Phylogenomics with reference genomes

Introduction to

Biodiversity Genomics

Tena, Ecuador

2024

Phylogenomics using reference genomes

- Use genome-scale sequence data to reconstruct evolutionary relationships

- Create species trees using one individual for each species, but thousands of markers

- "Phylogenomics draws information by comparing entire genomes, or at least large portions of genomes. Phylogenetics compares and analyzes the sequences of single genes, or a small number of genes,…”

Formenti et al. 2022

Phylogenomic exercise

- Download genomes from ncbi
- Extract BUSCO genes
- Cluster orthogroups and select single copy orthologs
  - Orthofinder
- Alignment - we need to compare homologous sites
  - mafft
- Filtering - remove gaps
  - trimal
- Phylogeny reconstruction, tree building
  - IQ-tree
- Visualisation
  - ggtree

Substitution models

- Substitution models
  - observed differences -> actual number of substitutions during the evolutionary history

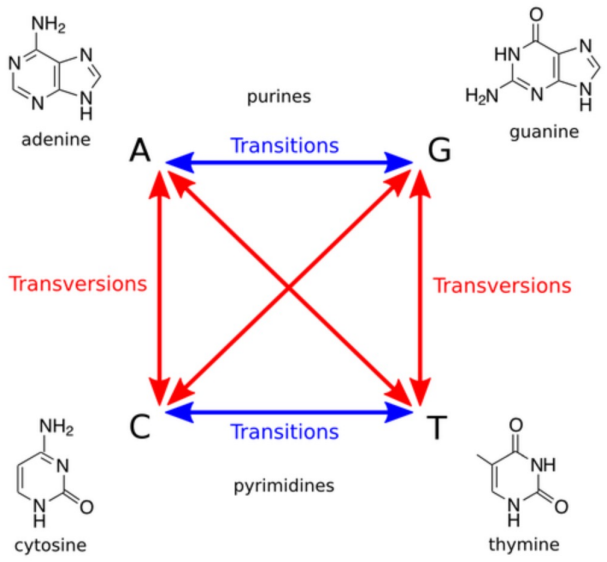
Substitution matrix

$$M_{ij} = \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} 1-3\epsilon & \epsilon & \epsilon & \epsilon \\ \epsilon & 1-3\epsilon & \epsilon & \epsilon \\ \epsilon & \epsilon & 1-3\epsilon & \epsilon \\ \epsilon & \epsilon & \epsilon & 1-3\epsilon \end{pmatrix} \end{array} \begin{array}{c} \\ A \\ C \\ G \\ T \end{array}$$

A-C, A-G, A-T, C-G, C-T and G-T

Substitution models, examples

JC or JC69	0	Equal substitution rates and equal base frequencies (Jukes and Cantor, 1969).
F81	3	Equal rates but unequal base freq. (Felsenstein, 1981).
K80 or K2P	1	Unequal transition/transversion rates and equal base freq. (Kimura, 1980).
HKY or HKY85	4	Unequal transition/transversion rates and unequal base freq. (Hasegawa, Kishino and Yano, 1985).
TN or TN93	5	Like HKY but unequal purine/pyrimidine rates (Tamura and Nei, 1993).
TNe	2	Like TN but equal base freq.
K81 or K3P	2	Three substitution types model and equal base freq. (Kimura, 1981).
K81u	5	Like K81 but unequal base freq.
GTR	8	General time reversal model, unequal rates and unequal base freq. (Tavare, 1986).



$$M_{ij} = \begin{pmatrix} 1-3\epsilon & \epsilon & \epsilon & \epsilon \\ \epsilon & 1-3\epsilon & \epsilon & \epsilon \\ \epsilon & \epsilon & 1-3\epsilon & \epsilon \\ \epsilon & \epsilon & \epsilon & 1-3\epsilon \end{pmatrix}$$

