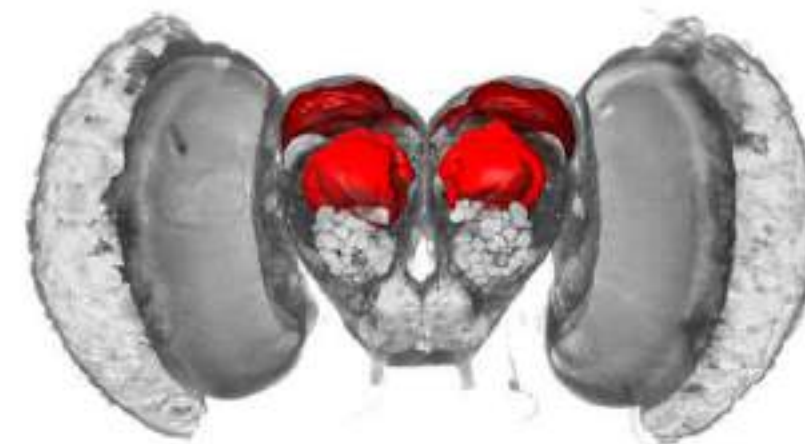


Day 9- Comparative Genomics

Attempting to make sense of how evolution works

F. Cicconardi, PhD



EBaB lab

2025 WORKSHOP ON GENOMICS, CESKY KRUMLOV

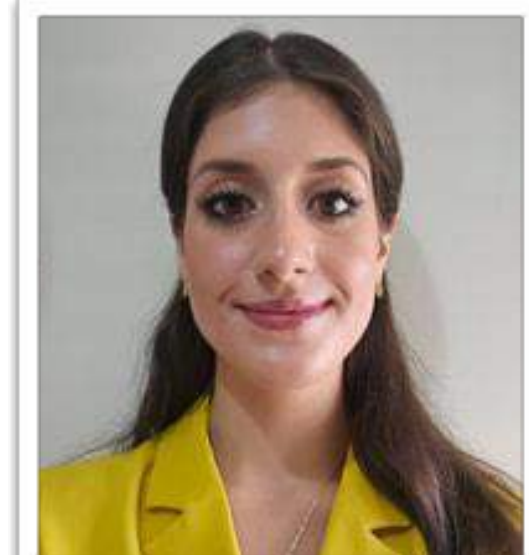
Evolution of Brain and Behaviour Lab



Dr. Stephen Montgomery

EBAB LAB

[HOME](#) [PEOPLE](#) [RESEARCH](#) [PUBLICATIONS](#) [OUTREACH](#) [OPPORTUNITIES](#)



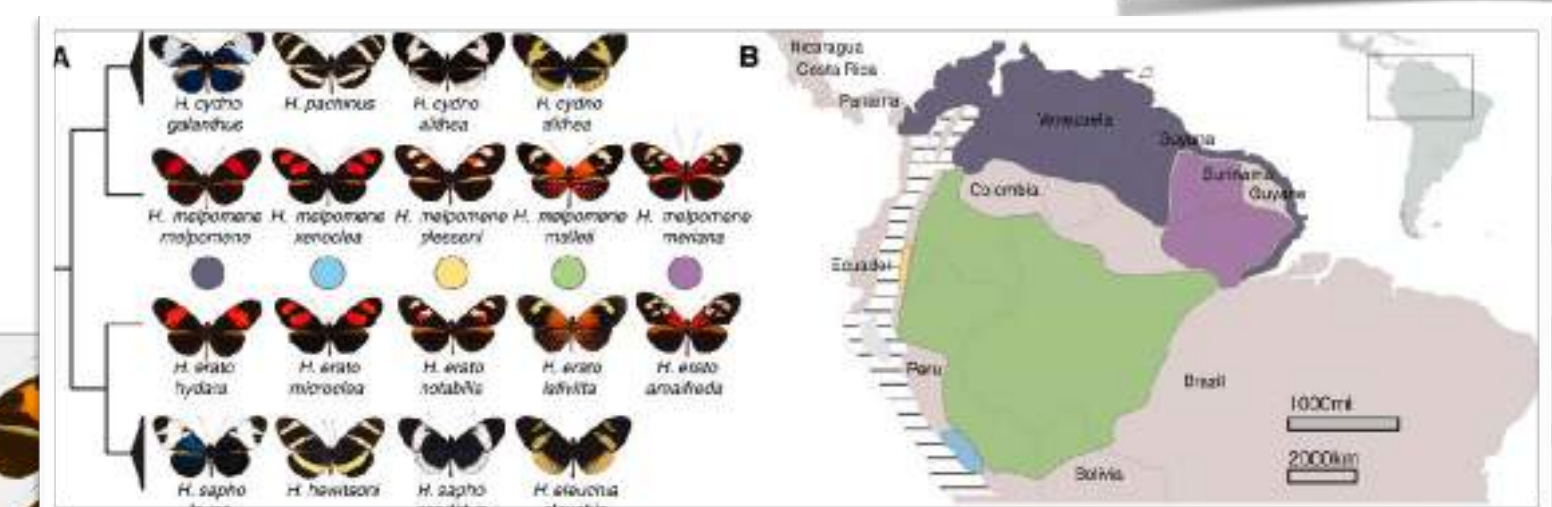
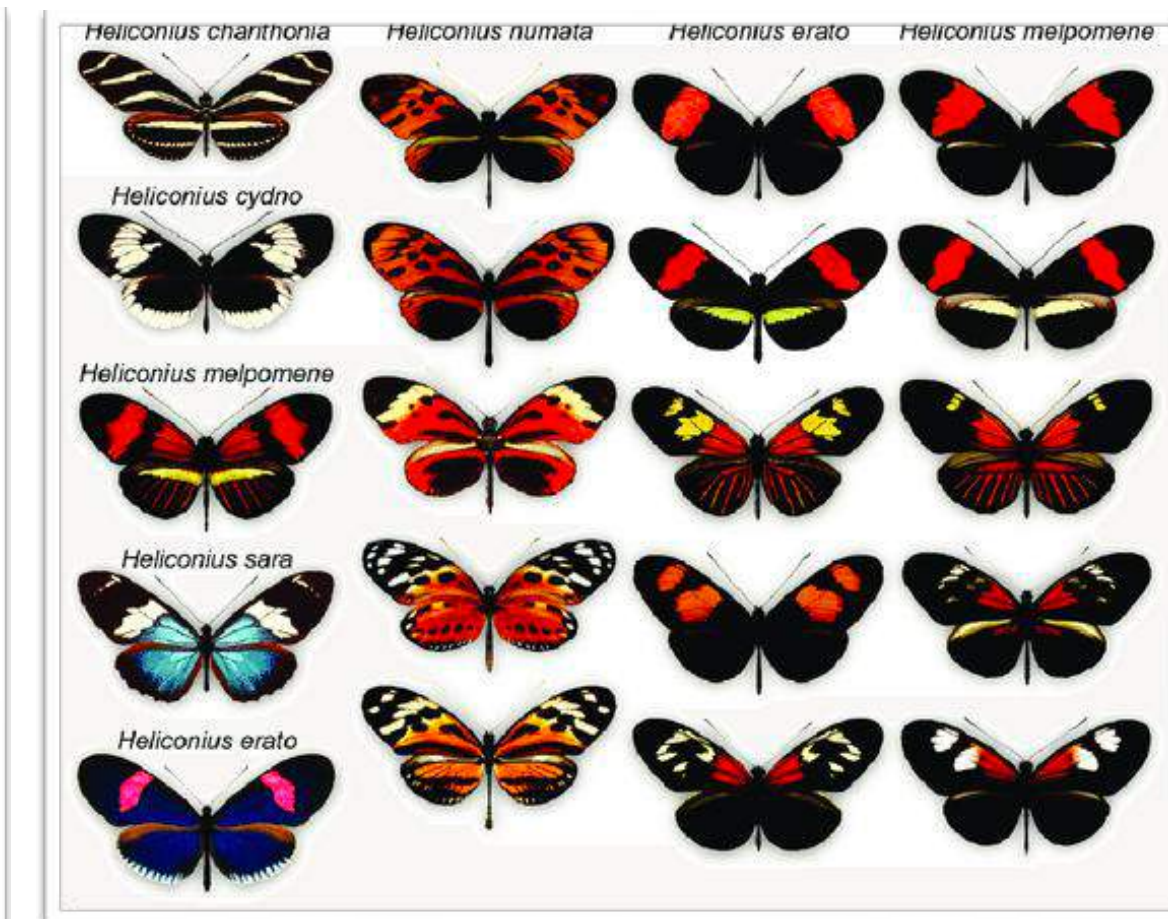
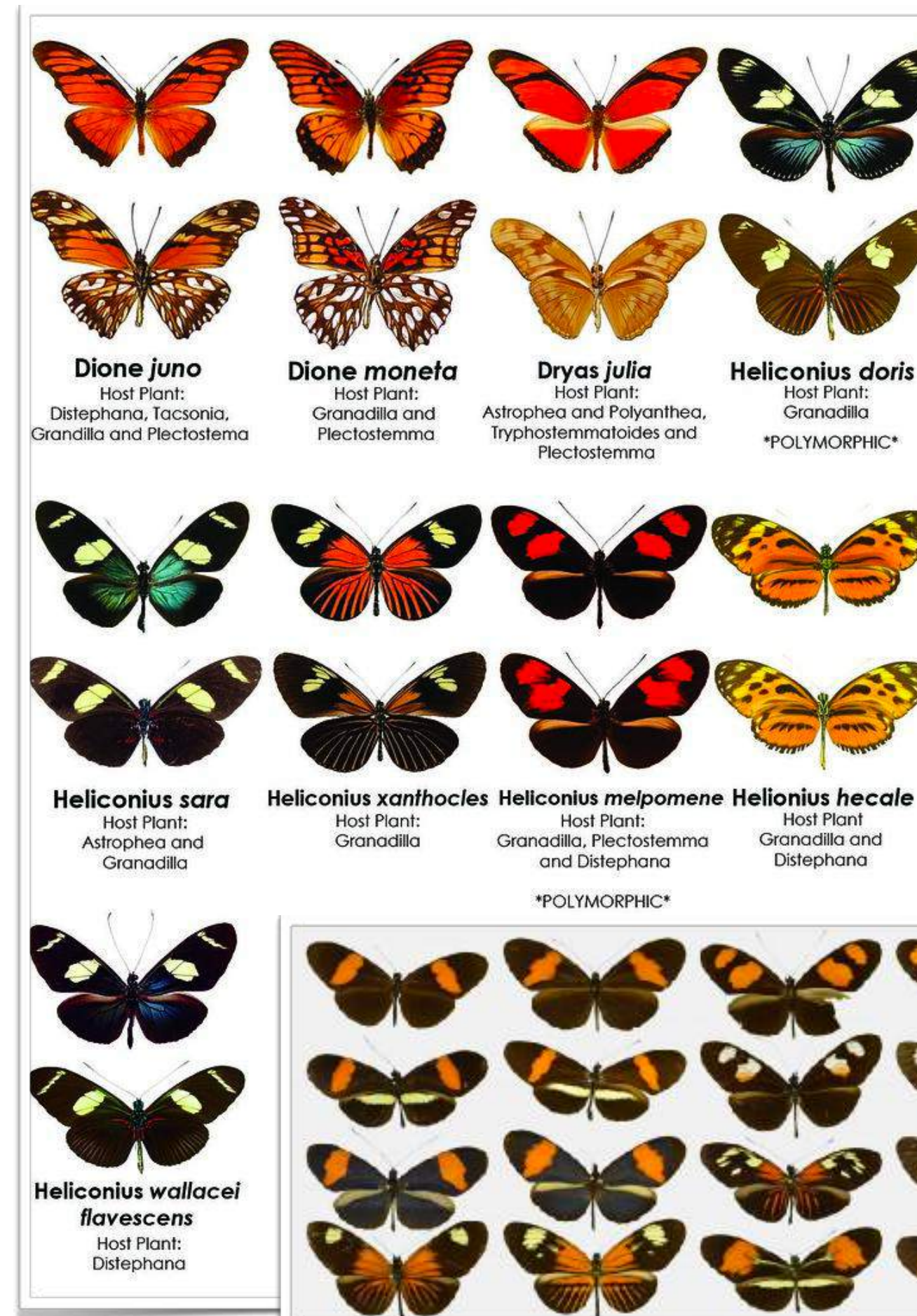
www.shmontgomery.co.uk

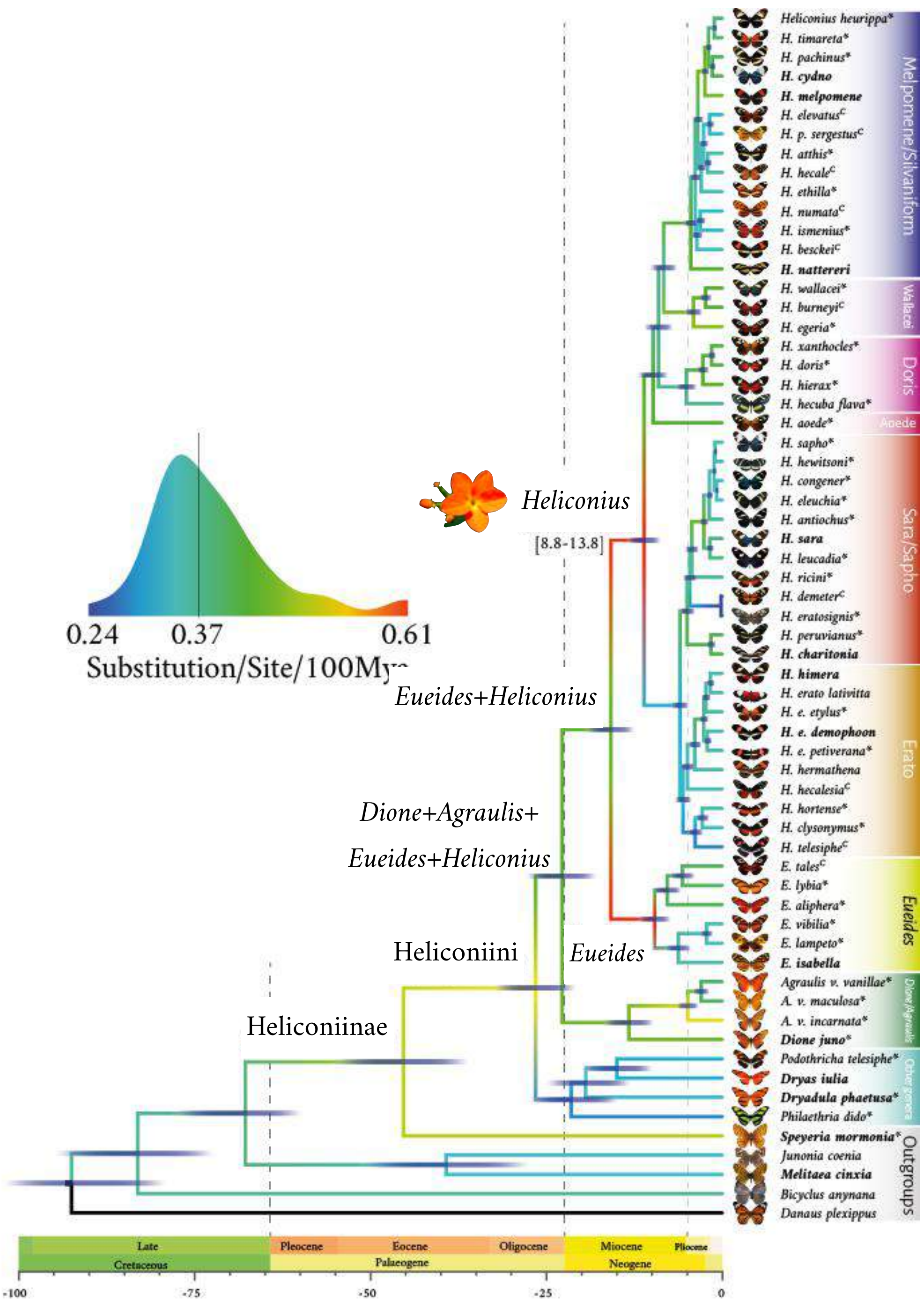


Photo credit: @mena_sebas

A close-up photograph of two butterflies on a bright red, fleshy flower. The butterfly on the left has brown wings with a lighter brown band. The butterfly on the right has black wings with a prominent white band and red spots on its hindwings. The background is dark and out of focus, showing green leaves.

- ***~440 sub-species***

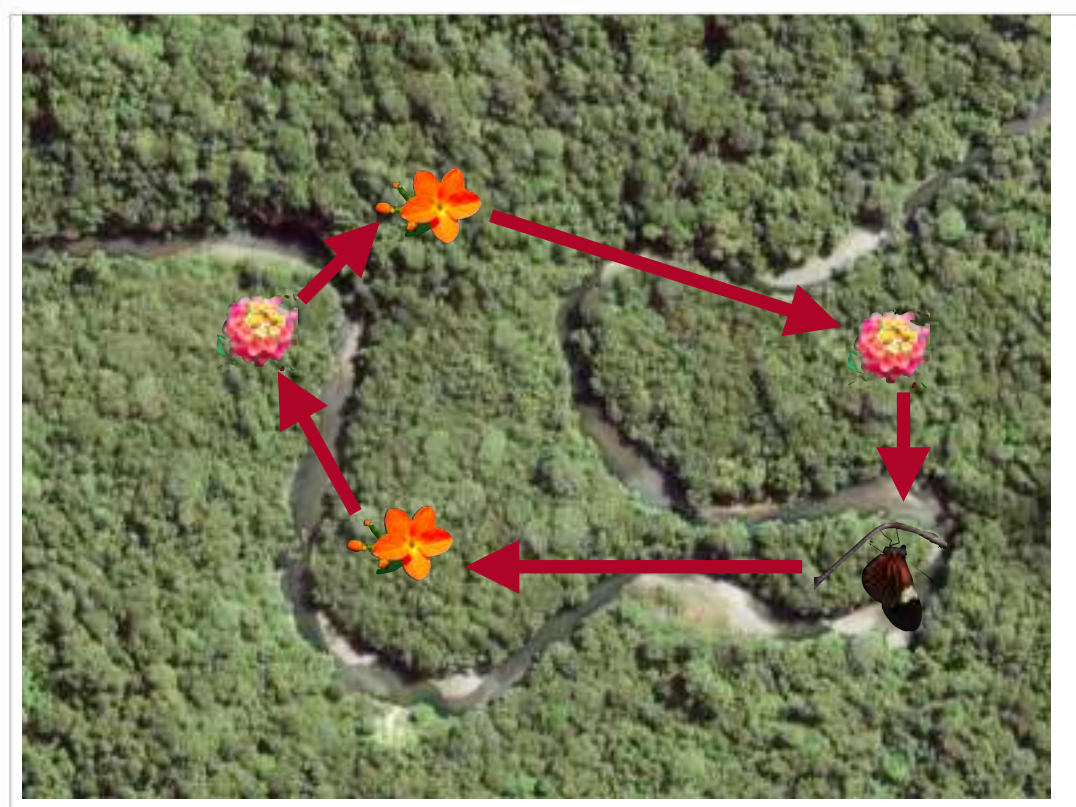




» Pollen-feeding

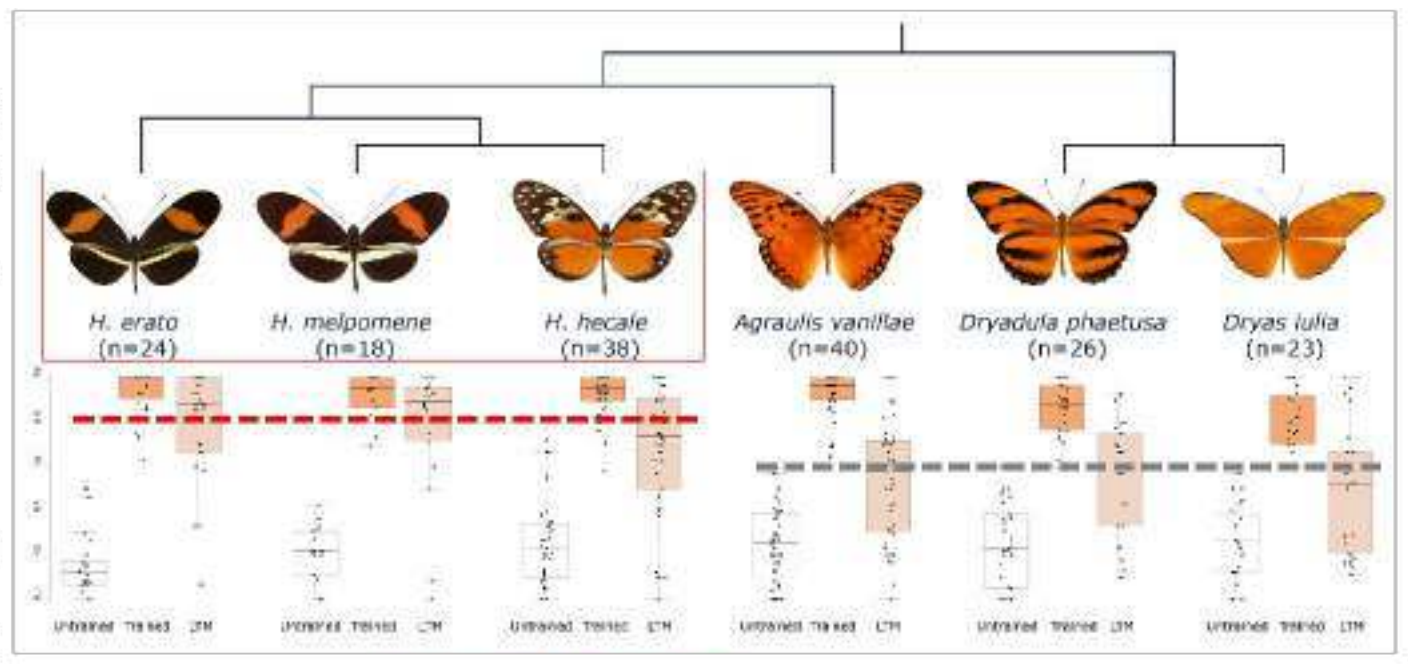
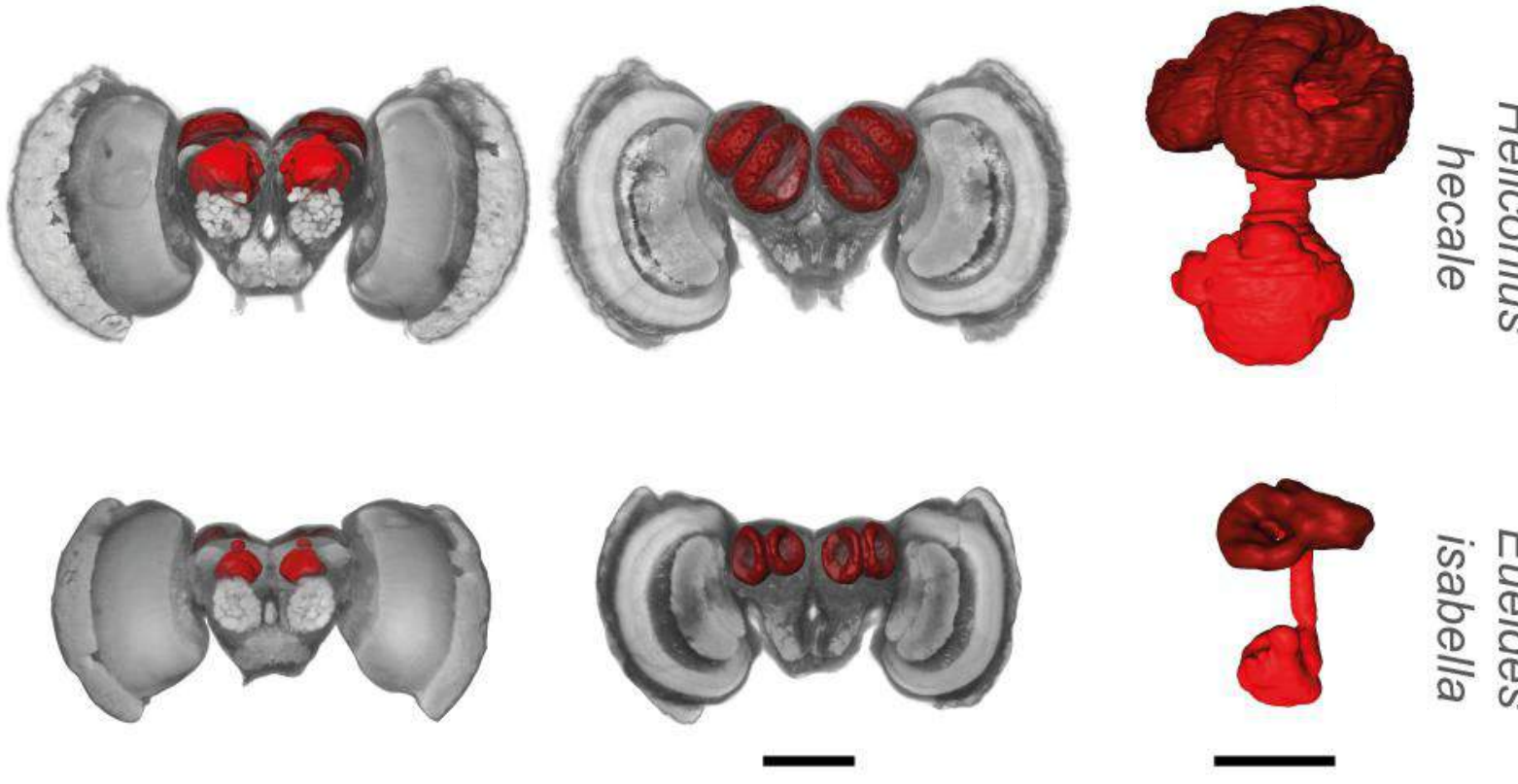


» Trap-lining behaviour



Moura et al., 2021. Functional Ecology.

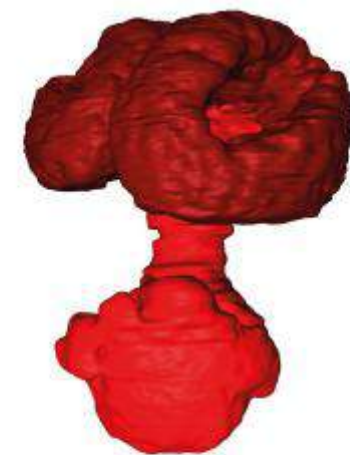
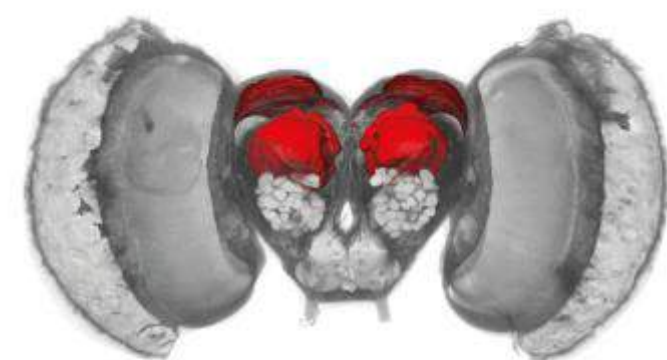
» Mushroom body expansion (Brain)



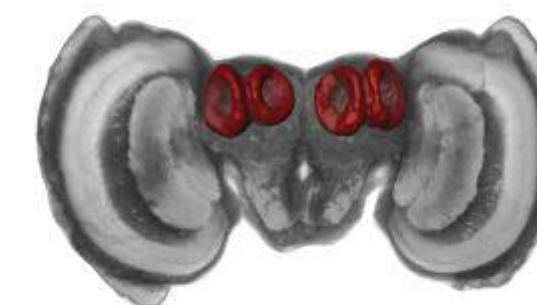
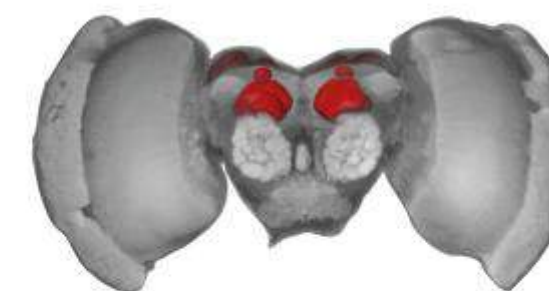
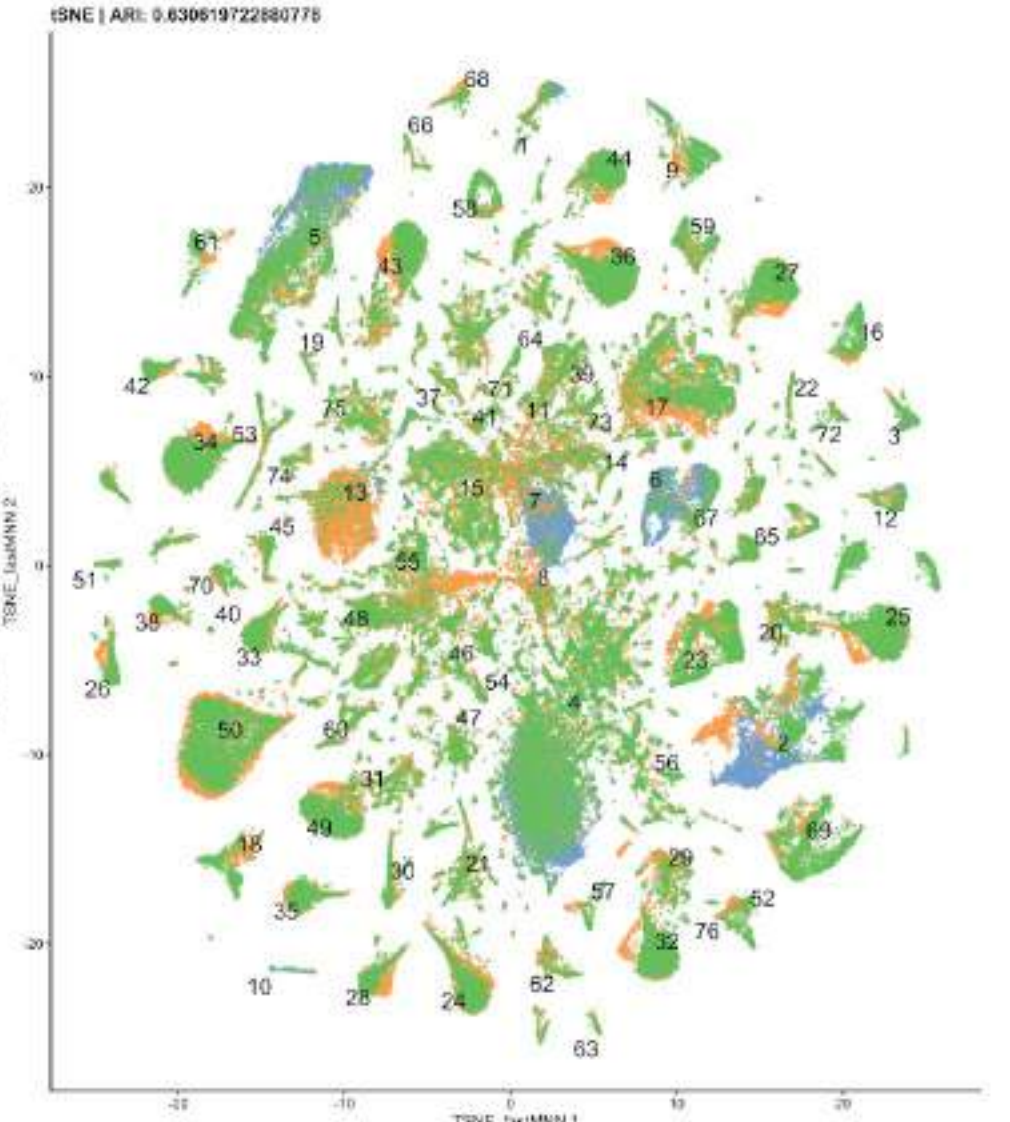
Interspecific variation: $X^2 = 76.2, p < 0.001$
 Small vs. big MBs: $X^2 = 182.8, p < 0.001$

A butterfly with dark wings and bright orange markings. The wings are spread, showing a large orange patch on each forewing and a smaller one on the hindwing. The body is dark and slender.

A t-SNE plot visualizing gene clusters. The x-axis is labeled "TSNE_dimMN 1" and ranges from -20 to 20. The y-axis is labeled "TSNE_dimMN 2" and ranges from -20 to 20. Numerous clusters are visible, each represented by a distinct color and labeled with a number. The clusters are distributed across the plot area, with some larger central clusters and many smaller peripheral ones. The colors used include shades of purple, blue, green, yellow, orange, red, and pink.



A detailed illustration of a butterfly, possibly a Papilio polytes, shown from a dorsal view with its wings spread. The wings are primarily a bright orange color, with dark brown or black markings along the edges and veins. The body is dark and segmented, with two long, thin antennae extending from the head. The background is plain white.



» 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.

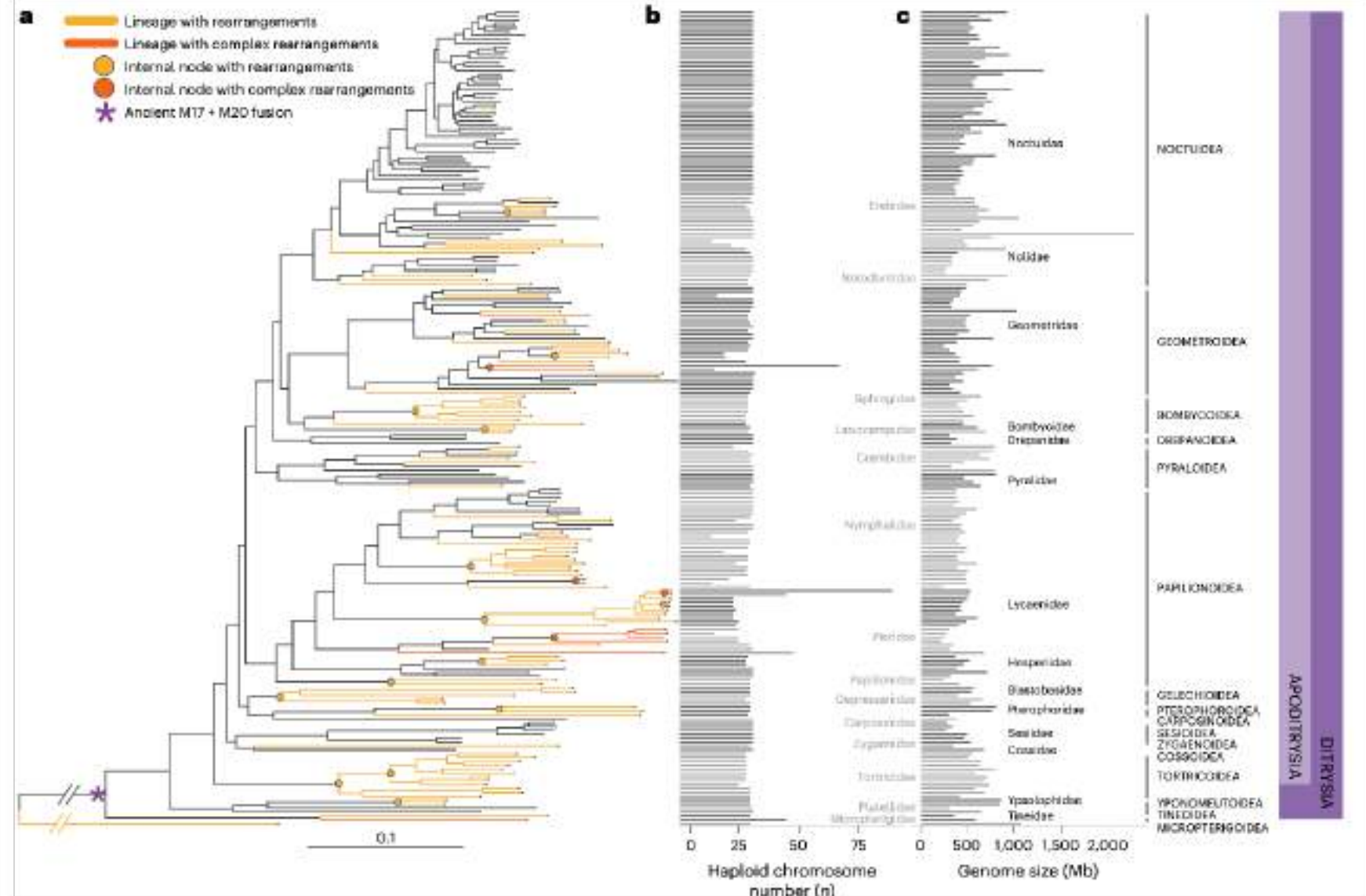
Genetic differences **within one species** such as our own can reveal variants with a role in disease.

Article

Comparative genomics reveals the dynamics of chromosome evolution in Lepidoptera

Received: 9 October 2023
Accepted: 12 January 2024
Published online: 21 February 2024

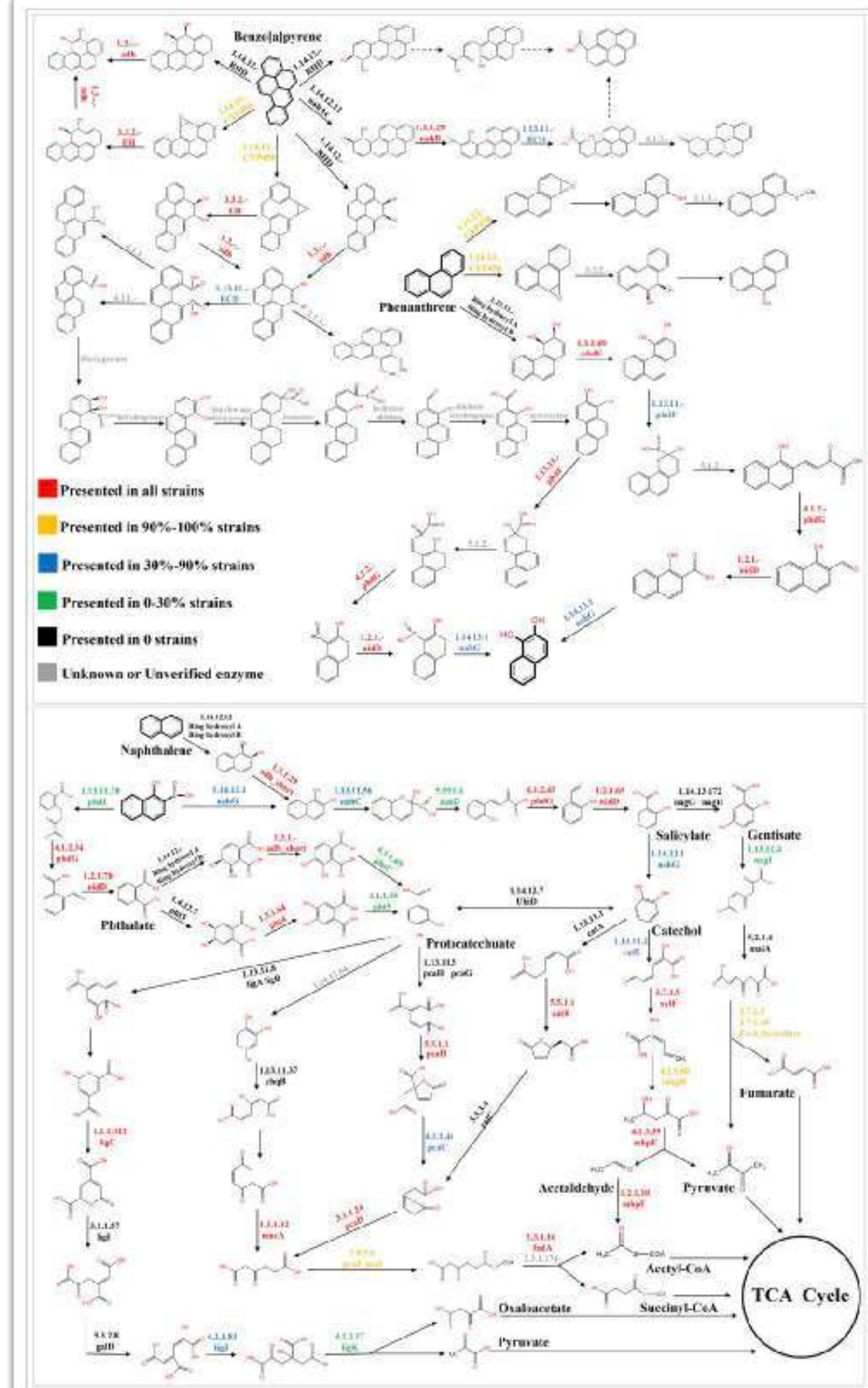
Charlotte J. Wright¹, Lewis Stevens¹, Alexander Mackintosh², Mara Lawnczak¹ & Mark Blaxter¹



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,c}, Zhong Hu^{a,b,*}

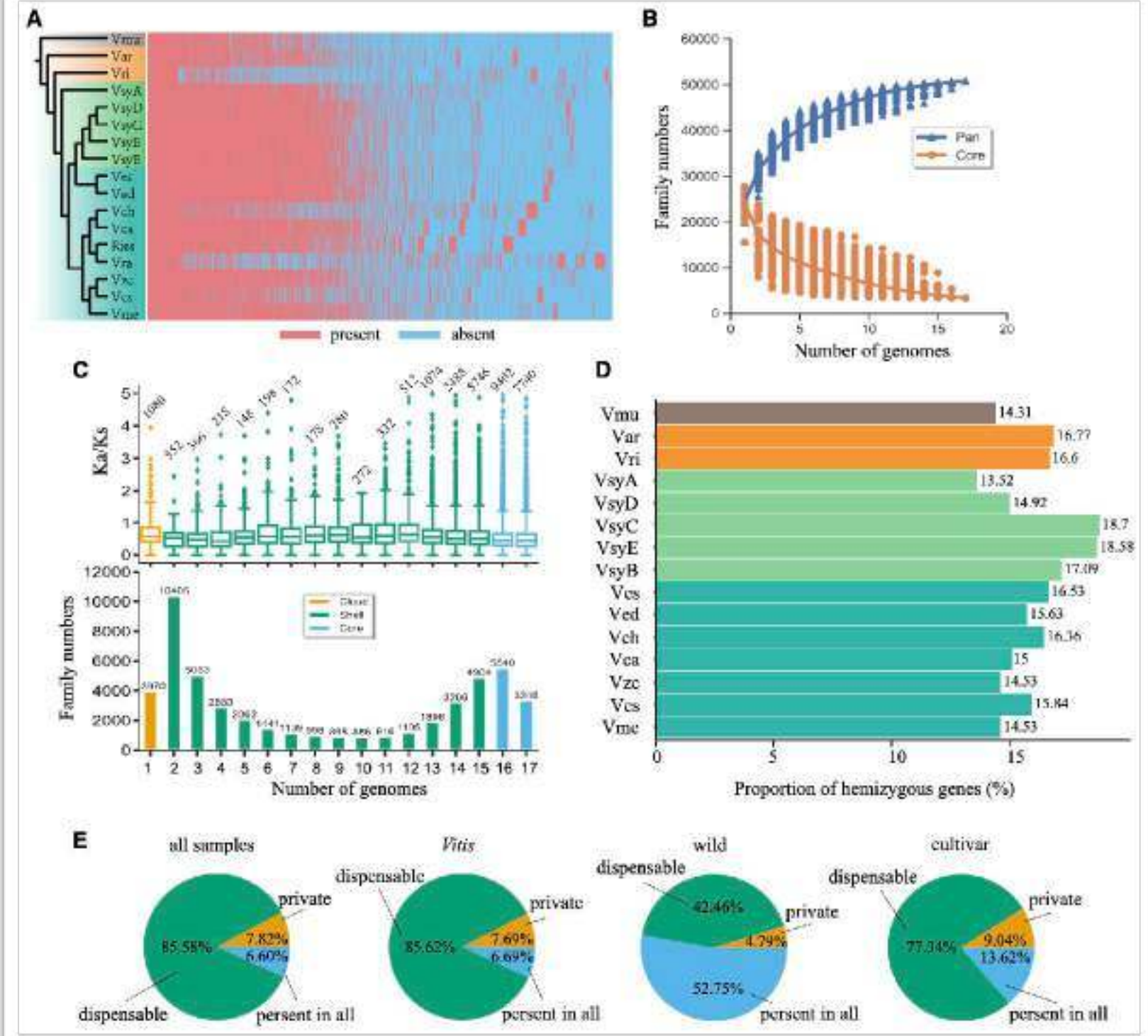
^a Department of Biology, Shantou University, Shantou, Guangdong 515063, PR China
^b Guangdong Research Center of Offshore Environmental Pollution Control Engineering, Shantou University, Shantou 515063, Guangdong, PR China



