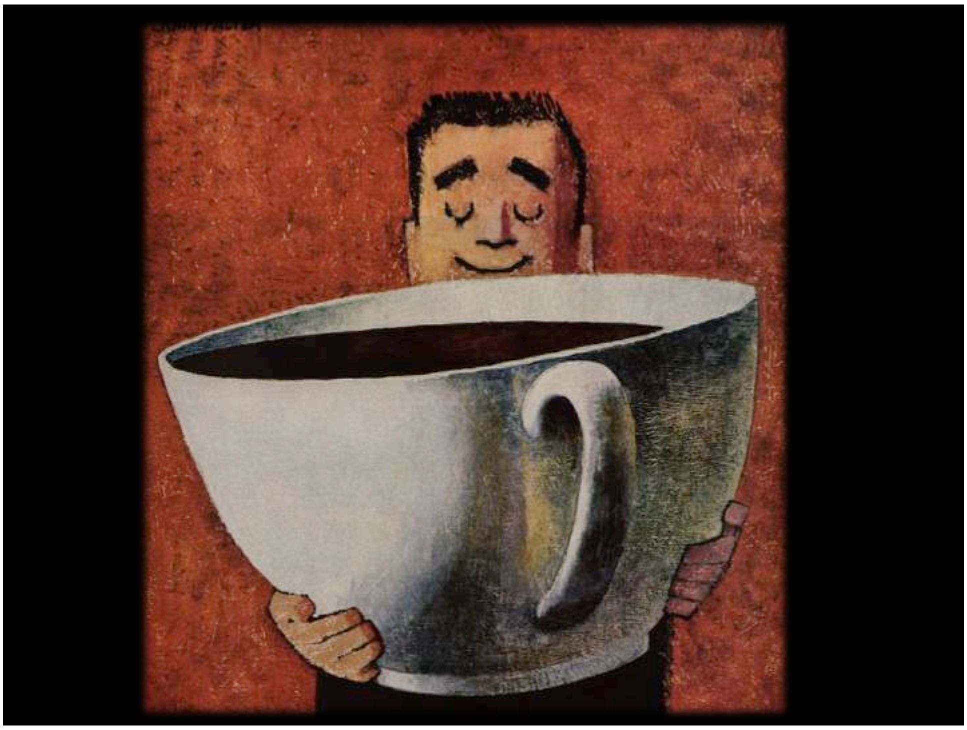
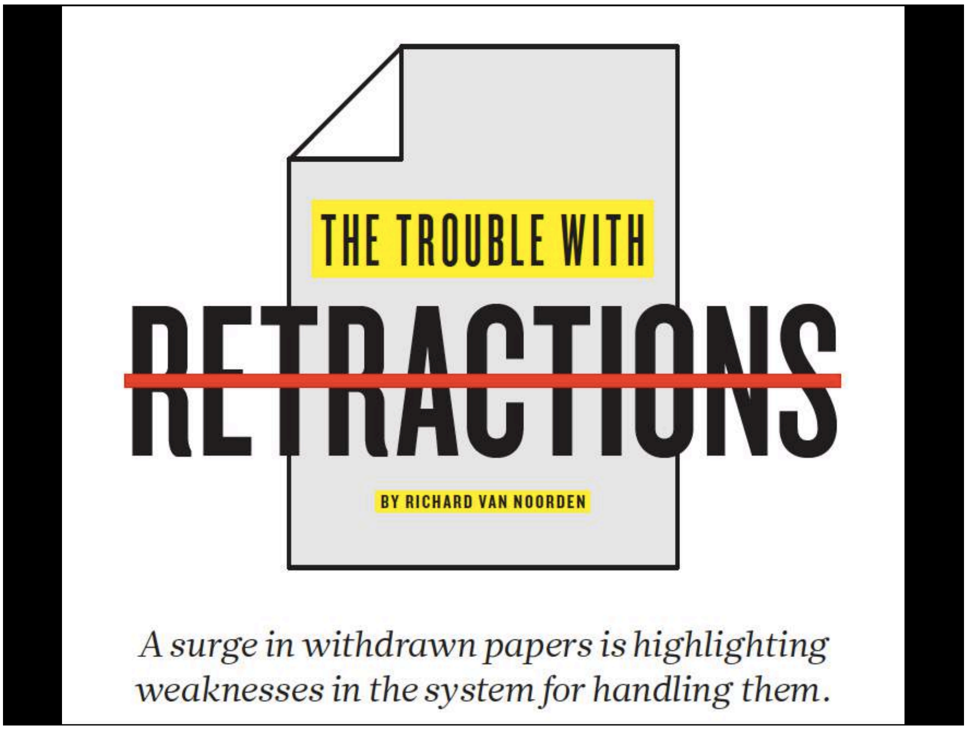
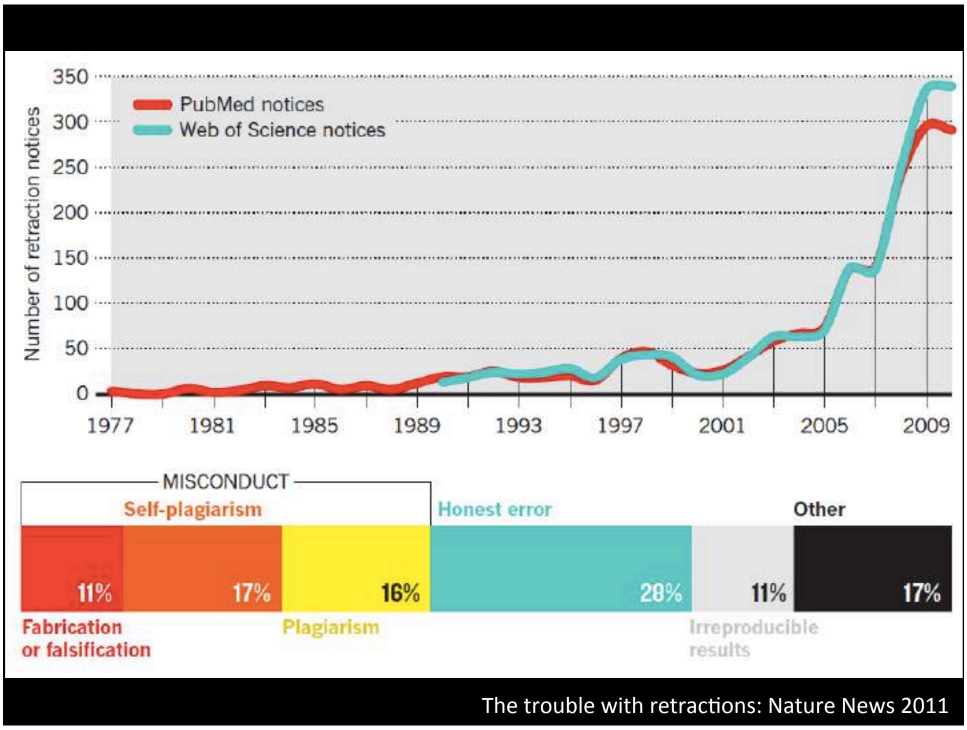




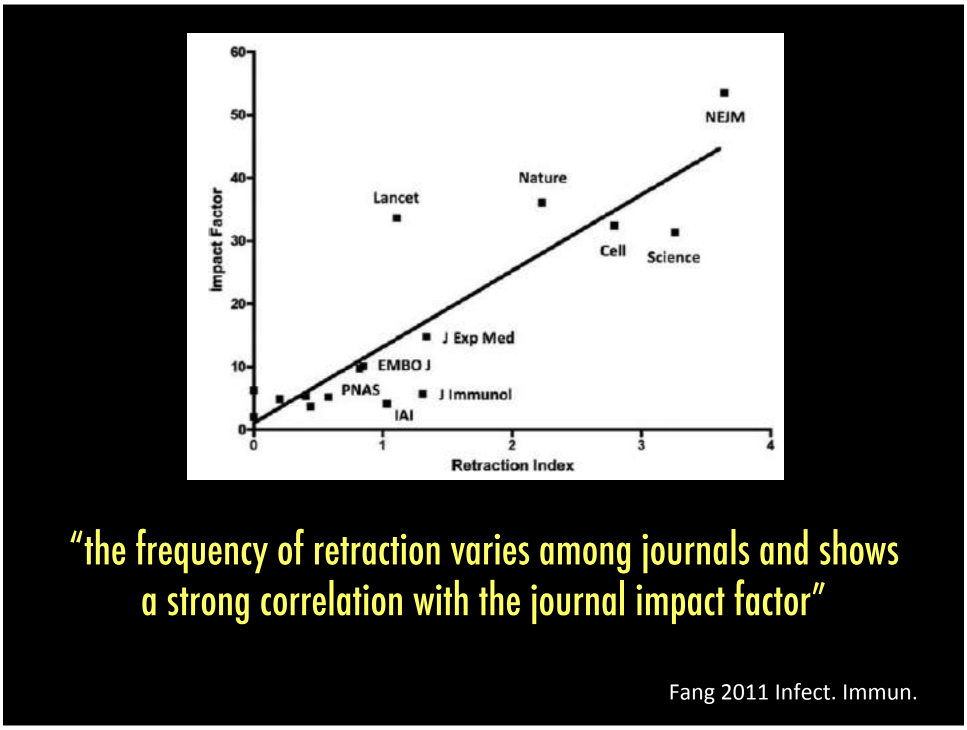
THE TROUBLE WITH ~~RETRACTIONS~~

BY RICHARD VAN NOORDEN

A surge in withdrawn papers is highlighting weaknesses in the system for handling them.

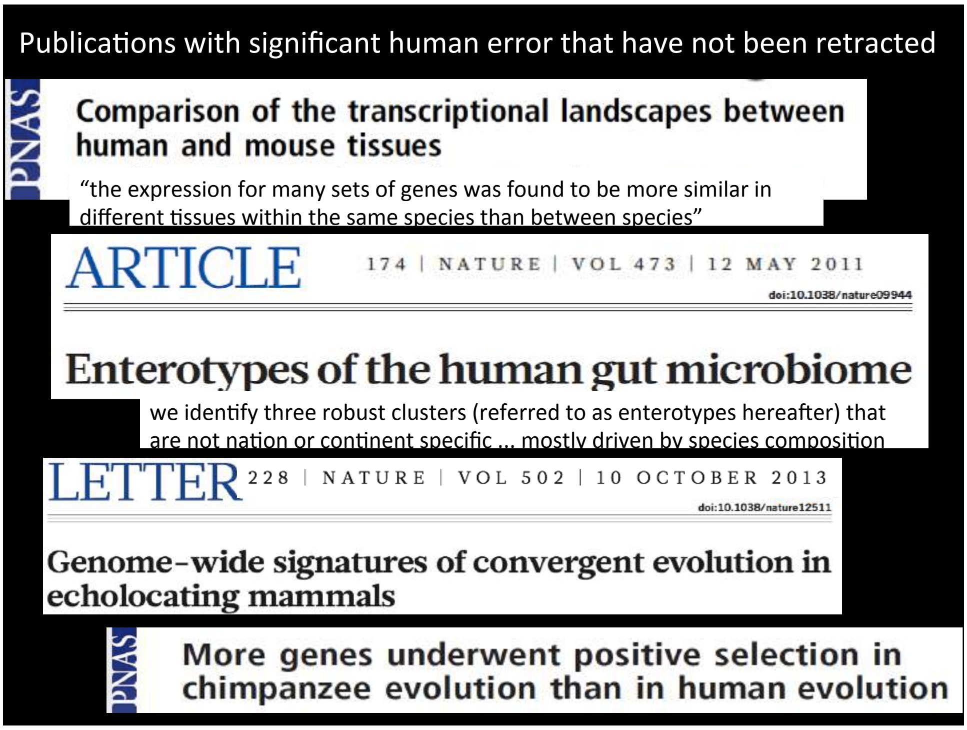


The trouble with retractions: Nature News 2011



“the frequency of retraction varies among journals and shows a strong correlation with the journal impact factor”

Fang 2011 Infect. Immun.



Publications with significant human error that have not been retracted

PNAS

Comparison of the transcriptional landscapes between human and mouse tissues

“the expression for many sets of genes was found to be more similar in different tissues within the same species than between species”

ARTICLE 174 | NATURE | VOL 473 | 12 MAY 2011

doi:10.1038/nature09944

Enterotypes of the human gut microbiome

we identify three robust clusters (referred to as enterotypes hereafter) that are not nation or continent specific ... mostly driven by species composition

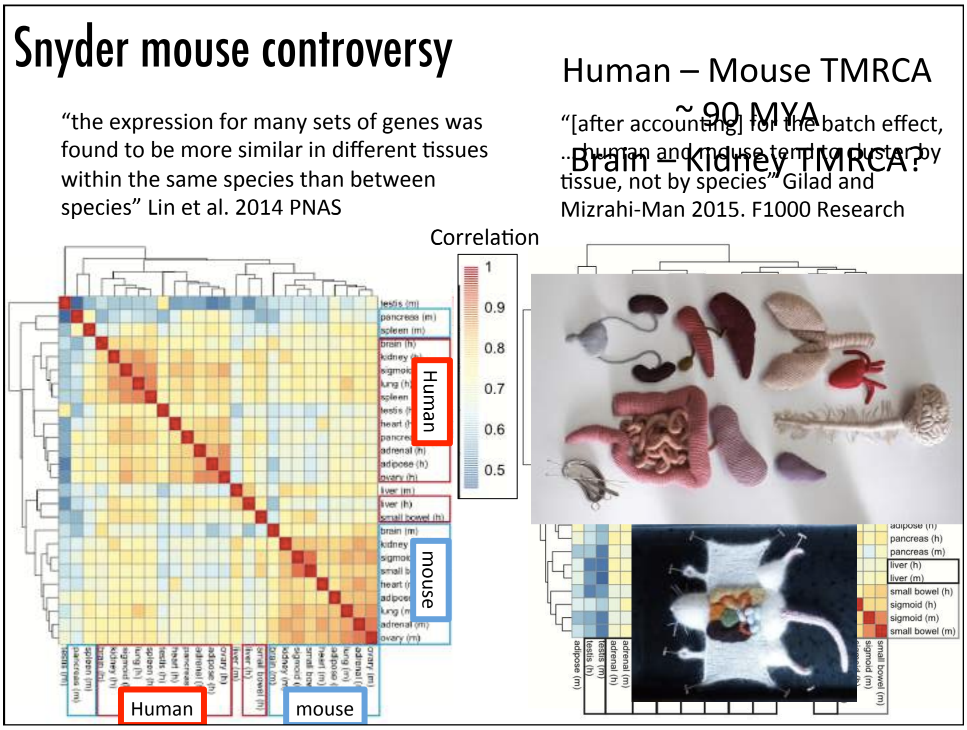
LETTER 228 | NATURE | VOL 502 | 10 OCTOBER 2013

doi:10.1038/nature12511

Genome-wide signatures of convergent evolution in echolocating mammals

PNAS

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



Snyder mouse controversy

“the expression for many sets of genes was found to be more similar in different tissues within the same species than between species” Lin et al. 2014 PNAS

Human – Mouse TMRCA ~ 90 MYA

Brain – Kidney TMRCA?

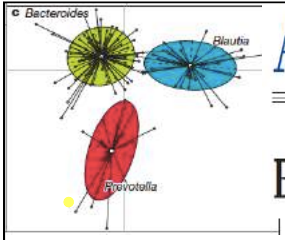
“[after accounting] for the batch effect, ..human and mouse tend to cluster by tissue, not by species” Gilad and Mizrahi-Man 2015. F1000 Research

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	● Human
testis		pancreas		● Mouse

Solution = Keep technical effects orthogonal to biological

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



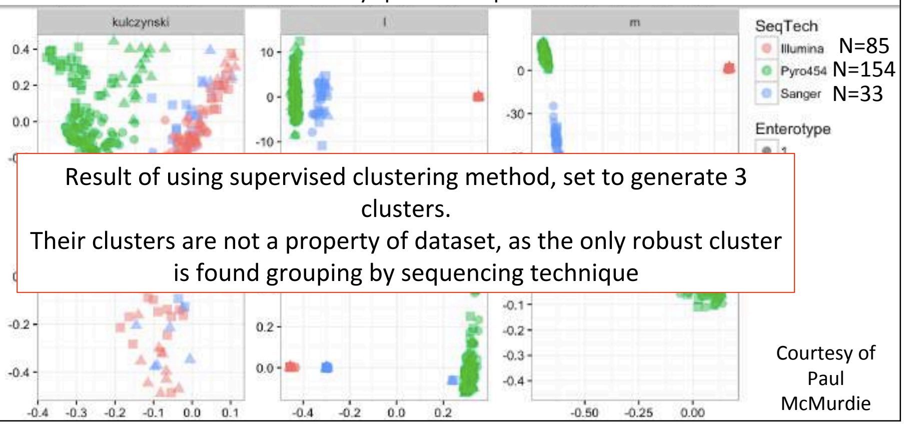
154 pyrosequencing-based 16S sequences

ARTICLE 174 | NATURE | VOL 473 | 12 MAY 2011

doi:10.1038/nature09944

Enterotypes of the human gut microbiome

we identify three robust clusters (referred to as enterotypes hereafter) that are not nation or continent specific ... mostly driven by species composition



Result of using supervised clustering method, set to generate 3 clusters.
Their clusters are not a property of dataset, as the only robust cluster is found grouping by sequencing technique

Courtesy of Paul McMurdie

LETTER 228 | NATURE | VOL 502 | 10 OCTOBER 2013

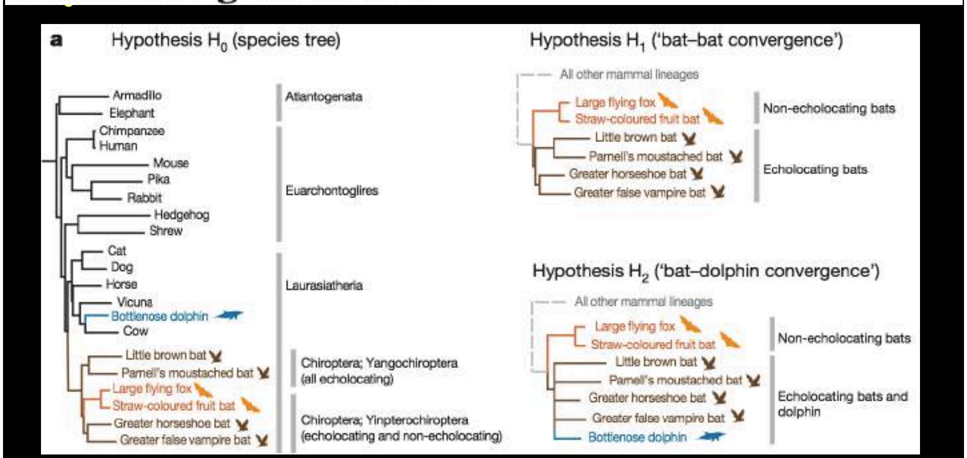
doi:10.1038/nature12511

Genome-wide signatures of convergent evolution in echolocating mammals

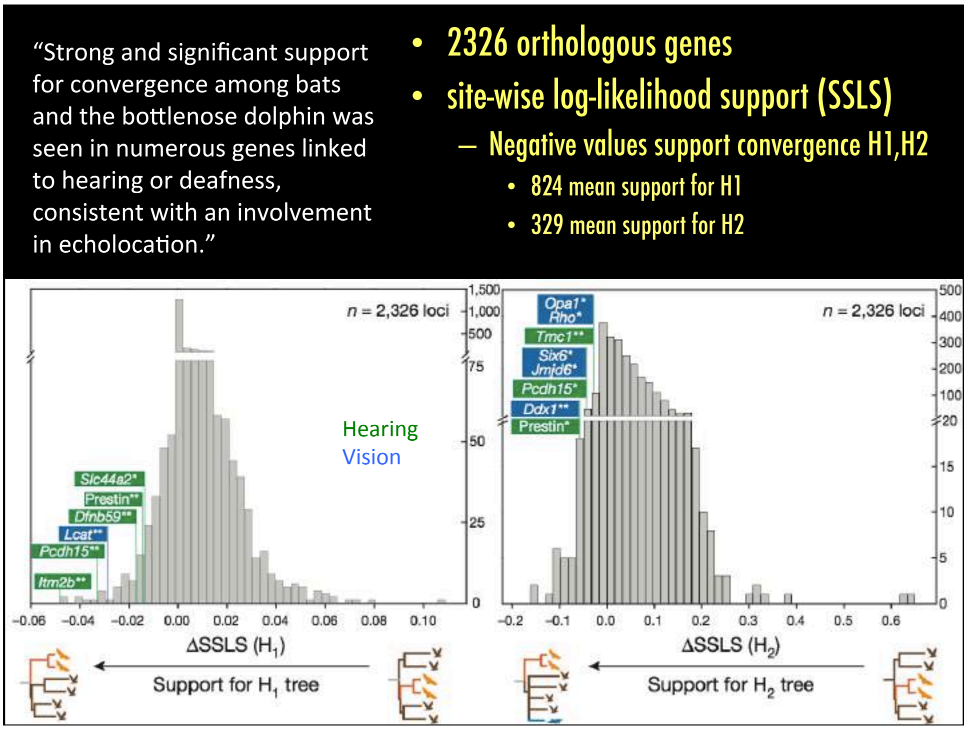
LETTER 228 | NATURE | VOL 502 | 10 OCTOBER 2013

doi:10.1038/nature12511

Genome-wide signatures of convergent evolution in echolocating mammals

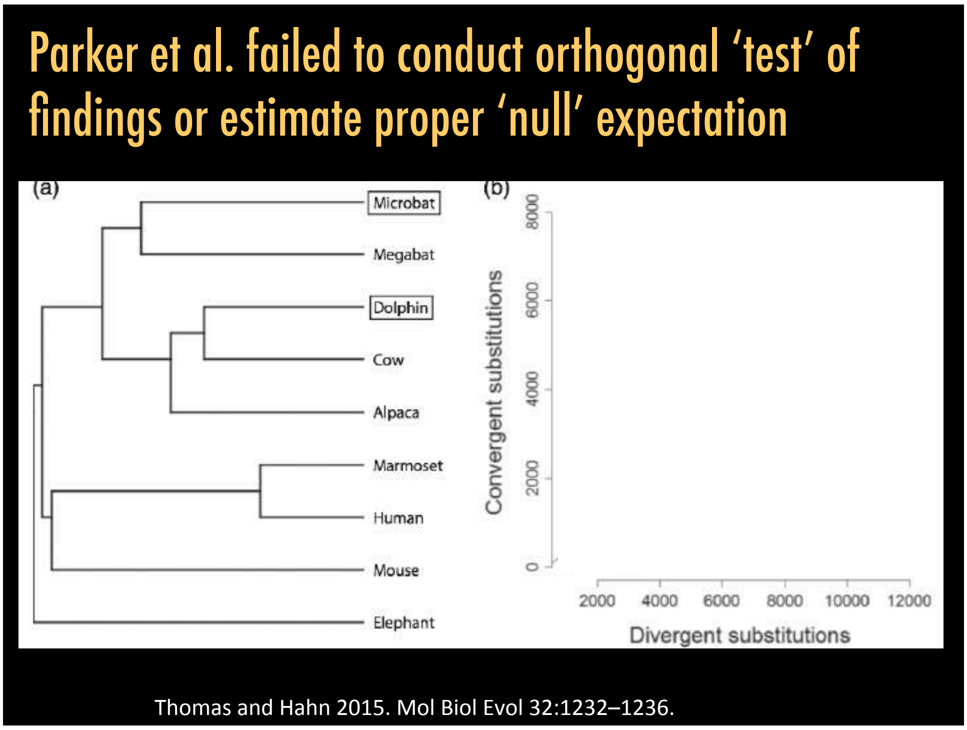


Parker et al. 2013. Nature 502:228–231.



"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

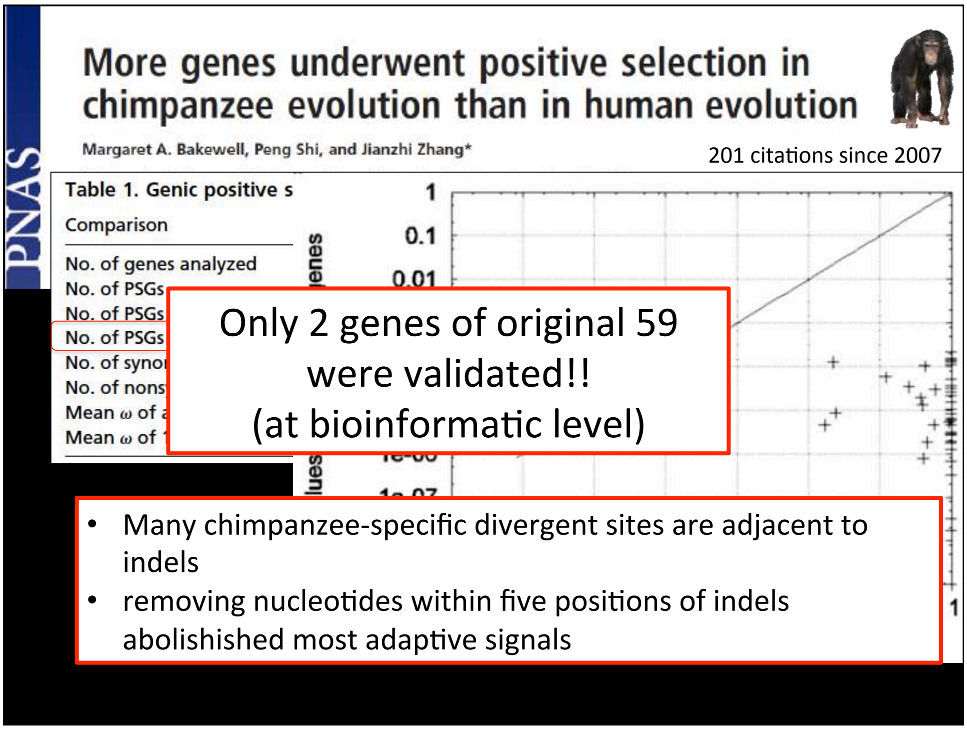


Parker et al. failed to conduct orthogonal 'test' of findings or estimate proper 'null' expectation

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

What makes us difference from chimps?

Is it really just 2%



PNAS

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

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

201 citations since 2007

Table 1. Genic positive s

Comparison

No. of genes analyzed

No. of PSGs

No. of PSGs

No. of PSGs

No. of synor

No. of nons

Mean ω of a

Mean ω of 1

Only 2 genes of original 59 were validated!!
(at bioinformatic level)

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

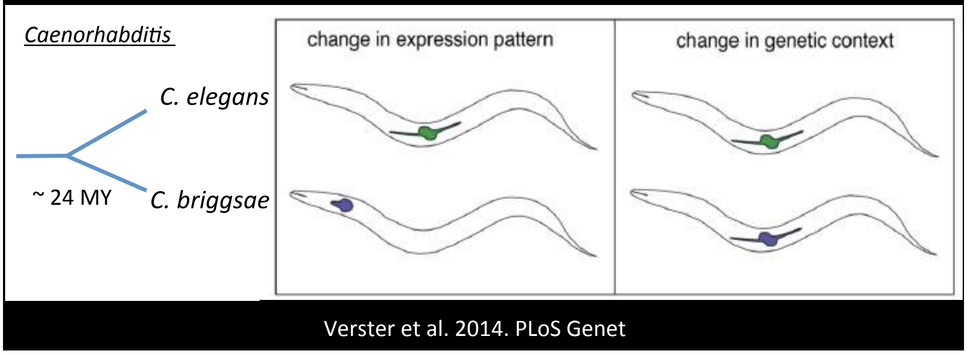
Evolutionary Inference = House of Cards?

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 for1,300 genes in two nematodes
  - TMRA ~24 MYA
  - 7% had divergent phenotypic effects (in lab, etc.)
  - Likely higher in nature



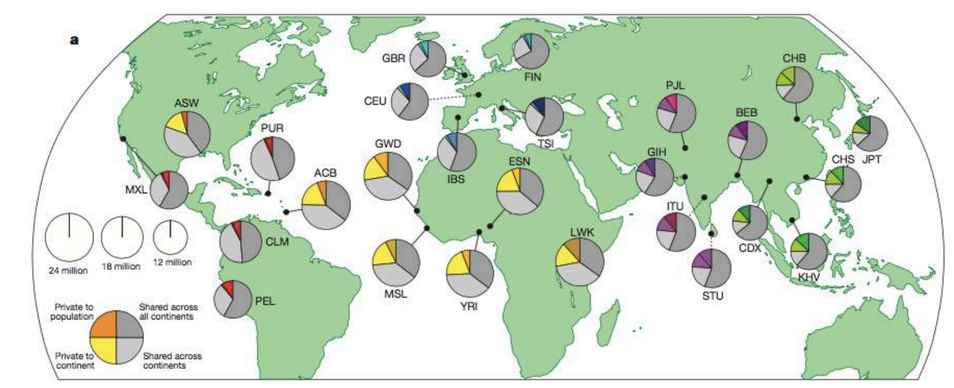
Verster et al. 2014. PLoS Genet

So … how many of you are sequencing a genome?

- What does that mean?
- What kind of genome are you generating?
- What is your question?
  - Short term vs. long term goals?
  - Are these in conflict?

Is there a genome for humans?

Genomes of 2,504 individuals from 26 populations



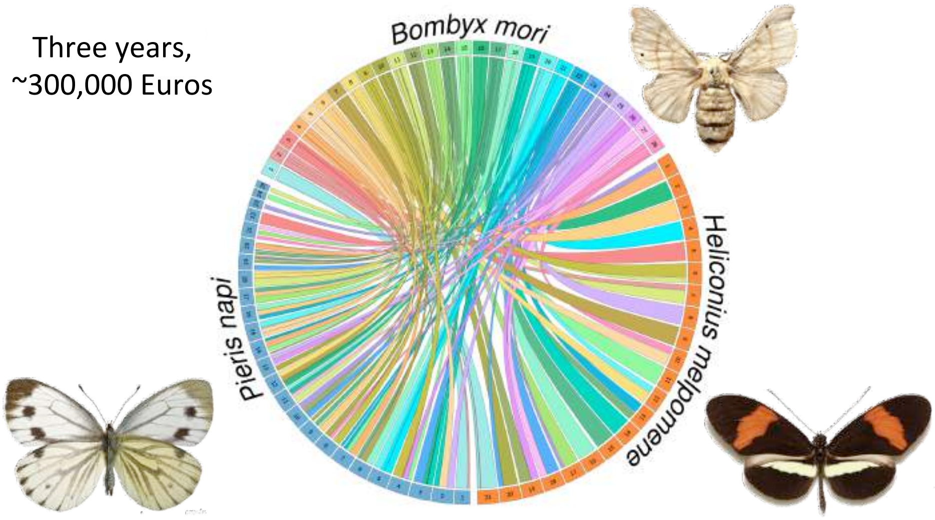
- You differ from references on average at:
  - 4 to 5 M SNPs
  - ~2k structural variants covering ~20 M bp

1000 Genomes Project Consortium (2015) Nature

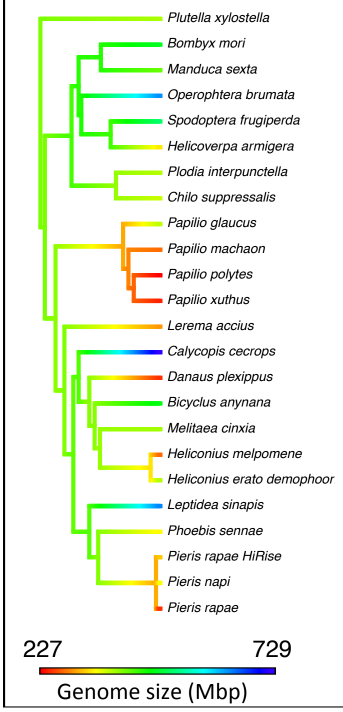
What does this mean

- Most species have lots of genomic polymorhism
  - SNPs are just the tip of the iceburg, lots of structural changes
  - Characterizing all the variation is very expensive
- But
- Very rarely will your questions require chromosomal level assembly
  - Thus you can get to your answers much faster and cheaper if you generate what you need rather than working for an ideal you don't need

Three years,
~300,000 Euros



Hill et al., in prep.



