In Comparative Genomics, All Roads Lead to HOGs

Reviews in Quantitative Biology

Natasha Glover

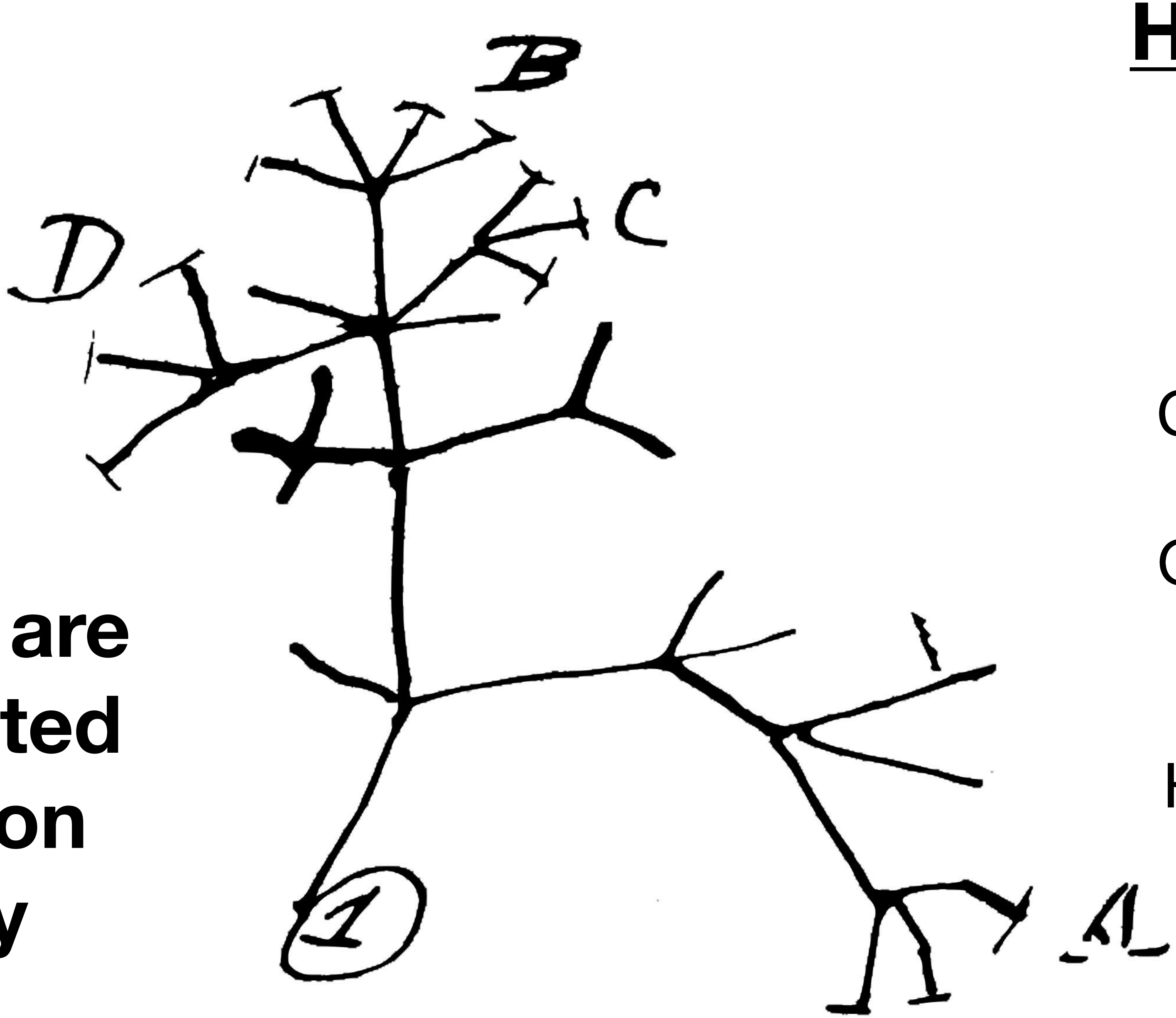
6 Nov 2020

Target audience

- Biologists interested in gene families, comparative genomics, phylogenetics, evolutionary biology
- Not a talk structured on methodology, but more of motivation and use cases for Hierarchical Orthologous Groups (HOGs)

Orthology and Paralogy

Homologs are genes related by common ancestry



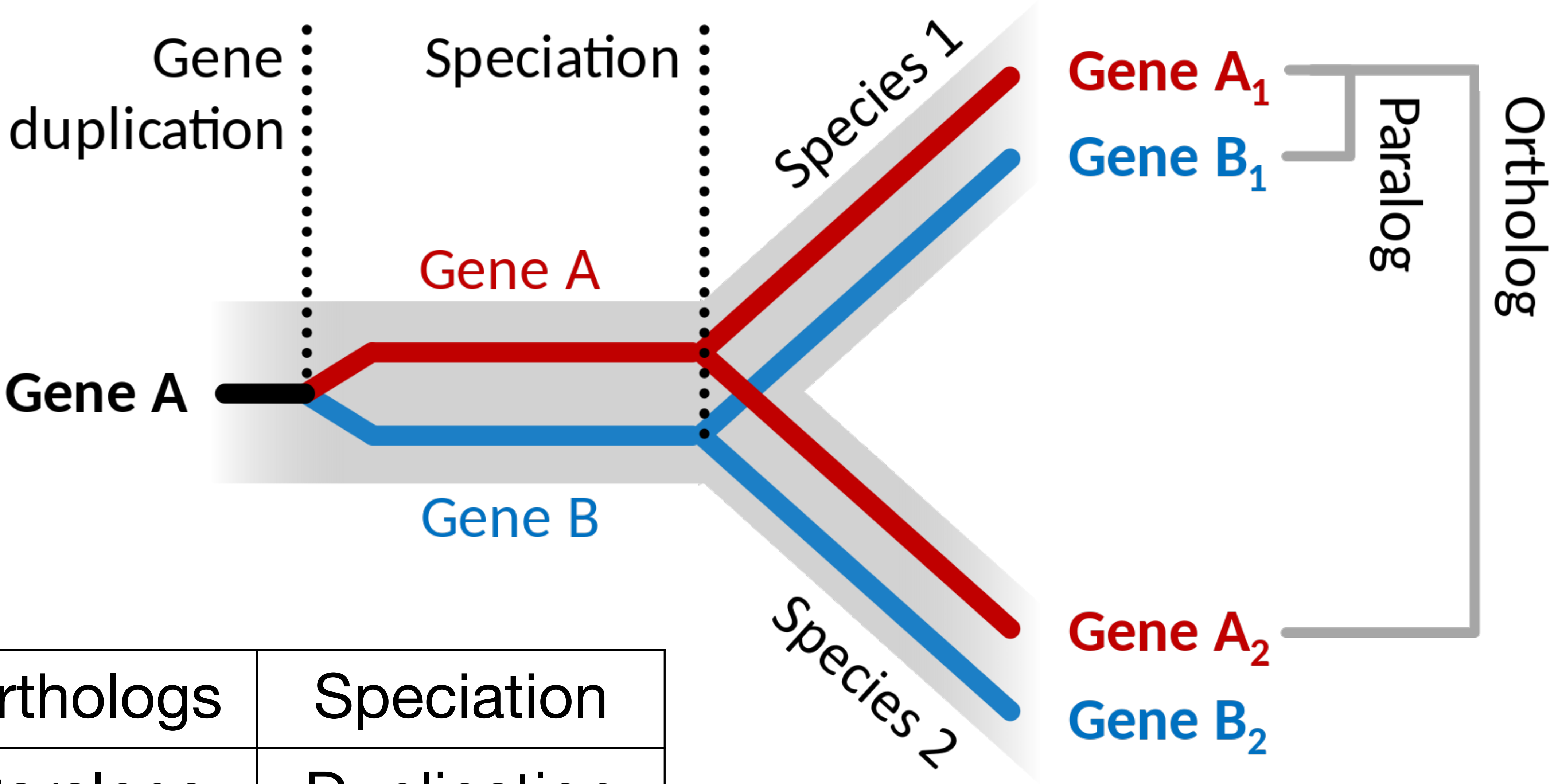
Homologs

Ortholog
Paralog
Ohnolog
Xenolog
Co-ortholog
In-paralog
Out-paralog
Syntelog
Paleolog
Homoeolog

Definition of orthology

- The concepts of orthology and paralogy were introduced in by Walter Fitch in 1970.
- Orthologous genes are the result of speciation so that the history of the gene reflects the history of the species.
  - *ortho=exact*
- Paralogous genes are the result of gene duplication. Both copies have descended in parallel during the history of an organism.
  - *para = next to*

Fitch 1970. https://doi.org/10.2307/2412448

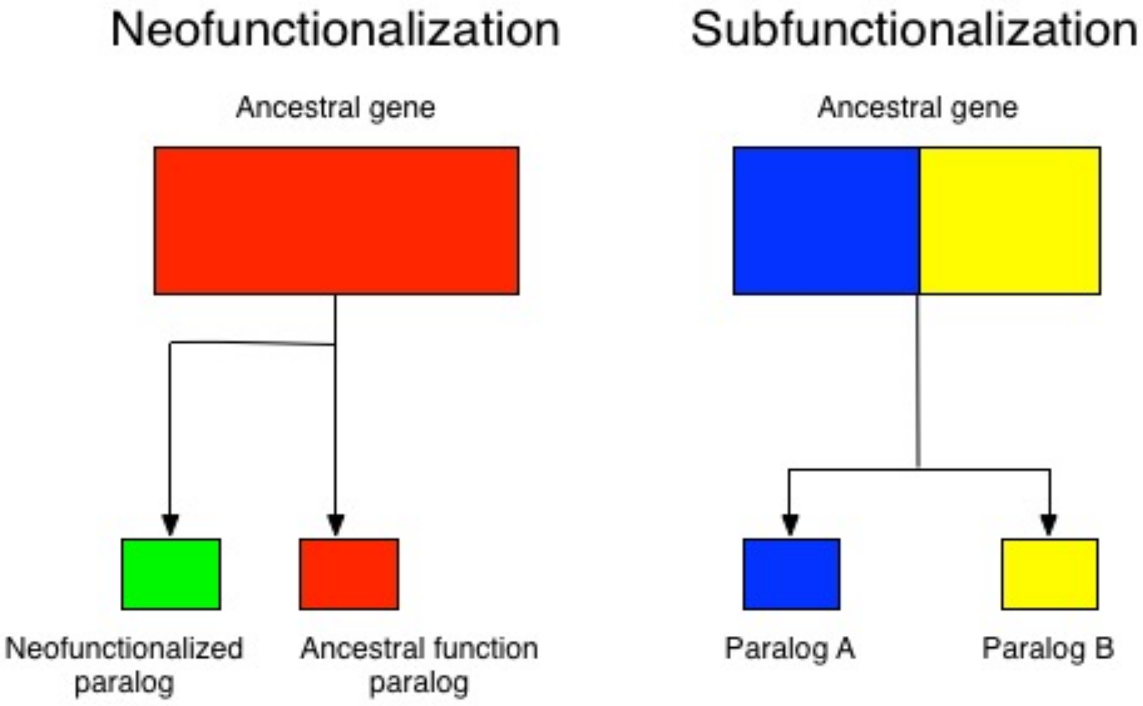


Orthologs	Speciation
Paralogs	Duplication

https://en.wikipedia.org/wiki/Sequence_homology

The value of distinguishing orthologs vs. paralogs

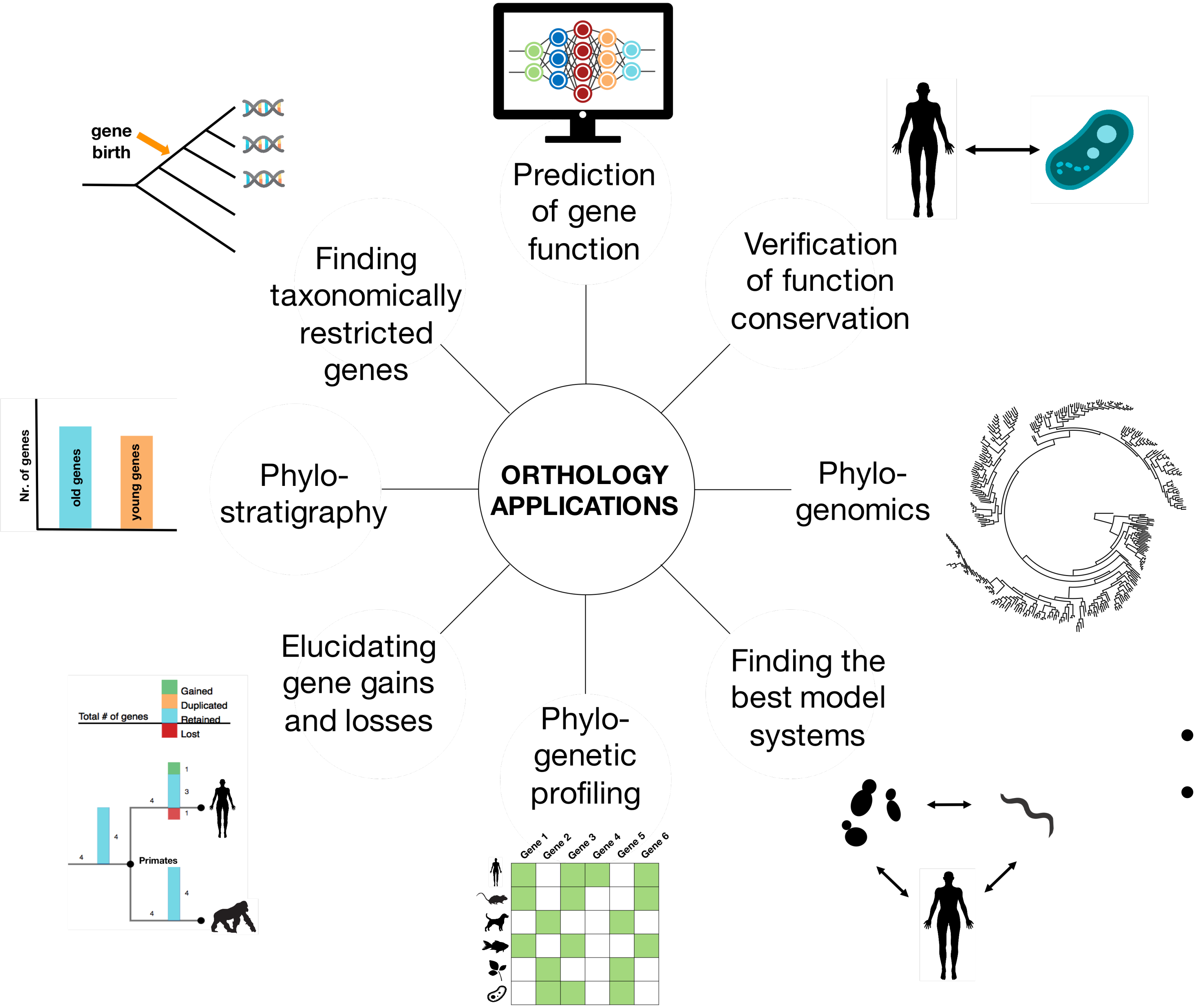
- Since orthologs arise by speciation, orthologs reflect the same evolutionary history as the underlying species
    - Can be used to make phylogenetic species trees
- True orthologs are likely to retain the same function over evolutionary time (probably)
    - Paralogs are more likely to diverge in function



