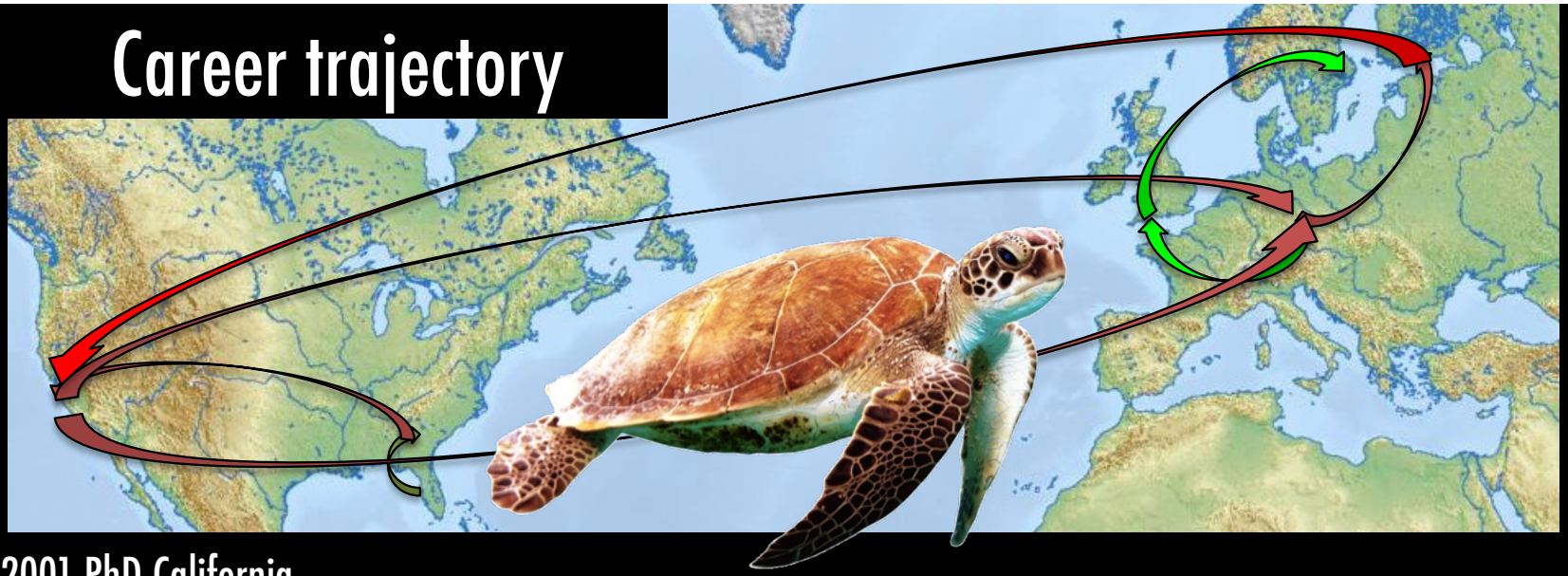
Lies, damn lies, and .... genomics

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

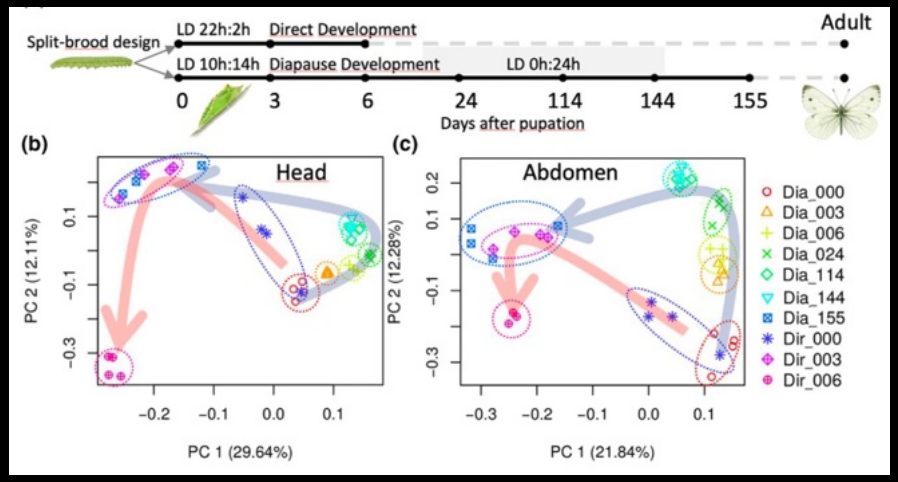
I was able to put science first & have fun along the way

Something you likely would never know about me

I am a Judge for the American Field Trial Clubs of America (since 2003)

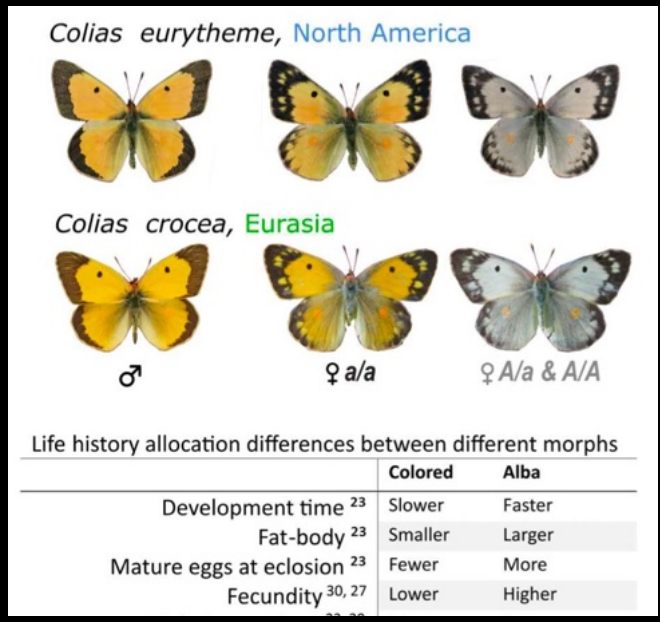
Ecological & Evolutionary Functional Genomics

Circadian and seasonal clock evolution



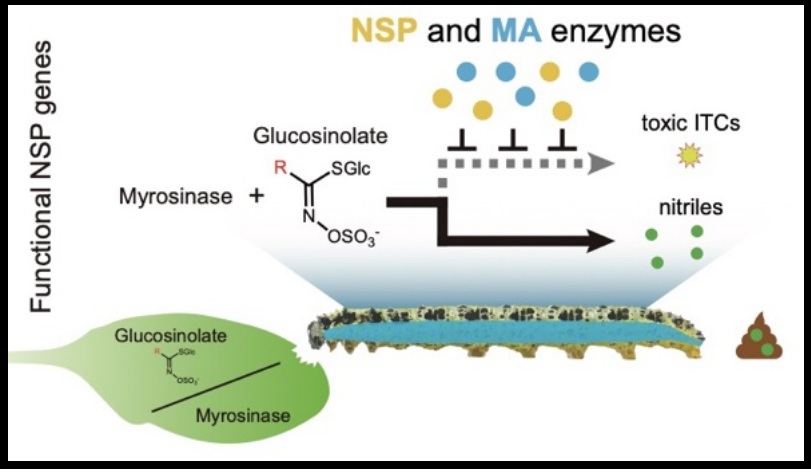
Butterfly-plant coevolution dynamics

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



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 to 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 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.

	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

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

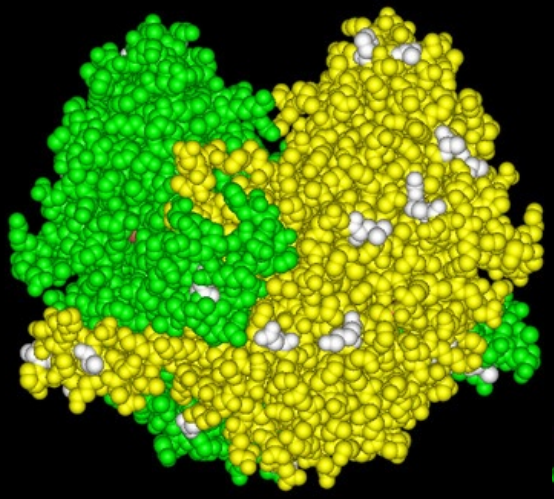
Colias eurytheme

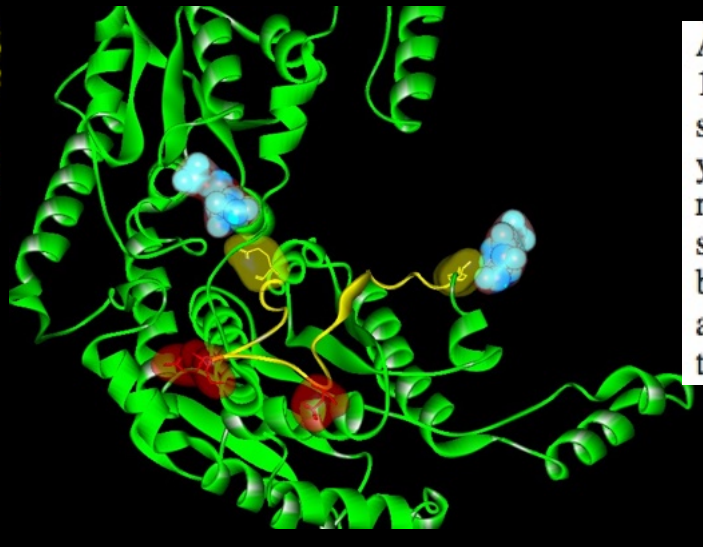
My PhD: use this DNA based molecular test of selection on a classic example of balancing selection from allozyme era

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

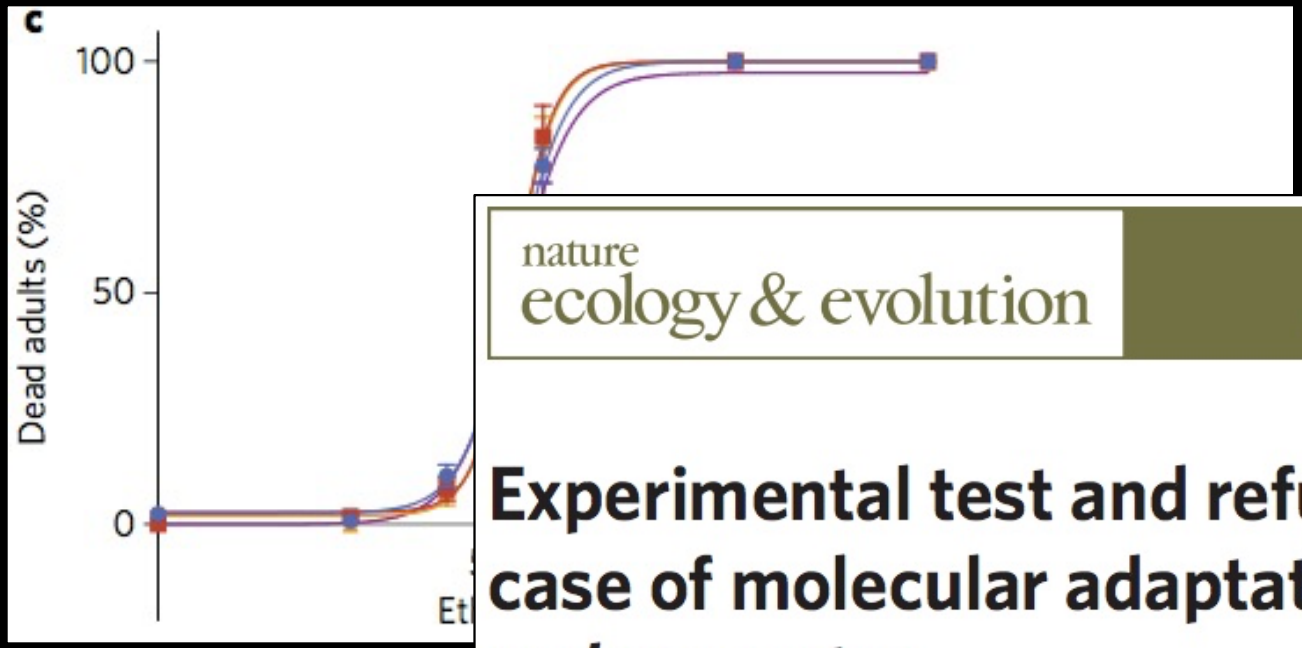


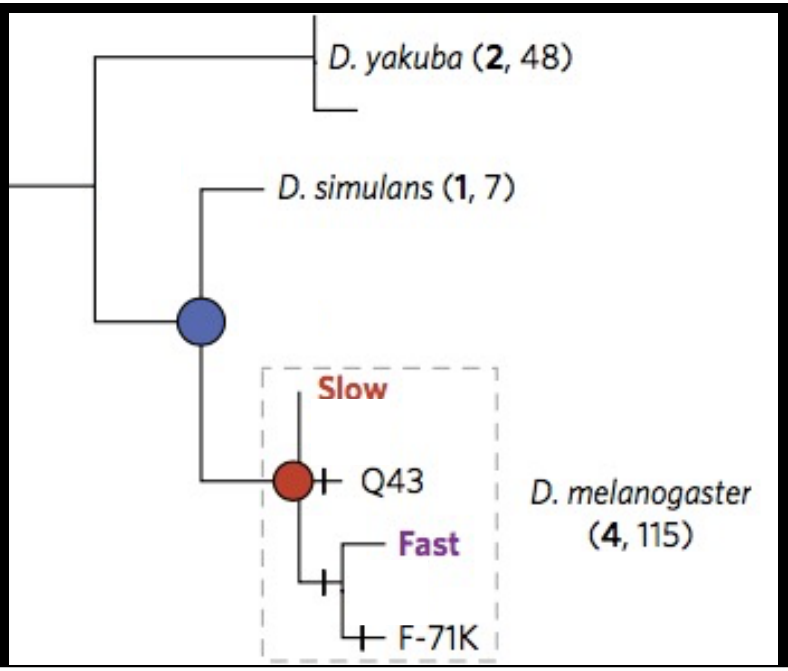


Among *C. eurytheme* and *C. meadii* PGI sequences, we find 126 synonymous and 20 nonsynonymous polymorphic sites. From their ratio, 6.3:1, neutrality predicts ~13 synonymous fixations alongside the two observed interspecies nonsynonymous fixations. But, *no* fixed synonymous sites were found (abo a 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

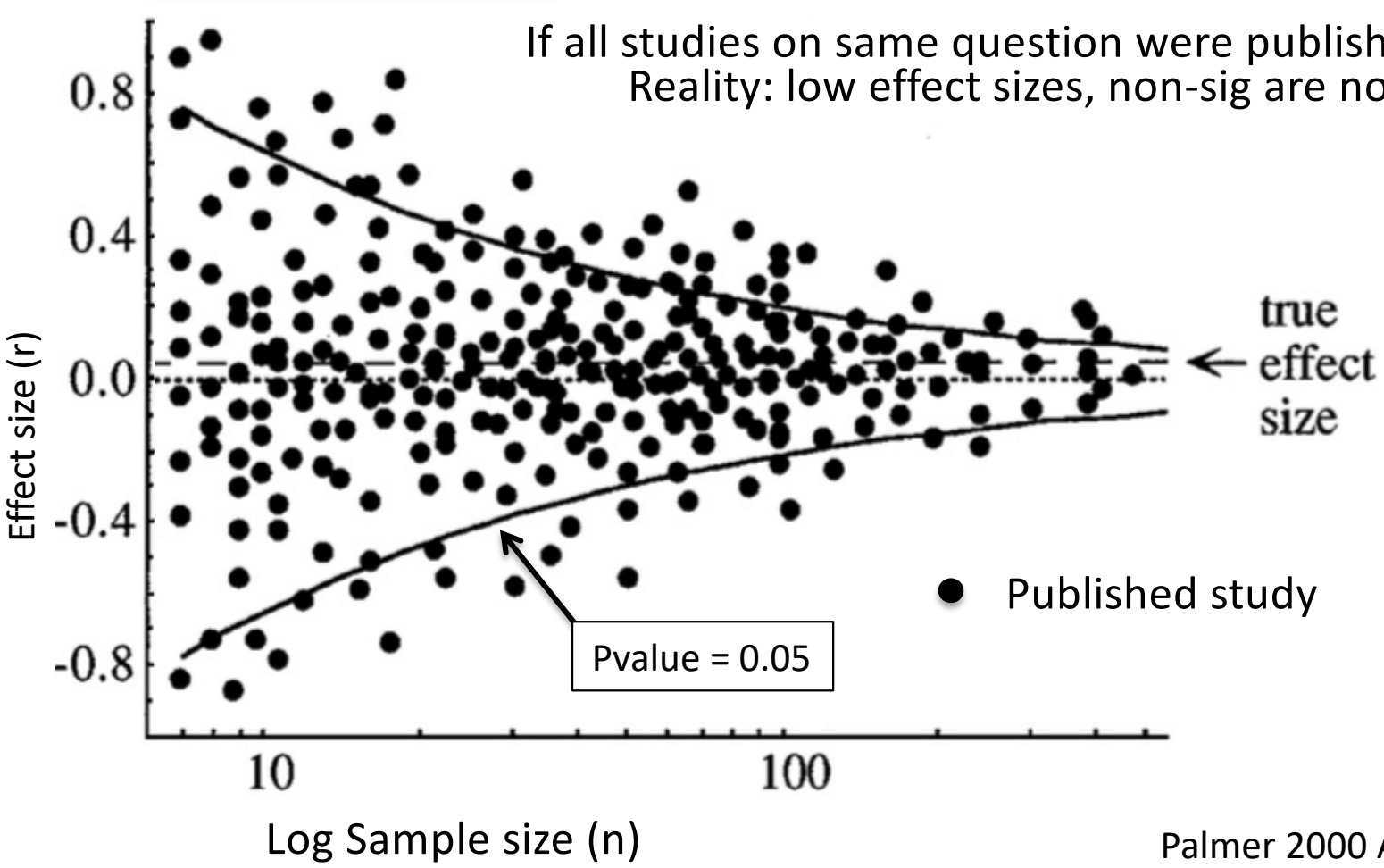
Publication replication failures

- 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

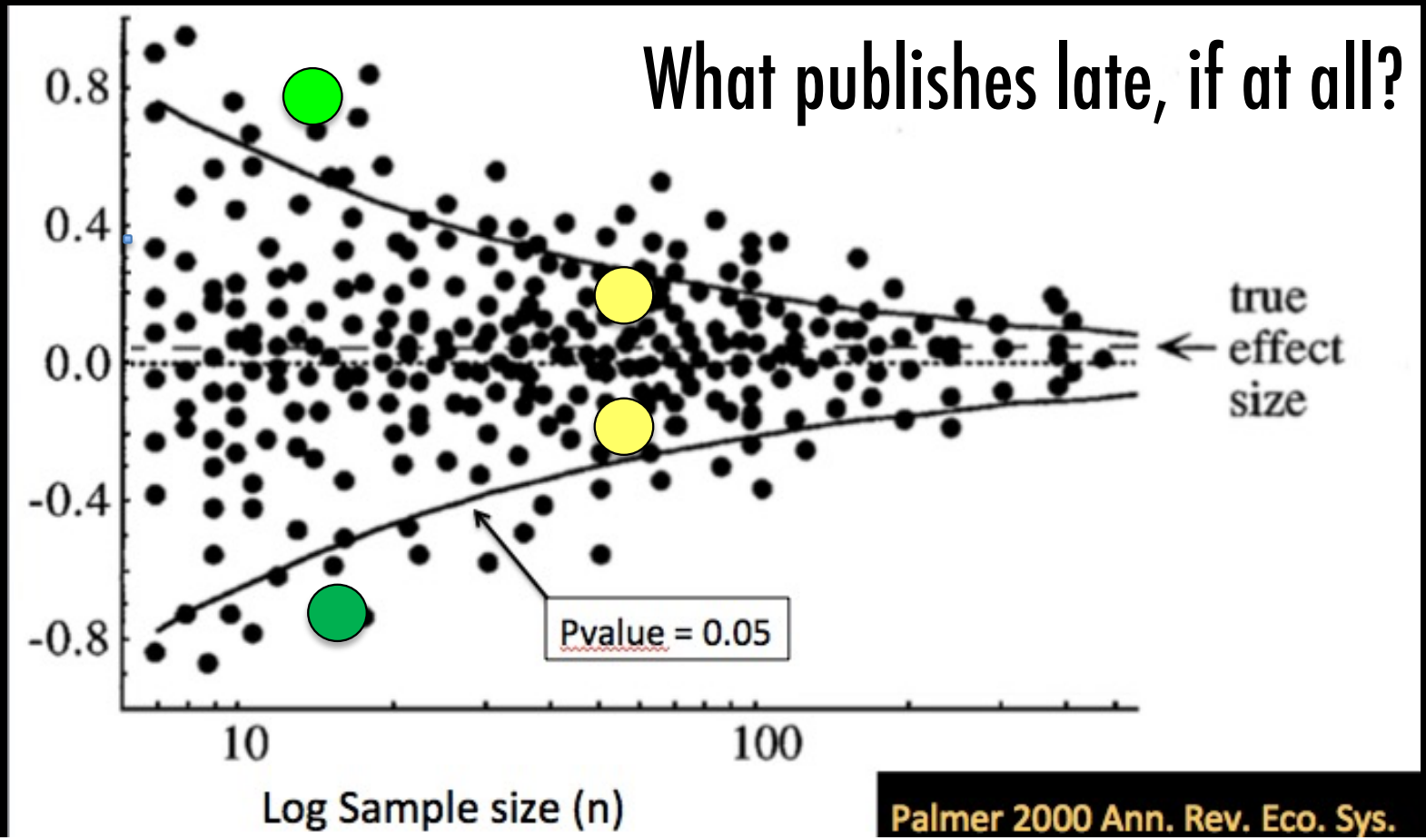
Publication bias can increase effect size



Palmer 2000 Ann. Rev. Eco. Sys.

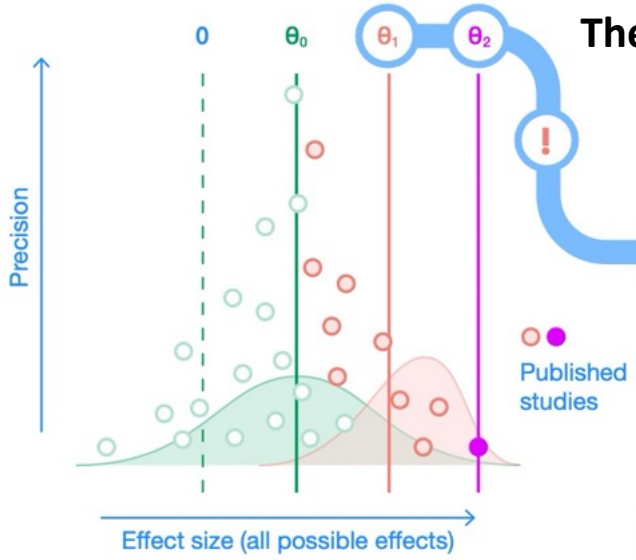
What if there is no replication?

What is most likely to publish first & where?



Palmer 2000 Ann. Rev. Eco. Sys.

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

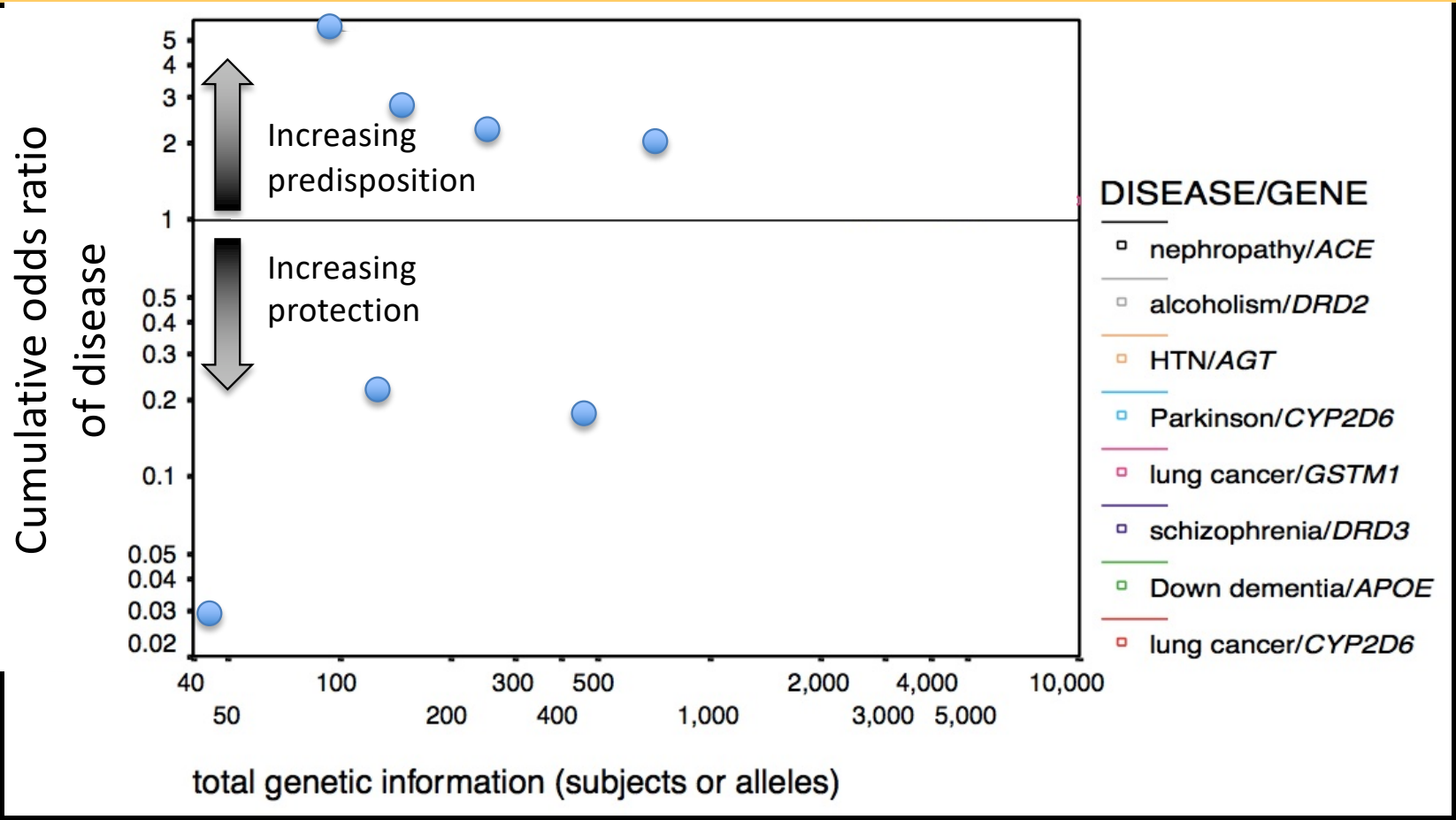
But ...
surely, this doesn't apply to genomics

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

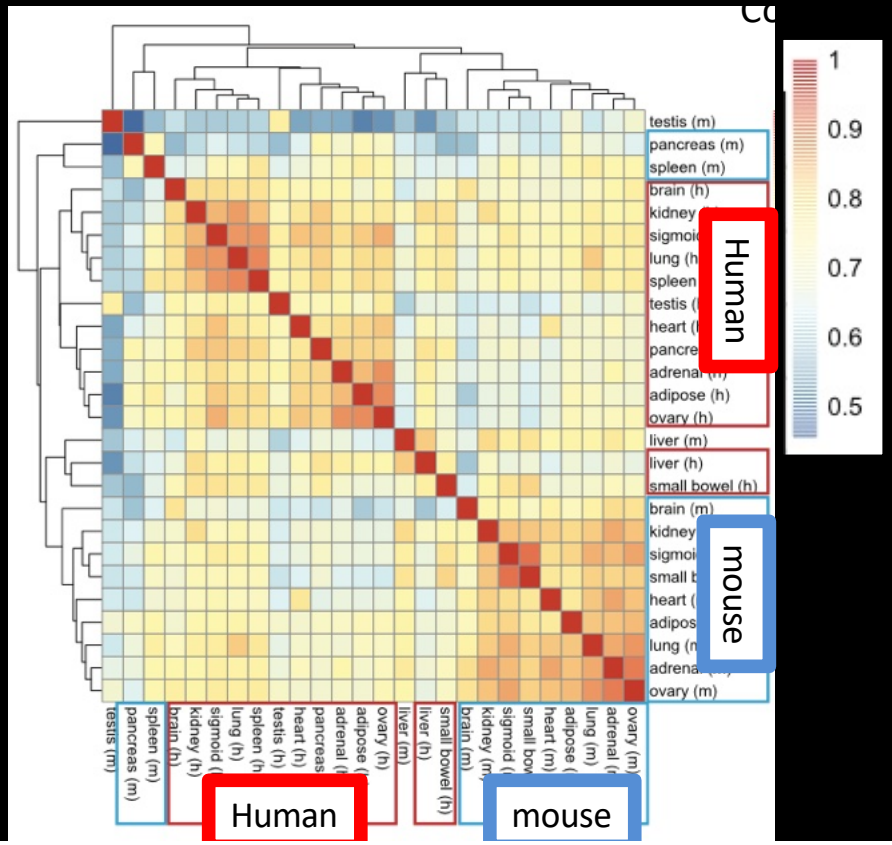
Comparison of the transcriptional landscapes between human and mouse tissues

"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

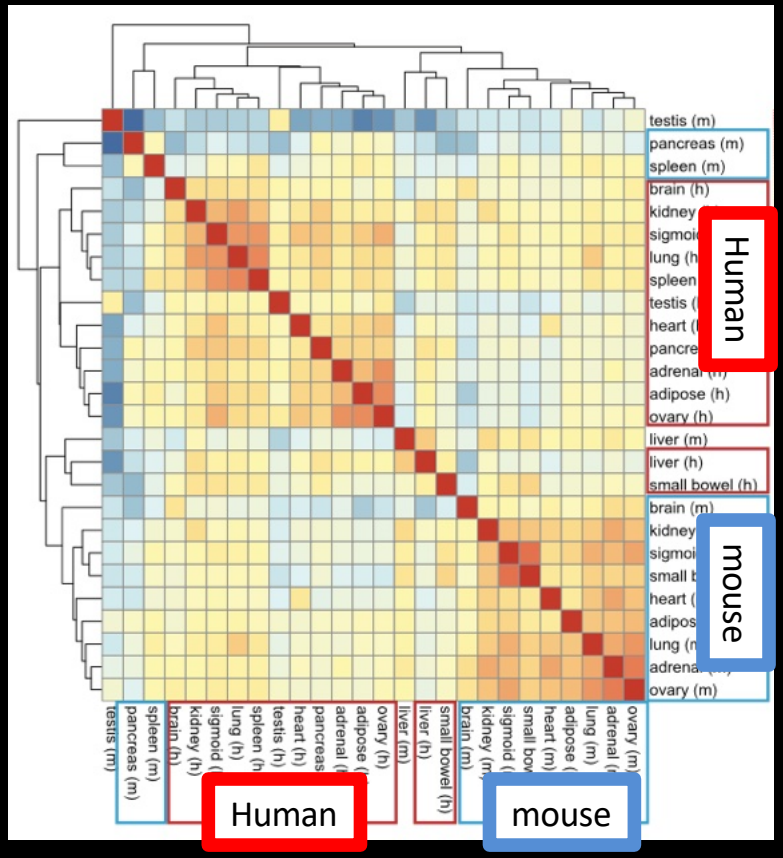
Authors found strong grouping of all organs by species, not by organ



Should gene expression patterns group by species or tissues?

