

# **Lies, damn lies, and .... genomics**

**you, your data, your perceptions and  
reality**

**Christopher West Wheat**



## **Goal of this lecture**

- **Present a critical view of ecological genomics**
- **Make you uncomfortable by sharing my nightmares**
- **Encourage you to critically assess findings and your expectations in light of publication biases**

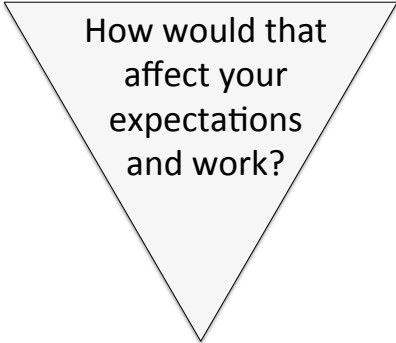
# Disclaimer

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 .....



How would that  
affect your  
expectations  
and work?



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

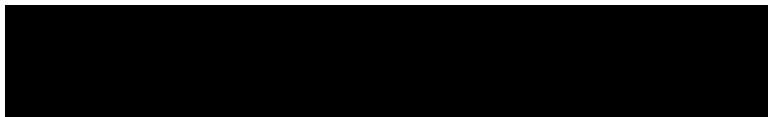
If the biomedical science has the most money and oversight, then ....

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 clinical studies, 45 showed intervention was effective
  - Most were randomized control studies (robust design)

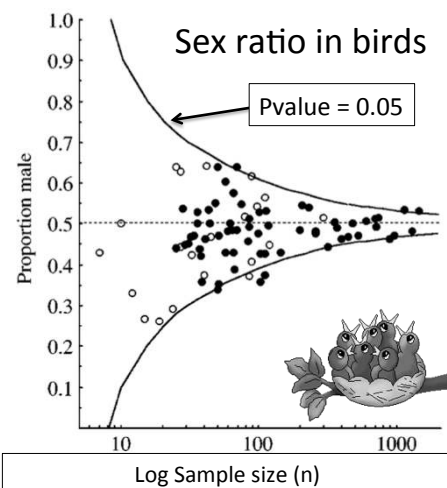


- Mouse cocaine effect study, replicated in three cities
  - Highly standardized study



Ioannidis 2005 JAMA; Lehrer 2010

## Assessing reality using funnel plots



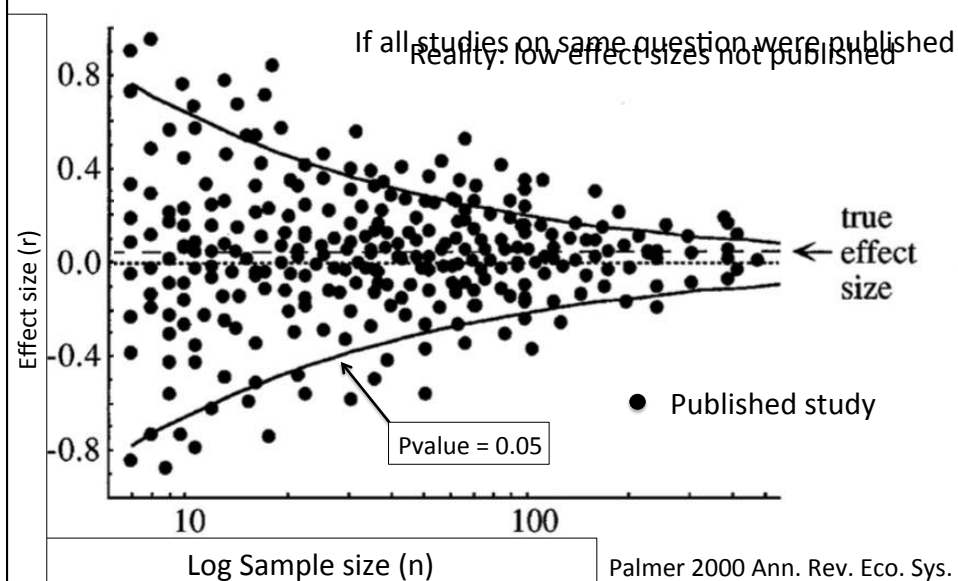
Small sample sizes affect measurement accuracy

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

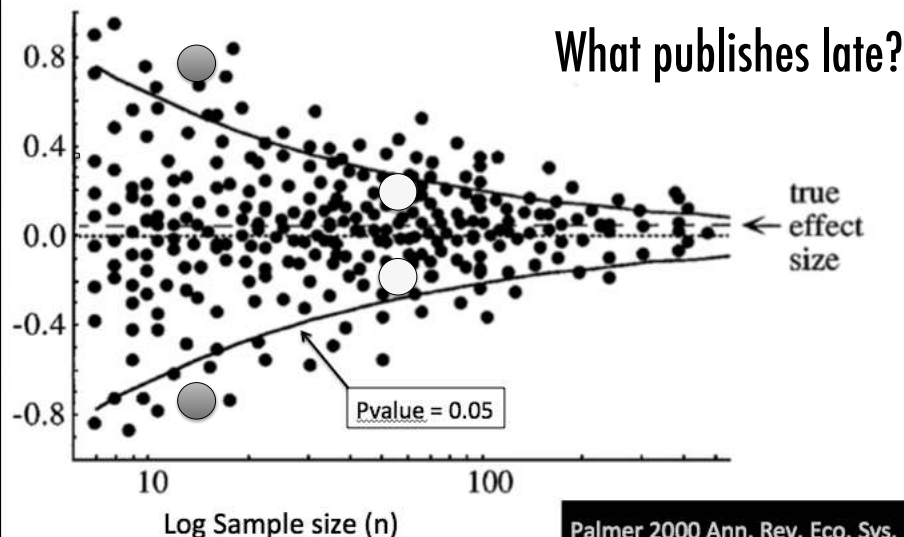
## Publication bias increases effect size



## What if there is no replication?

What is most likely to publish first & where?

What publishes late?



## Why Most Published Research Findings Are False

A research finding is less likely to be true when:

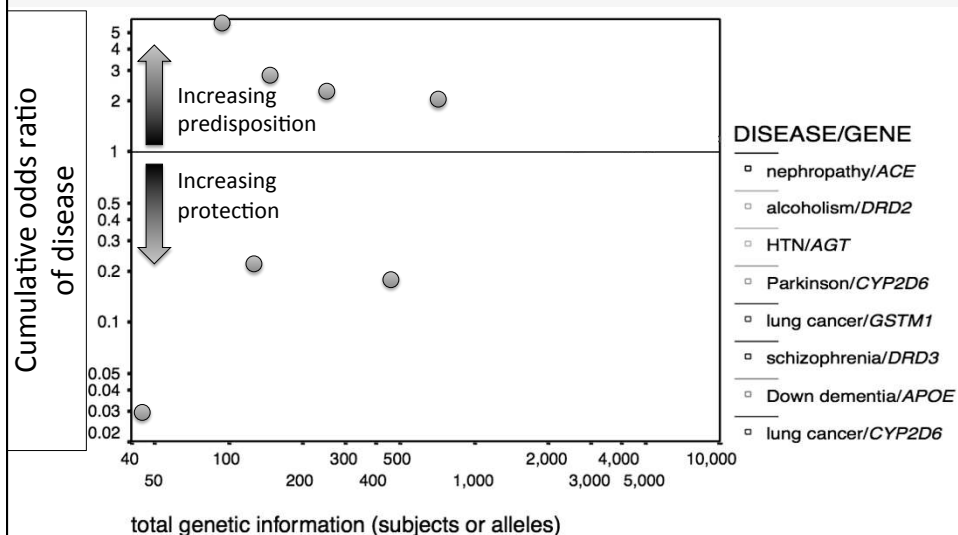
- ✓ the studies conducted in a field have a small sample size
- ✓ when effect sizes are small
- ✓ when there are many tested relationships using tests without *a priori* selection
- ✓ where there is greater flexibility in designs, definitions, outcomes, and analytical modes
- ✓ when there is greater 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?

## 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
  -
- Publication bias
  - 
  - 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

A universal human tendency to seek patterns in random information and view this as important



Story telling of Type 1 errors

Celebration of the false positives

# Outline

- Are there biases understanding the genomic architecture of adaptations?
- What is the power of molecular tests of selection?
- What does the dissection of some classic comparative genomics study reveal?

