

Lies, damn lies, and genomics

Navigating your data, your perceptions and reality

Christopher West Wheat
Professor at Department of Zoology



1

Career trajectory

A map showing the career trajectory of Christopher West Wheat. A sea turtle is positioned over the Atlantic Ocean. Red arrows indicate a path from California to Germany, then to Finland, and finally to Stockholm. A green arrow indicates a return path from Stockholm to California.

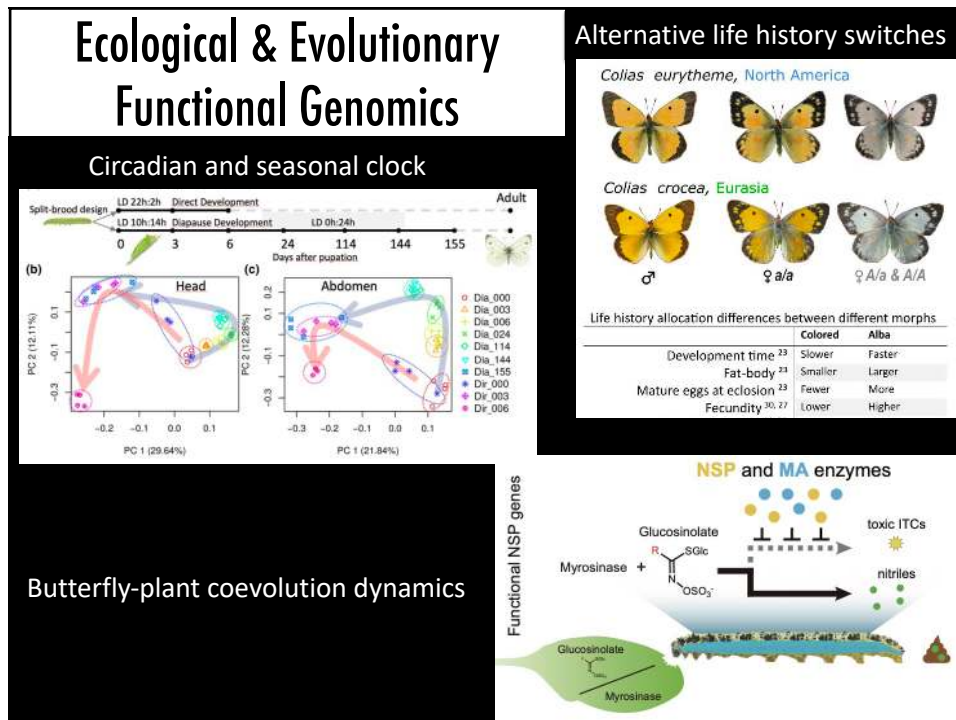
- 1995 – 2001 PhD California
- 2002 – 2005 Postdoc Germany
- 2005 – 2008 Postdoc Finland
- 2009 – unemployed 4 month, spent all savings
 - > 50 job applications, 1 grant application
- 2009 – visiting scientist Germany
 - 1 job offer UK
 - 1 grant Finland
- 2012 – Assistant Prof. at Stockholm University
- 2022 – Full Professor

What was important?

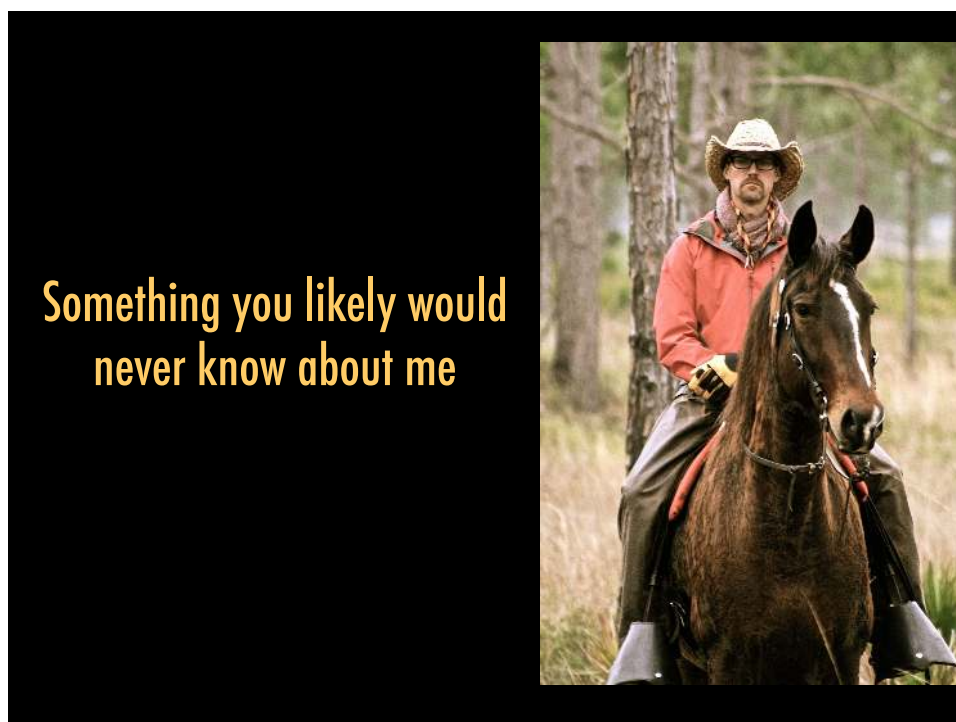
- Being able to move, chase the money & get new skills
- Learning how to believe in my ideas/skills

I was able to put science first, and had lots of fun along the way

2



3



4

I am a Judge of Field Trials,
for the American Field Trial Clubs of America, since 2003



5

Goals of this lecture

- Present a critical view of things genomic
- Make you uncomfortable by sharing some of my nightmares with you
- Critically assess findings and expectations in light of easy errors and publication biases
- Encourage you to be part of the solution