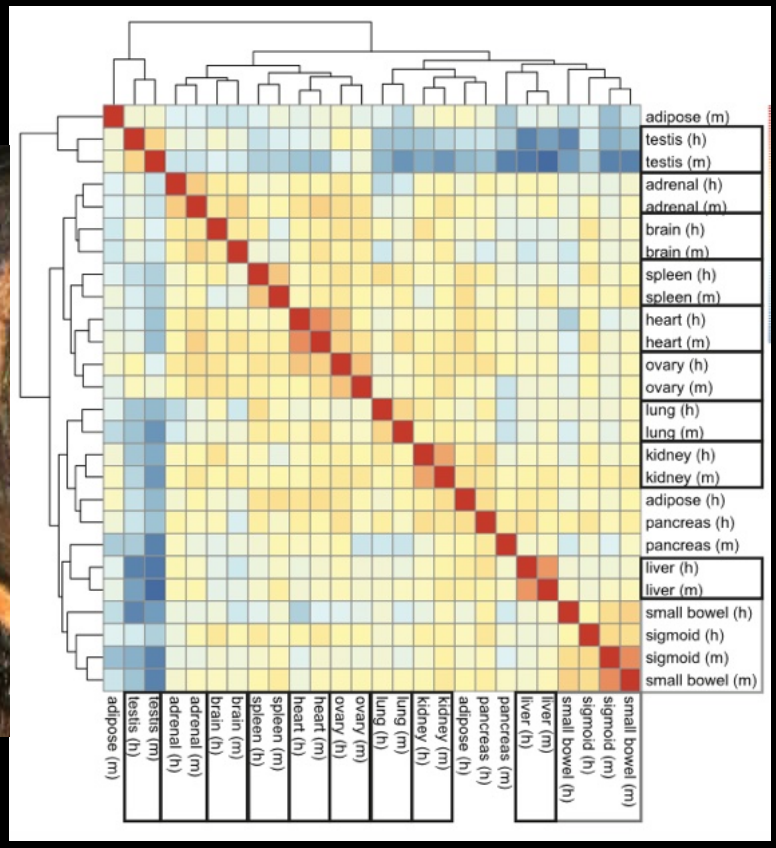
What do we expect from first principals, evolutionary relationships?

"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



Cause = batch effect at sequencing core of 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

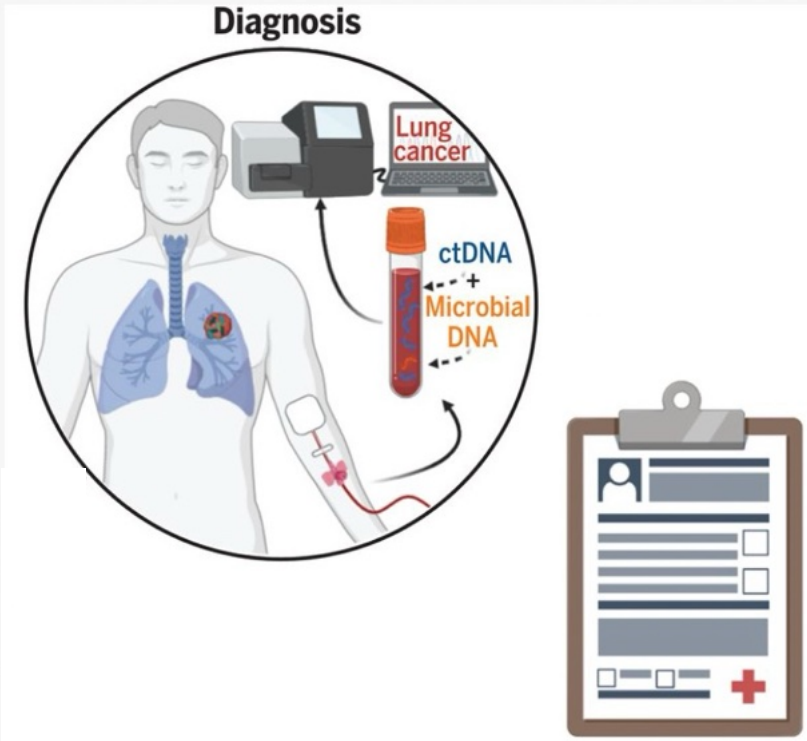
Process samples together, sequence all samples together

Article

Microbiome analyses of blood and tissues suggest cancer diagnostic approach

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

Poore et al. 2020; Spich-Poore et al. 2021; Gihawi et al. 2023



Opportunities for microbes to affect cancer care.

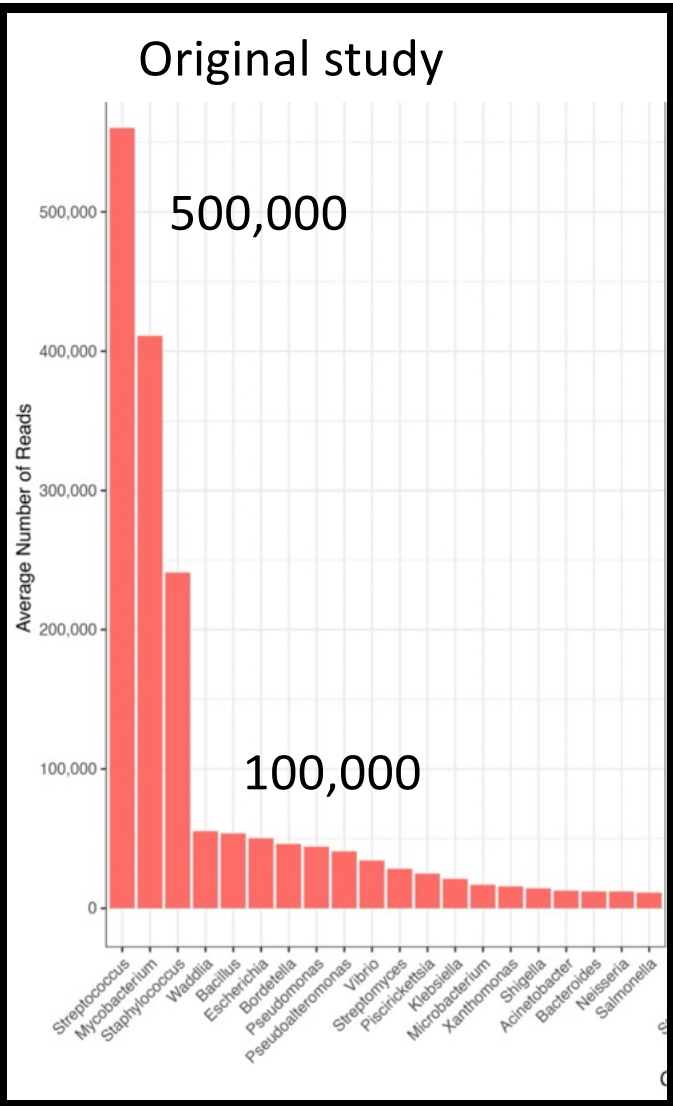
Reported findings:

- 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 likely to be invalid.

Gihawi et al. 2023 for text above

Over-counts of bacteria were due to human reads that erroneously matched bacteria

A huge artifact arose from omitting the human genome from the analysis database (Kraken)

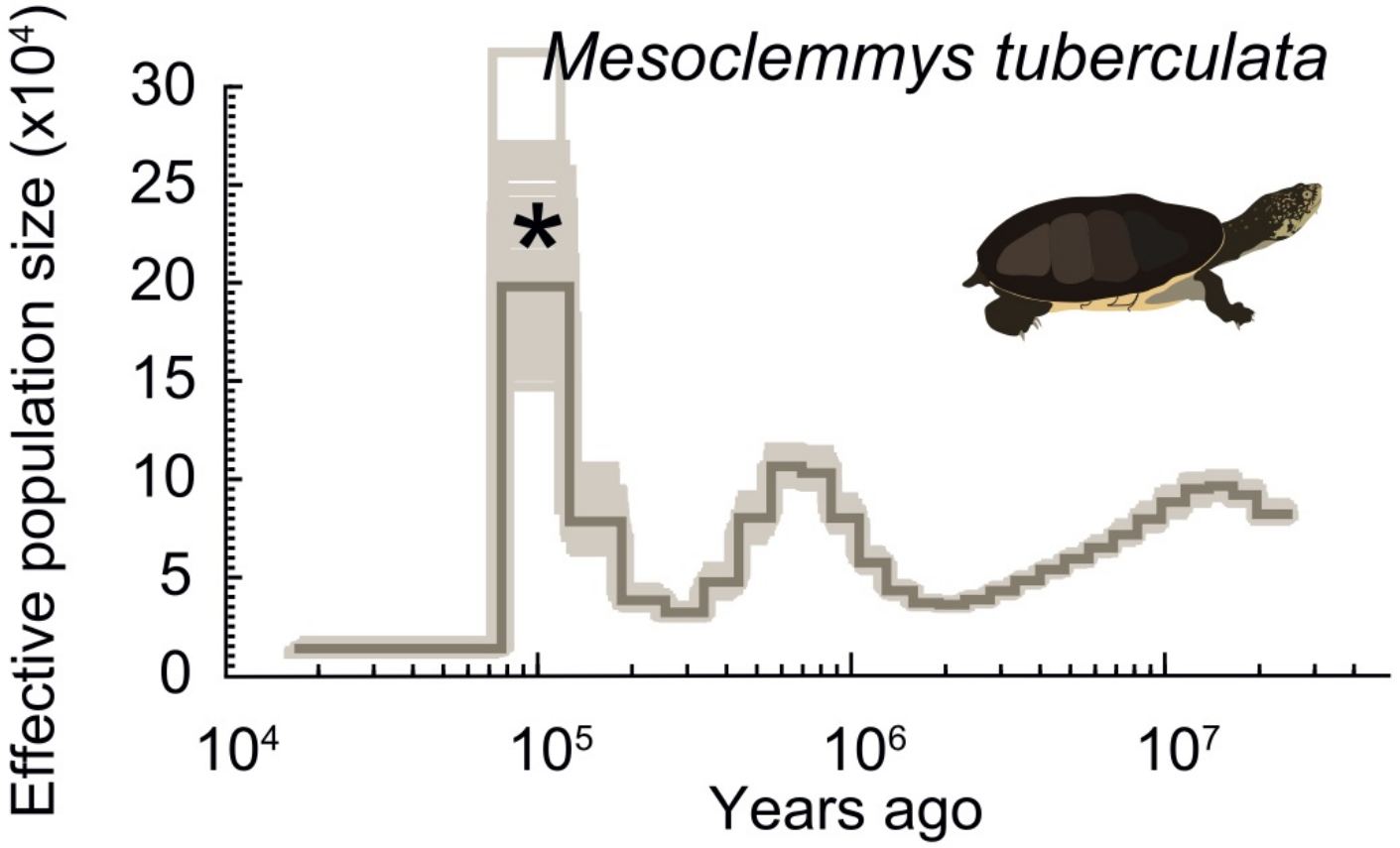


Gihawi et al. 2023 for text above

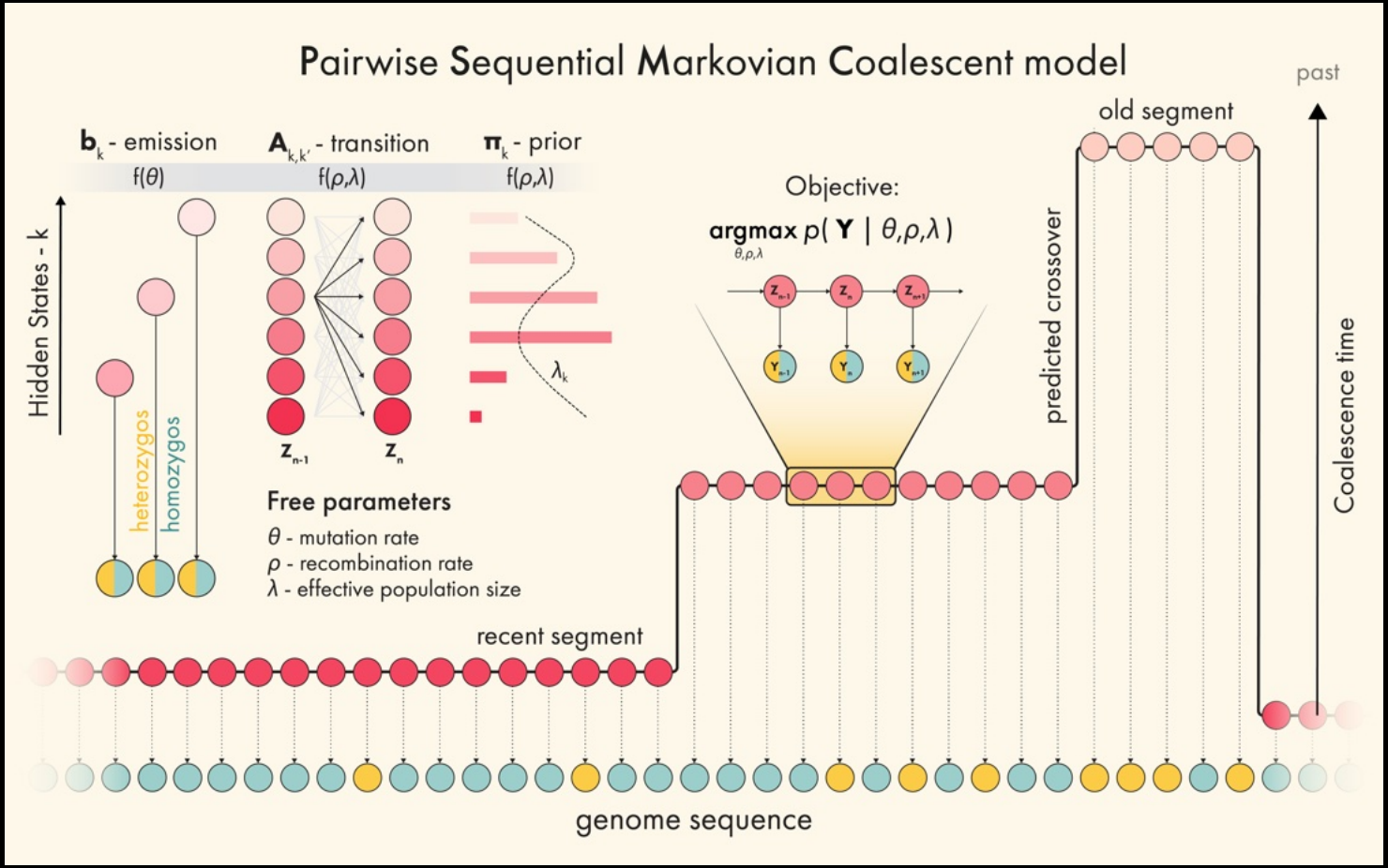
- Published 11 March 2020, retracted on 26 June 2024
- 4 years is actually fast, due largely to the open access to data & methods
- This represents actual progress in the genomics field

What type of analysis is this?

Anyone planning on running this for their species?



Many different methods using individual genomes for Ne/time



RESEARCH ARTICLE | HUMAN EVOLUTION

Genomic inference of a severe human bottleneck during the Early to Middle Pleistocene transition

WANGJIE HU, ZIQIAN HAO, PENGYUAN DU, FABIO DI VINCENZO, GIORGIO MANZI, JIALONG CUI, YUN-XIN FU, YI-HSUAN PAN, AND HAIPENG LI Authors Info & Affiliations

SCIENCE • 31 Aug 2023 • Vol 381, Issue 6661 • pp. 979-984 • DOI: 10.1126/science.abq7487

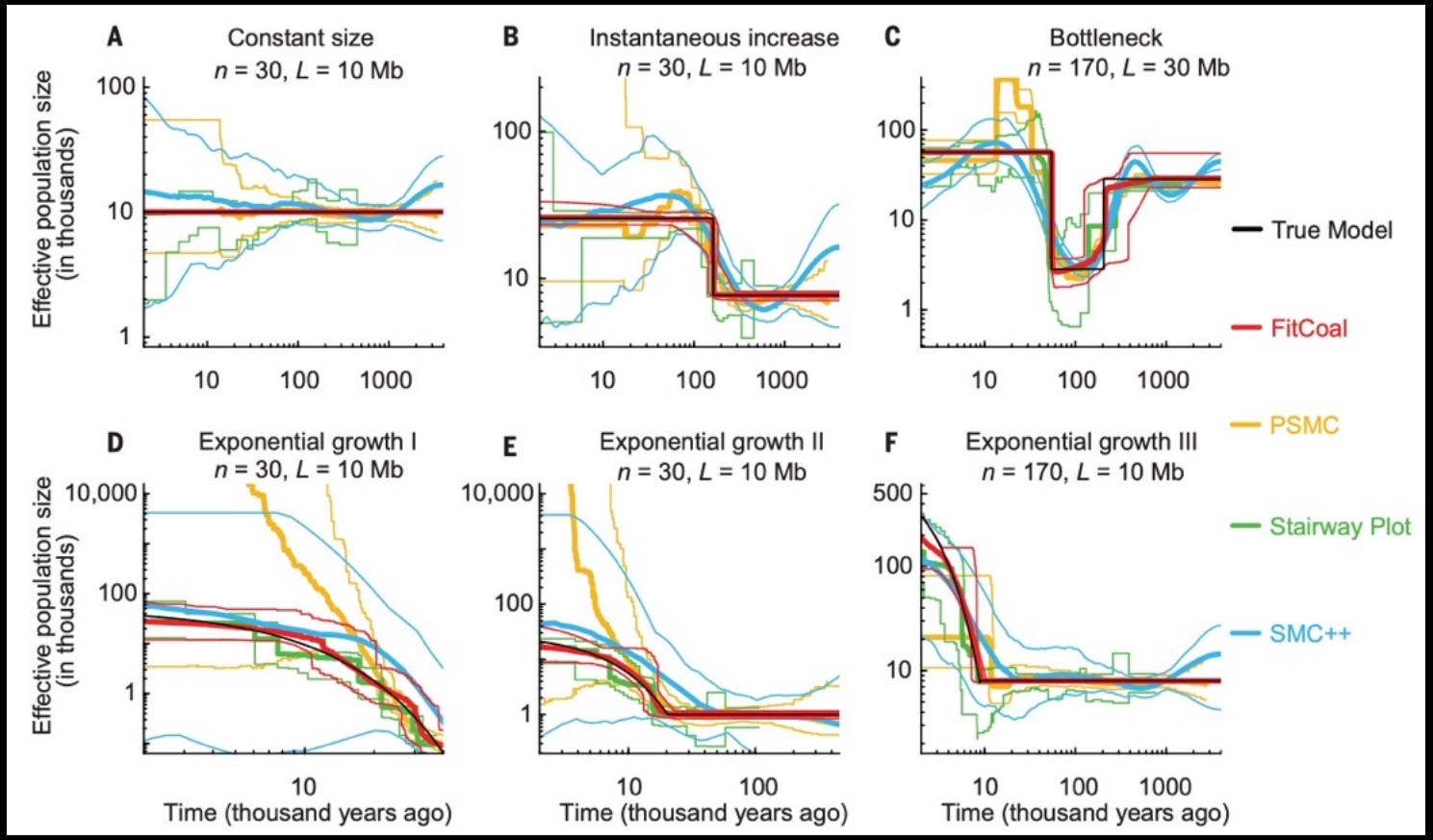
Science

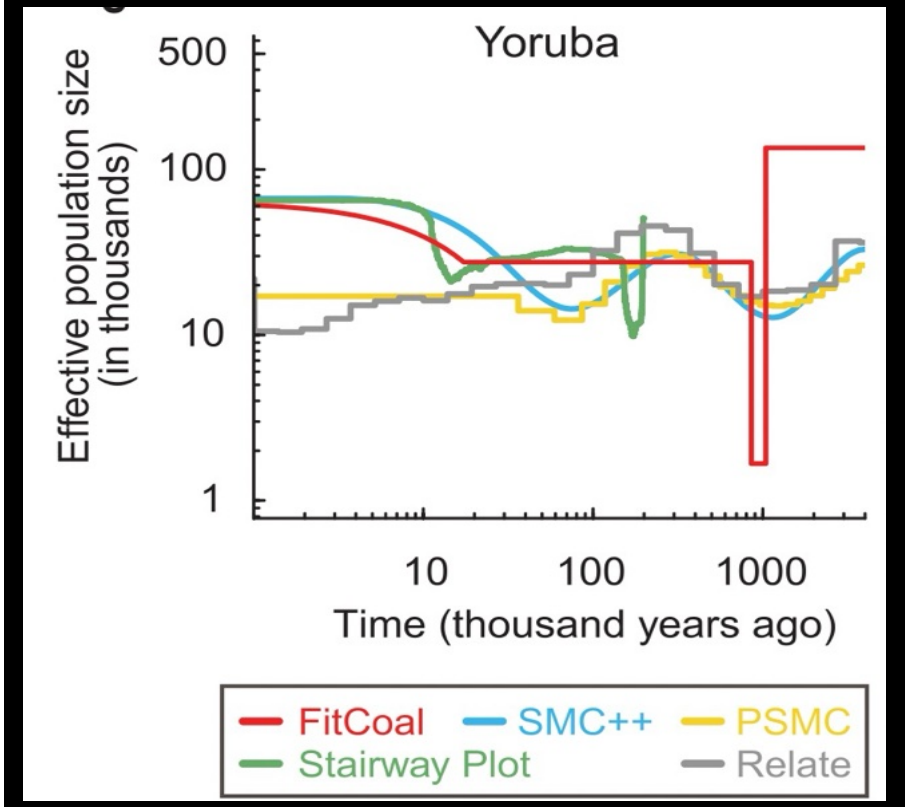
Current Issue First release papers Archive

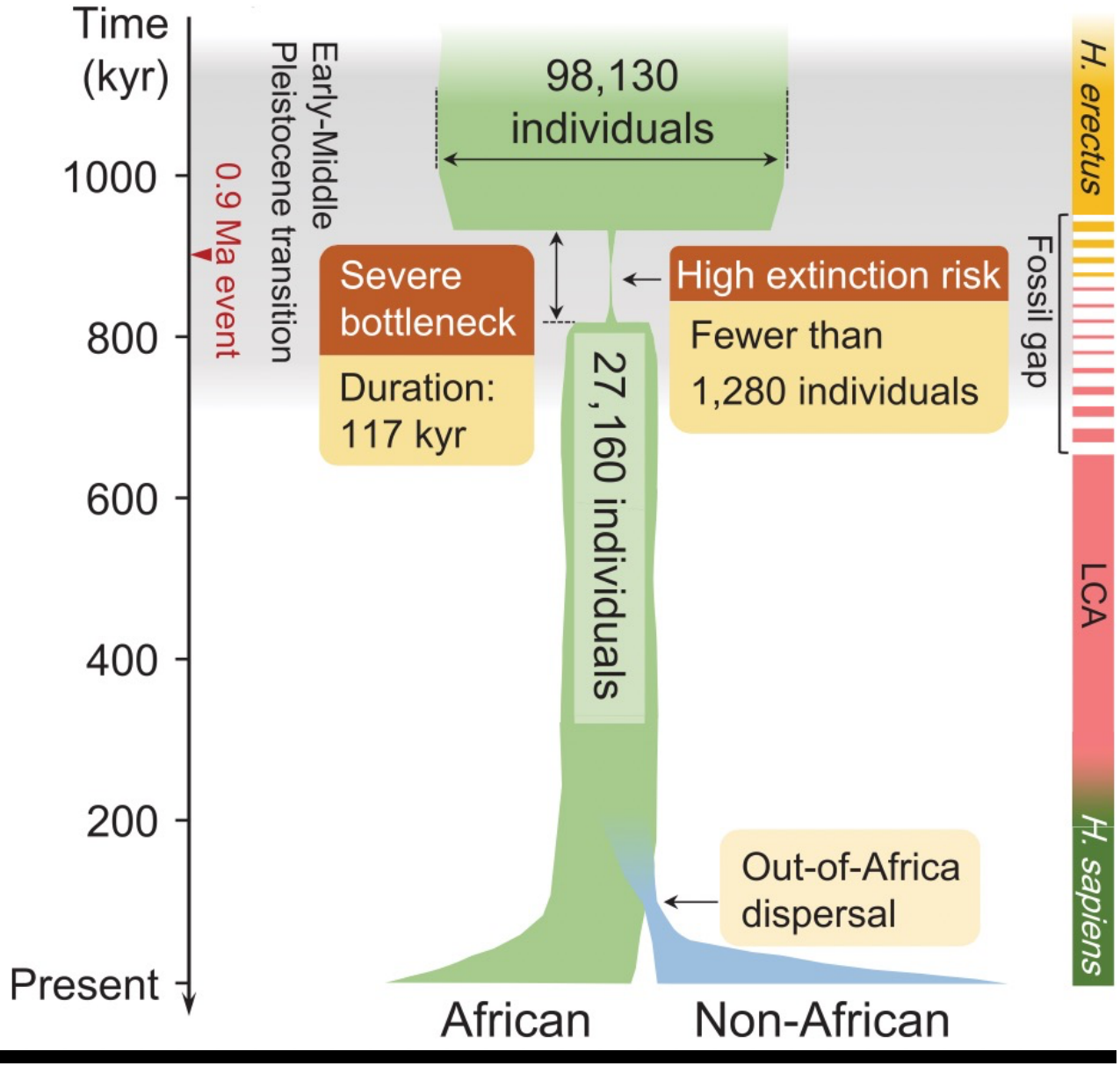
Did our ancestors nearly die out?

Genetic analyses suggest an ancient human population crash 900,000 years ago

The new method appeared to be a major advance in inferring historical population dynamics from genomes




Yoruba
Effective population size (in thousands)
500
100
10
1
10
100
1000
Time (thousand years ago)
FitCoal
SMC++
PSMC
Stairway Plot
Relate


Time (kyr)
1000
800
600
400
200
Present
Early-Middle Pleistocene transition
0.9 Ma event
98,130 individuals
Severe bottleneck
Duration: 117 kyr
High extinction risk
Fewer than 1,280 individuals
27,160 individuals
Out-of-Africa dispersal
African
Non-African
H. erectus
Fossil gap
LCA
H. sapiens

Nature
https://www.nature.com › nature podcast

Our ancestors lost nearly 99% of their population, ...

Sep 6, 2023 — Around 900,000 years ago the ancestors of modern **humans** were pushed to the brink of extinction, according to new research. Genetic studies ...

The New York Times
https://www.nytimes.com › human-survival-bottleneck

Humanity's Ancestors Nearly Died Out, Genetic Study ...

Sep 4, 2023 — Researchers in China have found evidence suggesting that 930,000 years ago, the ancestors of modern **humans** suffered a massive population crash.

EL PAÍS English
https://english.elpais.com › Science & Tech

Only 1200 people left: The moment humanity almost went ...

Aug 31, 2023

Smithsonian Magazine
https://www.smithsonianmag.com › science-nature › ge...

Our Human Ancestors Very Nearly Went Extinct ...

Aug 31, 2023 — Based on the study's estimates, some 98.7 percent of our human ancestors were wiped **out**. “The estimated population size for our ancestral ...

