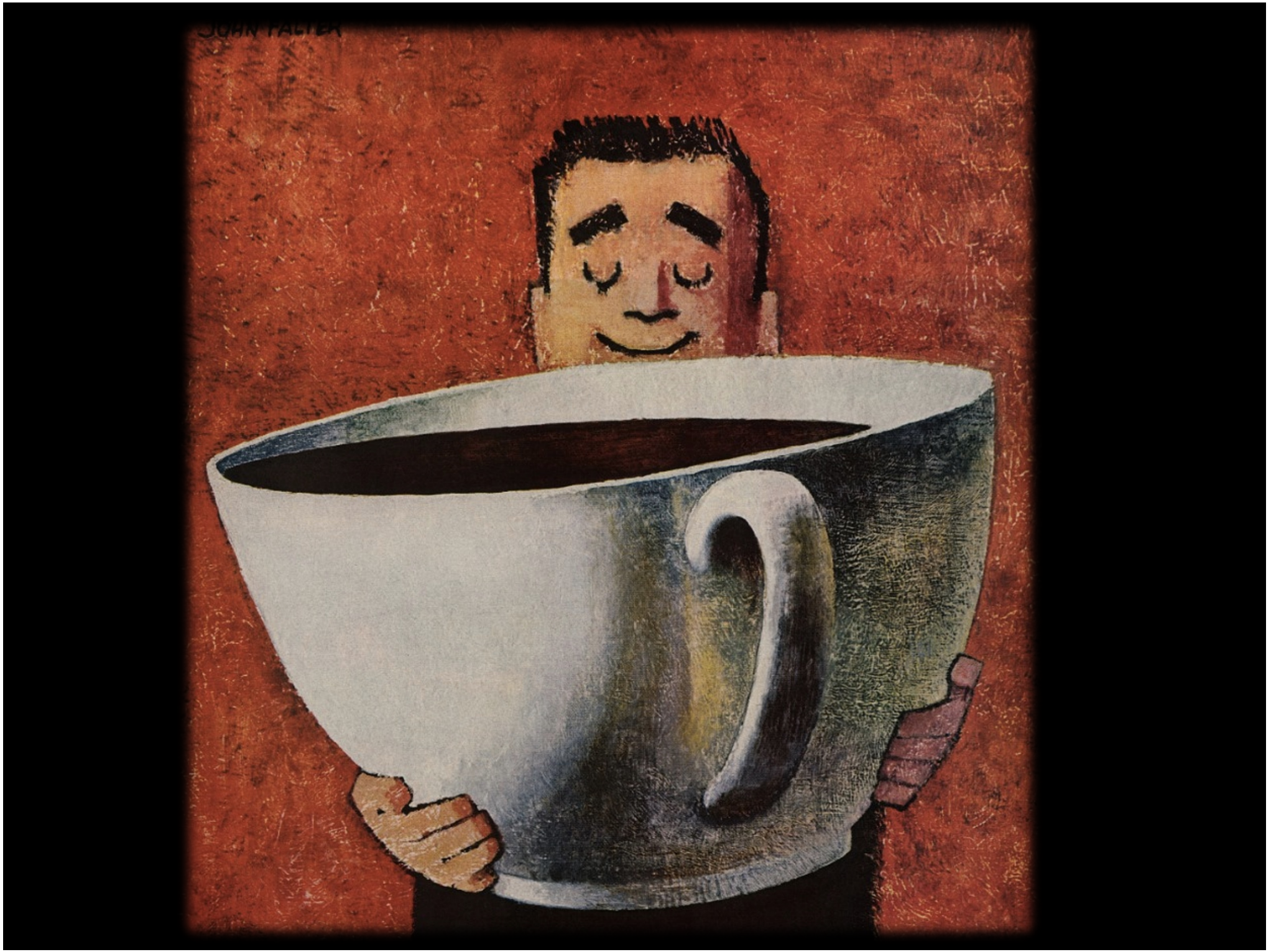

JOHN FALTER

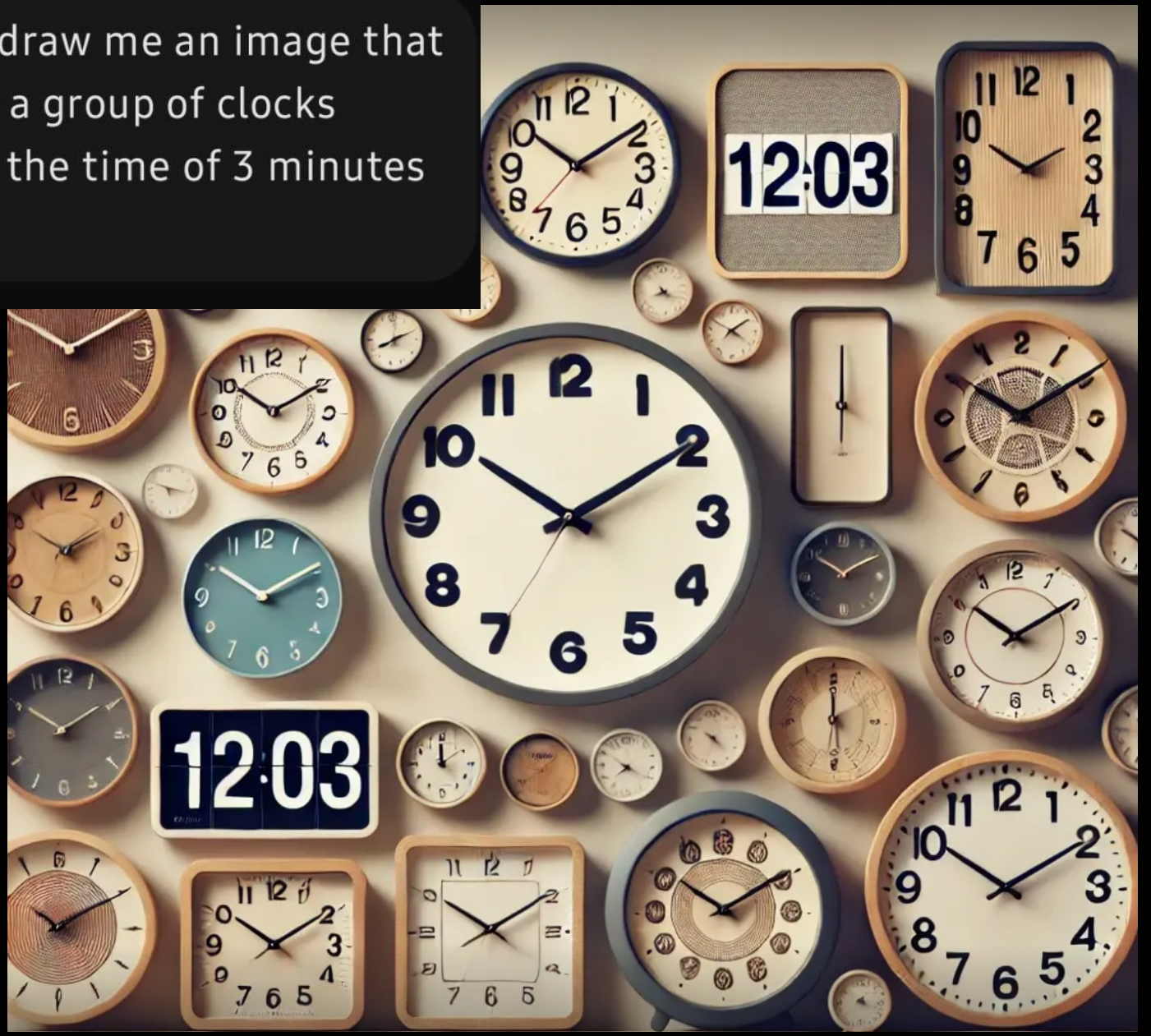
How many have used AI to help with their coding?

Where does the training set for these answers come from?

Does that matter?

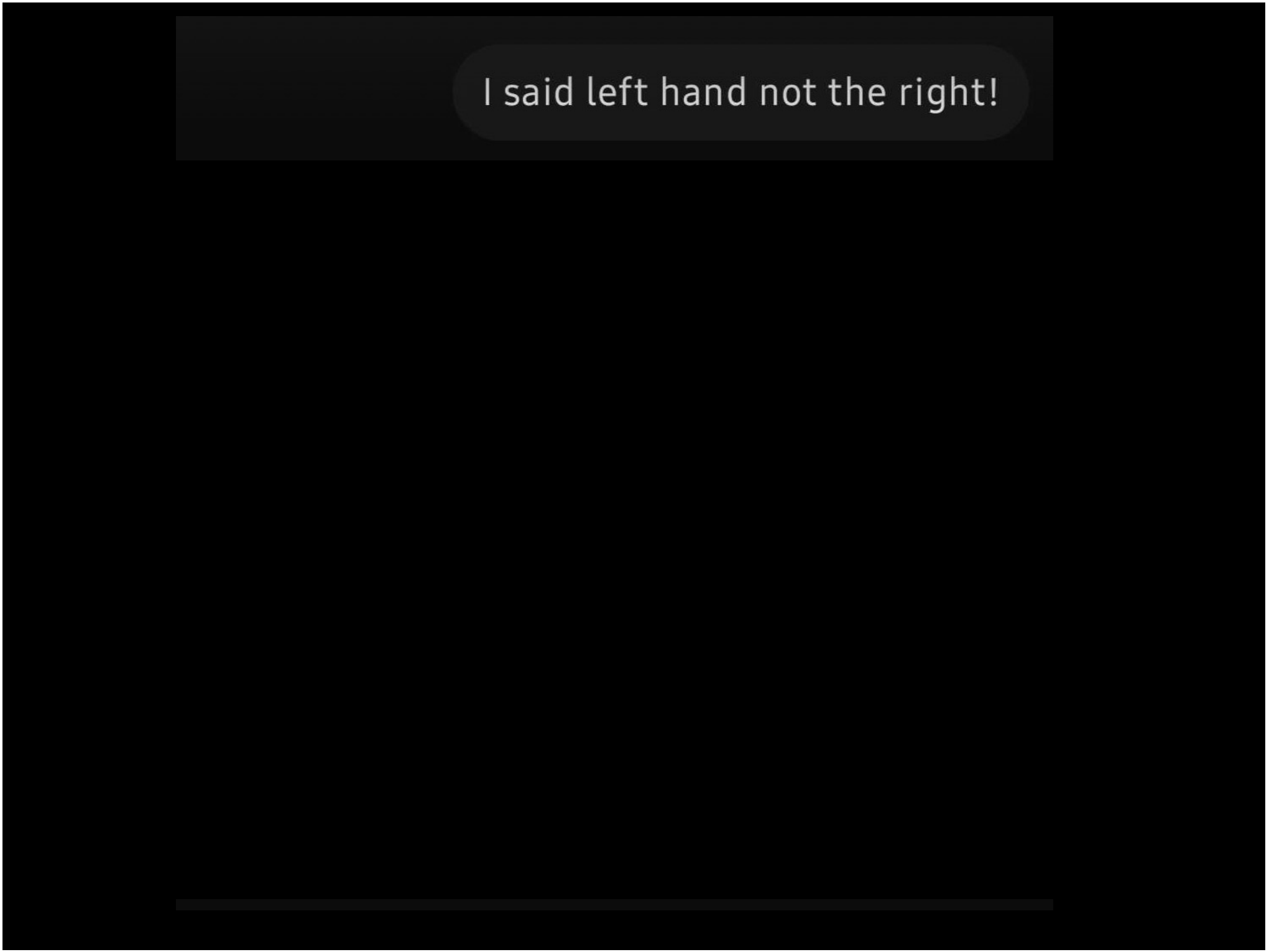
ChatGuyPT

Can you draw me an image that contains a group of clocks showing the time of 3 minutes past 12



ChatGuyPT

Draw me a picture of a person writing with their left hand.


I said left hand not the right!

Why is AI getting these simple things wrong ..

It's the training set

What species has the largest bioinformatics community?

Human biomedical studies drive the training set that inform all of your bioinformatic answers from ChapGPT

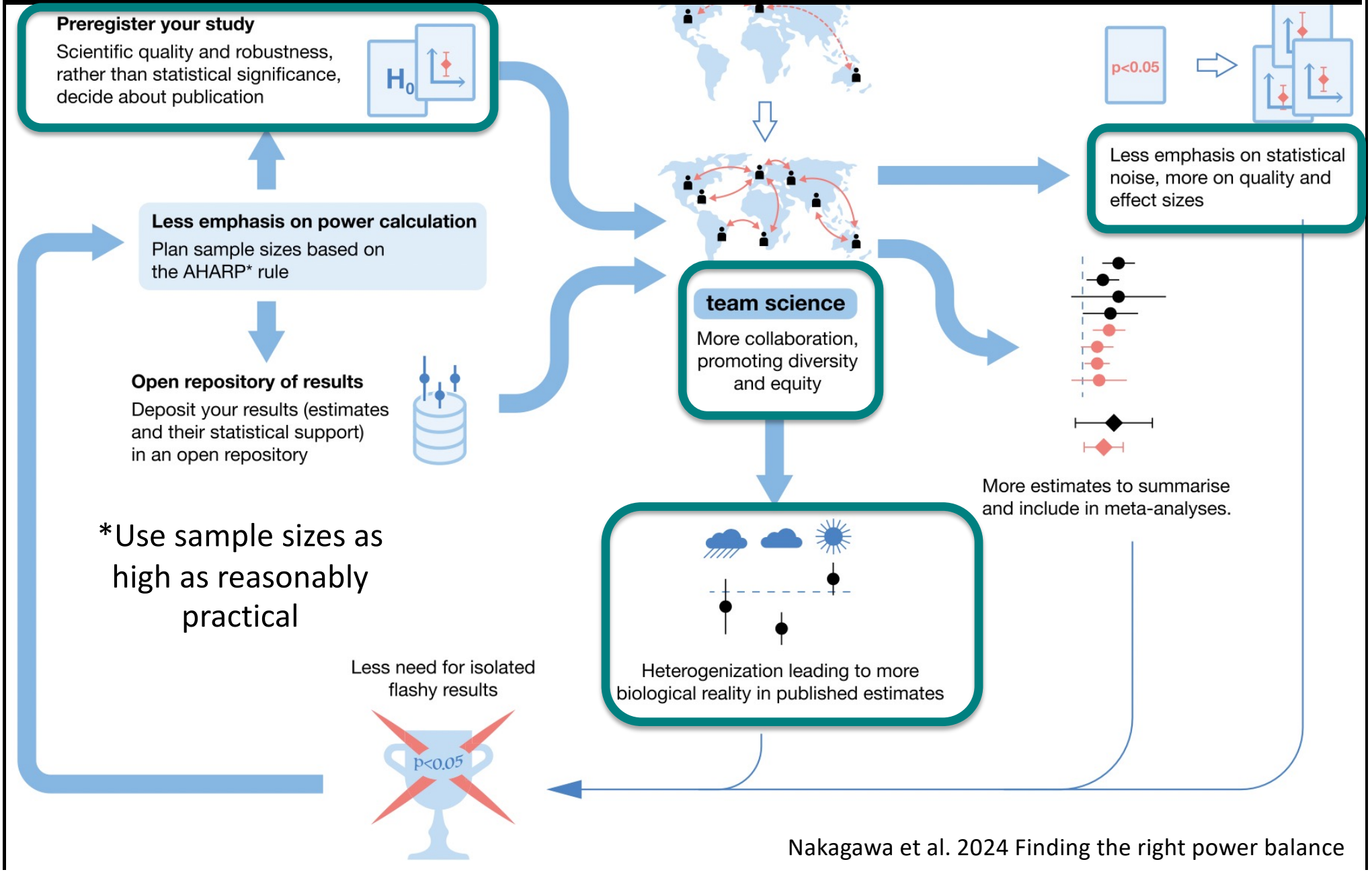
Use it, verify it works, modify for your species

This is a piece of toast

So .. how do we avoid Apophenia?

