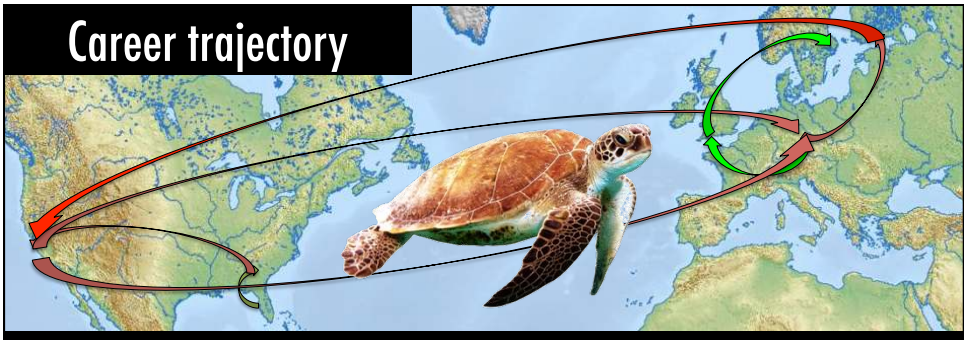
Lies, damn lies, and .... genomics

you, your data, your perceptions and reality

Christopher West Wheat

Career trajectory



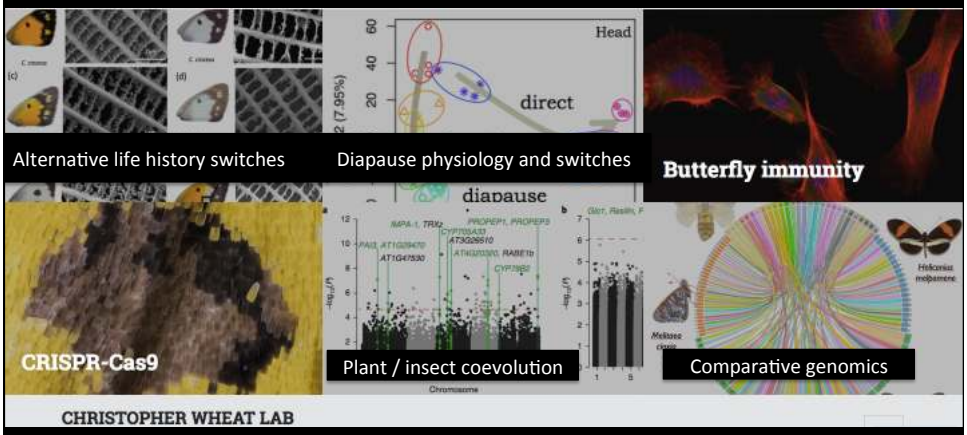
- 1995 – 2001 PhD California
- 2002 – 2005 Postdoc Germany
- 2005 – 2008 Postdoc Finland
- 2009 – unemployed 4 month, spent all savings
  - > 50 job applications, 1 grant application
- 2009 – visiting scientist Germany
  - 1 job offer UK
  - 1 grant Finland
- 2012 – started tenure track Sweden

What was important?

- Being able to move
- Chasing the money & skills
- Learning how to:
  - Write publications, grants
  - Believe in my ideas/skills

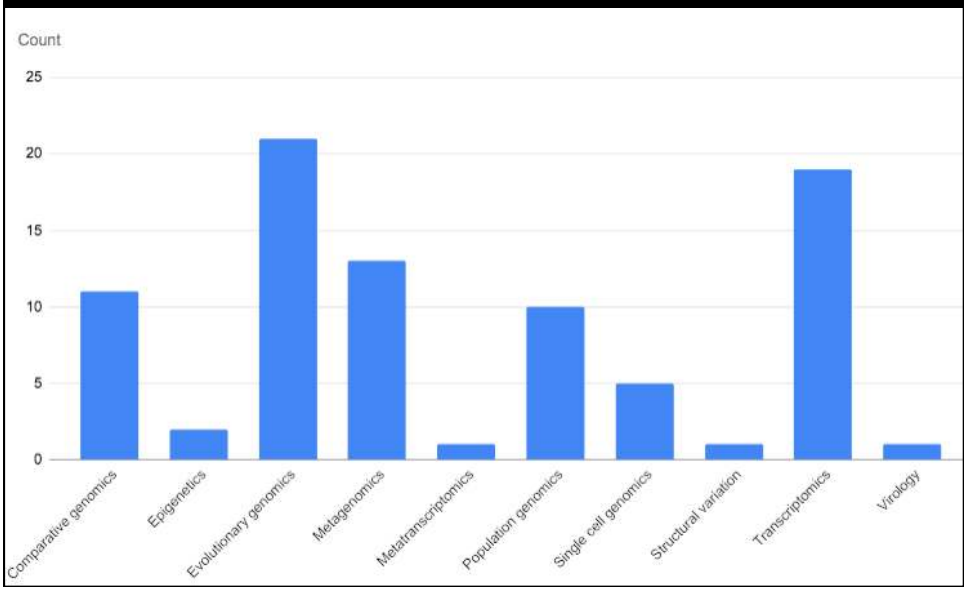
Needed to put science first, while having lots of fun along the way

Ecological & Evolutionary Functional Genomics



https://christopherwheatlab.net/

Now .... Who are you?



What do you study?

Rough totals:

- Invertebrates: 7
- Fish: 8
- Mammals: 5
- Microbes: 16
- Plants: 4
- Humans: 4

What are your goals?

- Finding and study genomic regions that matter
- Investigating ecological processes
  - metagenomics
- Investigating physiology
  - RNAseq

