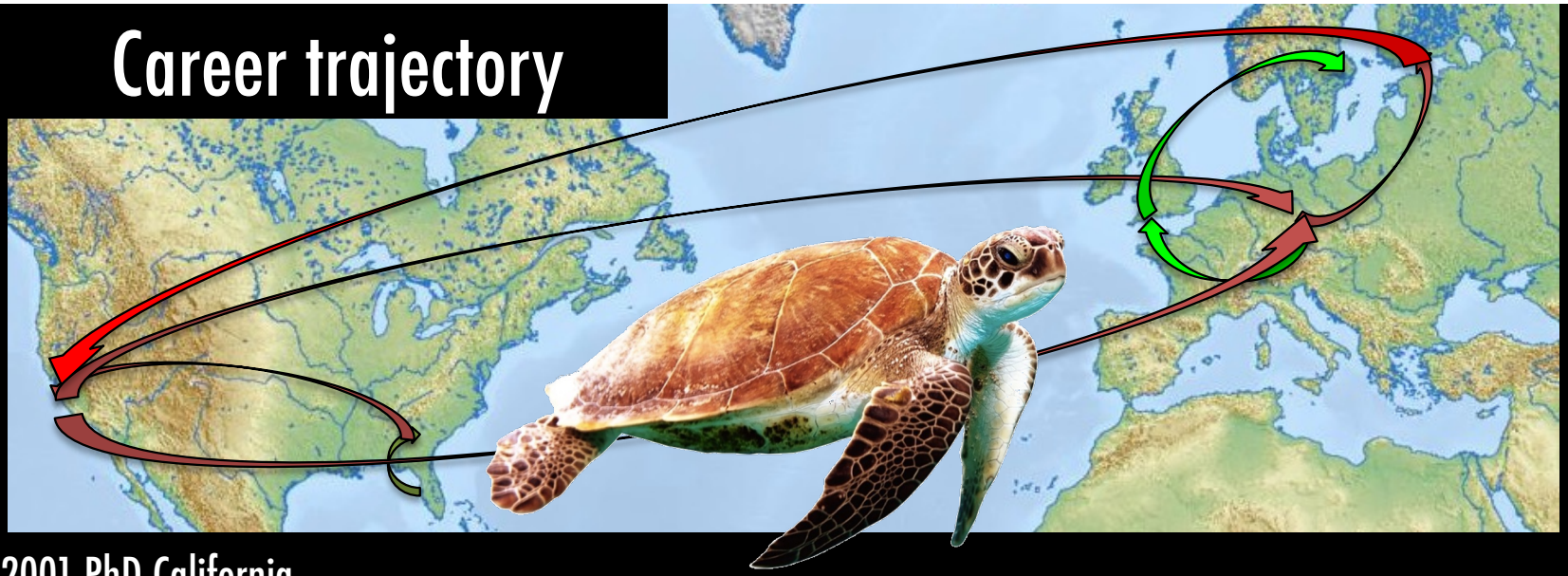
Lies, damn lies, and .... genomics

Navigating your data, your perceptions and reality

Christopher West Wheat
Professor at Department of Zoology

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 in Finland
- 2012 – Assistant Prof. at Stockholm University
- 2022 – Full Professor

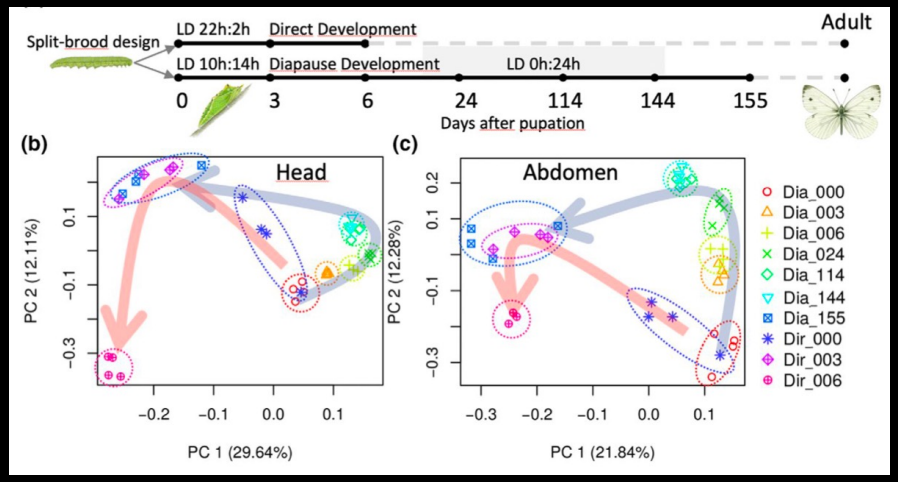
What was important?

- Being able to move, chase the money & get skills
- Learning how to believe in my ideas/skills
- Writing lots of grants, get used to rejections

I was able to put science first & have fun along the way

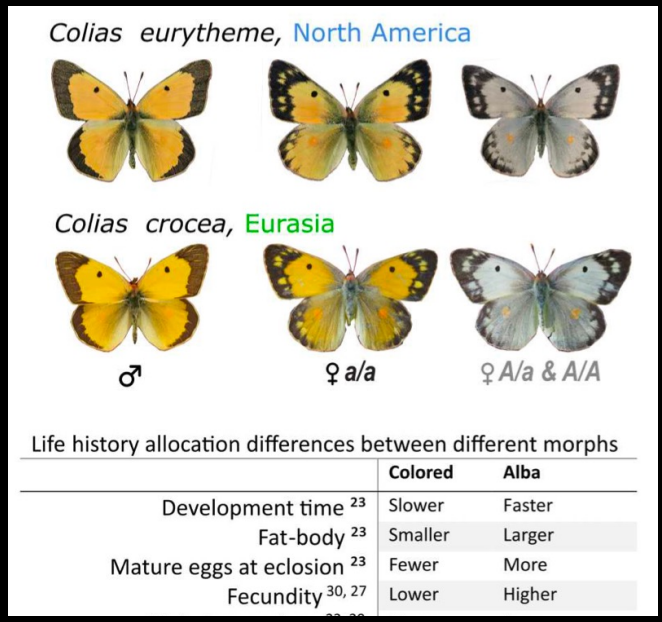
Ecological & Evolutionary Functional Genomics

Circadian and seasonal clock evolution



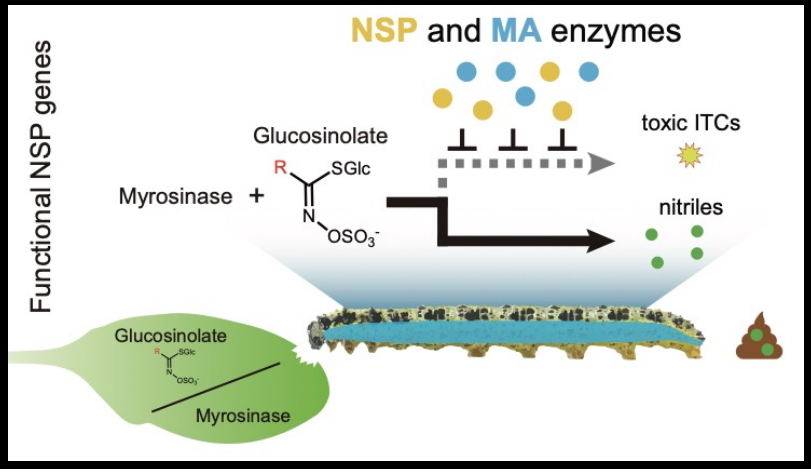
Butterfly-plant coevolution dynamics

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



Something you likely would never know about me

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

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

Disclaimer

I'm a positive person

I love my job and the work we all do

My goal here is to provoke you into think critically

What if .....

50% of your favorite studies were not repeatable?

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

TABLE 1 Variable nucleotides from the coding region of the *Adh* locus in *D. melanogaster*, *D. simulans* and *D. yakuba*

	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 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
950	G	- - - - - - - - - - - -	- A - - - -	- - - - - - - - - - - -	Syn.	Poly.
974	G	- - - - - - - - - - - -	T - T T T T	- - - - - - - - - - - -	Syn.	Poly.
983	T	- - - - - - - - - - - -	- - - - - -	C C C C C C C C C C C C	Syn.	Fixed
1019	C	- - - - - - - - - - - -	- - - - - -	- - - - A - - - - - - -	Syn.	Poly.
1031	C	- - - - - - - - - - - -	- - - - - -	- - - - - - - - A - - -	Syn.	Poly.
1034	T	- - - - - - - - - - - -	- - - - - -	- C C C C - - C - C C	Syn.	Poly.
1043	C	- - - - - - - - - - - -	- - - - - -	- - - - A - - - - - - -	Syn.	Poly.
1068	C	T T - - - - - - - - - -	- - - - - -	- - - - - - - - - - - -	Syn.	Poly.
1089	C	- - - - - - - - - - - -	A A A A A A	- - - - - - - - - - - -	Repl.	Fixed
1101	G	- - - - - - - - - - - -	- - - - - -	A A A A A A A A A A A A	Repl.	Fixed
1127	T	- - - - - - - - - - - -	- - - - - -	C C C C C C C C C C C C	Syn.	Fixed
1131	C	- - - - - - - - - - - -	- - - - - -	- - - - T - - - - - - -	Syn.	Poly.
1160	T	- - - - - - - - - - - -	- - - - - -	C C C C C C C C C C C C	Syn.	Fixed
1175	T	- - - - - - - - - - - -	- - - - - -	C C C C C C C C C C C C	Syn.	Fixed
1178	C	- - - - - - - - - - - -	- - - - - -	- - - - - - - - - - A -	Syn.	Poly.
1184	C	- - - - - - - - - - - -	- - - - - -	G G G G G G G G G G G G	Syn.	Fixed
1190	C	- - - - - - - - - - - -	- - - - - -	- - A - - - - - - - - -	Syn.	Poly.
1196	G	- - - - - - - - - - - -	- - - - - -	T T T - T T T - T - - -	Syn.	Poly.
1199	C	- T - - - - - - - - - -	- - - - - -	- - - - - - - - - - - -	Syn.	Poly.
1202	T	- - - - - - - - - - - -	- - - - - -	C C C C C C C C C C C C	Syn.	Fixed
1203	C	- - - - - - - - - - - -	- - T - - -	- - - - - - - - - - - -	Syn.	Poly.
1229	T	- - C C C C C C C C C C	- - - - - -	- - - - - - - - - - - -	Syn.	Poly.
1232	T	- - - - - - - - - - - -	- - - - - -	A A A A A A A A A A A A	Syn.	Fixed
1235	C	- - - - - - - - - A - -	- - - - - -	- - - - - - - - - - - -	Syn.	Poly.
1244	C	- - - - - - - - - - - -	- - - - - -	- - A - - - - - - - - -	Syn.	Poly.
1265	C	- - - - - - - - - - - -	- - - - - -	G G G G G G G G G G G G	Syn.	Fixed
1271	A	- - - - - - - - - - - -	- T - T - -	- - - - - - - - - - - -	Syn.	Poly.
1277	T	- - - - - - - - - - - -	- - - - - -	C C C C C C C C C C C C	Syn.	Fixed
1283	C	A A - - - - - - - - - -	- - - - - -	- - - - - - - - - - - -	Syn.	Poly.
1298	C	- - - - - - - - - - - -	- - - - - -	T T T T T T T T T T T T	Syn.	Fixed
1304	C	- - - - - - - - - - - -	- - - - T -	- - - - - - - - - - - -	Syn.	Poly.
1316	C	- - - - - - - - - - - -	- - T T - -	T T T T T T T T T T T T	Syn.	Poly.
1425	C	A A - - - - - - - - - -	- - - - - -	- - - - - - - - - - - -	Syn.	Poly.