https://liorpachter.wordpress.com/tag/neofunctionalization/

The ortholog conjecture
Stamboulian et al 2020 https://doi.org/10.1093/bioinformatics/btaa468
Altenhoff et al 2012 https://doi.org/10.1371/journal.pcbi.1002514

There are many applications of orthology and paralogy



ORTHOLOGY APPLICATIONS

Prediction of gene function

Verification of function conservation

Phylo-genomics

Finding the best model systems

Phylo-genetic profiling

Elucidating gene gains and losses

Phylo-stratigraphy

Finding taxonomically restricted genes

- synteny
- gene families

Glover et al 2019 https://doi.org/10.1093/molbev/msz150

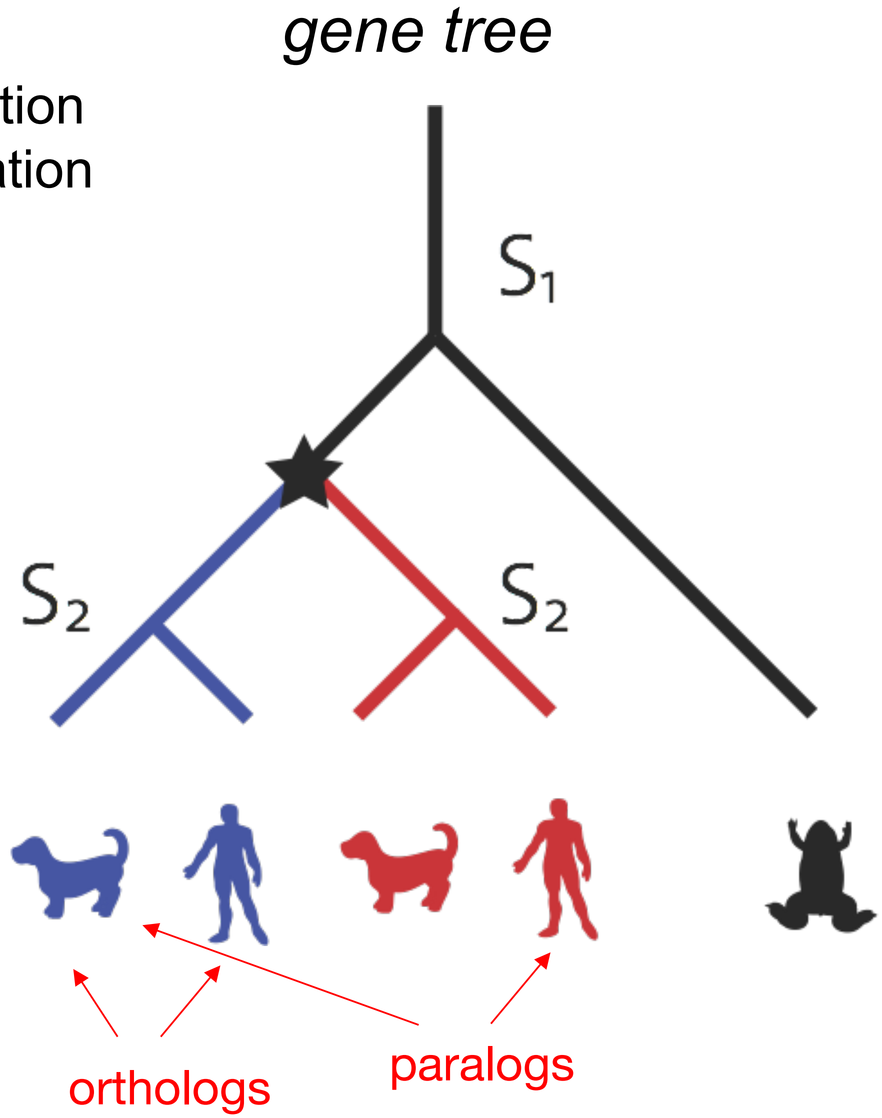
But it’s not so easy…

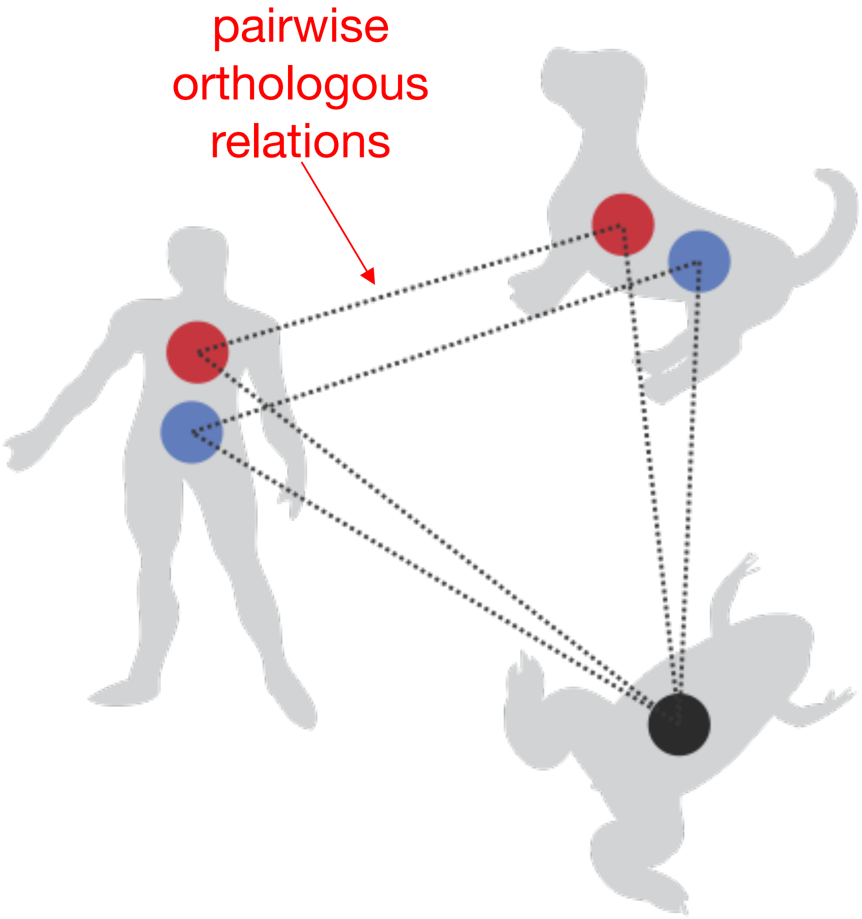
- Evolutionary scenarios and relationships become complicated when considering more than a pair of genes (multiple paralogs or species involved), with complex combinations of lineage-specific gene duplications, losses (and even horizontal gene transfer when speaking of bacteria)

Orthology

S = speciation

☆ = duplication



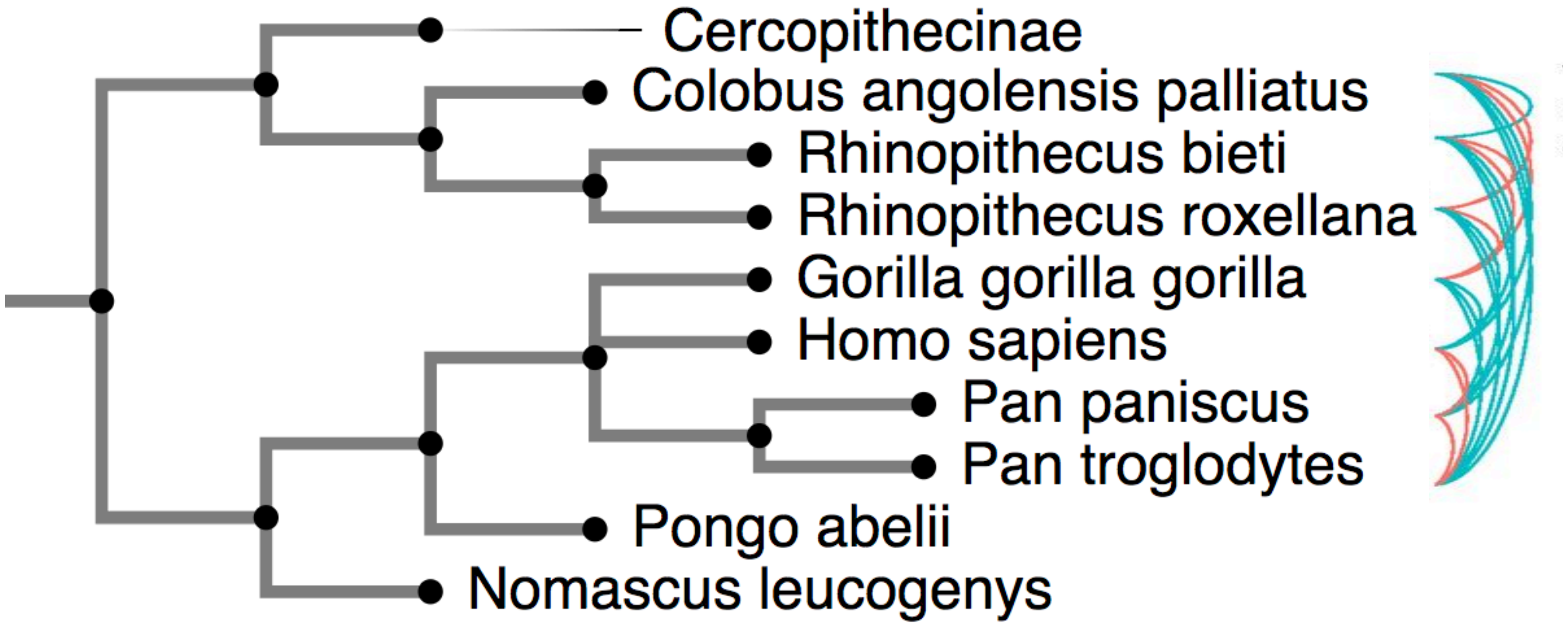


Altenhoff et al 2019 doi: 10.1007/978-1-4939-9074-0_5.

Why do we need Hierarchical Orthologous Groups?