Goal of this lecture

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

Disclaimer

I'm a positive person

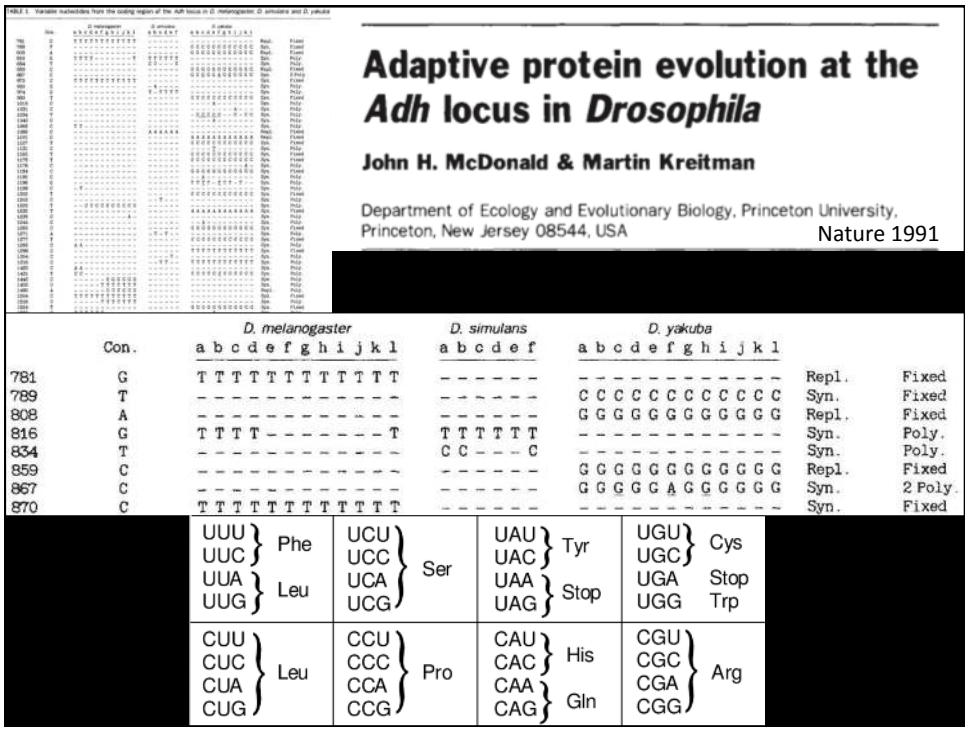
I love my job and the work we all do

I'm just sharing scrumptious food for thought

What if .....

50% of your favorite studies had conclusions that were just wrong?

How would that affect your expectations and work?



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

John H. McDonald & Martin Kreitman

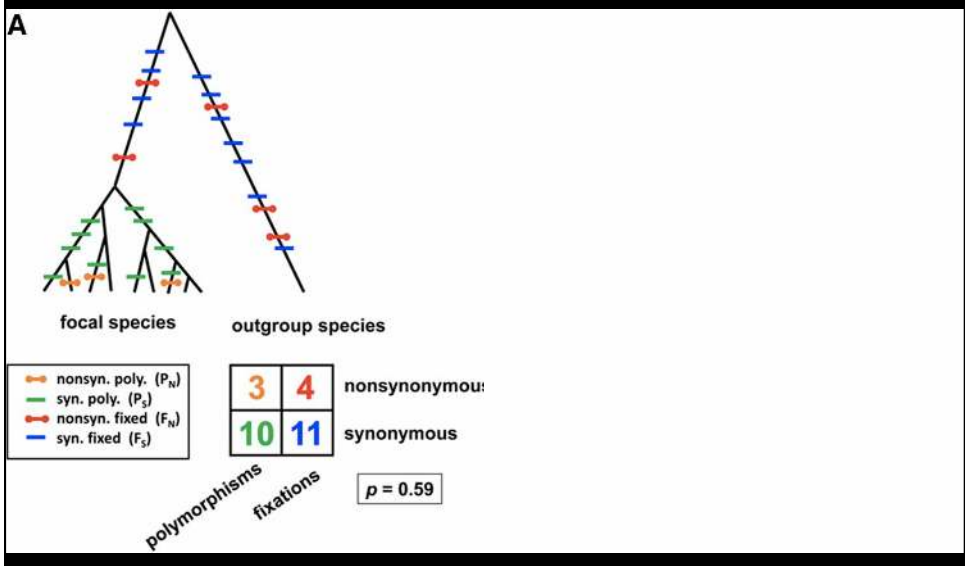
Department of Ecology and Evolutionary Biology, Princeton University, Princeton, New Jersey 08544, USA

Nature 1991

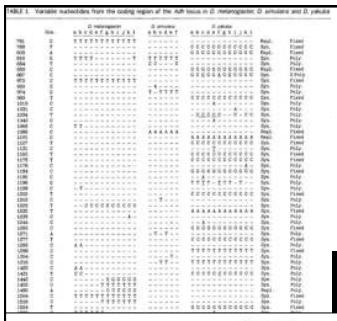
	Con.	*D. melanogaster* a b c d e f g h i j k l	*D. simulans* a b c d e f	*D. yakuba* a b c d e f g h i j k l		
781	G	T T T T T T T T T T T T	– – – – – –	– – – – – – – – – – – –	Repl.	Fixed
789	T	– – – – – – – – – – – –	– – – – – –	C C C C C C C C C C C C	Syn.	Fixed
808	A	– – – – – – – – – – – –	– – –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