	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 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

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

John H. McDonald & Martin Kreitman

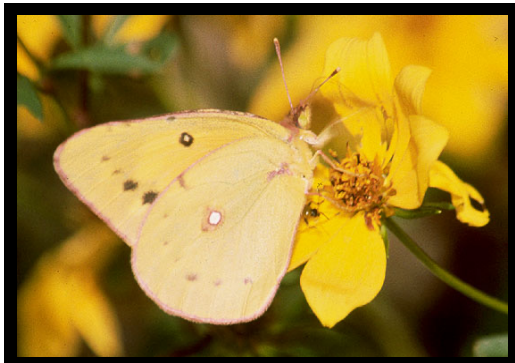
Department of Ecology and Evolutionary Biology, Princeton University, Princeton, New Jersey 08544, USA

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. $G=7.43$, $P=0.006$.



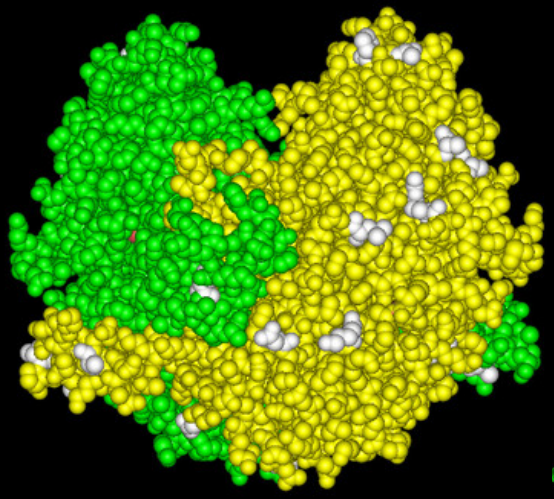
Colias eurytheme

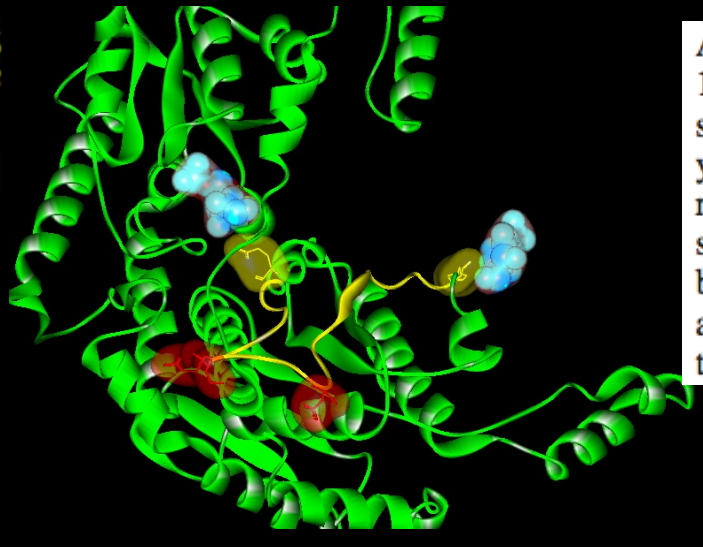
My PhD: use this DNA based molecular test of selection on a classic example of balancing selection from allozyme era

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

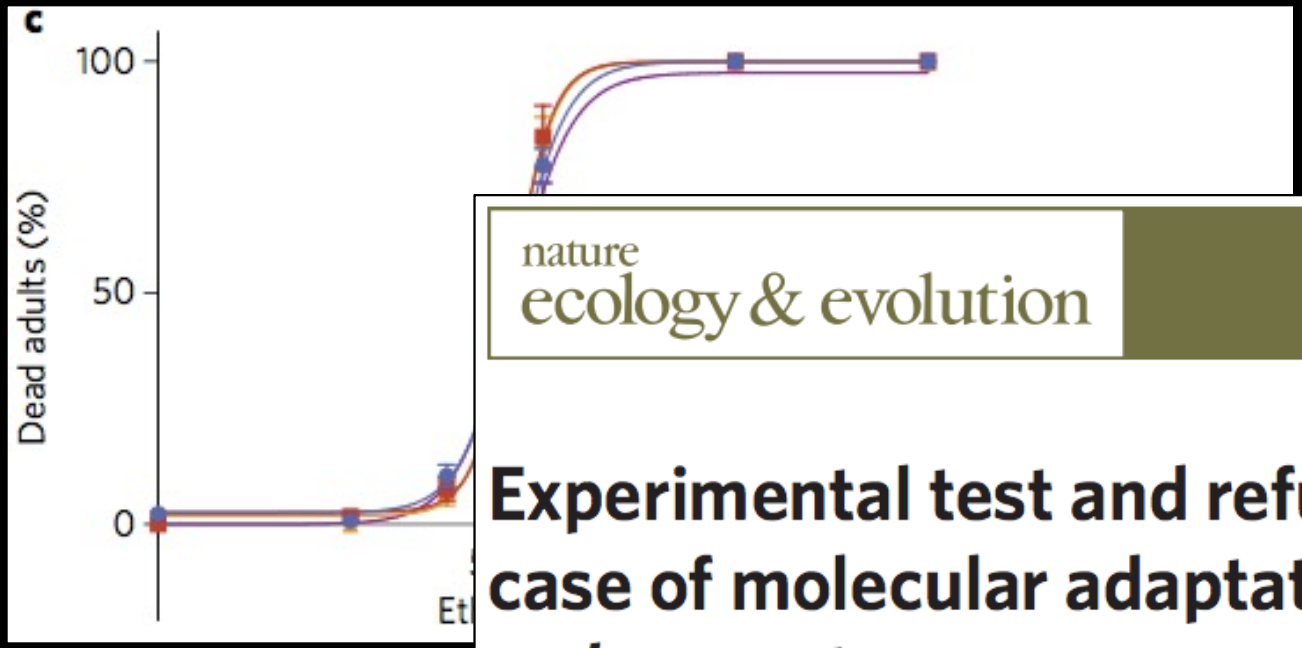


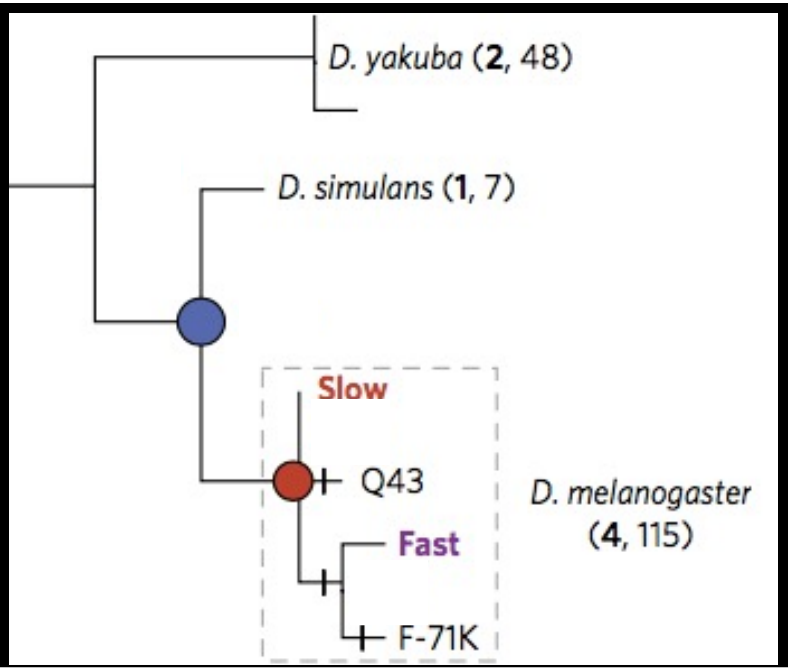


Among *C. eurytheme* and *C. meadii* PGI sequences, we find 126 synonymous and 20 nonsynonymous polymorphic sites. From their ratio, 6.3:1, neutrality predicts ~13 synonymous fixations alongside the two observed interspecies nonsynonymous fixations. But, *no* fixed synonymous sites were found (abo… …a differ significantly by Fisher's exact test $P = 0.021$, following Moriyama and Powell (1996) and by Goldstein's (1964) exact binomial test, $x^* = 3.41$, $P = 0.0006$.

Wheat et al. 2005

30 years later, these MK test results in *Drosophila melanogaster* were revisited …





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*

So .....

