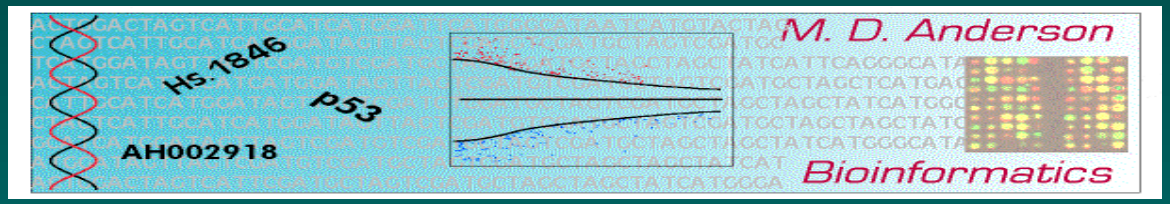
Reproducible Research in Bioinformatics

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

SAMSI, Sep 10, 2014



Why is Reproducibility Important in Genomics? With "Big Data" In General?

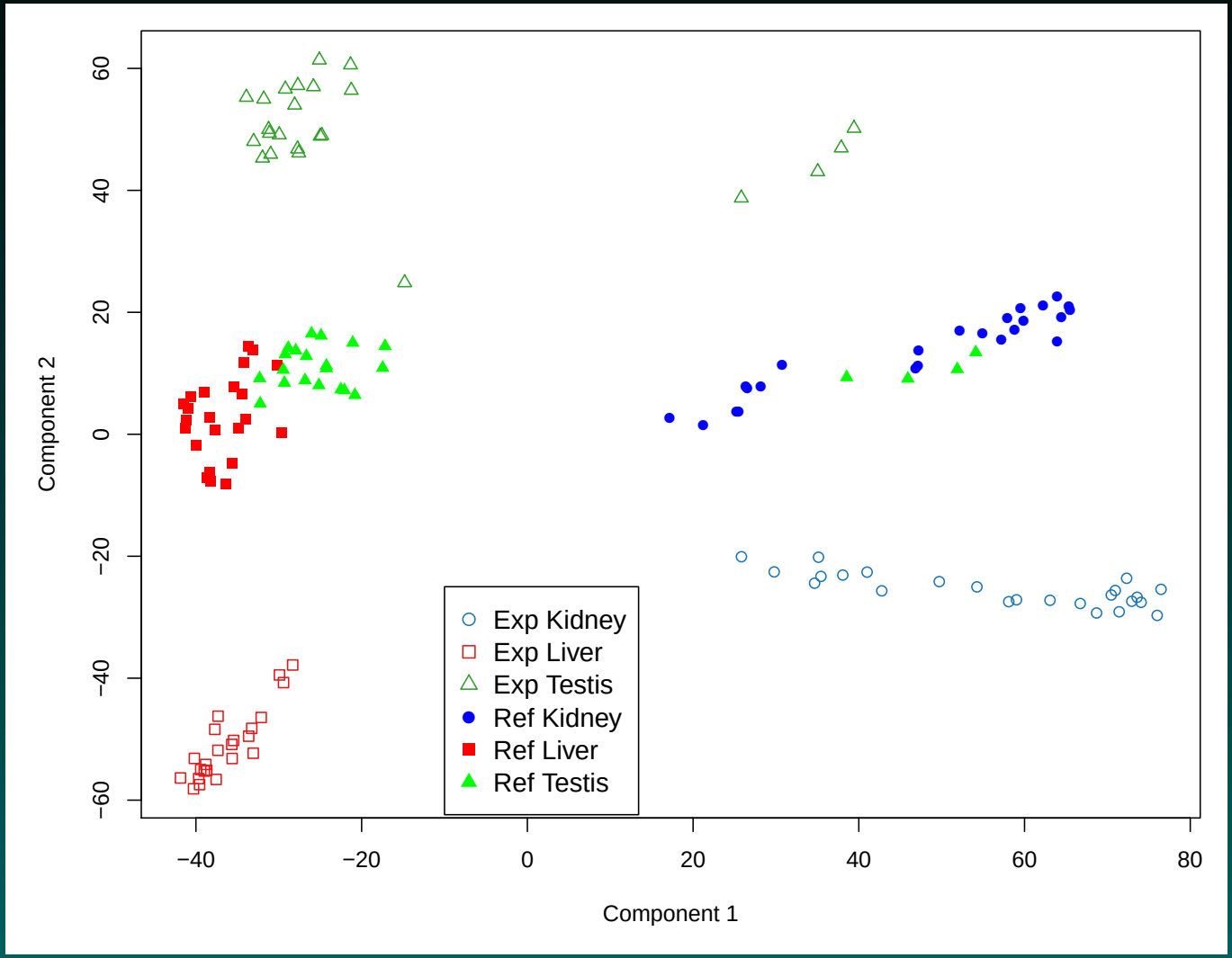
Our intuition about what "makes sense" is very poor in high dimensions.