## Metabolic Pathways

**How do we find the genes that matter?**

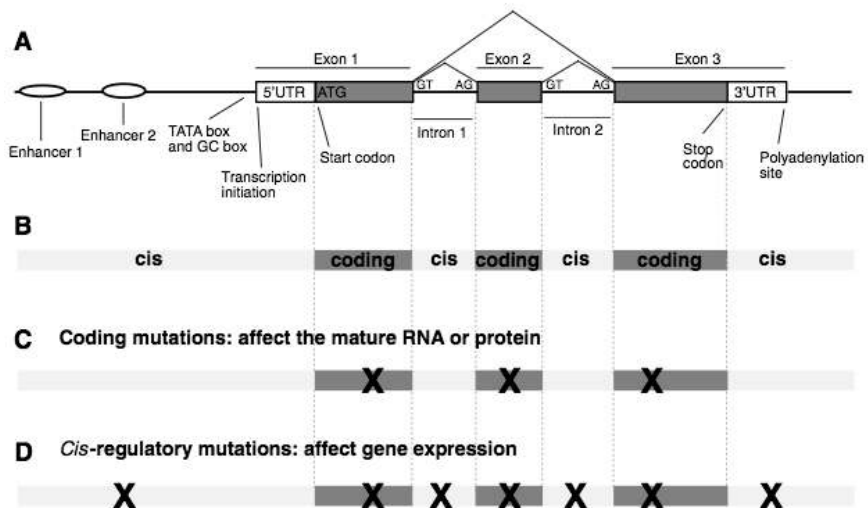
**Publications using molecular tests demonstrate we can sequence our way to answers**

**Current paradigm:**

**Sequence, map, find sig. patterns, make causal story, move on**

.....

## What is the architecture of a causal variant?



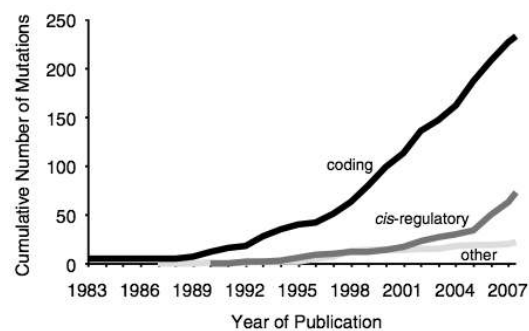
## What type of variant?

– SNP, indel, TE, inversion, CNV?

Stern & Orgogozo 2008 Evolution

## How predictable are adaptations?

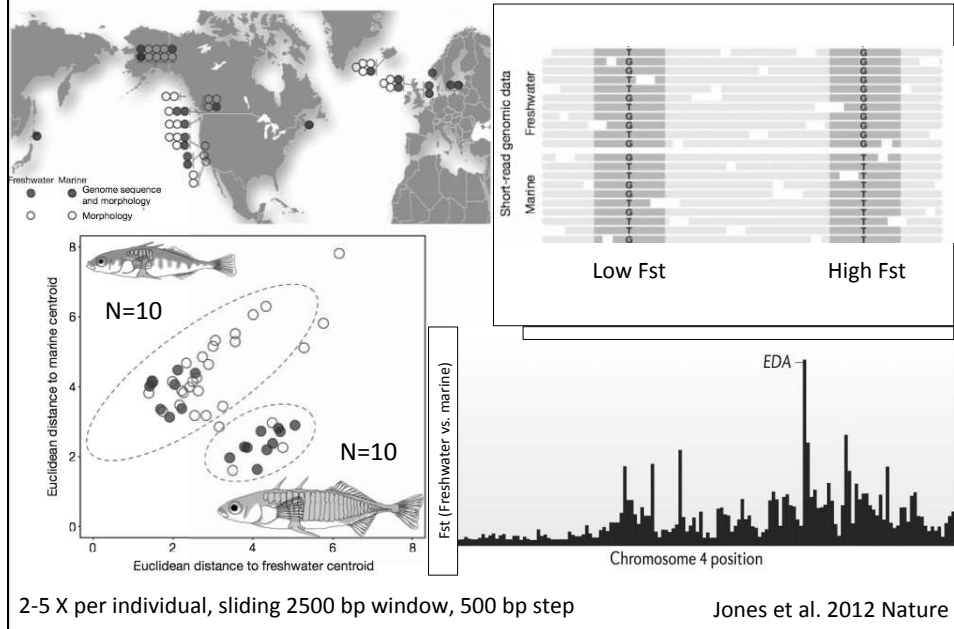
|                     | Plants | Animals |
|---------------------|--------|---------|
| Coding <sup>1</sup> | 71     | 163     |
| Cis-regulatory      | 26     | 48      |
| Other <sup>2</sup>  | 16     | 7       |
| Total               | 113    | 218     |
| Null <sup>3</sup>   | 67     | 32      |



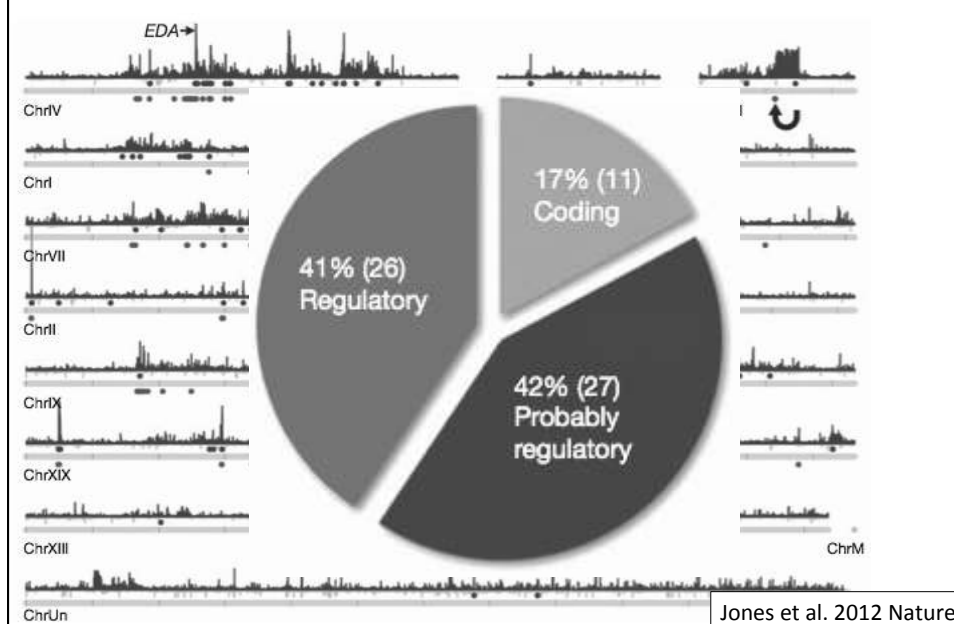
|                     | Morphology | Physiology | Behavior |
|---------------------|------------|------------|----------|
| Coding <sup>3</sup> | 62         | 170        | 2        |
| Cis-regulatory      | 43         | 29         | 2        |
| Other <sup>4</sup>  | 3          | 20         | 0        |
| Total               | 108        | 219        | 4        |
| Null <sup>5</sup>   | 41         | 58         | 0        |

Stern & Orgogozo 2008 Evolution

## Individual genome sequencing: powerful insights

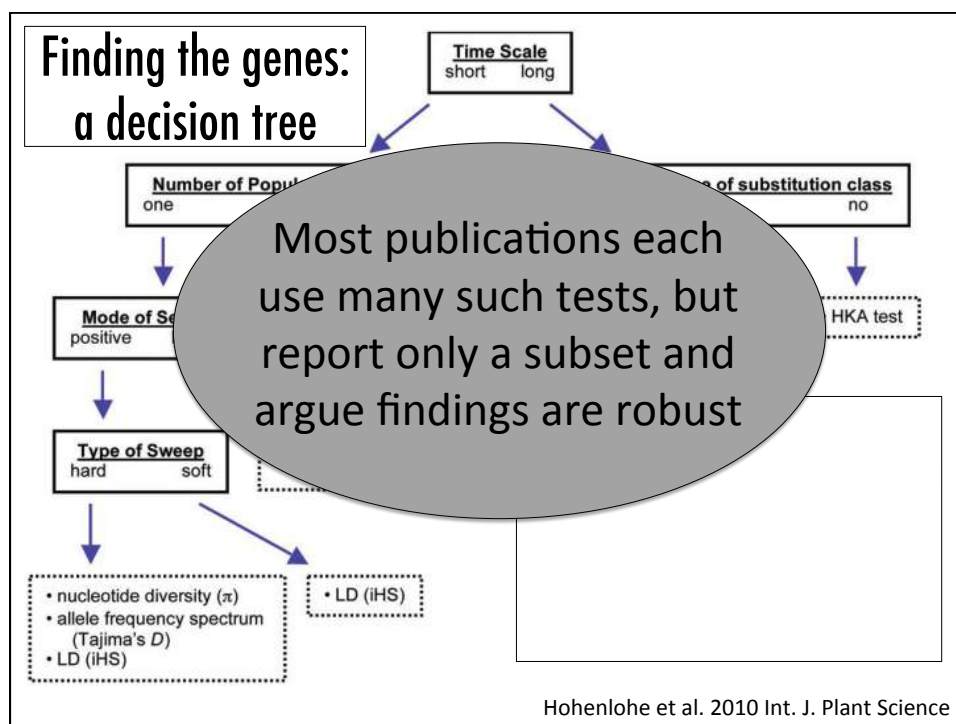


## Which regions are more important? Coding or expression?



## How do we identify the genes that matter?

- Molecular tests of selection are popular, but ...
  - What are their assumptions and power?
- What are these tests detecting?
  - What is a footprint of selection?
    - How are they formed?
    - How large are they?
    - How long do they last?