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



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

D. melanogaster *D. simulans* *D. yakuba*

Con. a b c d e f g h i j k l a b c d e f a b c d e f g h i j k l

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

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

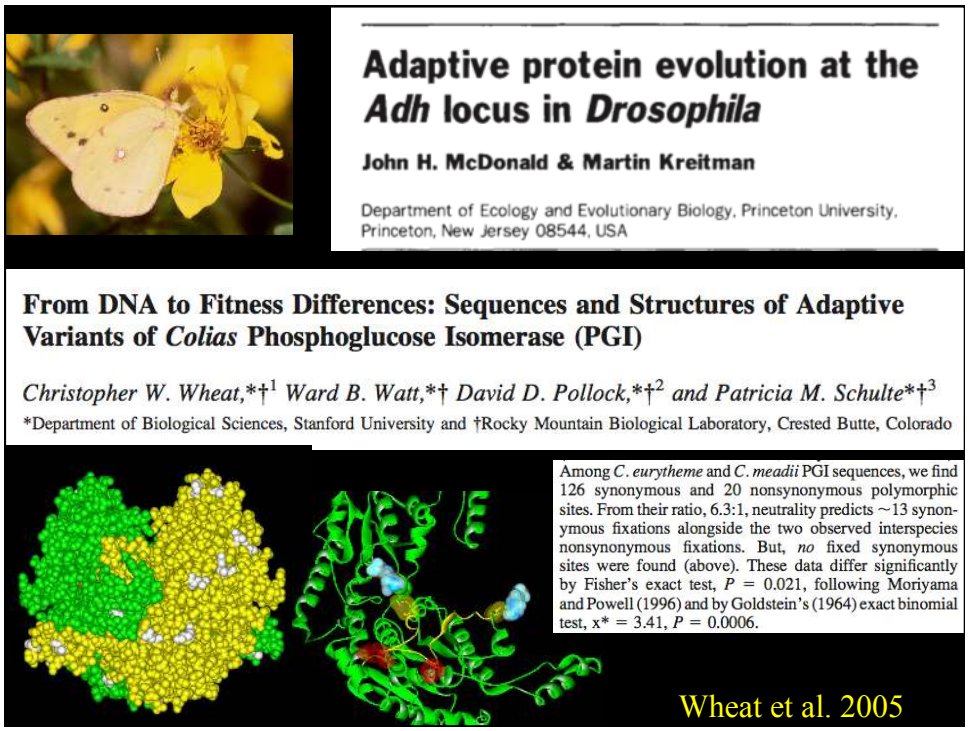
Nature 1991

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

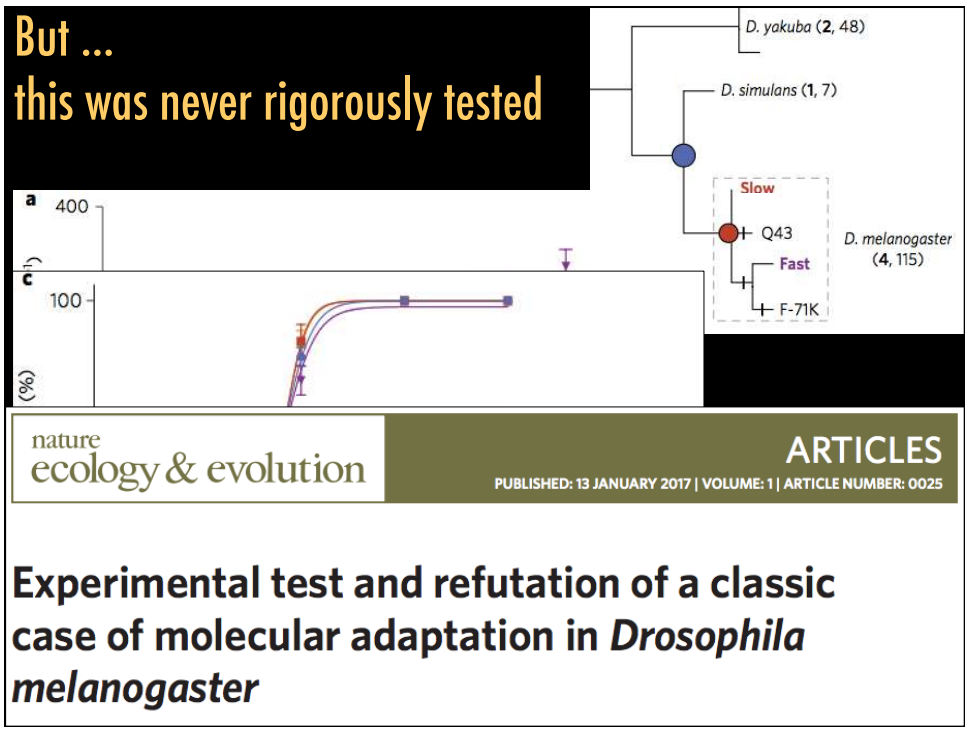
From DNA to Fitness Differences: Sequences and Structures of Adaptive Variants of *Colias* Phosphoglucose Isomerase (PGI)

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



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

Their findings should be robust:

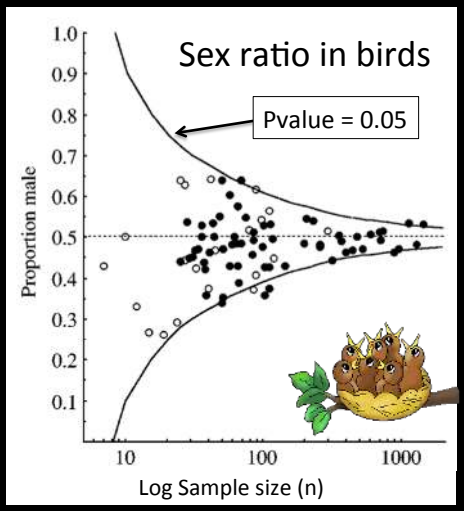
- Repeatable effect sizes
- The same across different labs
- The same across years

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



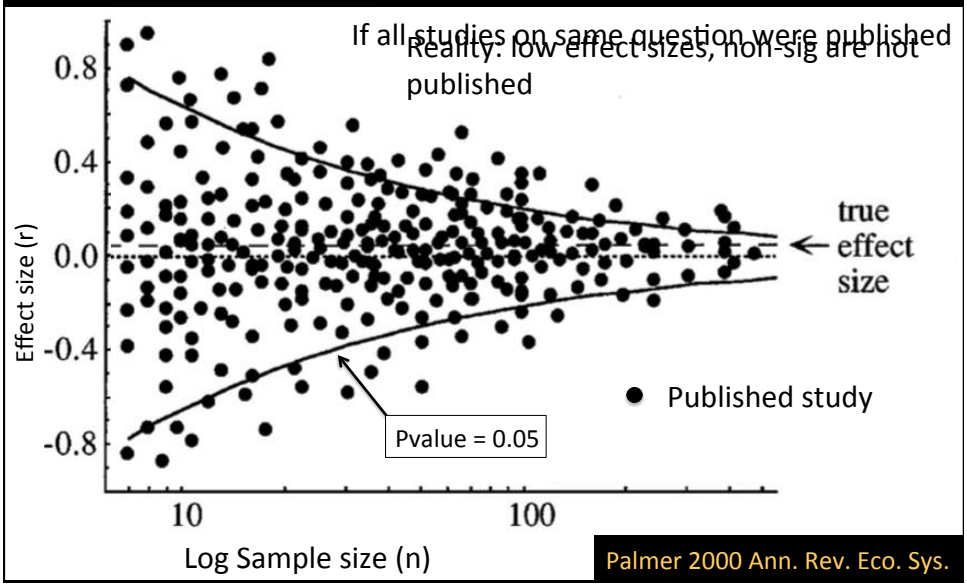
Small sample sizes affect measurement accuracy

