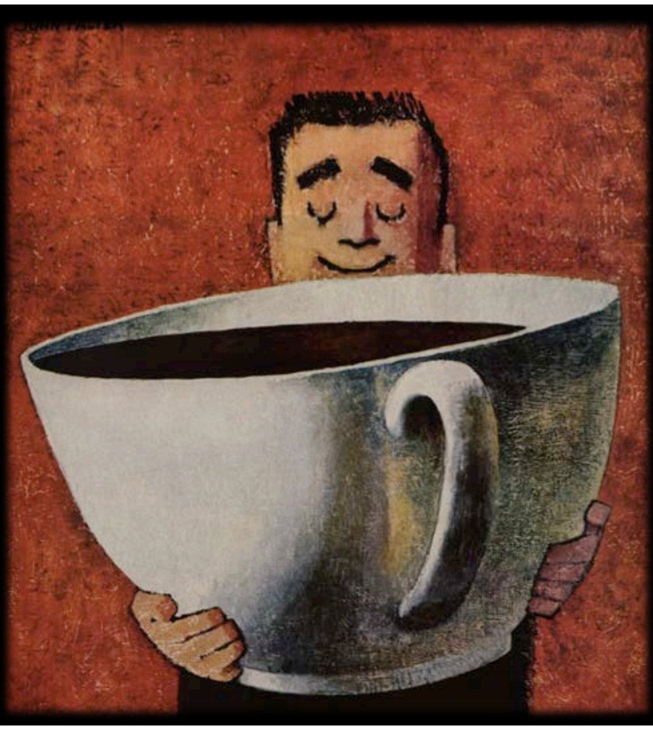


Outline

- What other biases do we suffer from?
- Here come the genomes ....
- Assembly errors and where they come from
- Annotation concerns
- RNAseq, reality and you ....

What other biases might we suffer from?

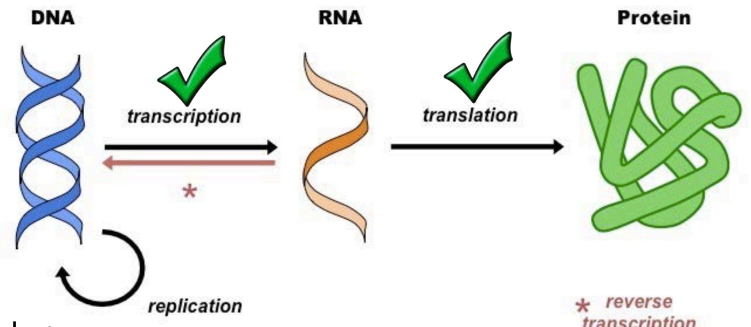
https://www.babyanimalprints.com/collections/monkeys-and-apes-black-and-white/chimpanzee

We're basically a rather lost, self domesticated chimp

We're very likely to :

- see patterns when none exist
- think we can predict the future, cause we think we know how things work … like:
  - gravity, your car, sunsets
  - weather, the stock market, Trump …
  - the central dogma …..

The central dogma



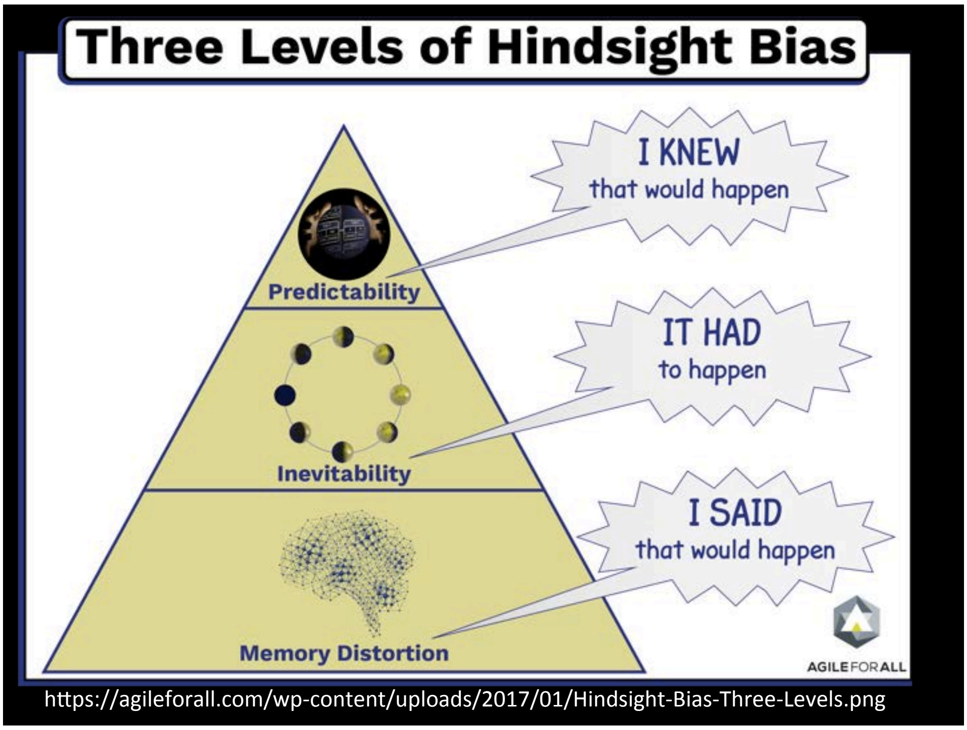
What about:

- Gene expression level based upon enhancer region? 🚫
- When and where a gene will be expressed from enhancer region? 🚫
- How will RNA sequence will fold into a 2^{o} structure? 🚫
- How will a protein sequence will fold into a 3^{o} structure? 🚫
- Function of an enzyme based upon its structure? 🚫
- Write a protein that will fold and do a specific enzymatic task? 🚫

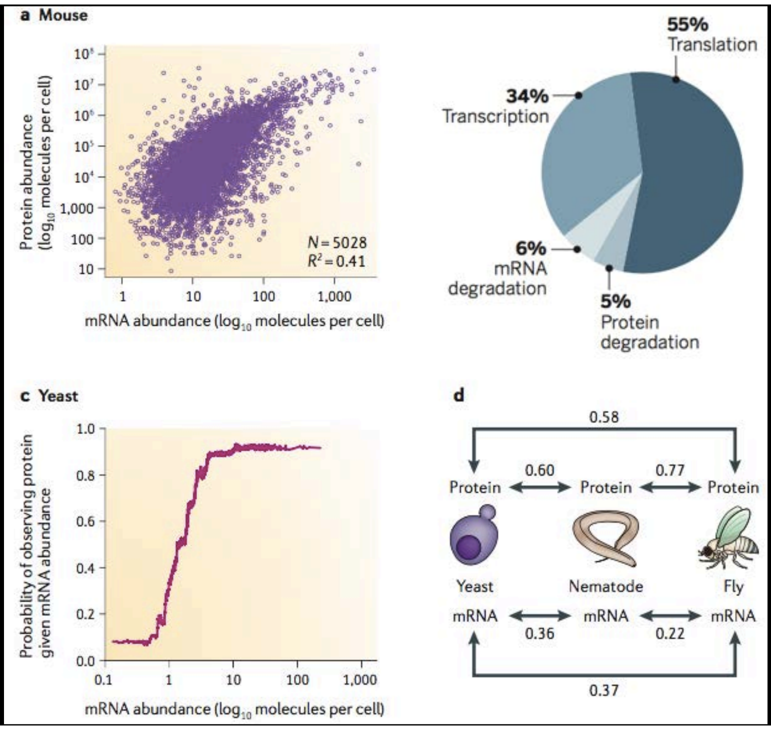
Hindsight bias

the knew-it-all-along effect

the inclination, after an event has occurred, to see the event as having been predictable, despite there having been little or no objective basis for predicting it.

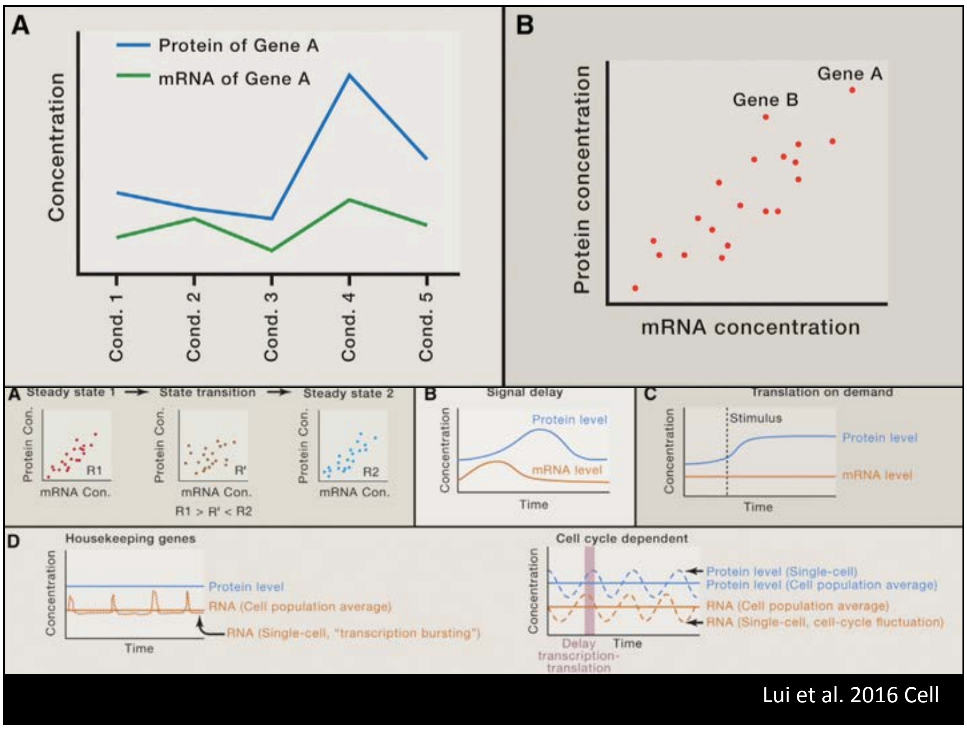


https://agileforall.com/wp-content/uploads/2017/01/Hindsight-Bias-Three-Levels.png



Li and Biggin 2015

Vogel and Marcotte 2012 NRG

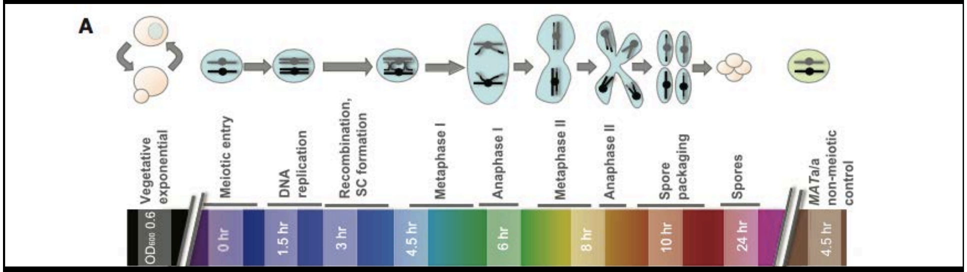


Lui et al. 2016 Cell

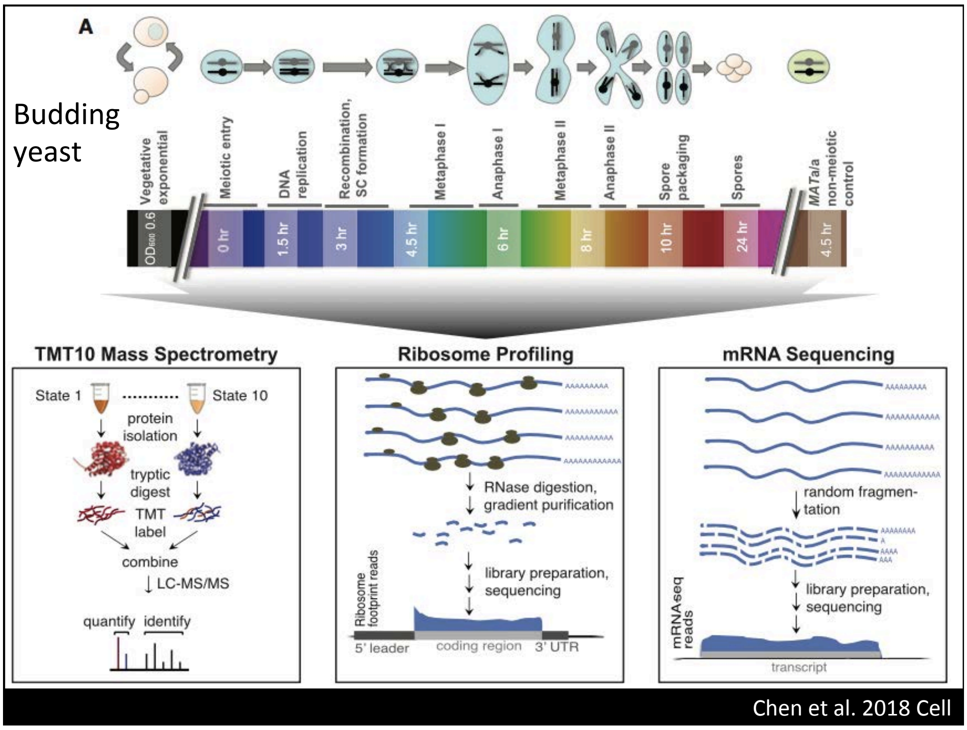
Cell

Article

Pervasive, Coordinated Protein-Level Changes Driven by Transcript Isoform Switching during Meiosis



- What's the mRNA to protein relationship in yeast
  - Across standardized cell cycles



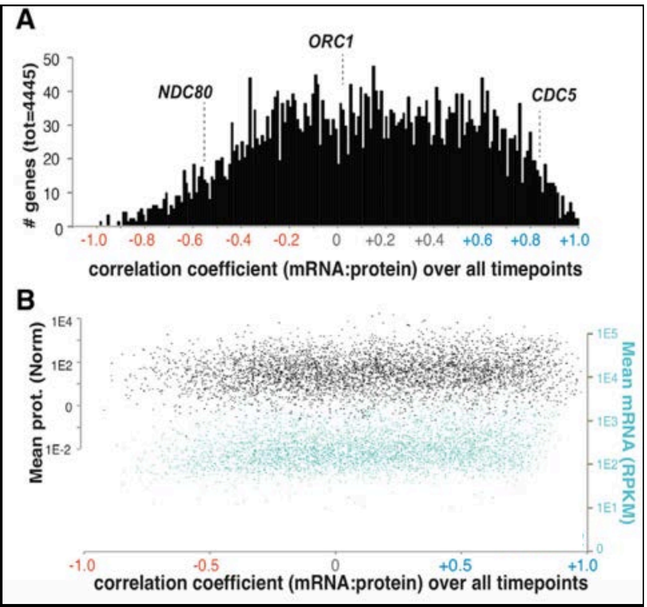
Budding yeast

Chen et al. 2018 Cell

Simultaneous measurements of:

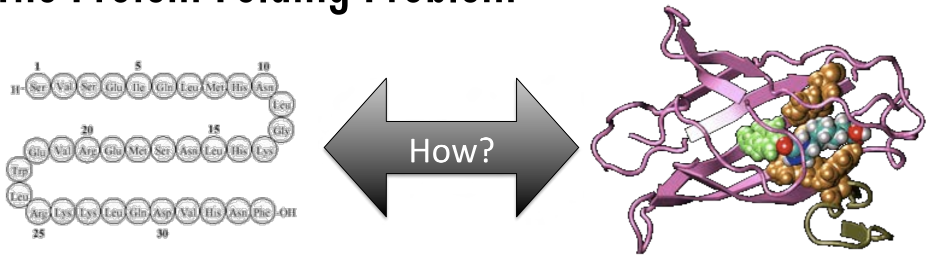
- mRNA
- translation
- Protein abundance

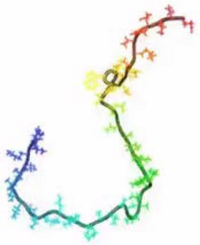
through meiotic differentiation in budding yeast

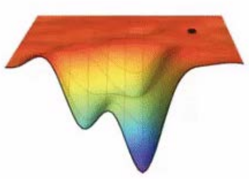


Chen et al. 2018 Cell

The Protein Folding Problem

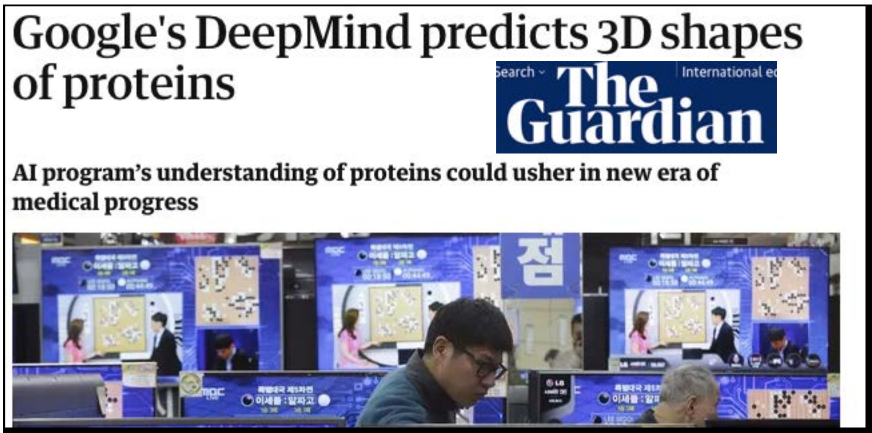

1
5
10
H-Ser Val Ser Glu Ile Gln Leu Met His Asn
Leu
Gly
15
20
Glu Val Arg Glu Met Ser Asn Leu His Lys
Trp
Leu
Arg Lys Lys Leu Gln Asp Val His Asn Phe-OH
25
30
How?





https://gfycat.com/greenpertinentkomododragon

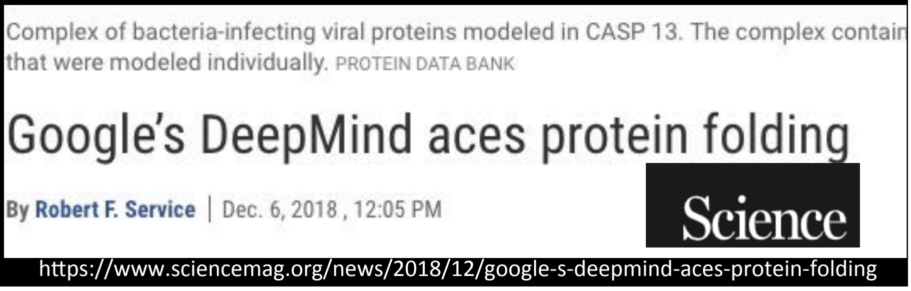
https://zhanglab.ccmb.med.umich.edu/image/Protein_design.gif



Google's DeepMind predicts 3D shapes of proteins

Search ⌄ The Guardian International e

AI program's understanding of proteins could usher in new era of medical progress