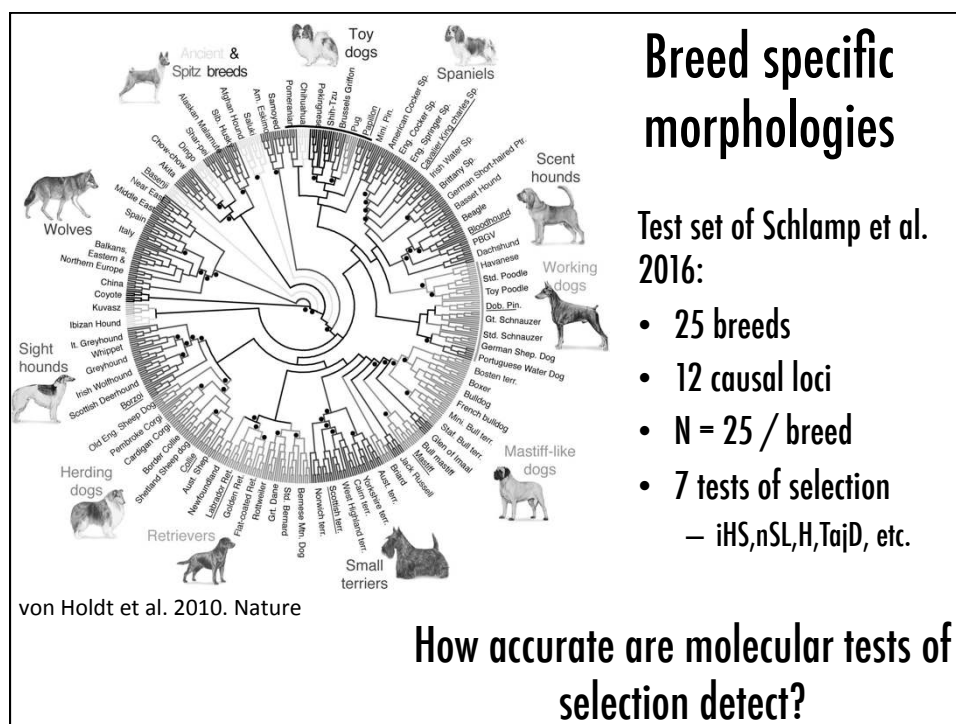


What power do we  
have to detect  
evolution by  
natural selection?

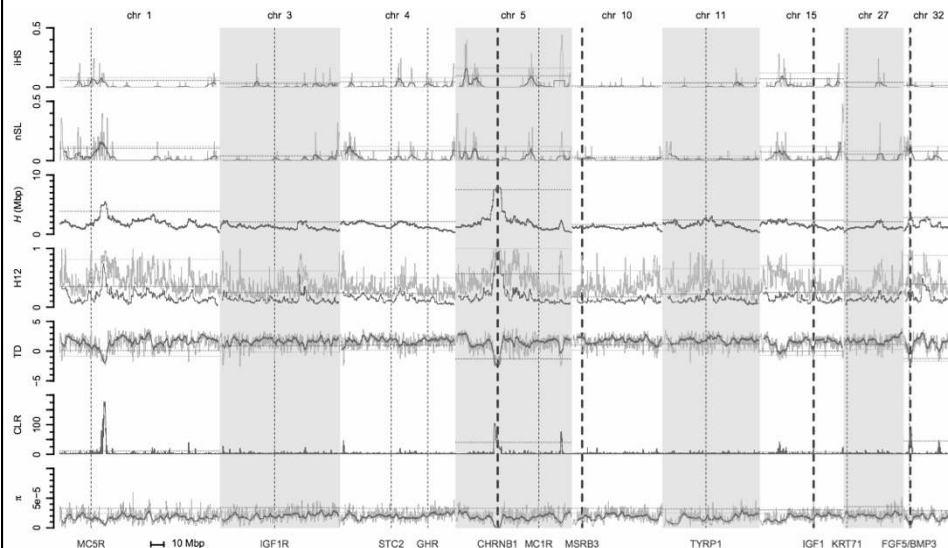
What is  
statistical  
power?

Power is the probability that the test will reject the  
null hypothesis when the alternative hypothesis is  
**TRUE**

Using a t-test, you would want power  $> 90\%$  at  
reasonable sample size, right?



## French Bulldog sample: low power (high type II error)



Schlamp et al. 2016. Evaluating the performance of selection scans to detect selective sweeps in domestic dogs. *Molecular Ecology* 25:342–356.

# Why don't these tests have much power?

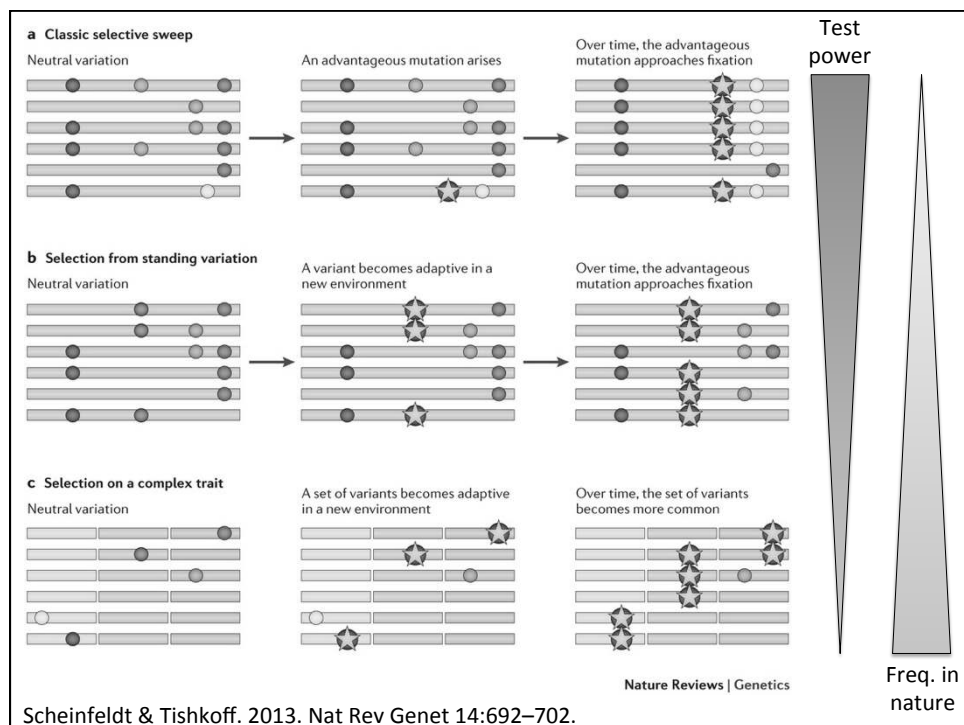
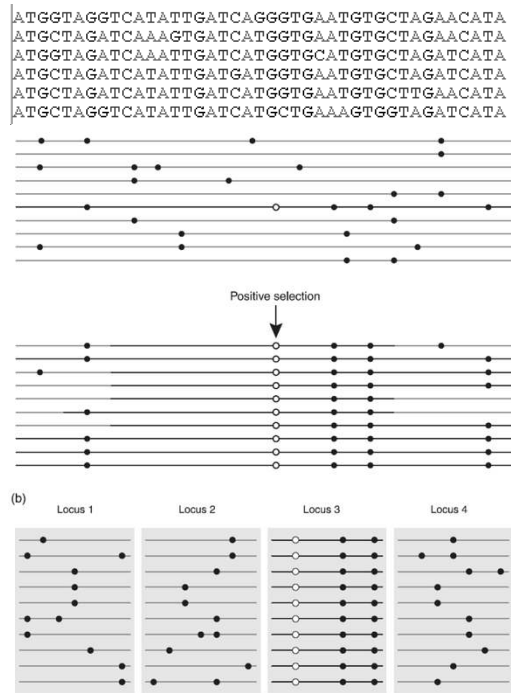
Biological reality  
vs.  
theoretical population genetics?

