Ecological Genomics, pt. 2

you, your data, your perception and the hard realities

Christopher West Wheat

Gordon Freeman

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
  - Assessing assemblies, mapping, and expression
  - What is validation?
- Insights from candidate genes
  - Can Second Gen methods get us there?

Core facilities and non-model species

Commonly heard statements that are not true:

- You can't do RNA-Seq without a genome
- The best metric for Transcriptome Assembly assessment is N50 & # of contigs
- We'll have your data back in < 1 month

Assessing transcriptome assembly

- Assessment metrics
  - Non-biological
    - N50, # of contigs
  - Biologically informative
    - # of orthologs identified
    - Ortholog hit ratio (OHR)

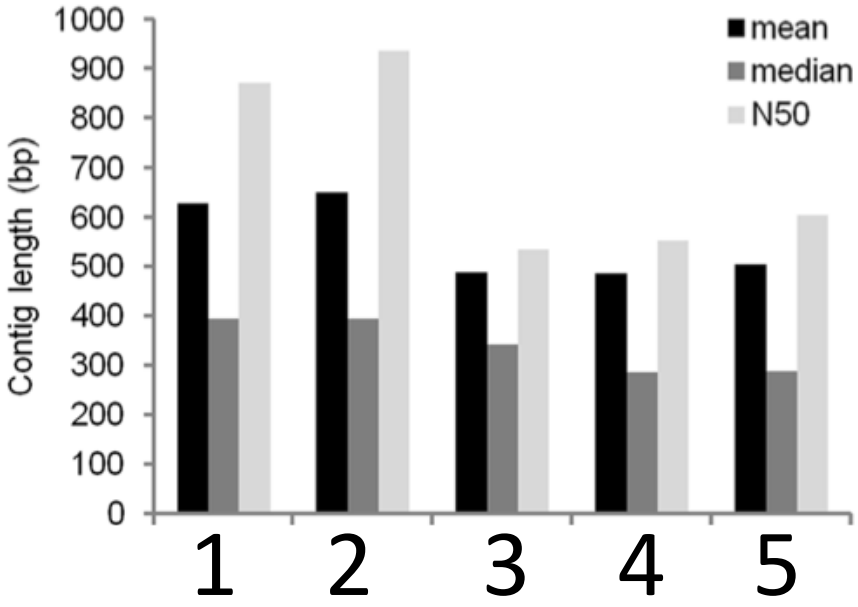
α / β :
1 = complete
< 1 = % covered

$$\alpha / \beta = \frac{\text{TA contig}}{\text{Ortholog}}$$

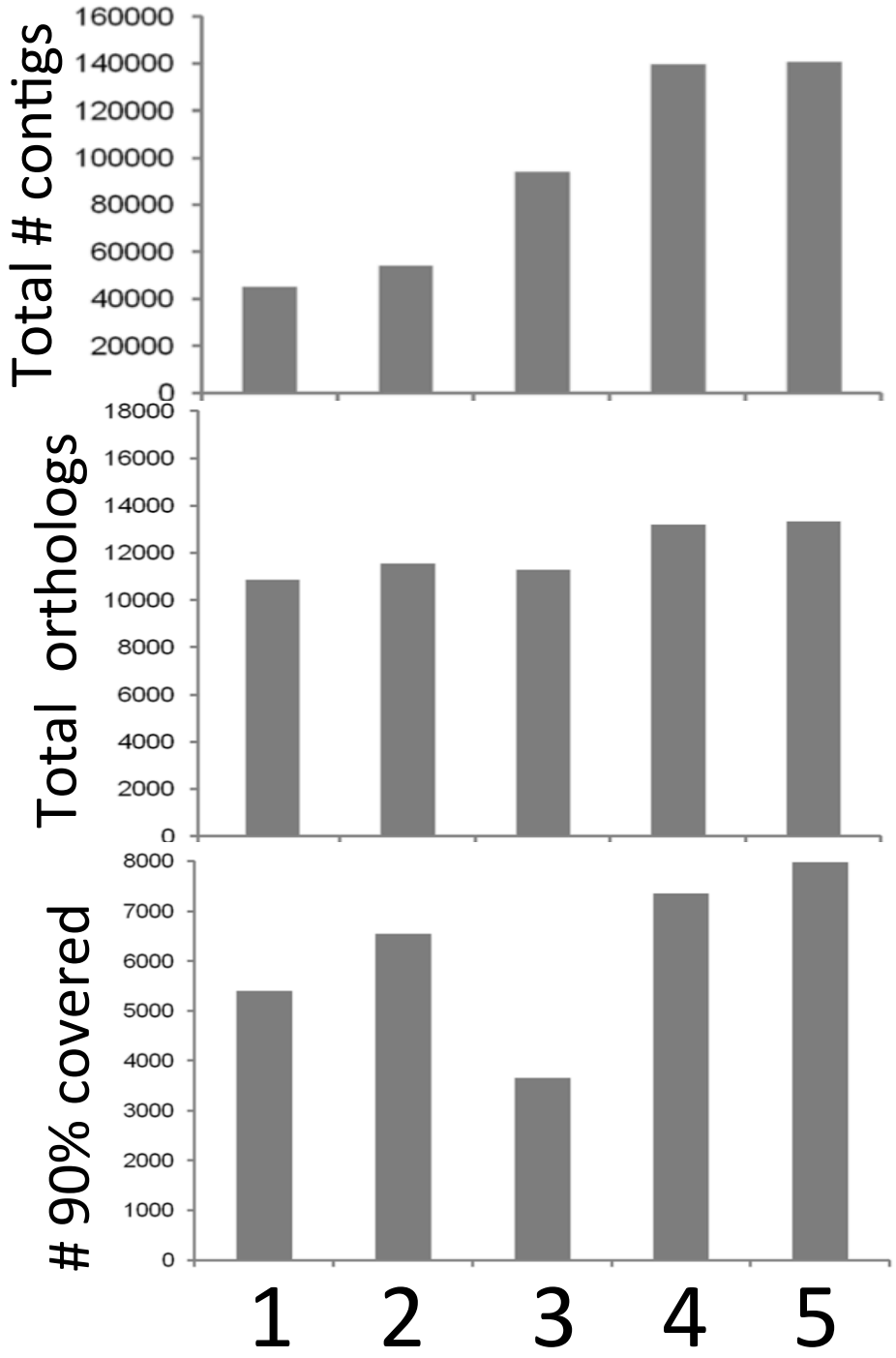
Length = α

Length = β

Hornett & Wheat 2012; O'neil & Emrich 2013 *BMC Genomics*



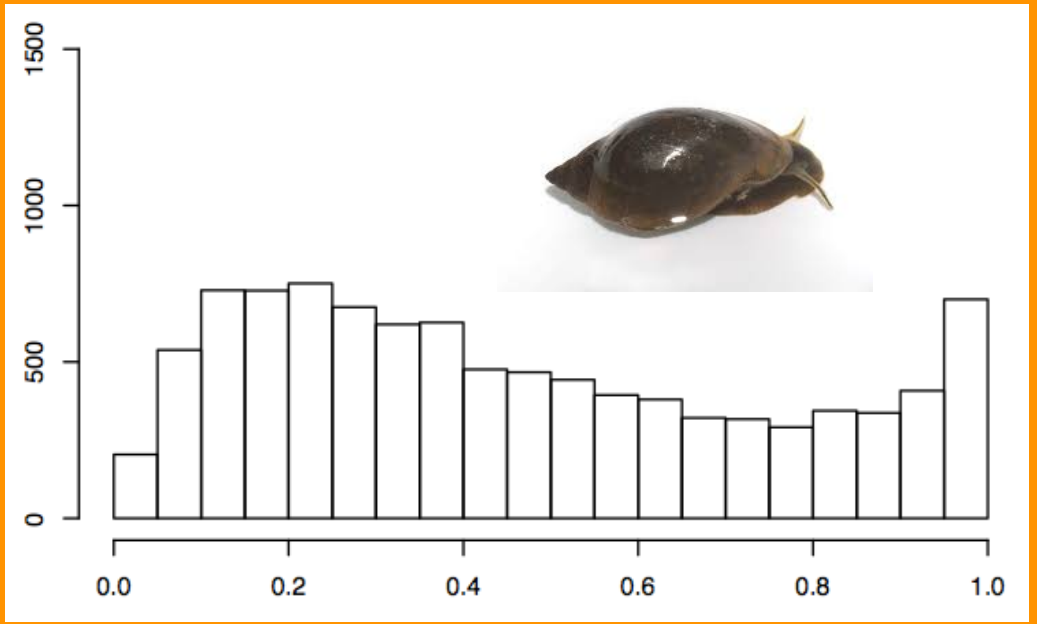
- 5 different TAs
- TA 2
  - Best N50, fewest contigs

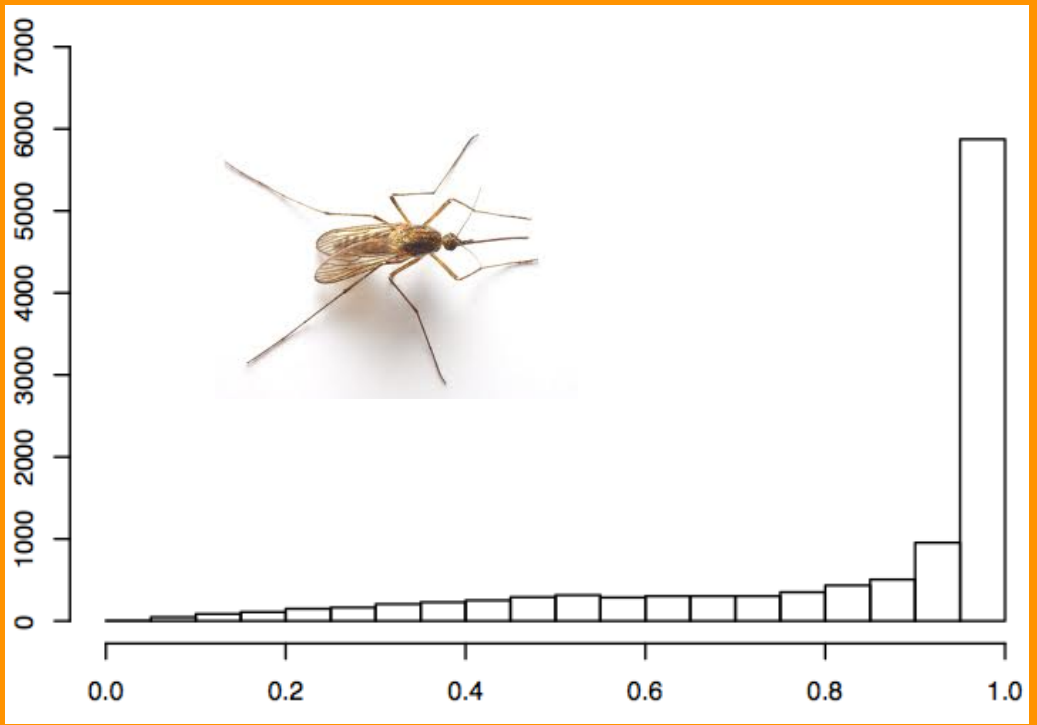


Hornett & Wheat 2012 BMC Genomics

OHR graphs

- Shows the number of unique orthologs hit
- Distribution of their reconstructed length

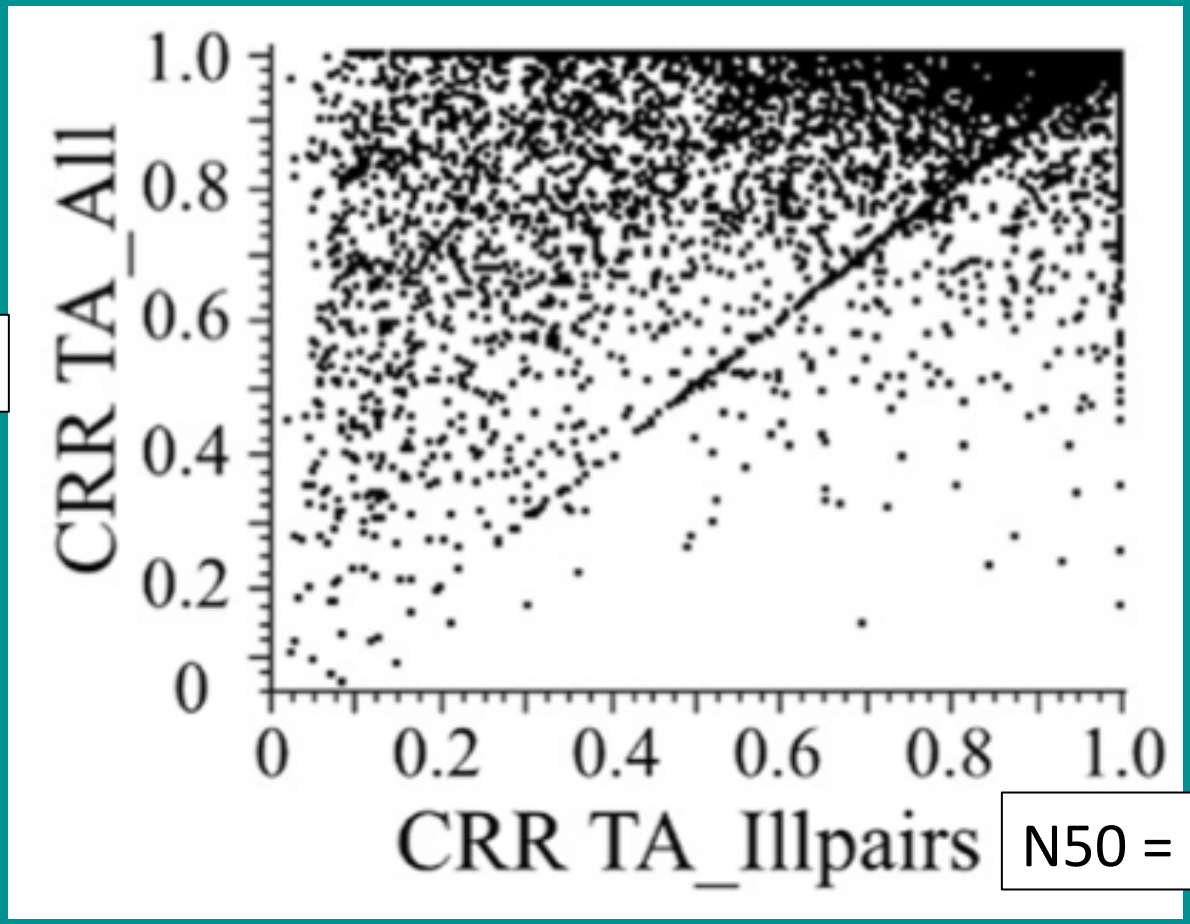




Comparative OHR

- Compare longest contig per ortholog for two assemblies
- Plot them against each other

