

Lies, damn lies, and genomics

you, your data, your perception and
the reality

Christopher West Wheat



Goal of this lecture

- Present a non-typical view of ecological genomics
- Make you uncomfortable by sharing my nightmares
- Encourage you to critically assess your results in light of publication biases

Disclaimer

I'm a positive person

I like 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

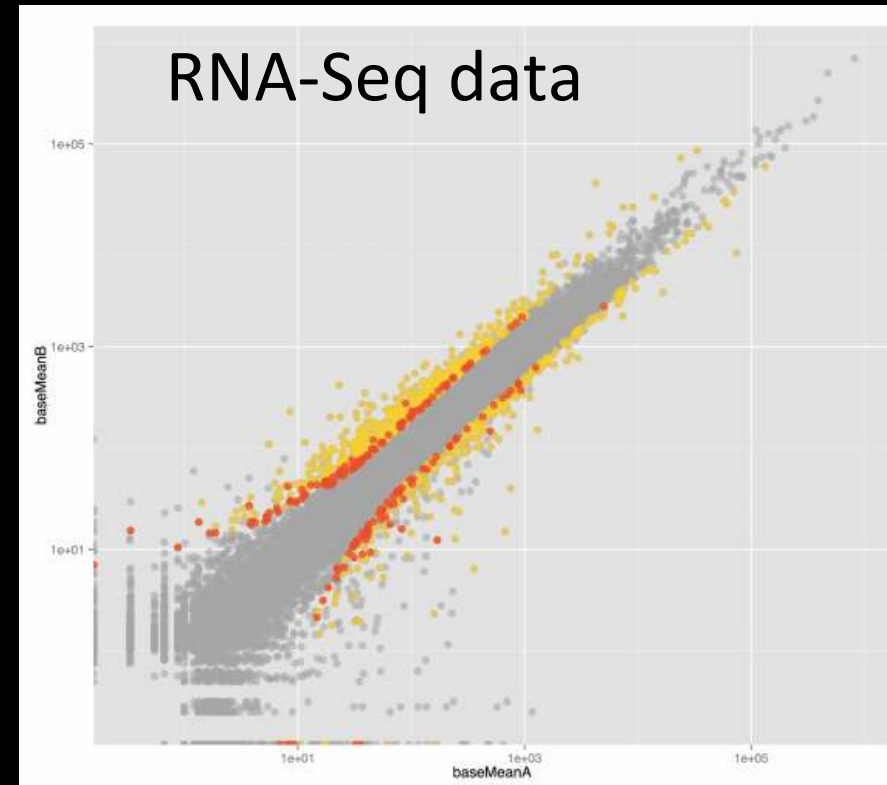
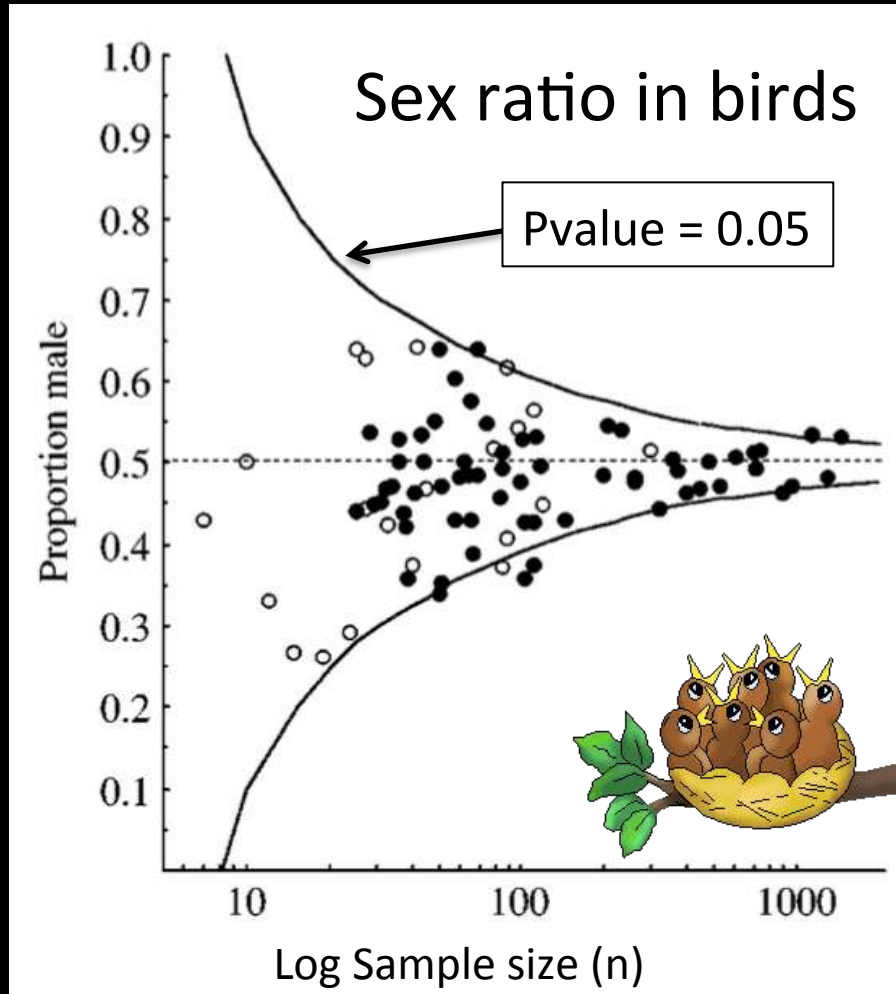
Their findings should be reasonably robust:

- Repeatable effect sizes
 - The same in different labs
 - The same over time

Publication replication failures

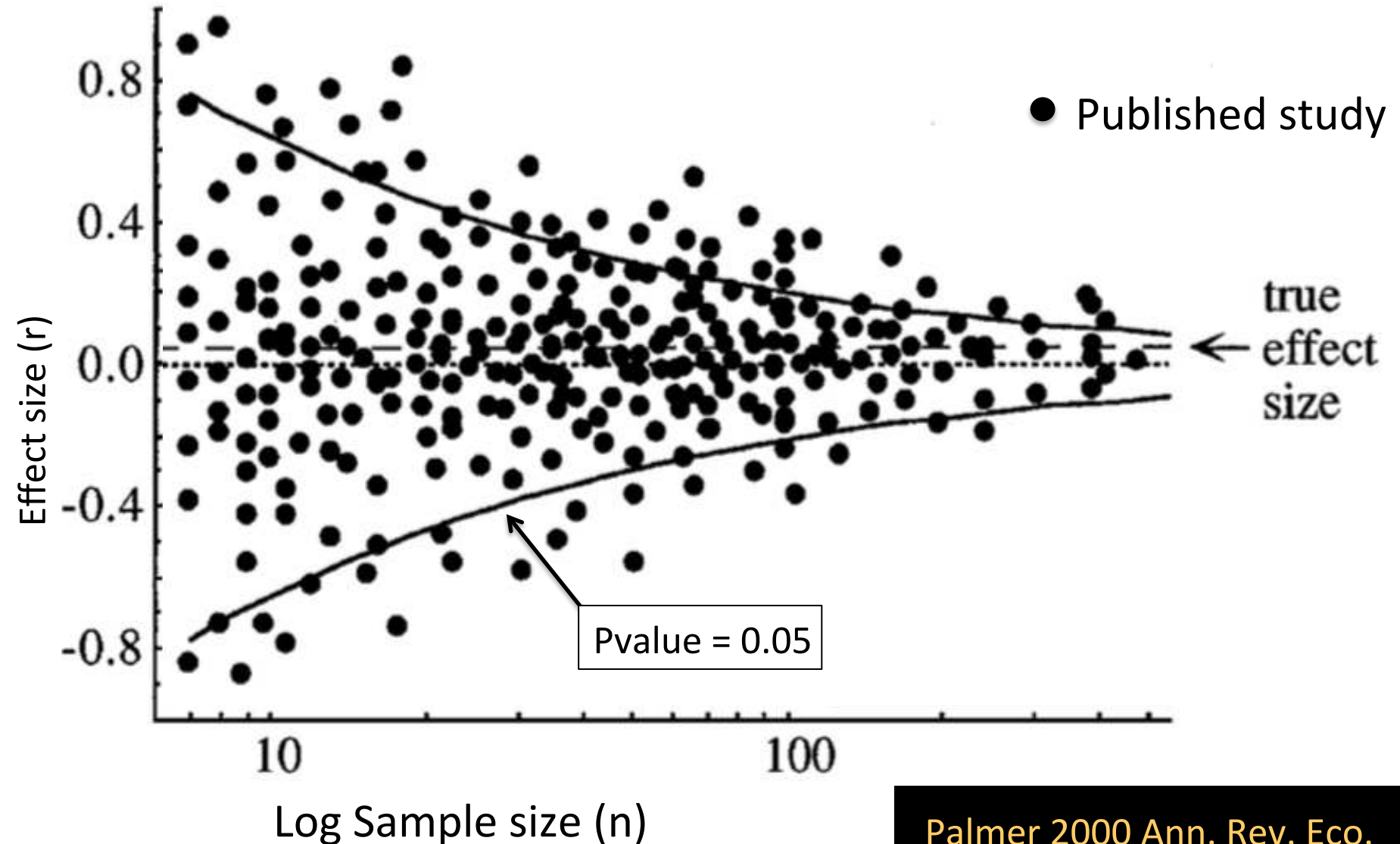
- Biomedical studies
 - Of 49 most cited clinical studies, 45 showed intervention was effective
 - Most were randomized control studies (robust design)
 - Of the 34 that were later replicated, 41% were directly contradicted or had much lower effect sizes.
- Mouse cocaine effect study, replicated in three cities
 - Highly standardized study
 - Average movement was 600 cm, 701 cm, and > 5000 cm in the three study sites

Assessing reality using funnel plots

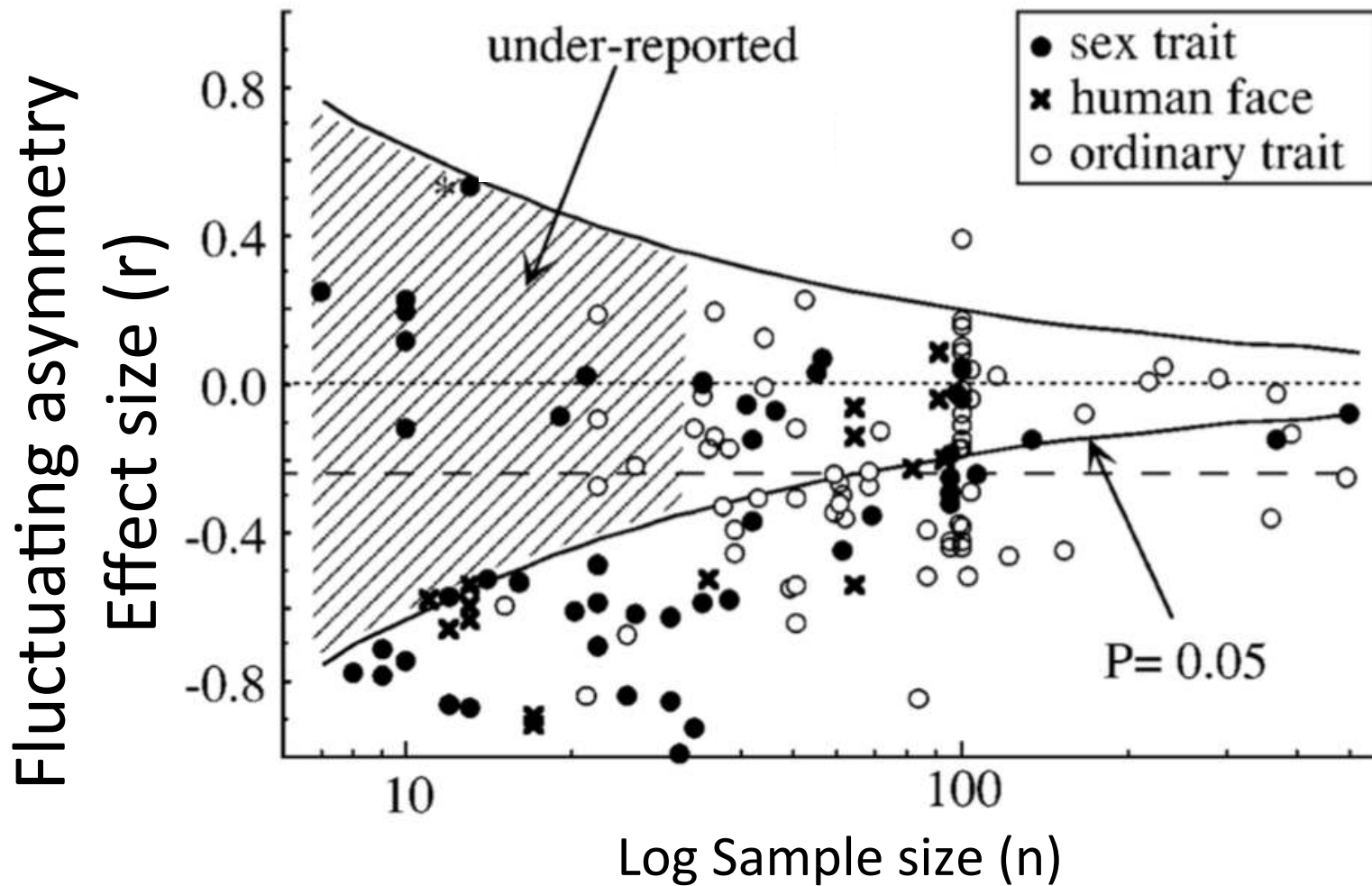


Small sample sizes
affect measurements

Publication bias increases effect size



Fluctuating asymmetry and mate preference: a correlation between effect size and sample size

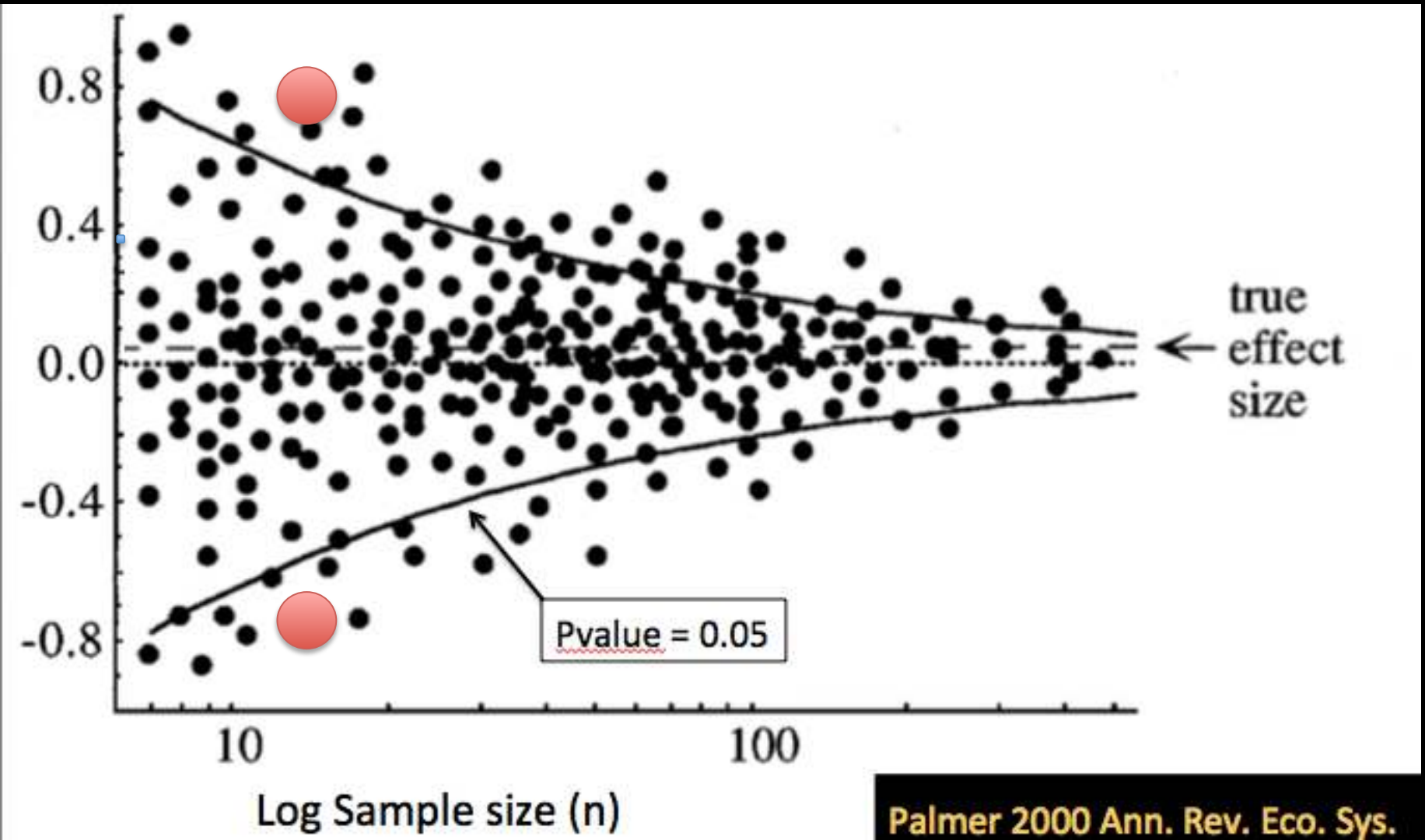


r_{bias} sex + face (Spearman's $\rho=0.38$, Pvalue=0.002)

r_{bias} ordinate trait: n.s.

What if there is no replication?

What is most likely to publish?



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 is a greater number of tested relationships using tests with *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

There are lies, damn lies, and

Are datasets too big to fail?

What do follow-up studies reveal?

How can we gain confidence in our work?

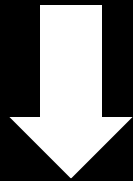
Outline

- What is the genomic architecture of phenotypes?
- What is the power of molecular tests of selection?
- What does the dissection of a classic comparative genomics study reveal?

Non – adaptive

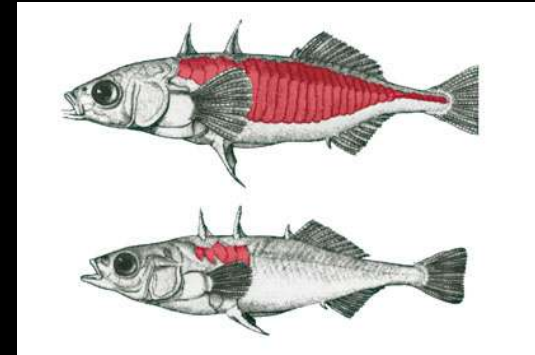


disease, aging, height, etc.

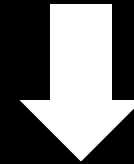


1000's of loci, each of
small effect size

Adaptive



salinity, color, resistance, etc.



