

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

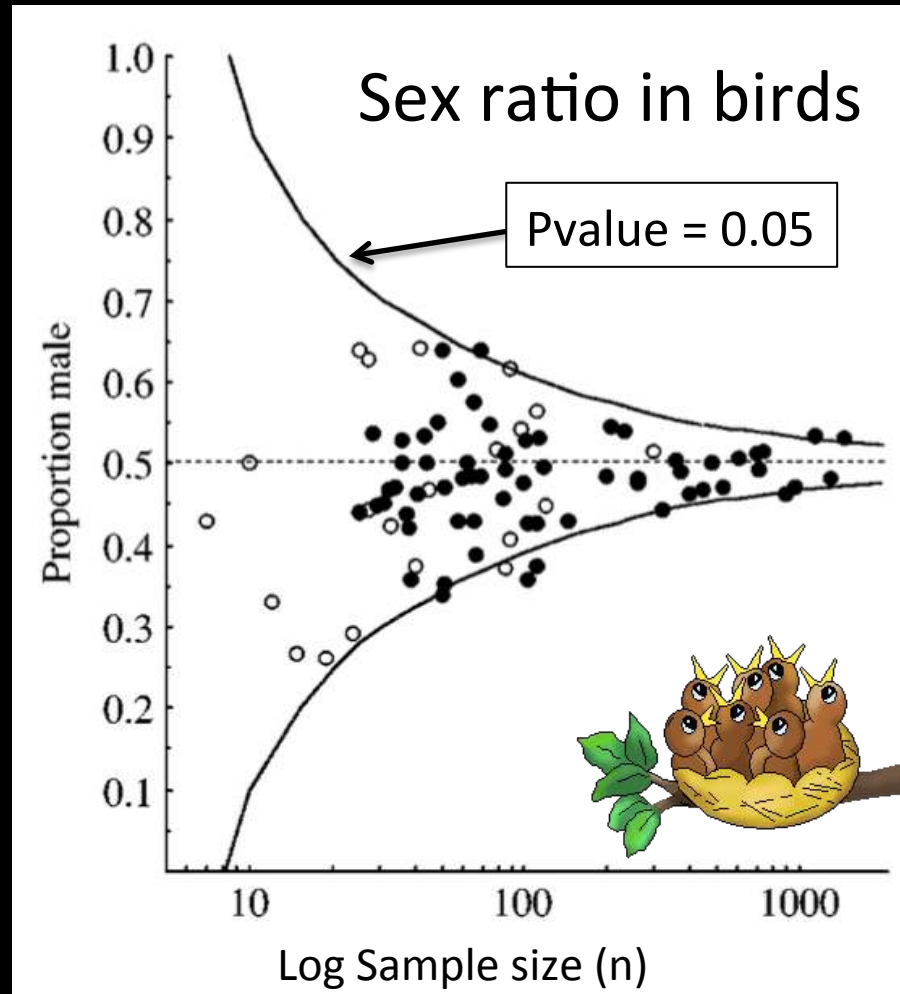
Their findings should be robust:

- **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

Assessing reality using funnel plots



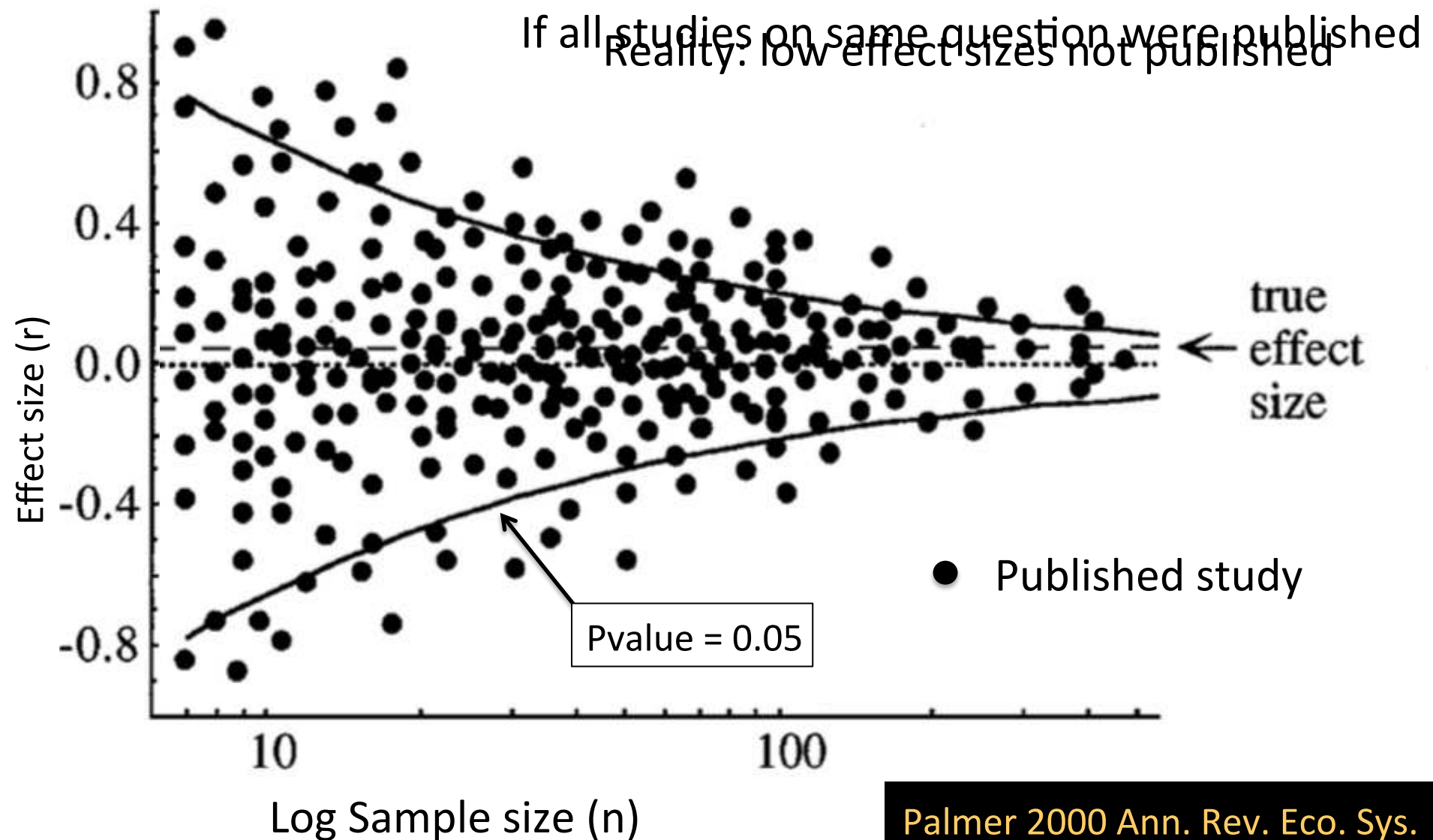
Small sample sizes affect measurement accuracy

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

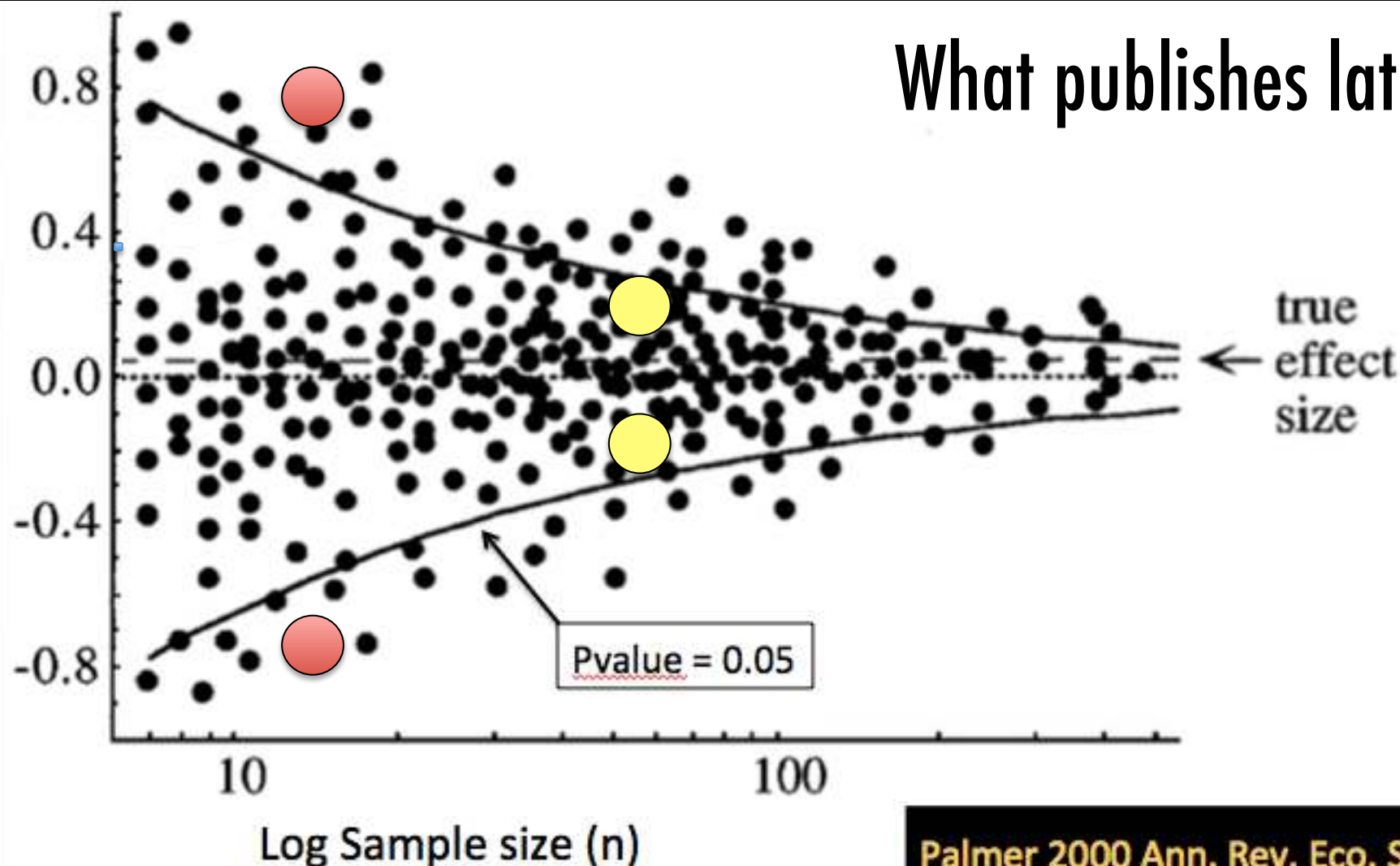
Publication bias increases effect size



What if there is no replication?

What is most likely to publish first & where?

What publishes late?



Why Most Published Research Findings Are False

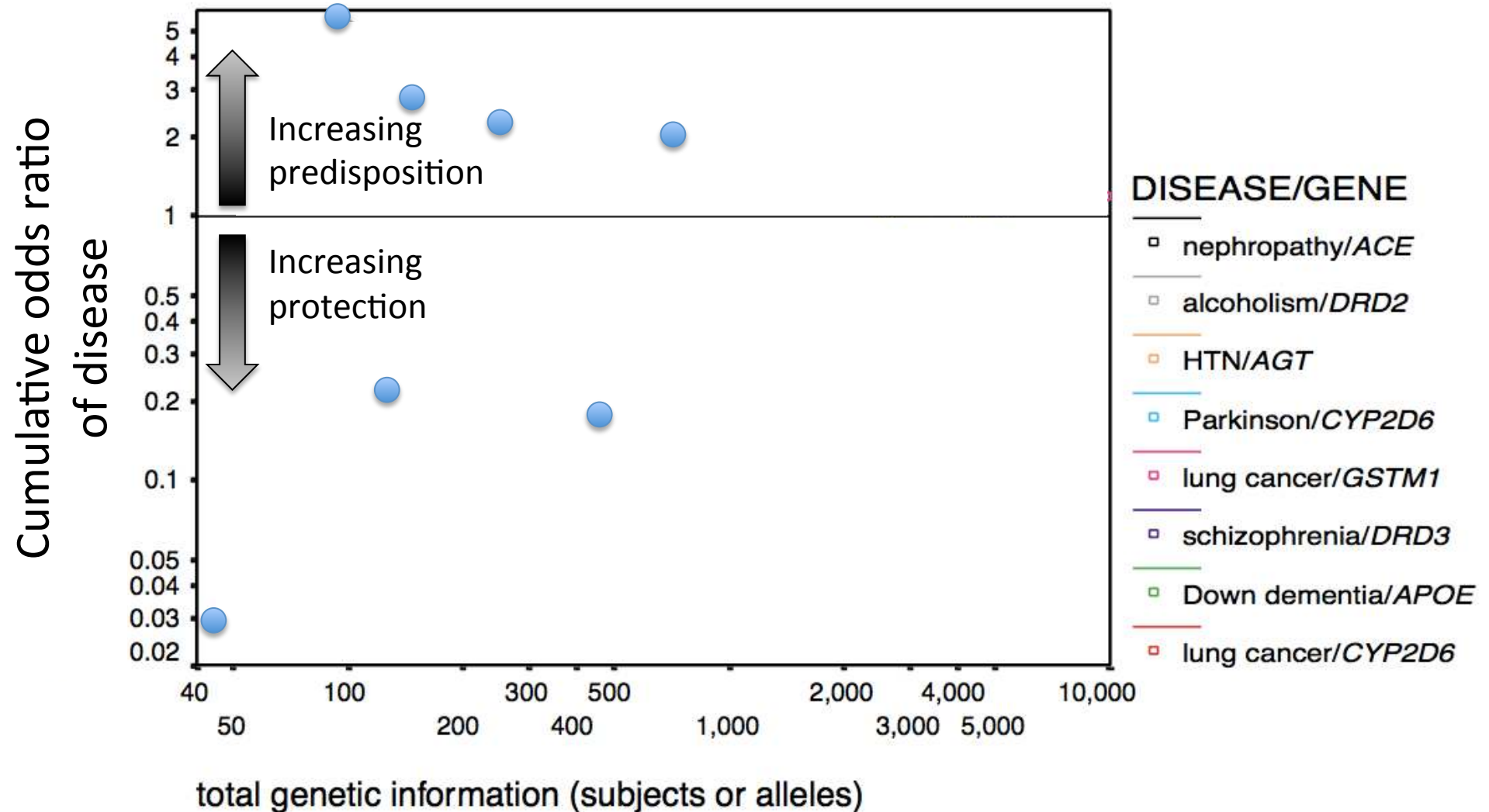
A research finding is less likely to be true when:

- ✓ the studies conducted in a field have a small sample size
- ✓ when effect sizes are small
- ✓ when there 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

**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
 - Space and time
- Publication bias
 - 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

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



- Similar to Type 1 error
 - false positive
- Opposite from Type 2 error
 - false negative

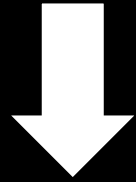
Outline

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

Non – adaptive



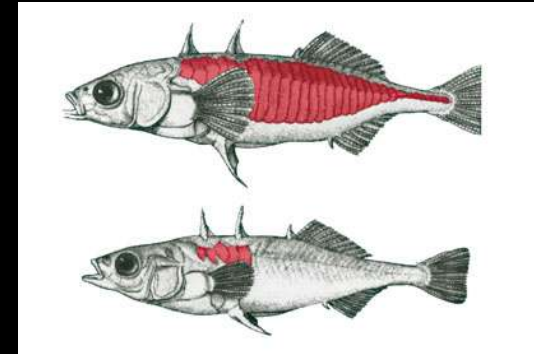
disease, aging, height, etc.



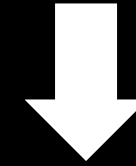
1000's of loci, each of
small effect size

generally ...

Adaptive



salinity, color, resistance, etc.

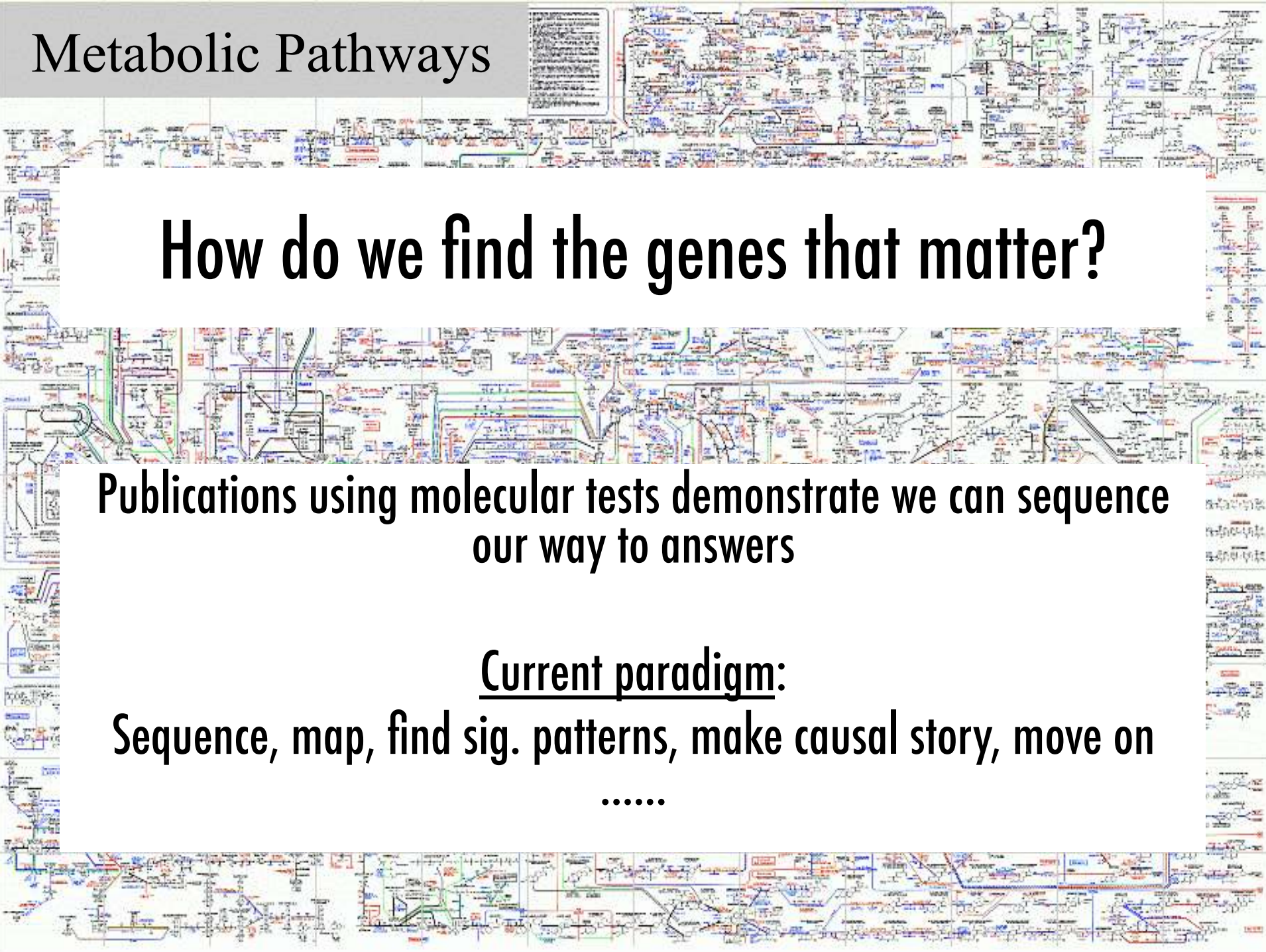


One or several loci of large
effect

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?

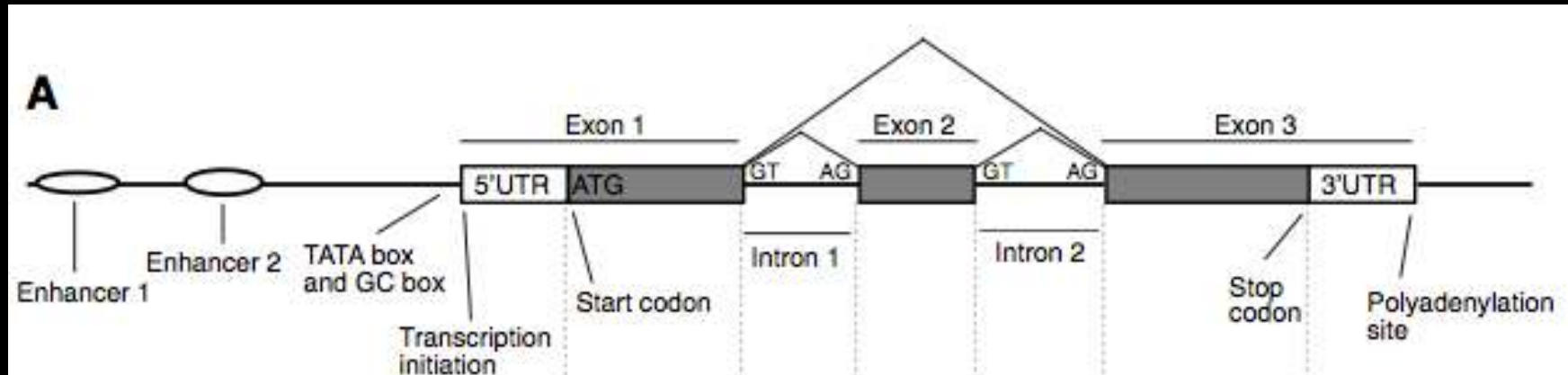
Publications using molecular tests demonstrate we can sequence our way to answers

Current paradigm:

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

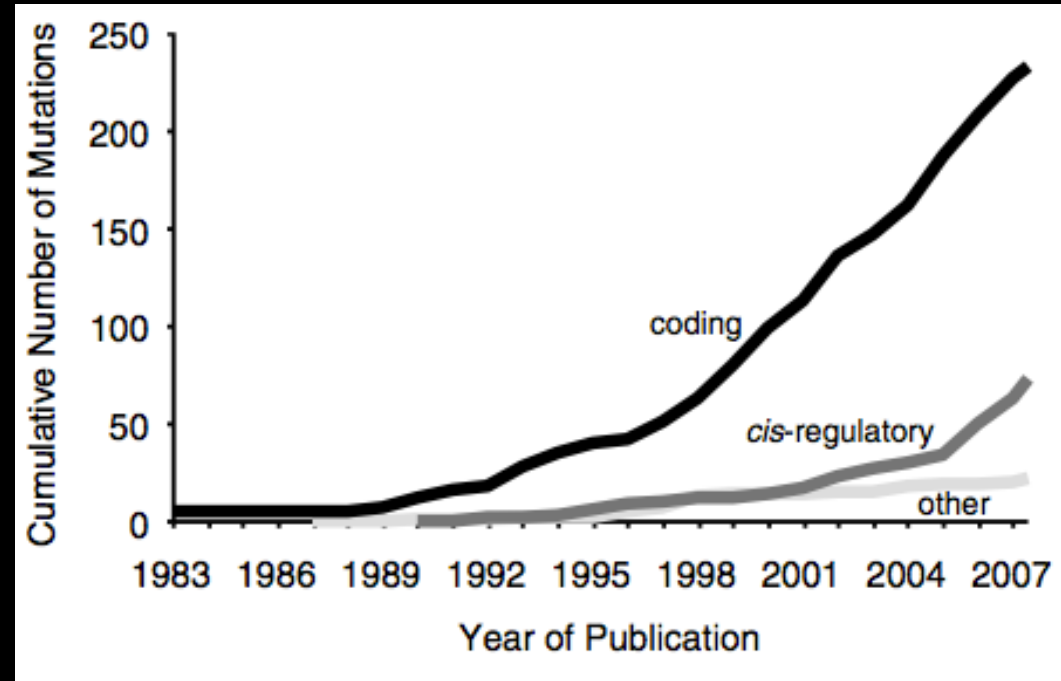
.....

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

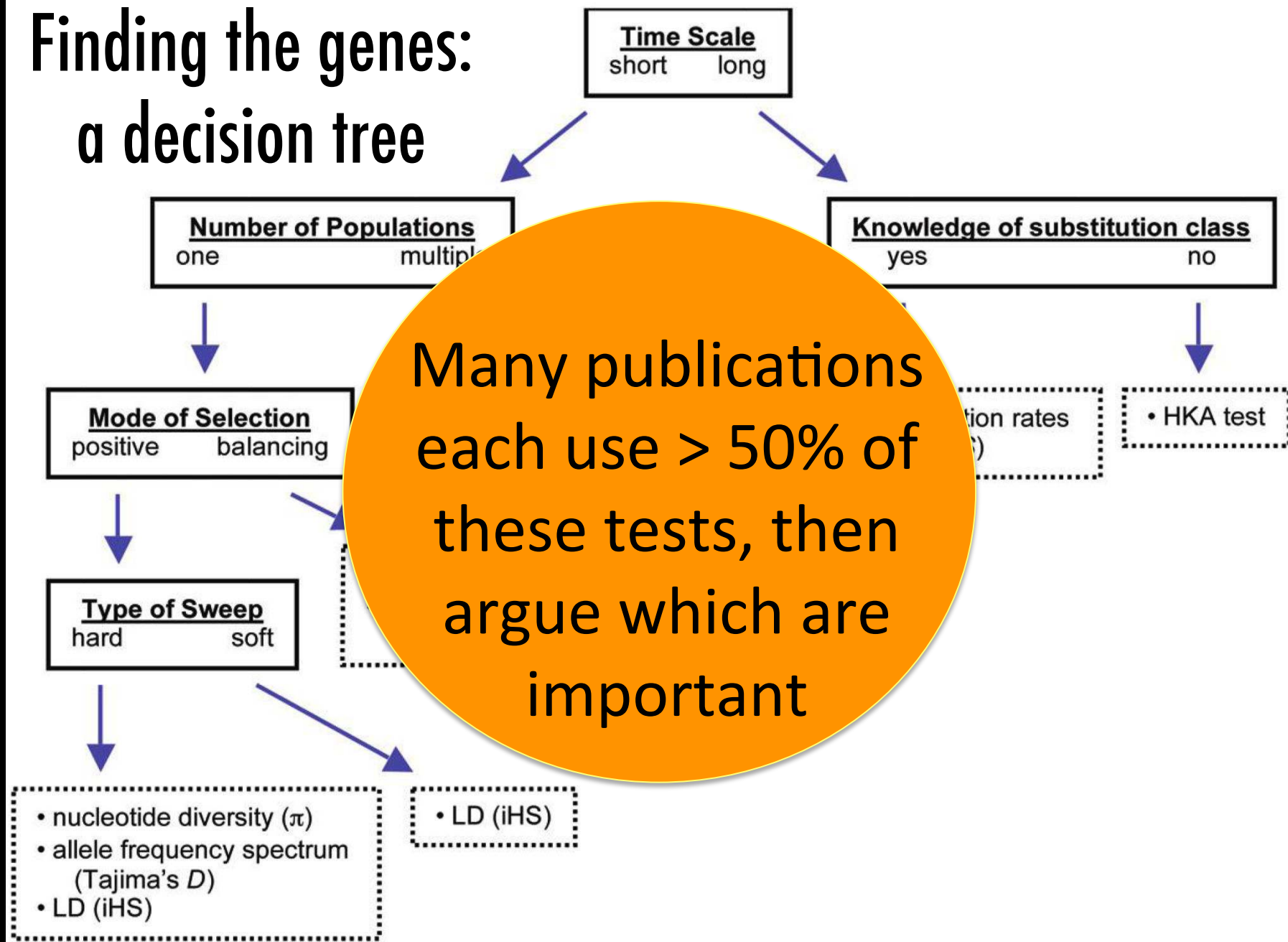


	Morphology	Physiology	Behavior
Coding ³	62	170	2
<i>Cis</i> -regulatory	43	29	2
Other ⁴	3	20	0
Total	108	219	4
Null ⁵	41	58	0

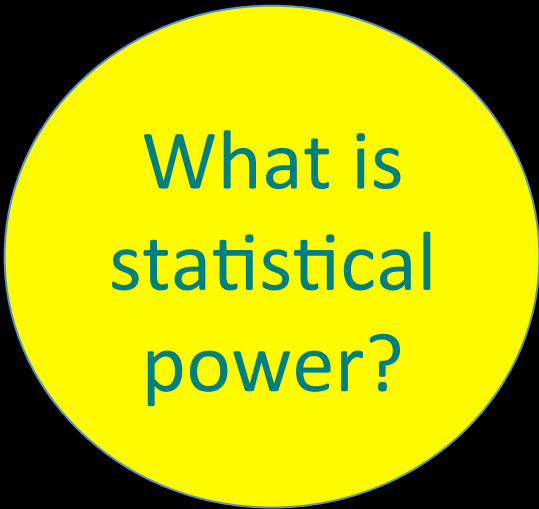
How do we identify the genes that matter?

- 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
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?

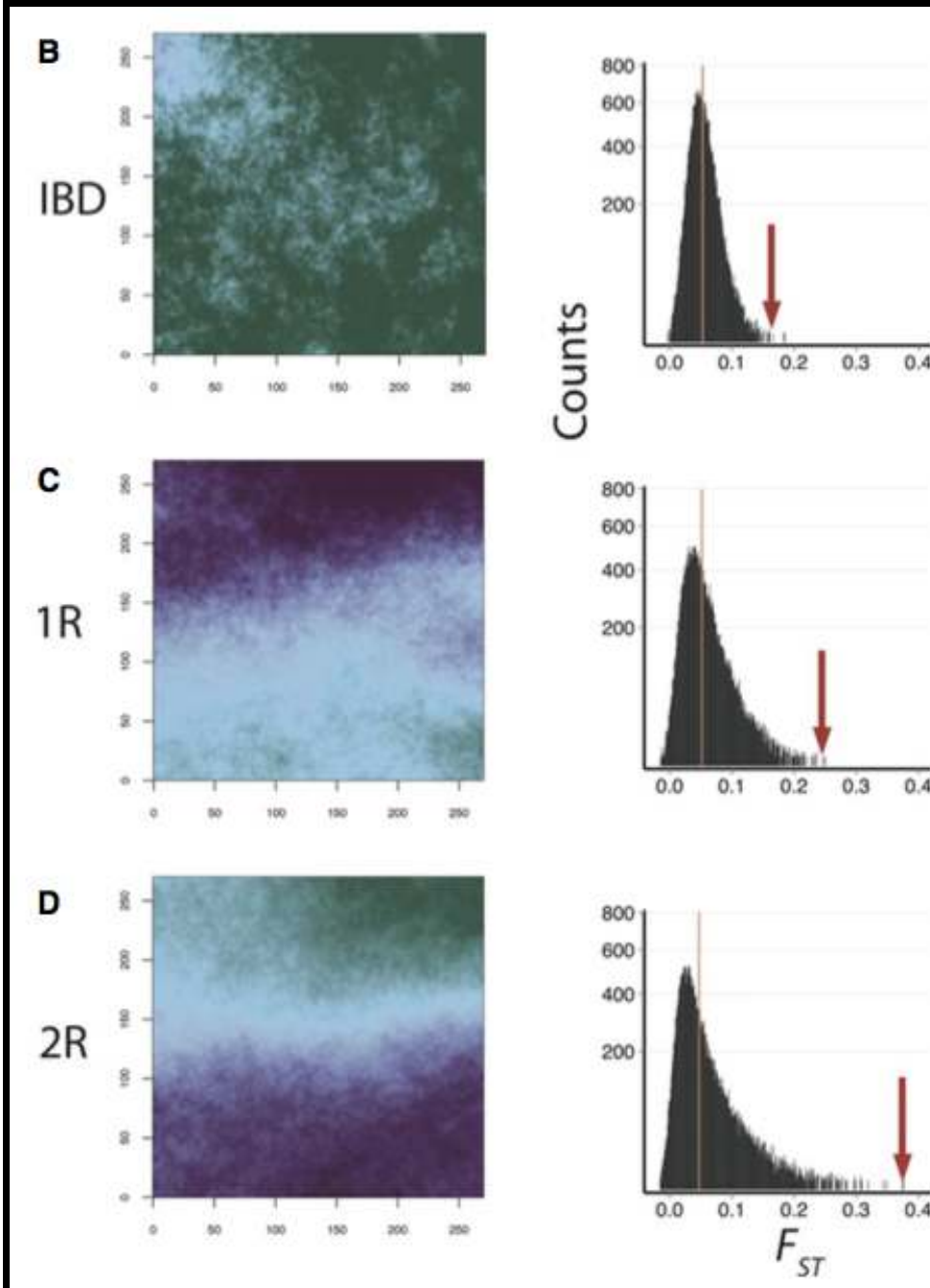
Directional selection: an example of the expectations of hard selection

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

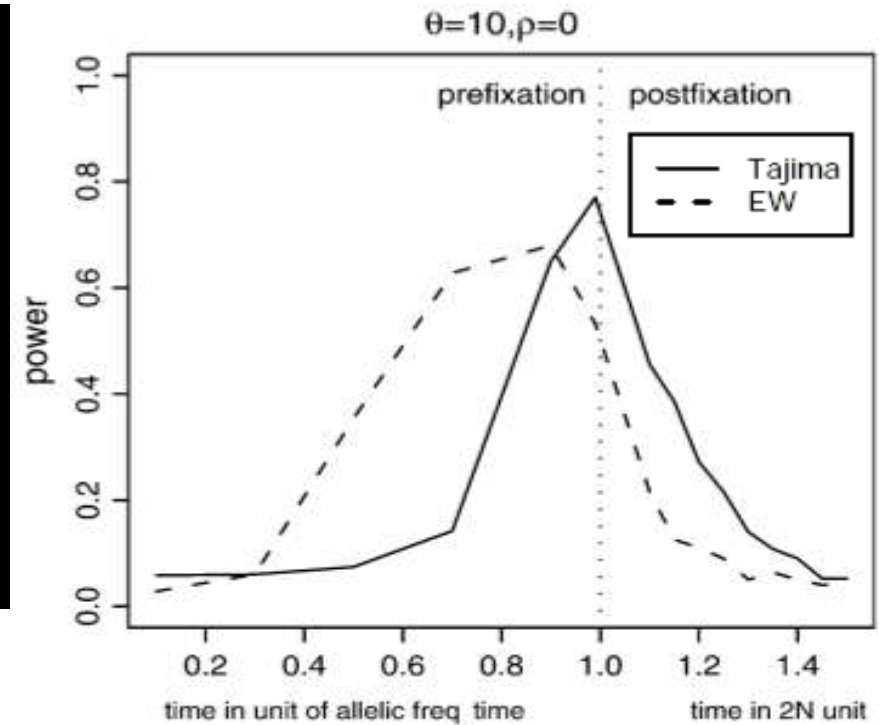
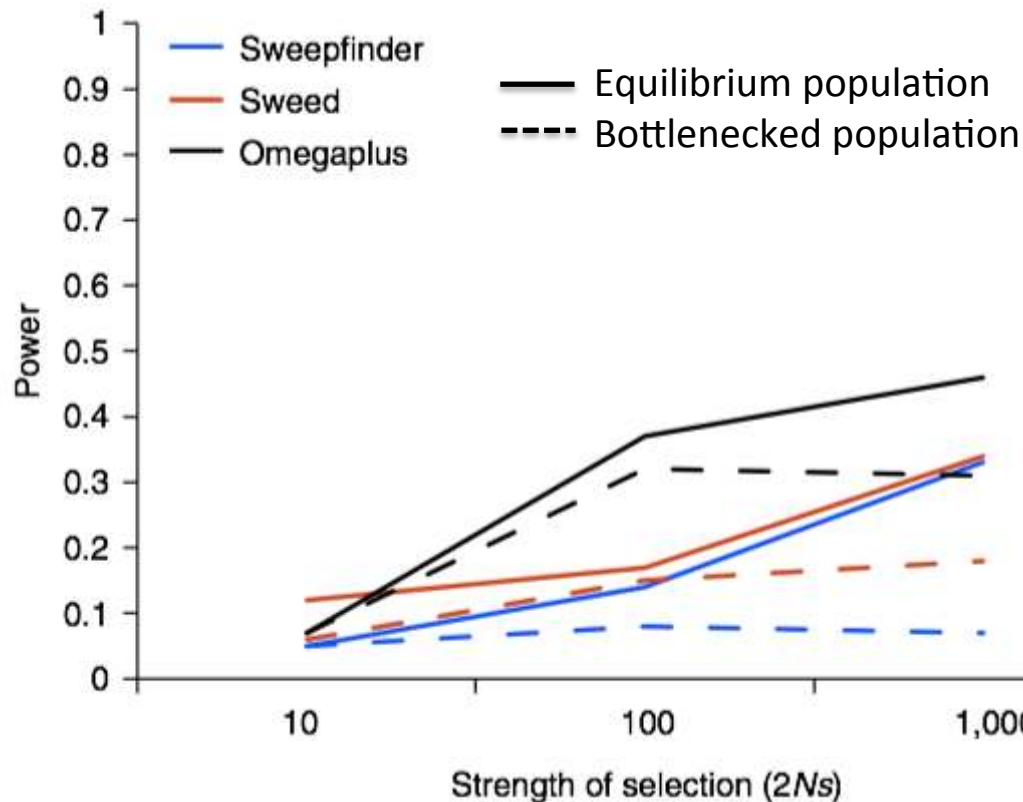
Fst outlier analysis



9900 Neutral, 100 selected sites
N=1500 (20 ind. per 75 populations)

■ Bayescan ■ FDIST2 ■ FLK

What is our power to detect hard sweeps within a population?