Published genomes vary dramatically in quality

Which do you need NOW?

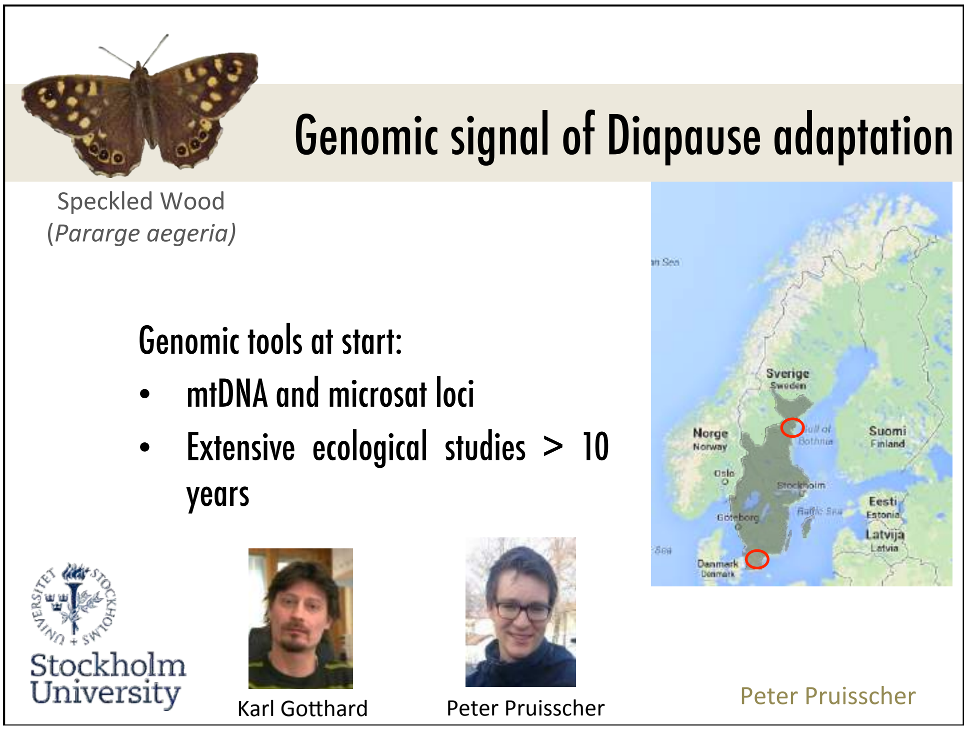
Hill et al., in prep.

Depending on your question

Just sequence lots of genomes

Generate hypotheses

Test them



Genomic signal of Diapause adaptation

Speckled Wood
(*Pararge aegeria*)

Genomic tools at start:

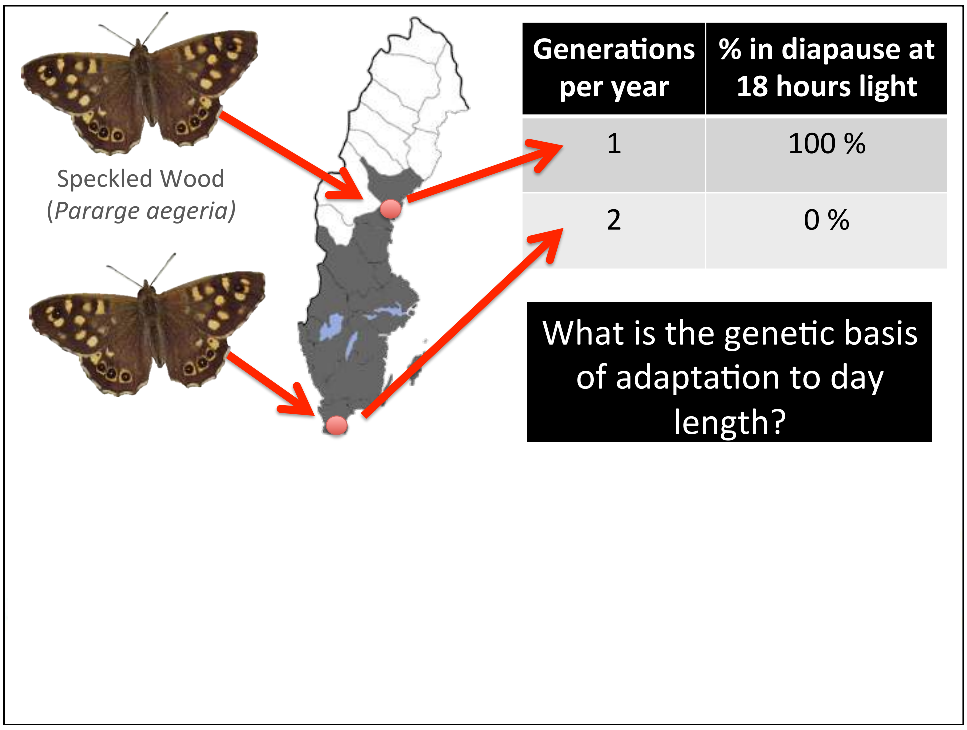
- mtDNA and microsat loci
- Extensive ecological studies > 10 years

Stockholm University

Karl Gotthard

Peter Pruisscher

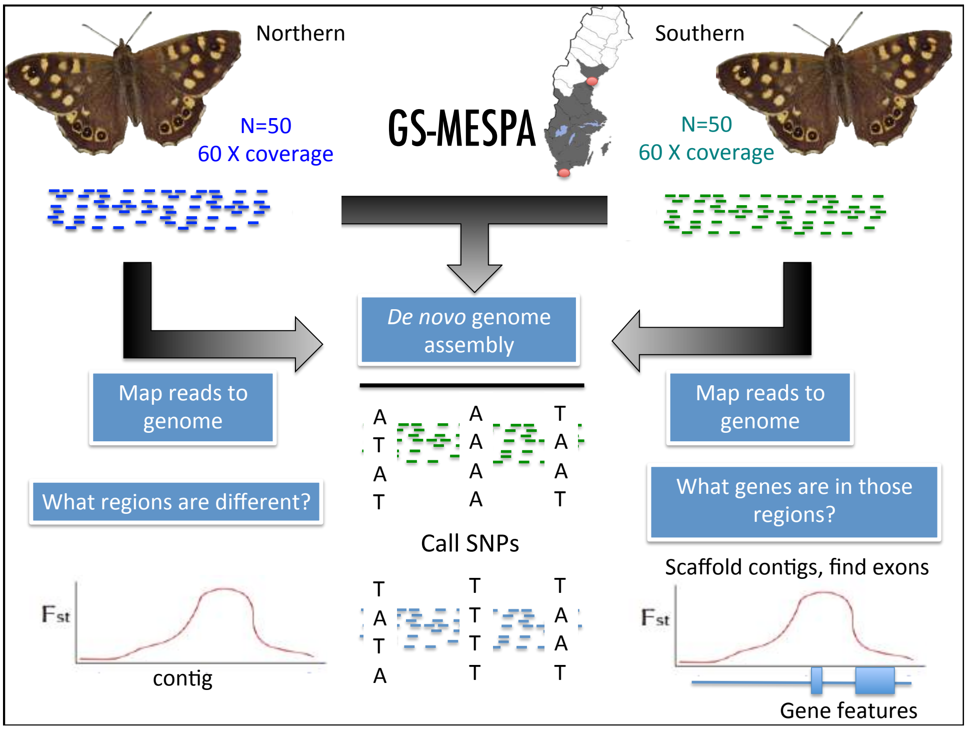
Peter Pruisscher



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

Northern

N=50
60 X coverage

Southern

N=50
60 X coverage

De novo genome assembly

Map reads to genome

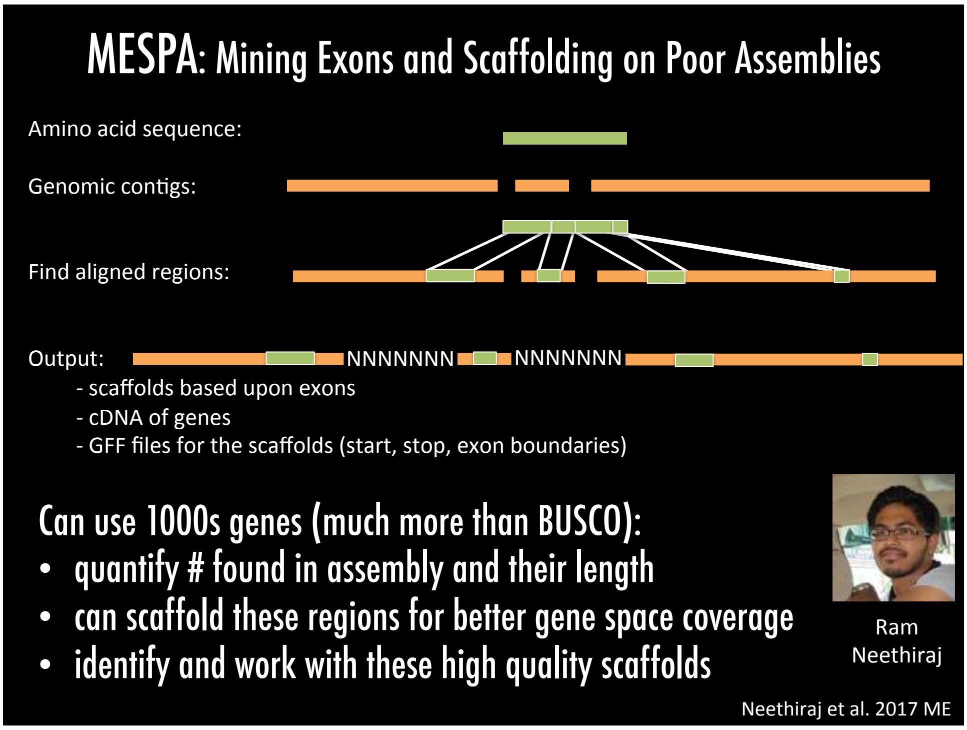
Map reads to genome

What regions are different?

Call SNPs

What genes are in those regions?

Scaffold contigs, find exons



MESPA: Mining Exons and Scaffolding on Poor Assemblies

Amino acid sequence:

Genomic contigs:

Find aligned regions:

Output: NNNNNNN NNNNNNN

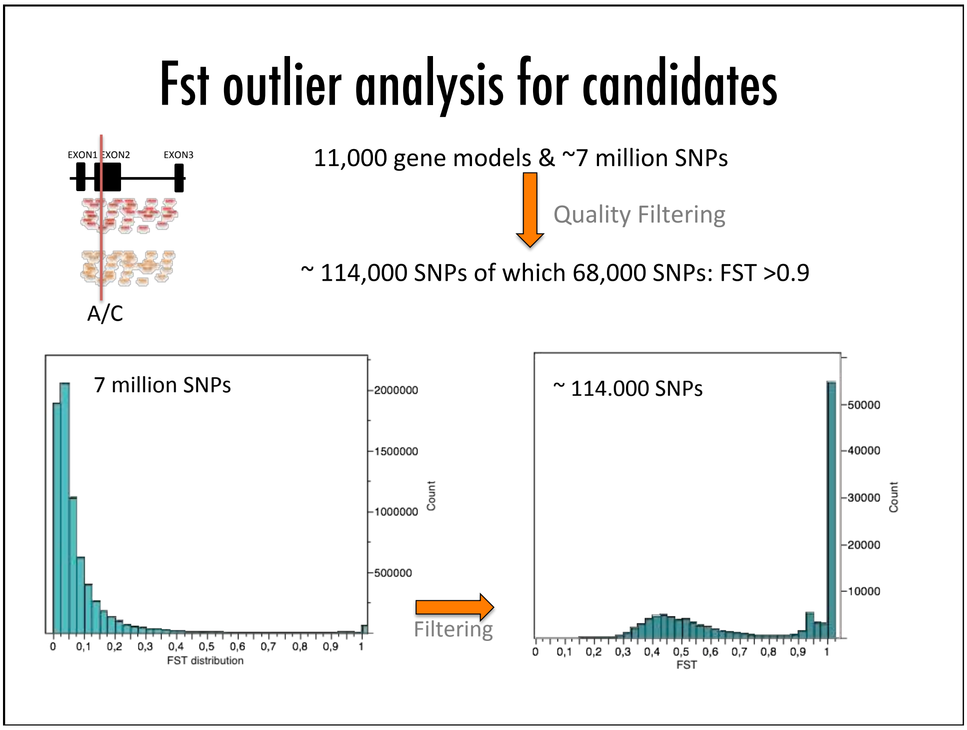
- scaffolds based upon exons
- cDNA of genes
- GFF files for the scaffolds (start, stop, exon boundaries)

Can use 1000s genes (much more than BUSCO):

- quantify # found in assembly and their length
- can scaffold these regions for better gene space coverage
- identify and work with these high quality scaffolds

Ram Neethiraj

Neethiraj et al. 2017 ME



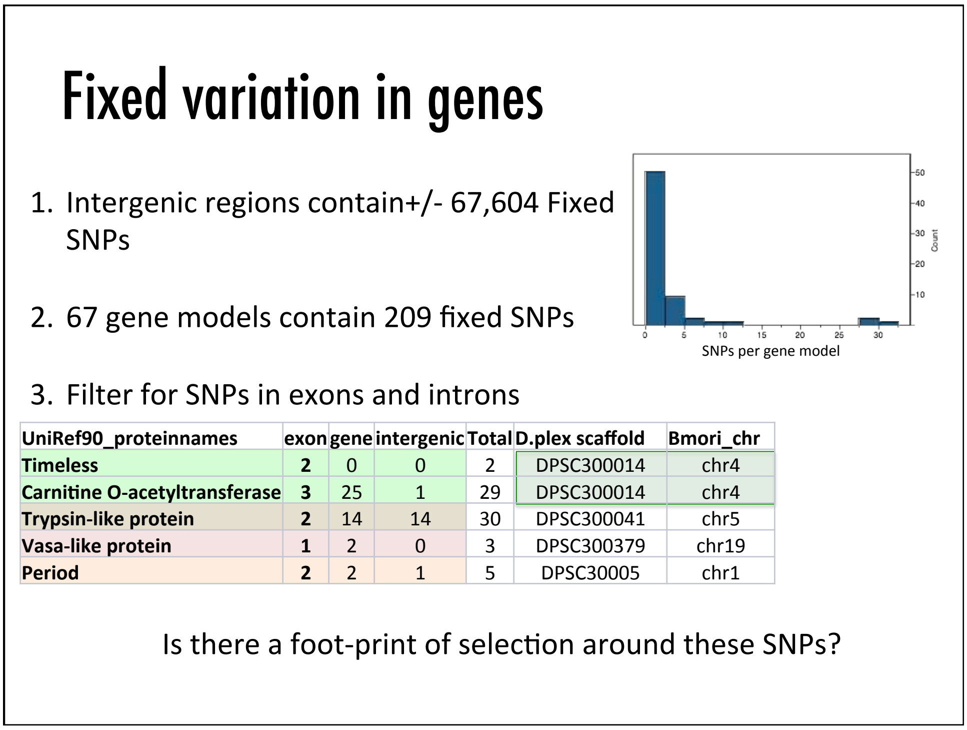
Fst outlier analysis for candidates

11,000 gene models & ~7 million SNPs

Quality Filtering

~ 114,000 SNPs of which 68,000 SNPs: FST >0.9

A/C

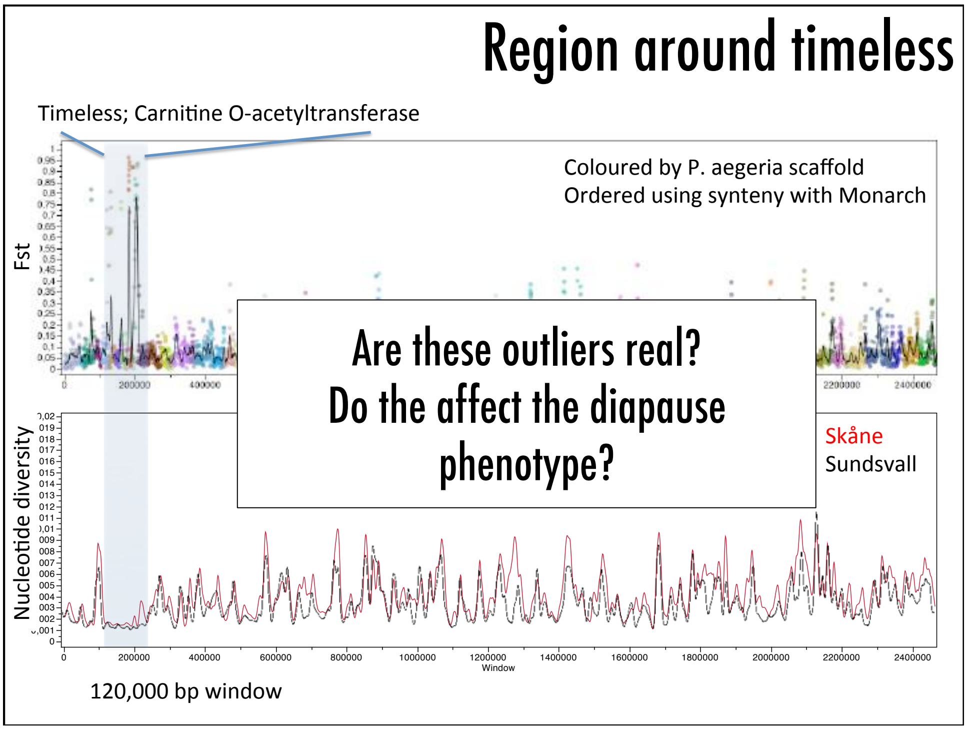


Fixed variation in genes

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	DPSC30005	chr1

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



Region around timeless

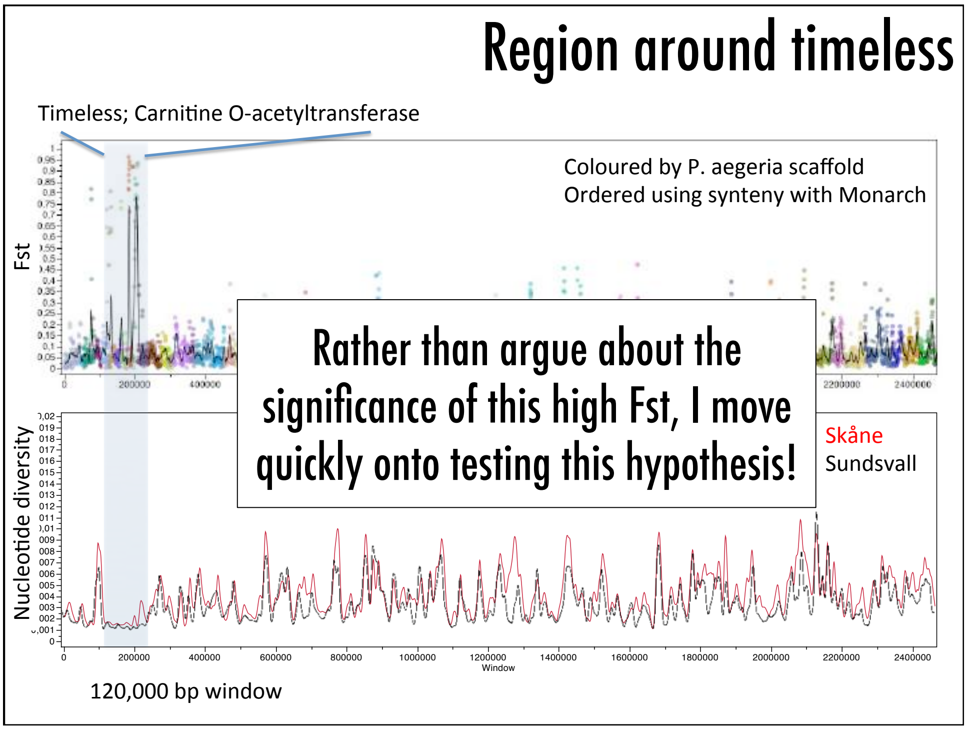
Timeless; Carnitine O-acetyltransferase

Coloured by P. aegeria scaffold
Ordered using synteny with Monarch

Are these outliers real?
Do the affect the diapause phenotype?

Skåne
Sundsvall

120,000 bp window



Region around timeless

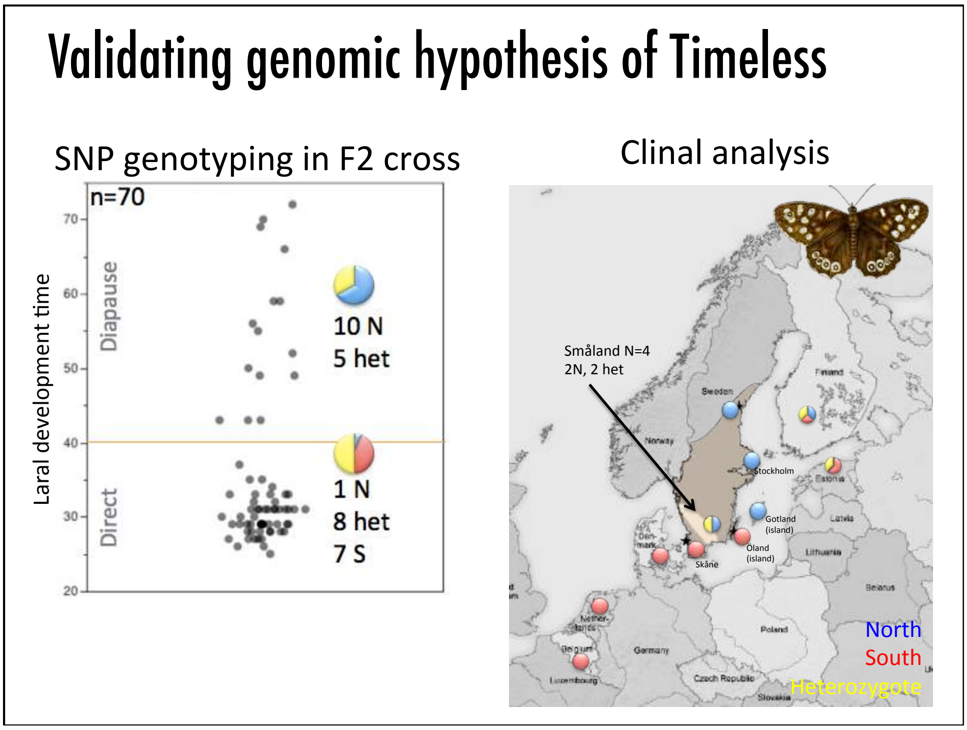
Timeless; Carnitine O-acetyltransferase

Coloured by P. aegeria scaffold
Ordered using synteny with Monarch

Rather than argue about the significance of this high Fst, I move quickly onto testing this hypothesis!

Skåne
Sundsvall

120,000 bp window



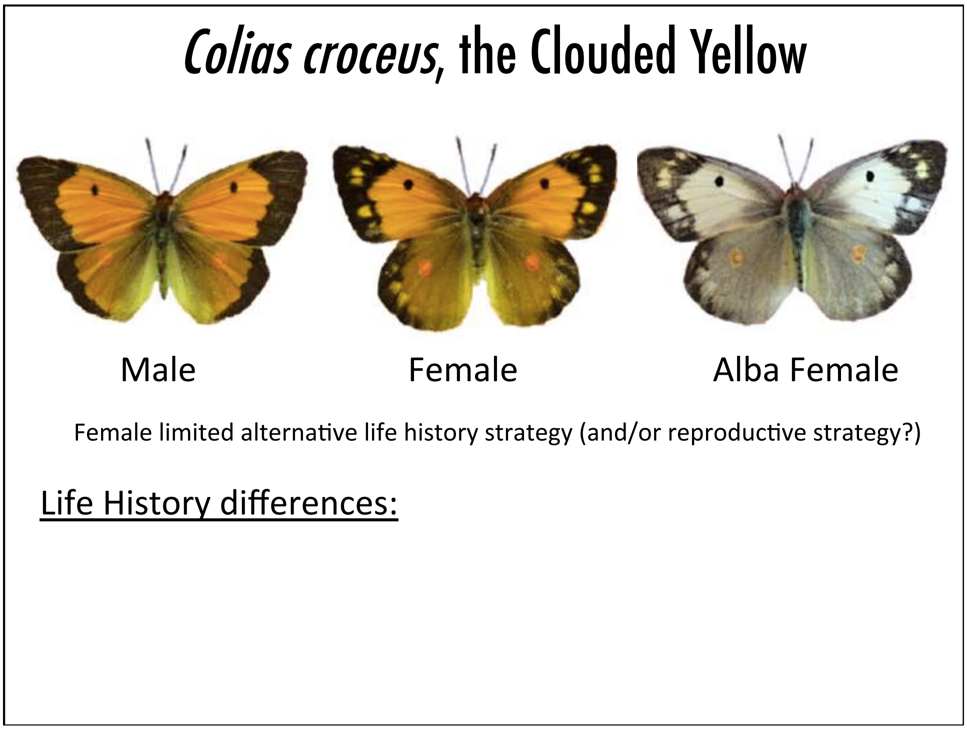
Validating genomic hypothesis of Timeless

SNP genotyping in F2 cross

Clinal analysis

Genomics to hypothesis to validation

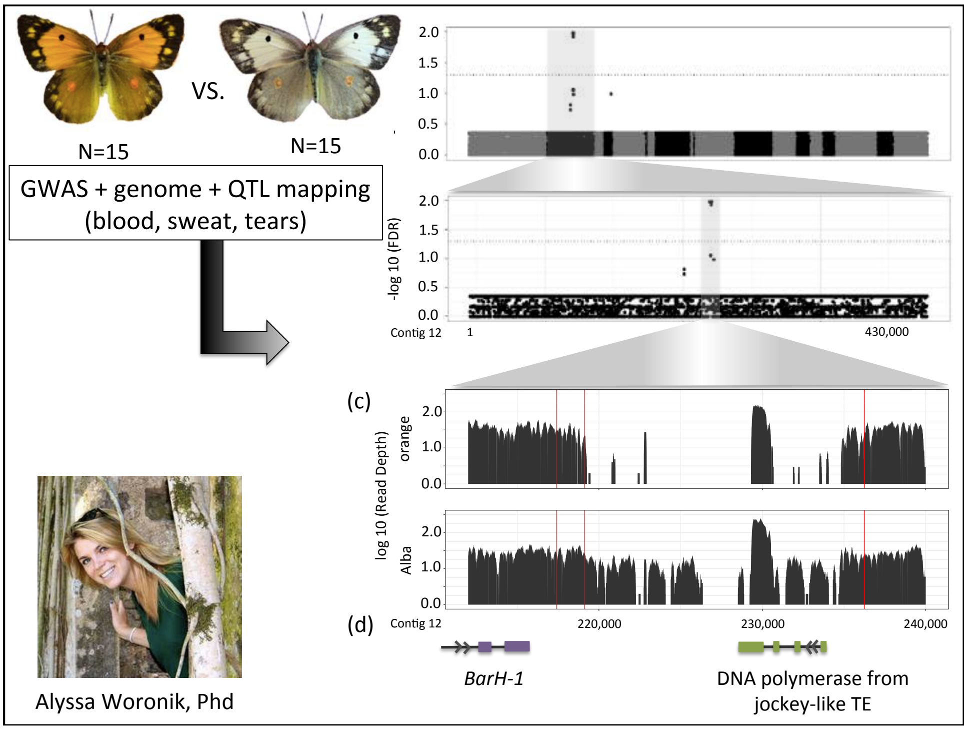
- Use genomics to generate robust hypothesis
  - Orthogonal methods
  - Stong signals
- Validate upwards
  - Use independent biological samples
  - Higher level of biological organization
  - Simultaneously test hypothesis and its generality



Colias croceus, the Clouded Yellow

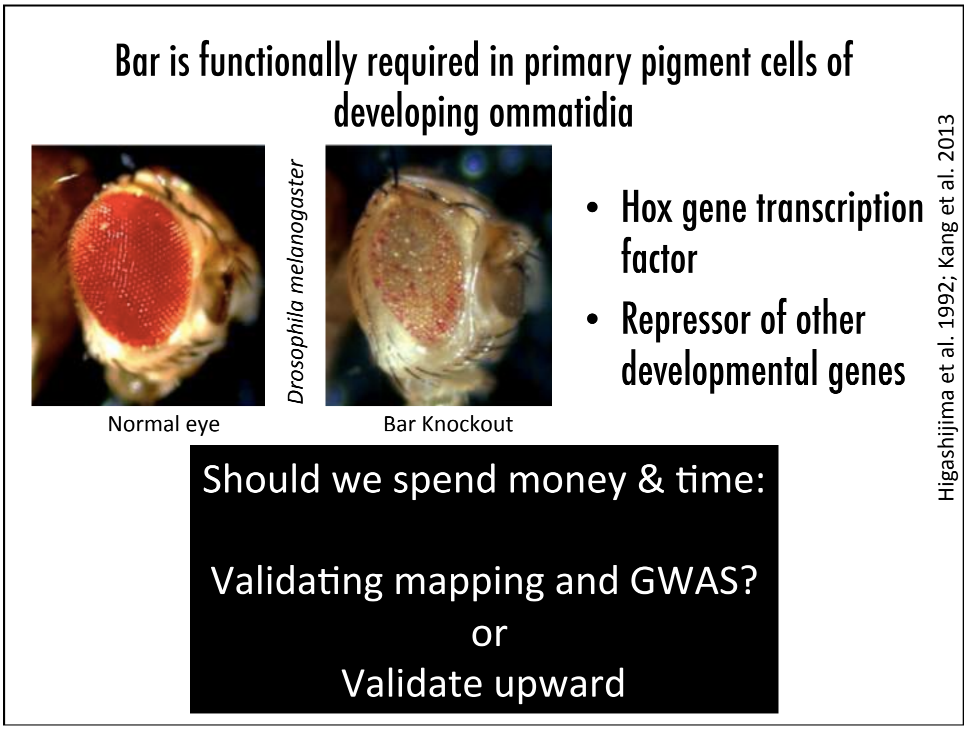
Female limited alternative life history strategy (and/or reproductive strategy?)

