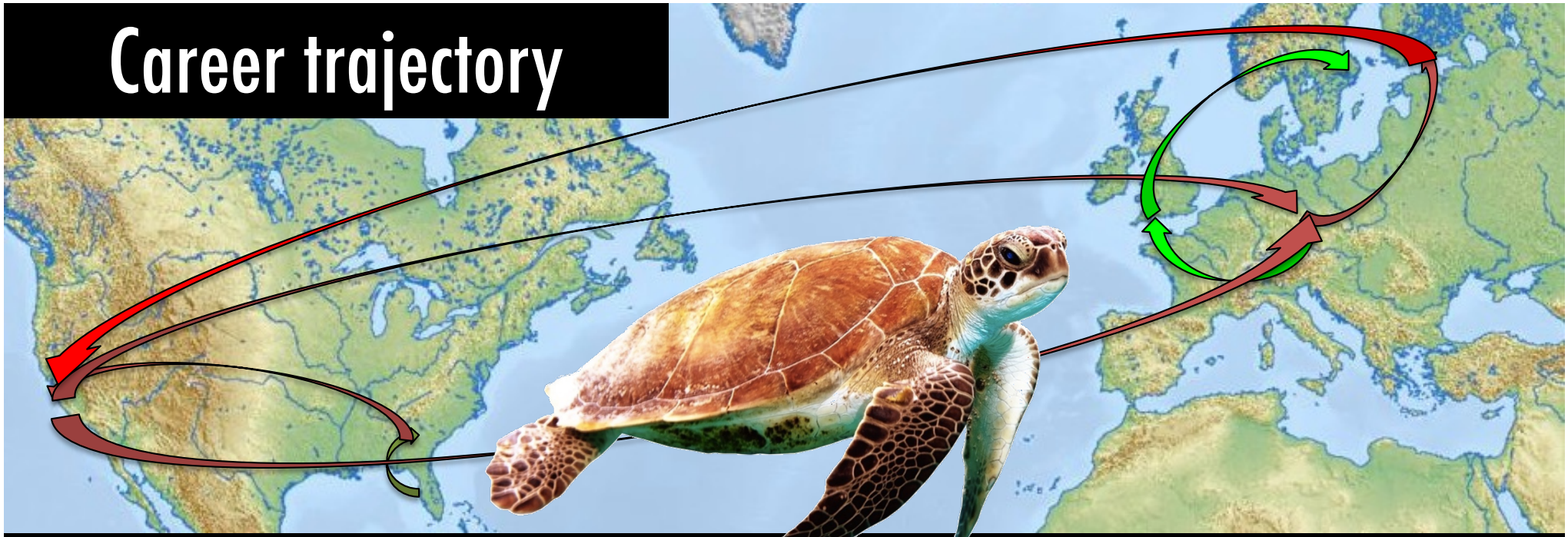
Lies, damn lies, and .... genomics

Navigating your data, your perceptions and reality

Christopher West Wheat

Career trajectory



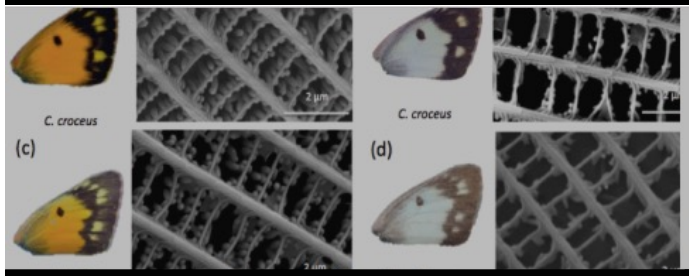
- 1995 – 2001 PhD California
- 2002 – 2005 Postdoc Germany
- 2005 – 2008 Postdoc Finland
- 2009 – unemployed 4 month, spent all savings
  - > 50 job applications, 1 grant application
- 2009 – visiting scientist Germany
  - 1 job offer UK
  - 1 grant Finland
- 2012 –Started at Stockholm University
- 2022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

What was important?

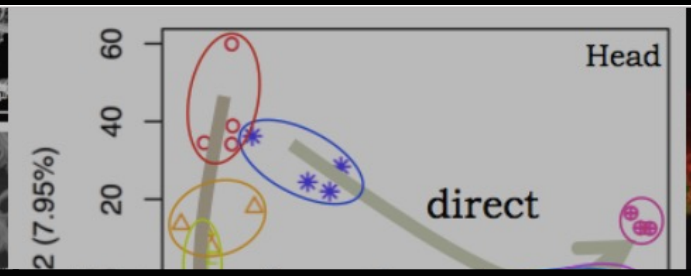
- Being able to move, chase the money & get new skills
- Learning how toBelieve in my ideas/skills

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

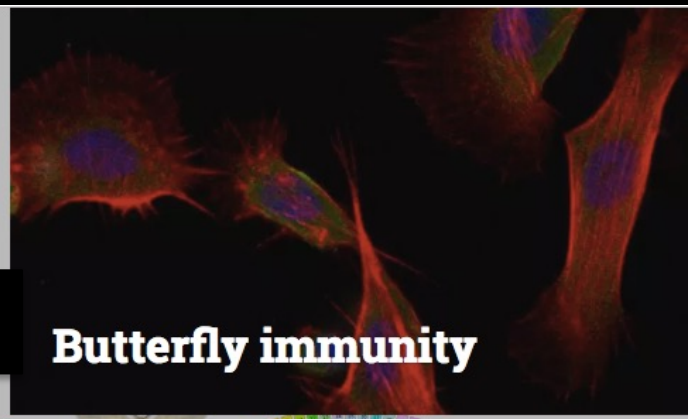
Ecological & Evolutionary Functional Genomics



Alternative life history switches

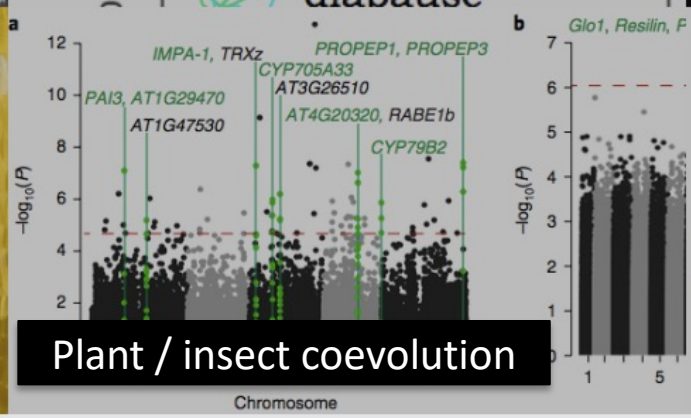


Circadian and seasonal clock

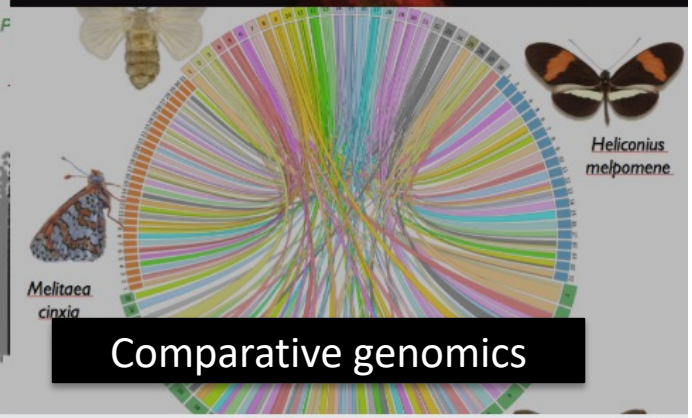


Butterfly immunity

CRISPR-Cas9



Plant / insect coevolution



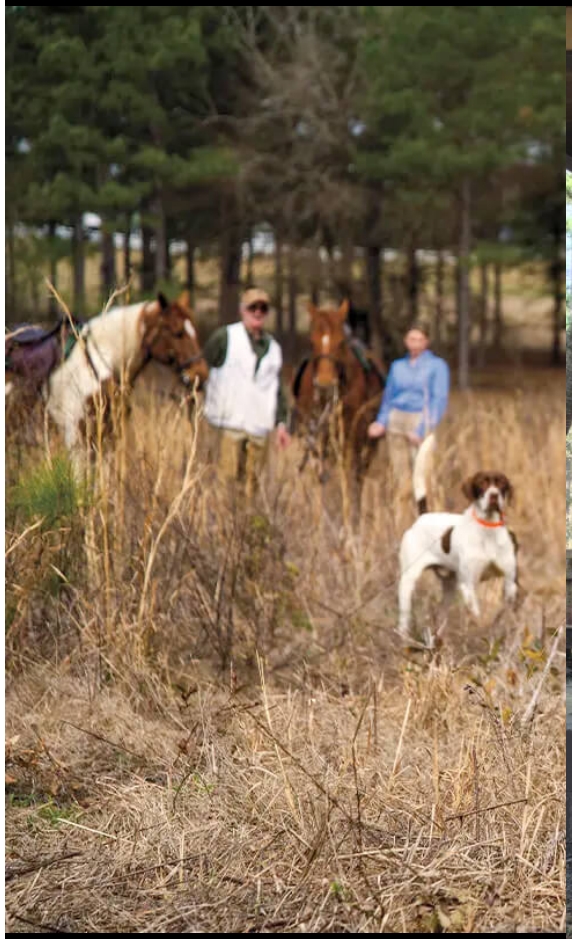
Comparative genomics

CHRISTOPHER WHEAT LAB

https://christopherwheatlab.net/

Something you don't know about me

I am a Judge of Field Trials, in the American Field Trial Clubs of America for over 20 years



Goal of this lecture

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

Disclaimer

I'm a positive person

I love my job and the work we all do

I'm just sharing scrumptious food for thought

What if .....

50% of your favorite studies were not repeatable?

Would that impact your science?

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 A G G G	Repl.	Fixed
867	C	- - - - - - - - - - - -	- - - - - -	G G G G A G 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 - 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.
1431	T	C C - - - - - - - - - -	- - - - - -	C C C C C C C C C C C C	Syn.	Poly.
1443	C	- - - - - - - G G G G G	- - - - - -	- - - - - - - - - - - -	Syn.	Poly.
1452	C	- - - - - - T T T T T T	- - - - - -	- - - - - - - - - - - -	Syn.	Poly.
1490	A	- - - - - - - C C C C C	- - - - - -	- - - - - - - - - - - -	Repl.	Poly.
1504	C	T T T T T T T T T T T T	- - - - - -	- - - - - - - - - - - -	Syn.	Fixed
1518	C	- - - - - - T T T T T T	- - - - - -	- - - - - - - - - - - -	Syn.	Poly.
1524	T	- - - - - - - - - - - -	- - - - - -	G G G G G G G G G G G G	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

Nature 1991

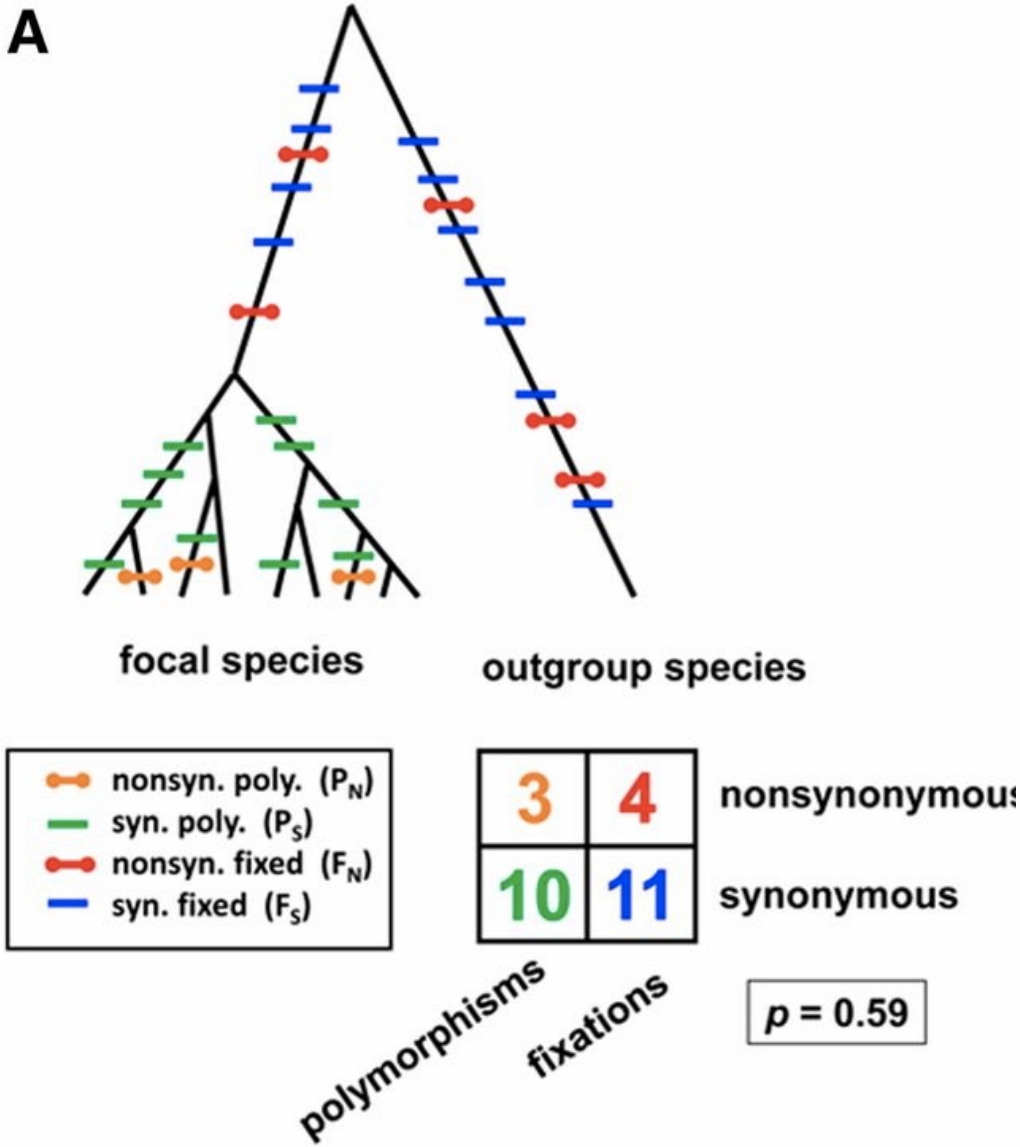
[redacted]

	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

[redacted]