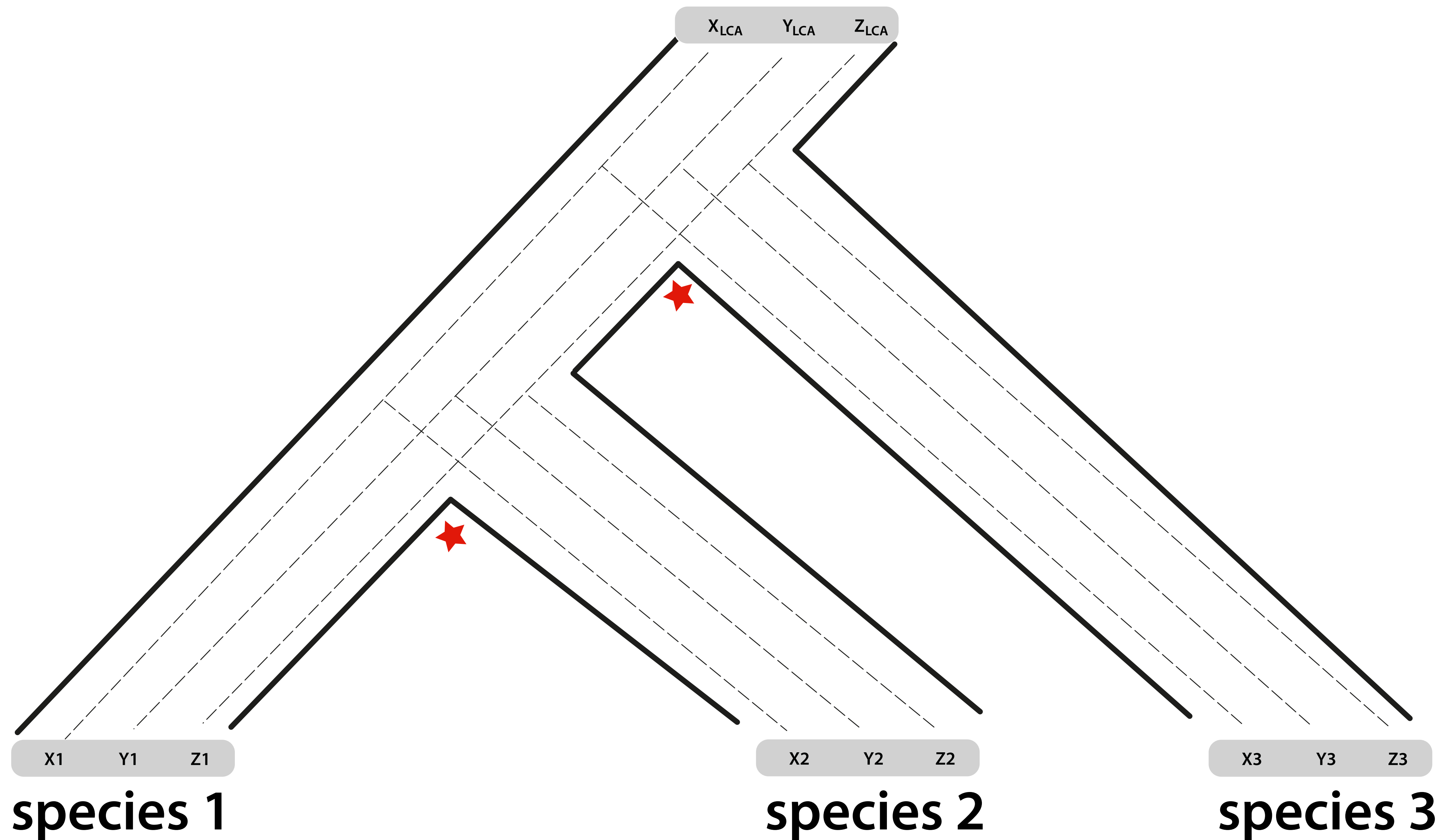
Plant Physiology®

Population comparative genomics discovers gene gain and loss during grapevine domestication

Qiming Long^{1,†}, Shuo Cao^{1,2,†}, Guizhou Huang¹, Xu Wang^{1,3}, Zhongjie Liu¹, Wenwen Liu¹, Yiwen Wang¹, Hua Xiao¹, Yanling Peng^{1,*} and Yongfeng Zhou^{1,4,*}

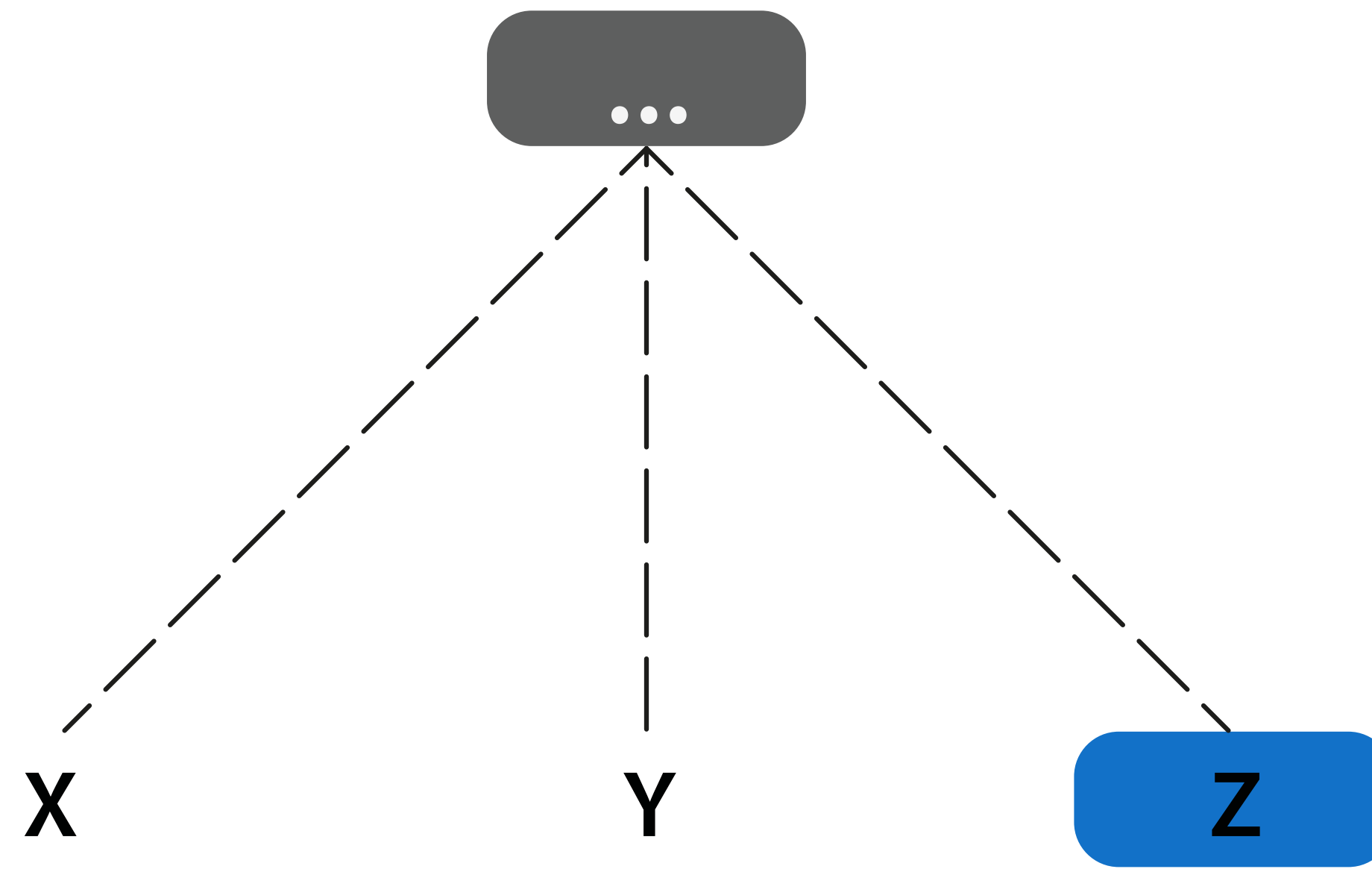


» ... Some definitions ... »



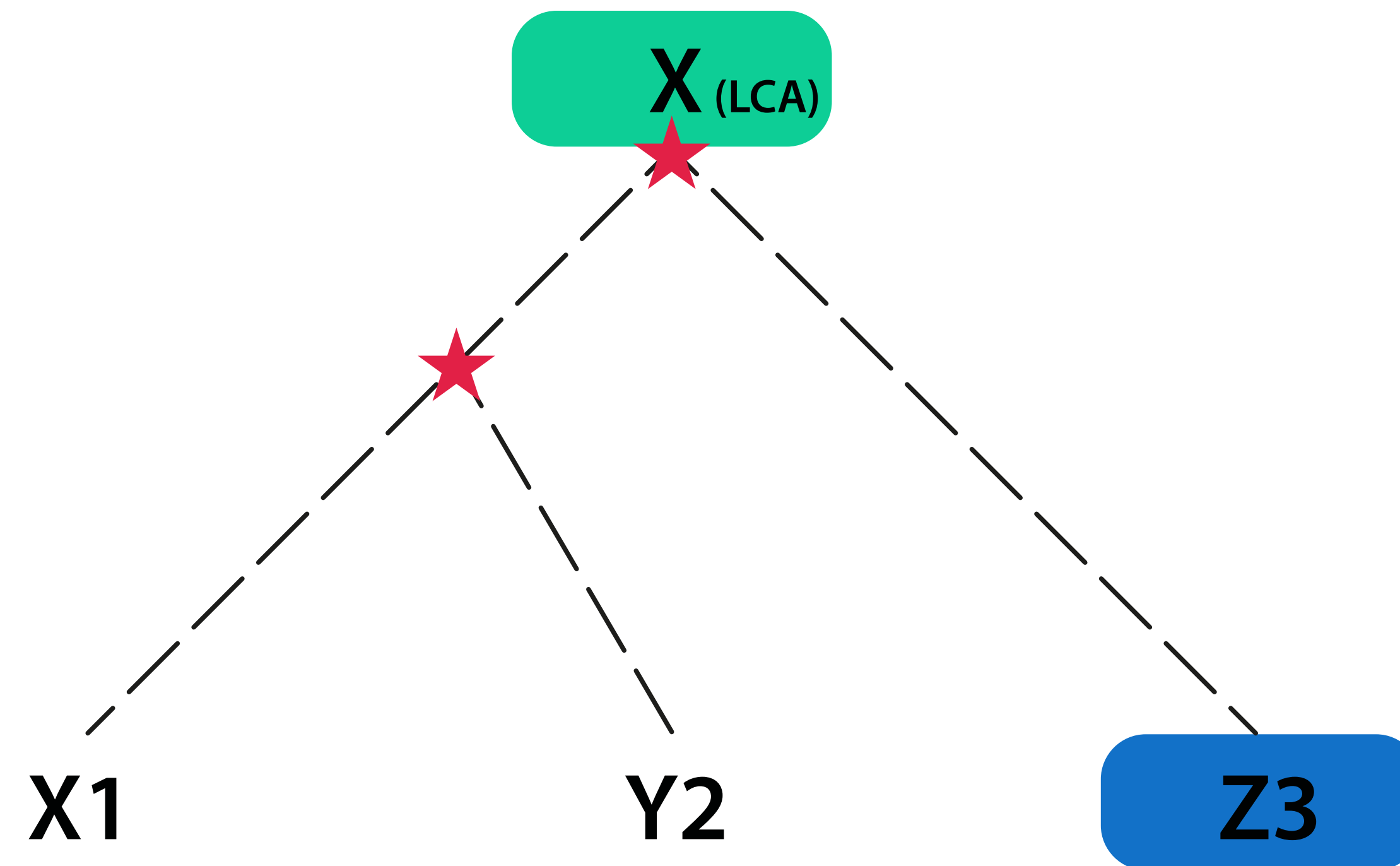
» Homology/Orthology definition »

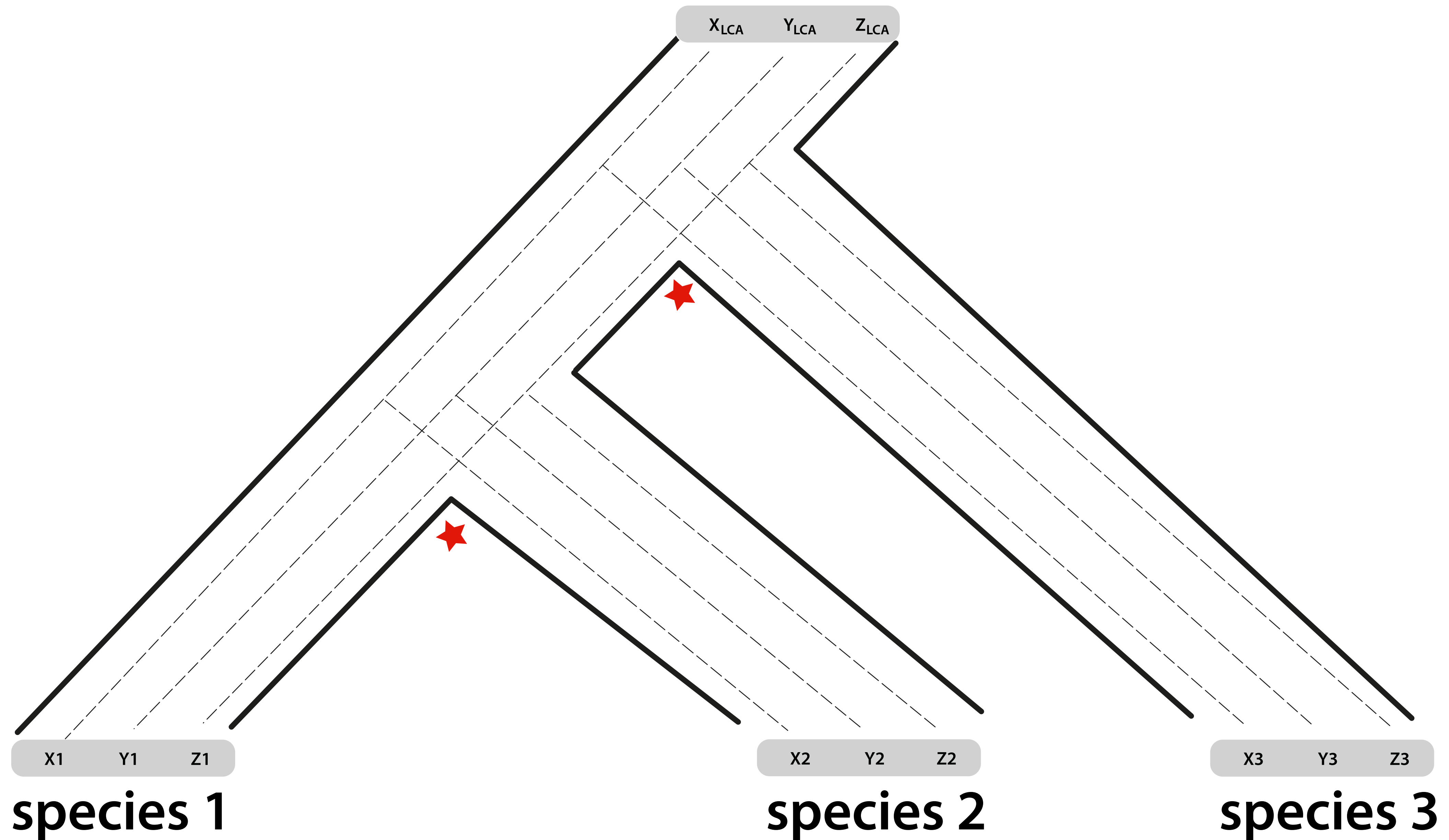
Homology: it describes descent from a common evolutionary origin: *two loci (genes) are homologous if they derive from the same ancestral locus (gene).*

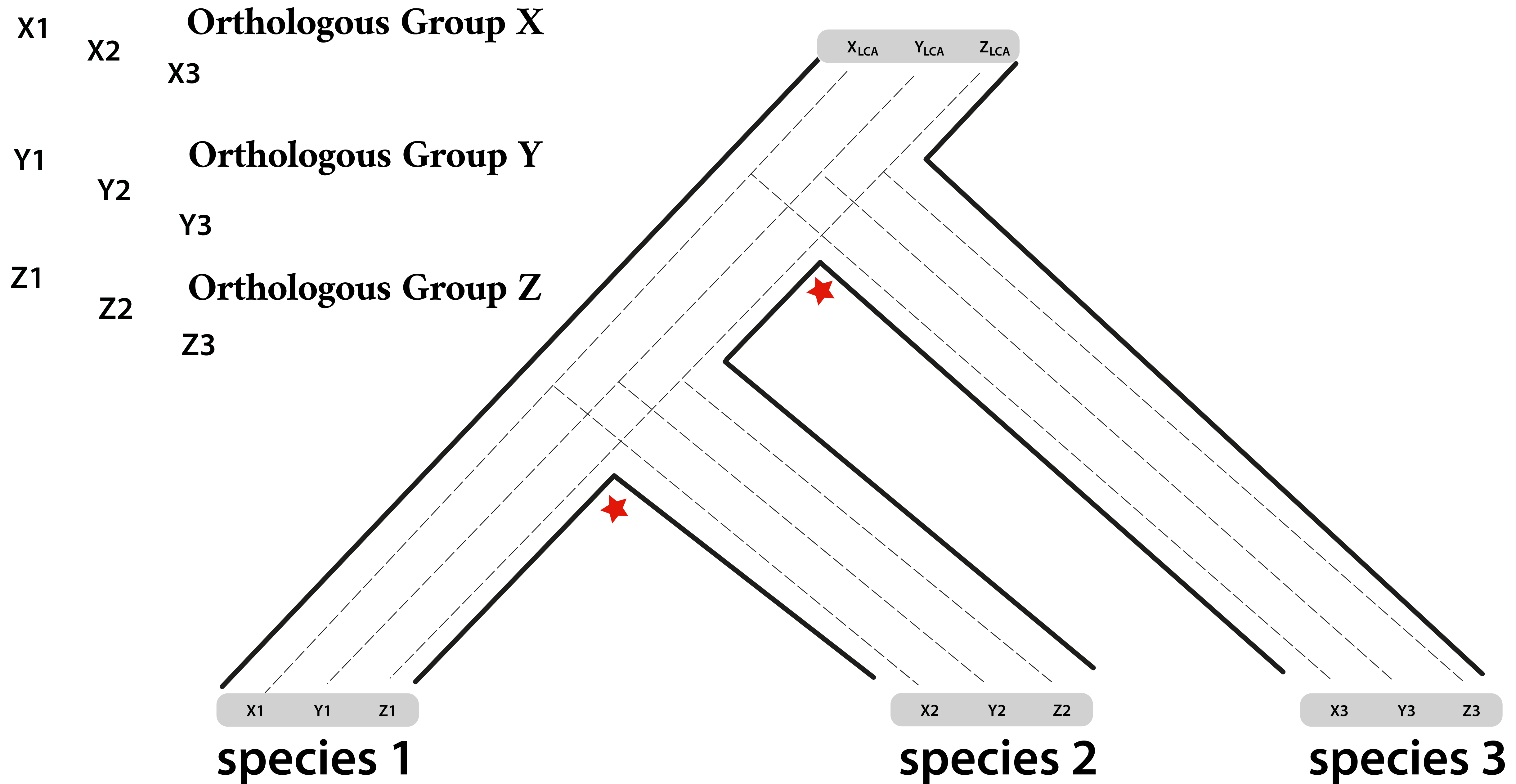


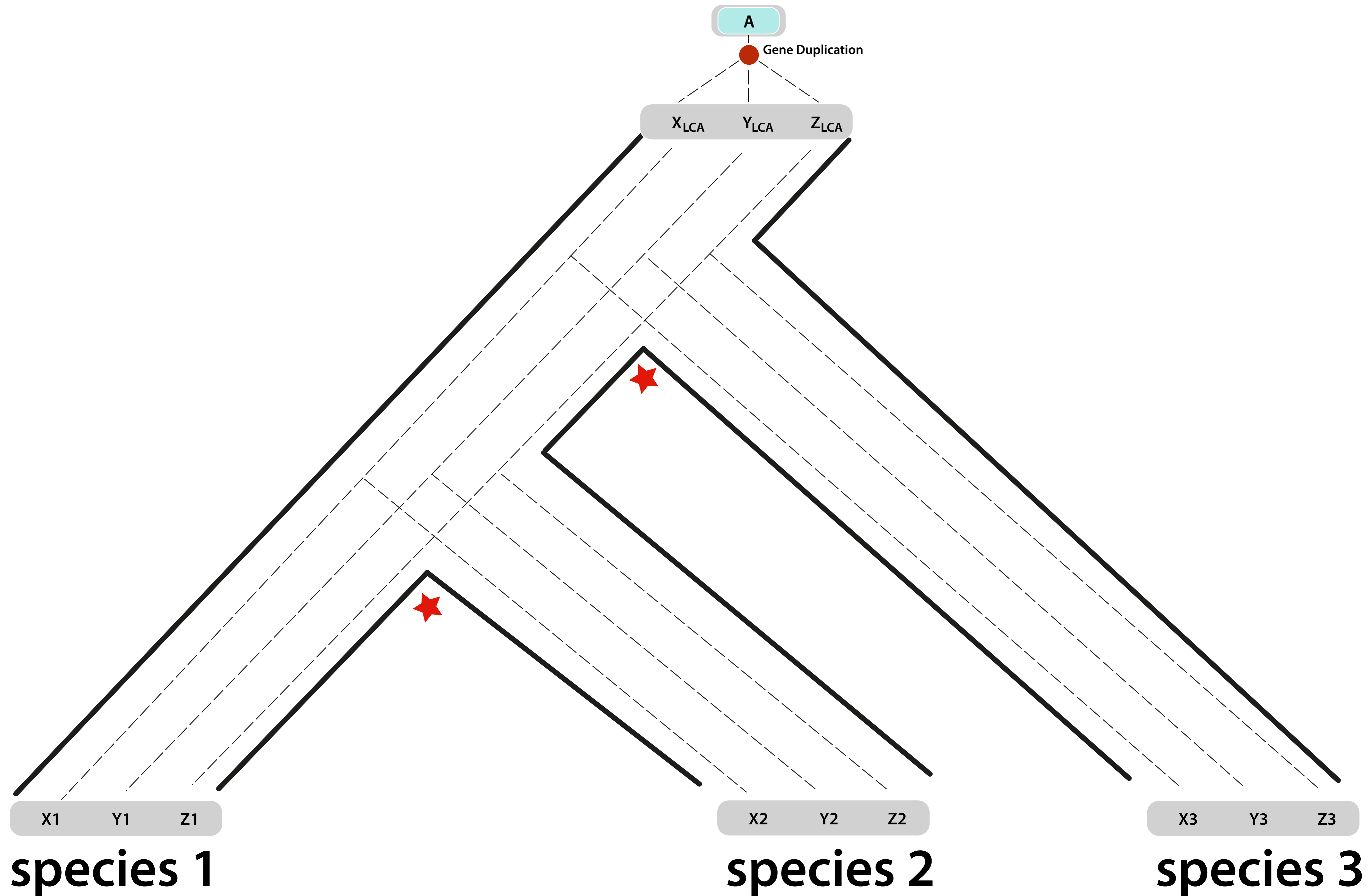
» Homology/Orthology definition »

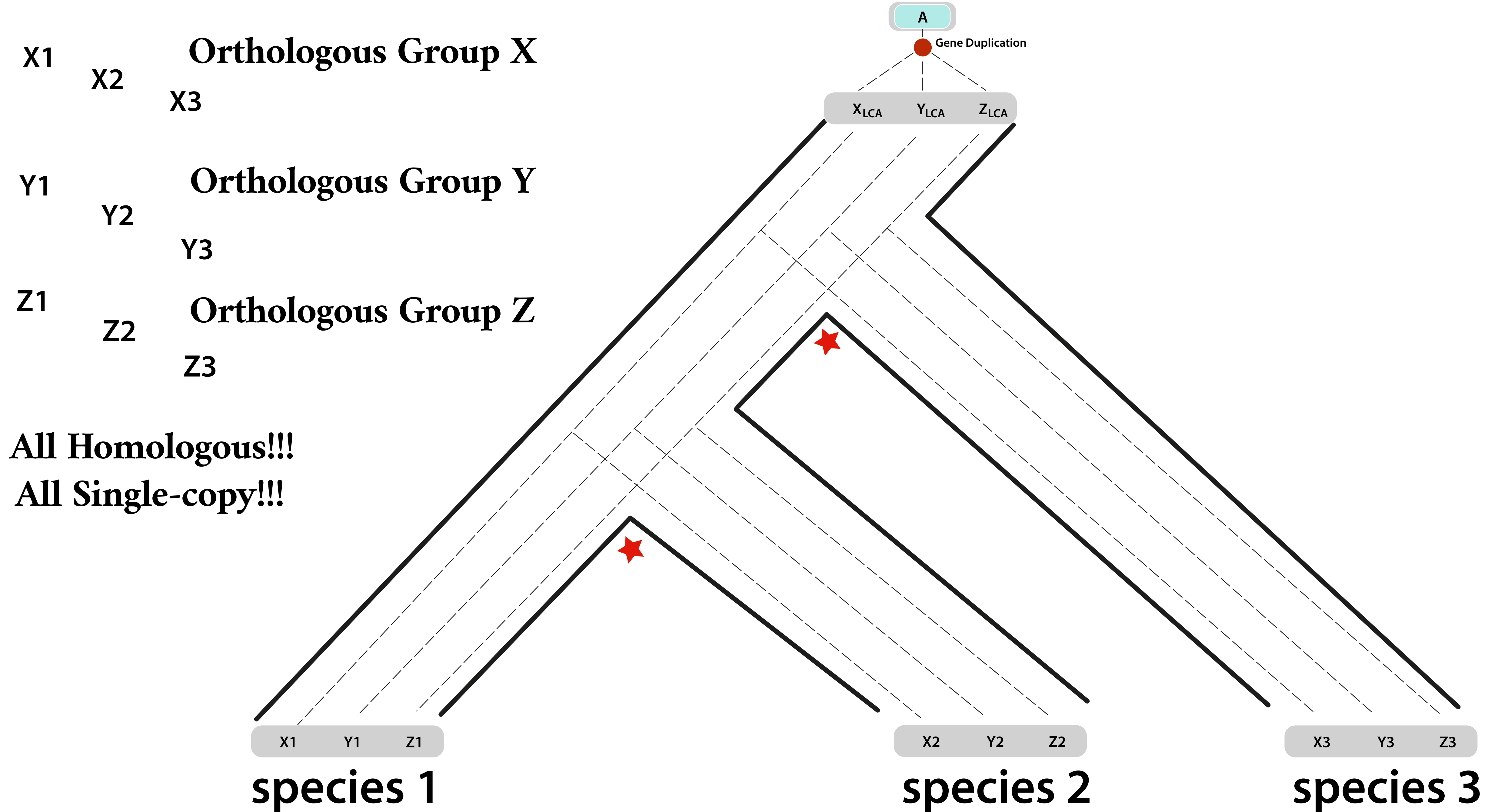
Orthology: it describes a relationship due to a speciation event:
two loci (genes) are orthologous if they derive from speciation.











» Gene loss »

X Y Z

X1

» Gene loss & duplication »

X_1 X_{a2} X_{b2} Orthologous Group X

Y_1 Y_2 Y_3 Orthologous Group X

Z_{a1} Z_{b1} Z_2 Z_3 Orthologous Group X



» Gene loss & duplication »

Za1
Zb1 ??

Xa2
Xb2 ??



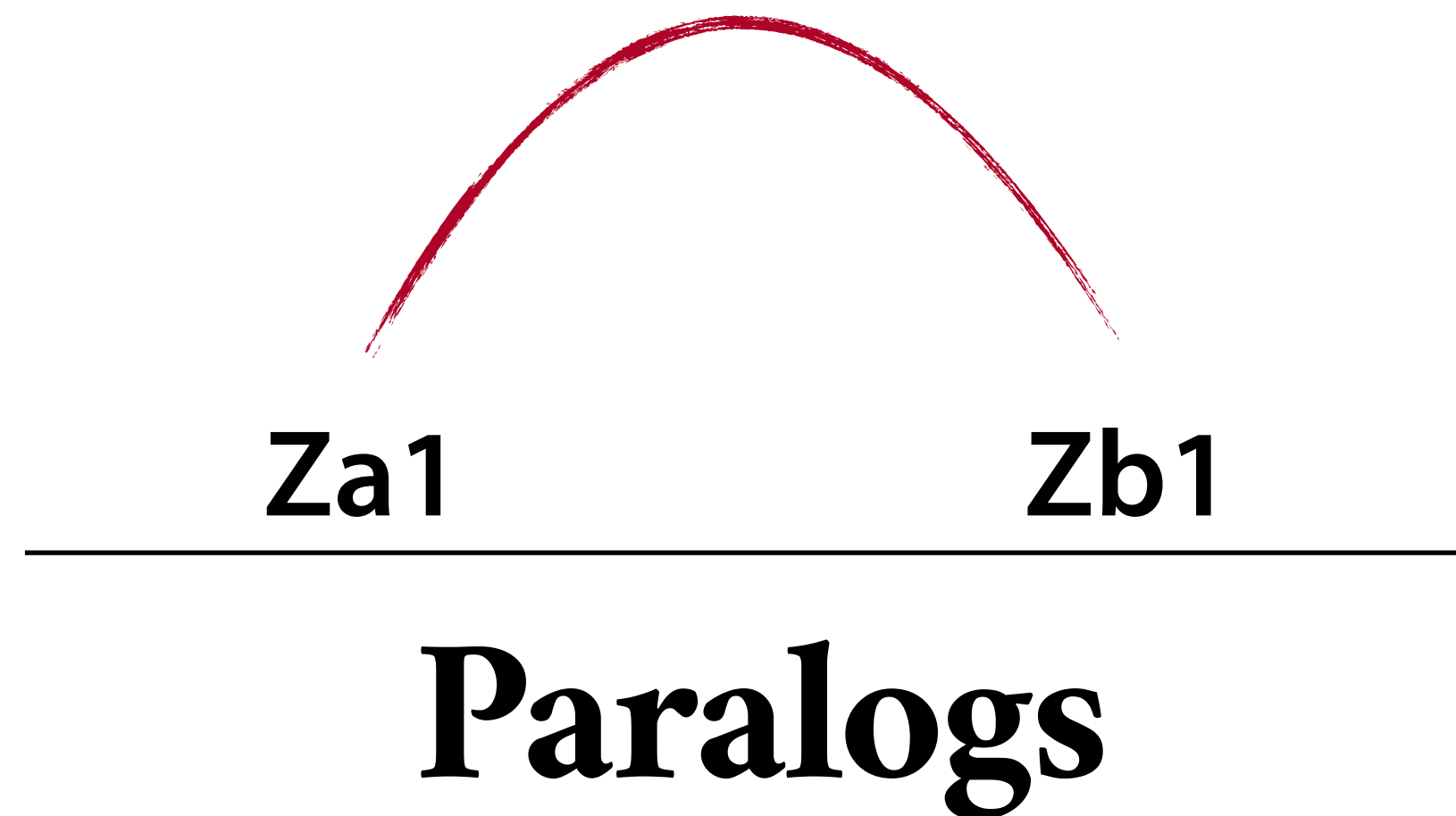
» Gene loss & duplication »

Za1
Zb1 **Paralogs**

Xa2
Xb2 **Paralogs**



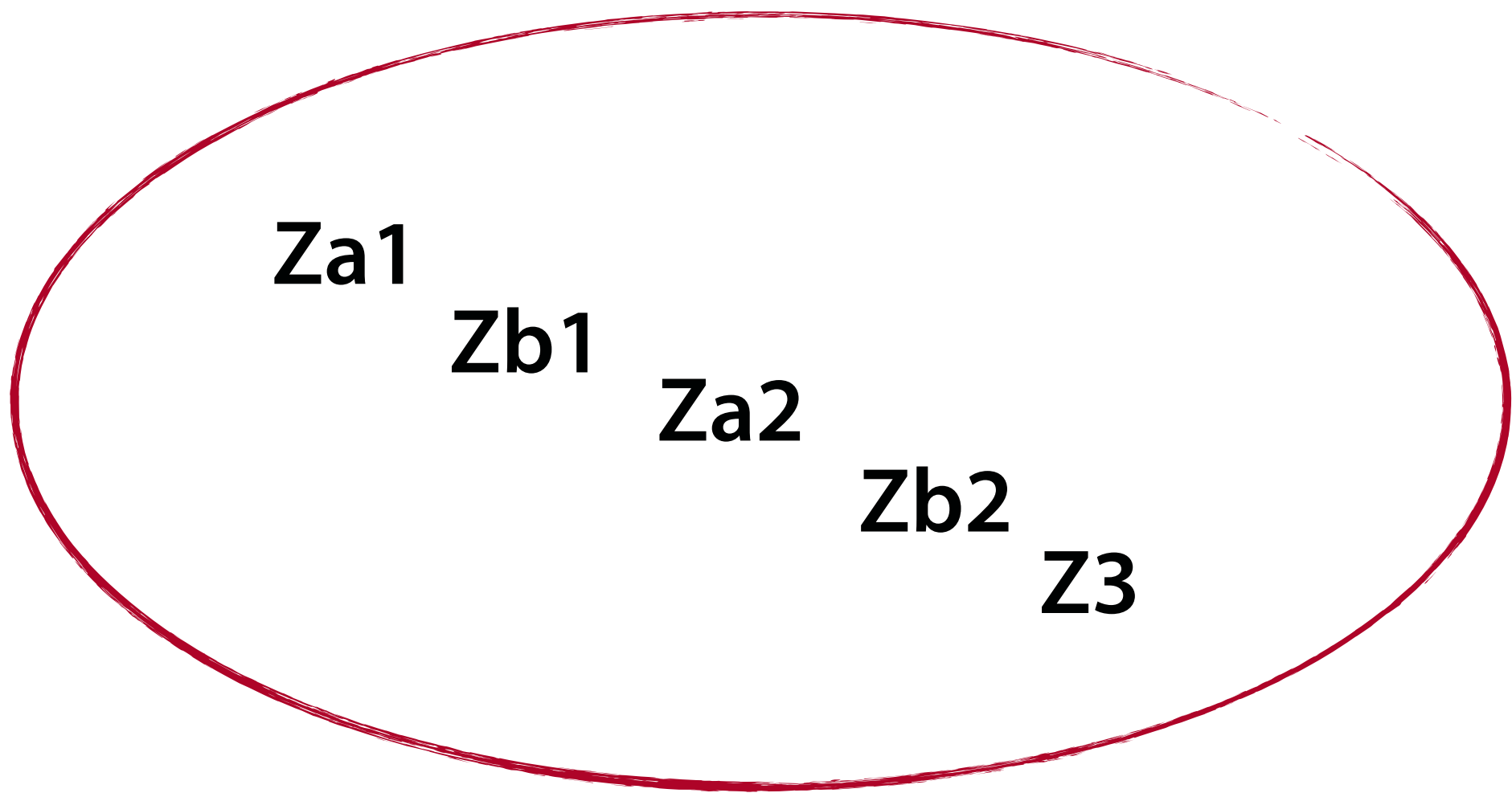
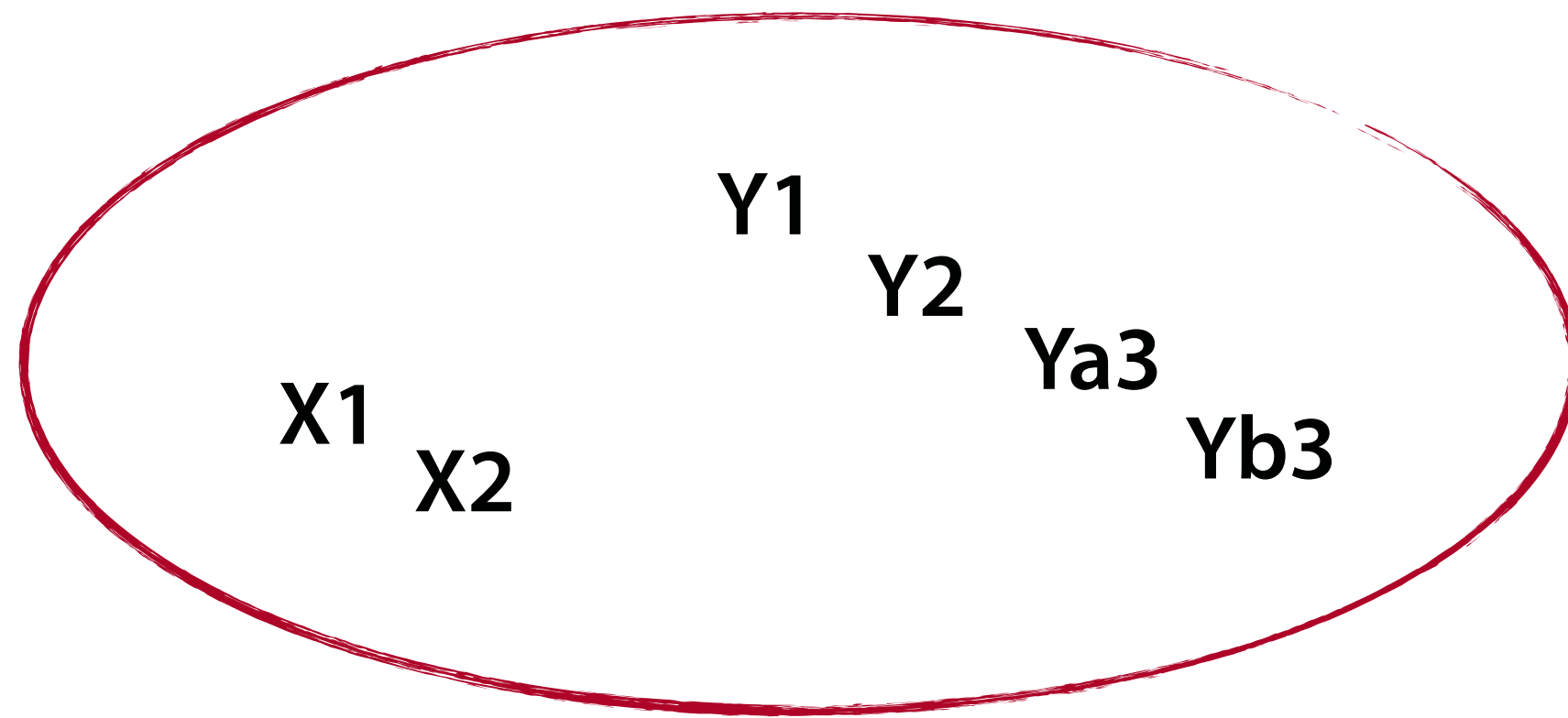
Paralogy: it describes a relationship that involve a duplication:
if a locus (gene) is generated by an event of tandem duplication.