If we want to use patterns we find in the data to direct patient care, we need to know they've been assembled correctly.

If the results aren't readily checkable, we may need to employ (lengthy!) *forensic bioinformatics* to reconstruct what was done.

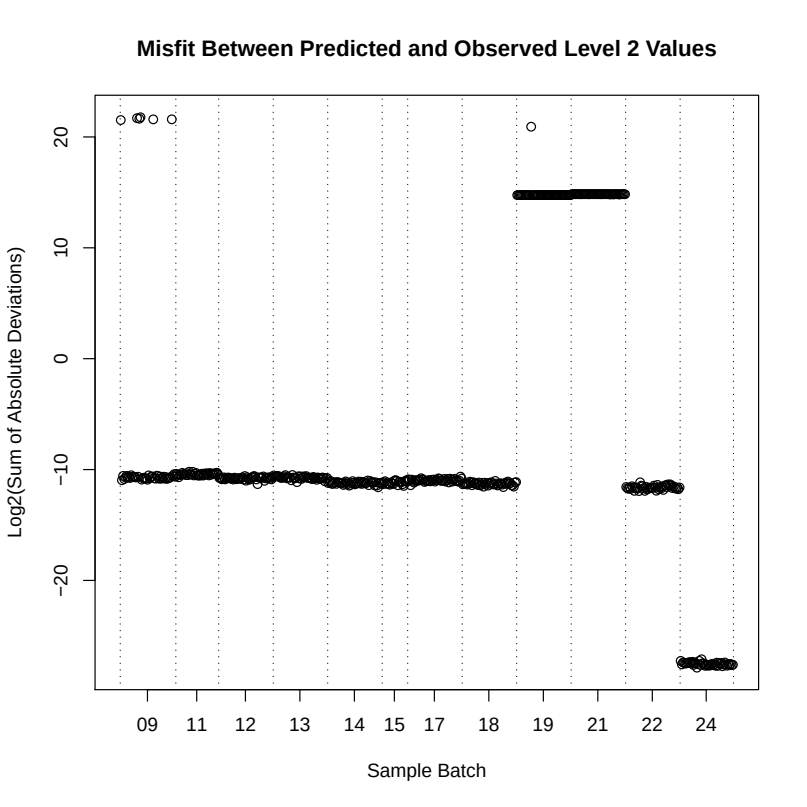
For better or worse, we've done this a lot.