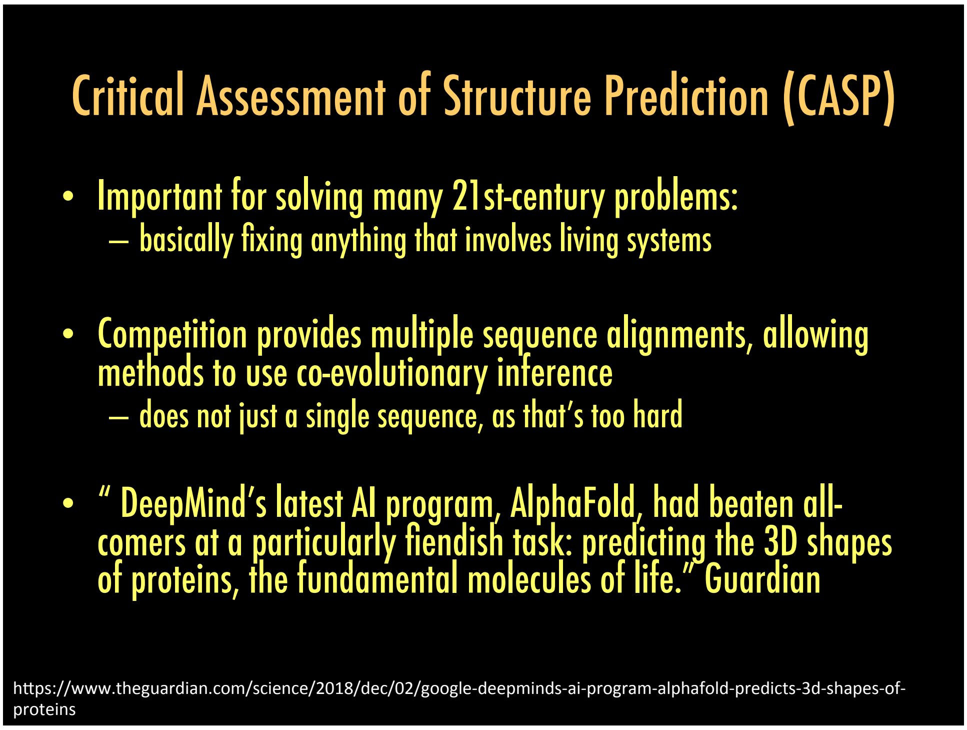
Complex of bacteria-infecting viral proteins modeled in CASP 13. The complex contain that were modeled individually. PROTEIN DATA BANK

Google's DeepMind aces protein folding

By Robert F. Service | Dec. 6, 2018 , 12:05 PM

Science

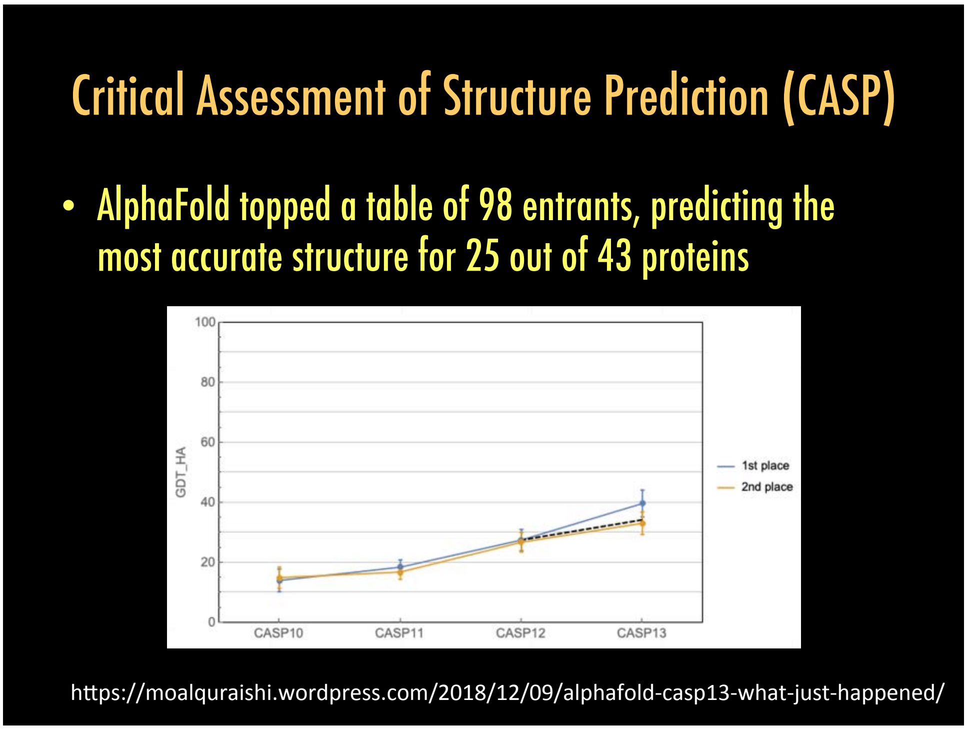
https://www.sciencemag.org/news/2018/12/google-s-deepmind-aces-protein-folding



Critical Assessment of Structure Prediction (CASP)

- Important for solving many 21st-century problems:
  - basically fixing anything that involves living systems
- Competition provides multiple sequence alignments, allowing methods to use co-evolutionary inference
  - does not just a single sequence, as that's too hard
- " DeepMind's latest AI program, AlphaFold, had beaten all-comers at a particularly fiendish task: predicting the 3D shapes of proteins, the fundamental molecules of life." Guardian

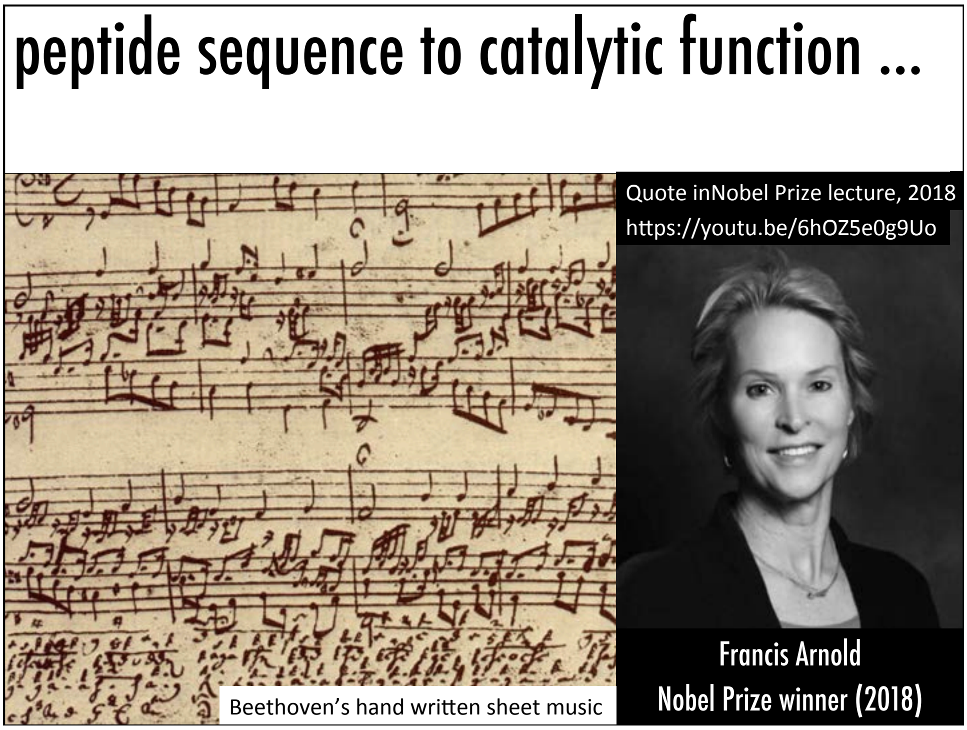
https://www.theguardian.com/science/2018/dec/02/google-deepminds-ai-program-alphafold-predicts-3d-shapes-of-proteins



Critical Assessment of Structure Prediction (CASP)

- AlphaFold topped a table of 98 entrants, predicting the most accurate structure for 25 out of 43 proteins

https://moalquraishi.wordpress.com/2018/12/09/alphafold-casp13-what-just-happened/

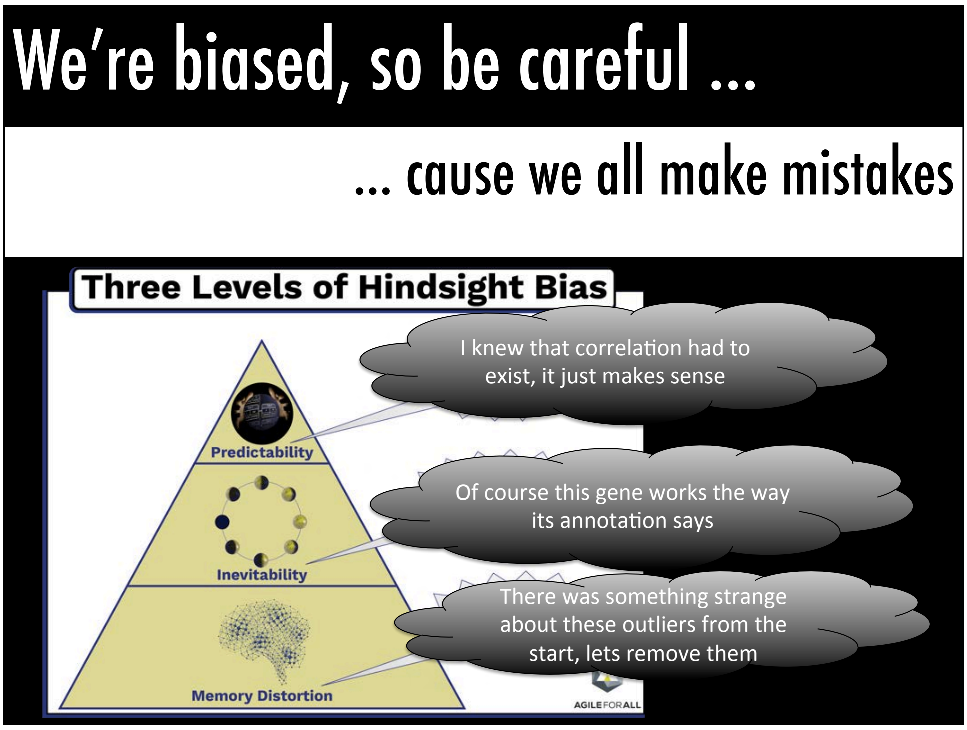


peptide sequence to catalytic function …

Quote inNobel Prize lecture, 2018
https://youtu.be/6hOZ5e0g9Uo

Beethoven’s hand written sheet music

Francis Arnold
Nobel Prize winner (2018)



We’re biased, so be careful …

… cause we all make mistakes

I knew that correlation had to exist, it just makes sense

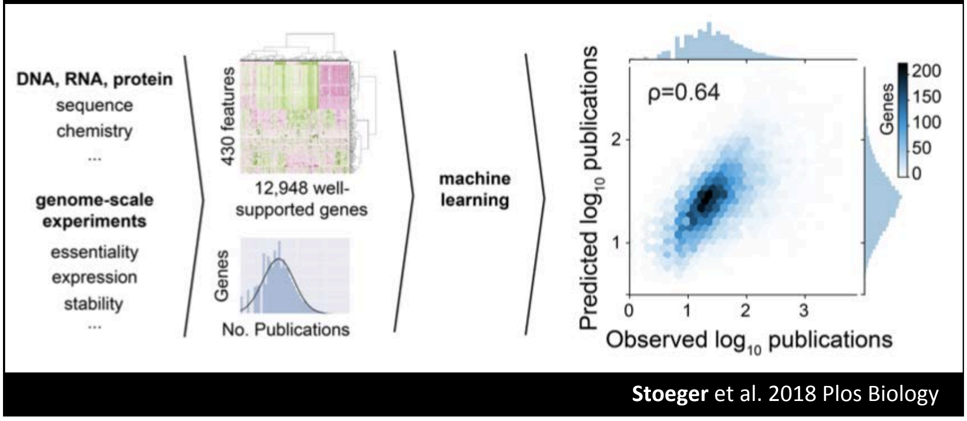
Of course this gene works the way its annotation says

There was something strange about these outliers from the start, lets remove them

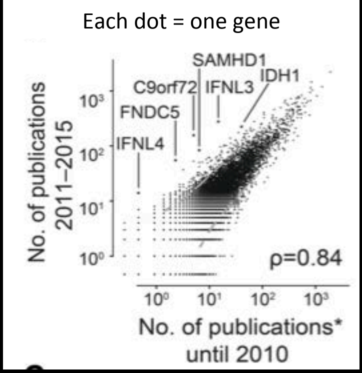
Do researchers distribute their attention equally across all genes?

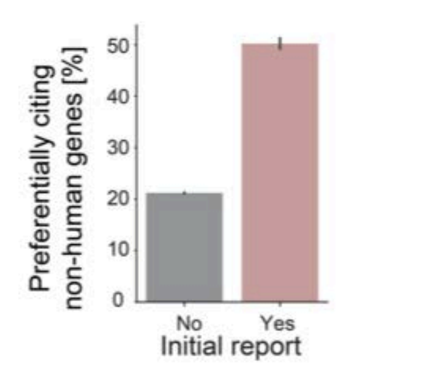
Do we ever conduct "unbiased" investigations?

What if we looked at investigations by gene, over time



Stoeger et al. 2018 Plos Biology


Each dot = one gene



- 30 percent of all genes have never been the focus of a scientific study
- less than 10 percent of genes are the subject of more than 90 percent of published papers
- historical and biological reasons rather then relevance drive study

Stoeger et al. 2018 Plos Biology

Get ready, here come the 1000^n genomes

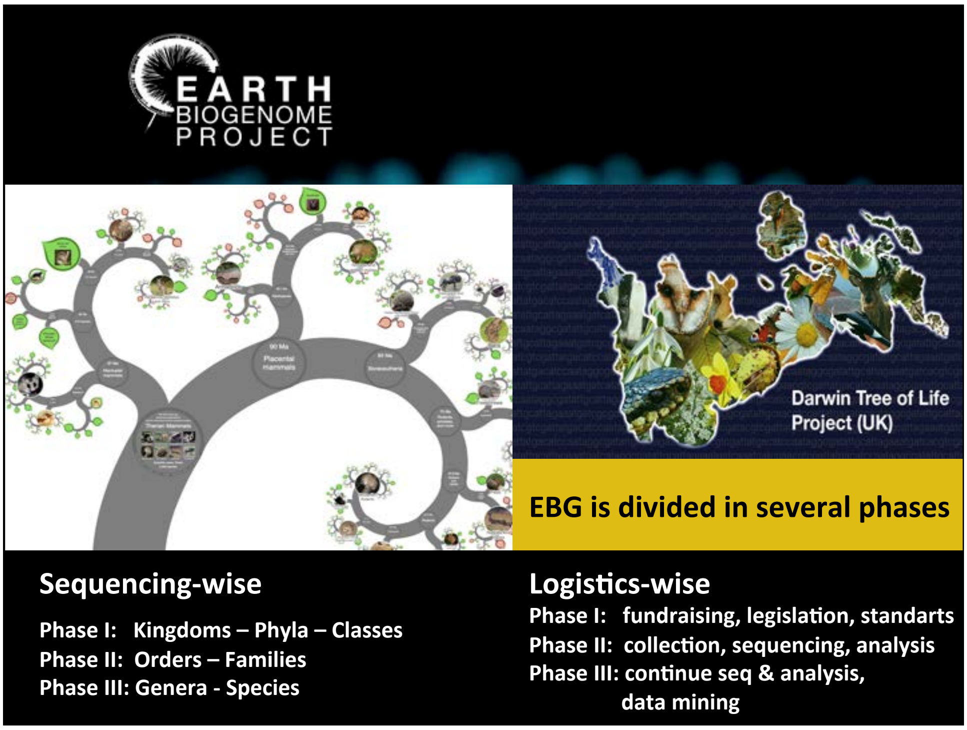
An unprecedented opportunity for large scale errors?

dying:

ships

– ion

– Functional insights into genes and genomic features (e.g. regulation and inheritance)



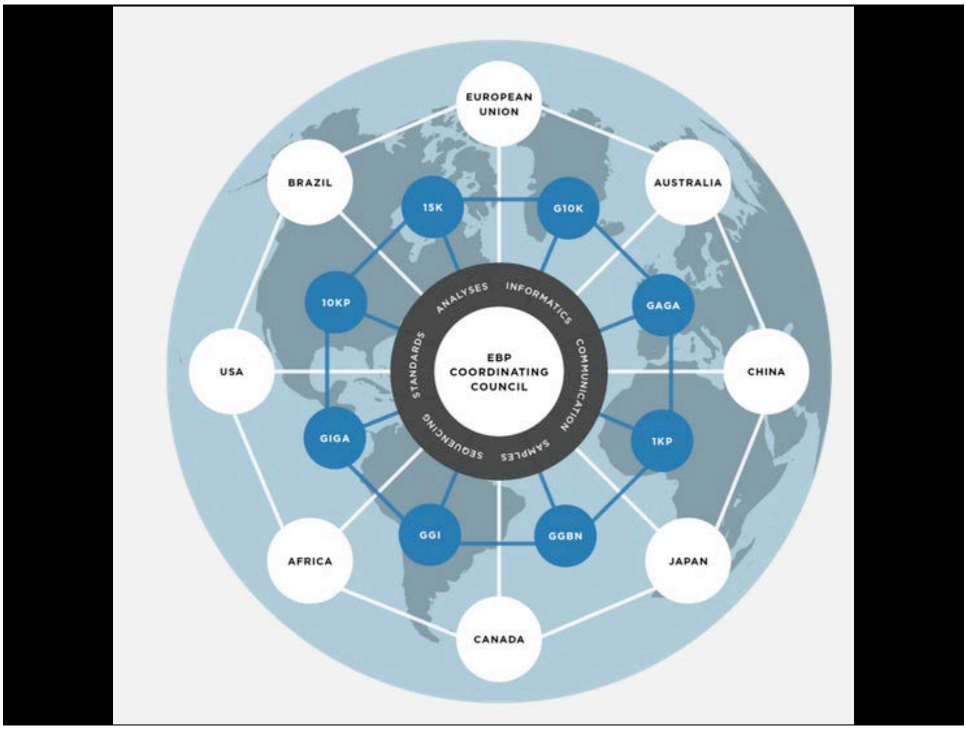
EBG is divided in several phases

Sequencing-wise

Phase I: Kingdoms – Phyla – Classes
Phase II: Orders – Families
Phase III: Genera - Species

Logistics-wise

Phase I: fundraising, legislation, standarts
Phase II: collection, sequencing, analysis
Phase III: continue seq & analysis, data mining



EARTH BIOGENOME PROJECT

VGP

54	58	32	30	90	4
BIRDS	MAMMALS	REPTILES	AMPHIBIANS	FISH	INVERTEBRATES

OVER 260 SPECIES

Standard: 3.4.2.QV40 PHASED

3: Contig N50 ≥ *1 Mb*
4: Scaffold N50 ≥ *10 Mb*
2: >90% of scaffolds are assigned to chromosomes, confirmed by two independent sources
QV40: Phred-score of average base quality is 1 error per 10 kb of sequence
Phased: Individual haplotypes should be resolved

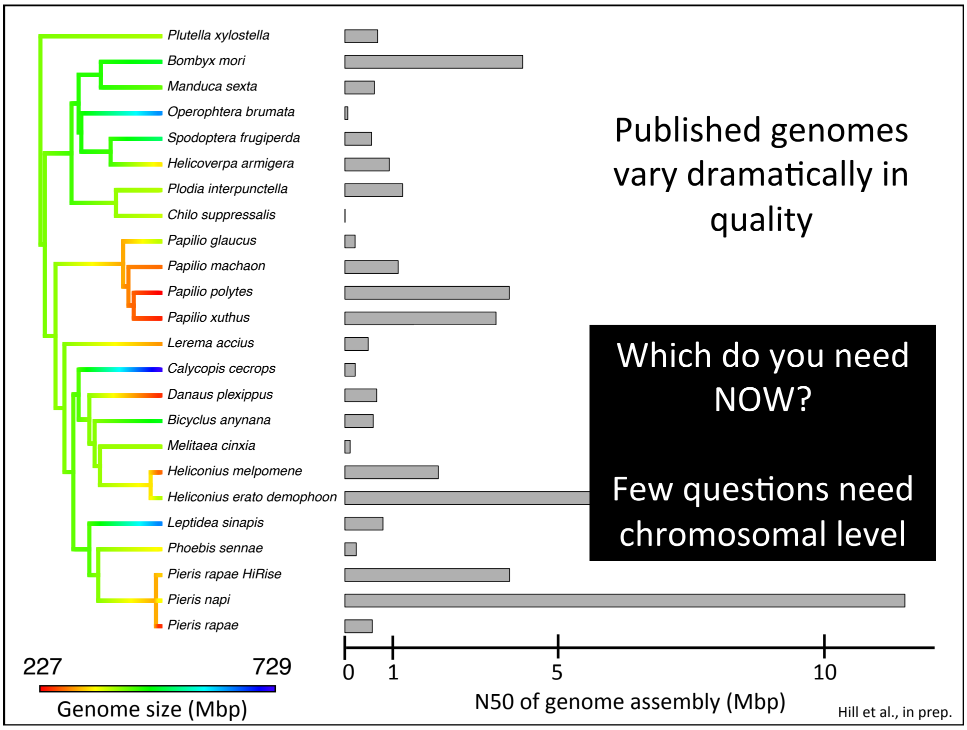
Sequencing & assembly:
PacBio (min 60x) + 10x Genomics + BioNano + HiC
(+ *ultralong* Nanopore reads)

Annotation (NCBI):
mRNA from at least 3 tissues (preferred: ***brain***, spleen, testis/ovaries), Illumina + Iso-Seq
miRNA
lncRNA

Slides from : Olga Pettersson, SciLifeLab, Sweden

So … how many of you are sequencing a genome?