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

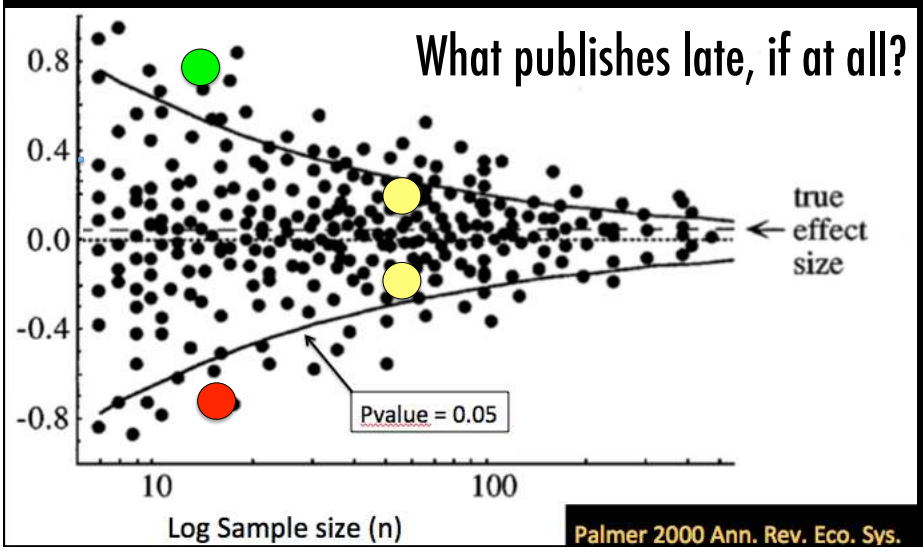
Publication bias increases effect size

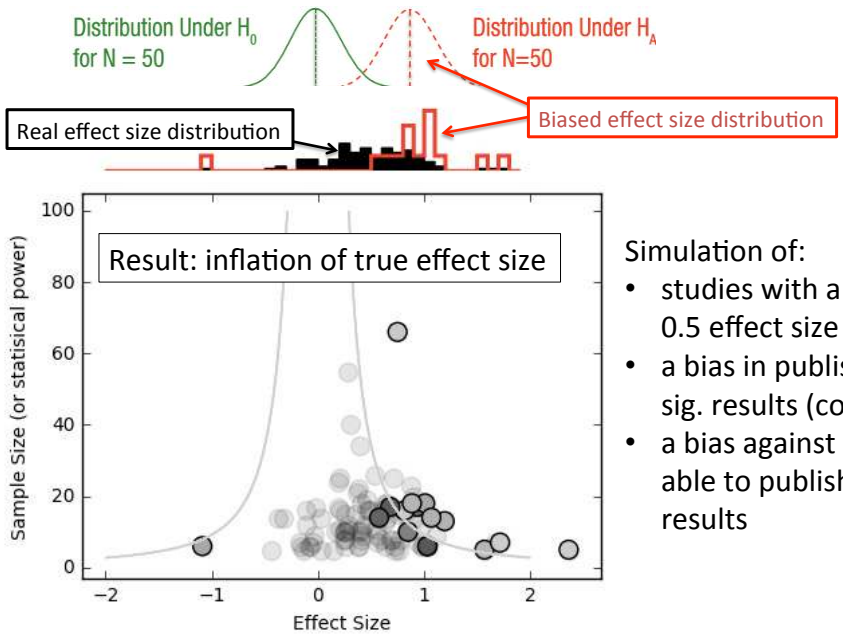


Palmer 2000 Ann. Rev. Eco. Sys.

What if there is no replication?

What is most likely to publish first & where?





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.

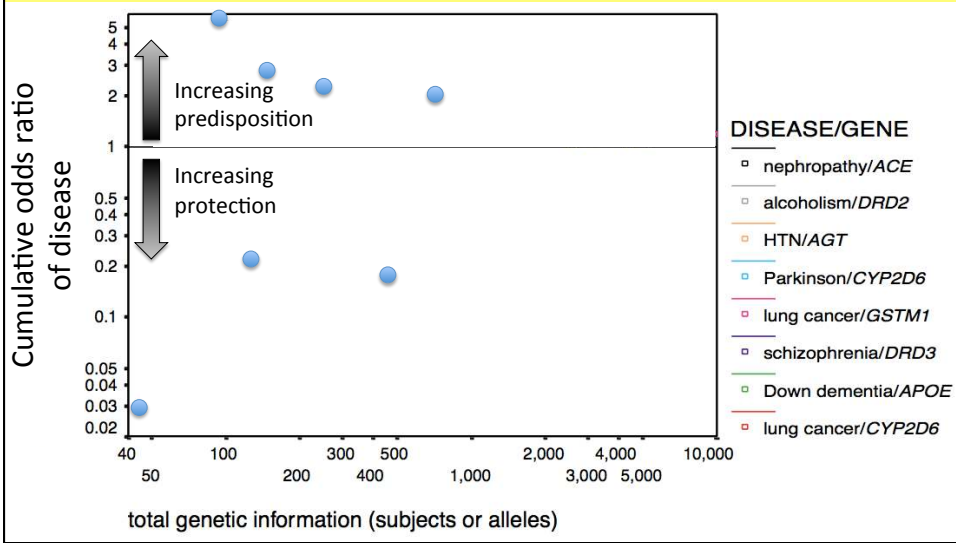
But surely, this doesn't apply to genomics ....

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 Type 1 errors

Celebration of the false positives

Genomics is too big to fail

- Making errors is extremely easy
- Results will very likely be significant, and sometimes dramatically so
- In non-model systems, rarely have replication studies
- You must always question your bioinformatics before falling in love with your results

When results are better than you could have dreamed,

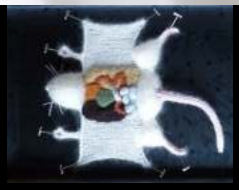
PNAS

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:

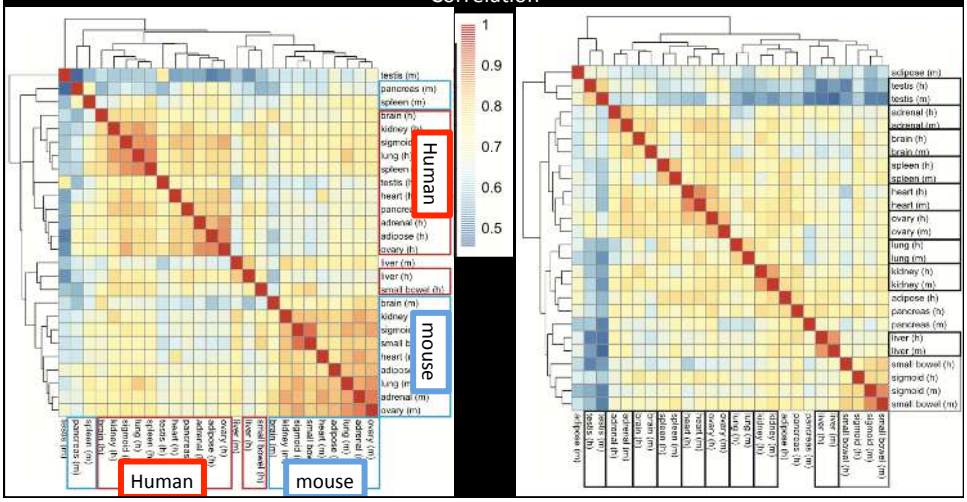
Human and Mouse



Snyder mouse controversy

"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 technical artifact called a batch effect. confounded sequencing grouping with biological grouping

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

Solution = Keep technical effects orthogonal to biological

- Mouse & Human in same lane, same tissues in same lane
- Will your Core facility know to do this for you?

