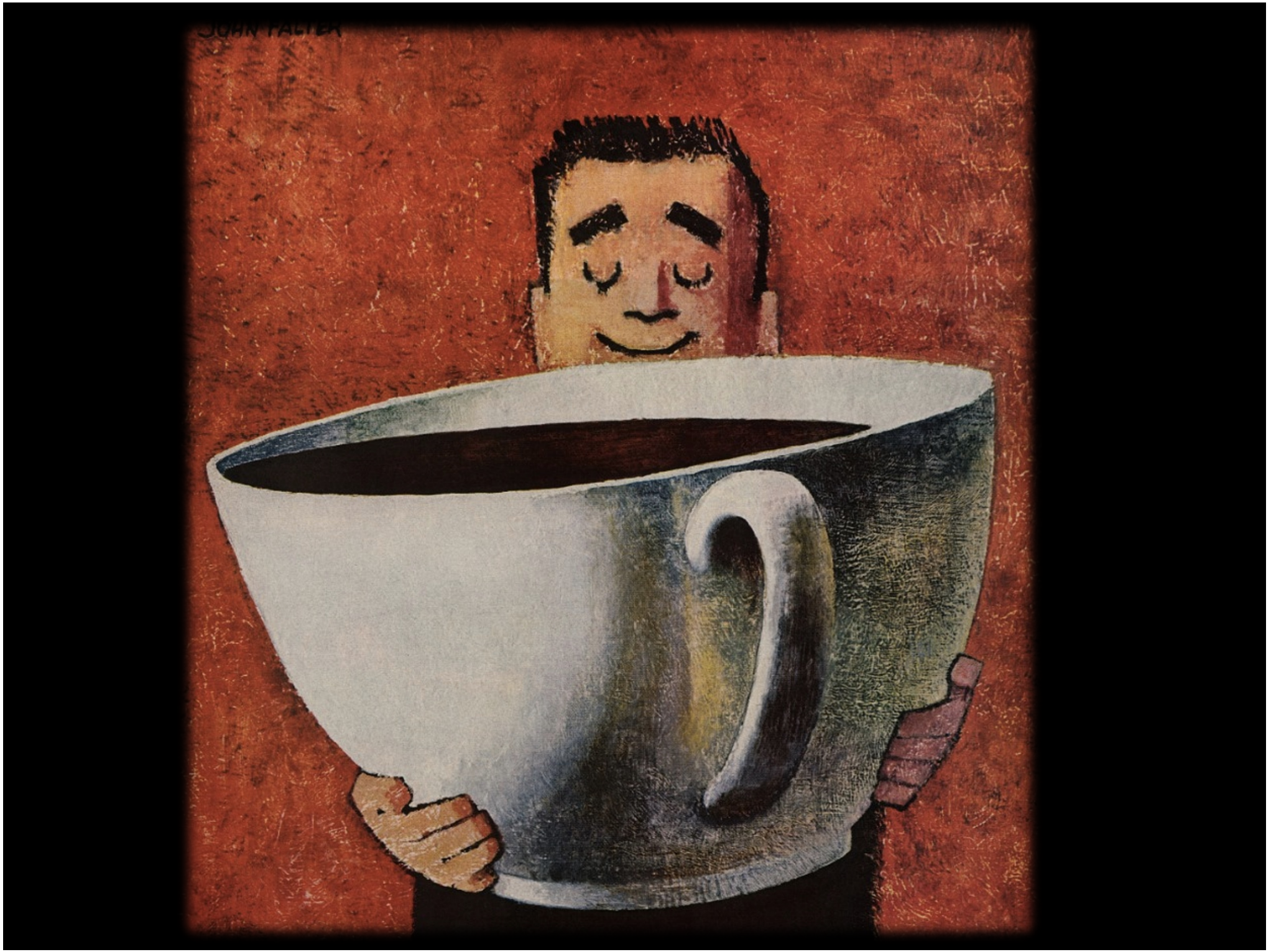

JOHN FALTER

So .. how do we avoid Apophenia?

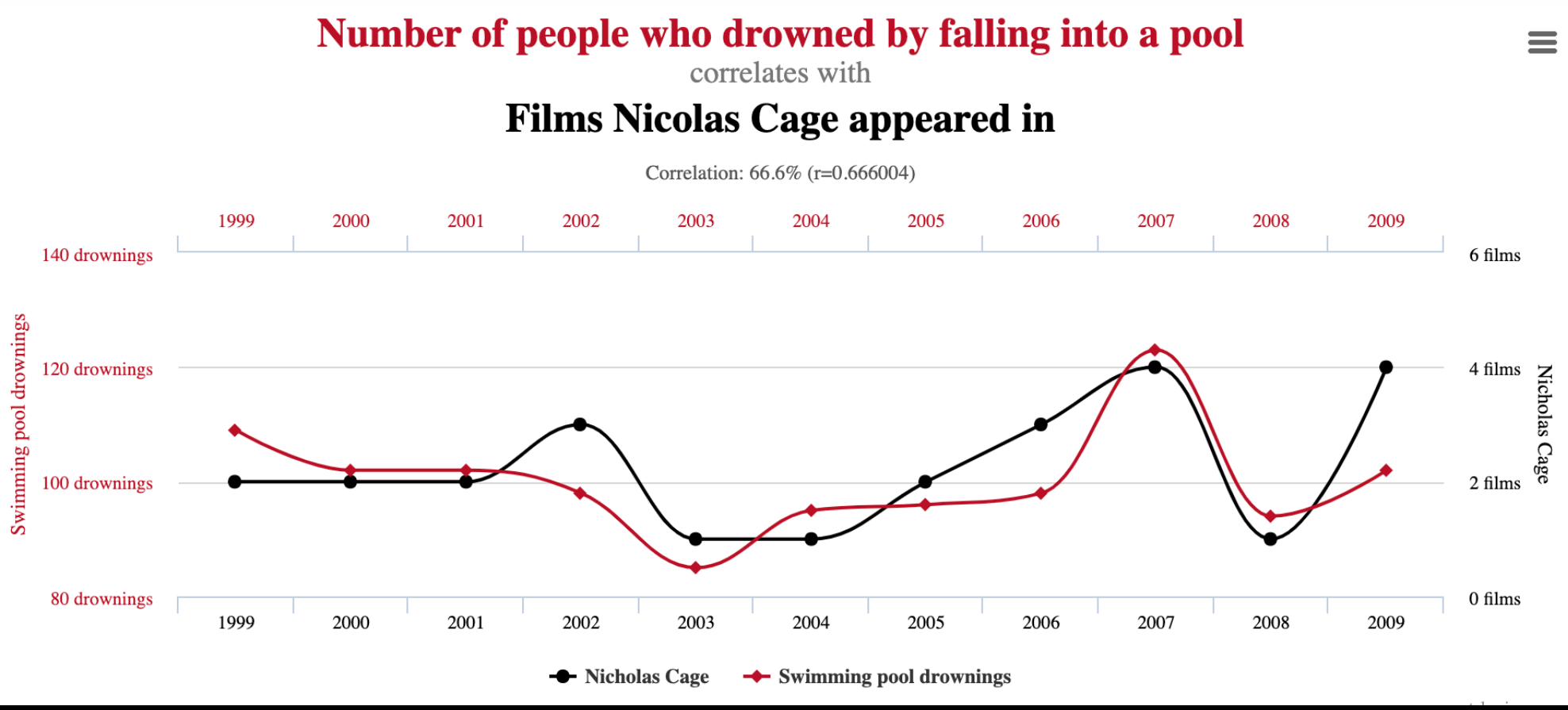
- Well … lets ask ChatGPT

  –how can humans overcome biases from apophenia

Test your hypotheses in independent ways

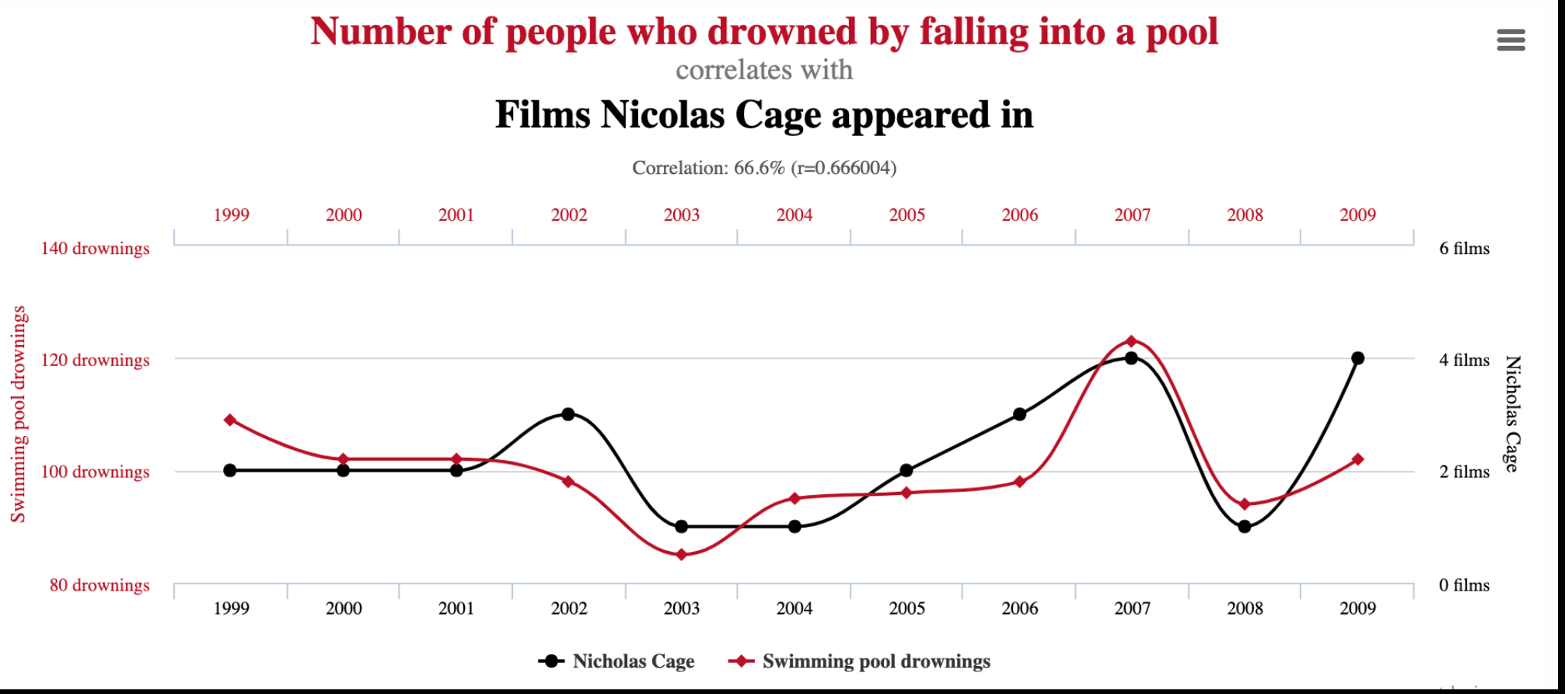
- Genomic datasets:
  - These are really observational data where patterns we observe have been created by things we barely understand
  - This is similar to all studies using observational data
    - Very susceptible to false positives

Genomic analyses easily find spurious correlations



- https://www.tylervigen.com/spurious-correlations

Genomic analyses easily find spurious correlations

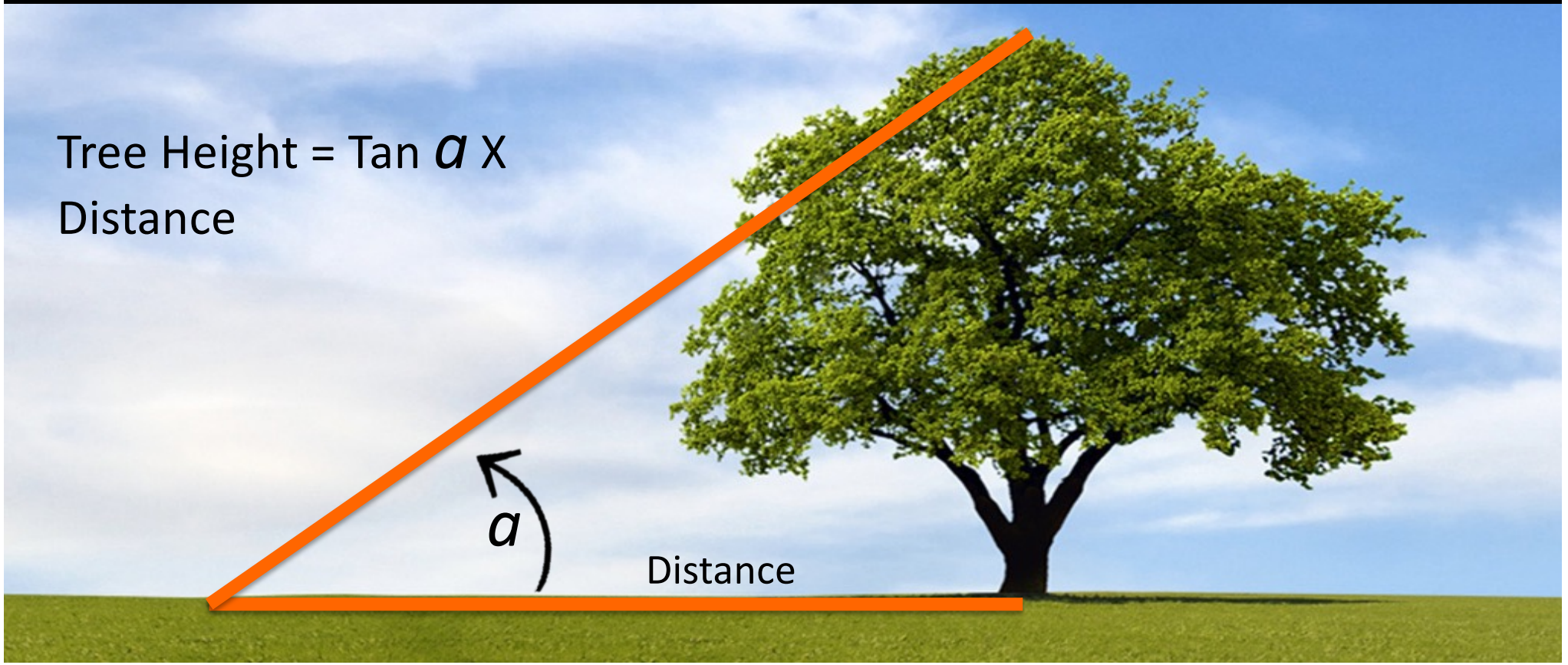


- https://www.tylervigen.com/spurious-correlations

Test your hypotheses in independent ways

- Genomic datasets:
  - These are really observational data where patterns we observe have been created by things we can barely envision
  - This is similar to all studies using observational data
    - Very susceptible to false positives
- Manipulation: functional validation via manipulation of genes, pathways, environments … real hypothesis testing!!
  - Experimental evolution, CRISPR KOs, environmental perturbations
- If you can't manipulate, at least triangulate!

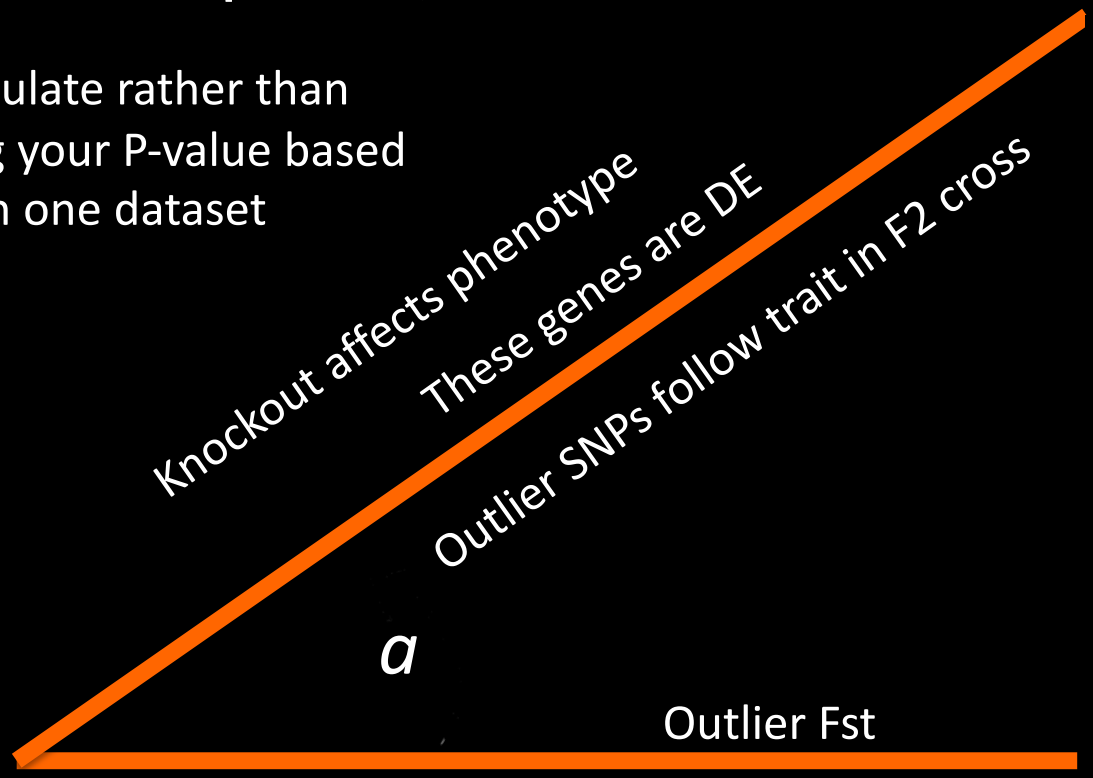
Triangulation for building evidence



Triangulation for building evidence

- Combine insights from independent axes of insight
  - biological replicates, test RNA patterns using proteins, etc.
- Challenge is maintaining genomic scale
  - Genome wide SNP scan for outliers, QTL mapping, RNA-Seq, knockouts, manipulations, etc.

Triangulate rather than justifying your P-value based on one dataset



Is it an adaptation?

What was ancestral state?

Is there any clinal variation?

Phenotype respond to chemical manipulation?

Response to selection experiment?

Three examples of triangulation in non-model species

- Population Genomics investigation of an adaptive phenotype
  - Independent genomic datasets
  - Orthogonal analyses
- Bioinformatic analysis of miRNA targets
  - Comparison across bioinformatic tools to assess consistency
  - Developing novel metric for biological signal in results
  - Comparative analysis for general insights and cross-check
- Functional genomic study of phenotypic plasticity
  - Experimental evolution
  - GWAS, RNAseq
  - CRISPR-Cas gene KO

Local adaptation

- Genomic scans may not be related to the trait you are focused upon
  - Large effect alleles at few loci
    - hard sweeps easy to detect via Fst, many tests
  - Many small effect alleles at many loci
    - Soft sweeps very, very difficult to reliably detect
- What is the genomic architecture of your trait of interest?

Population Genomics investigation of an adaptive phenotype

- Why does this dark morph exist?
  - Why female limited?
  - Is this an adaptive phenotype?
  - How and why did it evolve?

Pieris napi adalwinda

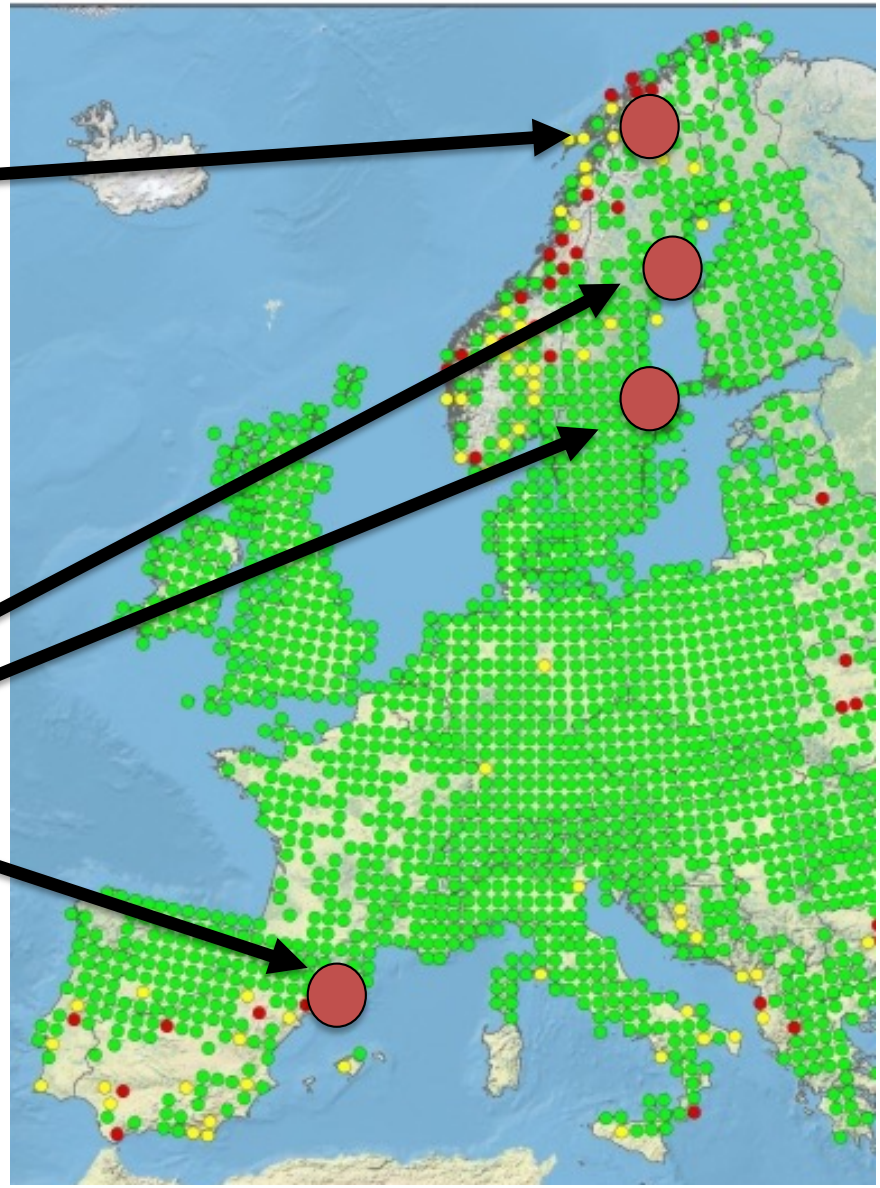
- Goal: find the genes, study their function
  - Connect genes to ecology