Life History differences:



GWAS + genome + QTL mapping
(blood, sweat, tears)

Alyssa Woronik, Phd



Bar is functionally required in primary pigment cells of developing ommatidia

Drosophila melanogaster

Normal eye

Bar Knockout

- Hox gene transcription factor
- Repressor of other developmental genes

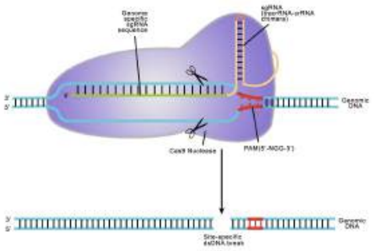
Higashijima et al. 1992; Kang et al. 2013

Should we spend money & time:

Validating mapping and GWAS?
or
Validate upward

Video

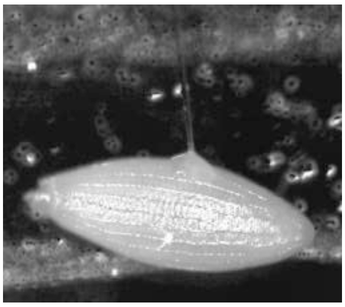
Testing BarH1 hypothesis: CRISPR/Cas9 knockout of Bar


Genome specific sgRNA sequence
sgRNA (tracrRNA-crRNA chimera)
3'
5'
Genomic DNA
Cas9 Nuclease
PAM(5'-NGG-3')
Site-specific dsDNA break

Allows us to remove function of BarH1

Up and working in 2 months

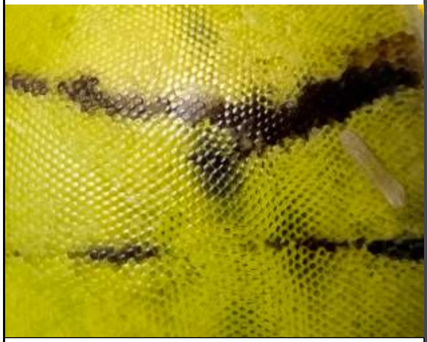
Masters project in lab



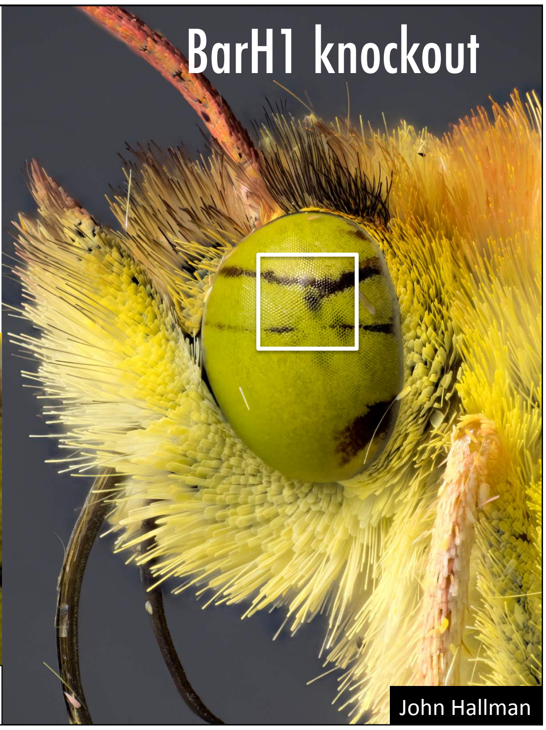
Injection of Cas9 protein + guideRNA into Colias croceus egg

Developmental defects:

- Lack of pigment formation within ommatidia
- Equivalent to *Drosophila* phenotype

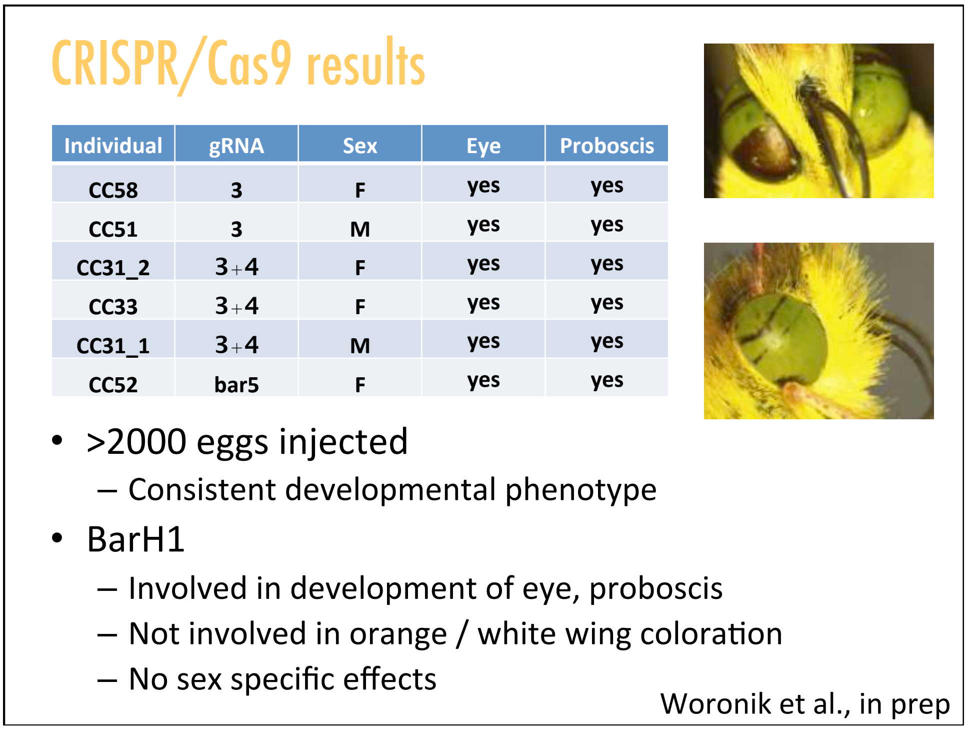


Phenotypes observed using 2 separate gRNA constructs, awaiting PCR validation



BarH1 knockout

John Hallman

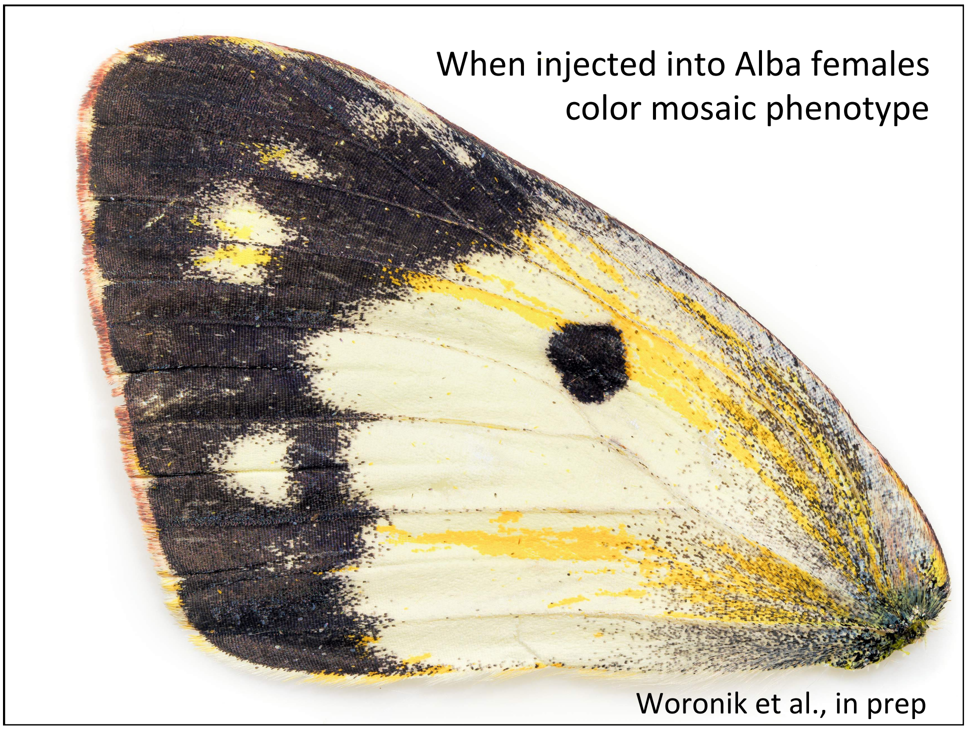


CRISPR/Cas9 results

Individual	gRNA	Sex	Eye	Proboscis
CC58	3	F	yes	yes
CC51	3	M	yes	yes
CC31_2	3+4	F	yes	yes
CC33	3+4	F	yes	yes
CC31_1	3+4	M	yes	yes
CC52	bar5	F	yes	yes

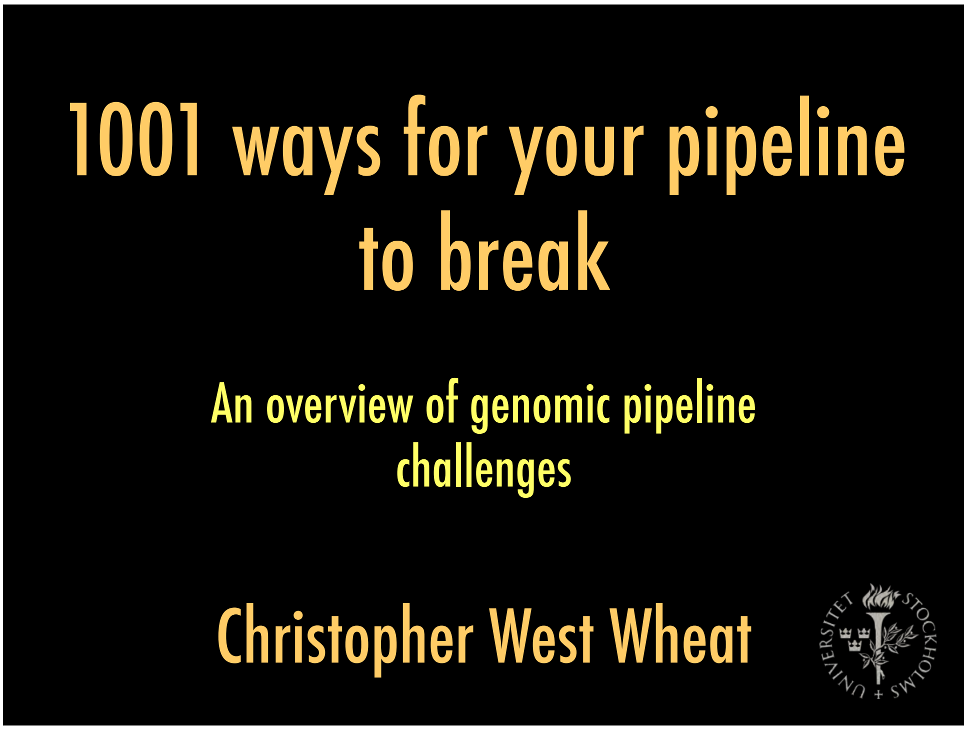
- >2000 eggs injected
  - Consistent developmental phenotype
- BarH1
  - Involved in development of eye, proboscis
  - Not involved in orange / white wing coloration
  - No sex specific effects

Woronik et al., in prep



When injected into Alba females
color mosaic phenotype

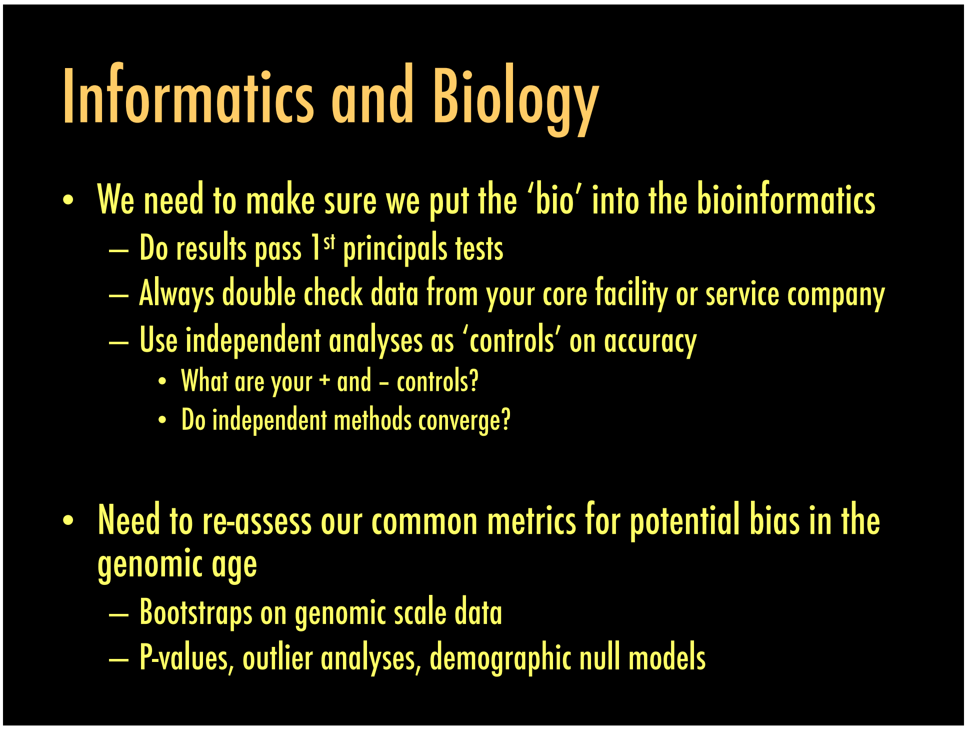
Woronik et al., in prep



1001 ways for your pipeline to break

An overview of genomic pipeline challenges

Christopher West Wheat



Informatics and Biology

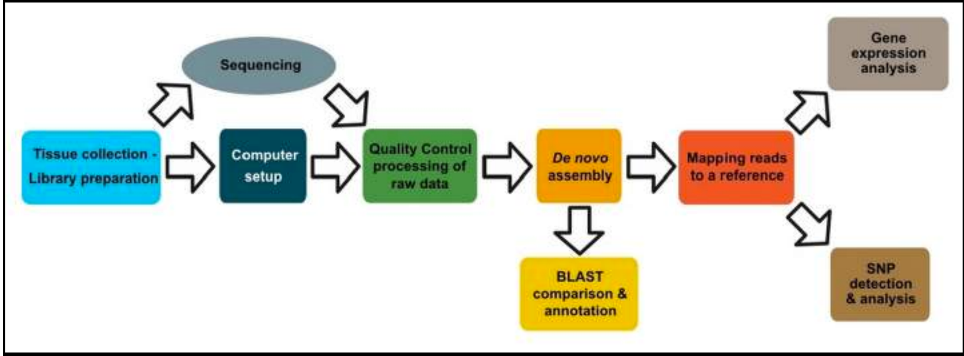
- We need to make sure we put the 'bio' into the bioinformatics
  - Do results pass 1st principals tests
  - Always double check data from your core facility or service company
  - Use independent analyses as 'controls' on accuracy
    - What are your + and – controls?
    - Do independent methods converge?
- 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



Sequencing

Tissue collection - Library preparation

Computer setup

Quality Control processing of raw data

De novo assembly

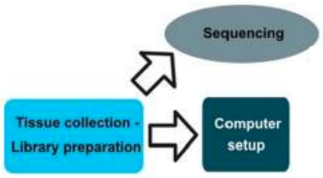
BLAST comparison & annotation

Mapping reads to a reference

Gene expression analysis

SNP detection & analysis

Pipeline Overview



How can I study my data using open source?

How much RAM do I need?

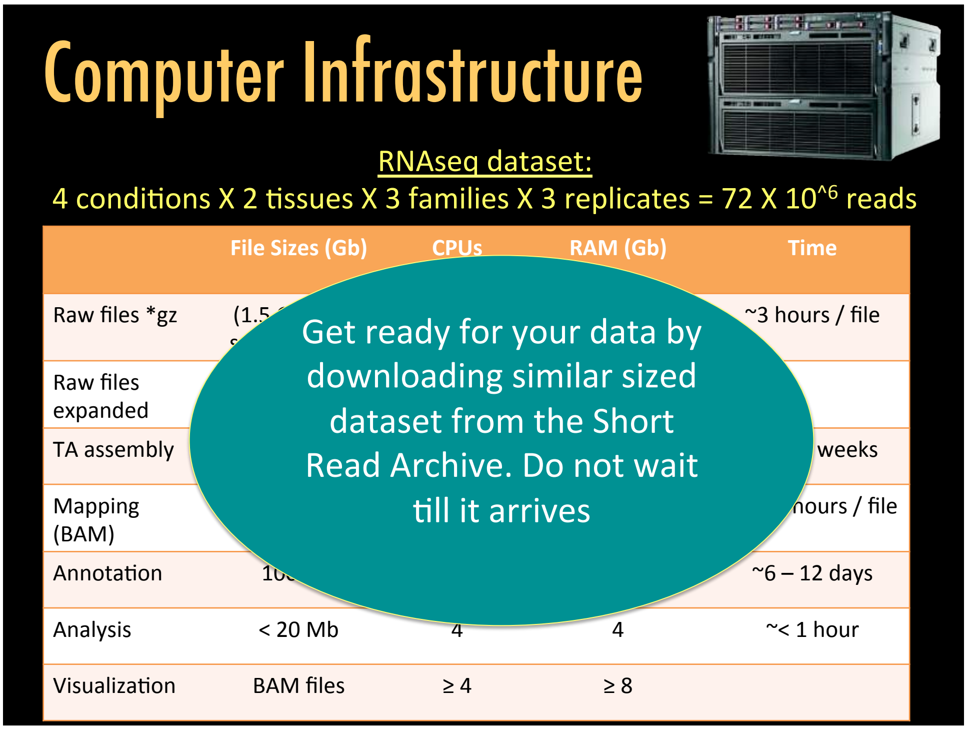
Are 16 cores enough?

Can I use my laptop?

What software & how do I get it?

Why Linux?

How much HD space is needed?



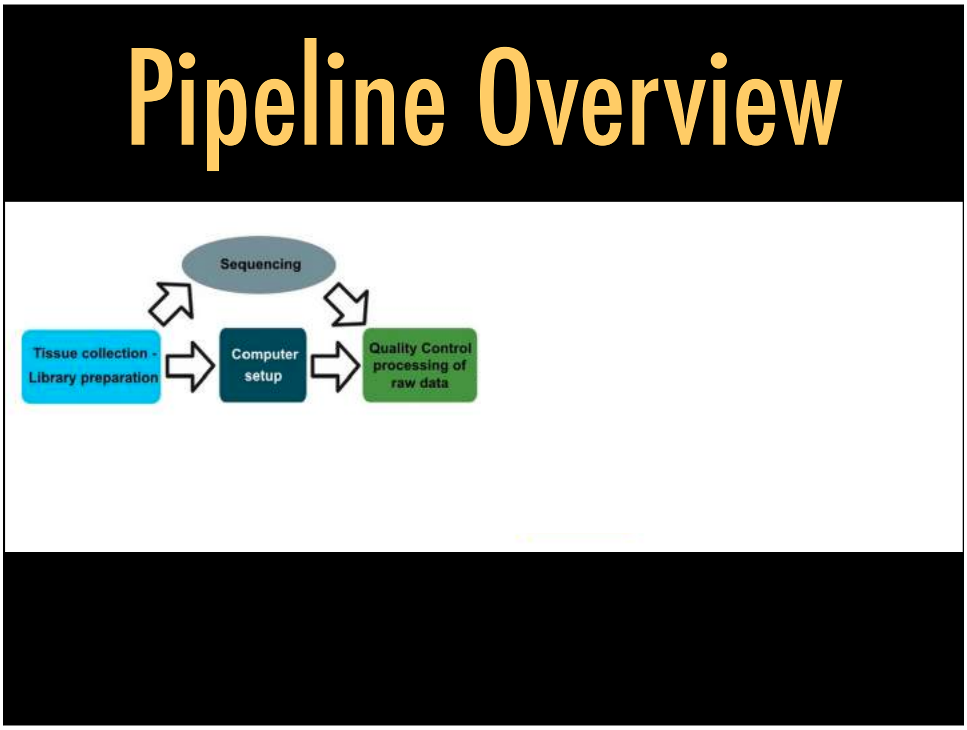
Computer Infrastructure

RNAseq dataset:

4 conditions X 2 tissues X 3 families X 3 replicates = 72 X 10$^{\wedge 6}$ reads

	File Sizes (Gb)	CPUs	RAM (Gb)	Time
Raw files *gz	(1.5			~3 hours / file
Raw files expanded				
TA assembly				weeks
Mapping (BAM)				hours / file
Annotation	10			~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

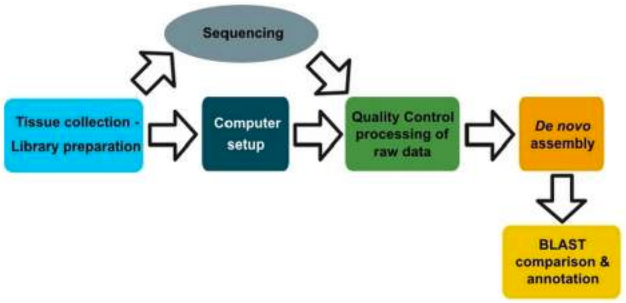
Sequencing
Tissue collection - Library preparation
Computer setup
Quality Control processing of raw data

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



Sequencing

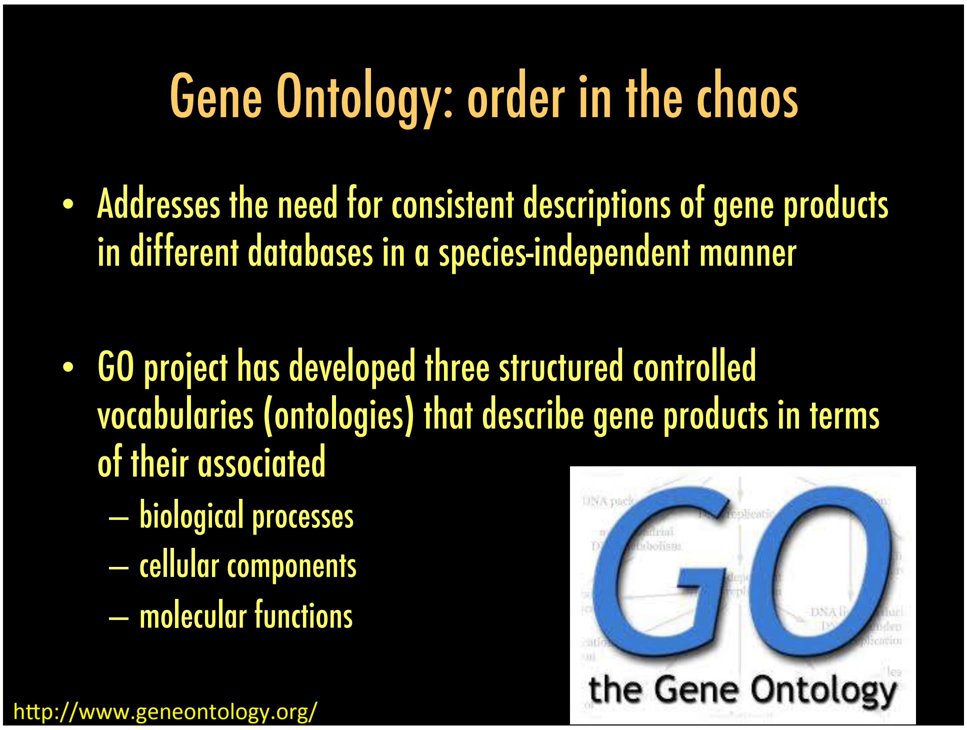
Tissue collection - Library preparation

Computer setup

Quality Control processing of raw data

De novo assembly

BLAST comparison & annotation

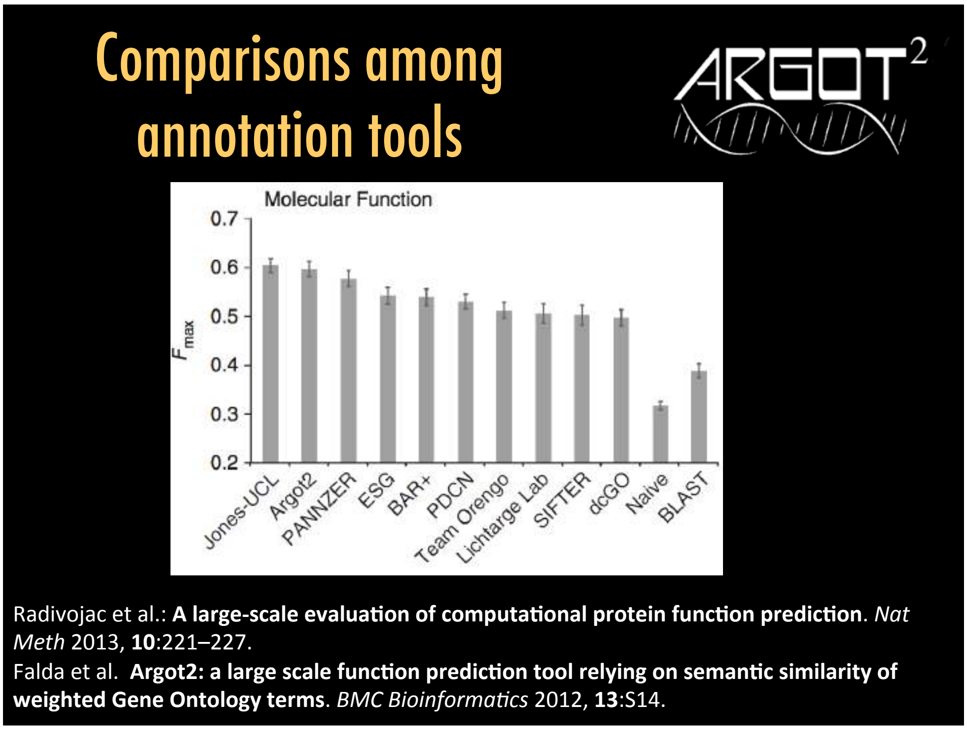


Gene Ontology: order in the chaos

- 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

GO
the Gene Ontology

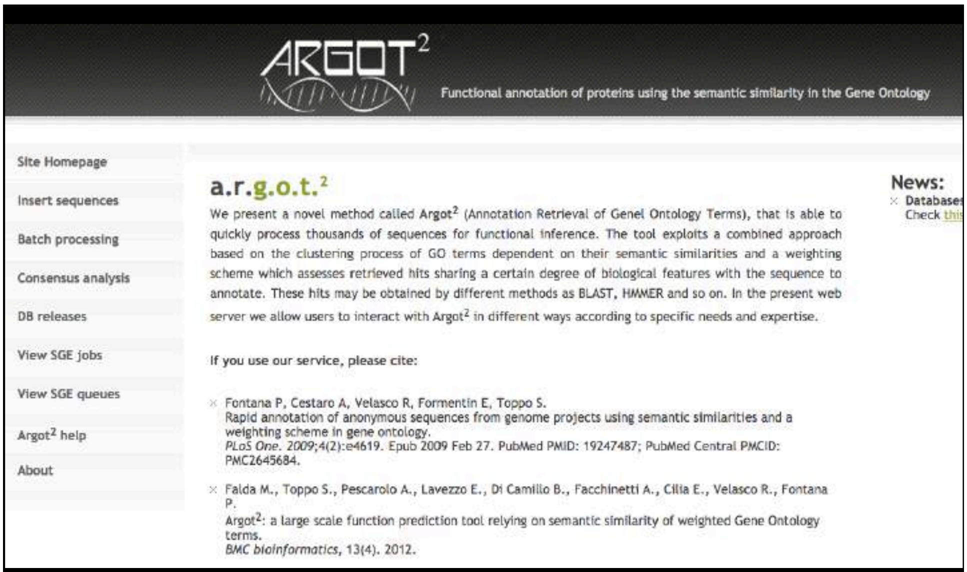
http://www.geneontology.org/



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.



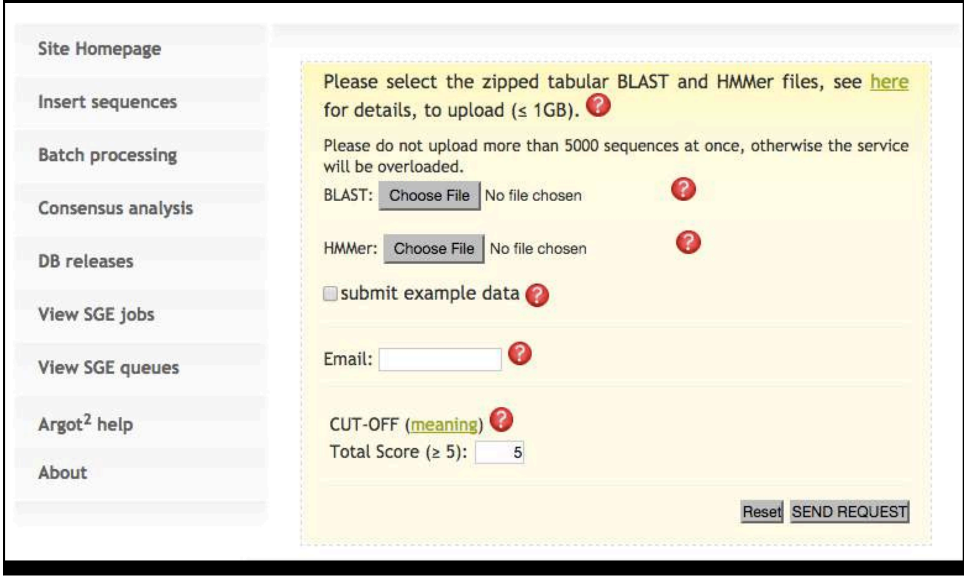
a.r.g.o.t.^{2}

We present a novel method called Argot^2 (Annotation Retrieval of Genel 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 Argot^2 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.
Argot^2: a large scale function prediction tool relying on semantic similarity of weighted Gene Ontology terms.
BMC bioinformatics, 13(4). 2012.

Batch processing for GO terms



Please select the zipped tabular BLAST and HMMer files, see here for details, to upload (≤ 1GB). ?

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

BLAST: Choose File No file chosen ?

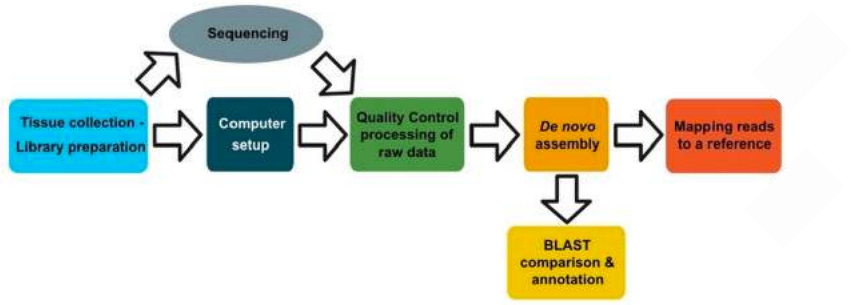
HMMer: Choose File No file chosen ?

□ submit example data ?

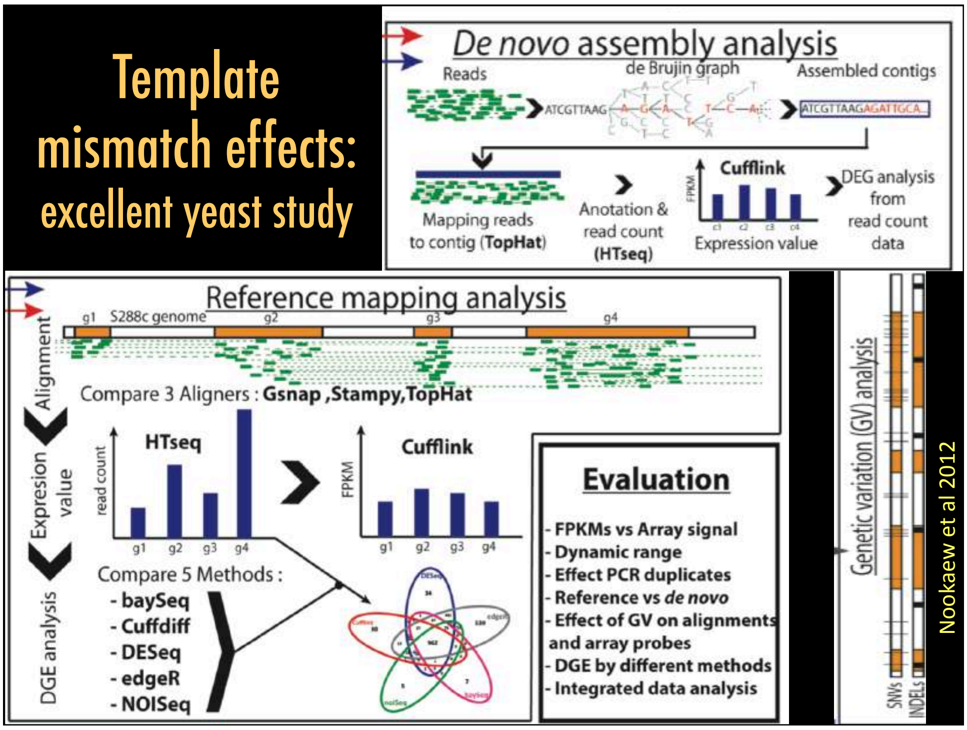
Email: ?

