

# ***Workshop on Genomics***

## ***Český Krumlov, May 22 2023***

**Comparative Genomics - Morning Session: 09:00-12:00**

**Robert M. Waterhouse**

*Department of Ecology & Evolution, University of Lausanne, Swiss Institute of Bioinformatics, Switzerland*



Swiss Institute of  
Bioinformatics



FONDS NATIONAL SUISSE  
SCHWEIZERISCHER NATIONALFONDS  
FONDO NAZIONALE SVIZZERO  
SWISS NATIONAL SCIENCE FOUNDATION

✉ [robert.waterhouse@gmail.com](mailto:robert.waterhouse@gmail.com)

🐦 [@rmwaterhouse](https://twitter.com/rmwaterhouse)

🌐 [www.rmwaterhouse.org](http://www.rmwaterhouse.org)

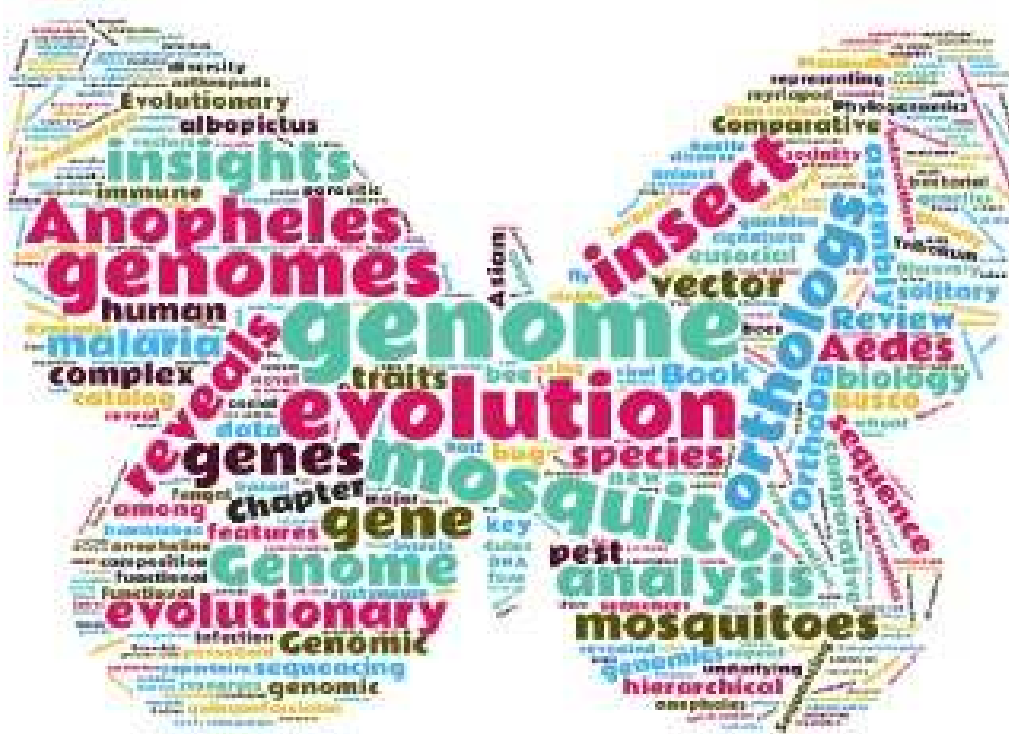


# ***Instructor Biography-Introduction***

- 2017- SNF Assistant Professor  
University of Lausanne
- 2015-16 Marie Curie Fellow & Maître assistant  
University of Geneva *ZDOBN OV*
- 2013-14 Marie Curie Outgoing Fellow  
Massachusetts Institute of Technology *KELLIS*
- 2009-12 Postdoctoral Researcher  
University of Geneva *ZDOBN OV*
- 2005-09 Wellcome Trust PhD  
Imperial College London *CHRISTOPHIDES*
- 2004-05 Wellcome Trust MSc Bioinformatics  
Imperial College London
- 2000-04 MBioch Biochemistry  
University of Oxford