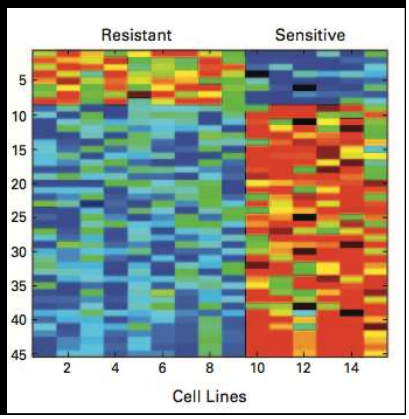
Personalized medicine: via an excel error

- 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

- Led to additional funding
  - Planned clinical trials with drugs

Hsu et al. 2007



FORENSIC BIOINFORMATICS AND REPRODUCIBLE RESEARCH IN HIGH-THROUGHPUT BIOLOGY

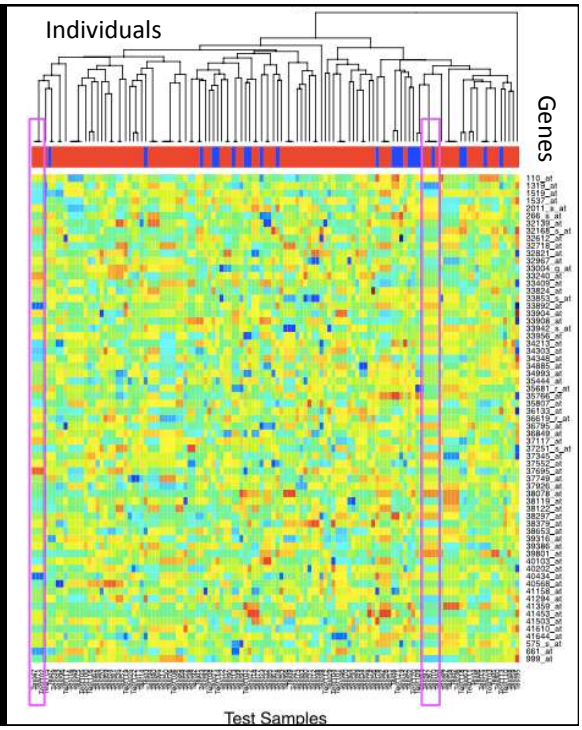
"Data processing, however, is often not described well enough to allow for exact reproduction of the results,

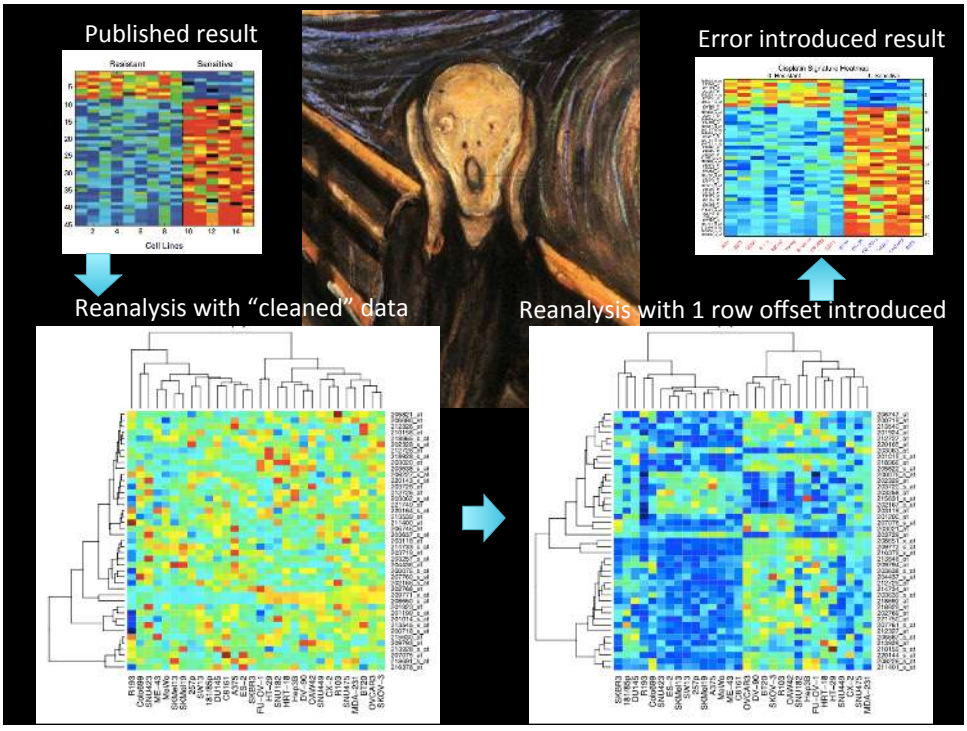
Thanks: Malachi Griffith

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

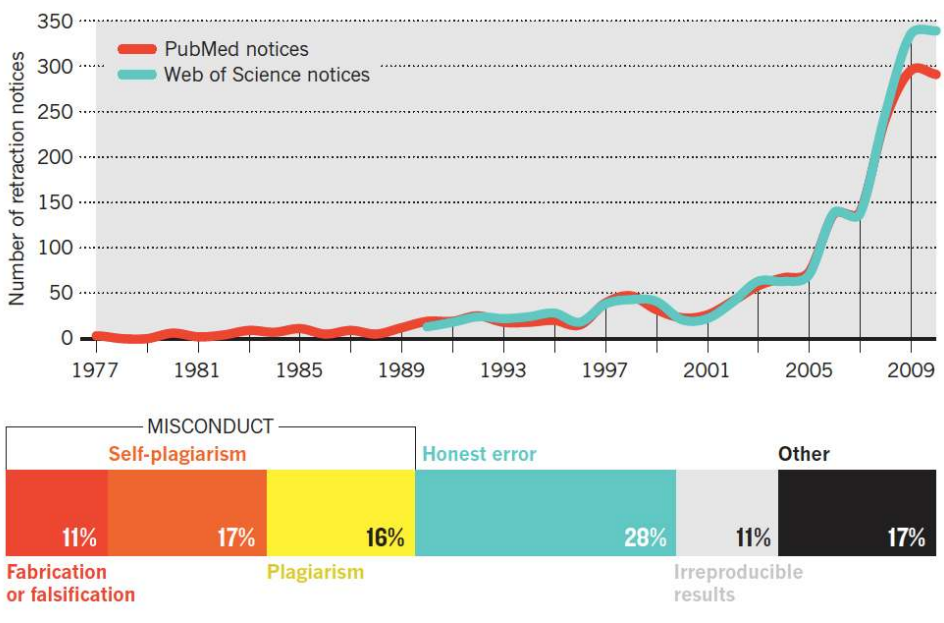




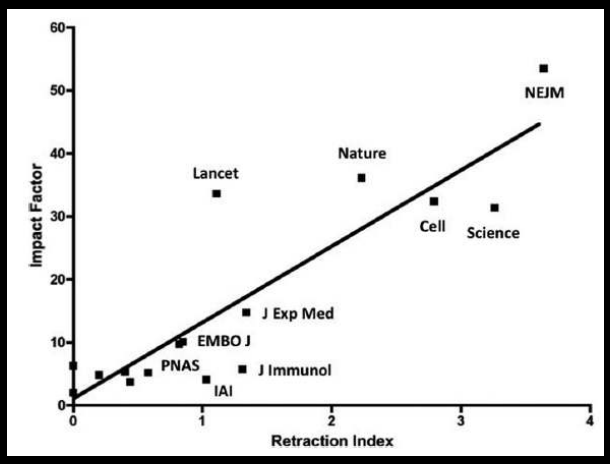
THE TROUBLE WITH RETRACTIONS

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

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 [illegible] enzymatic synthe[illegible] reproducible. scie[illegible]

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

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.