CUT-OFF (meaning) ?
Total Score (≥ 5): 5

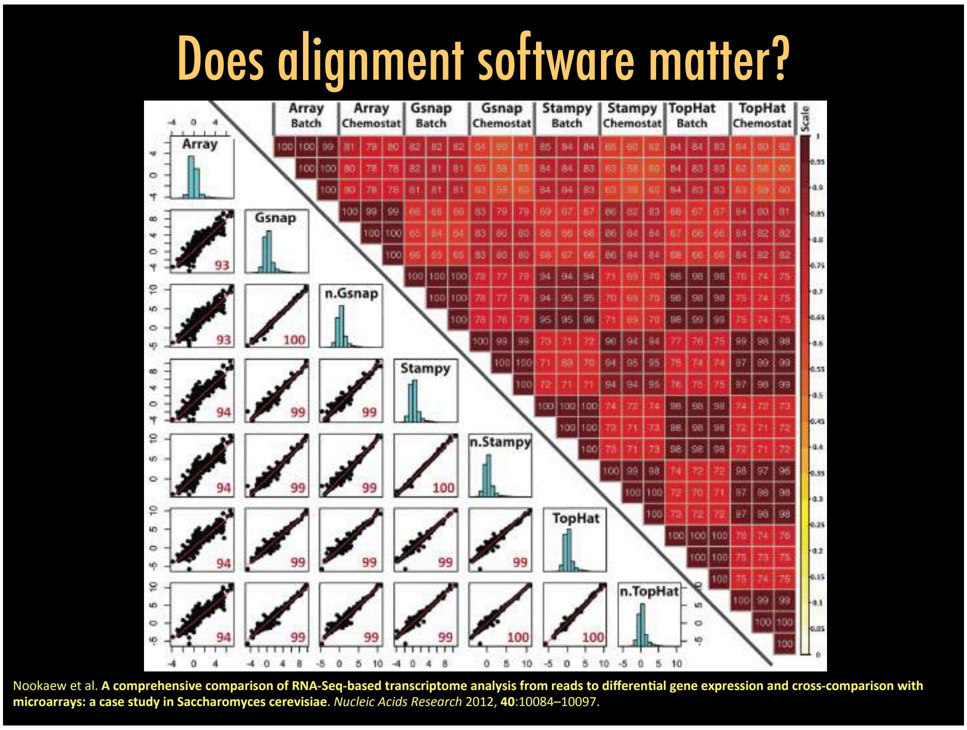
Pipeline Overview


Sequencing
Tissue collection - Library preparation
Computer setup
Quality Control processing of raw data
De novo assembly
Mapping reads to a reference
BLAST comparison & annotation

Template mismatch effects: excellent yeast study

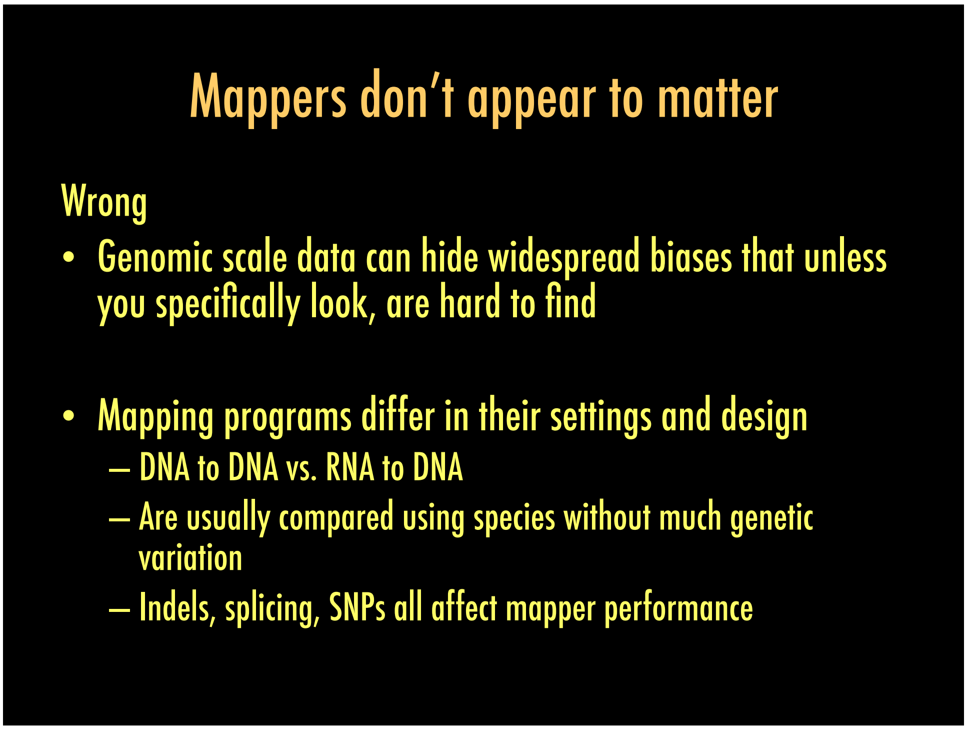


Nookaew et al 2012



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.



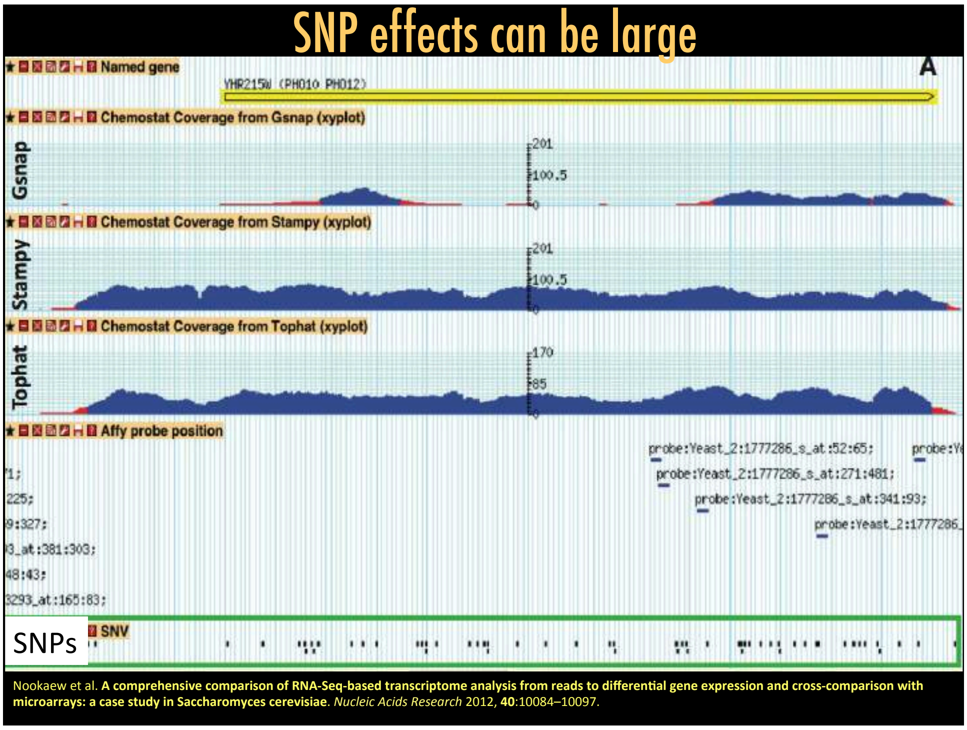
Mappers don't appear to matter

Wrong

- Genomic scale data can hide widespread biases that unless you specifically look, are hard to find

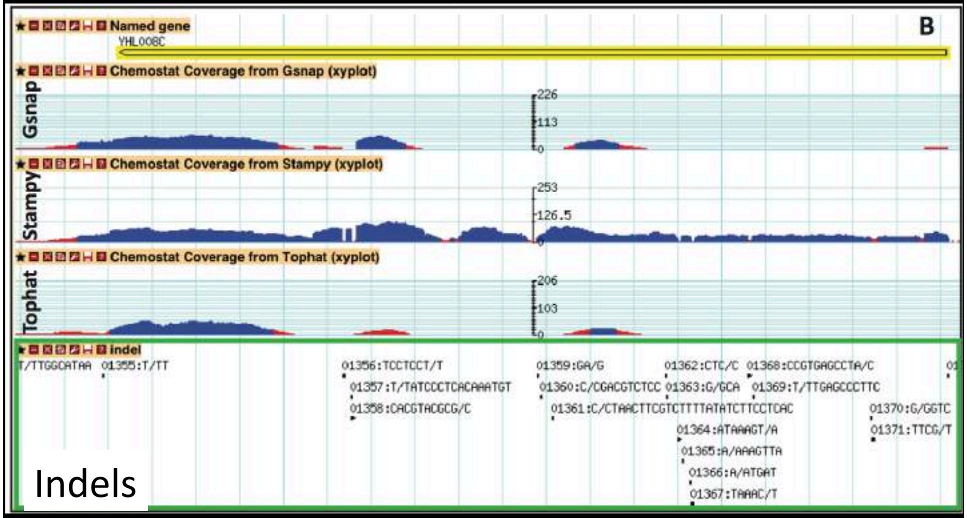
- Mapping programs differ in their settings and design
  - DNA to DNA vs. RNA to DNA
  - Are usually compared using species without much genetic variation
  - Indels, splicing, SNPs all affect mapper performance

SNP effects can be large

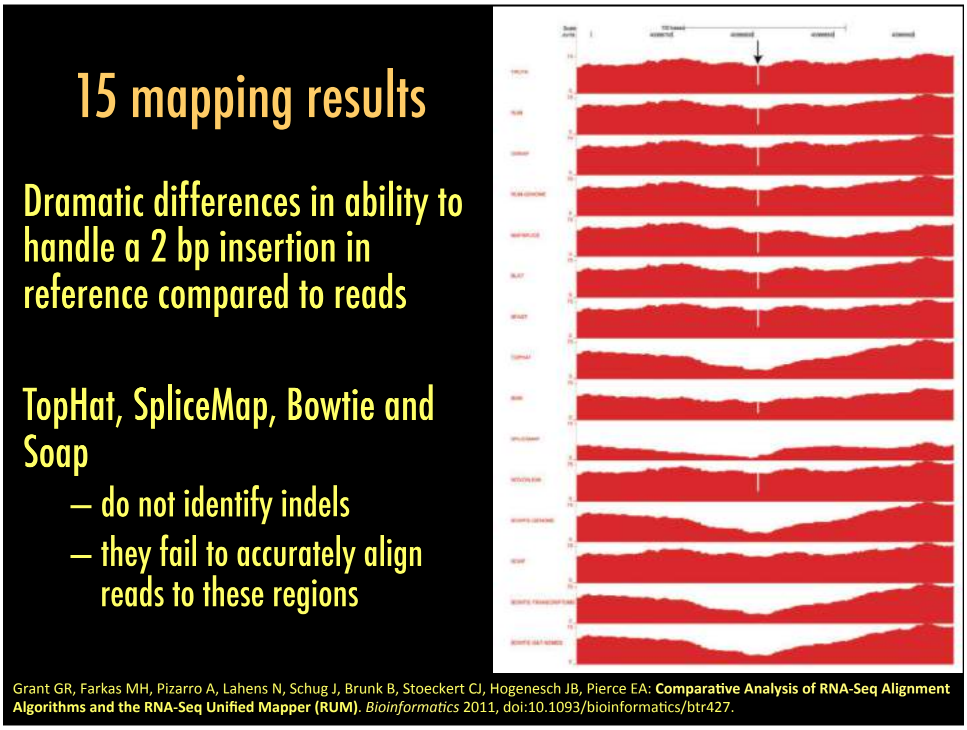

A
Named gene
YHR215W (PHO10 PHO12)
Chemostat Coverage from Gsnap (xyplot)
Gsnap
201
100.5
0
Chemostat Coverage from Stampy (xyplot)
Stampy
201
100.5
0
Chemostat Coverage from Tophat (xyplot)
Tophat
170
85
0
Affy probe position
probe:Yeast_2:1777286_s_at:52:65;
probe:Ye
probe:Yeast_2:1777286_s_at:271:481;
probe:Yeast_2:1777286_s_at:341:93;
probe:Yeast_2:1777286_
1;
225;
9:327;
3_at:381:303;
48:43;
3293_at:165:83;
SNV
SNPs

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



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.



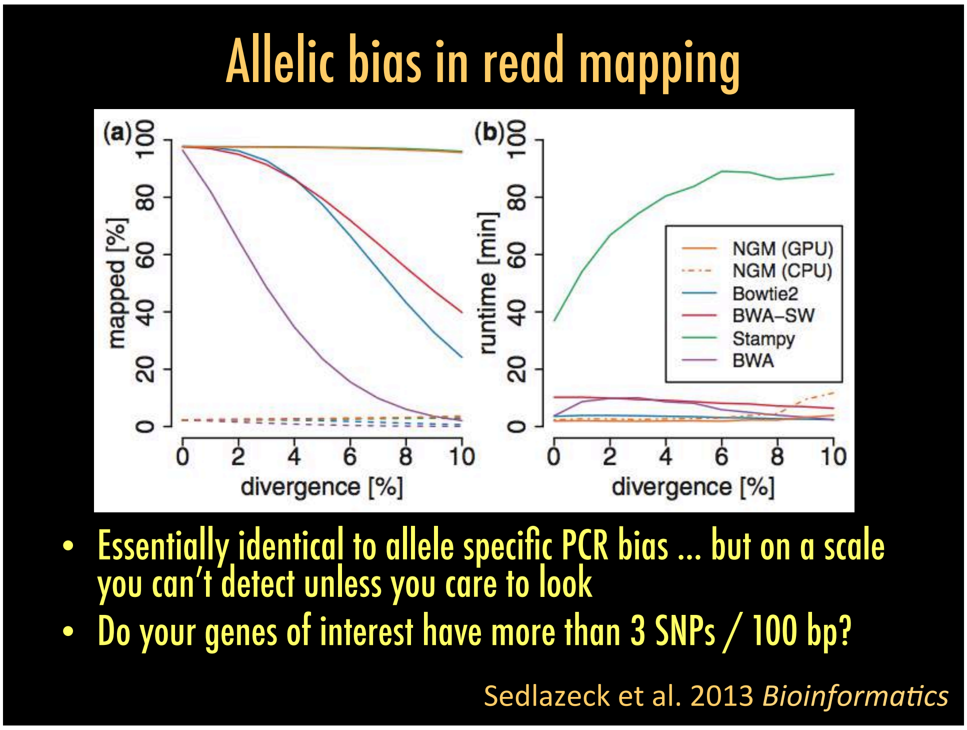
15 mapping results

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

Grant GR, Farkas MH, Pizarro A, Lahens N, Schug J, Brunk B, Stoeckert CJ, Hogenesch JB, Pierce EA: **Comparative Analysis of RNA-Seq Alignment Algorithms and the RNA-Seq Unified Mapper (RUM)**. *Bioinformatics* 2011, doi:10.1093/bioinformatics/btr427.

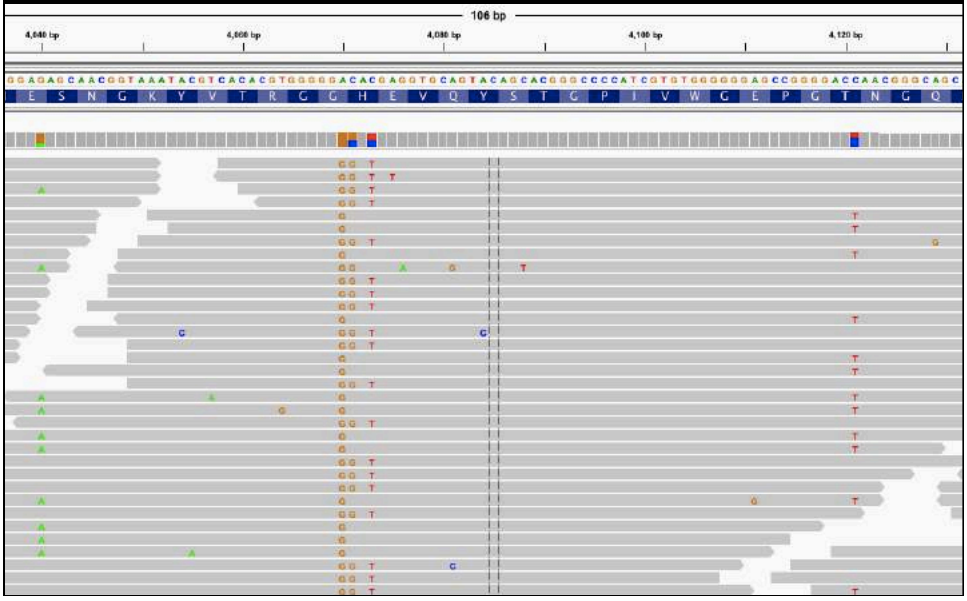


Allelic bias in read mapping

- 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?

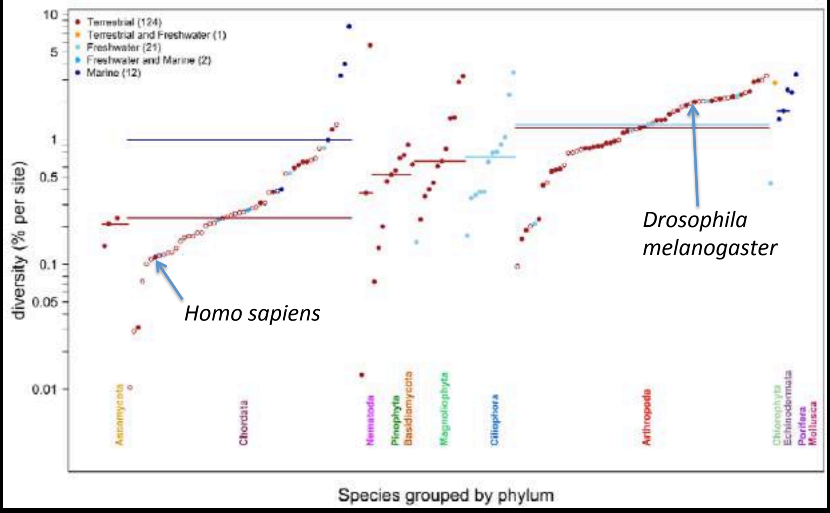
Sedlazeck et al. 2013 *Bioinformatics*

100 bp window with 4 – 5 SNPs differing from reference


106 bp
4,040 bp
4,060 bp
4,080 bp
4,100 bp
4,120 bp
E S N G K Y V T R G G H E V Q Y S T G P I V W G E P G T N G Q

Mapping reads in outbred species

Average genome polymorphism levels (ignores indels)

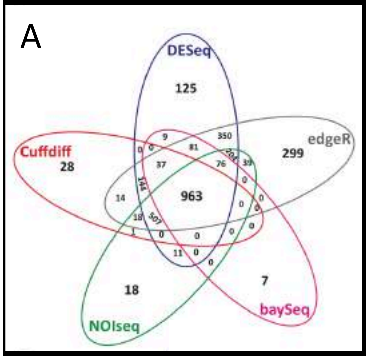


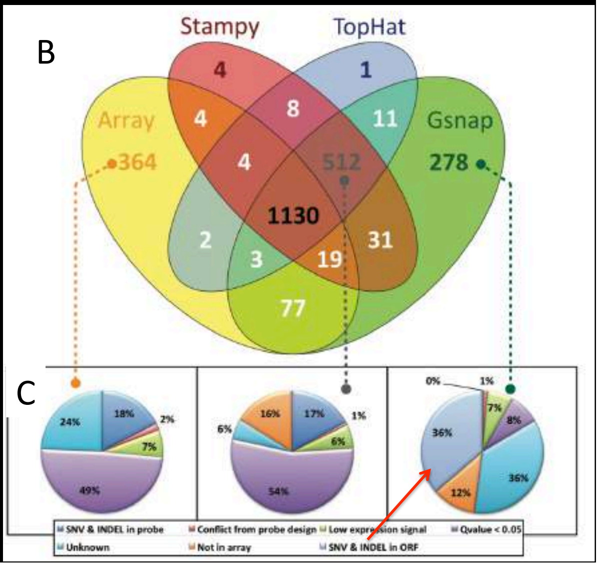
Homo sapiens

Drosophila melanogaster

Leffler *et al.* 2012 *Plos Biol*

Sig. expression differences by method





A: Stampy mapping

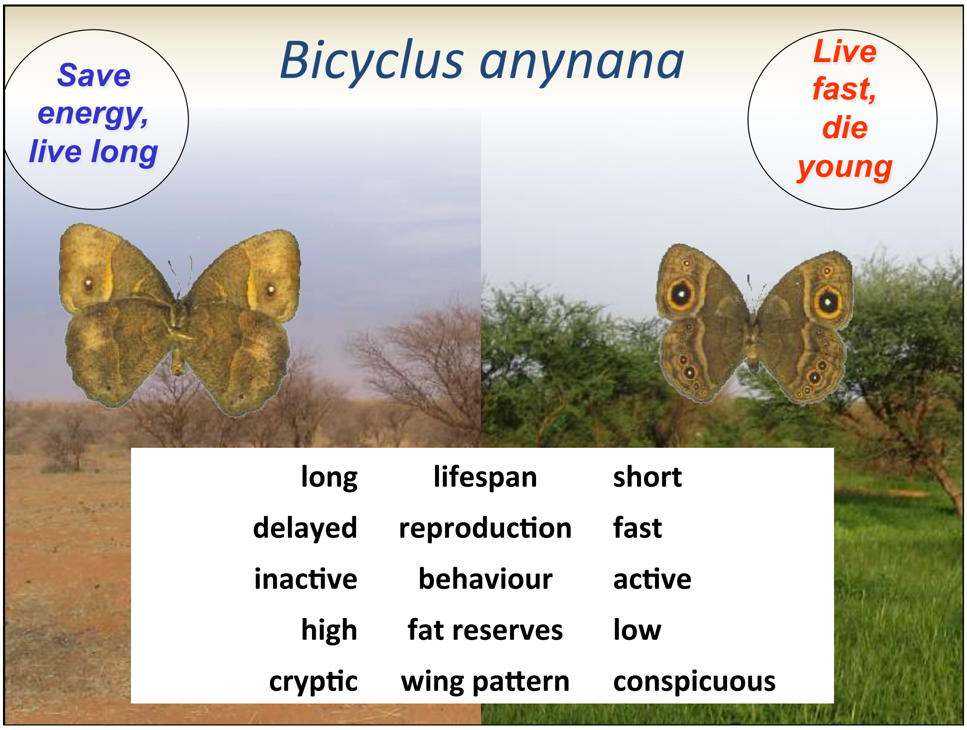
B: Cuffdiff analysis

C: Likely error source

RNA-Seq

Real world example

2 factor analysis with family effects

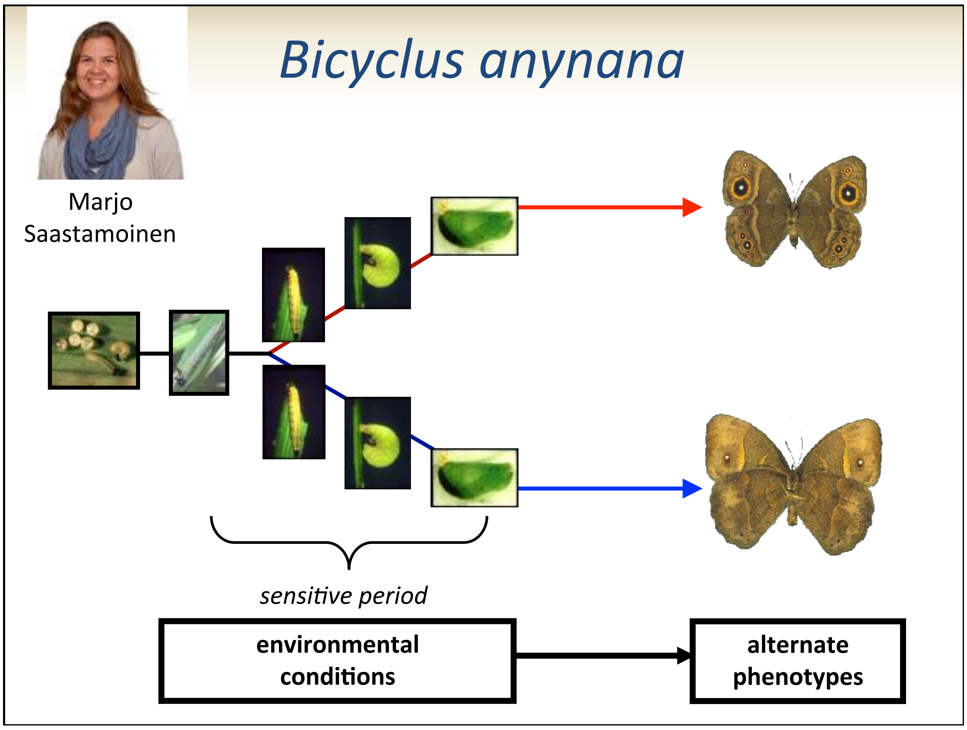


Bicyclus anynana

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

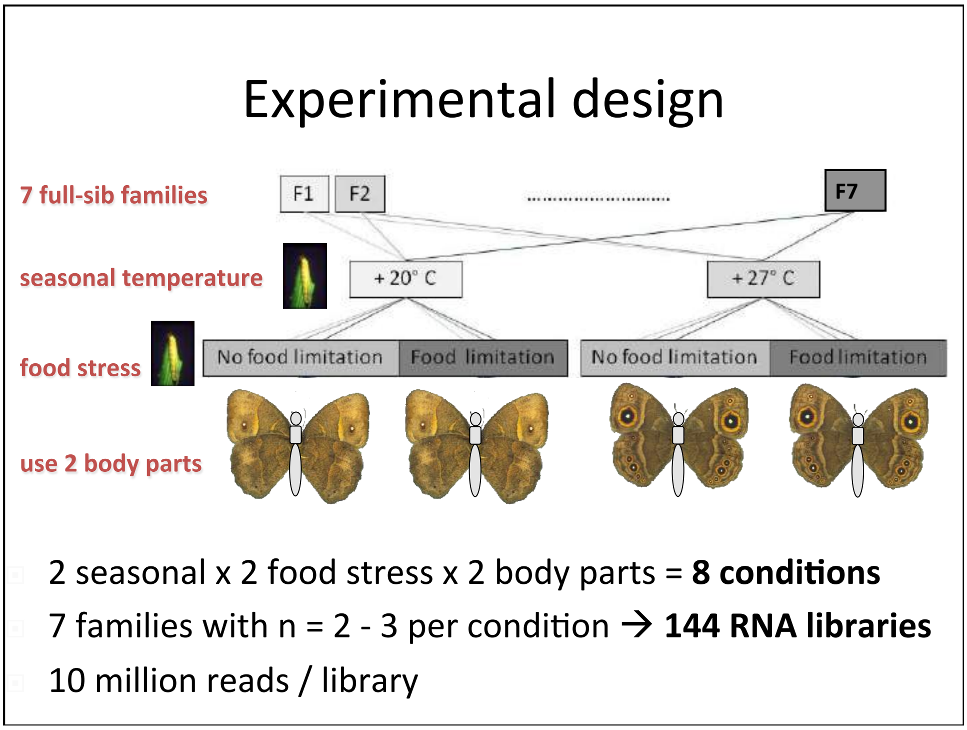


Bicyclus anynana

Marjo Saastamoinen

sensitive period

environmental conditions → alternate phenotypes



Experimental design

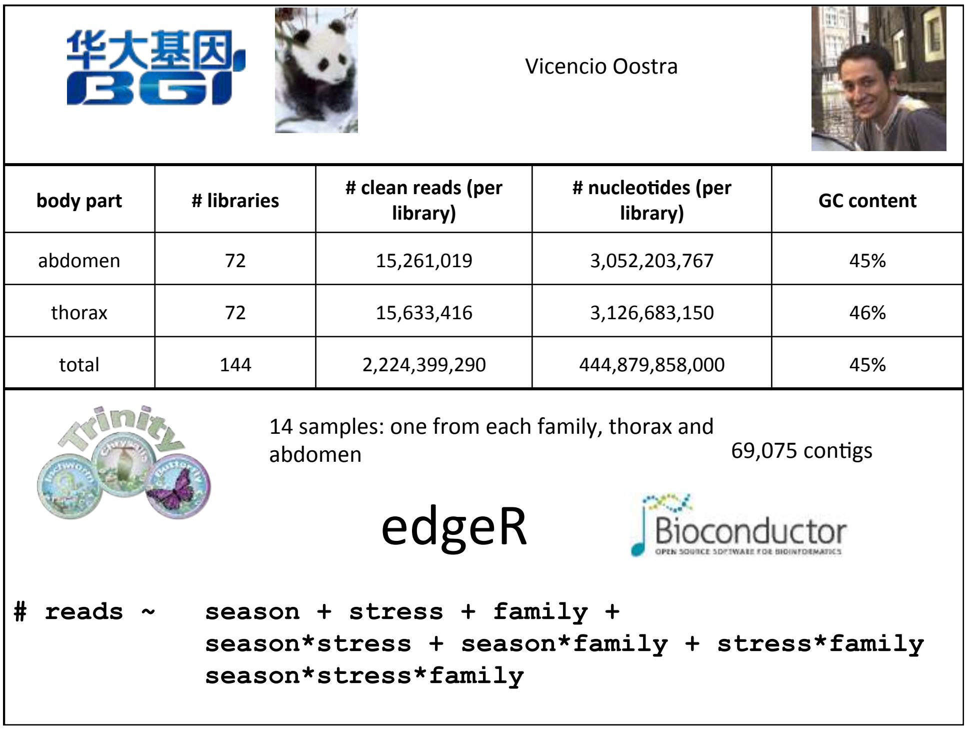
7 full-sib families

seasonal temperature

food stress

use 2 body parts

- 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%

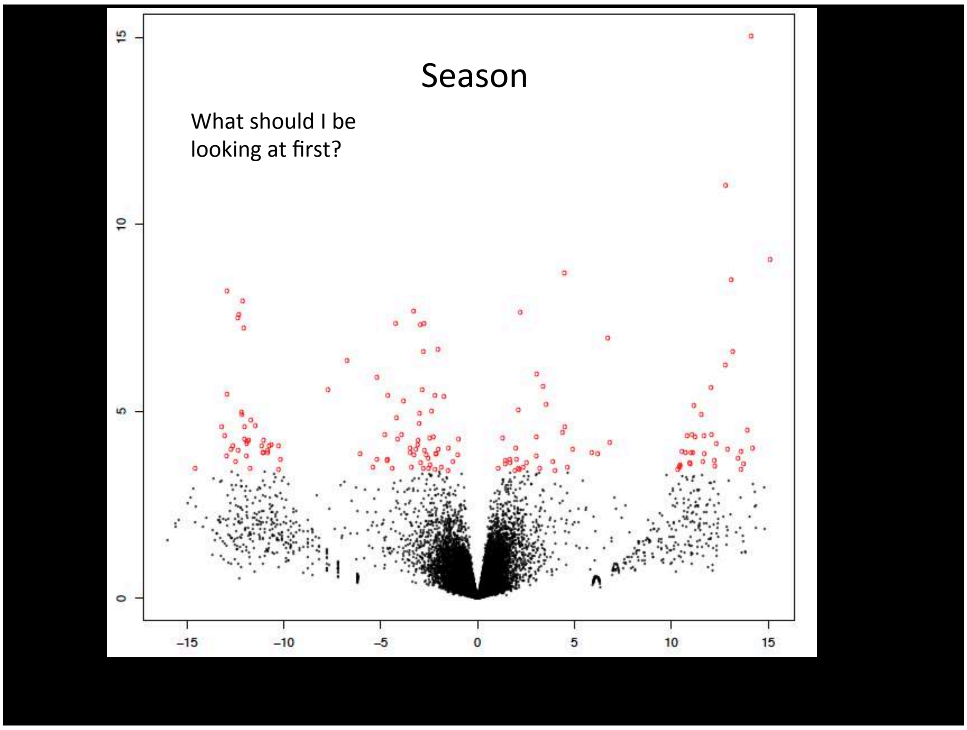
14 samples: one from each family, thorax and abdomen