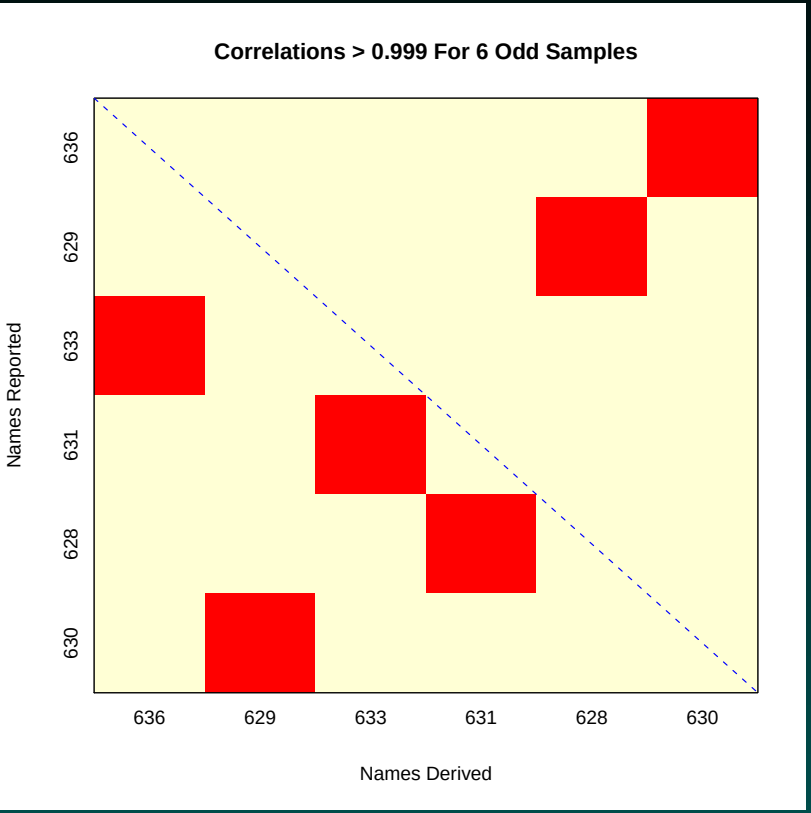
What are the Genes?: CAMDA 2002



There should be 4 clusters. Excel split two.