# ***Instructor Biography-Introduction***



# OrthoDB BUSCO

Arthropoda  
Assembly  
Assessment  
Catalogue



# Arthropod Evolutionary Genomics

## Insect Immunity Gene Evolution

# Orthology Delineation Quality Assessments

# ***i5k Arthropod Genomics Community***



Coordinating the sequencing and analysis of  
5'000 insects and other arthropods

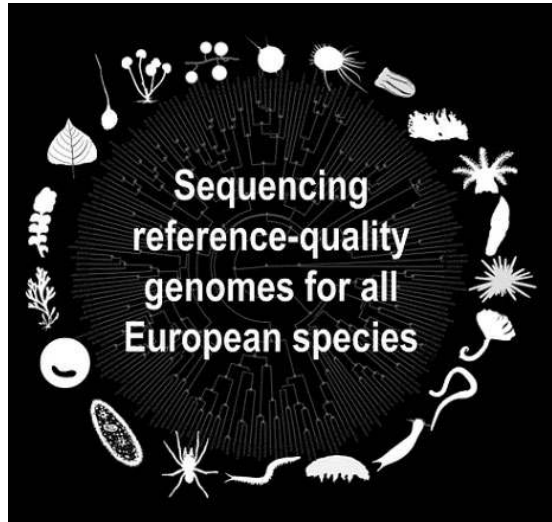
If you use genomics to study arthropods,  
you are an i5k member!

i5k pilot project: 28 species  
*Gene content evolution in the arthropods*  
Genome Biology 2020

A diverse international group of researchers  
Join the community on the i5k Slack Workspace: [bit.ly/artgen20](https://bit.ly/artgen20)



# ***The European Reference Genome Atlas ERGA***



Coordinating the sequencing and analysis of European eukaryote species

If you are building genome resources for European eukaryotes, you should join!

ERGA Pilot Project: 98 species  
*A distributed model of genome generation across 34 countries in Europe*

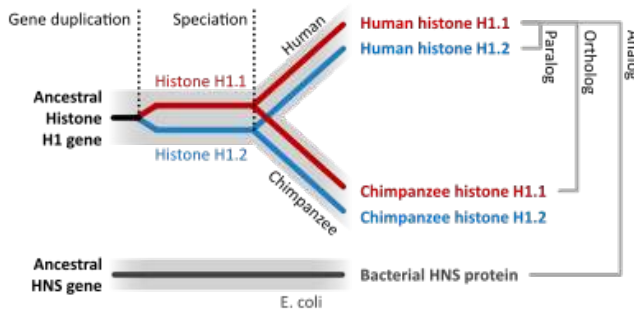
A diverse group of researchers in Europe  
Join the community: [www.erga-biodiversity.eu](http://www.erga-biodiversity.eu)



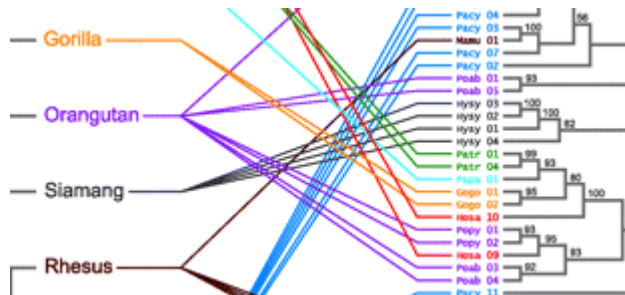
# Comparative Genomics Hands-On: Concepts and Applications



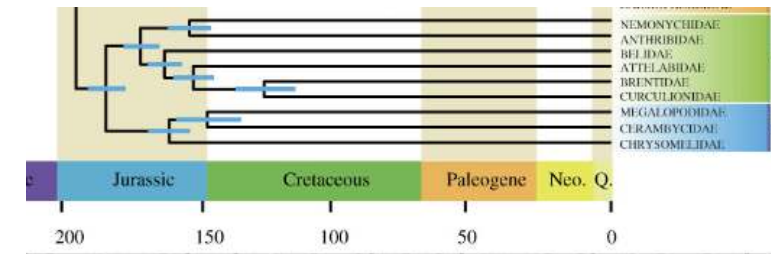
## Orthology delineation



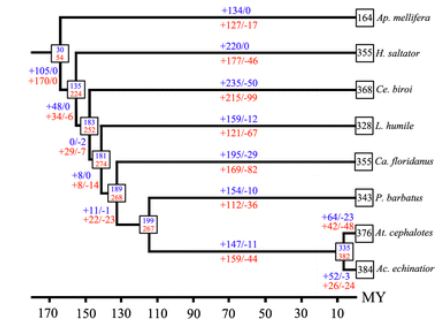
## Gene-tree-species-tree reconciliation



