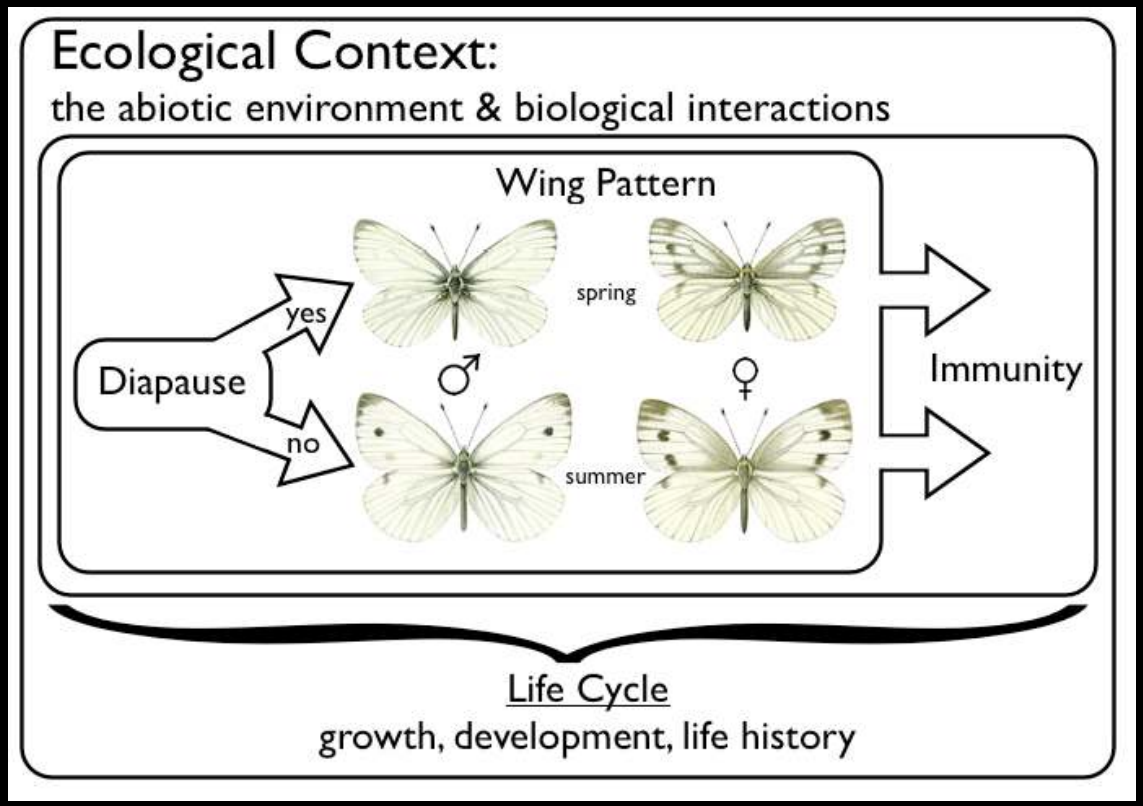
The second half

- Pool-Sequencing in species without a genome
  - What it is
  - Validation
  - An example from my lab

- RNA-Seq
  - Things to think about for those working in non-model species

Insect Life Cycle Functional Genomics



Pararge aegeria

Polygonia c-album

Lycaena tityrus

Colias eurytheme

Euphydryas editha

Our goal

- Find high quality candidate SNPs directly associated with local adaptations
- Spend our effort studying functional effects of SNPs on fitness in the lab and wild, not on finding the SNPs
- Learn how these adaptations are integrated

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

Pool-Seq approach in model species

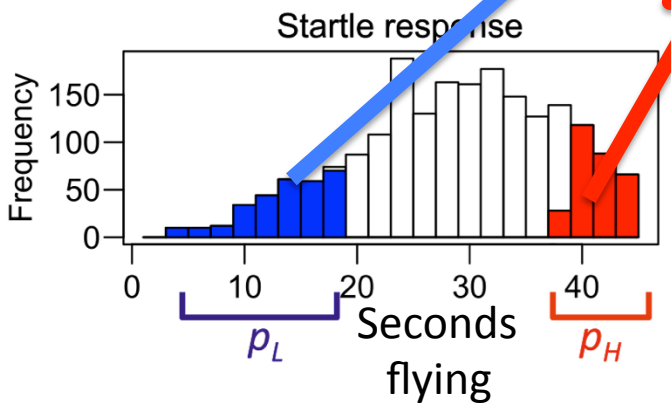
Low pool (n=300)

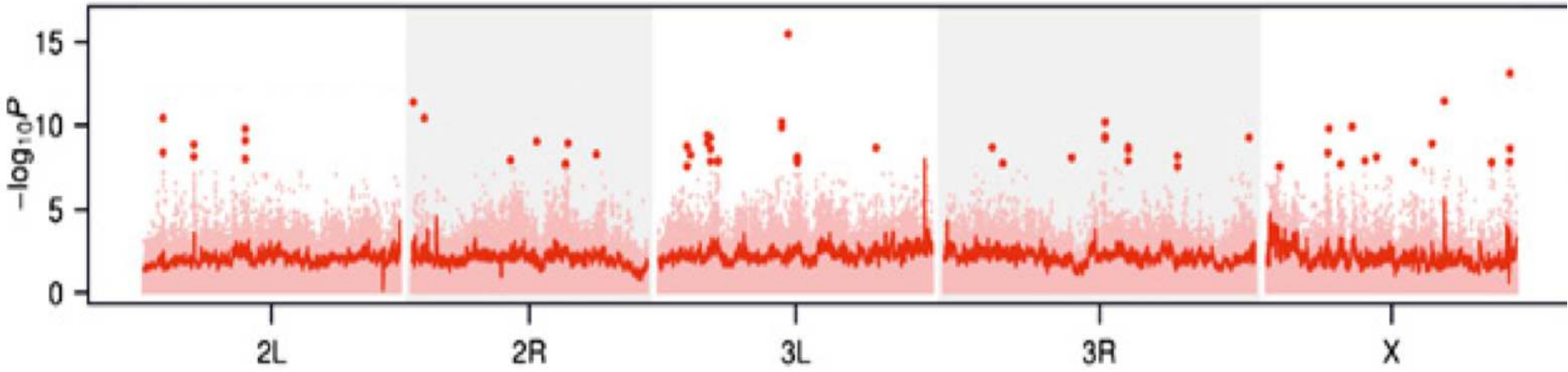
High pool (n=300)

Map to genome

Call SNPs

Scan genome for sig. allele freq. changes between groups





61 SNPs show sig association with startle response

Huang et al. 2012 PNAS

1001 ways for your pipeline to break

An overview of genomic pipeline challenges

Christopher West Wheat

Informatics and Biology

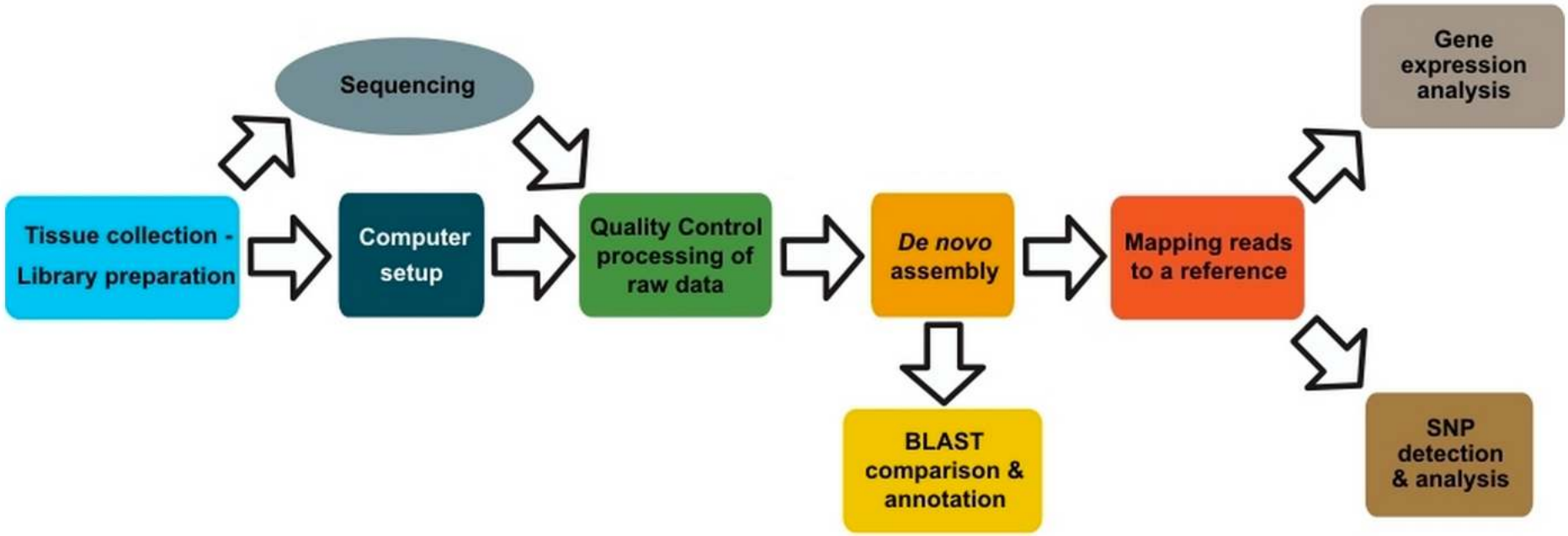
- We need to make sure we put the 'bio' into the bioinformatics
  - Do results pass 1st principals tests
  - Always double check data from your core facility or service company
  - Use independent analyses as 'controls' on accuracy
    - What are your + and – controls?
    - Do independent methods converge?

- Need to re-assess our common metrics for potential bias in the genomic age
  - Bootstraps on genomic scale data
  - P-values, outlier analyses, demographic null models

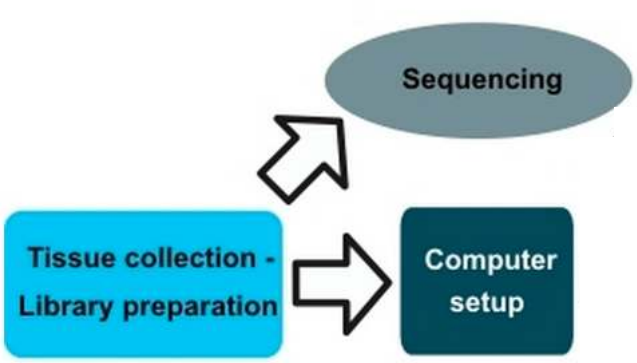
Outline

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

Pipeline Overview



Pipeline Overview



How can I study my data using open source?

How much RAM do I need?

Are 16 cores enough?

Can I use my laptop?

What software & how do I get it?

Why Linux?

How much HD space is needed?

Computer Infrastructure

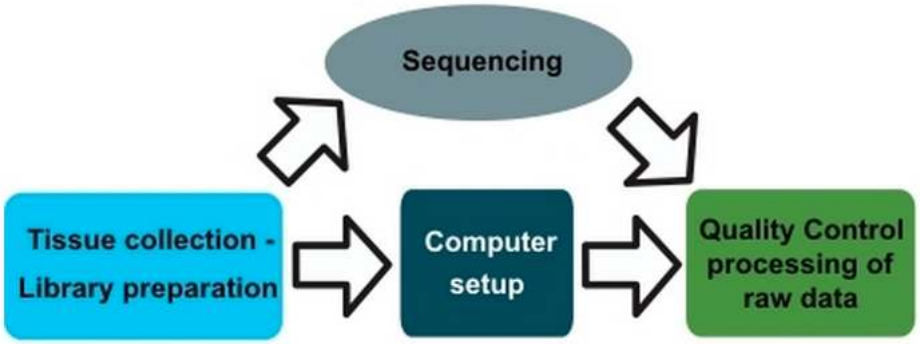
RNAseq dataset:

4 conditions X 2 tissues X 3 families X 3 replicates = 72 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

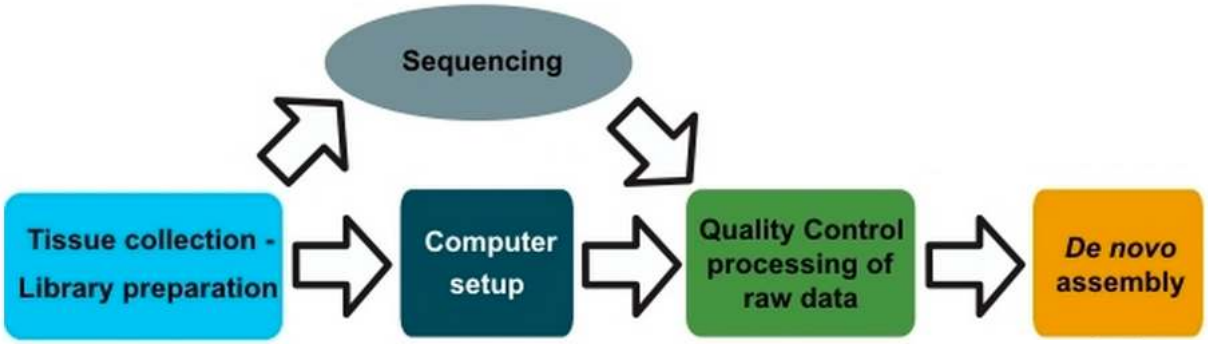
Statements from core facilities that are not true:

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

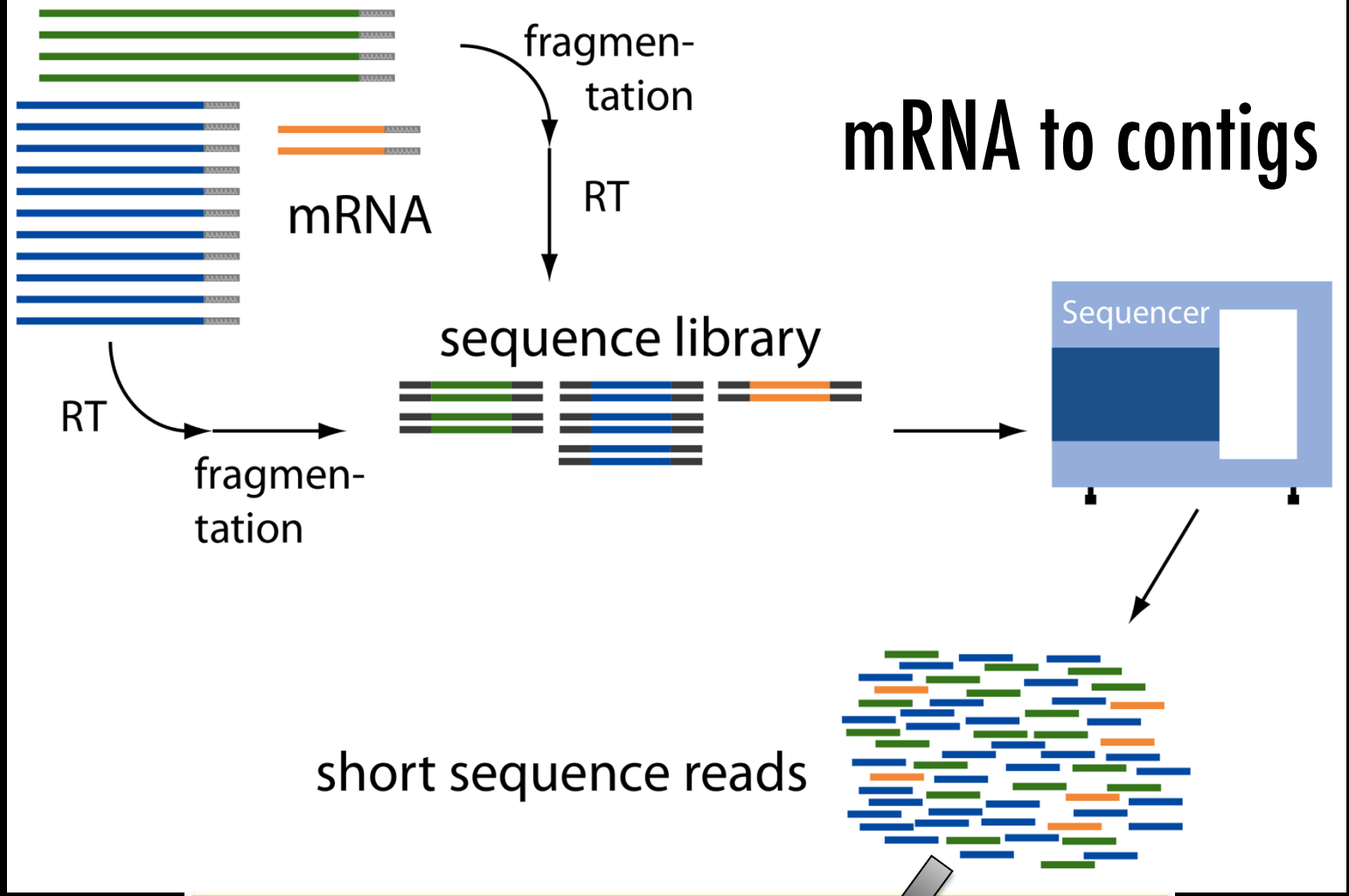
Duplication levels in RNA-Seq data

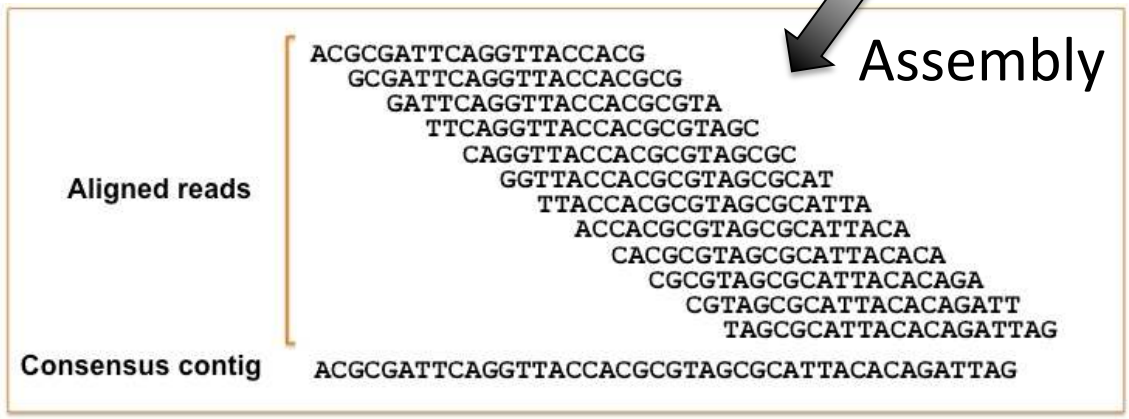
- Common in transcriptome work
- Starting with lots of high quality RNA increases
  - mRNA amount for sequencing
  - Decreases need of core facility to PCR your sample
- Moderate amounts of PCR duplication are OK
  - ~ 20% expected
  - > 50% perhaps problematic if correlated with experimental design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