N50 = 610

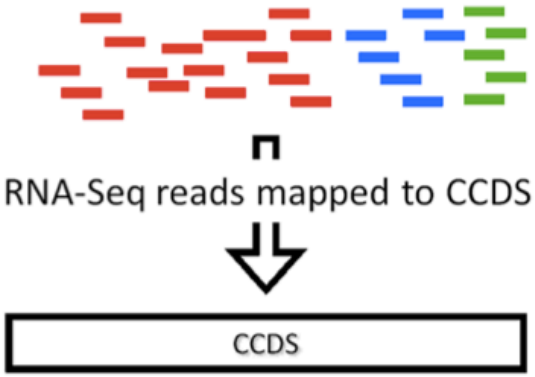


N50 = 930

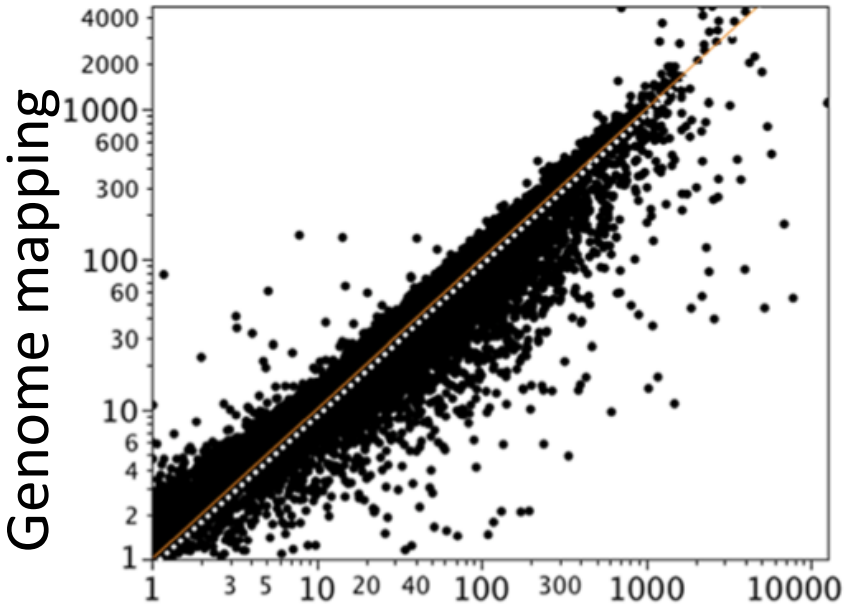
Hornett & Wheat 2012 BMC Genomics

RNA-Seq mapping: comparing genome vs. TA

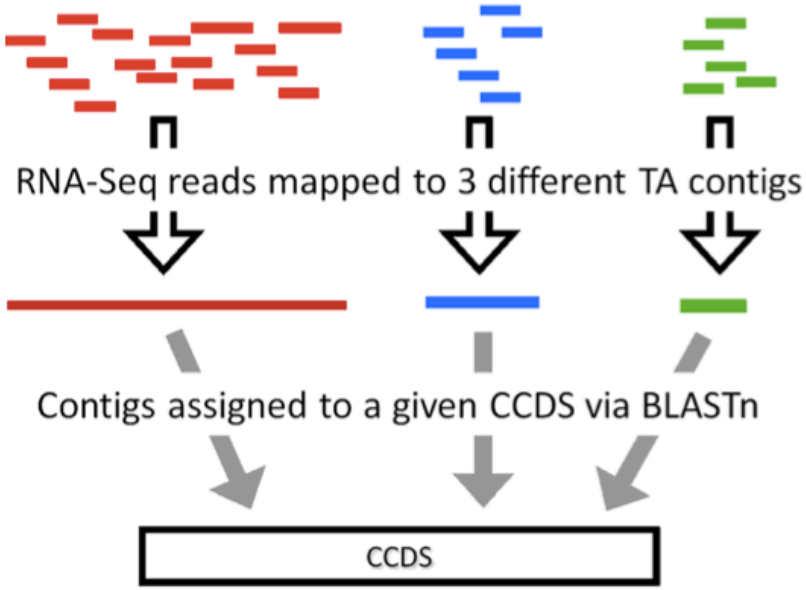
Genome mapping



Spearman's ρ = 0.95, P < 0.0001



Summed TA mapping

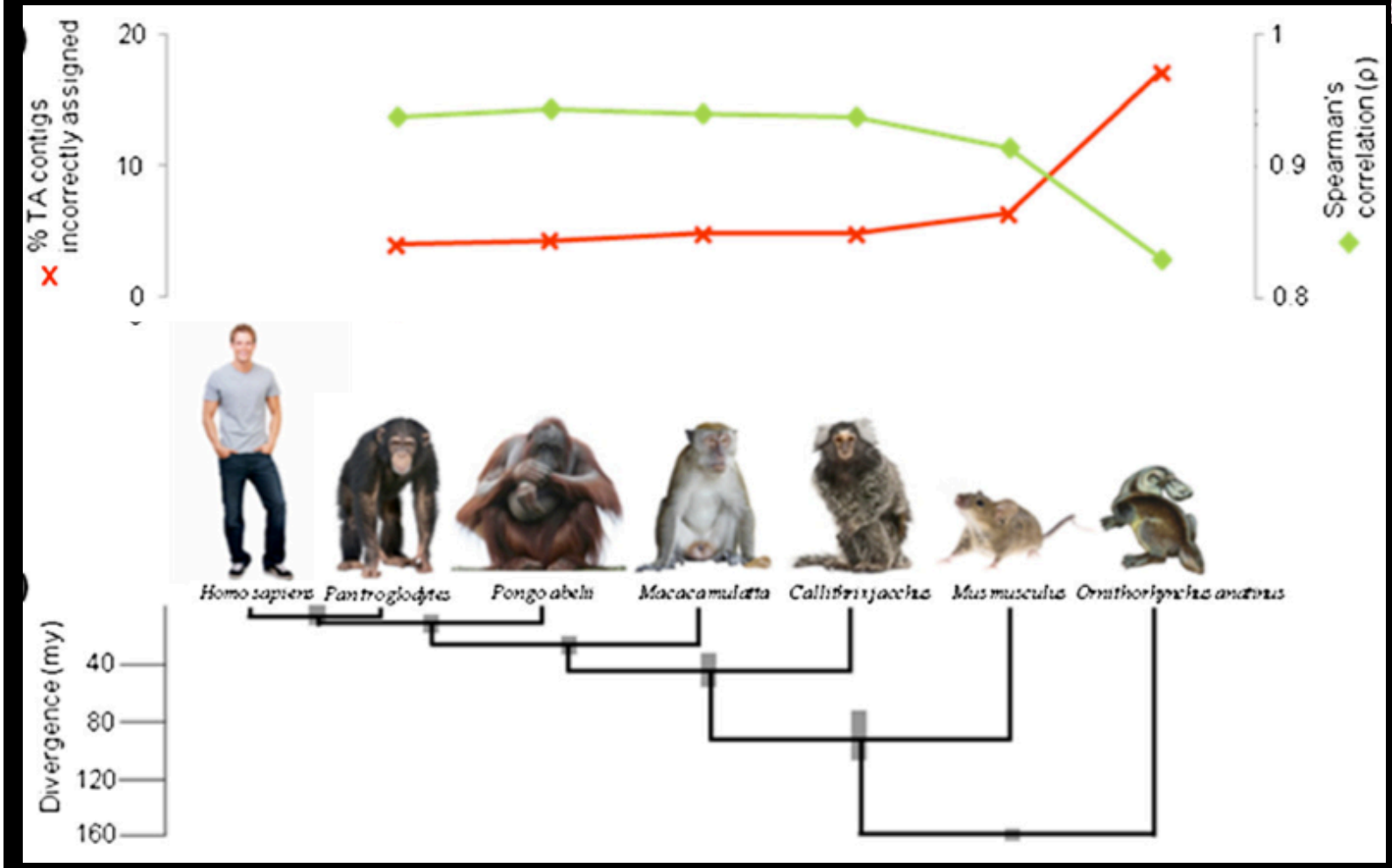


You can generate high quality data without a genome, for much of the transcriptome

Hornett & Wheat 2012 BMC Genomics

OHR can be calculated using predicted genes from divergent species

Emily Hornett

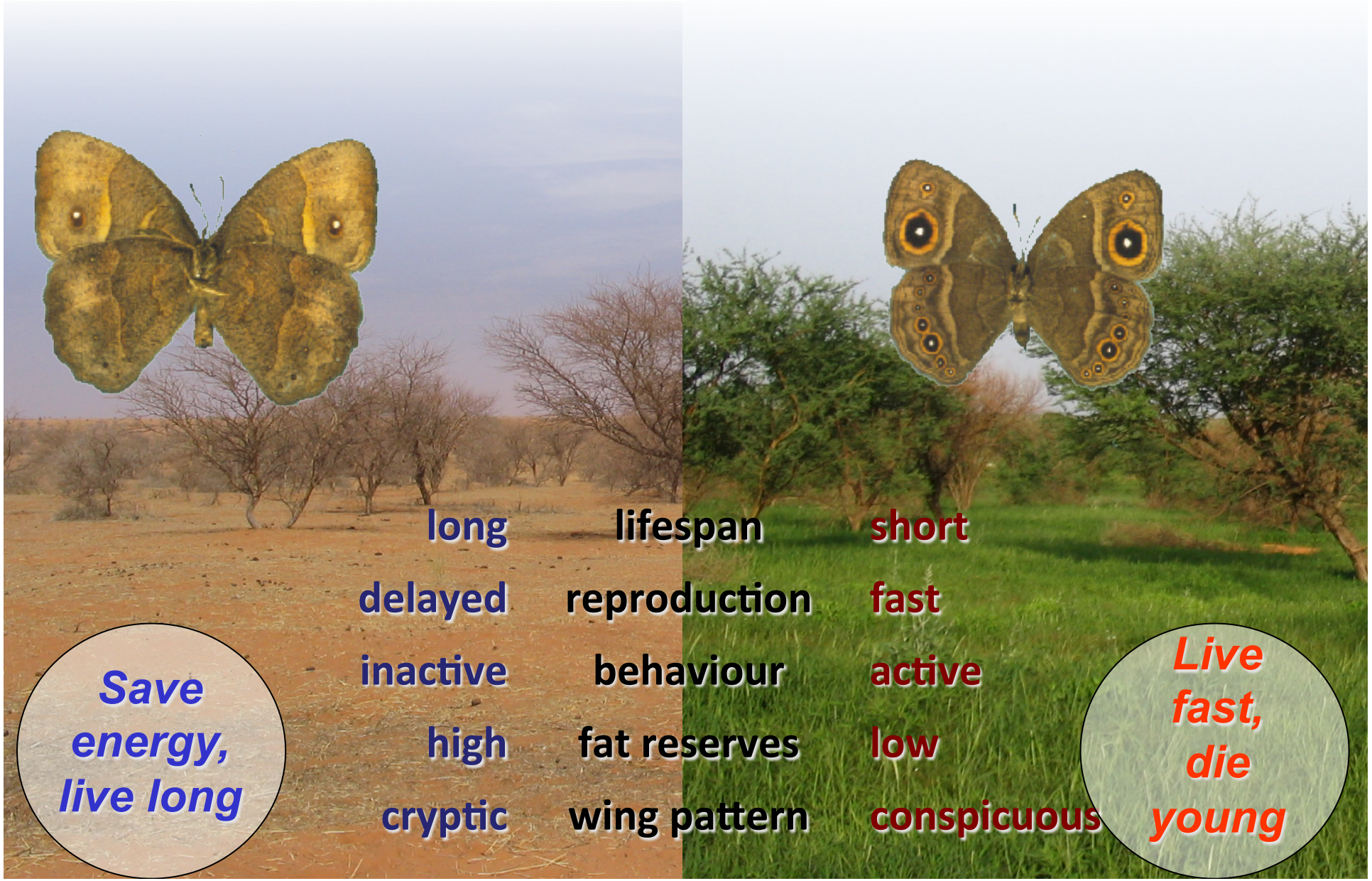


Hornett and Wheat 2012

RNA-Seq

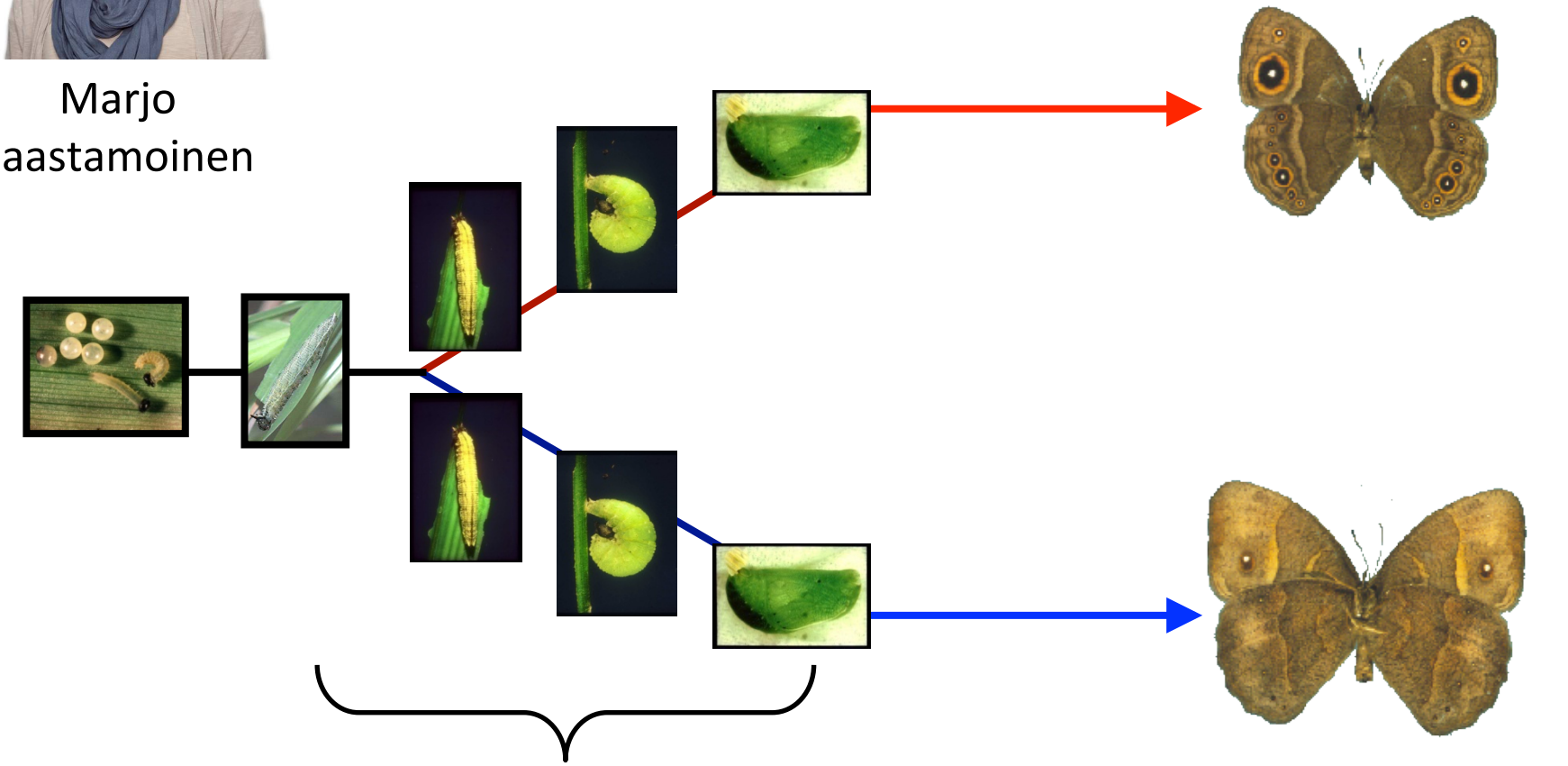
- Now we have a good assembly
- Ready for quantitative gene expression analysis
- 2 factor analysis with family effects

Bicyclus anynana



Bicyclus anynana

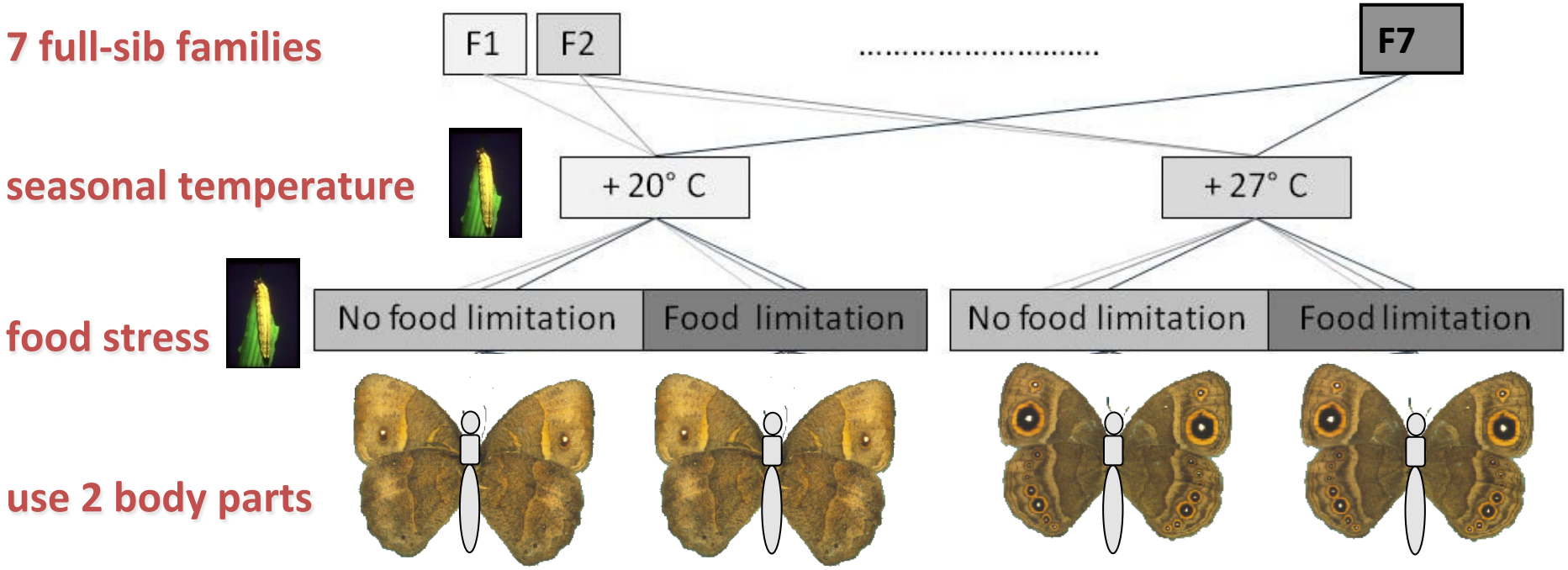
Marjo Saastamoinen



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

华大基因 BGI

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%

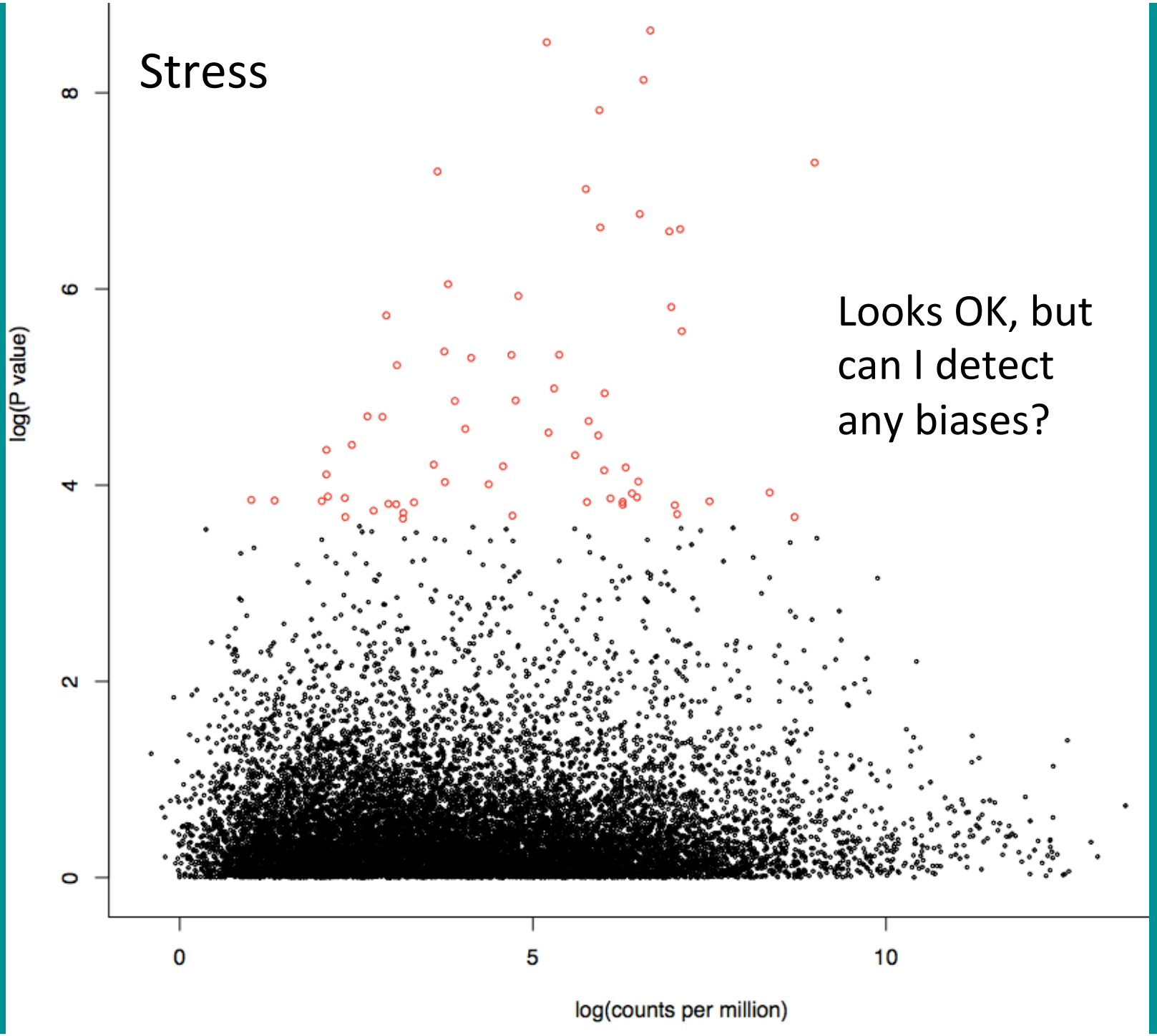
14 samples: one from each family, thorax and abdomen