## Species tree estimation

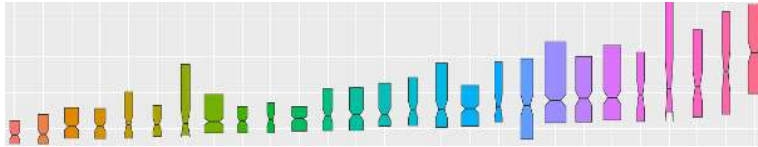


## Gene ancestral state reconstruction

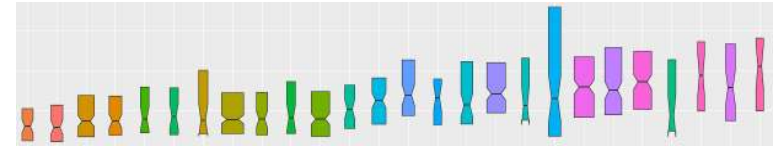


# CompGeno: gene family evolutionary dynamics

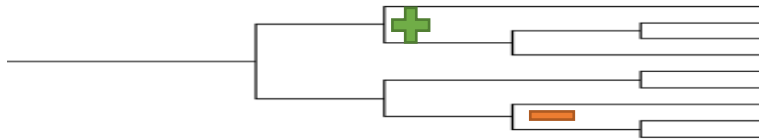
## Protein Sequence Divergence



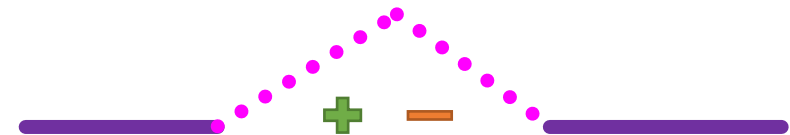
## DNA Selection/Constraint



## Gene Gain/Loss Rates



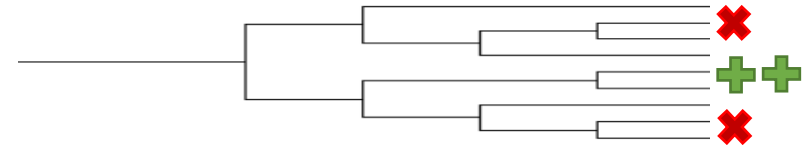
## Intron Gain/Loss Rates



## Stop-Codon Readthrough

|                      |                                                                                                                               |
|----------------------|-------------------------------------------------------------------------------------------------------------------------------|
| <i>B. hemeris_ga</i> | V L K Q P F L S H V * R H Q B A G G D M N T A G C D Q Q Q Q                                                                   |
| <i>B. hemeris</i>    | GTC CTC AAA CAA CCF CCF CTG TCG CAG GGA TGA CGT CAC CAG TCG GCG GGG GGC GAC ATG AAC ACC GCG GGC TGC GAC CAG CAA CAA CAA (     |
| <i>B. ignitus</i>    | GTC CTC AAA CAA CCF CCF CTG TCG CAG GGA TGA CGT CAC CAG TCG GCG GGG GGC GAC ATG AAC ACC GCG GGC TGC GAC CAG CAA CAA CAA (     |
| <i>B. polaris</i>    | GTC CTC AAA CAA CCF CCF CTG TCG CAG GGA TGA CGT CAC CAG TCG GCG GGG GGC GAC ATG AAC ACC GCG GGC TGC GAC CAG CAG CAA CAA CAA ( |
| <i>B. pipipes</i>    | GTC CTC AAA CAA CCF CCF CTG TCG CAG GGA TGA CGT CAC CAG TCG GCG GGG GGC GAC ATG AAC ACC GCG GGC TGC GAC CAG CAG CAA CAA CAA ( |
| <i>B. pyrosoma</i>   | GTC CTC AAA CAA CCF CCF CTG TCG CAG GGA TGA CGT CAC CAG TCG GCG GGG GGC GAC ATG AAC ACC GCG GGC TGC GAC CAG CAG CAA CAA CAA ( |
| <i>B. brevipes</i>   | GTC CTC AAA CAA CCF CCF CTG TCG CAG GGA TGA CGT CAC CAG TCG GCG GGG GGC GAC ATG AAC ACC GCG GGC TGC GAC CAG CAG CAA CAA CAA ( |

## Copy-Number/Universality



Comprehensive quantifications using multiple complementary approaches distinguish conserved/stable from divergent/dynamic gene families



# ***Orthology Delineation***

***What is orthology?***

***How do we delineate orthologs?***

***And why do we need to?  
(species/gene trees)***



# *Orthology – what is it?*

*Homology*



*Orthology*



# *Orthology – what is it?*

## *Homology*

“designates a relationship of **common descent** between any entities, without further specification of the evolutionary scenario”