Zhai, Nielsen & Slatkin 2008 MBE

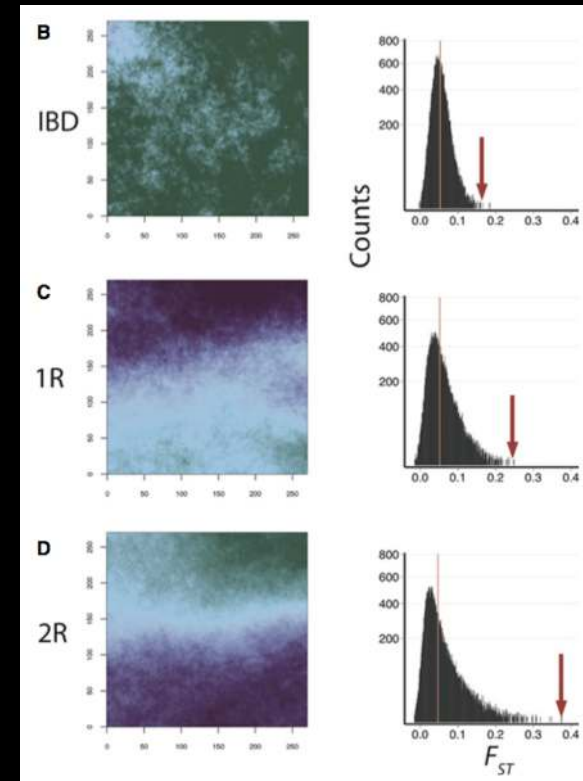
When did selection act on your phenotype?

What's the demographic history of your population?

Jensen 2014. Nature Communications 5:1–10.

What's a good way to assess molecular tests?

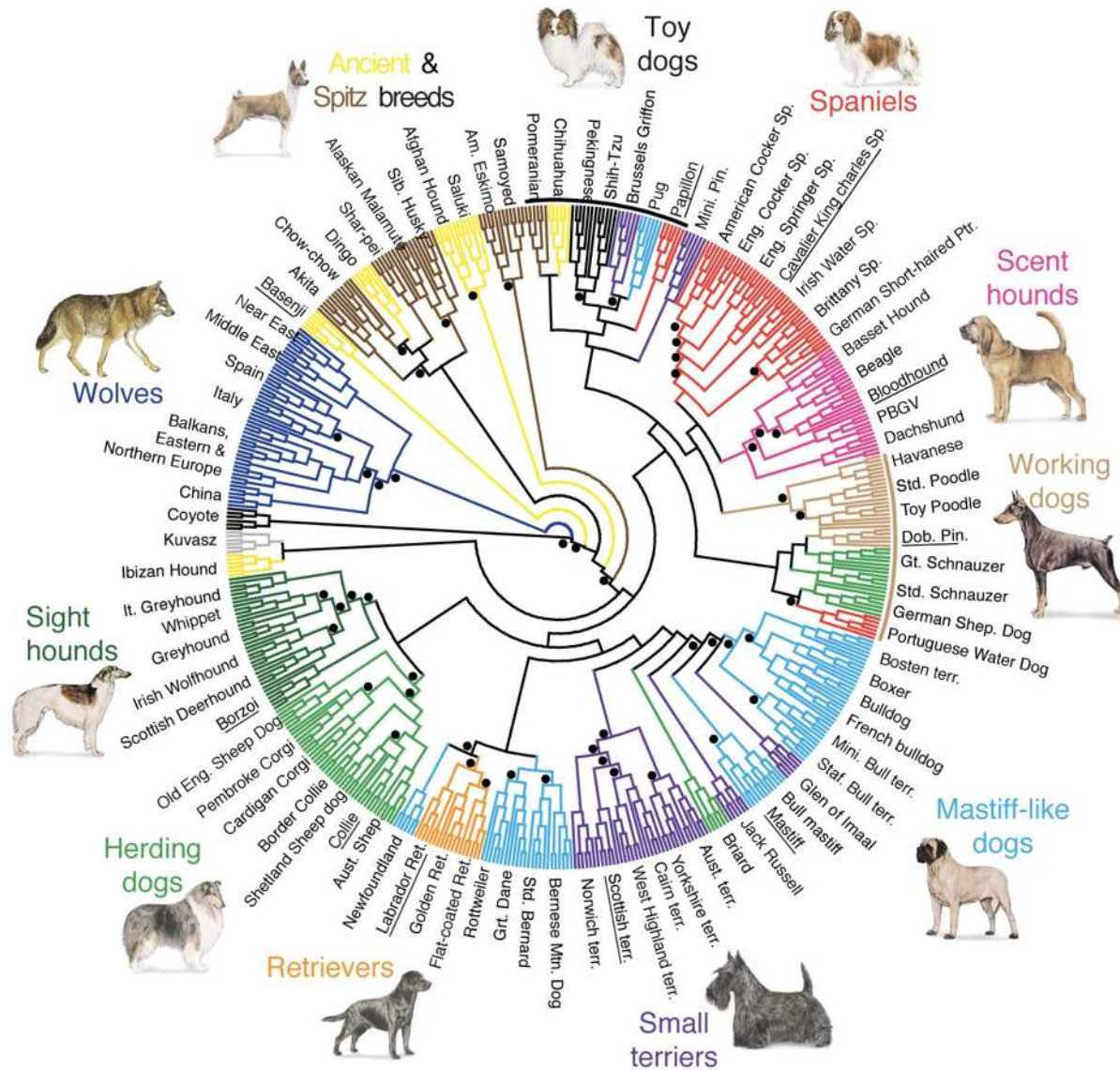
- Computer simulations of evolution
 - Across range of demographic scenarios
- What else?
- Testing them on real data where we know the targets of selection = real world validation
 - Which ones work and when
 - We could then use this to make better tests, right? (very rare)



Breed specific morphologies

Test set of Schlamp et al. 2016:

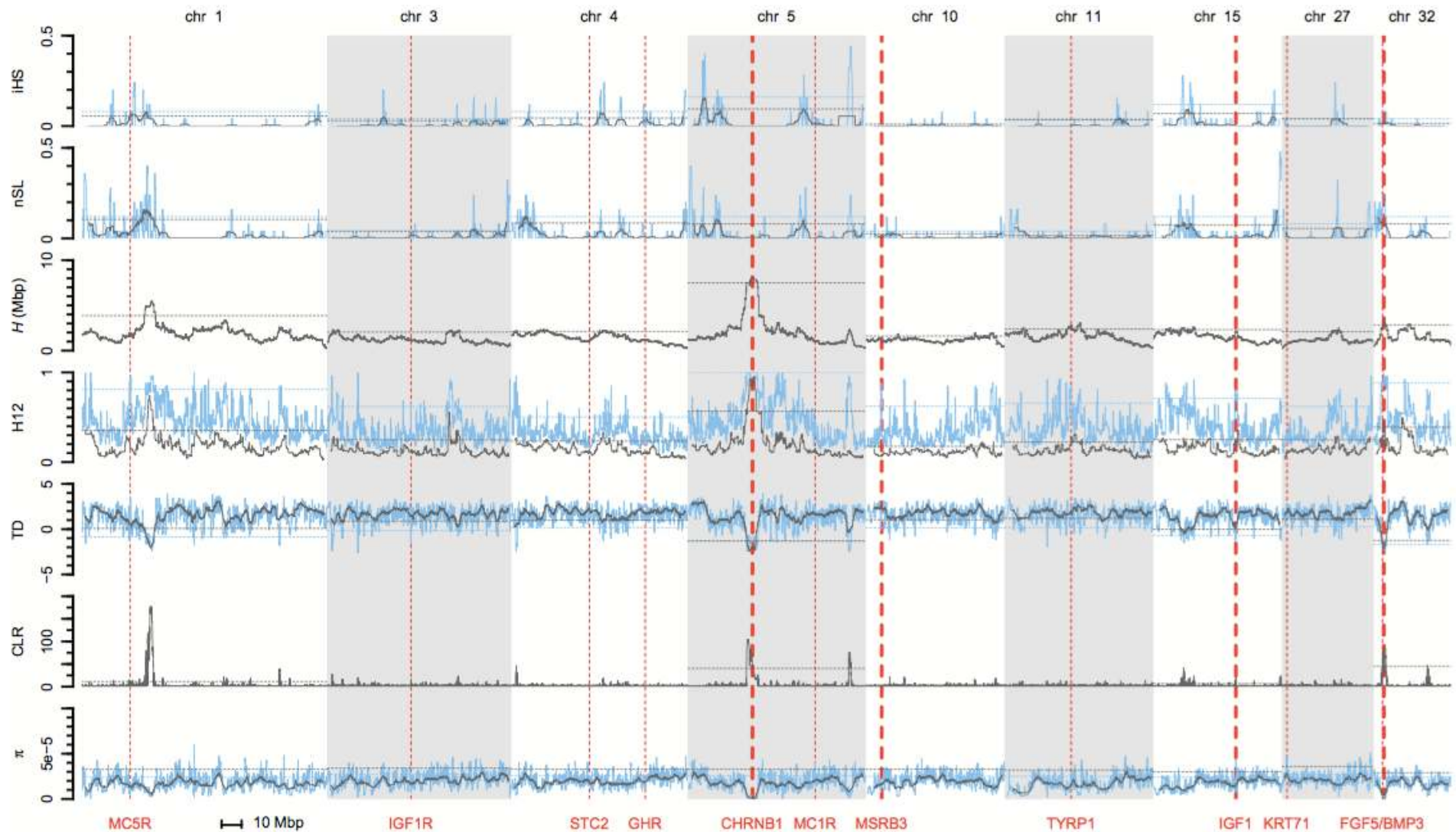
- 25 breeds
- 12 causal loci
- $N = 25$ / breed
- 7 tests of selection
 - $iHS, nSL, H, TajD$, etc.



von Holdt et al. 2010. Nature

What can state of the art molecular tests of selection detect?

French Bulldog sample: low power, high type I & II error



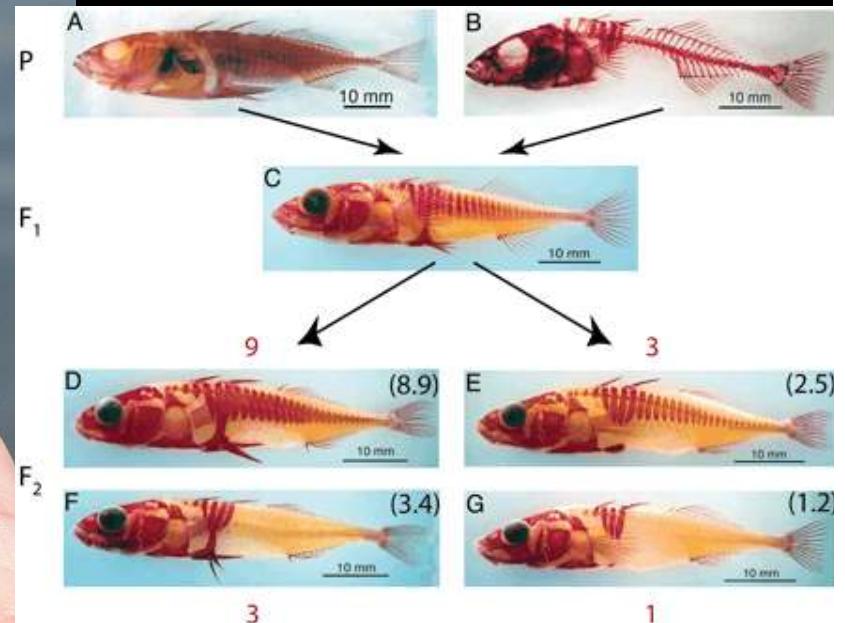
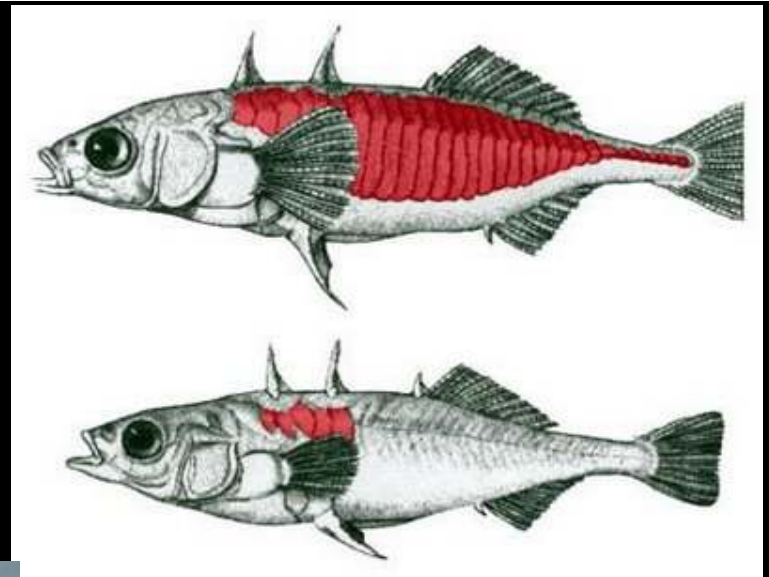
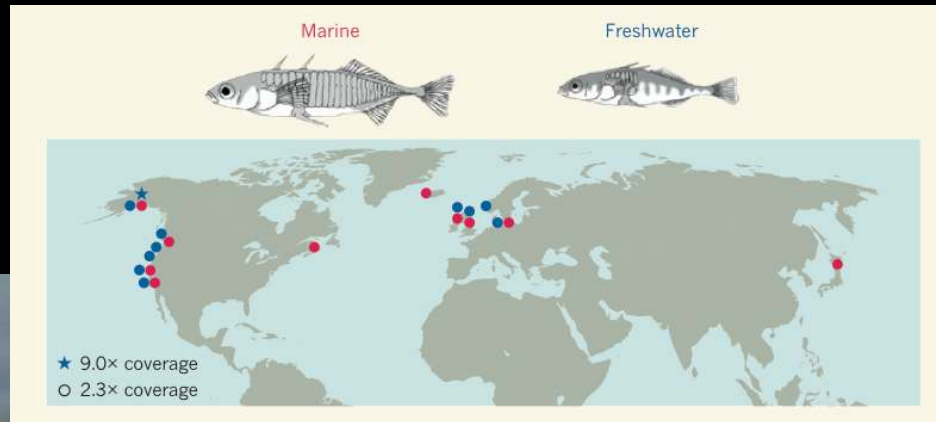
Schlamp et al. 2016. Evaluating the performance of selection scans to detect selective sweeps in domestic dogs. *Molecular Ecology* 25:342–356.

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
 - Have high false positive rates

Hard selection case example: threespine stickleback fish



Threespine stickleback fish

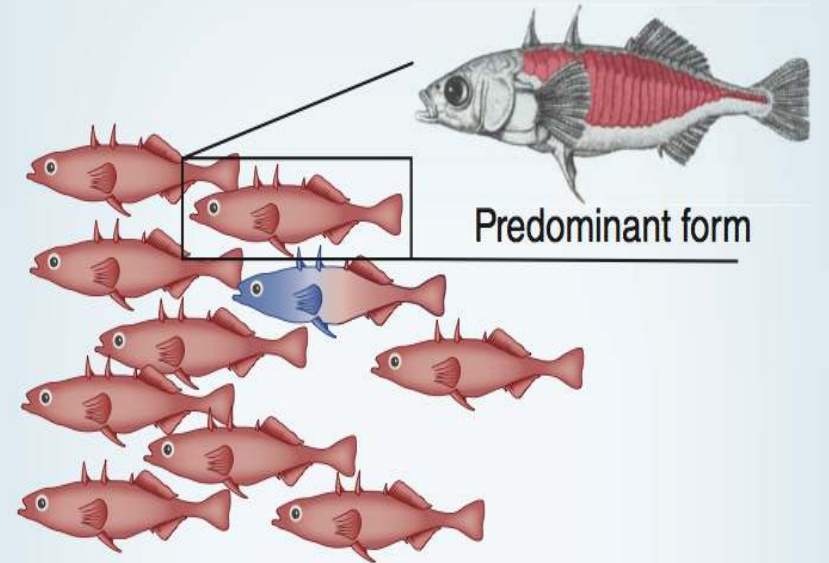
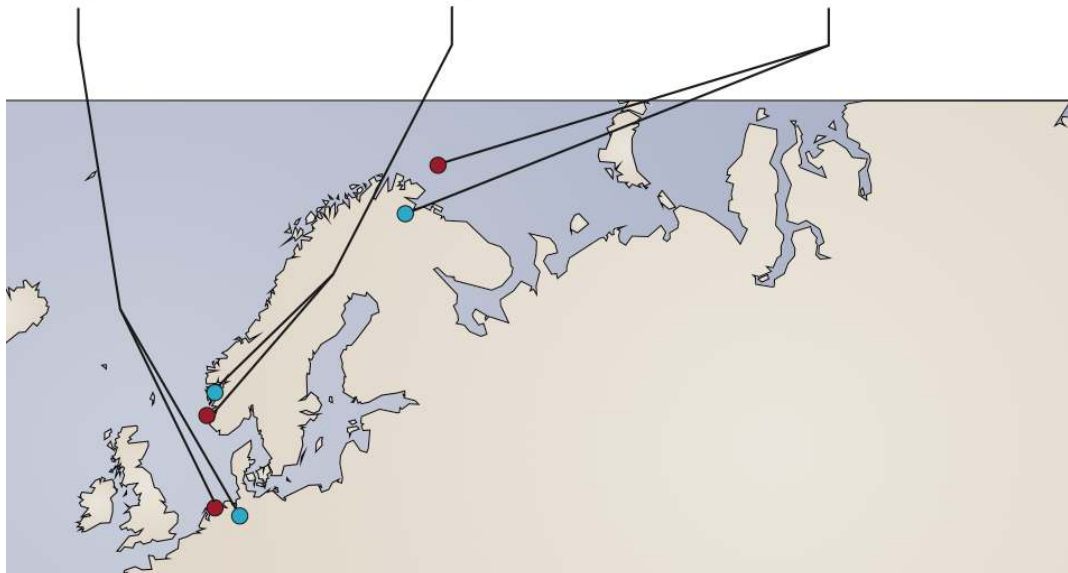
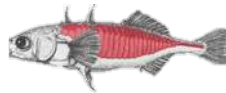
(*Gasterosteus aculeatus*)

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

Lakes

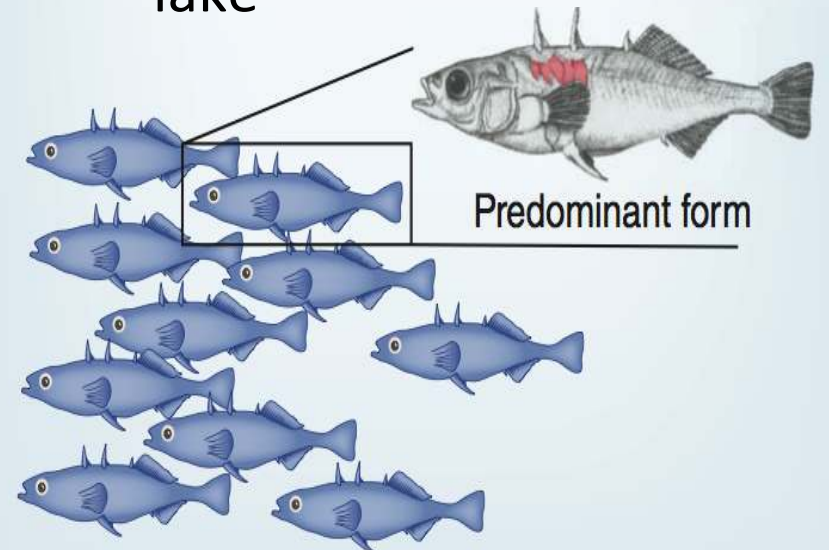


Ocean



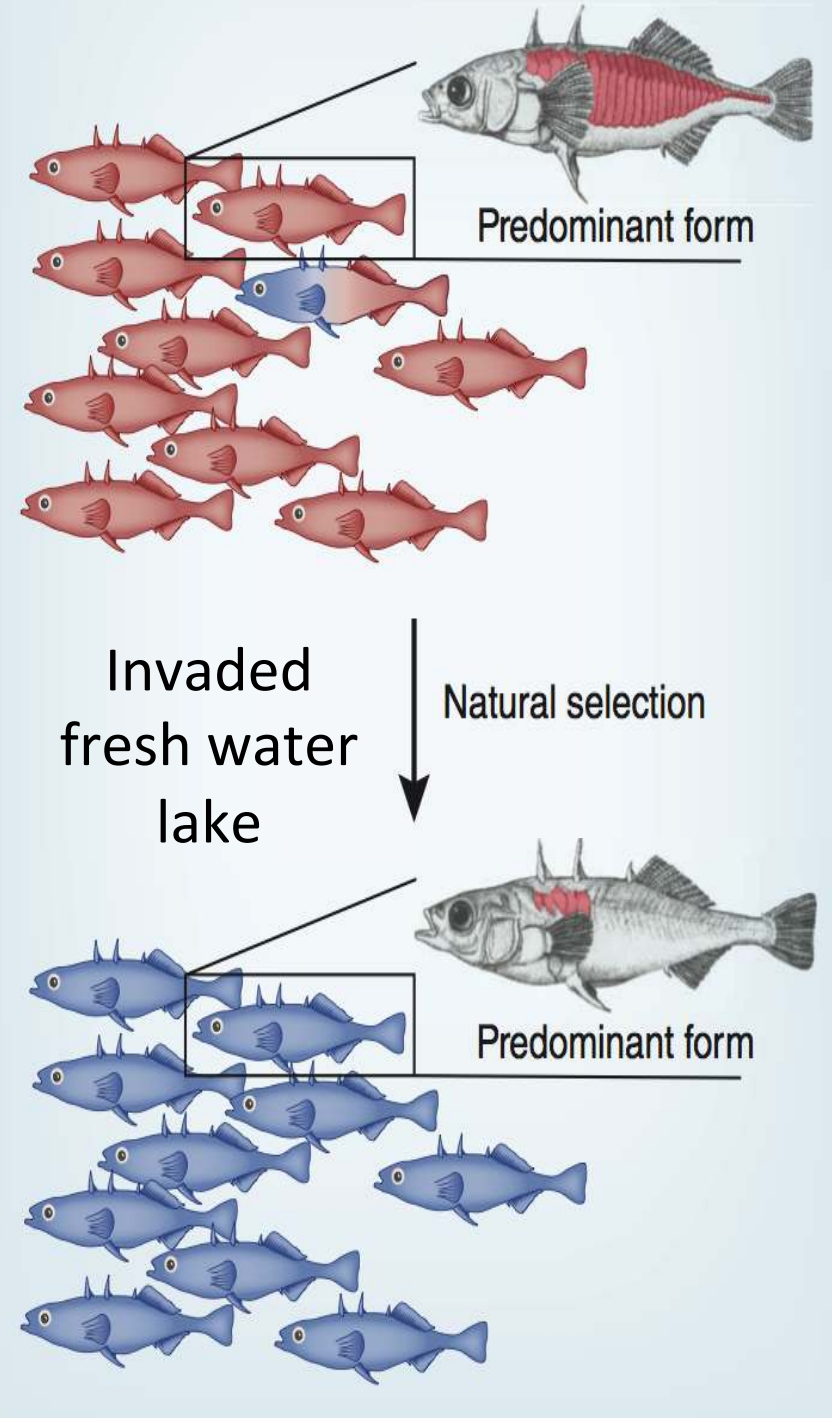
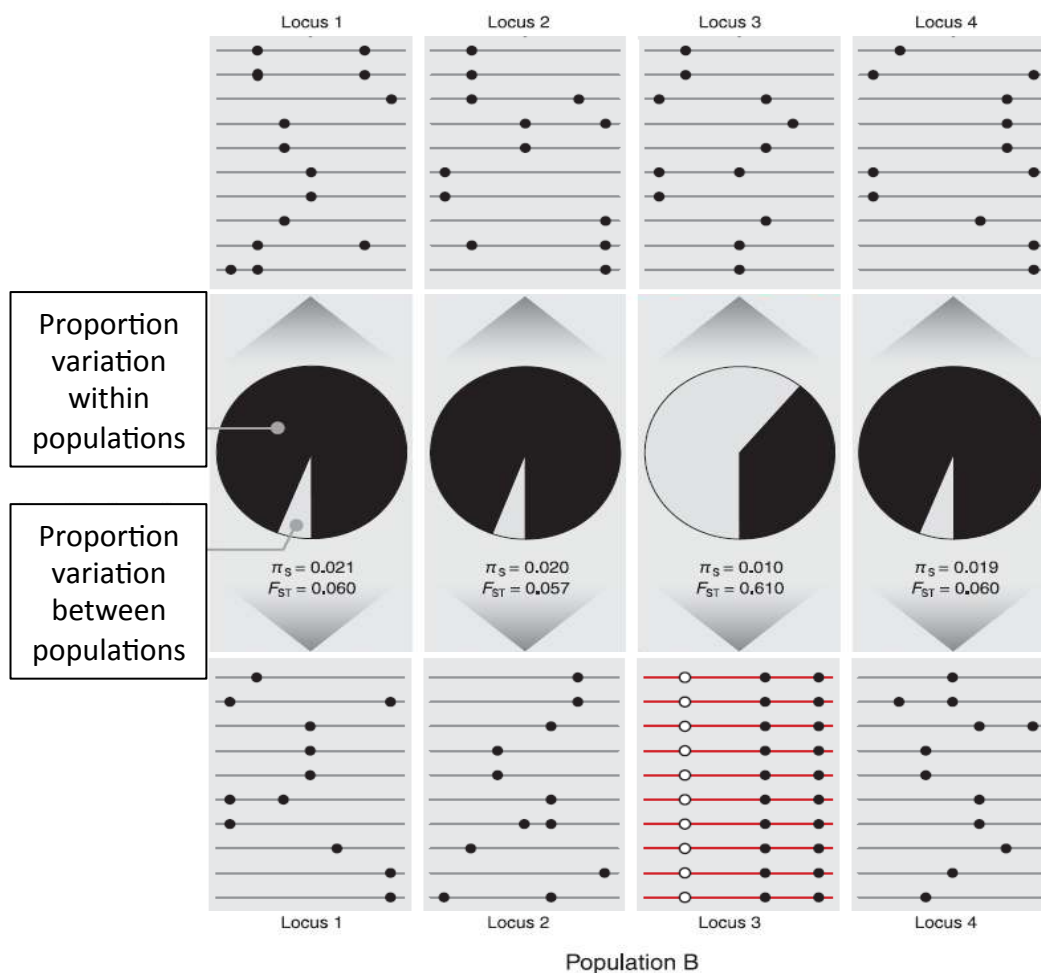
Invaded
fresh water
lake

Natural selection

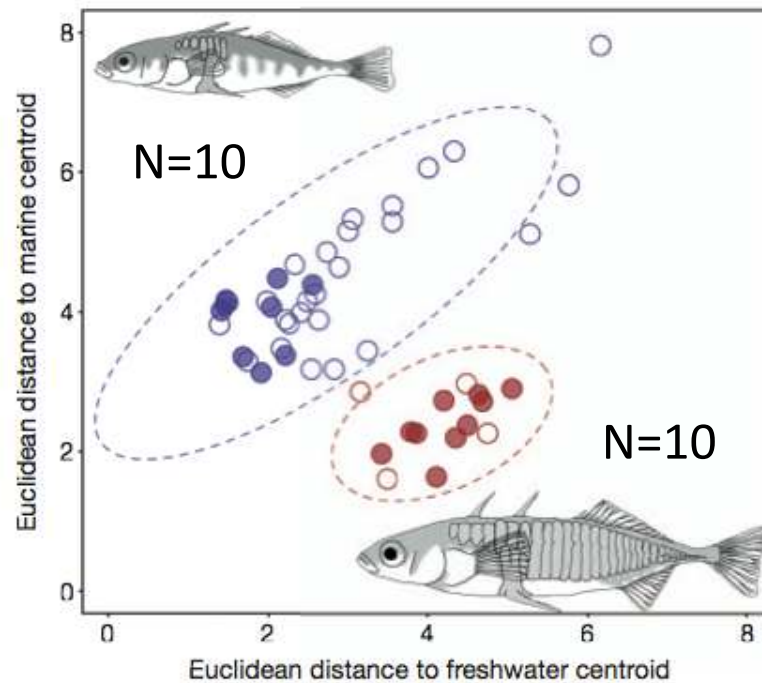
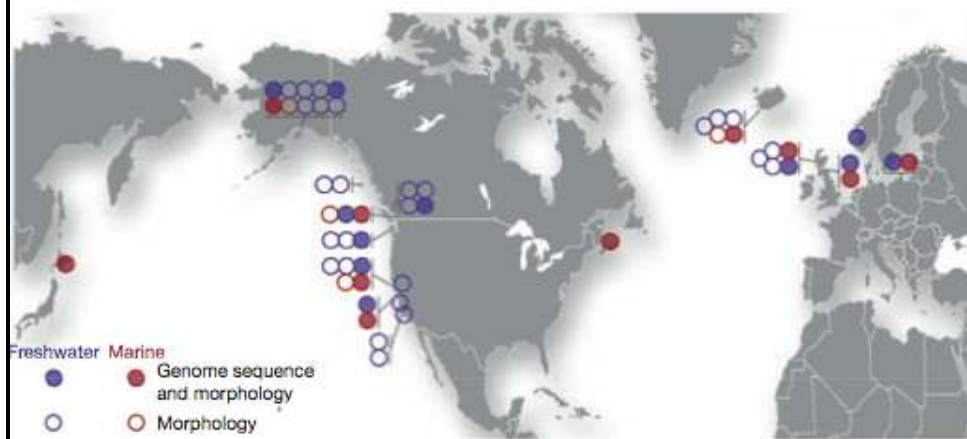


Parallel adaptation in fresh water lakes via hard sweeps

Marine population



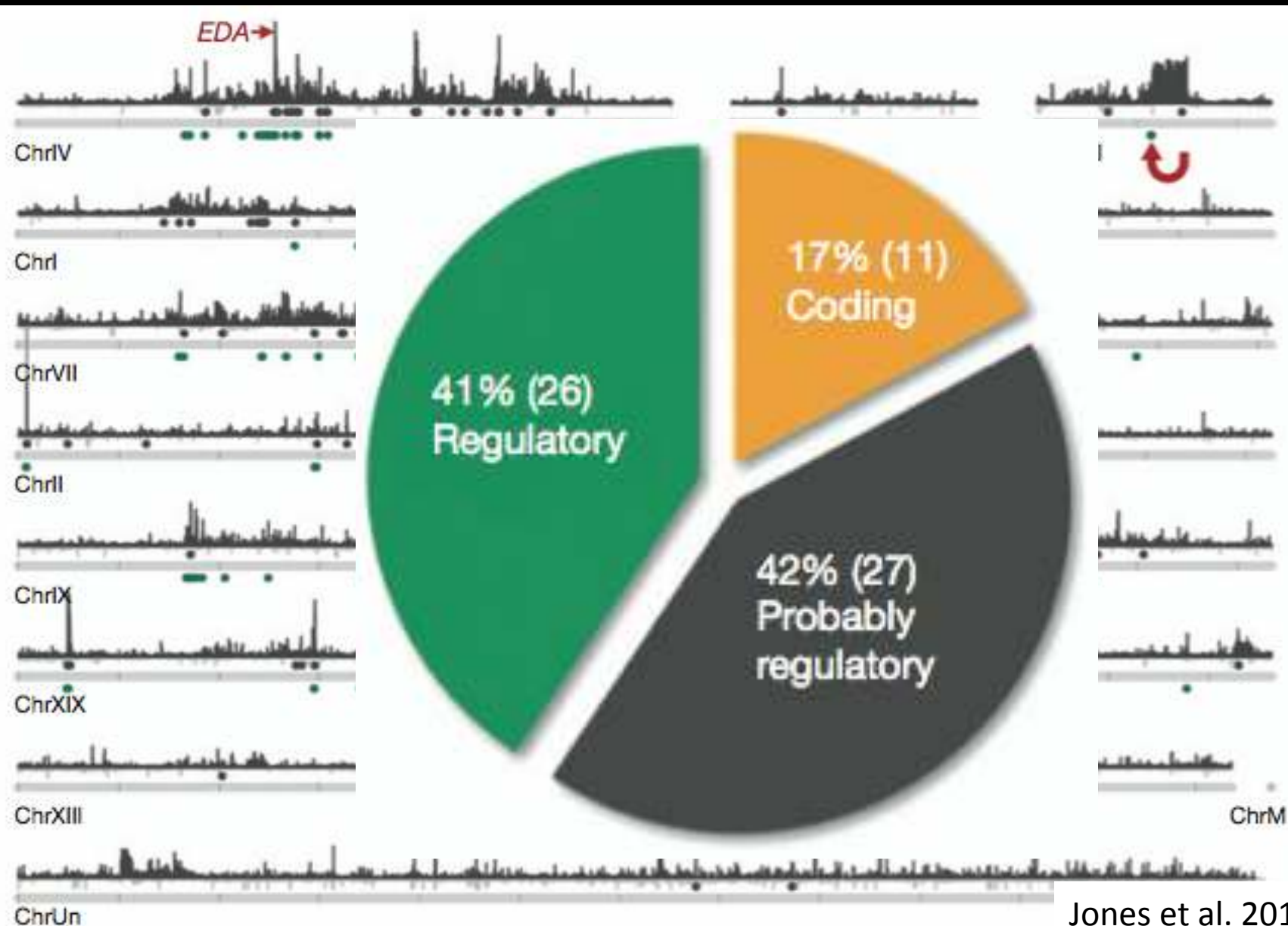
Individual genome sequencing: powerful insights



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

Jones et al. 2012 Nature

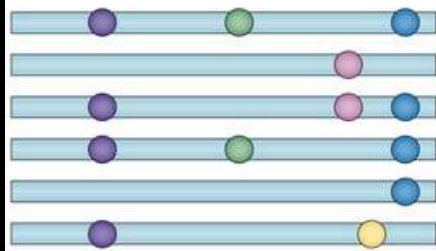
Which regions are more important? Coding or expression?



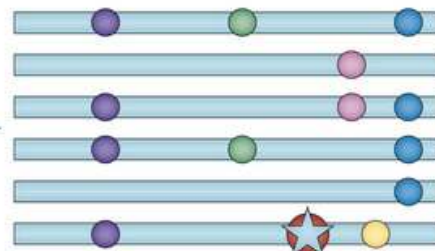
Jones et al. 2012 Nature

a Classic selective sweep

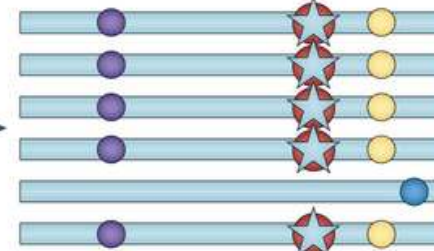
Neutral variation



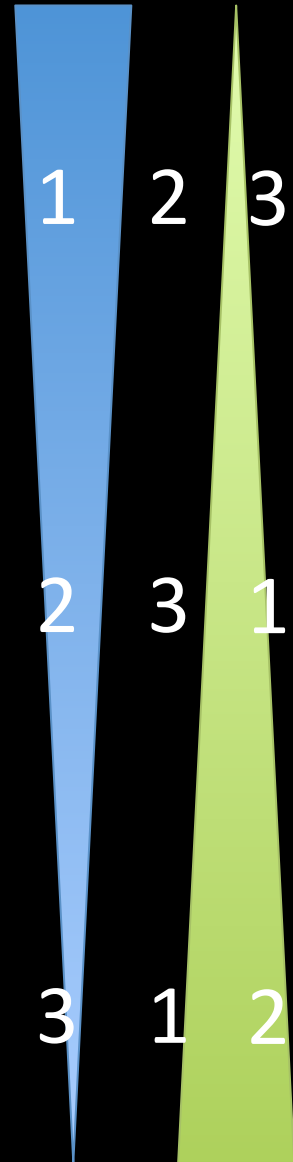
An advantageous mutation arises



Over time, the advantageous mutation approaches fixation



Test
power



Freq. in
nature

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 ‘incomplete’ [or] ‘soft’ sweep models.”

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



- This is an important read, critical of
 - assumptions underlying soft sweep
 - low power of molecular tests to detect hard & soft sweeps

How common are soft sweeps in your species?