- What does that mean? Have you told your mom?
- What kind of genome are you generating?
- What do you need, what is your question?
  - Short term vs. long term goals?
  - Are these in conflict?

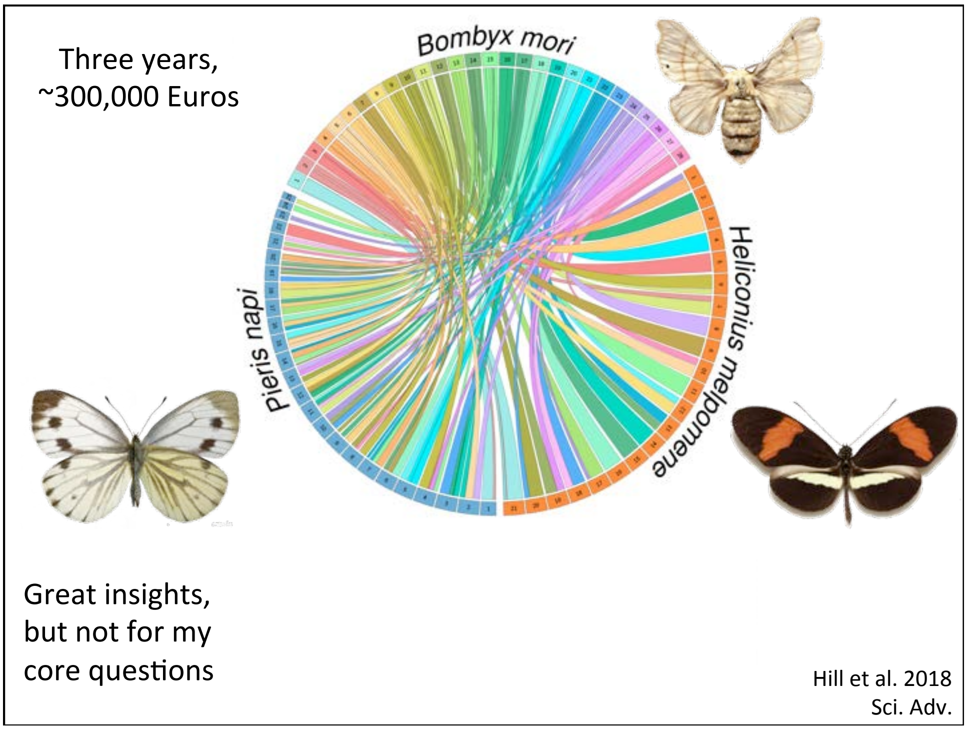


Published genomes vary dramatically in quality

Which do you need NOW?

Few questions need chromosomal level

Hill et al., in prep.



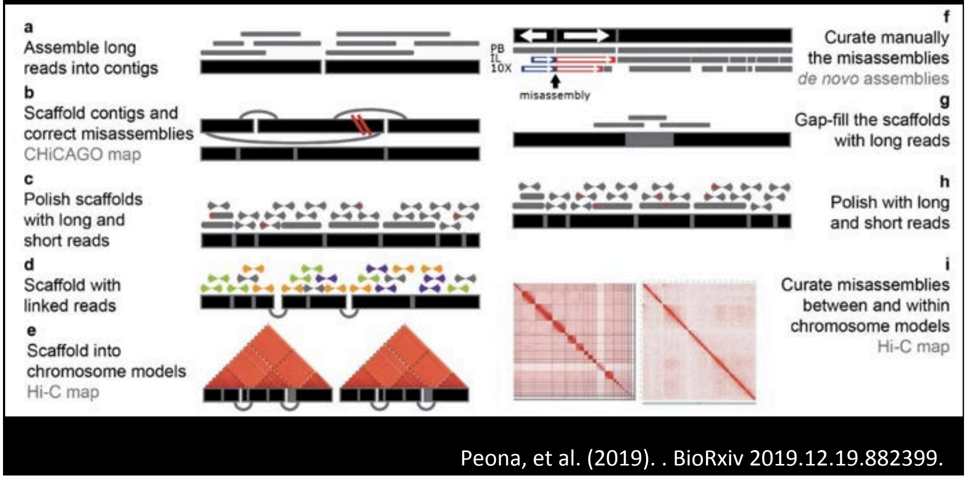
Three years,
~300,000 Euros

Great insights,
but not for my
core questions

Hill et al. 2018
Sci. Adv.

What determines genome quailty?

Identifying the causes and consequences of assembly gaps using a multiplatform genome assembly of a bird-of-paradise



a
Assemble long reads into contigs

b
Scaffold contigs and correct misassemblies
CHiCAGO map

c
Polish scaffolds with long and short reads

d
Scaffold with linked reads

e
Scaffold into chromosome models
Hi-C map

f
Curate manually the misassemblies
de novo assemblies

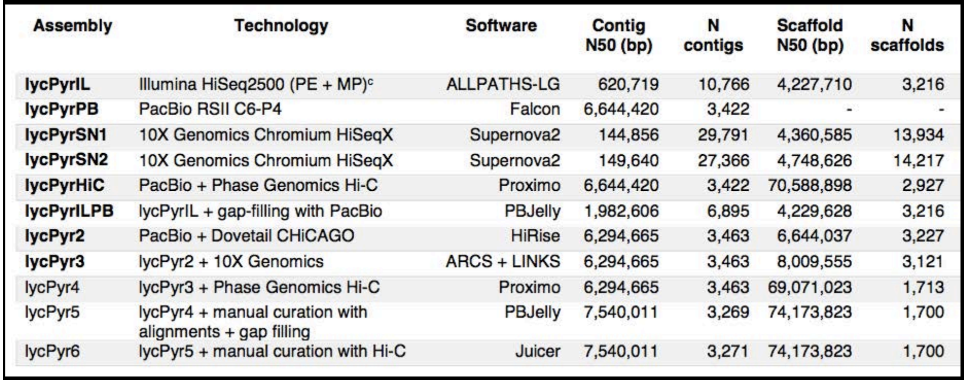
g
Gap-fill the scaffolds with long reads

h
Polish with long and short reads

i
Curate misassemblies between and within chromosome models
Hi-C map

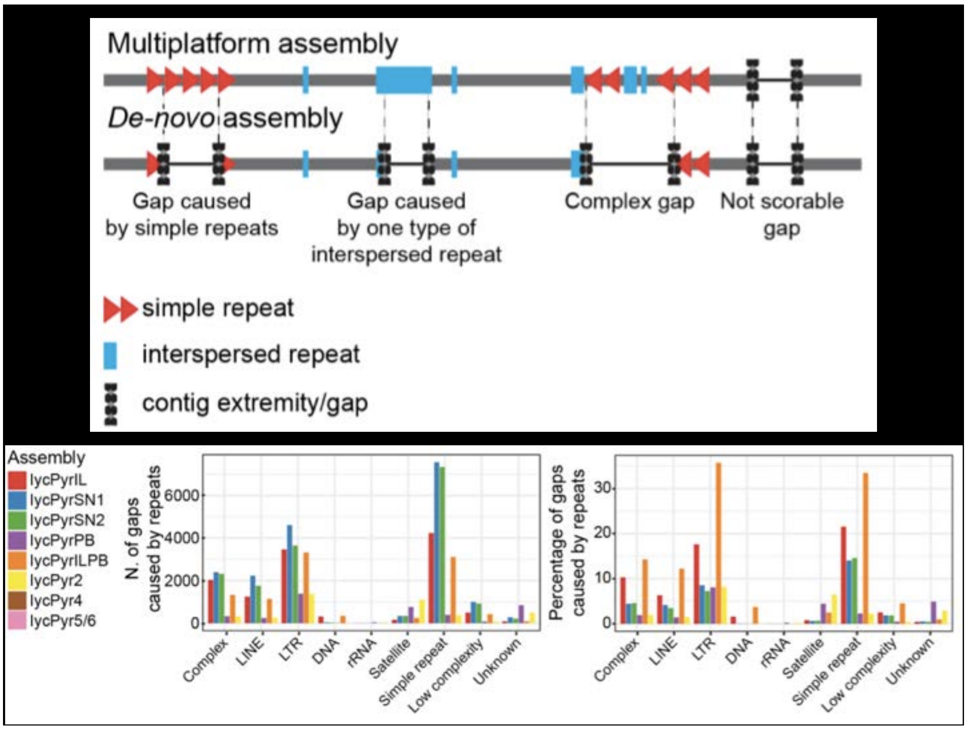
Peona, et al. (2019). . BioRxiv 2019.12.19.882399.

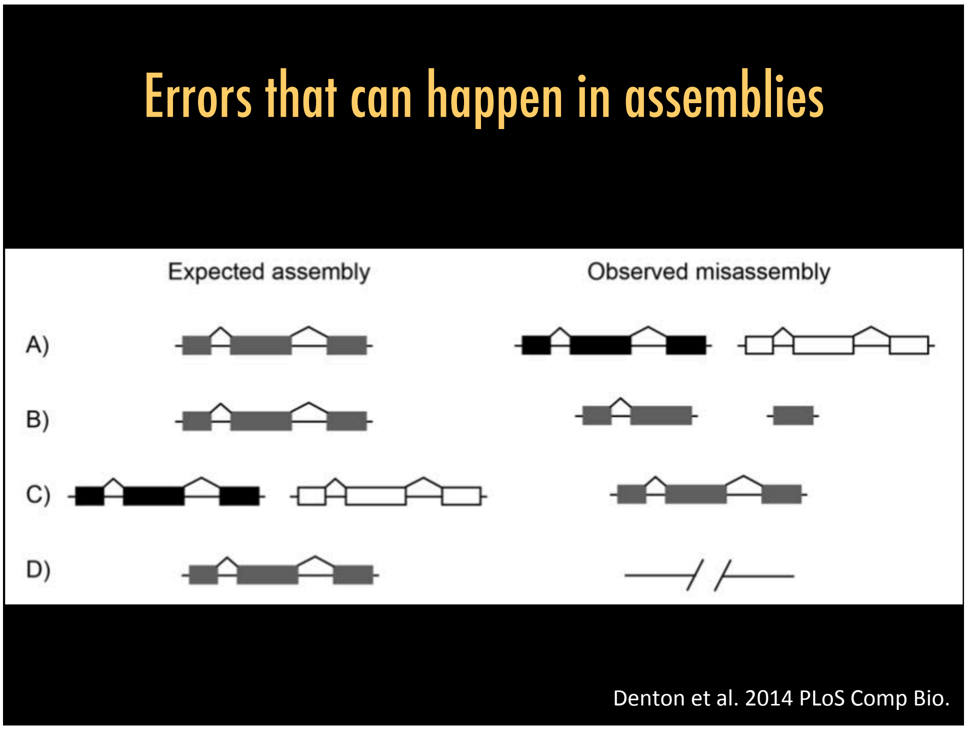
They made lots of assemblies along the way



Assembly	Technology	Software	Contig N50 (bp)	N contigs	Scaffold N50 (bp)	N scaffolds
lycPyrIL	Illumina HiSeq2500 (PE + MP)[c]	ALLPATHS-LG	620,719	10,766	4,227,710	3,216
lycPyrPB	PacBio RSII C6-P4	Falcon	6,644,420	3,422	-	-
lycPyrSN1	10X Genomics Chromium HiSeqX	Supernova2	144,856	29,791	4,360,585	13,934
lycPyrSN2	10X Genomics Chromium HiSeqX	Supernova2	149,640	27,366	4,748,626	14,217
lycPyrHiC	PacBio + Phase Genomics Hi-C	Proximo	6,644,420	3,422	70,588,898	2,927
lycPyrILPB	lycPyrIL + gap-filling with PacBio	PBJelly	1,982,606	6,895	4,229,628	3,216
lycPyr2	PacBio + Dovetail CHiCAGO	HiRise	6,294,665	3,463	6,644,037	3,227
lycPyr3	lycPyr2 + 10X Genomics	ARCS + LINKS	6,294,665	3,463	8,009,555	3,121
lycPyr4	lycPyr3 + Phase Genomics Hi-C	Proximo	6,294,665	3,463	69,071,023	1,713
lycPyr5	lycPyr4 + manual curation with alignments + gap filling	PBJelly	7,540,011	3,269	74,173,823	1,700
lycPyr6	lycPyr5 + manual curation with Hi-C	Juicer	7,540,011	3,271	74,173,823	1,700

Peona, et al. (2019). . BioRxiv 2019.12.19.882399.



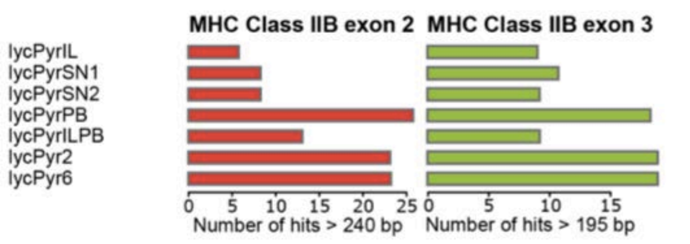


Errors that can happen in assemblies

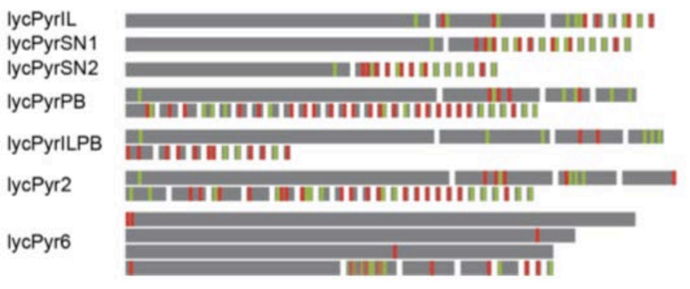
Denton et al. 2014 PLoS Comp Bio.

MHC IIB: complex tandem repeats as a case study

High diversity Low diversity



Highly diverse copies are very difficult to assemble



All loci are from same chromosome

Very challenging to place them accurately

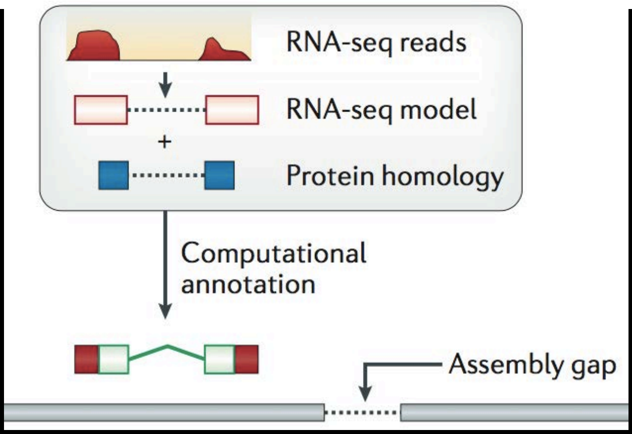
Post-genomics challenge

"What we can measure is by definition uninteresting and what we are interested in is by definition immeasurable"
- Lewontin 1974

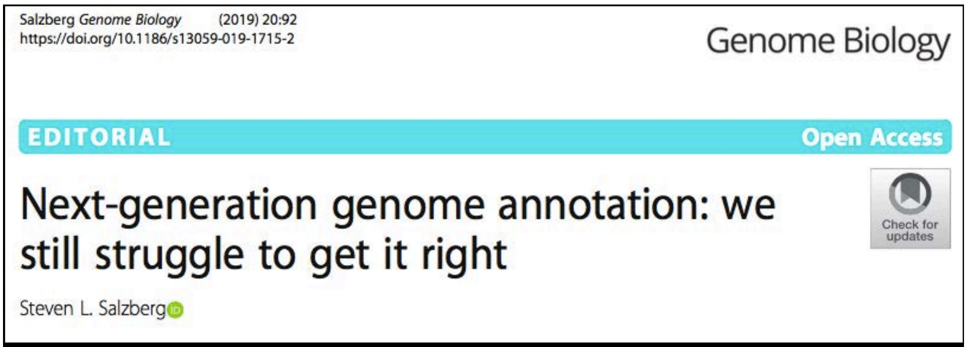
"What we understand of the genome is by definition uninteresting and what we are interested in is by definition very damn difficult to sequence and assemble and annotate and analyze at the genomic scale"
- Wheat 2015

For example:
- indels & inversions & repeats
- gene family dynamics
- evolutionary divergence

Genome annotation



- Using RNAseq and protein alignments to identify gene regions and exon boundaries



Salzberg *Genome Biology* (2019) 20:92
https://doi.org/10.1186/s13059-019-1715-2

Genome Biology

EDITORIAL Open Access

Next-generation genome annotation: we still struggle to get it right

Steven L. Salzberg

- Bacterial genomes, ~90% of genome is genic content
- Eukaryotes, << 2% is genic
  - – Gene prediction is very difficult, low accuracy

“even after 18 years of effort, the precise exon–intron structure of many human protein-coding genes is not settled”

How well does annotation work?