A-C, A-G, A-T, C-G, C-T and G-T

+F base frequences
+I invariate sites
+G gamma variation

http://www.iqtree.org/doc/Substitution-Models

Maximum Likelihood trees

- Starting tree (random, NJ (distance based) or parsimony)
- Calculates the likelihood – how well the model explains the data
- Rearranges the tree and recalculates the likelihood
  - NNI: Nearest Neighbour Interchange
  - SPR: Subtree Pruning and Re-grafting
  - TBR: Tree Bisection and Reconnection

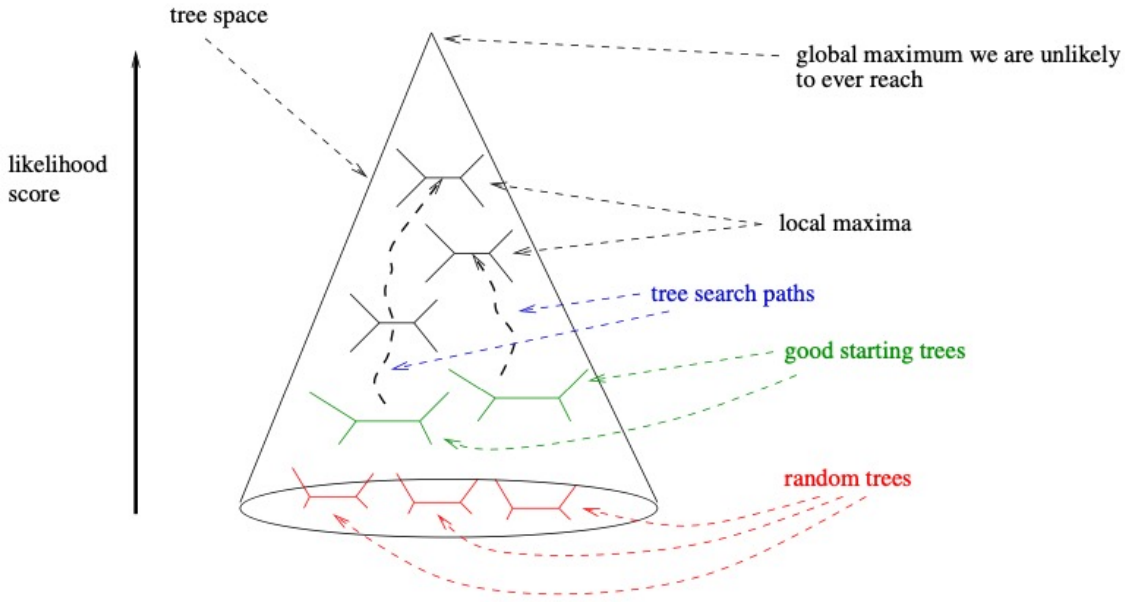
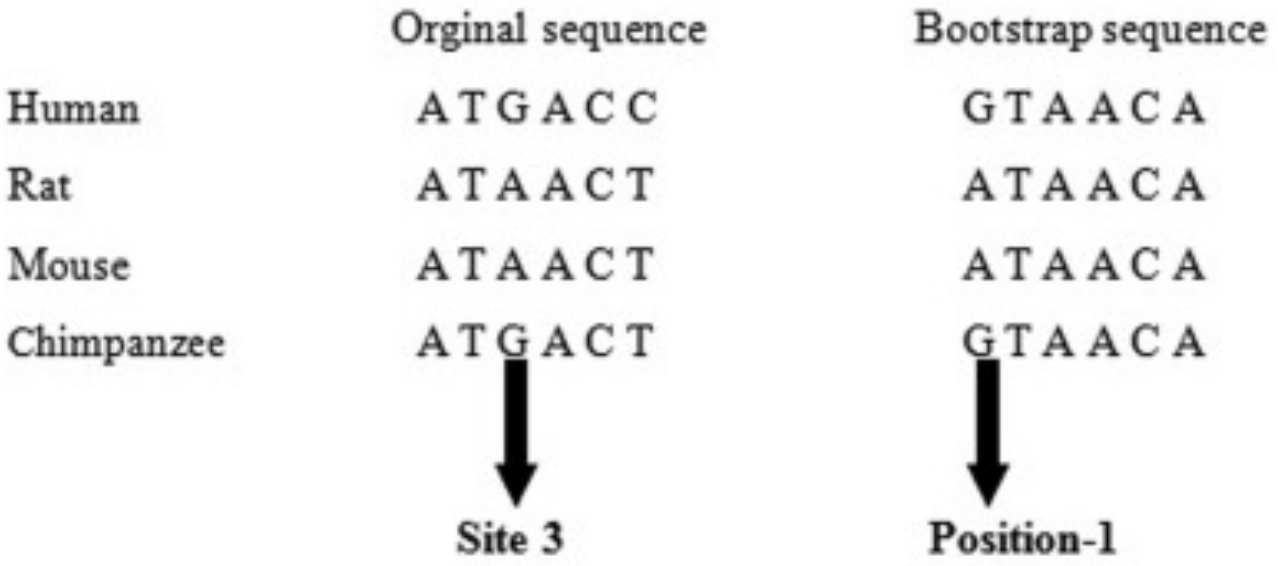