Pieris napi napi

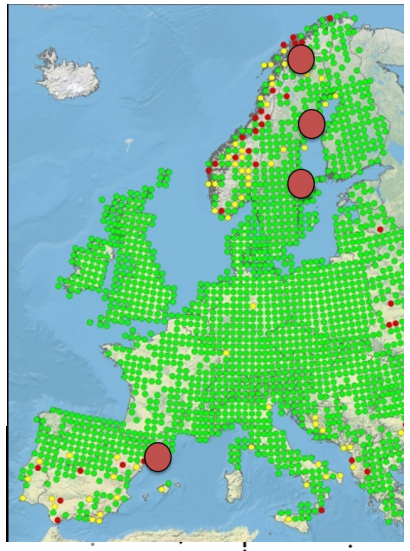
- Natural history
  - Common butterfly across Eurasia
  - Subspecies with female only dark morph in northern range limits (Sweden, Norway, Finland)

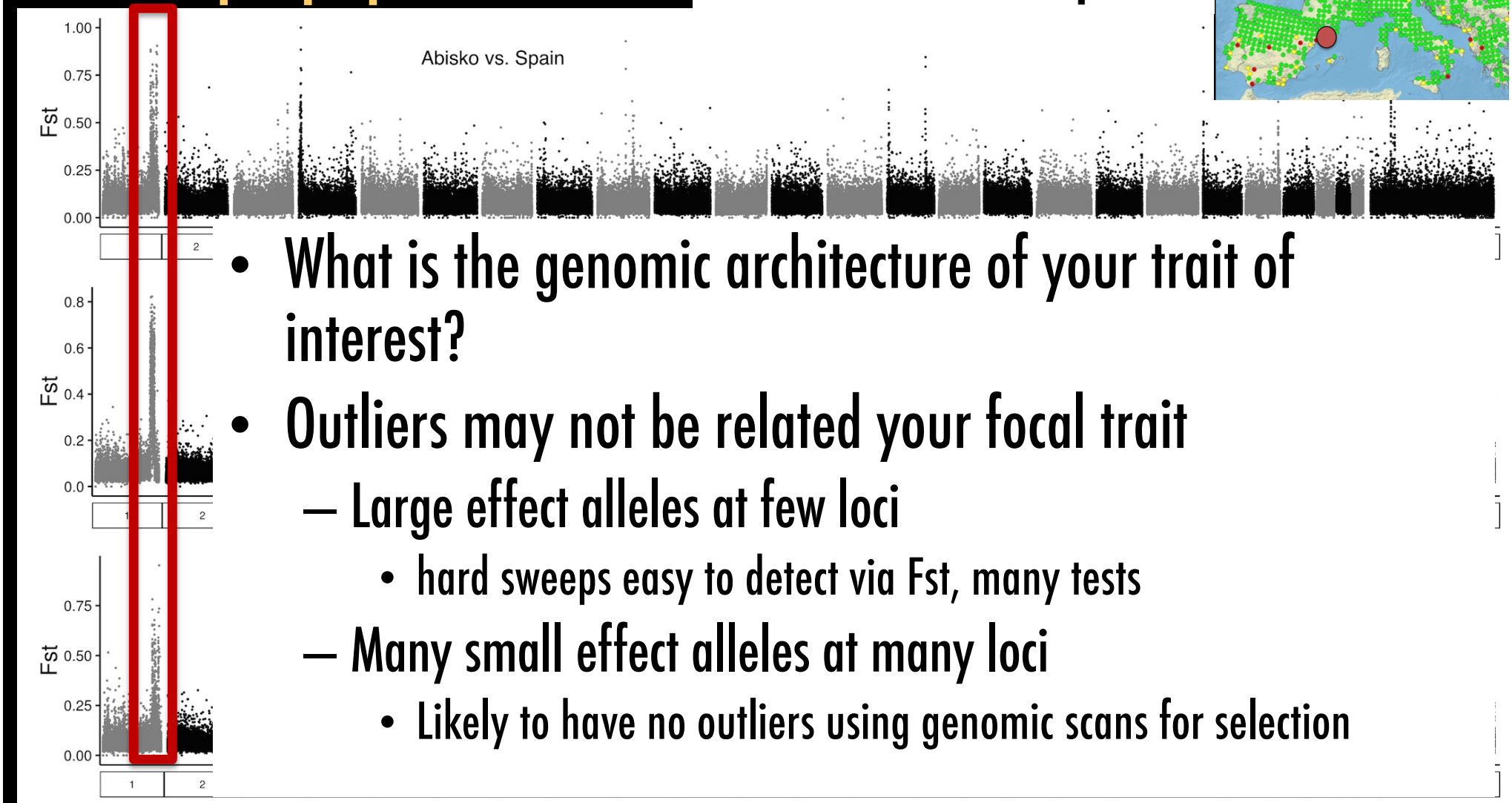
Genomic scans for local adaptation

Population re-sequencing using Pool-Seq across Europe (n=24 each, thorax)



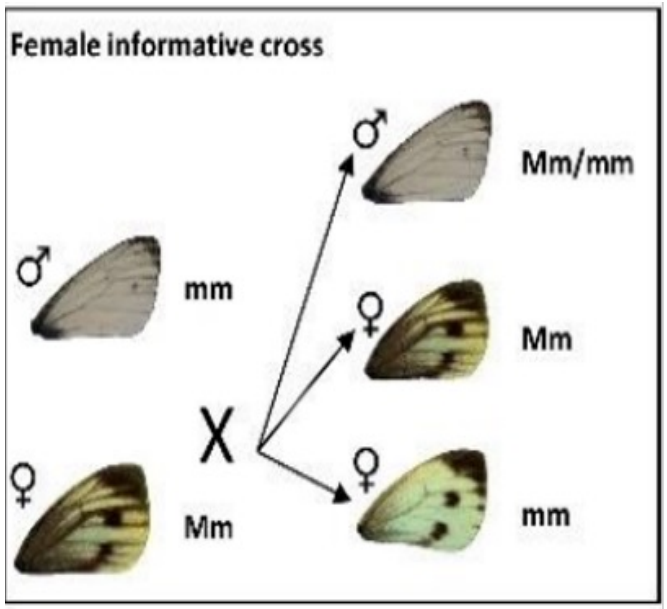
Fst of each population compared to dark morph population

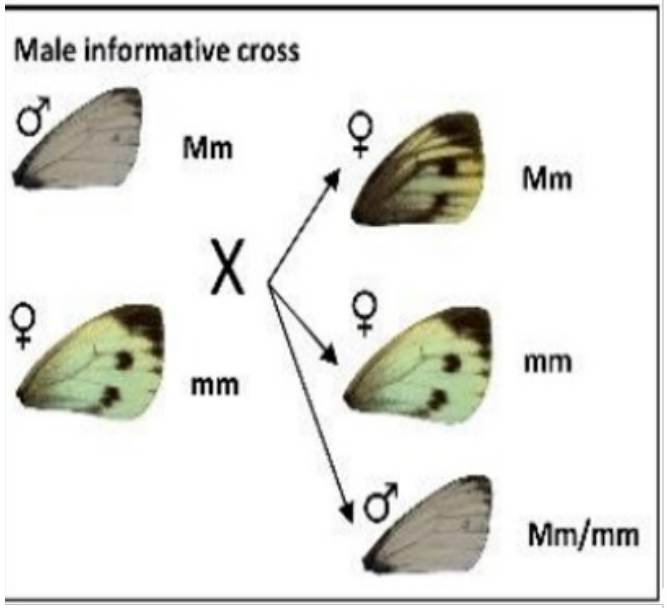


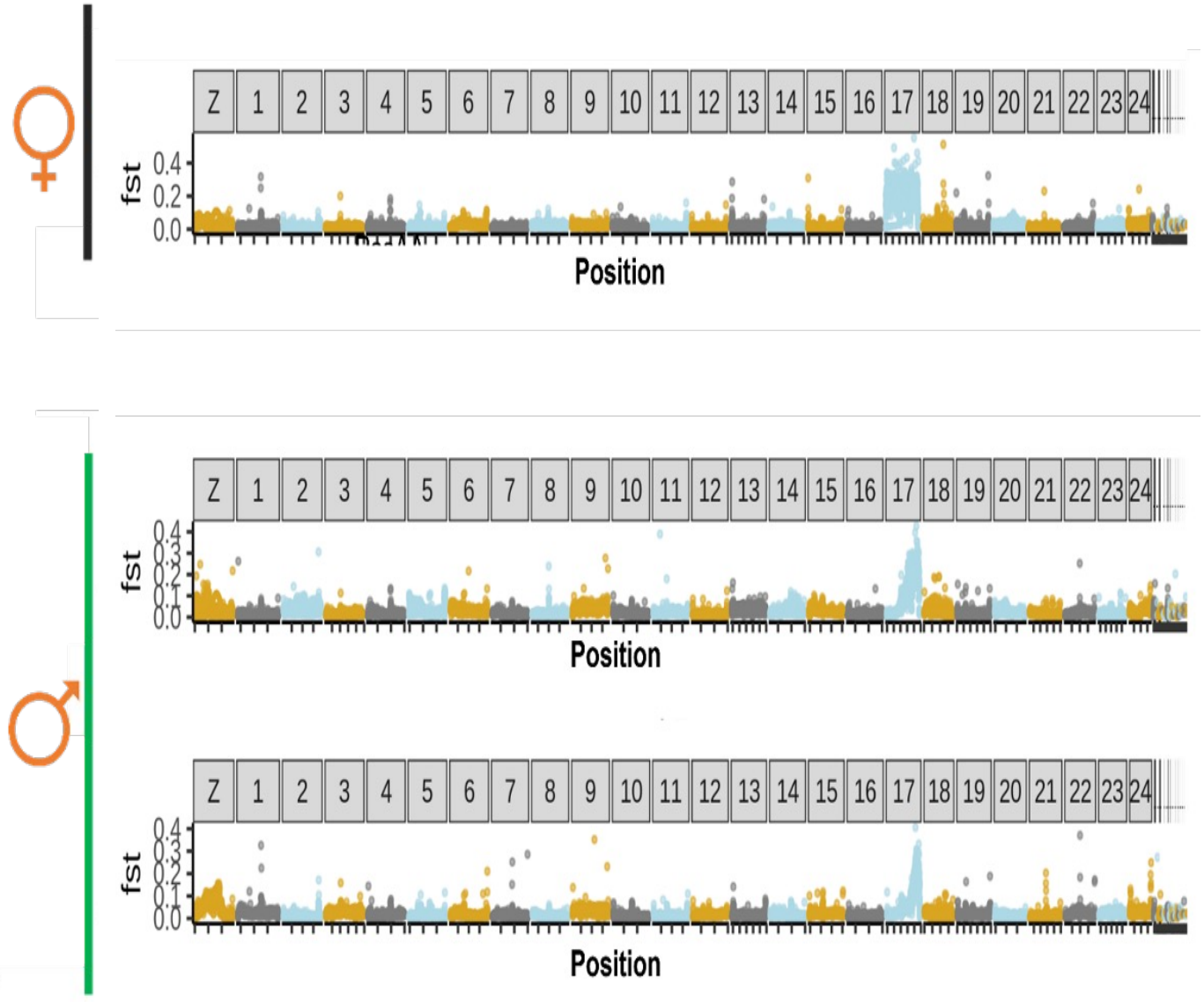


- What is the genomic architecture of your trait of interest?
- Outliers may not be related your focal trait
  - Large effect alleles at few loci
    - hard sweeps easy to detect via Fst, many tests
  - Many small effect alleles at many loci
    - Likely to have no outliers using genomic scans for selection

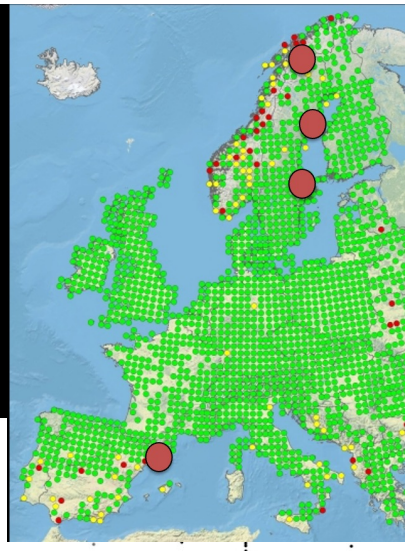
Test hypothesis using independent method: crosses

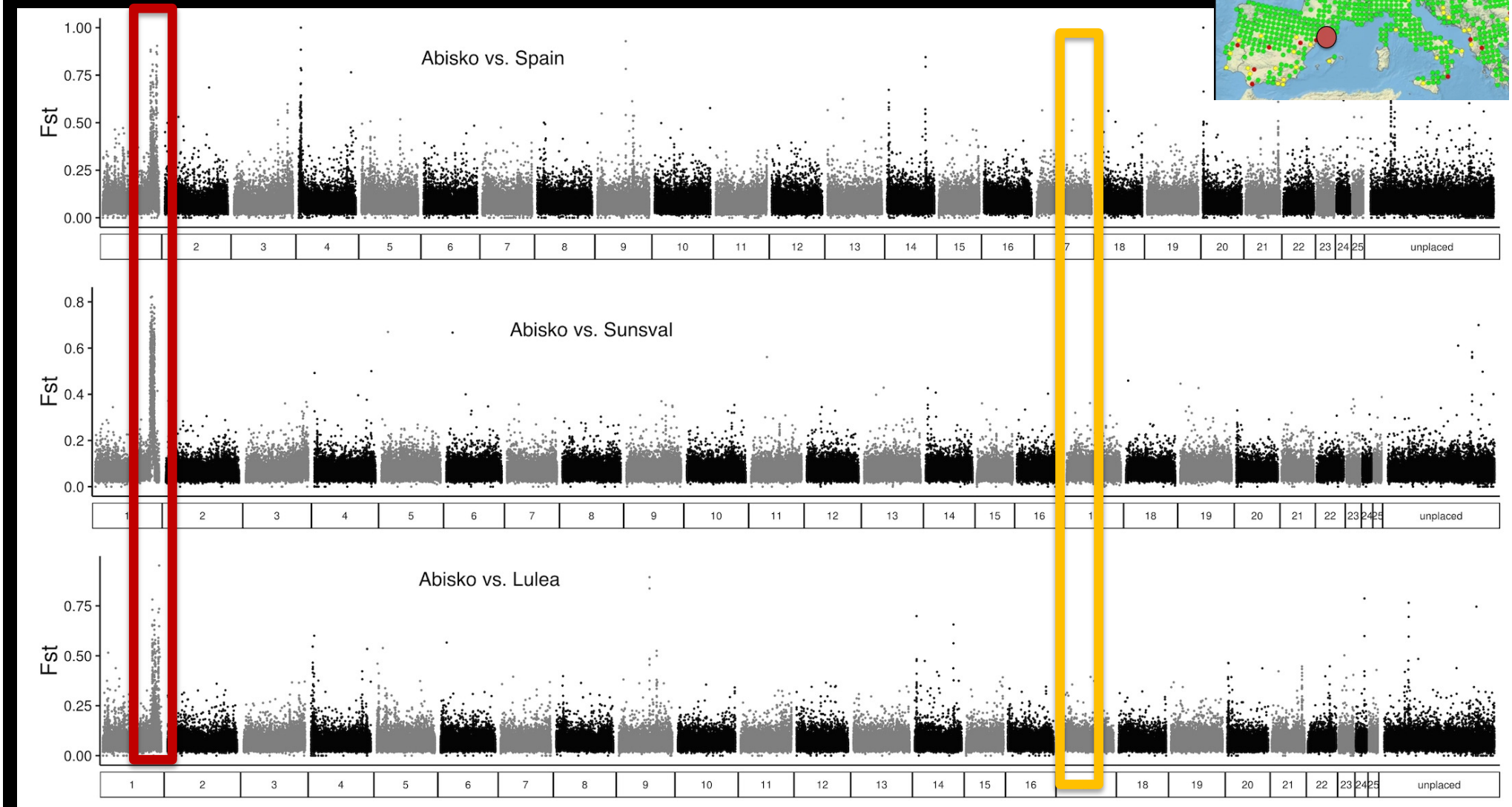




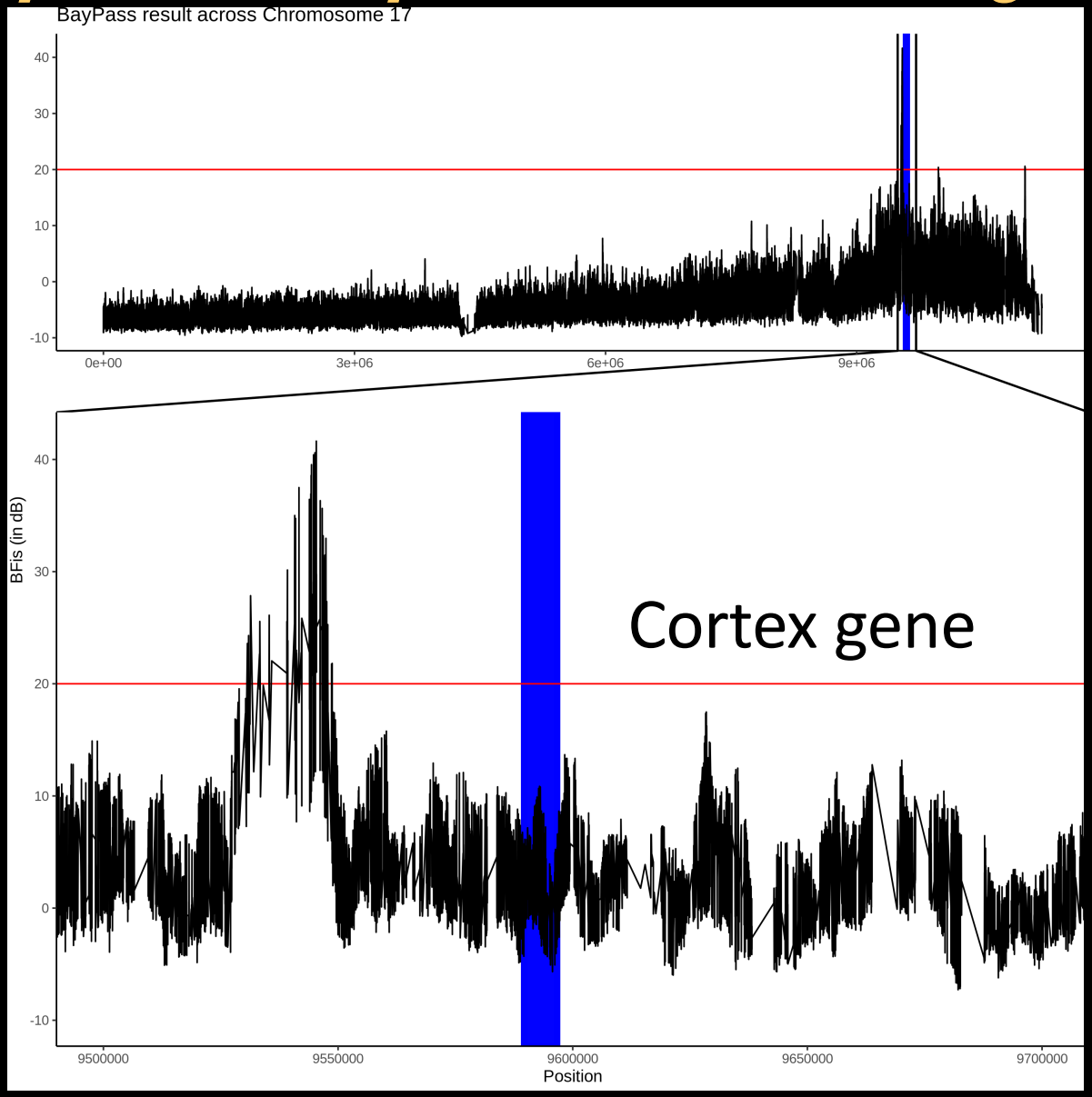


Fst of each population compared to dark morph population





Baysian analysis of all crossing data

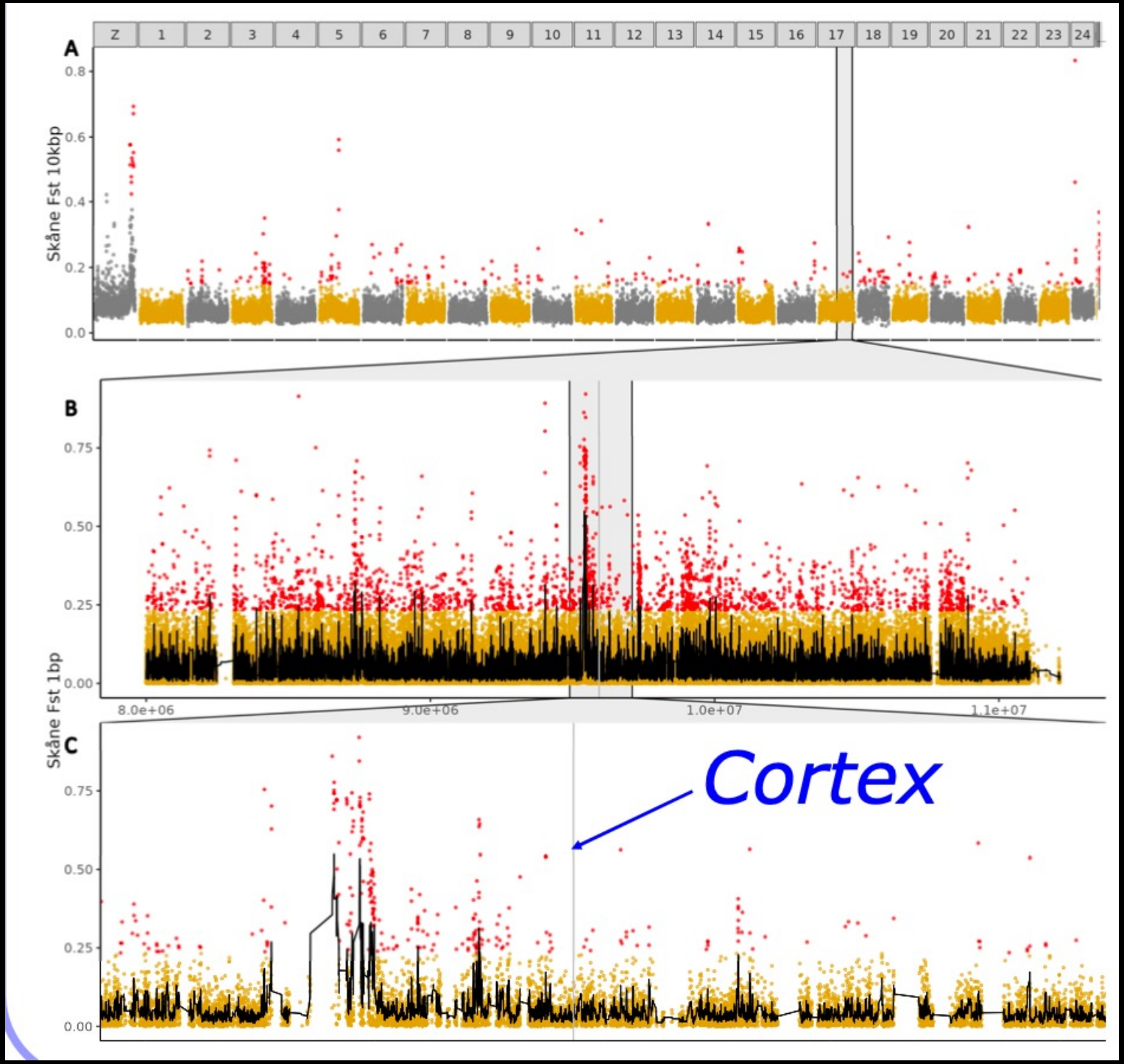


Gautier M. 2015. Genome-Wide Scan for Adaptive Divergence and Association with Population-Specific Covariates. doi: 10.1534/genetics.115.181453.