- Hard to say, no recent comparisons among methods
- Primarily depends upon the training dataset you use

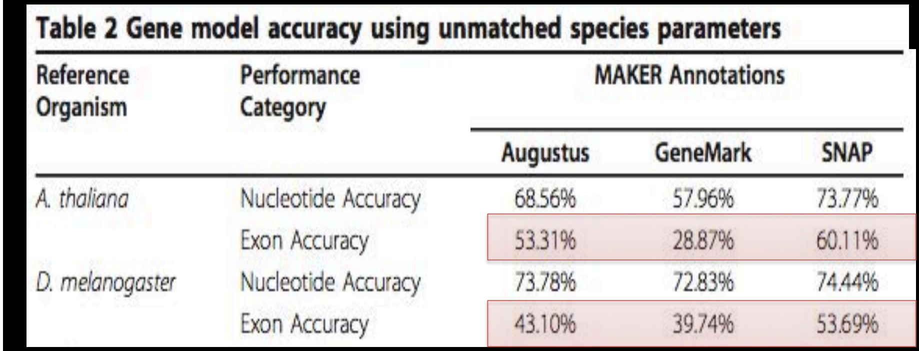


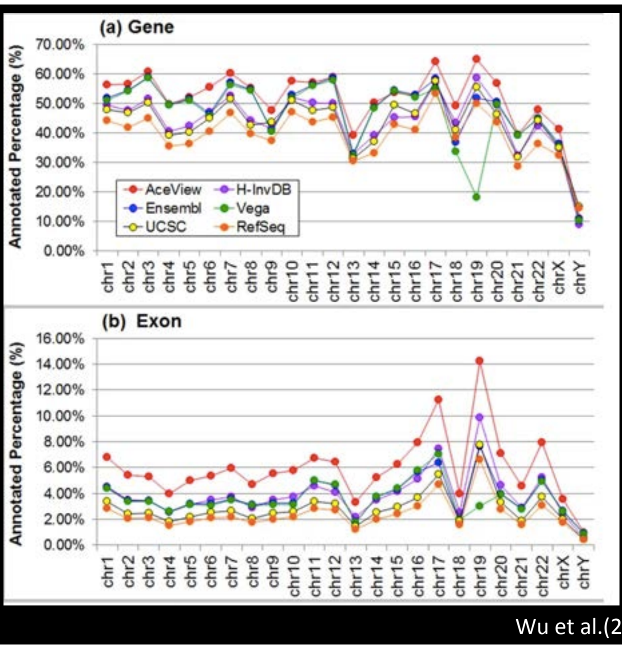
Table 2 Gene model accuracy using unmatched species parameters

Reference Organism	Performance Category	MAKER Annotations		
		Augustus	GeneMark	SNAP
A. thaliana	Nucleotide Accuracy	68.56%	57.96%	73.77%
	Exon Accuracy	53.31%	28.87%	60.11%
D. melanogaster	Nucleotide Accuracy	73.78%	72.83%	74.44%
	Exon Accuracy	43.10%	39.74%	53.69%

Holt and Yandell (2011). MAKER2: BMC Bioinf.

Variation among annotation datasets

Annotated % per chromosome by annotation



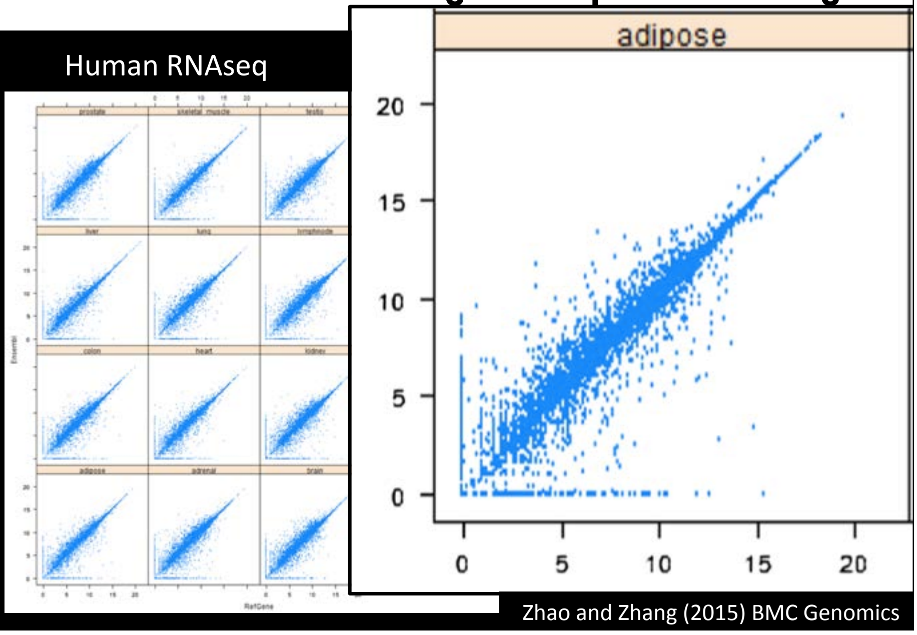
Wu et al.(2013). BMC Bioinformatics

Annotation choice affects gene expression insights

- Is mapping to the genome better than the assembled transcriptome?
- Biases
  - Genome assembly
    - might lack your gene of interest
    - Annotation for your genes might be bad
  - Transcriptome assembly
    - Will have all expressed genes
    - Assembly might have problems (fragments, duplicates, isoforms)

Annotation choice affects gene expression insights

Human RNAseq



Zhao and Zhang (2015) BMC Genomics

So, annotations matter … how do we get a good annotation?

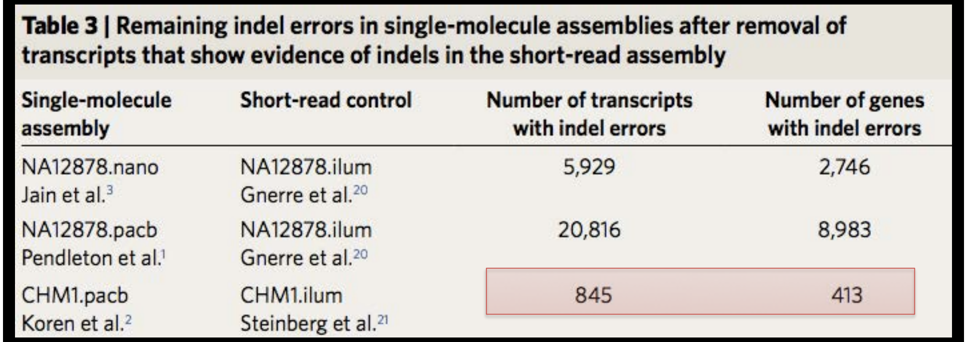


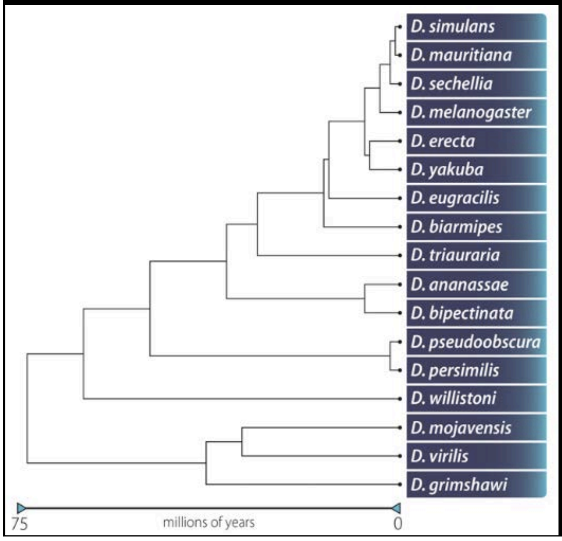
Table 3 | Remaining indel errors in single-molecule assemblies after removal of transcripts that show evidence of indels in the short-read assembly

Single-molecule assembly	Short-read control	Number of transcripts with indel errors	Number of genes with indel errors
NA12878.nano Jain et al.[3]	NA12878.ilum Gnerre et al.[20]	5,929	2,746
NA12878.pacb Pendleton et al.[1]	NA12878.ilum Gnerre et al.[20]	20,816	8,983
CHM1.pacb Koren et al.[2]	CHM1.ilum Steinberg et al.[21]	845	413

the group generated 142-fold coverage and used two rounds of Quiver polishing

Watson and Warr 2019 Nat. Biotech.

15 Drosophila genomes via Oxford Nanopore



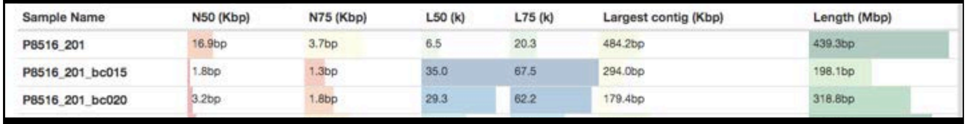
average of 29x coverage

average contig N50 of 4.4 Mb

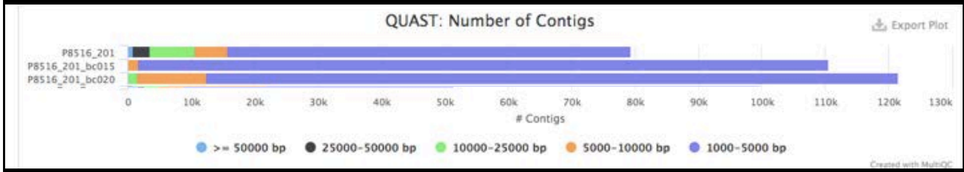
Miller et al. 2018 G3

Genome assembly assessment: metrics vs. biometrics

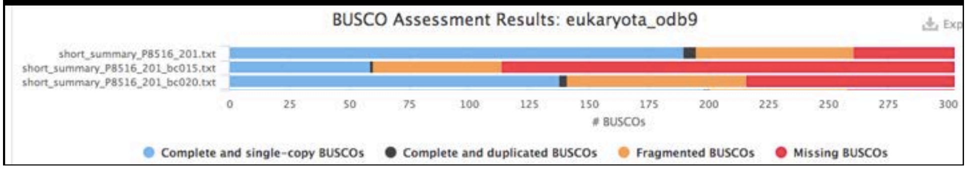
- Length and contiguity
- Gene content: number, completeness, fragmentation



Sample Name	N50 (Kbp)	N75 (Kbp)	L50 (k)	L75 (k)	Largest contig (Kbp)	Length (Mbp)
P8516_201	16.9bp	3.7bp	6.5	20.3	484.2bp	439.3bp
P8516_201_bc015	1.8bp	1.3bp	35.0	67.5	294.0bp	198.1bp
P8516_201_bc020	3.2bp	1.8bp	29.3	62.2	179.4bp	318.8bp

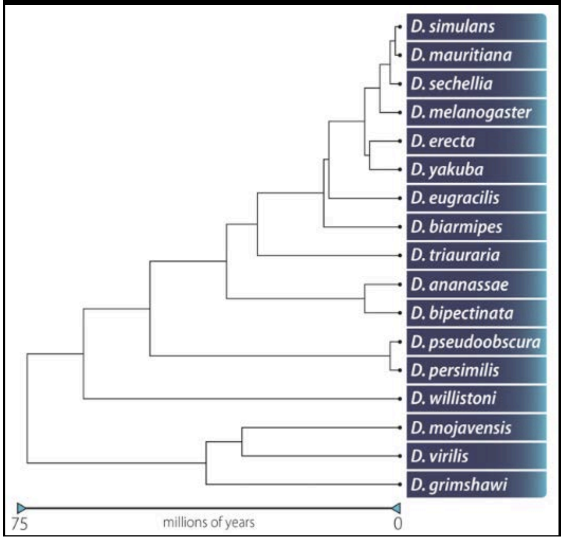


QUAST: Number of Contigs



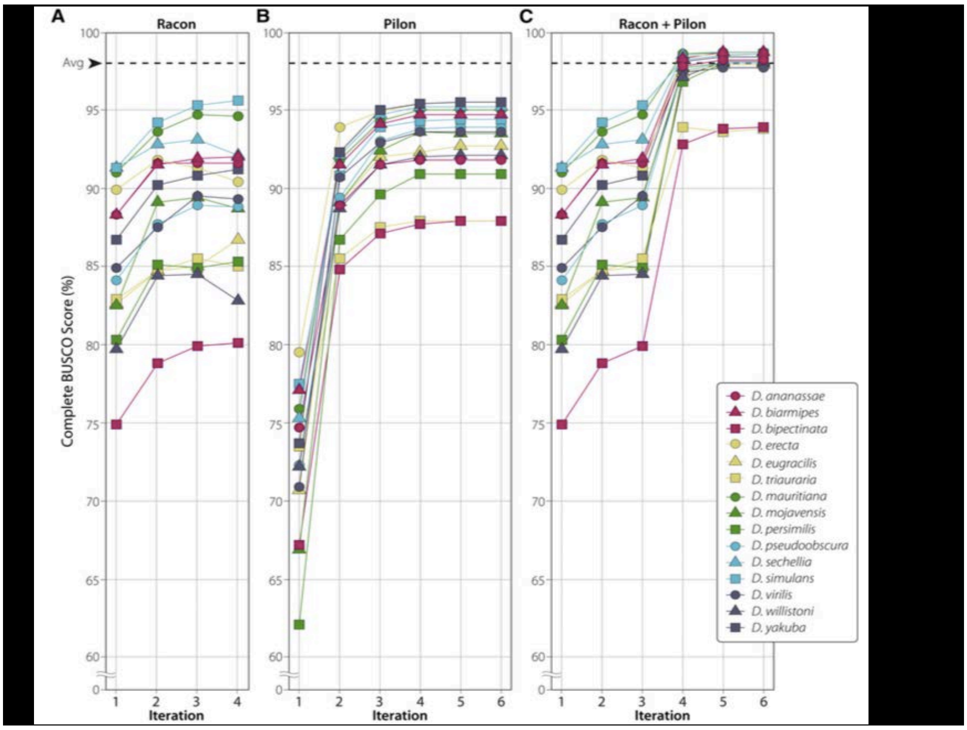
BUSCO Assessment Results: eukaryota_odb9

15 Drosophila genomes via Oxford Nanopore



So what did their BUSCO look like?

Miller et al. 2018 G3



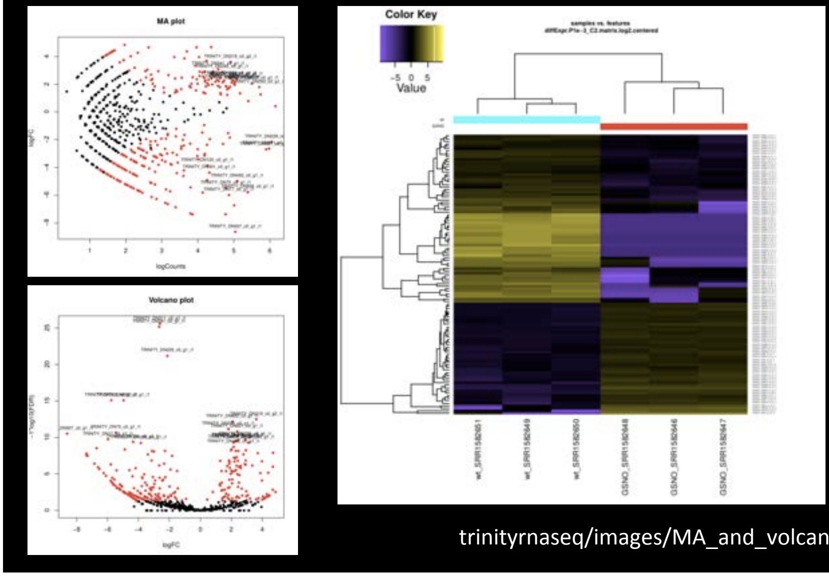
Want a nice genome? polish it ... a lot

RNAseq: ....

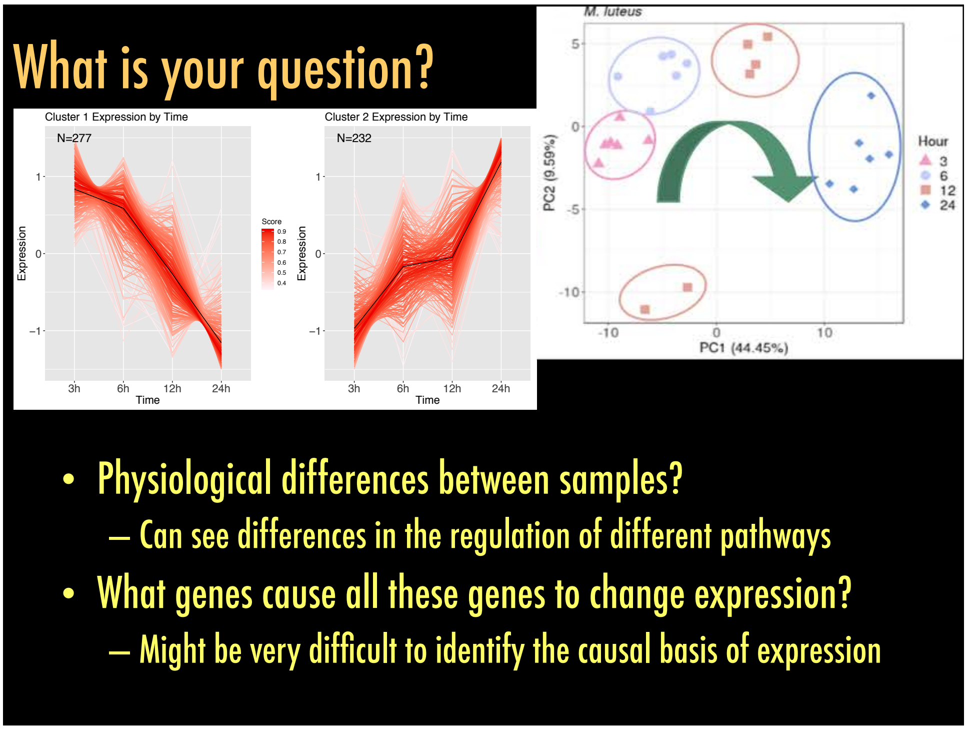
Are you measuring what you think you are measuring?

Why type of conclusions are you drawing?

What does a significant DE gene mean?

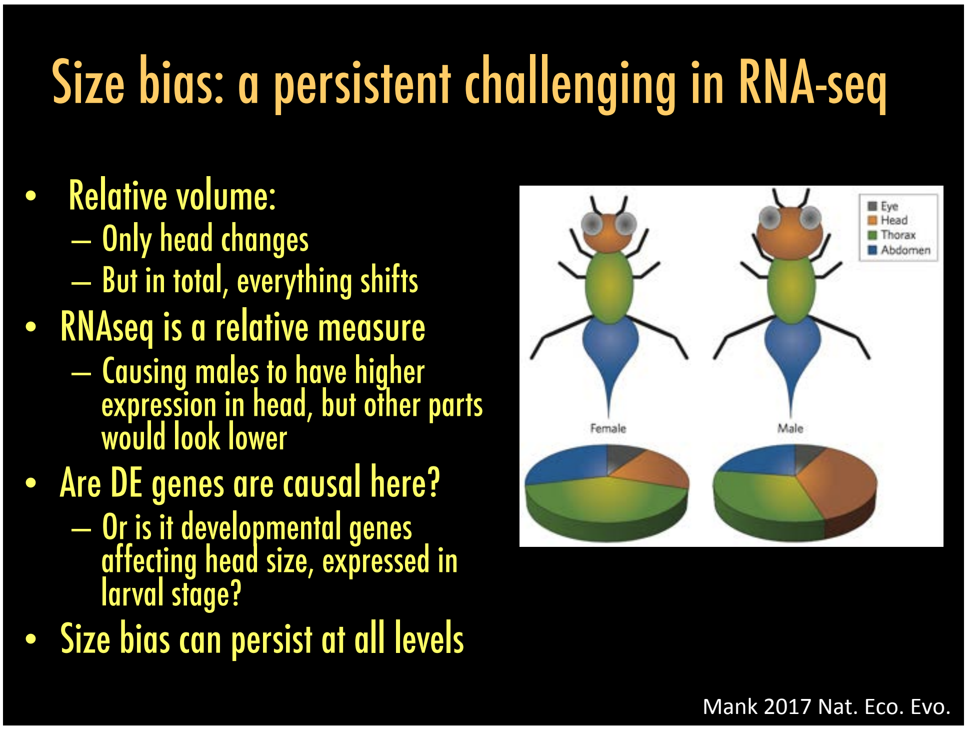


trinityrnaseq/images/MA_and_volcano_plot.png



What is your question?

