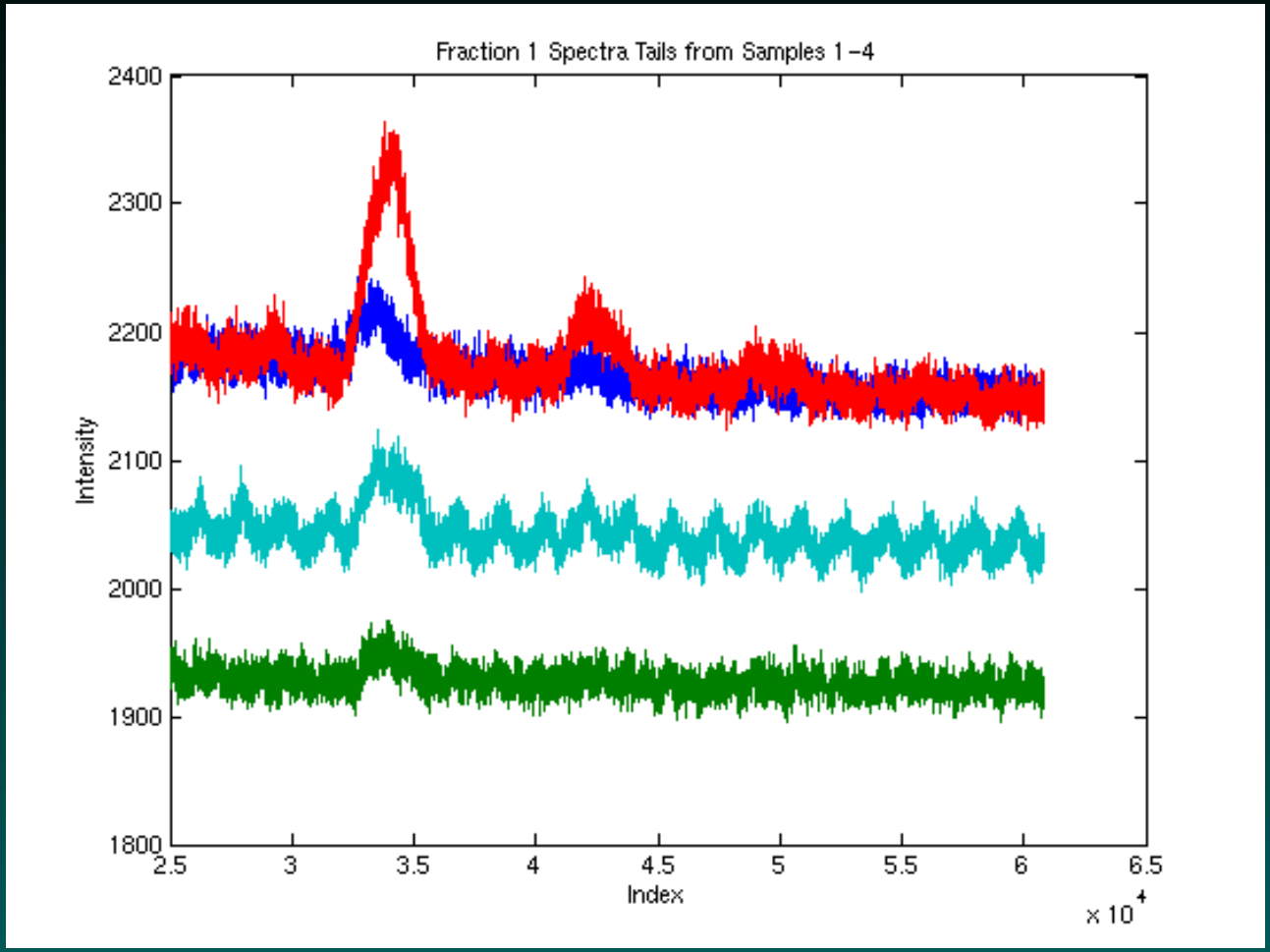
(We also knew what the disease was.)

