Replicability in Science: 2: The National Academies 2019 Report

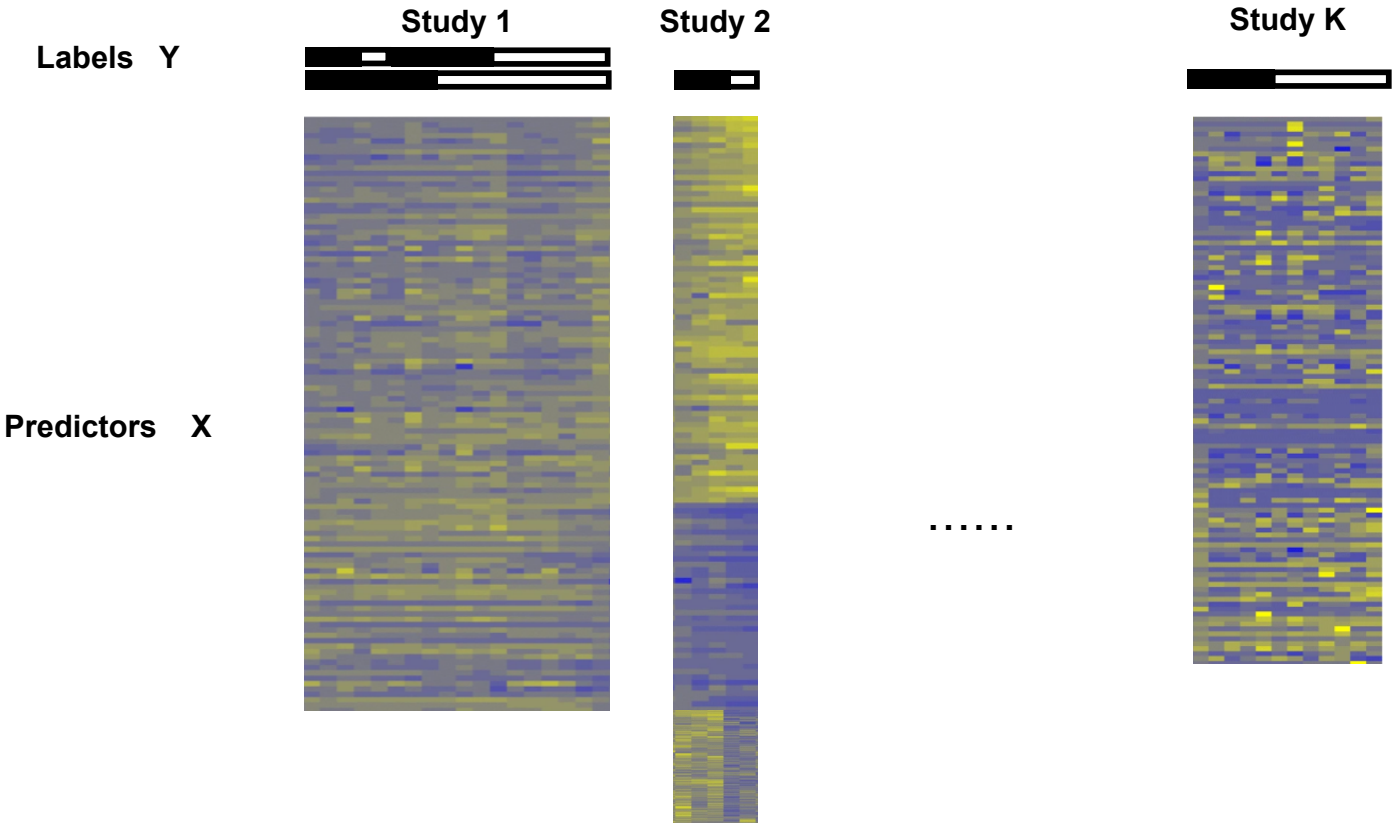
giovanni_parmigiani@dfci.harvard.edu

Padova, September 19, 2023

outline

- Left-overs
- Discussion 1: Who celebrates
- Discussion 2: Apples and CDs
- Extent of Non-Replicability in Research
- Sources of Non-Replicability in Research
- Improving Replicability in Research

data structure



False Discovery Rates

False Discovery Rates

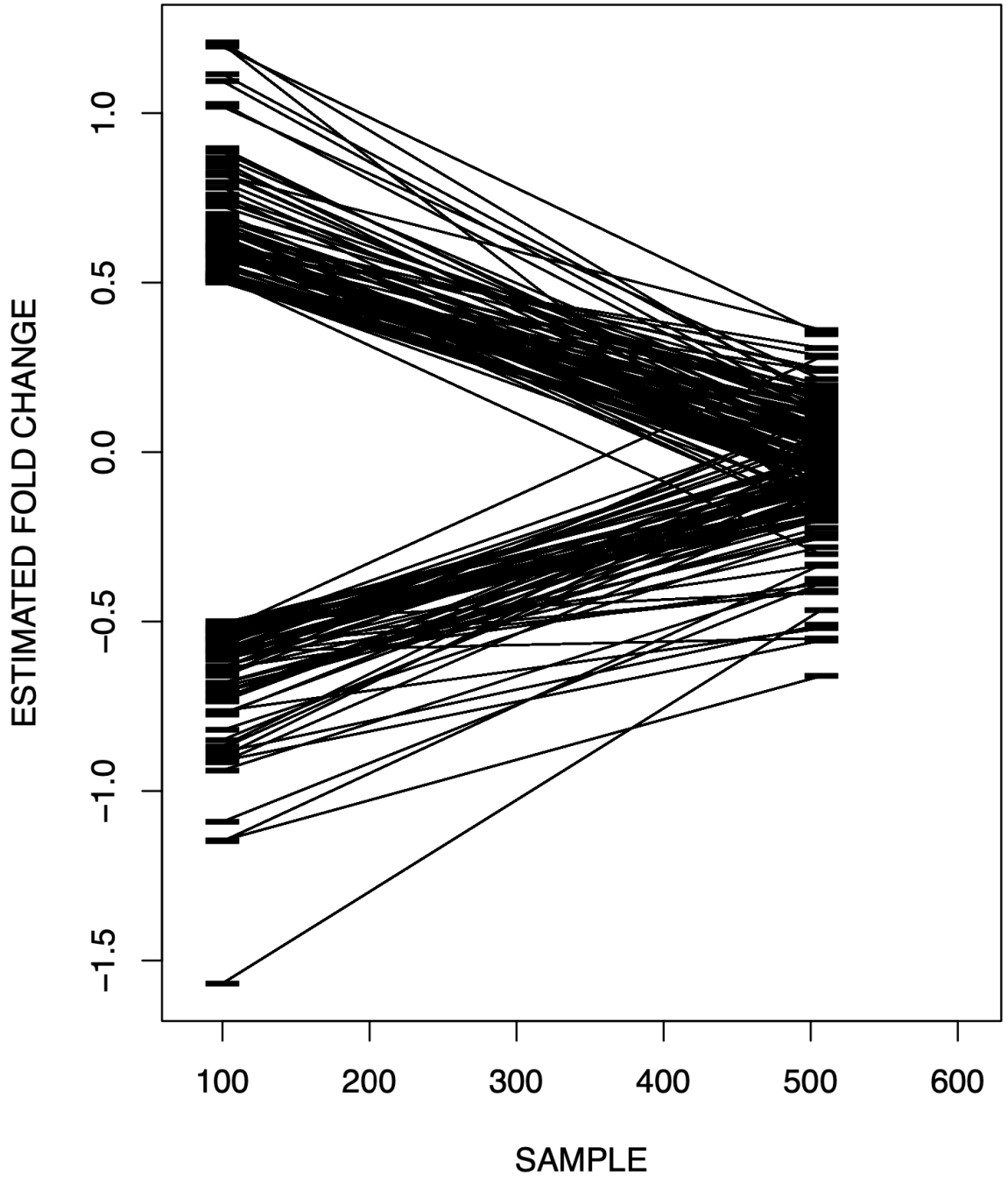
A simple way to think about the discovery process is to focus on a list of promising candidates (or "discoveries"), as we did in our previous lecture. If we know the true data generating model for each gene, we could, for example, look at this table:

	No Discovery	Discovery	Total
Null hypotheses	U	R_0	G_0
Alternative hypotheses	T	R_A	$G - G_0$
Total	$G - R$	R	G

It is common to study the proportion R_0/R of discoveries for which the true data generating model is null.

> **Note**
>
> R_0/R is unknown. It depends on both the data and the parameter (the vector of indicators of whether eagh gene is null)
>
> - When is R_0/R an appropriate quantity to focus on?

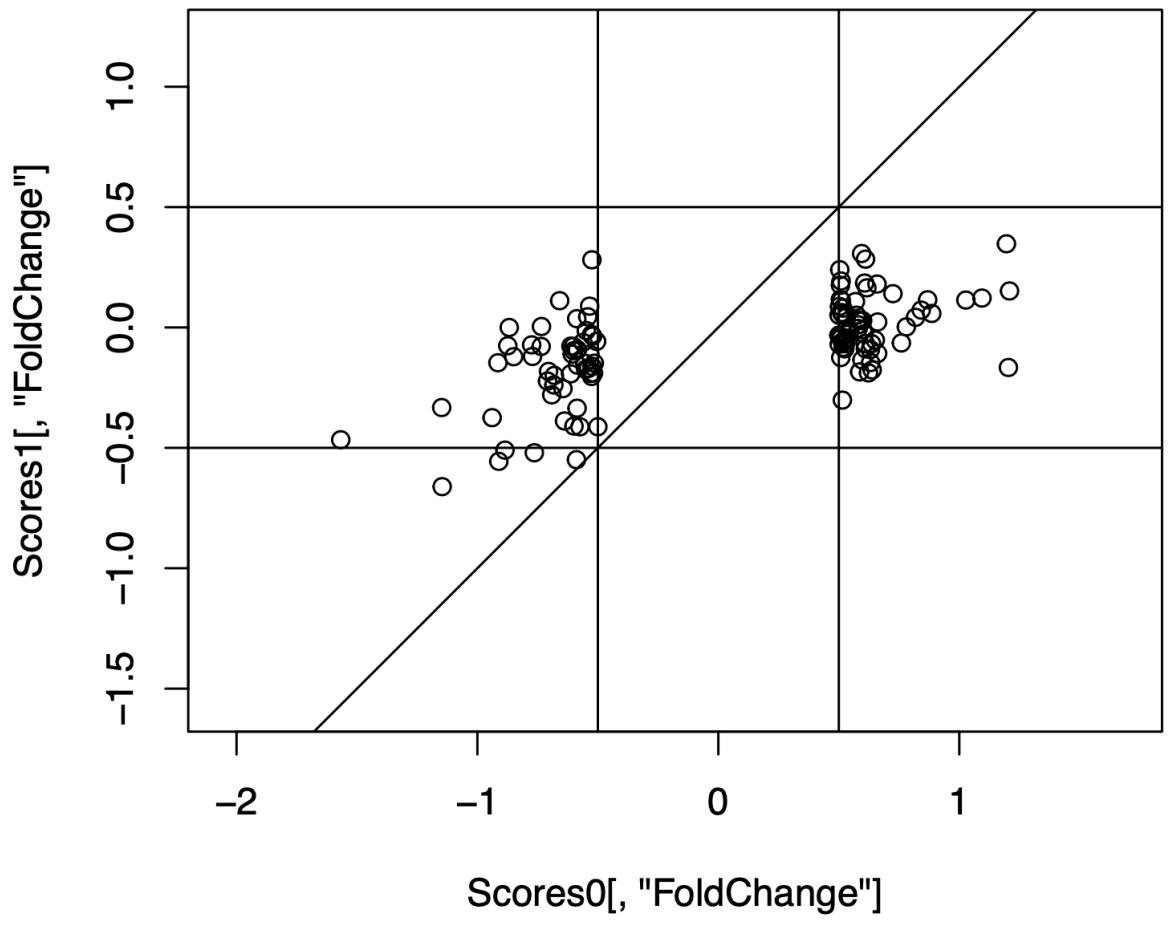
"half the effect size"



"half the effect size part 2: shrinkage"