Fst for 10 kb windows

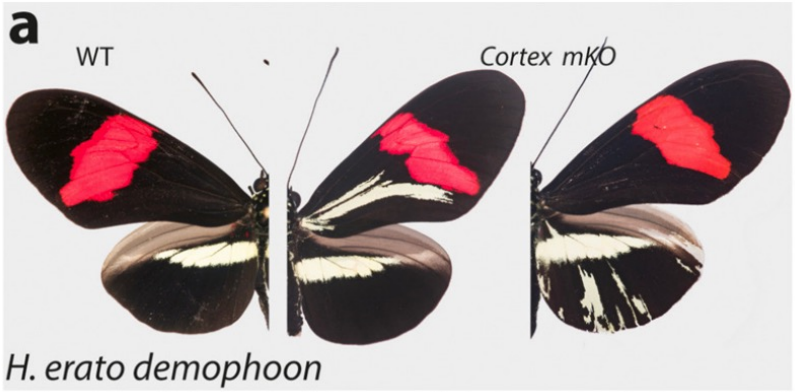
Fst for each bp

Fst for each bp



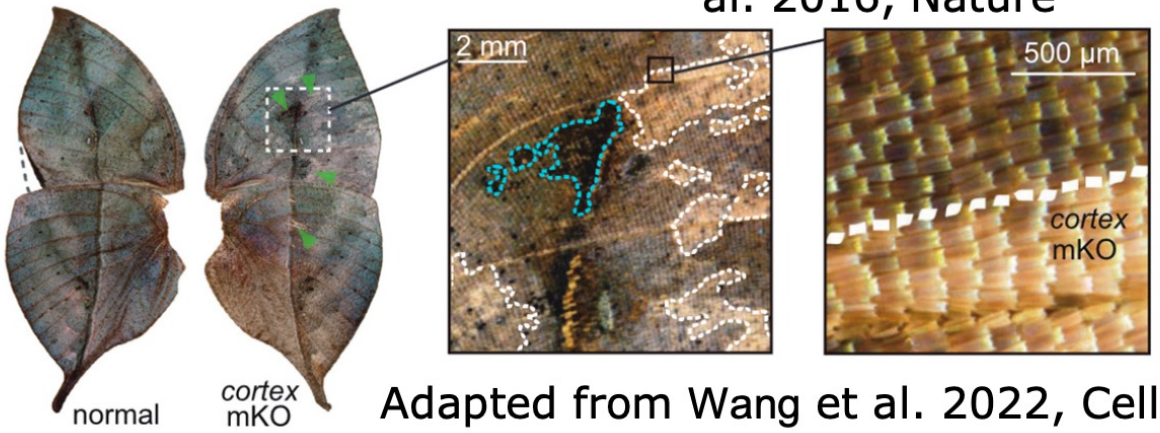
Cortex

cell-cycle regulator



Adapted from Livraghi et al. 2021, elife

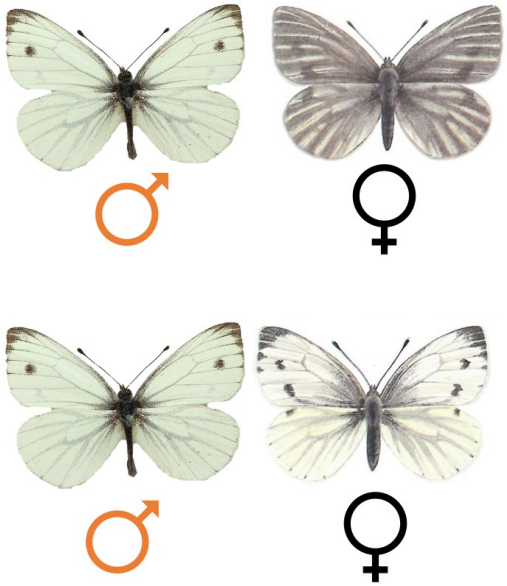
Adapted from van't Hof et al. 2016, Nature



Adapted from Wang et al. 2022, Cell

Has a common role in wing pigmentation and patterning across Lepidoptera, likely via scale cell developmental processes

Tunström et al. (in prep)

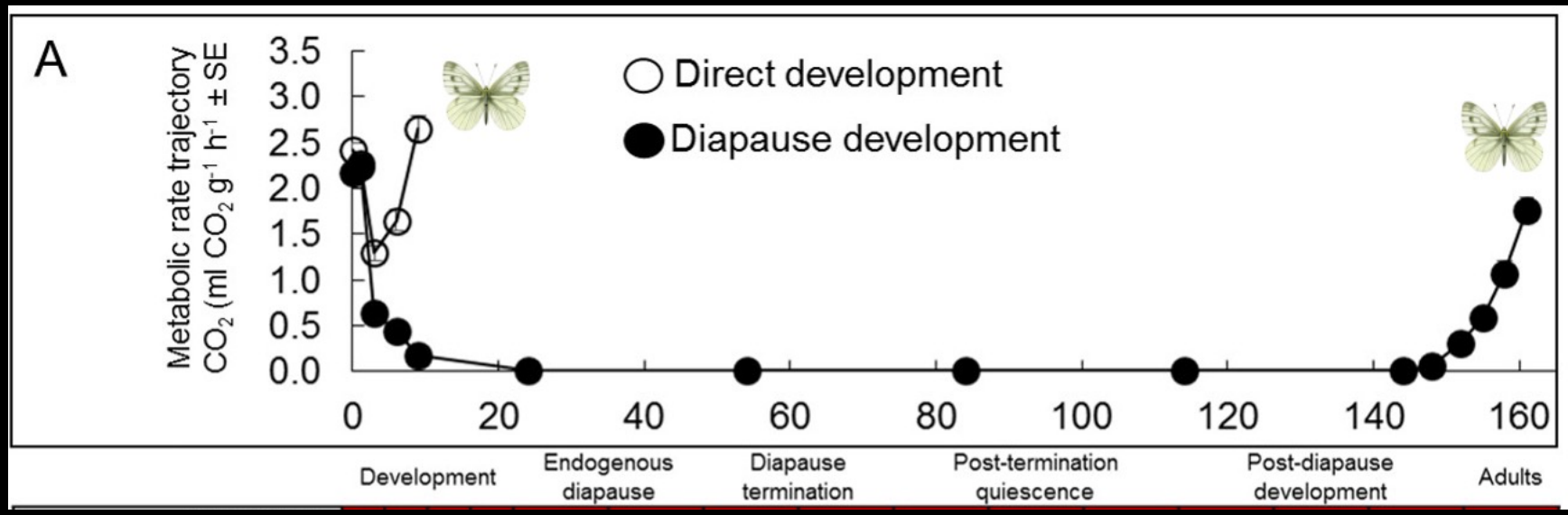


Population Genomics investigation of an adaptive phenotype

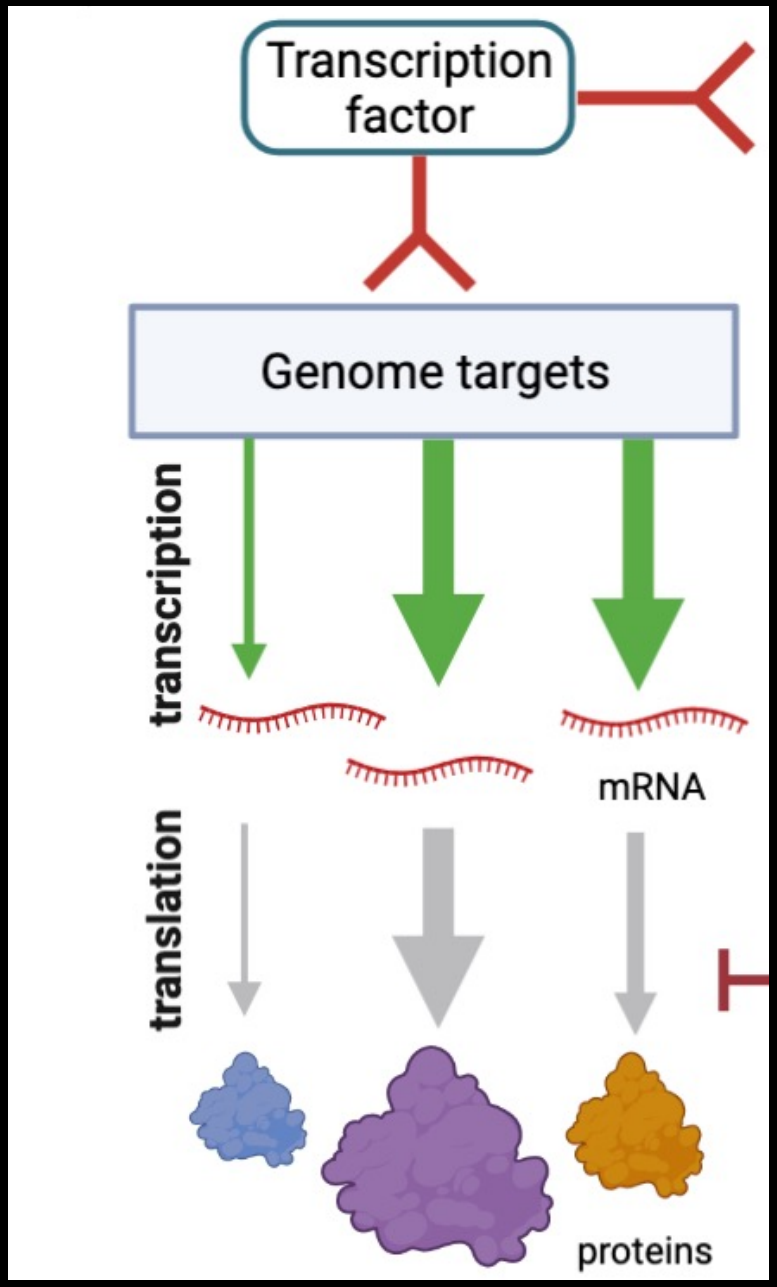
- Outliers may have nothing to do with your view of how things work
- Intersection with orthogonal analysis is critical to gain deeper causal insights
  - Without validation steps, naked conclusions make weak contributions to the literature
- Here: intersection between genomic scan and crosses localized adaptation signal to single genomic region

Bioinformatic analysis of miRNA targets

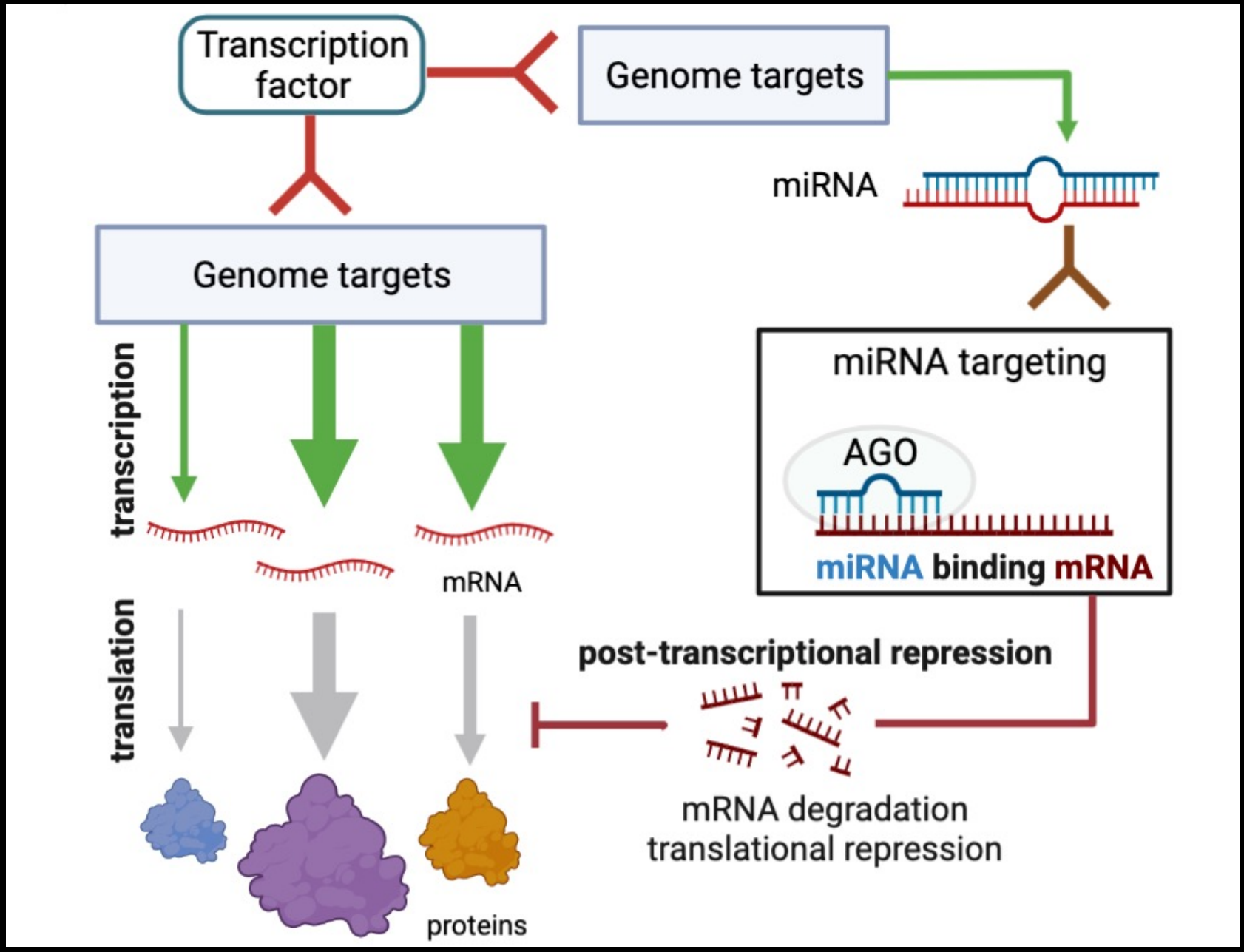
Does miRNA play a role in diapause progression in Pieris napi



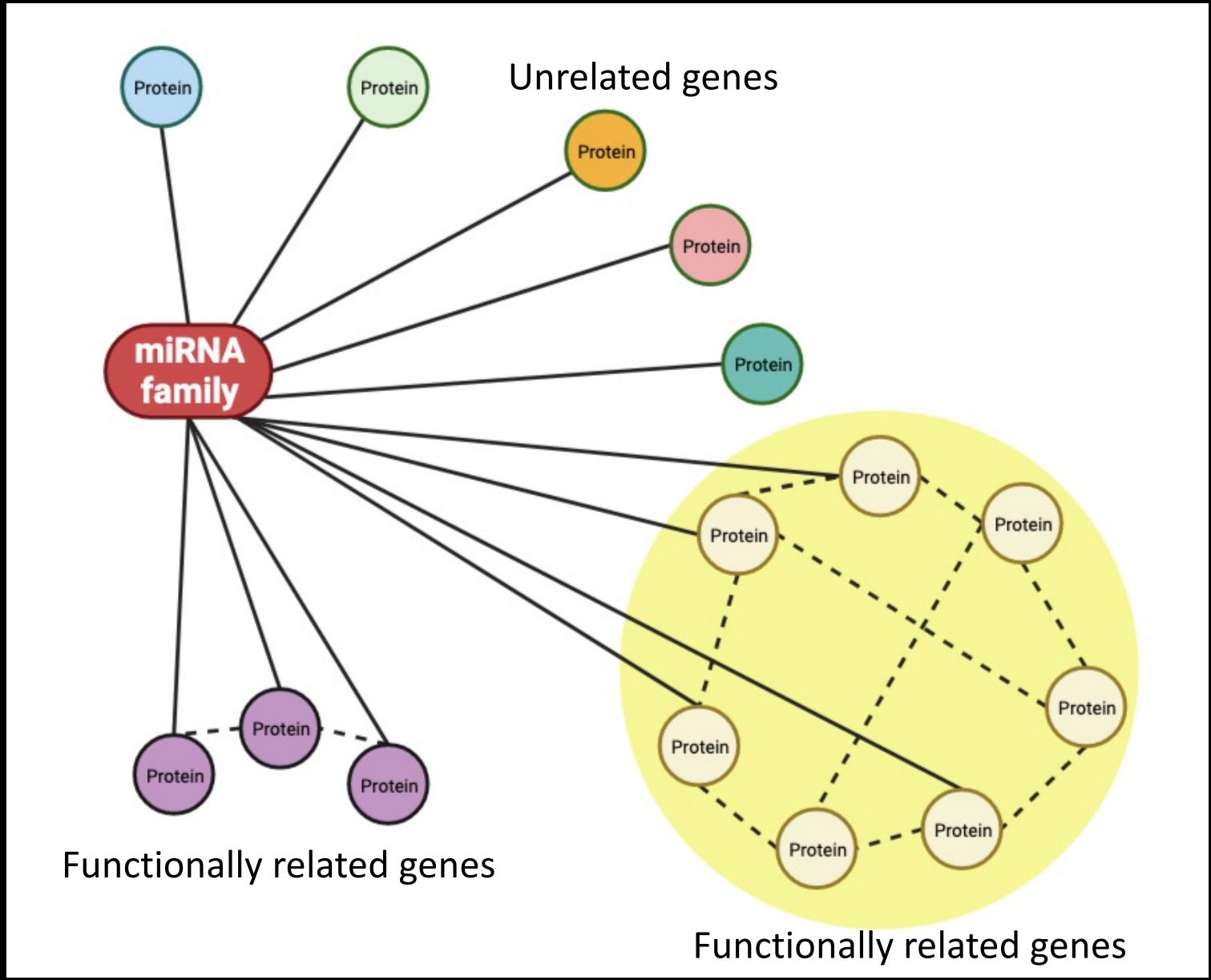
The role of miRNA in sculpting the transcriptome



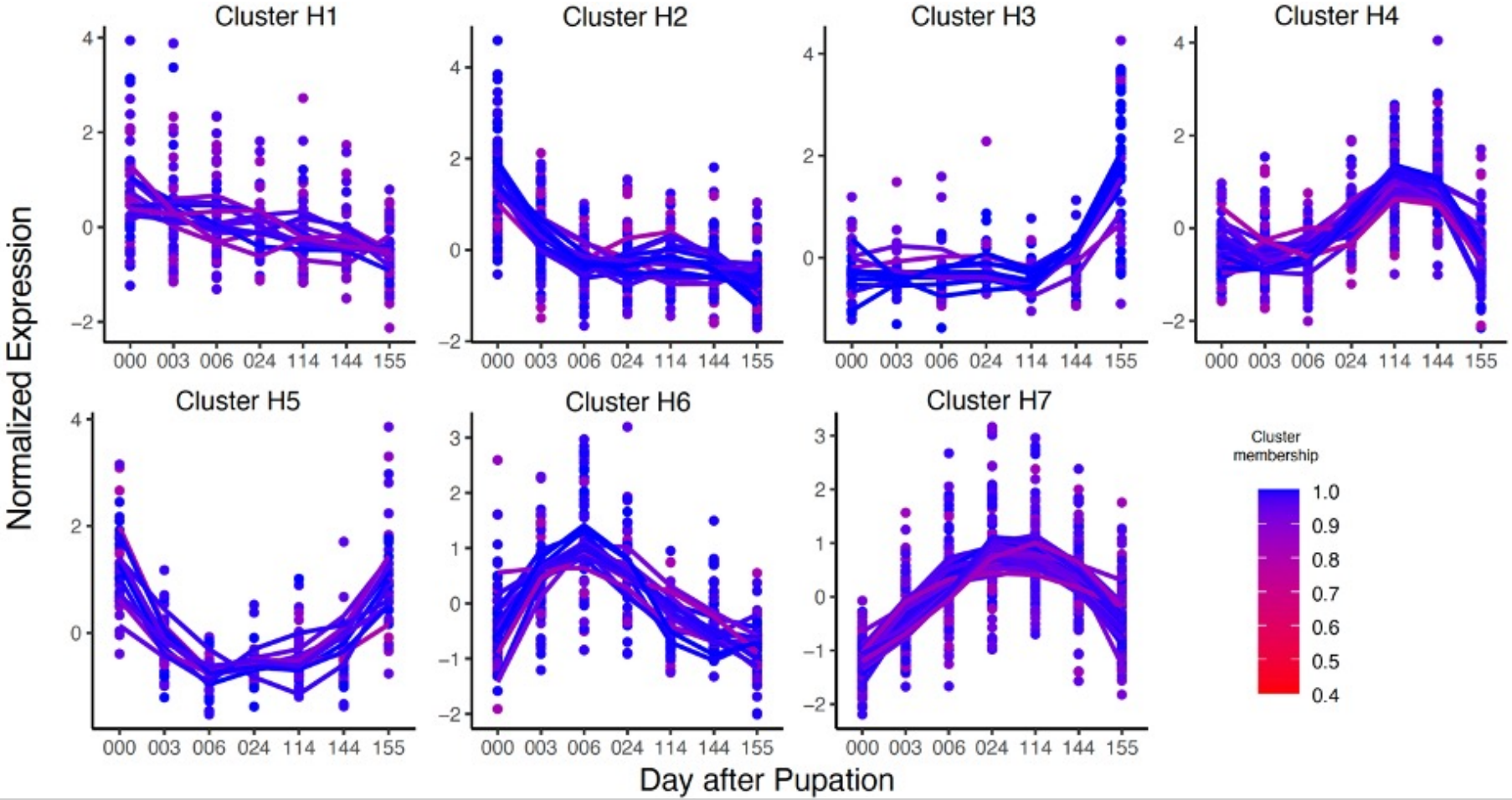
The role of miRNA in sculpting the transcriptome