JOURNAL ARTICLE FEATURED

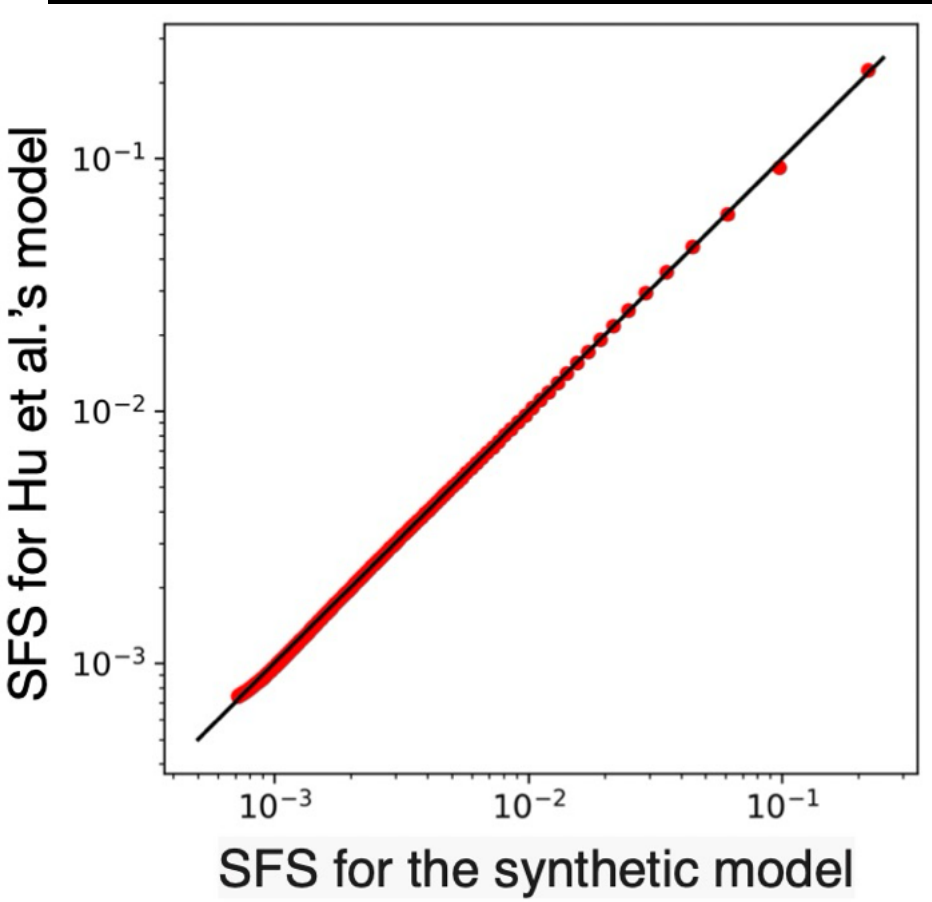
A previously reported bottleneck in human ancestry 900 kya is likely a statistical artifact

Get access >

Yun Deng , Rasmus Nielsen ✉ , Yun S Song ✉

Genetics, Volume 229, Issue 1, January 2025, iyae192,
https://doi.org/10.1093/genetics/iyae192
Published: 16 December 2024 **Article history** ▾

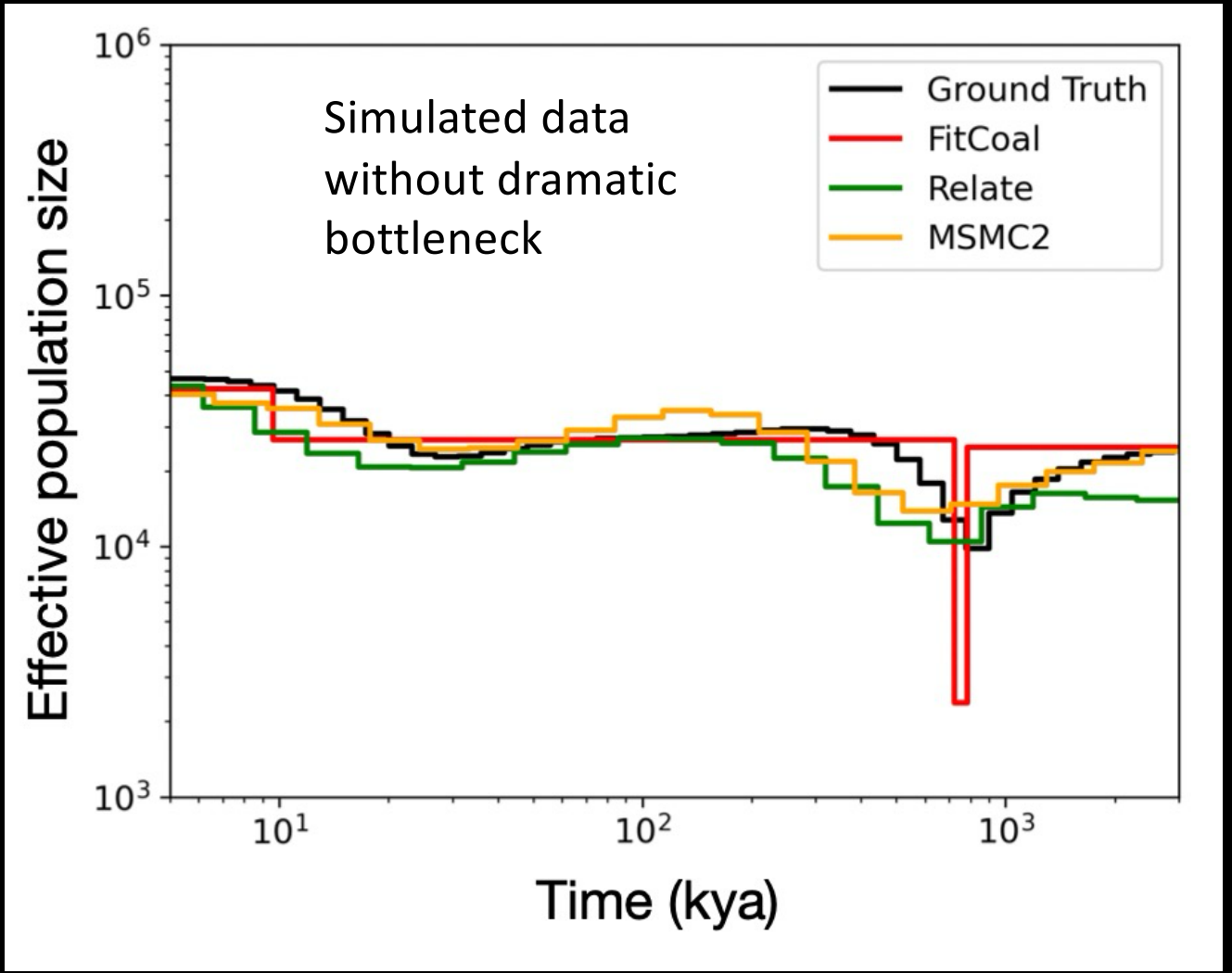
- Simulated data with nearly identical SFS to publication
  - but under demographic scenarios that did experience dramatic bottleneck
- Models with or without the bottleneck can result in very similar expected SFSs



FitCoal artifactually tends to infer a sharp bottleneck when there in fact is none

- Reported bottleneck is likely a statistical artifact

Providing valid statistical measures of uncertainty for inferences of demographic models is an important research challenge in computational population genetics



Frances Arnold
@francesarnold

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

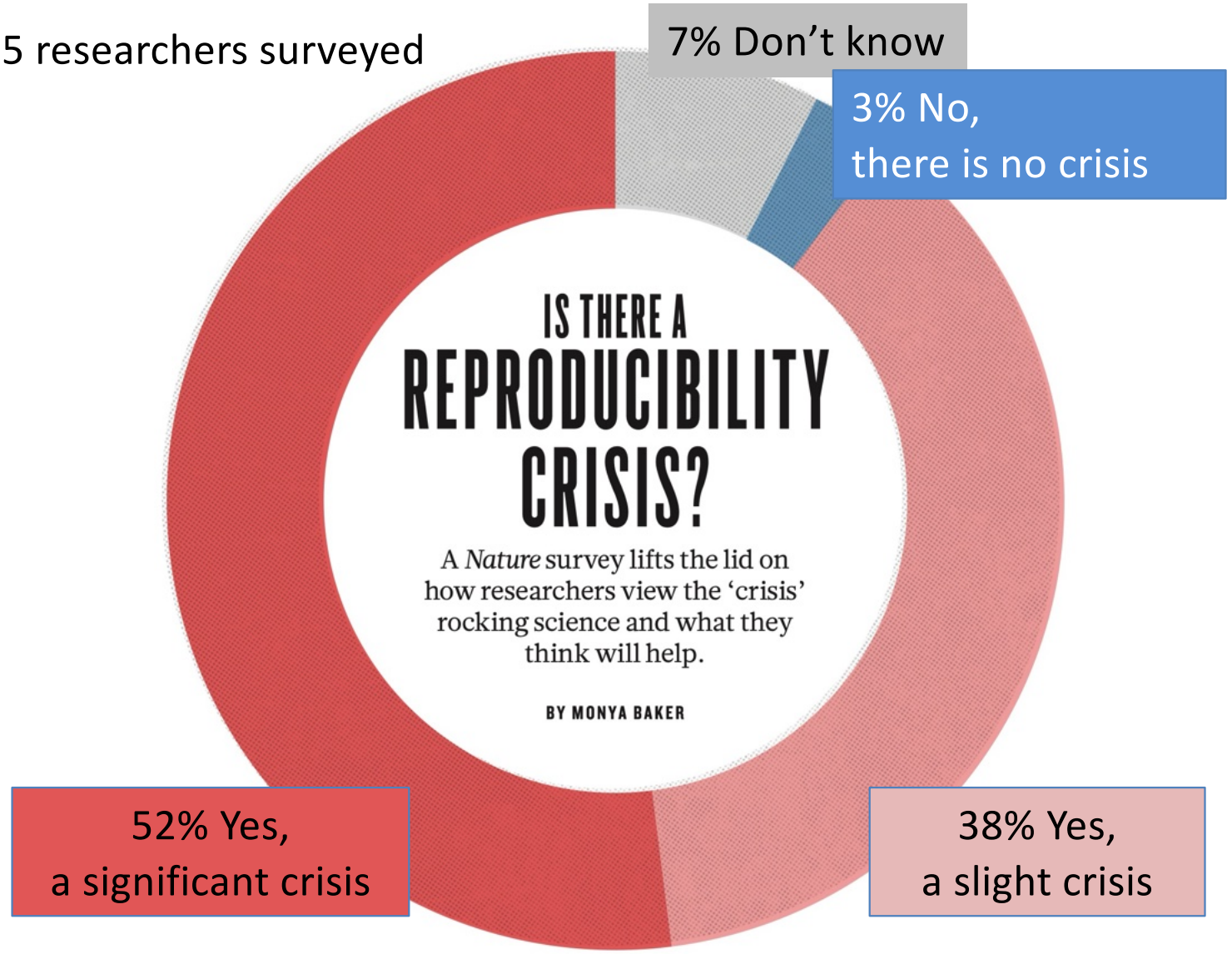
Site-selective en
Enzymes excel at
sites. With approp
science.sciencem

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

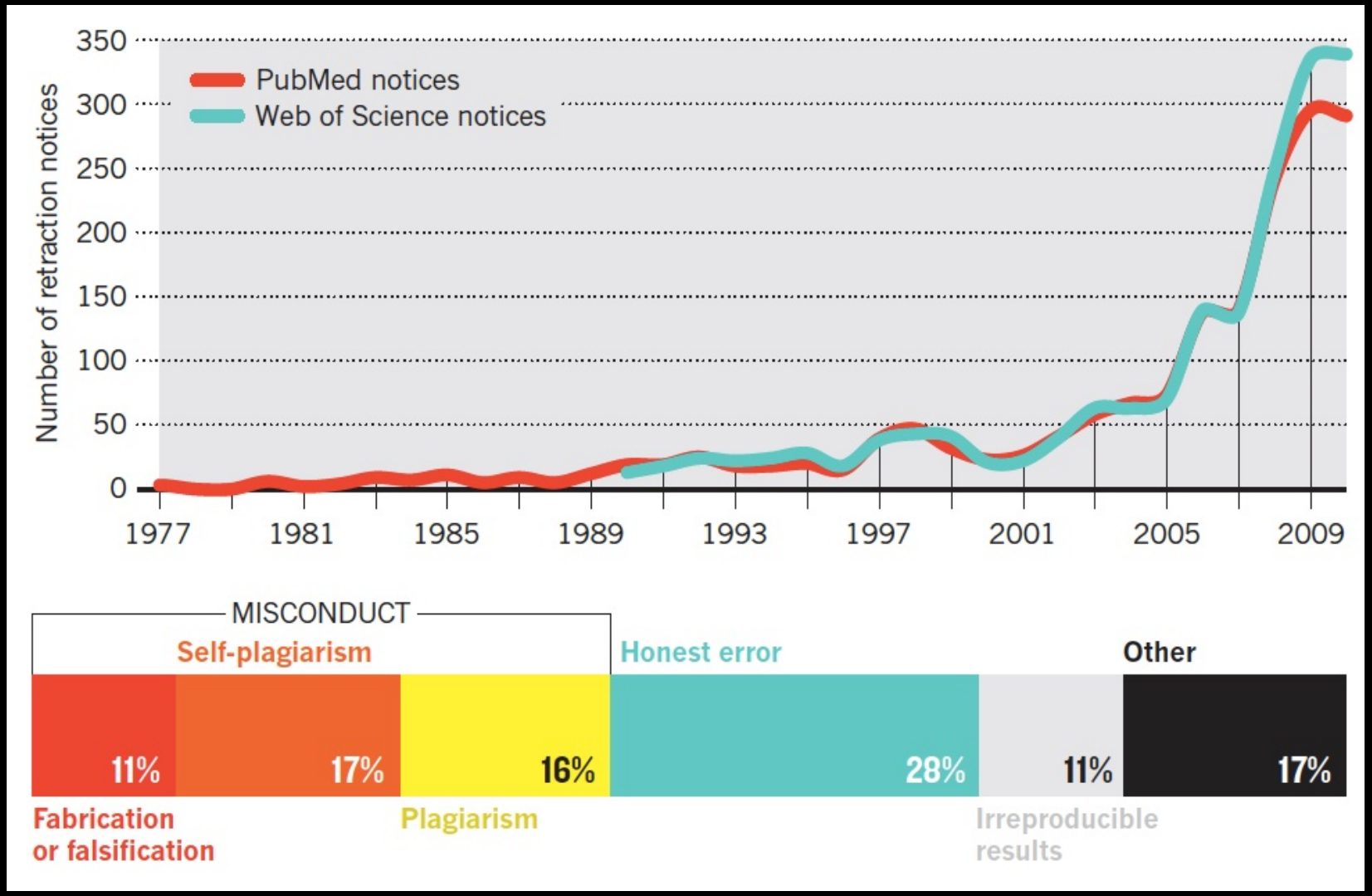
Lynn Kamerlin @kamerlinlab · Jan 2
Replying to @francesarnold
Sorry about the problems, but kudos for doing the right thing, and setting a good example.
1 1 178

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 reproducibility crisis?

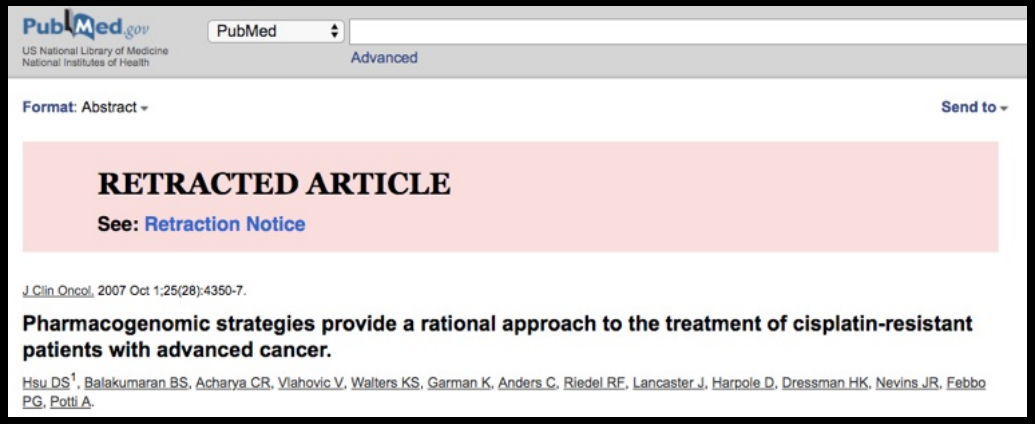


The trouble with retractions: Nature News 2011

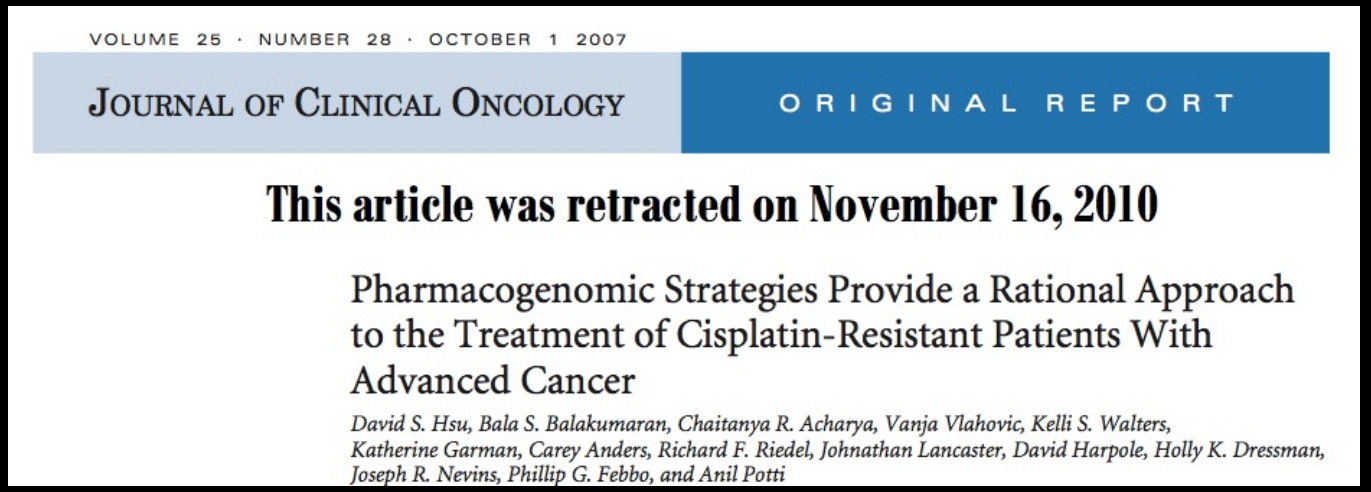
- Keeps community updated
- Help kill zombie papers that keep getting cited when they should not
- Starting to get integrated into websites and ref managers

- Be sure you are never keeping zombies alive

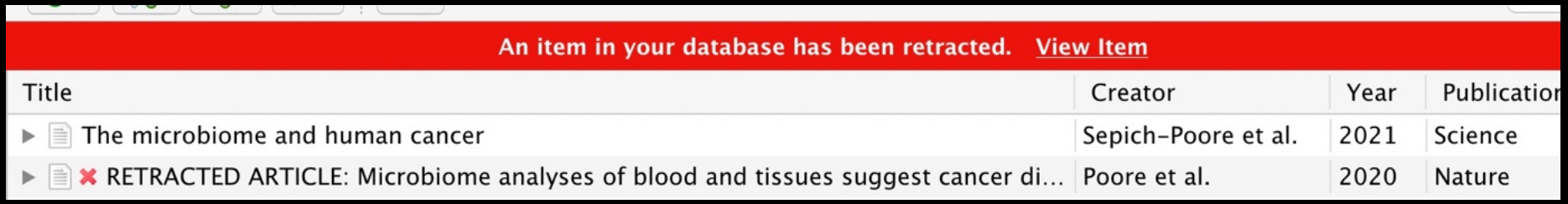
PubMed



Journal



Zotero

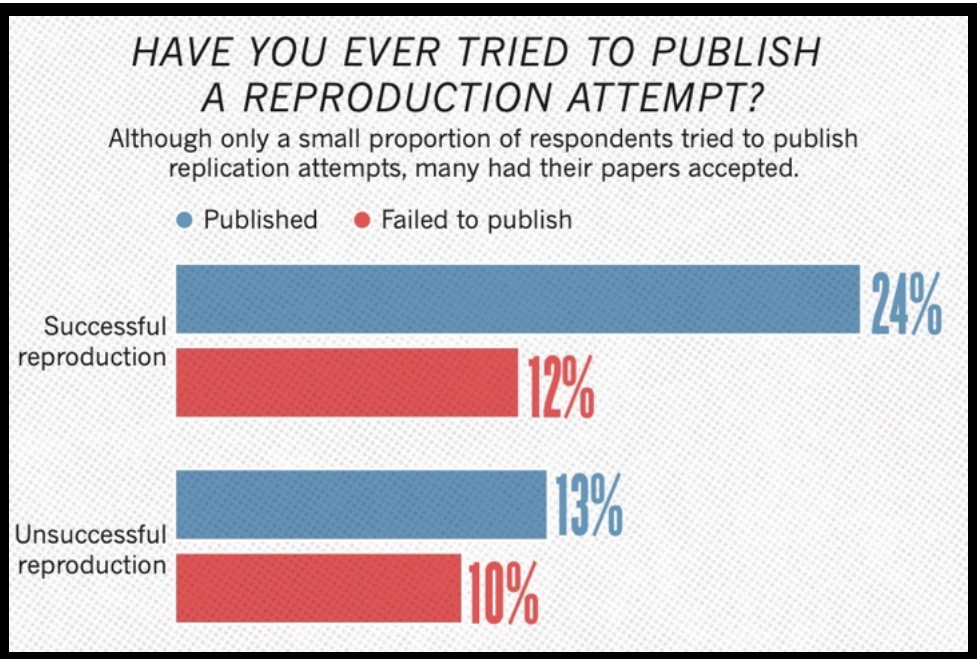


How can we improve reproducible findings?

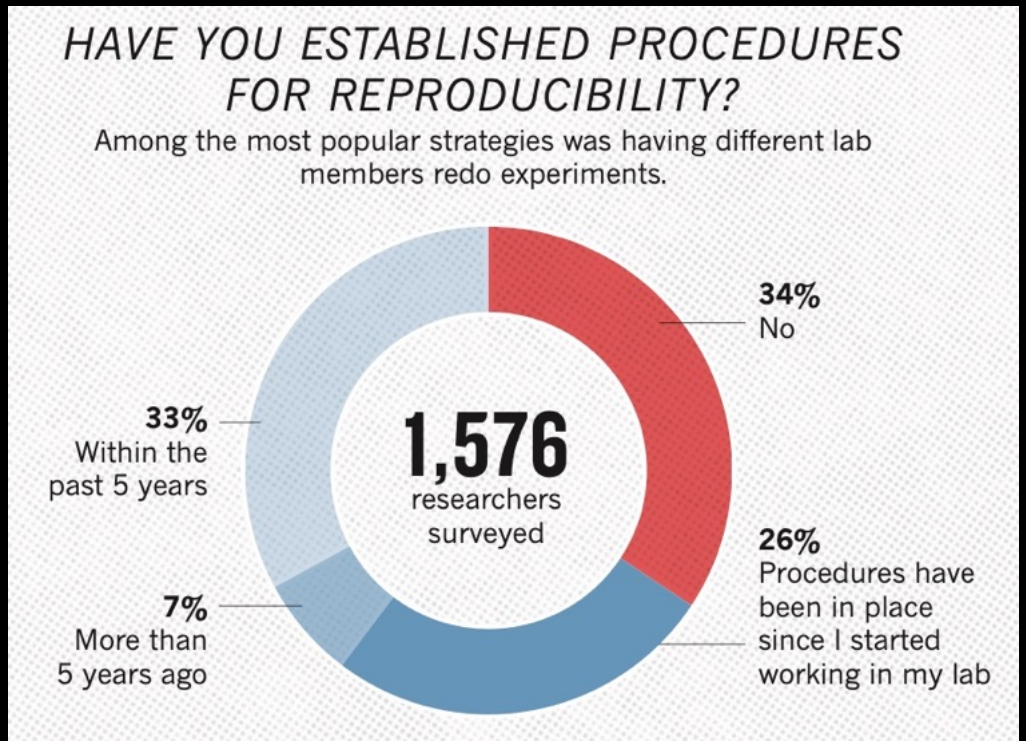
Work better as a community, check each others code and post our code

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



Of the few that 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, non-repeatable studies out there …

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

Thus .. calibrate your expectations, approaches, be careful & stay humble

What is your personal error rate?

I assume mine is 12%

therefore I perform many sanity & error checks to catch the errors I WILL MAKE, so my effective rate is much lower

"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

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

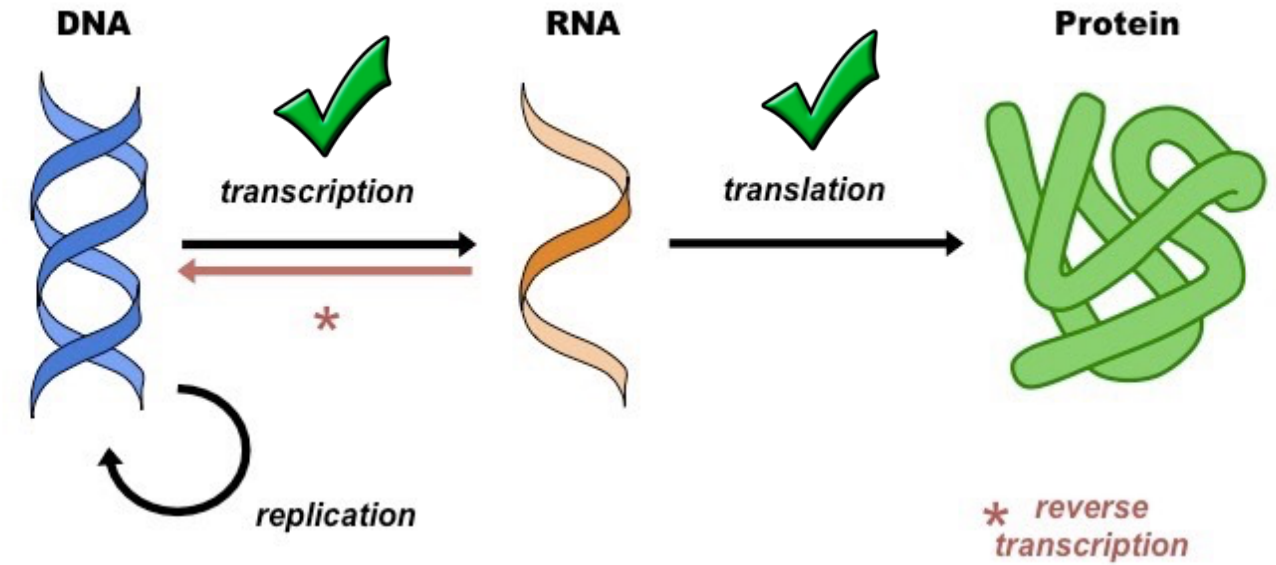
the I knew-it-all-along effect

the tendency,
after an event has occurred,
to see the event as having been predictable,
despite there having been no objective basis
for predicting it.

