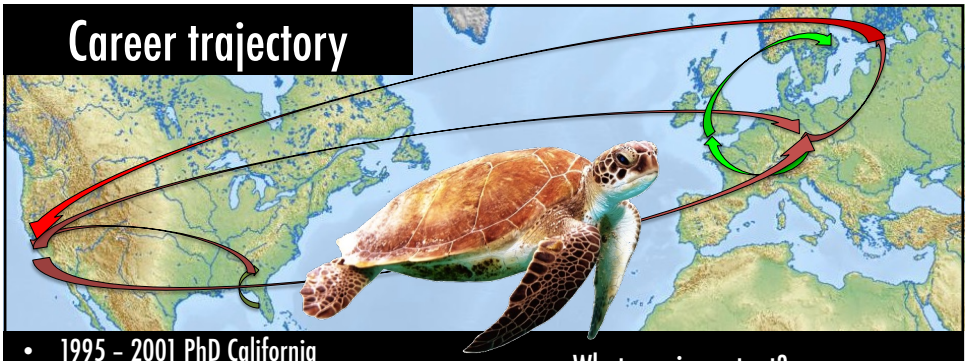
Lies, damn lies, and .... genomics

Navigating your data, your perceptions and reality

Christopher West Wheat
Professor at Department of Zoology

1

Career trajectory



- 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?

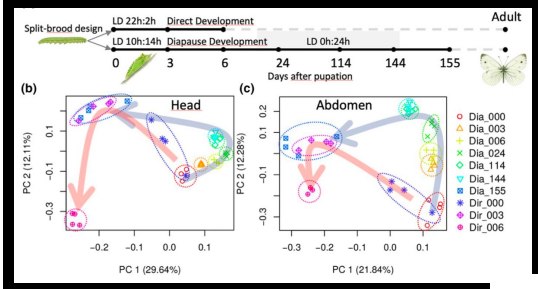
- Being able to move, chase the money & get new skills
- Learning how to believe in my ideas/skills

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

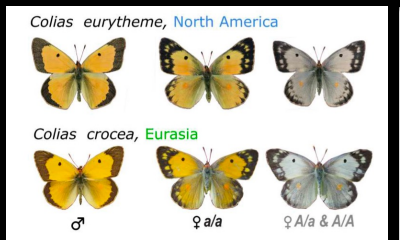
2

Ecological & Evolutionary Functional Genomics

Circadian and seasonal clock



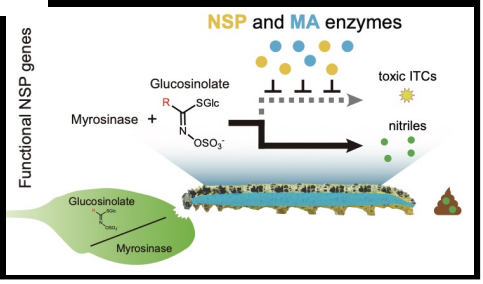
Alternative life history switches



Life history allocation differences between different morphs

	Colored	Alba
Development time [23]	Slower	Faster
Fat-body [23]	Smaller	Larger
Mature eggs at eclosion [23]	Fewer	More
Fecundity [30, 27]	Lower	Higher

Butterfly-plant coevolution dynamics



3

Something you likely would never know about me

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

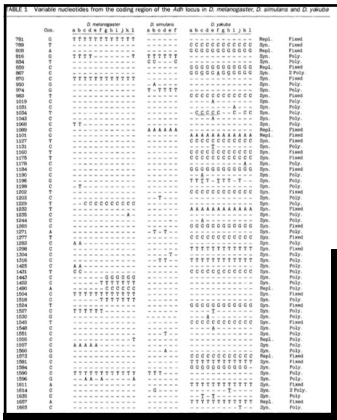
7

What if .....

50% of your favorite studies were not repeatable?

Would that impact your science?

8



Adaptive protein evolution at the *Adh* locus in *Drosophila*

John H. McDonald & Martin Kreitman

Department of Ecology and Evolutionary Biology, Princeton University, Princeton, New Jersey 08544, USA

Nature 1991

UUU, UUC } Phe UUA, UUG } Leu	UCU, UCC, UCA, UCG } Ser	UAU, UAC } Tyr UAA, UAG } Stop	UGU, UGC } Cys UGA Stop UGG Trp
CUU, CUC, CUA, CUG } Leu	CCU, CCC, CCA, CCG } Pro	CAU, CAC } His CAA, CAG } Gln	CGU, CGC, CGA, CGG } Arg

	Con.	*D. melanogaster* a b c d e f g h i j k l	*D. simulans* a b c d e f	*D. yakuba* a b c d e f g h i j k l		
781	G	T T T T T T T T T T T T	– – – – – –	– – – – – – – – – – – –	Repl.	Fixed
789	T	– – – – – – – – – – – –	– – – – – –	C C C C C C C C C C C C	Syn.	Fixed
808	A	– – – – – – – – – – – –	– – – – – –	G G G G G G G G G G G G	Repl.	Fixed
816	G	T T T T – – – – – – – T	T T T T T T	– – – – – – – – – – – –	Syn.	Poly.
834	T	– – – – – – – – – – – –	C C – – – C	– – – – – – – – – – – –	Syn.	Poly.
859	C	– – – – – – – – – – – –	– – – – – –	G G G G G G G G G G G G	Repl.	Fixed
867	C	– – – – – – – – – – – –	– – – – – –	G G G G G A G G G G G G	Syn.	2 Poly.
870	C	T T T T T T T T T T T T	– – – – – –	– – – – – – – – – – – –	Syn.	Fixed

9

Adaptive protein evolution at the *Adh* locus in *Drosophila*

John H. McDonald & Martin Kreitman

Department of Ecology and Evolutionary Biology, Princeton University, Princeton, New Jersey 08544, USA

Nature 1991

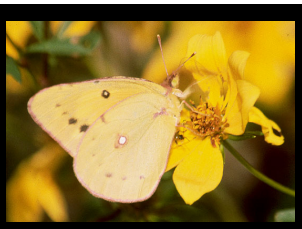
We suggest that these excess replacement substitutions result from adaptive fixation of selectively advantageous mutations.

TABLE 2 Number of replacement and synonymous substitutions for fixed differences between species and polymorphisms within species

	Fixed	Polymorphic
Replacement	7	2
Synonymous	17	42

A *G*-test of independence (with the Williams correction for continuity)[1] was used to test the null hypothesis, that the proportion of replacement substitutions is independent of whether the substitutions are fixed or polymorphic. $G=7.43$, $P=0.006$.