69,075 contigs

edgeR

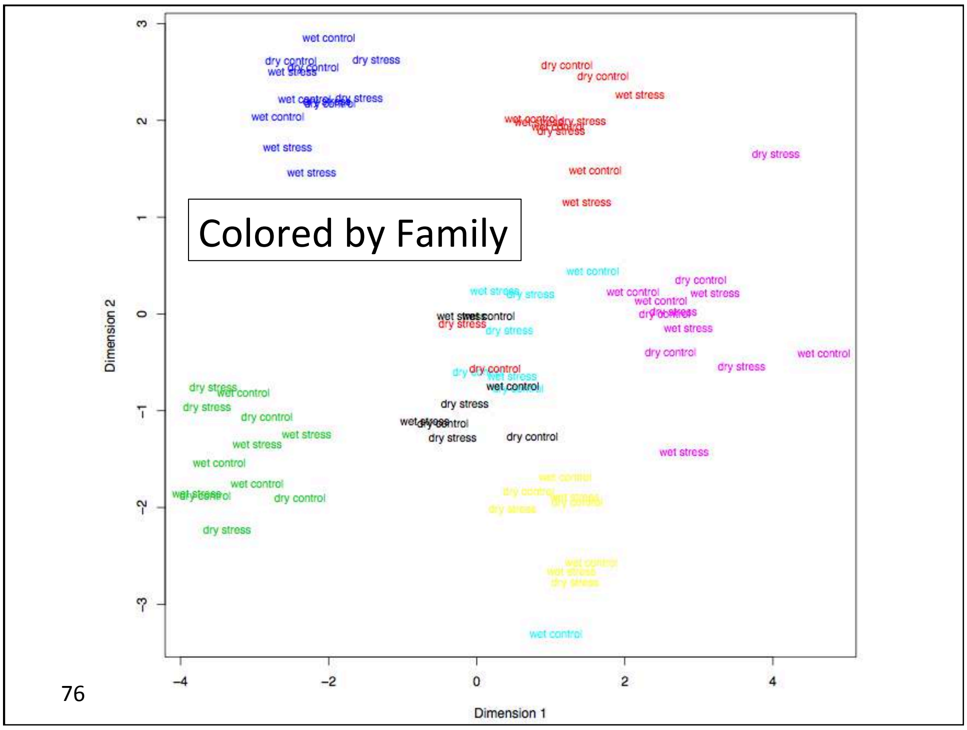
Bioconductor

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



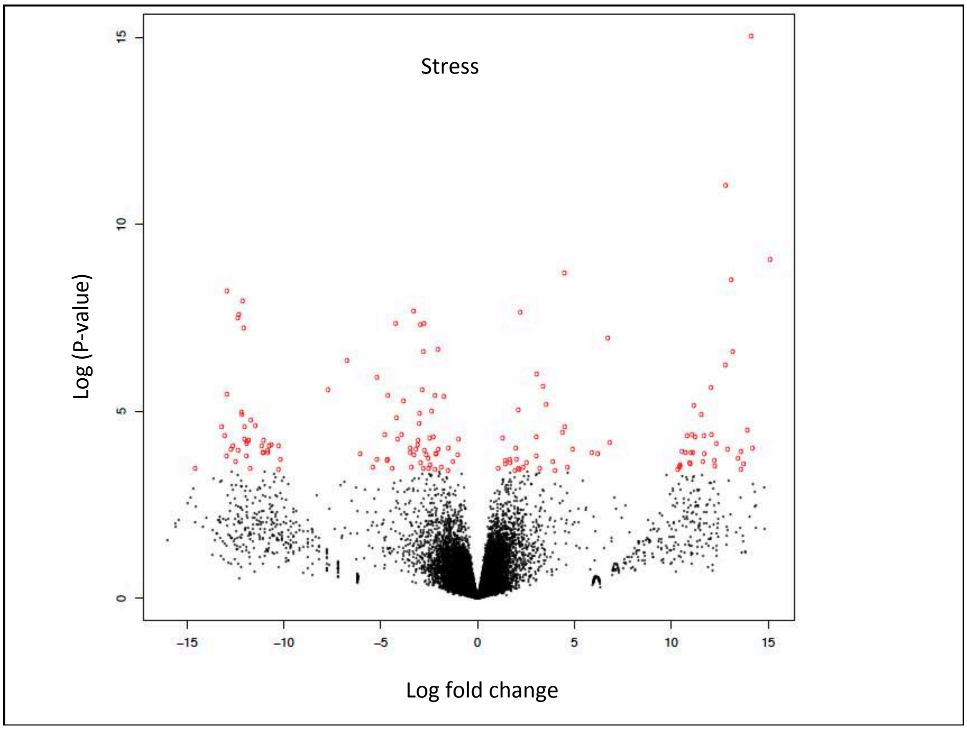
Season

What should I be looking at first?



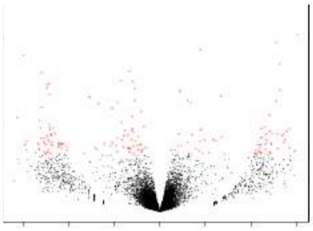
Colored by Family

76

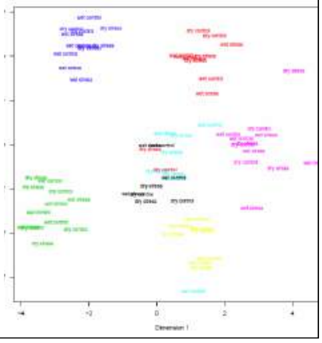

Stress
Log (P-value)
Log fold change
15
10
5
0
−15
−10
−5
0
5
10
15

Log (P-value)
Log fold change

Effect of filtering the mapping to Trinity contigs



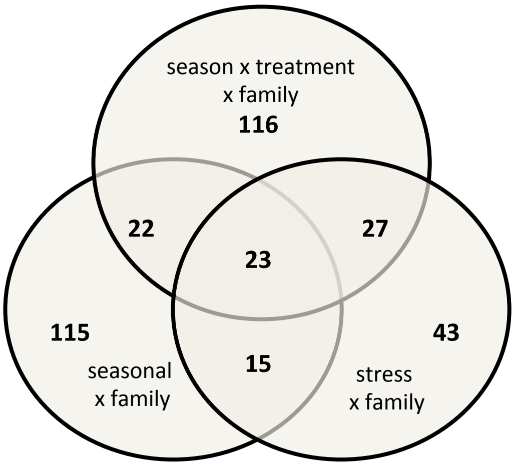
71 zero-read samples allowed



GLM results

- Plastic responses:
  - Effects without any interaction with Family

- Genetic response:
  - ○ Effects that have an interaction with family
  - ○ Potential targets of natural selection



season x treatment x family **116**

22

27

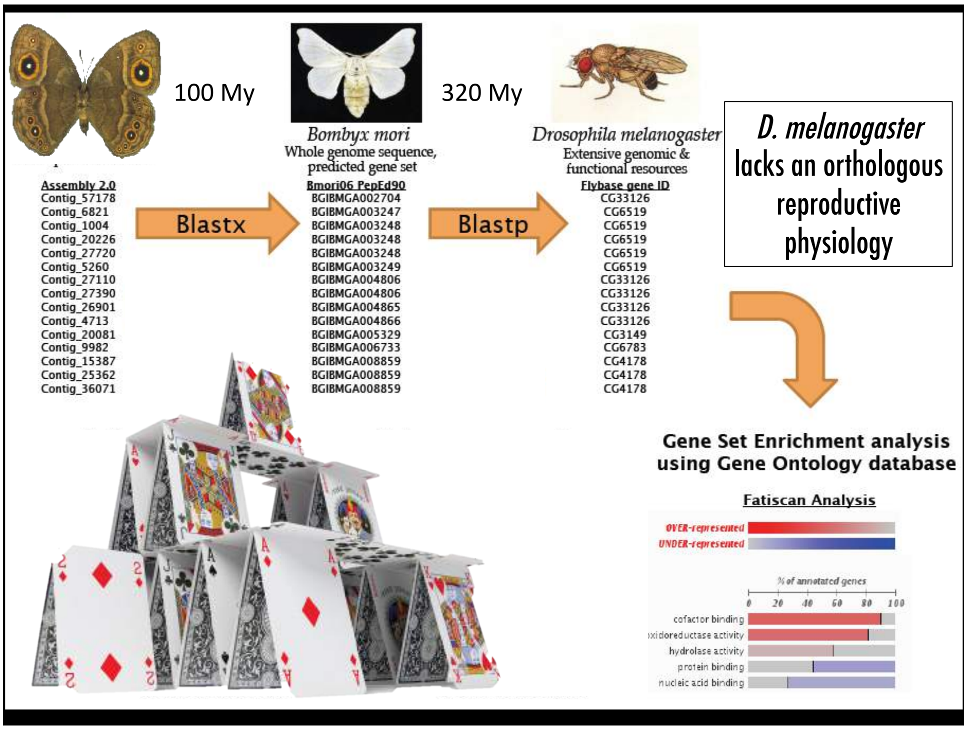
23

115 seasonal x family

15

43 stress x family

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



Bombyx mori
Whole genome sequence, predicted gene set

Drosophila melanogaster
Extensive genomic & functional resources

Assembly 2.0	Bmori06 PepEd90	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

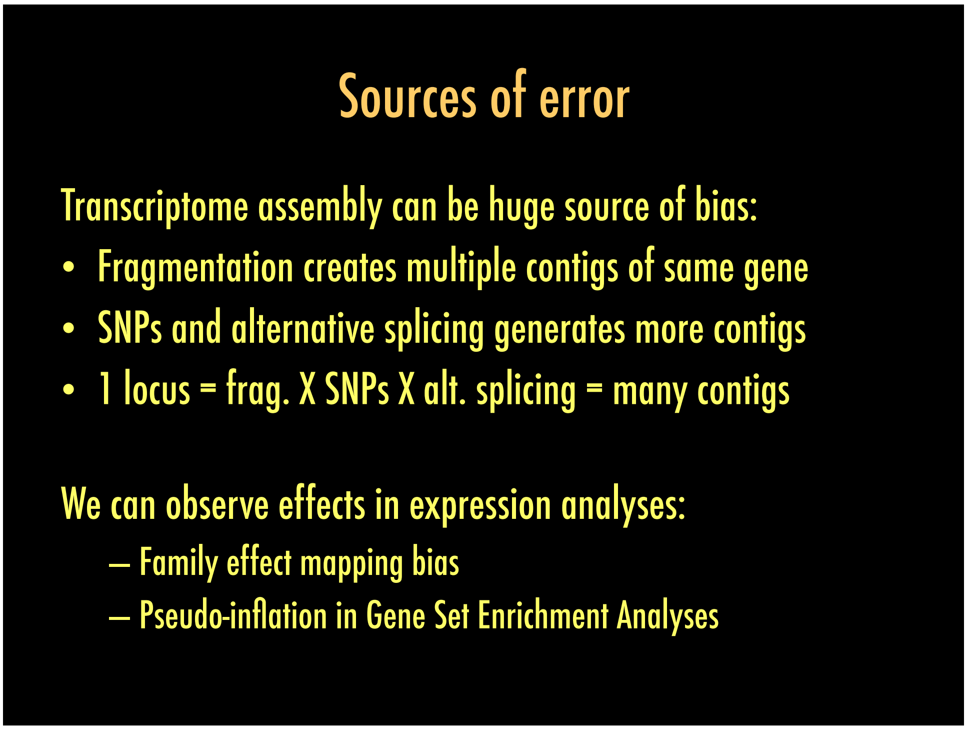
Fatiscan Analysis

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 I	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



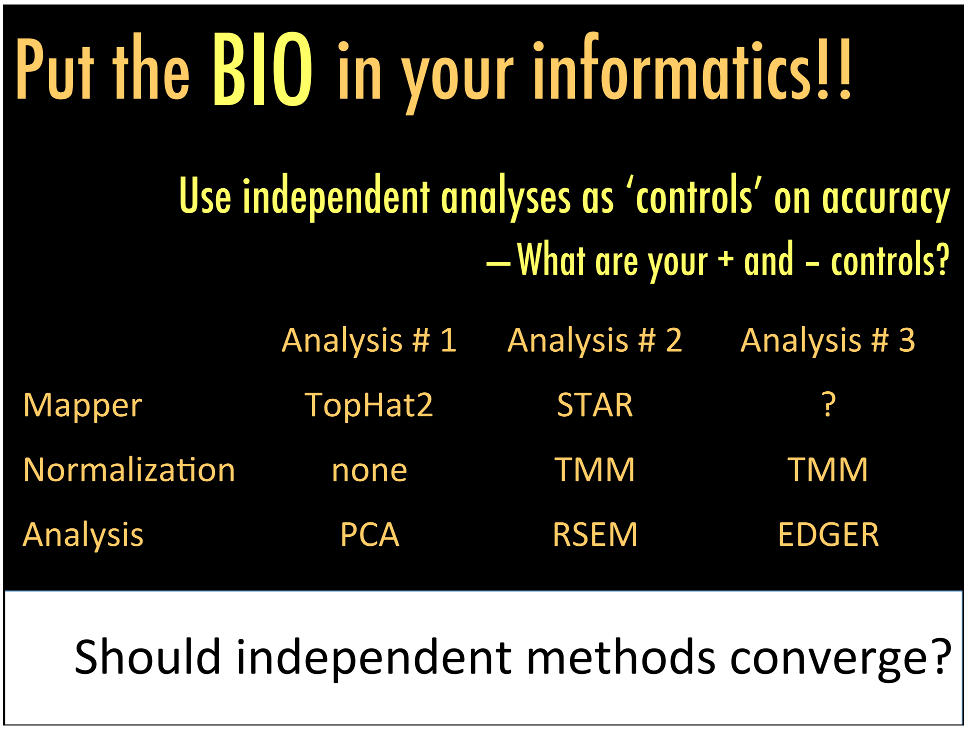
Sources of error

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



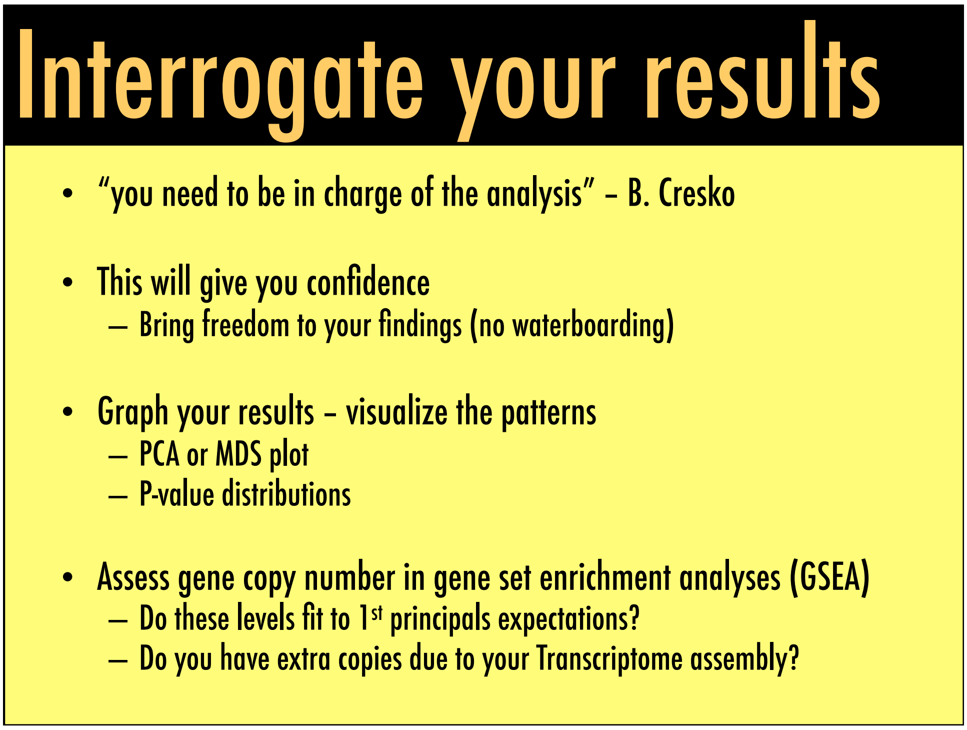
Put the BIO in your informatics!!

Use independent analyses as 'controls' on accuracy

– What are your + and - controls?

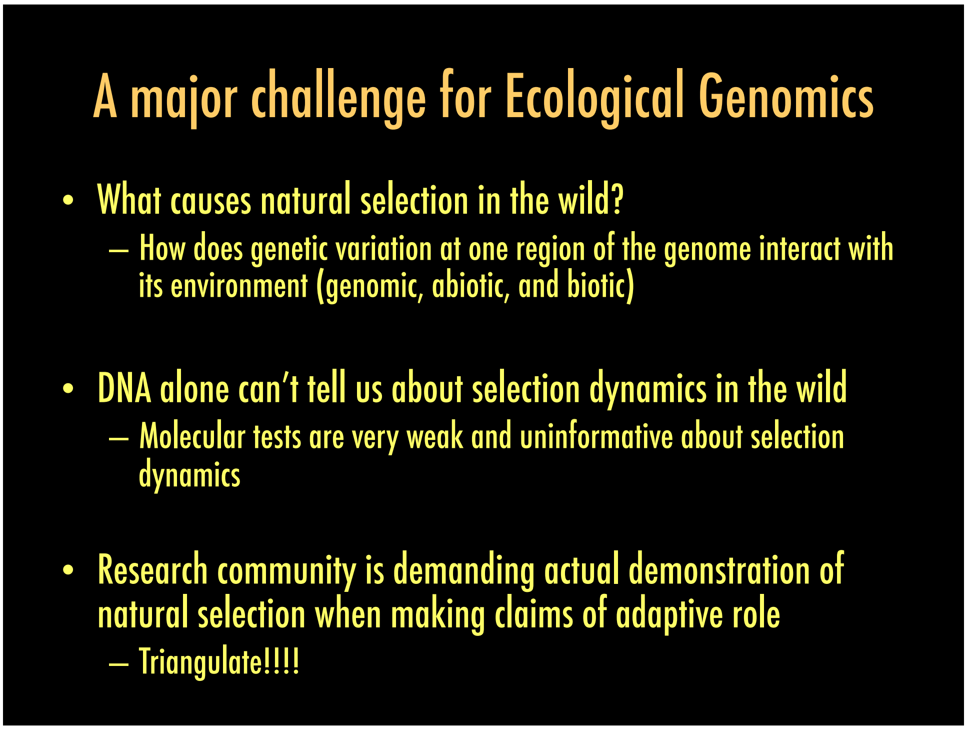
	Analysis # 1	Analysis # 2	Analysis # 3
Mapper	TopHat2	STAR	?
Normalization	none	TMM	TMM
Analysis	PCA	RSEM	EDGER

Should independent methods converge?



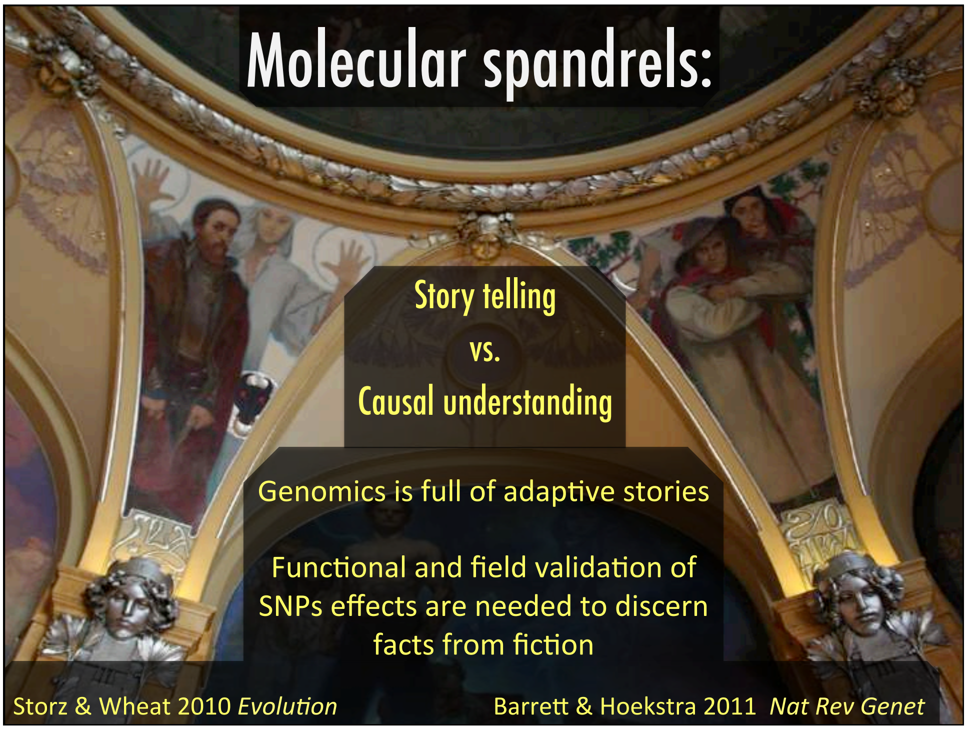
Interrogate your results

- "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*

Team Alba

Constantí Stefanescu

Alyssa Woronik

Mike Perry

Kalle Tunström

Maria Celorio

John Hallmén

Philipp Lehmann

Erik Philip-Sörensens stiftelse
FÖR FRAMJANDET AV GENETISK OCH HUMANISTISK VETENSKAPLIG FORSKNING

Stockholm University

Knut och Alice Wallenbergs Stiftelse

Thanks!

Stockholm University

Academy of Finland

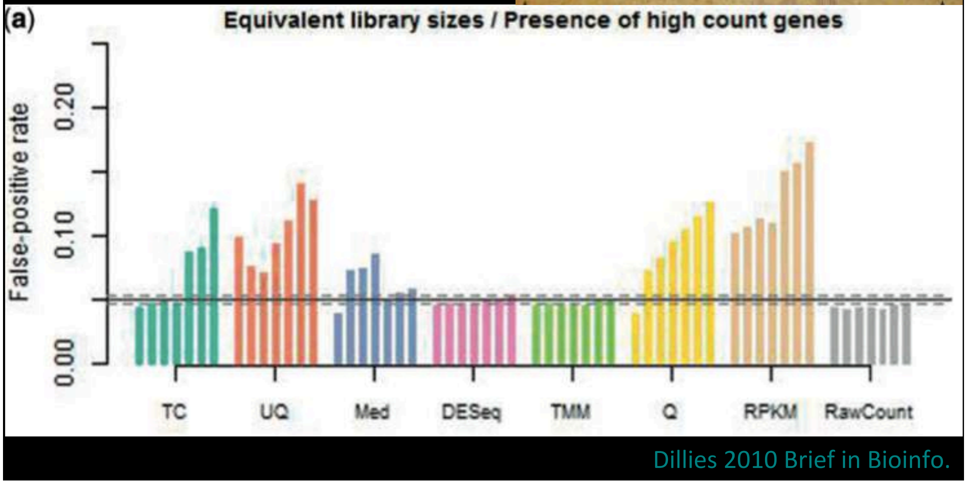
Vetenskapsrådet

Knut och Alice Wallenbergs Stiftelse

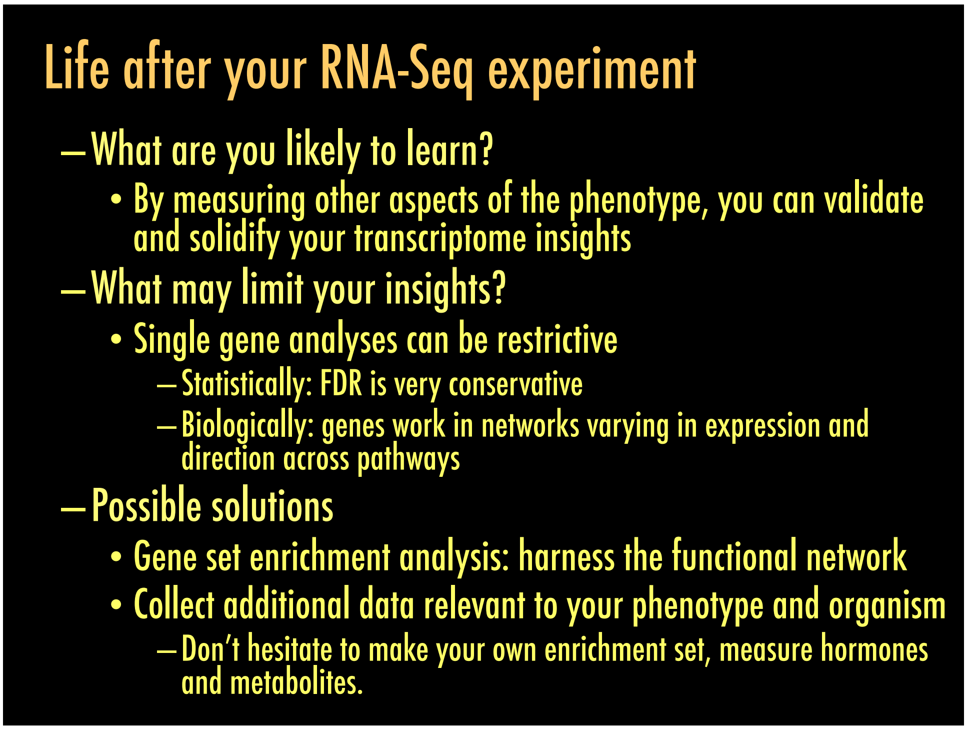
Common mistakes

- Blindly trusting bioinformaticians: look at your data!!!
- Mapping reads to a very divergent genome
  - Only most conserved genes map: bias due to divergence and mapping threholds
- Not accurately assessing a TA
  - Your template determines quality of results
- Not enough reads, replication, or statistical power
  - Large amounts of data to not change fundamental statistics (never pool unless necessary)
- Not assessing likely biases in analyses
  - Try different mapping thresholds & analysis methods to assess convergence of biological signal
  - Assess alternative splicing and duplication potential in findings
- Data size and computational power are demanding
  - Download data and work with it before your real data comes.

Normalization matters

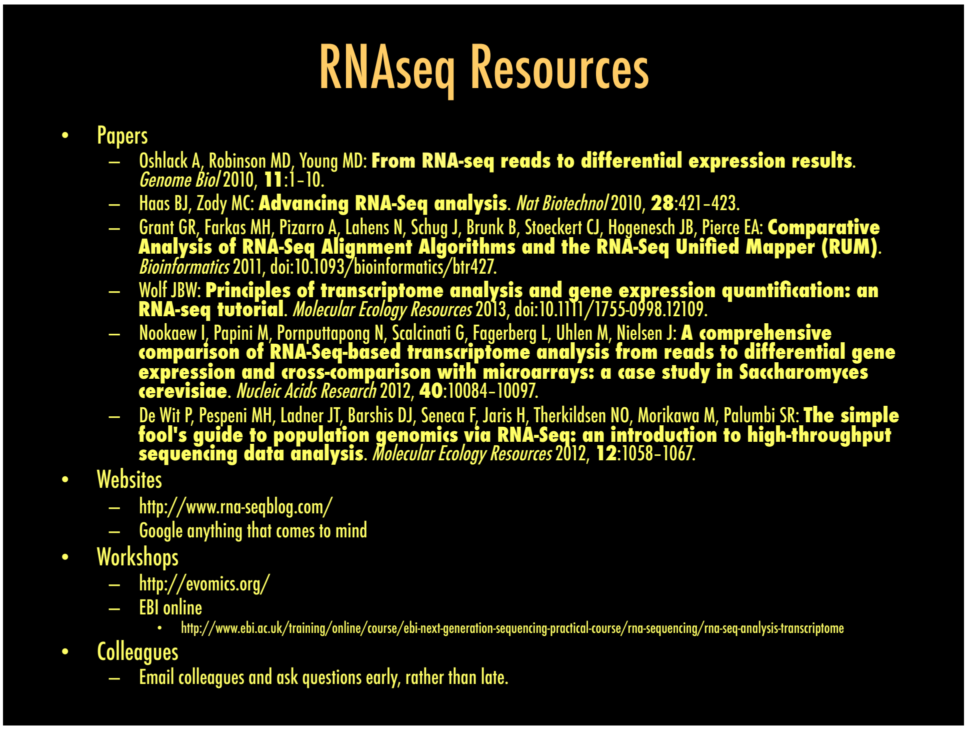


Dillies 2010 Brief in Bioinfo.



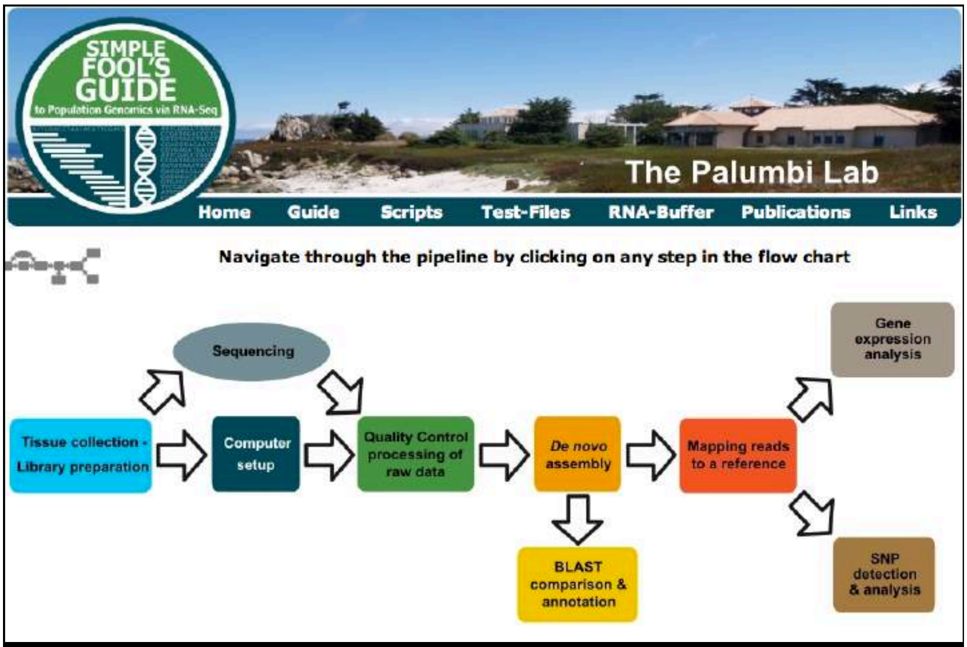
Life after your RNA-Seq experiment

- What are you likely to learn?
  - By measuring other aspects of the phenotype, you can validate and solidify your 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
  - Collect additional data relevant to your phenotype and organism
    - Don't hesitate to make your own enrichment set, measure hormones and metabolites.