But there are lots of errors out there ...

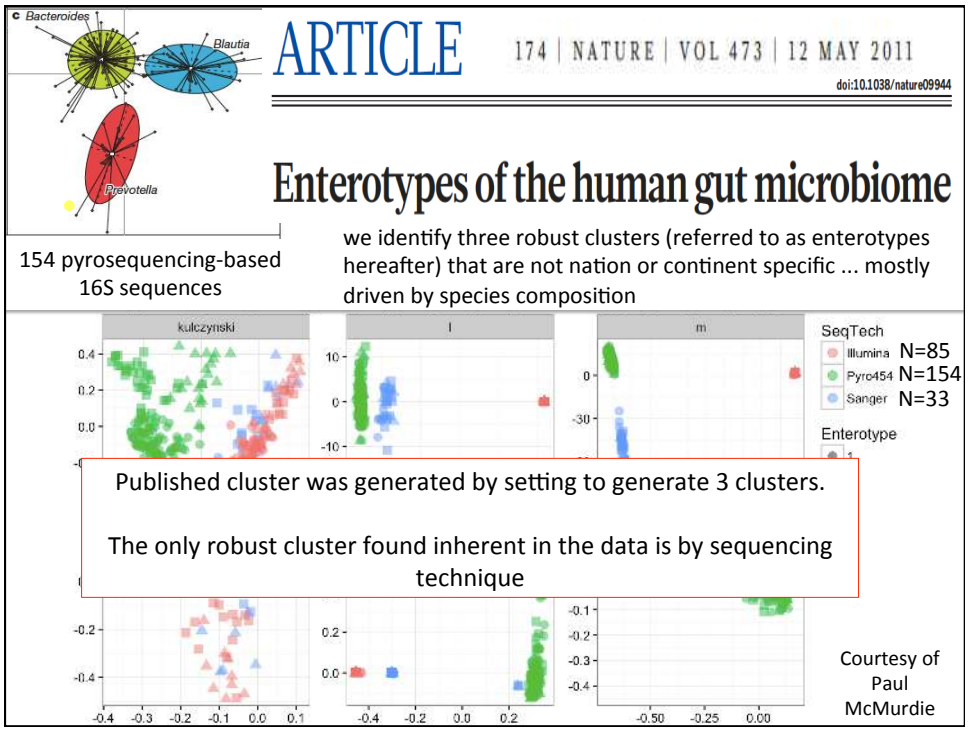
In most instances, this is scientific progress ...

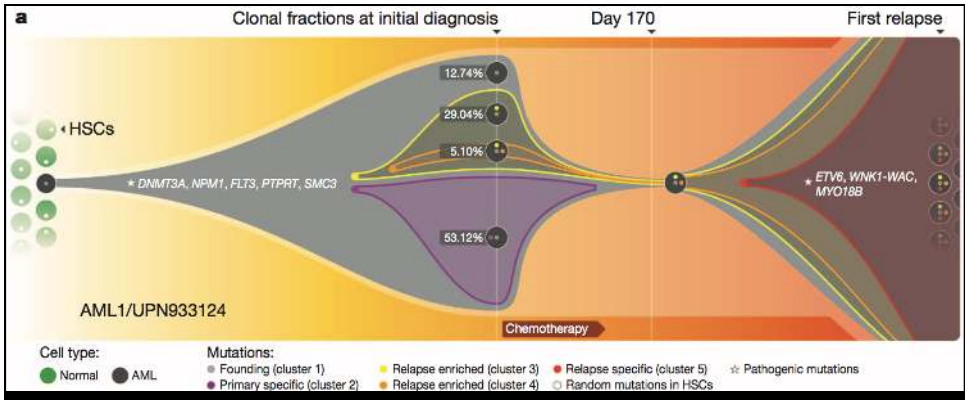
But, you must navigate these to calibrate your expections and approaches

Bioinformatics: get it right!

Can happen using the most basic tools / steps in genomics:

- Clustering of groups
- Mapping of reads against genome
- Comparative sequence alignment

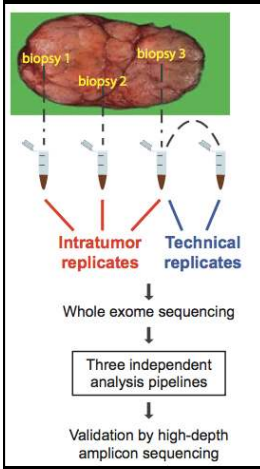




- Acute myeloid leukemia (AML) is a cancer of myeloid blood cells
  - sequencing the complete genomes of primary tumor, relapsed tumor, and matched normal (skin) samples
- AML relapse is associated with the addition of new mutations and clonal evolution, which is shaped, in part, by the chemotherapy
- AML genome in an individual patient is clearly a 'moving target'; eradication of the founding clone and all of its subclones will be required to achieve cures.

Ding et al. 2012 Nature

How well can we track early stages of relapse?