10



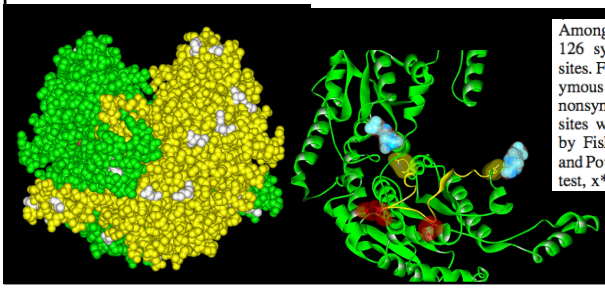
Colias eurytheme

I wanted to us 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

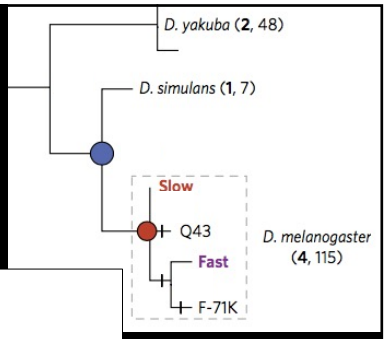


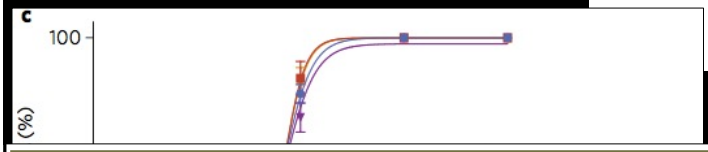
Among *C. eurytheme* and *C. meadii* PGI sequences, we find 126 synonymous and 20 nonsynonymous polymorphic sites. From their ratio, 6.3:1, neutrality predicts ~13 synonymous fixations alongside the two observed interspecies nonsynonymous fixations. But, *no* fixed synonymous sites were found (above). These data differ significantly by Fisher's exact test, $P = 0.021$, following Moriyama and Powell (199[illegible] stein'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.....

my PhD chased an adaptive story lacking a rigorous foundation

Does this happen only in bugs?

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

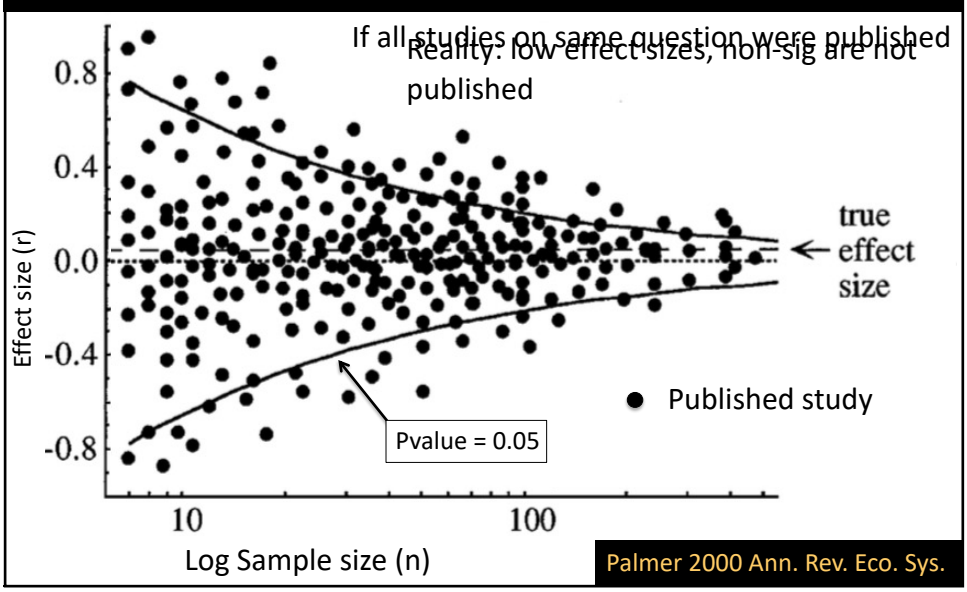
- Biomedical studies
  - Of 49 most cited clinical studies, 45 showed intervention was effective
  - Most were randomized control studies (robust design)

- Mouse cocaine effect study, replicated in three cities
  - Highly standardized study

Ioannidis 2005 *JAMA;* Lehrer 2010

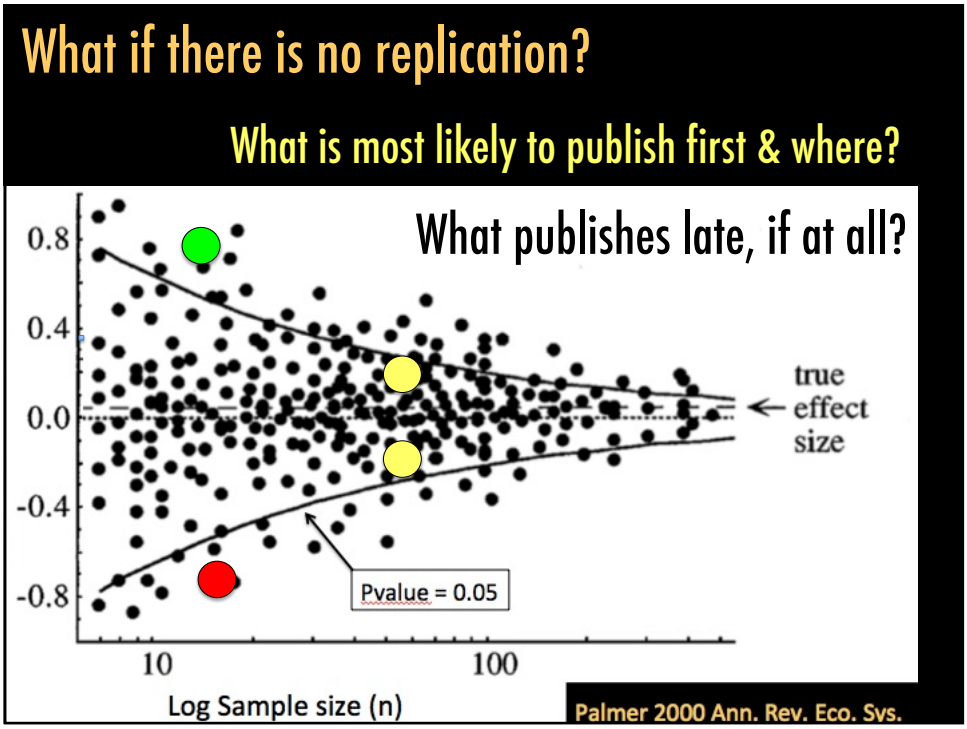
15

Publication bias increases effect size

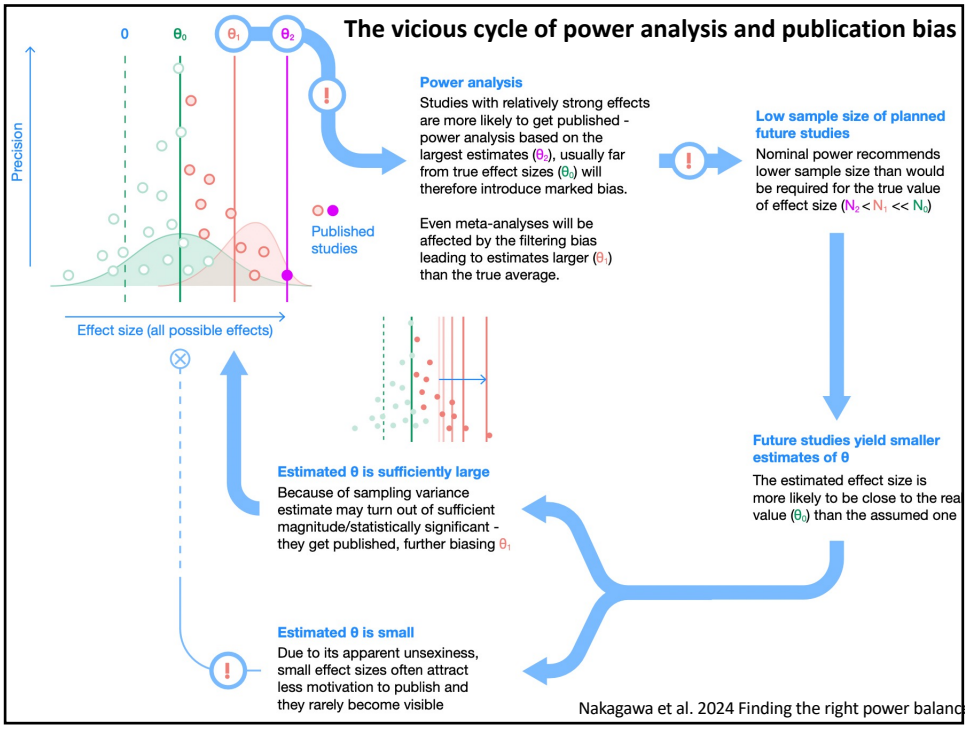


Palmer 2000 Ann. Rev. Eco. Sys.