» Homologs ? »

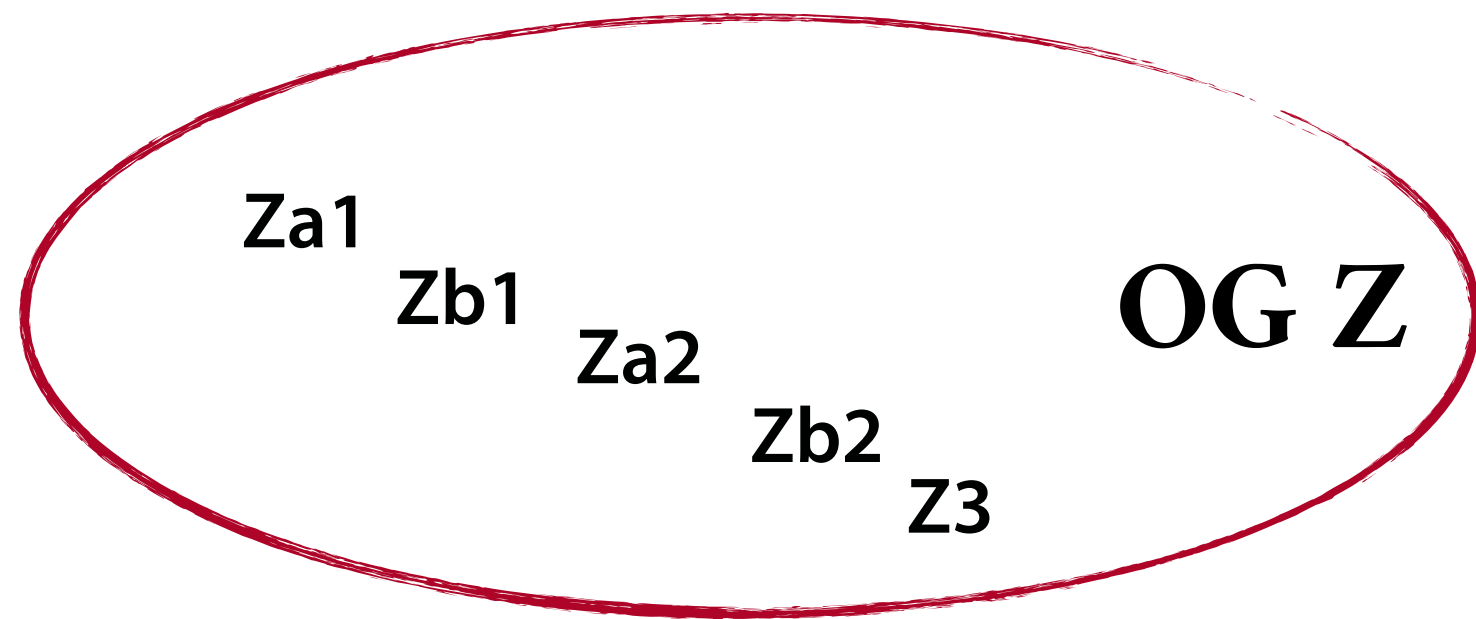
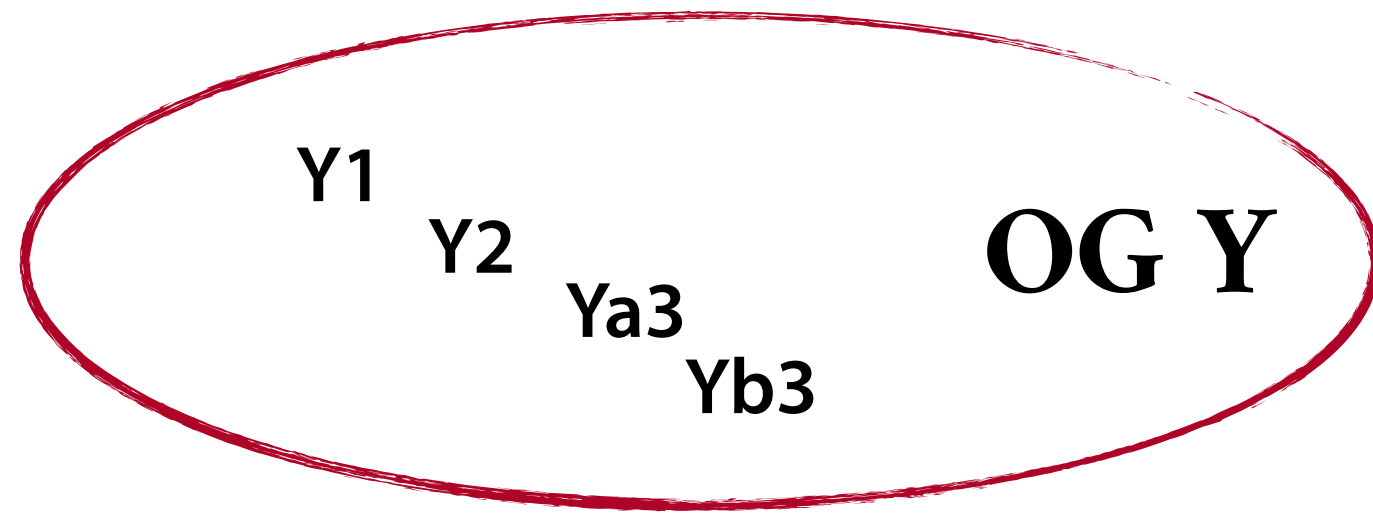
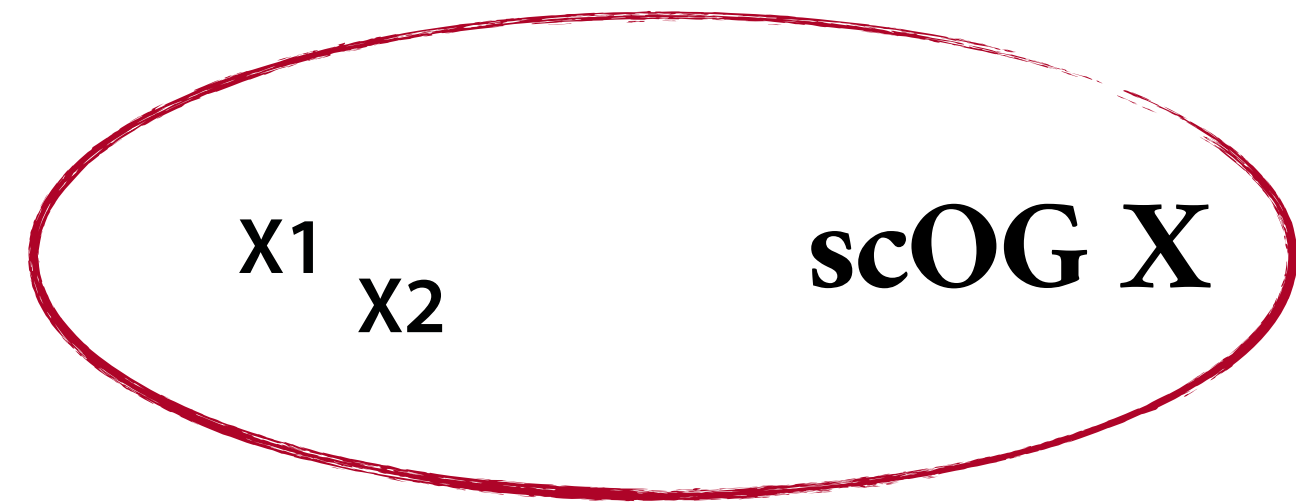


» Homologs ? »



» Orthologs ? »

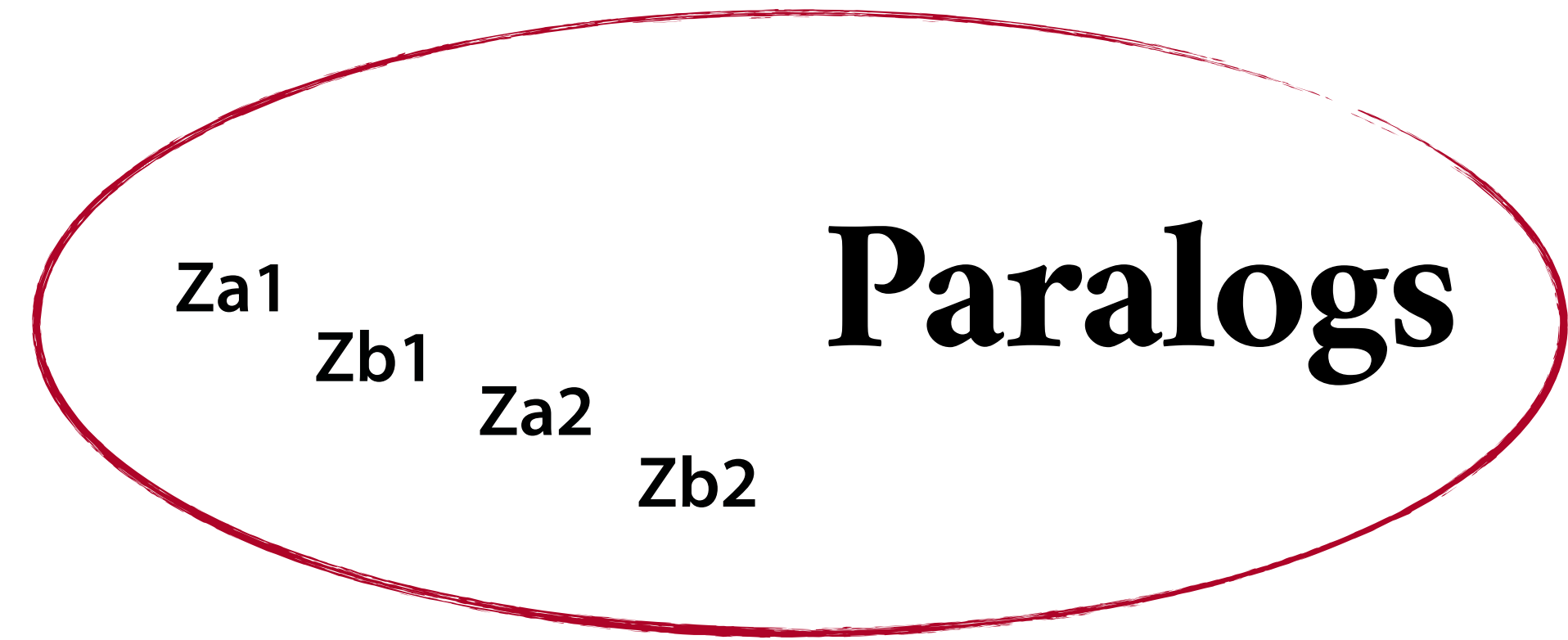
» Orthologs ? »



» Paralogs ? »



» Paralog ? »



» Paralog ? »

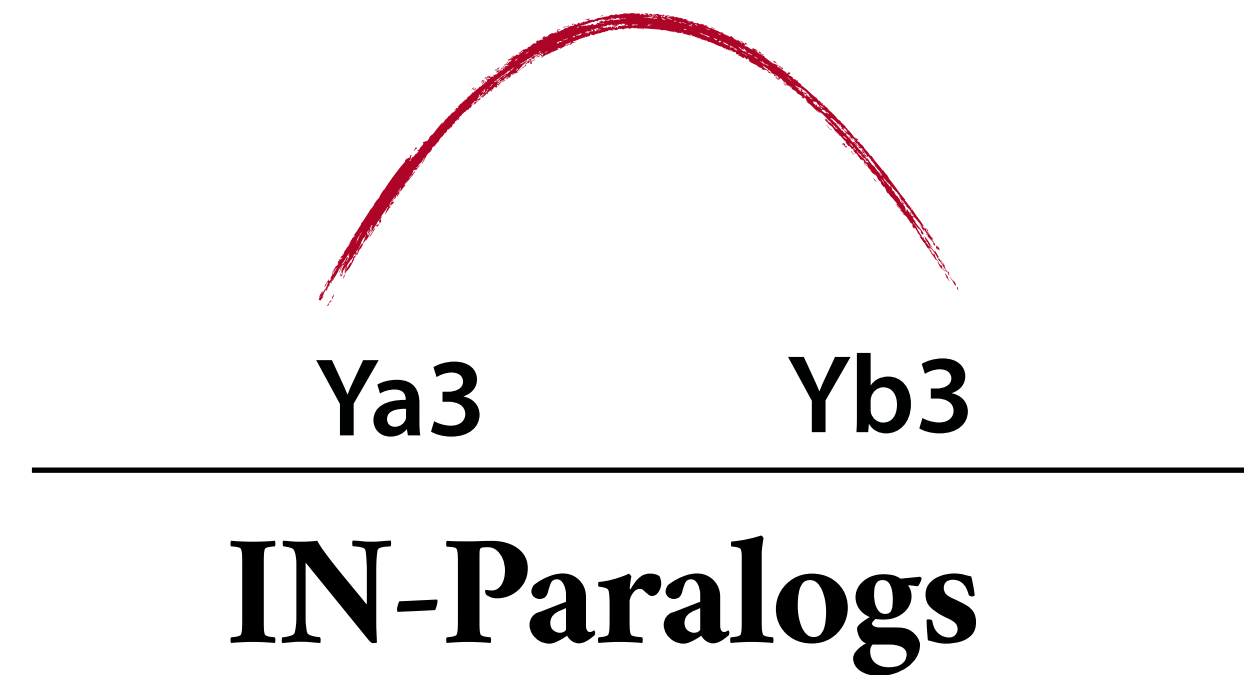
Out-Paralogs

Za1
Zb1
Za2
Zb2

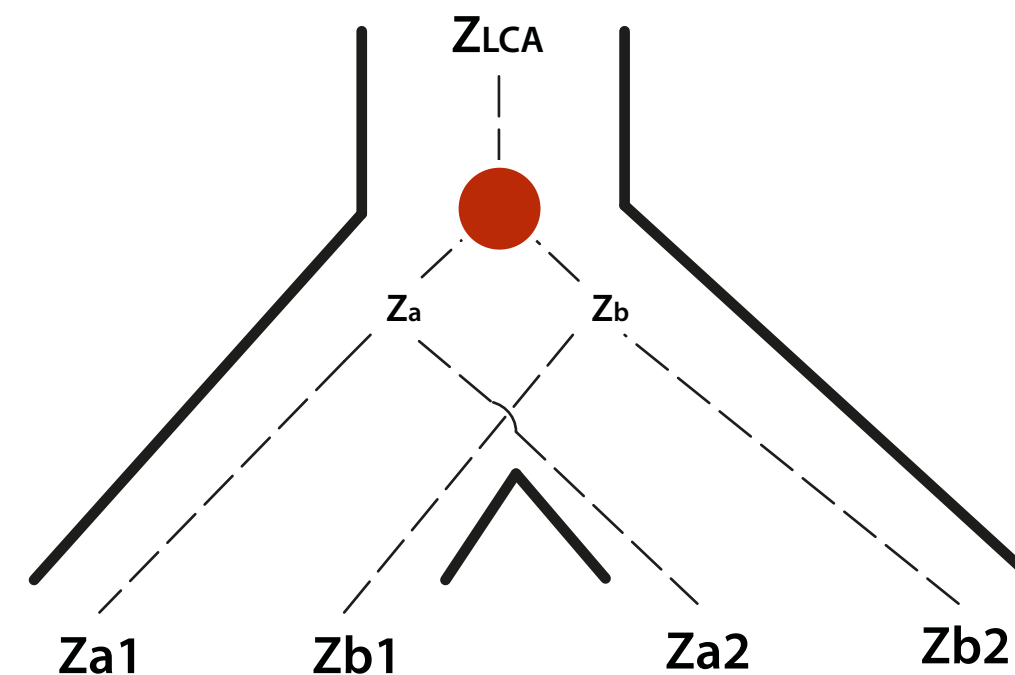
In-Paralogs

Ya3
Yb3

IN-Paralogy: it describes a relationship that involve a duplication occurred within a species

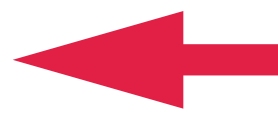


OUT-Paralogy: it describes a relationship that involve a duplication occurred in one of the ancestor



» Horizontal Gene Transfer (HGT) »

Xenologs: locus acquired through horizontal gene transfer



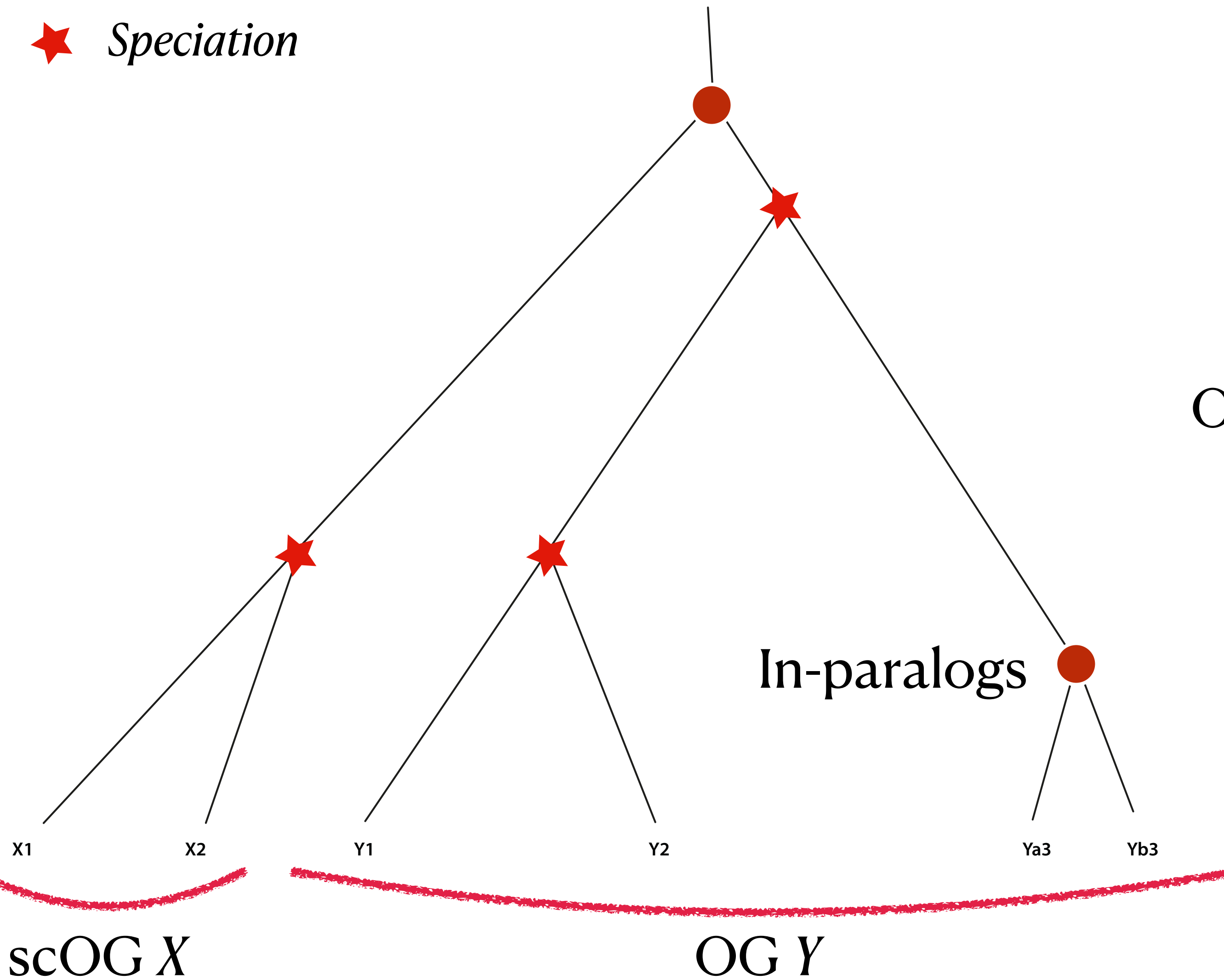
» Gene Family Tree »



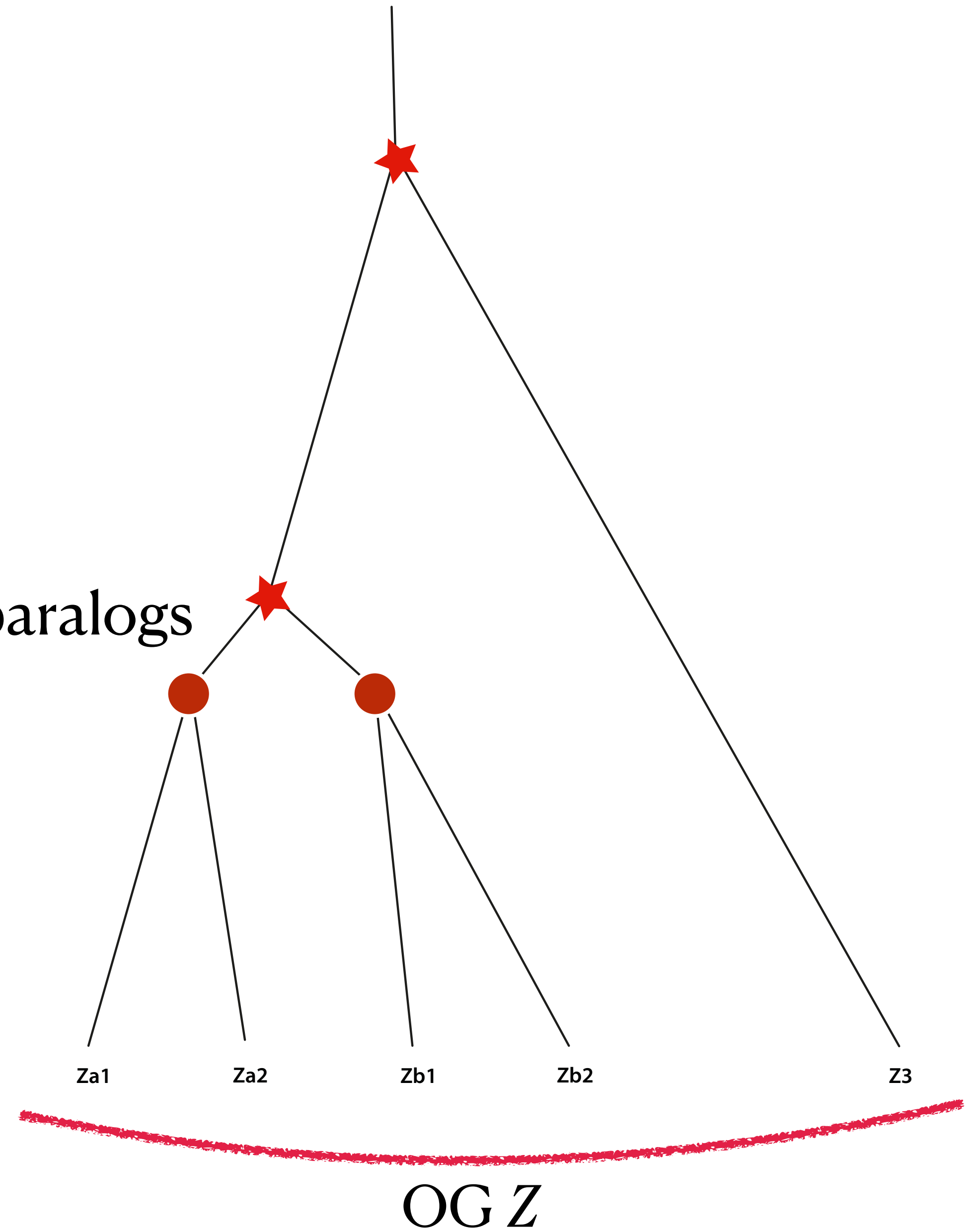
» Gene Family Trees »

● *Gene duplication*

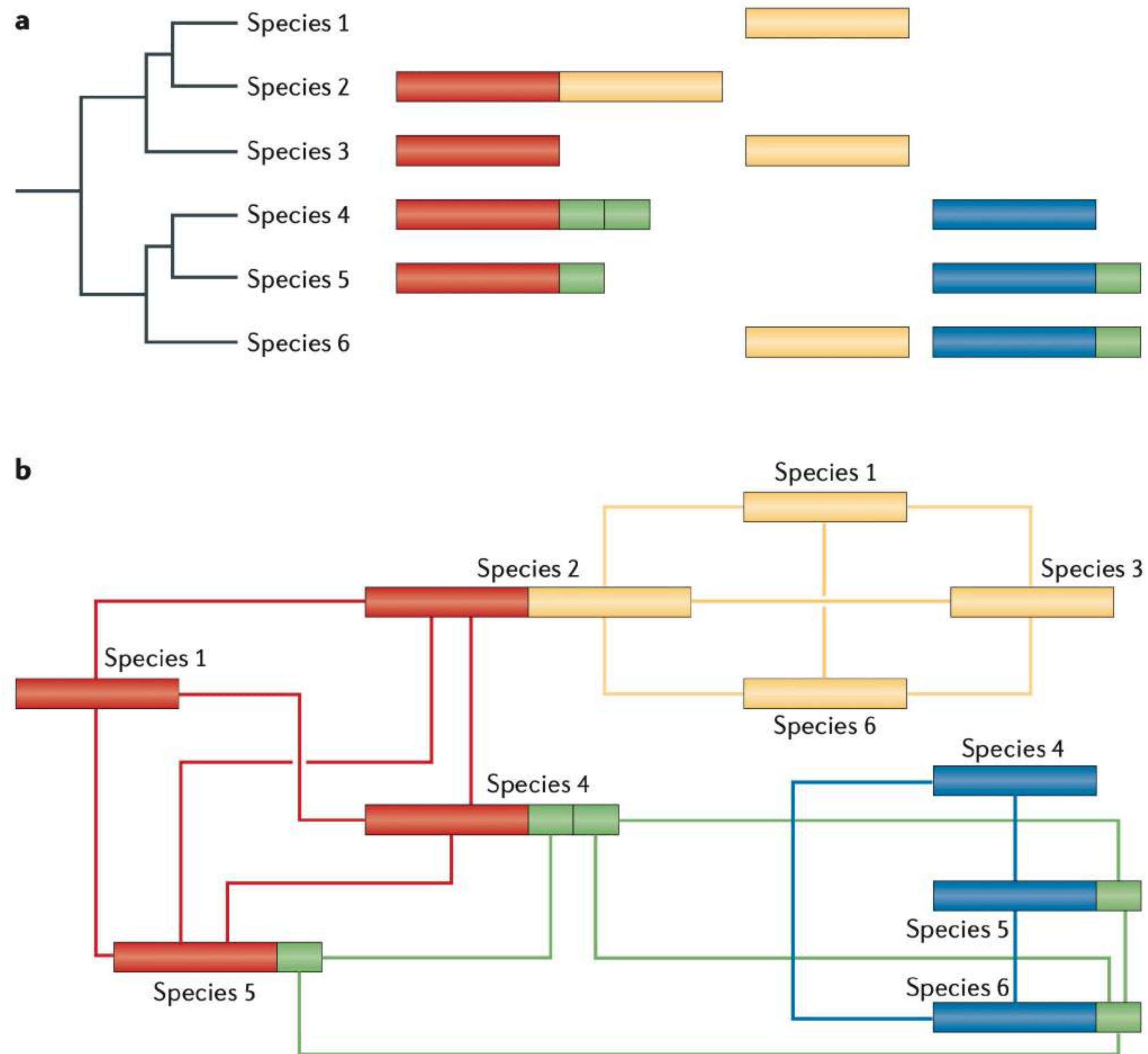
★ *Speciation*



Out-paralogs

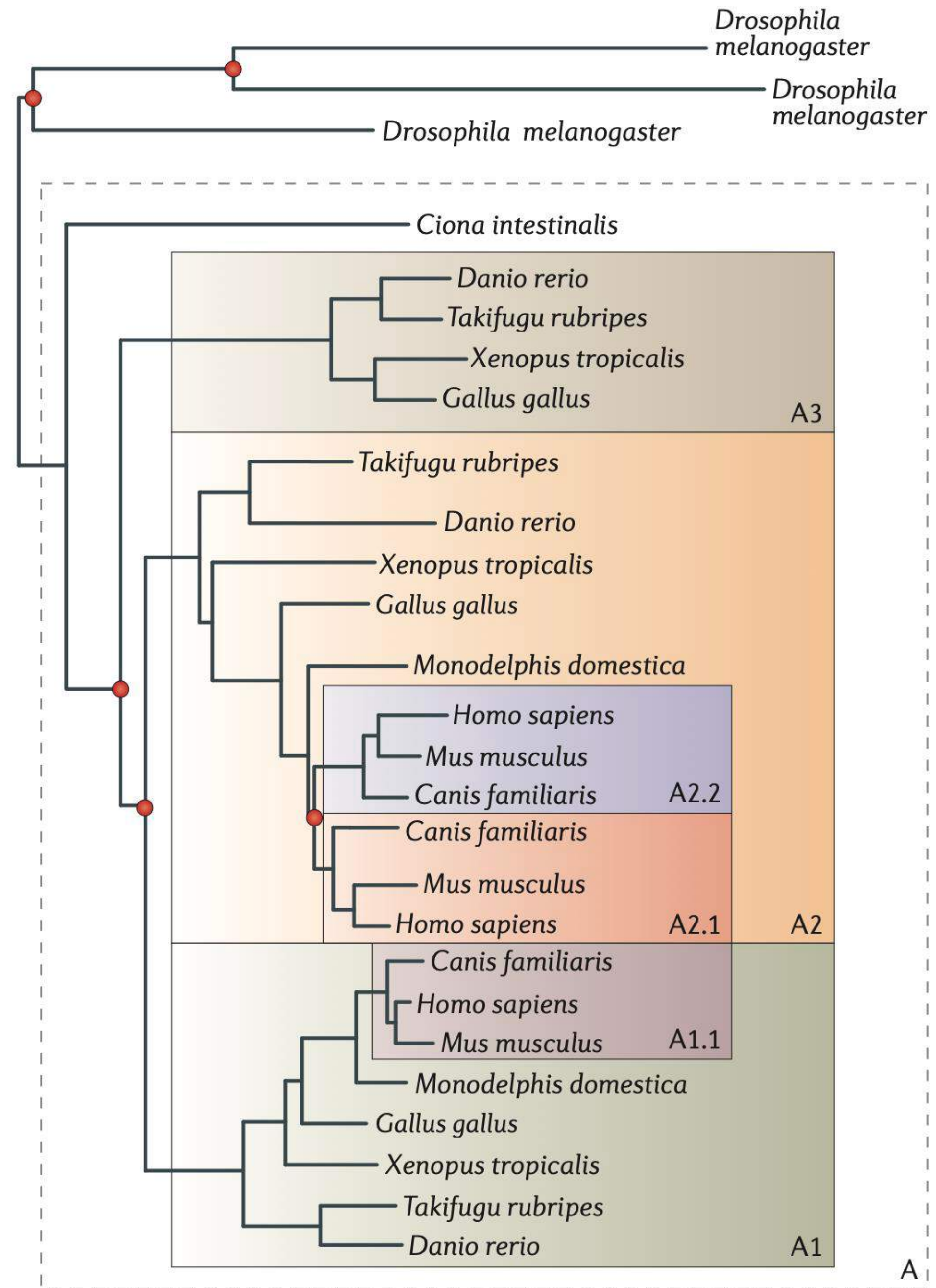


» Units of orthology »



» Real examples »

Transferrin family

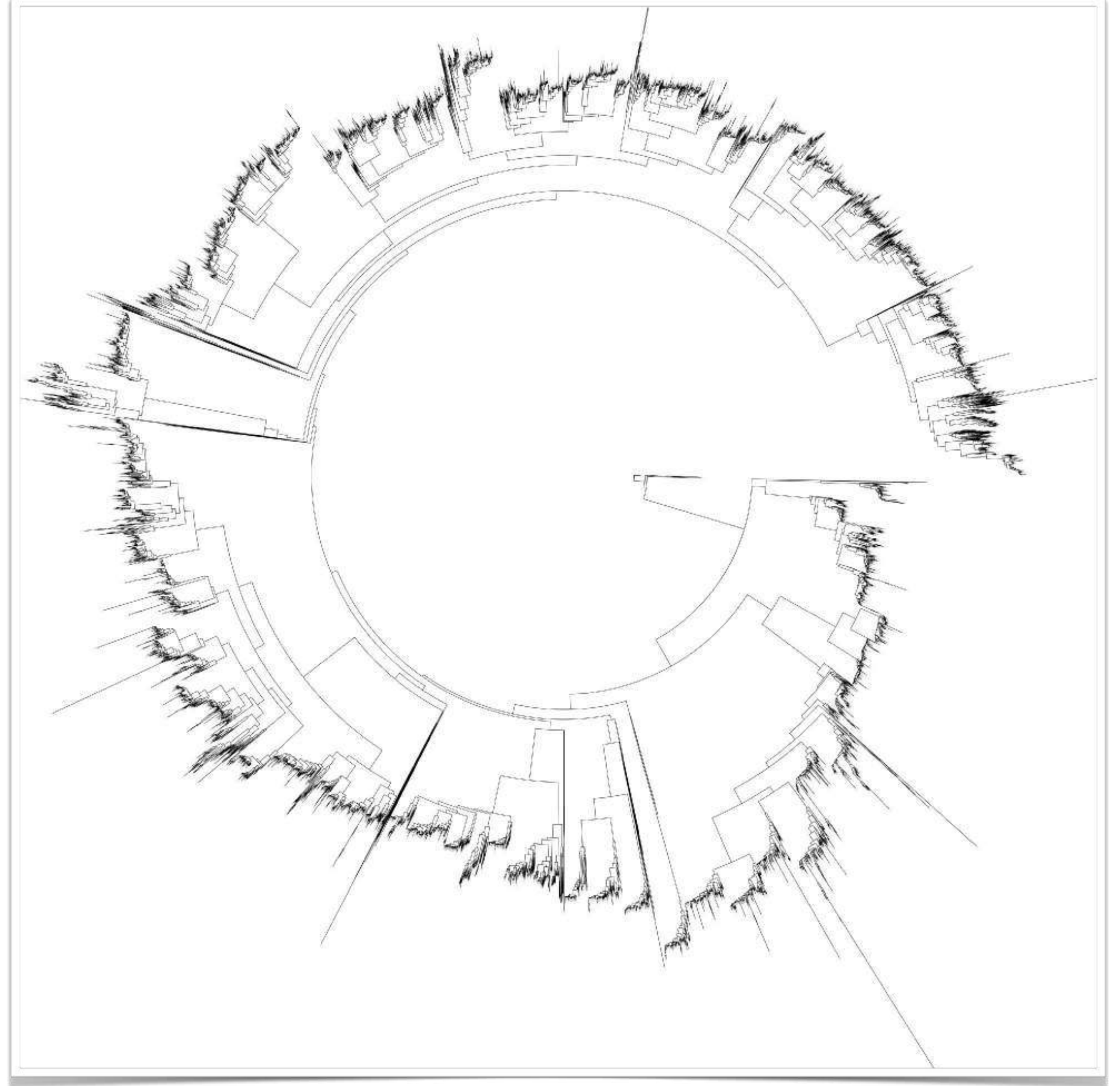


» Real examples »

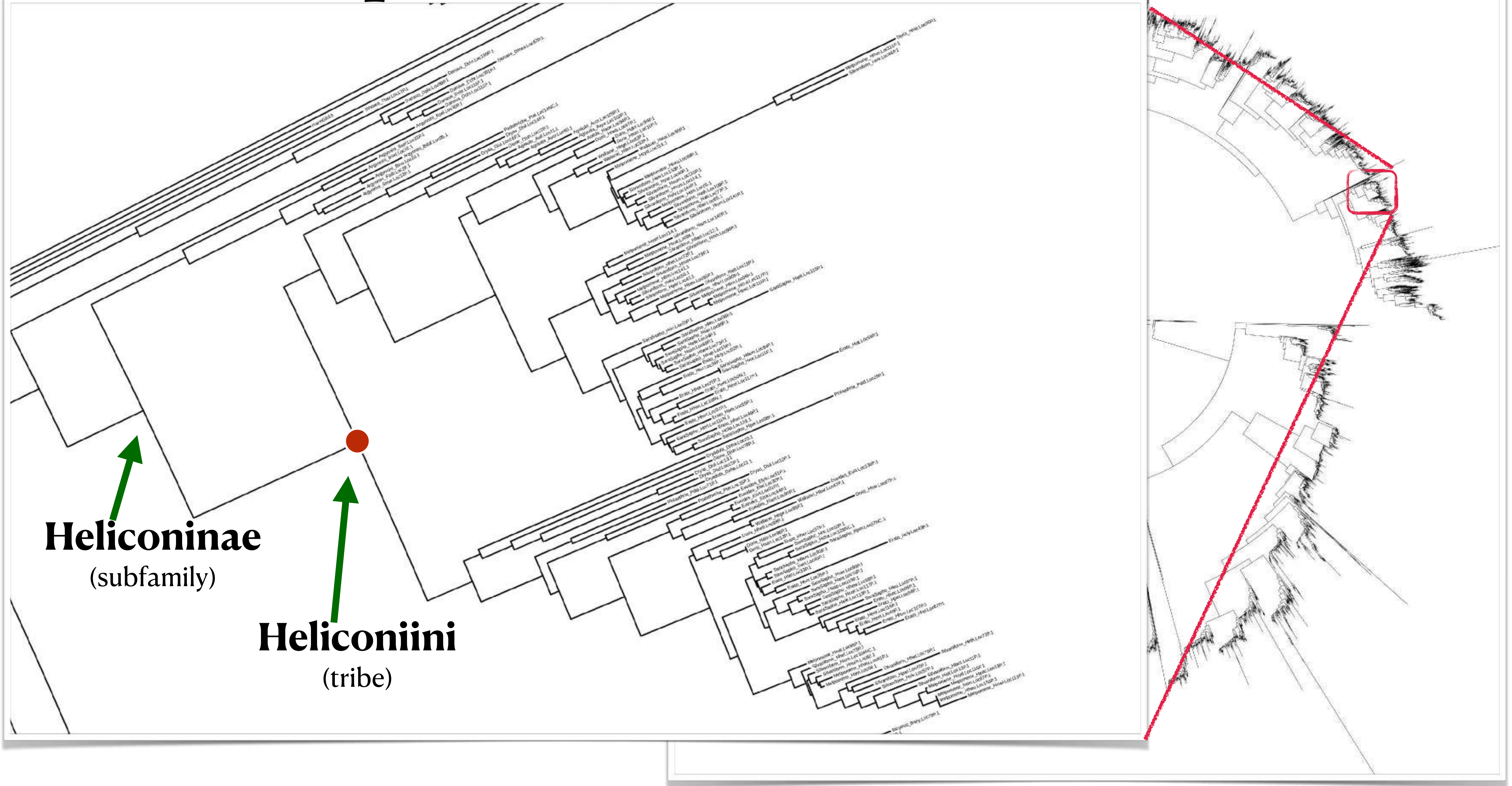
Gustatory receptors

8,320 tips

16,638 nodes

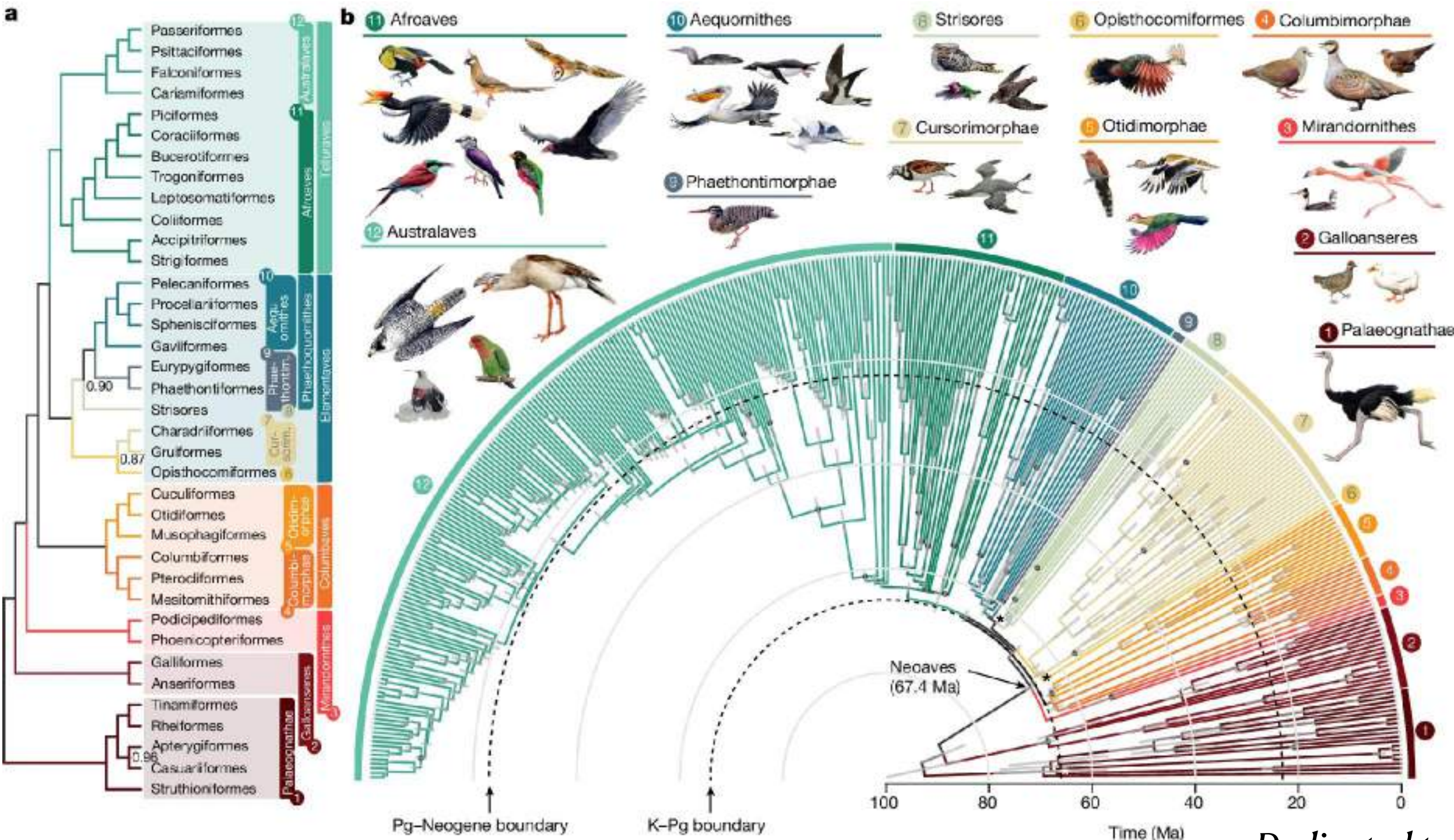
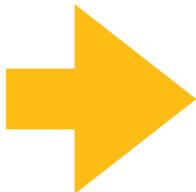
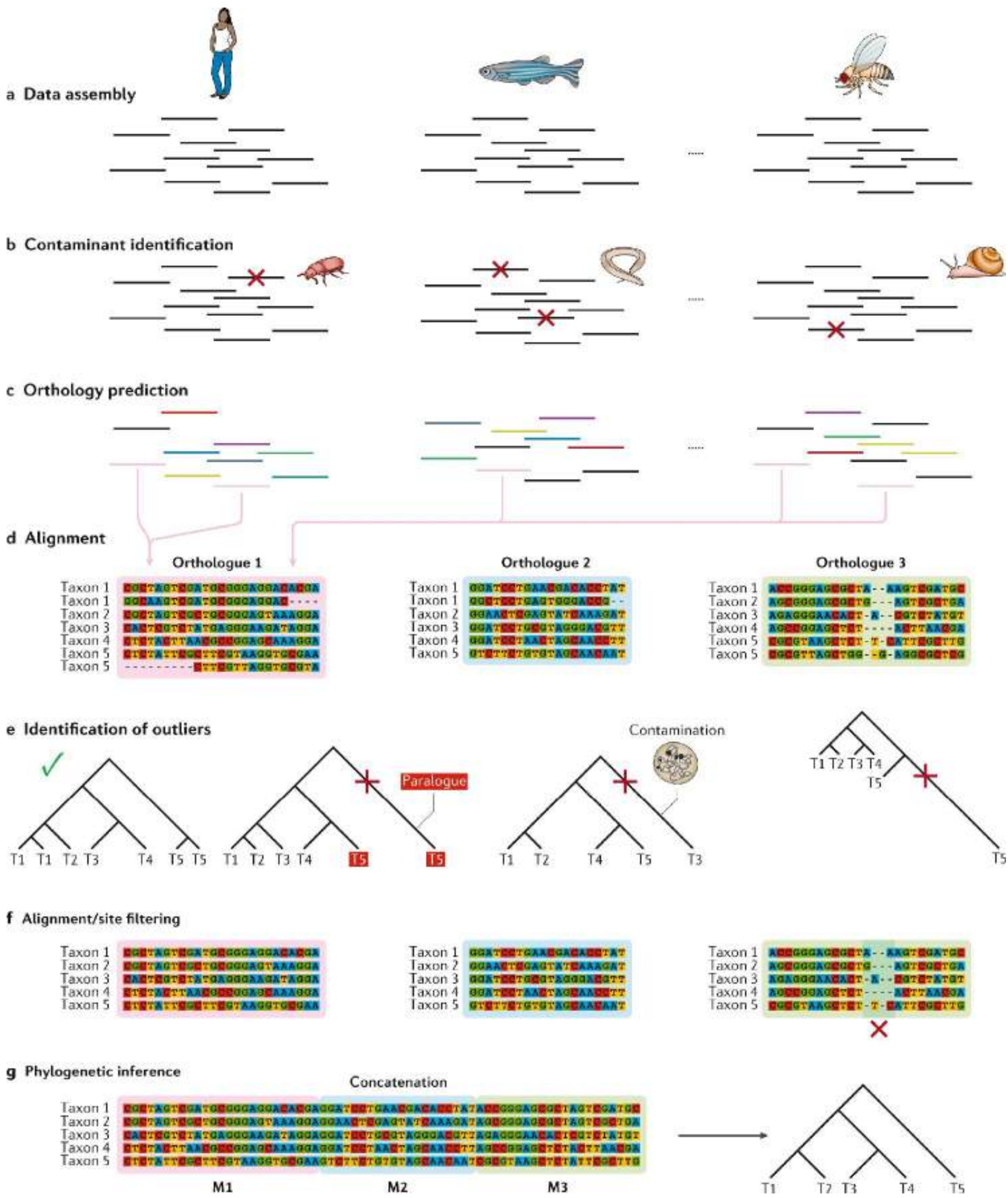


» Real examples »