6

Disclaimer

I'm a positive person

I love my job and the work we all do

I'm just sharing scrumptious food for thought

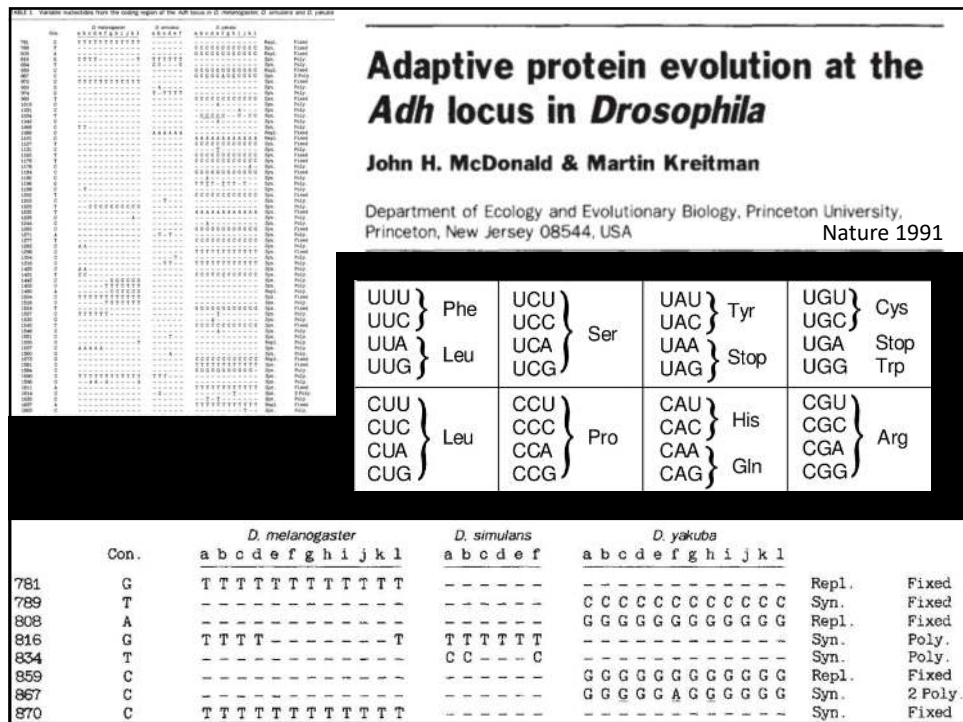
7

What if

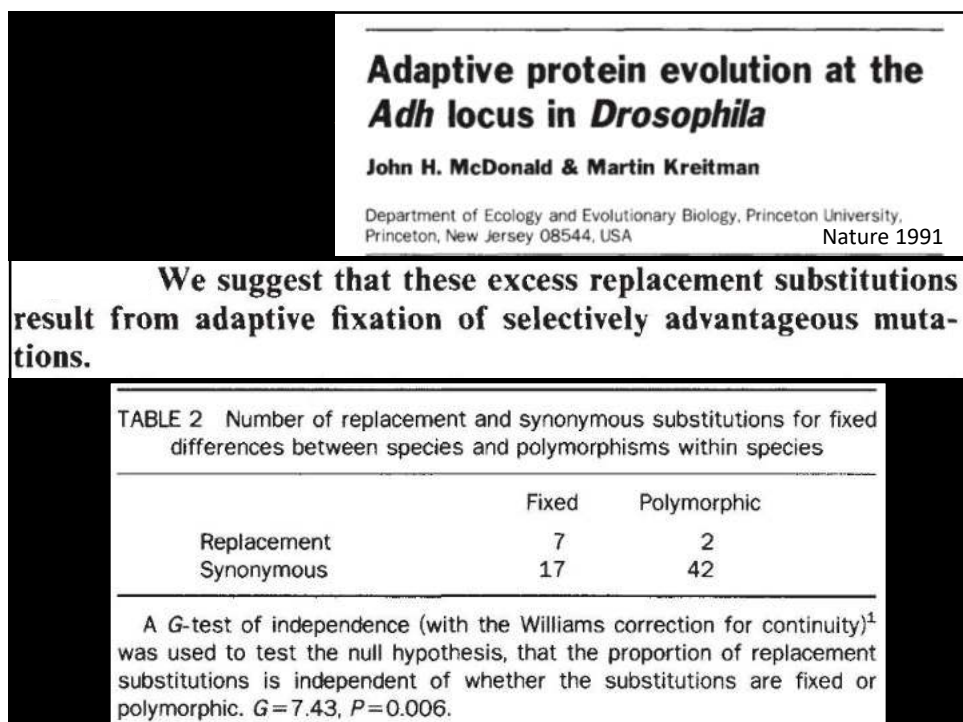
Would that
impact your
science?

50% of your
favorite studies
were not
repeatable?

8



9



10



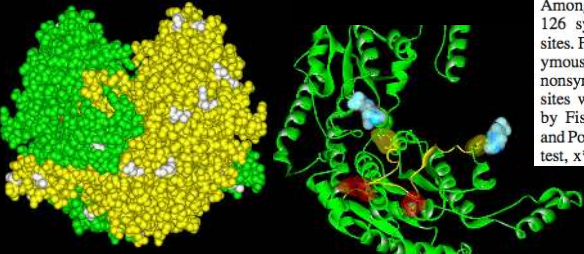
Colias eurytheme

I wanted to use this new molecular test of selection on a classic example of balancing selection from allozyme era

From DNA to Fitness Differences: Sequences and Structures of Adaptive Variants of *Colias* Phosphoglucose Isomerase (PGI)

Christopher W. Wheat,*^{†1} Ward B. Watt,*[†] David D. Pollock,*^{†2} and Patricia M. Schulte*^{†3}

*Department of Biological Sciences, Stanford University and [†]Rocky Mountain Biological Laboratory, Crested Butte, Colorado

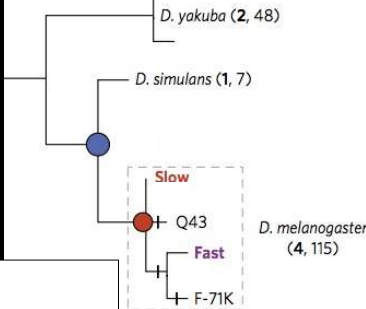
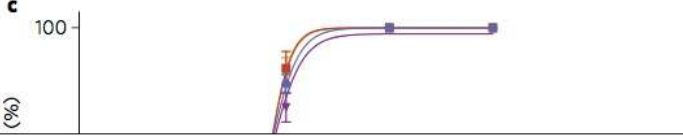


Among *C. eurytheme* and *C. meadii* PGI sequences, we find 126 synonymous and 20 nonsynonymous polymorphic sites. From their ratio, 6.3:1, neutrality predicts ~13 synonymous fixations alongside the two observed interspecies nonsynonymous fixations. But, *no* fixed synonymous sites were found (above). These data differ significantly by Fisher's exact test, $P = 0.021$, following Moriyama and Powell (1987) and by Goldman's (1964) exact binomial test, $x^* = 3.41$, $P = 0.0006$.

Wheat et al. 2005

11

But ... the implications of the MK test results in *Drosophila melanogaster* were never rigorously investigated till 30 years later ...

nature ecology & evolution

ARTICLES

PUBLISHED: 13 JANUARY 2017 | VOLUME: 1 | ARTICLE NUMBER: 0025

Experimental test and refutation of a classic case of molecular adaptation in *Drosophila melanogaster*

12

So.....

Does this happen only in bugs?

my PhD chased an adaptive story lacking a rigorous foundation

13

If the biomedical science has the most money and oversight, then

Their findings should be robust:

- Repeatable effect sizes
- The same across different labs
- The same across years

14

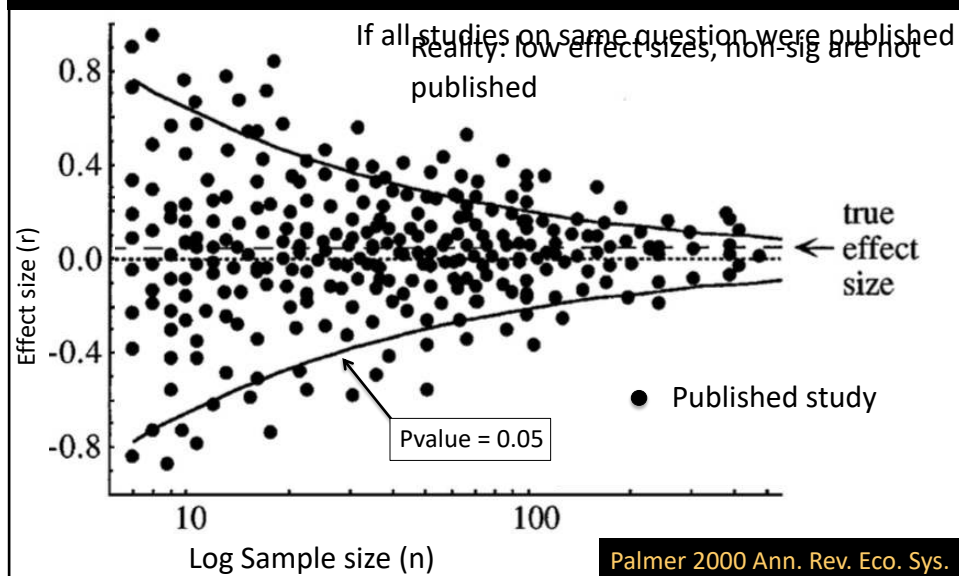
Publication replication failures

- Biomedical studies
 - Of 49 most cited clinical studies, 45 showed intervention was effective
 - Most were randomized control studies (robust design)
- Mouse cocaine effect study, replicated in three cities
 - Highly standardized study