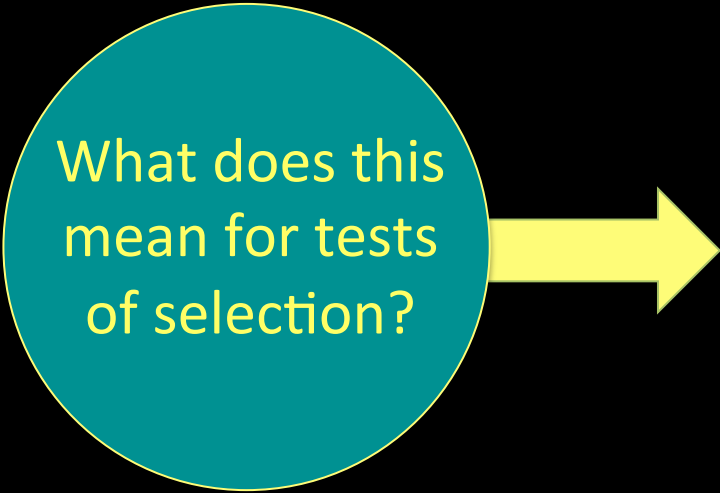
Thought experiment:

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?



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 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 lightning rod for 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?

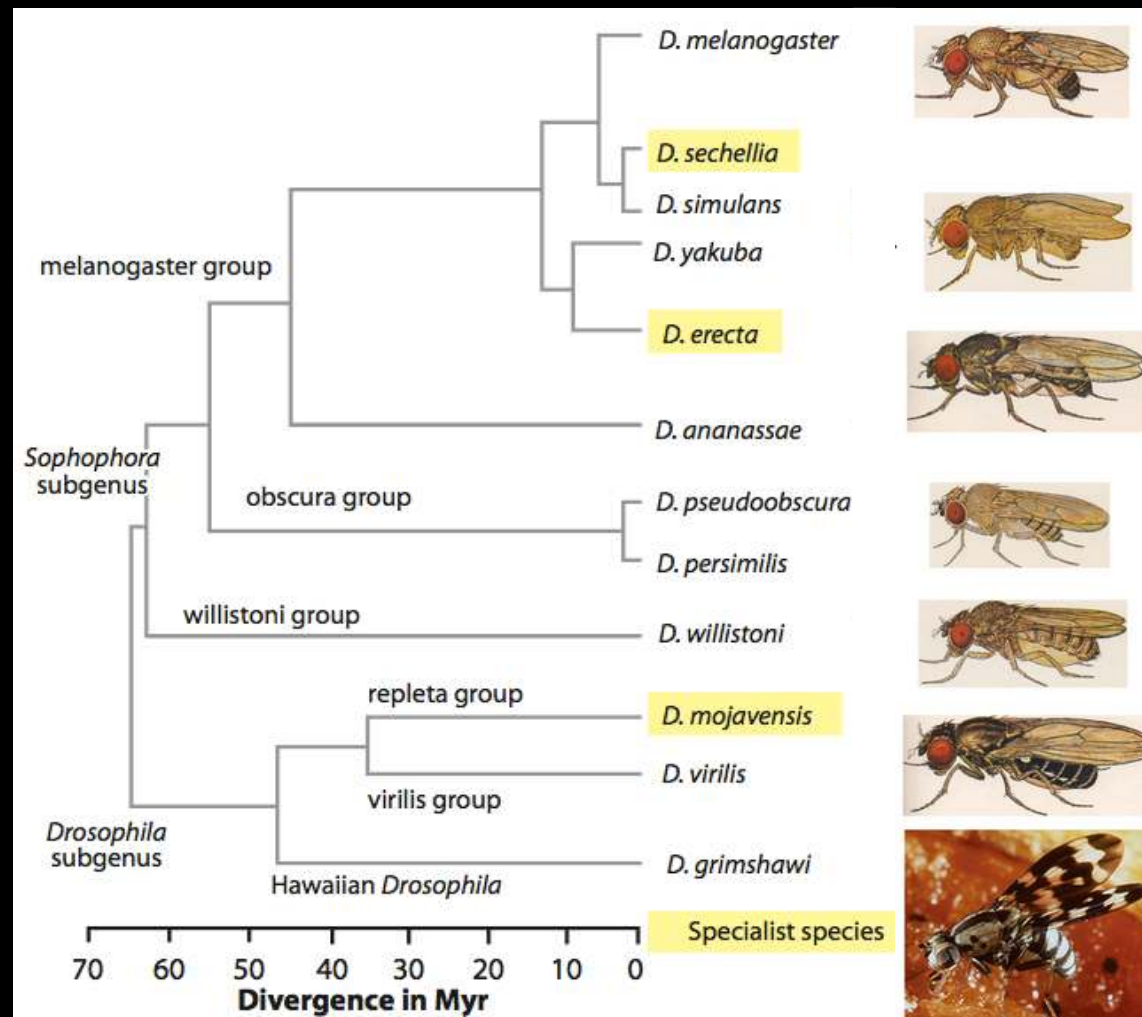


studying:

relationships

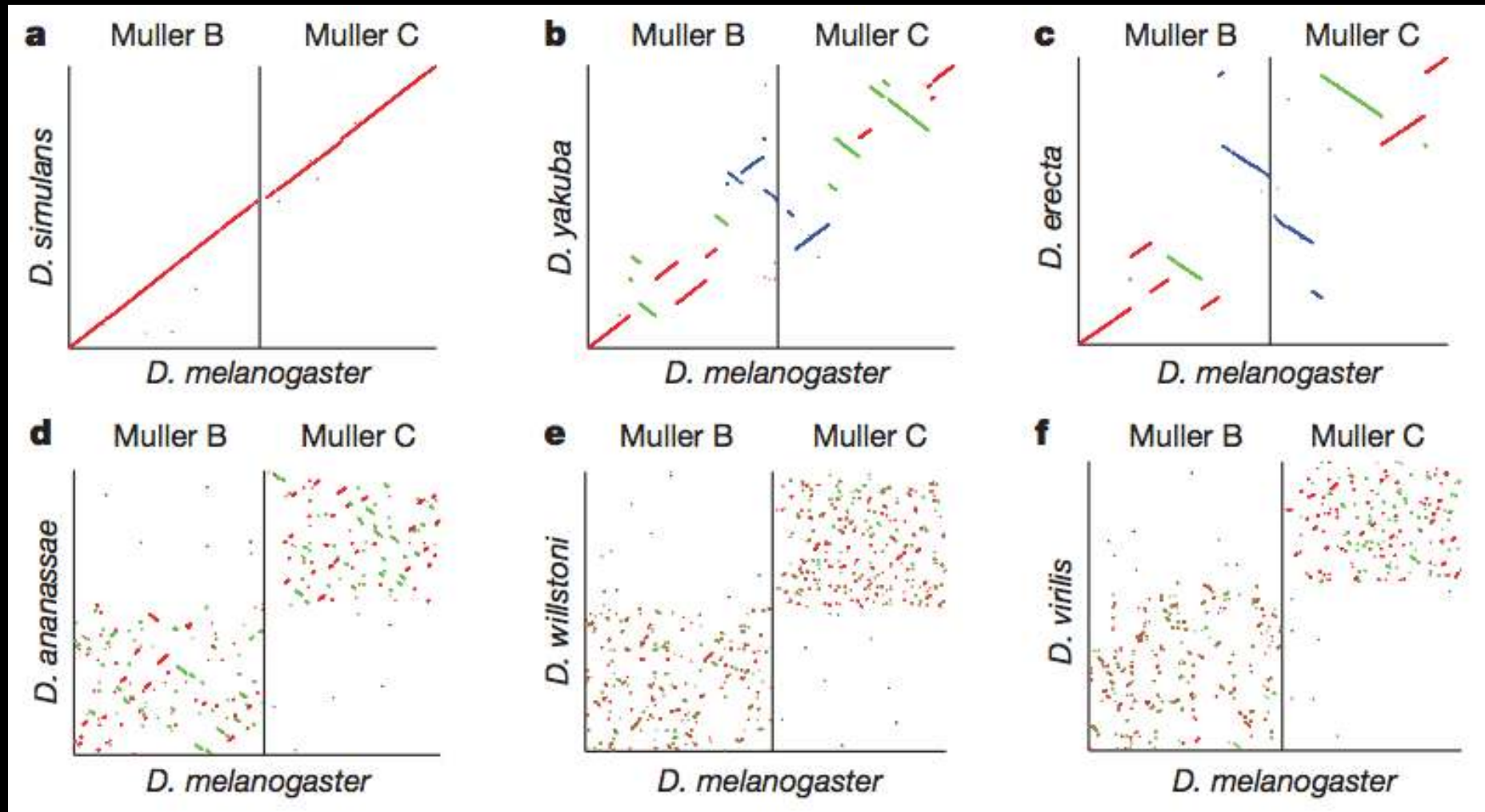
- Genome evolution
- Functional insights into genes and genomic features (e.g. regulation and inheritance)

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



Drosophila 12 Genomes Consortium 2007 Nature

Tempo and mode of chromosome evolution



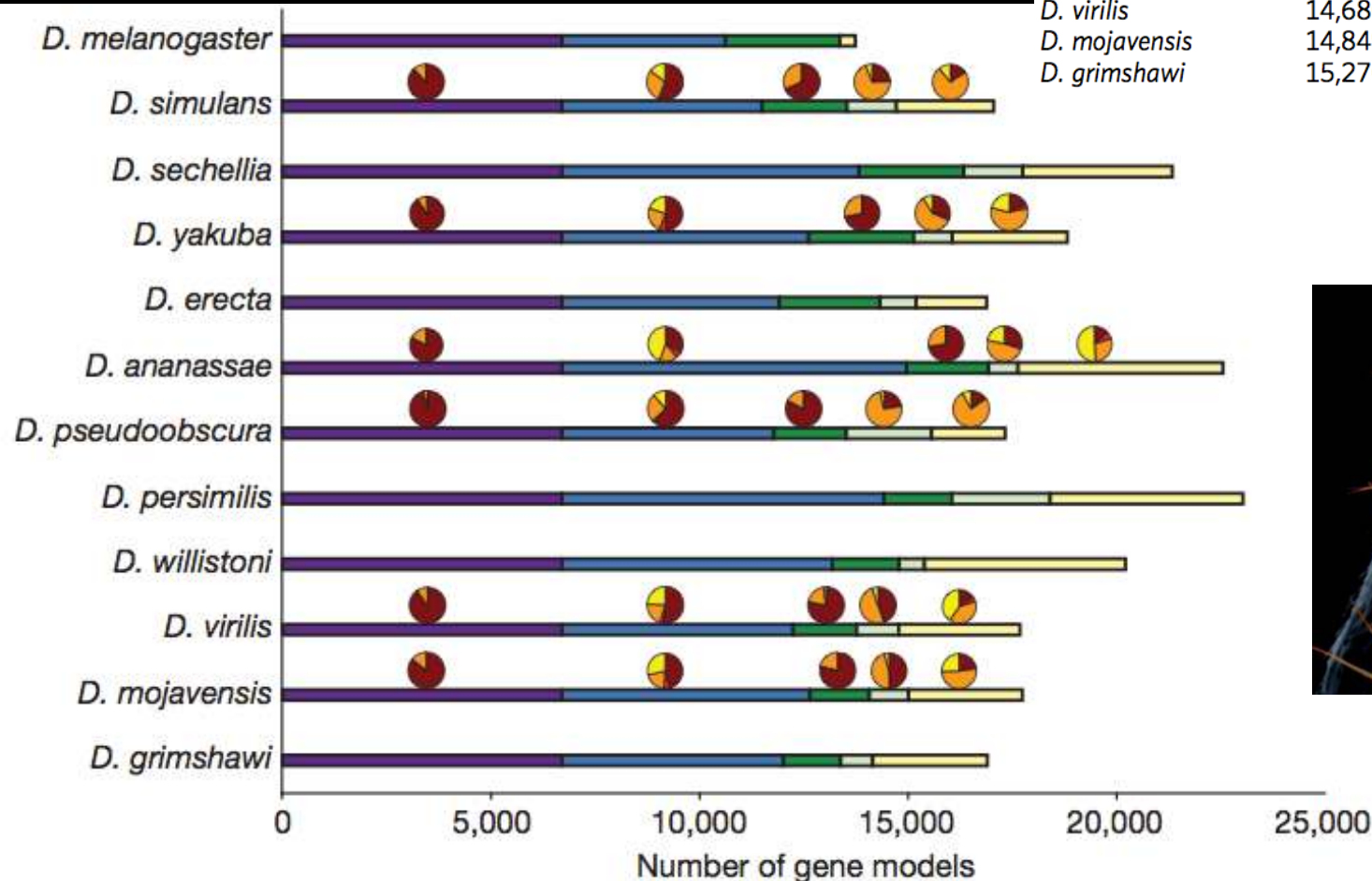
- > 20 My, chromosomal order completely reshuffled in Diptera

Drosophila 12 Genomes Consortium 2007 Nature

Genome evolution

Drosophila 12 Genomes Consortium 2007 Nature

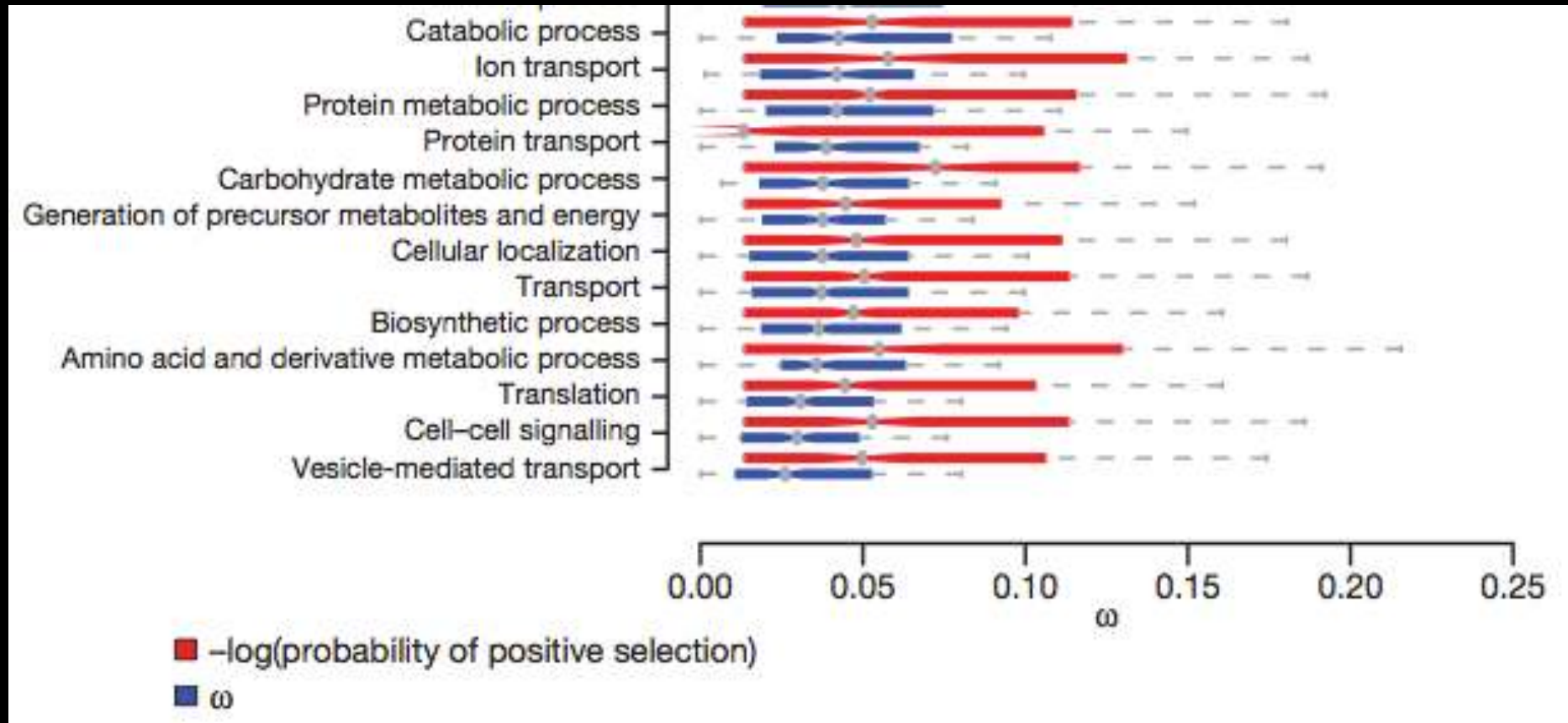
	Total no. of protein- coding genes (per cent with <i>D. melanogaster</i> homologue)	Coding sequence/ intron (Mb)
<i>D. melanogaster</i>	13,733 (100%)	38.9/21.8
<i>D. simulans</i>	15,983 (80.0%)	45.8/19.6
<i>D. sechellia</i>	16,884 (81.2%)	47.9/21.9
<i>D. yakuba</i>	16,423 (82.5%)	50.8/22.9
<i>D. erecta</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



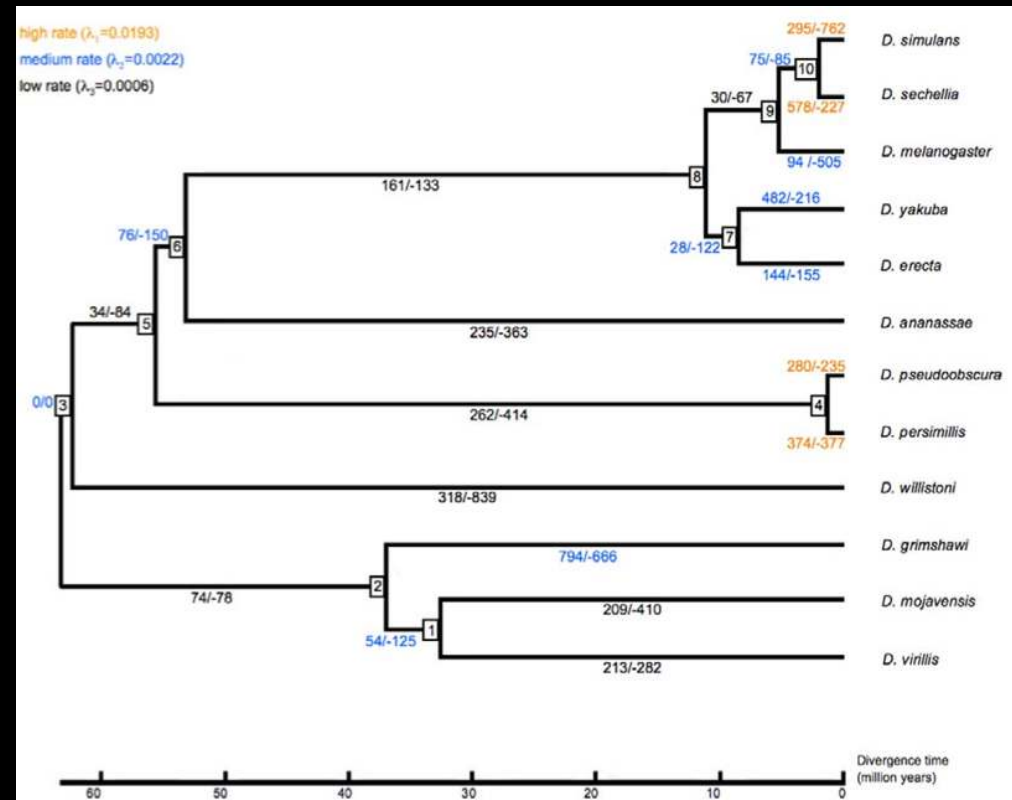
Selection dynamics across functional categories



- 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 ...
 - 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

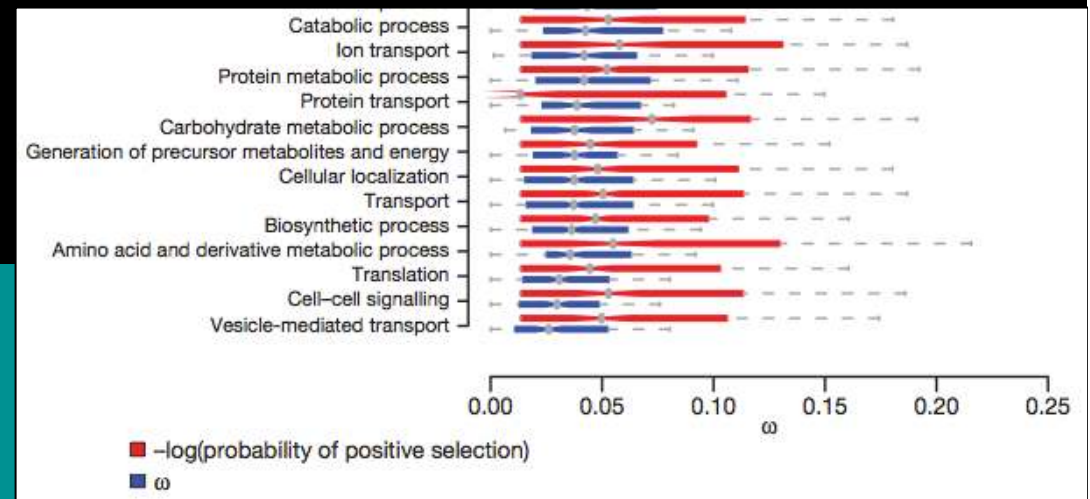
- Gene alignments in detecting positive selection
- Calibrations in temporal analysis

Established studies allow ...

Follow up studies to reveal limitations

Robust findings to emerge with age

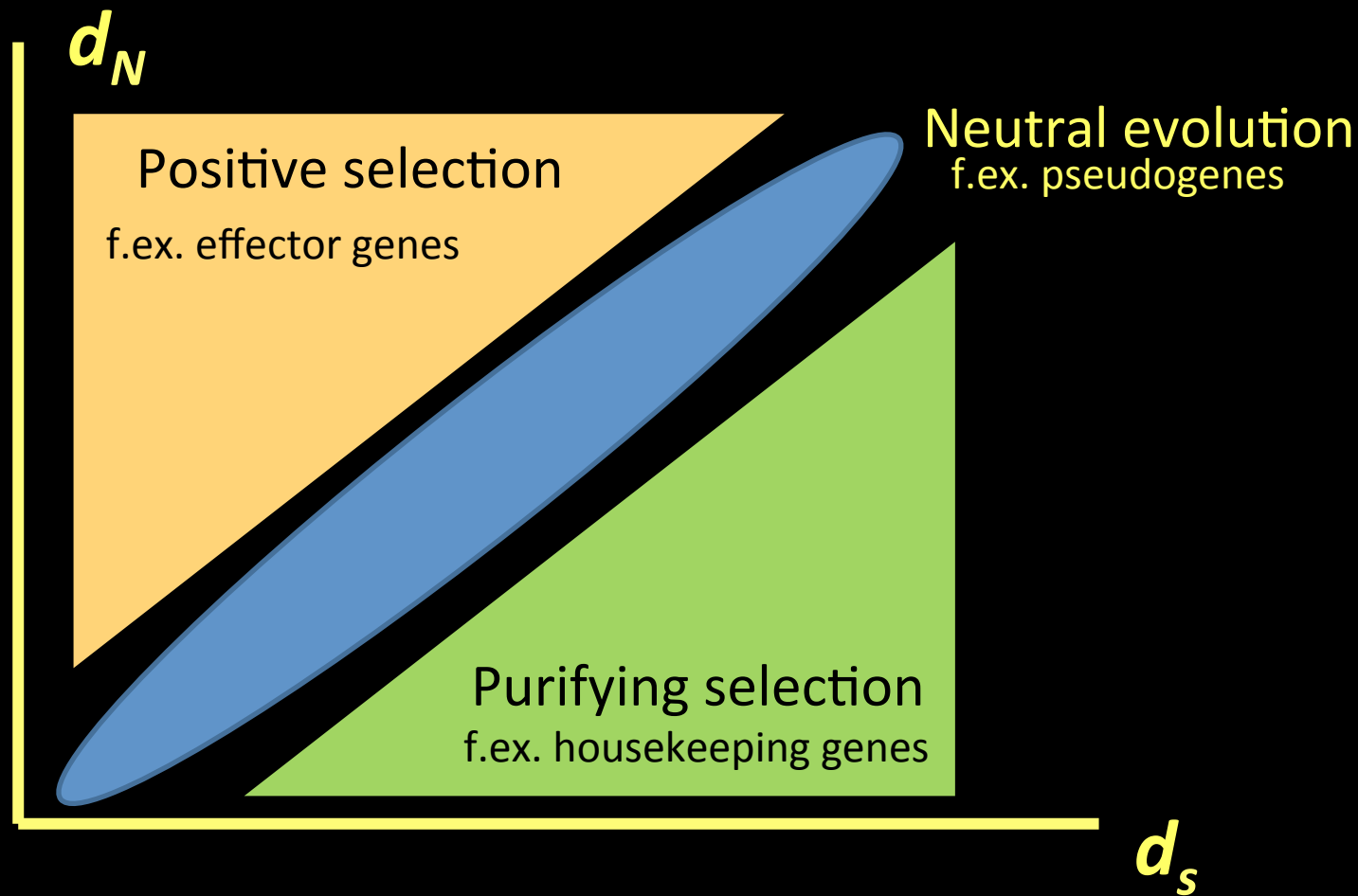
Inferring selection dynamics:



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

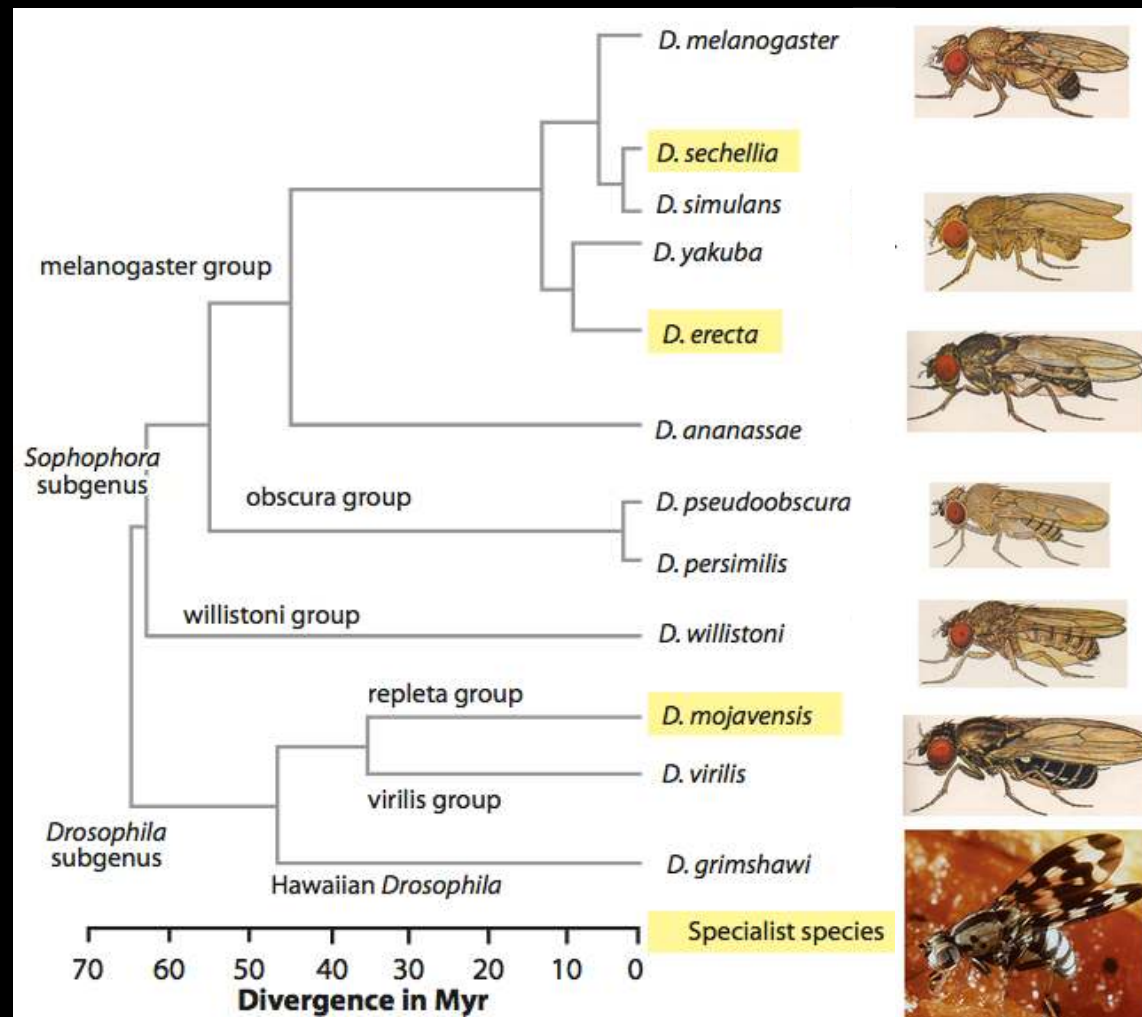


d_N / d_S
ratio

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

IMPRS workshop,
Comparative Genomics

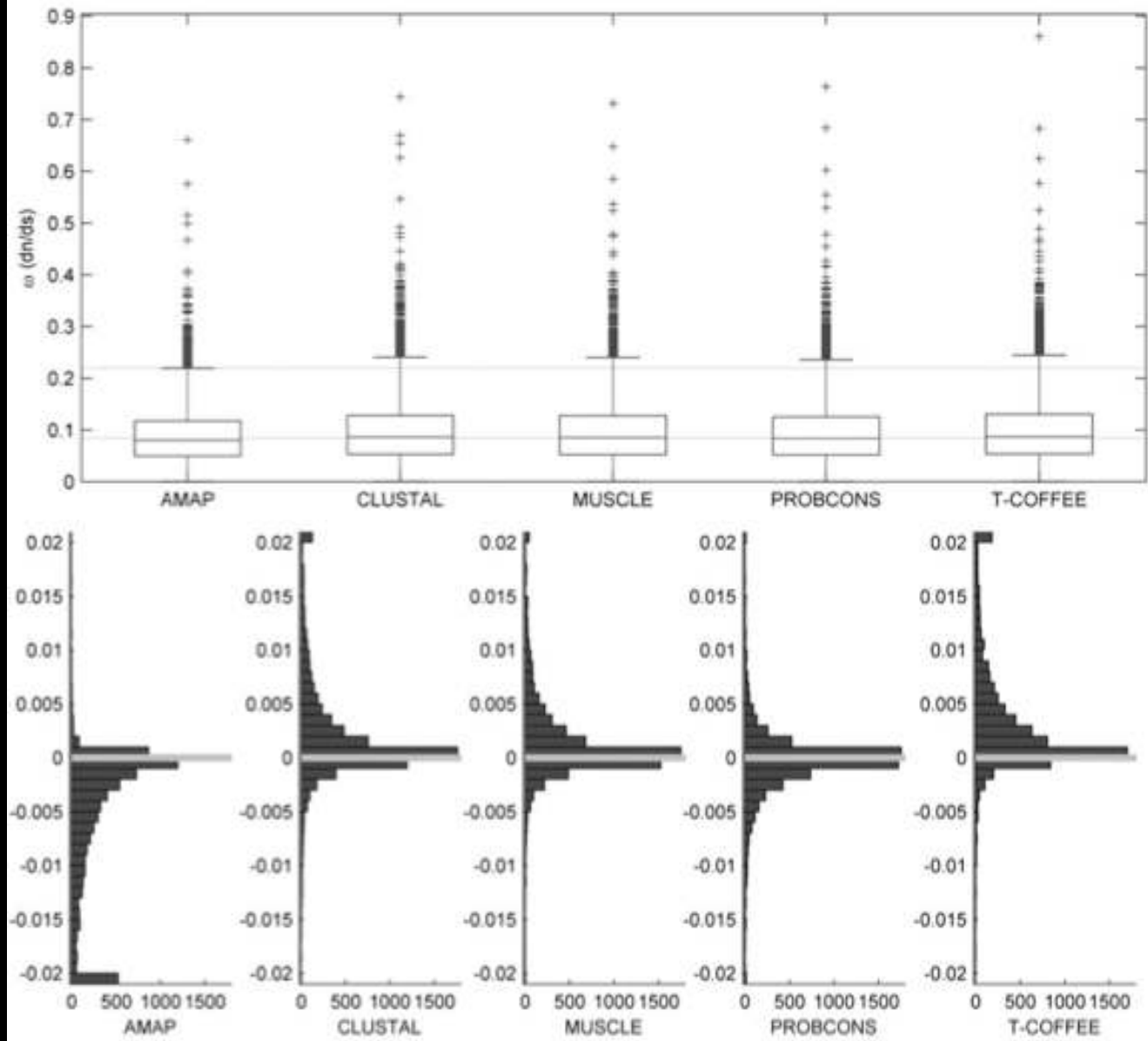
Evolution of genes and genomes on the *Drosophila* phylogeny



Drosophila 12 Genomes Consortium 2007 Nature

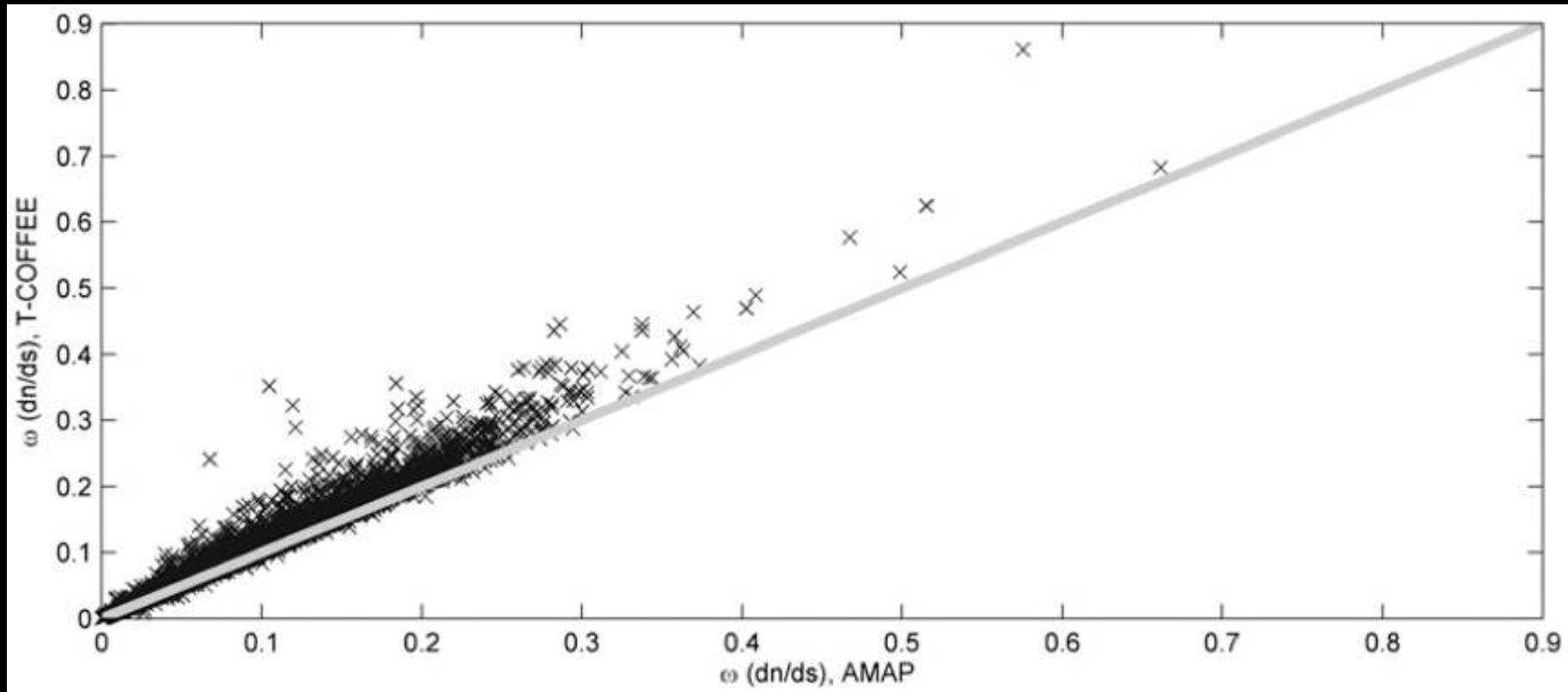
dN/dS estimates by aligner

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



Comparing results across methods is responsible bioinformatics!!!!