Regulatory network view of miRNA impacts

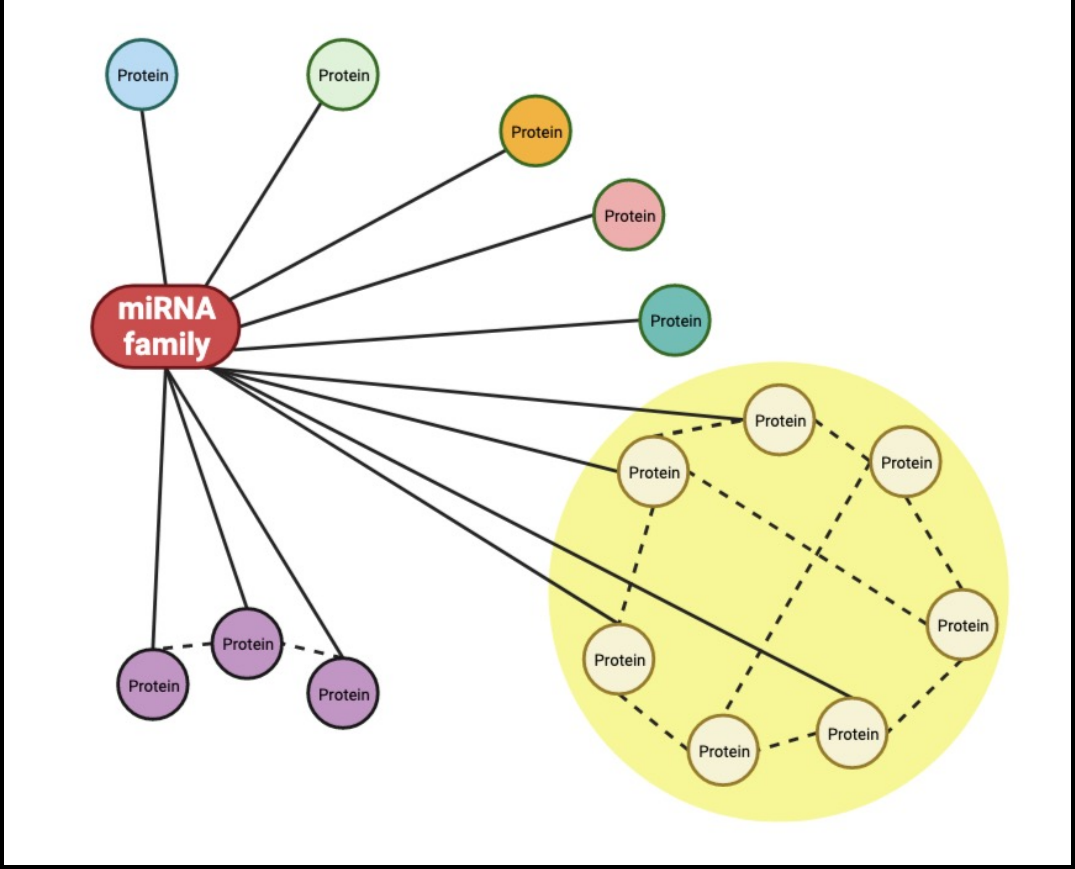


miRNA expression changes



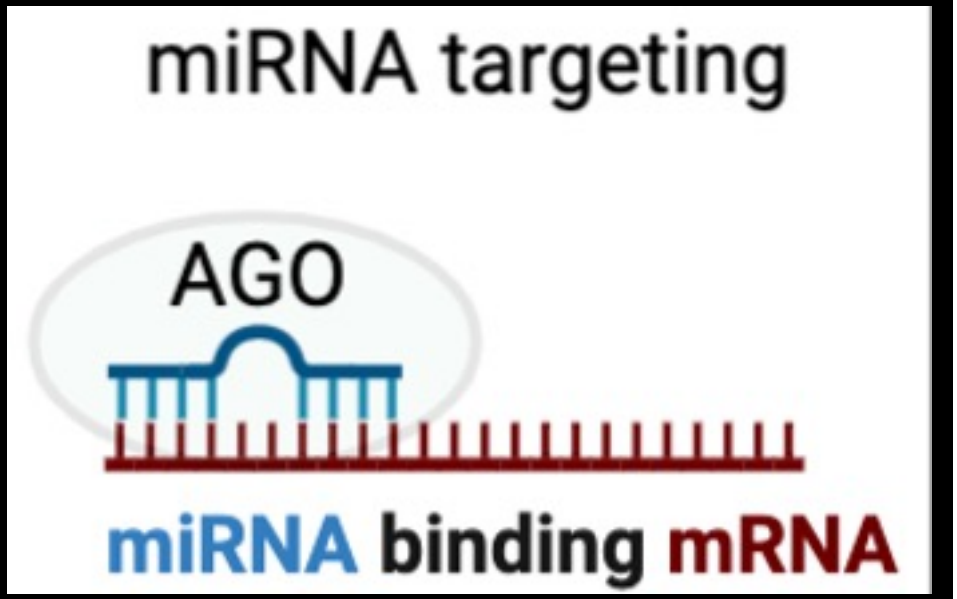
OK, so some miRNAs are changing through time..

Where are they targeting?
What are they doing?



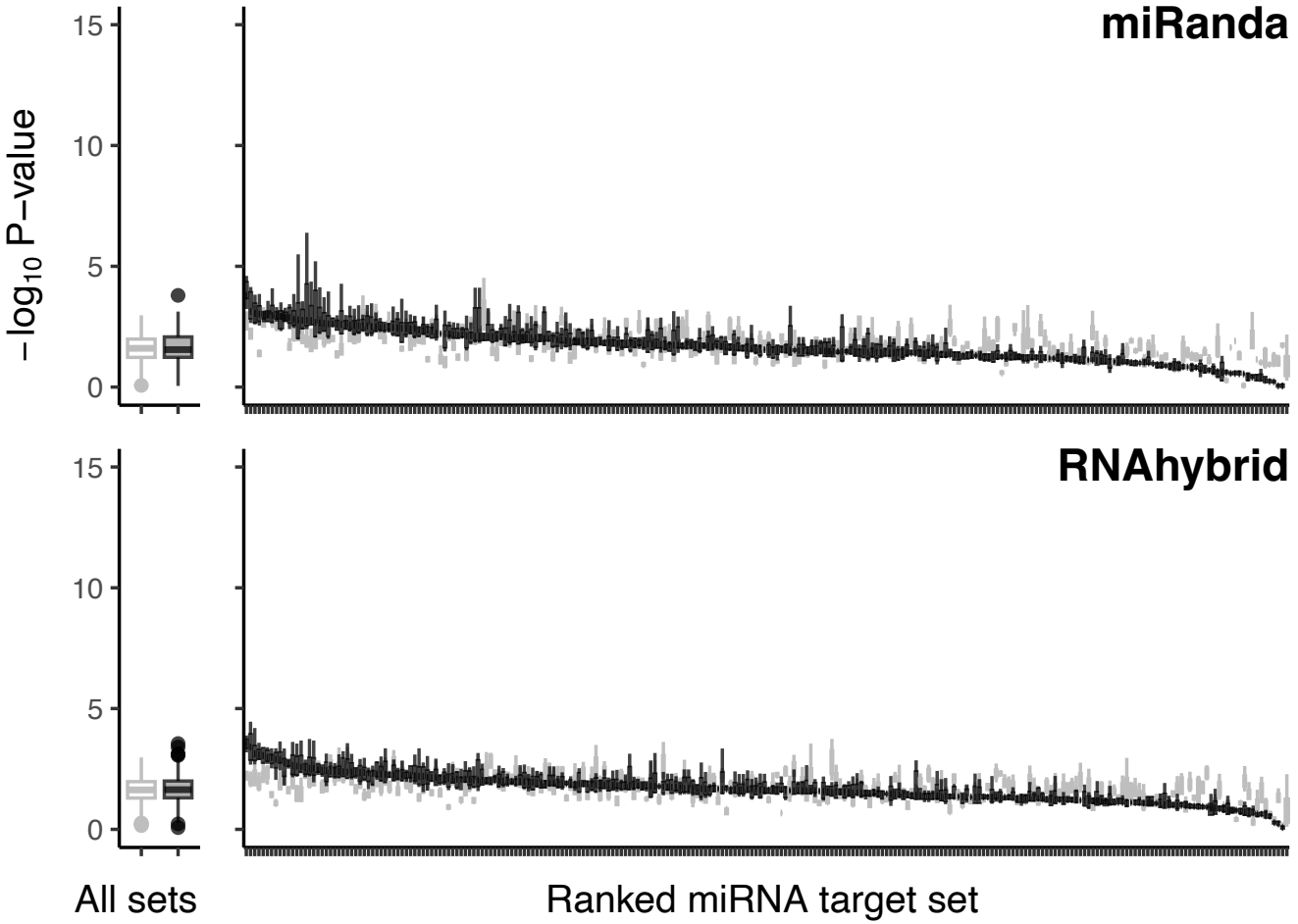
What functional groups or pathways might they regulate?

miRNA target detection

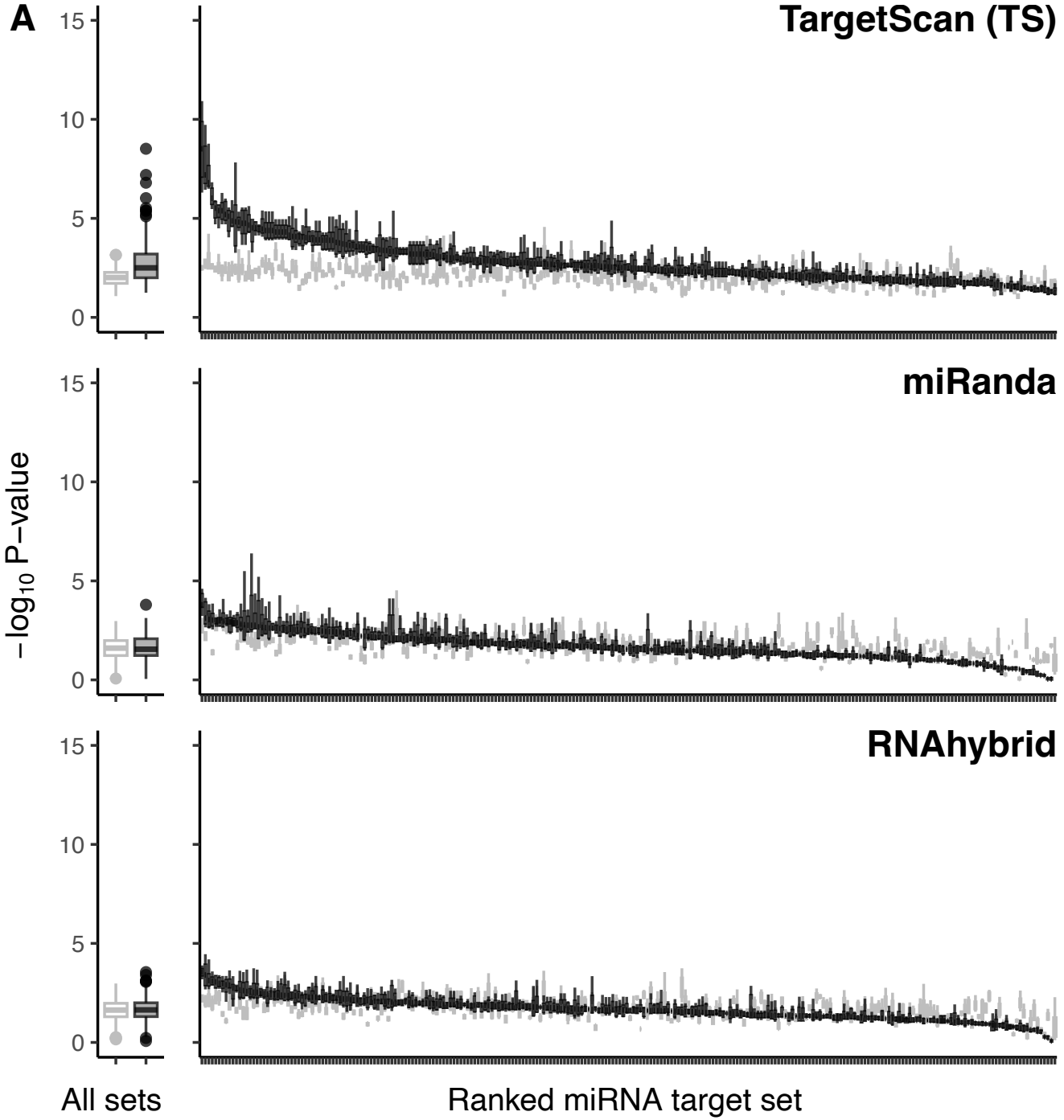


- miRNAs primarily bind a very short, ±7 bp region of the 3'UTR of mRNA
- This binding ultimately leads to a decrease of translated proteins
- There are 100,000's of 7 bp motifs in genome, of which miRNAs bind small fraction

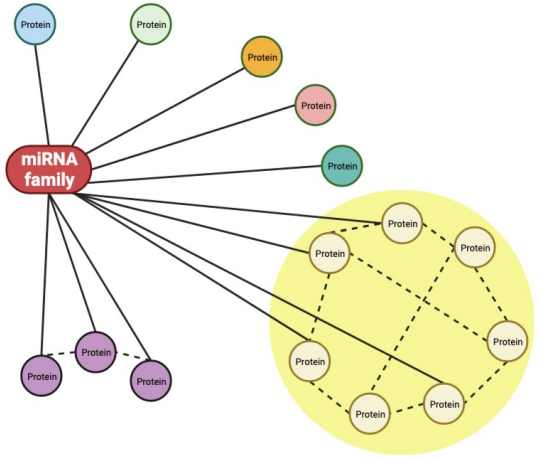
Assessing functional enrichment for targets of each predicted miRNA gene



Assessing functional enrichment for targets of each predicted miRNA gene



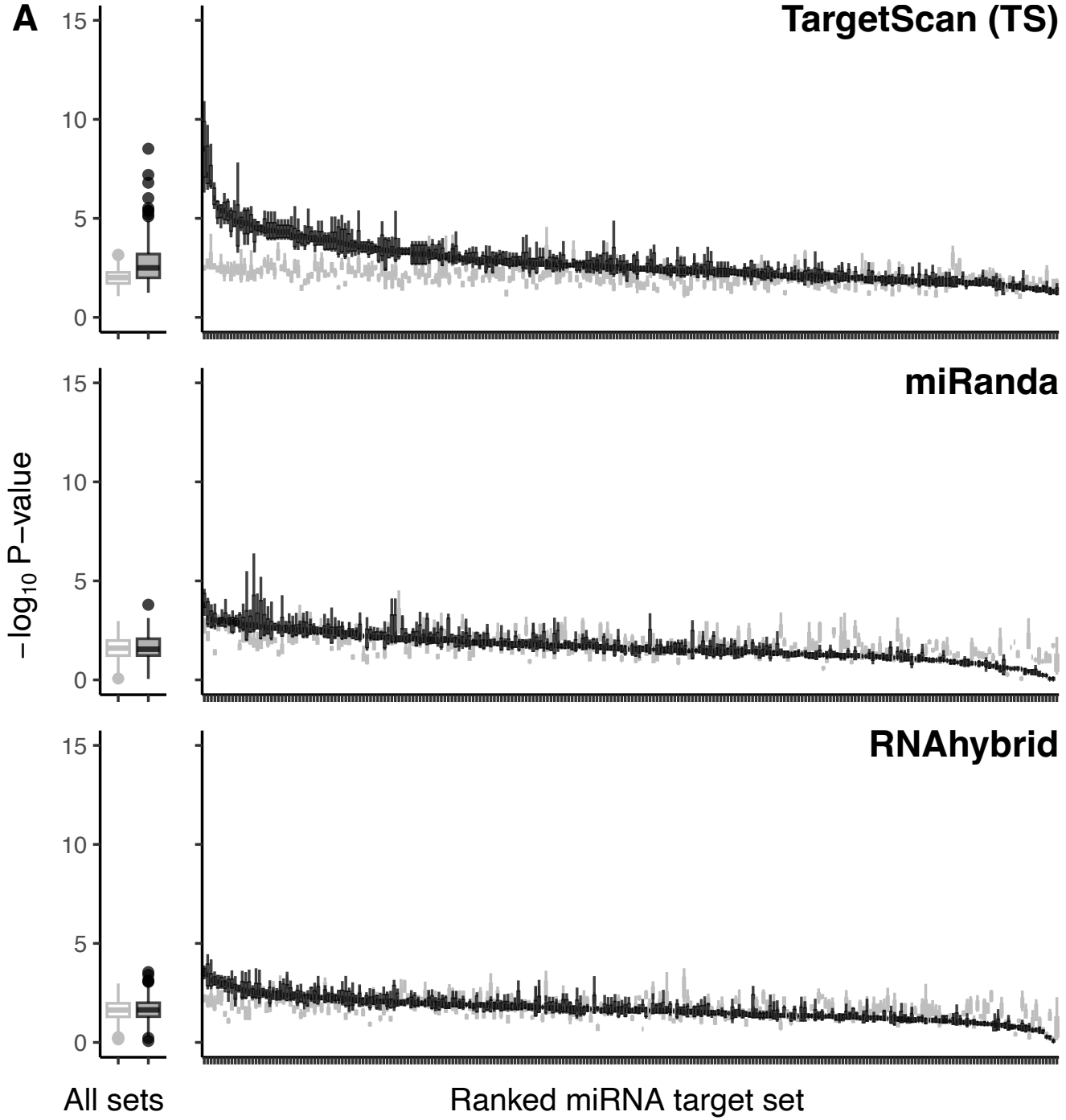
Why variation in functional enrichment in targets



- Targetscan was run using 7 species alignment of 3'UTRs, identifying 7 bp motifs that were identical
  - Under strong purifying selection, a expected when functional
- miRanda, RNAhybrid
  - Run on only 1 species, appear to have a very false positive rate
  - This is well documented in literature

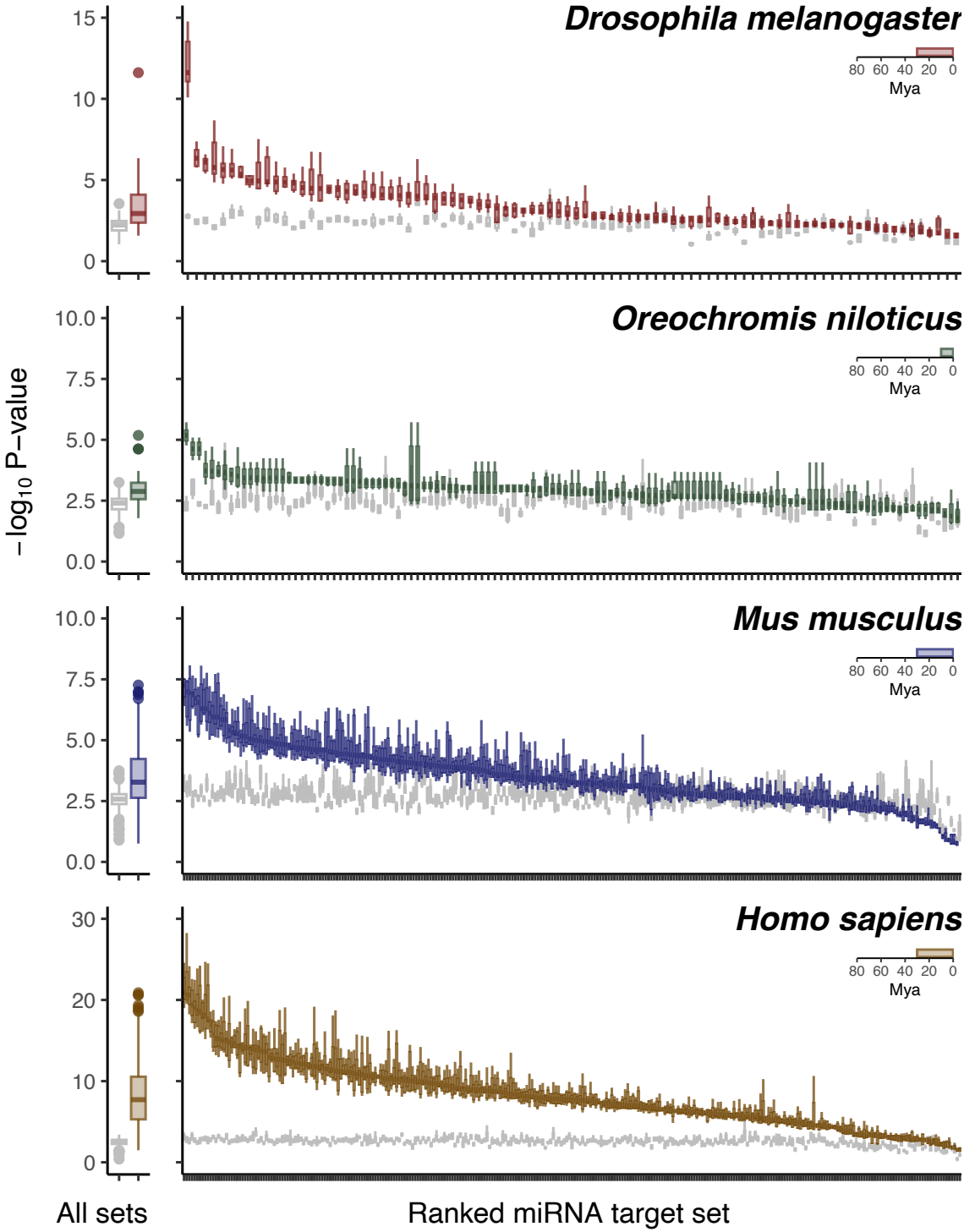
- Pinzón N et al. 2017. microRNA target prediction programs predict many false positives. Genome Res. 27:234–245.
- Ritchie W, Flamant S, Rasko JEJ. 2009. Predicting microRNA targets and functions: traps for the unwary. Nat Methods. 6:397–398.

So, if Targetscan is really doing better, can I find functional enrichment in other species target sets?