## Directional selection: an example of the expectations of hard selection

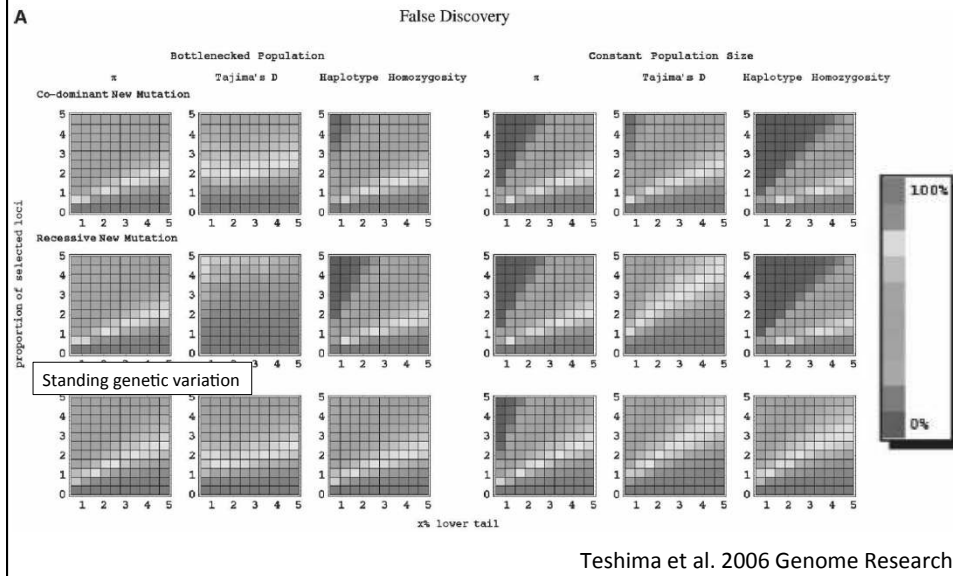
Population genomics has been dominated by developing methods to detect hard sweeps for past two decades

- But a proper 'null model' continues to be elusive, resulting in a high false positive rate since their inception

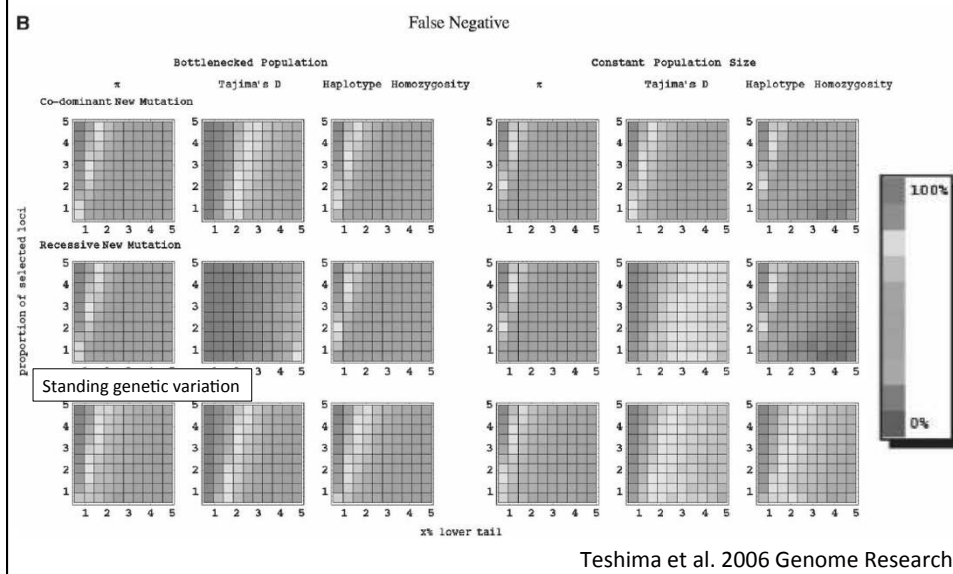
Storz 2005 Mol. Ecology



## Estimate of error rates using , Tajima's D, and haplotype homozygosity under the models for a human population



## Estimate of error rates using , Tajima's D, and haplotype homozygosity under the models for a human population



## Simulation conclusions

- Simulations suggest
  - empirical approaches will identify several interesting candidates
  - But will also miss many—in some cases, most—loci of interest
- False-discovery rate is higher when
  - directional selection involves a recessive rather than a co-dominant allele
  - when it acts on a previously neutral rather than a new allele
  - Demographic size changes rather than constant population size

Genomic scans yield an unrepresentative subset of loci that contribute to adaptations

## Molecular tests ...

BASED ON 20 YEARS OF PUBLICATIONS

- Are still chasing an elusive null model ....
  - Each performs better than previous ones under a specific set of conditions, all have poor null model
- But ... under realistic biological conditions, they all
  - Have very low power (high type II error rates)
  - Have high false positive rates

## How common are hard sweeps in nature?

- “we argue that soft sweeps might be the dominant mode of adaptation in many species”  
Messer and Petrov 2013 TREE

## The lab?

- “Signatures of selection ... [are] not associated with ‘classic’ sweeps ... More parsimonious explanations include [selection on standing variation]”  
Burke et al. 2010 Nature

## How common were hard sweeps in our history?

- “classic sweeps were not a dominant mode of human adaptation over the past 250,000 years”
- “much local adaptation has occurred by selection acting on existing variation rather than new mutation”  
1000 Genomes PC 2010 Science  
Hernandez et al. 2011 Science

## Certainly not everyone agrees ....



### REVIEW

Received 24 Mar 2014 | Accepted 17 Sep 2014 | Published 27 Oct 2014

DOI: 10.1038/ncomms6281

### On the unfounded enthusiasm for soft selective sweeps

Jeffrey D. Jensen<sup>1,2</sup>

- This is an important read, critical of
  - assumptions underlying soft sweep (selection on standing variation)
  - the low power of molecular tests to detect hard & soft sweeps

## How likely does natural selection use standing variation in your species?

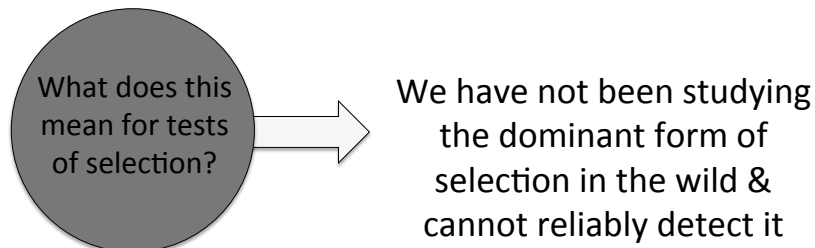
### Thought experiment:

What fraction of species respond to selection in the lab? ██████████

Why? ████████████████████

If populations have variation, how likely is selection to use it? ██████████

What's likelihood of selection on standing variation in wild? ██████████



## Age and type of selection matters

- Novel mutation, large effect, hard sweep that goes to fixation
  - Probability of detection 20 – 90%, depending on demography, etc.
- Old mutation and / or polygenetic that does not sweep to fixation
  - Probability of detection close to 0
- Finding the causal mechanism
  - Coding > expression (but allele specific expression can be lightening rod for expression)
  - SNPs > more complex mutations (indel, TE, CNV)
  - Ongoing gene flow & grouping by phenotype across replicate populations helps a lot
- Demographic effects
  - Nearly all species have experienced a major demographic change in the past 10,000 generations
  - Demographic change significantly reduces power and increases false positive rates.
- What is the relative frequency of these?
  - What will be the architecture of your phenotype?
  - What does your method have the highest power to detect?

Get ready, here come the  
1000<sup>n</sup> genomes

- Roughly 20 arthropods sequenced to date  
— plans to sequence many more
- Many other large genomes being sequenced

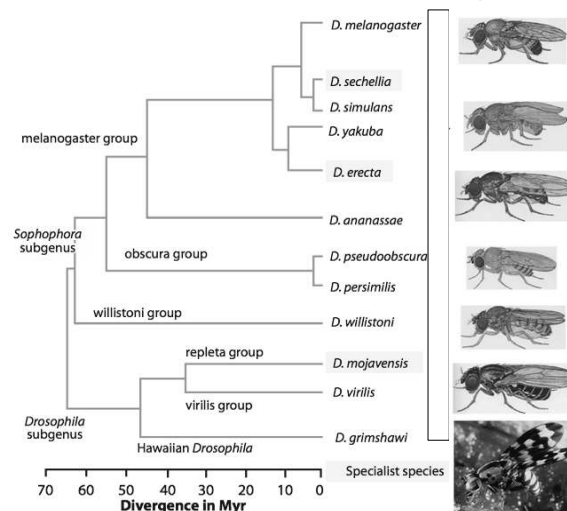
An unprecedented opportunity for large scale errors?