Figure 3 Our way of imagining tree search space, including random starting trees, "good" starting trees, and tree search paths that take us closer to the desired global maximum, that is *the* ML tree.

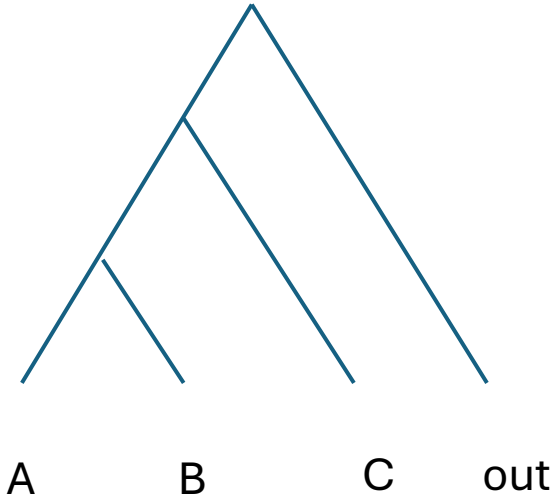
Testing robustness of the tree

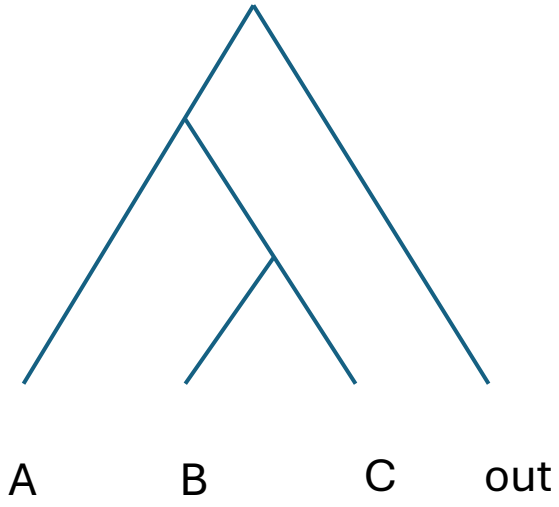
- Bootstrap
  - Resampling with replacement
  - How many times out of 100 the same branch is observed when repeating the generation of a phylogenetic tree on resampled sets of data.