my PhD chased an adaptive story lacking a rigorous foundation

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

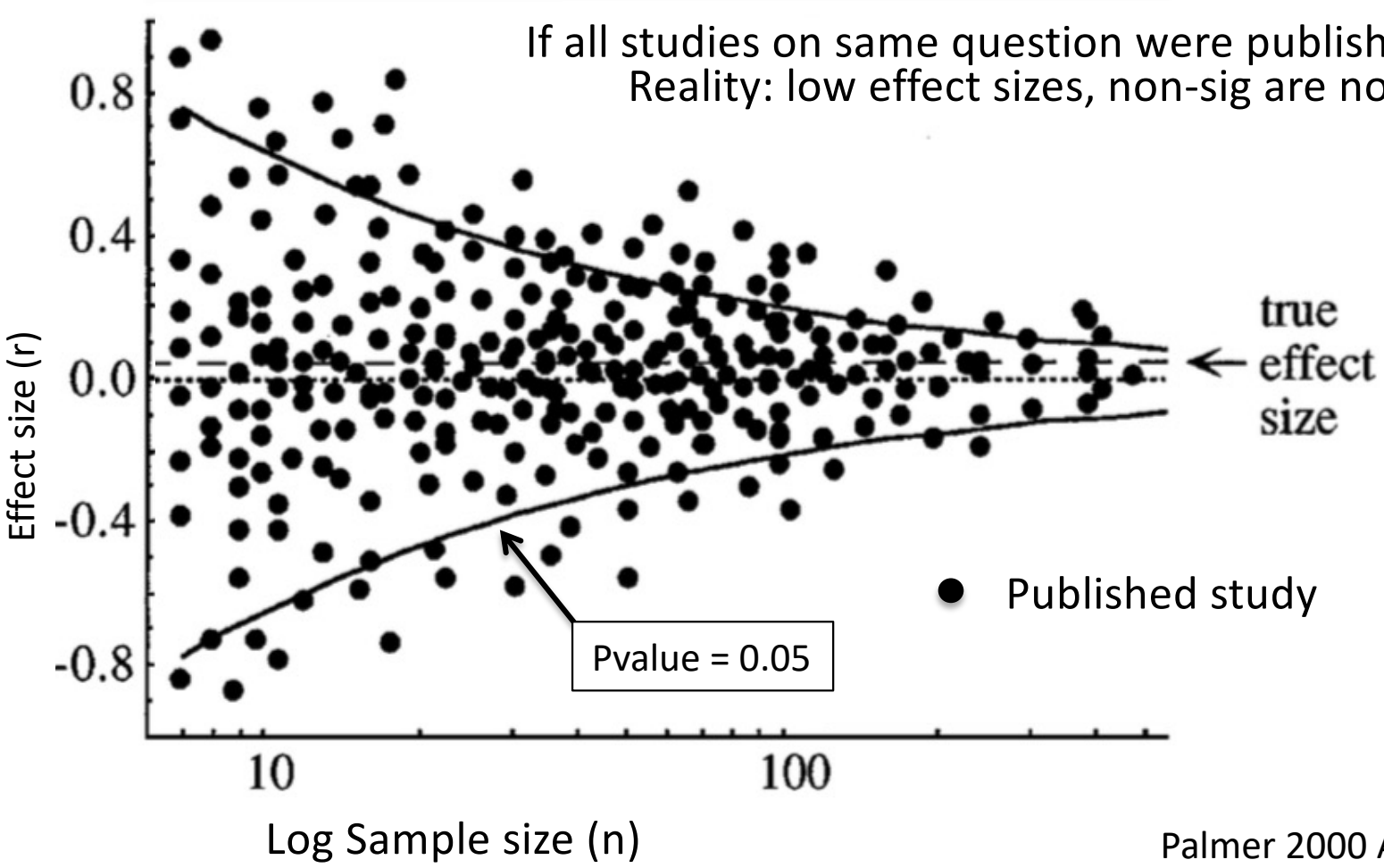
Publication replication failures

- 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

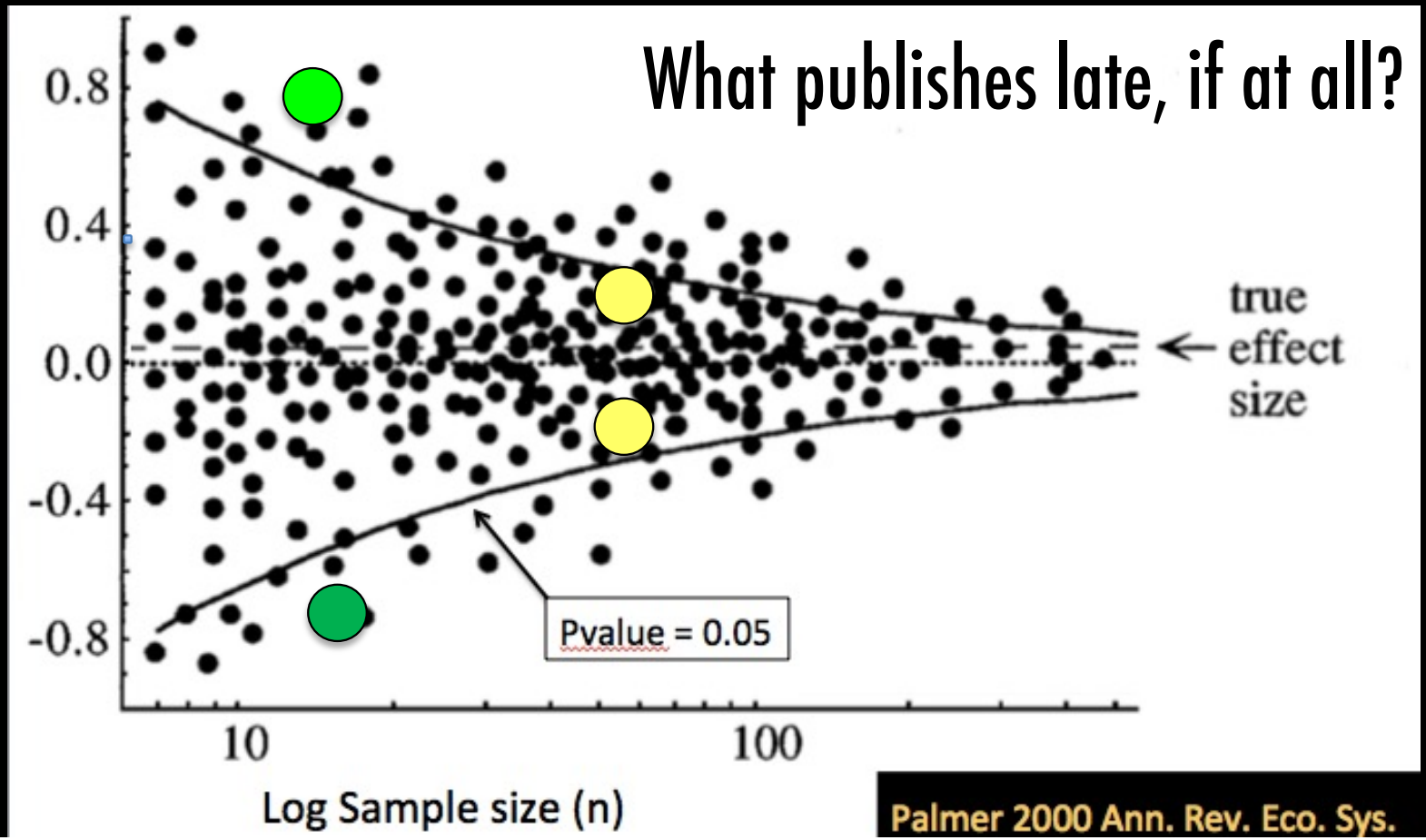
Publication bias can increase effect size



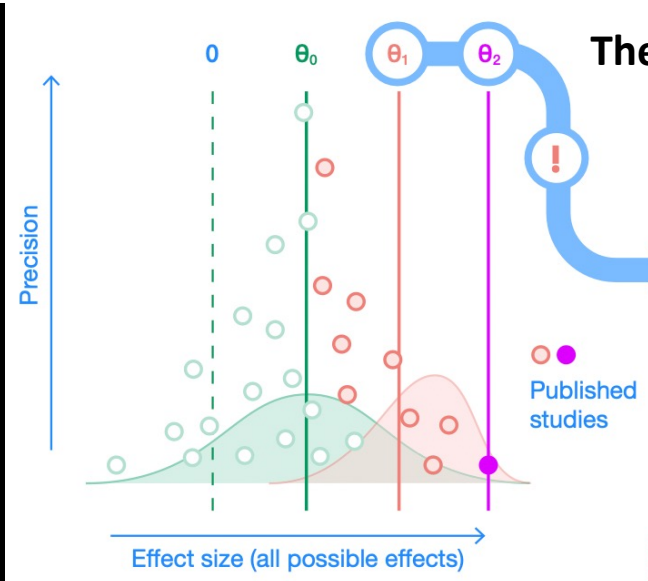
Palmer 2000 Ann. Rev. Eco. Sys.

What if there is no replication?

What is most likely to publish first & where?



The vicious cycle of power analysis and publication bias



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 real 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 balance

Why Most Published Research Findings Are False

Ioannidis 2005 Plos Med.

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, & analyses
- when there is greater financial and other interest and prejudice
- when more teams are involved, all chasing after statistical significance by using different tests

Which of these apply to genomics?

- ✓ 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, & analyses
- when there is greater financial and other interest and prejudice
- ✓ when more teams are involved, all chasing after statistical significance by using different tests

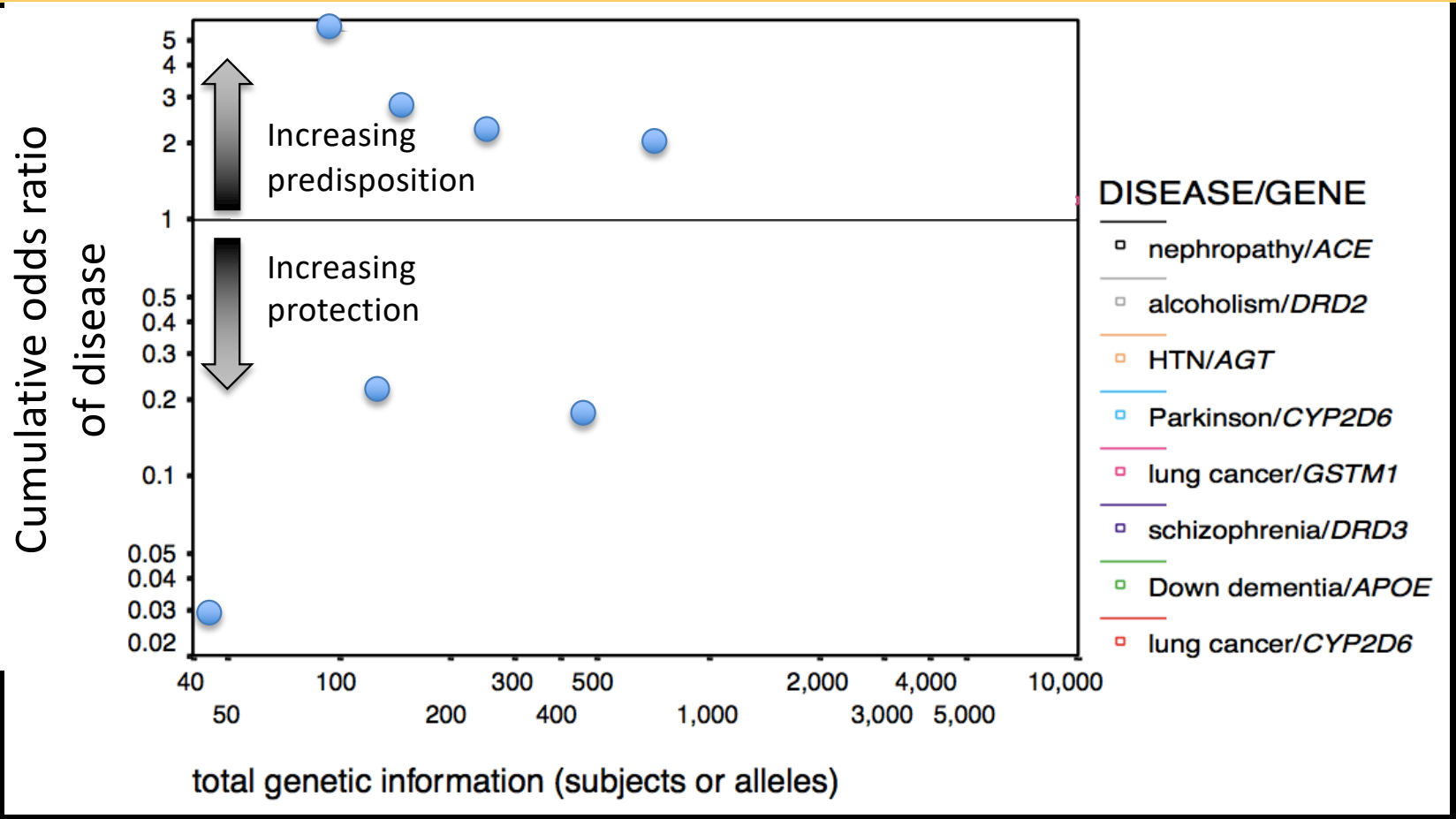
But ...
surely, this doesn't apply to genomics

or does it?