Pipeline Overview

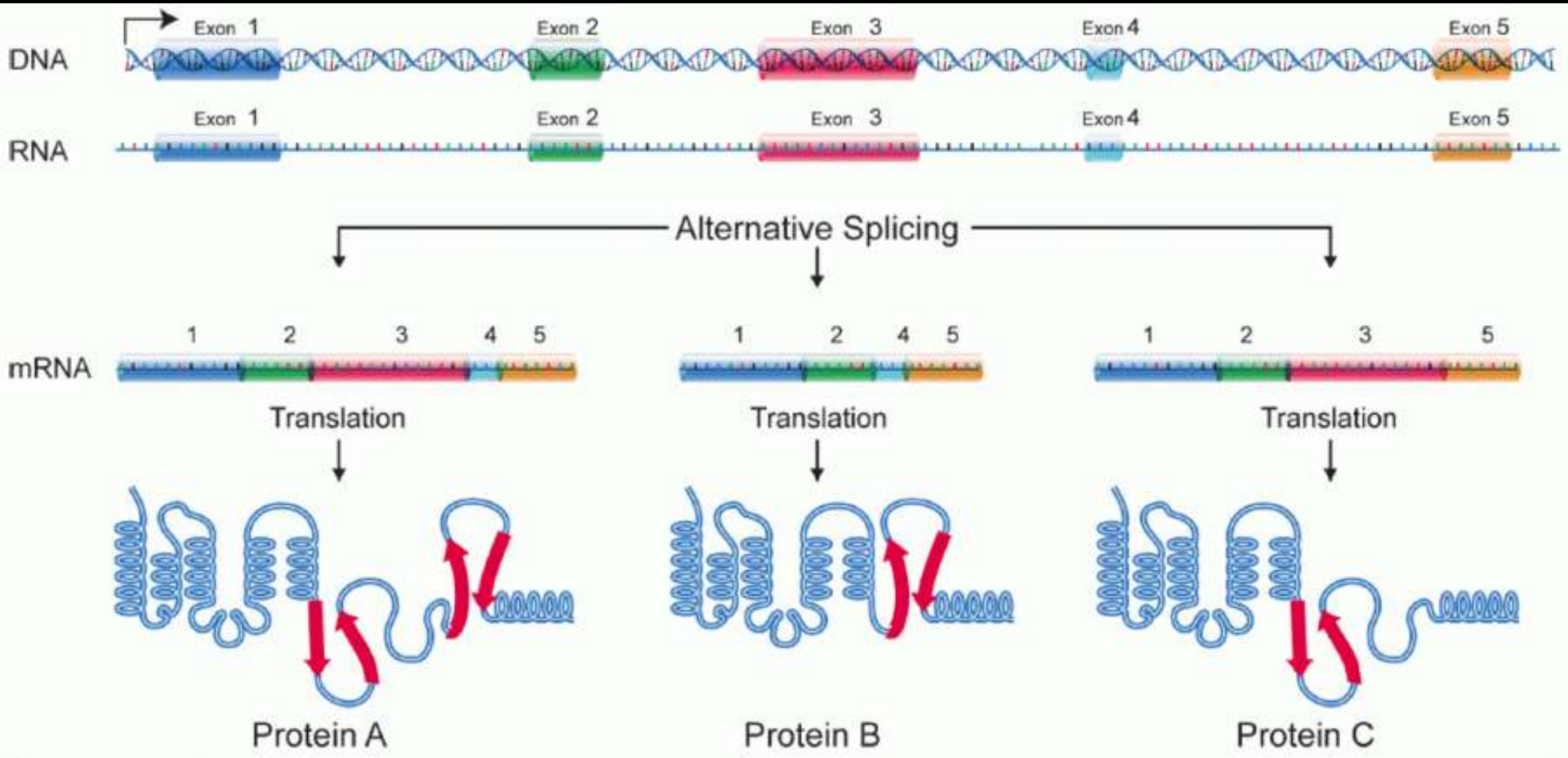


mRNA to contigs





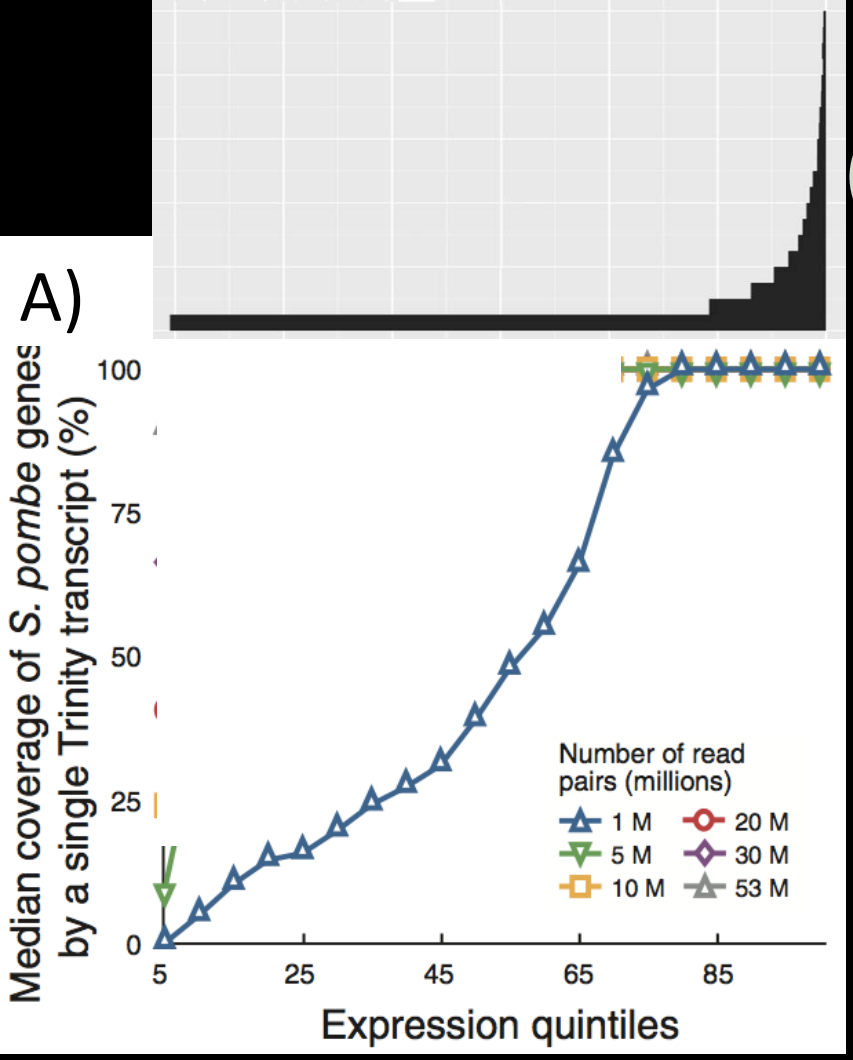
Alternative splicing complicates everything

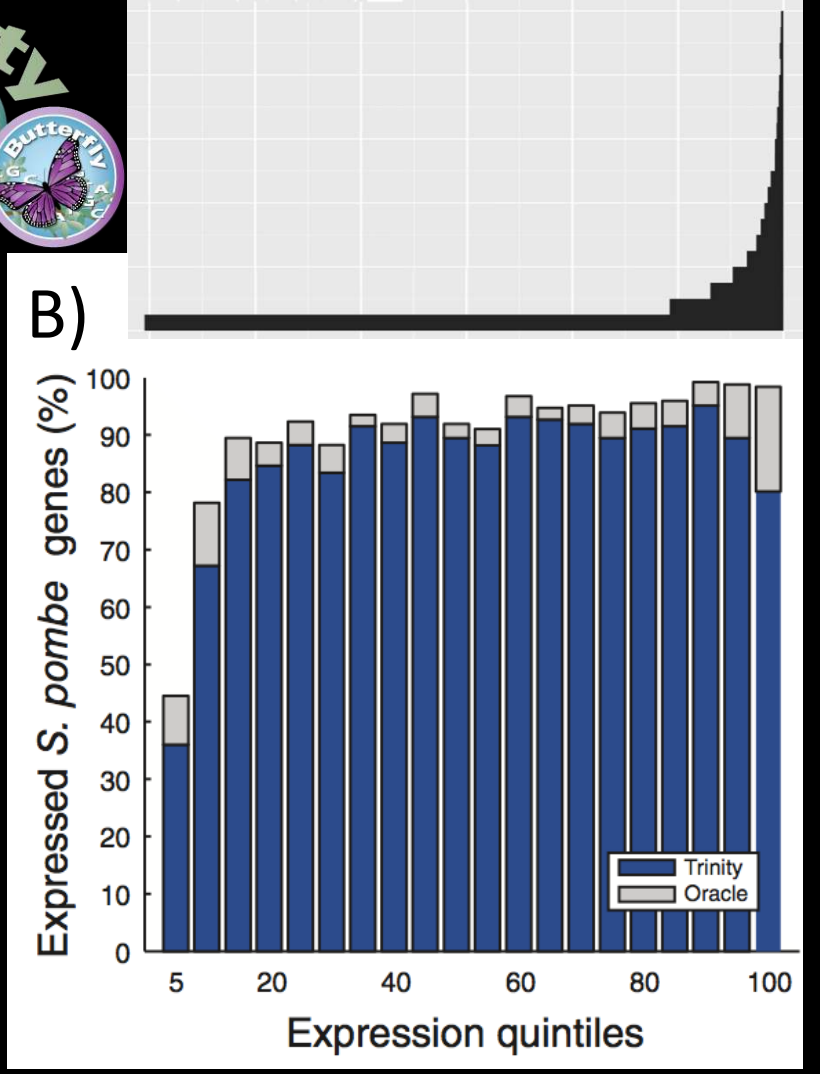


H. sapiens: > 95% of multi-exonic genes are spliced

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

BUT, when Trinity finishes ....

HOW do we know we did it right?

Assessing transcriptome assembly

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

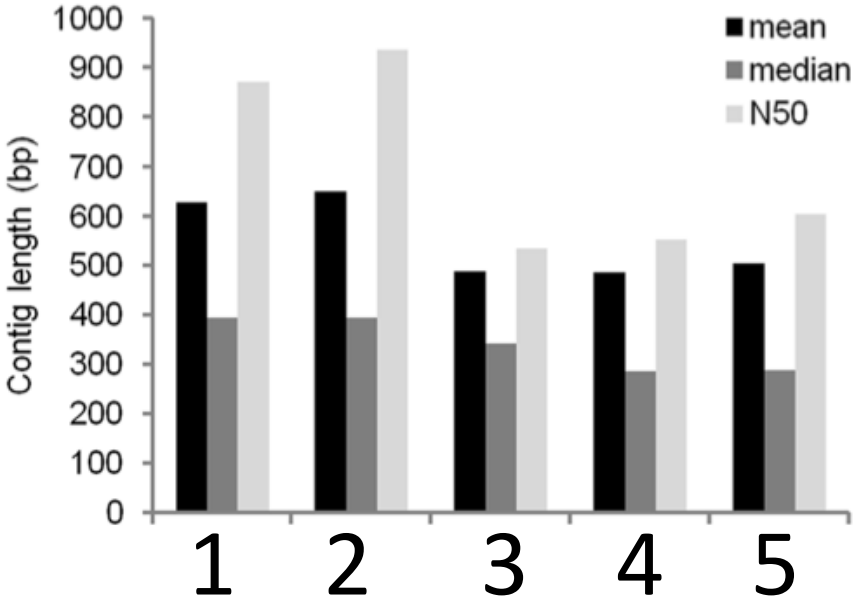
α / β :
1 = complete
< 1 = % covered

$$\alpha / \beta = \frac{\text{Length} = \alpha}{\text{Length} = \beta}$$

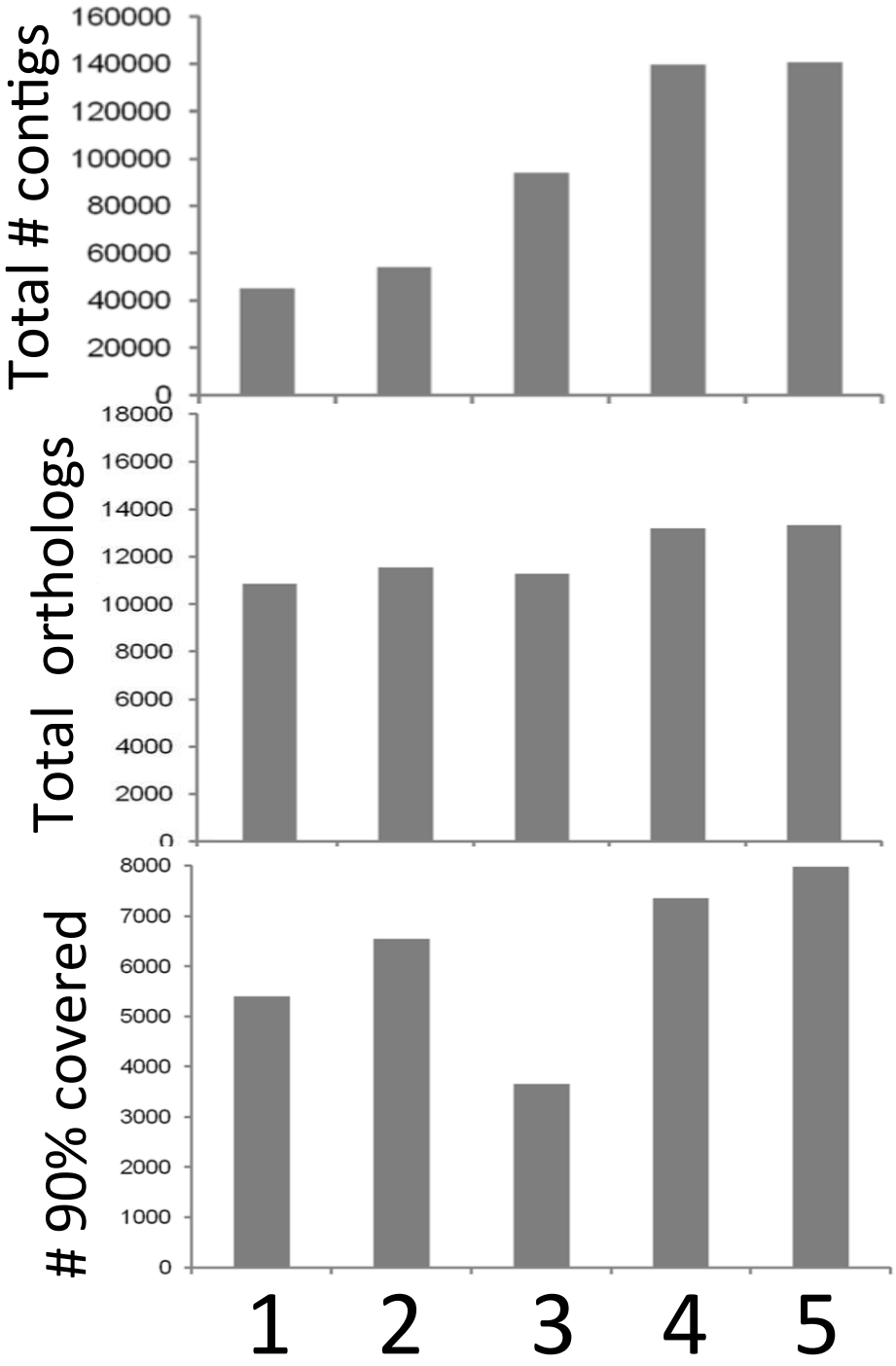
TA contig

Ortholog

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



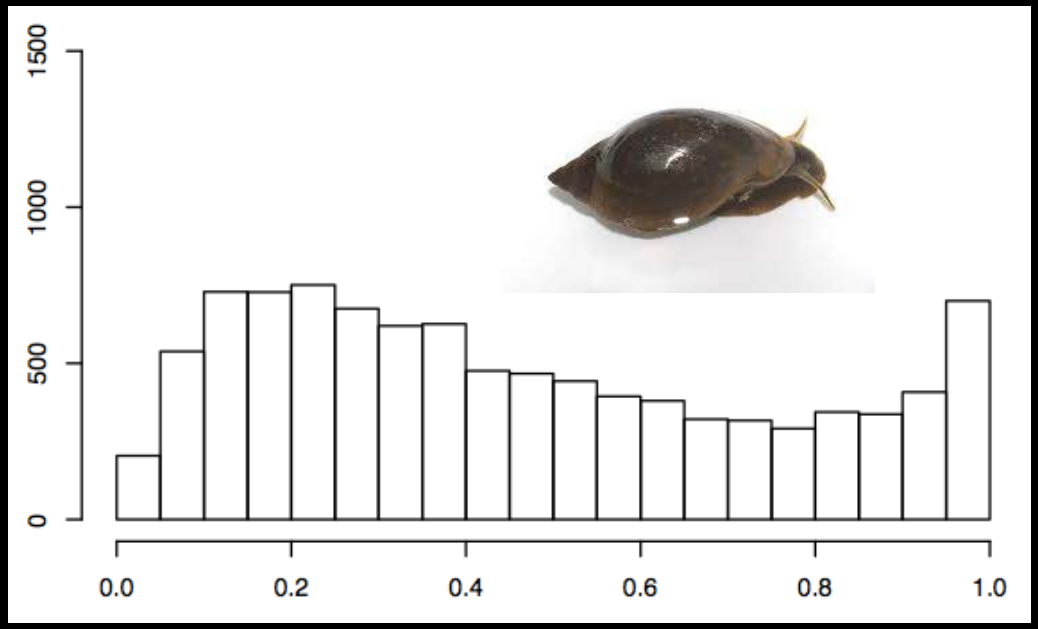
- 5 different TAs
- TA 2
  - Best N50, fewest contigs