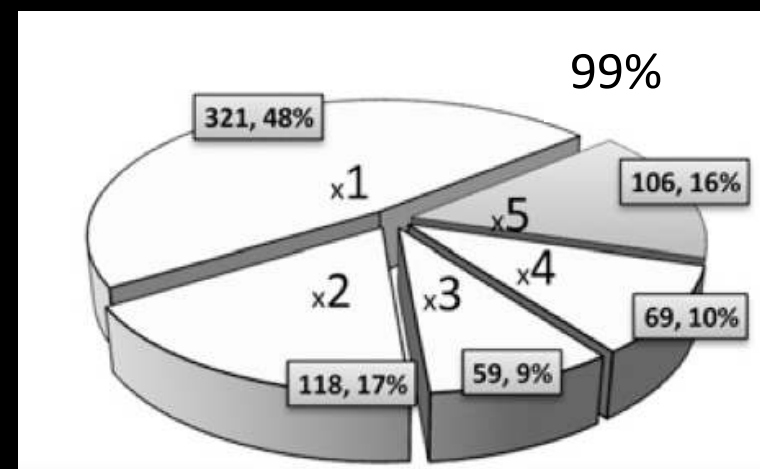
Since we can't look at our data, we need approaches that
allow 1st principal assessments



Aligner tool has a larger effect than biology

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

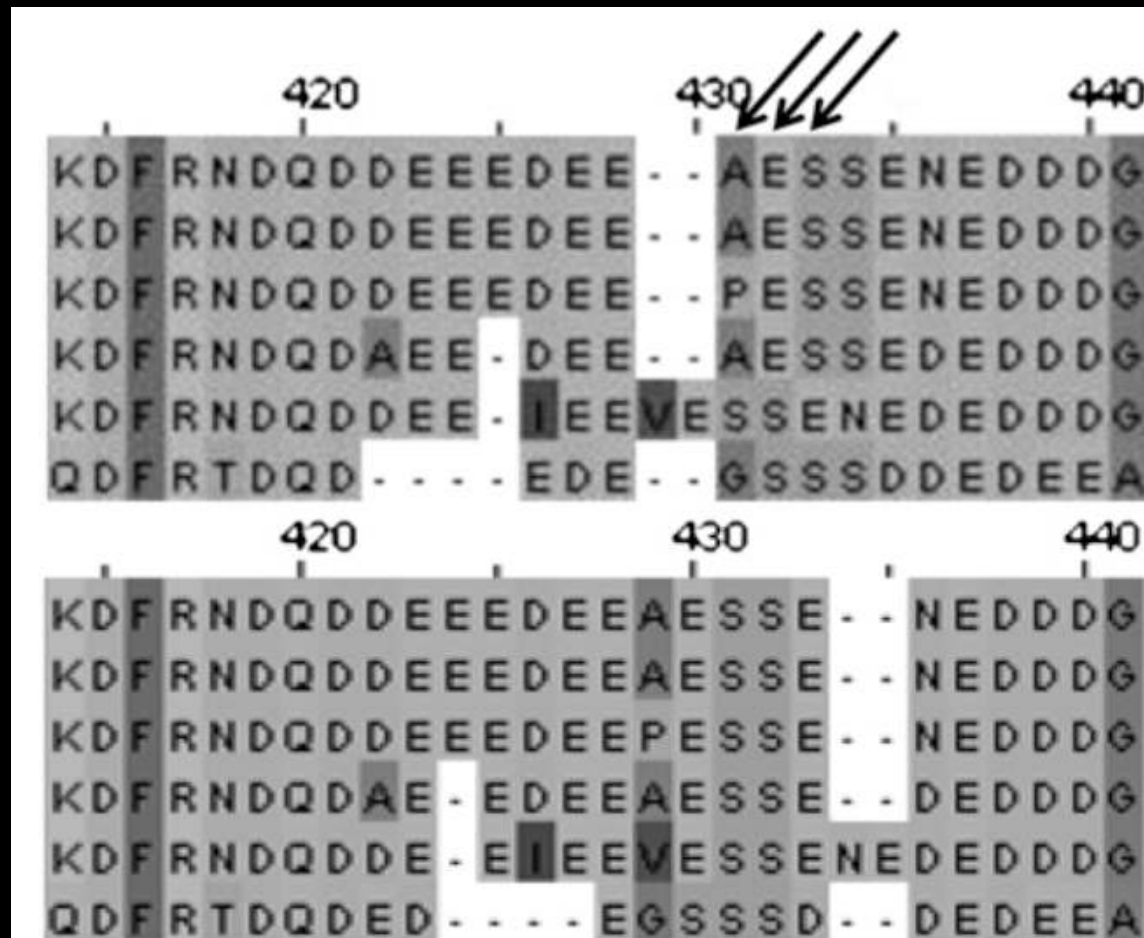
Number of significant genes in common across 1, 2, 3, 4, or all 5 of the alignment methods



Alignment results highlight importance of alignment score!

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

ProbCons Tcoffee



What about recent genomes?

Surely they are better?

and mammals ... they have good genomes

and alignment problems rarely happen

... right?

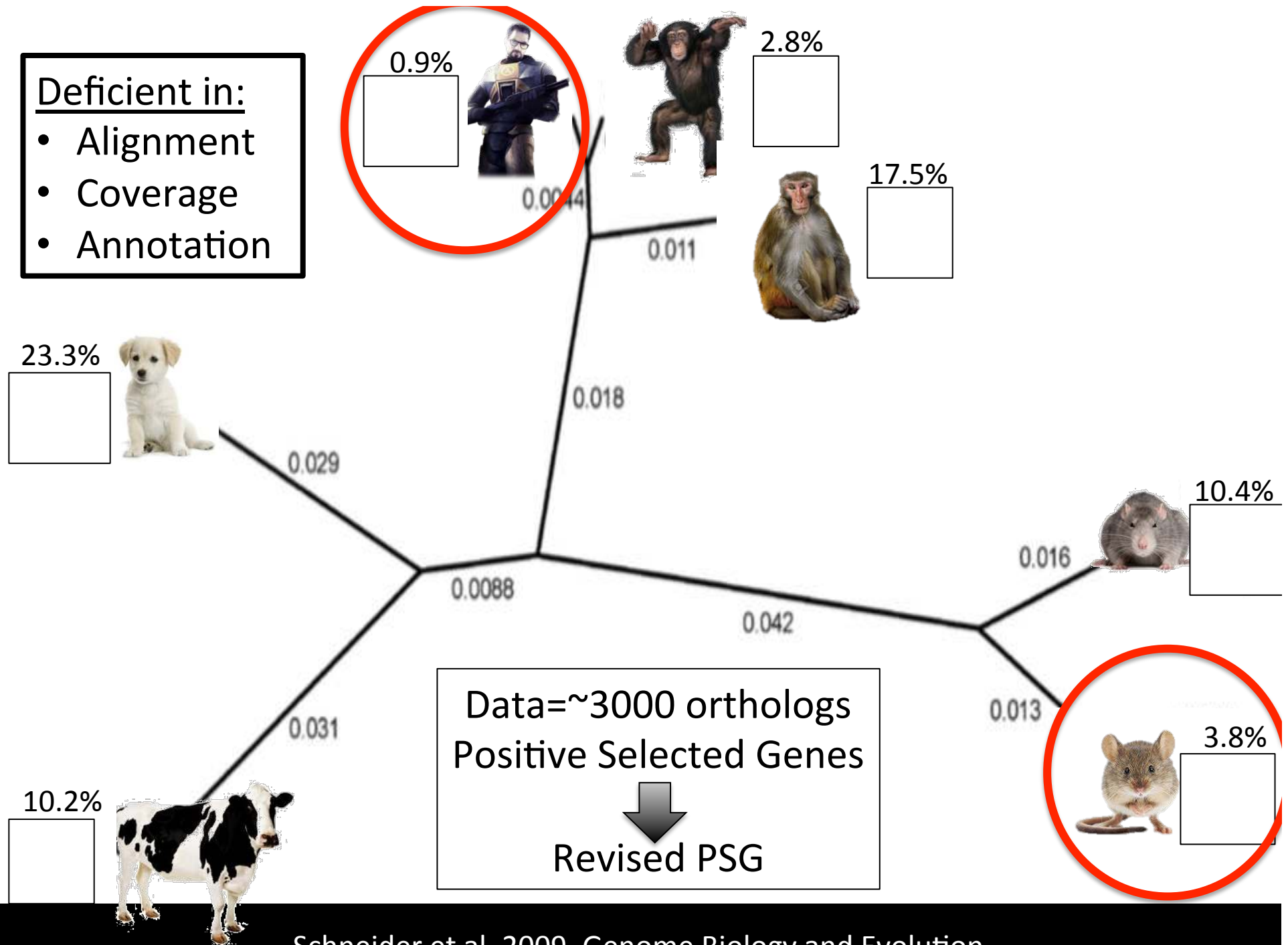


How did I
evolve to
be so cute?



Deficient in:

- Alignment
- Coverage
- Annotation



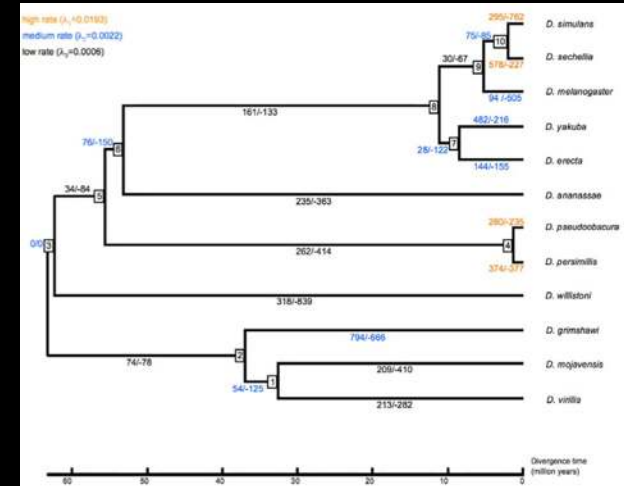
Temporal inference:

fact or fiction?

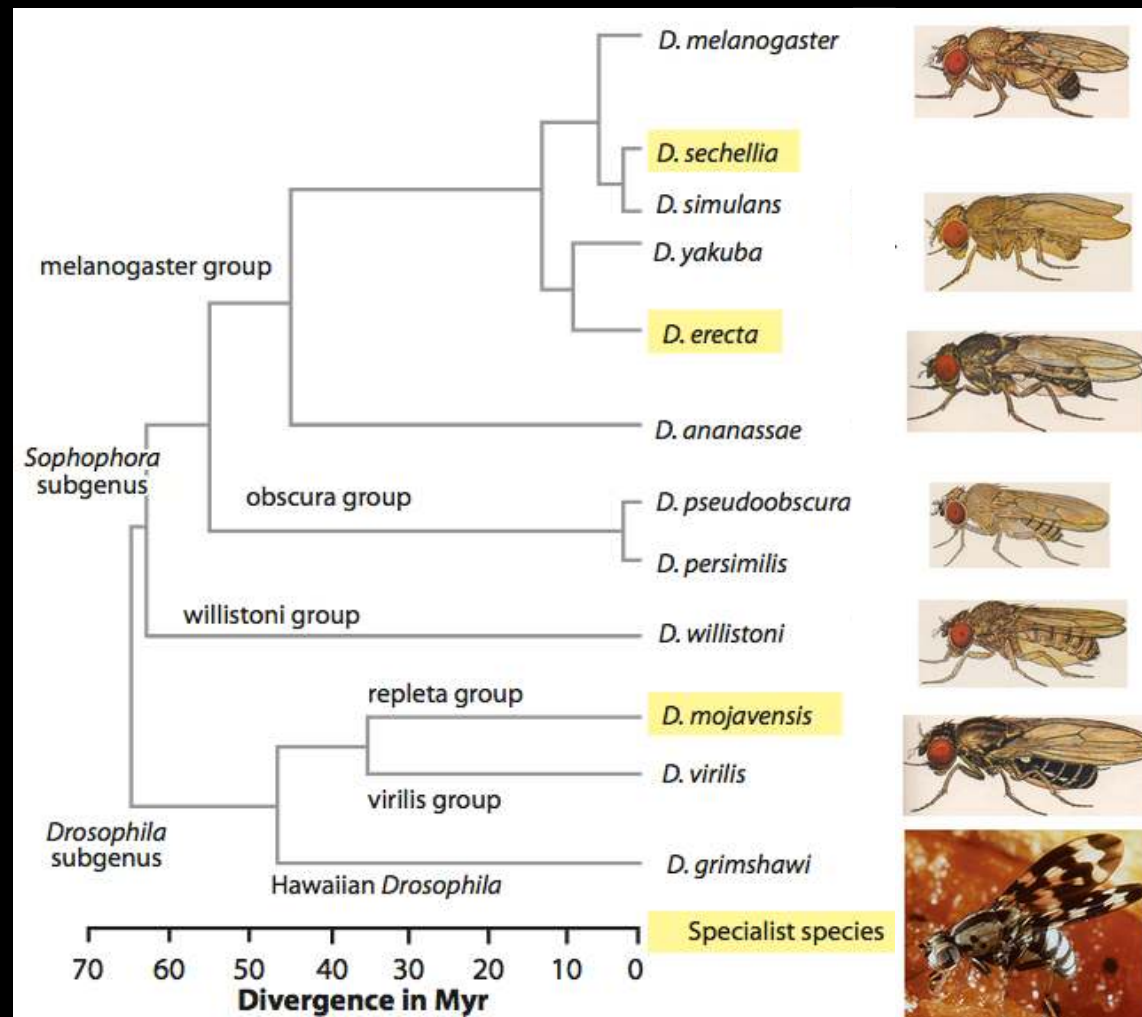


Timing of divergence

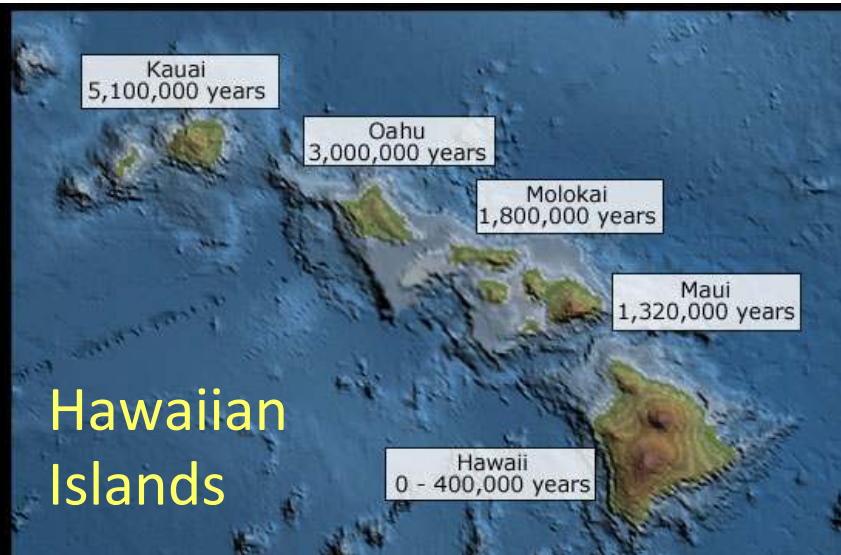
- Directly affects rate estimates
- Deriving unbiased dates from molecular data
 - Large field of software development
- Bayesian methods, while potentially informative and unbiased
 - Can be easily, and are routinely, abused



Evolution of genes and genomes on the *Drosophila* phylogeny



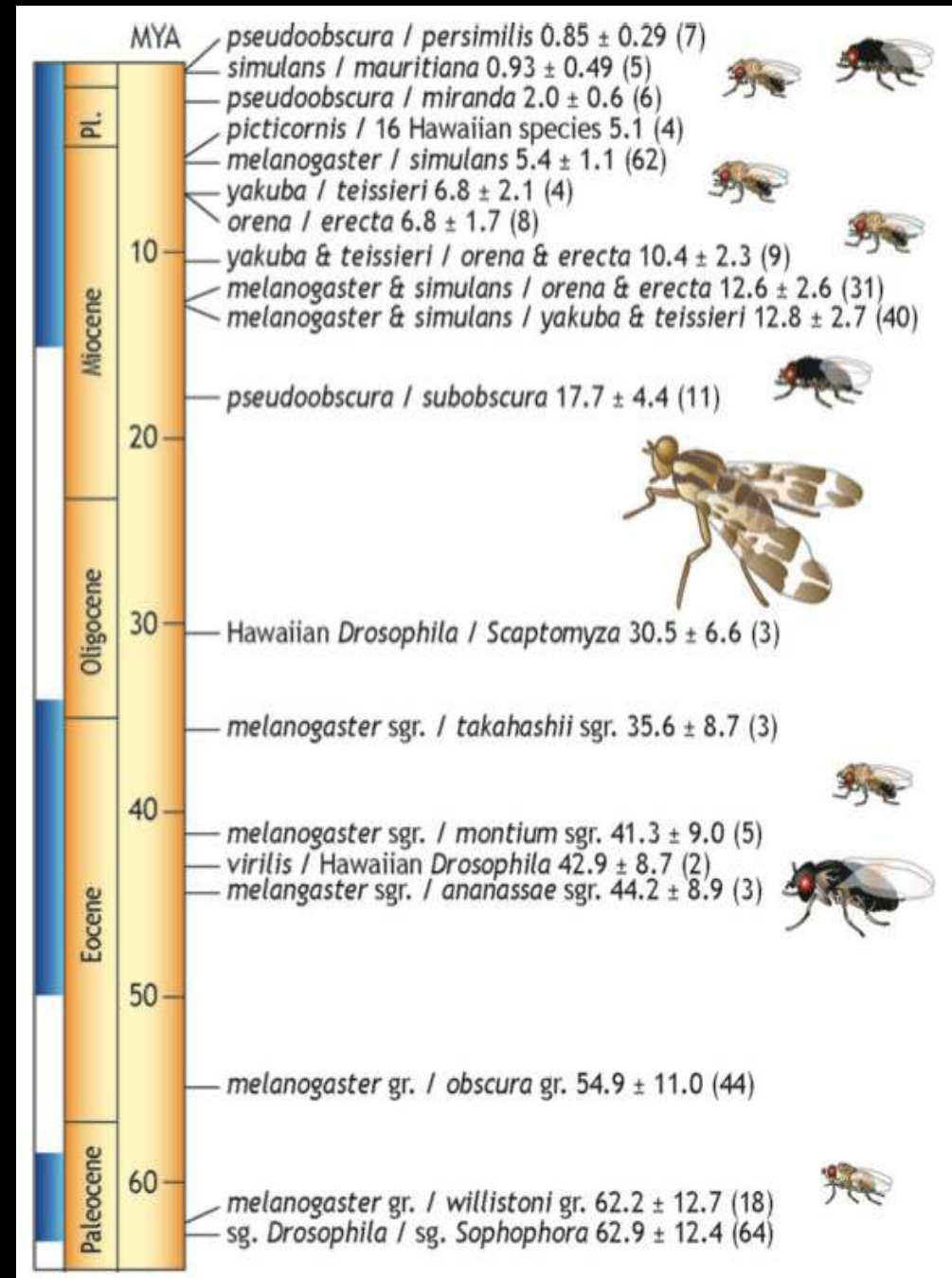
Drosophila 12 Genomes Consortium 2007 Nature



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)

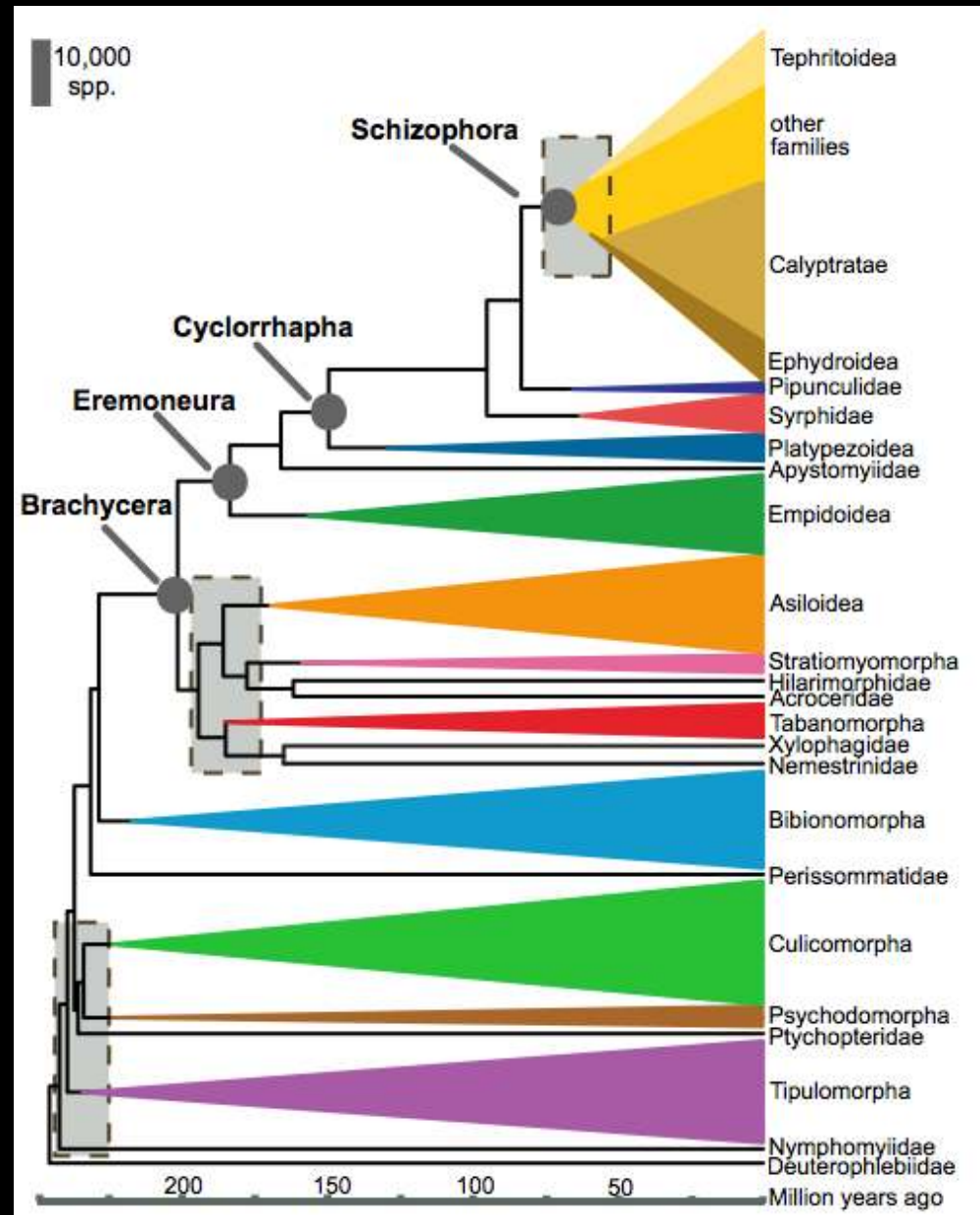




Drosophila clade:

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

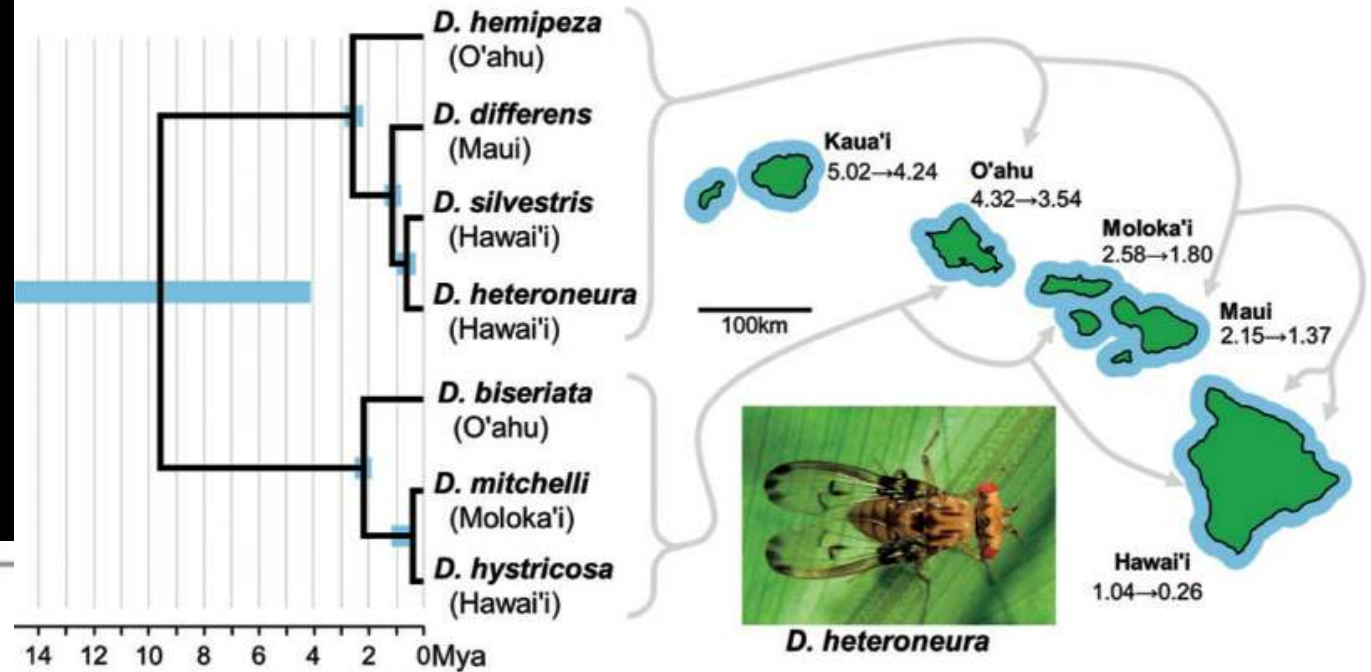
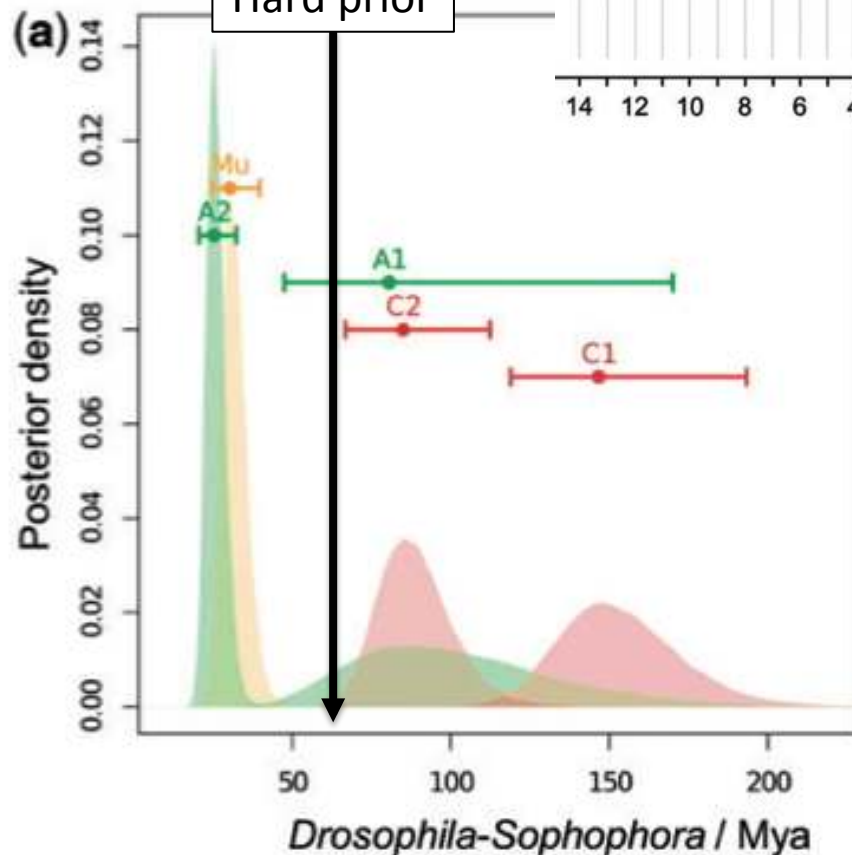
What is reality?



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

Determining objective priors is challenging

Hard prior



Priors in Bayesian rel. clock analysis:

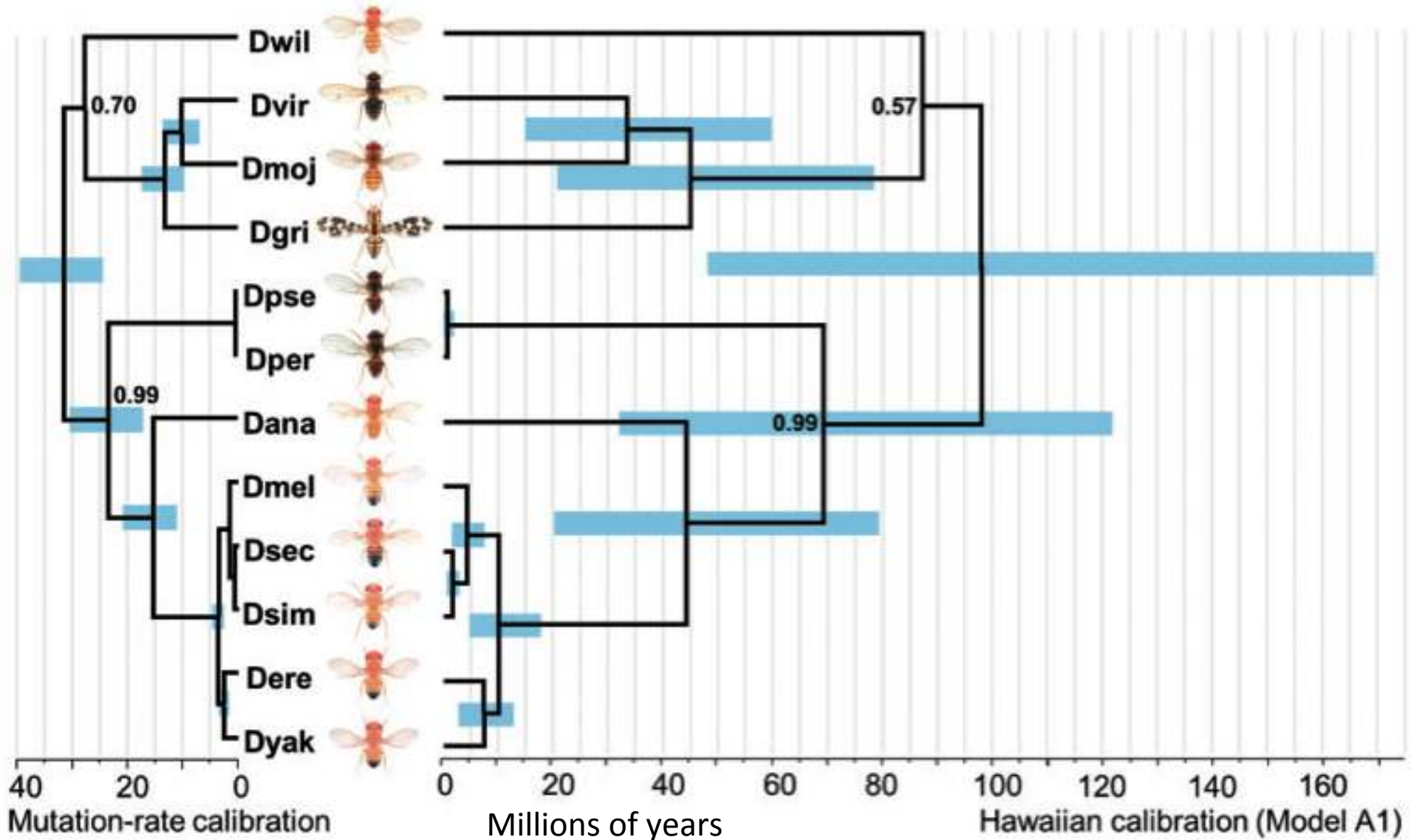
Mu = lab observed mutation rate

A1,2 = geological calibration, small Ne

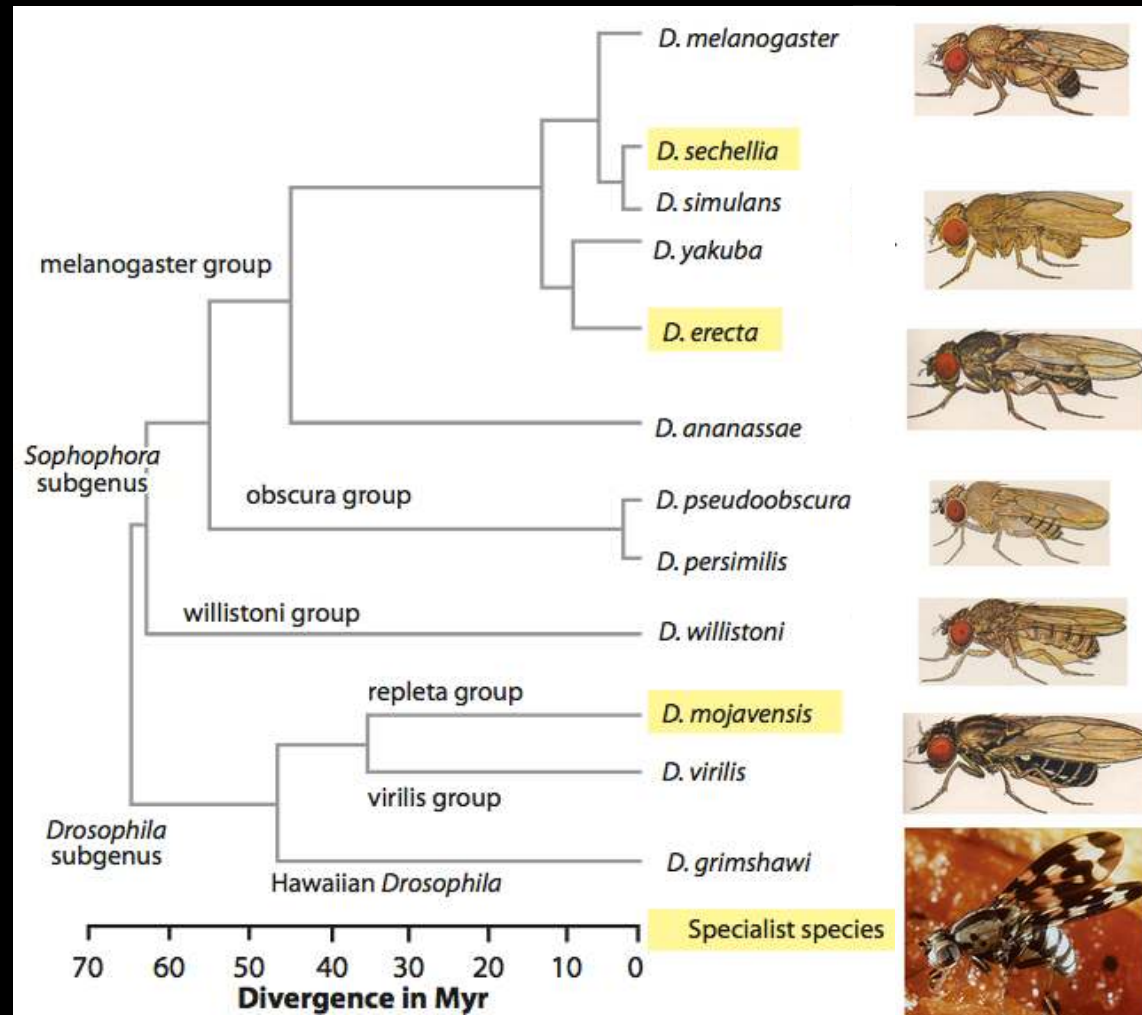
C1,2 = geological calibration, large Ne

Obbard et al. 2012 Mol. Biol. Evol.

Priors directly influence posteriors



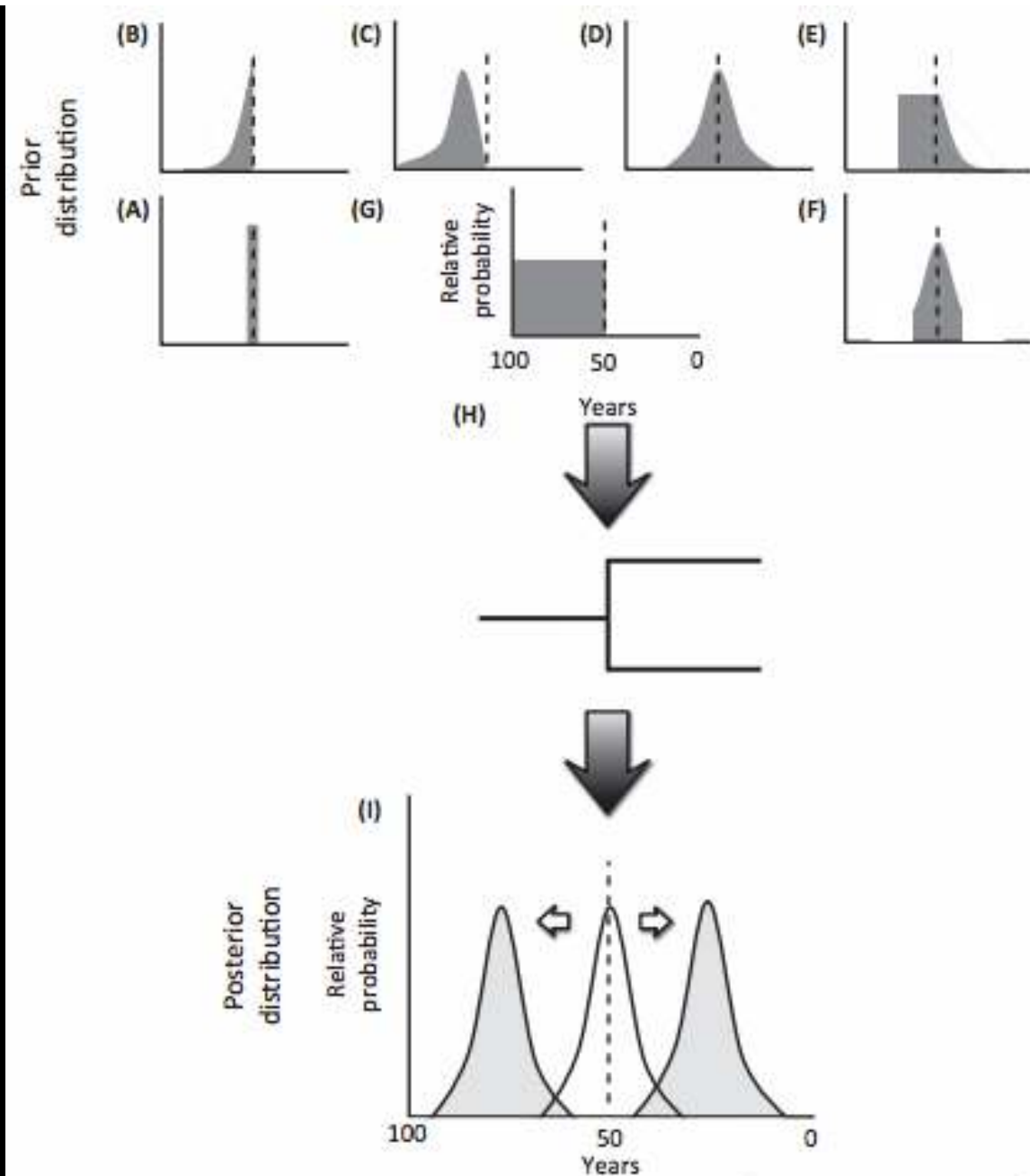
Thus, the age of this clade is fiction



Drosophila 12 Genomes Consortium 2007 Nature

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



Wheat and Wahlberg 2013
Trends Ecology & Evolution

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**

What does a
good
P-value
really tell
you?