Roadblock 1: Pairwise genome comparisons

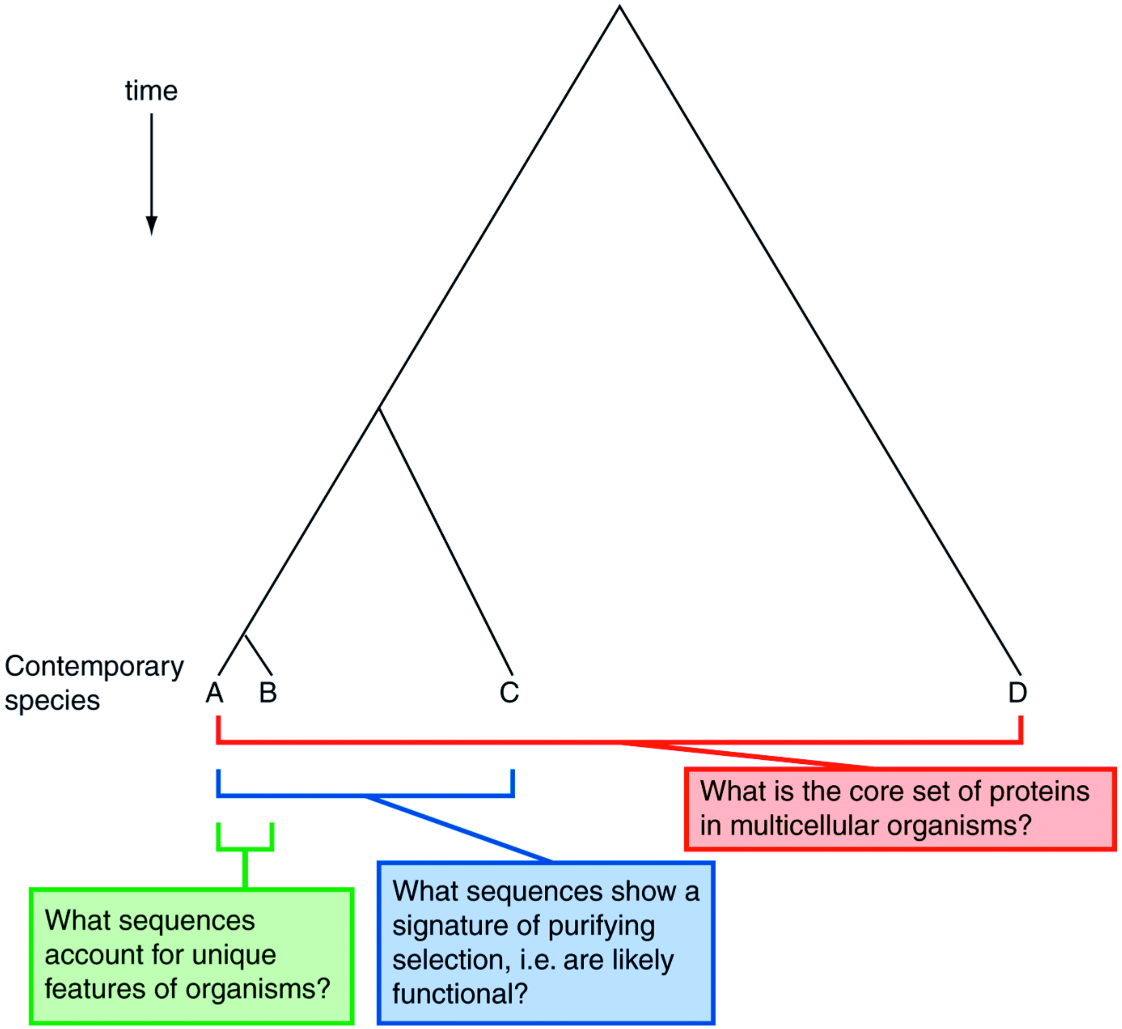
Pairwise genome comparisons



Best bidirectional hits between pairs of genomes considered as orthologs

The problem with pairwise genome comparisons

- Many analyses require orthologous relations over more than 2 genomes at a time
    - Comparative genomics, phylogenetics

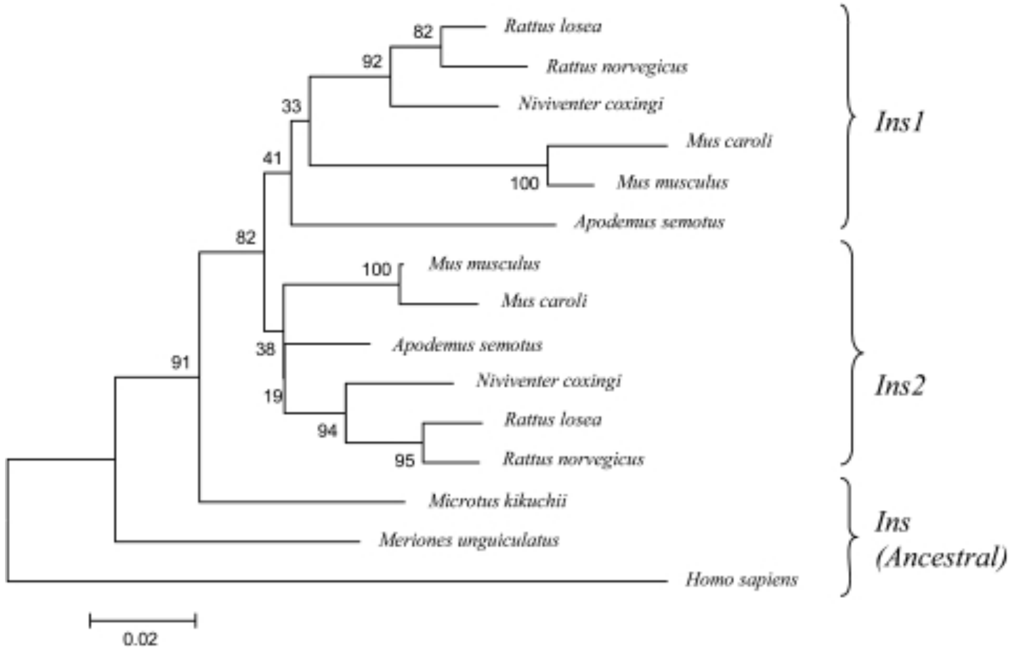

time
Contemporary species
A
B
C
D
What is the core set of proteins in multicellular organisms?
What sequences show a signature of purifying selection, i.e. are likely functional?
What sequences account for unique features of organisms?

Hardison, 2003. https://doi.org/10.1371/journal.pbio.0000058

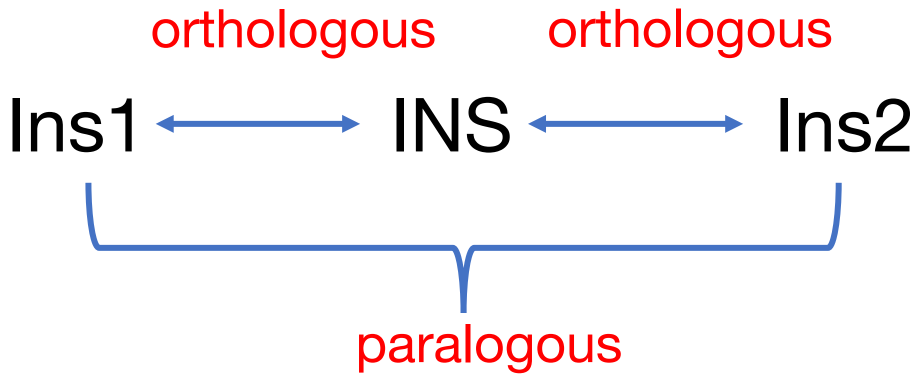
The problem with pairwise genome comparisons

- Orthology relationships are non-transitive
- If gene A is orthologous to B, and B is orthologous to C, it does not mean that A and B are orthologous to each other.

Fernández et al 2019. https://arxiv.org/pdf/1903.04530.pdf



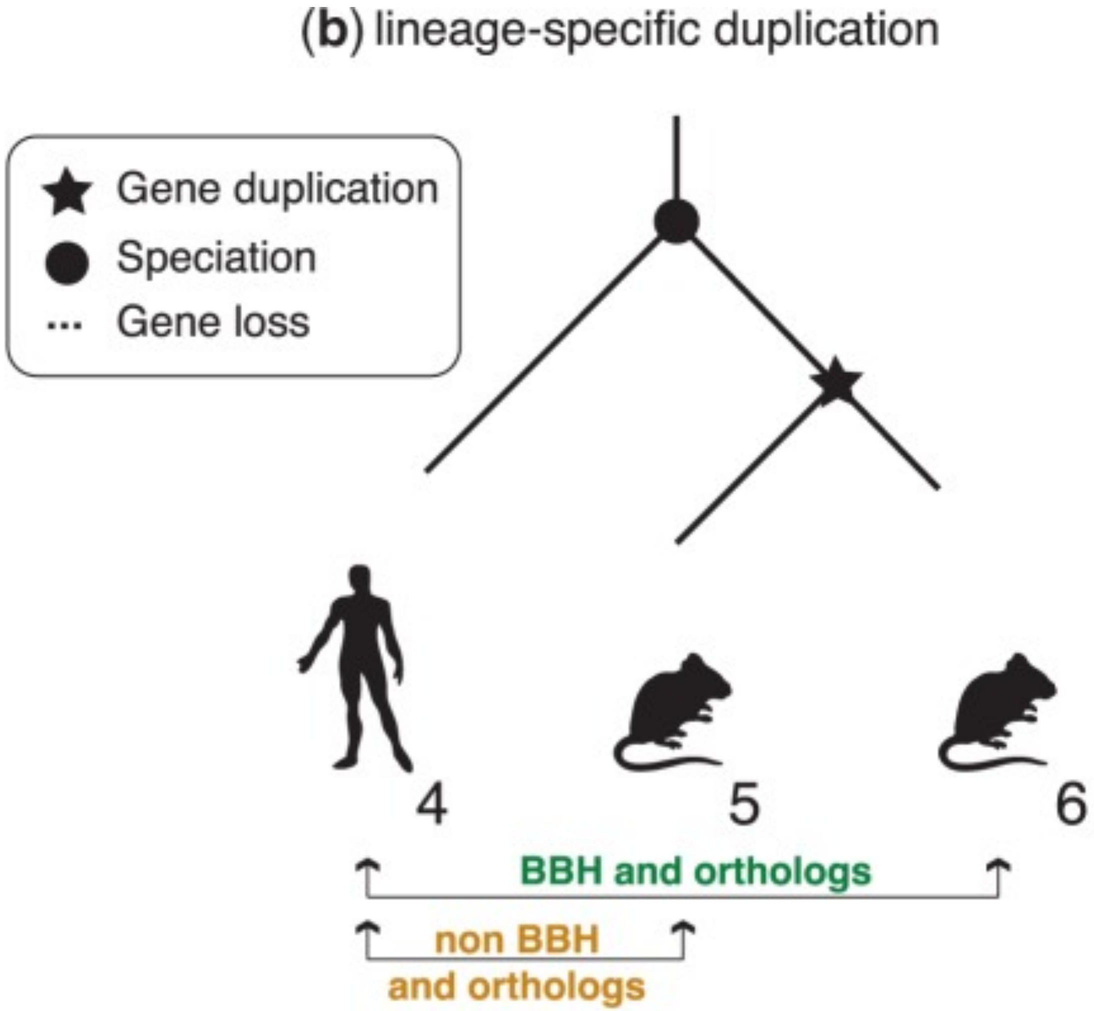
Shiao et al. 2008. doi: 10.1534/genetics.108.087023



The generalisation between multiple orthologs and paralogs is not a straightforward

The problem with pairwise genome comparisons

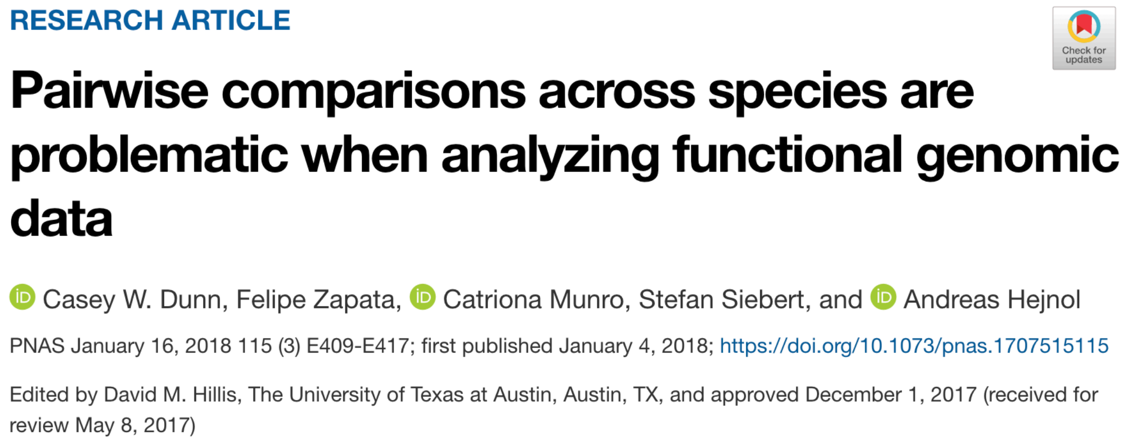
- Pairwise comparisons are likely to have false negatives (and sometimes false positives)
- This is due to high levels of duplication, differential gene loss, or variability in the rate of gene evolution



Dalquen et al 2013. doi: 10.1093/gbe/evt132

The problem with pairwise genome comparisons

- Pairwise methods can bias results
- Pairwise comparisons (as opposed to phylogenetic comparisons) are not independent, i.e. they repeatedly sample the same evolutionary changes
- Pairwise comparisons show current patterns, rather than historical processes



RESEARCH ARTICLE

Pairwise comparisons across species are problematic when analyzing functional genomic data