Three Levels of Hindsight Bias

I KNEW
that would happen

The central dogma

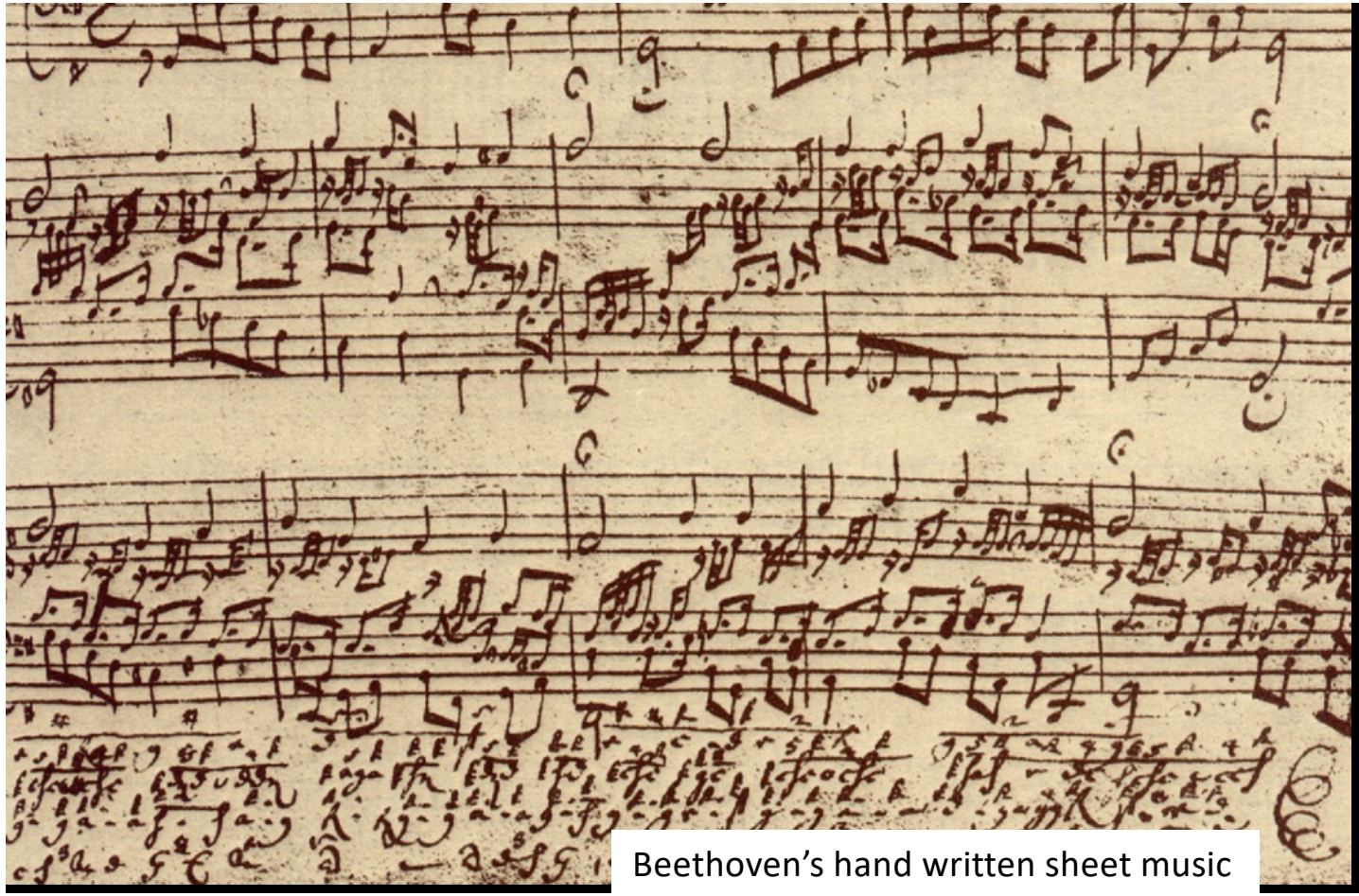


<u>But, can we, in a novel species :</u>

- 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"



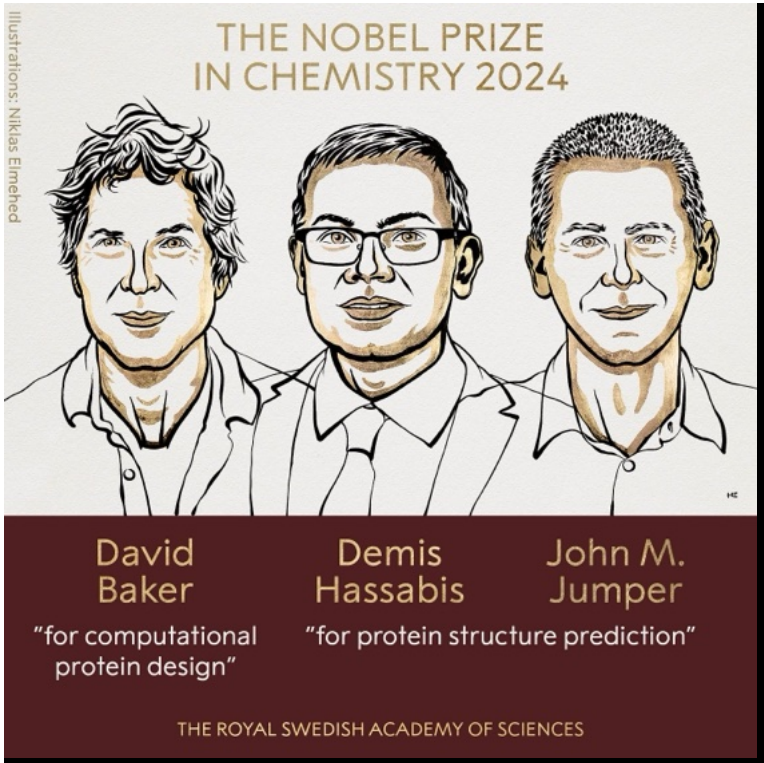
Beethoven's hand written sheet music

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

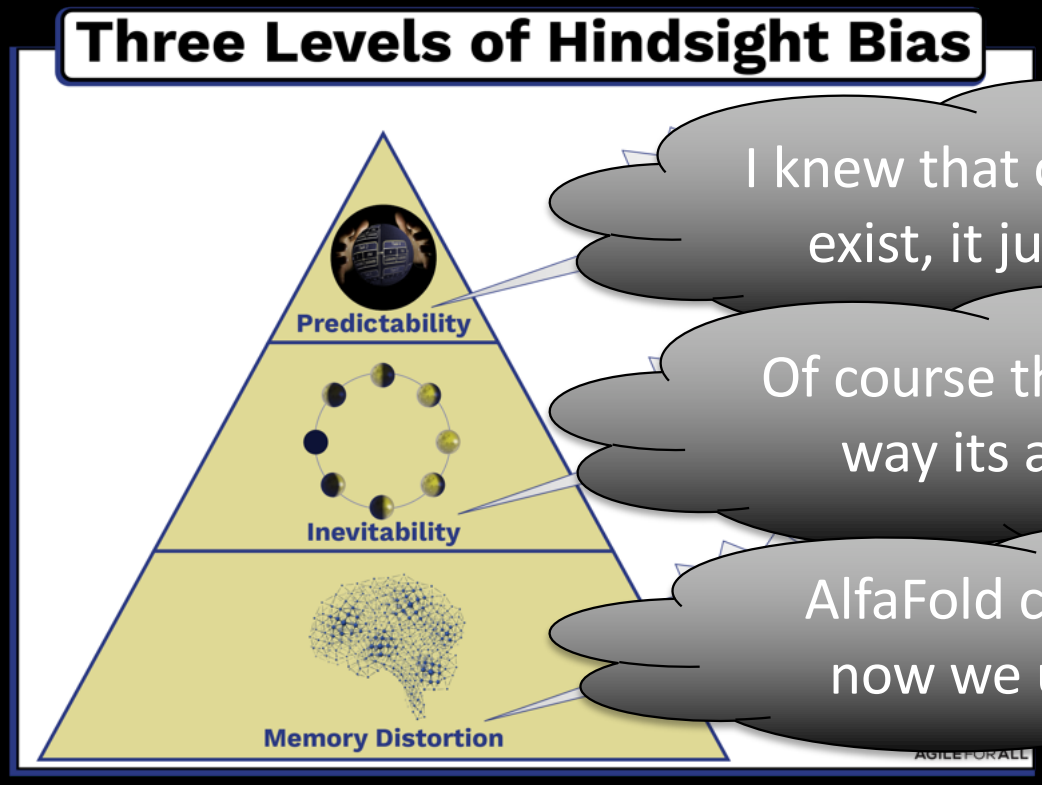
"If we can predict how proteins fold without understanding how they do it, are we even legitimately doing science anymore, or is it something different?"

"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

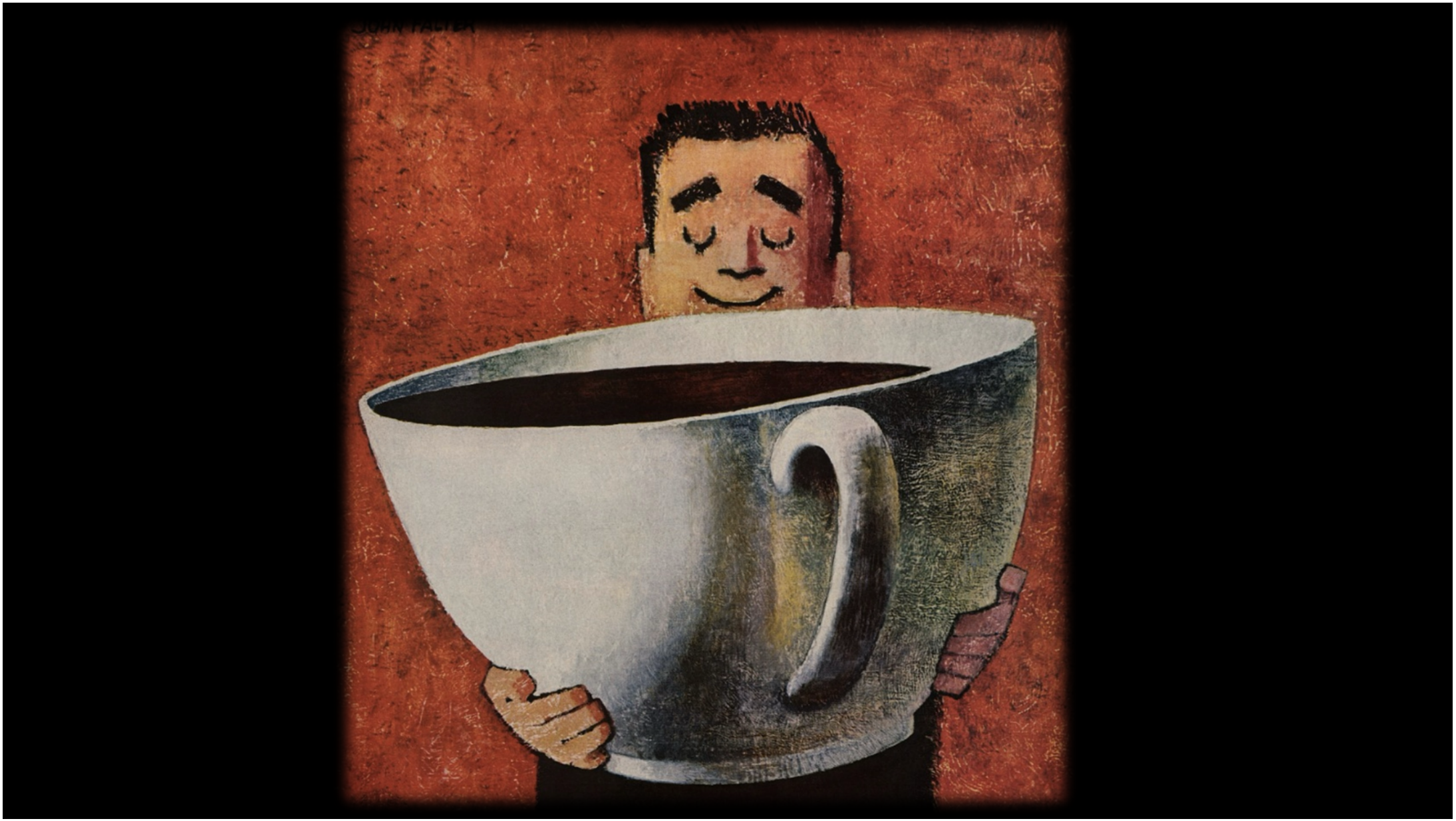


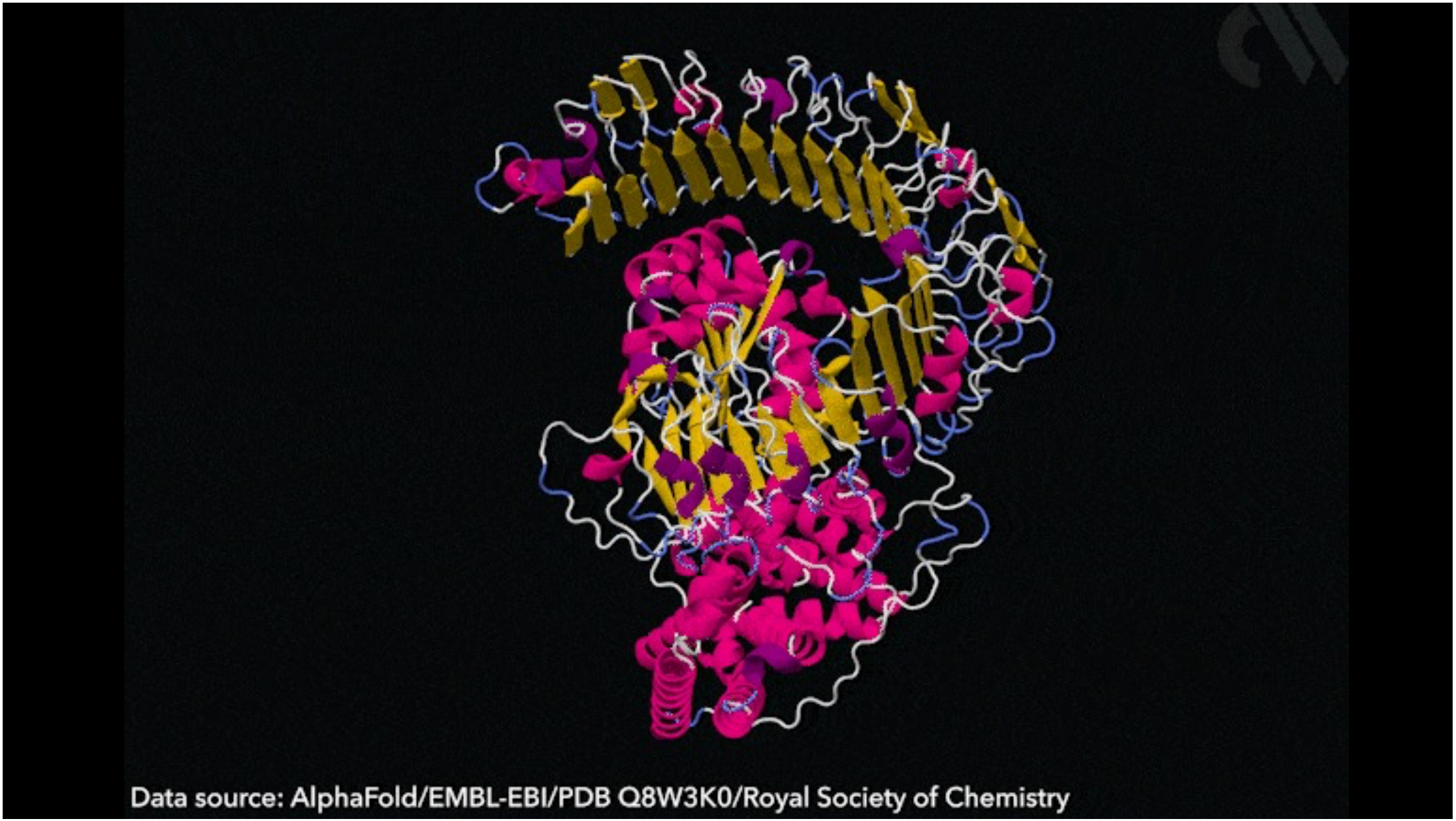
I knew that correlation had to exist, it just makes sense

Of course this gene works the way its annotation says

AlfaFold can predict structures, now we understand enzymes


sallyedelsteincollage.com




Data source: AlphaFold/EMBL-EBI/PDB Q8W3K0/Royal Society of Chemistry

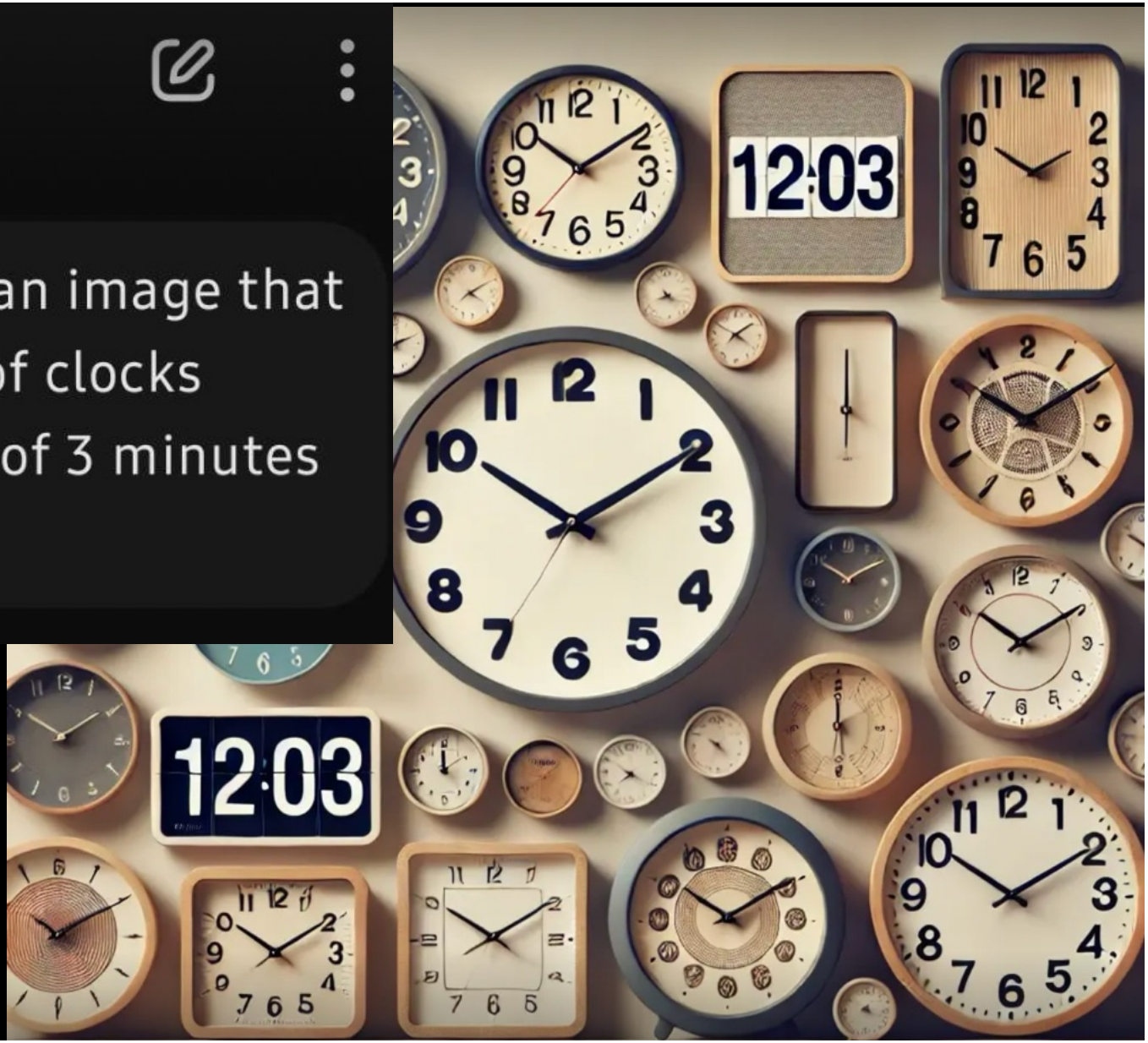
How many have used AI to help with their coding?

Where does the training set for these answers come from?

Does that matter?

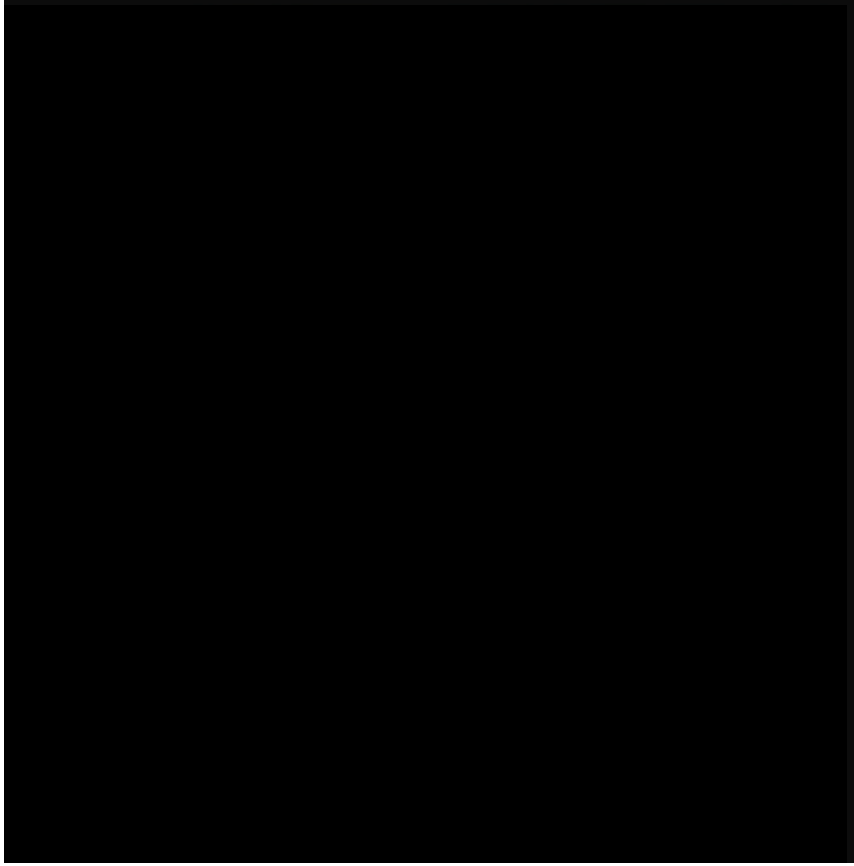
ChatGPT 4o

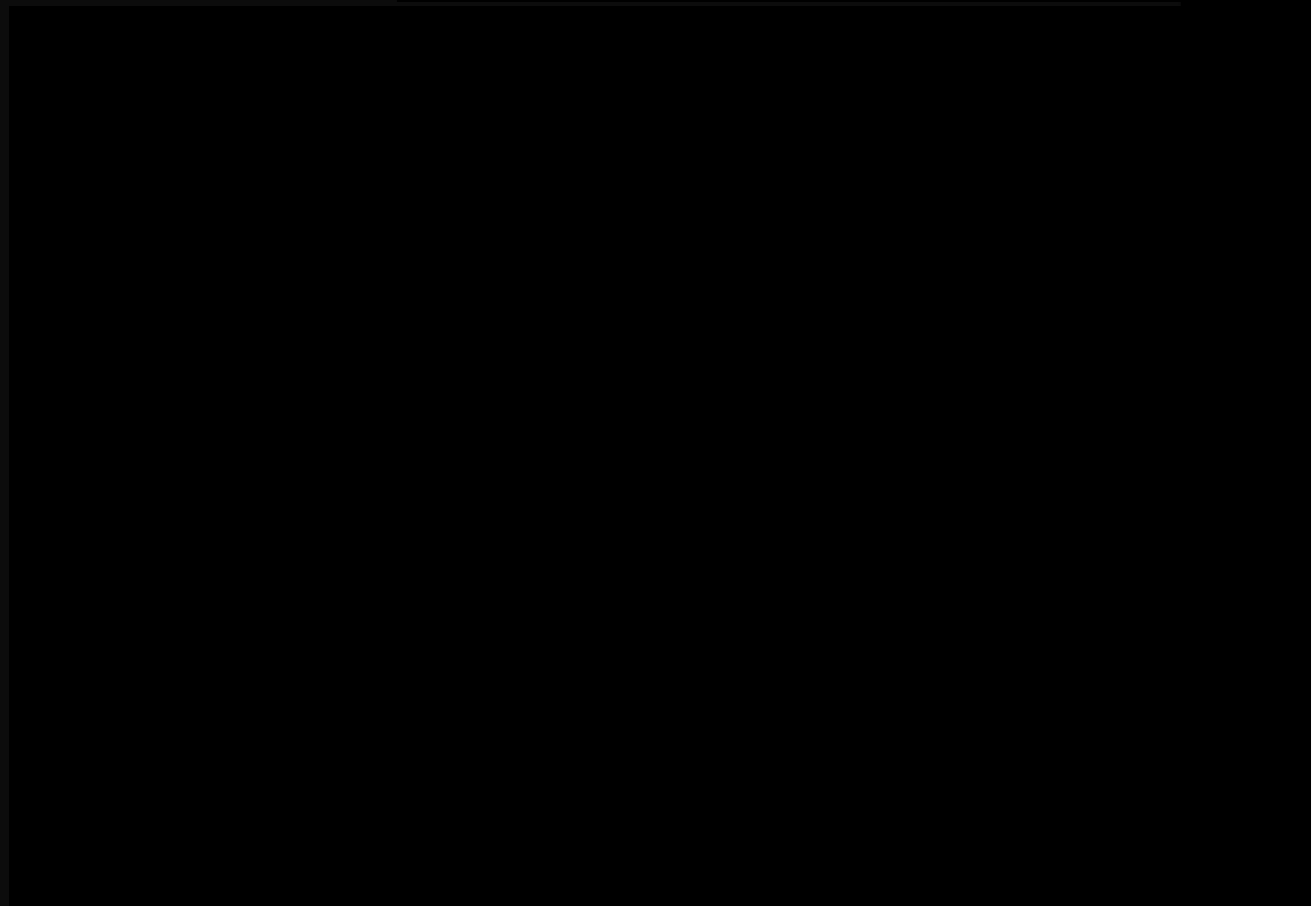
Can you draw me an image that contains a group of clocks showing the time of 3 minutes past 12



ChatGPT

Draw me a picture of a person writing with their left hand.






I said left hand not the right!

Why is AI getting these simple things wrong ..

It's the training set and that has huge implications for most of your work

What species has the largest bioinformatics community?

Human biomedical studies drive the training set that inform all of your bioinformatic answers from ChapGPT

Use AI, verify the code works, then modify for your species

This is a piece of toast

Non-random patterns are abundant in genome scale data

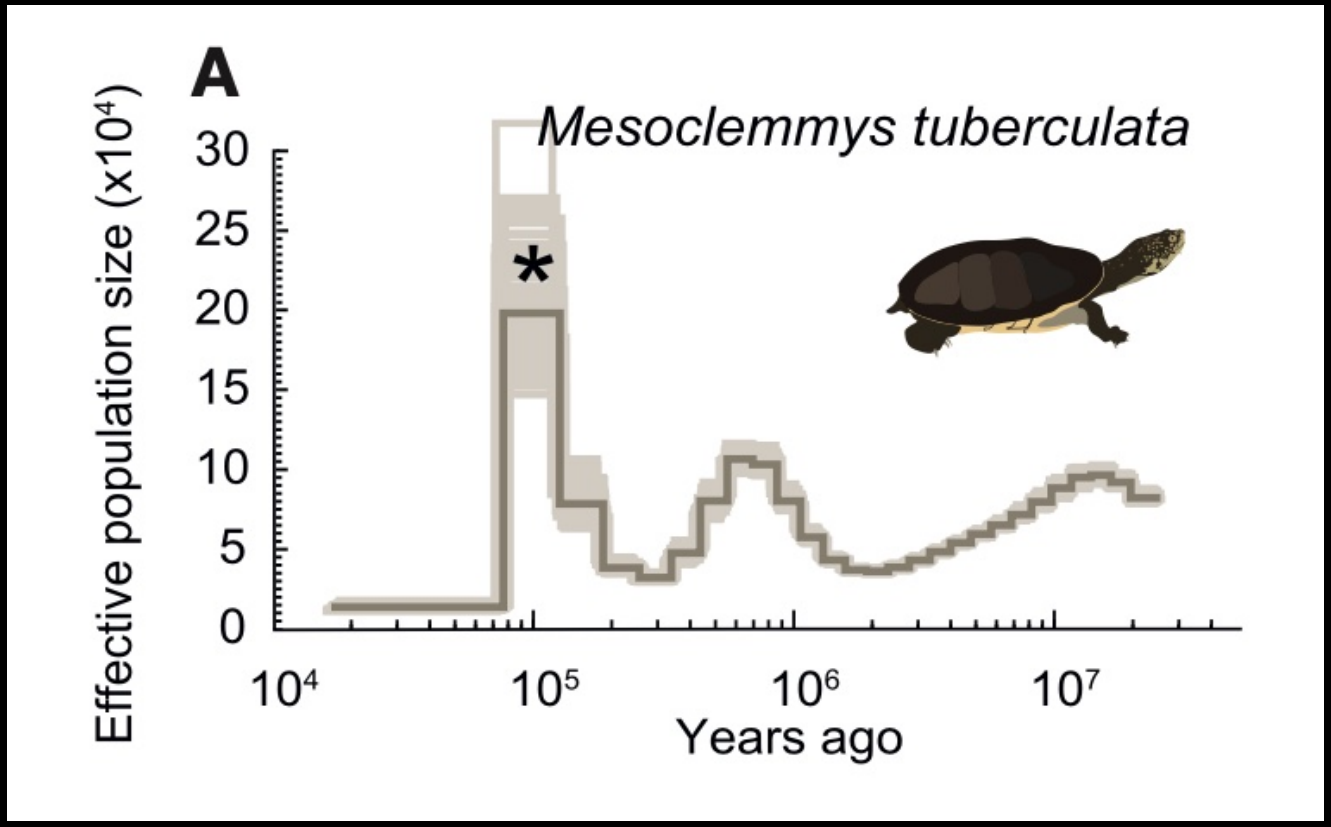
Null models & controls are very difficult to get “right”

Analyses easily generate results that arise from
diverse non-biological causes

So .. how do we avoid Apophenia?

- Double check your data, analysis parameters / settings
  - Plot your data, look at it, does it make sense on 1st principals?
- Genome scale patterns are just hypotheses
  - Explore independent datasets, tools, approaches
    - biological samples, simulations, DNA vs. protein results
  - Manipulation: functional validation via manipulation of genes, pathways
    - Experimental evolution, CRISPR KOs, environmental perturbations

PSMC analysis of single genome for demographic trends

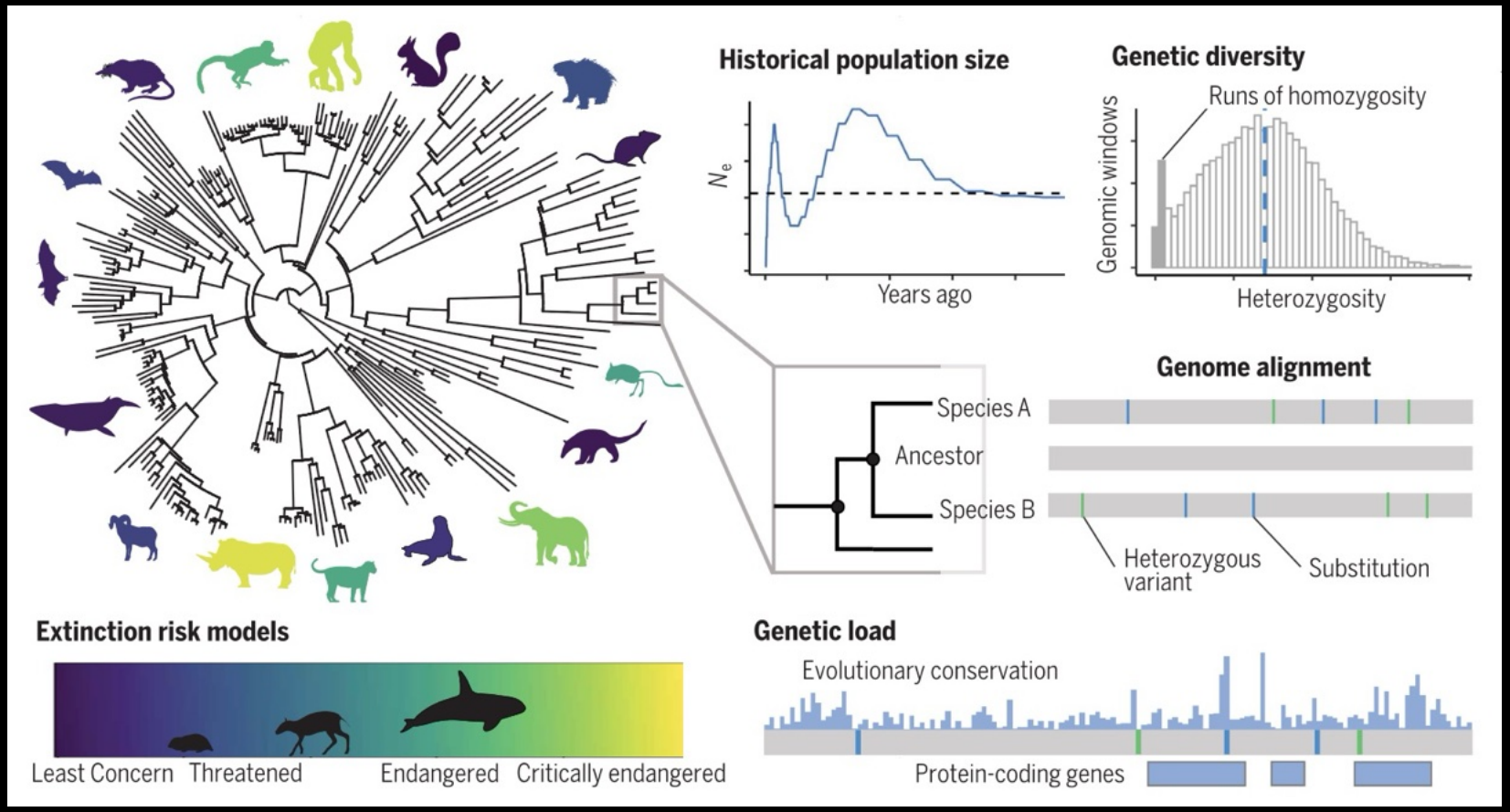


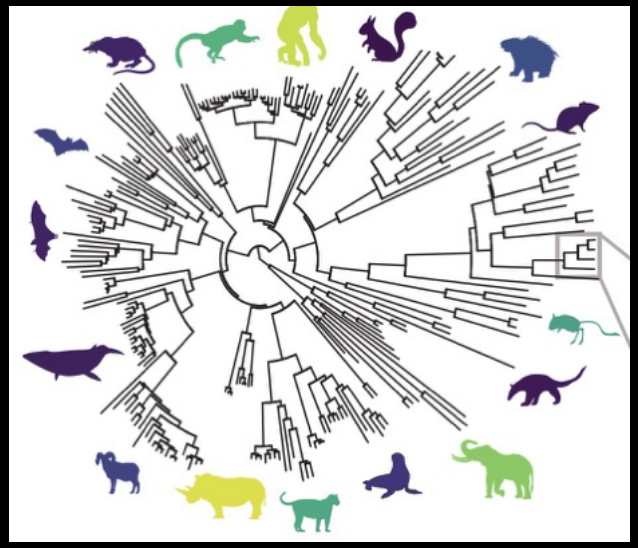
a very common analysis in the literature
basis of many evolutionary stories, conservation concerns

Hilgers et al. 2024 Current Biology

What can you do with 240 mammalian genomes?