- Physiological differences between samples?
  – Can see differences in the regulation of different pathways
- What genes cause all these genes to change expression?
  – Might be very difficult to identify the causal basis of expression



Size bias: a persistent challenging in RNA-seq

- Relative volume:
  – Only head changes
  – But in total, everything shifts
- RNAseq is a relative measure
  – Causing males to have higher expression in head, but other parts would look lower
- Are DE genes are causal here?
  – Or is it developmental genes affecting head size, expressed in larval stage?
- Size bias can persist at all levels

Mank 2017 Nat. Eco. Evo.

Temporal changes in cell types over time

Thus, if we observed DE in a tissue between time points, is this due to:

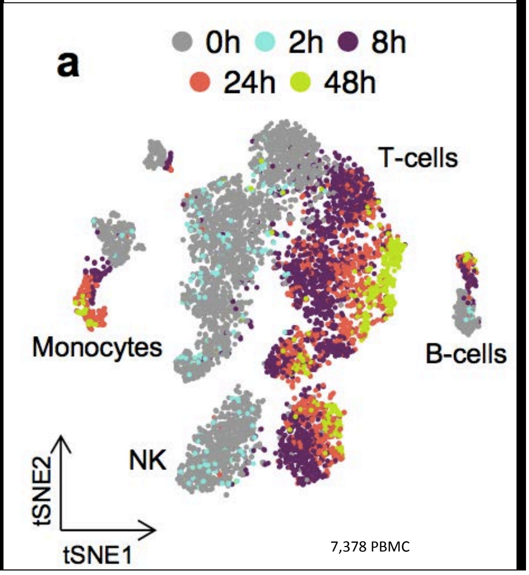
- Regulation of the DE genes in the same cells
- Differences in abundance of cell types due to proliferation

These can give the same results

- But have different causes

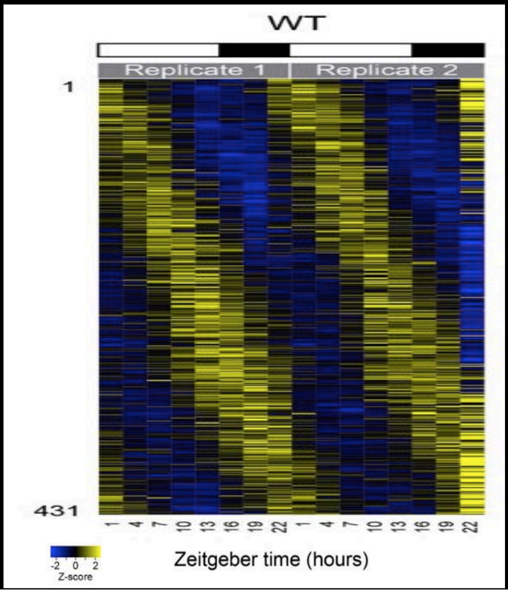
Massoni-Badosa et al. (2020) BioRxiv

Blood mononuclear cells from healthy donors



7,378 PBMC

When are you sampling?

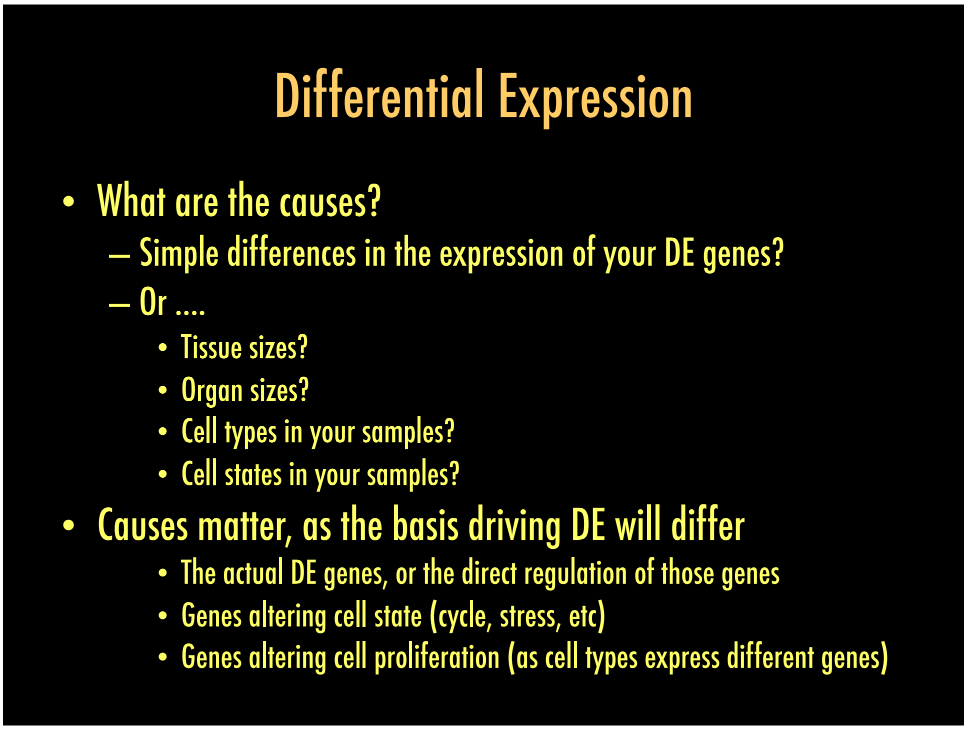


Butterfly brains collected every 3 hours

Time of day affects transcriptome

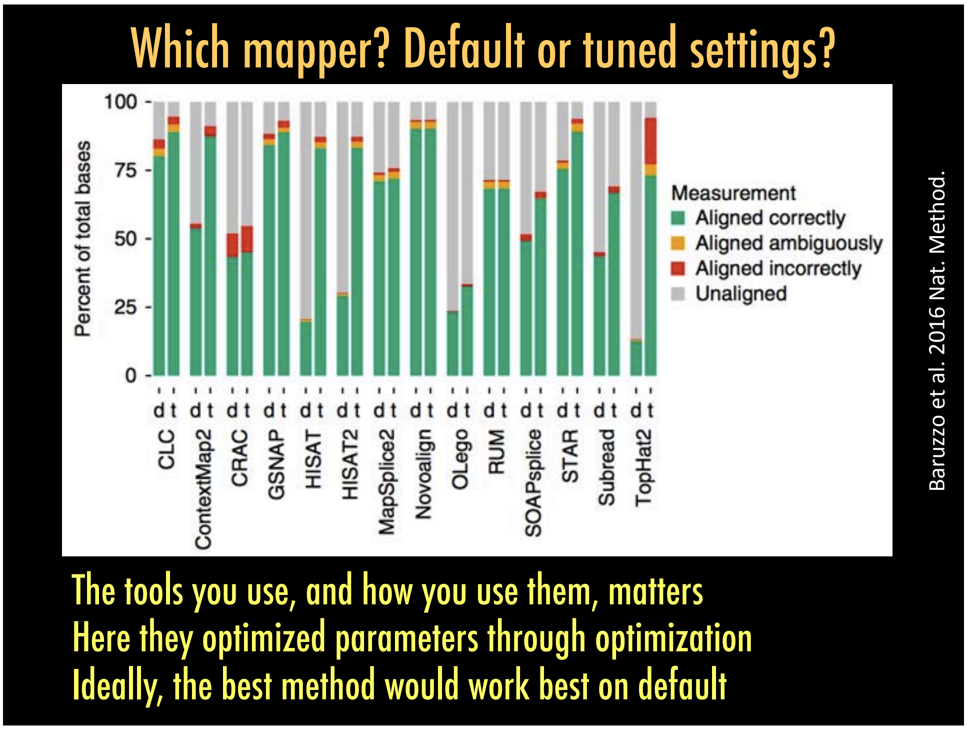
Standardize your time sampling

Lugena et al. 2019 PLOS Gen.



Differential Expression

- What are the causes?
  - Simple differences in the expression of your DE genes?
  - Or ….
    - Tissue sizes?
    - Organ sizes?
    - Cell types in your samples?
    - Cell states in your samples?
- Causes matter, as the basis driving DE will differ
    - The actual DE genes, or the direct regulation of those genes
    - Genes altering cell state (cycle, stress, etc)
    - Genes altering cell proliferation (as cell types express different genes)



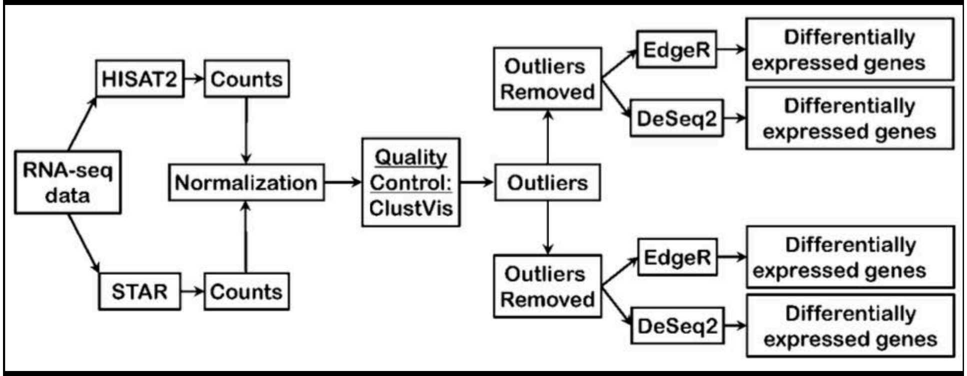
Which mapper? Default or tuned settings?

Baruzzo et al. 2016 Nat. Method.

The tools you use, and how you use them, matters
Here they optimized parameters through optimization
Ideally, the best method would work best on default

Clinical breast cancer datasets

Pipeline for bioinformatic comparisons



RNA-seq data

HISAT2 → Counts

STAR → Counts

Normalization → Quality Control: ClustVis → Outliers

Outliers Removed

EdgeR → Differentially expressed genes

DeSeq2 → Differentially expressed genes

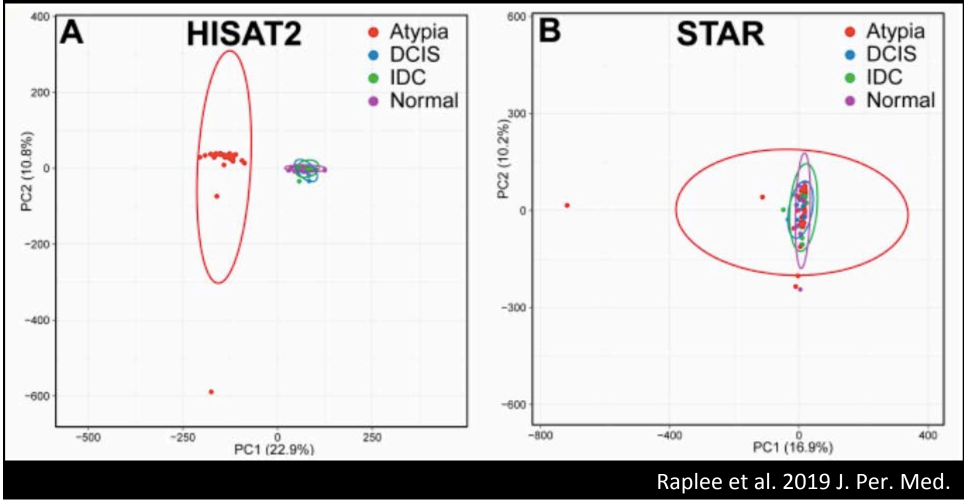
Outliers Removed

EdgeR → Differentially expressed genes

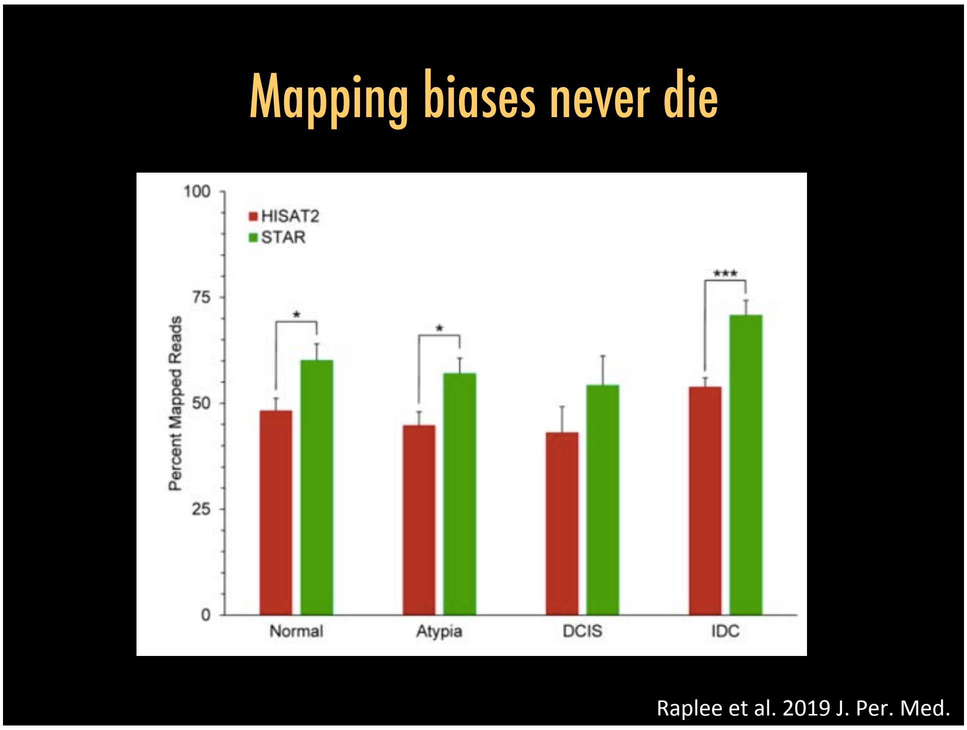
DeSeq2 → Differentially expressed genes

Raplee 2019 J. of Personalized Medicine

Mapper effects are real

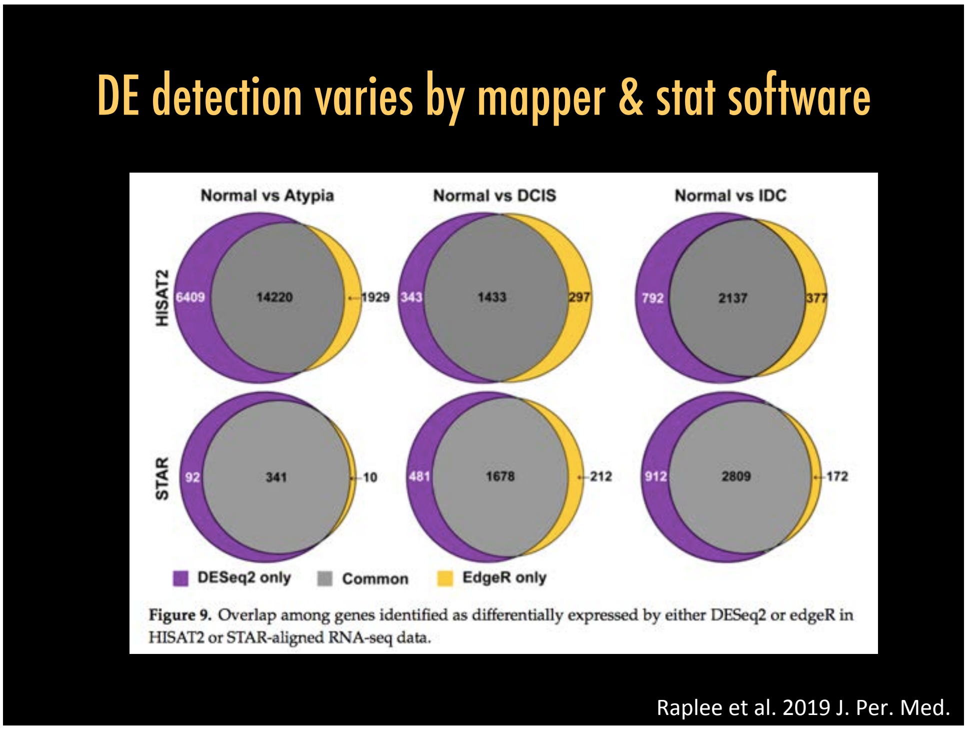


Raplee et al. 2019 J. Per. Med.



Mapping biases never die

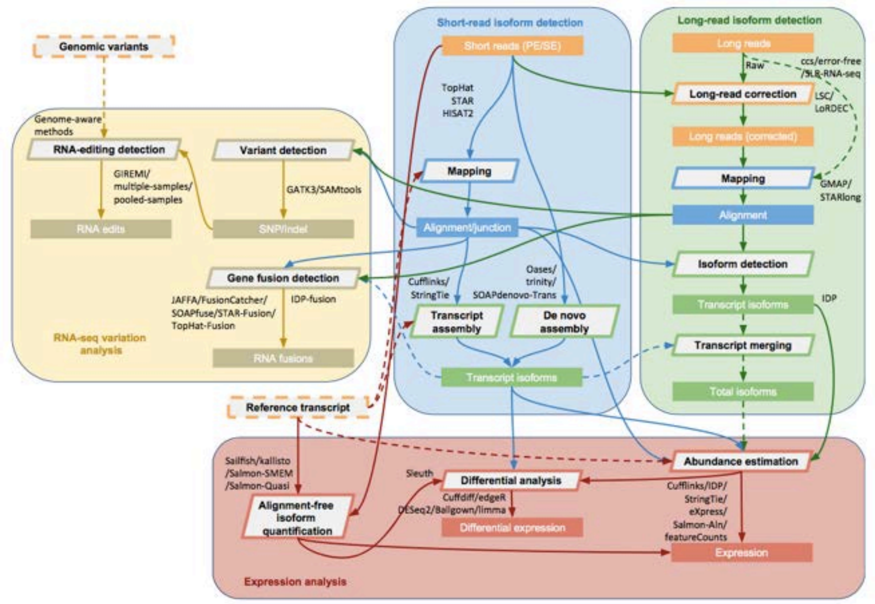
Raplee et al. 2019 J. Per. Med.



DE detection varies by mapper & stat software

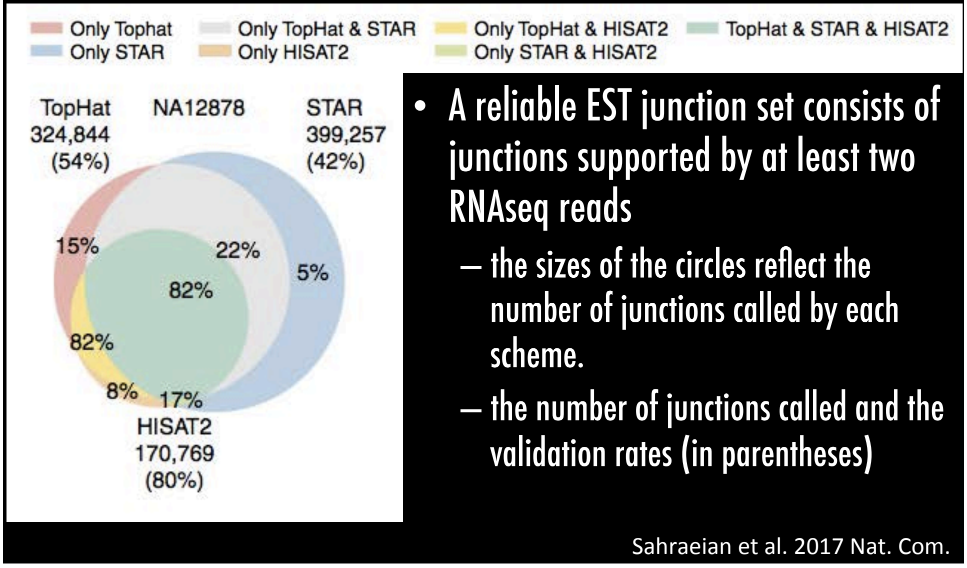
Figure 9. Overlap among genes identified as differentially expressed by either DESeq2 or edgeR in HISAT2 or STAR-aligned RNA-seq data.