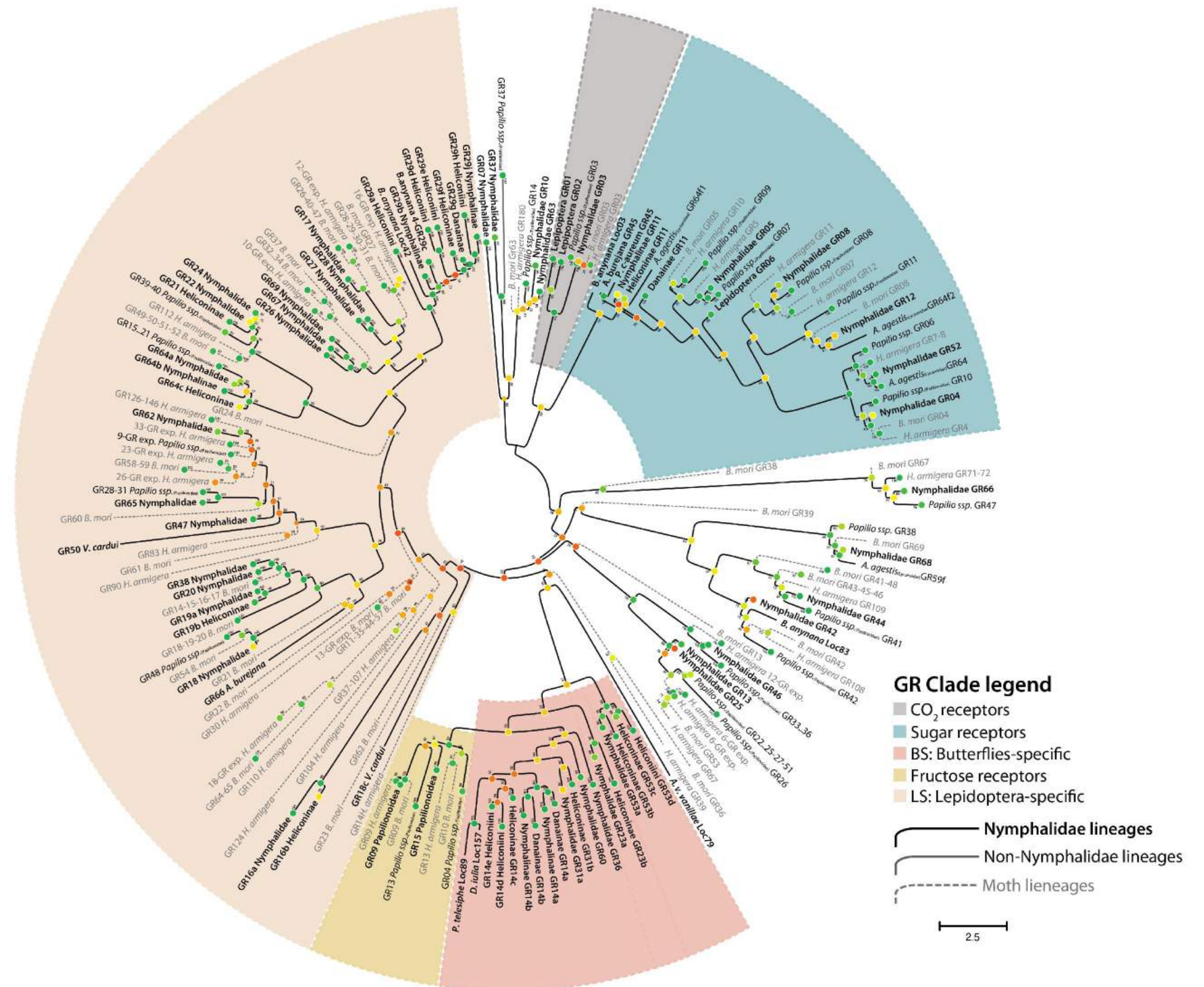
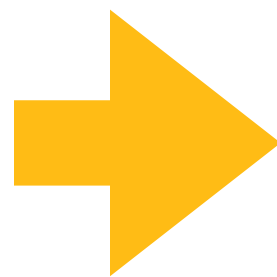
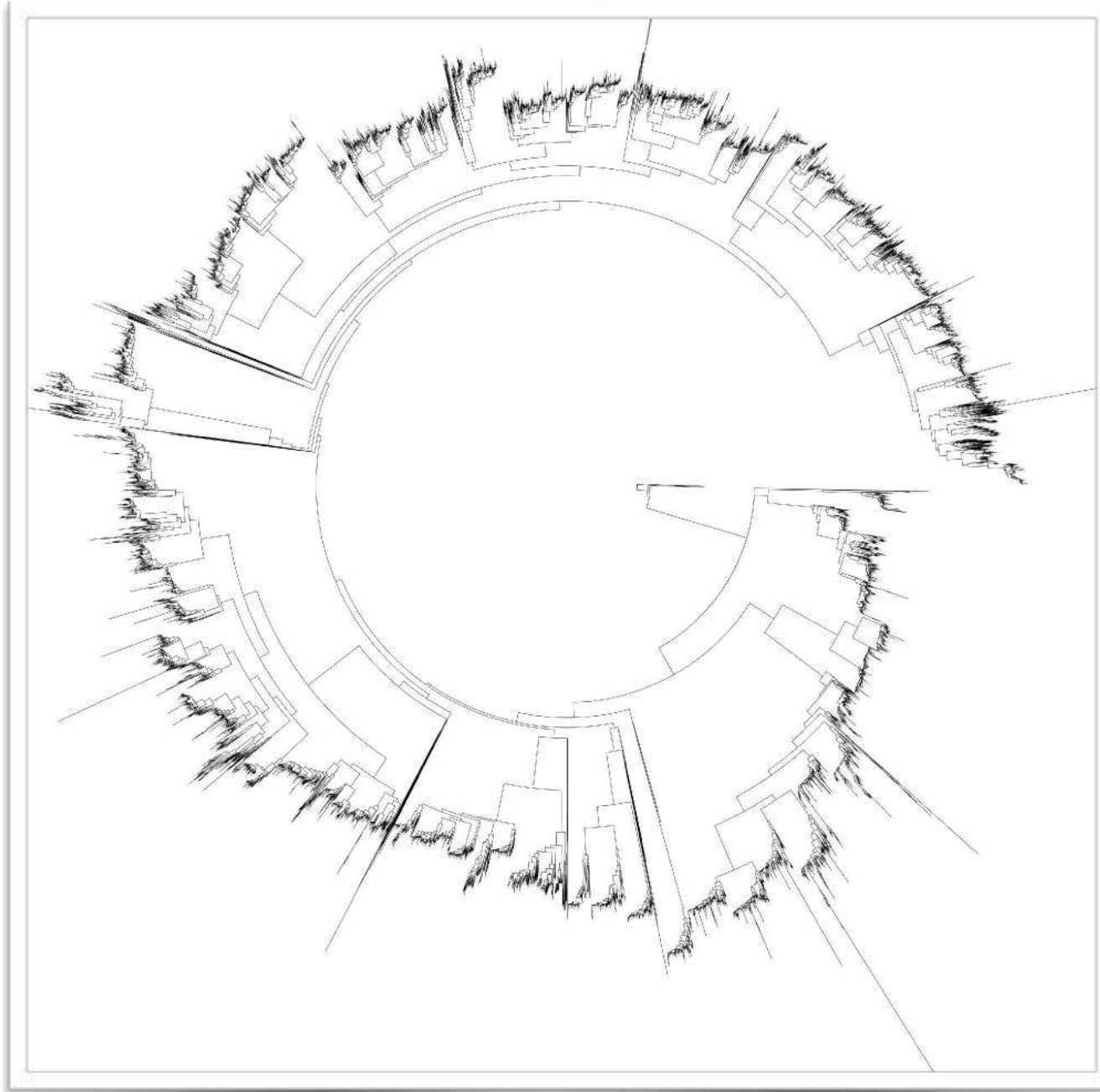
» **Why this is relevant ?** »

» Species Tree Estimation »

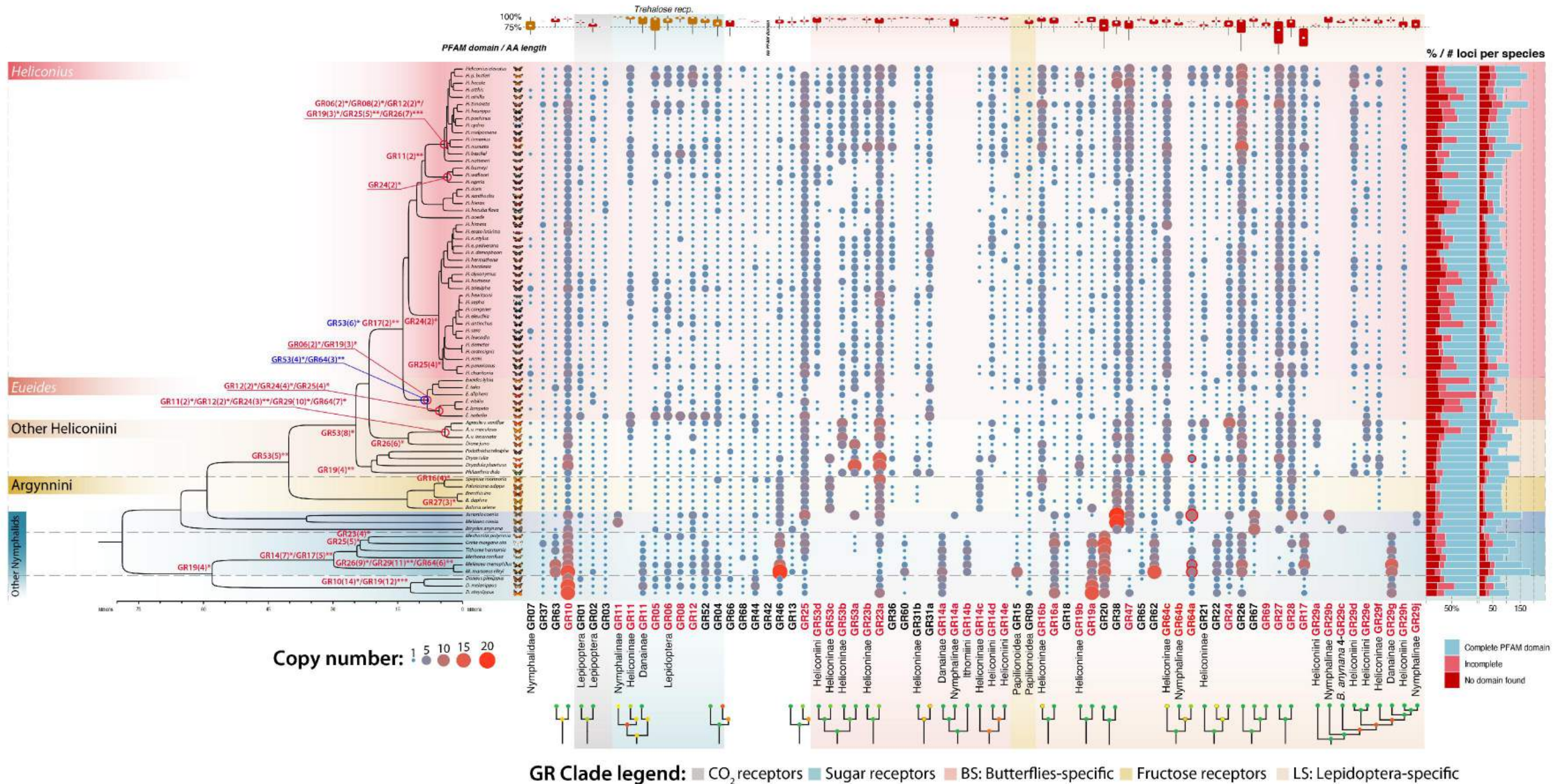


Dedicated to Joan

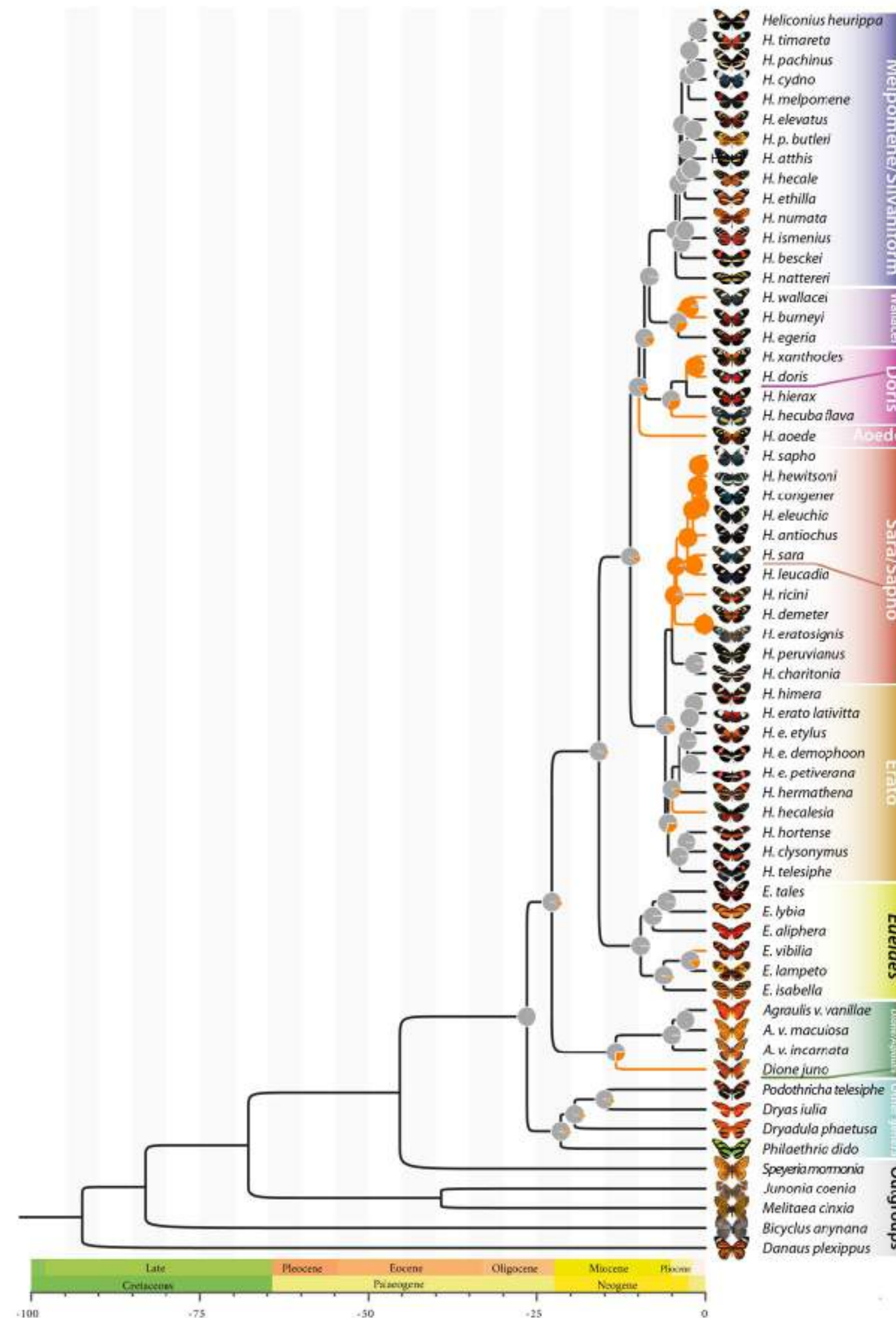
» Gene family reconstruction | expansions/contractions »



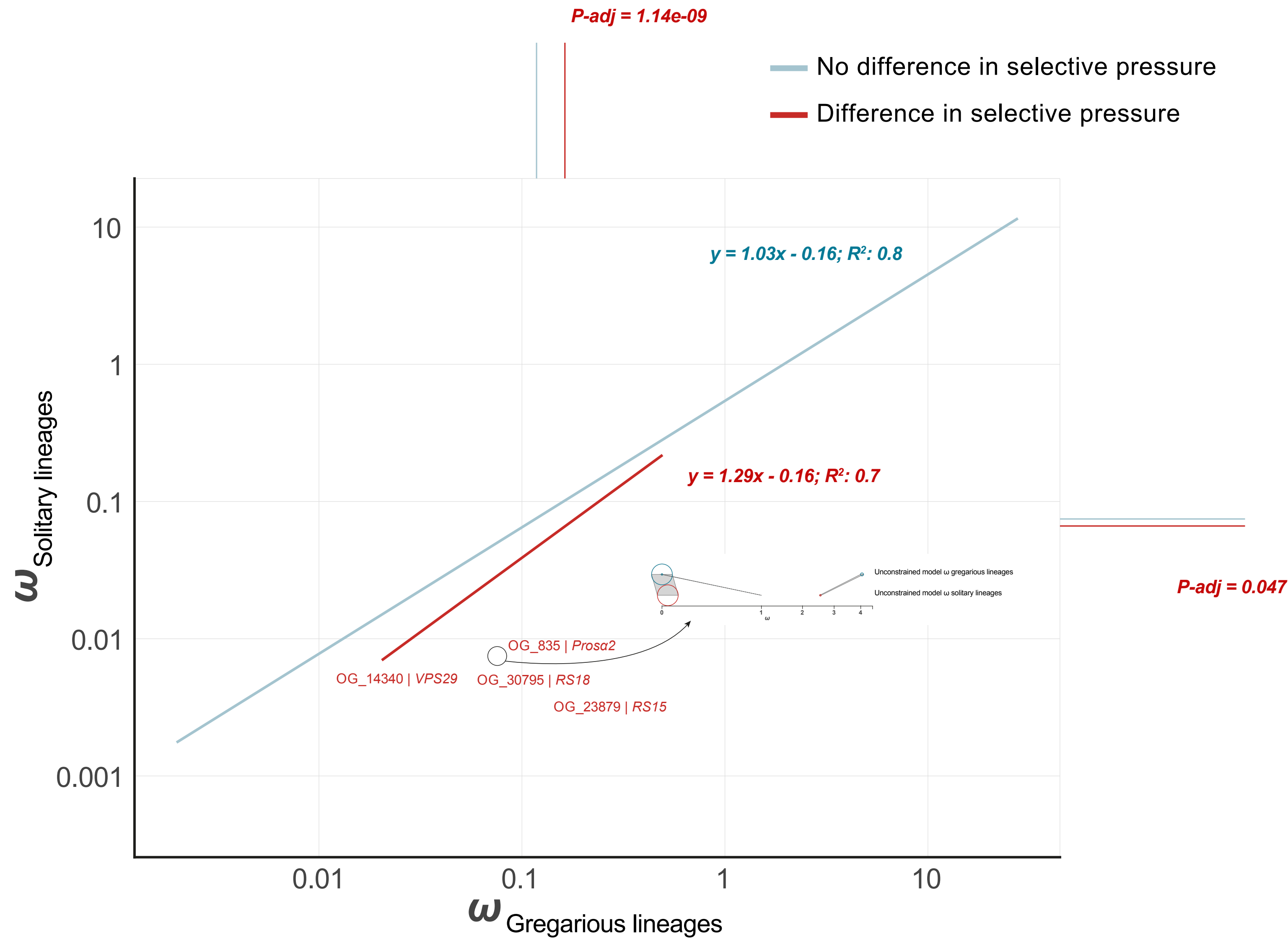
» Gene family reconstruction | expansions/contractions »



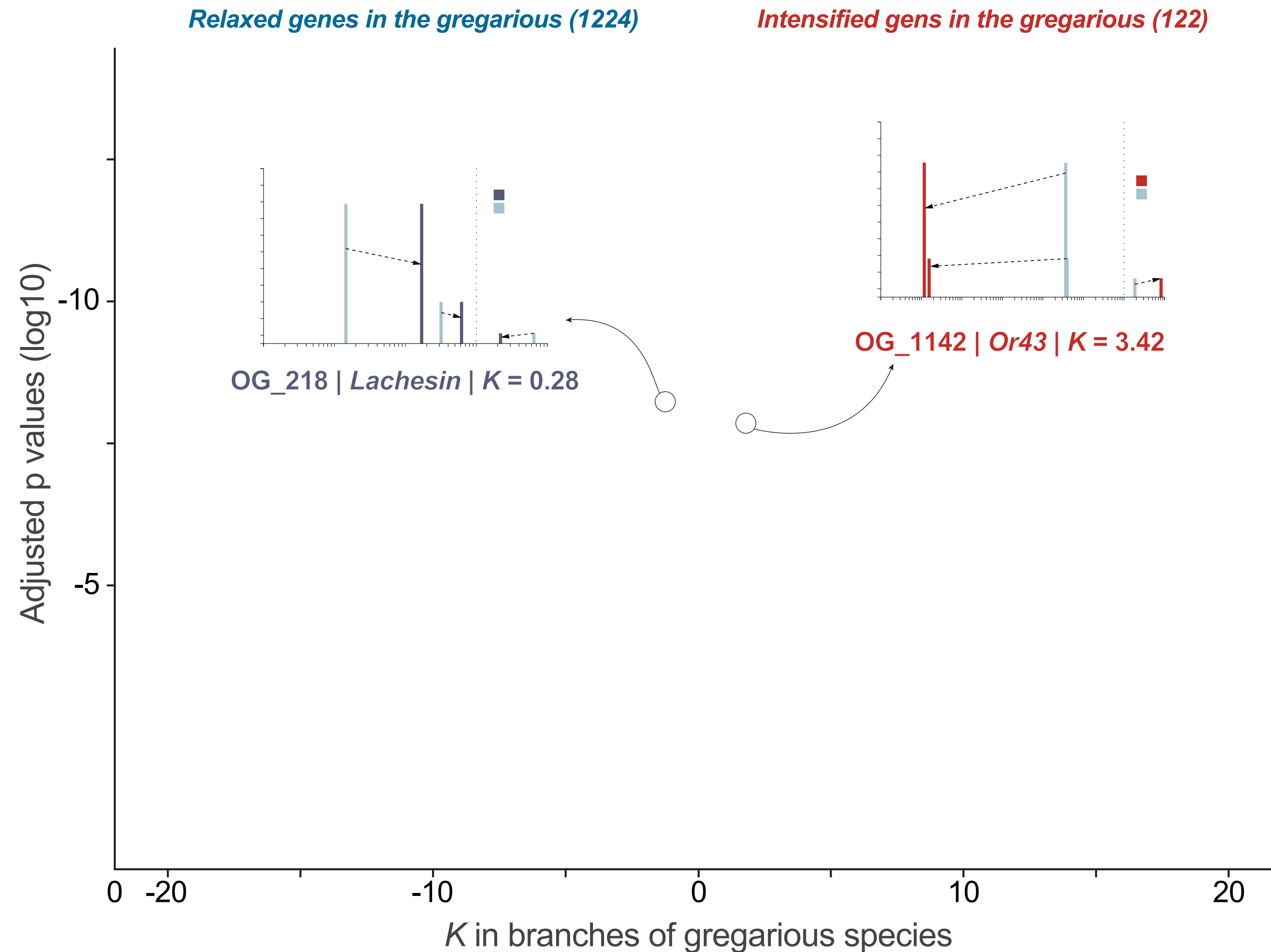
» Selecting pressures: Positive/Purifying/Convergence »



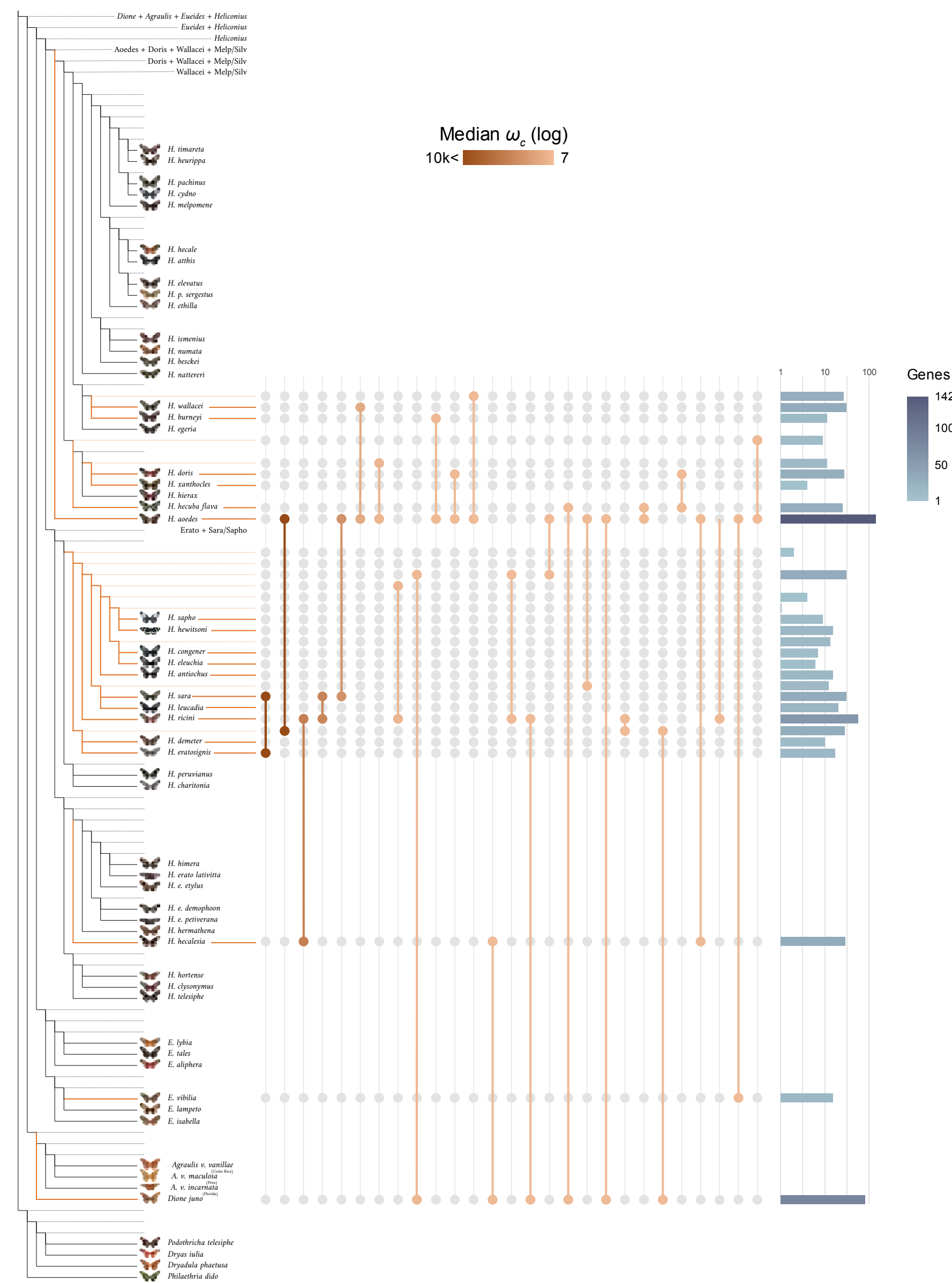
» Selecting pressures: Positive/Purifying/Convergence »



» Selecting pressures: Positive/Purifying/Convergence »

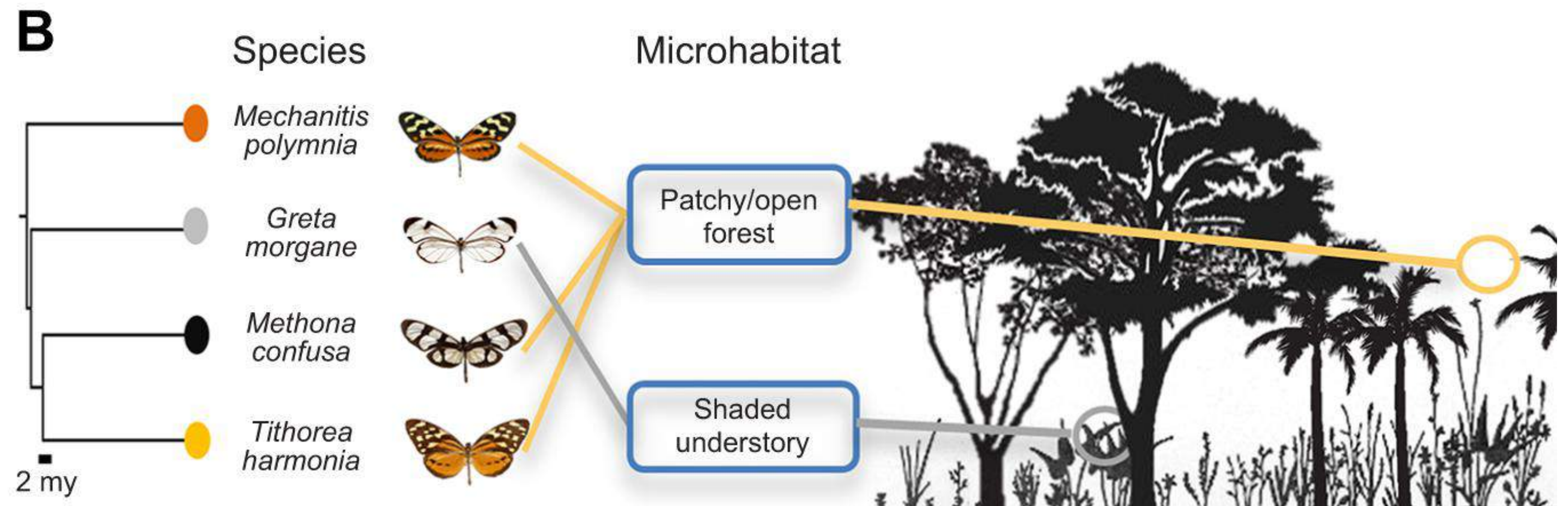
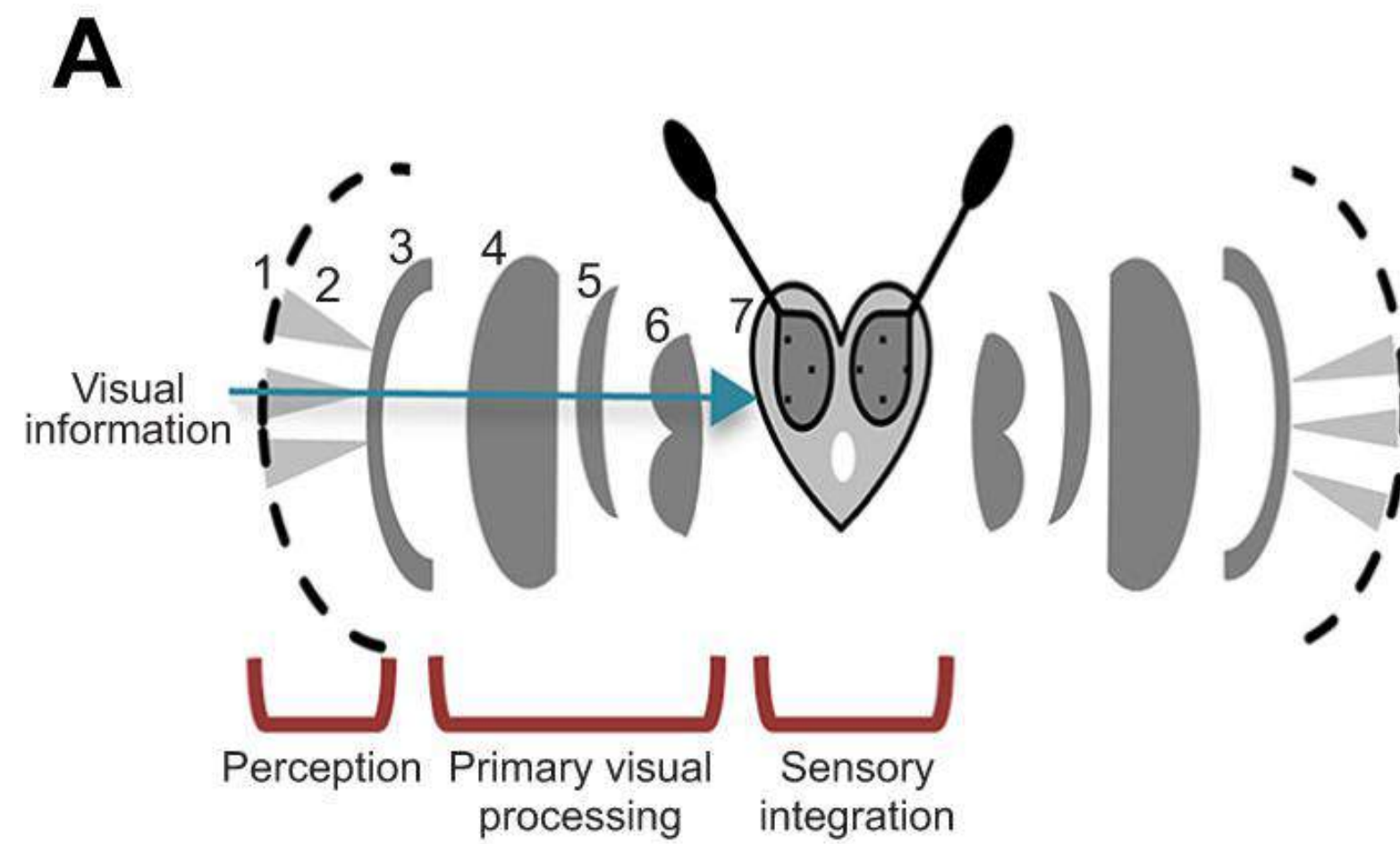


» Selecting pressures: Positive/Purifying/Convergence »

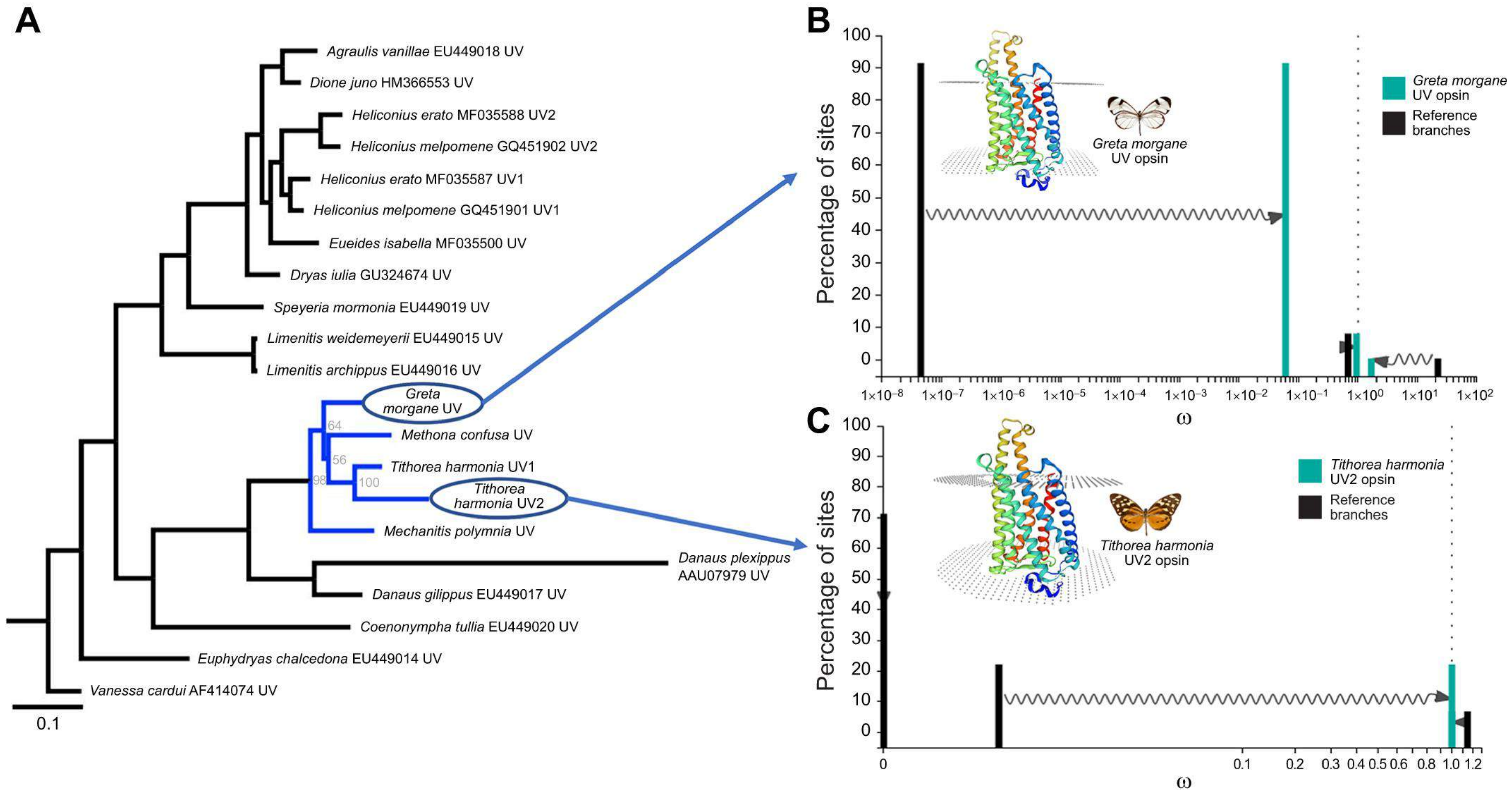


» Loss / Gain of function »

» Loss of function »



» Loss of function »

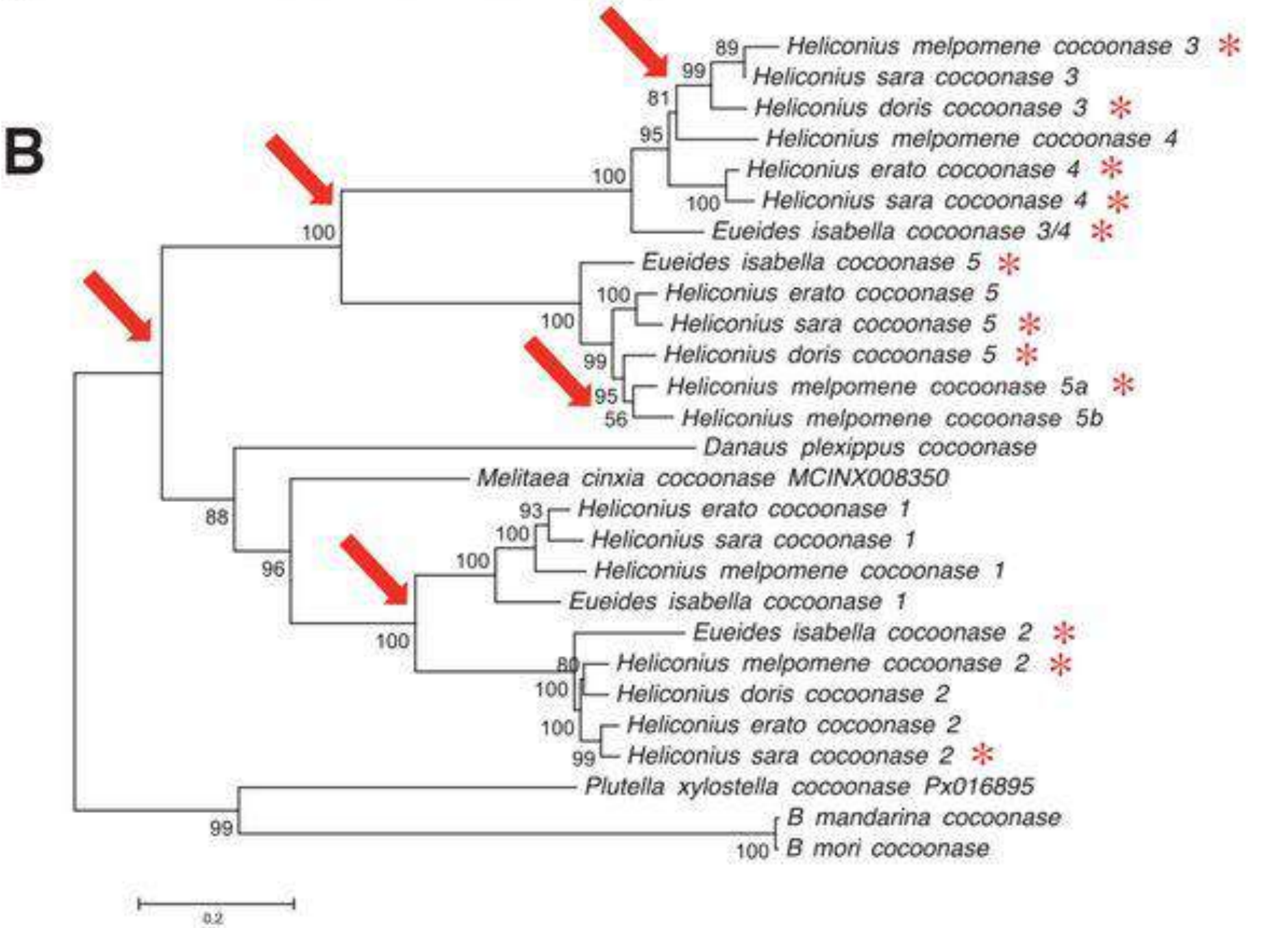
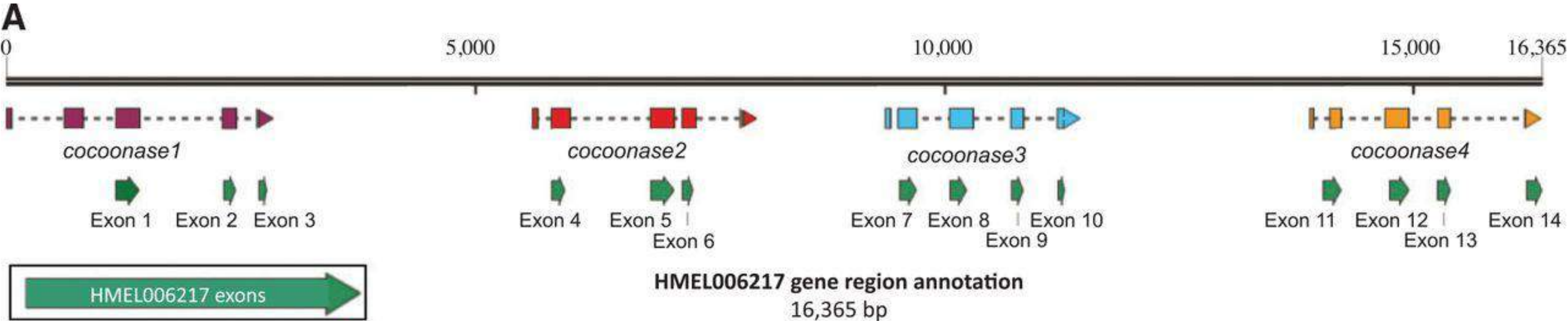


» *(Putative)* Gain of function »

Cocoonase: a protease secreted during the emergence of silk moths



» (Putative) Gain of function »