Multiple trees - concordance

- Consensus tree
- Concatenation
- (Dis)agreement between trees
  - Gene tree discordance
    - number of trees supporting the branch
  - Robinson-Foulds (RF) distance
    - number of partitions in each tree that is not found in the other





Ancestral reconstruction

- Ancestral reconstruction allows inference of rates and patterns of evolutionary change through time

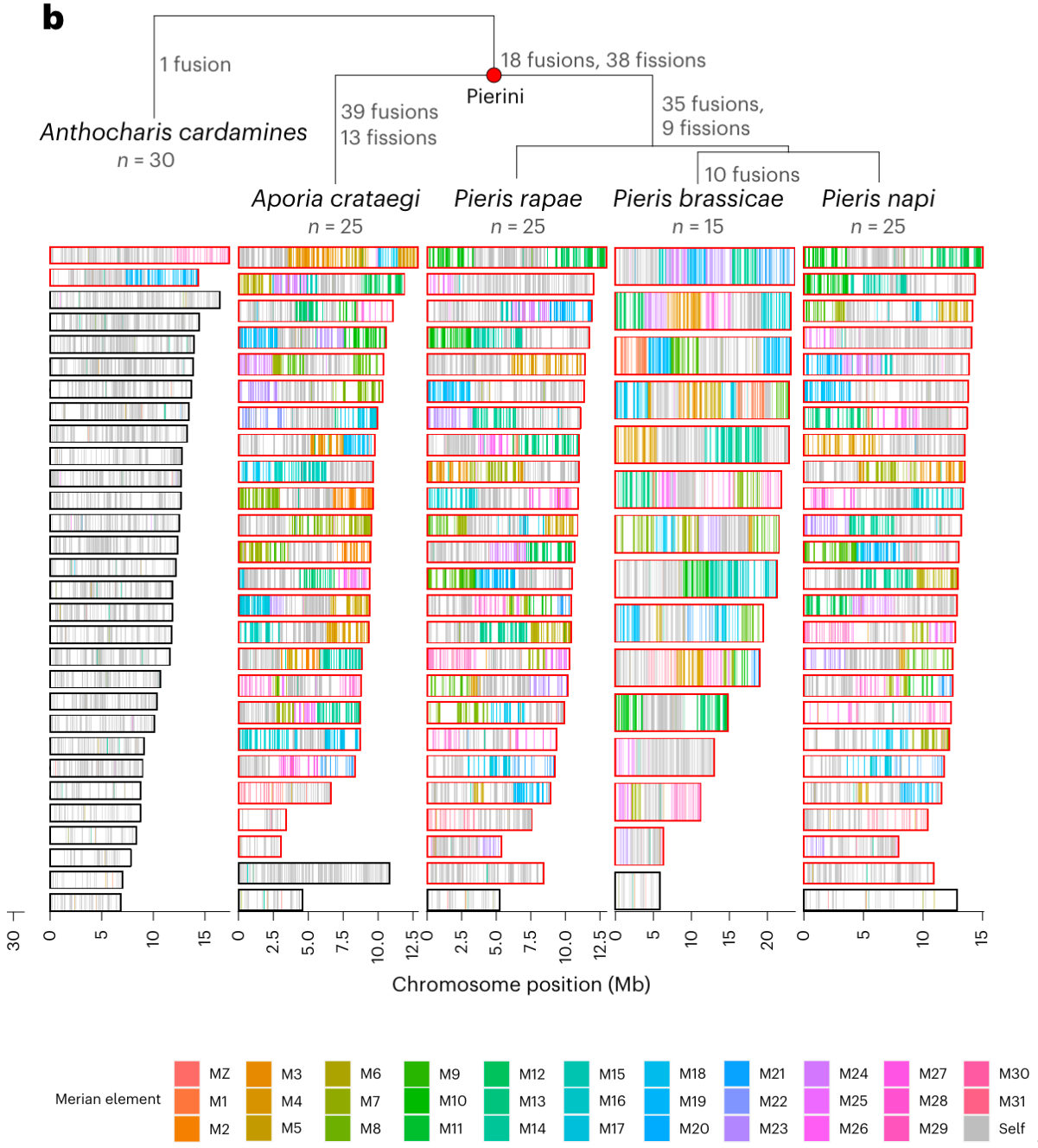
- Reconstruct
  - Nucleotide sequences
  - Genome composition
  - Synteny