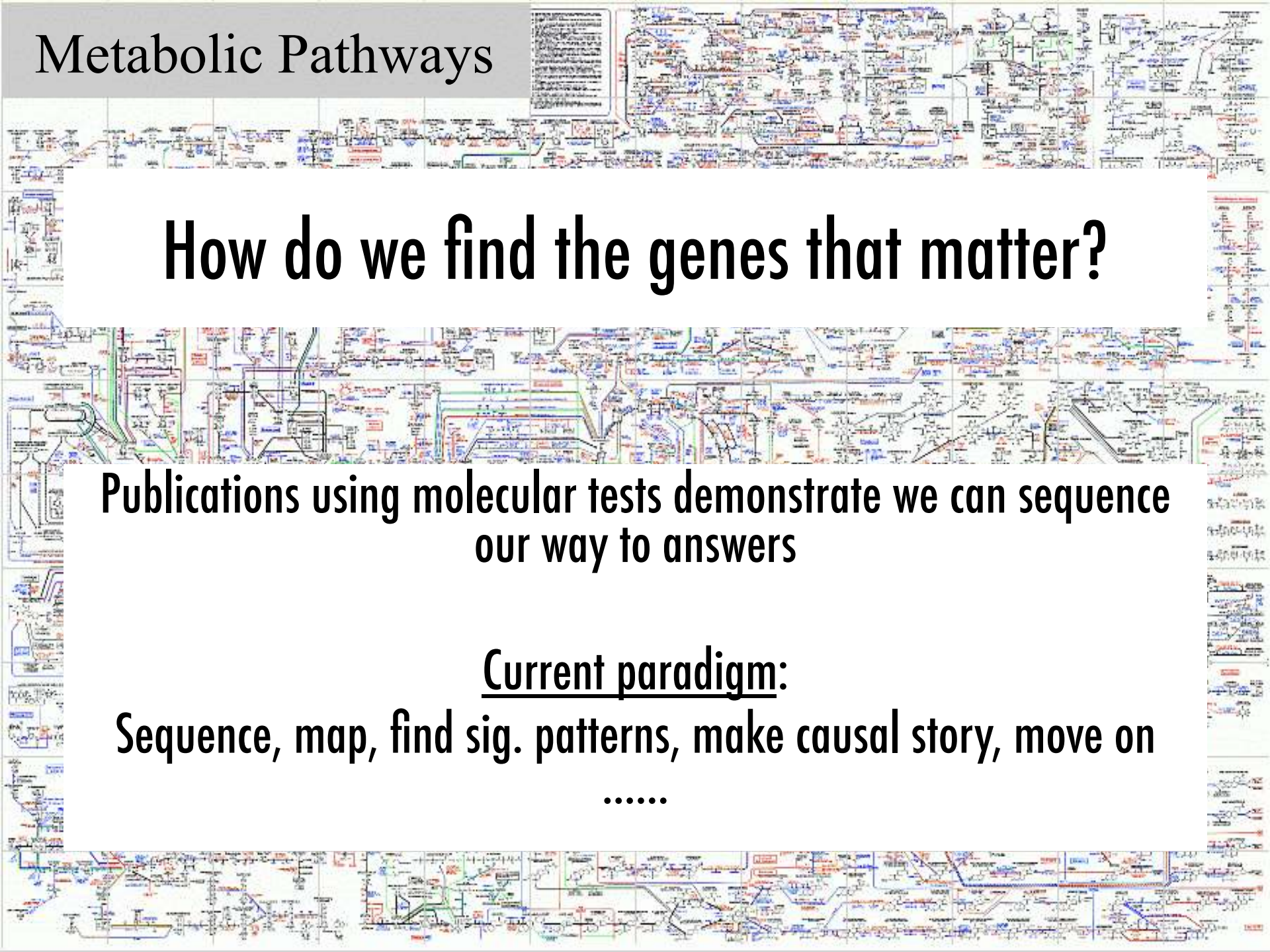
One or several loci of large
effect

generally ...

Is this a publication bias?

Will your trait have 1000's of small effect
genes, or a few genes of large effect?

Metabolic Pathways



How do we find the genes that matter?

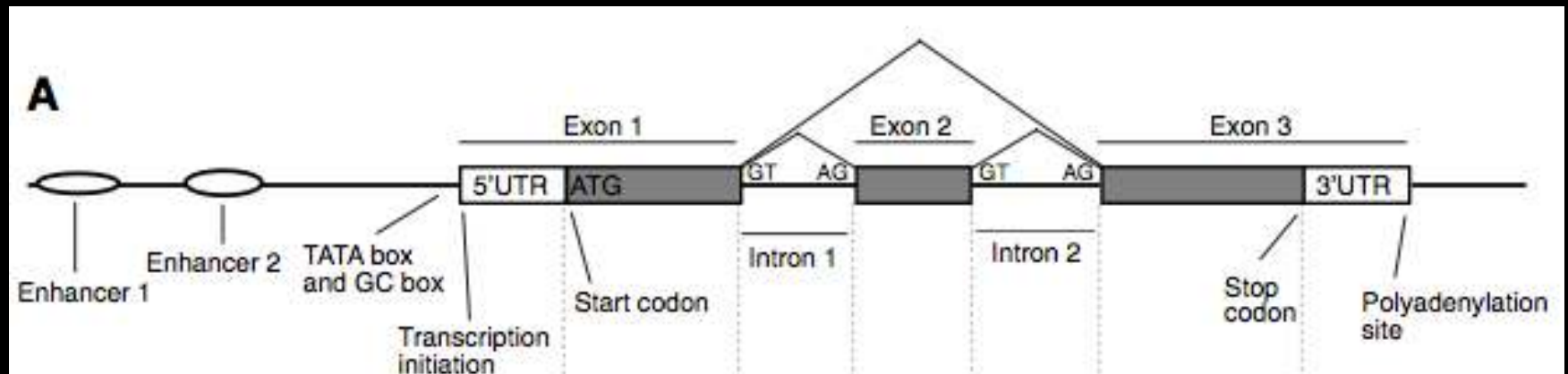
Publications using molecular tests demonstrate we can sequence our way to answers

Current paradigm:

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

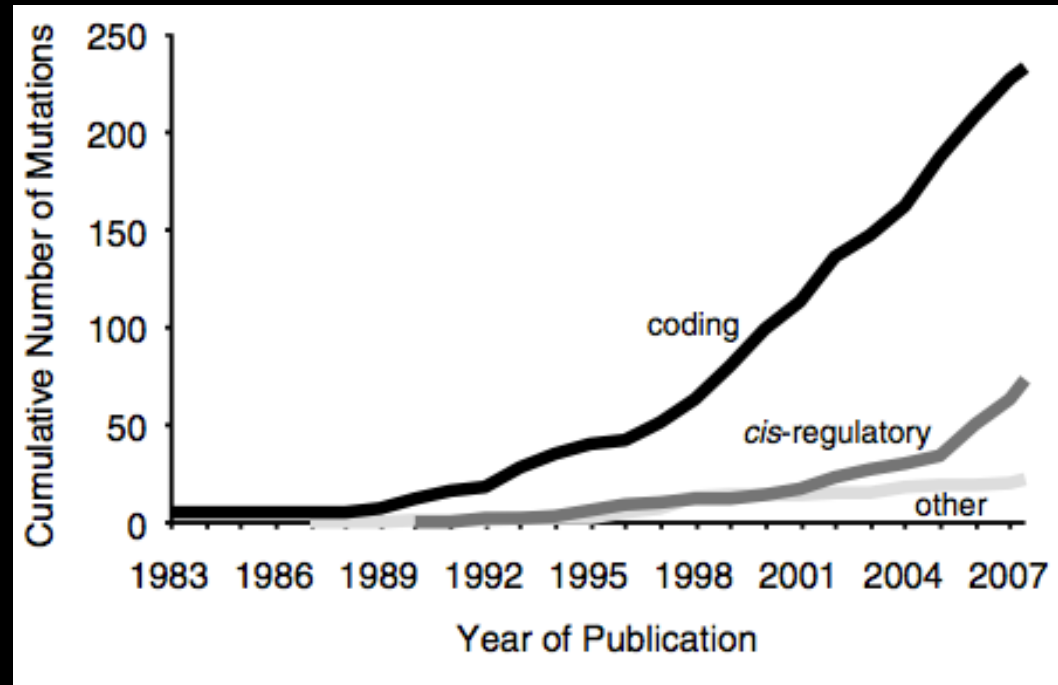
.....

What is the architecture of a causal variant?



How predictable are adaptations?

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

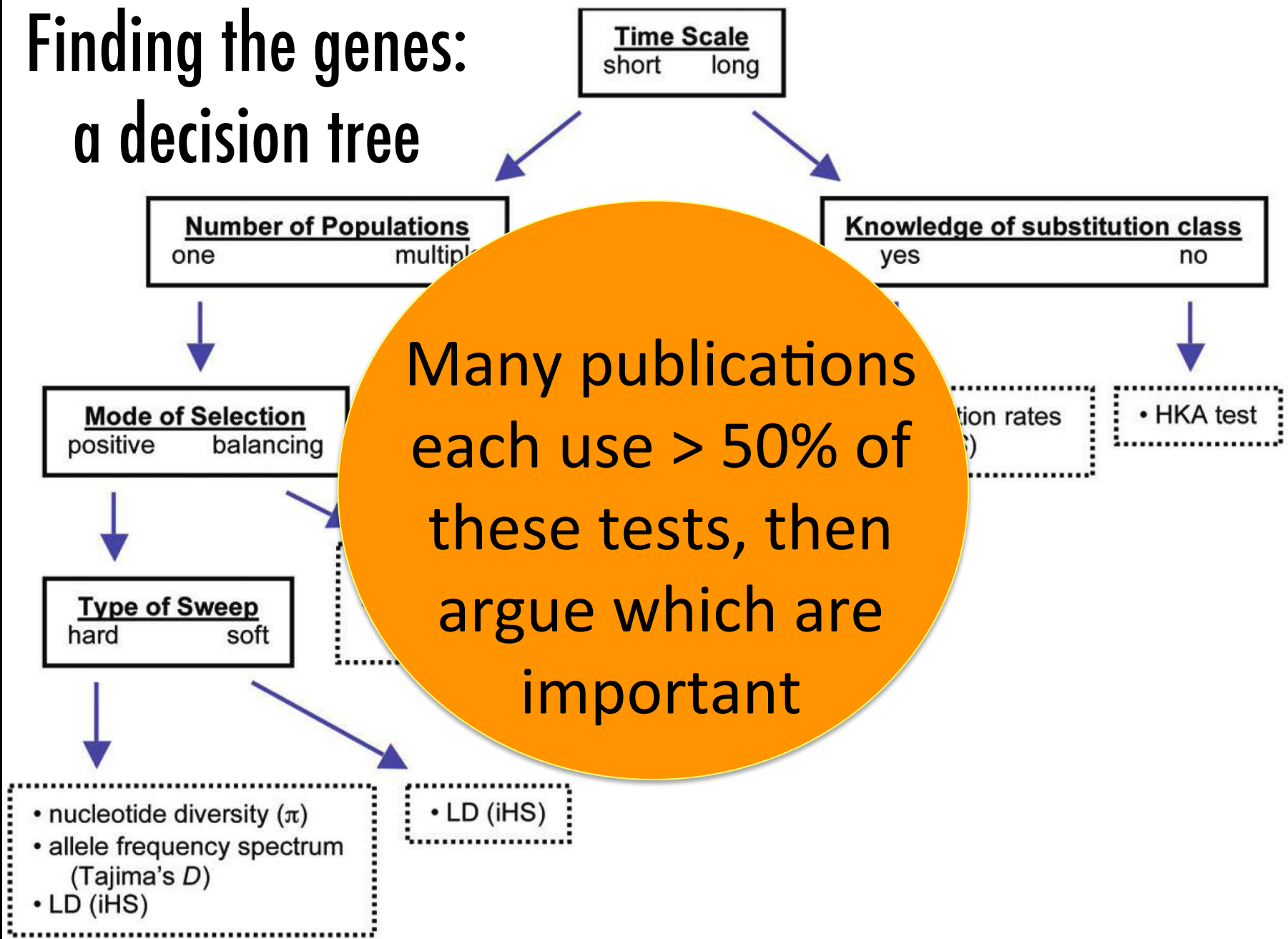


	Morphology	Physiology	Behavior
Coding ³	62	170	2
<i>Cis</i> -regulatory	43	29	2
Other ⁴	3	20	0
Total	108	219	4
Null ⁵	41	58	0

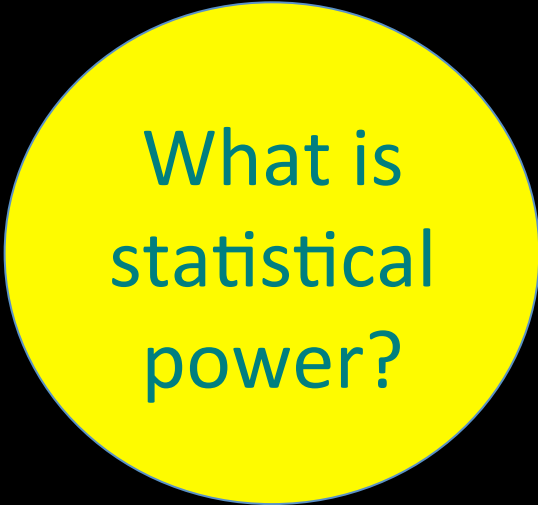
How do we identify the genes that matter?

- 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?

Finding the genes: a decision tree



What power do we
have to detect
balancing
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 want power $> 90\%$ at reasonable
sample size, right?

What power do we have to detect balancing selection?

ρ	Width of window (bp)				
	25	50	100	200	1000
1	85.6	90.2	92.8	93.5	83.8
3	80.8	85.3	86.3	83.5	44.7
10	69.0	69.9	64.5	51.0	4.1
30	48.1	42.5	31.0	15.7	0.1
100	20.5	15.6	8.9	2.4	0.0

Tajima's D
% finding selection of 5000 simulations

- For *Drosophila melanogaster*, power = 50% with window size of 200 bp, using 24 diploid individuals.
- For species with larger population size, power likely lower
- Recombination and gene conversion destroy 'footprint' rather quickly

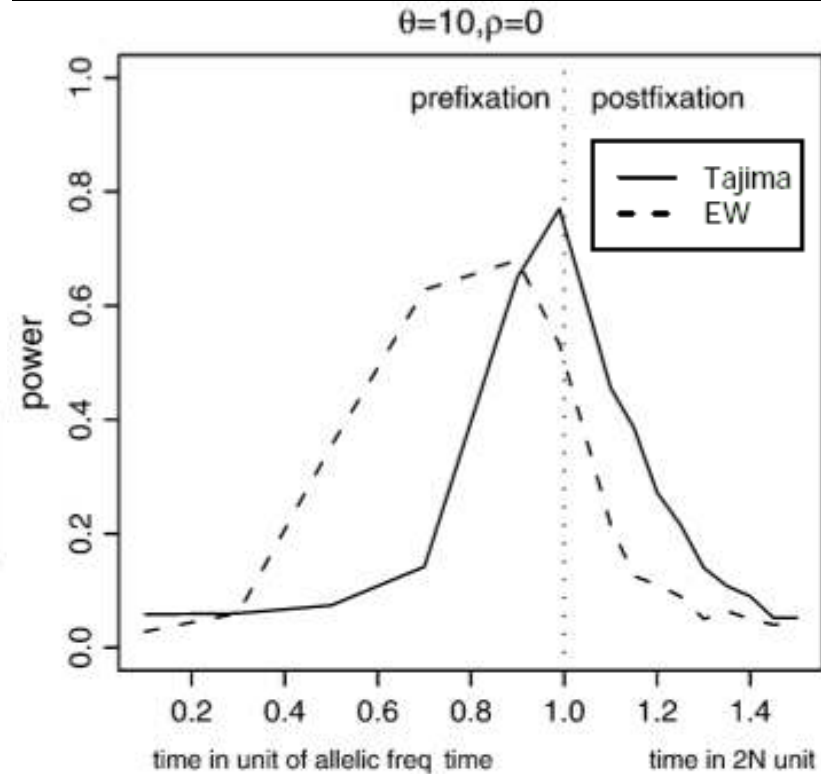
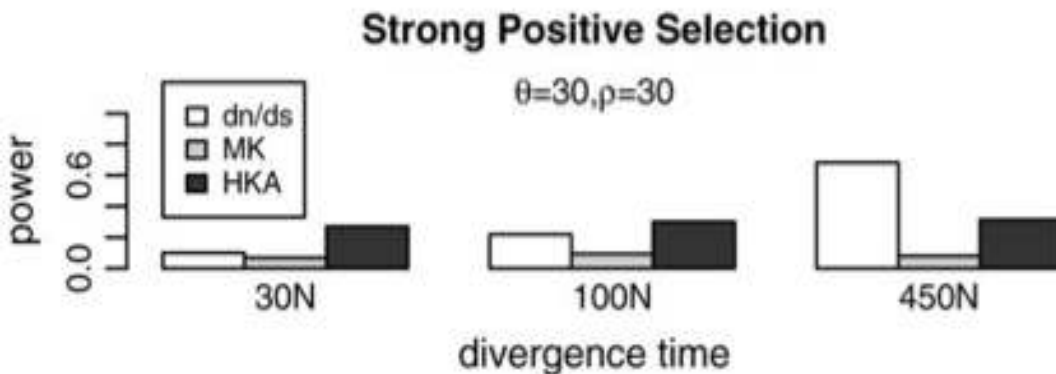
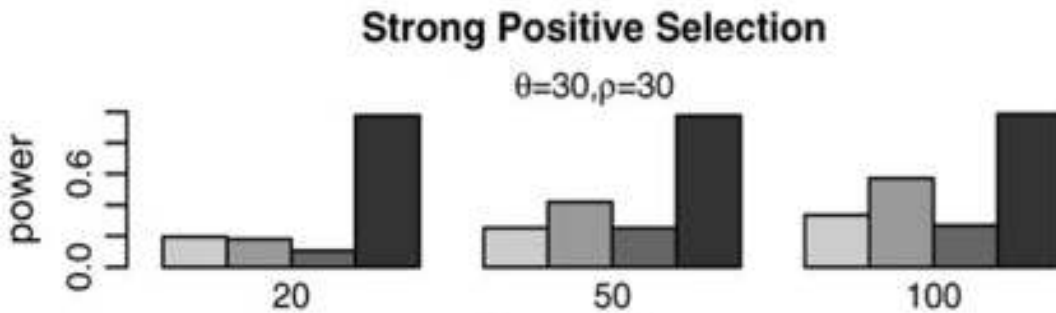
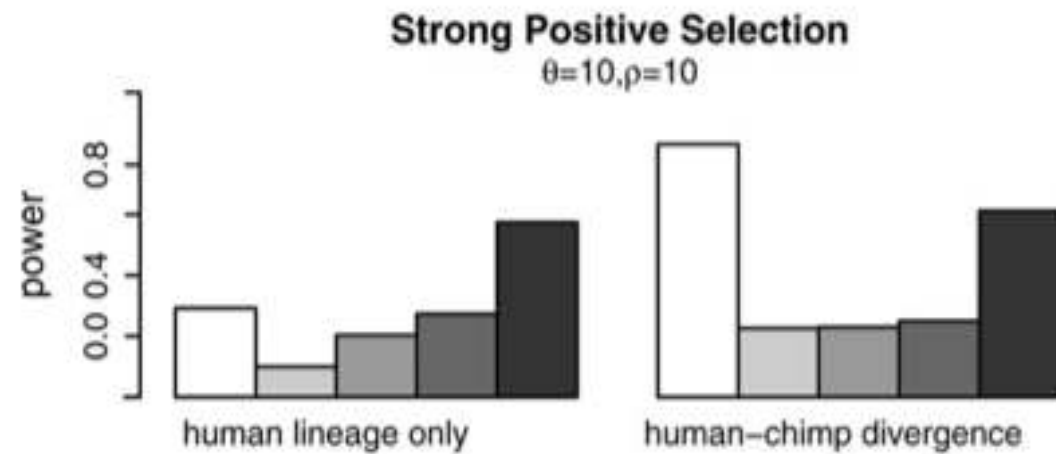
Directional selection: an example of the expectations of hard selection

```
A T G G T A G G T C A T A T T G A T C A G G G T G A A T G T G C T A G A A C A T A
A T G C T A G A T C A A A G T G A T C A T G G T G A A T G T G C T A G A A C A T A
A T G G T A G A T C A A A T T G A T C A T G G T G C A T G T G C T A G A T C A T A
A T G C T A G A T C A T A T T G A T G A T G G T G A A T G T G C T A G A T C A T A
A T G C T A G A T C A T A T T G A T C A T G G T G A A T G T G C T T G A A C A T A
A T G C T A G G T C A T A T T G A T C A T G C T G A A A G T G G T A G A T C A T A
```

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

- But a 'null model' has been elusive, resulting in many false positives

What is our power to detect hard sweeps?