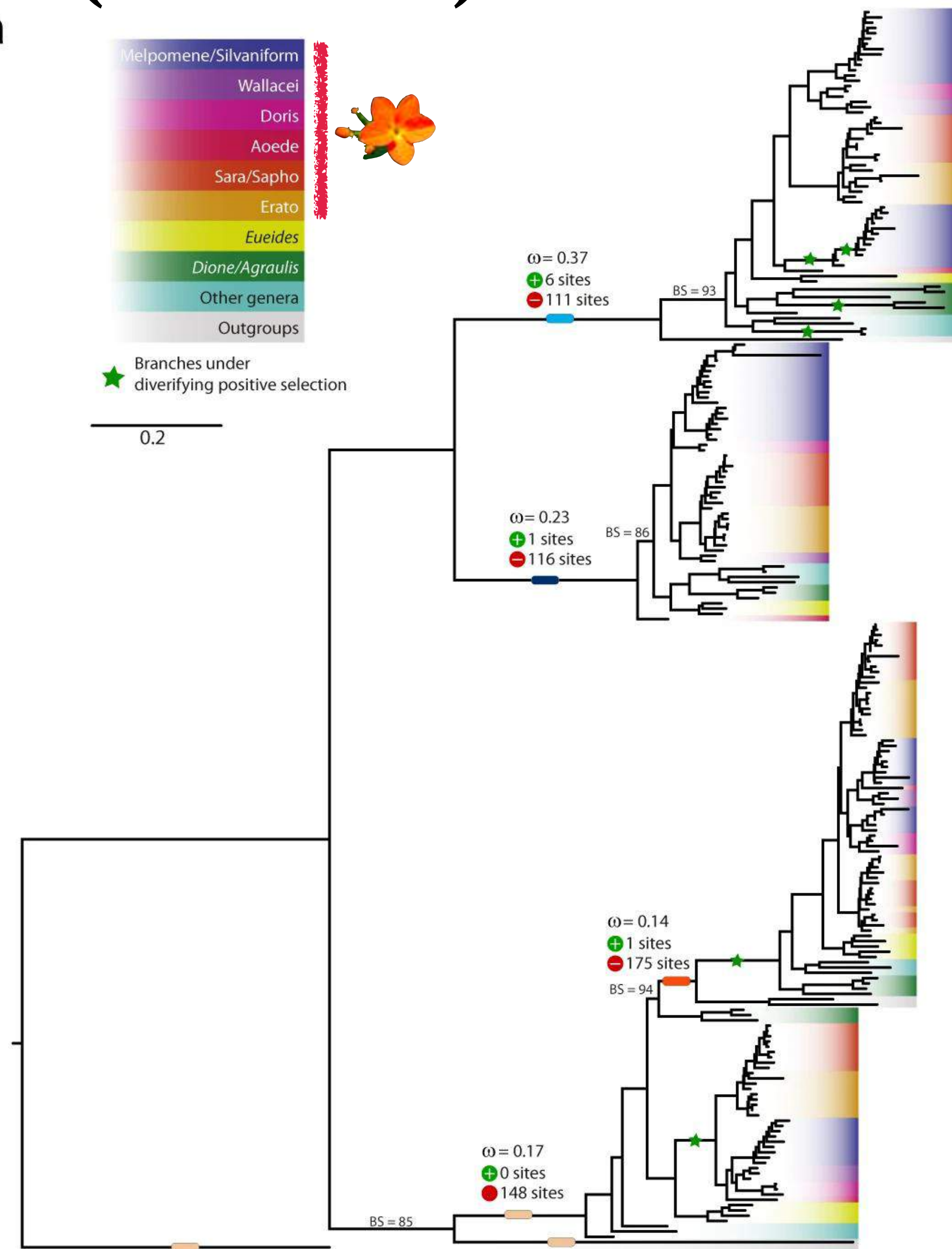


C

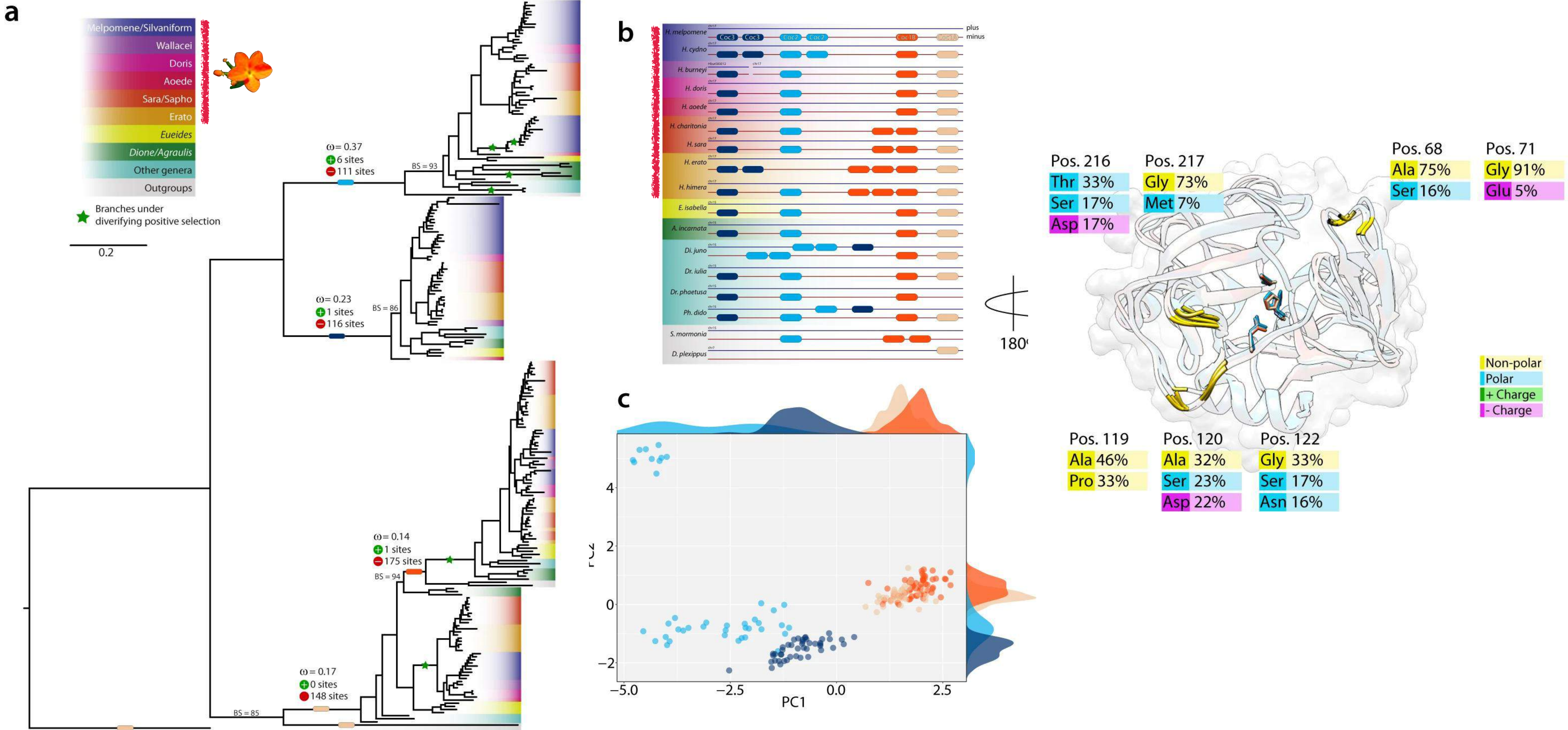
Gene name	Length (aa)
<i>cocoonase</i> (<i>Coc</i> ; <i>B. mori</i>)	260
<i>cocoonase 1</i> (<i>Coc1</i> ; <i>H. melpomene</i>)	262
<i>cocoonase 2</i> (<i>Coc2</i> ; <i>H. melpomene</i>)	262
<i>cocoonase 3</i> (<i>Coc3</i> ; <i>H. melpomene</i>)	264
<i>cocoonase 4</i> (<i>Coc4</i> ; <i>H. melpomene</i>)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

» (Putative) Gain of function »

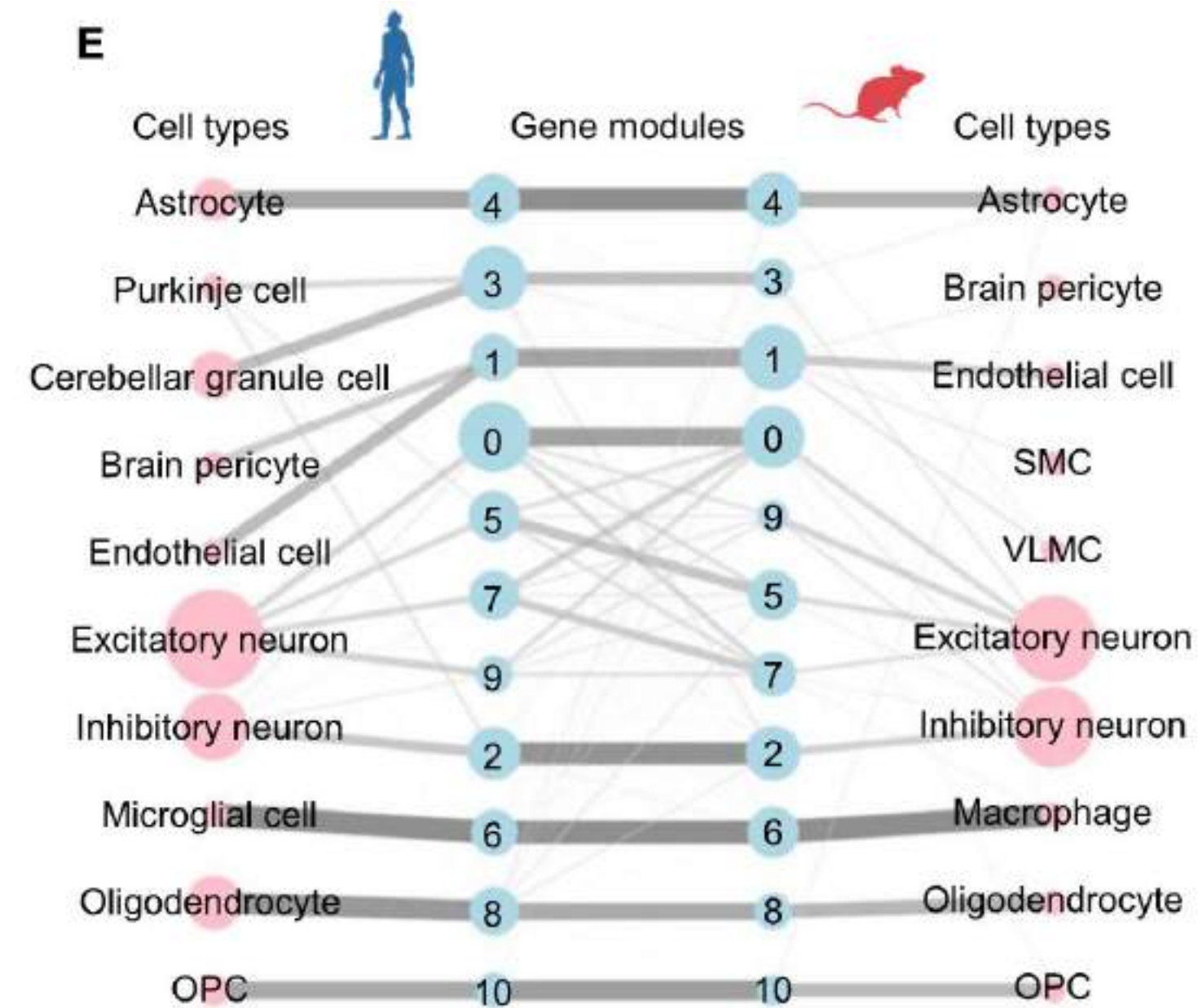
a



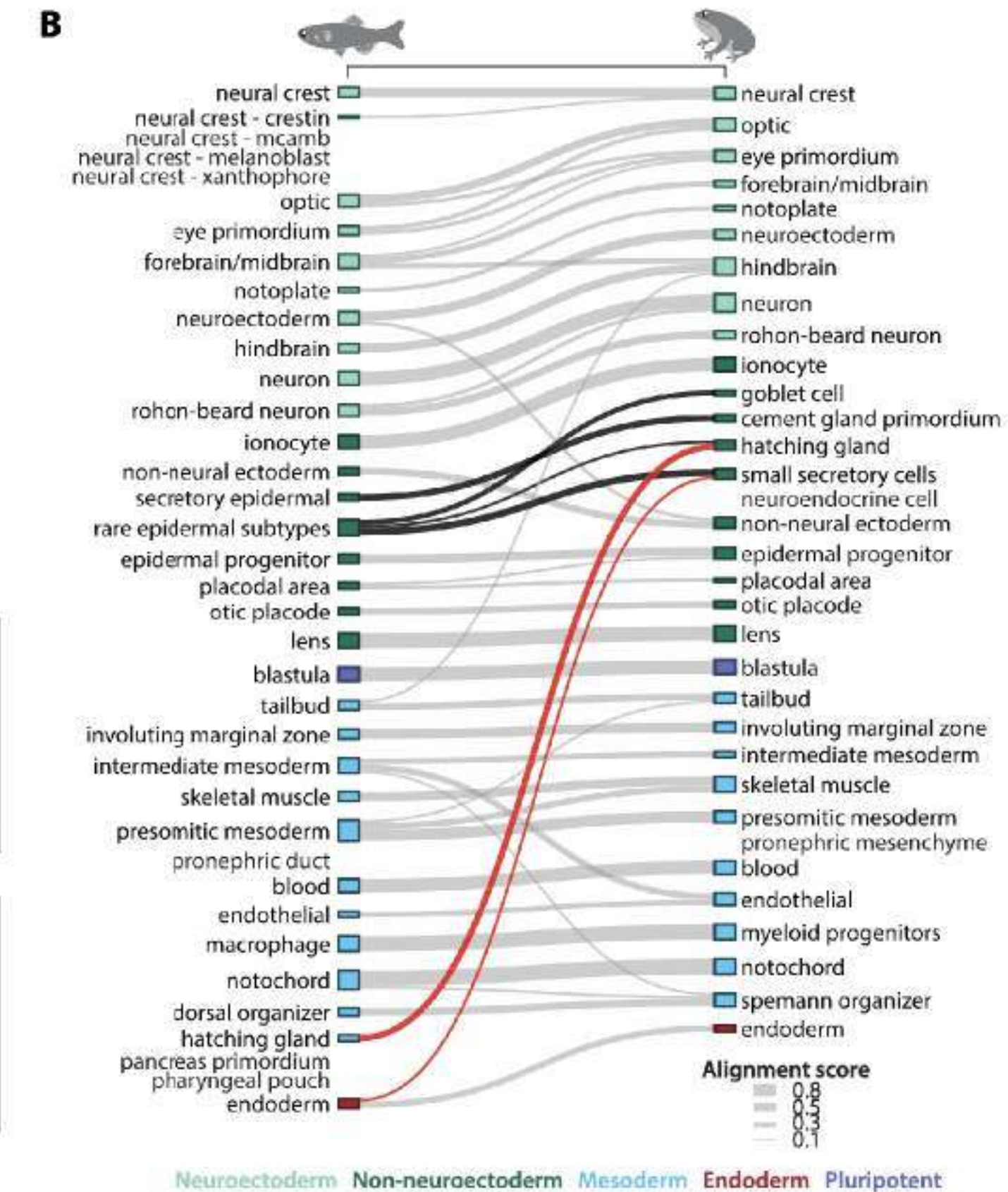
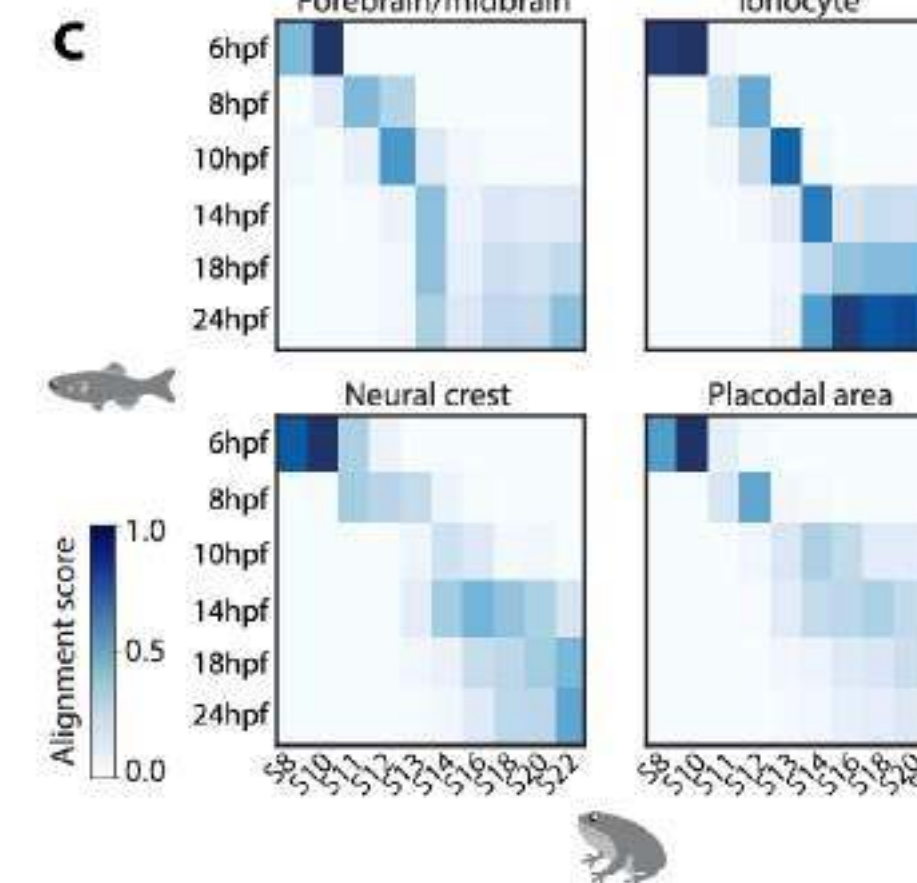
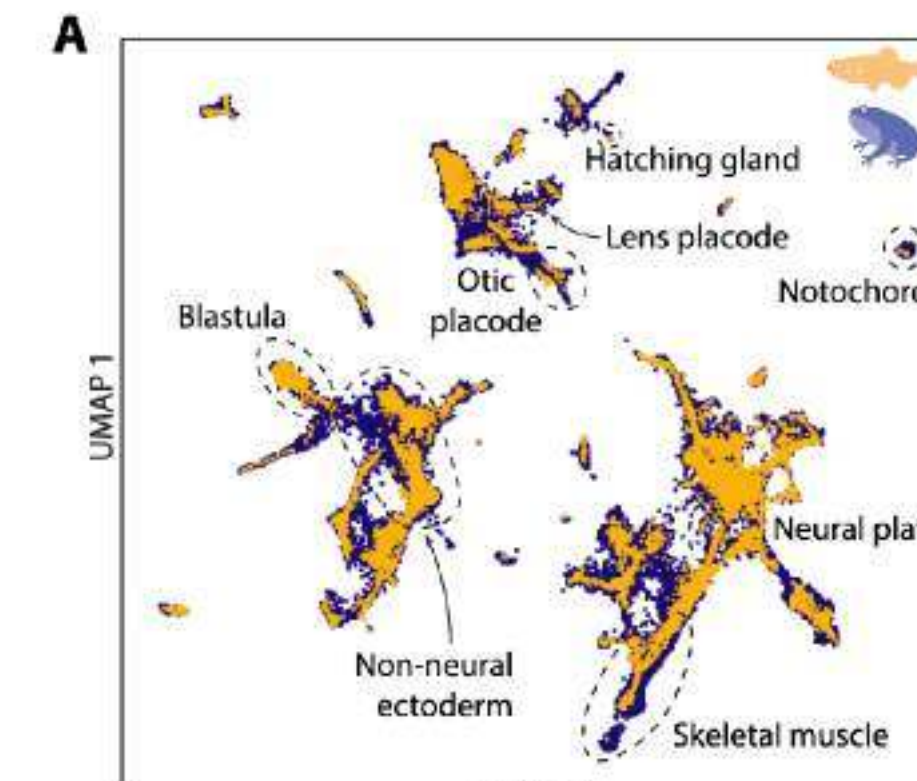
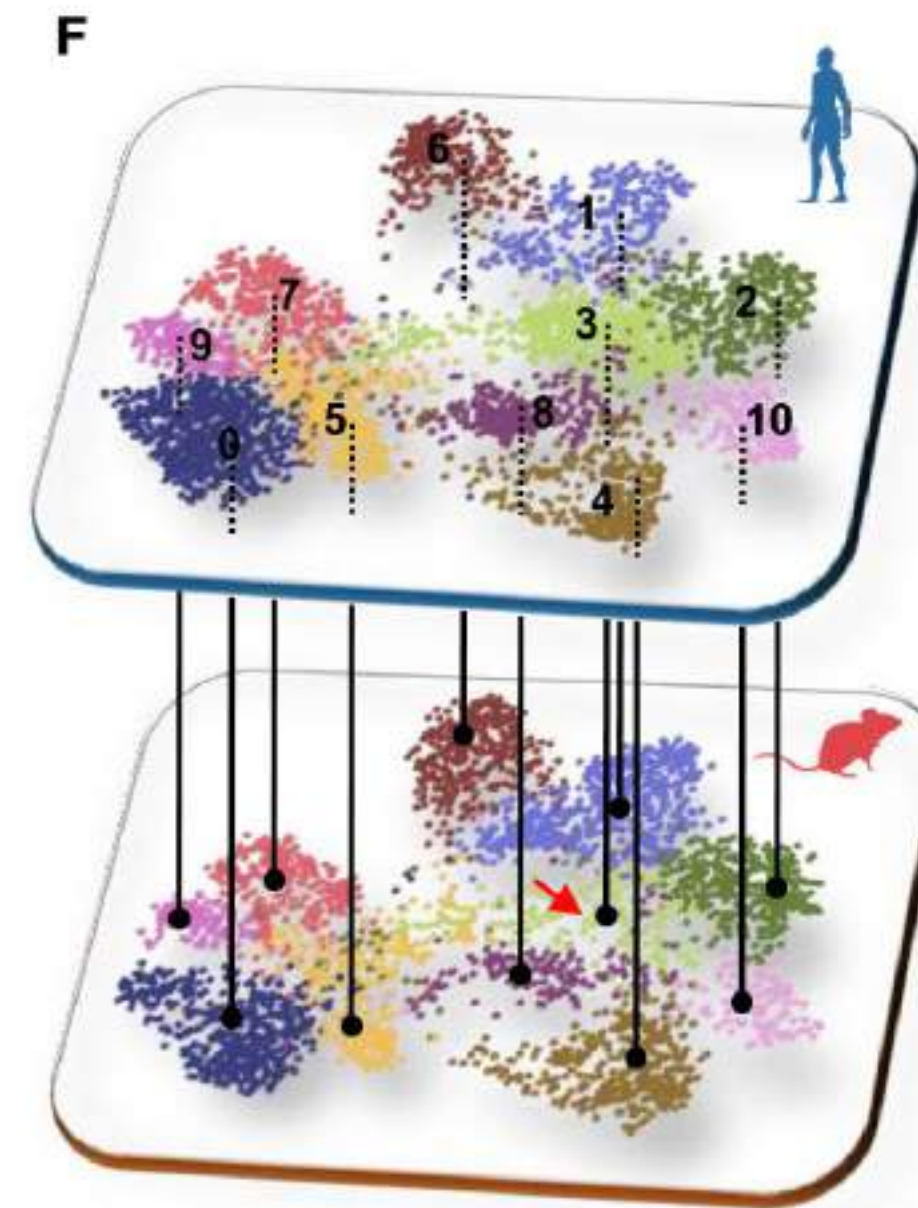
» *(Putative)* Gain of function »



» Single cell RNA data integration »



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



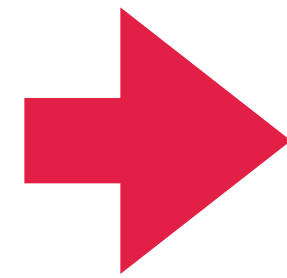
<https://elifesciences.org/articles/66747>

» **Questions ?** »

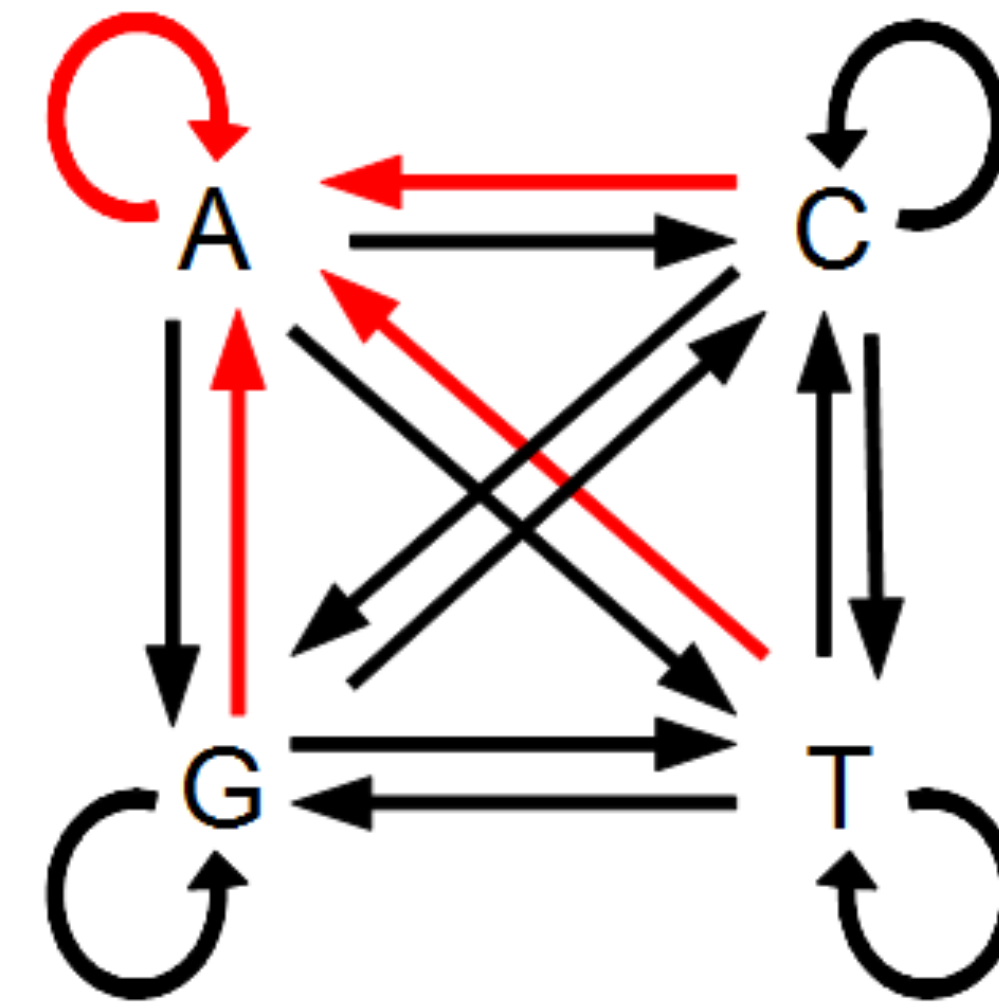
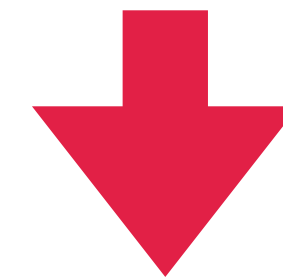
» **Methods** »

» Phylogeny on Homologous loci (Tree-based) »

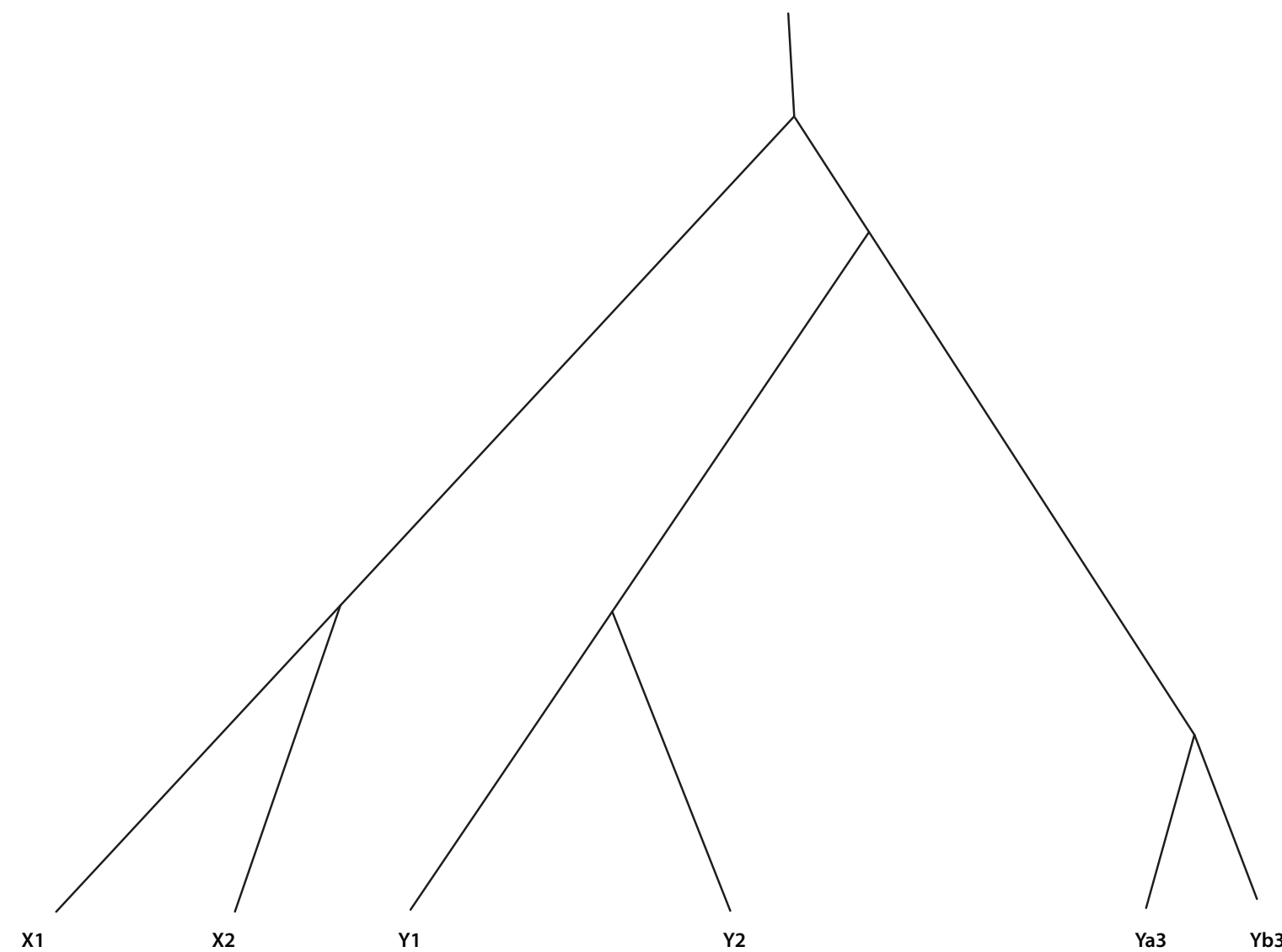
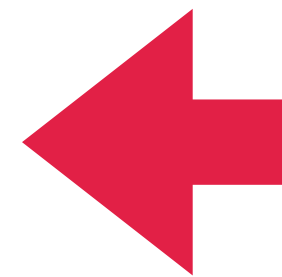
X1 ATGCTTAGGTCAGCTAGATAGTGGCTAATACCTAGCAGTTGAGTAA
Y1 ATGCTTAGGTCAGCATAGTGGCTAATACCTAGCAGTTGAGTAA
X2 ATGCTTAGGTCAGCTAGATAGTGGCTAATACCTAGCAGTTGAGTAA
Y2 ATGCTTAGGTCAGCATAGTGGCTAGTACCTAGCAGTTGAGTAA
Ya3 ATGCTTAGGTCAGCTAGATAGTGGCTAATACCTAGCAGTTGAGTAA
Yb3 ATGCTTAGGTCAGCATAGTGGCTAATACCTAGCAGTTGAGTAA



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



Substitution model



» Homology based-methods »

BLAST (Basic Local Alignment Search Tool) 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](#)

N
E
W
S

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



Nucleotide BLAST
nucleotide ► nucleotide

blastx
translated nucleotide ► protein

tblastn
protein ► translated nucleotide



Protein BLAST
protein ► protein

» 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*!!!

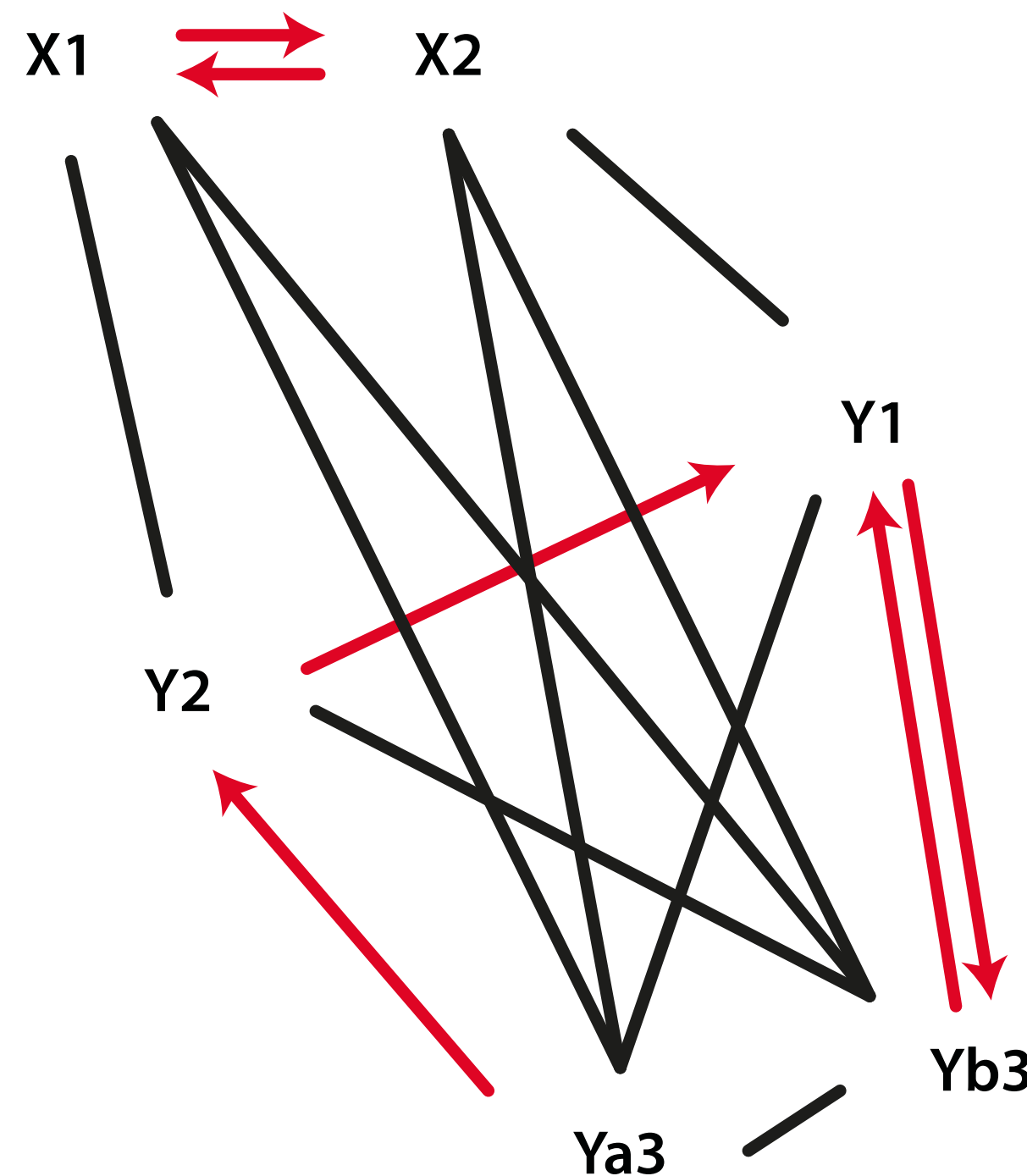
It is a **sequence similarity search tool**, and while it can help identify genes with similar sequences, it does not directly determine **orthology**



» Homology based-methods »

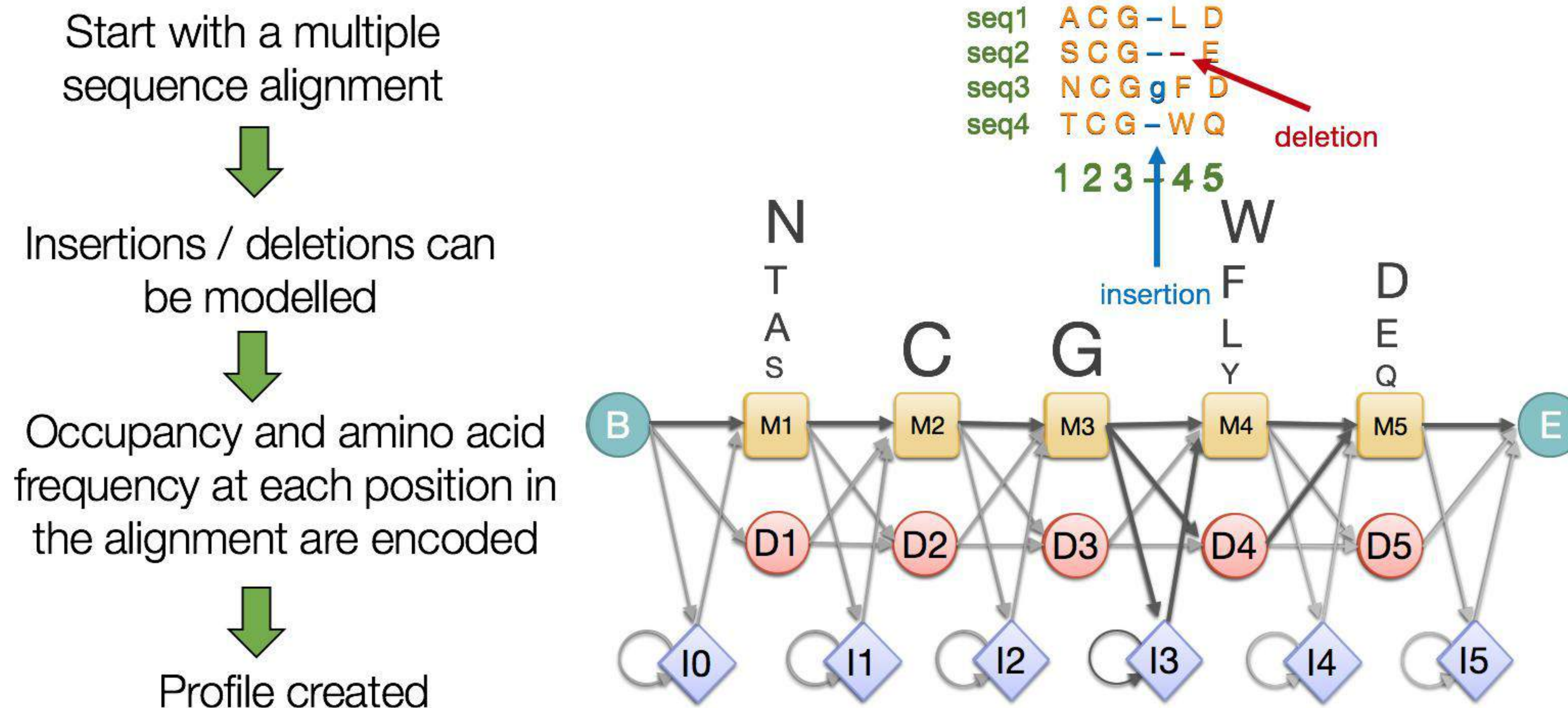
You can implement a **reciprocal Best Hits (RBH)** to identify two genes which are each other's best hits across species, but this is an approximation and not proof.

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.