What type
of
selection?

Is method
mismatched
to
mechanism?

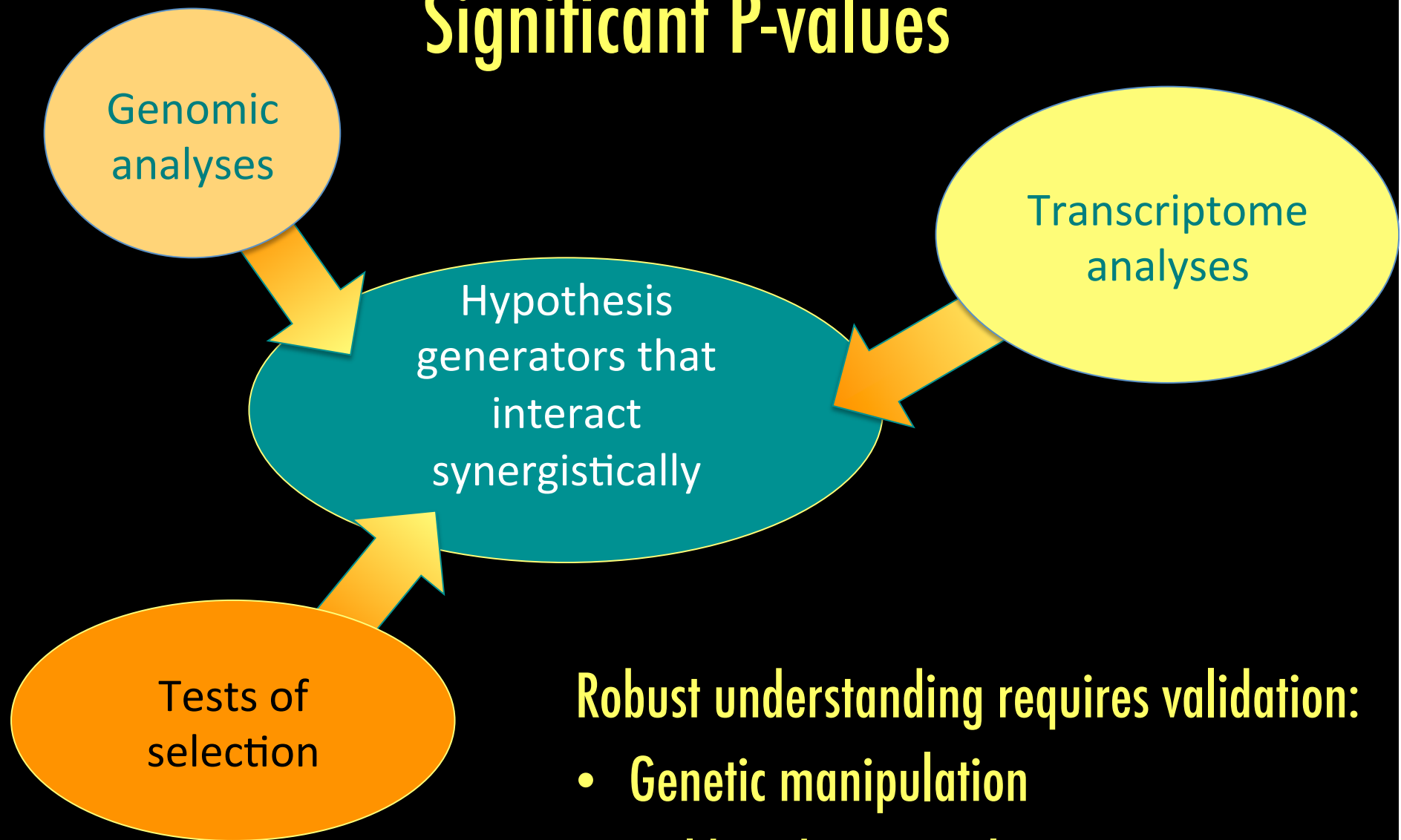
Are you
chasing a
good P-
value?

When
did
selection
happen?

What does a
bad
P-value
really tell
you?



Significant P-values



Robust understanding requires validation:

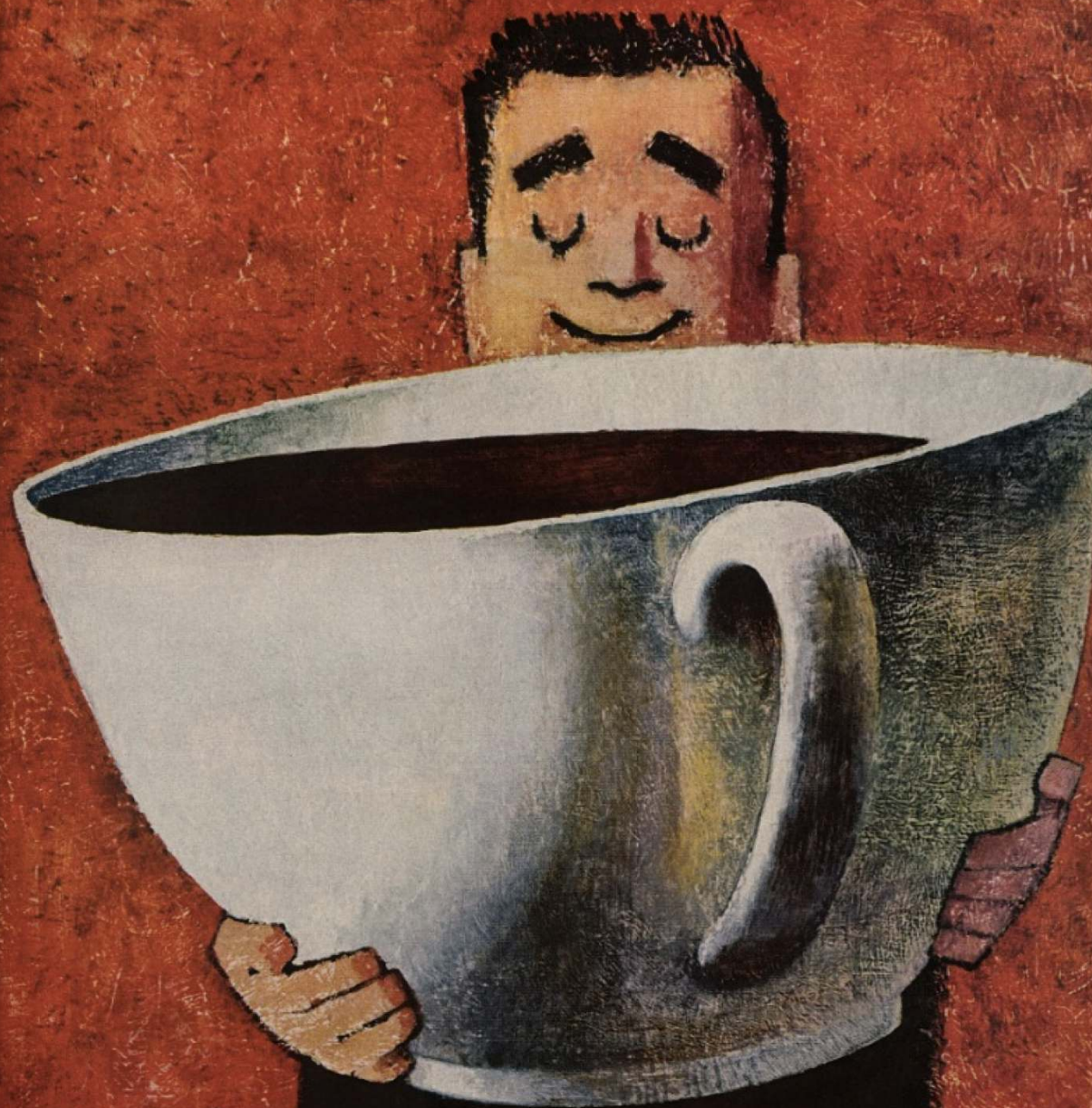
- Genetic manipulation
- Field study manipulations

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



JOHN FALIER

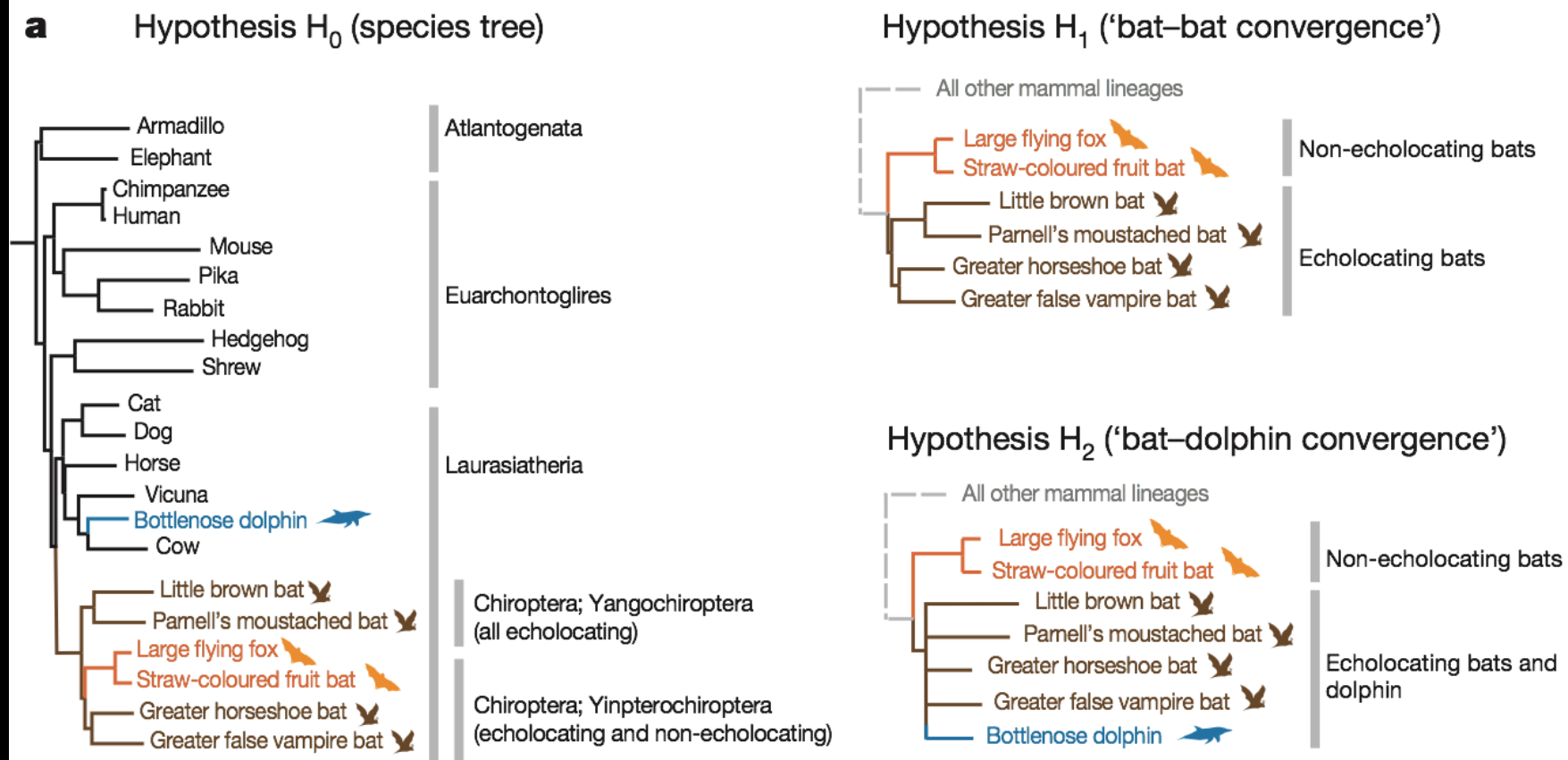


Outline

- Type I errors in studies
- How I try and avoid this
- RNA-Seq gone wrong

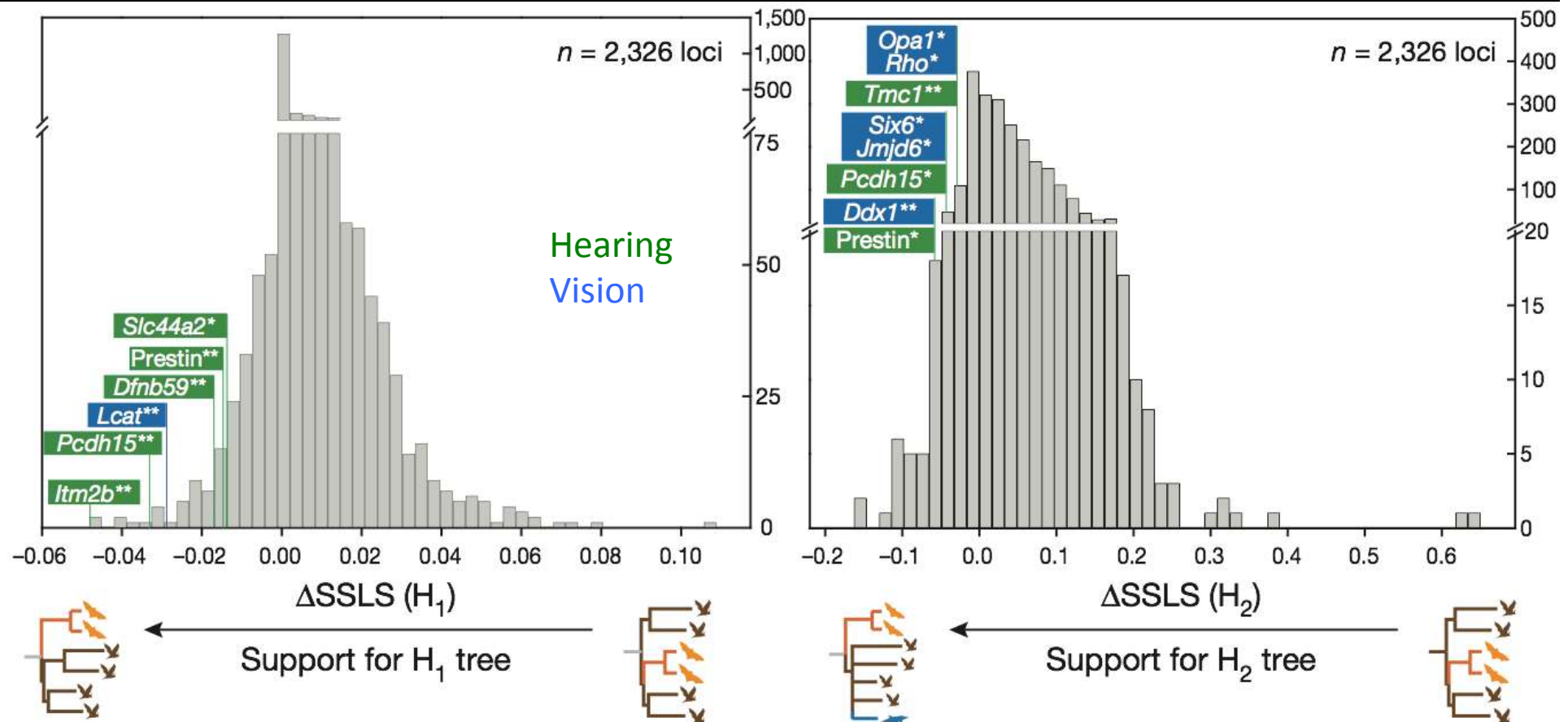


Genome-wide signatures of convergent evolution in echolocating mammals

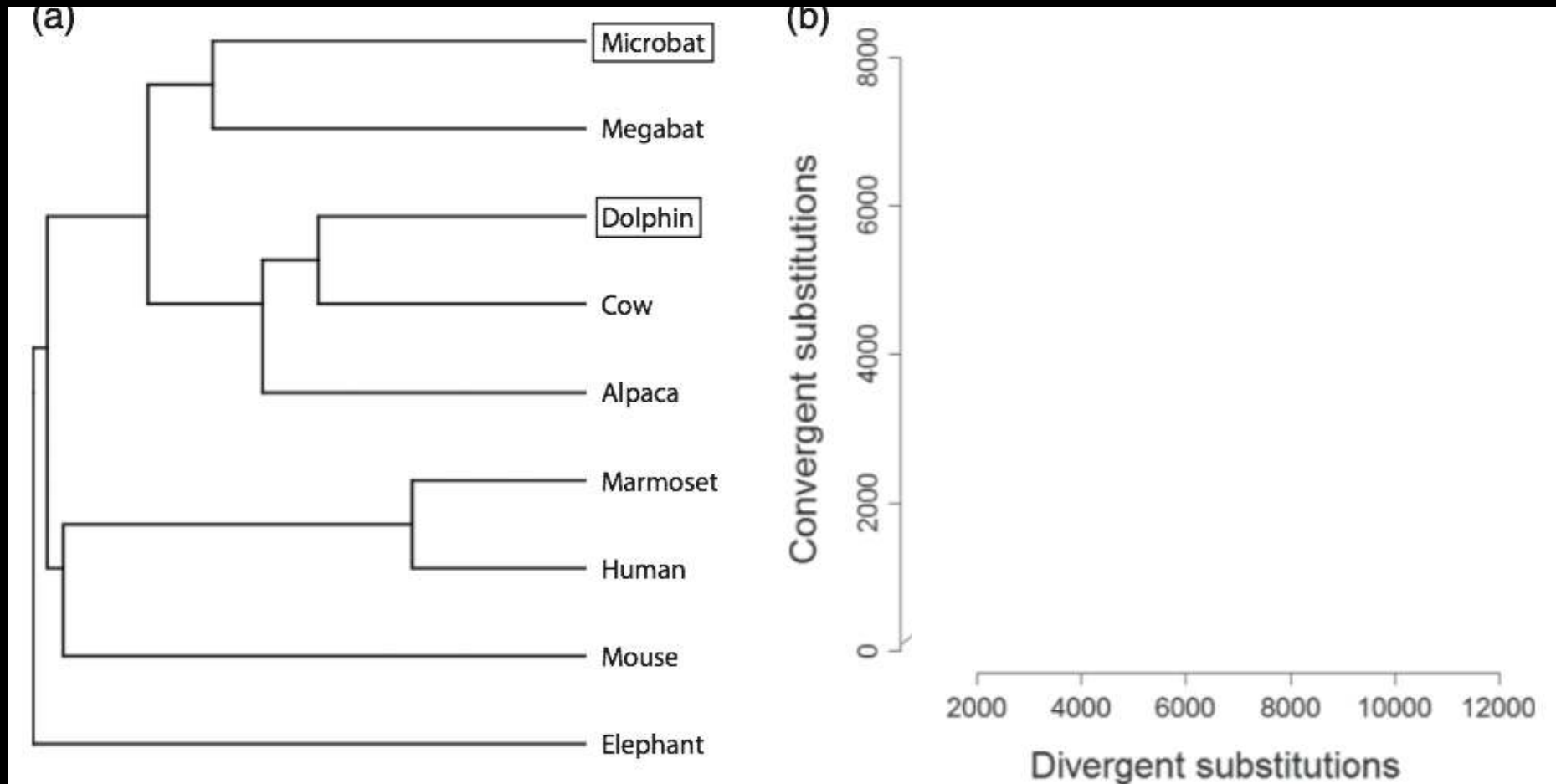


“Strong and significant support for convergence among bats and the bottlenose dolphin was seen in numerous genes linked to hearing or deafness, consistent with an involvement in echolocation.”

- 2326 orthologous genes
- site-wise log-likelihood support (SSLS)
 - Negative values support convergence H1,H2
 - 824 mean support for H1
 - 329 mean support for H2



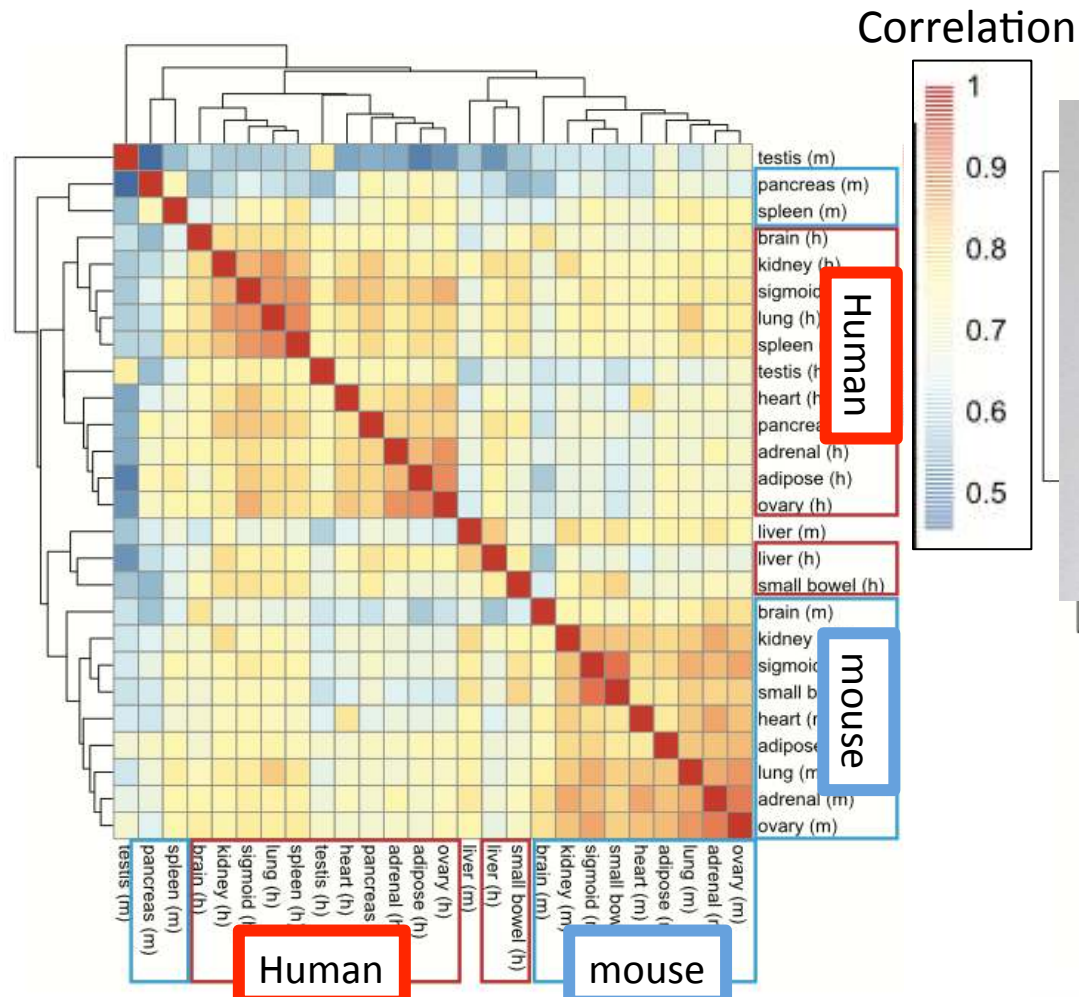
Palmer failed to conduct orthogonal 'test' of findings or estimate proper 'null' expectation



Thomas and Hahn 2015. Mol Biol Evol 32:1232–1236.

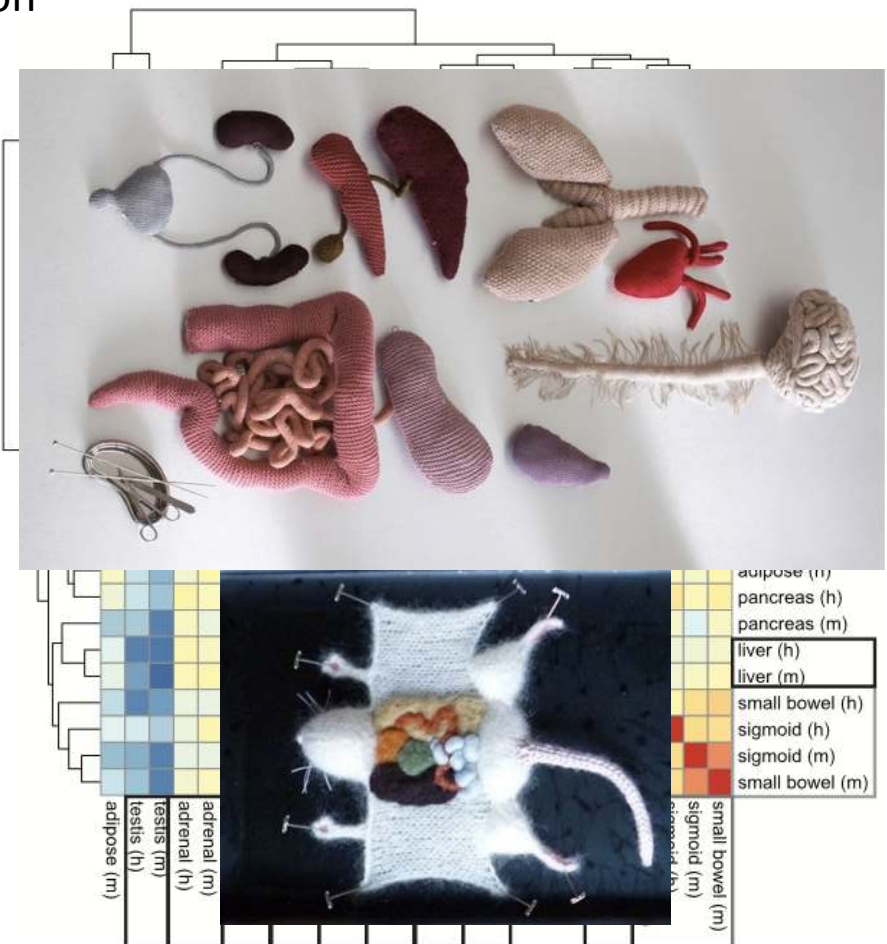
Synder mouse controversy

“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



Human – Mouse TMRCA

~90 MYA
 “[after accounting] for the batch effect, ... human and mouse tend to cluster by tissue, not by species” Gilad and Mizrahi-Man 2015. F1000 Research



Batch effect: confounding 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	
testis		pancreas		
				● Human
				● 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?

Evolutionary Inference = House of Cards?

The quality of our evolutionary inference

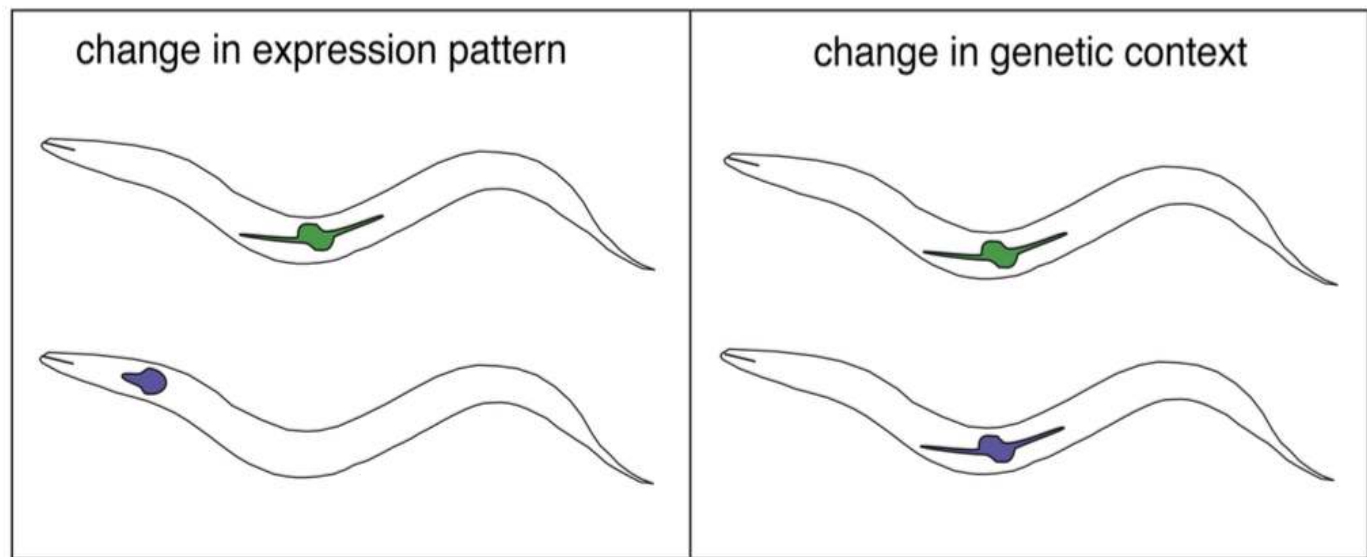
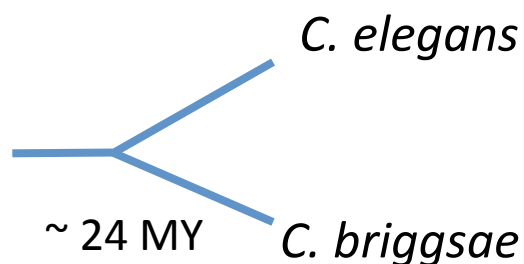
Is proportional to assumptions of orthology



Orthologous genes ... can their phenotypic effects drift over evolutionary time?

- RNAi phenotypes assessed for 1,300 genes in two nematodes
 - TMRA ~24 MYA
 - 7% had divergent phenotypic effects (in lab, etc.)
 - Likely higher in nature

Caenorhabditis



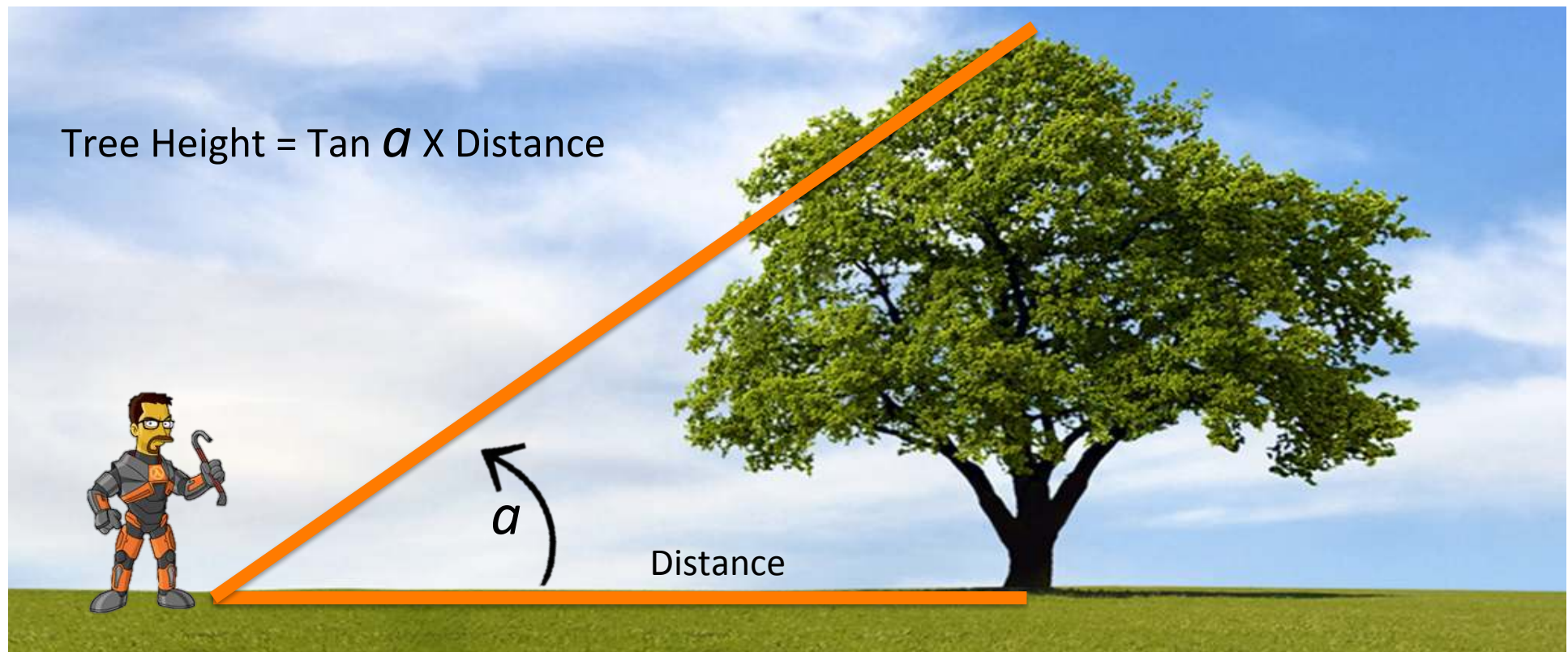
If I'm talking about all these errors ...

How do I work to minimize making type I errors?

- **I try and avoid over stating my work**
- **I 'triangulate'**

Triangulation for building evidence

- Use more than one independent set of evidence
 - Derived from independent biological replicates
- Challenge is maintaining genomic scale
 - Genome wide SNP scan for outliers, QTL mapping, RNA-Seq, knockouts, manipulations, etc.



Triangulation for building evidence

- Use more than one independent set of evidence
 - Derived from independent biological replicates
- Challenge is maintaining genomic scale
 - Genome wide SNP scan for outliers, QTL mapping, RNA-Seq, knockouts, manipulations, etc.

Move onto Triangulation quickly rather than justifying your P-value based on one dataset



These genes are DE

Outlier SNPs follow trait in F2 cross

Outlier Fst

Knockout affects phenotype

Is it an adaptation?

What was ancestral state?

Is there any clinal variation?

Phenotype respond to chemical manipulation?

Response to selection experiment?



Speckled Wood
(*Pararge aegeria*)

Genomic signal of Diapause adaptation

15 months ago, only :

- mtDNA and microsat loci
- Extensive ecological studies > 10 years



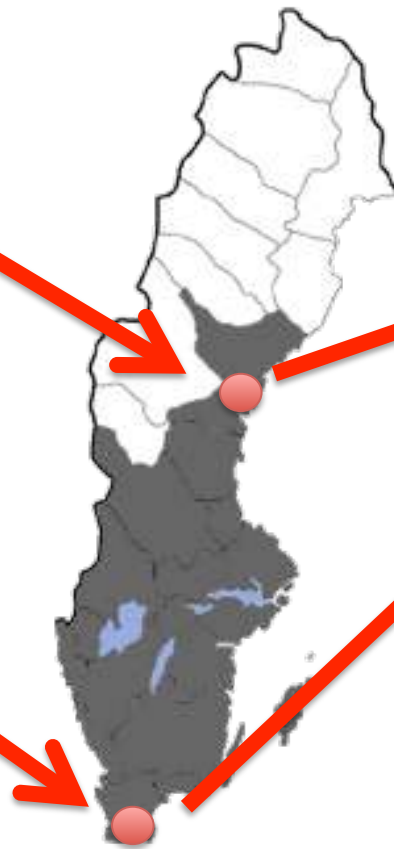
Stockholm
University



Peter Pruisscher

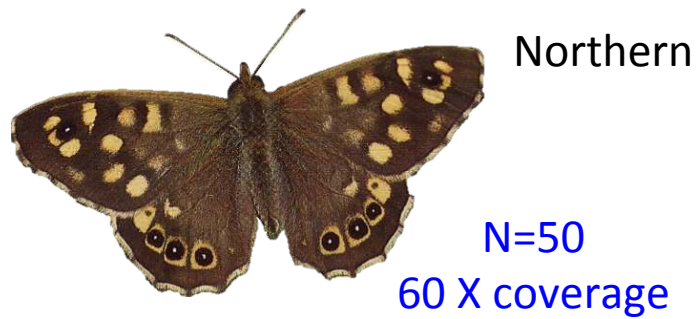


Speckled Wood
(*Pararge aegeria*)

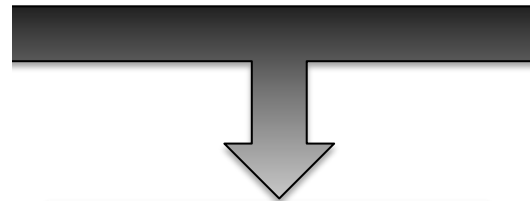
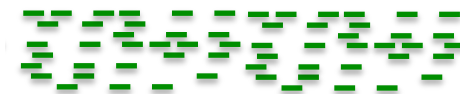
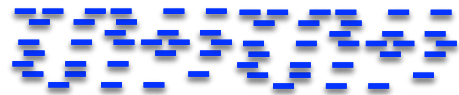


Generations per year	% in diapause at 18 hours light
1	100 %
2	0 %

What is the genetic basis
of adaptation to day
length?