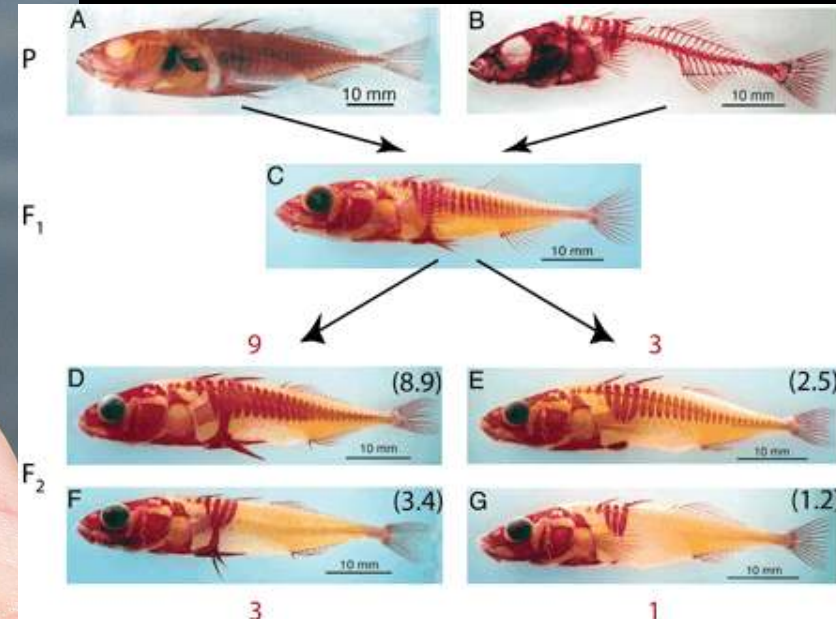
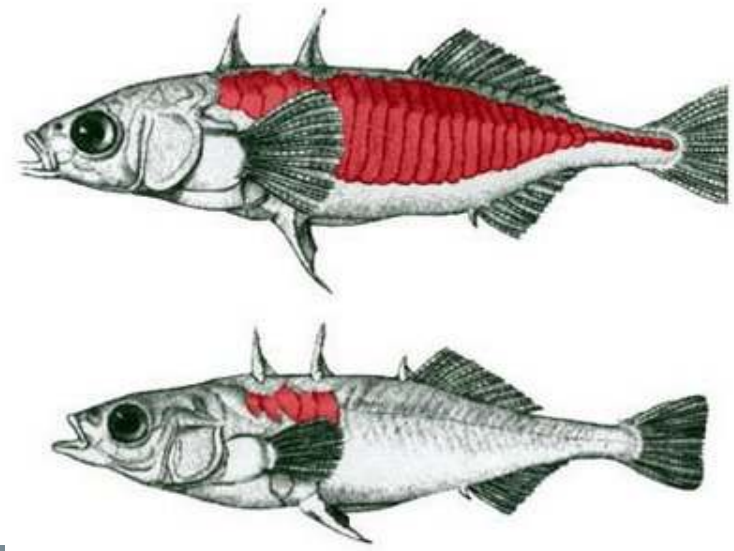
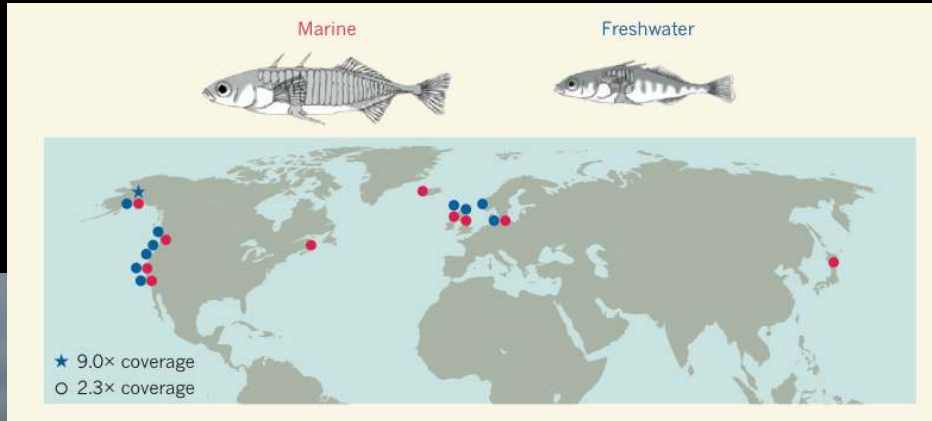
When did selection act on your phenotype?

Zhai, Nielsen & Slatkin 2008 MBE

Clear pattern

- There are many molecular tests of selection
 - Each performs better under a specific set of conditions
- Their power is very low under a range of realistic biological conditions
- Their false positive rate can be very high across a range of realistic biological conditions

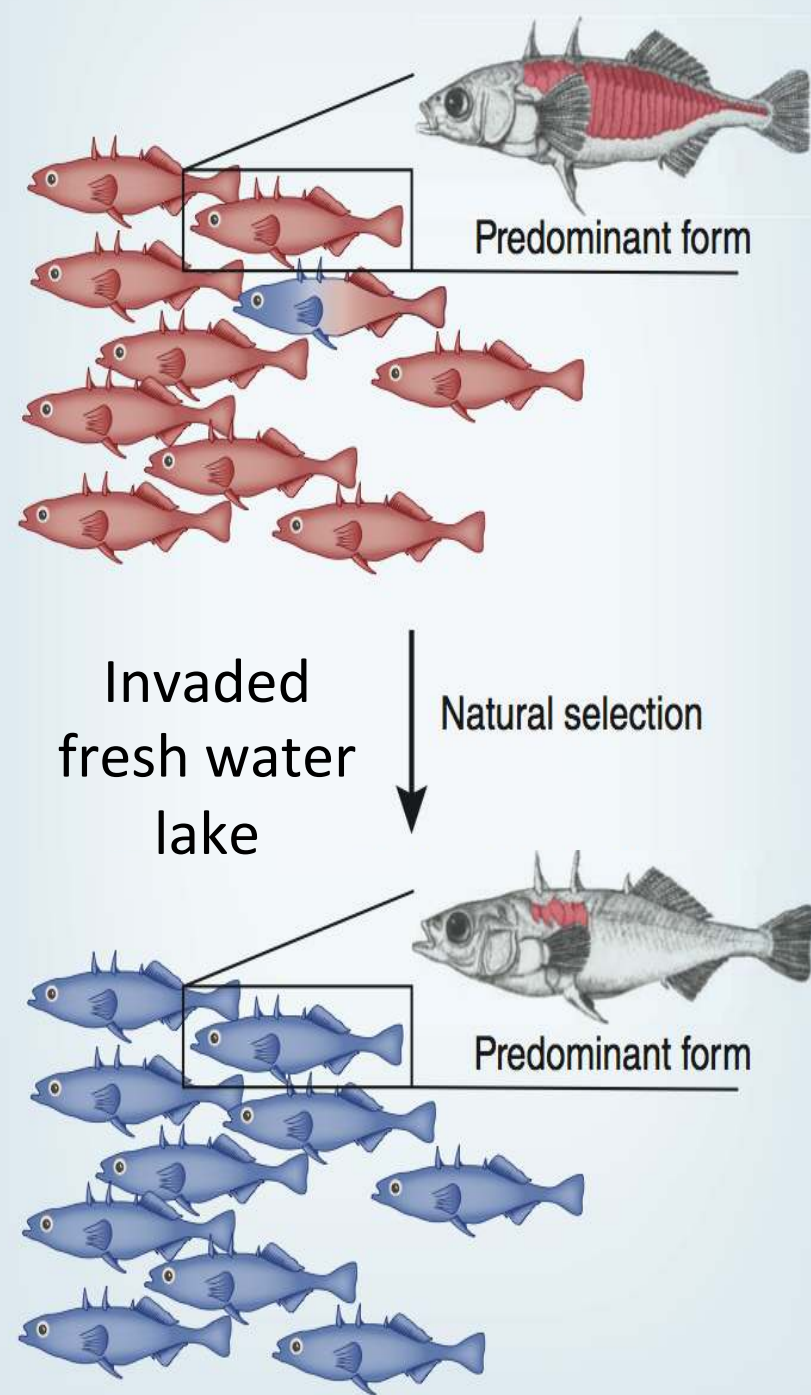
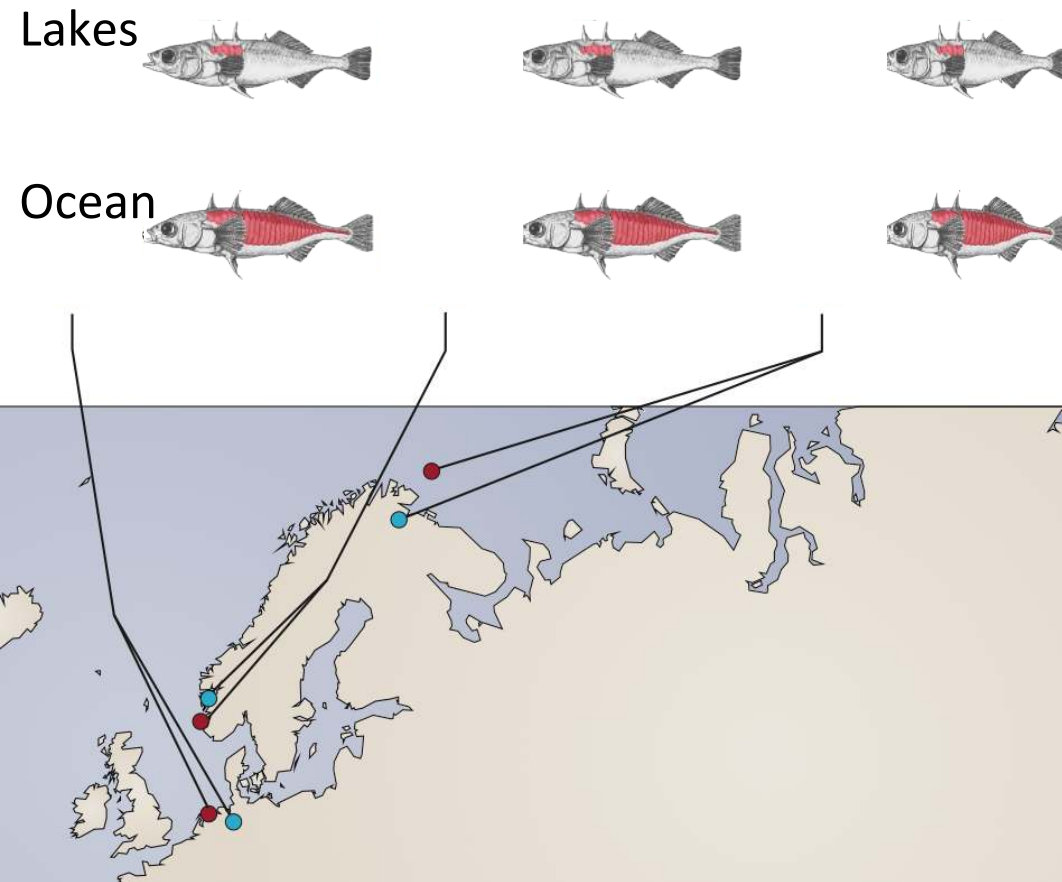
Hard selection case example: threespine stickleback fish



Threespine stickleback fish

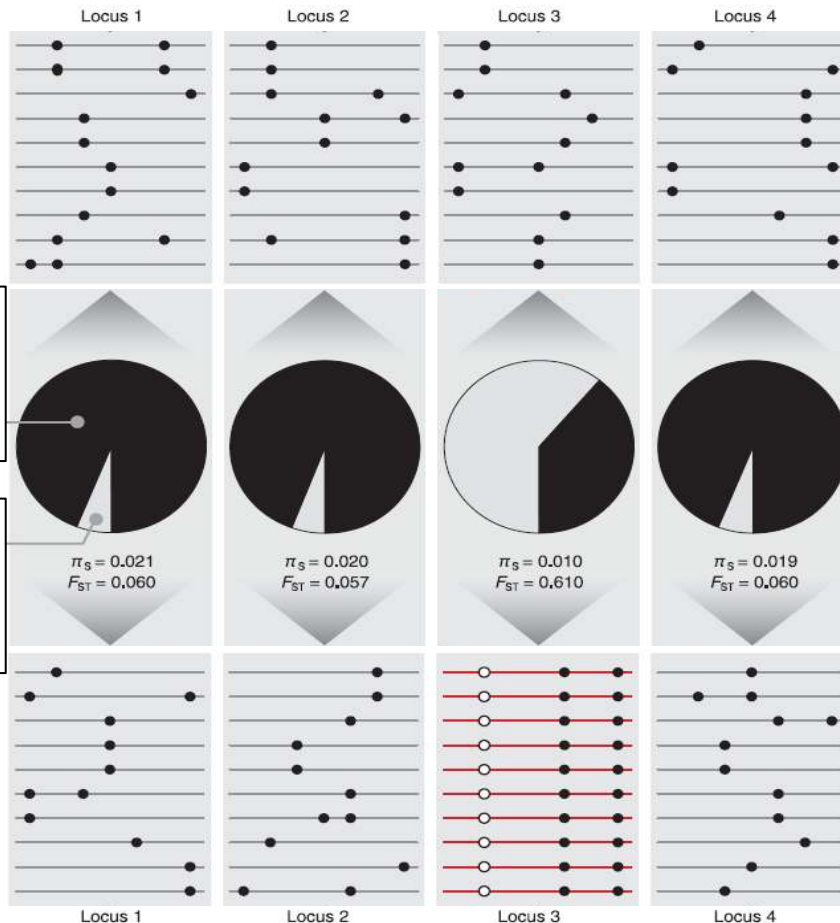
(*Gasterosteus aculeatus*)

- Has body armor in the ocean
- Loses almost all armor in lakes



Parallel adaptation in fresh water lakes via hard sweeps

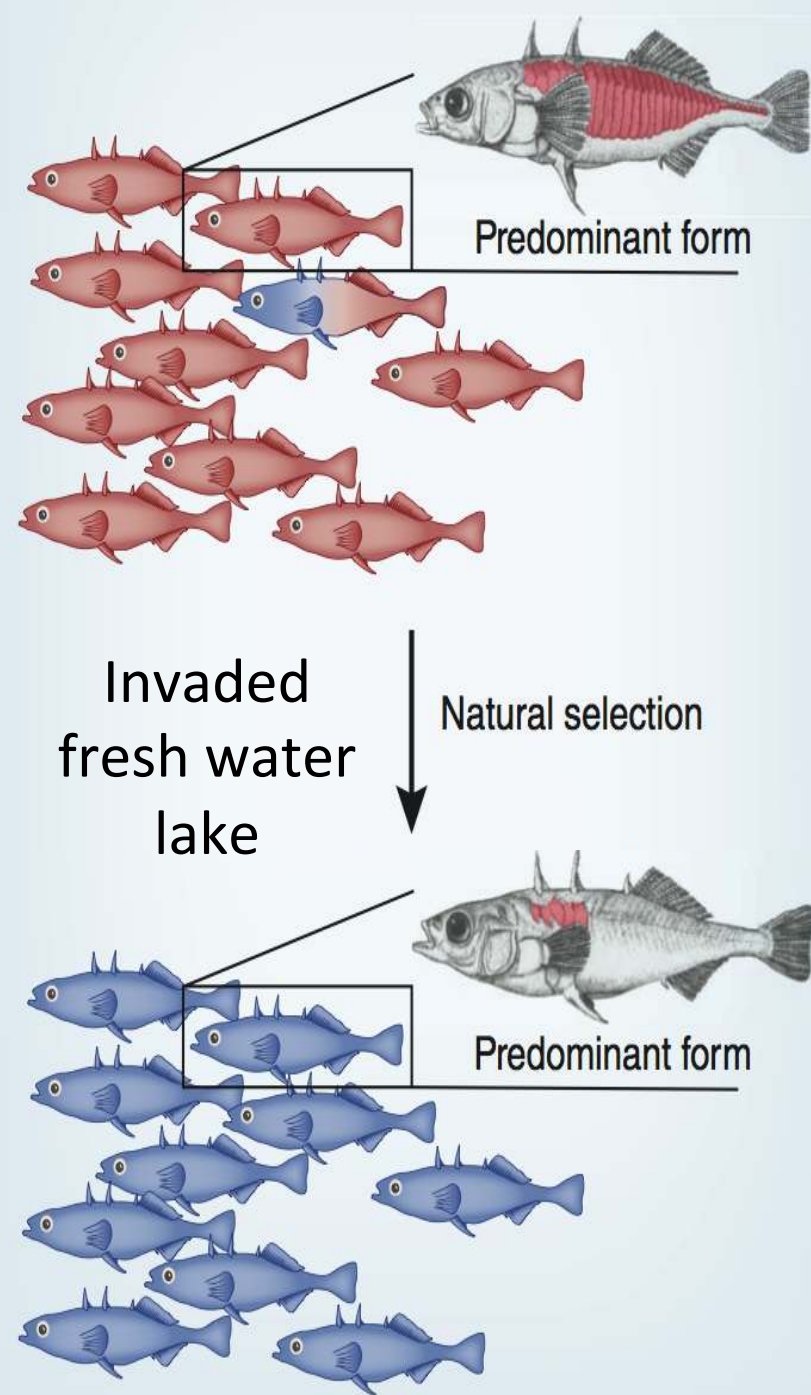
Marine population



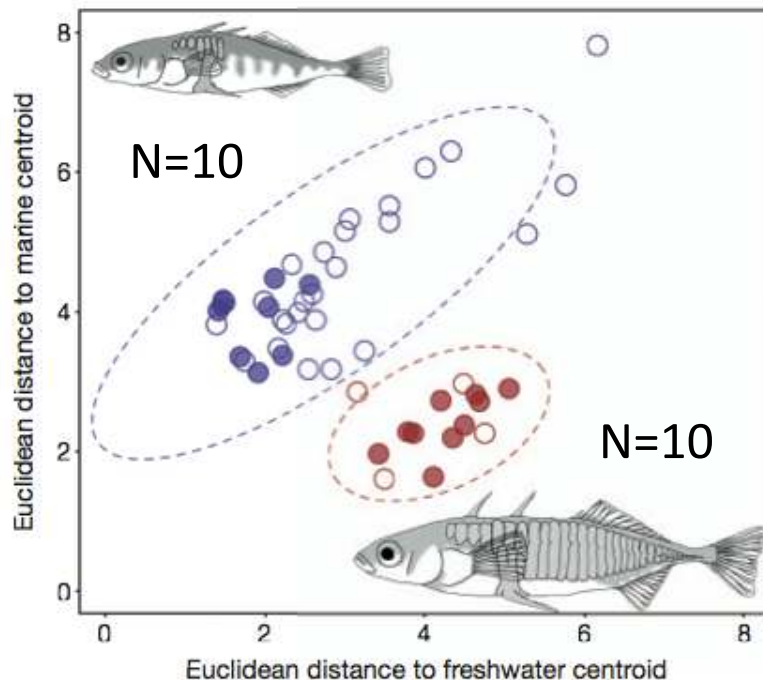
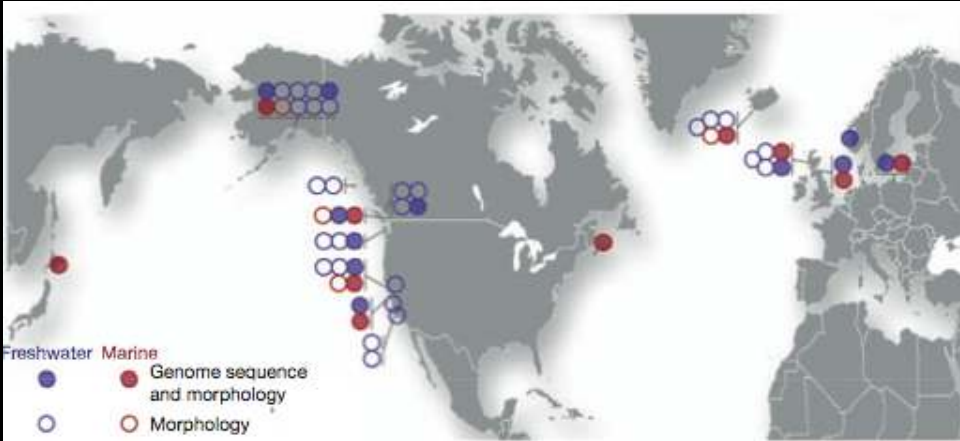
Proportion variation within populations