16


What if there is no replication?
What is most likely to publish first & where?
What publishes late, if at all?
0.8
0.4
0.0
-0.4
-0.8
true effect size
Pvalue = 0.05
10
100
Log Sample size (n)
Palmer 2000 Ann. Rev. Eco. Sys.

17


The vicious cycle of power analysis and publication bias
Precision
Effect size (all possible effects)
Published studies
Power analysis
Studies with relatively strong effects are more likely to get published - power analysis based on the largest estimates (θ_2), usually far from true effect sizes (θ_0) will therefore introduce marked bias.
Even meta-analyses will be affected by the filtering bias leading to estimates larger (θ_1) than the true average.
Low sample size of planned future studies
Nominal power recommends lower sample size than would be required for the true value of effect size ($N_2 < N_1 << N_0$)
Future studies yield smaller estimates of θ
The estimated effect size is more likely to be close to the rea value (θ_0) than the assumed one
Estimated θ is sufficiently large
Because of sampling variance estimate may turn out of sufficient magnitude/statistically significant - they get published, further biasing θ_1
Estimated θ is small
Due to its apparent unsexiness, small effect sizes often attract less motivation to publish and they rarely become visible
Nakagawa et al. 2024 Finding the right power balanc

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 *a 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?

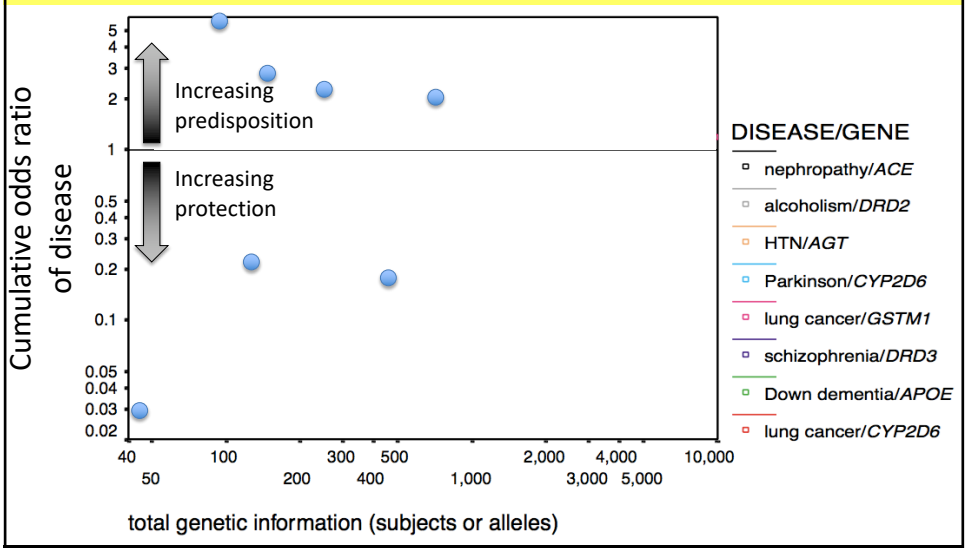
20

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

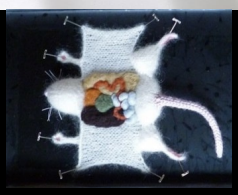
PNAS

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"

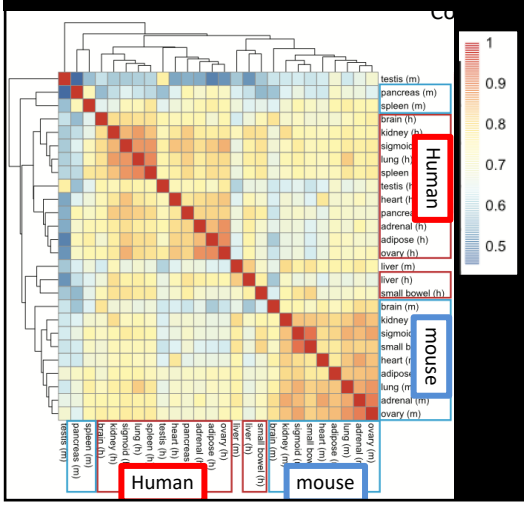
Time of the most recent common ancestor:

Human and Mouse



29

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



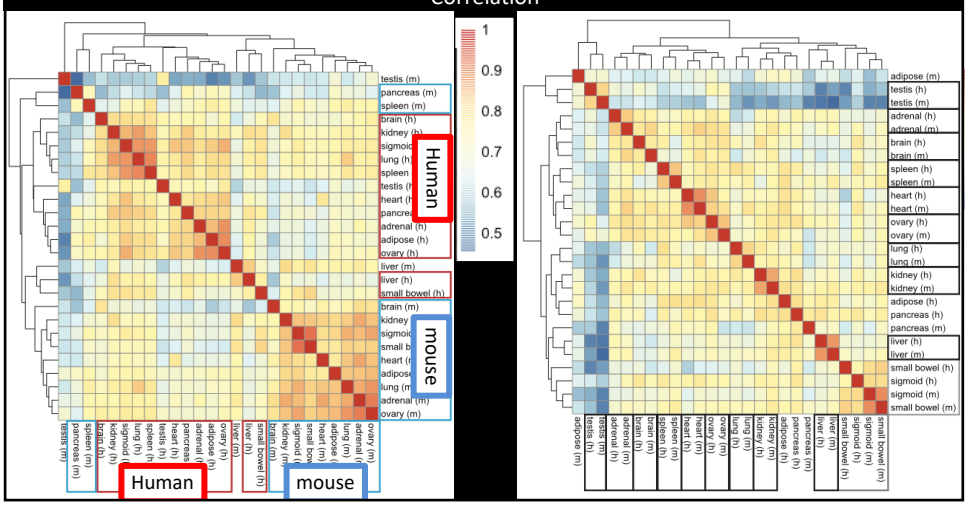
Should gene expression patterns group by species or tissues?

What do we expect from first principals, evolutionary relationships?

