University of
BRISTOL

Day 9 - Comparative Genomics

Attempting to make sense of how evolution works

F. Cicconardi, PhD

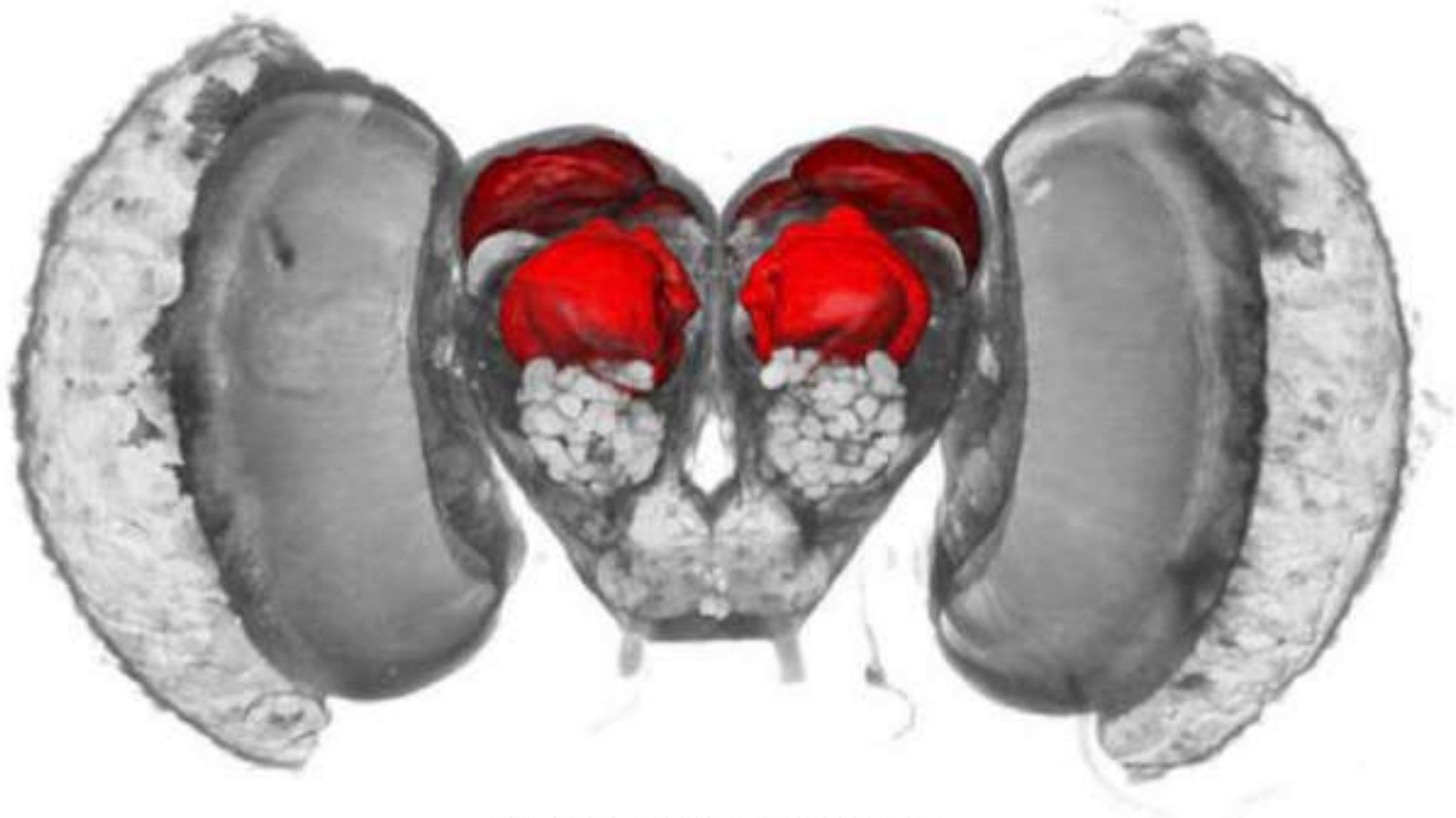
EBaB lab

2025 WORKSHOP ON GENOMICS, CESKY KRUMLOV

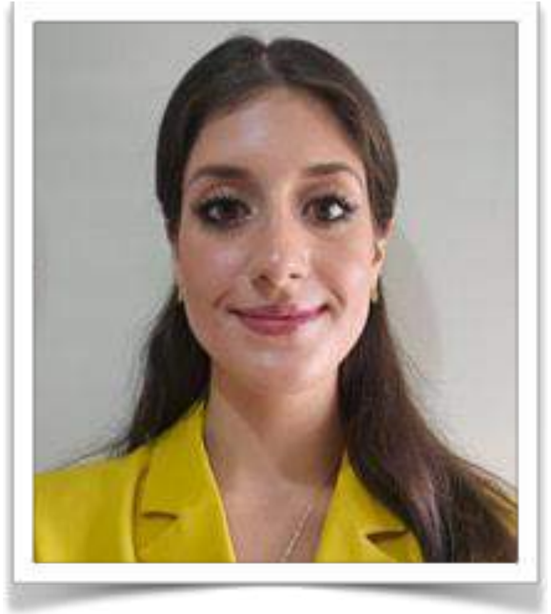
Evolution of Brain and Behaviour Lab

EBAB LAB

Dr. Stephen Montgomery



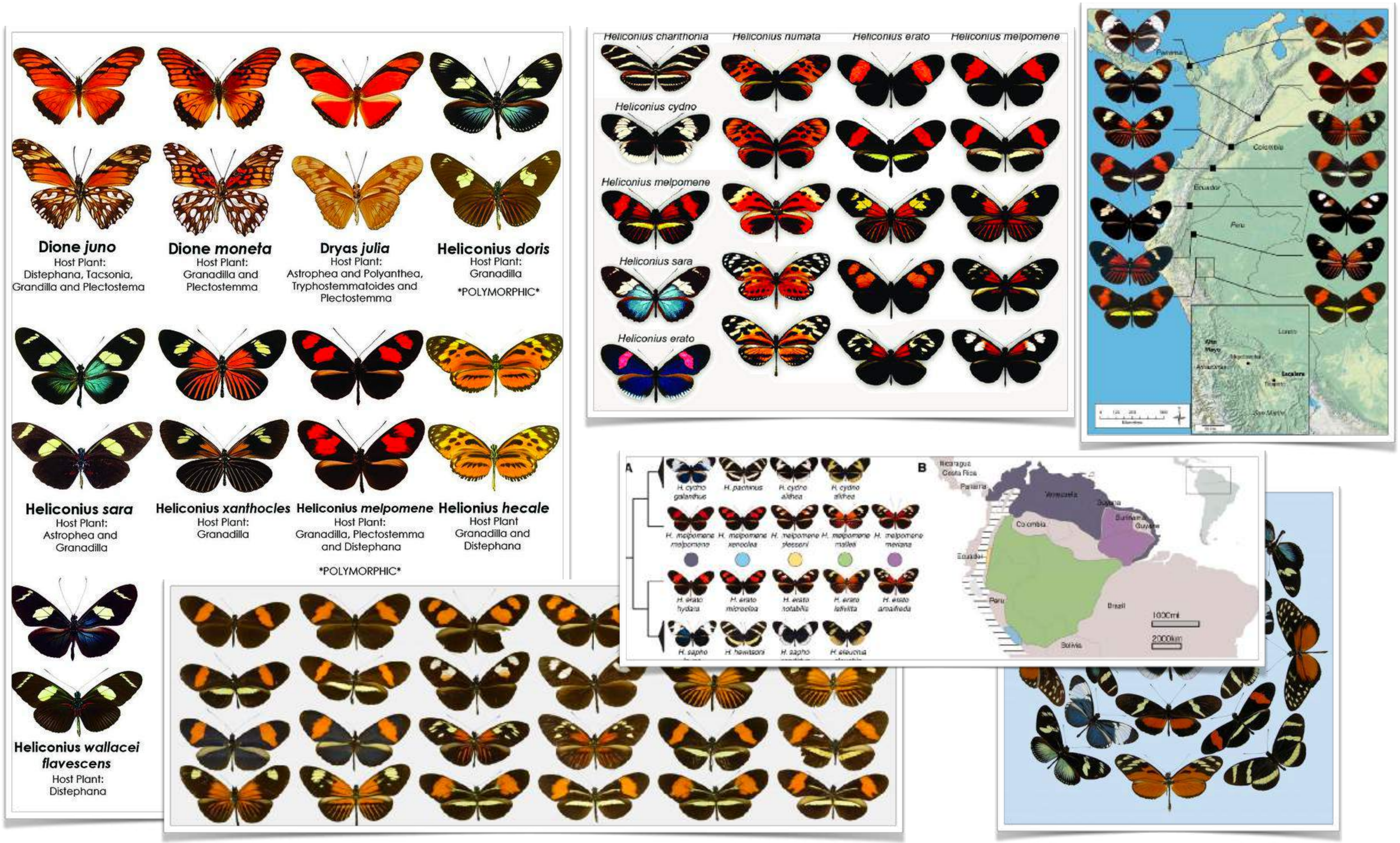
www.shmontgomery.co.uk

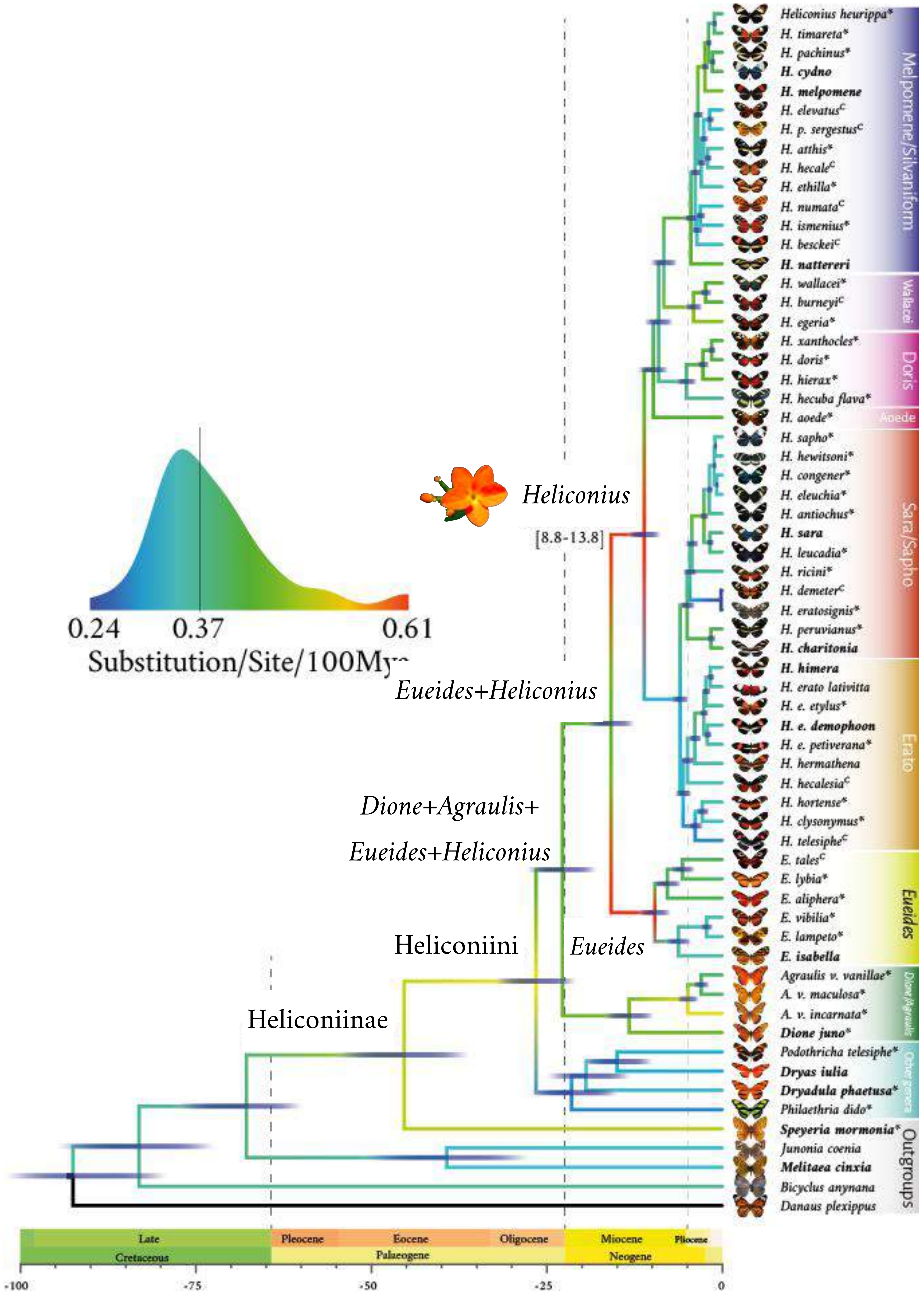


Adaptive Radiation of Heliconiini (Family: Nymphalidae)

Photo credit: @mena_sebas

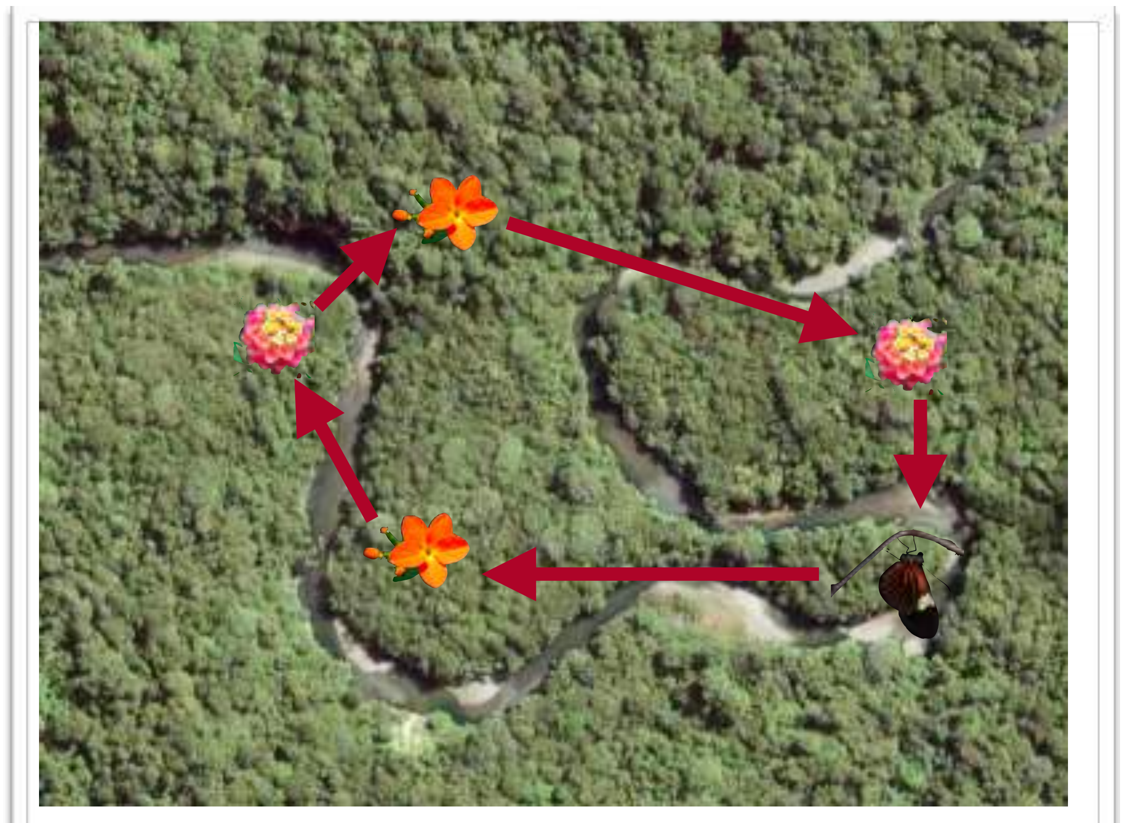
- *8 Genera*
- *87 Species*
- *~440 sub-species*





» Pollen-feeding

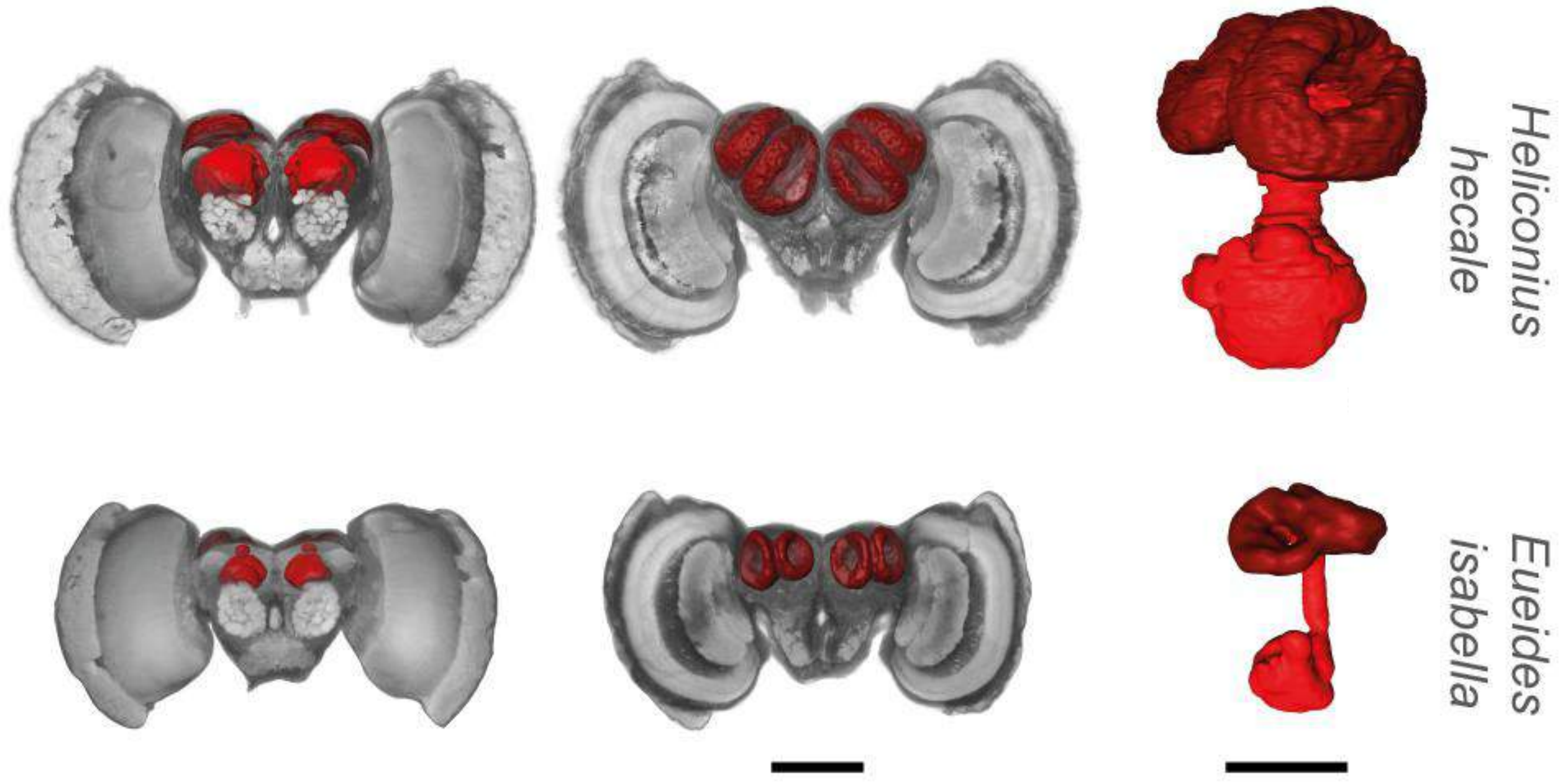
» Trap-lining behaviour

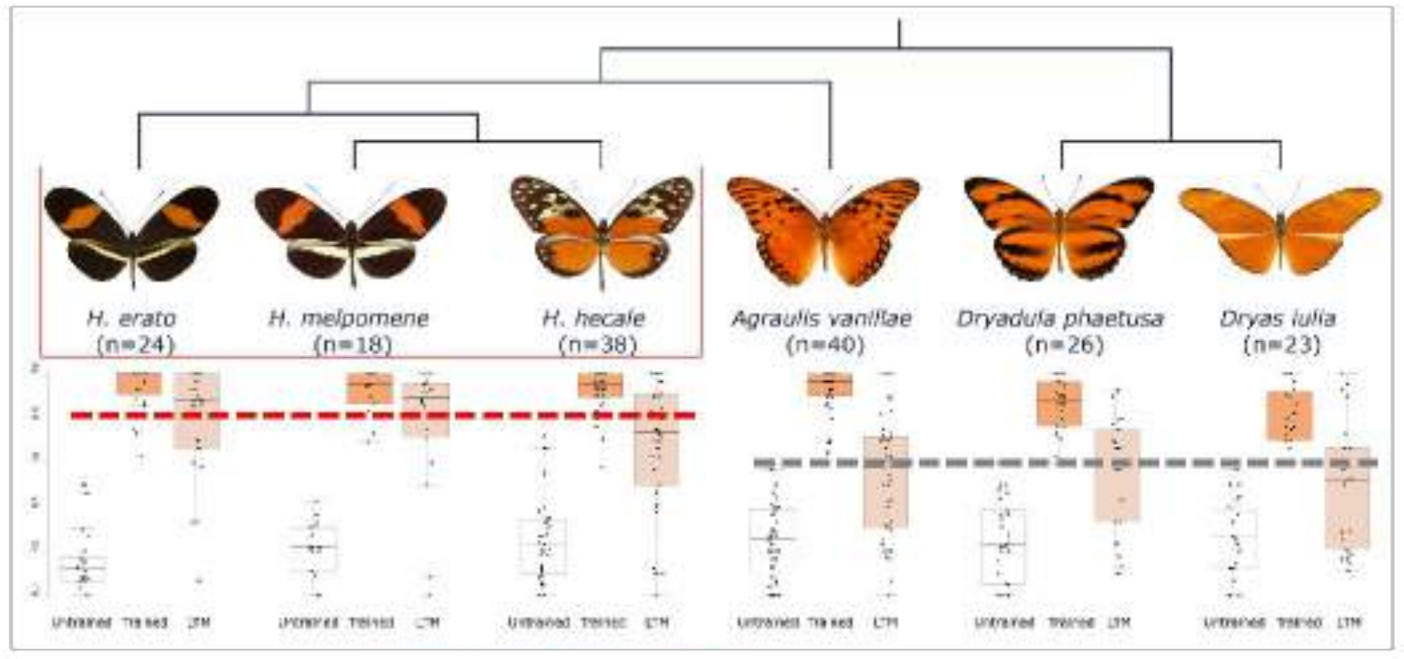


Allocentric trap-lining in home ranges up to 1 km^2

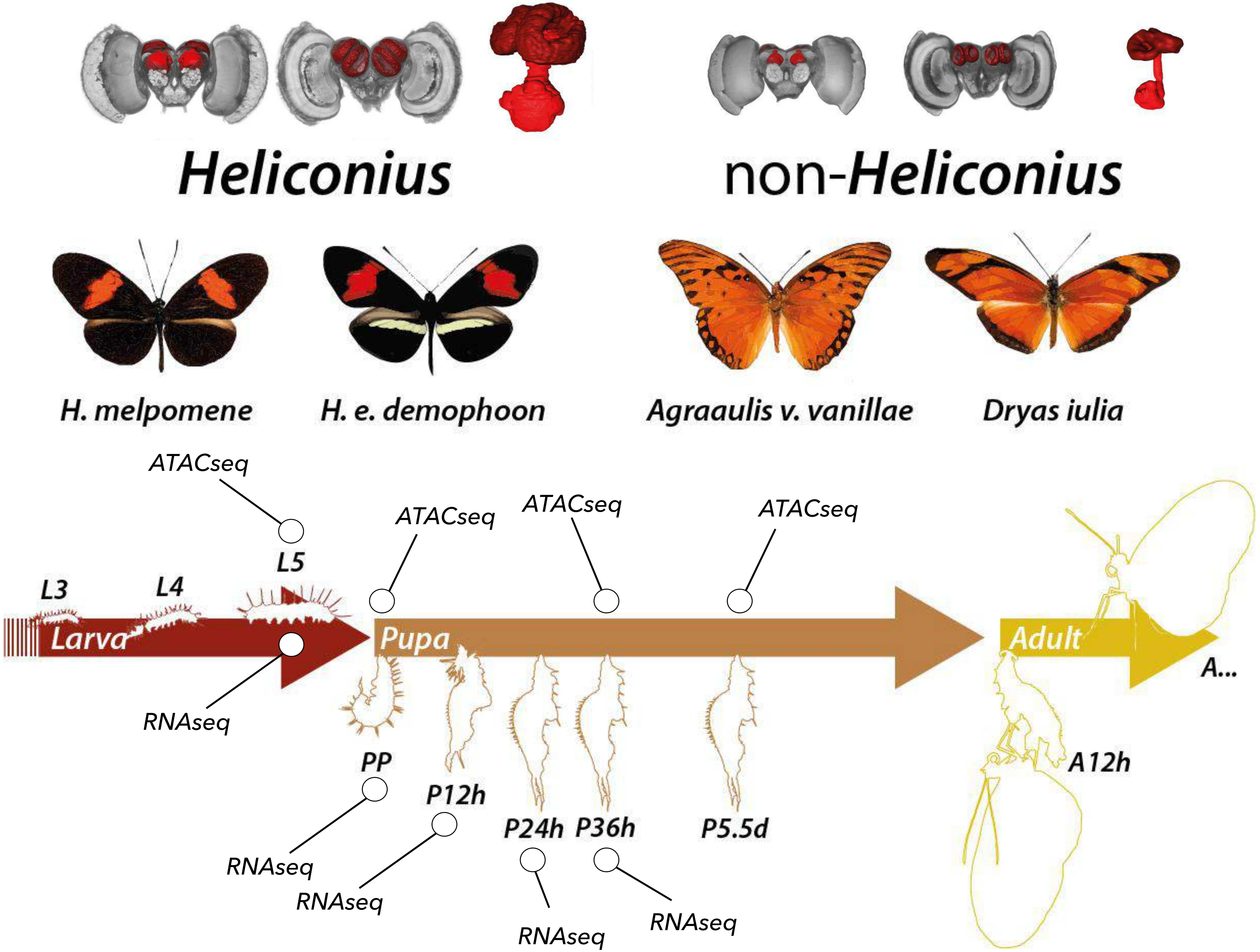
Moura et al.. **2021**. *Functional Ecoloav.*

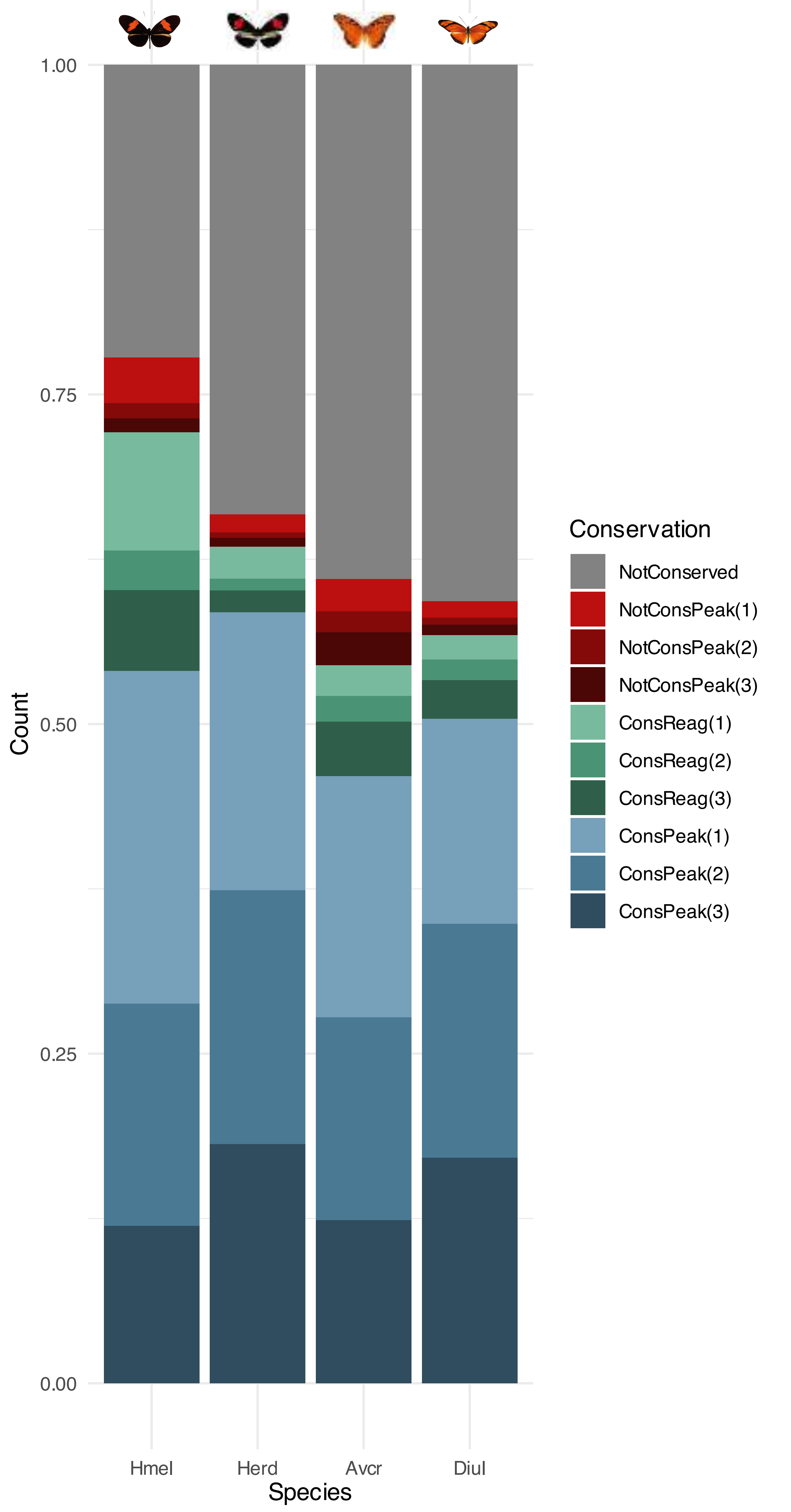
» Mushroom body expansion (Brain)



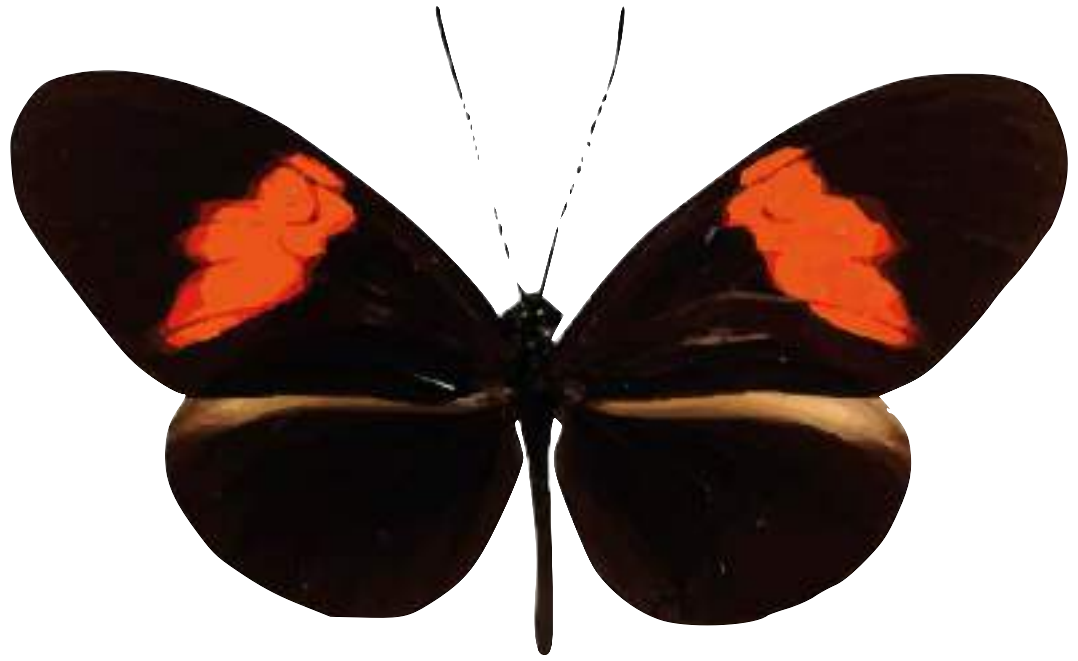


Interspecific variation: $X_2 = 76.2$, $p < 0.001$
Small vs. big MBs: $X_8 = 182.8$, $p < 0.001$

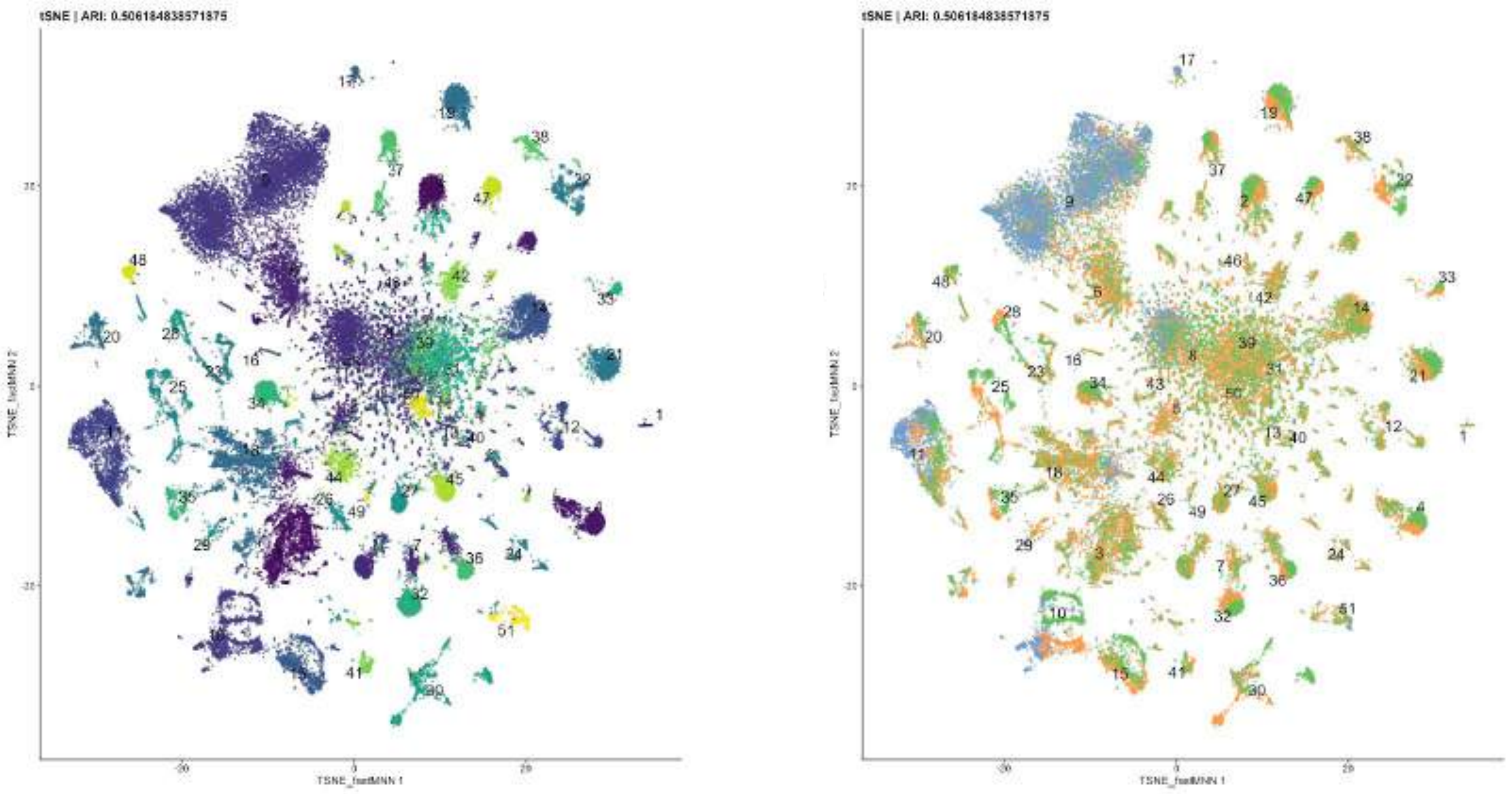


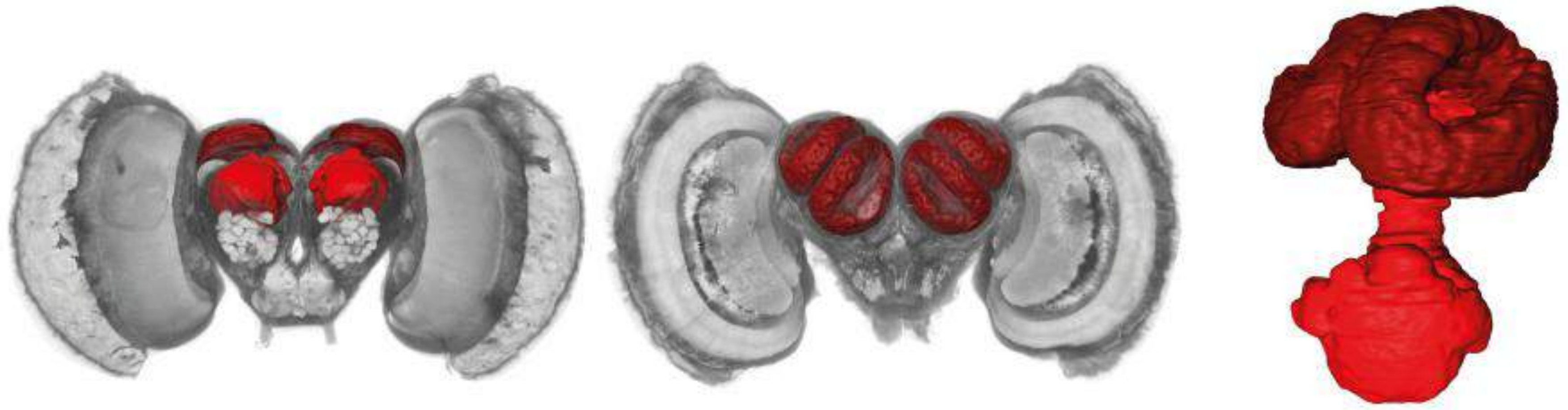

Stacked Histogram of Conservation Levels by Species

Heliconius



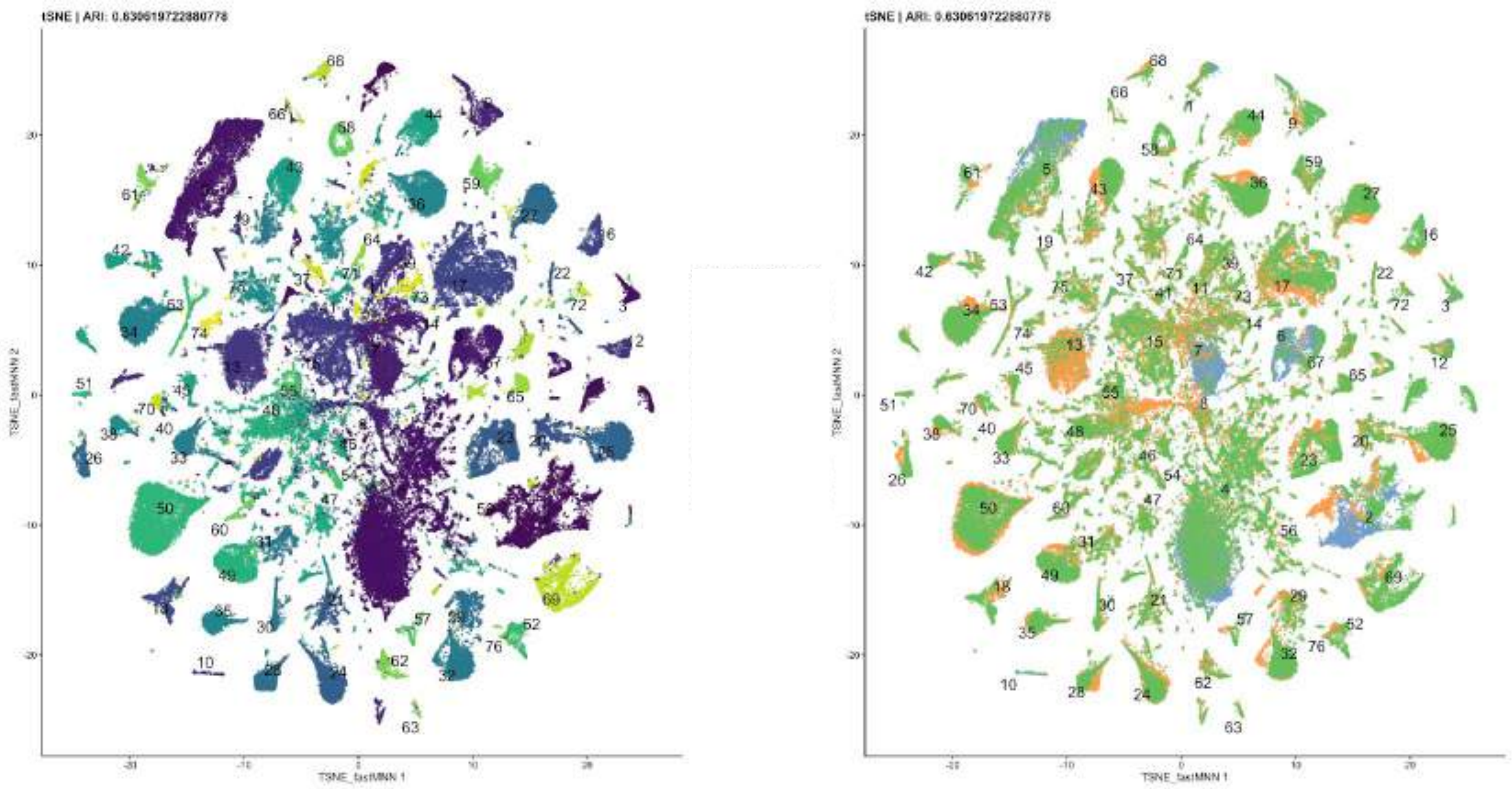
225,859 nuclei of *H. melpomene* brains

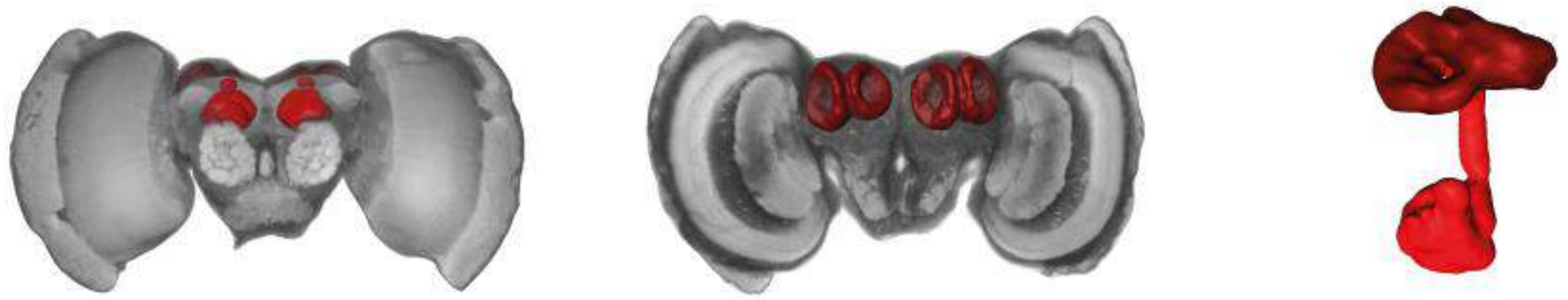




non-*Heliconius*

176,804 nuclei of *D. iulia* brains





» But how do you tackle these problems? »

» How do we make sense of how evolution works? »

» But how do you tackle these problems? »

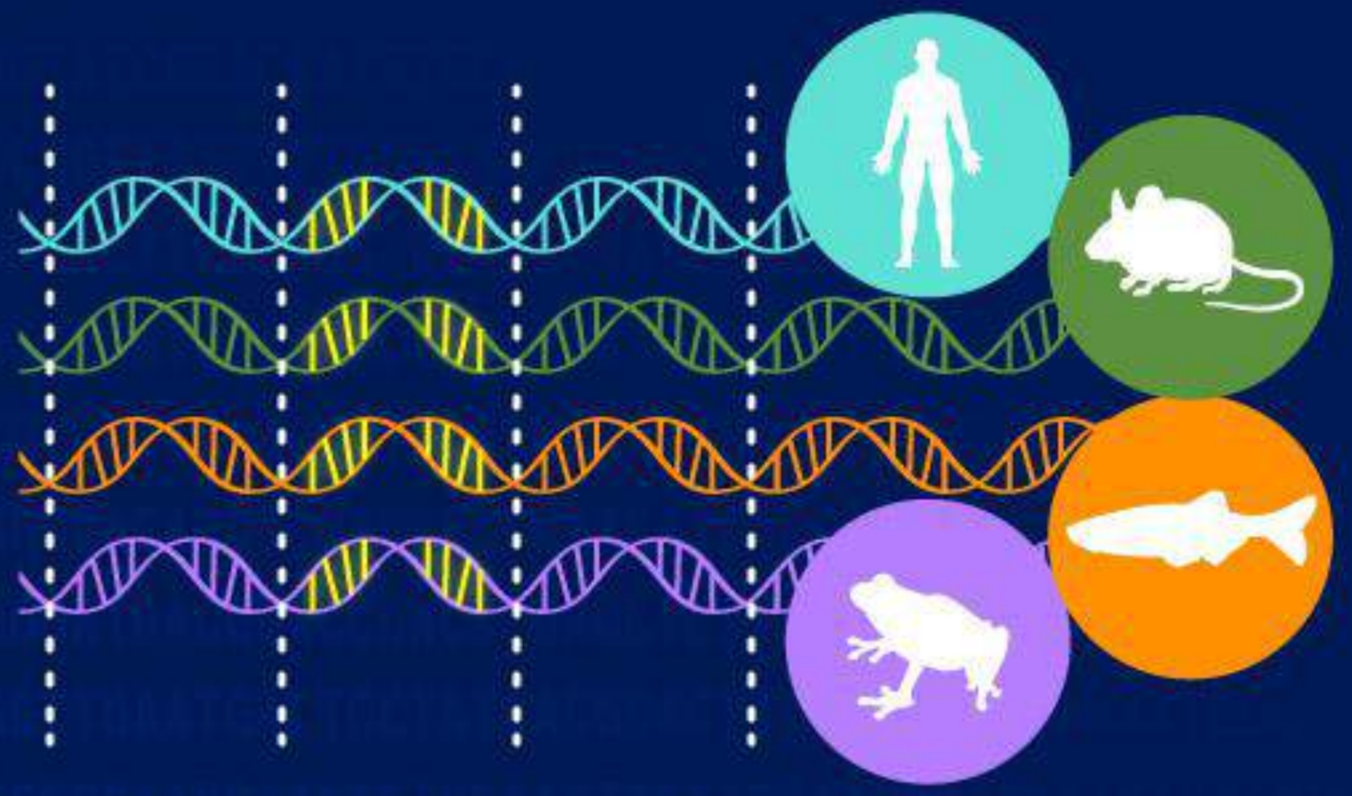
» How do we make sense of how evolution works? »

» **Comparing “things”!** »

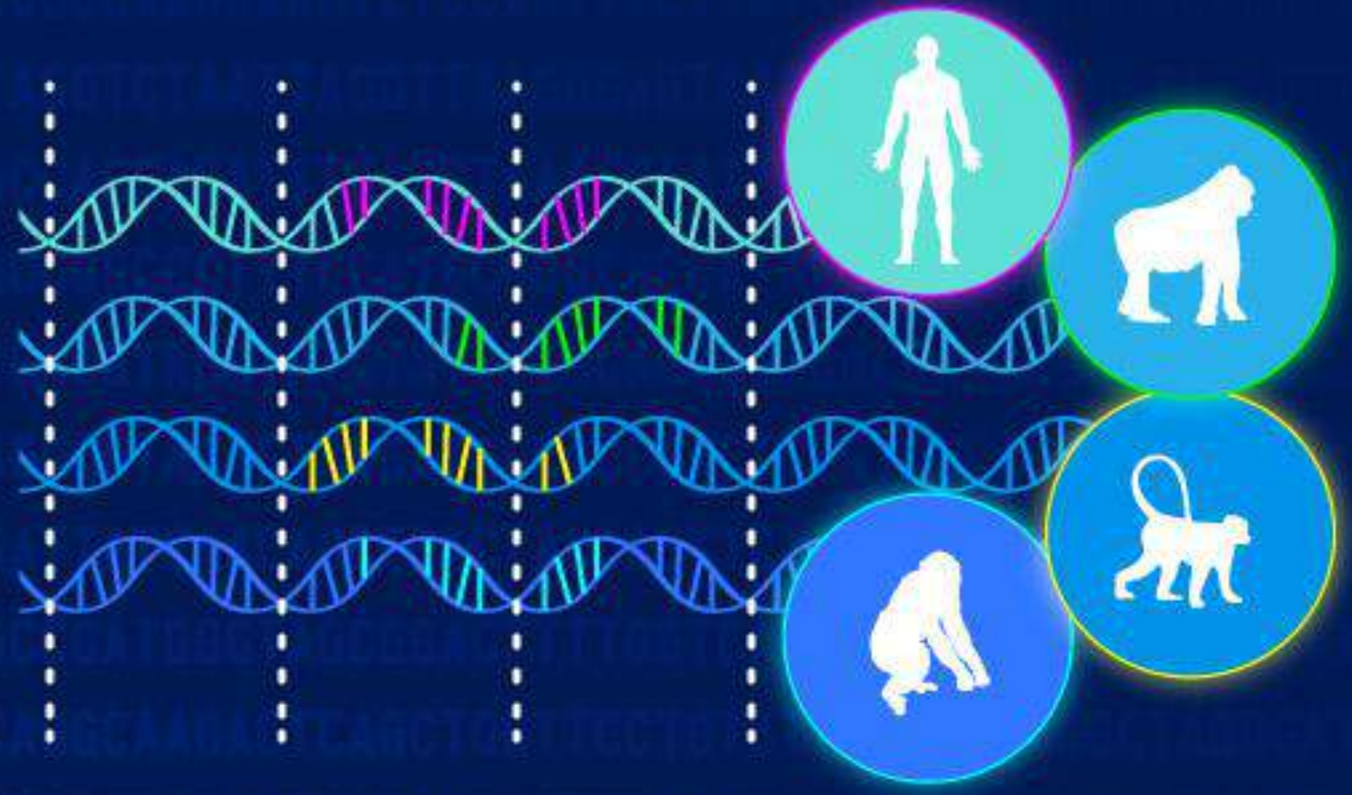
Comparative Genomics

NHGRI FACT SHEETS
genome.gov

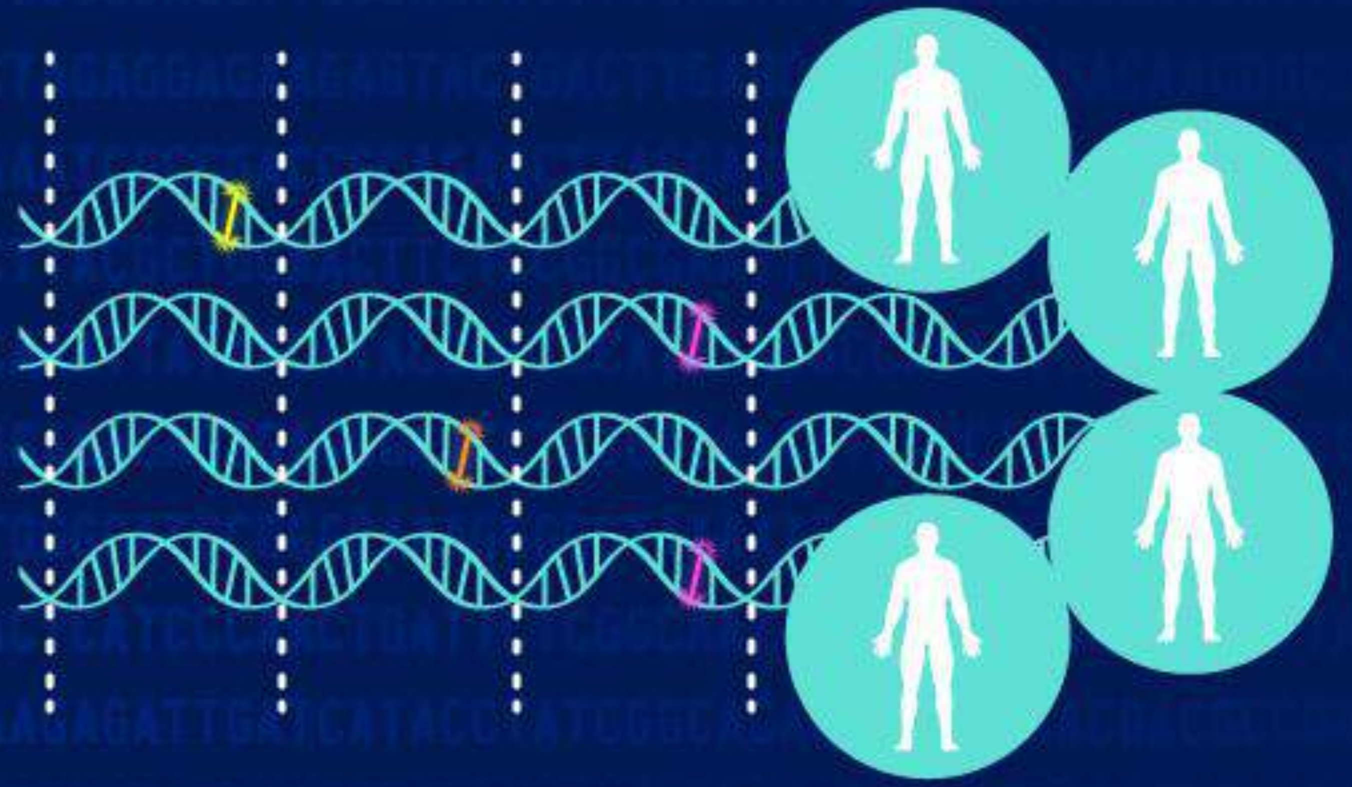
Researchers choose the appropriate time-scale of evolutionary conservation for the question being addressed.