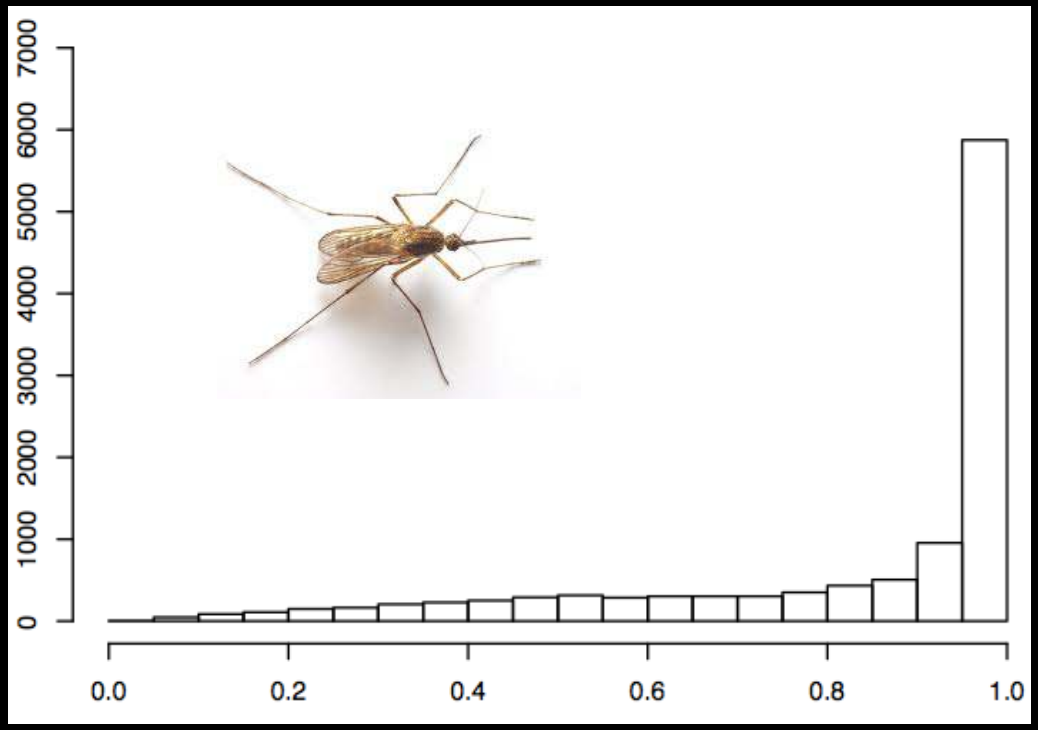


Hornett & Wheat 2012 BMC Genomics

OHR graphs

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

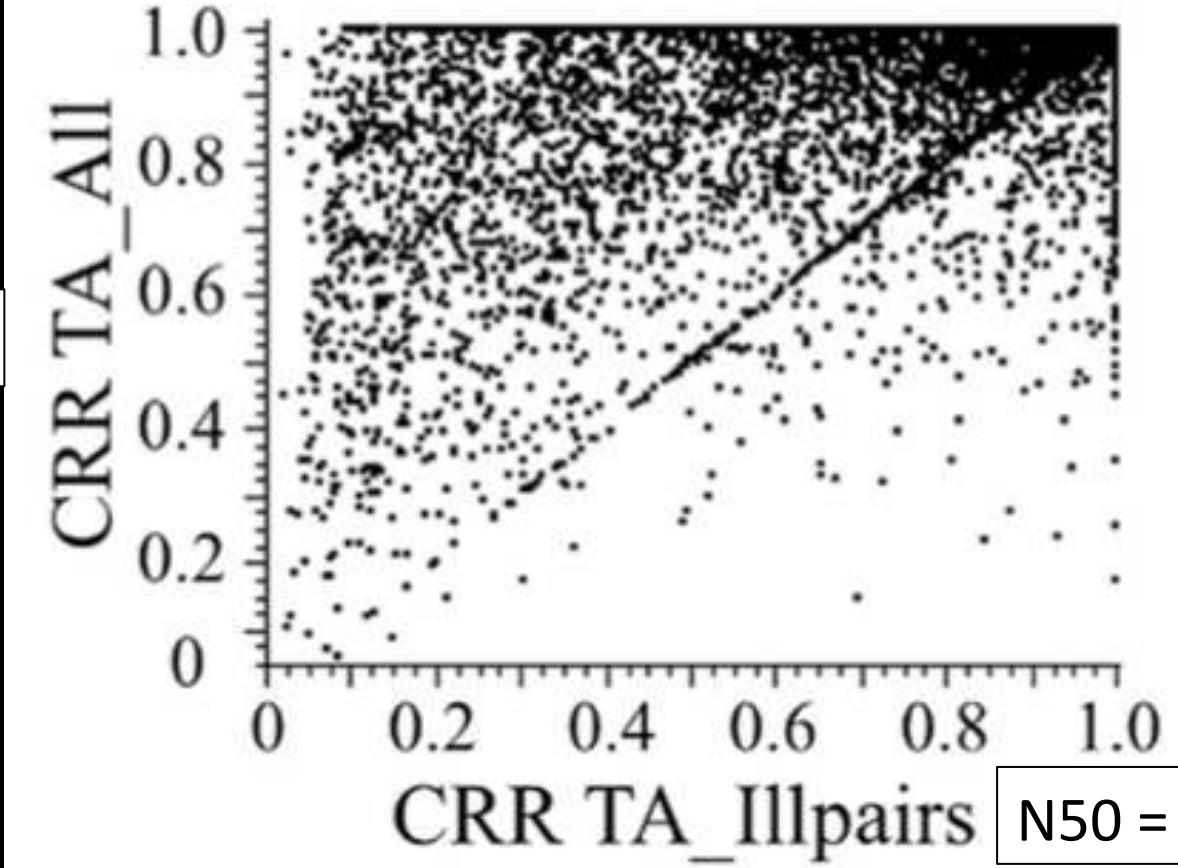




Comparative OHR

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

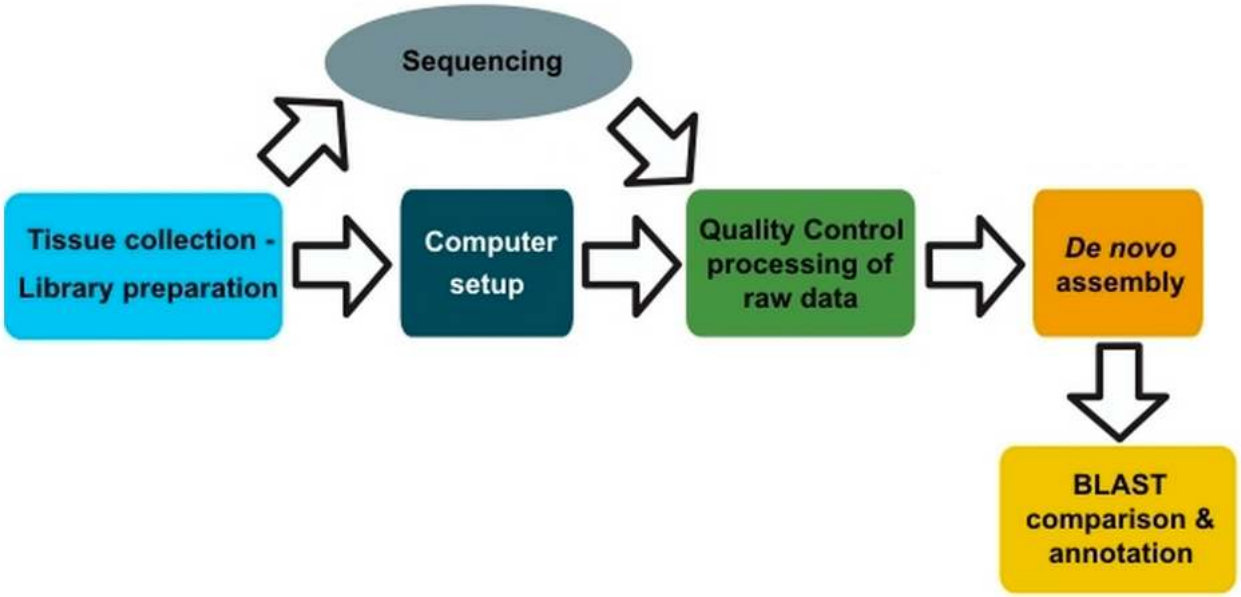
N50 = 610



N50 = 930

Hornett & Wheat 2012 BMC Genomics

Pipeline Overview



Blast: an alignment tool for identification

1) Input

- DNA or protein sequence
- Database of DNA or protein

2) Searches database

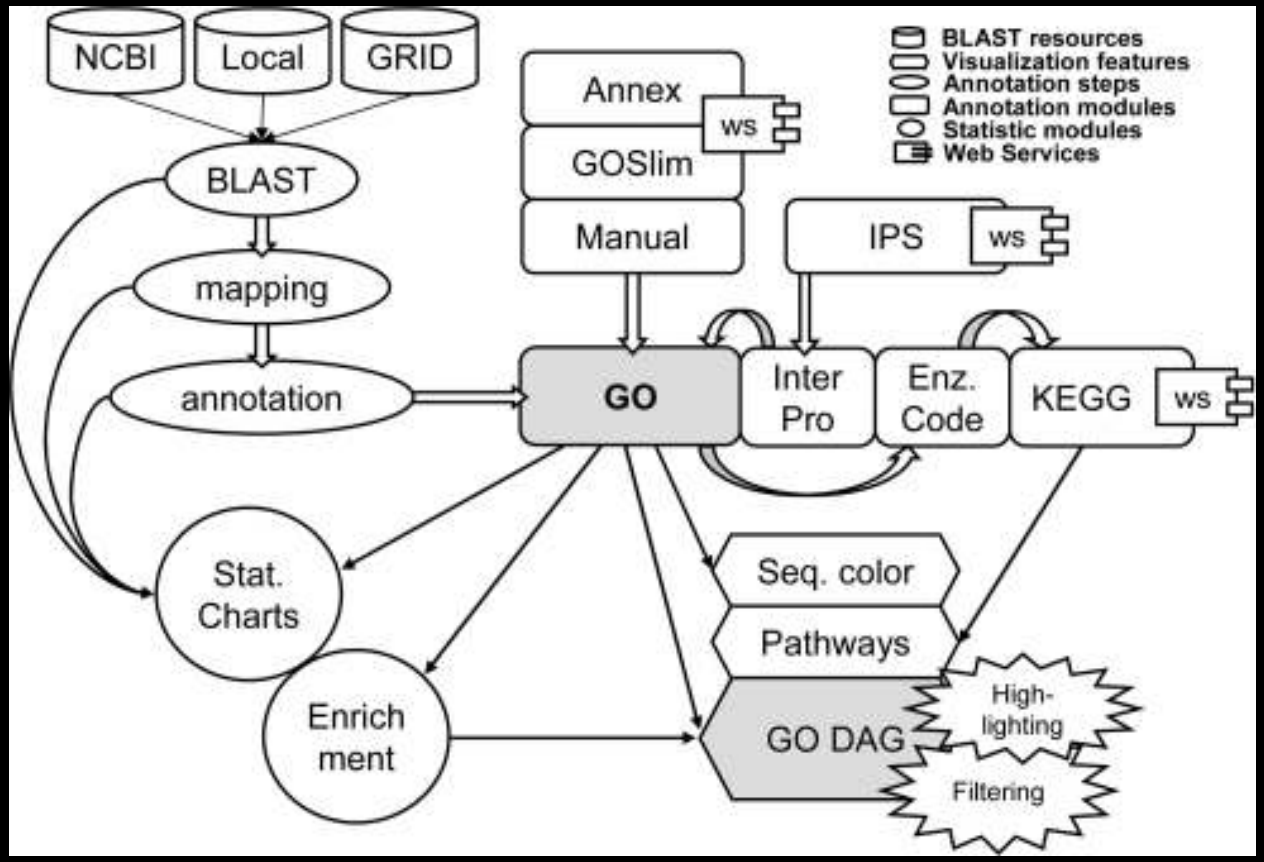
- Returns detailed matching information

Description	Max score	Total score	Query cover	E value	Ident	Accession
RecName: Full=Glucose-6-phosphate isomerase; Short=GPI; AltName: Full=Phosphoglucose isomerase; Short=PGI; AltName: Full=Phosphohexose isom	885	885	99%	0.0	76%	P52031.1
RecName: Full=Glucose-6-phosphate isomerase; Short=GPI; AltName: Full=Phosphoglucose isomerase; Short=PGI; AltName: Full=Phosphohexose isom	885	885	99%	0.0	76%	P52030.1
RecName: Full=Glucose-6-phosphate isomerase; Short=GPI; AltName: Full=Phosphoglucose isomerase; Short=PGI; AltName: Full=Phosphohexose isom	882	882	98%	0.0	76%	P52029.2
RecName: Full=Glucose-6-phosphate isomerase; Short=GPI; AltName: Full=Autocrine motility factor; Short=AMF; AltName: Full=Neuroleukin; Short=NLK;	839	839	98%	0.0	72%	P08059.3
RecName: Full=Glucose-6-phosphate isomerase; Short=GPI; AltName: Full=Autocrine motility factor; Short=AMF; AltName: Full=Neuroleukin; Short=NLK;	836	836	98%	0.0	71%	Q3ZBD7.4

```
Query  7    PKVNLKQDPAYQKLQEYYNNNADKINILQLFQQDADRFIKYSLRIPTPNDGEILLDYSKN  186
            P   L Q+ A+QKLQEYY+++    +NI  LF +DA RF KYSLR+ T NDGEILLDYSKN
Sbjct  4    PLPPLNQEAAFQKLQEYYDSSGKDLNIKDLFVKDAKRFSKYSLRLHTQNDGEILLDYSKN  63

Query  187  RIDDTTFSLLLNLAKSRNVEKARDAMFAGEKINFTEDRAVLHVALRNRQNRPIMVNGKDV  366
            RI+D  + LLL LAK R V+ ARDAMF+G+ IN TE+RAVLH ALRNR   P++V+ KDV
Sbjct  64   RINDEVWDLLLALAKVRRVDAARDAMFSGQHINITENRAVLHTALRNRGTDPVLVDDKDV  123

Query  367  TPDVNGVLAHMKEFSTQVISGAWKGYTGKPITDVINIGIGGSDLGPLMVTEALKPYANHL  546
              PDV   LAHMKEF+  VISG W+G TGK ITDV+NIGIGGSDLGPLMVTEALKPY   L
Sbjct  124  MPDVRAELAHMKEFTNMVISGVWRGCTGKQITDVVNIGIGGSDLGPLMVTEALKPYGKGL  183

Query  547  KVHFVSNIDGTHLAEVLKRLNPETALFIIASKTFTTQETITNATSAKTWFLEAAKDPAAV  726
              HFVSNIDGTHLAEVLK++N ET LFI+ASKTFTTQETITNATSAKTW LE +K+P +V
Sbjct  184  HSHFVSNIDGTHLAEVLKKVNYETTLFIVASKTFTTQETITNATSAKTWLLEHSKEPESV  243

Query  727  SKHFVALSTNGEKVTAFGIDPKNMFGFWDWVGGRYSLWSAIGLSISLYIGFENFEKLLDG  906
            +KHFVALSTN EKVT FGID  NMFGFWDWVGGRYSLWSAIGLSI L IGFENFE+LLDG
Sbjct  244  AKHFVALSTNKEKVTEFGIDSTNMFGFWDWVGGRYSLWSAIGLSICLSIGFENFEQLLDG  303
```

Annotation

- TA contigs
  - DNA
  - Search database of known proteins
- BlastX
  - Makes 6 frame translation of DNA into protein
  - Searches DB 6 times



blast2go

Yandell M, Ence D: **A beginner's guide to eukaryotic genome annotation.** *Nat Rev Genet* 2012, **13**:329–342.

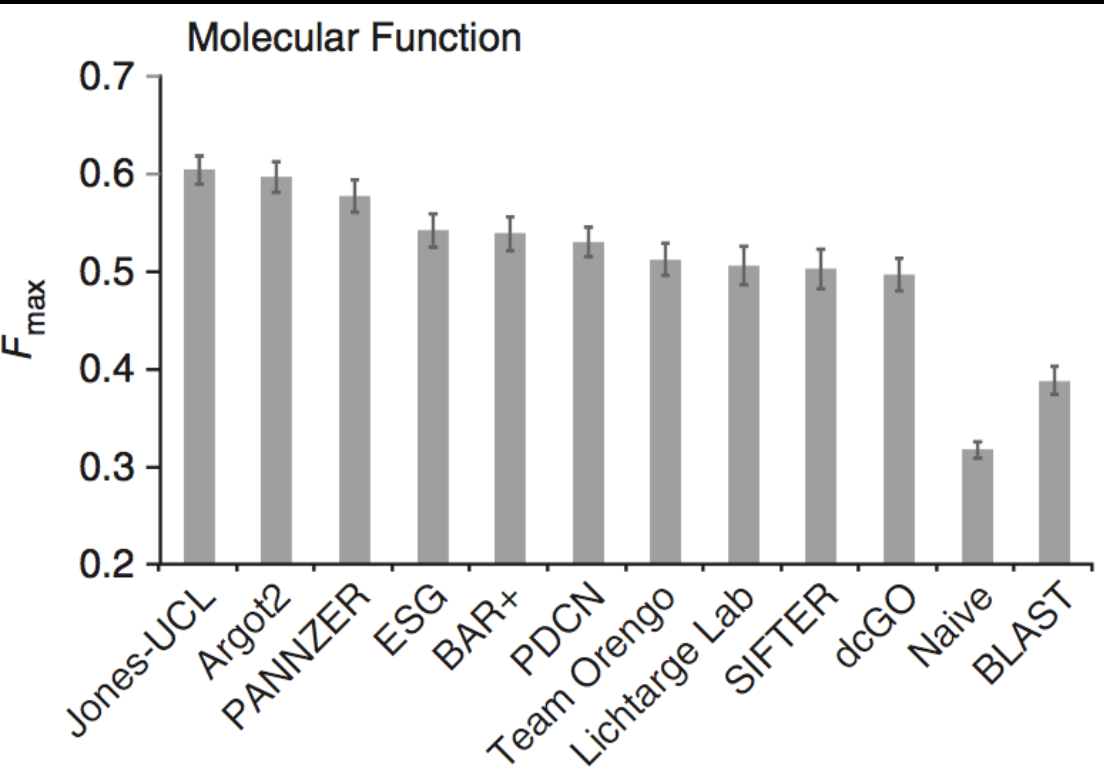
Gene Ontology: order in the chaos

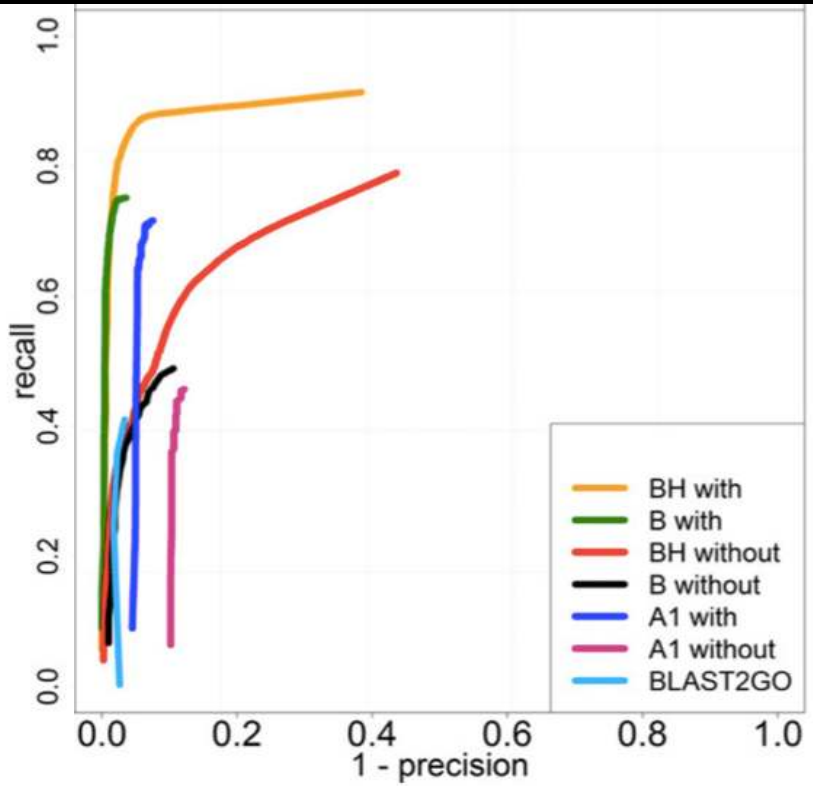
- addresses the need for consistent descriptions of gene products in different databases in a species-independent manner

- GO project has developed three structured controlled vocabularies (ontologies) that describe gene products in terms of their associated
  - biological processes
  - cellular components
  - molecular functions

http://www.geneontology.org/

Comparisons among annotation tools

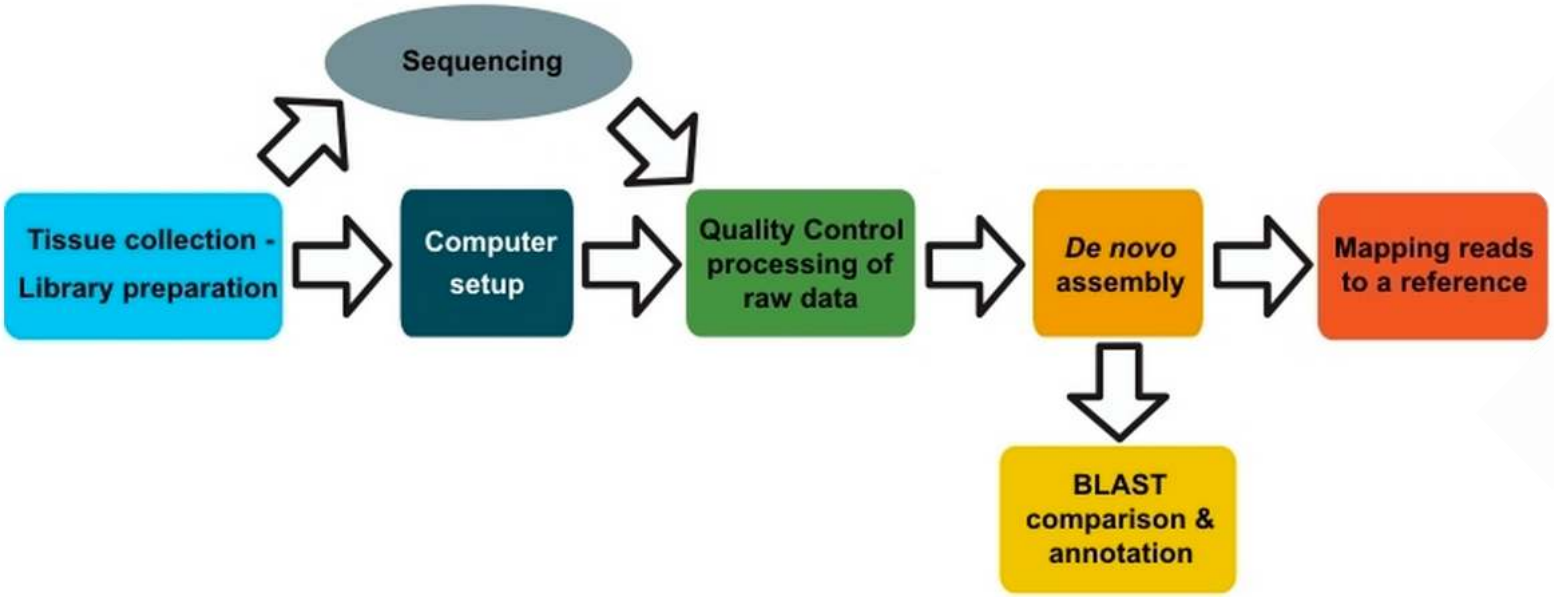


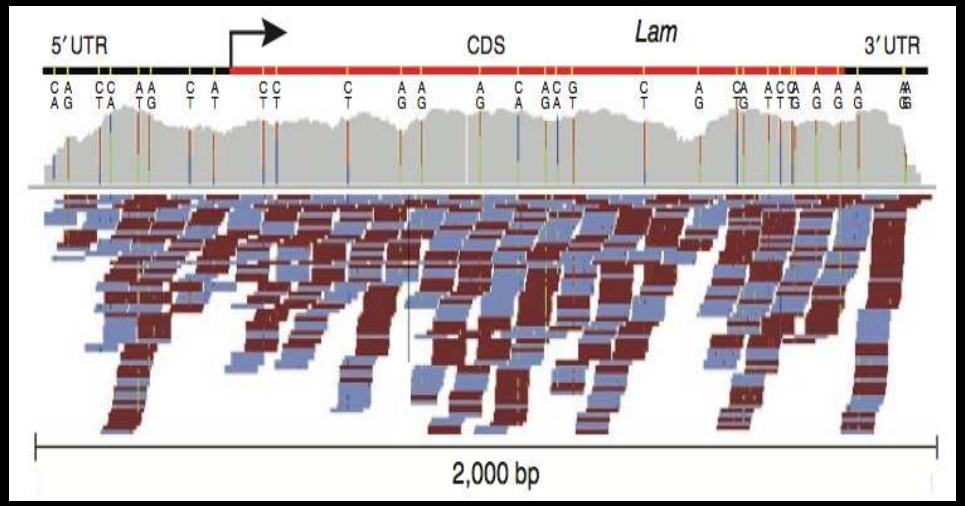


Radivojac et al.: **A large-scale evaluation of computational protein function prediction**. *Nat Meth* 2013, **10**:221–227.