Outline

- Why replication failures are happening in genomics
- Why we are responsible for most of this
- Steps we can implement to overcome these problems

8 disease genes 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.

There are lies, damn lies, and ....

But wait, is that fair?

Are these really lies?

Where does this replication problem 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
  - Technology and methods move faster than rigorous error modeling

Where does this MOST bias come from?

YOU!!

And me …. All of us

Its arises from humans doing science

The way we think

The way our institutions work

Apophenia

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

Story telling of the false positives

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

Question your bioinformatics before falling in 

with your results

When results are better than you could have dreamed,

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

PNAS

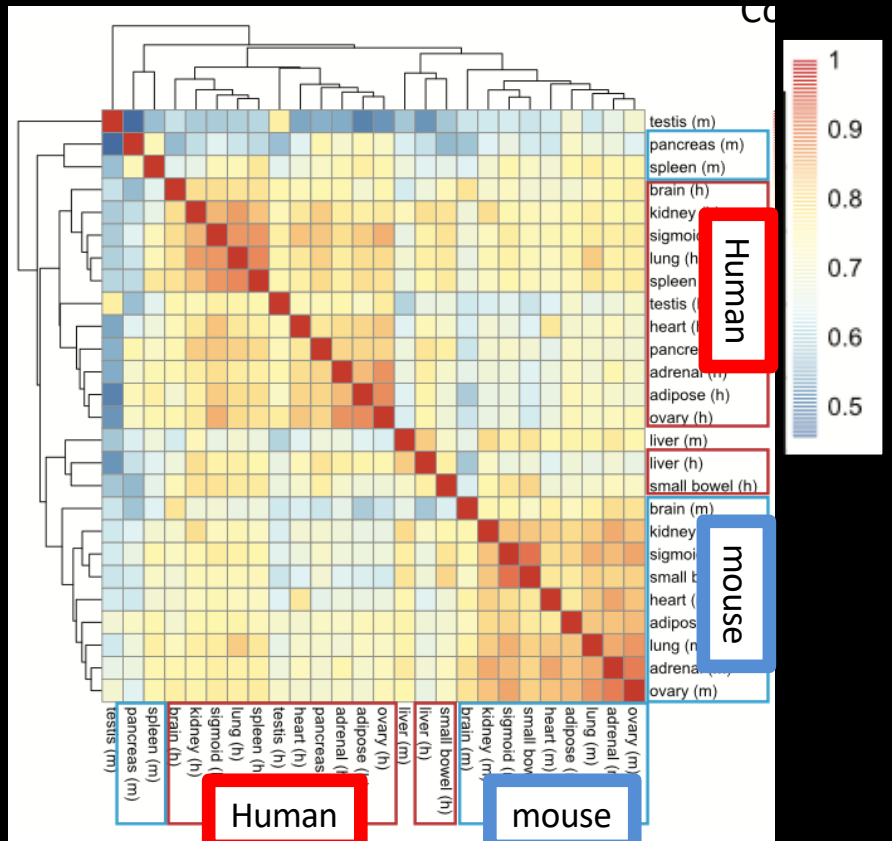
Comparison of the transcriptional landscapes between human and mouse tissues

"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

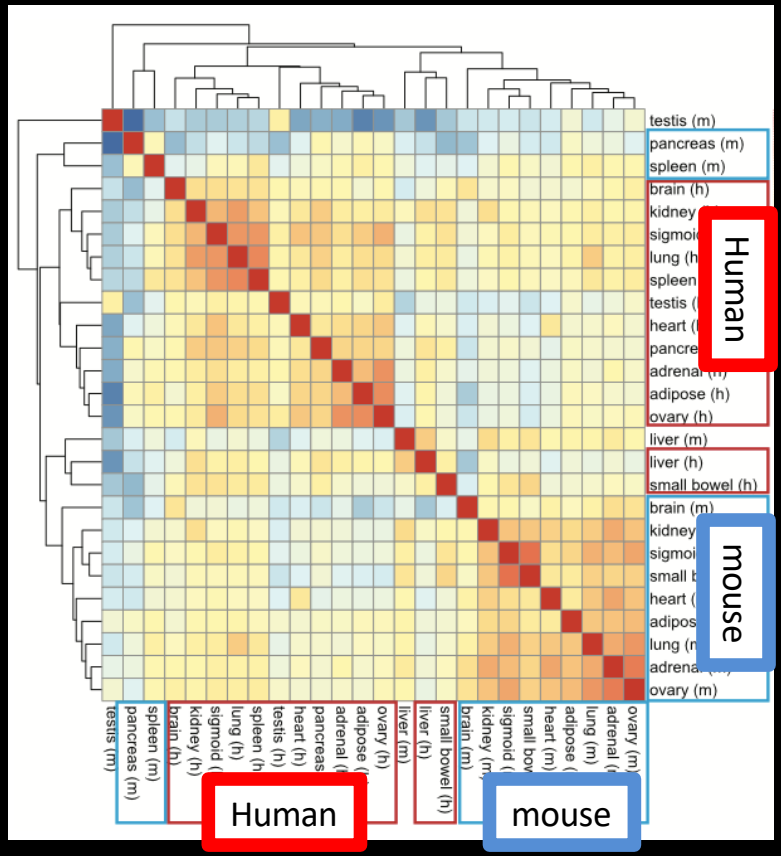
Authors found strong grouping of all organs by species, not by organ



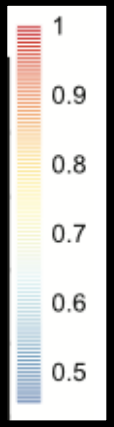
Should gene expression patterns group by species or tissues?

What do we expect from first principals, evolutionary relationships?

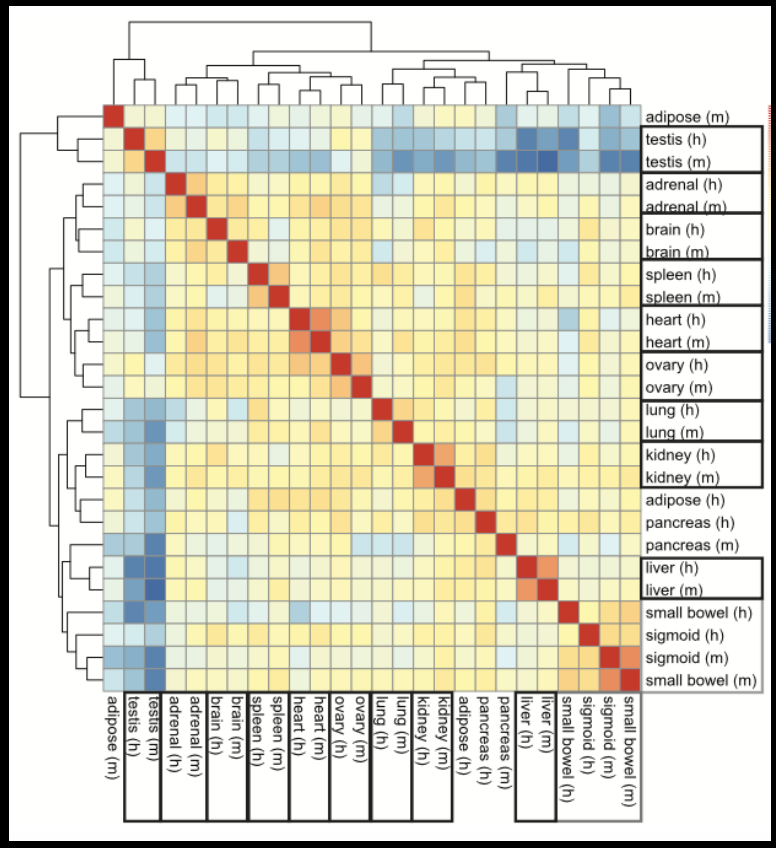
"the expression for many sets of genes was found to be more similar in different tissues within the same species than between species" Lin et al. 2014 PNAS



Correlation



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



Why? a batch effect 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

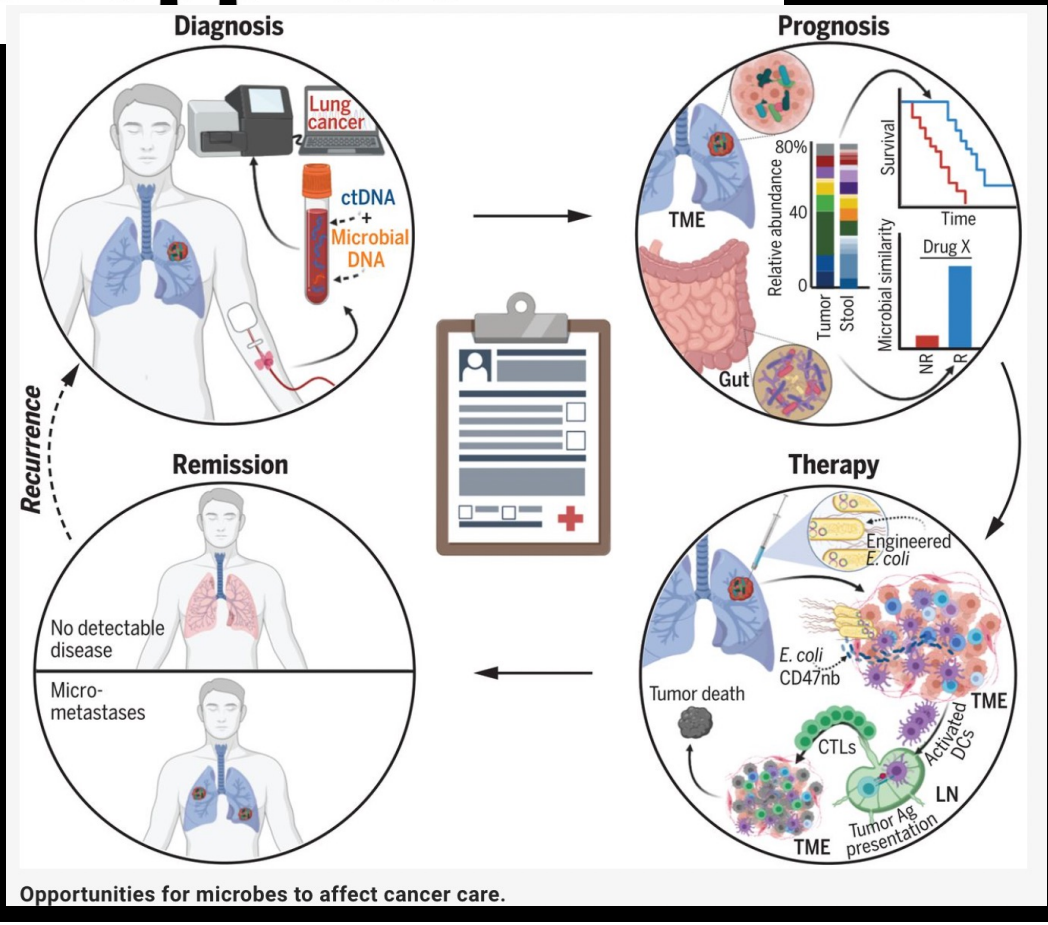
Process samples together, sequence all samples together

