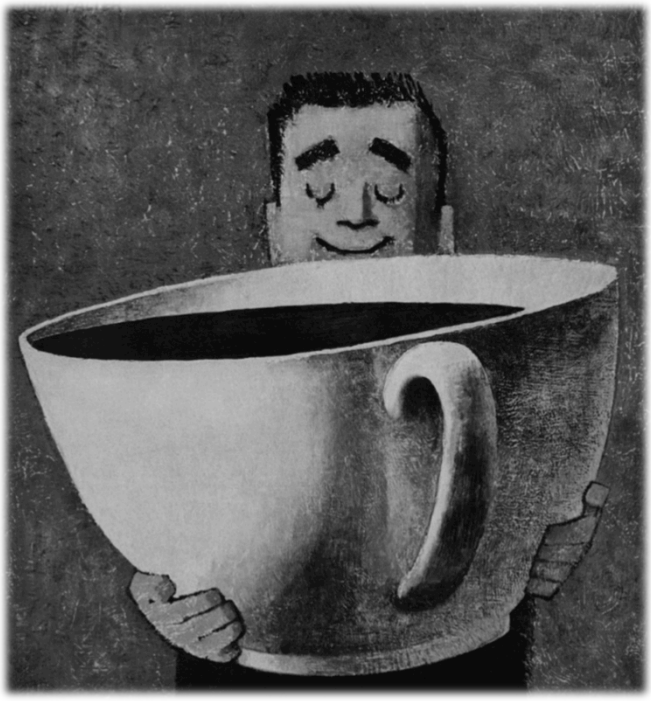
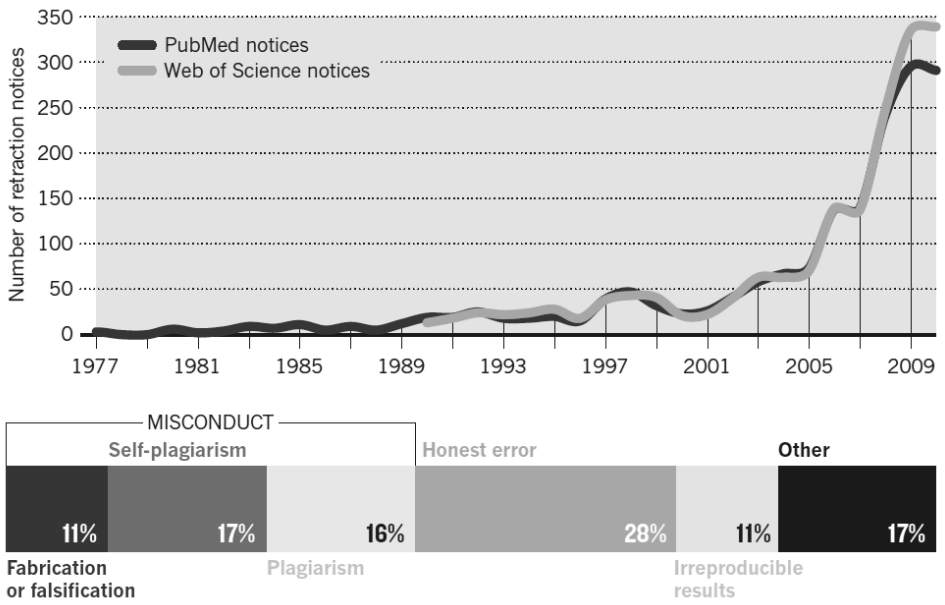
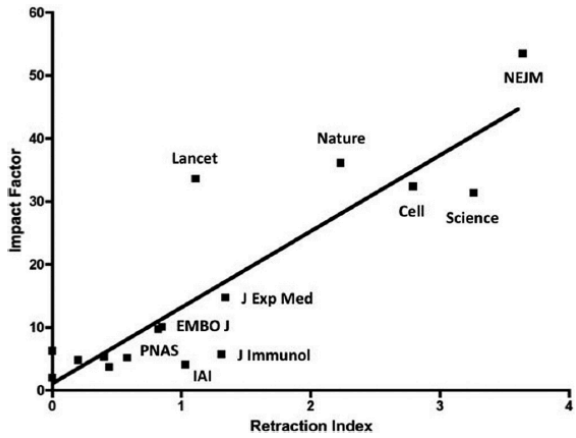


A surge in withdrawn papers is highlighting weaknesses in the system for handling them.



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Publications with significant human error that have not been retracted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

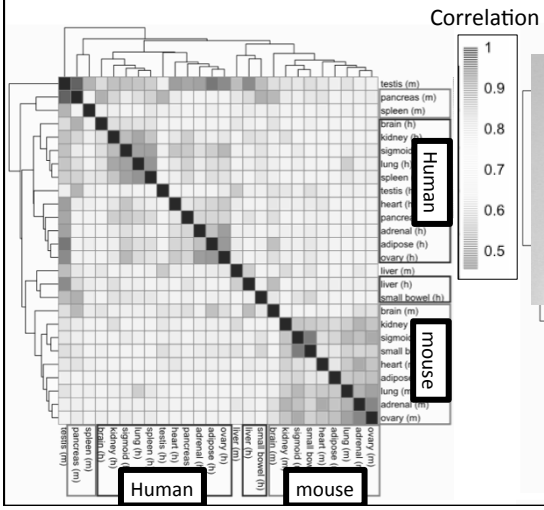
Snyder mouse controversy

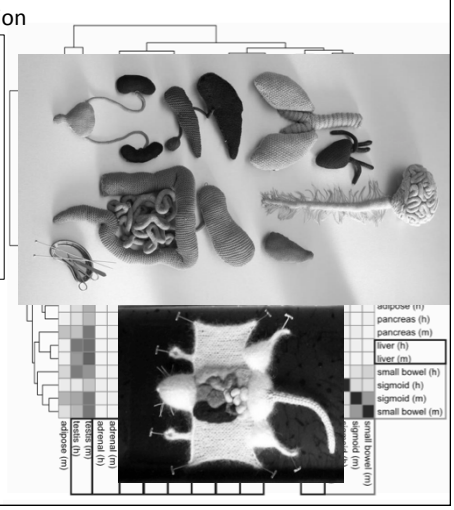
"the expression for many sets of genes was found to be more similar in different tissues within the same species than between species" Lin et al. 2014 PNAS

Human – Mouse TMRCA ~ 90 MYA

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

Brain – Kidney TMRCA?



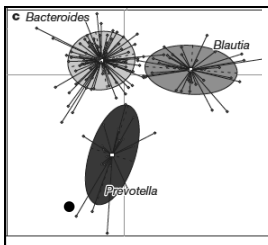


Batch effect: confounding sequencing grouping with biological grouping

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?