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

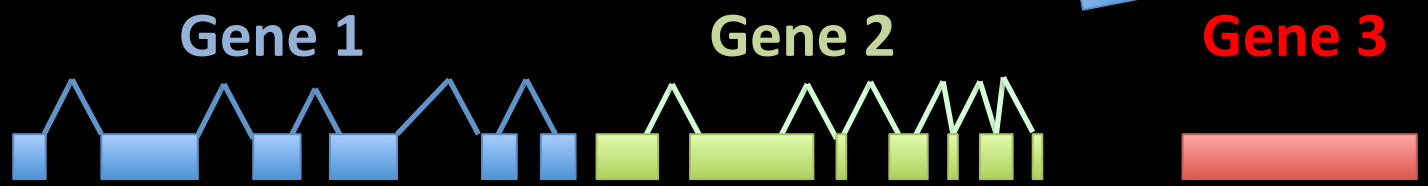
Pipeline Overview

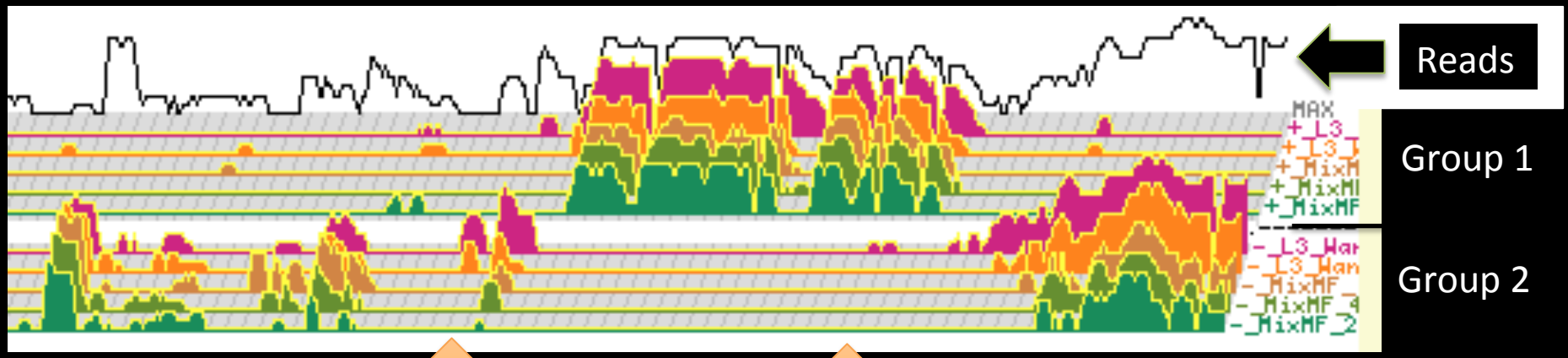




Whole gene level

Exon level





Alternative splicing

Expression difference

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)

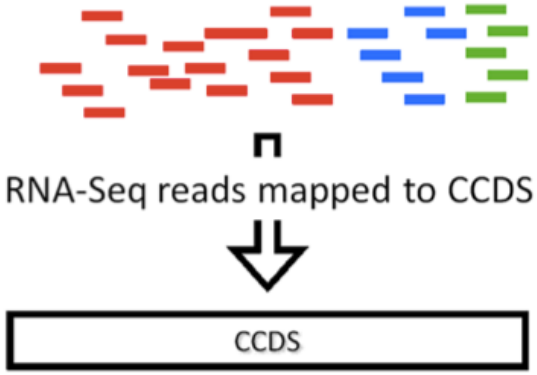
Map to TA vs. Genome:

which is better?

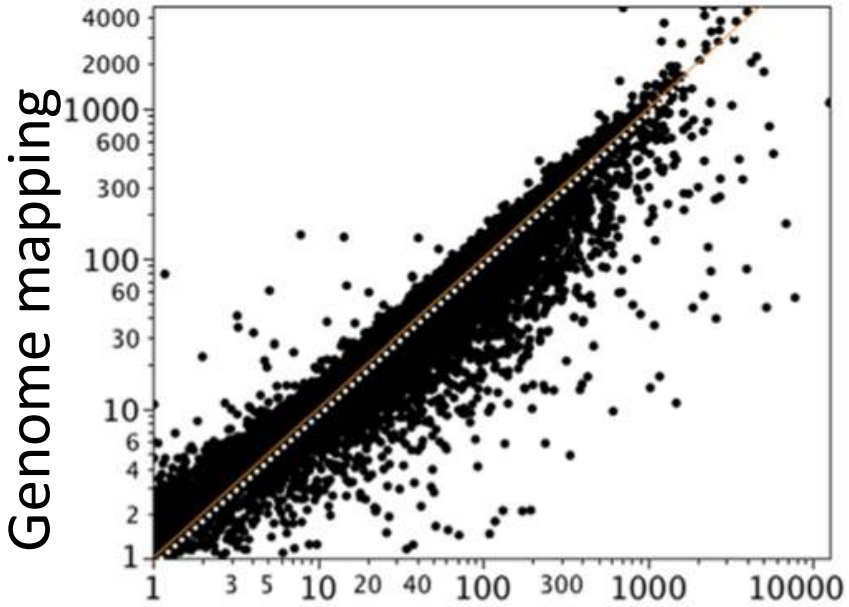
Template effects:

- 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

Genome mapping



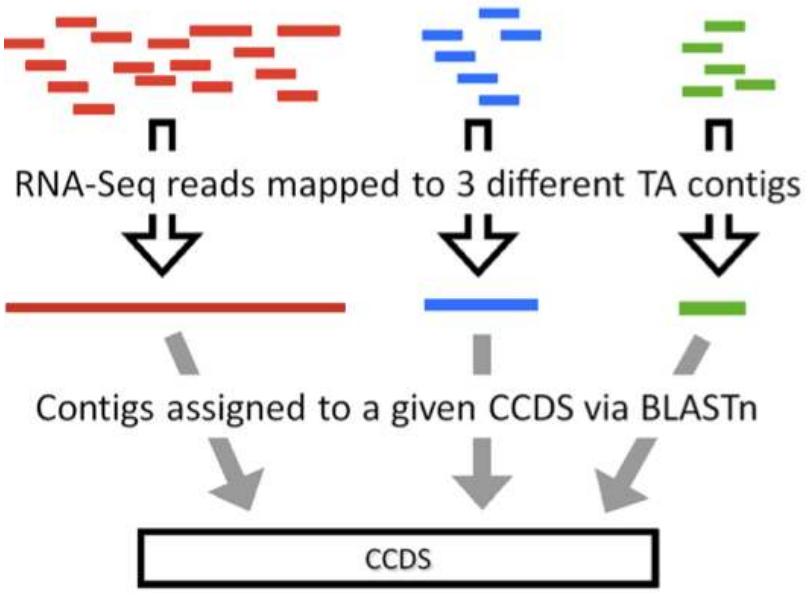
Spearman's ρ = 0.95, P < 0.0001



RNA-Seq mapping: comparing genome vs. TA

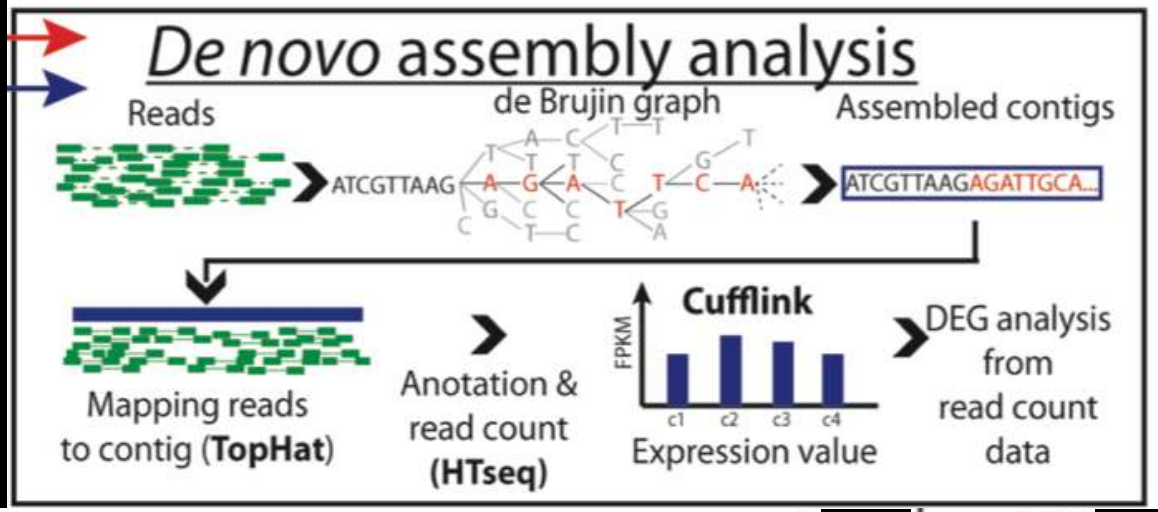
You can generate high quality data without a genome, for much of the transcriptome

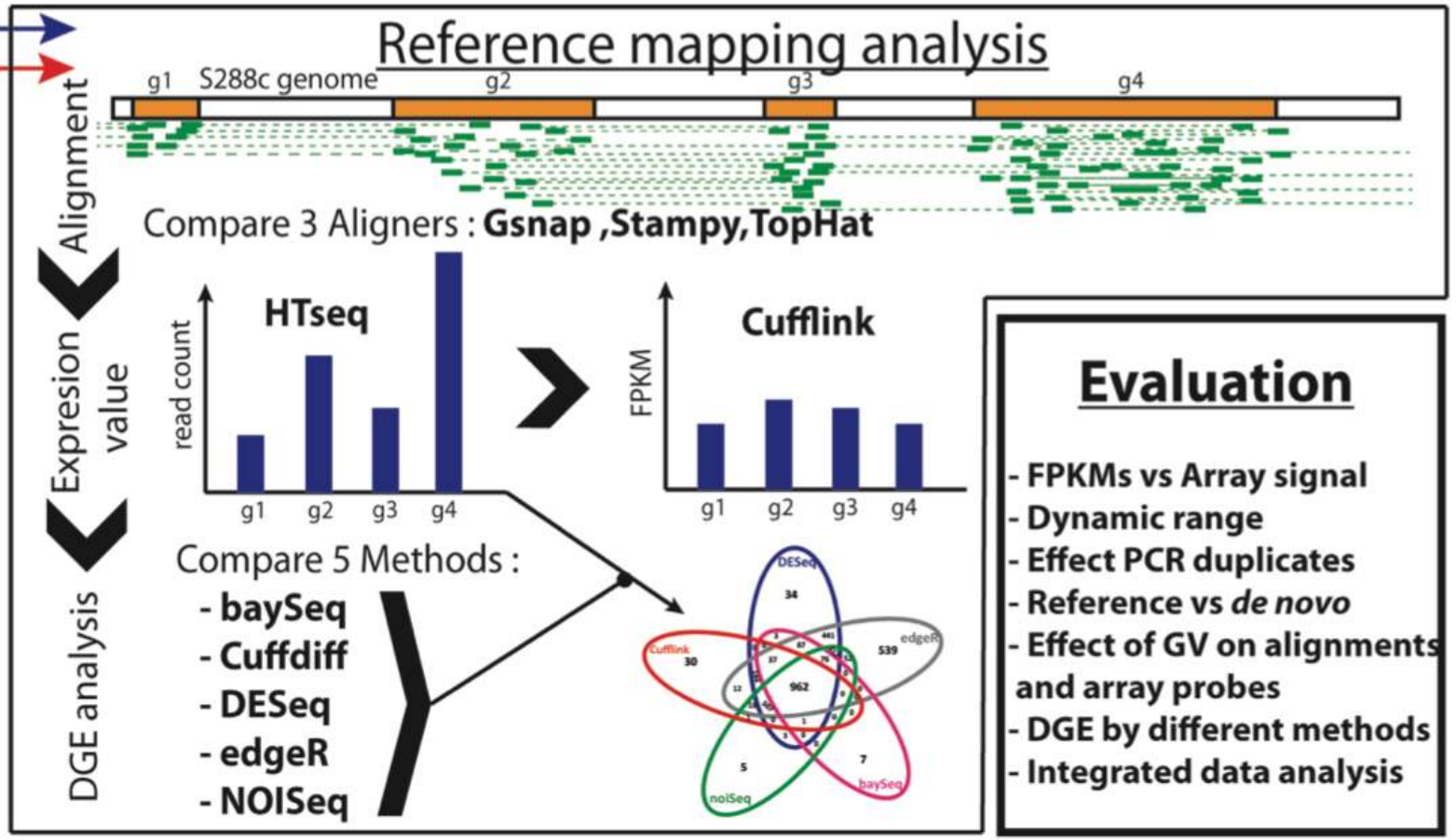
Summed TA mapping

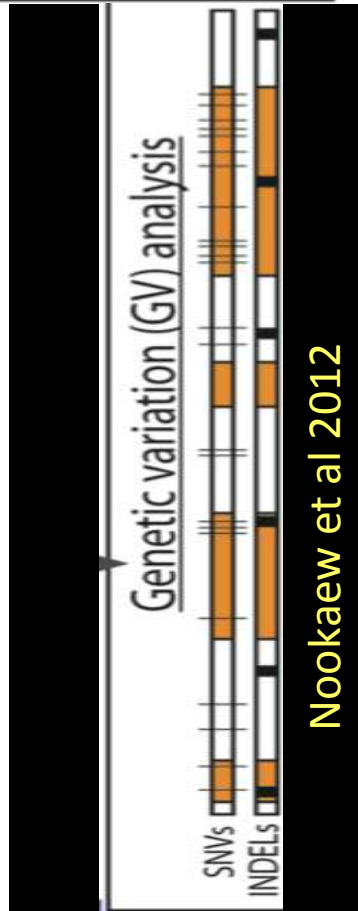


Hornett & Wheat 2012 BMC Genomics

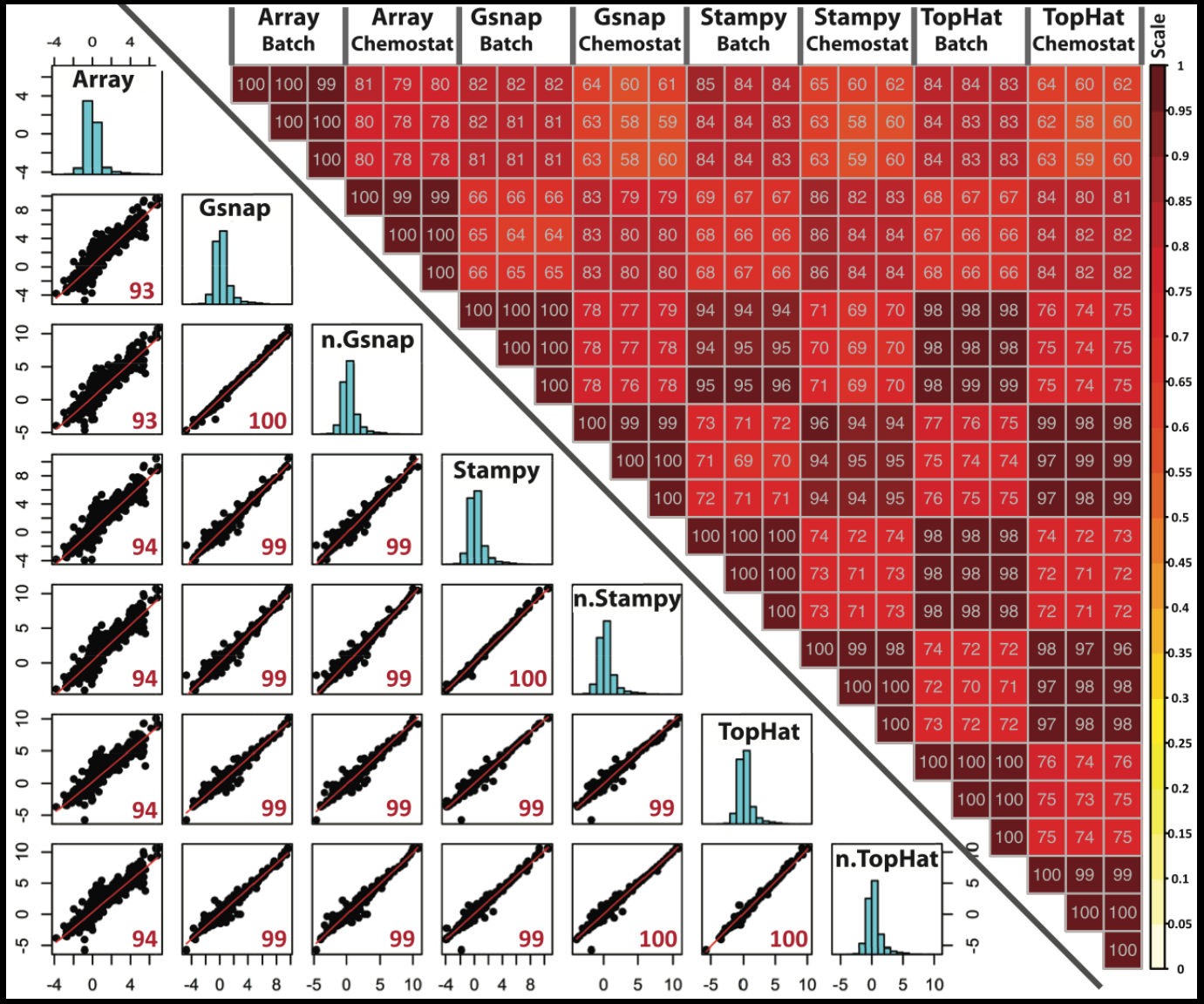
Template mismatch effects: excellent yeast study







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.