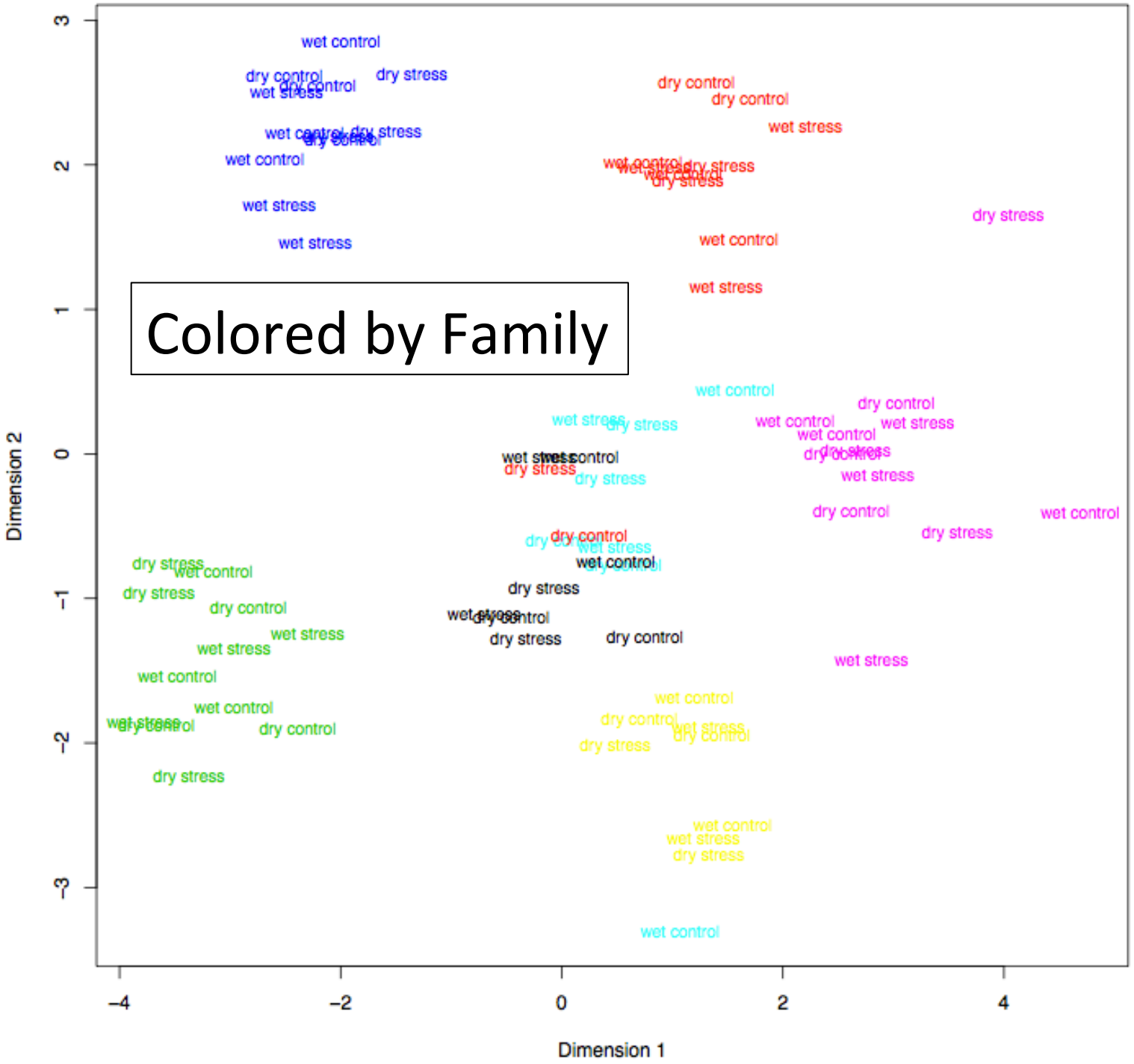
69,075 contigs

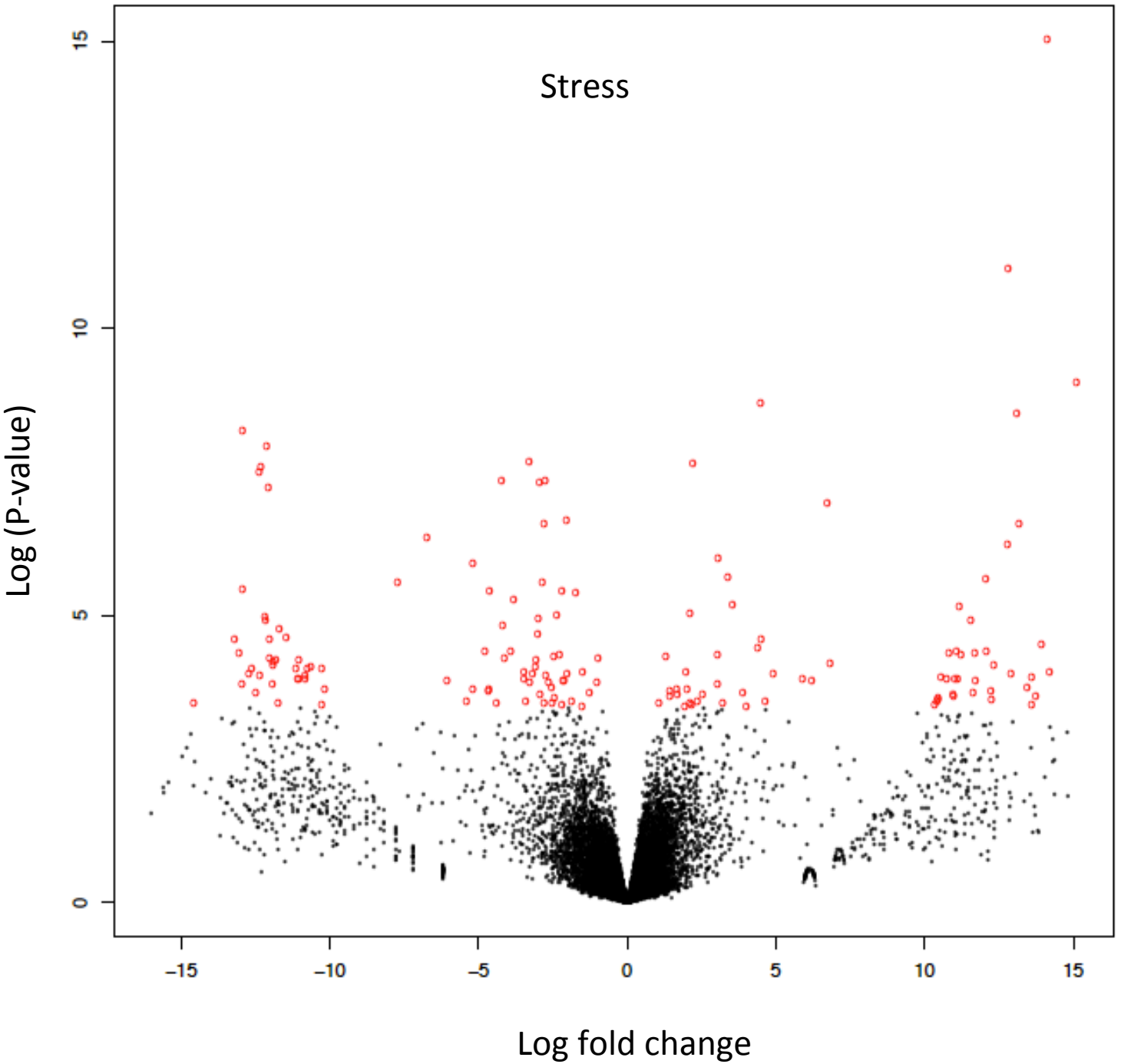
edgeR

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

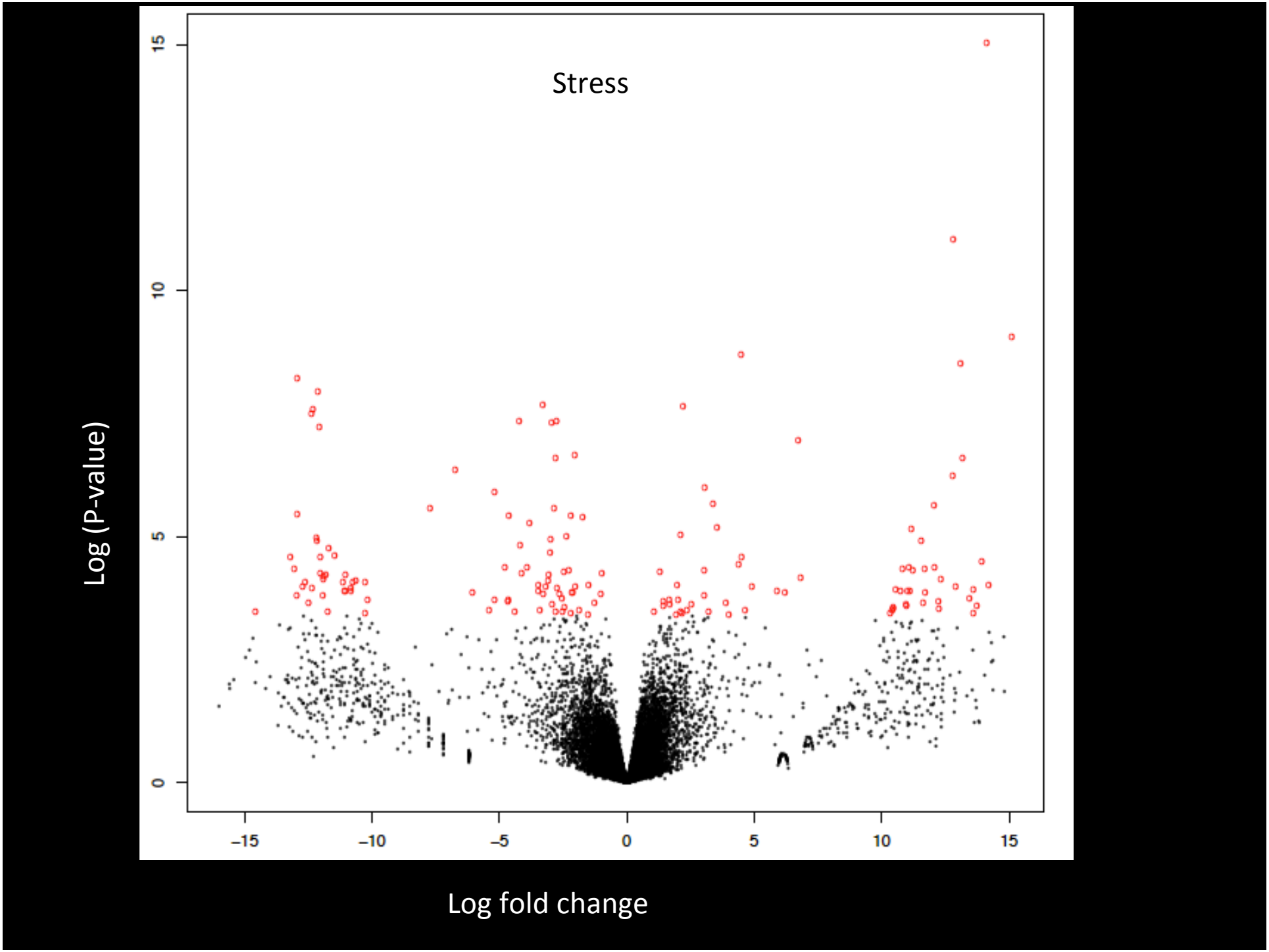
```
# reads ~   season + stress + family +
          season*stress + season*family + stress*family
          season*stress*family
```


Stress
log(P value)
0
2
4
6
8
0
5
10
log(counts per million)
Looks OK, but
can I detect
any biases?

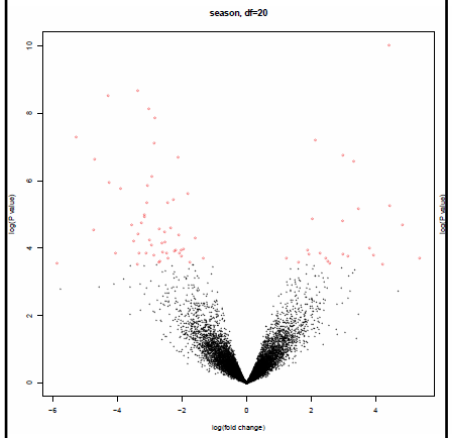
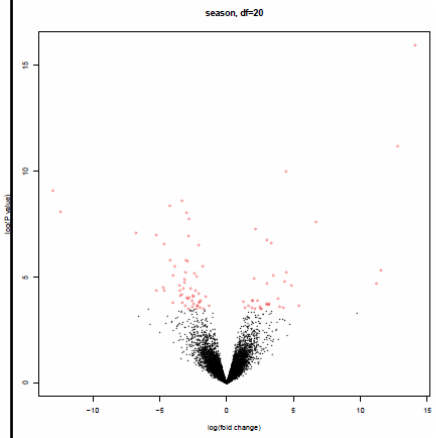
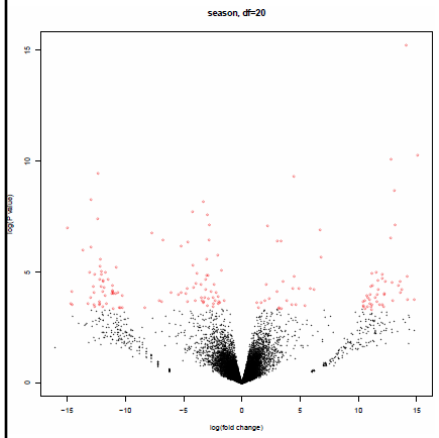
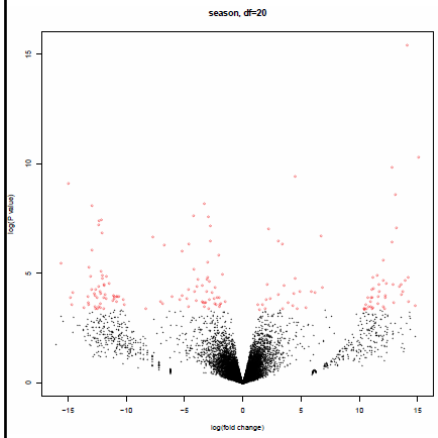
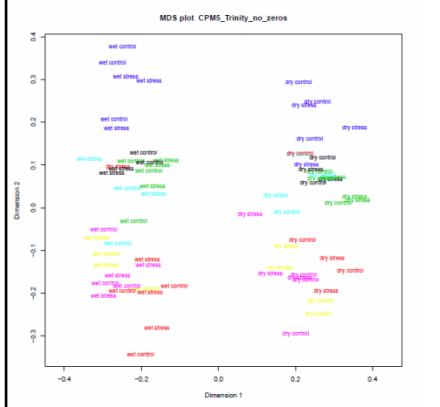
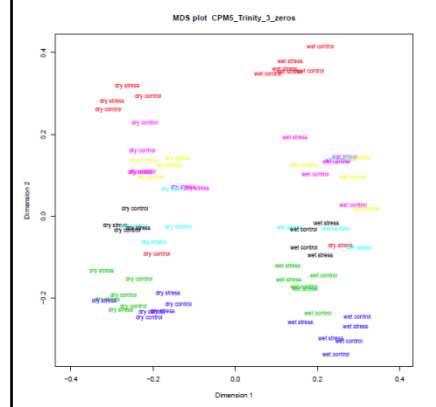
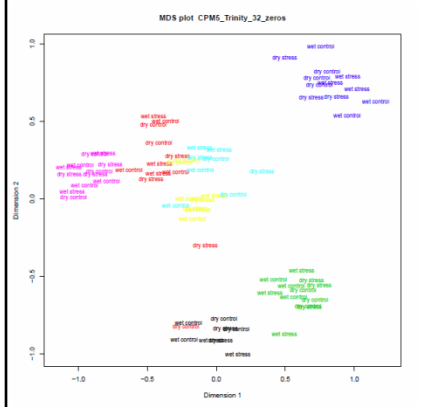
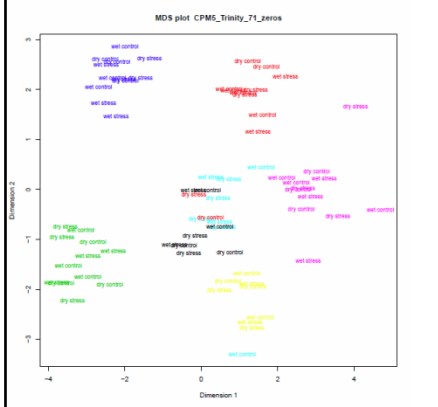

Colored by Family
Dimension 1
Dimension 2
−4
−2
0
2
4
3
2
1
0
−1
−2
−3
wet control
dry control
dry stress
wet stress