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

McDonald Kreitman test



Lazzaro 2018 Genetics

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 A 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
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 - 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.
1431	T	C C - - - - - - - - - -	- - - - - -	C C C C C C C C C C C C	Syn.	Poly.
1443	C	- - - - - - - G G G G G G	- - - - - -	- - - - - - - - - - - -	Syn.	Poly.
1452	C	- - - - - - T T T T T T T	- - - - - -	- - - - - - - - - - - -	Syn.	Poly.
1490	A	- - - - - - - C C C C C C	- - - - - -	- - - - - - - - - - - -	Repl.	Poly.
1504	C	T T T T T T T T T T T T	- - - - - -	- - - - - - - - - - - -	Syn.	Fixed
1518	C	- - - - - - T T T T T T T	- - - - - -	- - - - - - - - - - - -	Syn.	Poly.
1524	T	- - - - - - - - - - - -	- - - - - -	G G G G G G G G G G G G	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

Nature 1991

[redacted]

	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						Fixed
808						Fixed
816						Poly.
834						Poly.
859						Fixed
867						2 Poly.
870						Fixed

[redacted]

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

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

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

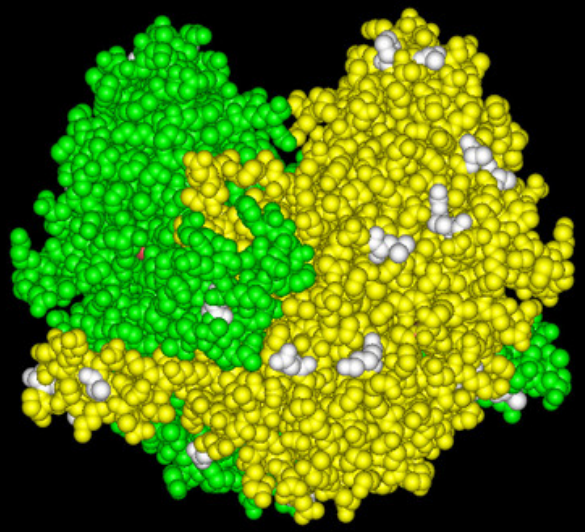
John H. McDonald & Martin Kreitman

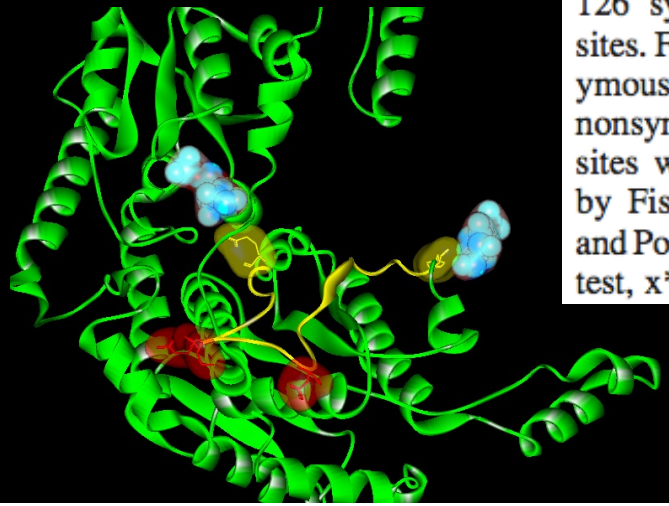
Department of Ecology and Evolutionary Biology, Princeton University, Princeton, New Jersey 08544, USA

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

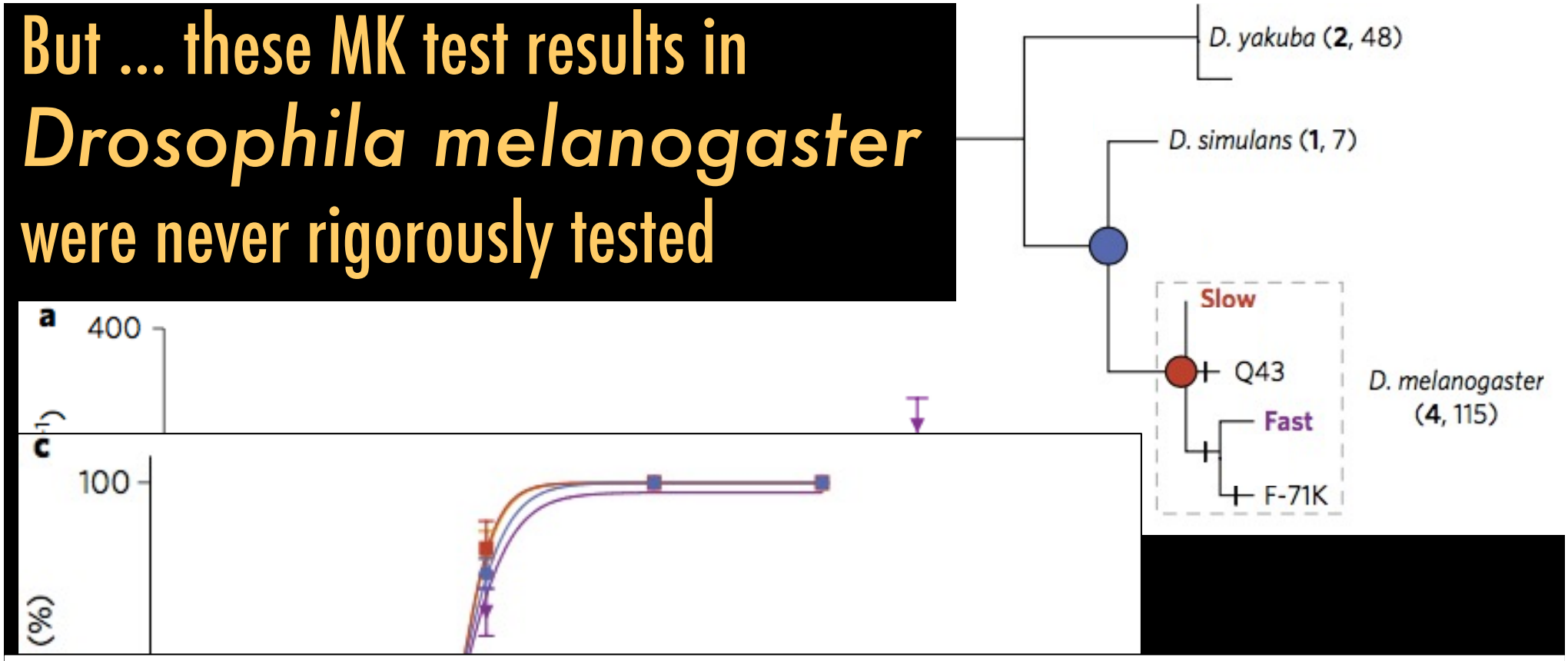




Among *C. eurytheme* and *C. meadii* PGI sequences, we find 126 synonymous and 20 nonsynonymous polymorphic sites. From their ratio, 6.3:1, neutrality predicts ~13 synonymous fixations alongside the two observed interspecies nonsynonymous fixations. But, *no* fixed synonymous sites were found (above).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

Wheat et al. 2005

But ... these MK test results in *Drosophila melanogaster* were never rigorously tested



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.....

Does this happen only in bugs?

My PhD was chasing results based upon an weak framework?

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

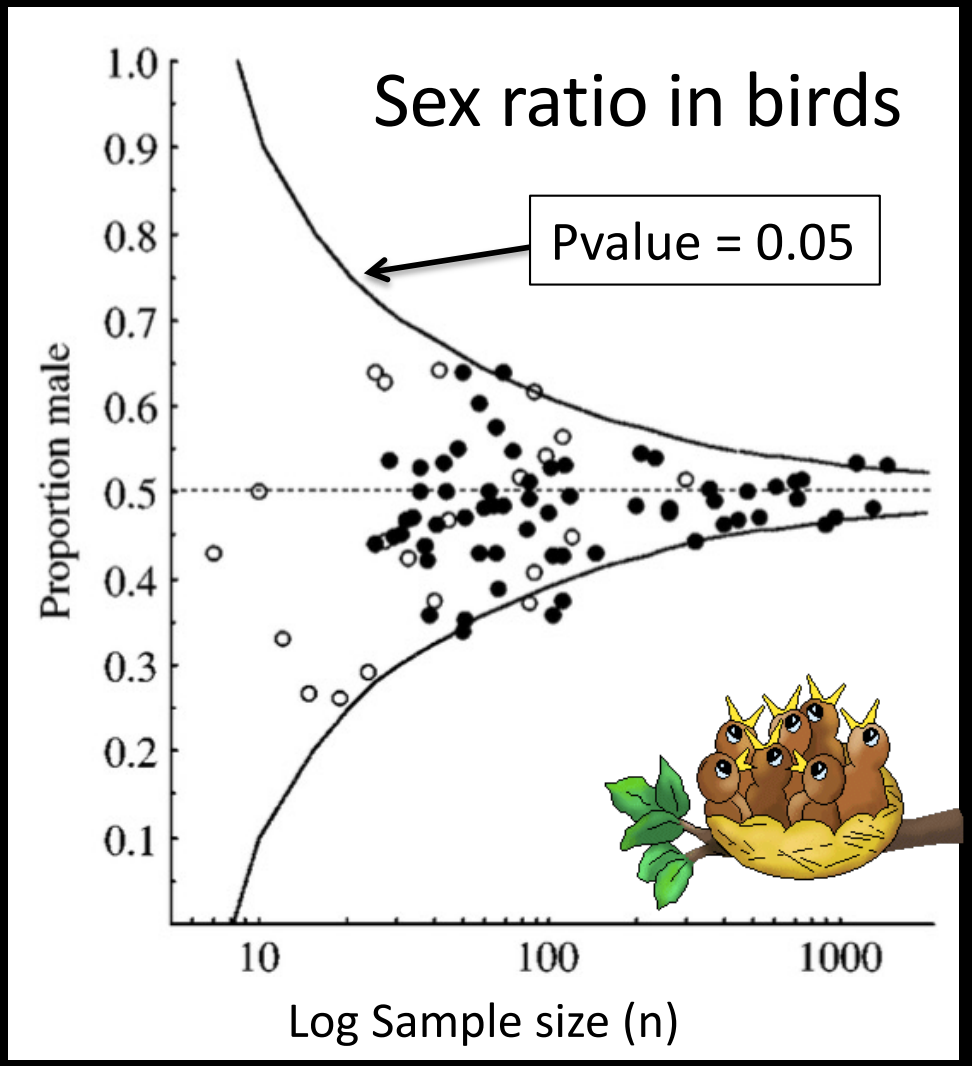
Publication replication failures

- Biomedical studies
  - Of 49 most cited clincal 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

Assessing reality using funnel plots



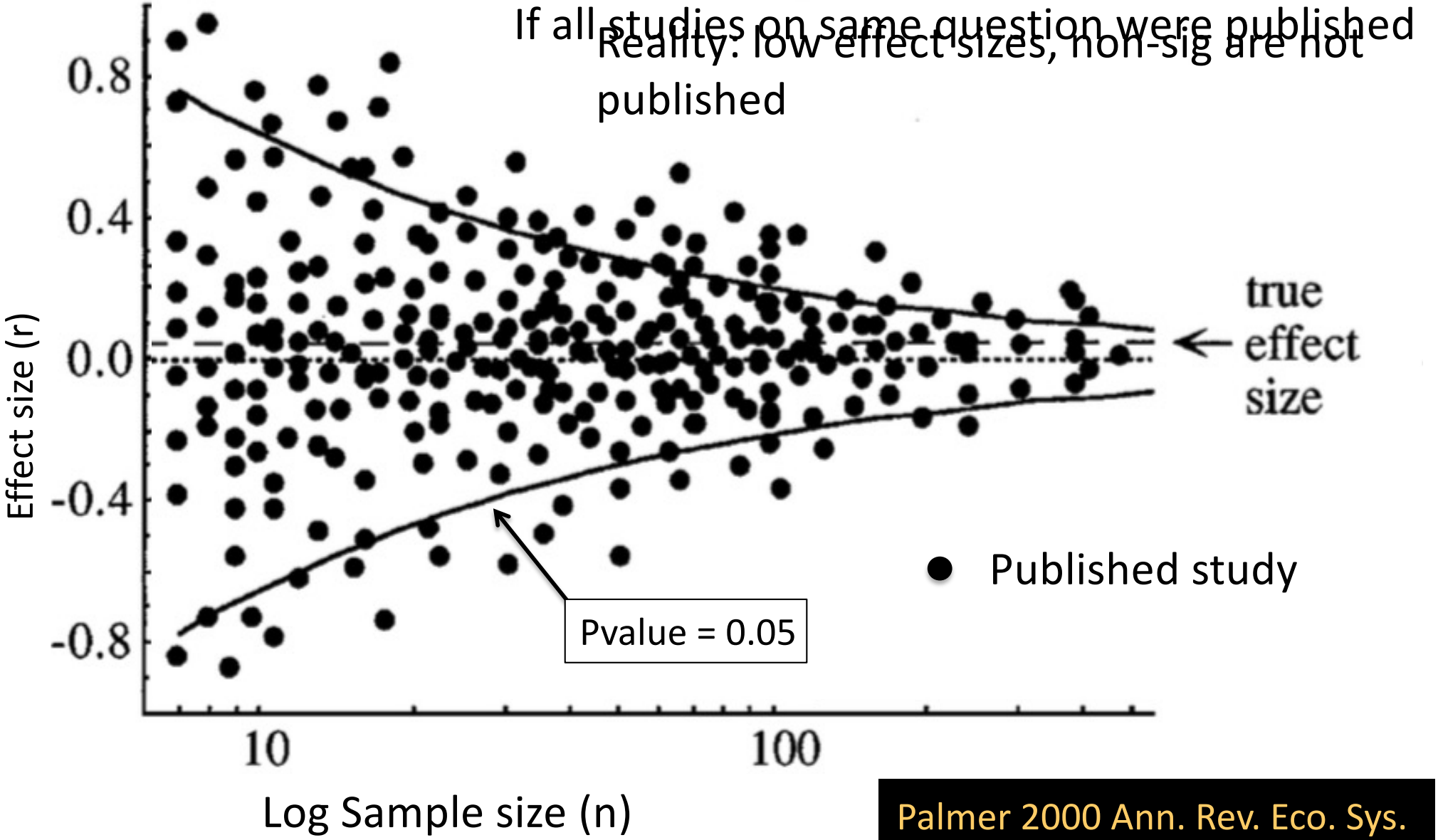
Small sample sizes affect measurement accuracy