Ioannidis 2005 JAMA; Lehrer 2010

15

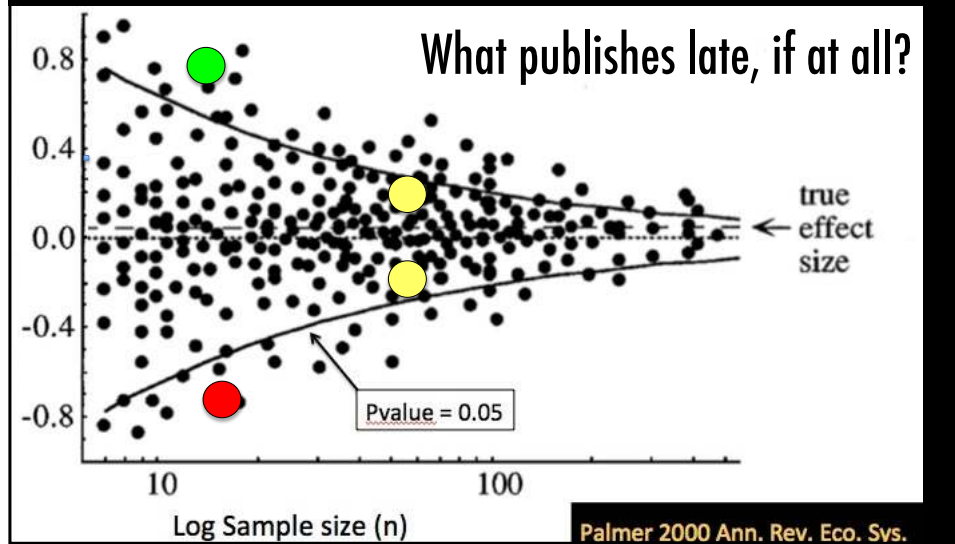
Publication bias increases effect size



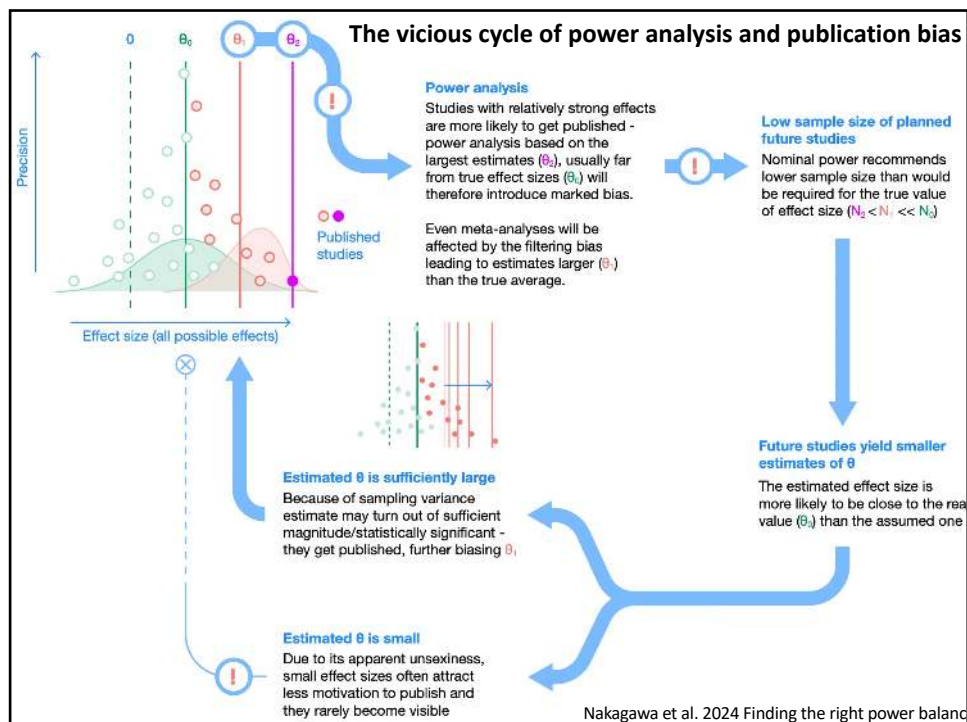
16

What if there is no replication?

What is most likely to publish first & where?



17



18

Why Most Published Research Findings Are False

A research finding is less likely to be true when:

- ✓ the studies conducted in a field have a small sample size
- ✓ when effect sizes are small
- ✓ when there are many tested relationships using tests without α priori selection
- ✓ where there is greater flexibility in designs, definitions, outcomes, and analytical modes
- ✓ when there is greater financial and other interest and prejudice
- ✓ when more teams are involved in a scientific field, all chasing after statistical significance by using different tests

Ioannidis 2005 Plos Med.

19

But surely, this doesn't
apply to genomics

Or does it?

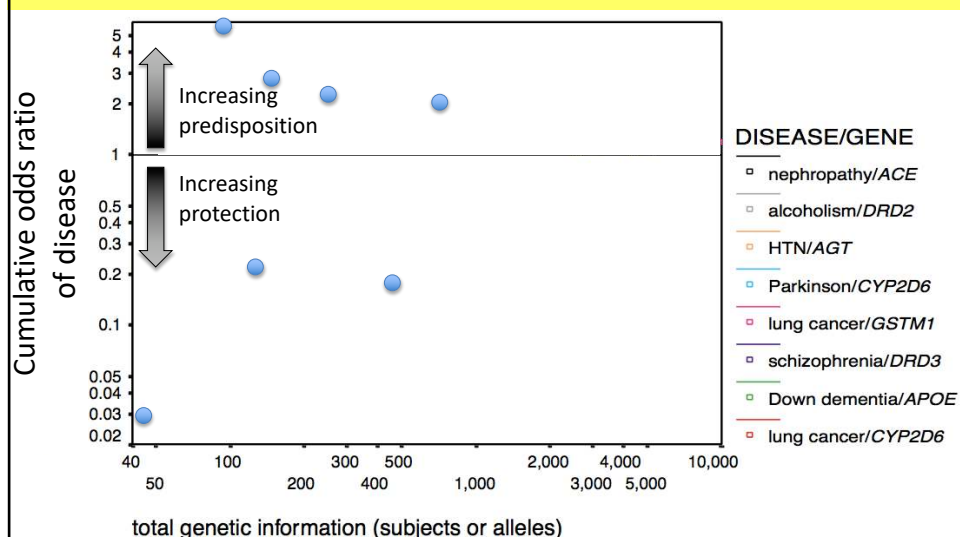
20

Outline

- Are these biases inherent in genomic studies?
- Why is this happening?
- How can we try and overcome these problems?

21

8 topics first reported with $P < 0.05$



Ioannidis, J. P., E. E. Ntzani, T. A. Trikalinos, and D. G. Contopoulos-Ioannidis. 2001. Replication validity of genetic association studies. *Nat Genet* 29:306–309.

22

There are lies, damn lies,
and

But wait, is that fair?

Are these really lies?

23

Where does this bias come from?

- Population heterogeneity
 - Space and time
- Publication culture
 - Large & significant effects publish fast with high impact
 - Small & non-significant effects publish slow, rarely, and with low impact

24

Where does this bias come from?



YOU!!

And me All of us

Its arises from humans doing science

The way we think
The way our institutions work

25

Apophenia

The tendency to seek and see patterns in random information and view this as important




Story telling of the false positives

26

Genomics is too big to fail

- Making errors is extremely common
- Errors almost always result in highly significant results
- Studies in non-model species are rarely replicated

Thus, always question your bioinformatics before falling in love with your results

When results are better than you could have dreamed,

27

Publications with significant human error that have not been retracted

PNAS

Comparison of the transcriptional landscapes between human and mouse tissues

“the expression for many sets of genes was found to be more similar in different tissues within the same species than between species”

ARTICLE

174 | NATURE | VOL 473 | 12 MAY 2011

doi:10.1038/nature09944

Enterotypes of the human gut microbiome

we identify three robust clusters (referred to as enterotypes hereafter) that are not nation or continent specific ... mostly driven by species composition

LETTER

228 | NATURE | VOL 502 | 10 OCTOBER 2013

doi:10.1038/nature12511

Genome-wide signatures of convergent evolution in echolocating mammals

PNAS

More genes underwent positive selection in chimpanzee evolution than in human evolution

28

PNAS

Comparison of the transcriptional landscapes between human and mouse tissues

"the expression for many sets of genes was found to be more similar in different tissues within the same species than between species"

Time of the most recent common ancestor:

Human and Mouse

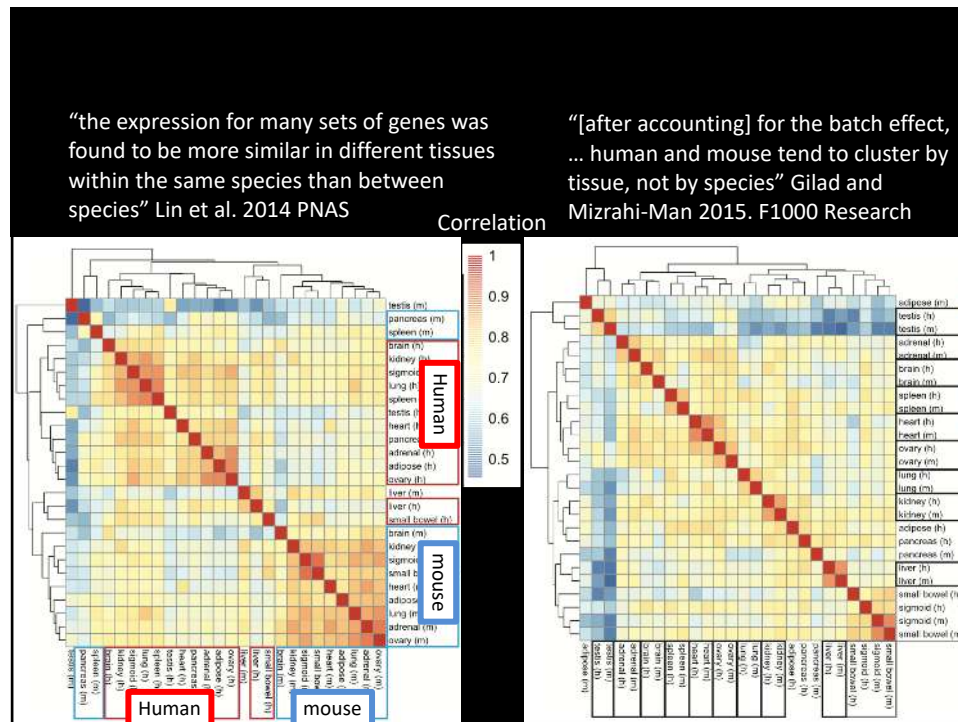
29

Authors found strong grouping of all organs by species, not by organ

Should gene expression patterns group by species or tissues?

What do we expect from first principals, evolutionary relationships?

30



31

Why? this was a batch effect, which confounded sequencing grouping with biological grouping

D87PMJN1 (run 253, flow cell D2GUAACXX, lane 7)	D87PMJN1 (run 253, flow cell D2GUAACXX , lane 8)	D4LHBFN1 (run 276, flow cell C2HKJACXX , lane 4)	MONK (run 312, flow cell C2GR3ACXX , lane 6)	HWI-ST373 (run 375, flow cell C3172ACXX , lane 7)
heart	adipose	adipose	heart	brain
kidney	adrenal	adrenal	kidney	pancreas
liver	sigmoid colon	sigmoid colon	liver	brain
small bowel	lung	lung	small bowel	spleen
spleen	ovary	ovary	testis	
testis		pancreas		

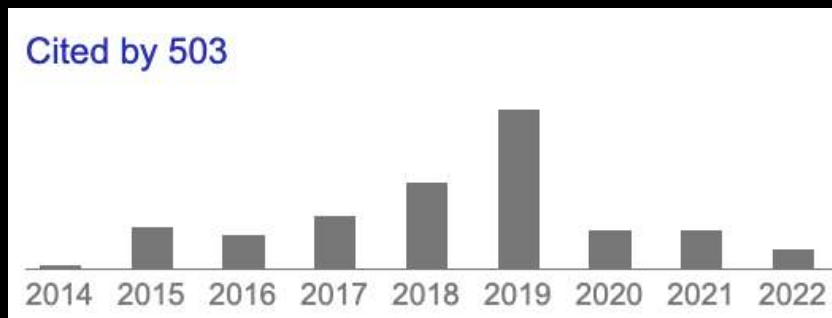
● Human
● Mouse

Solution = Keep technical effects orthogonal to biological

- Process samples together, both species in same lane, same tissues in same lane
- Will your Core facility know to do this for you?

32

.... why is this still being cited?

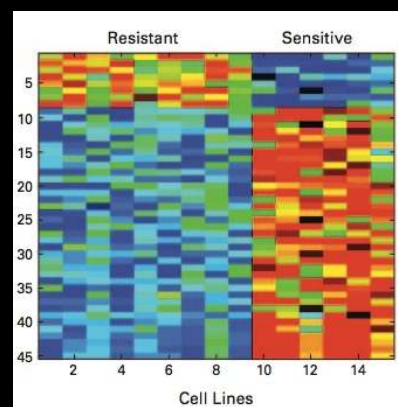


33

Do you want significant results? use Excel

- Personal medicine study, searching for gene expression signatures predicting sensitivity to specific cancer drugs, as patients show highly variable response to drug called cisplatin
 - treatment for advanced non-small-cell lung cancer
- Found strong signature in transcriptome between resistant vs. responsive cells to cisplatin
- Leading to additional funding
 - Prescreen patients, get better outcome
 - Planned clinical trials with drugs

Hsu et al. 2007



34

FORENSIC BIOINFORMATICS AND REPRODUCIBLE RESEARCH IN HIGH-THROUGHPUT BIOLOGY

"Data processing, however, is often not described well enough to allow for exact reproduction of the results,

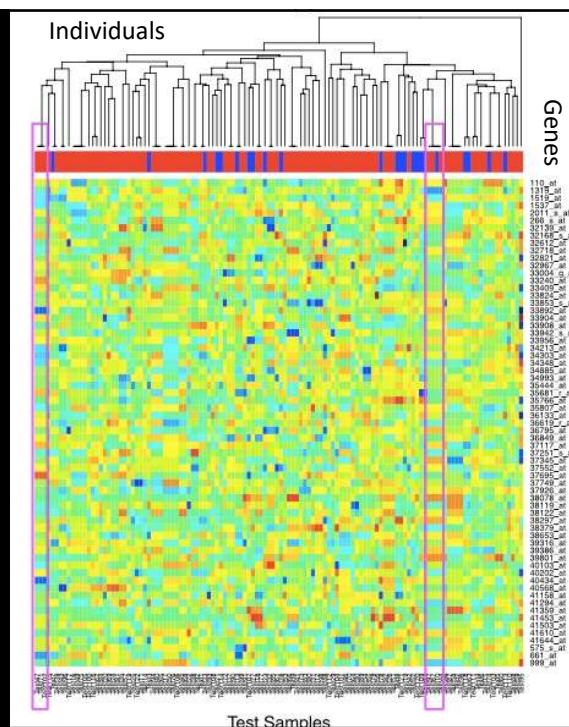
Thanks: Malachi Griffith

Baggerly and Coombes 2009

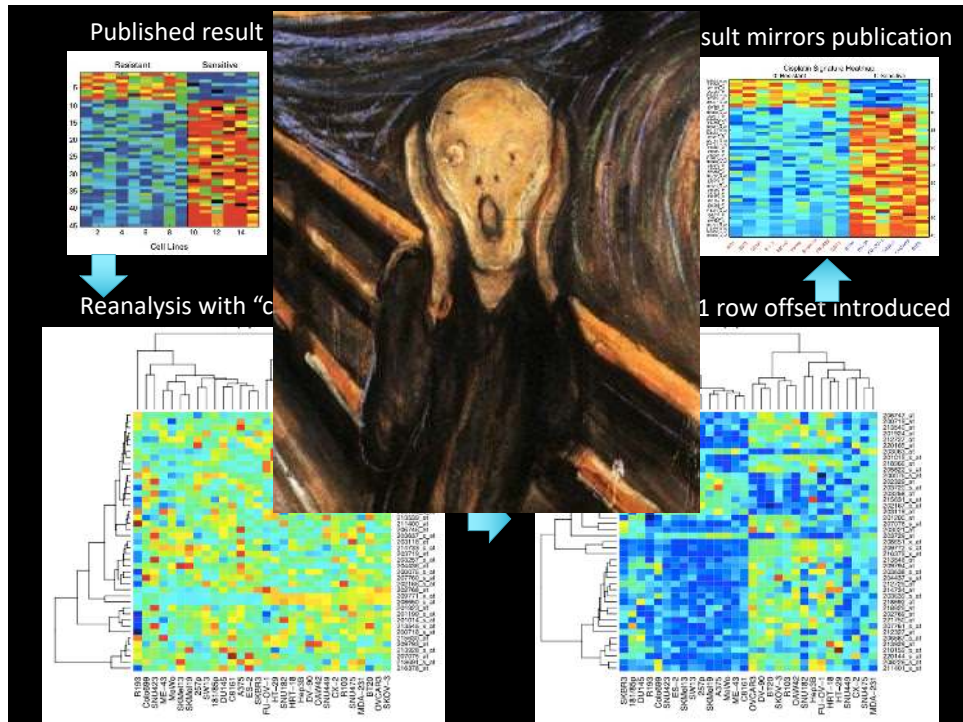
35

Digging revealed:

- Instances of repeated sampled data
- Only 84/122 test samples were distinct
- Some repeated samples labeled both sensitive and resistant
- Row offset in data table



36



37

VOLUME 25 • NUMBER 28 • OCTOBER 1 2007

JOURNAL OF CLINICAL ONCOLOGY ORIGINAL REPORT

This article was retracted on November 16, 2010

Pharmacogenomic Strategies Provide a Rational Approach to the Treatment of Cisplatin-Resistant Patients With Advanced Cancer

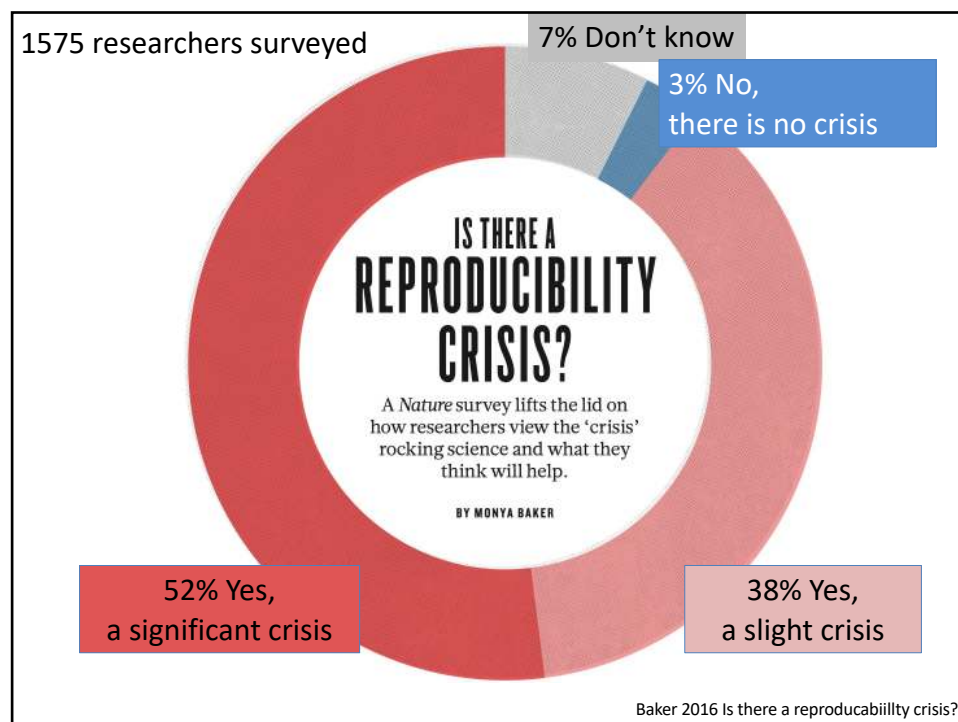
38

Can we reduce these type of publications?

YES!!!!

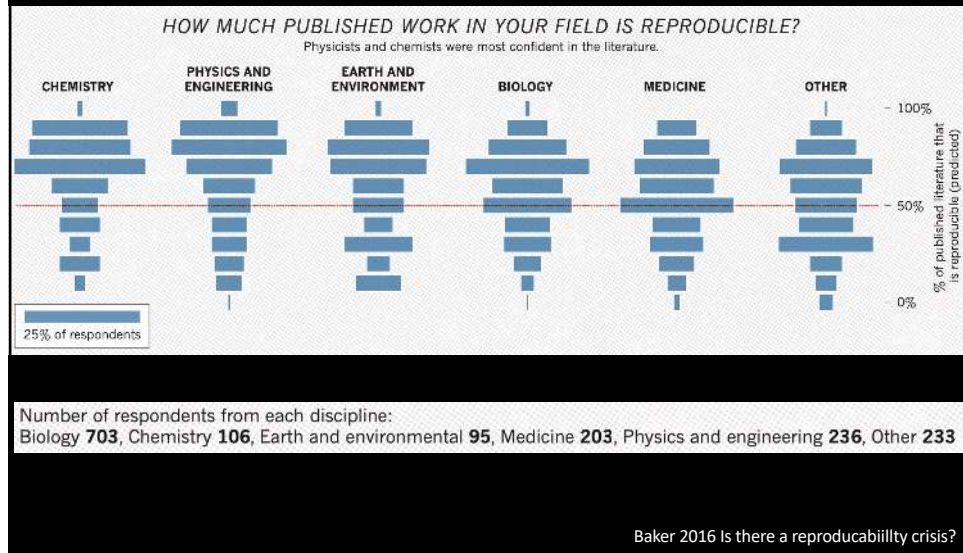
- Work better as a community, check each others code
- As author, as supervisor, as reviewer, as Associate Editor, make sure all studies you touch :
 - Have all code and raw data open source
 - Analyzed datasets open source
 - Methods clearly described