Common features of different organisms such as humans and fish are often encoded within the DNA evolutionarily conserved between them.

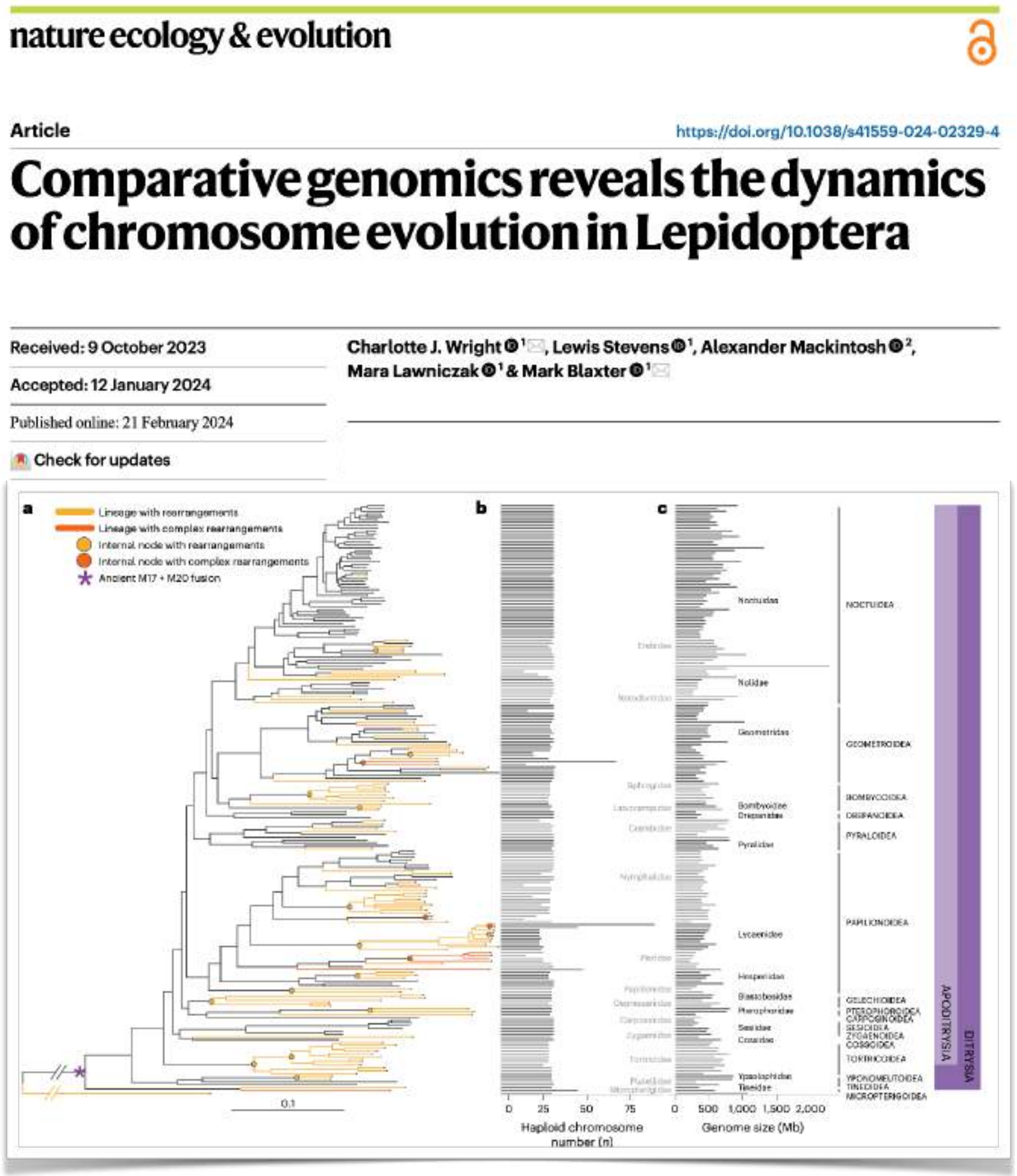


Looking at closely related species such as humans and chimpanzees shows which genomic elements are unique to each.



Genetic differences within one species such as our own can reveal variants with a role in disease.

NIH National Human Genome Research Institute



nature ecology & evolution

Article

https://doi.org/10.1038/s41559-024-02329-4

Comparative genomics reveals the dynamics of chromosome evolution in Lepidoptera

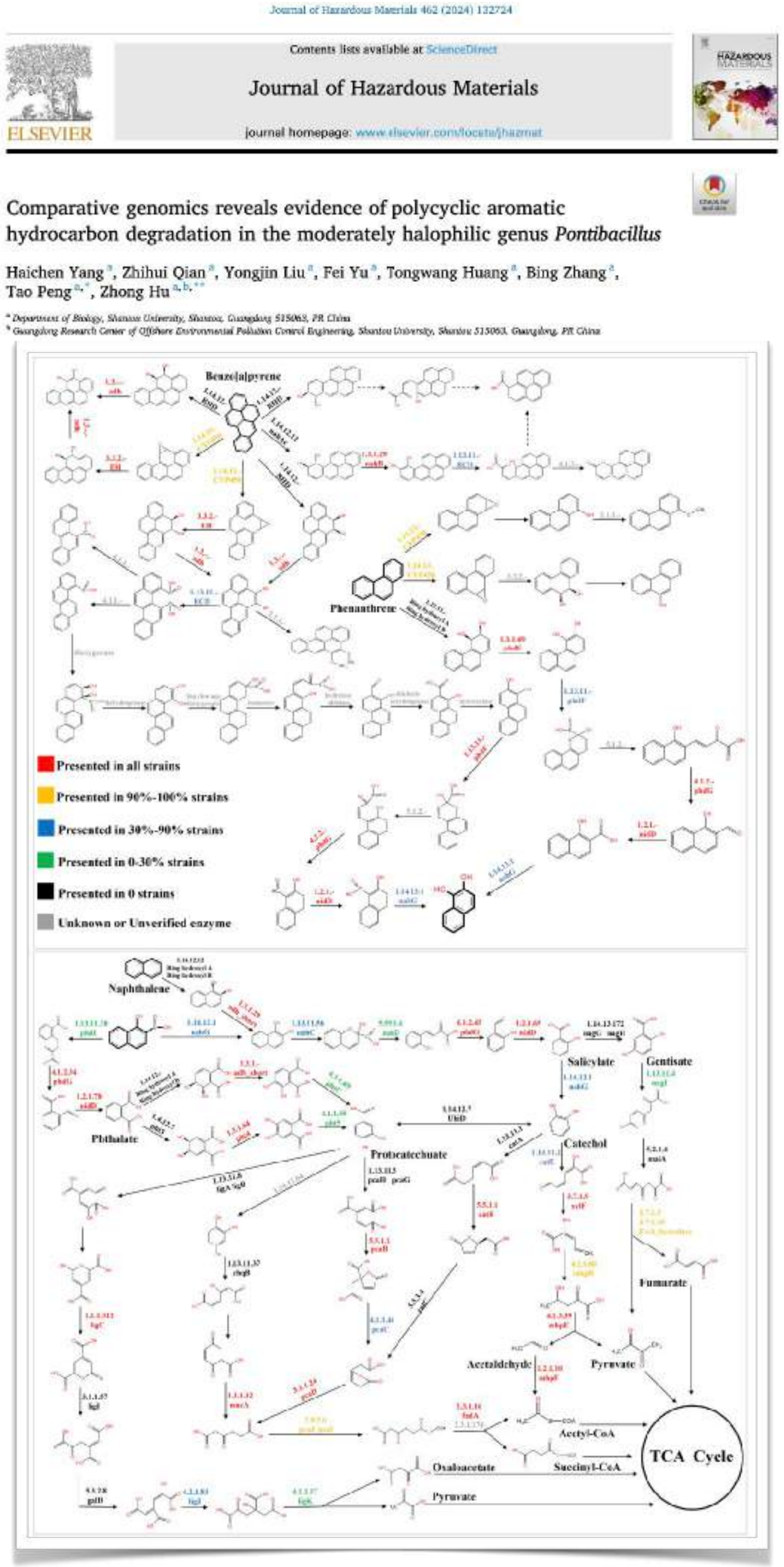
Received: 9 October 2023

Accepted: 12 January 2024

Published online: 21 February 2024

Check for updates

Charlotte J. Wright[1], Lewis Stevens[1], Alexander Mackintosh[2], Mara Lawniczak[1] & Mark Blaxter[1]



Journal of Hazardous Materials 462 (2024) 132724

Contents lists available at ScienceDirect

Journal of Hazardous Materials

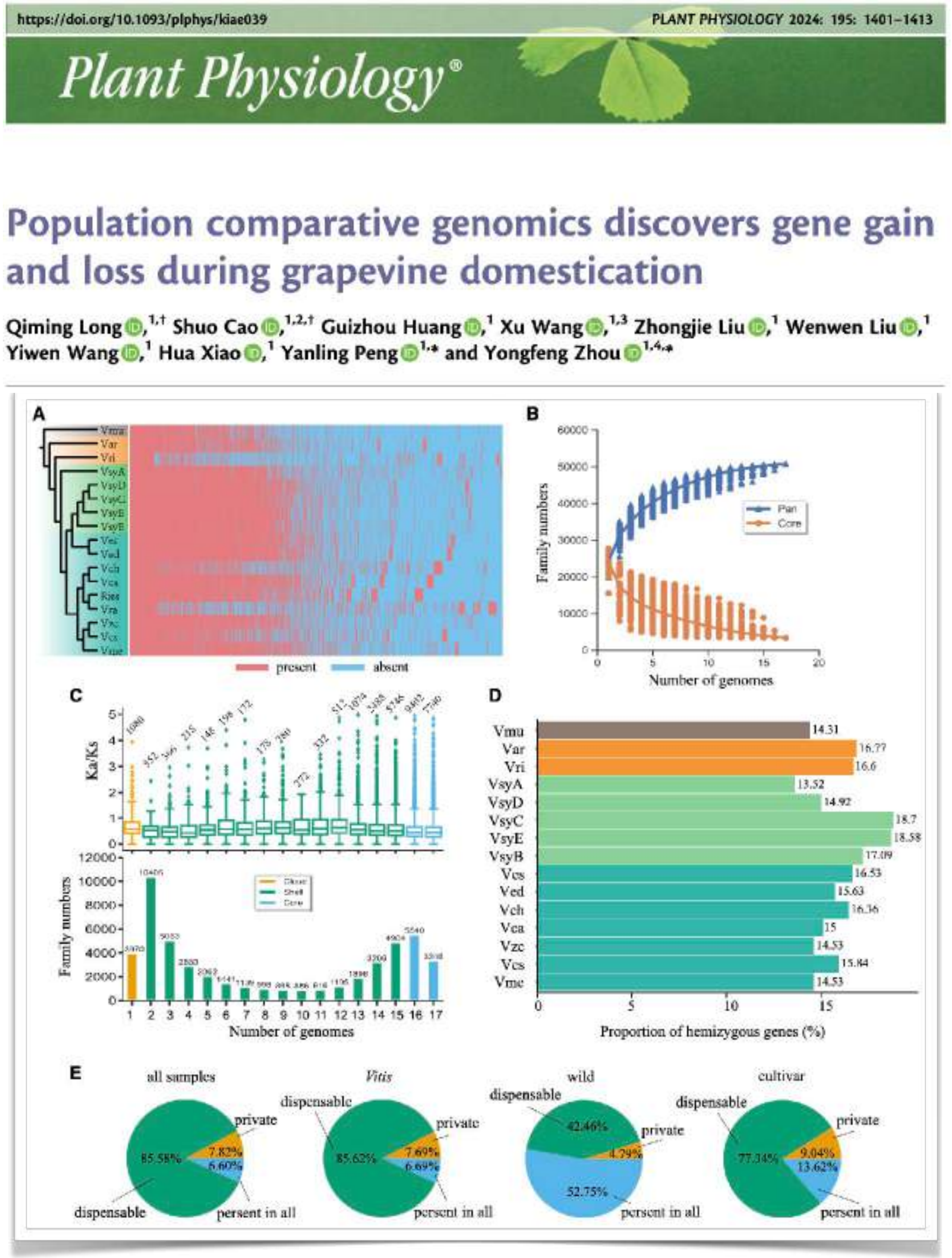
journal homepage: www.elsevier.com/locate/jhazmat

Comparative genomics reveals evidence of polycyclic aromatic hydrocarbon degradation in the moderately halophilic genus *Pontibacillus*

Haichen Yang[a], Zhihui Qian[a], Yongjin Liu[a], Fei Yu[a], Tongwang Huang[a], Bing Zhang[a], Tao Peng[a,*], Zhong Hu[a,b,**]

[a] Department of Biology, Shantou University, Shantou, Guangdong 515063, PR China

[b] Guangdong Research Center of Offshore Environmental Pollution Control Engineering, Shantou University, Shantou 515063, Guangdong, PR China



https://doi.org/10.1093/plphys/kiae039

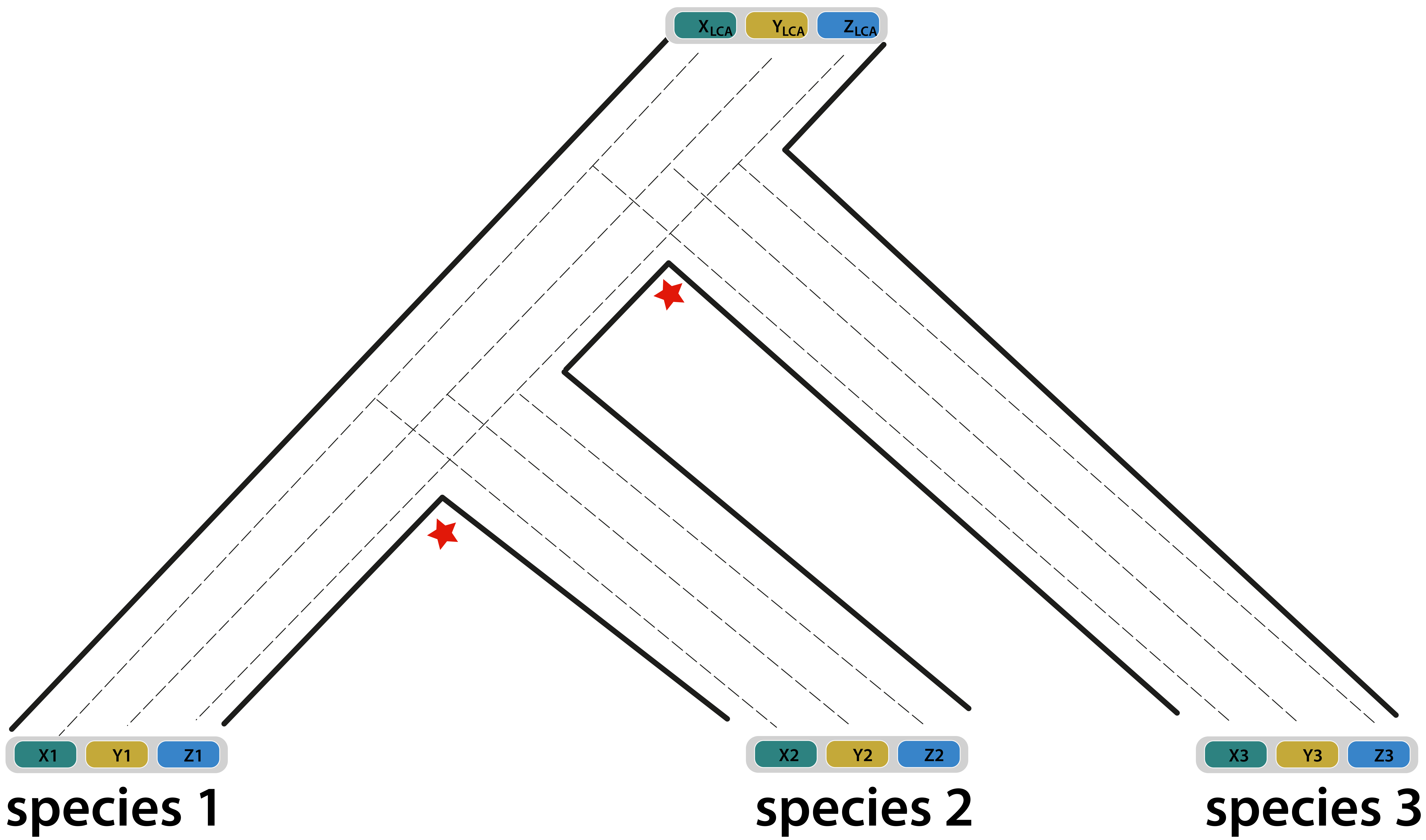
PLANT PHYSIOLOGY 2024: 195: 1401–1413

Plant Physiology®

Population comparative genomics discovers gene gain and loss during grapevine domestication

Qiming Long[1,†], Shuo Cao[1,2,†], Guizhou Huang[1], Xu Wang[1,3], Zhongjie Liu[1], Wenwen Liu[1], Yiwen Wang[1], Hua Xiao[1], Yanling Peng[1,*] and Yongfeng Zhou[1,4,*]

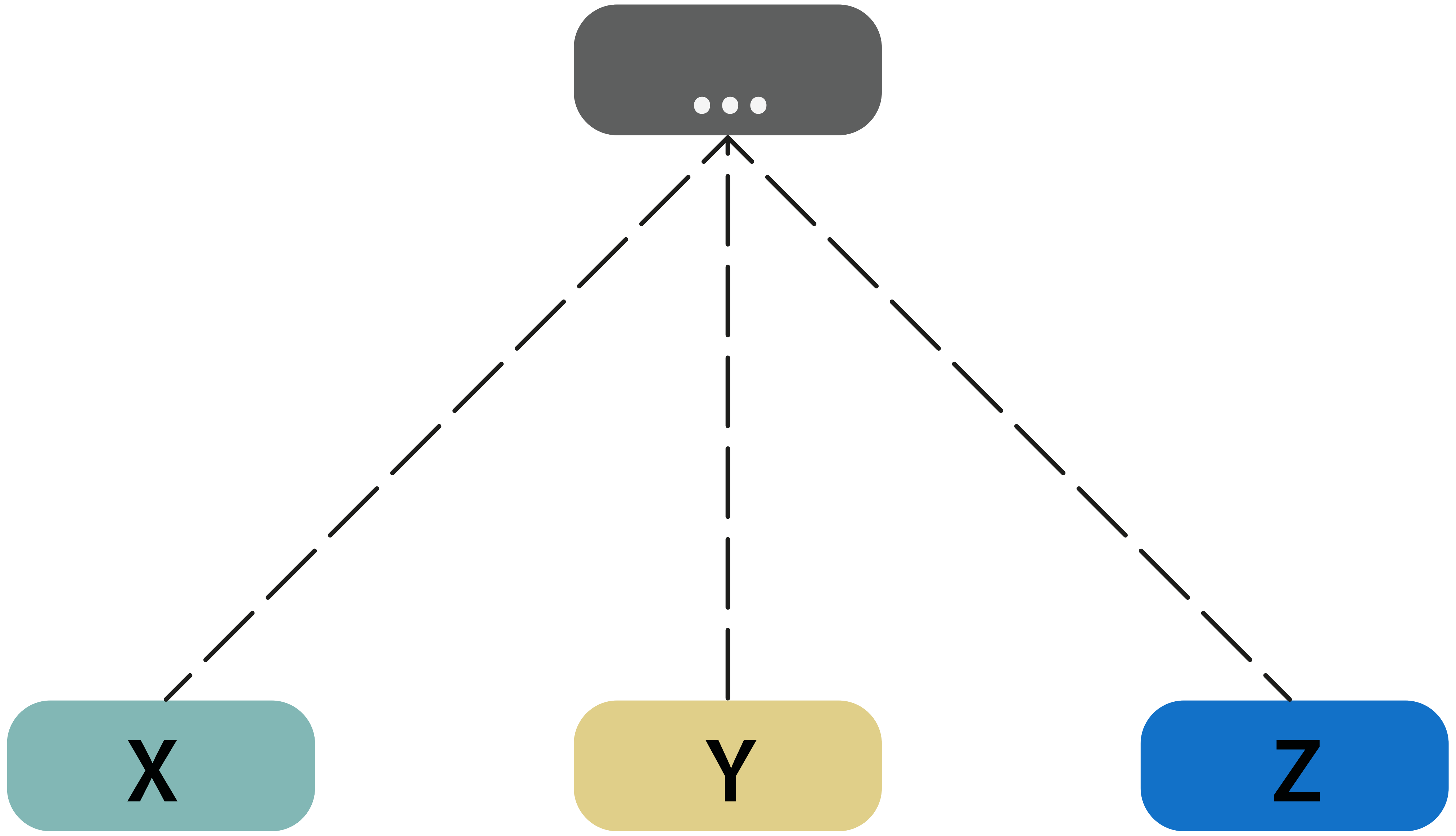
» … Some definitions … »



Orthologs, Paralogs, and Evolutionary Genomics | https://www.annualreviews.org/content/journals/10.1146/annurev.genet.39.073003.114725
Functional and evolutionary implications of gene orthology | https://www.nature.com/articles/nrg3456

» Homology/Orthology definition »

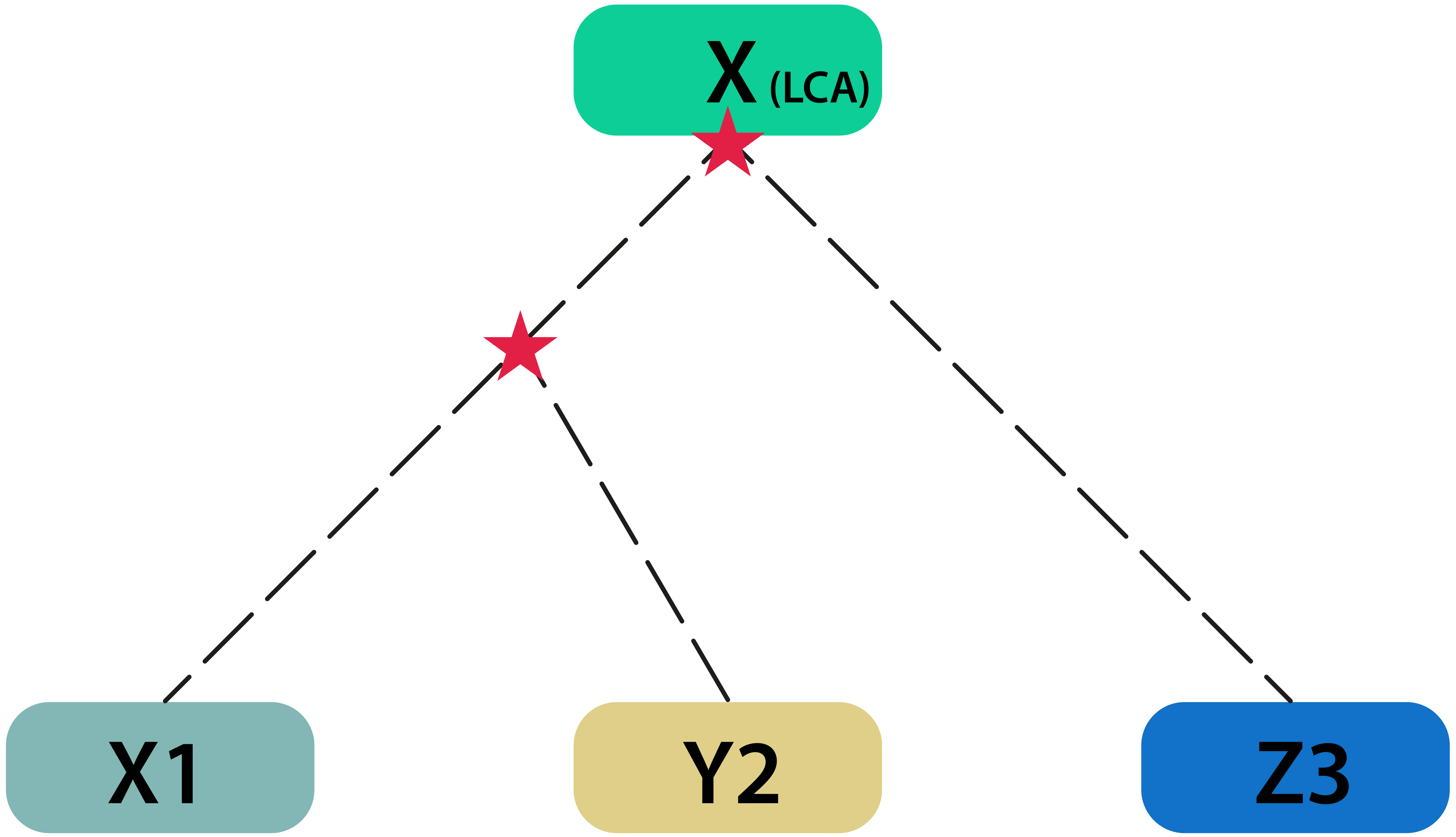
Homology: it describes descent from a common evolutionary origin: *two loci (genes) are homologous if they derive from the same ancestral locus (gene).*



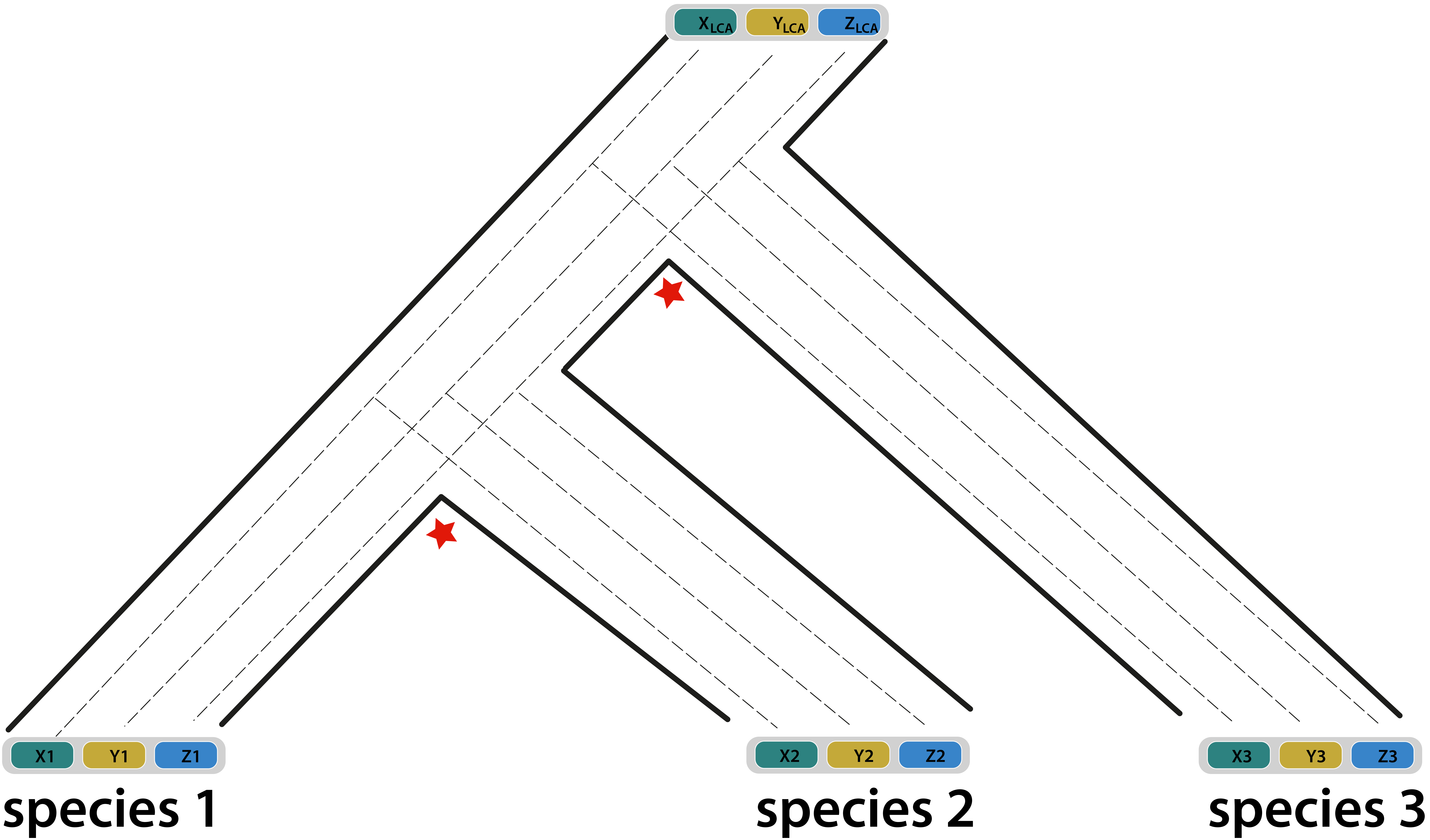
Orthologs, Paralogs, and Evolutionary Genomics | https://www.annualreviews.org/content/journals/10.1146/annurev.genet.39.073003.114725
Functional and evolutionary implications of gene orthology | https://www.nature.com/articles/nrg3456

» Homology/Orthology definition »

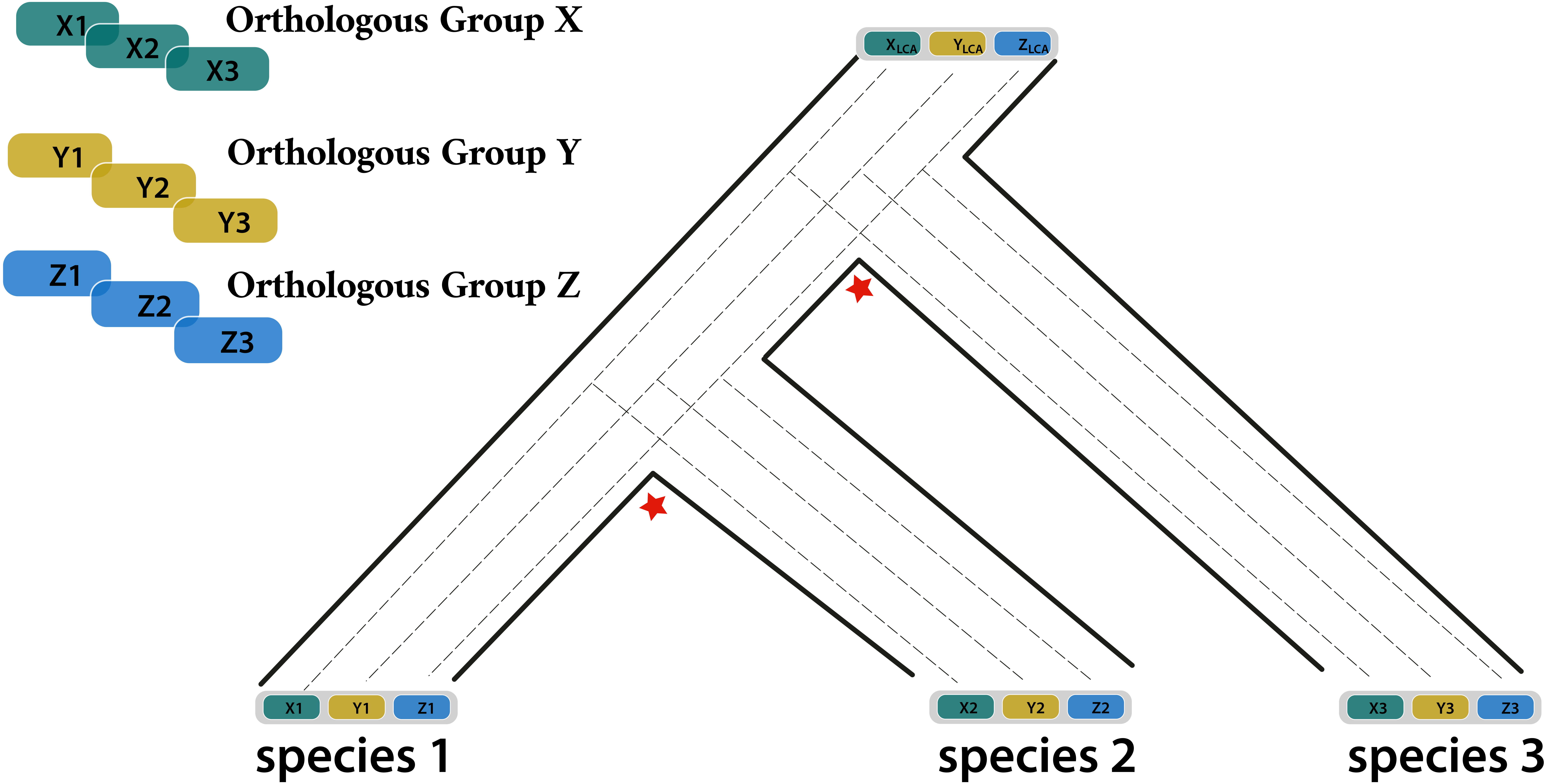
Orthology: it describes a relationship due to a <u>speciation event</u>:
two loci (genes) are orthologous if they derive from speciation.



Orthologs, Paralogs, and Evolutionary Genomics | https://www.annualreviews.org/content/journals/10.1146/annurev.genet.39.073003.114725
Functional and evolutionary implications of gene orthology | https://www.nature.com/articles/nrg3456



Orthologs, Paralogs, and Evolutionary Genomics | https://www.annualreviews.org/content/journals/10.1146/annurev.genet.39.073003.114725
Functional and evolutionary implications of gene orthology | https://www.nature.com/articles/nrg3456



Orthologous Group X

Orthologous Group Y

Orthologous Group Z

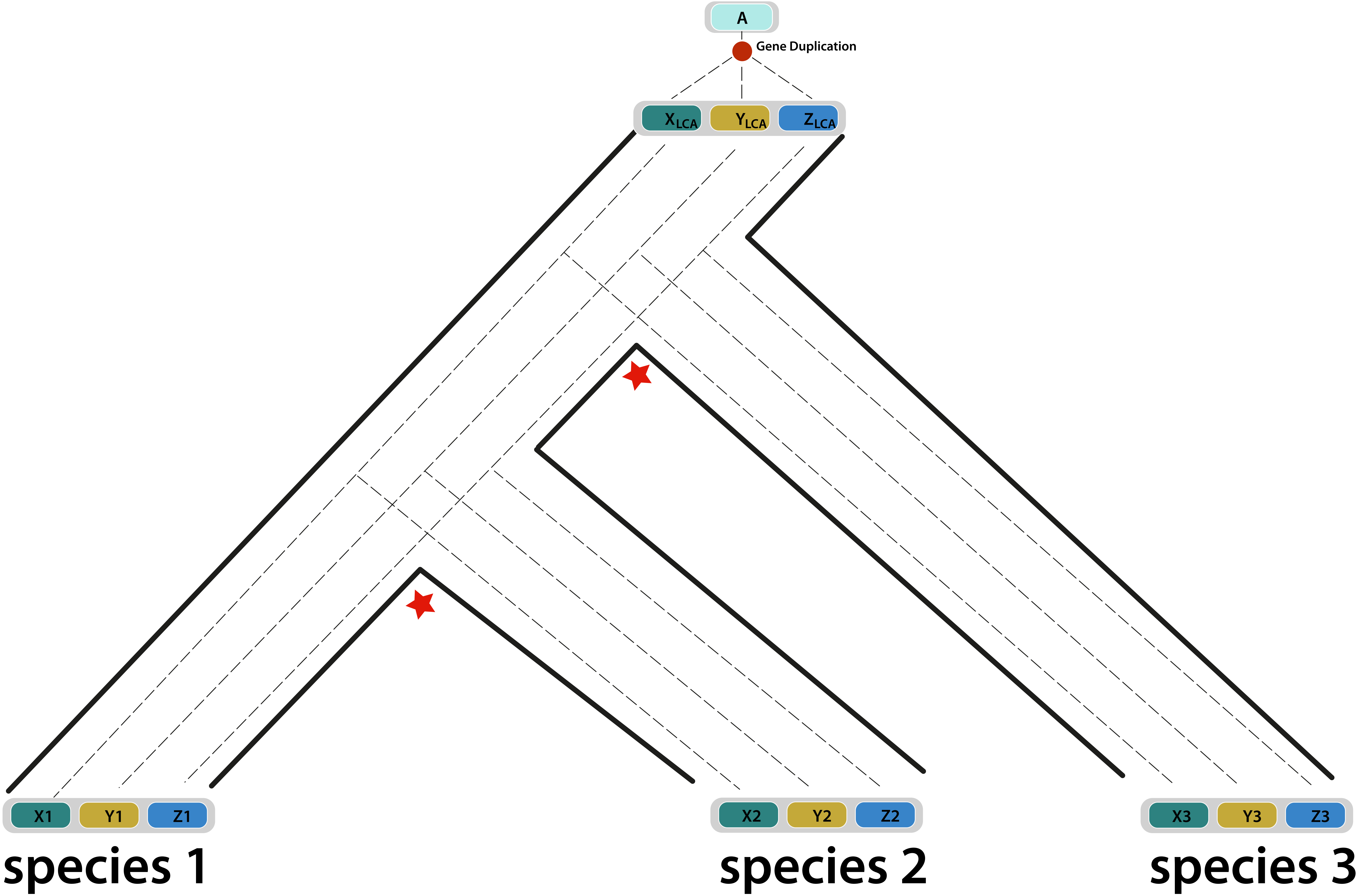
species 1

species 2

species 3

Orthologs, Paralogs, and Evolutionary Genomics | https://www.annualreviews.org/content/journals/10.1146/annurev.genet.39.073003.114725

Functional and evolutionary implications of gene orthology | https://www.nature.com/articles/nrg3456



Orthologs, Paralogs, and Evolutionary Genomics | https://www.annualreviews.org/content/journals/10.1146/annurev.genet.39.073003.114725

Functional and evolutionary implications of gene orthology | https://www.nature.com/articles/nrg3456

X1 X2 X3

Orthologous Group X

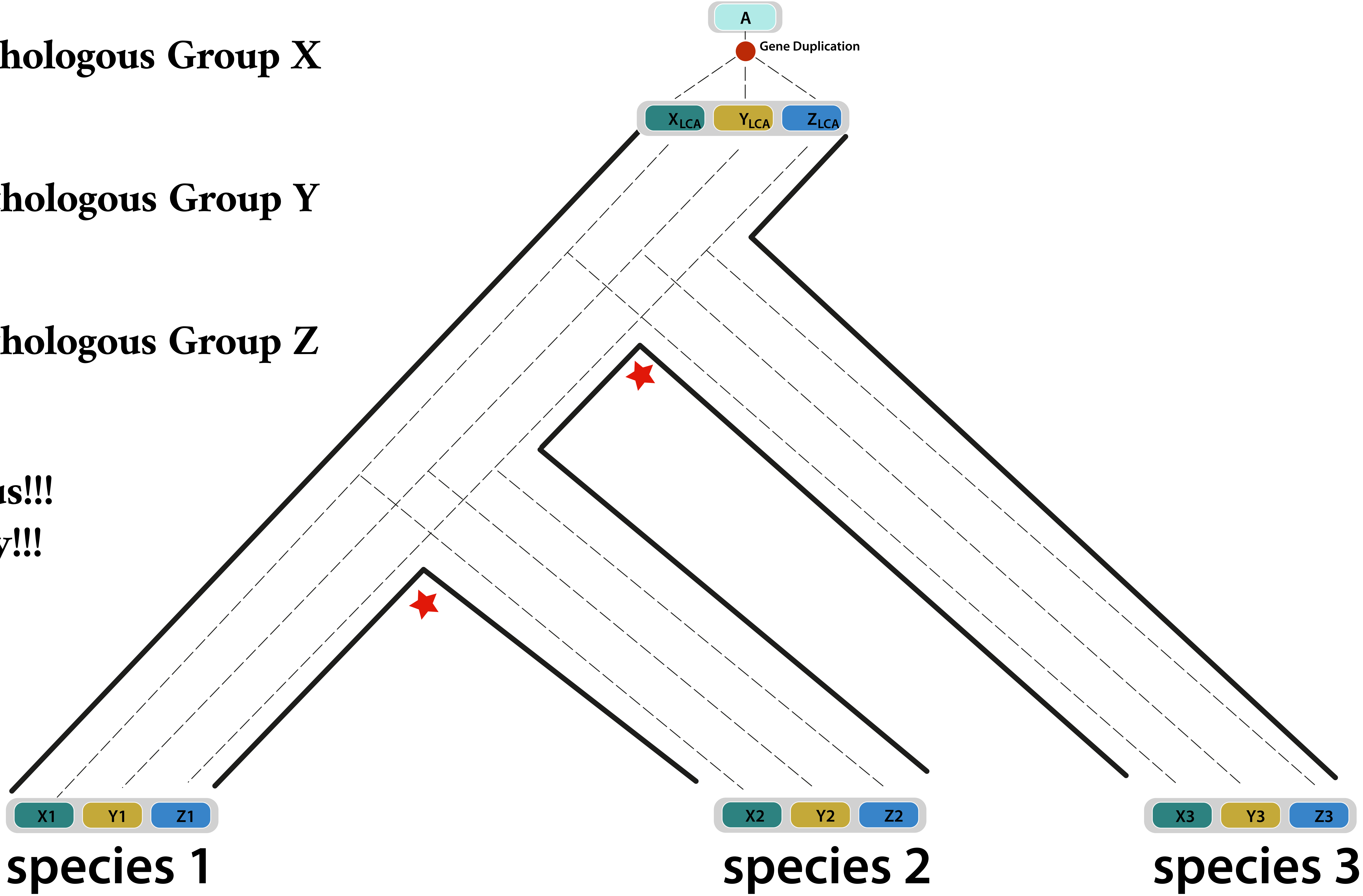
Y1 Y2 Y3

Orthologous Group Y

Z1 Z2 Z3

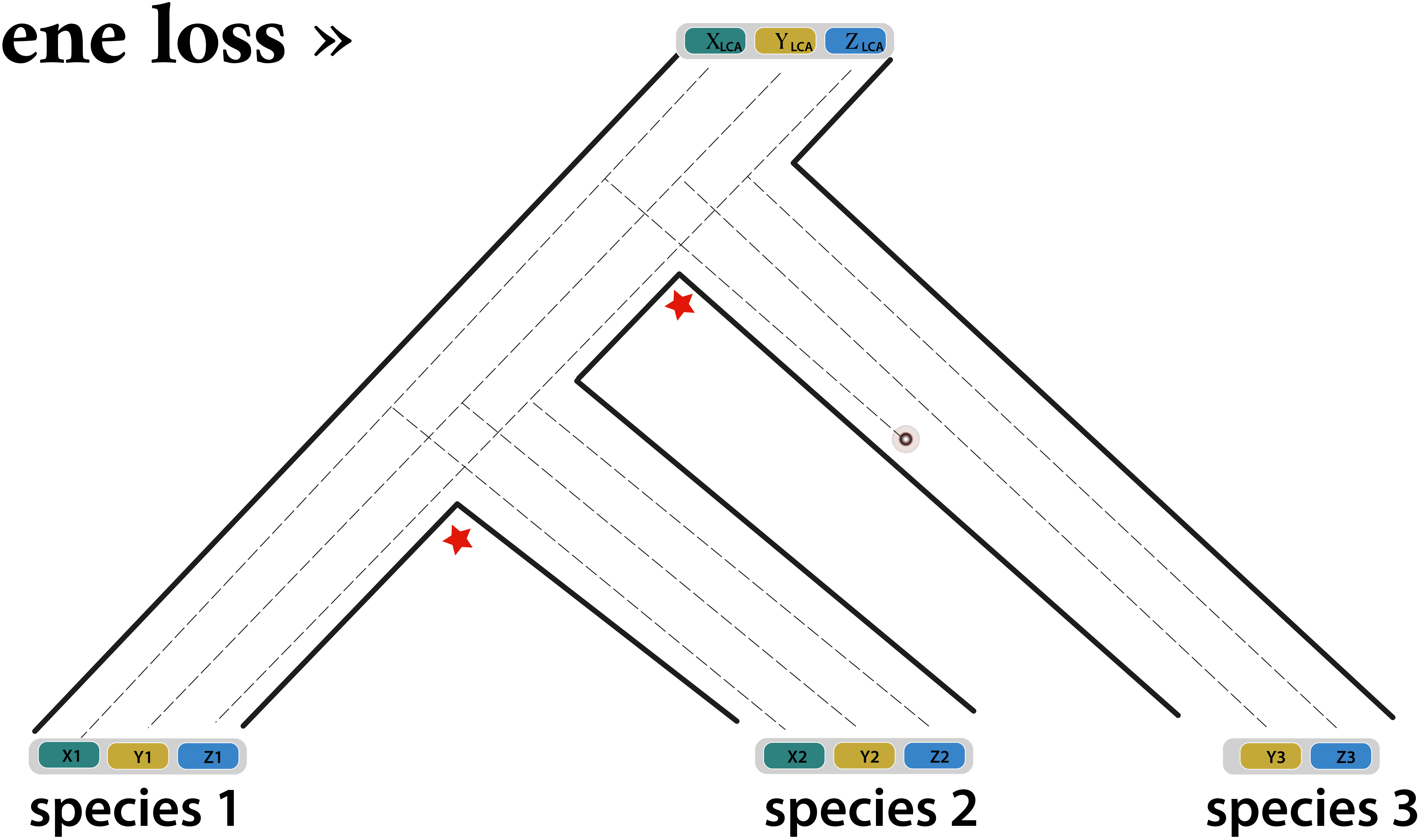
Orthologous Group Z

All Homologous!!!
All Single-copy!!!