Stress
Log (P-value)
15
10
5
0
−15
−10
−5
0
5
10
15
Log fold change

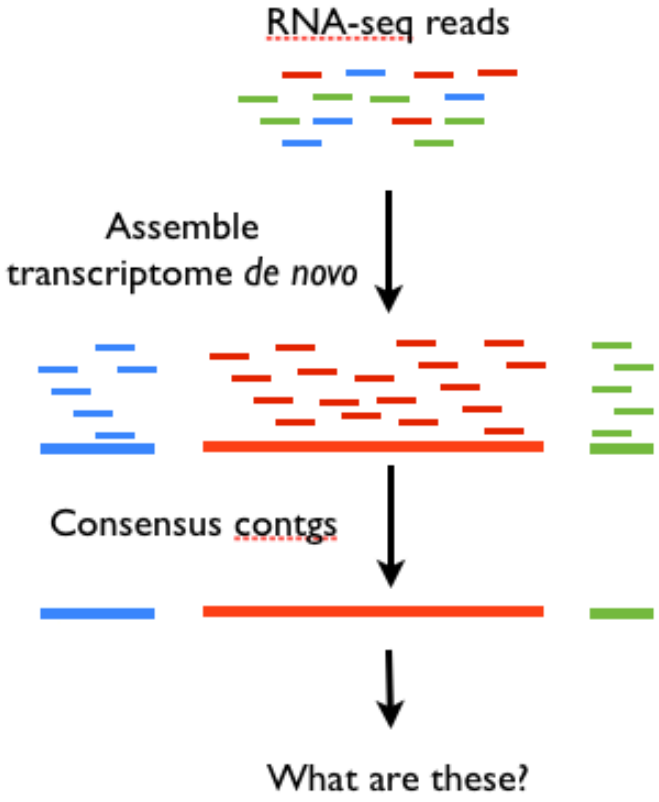

Log (P-value)
Log fold change

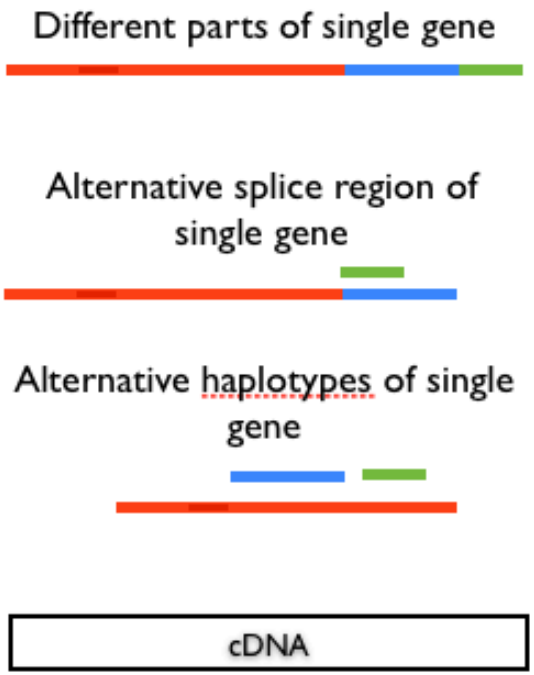

Stress
Log (P-value)
Log fold change
15
10
5
0
−15
−10
−5
0
5
10
15

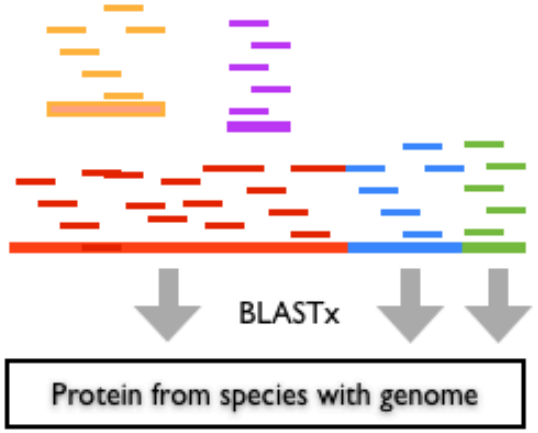
Effect of filtering, mapping to Trinity contigs

	0 zero-read samples allowed	3 zero-read samples allowed	32 zero-read samples allowed	71 zero-read samples allowed
				
				

What's happening?



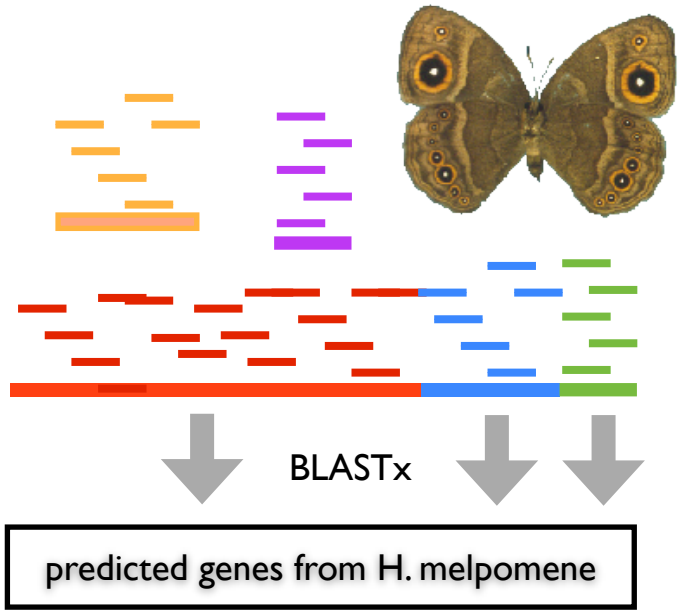


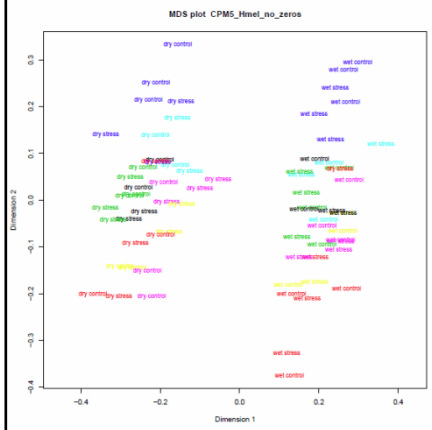
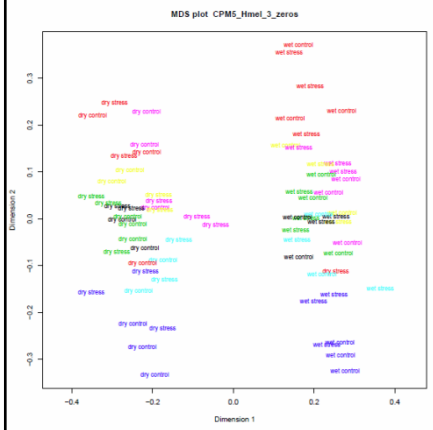
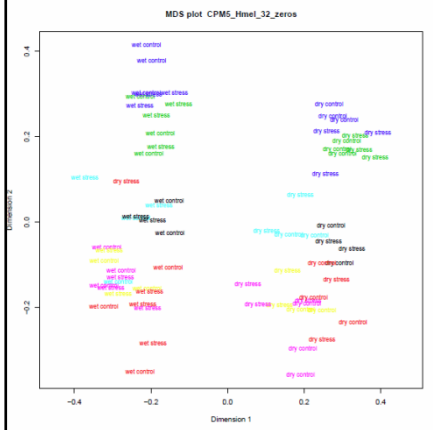
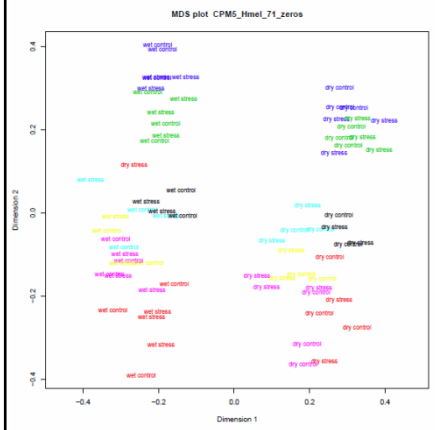


solution:
take the sum

Separate contigs made during assembly: SNPs X splicing
Creates bias in expression pattern, with large family effect
Summing by ortholog corrects this bias

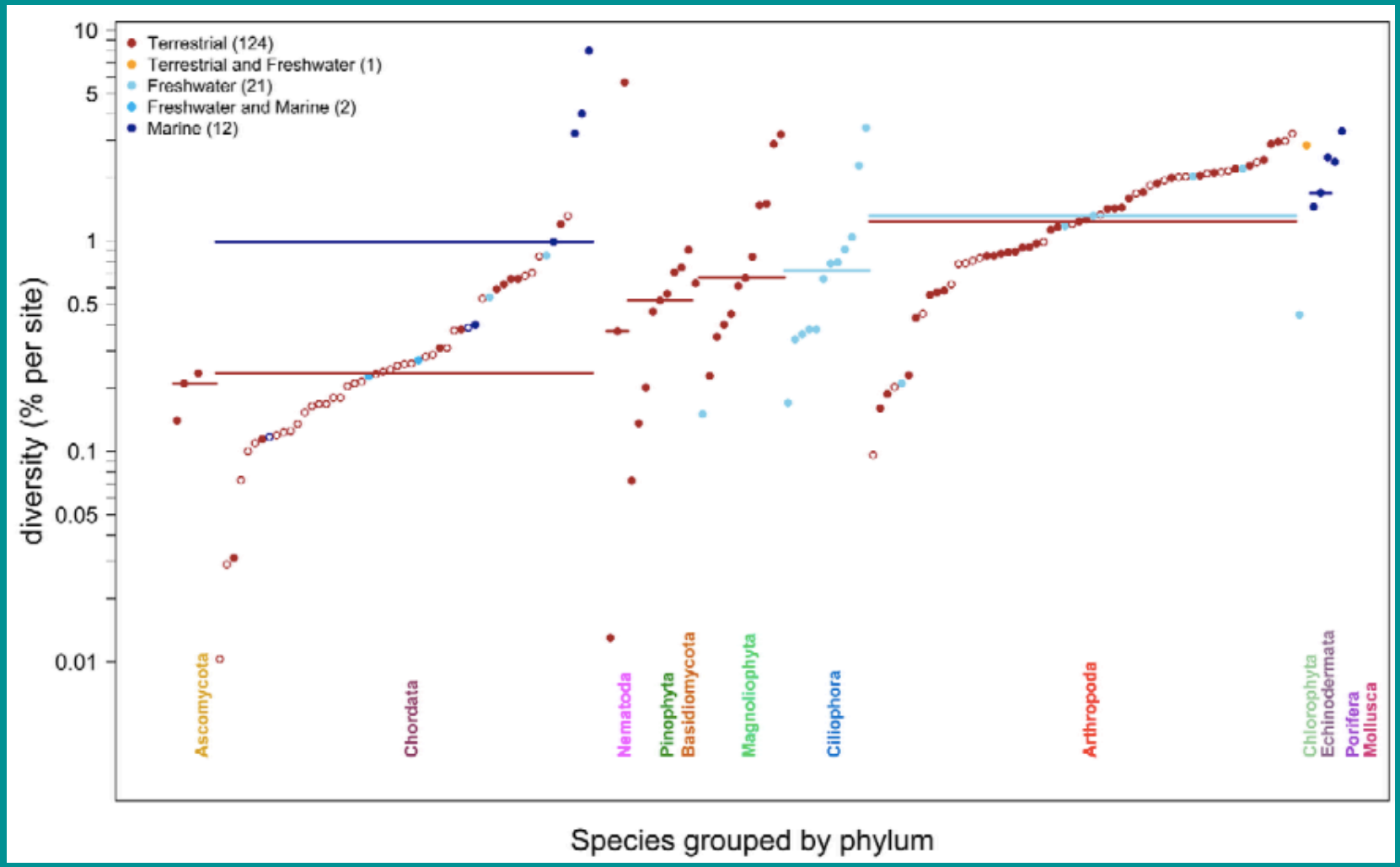
Effect of filtering when using sum method



	0 zero-read samples allowed	3 zero-read samples allowed	32 zero-read samples allowed	71 zero-read samples allowed
CPM > 5				

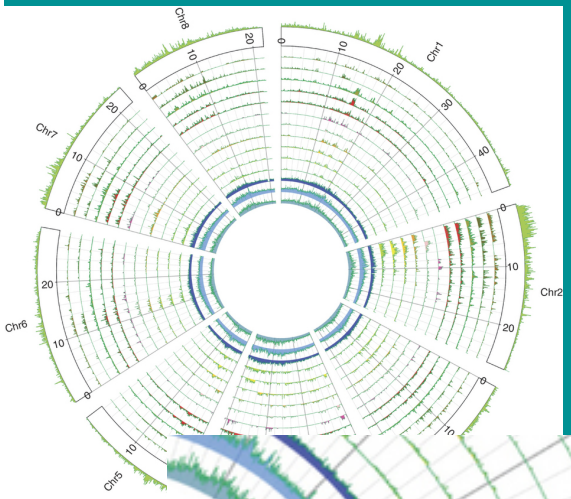
Mapping reads in outbred species

Average genome polymorphism levels

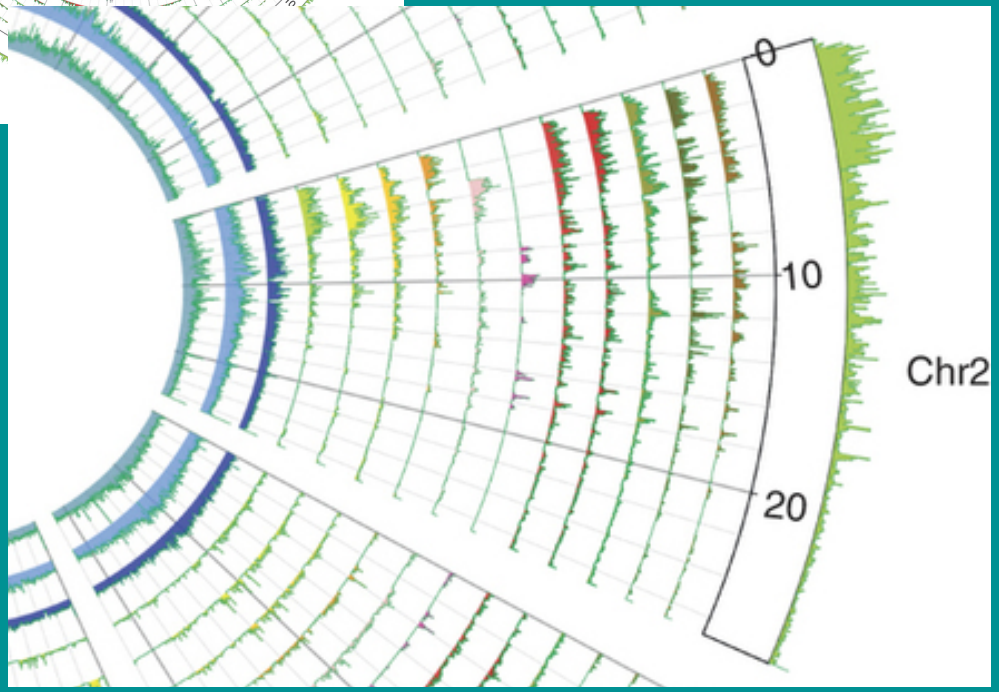


Leffler *et al.* 2012 *Plos Biol*

Is the mean where you want to look?



- Genes of interest are likely to have SNPs densities >> genomic average
- These are not likely to get mapped
- Leads to biased expression values

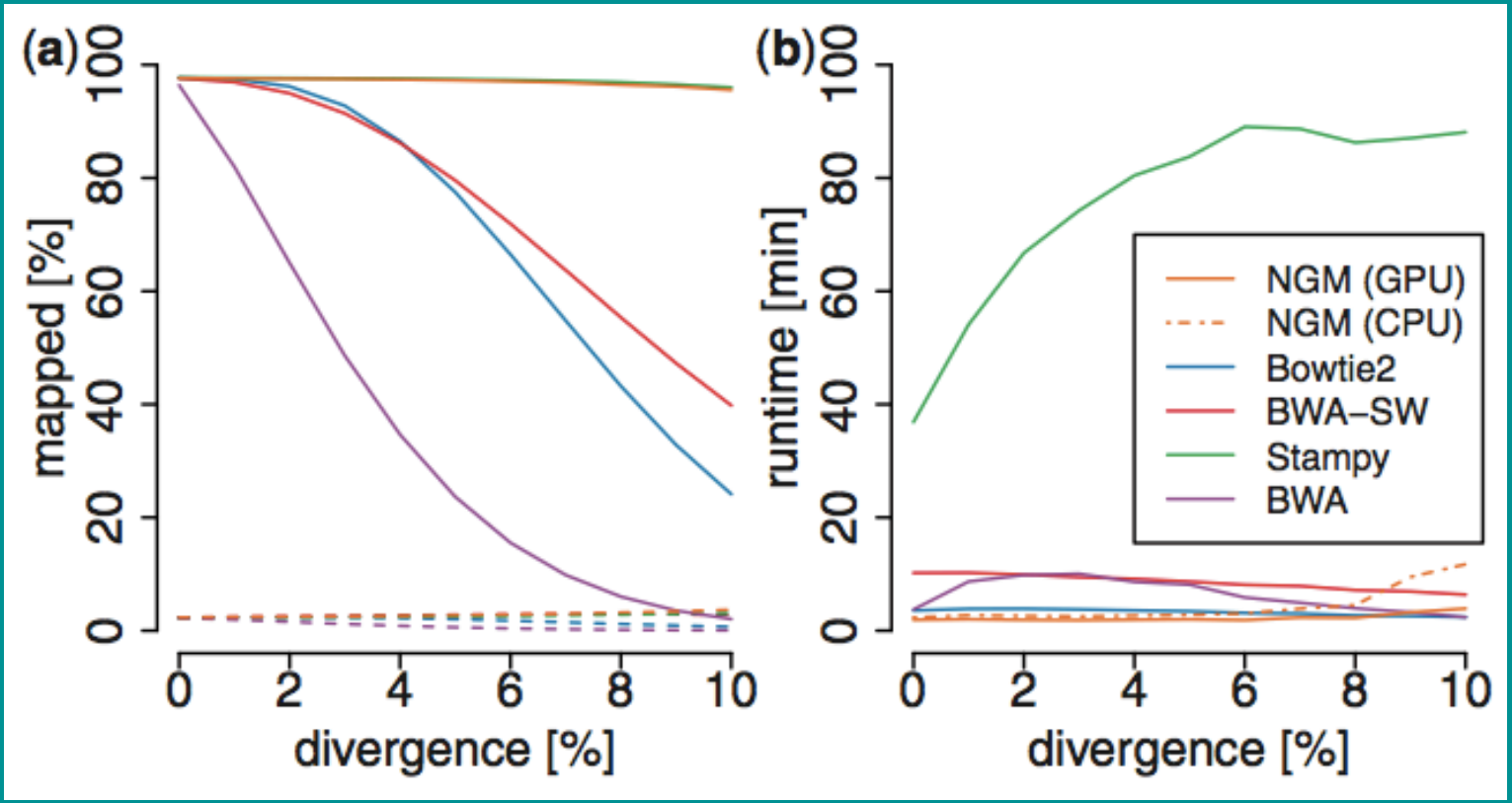


- Resequencing data will be allele biased
- But perhaps only in small fraction of genome

50-kb nonoverlapping sliding windows estimated from a sample of 23 haploid Peach lines

The International Peach Genome Initiative 2013 Nat. Gen.