Casey W. Dunn, Felipe Zapata, Catriona Munro, Stefan Siebert, and Andreas Hejnol

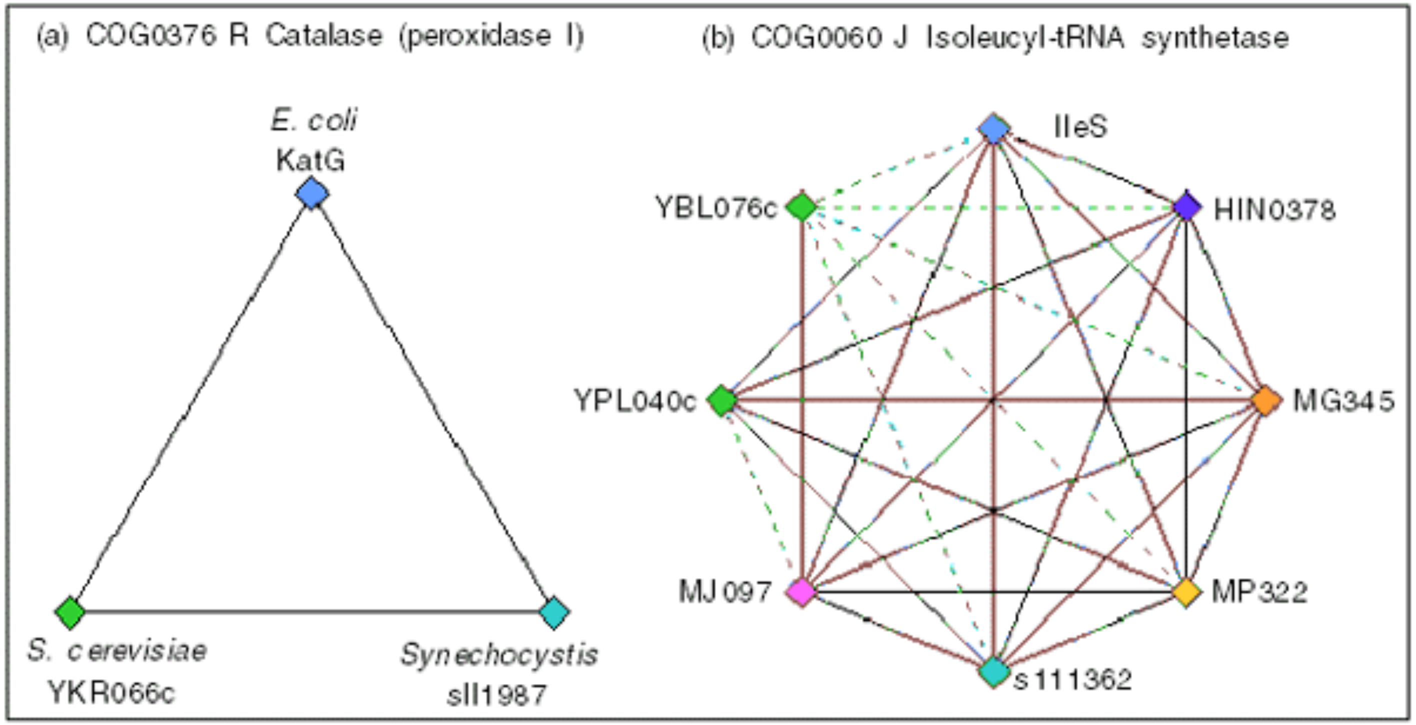
PNAS January 16, 2018 115 (3) E409-E417; first published January 4, 2018; https://doi.org/10.1073/pnas.1707515115

Edited by David M. Hillis, The University of Texas at Austin, Austin, TX, and approved December 1, 2017 (received for review May 8, 2017)

https://www.pnas.org/content/115/3/E409

A solution to pairwise comparisons

- It is useful to go from pairs to orthologous groups.
- Orthologous groups are clusters of orthologs and paralogs from multiple species

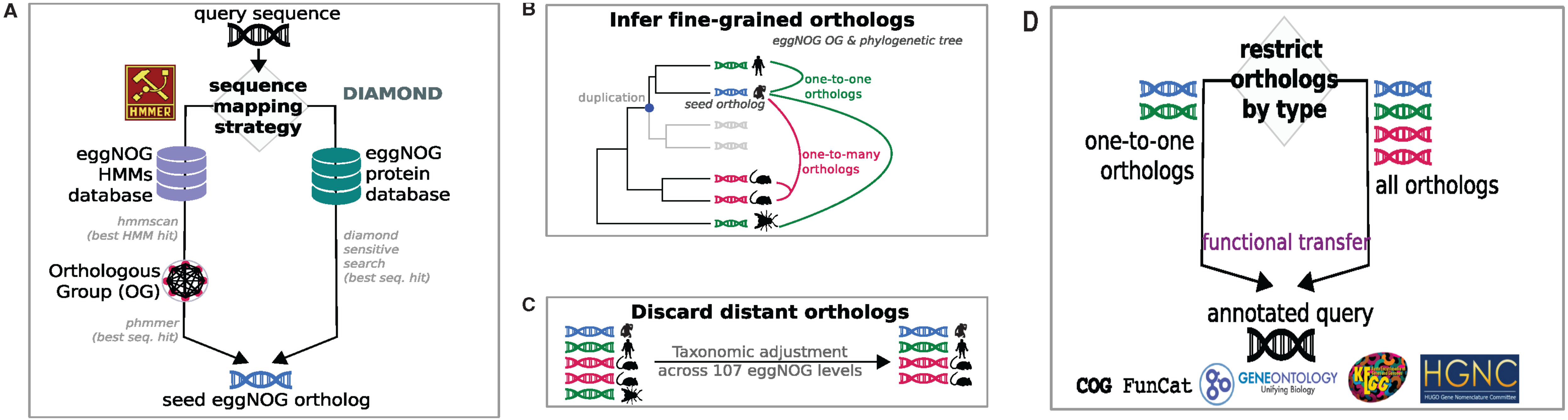


https://www.ncbi.nlm.nih.gov/Web/Newsltr/feb98.html
Tatusov et al 2000 doi: 10.1093/nar/28.1.33

Benefits and uses of Orthologous Groups

- Combines information from multiple species
- Can highlight divergence and conservation of gene families and biological processes
- Can improve orthology inference
- Can come to a consensus on information based on properties of genes in the orthologous group
  - Useful in functional annotation

Orthologous Groups useful for functional annotation



map a gene of unknown function to an ortholog in the database

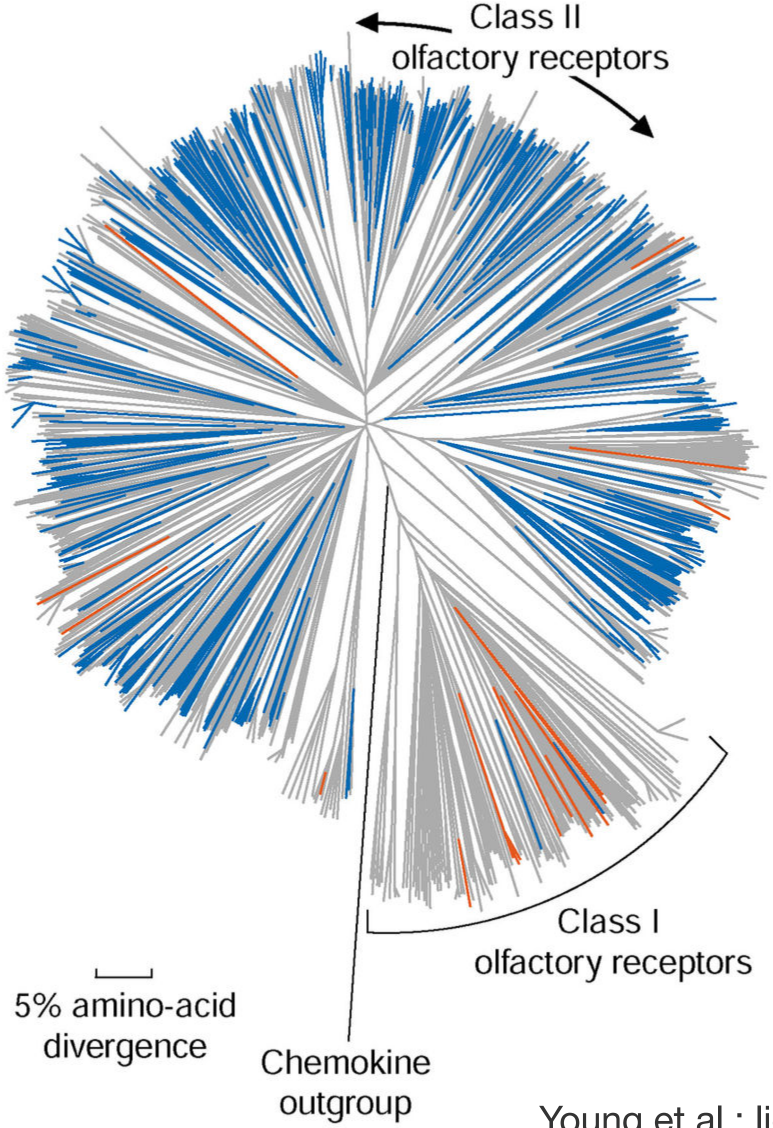
do some filtering

propagate the functional annotation to the query gene

Huerta-Cepas et al 2017. https://doi.org/10.1093/molbev/msx148

Benefits and uses of Orthologous Groups

- Orthologous groups can be thought of as a gene family

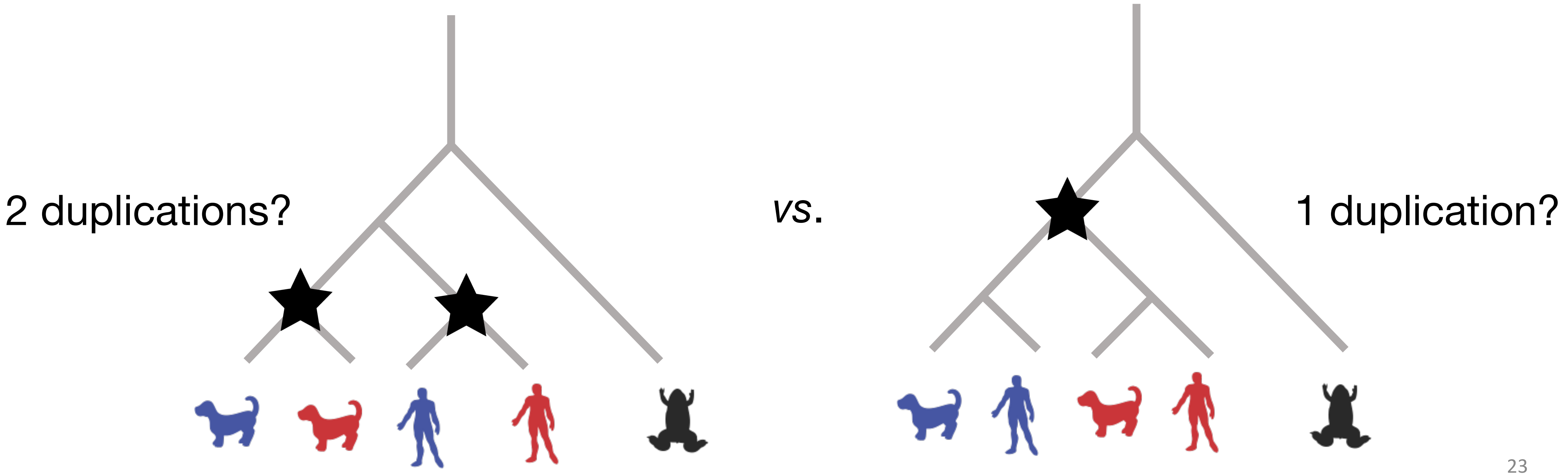


Young et al.; licensee BioMed Central Ltd. 2003

Roadblock 2: non-hierarchical orthologous groups

The problem with non-hierarchical orthologous groups

- Grouping does not have evolutionary meaning
- i.e. no information about speciation and duplication events



2 duplications?

vs.

1 duplication?

Hierarchical Orthologous Groups (HOGs)

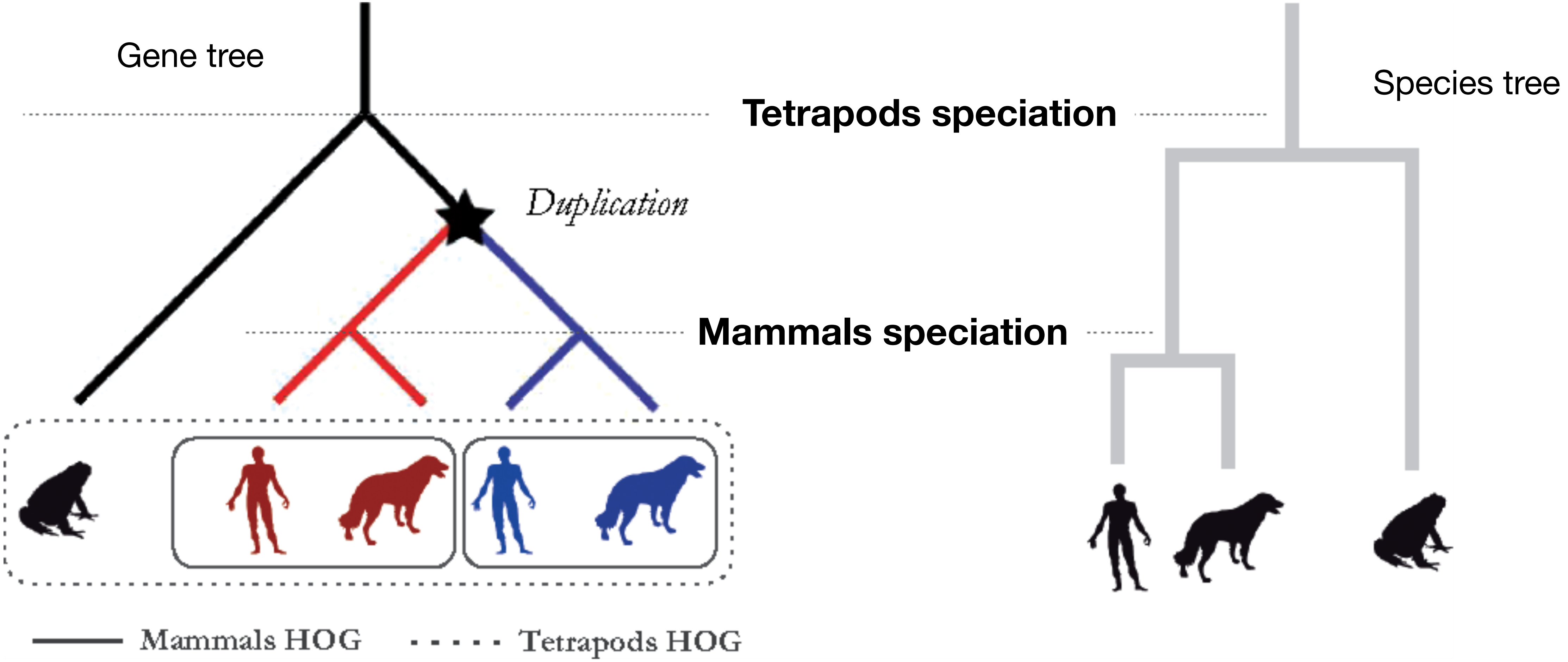
- Set of genes all descendant from a single common ancestral gene at a specific taxonomic range

or

- Sub-tree in a labelled gene tree rooted by a speciation node at a specific taxonomic range