For some miRNAs, their targets are enriched for genes of related function

Wheat et al. (in prep)

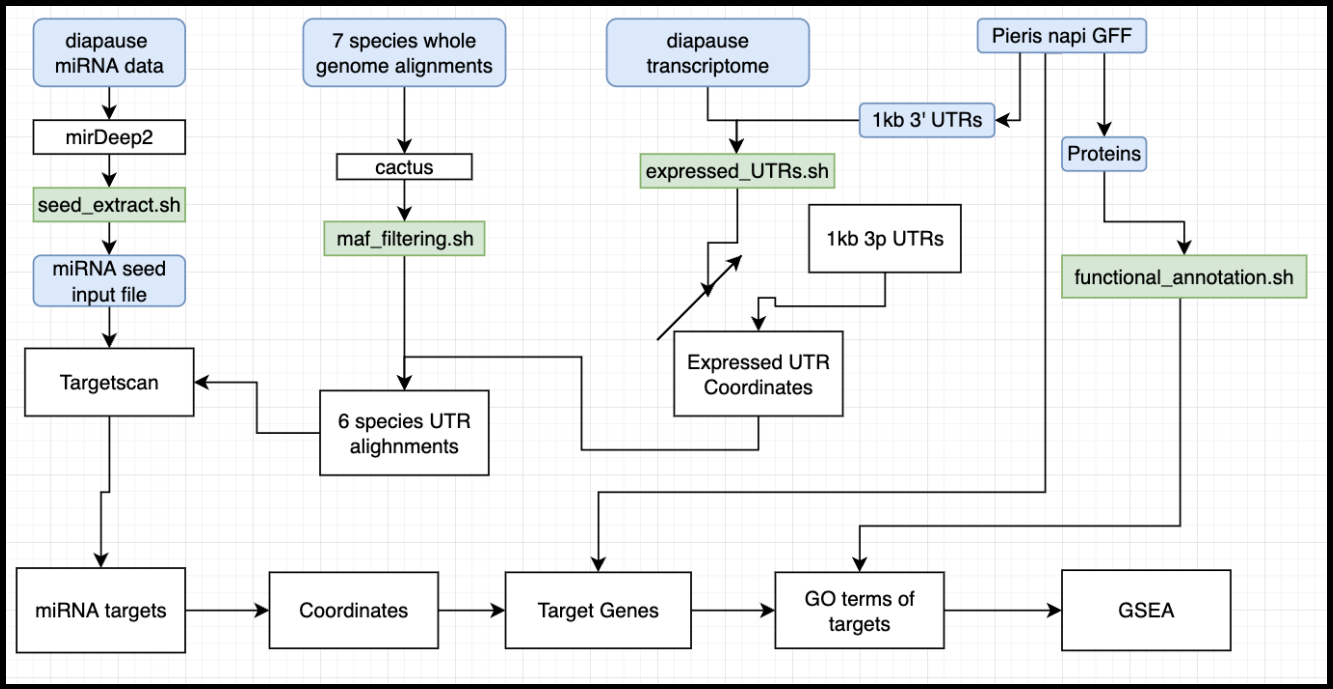


So, why don't more people use Targetscan with alignments?

- Running miRanda:
  - Download, load 3'UTR data from your species, load miRNA seed sites, run

So, why don't more people use Targetscan with alignments?

- Running TargetScan7 with alignments
  - Download scripts, generate 3'UTR alignments for 7 species, load miRNA seed sites, run



Bioinformatic analysis of miRNA targets

- Detecting miRNA expression changes is easy, but target detection is inherently very difficult
- Intersection
  - Comparison across bioinformatic tools
    - Revealed inconsistent results, primarily because used VERY different methods (e.g. using vs. not using alignments)
  - Developed novel metric for assess biological signal in results
  - Used cross species comparisons for cross-check & generality
- Here: intersection across divergent methods, 1st principals metric, and comparative analysis using other data

This is a piece of toast

Functional genomic study of phenotypic plasticity

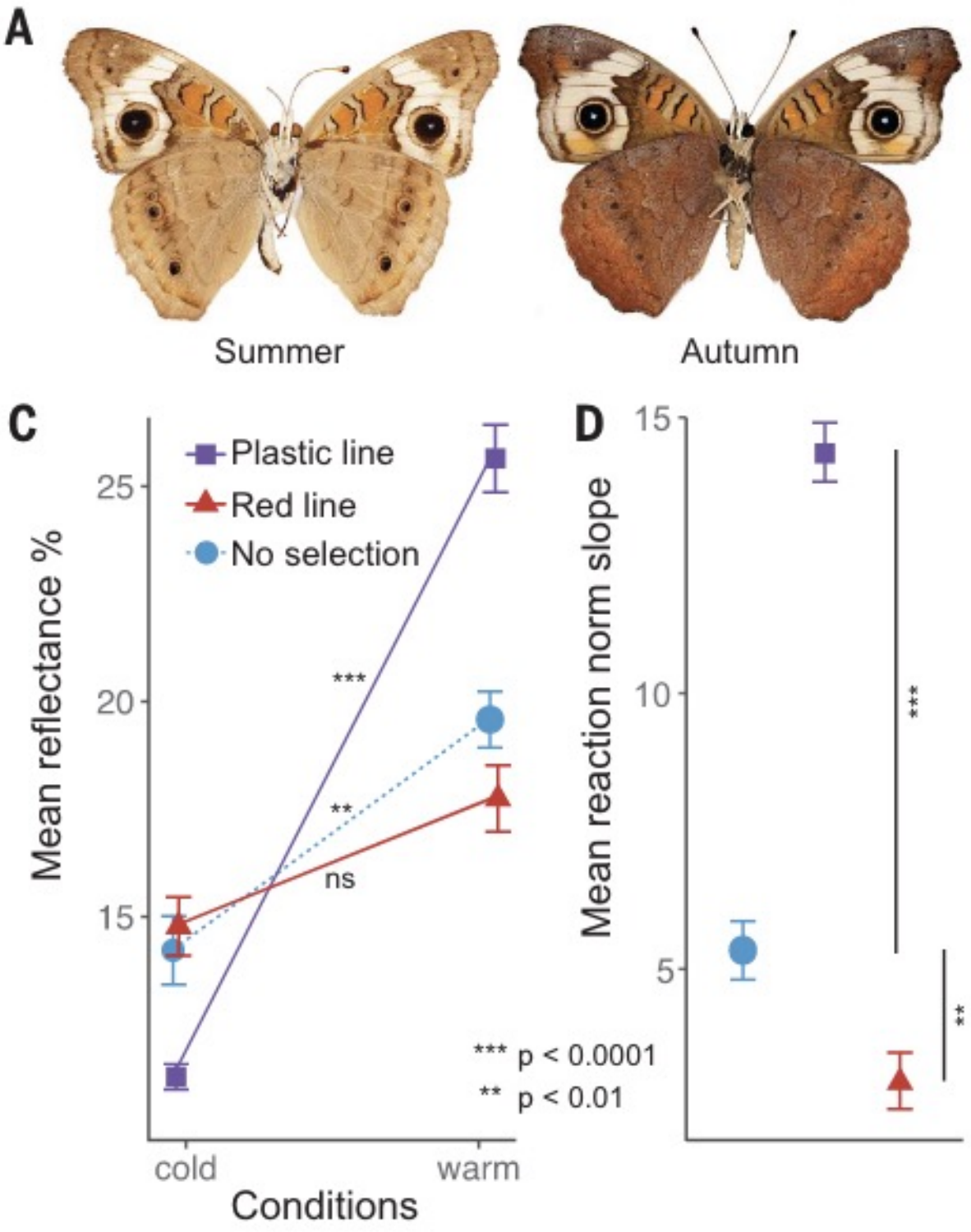
- Identifying the genetic basis of plastic phenotypes is very challenging

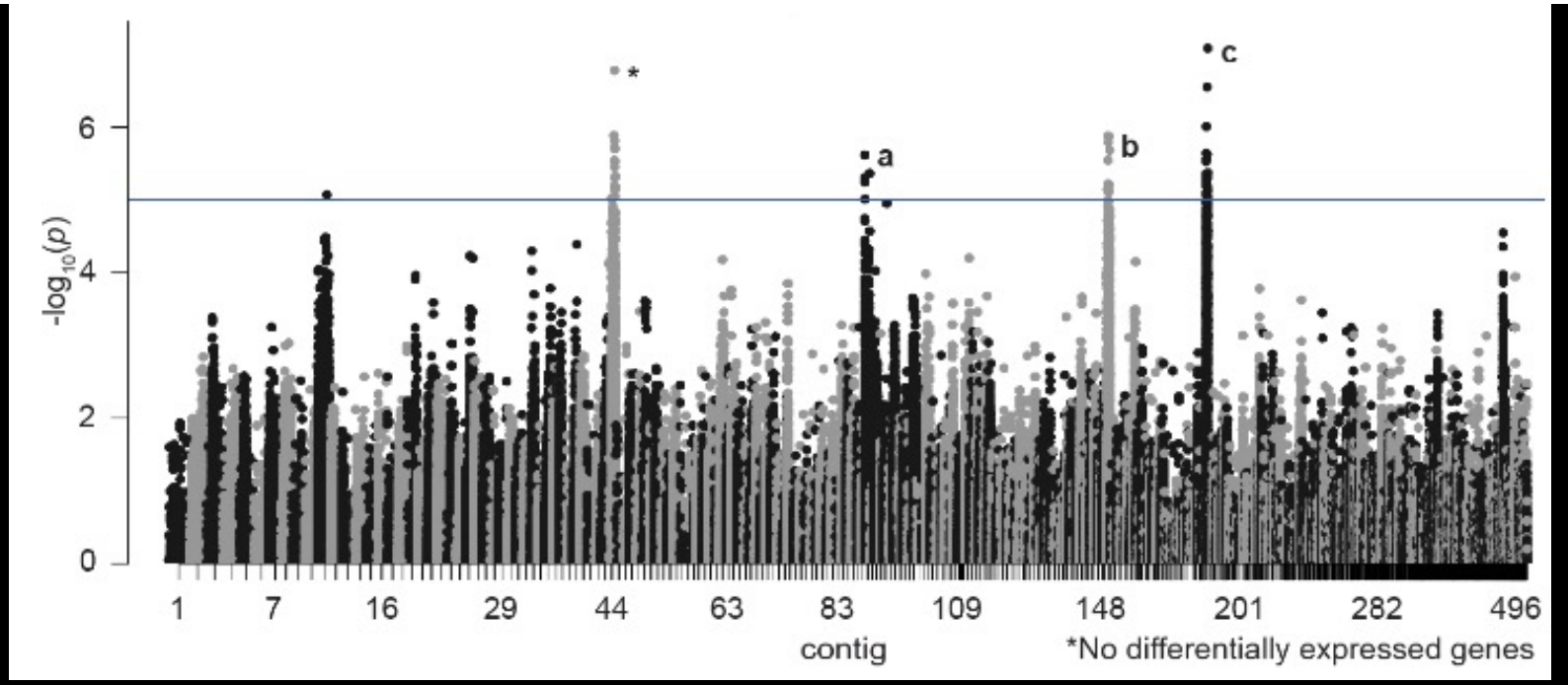
- Here researchers used
  - Experimental evolution to fix trait so they could map it
  - GWAS between the alternative lines of high vs. low trait
  - RNAseq between the alternative lines of high vs. low trait
  - CRISPR-Cas gene KO to test candidate genes

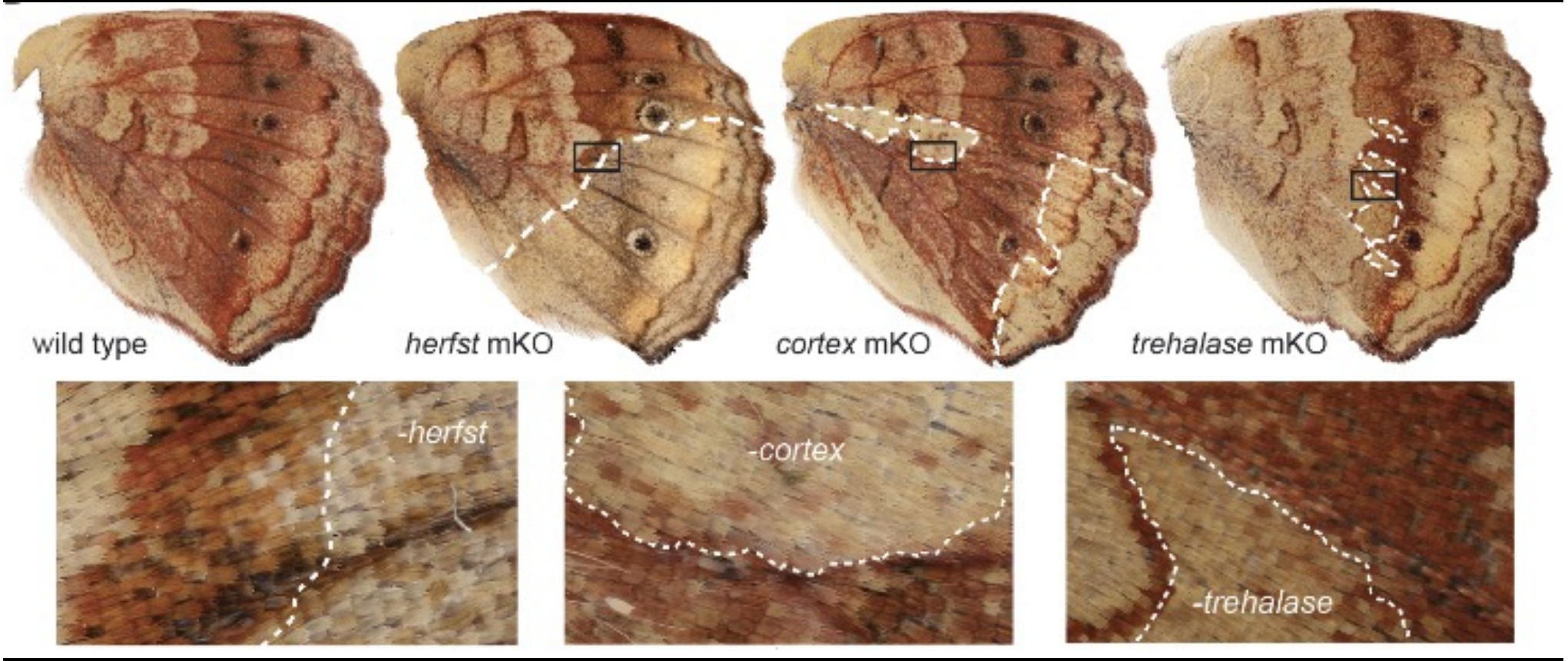
Genomic architecture of a genetically assimilated seasonal color pattern

- Made selection line having no plastic response
- Crossed back to plastic line
- GWAS on offspring for plastic response

Burg et al. 2020. Science.




*
a
b
c
6
4
2
0
-log10(p)
1
7
16
29
44
63
83
109
148
201
282
496
contig
*No differentially expressed genes


wild type
herfst mKO
cortex mKO
trehalase mKO
-herfst
-cortex
-trehalase

Functional genomic study of phenotypic plasticity

- An integrated study identified several genes underlying a plastic phenotype
- Integration involved
  - Manipulation of trait using experimental evolution
  - Intersecting GWAS and RNAseq results
  - Functional validation using gene KOs
- Importantly
  - Investigated gene without annotation, found functional association, increased knowledge of phenotype for future studies

On the importance of functional validation

- P-values do not indicate effect size
- Genes likely do not function the way we image
- Organisms are gloriously more complex than we can imagine

Without functional validation, we let past glimpses of insight retard progress towards deeper understanding

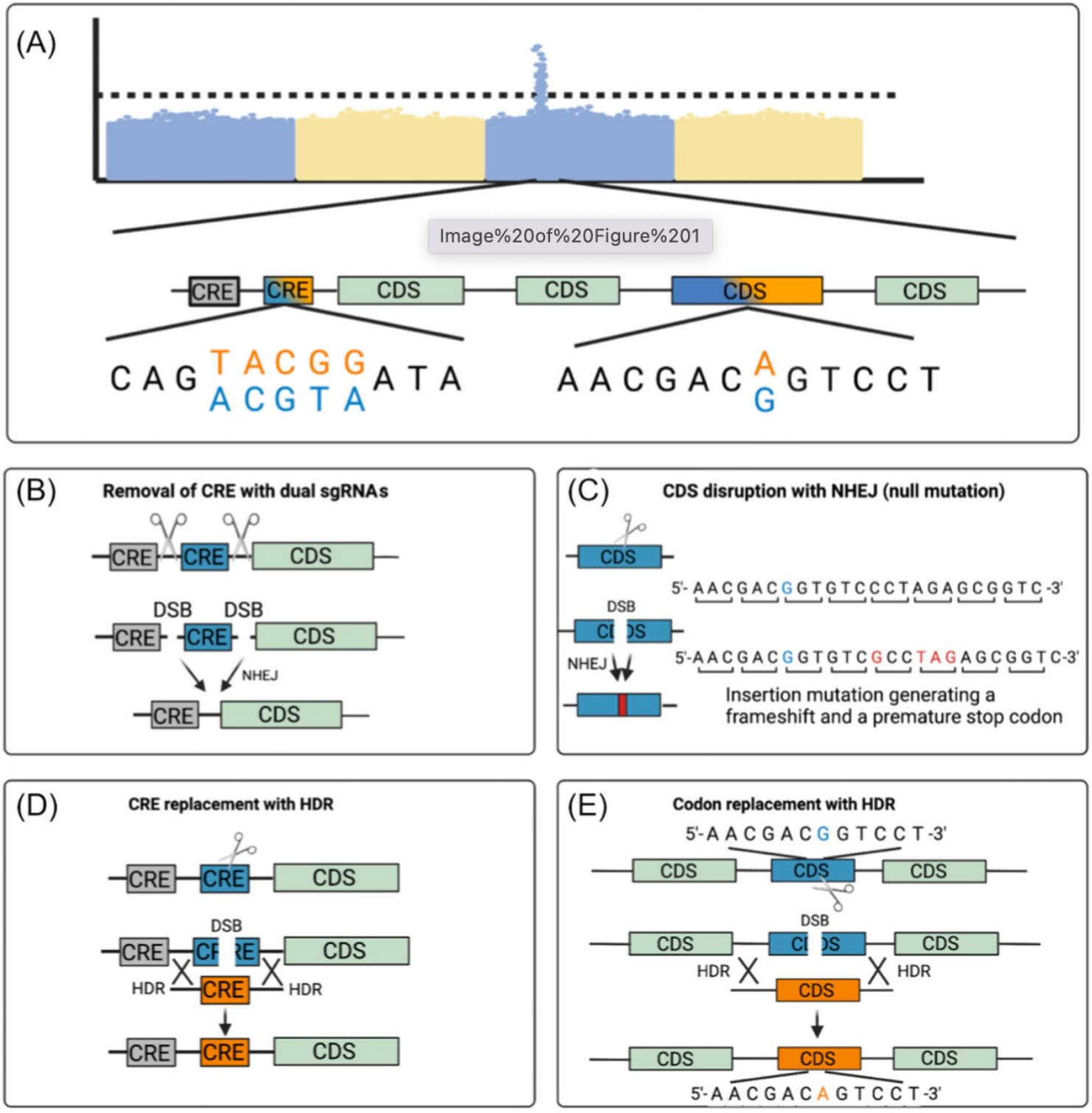
CellPress
OPEN ACCESS

Trends in
Ecology & Evolution

Review

Functional genomic tools for emerging model species

Erik Gudmunds,[1,*] Christopher W. Wheat,[2] Abderrahman Khila,[1,3] and Arild Husby[1,*]


(A)
Image%20of%20Figure%201
CRE
CRE
CDS
CDS
CDS
CDS
CAG TACGG ATA
ACGTA
AACGAC A GTCCT
G
(B)
Removal of CRE with dual sgRNAs
CRE
CRE
CDS
DSB
DSB
CRE
CRE
CDS
NHEJ
CRE
CDS
(C)
CDS disruption with NHEJ (null mutation)
CDS
5'- AACGACGGTGTCCCTAGAGCGGTC -3'
DSB
NHEJ
5'-AACGACGGTGTCGCCTAGAGCGGTC-3'
Insertion mutation generating a
frameshift and a premature stop codon
(D)
CRE replacement with HDR
CRE
CRE
CDS
DSB
CRE
CDS
HDR
CRE
HDR
CRE
CRE
CDS
(E)
Codon replacement with HDR
5'-AACGACGGTCCT-3'
CDS
CDS
CDS
DSB
CDS
CDS
HDR
HDR
CDS
CDS
CDS
CDS
5'-AACGACAGTCCT-3'

Chicken Church!!!!

Chicken Church!!!!

Bioinformatic wisdom, pt. 1

- Expect errors and noise
  - Analysis results need many rounds of refinement
  - Invoke biological causes of results last
- 70% of your time will be troubleshooting
  - This is normal, keep a notebook, intermediate files
- Fear the new and shiny programs that will simplify your life
  - 80% of all new software will not be usable
    - Un-installable, no manual, no test examples, not repeatable

Cookbooking ...

- Google and AI are your friends
- Use them, but don't trust them ..
- Test what you use, then learn from it.