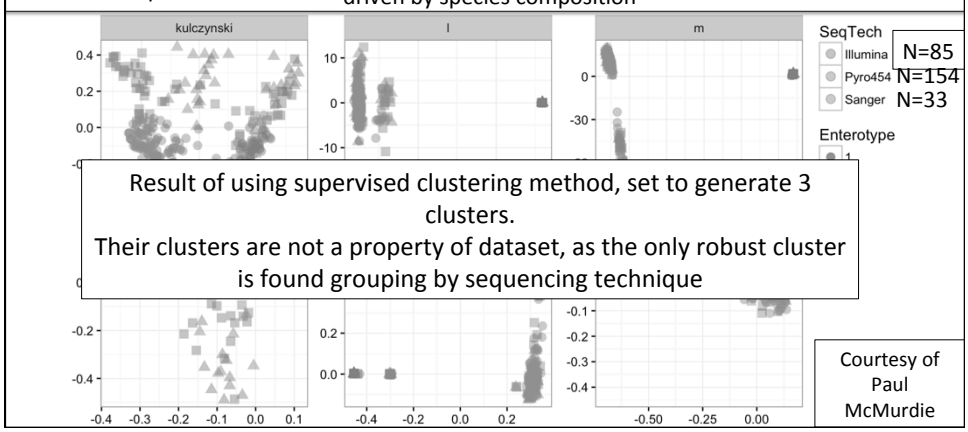
ARTICLE 174 | NATURE | VOL 473 | 12 MAY 2011

doi:10.1038/nature09944

Enterotypes of the human gut microbiome

154 pyrosequencing-based 16S sequences

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

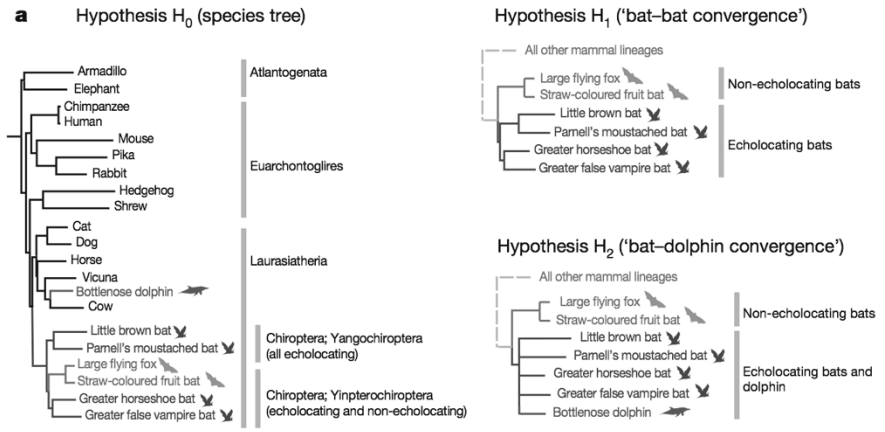


Result of using supervised clustering method, set to generate 3 clusters.
Their clusters are not a property of dataset, as the only robust cluster is found grouping by sequencing technique

Courtesy of Paul McMurdie

LETTER 228 | NATURE | VOL 502 | 10 OCTOBER 2013
doi:10.1038/nature12511

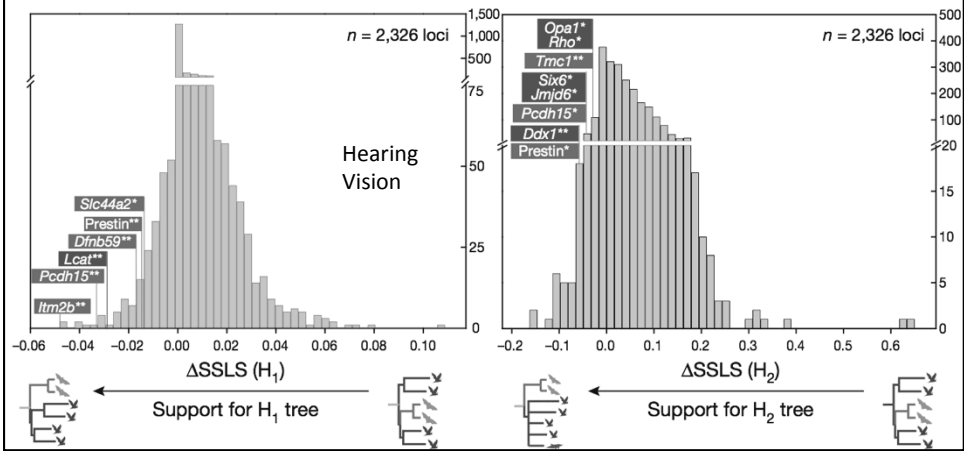
Genome-wide signatures of convergent evolution in echolocating mammals



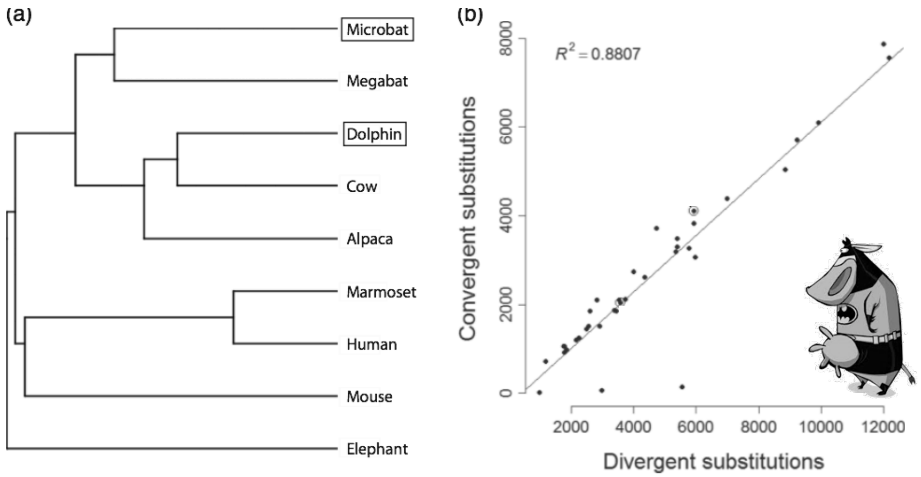
Parker et al. 2013. Nature 502:228–231.

"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%

[redacted]

PNAS

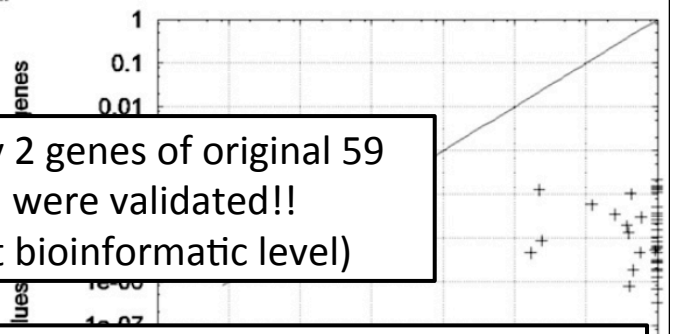
More genes underwent positive selection in chimpanzee evolution than in human evolution

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 syno
No. of nons
Mean ω of
Mean ω of



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

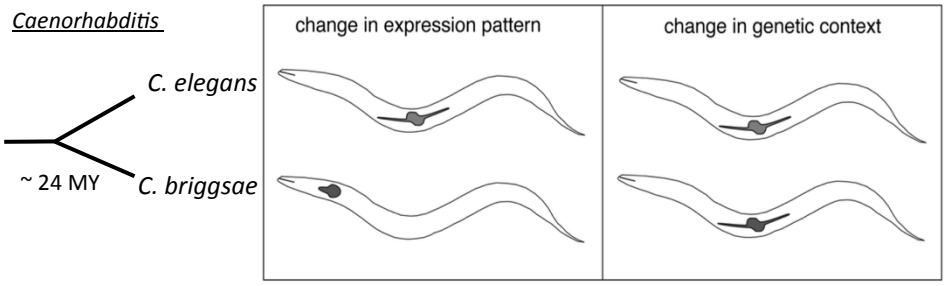
Evolutionary Inference = House of Cards?