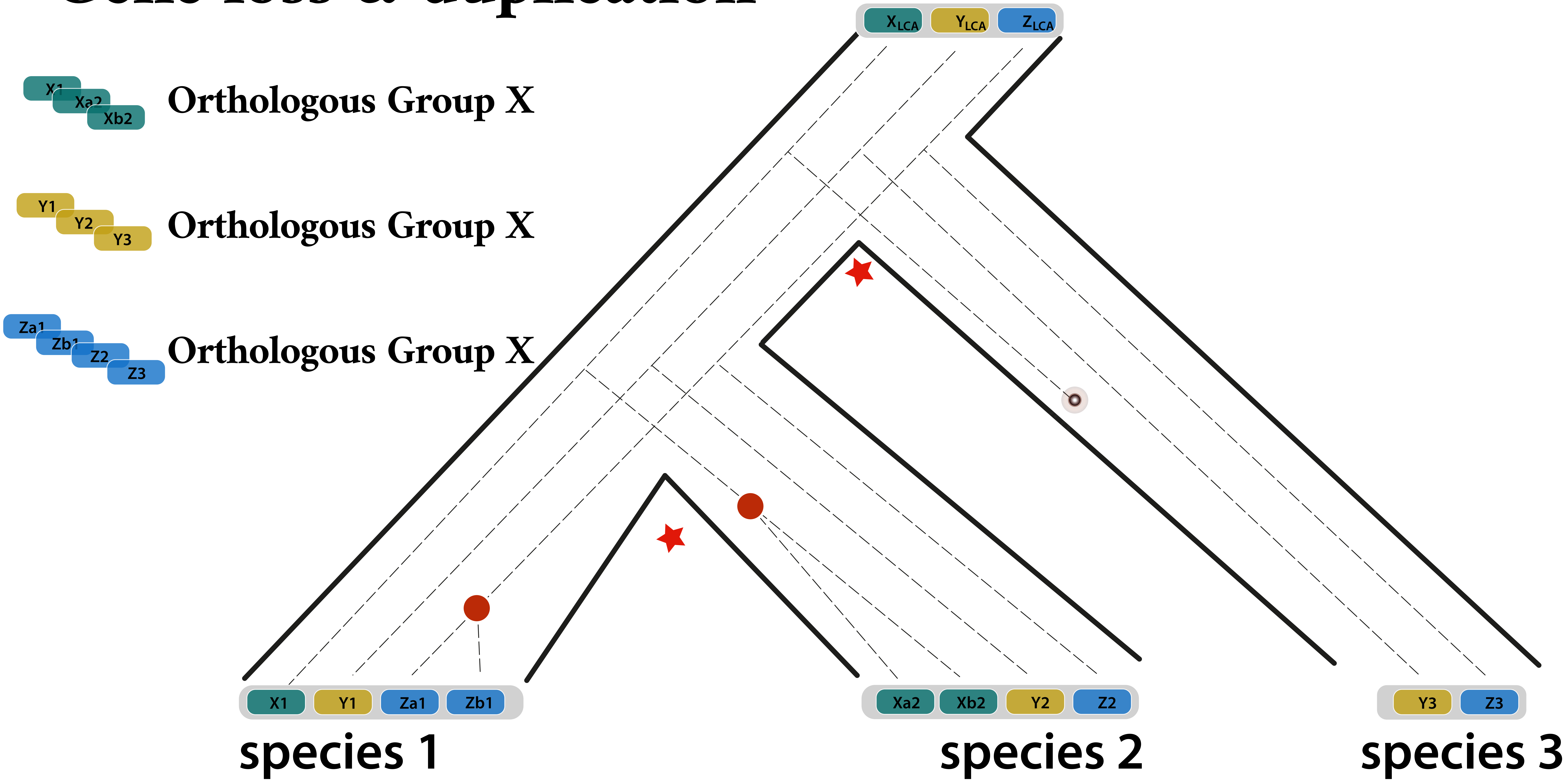
Orthologs, Paralogs, and Evolutionary Genomics | *https://www.annualreviews.org/content/journals/10.1146/annurev.genet.39.073003.114725*
Functional and evolutionary implications of gene orthology | *https://www.nature.com/articles/nrg3456*

» Gene loss »



Orthologs, Paralogs, and Evolutionary Genomics | https://www.annualreviews.org/content/journals/10.1146/annurev.genet.39.073003.114725
Functional and evolutionary implications of gene orthology | https://www.nature.com/articles/nrg3456

» Gene loss & duplication »



Orthologous Group X

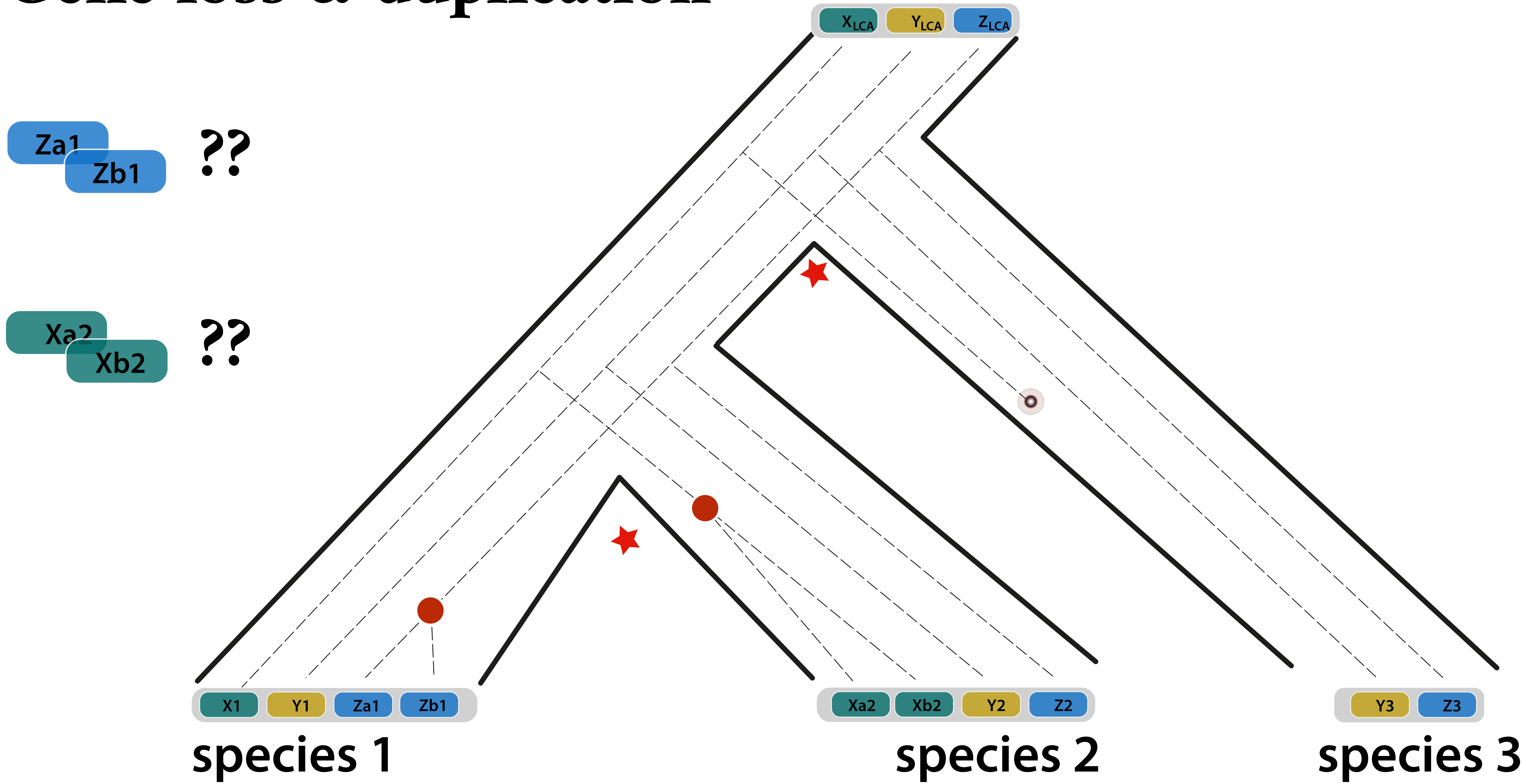
Orthologous Group X

Orthologous Group X

Orthologs, Paralogs, and Evolutionary Genomics | https://www.annualreviews.org/content/journals/10.1146/annurev.genet.39.073003.114725

Functional and evolutionary implications of gene orthology | https://www.nature.com/articles/nrg3456

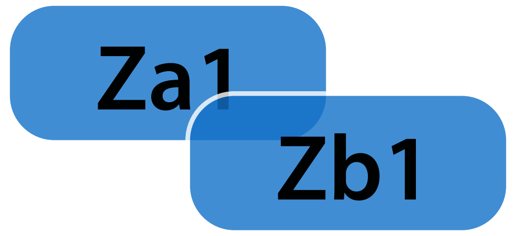
» Gene loss & duplication »



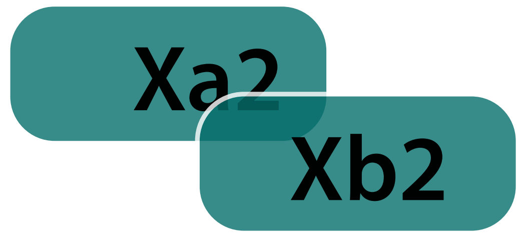
Orthologs, Paralogs, and Evolutionary Genomics | https://www.annualreviews.org/content/journals/10.1146/annurev.genet.39.073003.114725

Functional and evolutionary implications of gene orthology | https://www.nature.com/articles/nrg3456

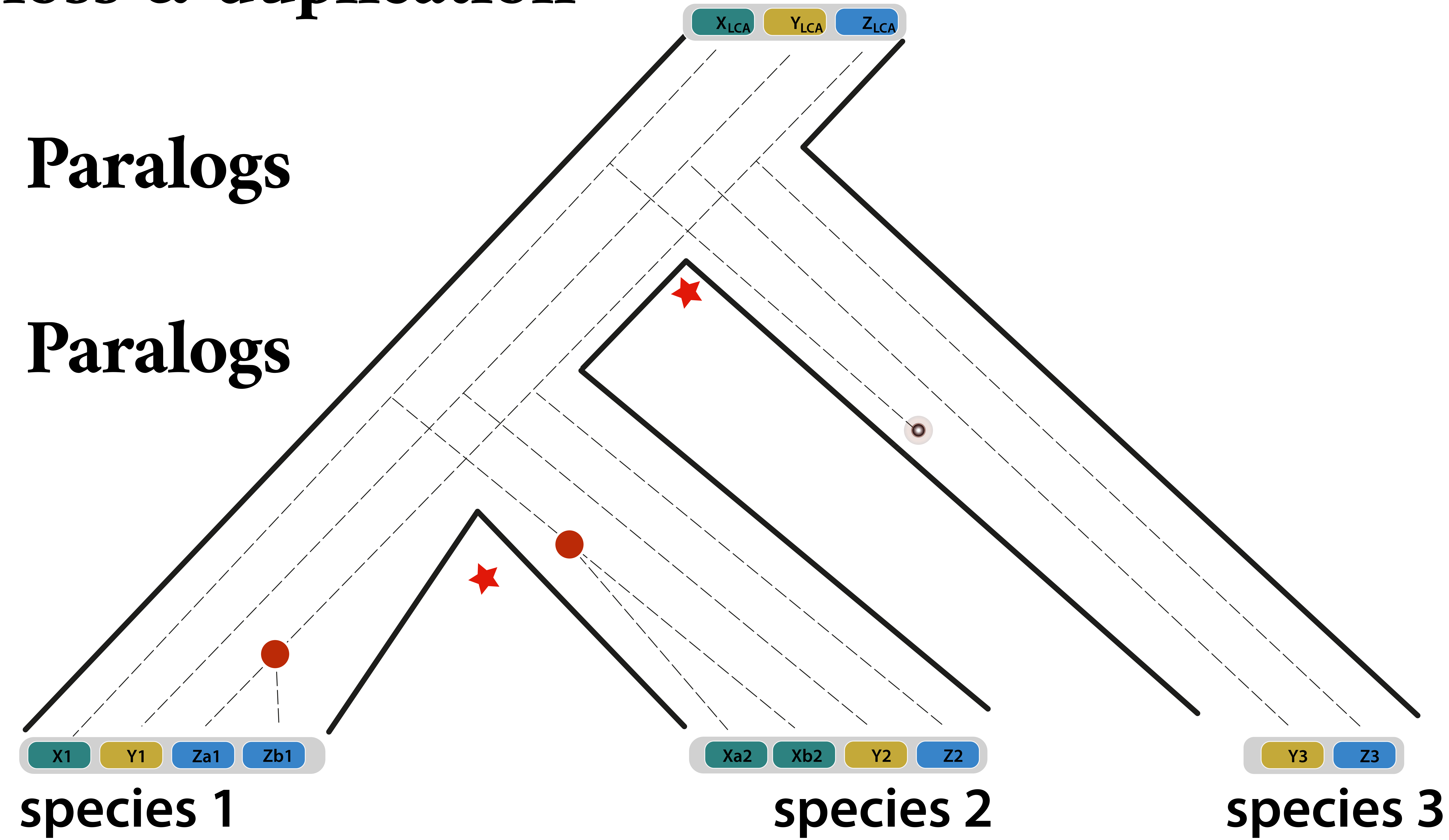
» Gene loss & duplication »



Paralogs

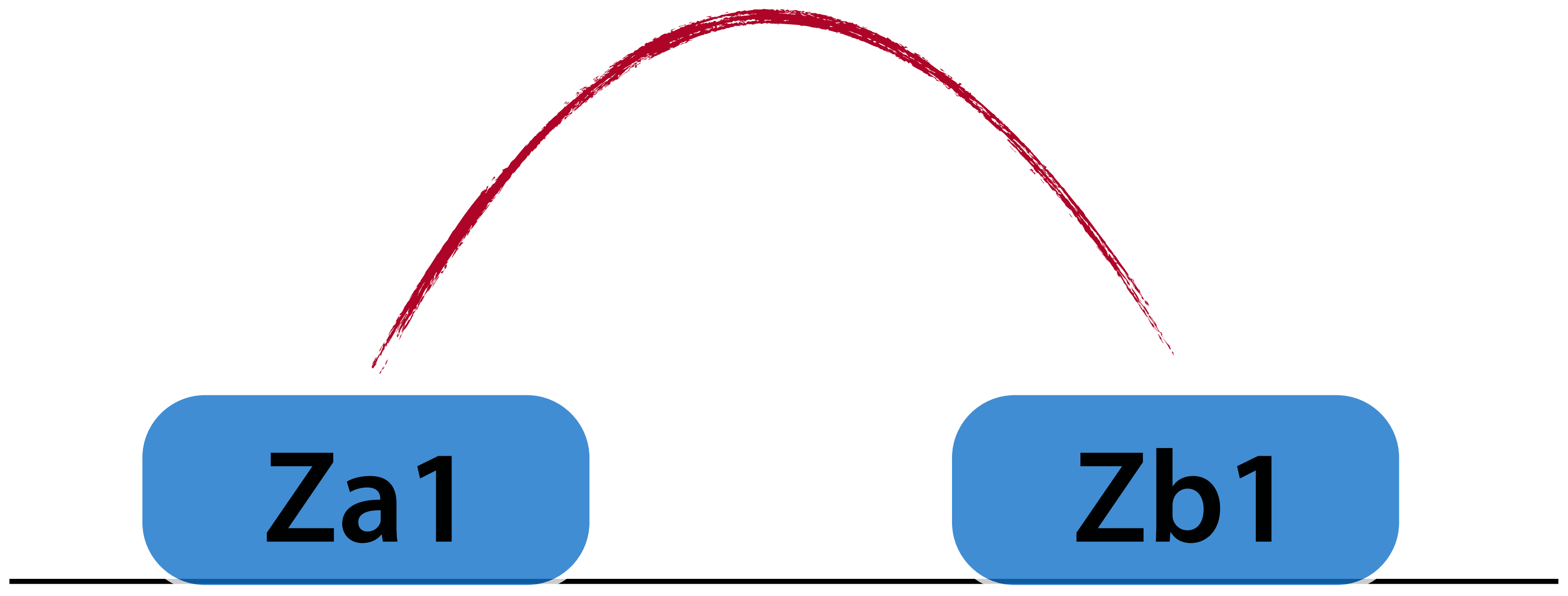


Paralogs



Orthologs, Paralogs, and Evolutionary Genomics | https://www.annualreviews.org/content/journals/10.1146/annurev.genet.39.073003.114725
Functional and evolutionary implications of gene orthology | https://www.nature.com/articles/nrg3456

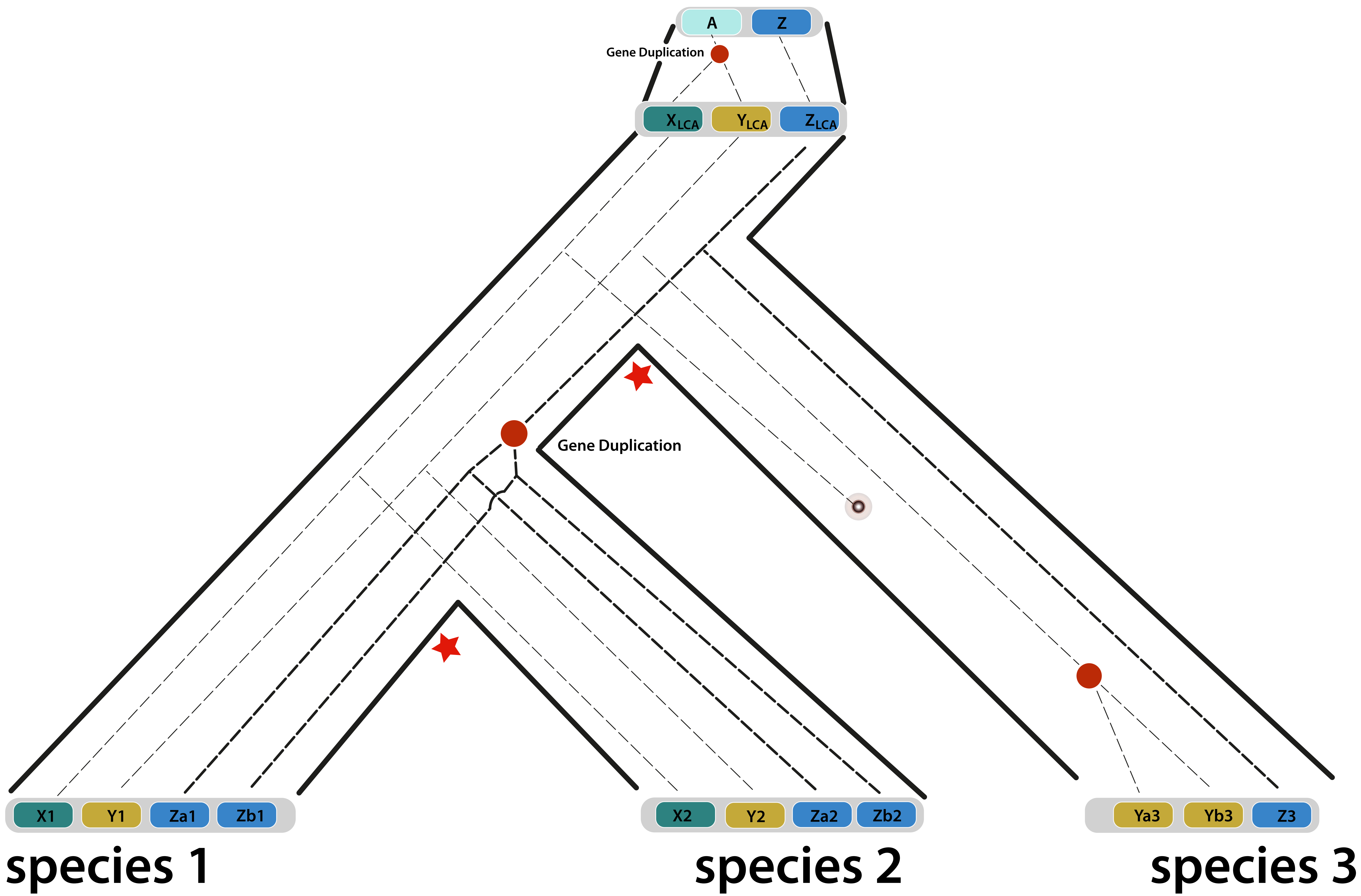
Paralogy: it describes a relationship that involve a duplication: *if a locus (gene) is generated by an event of tandem duplication*.



Paralogs

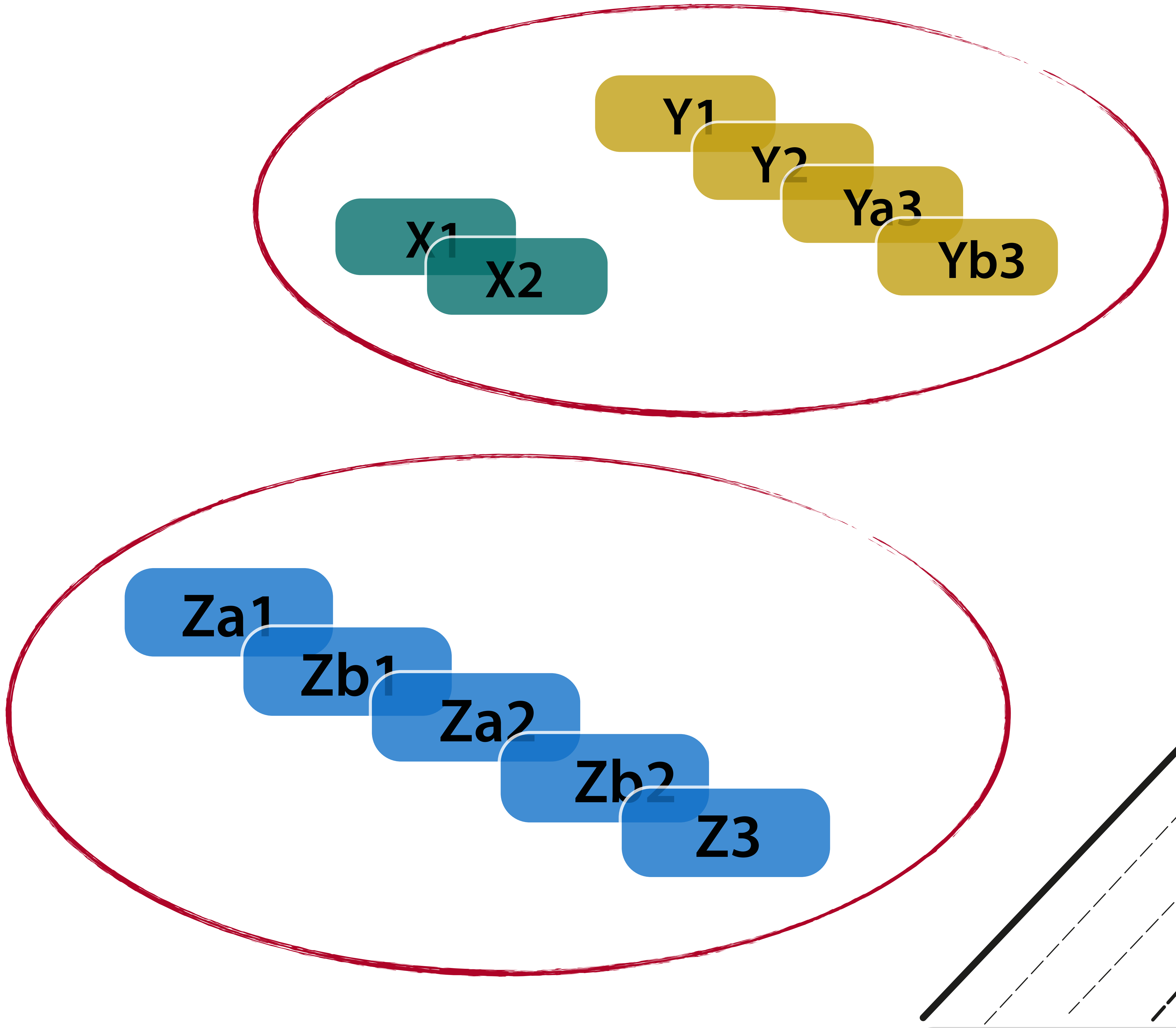
Orthologs, Paralogs, and Evolutionary Genomics | https://www.annualreviews.org/content/journals/10.1146/annurev.genet.39.073003.114725
Functional and evolutionary implications of gene orthology | https://www.nature.com/articles/nrg3456

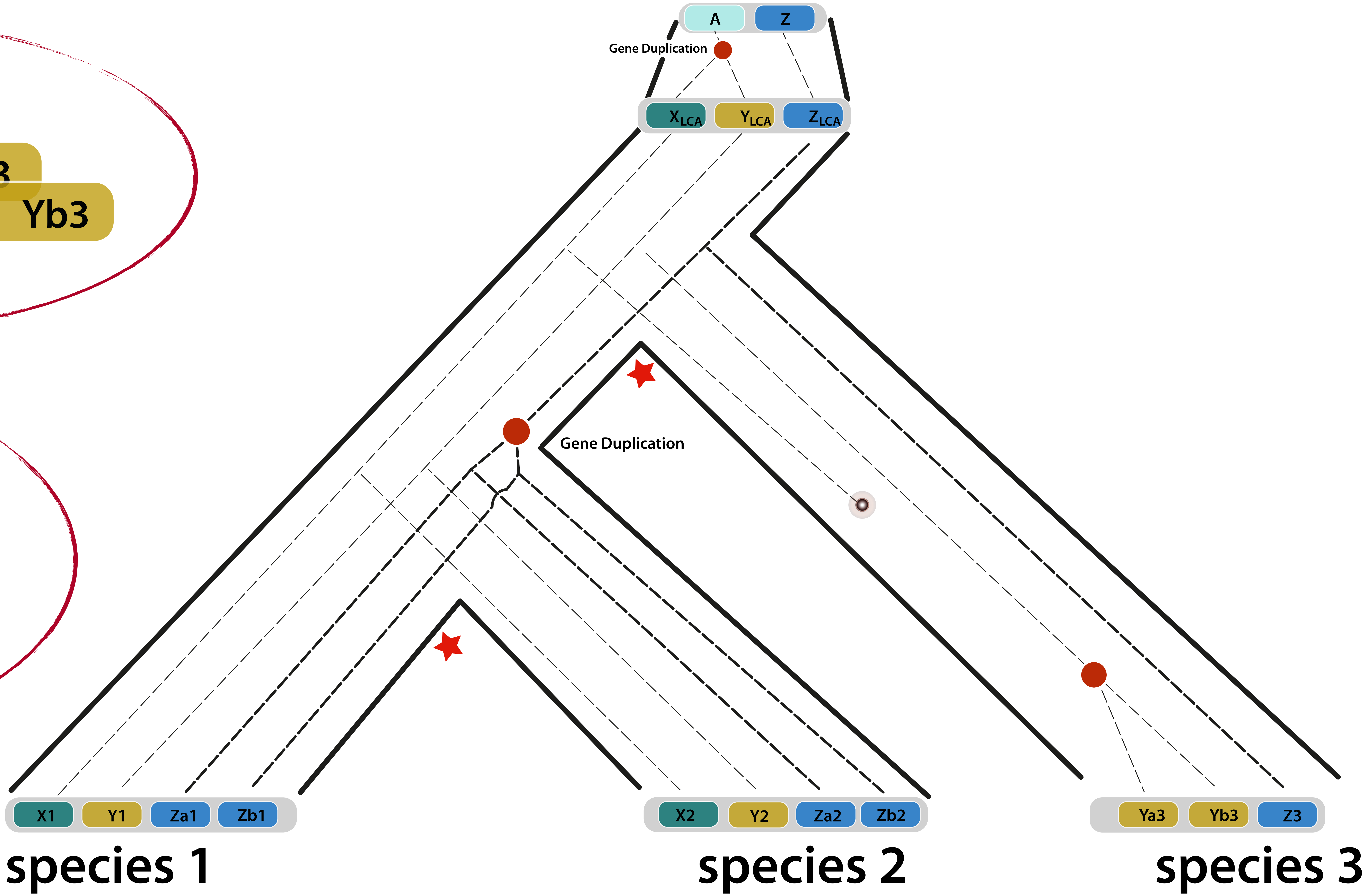
» Homologs ? »



Orthologs, Paralogs, and Evolutionary Genomics | https://www.annualreviews.org/content/journals/10.1146/annurev.genet.39.073003.114725
Functional and evolutionary implications of gene orthology | https://www.nature.com/articles/nrg3456

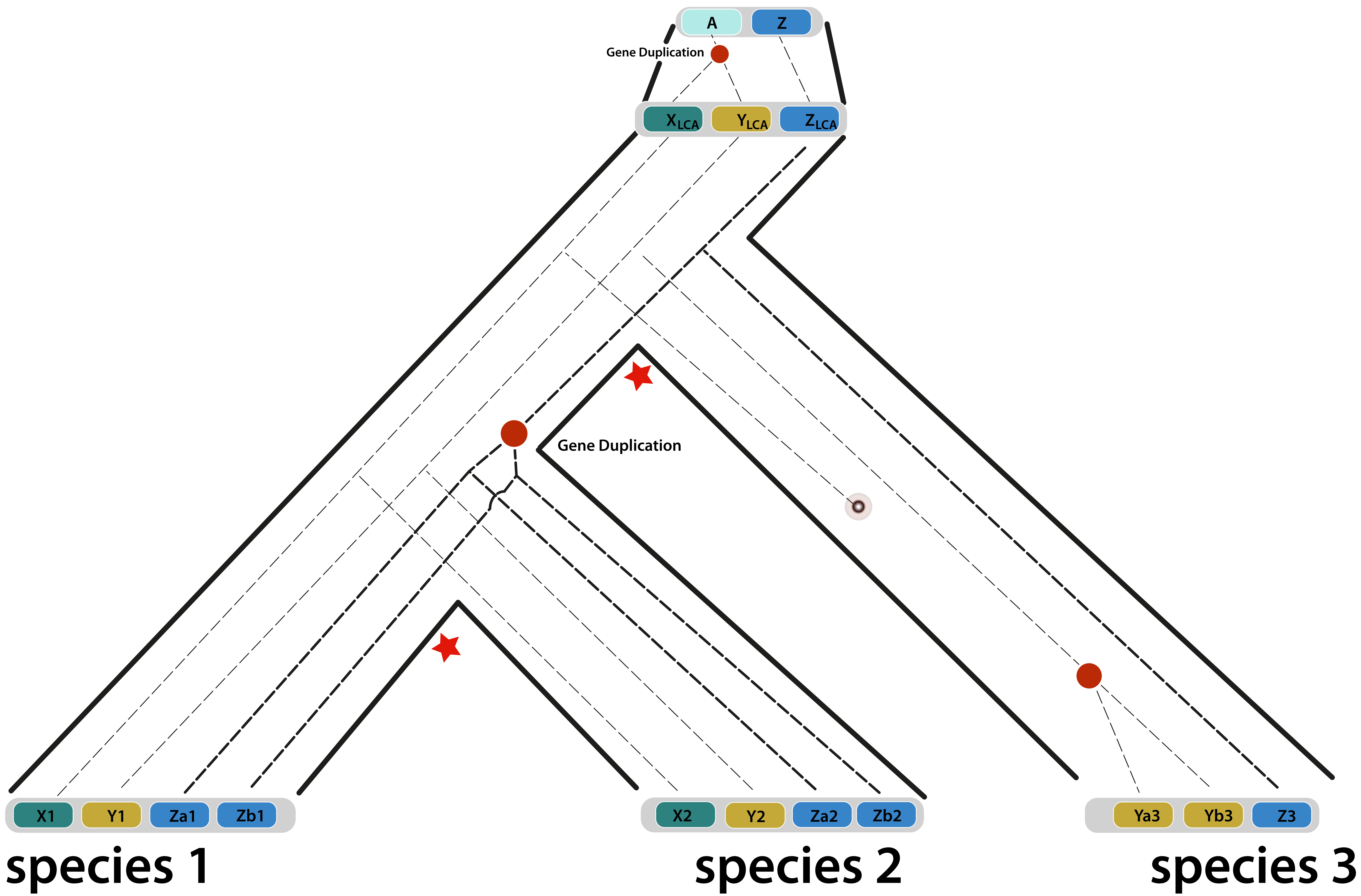
» Homologs ? »





Orthologs, Paralogs, and Evolutionary Genomics | https://www.annualreviews.org/content/journals/10.1146/annurev.genet.39.073003.114725
Functional and evolutionary implications of gene orthology | https://www.nature.com/articles/nrg3456

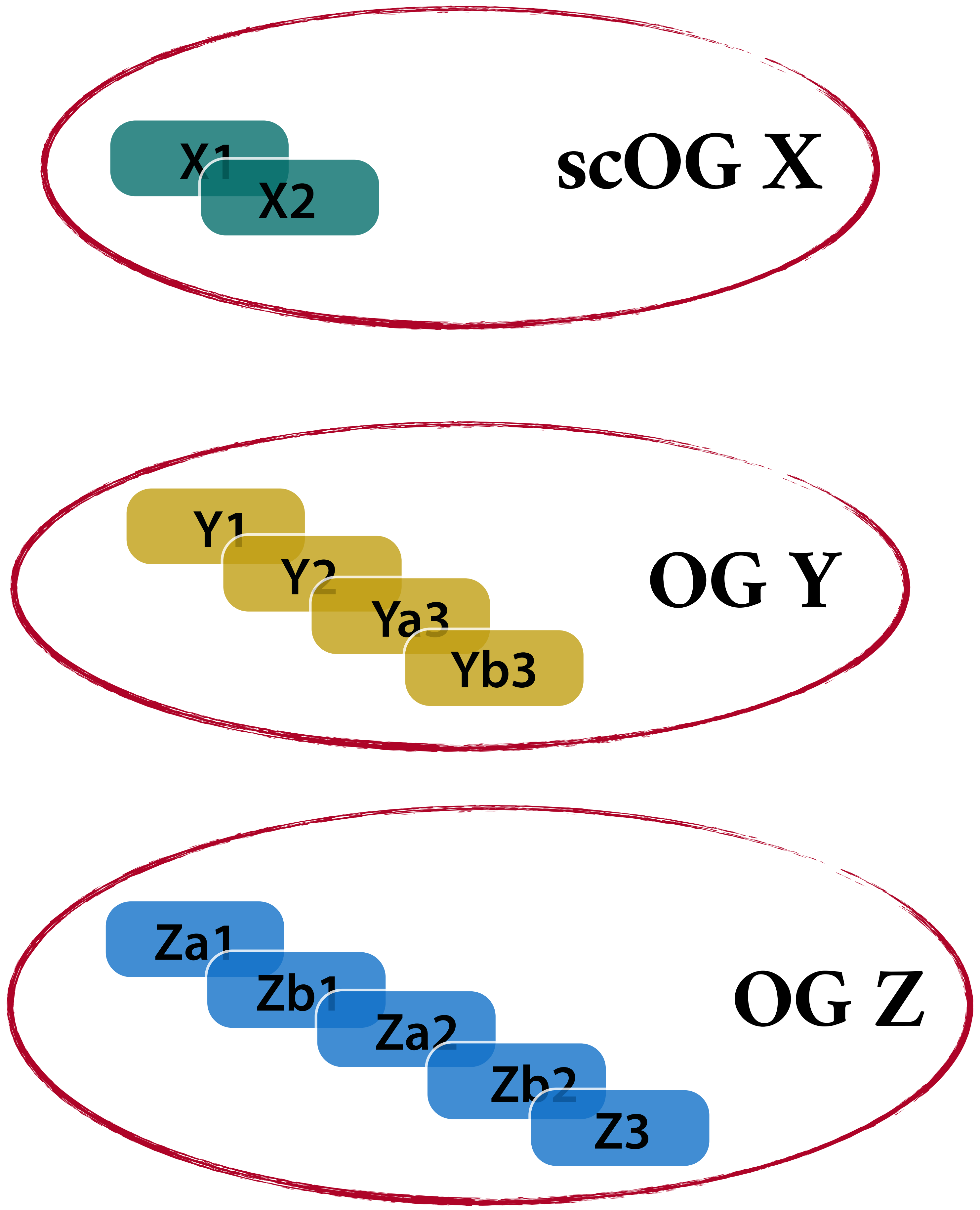
» Orthologs ? »

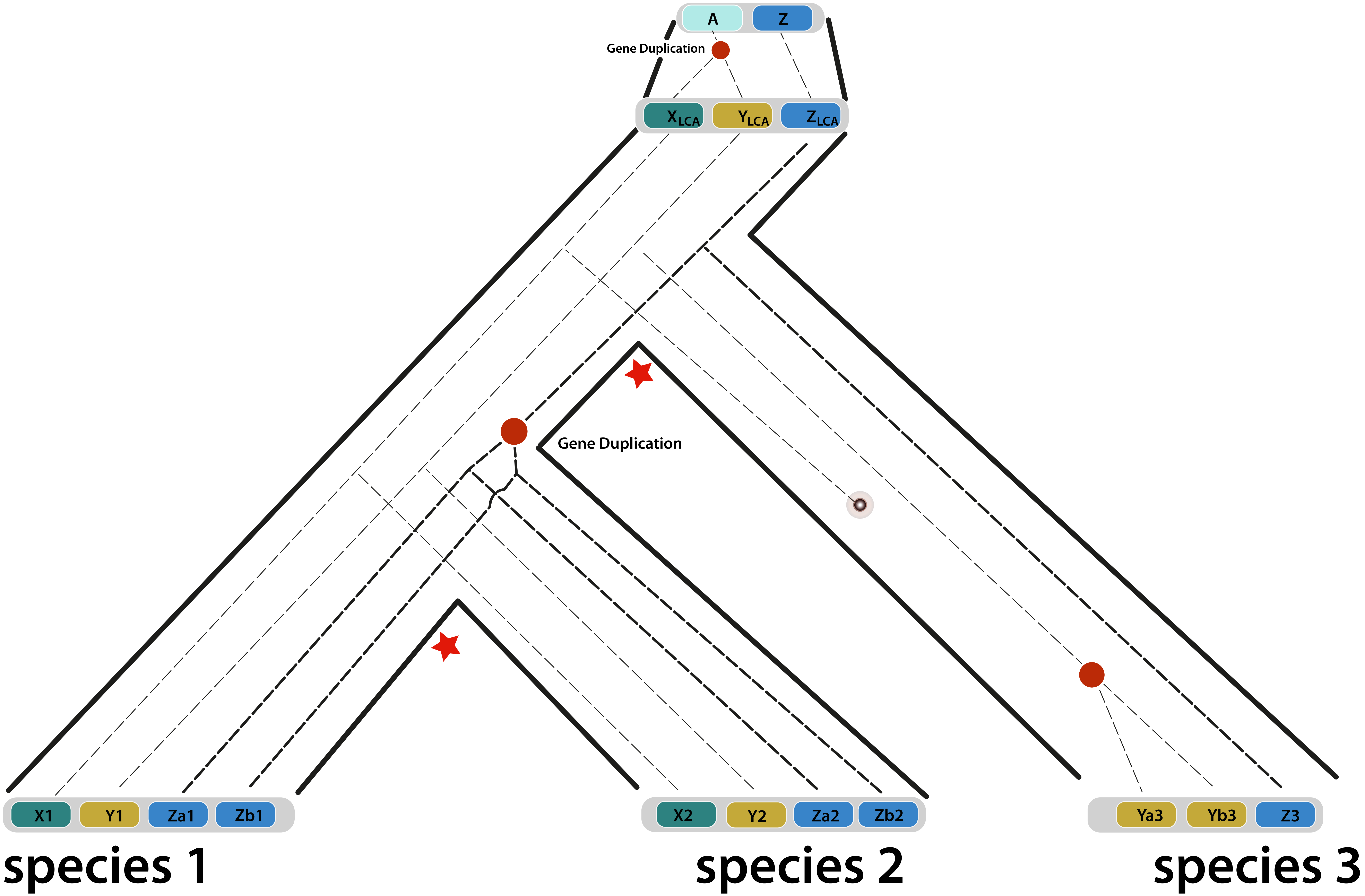


Orthologs, Paralogs, and Evolutionary Genomics | https://www.annualreviews.org/content/journals/10.1146/annurev.genet.39.073003.114725

Functional and evolutionary implications of gene orthology | https://www.nature.com/articles/nrg3456

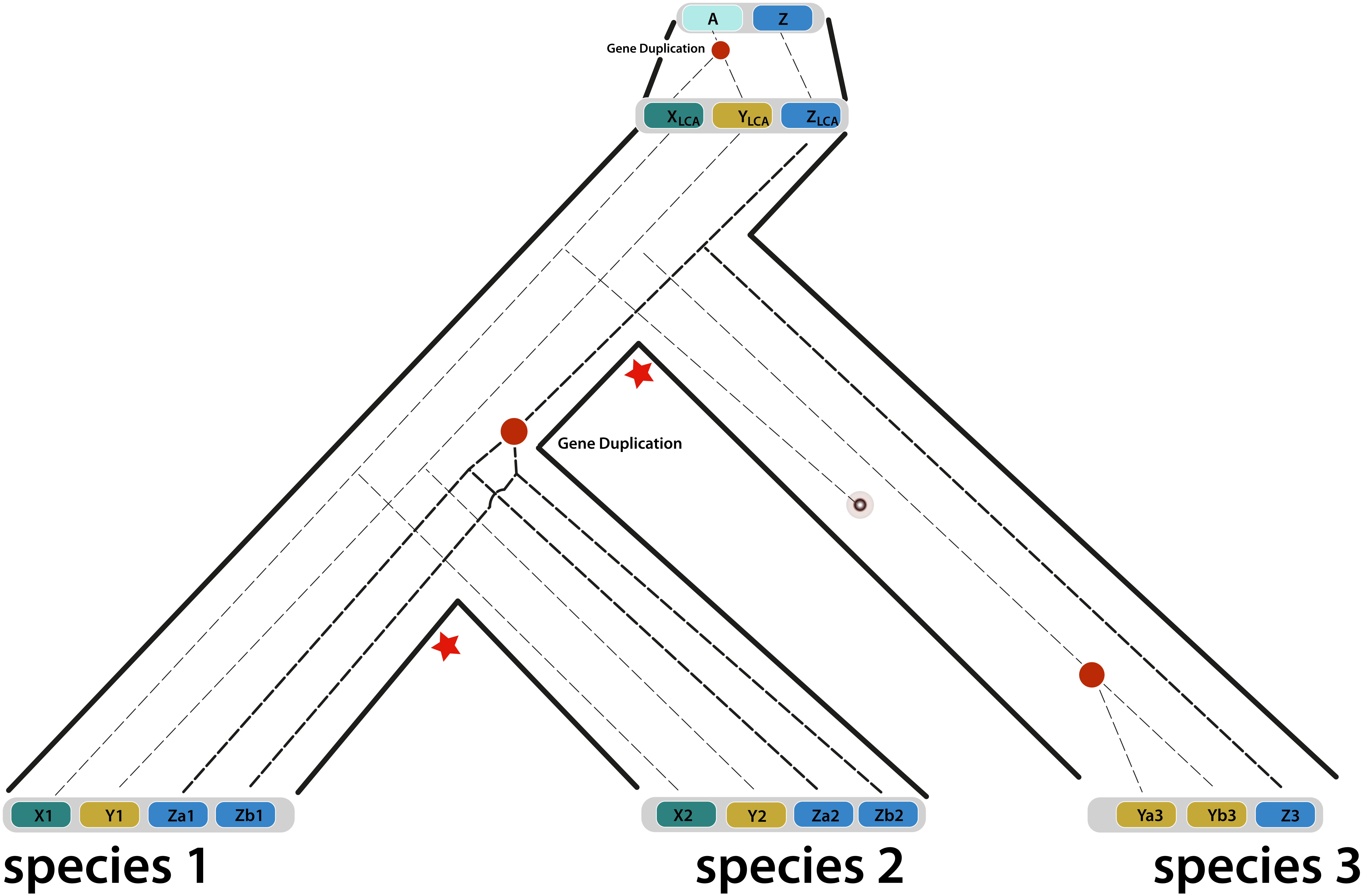
» Orthologs ? »





Orthologs, Paralogs, and Evolutionary Genomics | https://www.annualreviews.org/content/journals/10.1146/annurev.genet.39.073003.114725
Functional and evolutionary implications of gene orthology | https://www.nature.com/articles/nrg3456

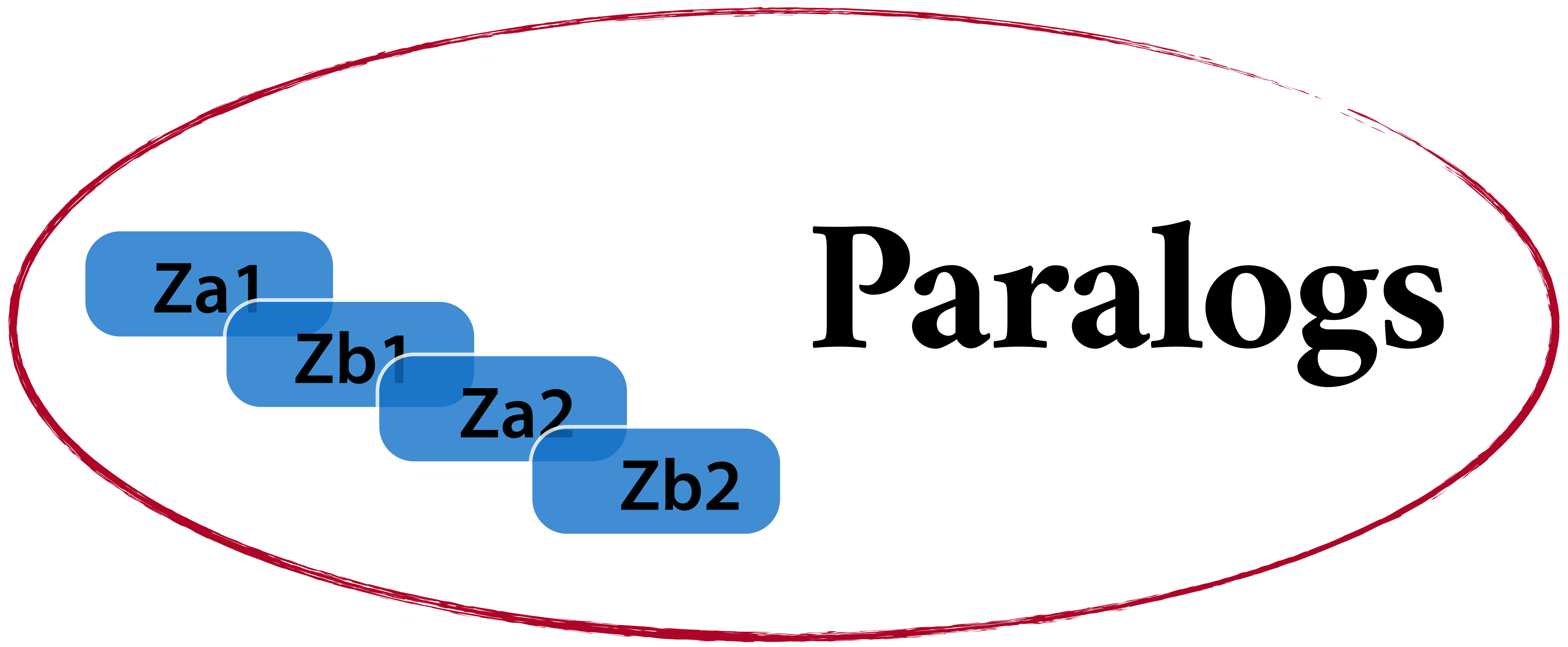
» Paralogs ? »



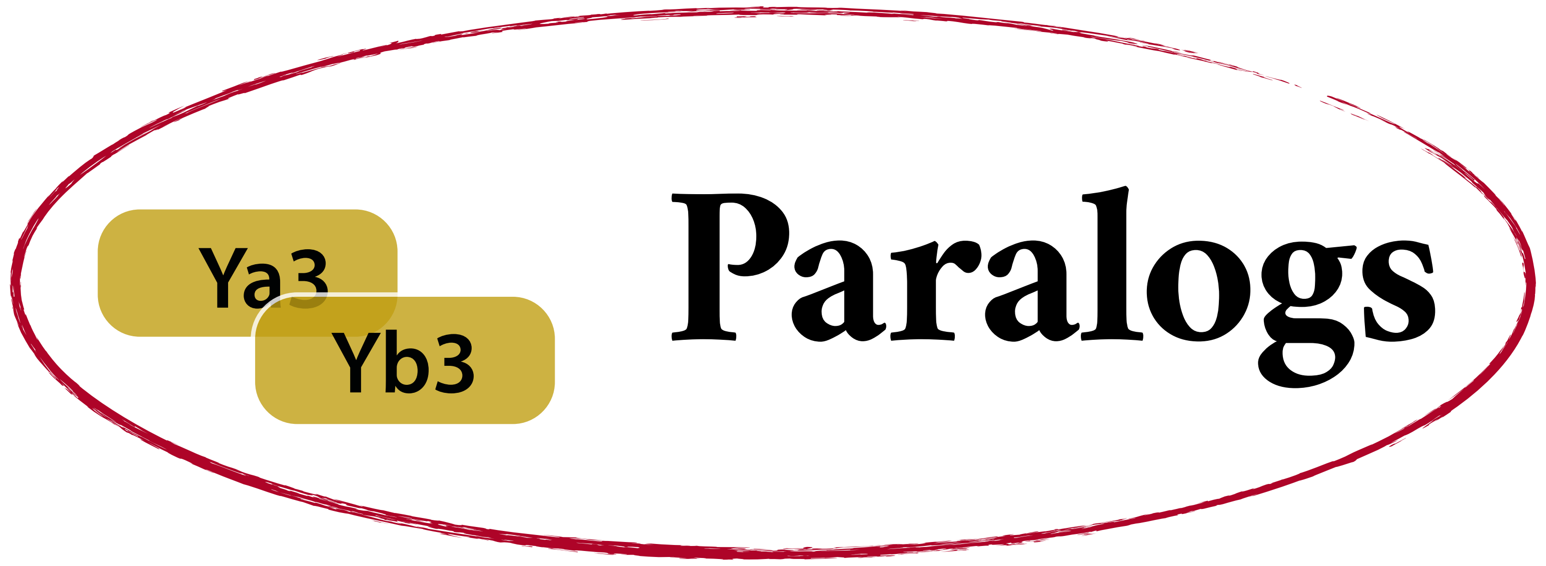
Orthologs, Paralogs, and Evolutionary Genomics | *https://www.annualreviews.org/content/journals/10.1146/annurev.genet.39.073003.114725*

Functional and evolutionary implications of gene orthology | *https://www.nature.com/articles/nrg3456*

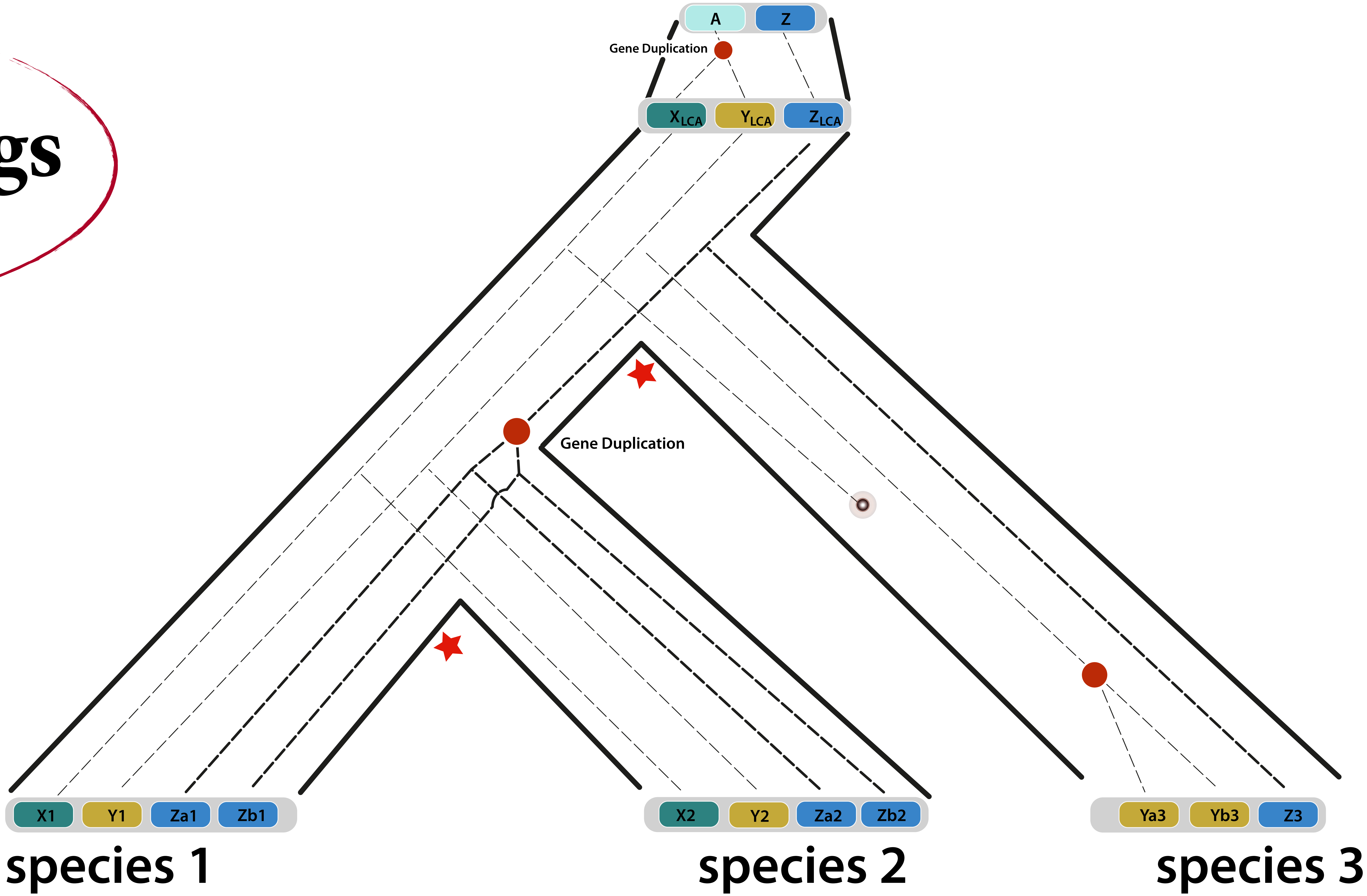
» Paralogs ? »