RNAseq Resources

- Papers
  - Oshlack A, Robinson MD, Young MD: **From RNA-seq reads to differential expression results**. *Genome Biol* 2010, **11**:1–10.
  - Haas BJ, Zody MC: **Advancing RNA-Seq analysis**. *Nat Biotechnol* 2010, **28**:421–423.
  - Grant GR, Farkas MH, Pizarro A, Lahens N, Schug J, Brunk B, Stoeckert CJ, Hogenesch JB, Pierce EA: **Comparative Analysis of RNA-Seq Alignment Algorithms and the RNA-Seq Unified Mapper (RUM)**. *Bioinformatics* 2011, doi:10.1093/bioinformatics/btr427.
  - Wolf JBW: **Principles of transcriptome analysis and gene expression quantification: an RNA-seq tutorial**. *Molecular Ecology Resources* 2013, doi:10.1111/1755-0998.12109.
  - Nookaew I, Papini M, Pornputtapong N, Scalcinati G, Fagerberg L, Uhlen M, Nielsen J: **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.
  - De Wit P, Pespeni MH, Ladner JT, Barshis DJ, Seneca F, Jaris H, Therkildsen NO, Morikawa M, Palumbi SR: **The simple fool's guide to population genomics via RNA-Seq: an introduction to high-throughput sequencing data analysis**. *Molecular Ecology Resources* 2012, **12**:1058–1067.
- Websites
  - http://www.rna-seqblog.com/
  - Google anything that comes to mind
- Workshops
  - http://evomics.org/
  - EBI online
    - http://www.ebi.ac.uk/training/online/course/ebi-next-generation-sequencing-practical-course/rna-sequencing/rna-seq-analysis-transcriptome
- Colleagues
  - Email colleagues and ask questions early, rather than late.

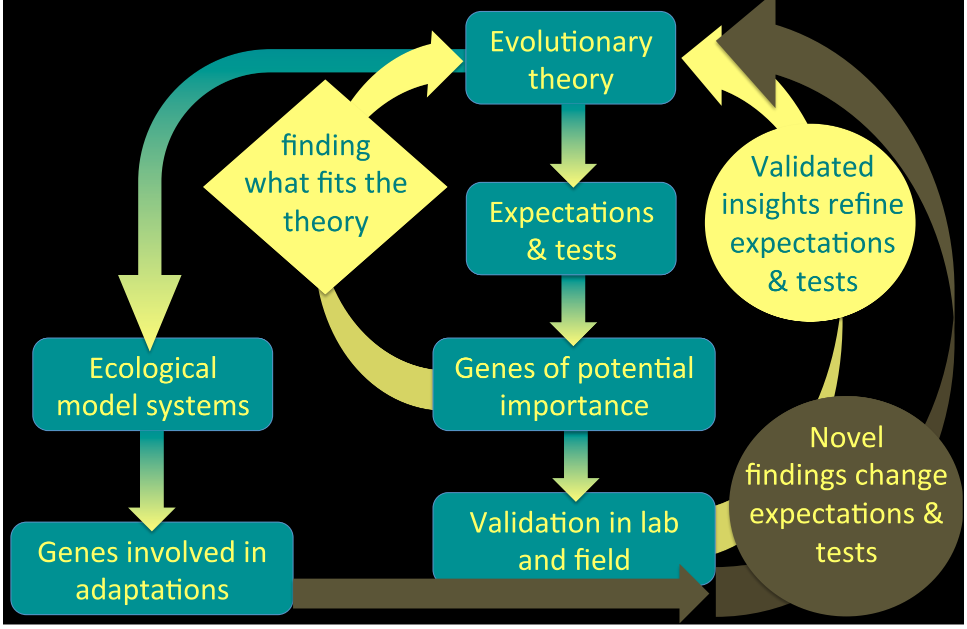


Navigate through the pipeline by clicking on any step in the flow chart

Sequencing
Tissue collection - Library preparation
Computer setup
Quality Control processing of raw data
De novo assembly
Mapping reads to a reference
Gene expression analysis
SNP detection & analysis
BLAST comparison & annotation

http://sfg.stanford.edu/guide.html

Validating candiate genes moves us forward:



Evolutionary theory

finding what fits the theory

Expectations & tests

Validated insights refine expectations & tests

Ecological model systems

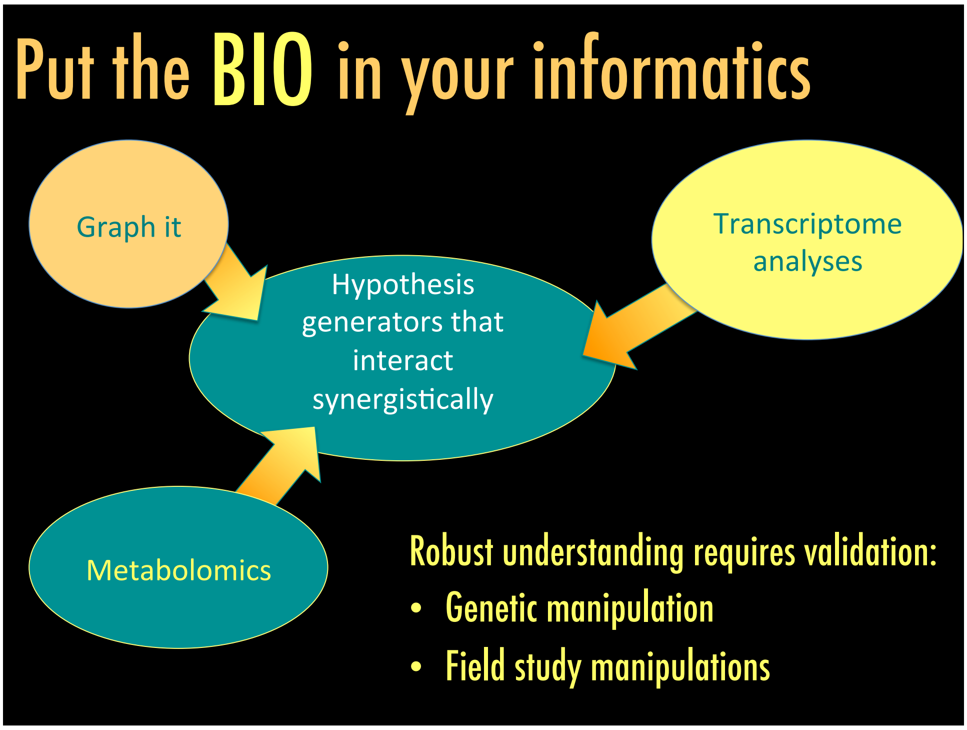
Genes of potential importance

Novel findings change expectations & tests

Genes involved in adaptations

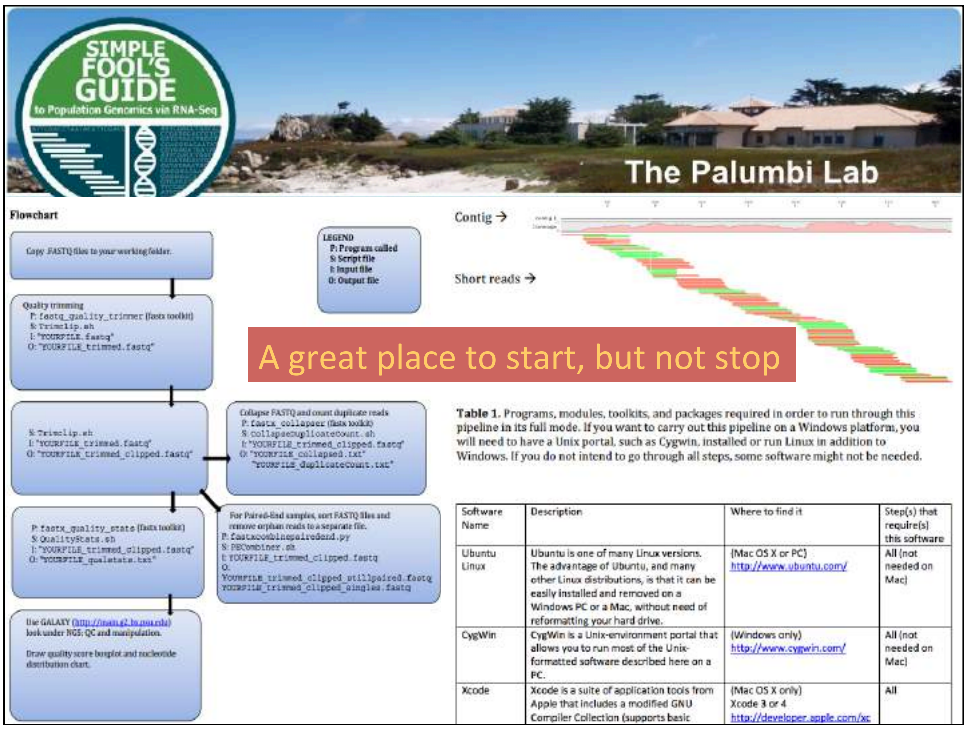
Validation in lab and field

Put the BIO in your informatics



Robust understanding requires validation:

- Genetic manipulation
- Field study manipulations



A great place to start, but not stop

Flowchart

Copy .FASTQ files to your working folder.

Quality trimming
P: fastq_quality_trimmer (fastx toolkit)
S: Trimclip.sh
I: "YOURFILE.fastq"
O: "YOURFILE_trimmed.fastq"

S: Trimclip.sh
I: "YOURFILE_trimmed.fastq"
O: "YOURFILE_trimmed_clipped.fastq"

Collapse FASTQ and count duplicate reads
P: fastx_collapser (fastx toolkit)
S: collapseDuplicateCount.sh
I: "YOURFILE_trimmed_clipped.fastq"
O: "YOURFILE_collapsed.txt"
"YOURFILE_duplicateCount.txt"

P: fastx_quality_stats (fastx toolkit)
S: QualityStats.sh
I: "YOURFILE_trimmed_clipped.fastq"
O: "YOURFILE_qualstats.txt"

For Paired-End samples, sort FASTQ files and remove orphan reads to a separate file.
P: fastacombinepairedend.py
S: PECombiner.sh
I: YOURFILE_trimmed_clipped.fastq
O:
YOURFILE_trimmed_clipped_stillpaired.fastq
YOURFILE_trimmed_clipped_singles.fastq

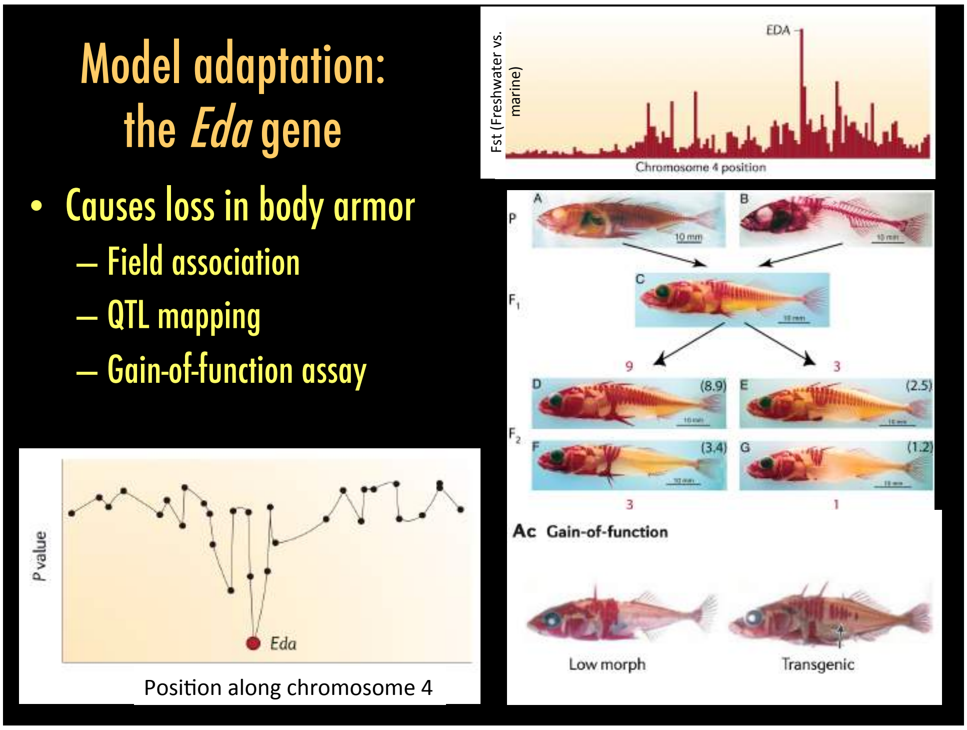
Use GALAXY (http://main.g2.bx.psu.edu) look under NGS: QC and manipulation.

Draw quality score boxplot and nucleotide distribution chart.

LEGEND
P: Program called
S: Script file
I: Input file
O: Output file

Table 1. Programs, modules, toolkits, and packages required in order to run through this pipeline in its full mode. If you want to carry out this pipeline on a Windows platform, you will need to have a Unix portal, such as Cygwin, installed or run Linux in addition to Windows. If you do not intend to go through all steps, some software might not be needed.

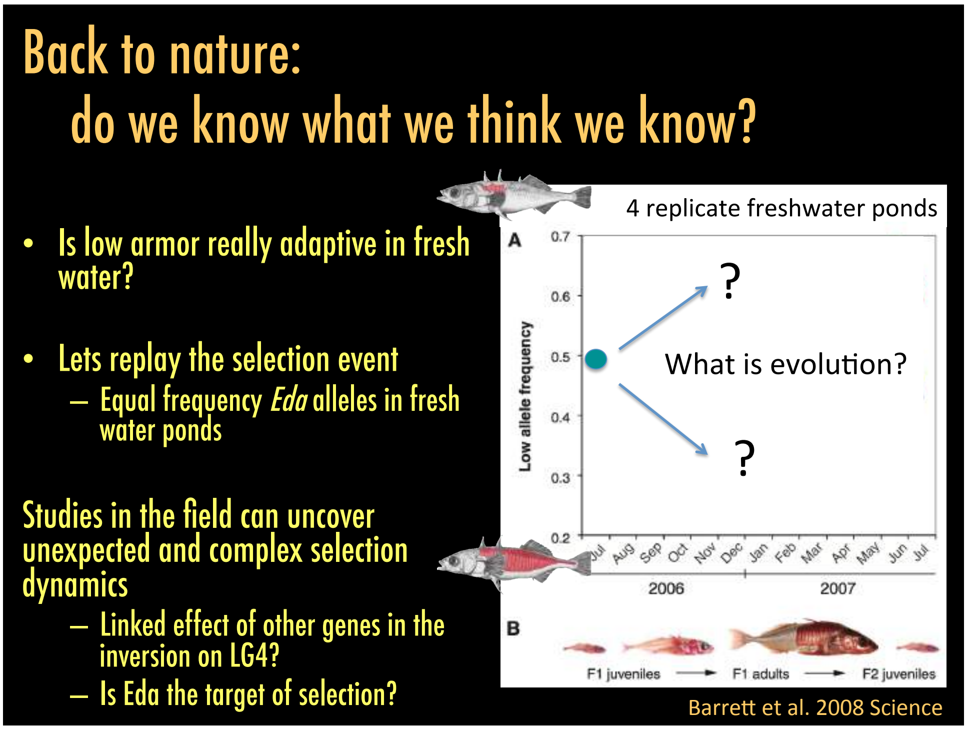
Software Name	Description	Where to find it	Step(s) that require(s) this software
Ubuntu Linux	Ubuntu is one of many Linux versions. The advantage of Ubuntu, and many other Linux distributions, is that it can be easily installed and removed on a Windows PC or a Mac, without need of reformatting your hard drive.	(Mac OS X or PC) http://www.ubuntu.com/	All (not needed on Mac)
CygWin	CygWin is a Unix-environment portal that allows you to run most of the Unix-formatted software described here on a PC.	(Windows only) http://www.cygwin.com/	All (not needed on Mac)
Xcode	Xcode is a suite of application tools from Apple that includes a modified GNU Compiler Collection (supports basic	(Mac OS X only) Xcode 3 or 4 http://developer.apple.com/xc	All



Model adaptation: the *Eda* gene

- Causes loss in body armor
  - Field association
  - QTL mapping
  - Gain-of-function assay

Fst (Freshwater vs. marine)
EDA
Chromosome 4 position
P value
Eda
Position along chromosome 4
P
A
10 mm
B
10 mm
F1
C
10 mm
9
3
D
(8.9)
E
(2.5)
F2
F
(3.4)
G
(1.2)
3
1
Ac Gain-of-function
Low morph
Transgenic



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

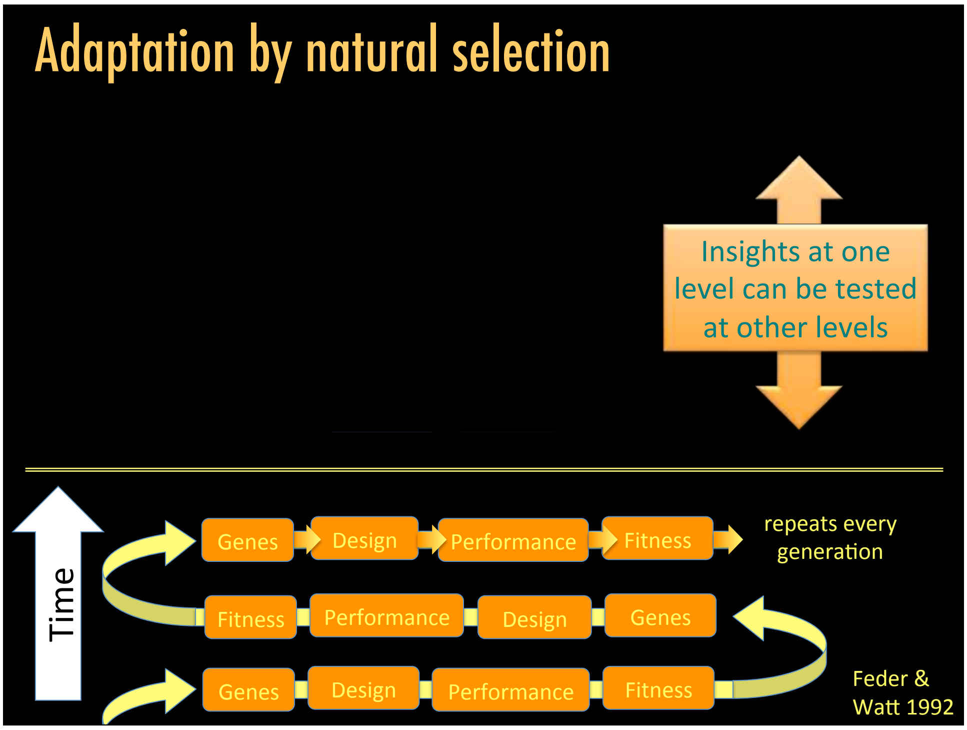
Studies in the field can uncover unexpected and complex selection dynamics

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

4 replicate freshwater ponds

What is evolution?

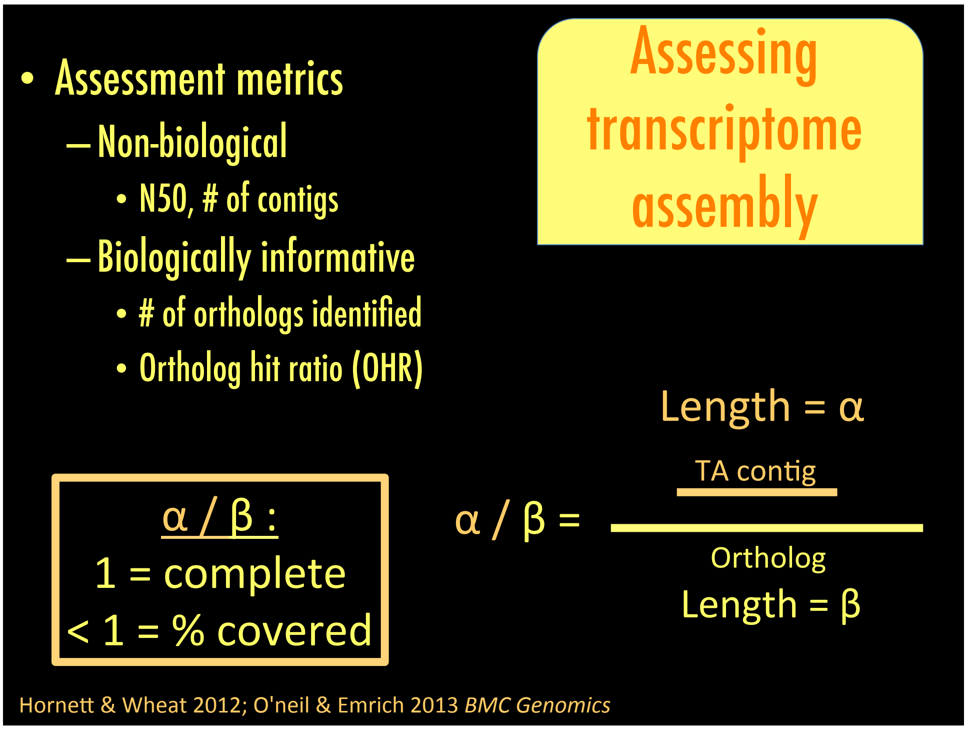
Barrett et al. 2008 Science



Adaptation by natural selection

Insights at one level can be tested at other levels

Feder & Watt 1992



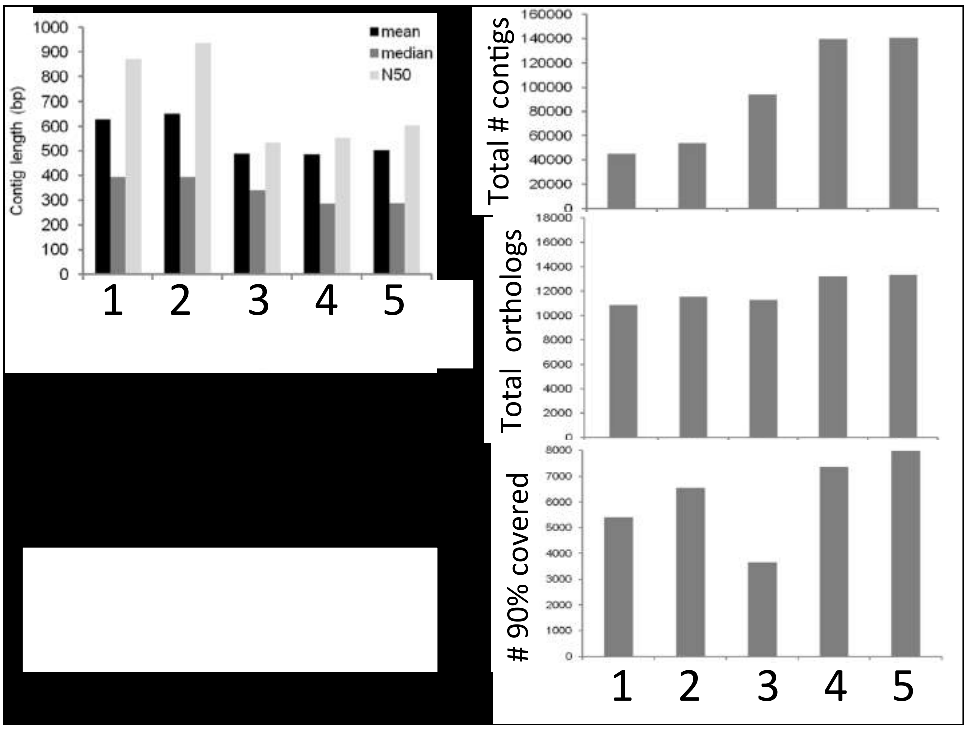
Assessing transcriptome assembly

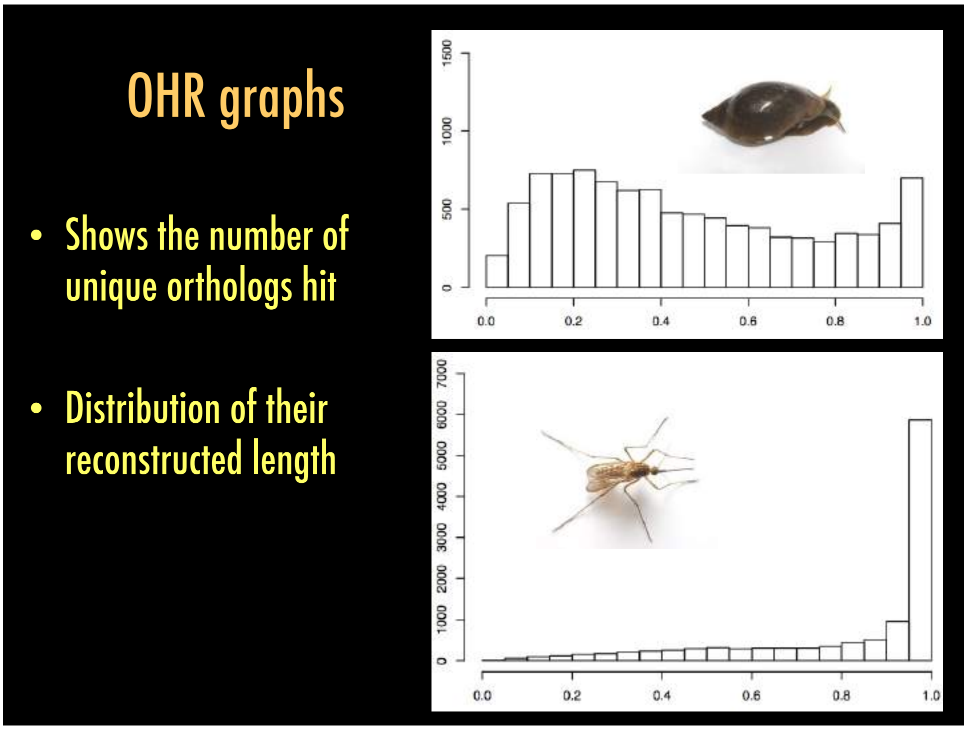
- Assessment metrics
  - Non-biological
    - N50, # of contigs
  - Biologically informative
    - # of orthologs identified
    - Ortholog hit ratio (OHR)

<u>α / β :</u>
1 = complete
< 1 = % covered

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

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



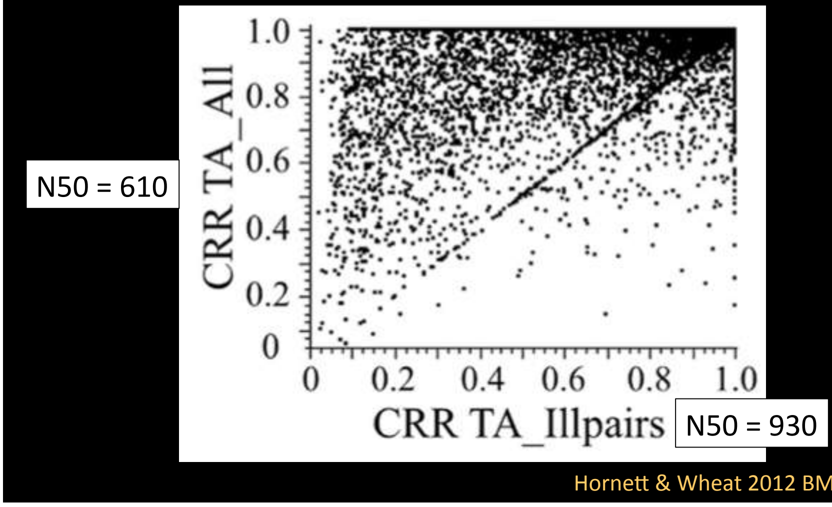


OHR graphs

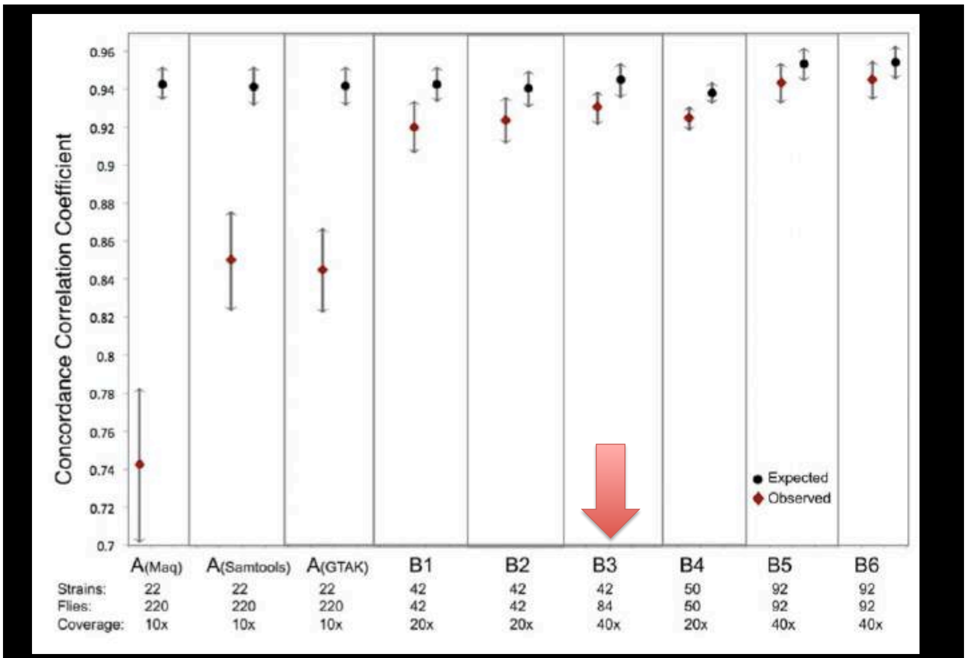
- Shows the number of unique orthologs hit
- Distribution of their reconstructed length