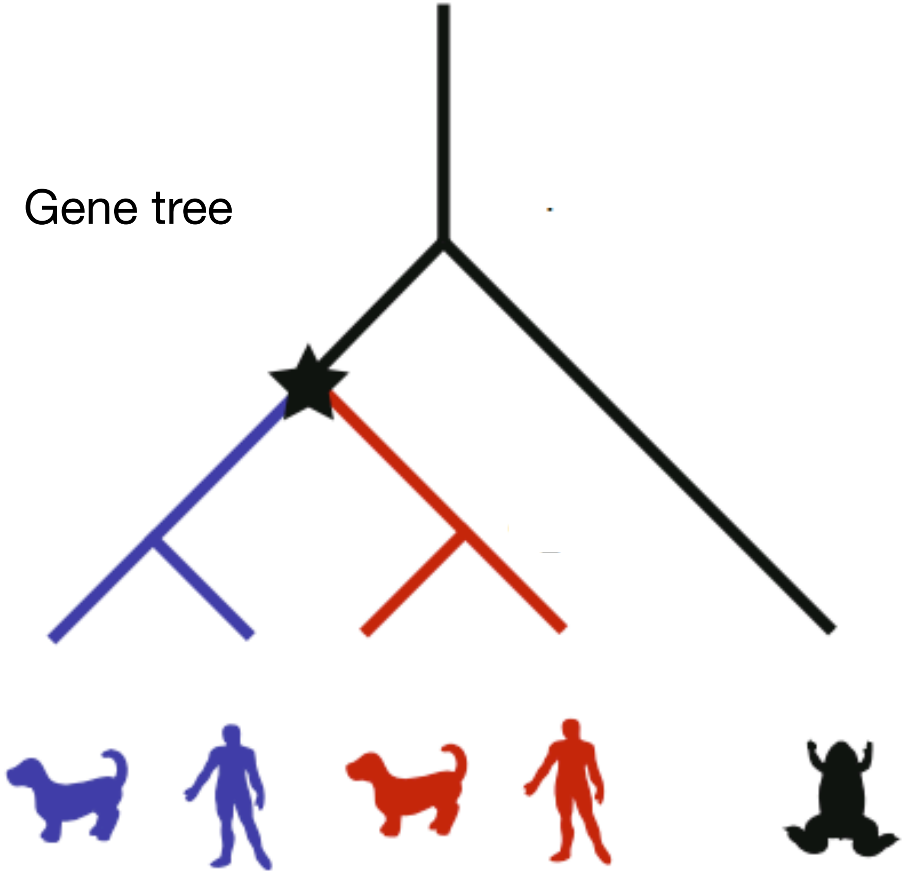
orthology relations are inherently hierarchical

Hierarchical Orthologous Groups

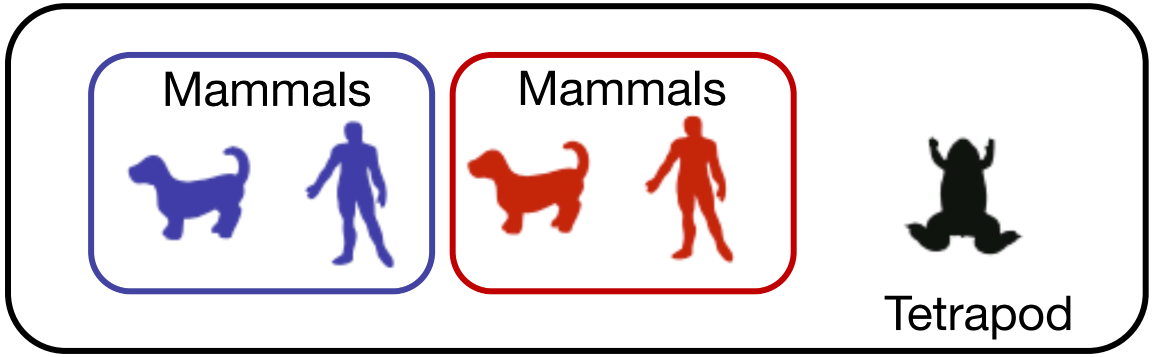


HOGs are defined at different taxonomic levels

HOGs



Hierarchical Orthologous Group(s)



With HOGs, the speciation and duplication information is encoded implicitly

Hierarchical orthologous groups

- Sets of genes that have descended from a common ancestral gene in a given ancestral species

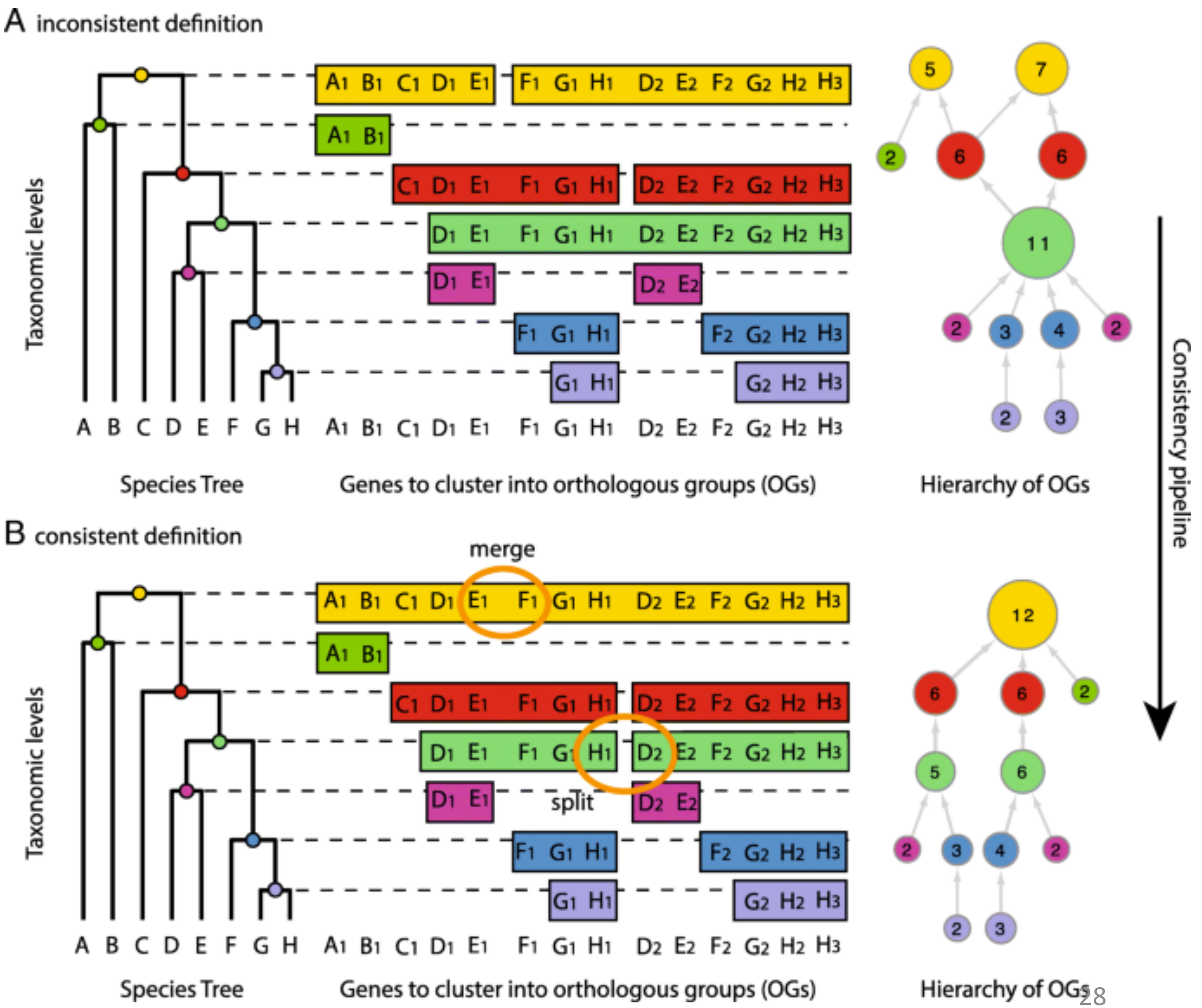
- Defined with respect to specific clades (taxonomic levels)

- Hierarchical in that groups are defined with respect to deeper clades that encompass multiple groups defined on their descendants
    - Basically nested subfamilies

HOGs are hierarchically consistent across taxonomic levels

Example from eggNOG hierarchical clustering:

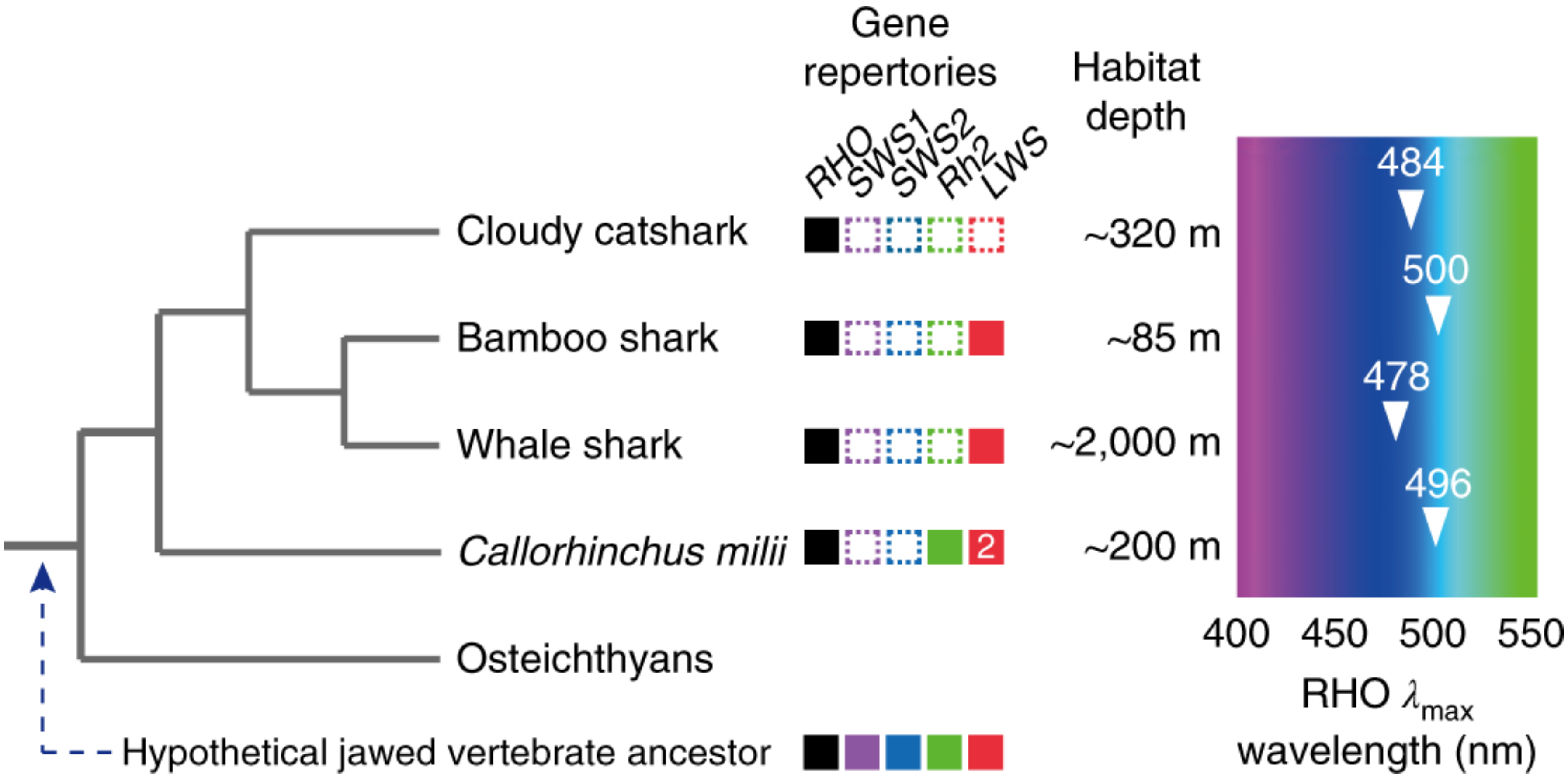
“The example shows how genes are clustered into OGs based on the chosen taxonomic level (dotted line) and how the independently computed levels can be joined into a hierarchy of OGs (right side)”



Heller et al. 2019 https://doi.org/10.1186/s12859-019-2828-z

HOGs help with interpretation of gene families

- Direct relationship between phylogenetic gene trees and HOGs
- Evolutionary history of shark visual opsin gene loss and duplication
- “The absence of LWS might be due to an evolutionary gene loss that was permitted in the catshark ancestor by its possible exclusive deep-sea habitat”