Paralogs



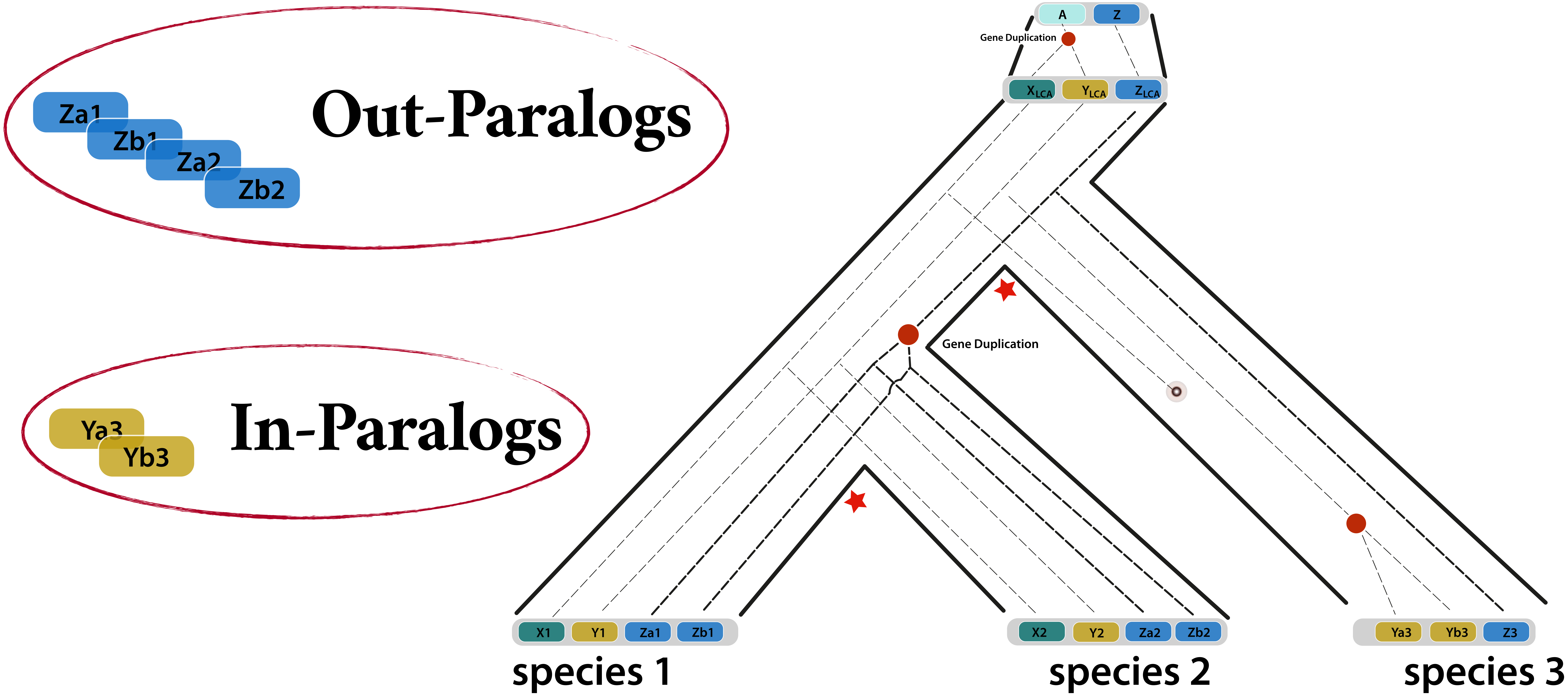
Paralogs



Orthologs, Paralogs, and Evolutionary Genomics | https://www.annualreviews.org/content/journals/10.1146/annurev.genet.39.073003.114725

Functional and evolutionary implications of gene orthology | https://www.nature.com/articles/nrg3456

» Paralogs ? »

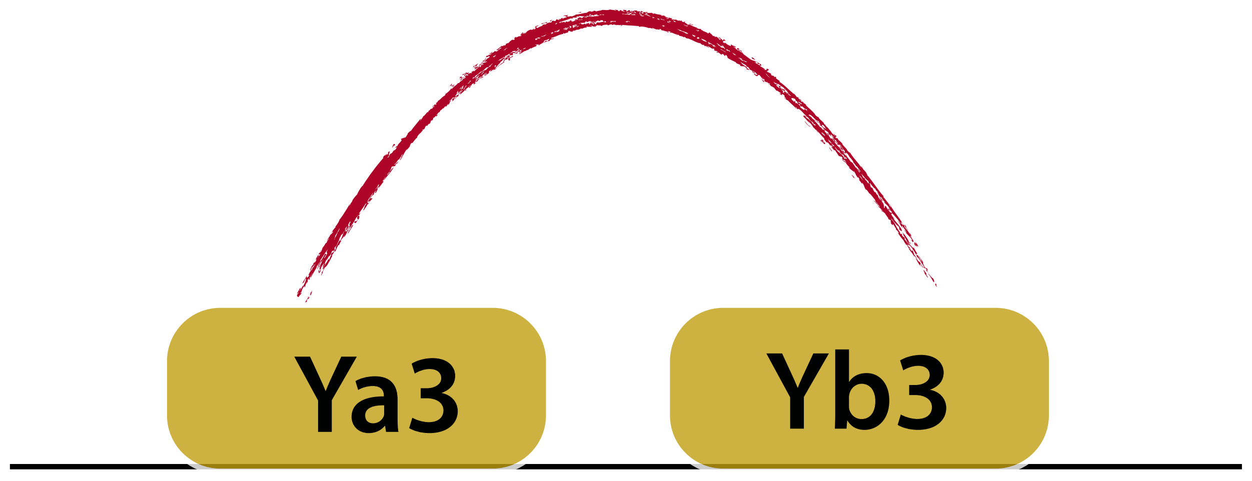


Out-Paralogs

In-Paralogs

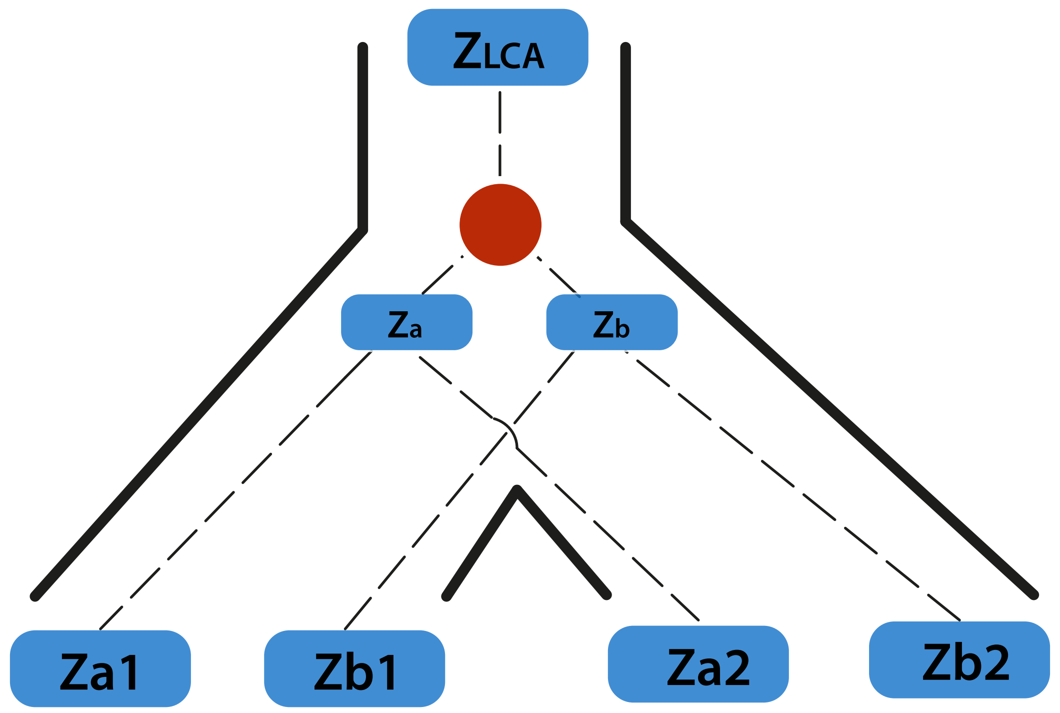
Orthologs, Paralogs, and Evolutionary Genomics | https://www.annualreviews.org/content/journals/10.1146/annurev.genet.39.073003.114725
Functional and evolutionary implications of gene orthology | https://www.nature.com/articles/nrg3456

IN-Paralogy: it describes a relationship that involve a duplication occurred <u>within a species</u>



IN-Paralogs

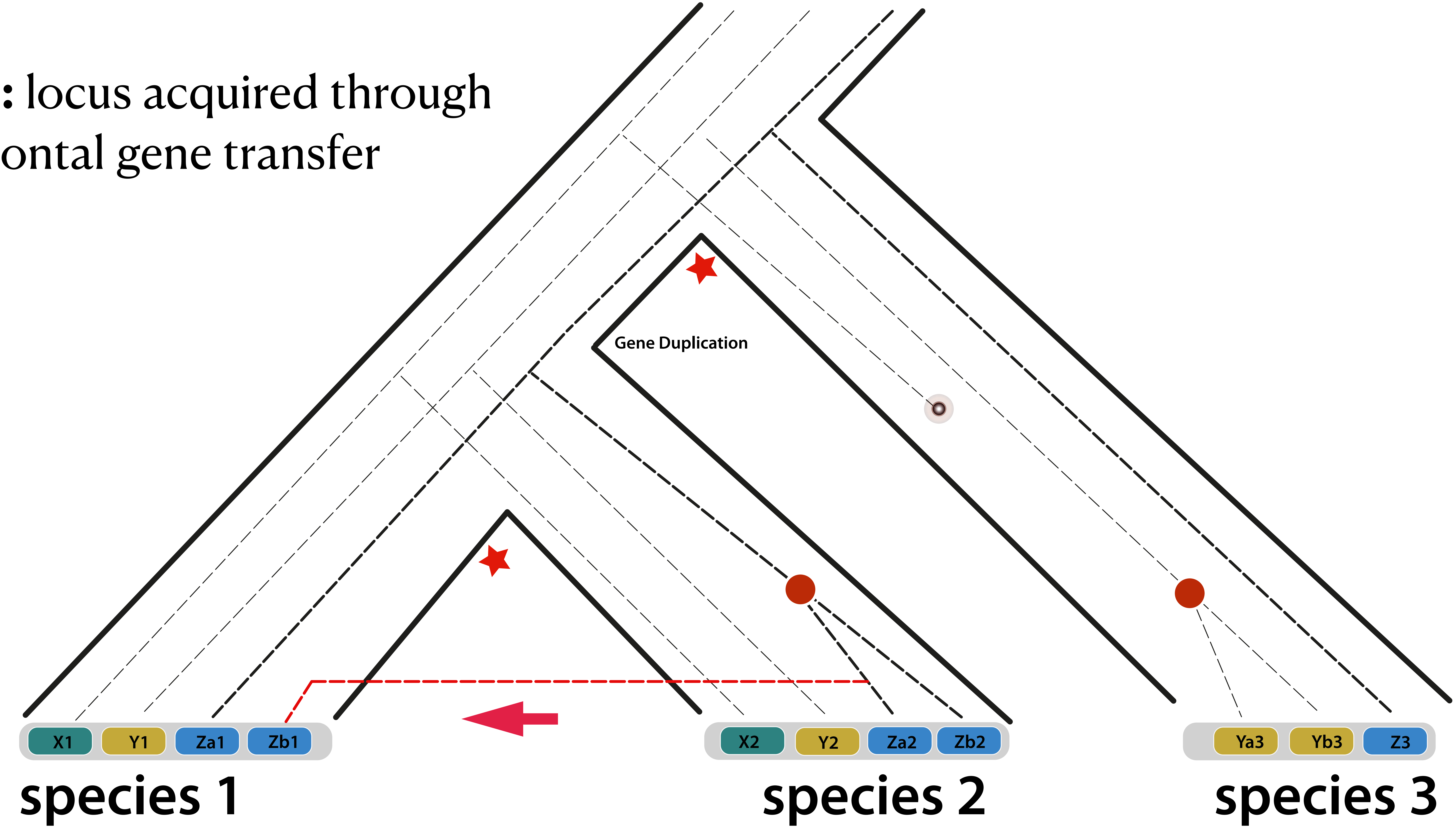
OUT-Paralogy: it describes a relationship that involve a duplication occurred <u>in one of the ancestor</u>



Orthologs, Paralogs, and Evolutionary Genomics | *https://www.annualreviews.org/content/journals/10.1146/annurev.genet.39.073003.114725*
Functional and evolutionary implications of gene orthology | *https://www.nature.com/articles/nrg3456*

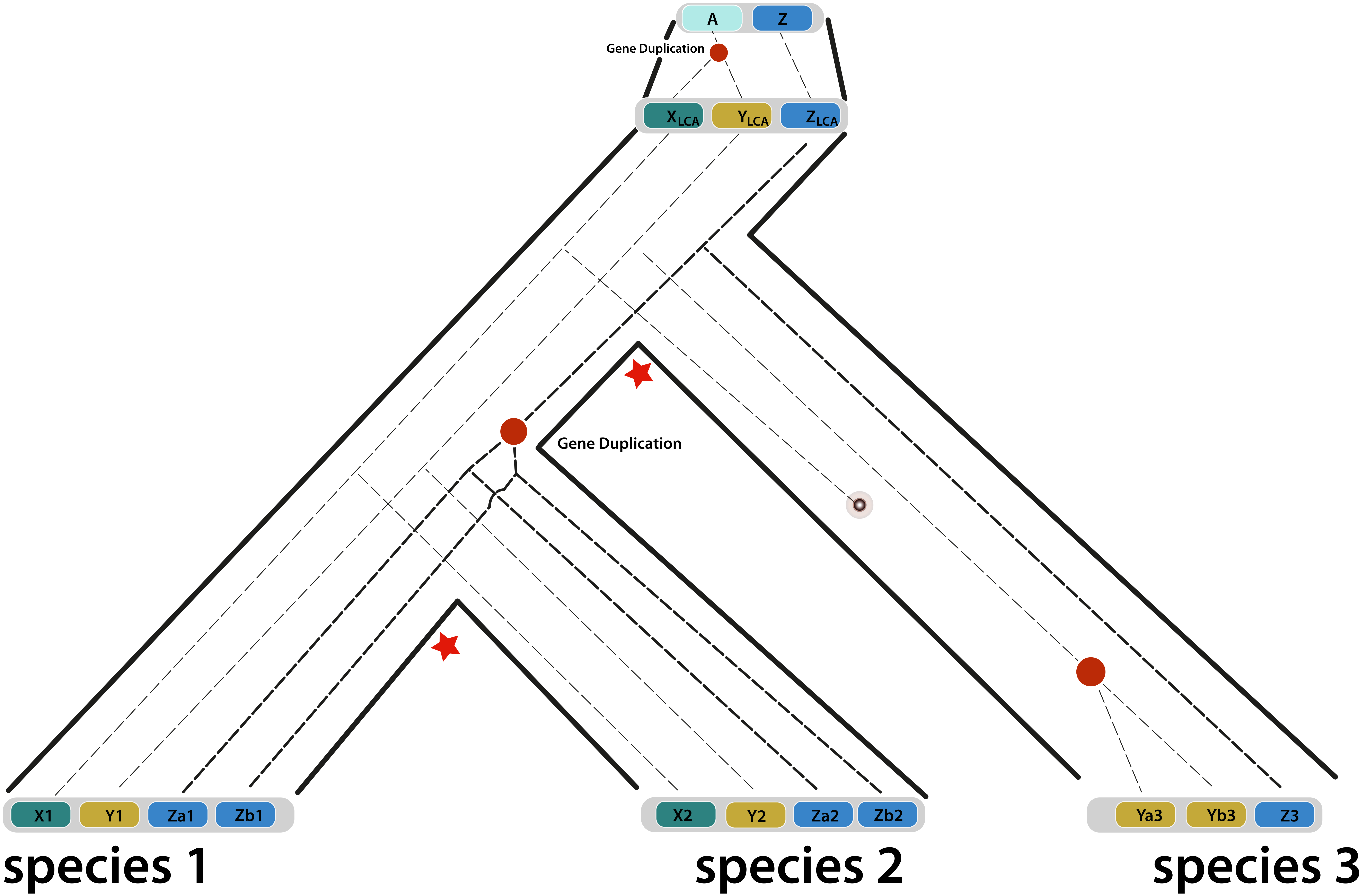
» Horizontal Gene Transfer (HGT) »

Xenologs: locus acquired through horizontal gene transfer

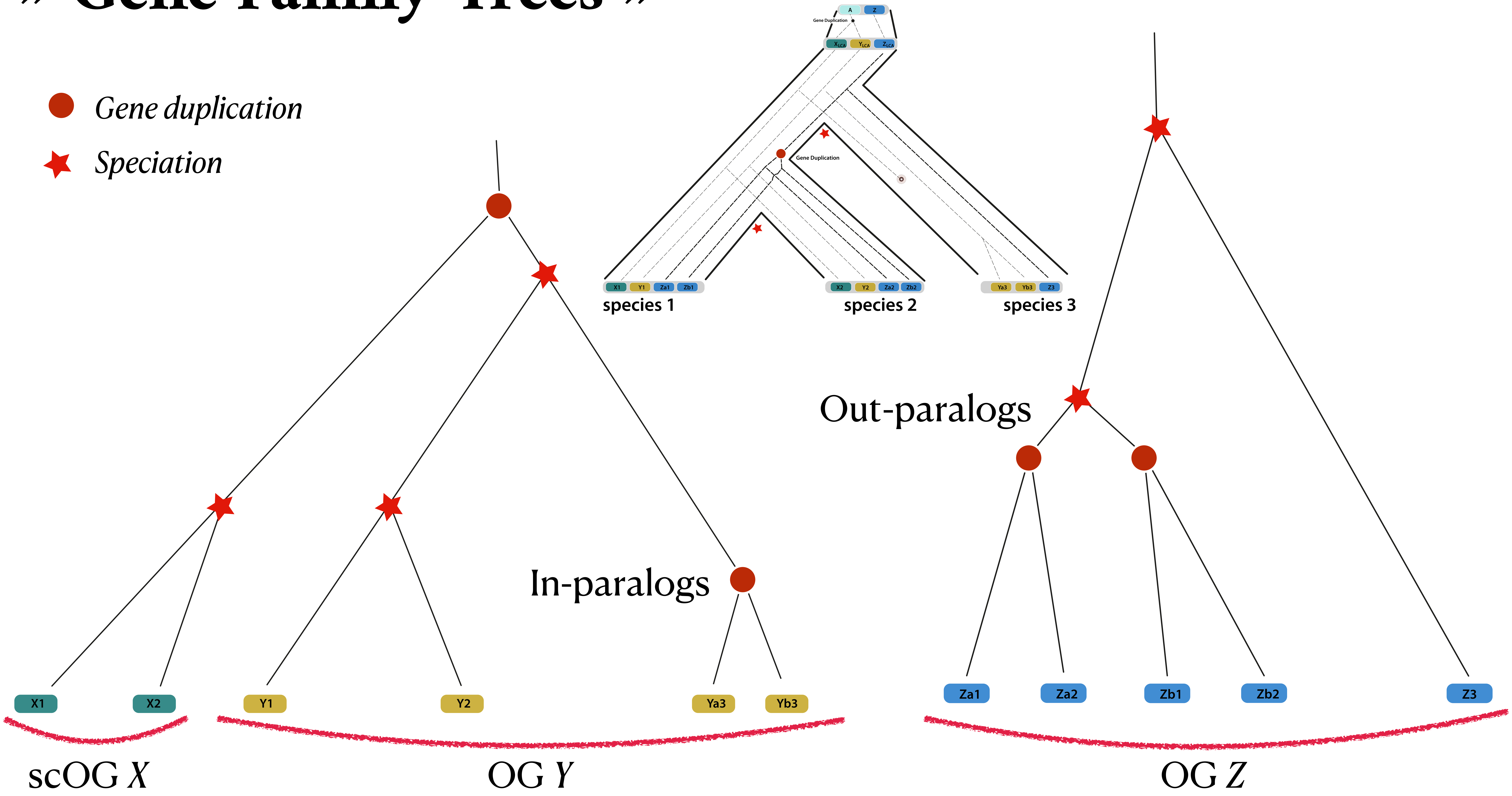


Orthologs, Paralogs, and Evolutionary Genomics | https://www.annualreviews.org/content/journals/10.1146/annurev.genet.39.073003.114725
Functional and evolutionary implications of gene orthology | https://www.nature.com/articles/nrg3456

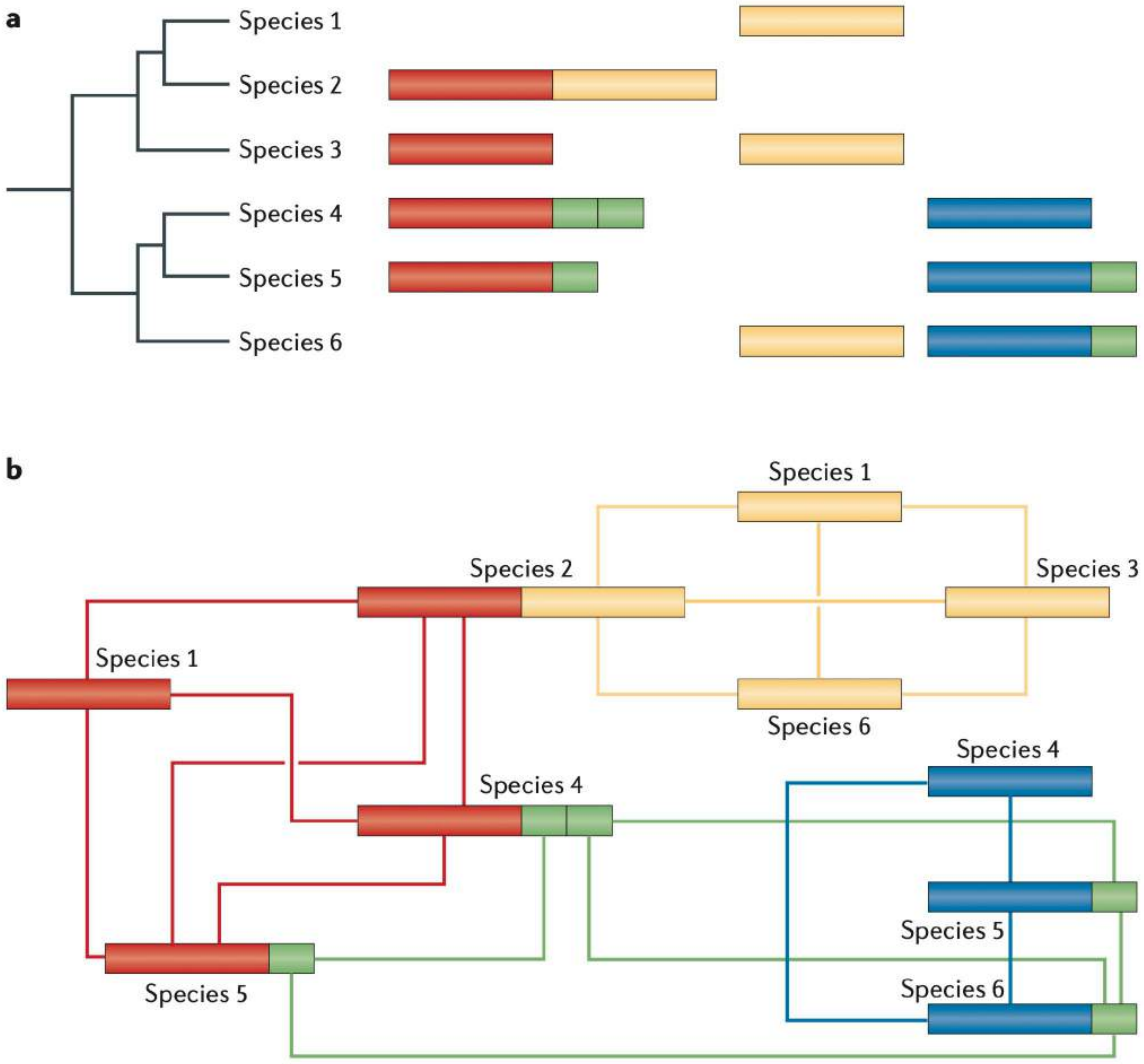
» Gene Family Tree »



» Gene Family Trees »



» Units of orthology »

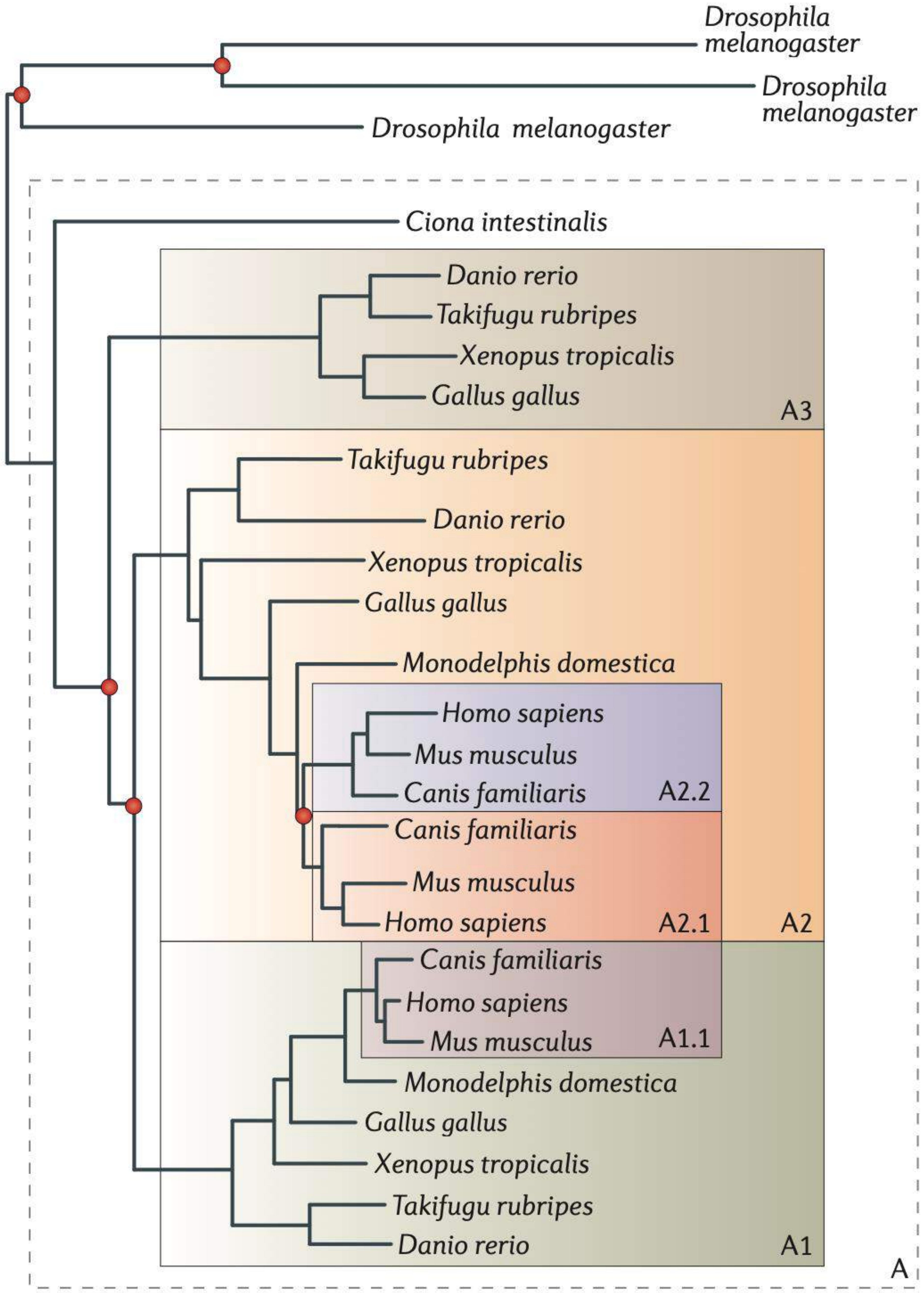


Orthologs, Paralogs, and Evolutionary Genomics | https://www.annualreviews.org/content/journals/10.1146/annurev.genet.39.073003.114725

Functional and evolutionary implications of gene orthology | https://www.nature.com/articles/nrg3456

» Real examples »

Transferrin family



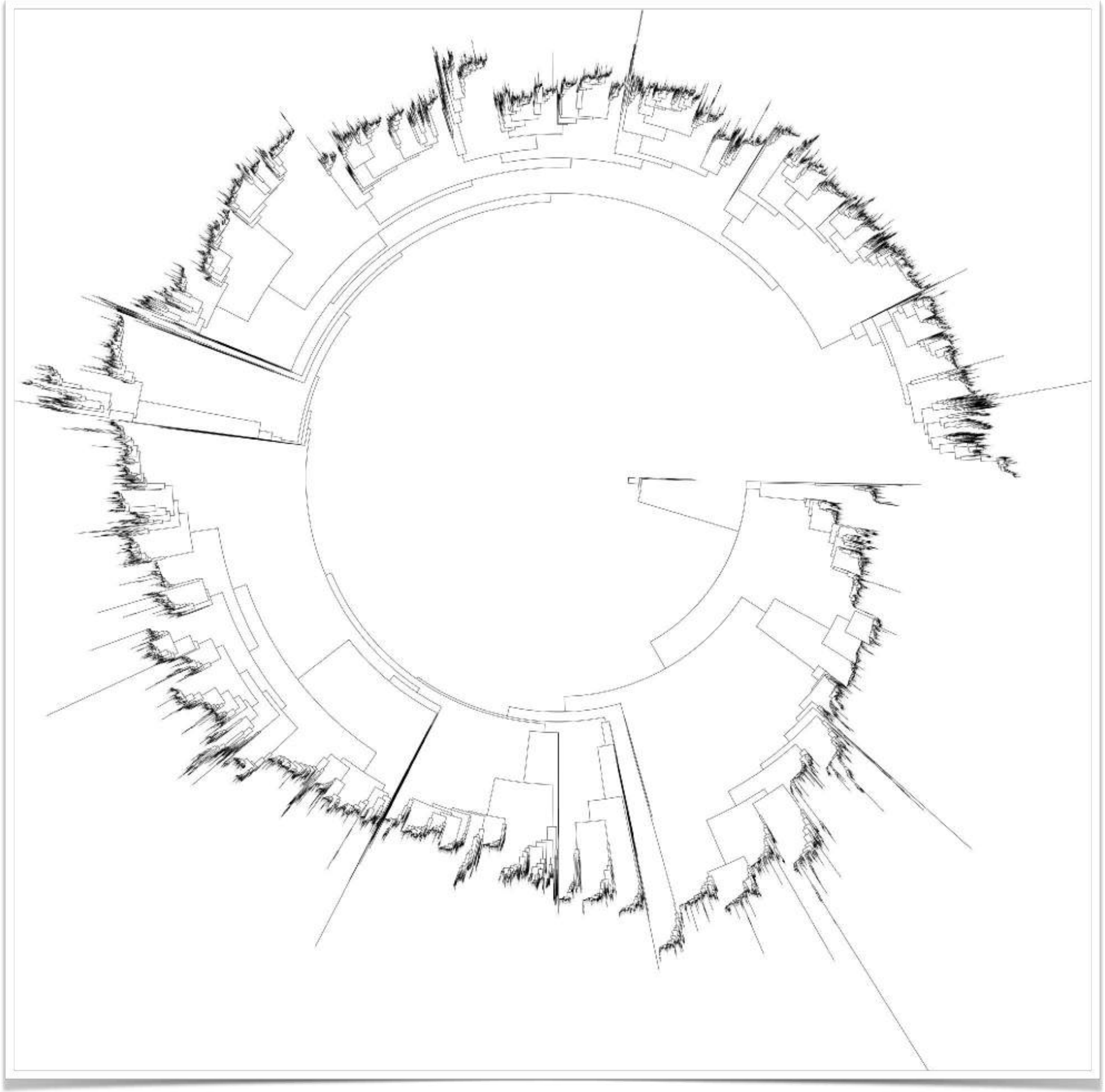
Orthologs, Paralogs, and Evolutionary Genomics | https://www.annualreviews.org/content/journals/10.1146/annurev.genet.39.073003.114725
Functional and evolutionary implications of gene orthology | https://www.nature.com/articles/nrg3456

» Real examples »

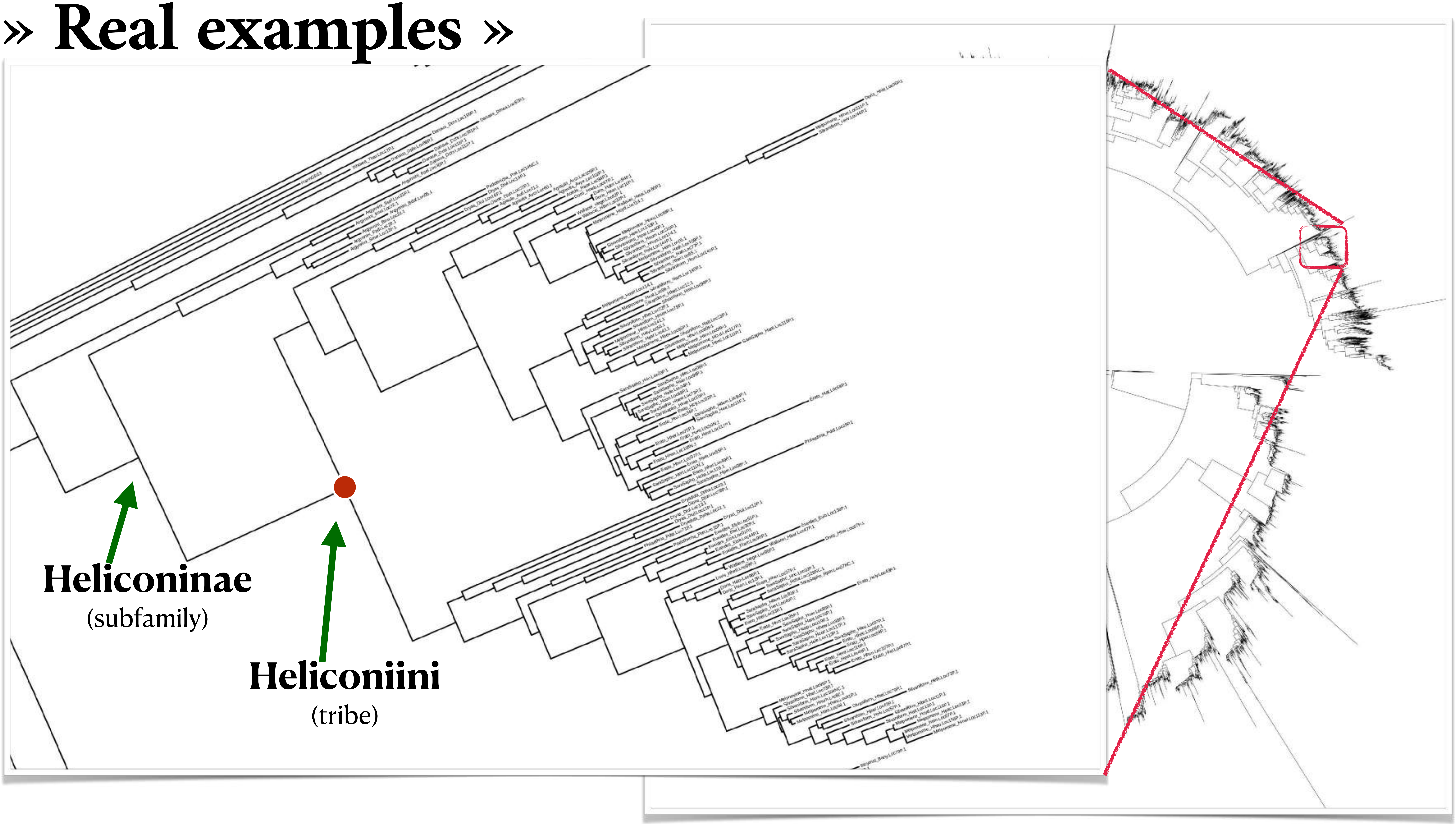
Gustatory receptors

***8,320** tips*

***16,638** nodes*



» Real examples »

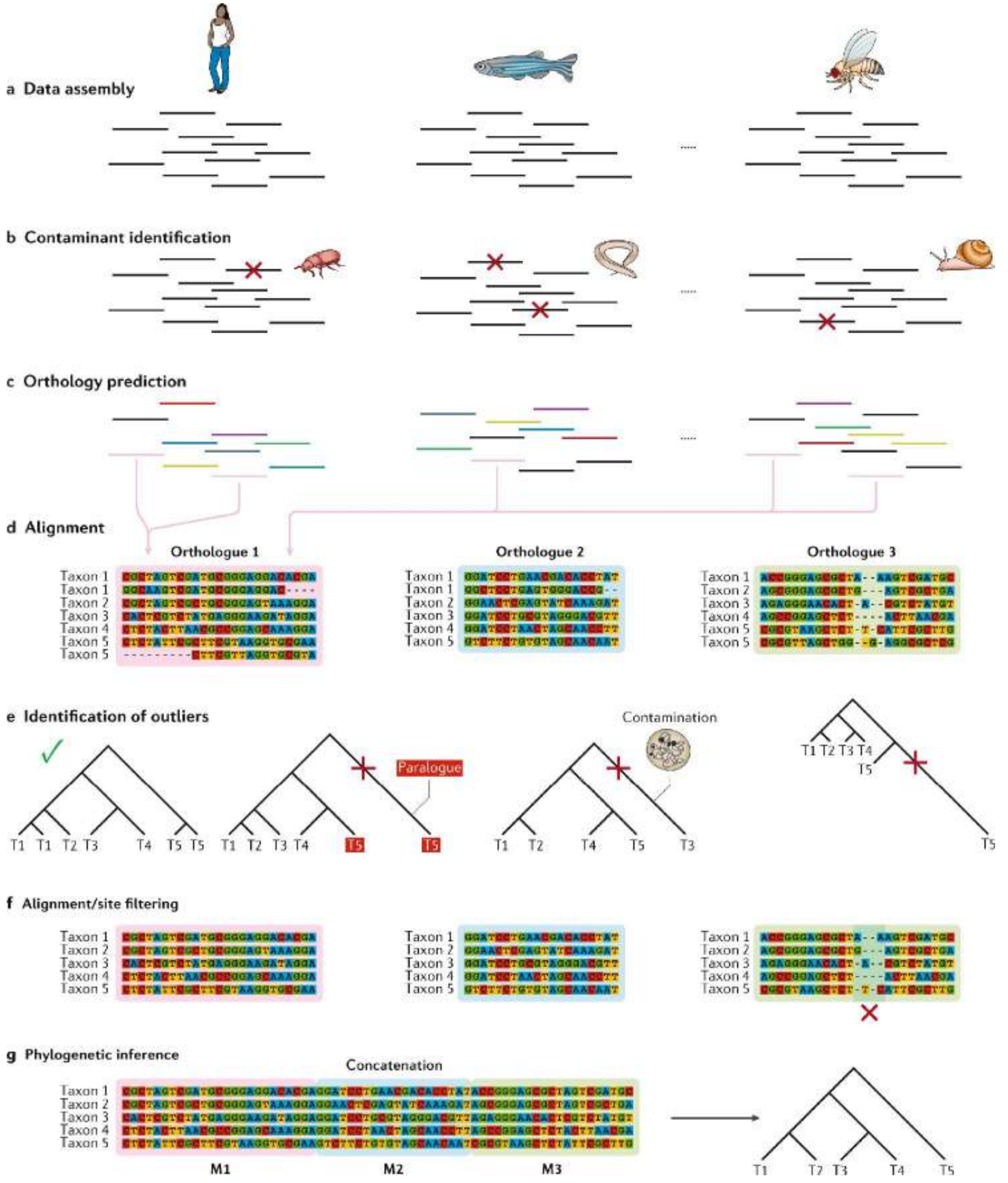


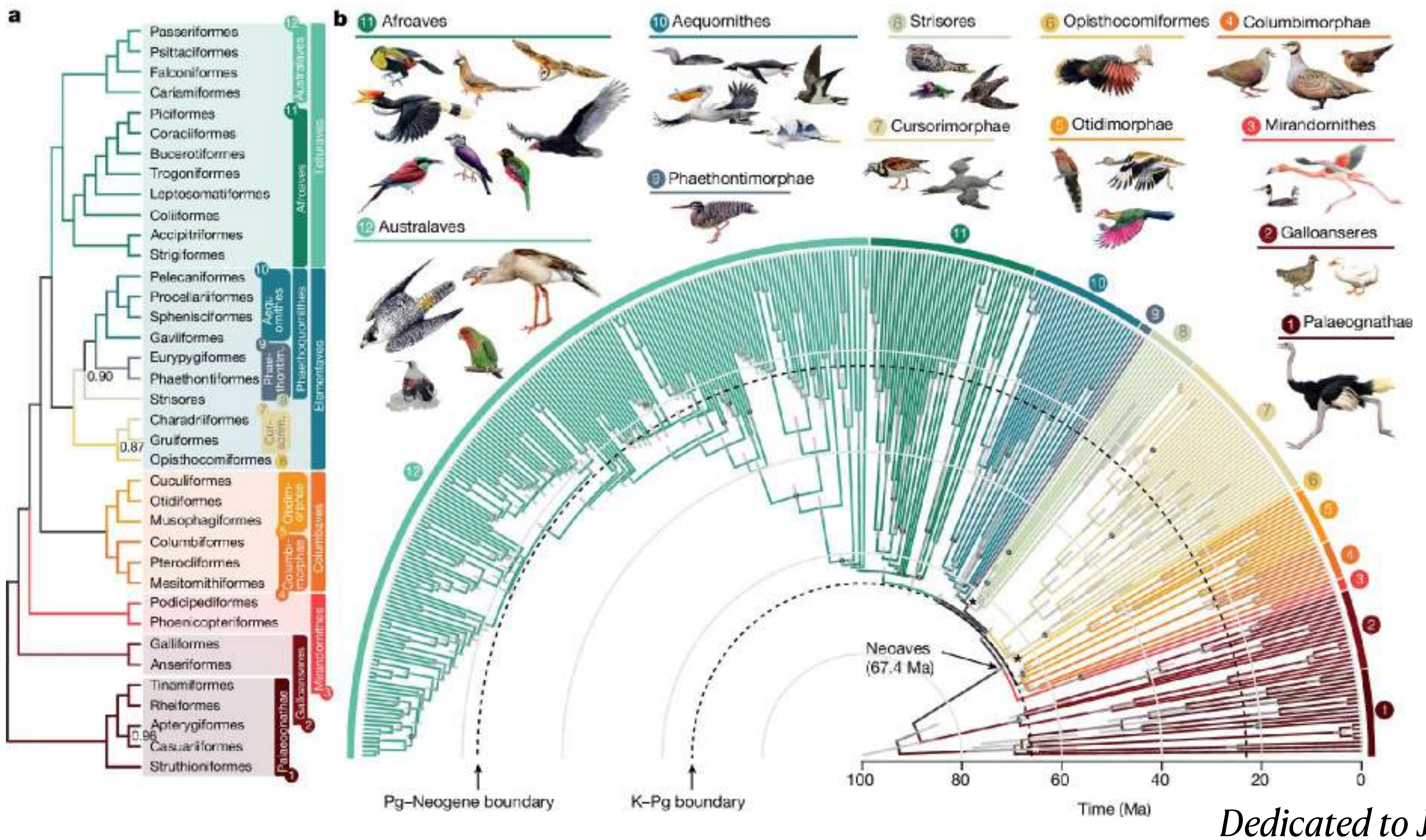
Heliconinae
(subfamily)

Heliconiini
(tribe)

» Why this is relevant ? »

» Species Tree Estimation »



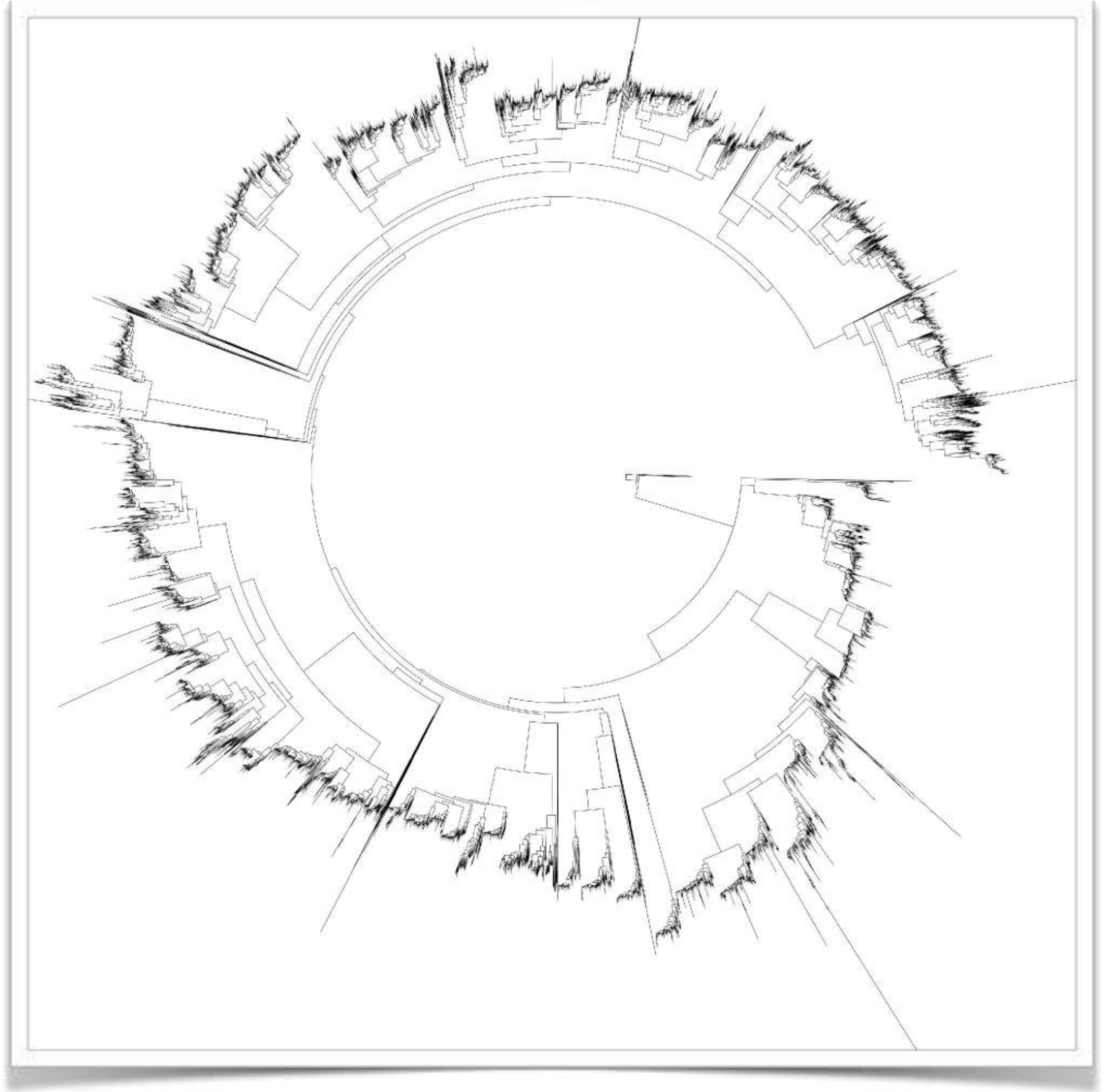


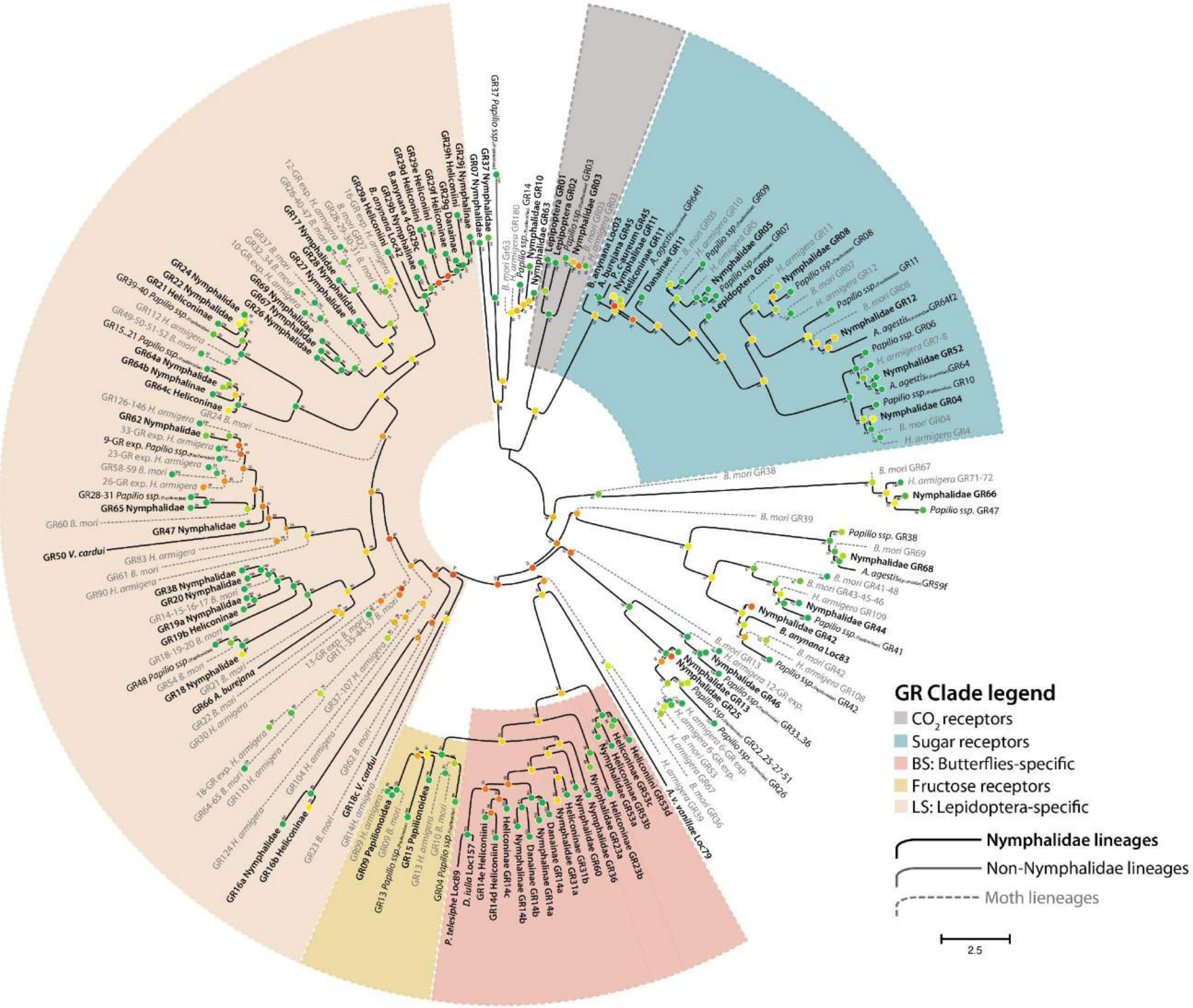
https://www.nature.com/articles/s41586-024-07323-1