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

1001 ways for your pipeline to break, or feed you sewage

An overview of genomic pipeline challenges

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

Batcow says, take a break!!!!

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

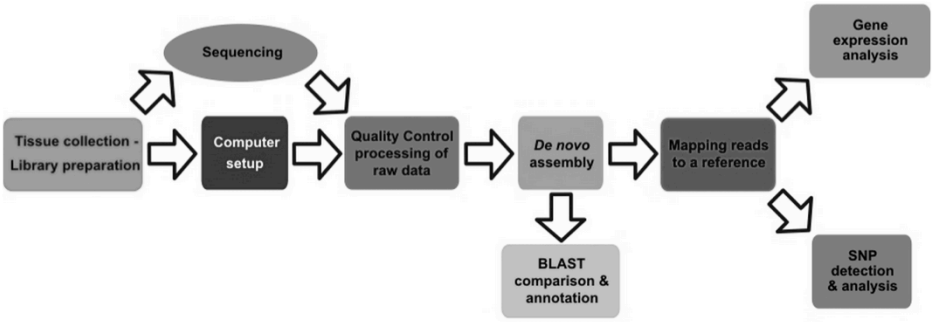
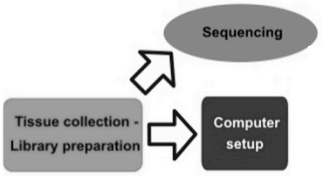


Image from http://sfg.stanford.edu/guide.html

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

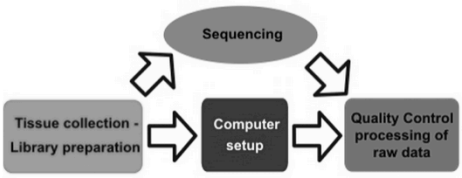
RNAseq dataset:

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

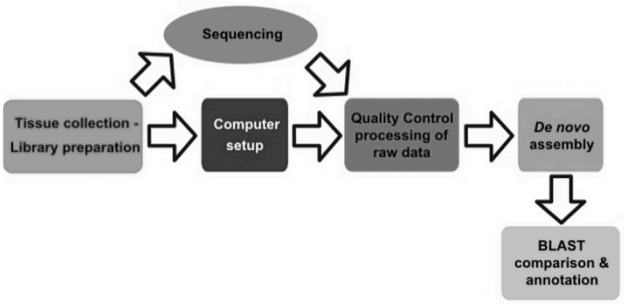


Core facilities and non-model species

Statements from core facilities that are not true:

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

Pipeline Overview

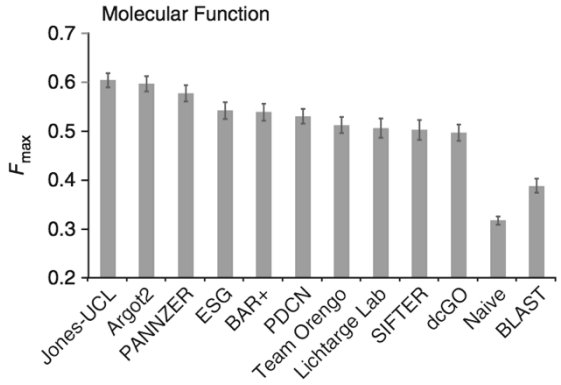


Gene Ontology: order in the chaos

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

http://www.geneontology.org/

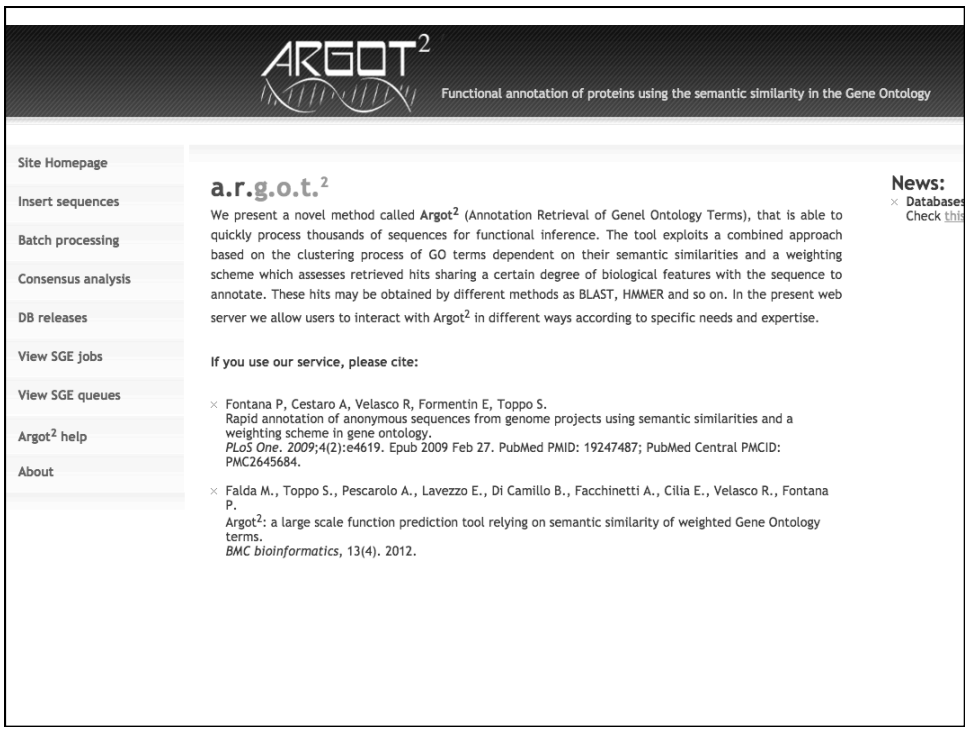
Comparisons among annotation tools



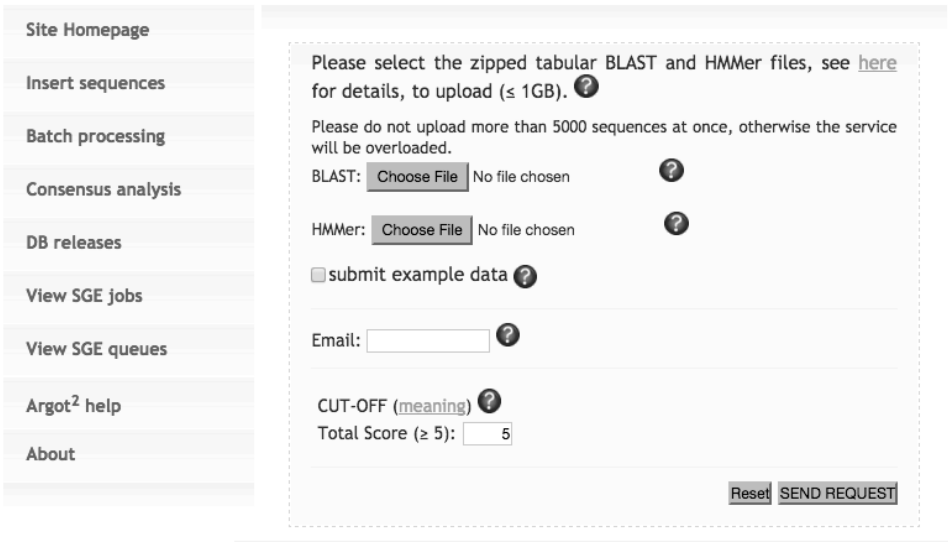
Radivojac et al.: . *Nat Meth* 2013, :221–227.