Raw vs Processed – Use Raw



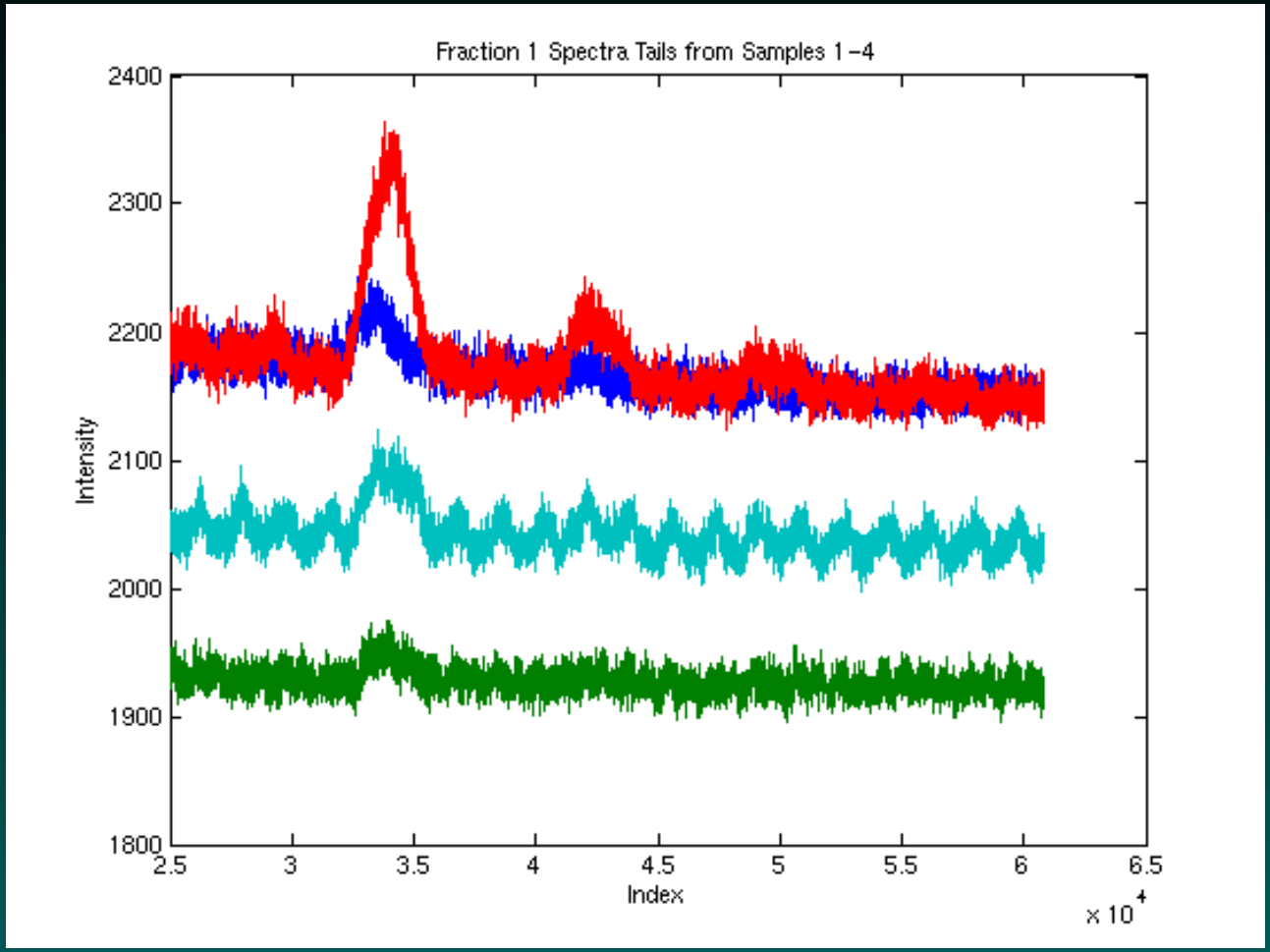
Note the need for baseline correction!

Oscillatory Behavior...



Half the spectra are “wiggly”!

Oscillatory Behavior...



Half the spectra are “wiggly”! It’s the A/C power cord.
So, what happened?

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!