30

"the expression for many sets of genes was found to be more similar in different tissues within the same species than between species" Lin et al. 2014 PNAS

"[after accounting] for the batch effect, … human and mouse tend to cluster by tissue, not by species" Gilad and Mizrahi-Man 2015. F1000 Research



31

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

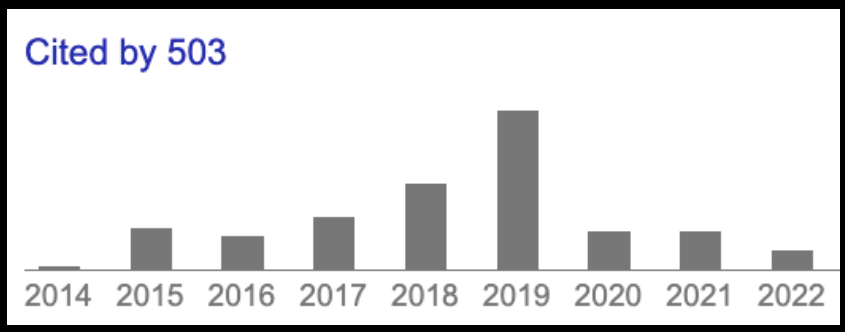
D87PMJN1 (run 253, flow cell D2GUAACXX, lane 7)	D87PMJN1 (run 253, flow cell D2GUAACXX , lane 8)	D4LHBFN1 (run 276, flow cell C2HKJACXX , lane 4)	MONK (run 312, flow cell C2GR3ACXX , lane 6)	HWI-ST373 (run 375, flow cell C3172ACXX , lane 7)
heart	adipose	adipose	heart	brain
kidney	adrenal	adrenal	kidney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

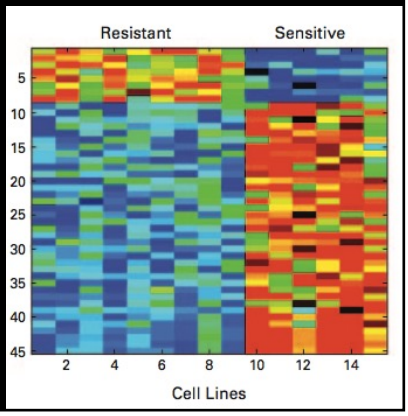


33

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. reponsive 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,

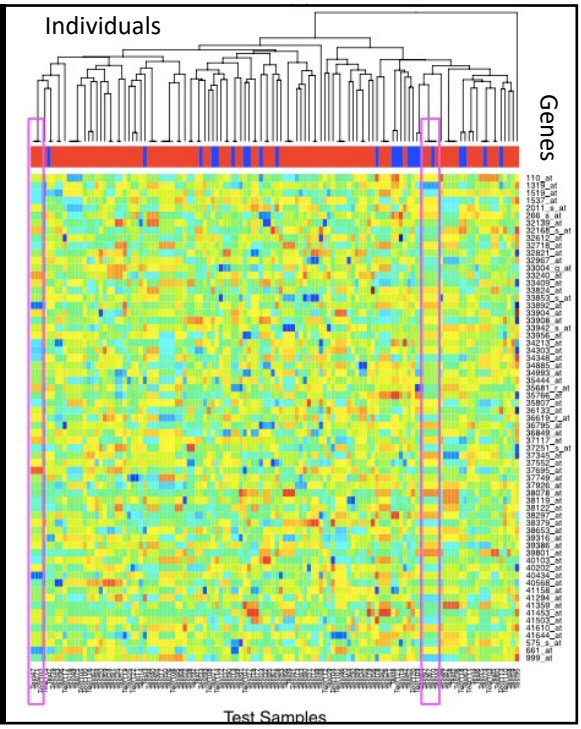
Thanks: Malachi Griffith

Baggerly and Coombes 2009

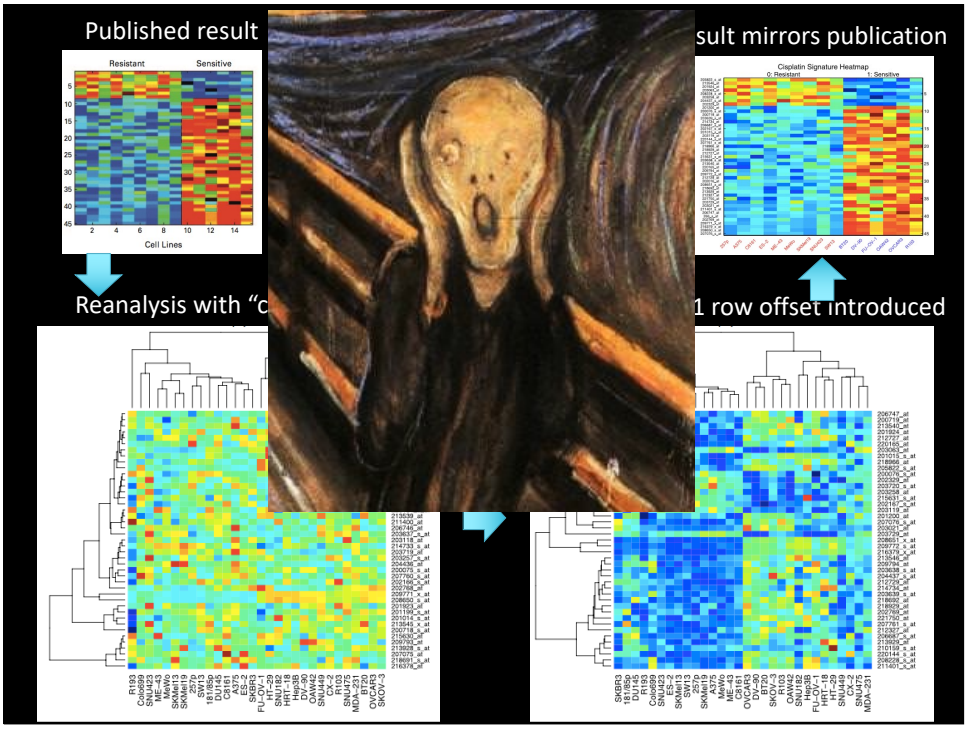
35

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

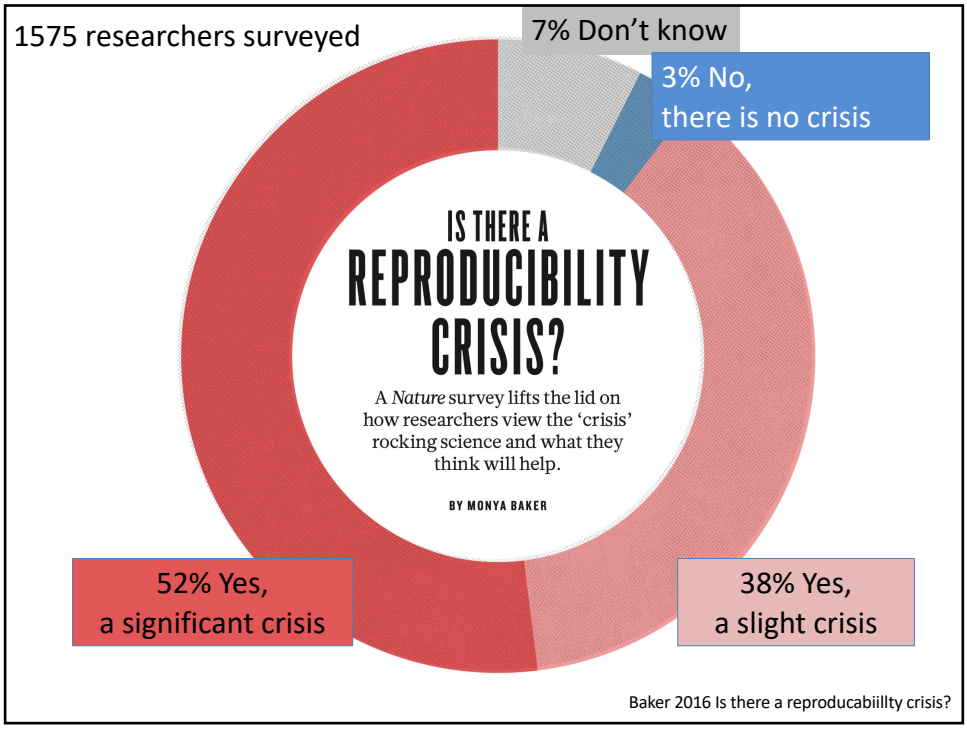
38

Can we reduce these type of publications?