Each dot = a study and has error

Study estimates are randomly distributed about the real value

Your study is just a random estimate of some idealized value

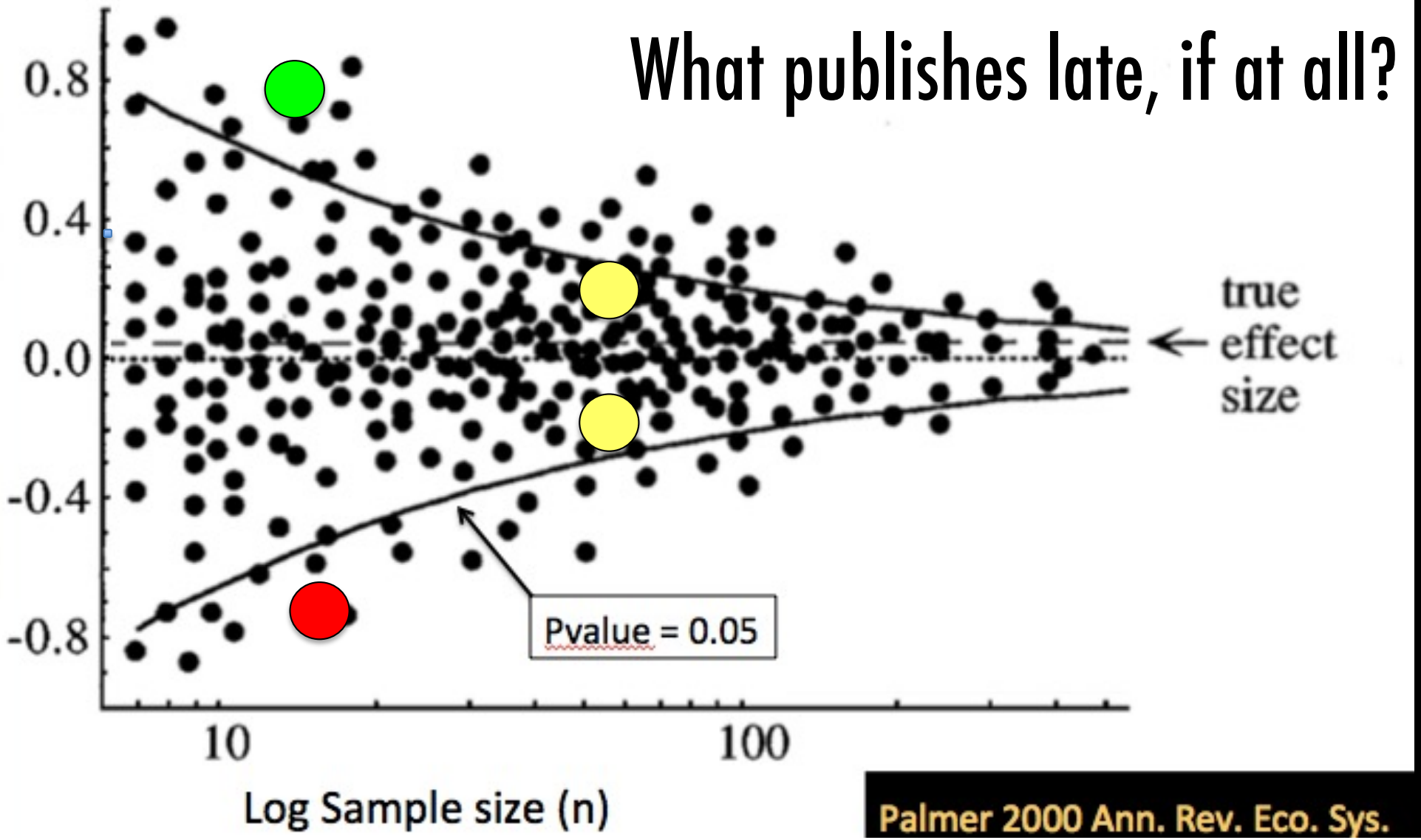
Publication bias increases effect size



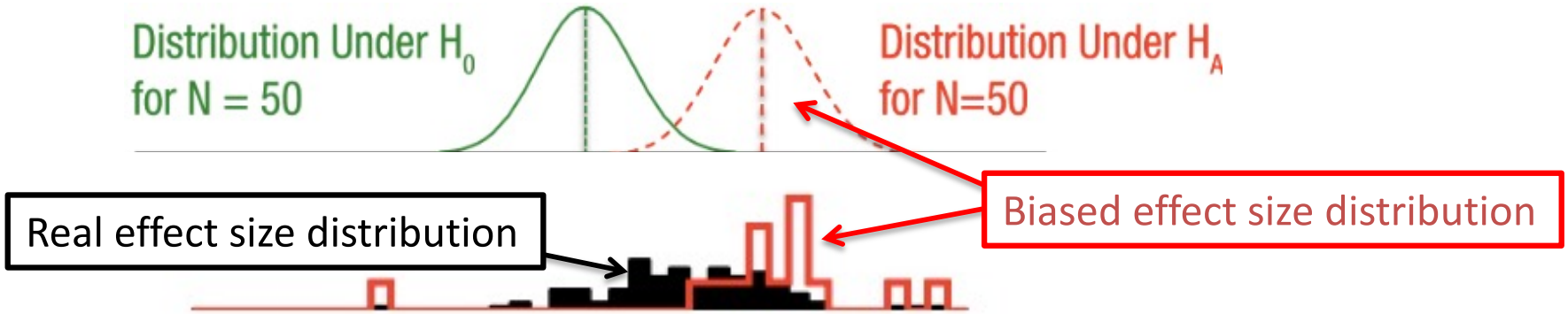
Palmer 2000 Ann. Rev. Eco. Sys.

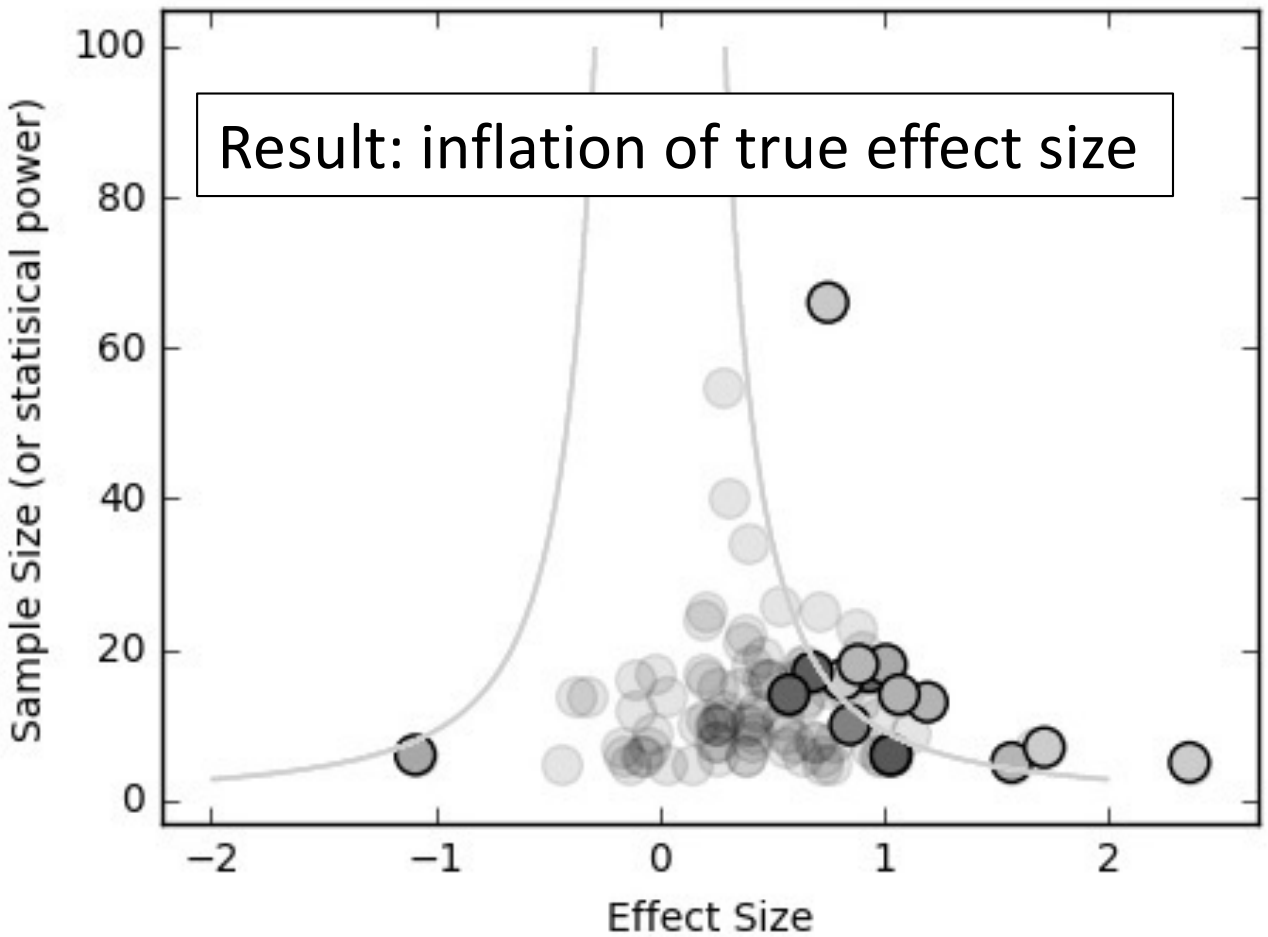
What if there is no replication?

What is most likely to publish first & where?



Palmer 2000 Ann. Rev. Eco. Sys.





Simulation of:

- studies with a low N, 0.5 effect size
- a bias in publishing sig. results (colored)
- a bias against being able to publish null results

https://bids.berkeley.edu/news/visualizing-publication-bias-case-funnel-plots

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.

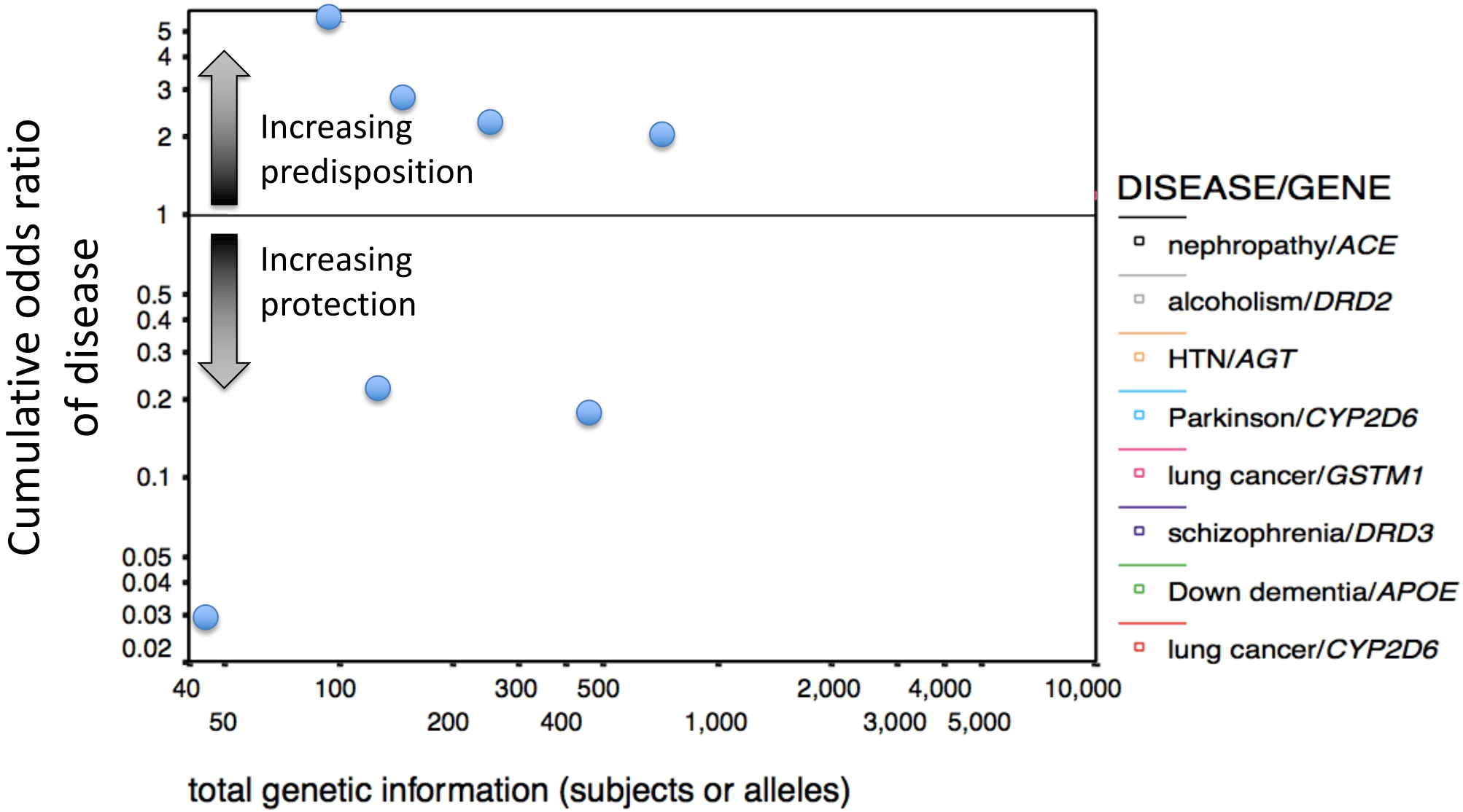
But surely, this doesn't apply to genomics ....

Or does it?

Outline

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

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.

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

But wait, is that fair?

Are these really lies?

Where does this bias come from?