Dedicated to Joan

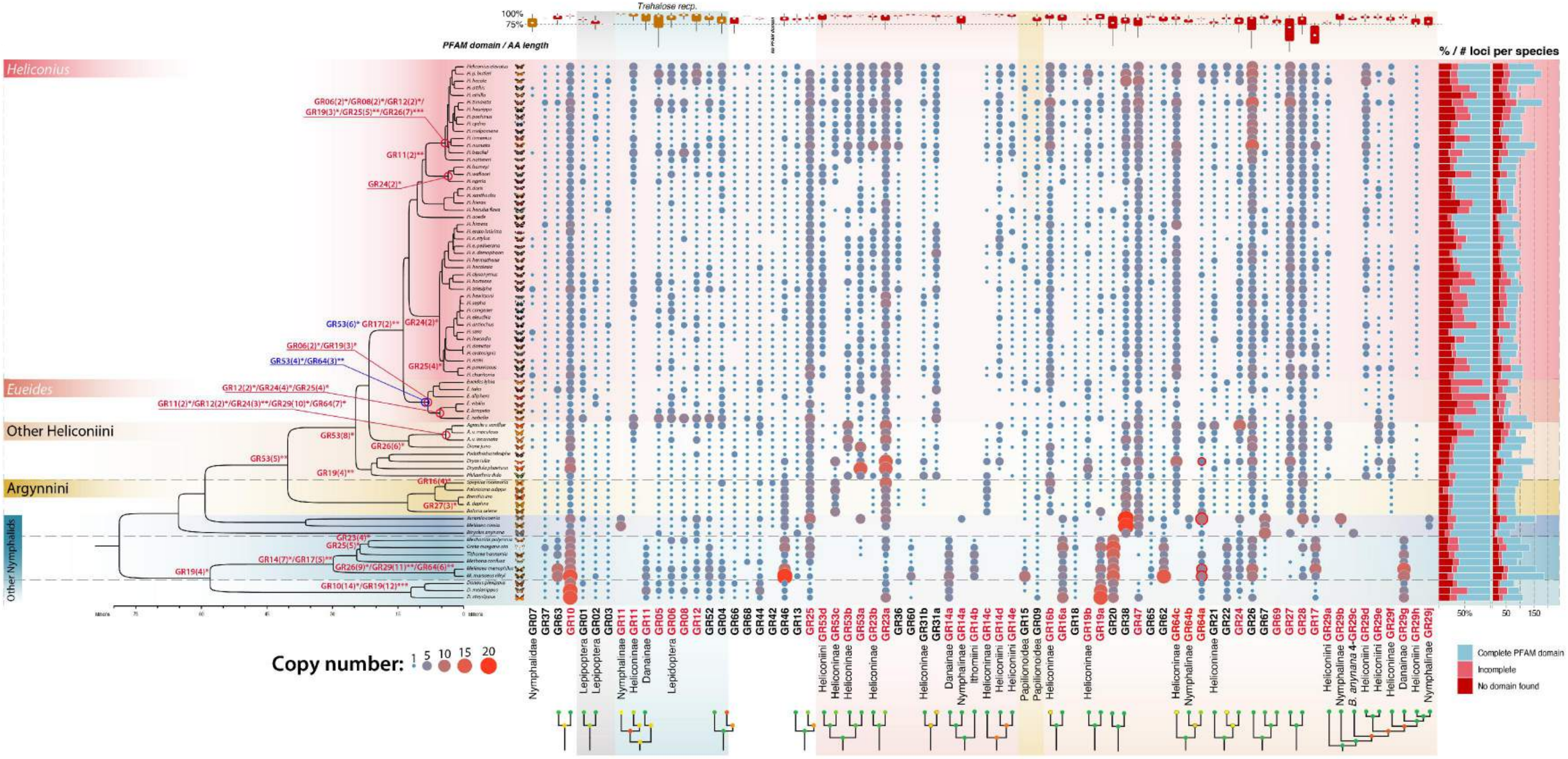
Phylogenetic tree building in the genomic age | https://www.nature.com/articles/s41576-020-0233-0

» Gene family reconstruction | expansions/contractions »



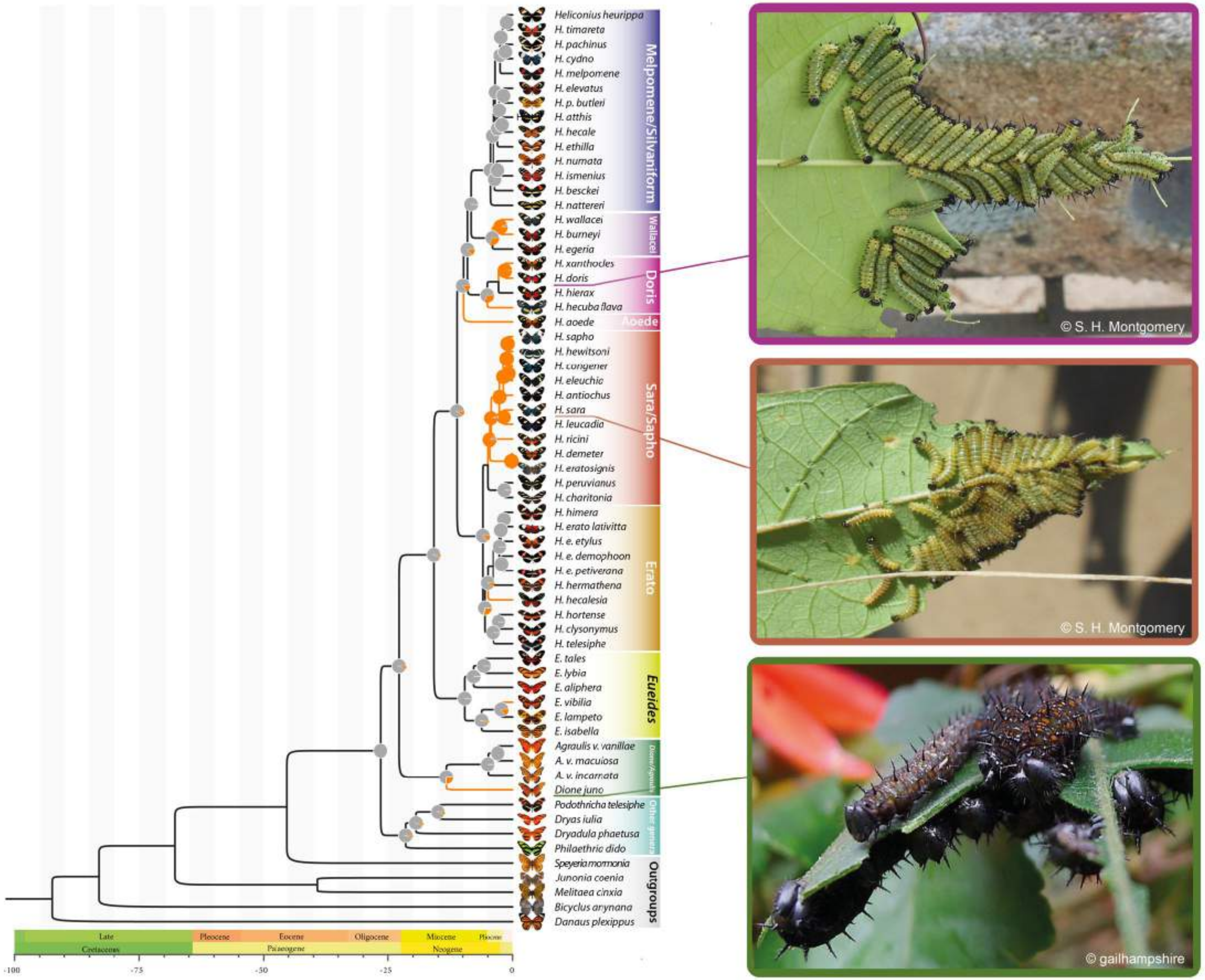


» Gene family reconstruction | expansions/contractions »

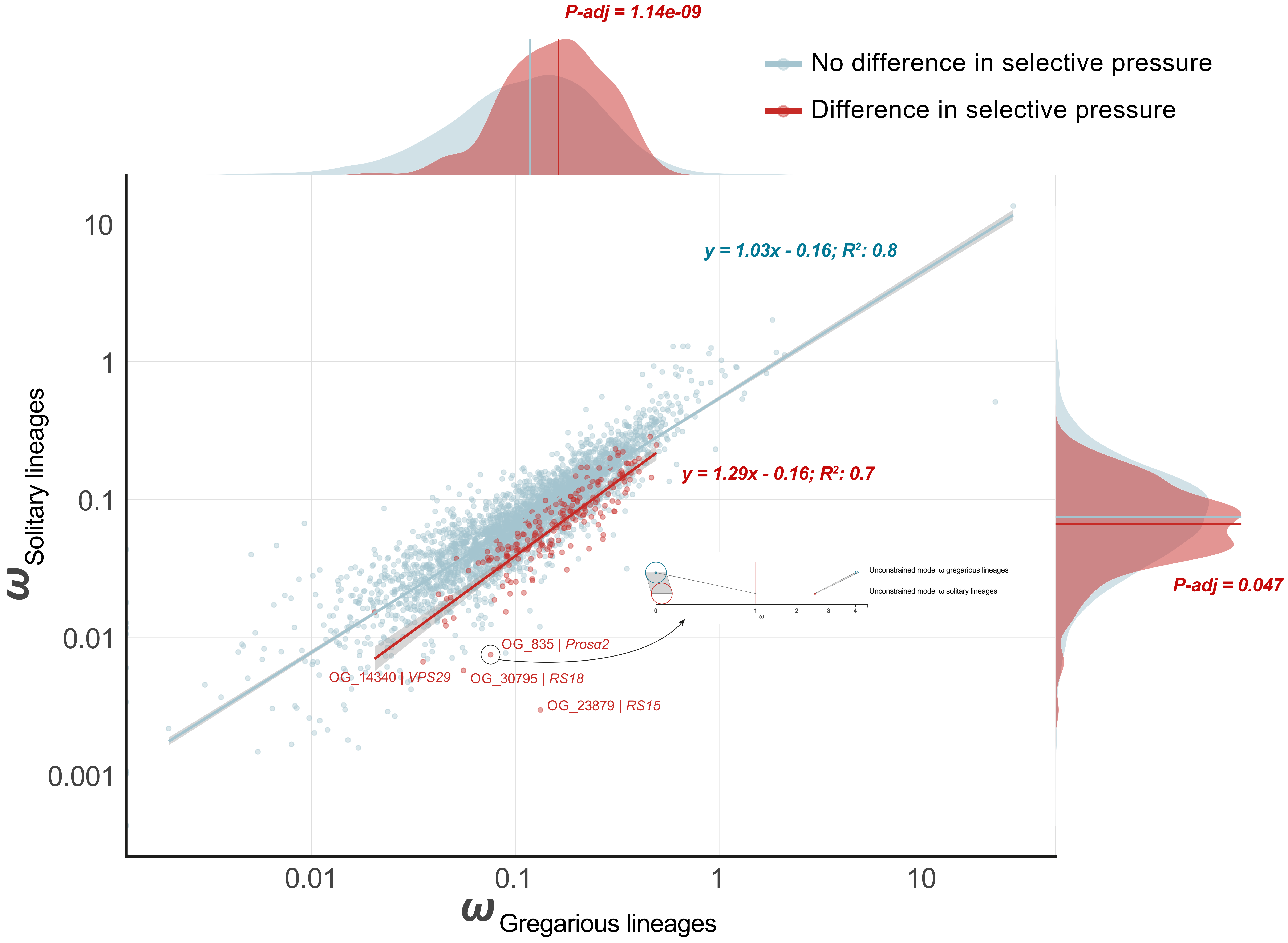


GR Clade legend: ■ CO_2 receptors ■ Sugar receptors ■ BS: Butterflies-specific ■ Fructose receptors ■ LS: Lepidoptera-specific

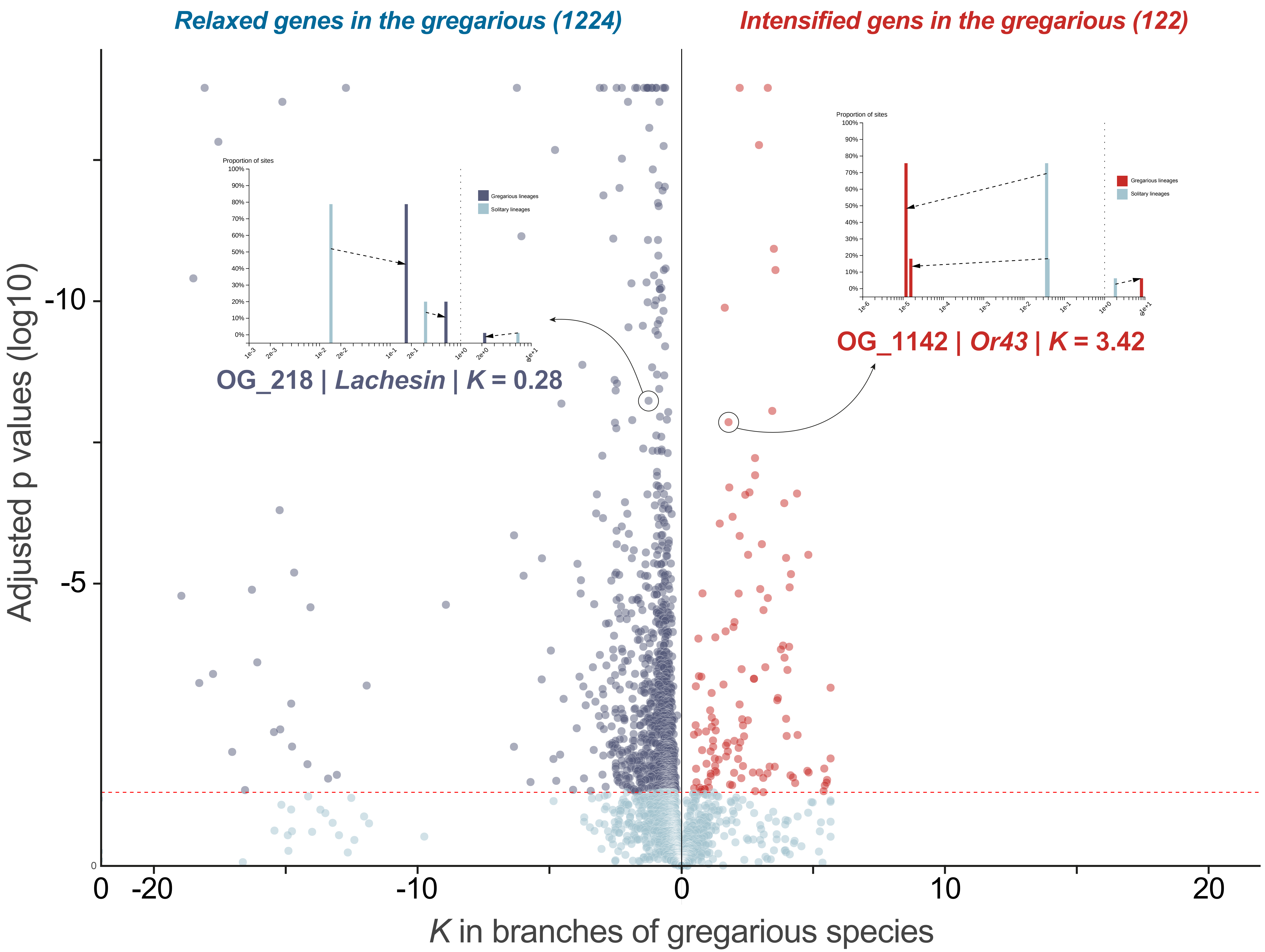
» Selecting pressures: Positive/Purifying/Convergence »



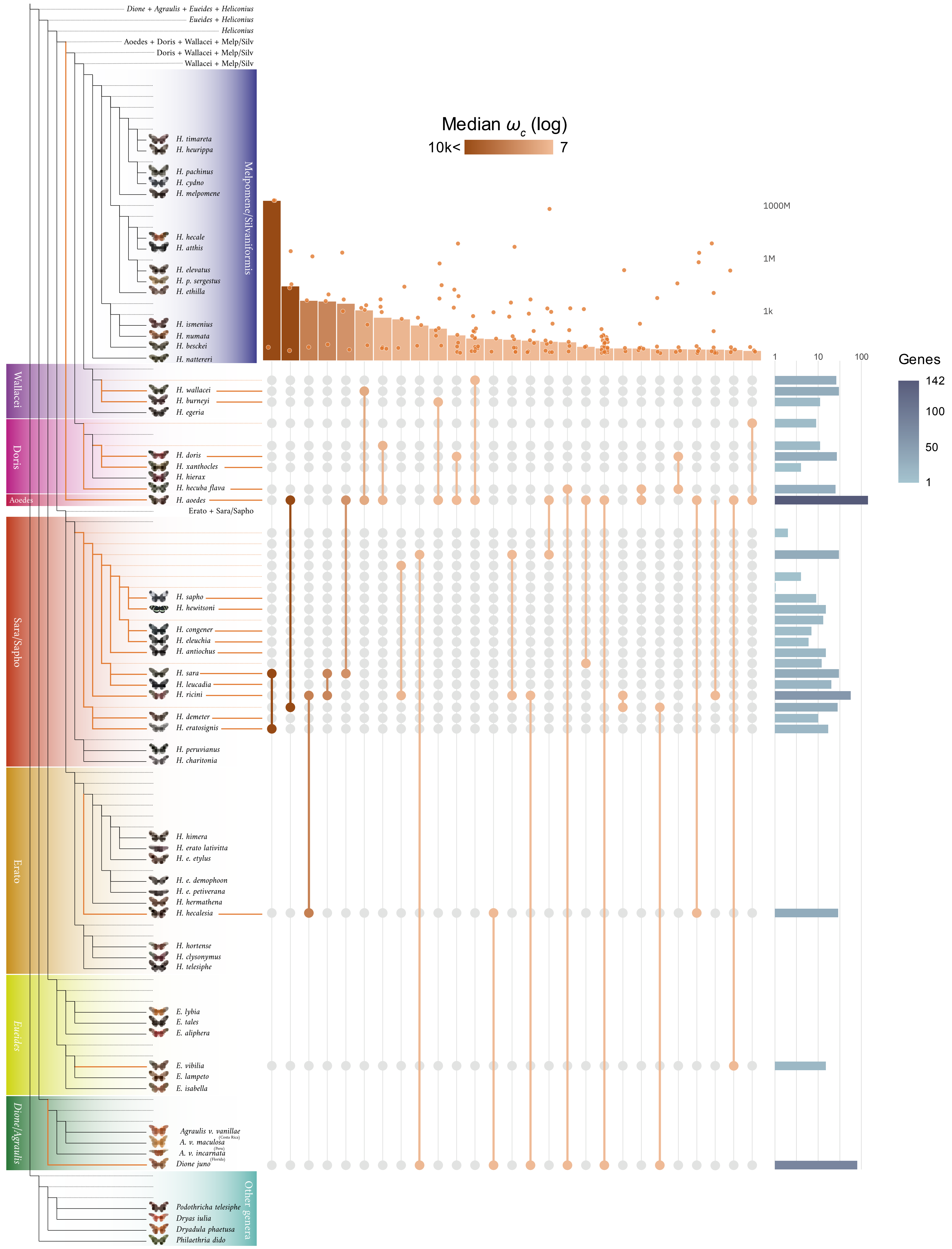
» Selecting pressures: Positive/Purifying/Convergence »



» Selecting pressures: Positive/Purifying/Convergence »



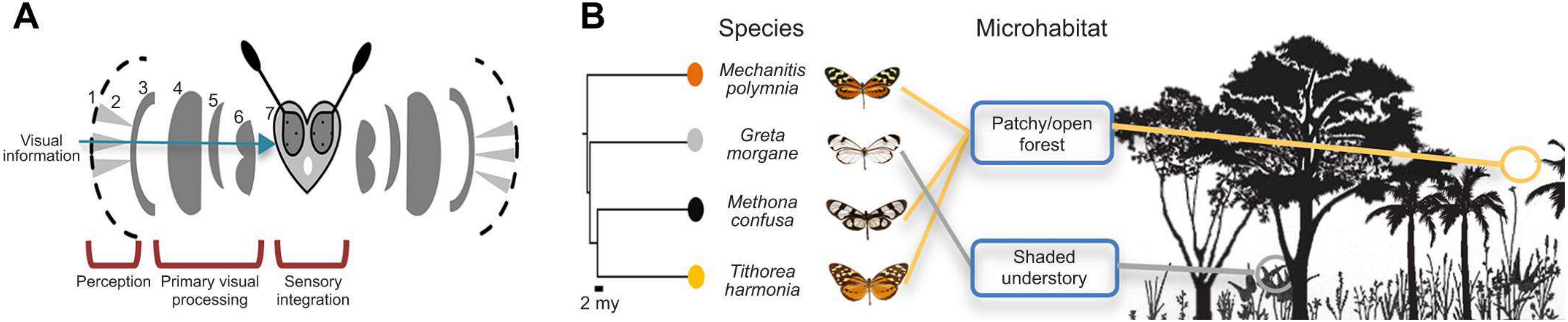
» Selecting pressures: Positive/Purifying/Convergence »



» Loss / Gain of function »

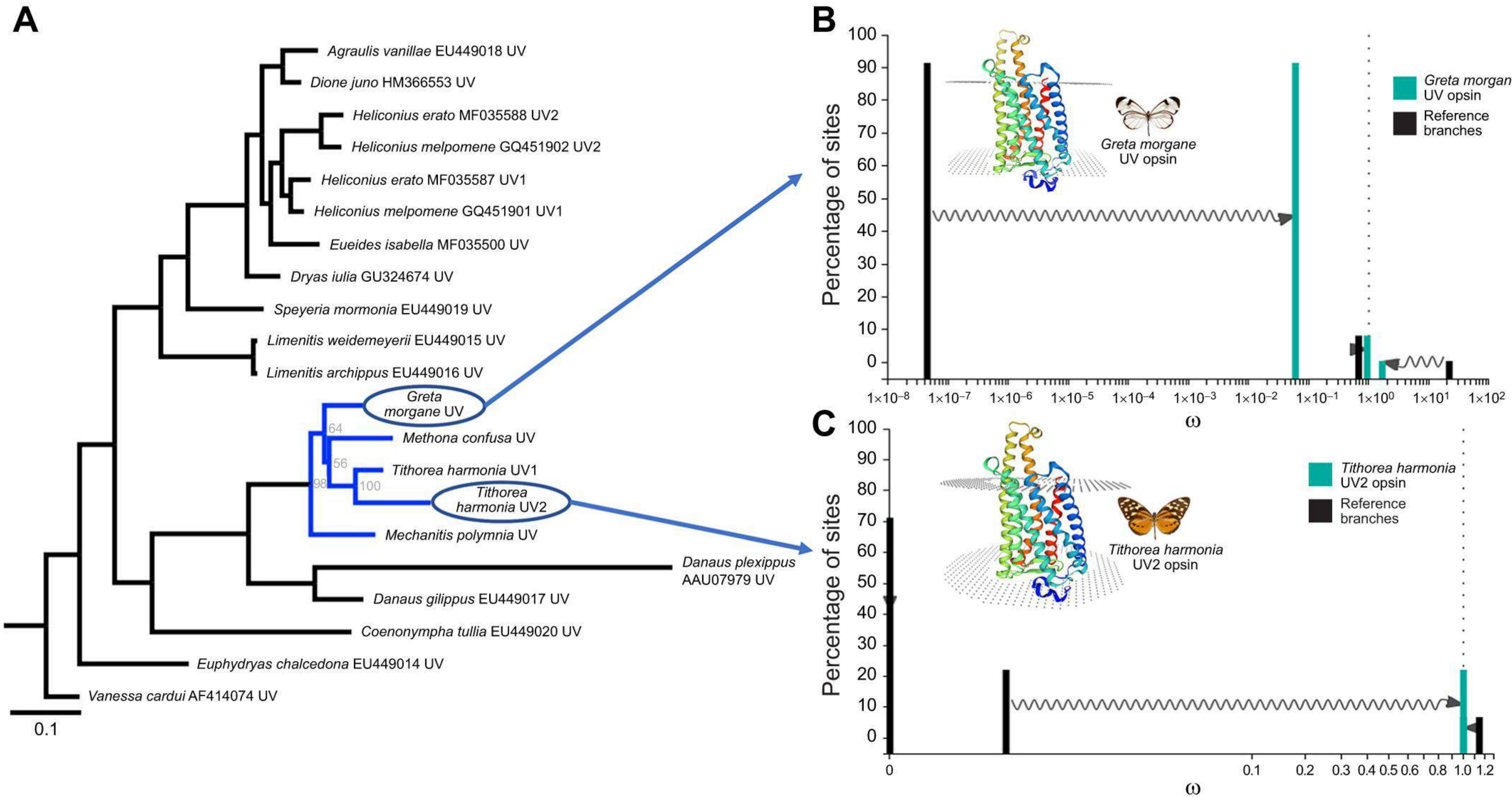
Multiple axes of visual system diversity in Ithomiini, an ecologically diverse tribe of mimetic butterflies | *https://journals.biologists.com/jeb/article/226/24/jeb246423/336534*

» Loss of function »



Multiple axes of visual system diversity in Ithomiini, an ecologically diverse tribe of mimetic butterflies | https://journals.biologists.com/jeb/article/226/24/jeb246423/336534

» Loss of function »



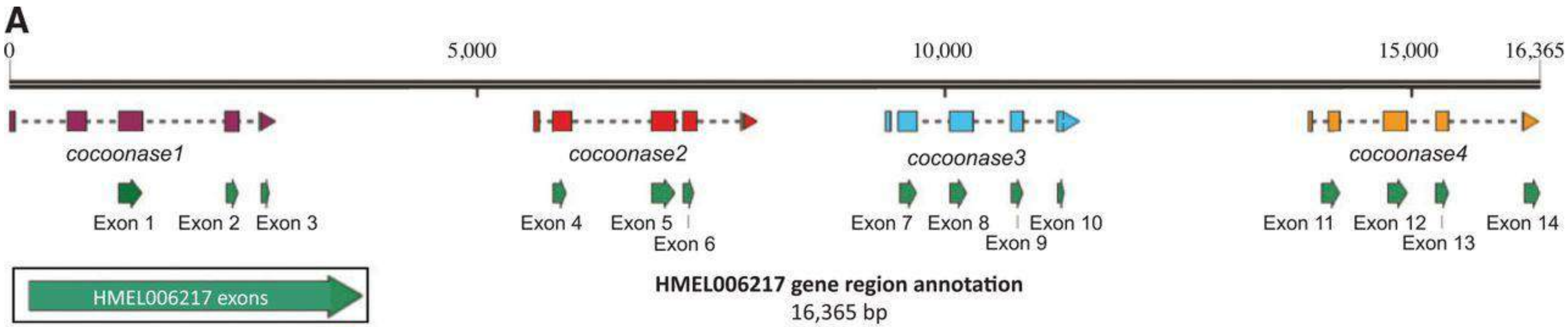
Multiple axes of visual system diversity in Ithomiini, an ecologically diverse tribe of mimetic butterflies | https://journals.biologists.com/jeb/article/220/24/jeb240423/330534

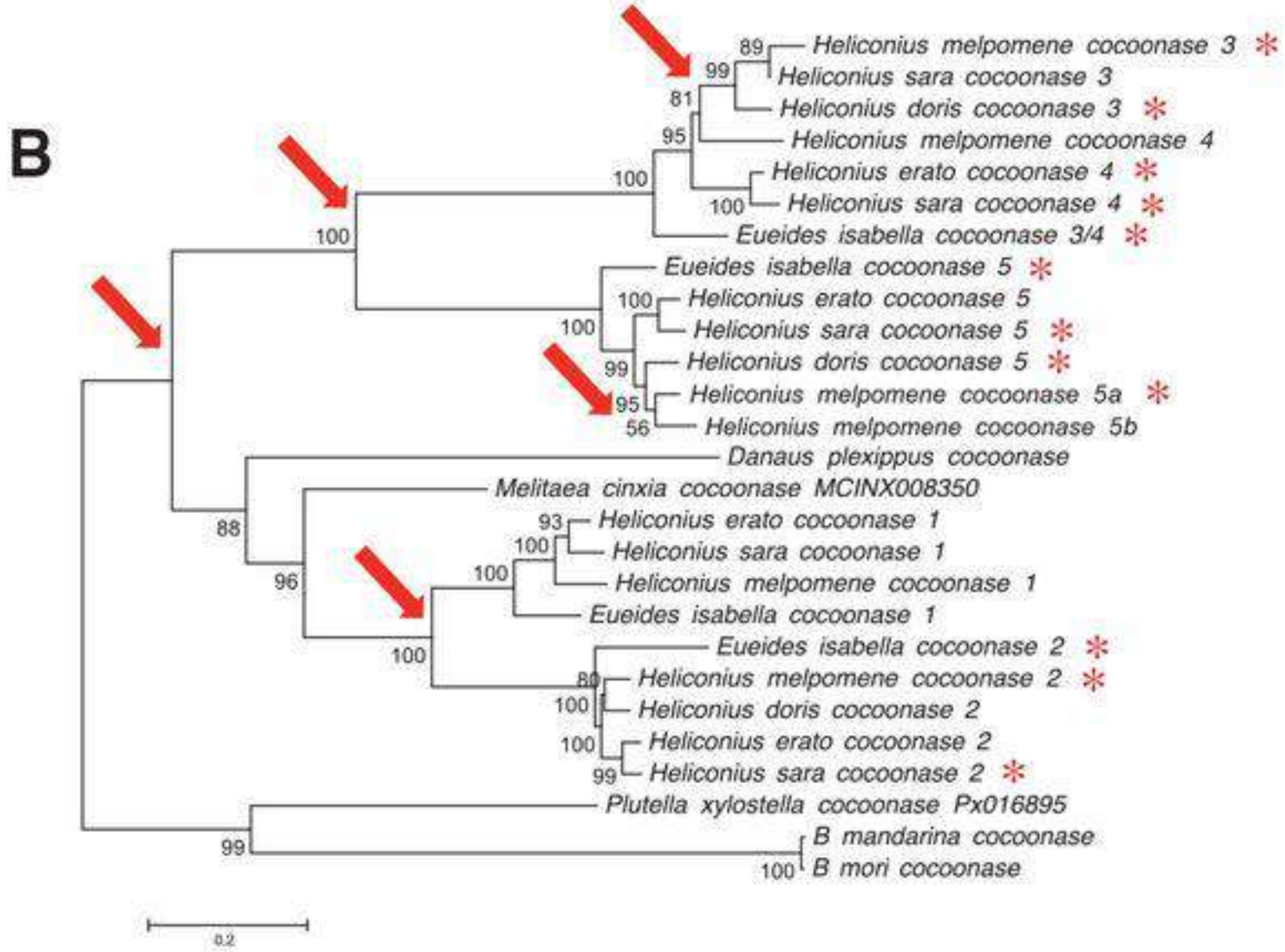
» (*Putative*) Gain of function »

Cocoonase: a protease secreted during the emergence of silk moths

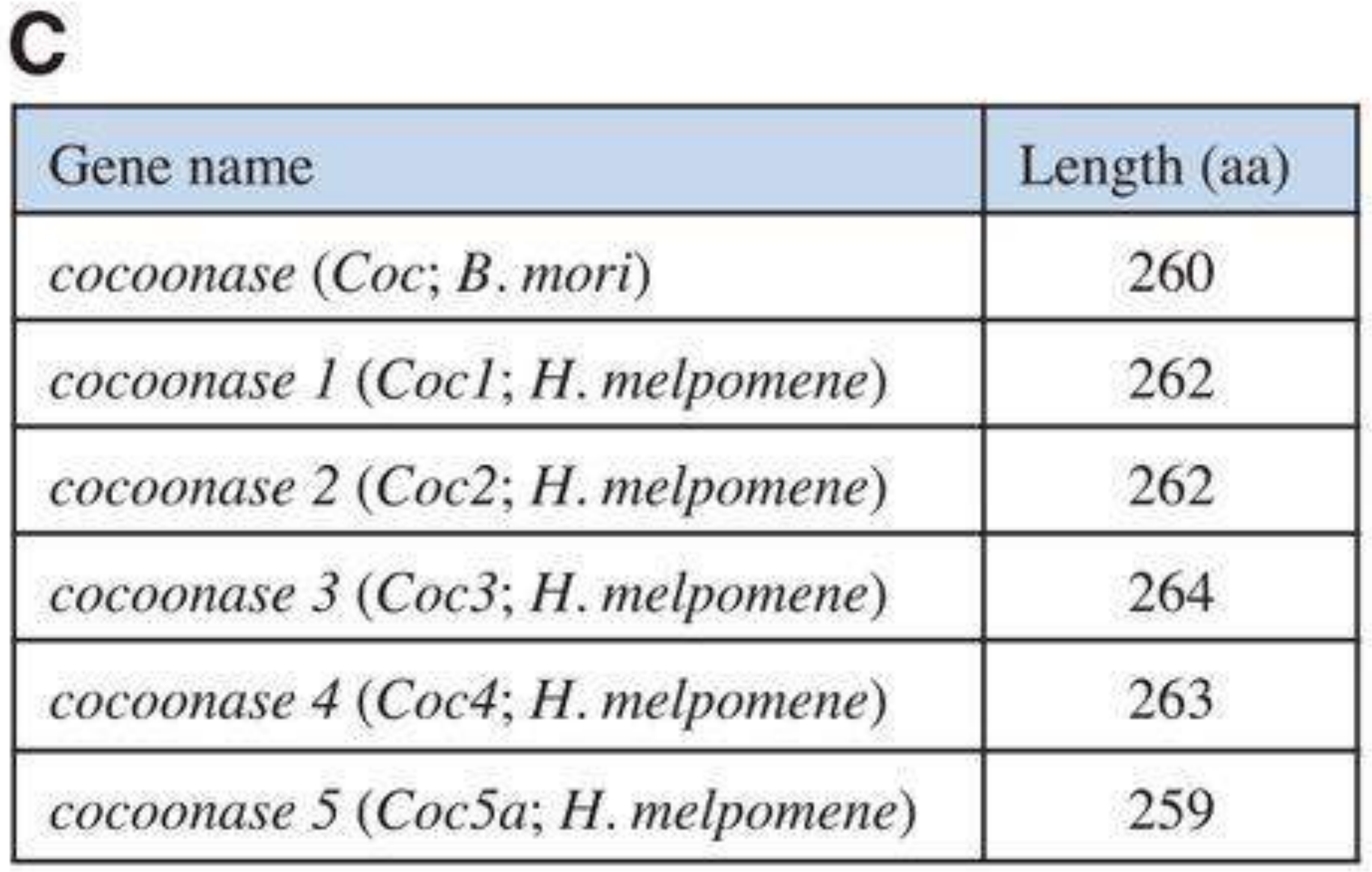
Purification and characterization of cocoonase from the silkworm Bombyx mori | *https://academic.oup.com/bbb/article/78/2/202/5939598*

» (*Putative*) Gain of function »




B
Heliconius melpomene cocoonase 3 *
Heliconius sara cocoonase 3
Heliconius doris cocoonase 3 *
Heliconius melpomene cocoonase 4
Heliconius erato cocoonase 4 *
Heliconius sara cocoonase 4 *
Eueides isabella cocoonase 3/4 *
Eueides isabella cocoonase 5 *
Heliconius erato cocoonase 5
Heliconius sara cocoonase 5 *
Heliconius doris cocoonase 5 *
Heliconius melpomene cocoonase 5a *
Heliconius melpomene cocoonase 5b
Danaus plexippus cocoonase
Melitaea cinxia cocoonase MCINX008350
Heliconius erato cocoonase 1
Heliconius sara cocoonase 1
Heliconius melpomene cocoonase 1
Eueides isabella cocoonase 1
Eueides isabella cocoonase 2 *
Heliconius melpomene cocoonase 2 *
Heliconius doris cocoonase 2
Heliconius erato cocoonase 2
Heliconius sara cocoonase 2 *
Plutella xylostella cocoonase Px016895
B mandarina cocoonase
B mori cocoonase
89
99
81
95
100
100
100
100
100
99
95
56
88
96
93
100
100
100
80
100
100
99
99
100
0.2

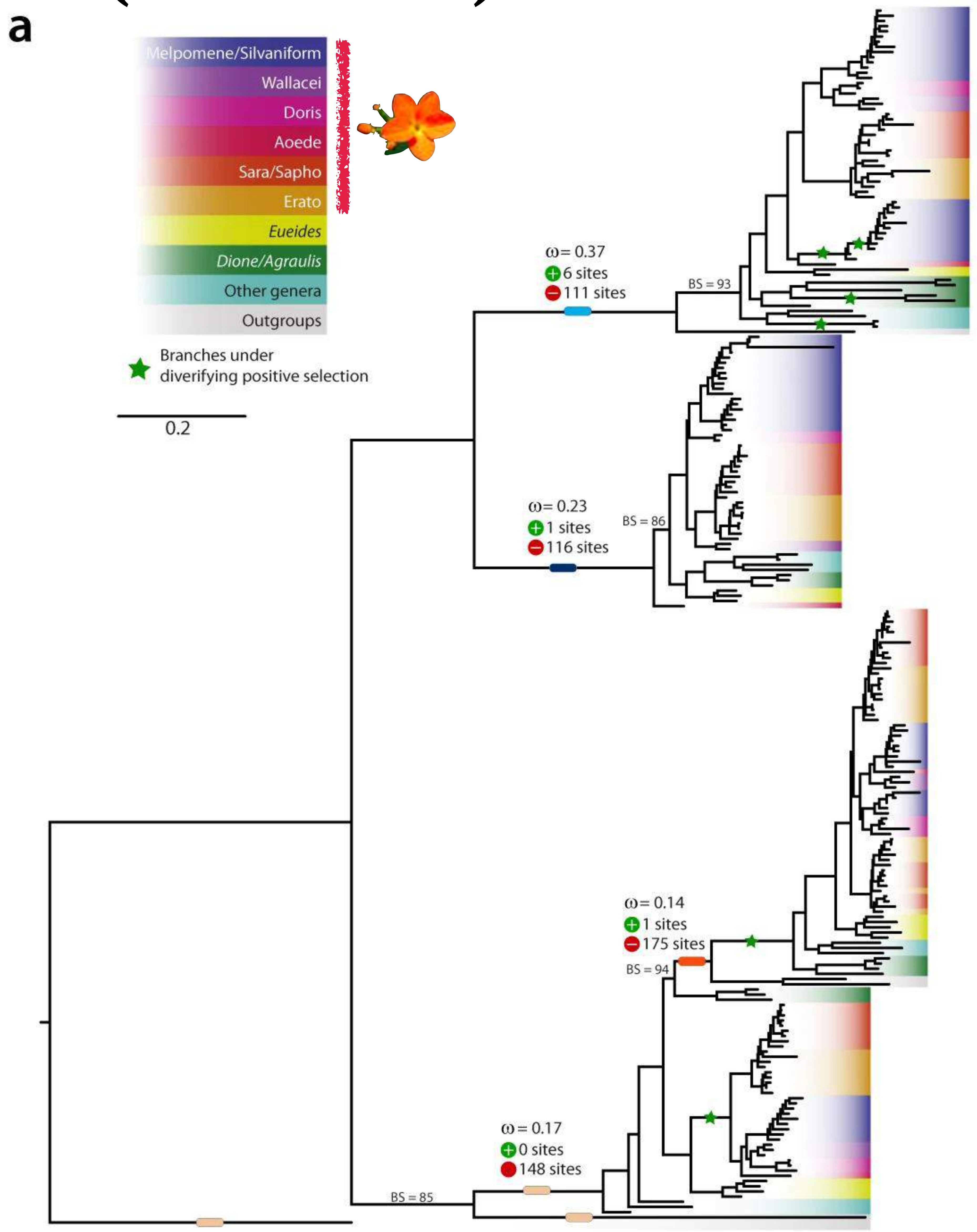
C



Gene name	Length (aa)
cocoonase (*Coc*; *B. mori*)	260
cocoonase 1 (*Coc1*; *H. melpomene*)	262
cocoonase 2 (*Coc2*; *H. melpomene*)	262
cocoonase 3 (*Coc3*; *H. melpomene*)	264
cocoonase 4 (*Coc4*; *H. melpomene*)	263
cocoonase 5 (*Coc5a*; *H. melpomene*)	259

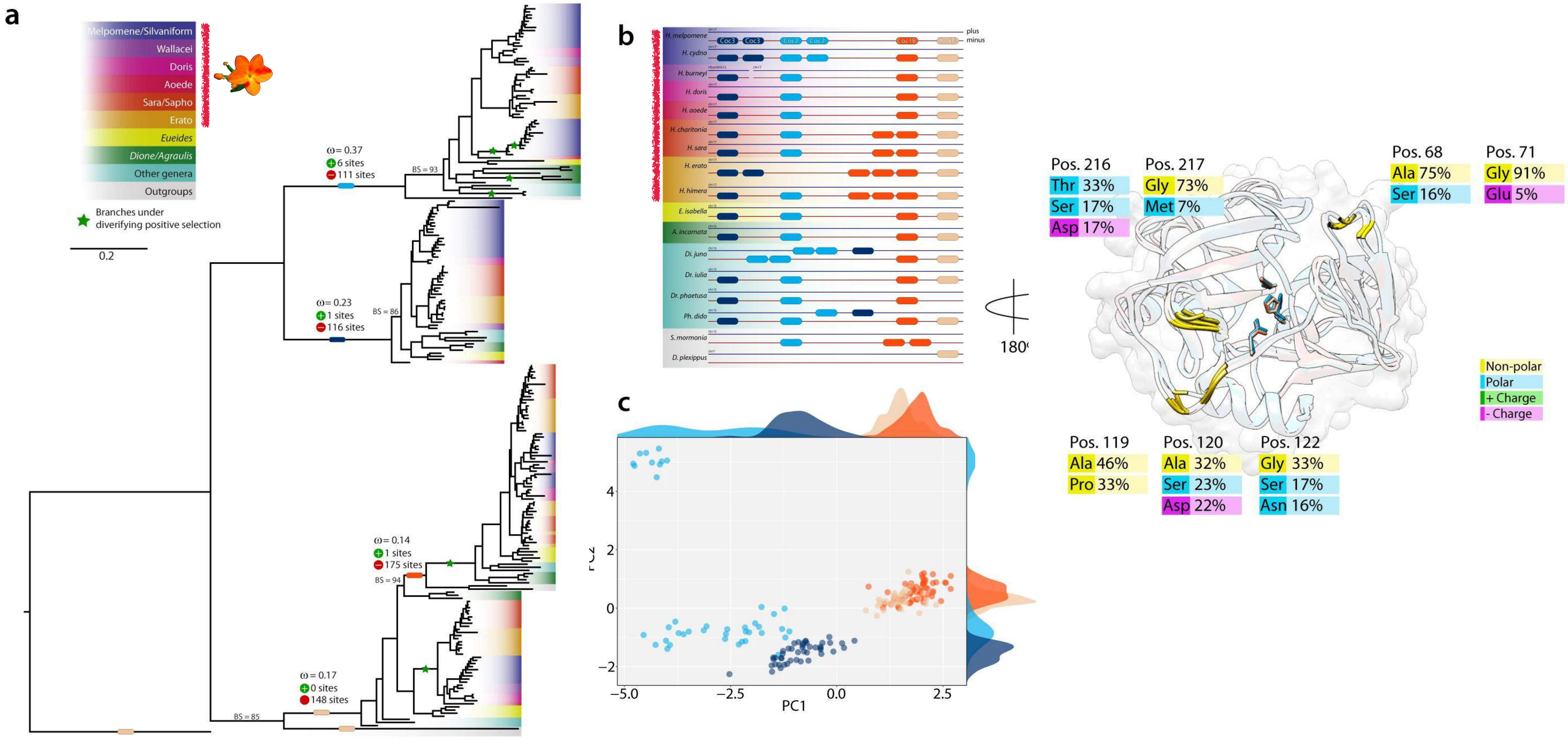
Gene Duplication and Gene Expression Changes Play a Role in the Evolution of Candidate Pollen Feeding Genes in Heliconius Butterflies | https://academic.oup.com/gbe/article/8/8/2581/2198349

» (*Putative*) Gain of function »



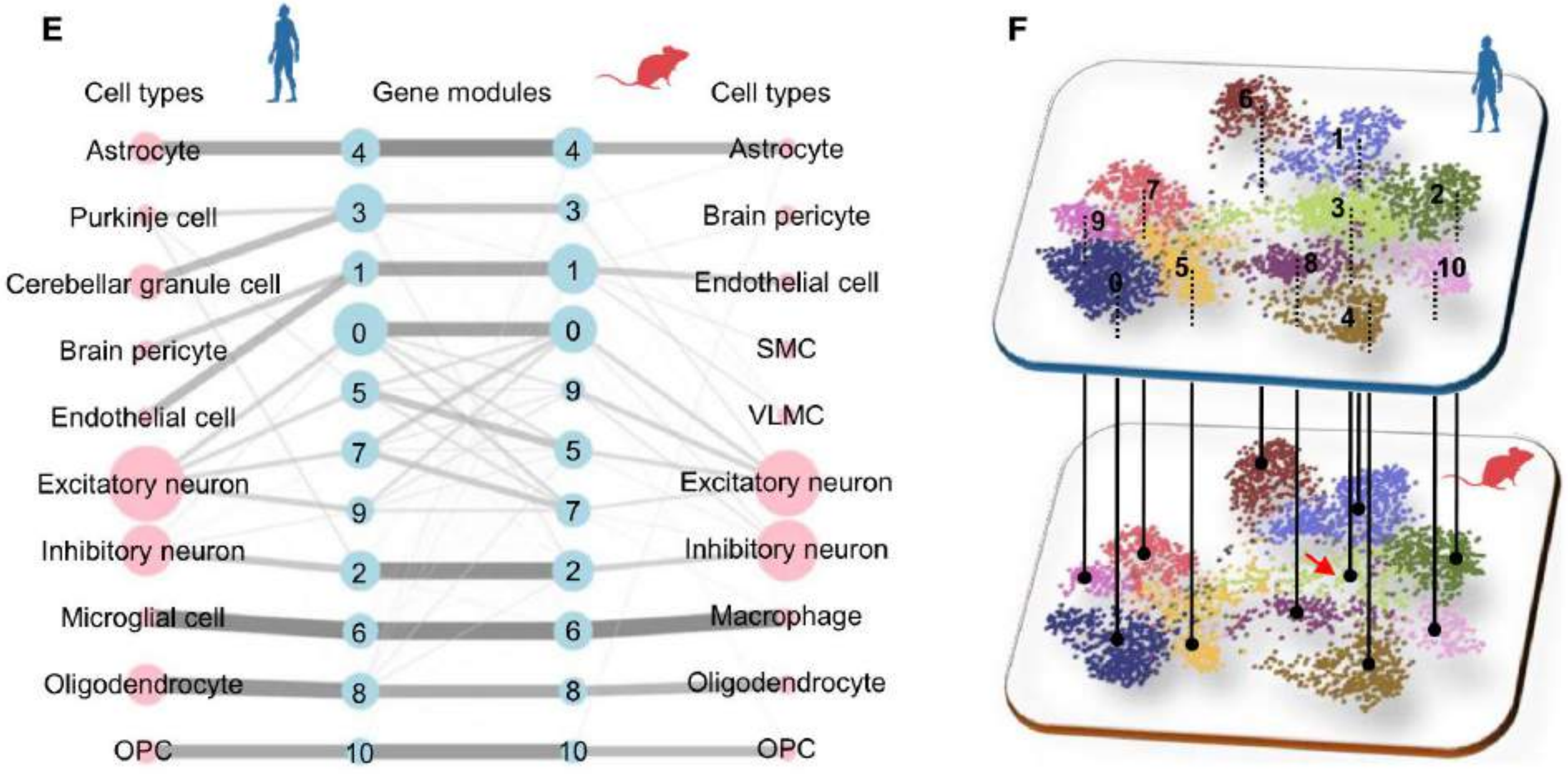
Evolutionary dynamics of genome size and content during the adaptive radiation of Heliconiini butterflies | *https://www.nature.com/articles/s41467-023-41412-5*

» (*Putative*) Gain of function »

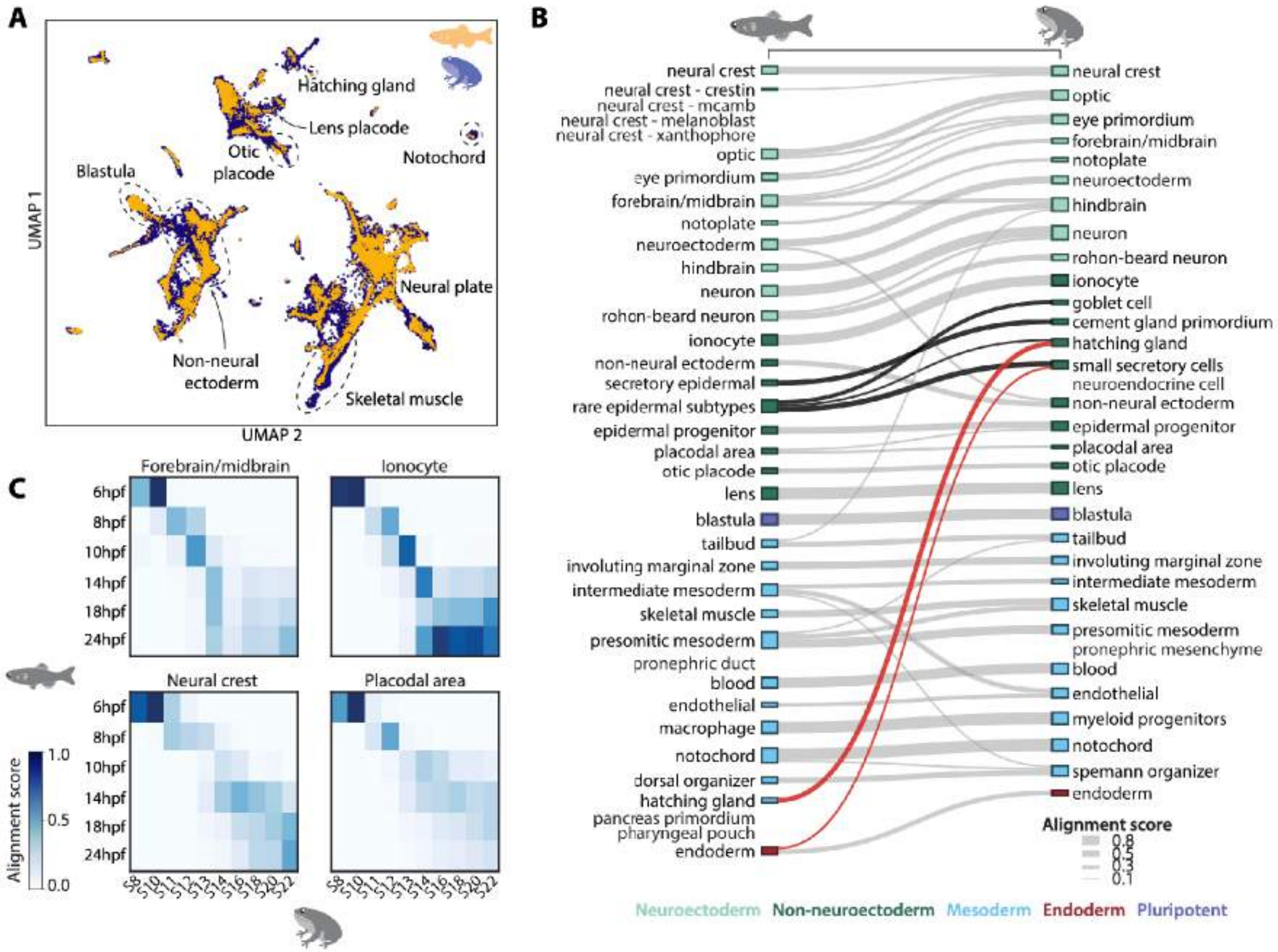


Evolutionary dynamics of genome size and content during the adaptive radiation of Heliconiini butterflies | https://www.nature.com/articles/s41467-023-41412-5