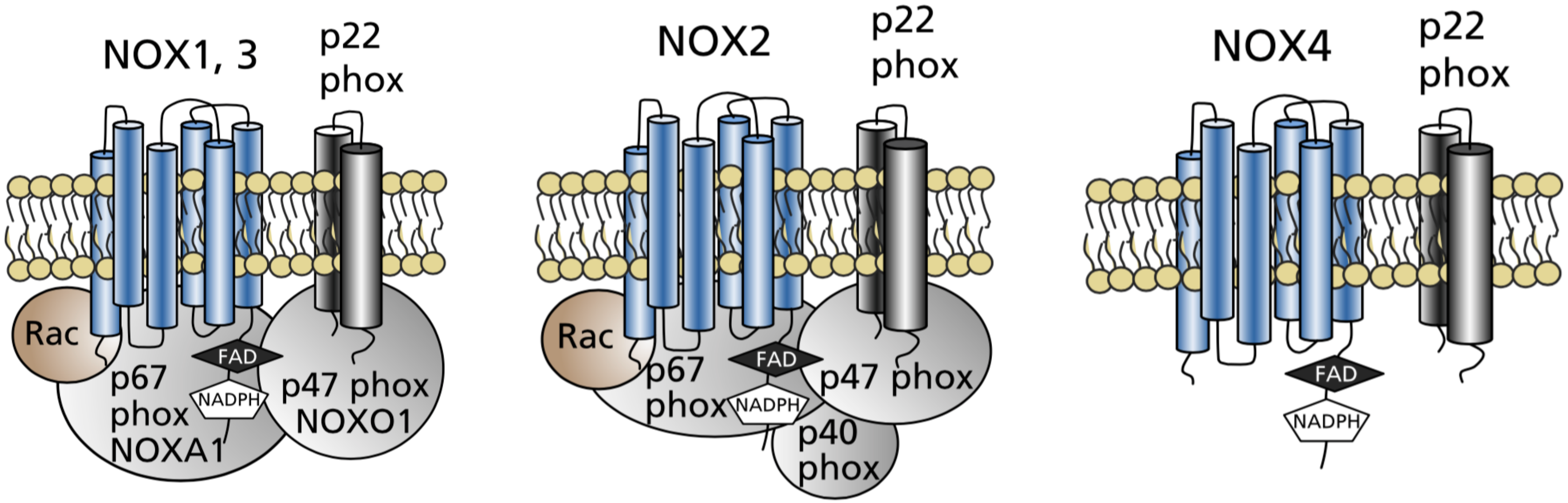


Hara et al. 2018. https://doi.org/10.1038/s41559-018-0673-5

Hierarchical orthologous groups

- HOGs allow for a more fine-grained level of analysis, and this can affect the biological relevance/interpretation
- For example, genes which duplicated at a certain taxonomic level may have subfunctionalized. This information may not be revealed if looking at an orthologous group at an older taxonomic level

Example on the NADPH oxidase family



Related disorders

hypertension, aortic dissection (aneurysm), neointima formation

inflammatory pain, cerebral ischemia, neuroinflammation

hyperoxia-induced acute lung injury

colorectal cancer

development of the otoconia

hearing loss

insulin resistance (diabetes)

cardiac hypertrophy, fibrosis, heart failure

myocardial infarction, neovascularization (ischemic cardiovascular diseases)

Alzheimer's disease, Parkinson's disease, ischemic stroke

neuropathic pain (peripheral nerve injury)

glutamate release (schizophrenia)

liver fibrosis, Liver ischemia and reperfusion injury

amyotrophic lateral sclerosis (ALS)

mitochondrial dysfunction (cardiac hypertrophy, interstitial fibrosis)

sympathetic nerve activity (heart failure, myocardial infarction)

pulmonary fibrosis, pulmonary hypertension

diabetic nephropathy, renal cancer

Looking at the genes at a deeper taxonomic level would merge all these functions

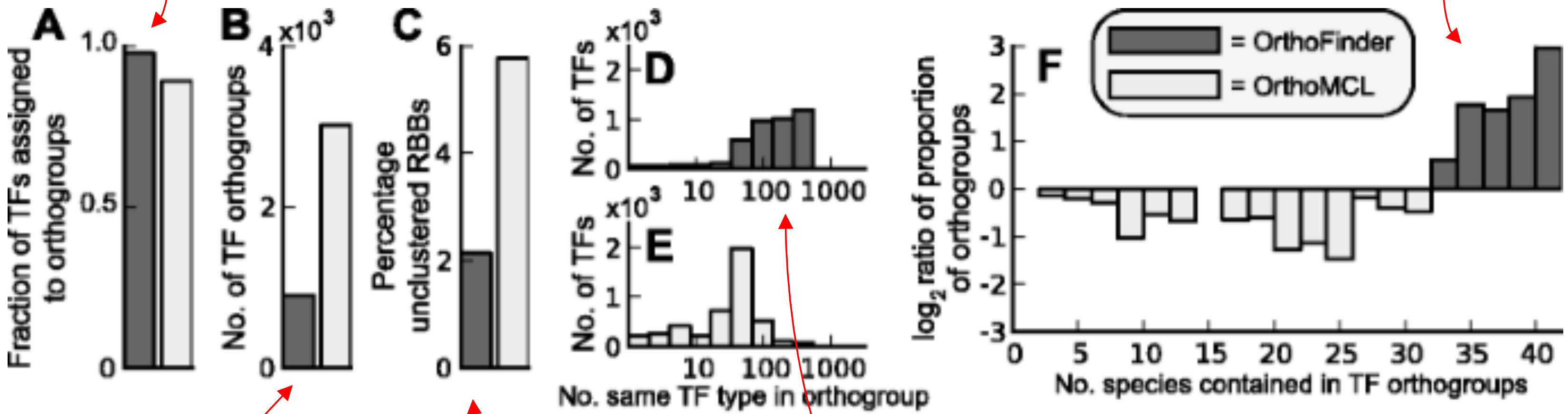
Katsuyama et al. J. Clin. Biochem. Nutr. 2012

The problem with non-hierarchical orthologous groups

- Too inclusive or not inclusive enough orthologous groups
- Clustering based on percent identity, OrthoMCL inflation parameter

OrthoFinder vs. OrthoMCL

8.5 % more transcription factors placed in OGs



Encompass a larger number of species (can find orthologs over greater phylogenetic distances)

Less fragmented OGs

Missing fewer reciprocal best hits

Clustered more of the same type of transcription factor together

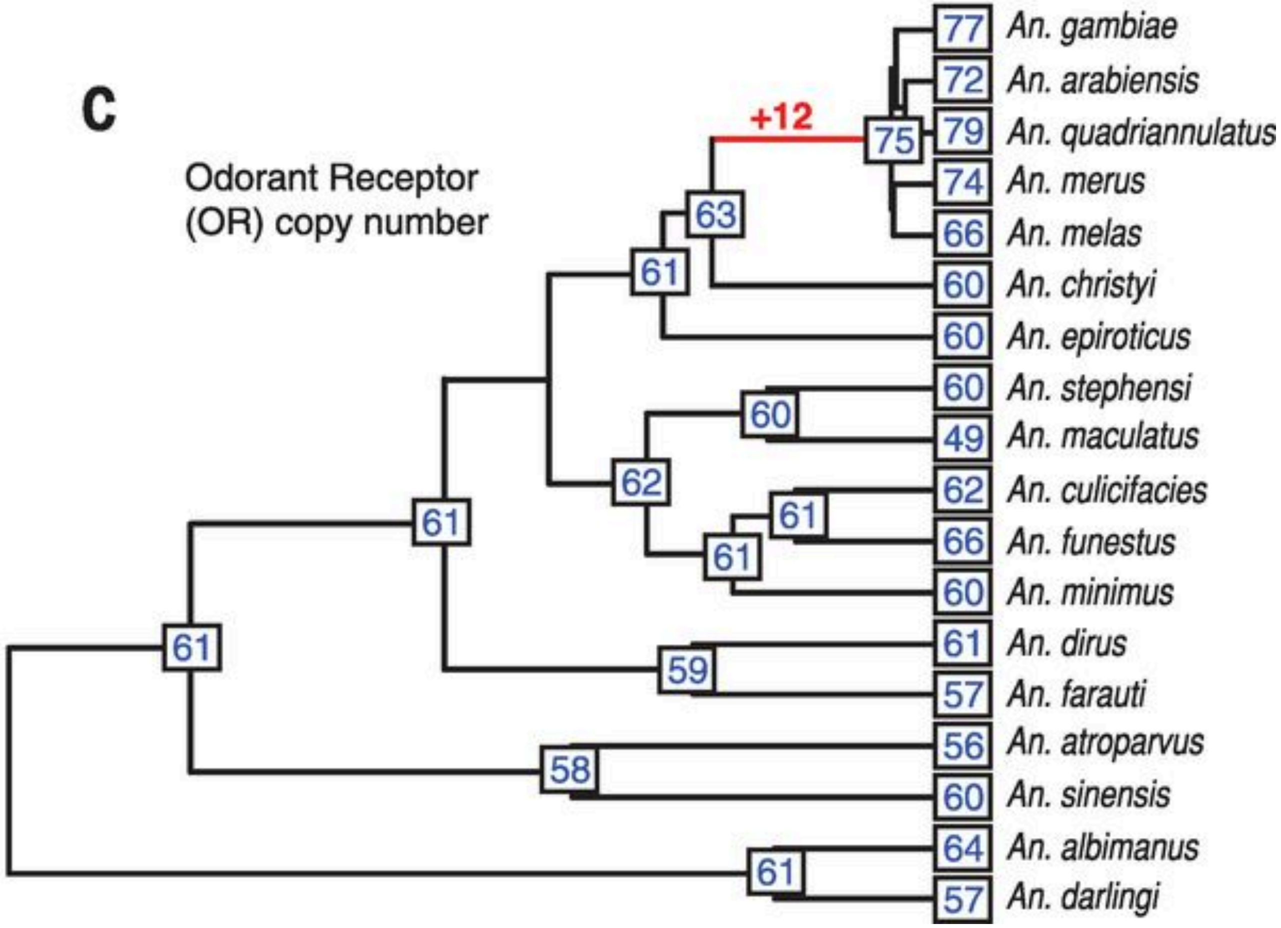
Emms et al. 2015 https://doi.org/10.1186/s13059-015-0721-2

The problem with non-hierarchical orthologous groups

- Orthologous groups (clusters) are static
- Only gives 1 ancestral level: that which relates all the species used in the analysis
- Cannot study evolutionary history of genes over time
- Need different levels of resolution for functional and evolutionary analysis

HOGs can be used to trace gene families

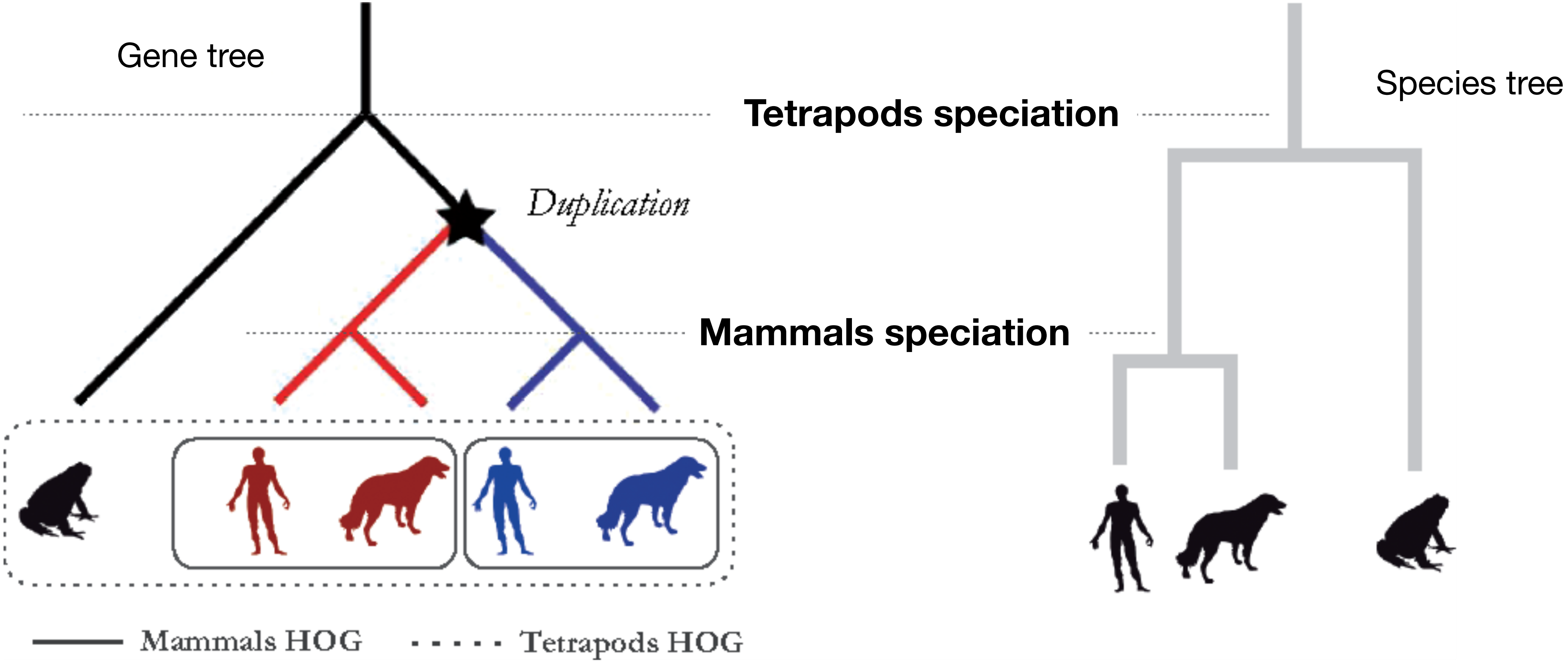
- Can study the evolutionary history of gene families
- Neafsey et al. used OrthoDB to delineate HOGs at each last common ancestor of the species phylogeny in 43 insects
- Detected where odorant receptors were gained and lost along the phylogenetic tree