Falda et al.

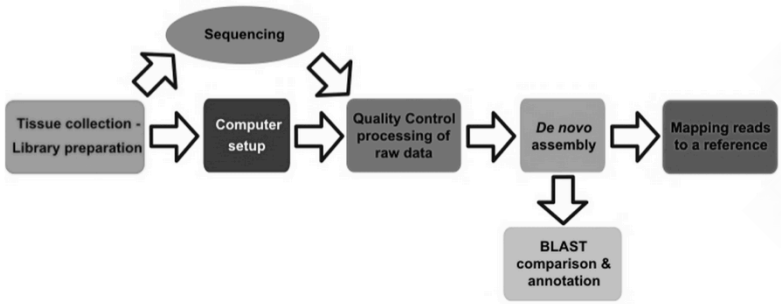
weighted Gene Ontology terms. *BMC Bioinformatics* 2012, **13**:S14.



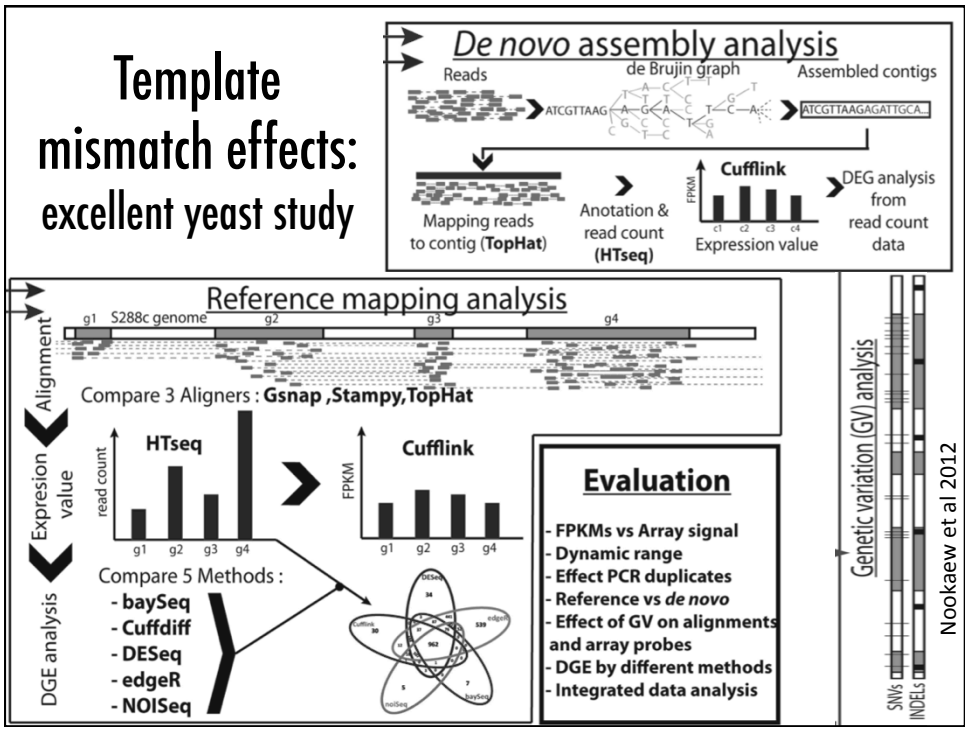
Batch processing for GO terms



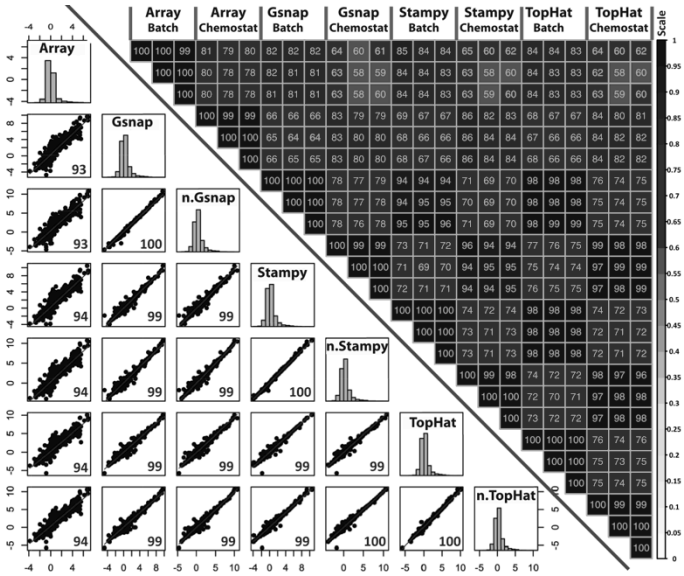
Pipeline Overview



Template mismatch effects: excellent yeast study



Does alignment software matter?



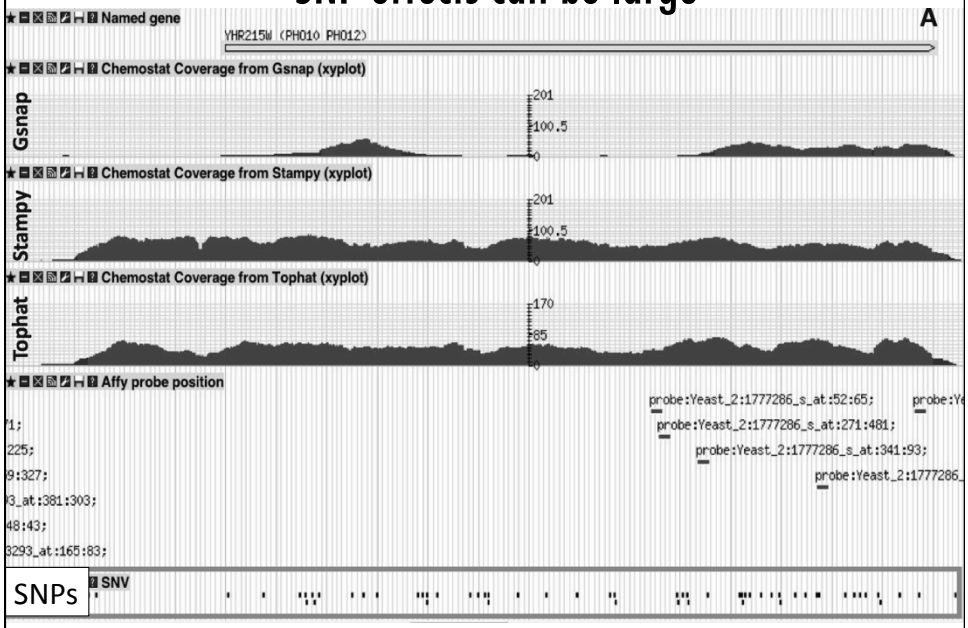
Nookaew et al. **A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae**. *Nucleic Acids Research* 2012, **40**:10084–10097.