» Single cell RNA data integration »



https://genome.cshlp.org/content/33/1/96



https://elifesciences.org/articles/66747

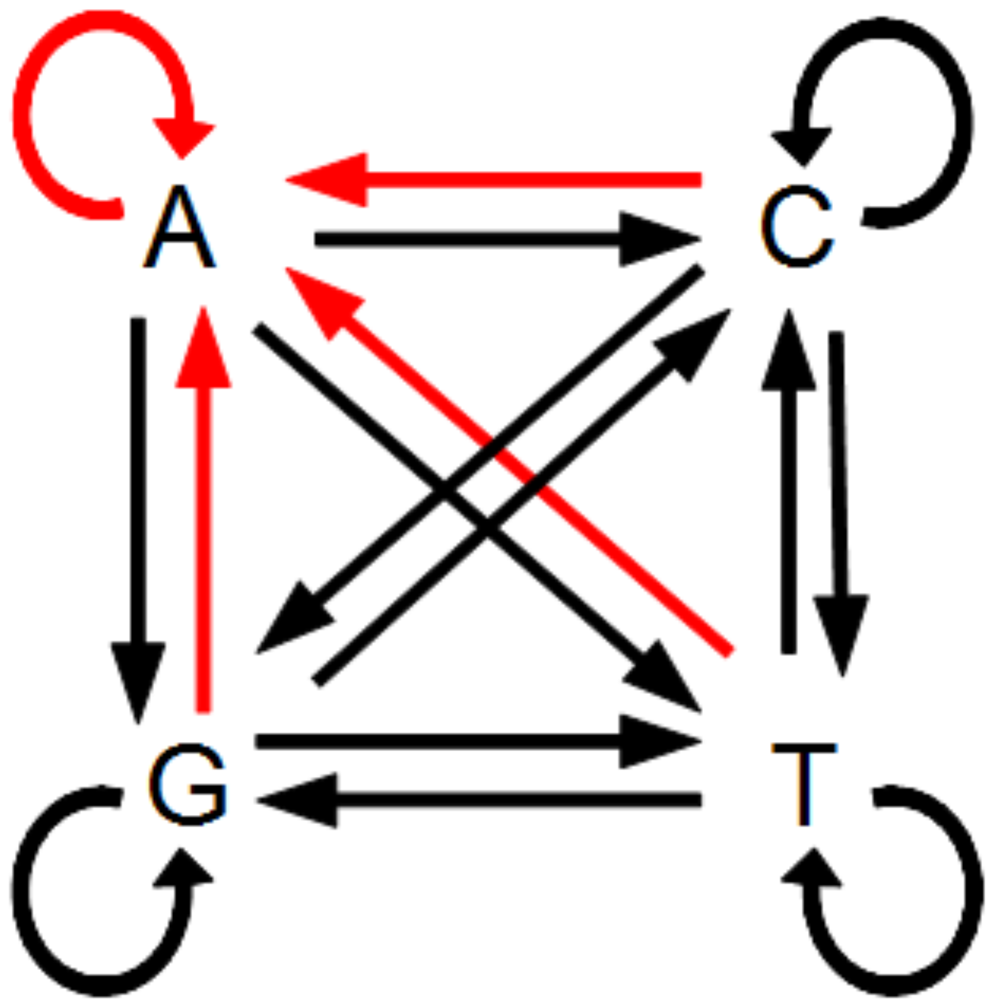
» **Questions ?** »

» Methods »

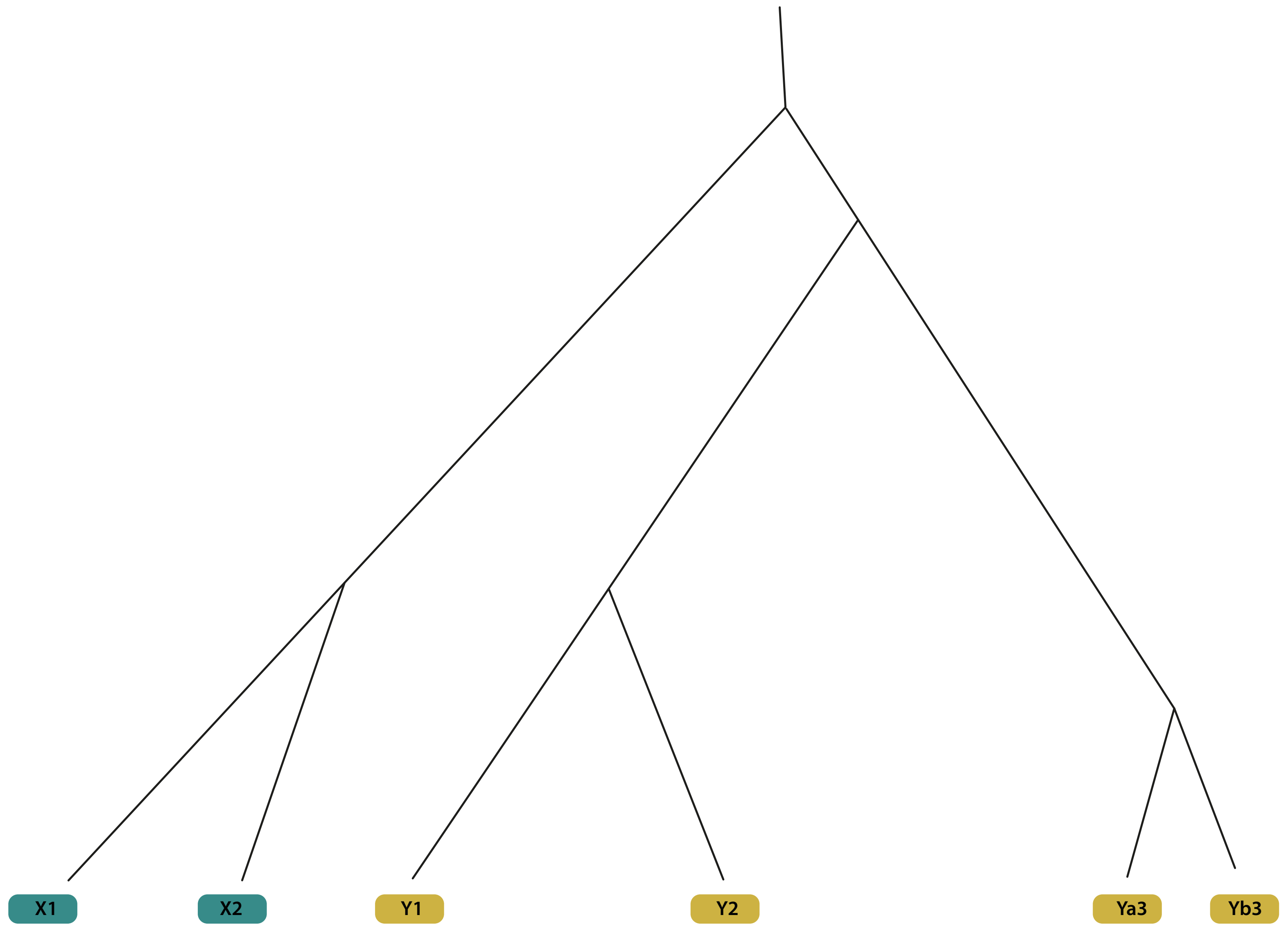
» Phylogeny on Homologous loci (Tree-based) »

```
X1  ATGCTTAGGTCAGCTAGATAGTGGCTAATACCTAGCAGTTGAGTAA
Y1  ATGCTTAGGTCAGCATAGTGGCTAATACCTAGCAGTTGAGTAA
X2  ATGCTTAGGTCAGCTAGATAGTGGCTAATACCTAGCAGTTGAGTAA
Y2  ATGCTTAGGTCAGCATAGTGGCTAGTACCTAGCAGTTGAGTAA
Ya3 ATGCTTAGGTCAGCTAGATAGTGGCTAATACCTAGCAGTTGAGTAA
Yb3 ATGCTTAGGTCAGCATAGTGGCTAATACCTAGCAGTTGAGTAA
```

```
X1  ATGCTTAGGTCAGCTAGATAGTGGCTAA---TACCTAGCAGTTGAGTAA
Y1  ATGCTTAGGTCAGC---ATAGTGGCTAA---TACCTAGCAGTTGAGTAA
X2  ATGCTTAGGTCAGCTAGATAGTGGCTAA---TACCTAGCAGTTGAGTAA
Y2  ATGCTTAGGTCAGC---ATAGTGGCTAG---TACCTAGCAGTTGAGTAA
Ya3 ATGCTTAGGTCAGCTAGATAGTGGCTAA---TACCTAGCAGTTGAGTAA
Yb3 ATGCTTAGGTCAGC---ATAGTGGCTAA---TACCTAGCAGTTGAGTAA
```



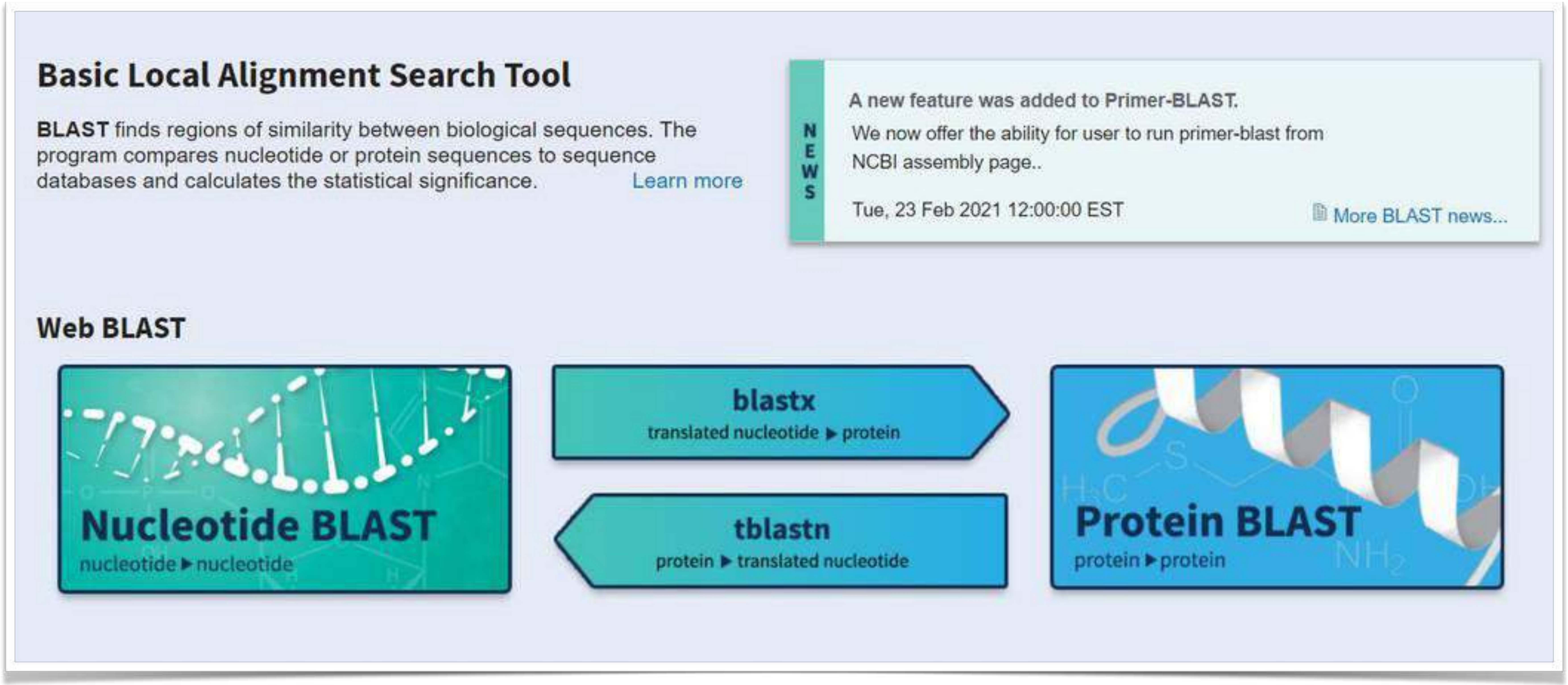
Substitution model



Phylogenetic tree building in the genomic age | https://www.nature.com/articles/s41576-020-0233-0

» Homology based-methods »

BLAST (**B**asic **L**ocal **A**lignment **S**earch **T**ool) is a widely used algorithm for comparing biological sequences.



Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. Learn more

NEWS

A new feature was added to Primer-BLAST.

We now offer the ability for user to run primer-blast from NCBI assembly page..

Tue, 23 Feb 2021 12:00:00 EST

More BLAST news...

Web BLAST

» Homology based-methods »

BLAST identifies regions of *local similarity* between sequences by breaking the query and database sequences into smaller segments (called words) and then aligning these segments.

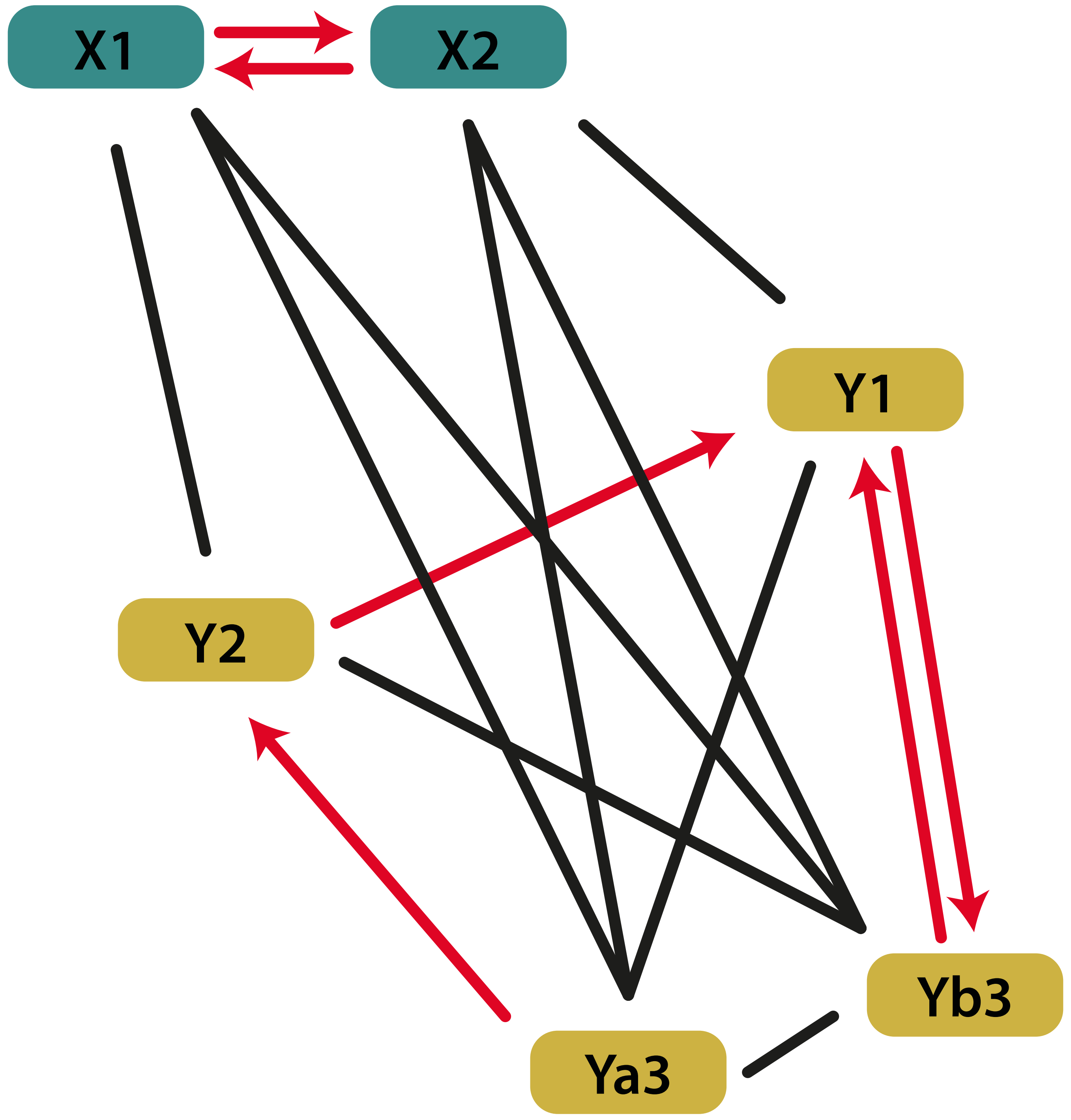
It does not identify ORTHOLOGY *per se!!!*

It is a **sequence similarity search tool**, and while it can help identify genes with similar sequences, it does not directly determine **orthology**

» Homology based-methods »

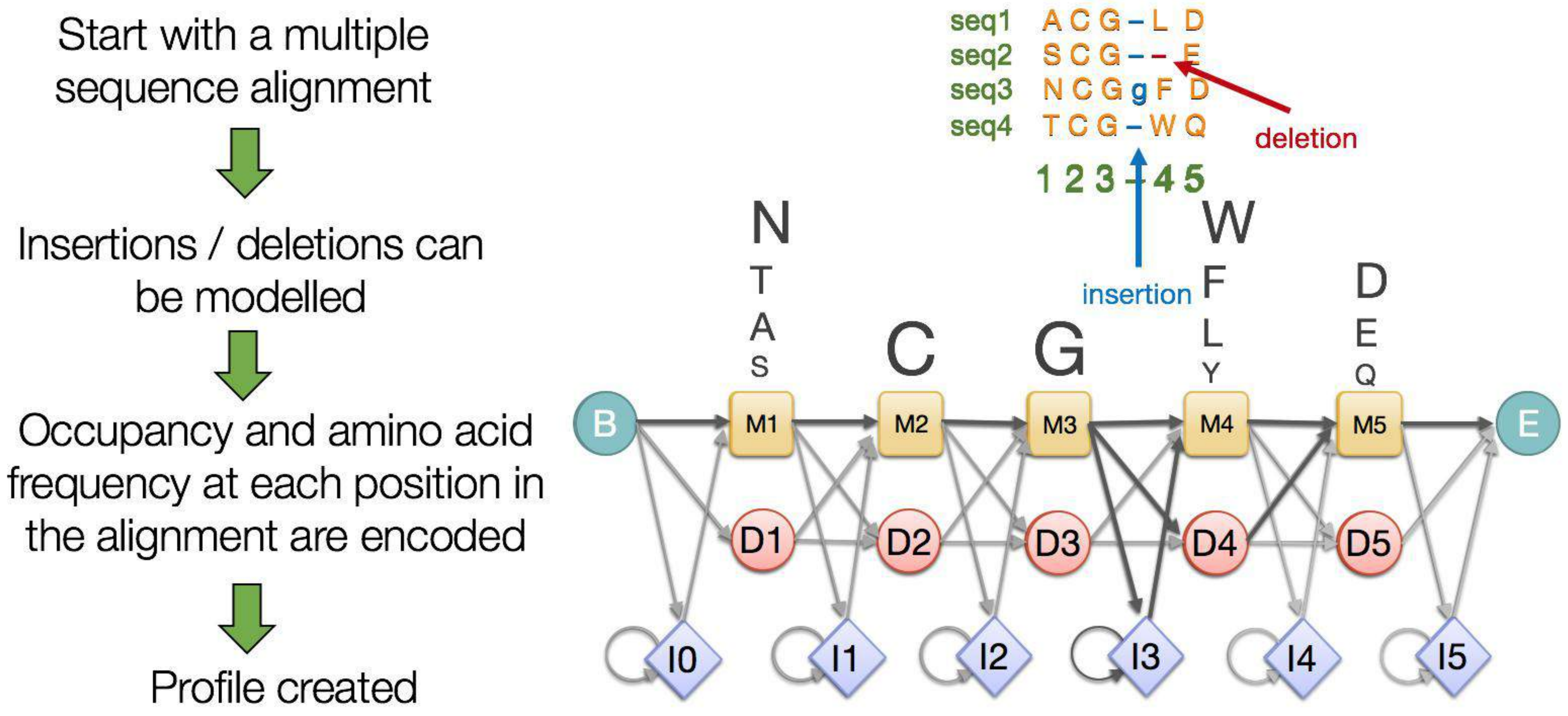
You can implement a **reciprocal Best Hits (RBH)** to identify two genes which are each other's best hits across species, but this is an approximation and not proof.

It can fail in cases like gene duplication, loss, or incomplete genomes.



» Hidden Markov Models (HMMs) »

It's a ***probabilistic model*** used to describe sequences (DNA, RNA, or proteins) based on their underlying statistical properties.



Start with a multiple sequence alignment

Insertions / deletions can be modelled

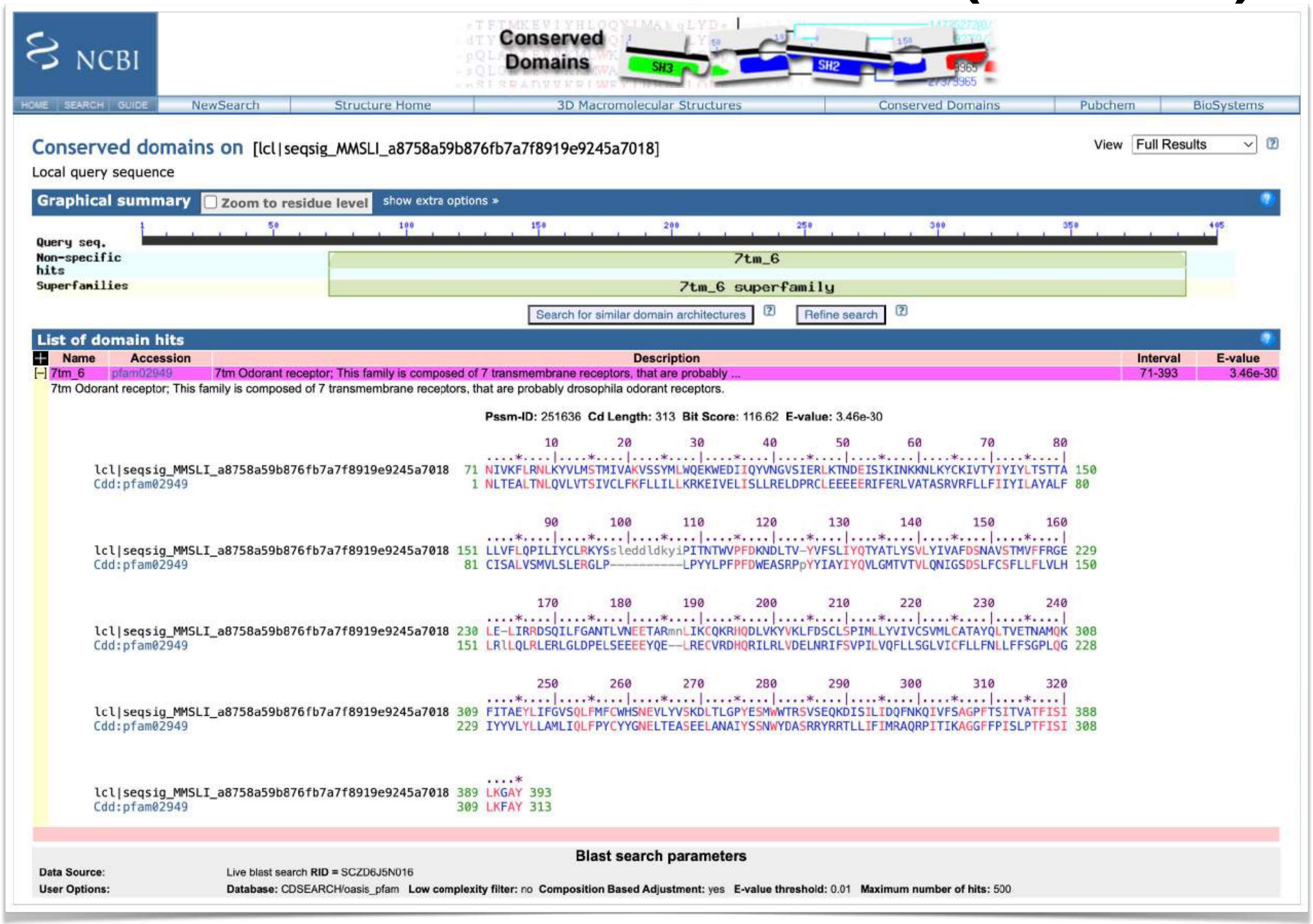
Occupancy and amino acid frequency at each position in the alignment are encoded

Profile created

```
seq1  A C G - L D
seq2  S C G - - E
seq3  N C G g F D
seq4  T C G - W Q
      1 2 3   4 5
```

https://www.ebi.ac.uk/training/online/courses/pfam-creating-protein-families/what-are-profile-hidden-markov-models-hmms/

» Hidden Markov Models (HMMs) »



Conserved domains on [lcl|seqsig_MMSLI_a8758a59b876fb7a7f8919e9245a7018]

Local query sequence

View: Full Results

Graphical summary □ Zoom to residue level show extra options »

List of domain hits

Name	Accession	Description	Interval	E-value
7tm_6	pfam02949	7tm Odorant receptor; This family is composed of 7 transmembrane receptors, that are probably ...	71-393	3.46e-30

7tm Odorant receptor; This family is composed of 7 transmembrane receptors, that are probably drosophila odorant receptors.

Pssm-ID: 251636 **Cd Length:** 313 **Bit Score:** 116.62 **E-value:** 3.46e-30

```
                                                         10        20        30        40        50        60        70        80
                                                     ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
lcl|seqsig_MMSLI_a8758a59b876fb7a7f8919e9245a7018  71 NIVKFLRNLKYVLMSTMIVAKVSSYMLWQEKWEDIIQYVNGVSIERLKTNDEISIKINKKNLKYCKIVTYIYIYLTSTTA 150
Cdd:pfam02949                                        1 NLTEALTNLQVLVTSIVCLFKFLLILLKRKEIVELISLLRELDPRCLEEEERIFERLVATASRVRFLLFIIYILAYALF 80

                                                         90       100       110       120       130       140       150       160
                                                     ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
lcl|seqsig_MMSLI_a8758a59b876fb7a7f8919e9245a7018 151 LLVFLQPILIYCLRKYSsleddldkyiPITNTWVPFDKNDLTV-YVFSLIYQTYATLYSVLYIVAFDSNAVSTMVFFRGE 229
Cdd:pfam02949                                       81 CISALVSMVLSLERGLP----------LPYYLPFPFDWEASRPpYYIAYIYQVLGMTVTVLQNIGSDSLFCSFLLFLVLH 150

                                                        170       180       190       200       210       220       230       240
                                                     ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
lcl|seqsig_MMSLI_a8758a59b876fb7a7f8919e9245a7018 230 LE-LIRRDSQILFGANTLVNEETARmnLIKCQKRHQDLVKYVKLFDSCLSPIMLLYVIVCSVMLCATAYQLTVETNAMQK 308
Cdd:pfam02949                                      151 LRlLQLRLERLGLDPELSEEEEYQE--LRECVRDHQRILRLVDELNRIFSVPILVQFLLSGLVICFLLFNLLFFSGPLQG 228

                                                        250       260       270       280       290       300       310       320
                                                     ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
lcl|seqsig_MMSLI_a8758a59b876fb7a7f8919e9245a7018 309 FITAEYLIFGVSQLFMFCWHSNEVLYVSKDLTLGPYESMWWTRSVSEQKDISILIDQFNKQIVFSAGPFTSITVATFISI 388
Cdd:pfam02949                                      229 IYYVLYLLAMLIQLFPYCYYGNELTEASEELANAIYSSNWYDASRRYRRTLLIFIMRAQRPITIKAGGFFPISLPTFISI 308

                                                     ....*
lcl|seqsig_MMSLI_a8758a59b876fb7a7f8919e9245a7018 389 LKGAY 393
Cdd:pfam02949                                      309 LKFAY 313
```

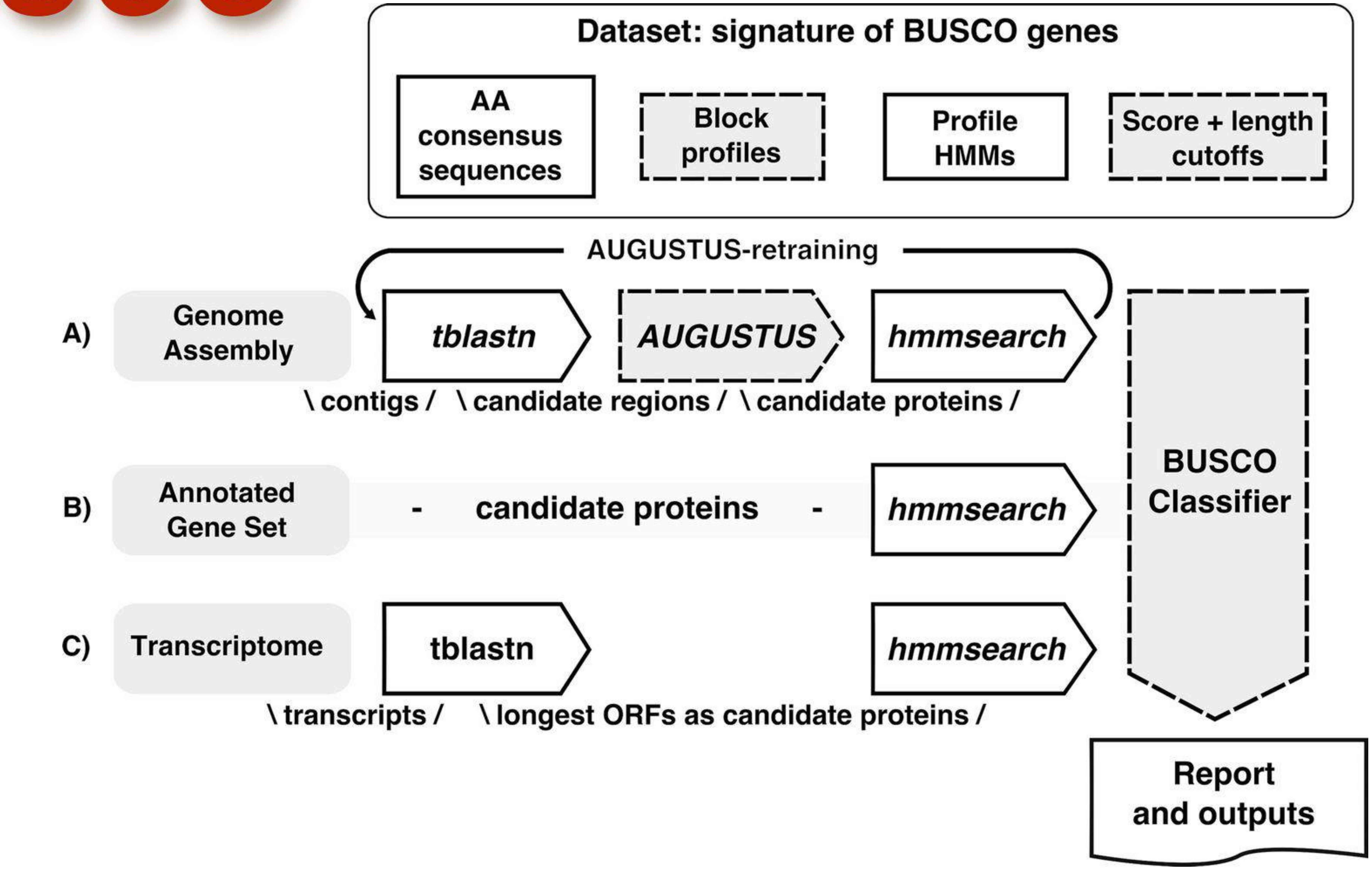
Blast search parameters

Data Source: Live blast search **RID** = SCZD6J5N016

User Options: **Database:** CDSEARCH/oasis_pfam **Low complexity filter:** no **Composition Based Adjustment:** yes **E-value threshold:** 0.01 **Maximum number of hits:** 500

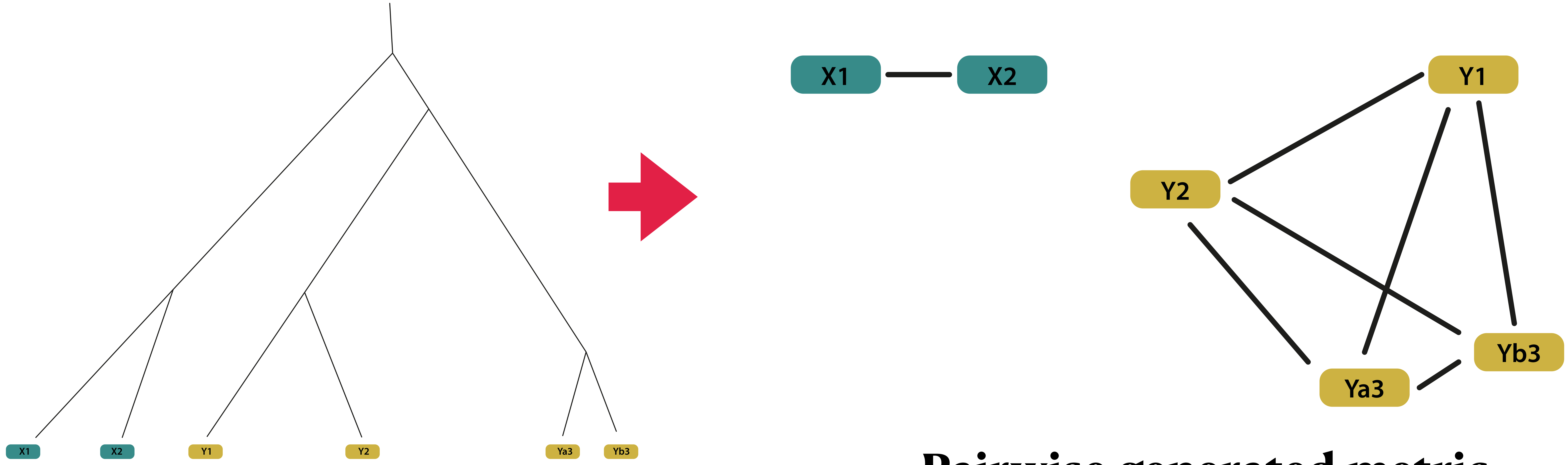
https://www.ebi.ac.uk/training/online/courses/pfam-creating-protein-families/what-are-profile-hidden-markov-models-hmms/

BUSCO



BUSCO: Assessing Genome Assembly and Annotation Completeness | https://link.springer.com/protocol/10.1007/978-1-4939-9173-0_14

» Methods | Network based (with some integration of the phylogeny) »

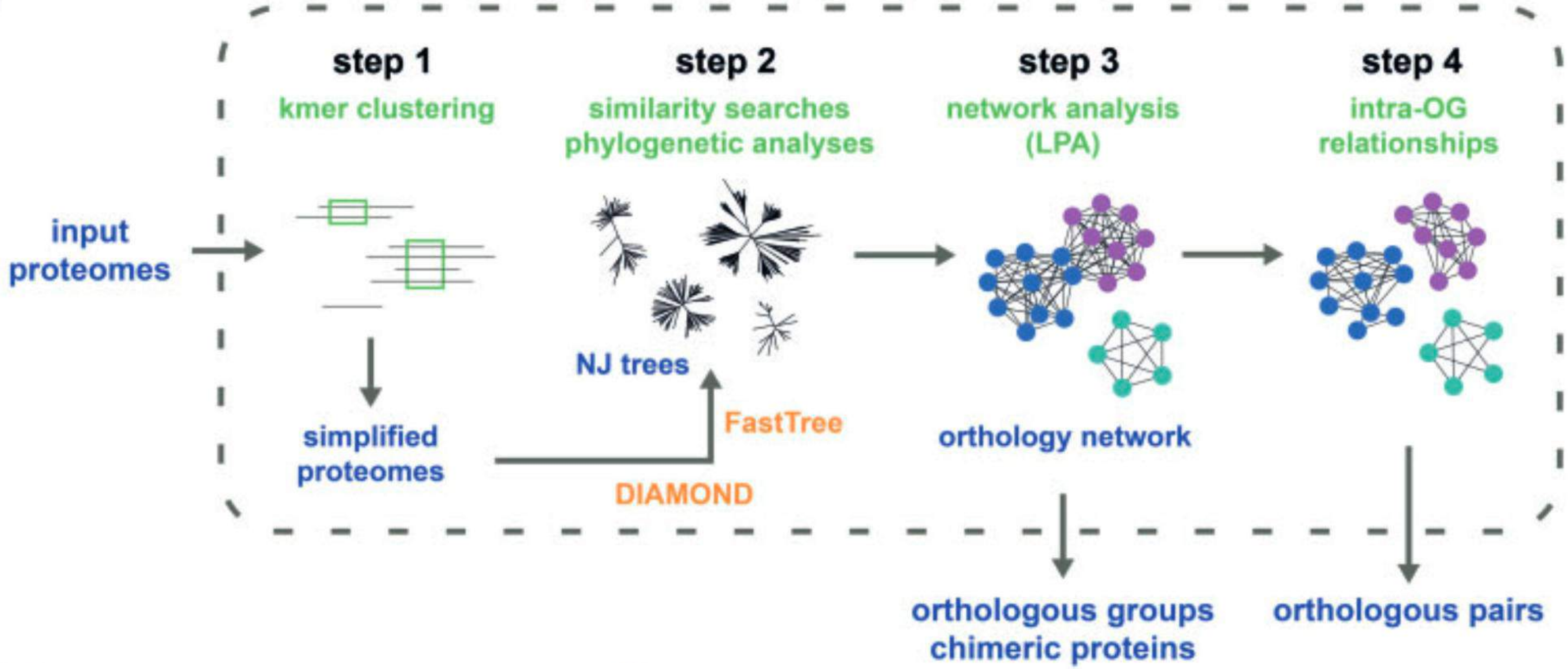


Pairwise generated metric

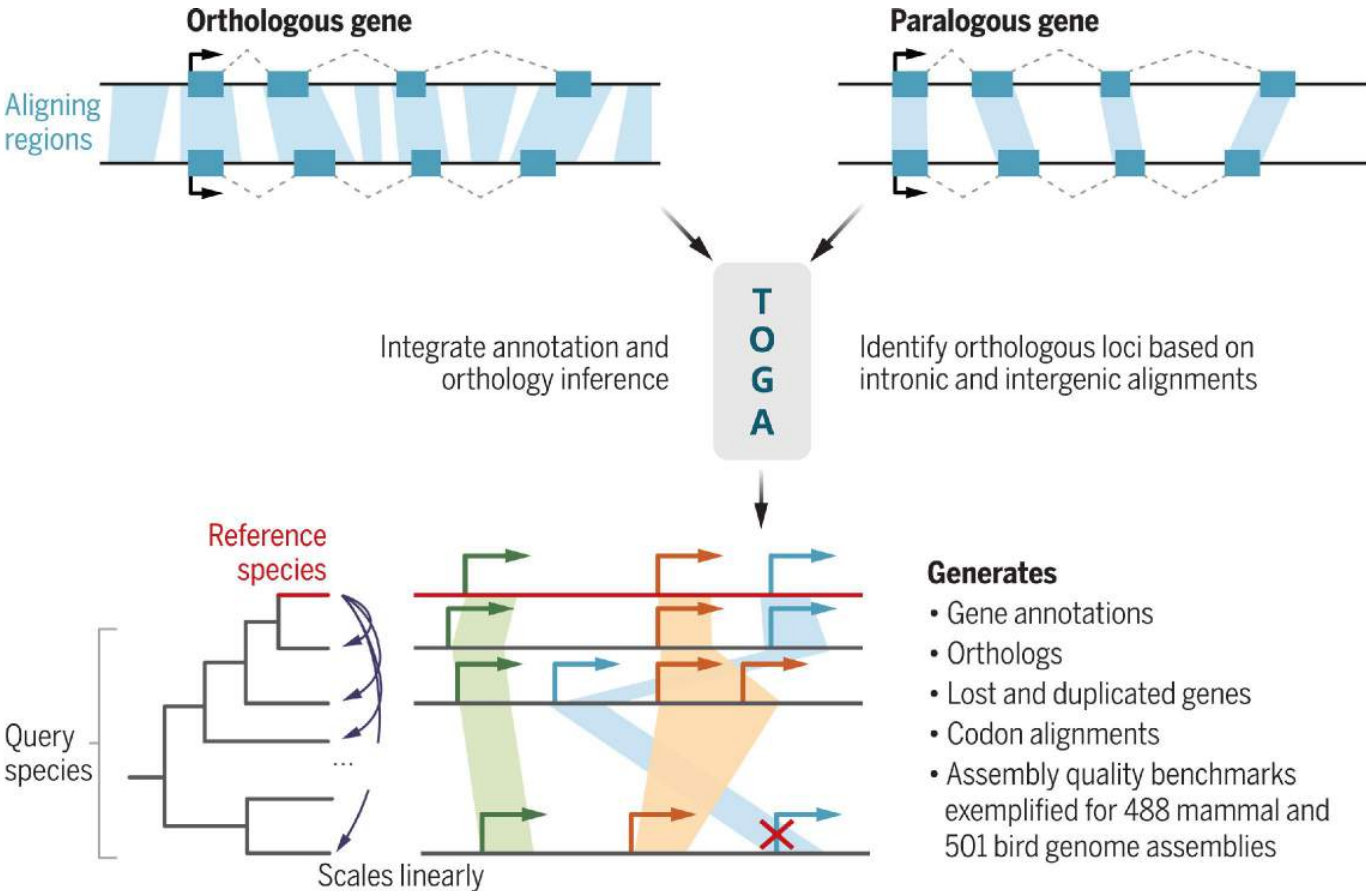
E-value: the number of alignments you would expect to find by chance in a database of a given size.
Bitscore: a normalized score that represents the quality of the alignment. It's independent of the database size.

» Methods | Network based (with some integration of the phylogeny) »

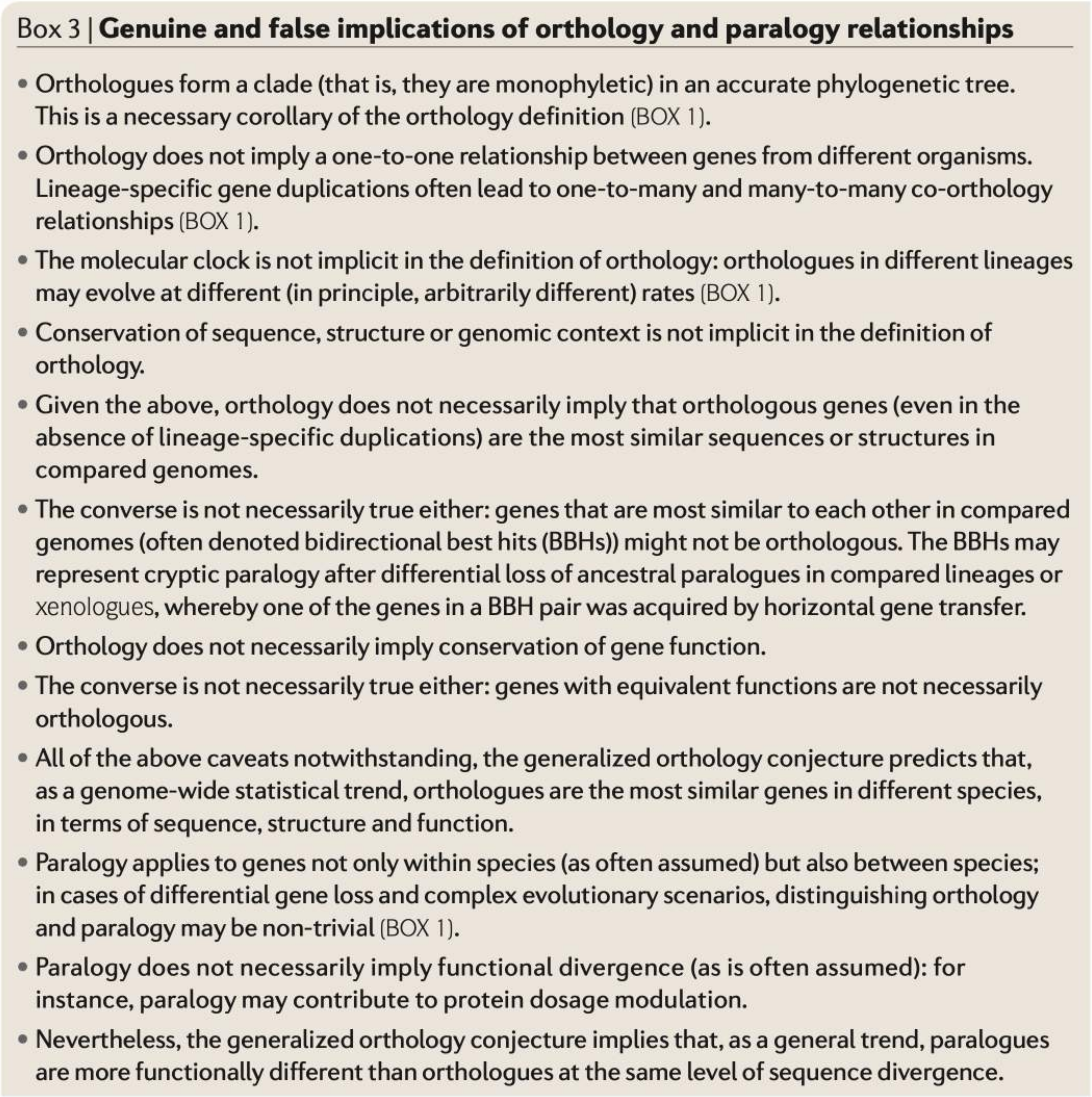
https://academic.oup.com/mbe/article/37/11/3389/5865275



» Methods | Whole-genome alignment based »


Orthologous gene
Paralogous gene
Aligning regions
TOGA
Integrate annotation and orthology inference
Identify orthologous loci based on intronic and intergenic alignments
Reference species
Query species
...
Scales linearly
Generates
• Gene annotations
• Orthologs
• Lost and duplicated genes
• Codon alignments
• Assembly quality benchmarks exemplified for 488 mammal and 501 bird genome assemblies

https://www.science.org/doi/10.1126/science.abn3107



Box 3 | Genuine and false implications of orthology and paralogy relationships

- Orthologues form a clade (that is, they are monophyletic) in an accurate phylogenetic tree. This is a necessary corollary of the orthology definition (BOX 1).
- Orthology does not imply a one-to-one relationship between genes from different organisms. Lineage-specific gene duplications often lead to one-to-many and many-to-many co-orthology relationships (BOX 1).
- The molecular clock is not implicit in the definition of orthology: orthologues in different lineages may evolve at different (in principle, arbitrarily different) rates (BOX 1).
- Conservation of sequence, structure or genomic context is not implicit in the definition of orthology.
- Given the above, orthology does not necessarily imply that orthologous genes (even in the absence of lineage-specific duplications) are the most similar sequences or structures in compared genomes.
- The converse is not necessarily true either: genes that are most similar to each other in compared genomes (often denoted bidirectional best hits (BBHs)) might not be orthologous. The BBHs may represent cryptic paralogy after differential loss of ancestral paralogues in compared lineages or xenologues, whereby one of the genes in a BBH pair was acquired by horizontal gene transfer.
- Orthology does not necessarily imply conservation of gene function.
- The converse is not necessarily true either: genes with equivalent functions are not necessarily orthologous.
- All of the above caveats notwithstanding, the generalized orthology conjecture predicts that, as a genome-wide statistical trend, orthologues are the most similar genes in different species, in terms of sequence, structure and function.
- Paralogy applies to genes not only within species (as often assumed) but also between species; in cases of differential gene loss and complex evolutionary scenarios, distinguishing orthology and paralogy may be non-trivial (BOX 1).
- Paralogy does not necessarily imply functional divergence (as is often assumed): for instance, paralogy may contribute to protein dosage modulation.
- Nevertheless, the generalized orthology conjecture implies that, as a general trend, paralogues are more functionally different than orthologues at the same level of sequence divergence.

Orthologs, Paralogs, and Evolutionary Genomics | https://www.annualreviews.org/content/journals/10.1146/annurev.genet.39.073003.114725

Functional and evolutionary implications of gene orthology | https://www.nature.com/articles/nrg3456

» **Questions ? / Break ?** »

» Whole Genome alignment »

» Conserved non-coding Elements (CNEs) »

» CNEs enrichment (extra) »

» Whole Genome alignment »

» Conserved non-coding Elements (CNEs) »

» CNEEs enrichment (extra) »

» Whole Genome alignment »

» Whole Genome alignment »

> **Deeper Species-tree inference**
> **Comparative gene annotation**
> **Detection of selection / Conservation**

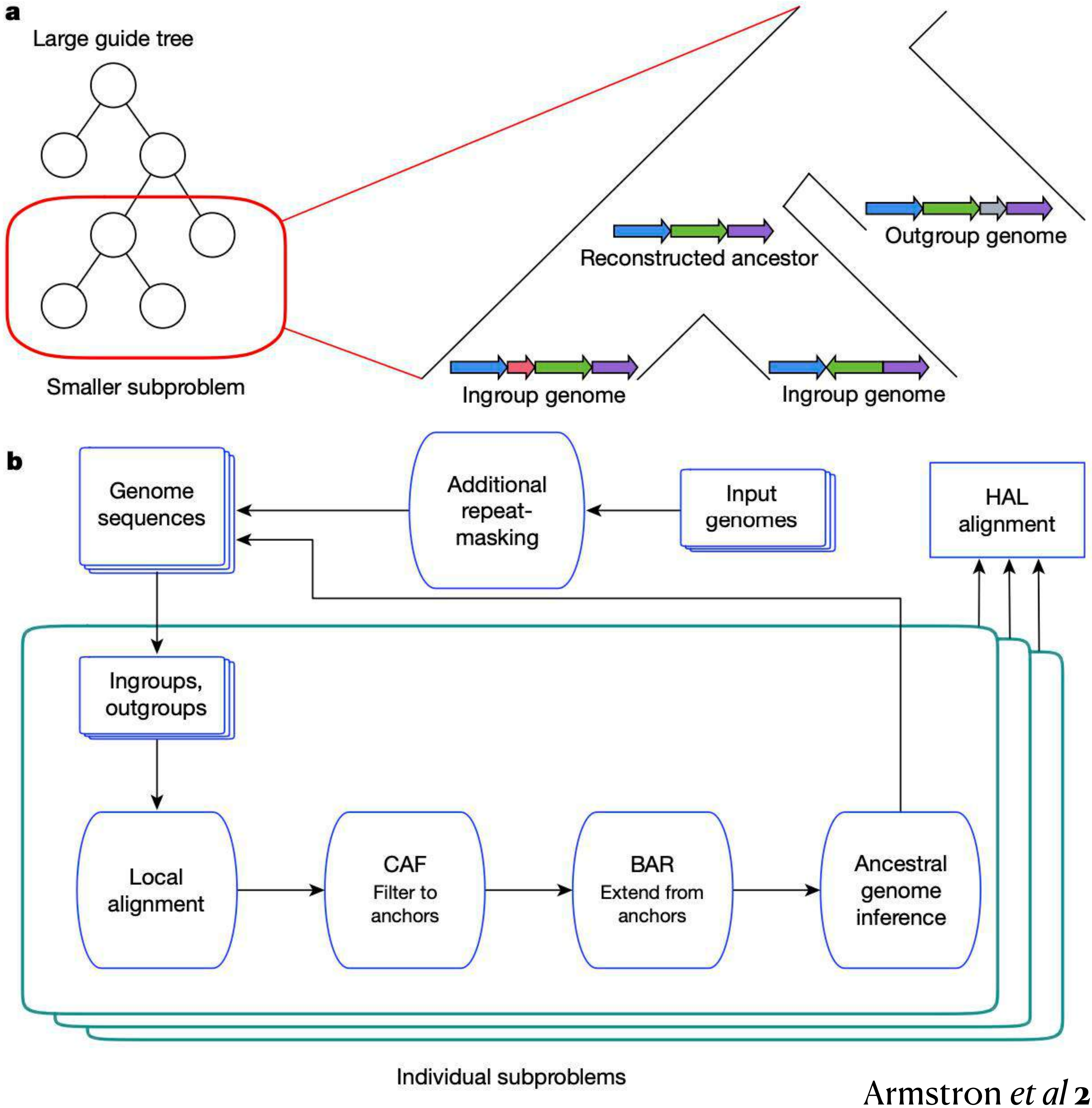
- **Multi-species map of genomic regions to a corresponding region in each other genome.**

- **Taking into account complex rearrangements and copy number changes.**

Common limitations are 'reference bias':
- Constrains a multiple alignment to only regions present in reference genome.
- Restricting the alignment to be 'single-copy', determining miss multiple-orthology relationships.