39

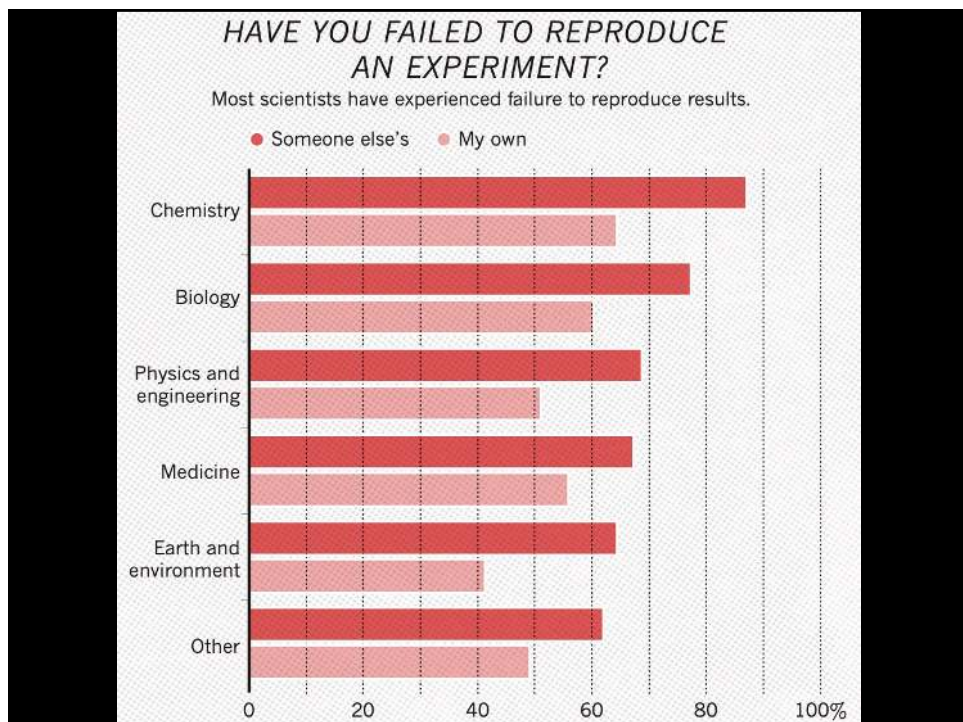


40

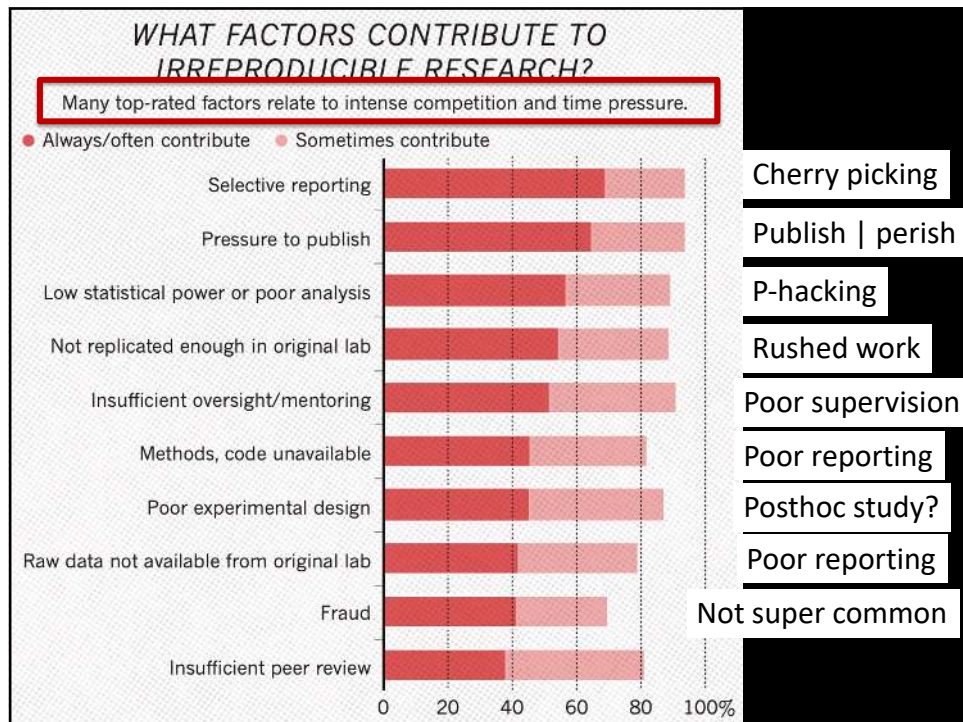
Breakdown by research field



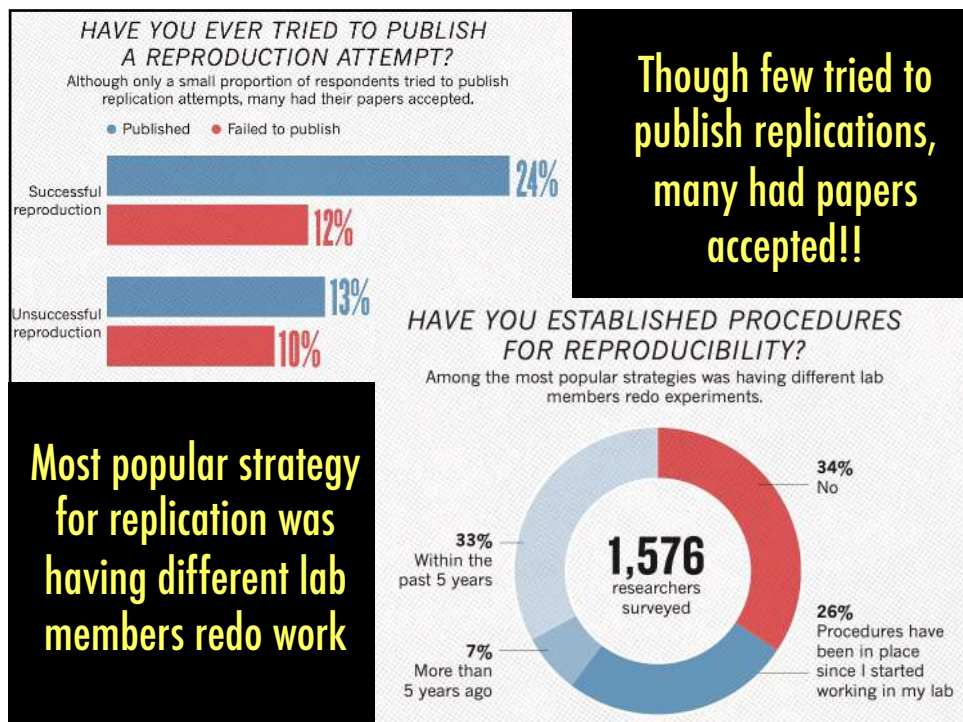
41



42



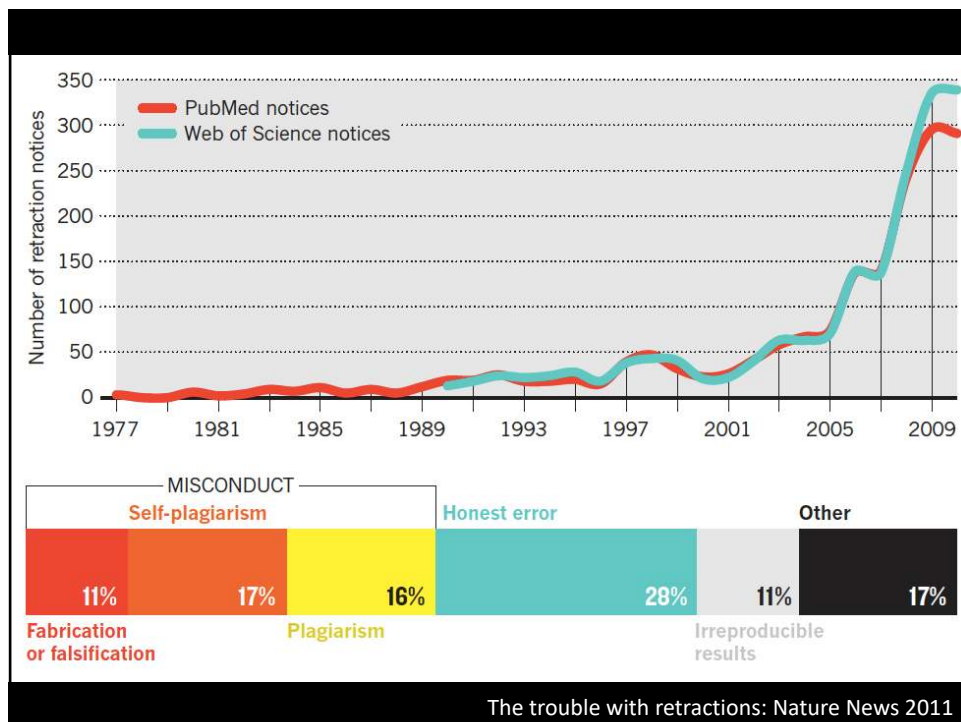
43



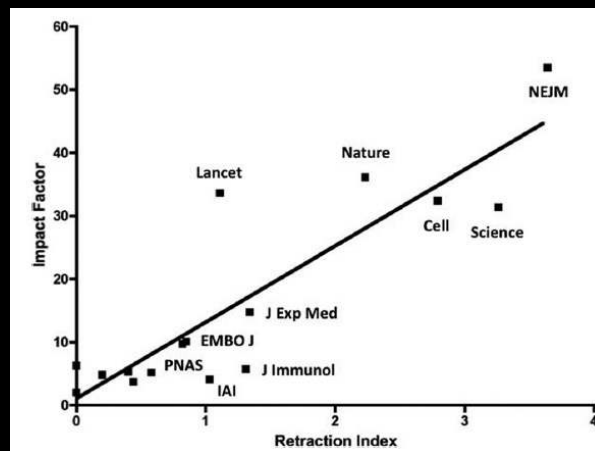
44



45



46



"the frequency of retraction varies among journals and shows a strong correlation with the journal impact factor"

Fang 2011 Infect. Immun.