Article

Microbiome analyses of blood and tissues suggest cancer diagnostic approach

- strong association between microbial species and 33 different cancer types were based on a large collection of DNA and RNA sequencing samples taken from human cancers and from matched normal tissues
- processed by a sophisticated machine-learning method to create highly accurate classifiers that could distinguish among tumor types and could distinguish tumor from normal tissue

Poore et al. 2020; Spich-Poore et al. 2021; Gihawi et al. 2023 for text above

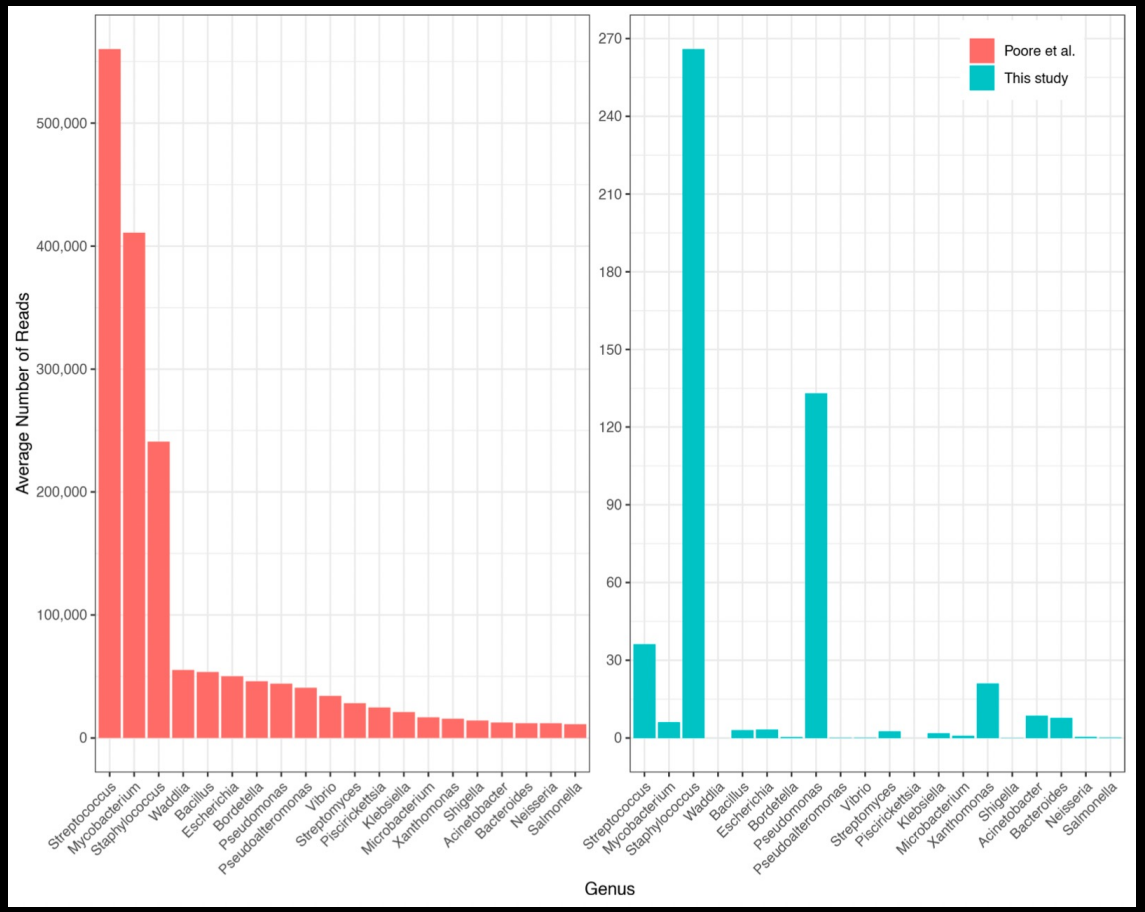


Opportunities for microbes to affect cancer care.

- led to a flurry of papers describing microbial signatures of different cancer types.
- Many of these reports are based on flawed data that, upon re-analysis, completely overturns the original findings.
- re-analysis shows that most of the microbes originally reported as associated with cancer were not present at all in the samples.
- The original report of a cancer microbiome and more than a dozen follow-up studies are, therefore, likely to be invalid.

Gihawi et al. 2023 for text above

- over-counts were due to human reads that erroneously matched bacteria
- A huge effect arising from omitting the human genome from the analysis database (Kraken)



Gihawi et al. 2023 for text above

- Published 11 March 2020, retracted on 26 June 2024
- 4 years is actually fast, due largely to the open access to data & methods
- This represents progress in the genomics field.

Frances Arnold
@francesarnold

For my first work-related tweet of 2020, I am totally bummed to announce that we
enzymatic synthe
reproducible. scie

Site-selective en
Enzymes excel at
sites. With approp
science.sciencem

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

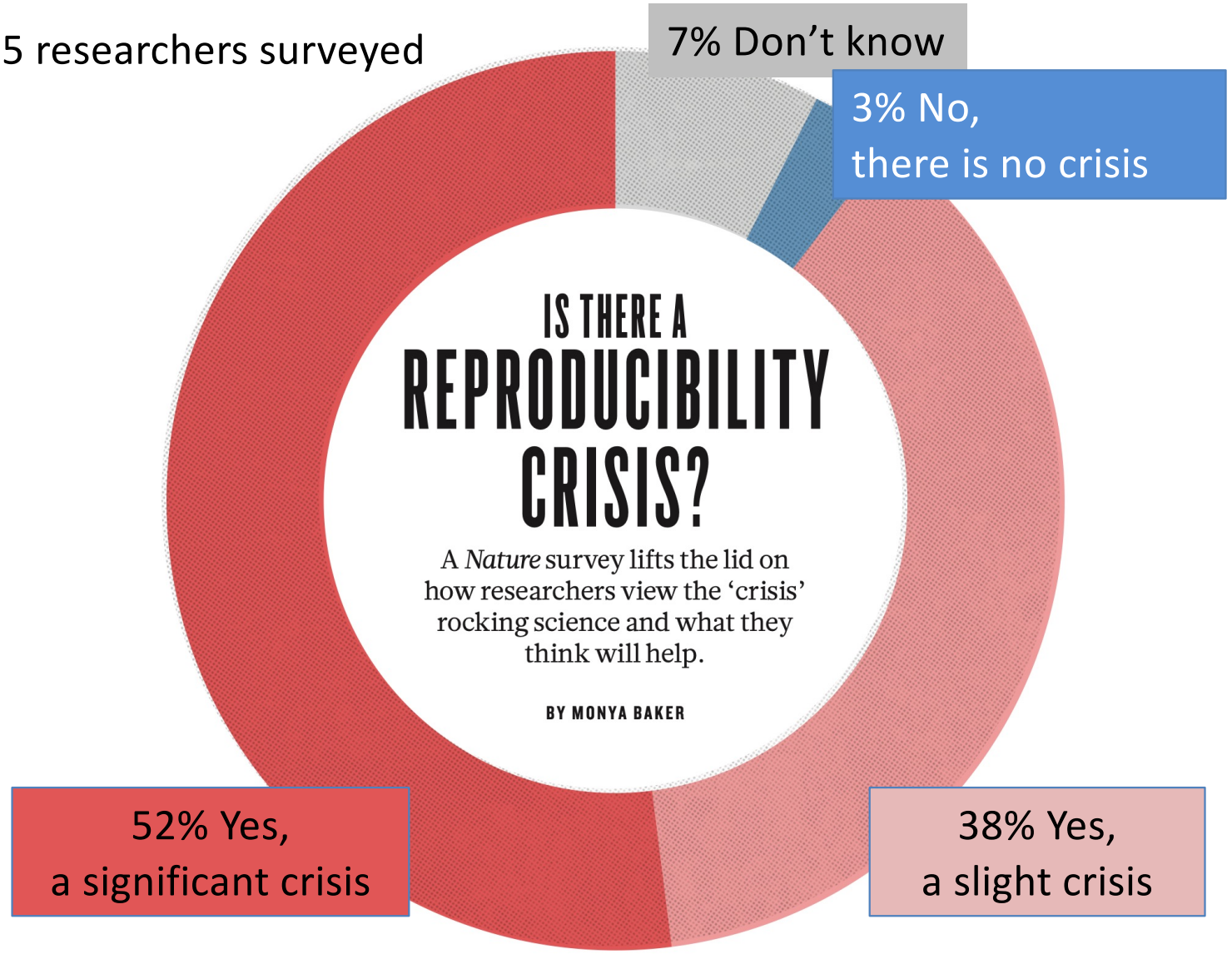
1 more reply

Lynn Kamerlin @kamerlinlab · Jan 2
Replying to @francesarnold
Sorry about the problems, but kudos for doing the right thing, and setting a good example.