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

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

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

Thus, always question your bioinformatics before falling in love 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

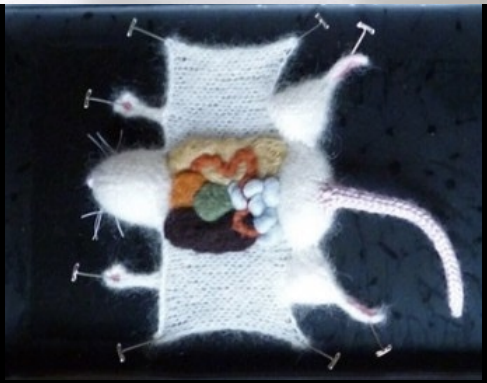
More genes underwent positive selection in chimpanzee evolution than in human evolution

Comparison of the transcriptional landscapes between human and mouse tissues

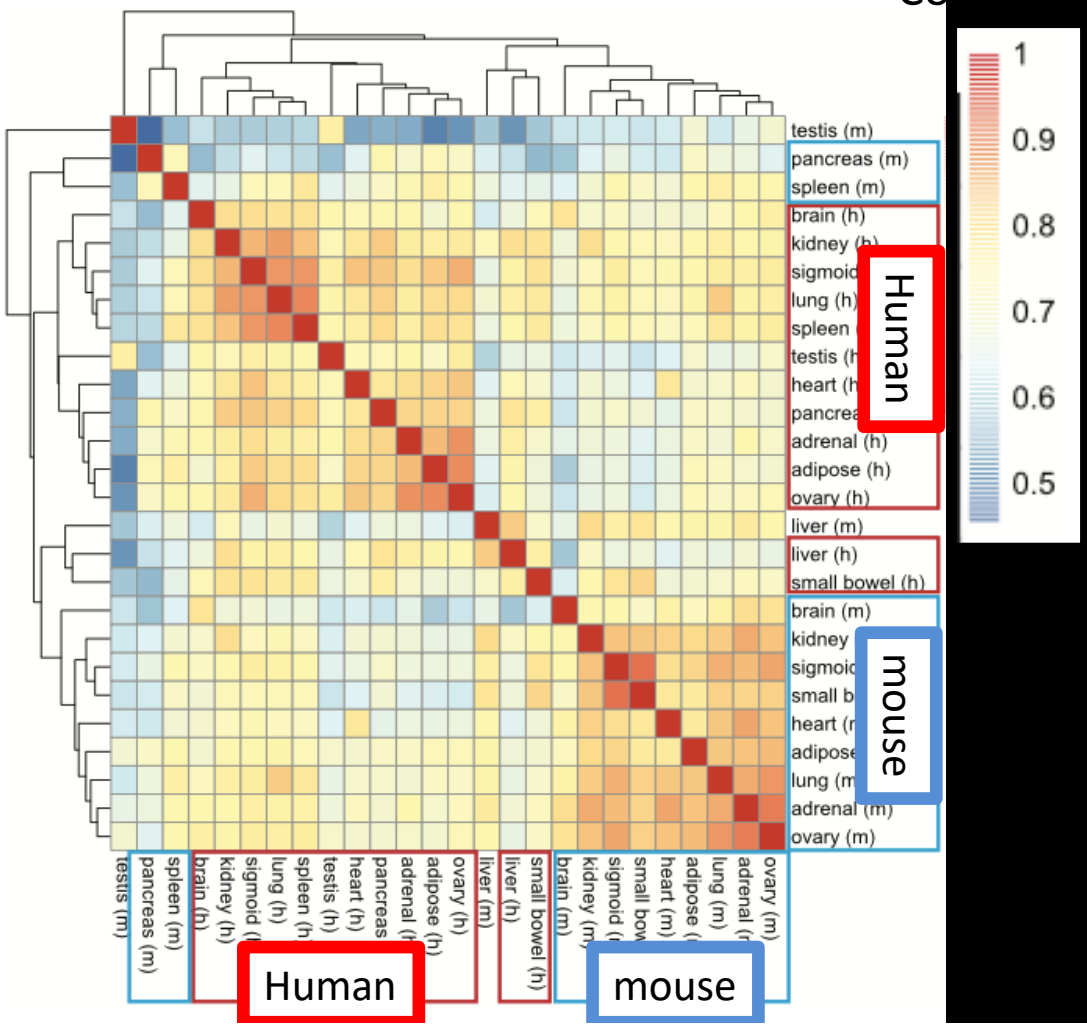
"the expression for many sets of genes was found to be more similar in different tissues within the same species than between species"

Time of the most recent common ancestor:

Human and Mouse



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

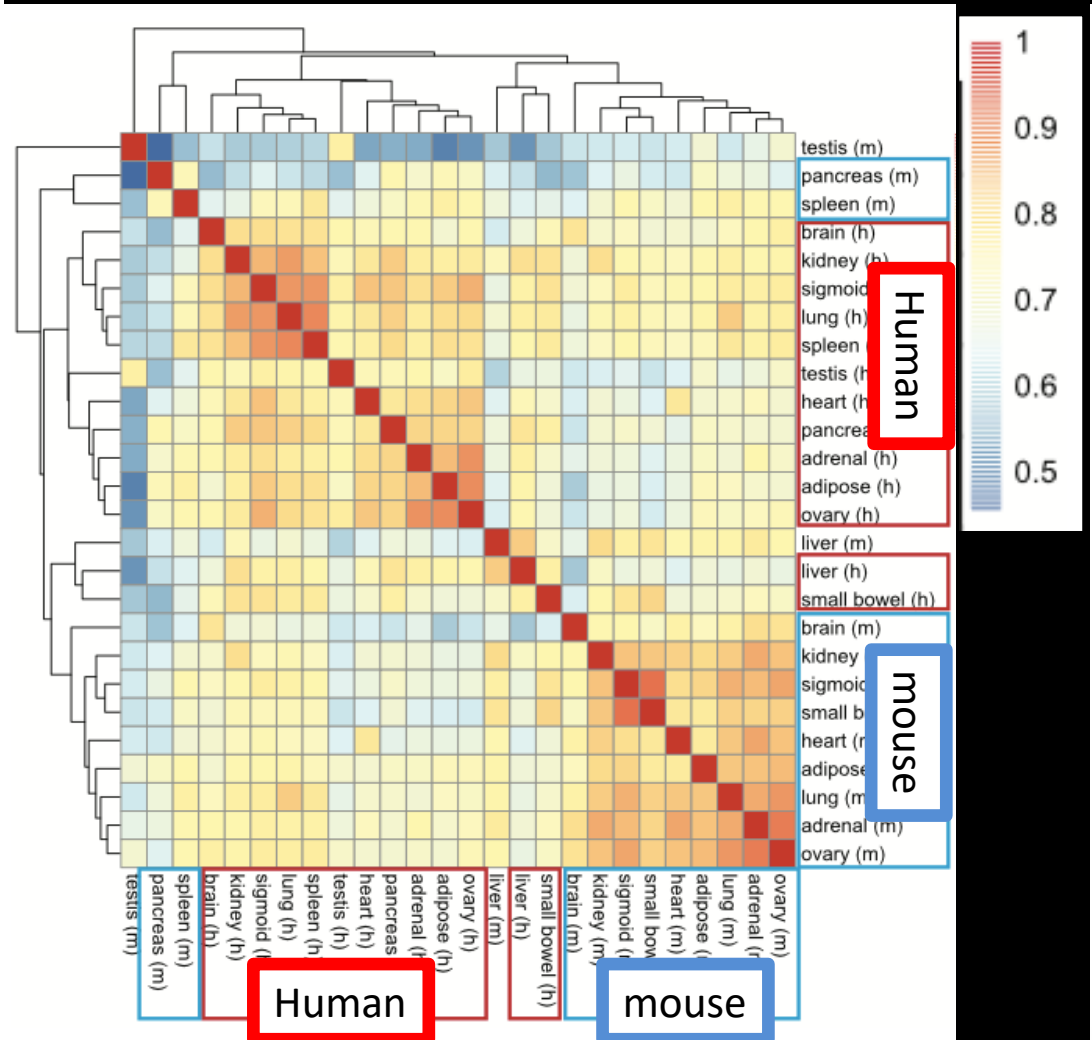


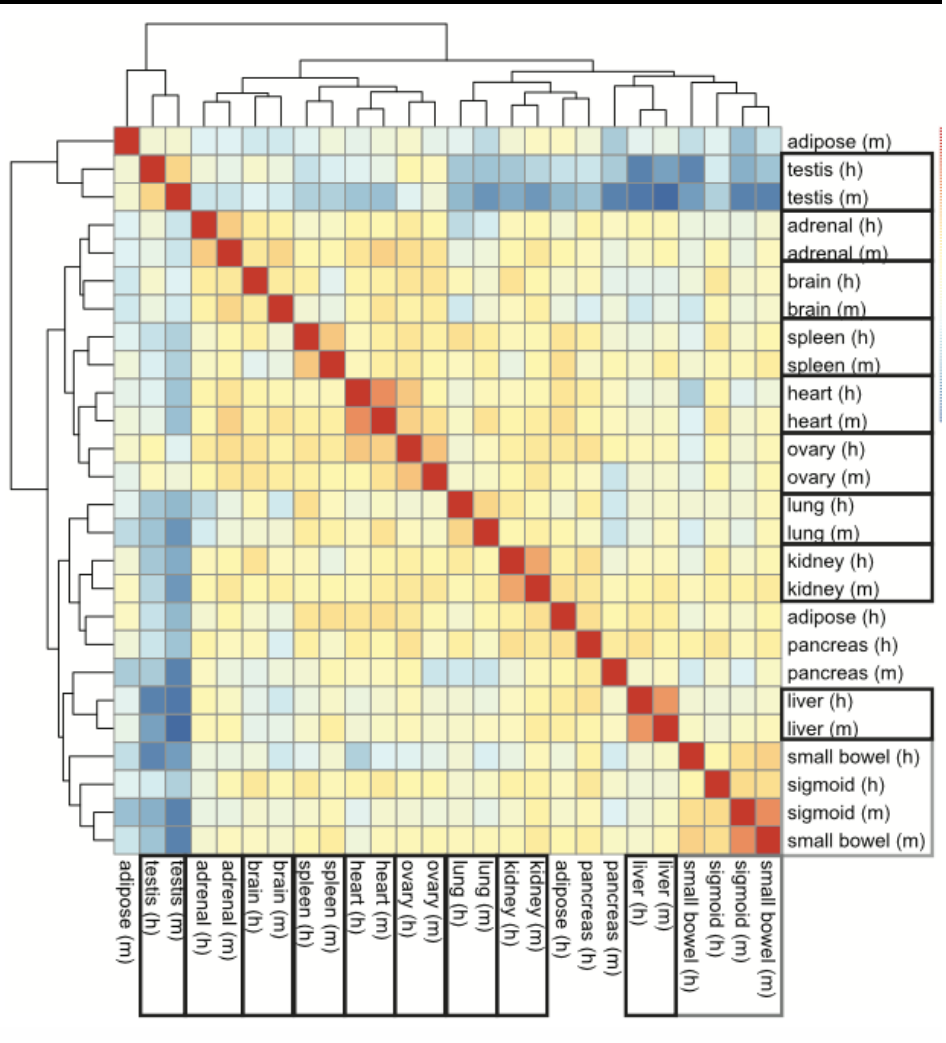
Should gene expression patterns group by species or tissues?

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

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





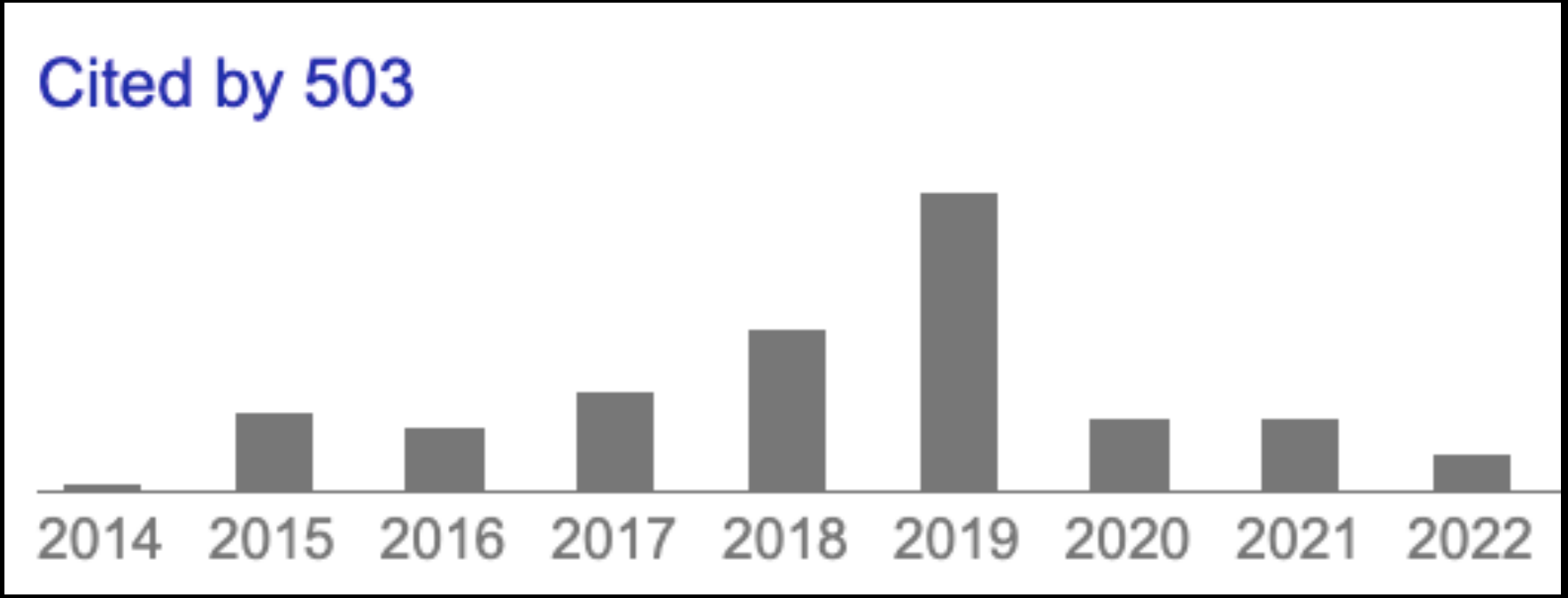
Why? this was a batch effect, which confounded sequencing grouping with biological grouping

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?

.... why is this still being cited?