47

- Website shows retraction

PubMed.gov
US National Library of Medicine
National Institutes of Health

PubMed Advanced

Format: Abstract + Send to +

RETRACTED ARTICLE
See: [Retraction Notice](#)

J Clin Oncol. 2007 Oct 1;25(28):4350-7.

Pharmacogenomic strategies provide a rational approach to the treatment of cisplatin-resistant patients with advanced cancer.

Hsu DS¹, Balakumaran BS, Acharya CB, Vlahovic V, Walters KS, Garman K, Anders C, Riedel RF, Lancaster J, Harpole D, Dressman HK, Nevins JR, Febbo PG, Portis A.

48



- Keep community updated
- Help kill zombie papers that keep getting cited when they should not
- Starting to get integrated into different websites for automatic scans
- Be sure you are never keeping zombies alive



49



Frances Arnold
@francesarnold



For my first work-related tweet of 2020, I am totally bummed to announce that we have retracted last year's paper on enzymatic synthesis of a complex molecule. The results are not reproducible. [science.sciencemag.org](#)



DEEP-SEA RESEARCH
Rod opsins and color vision in deep-sea fishes pp. 520-528

Site-selective enzymatic synthesis
Enzymes excel at site-selective synthesis. With appropriate engineering, they can be used to synthesize complex molecules.

[science.sciencemag.org](#)



Prof. Lee Cronin @leecronin · Jan 2

Replying to @francesarnold

First class. Sometimes things appear to work, then they don't. Science should be a process, not winner takes all whatever the cost. Entrepreneurs are encouraged to fail well, but in science it's still taboo. I hope when I slip up I'm able to do it so openly & well.

4
13
262

1 more reply



Lynn Kamerlin @kamerlinlab · Jan 2

Replying to @francesarnold

Sorry about the problems, but kudos for doing the right thing, and setting a good example.

1
1
178



Waheed Ahmed @WaheedURAhmed1 · Jan 3

Honesty is so important and unfortunately, pretty underrated. Lots of respect and admiration for your actions.

50

25

So ... there are lots of high-profile errors out there ...

Much of this is scientific progress ... we are not perfect, just doing what we can

Thus you must calibrate your expectations, approaches, and stay humble

51

What is your personal error rate?

I assume mine is 12%

therefore I perform many sanity & error checks to catch errors that I KNOW I WILL MAKE

52

What other biases might we suffer from?



53

We're basically a rather lost, self domesticated chimp

We're very likely to :

- see patterns when none exist
- think we can predict the future, cause we think we know how things work ... like:
 - gravity, your car, sunsets
 - weather, the stock market, Covid ...
 - the central dogma

54

Hindsight bias

the knew-it-all-along effect

55

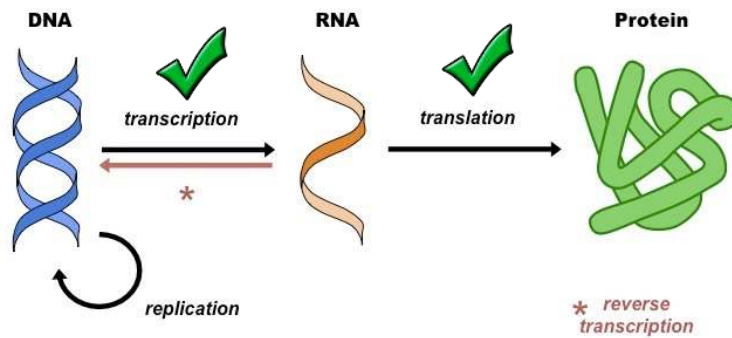
Three Levels of Hindsight Bias



I KNEW
that would happen

56

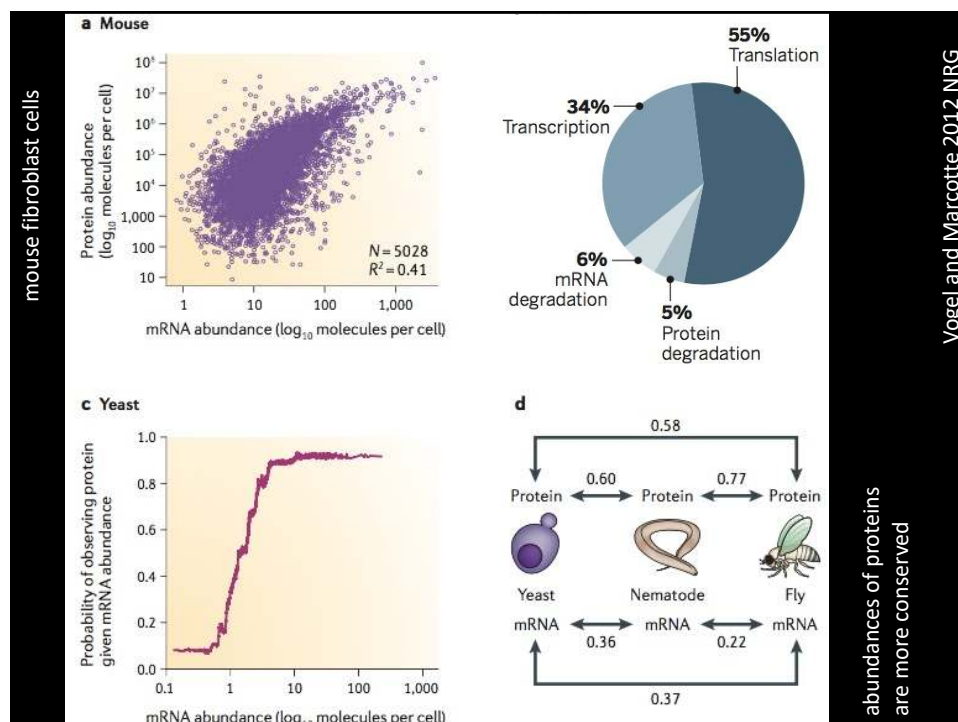
The central dogma



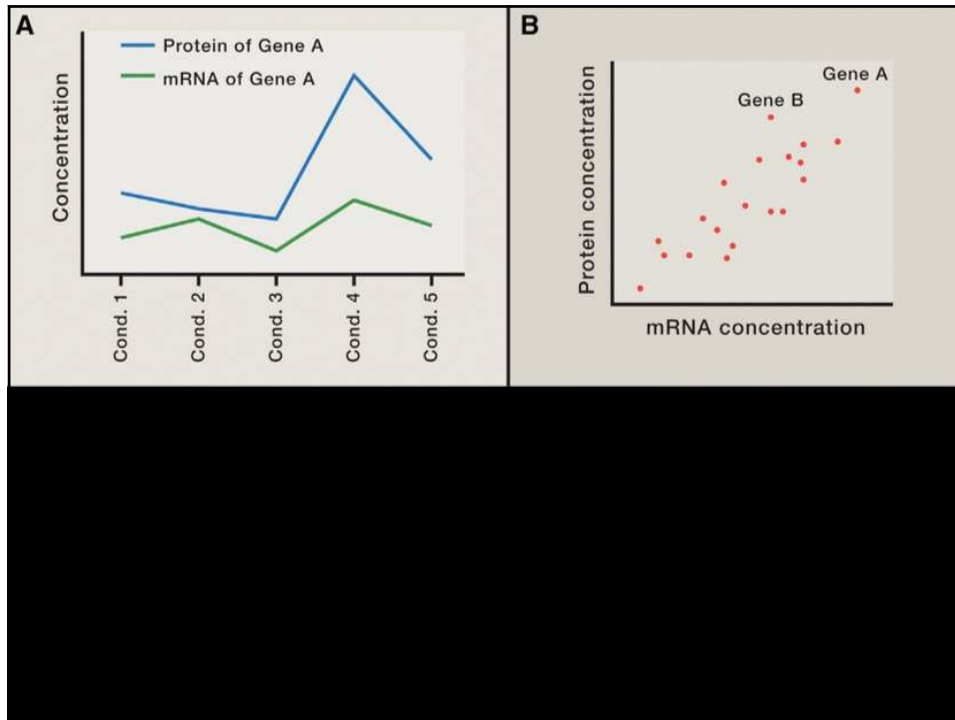
But, can we, in a novel species :

- Predict gene expression level from DNA alone?
- Predict when / where a gene will be expressed from DNA alone?
- Write a protein that will do a specific enzymatic reaction, or several?