- Non-random patterns are abundant in genome scale data
  - We generally lack ability to calibrate our expectations
  - Null models, controls are very difficult to get "right"
- Double check your data and analyses
  - Plot your data, look at it, does it make sense on 1st principals?
- Test your hypotheses in independent ways
  - Genomics: independent datasets, independent analyses, across levels
    - Independent biological samples, GWAS vs. K-mer GWAS, mRNA vs. protein
  - Manipulation: functional validation via manipulation of genes, pathways
    - Experimental evolution, CRISPR KOs, environmental perturbations

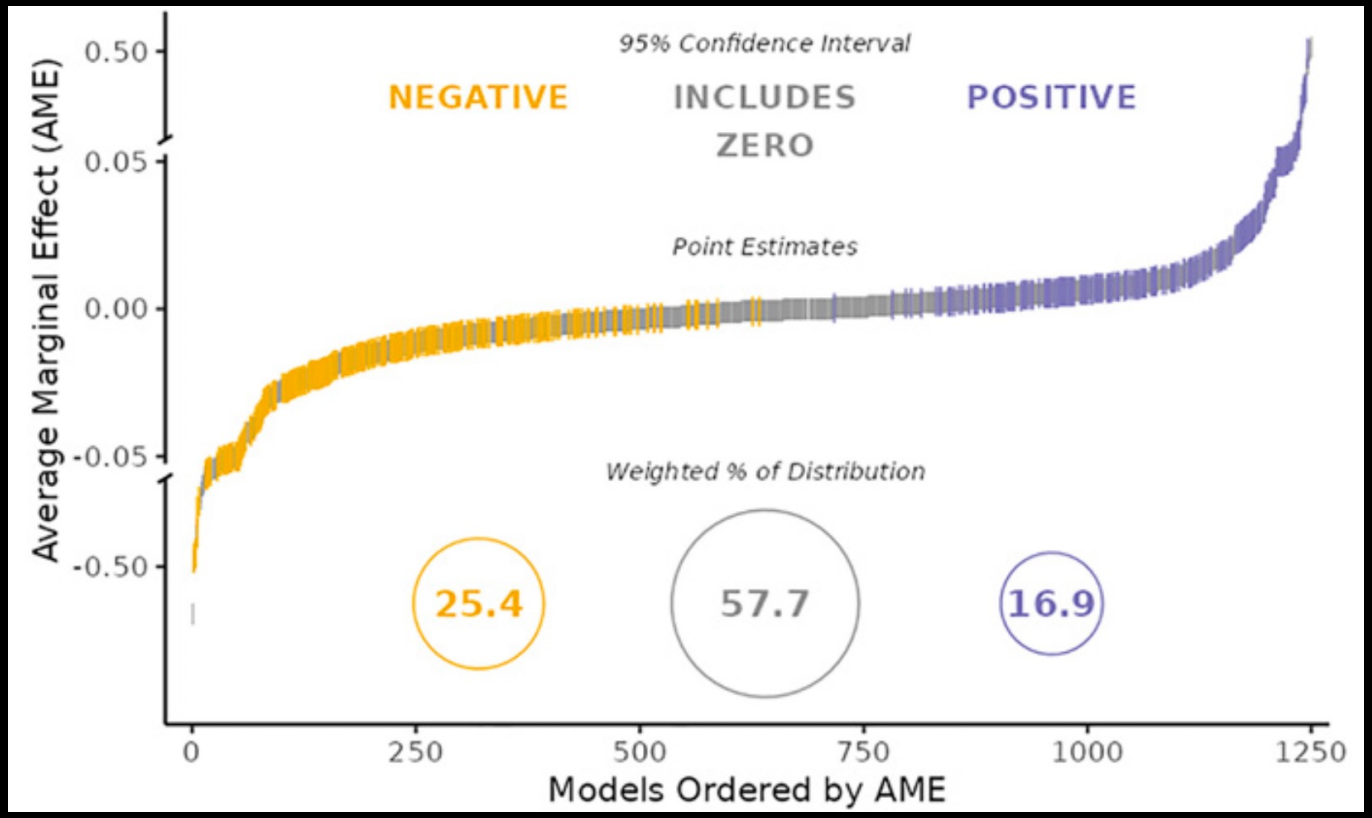
One way out of the vicious cycle of power analysis and publication bias



Nakagawa et al. 2024 Finding the right power balance

Large scale replication study in social sciences

1 dataset analyzed by 161 researchers in 73 research teams



Do you expect the same variation in outcomes if this was repeated in your field of genomics?

Breznau et al 2022 Observing many researchers using the same data and hypothesis reveals …

How many of you are trying to find genomic regions of importance for your phenotypes?

Are you using molecular tests of selection?

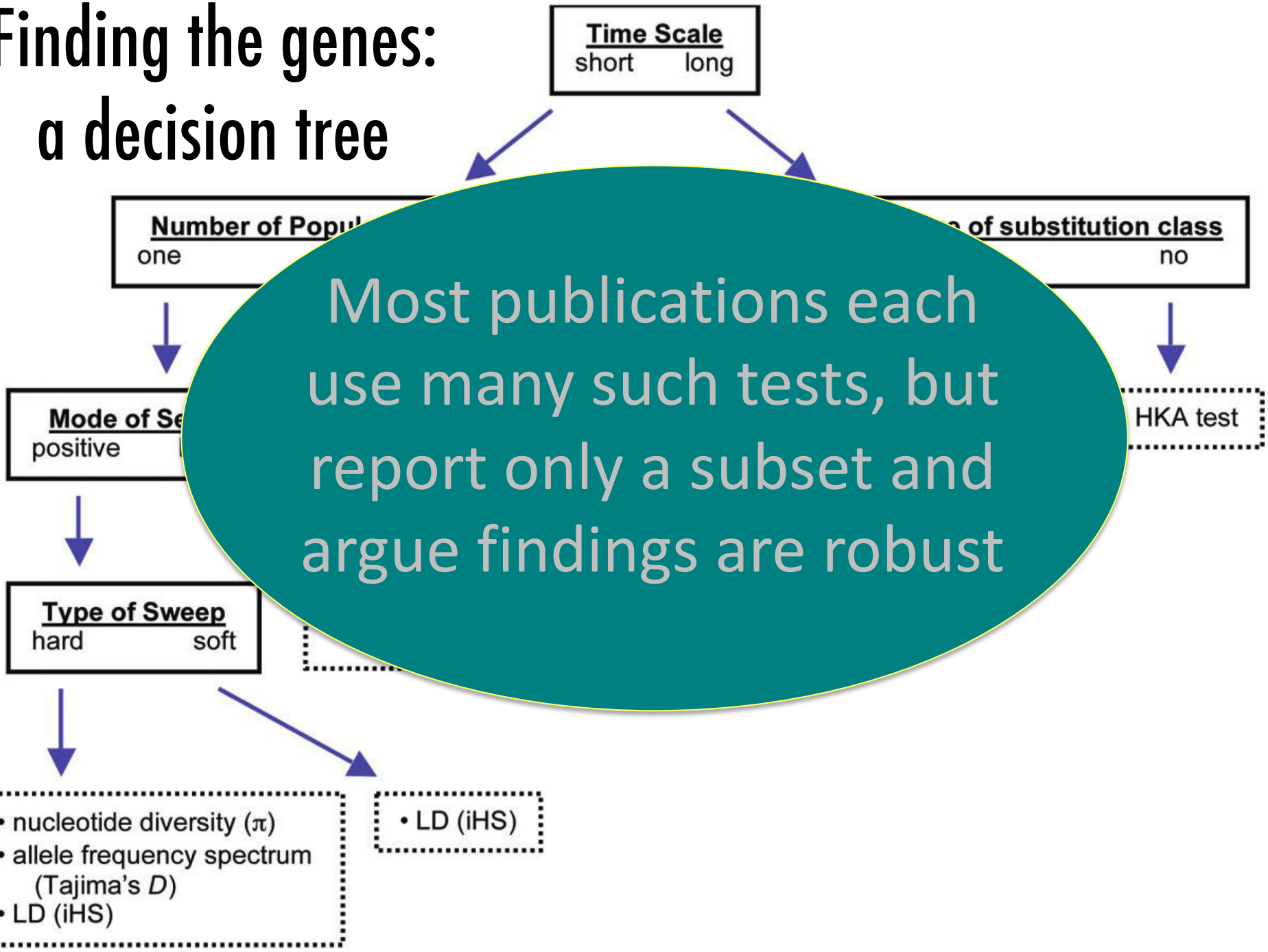
How do you decide what tools to use?
Does it matter?
When do you stop running tests?
Which do you report?

How do we identify the genes that matter?

- Molecular tests of selection are popular, but ...
  - What are their assumptions and statistical power?

- What are these tests detecting?
  - What is a footprint of selection?
    - How are they formed?
    - How large are they?, how long do the last?
    - How are they impacted by demographic history? Introgression? Aliens?
    - Should we even expect "footprints"?

Finding the genes: a decision tree



Most publications each use many such tests, but report only a subset and argue findings are robust

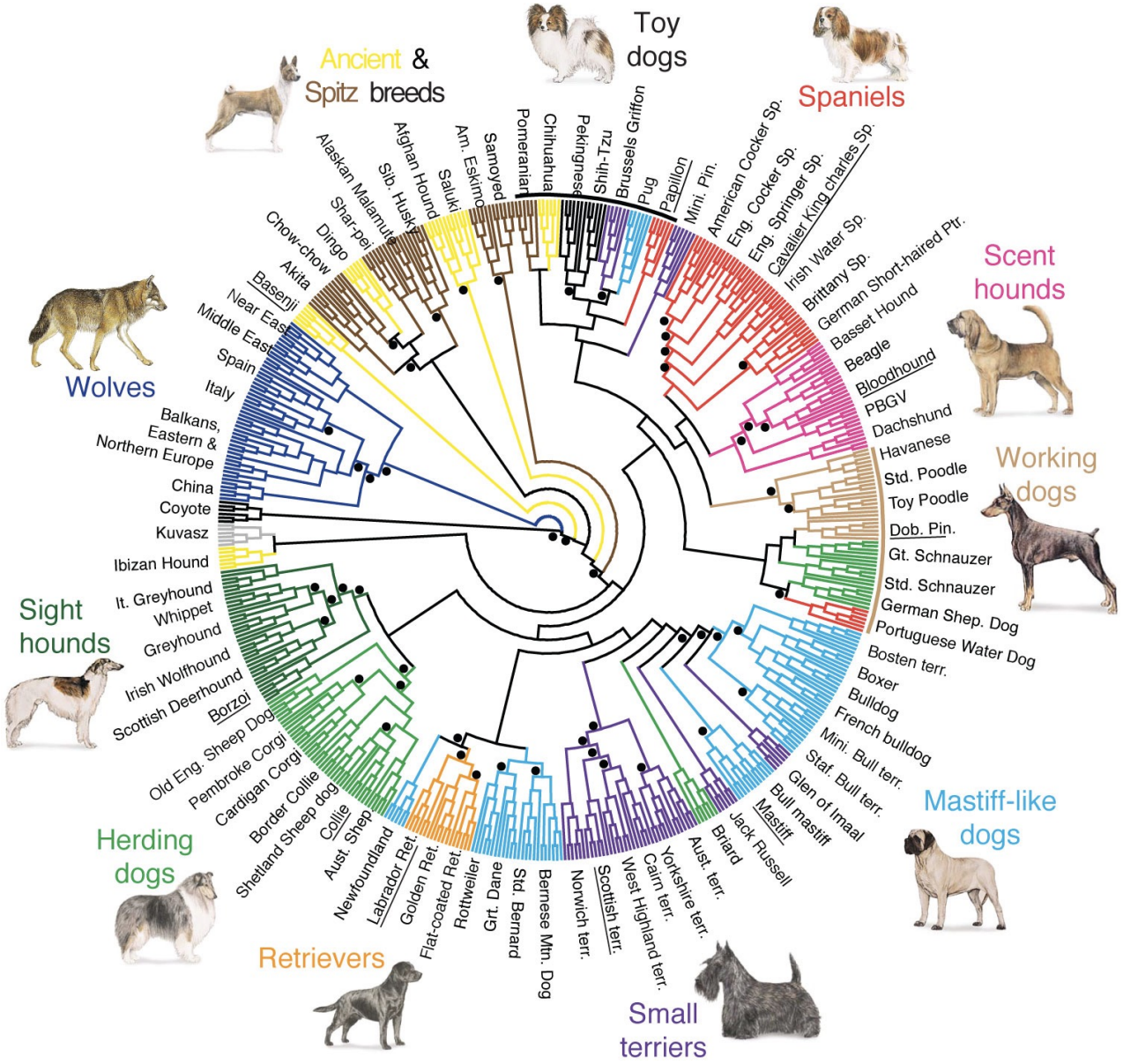
Hohenlohe et al. 2010 Int. J. Plant Science

What power do we have to detect evolution by natural selection?

What is statistical power?