Neafsey et al 2015 DOI: 10.1126/science.1258522

Roadblock 3: studying ancestral genomes and evolutionary histories

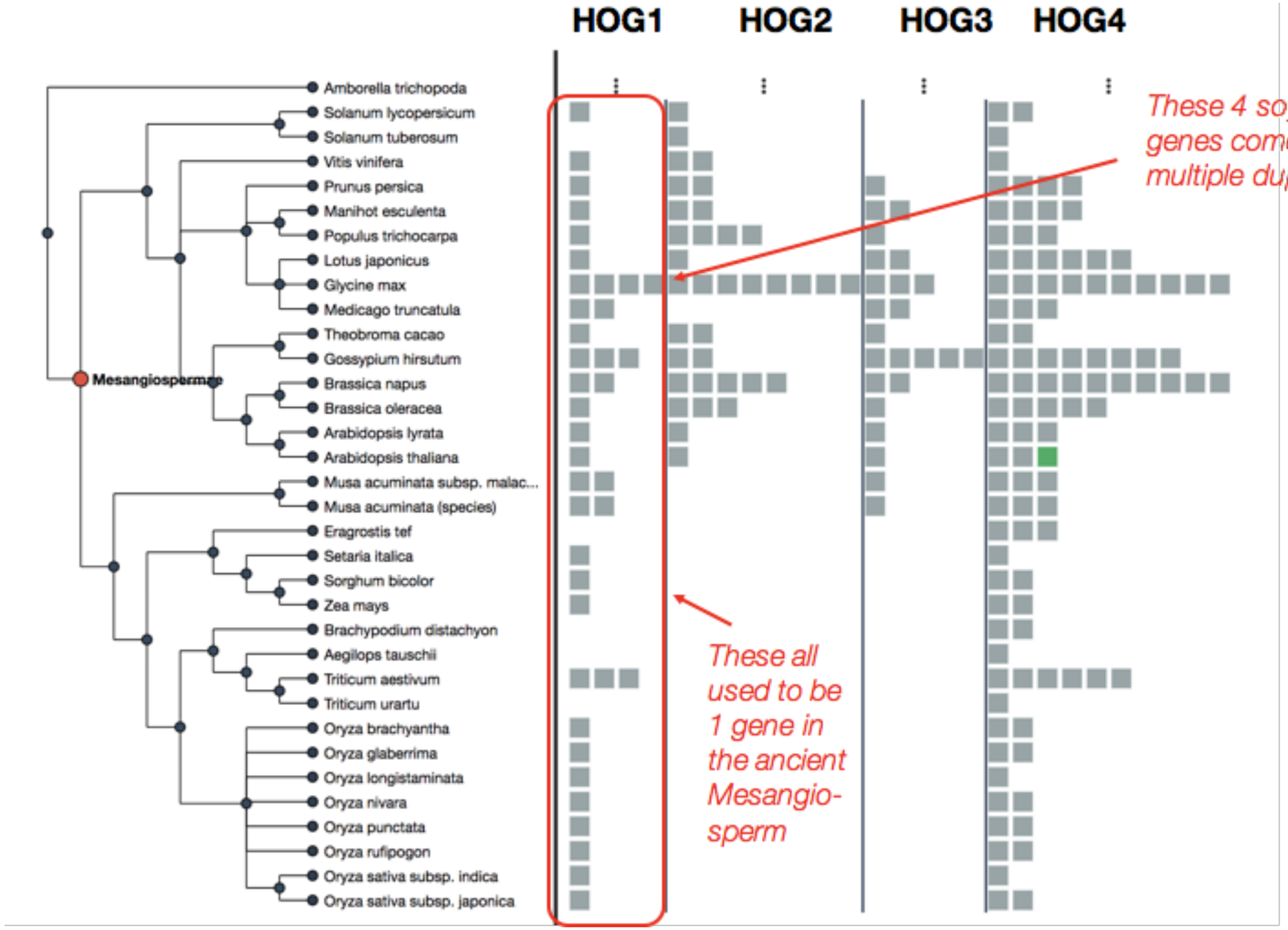
Hierarchical Orthologous Groups



Each HOG is an ancestral gene at a given taxonomic level

HOGs are ancestral genes

- HOGs by definition all descended from a common ancestral gene
- Thus, at each taxonomic level, the ancestral genome is comprised of all the HOGs *at that level*



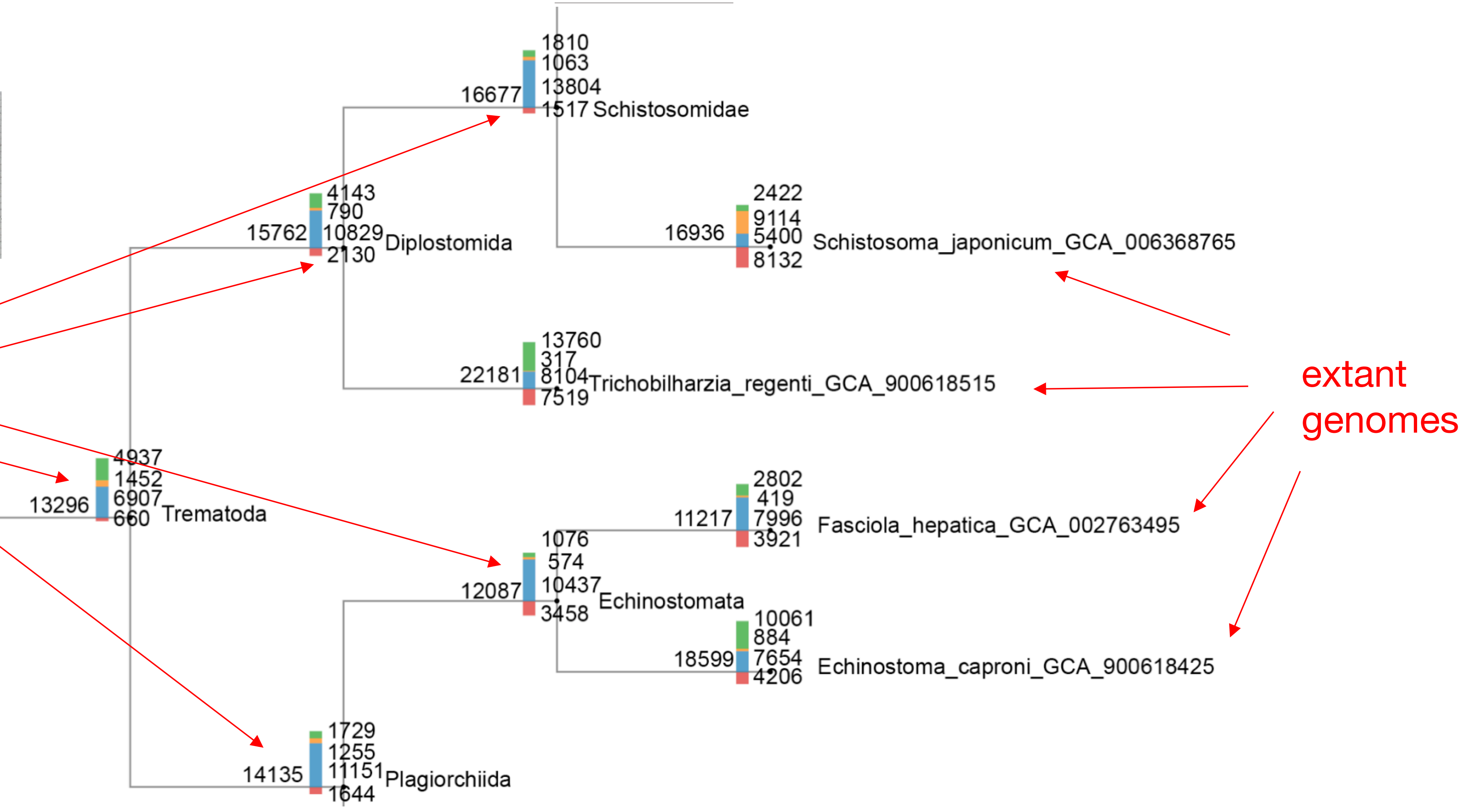
These 4 soybean genes come from multiple duplications