Methods

- InferCAR
- AGORA
- Syngraph
- Deschrambler
- Cactus, HAL-toolset

Ancestral reconstruction

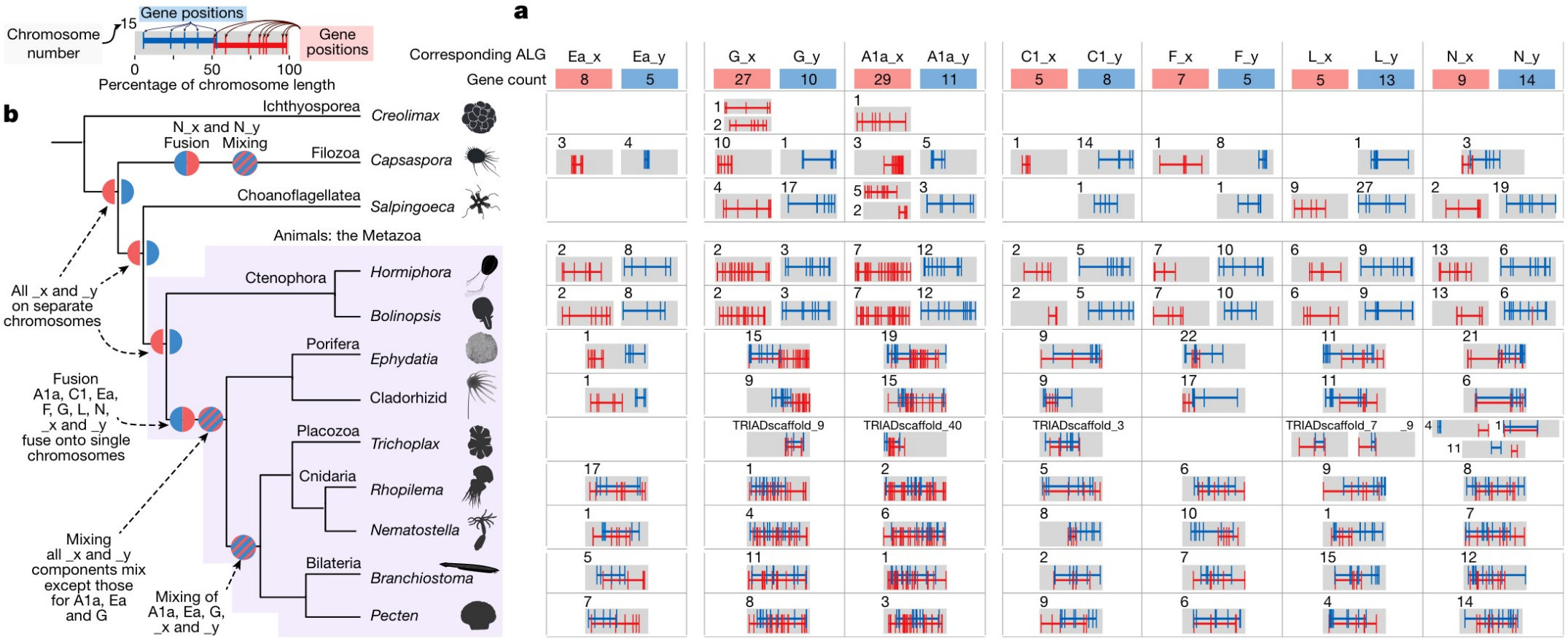
- Wright et al 2024
- BUSCO to identify single copy orthologues (SCO)
- OrthoFinder cluster the SCO into orthogroups
- Syngraph – parsimony inference of linkage groups in the last common ancestor