» Hidden Markov Models (HMMs) »

NCBI

Conserved Domains

HOME SEARCH GUIDE NewSearch Structure Home 3D Macromolecular Structures Conserved Domains Pubchem BioSystems

Conserved domains on [lcl|seqsig_MMSLI_a8758a59b876fb7a7f8919e9245a7018] View Full Results

Local query sequence

Graphical summary ☐ Zoom to residue level show extra options

Query seq. Non-specific hits Superfamilies

7tm_6

7tm_6 superfamily

Search for similar domain architectures Refine search

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

lcl|seqsig_MMSLI_a8758a59b876fb7a7f8919e9245a7018 71 NIVKFLRNLYVLMSTIVAKVSSYMLWQEKWEDIQYVNGVSIERLKTND EISIKINKNLKYCKIVTYIYIYLTSTTA 150
Cdd:pfam02949 1 NLTEALTNLQVLVTSIVCLFKLLILKKRKEIVELISLLRELDPRCL EEEEEERIFERLVATASRVRFLFIYIYLAYALF 80

lcl|seqsig_MMSLI_a8758a59b876fb7a7f8919e9245a7018 151 LLVFLQPIIYCLRKYSsleddldkyiPITNTWVPFDKNDLTV-YVPSLIYQTYATLYSVLYIVAFDSNAVSTMVFFRGE 229
Cdd:pfam02949 81 CISALVSMVLSLERGLP-----LPYYLPFPFDWEASRPpYYIAYIQVLGMTVTVLQNIGSDSLFCSFLLFLVLH 150

lcl|seqsig_MMSLI_a8758a59b876fb7a7f8919e9245a7018 230 LE-LIRRDQILFGANTLWNEETARnnLIKQKRHQDLVKYVKLFDSCSPIMLYVIVCSVMLCATAYQLTVETNAMQK 308
Cdd:pfam02949 151 LRLLQLRLERLGLDPELSEEEYQE--LRECVRDHQRLRLVDELNRIFSVPILVQFLLSGLVICFLLFNLLFFSGPLQG 228

lcl|seqsig_MMSLI_a8758a59b876fb7a7f8919e9245a7018 309 FITAEYLIFGVSQLFMFCWHSNEVLVYVKDLTLGPYESMwWTRSVSEQKDISILIDQFNKQIVFSAGPFTSITVATFISI 388
Cdd:pfam02949 229 IYYVLYLLAMLILQFPYCYYGNELTEASEELANAIYSSNwyDASRRYRRTLLIFIMRAQRITIKAGGFPIISLPFTFISI 308

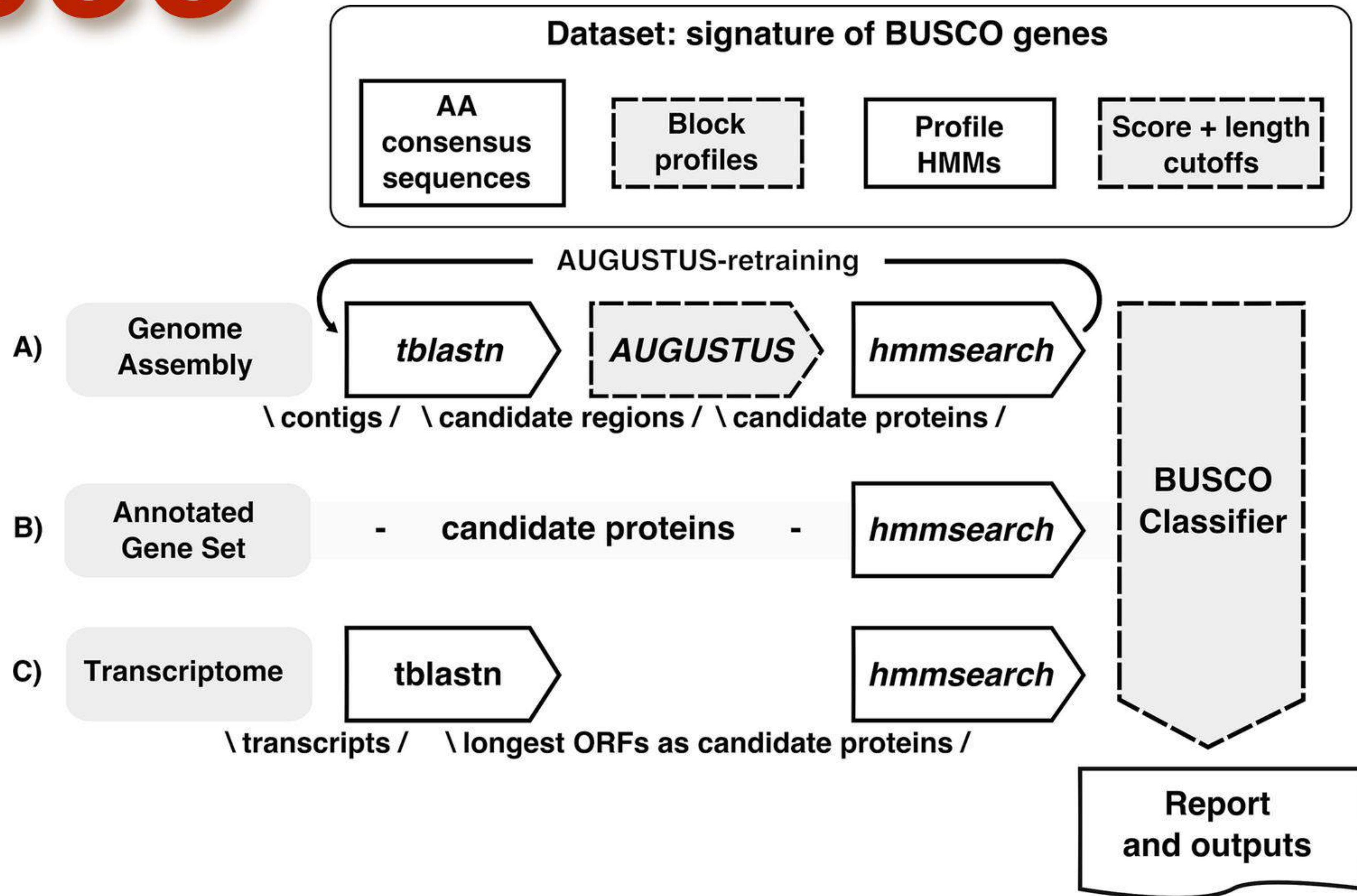
lcl|seqsig_MMSLI_a8758a59b876fb7a7f8919e9245a7018 389 LKGAY 393
Cdd:pfam02949 309 LKFAY 313

Blast search parameters

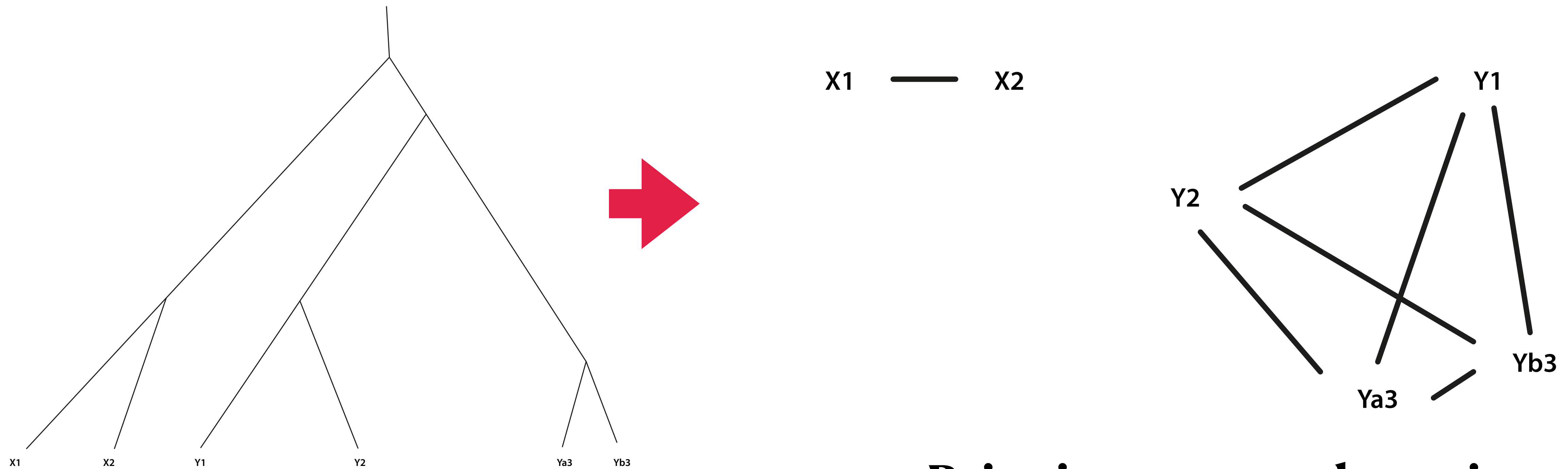
Data Source: Live blast search RID = SCZD6J5N016

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

BUSCO



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



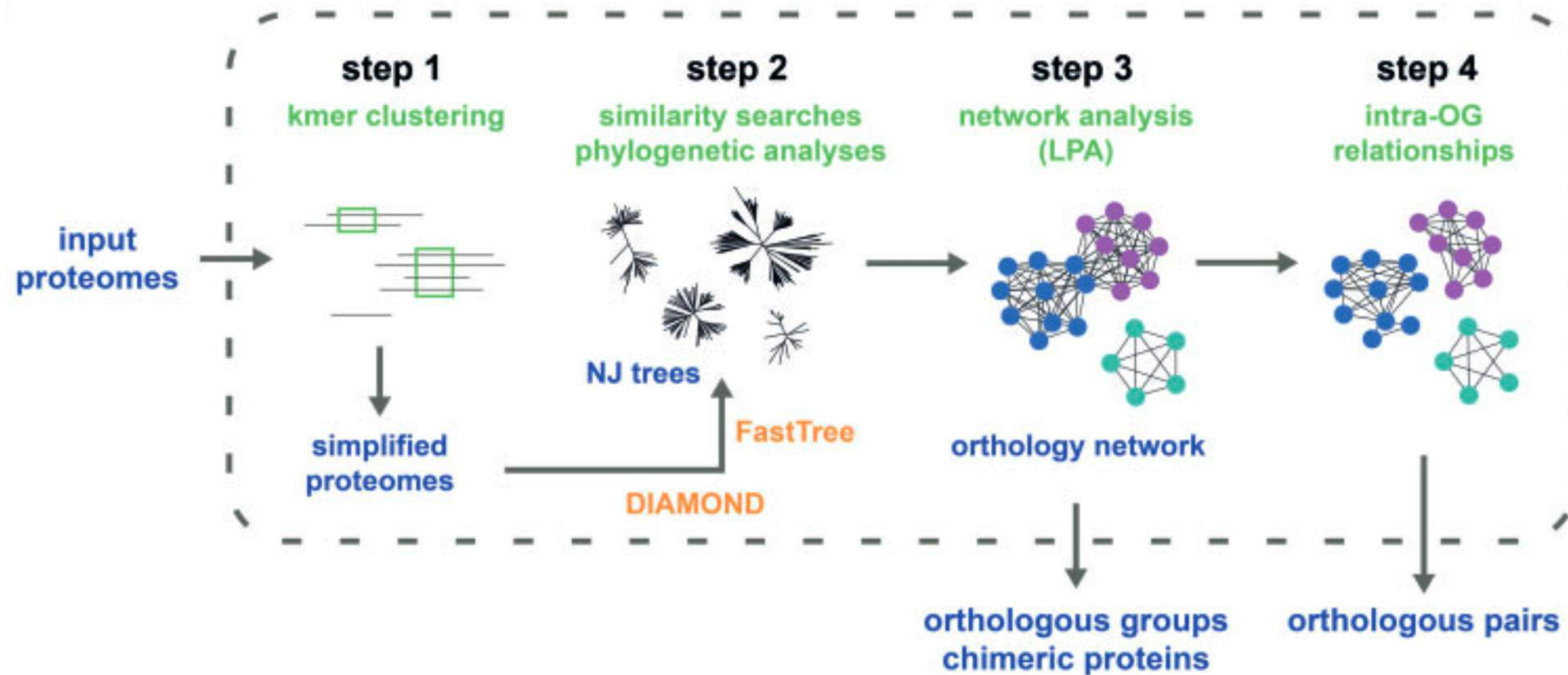
Pairwise generated metric

E-value: the number of alignments you would expect to find by chance in a database of a given size.

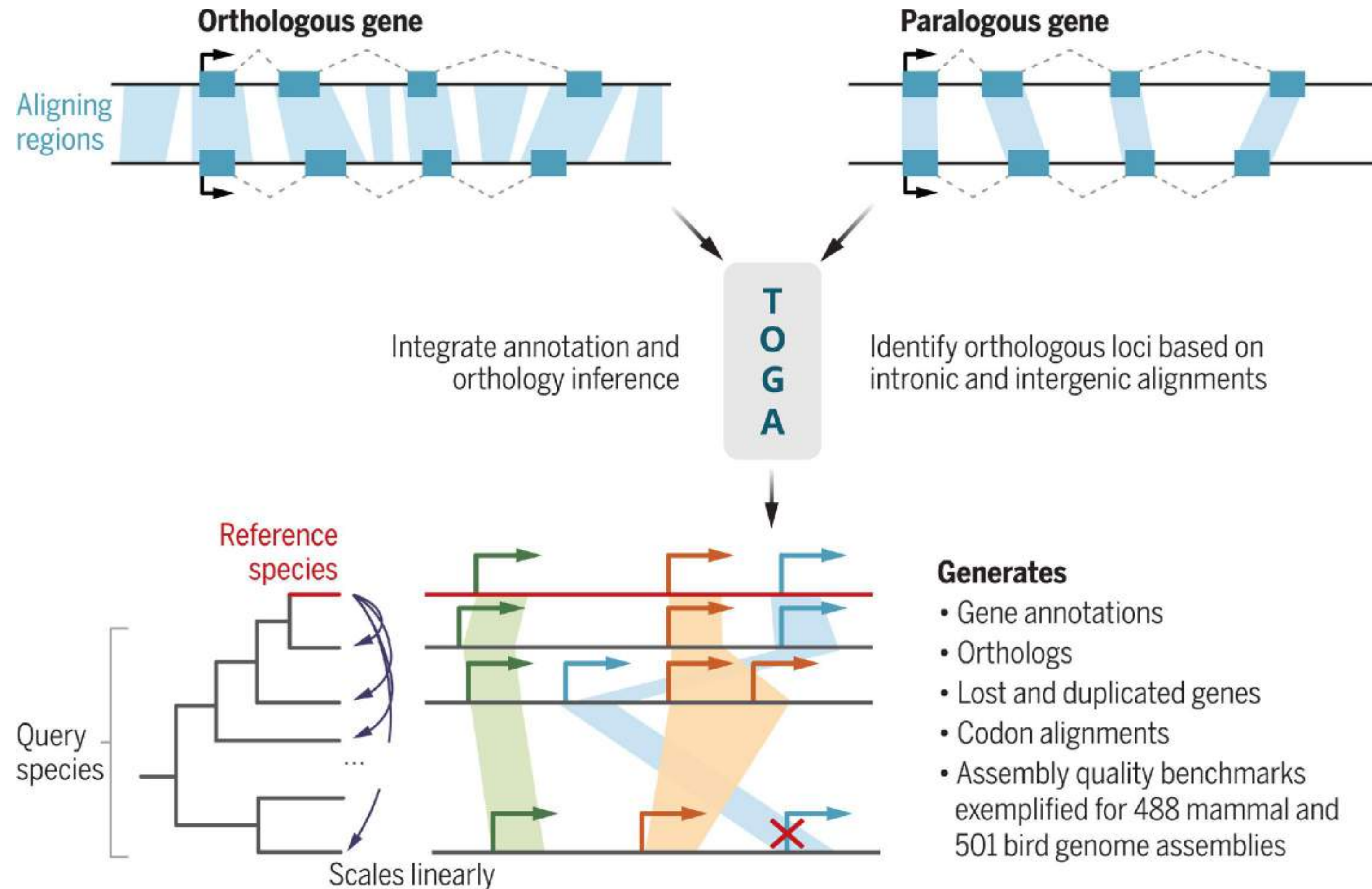
Bitscore: a normalized score that represents the quality of the alignment. It's independent of the database size.

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

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



» Methods | Whole-genome alignment based »



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.

» 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.**

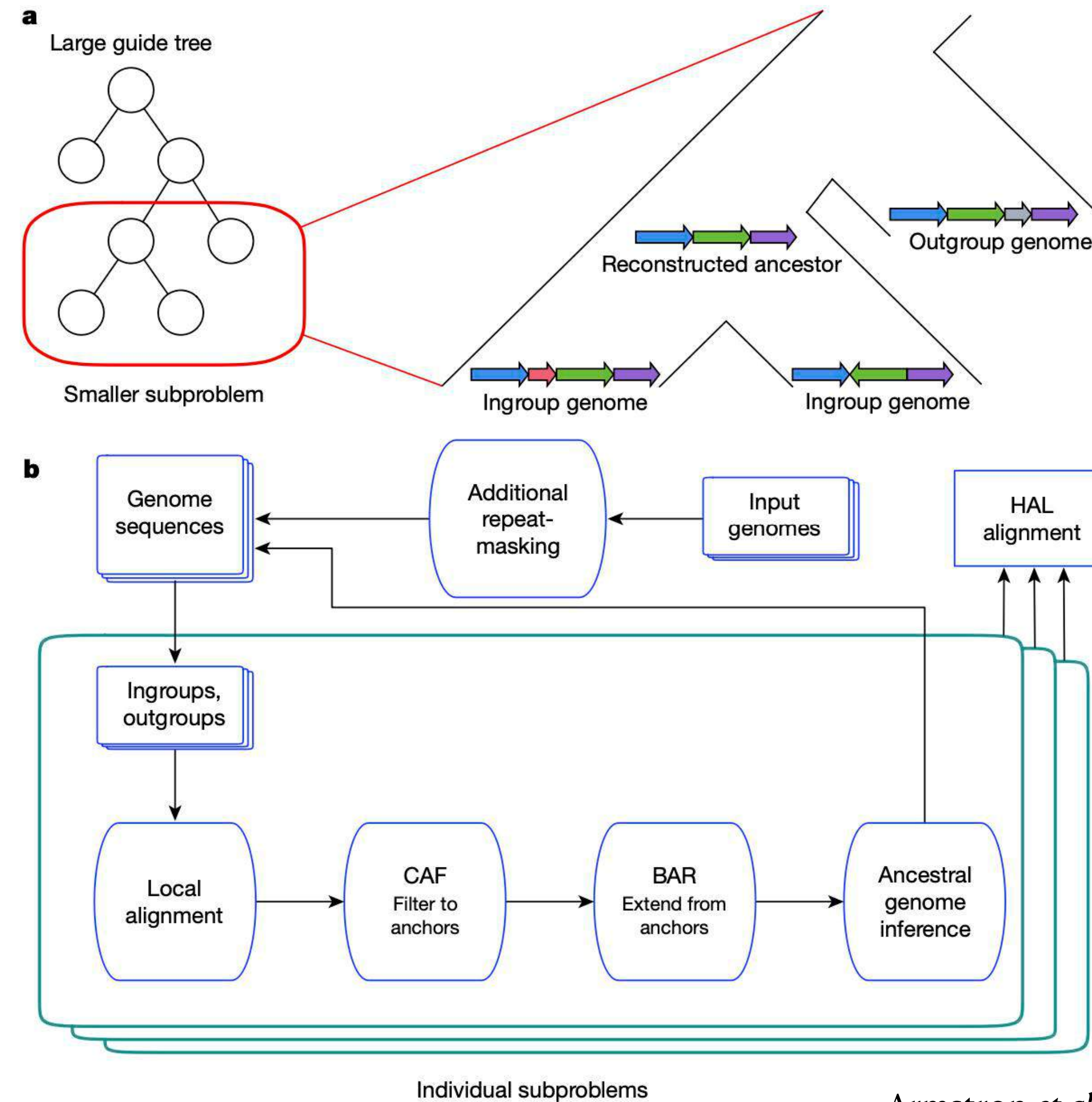
Common limitations are ‘reference bias’:

- Constrains a multiple alignment to only regions present in reference genome.
- Restricting the alignment to be ‘single-copy’, determining miss multiple-orthology relationships.

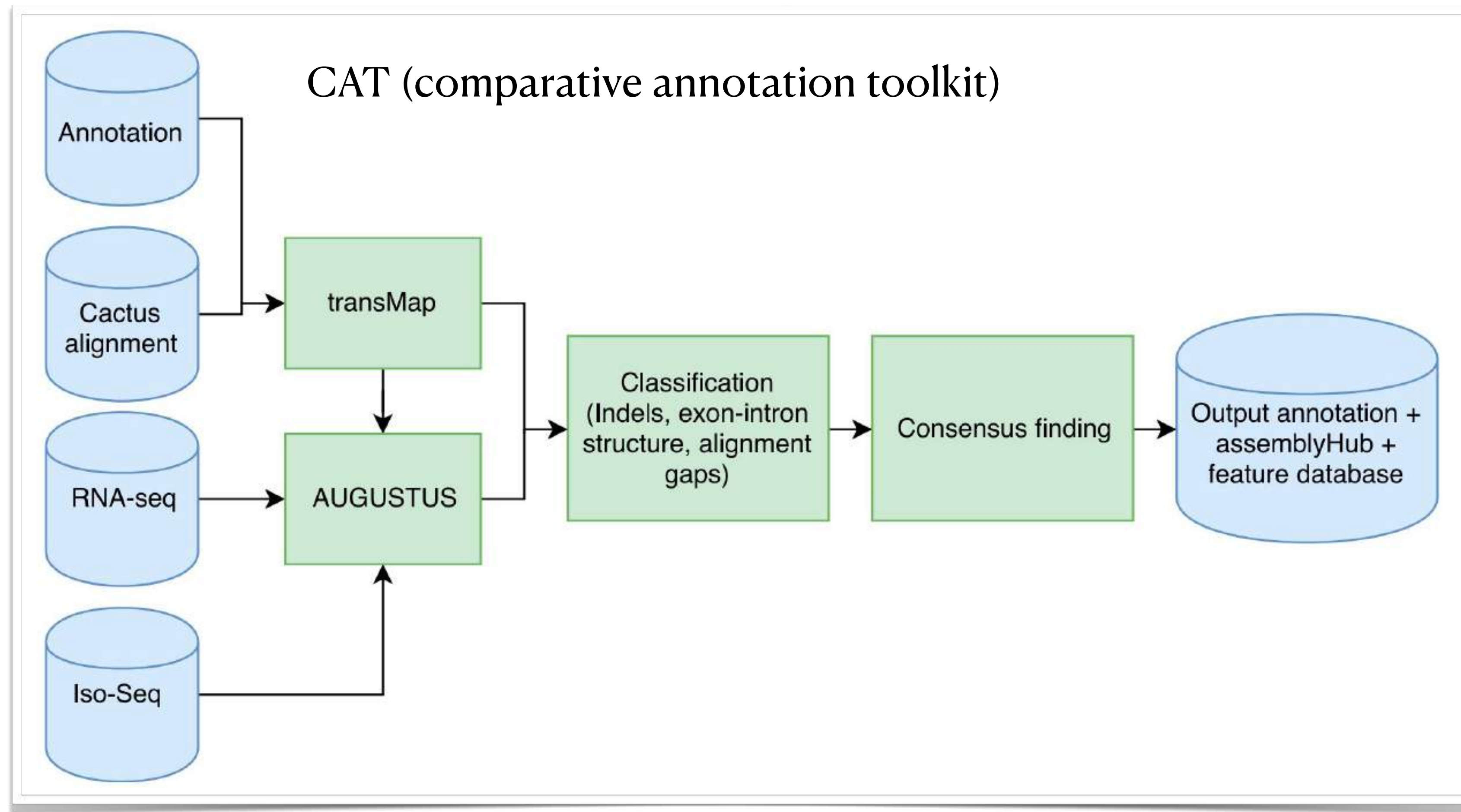
Cactus (ProgressiveCactus) is a “reference *free*” whole genome aligner.



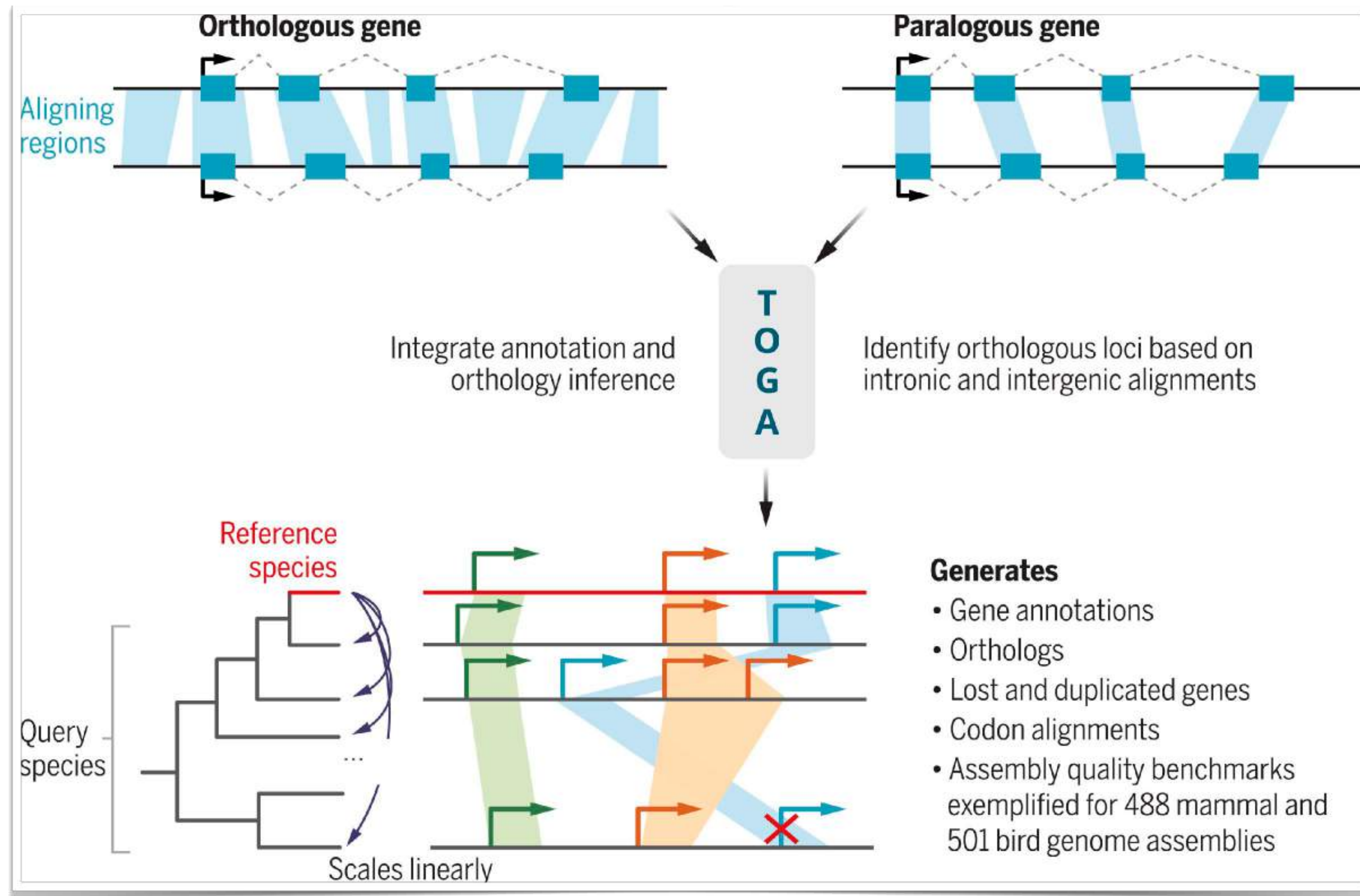
» Whole Genome alignment »



» Some of the applications (annotation) »

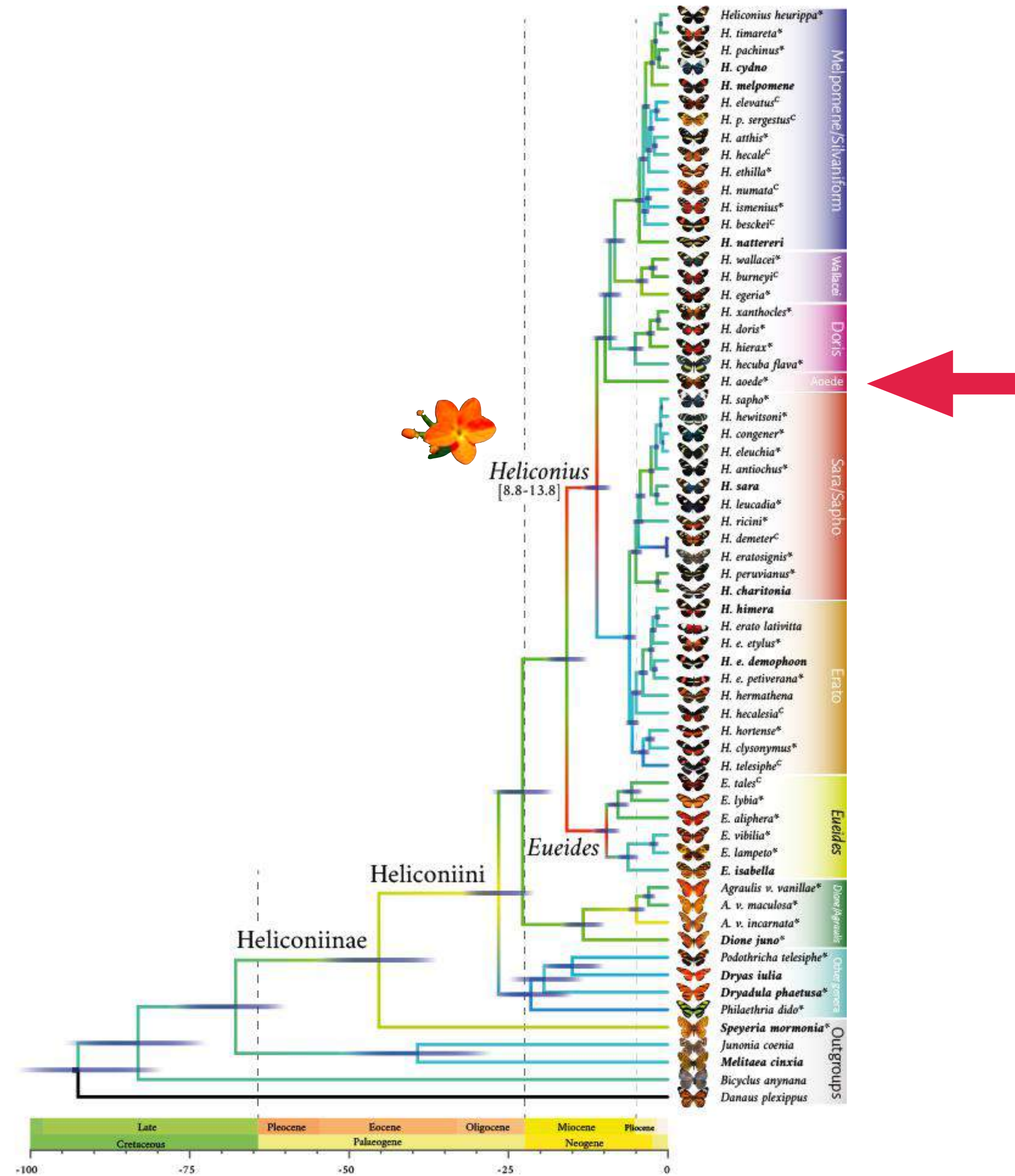


» Some of the applications (orthology inference) »

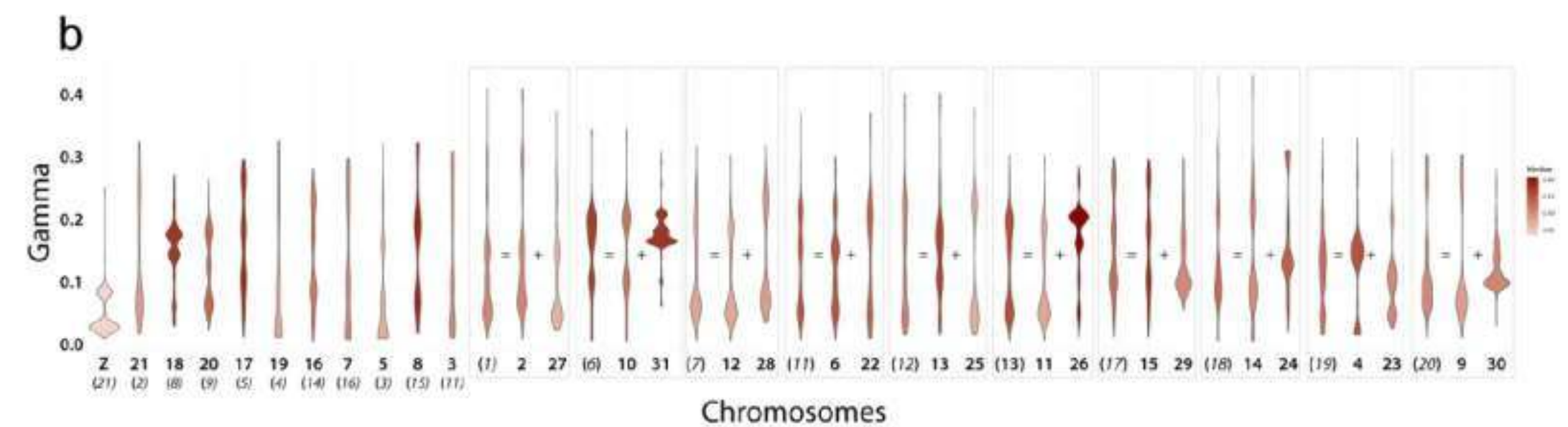
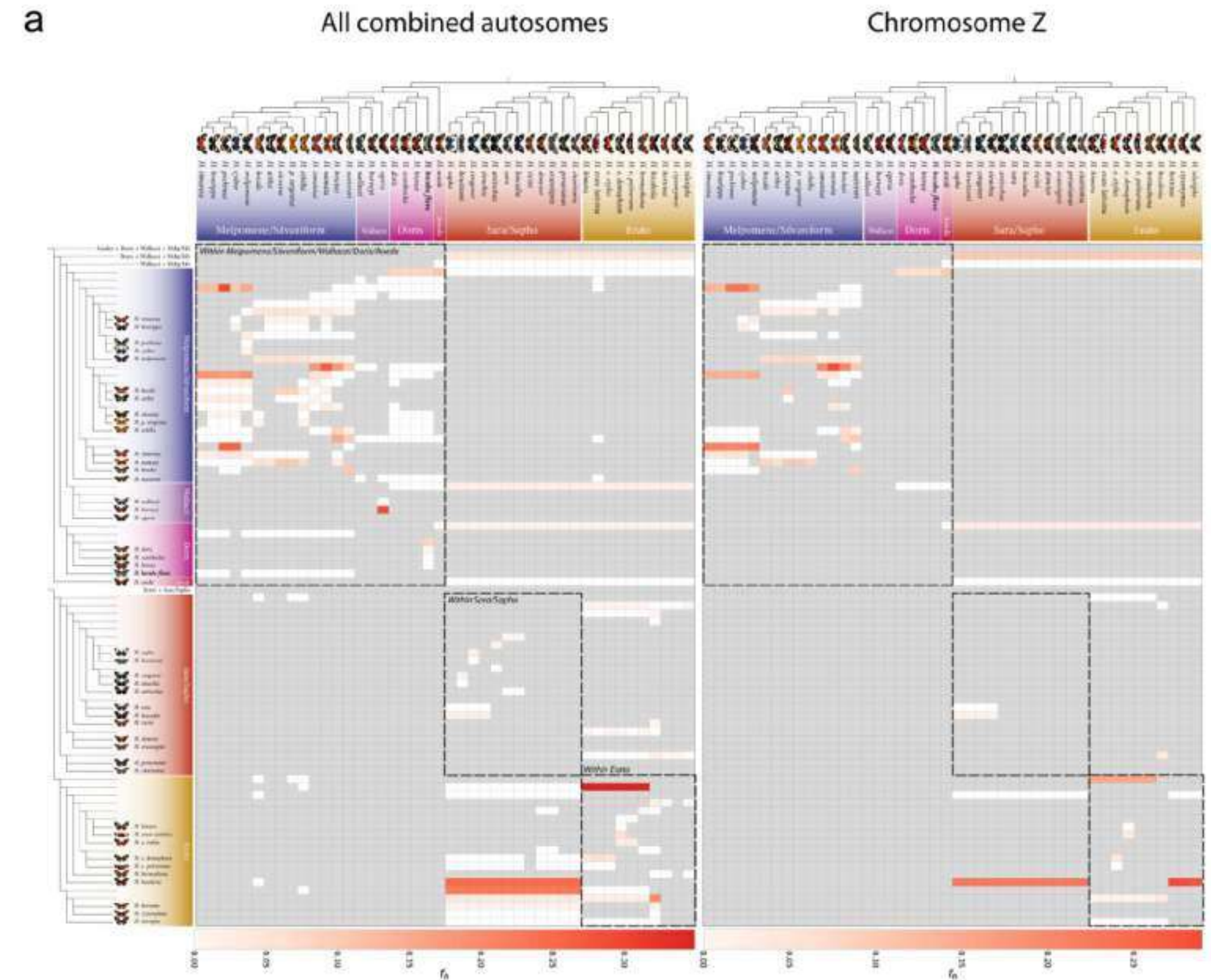
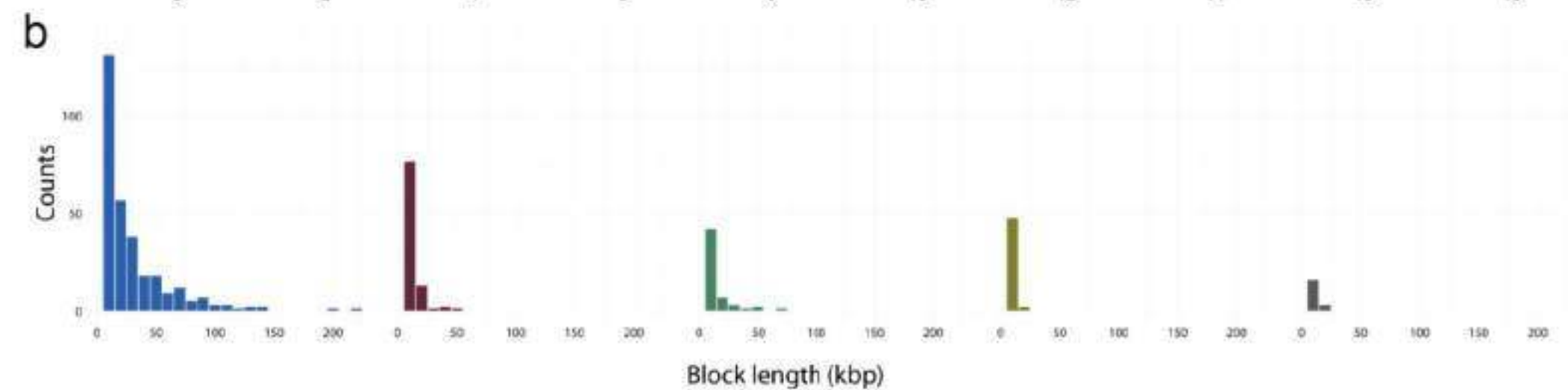
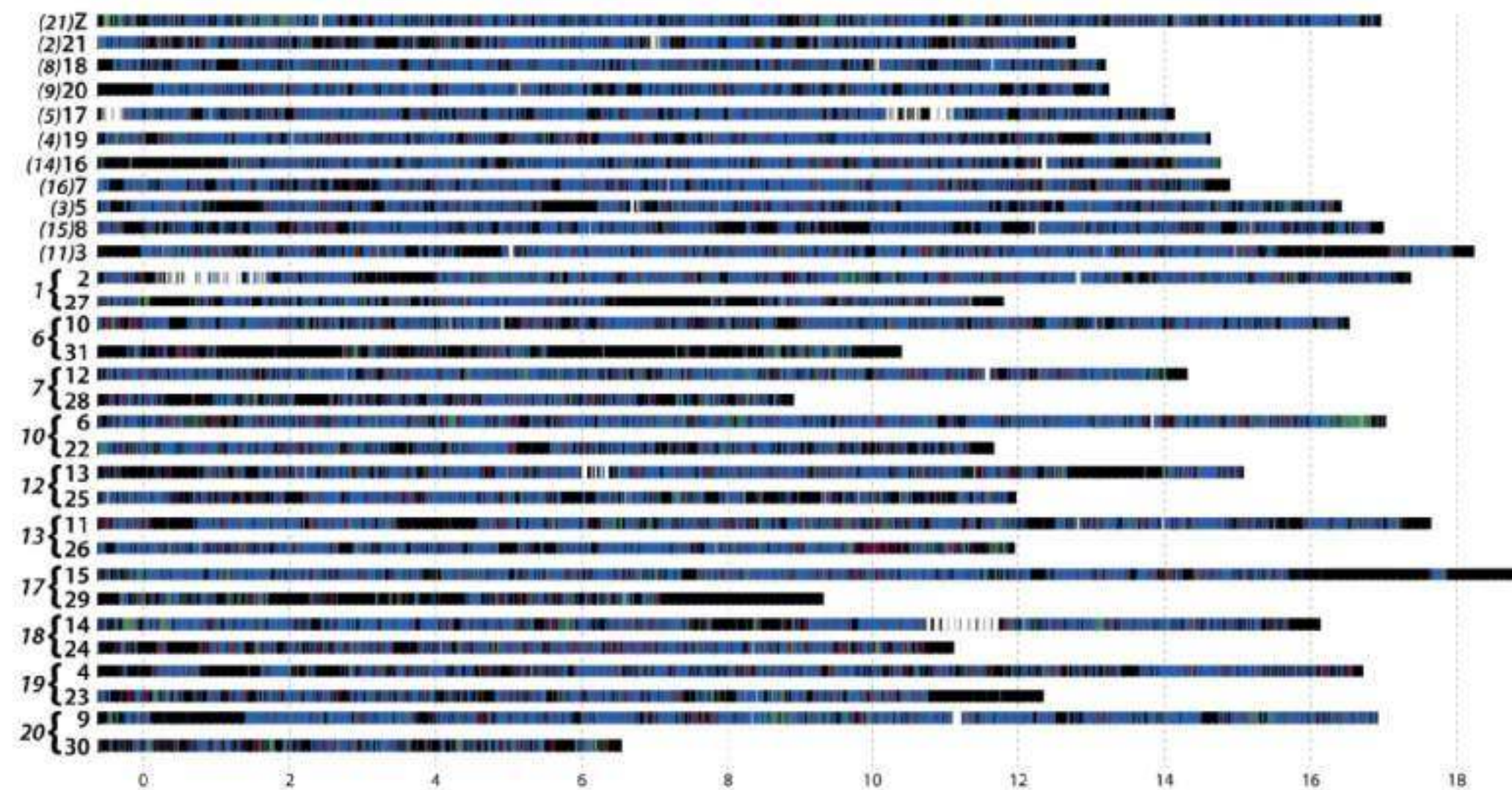
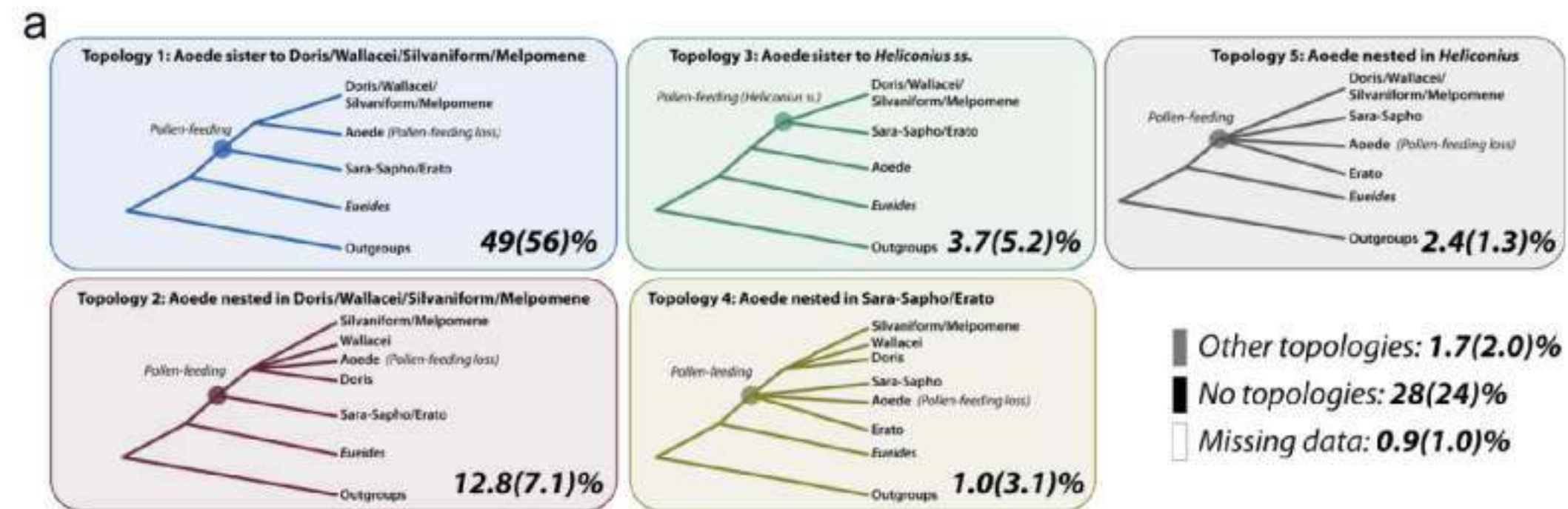


» Phylogeny / Introgression »

» Phylogeny / Introgression »

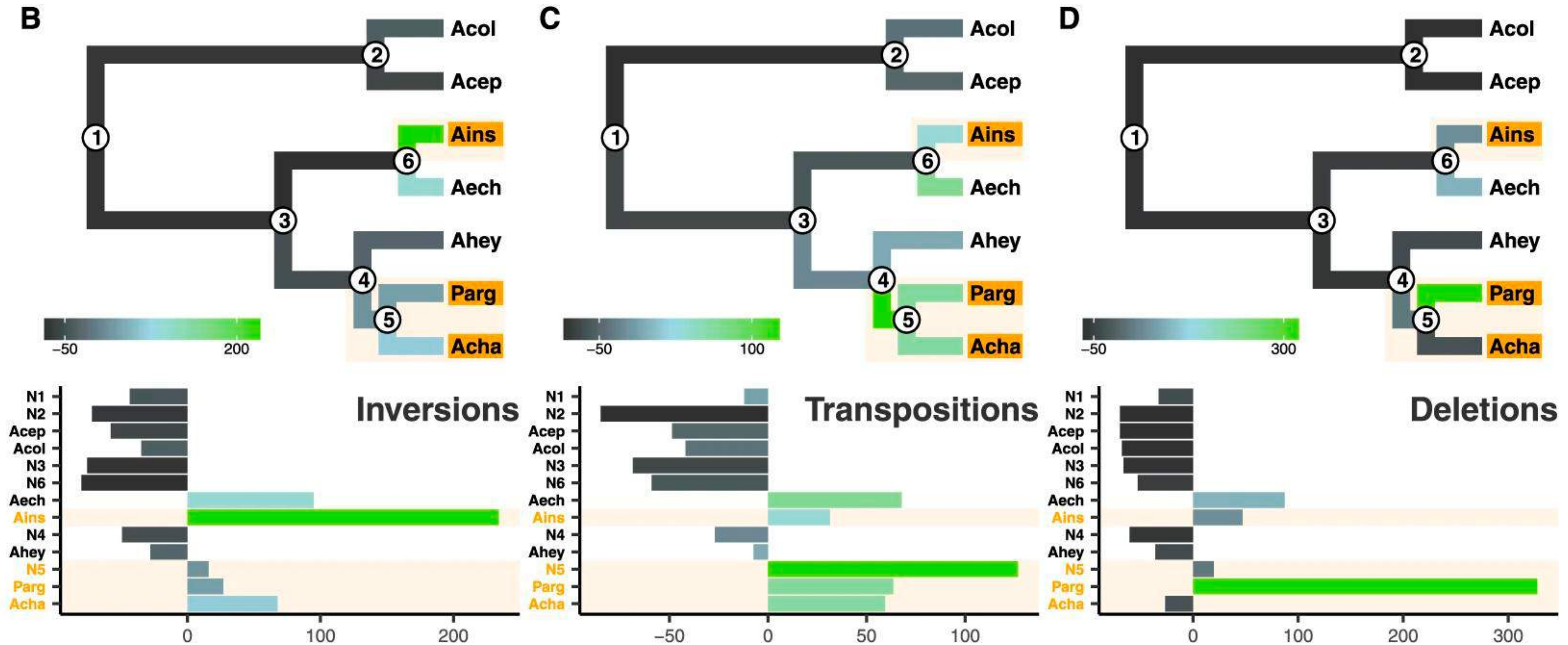


» Phylogeny / Introgression »