Power is the probability that the test will reject the null hypothesis when the alternative hypothesis is TRUE

Should independent molecular tests converge?



von Holdt et al. 2010. Nature

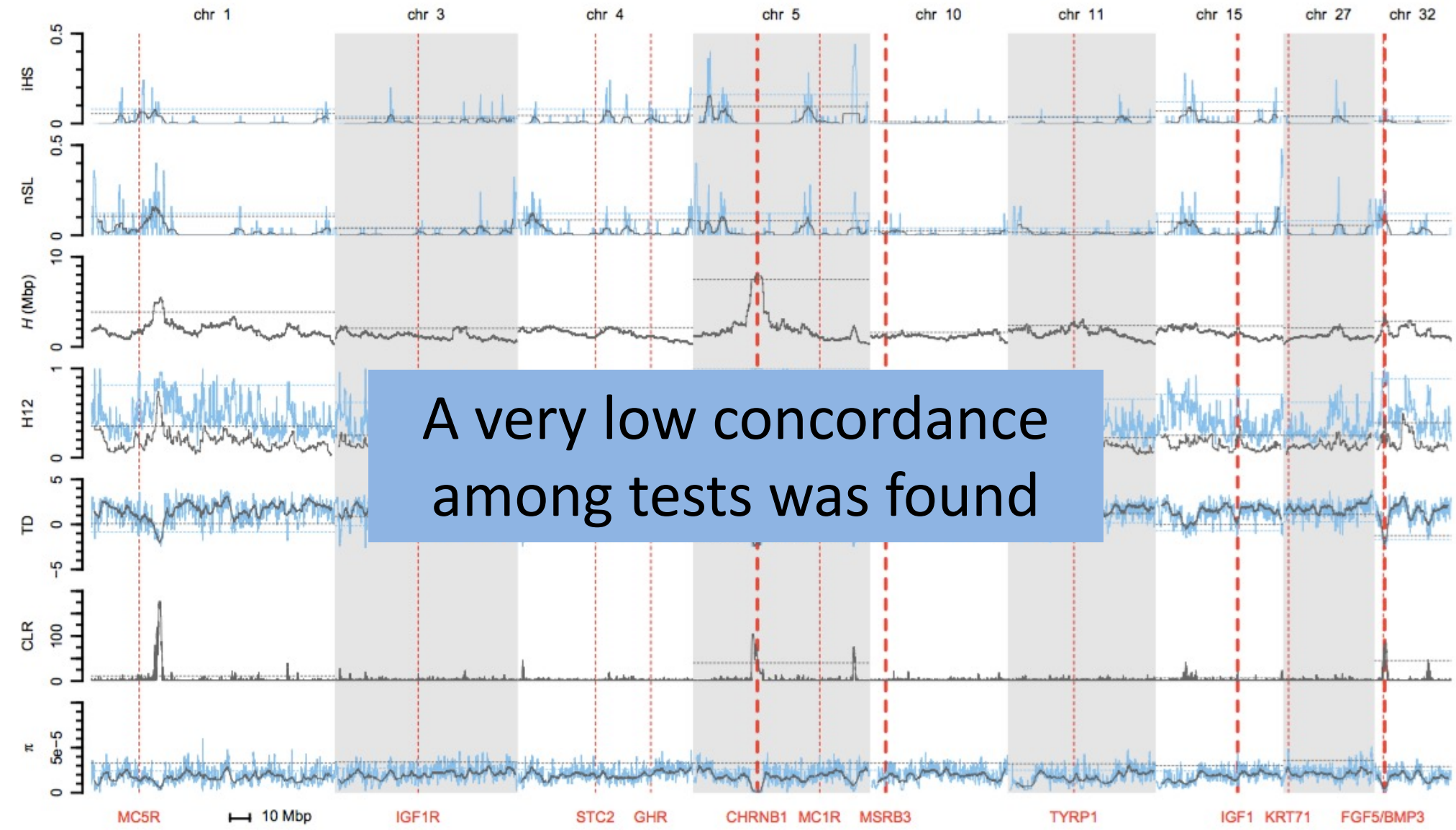
Breed specific morphologies

Test set of Schlamp et al. 2016:

- 25 breeds
- 12 causal loci identified by QTLs
- N = 25 / breed
- 7 tests of selection
  - iHS,nSL,H,TajD, etc.

How concordant are molecular tests of selection detect?

French Bulldog sample: red lines are causal QTL loci



Schlamp et al. 2016. Evaluating the performance of selection scans to detect selective sweeps in domestic dogs. Molecular Ecology 25:342–356.

Why don't these these tests agree?

biological reality
or
our expectations
or
theoretical population genetics

What if different groups used different tests?

Would that skew the literature?

Is this common? Should we worry?

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
  - Extremely large P-values can arise from extremely weak patterns, so ask yourself, does the effect and effect size have biological meaning?

Test your hypotheses in independent ways

- Derive hypotheses from your genomic results, then
- Test these hypotheses using relevant manipulations
  - 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

Robust research needs many lines of evidence

Replication is not enough

Munafò and Smith 2018 Robust research needs many lines of evidence

Triangulation

Robust research needs many lines of evidence

Replication is not enough

Munafò and Smith 2018 Robust research needs many lines of evidence

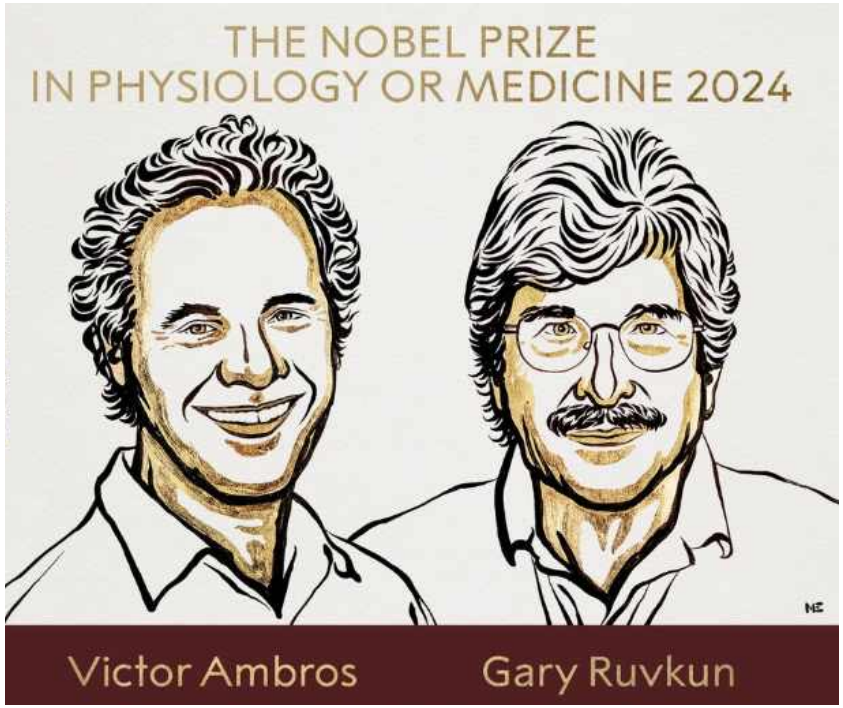
Triangulation — a checklist

- Use different approaches to address the same hypothesis, or extensions of hypothesis
- Sources of bias for each approach should be explicitly acknowledged, in opposite directions, and independent
- Results from more than two approaches are ideally compared

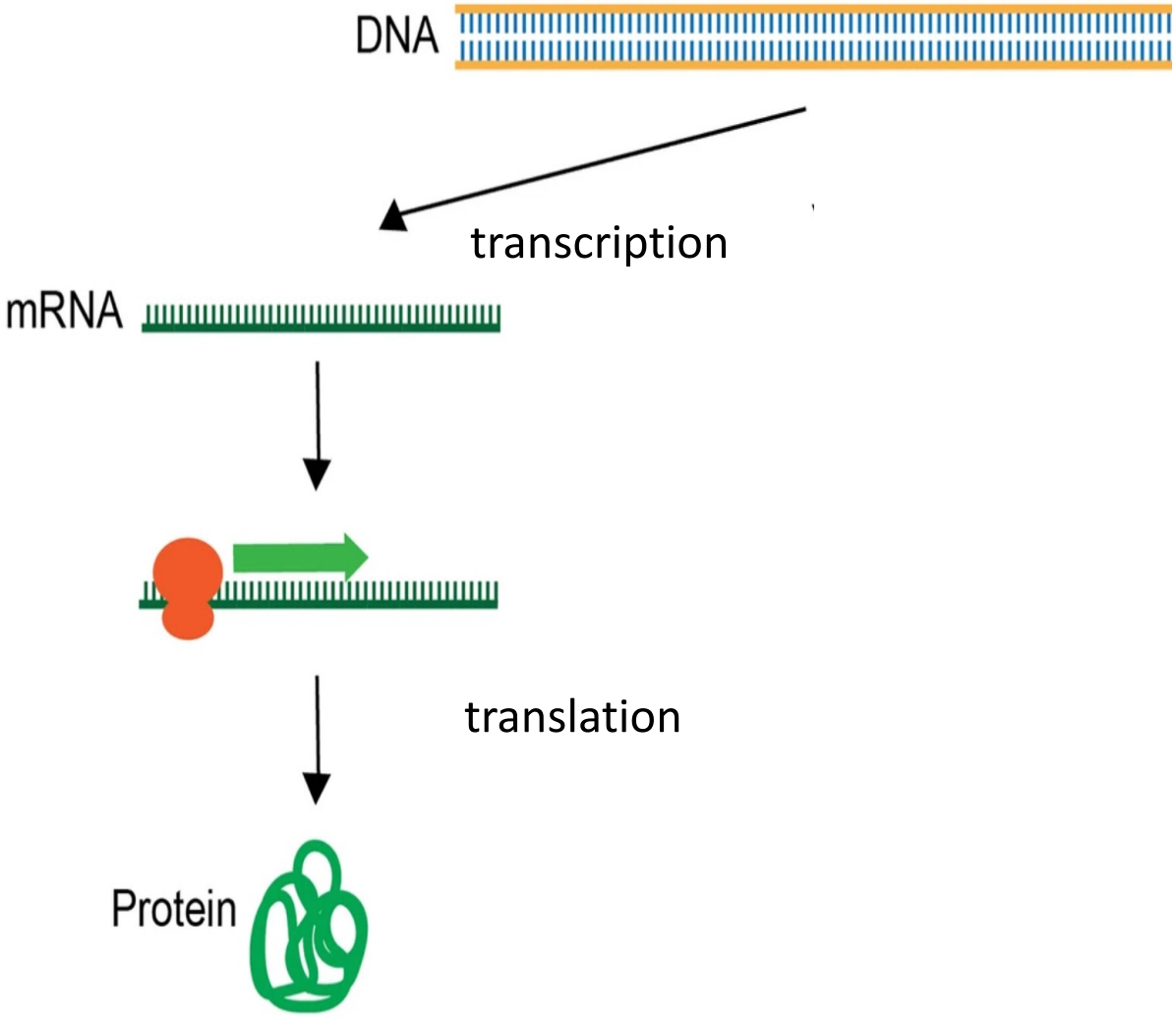
Munafò and Smith 2018 Robust research needs many lines of evidence

An example

- Using a new genomics technique
  - miRNA
- Trying to understand what is best practice
- Worked hard to triangulate upon what's a biological signal vs. bioinformatic artifact
- Uncovered serious problem in the non-model community



"for the discovery of microRNA and its role in post-transcriptional gene regulation"