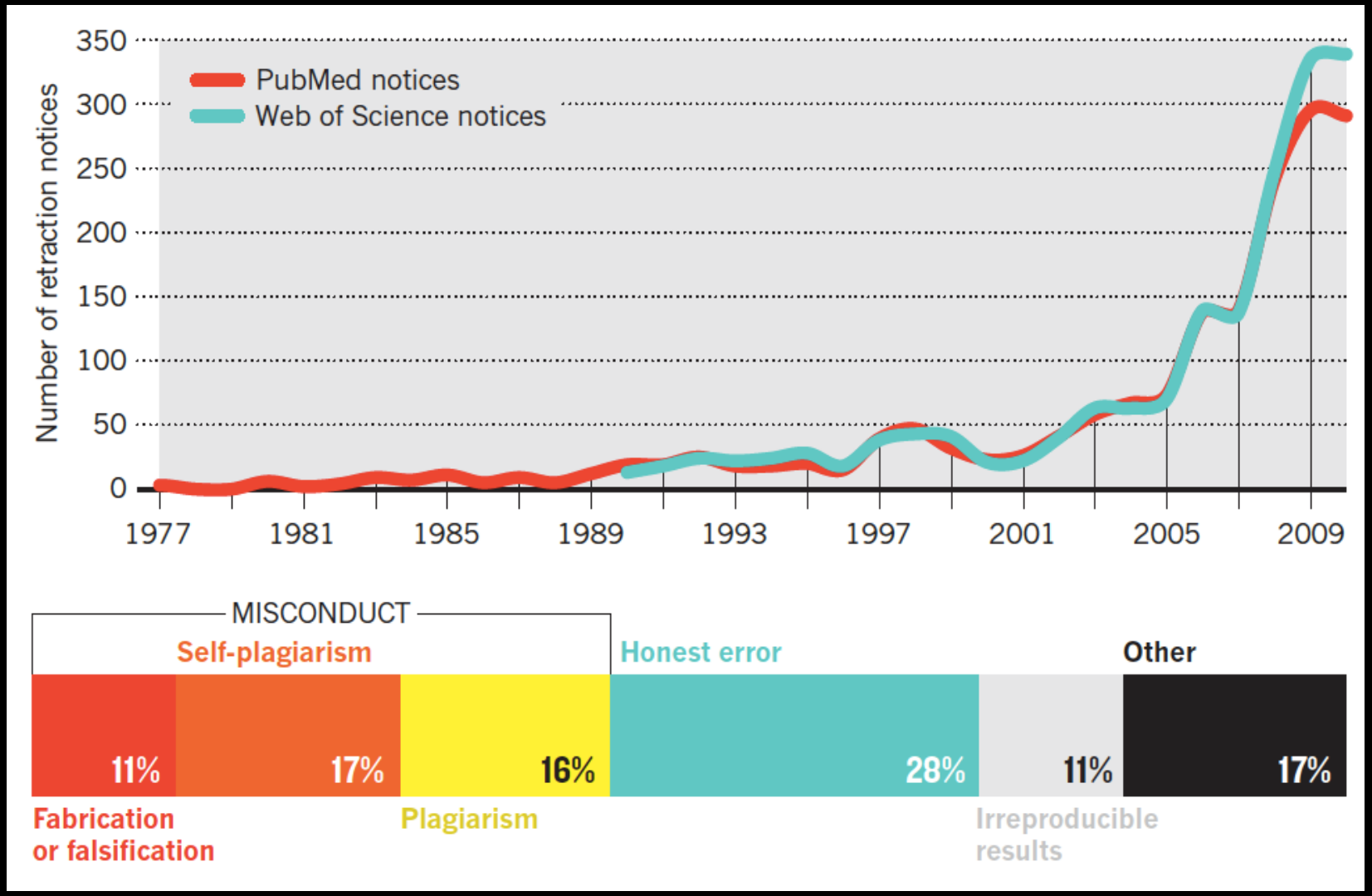
1 1 178

Waheed Ahmed @WaheedURAhmed1 · Jan 3
Honesty is so important and unfortunately, pretty underrated. Lots of respect and admiration for your actions.

1575 researchers surveyed



Baker 2016 Is there a reproducabiillty crisis?



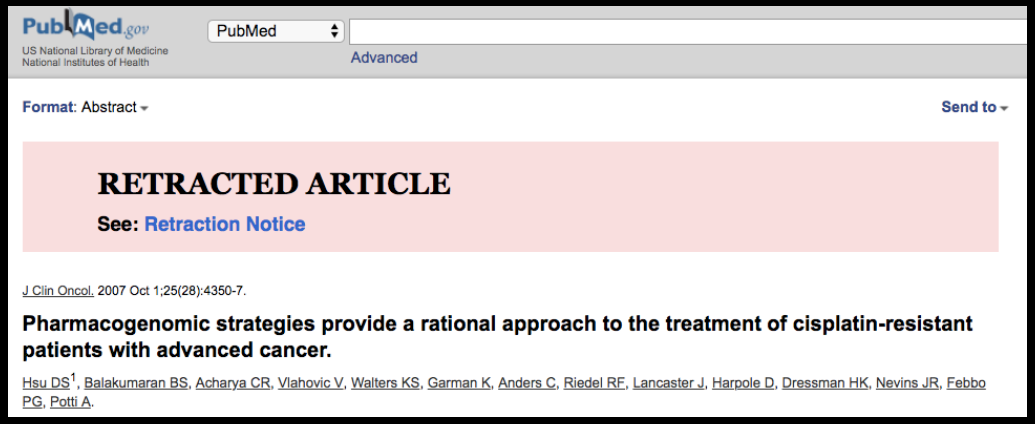
The trouble with retractions: Nature News 2011

Retraction Watch

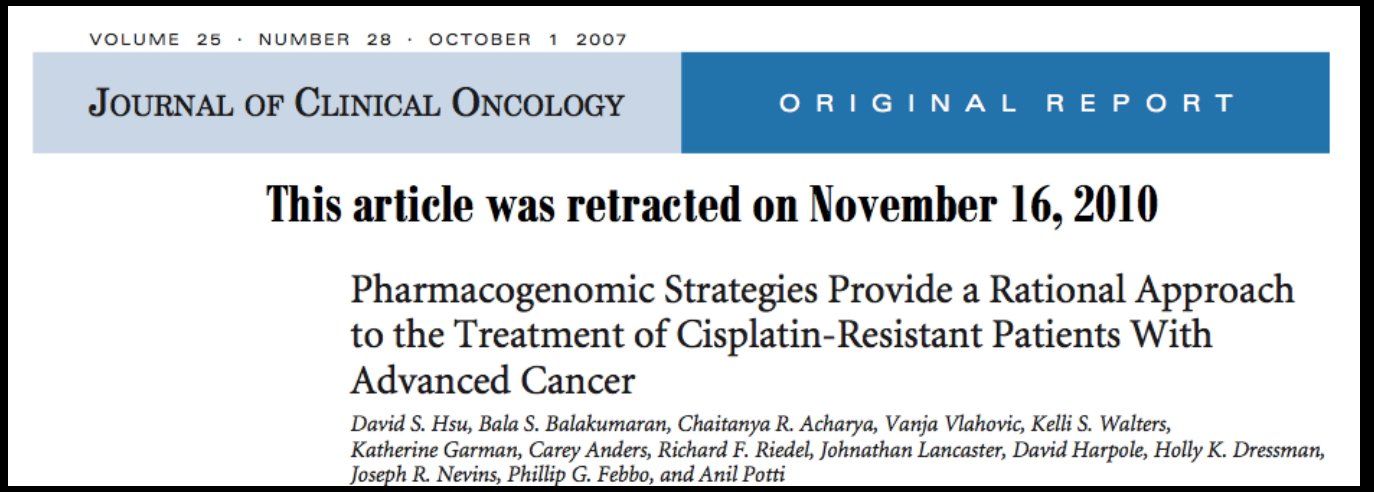
- Keeps community updated
- Help kill zombie papers that keep getting cited when they should not
- Starting to get integrated into websites and ref managers

- Be sure you are never keeping zombies alive

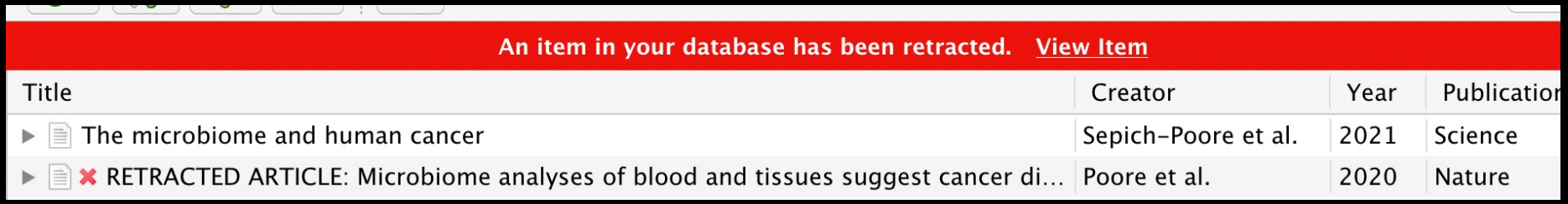
PubMed



Journal



Zotero

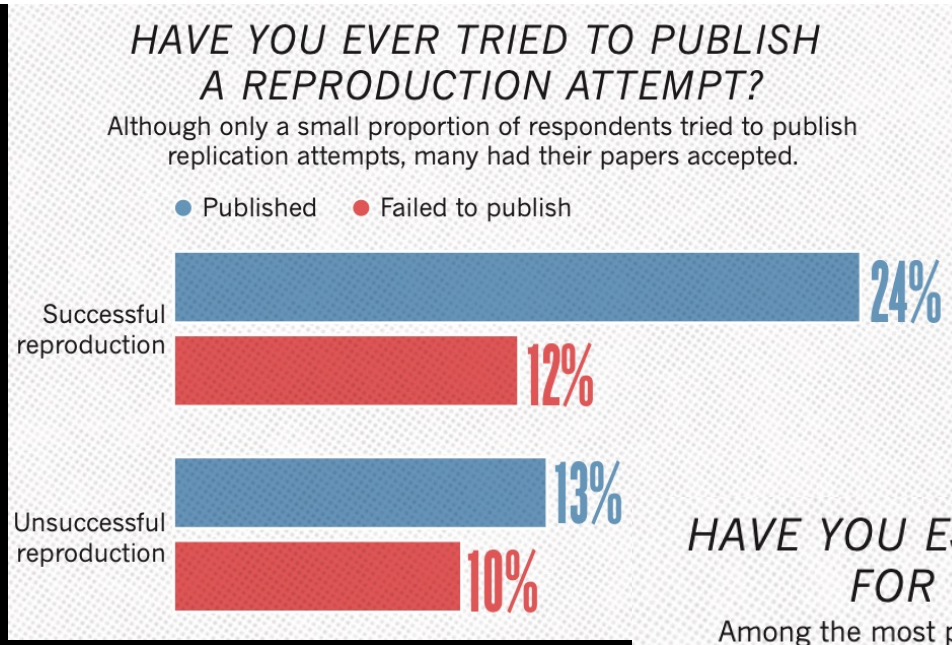


How can we improve reproducible findings?