Keep good bioinformatic notes

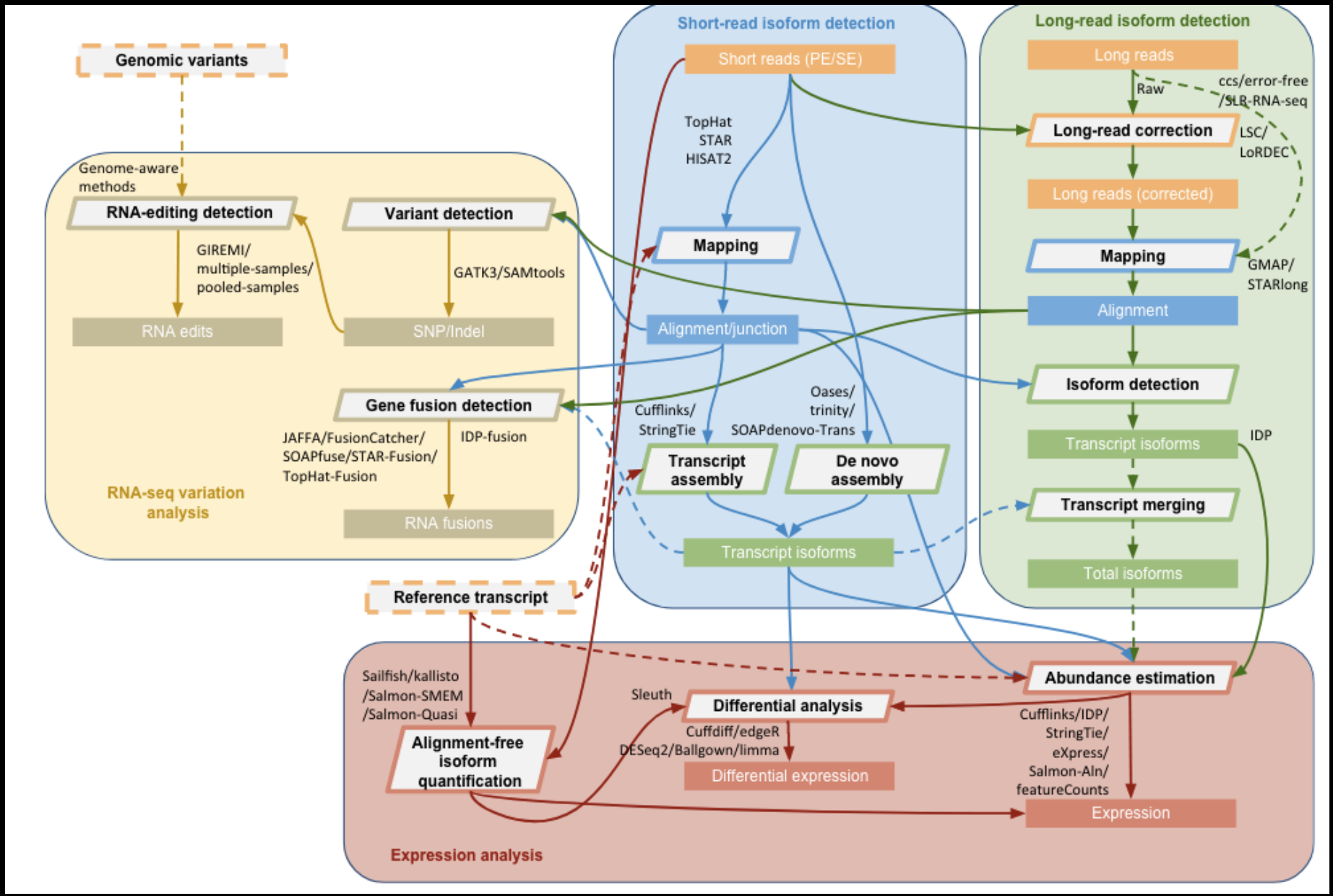
- I keep a special file with commands I learned and like
  - use it to quickly find commands, refresh memory
- Use positive and negative controls to test the output of the commands you run
  - I call these sanity checks
  - Always test to make code is working correctly
    - Great reason to use > 1 method, right?
- Read up on good file structure, version control, and how to parallelize your commands (Doug's lecture was awesome)

Publish your code, no matter how messy

Yours is without a doubt the worst code I've ever run

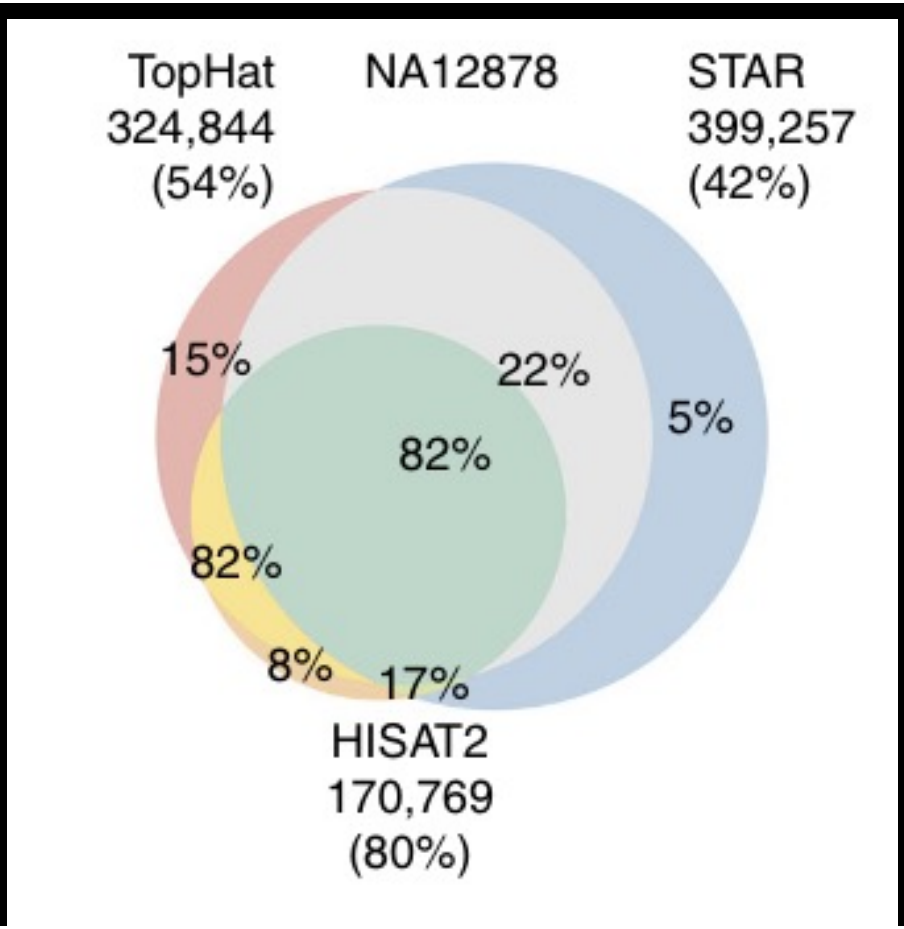
But it runs

Many different ways to make a pipeline

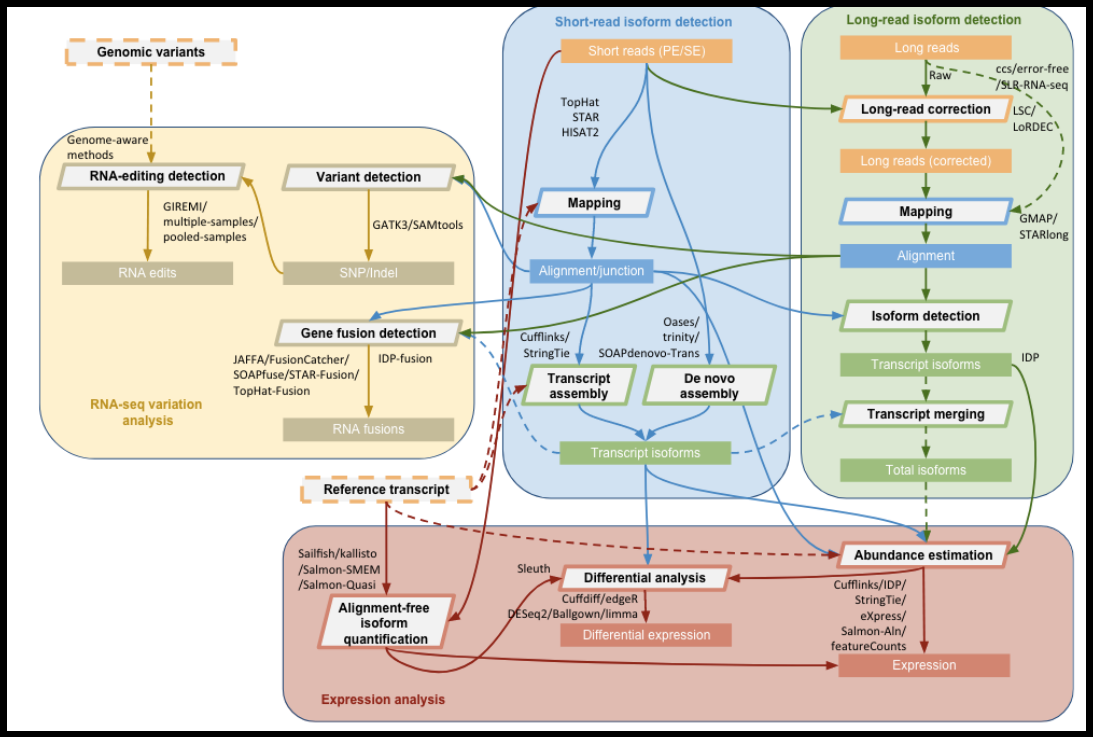


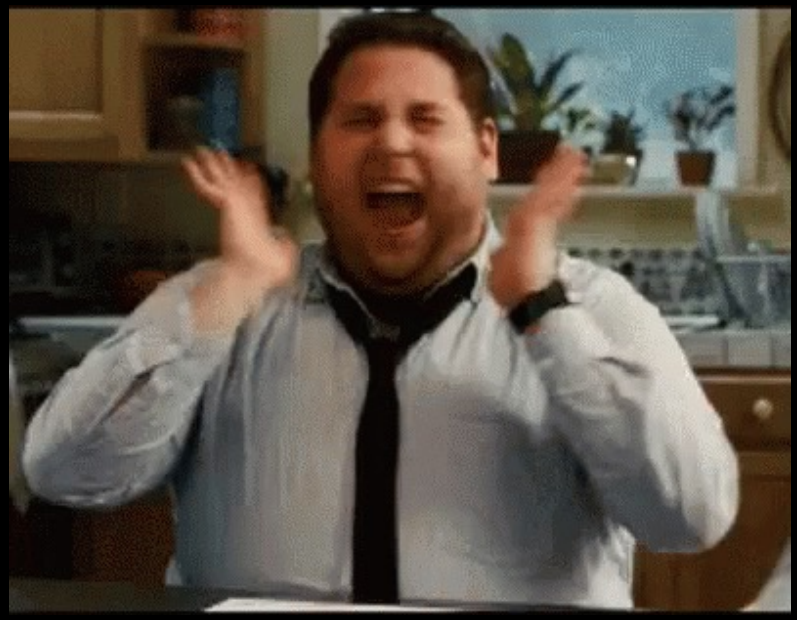
Sahraeian SME et al. 2017. Nat Commun.

Many ways to run, performance varies across species, samples, etc.



Analysis paralysis is common





Which is the right way?

- Just get through a single pipeline
- Then try different approach to assess your first results

Bioinformatic wisdom, pt. 2

- If all publications provided all their code, science would advance faster, with more accuracy
- Provide your code with all your publications, along with all your data. Be part of the solution.
- Look at others code:
  - Discover new ways of coding, reporting
  - Become frustrated that other published work is not repeatable
- If work is not reproducible, how much can we trust it?

Bioinformatic wisdom, pt. 3

- Data management
  - Get your raw data uploaded to ENA as soon as possible.
  - Its a free backup and you can set embargo date
    - keep pushing the date on the embargo
- Reproducibility is super important
  - Know about Snakemake or Nextflow ... but
  - Be careful of how you invest your time, as some people will try to convince you to learn their pipeline ... that you use once ...
- Is the pipeline for
  - you, or others
  - A few, or many samples?

Get ready ...

Here come the genomes .... and all their glorious errors ...

- –Annotation
- –Gene alignment
- –Functional annotation

Get ready, here come the 1000^{n} genomes

GENOM

5k

An unprecedented opportunity for large scale errors?

Sequencing of Life

dying:

ships

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

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?
- How accurate do you need your genome to be?
  - Short term vs. long term goals?
  - Are these in conflict?

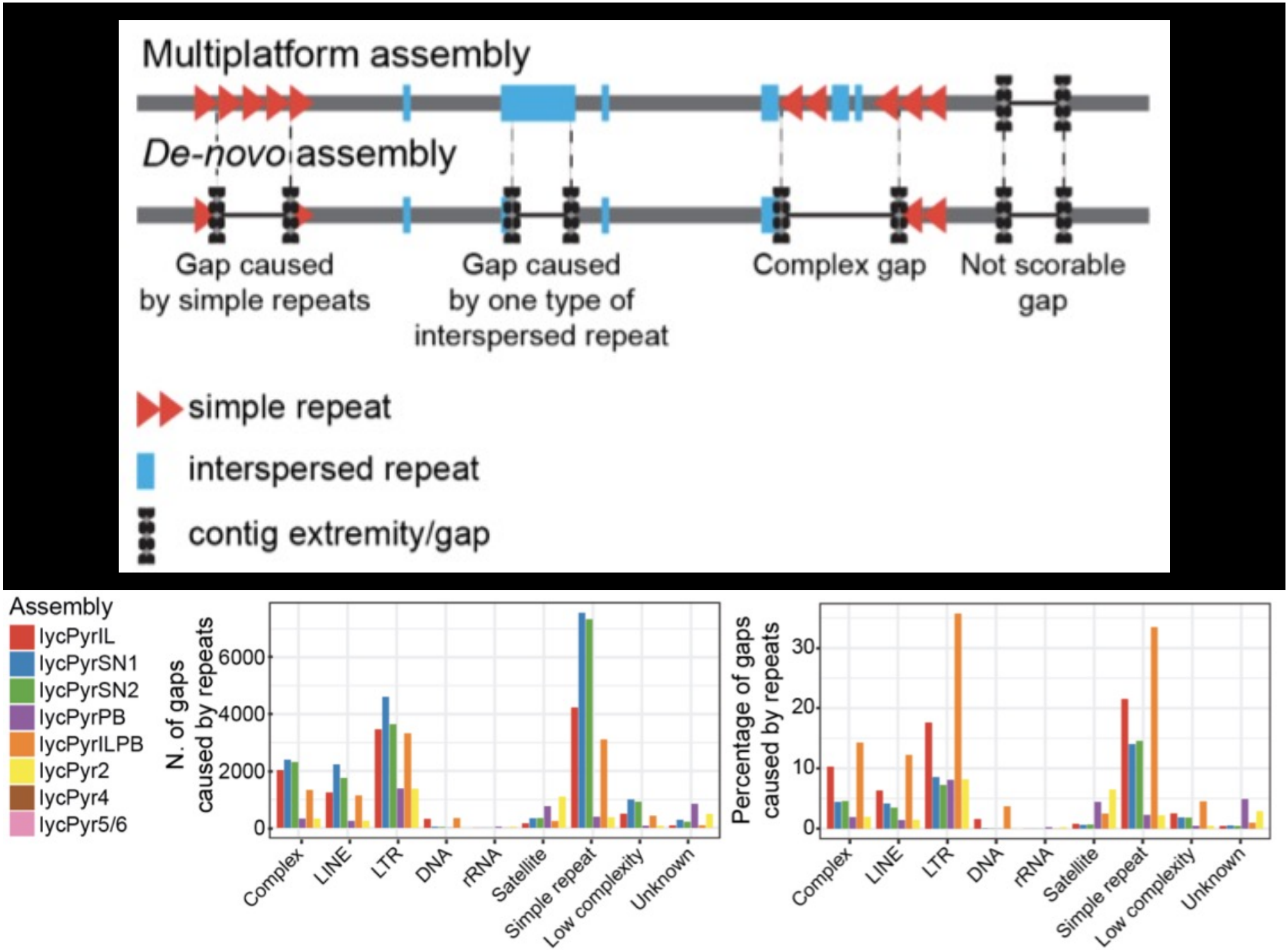
Identifying the causes and consequences of assembly gaps using a multiplatform genome assembly of a bird-of-paradise

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

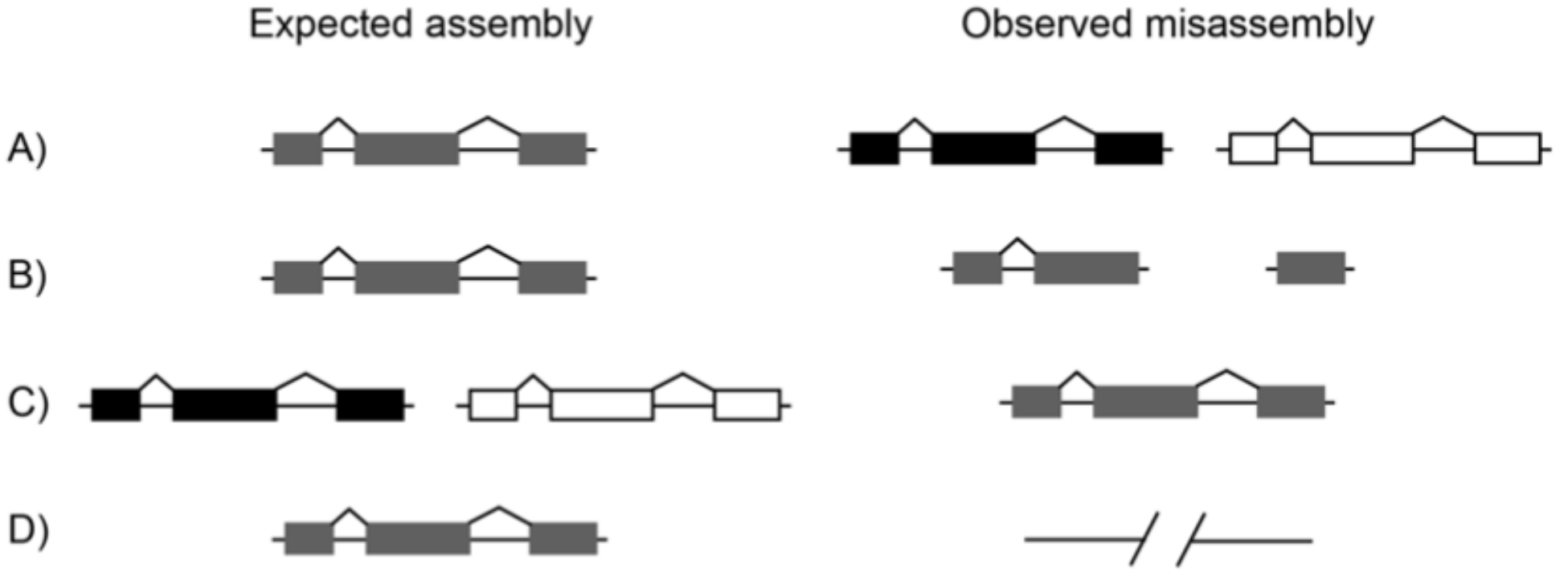
They made lots of assemblies along the way

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.


Multiplatform assembly
De-novo assembly
Gap caused by simple repeats
Gap caused by one type of interspersed repeat
Complex gap
Not scorable gap
simple repeat
interspersed repeat
contig extremity/gap
Assembly
IycPyrIL
IycPyrSN1
IycPyrSN2
IycPyrPB
IycPyrILPB
IycPyr2
IycPyr4
IycPyr5/6
N. of gaps caused by repeats
0
2000
4000
6000
Percentage of gaps caused by repeats
0
10
20
30
Complex
LINE
LTR
DNA
rRNA
Satellite
Simple repeat
Low complexity
Unknown

Errors that can happen in assemblies

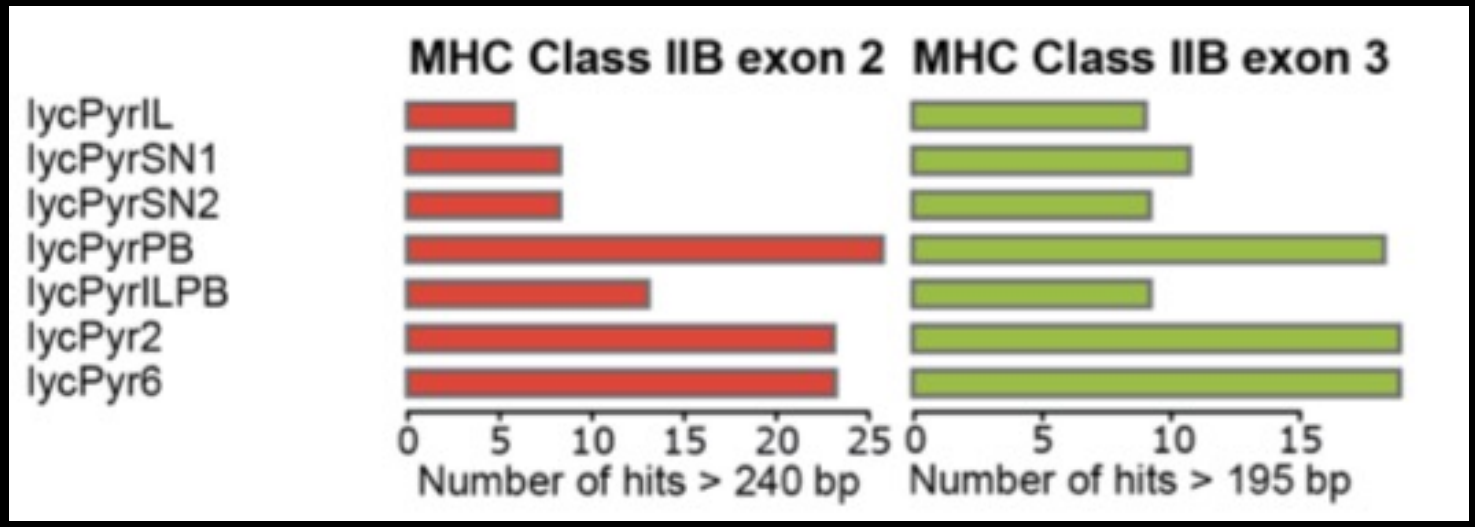


Genomes are scary and messy, especially when we re-assembly them with crude tools

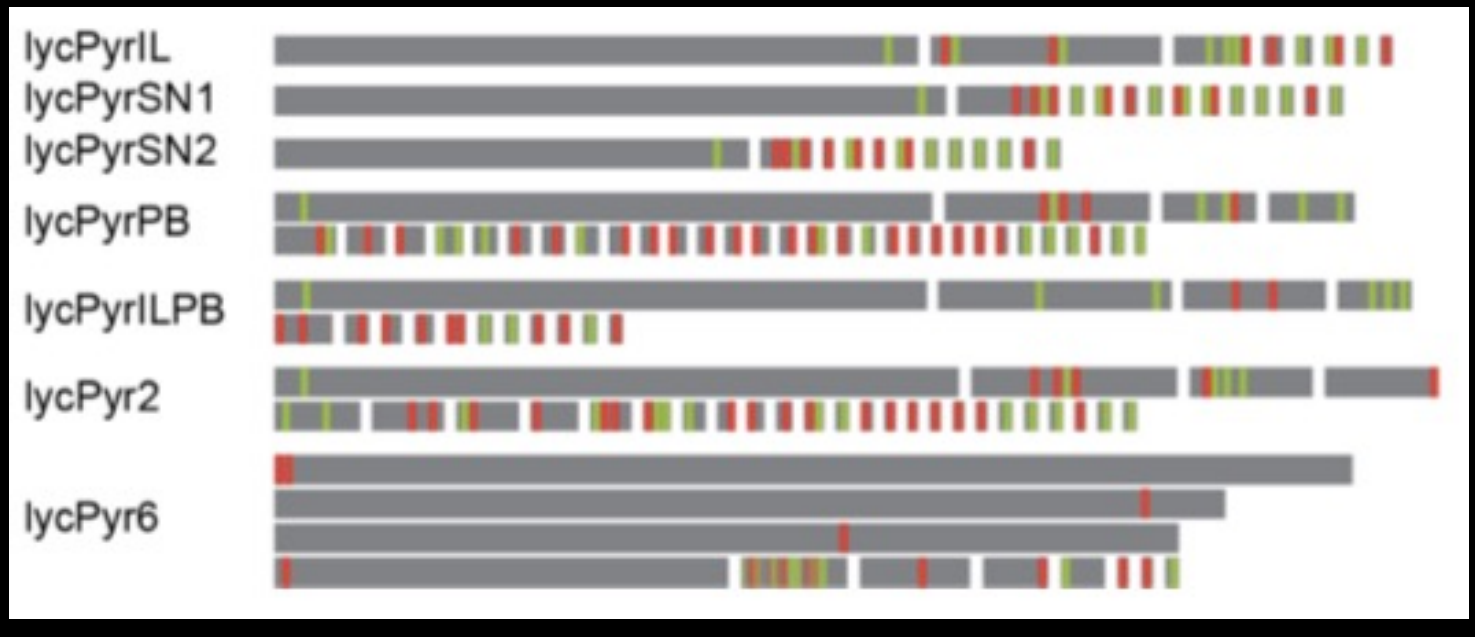
Denton et al. 2014 PLoS Comp Bio.