Zoonomia

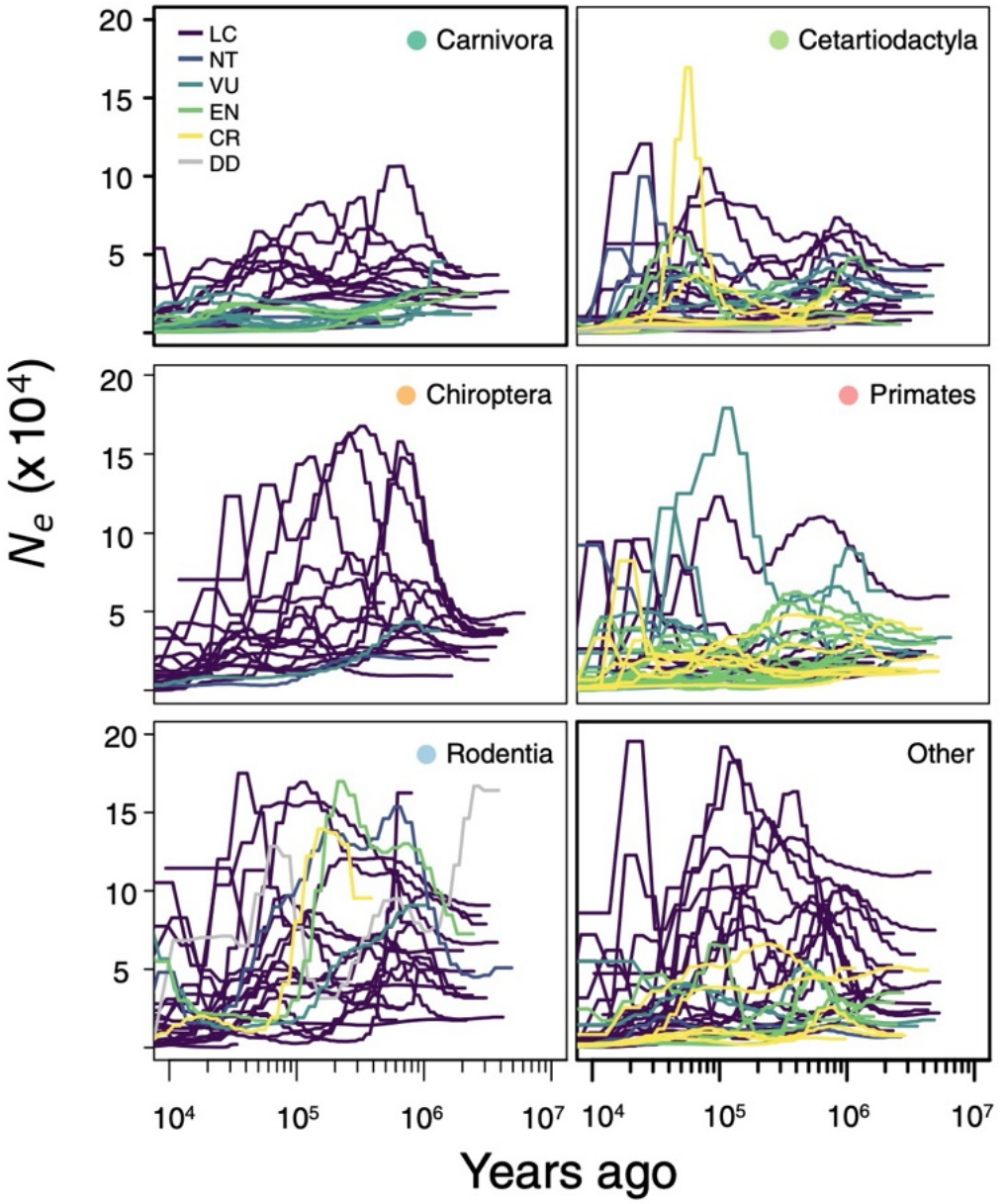




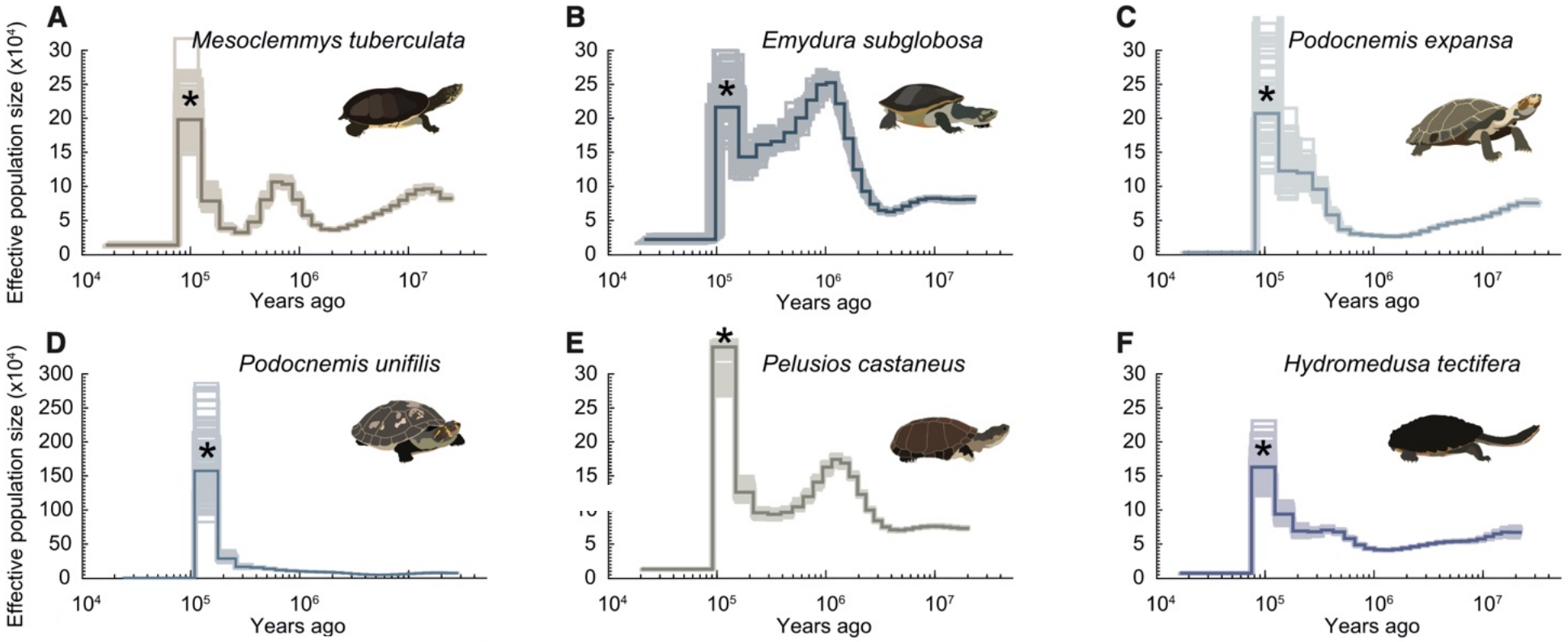
Assessed how historical effective population size (Ne)

impacts heterozygosity

in relation to IUCN conservation status



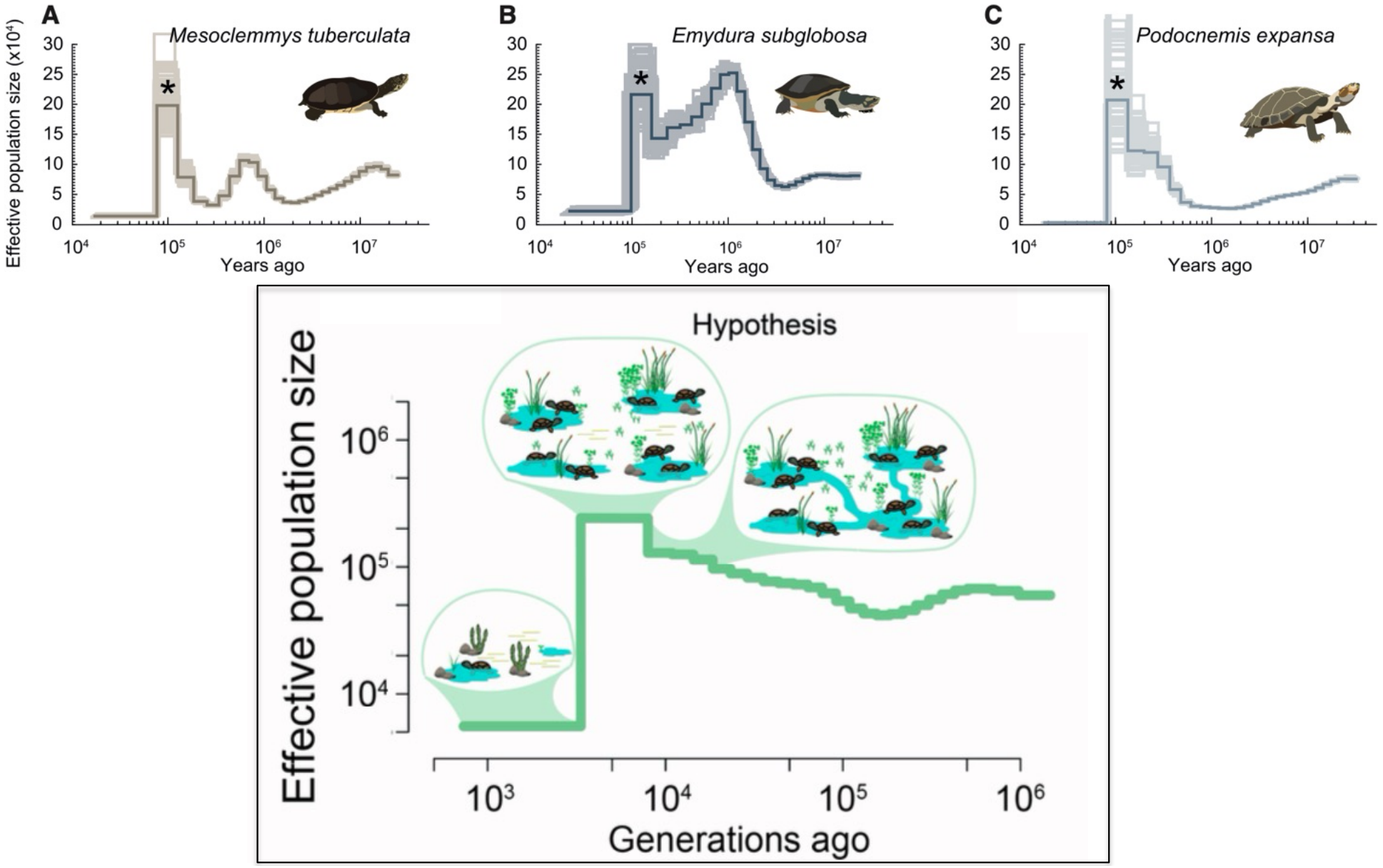
Wilder et al. 2023, Science

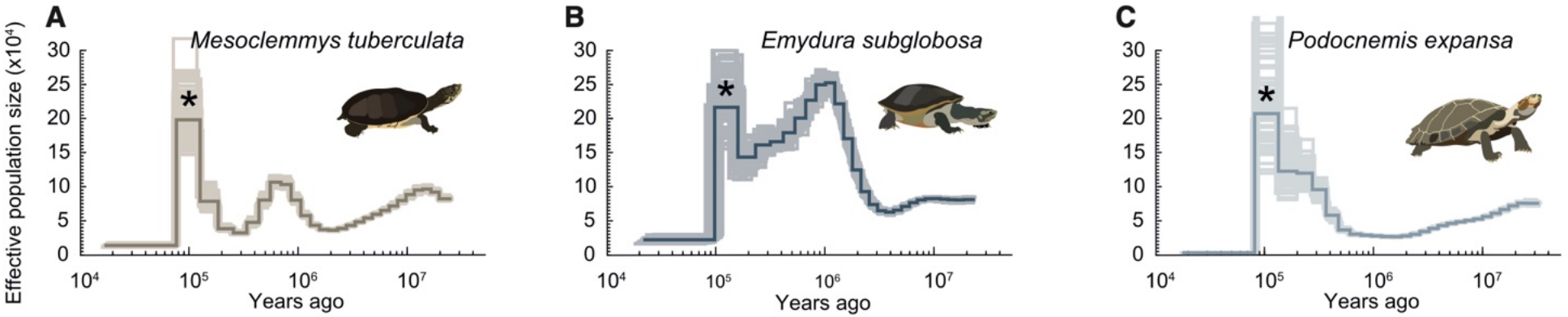


Dramatic peaks followed by even more extreme population collapses

- pattern occurred across
- distantly related turtle species around the globe and its existence
- was consistently supported based on bootstrap replicates

Hilgers et al. 2024 Current Biology


A
Mesoclemmys tuberculata
B
Emydura subglobosa
C
Podocnemis expansa
Effective population size (x10⁴)
Years ago
Hypothesis
Effective population size
Generations ago

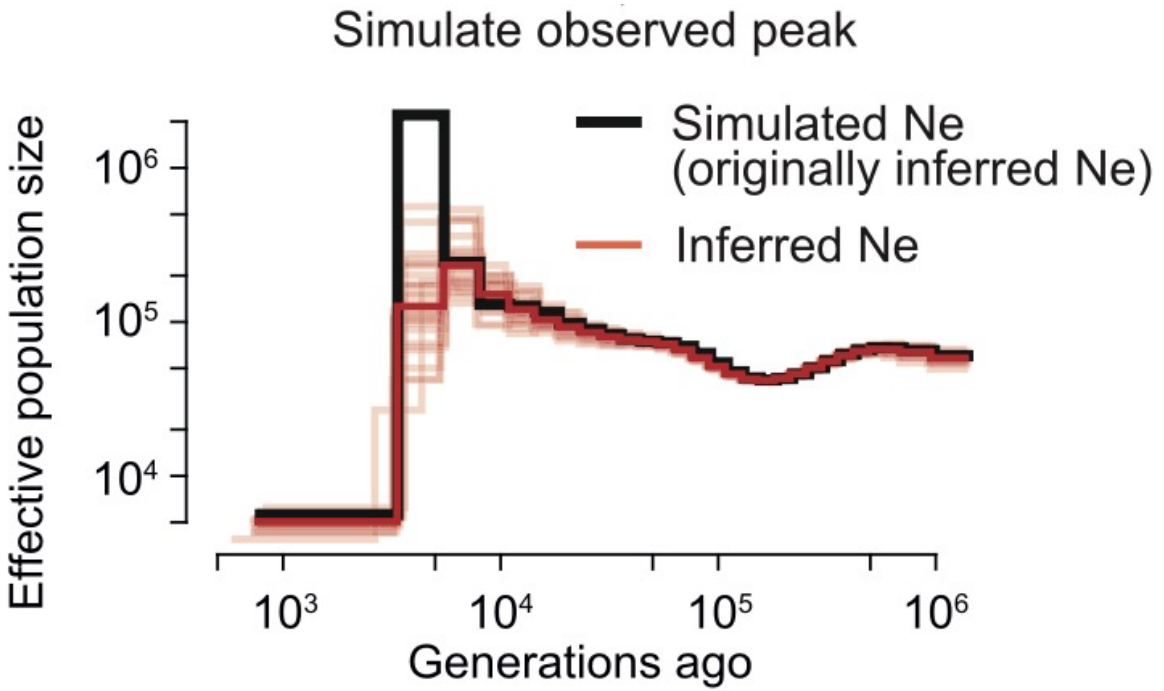


Authors decided to explore if their bioinformatic choices could be generating cross species pattern

- Input
  - used simulations of genomic data
- Tool settings
  - explored range of run settings (default vs. others)

Dataset simulations were used to explore what is being detected

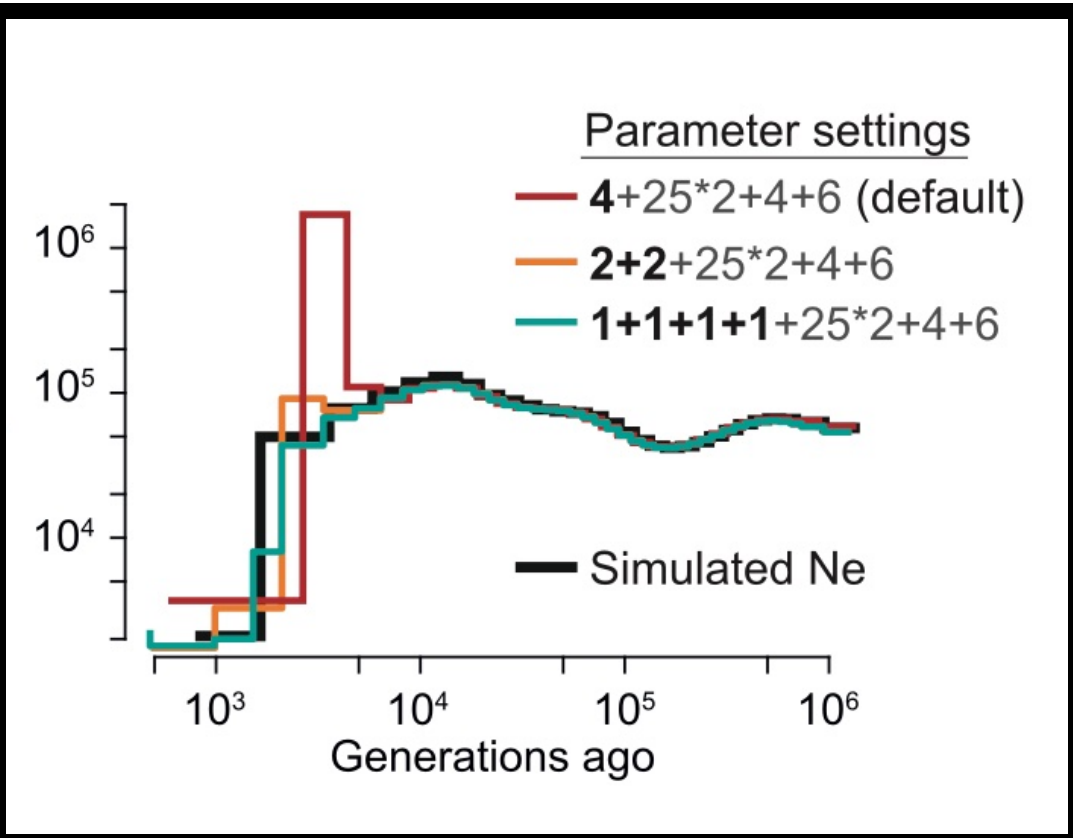
- simulated diploid genomes (n=30) from a population with a population history having pronounced increase / decrease (black)
- Analysis never recovered peak on simulated data, even though it had that exact history (red)



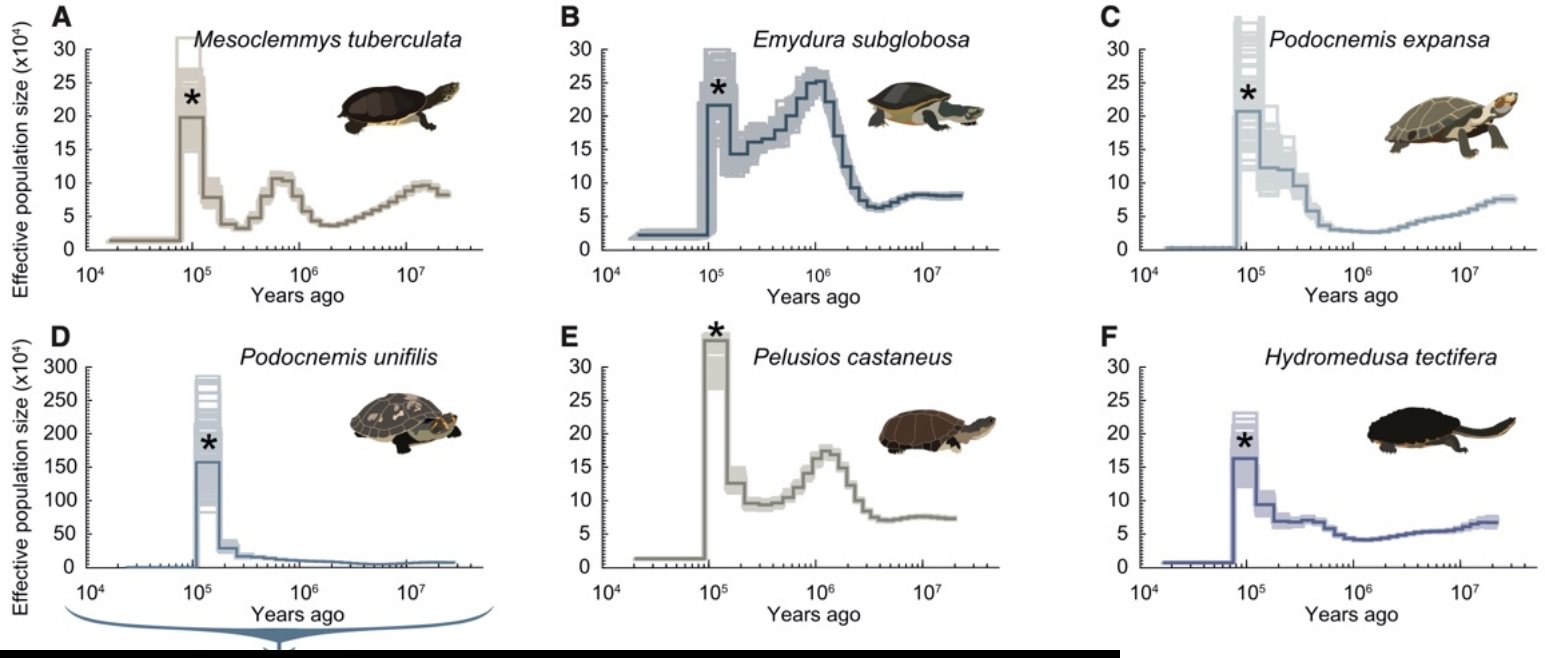
Method not working the way it should …

Explored PSMC run parameters

- default settings optimized for humans
- default setting fixes the first 4 atomic time intervals to first time window.
- infers a single Ne for a large first time window and cannot model population size changes during this time.



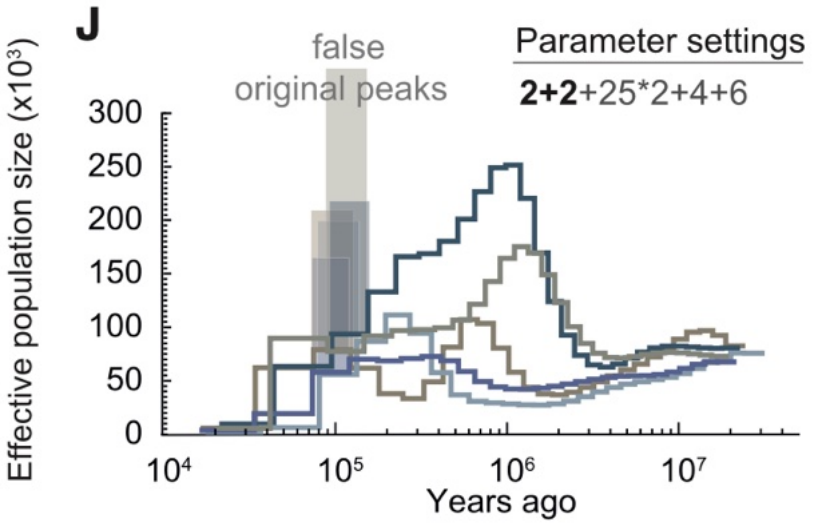
if population declines within the first 4 time intervals, the model likely overcompensates by producing exaggerated Ne estimates in the previous time window



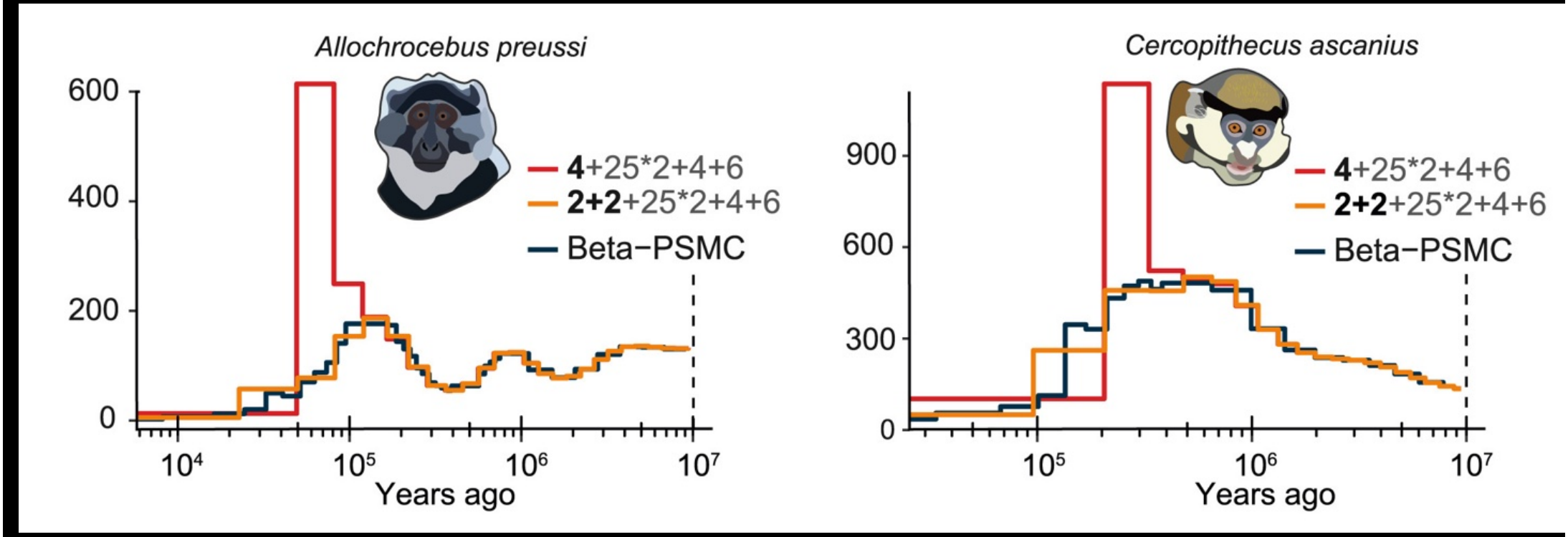
when changing default settings to those that better fit their species

all peaks/crashes go away

see the paper for analysis suggestions



They also re-ran published data



documenting widespread problem in the literature ...