Mappers don't appear to matter

Wrong

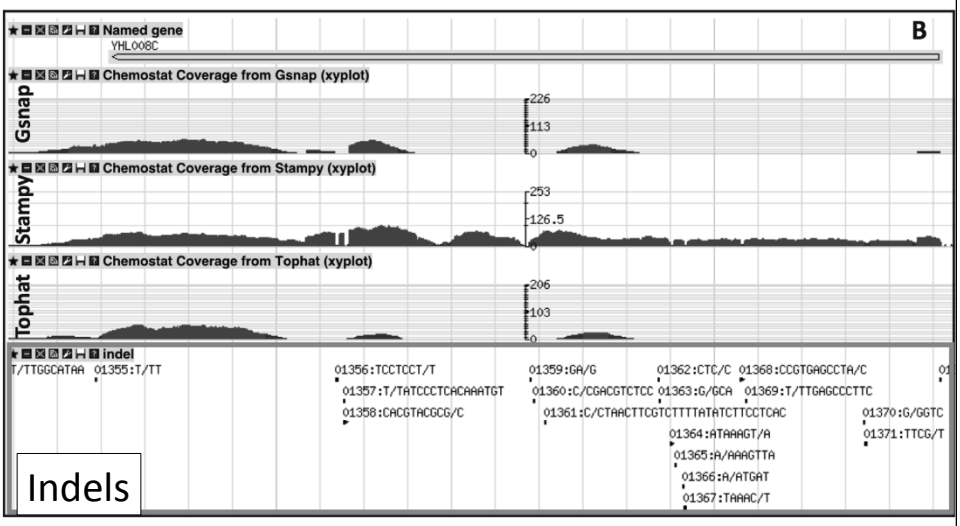
- Genomic scale data can hide widespread biases that unless you specifically look, are hard to find
- Mapping programs differ in their settings and design
  - DNA to DNA vs. RNA to DNA
  - Are usually compared using species without much genetic variation
  - Indels, splicing, SNPs all affect mapper performance

SNP effects can be large



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

Insertions & deletions (indels) have large effects



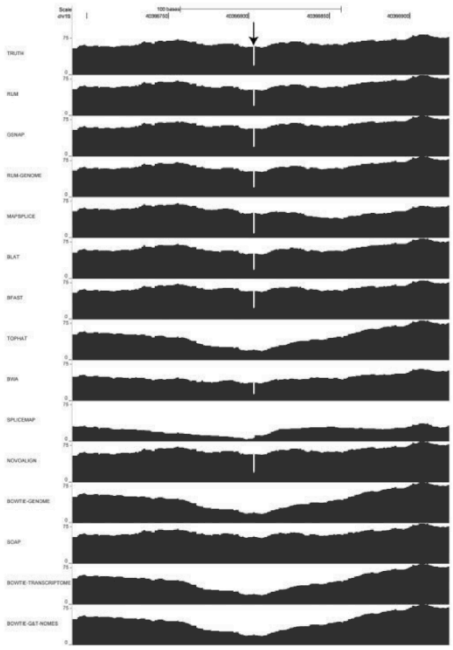
Nookaew et al. **A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae**. *Nucleic Acids Research* 2012, **40**:10084–10097.

15 mapping results

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

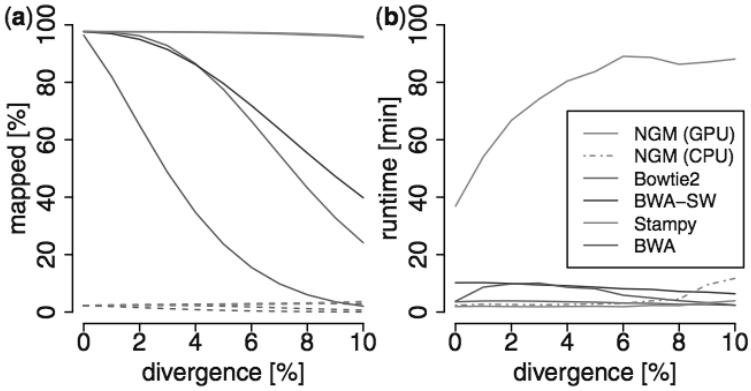
TopHat, SpliceMap, Bowtie and Soap

- 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.

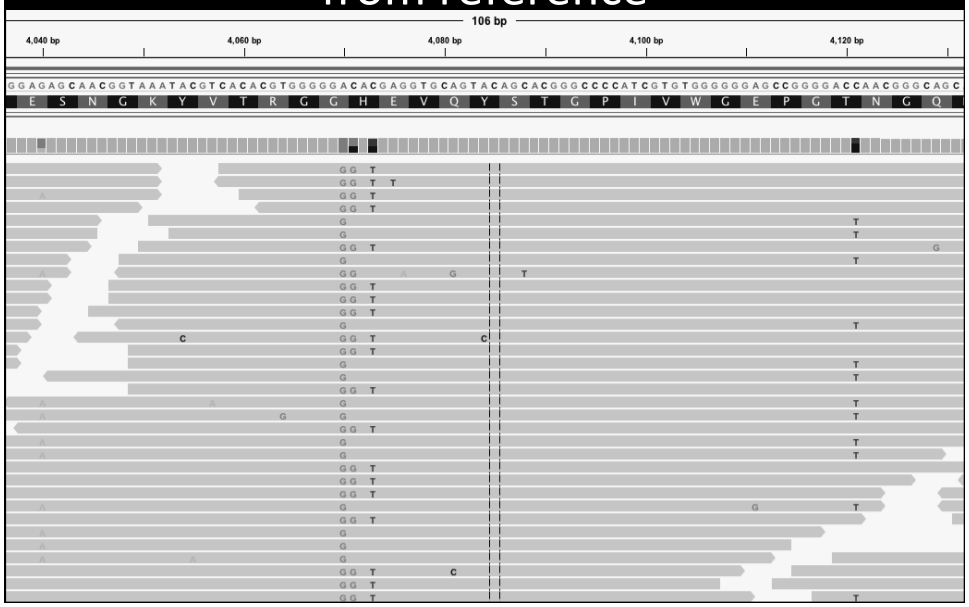
Allelic bias in read mapping



- 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?