Proportion variation between populations

Population B



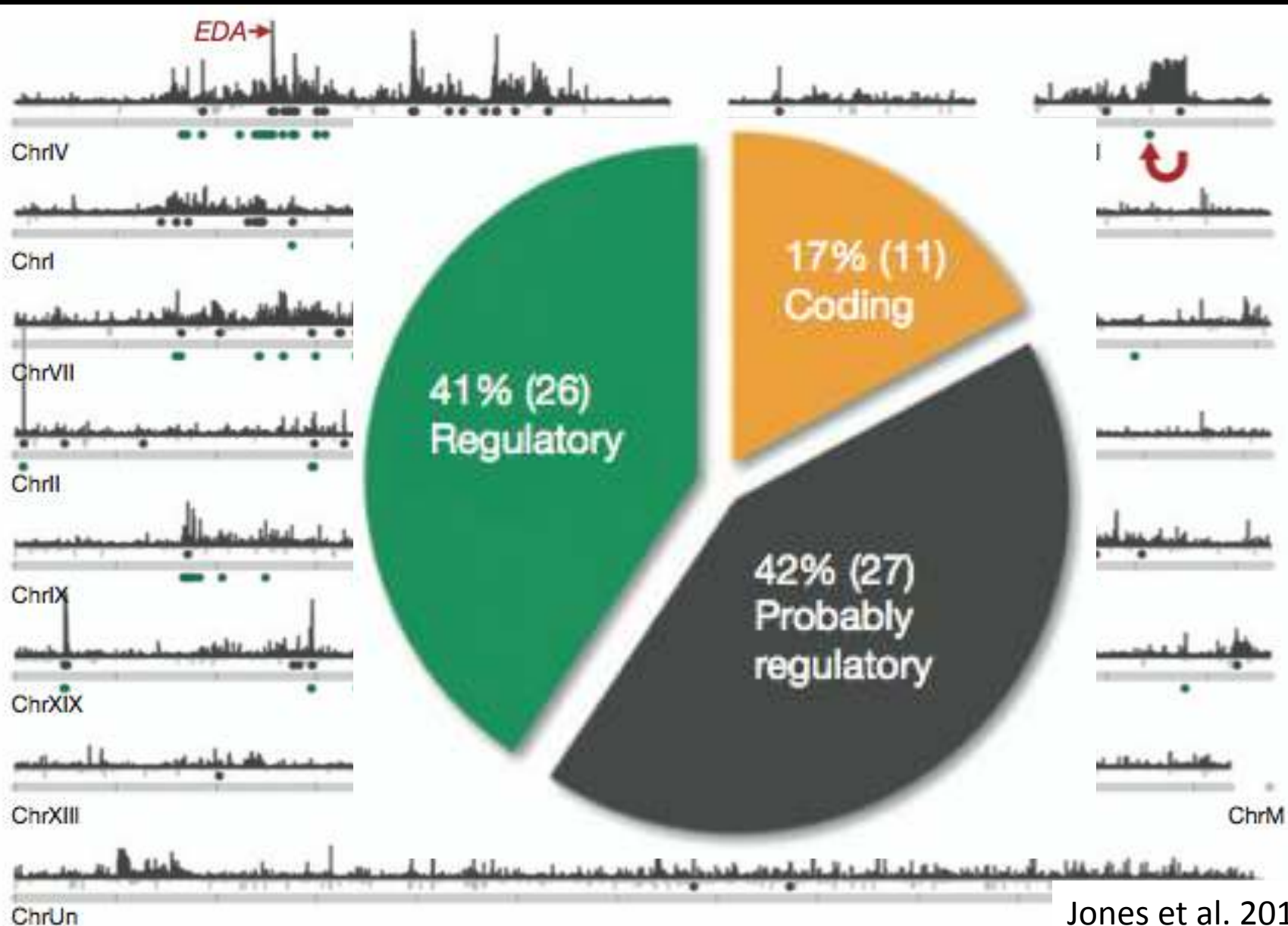
Individual genome sequencing: powerful insights



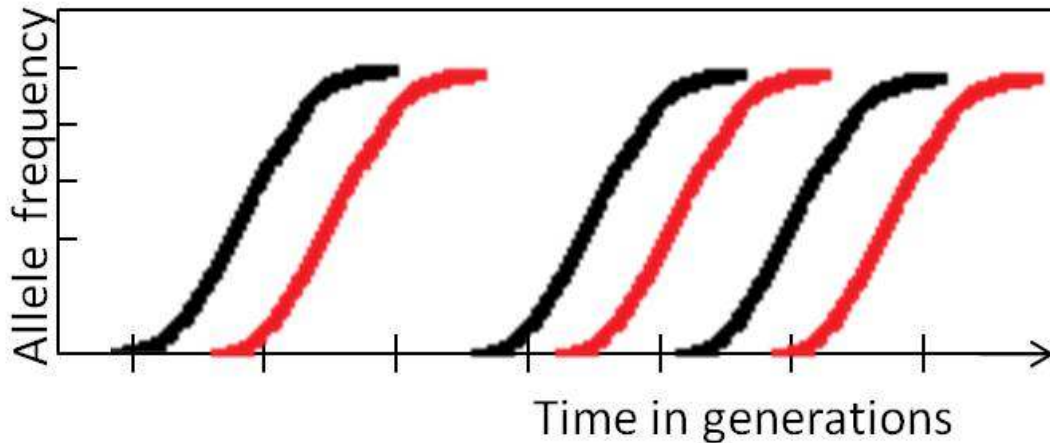
2-5 X per individual, sliding 2500 bp window, 500 bp step

Jones et al. 2012 Nature

What regions are important? Coding or expression?

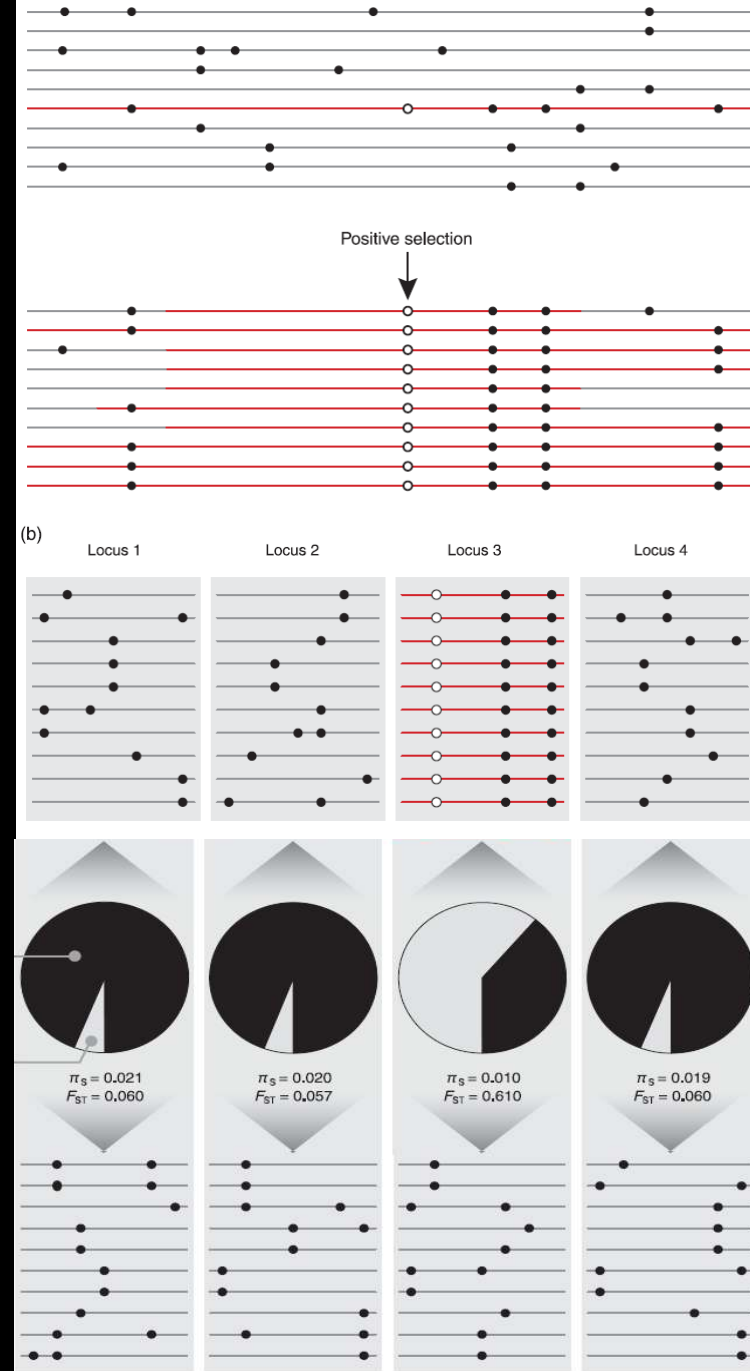


How common are such hard selective sweeps?

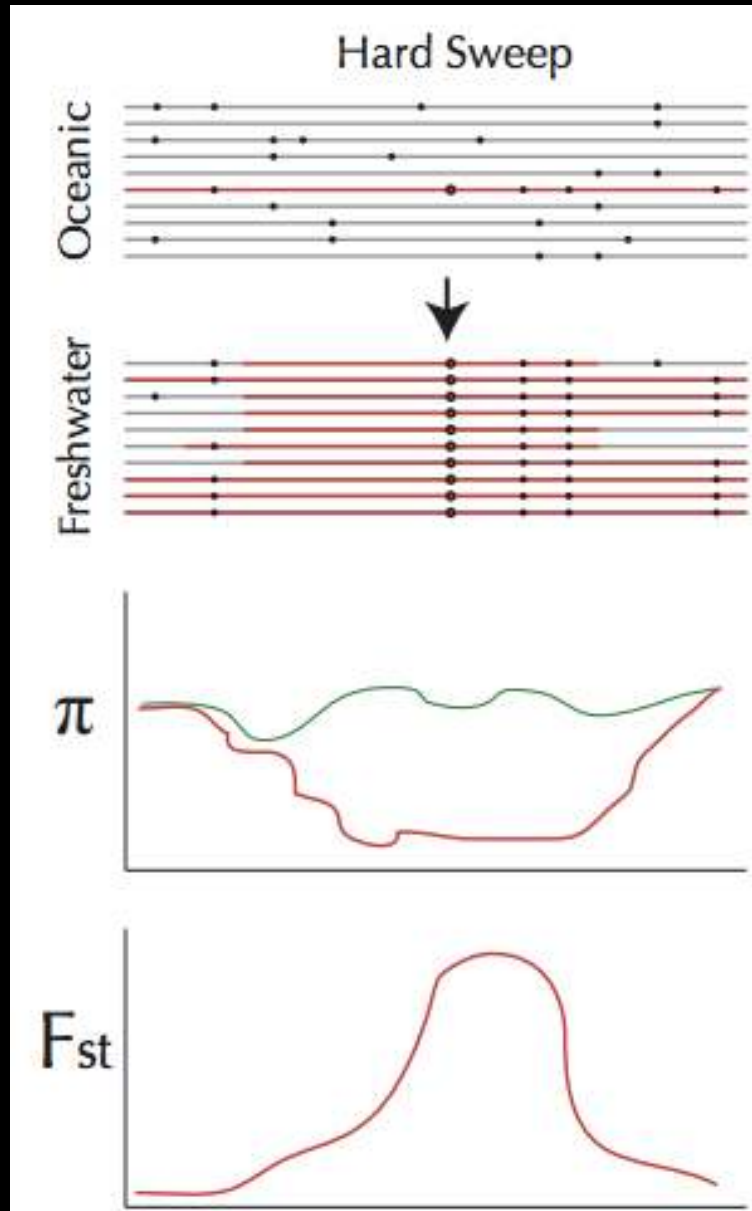


Does your favorite test for selection rely upon one or many sweep events?

- MK-test needs repeated events
- F_{ST} outlier, EHH, Tajima's D , etc.



Hard vs. soft or incomplete sweeps in populations



How common were hard sweeps in our history?

- “we argue that soft sweeps might be the dominant mode of adaptation in many species”

Messer and Petrov 2013 TREE

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

How common are soft sweeps in your species?

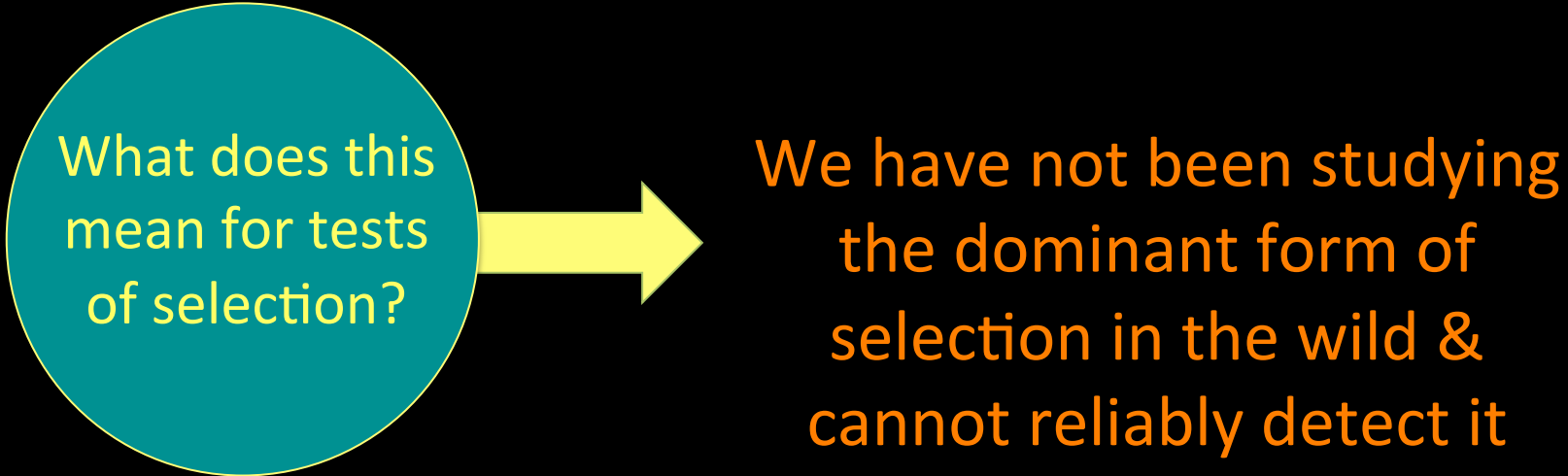
Thought experiment:

Do most species respond to selection in the lab? **Yes**

Why? **existing variation in population**

If populations have variation, can selection act on it? **Yes**

What does this tell us about frequency of soft selection in wild?



What does this
mean for tests
of selection?

We have not been studying
the dominant form of
selection in the wild &
cannot reliably detect it

Age and type of selection matters

- Novel mutation, large mutation, hard sweep selected to fixation
 - High probability of detection
- Old mutation, polygenetic, soft sweep of incomplete fixation
 - Very low probability of detection
- Finding the causal mechanism
 - Coding > expression
 - SNPs > more complex mutations (indel, TE, CNV)
 - Ongoing gene flow & grouping by phenotype across replicate populations helps a lot
- 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 1000ⁿ genomes

- Roughly 20 arthropods sequenced to date
 - plans to sequence more
- Many other large



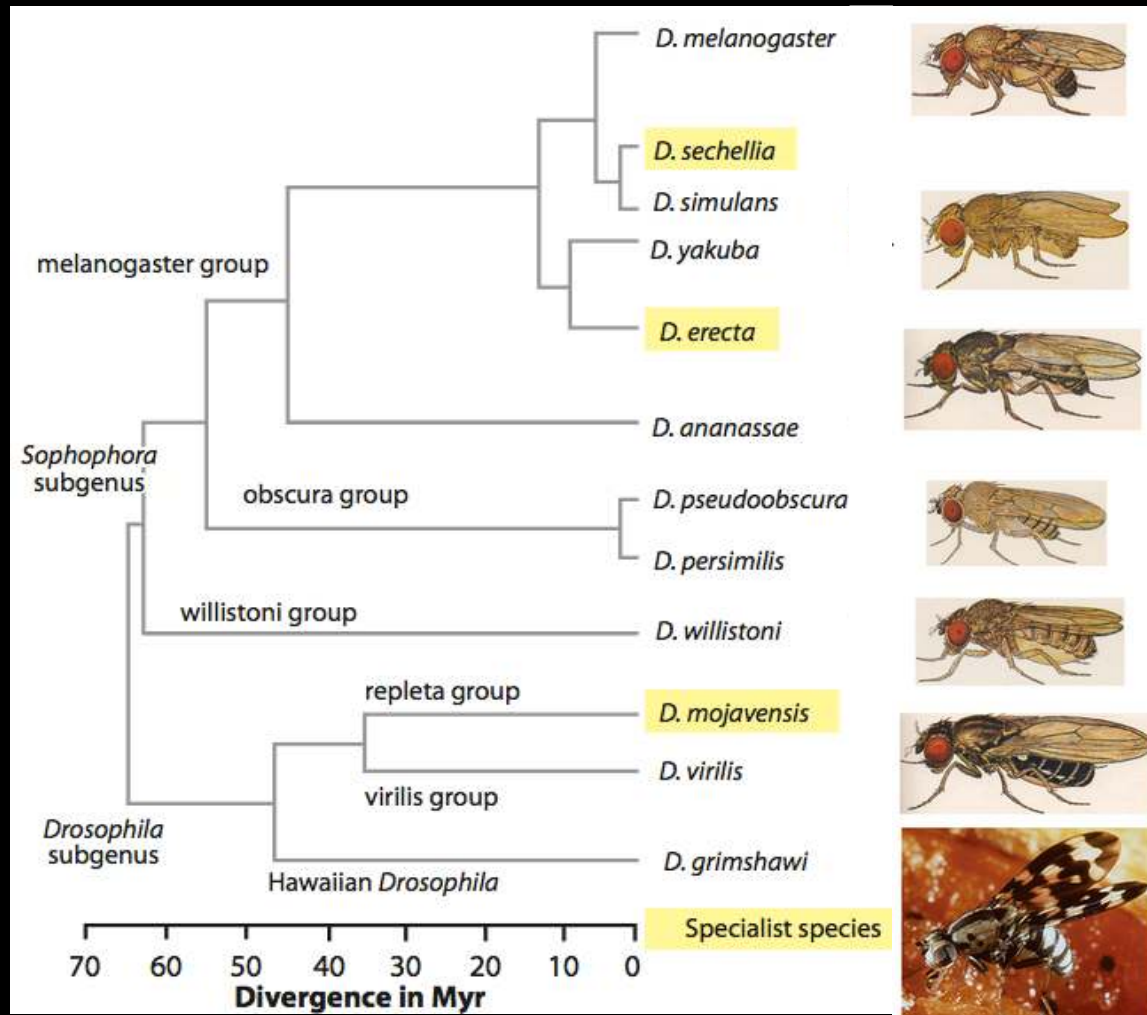
An unprecedented
opportunity for
large scale errors?