Report

Current Biology

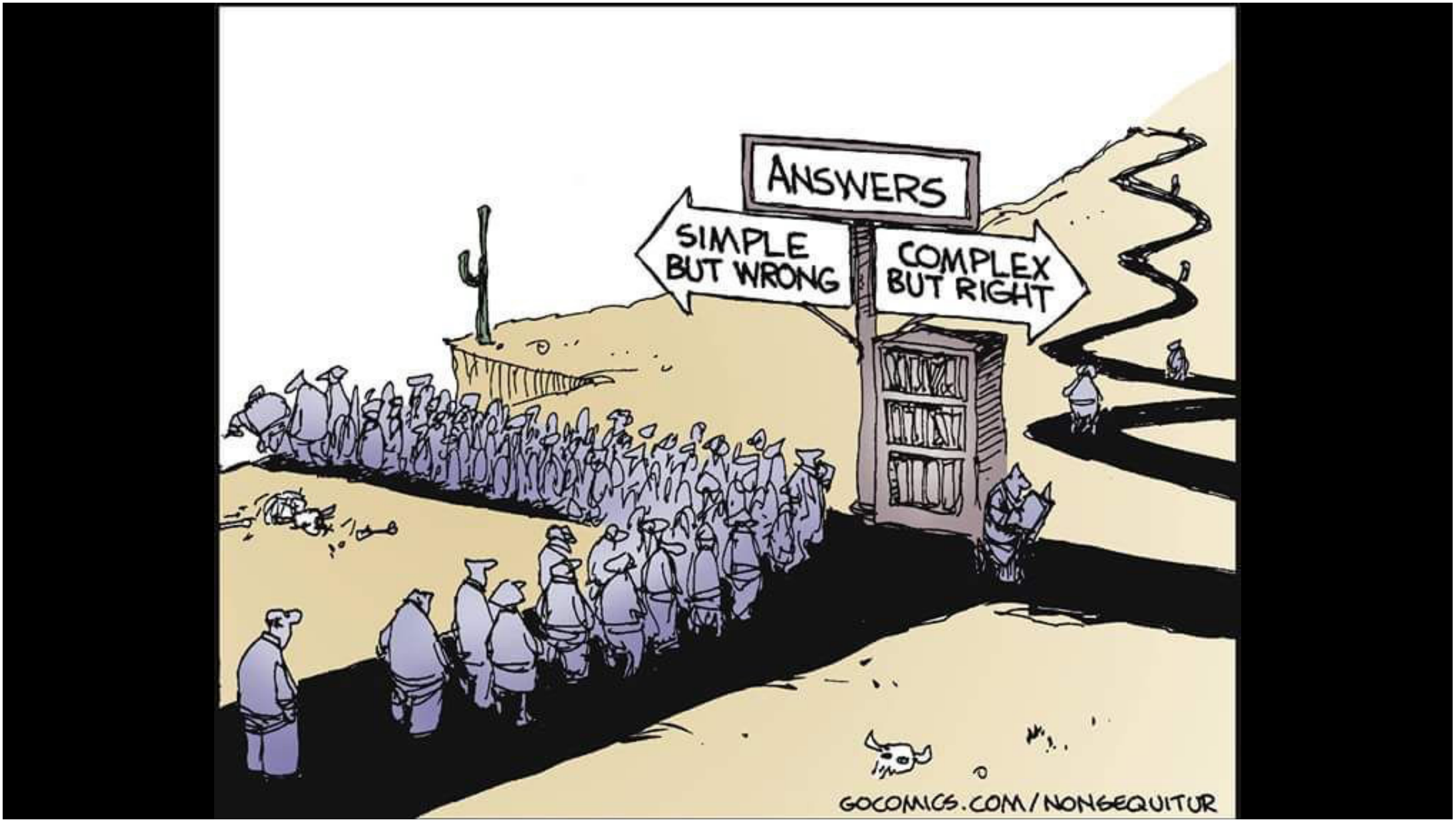
Avoidable false PSMC population size peaks occur across numerous studies

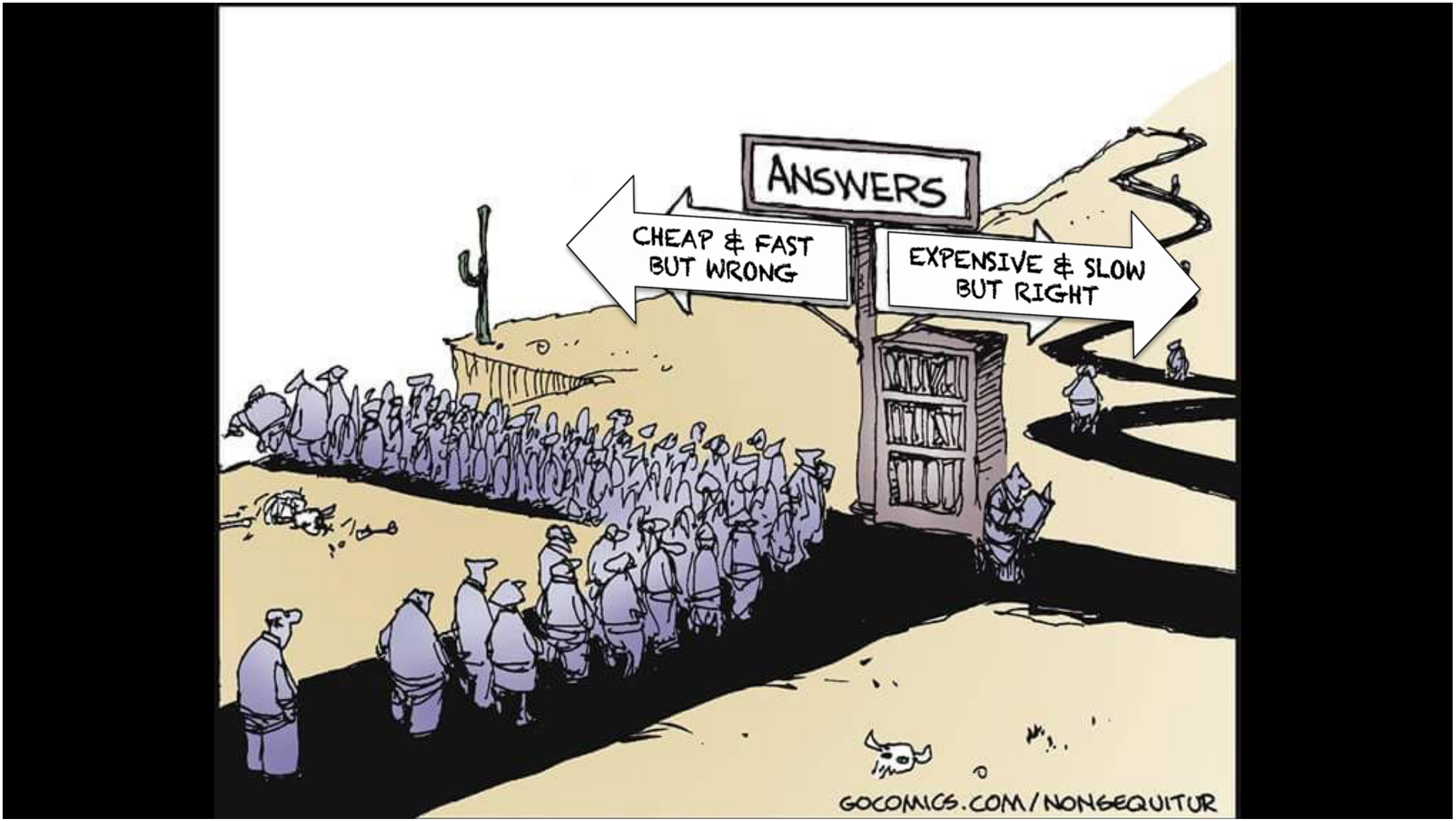
Highlights

- We detected and solved a common artifact in PSMC and related methods
- Default PSMC parameters frequently cause false peaks in recent population histories
- New methods perform better but do not always avoid this artifact
- Users should test multiple parameter settings that split the first time window

Authors

Leon Hilgers, Shenglin Liu, Axel Jensen, ..., Regev Schweiger, Katerina Guschanski, Michael Hiller


ANSWERS
SIMPLE BUT WRONG
COMPLEX BUT RIGHT
GOCOMICS.COM/NONSEQUITUR


ANSWERS
CHEAP & FAST
BUT WRONG
EXPENSIVE & SLOW
BUT RIGHT
GOCOMICS.COM/NONSEQUITUR

Test your hypotheses in independent ways

- Genomic datasets:
  - Are observational data where patterns observed have been created by events unknown to us
  - This is similar to all studies using observational data
    - Very susceptible to false positives and just so stories
  - Extremely large P-values can arise from extremely weak patterns, so ask yourself, does the effect and effect size have biological meaning?

Test your hypotheses in independent ways

- Derive hypotheses from your genomic results, then

- Test these hypotheses using relevant manipulations
  - Simulated data, parameter exploration
  - Functional validation via manipulation of genes, pathways, environments ... real hypothesis testing!!
  - Experimental evolution, CRISPR KOs, environmental perturbations
- If you can't manipulate, at least triangulate!

Triangulation

Robust research needs many lines of evidence

Replication is not enough

Munafò and Smith 2018 Robust research needs many lines of evidence

Triangulation

Robust research needs many lines of evidence

Replication is not enough

Munafò and Smith 2018 Robust research needs many lines of evidence

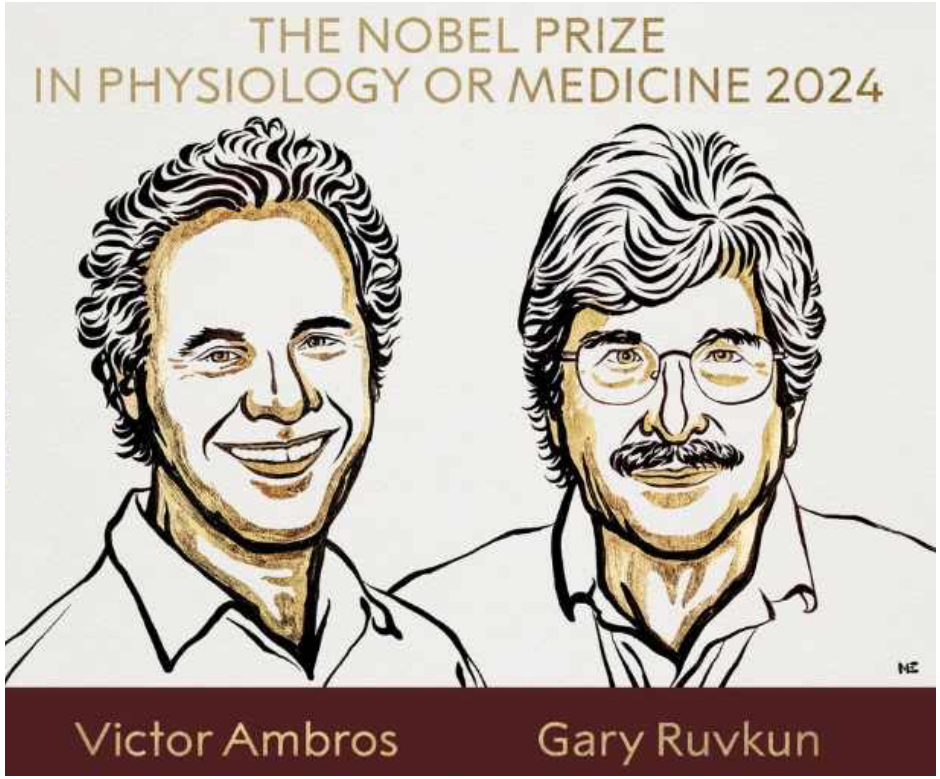
Triangulation

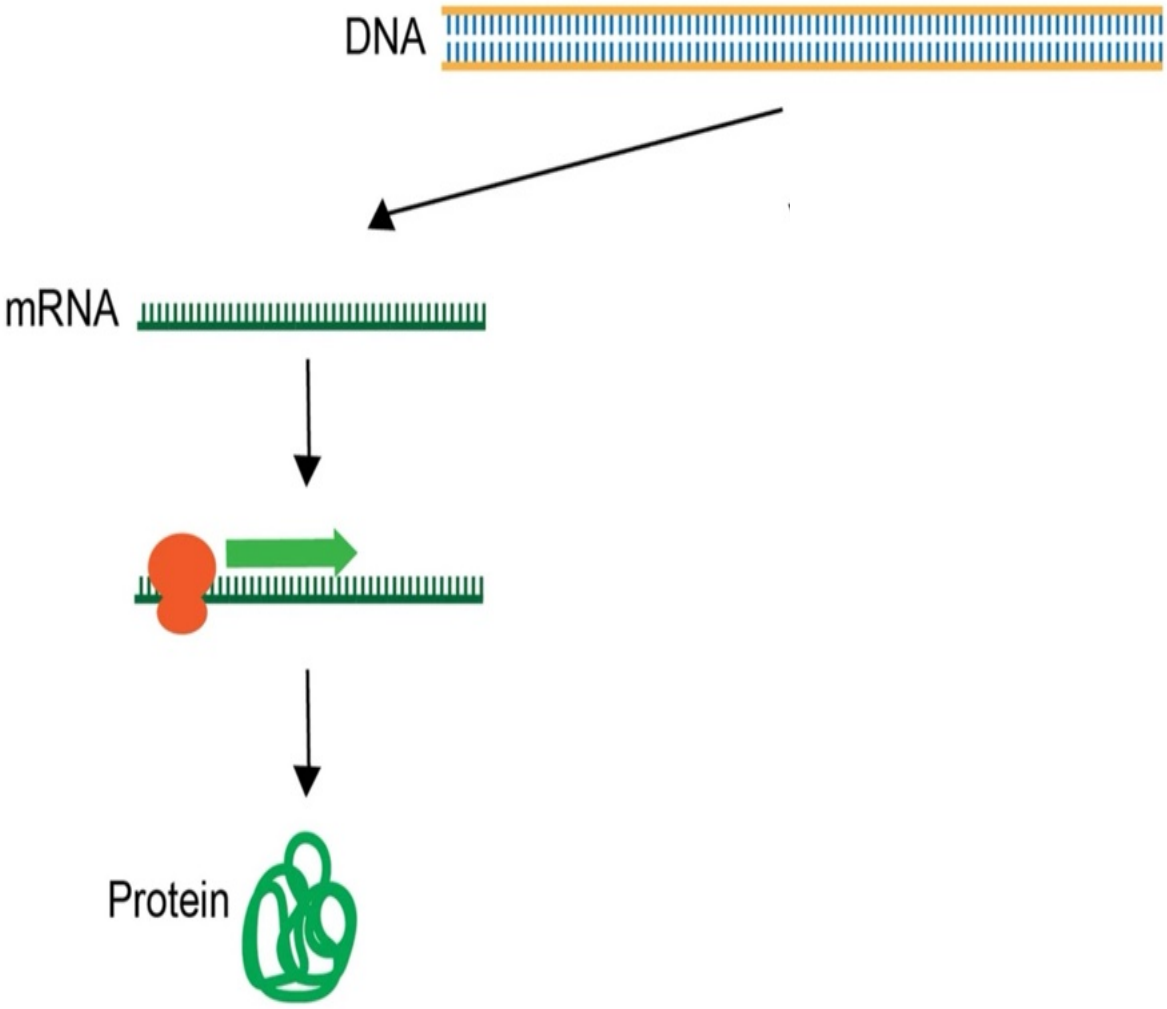
- Use different approaches to address the same hypothesis, or extensions of hypothesis
- Sources of bias for each approach should be explicitly acknowledged, in opposite directions, and independent
- Results from more than two approaches are ideally compared
- As genomics gets faster/cheaper, invest savings in validation

Munafò and Smith 2018 Robust research needs many lines of evidence

Buckle-Up for another bioinformatic journey

- Using a new genomics technique
  - miRNA
- Trying to understand what is best practice
- Worked hard to triangulate upon what's a biological signal vs. bioinformatic artifact
- Uncovered serious problem in the non-model community





"for the discovery of microRNA and its role in post-transcriptional gene regulation"

miRNAs

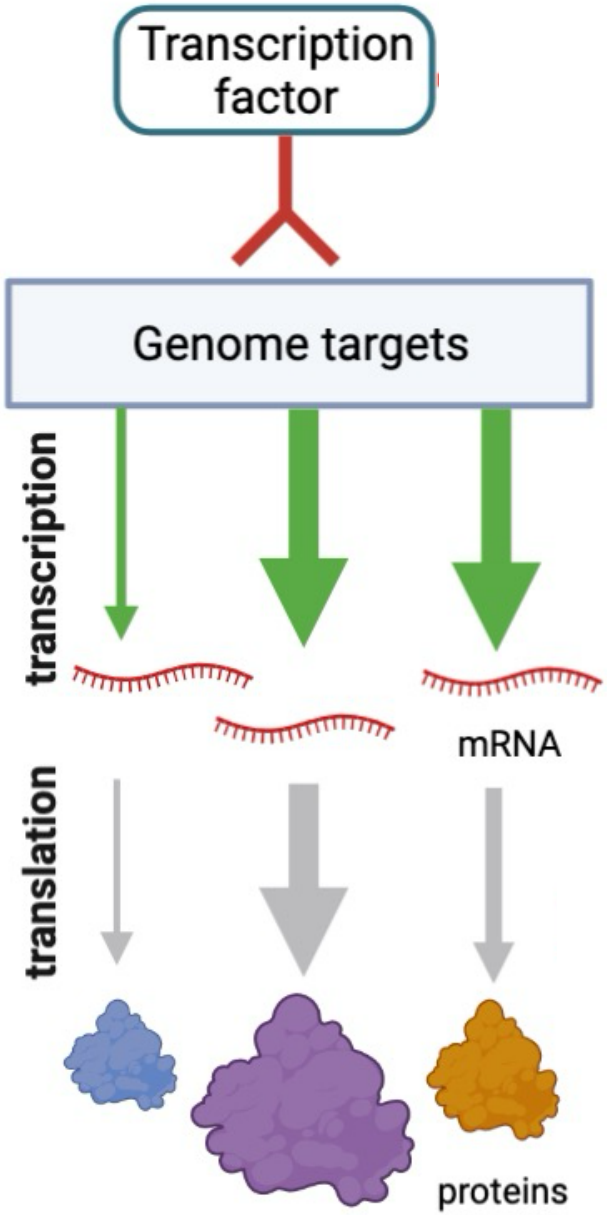
destabilize mRNA

sculpt the pool of mRNA

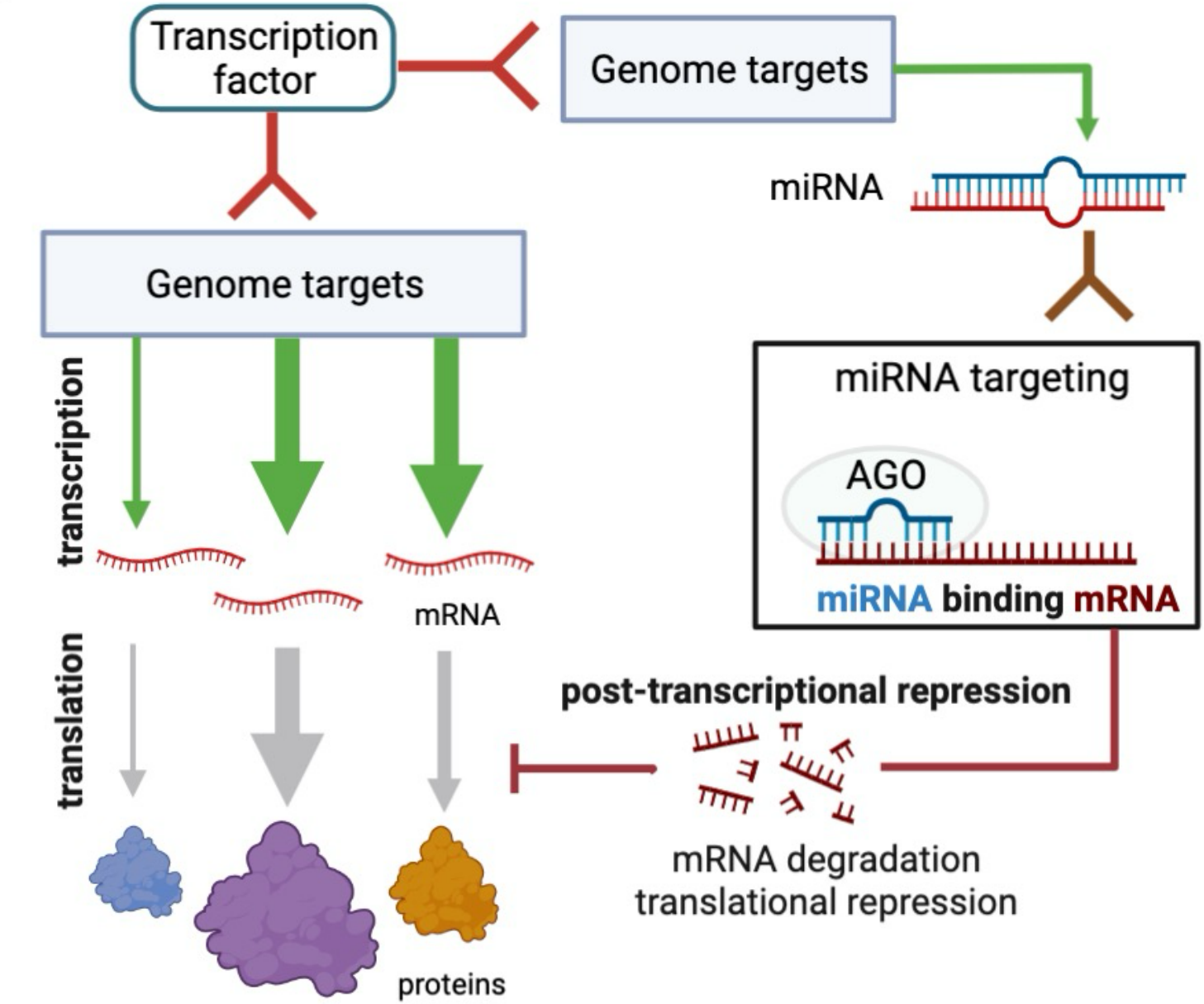
are a key part of regulatory networks

metazoans can't live without'em

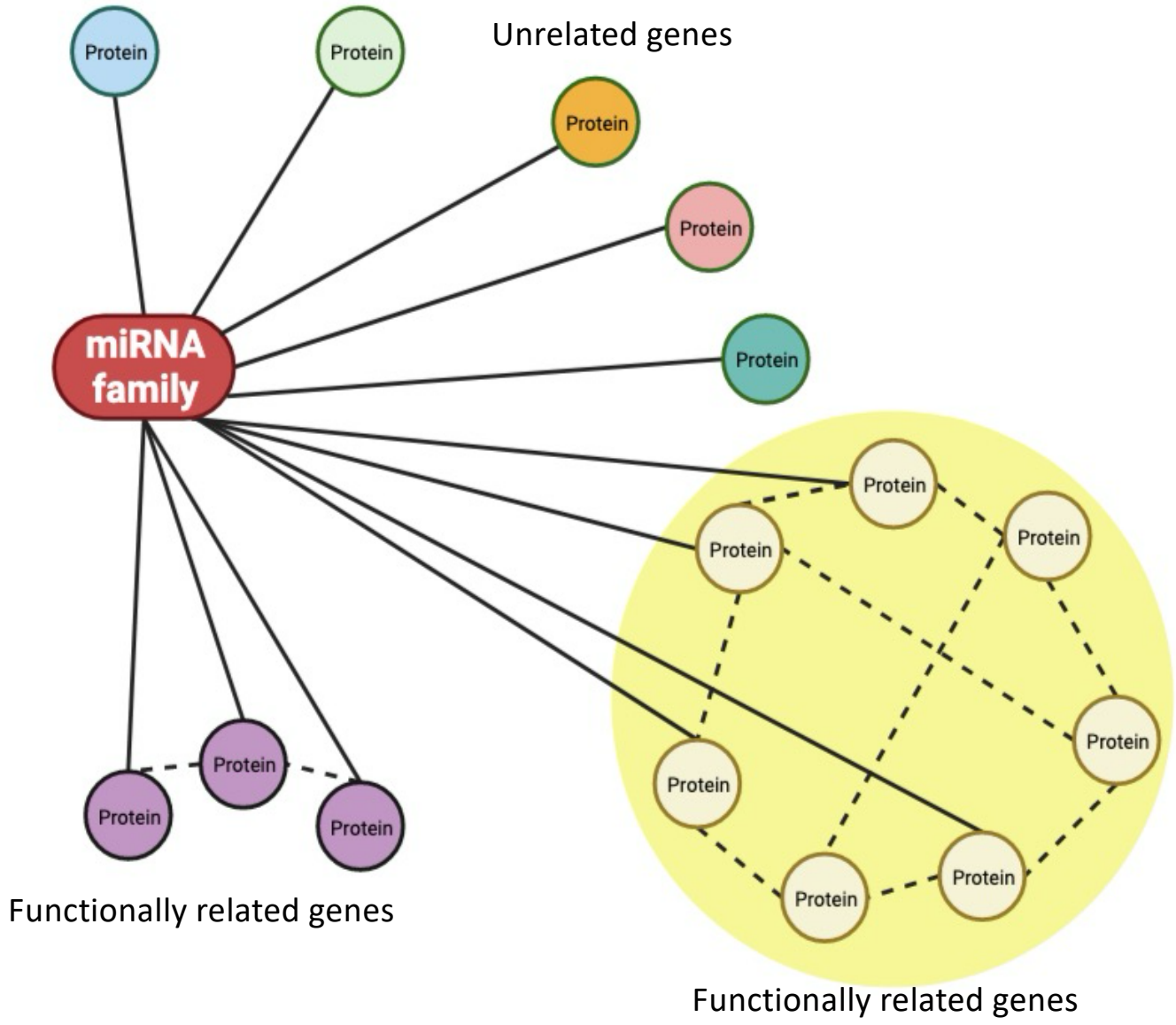
The role of miRNA in sculpting the transcriptome



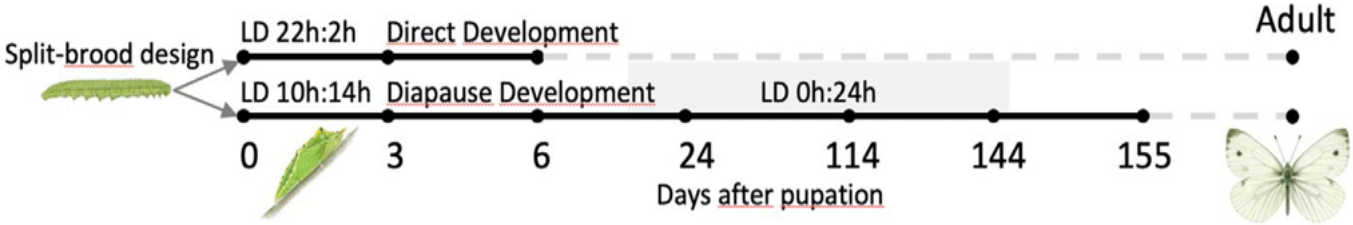
The role of miRNA in sculpting the transcriptome



Regulatory network view of miRNA impacts



Dynamic microRNA expression across diapause

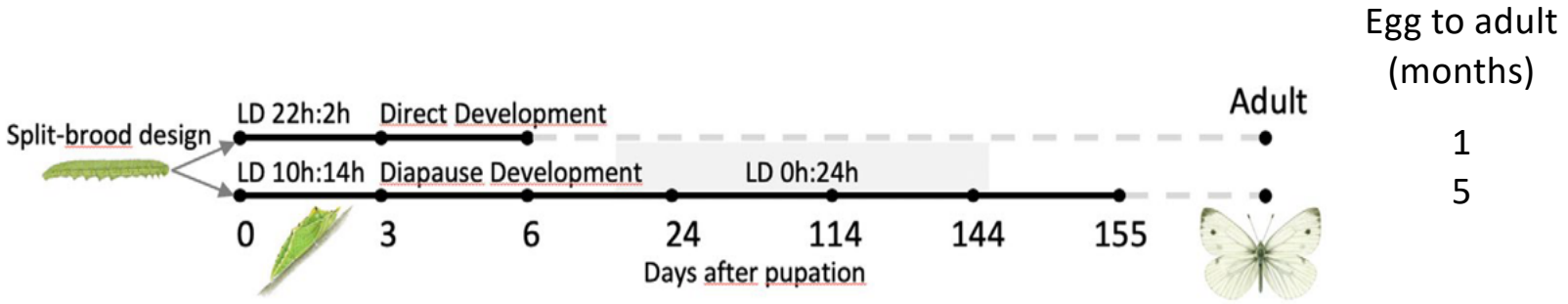


Egg to adult (months)

1

5

Dynamic microRNA expression across diapause



Pieris napi

miRDeep2 identified 188 microRNAs

73 libraries (±6.9M reads / lib):

- 12 time points
- 2 tissues
- 3 replicates

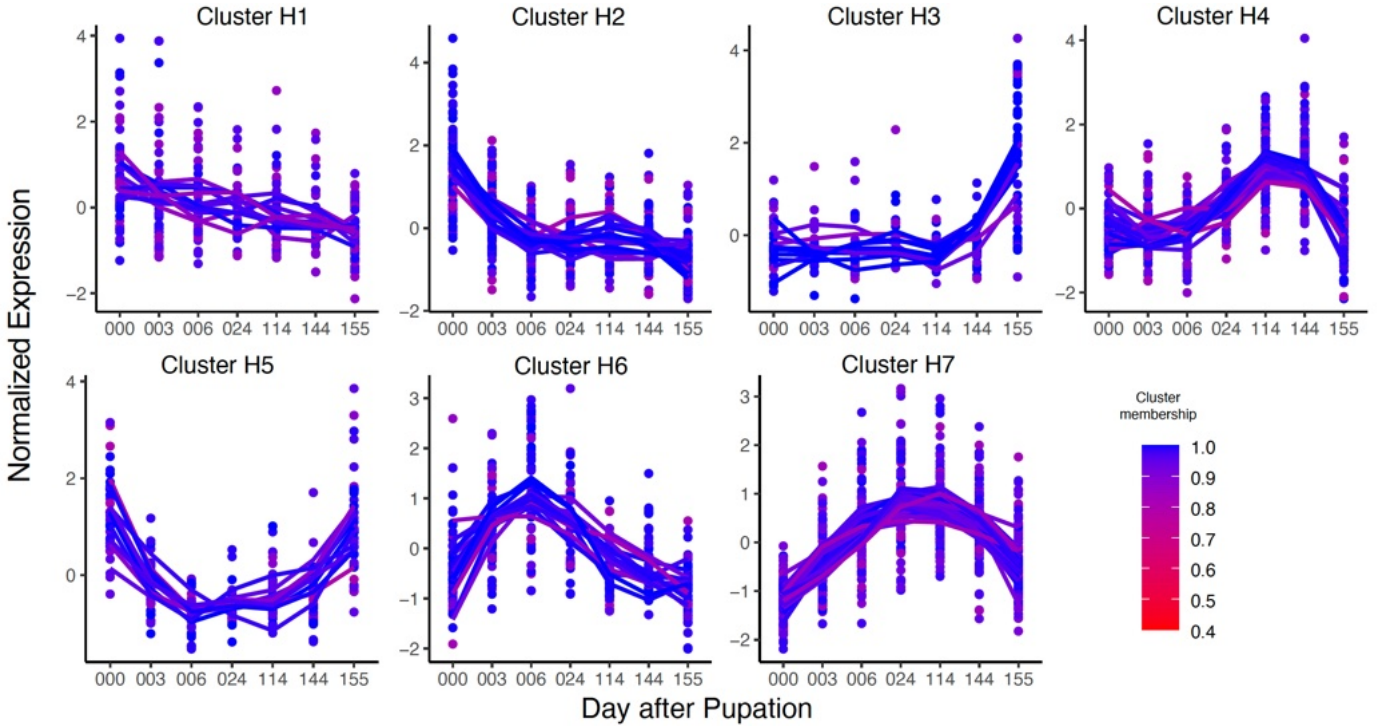
Roberts et al. 2024 Molecular Ecology

Dynamic microRNA expression across diapause

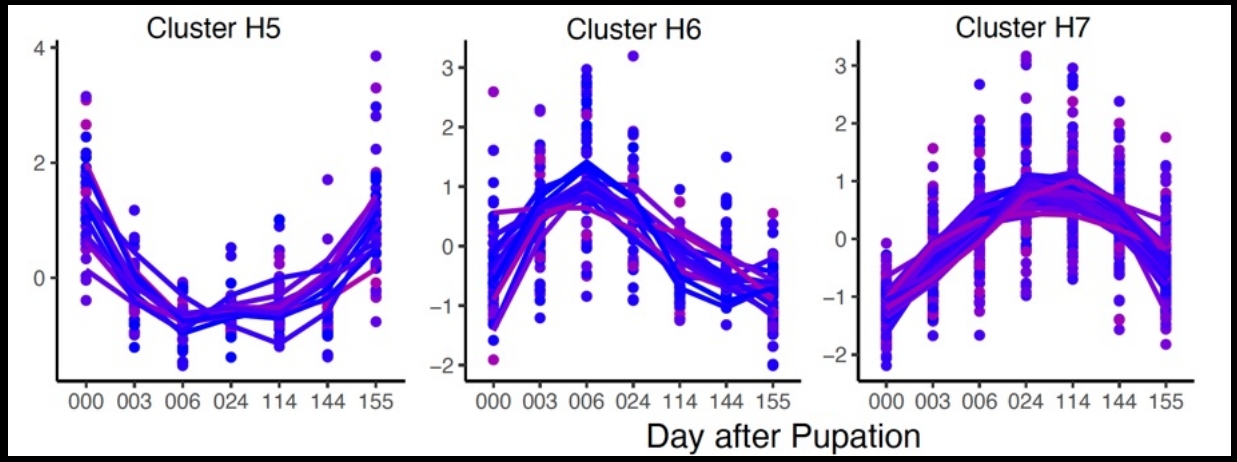
Egg to adult (months)

1

5

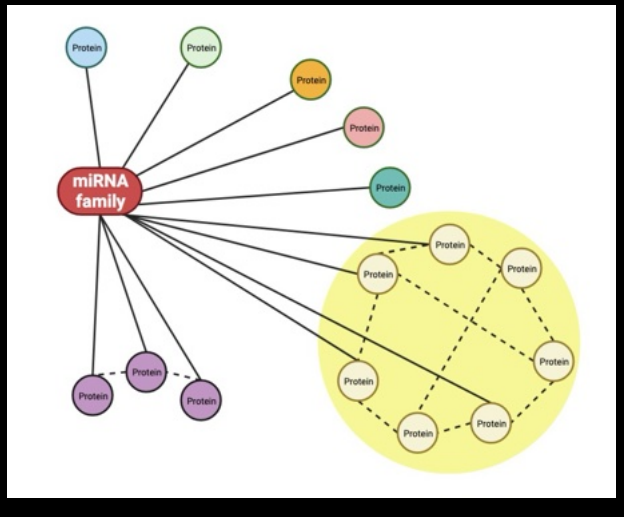


OK, so some miRNAs are changing through time ...