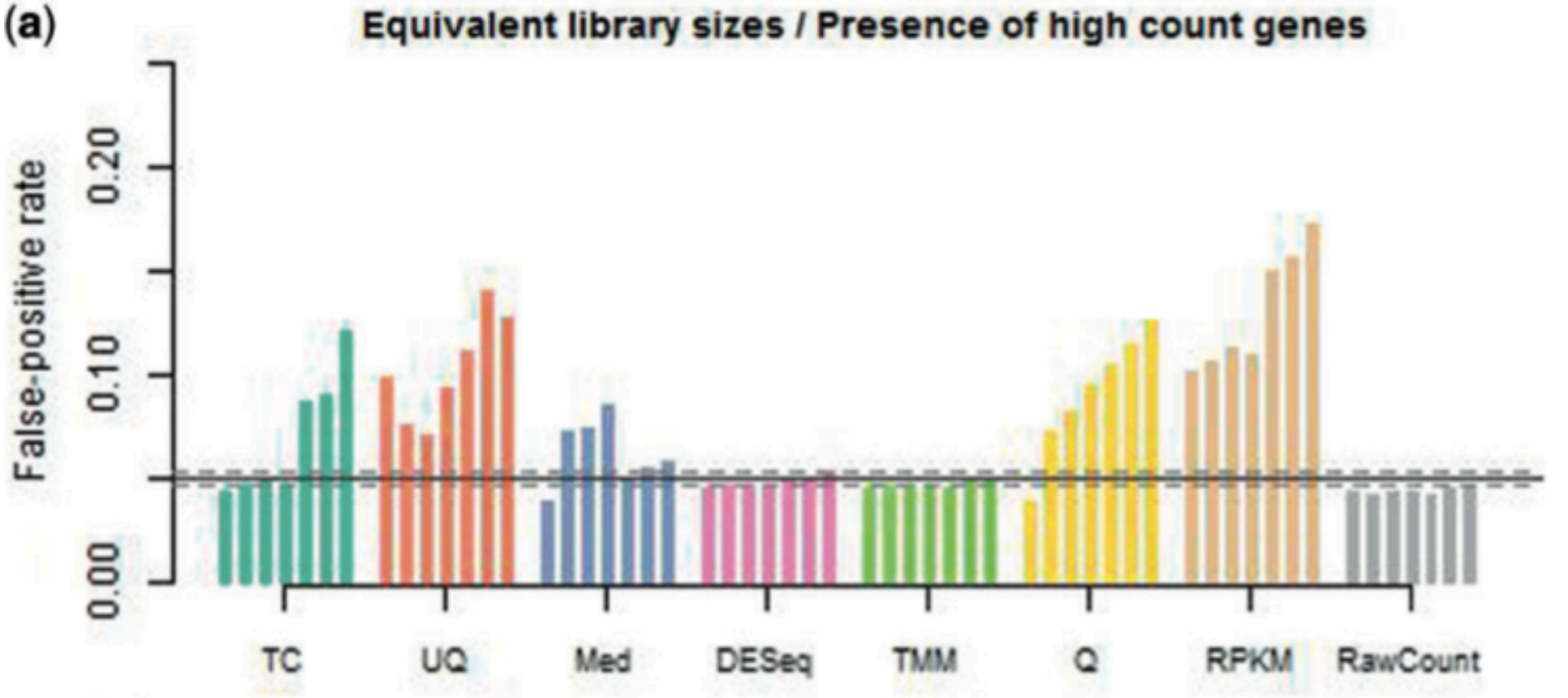
Allelic bias in read mapping



- 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?

Sedlazeck et al. 2013 *Bioinformatics*

Normalization

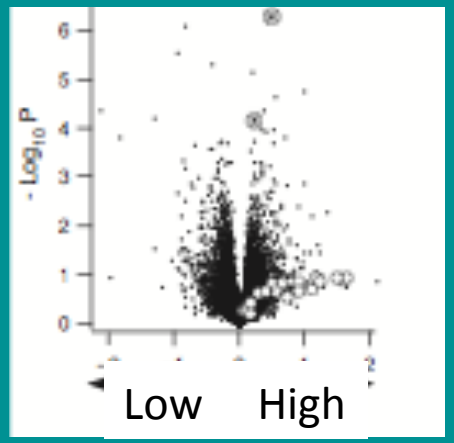


Dillies 2010 Brief in Bioinfo.

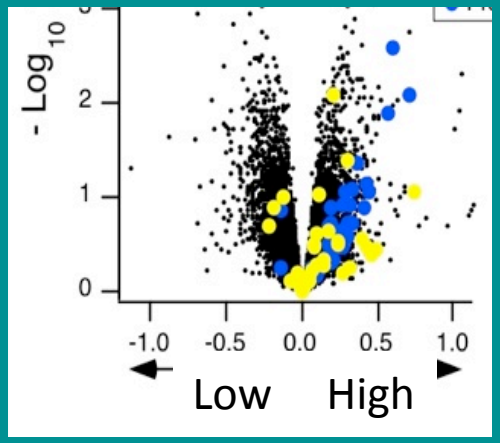
Can genetic variation affect dispersal?

- Identifying such variation could help
  - Ecological & evolutionary study (theoretical models)
  - Conservation biology (captive breeding, predictions)

Abdomen

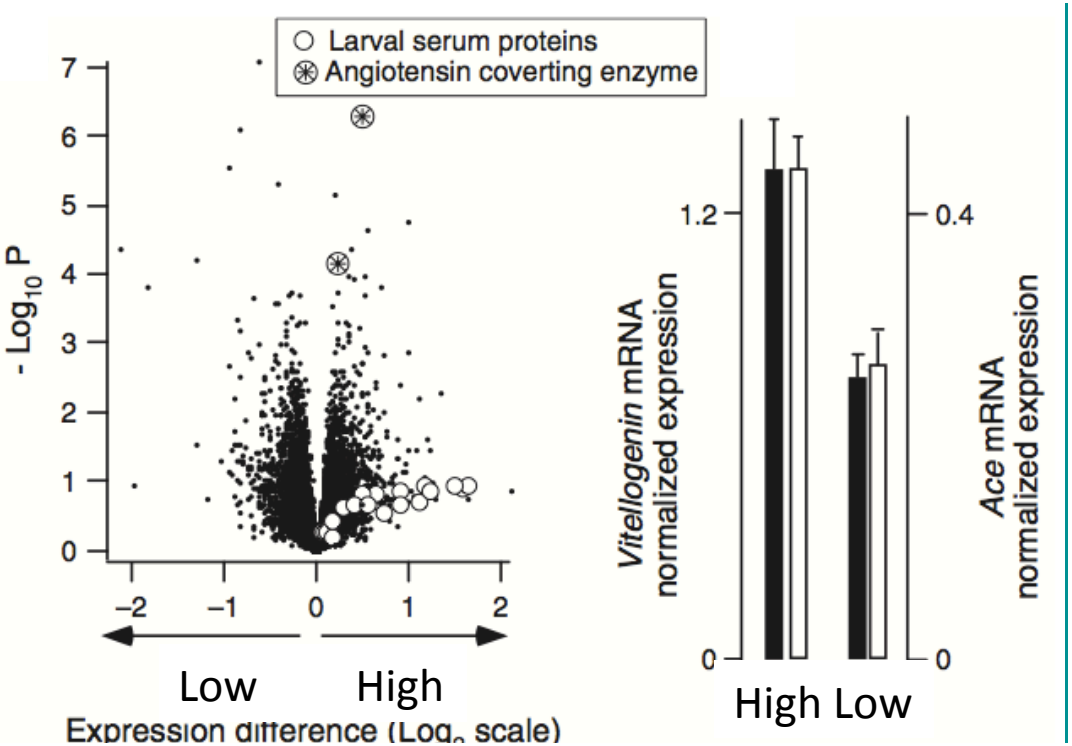


Thorax

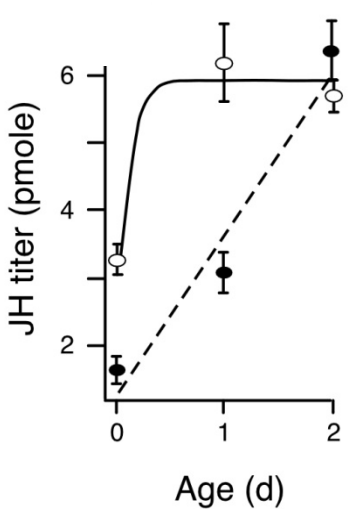


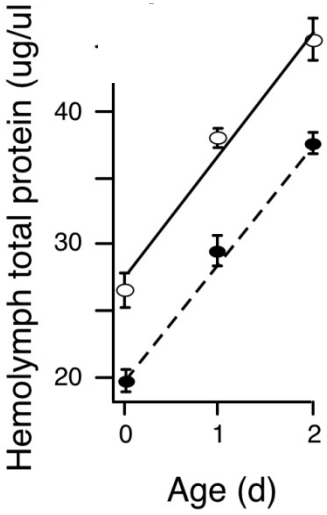
- Gene expression changes are primarily involved in protein allocation
  - Thorax - flight muscle performance
  - Abdomen - reproductive physiology
- A single gene could cause all these expression differences, but which gene?

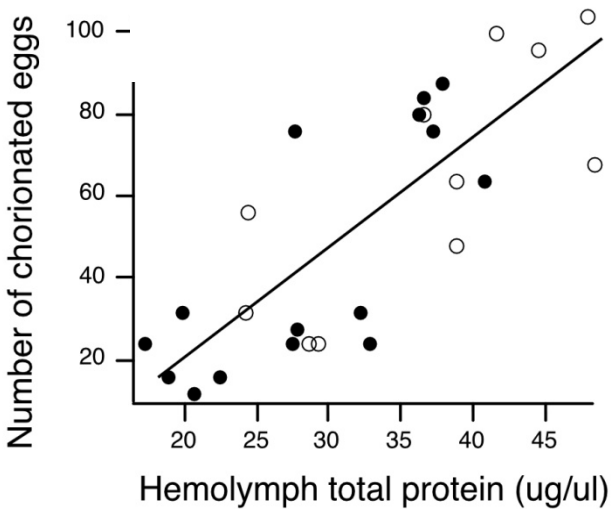
Butterfly dispersal genetics

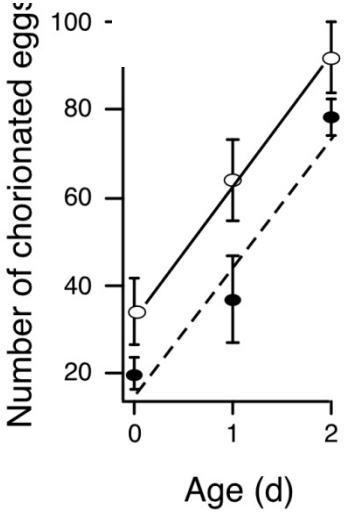


● Low dispersal ○ High dispersal









Wheat et al. 2010 Mol. Ecology

Most studies are annotation limited

- What is the biological meaning of the top Pvalue genes?
- Low Pvalue or expression genes are certainly important
- Gene set enrichments are key to insights
  - need network and regulatory insights relevant to the questions

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, wh	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-relatec	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

Melitaea cinxia

454 sequence database

100 My

Bombyx mori

Whole genome sequence, predicted gene set

320 My

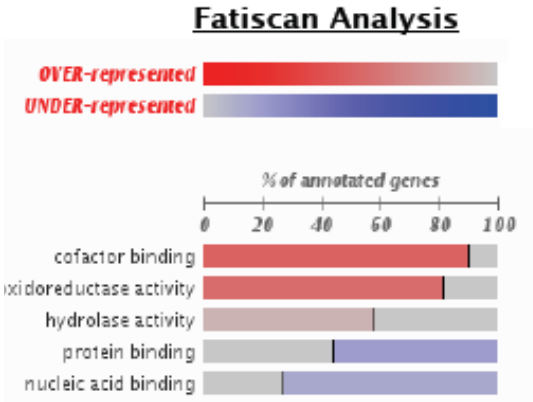
Drosophila melanogaster

Extensive genomic & functional resources

Assembly 2.0	Blastx	Bmori06 PepEd90	Blastp	Flybase gene ID
Contig_57178		BGIBMGA002704		CG33126
Contig_6821		BGIBMGA003247		CG6519
Contig_1004		BGIBMGA003248		CG6519
Contig_20226		BGIBMGA003248		CG6519
Contig_27720		BGIBMGA003248		CG6519
Contig_5260		BGIBMGA003249		CG6519
Contig_27110		BGIBMGA004806		CG33126
Contig_27390		BGIBMGA004806		CG33126
Contig_26901		BGIBMGA004865		CG33126
Contig_4713		BGIBMGA004866		CG33126
Contig_20081		BGIBMGA005329		CG3149
Contig_9982		BGIBMGA006733		CG6783
Contig_15387		BGIBMGA008859		CG4178
Contig_25362		BGIBMGA008859		CG4178
Contig_36071		BGIBMGA008859		CG4178

D. melanogaster lacks an orthologous reproductive physiology

Gene Set Enrichment analysis using Gene Ontology database



Life after your RNA-Seq experiment

- What are you likely to learn?
  - By measuring other aspects of the phenotype, we could at least validate and solidify our transcriptome insights
- What may limit your insights?
  - Single gene analyses can be restrictive
    - Statistically: FDR is very conservative
    - Biologically: genes work in networks varying in expression and direction across pathways
- Possible solutions
  - Gene set enrichment analysis: harness the functional network
  - Need data relevant to your phenotype and organism
    - Don't hesitate to make your own enrichment set

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

To address these we need to develop functional genomic insights in species with well understood ecologies that can be manipulated in the lab and in the field

Science News ... from universities, journals, and other research organizatio

Widespread Cannibalism May Have Caused Prehistoric Prion Disease Epidemics, Science Study Suggests

Apr. 11, 2003 — Human flesh may have been a fairly regular menu item for our prehistoric ancestors, according to researchers. They say it's the most likely explanation for their discovery that genes protecting against prion diseases -- which can be spread by eating contaminated flesh -- have long been widespread throughout the world

Opinion TRENDS in Genetics Vol.20 No.7 July 2004

ELSEVIER

Balancing claims for balancing selection

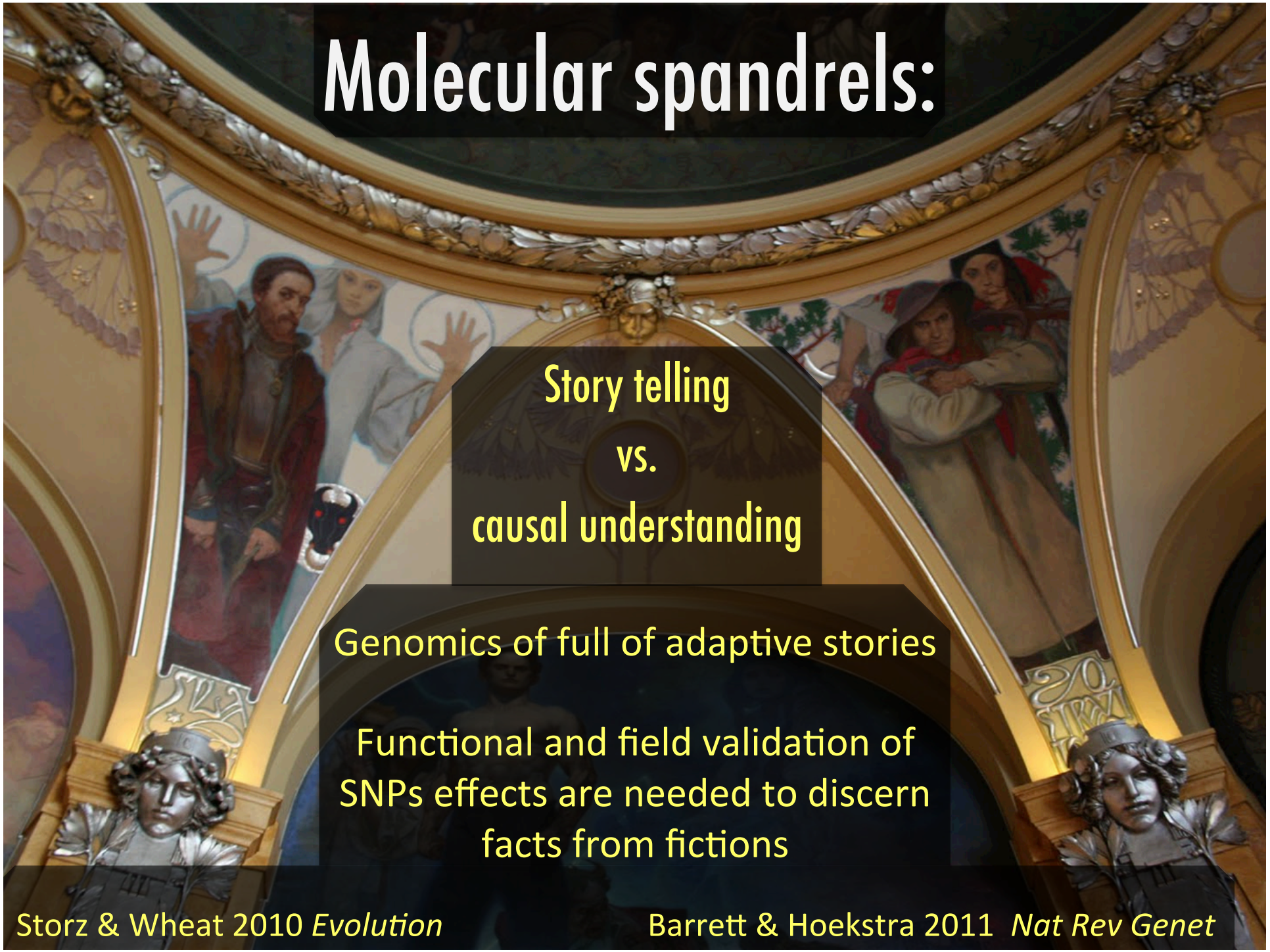
Martin Kreitman[1] and Anna Di Rienzo[2]

Letter

Assessing the signatures of selection in *PRNP* from polymorphism data: results support Kreitman and Di Rienzo's opinion