What are the Samples?: TCGA 2010



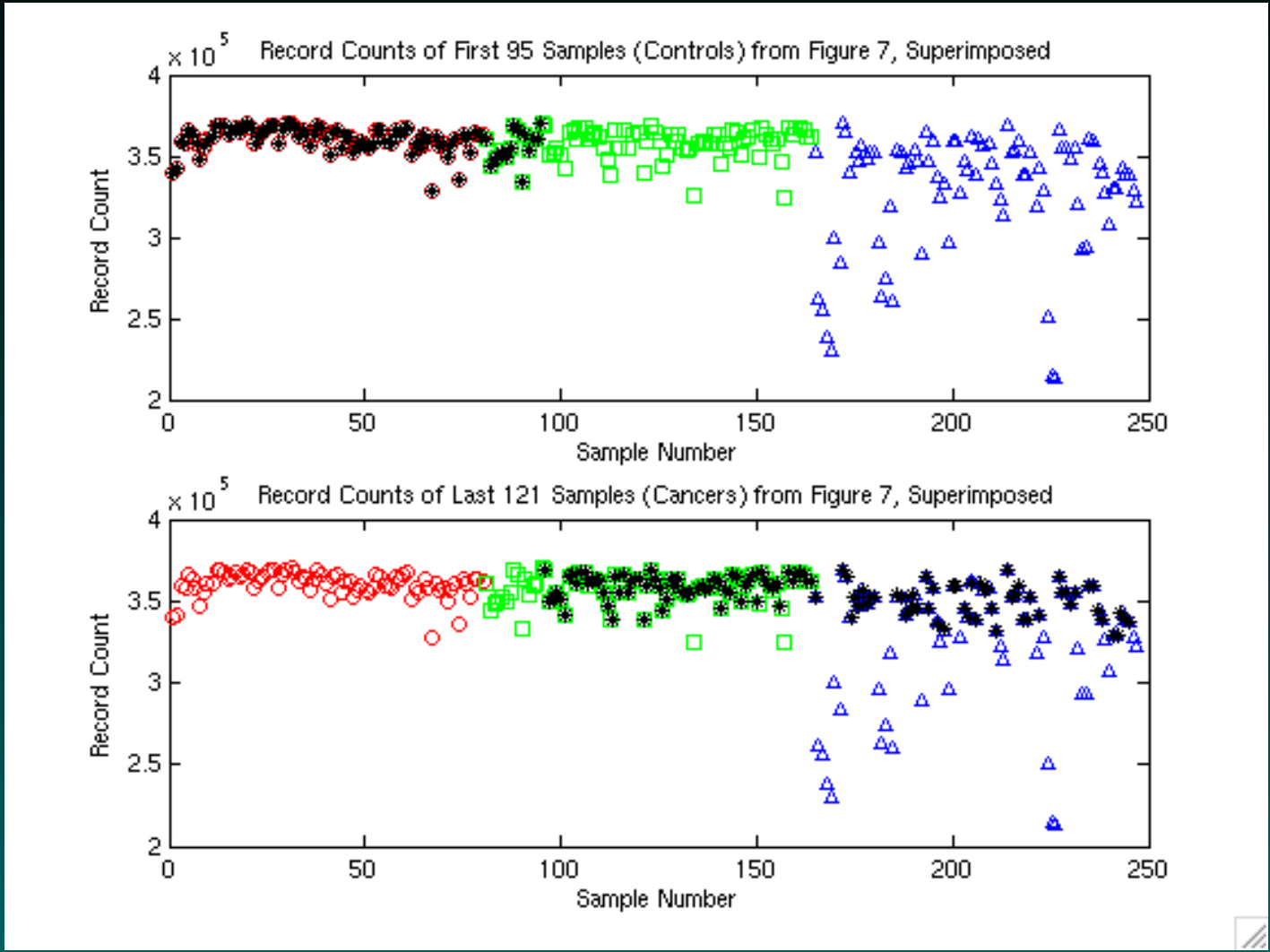


Ovarian miR Data; Level 1 to Level 2

Correlations between named and derived profiles

How were the data processed?

How was the Experiment Performed? Prot 2004



Machine breaking. All controls run first.

Are we Learning? NYT 04, 08, Nat Gen 10

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.

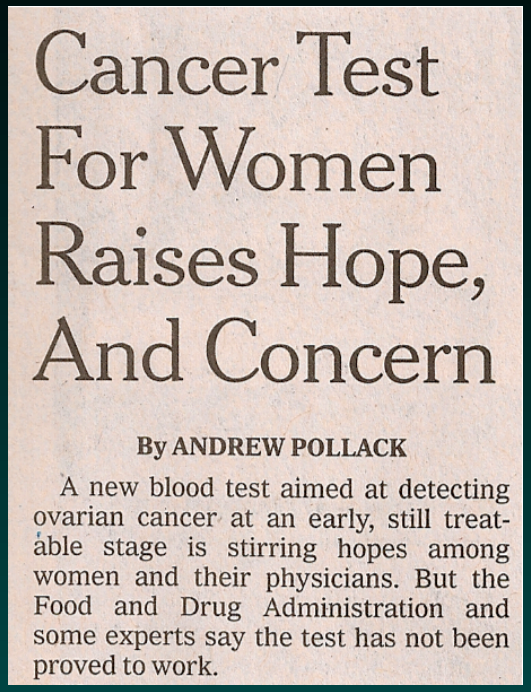
"If it's going to happen to me or anyone I know, I want it to be caught at an early stage," said Ms. Doimer, who lives in Louisville.

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.

Feb 4, 2004

Aug 26, 2008

It's not just one assay.

Leek et al. (2010), Nat Gen 11:733-9.

Batch effects are pervasive.

Is This 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 validation of clinical "breakthroughs" prior to further study. Validated 6/53.

Recent Links

Science, March 6, 2013 http://www.aaas.org/news/releases/2013/0311_alberts.shtml

Nature, April 24, 2013 http://www.nature.com/news/announcement-reducing-our-irreproducibility-1.12852

Colbert report, April 23, 2013 http://www.colbertnation.com/the-colbert-report-videos/425749/april-23-2013/austerity-s-spreadsheet-error---thomas-herndon

Nature, BMC Medicine, Oct 17, 2013 http://www.nature.com/nature/journal/v502/n7471/full/nature12564.html, http://www.biomedcentral.com/1741-7015/11/220

Everything Old is New Again...

"To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of." RA Fisher, 1938. Sankhya 4, 14-17.

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

So, How Do We Do It?

By fiat

Literate Programming

R

Sweave (R + Latex)

RStudio/knitr/markdown - very low barriers to entry!

Producing reports on a regular basis

Structuring Reports

"We have the duty of formulating, of summarizing, and of communicating our conclusions, in intelligible form, in recognition of the right of other free minds to utilize them in making their own decisions." RA Fisher, Statistical methods and scientific induction. JRSS-B, 17, 69-78, 1955.

Introduction
Data and Methods
Results
Conclusions
Appendices

Challenges

Reports vs Projects (Elizabeth!)

Freezes, Annotation builds

Sampling when Reproduction is Infeasible (Susan Holmes)

Data Volumes

Where Can We Learn More?

Free course at Coursera

https://www.coursera.org/course/repdata

Data Scientist's Toolbox

https://www.coursera.org/course/datascitoolbox

Yihui Xie's book on knitr, and his presentation to the LA R User Group from May 2014

https://www.youtube.com/watch?v=2yvW0O_7xOg

Christopher Gandrud's book

GitHub