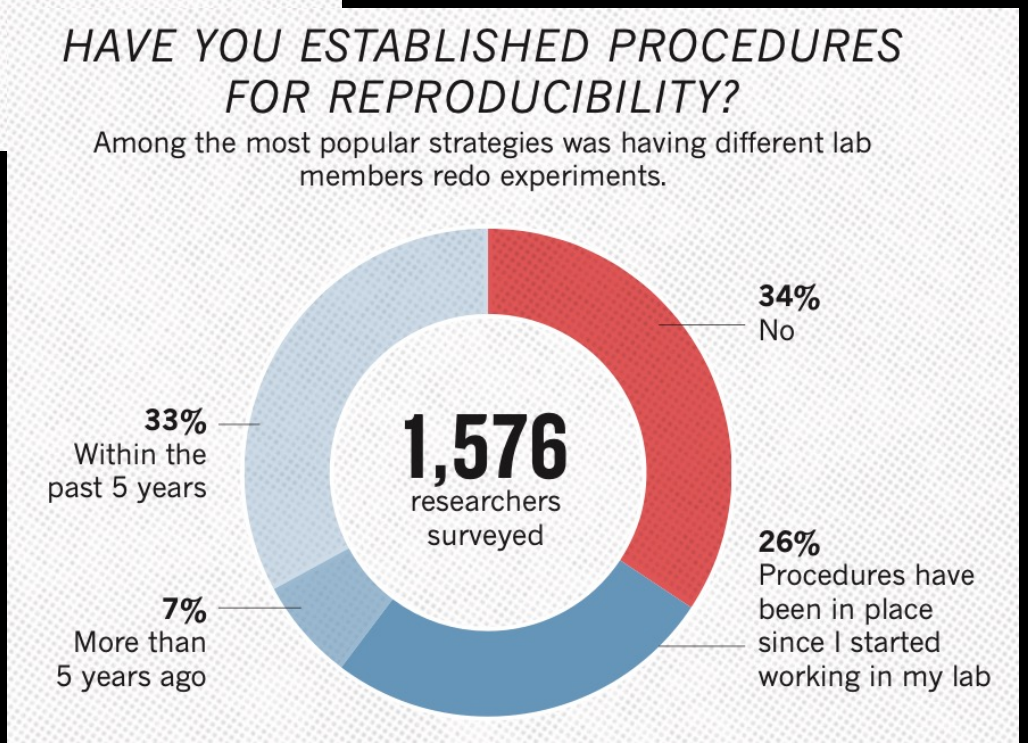
Work better as a community, check each others code and post our code

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



Though few tried to publish replications, many had papers accepted!!



Most popular strategy for replication was having different lab members redo work

Baker 2016 Is there a reproducibility crisis?

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

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

"You have to validate what you create"
Erik Garrison

What other biases might we suffer from?

https://www.babyanimalprints.com/collections/monkeys-and-apes-black-and-white/chimpanzee

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 .....

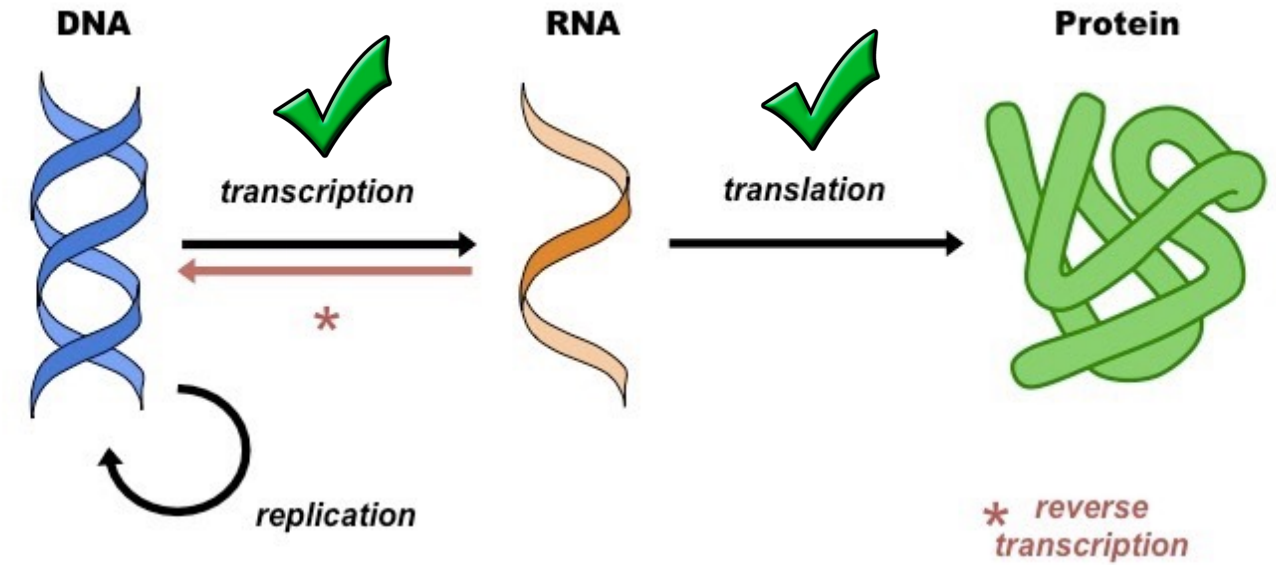
Hindsight bias

the knew-it-all-along effect

Three Levels of Hindsight Bias

I KNEW
that would happen

The central dogma

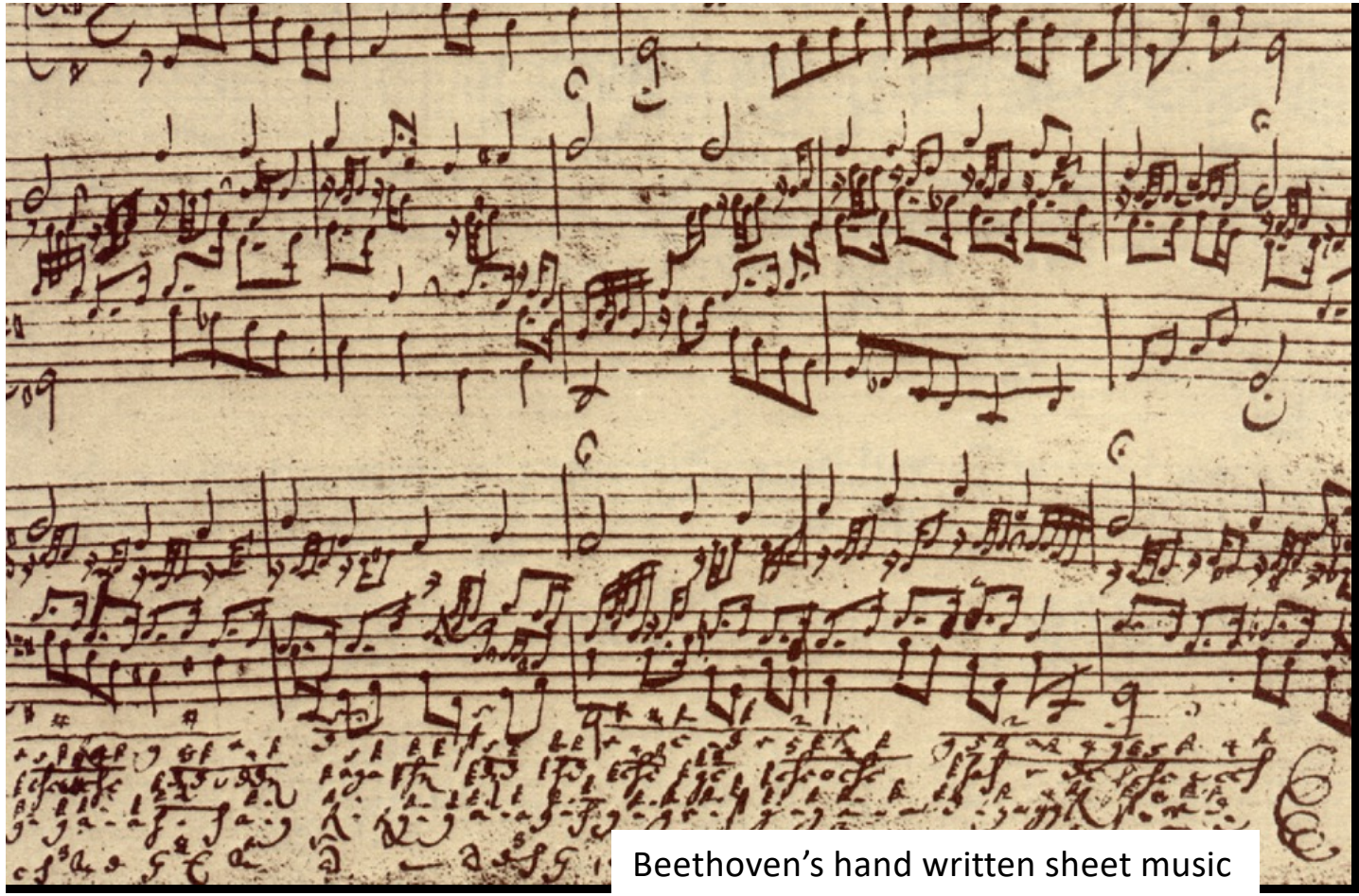


<u>But, can we, in a novel species :</u>

- 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?

Going from peptide sequence to catalytic function ...

"We don't know how to write that way"



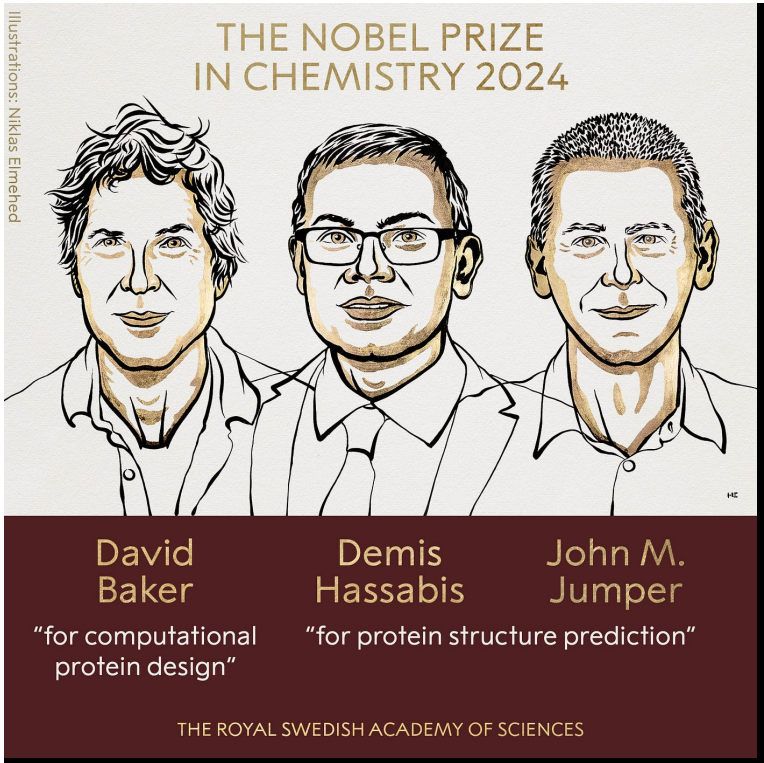
Beethoven's hand written sheet music

Quote in Nobel Prize lecture, 2018

https://youtu.be/6hOZ5e0g9Uo

Francis Arnold

Nobel Prize winner (2018)



inventors of Alphafold were awarded the Nobel Prize for developing an AI model to solve a 50-year-old problem: predicting proteins' complex structures

nature

Article | Open access | Published: 08 May 2024

Accurate structure prediction of biomolecular interactions with AlphaFold 3

Can model protein protein interactions, along with other molecules

Did AI Solve the Protein-Folding Problem?