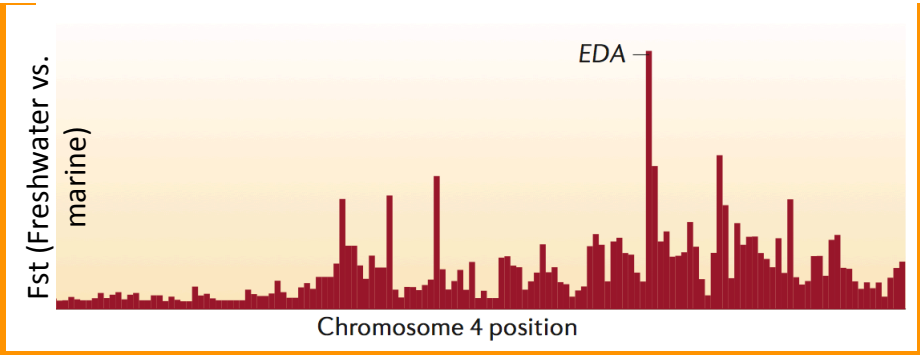
Marta Soldevila[1], Francesc Calafell[1], Agnar Helgason[2], Kári Stefánsson[2] and Jaume Bertranpetit[1]

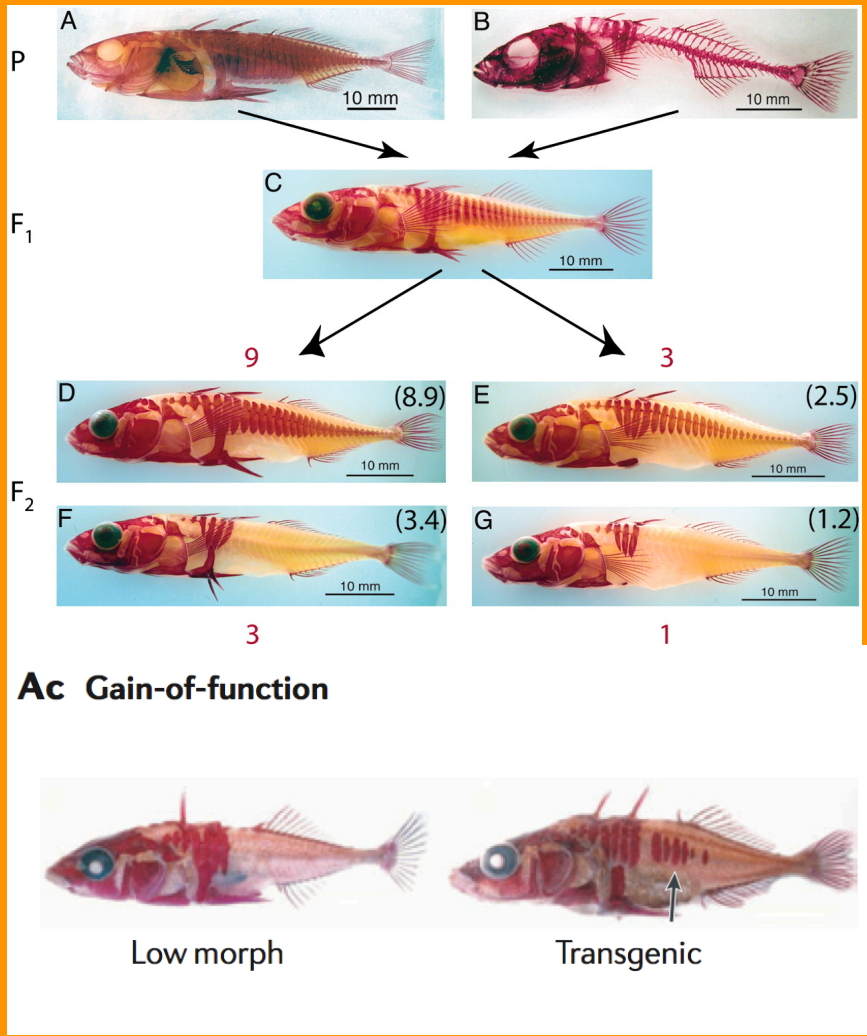
Story time in genomics land .....

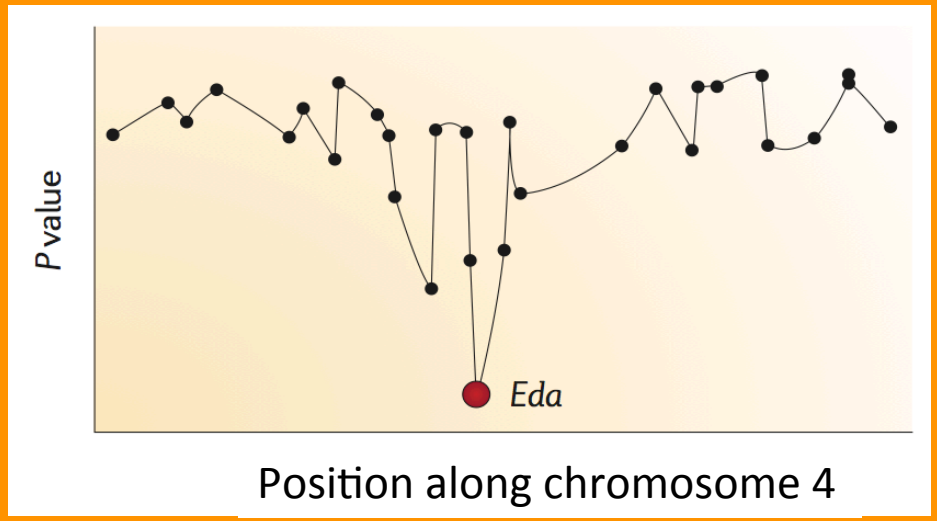

Molecular spandrels:
Story telling
vs.
causal understanding
Genomics of full of adaptive stories
Functional and field validation of SNPs effects are needed to discern facts from fictions
Storz & Wheat 2010 *Evolution*
Barrett & Hoekstra 2011 *Nat Rev Genet*

Model adaptation: the *Eda* gene

- Causes loss in body armor
  - Field association
  - QTL mapping





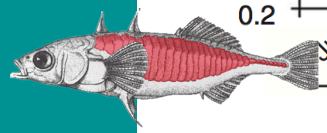


Back to nature: do we know what we think we know?

- Is low armor really adaptive in fresh water?
- Lets replay the selection event
  - Equal frequency *Eda* alleles in fresh water ponds

Studies in the field can uncover unexpected and complex selection dynamics

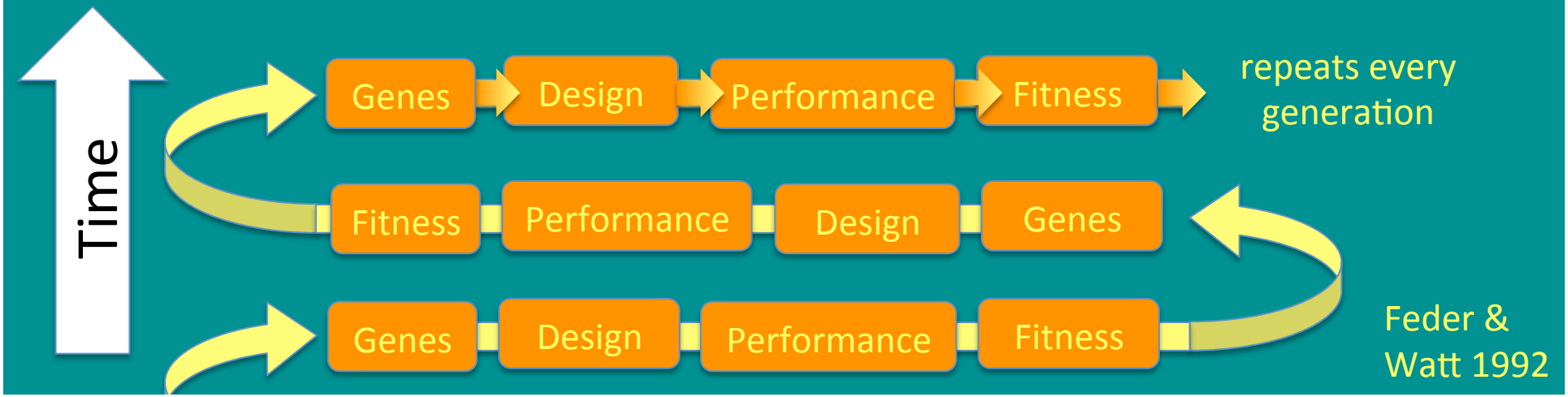
- Linked effect of other genes in the inversion on LG4?
- Is Eda even a target of selection?



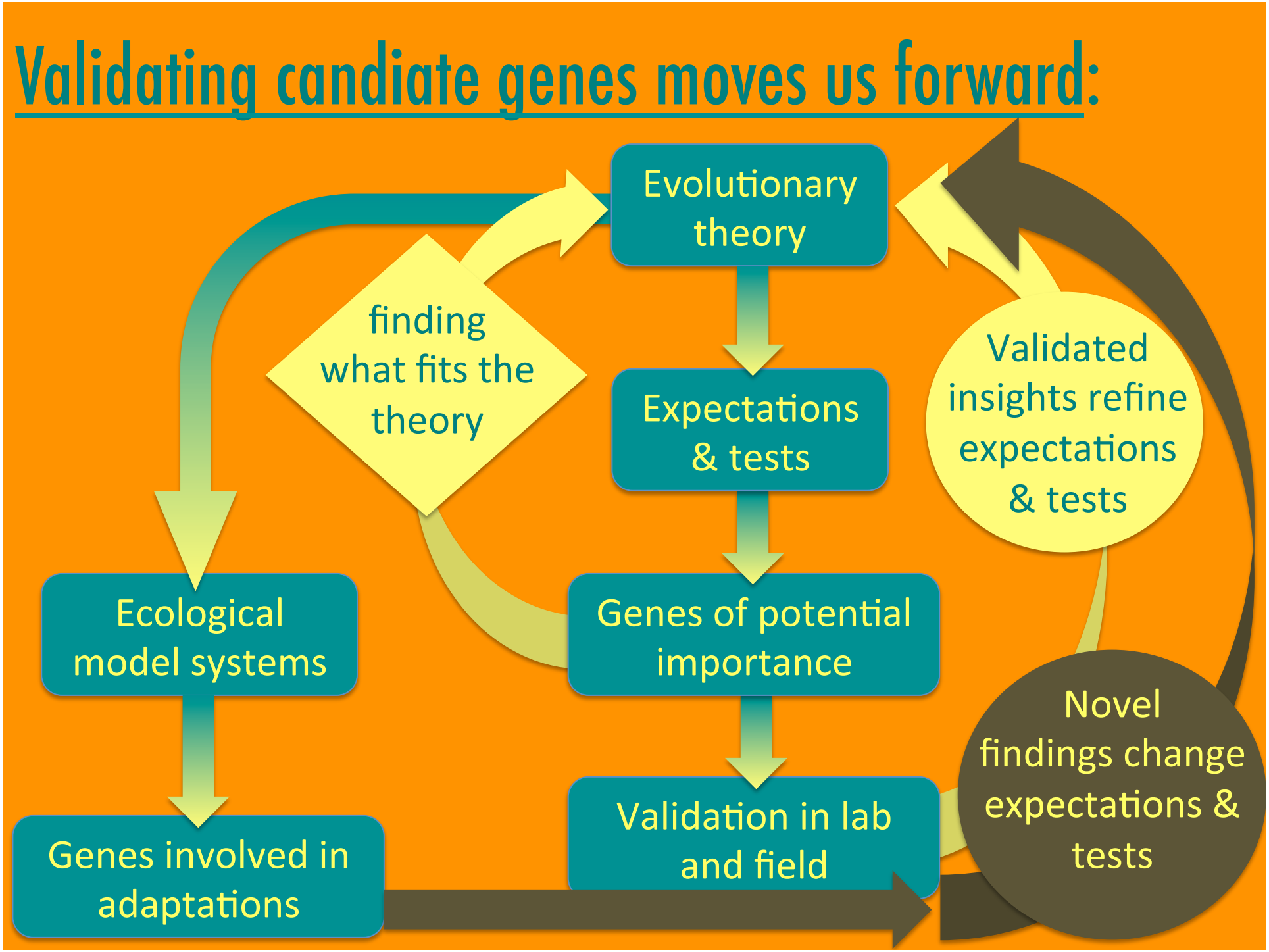
Barrett et al. 2008 Science

Adaptation by natural selection

Insights at one level can be tested at other levels



Validating candiate genes moves us forward:

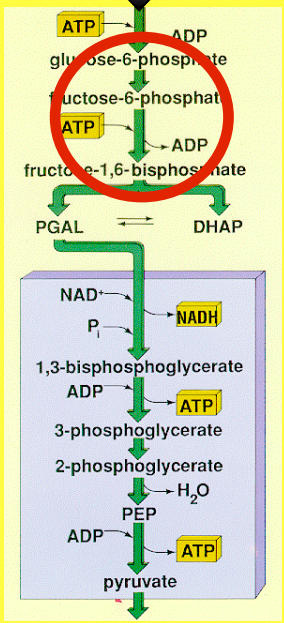


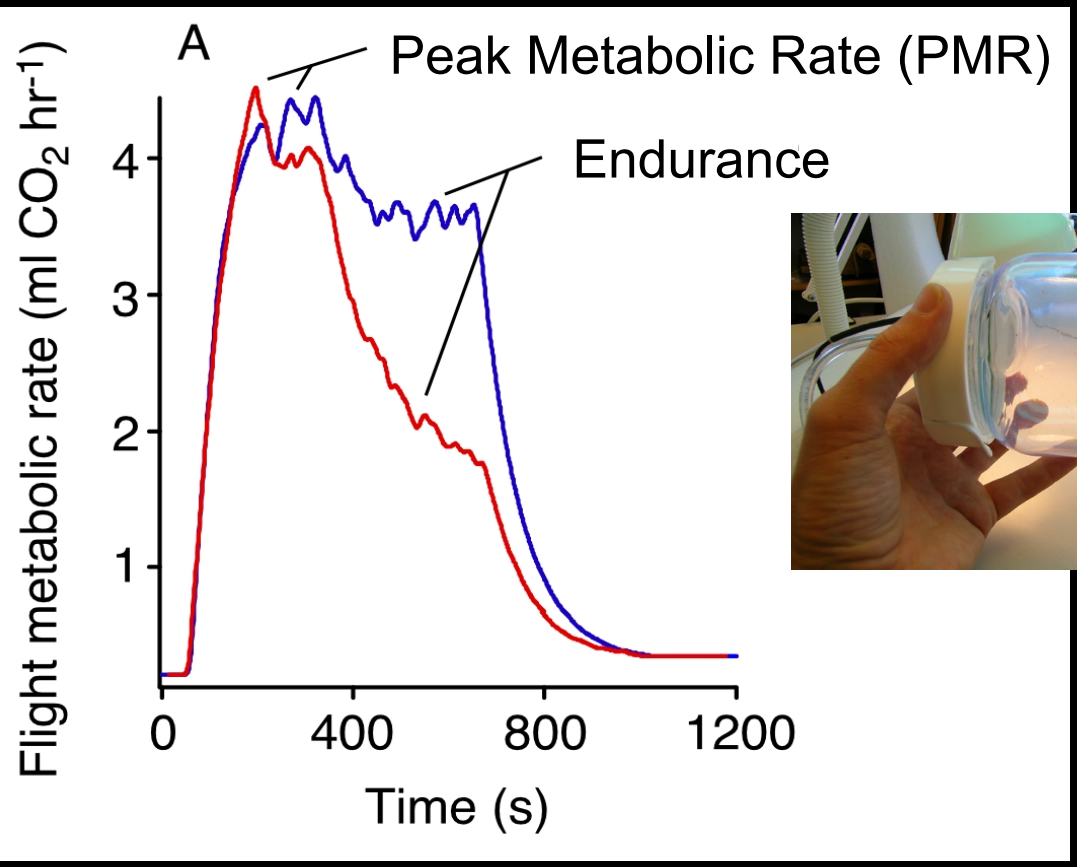
Can you get there from here?