miRNAs

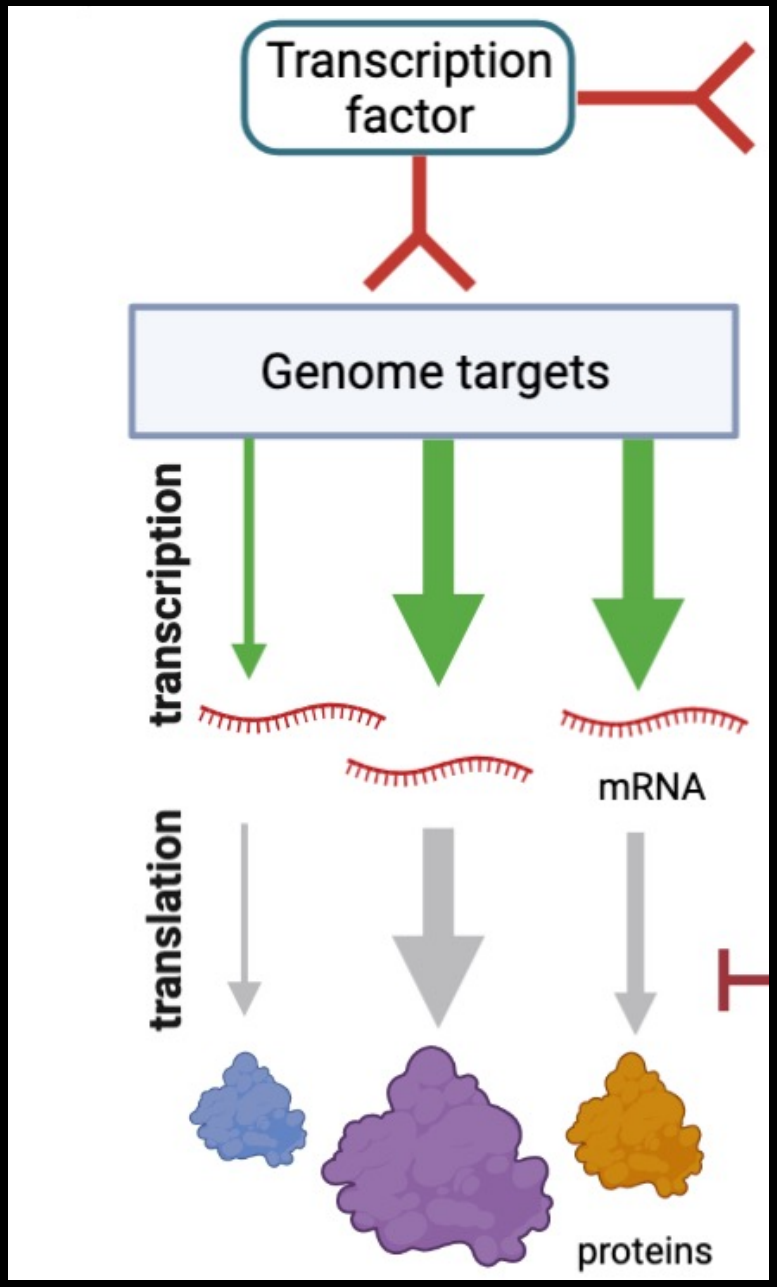
destabilizes mRNA

sculpts the pool of mRNA

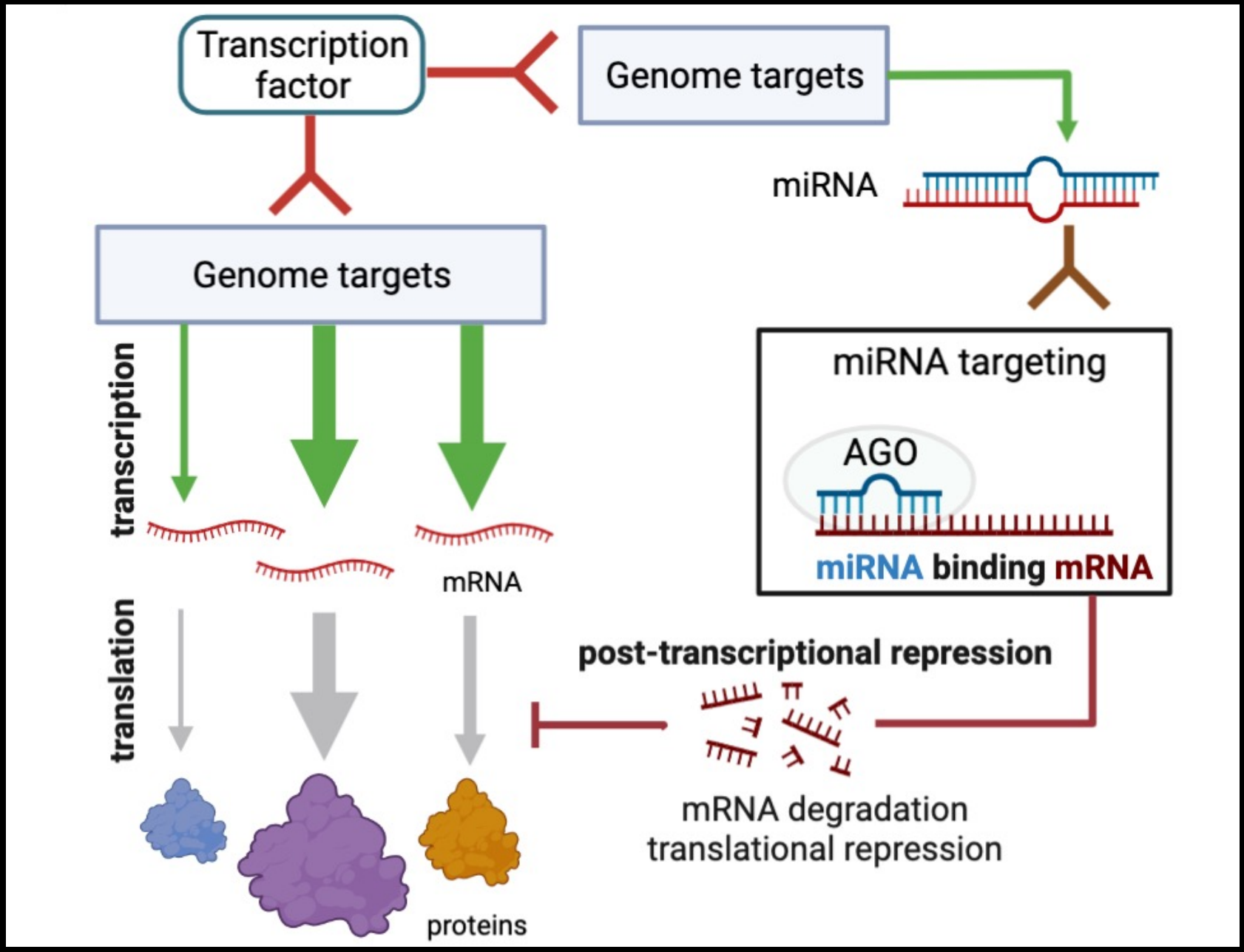
key part of regulatory networks

metazoans can't live without'em

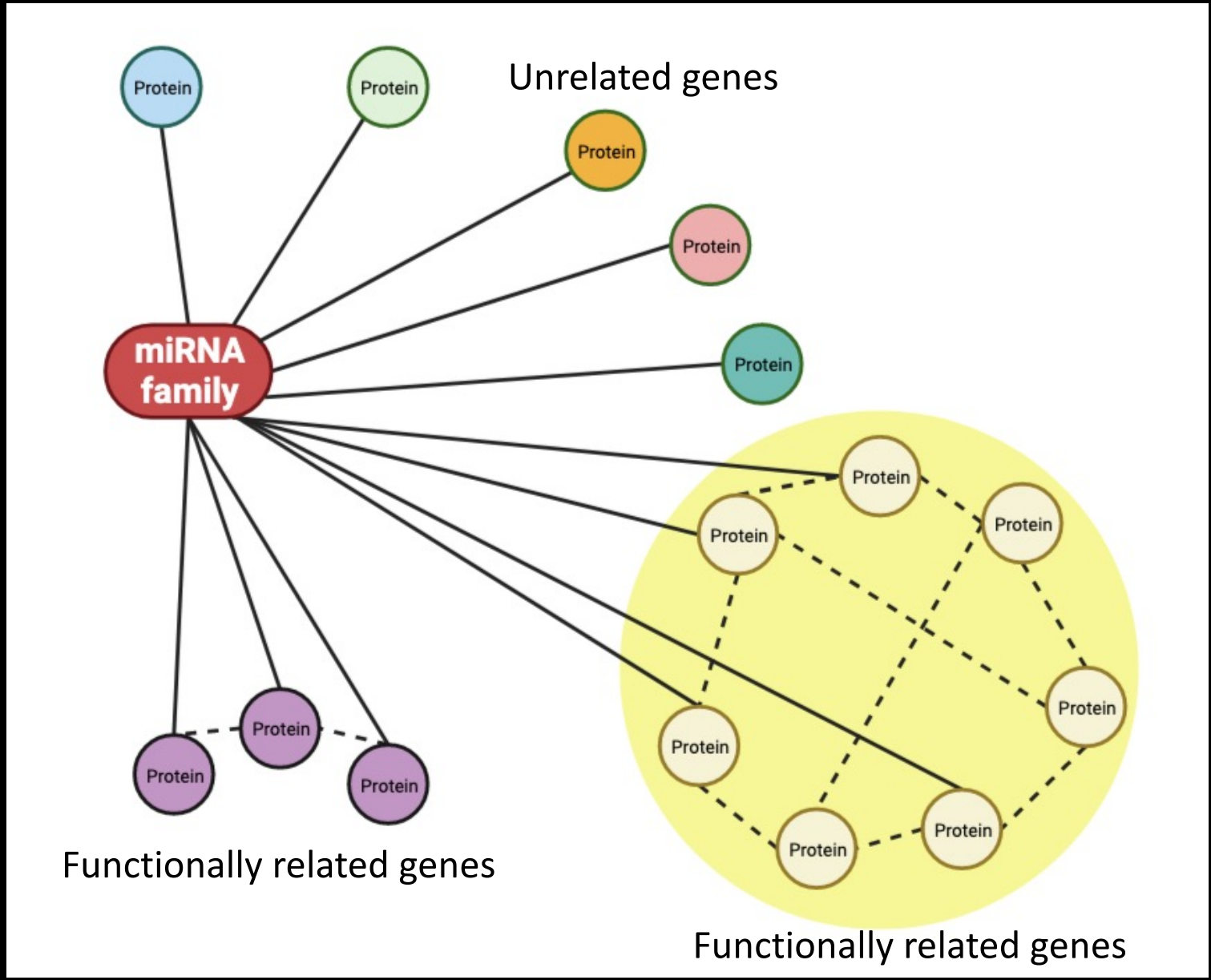
The role of miRNA in sculpting the transcriptome



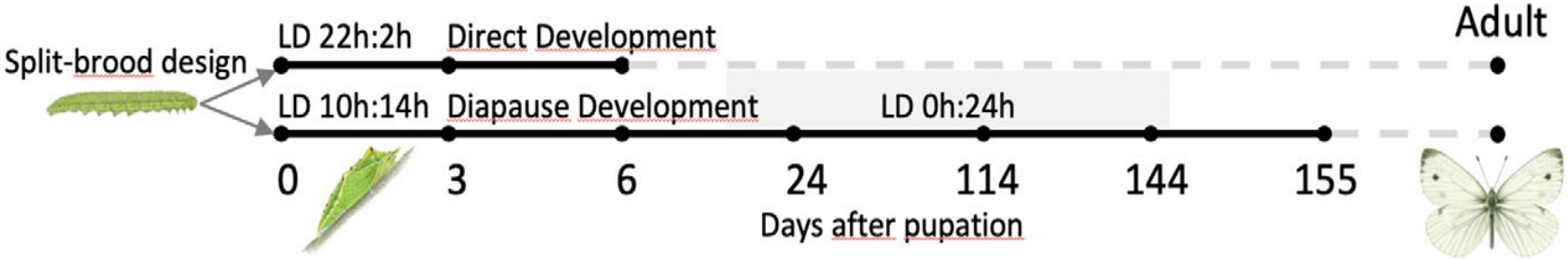
The role of miRNA in sculpting the transcriptome



Regulatory network view of miRNA impacts



Dynamic microRNA expression across diapause

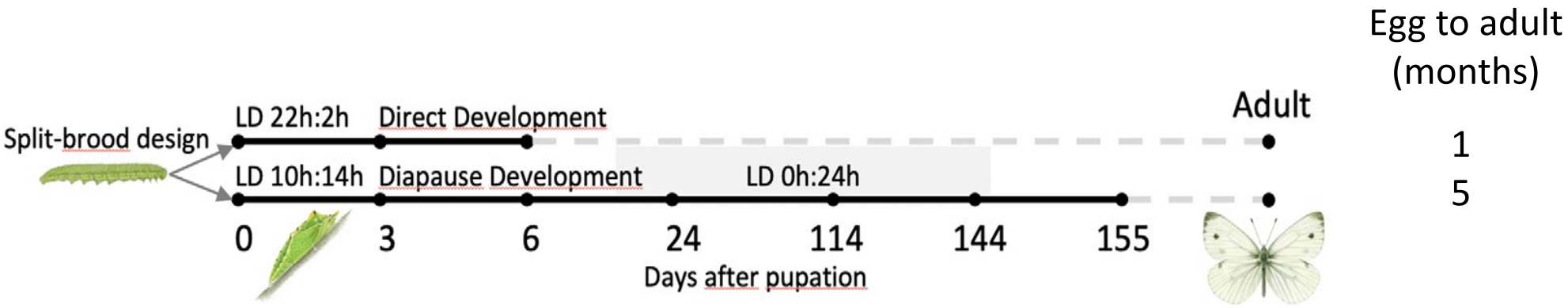


Egg to adult (months)

1

5

Dynamic microRNA expression across diapause



Pieris napi

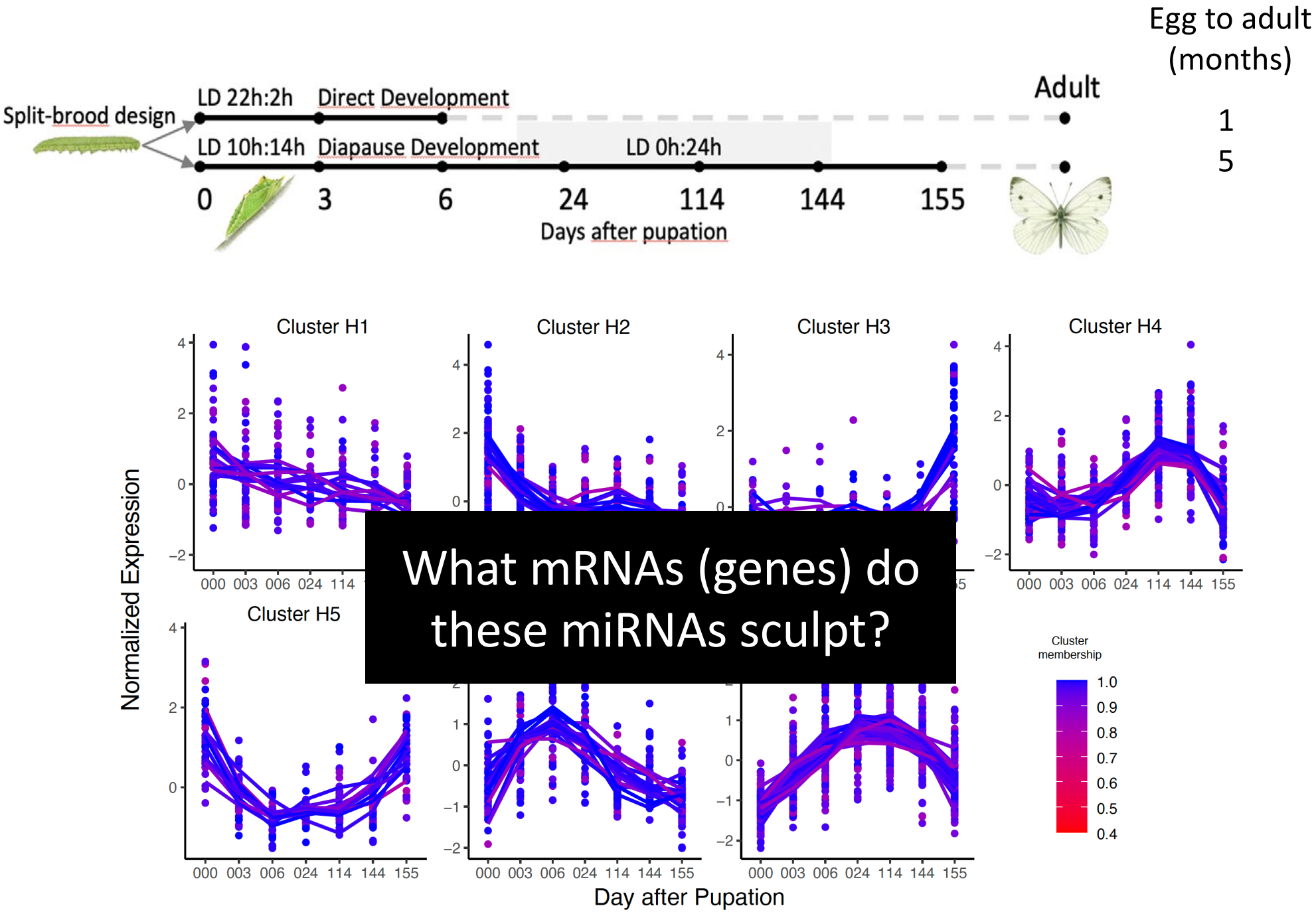
miRDeep2 identified 188 microRNAs

<u>73 libraries (</u>±6.9M reads / lib):

- 12 time points
- 2 tissues
- 3 replicates

Roberts et al. 2024 Molecular Ecology

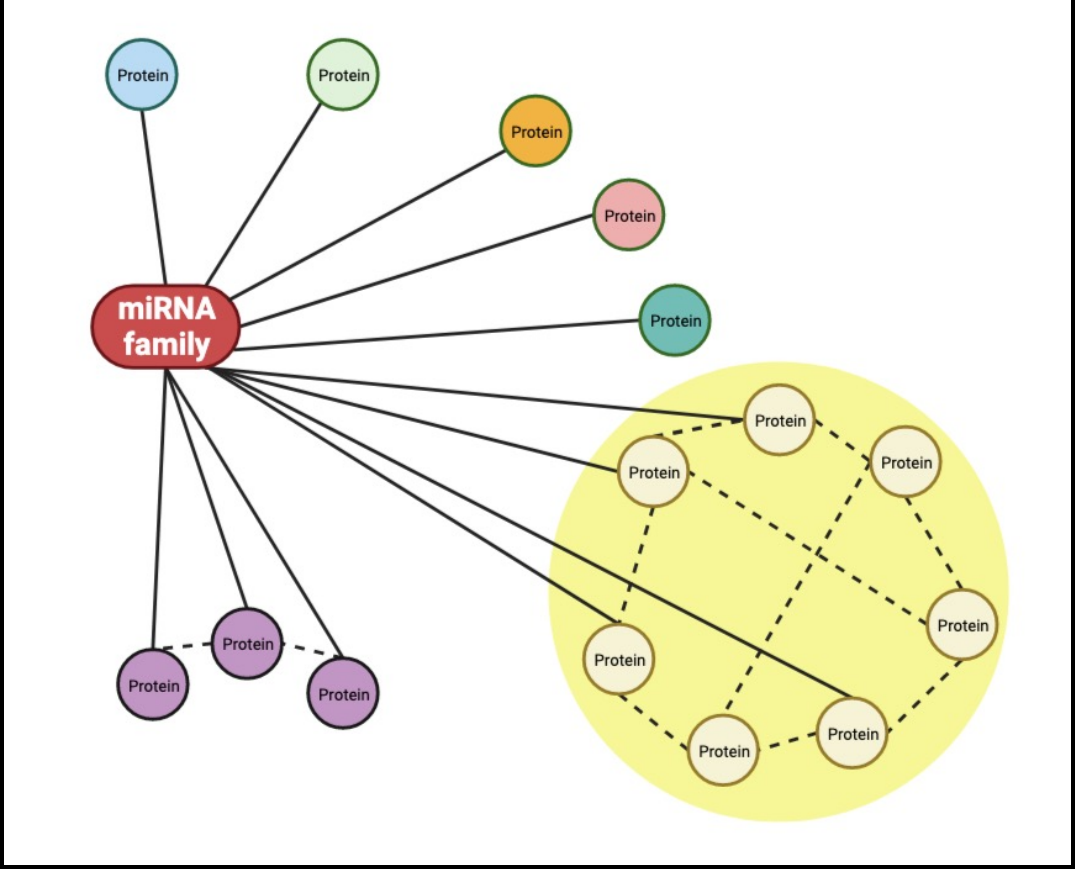
Dynamic microRNA expression across diapause



What mRNAs (genes) do these miRNAs sculpt?