Open question is whether AlphaFold has actually discovered something meaningful about the physics of protein folding that humans haven't

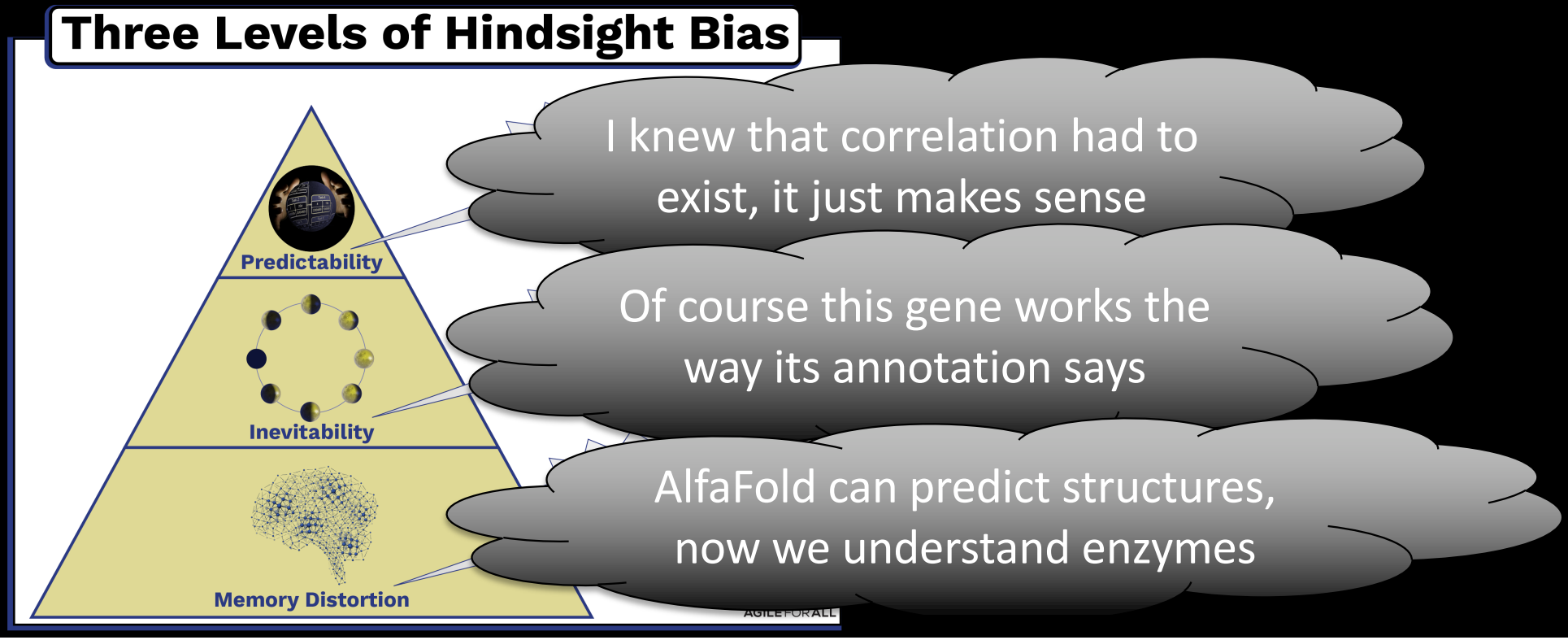
"If we can predict how proteins fold without understanding how they do it, are we even legitimately doing science anymore, or is it something different?"

"We're able to get the practical benefits, but we're not necessarily gaining intellectual benefits"

https://magazine.hms.harvard.edu/articles/did-ai-solve-protein-folding-problem

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

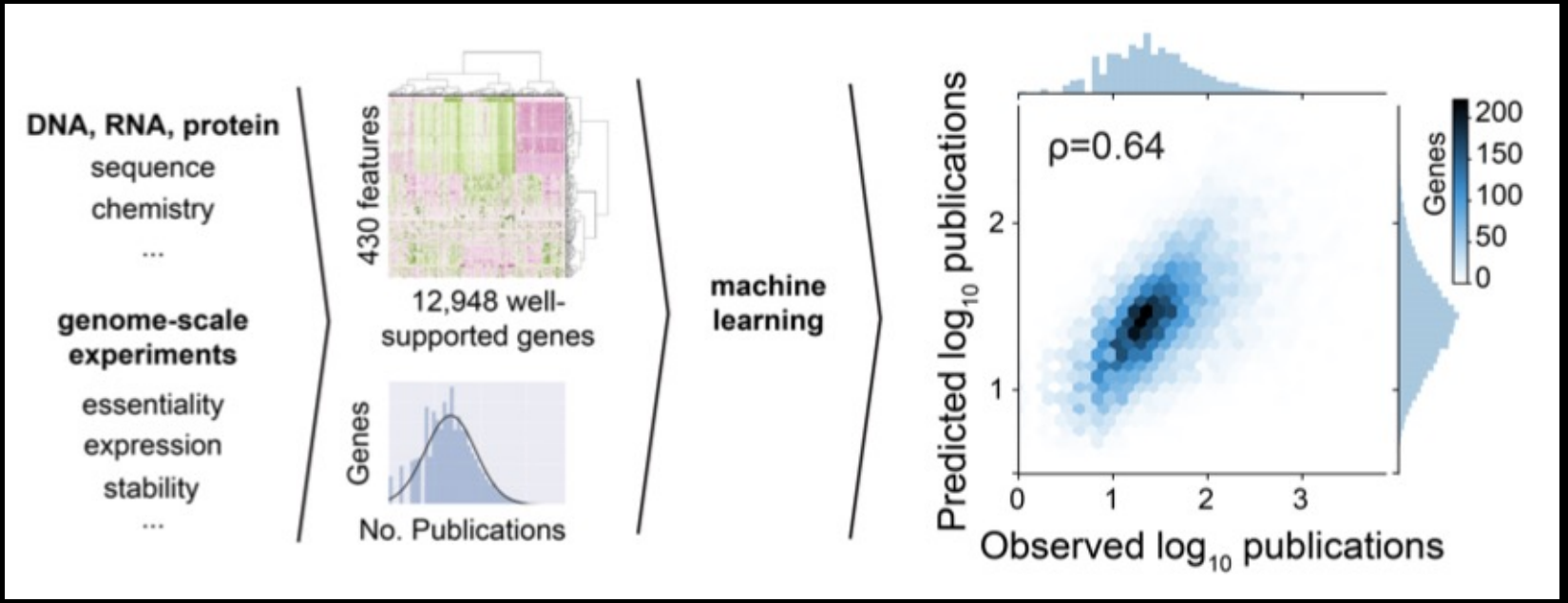
... but biology is exceptionally complex



What about the genes we study?

Do we ever conduct "unbiased" investigations?

What if we looked at investigations by gene, over time



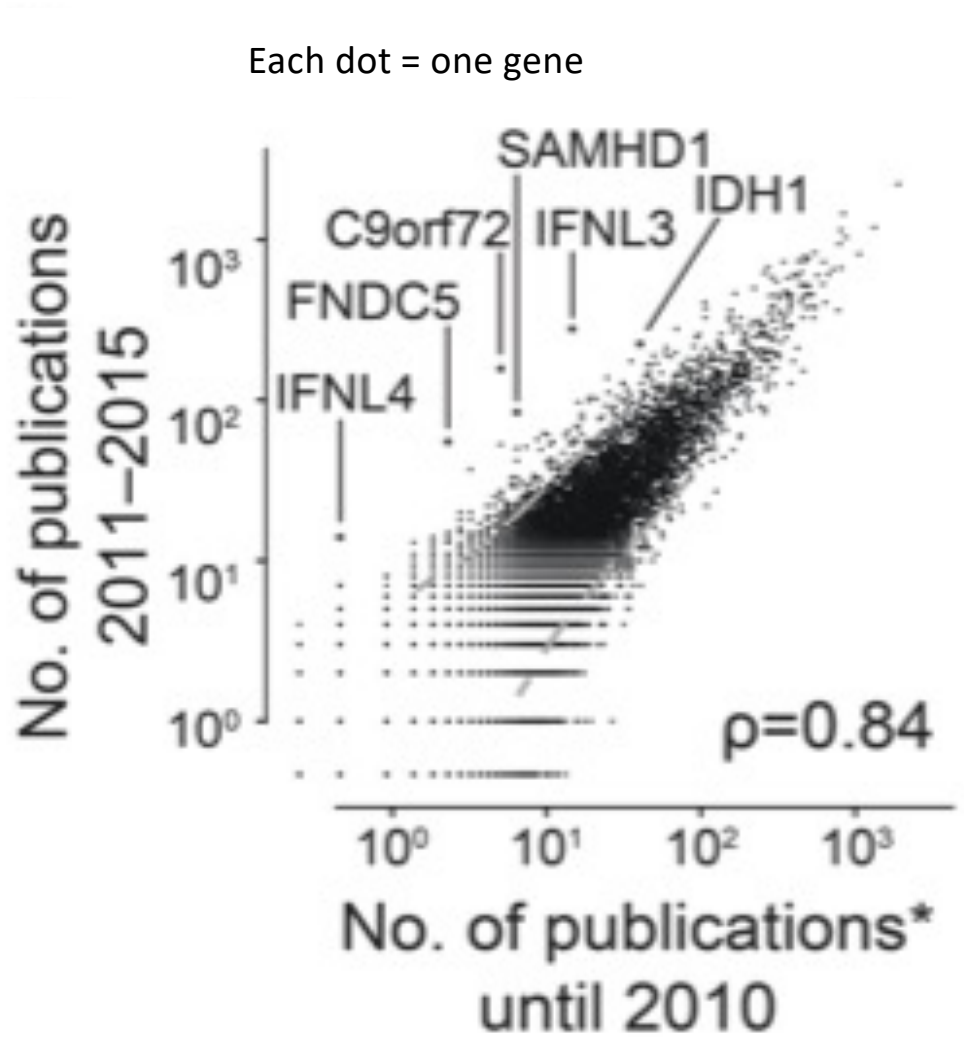
Stoeger et al. 2018 Plos Biology

Historical precedence drives what genes get detailed study

30 % of all genes have never been the focus of a scientific study

< 10 % of genes are the subject of > 90 % of published papers

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



Stoeger et al. 2018 Plos Biology


sallyedelsteincollage.com