Raplee et al. 2019 J. Per. Med.



Sahraeian et al. 2017 Nat. Com.

Detected exon splice junctions by different schemes



- A reliable EST junction set consists of junctions supported by at least two RNAseq reads
  - – the sizes of the circles reflect the number of junctions called by each scheme.
  - – the number of junctions called and the validation rates (in parentheses)

Sahraeian et al. 2017 Nat. Com.

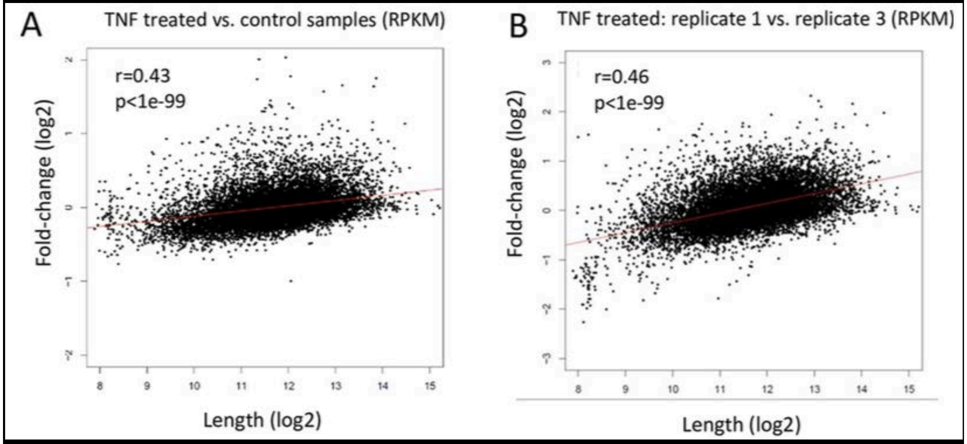
Persistence of bias in RNAseq studies: length

- Brian Haas already talked about how we standardize expression for gene length
  - FPKM

Reported as: Number of RNA-Seq **F**ragments **P**er **K**ilobase of transcript per total **M**illion fragments mapped

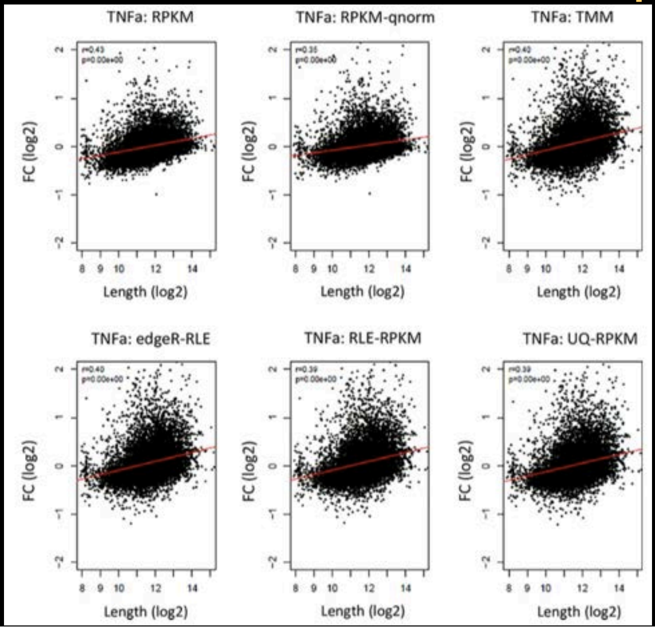
FPKM

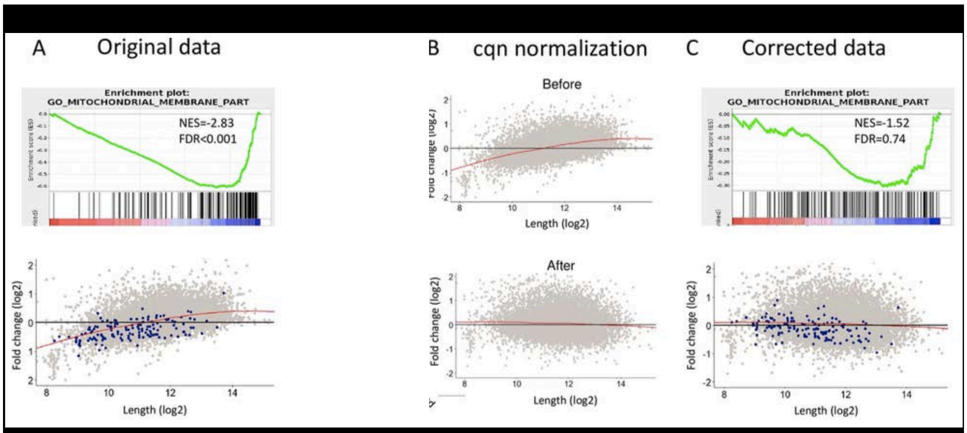
Persistence of bias in RNAseq studies: length

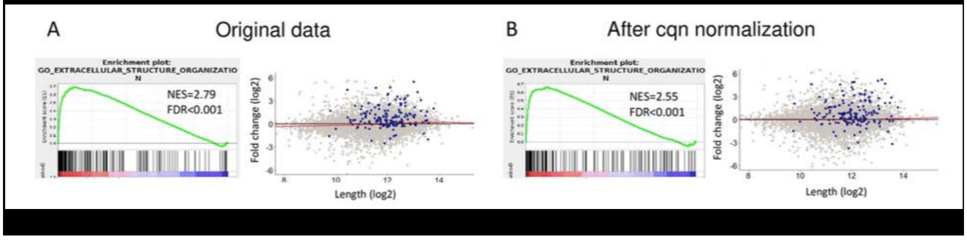


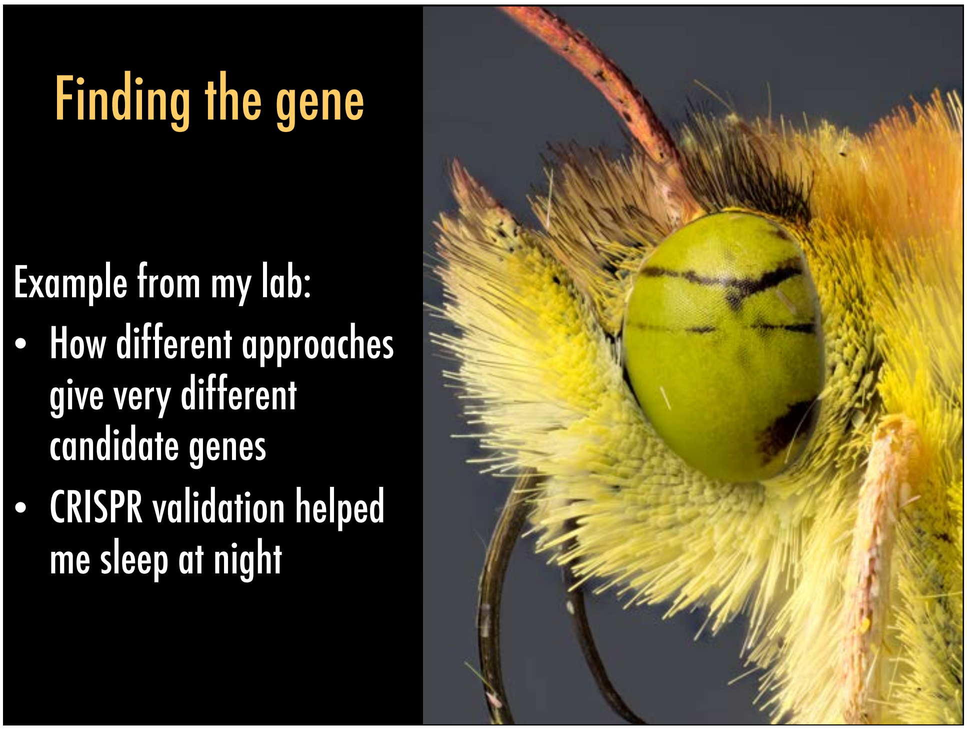
Mandelboum et al. 2019 PLOS Biology

Standard normalization doesn’t help





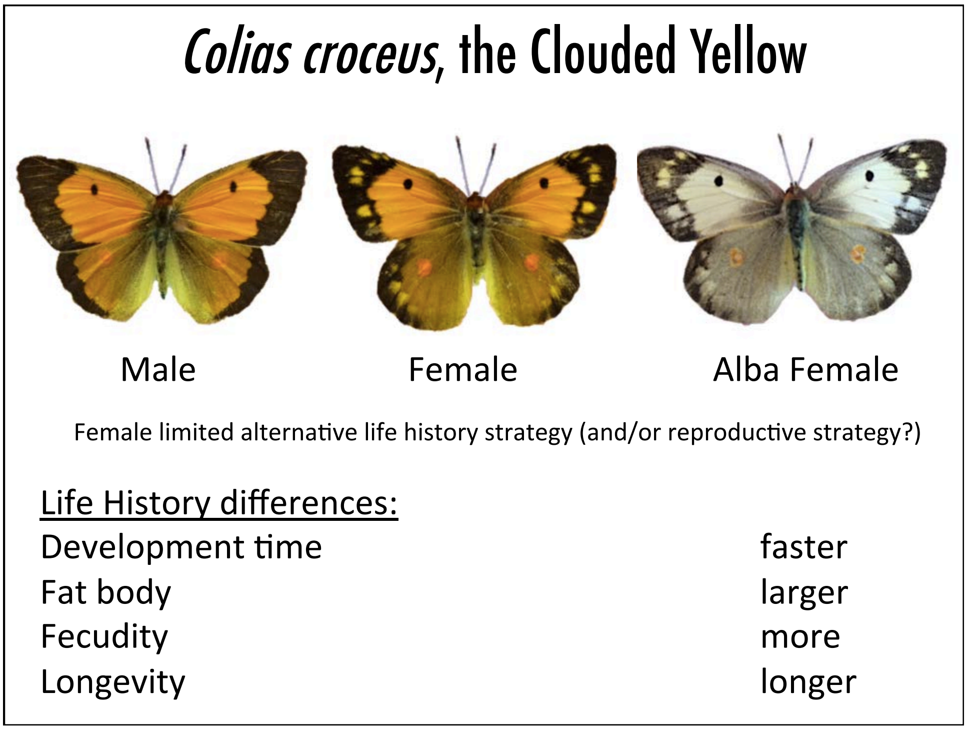




Finding the gene

Example from my lab:

- How different approaches give very different candidate genes
- CRISPR validation helped me sleep at night



Colias croceus, the Clouded Yellow

Male

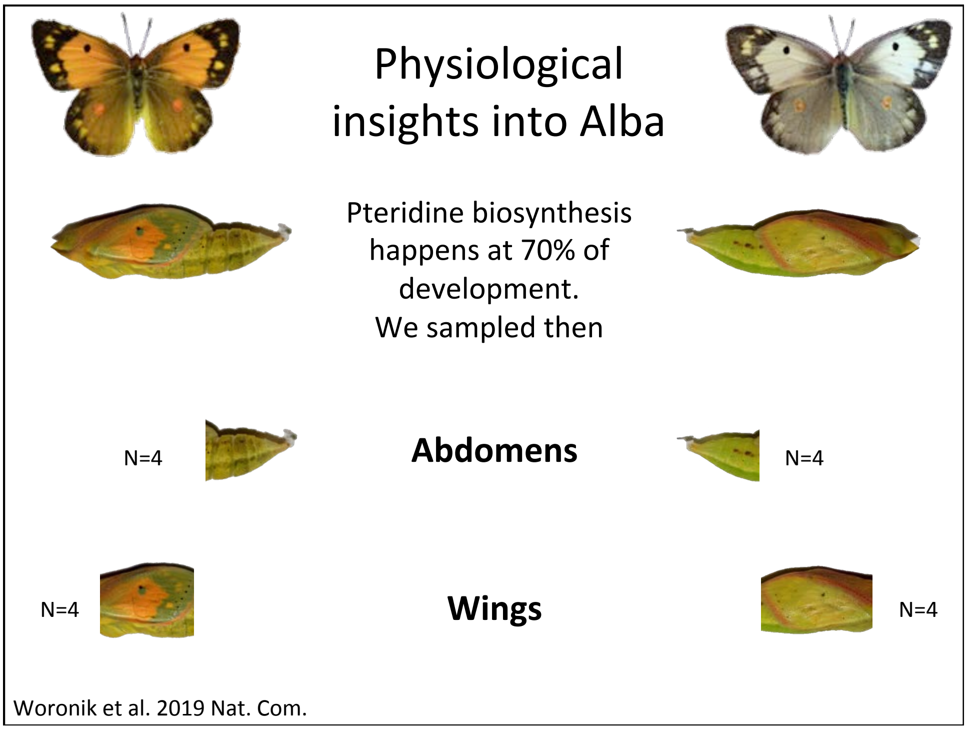
Female

Alba Female

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

Life History differences:

Development time	faster
Fat body	larger
Fecudity	more
Longevity	longer



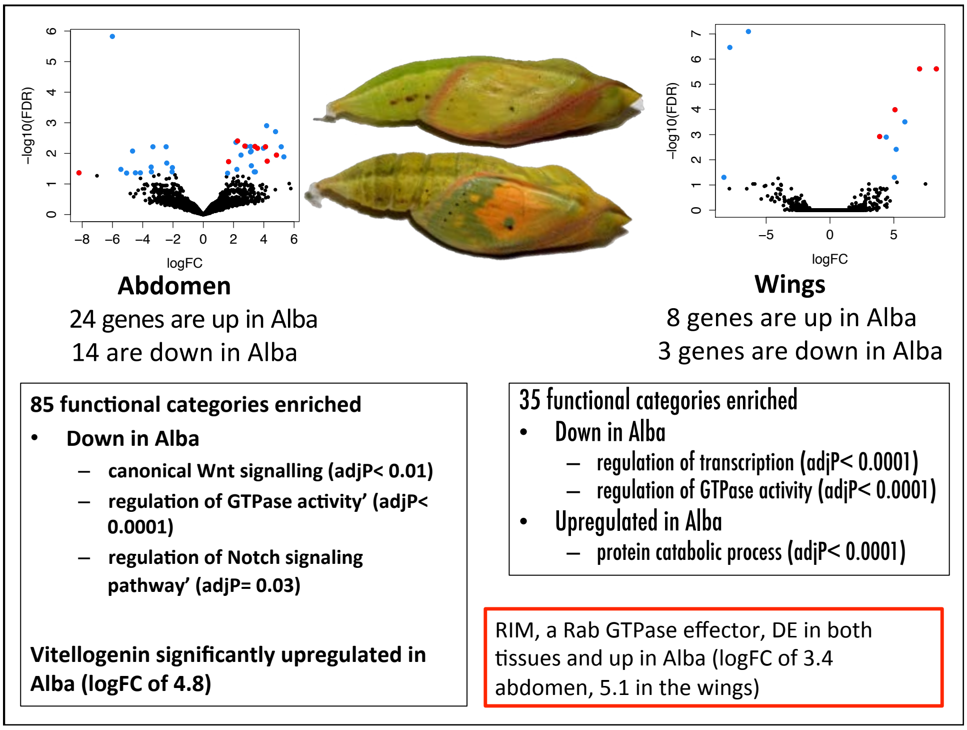
Physiological insights into Alba

Pteridine biosynthesis happens at 70% of development.
We sampled then

N=4 **Abdomens** N=4

N=4 **Wings** N=4

Woronik et al. 2019 Nat. Com.



Abdomen
24 genes are up in Alba
14 are down in Alba

Wings
8 genes are up in Alba
3 genes are down in Alba

85 functional categories enriched

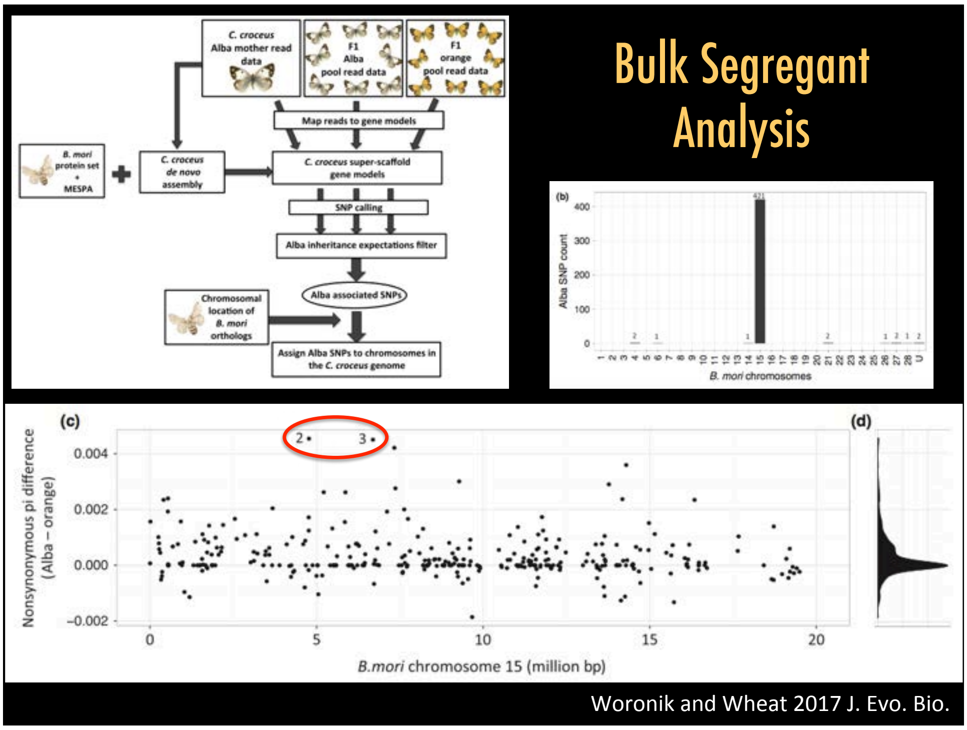
- **Down in Alba**
  - **canonical Wnt signalling (adjP< 0.01)**
  - **regulation of GTPase activity' (adjP< 0.0001)**
  - **regulation of Notch signaling pathway' (adjP= 0.03)**

Vitellogenin significantly upregulated in Alba (logFC of 4.8)

35 functional categories enriched

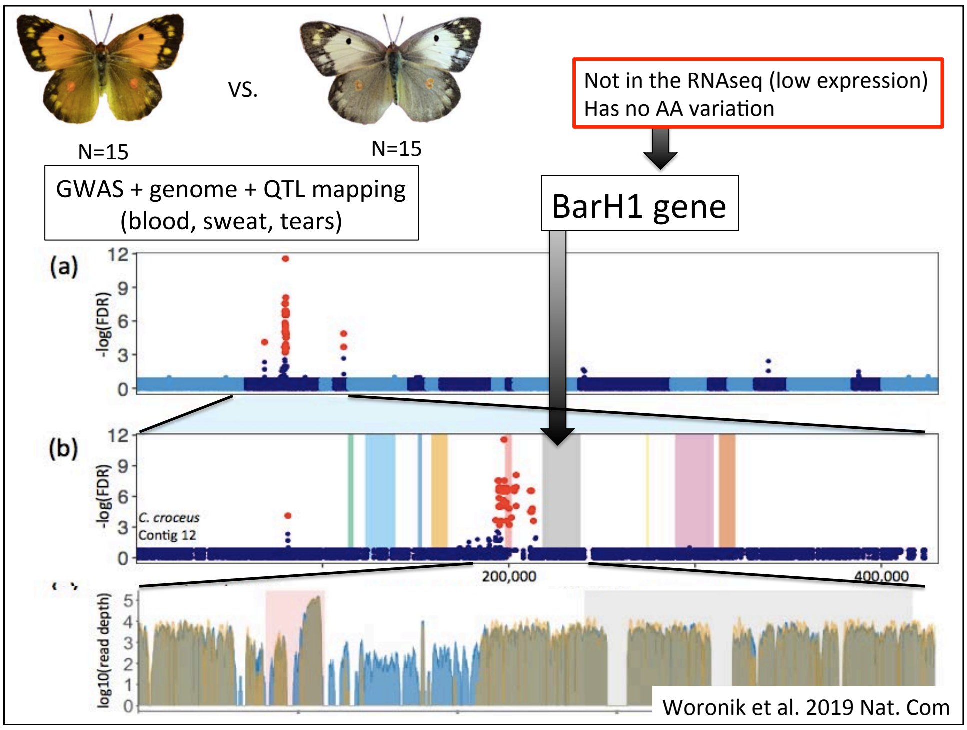
- Down in Alba
  - regulation of transcription (adjP< 0.0001)
  - regulation of GTPase activity (adjP< 0.0001)
- Upregulated in Alba
  - protein catabolic process (adjP< 0.0001)

RIM, a Rab GTPase effector, DE in both tissues and up in Alba (logFC of 3.4 abdomen, 5.1 in the wings)



Bulk Segregant Analysis

Woronik and Wheat 2017 J. Evo. Bio.