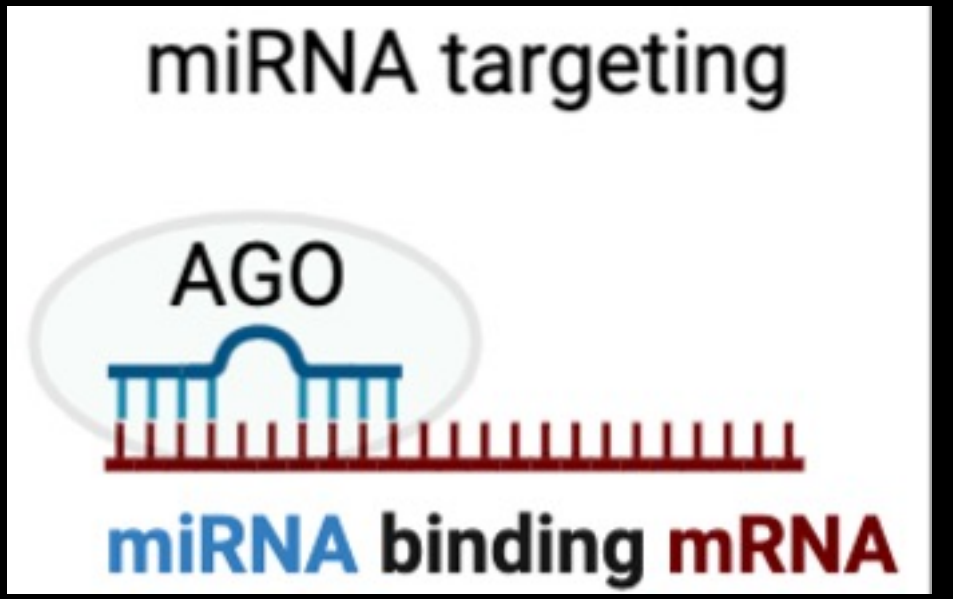
OK, so some miRNAs are changing through time..

Where are they targeting?
What are they doing?

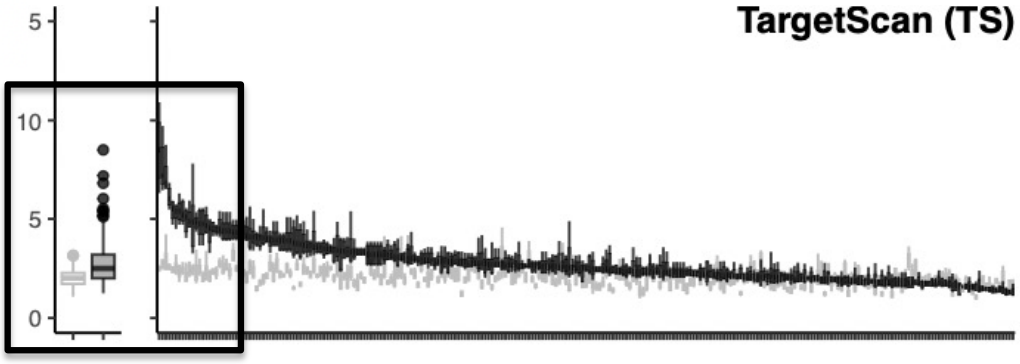


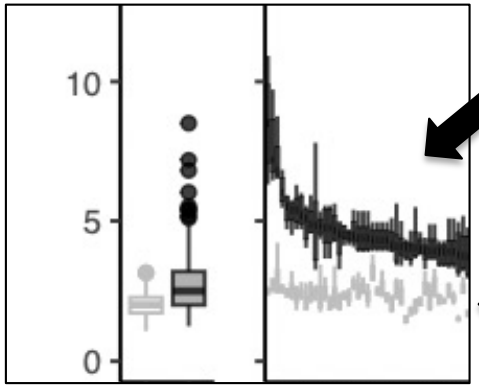
What functional groups or pathways might they regulate?

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

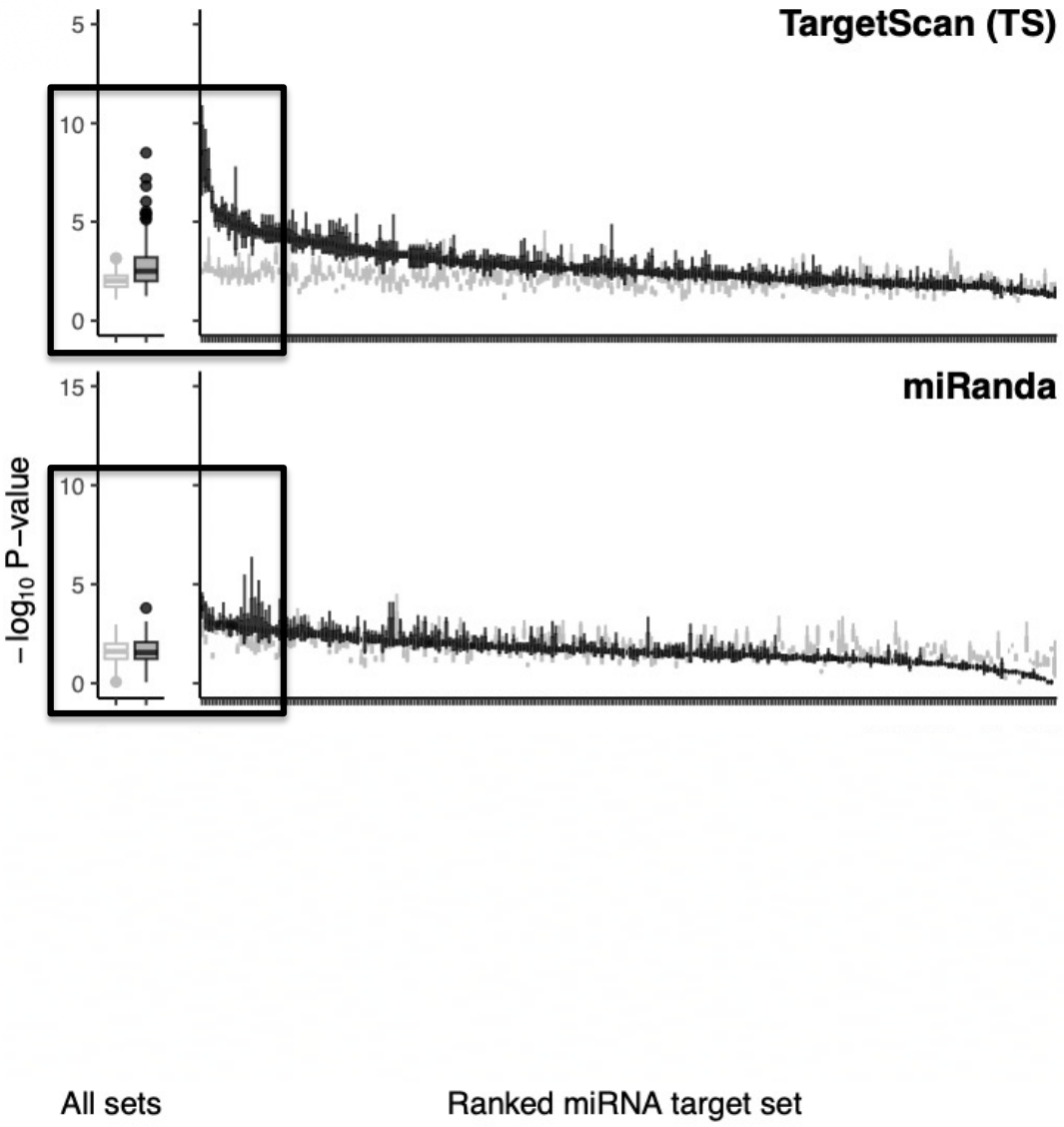


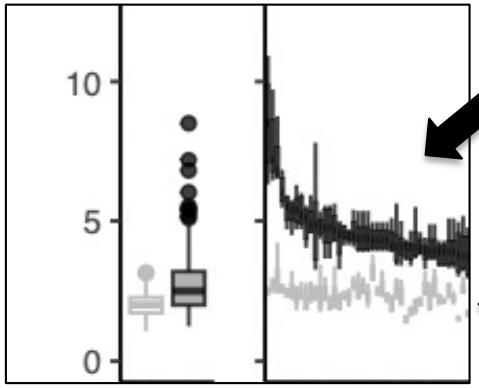


Targetscan using 6 species 3’UTRs alignments

Random sets

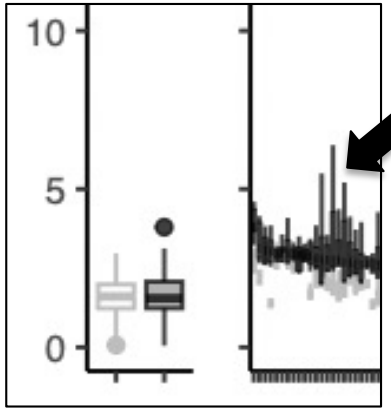
Wheat et al., in submission





Targetscan using 6 species 3’UTRs alignments

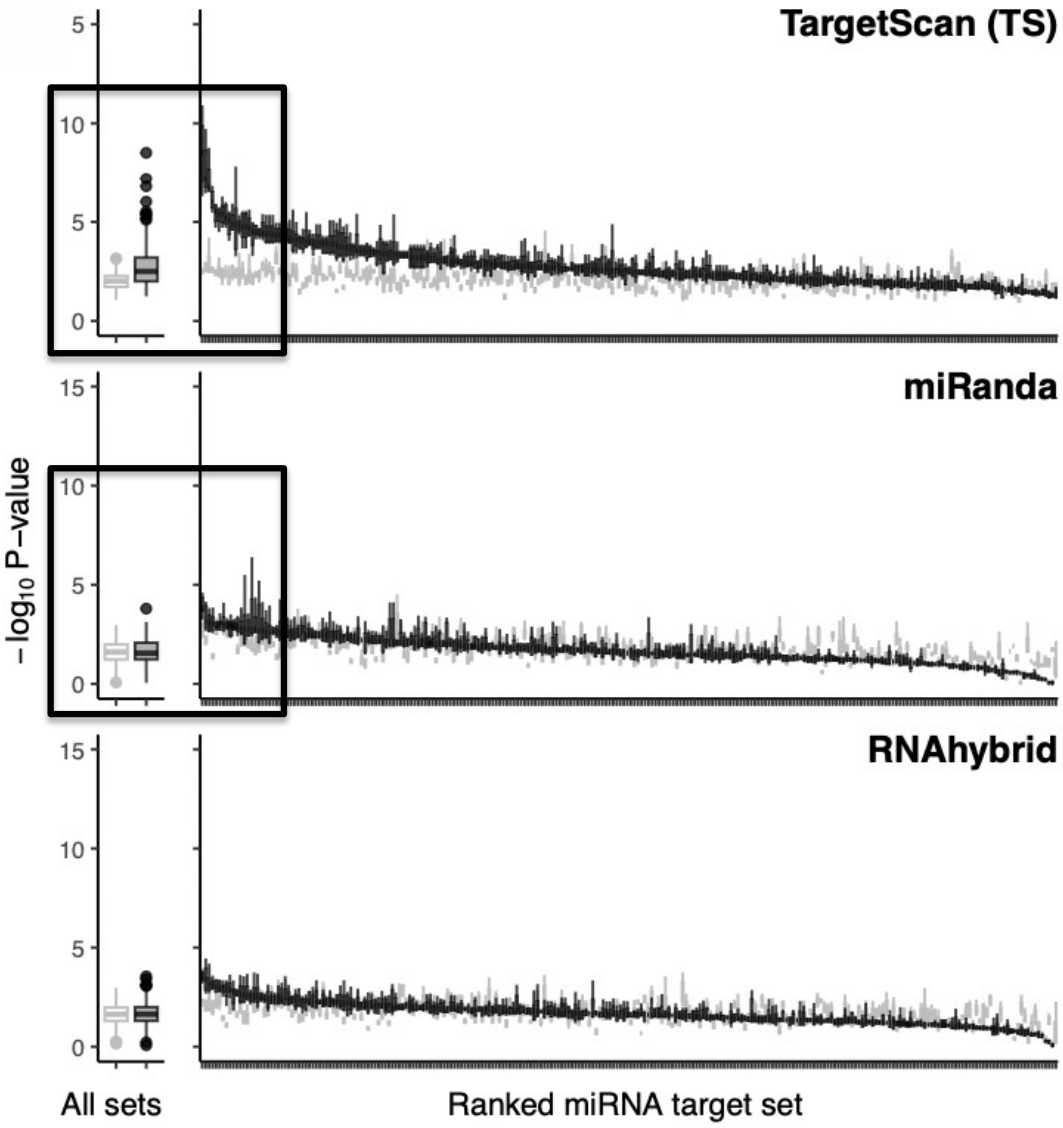
Random sets

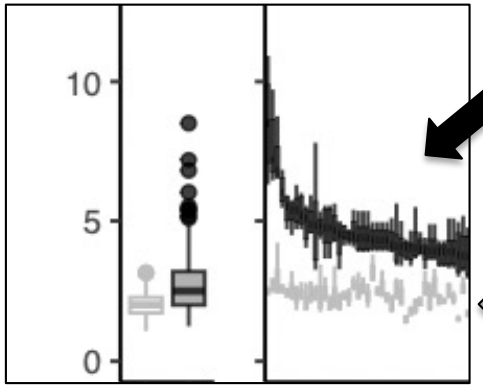


miRanda using one species 3’UTRs

Random sets

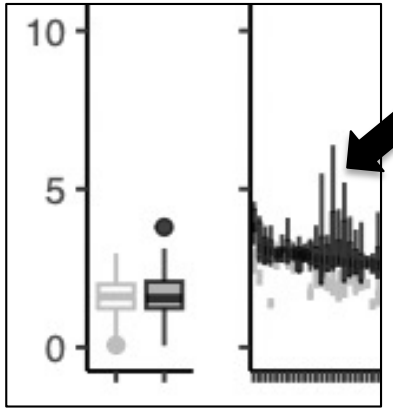
Wheat et al., in submission





Targetscan using 6 species 3’UTRs alignments

Random sets



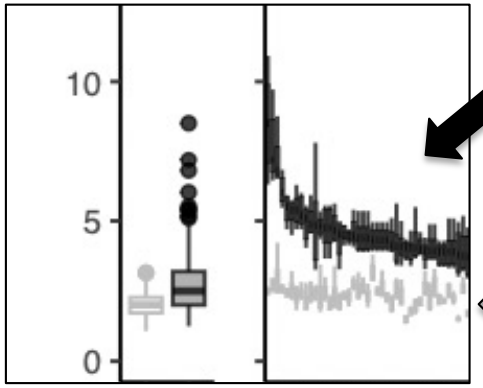
miRanda using one species 3’UTRs

Random sets

Wheat et al., in submission

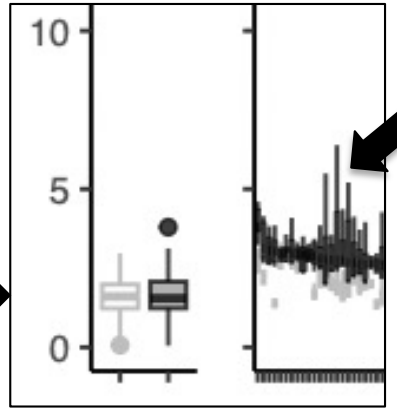
Large bioinformatic effect

>90% of miRNA literature in ecology and evolution uses miRanda to assess miRNA impacts…