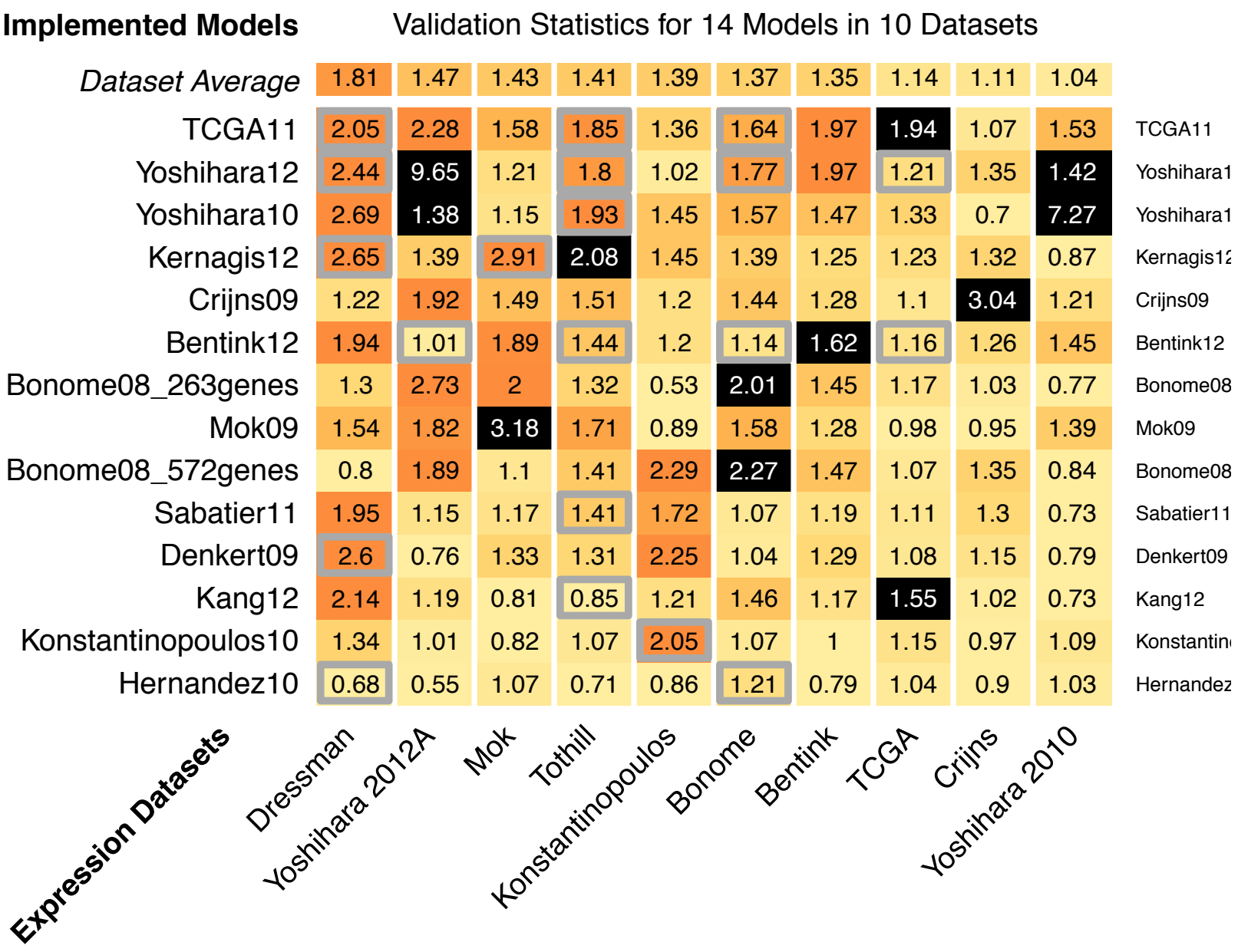
Scatterplot version of the same data, to emphasize "shrinkage".

```
Scores0 = ScoresSub[fcOnly,]
Scores1 = CompScores(XXX[fcOnly,],YYY)
plot(Scores0[,"FoldChange"],Scores1[,"FoldChange"],
     asp=1,ylim=range(Scores0[,"FoldChange"]))
abline(h=fcut); abline(h=-fcut); abline(v=fcut)
abline(v=-fcut); abline(0,1)
```