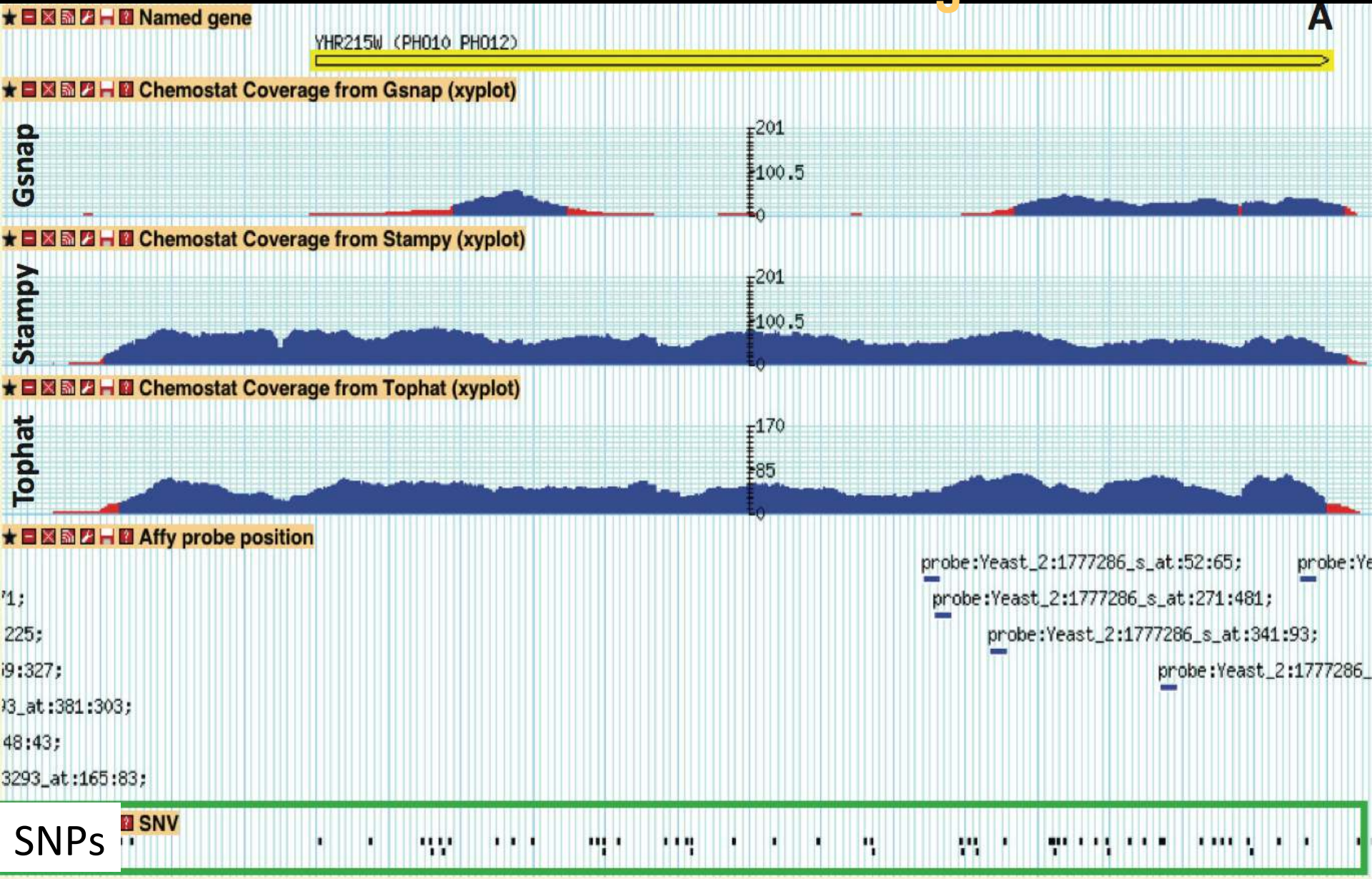
Mappers don't appear to matter

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

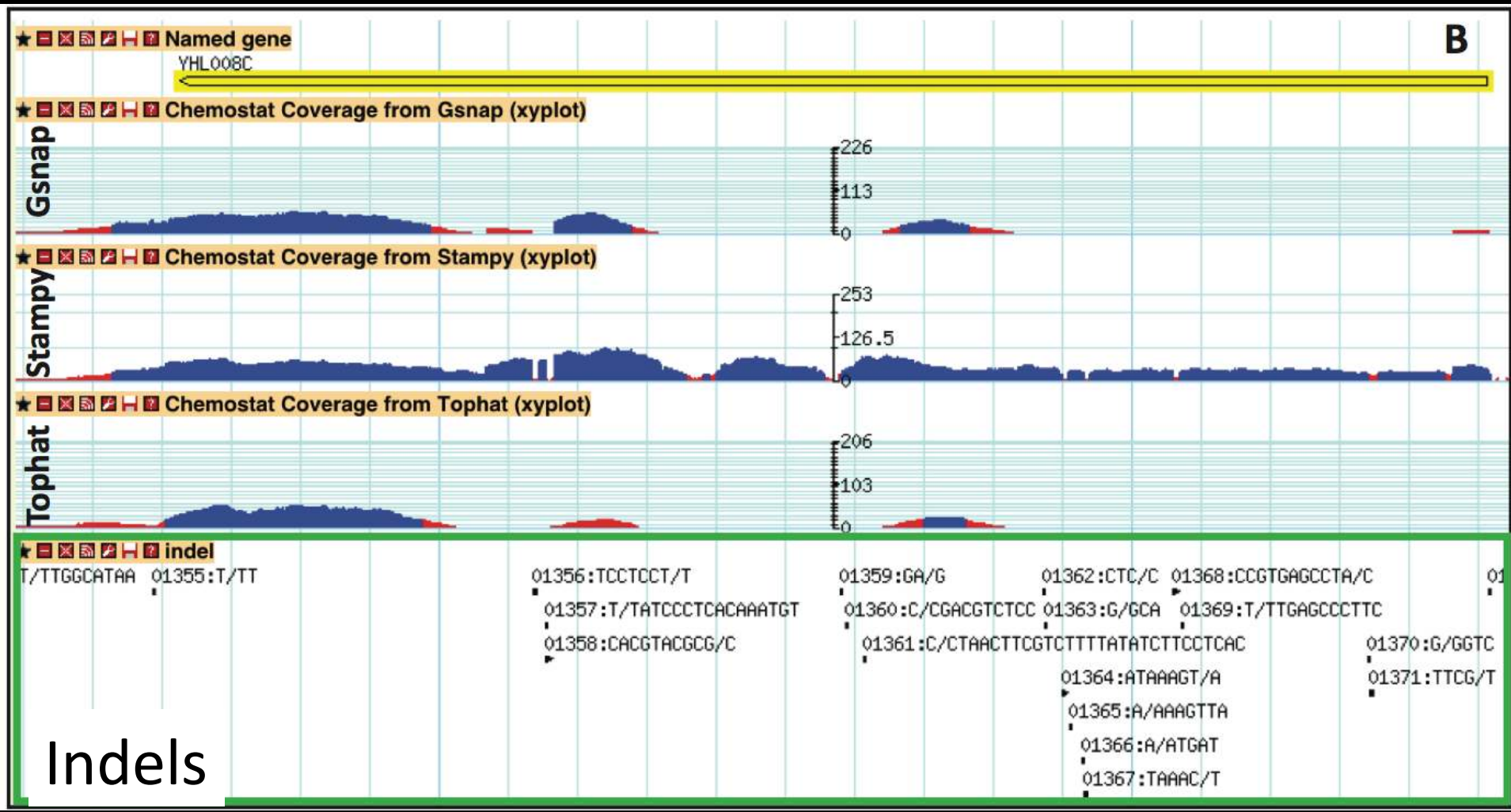
SNP effects can be large



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



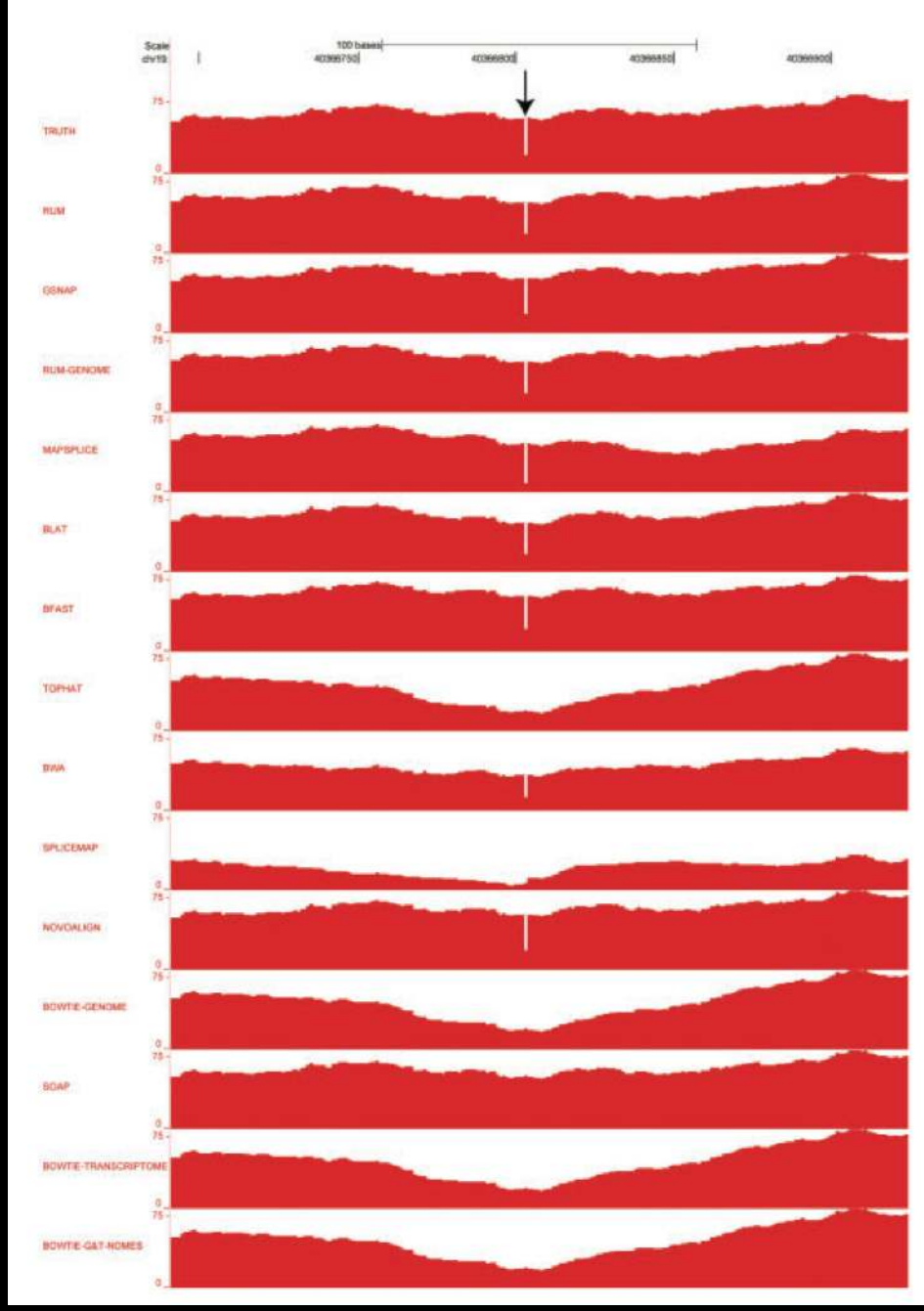
Nookaew et al. **A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae**. *Nucleic Acids Research* 2012, **40**:10084–10097.

15 mapping results

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

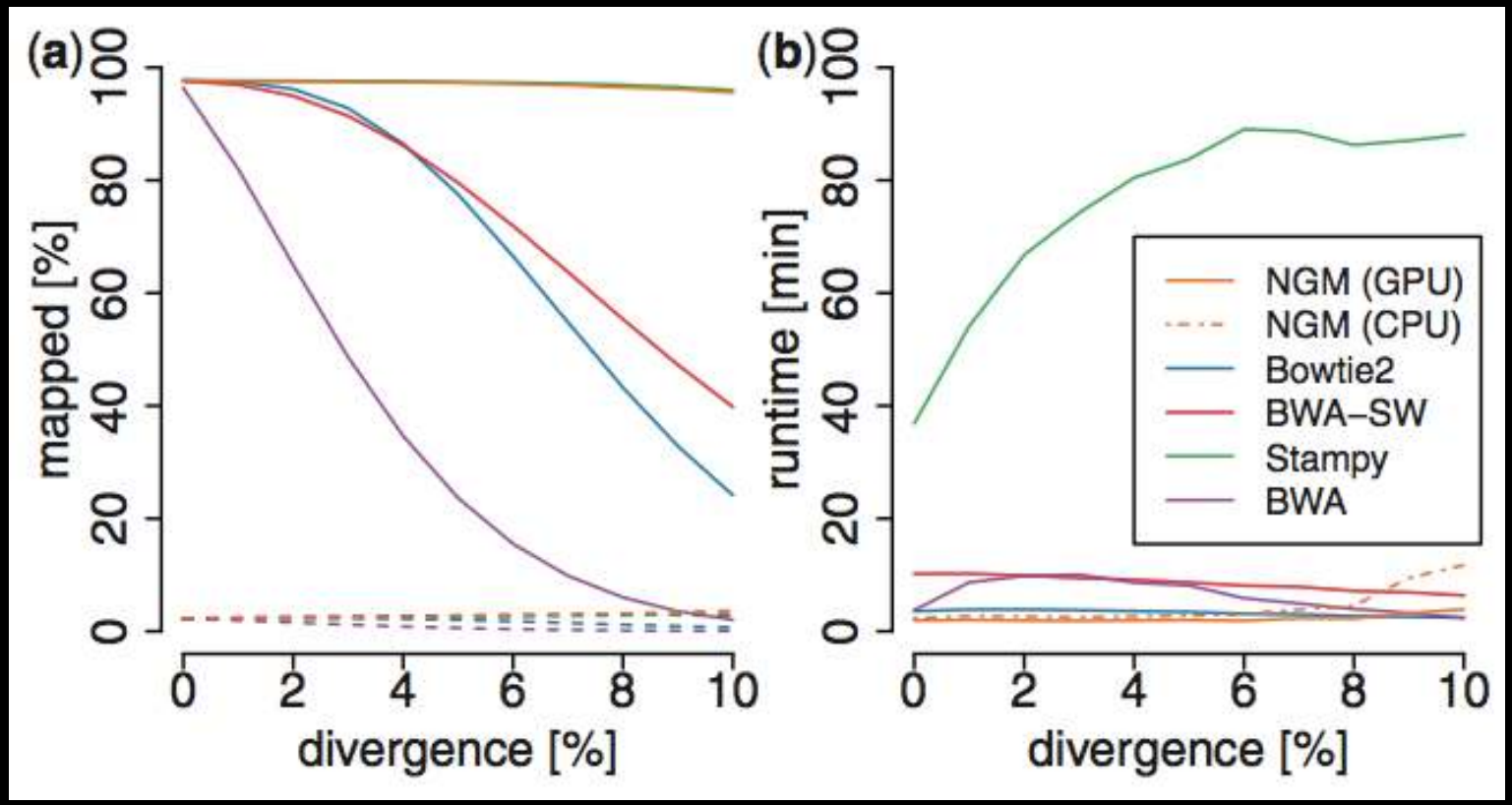
TopHat, SpliceMap, Bowtie and Soap

- 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.

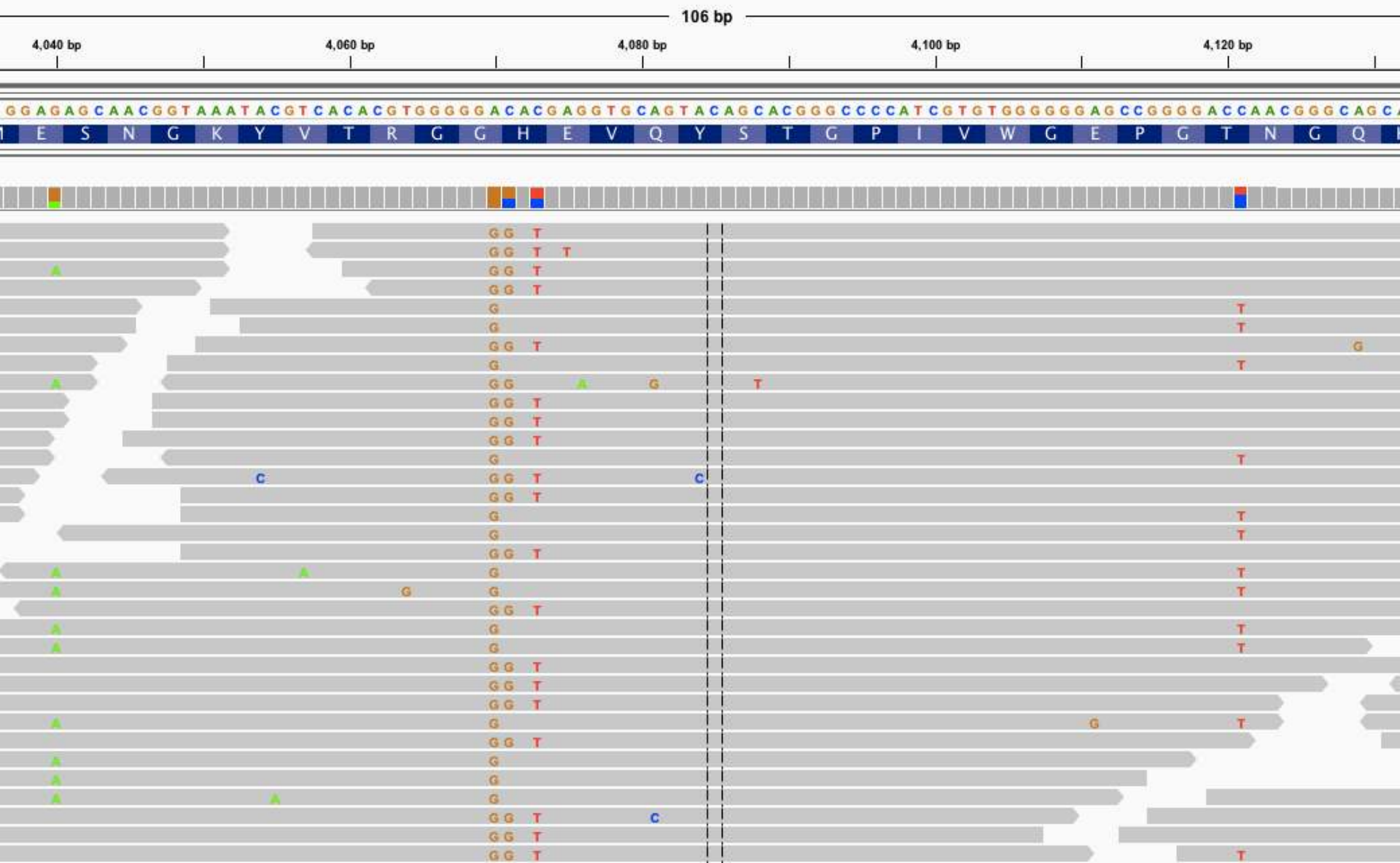
Allelic bias in read mapping



- 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?