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"

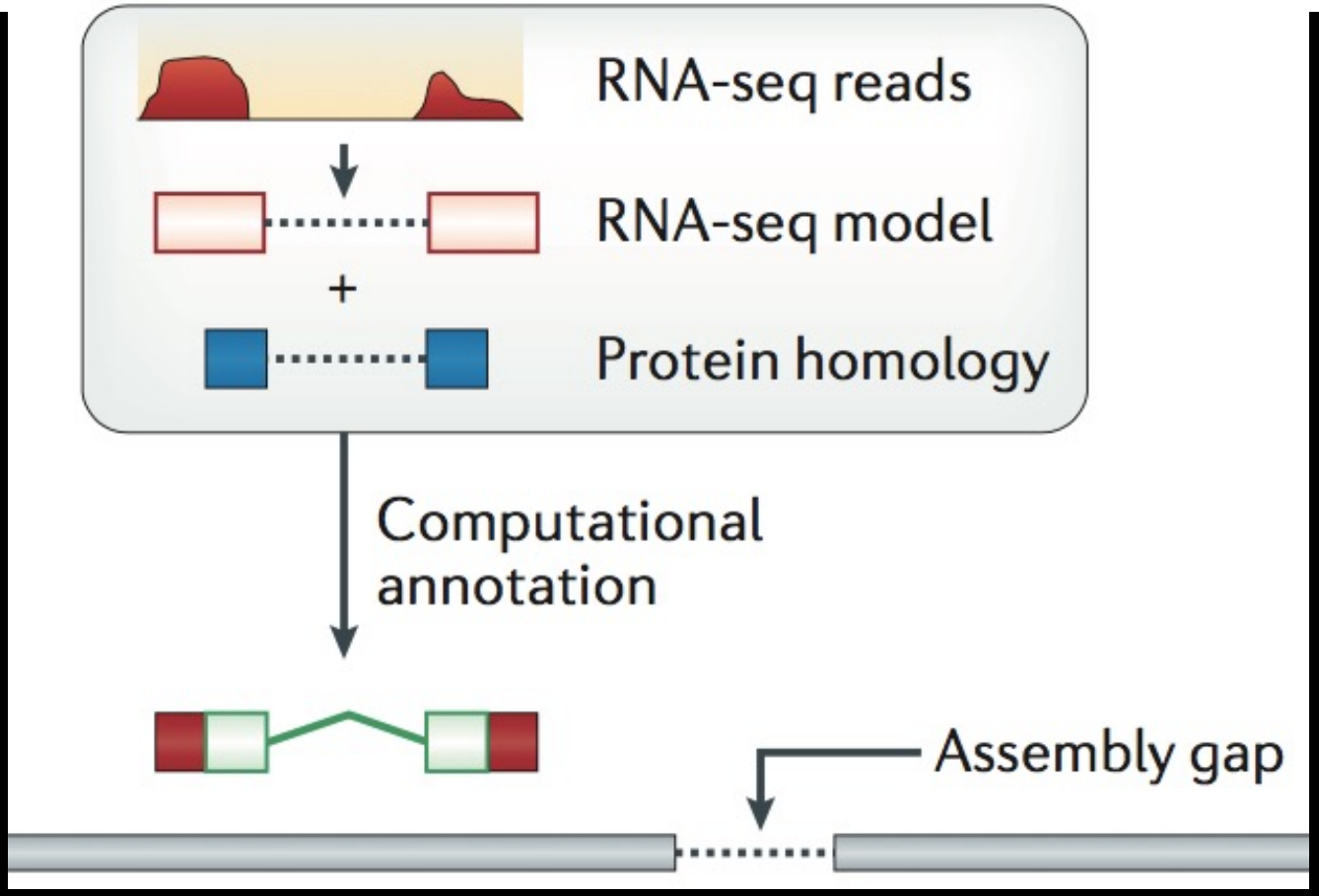
- Wheat 2015

For example:

- structural variants

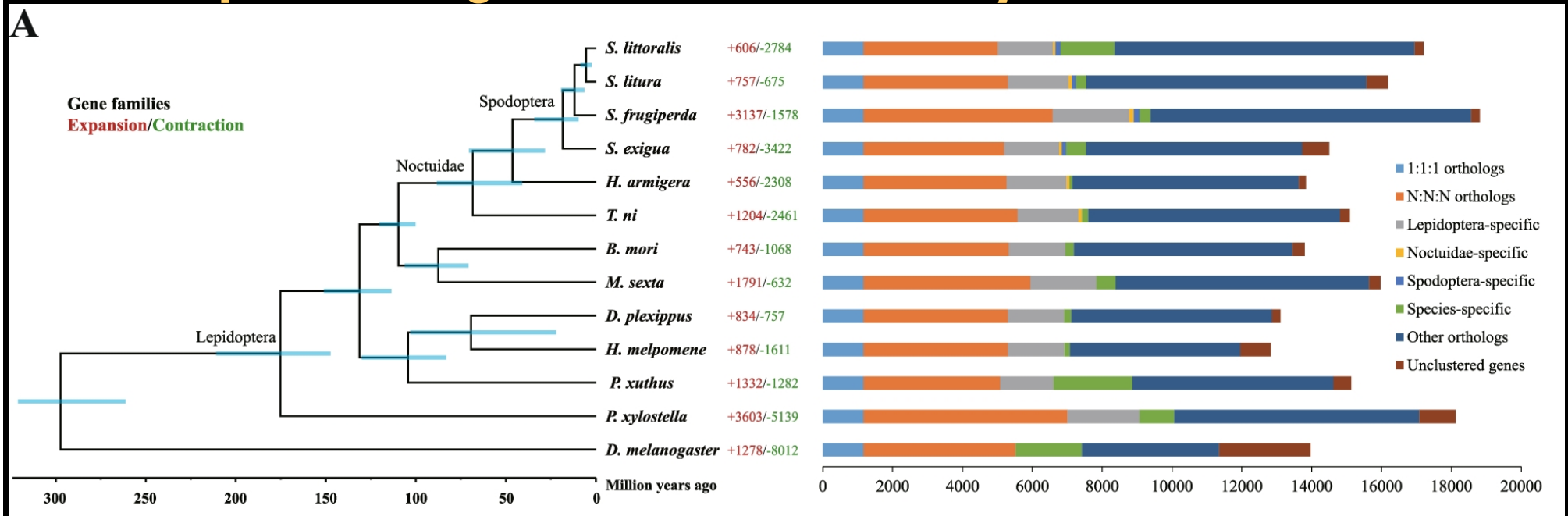
.. but revisit Evan Eichler's talk, there is hope for the future!

Genome annotation



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

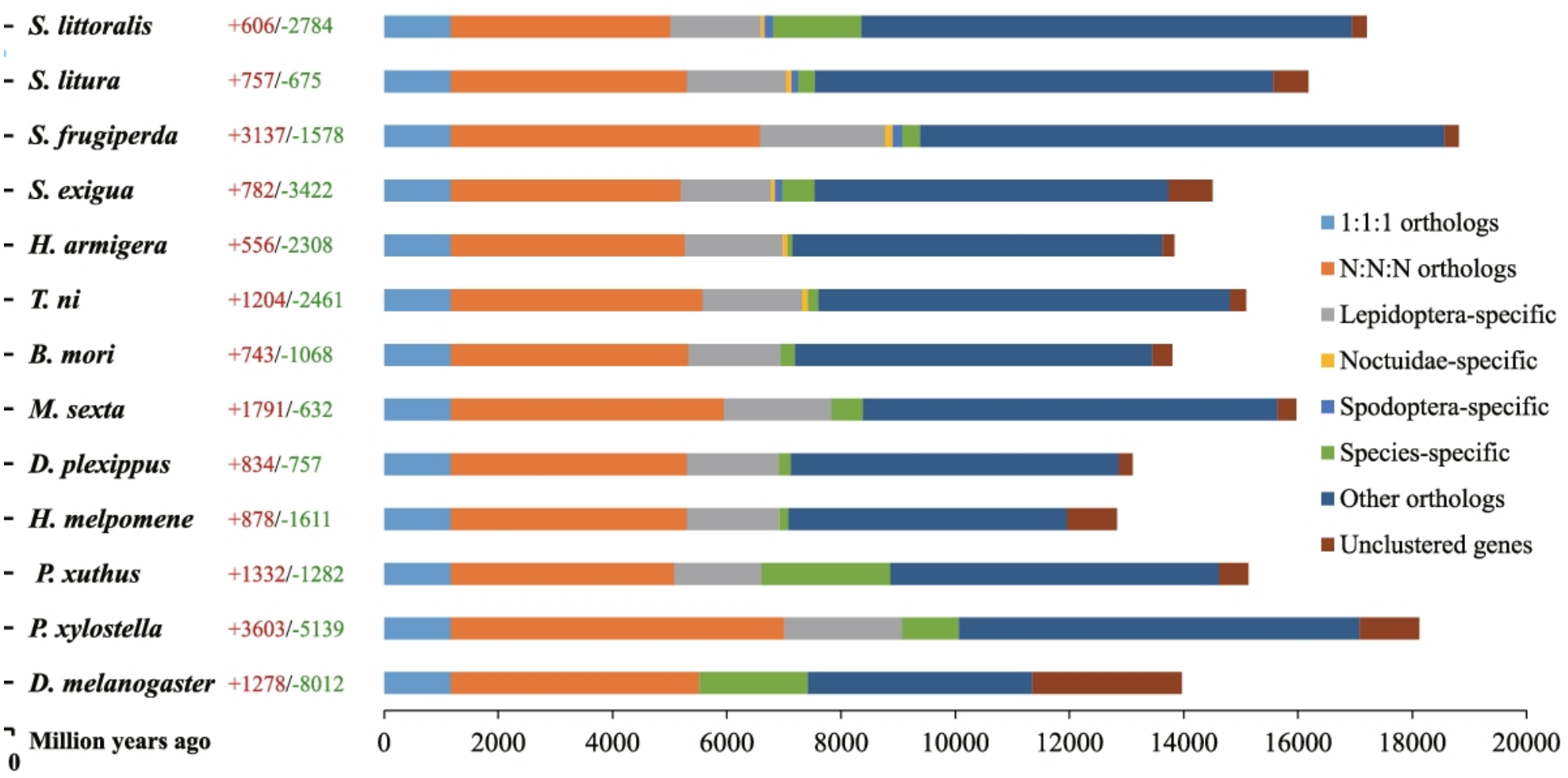
Comparative genomics commonly use annotations



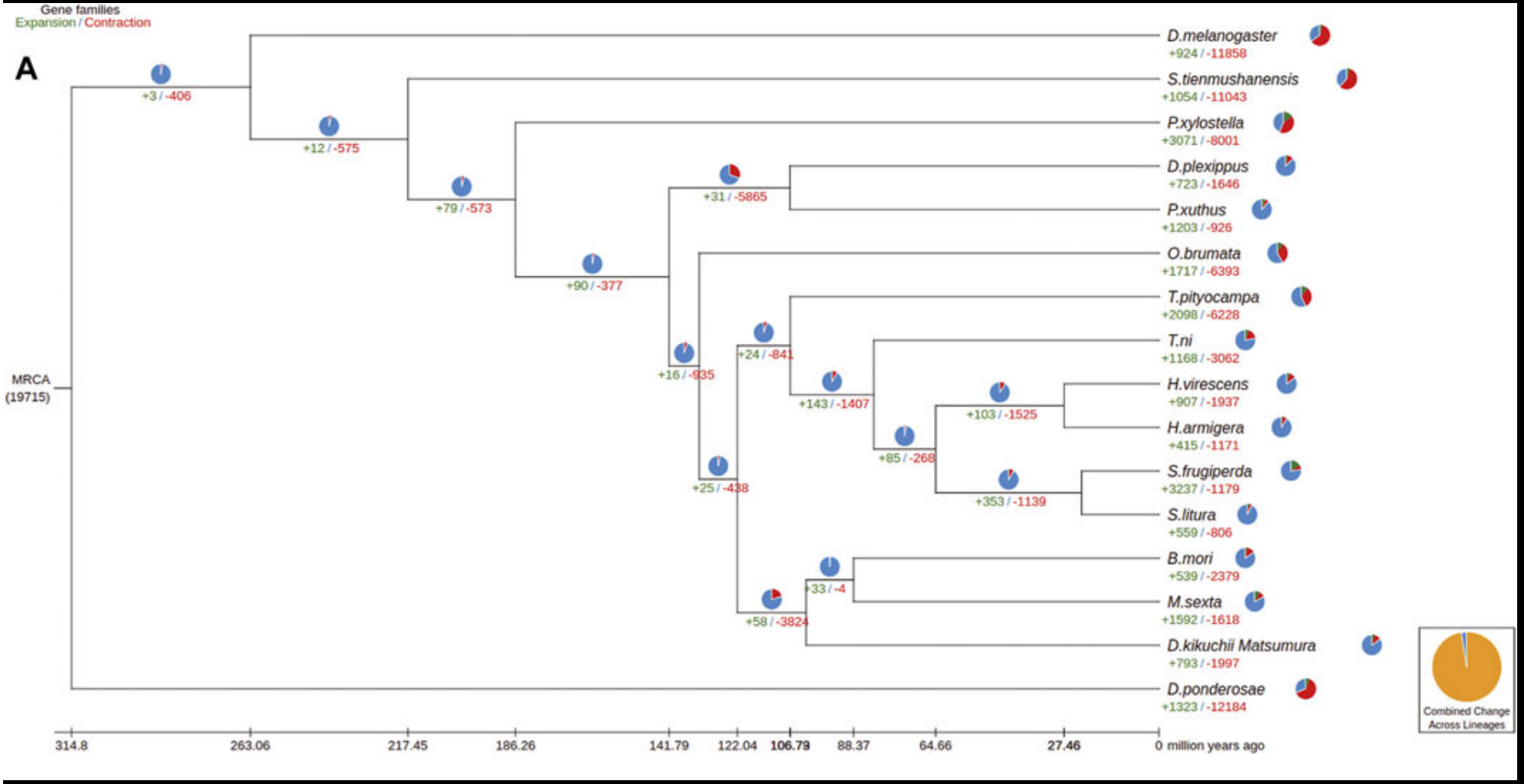
Typical genome report comparing gene content among species

- Rates of birth, death
- Lineage specific genes

Estimates of gene evolution rely upon good annotations

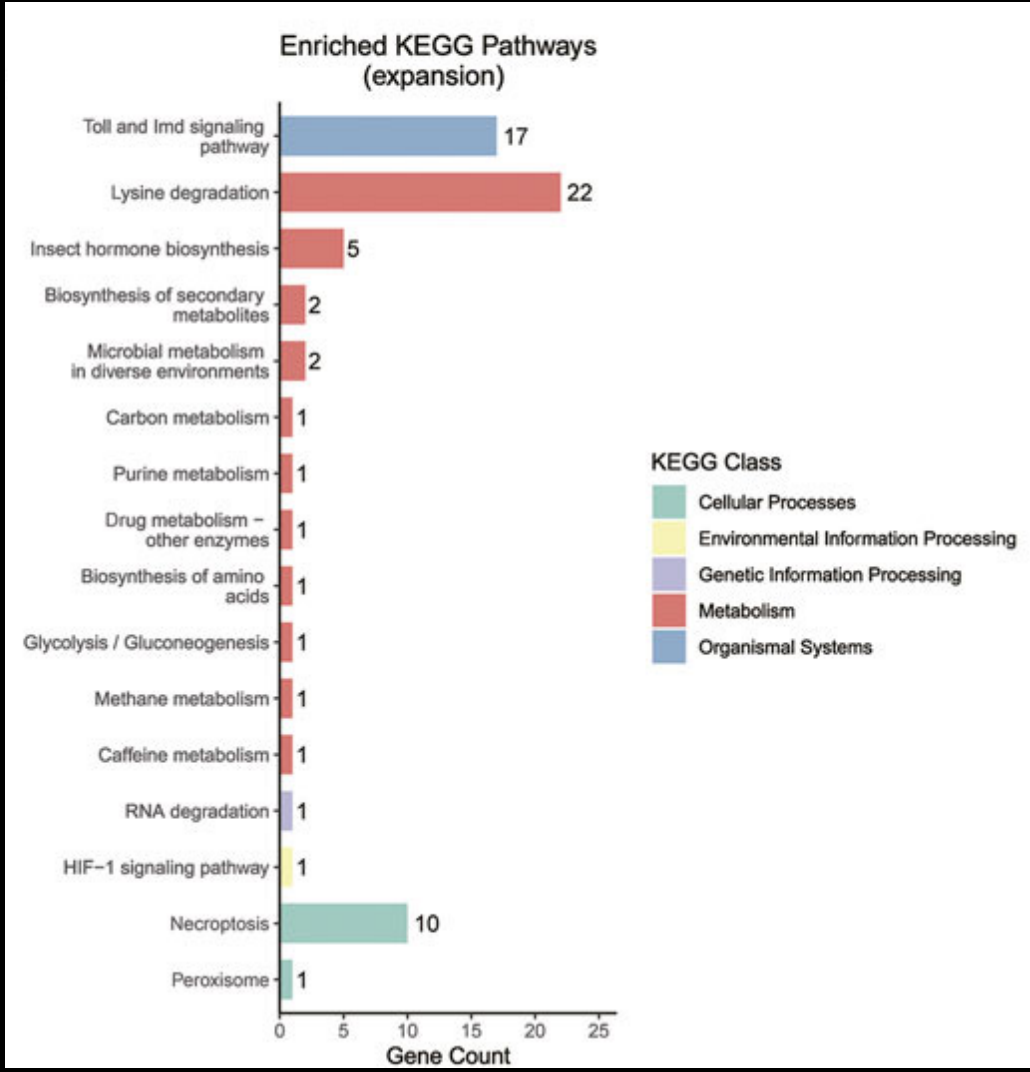


Gene birth-death dynamics



Gene birth-death dynamics

- Do changes in gene numbers have physiological meaning?
- Fundamental and important evolutionary question
- Very difficult to assess accurately
  - Need good genomes, annotations
  - Then good analyses



Are all annotations equal among species?

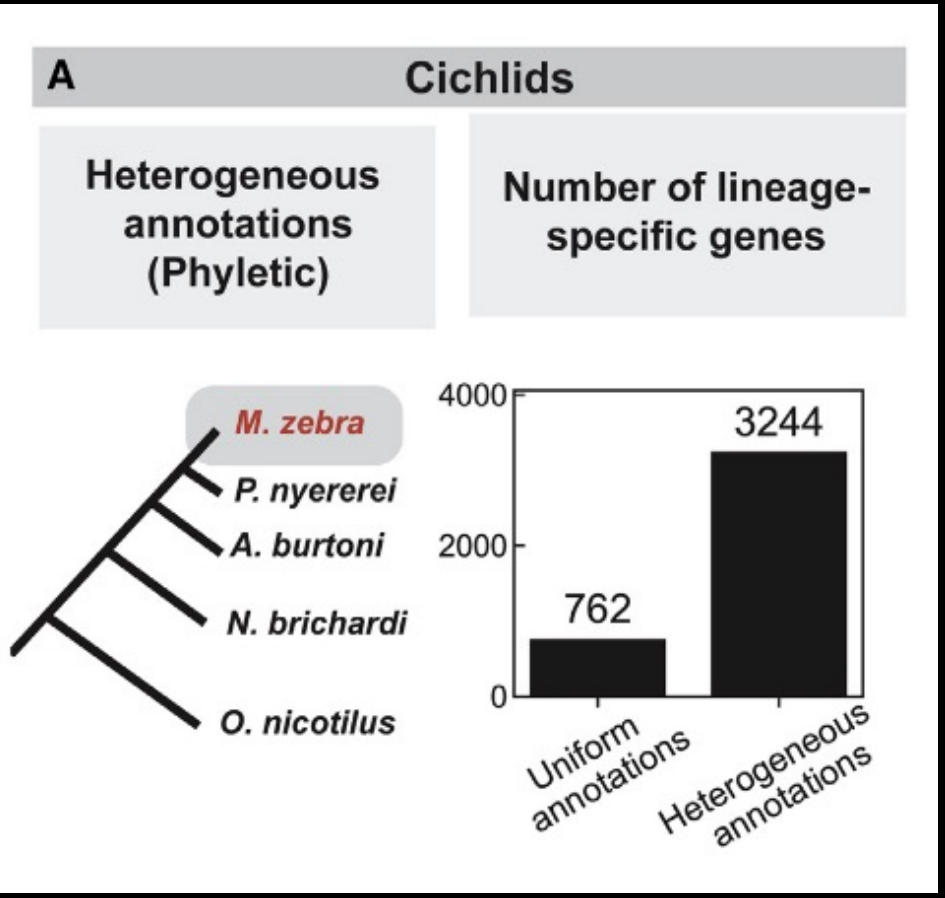
- Do species genomes differ in:
  - When they were sequenced, thus technology?
  - The quality of their assembly (e.g. N50, haploid state)?
  - How they did their annotation (proteins only vs. lots of RNAseq)?