YES!!!!!

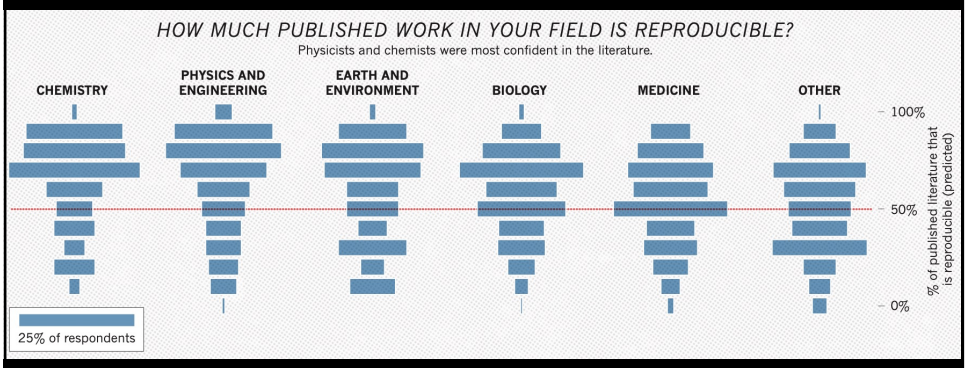
- Work better as a community, check each others code
- As author, as supervisor, as reviewer, as Associate Editor, make sure all studies you touch :
  - Have all code and raw data open source
  - Analyzed datasets open source
  - Methods clearly described

39



40

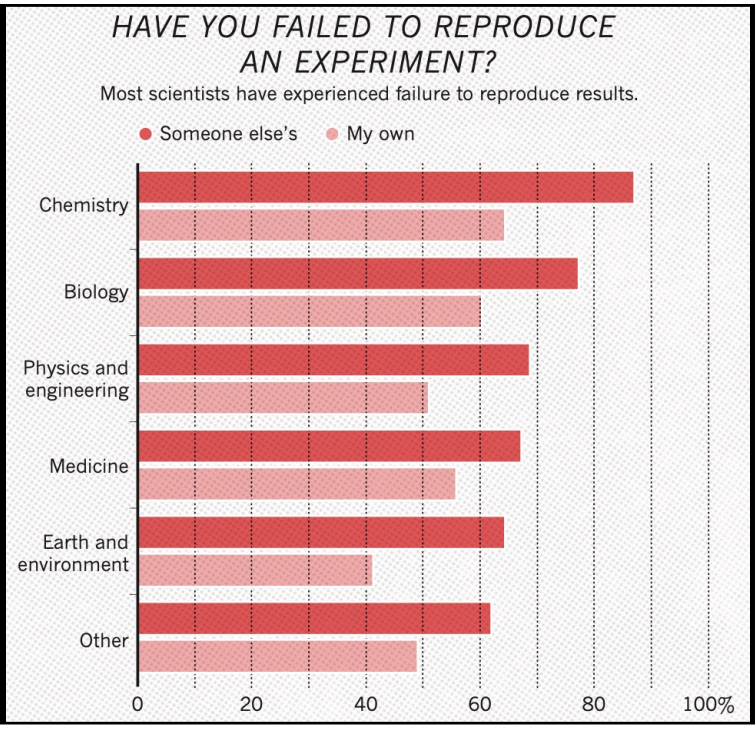
Breakdown by research field



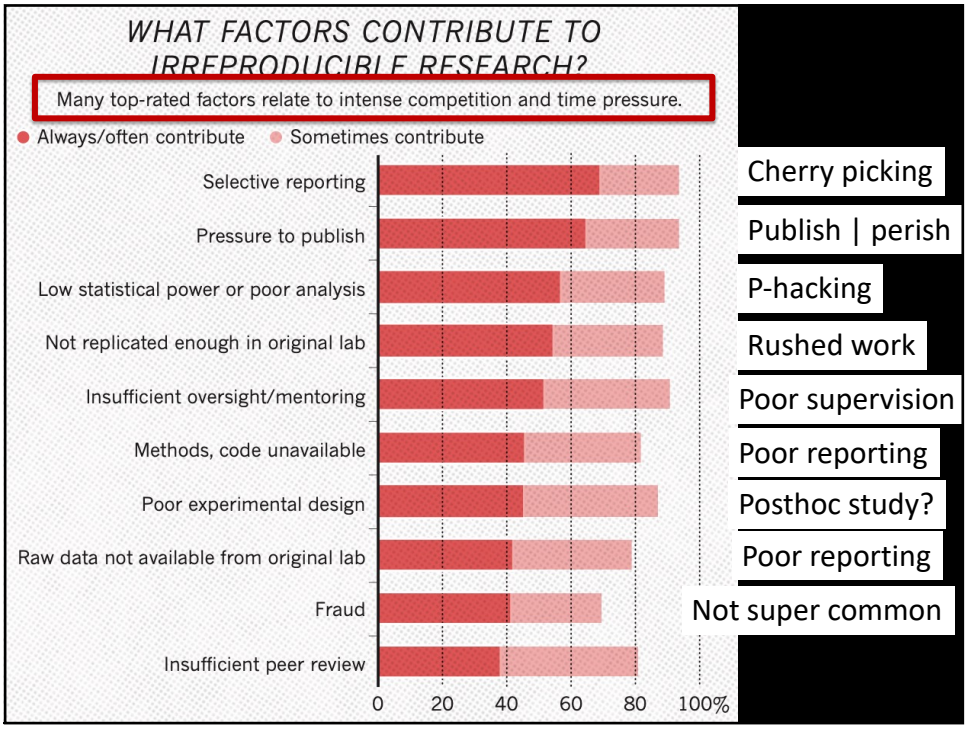
Number of respondents from each discipline:
Biology **703**, Chemistry **106**, Earth and environmental **95**, Medicine **203**, Physics and engineering **236**, Other **233**

Baker 2016 Is there a reproducabiillty crisis?