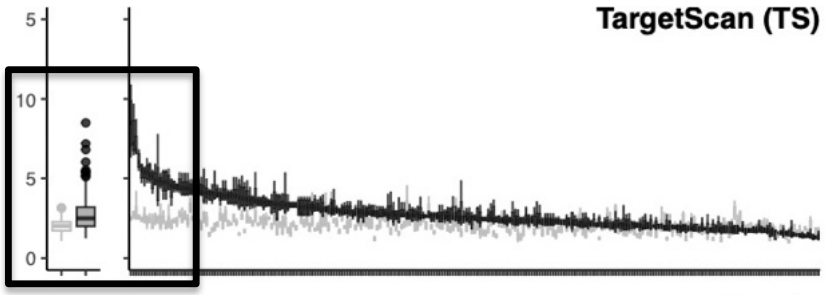
Where are they targeting?
What are they doing?

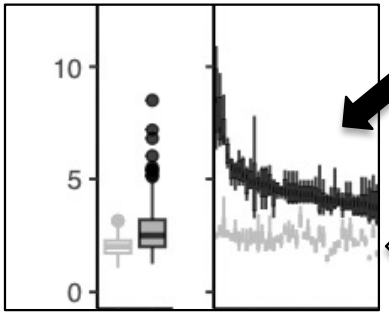
What functional groups or pathways might they regulate?



miRNA target detection

- miRNAs primarily bind a very short, ±7 bp region of the 3'UTR of mRNA
- This binding ultimately leads to a decrease of translated proteins
- There are 100,000's of 7 bp motifs in genome, of which miRNAs bind small fraction

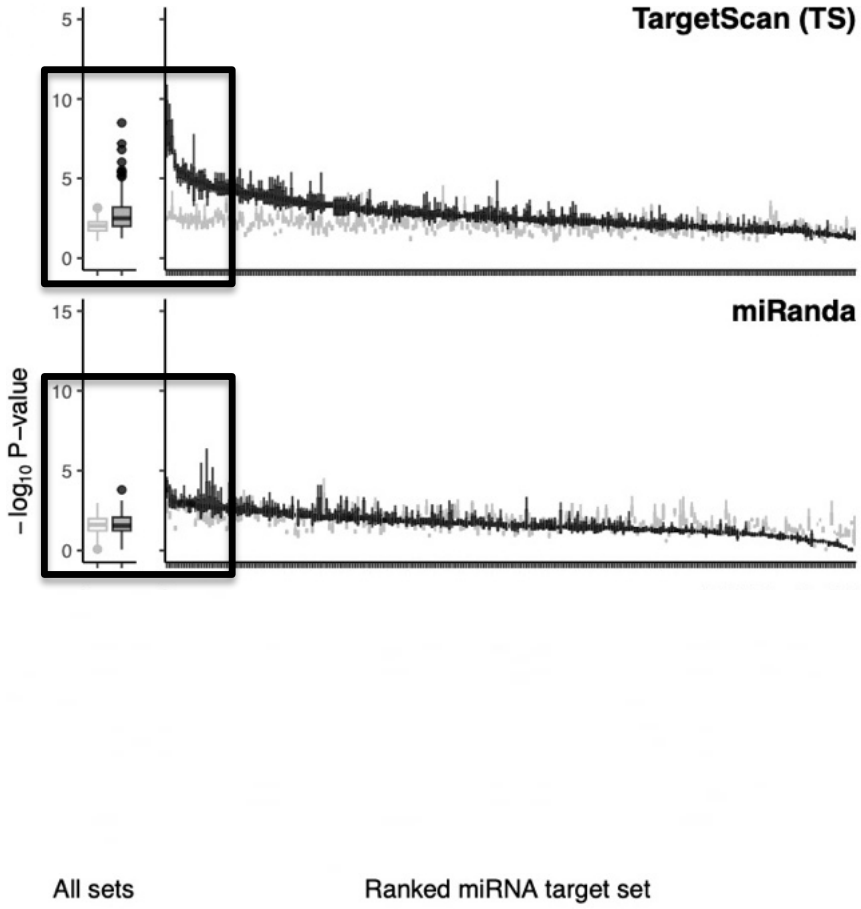


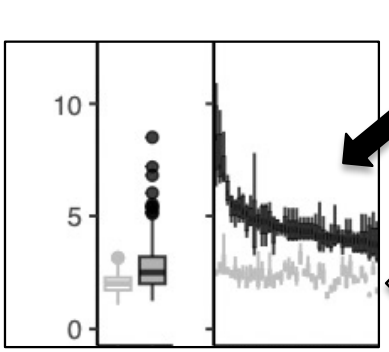


Targetscan using 6 species 3'UTRs alignments

Random sets

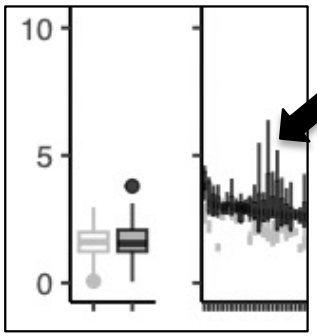
Wheat et al., in submission





Targetscan using 6 species 3'UTRs alignments

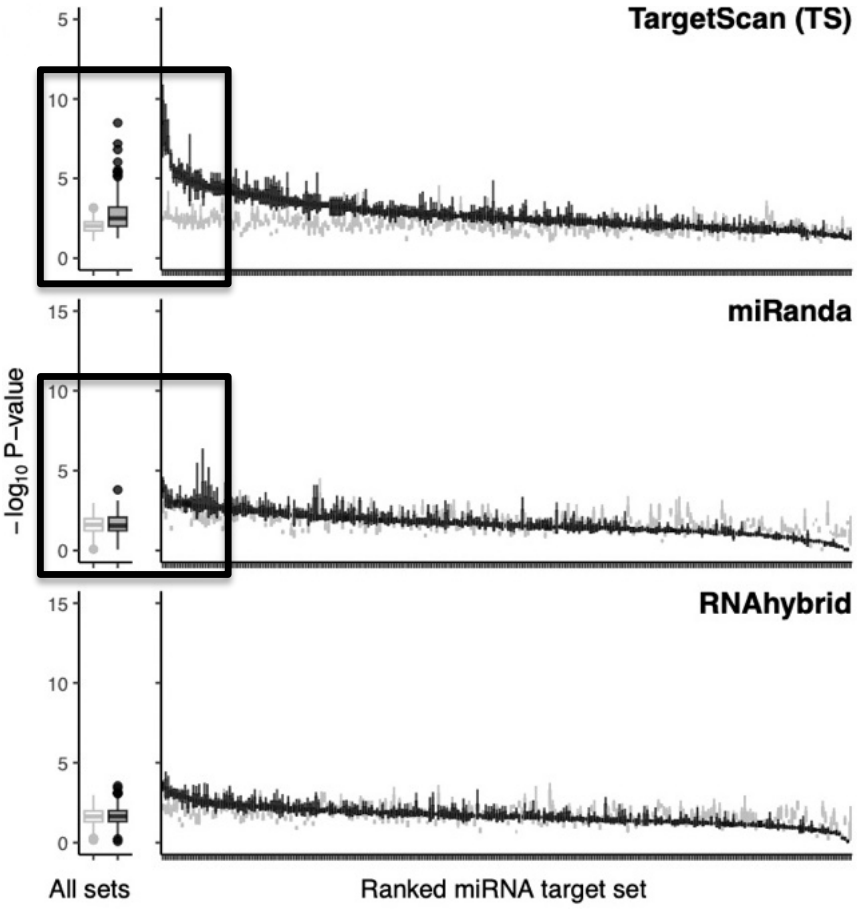
Random sets

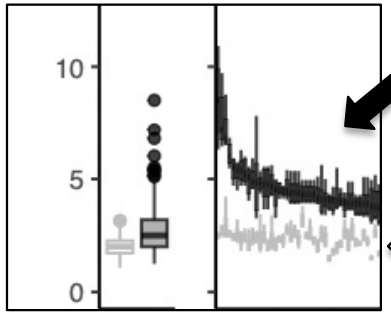


miRanda using one species 3'UTRs

Random sets

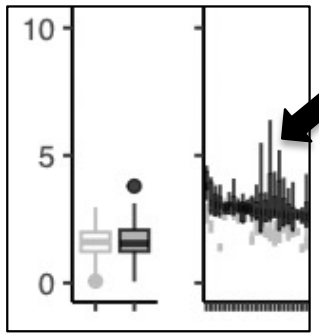
Wheat et al., in submission





Targetscan using 6 species 3'UTRs alignments

Random sets



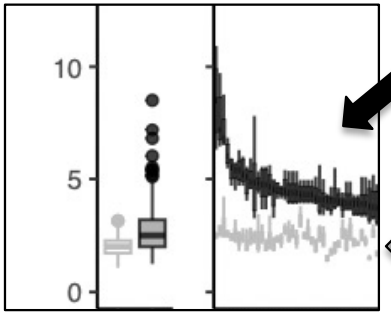
miRanda using one species 3'UTRs

Random sets

Wheat et al., in submission

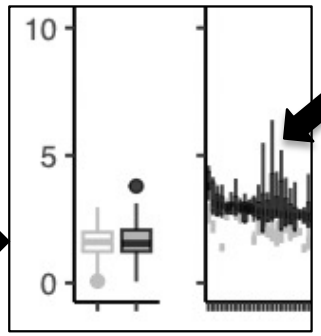
Large bioinformatic effect

>90% of miRNA literature in ecology and evolution uses miRanda to assess miRNA impacts…



Targetscan using 6 species 3'UTRs alignments

Random sets



miRanda using one species 3'UTRs

Random sets

Wheat et al., in submission

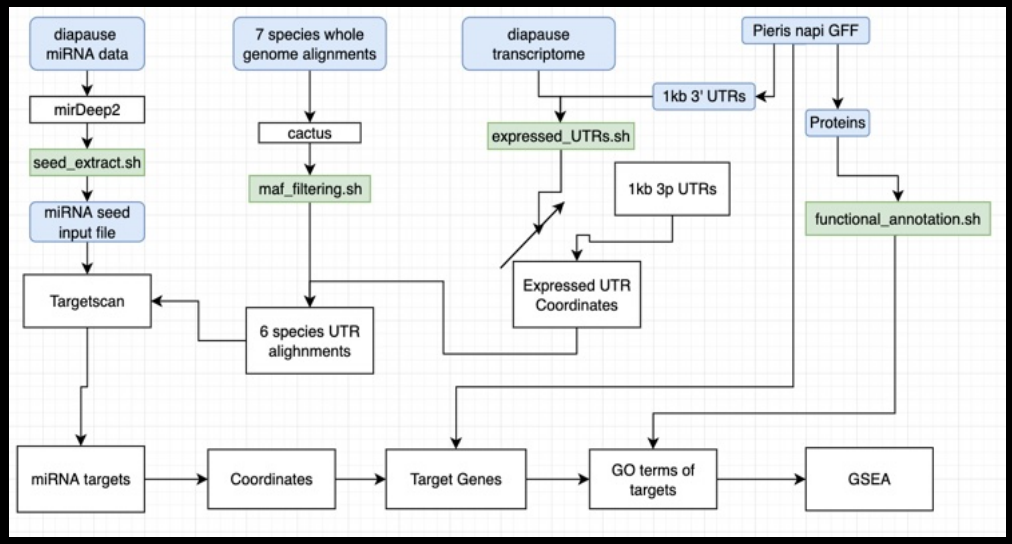
So, why don't more people use Targetscan approach?

Running miRanda is quick and easy

- Download, load 3'UTR data from your species, load miRNA seed sites, run

Running TargetScan7 with alignments is a lot of work

- Download scripts, generate 3'UTR alignments for 7 species, load miRNA seed sites, etc.



Bioinformatic analysis of miRNA targets

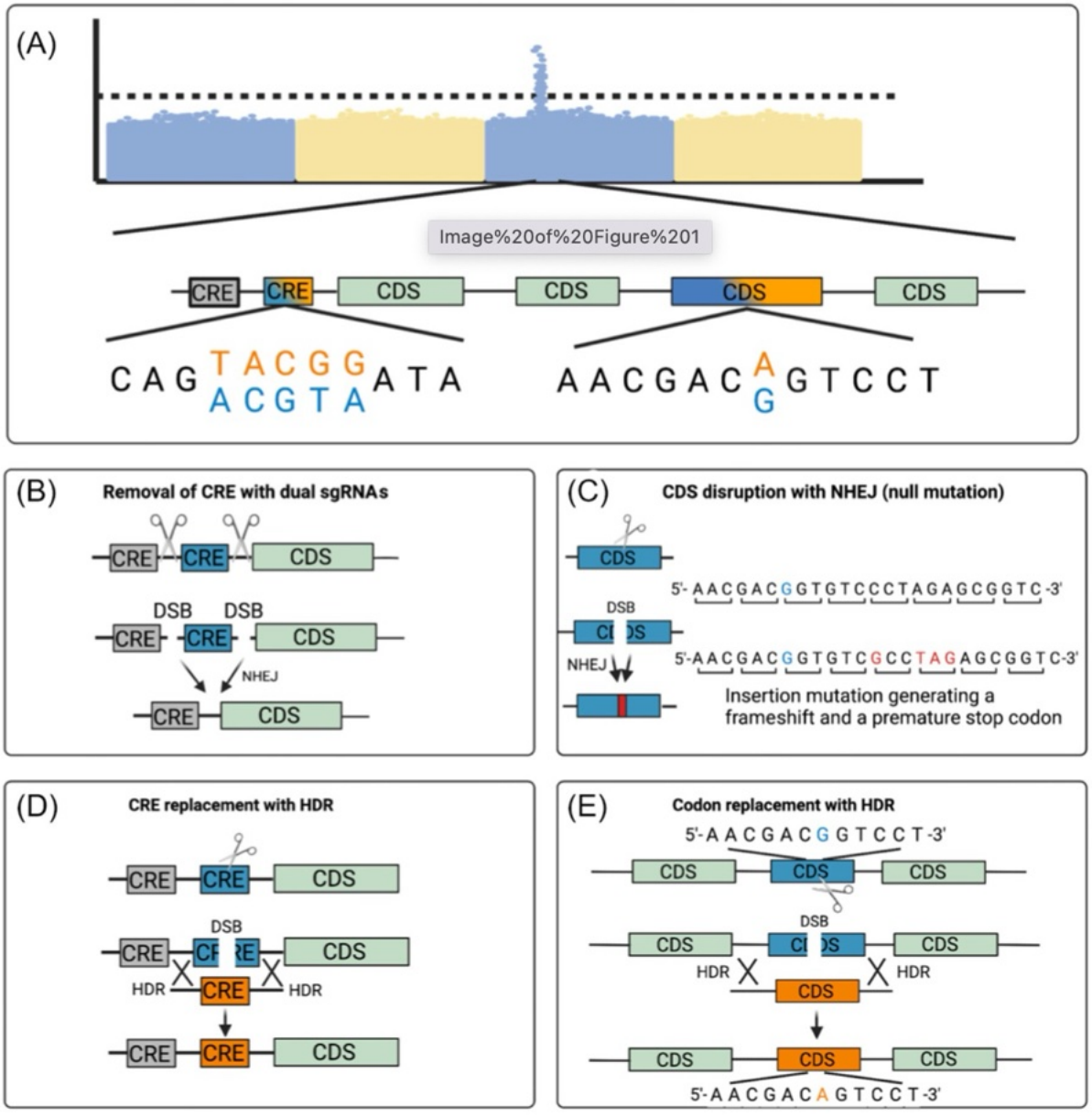
Detecting miRNA expression changes is easy, but target detection is inherently very difficult

- Intersection
  - Comparison across bioinformatic tools
    - Revealed inconsistent results, primarily because used VERY different methods (e.g. using vs. not using alignments)
  - Developed novel metric for assess biological signal in results
  - Species comparisons for cross-check & generality

Sum: intersection across divergent methods, 1st principals metric, and comparative analysis revealed believable results

Recent review covering diverse means of validation across diverse taxa

As genomics gets cheaper, invest more in validation instead of just more sequencing!!!!!


(A)
Image%20of%20Figure%201
CRE
CRE
CDS
CDS
CDS
CDS
CAG TACGG ATA
ACGTA
AACGAC AGTCCT
G
(B)
Removal of CRE with dual sgRNAs
CRE
CRE
CDS
DSB
DSB
CRE
CRE
CDS
NHEJ
CRE
CDS
(C)
CDS disruption with NHEJ (null mutation)
CDS
5'-AACGACGGTGTCCCTAGAGCGGTC-3'
DSB
5'-AACGACGGTGTCGCCTAGAGCGGTC-3'
NHEJ
Insertion mutation generating a
frameshift and a premature stop codon
(D)
CRE replacement with HDR
CRE
CRE
CDS
DSB
CRE
CDS
HDR
CRE
HDR
CRE
CRE
CDS
(E)
Codon replacement with HDR
5'-AACGACGGTCCT-3'
CDS
CDS
CDS
DSB
CDS
CDS
HDR
HDR
CDS
CDS
CDS
CDS
5'-AACGACAGTCCT-3'

A bit of bioinformatic wisdom

- Expect errors and noise
  - Analysis results need many rounds of refinement
  - Invoke biological causes of results last
- 70% of your time will be troubleshooting (AI might help reduce this)
  - This is normal, keep a notebook, intermediate files
- Fear the new and shiny programs that will simplify your life
  - 80% of all new software will not be usable
    - Un-installable, no manual, no test examples, not repeatable
    - Beware of these red flags, as many authors only seek a publication and won't help

Cookbooking ...

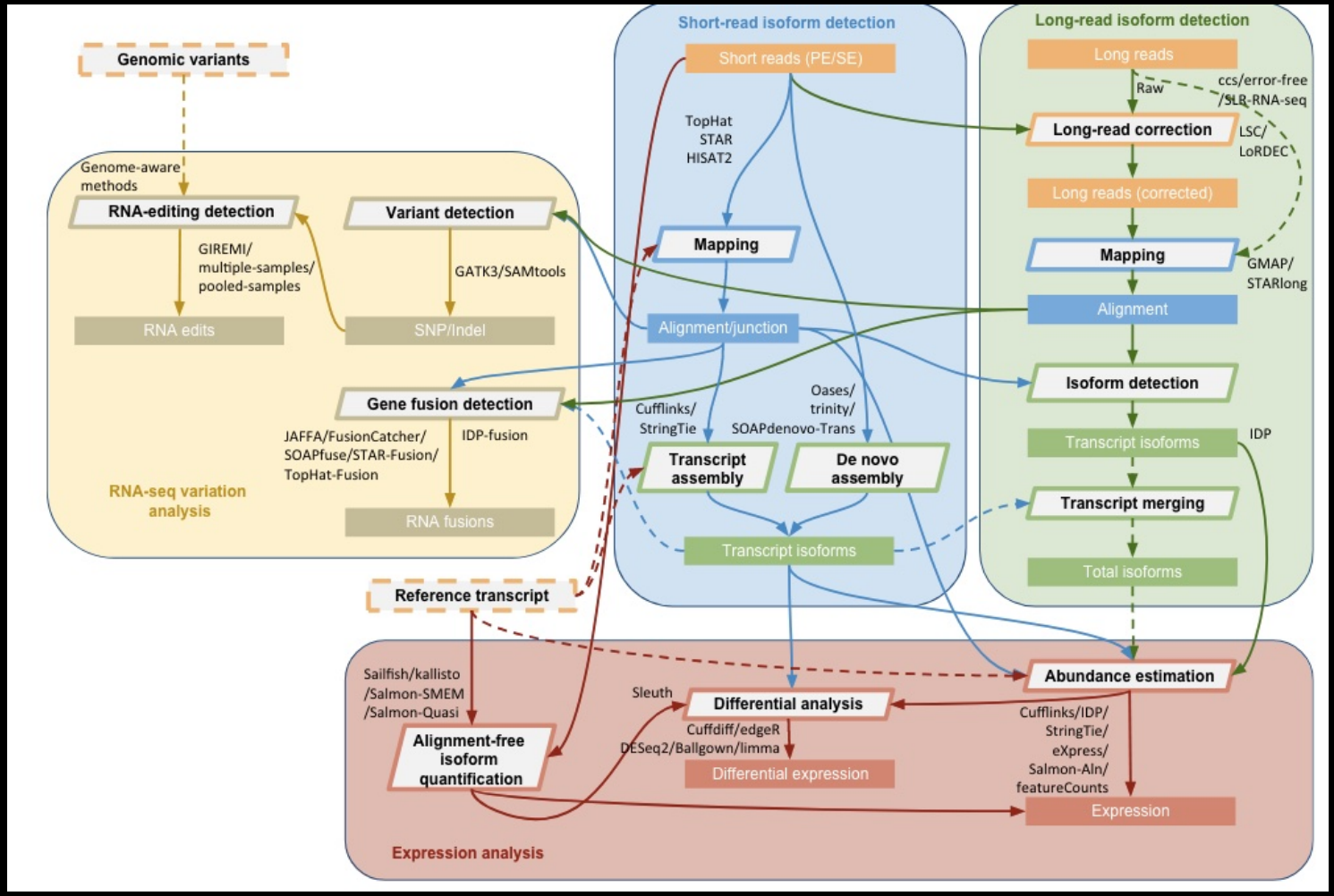
- Google and AI are your friends
- Use them, but don't trust them
- Test what you use, learn from it, build your own toolbox
- Copying without learning makes you easily replaced and open to massive errors

Keep good bioinformatic notes

- I keep a special file with commands I learned, like and validate
  - use it to quickly find commands, refresh memory
- Use positive and negative controls to test the output of the commands you run (like all experimental biology)
  - I call these sanity checks
  - Always test to make code is working correctly
    - Great reason to use > 1 method, right?
- Read up on good file structure, version control, and how to parallelize your commands

Publish your code, no matter how messy

Many different ways to make a pipeline



Sahraeian SME et al. 2017. Nat Commun.

Many tools, performance varies across species, samples. This is no BEST tool or setting across all species

Differential expression detection can vary by:

- Mapper
- Analysis software
- Reference genome
- Species

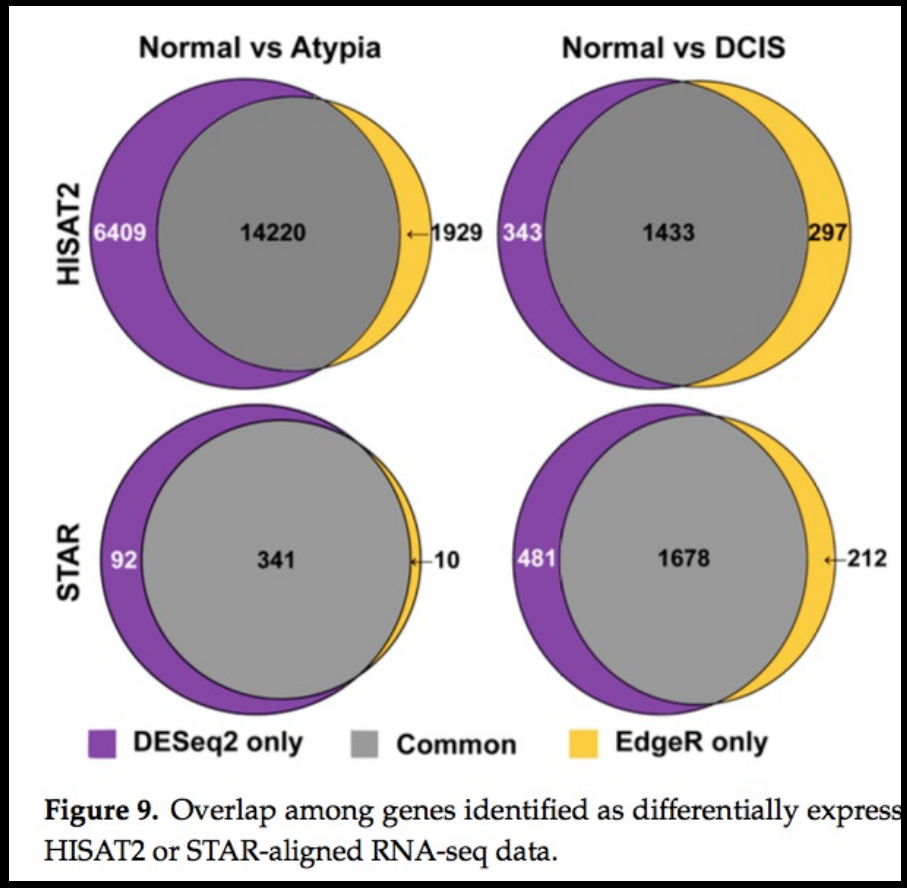
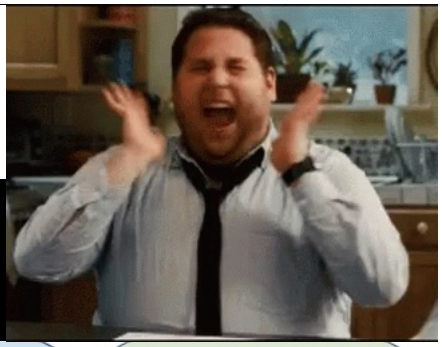


Figure 9. Overlap among genes identified as differentially express
HISAT2 or STAR-aligned RNA-seq data.

Raplee et al. 2019 J. Per. Med.