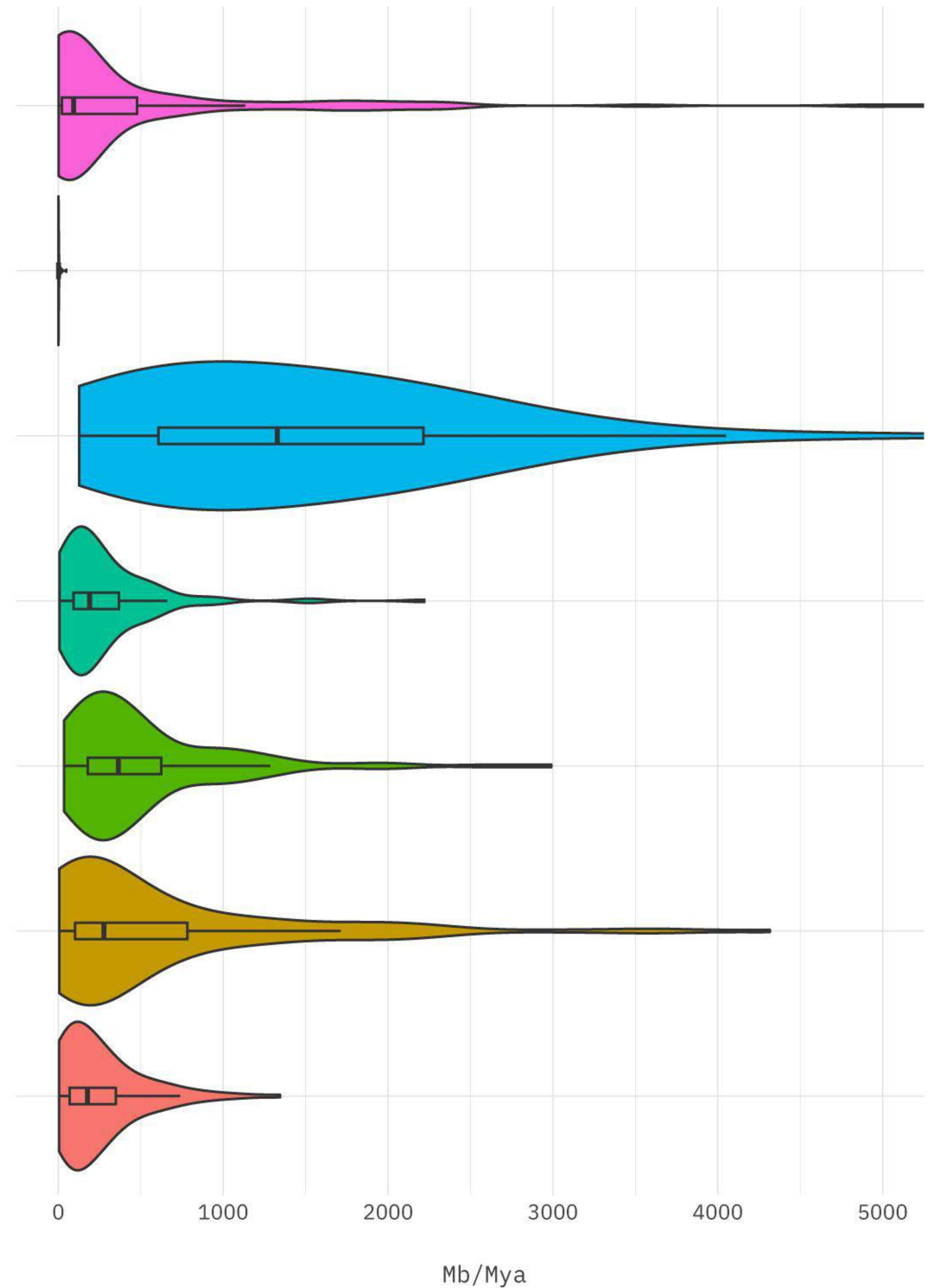
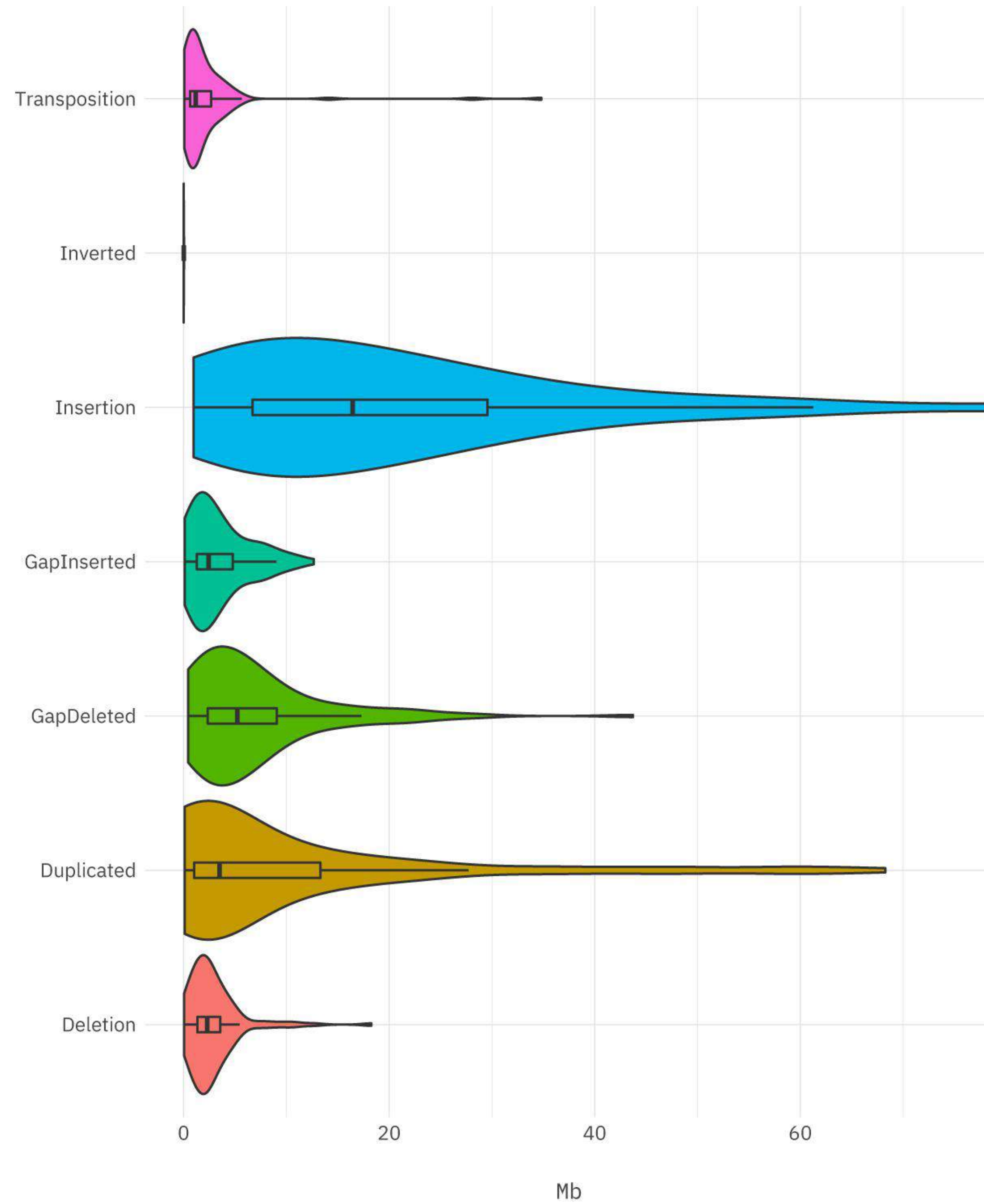


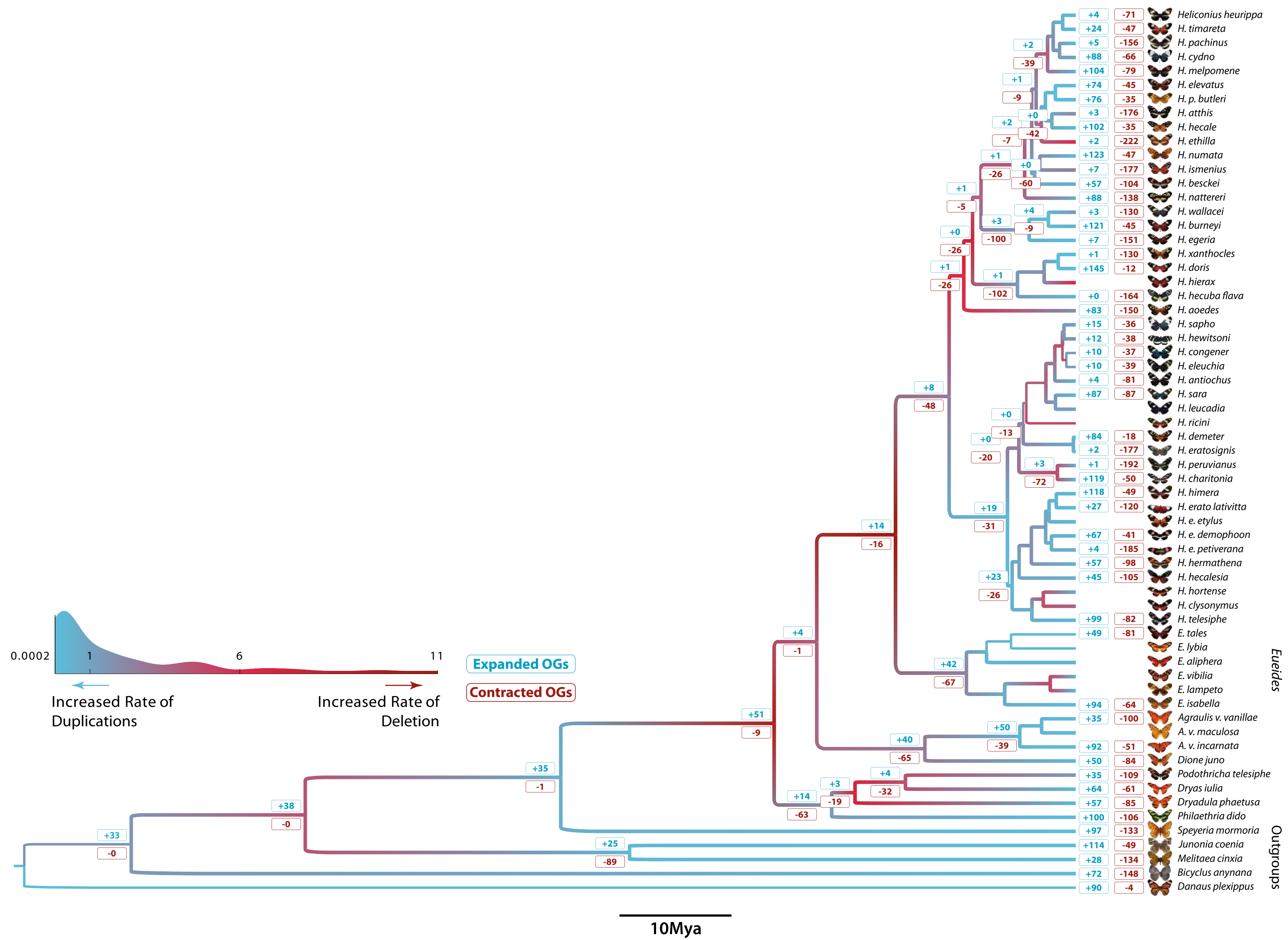
» Structural Rearrangements »

» Structural Rearrangements »



Schrader et al 2021 Nature Comm.

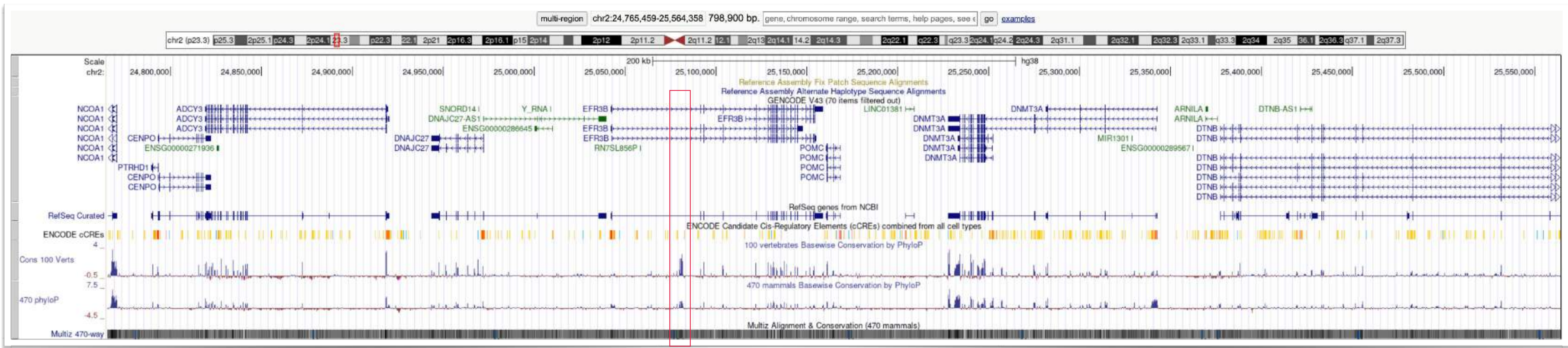




» Conserved non-coding Elements (CNEs) »

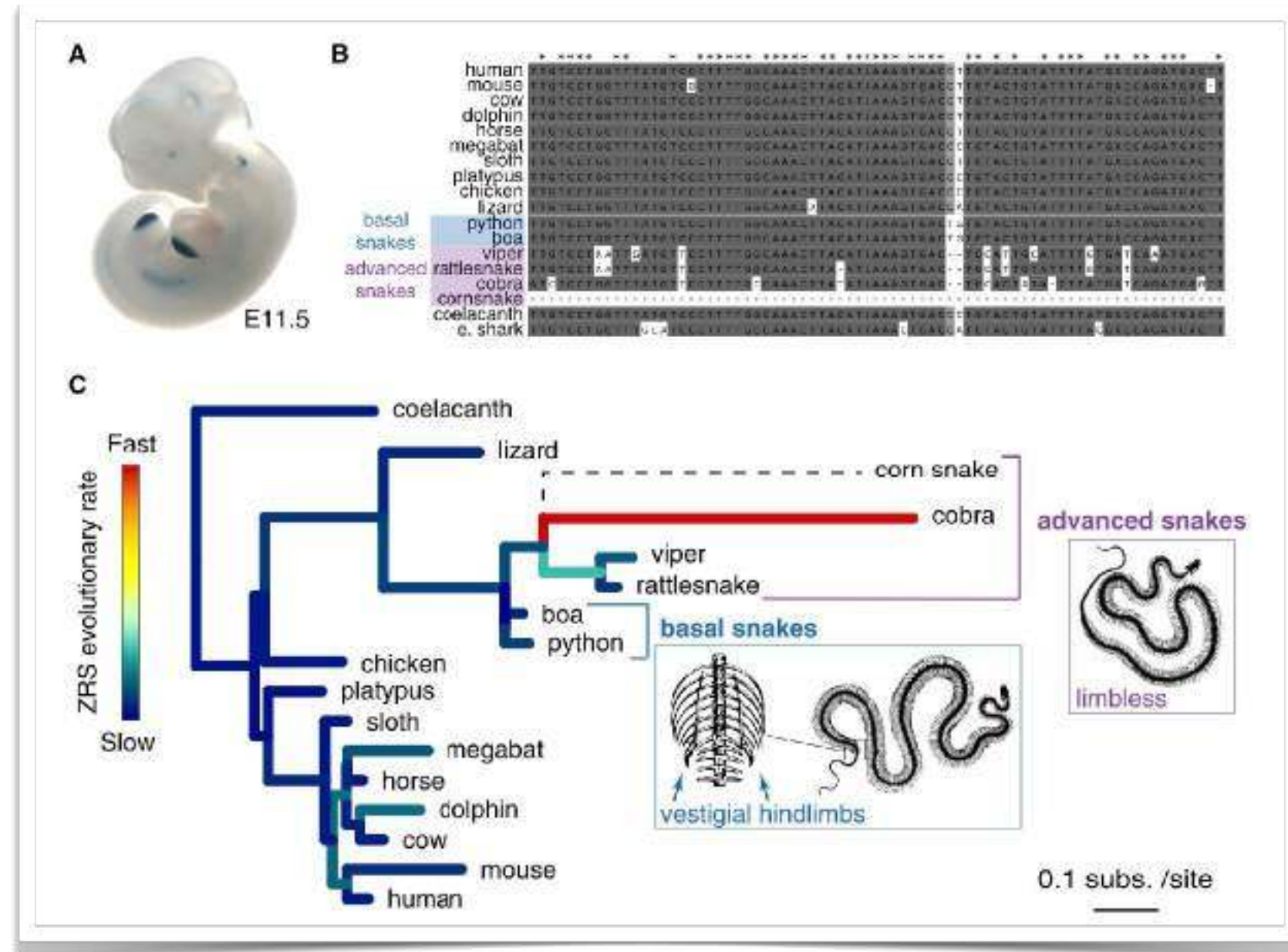
» Conserved non-coding Elements (CNEs) »

- > 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.

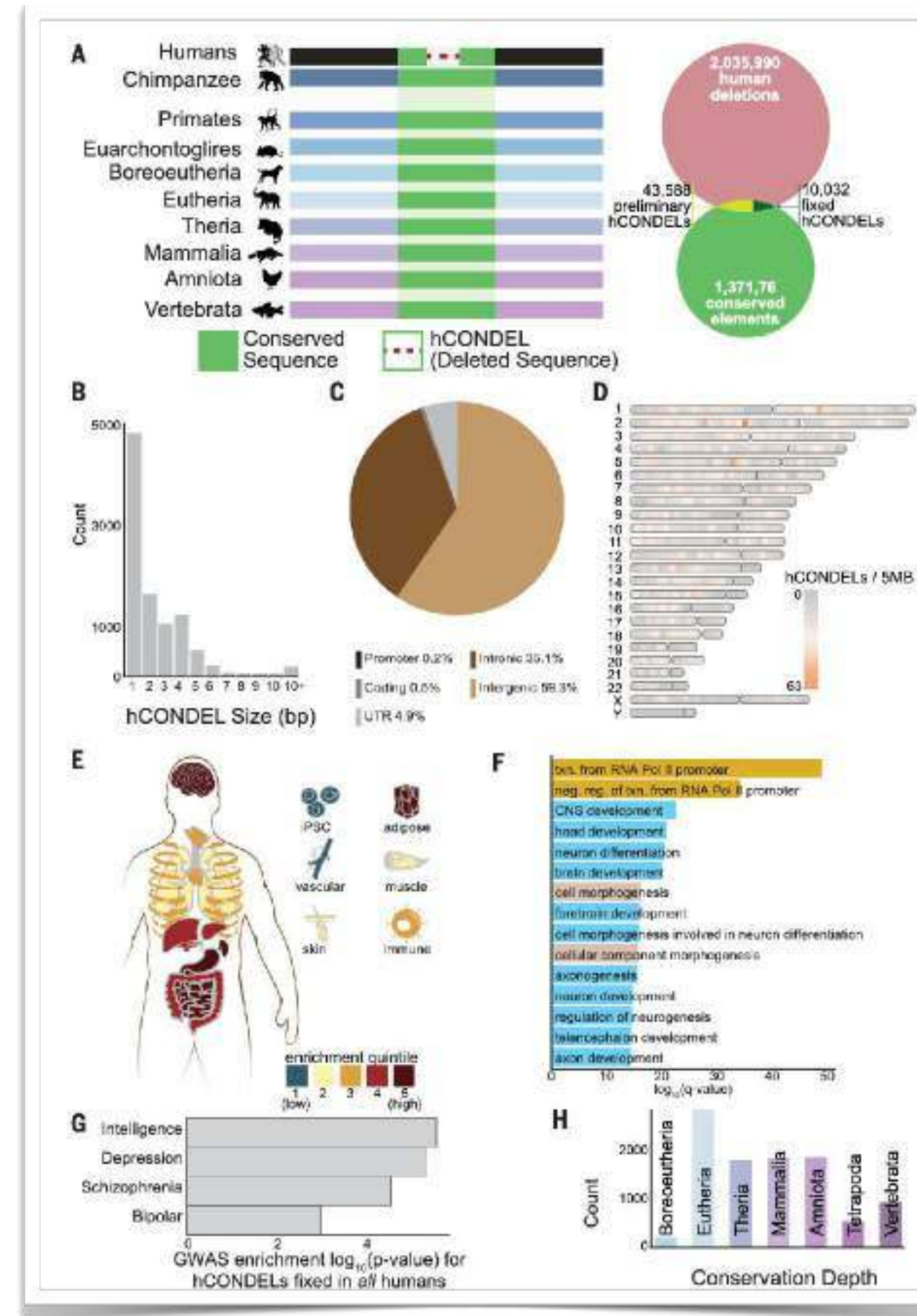


Candidate Cis-Regulatory Elements (cCREs)

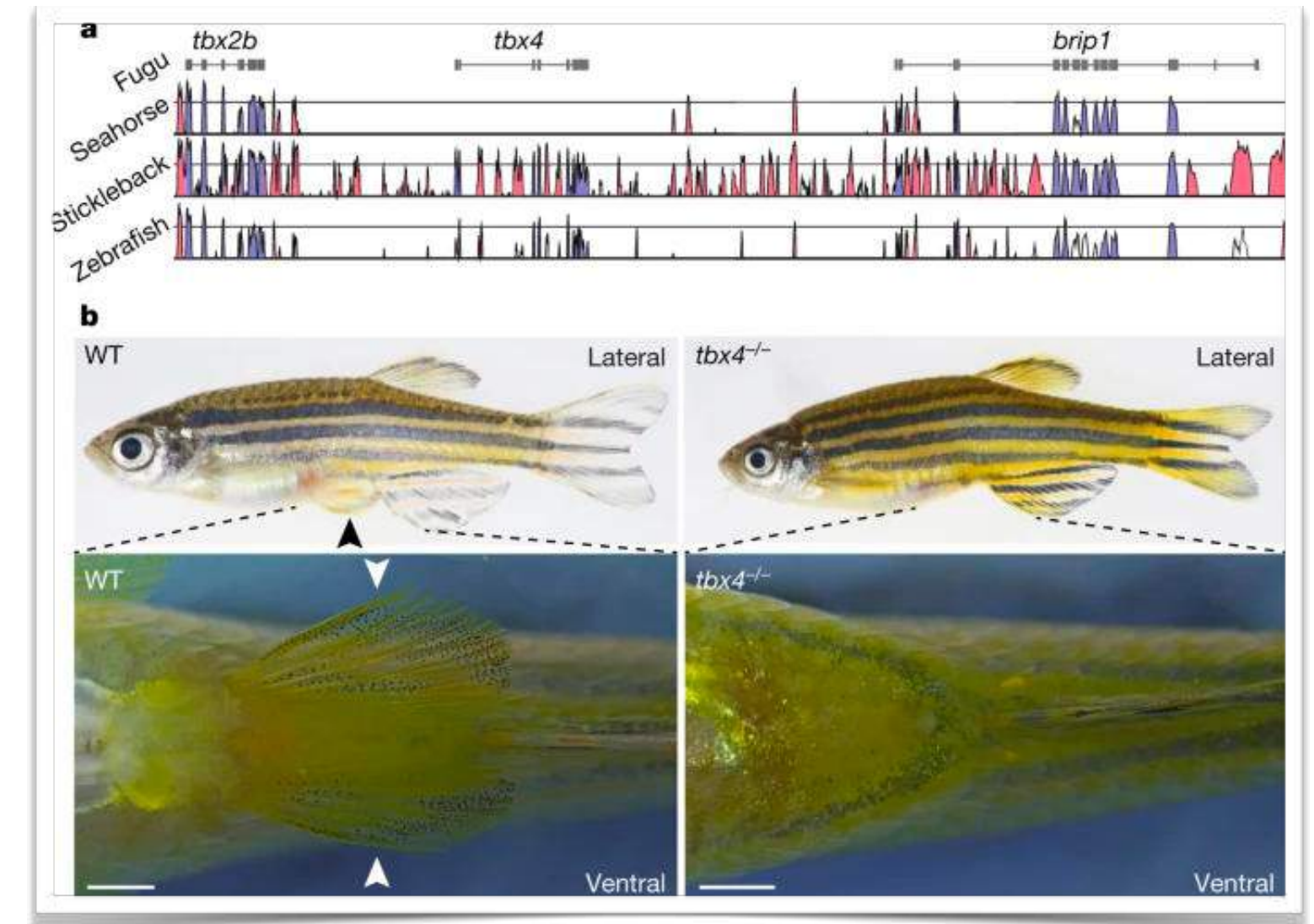
» Conserved non-coding Elements (CNEs) »



Kvon et al 2016 Cell



Xue et al 2023 Science



Lin et al 2016 Nature

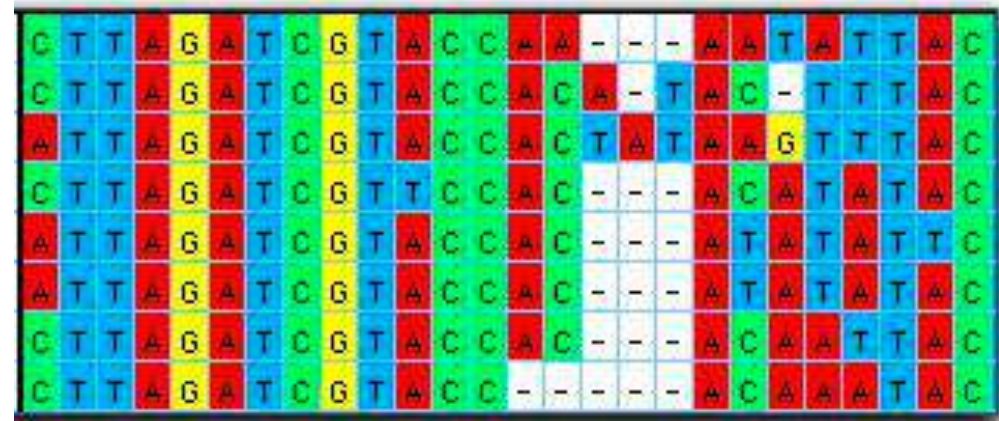
» How to identify these regions »

1. A Model of DNA sequence evolution

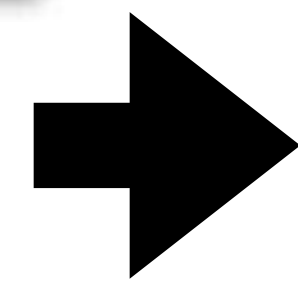
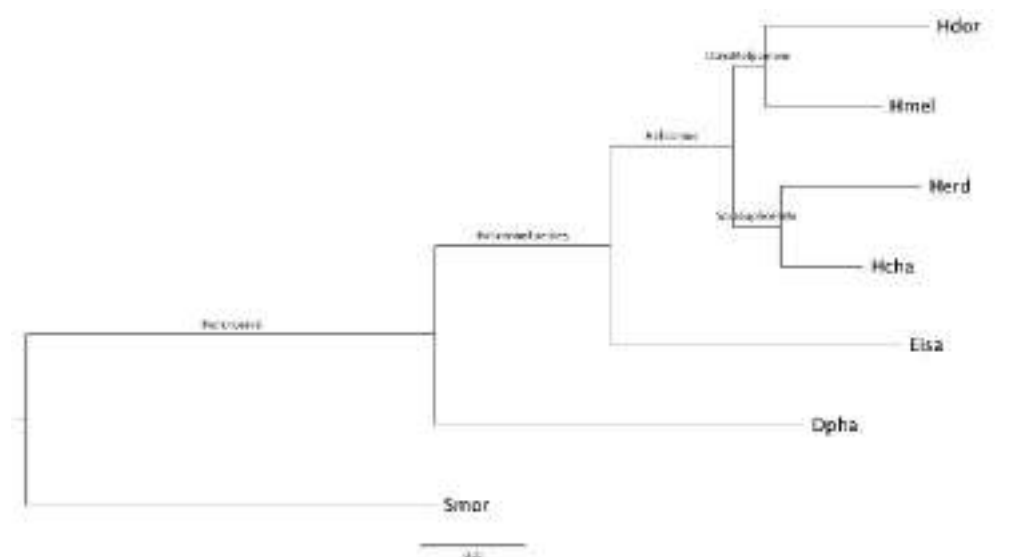
2. Phylogenetic tree

» How to identify these regions »

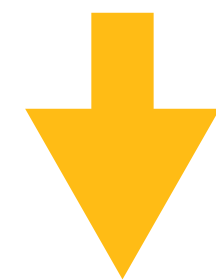
Neutral evolving regions



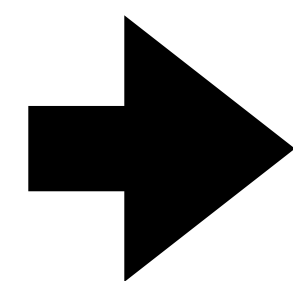
+



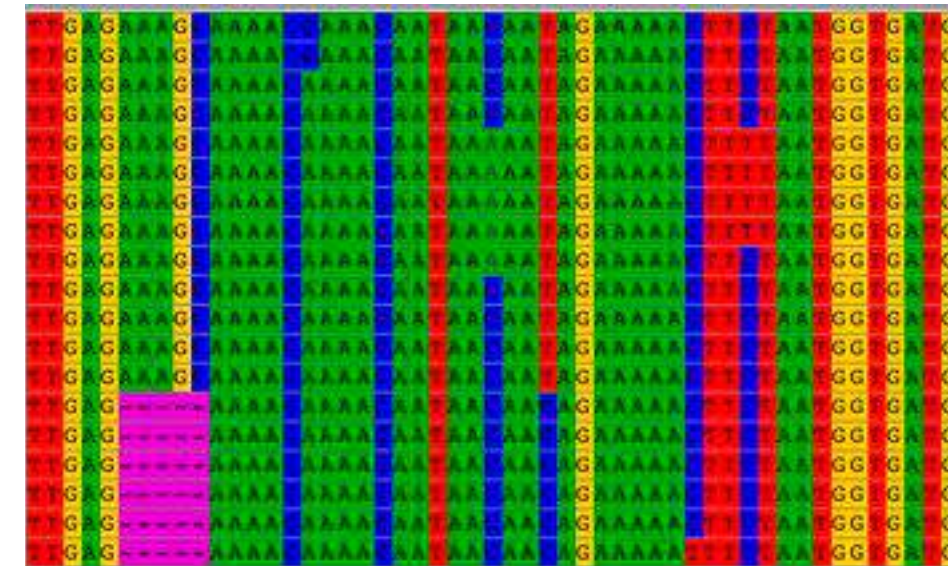
Neutral model



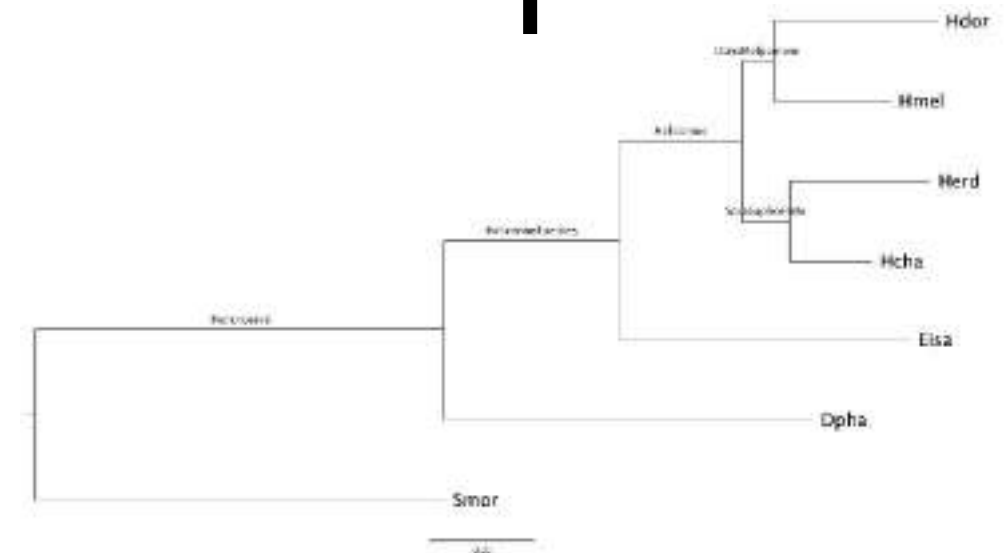
Other metrics
(*PhyloP*)



WGA

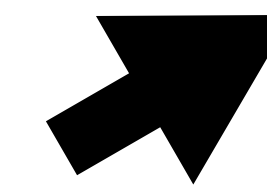


+

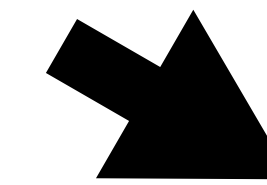


+

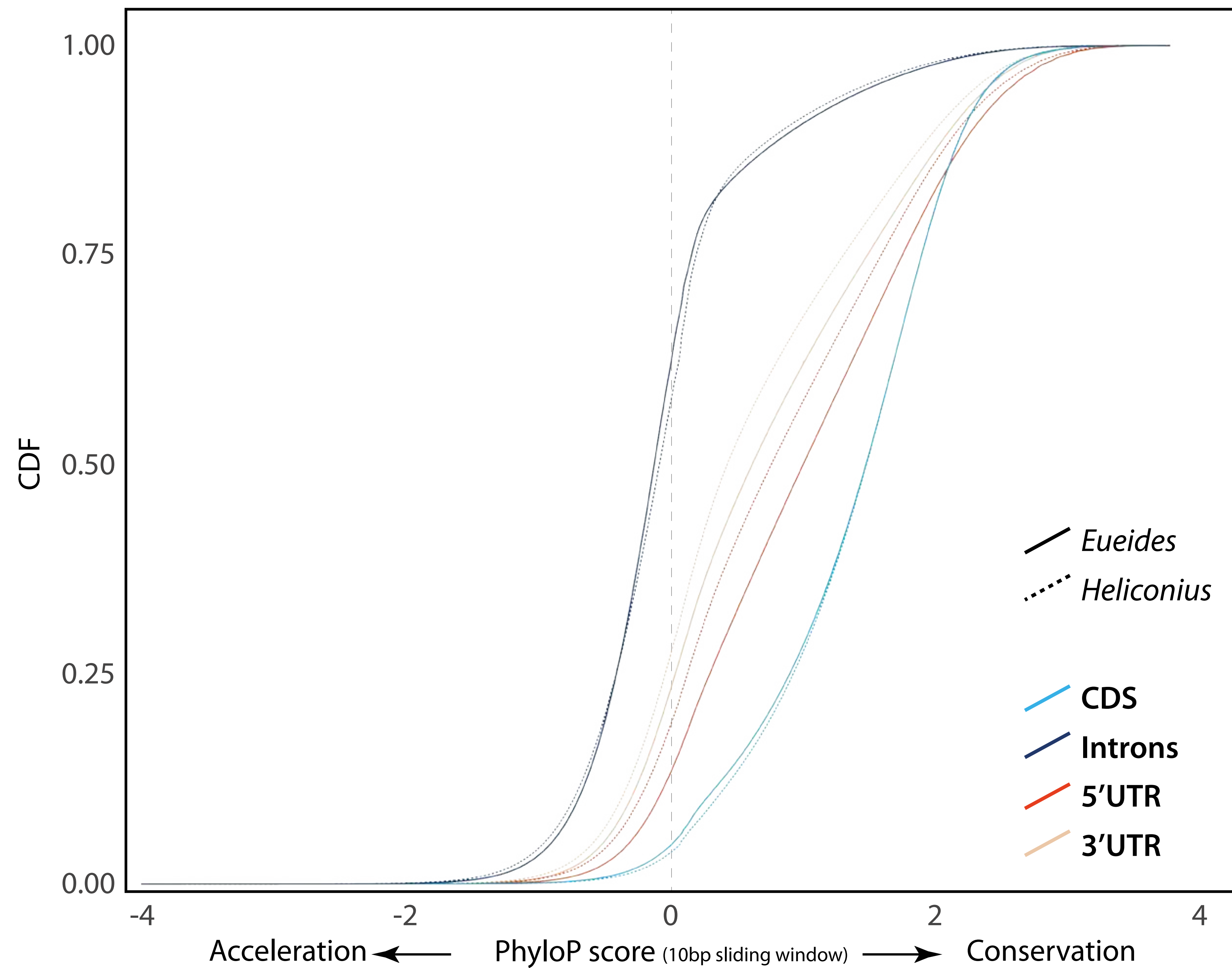
neutral model



Conservation model



Non-conserved model



» 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 (Assay for Transposase-Accessible Chromatin using sequencing) »



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

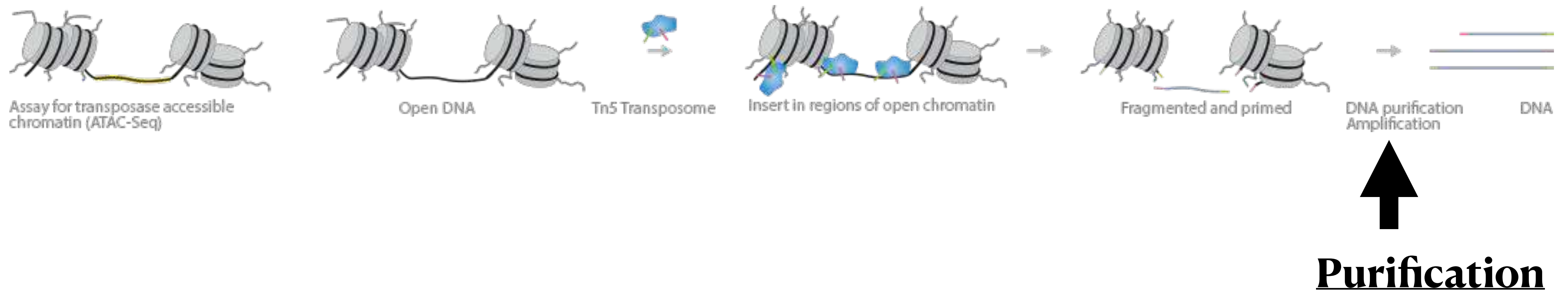
» ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing) »



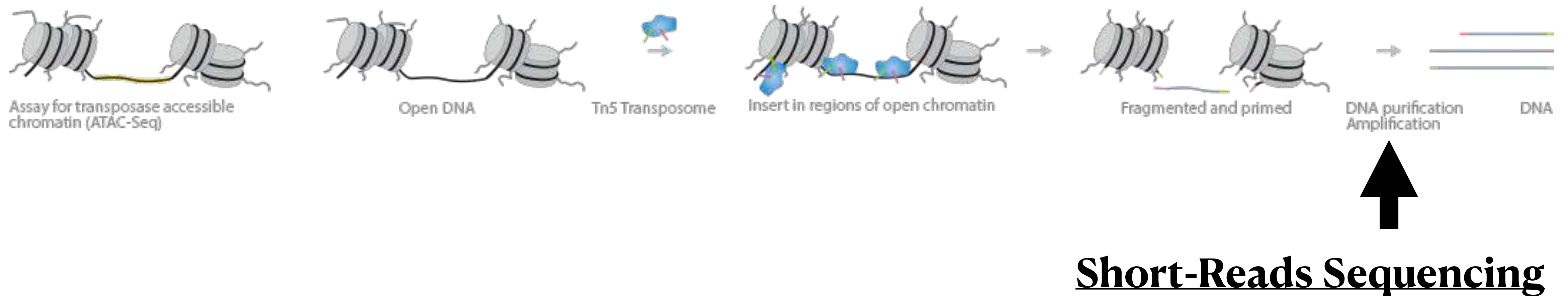
Tagmentation

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

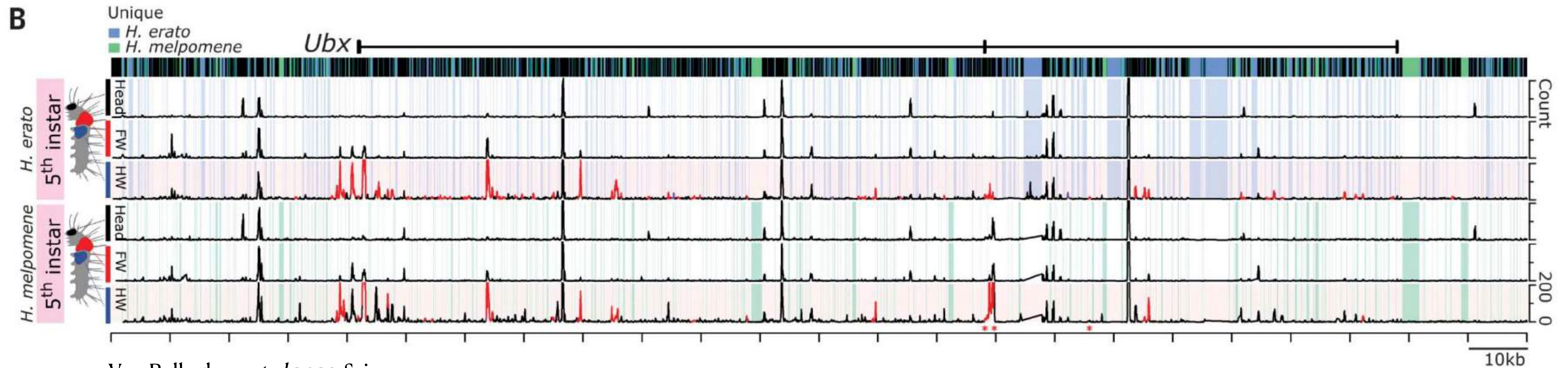
» ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing) »



» ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing) »



» ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing) »



Van Belleghem *et al* 2023 *Science*

» Let's continue »

» Notes »