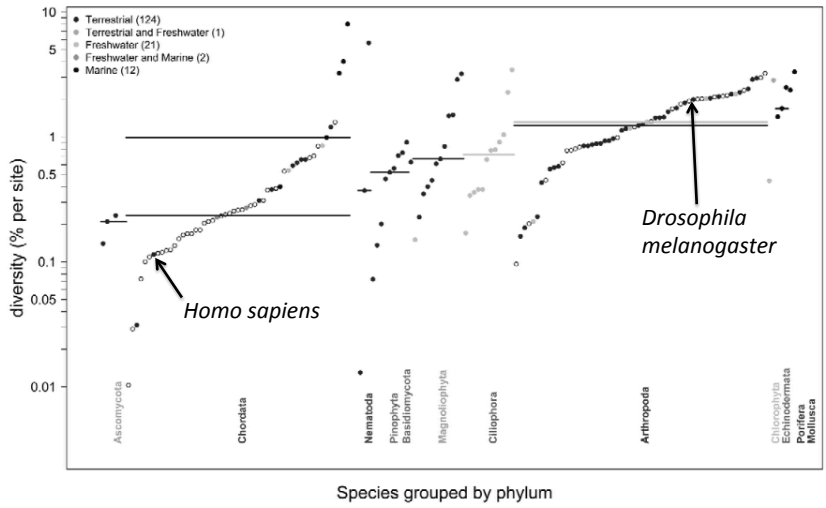
Sedlazeck et al. 2013 *Bioinformatics*

100 bp window with 4 – 5 SNPs differing from reference



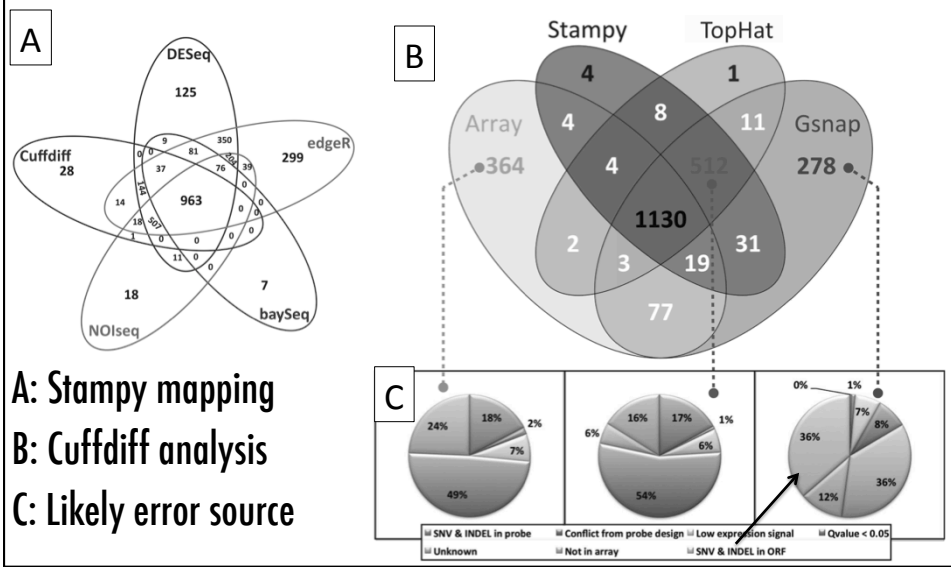
Mapping reads in outbred species

Average genome polymorphism levels (ignores indels)



Leffler *et al.* 2012 *Plos Biol*

Sig. expression differences by method

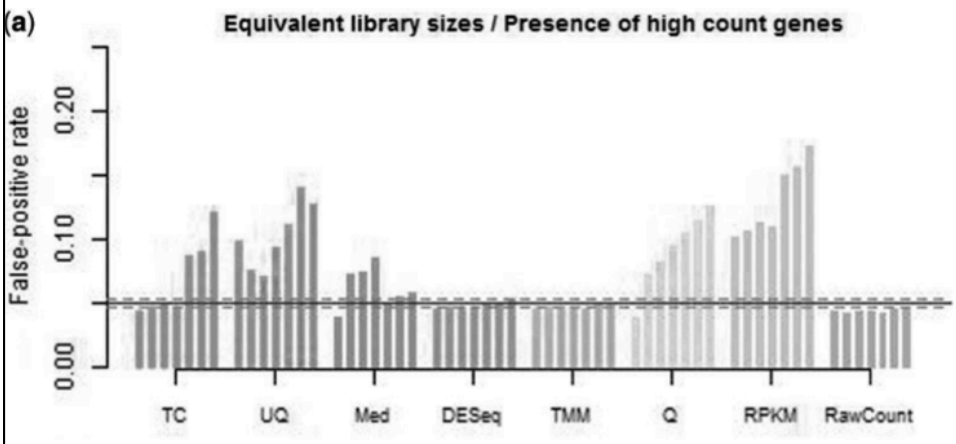


A: Stampy mapping

B: Cuffdiff analysis

C: Likely error source

Normalization matters, as it directly affects false-positive rate

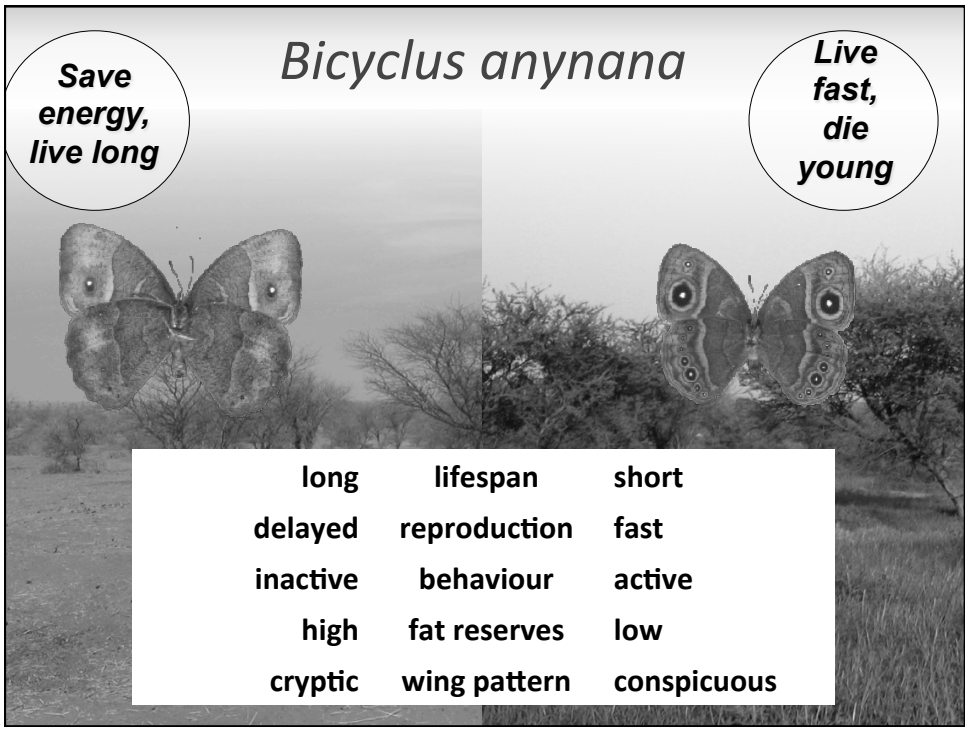


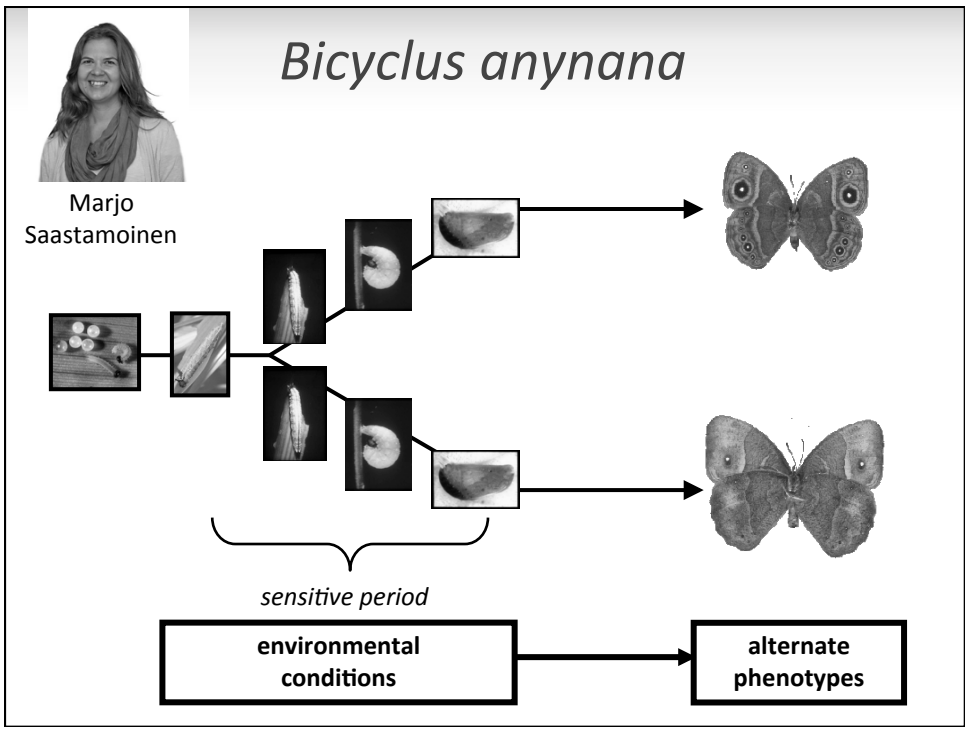
Dillies 2010 Brief in Bioinfo.