studying:
relationships

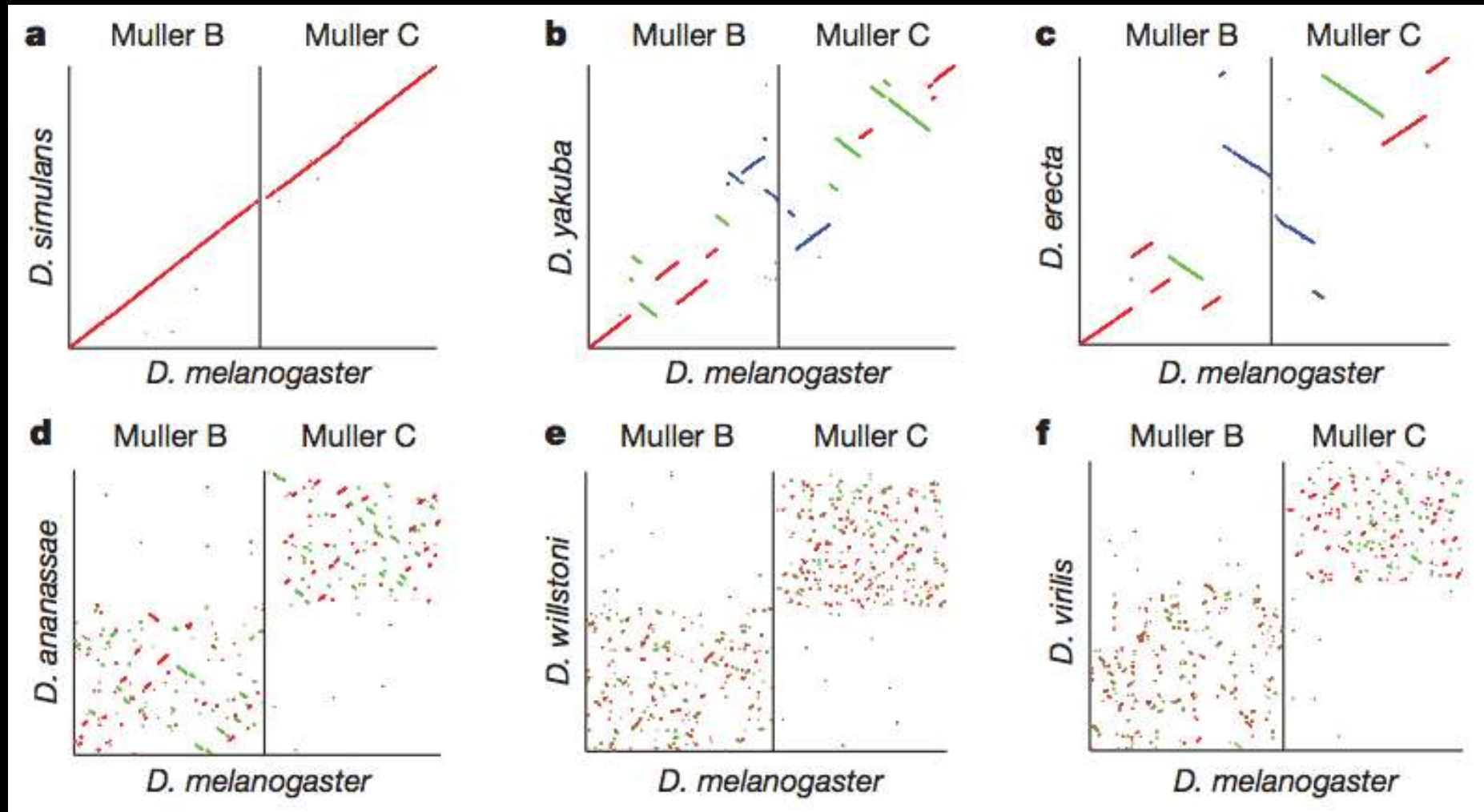
- Genome evolution
- Functional insights into genes and genomic features (e.g. regulation and inheritance)

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



Drosophila 12 Genomes Consortium 2007 Nature

Tempo and mode of chromosome evolution



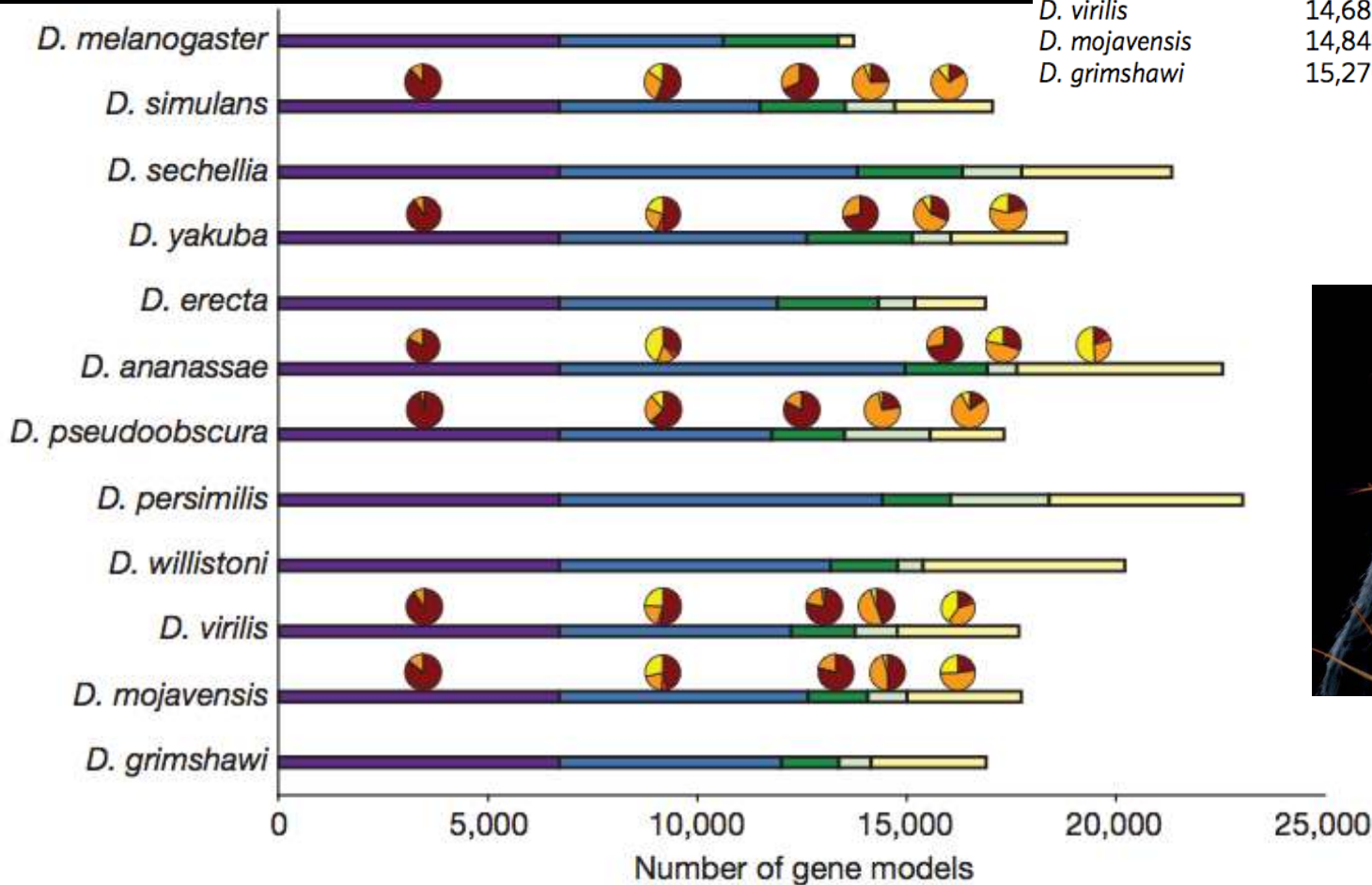
- > 20 My, chromosomal order completely reshuffled in Diptera

Drosophila 12 Genomes Consortium 2007 Nature

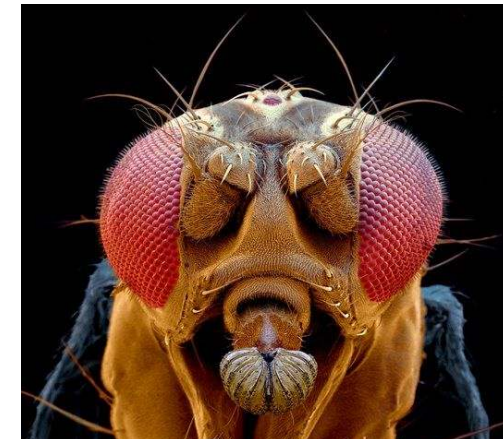
Genome evolution

Drosophila 12 Genomes Consortium 2007 Nature

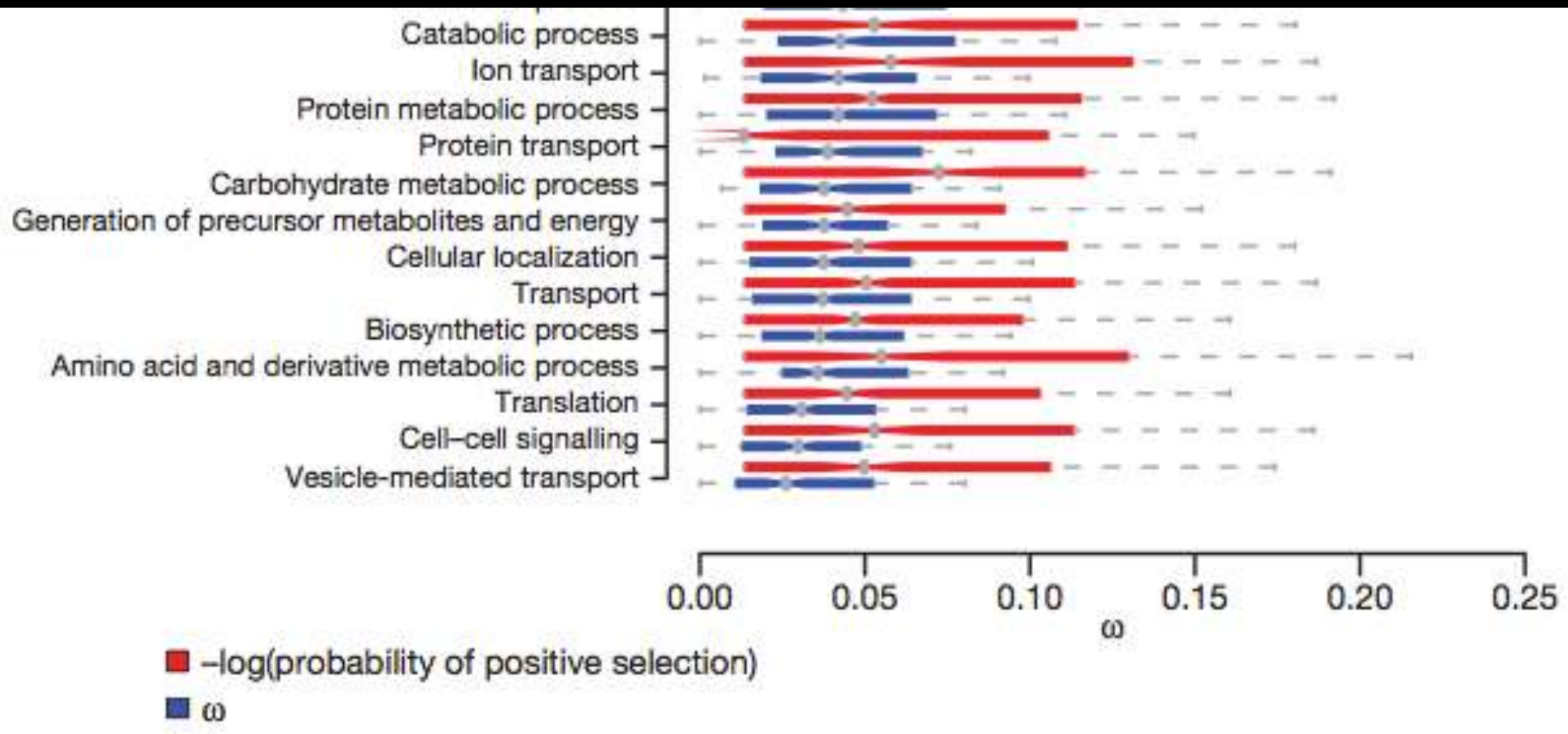
	Total no. of protein- coding genes (per cent with <i>D. melanogaster</i> homologue)	Coding sequence/ intron (Mb)
<i>D. melanogaster</i>	13,733 (100%)	38.9/21.8
<i>D. simulans</i>	15,983 (80.0%)	45.8/19.6
<i>D. sechellia</i>	16,884 (81.2%)	47.9/21.9
<i>D. yakuba</i>	16,423 (82.5%)	50.8/22.9
<i>D. erecta</i>	15,324 (86.4%)	49.1/22.0
<i>D. ananassae</i>	15,276 (83.0%)	57.3/22.3
<i>D. pseudoobscura</i>	16,363 (78.2%)	49.7/24.0
<i>D. persimilis</i>	17,325 (72.6%)	54.0/21.9
<i>D. willistoni</i>	15,816 (78.8%)	65.4/23.5
<i>D. virilis</i>	14,680 (82.7%)	57.9/21.7
<i>D. mojavensis</i>	14,849 (80.8%)	57.8/21.9
<i>D. grimshawi</i>	15,270 (81.3%)	54.9/22.5



■ Single-copy orthologues ■ Conserved homologues ■ Patchy homologues (with *mel.*) ■ Patchy homologues (no *mel.*) ■ Lineage specific



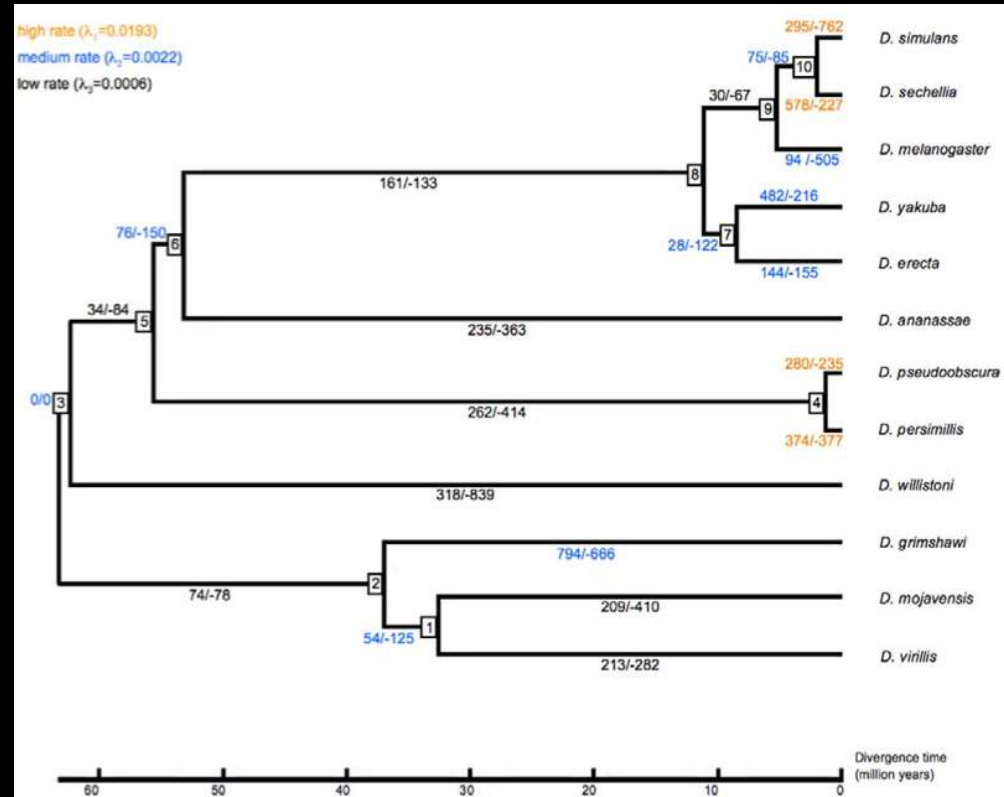
Selection dynamics across functional categories



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

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 ...
 - Its likely 50% of what we think we know is wrong
- All conclusions, at some stage, rest upon
 - Simple bioinformatics
 - Assumptions that get incorporated into seemingly unbiased methods
- Exploring two pillars of these studies, their error and repercussions
 - Gene alignments in detecting positive selection
 - Calibrations in temporal analysis

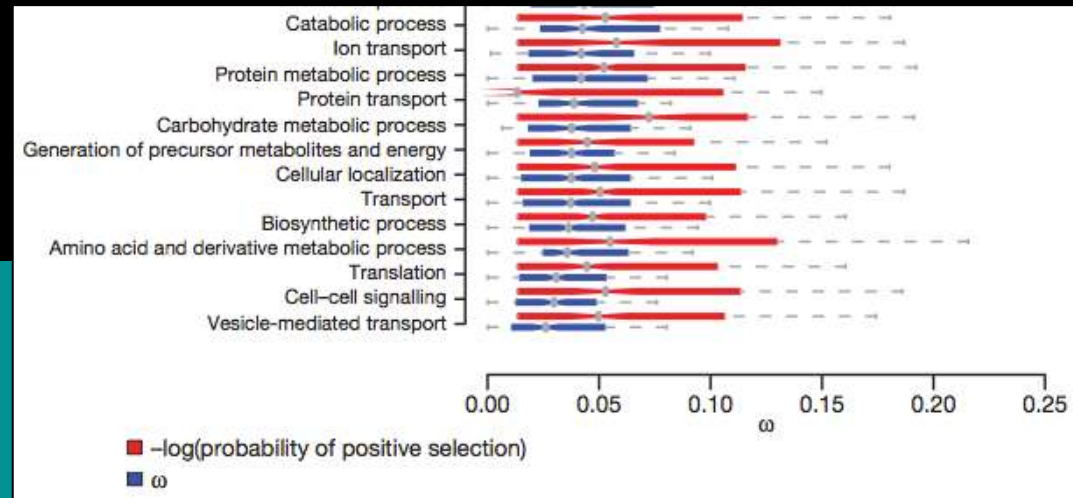


Established studies allow ...