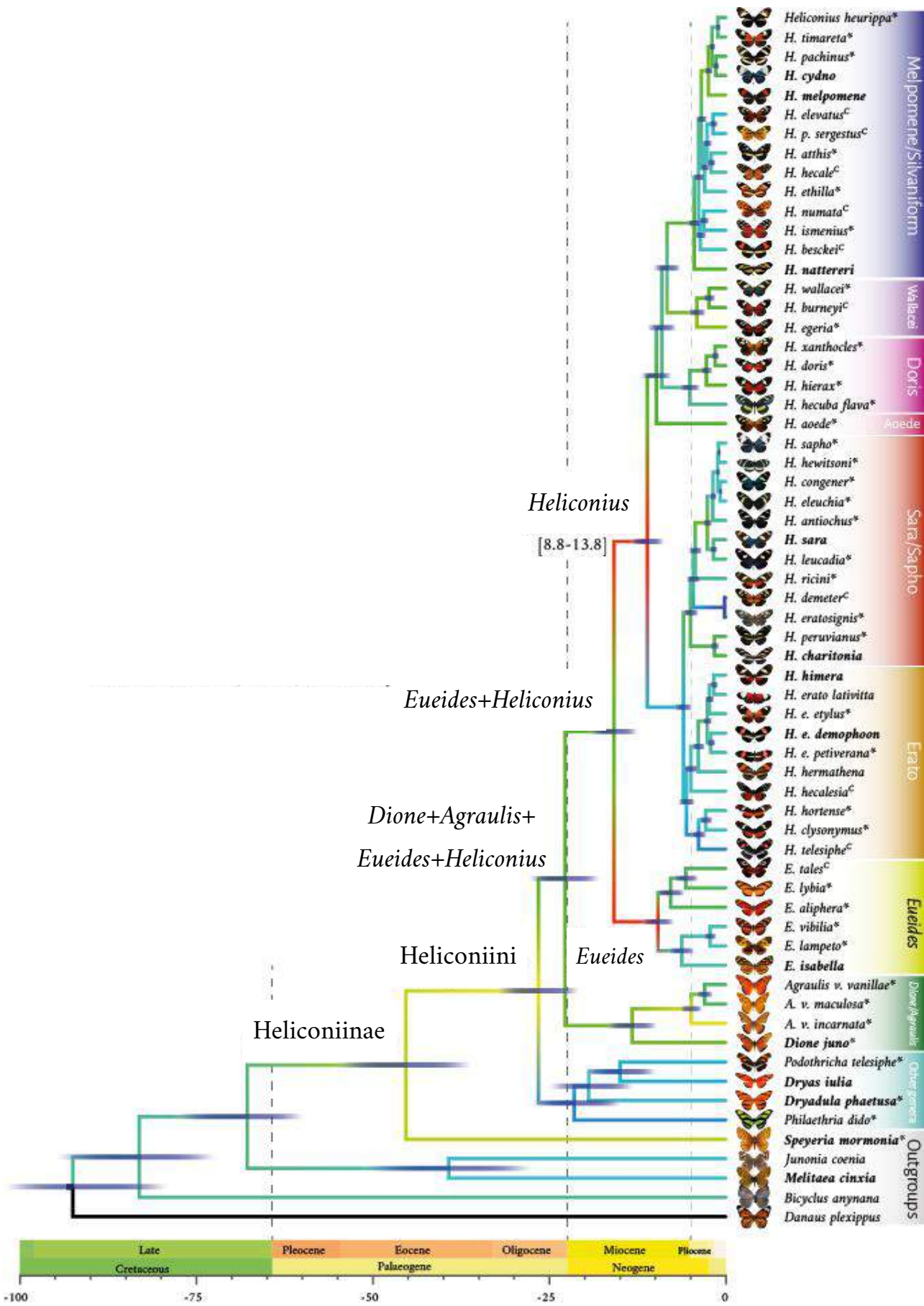
Some UNIX tools you probably going to need:

- “>” Redirect
- “|” Pipe
- “sed” a stream editor

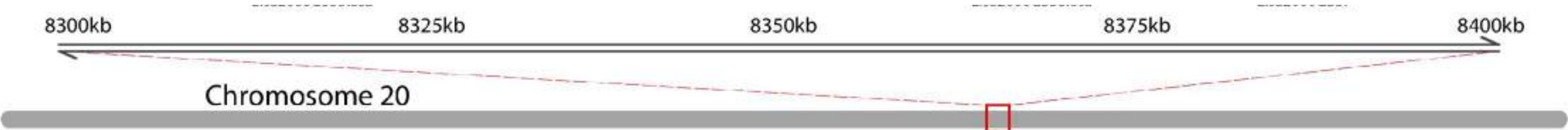
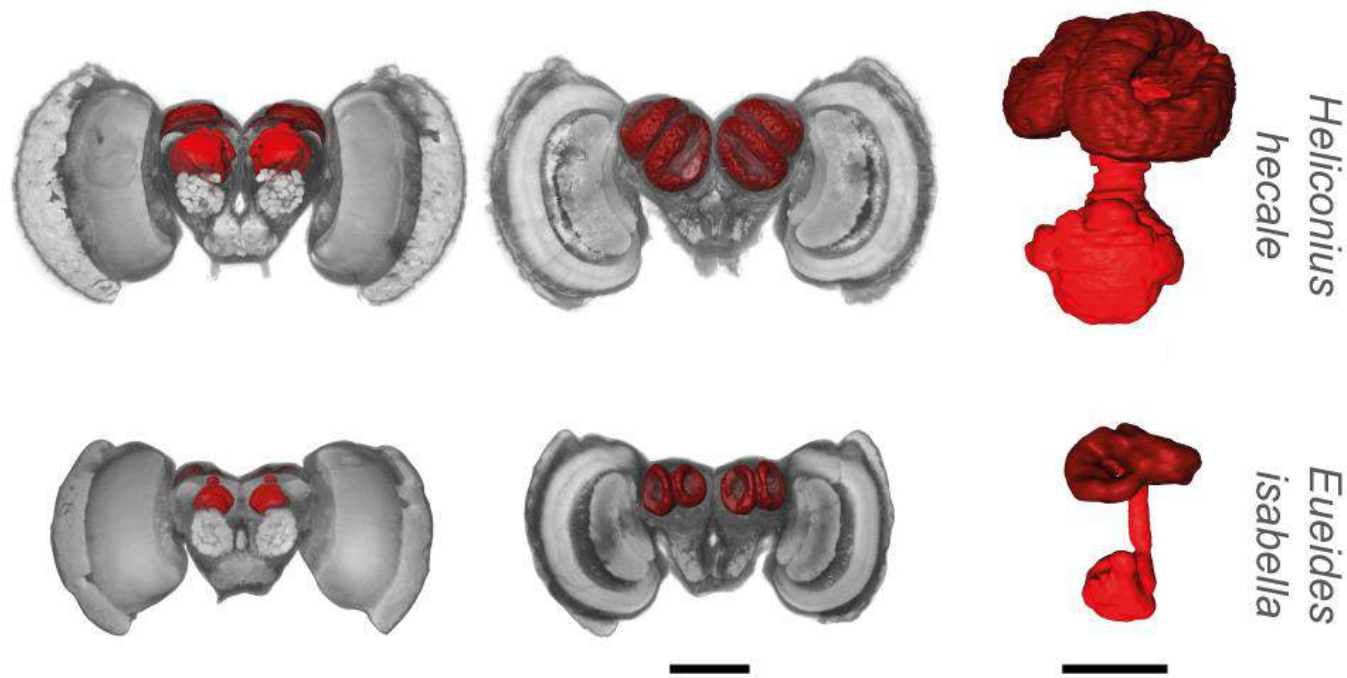
For the extra task:

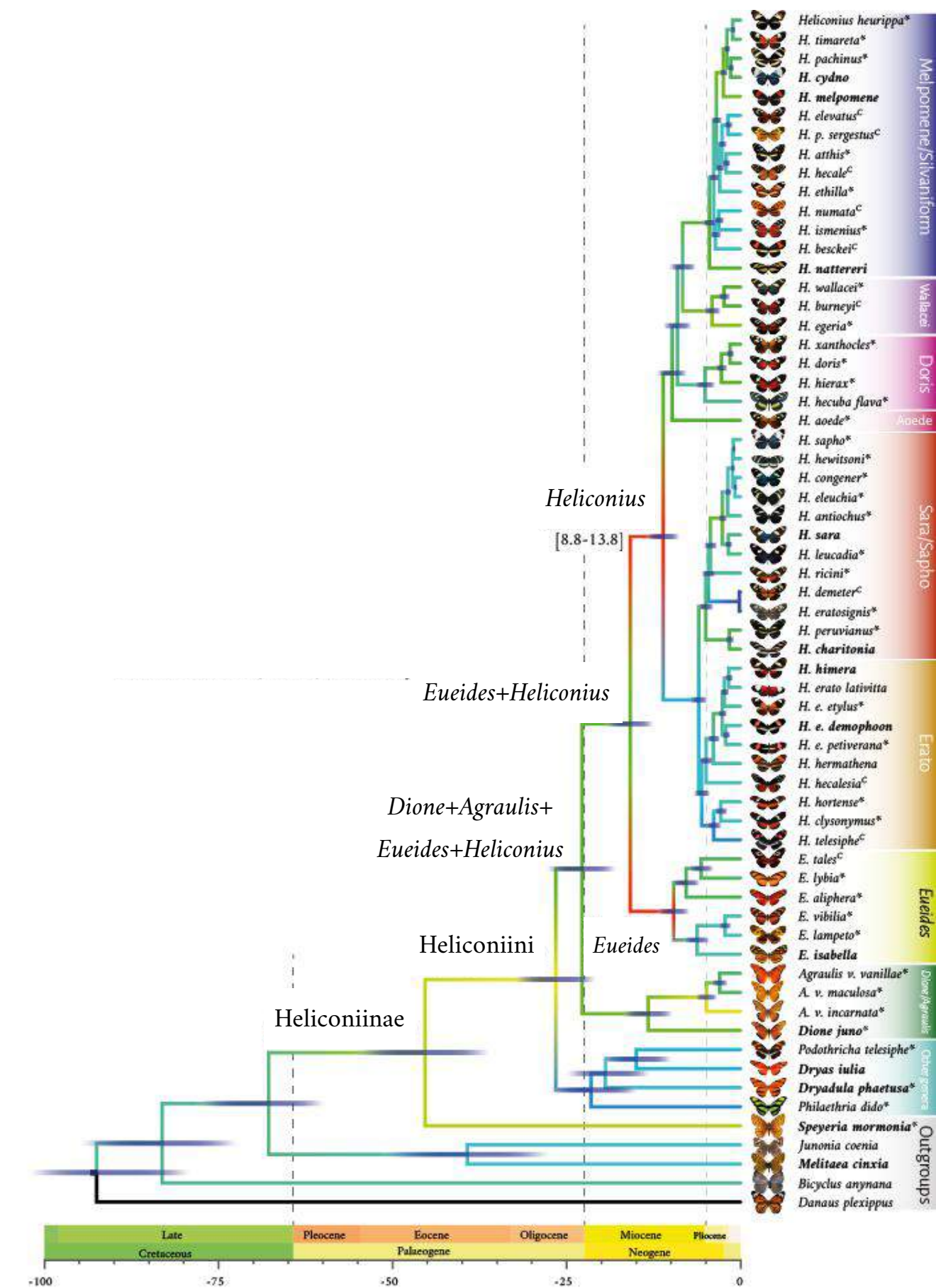
- “seq” unix command to generate a sequence of numbers
- *For loop*

» Test for Acceleration »

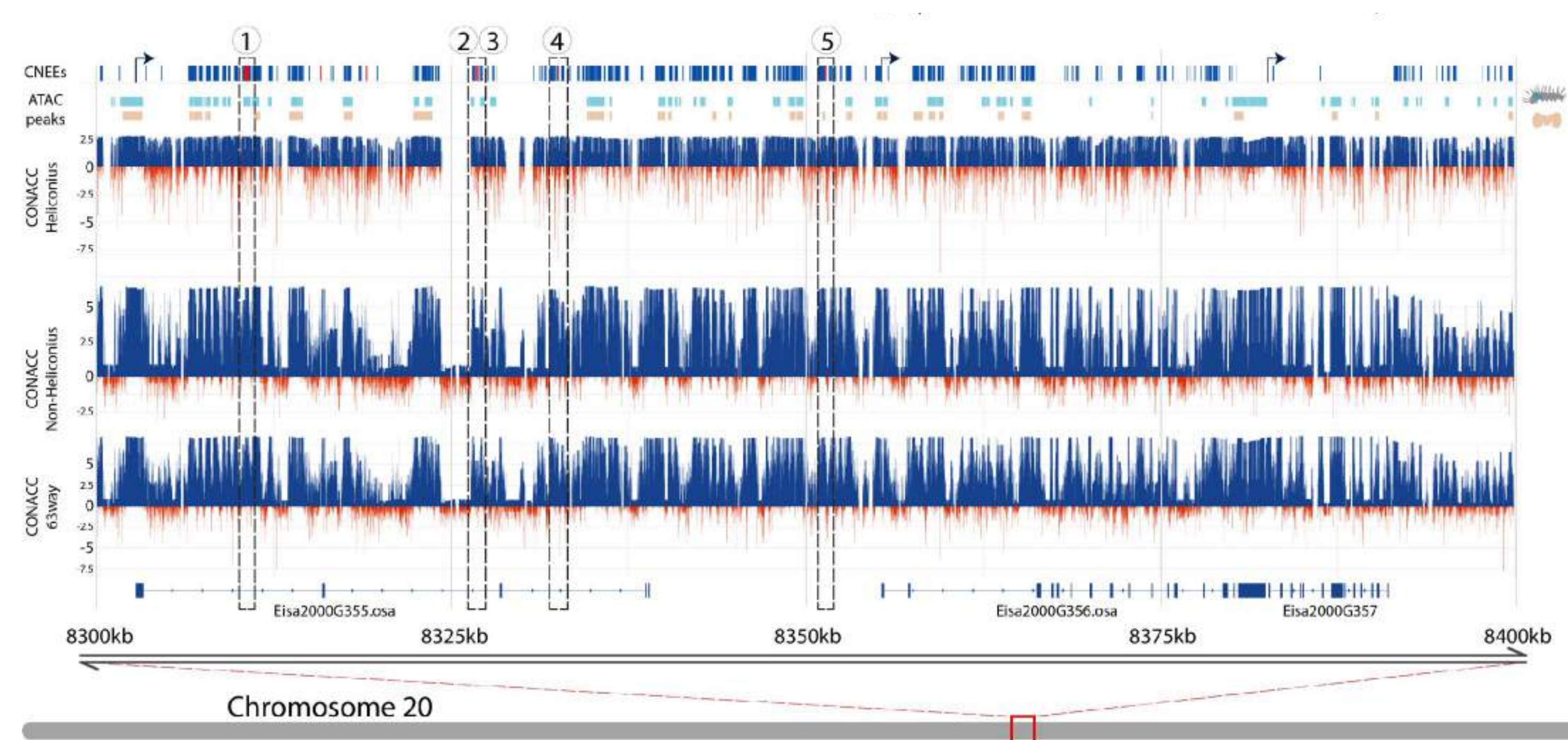
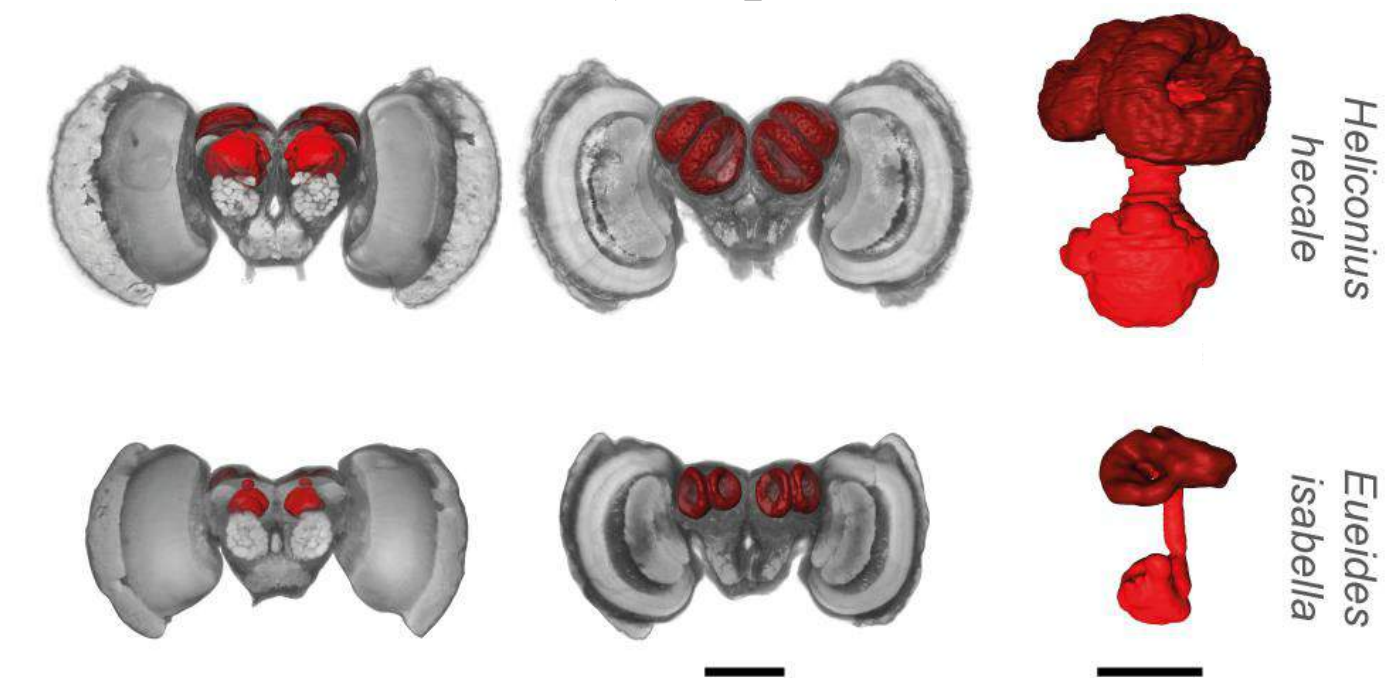


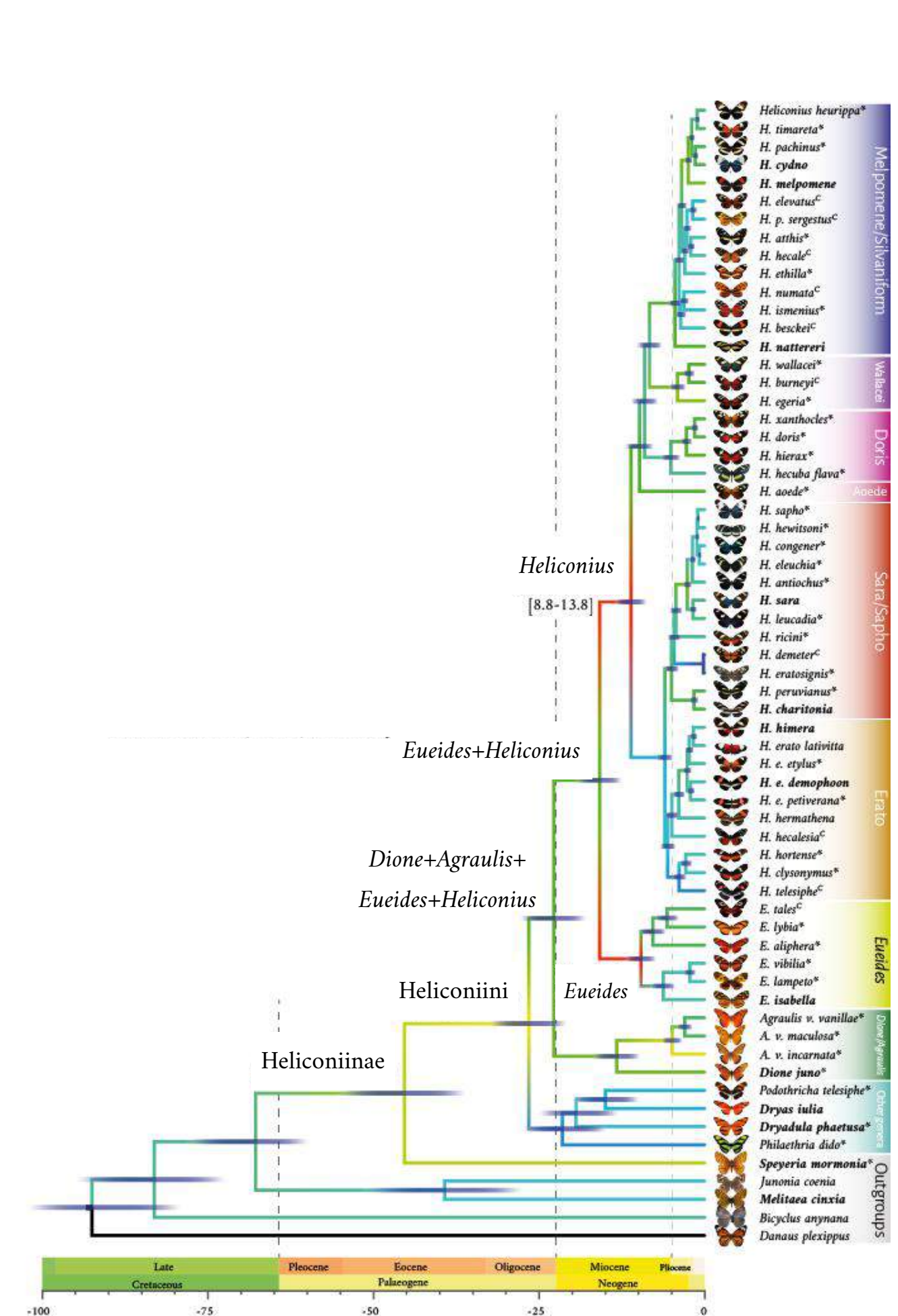
» *Mushroom body expansion (Brain)*



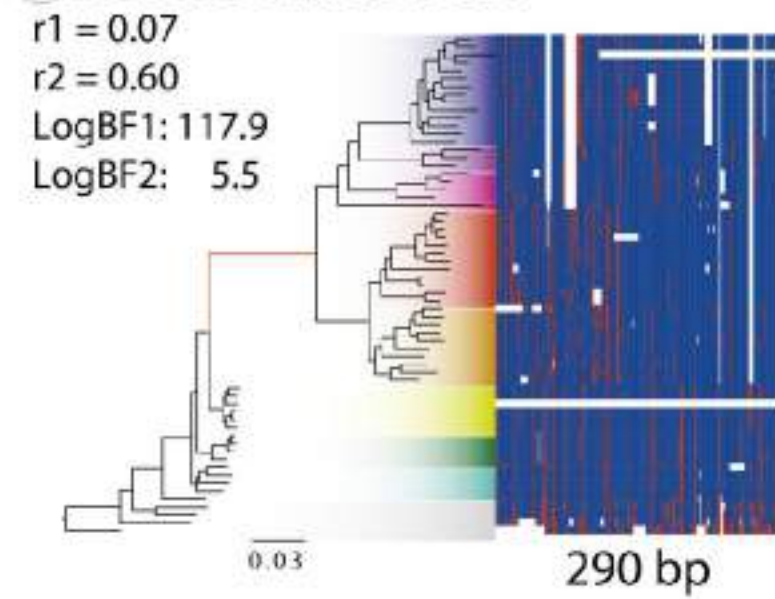


» Mushroom body expansion (Brain)

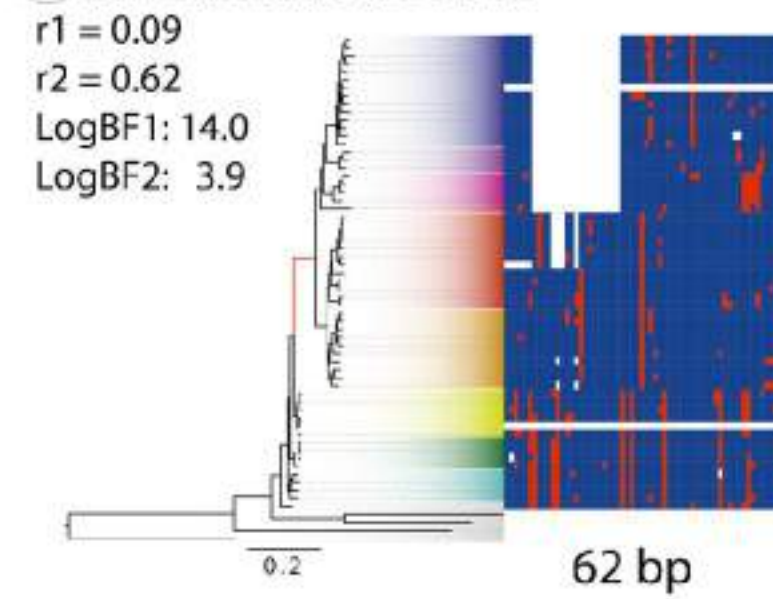




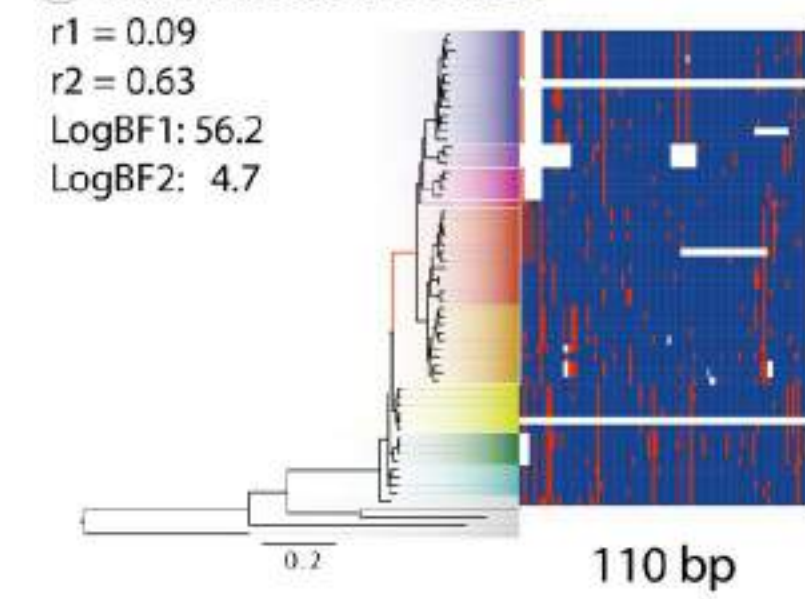
1 intronicCNEE156893



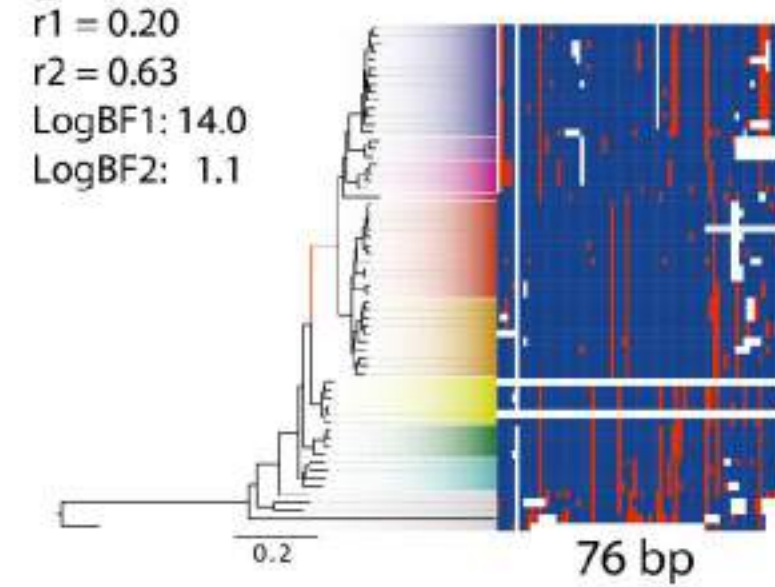
2 intronicCNEE156926



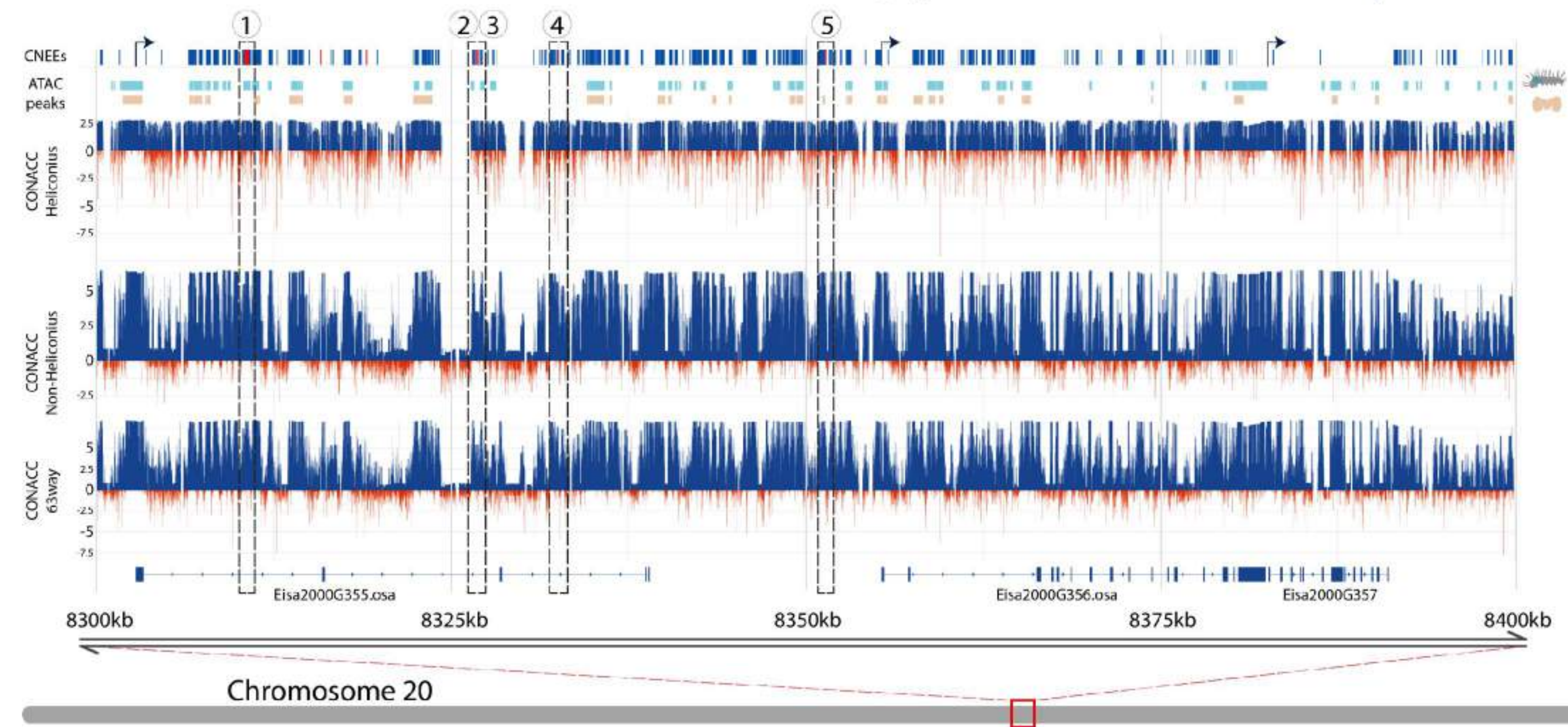
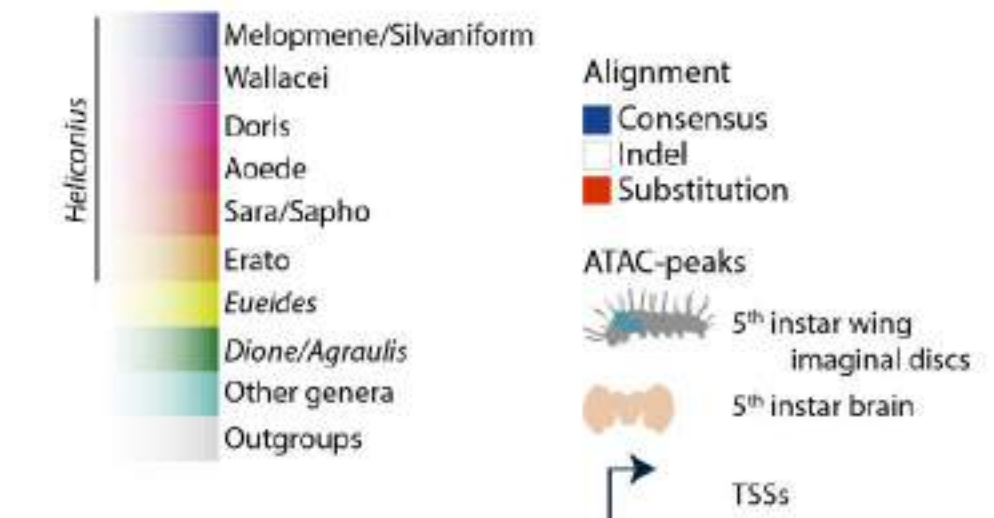
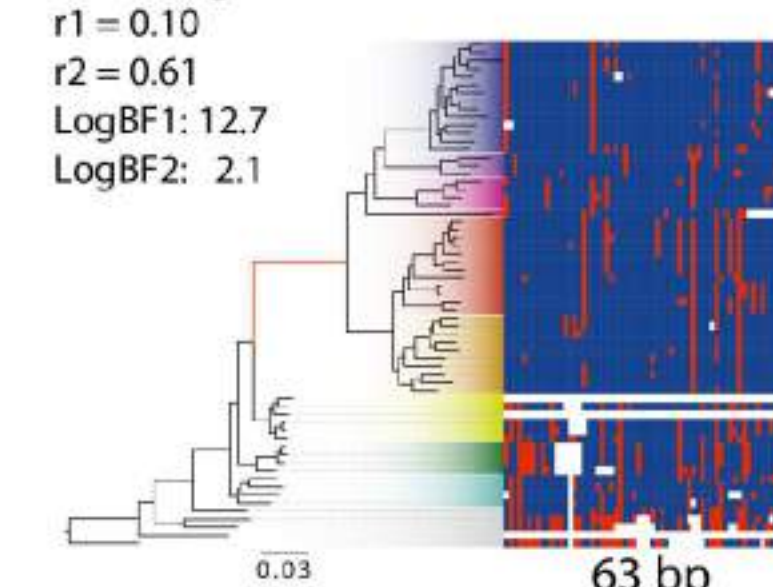
3 intronicCNEE156927



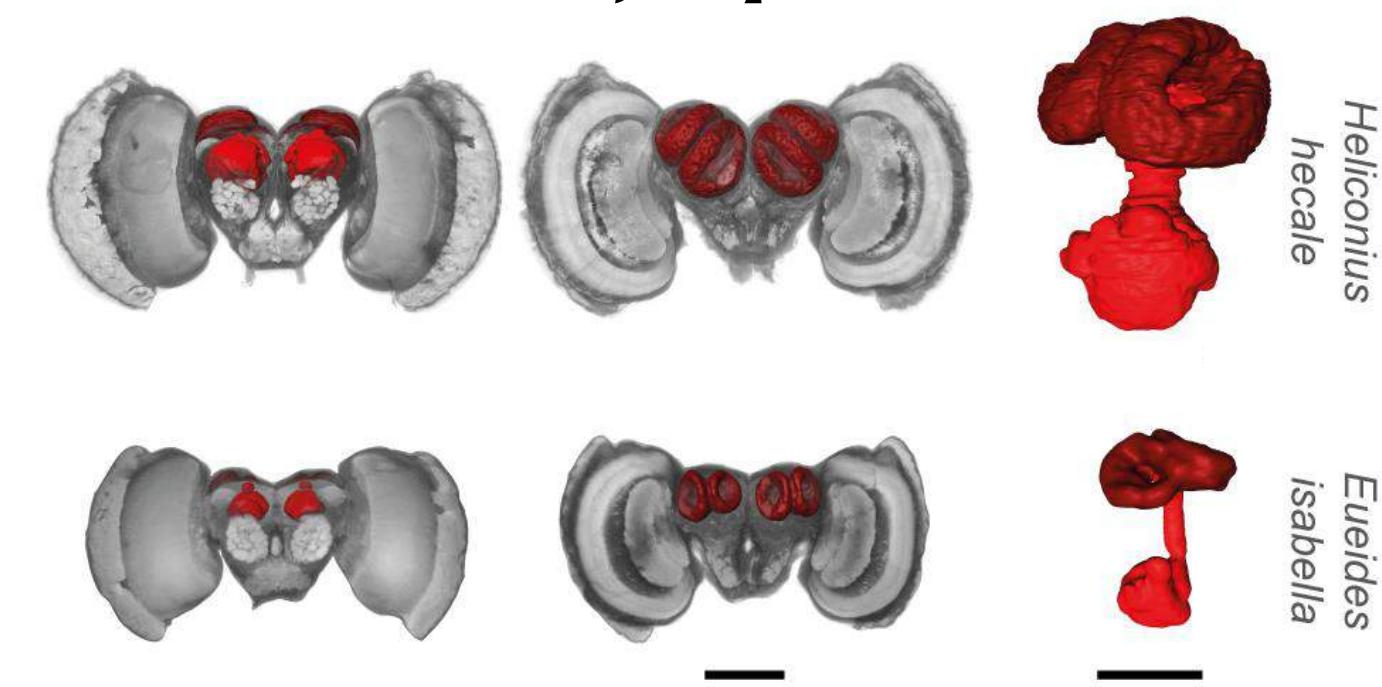
4 intronicCNEE156944



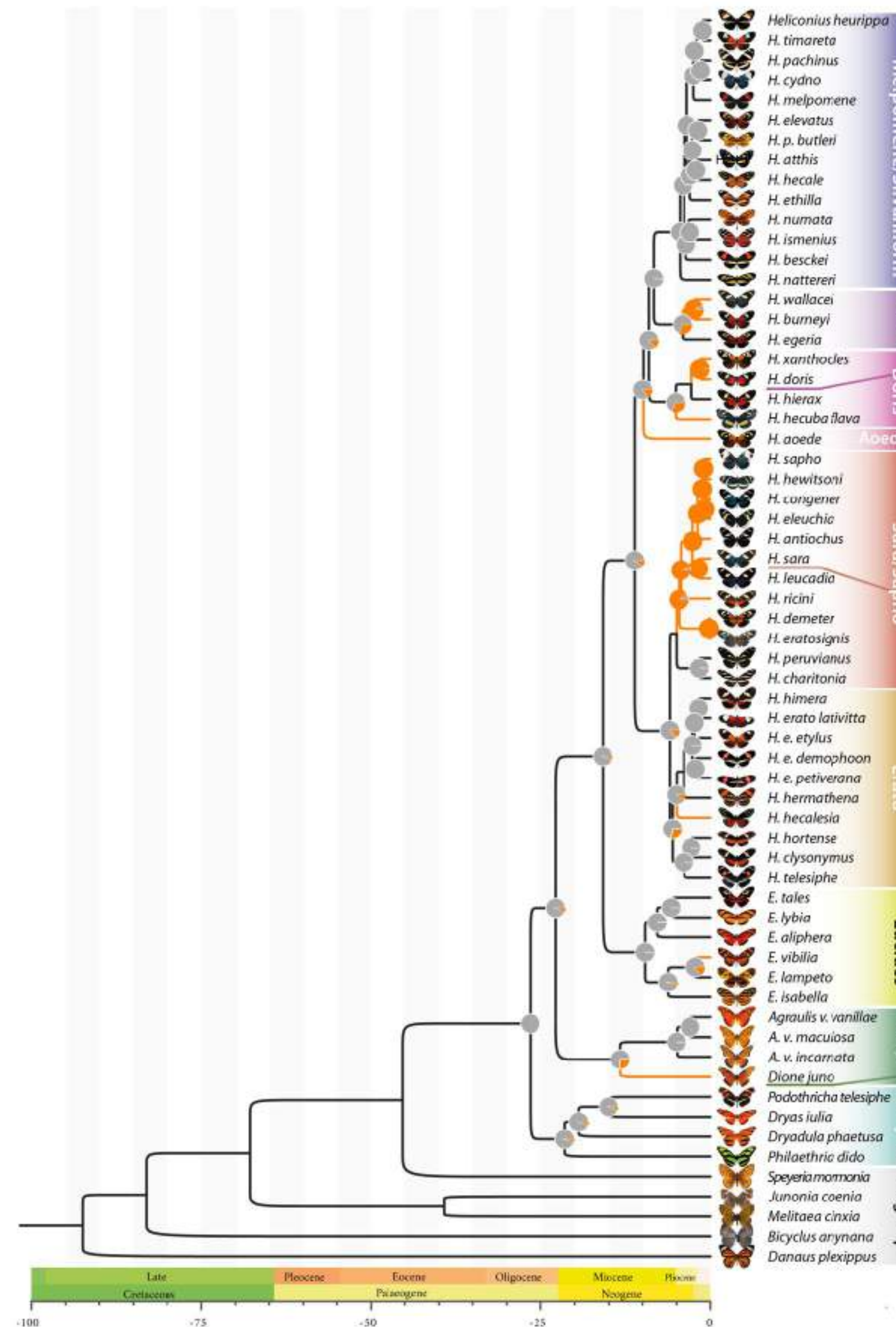
5 intergCNEE179234



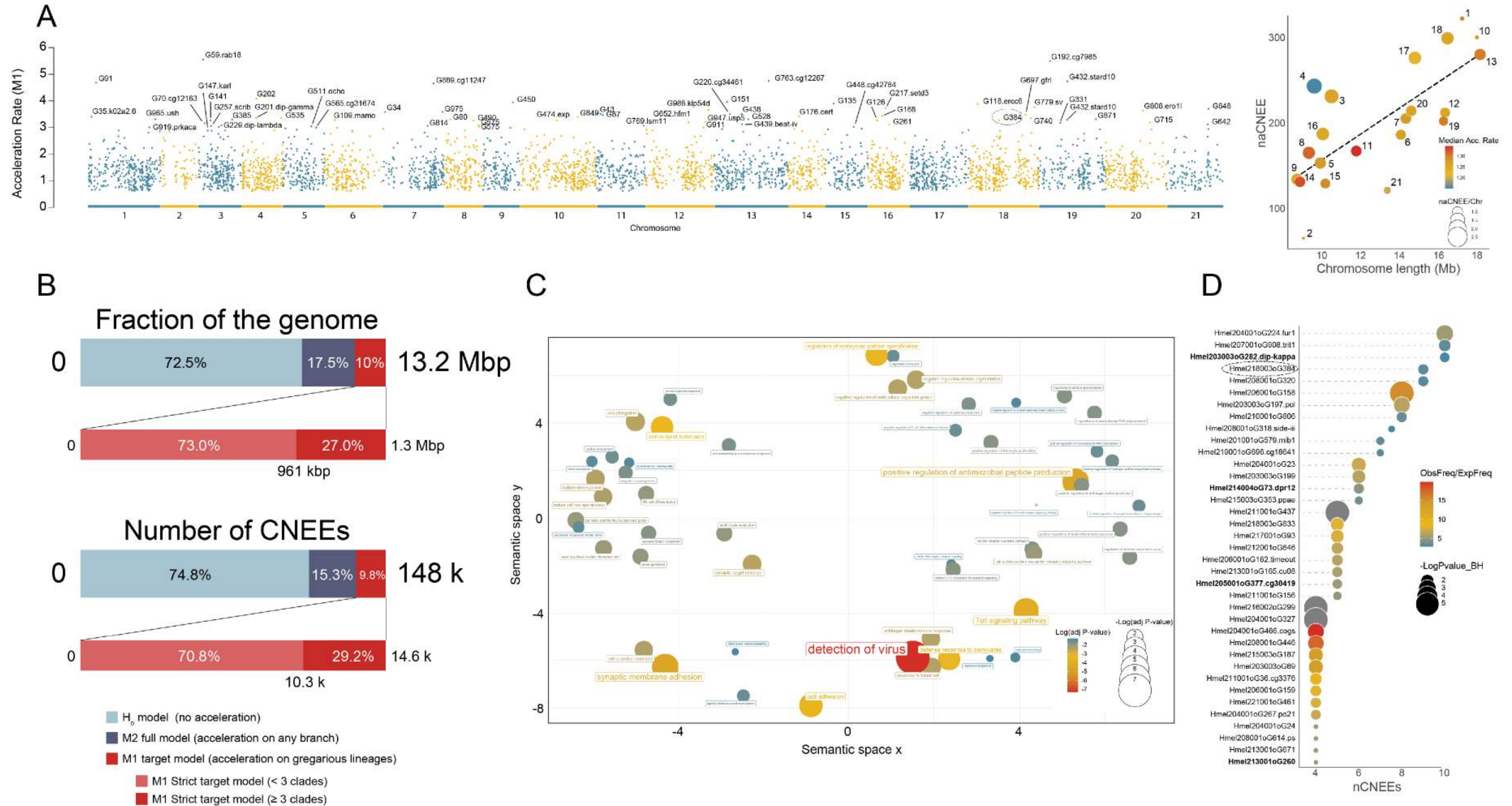
» Mushroom body expansion (Brain)



» Test for Convergence »



» Test for Convergence »



» Test for Convergence »