Then resulting annotation protein sets likely differ due to technology, not biology

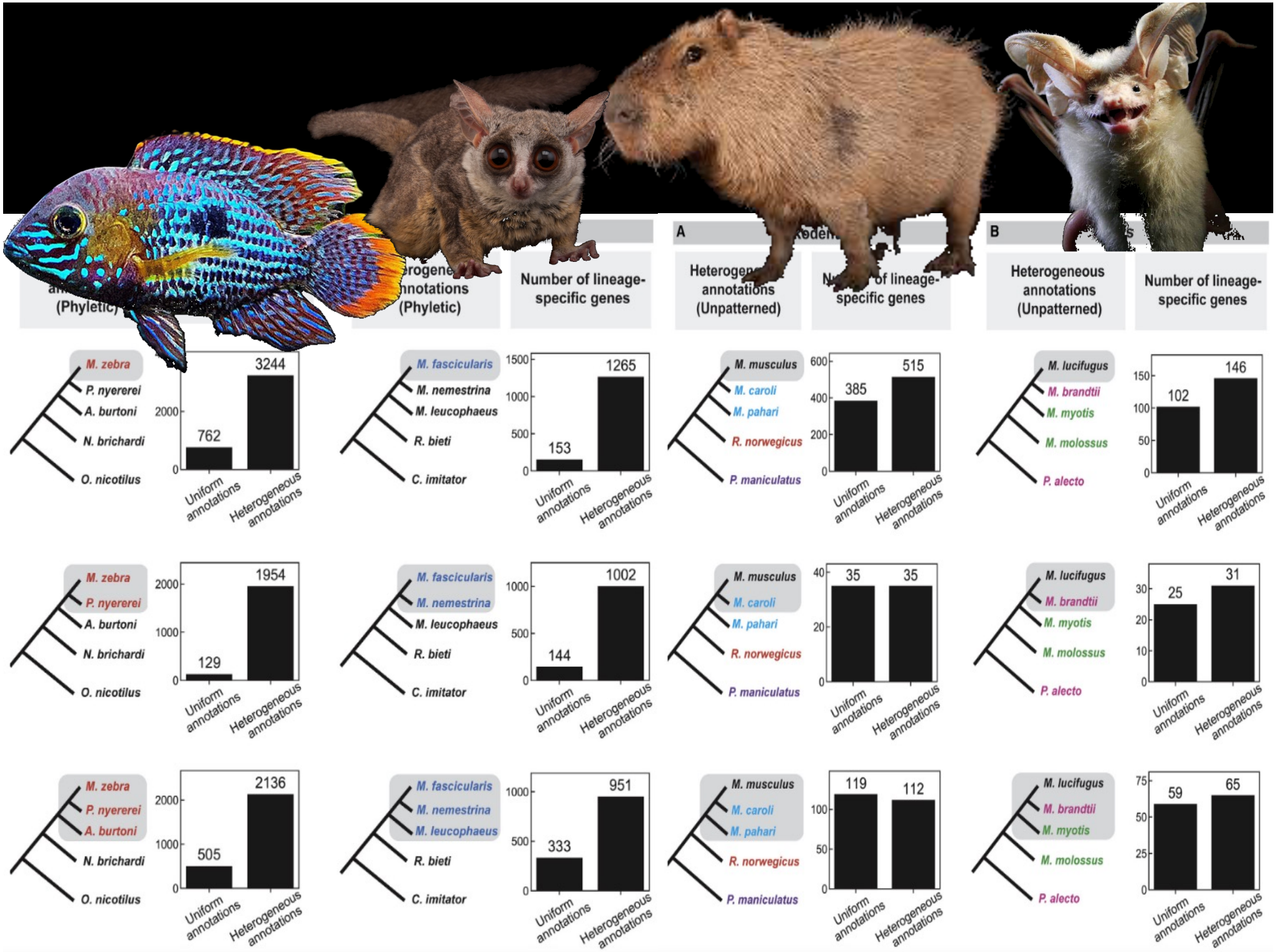
Will this impact analyses that rely upon accurate protein sets?

Non-standard annotations introduce major artifacts

- Lineage specific genes inflated by
  - 10 to 1000's of genes, with increases up to 15 fold



Weisman et al. 2022. Current Biology


(Phyletic)
Number of lineage-specific genes
annotations (Phyletic)
Number of lineage-specific genes
A
Heterogeneous annotations (Unpatterned)
Number of lineage-specific genes
B
Heterogeneous annotations (Unpatterned)
Number of lineage-specific genes
M. zebra
P. nyererei
A. burtoni
N. brichardi
O. nicotilus
762
3244
M. fascicularis
M. nemestrina
M. leucophaeus
R. bieti
C. imitator
153
1265
M. musculus
M. caroli
M. pahari
R. norwegicus
P. maniculatus
385
515
M. lucifugus
M. brandtii
M. myotis
M. molossus
P. alecto
102
146
129
1954
144
1002
35
35
25
31
505
2136
333
951
119
112
59
65
Uniform annotations
Heterogeneous annotations

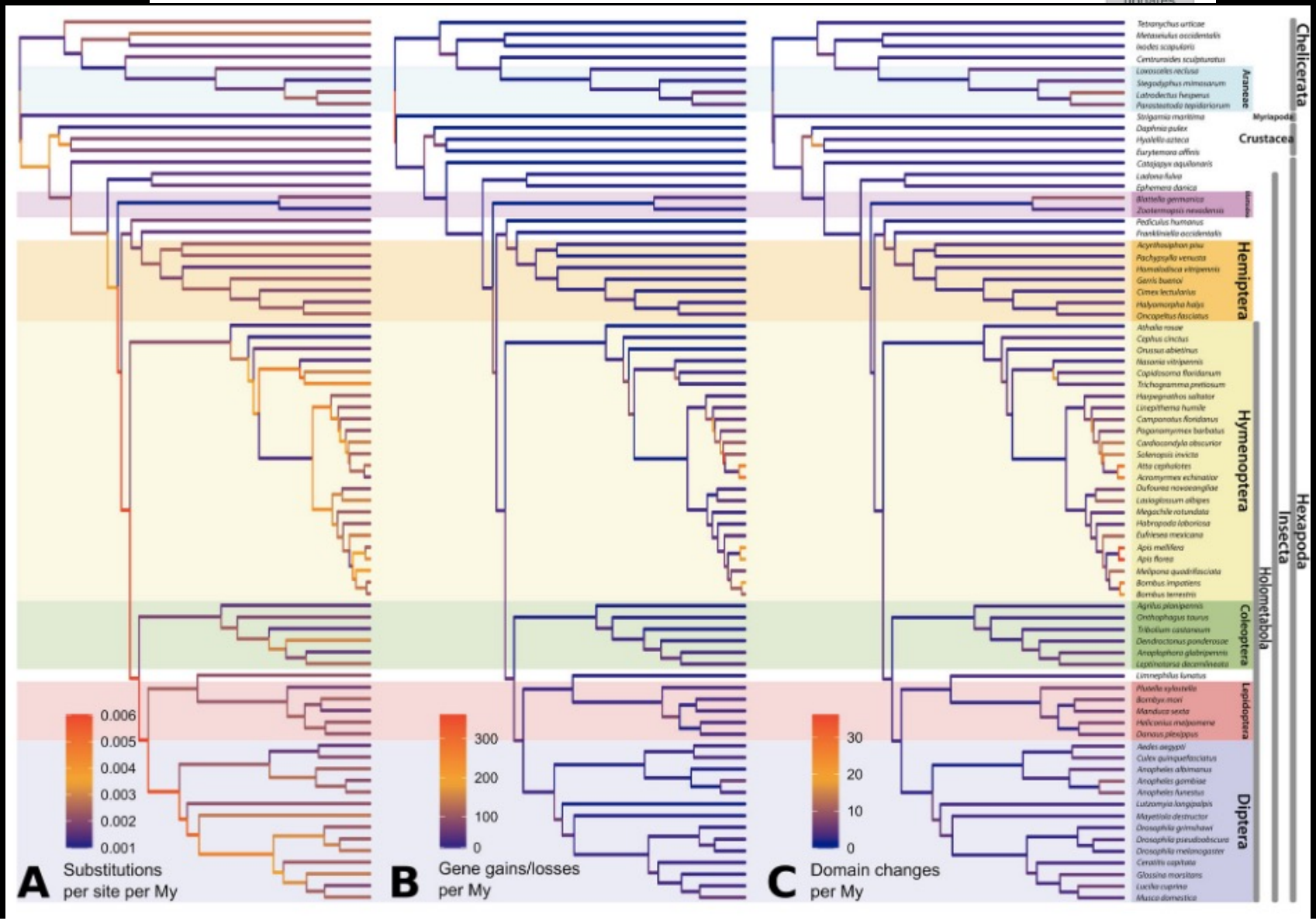

What are the ramifications?

Thomas *et al. Genome Biology* (2020) 21:15
https://doi.org/10.1186/s13059-019-1925-7

Genome Biology

RESEARCH Open Access

Gene content evolution in the arthropods



Some major conclusions of the paper

A

Last Insect Common Ancestor: 147 emergent gene families

Function	Emergent families
Wing morphogenesis	EOG86HJQQ EOG8TMTG9 EOG80ZTDS
Exoskeleton development and pigmentation	EOG8Q2GZG EOG8RZ1DS EOG8VDSCK EOG8WHC14 EOG8XPT03
Adaptation to terrestrial environment	EOG83XXJ1 EOG82VBZ4 EOG8PVRGC EOG8HTC7X
Larval behavior	EOG81K1SK

B

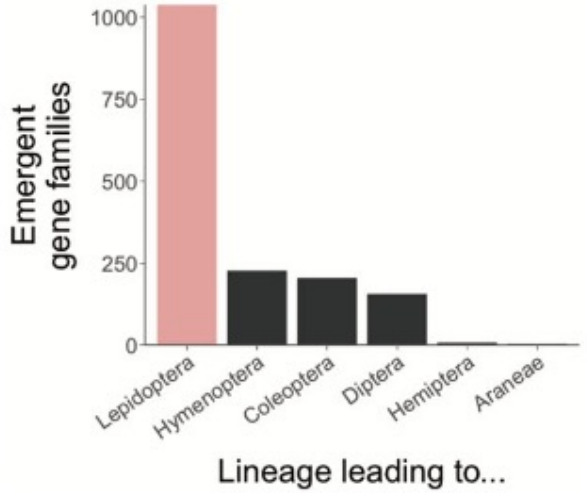
Last Holometabolous Common Ancestor: 10 emergent gene families

Function	Emergent families
Anterior head segmentation	EOG8HDW8X
Nucleosome assembly	EOG8G1PZD
Transporter activity	EOG847J8K
Transferase activity	EOG8ZPH98
Serine-type endopeptidase	EOG8QJV3F

+ 5 families with no known function

C

Last Lepidopteran common ancestor: 1,038 emergent gene families



"Although the majority of these gene sets were built using MAKER, variation in annotation pipelines and supporting data, introduce a potential source of technical gene content error in our analysis."

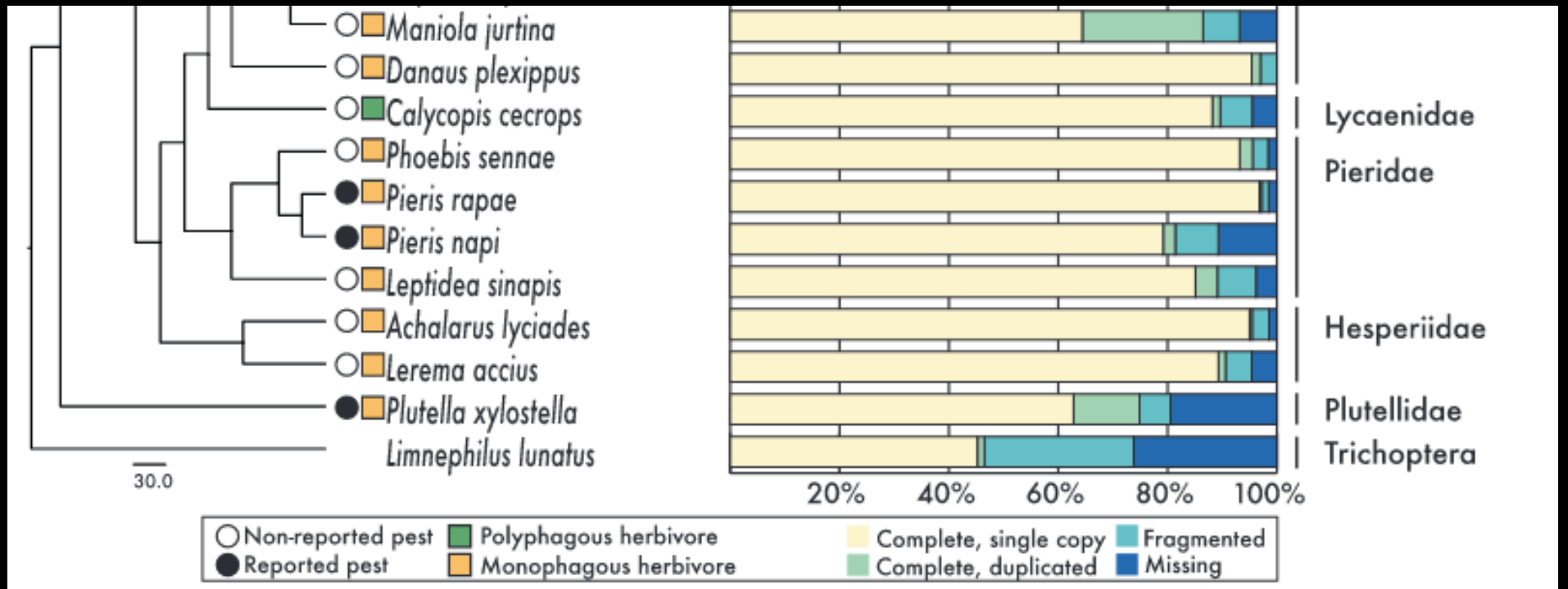
Proteins sets: a mixed bag of isoforms and pseudo-duplicates

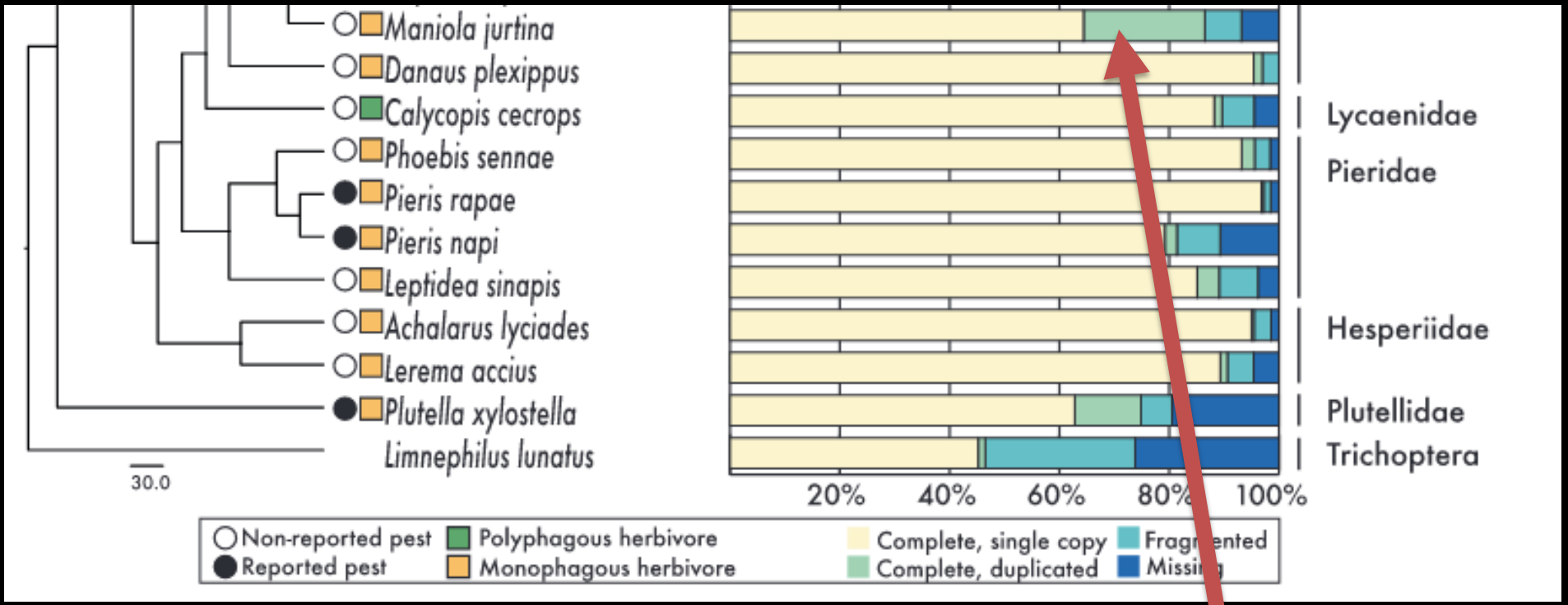
- Unfortunately, many studies are not isoform filtering their protein sets prior to analysis
  - Using raw protein sets from genome projects must always be filtered down to one protein per locus
  - This will have ramifications at all levels
    - Will severely impact ortholog assessments, gene birth death analysis

- Some genomes are not properly haploidified
  - Causes a pseudo-inflation of predicted genes
  - Creates artifacts in analyses

BUSCOs, when used properly, are very helpful

- Never report only complete BUSCO estimates
- Single copy and duplicated components are important
  - single copy indicates completeness
  - duplicated indicates haploid status
    - If not haploid, mapping your data to it will be very problematic





Species with nearly 2x gene content has high duplicated %

Family	Species	Predicted Genes
Nymphalidae	*Heliconius melpomene*	20075
	Heliconius erato lativitta	14613
	Heliconius erato demophoon	14517
	Junonia coenia	19234
	Melitaea cinxia	16667
	Bicyclus anynana	22642
	Maniola jurtina	36294
	Danaus plexippus	15130


THE
END
IS
NEAR

Put the BIO in your informatics!!

Use independent analyses as 'controls'

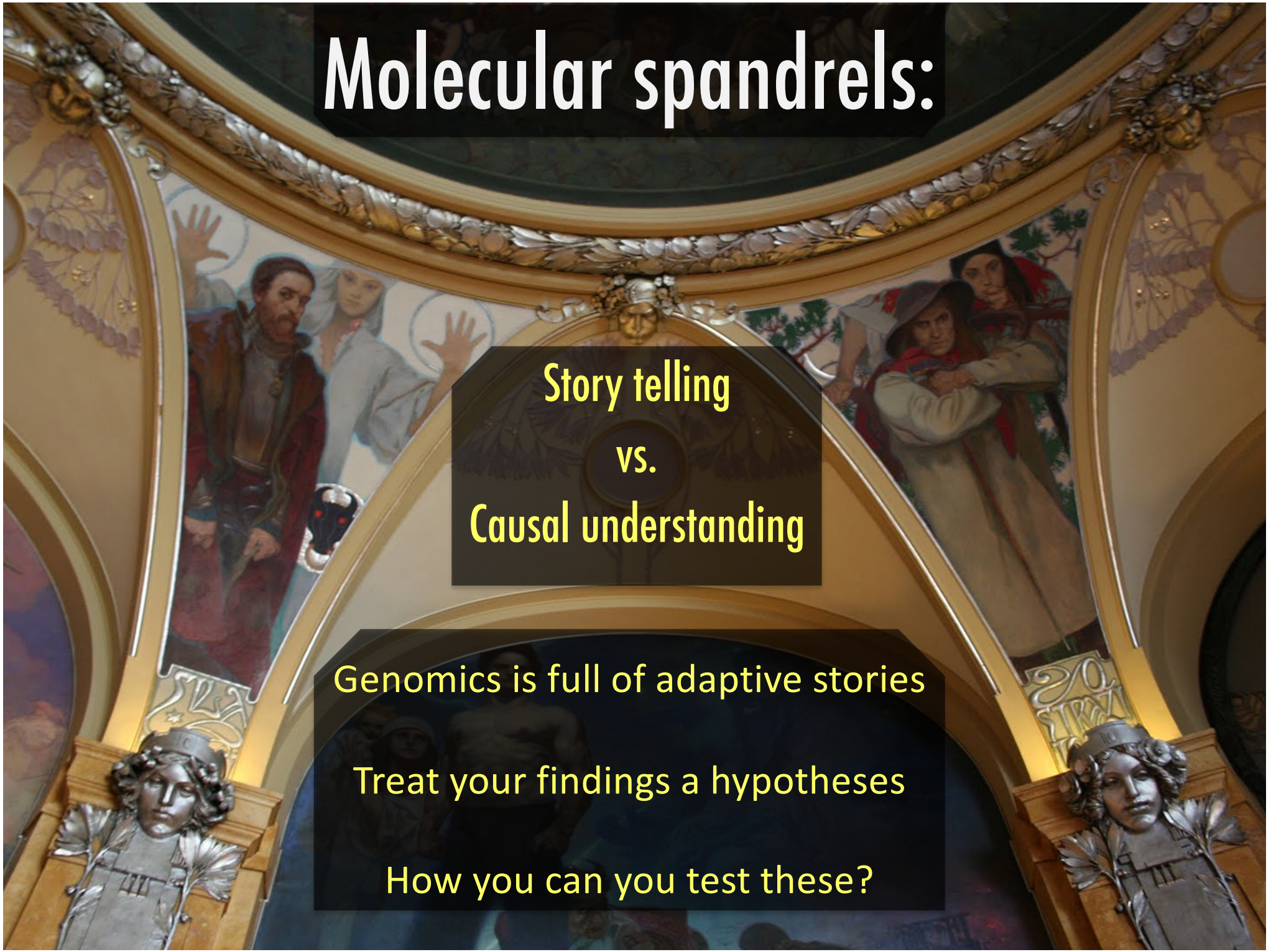
– What are your + and – controls?

	Analysis # 1	Analysis # 2	Analysis # 3
Mapper	HiSat2	Bwa-mem2	STAR
Normalization	none	TMM	TMM
Analysis	PCA	DEseq	EDGER

Should independent methods converge?

Interrogate your results

- "you need to be in charge of the analysis"
- The more you analyze your data, your confidence will grow
  - Let your findings talk to you in different ways
- Graph your results – visualize the patterns, assess 1st principals
  - Always start with PCA or MDS plot (how do your samples cluster?)
  - Compare with your different analysis results
- If you find interesting genes or patterns, can you test this hypothesis?
  - Using independent samples?
  - At a higher level of biological organization?
  - In some manipulative, 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

Come say hi if you're in town!

Stockholm University