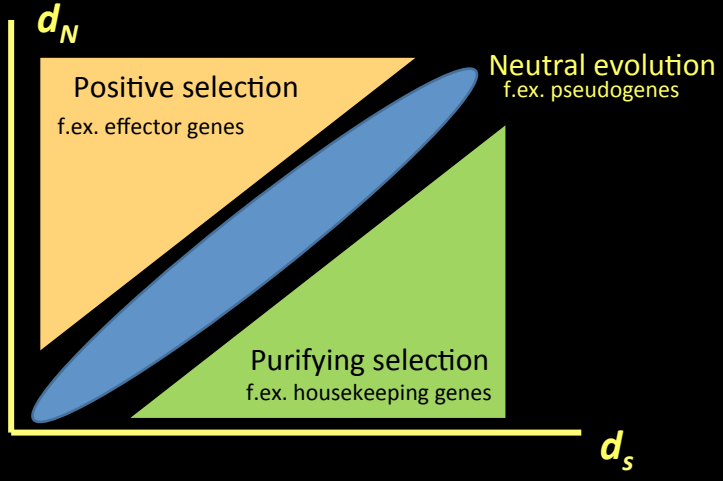
Weiwei 2018 Cell Reports

Intratumor genetic heterogeneity (ITGH)

- the coexistence of genetically distinct but clonally related cancer cells within the same patient
- 34%–80% of the discordant somatic variants, which could be interpreted as ITGH, were found to constitute technical noise
- Excluding mutations affecting low mappability regions or occurring in certain mutational contexts was found to reduce artifacts

Weiwei 2018 Cell Reports

Codon based tests of selection

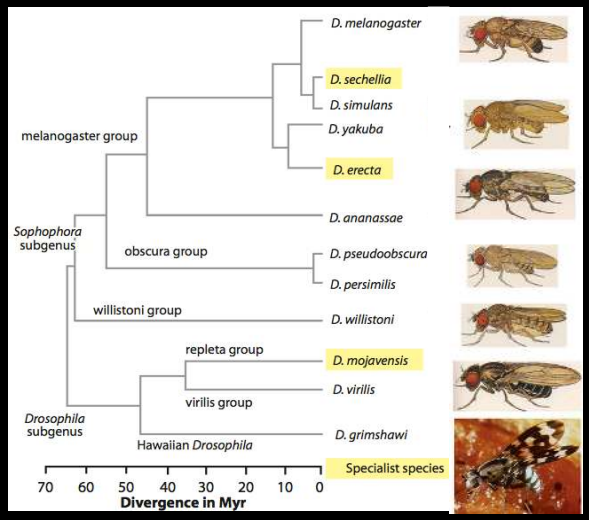


d_N / d_s ratio

```
> 1 positive sel.
= 1 neutral
< 1 purifying sel.
```

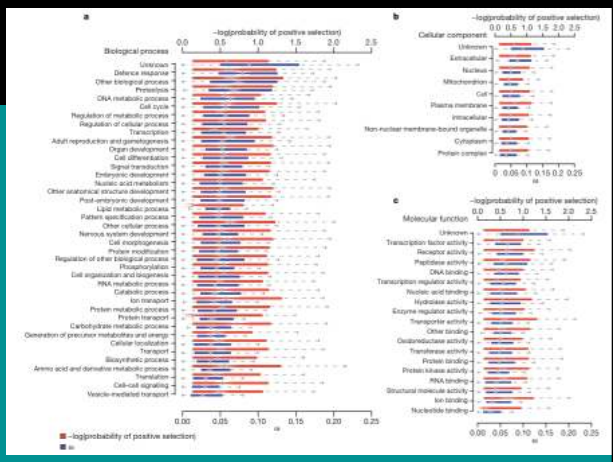
IMPRS workshop, Comparative Genomics

Evolution of genes and genomes on the *Drosophila* phylogeny



Drosophila 12 Genomes Consortium 2007 Nature

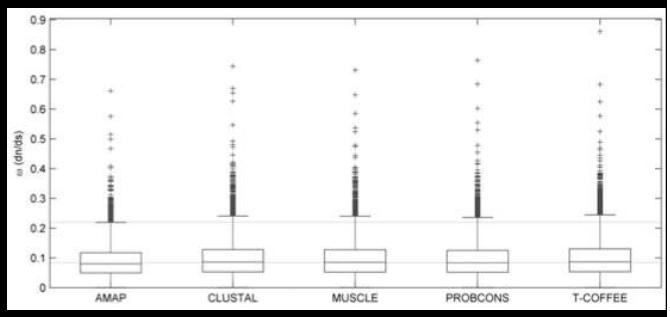
Genome-wide selection dynamics:



How robust are these conclusions?

dN/dS estimates by aligner

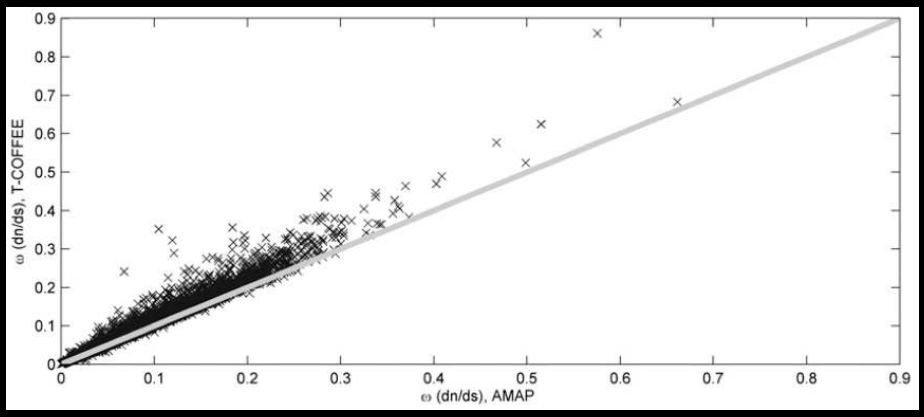
- 6690 orthologs
- 5 alignment methods
- Alignment methods affect dN/dS estimates



Markova-Raina & Petrov 2011 Genome Biology

Comparing results across methods is responsible bioinformatics!!!!!

Since we can't look at our data, we need approaches that allow 1st principal assessments



Markova-Raina & Petrov 2011 Genome Biology

Aligner has a larger effect than biological signal

Number of significant genes

	12 genomes, M7/8	
Aligner	95% (a)	99% (b)
AMAP	817	213
MUSCLE	1043	306
ProbCons	1013	281
T-Coffee	1290	479
ClustalW	902	261
Total in 5	1902	673
PRANK	468	49

Markova-Raina & Petrov 2011 Genome Biology

Alignment results highlight importance of alignment score!

- Tcoffee finds 3 selected sites indicated by arrows
- ProbCons identifies region with low alignment score, not used
- Removing these regions doesn't fix all problems (Gblocks)

Markova-Raina & Petrov 2011 Genome Biology

What makes us difference from chimps?