Studying:  
— Evolutionary relationships  
— Genome evolution  
— Functional insights into genes and genomic features (e.g. regulation and inheritance)



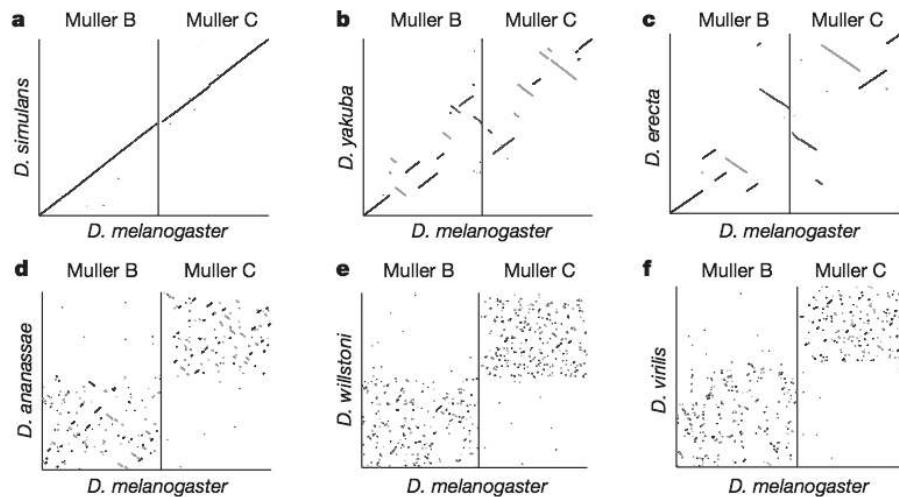


## Classic study: Evolution of genes and genomes on the *Drosophila* phylogeny



Drosophila 12 Genomes Consortium 2007 Nature

## Tempo and mode of chromosome evolution

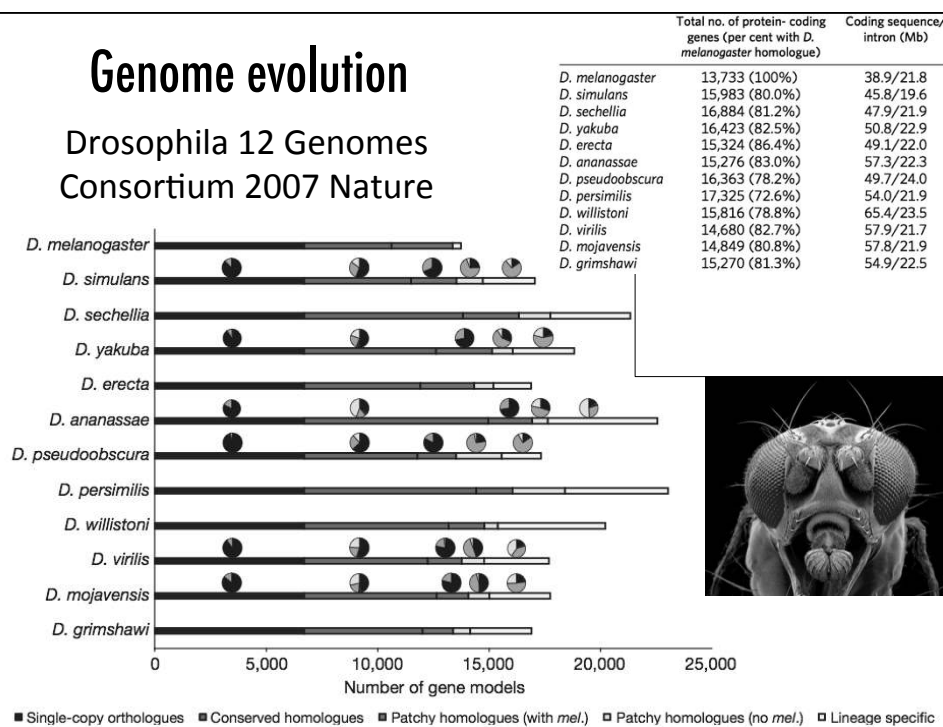


- > 20 My, chromosomal order completely reshuffled in Diptera

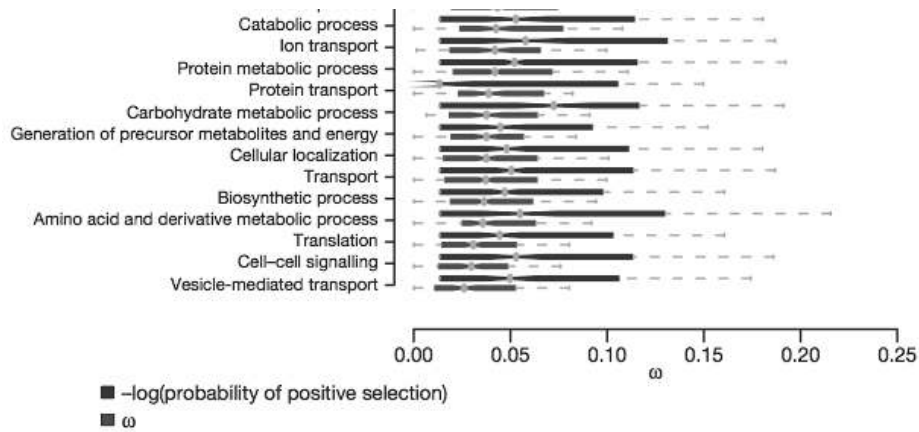
Drosophila 12 Genomes Consortium 2007 Nature

## Genome evolution

Drosophila 12 Genomes Consortium 2007 Nature



## Selection dynamics across functional categories

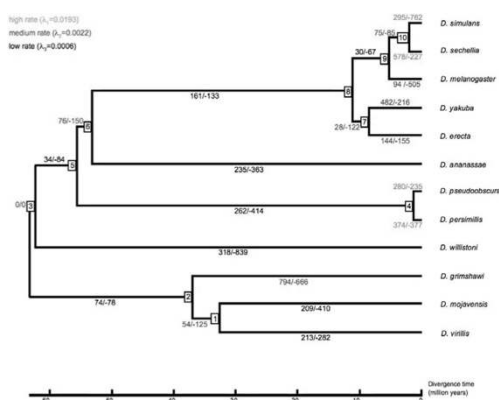


- 33.1% of single-copy orthologues have experienced positive selection on at least a subset of codons.

Drosophila 12 Genomes Consortium 2007 Nature

## Gene Family Evolution across 12 Drosophila Genomes

- One fixed gene gain/ loss across the genome every 60,000 yr
- 17 genes are estimated to be duplicated and fixed in a genome every million years



Drosophila 12 Genomes Consortium 2007 Nature  
Hahn et al. 2007 Plos Genetics

## Comparative Genomics : a house of cards?

- Data scale is too large to thoroughly assess errors ...
  - Perhaps the findings are just .... wrong
- All conclusions, at some stage, rest upon
  - Simple bioinformatics
  - Assumptions that get incorporated into seemingly unbiased methods



Lets exploring two pillars of these studies, their error and repercussions

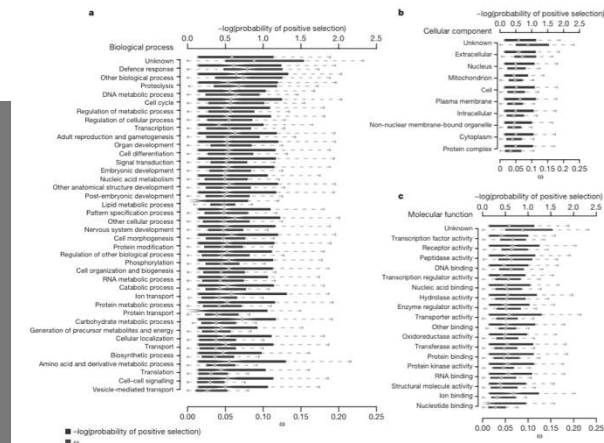
- Gene alignments in detecting positive selection
- Calibrations in temporal analysis