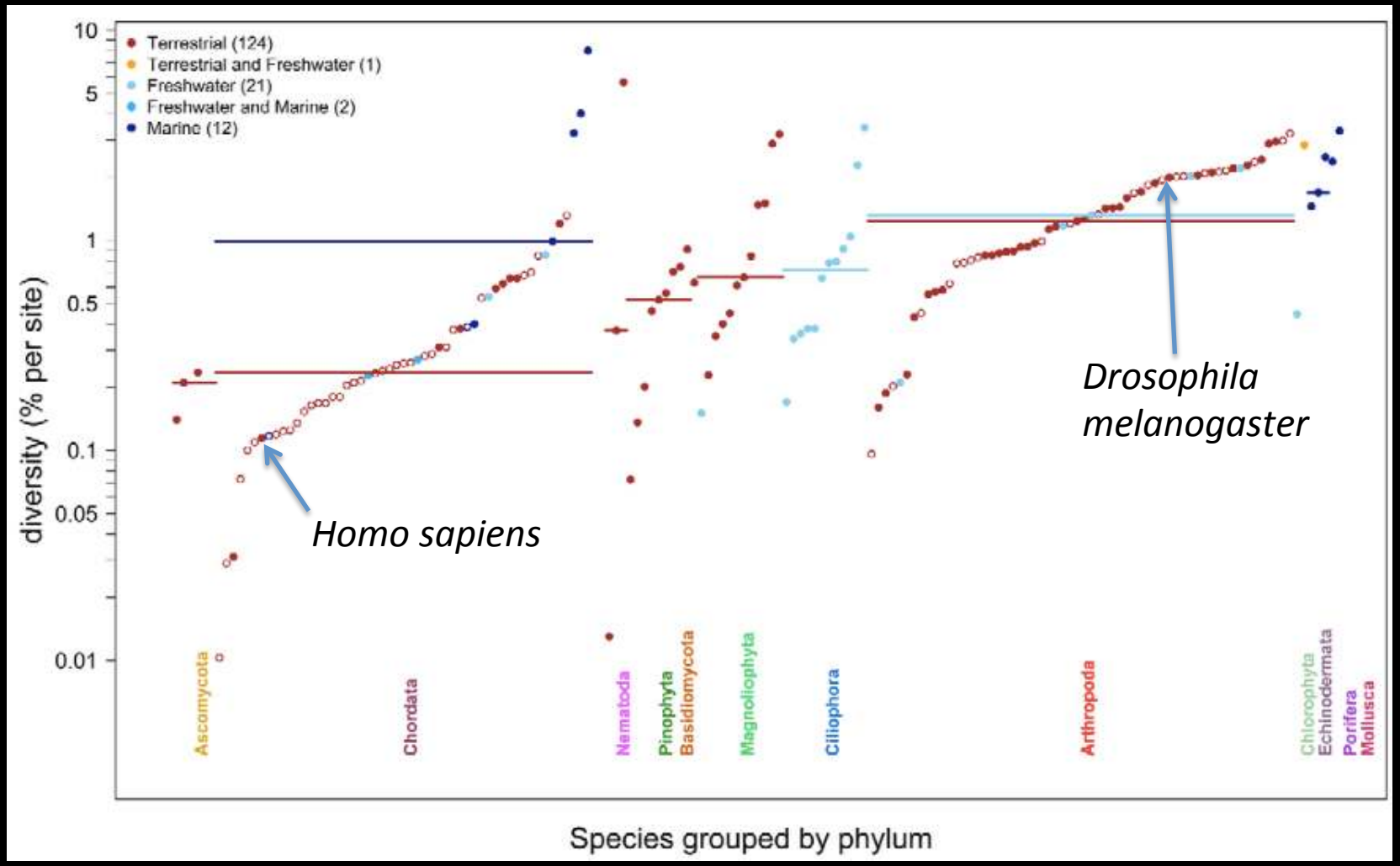
Sedlazeck et al. 2013 *Bioinformatics*

100 bp window with 4 – 5 SNPs differing from reference



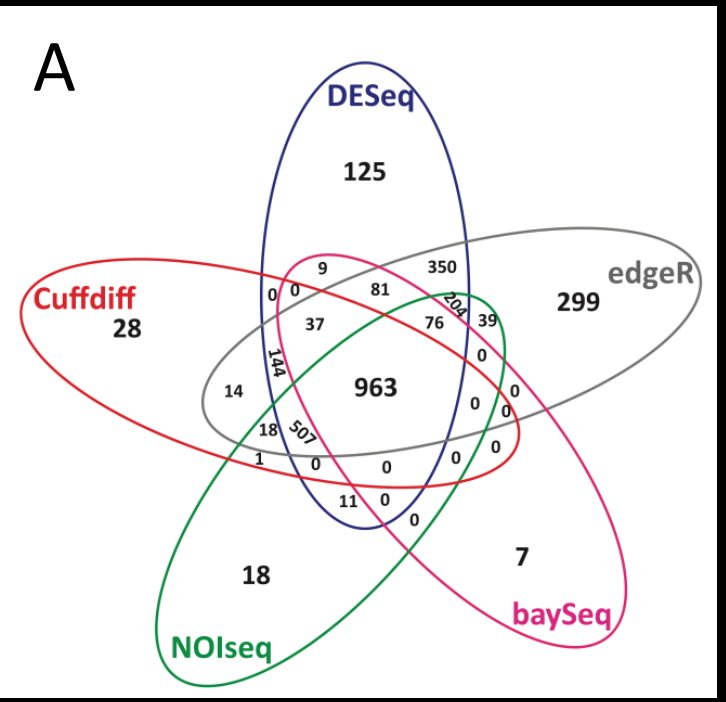
Mapping reads in outbred species

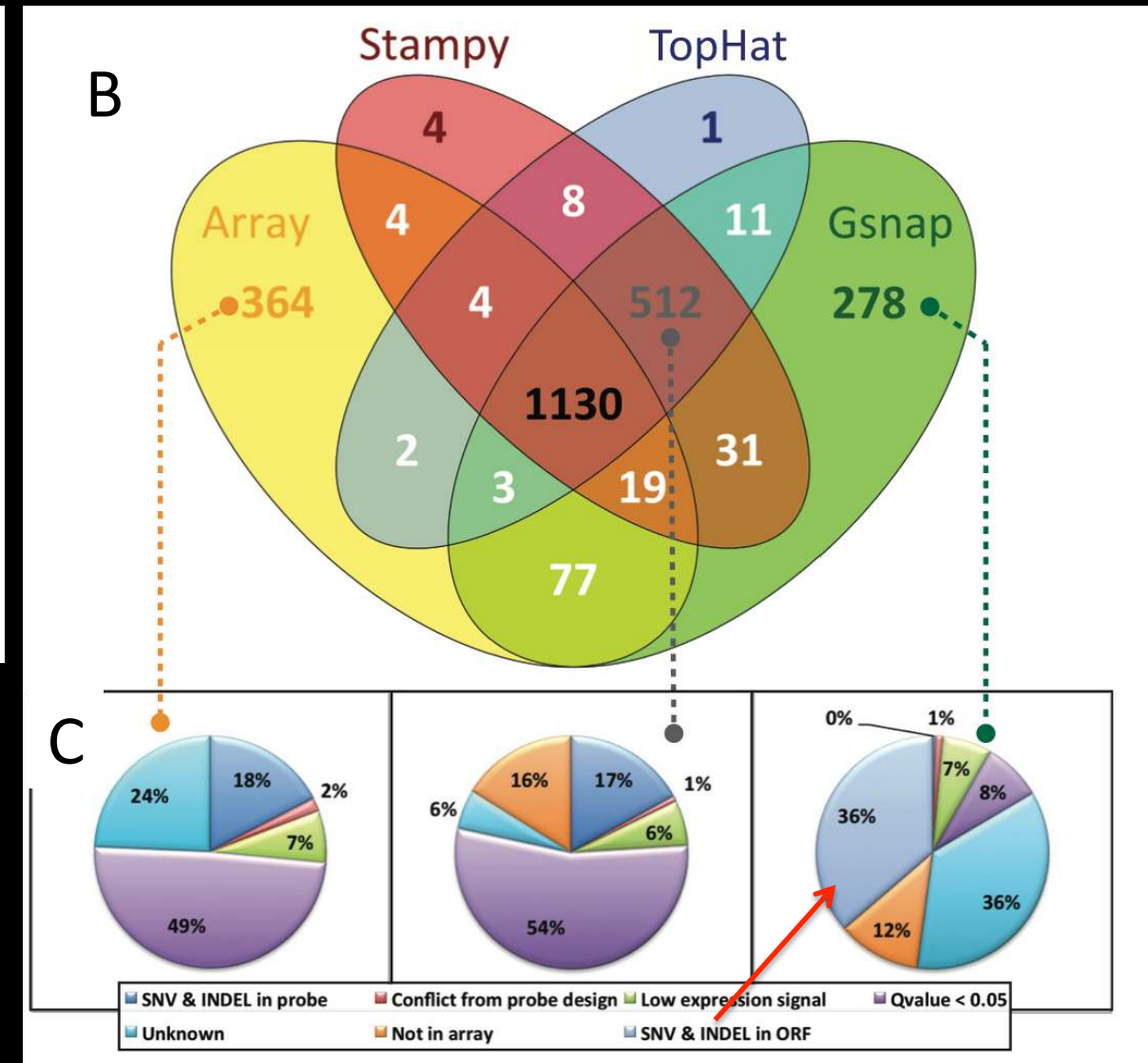
Average genome polymorphism levels (ignores indels)



Leffler *et al.* 2012 *Plos Biol*

Sig. expression differences by method





A: Stampy mapping

B: Cuffdiff analysis

C: Likely error source

RNA-Seq

Real world example

2 factor analysis with family effects

Bicyclus anynana

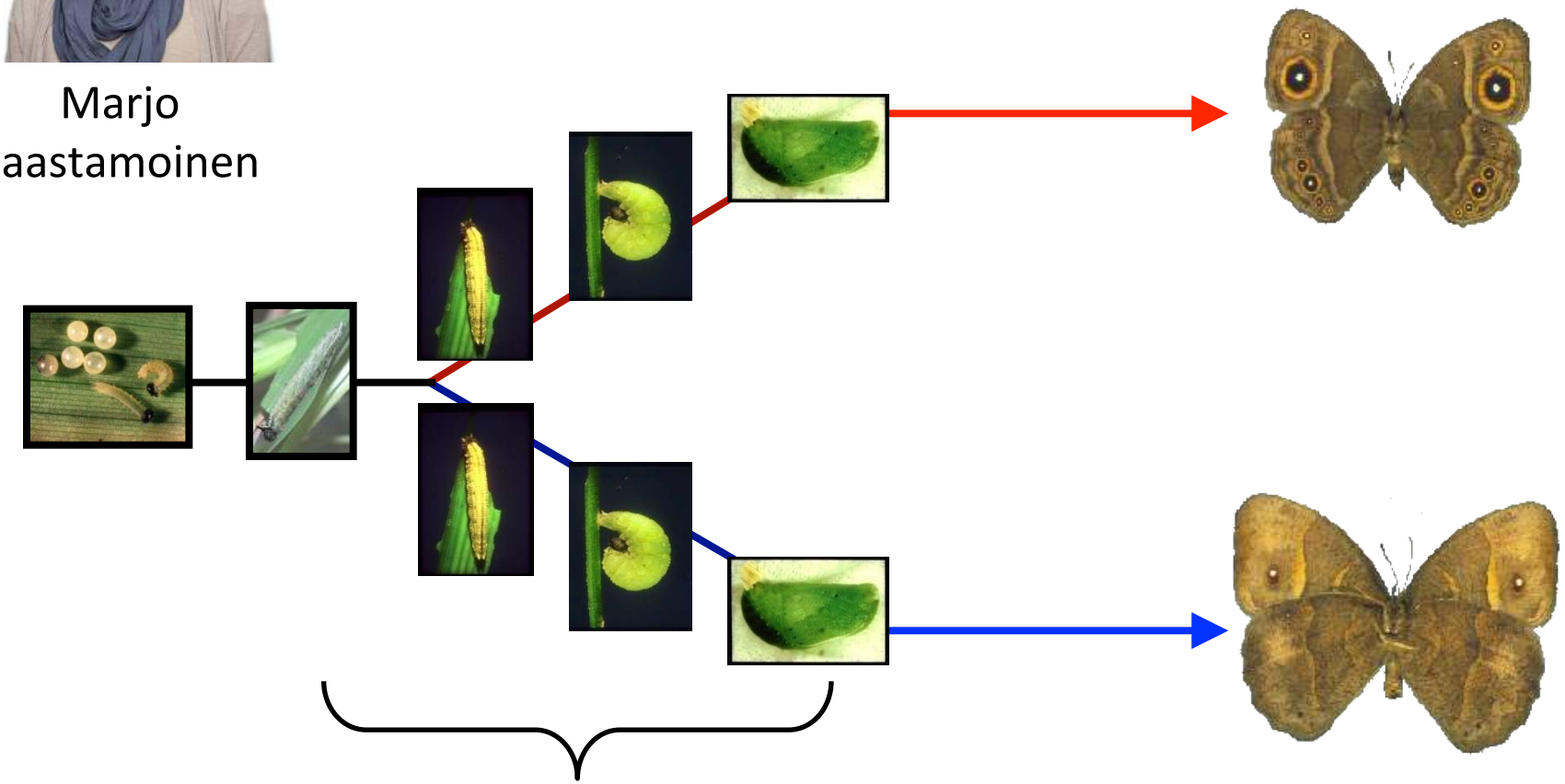
Save energy, live long

Live fast, die young

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

Bicyclus anynana

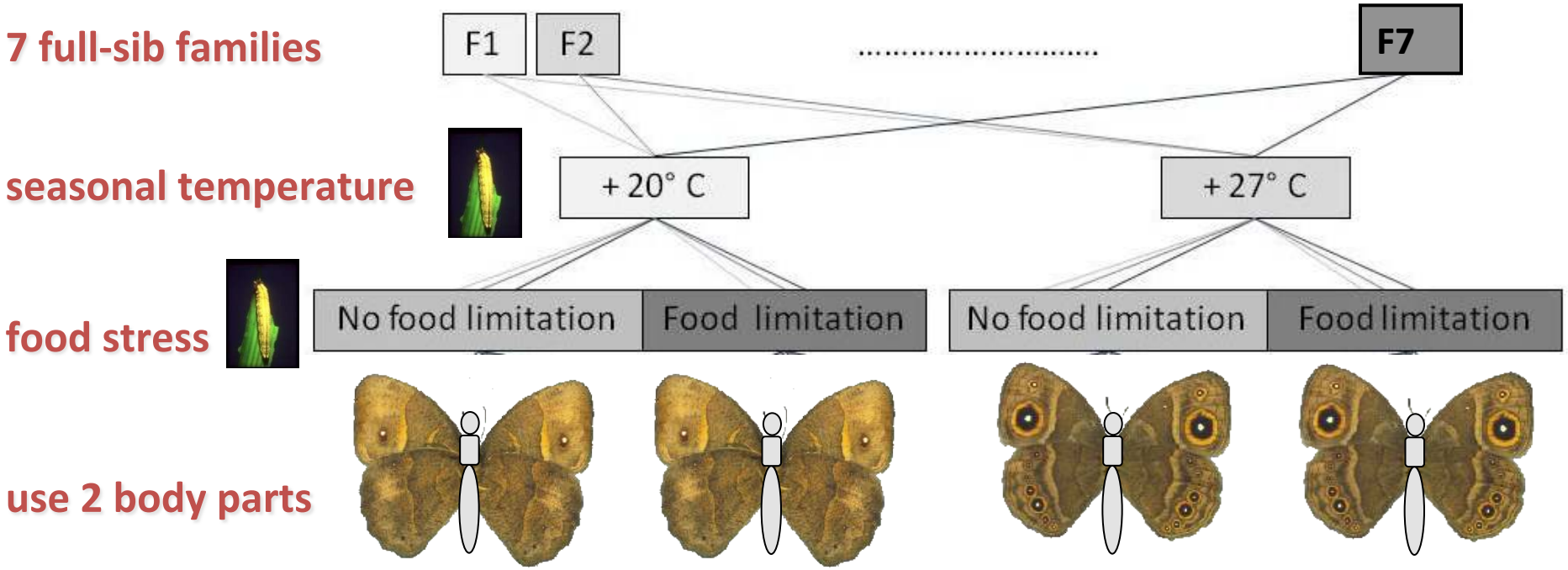
Marjo Saastamoinen



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

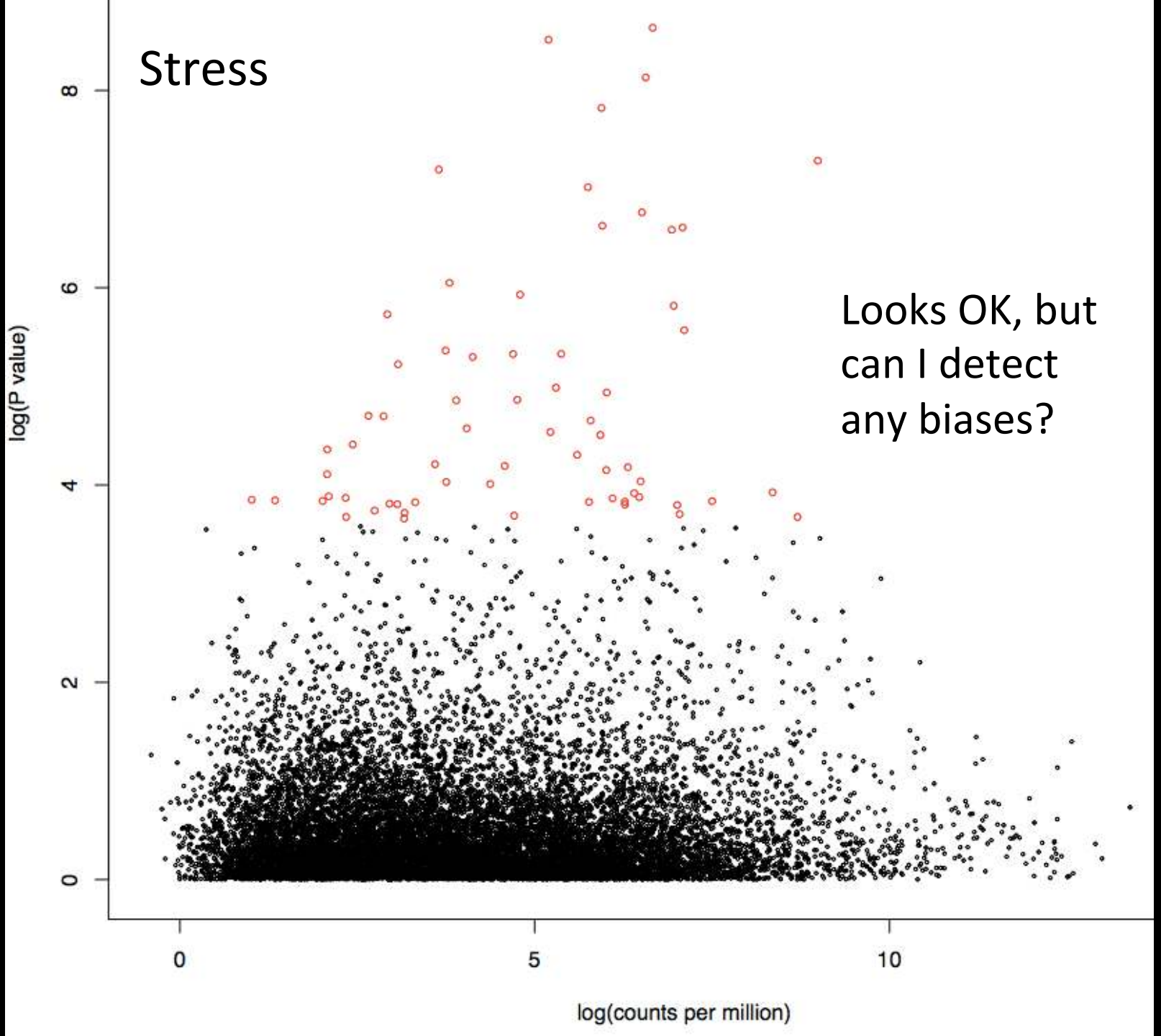
body part	# libraries	# clean reads (per library)	# nucleotides (per library)	GC content
abdomen	72	15,261,019	3,052,203,767	45%
thorax	72	15,633,416	3,126,683,150	46%
total	144	2,224,399,290	444,879,858,000	45%