Targetscan using 6 species 3'UTRs alignments

Random sets



miRanda using one species 3'UTRs

Random sets

Wheat et al., in submission

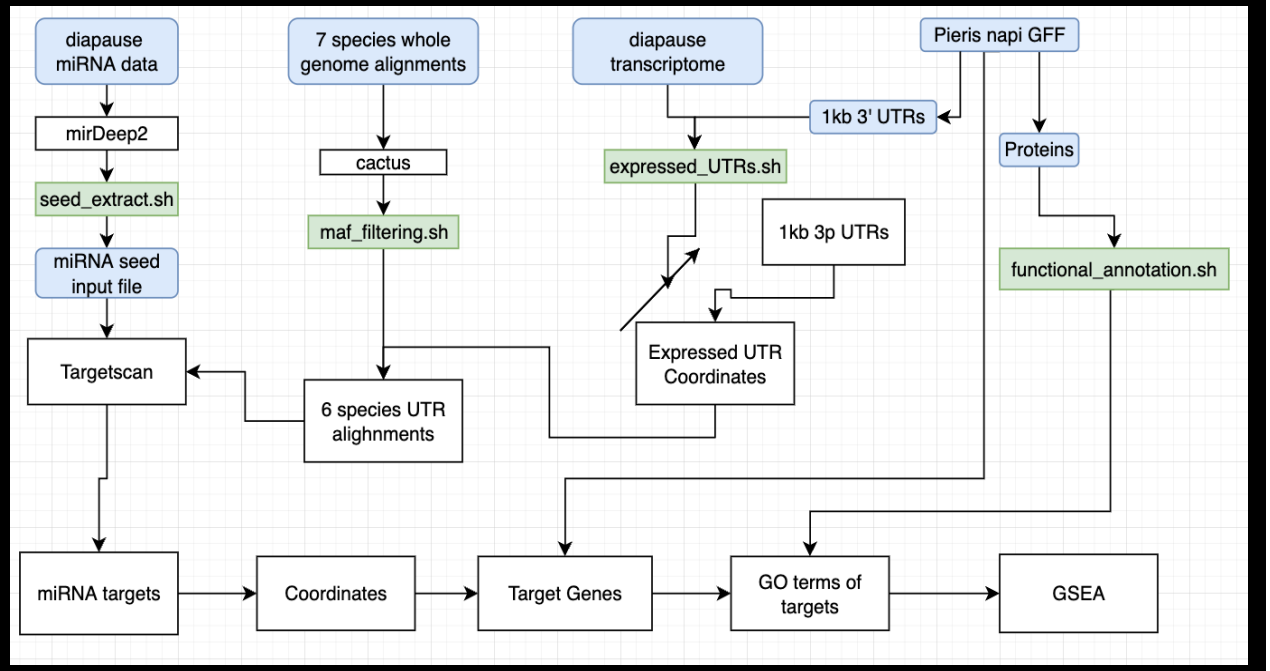
So, why don't more people use Targetscan approach?

Running miRanda is quick and easy

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

Running TargetScan7 with alignments is a lot of work

- Download scripts, generate 3'UTR alignments for 7 species, load miRNA seed sites, etc.



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
  - Species comparisons for cross-check & generality

Sum: intersection across divergent methods, 1st principals metric, and comparative analysis revealed believable results

CellPress
OPEN ACCESS

Trends in
Ecology & Evolution

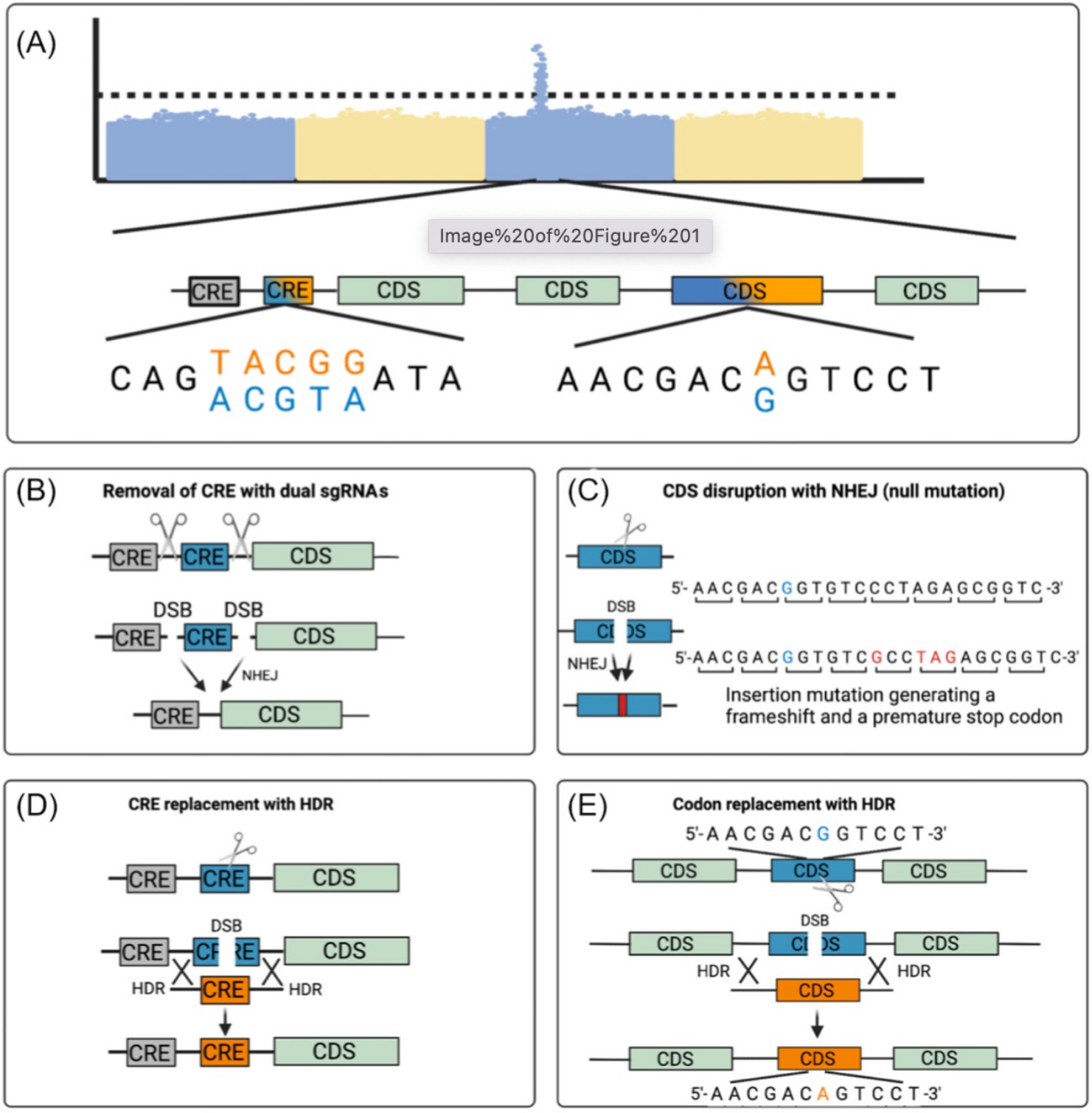
Review

Functional genomic tools for emerging model species

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

Recent review covering diverse means of validation across diverse taxa

As genomics gets cheaper, invest more in validation instead of just more sequencing!!!!!


(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'

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
    - Beware of these red flags, as many authors only seek a publication and won't help

Cookbooking ...

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

Keep good bioinformatic notes

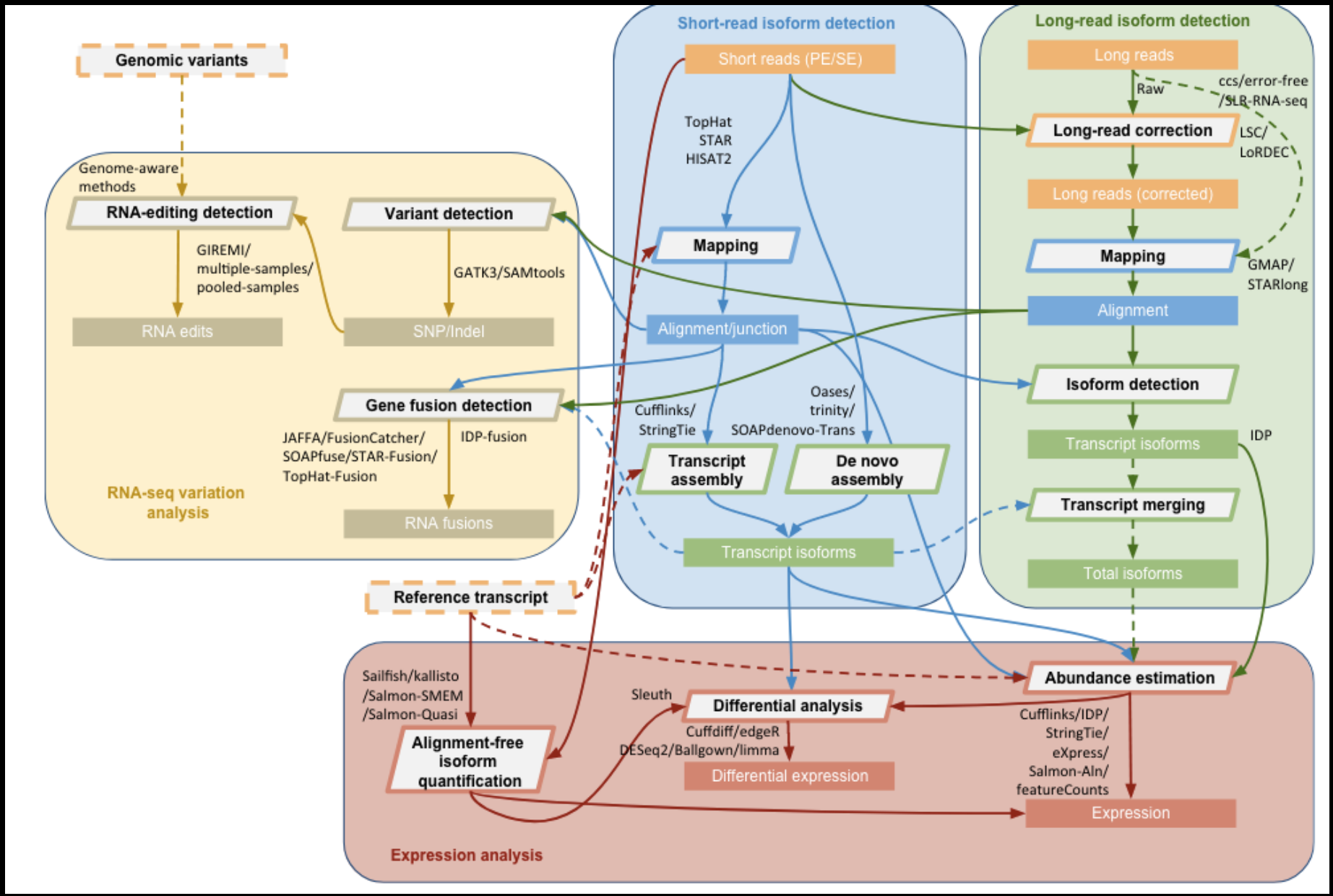
- I keep a special file with commands I learned, like and validate
  - use it to quickly find commands, refresh memory
- Use positive and negative controls to test the output of the commands you run (like all experimental biology)
  - I call these sanity checks
  - Always test to make code is working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

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 tools, performance varies across species, samples. This is no BEST tool or setting across species

Differential expression detection can vary by:

- Mapper
- Analysis software
- Reference genome
- Species

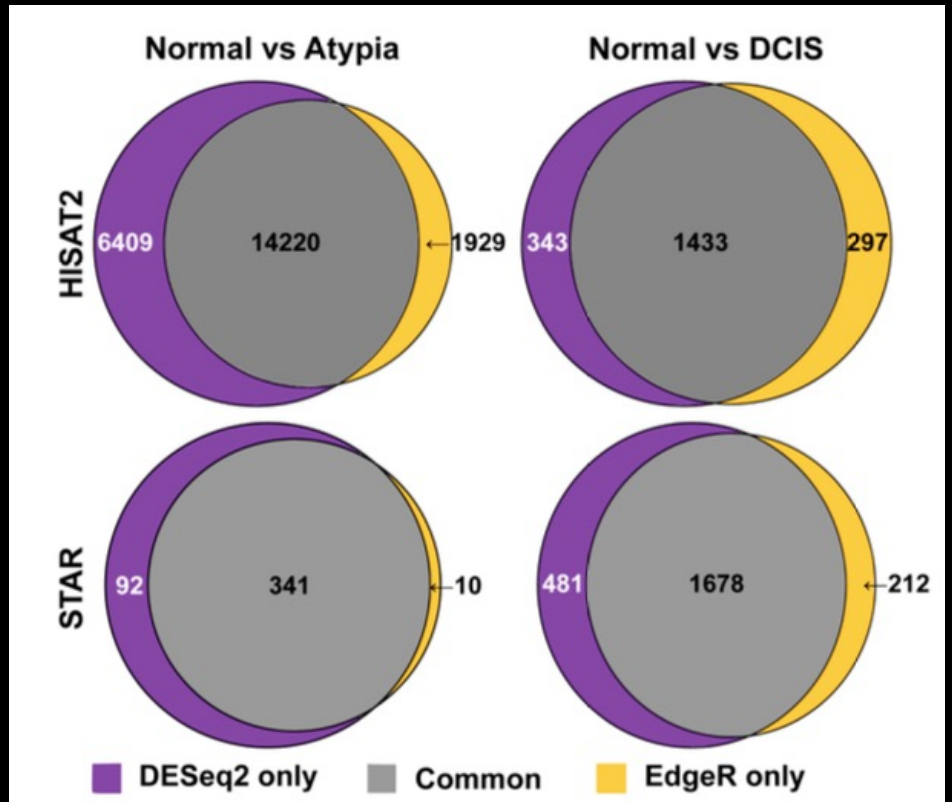
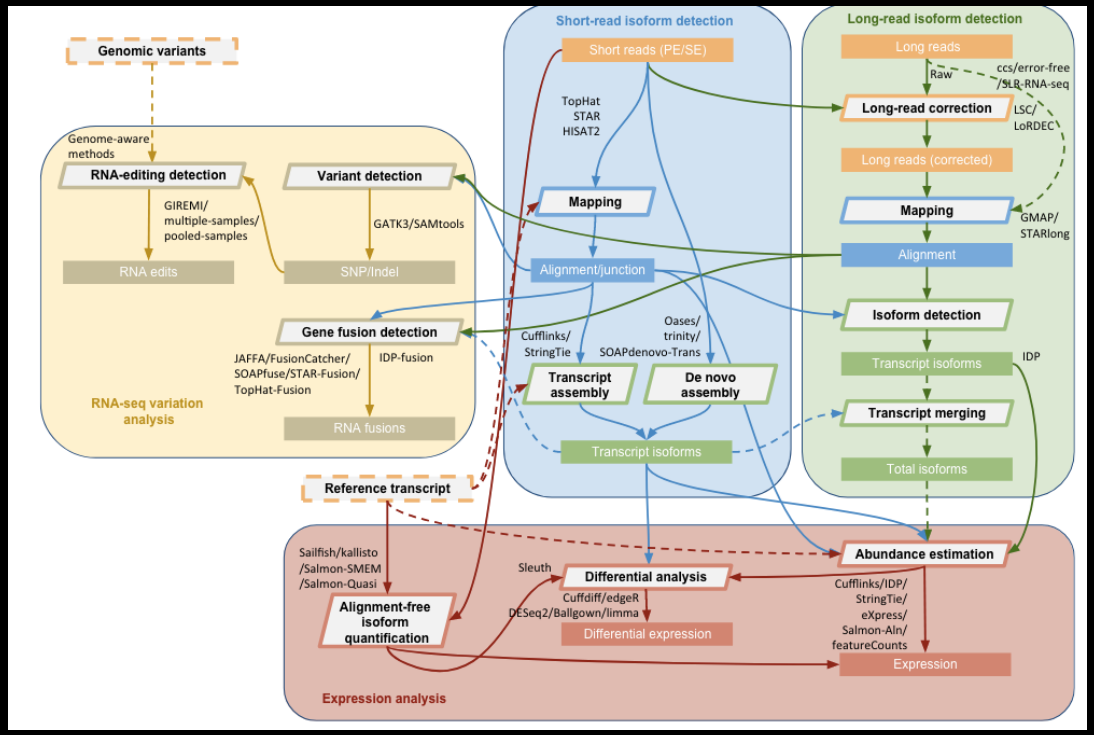
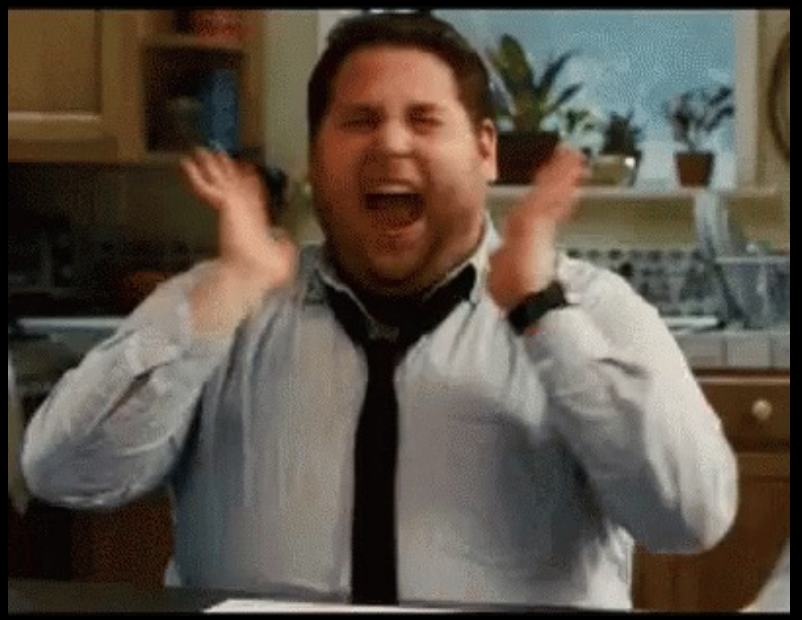


Figure 9. Overlap among genes identified as differentially express
HISAT2 or STAR-aligned RNA-seq data.

Raplee et al. 2019 J. Per. Med.

Doing many analyses ? analysis paralysis is common





Which is the right way?

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

Used published data & code, then try additional approaches

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.

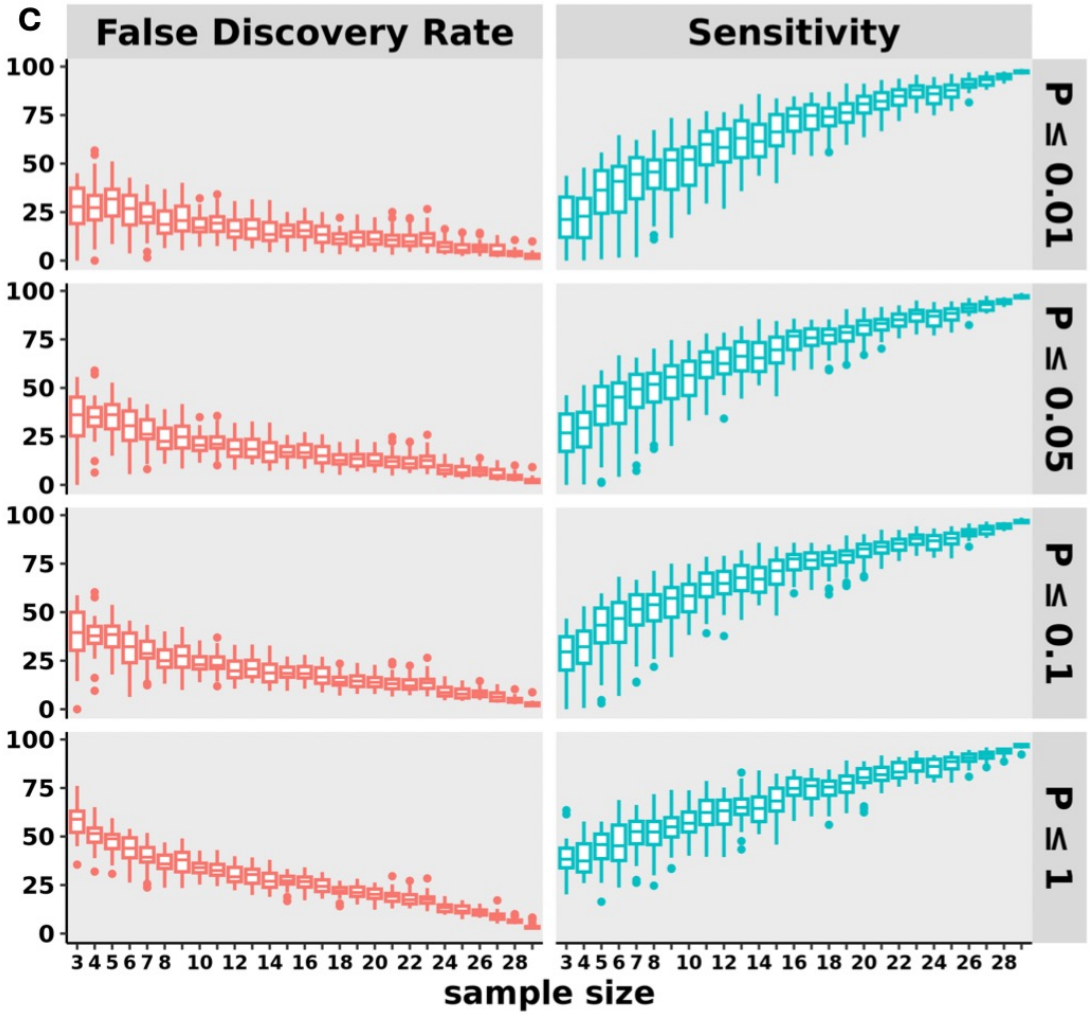
Look at others other peoples online, open access code & pipelines:

- Discover new ways of coding, reporting, working
- Gain confidence in your skills
- Become frustrated that other published work is not repeatable

If bioinfo work is not reproducible, how much can we trust it?

RNAseq and sample size

- wild-type mice and mice in which one copy of a gene had been deleted
- N=30 each group



Halasz, et al. bioRxiv July 11, 2024, p 2024.07.08.602525. https://doi.org/10.1101/2024.07.08.602525.

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 you want to invest months in ... for
  - you, or others?
  - A few, or many samples?
  - A way to help you advance your science and career?

Entering the mega-genomes era

GENOME

Darwin TREE of LIFE

Sequencing of Life

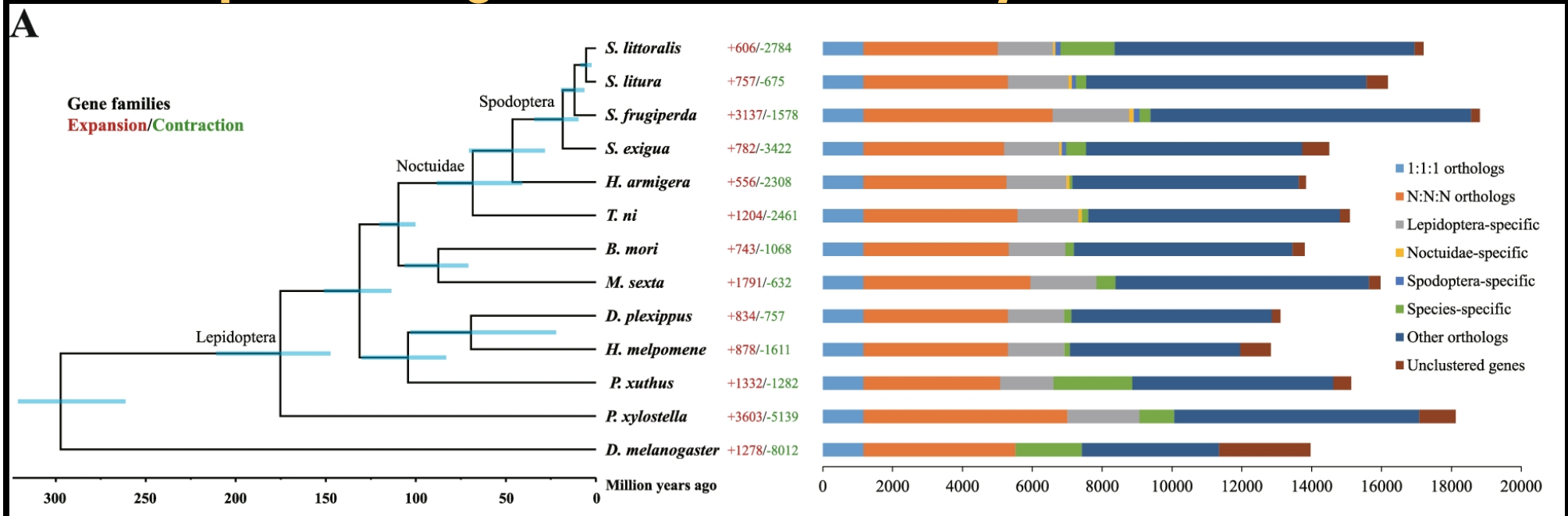
An unprecedented opportunity for large scale errors?

udying:

ships

– Functional insights into genes and their regulation

Comparative genomics commonly use annotations



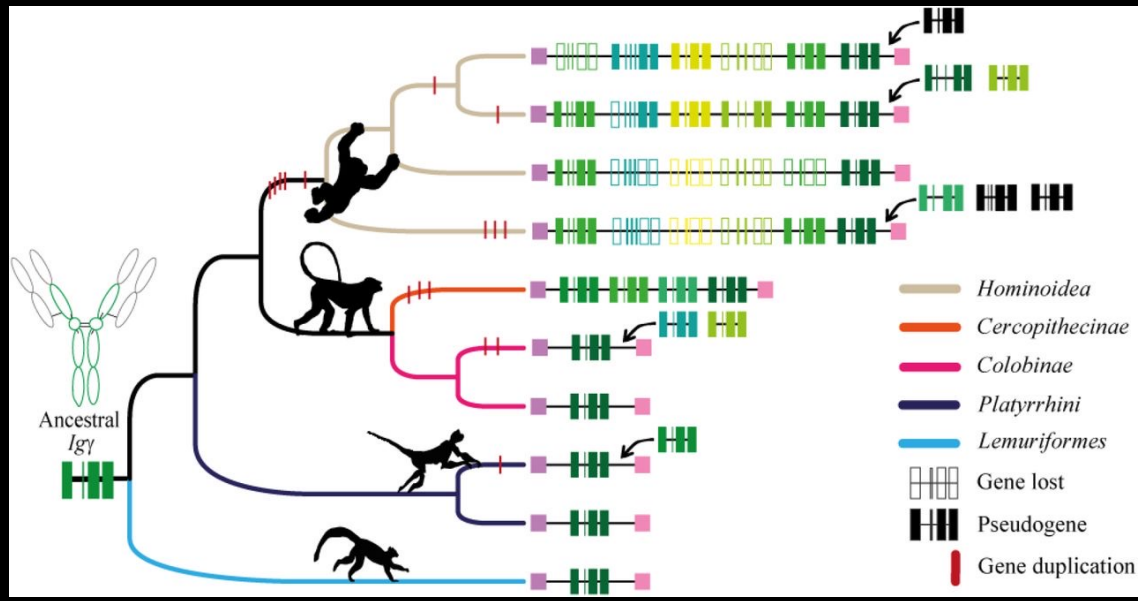
Typical genome report comparing gene content among species

- Rates of birth, death
- Lineage specific genes

Gene birth-death dynamics: biased, artifacts, or meaningful results?

- Are changes in gene numbers across species meaningful?
- Fundamental and important evolutionary question
- Very difficult to assess accurately
  - Need good genomes, annotations
  - Then good analyses

Immunoglobulin heavy constant gamma gene evolution



Garzón-Ospina & Buitrago 2022

Are all annotations equal among species?