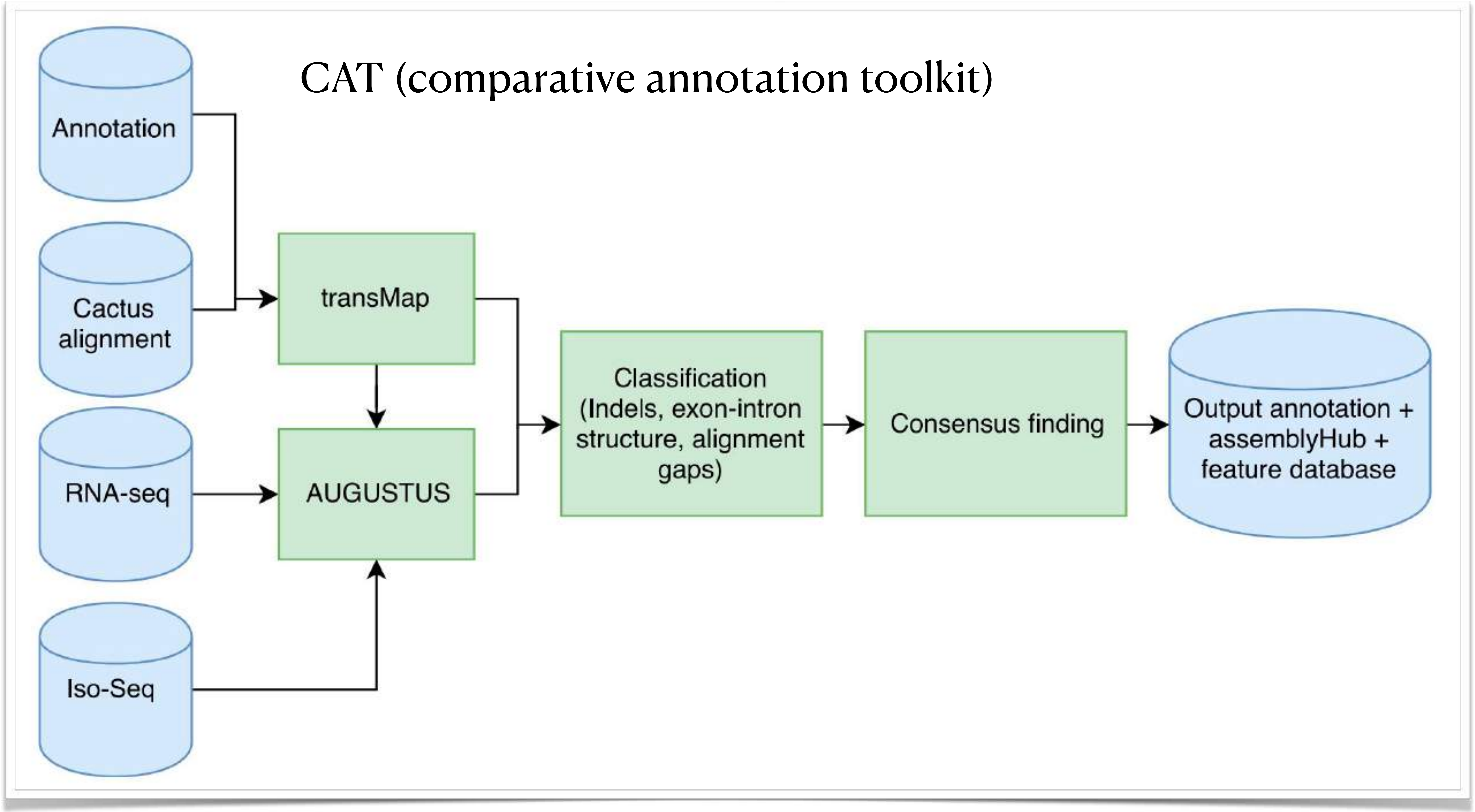
Cactus (ProgressiveCactus) is a "reference *free*" whole genome aligner.

» Whole Genome alignment »



Armstron *et al* **2020** *Nature*

» Some of the applications (annotation) »

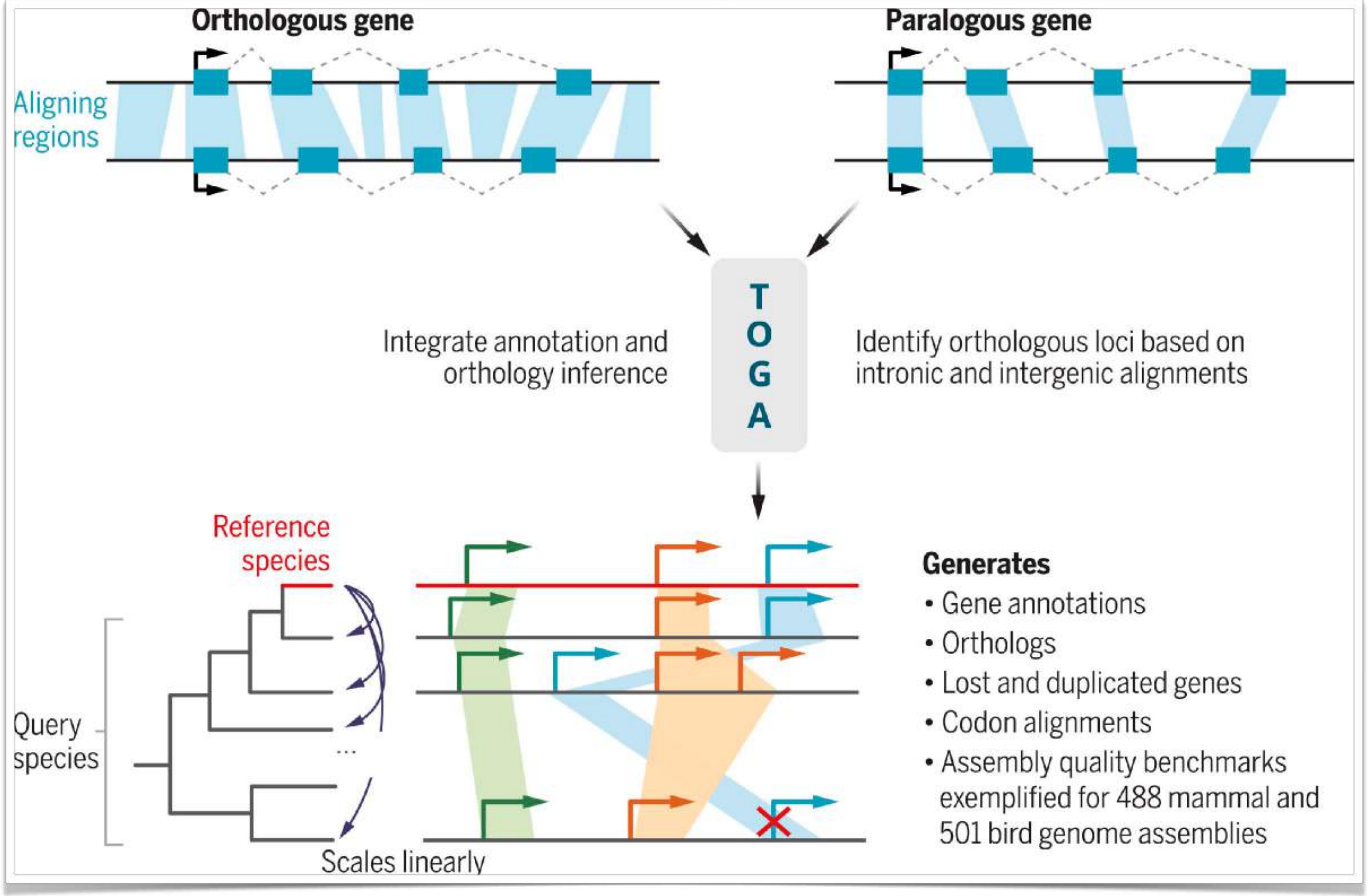


CAT (comparative annotation toolkit)

Annotation
Cactus alignment
RNA-seq
Iso-Seq
transMap
AUGUSTUS
Classification (Indels, exon-intron structure, alignment gaps)
Consensus finding
Output annotation + assemblyHub + feature database

Fiddes *et al* **2018** *Genome Research*

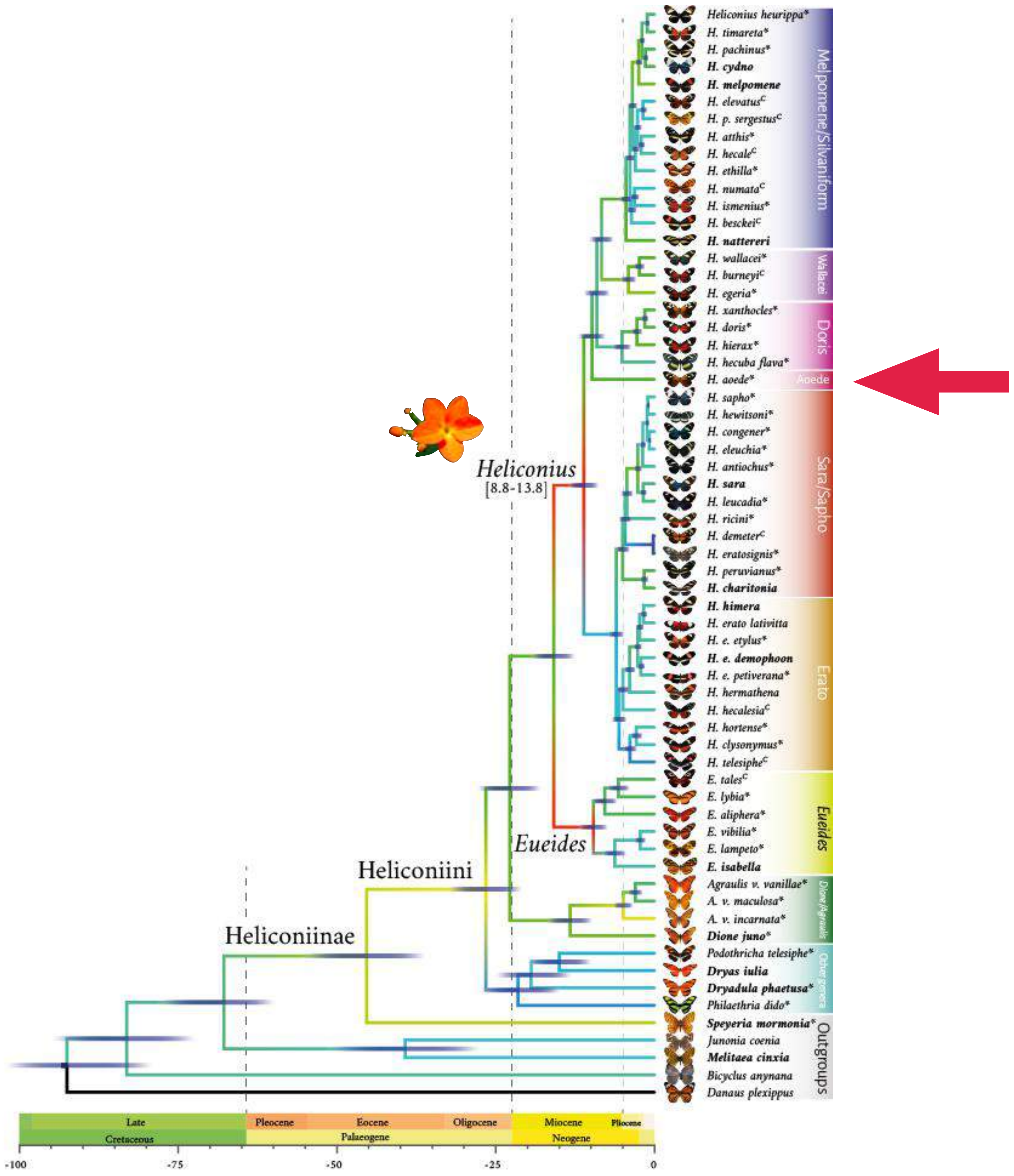
» Some of the applications (orthology inference) »


Orthologous gene
Paralogous gene
Aligning regions
Integrate annotation and orthology inference
T O G A
Identify orthologous loci based on intronic and intergenic alignments
Reference species
Query species
...
Scales linearly
Generates
• Gene annotations
• Orthologs
• Lost and duplicated genes
• Codon alignments
• Assembly quality benchmarks exemplified for 488 mammal and 501 bird genome assemblies

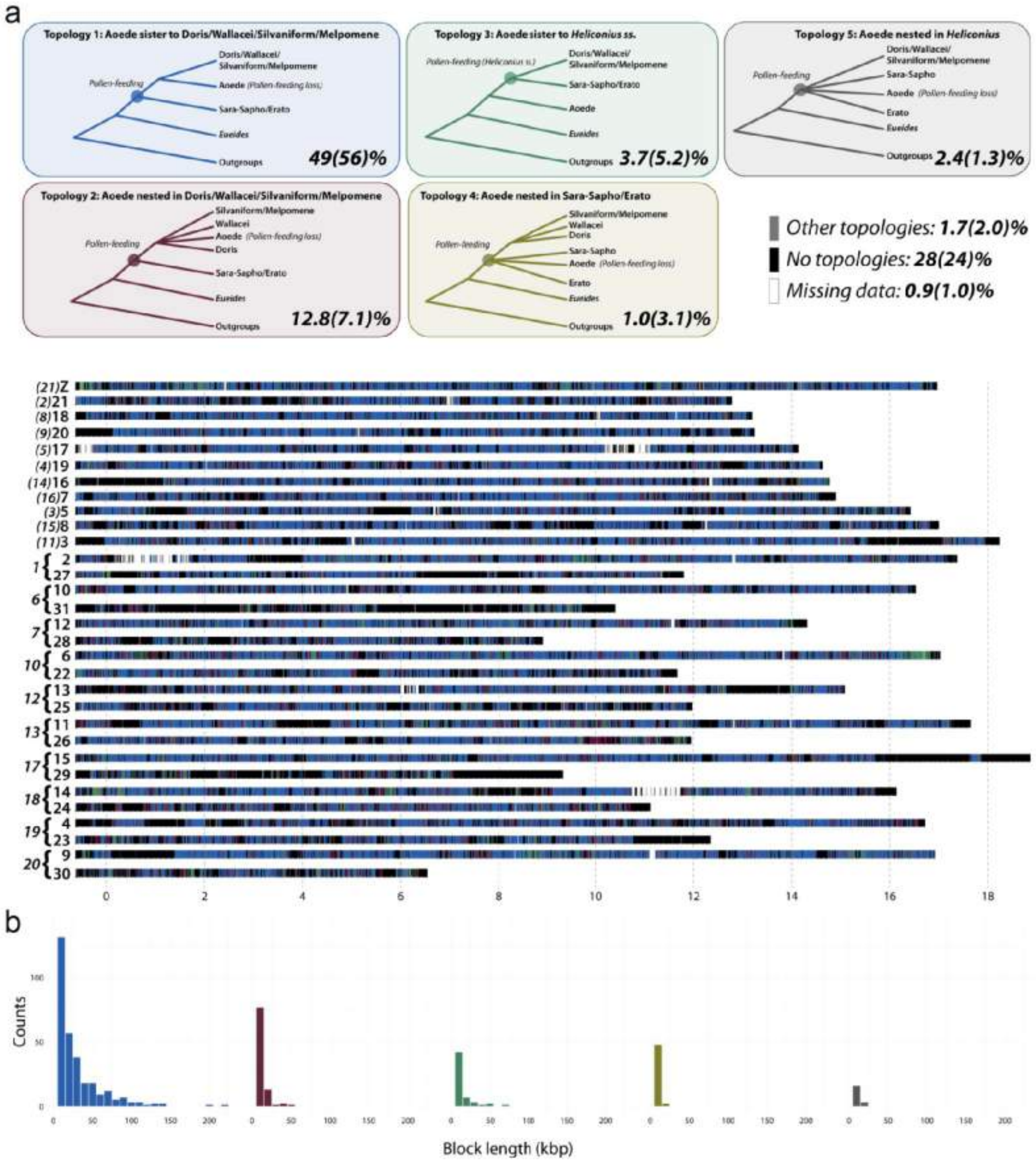
Kirilenco *et al* **2023** *Science*

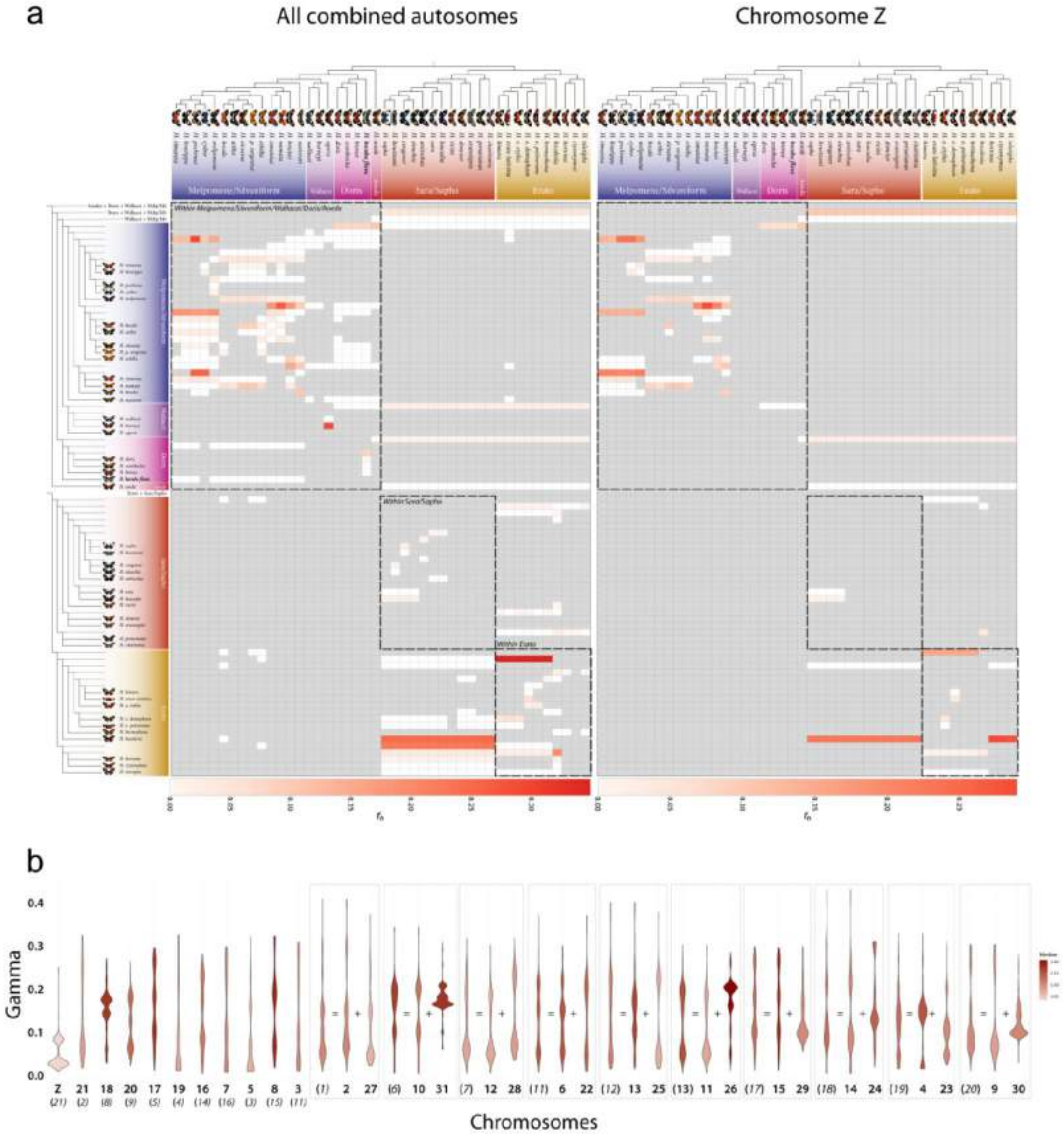
» Phylogeny / Introgression »

» Phylogeny / Introgression »



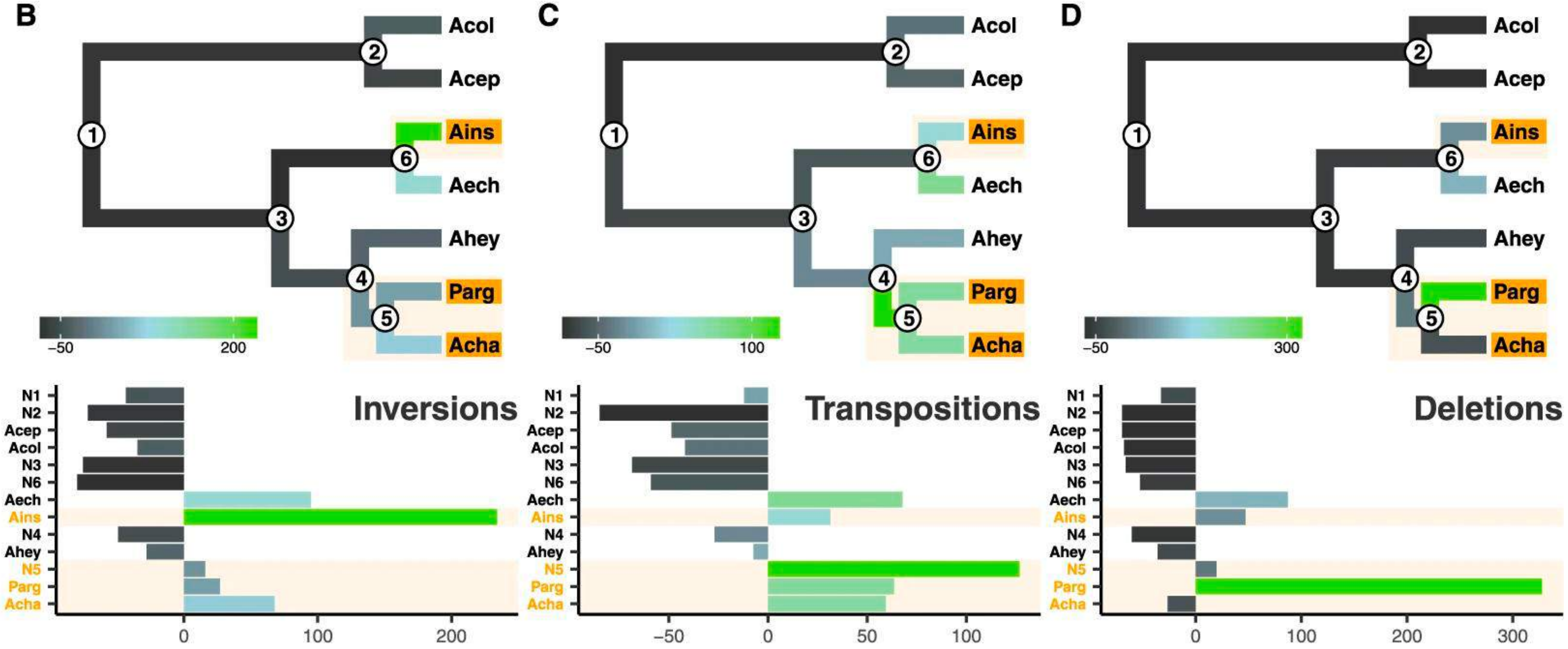
» Phylogeny / Introgression »



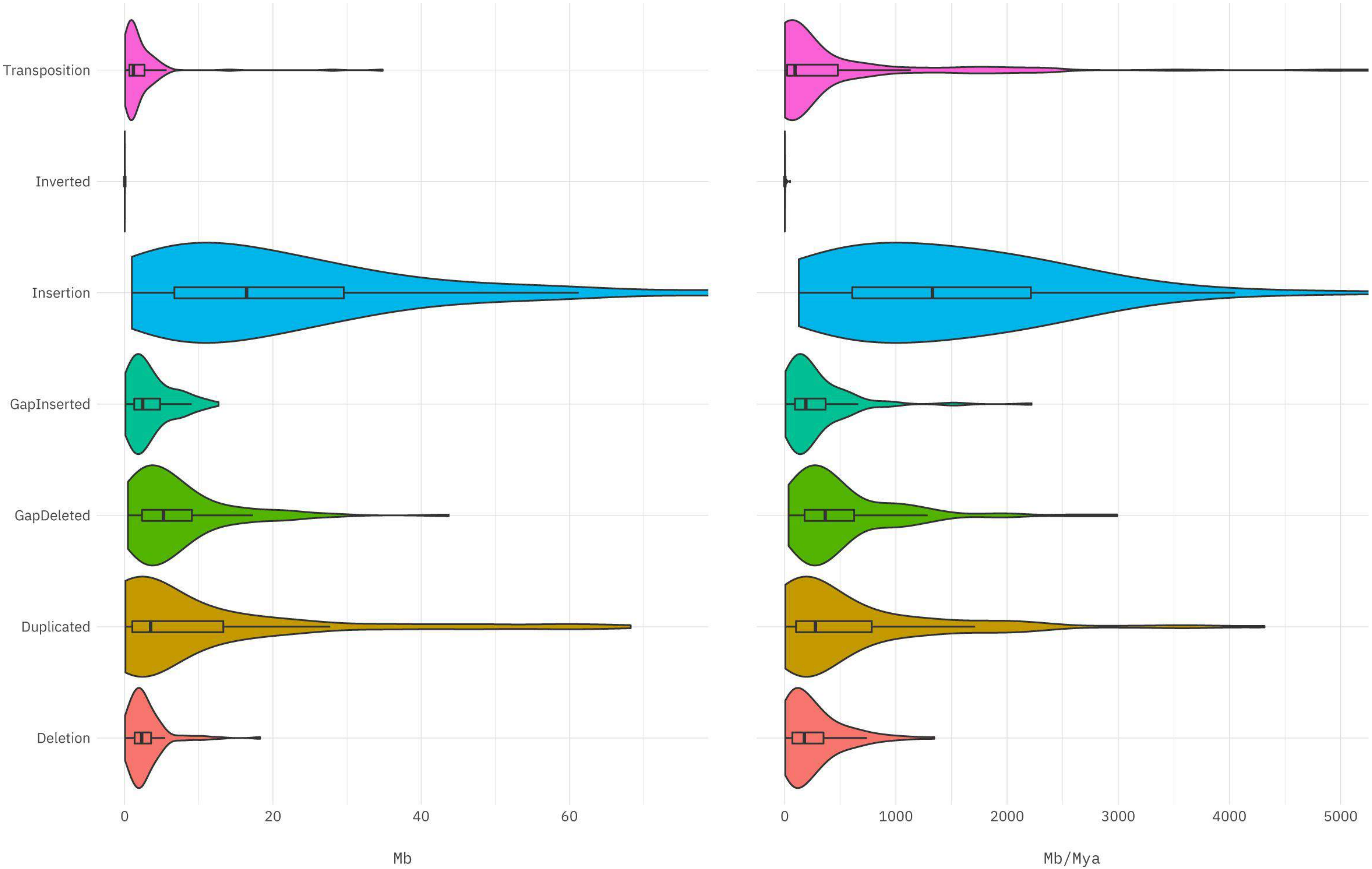


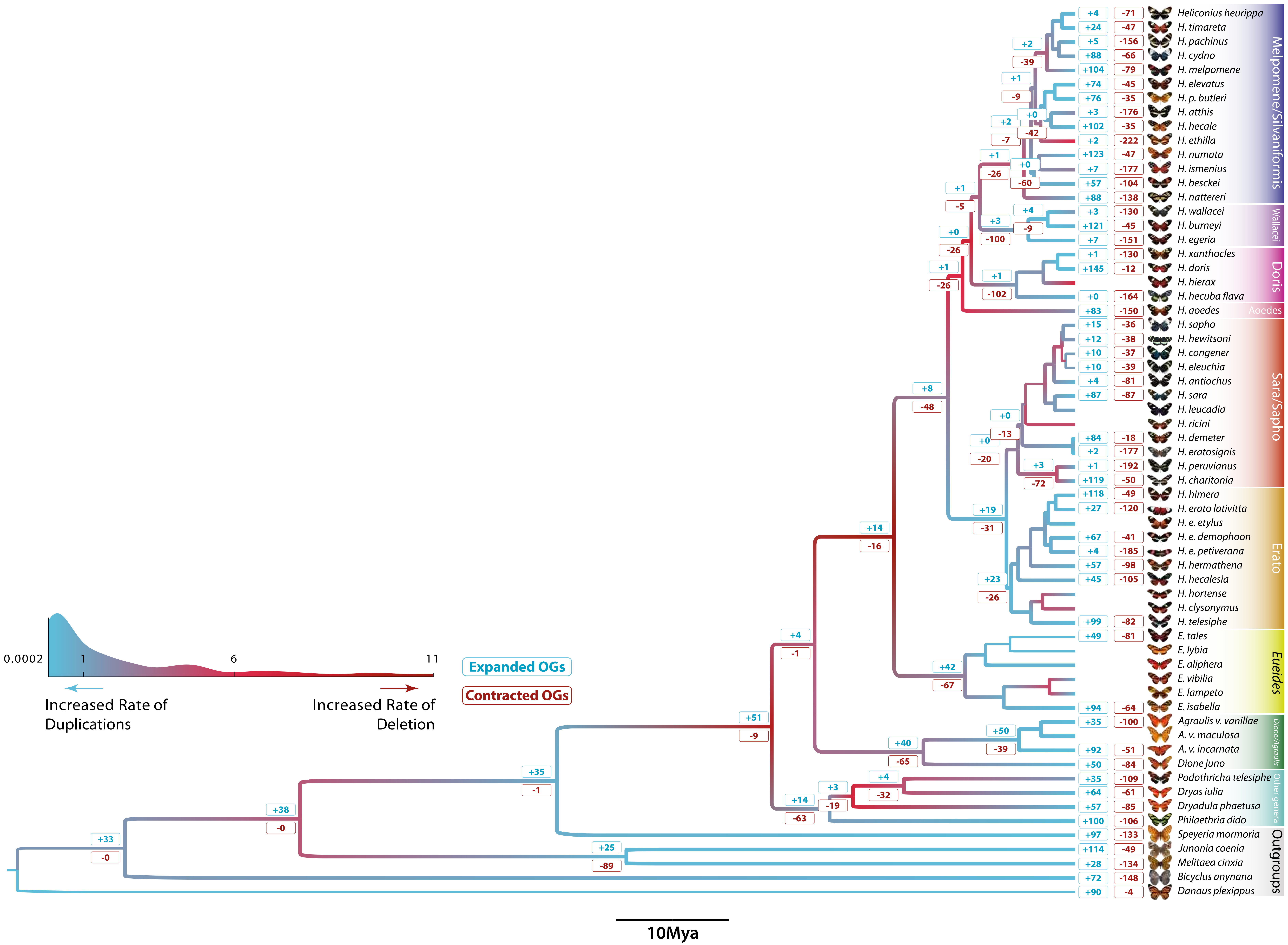
» Structural Rearrangements »

» Structural Rearrangements »



Schrader *et al* **2021** *Nature Comm.*


Transposition
Inverted
Insertion
GapInserted
GapDeleted
Duplicated
Deletion
0
20
40
60
Mb
0
1000
2000
3000
4000
5000
Mb/Mya



+4 -71 Heliconius heurippa
+24 -47 H. timareta
+5 -156 H. pachinus
+88 -66 H. cydno
+104 -79 H. melpomene
+74 -45 H. elevatus
+76 -35 H. p. butleri
+3 -176 H. atthis
+102 -35 H. hecale
+2 -222 H. ethilla
+123 -47 H. numata
+7 -177 H. ismenius
+57 -104 H. besckei
+88 -138 H. nattereri
+3 -130 H. wallacei
+121 -45 H. burneyi
+7 -151 H. egeria
+1 -130 H. xanthocles
+145 -12 H. doris
H. hierax
+0 -164 H. hecuba flava
+83 -150 H. aoedes
+15 -36 H. sapho
+12 -38 H. hewitsoni
+10 -37 H. congener
+10 -39 H. eleuchia
+4 -81 H. antiochus
+87 -87 H. sara
H. leucadia
H. ricini
+84 -18 H. demeter
+2 -177 H. eratosignis
+1 -192 H. peruvianus
+119 -50 H. charitonia
+118 -49 H. himera
+27 -120 H. erato lativitta
H. e. etylus
+67 -41 H. e. demophoon
+4 -185 H. e. petiverana
+57 -98 H. hermathena
+45 -105 H. hecalesia
H. hortense
H. clysonymus
+99 -82 H. telesiphe
+49 -81 E. tales
E. lybia
E. aliphera
E. vibilia
E. lampeto
+94 -64 E. isabella
+35 -100 Agraulis v. vanillae
A. v. maculosa
+92 -51 A. v. incarnata
+50 -84 Dione juno
+35 -109 Podothricha telesiphe
+64 -61 Dryas iulia
+57 -85 Dryadula phaetusa
+100 -106 Philaethria dido
+97 -133 Speyeria mormoria
+114 -49 Junonia coenia
+28 -134 Melitaea cinxia
+72 -148 Bicyclus anynana
+90 -4 Danaus plexippus
Melpomene/Silvaniformis
Wallacei
Doris
Aoedes
Sara/Sapho
Erato
Eueides
Dione/Agraulis
Other genera
Outgroups
+2
-39
+1
-9
+0
-42
+2
-7
+1
-26
+0
-60
+1
-5
+4
-9
+3
-100
+0
-26
+1
-102
+1
-26
+8
-48
+0
-13
+0
-20
+3
-72
+19
-31
+23
-26
+14
-16
+42
-67
+4
-1
+50
-39
+40
-65
+51
-9
+4
-32
+3
-19
+14
-63
+35
-1
+38
-0
+25
-89
+33
-0
0.0002
1
6
11
Increased Rate of Duplications
Increased Rate of Deletion
Expanded OGs
Contracted OGs
10Mya

» Conserved non-coding Elements (CNEs) »

» Conserved non-coding Elements (CNEs) »

> A class of *non-protein-coding* genomic sequences with elevated degree of conservation.

> CNEs are non-randomly distributed, clustering in the vicinity of genes with regulatory roles.

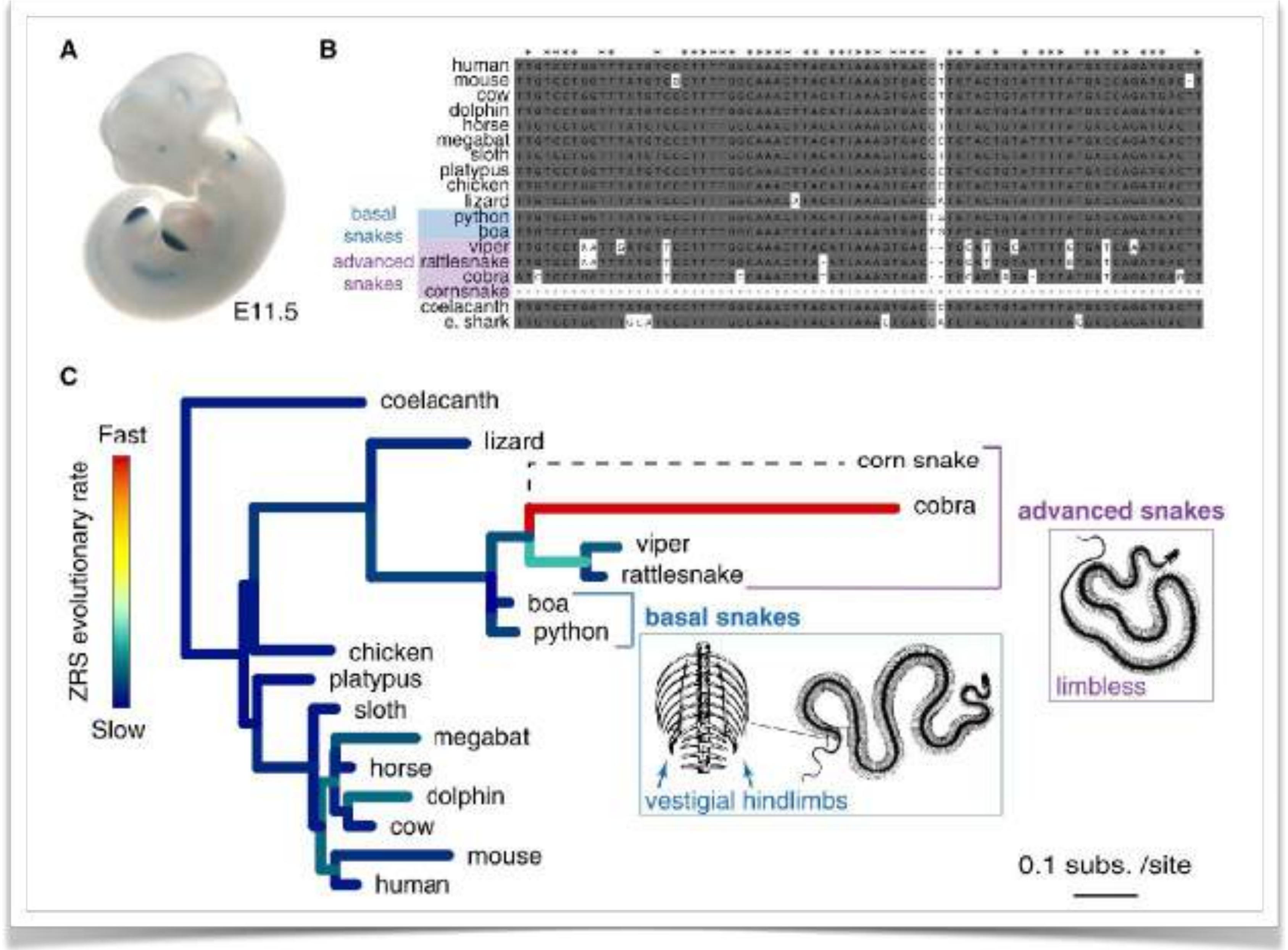
> Organised into functional ensembles (regulatory blocks), which coordinate the expression of shared target genes.

> The disruption of these elements contribute to diseases linked with development, and cancer.

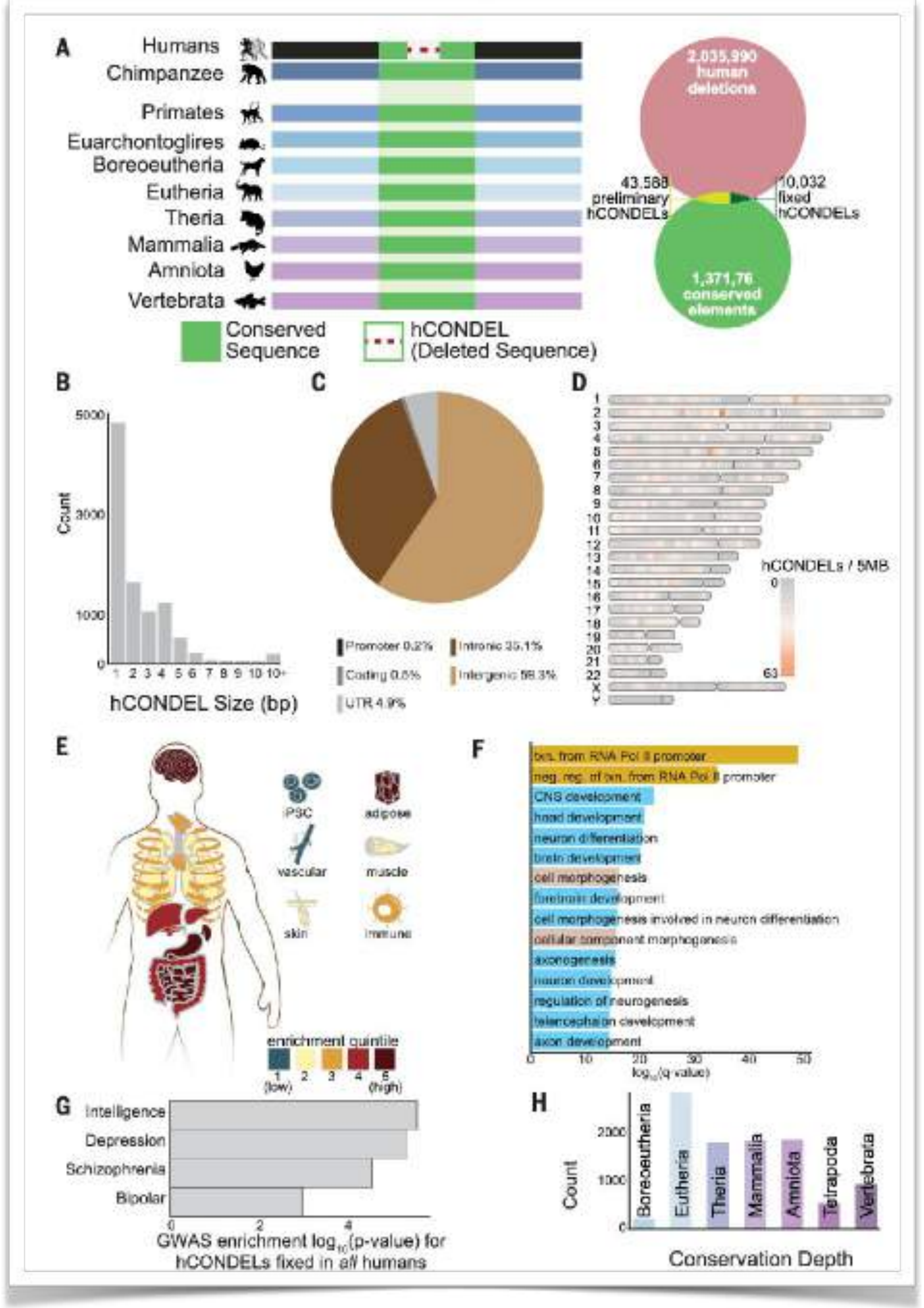
multi-region
chr2:24,765,459-25,564,358 798,900 bp.
gene, chromosome range, search terms, help pages, see e
go
examples
chr2 (p23.3)
Scale
chr2:
200 kb
hg38
24,800,000
24,850,000
24,900,000
24,950,000
25,000,000
25,050,000
25,100,000
25,150,000
25,200,000
25,250,000
25,300,000
25,350,000
25,400,000
25,450,000
25,500,000
25,550,000
Reference Assembly Fix Patch Sequence Alignments
Reference Assembly Alternate Haplotype Sequence Alignments
GENCODE V43 (70 items filtered out)
NCOA1
ADCY3
CENPO
ENSG00000271936
PTRHD1
SNORD14
DNAJC27-AS1
Y_RNA
ENSG00000286645
DNAJC27
EFR3B
RN7SL856P
LINC01381
POMC
DNMT3A
MIR1301
ARNILA
DTNB-AS1
DTNB
ENSG00000289567
RefSeq Curated
RefSeq genes from NCBI
ENCODE cCREs
ENCODE Candidate Cis-Regulatory Elements (cCREs) combined from all cell types
100 vertebrates Basewise Conservation by PhyloP
Cons 100 Verts
4
-0.5
7.5
470 phyloP
-4.5
470 mammals Basewise Conservation by PhyloP
Multiz Alignment & Conservation (470 mammals)
Multiz 470-way

Candidate Cis-Regulatory Elements (cCREs)

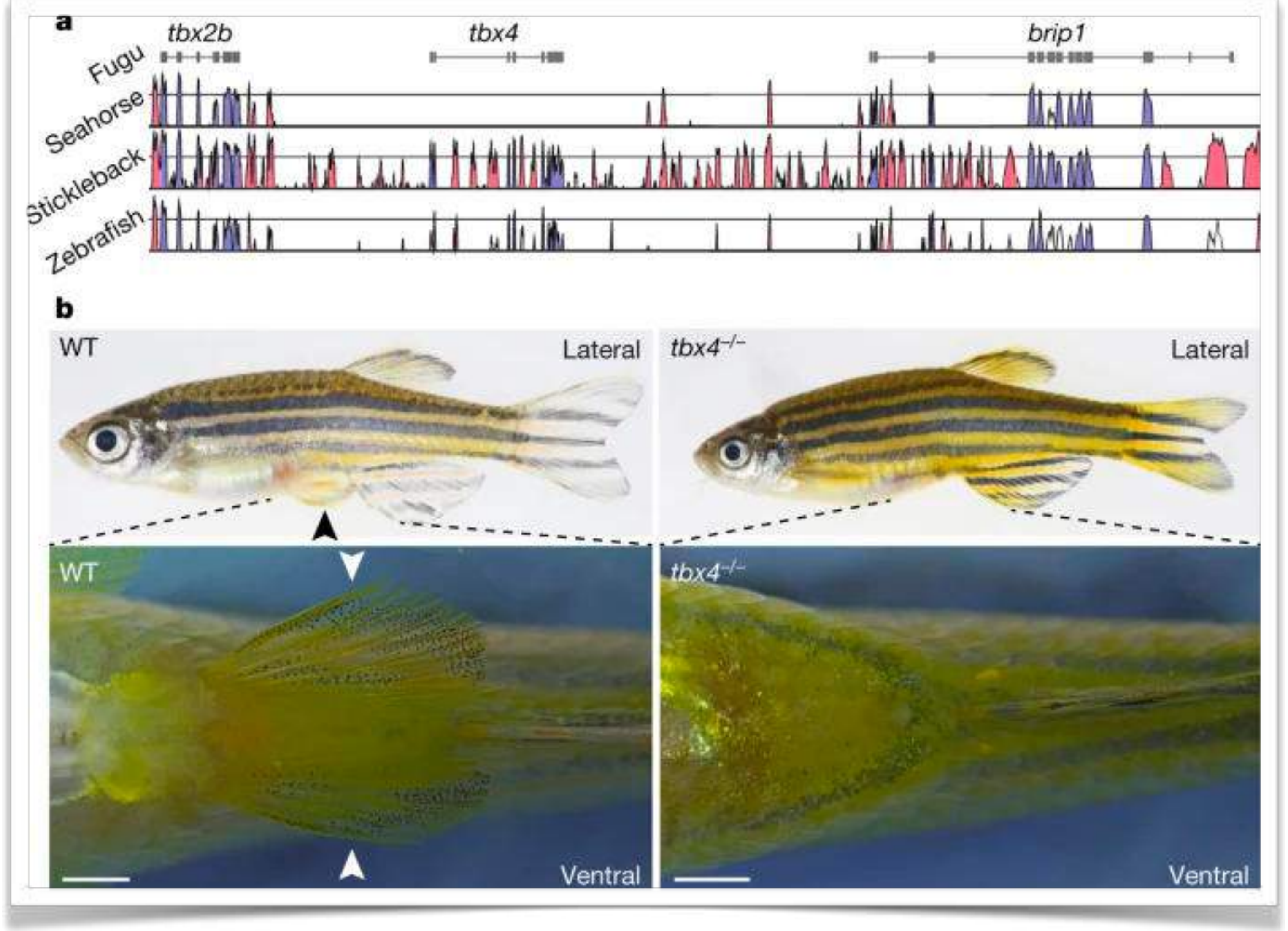
» Conserved non-coding Elements (CNEs) »



Kvon *et al* **2016** *Cell*



Xue *et al* **2023** *Science*



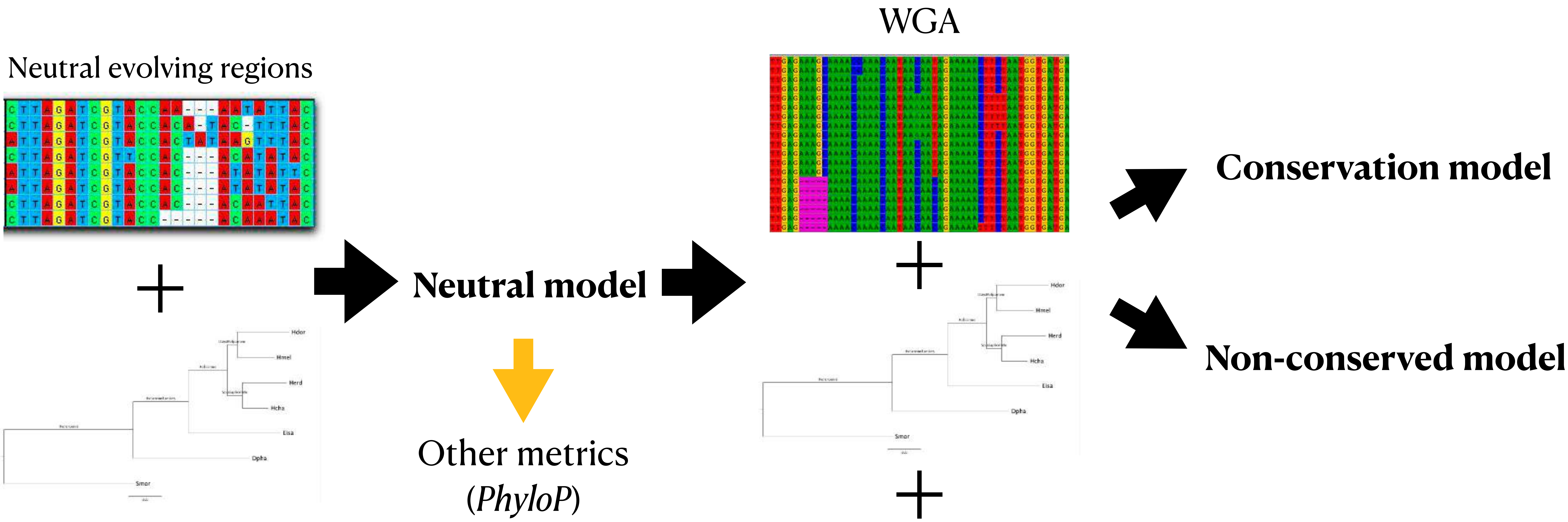
Lin *et al* **2016** *Nature*

» How to identify these regions »

1. **A Model of DNA sequence evolution**

2. **Phylogenetic tree**

» How to identify these regions »



Neutral evolving regions

+

Neutral model

Other metrics

(*PhyloP*)

WGA

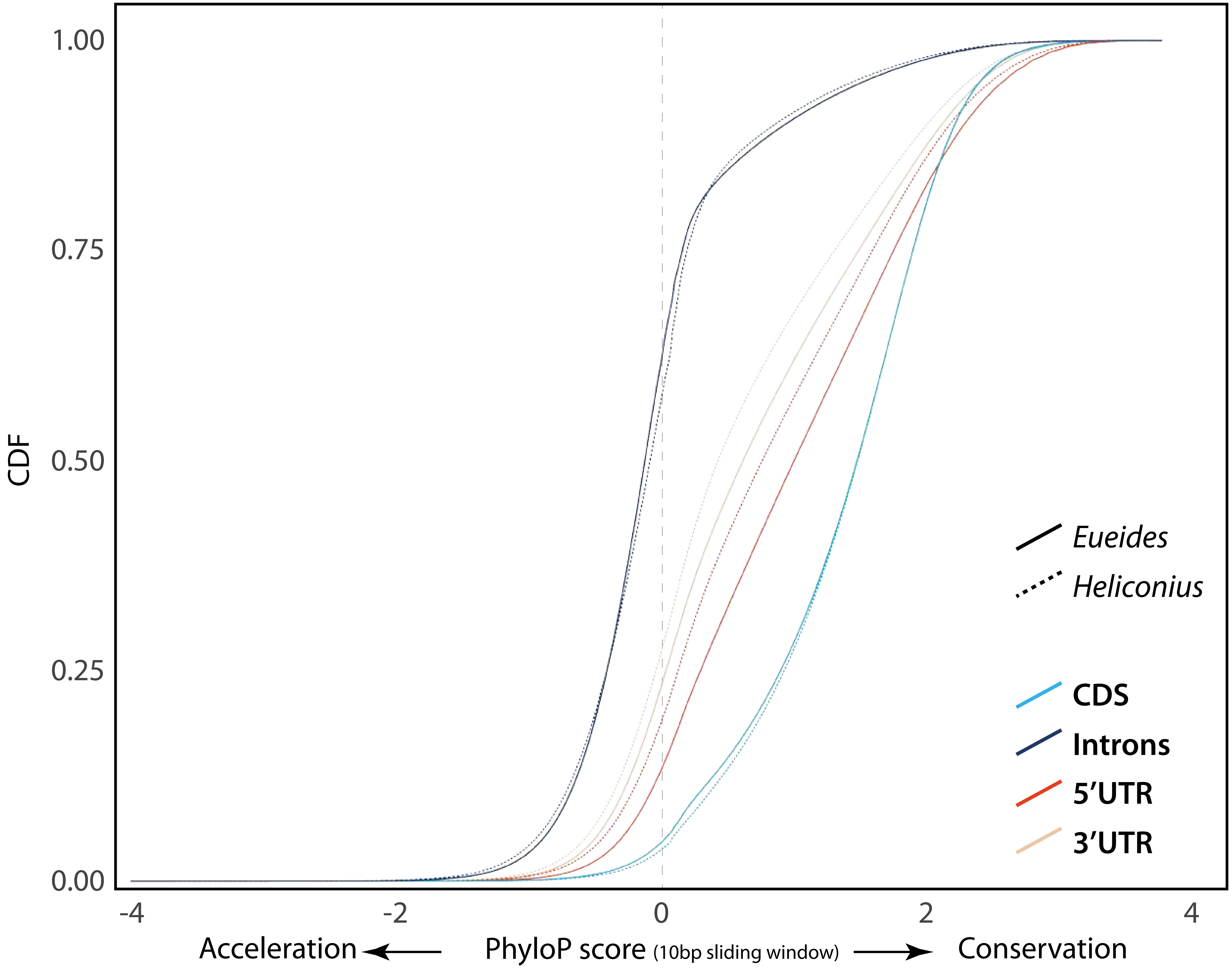
+

+

neutral model

Conservation model

Non-conserved model


1.00
0.75
0.50
0.25
0.00
CDF
-4
-2
0
2
4
Acceleration
PhyloP score (10bp sliding window)
Conservation
Eueides
Heliconius
CDS
Introns
5'UTR
3'UTR

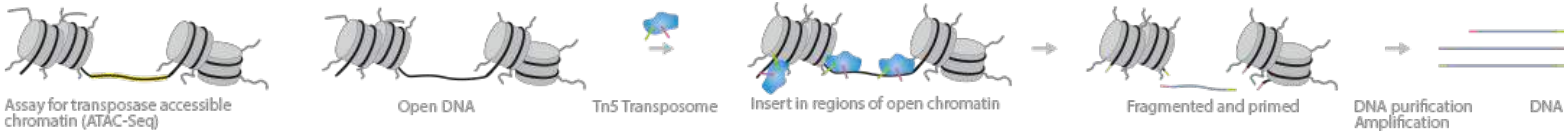
» Let’s start the exercise »

Some tools you probably going to need:

- Cactus
- halSummarizeMutations
- halAlignmentDepth
- wigToBigWig
- samtools
- .... and more ...