GS-MESPA



De novo genome assembly



Map reads to genome



Map reads to genome

What regions are different?

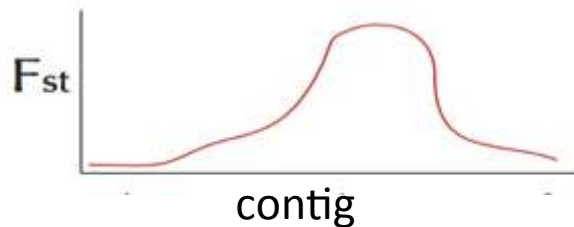
A	A	T
T	A	A
A	A	A
T	A	T

Call SNPs

T	T	T
A	T	A
T	T	A
A	T	T

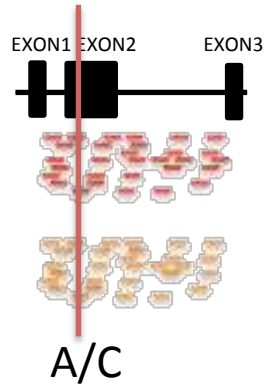
What genes are in those regions?

Scaffold contigs, find exons



Gene features

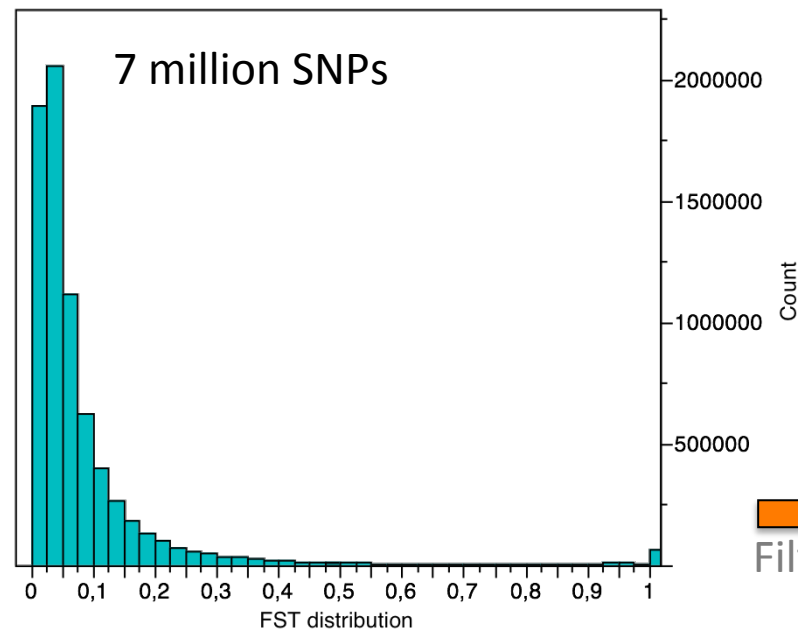
Fst outlier analysis for candidates



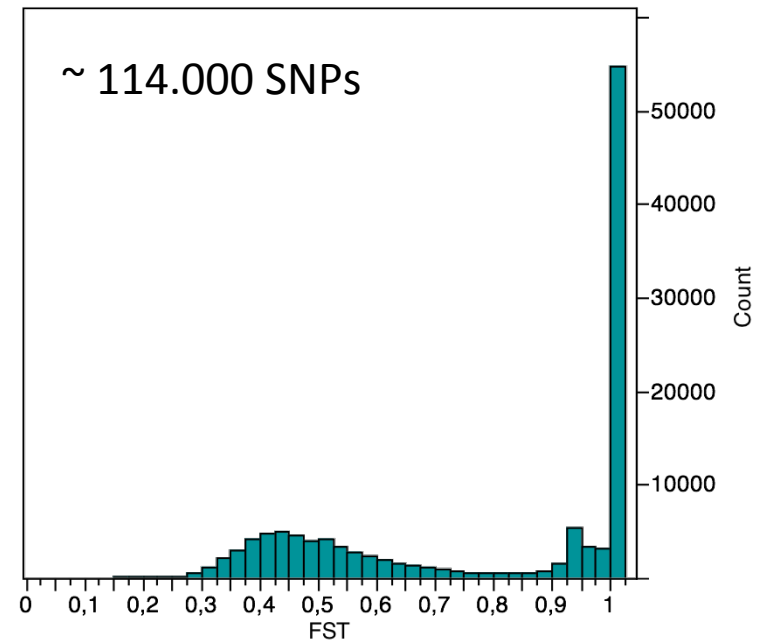
11,000 gene models & ~7 million SNPs

Quality Filtering

~ 114,000 SNPs of which 68,000 SNPs: $F_{ST} > 0.9$

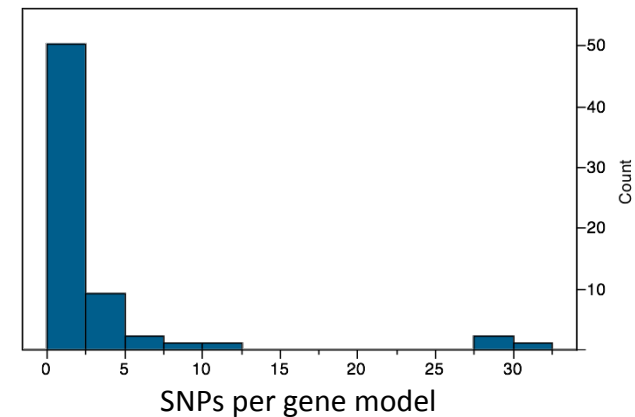


Filtering



Fixed variation in genes

1. Intergenic regions contain +/- 67,604 Fixed SNPs
2. 67 gene models contain 209 fixed SNPs
3. Filter for SNPs in exons and introns

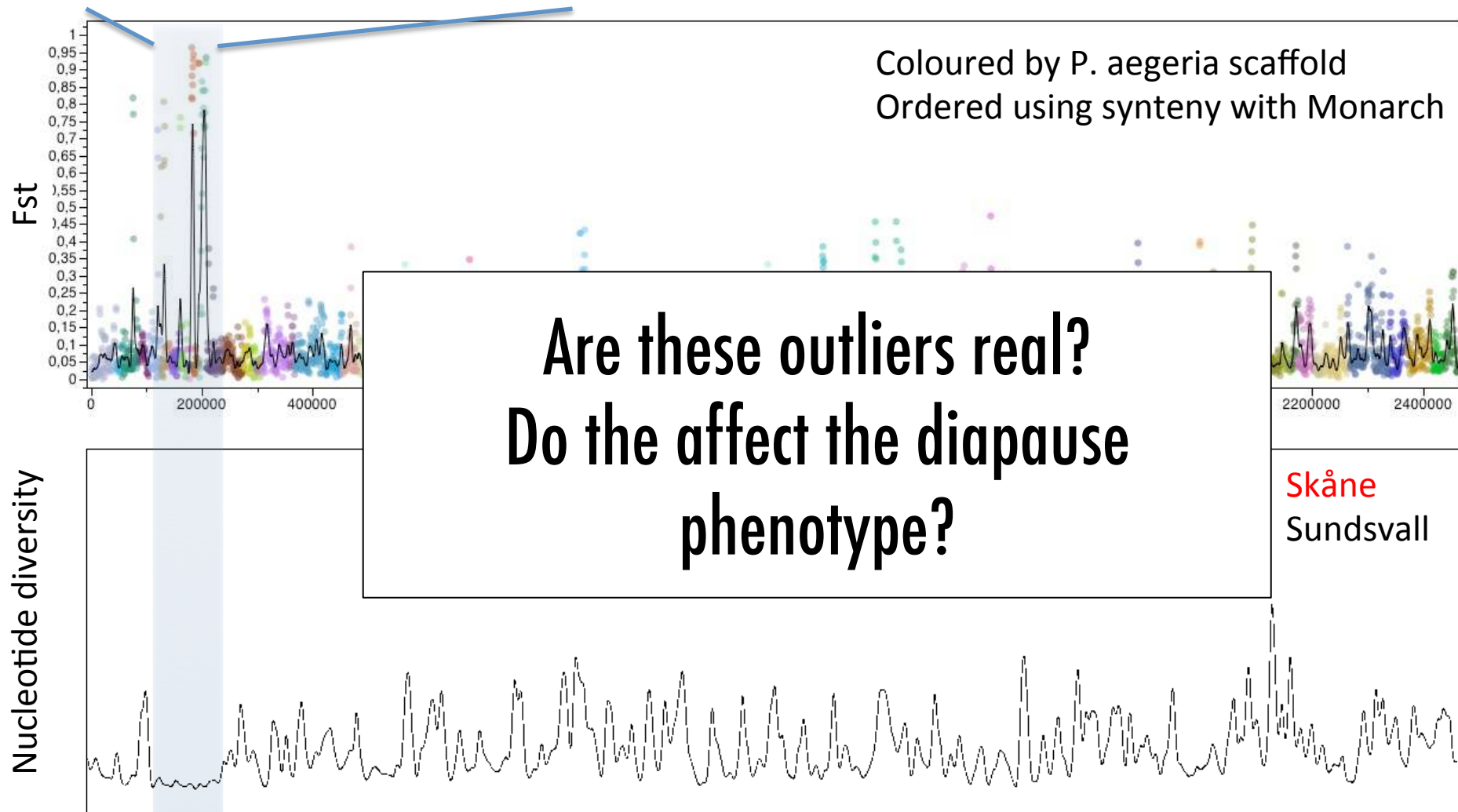


UniRef90_proteinnames	exon	gene	intergenic	Total	D.plex scaffold	Bmori_chr
Timeless	2	0	0	2	DPSC300014	chr4
Carnitine O-acetyltransferase	3	25	1	29	DPSC300014	chr4
Trypsin-like protein	2	14	14	30	DPSC300041	chr5
Vasa-like protein	1	2	0	3	DPSC300379	chr19
Period	2	2	1	5	DPSC300005	chr1

Is there a foot-print of selection around these SNPs?

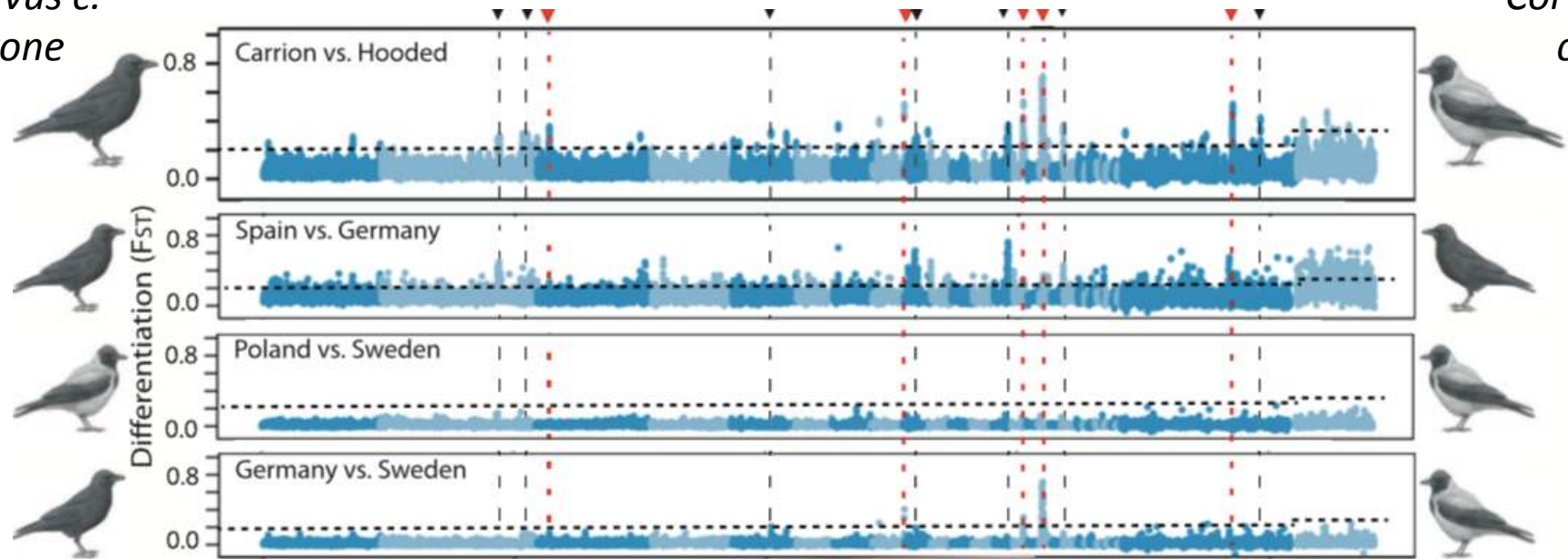
Region around timeless

Timeless; Carnitine O-acetyltransferase



120,000 bp window

*Corvus c.
corone*

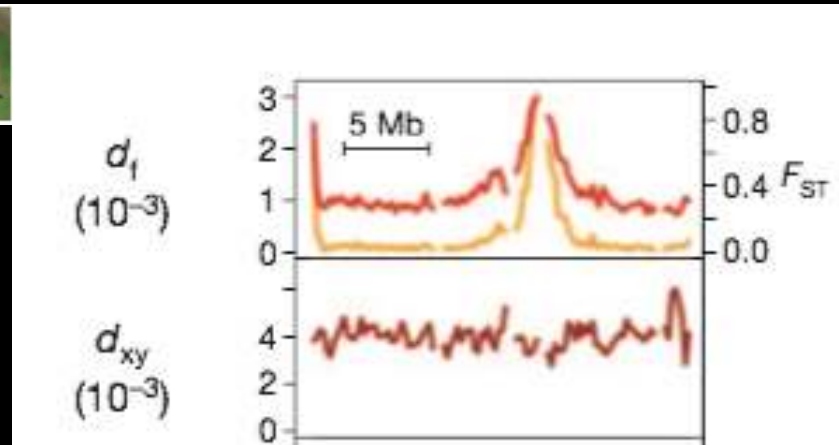


*Corvus c.
cornix*

Islands of speciation or background selection?



D_{xy} :
An absolute
measure of
differentiation,
increase due to
mutations

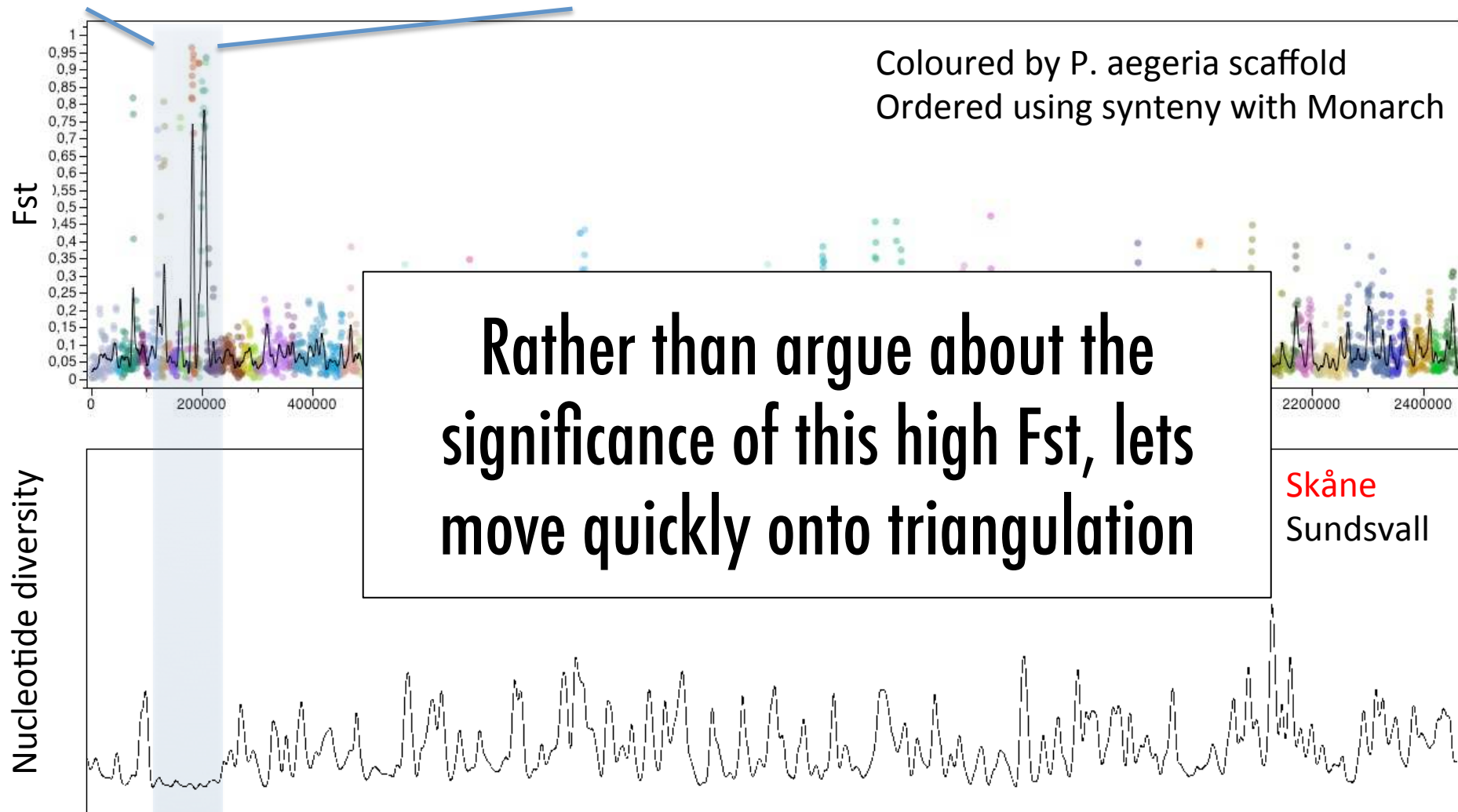


F_{ST} :
A relative measure
of differentiation,
increases due to
freq. change

The absence of high D_{xy} in regions of high F_{ST} suggest a role of background selection driving these patterns rather than genomic 'islands' driving speciation.

Region around timeless

Timeless; Carnitine O-acetyltransferase

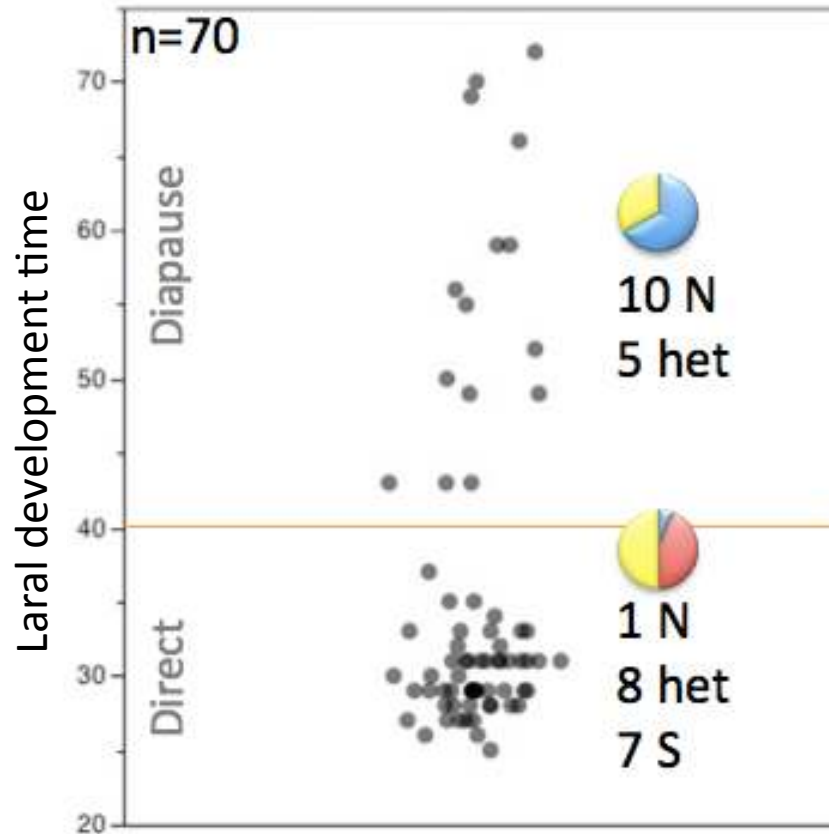


120,000 bp window

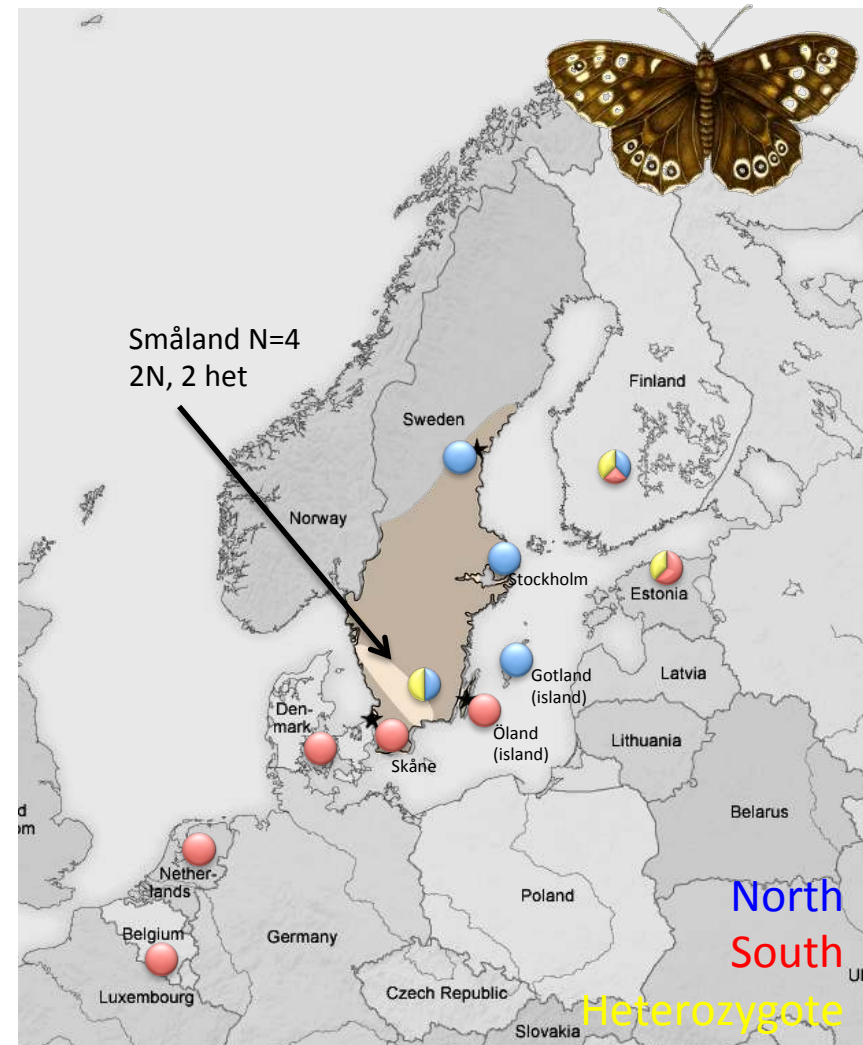
Window

Triangulating Timeless

SNP genotyping in F2 cross



Clinal analysis



1001 ways for your pipeline to break

An overview of genomic pipeline
challenges

Christopher West Wheat



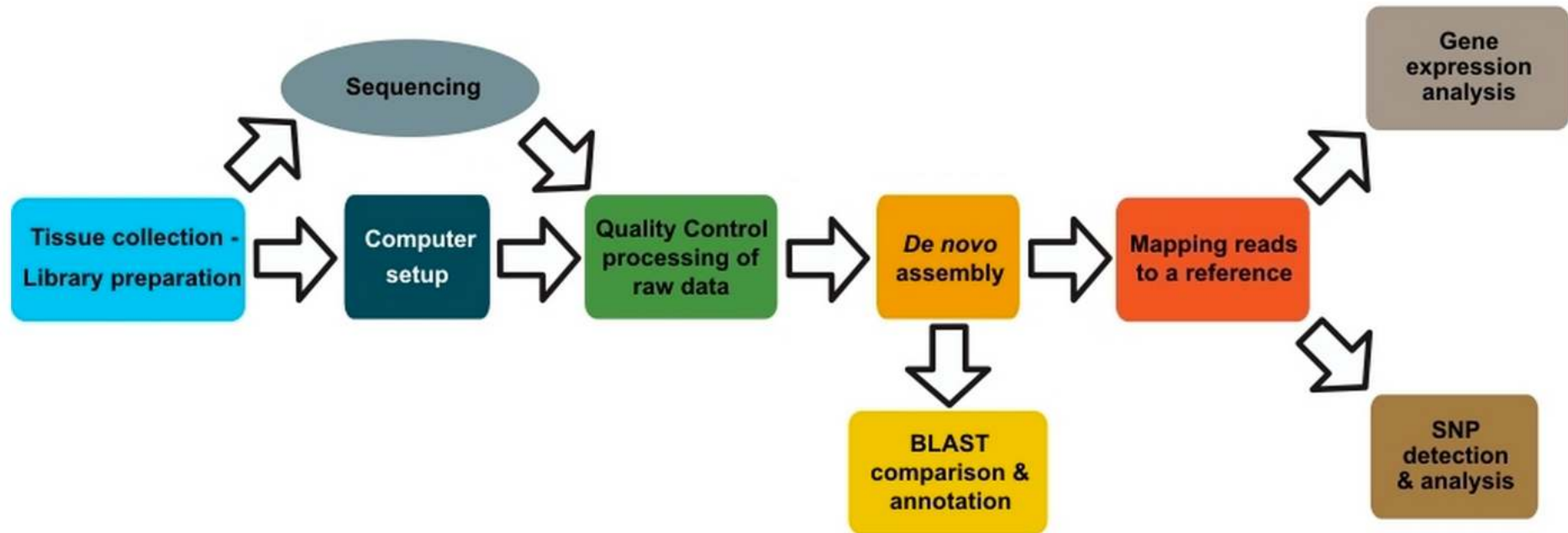
Informatics and Biology

- We need to make sure we put the 'bio' into the bioinformatics
 - Do results pass 1st principals tests
 - Always double check data from your core facility or service company
 - Use independent analyses as 'controls' on accuracy
 - What are your + and – controls?
 - Do independent methods converge?
- Need to re-assess our common metrics for potential bias in the genomic age
 - Bootstraps on genomic scale data
 - P-values, outlier analyses, demographic null models

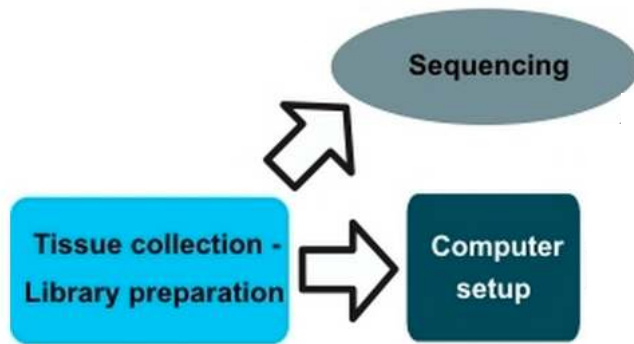
Outline

- Transcriptome analyses in non-model species
 - Walk through pipeline and highlight issues of concern
 - What is validation?
- Insights from candidate genes
 - Can Second Gen methods get us there?

Pipeline Overview



Pipeline Overview



How can I study
my data using
open source?

How
much
RAM do I
need?



Are 16 cores
enough?

Can I
use my
laptop?

What
software &
how do I
get it?



Why
Linux?

How much
HD space
is needed?

Computer Infrastructure



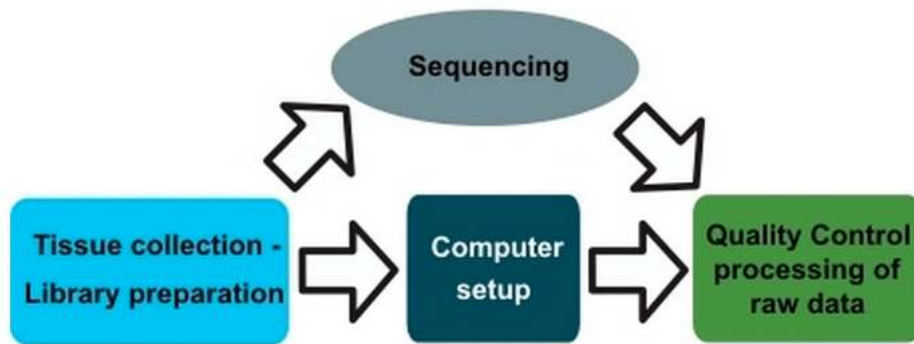
RNAseq dataset:

4 conditions X 2 tissues X 3 families X 3 replicates = 72×10^6 reads

	File Sizes (Gb)	CPUs	RAM (Gb)	Time
Raw files *.gz	(1.5 Gb)	8	8	~3 hours / file
Raw files expanded				
TA assembly				~2 weeks
Mapping (BAM)				~3 hours / file
Annotation	100 Mb	8	8	~6 – 12 days
Analysis	< 20 Mb	4	4	~< 1 hour
Visualization	BAM files	≥ 4	≥ 8	

Get ready for your data by downloading similar sized dataset from the Short Read Archive. Do not wait till it arrives

Pipeline Overview

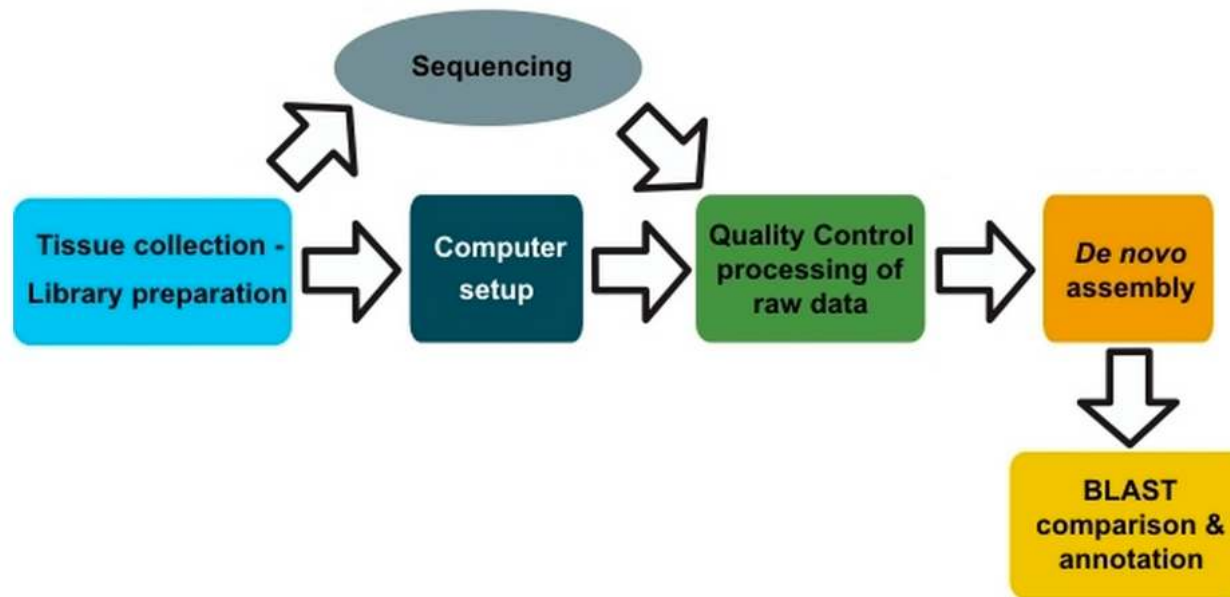


Core facilities and non-model species

Statements from core facilities that are not true:

- Here is your data
- You can't do RNA-Seq without a genome
- We'll have your data back in < 1 month

Pipeline Overview



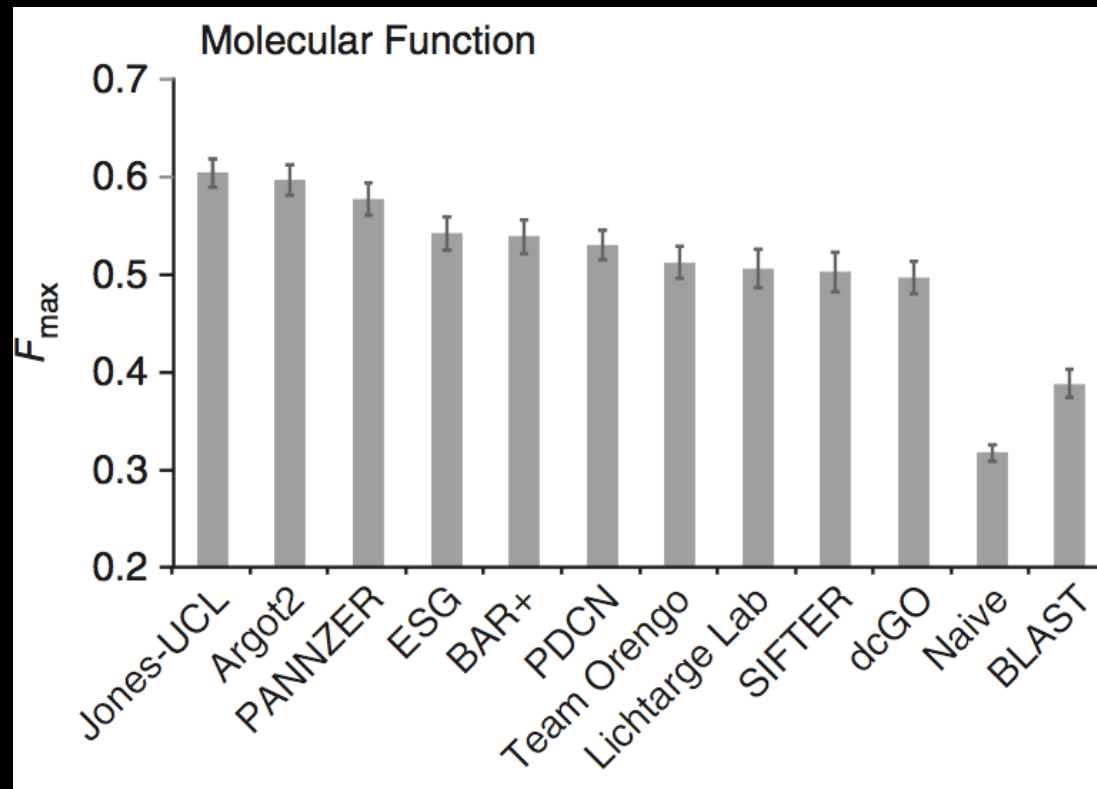
Gene Ontology: order in the chaos

- Addresses the need for consistent descriptions of gene products in different databases in a species-independent manner
- GO project has developed three structured controlled vocabularies (ontologies) that describe gene products in terms of their associated
 - biological processes
 - cellular components
 - molecular functions

<http://www.geneontology.org/>



Comparisons among annotation tools



Radivojac et al.: A large-scale evaluation of computational protein function prediction. *Nat Meth* 2013, **10**:221–227.

Falda et al. Argot2: a large scale function prediction tool relying on semantic similarity of weighted Gene Ontology terms. *BMC Bioinformatics* 2012, **13**:S14.



Functional annotation of proteins using the semantic similarity in the Gene Ontology

Site Homepage

Insert sequences

Batch processing

Consensus analysis

DB releases

View SGE jobs

View SGE queues

Argot² help

About

a.r.g.o.t.²

We present a novel method called **Argot²** (Annotation Retrieval of Gene Ontology Terms), that is able to quickly process thousands of sequences for functional inference. The tool exploits a combined approach based on the clustering process of GO terms dependent on their semantic similarities and a weighting scheme which assesses retrieved hits sharing a certain degree of biological features with the sequence to annotate. These hits may be obtained by different methods as BLAST, HMMER and so on. In the present web server we allow users to interact with Argot² in different ways according to specific needs and expertise.