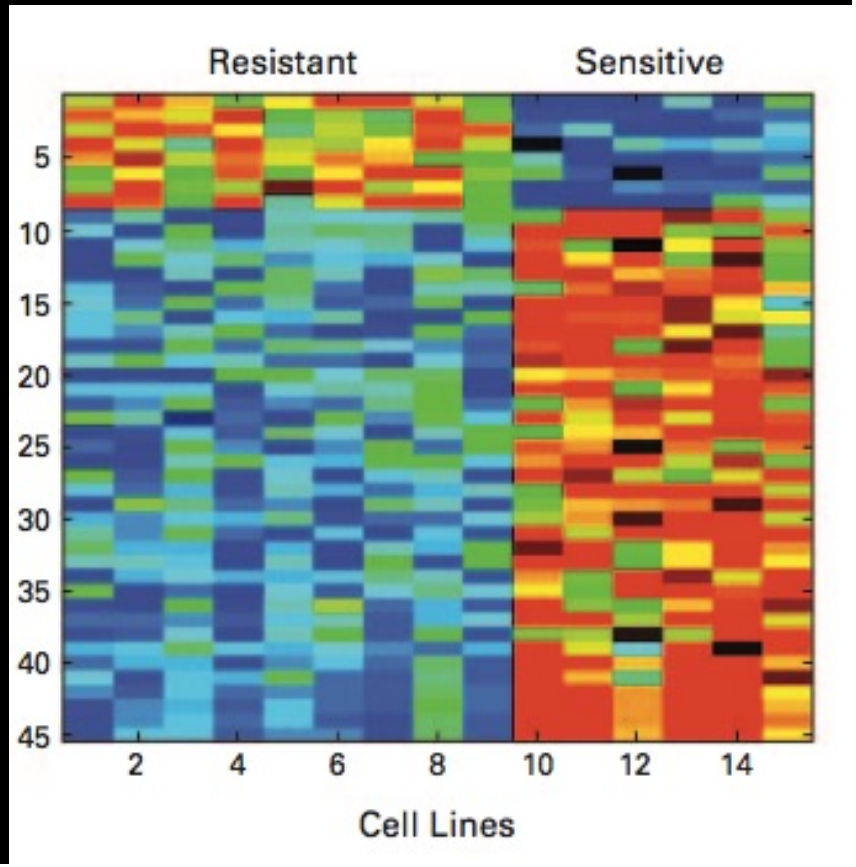


Do you want significant results? use Excel

- 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

FORENSIC BIOINFORMATICS AND REPRODUCIBLE RESEARCH IN HIGH-THROUGHPUT BIOLOGY

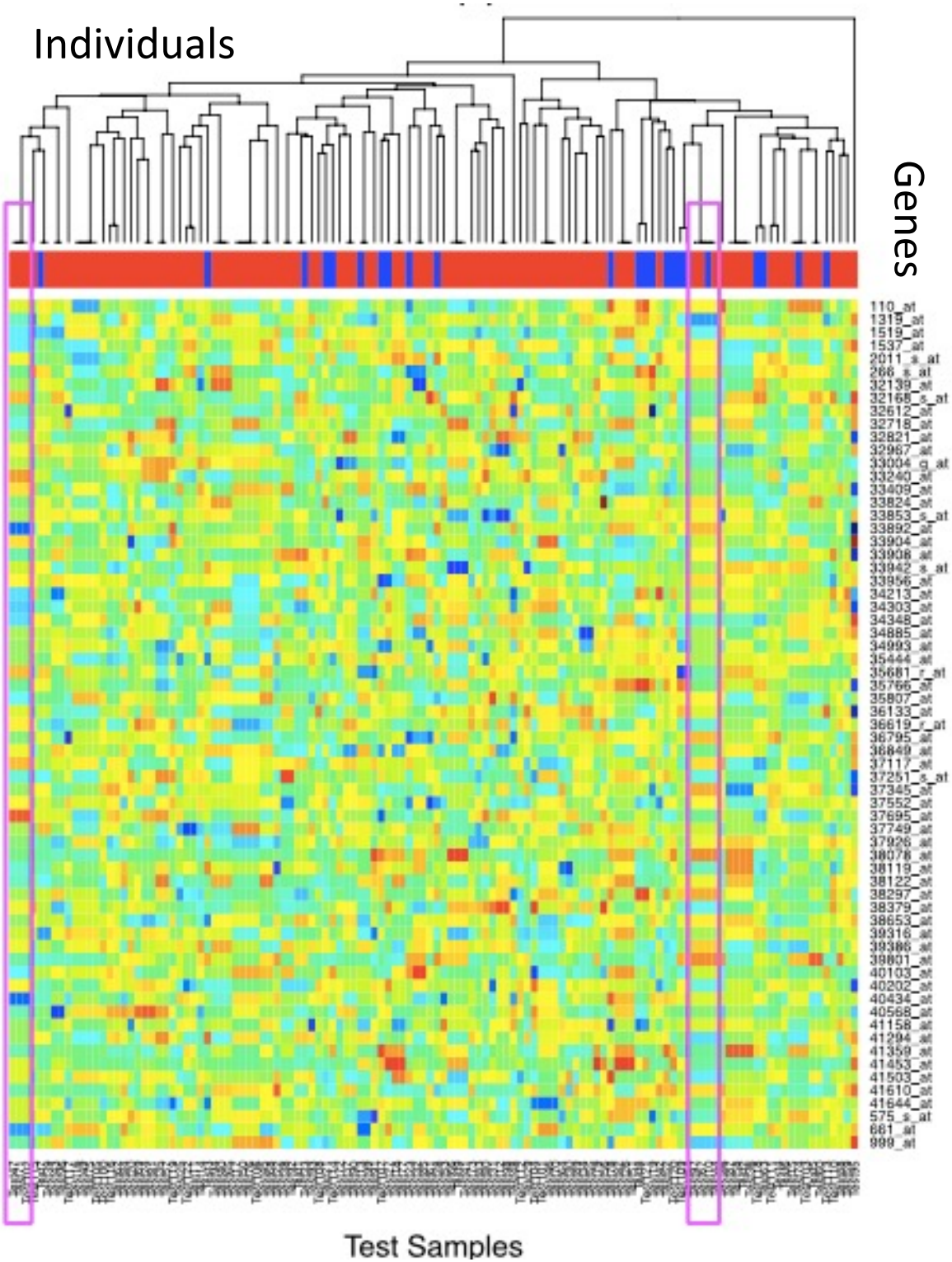
"Data processing, however, is often not described well enough to allow for exact reproduction of the results,

Thanks: Malachi Griffith

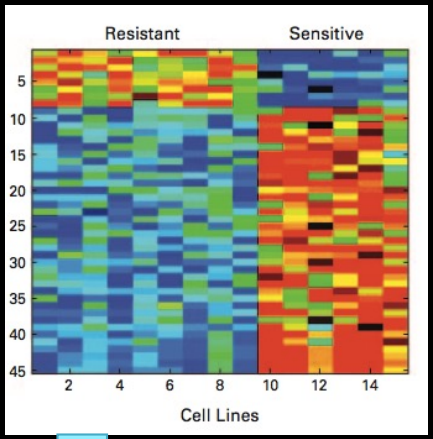
Baggerly and Coombes 2009

Digging revealed:

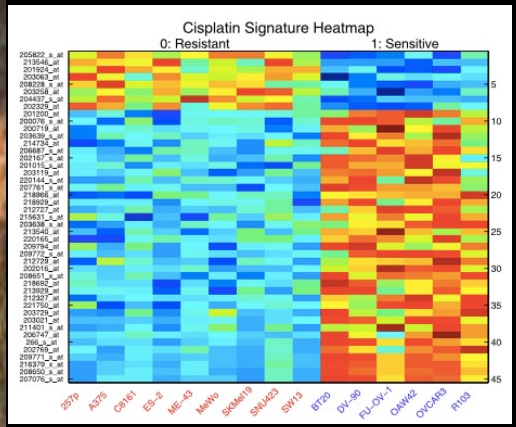
- Instances of repeated sampled data
- Only 84/122 test samples were distinct
- Some repeated samples labeled both sensitive and resistant
- Row offset in data table



Published result

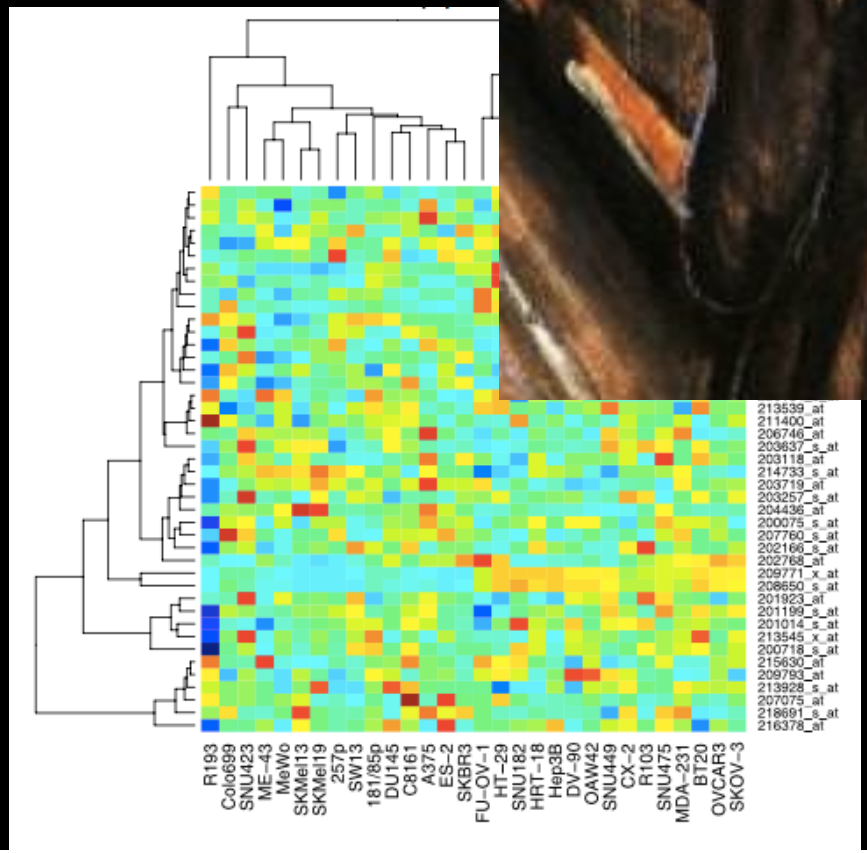


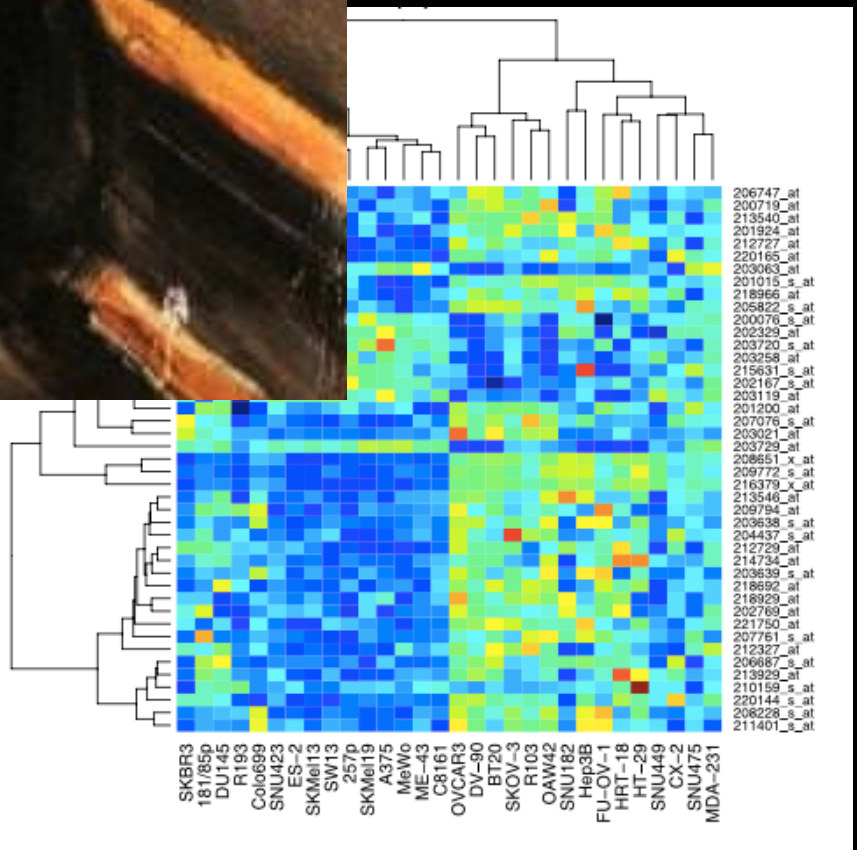
Error introduced result



Reanalysis with "c

1 row offset introduced





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

Can we reduce these type of publications?

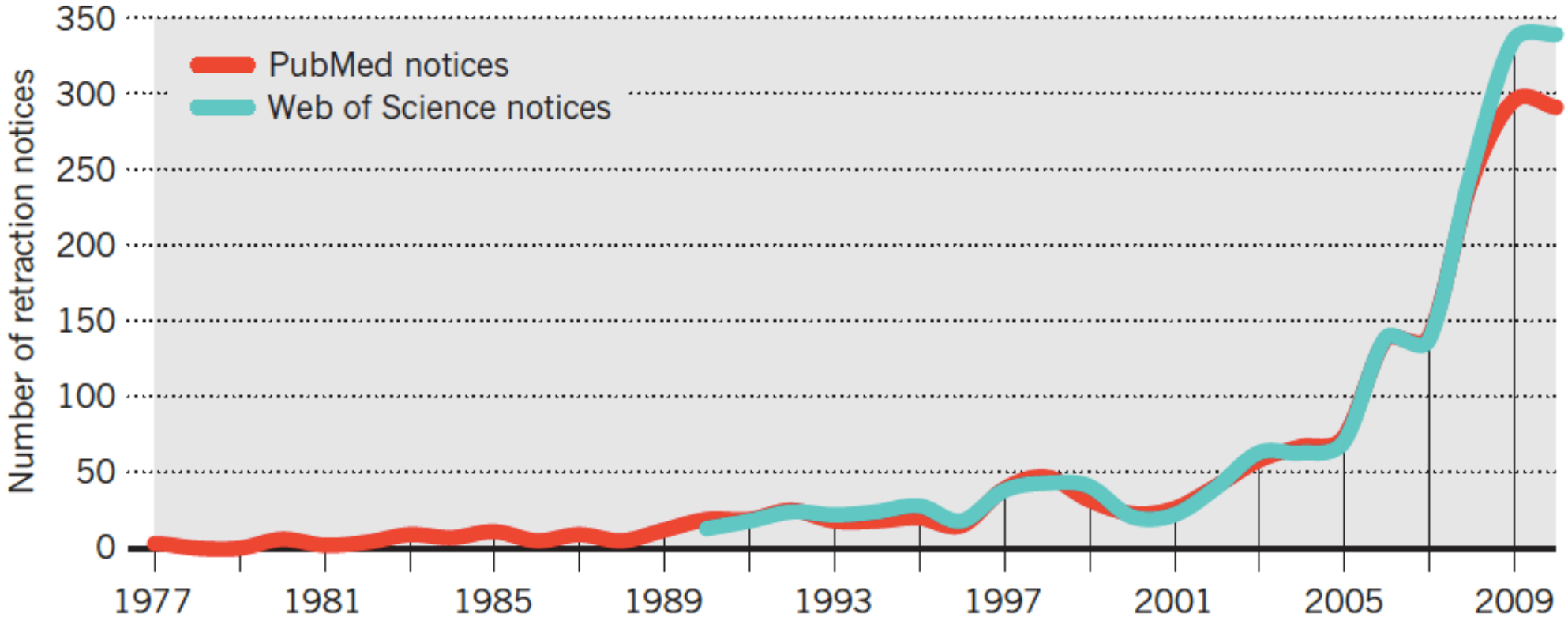
YES!!!!!