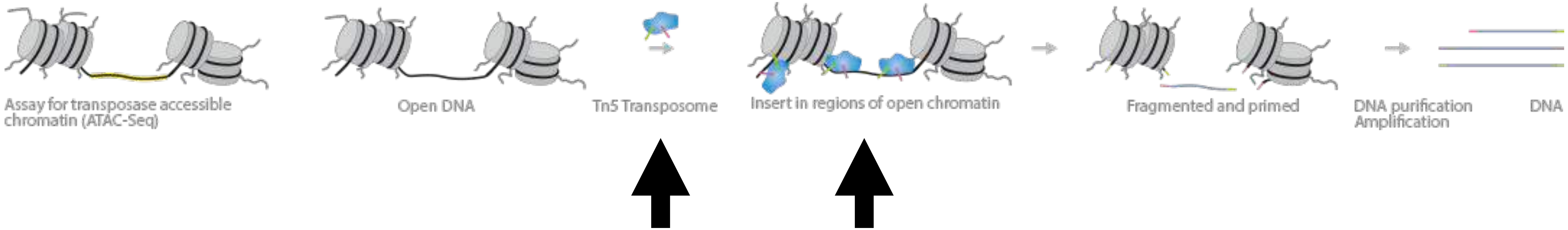
Don’t forget IGV!

» ATAC-seq (**A**ssay for **T**ransposase-**A**ccessible **C**hromatin using **seq**uencing) »


Assay for transposase accessible chromatin (ATAC-Seq)
Open DNA
Tn5 Transposome
Insert in regions of open chromatin
Fragmented and primed
DNA purification Amplification
DNA

» ATAC-seq (**A**ssay for **T**ransposase-**A**ccessible **C**hromatin using **seq**uencing) »

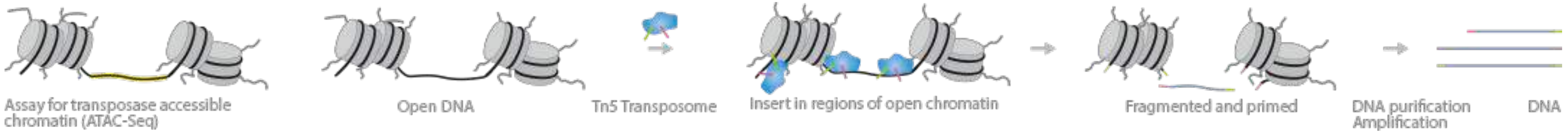
» ATAC-seq (**A**ssay for **T**ransposase-**A**ccessible Chromatin using **seq**uencing) »


Assay for transposase accessible chromatin (ATAC-Seq)
Open DNA
Tn5 Transposome
Insert in regions of open chromatin
Fragmented and primed
DNA purification Amplification
DNA

Tagmentation

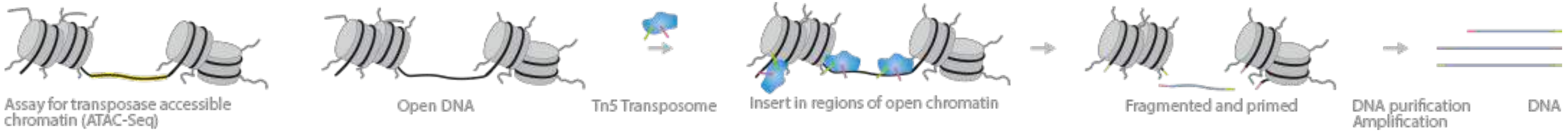
The activity Tn5 (hyperactive) Transposase that inserts sequencing adapters into open regions of the genome and cleaves

» ATAC-seq (**A**ssay for **T**ransposase-**A**ccessible Chromatin using **seq**uencing) »


Assay for transposase accessible chromatin (ATAC-Seq)
Open DNA
Tn5 Transposome
Insert in regions of open chromatin
Fragmented and primed
DNA purification Amplification
DNA

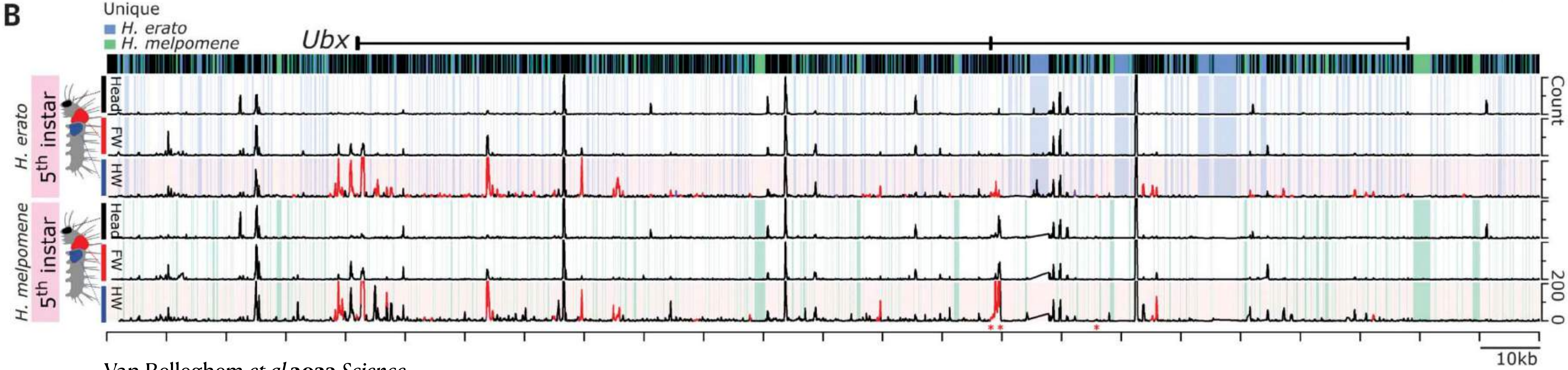
<u>Purification</u>

» ATAC-seq (**A**ssay for **T**ransposase-**A**ccessible **C**hromatin using **seq**uencing) »


Assay for transposase accessible chromatin (ATAC-Seq)
Open DNA
Tn5 Transposome
Insert in regions of open chromatin
Fragmented and primed
DNA purification Amplification
DNA

<u>Short-Reads Sequencing</u>

» ATAC-seq (**A**ssay for **T**ransposase-**A**ccessible **C**hromatin using **seq**uencing) »



Van Belleghem *et al* **2023** *Science*

» Let’s continue »

» Notes »

Some UNIX tools you probably going to need:

- “>” Redirect

- “|” Pipe

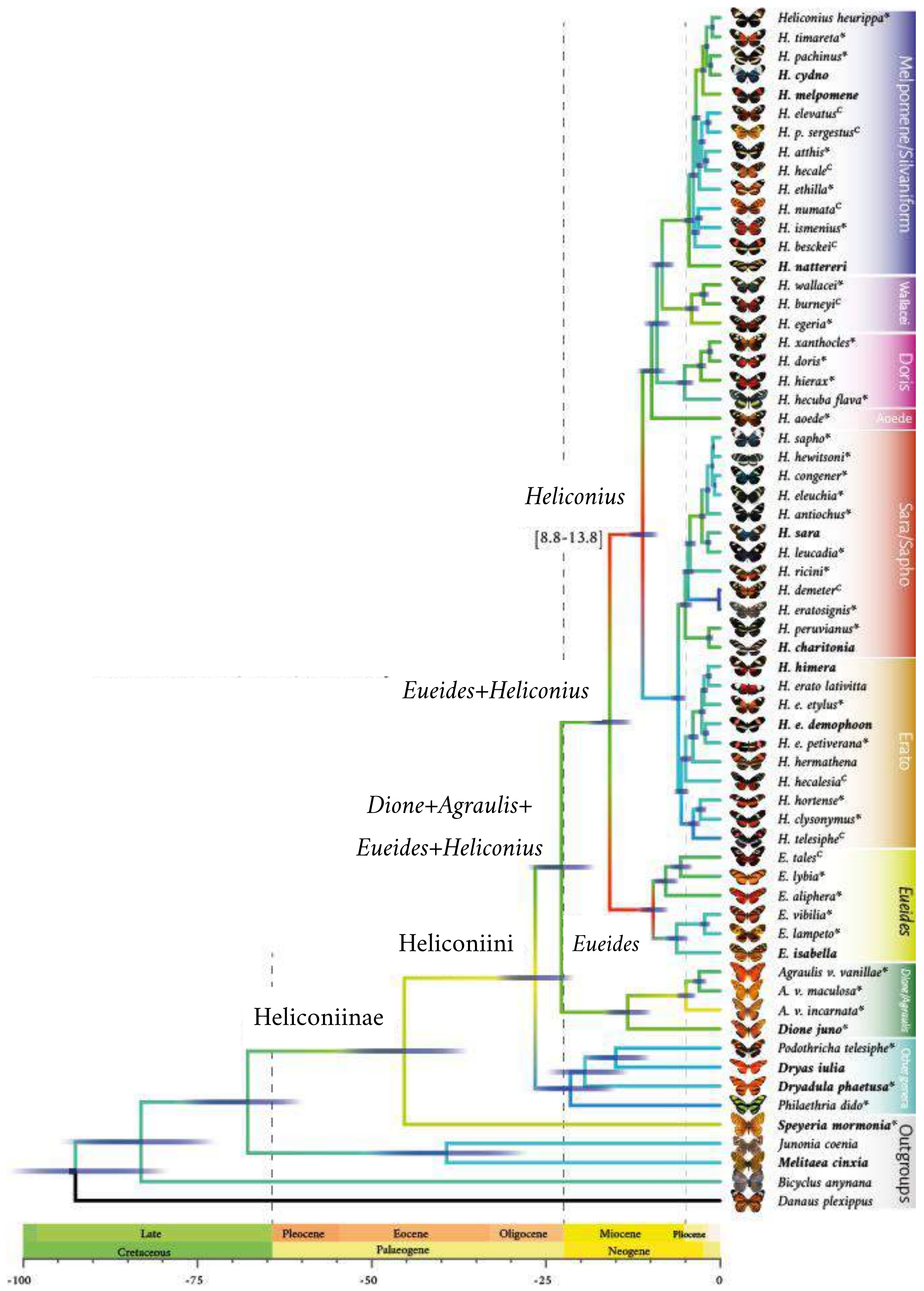
- “sed” a stream editor

For the extra task:

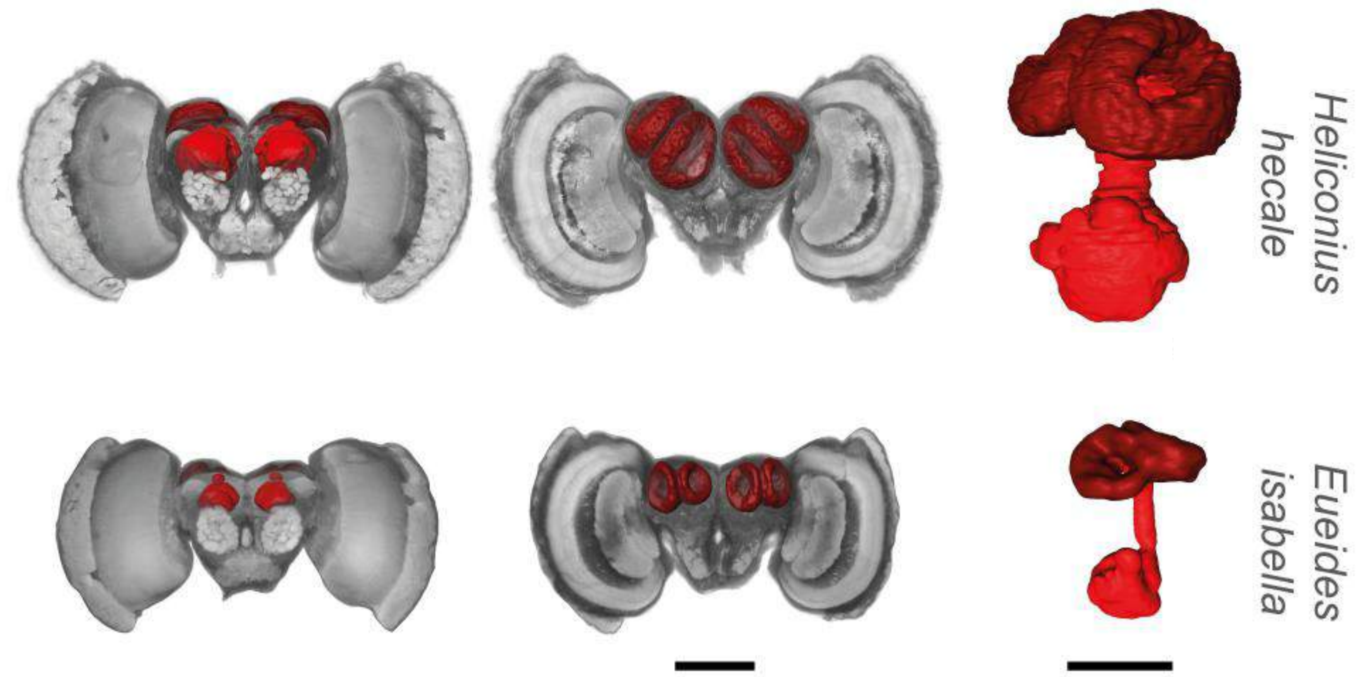
- “seq” unix command to generate a sequence of numbers

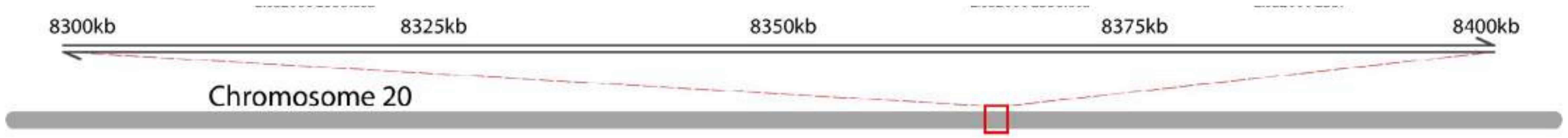
- For loop

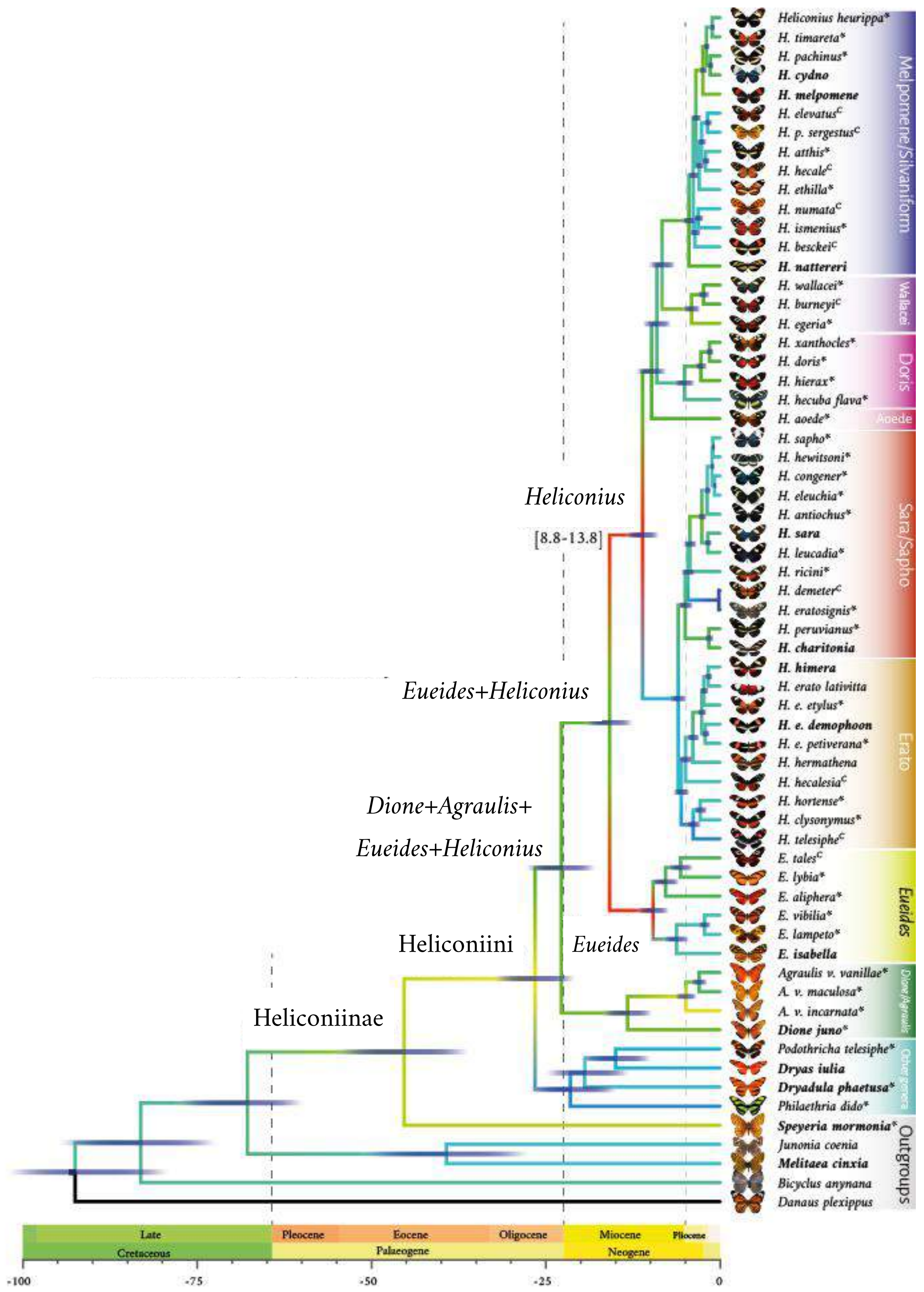
» Test for Acceleration »



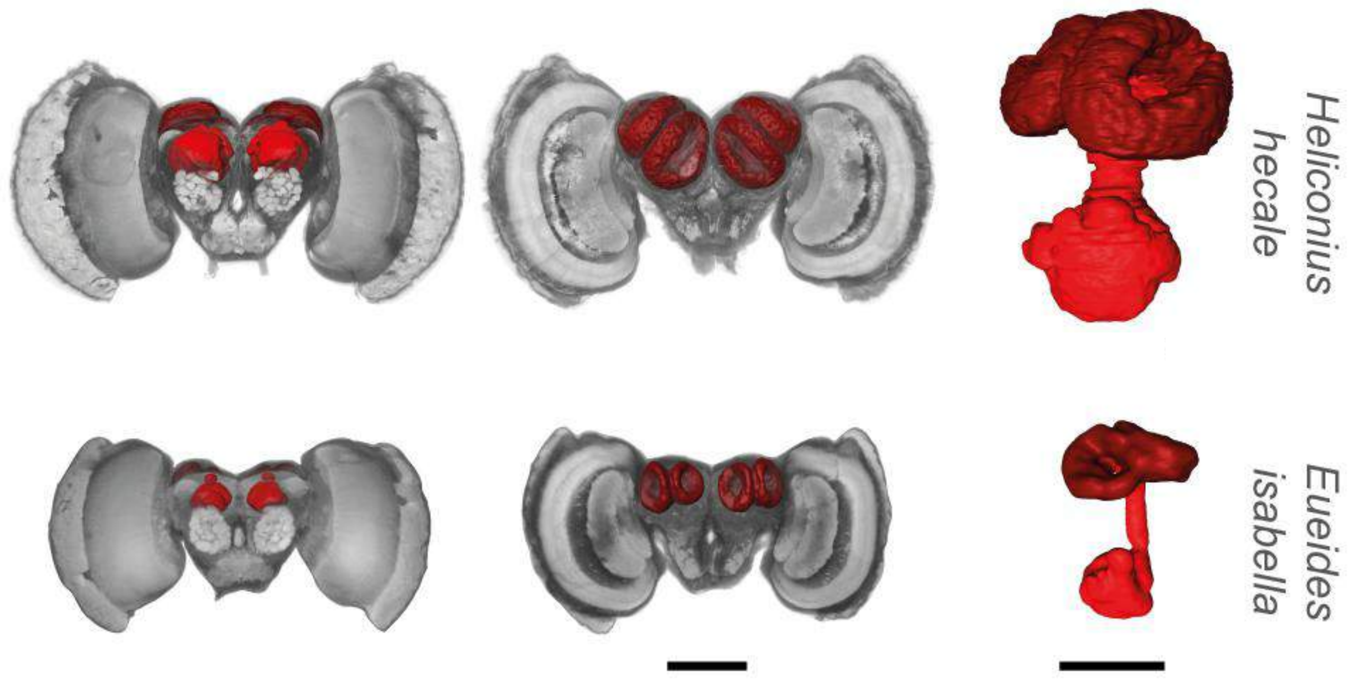
» *Mushroom body expansion (Brain)*

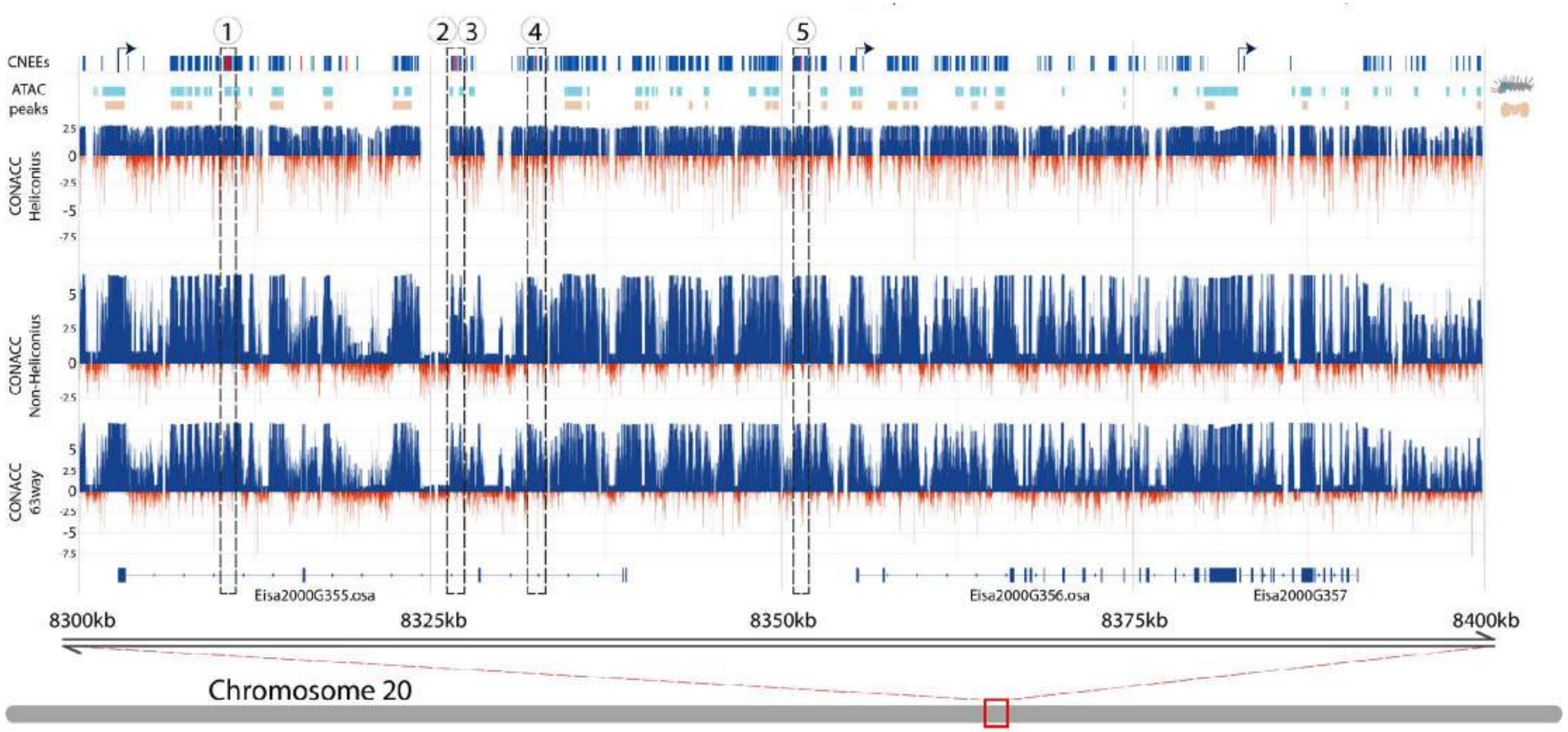


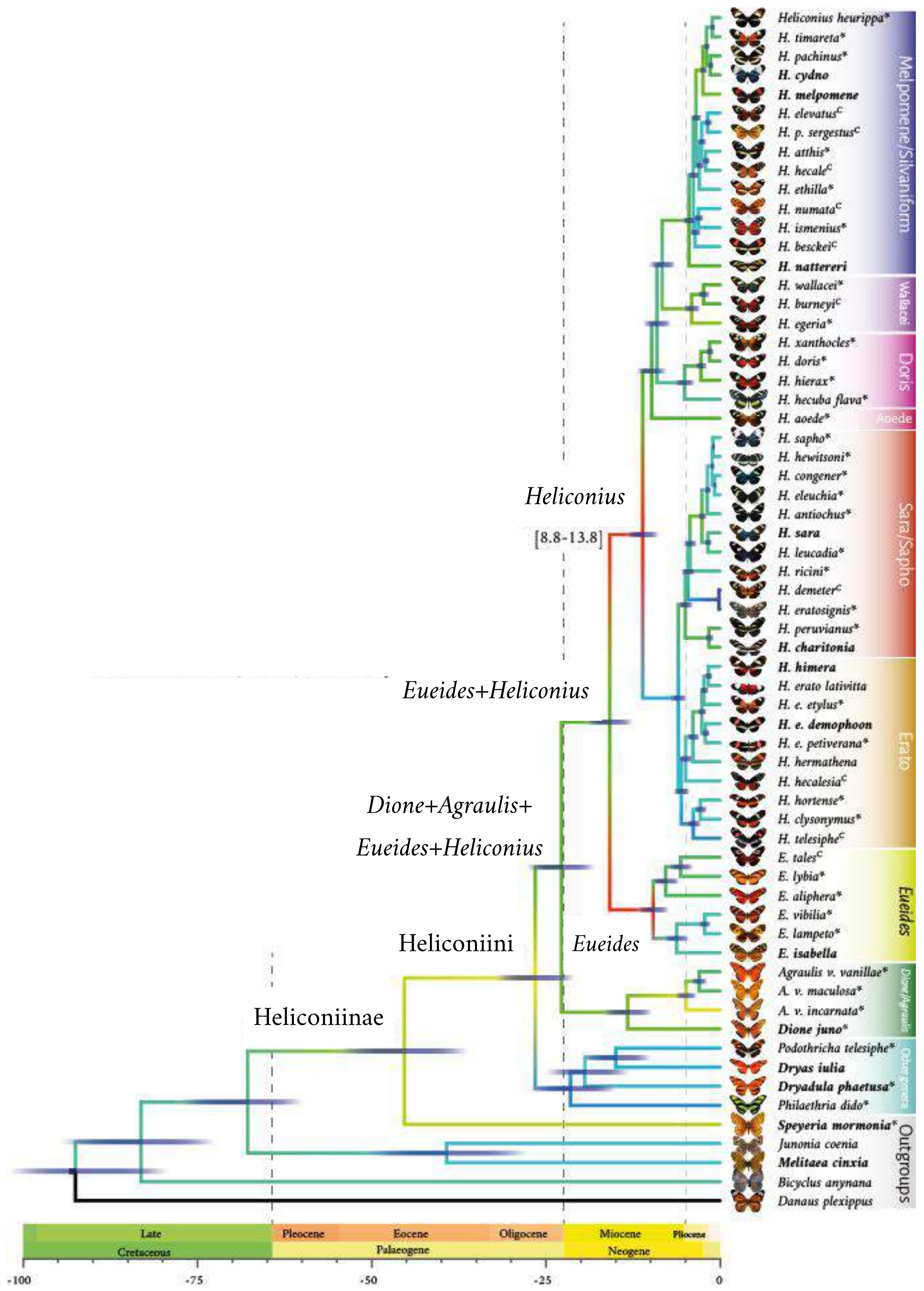




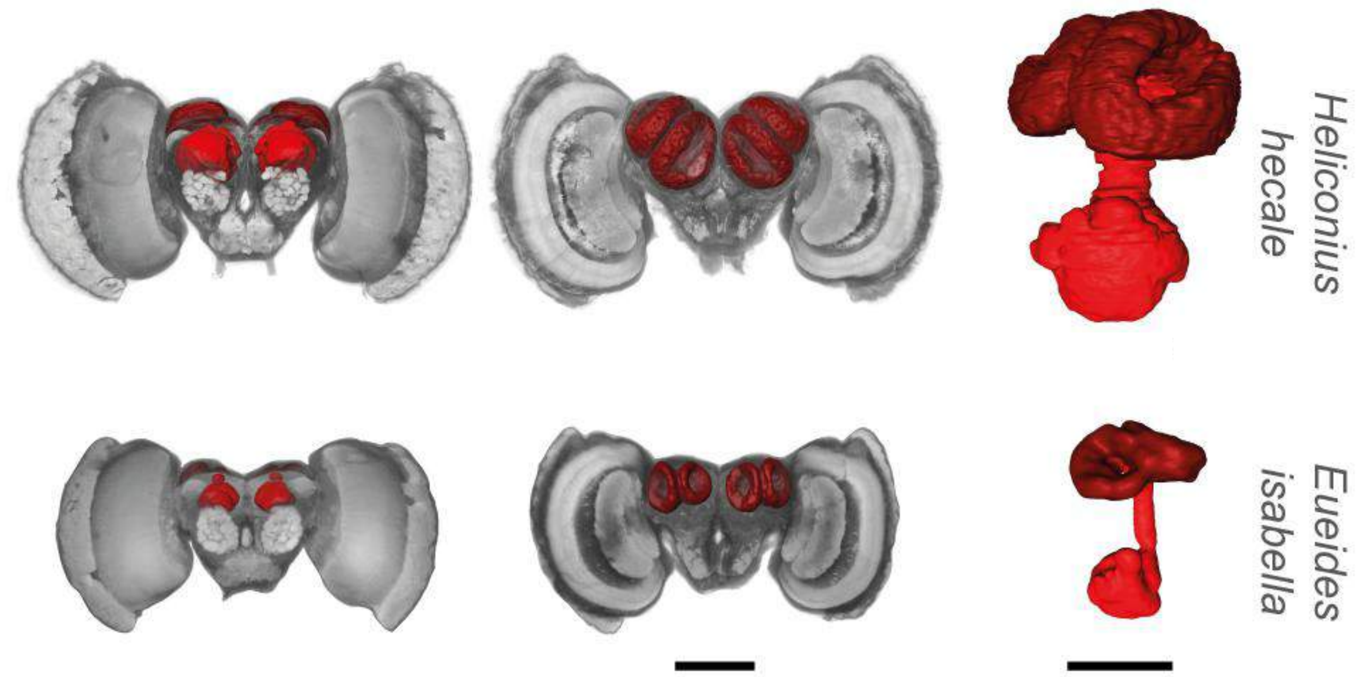
» Mushroom body expansion (Brain)

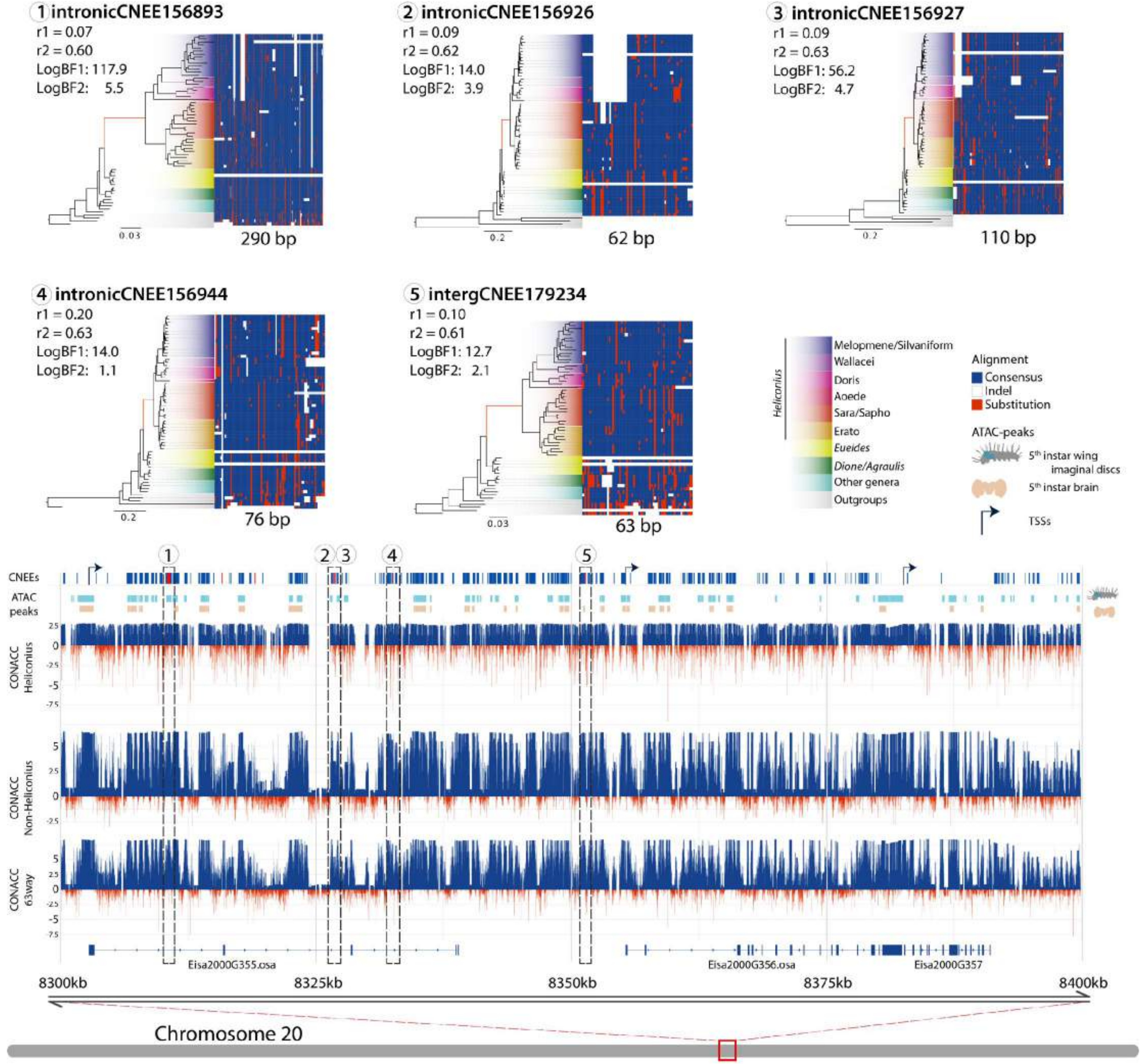




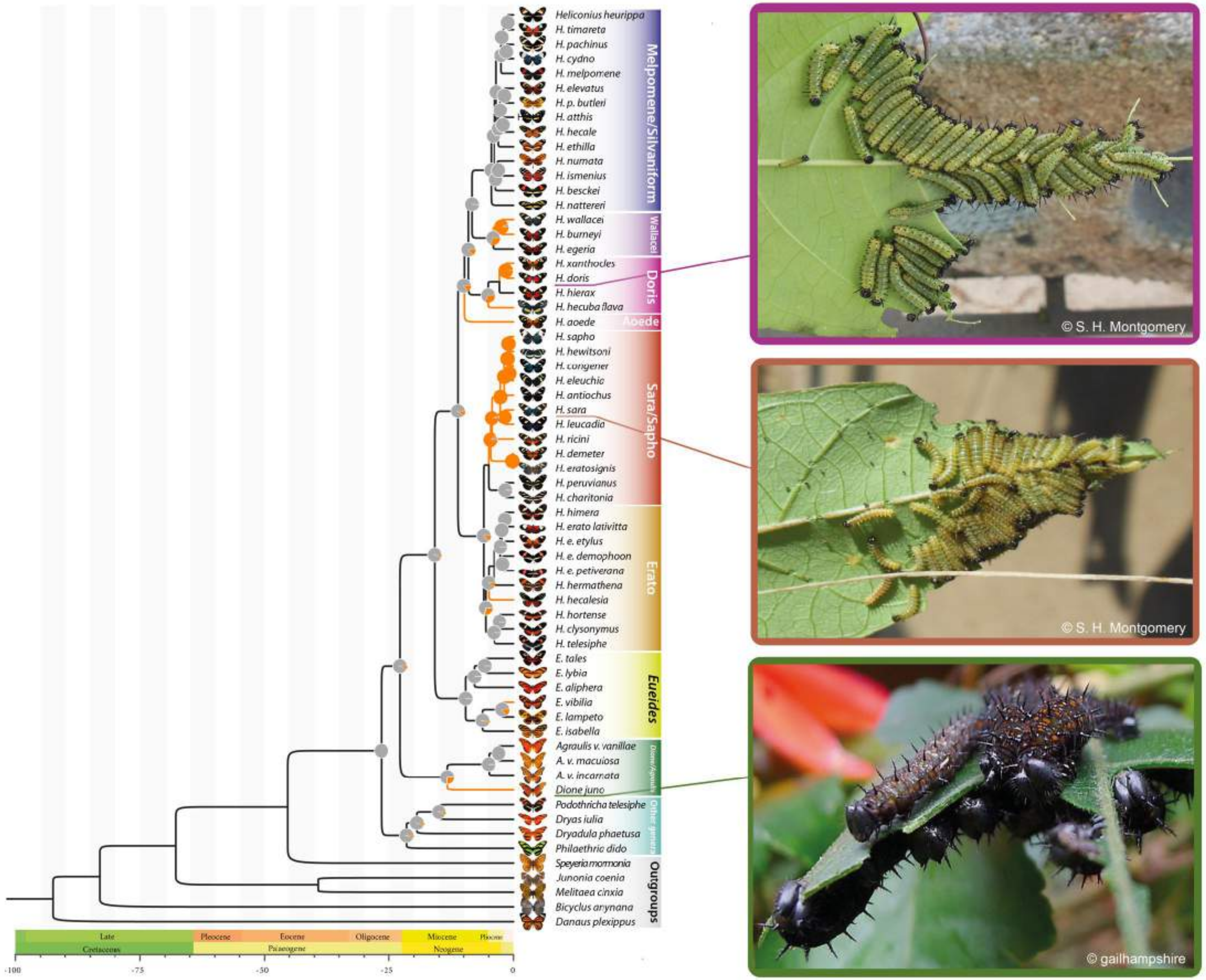


» Mushroom body expansion (Brain)

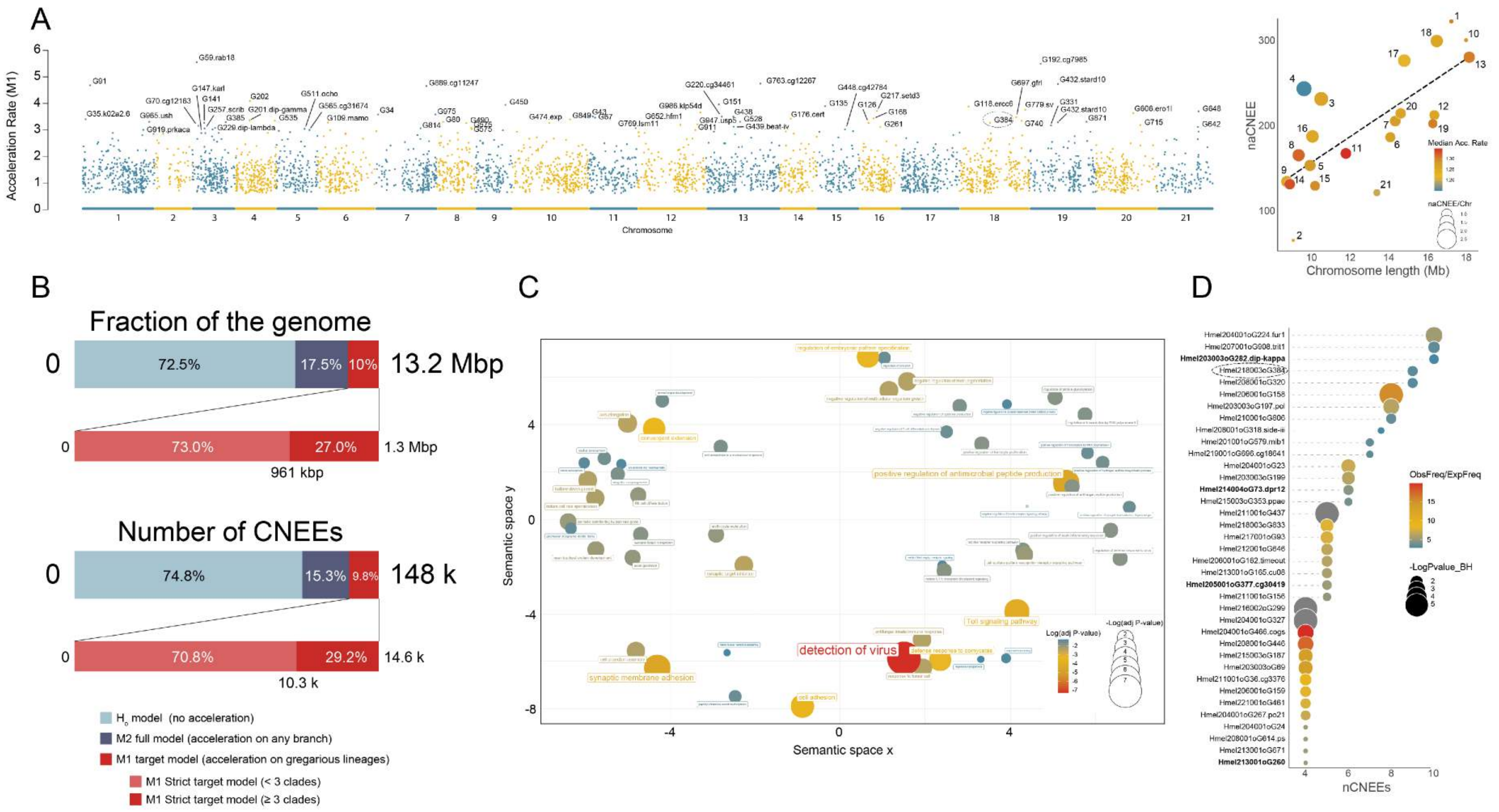




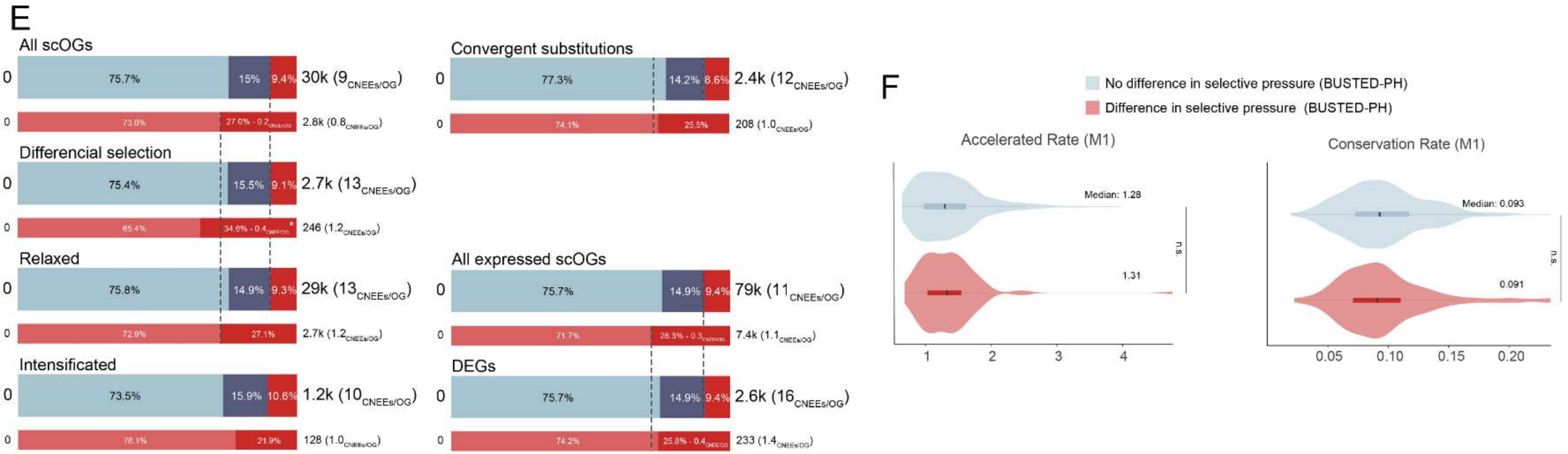
» Test for Convergence »



» Test for Convergence »



» Test for Convergence »


E
All scOGs
30k ($9_{CNEEs/OG}$)
27.0% - $0.2_{CNEEs/OG}$
2.8k ($0.8_{CNEEs/OG}$)
Differencial selection
2.7k ($13_{CNEEs/OG}$)
34.6% - $0.4_{CNEE/OG}$ *
246 ($1.2_{CNEEs/OG}$)
Relaxed
29k ($13_{CNEEs/OG}$)
2.7k ($1.2_{CNEEs/OG}$)
Intensificated
1.2k ($10_{CNEEs/OG}$)
128 ($1.0_{CNEEs/OG}$)
Convergent substitutions
2.4k ($12_{CNEEs/OG}$)
208 ($1.0_{CNEEs/OG}$)
All expressed scOGs
79k ($11_{CNEEs/OG}$)
28.3% - $0.3_{CNEE/OG}$
7.4k ($1.1_{CNEEs/OG}$)
DEGs
2.6k ($16_{CNEEs/OG}$)
25.8% - $0.4_{CNEE/OG}$
233 ($1.4_{CNEEs/OG}$)
F
No difference in selective pressure (BUSTED-PH)
Difference in selective pressure (BUSTED-PH)
Accelerated Rate (M1)
Median: 1.28
n.s.
Conservation Rate (M1)
Median: 0.093
n.s.