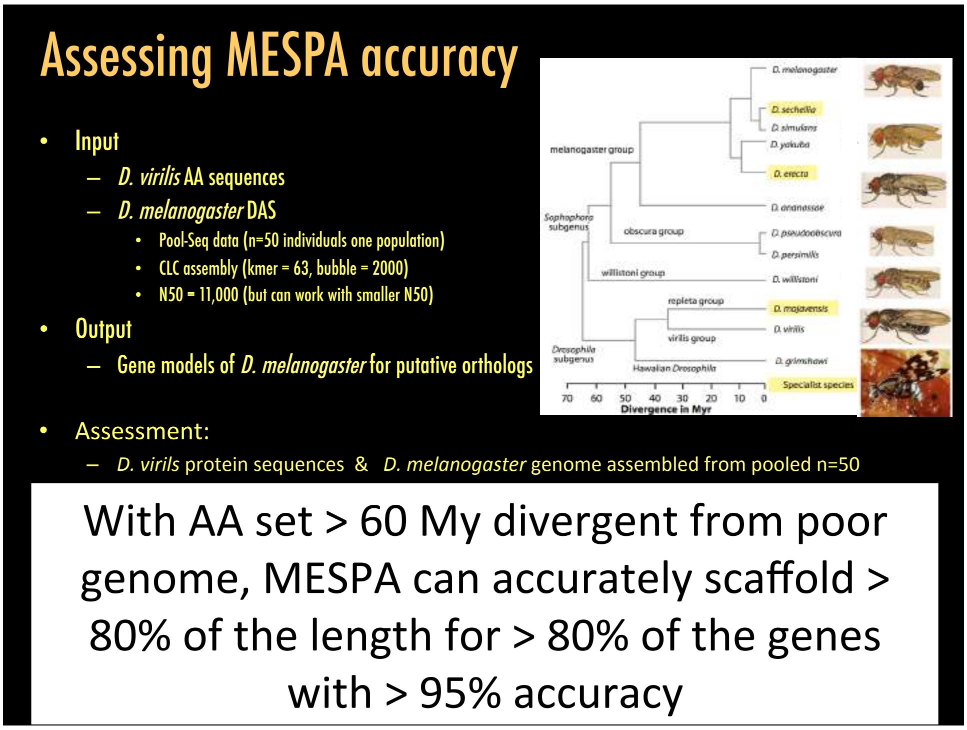
Comparative OHR

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



Hornett & Wheat 2012 BMC Genomics



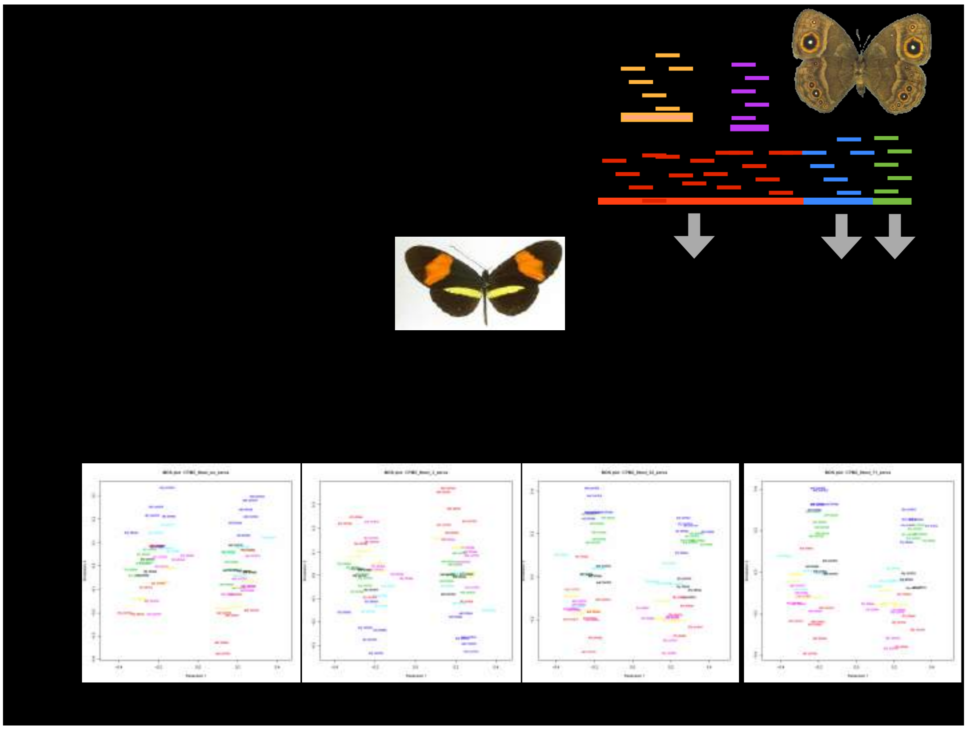


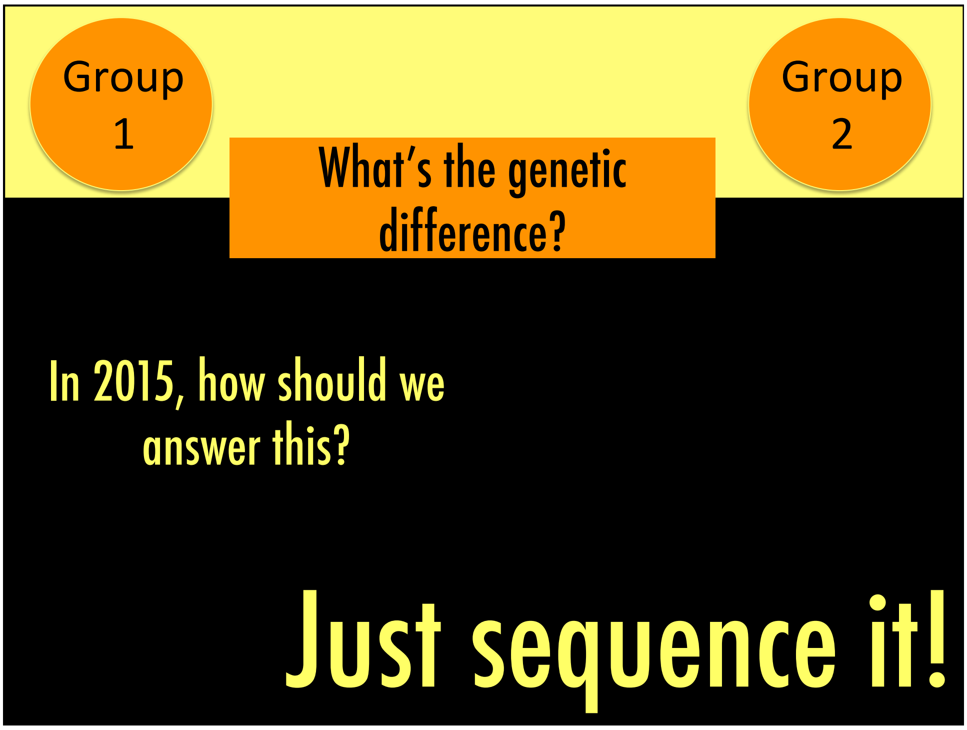
Assessing MESPA accuracy

- Input
  - *D. virilis* AA sequences
  - *D. melanogaster* DAS
    - Pool-Seq data (n=50 individuals one population)
    - CLC assembly (kmer = 63, bubble = 2000)
    - N50 = 11,000 (but can work with smaller N50)
- Output
  - Gene models of *D. melanogaster* for putative orthologs
- Assessment:
  - *D. virils* protein sequences & *D. melanogaster* genome assembled from pooled n=50

D. melanogaster
D. sechellia
D. simulans
D. yakuba
D. erecta
D. ananassae
D. pseudoobscura
D. persimilis
D. willistoni
D. mojavensis
D. virilis
D. grimshawi
Specialist species
melanogaster group
Sophophora subgenus
obscura group
willistoni group
repleta group
virilis group
Drosophila subgenus
Hawaiian Drosophila
70 60 50 40 30 20 10 0
Divergence in Myr

With AA set > 60 My divergent from poor genome, MESPA can accurately scaffold > 80% of the length for > 80% of the genes with > 95% accuracy





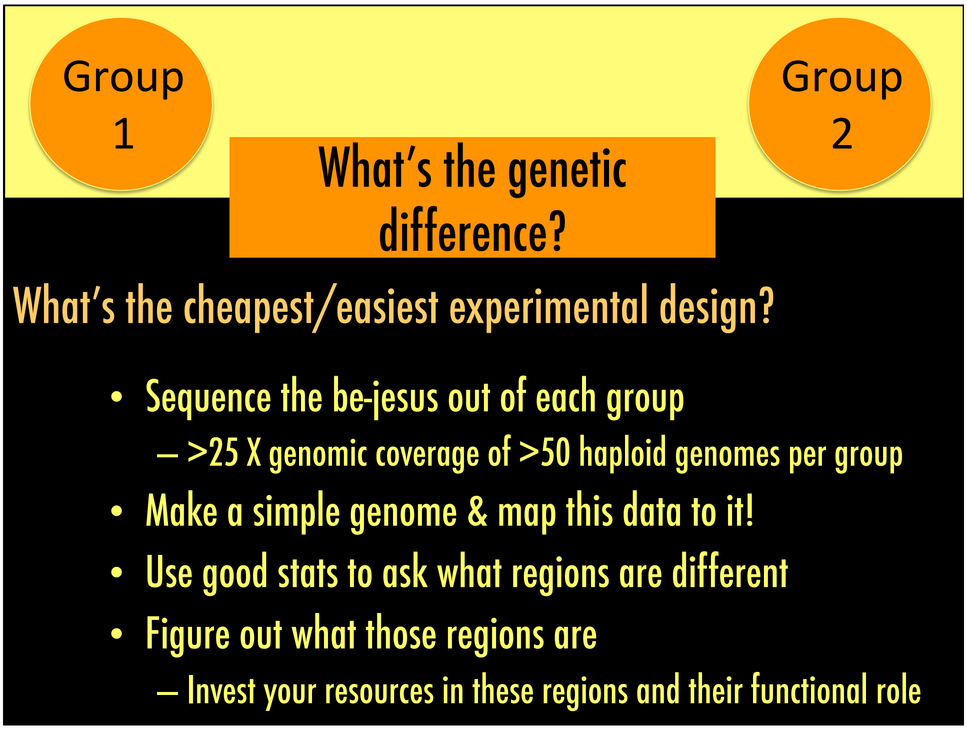
Group 1

Group 2

What's the genetic difference?

In 2015, how should we answer this?

Just sequence it!



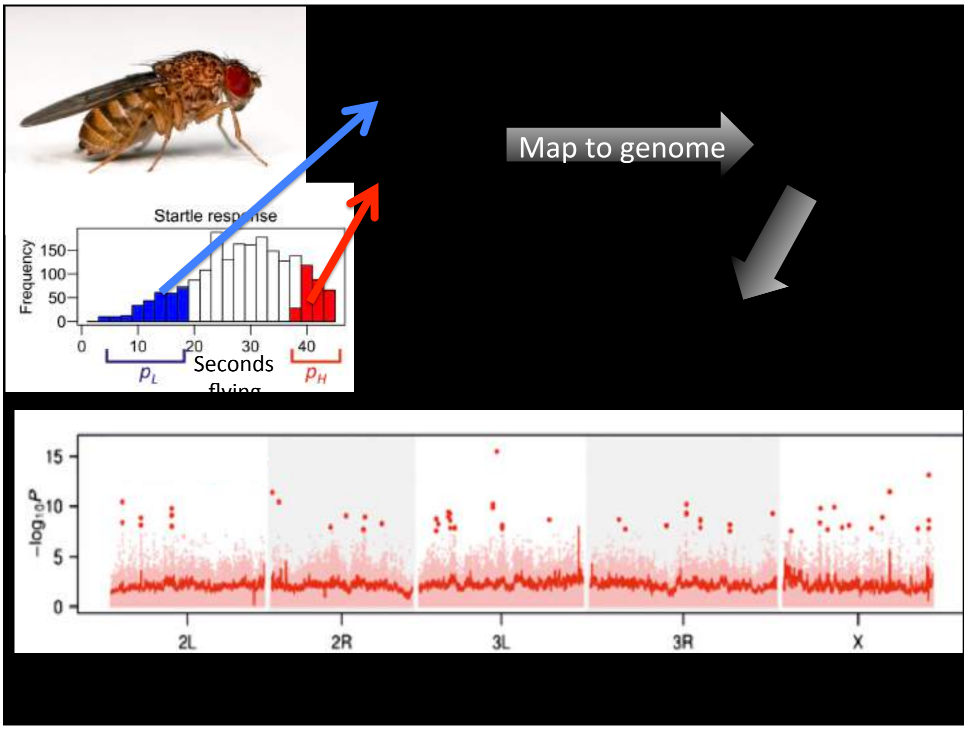
Group 1

Group 2

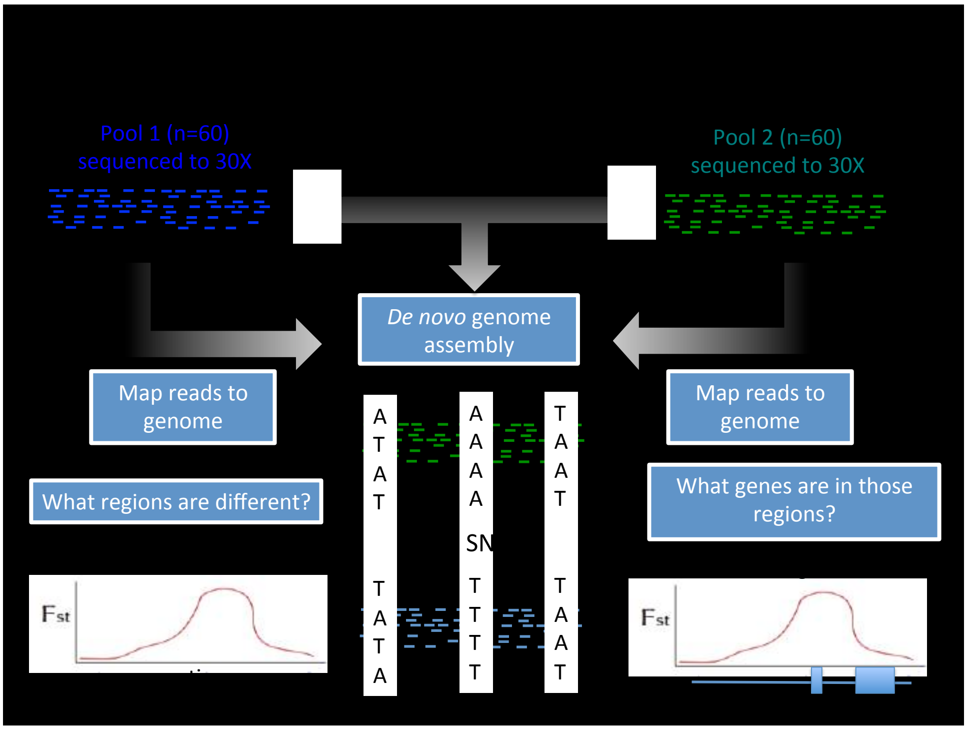
What's the genetic difference?

What's the cheapest/easiest experimental design?

- Sequence the be-jesus out of each group
  - >25 X genomic coverage of >50 haploid genomes per group
- Make a simple genome & map this data to it!
- Use good stats to ask what regions are different
- Figure out what those regions are
  - Invest your resources in these regions and their functional role



Map to genome



Pool 1 (n=60) sequenced to 30X

Pool 2 (n=60) sequenced to 30X

De novo genome assembly

Map reads to genome

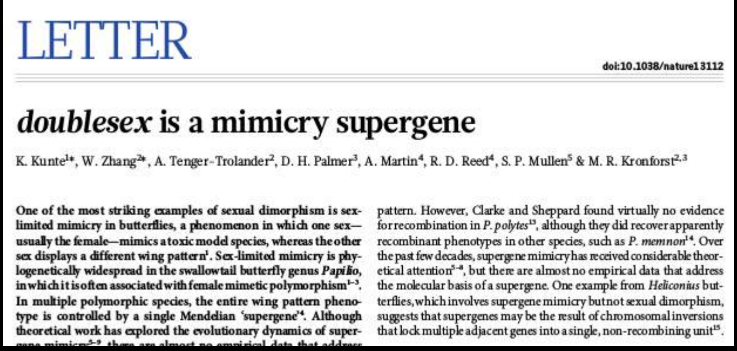
Map reads to genome

What regions are different?

What genes are in those regions?

Can this really work?

Case study # 1



LETTER

doi:10.1038/nature13112

doublesex is a mimicry supergene

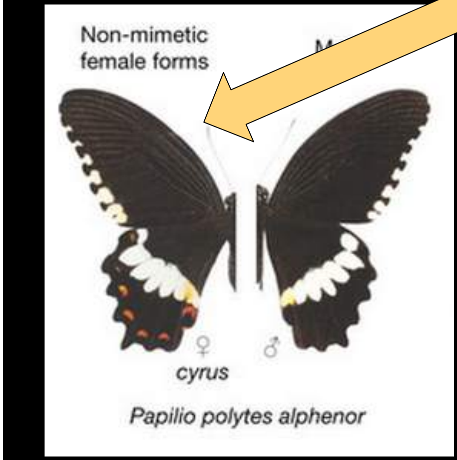
K. Kunte[1*], W. Zhang[2*], A. Tenger-Trolander[2], D. H. Palmer[3], A. Martin[4], R. D. Reed[4], S. P. Mullen[5] & M. R. Kronforst[2,3]

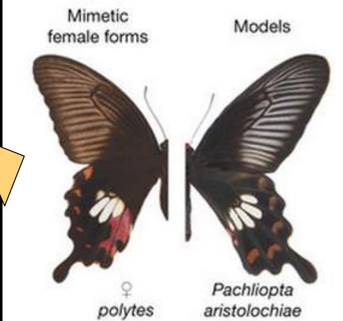
One of the most striking examples of sexual dimorphism is sex-limited mimicry in butterflies, a phenomenon in which one sex—usually the female—mimics a toxic model species, whereas the other sex displays a different wing pattern[1]. Sex-limited mimicry is phylogenetically widespread in the swallowtail butterfly genus *Papilio*, in which it is often associated with female mimetic polymorphism[1–3]. In multiple polymorphic species, the entire wing pattern phenotype is controlled by a single Mendelian 'supergene'[4]. Although theoretical work has explored the evolutionary dynamics of super-

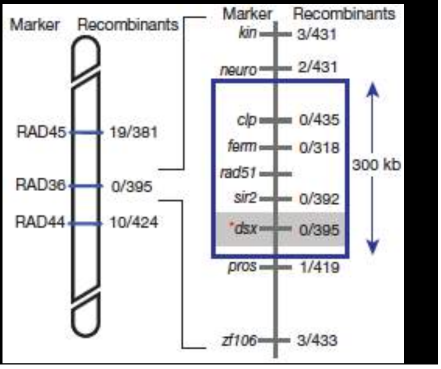
pattern. However, Clarke and Sheppard found virtually no evidence for recombination in *P. polytes*[13], although they did recover apparently recombinant phenotypes in other species, such as *P. memnon*[14]. Over the past few decades, supergene mimicry has received considerable theoretical attention[5–8], but there are almost no empirical data that address the molecular basis of a supergene. One example from *Heliconius* butterflies, which involves supergene mimicry but not sexual dimorphism, suggests that supergenes may be the result of chromosomal inversions that lock multiple adjacent genes into a single, non-recombining unit[15].

Kunte et al. 2014 Nature

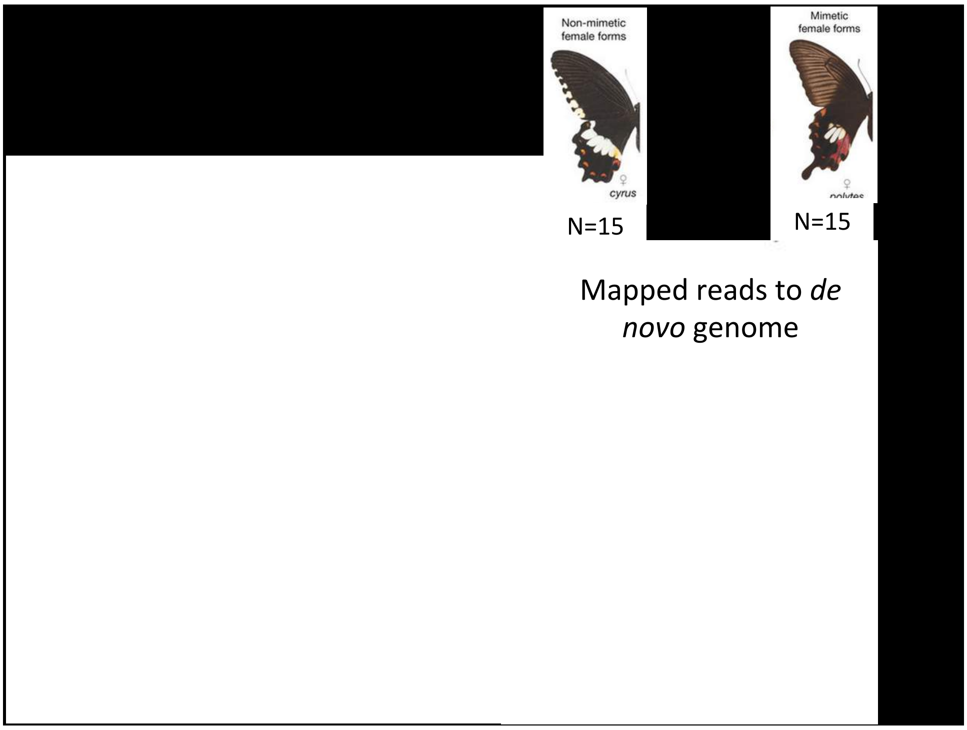
Polymorphic, sex-limited mimicry







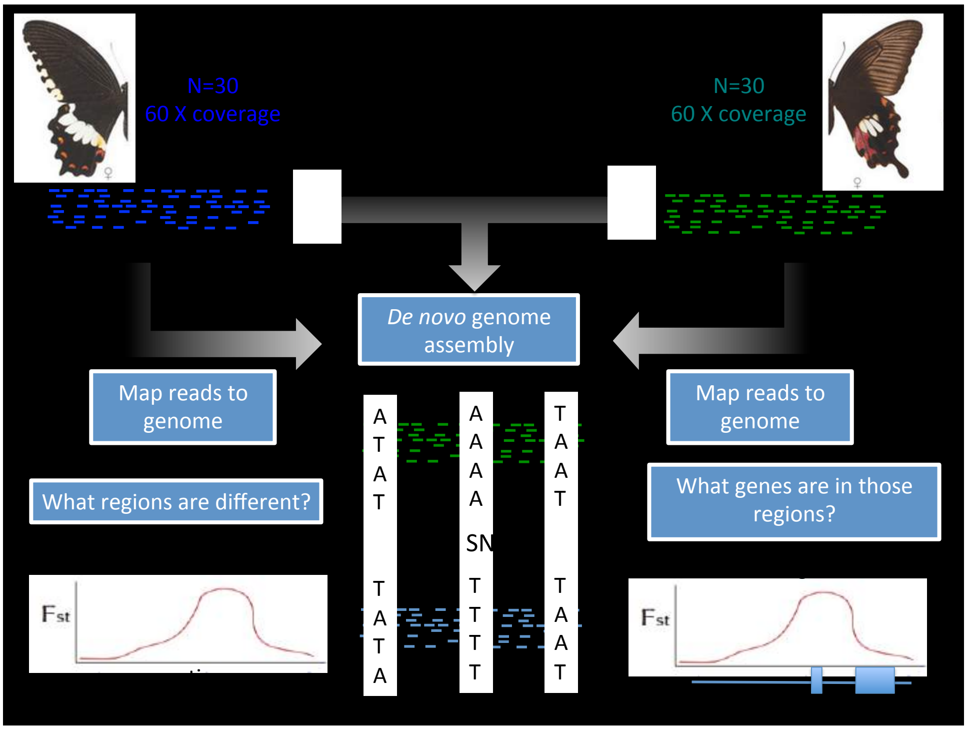
K Kunte et al. Nature **000**, 1-4 (2014) doi:10.1038/nature13112



N=15

N=15

Mapped reads to *de novo* genome



N=30
60 X coverage

N=30
60 X coverage

De novo genome assembly

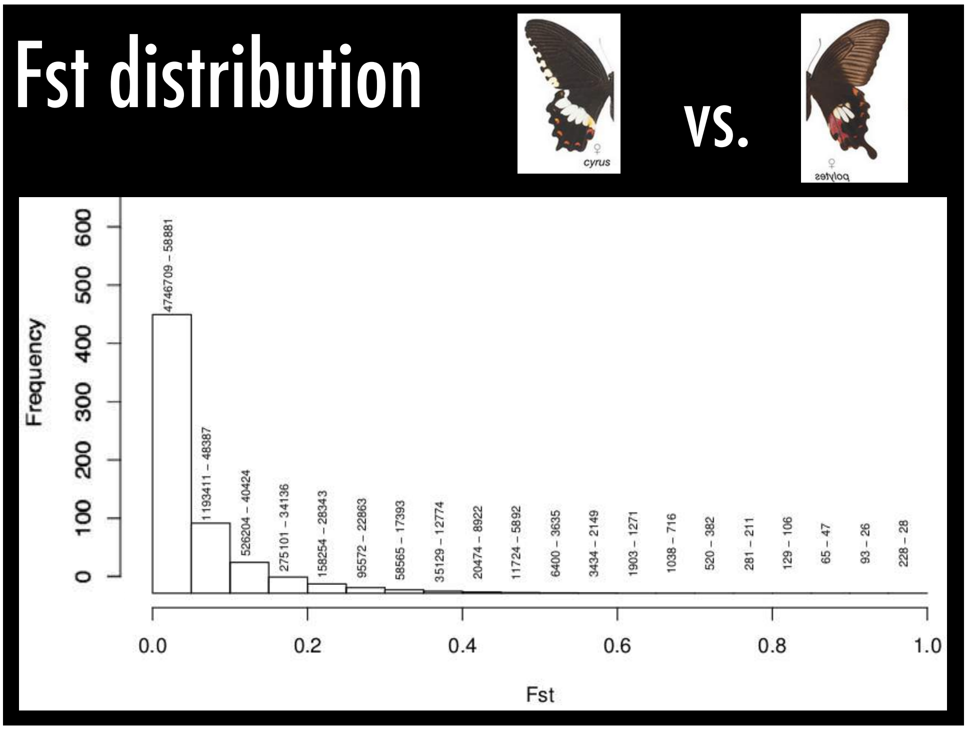
Map reads to genome

Map reads to genome

What regions are different?

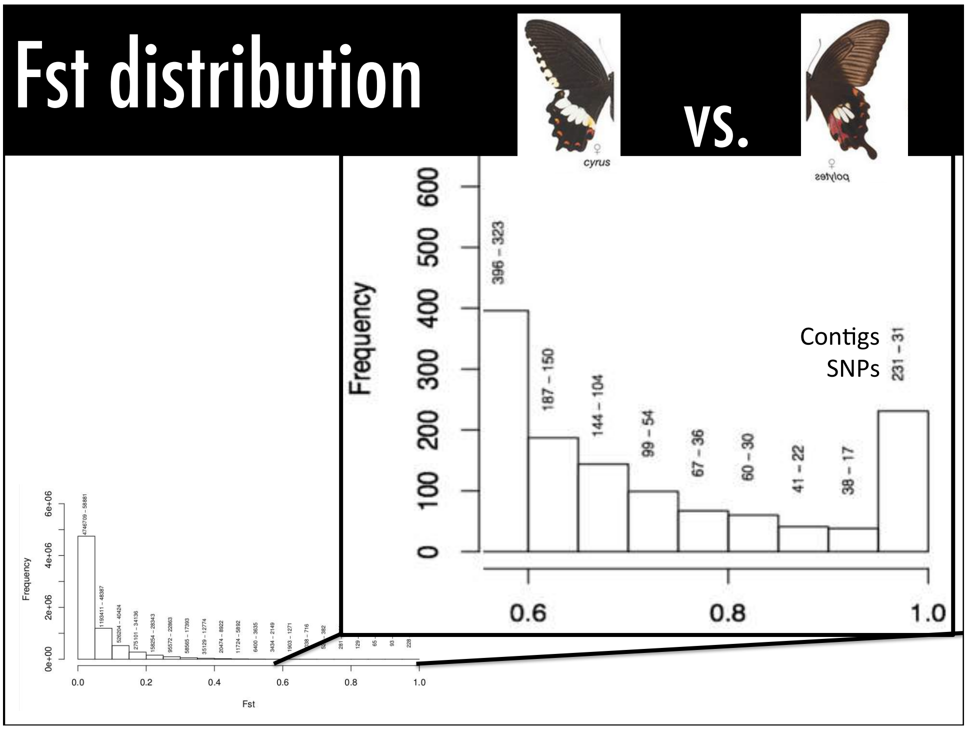
What genes are in those regions?

Can we find the same genomic regions?



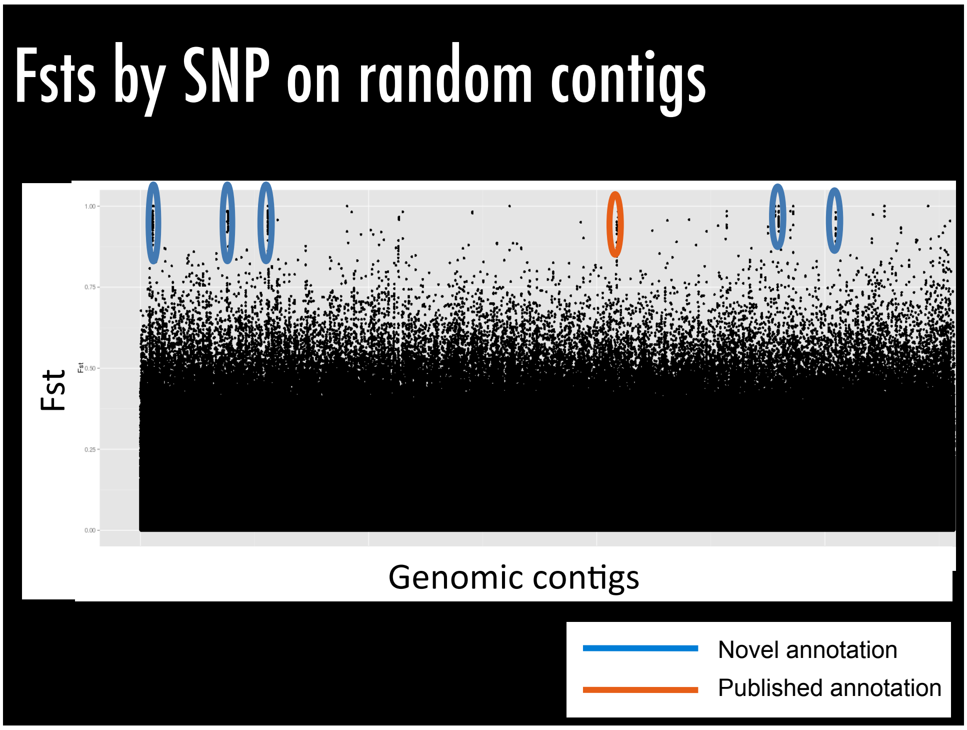
Fst distribution

vs.



Fst distribution

vs.