If you use our service, please cite:

- × Fontana P, Cestaro A, Velasco R, Formentin E, Toppo S.
Rapid annotation of anonymous sequences from genome projects using semantic similarities and a weighting scheme in gene ontology.
PLoS One. 2009;4(2):e4619. Epub 2009 Feb 27. PubMed PMID: 19247487; PubMed Central PMCID: PMC2645684.
- × Falda M., Toppo S., Pescarolo A., Lavezzo E., Di Camillo B., Facchinetti A., Cilia E., Velasco R., Fontana P.
Argot²: a large scale function prediction tool relying on semantic similarity of weighted Gene Ontology terms.
BMC bioinformatics, 13(4). 2012.

News:

× Databases
Check [this](#)

Batch processing for GO terms

Site Homepage

Insert sequences

Batch processing

Consensus analysis

DB releases

View SGE jobs

View SGE queues

Argot² help

About

Please select the zipped tabular BLAST and HMMer files, see [here](#) for details, to upload ($\leq 1\text{GB}$). ?

Please do not upload more than 5000 sequences at once, otherwise the service will be overloaded.

BLAST: No file chosen ?

HMMer: No file chosen ?

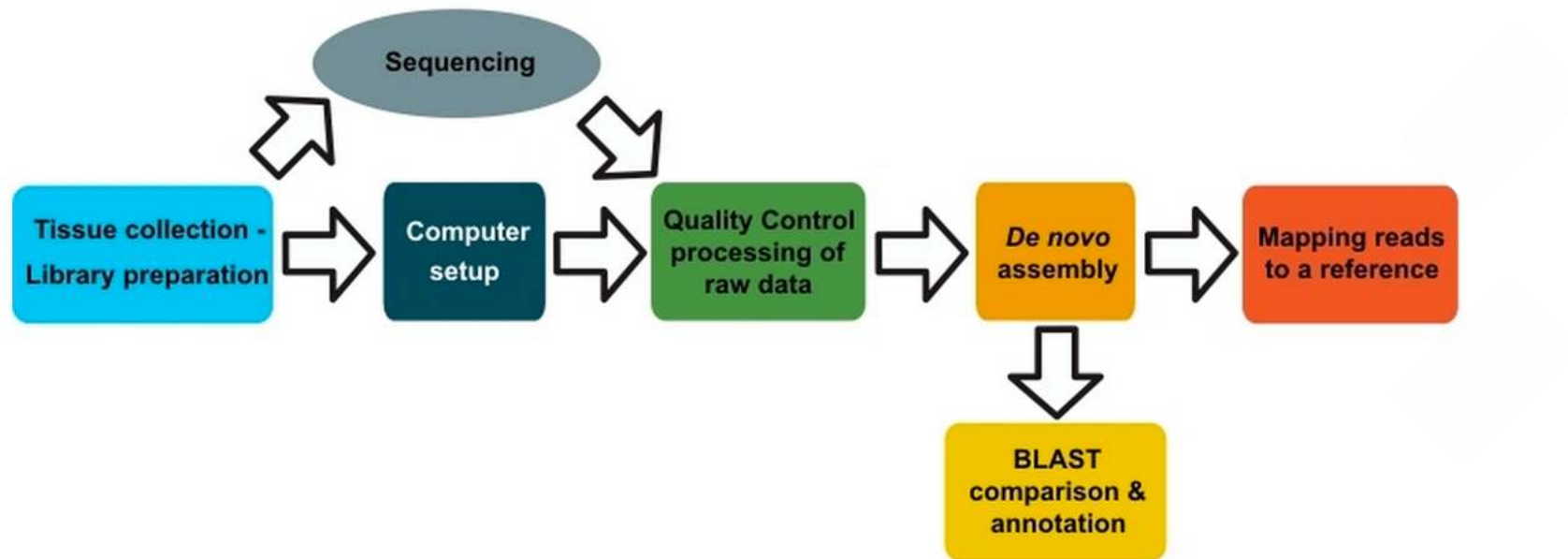
☐ submit example data ?

Email: ?

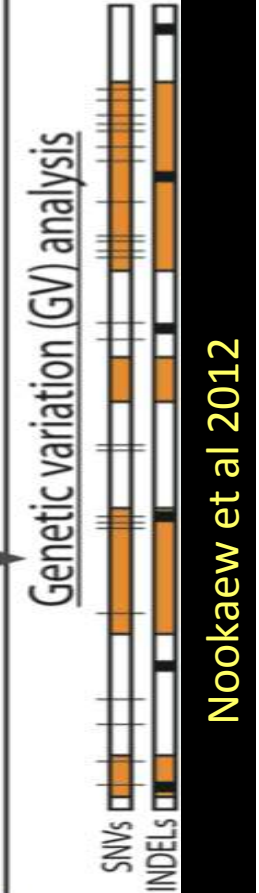
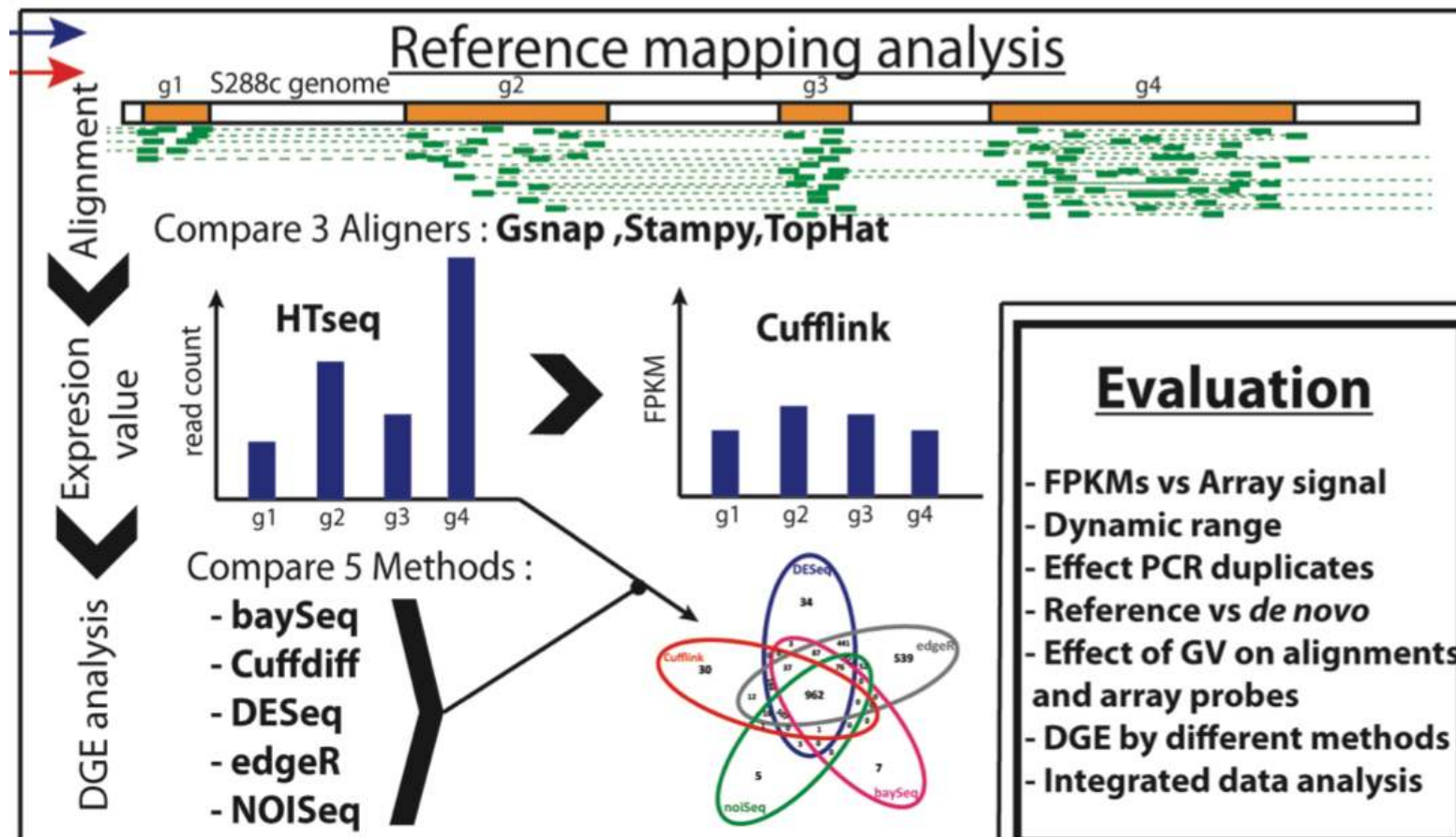
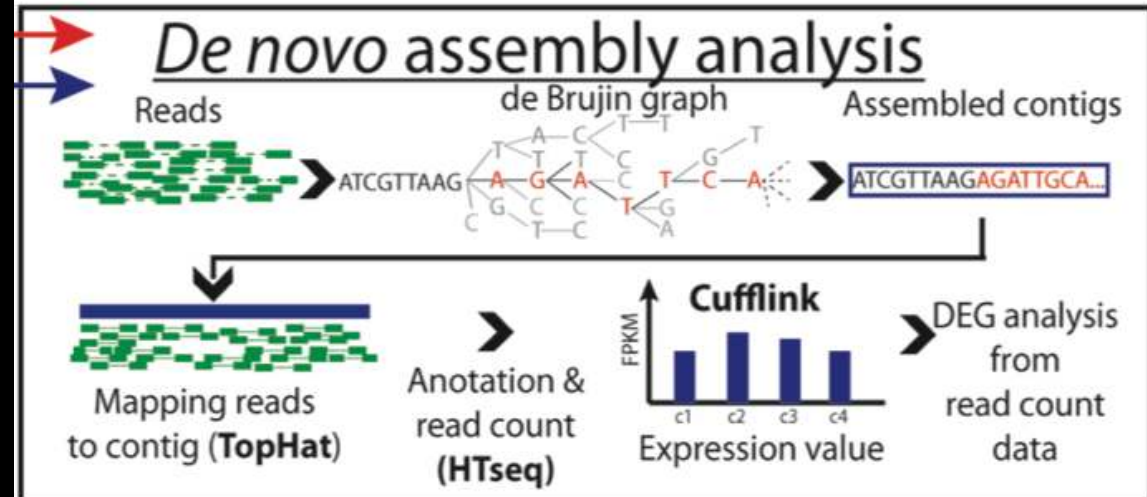
CUT-OFF ([meaning](#)) ?

Total Score (≥ 5):

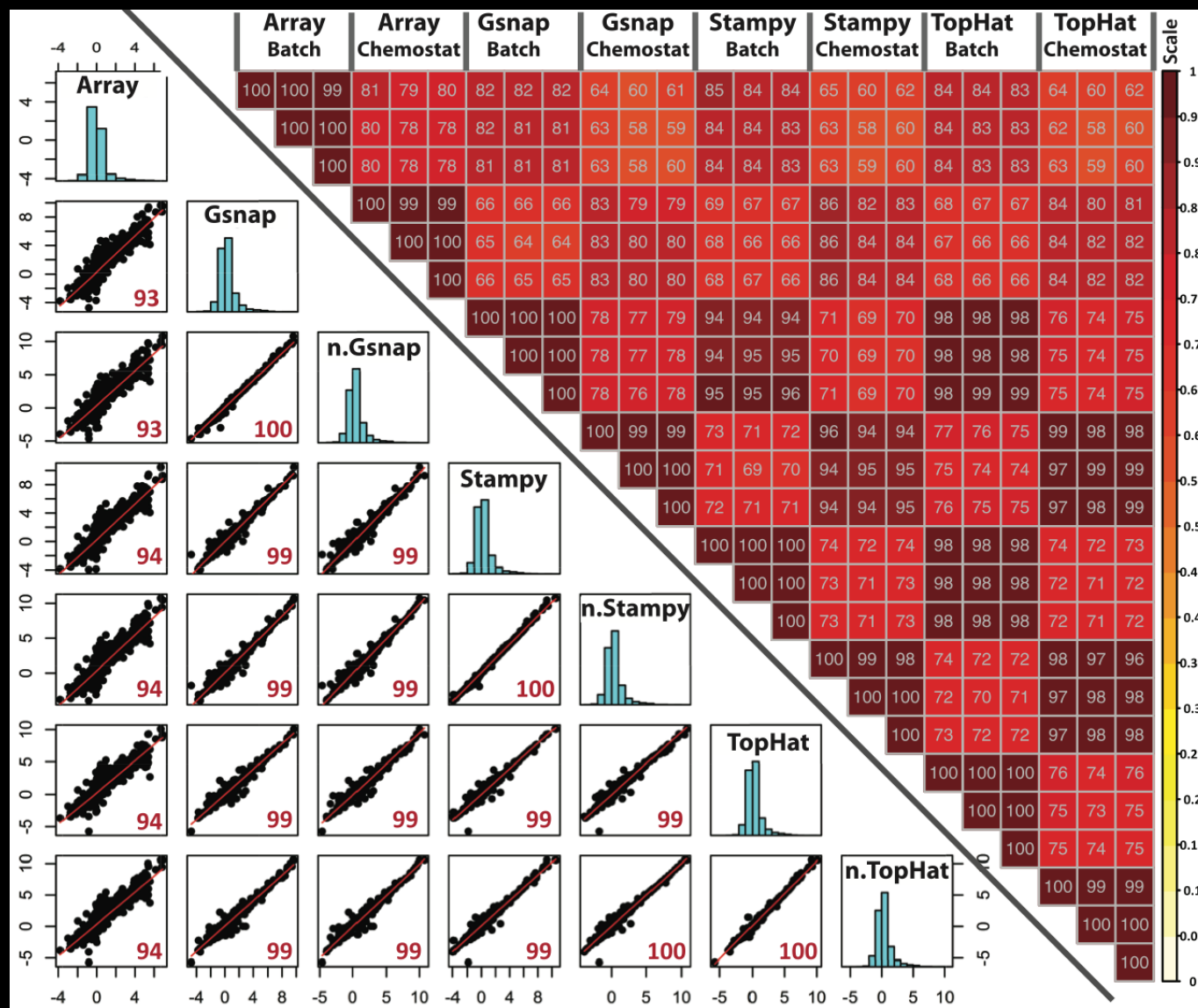
Pipeline Overview



Template mismatch effects: excellent yeast study



Does alignment software matter?



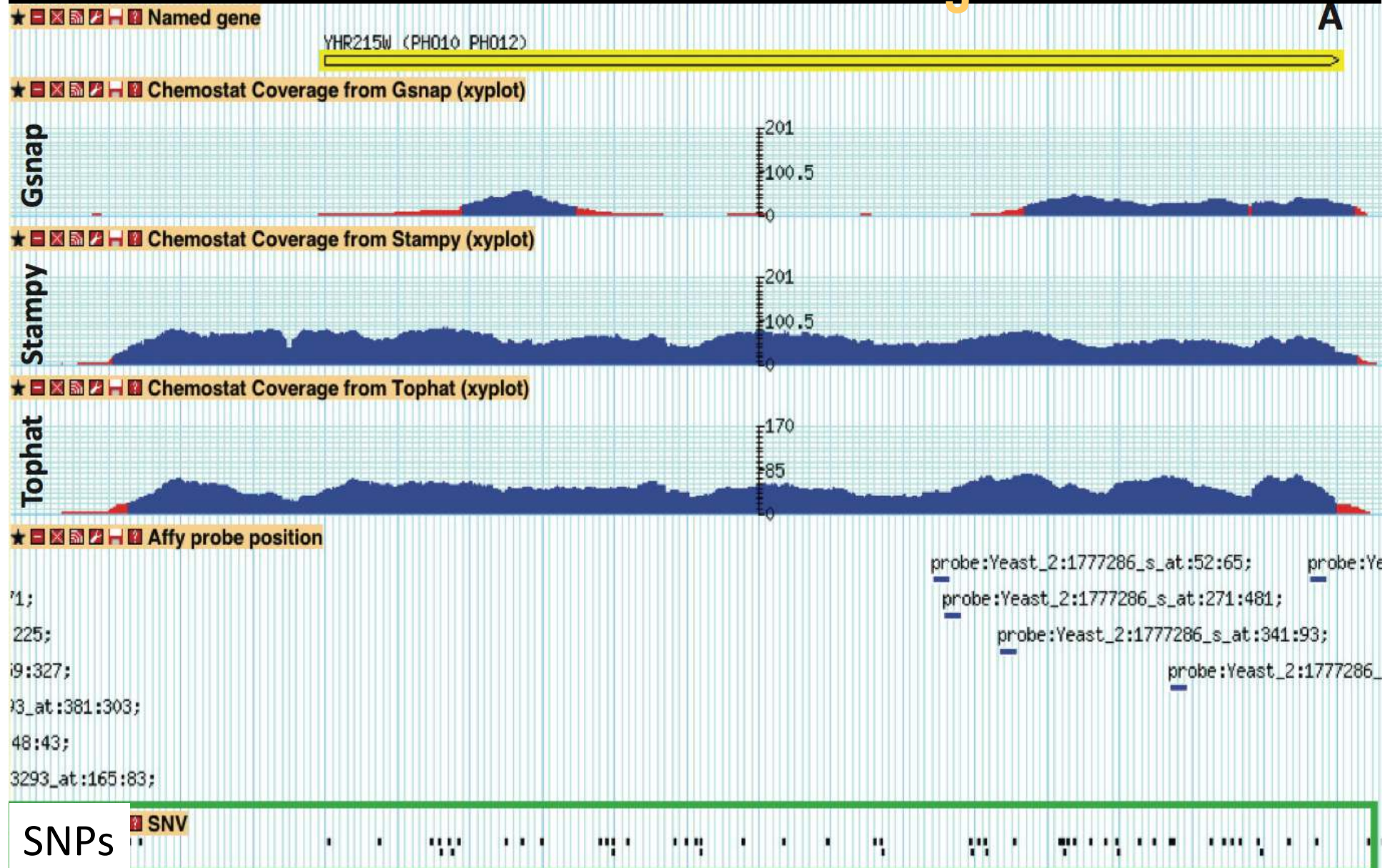
Nookaew et al. A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in *Saccharomyces cerevisiae*. *Nucleic Acids Research* 2012, **40**:10084–10097.

Mappers don't appear to matter

Wrong

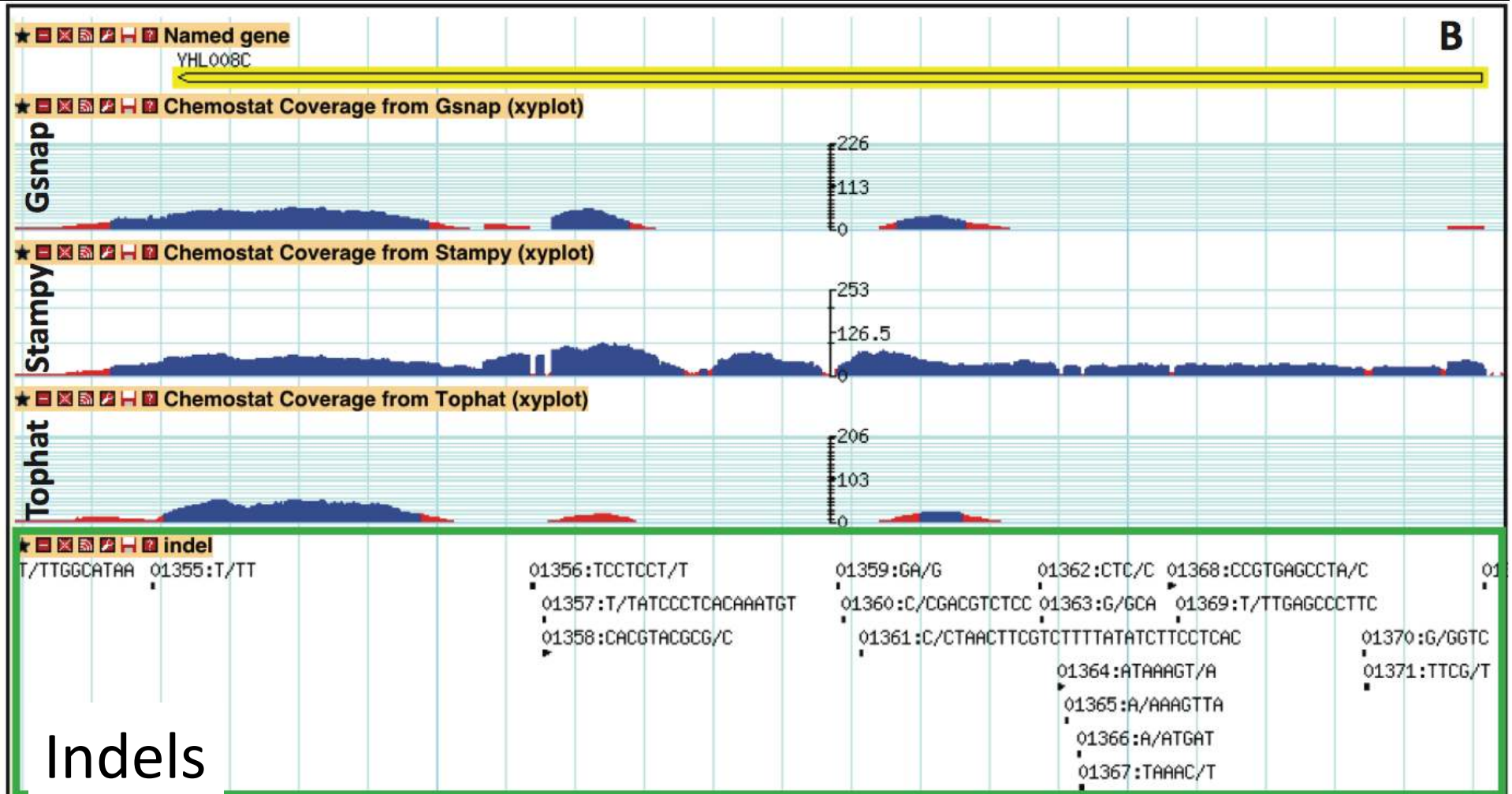
- Genomic scale data can hide widespread biases that unless you specifically look, are hard to find
- Mapping programs differ in their settings and design
 - DNA to DNA vs. RNA to DNA
 - Are usually compared using species without much genetic variation
 - Indels, splicing, SNPs all affect mapper performance

SNP effects can be large



Nookaew et al. A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in *Saccharomyces cerevisiae*. *Nucleic Acids Research* 2012, **40**:10084–10097.

Insertions & deletions (indels) have large effects



Indels

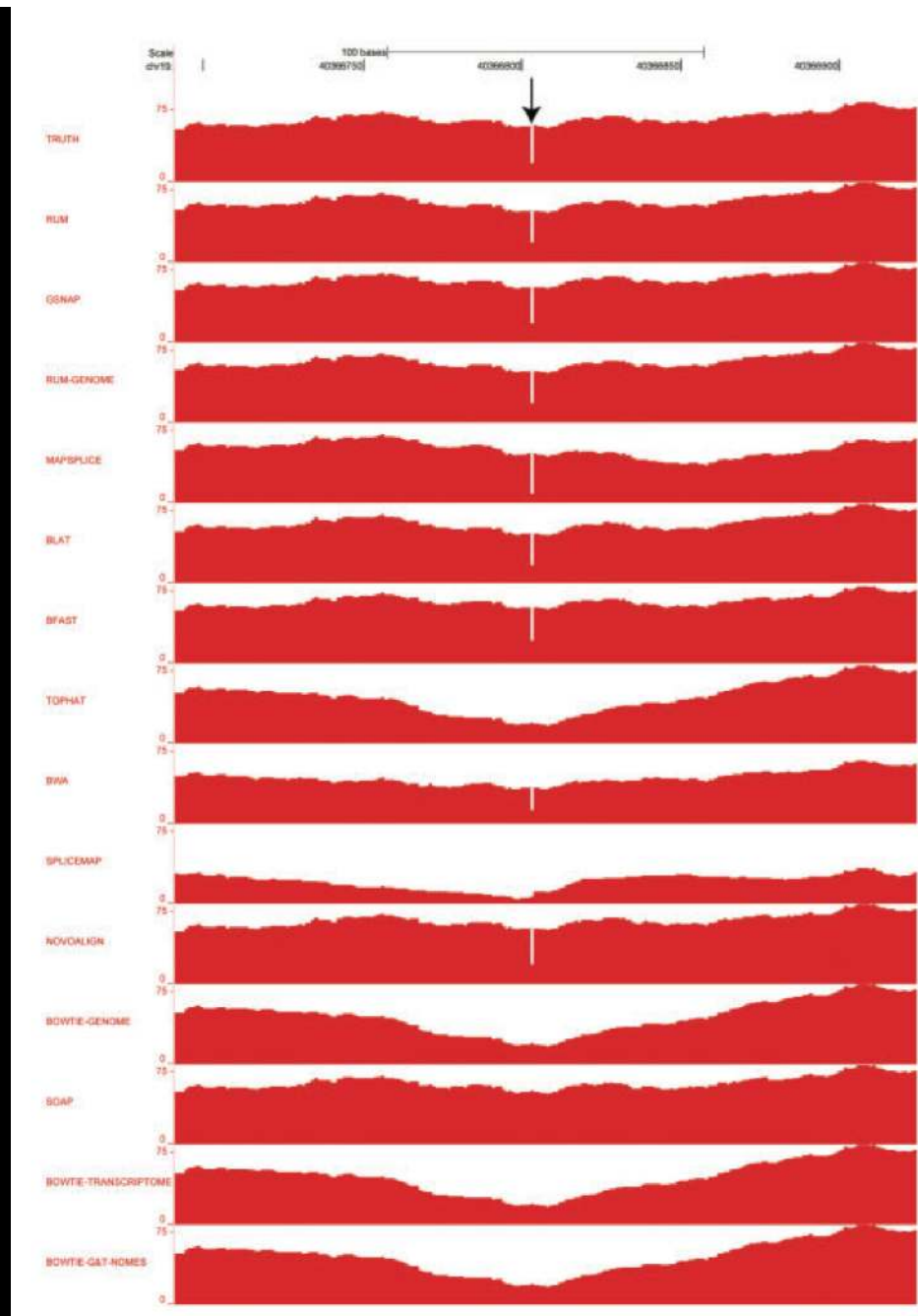
Nookaew et al. **A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in *Saccharomyces cerevisiae*.** *Nucleic Acids Research* 2012, **40**:10084–10097.

15 mapping results

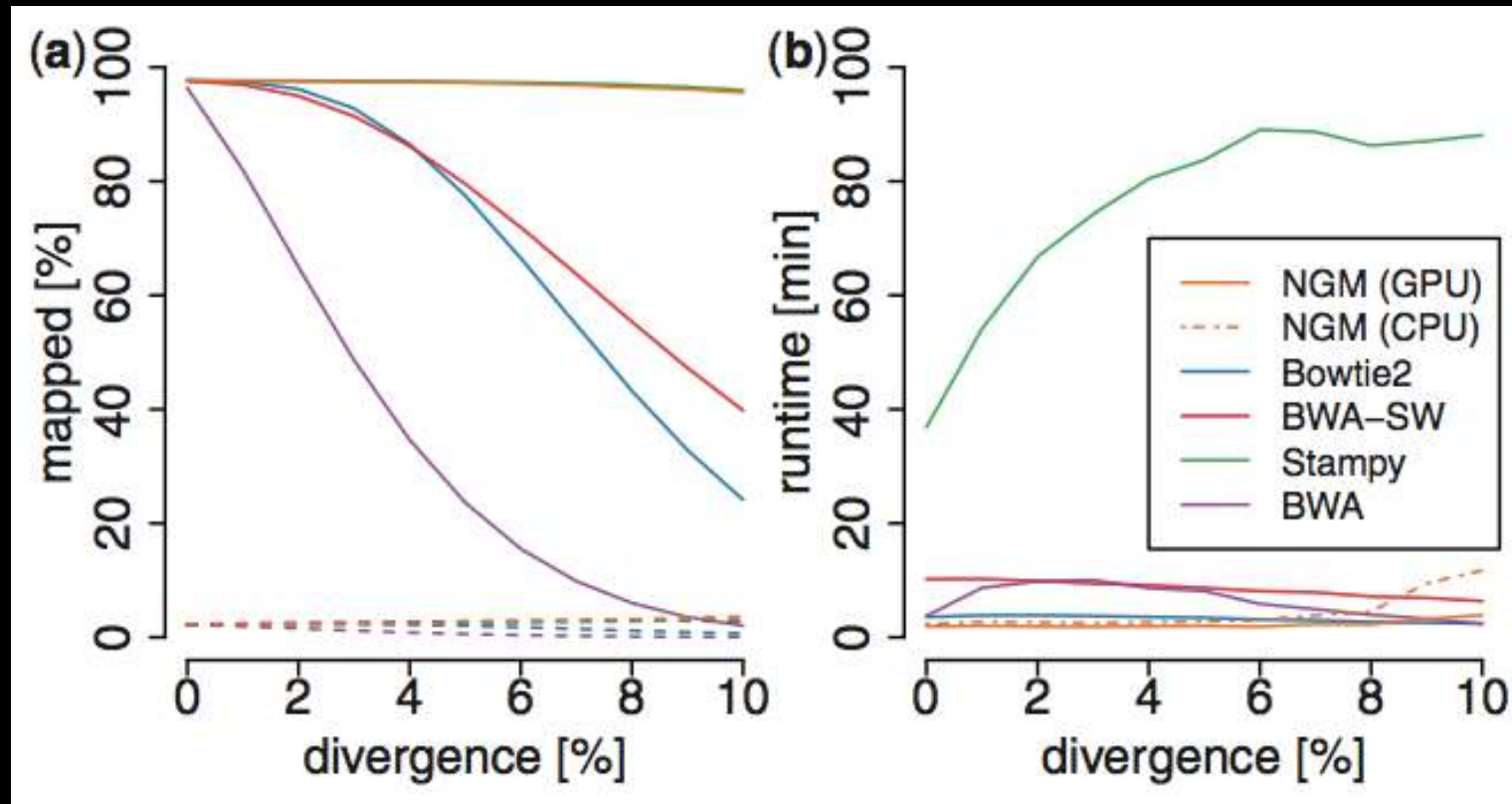
Dramatic differences in ability to handle a 2 bp insertion in reference compared to reads

TopHat, SpliceMap, Bowtie and Soap

- do not identify indels
- they fail to accurately align reads to these regions

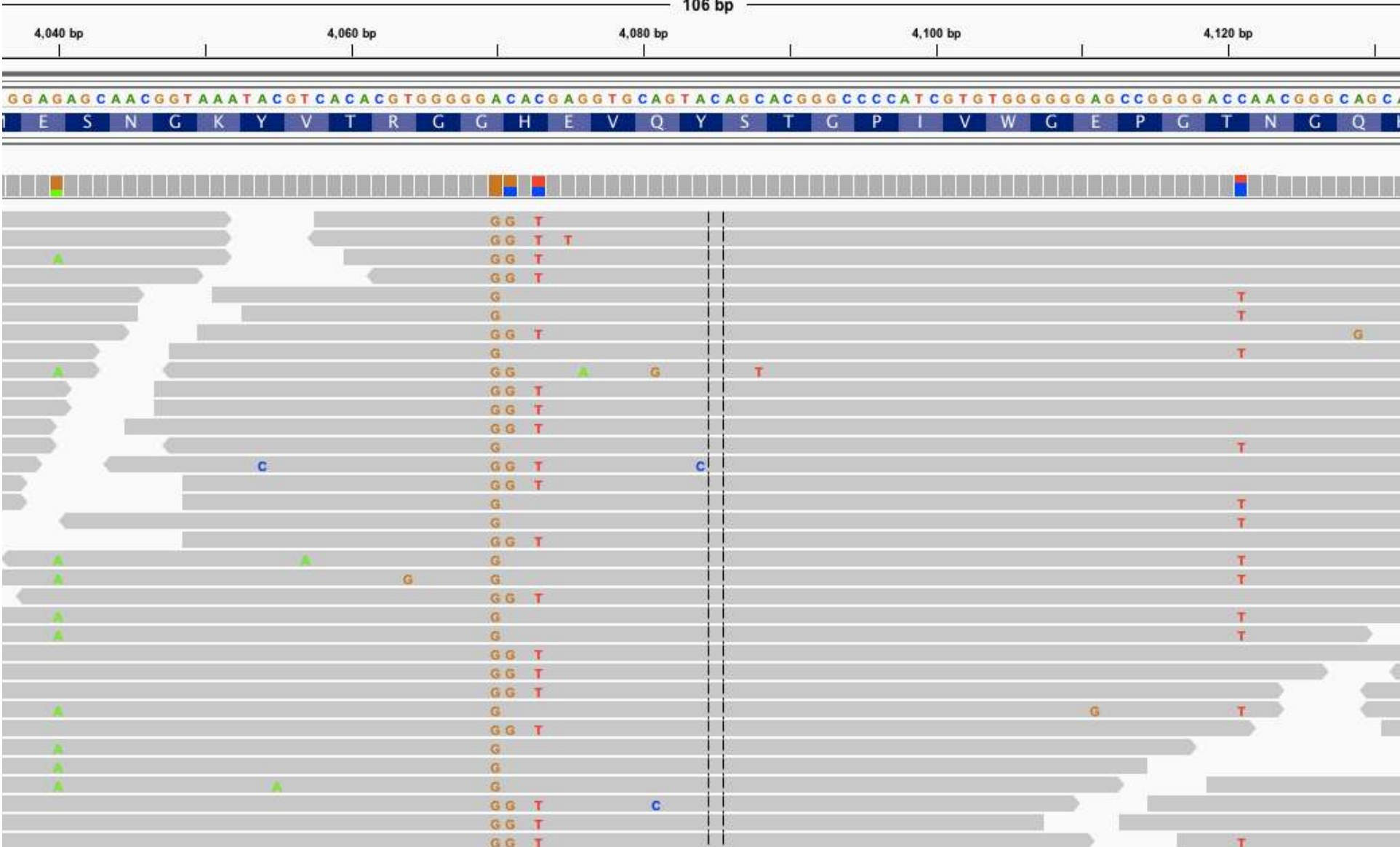


Allelic bias in read mapping



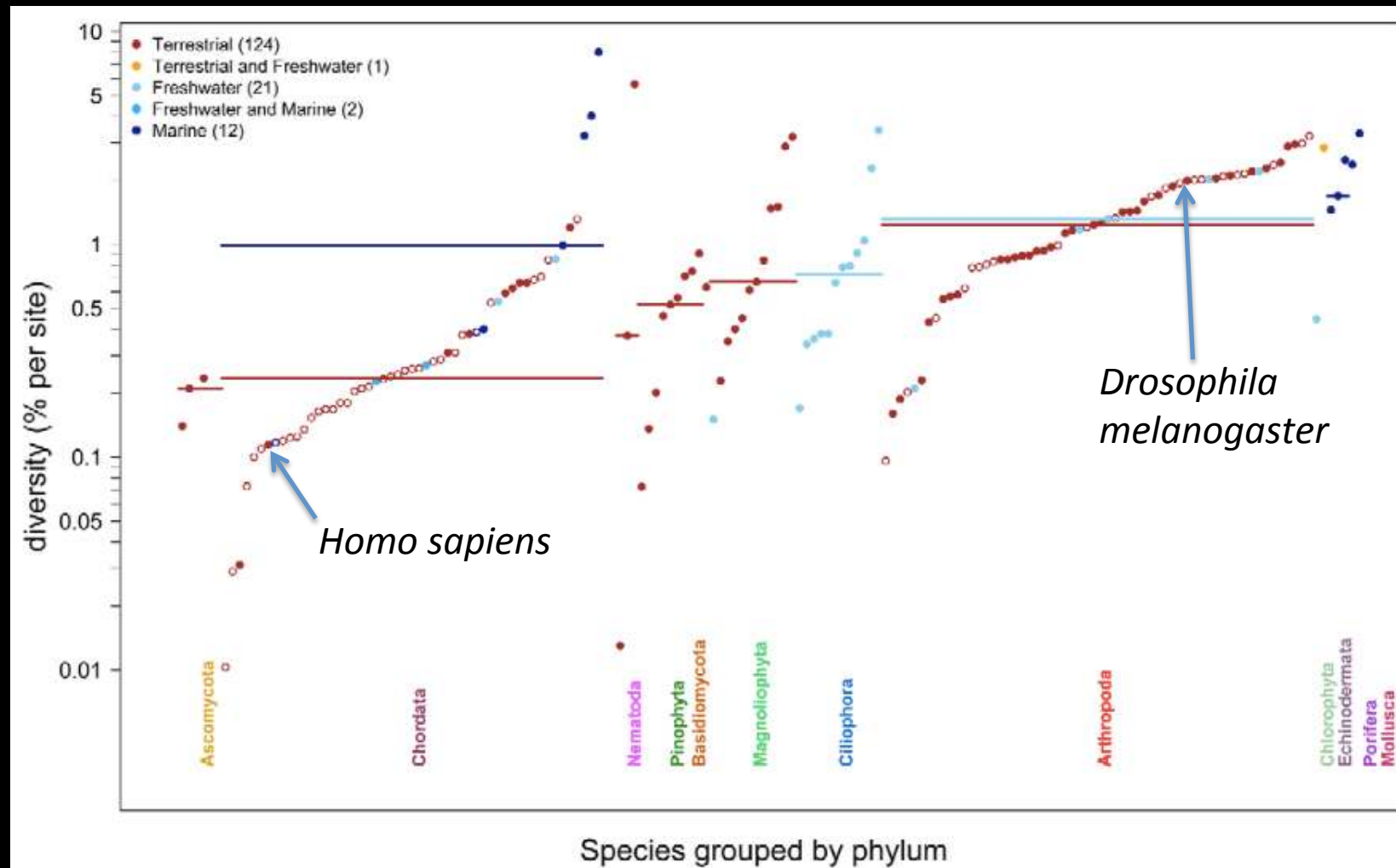
- Essentially identical to allele specific PCR bias ... but on a scale you can't detect unless you care to look
- Do your genes of interest have more than 3 SNPs / 100 bp?