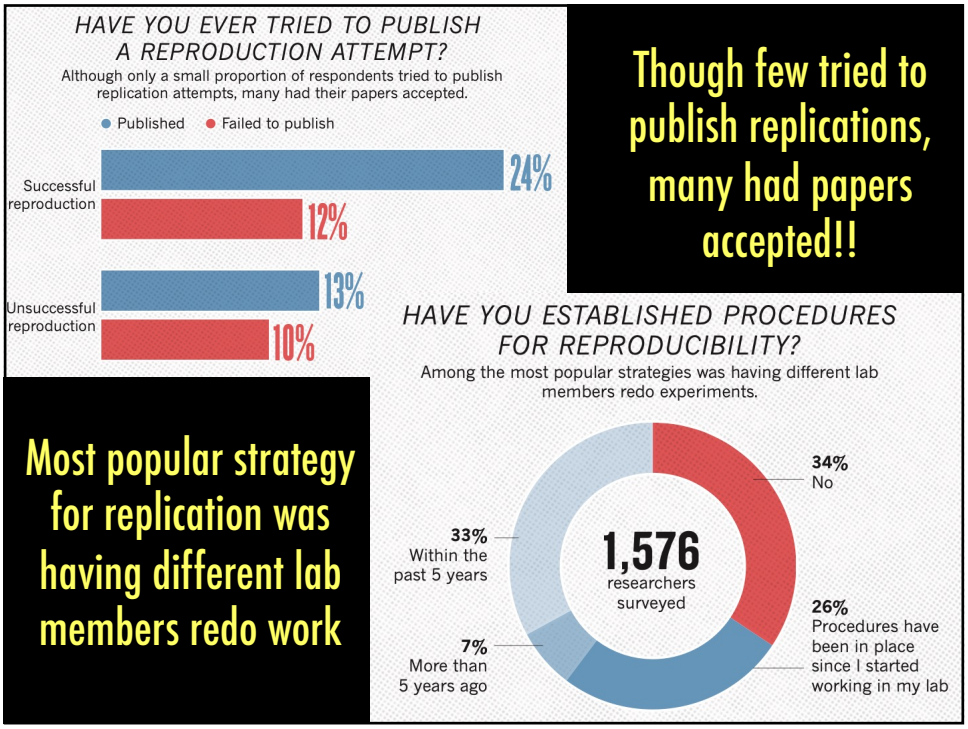
41



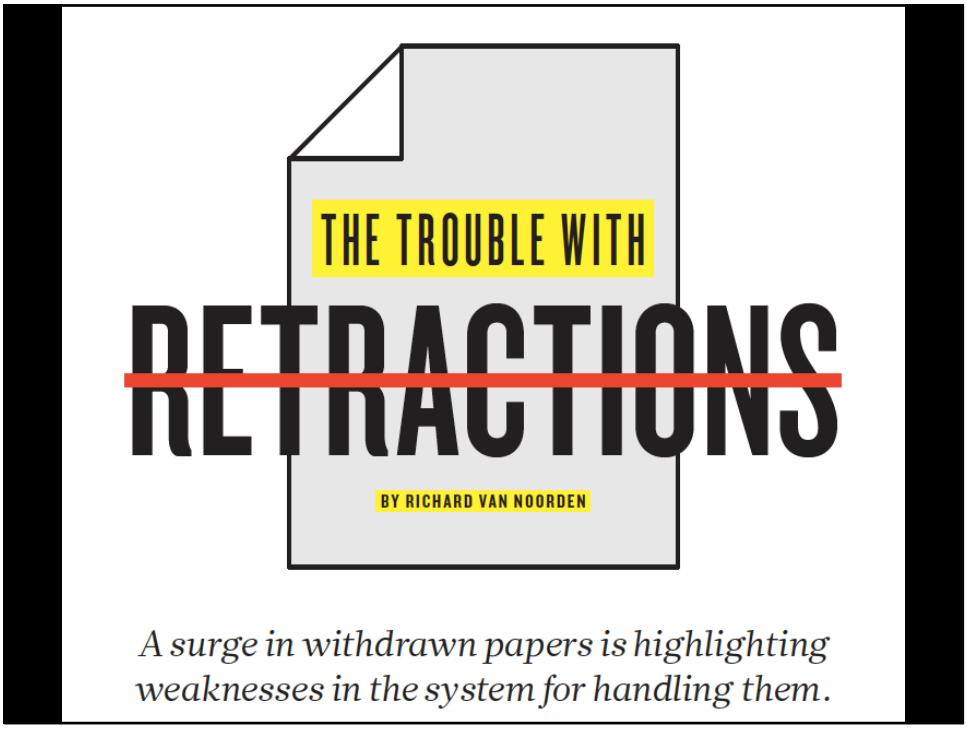
42


WHAT FACTORS CONTRIBUTE TO IRREPRODUCIBLE RESEARCH?
Many top-rated factors relate to intense competition and time pressure.
Always/often contribute
Sometimes contribute
Selective reporting
Pressure to publish
Low statistical power or poor analysis
Not replicated enough in original lab
Insufficient oversight/mentoring
Methods, code unavailable
Poor experimental design
Raw data not available from original lab
Fraud
Insufficient peer review
0
20
40
60
80
100%
Cherry picking
Publish | perish
P-hacking
Rushed work
Poor supervision
Poor reporting
Posthoc study?
Poor reporting
Not super common

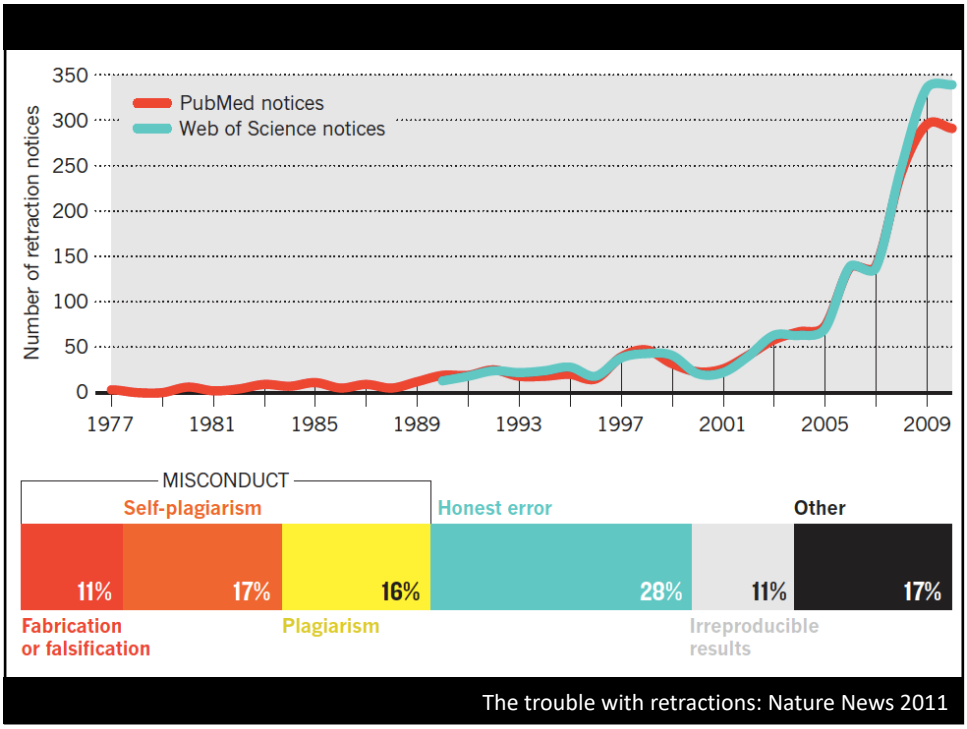
43


HAVE YOU EVER TRIED TO PUBLISH A REPRODUCTION ATTEMPT?
Although only a small proportion of respondents tried to publish replication attempts, many had their papers accepted.
Published
Failed to publish
Successful reproduction
24%
12%
Unsuccessful reproduction
13%
10%
Though few tried to publish replications, many had papers accepted!!
HAVE YOU ESTABLISHED PROCEDURES FOR REPRODUCIBILITY?
Among the most popular strategies was having different lab members redo experiments.
1,576
researchers surveyed
34%
No
26%
Procedures have been in place since I started working in my lab
33%
Within the past 5 years
7%
More than 5 years ago
Most popular strategy for replication was having different lab members redo work

44

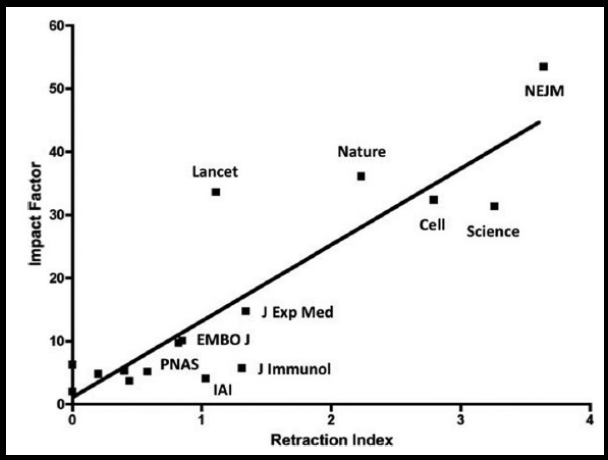


45



The trouble with retractions: Nature News 2011

46



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

Fang 2011 Infect. Immun.