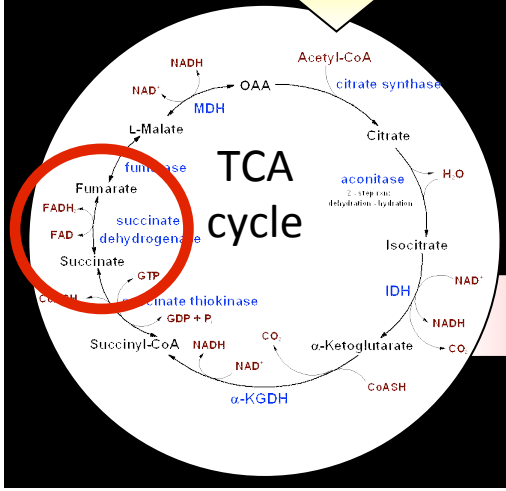
- Candidate genes associated with large effects on Darwinian fitness
  - Classic study systems in the wild
  - Validation process is still ongoing
- Can 'Second Generation' approachs find these same genes?
  - Sometimes not, or at least not easily
- Why?
  - Cause the modern tools aren't designed with such architectures in mind

Glucose

Glycolysis

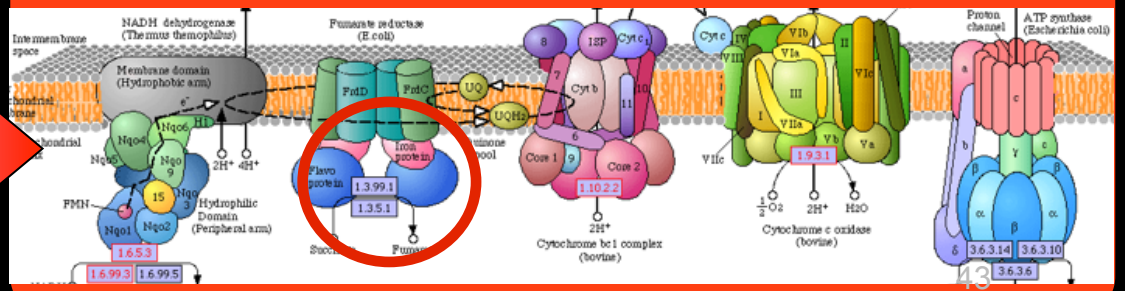






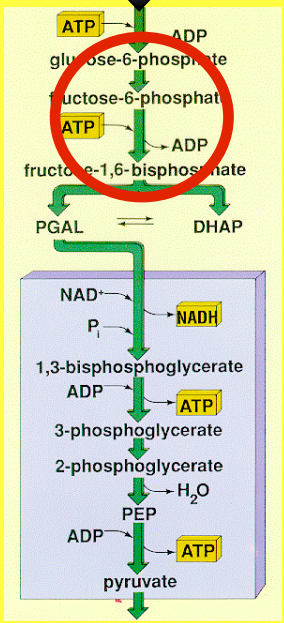
NADH

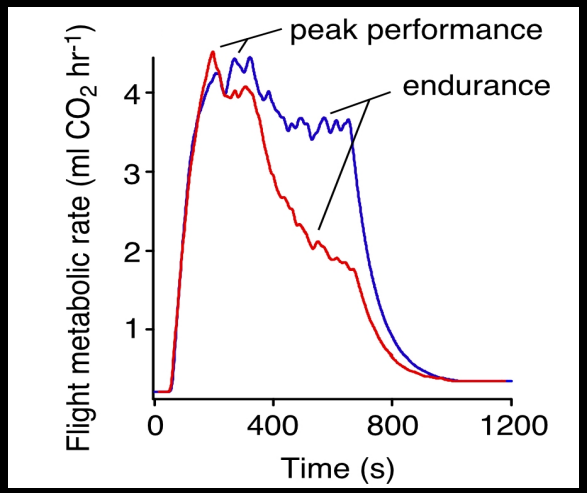
Electron Transport

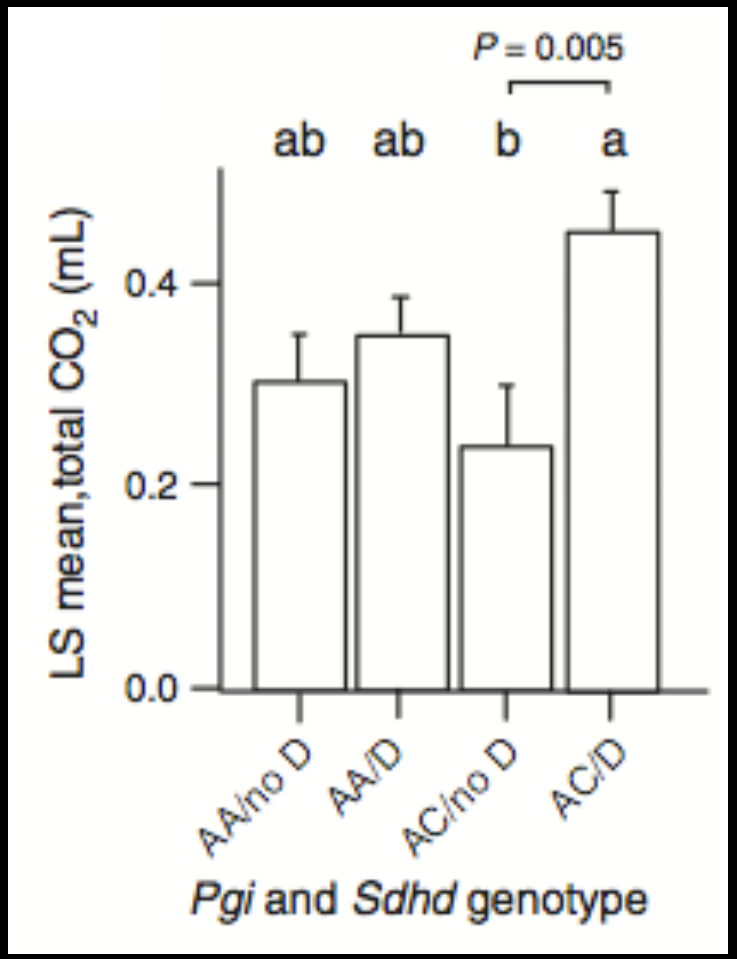


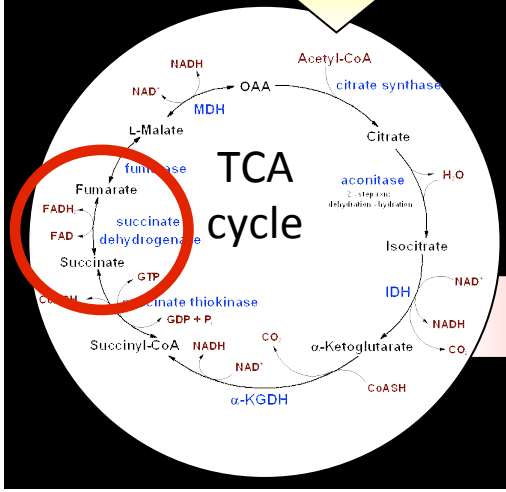
Glucose

Glycolysis



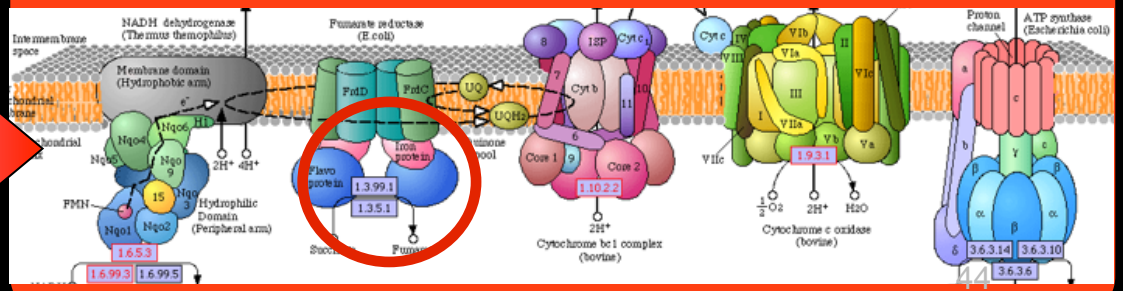






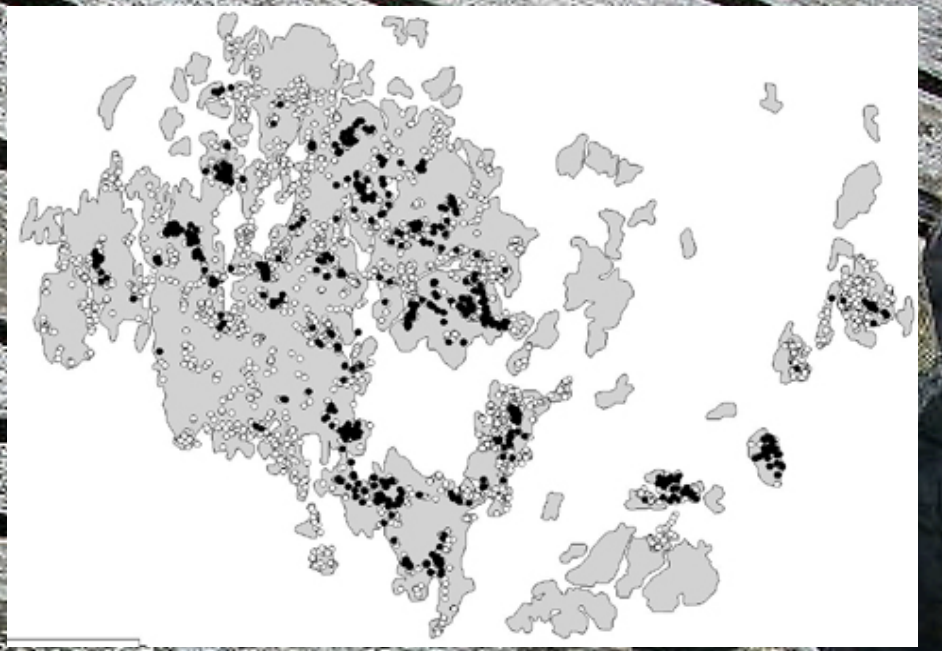
NADH

Electron Transport



Fitness effects in the wild

Year to year change in number of families living in 43 demes



Pgi & Sdhd SNPs are in linkage equilibrium

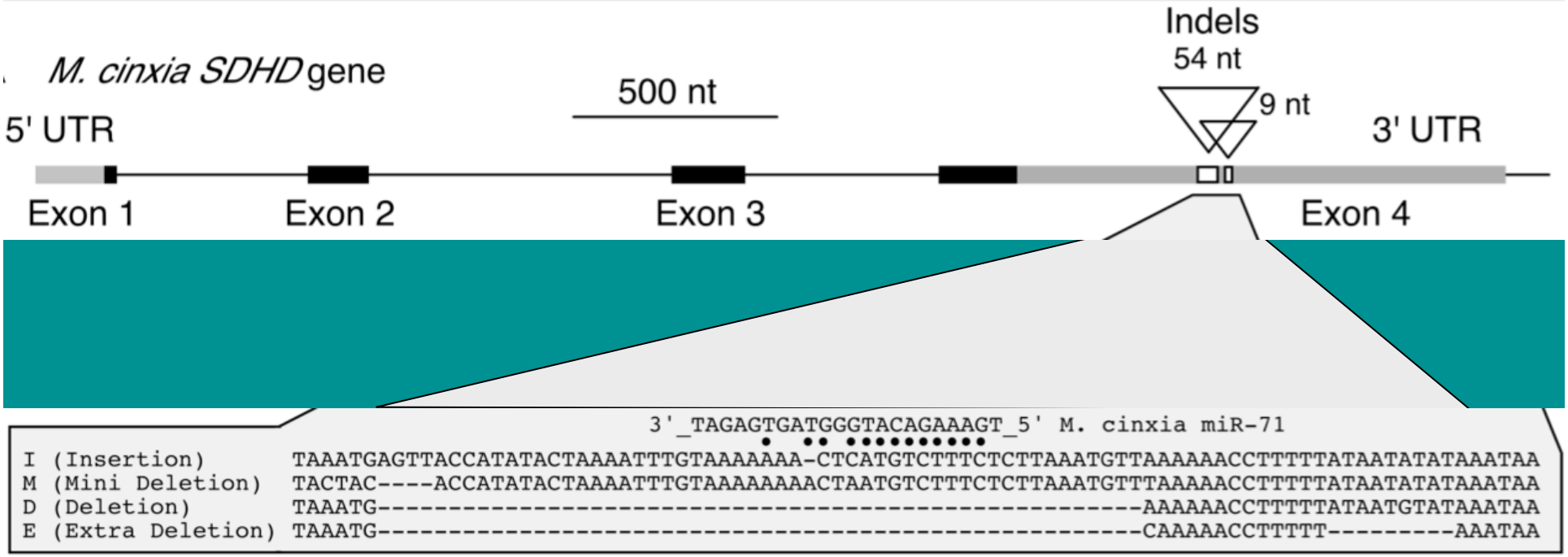
Source (Full model $R^2 = 0.64$)	d.f.	*F* ratio	*P*
Patch area	1	0.0001	0.99
Frequency *Pgi F*	1	10.9	0.002
Frequency *Pgi F* × Patch area	1	22.5	<0.0001
Frequency *Sdhd M* allele	1	19.1	0.0001
Frequency *Sdhd M* allele × Patch area	1	21.1	<0.0001

Wheat et al. 2011

Succinate dehydrogenase d (Sdhd)

3’ UTR indel associated with performance and fitness in 5 studies across 3 populations

James H. Marden
Penn State Univ.

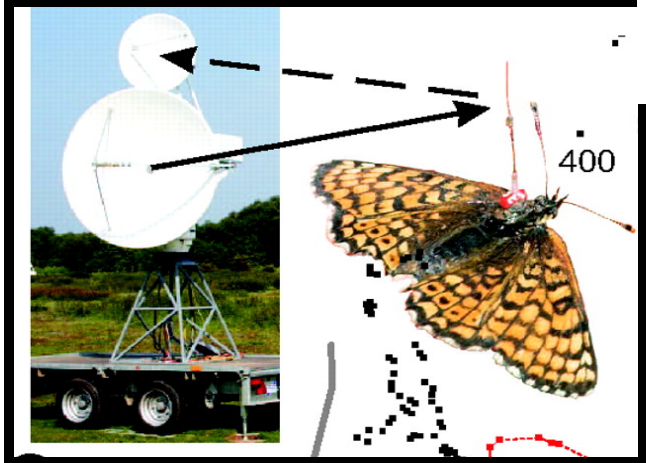


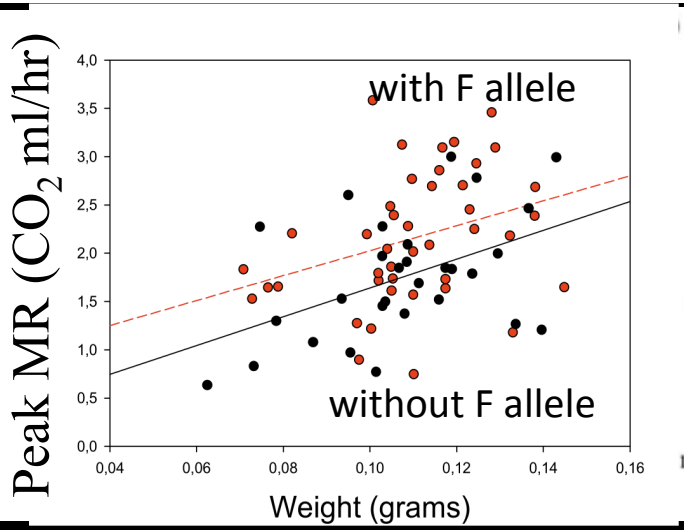
Assembly of this region would be biased to most common allele.

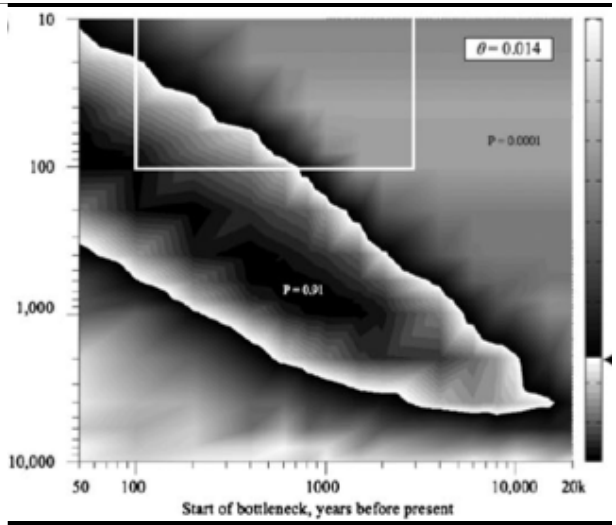
Mapping would be allelic biased

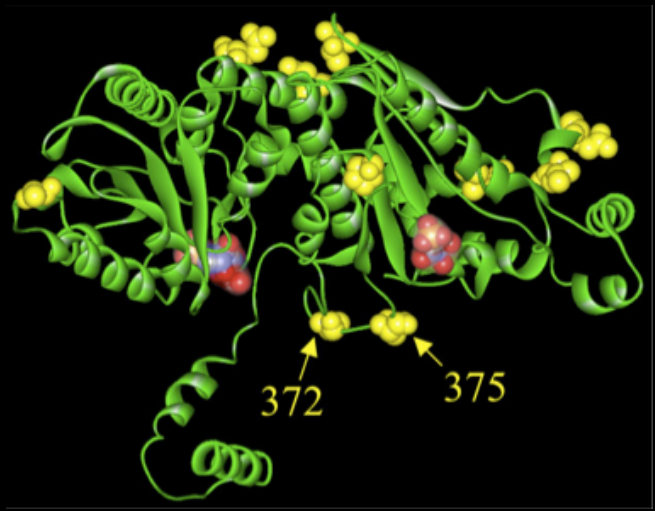
Wheat et al. 2011; Marden et al. 2013

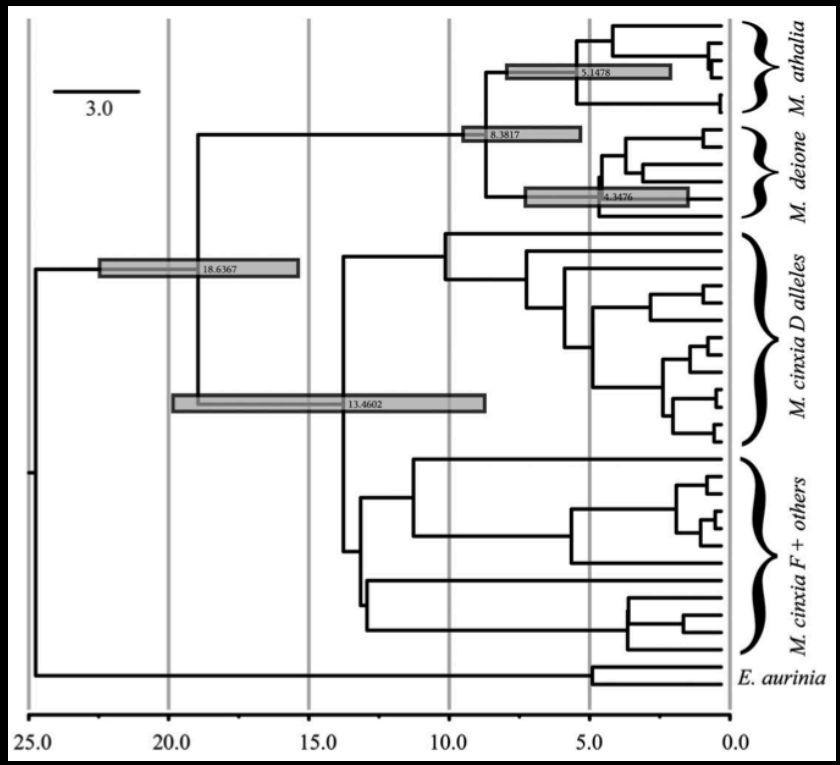
Phosphoglucose isomerase (*Pgi*)







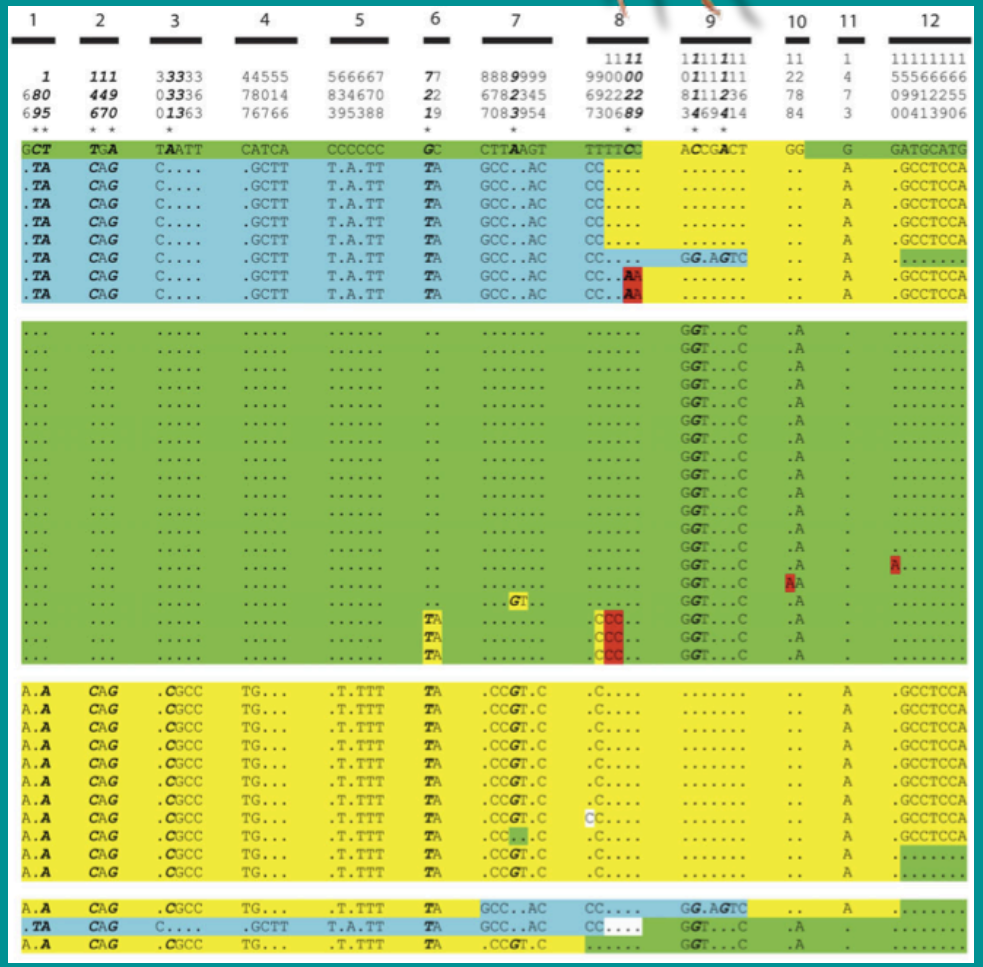




Niitepõld 2008; Wheat et al. 2010; Haag et al. 2005

Pgi genetic variation: 1 SNP/30 bp

- Reduced recombination
- Very divergent alleles at intermediate frequency
  - Older than age of sister clade of 5 species
- Most mapping software would exhibit strong allelic mapping bias



Wheat et al. 2010 MBE

Conclusions:

What do we really know about .....