100 bp window with 4 – 5 SNPs differing from reference

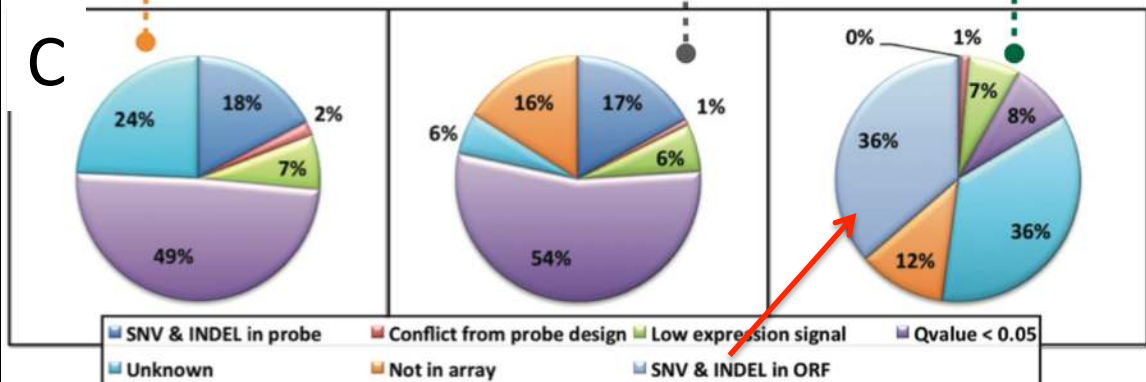
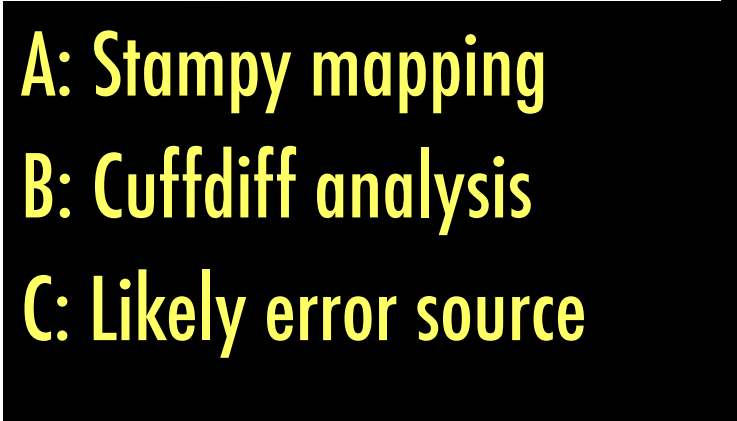


Mapping reads in outbred species

Average genome polymorphism levels (ignores indels)



Sig. expression differences by method



RNA-Seq



Real world example

2 factor analysis with family effects

Bicyclus anynana

**Save
energy,
live long**

**Live
fast,
die
young**

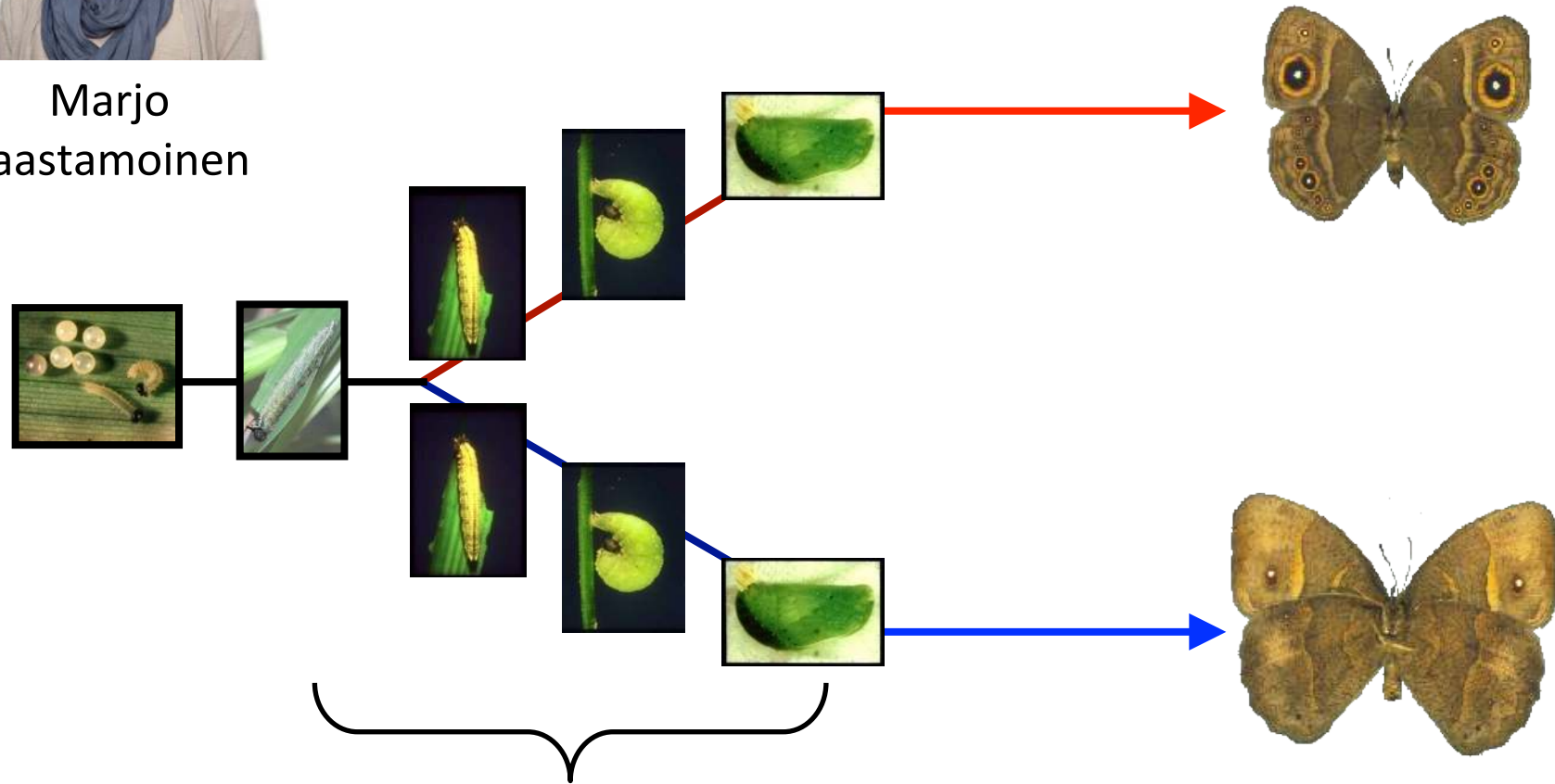


long	lifespan	short
delayed	reproduction	fast
inactive	behaviour	active
high	fat reserves	low
cryptic	wing pattern	conspicuous



Marjo
Saastamoinen

Bicyclus anynana

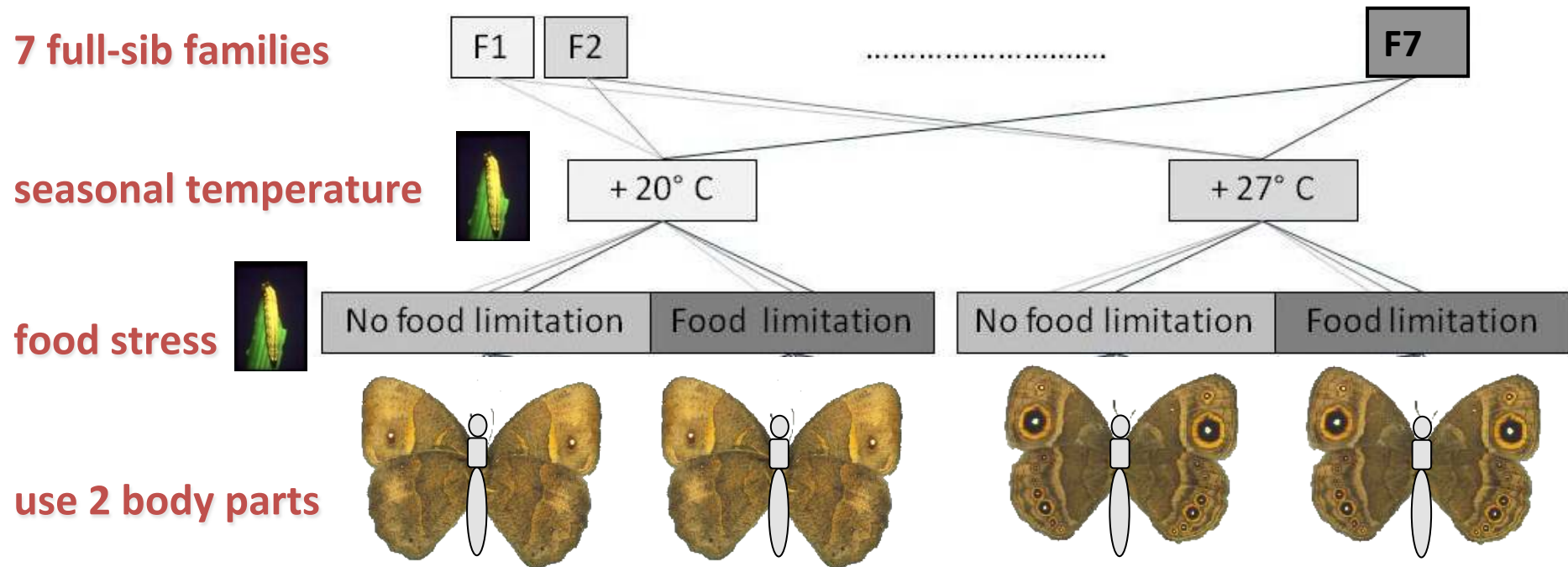


sensitive period

environmental
conditions

alternate
phenotypes

Experimental design



- 2 seasonal x 2 food stress x 2 body parts = **8 conditions**
- 7 families with $n = 2 - 3$ per condition → **144 RNA libraries**
- 10 million reads / library



Vicencio Oostra



body part	# libraries	# clean reads (per library)	# nucleotides (per library)	GC content
abdomen	72	15,261,019	3,052,203,767	45%
thorax	72	15,633,416	3,126,683,150	46%
total	144	2,224,399,290	444,879,858,000	45%



14 samples: one from each family, thorax and abdomen

69,075 contigs

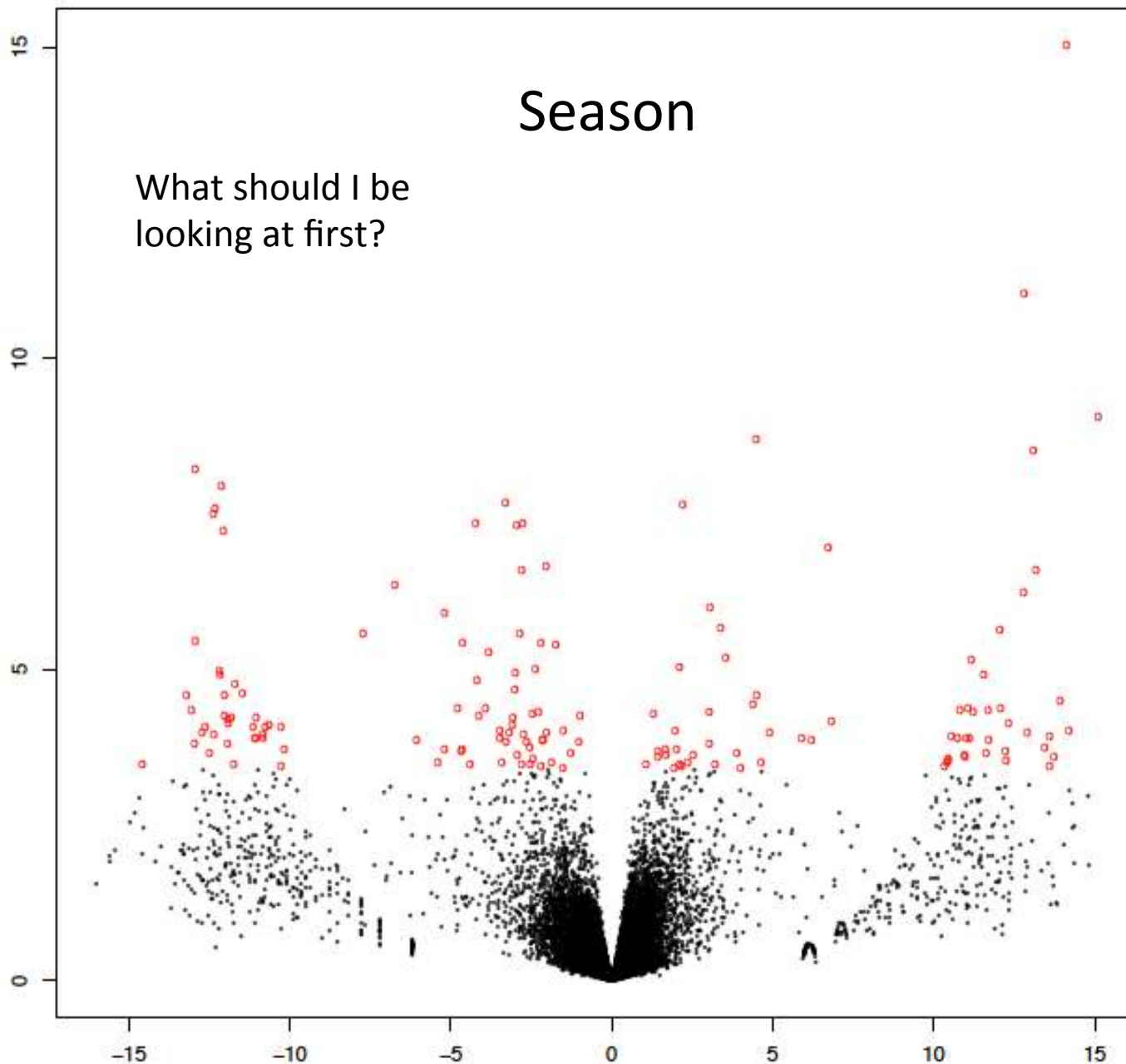
edgeR

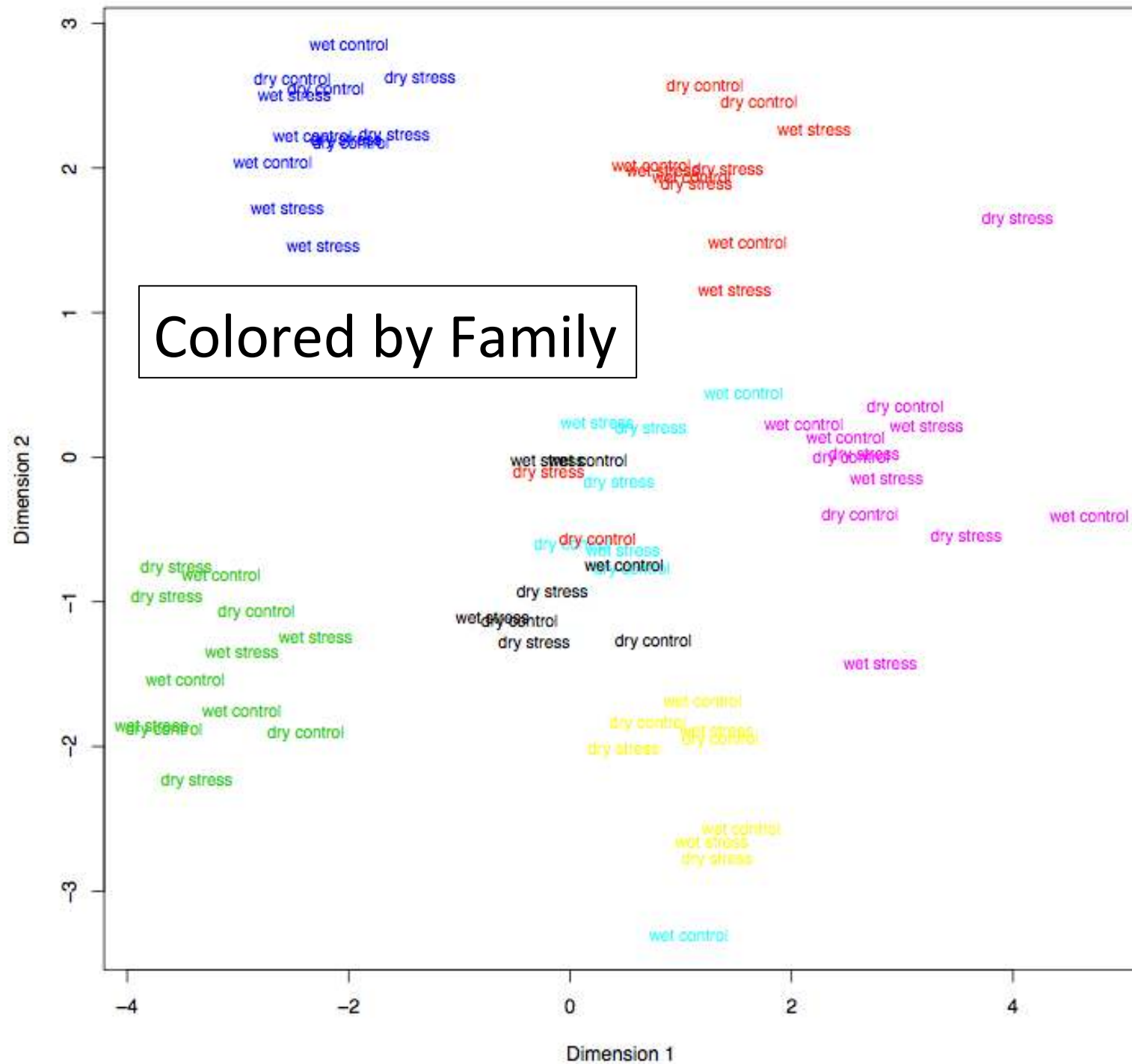


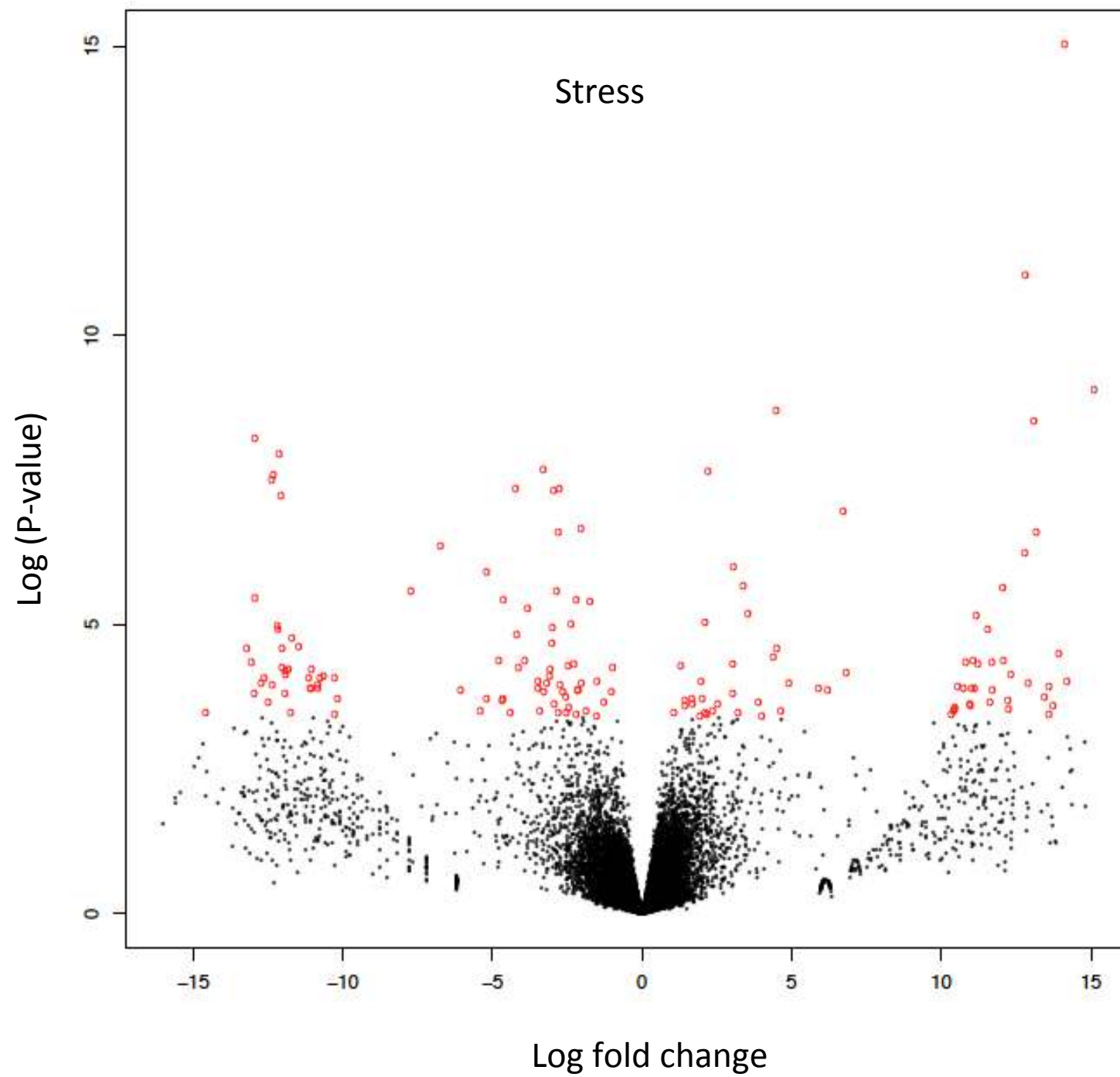
reads ~ season + stress + family +
season*stress + season*family + stress*family
season*stress*family

Season

What should I be
looking at first?





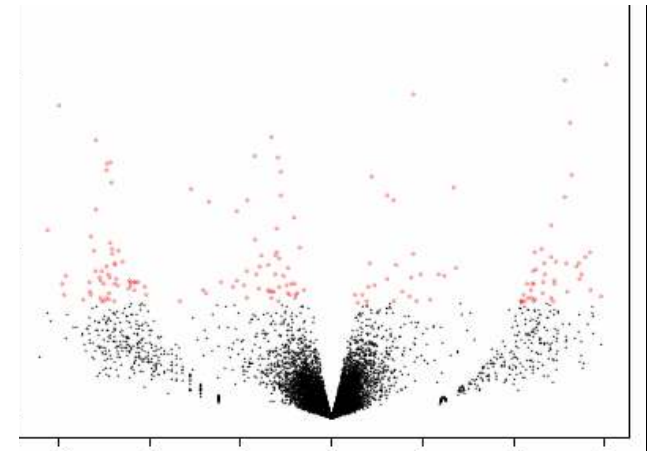


Log (P-value)

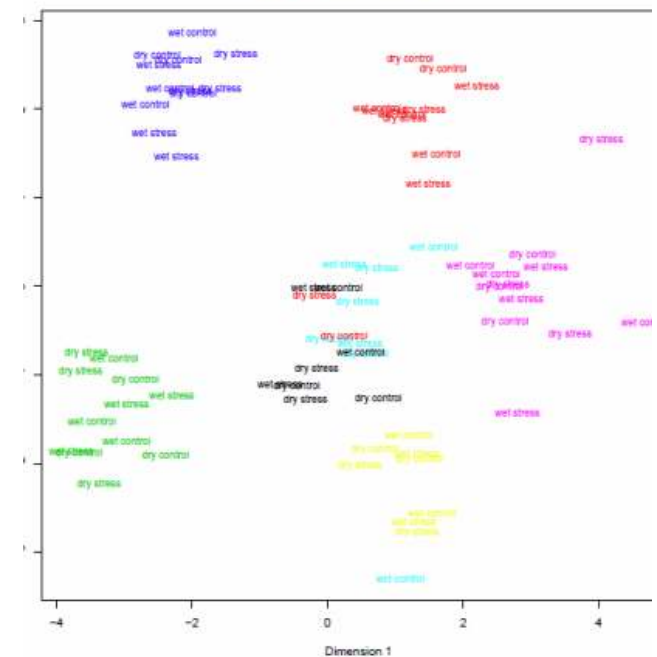


Log fold change

Effect of filtering the mapping to Trinity contigs

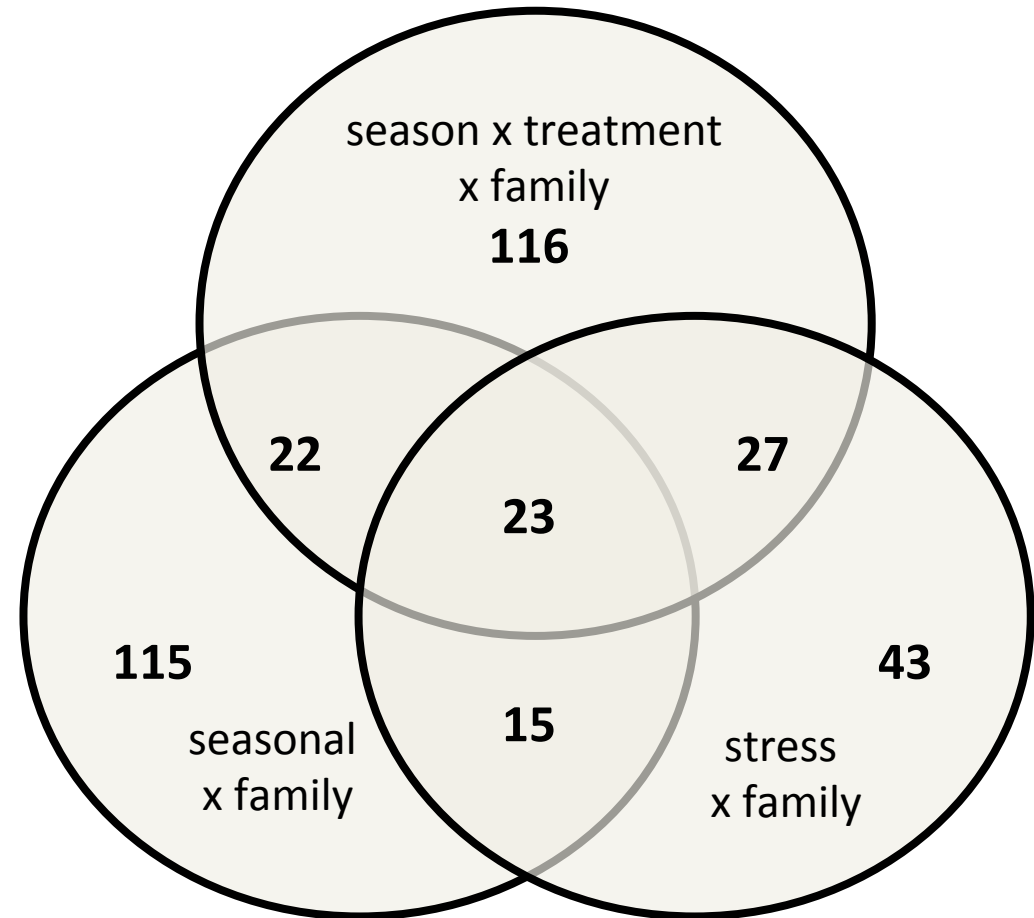


71 zero-read samples
allowed



GLM results

- Plastic responses:
 - Effects without any interaction with Family
- Genetic response:
 - Effects that have an interaction with family
 - Potential targets of natural selection



```
reads ~ season + stress + family + season*stress +  
season*family + stress*family + season*stress*family
```



100 My



Bombyx mori
Whole genome sequence,
predicted gene set

320 My



Drosophila melanogaster
Extensive genomic &
functional resources

Assembly 2.0
Contig_57178
Contig_6821
Contig_1004
Contig_20226
Contig_27720
Contig_5260
Contig_27110
Contig_27390
Contig_26901
Contig_4713
Contig_20081
Contig_9982
Contig_15387
Contig_25362
Contig_36071

Blastx

Bmori06 PepEd90
BGIBMGA002704
BGIBMGA003247
BGIBMGA003248
BGIBMGA003248
BGIBMGA003248
BGIBMGA003249
BGIBMGA004806
BGIBMGA004806
BGIBMGA004865
BGIBMGA004866
BGIBMGA005329
BGIBMGA006733
BGIBMGA008859
BGIBMGA008859
BGIBMGA008859

Blastp

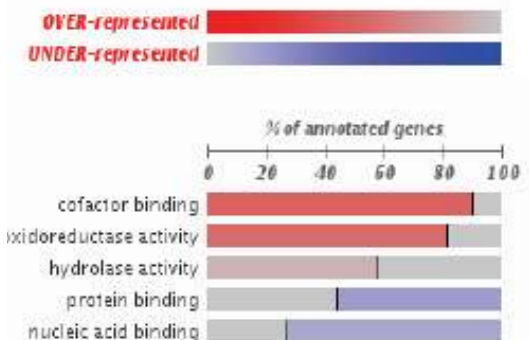
Flybase gene ID
CG33126
CG6519
CG6519
CG6519
CG6519
CG6519
CG33126
CG33126
CG33126
CG33126
CG3149
CG6783
CG4178
CG4178
CG4178

D. melanogaster
lacks an orthologous
reproductive
physiology



**Gene Set Enrichment analysis
using Gene Ontology database**

Fatiscan Analysis



Most studies are annotation limited

- What is the biological meaning of the top P-value genes?
- Low P-value or expression genes are certainly important
- Gene set enrichments are key to insights
 - Thus, annotation is very important

Description	Uniprot	-log10P
Oxidoreductase.	Q9VMH9	7.087008
Hypothetical protein.		6.993626
SD27140p.		6.315473
	Q8SXX2	6.300667
SD01790p.	Q95TI3	5.316371
Electron-transfer-flavoprotein	Q0KHZ6	5.1425
Pseudouridylate synthase.	Q9W282	4.784378
Hypothetical protein.	Q9VGX0	4.750469
CG14686-PA (RE68889p).	Q9VGX0	4.650051
Chromosome 11 SCAF14979, w	Q8T058	4.506043
		4.470413
, complete genome. (EC 1.6.5.5)		4.445501
RNA-binding protein.		4.374033
Hypothetical protein.	Q9VPL4	4.369727
Peptidoglycan recognition-like		4.206247
Angiotensin-converting-related	Q8SXX2	4.172776
Lachesin, putative.	Q9I7H7	4.056174
Secretory component.	Q9VVK5	3.981175
Putative adenosine deaminase	Q9VVK5	3.980728
		3.95787

7 of 20 (35%) no Uniprot ID

Sources of error

Transcriptome assembly can be huge source of bias:

- Fragmentation creates multiple contigs of same gene
- SNPs and alternative splicing generates more contigs
- 1 locus = frag. X SNPs X alt. splicing = many contigs

We can observe effects in expression analyses:

- Family effect mapping bias
- Pseudo-inflation in Gene Set Enrichment Analyses

Put the **BIO** in your informatics!!

Use independent analyses as 'controls' on accuracy
— What are your + and – controls?

	Analysis # 1	Analysis # 2	Analysis # 3
Mapper	TopHat2	STAR	?
Normalization	none	TMM	TMM
Analysis	PCA	RSEM	EDGER

Should independent methods converge?

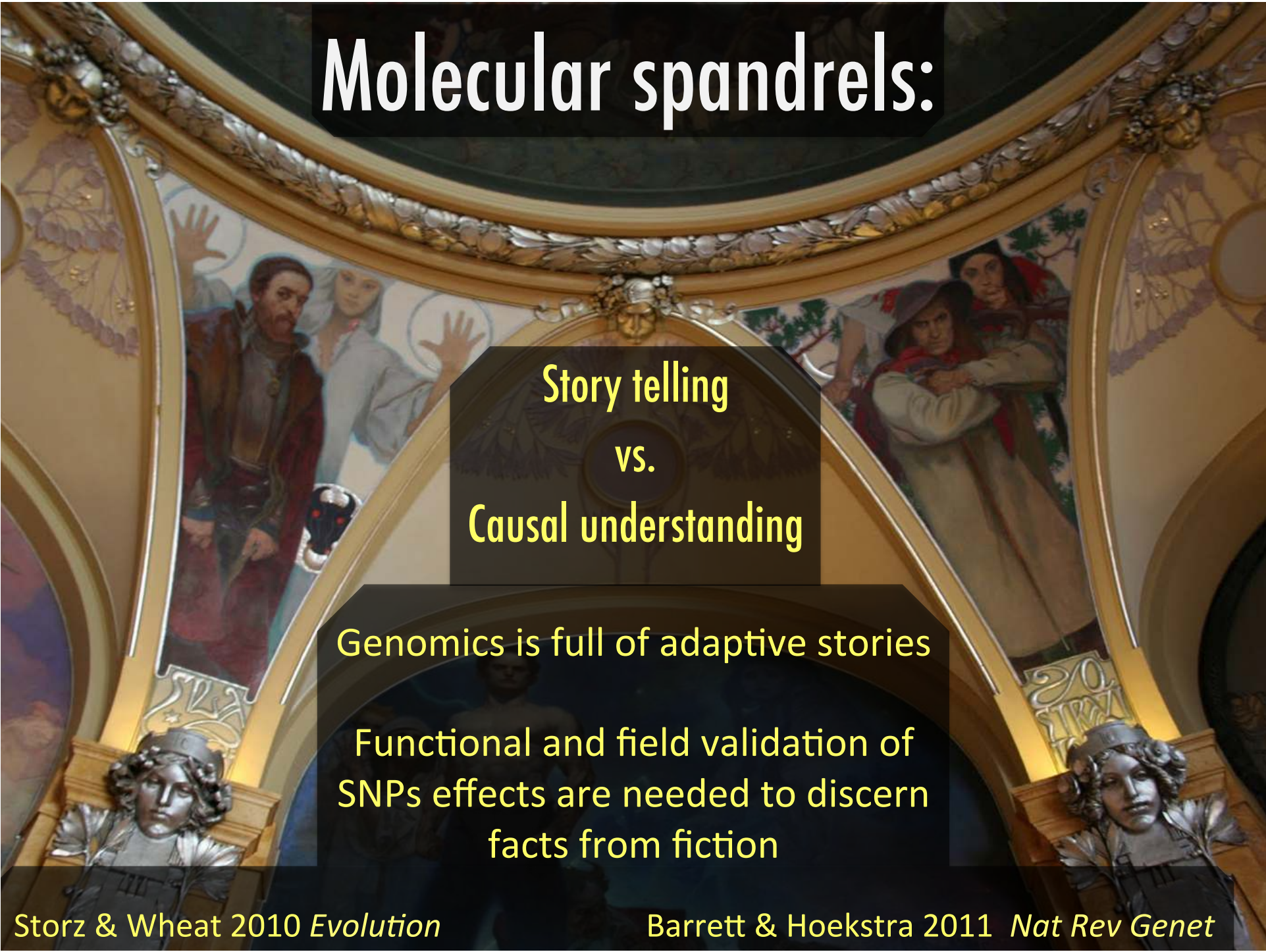
Interrogate your results

- “you need to be in charge of the analysis” – B. Cresko
- This will give you confidence
 - Bring freedom to your findings (no waterboarding)
- Graph your results – visualize the patterns
 - PCA or MDS plot
 - P-value distributions
- Assess gene copy number in gene set enrichment analyses (GSEA)
 - Do these levels fit to 1st principals expectations?
 - Do you have extra copies due to your Transcriptome assembly?

A major challenge for Ecological Genomics

- What causes natural selection in the wild?
 - How does genetic variation at one region of the genome interact with its environment (genomic, abiotic, and biotic)
- DNA alone can't tell us about selection dynamics in the wild
 - Molecular tests are very weak and uninformative about selection dynamics
- Research community is demanding actual demonstration of natural selection when making claims of adaptive role
 - Triangulate!!!!

Molecular spandrels:



Story telling
vs.
Causal understanding

Genomics is full of adaptive stories

Functional and field validation of
SNPs effects are needed to discern
facts from fiction

Storz & Wheat 2010 *Evolution*

Barrett & Hoekstra 2011 *Nat Rev Genet*

Ongoing work

- Currently trying to write commentary on biases in field
- Please send along other examples I might have missed
 - Feedback / critique is greatly appreciated



Karl Gotthard



Pararge aegeria



Peter Pruisscher



Ram Neethiraj