47

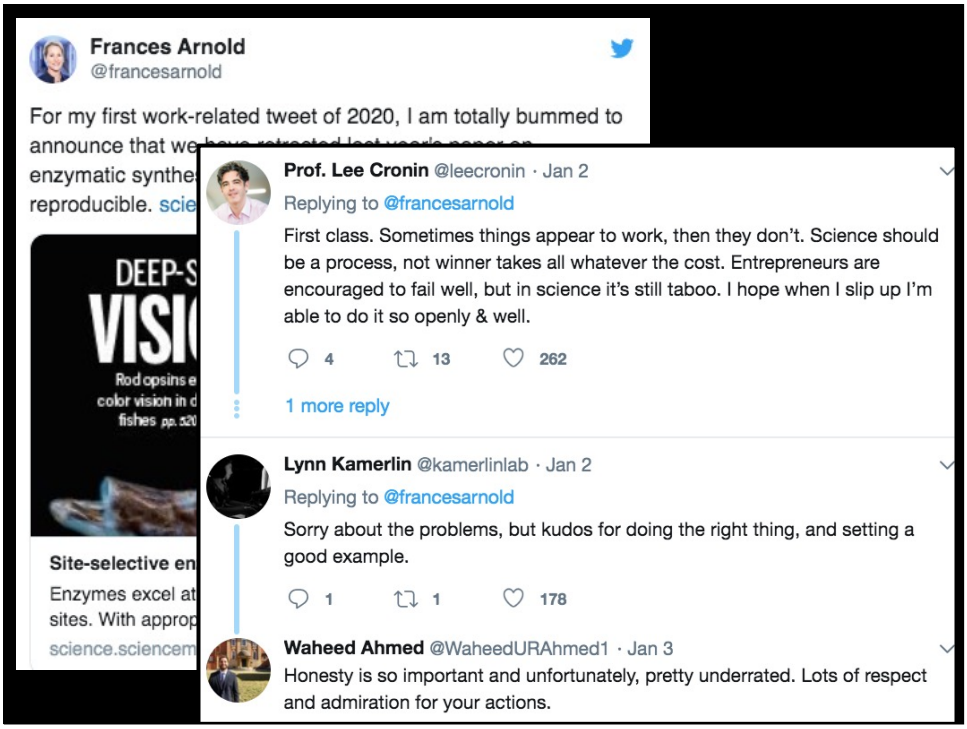
- Website shows retraction

48

Retraction Watch

- 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



50

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

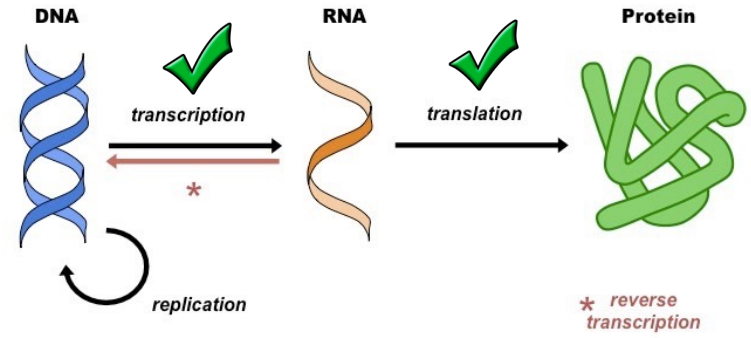
the knew-it-all-along effect

55

Three Levels of Hindsight Bias

56

The central dogma

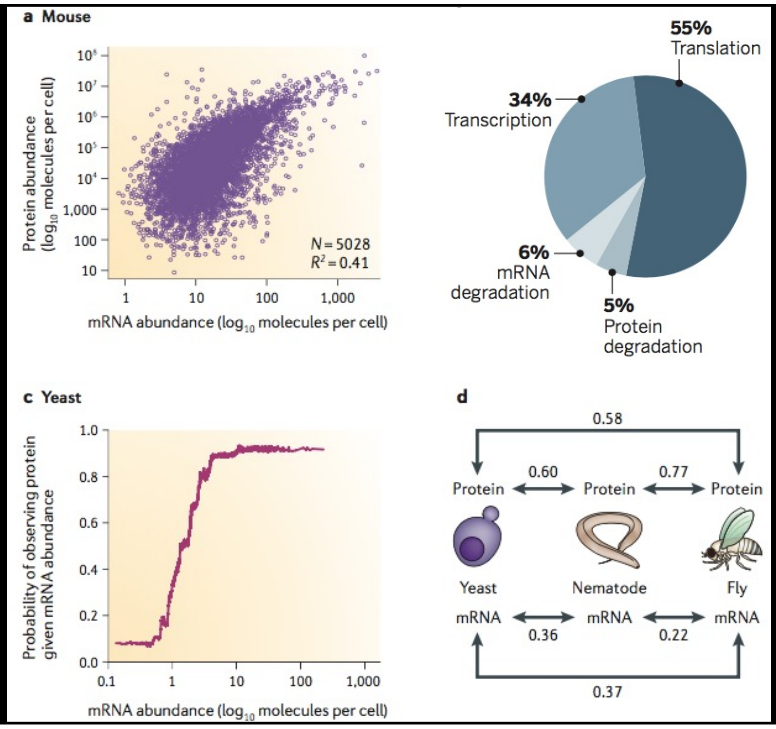


<u>But, can we, in a novel species</u> :