VS.

N=15

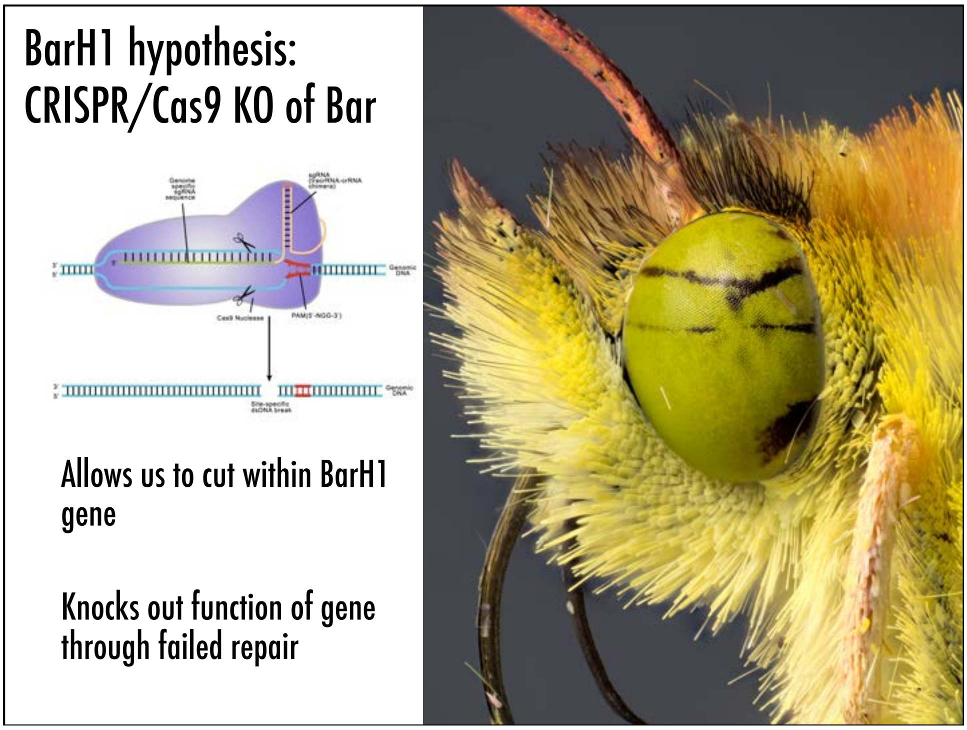
N=15

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

Not in the RNAseq (low expression)
Has no AA variation

BarH1 gene

Woronik et al. 2019 Nat. Com

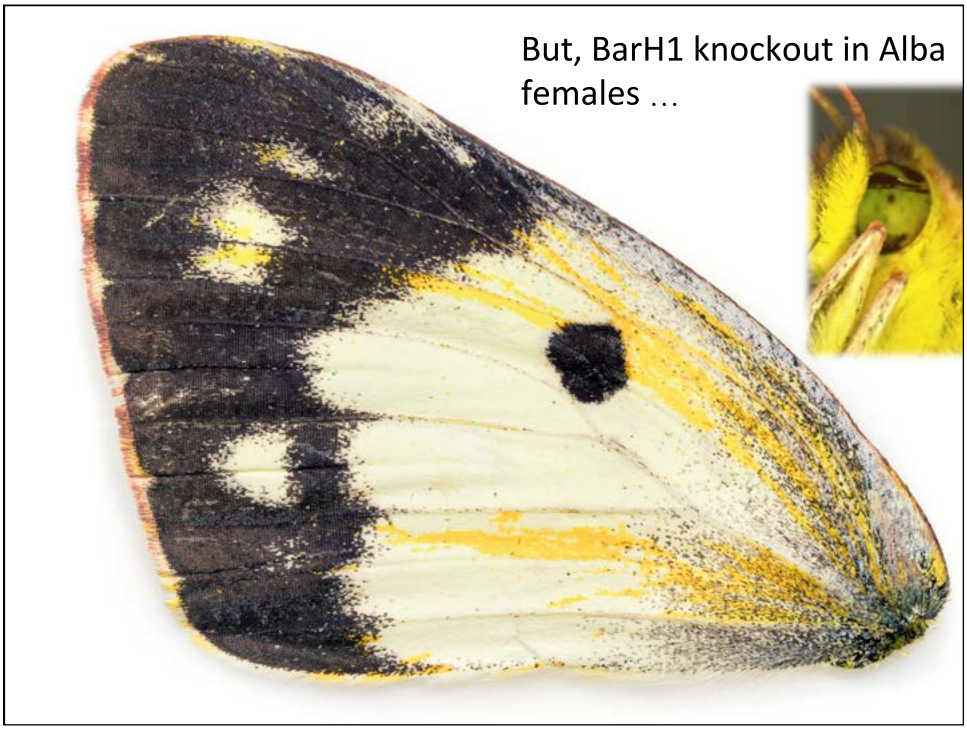


BarH1 hypothesis: CRISPR/Cas9 KO of Bar

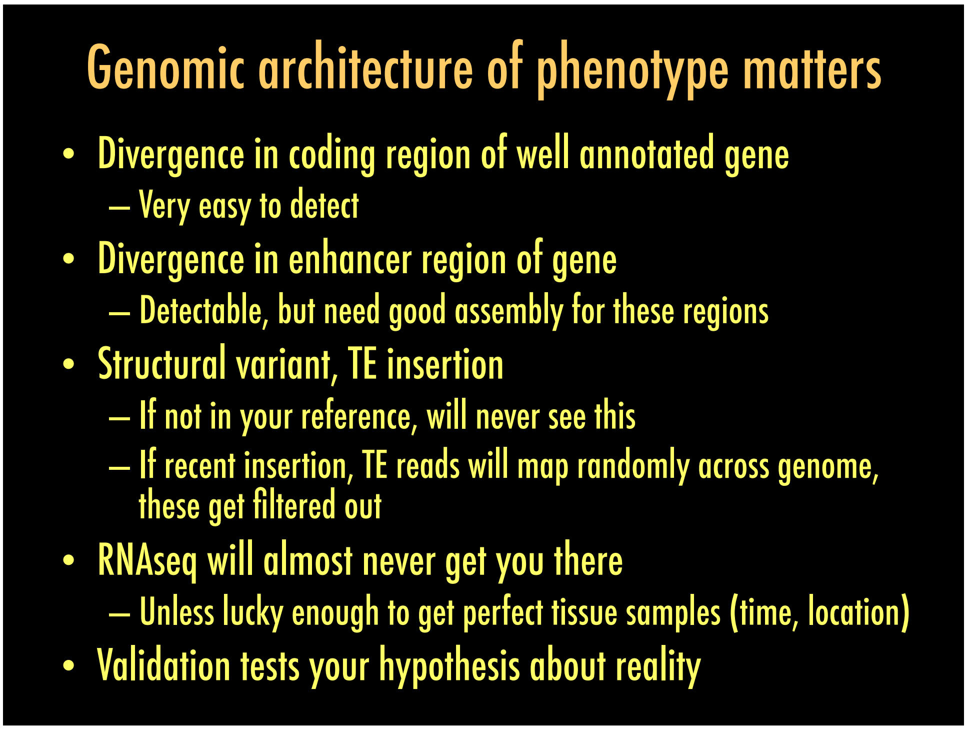
Genome specific sgRNA sequence
sgRNA (TracrRNA-crRNA chimera)
3'
5'
Genomic DNA
Cas9 Nuclease
PAM(5'-NGG-3')
3'
5'
Genomic DNA
Site-specific dsDNA break

Allows us to cut within BarH1 gene

Knocks out function of gene through failed repair

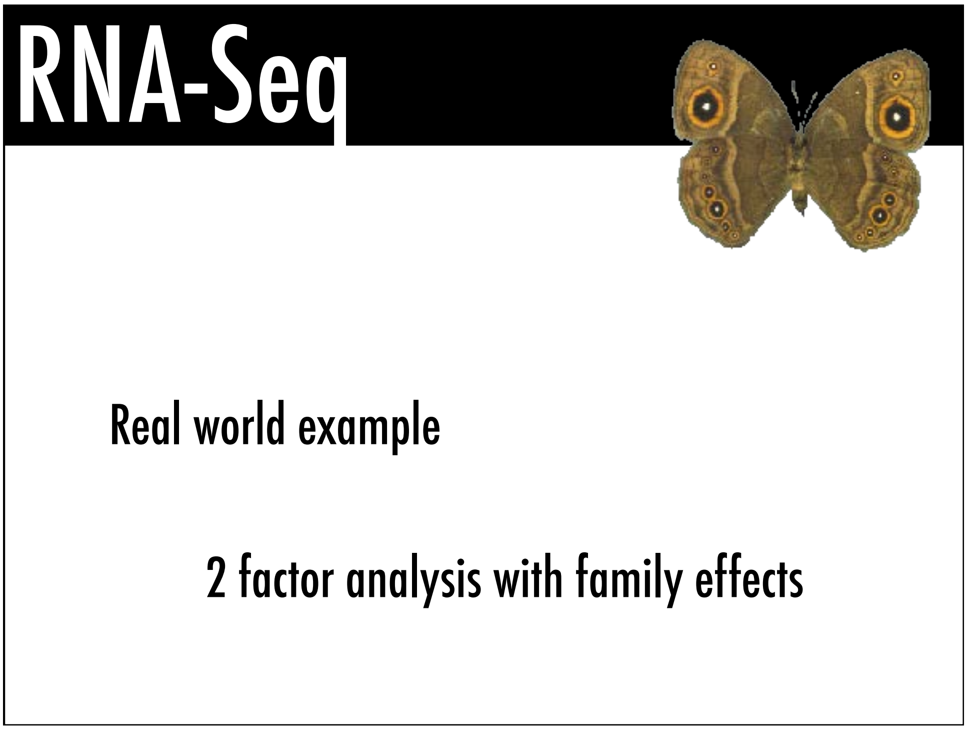


But, BarH1 knockout in Alba females …



Genomic architecture of phenotype matters

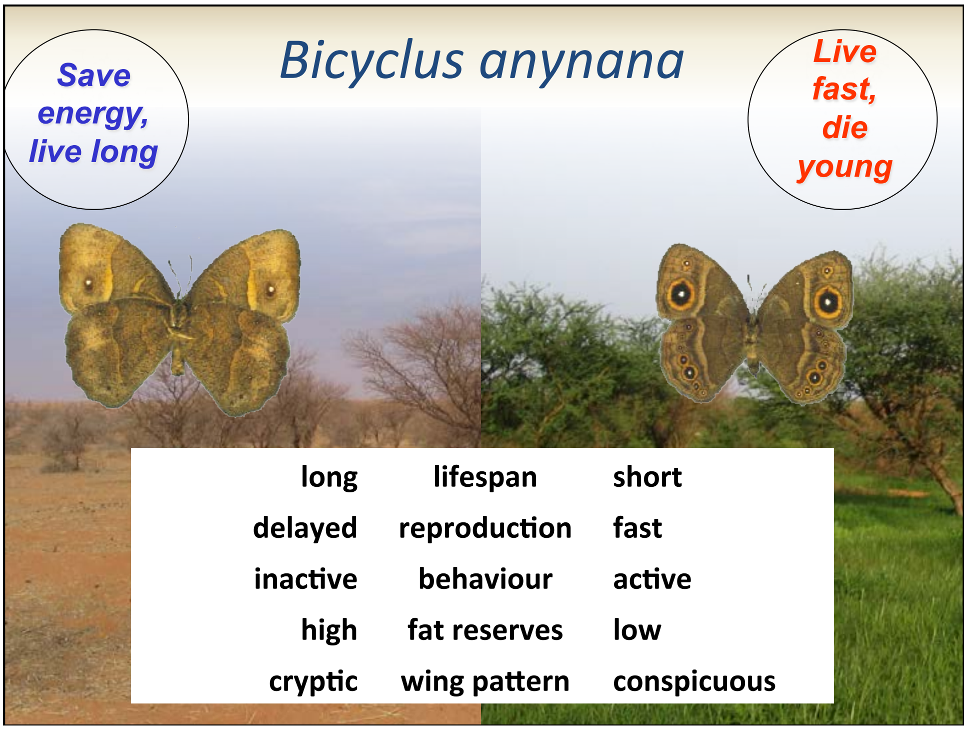
- Divergence in coding region of well annotated gene
  - Very easy to detect
- Divergence in enhancer region of gene
  - Detectable, but need good assembly for these regions
- Structural variant, TE insertion
  - If not in your reference, will never see this
  - If recent insertion, TE reads will map randomly across genome, these get filtered out
- RNAseq will almost never get you there
  - Unless lucky enough to get perfect tissue samples (time, location)
- Validation tests your hypothesis about reality