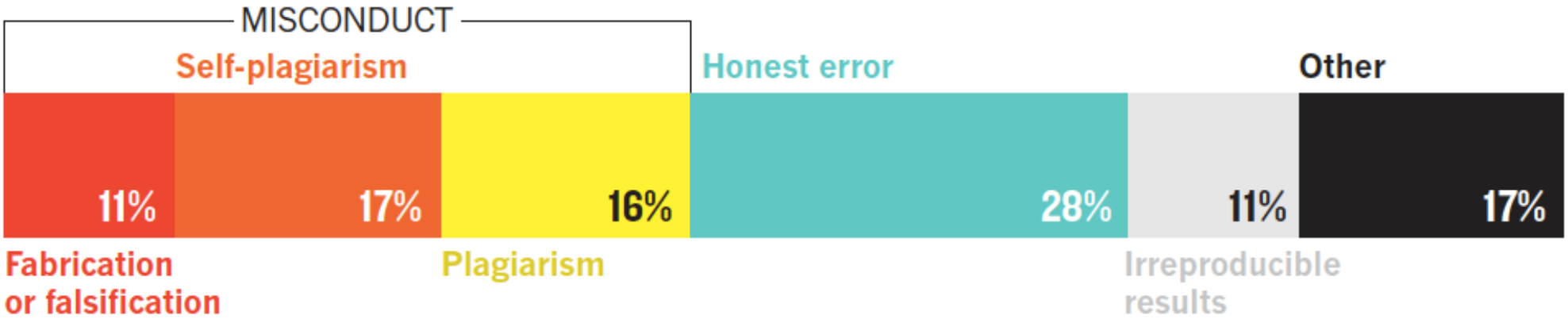
- Work better as a community, check each others code
- As author, as supervisor, as reviewer, as Associate Editor, make sure all studies you touch :
  - Have all code and raw data open source
  - Analyzed datasets open source
  - Methods clearly described

THE TROUBLE WITH RETRACTIONS

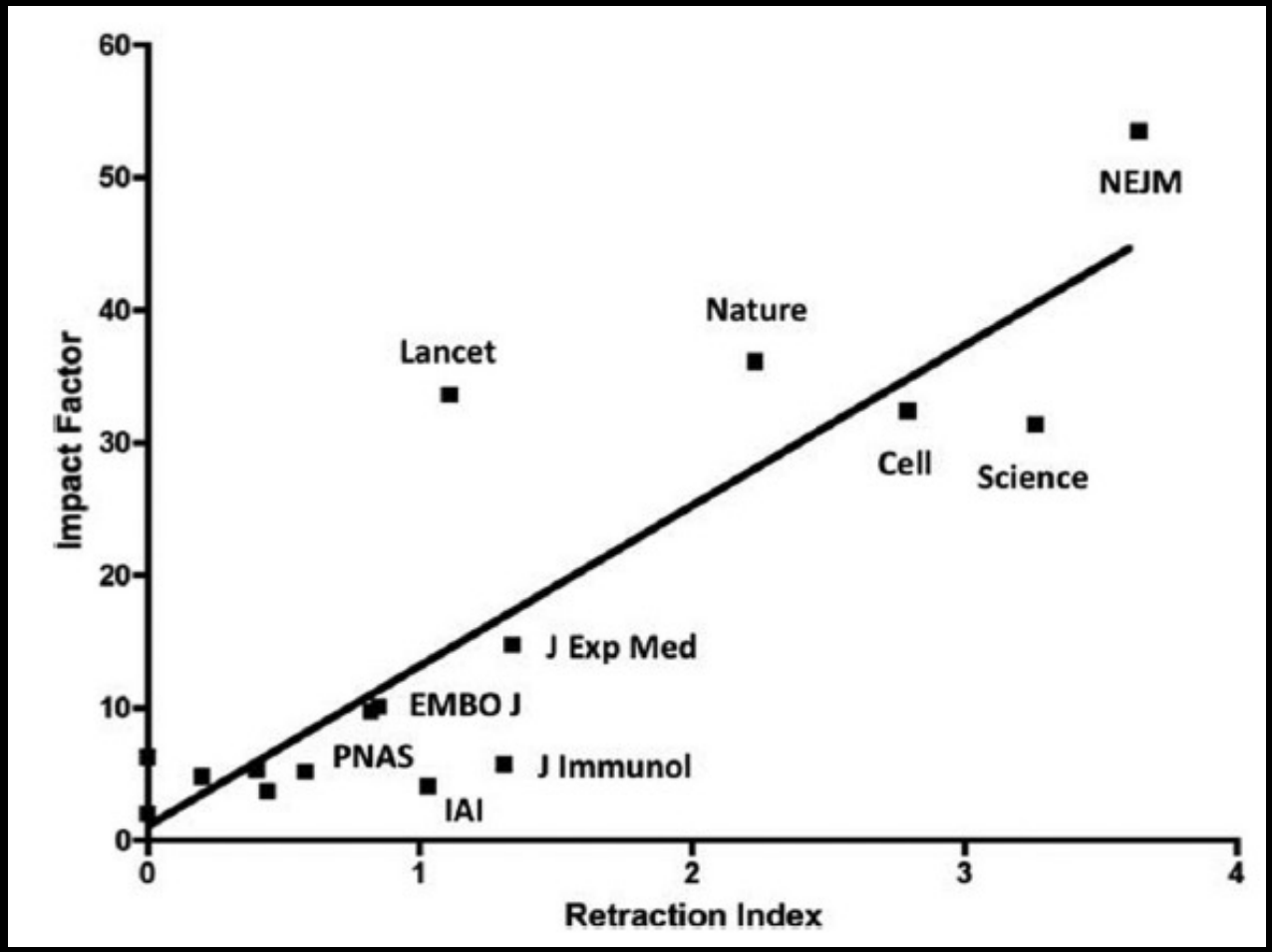
BY RICHARD VAN NOORDEN

A surge in withdrawn papers is highlighting weaknesses in the system for handling them.





The trouble with retractions: Nature News 2011



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

Fang 2011 Infect. Immun.

- 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 DS[1], Balakumaran BS, Acharya CR, Vlahovic V, Walters KS, Garman K, Anders C, Riedel RF, Lancaster J, Harpole D, Dressman HK, Nevins JR, Febbo PG, Potti A.

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

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 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.

So ... there are lots of 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 and error checks to catch errors the I KNOW I WILL MAKE

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

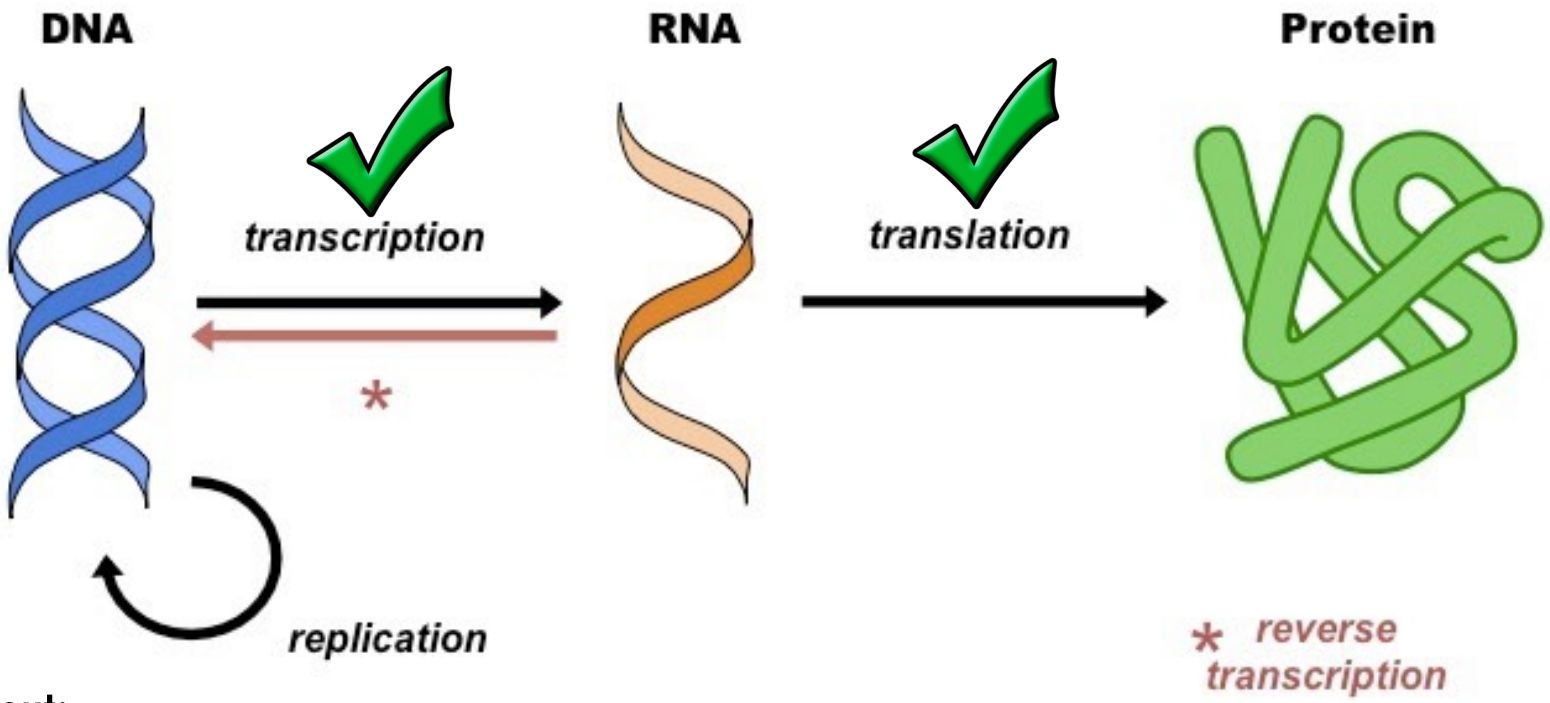
- think we can predict the future, cause we think we know how things work ... like:
  - gravity, your car, sunsets
  - weather, the stock market, Covid ...
  - the central dogma .....

Hindsight bias

the knew-it-all-along effect

Three Levels of Hindsight Bias

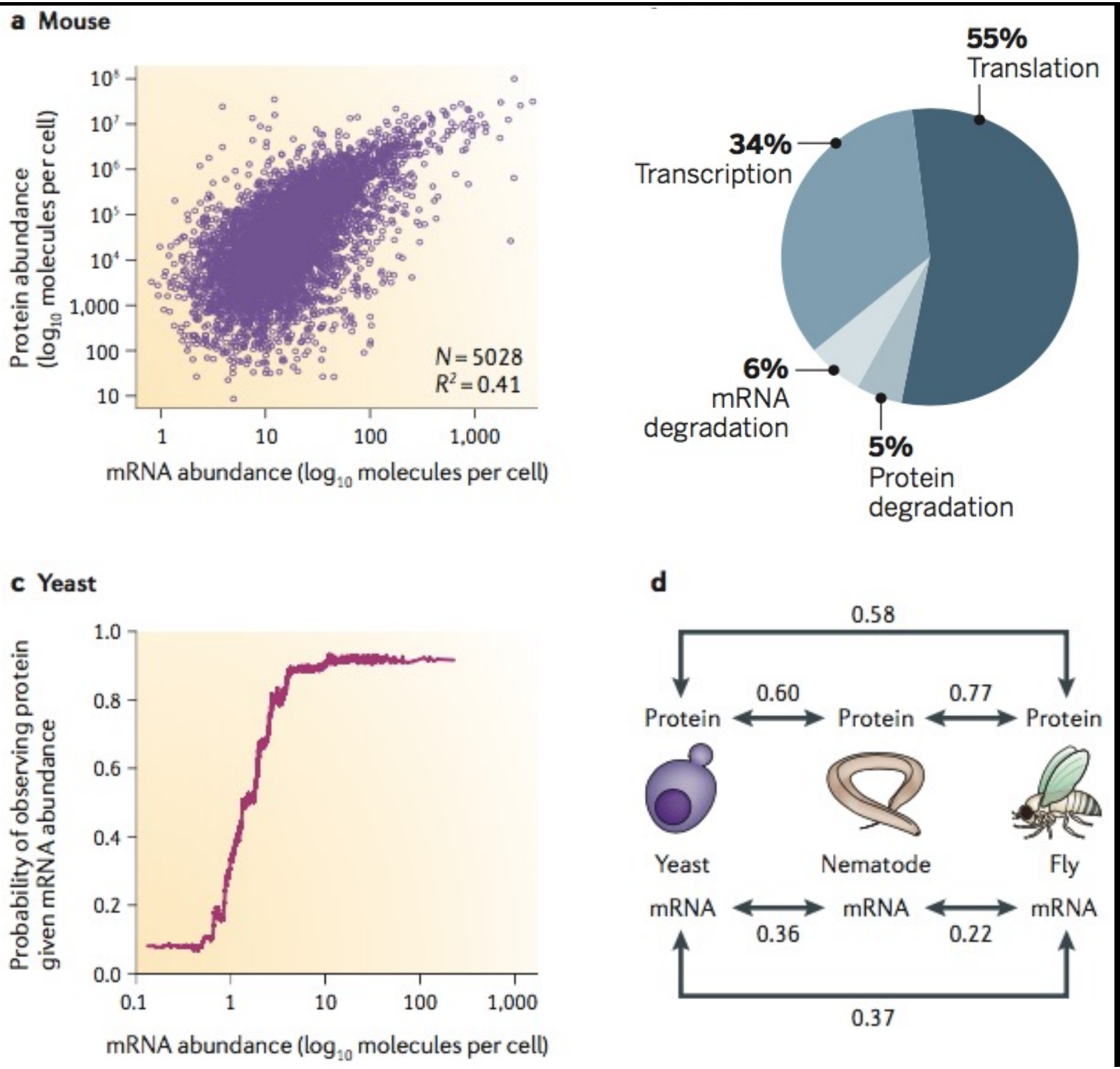
The central dogma



<u>What about</u>:

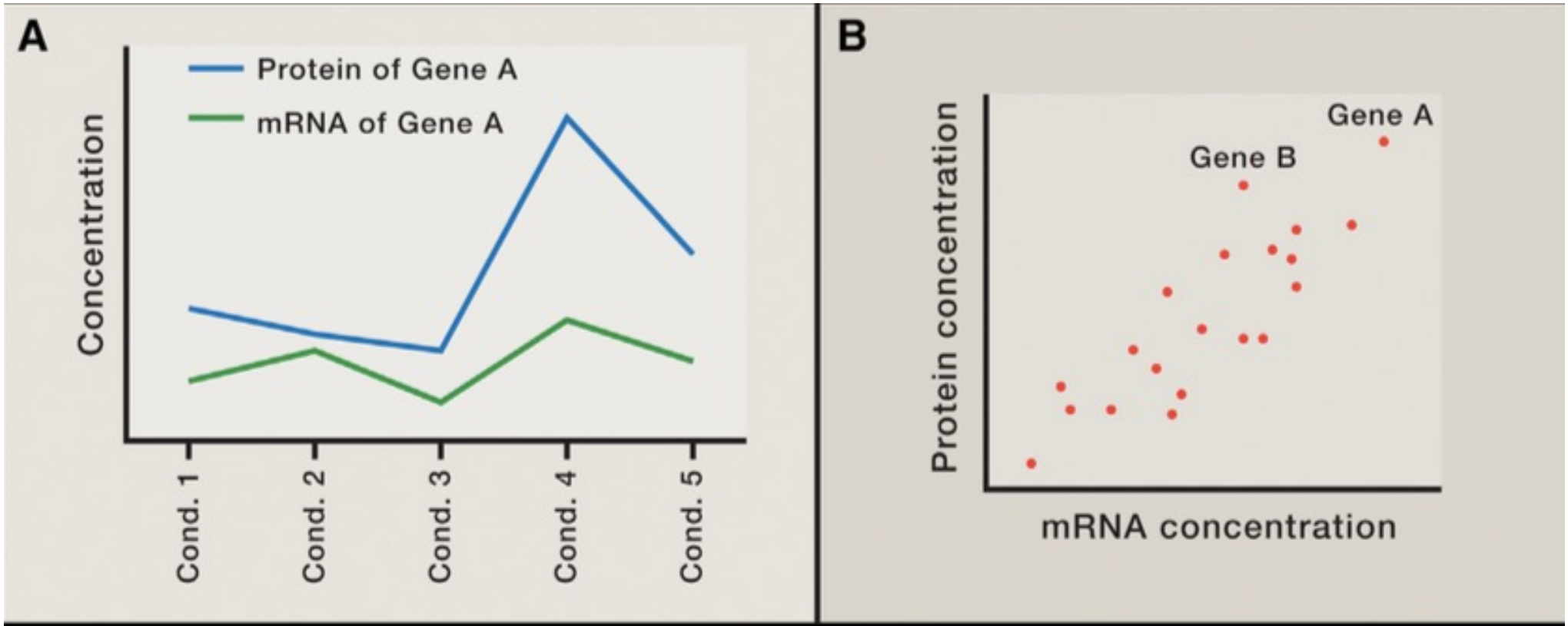
- Gene expression level from noncoding region?
- When and where a gene will be expressed from noncoding region?
- RNA to 2º structure?
- Amino acids to enzyme structure?
- Function based upon enzyme structure?
- Write a protein to do a specific enzymatic task?