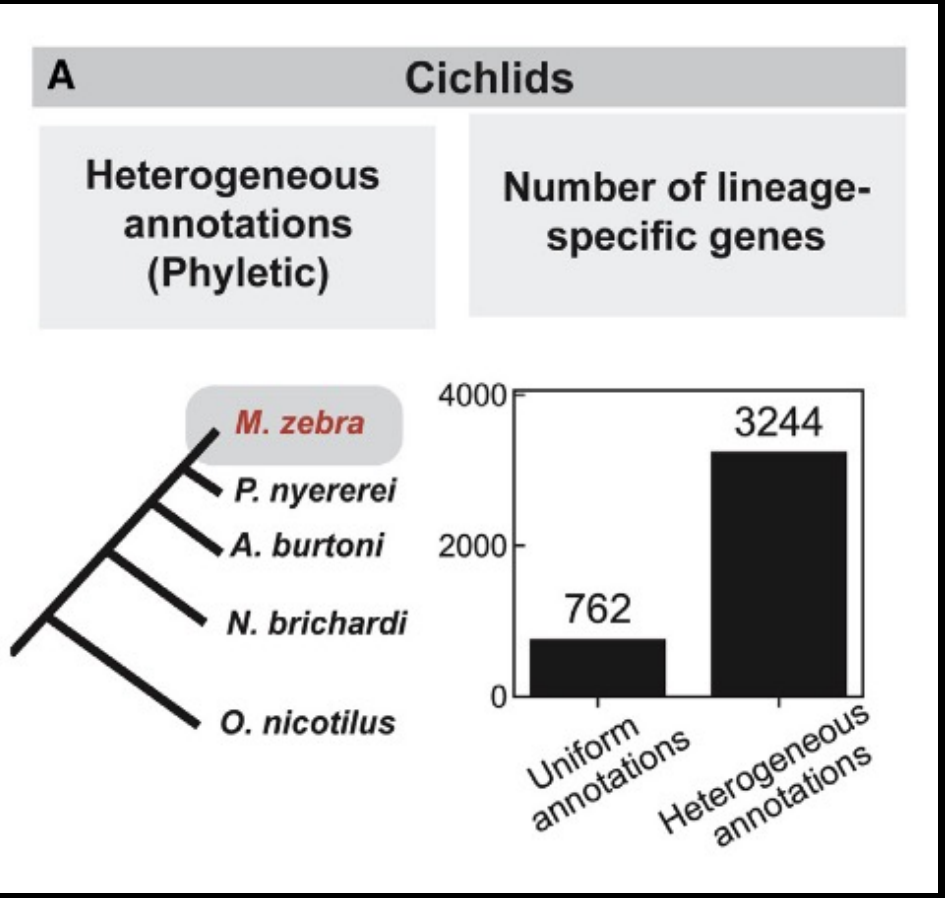
- Do species genomes differ in:
  - When they were sequenced, thus technology?
  - The quality of their assembly (e.g. N50, haploid state)?
  - How they did their annotation (proteins only vs. lots of RNAseq)?

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

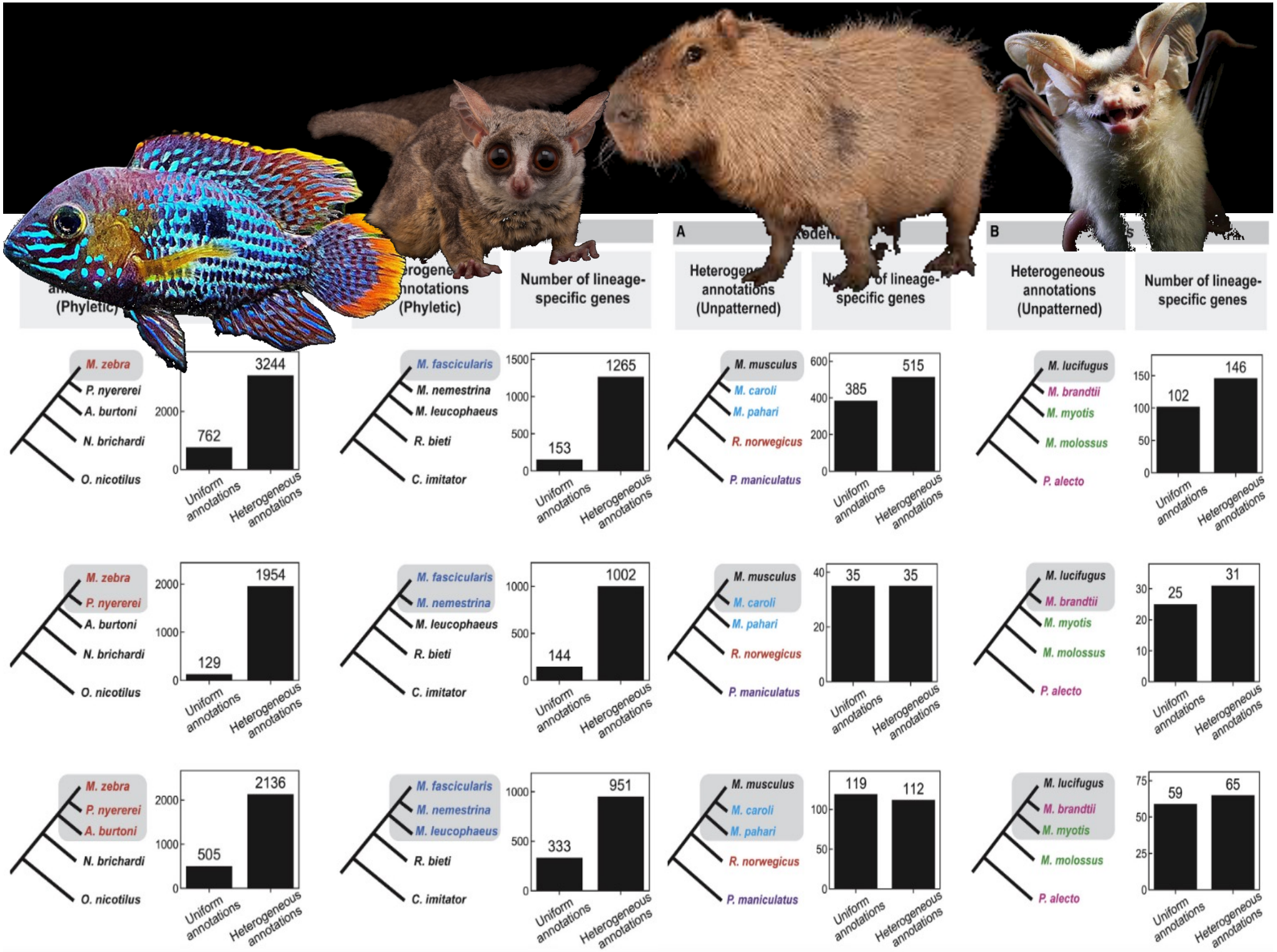
Will this impact analyses that rely upon accurate protein sets?

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

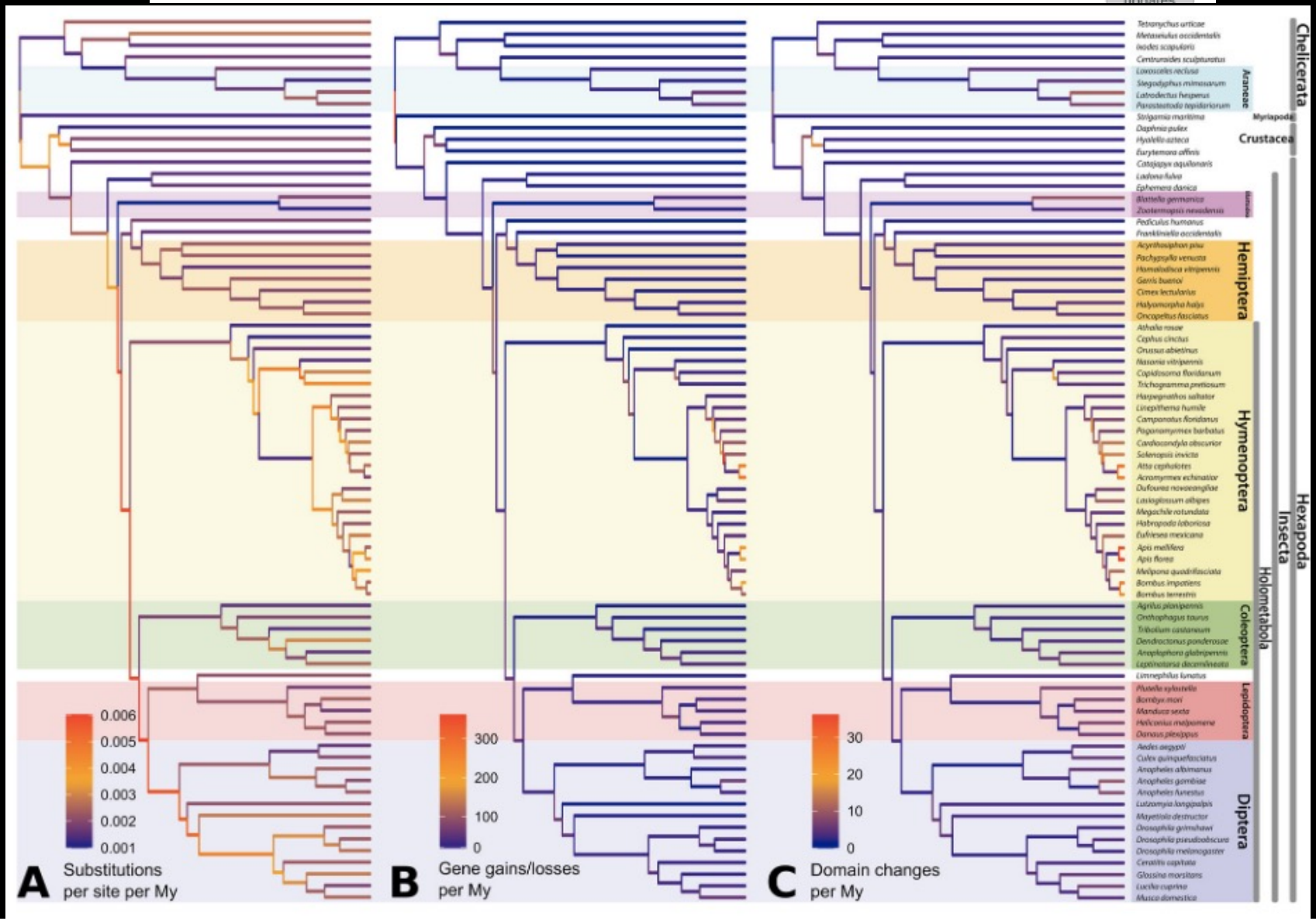

What are the ramifications?

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

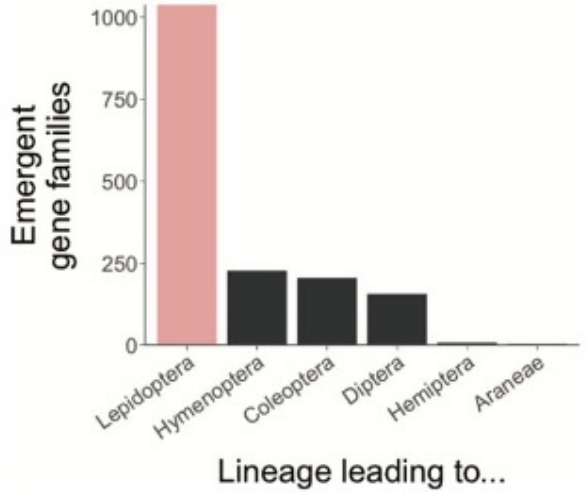
Last Holometabolous Common Ancestor: 10 emergent gene families

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."

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


THE
END
IS
NEAR

Post-genomics challenge

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

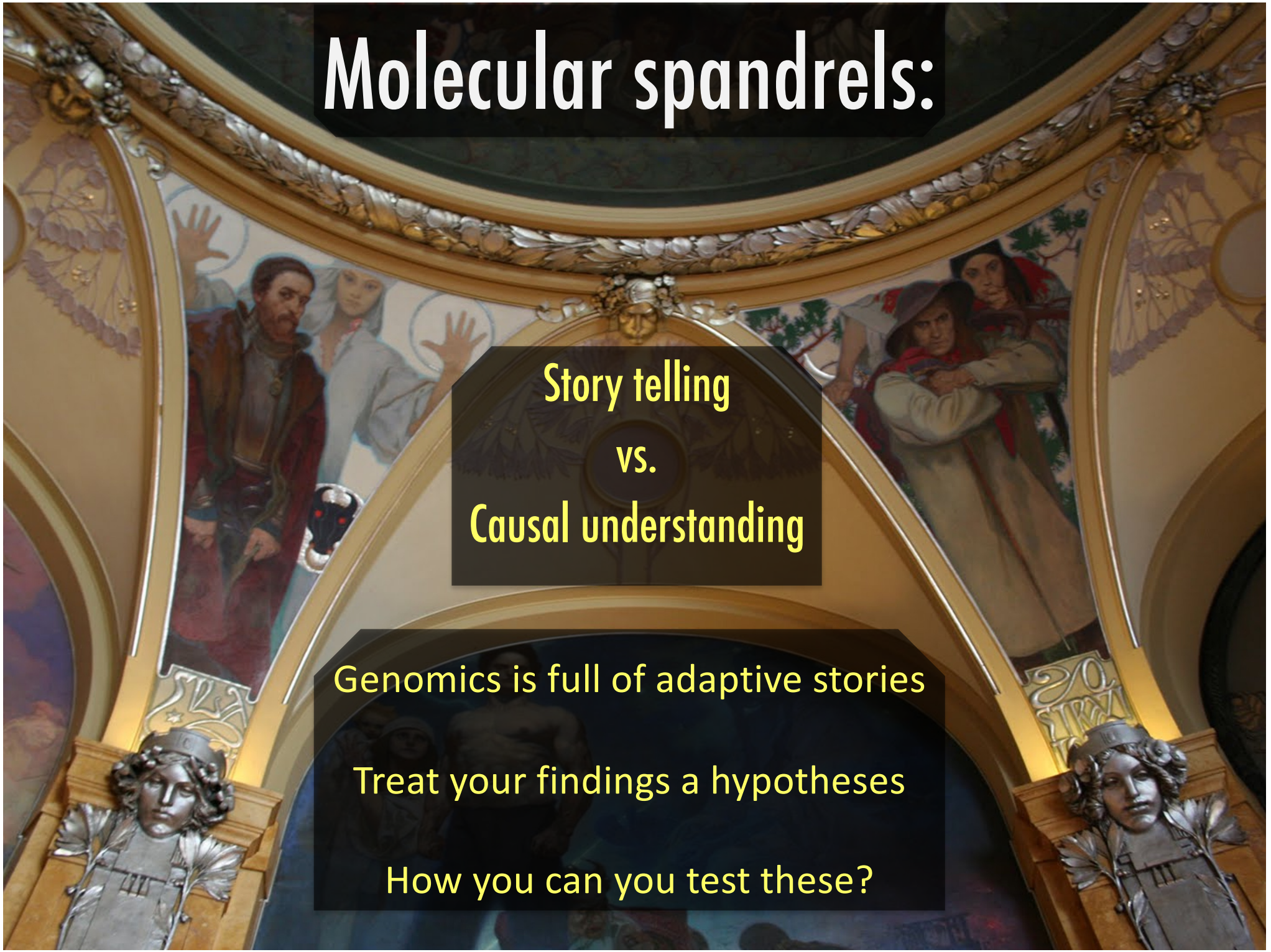
- Lewontin 1974

"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

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

Stockholm University

Thanks for a great workshop