Is it really just 2%

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

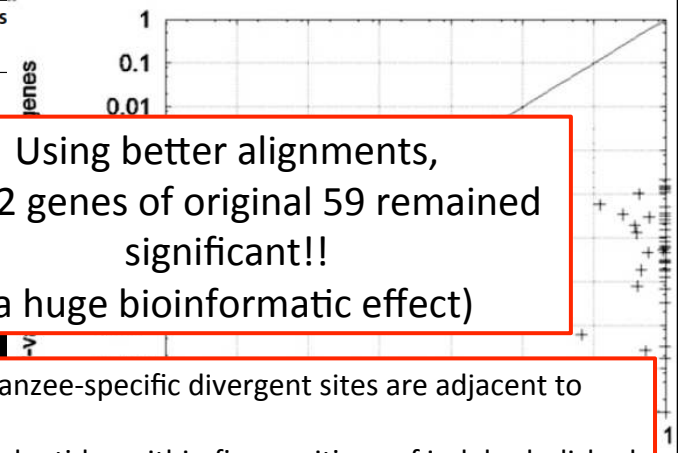
Margaret A. Bakewell, Peng Shi, and Jianzhi Zhang*

201 citations since 2007

PNAS

Table 1. Genic positive s

Comparison
No. of genes analyzed
No. of PSGs
No. of PSGs
No. of PSGs
No. of syno
No. of nons
Mean ω of
Mean ω of



Using better alignments,
only 2 genes of original 59 remained significant!!
(a huge bioinformatic effect)

- Many chimpanzee-specific divergent sites are adjacent to indels
- removing nucleotides within five positions of indels abolished most adaptive signals

How do we avoid Apophenia?

- Double check your tables and analyses
  - Plot your data, look at it, does it make sense?
- Test your hypotheses in an independent way
  - Test your findings using separate data and a different analysis
  - Functional Validation

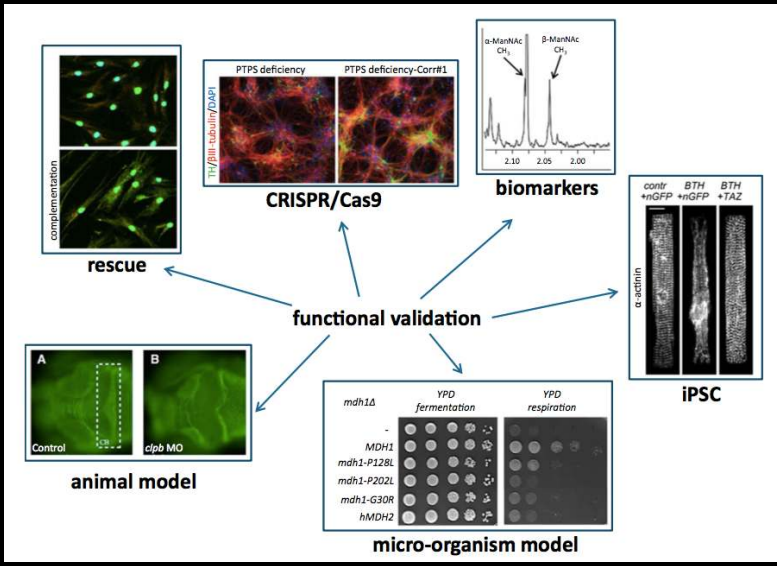
Published studies allow …

You to practice your bioinformatics

Assess their repeatability

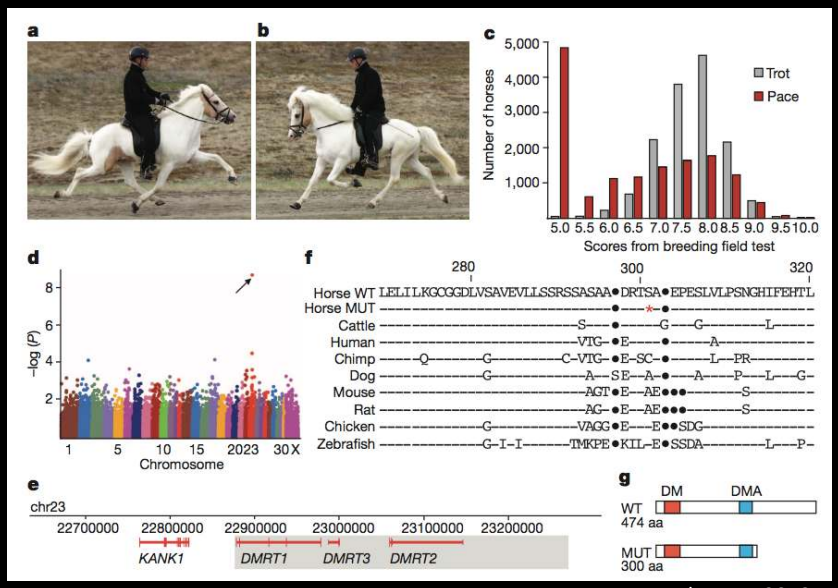
Papers need enough details for replication

Functional Validation

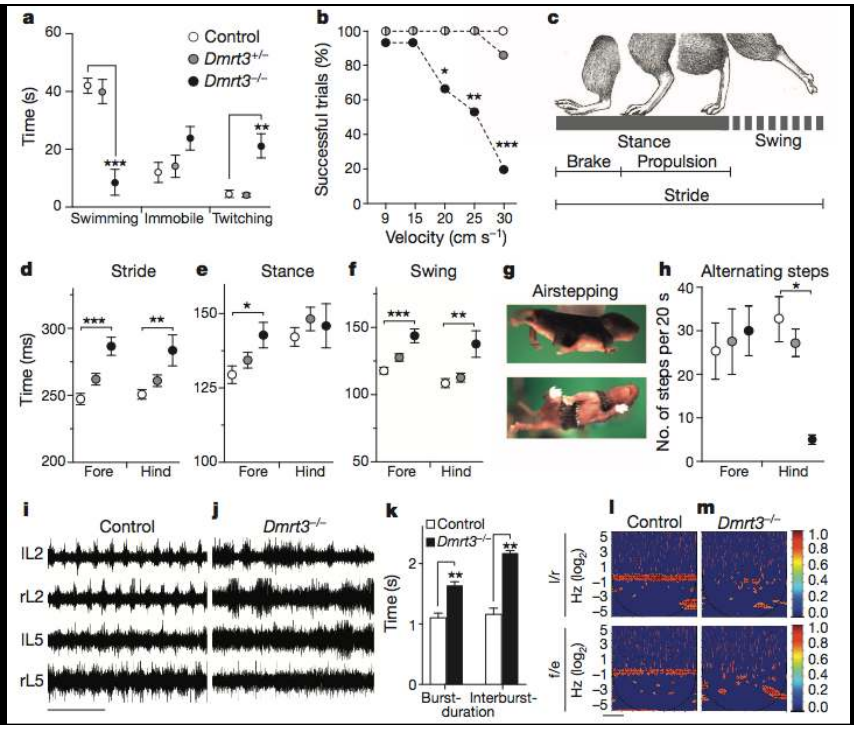


Rodenburg 2017

Gatedness in horses



Andersson 2012 Nature



On the importance of negative results

JOURNAL OF NEGATIVE RESULTS

- ECOLOGY & EVOLUTIONARY BIOLOGY -

- There is a great need, and little incentive to publish negative results
- How can we change this?
  - Free publication charges
  - Change the name from negative to .... ?
  - ????