57



58



59

Going from peptide sequence to catalytic function ...
"We don't know how to write that way"

Quote in Nobel Prize lecture, 2018
<https://youtu.be/6hOZ5e0g9Uo>

Beethoven's hand written sheet music

Francis Arnold
Nobel Prize winner (2018)

60

In sum, we think we know how things work...

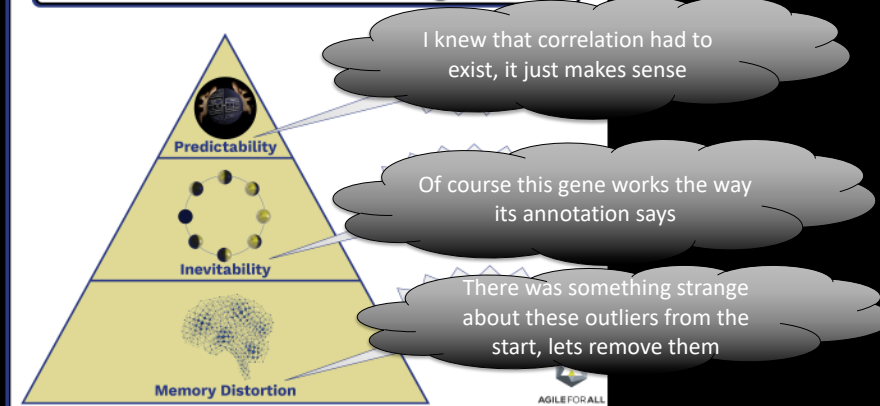
... but biology is exceptionally complex

61

In sum, we think we know how things work...

... but biology is exceptionally complex

Three Levels of Hindsight Bias

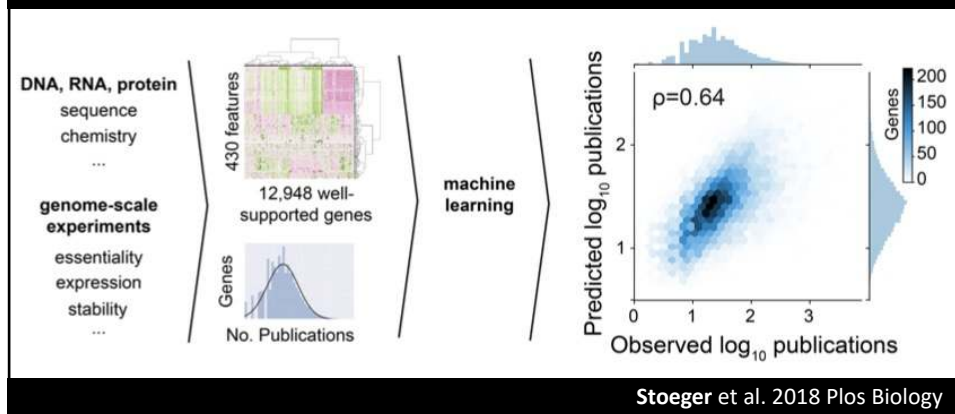


62

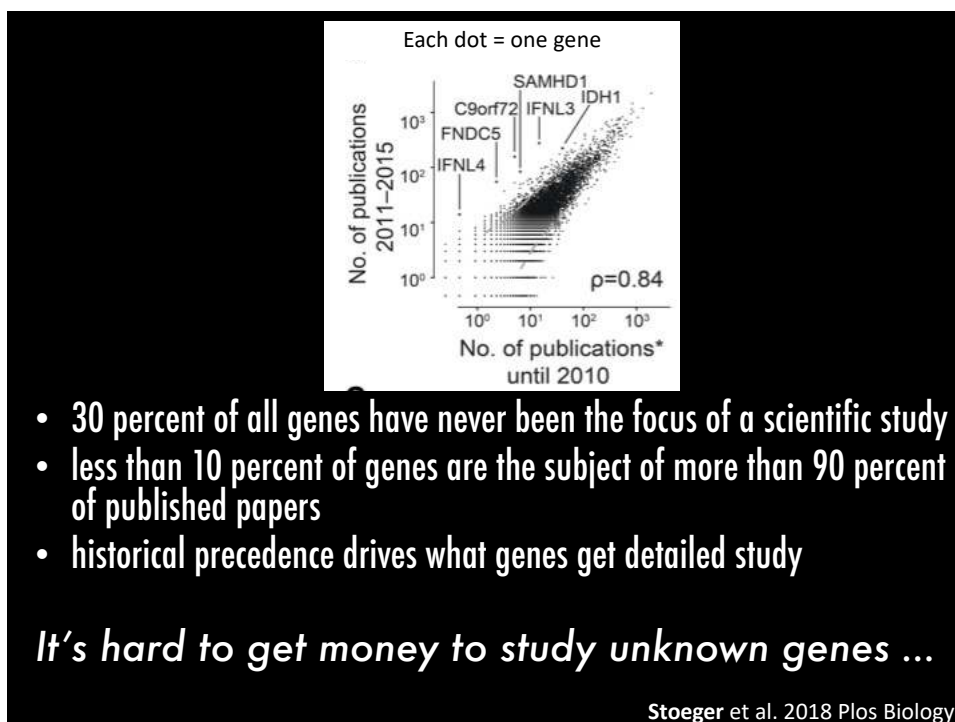
What about the genes we study?

Do we ever conduct “unbiased” investigations?

What if we looked at investigations by gene, over time

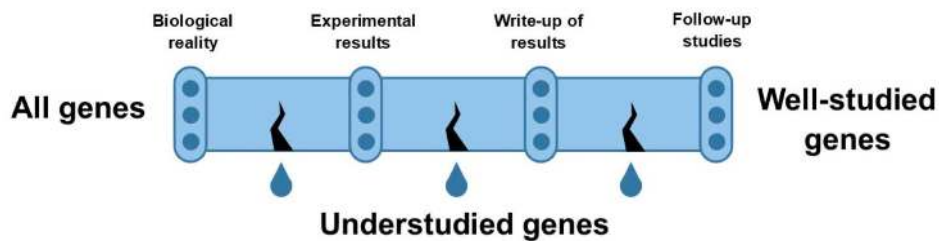


63



64

Understudied genes primarily dropped during writing stage, not due to later follow-up studies



The problem of understudied genes is a consequence of:

- How we view importance
- Draw conclusions
- Use limited space provided in publications in order to sell our story

Richardson et al. 2023 Meta-Research: understudied genes are lost in a leaky pipeline ..

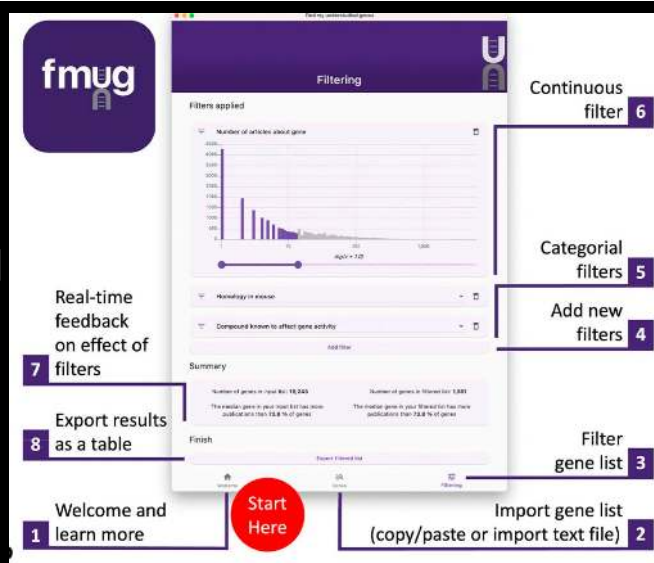
65

fmug “synthesizes data from an array of sources to allow users to identify understudied genes and characterize their tractability for future research”

fmug
Find
My
Understudied
Genes

NOTE: Humans only!!!!

*But good template for
comparative
framework*



66



67