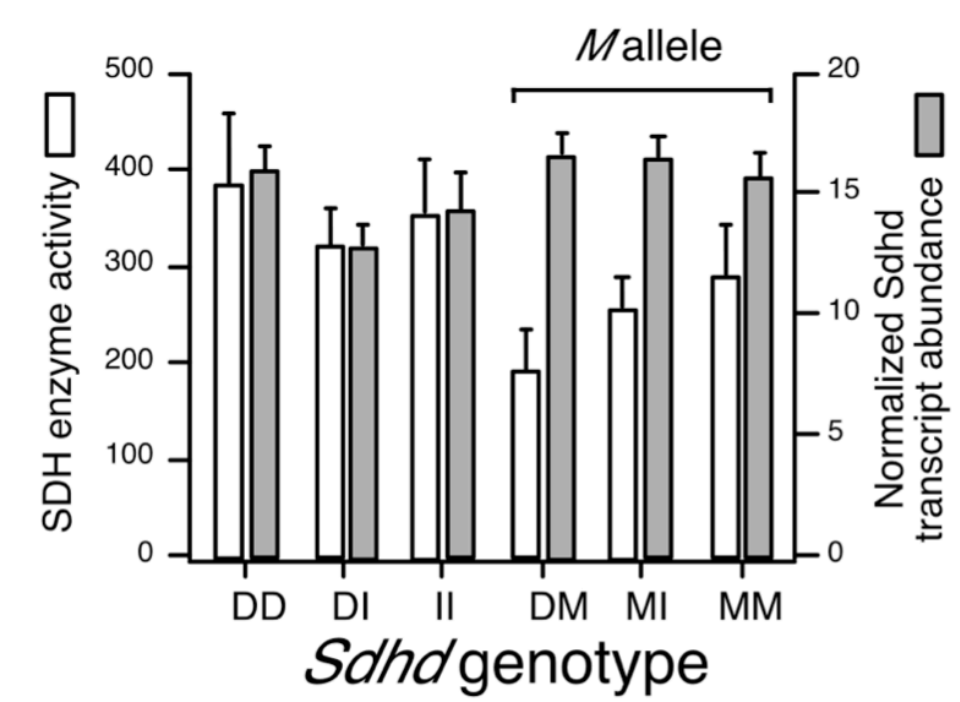
- molecular evolutionary dynamics?
- targets of selection in the wild?
- the age of species?
- transcriptomes?
- the performance of 2nd gen methods?

Funding sources

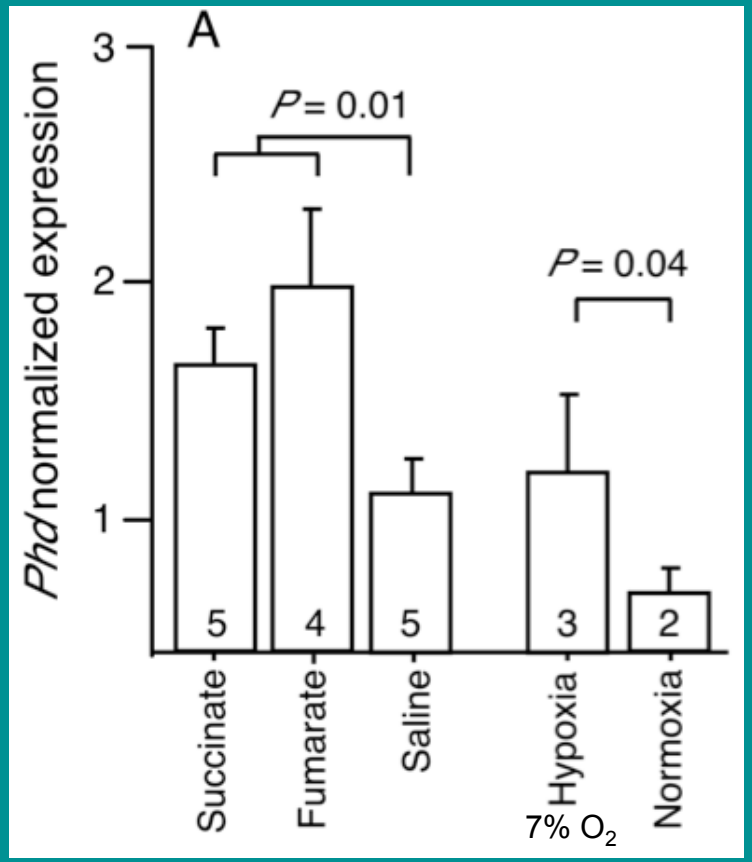
Finnish Academy of Sciences (Finland)
Vetenskapsrådet (Sweden)
Wallenberg Foundation (Sweden)

How could Sdhd affect flight?

- Loss of function studies in humans result in constitutive activation of HIF pathway
- Increased flanking metabolites result in hypoxia signaling



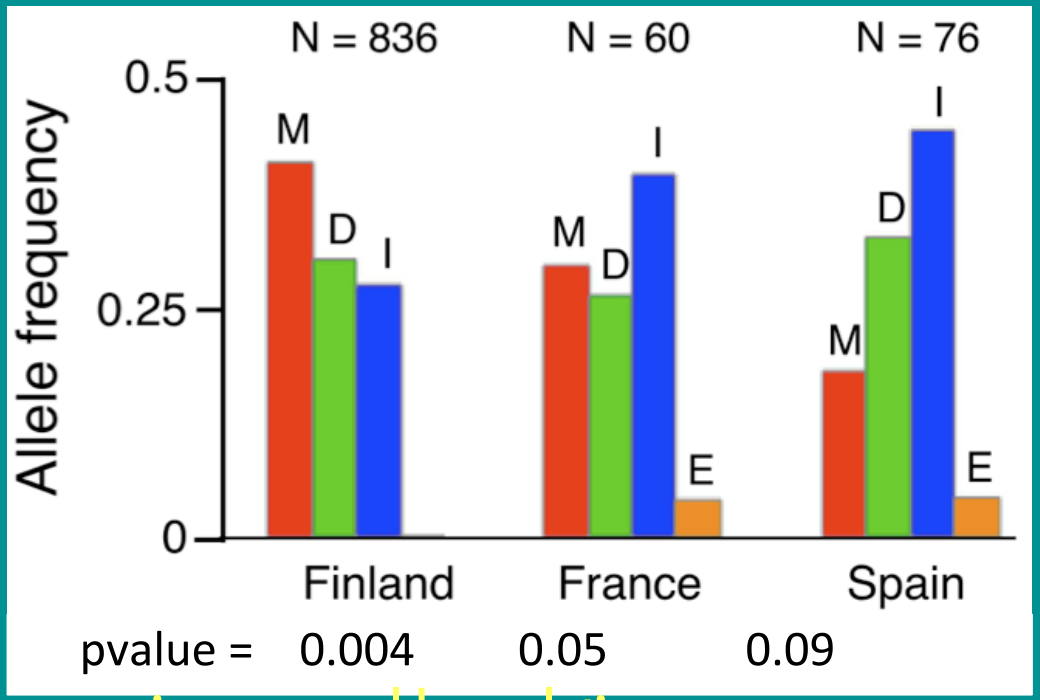
Suggests micro RNA down-regulation of SDH enzyme



Marden et al., accepted, Evolution

Sdhd indel alleles

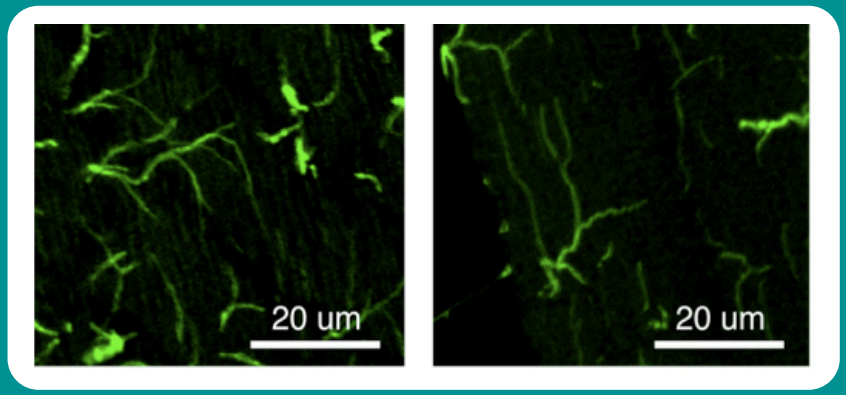
- Excess homozygosity / intermediate frequency within populations among alleles
  - Ewens Watterson tests suggestive of balancing selection



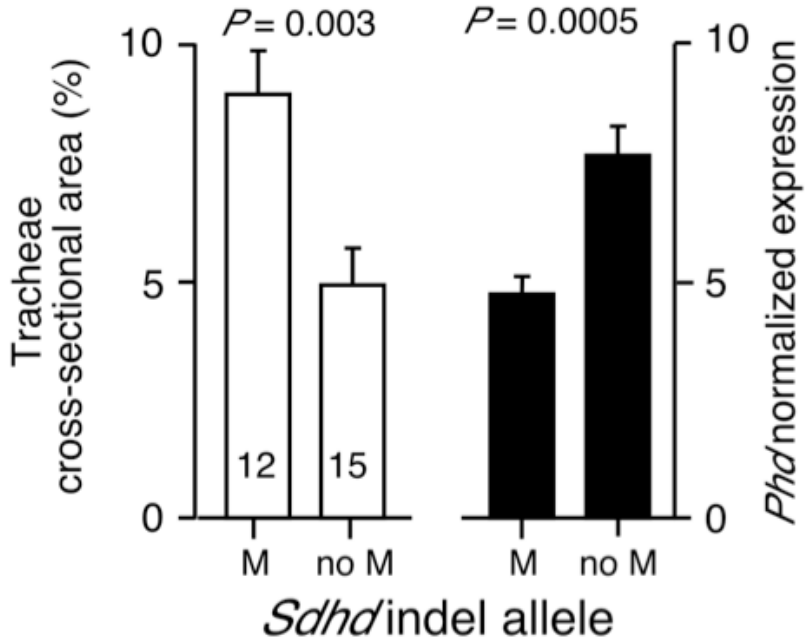
- M allele higher frequency in new vs. old populations
  - P = 0.04; N = 94 butterflies, 33 populations)

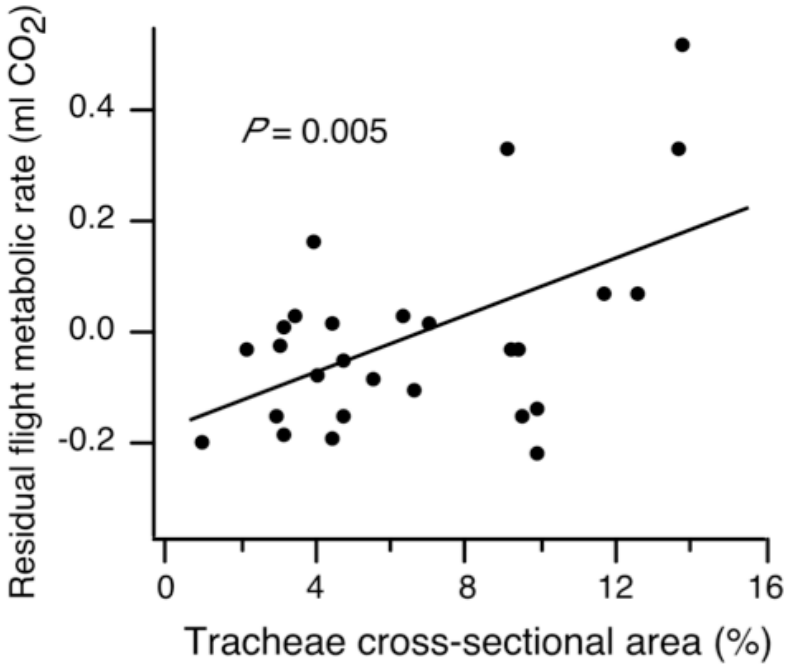
Wheat et al. 2011

Tracheael elaboration

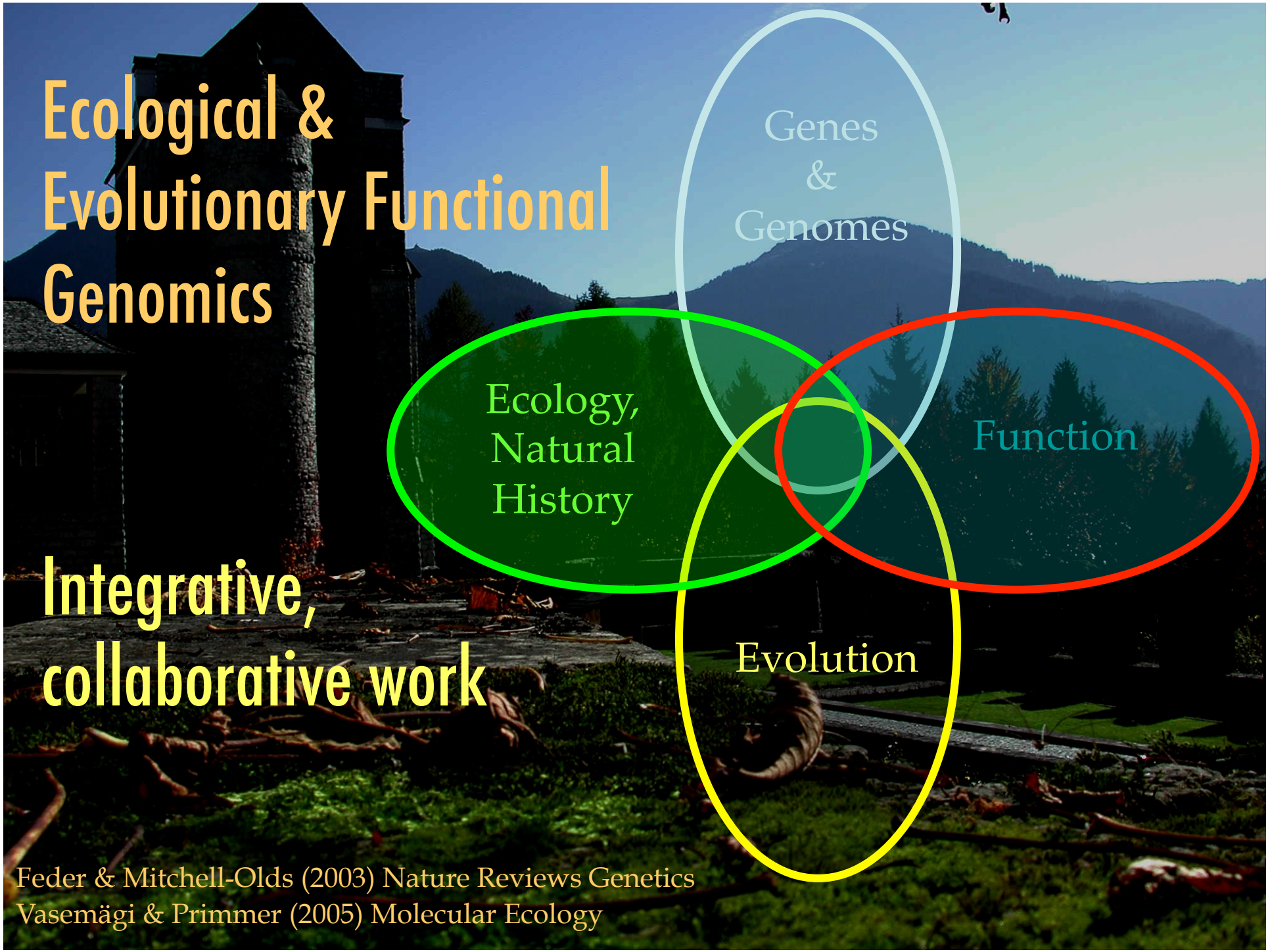


Two butterflies differing in tracheal elaboration





Marden et al., accepted, Evolution


Ecological &
Evolutionary Functional
Genomics
Genes
&
Genomes
Ecology,
Natural
History
Function
Evolution
Integrative,
collaborative work
Feder & Mitchell-Olds (2003) Nature Reviews Genetics
Vasemägi & Primmer (2005) Molecular Ecology