## Published studies allow ...

Follow up studies to reveal limitations

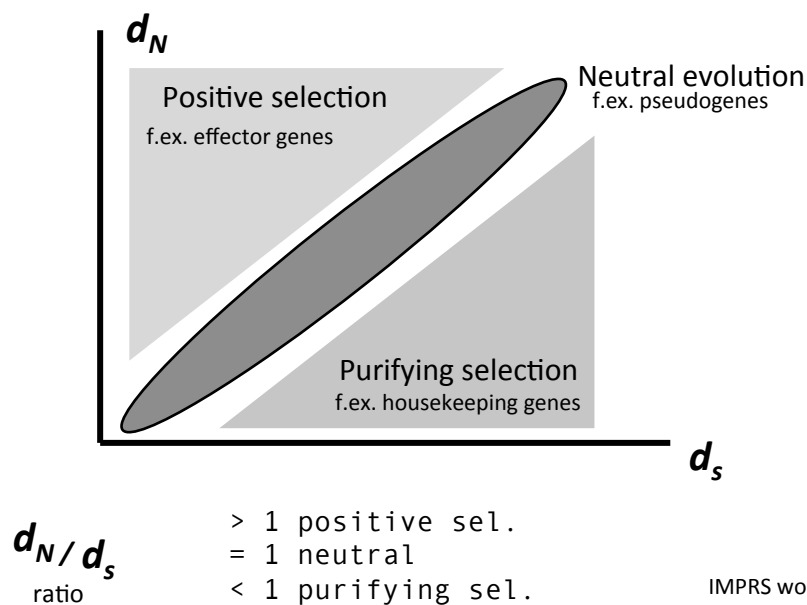
But, must have enough details to be repeatable

# Genome-wide selection dynamics:



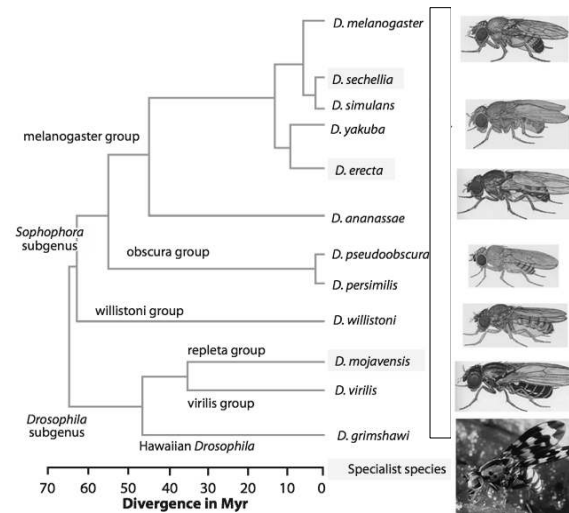
How robust are these conclusions?

## Codon based tests of selection



IMPRS workshop,  
Comparative Genomics

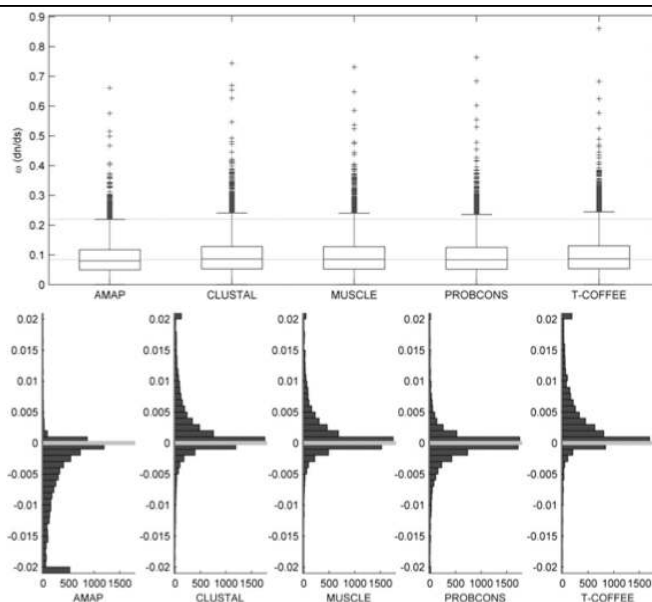
## Evolution of genes and genomes on the *Drosophila* phylogeny



Drosophila 12 Genomes Consortium 2007 Nature

## dN/dS estimates by aligner

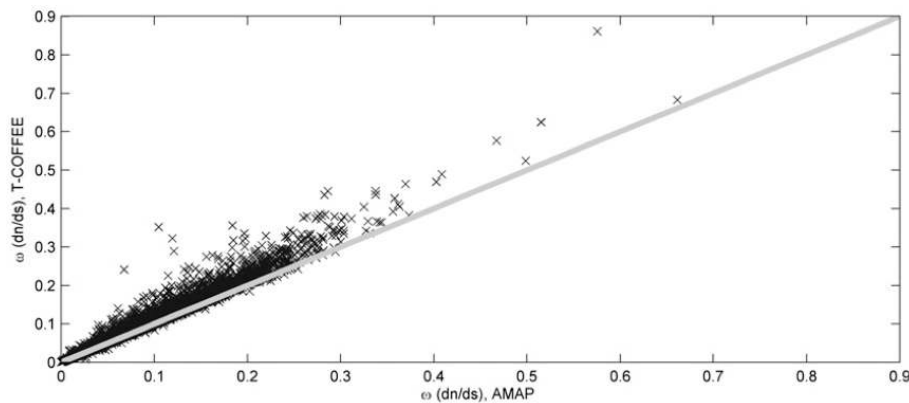
- 6690 orthologs
- 5 alignment methods
- Alignment methods affect dN/dS estimates



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 1<sup>st</sup> principal assessments

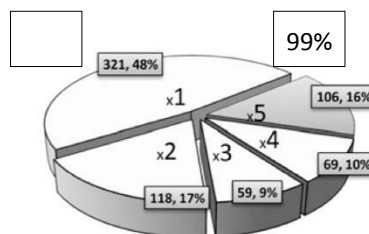


Markova-Raina & Petrov 2011 Genome Biology

## Aligner tool has a larger effect than biology

| Aligner    | 12 genomes, M7/8 |         | 12 genomes, M1a/2a |         | 12 genomes, M7/8, with removed gaps |         | Melanogaster group, M7/8 |           |
|------------|------------------|---------|--------------------|---------|-------------------------------------|---------|--------------------------|-----------|
|            | 95% (a)          | 99% (b) | 95% (c)            | 99% (d) | 95% (e)                             | 99% (f) | 95% (g)                  | 99% (h)   |
| AMAP       | 817              | 213     | 256                | 110     | 558                                 | 104     | 973                      | 257       |
| MUSCLE     | 1043             | 306     | 379                | 192     | 764                                 | 155     | 1134                     | 366       |
| ProbCons   | 1013             | 281     | 346                | 180     | 801                                 | 182     | 1128                     | 371       |
| T-Coffee   | 1290             | 479     | 612                | 353     | 824                                 | 173     | 1248 (909)               | 463 (218) |
| ClustalW   | 902              | 261     | 244                | 117     | 666                                 | 112     | 1269                     | 453       |
| Total in 5 | 1902             | 673     | 799                | 441     | 1562                                | 384     | 1737 (1723)              | 652 (620) |
| PRANK      | 468              | 49      | 49                 | 16      | 258                                 | 42      | 581                      | 70        |

Number of significant genes in common across 1, 2, 3, 4, or all 5 of the alignment methods