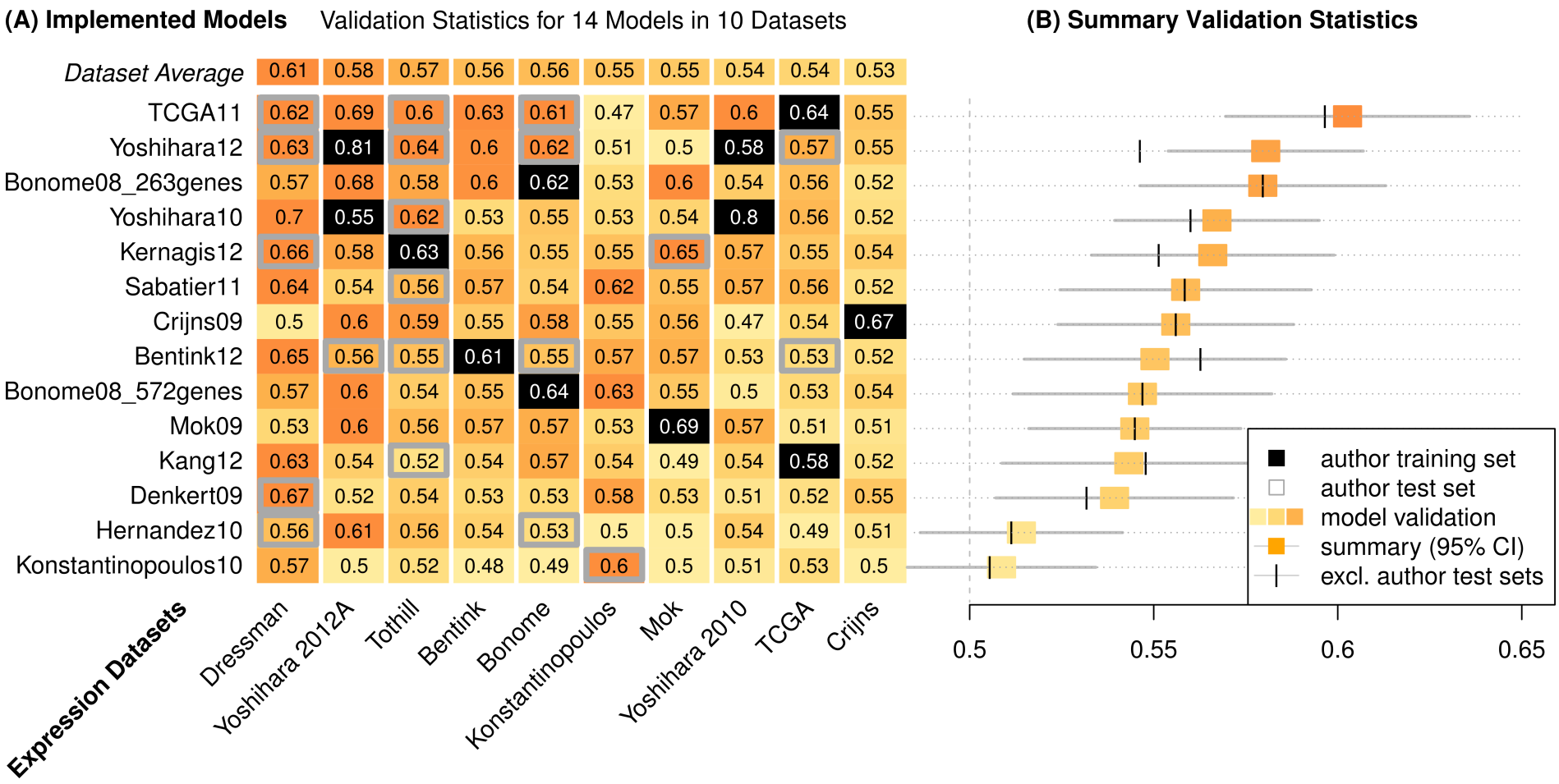


classifier cross-study validation matrix

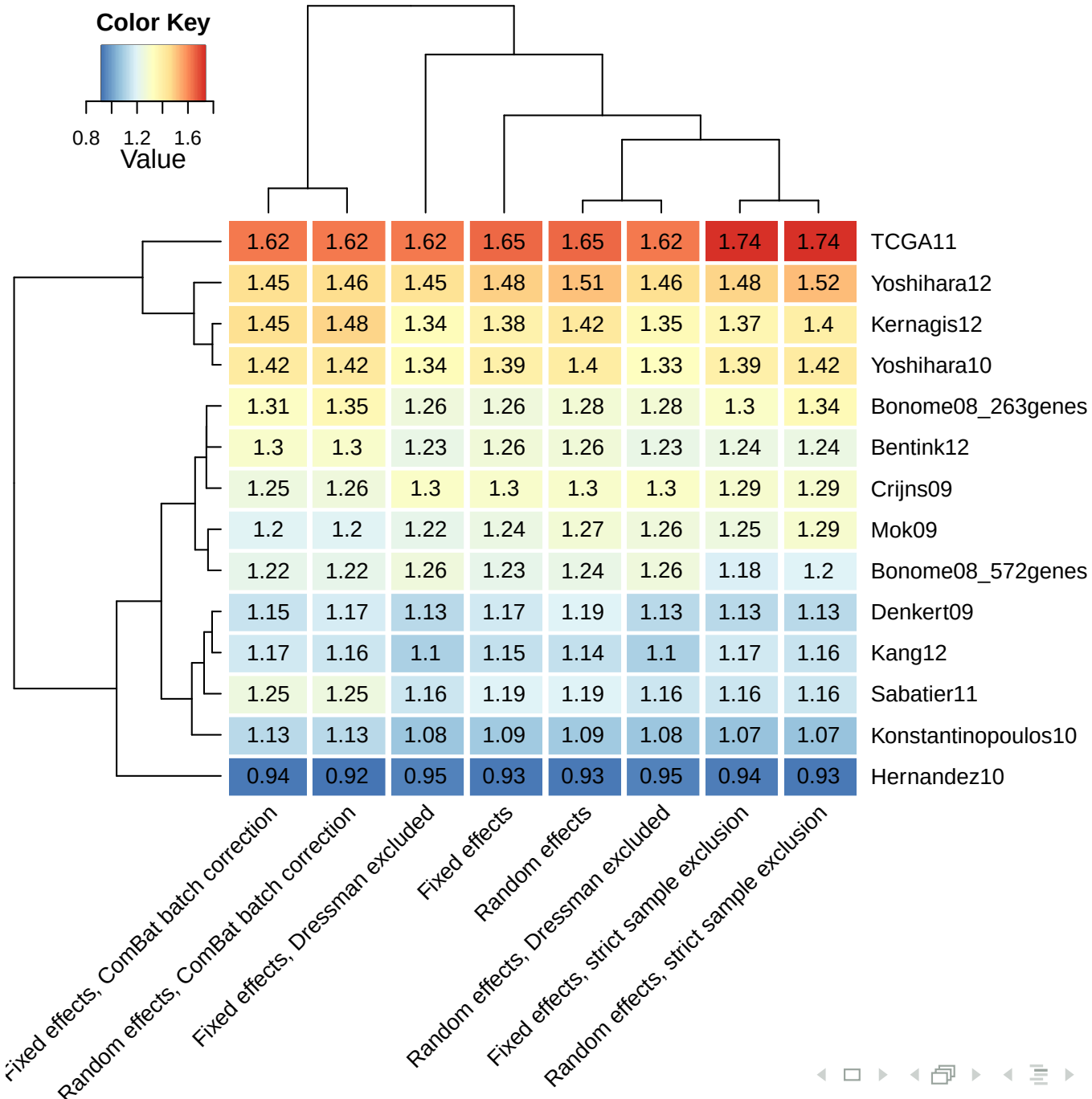
Waldron etal 2014



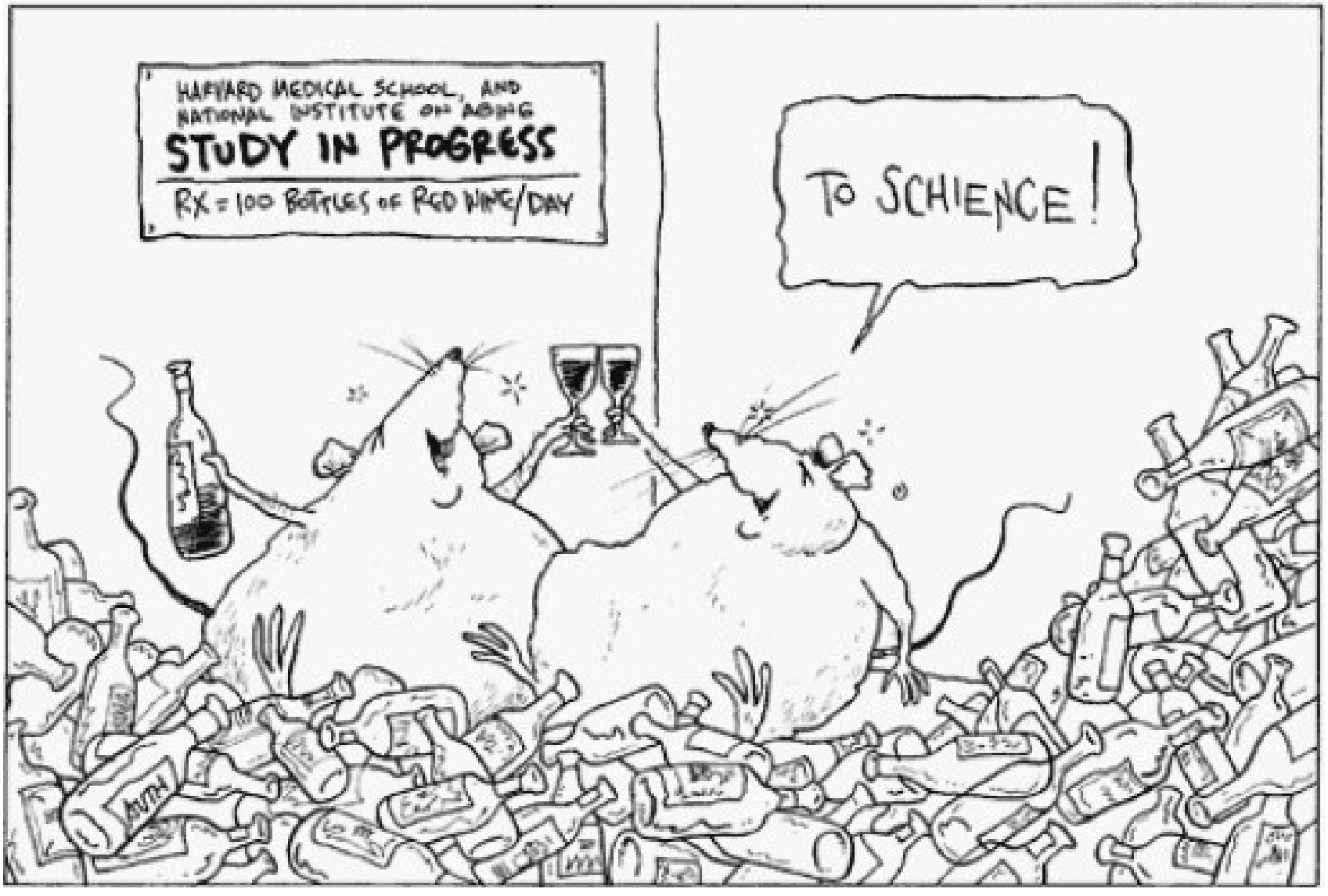
using c-stat instead



sensitivity analysis



DISCUSSION 1: WHO CELEBRATES?



no replicability. who wins?

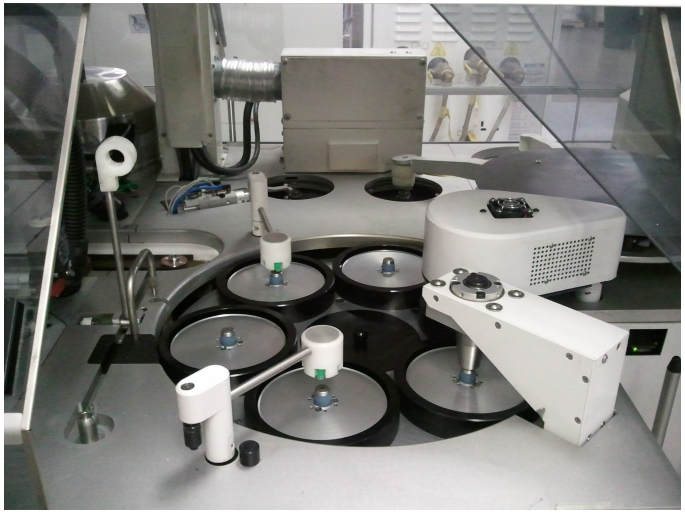
Imagine a world-wide committee of replicability experts came to the conclusion that a specific field of science is not replicable. Who would benefit?

replicability is built in. who wins?

Imagine the same world-wide committee of replicability experts came to the conclusion that a specific field of science is intrinsically self correcting and thus that replicability is no concern.
Who would benefit?