These all used to be 1 gene in the ancient Mesangio-sperm

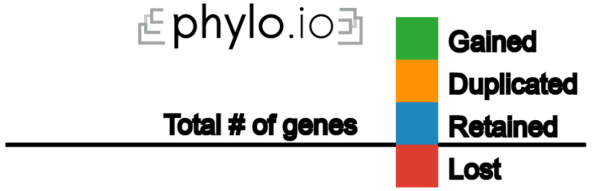
a gene family

HOGs can be used to trace evolution of genomes

ancestral genomes



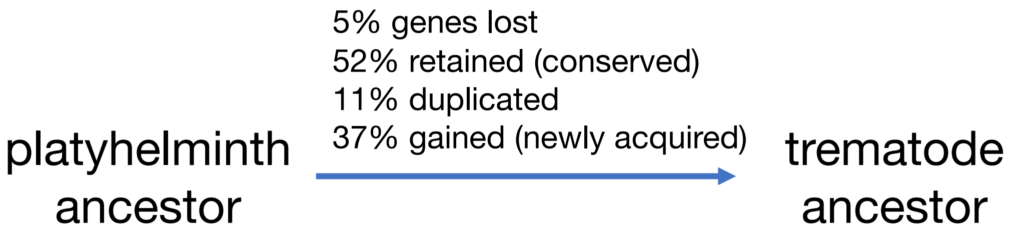
extant genomes

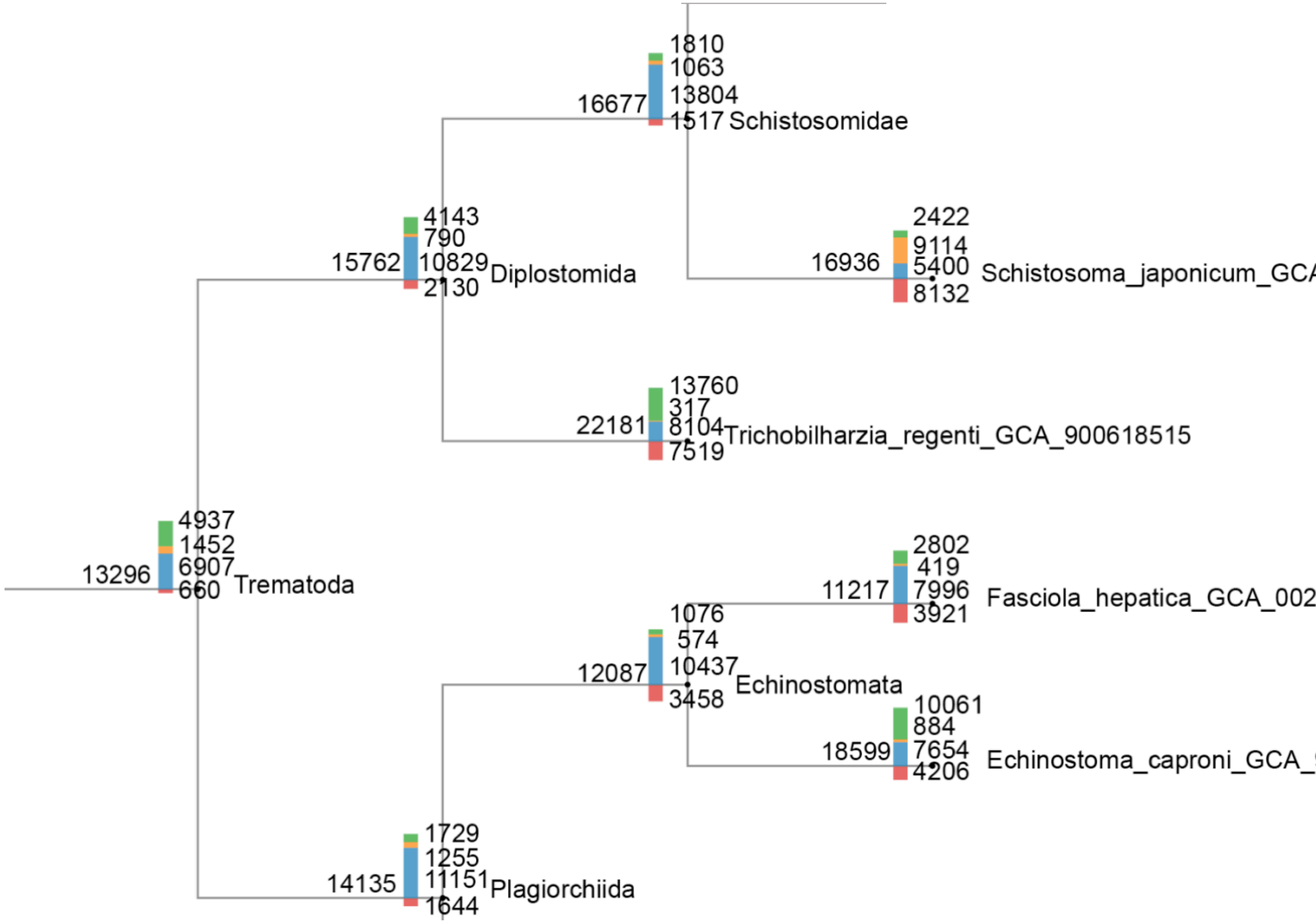


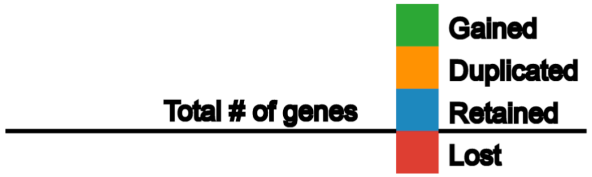
Zajac et al. 2020, GBE, under revision

HOGs can be used to trace evolution of genomes

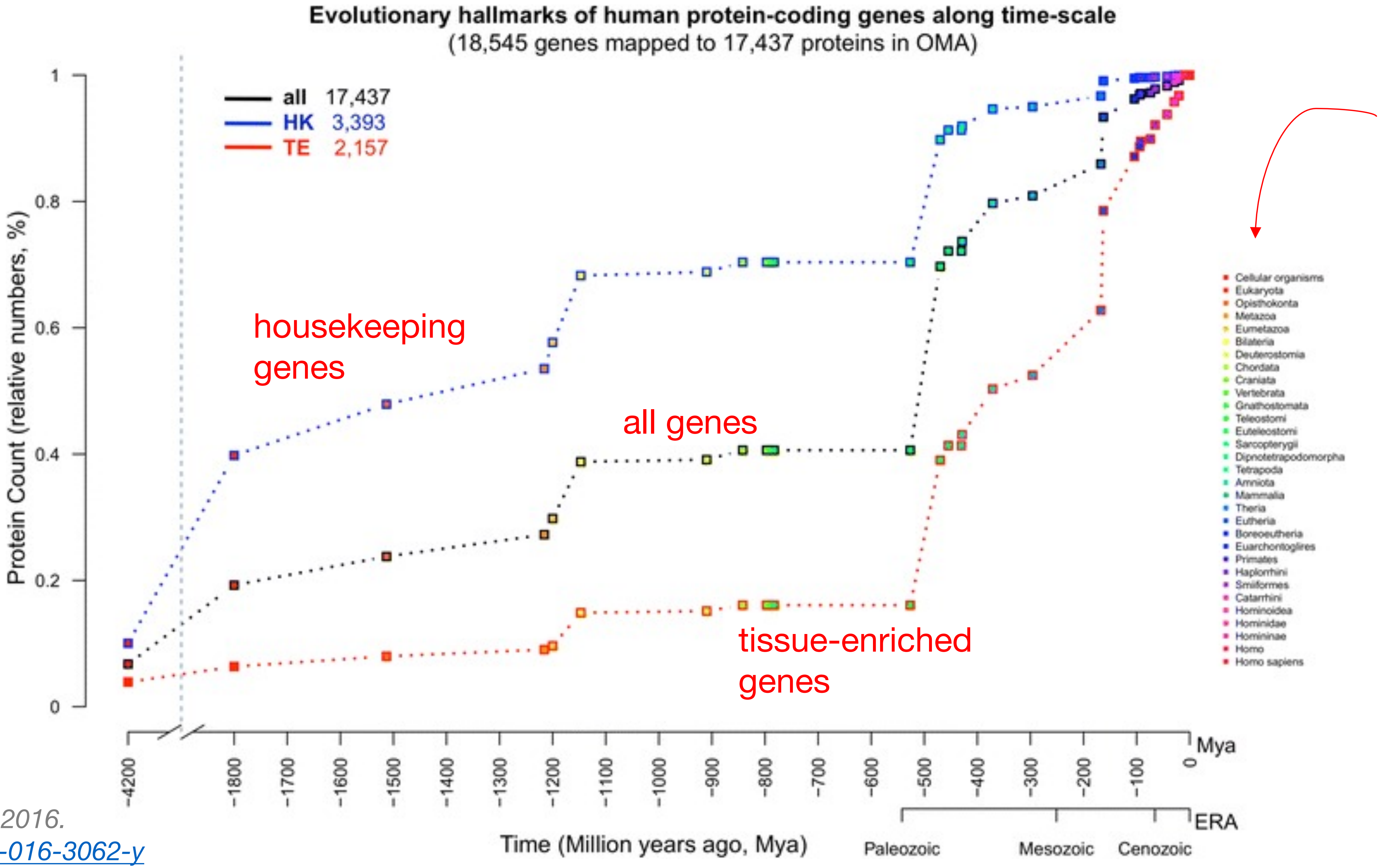
13296 HOGs at the trematoda level = the ancestral trematode genome had 13,296 genes







HOGs can be used to trace evolution of genomes



Evolutionary hallmarks of human protein-coding genes along time-scale
(18,545 genes mapped to 17,437 proteins in OMA)

housekeeping genes

all genes

tissue-enriched genes

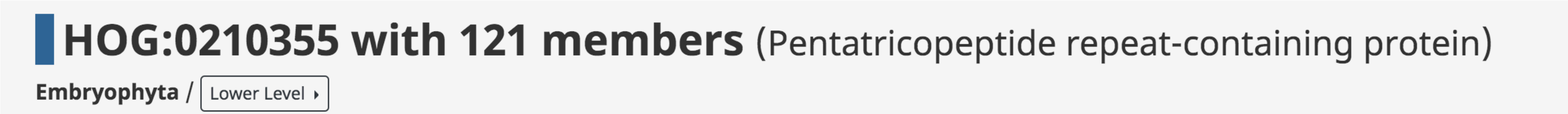
different taxonomic levels

de Pavia Lopes et al 2016.
doi: 10.1186/s12864-016-3062-y

Future directions

HOG visualization

- Live demo
- Example:



HOG:0210355 with 121 members (Pentatricopeptide repeat-containing protein)

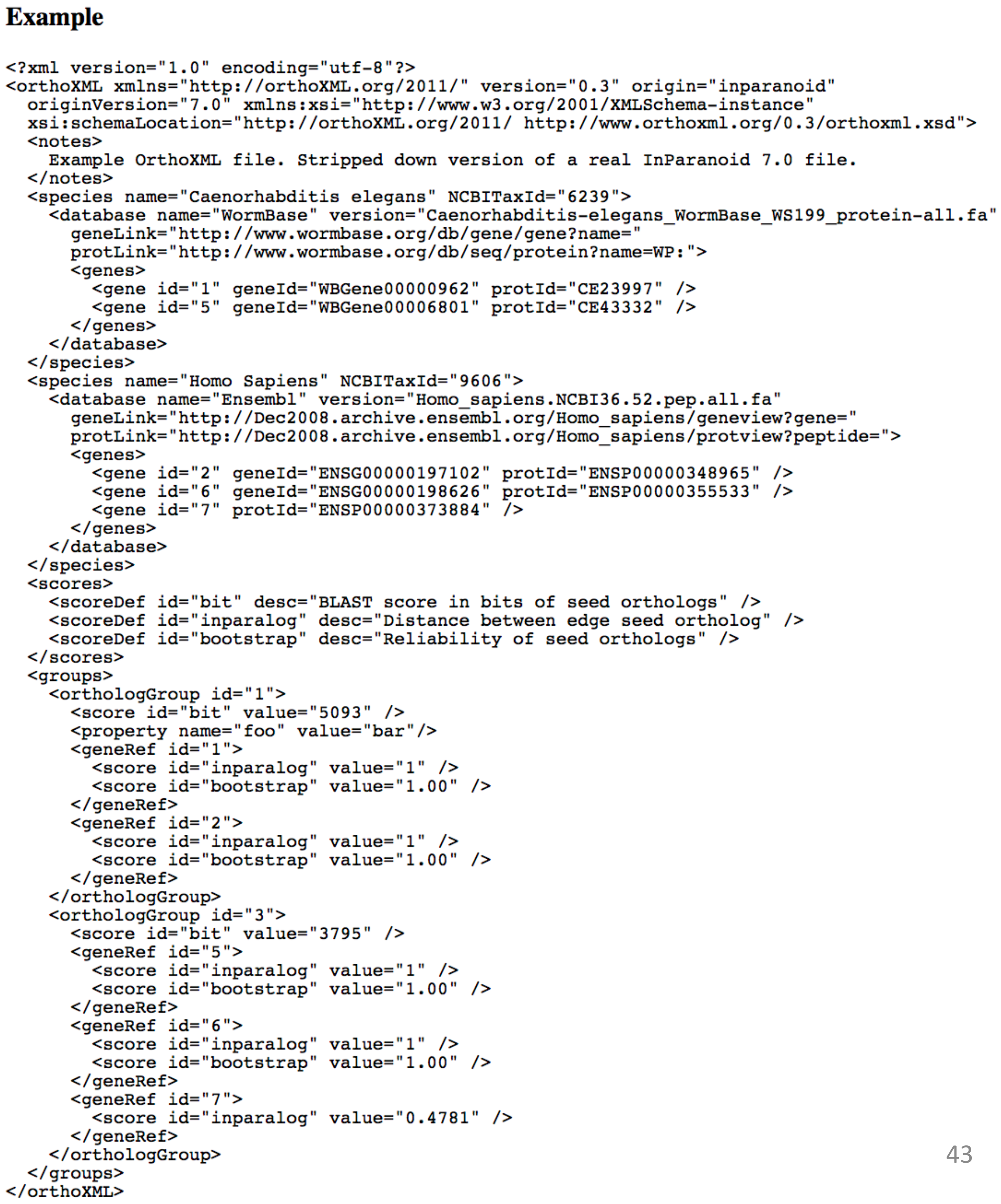
Embryophyta /

- https://omabrowser.org/oma/hog/HOG:0210355/iham/

Train et al 2019. https://doi.org/10.1093/bioinformatics/bty994
Konczal et al 2020. https://doi.org/10.1111/mec.15421

OrthoXML

- An XML schema designed to describe orthology relations
- Can store orthology data from different sources in a uniform manner
- Useful when working with HOG data



Example

```
<?xml version="1.0" encoding="utf-8"?>
<orthoXML xmlns="http://orthoXML.org/2011/" version="0.3" origin="inparanoid"
  originVersion="7.0" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://orthoXML.org/2011/ http://www.orthoxml.org/0.3/orthoxml.xsd">
  <notes>
    Example OrthoXML file. Stripped down version of a real InParanoid 7.0 file.
  </notes>
  <species name="Caenorhabditis elegans" NCBITaxId="6239">
    <database name="WormBase" version="Caenorhabditis-elegans_WormBase_WS199_protein-all.fa"
      geneLink="http://www.wormbase.org/db/gene/gene?name="
      protLink="http://www.wormbase.org/db/seq/protein?name=WP:">
      <genes>
        <gene id="1" geneId="WBGene00000962" protId="CE23997" />
        <gene id="5" geneId="WBGene00006801" protId="CE43332" />
      </genes>
    </database>
  </species>
  <species name="Homo Sapiens" NCBITaxId="9606">
    <database name="Ensembl" version="Homo_sapiens.NCBI36.52.pep.all.fa"
      geneLink="http://Dec2008.archive.ensembl.org/Homo_sapiens/geneview?gene="
      protLink="http://Dec2008.archive.ensembl.org/Homo_sapiens/protview?peptide=">
      <genes>
        <gene id="2" geneId="ENSG00000197102" protId="ENSP00000348965" />
        <gene id="6" geneId="ENSG00000198626" protId="ENSP00000355533" />
        <gene id="7" protId="ENSP00000373884" />
      </genes>
    </database>
  </species>
  <scores>
    <scoreDef id="bit" desc="BLAST score in bits of seed orthologs" />
    <scoreDef id="inparalog" desc="Distance between edge seed ortholog" />
    <scoreDef id="bootstrap" desc="Reliability of seed orthologs" />
  </scores>
  <groups>
    <orthologGroup id="1">
      <score id="bit" value="5093" />
      <property name="foo" value="bar"/>
      <geneRef id="1">
        <score id="inparalog" value="1" />
        <score id="bootstrap" value="1.00" />
      </geneRef>
      <geneRef id="2">
        <score id="inparalog" value="1" />
        <score id="bootstrap" value="1.00" />
      </geneRef>
    </orthologGroup>
    <orthologGroup id="3">
      <score id="bit" value="3795" />
      <geneRef id="5">
        <score id="inparalog" value="1" />
        <score id="bootstrap" value="1.00" />
      </geneRef>
      <geneRef id="6">
        <score id="inparalog" value="1" />
        <score id="bootstrap" value="1.00" />
      </geneRef>
      <geneRef id="7">
        <score id="inparalog" value="0.4781" />
      </geneRef>
    </orthologGroup>
  </groups>
</orthoXML>
```

Future directions

- Ancestral genome synteny (by ordering the HOGs)
- Improvement of HOG construction algorithms (faster, more accurate)
- Tools to covert HOGs to trees and vice versa
- Allowing for Horizontal Gene Transfer

The Rise of the HOGs

- Increasing number of resources provide HOGs, usually a topic at the Quest for Orthologs.
  - eggNOG (Heller et al. 2019)
  - OMA (Altenhoff et al. 2013)
  - OrthoDB (Waterhouse et al. 2013)
  - Hieranoid (Schreiber and Sonnhammer 2013; Kaduk and Sonnhammer 2017)
  - LOFT (van der Heijden et al. 2007)
  - OrthoFinder (Emms and Kelly 2019)

Thanks for listening!
natasha.glover@unil.ch