Follow up studies to reveal limitations

Robust findings to emerge with age

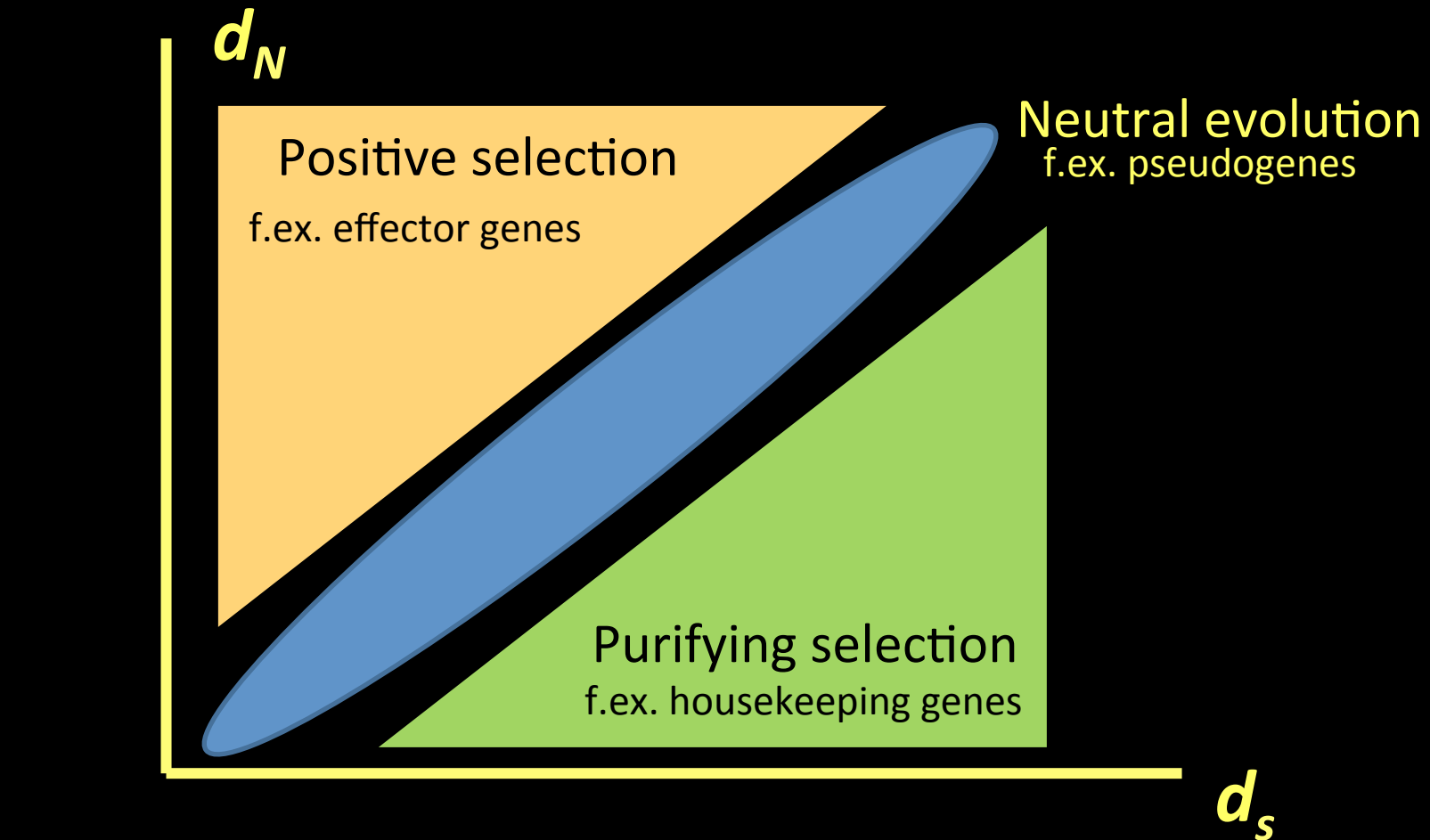
Inferring selection dynamics:



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

How robust are these conclusions?

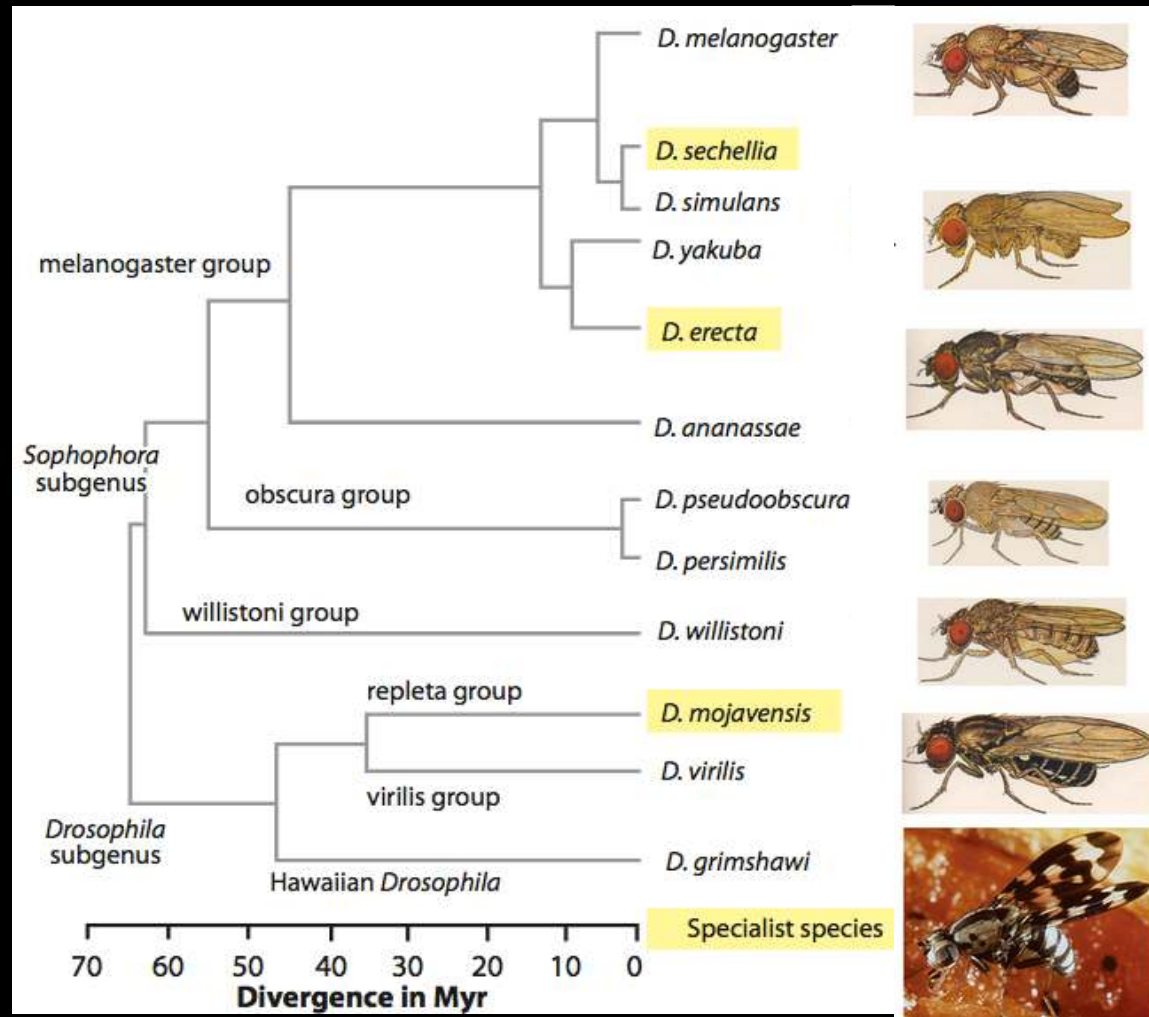
Codon based tests of selection



d_N / d_s
ratio

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

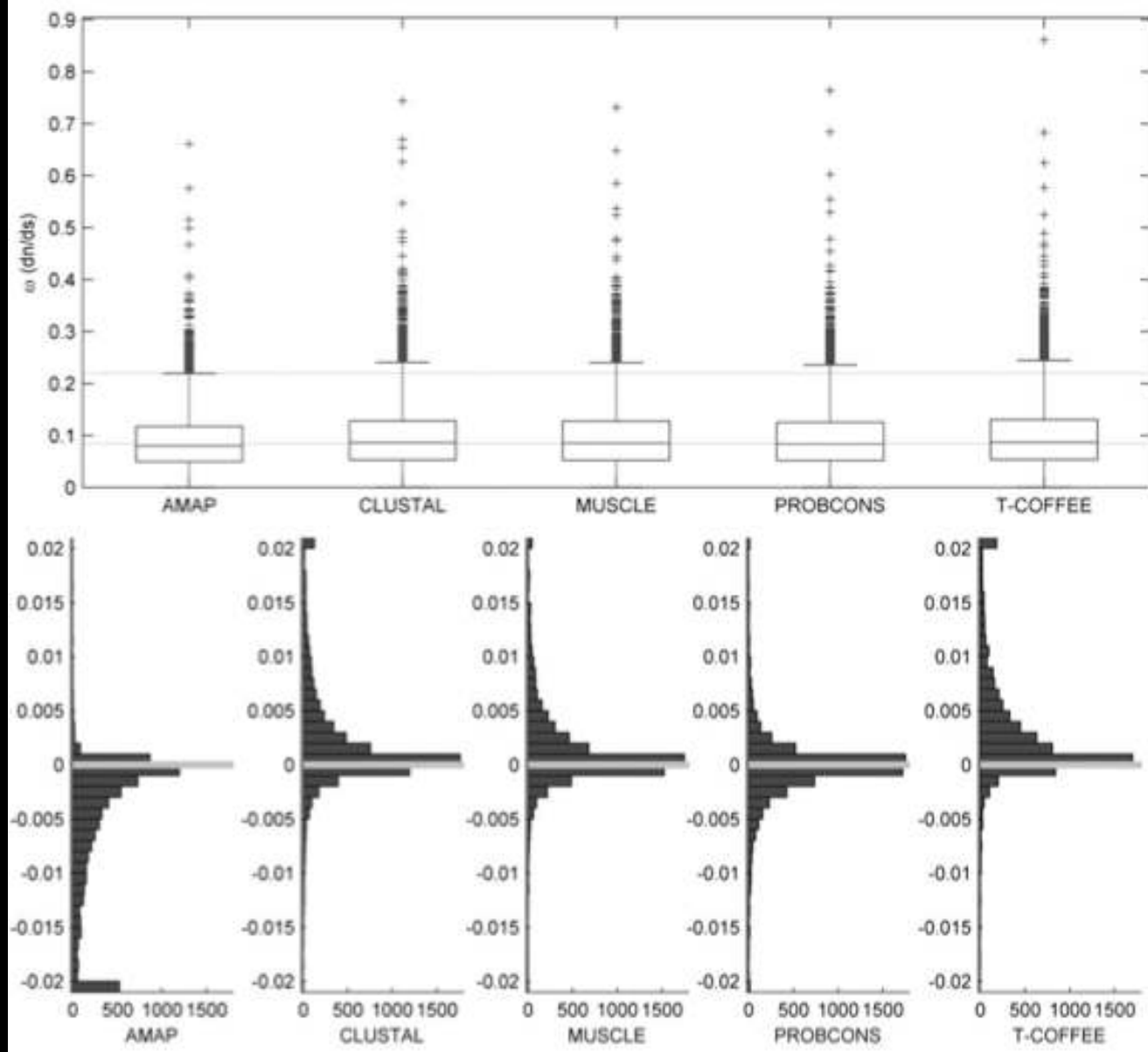
Evolution of genes and genomes on the *Drosophila* phylogeny



Drosophila 12 Genomes Consortium 2007 Nature

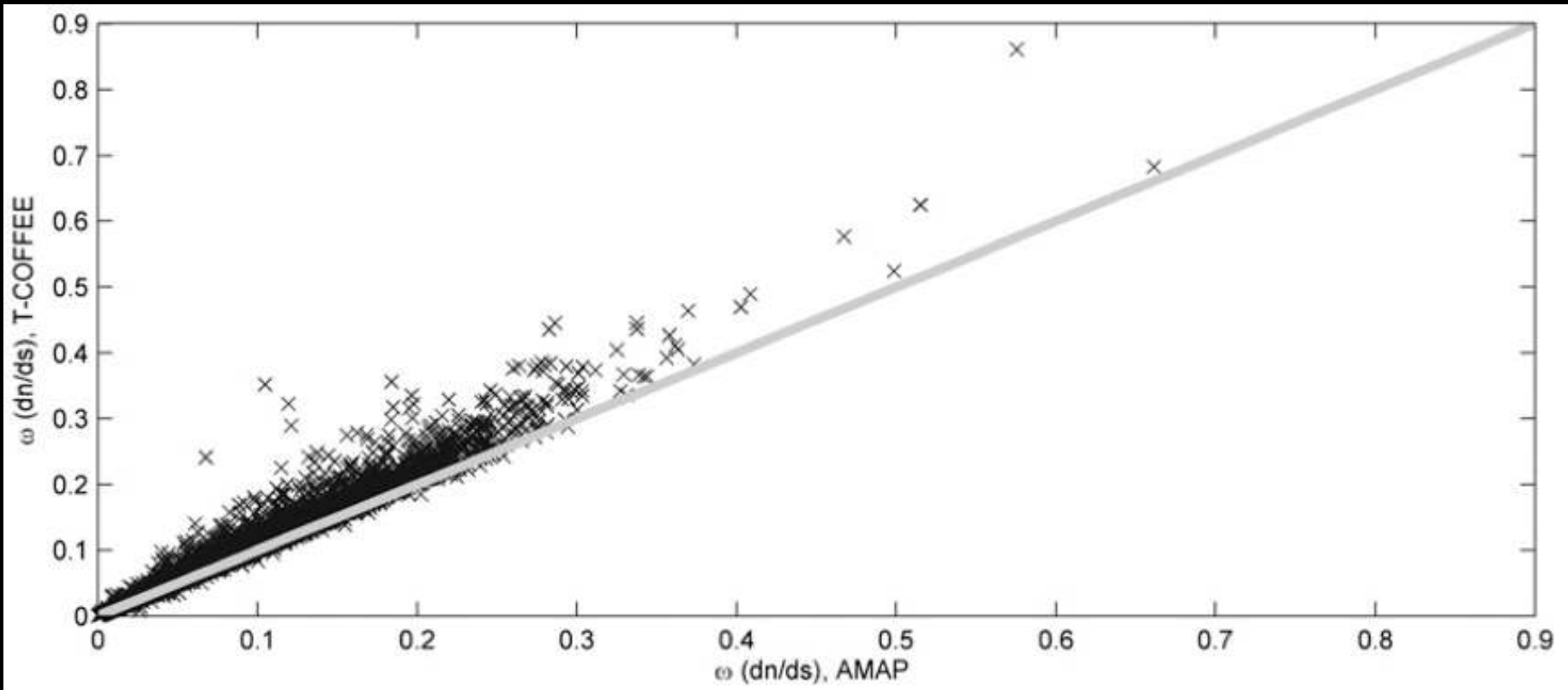
dN/dS estimates by aligner

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



Comparing results across methods is responsible bioinformatics!!!!

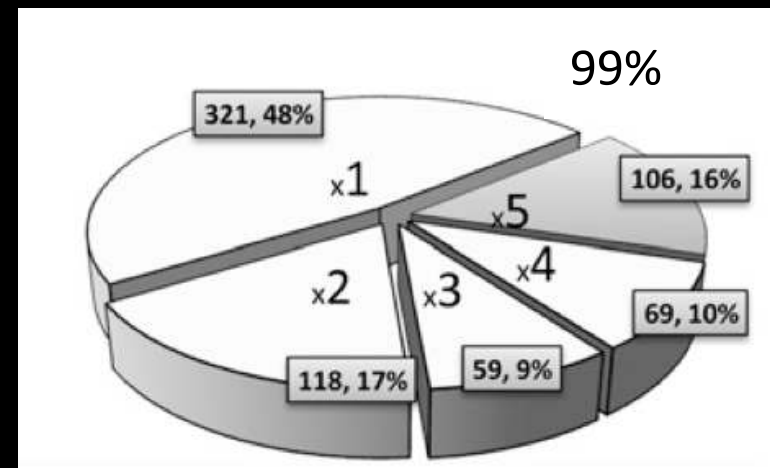
Since we can't look at our data, we need approaches that
allow 1st principal assessments



Aligner tool has a larger effect than biology

	12 genomes, M7/8		12 genomes, M1a/2a		12 genomes, M7/8, with removed gaps		<i>Melanogaster</i> group, M7/8	
Aligner	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

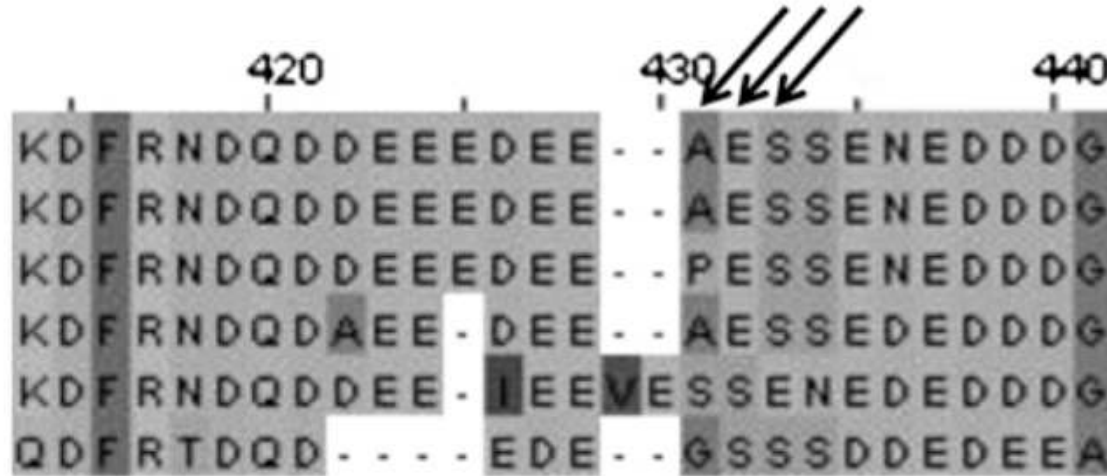
Number of significant genes in common across 1, 2, 3, 4, or all 5 of the alignment methods



Alignment results highlight importance of alignment score!

- Tcoffee finds 3 selected sites indicated by arrows
- ProbCons identifies region with low alignment score, not used

Tcoffee



ProbCons



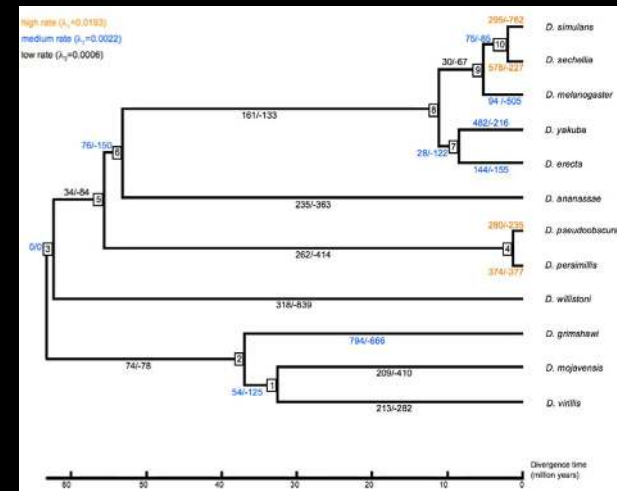
Temporal inference:

fact or fiction?