RNA-Seq

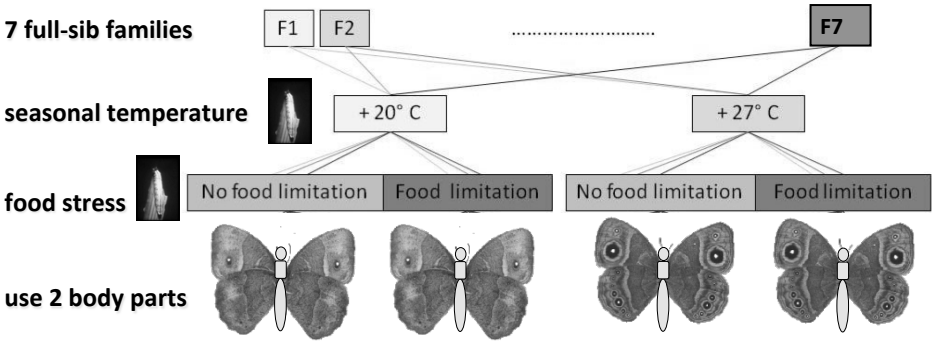
Real world example

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



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

Vicencio Oostra

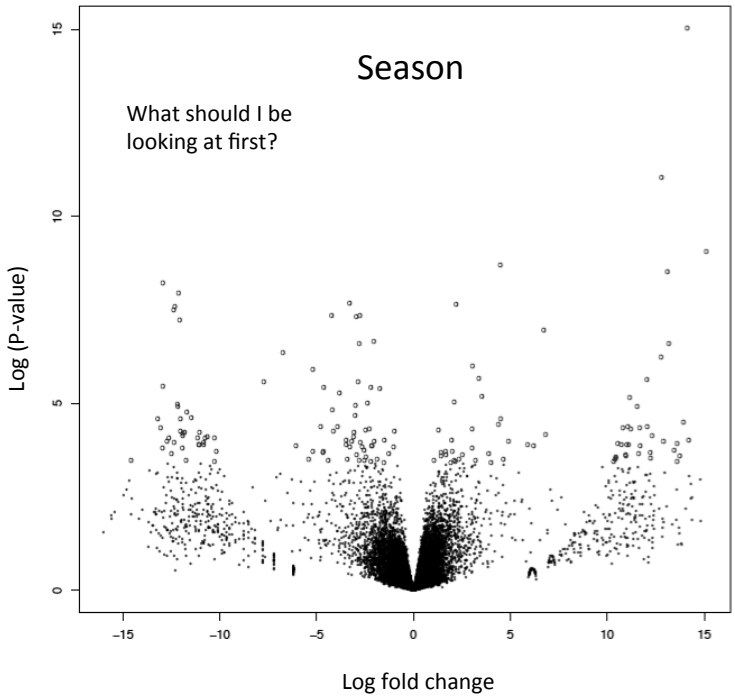
body part	# libraries	# clean reads (per library)	# nucleotides (per library)	GC content
abdomen	72	15,261,019	3,052,203,767	45%
thorax	72	15,633,416	3,126,683,150	46%
total	144	2,224,399,290	444,879,858,000	45%

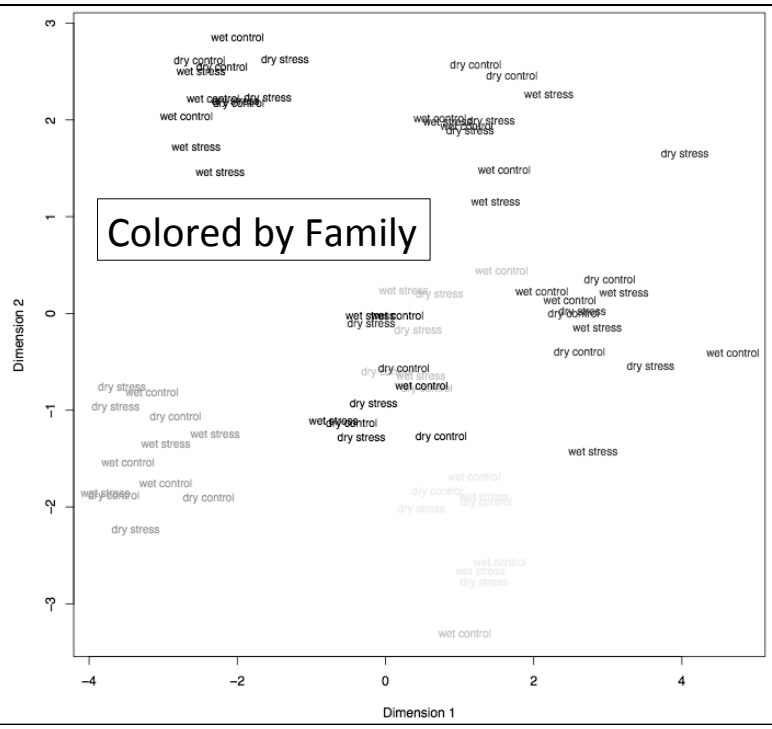
14 samples: one from each family, thorax and abdomen