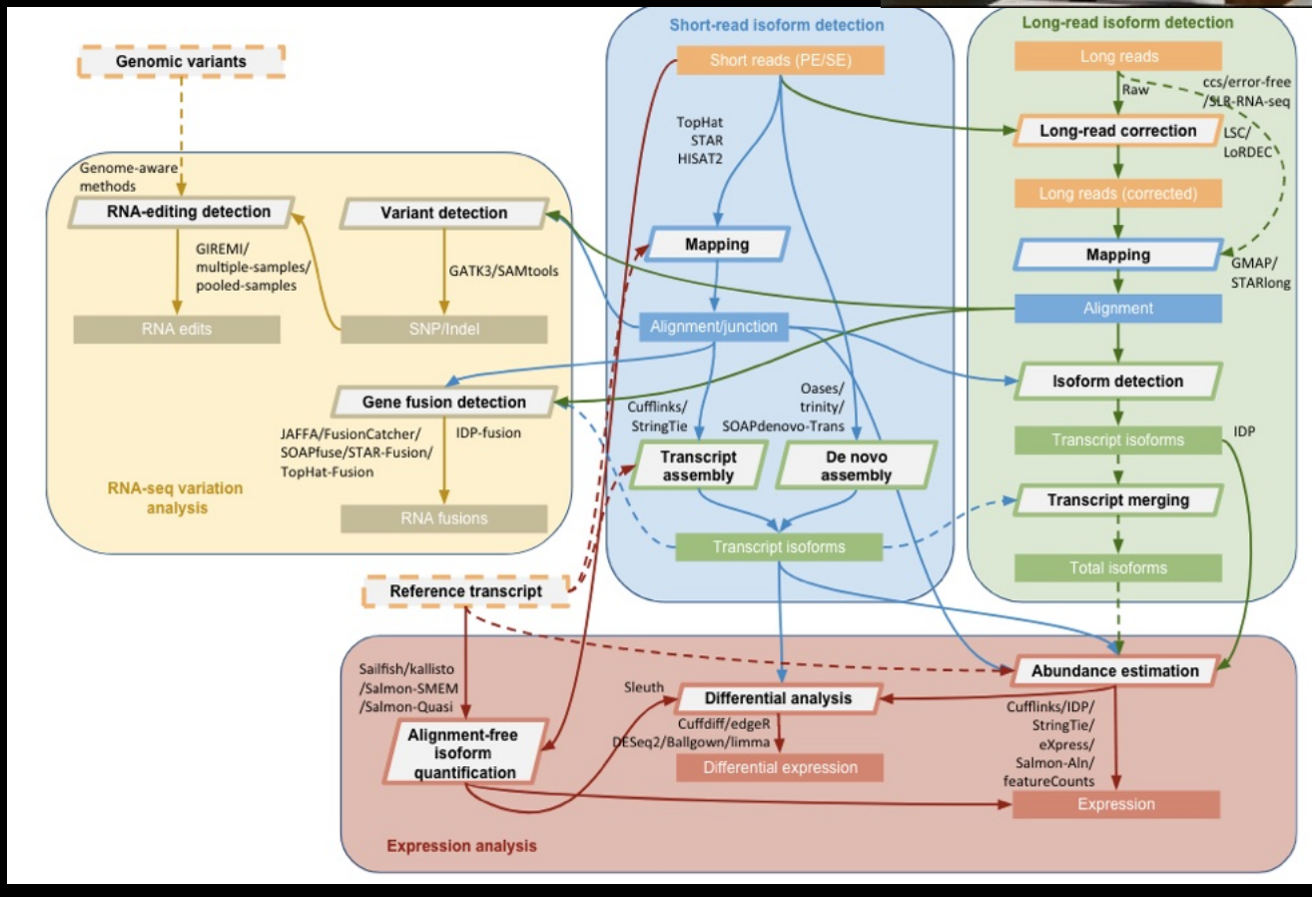
Analysis paralysis is common .. Which is the right way?



Just start by get through a single pipeline, start to end

Then try different approach to assess your first results

Used published data & code, then explore additional approaches



Experimental design matters: RNAseq and sample size

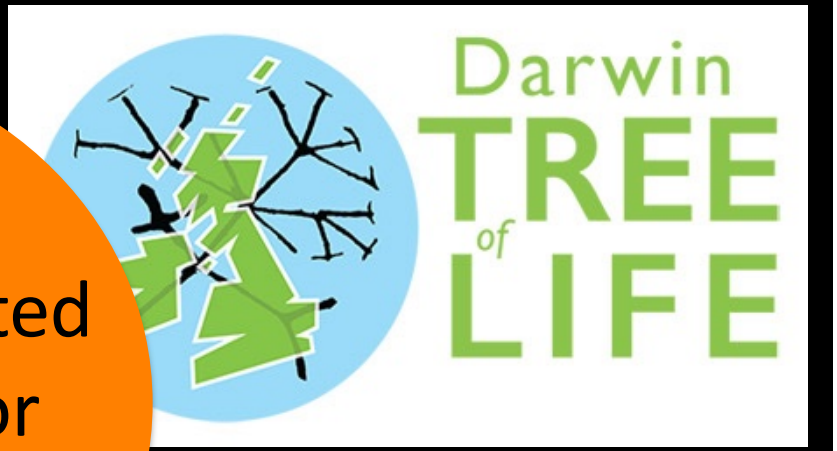
- Simple two group comparisons
  - 30 wild-type mice vs 30 mice in which one copy of a gene had been deleted
  - Full analysis: n=30 vs. n=30

FDR= % of DEG in subsample not found in full analysis

Sensitivity = % of genes found in subsample also seen in full analysis

Halasz, et al. bioRxiv July 11, 2024, p 2024.07.08.602525. https://doi.org/10.1101/2024.07.08.602525.

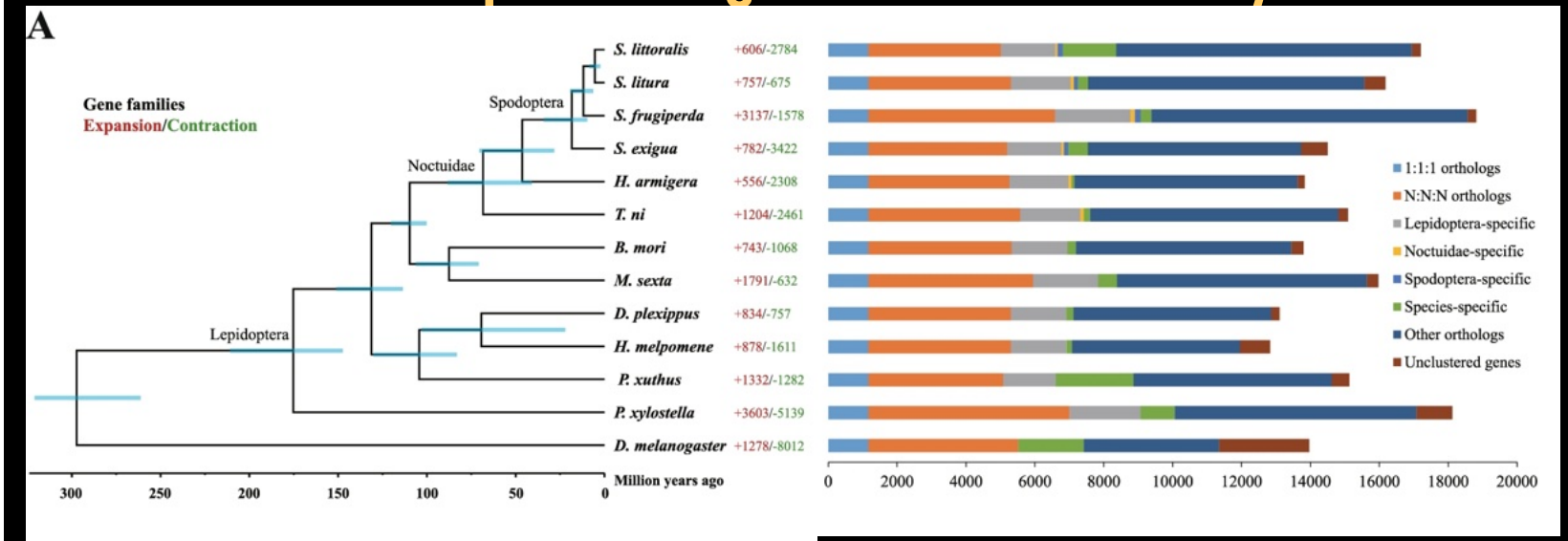
Entering the mega-genomes era

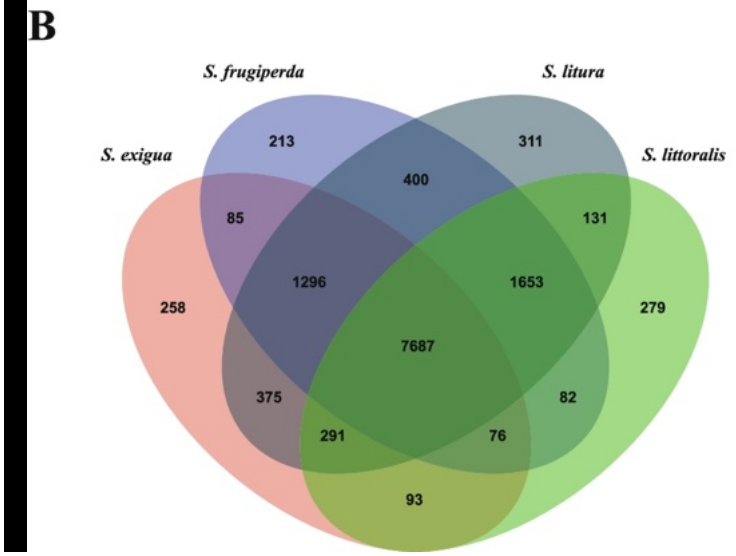


An unprecedented opportunity for large scale errors?

- Unp dying:
  - –Phylog onships
  - –Genome evolution
  - –Functional study of diverse adaptations

Comparative genomics commonly use annotations





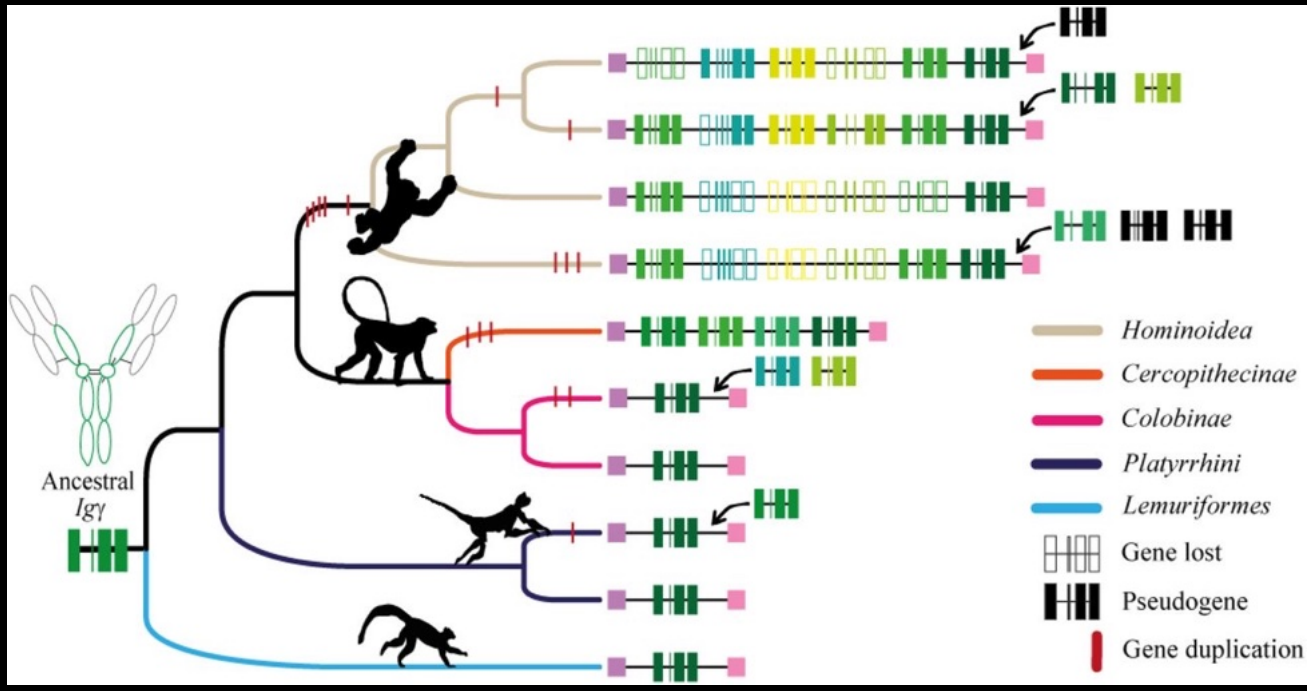
Typical genome report comparing gene content among species

- Number of genes
- Rates of gene birth, death
- # of lineage specific genes

Gene birth-death dynamics: biased, artifacts, or meaningful results?

- Are changes in gene numbers across species meaningful?
- Fundamental and important evolutionary question
- Very difficult to assess accurately
  - Need good genomes, annotations
  - Then good analyses

Immunoglobulin heavy constant gamma gene evolution



Garzón-Ospina & Buitrago 2022

Are all annotations equal among species?

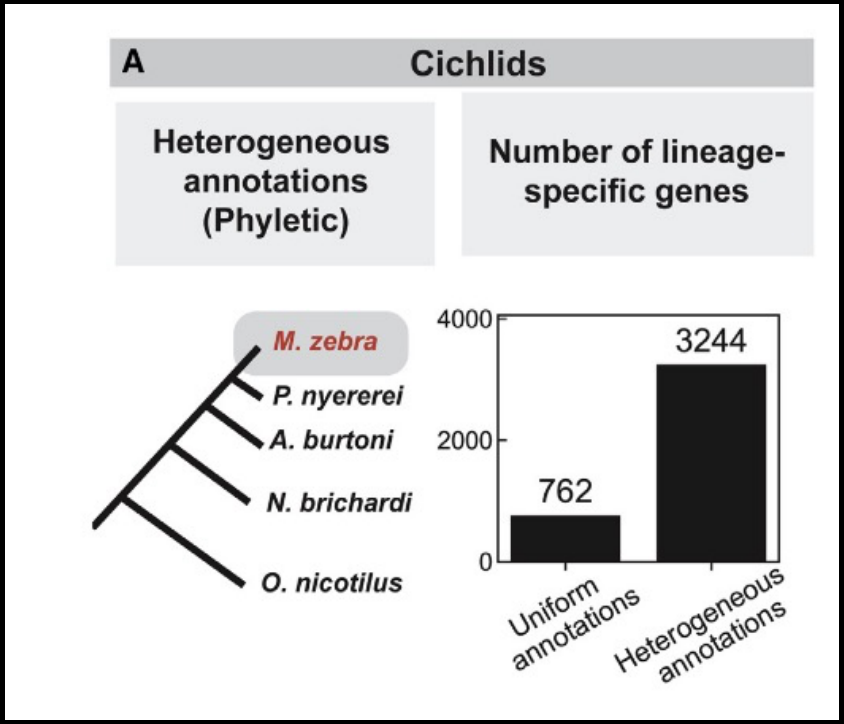
- Do species genomes differ in:
  - When they were sequenced, thus technology?
  - The quality of their assembly (e.g. N50, haploid state)?
  - How they did their annotation (proteins only vs. lots of RNAseq)?

Then resulting annotation protein sets likely differ due to technology, not biology

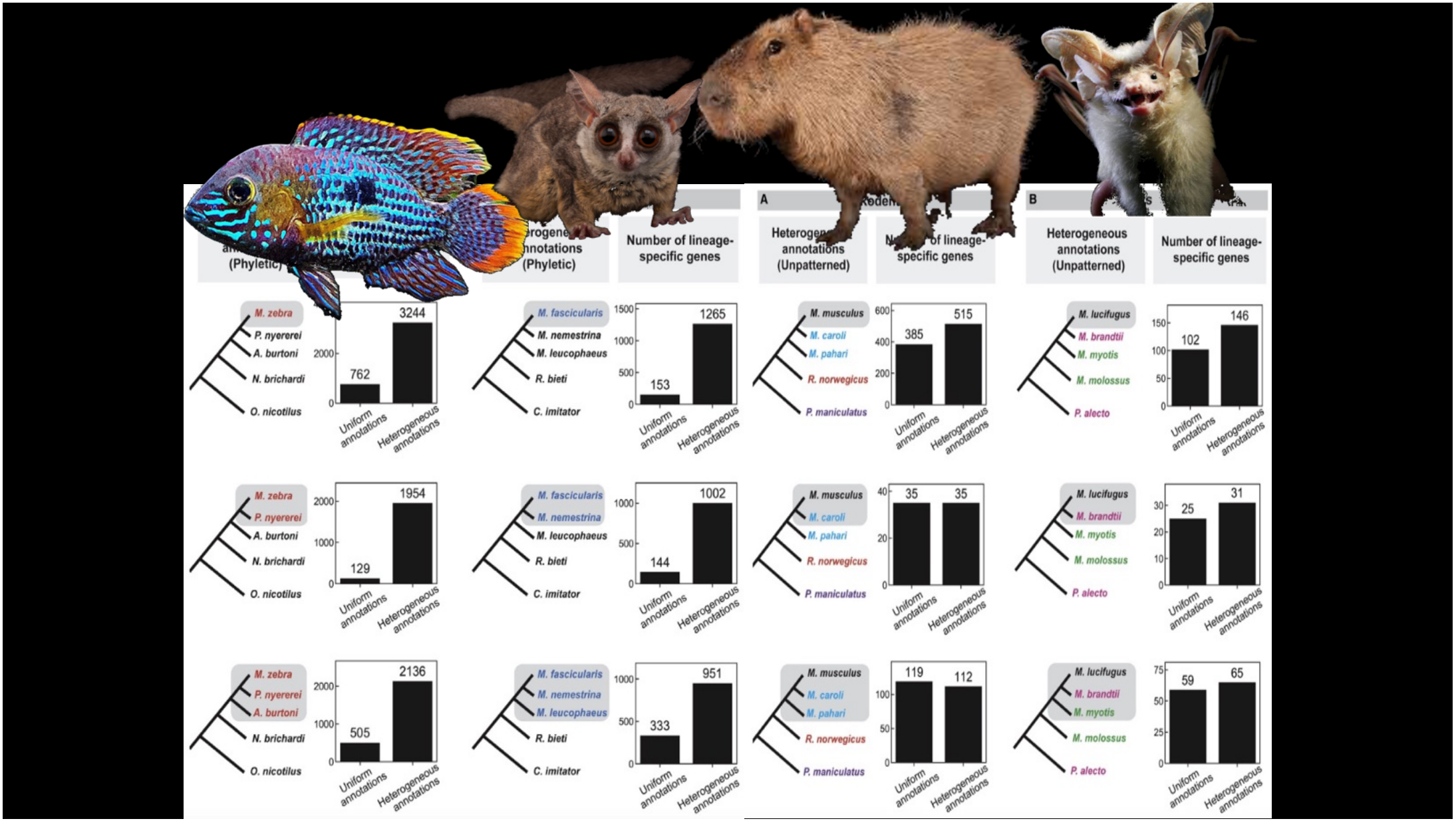
Will this impact analyses that rely upon accurate protein sets?

Non-standard annotations introduce major artifacts

- Lineage specific genes inflated by
  - 10 to 1000's of genes, with increases up to 15 fold



Weisman et al. 2022. Current Biology


Number of lineage-specific genes
Heterogeneous annotations (Unpatterned)
Number of lineage-specific genes
Heterogeneous annotations (Unpatterned)
Number of lineage-specific genes
M. zebra
P. nyererei
A. burtoni
N. brichardi
O. nicotilus
3244
762
1265
153
385
515
146
102
1954
129
1002
144
35
35
31
25
2136
505
951
333
119
112
65
59
M. fascicularis
M. nemestrina
M. leucophaeus
R. bieti
C. imitator
M. musculus
M. caroli
M. pahari
R. norwegicus
P. maniculatus
M. lucifugus
M. brandtii
M. myotis
M. molossus
P. alecto
Uniform annotations
Heterogeneous annotations

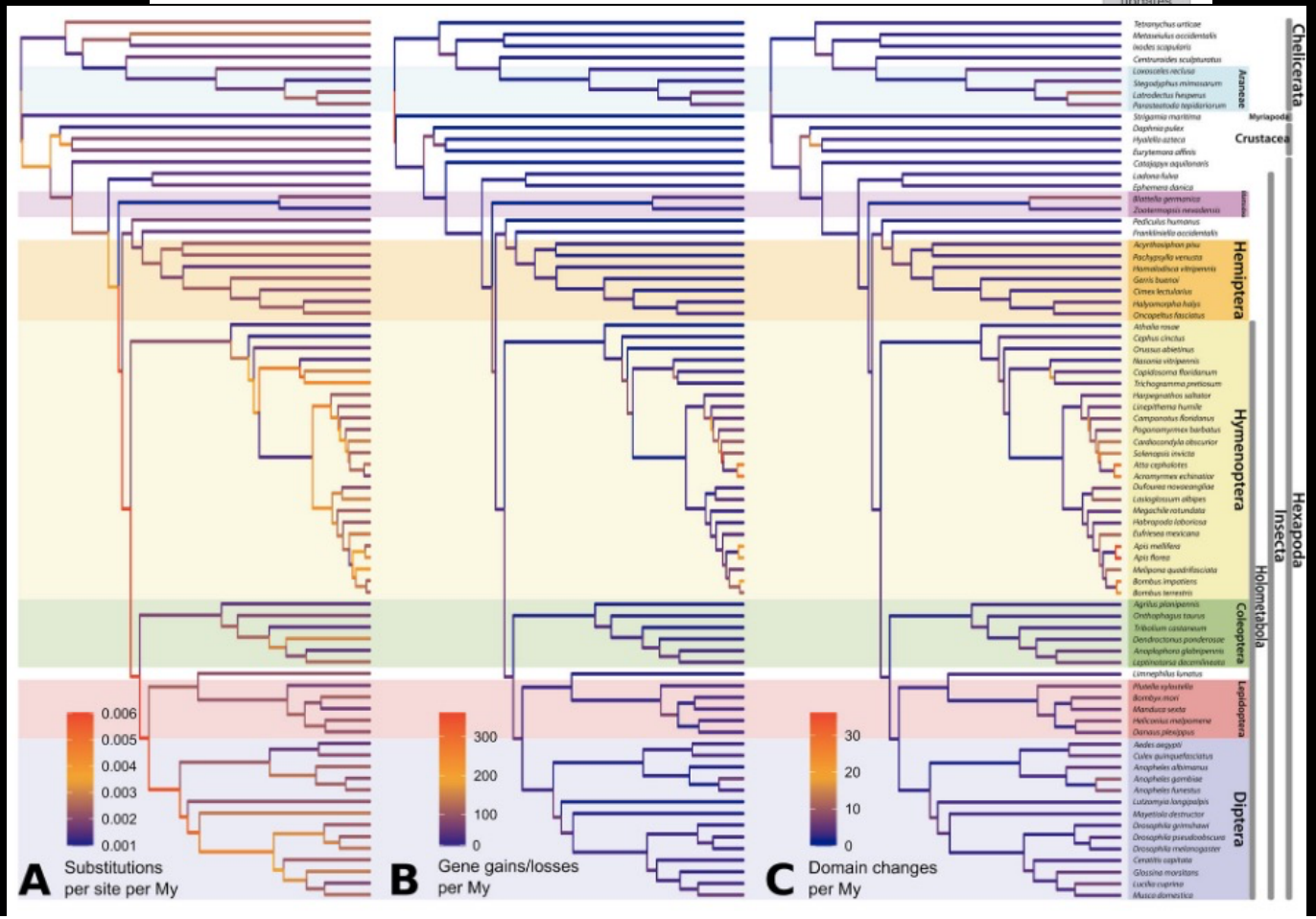
Comparative genomics and gene content evolution

Thomas *et al. Genome Biology* (2020) 21:15
https://doi.org/10.1186/s13059-019-1925-7

Genome Biology

RESEARCH Open Access

Gene content evolution in the arthropods



Some major conclusions of the paper

A

Last Insect Common Ancestor: 147 emergent gene families

Function	Emergent families
Wing morphogenesis	EOG86HJQQ EOG8TMTG9 EOG80ZTDS
Exoskeleton development and pigmentation	EOG8Q2GZG EOG8RZ1DS EOG8VDSCK EOG8WHC14 EOG8XPT03
Adaptation to terrestrial environment	EOG83XXJ1 EOG82VBZ4 EOG8PVRGC EOG8HTC7X
Larval behavior	EOG81K1SK

B

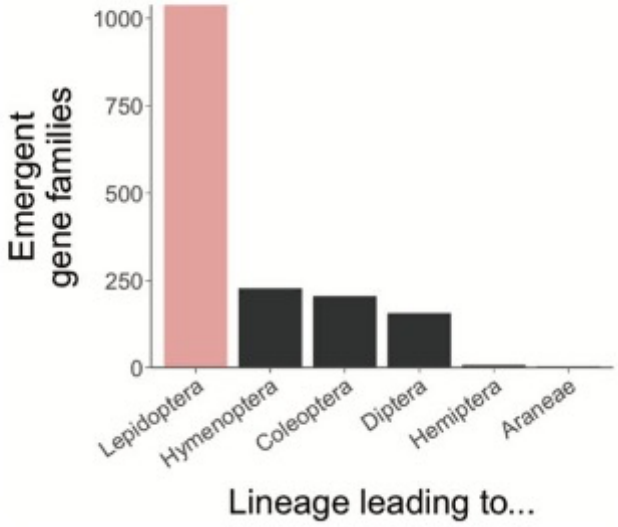
Last Holometabolous Common Ancestor: 10 emergent gene families

Function	Emergent families
Anterior head segmentation	EOG8HDW8X
Nucleosome assembly	EOG8G1PZD
Transporter activity	EOG847J8K
Transferase activity	EOG8ZPH98
Serine-type endopeptidase	EOG8QJV3F

+ 5 families with no known function

C

Last Lepidopteran common ancestor: 1,038 emergent gene families



"Although the majority of these gene sets were built using MAKER, variation in annotation pipelines and supporting data, introduce a potential source of technical gene content error in our analysis."

Post-genomics challenge

"What we can measure is by definition uninteresting and what we are interested in is by definition immeasurable"

- Lewontin 1974

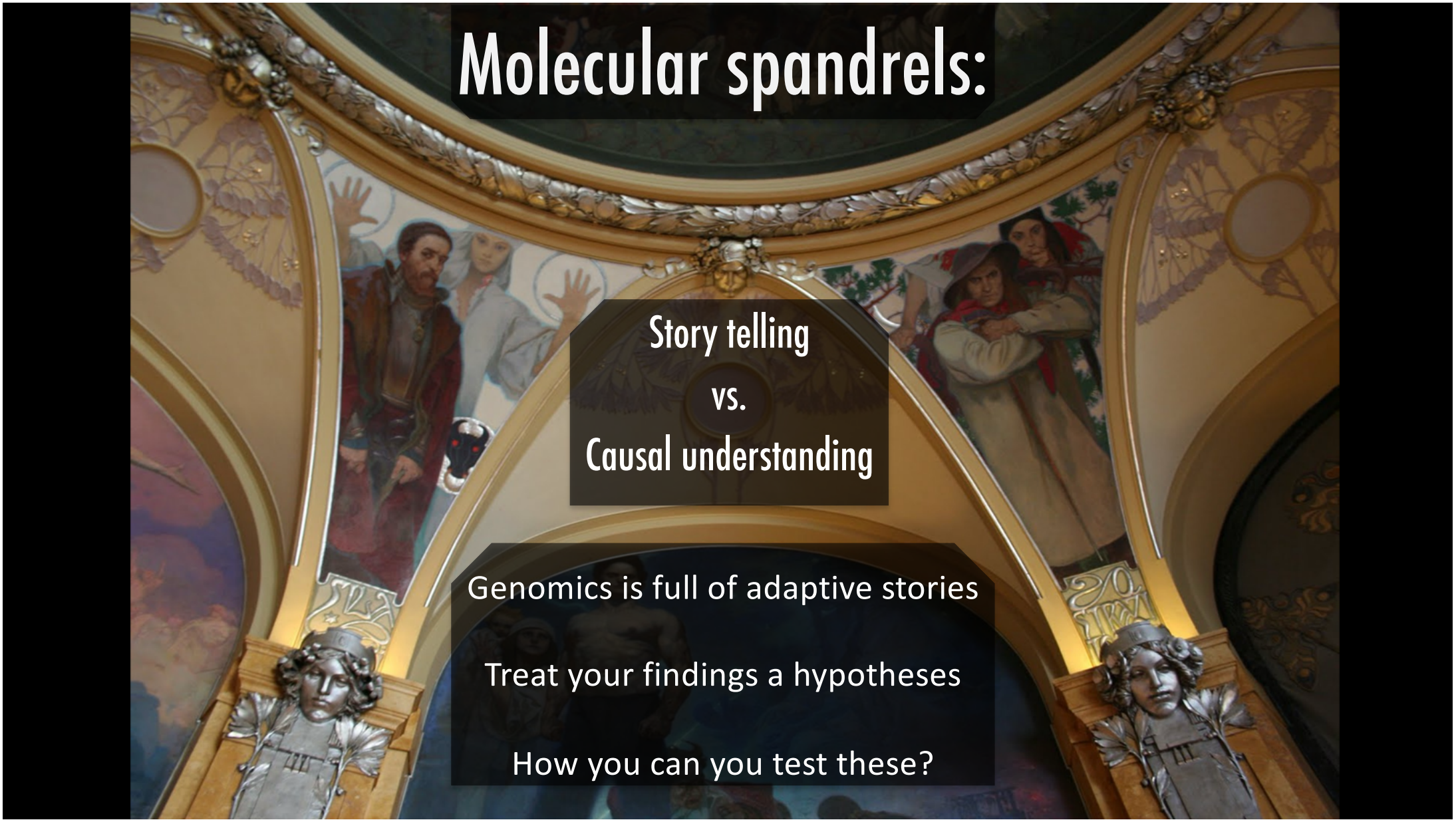
"What we understand of the genome is by definition uninteresting and what we are interested in is by definition very damn difficult to sequence and assemble and annotate and analyze at the genomic scale"

- Wheat 2015

Interrogate your results

"you need to be in charge of the analysis"

- The more you analyze your data, your confidence will grow
  - Let your findings talk to you in different ways
- Graph your results – visualize the patterns, assess 1st principals
  - Always start with PCA or MDS plot (how do your samples cluster?)
  - Compare with your different analysis results
- If you find interesting genes or patterns, can you test this hypothesis?
  - Using independent samples?
  - At a higher level of biological organization?
  - In some manipulative, functional way?


Molecular spandrels:
Story telling
vs.
Causal understanding
Genomics is full of adaptive stories
Treat your findings a hypotheses
How you can you test these?

Never forget your origins and biases

Genomic results are only hypotheses:
easy to get, likely misleading, need validation