DISCUSSION 2: APPLES AND CDS

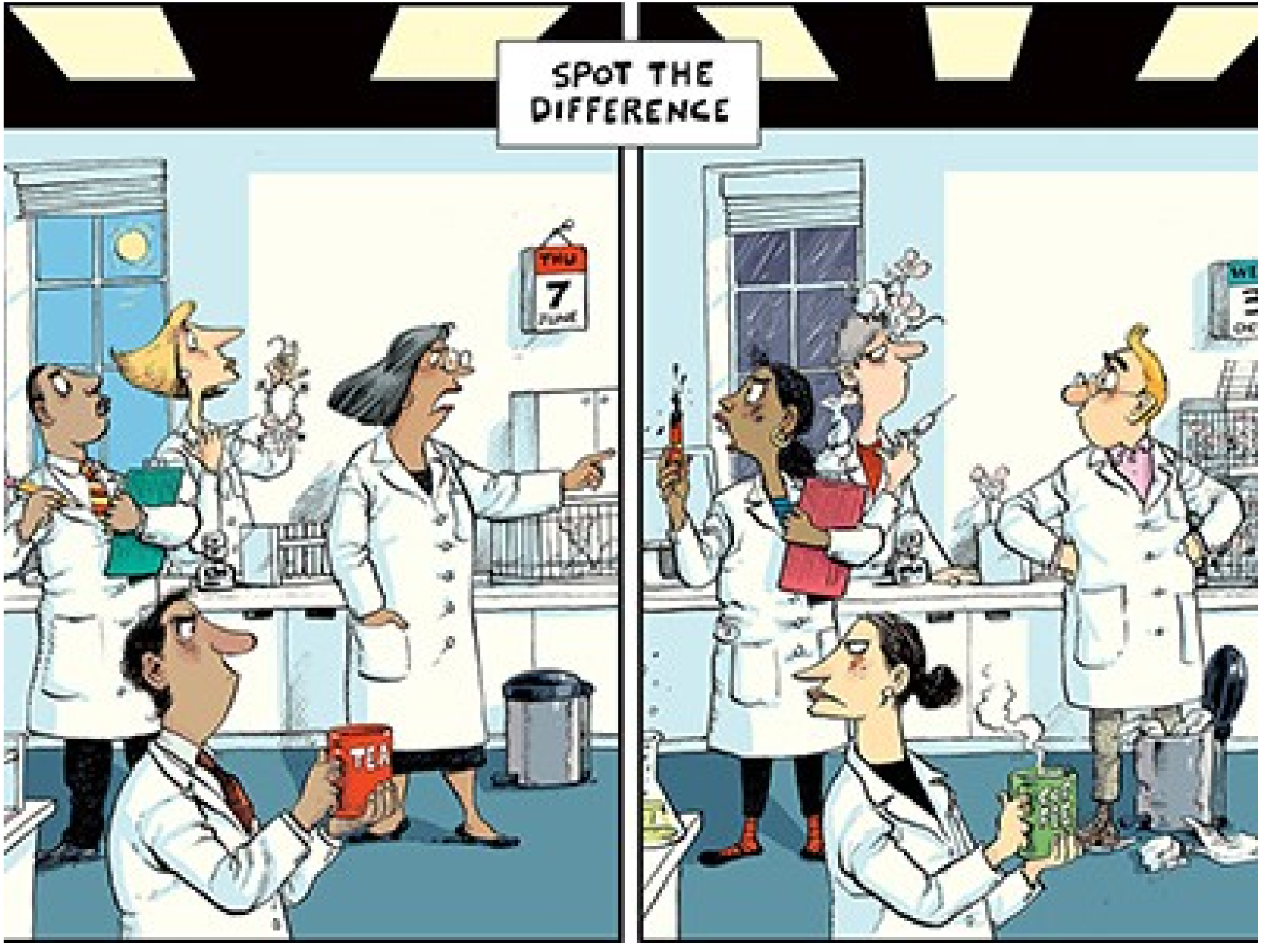
...



replicability or variability?

- No two experiments will ever be the same.
  Replicability is not a separate question from variability.
  Tweet: Replicability does not exist".
- "No knowledge can be safely used in practice in future instances if it not replicable.
  Tweet: Science does not exist without replicability."

EXTENT OF NON-REPLICABILITY



recommendation 5-1:

Researchers should, as applicable to the specific study, provide an accurate and appropriate characterization of relevant uncertainties when they report or publish their research. Researchers should thoughtfully

communicate all recognized uncertainties and

estimate or acknowledge other potential sources of uncertainty that bear on their results, including

stochastic uncertainties and

uncertainties in measurement, computation, knowledge, modeling, and methods of analysis.

hues of replicability

NAS: Importantly, the assessment of replicability may not result in a binary pass/fail answer; rather, the answer may best be expressed as the degree to which one result replicates another.

restrictive and unreliable

NAS: it is restrictive and unreliable to accept replication only when the results in both studies have attained "statistical significance," that is, when the p-values in both studies have exceeded a selected threshold.

... rather

NAS: Rather, in determining replication, it is important to consider the distributions of observations and to examine how similar these distributions are. This examination would include summary measures, such as proportions, means, standard deviations (uncertainties), and additional metrics tailored to the subject matter.