RNA-Seq

Real world example

2 factor analysis with family effects

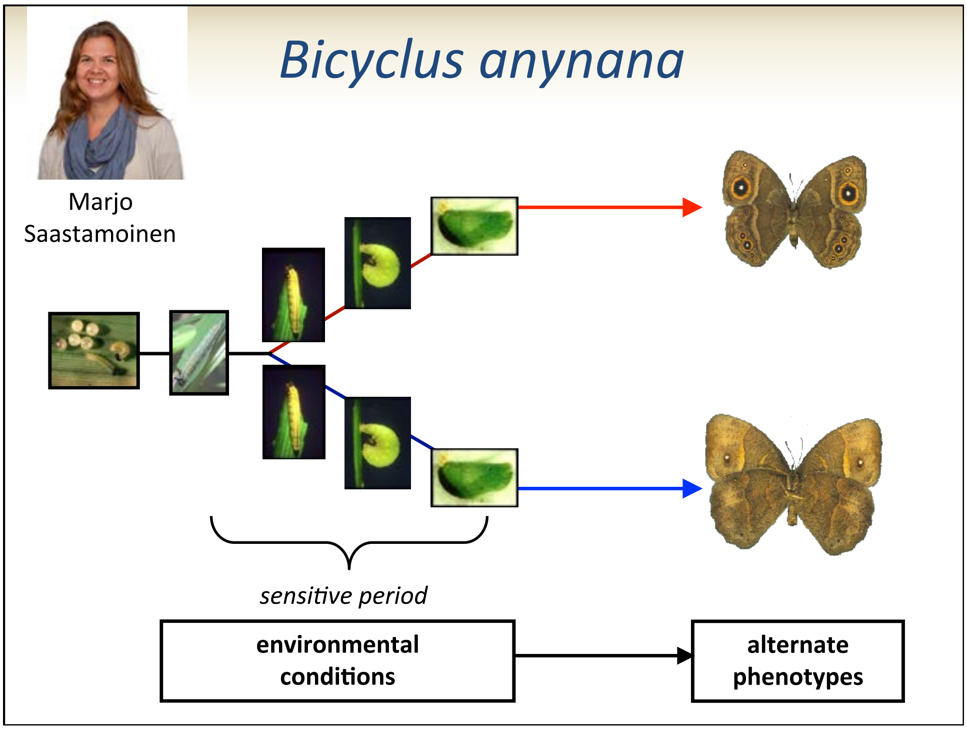


Bicyclus anynana

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



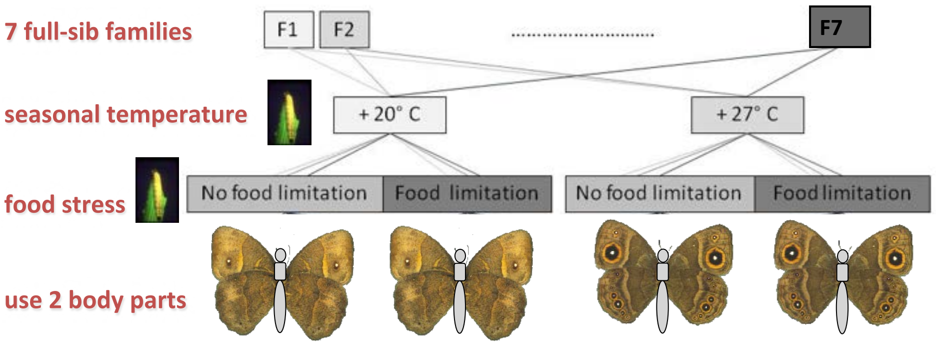
Bicyclus anynana

Marjo Saastamoinen

sensitive period

environmental conditions → alternate phenotypes

Experimental design



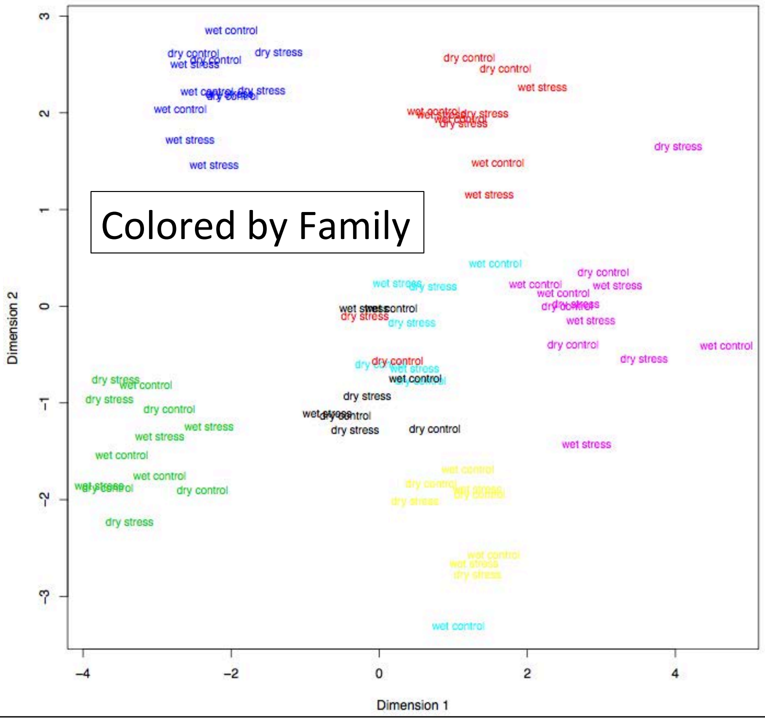
7 full-sib families

seasonal temperature

food stress

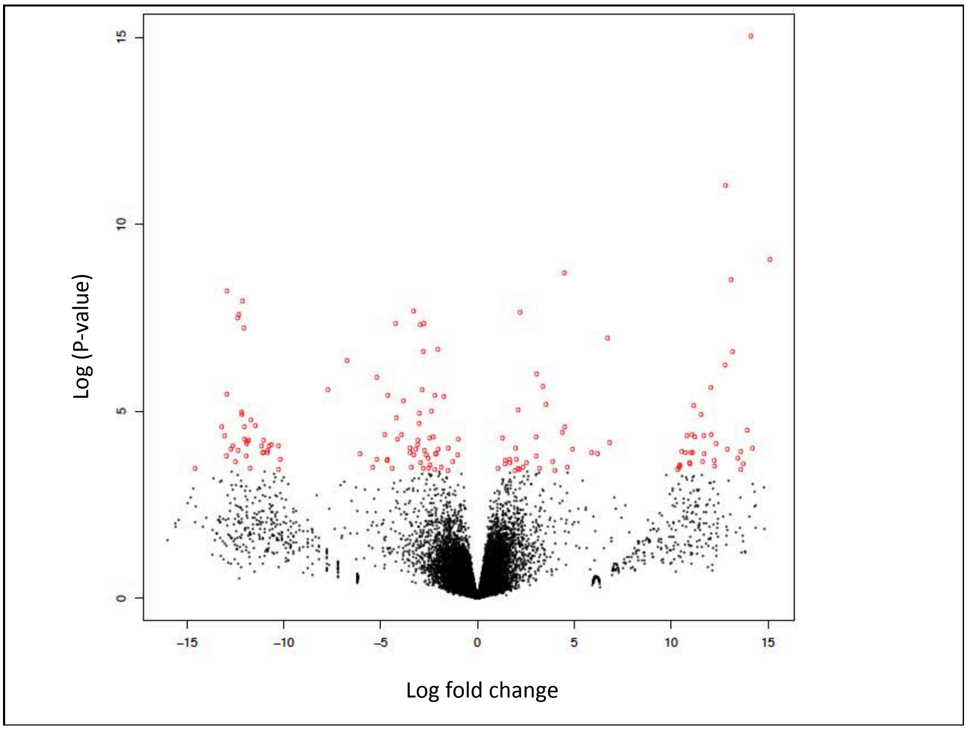
use 2 body parts

- 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



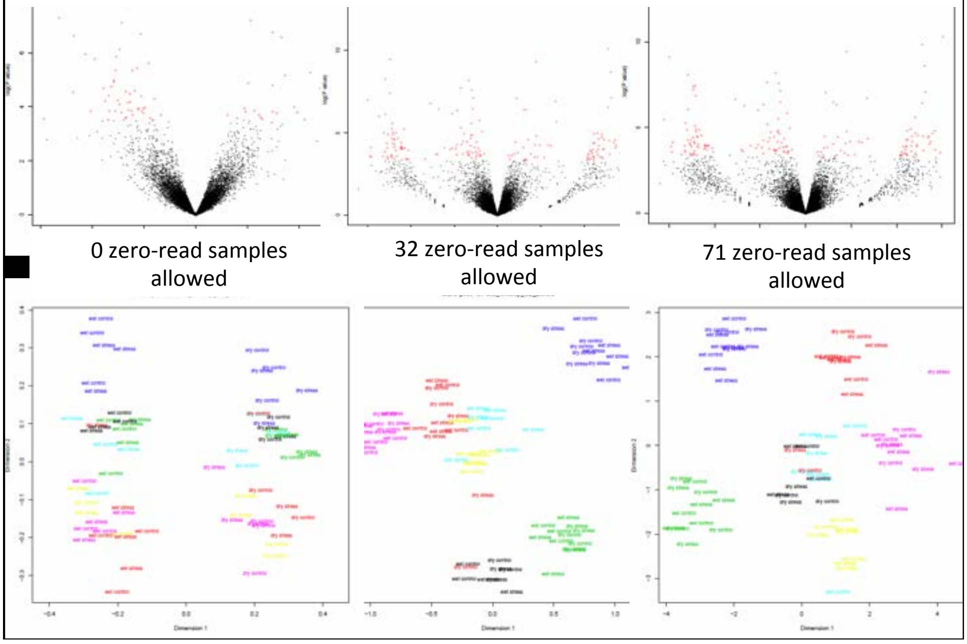
Colored by Family

78


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

Log (P-value)
Log fold change

Effect of filtering read mapping



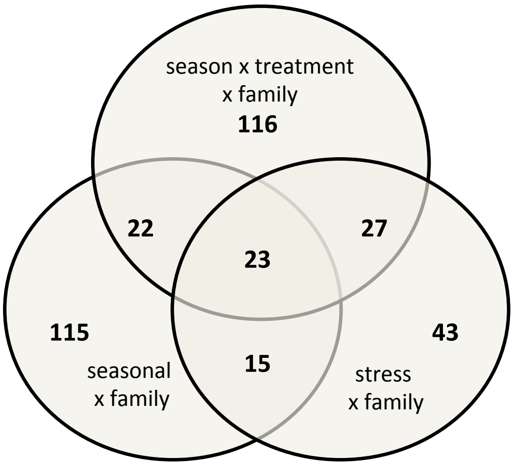
0 zero-read samples allowed

32 zero-read samples allowed

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

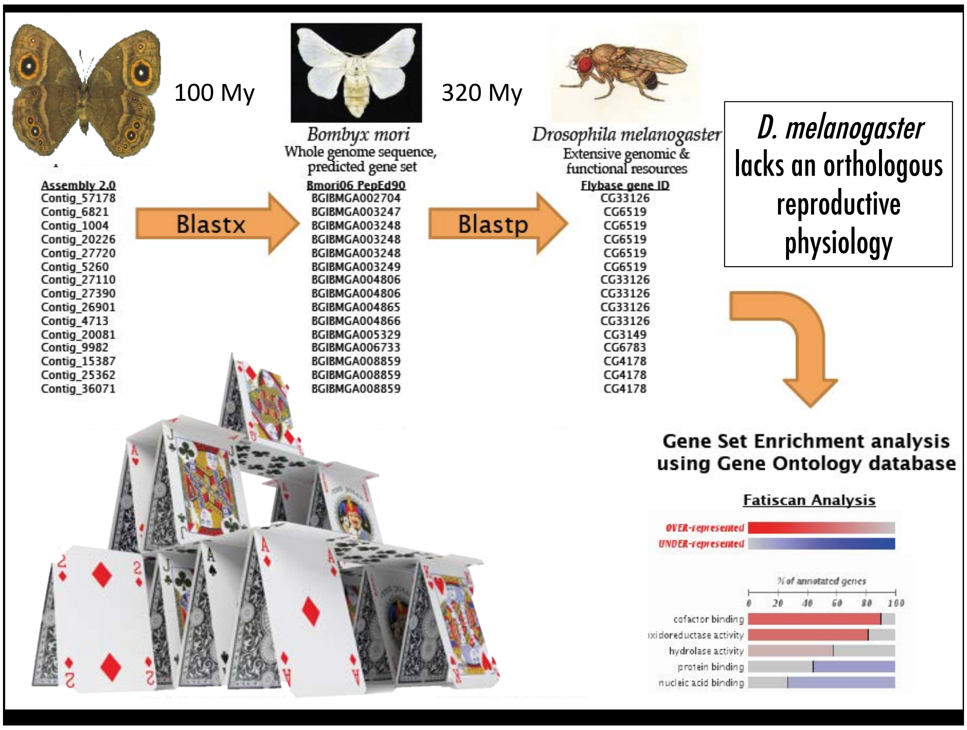
15

43

stress x family

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

Oostra et al. 2018 Nat. Com



100 My

320 My

Bombyx mori
Whole genome sequence, predicted gene set

Drosophila melanogaster
Extensive genomic & functional resources

Assembly 2.0		Bmori06 PepEd90		Flybase gene ID
Contig_57178	Blastx	BGIBMGA002704	Blastp	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

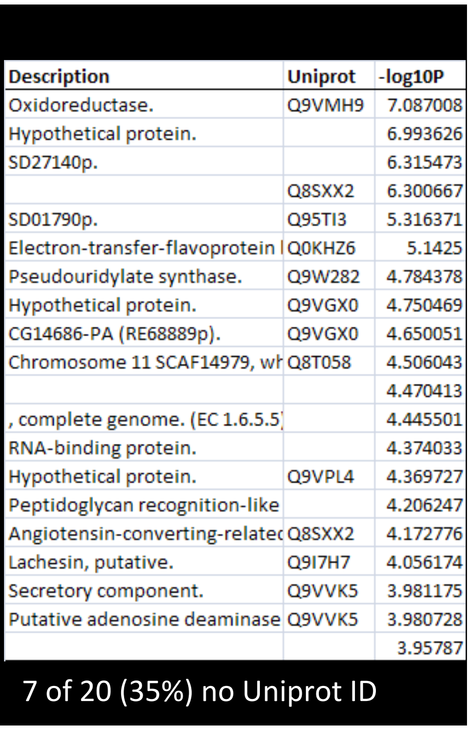
D. melanogaster lacks an orthologous reproductive physiology

Gene Set Enrichment analysis using Gene Ontology database

Fatiscan Analysis

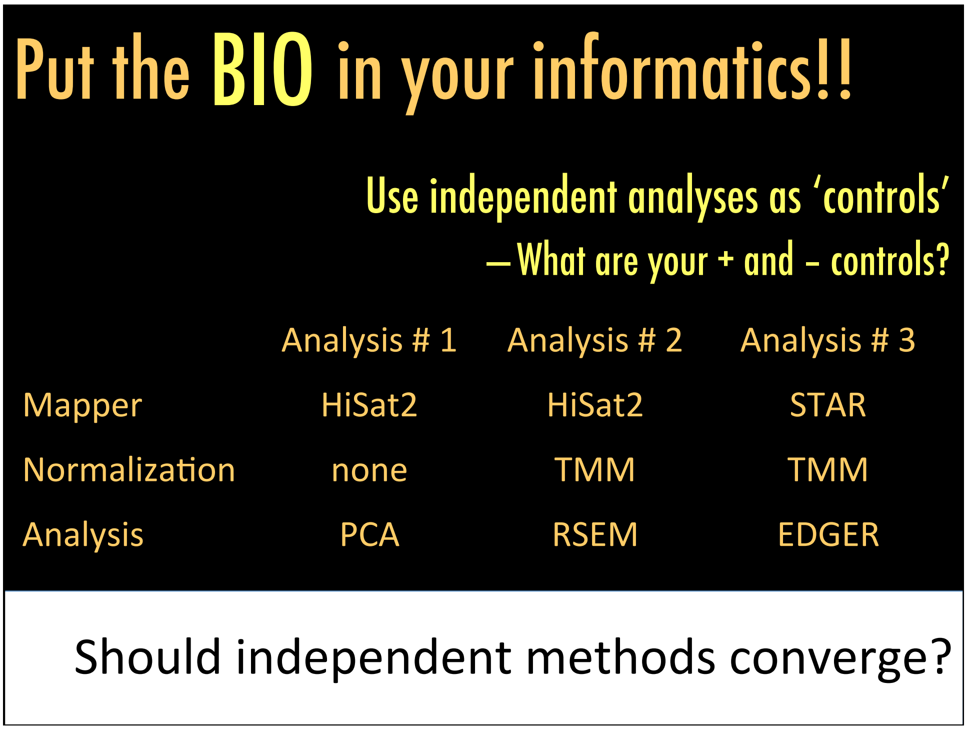
Most studies are annotation limited

- 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	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



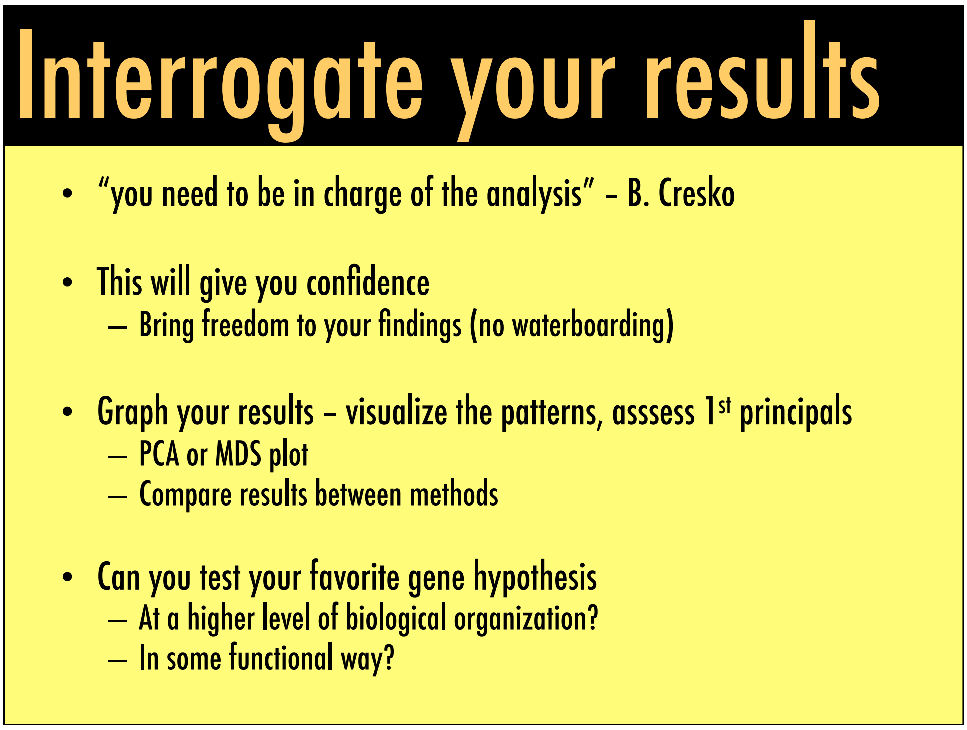
Put the BIO in your informatics!!

Use independent analyses as 'controls'

– What are your + and - controls?

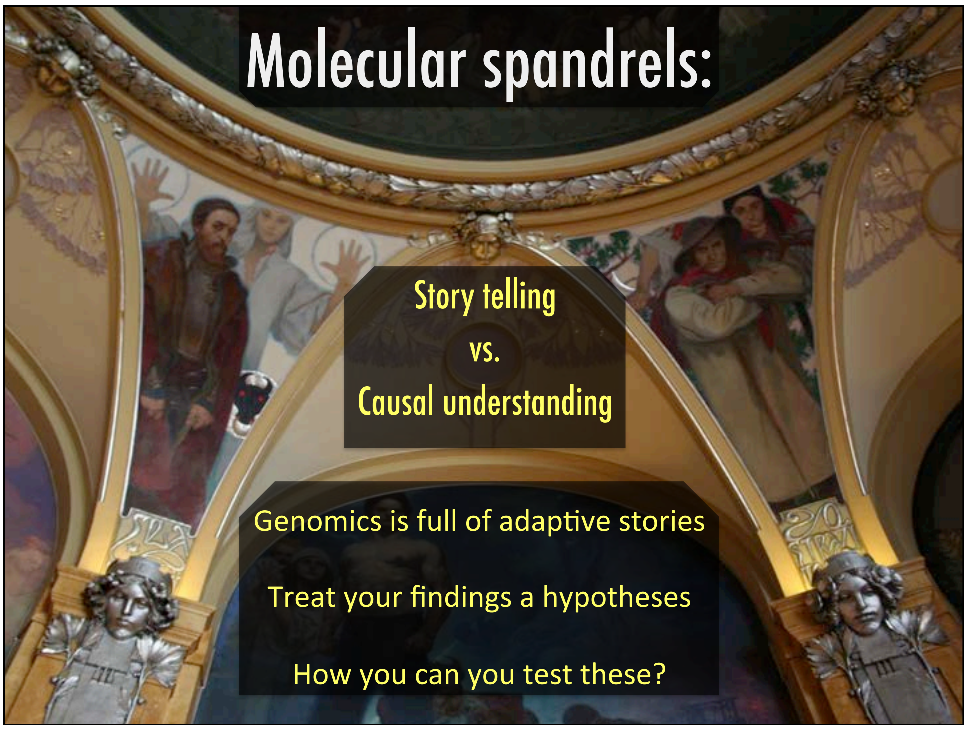
	Analysis # 1	Analysis # 2	Analysis # 3
Mapper	HiSat2	HiSat2	STAR
Normalization	none	TMM	TMM
Analysis	PCA	RSEM	EDGER

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, asssess 1st principals
  - PCA or MDS plot
  - Compare results between methods
- Can you test your favorite gene hypothesis
  - At a higher level of biological organization?
  - In some functional way?



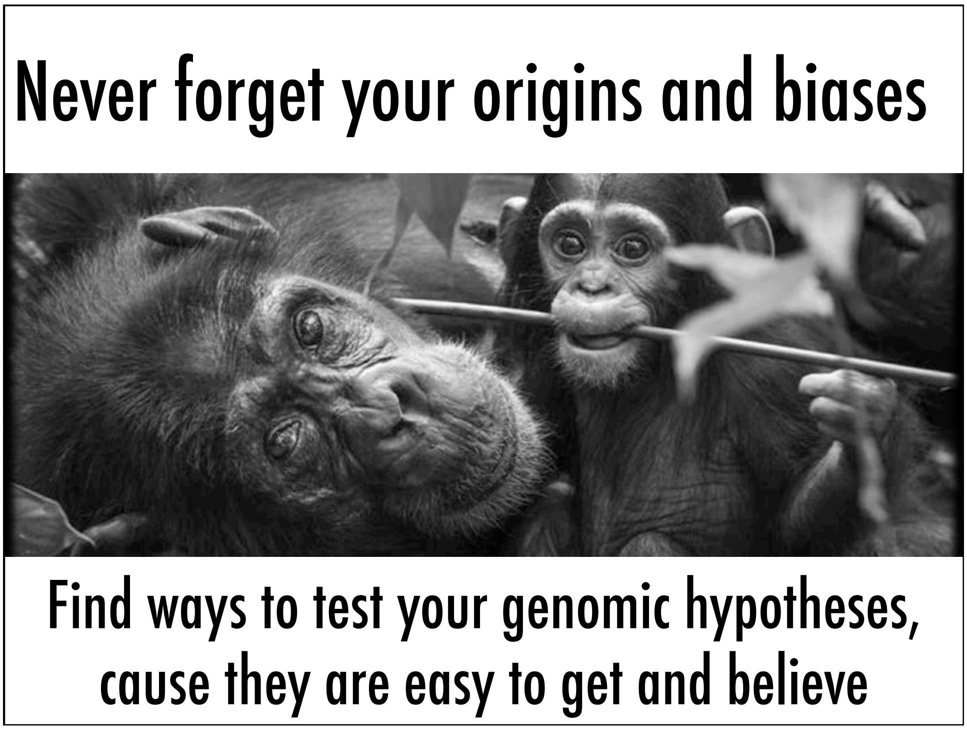
Molecular spandrels:

Story telling
vs.
Causal understanding

Genomics is full of adaptive stories

Treat your findings a hypotheses

How you can you test these?



Never forget your origins and biases

Find ways to test your genomic hypotheses, cause they are easy to get and believe

Stockholm
University