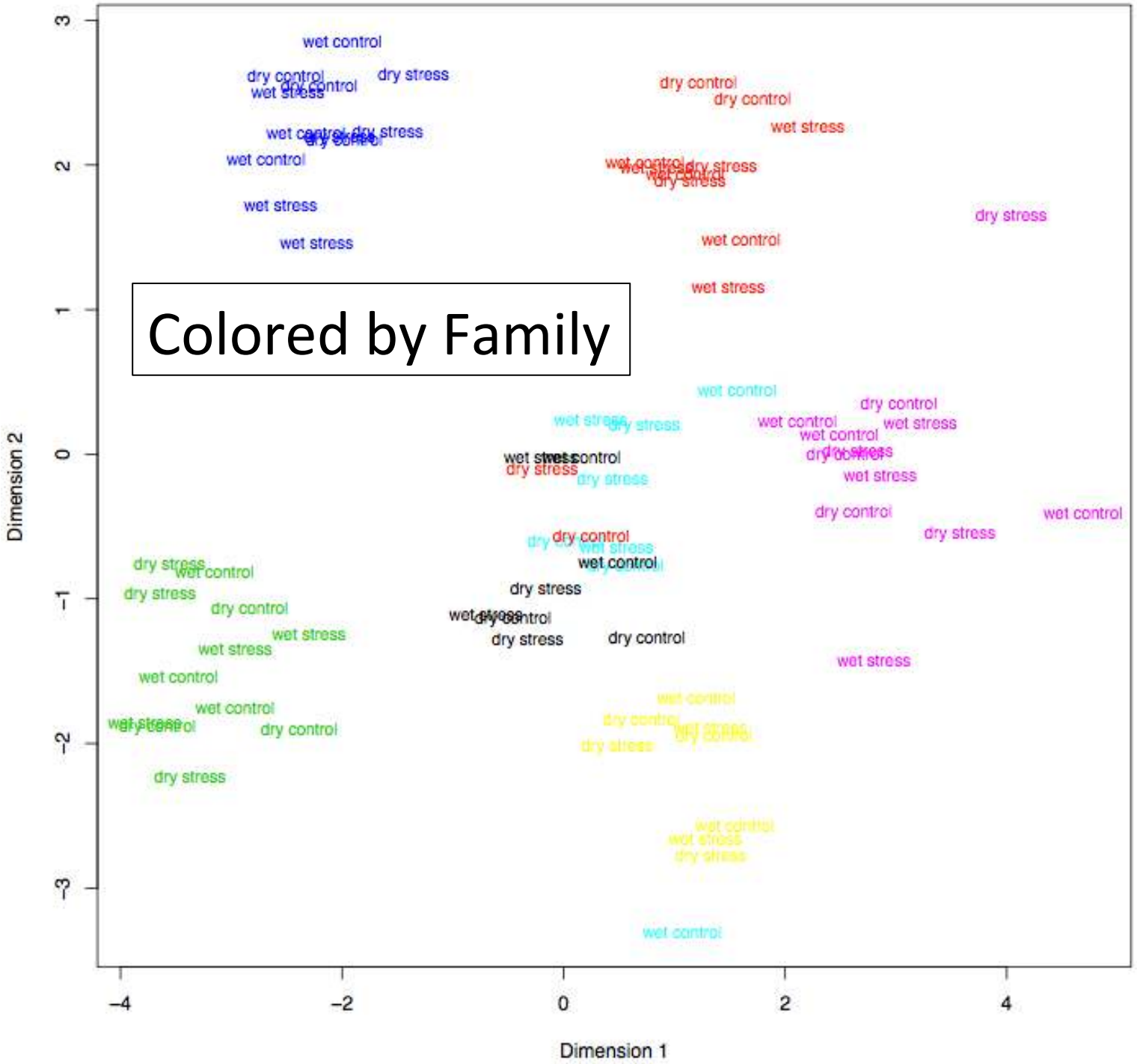
14 samples: one from each family, thorax and abdomen

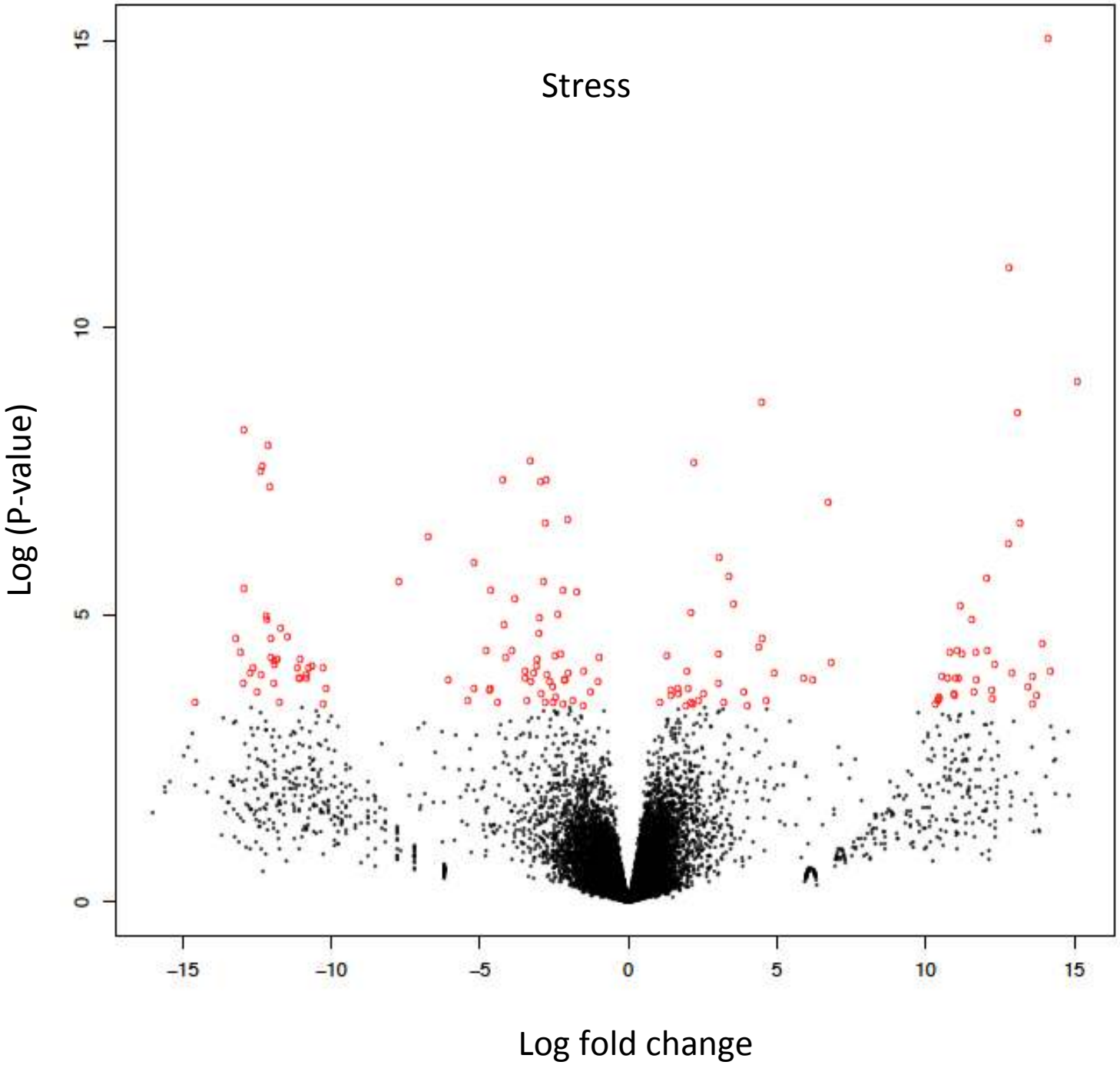
69,075 contigs

edgeR

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

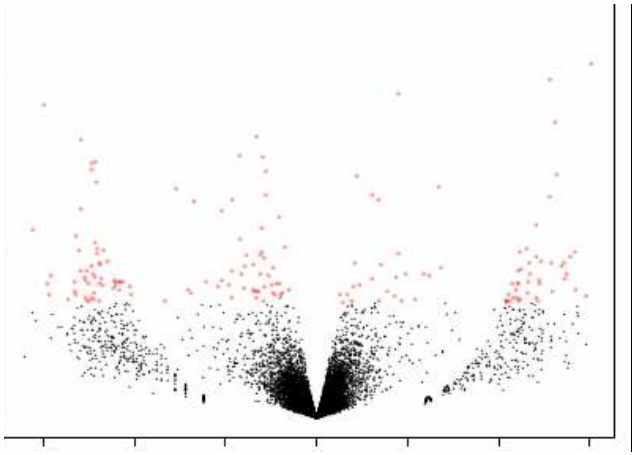

Stress
Looks OK, but can I detect any biases?
log(P value)
0
2
4
6
8
0
5
10
log(counts per million)


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

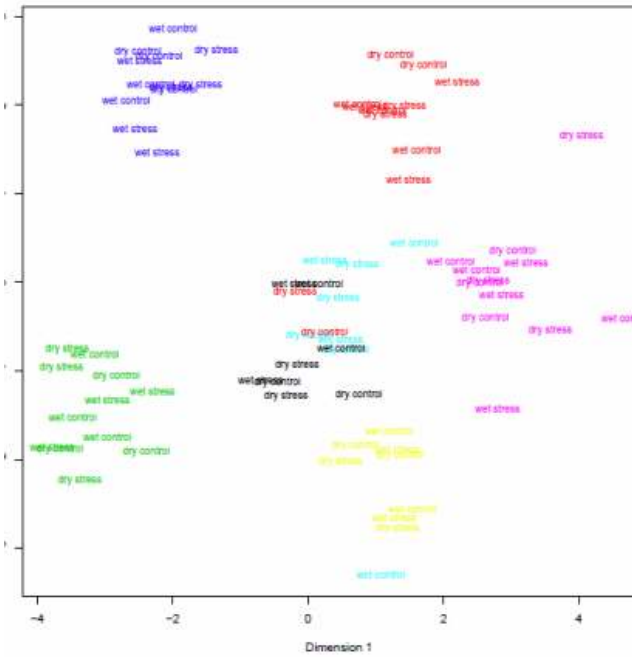

Stress
Log (P-value)
15
10
5
0
−15
−10
−5
0
5
10
15
Log fold change


Log (P-value)
Log fold change

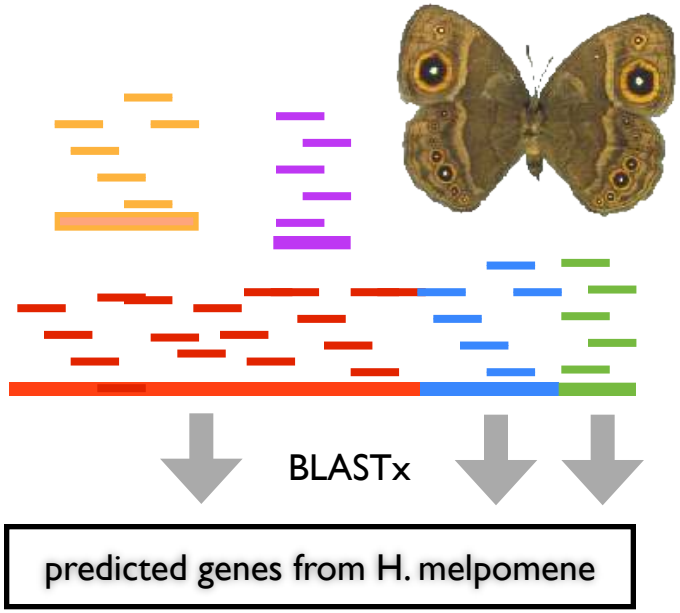
Effect of filtering, mapping to Trinity contigs

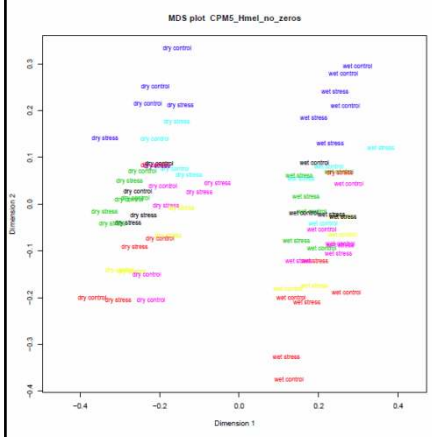
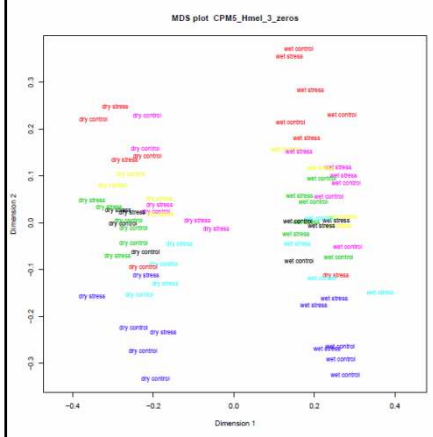
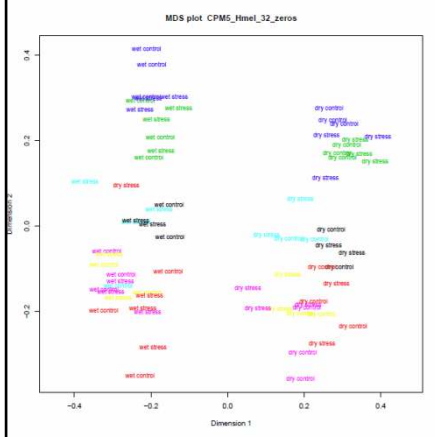
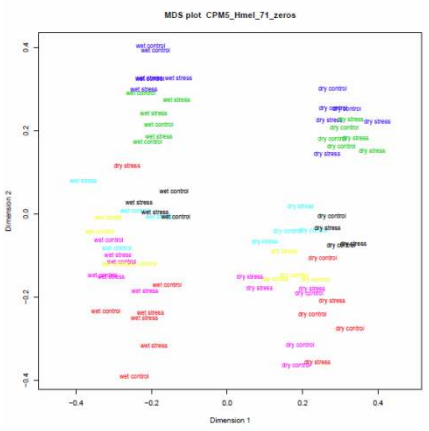


71 zero-read samples allowed



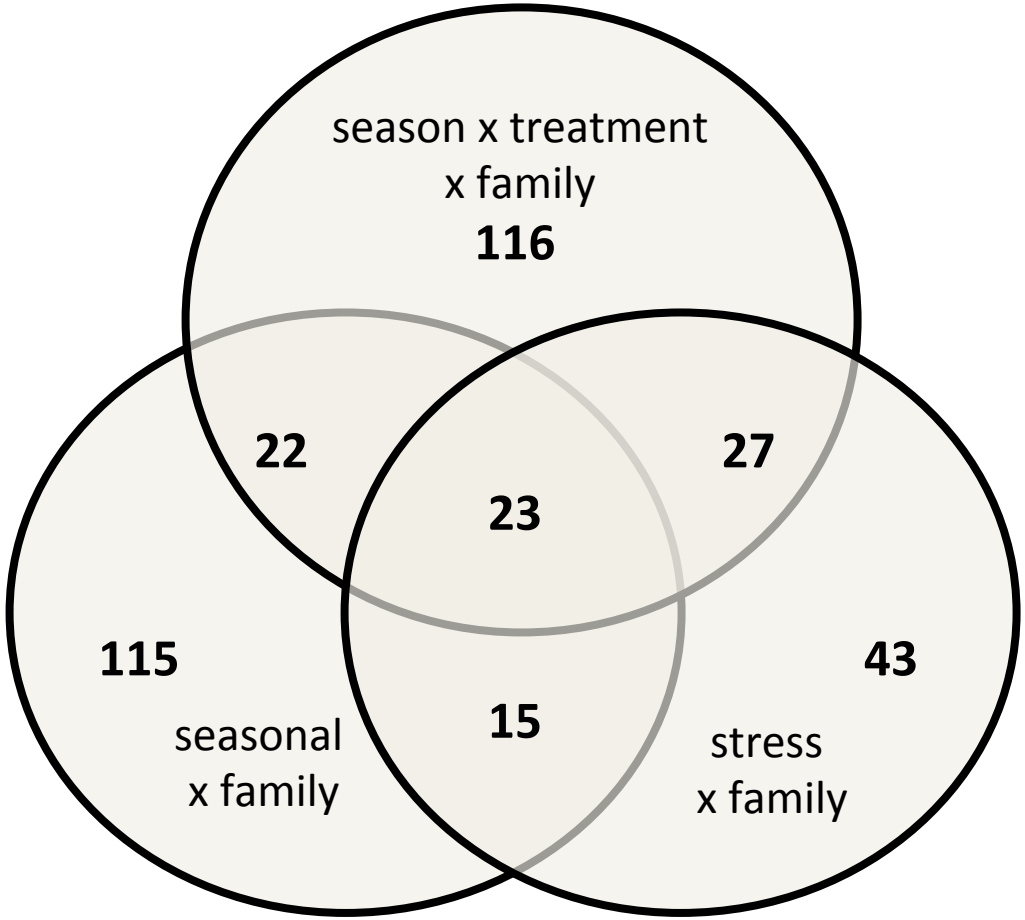
Effect of filtering when using sum method: whole gene expression



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

GLM results

- Plastic responses:
  - Effects without any interaction with Family



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

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

100 My

Bombyx mori
Whole genome sequence, predicted gene set

320 My

Drosophila melanogaster
Extensive genomic & functional resources

D. melanogaster lacks an orthologous reproductive physiology

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

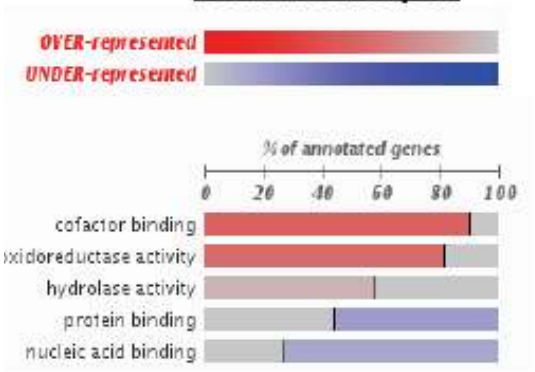
Blastx

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

Blastp

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

Gene Set Enrichment analysis using Gene Ontology database



Most studies are annotation limited

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

Description	Uniprot	-log10P
Oxidoreductase.	Q9VMH9	7.087008
Hypothetical protein.		6.993626
SD27140p.		6.315473
	Q8SXX2	6.300667
SD01790p.	Q95TI3	5.316371
Electron-transfer-flavoprotein	Q0KHZ6	5.1425
Pseudouridylate synthase.	Q9W282	4.784378
Hypothetical protein.	Q9VGX0	4.750469
CG14686-PA (RE68889p).	Q9VGX0	4.650051
Chromosome 11 SCAF14979, 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?