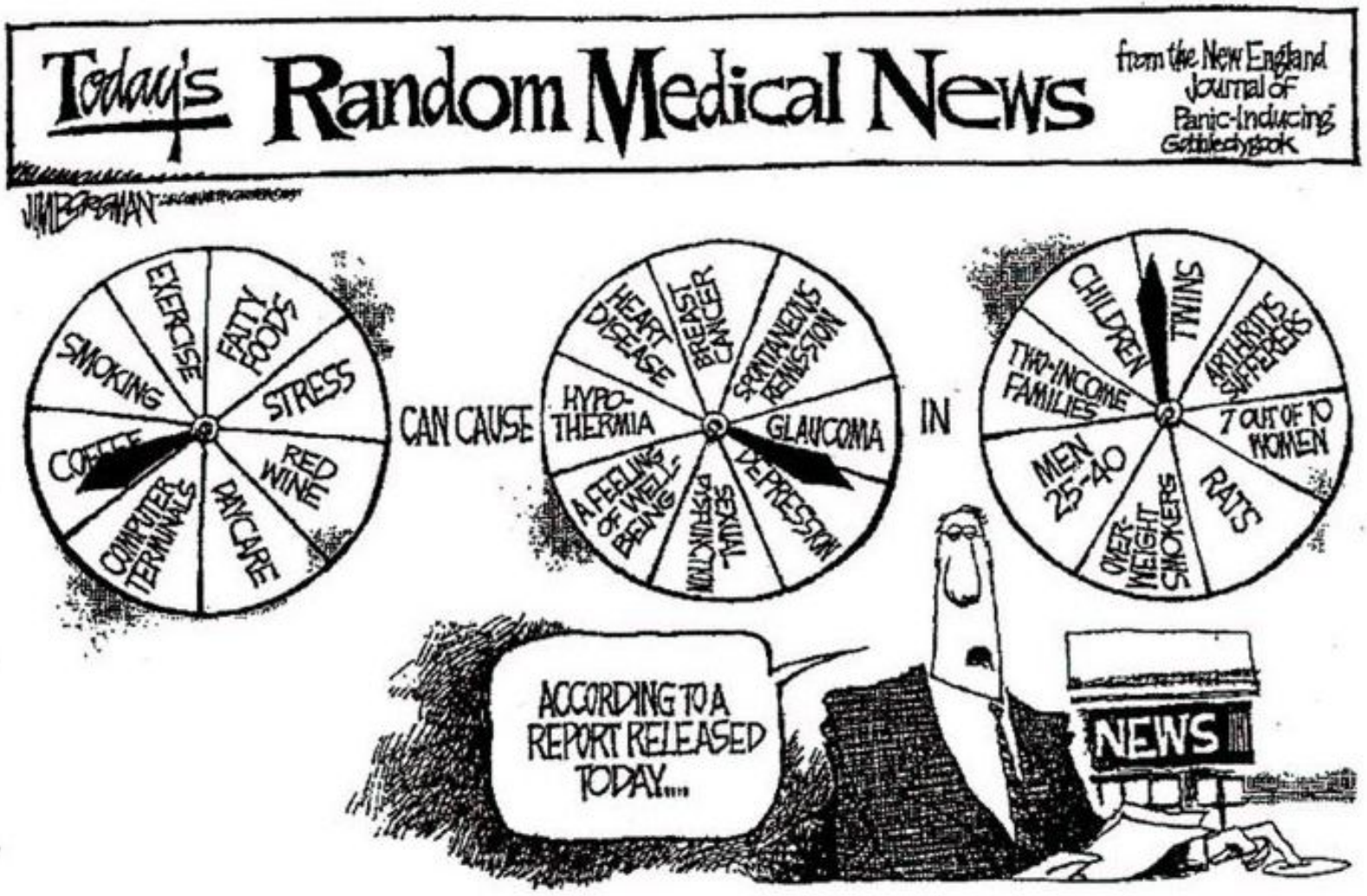
cumulative evidence

NAS: A predominant focus on the replicability of individual studies is an inefficient way to assure the reliability of scientific knowledge. Rather, reviews of cumulative evidence on a subject, to assess both the overall effect size and generalizability, is often a more useful way to gain confidence in the state of scientific knowledge.

many replication studies are not reported

NAS: Because many scientists routinely conduct replication tests as part of a follow-on experiment and do not report replication results separately, the evidence base of non-replicability across all science and engineering research is incomplete.

SOURCES OF NON-REPLICABILITY



helpful and not

NAS: Non-replicability occurs for a number of reasons that do not necessarily reflect that something is wrong. Some occurrences of non-replicability may be helpful to science —discovering previously unknown effects or sources of variability— while others, ranging from simple mistakes to methodological errors to bias and fraud, are not helpful.

IMPROVING REPLICABILITY

recommendation 6-7

Journals and scientific societies requesting submissions for conferences should disclose their policies relevant to achieving reproducibility and replicability. The strength of the claims made in a journal article or conference submission should reflect the reproducibility and replicability standards to which an article is held, with **stronger claims reserved for higher expected levels of reproducibility and replicability**.

recommendation 6-7, continued

1. set and implement desired standards of reproducibility and replicability and make this one of their priorities, such as deciding which level they wish to achieve for each Transparency and Openness Promotion guideline and working towards that goal;
2. adopt policies to reduce the likelihood of non-replicability, such as considering incentives or requirements for research materials transparency, design, and analysis plan transparency, enhanced review of statistical methods, study or analysis plan preregistration, and replication studies; and
3. require as a review criterion that all research reports include a thoughtful discussion of the uncertainty in measurements and conclusions. inferences.

recommendation 6-8

Many considerations enter into decisions about what types of scientific studies to fund, including striking a balance between exploratory and confirmatory research.

recommendation 6-8, continued

If private or public funders choose to invest in initiatives on reproducibility and replication, two areas may benefit from additional funding:

1. education and training initiatives to ensure that researchers have the knowledge, skills, and tools needed to conduct research in ways that adhere to the highest scientific standards; that describe methods clearly, specifically, and completely; and that express accurately and appropriately the uncertainty involved in the research; and
2. reviews of published work, such as testing the reproducibility of published research, conducting rigorous replication studies, and publishing sound critical commentaries.

recommendation 6-9

Funders should require a thoughtful **discussion in grant applications of how uncertainties will be evaluated**, along with any relevant issues regarding replicability and computational reproducibility.

Funders should introduce **review of reproducibility and replicability guidelines and activities** into their merit-review criteria, as a low-cost way to enhance both.

recommendation 6-10

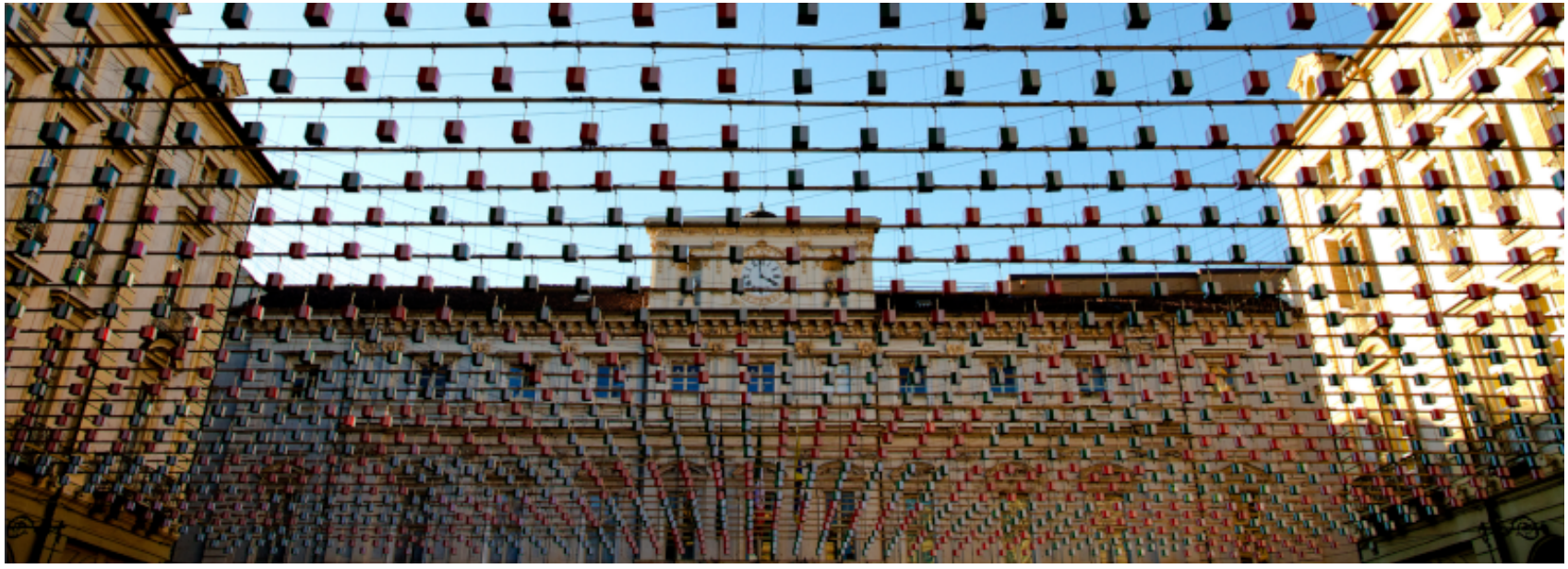
When funders, researchers, and other stakeholders are considering whether and where to direct resources for replication studies, they should consider the following criteria:

1. The scientific results are important for individual decision-making or for policy decisions.
2. The results have the potential to make a large contribution to basic scientific knowledge.
3. The original result is particularly surprising, that is, it is unexpected in light of previous evidence and knowledge.
4. There is controversy about the topic.

recommendation 6-10, continued

5. There was potential bias in the original investigation, due, for example, to the source of funding.
6. There was a weakness or flaw in the design, methods, or analysis of the original study.
7. The cost of a replication is offset by the potential value in reaffirming the original results.
8. Future expensive and important studies will build on the original scientific results.

too much replication?



DIGRESSIONS

the tale of the worms and the red wine

Phillips, P., Lithgow, G.J., and Driscoll, M. (2017).

A Long Journey to Reproducible Results. Nature, 548(7668), 387-388.

the tale of the worms and the red wine

Some research had found that resveratrol (found in red wine) could dramatically extend the life of worms in the lab, but other scientist had difficulty replicating the results.

Reasons for this lack of replicability:

Differences in lab **protocol:**

e.g. worms handled by gentle lab technicians lived a full day longer than others.

Difference in **measurement:**

e.g. one lab determined age on the basis of when an egg was laid; another began counting when it was hatched.

Once these were eliminated, researchers discovered inherent variability in the **model system:**

e.g. some cohorts of worms could partition into short-lived or long-lived modes of aging (previously unknown.)

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

Summary

Background New technologies for the detection of early-stage ovarian cancer are urgently needed. Pathological changes within an organ might be reflected in proteomic patterns in serum. We developed a bioinformatics tool and used it to identify proteomic patterns in serum that distinguish neoplastic from non-neoplastic disease within the ovary.

Methods Proteomic spectra were generated by mass spectroscopy (surface-enhanced laser desorption and ionisation). A preliminary "training" set of spectra derived from analysis of serum from 50 unaffected women and 50 patients with ovarian cancer were analysed by an iterative searching algorithm that identified a proteomic pattern that completely discriminated cancer from non-cancer. The discovered pattern was then used to classify an independent set of 116 masked serum samples: 50 from women with ovarian cancer, and 66 from unaffected women or those with non-malignant disorders.

Findings The algorithm identified a cluster pattern that, in the training set, completely segregated cancer from non-cancer. The discriminatory pattern correctly identified all 50 ovarian cancer cases in the masked set, including all 18 stage I cases. Of the 66 cases of non-malignant disease, 63 were recognised as not cancer. This result yielded a sensitivity of 100% (95% CI 93–100), specificity of 95% (87–99), and positive predictive value of 94% (84–99).

Introduction

Application of new technologies for detection of ovarian cancer could have an important effect on public health,[1] but to achieve this goal, specific and sensitive molecular markers are essential.[1–5] This need is especially urgent in women who have a high risk of ovarian cancer due to family or personal history of cancer, and for women with a genetic predisposition to cancer due to abnormalities in predisposition genes such as *BRCA1* and *BRCA2*. There are no effective screening options for this population.

Ovarian cancer presents at a late clinical stage in more than 80% of patients,[1] and is associated with a 5-year survival of 35% in this population. By contrast, the 5-year survival for patients with stage I ovarian cancer exceeds 90%, and most patients are cured of their disease by surgery alone.[1–6] Therefore, increasing the number of women diagnosed with stage I disease should have a direct effect on the mortality and economics of this cancer without the need to change surgical or chemotherapeutic approaches.

Cancer antigen 125 (CA125) is the most widely used biomarker for ovarian cancer.[1–6] Although concentrations of CA125 are abnormal in about 80% of patients with advanced-stage disease, they are increased in only 50–60% of patients with stage I ovarian cancer.[1–6] CA125 has a positive predictive value of less than 10% as a single marker, but the addition of ultrasound screening to CA125 measurement has improved the

Ovacheck

- The researchers, from NIH & FDA, won widespread acclaim.
- Congressional resolution urged further funding for their research.
- The magazine "Health" named the test one of the top ten medical advances of the year.
- Commercial rights to develop the test were licensed from the US government to Correlogic Systems.
- Correlogic granted licenses to Quest Diagnostics and the Laboratory Corporation of America, hoping to market the test under the brand name OvaCheck.

Slide based on a presentation by S. Goodman at the National Cancer Policy Forum, 2018

BIOINFORMATICS

Vol. 20 no. 5 2004, pages 777–785
DOI: 10.1093/bioinformatics/btg484

Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments

Keith A. Baggerly, Jeffrey S. Morris and Kevin R. Coombes*

Department of Biostatistics, U.T. M.D. Anderson Cancer Center, 1515 Holcombe Blvd, Box 447, Houston, TX 77030-4009, USA

Received on July 14, 2003; revised on October 14, 2003; accepted on October 16, 2003
Advance Access publication January 29, 2004

Compared SELDI proteomic spectra from serum from three experiments by the same group on separating ovarian cancer from normal tissue. These spectra are available on the web at http://clinicalproteomics.steem.com.

Baggerly et al results

Results: In general, the results were not reproducible across experiments. Baseline correction prevents reproduction of the results for two of the experiments. In one experiment, there is evidence of a **major shift in protocol mid-experiment** which could bias the results. In another, **structure in the noise regions of the spectra allows us to distinguish normal from cancer**, suggesting that the normals and cancers were processed differently. Sets of **features found to discriminate well in one experiment do not generalize** to other experiments. Finally, the mass calibration in all three experiments appears suspect. Taken together, these and other concerns suggest that much of the structure uncovered in these experiments could be due to artifacts of sample processing, not to the underlying biology of cancer. We provide some guidelines for design and analysis in experiments like these to ensure better reproducible, biologically meaningfully results.