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



Markova-Raina & Petrov 2011 Genome Biology

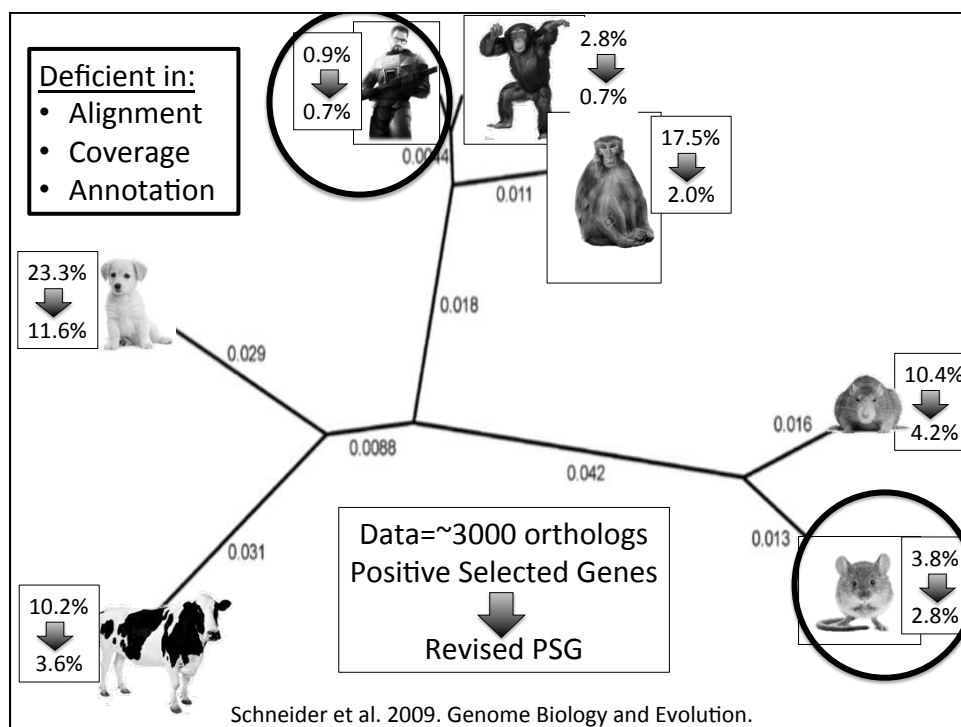
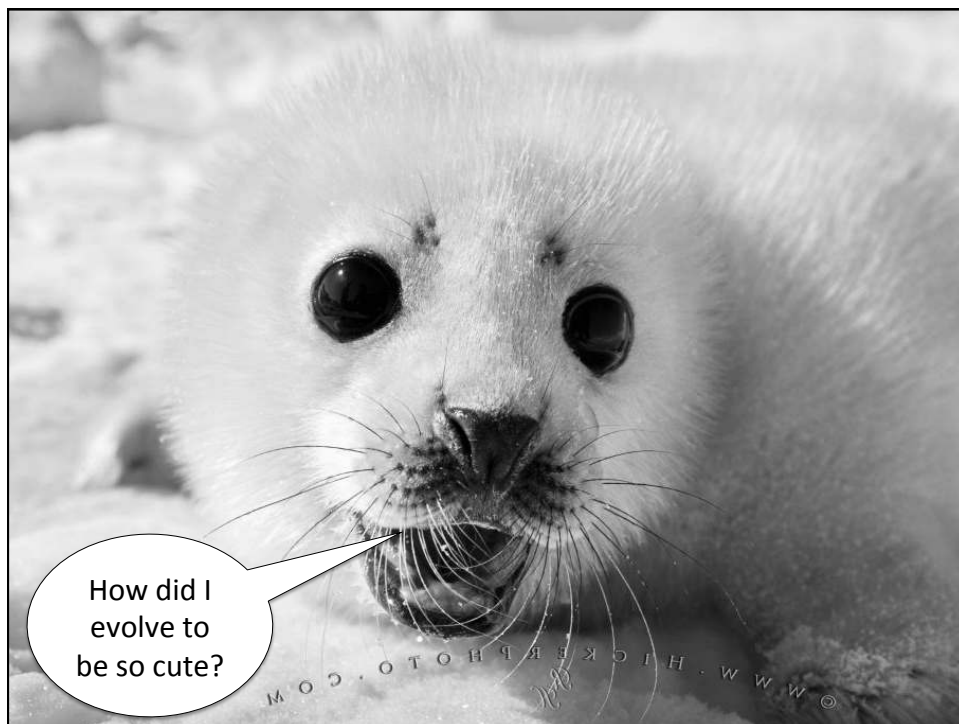
## What about recent genomes?

Surely they are better?

and mammals ... they have good genomes

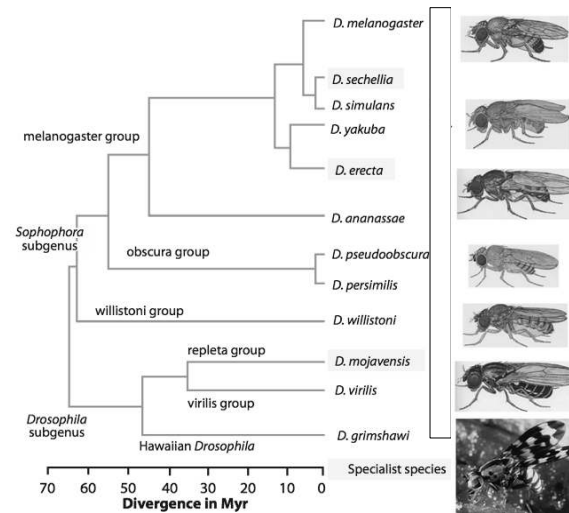
and alignment problems rarely happen

... right?

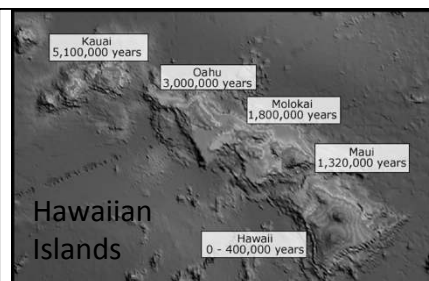




## Evolution of genes and genomes on the *Drosophila* phylogeny



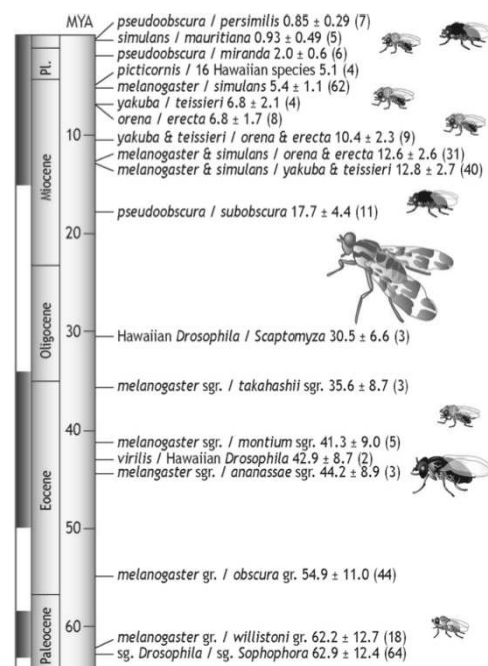
Drosophila 12 Genomes Consortium 2007 Nature



Calibration: Kauai age of 5.1 my for divergence of two Hawaiian species

1. No phylogeny
2. Fixed clock rate
3. Between 3 – 64 genes in pairwise comparisons

Temporal patterns in fruitflies (Tamura et al. 2004 MBE)

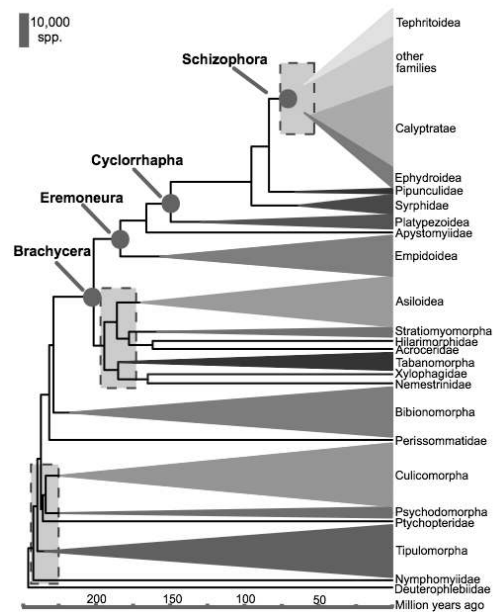




### Drosophila clade:

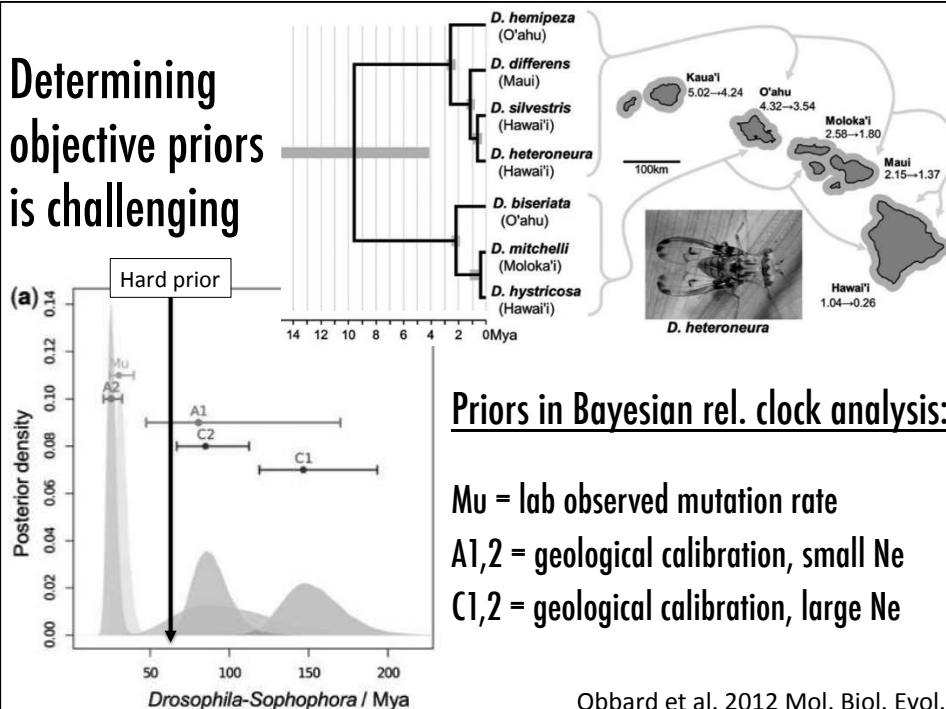
- Schizophora constrained to maximum of 70 Ma
- Without constraint, goes to 115 Ma

What is reality?