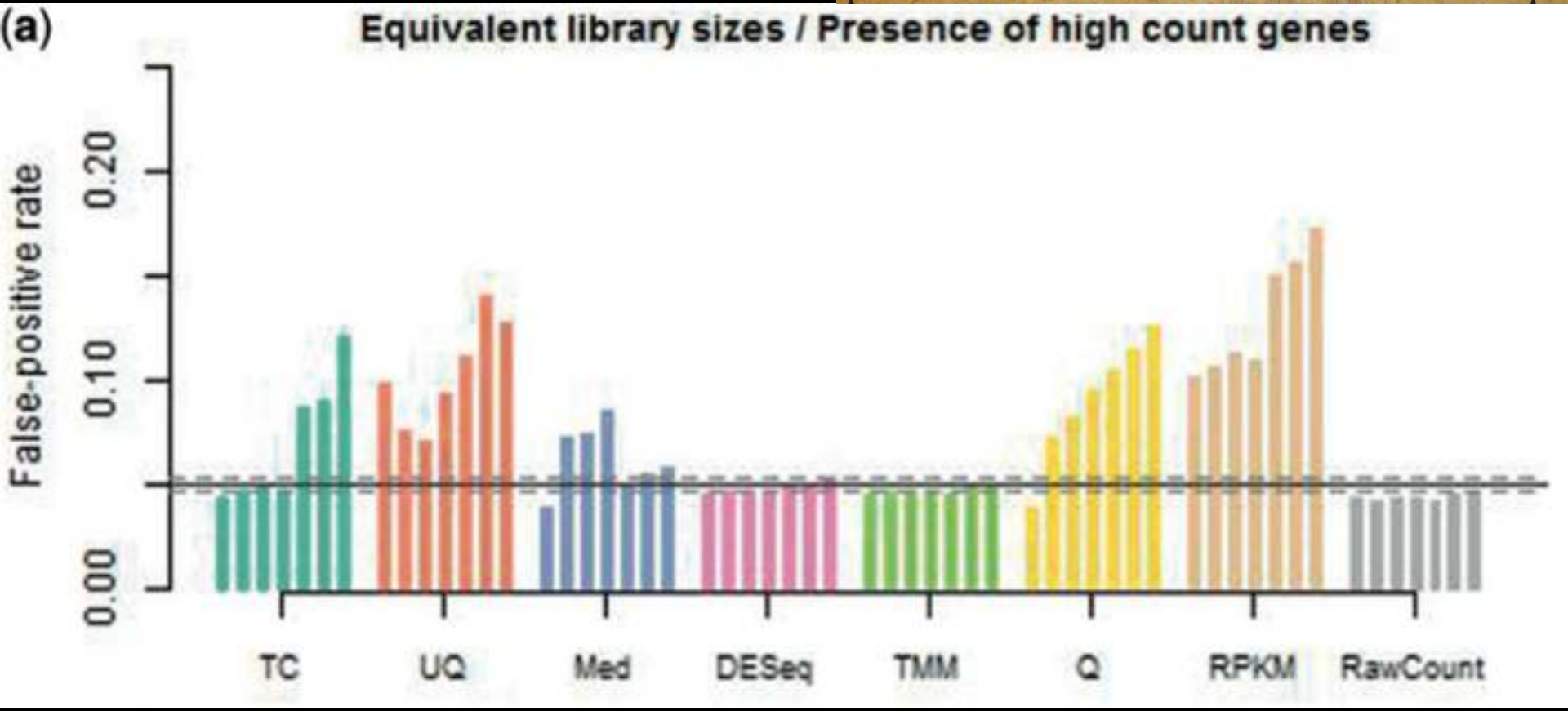
	Analysis # 1	Analysis # 2	Analysis # 3
Mapper	TopHat2	STAR	?
Normalization	none	RPKM	TPM
Analysis	PCA	RSEM	EDGER

Should independent methods converge?

Interrogate your results

- This will give you confidence
  - Bring freedom to your findings (no waterboarding)
- Graph your results
  - PCA plot or similar
  - 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?

Normalization & Analysis



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

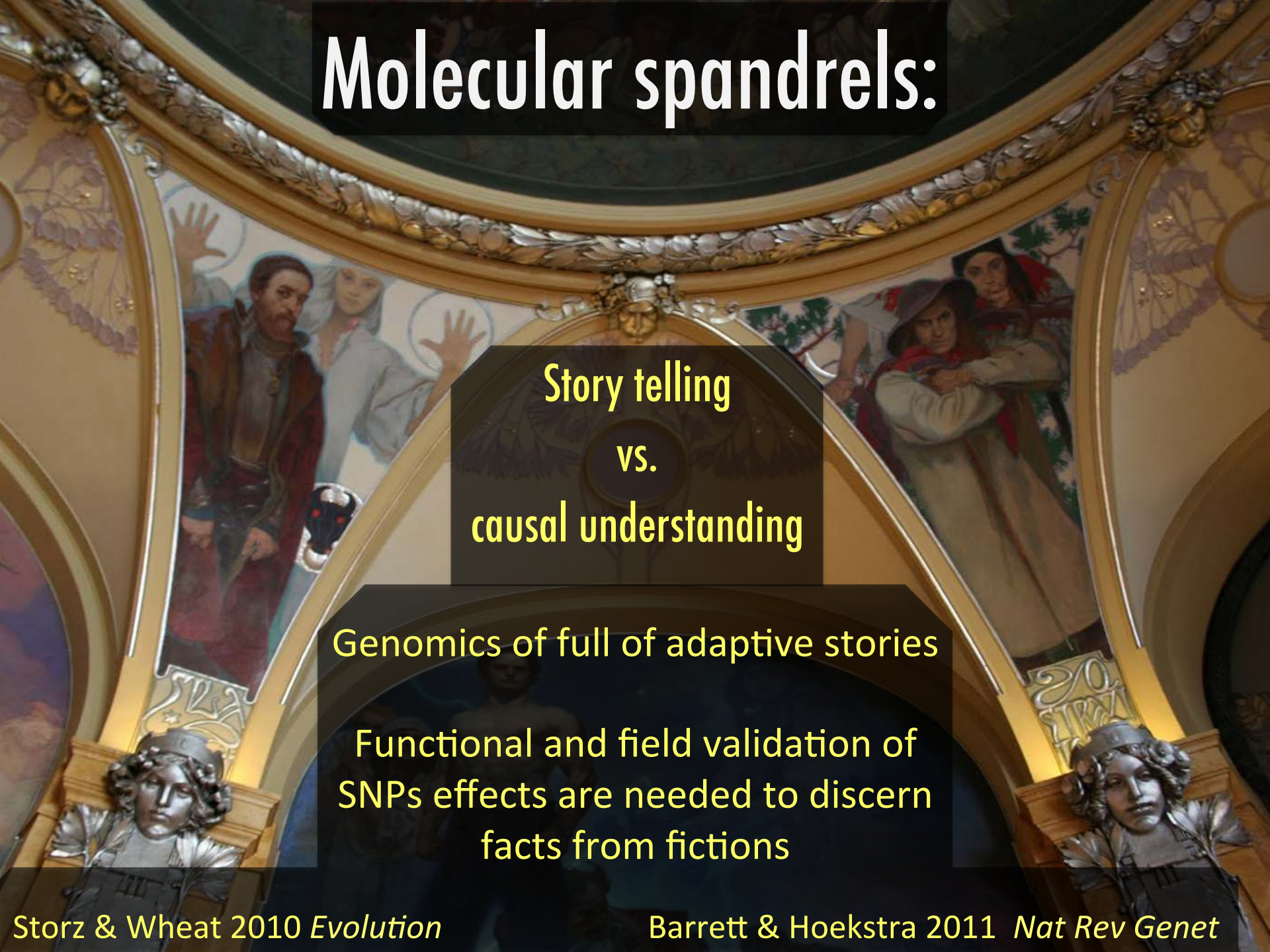
A major challenge for Ecological Genomics

- 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

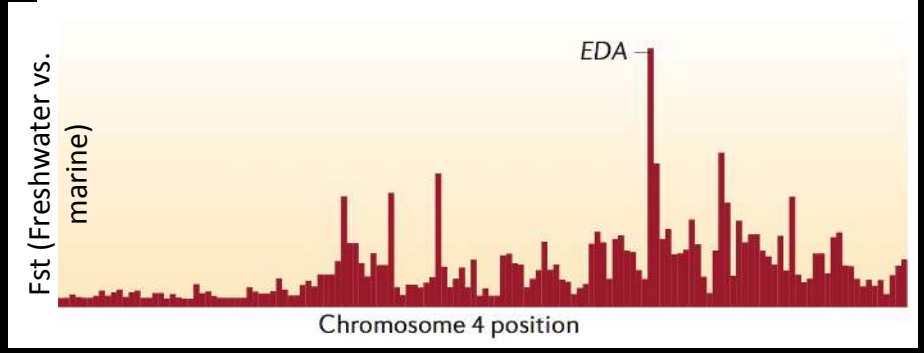
- Research community is demanding actual demonstration of natural selection when making claims of adaptive role

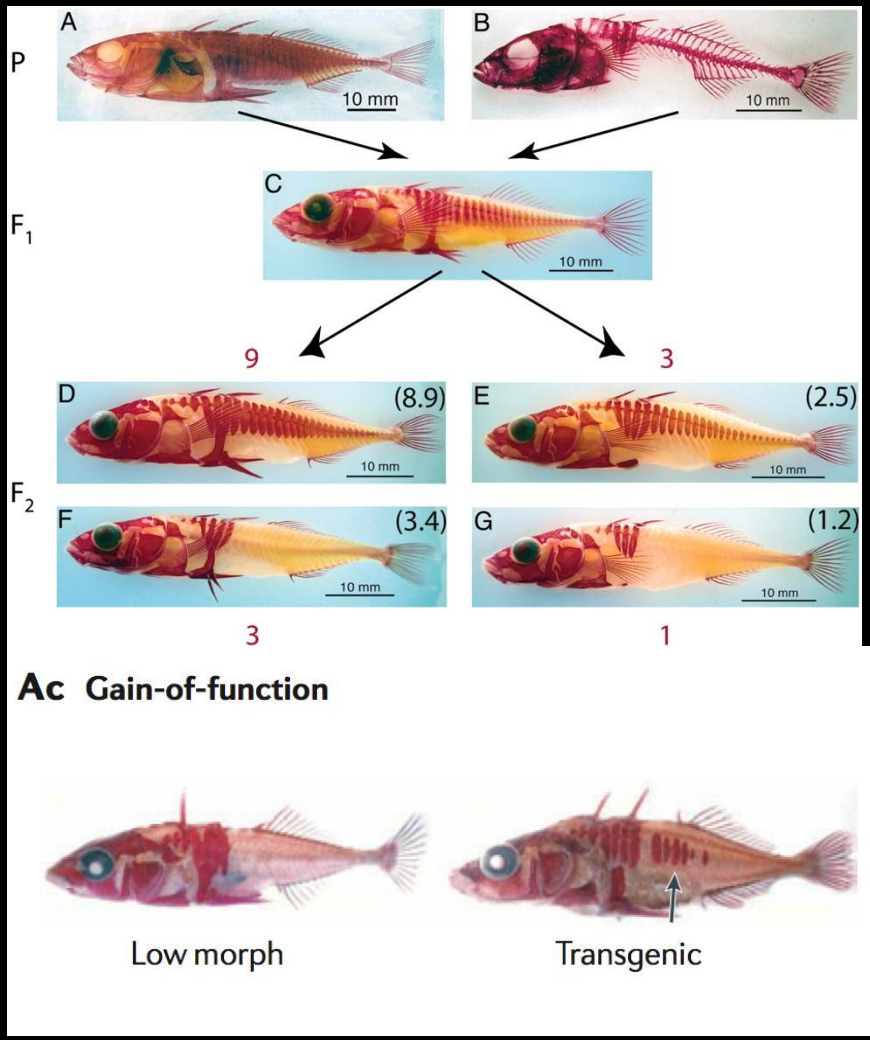
To address these we need to develop functional genomic insights in species with well understood ecologies that can be manipulated in the lab and in the field

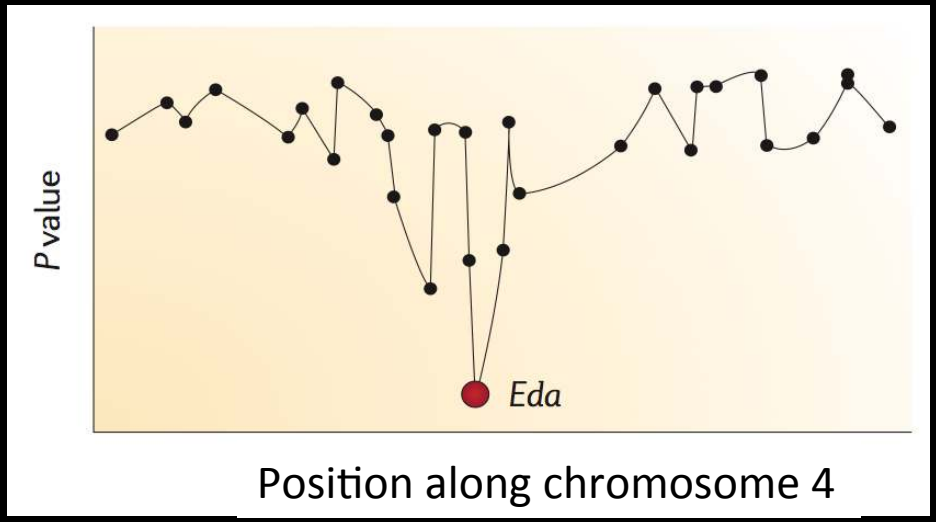

Molecular spandrels:
Story telling
vs.
causal understanding
Genomics of full of adaptive stories
Functional and field validation of SNPs effects are needed to discern facts from fictions
Storz & Wheat 2010 *Evolution*
Barrett & Hoekstra 2011 *Nat Rev Genet*

Model adaptation: the *Eda* gene

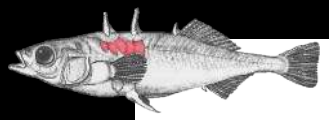
- Causes loss in body armor
  - Field association
  - QTL mapping
  - Gain-of-function assay







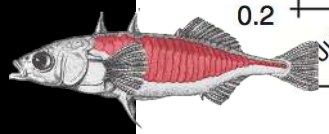
Back to nature: do we know what we think we know?



- 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

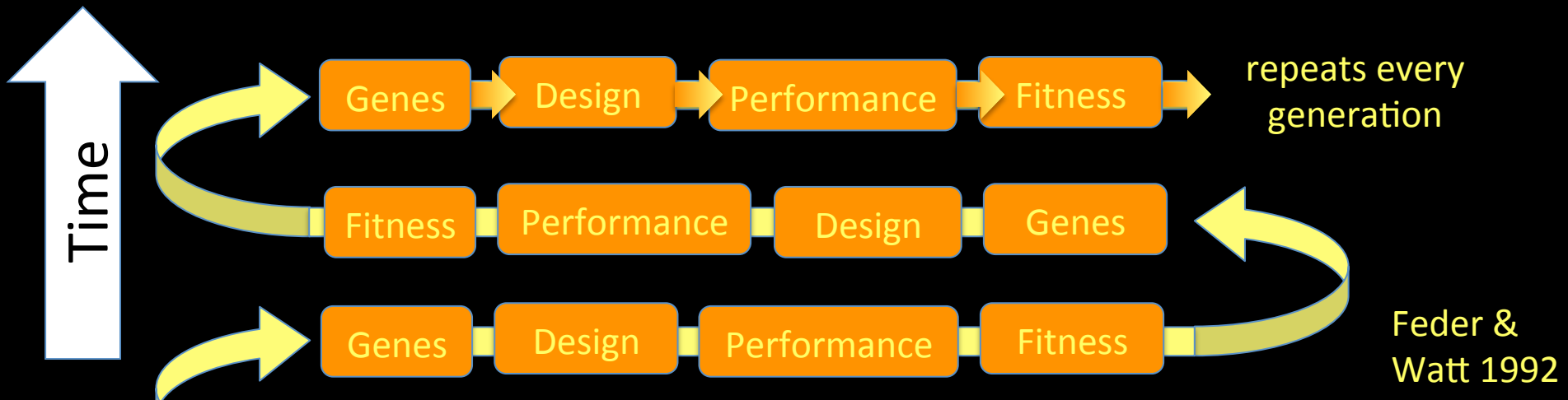
- Linked effect of other genes in the inversion on LG4?
- Is Eda the target of selection?



Barrett et al. 2008 Science

Adaptation by natural selection

Insights at one level can be tested at other levels



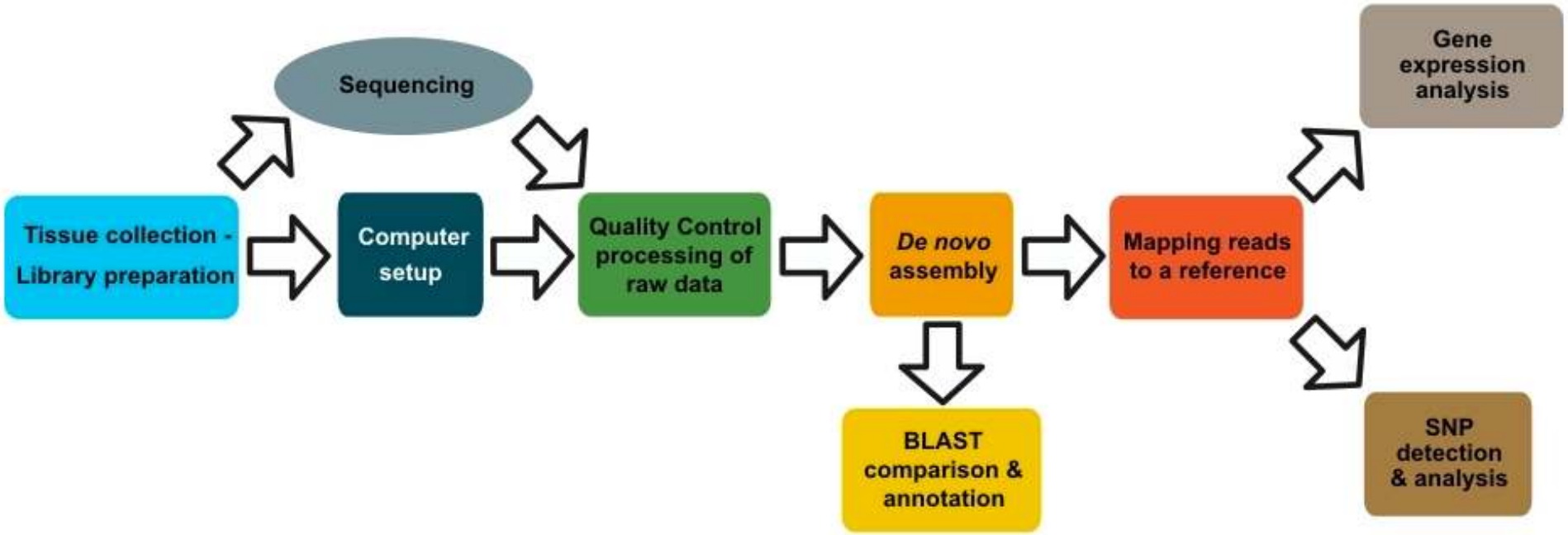
Common mistakes

- 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.

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



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

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"

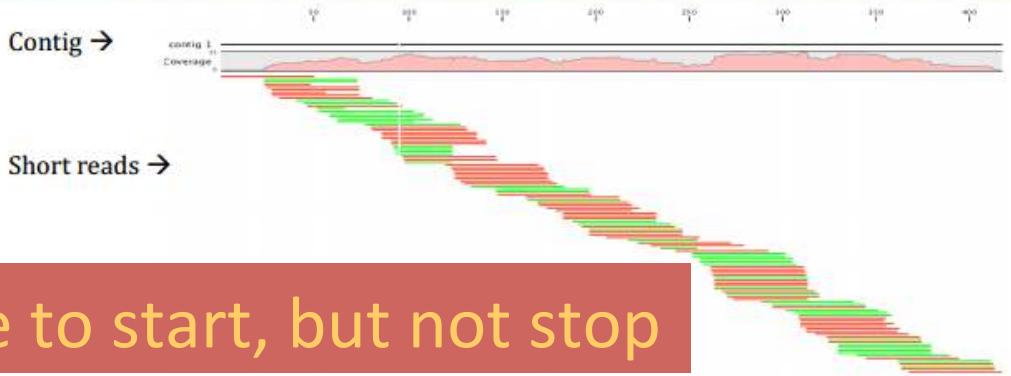
P: fastx_quality_stats (fastx toolkit)
S: QualityStats.sh
I: "YOURFILE_trimmed_clipped.fastq"
O: "YOURFILE_qualstats.txt"

For Paired-End samples, sort FASTQ files and remove orphan reads to a separate file.
P: fastxcombinepairedend.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



A great place to start, but not stop

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

Karl Gotthard

Pararge aegeria

Peter Pruisscher

GS-MESPA

Ram Neethiraj

Stockholm University

ACADEMY OF FINLAND

VETENSKAPSRÅDET
THE SWEDISH RESEARCH COUNCIL

Knut och Alice Wallenbergs Stiftelse