69,075 contigs

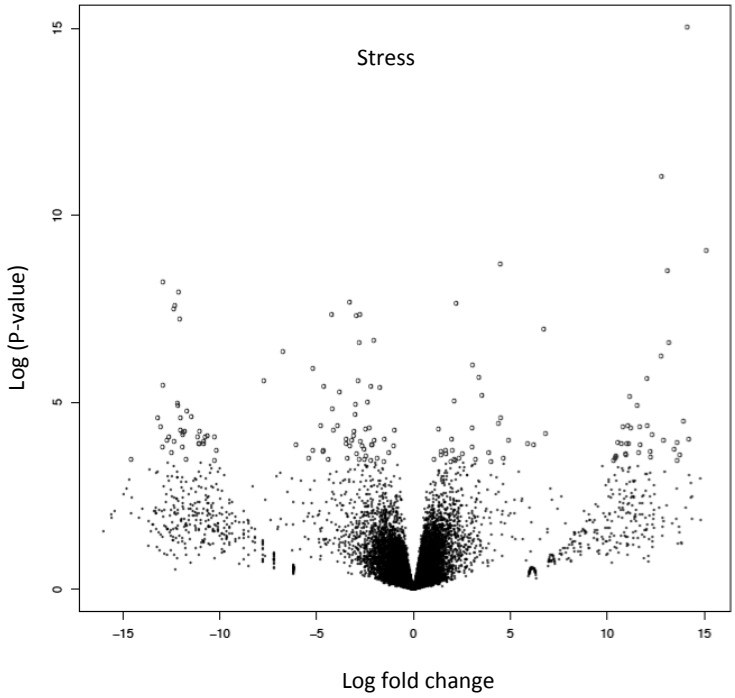
edgeR

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


Season
What should I be looking at first?
Log (P-value)
Log fold change

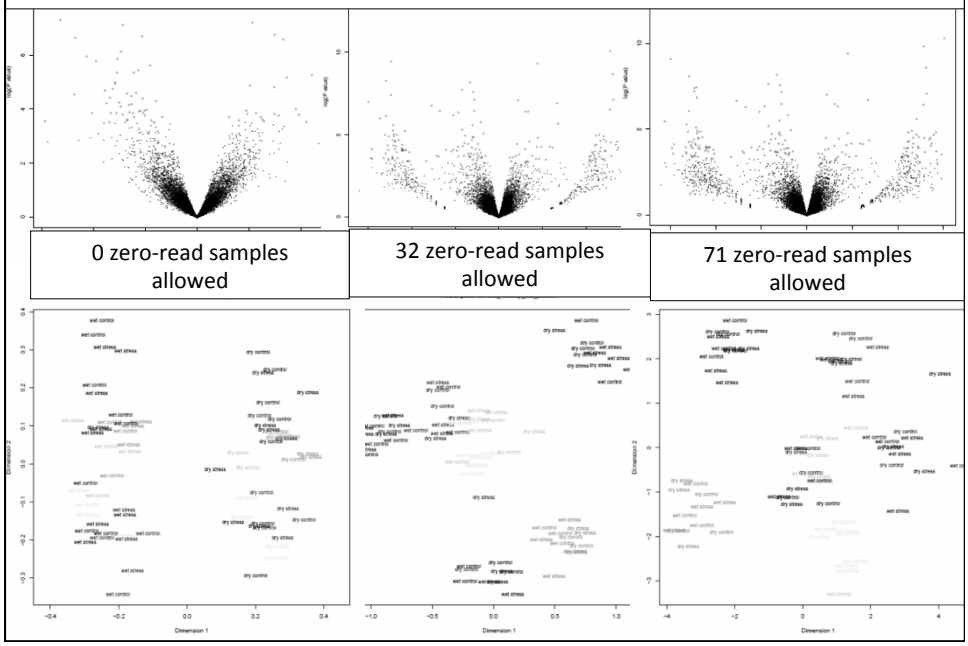

Colored by Family
Dimension 1
Dimension 2

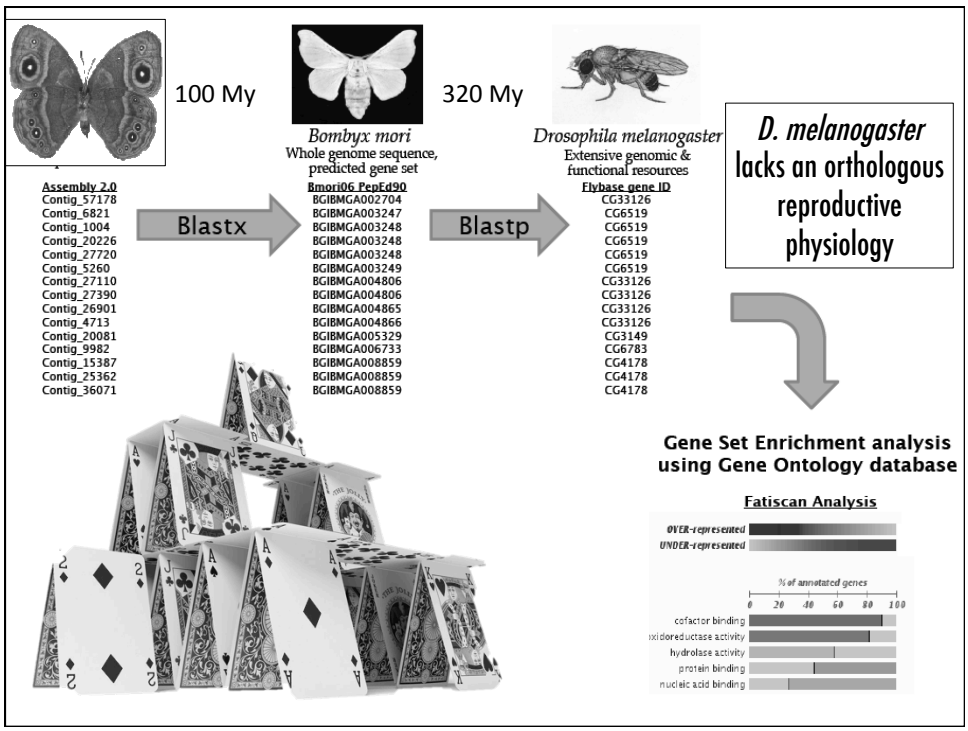
50


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


Log (P-value)
Log fold change

Effect of filtering the mapping to Trinity contigs





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 l	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

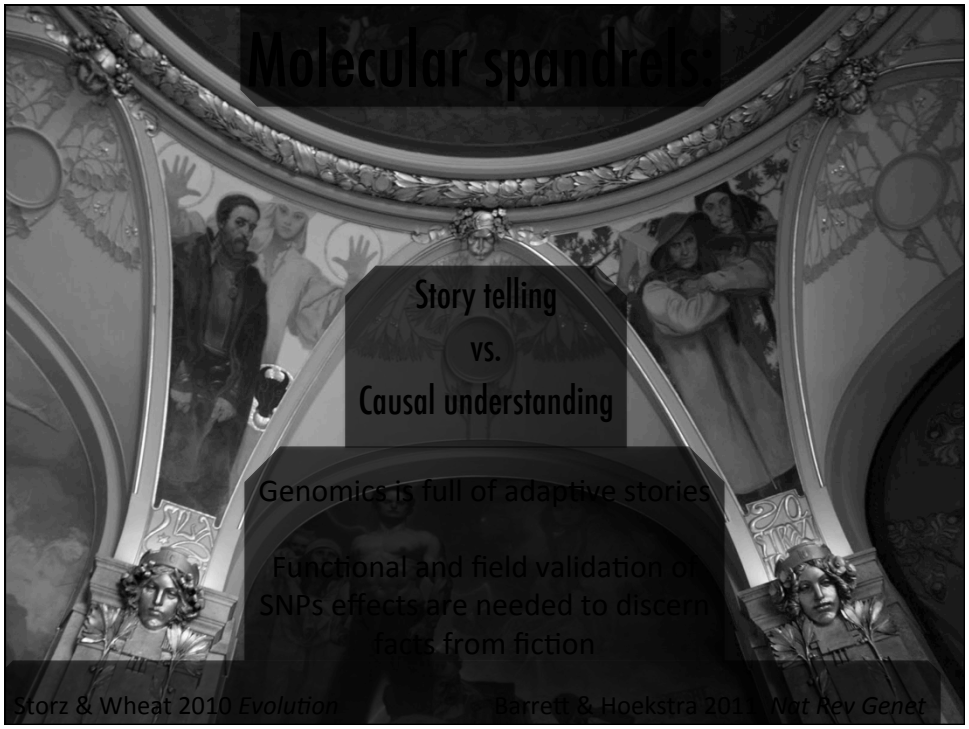
Should independent methods converge?

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!!!!



This is ongoing work

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

This is all due to the Workshop on Genomics gang, thanks to your unwavering support over the years!

Captain Skoot

The Gang

Stockholm University

Academy of Finland

Vetenskapsrådet
The Swedish Research Council

Knut och Alice Wallenbergs Stiftelse