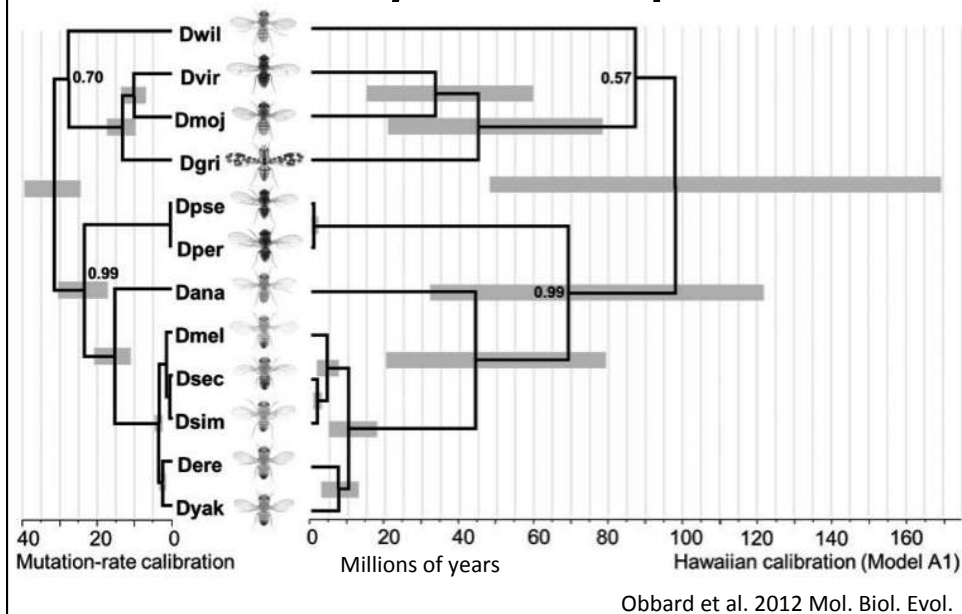


Episodic radiations in the fly tree of life  
(Wiegmann et al. 2011 PNAS)

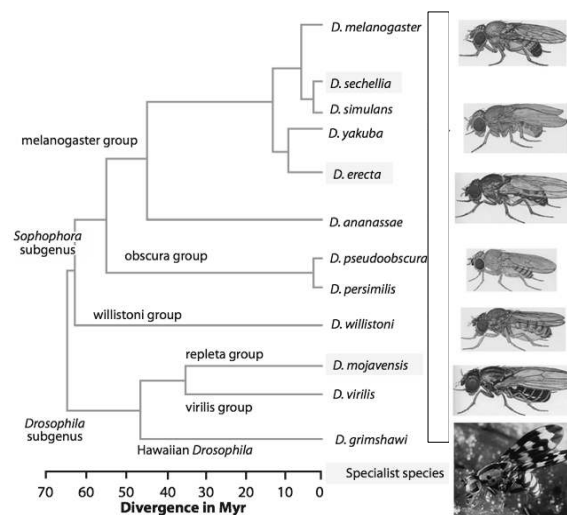
### Determining objective priors is challenging



## Priors directly influence posteriors



## Thus, the age of this clade is fiction



Drosophila 12 Genomes Consortium 2007 Nature

## Post-genomics challenge

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

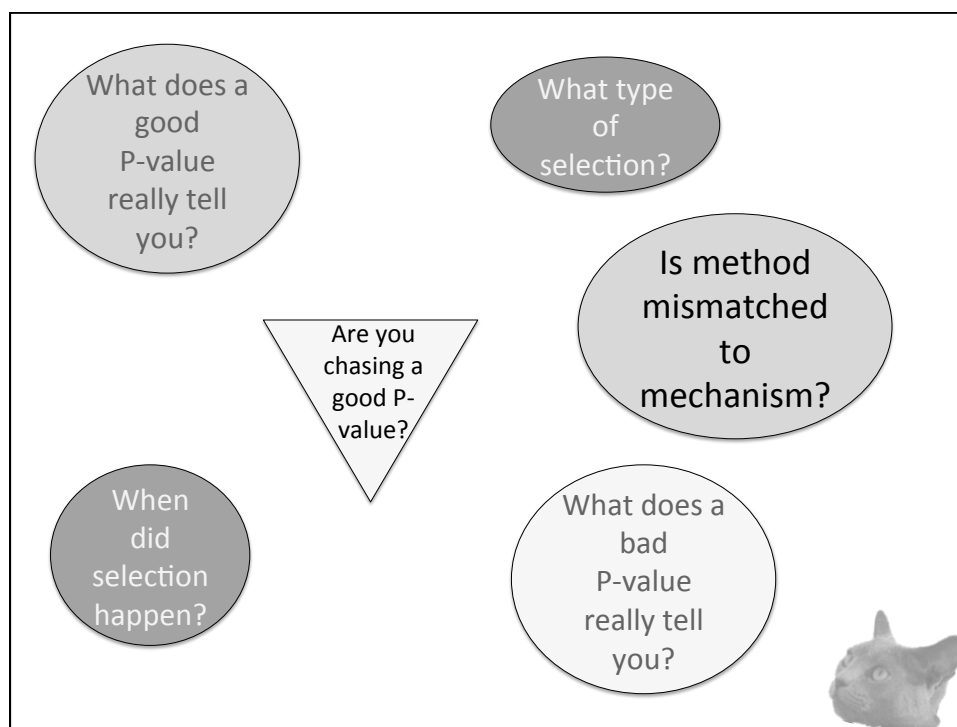
- Lewontin 1974

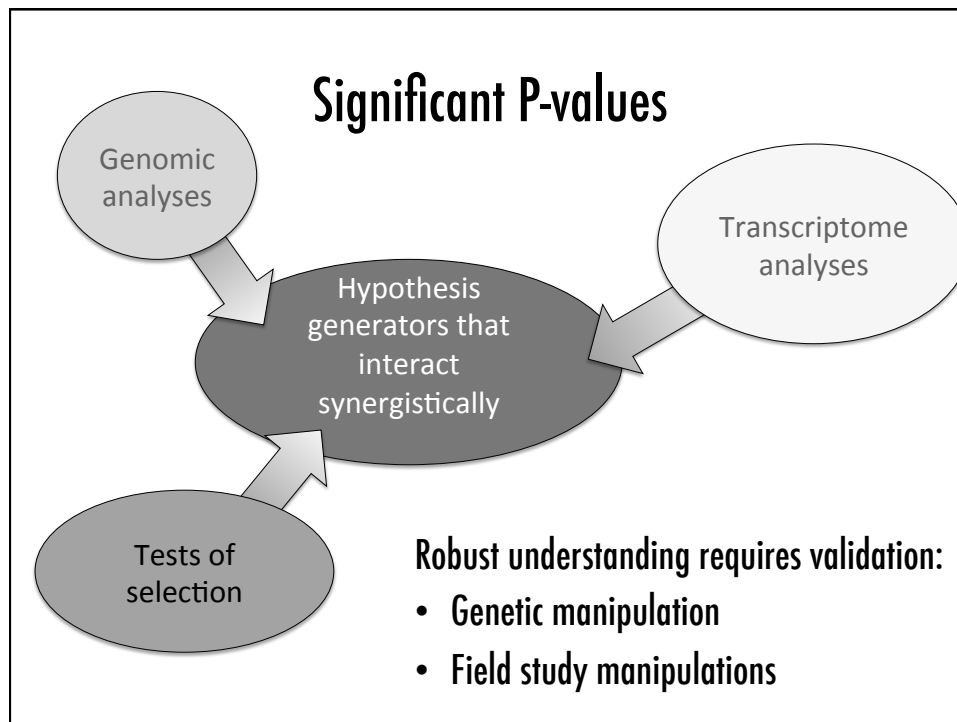
**"What we understand of the genome is by definition uninteresting and what we are interested in is by definition very damn difficult to sequence and assemble and annotate and analyze at genomic scale"**

- Wheat 2015

For example:

- indels & inversions
- gene family dynamics
- evolutionary dynamics





## Goal of this lecture

- Present a non-typical view of ecological genomics
  - So you have a more complete view of the field
- Make you uncomfortable
  - Provide a context for understanding your results
- Encourage you to rethink the reality presented by publication biases
  - Overcoming this bias is a continual challenge