Timing of divergence

- Directly affects rate estimates
- Deriving unbiased dates from molecular data
 - Large field of software development
- Bayesian methods, while potentially informative and unbiased
 - Can be easily, and are routinely, abused



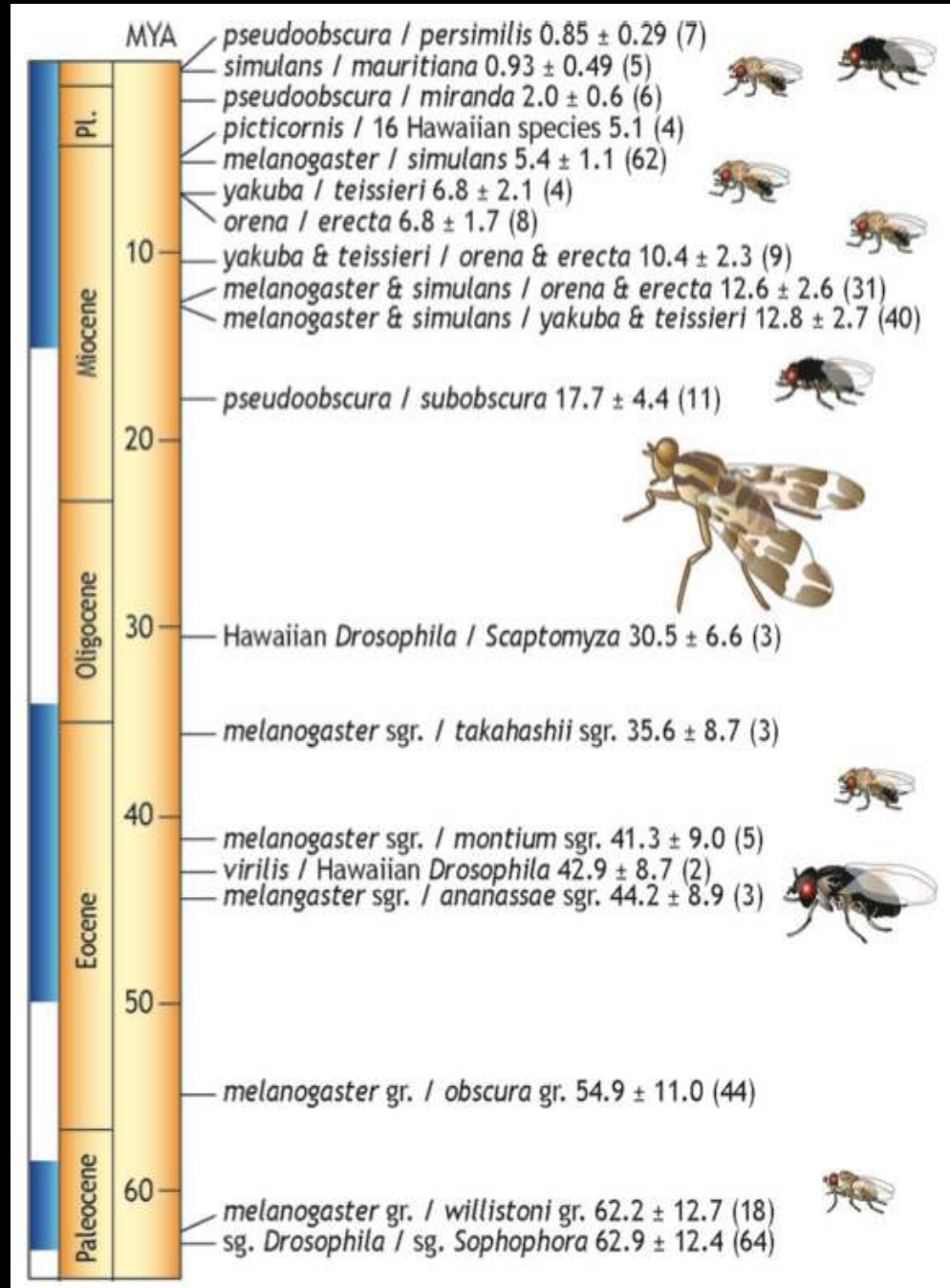


Hawaiian Islands

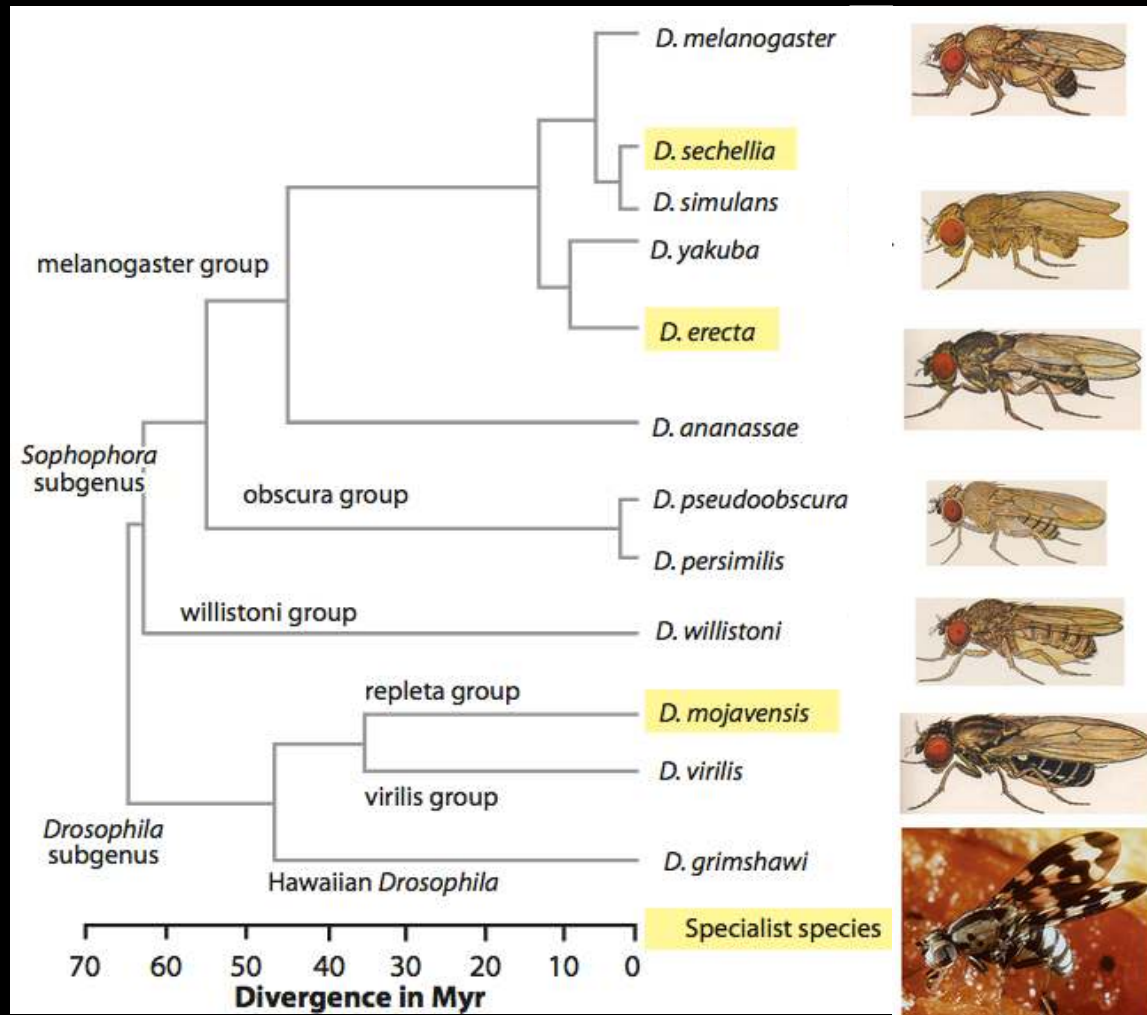
Calibration: Kauai age of 5.1 my for divergence of two Hawaiian species

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

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



Evolution of genes and genomes on the *Drosophila* phylogeny



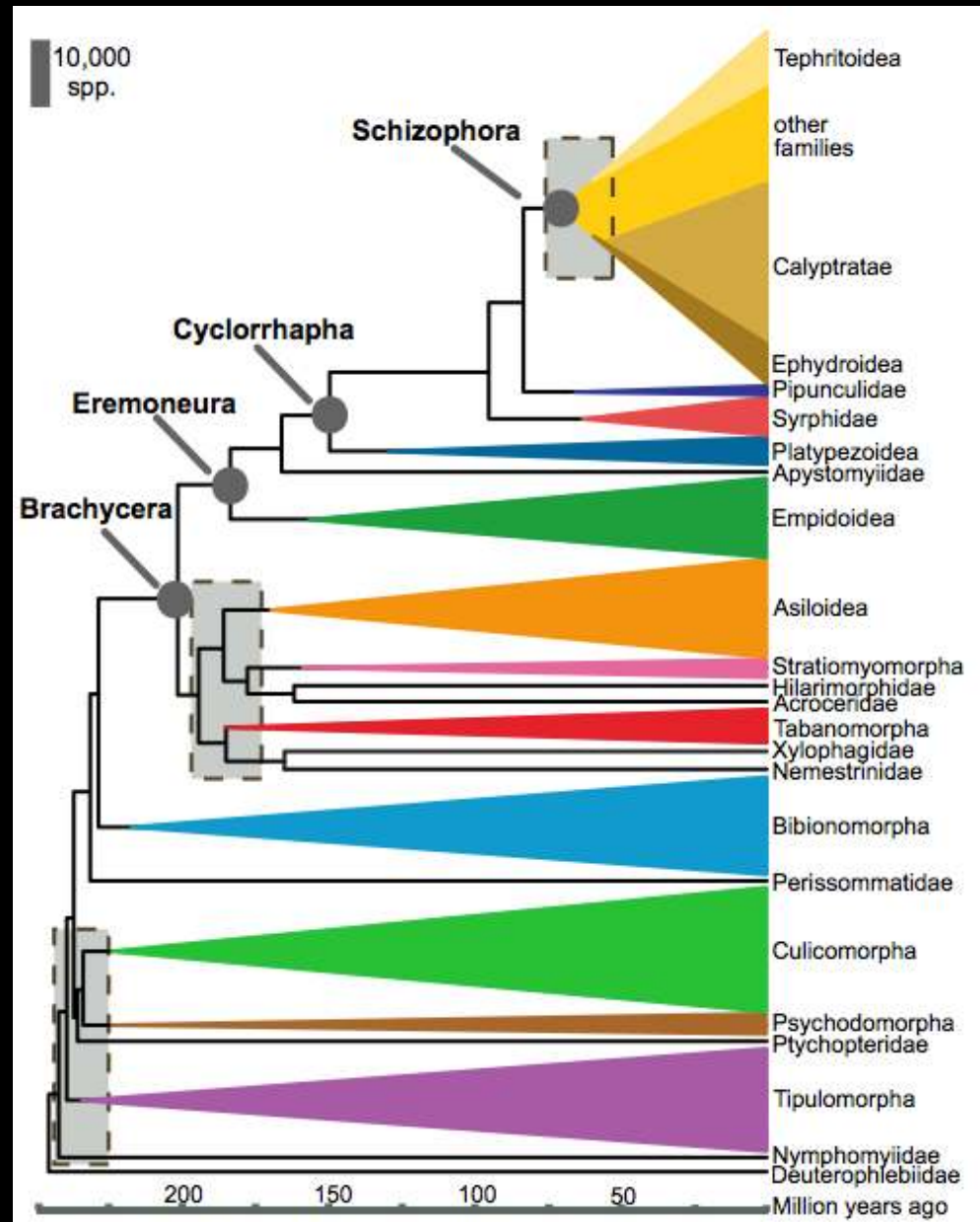
Drosophila 12 Genomes Consortium 2007 Nature



Drosophila clade:

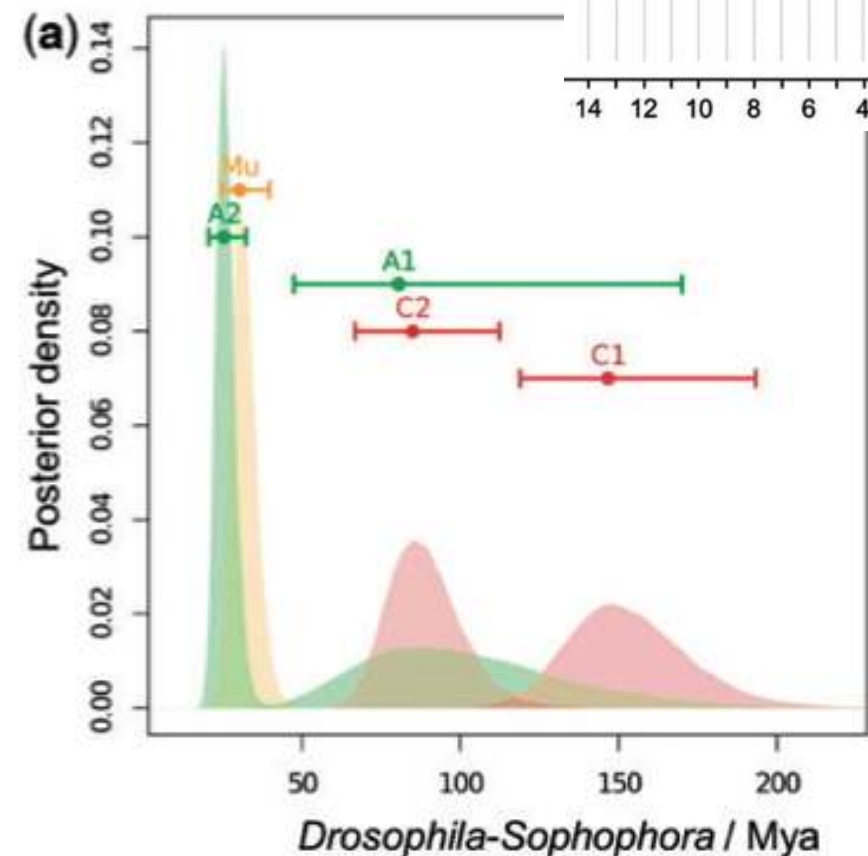
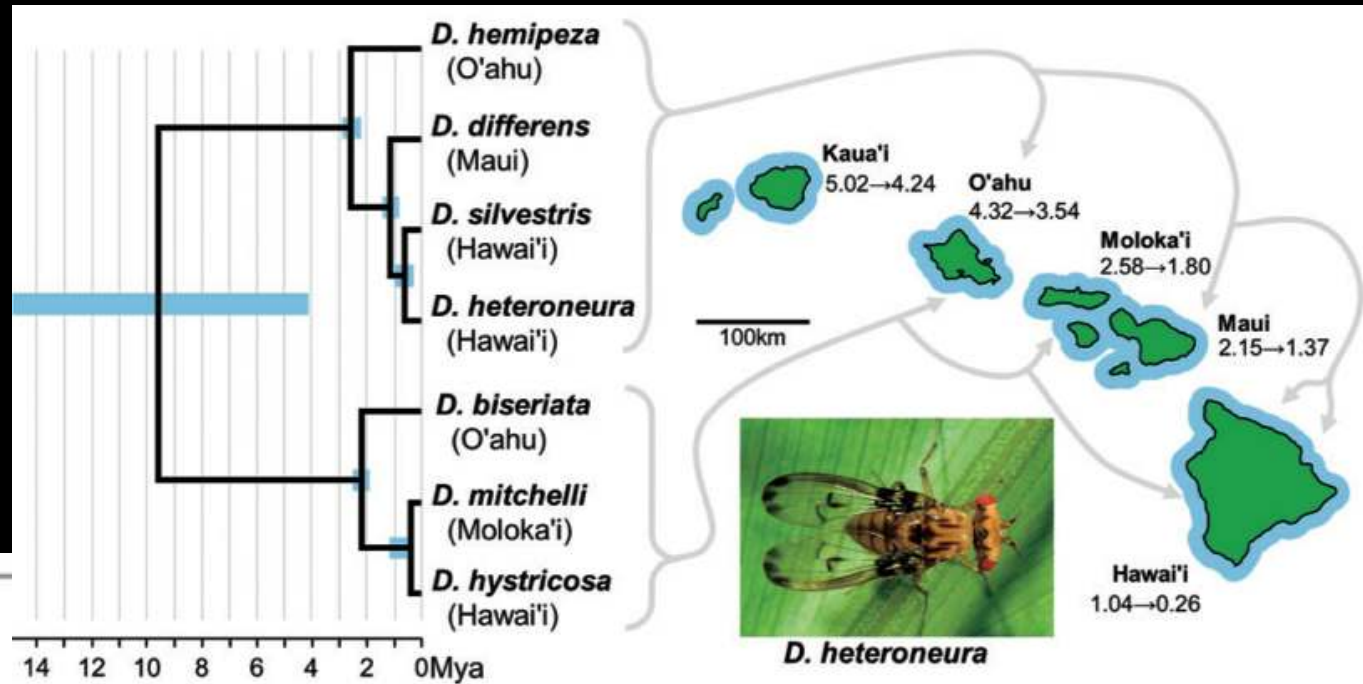
- Schizophora constrained to maximum of 70 Ma
- Without constraint, goes to 115 Ma

What is reality?



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

Determining objective priors is challenging



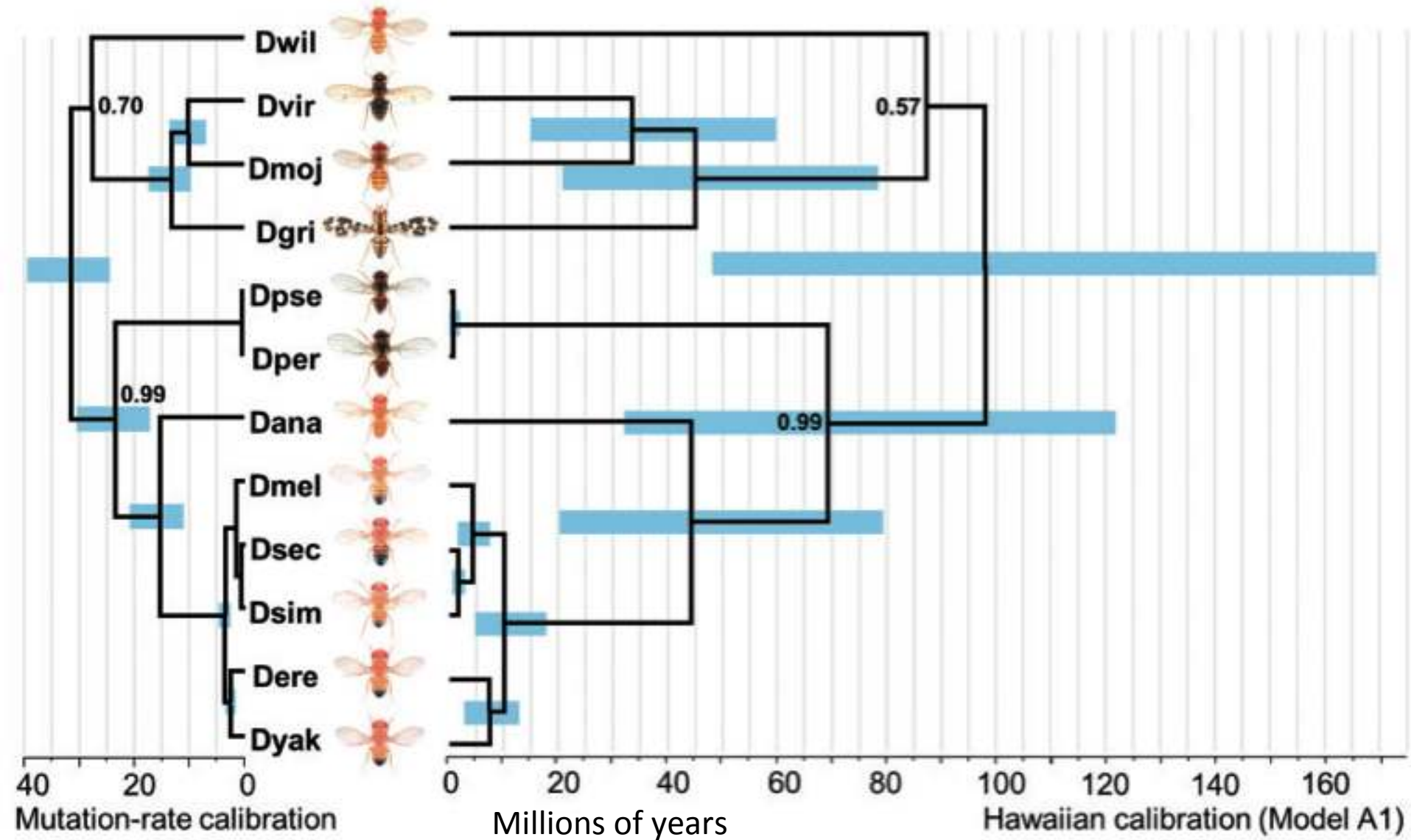
Priors in Bayesian rel. clock analysis:

μ = lab observed mutation rate

A1,2 = geological calibration, small N_e

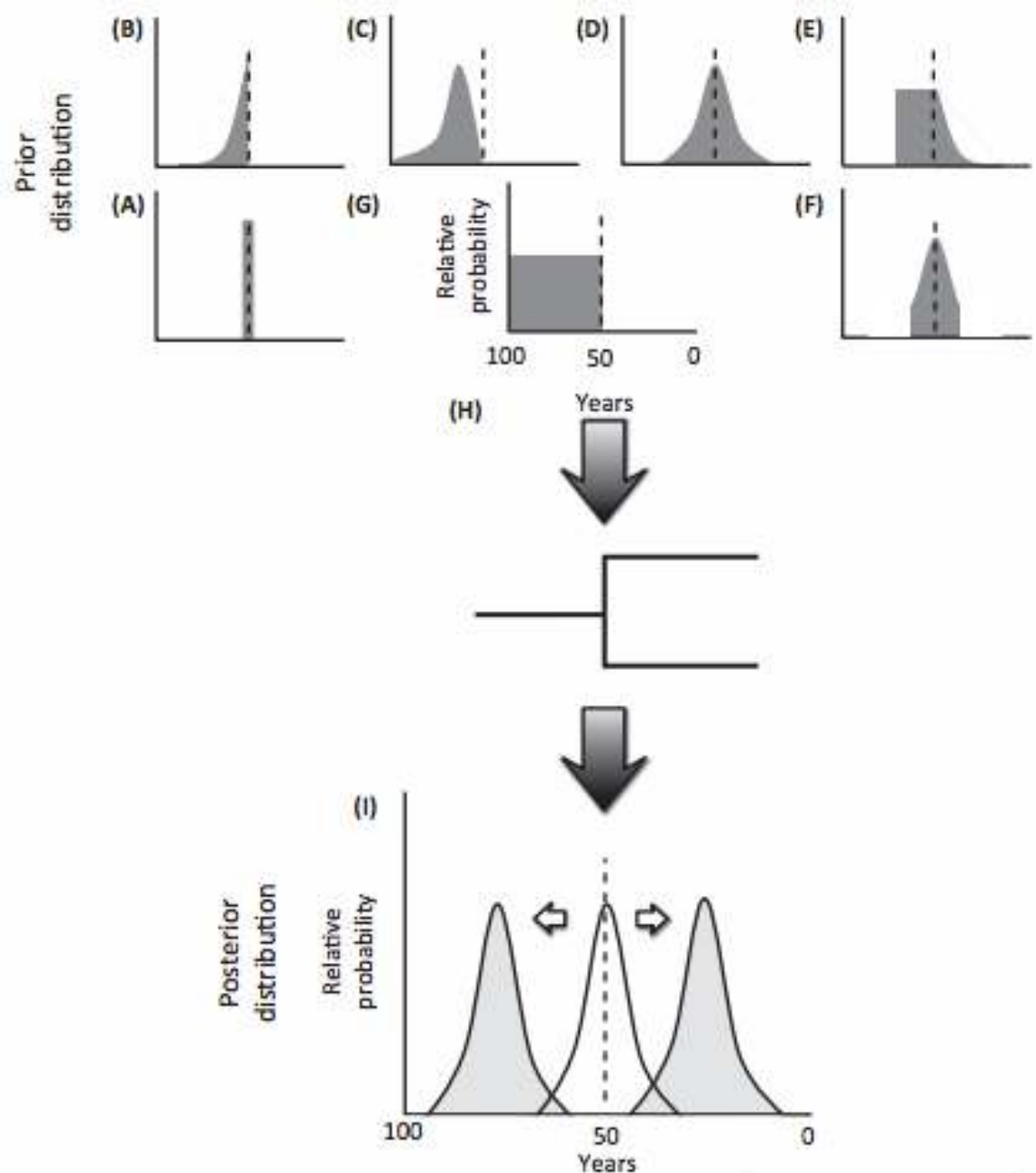
C1,2 = geological calibration, large N_e

Priors directly influence posteriors



Prior distributions matter

- Integrative science is challenging
- Discuss or collaborate with experts to evaluate your approach.



How do we gain dating confidence when we are in the dark?

- Fossils and DNA are likely to rarely agree
- How can we assess the temporal signal in the DNA in a robust manner?
 - Reducing prior biases and using lots of DNA data, while modeling likely violations of analysis models



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 quantify”

- Wheat 2015

For example:

- indels & inversions**
- gene family dynamics**
- evolutionary dynamics**

What does a
good
P-value
really tell
you?

What type
of
selection?

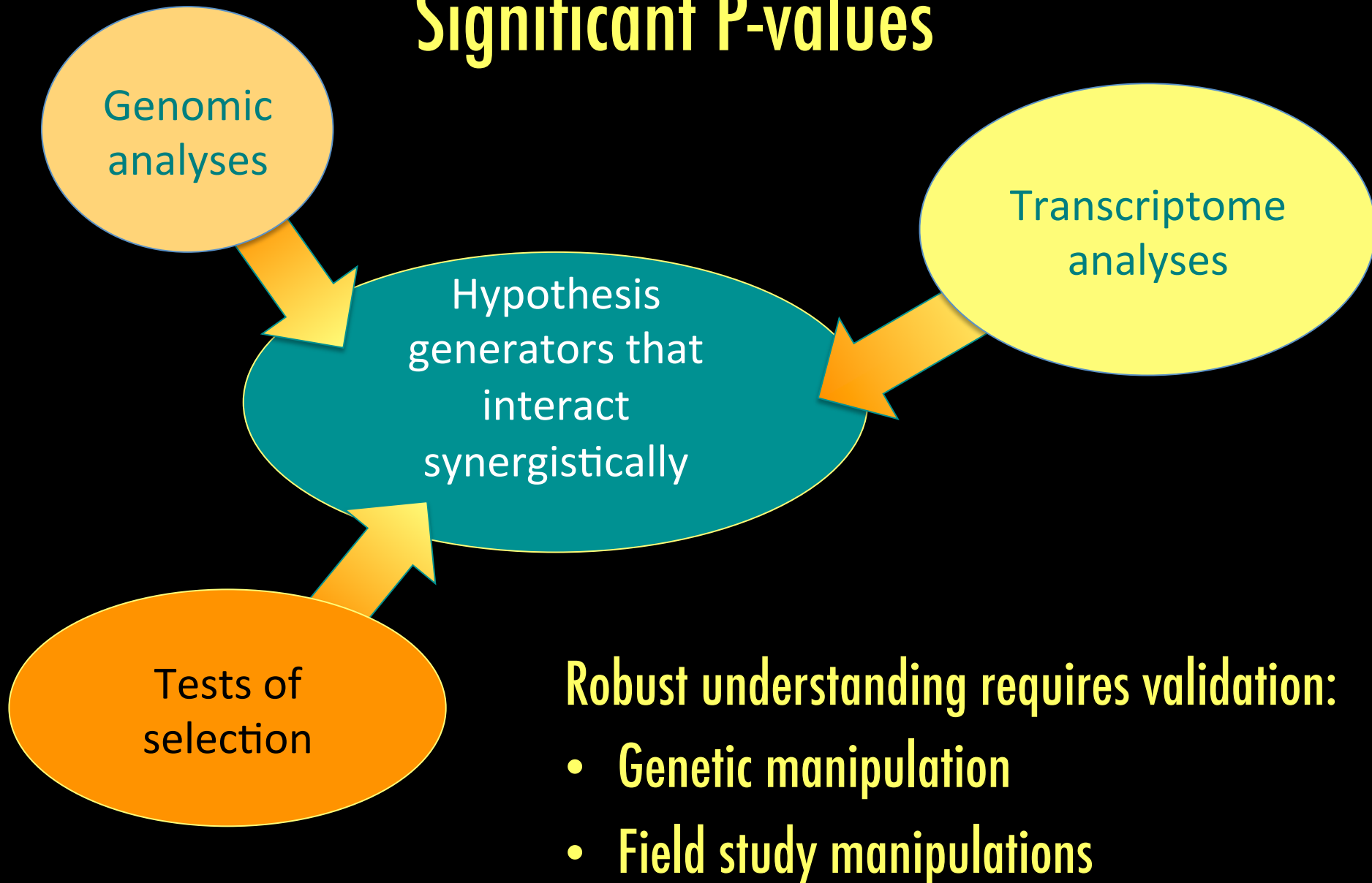
Is method
mismatched
to
mechanism?

Are you
chasing a
good P-
value?

When
did
selection
happen?

What does a
bad
P-value
really tell
you?

Significant P-values



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

JOURNAL OF NEGATIVE RESULTS

- ECOLOGY & EVOLUTIONARY BIOLOGY -

Now handling genomic data

<http://www.jnr-eeb.org/index.php/jnr>

Microevolution effects

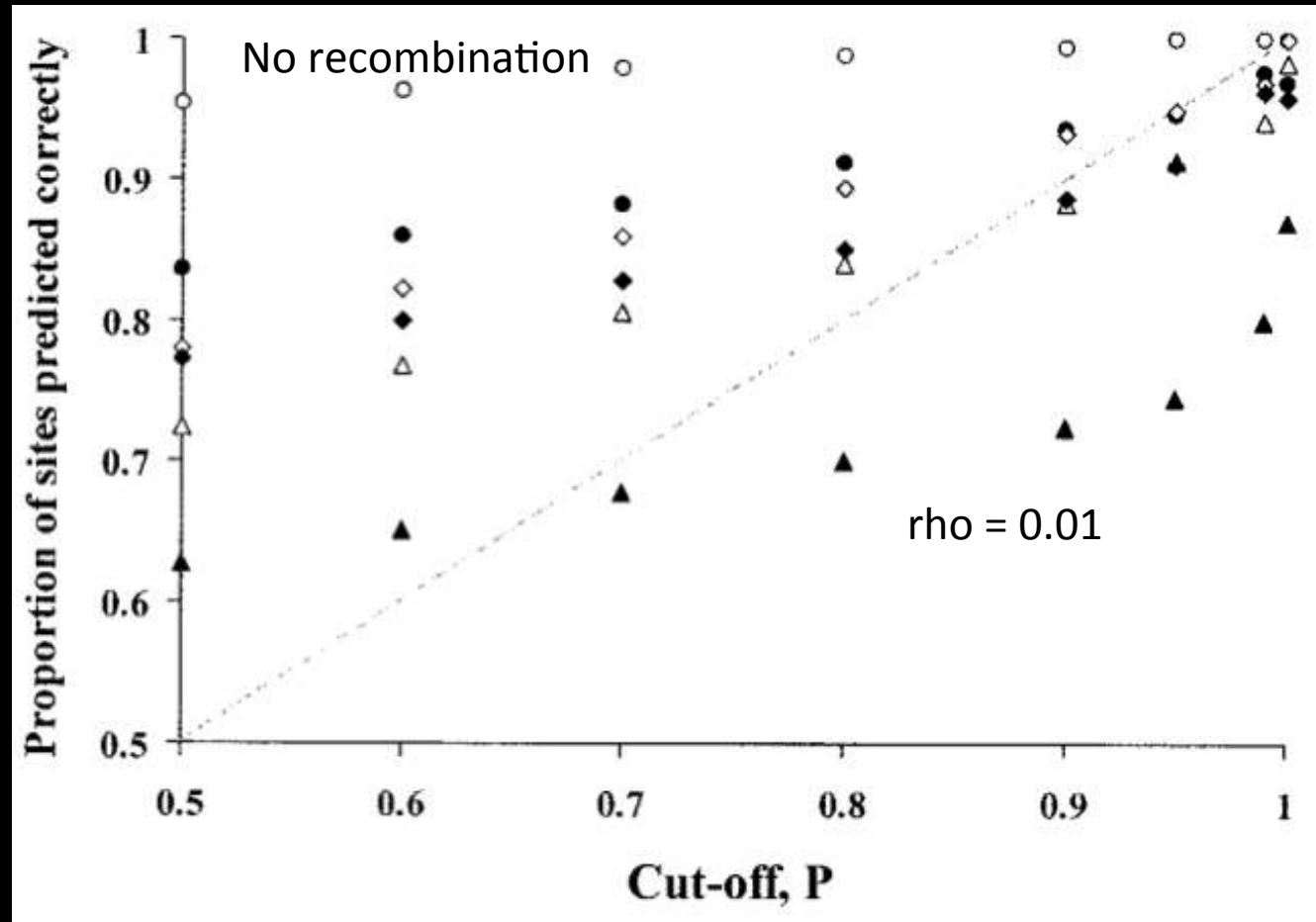
Previous examples were at deep evolutionary time scales

Surely such problems don't exist at the within genera level Right?

Recombination violates dN/dS tests

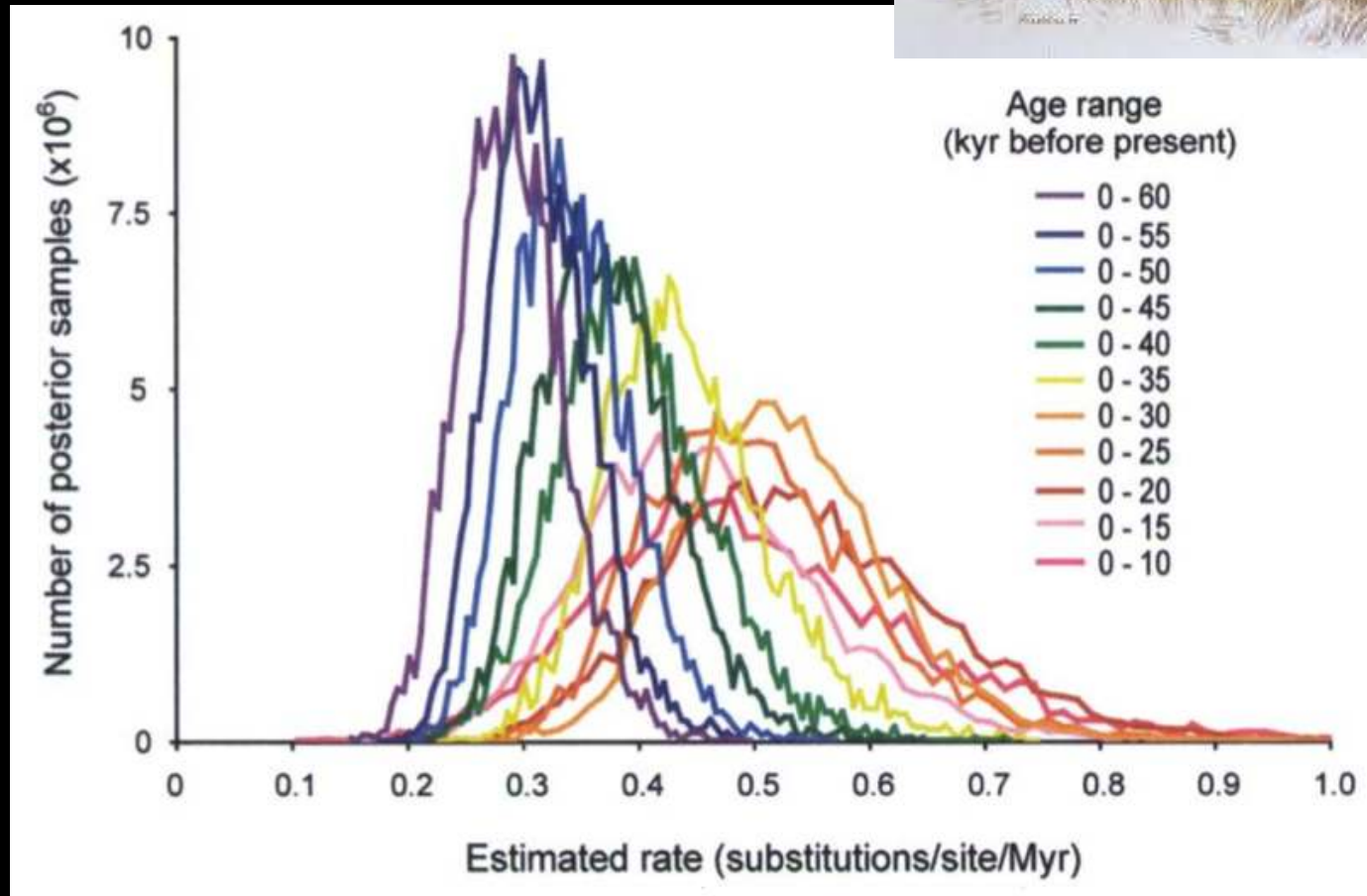
Codeml
inferred
selection:

False
positives can
increase to
over 30%



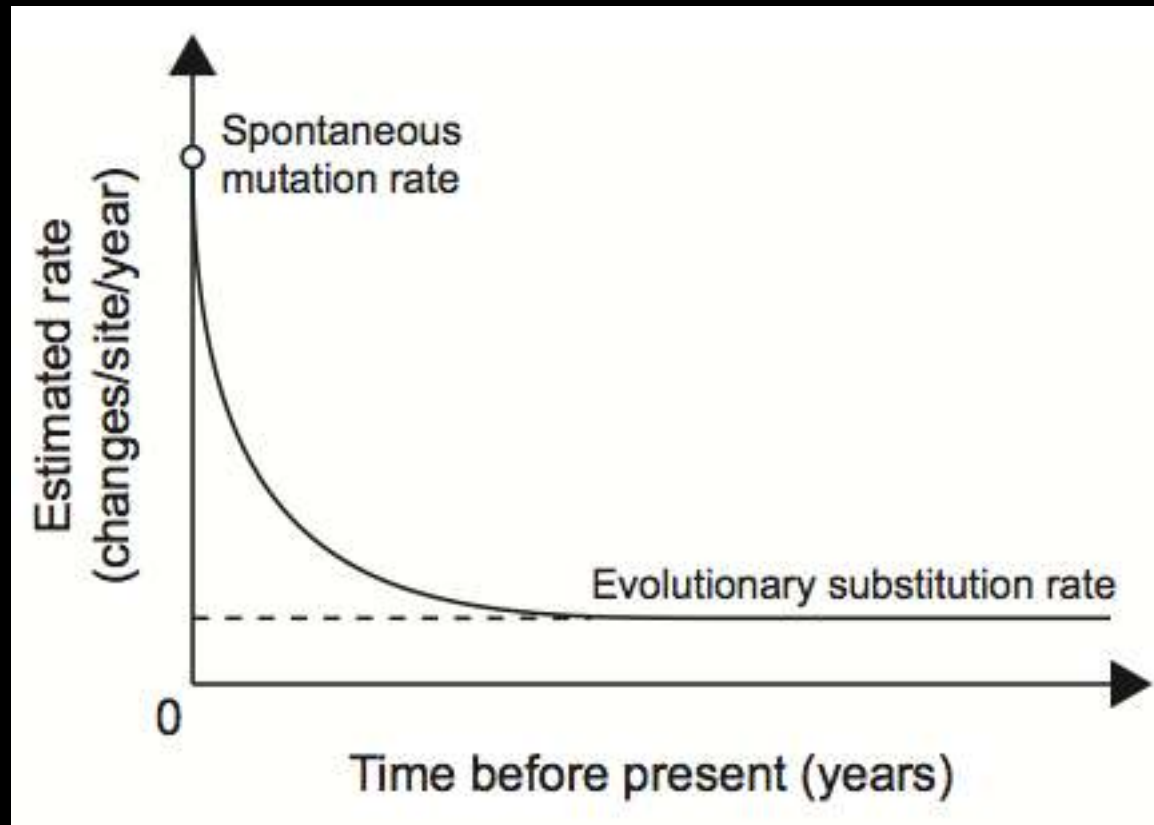
- 13% of sites simulated at $\omega = 2.5$
- Sample size = 30 sequences

Posterior distribution estimates of substitution rates from mitochondrial control region from Beringian bison



Time dependent rates of molecular evolution

Significant implications for phylogeographic studies that use fixed rates to assess demographic with environmental change



... and now for pt. 2