mouse fibroblast cells

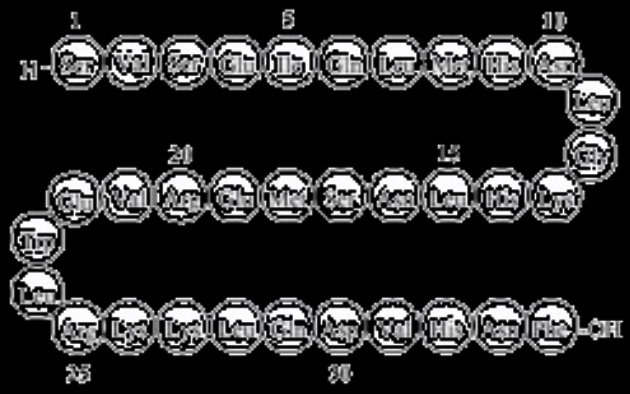


Vogel and Marcotte 2012 NRG

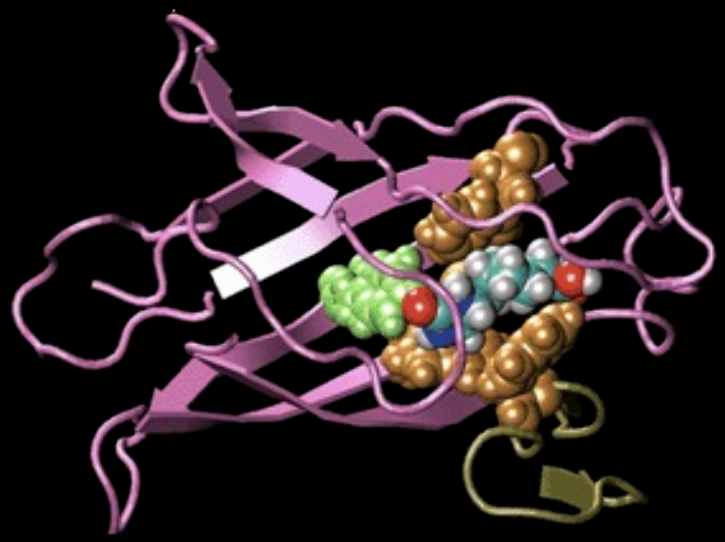
abundances of proteins are more conserved

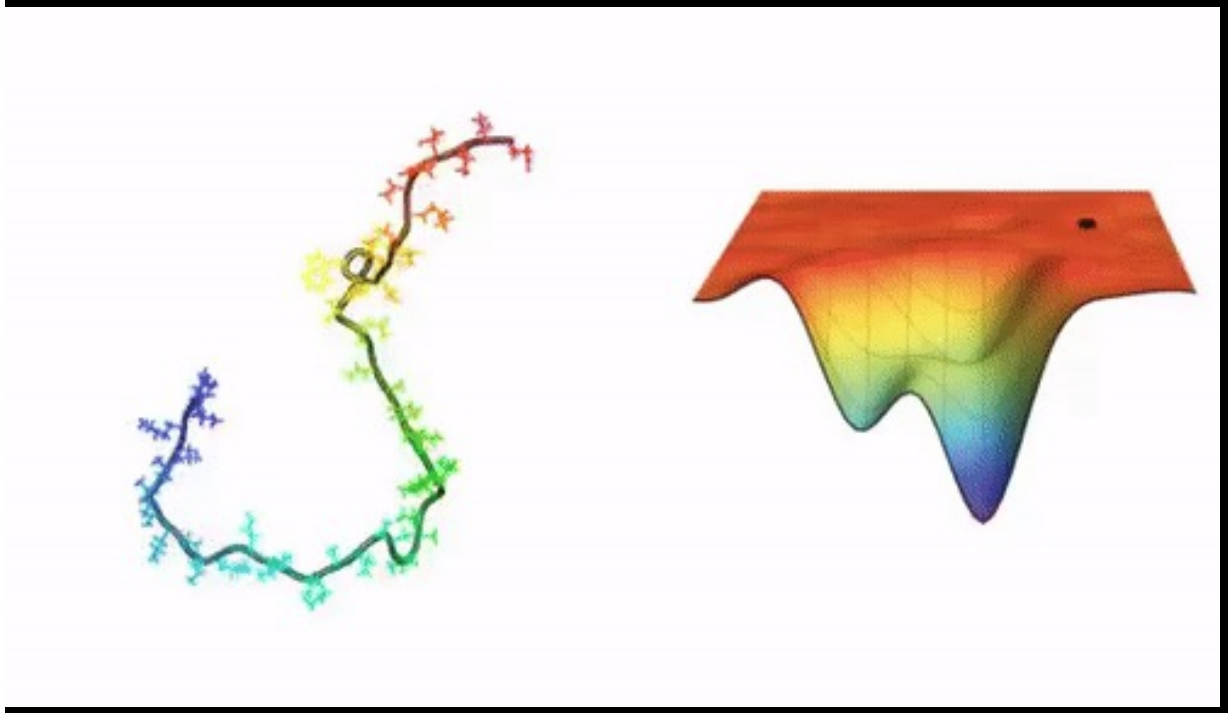

A
Protein of Gene A
mRNA of Gene A
Concentration
Cond. 1
Cond. 2
Cond. 3
Cond. 4
Cond. 5
B
Gene A
Gene B
Protein concentration
mRNA concentration

The Protein Folding Problem



How?

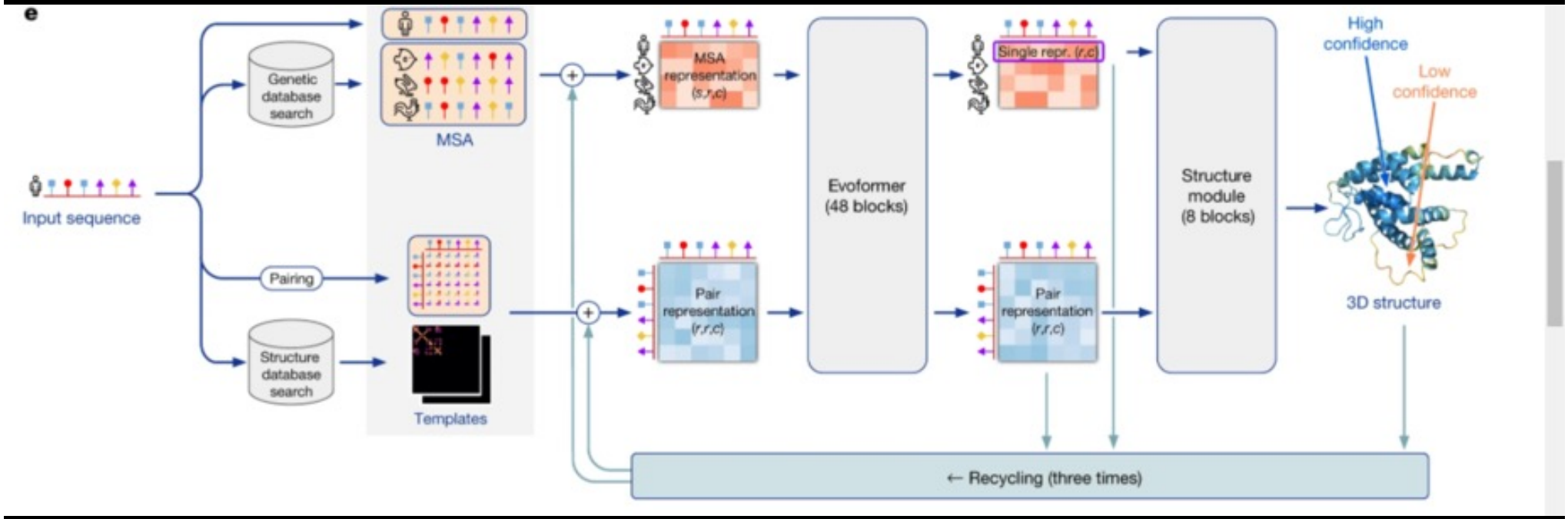




https://gfycat.com/greenpertinentkomododragon

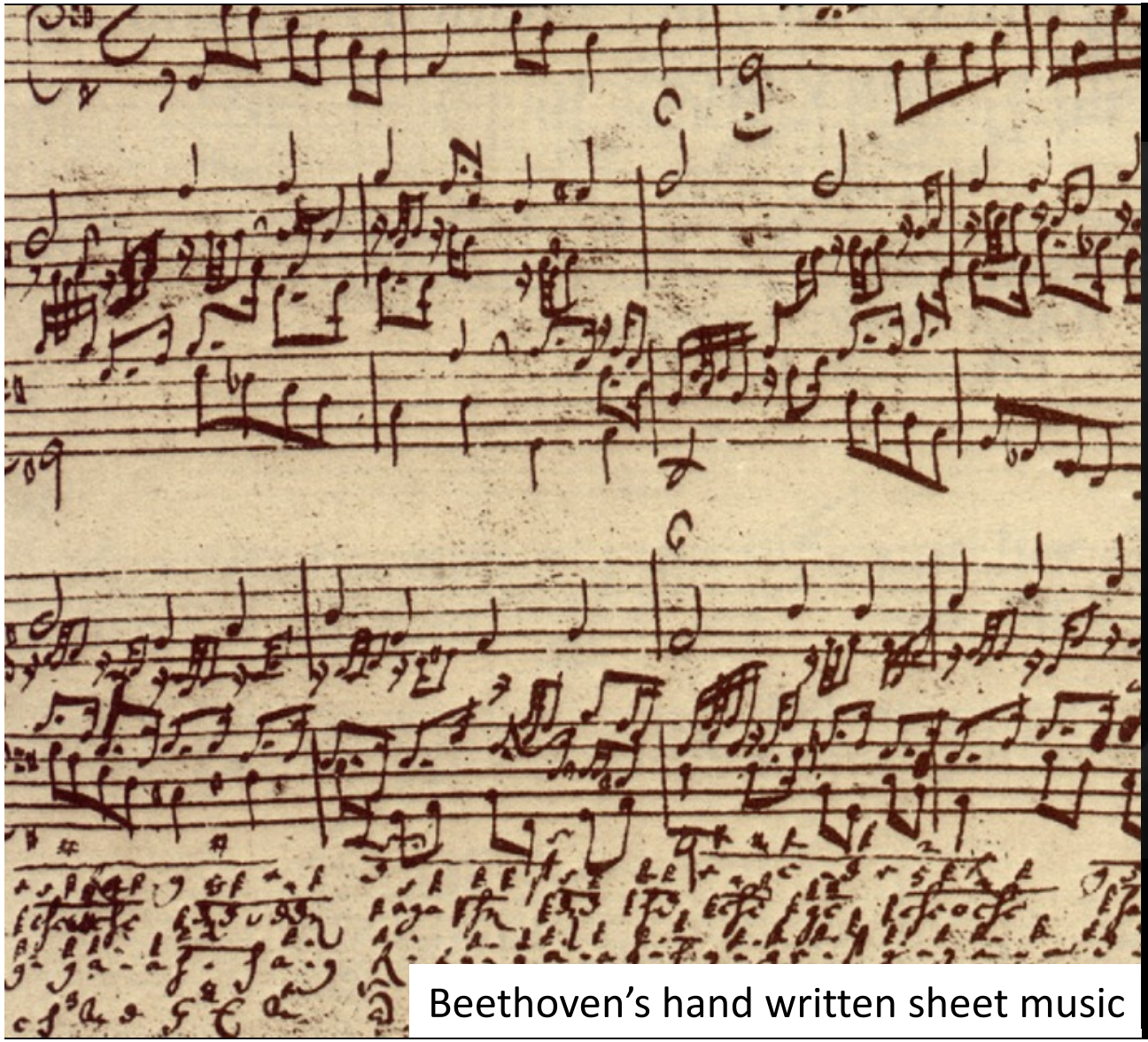
https://zhanglab.ccmb.med.umich.edu/image/Protein_design.gif

Alphafold 2



- Deep learning of existing patterns due to extensive observations
- Can predict most protein structures to high accuracy

But … peptide sequence to catalytic function …
"We don't know how to write that way"

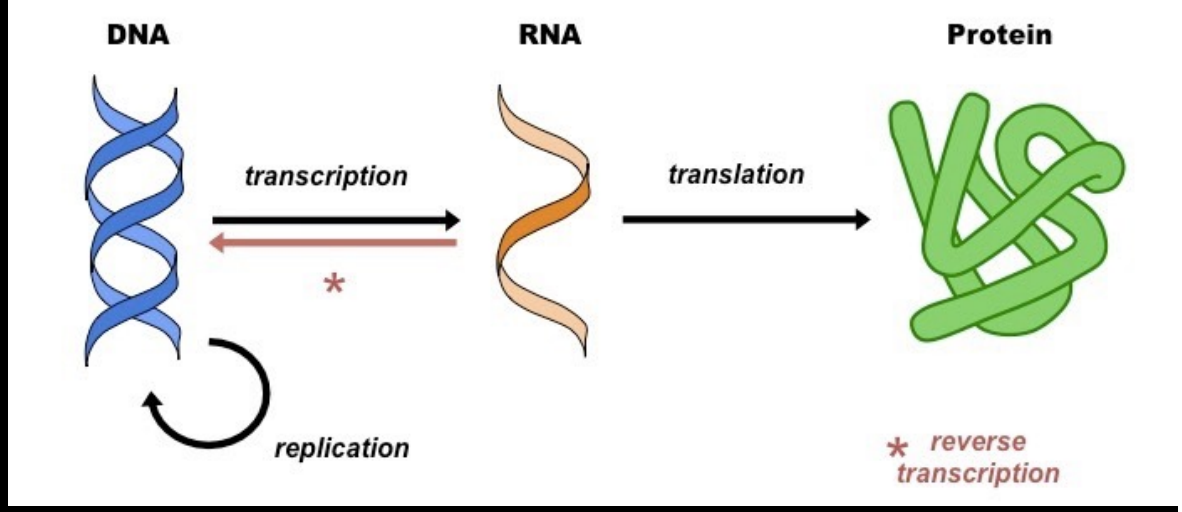


Beethoven's hand written sheet music

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

Francis Arnold
Nobel Prize winner (2018)