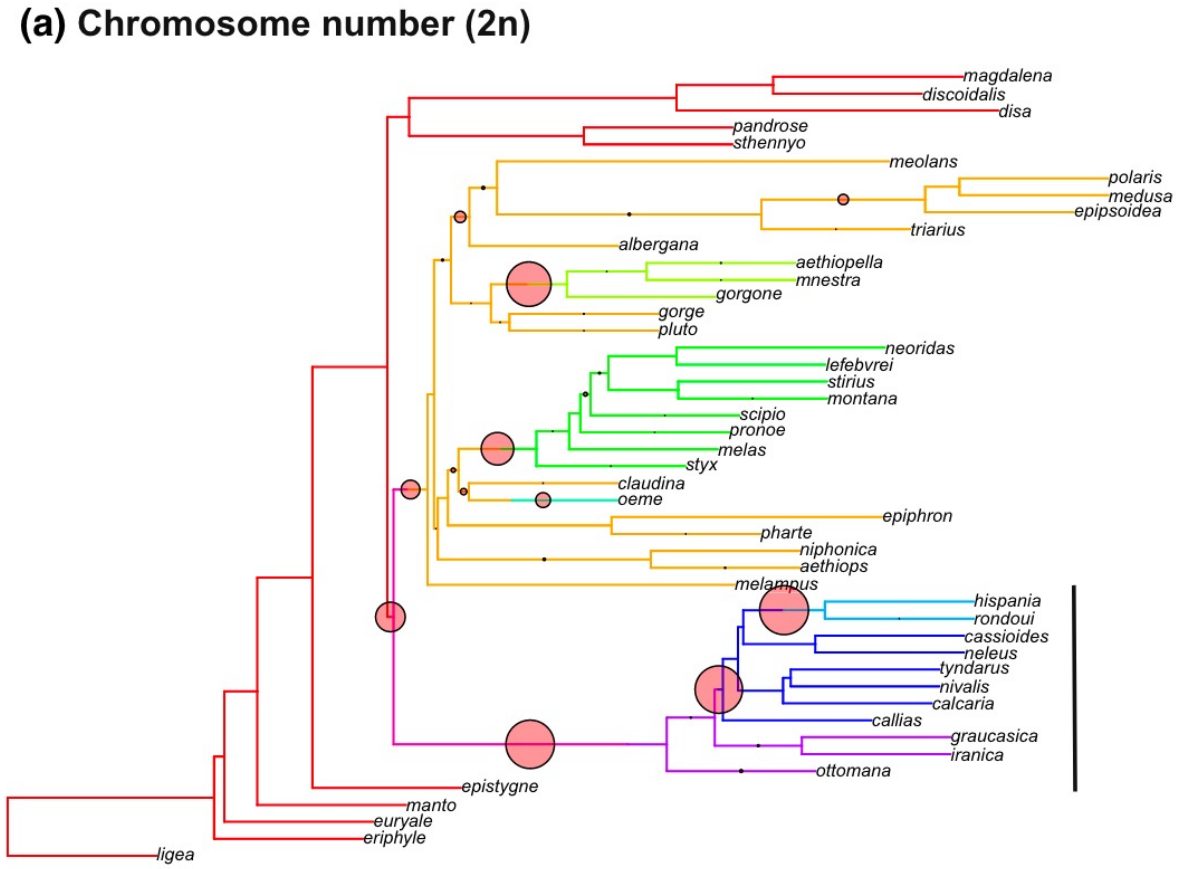
Wright et al 2024

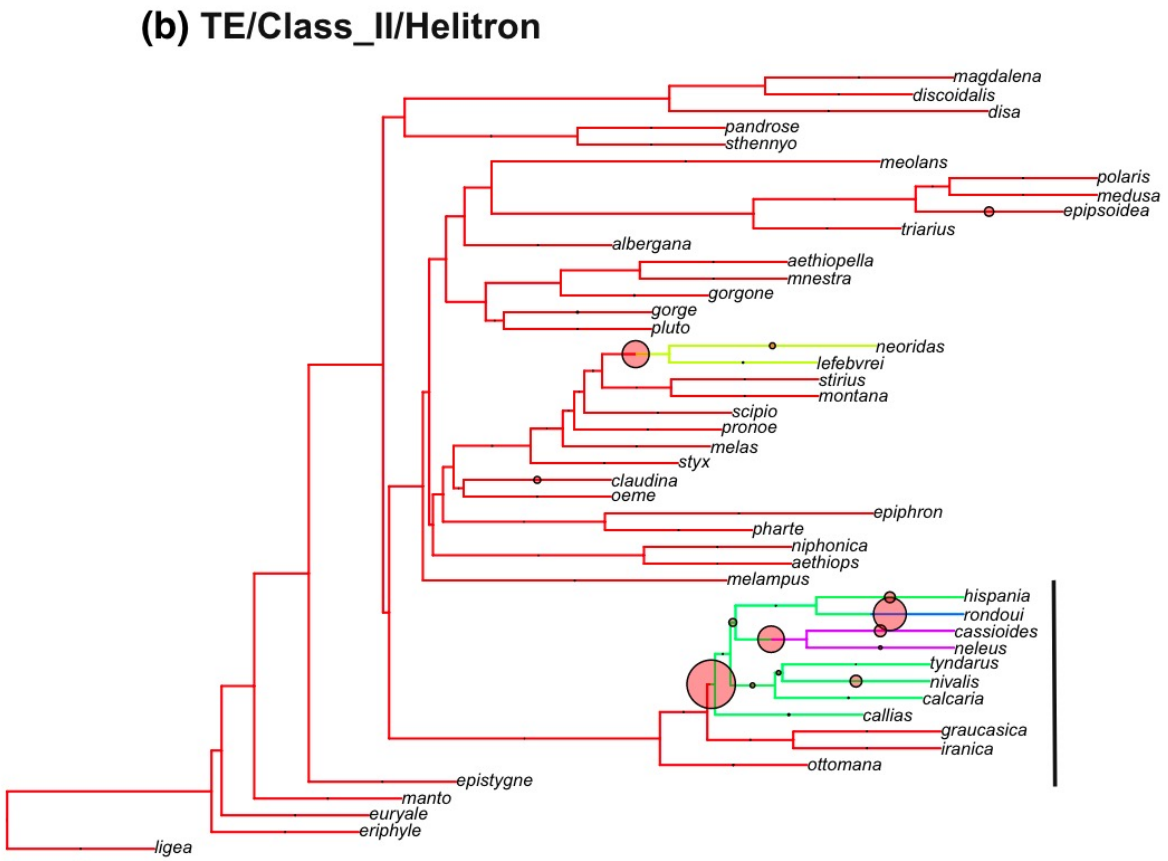
Resolve phylogenetic relationships

- Ctenophora likely basal to all other Metazoa



Dynamics through time





Cornet et al 2023

Inference of selection - d_N/d_S

- Genome from one individual per species can be used
- Number of substitutions per synonymous site (d_S)
- Number of substitutions per non-synonymous site (d_N)
- Usually inferred per gene or window based over multiple genes

- Under neutrality: $d_N = d_S$
- Purifying selection: $d_N/d_S < 1$
- Positive selection: $d_N/d_S > 1$

Inference of selection - d_N/d_S

- Medium divergence
  - total branch length > 0.5 (d_S)
  - Recently diverged –low number of fixed differences,
  - Very distant – substitution saturation (recurrent mutations)

- Codon alignment of coding nucleotide sequences

- paml – model comparison detecting positive selection
- mapNH – time heterogeneous models, account for non-stationary GC-content

Phylogenomic exercise

- BUSCO genes
- Cluster orthogroups and select single copy orthologs - Orthofinder
- Alignment - we need to align homologous sites - mafft
- Filtering - reduce trimal
- Phylogeny reconstruction, tree building
  - IQ-tree