- 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

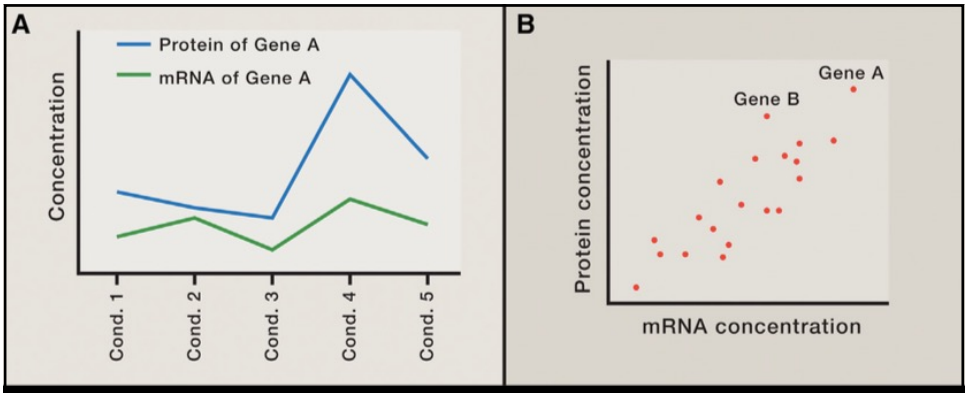
mouse fibroblast cells



Vogel and Marcotte 2012 NRG

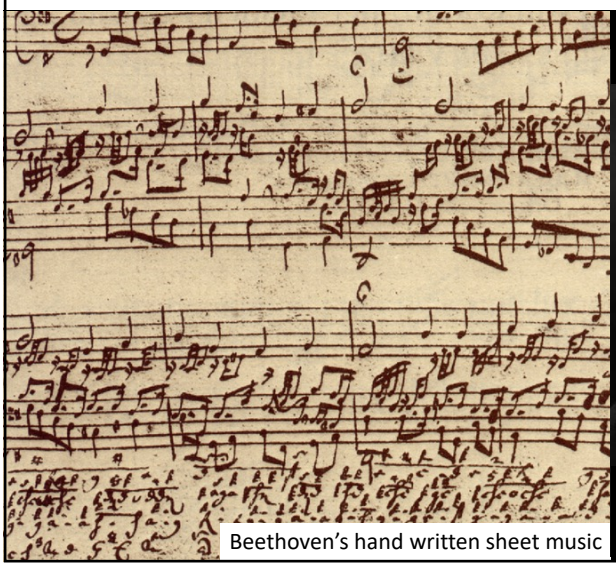
abundances of proteins are more conserved

58



59

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



Beethoven's hand written sheet music

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

Francis Arnold
Nobel Prize winner (2018)

60

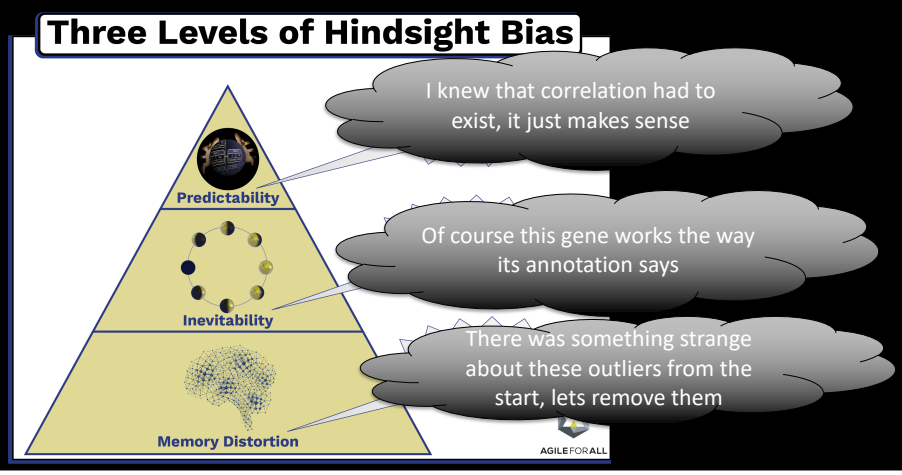
In sum, we think we how things work…

… but biology is exceptionally complex

61

In sum, we think we how things work…

… but biology is exceptionally complex

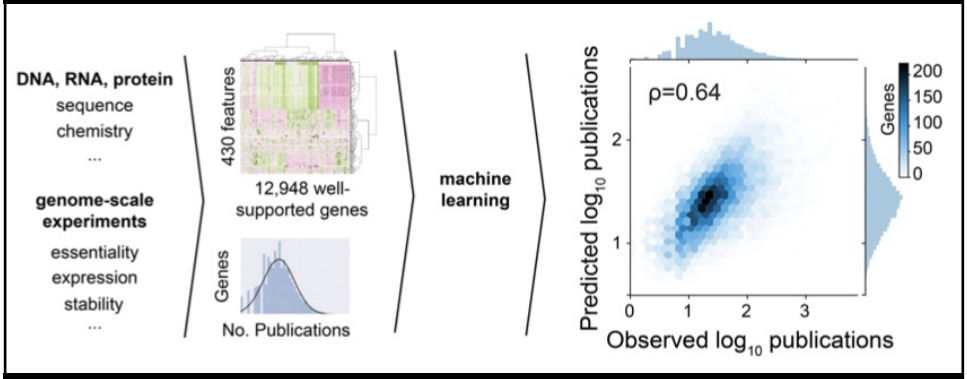


62

What about the genes we study?

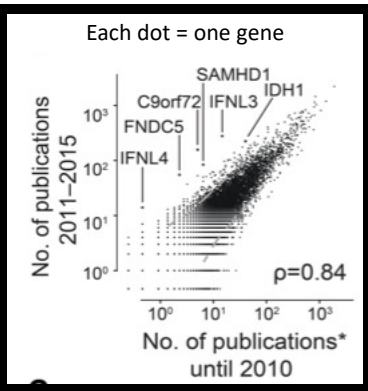
Do we ever conduct "unbiased" investigations?

What if we looked at investigations by gene, over time



Stoeger et al. 2018 Plos Biology

63



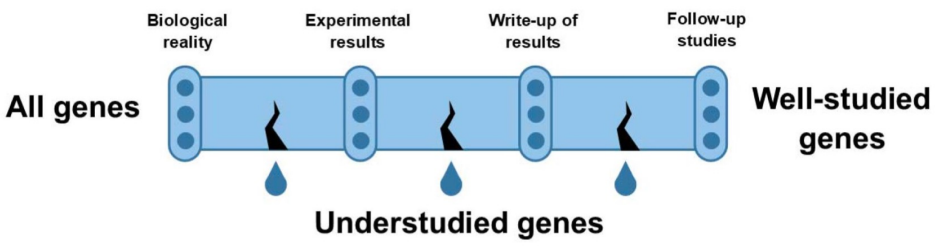
- 30 percent of all genes have never been the focus of a scientific study
- less than 10 percent of genes are the subject of more than 90 percent of published papers
- historical precedence drives what genes get detailed study

It's hard to get money to study unknown genes ...

Stoeger et al. 2018 Plos Biology

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

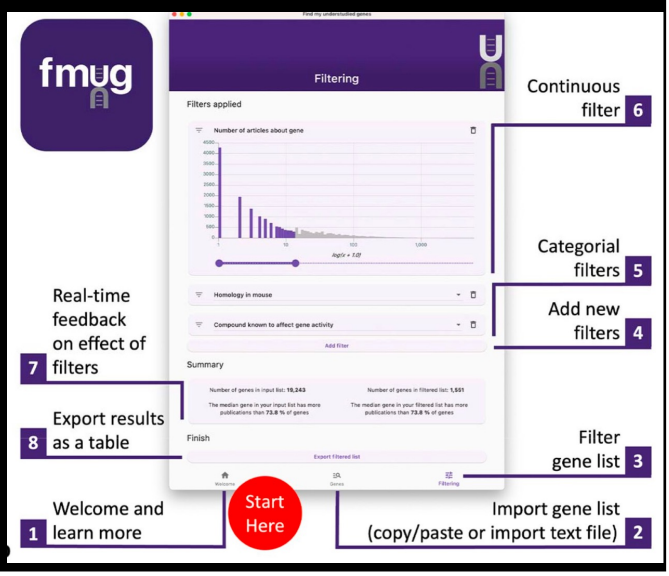
fmug "synthesizes data from an array of sources to allow users to identify understudied genes and characterize their tractability for future research"

fmug

Find
My
Understudied
Genes

NOTE: Humans only!!!!

But good template for comparative framework



66

67