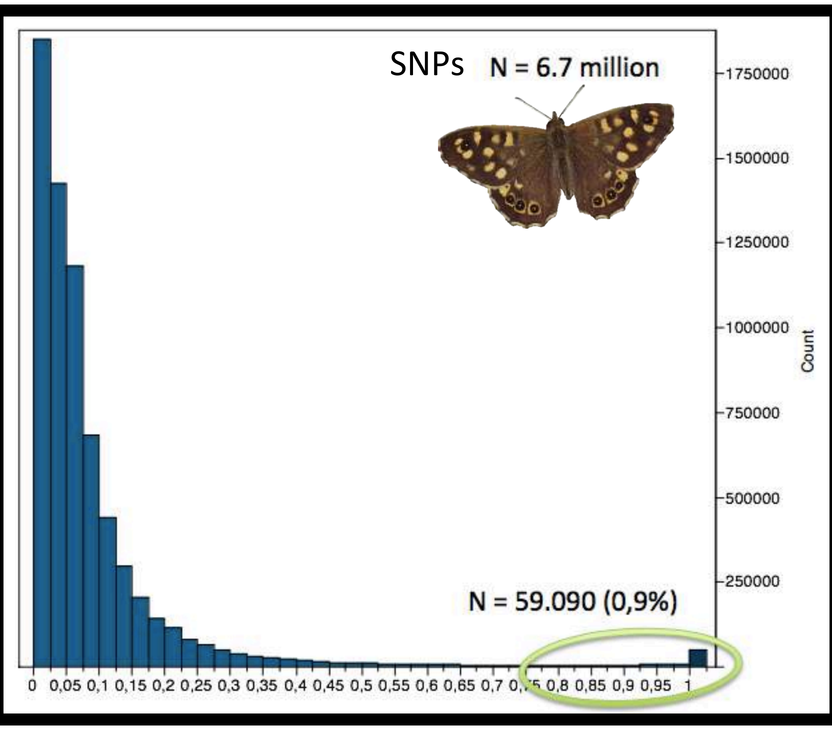
Fsts by SNP on random contigs

- Novel annotation
- Published annotation

Results

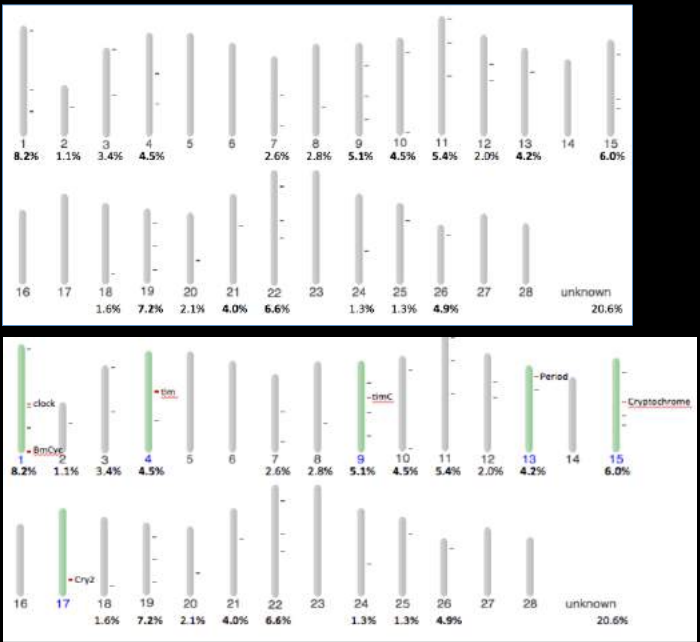
- Identified *doublesex* immediately
- Found new genes missed in pubication
  - p270
  - RNA directed DNA polymerase
  - Arginine/Serine rich coiled-coil protein 2

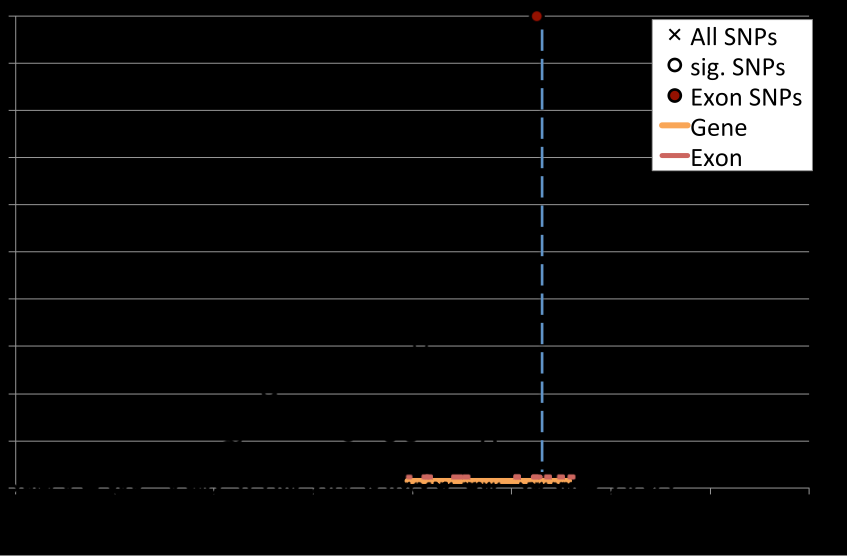
– Now searching to see if these are all near each other



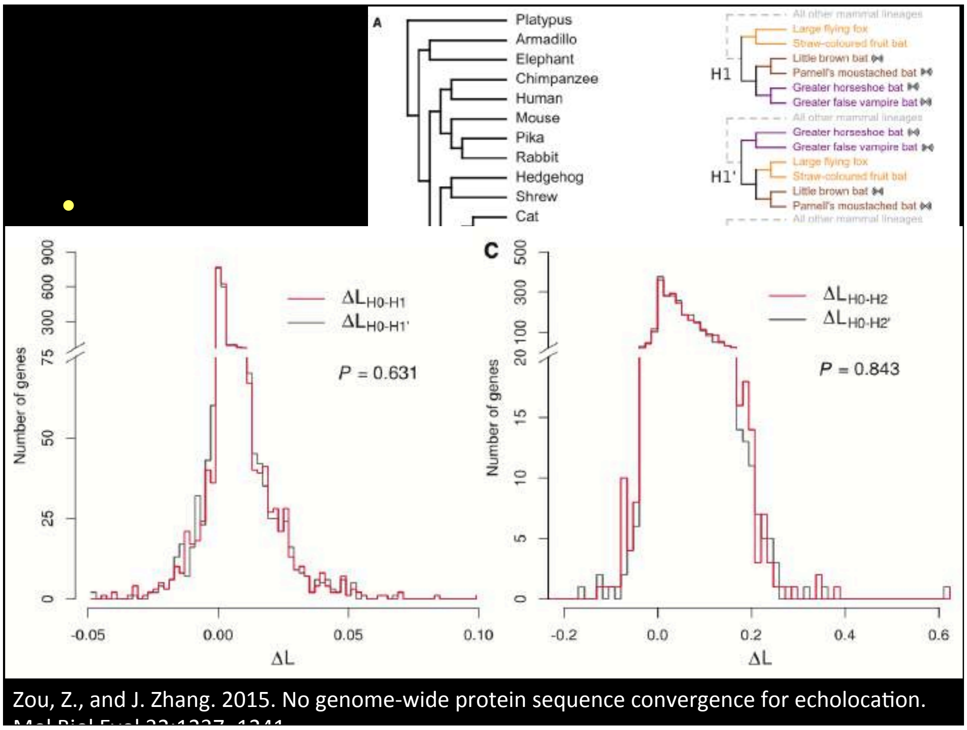
SNPs N = 6.7 million

N = 59.090 (0,9%)


1 8.2%
2 1.1%
3 3.4%
4 4.5%
5
6
7 2.6%
8 2.8%
9 5.1%
10 4.5%
11 5.4%
12 2.0%
13 4.2%
14
15 6.0%
16
17
18 1.6%
19 7.2%
20 2.1%
21 4.0%
22 6.6%
23
24 1.3%
25 1.3%
26 4.9%
27
28
unknown 20.6%
clock
BmCyc
tim
timC
Period
Cryptochrome
Cry2


× All SNPs
○ sig. SNPs
● Exon SNPs
Gene
Exon

•



Zou, Z., and J. Zhang. 2015. No genome-wide protein sequence convergence for echolocation. Mol Biol Evol 32:1237–1241

De novo RNA-Seq: Do you need a genome?

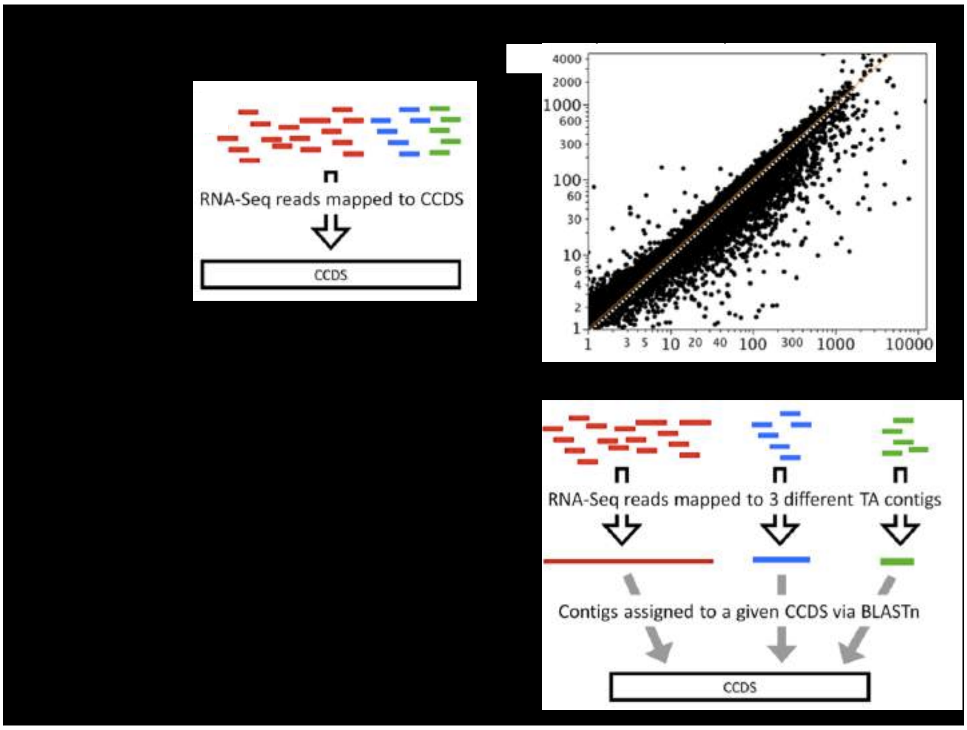
No, but there are important biases & limitations

- TA mapping limitations
  - No exon level resolution but this will change soon
  - No coding information on identified SNPs unless you build gene feature files on contigs
- TA mapping biases unique to it
  - Spicing may cause mapping problems if locus is collapsed, but generally OK to not assume a gene model
- TA mapping biases shared with genomic mapping
  - SNP and indel effects
  - gene duplication (are reads mapping to the right place)

Map to TA vs. Genome: which is better?

Template effects:

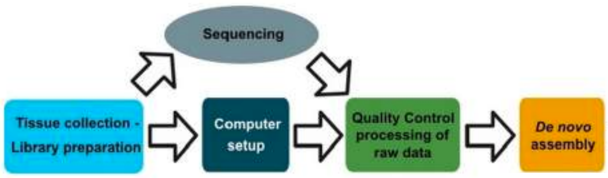
- Mismatch :
  - – SNPs (single nucleotide polymorphisms)
  - – Indels (insertion or deletion polymorphisms)
- Pseudo-inflation
  - – An increase in the copy number of a gene that arise from genome assembly errors or TA errors
- Gene model errors
  - – If the models in your genome are bad, this will affect results


RNA-Seq reads mapped to CCDS
CCDS
4000
2000
1000
600
300
100
60
30
10
6
4
2
1
1
3
5
10
20
40
100
300
1000
10000
RNA-Seq reads mapped to 3 different TA contigs
Contigs assigned to a given CCDS via BLASTn
CCDS

Duplication levels in RNA-Seq data

- Common in transcriptome work
- Starting with lots of high quality RNA increases
  - mRNA amount for sequencing
  - Decreases need of core facility to PCR your sample
- Moderate amounts of PCR duplication are OK
  - ~ 20% expected
  - > 50% perhaps problematic if correlated with experimental design
  - Clone_filter program in STACKS is excellent assessing this

Pipeline Overview



Sequencing

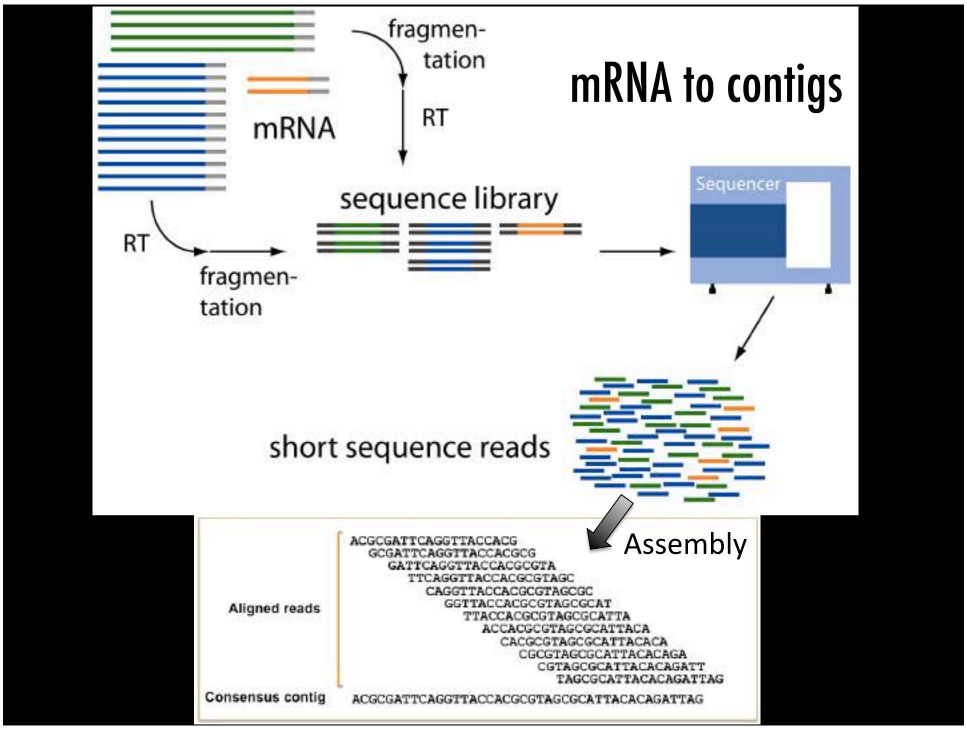
Tissue collection - Library preparation

Computer setup

Quality Control processing of raw data

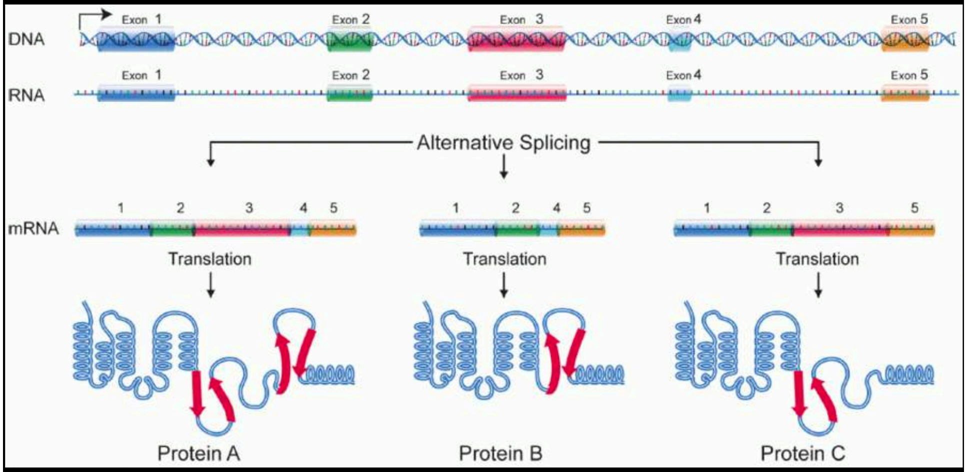
De novo assembly

mRNA to contigs



```
Aligned reads     ACGCGATTCAGGTTACCACG
                    GCGATTCAGGTTACCACGCG
                      GATTCAGGTTACCACGCGTA
                        TTCAGGTTACCACGCGTAGC
                          CAGGTTACCACGCGTAGCGC
                            GGTTACCACGCGTAGCGCAT
                              TTACCACGCGTAGCGCATTA
                                ACCACGCGTAGCGCATTACA
                                  CACGCGTAGCGCATTACACA
                                    CGCGTAGCGCATTACACAGA
                                      CGTAGCGCATTACACAGATT
                                        TAGCGCATTACACAGATTAG
Consensus contig  ACGCGATTCAGGTTACCACGCGTAGCGCATTACACAGATTAG
```

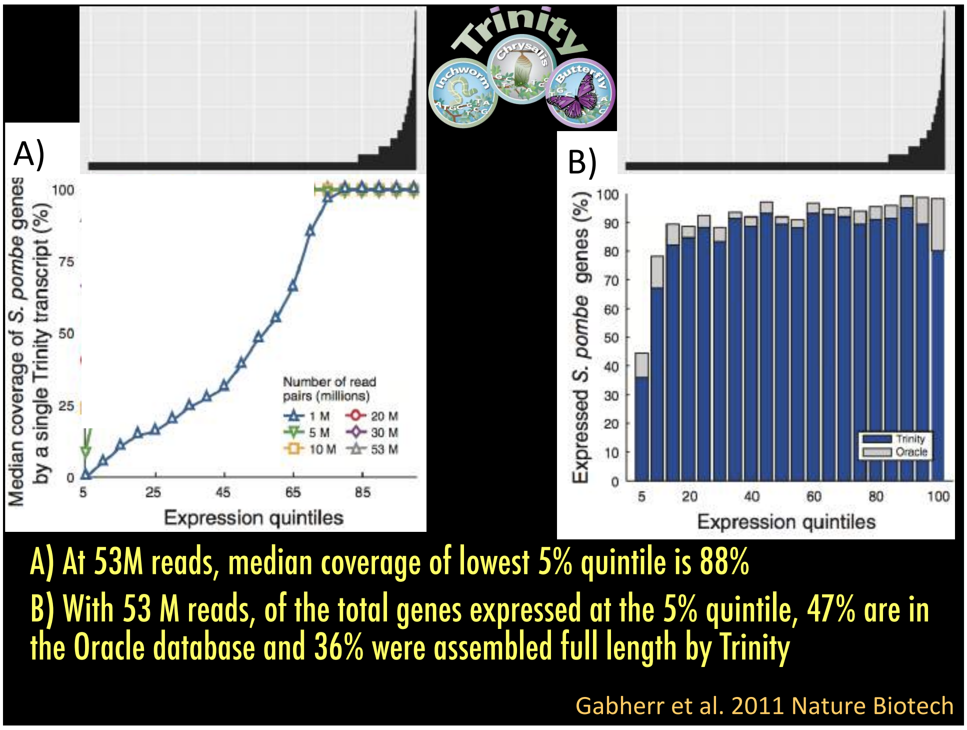
Alternative splicing complicates everything



H. sapiens: > 95% of multi-exonic genes are spliced

De novo transcritpome assembly

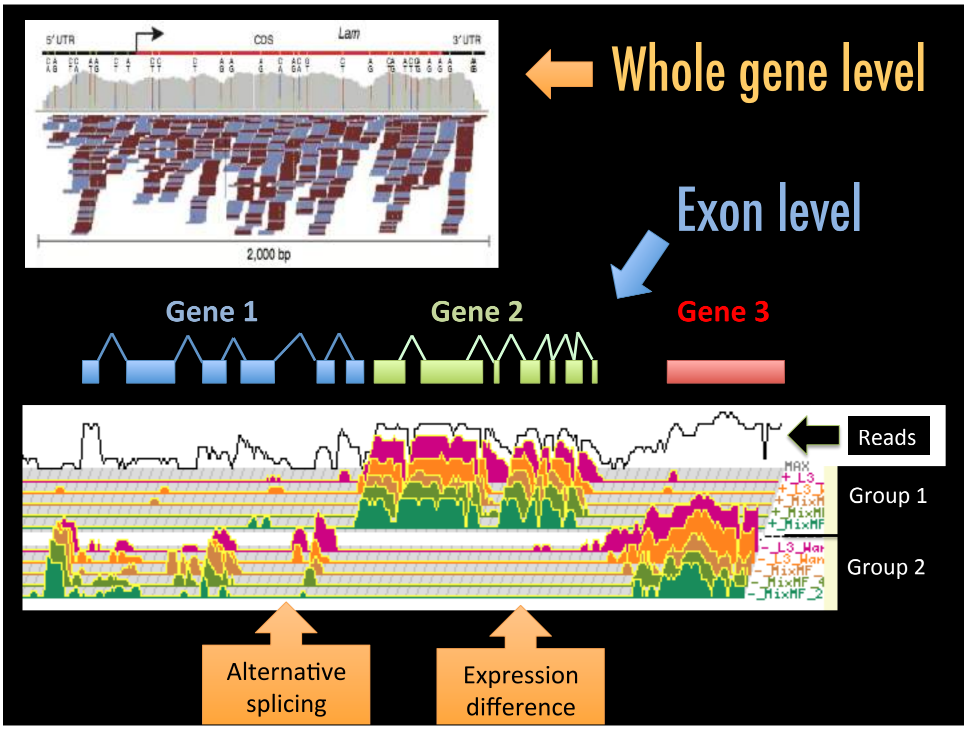
Reconstructs splice isoforms using PE Illumina data



A) At 53M reads, median coverage of lowest 5% quintile is 88%

B) With 53 M reads, of the total genes expressed at the 5% quintile, 47% are in the Oracle database and 36% were assembled full length by Trinity

Gabherr et al. 2011 Nature Biotech



Whole gene level

Exon level

Gene 1

Gene 2

Gene 3

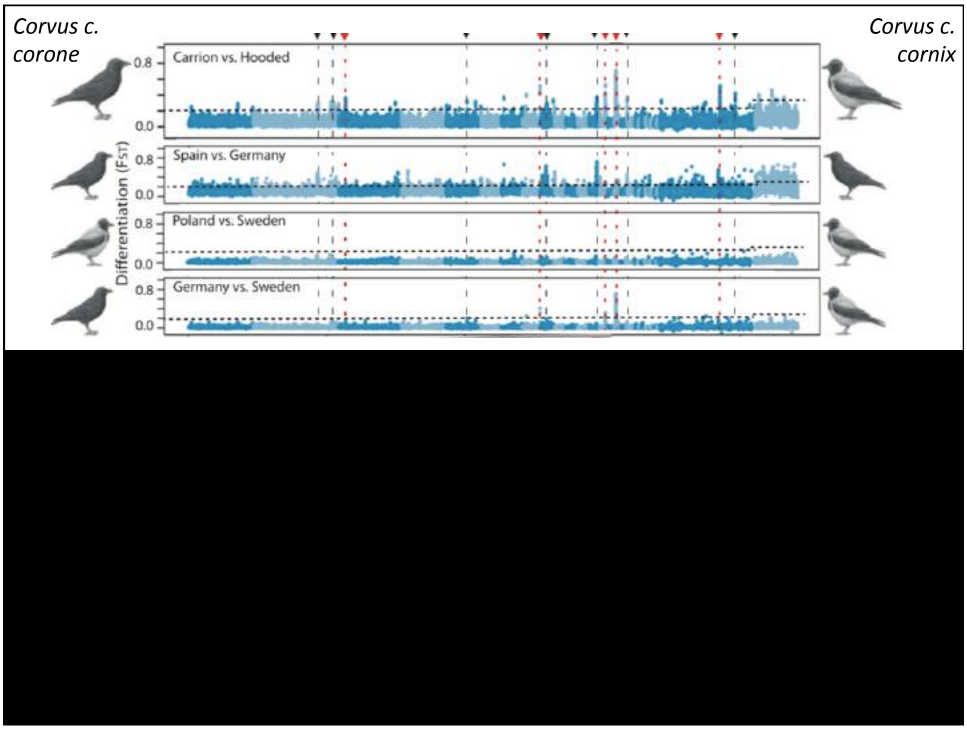
Reads

Group 1

Group 2

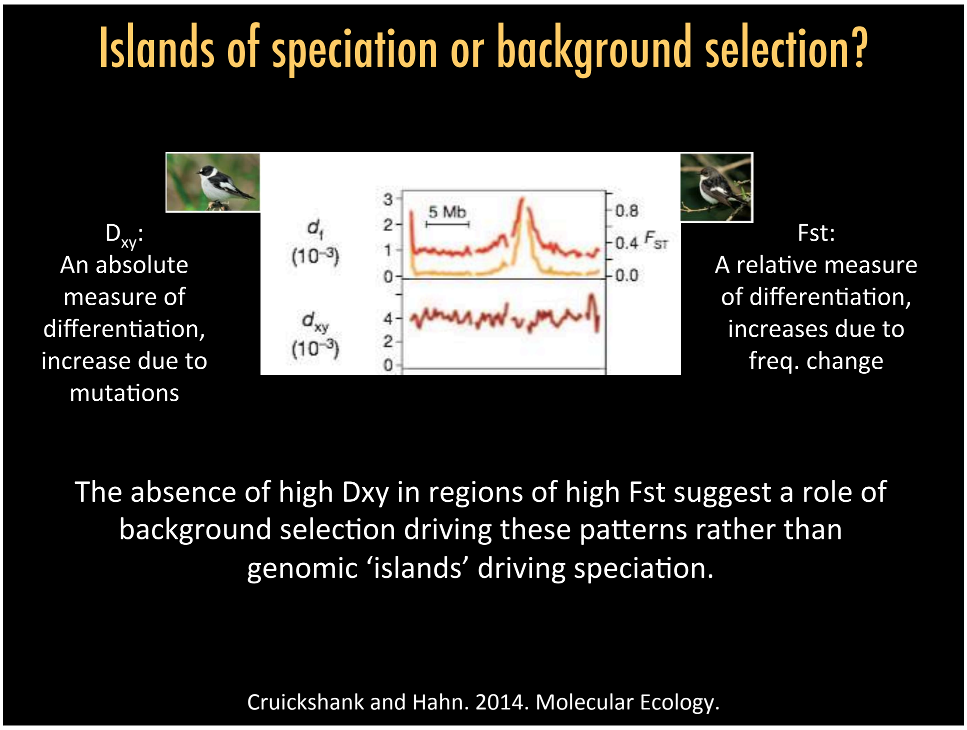
Alternative splicing

Expression difference



Corvus c. corone

Corvus c. cornix



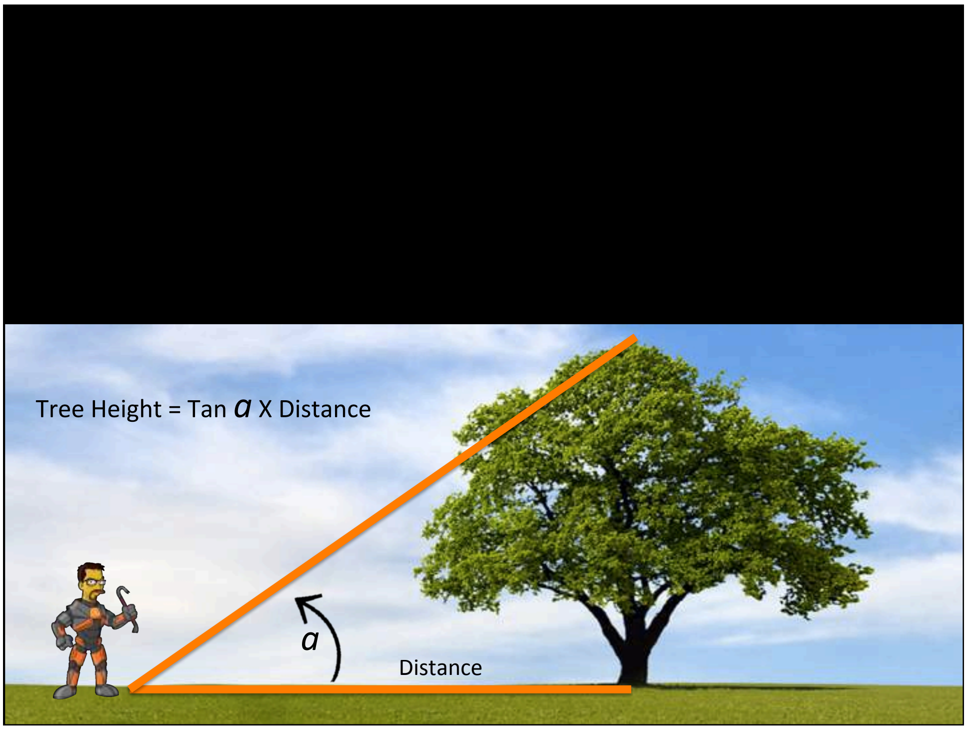
Islands of speciation or background selection?

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

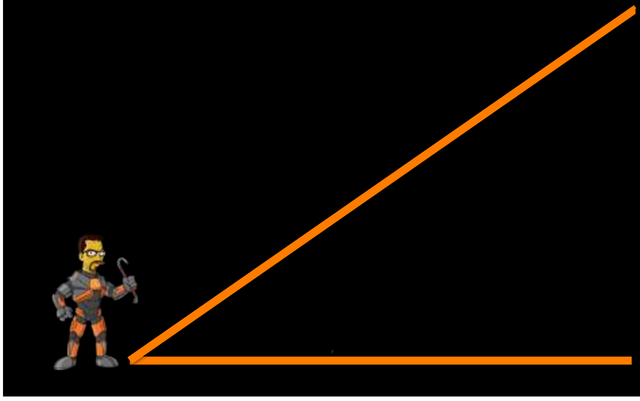
Fst:
A relative measure of differentiation, increases due to freq. change

The absence of high Dxy in regions of high Fst suggest a role of background selection driving these patterns rather than genomic ‘islands’ driving speciation.

Cruickshank and Hahn. 2014. Molecular Ecology.

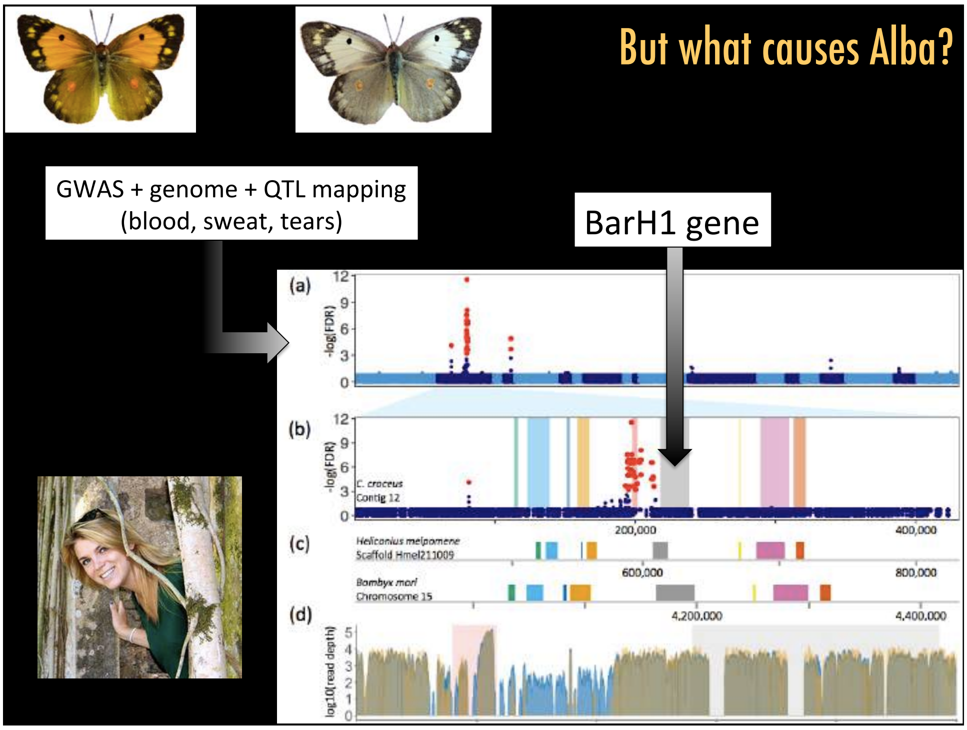


Tree Height = Tan a X Distance



Is it an adaptation?

Colias croceus, the Clouded Yellow



But what causes Alba?

GWAS + genome + QTL mapping
(blood, sweat, tears)

BarH1 gene