The central dogma



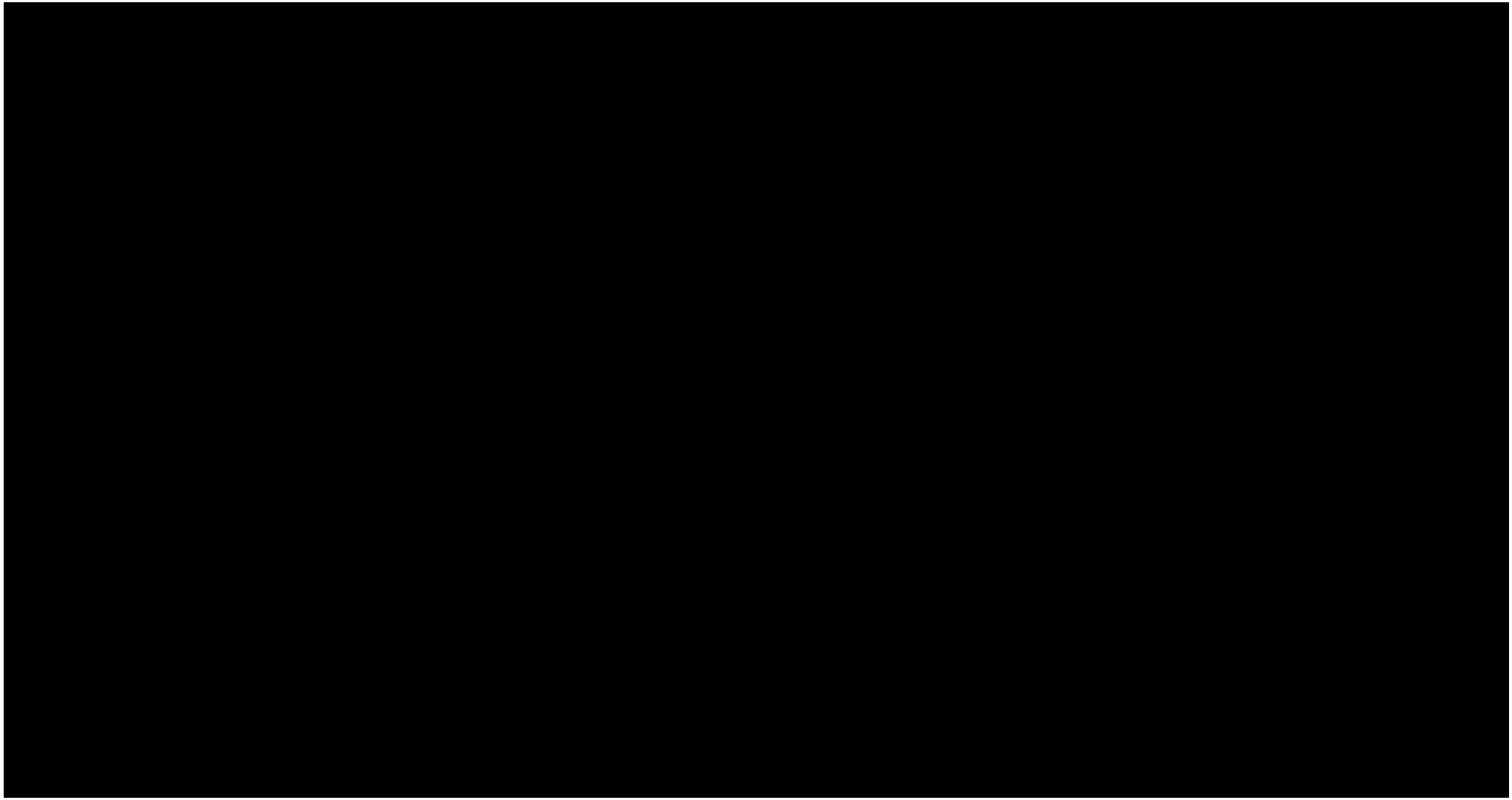
<u>What about</u>:

- Gene expression level from noncoding region?
- When and where a gene will be expressed from noncoding region?
- RNA to 2° structure?
- Amino acids to enzyme structure?
- Function based upon enzyme structure?
- Write a protein to do a specific enzymatic task?

- If AI can solve these, does that mean we understand?
- How limited to data input will solutions be?
- What about non-humans?

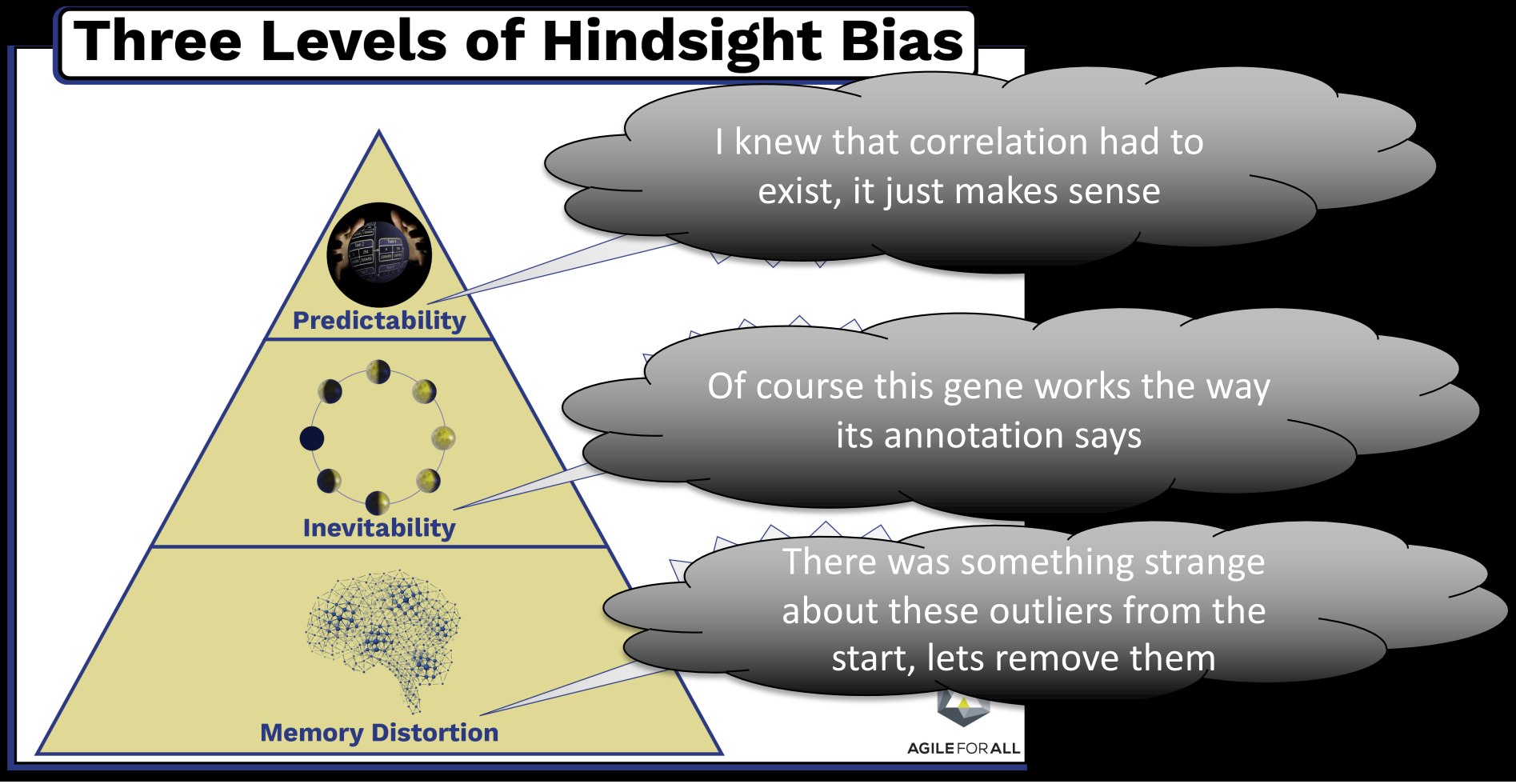
In sum, we think we how things work...

... but biology is exceptionally complex



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

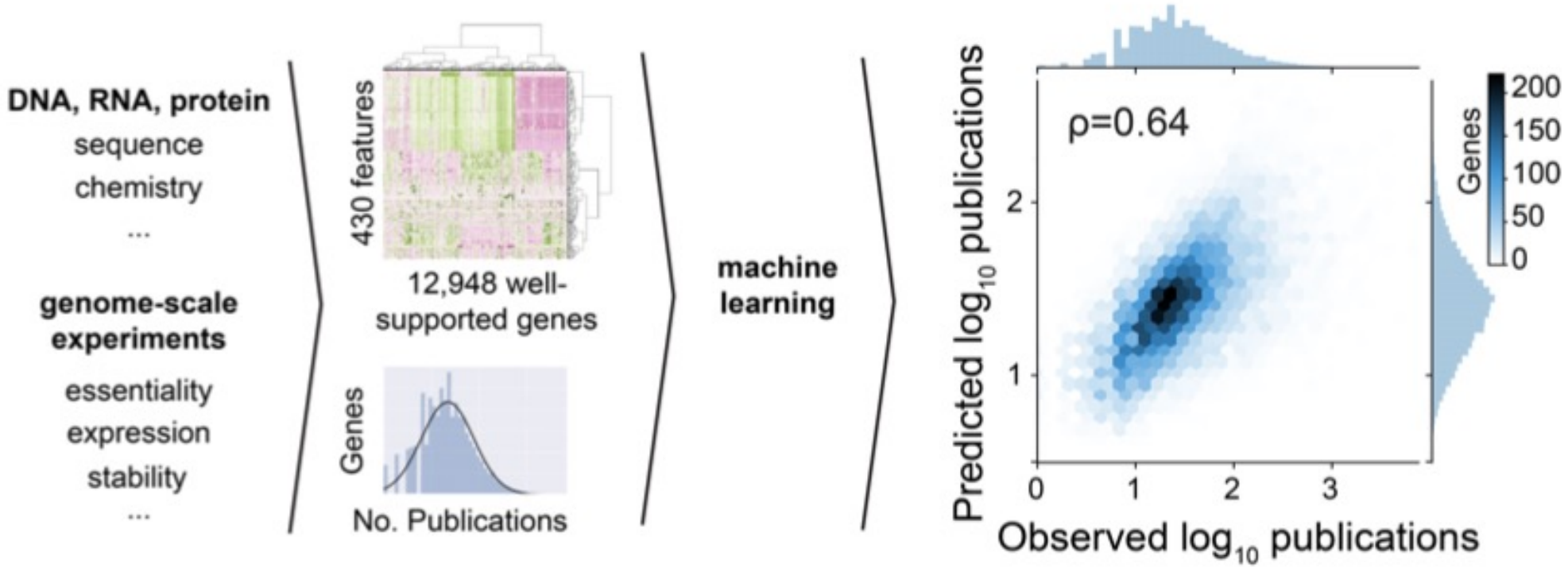
... but biology is exceptionally complex



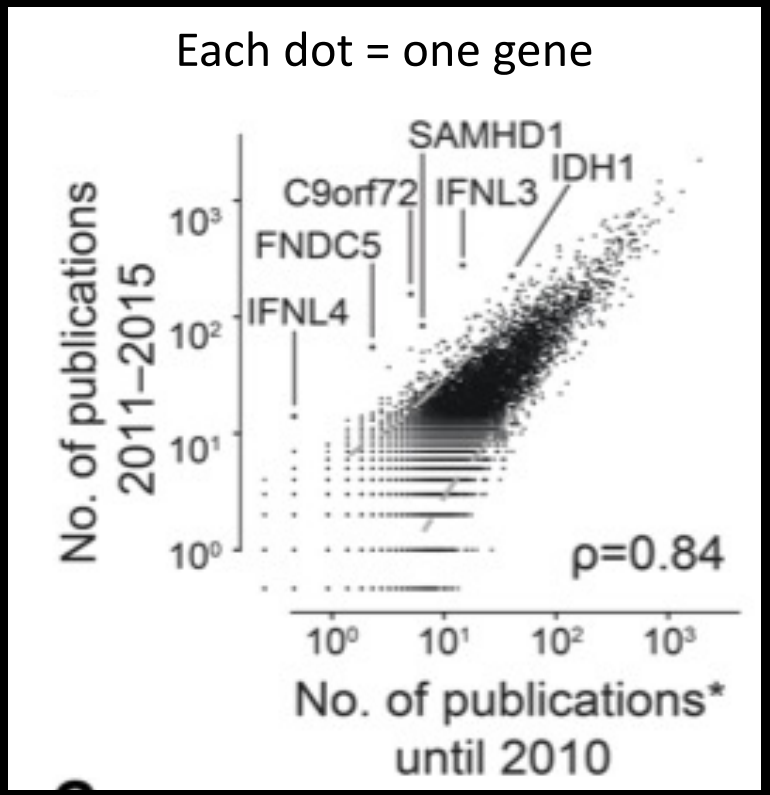
What about the genes we study?

Do we ever conduct "unbiased" investigations?

What if we looked at investigations by gene, over time



Stoeger et al. 2018 Plos Biology



- 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

Its hard to get money to study genes with unknown functions ...

Stoeger et al. 2018 Plos Biology

So .. how do we avoid Apophenia?

- Non-random patterns are abundant in genome scale data
  - We generally lack ability to calibrate our expectations
  - Null models, controls are very difficult to get "right"
- Double check your data and analyses
  - Plot your data, look at it, does it make sense on 1st principals?
- Test your hypotheses in independent ways
  - Genomics: independent datasets, alternative analyses, other levels
    - Separate collections, GWAS vs. K-mer GWAS, mRNA vs. protein
  - Manipulation: functional validation via manipulation of genes, pathways
    - Experimental evolution, CRISPR KOs, environmental perturbations


sallyedelsteincollage.com