Orthologs, Paralog, and  
Evolutionary Genomics<sup>1</sup>

Eugene V. Koonin

Annu. Rev. Genet.  
2005. 39:309–38



# ***Orthology – what is it?***

“genes originating from a single ancestral gene in the last common ancestor of the compared genomes”

***Orthology***

Orthologs, Paralog, and  
Evolutionary Genomics<sup>1</sup>

Eugene V. Koonin

Annu. Rev. Genet.  
2005. 39:309–38



# ***Orthology – what is it?***

“paralogs are  
genes related via duplication”

***Paralogy***

Orthologs, Paralogs, and  
Evolutionary Genomics<sup>1</sup>

Eugene V. Koonin

Annu. Rev. Genet.  
2005. 39:309–38



# ***Orthology – what is it?***

## ***Homologs***

Common Ancestor



## ***Orthologs***

Speciation  
Event

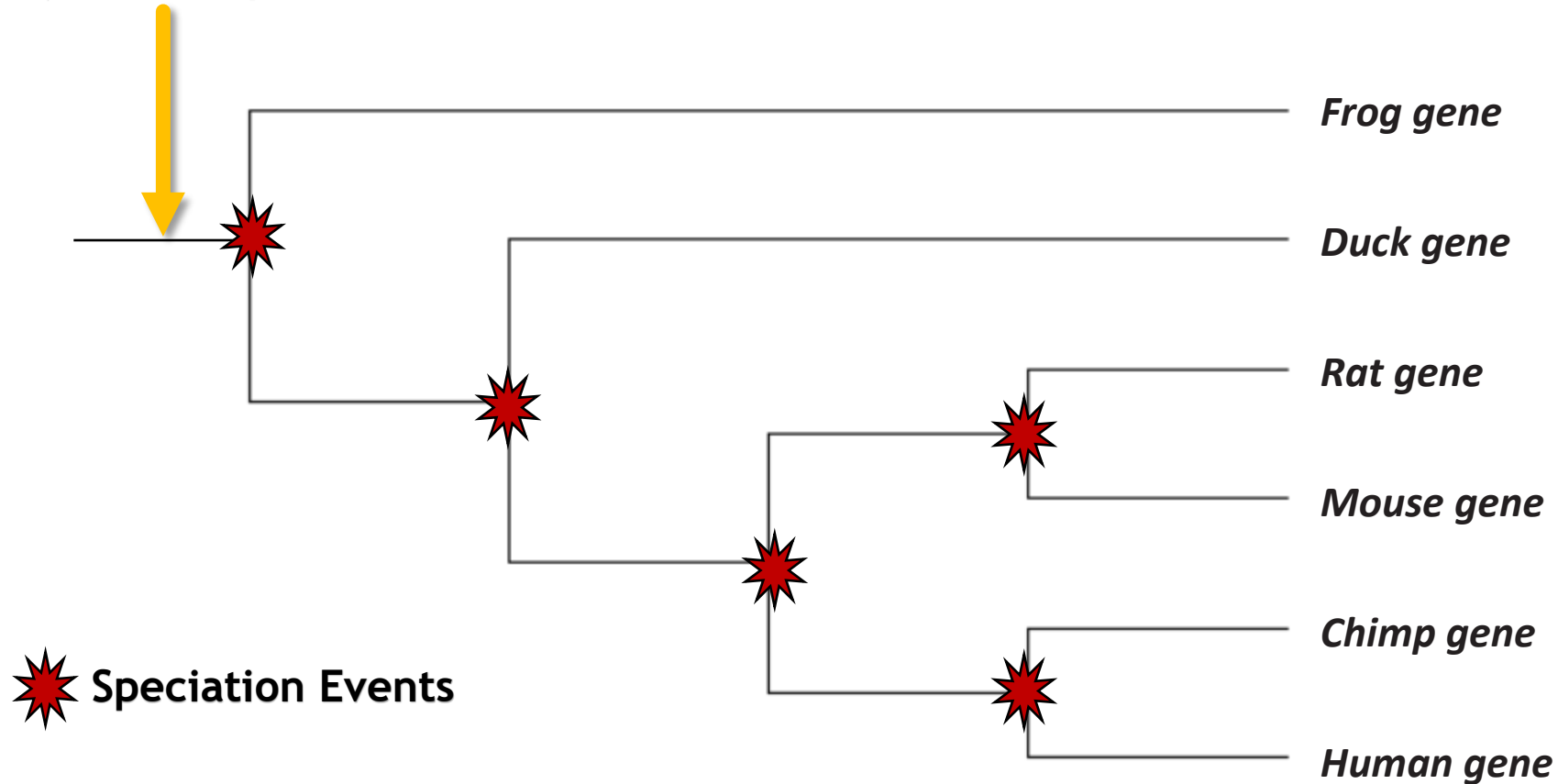
## ***Paralogs***

Duplication  
Event



# Orthology – simple scenario

Last Common Ancestor  
(LCA) of all 6 species

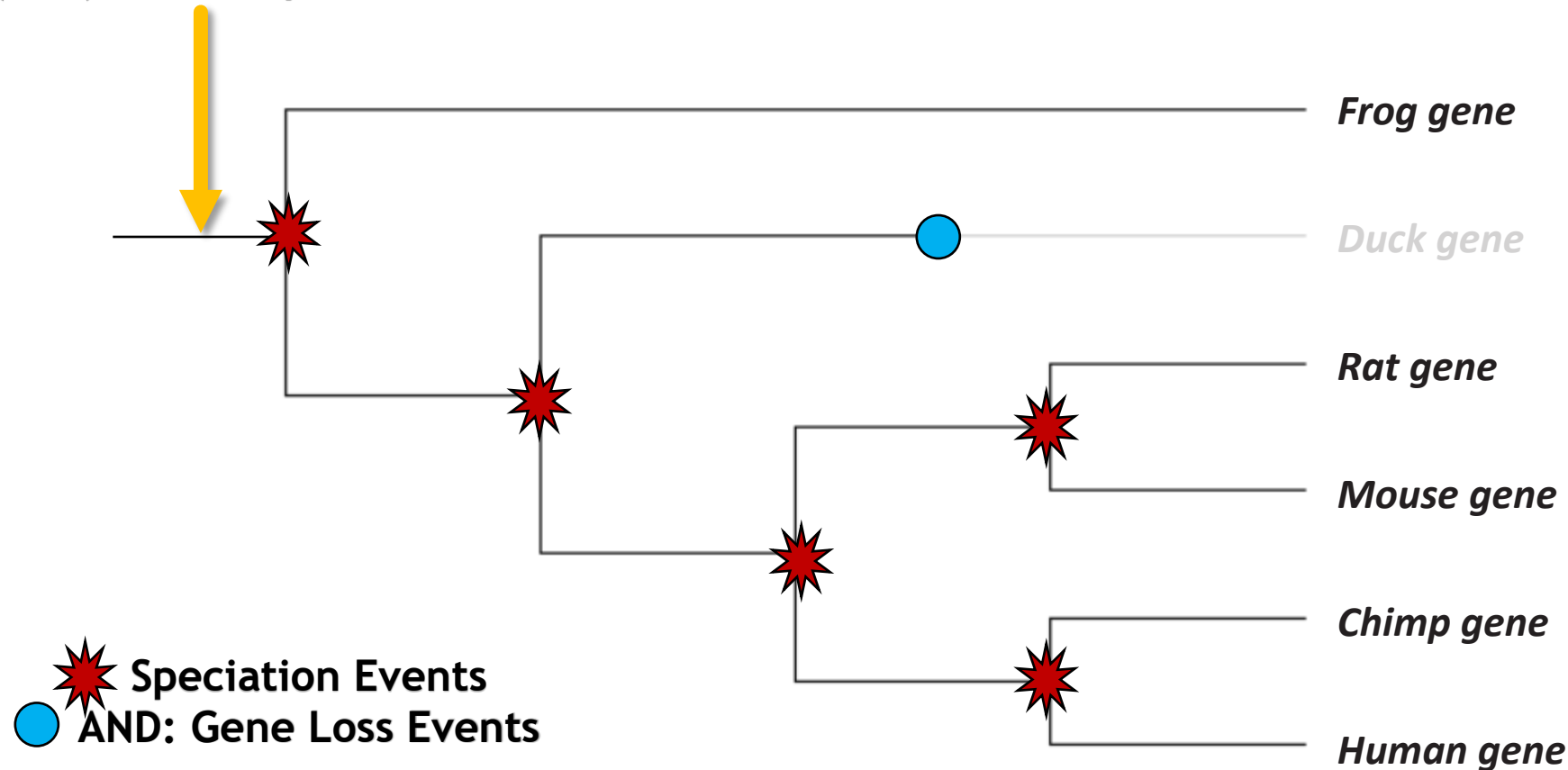


## Single-Copy Orthologs



# Evolution ≠ simple

Last Common Ancestor  
(LCA) of all 6 species

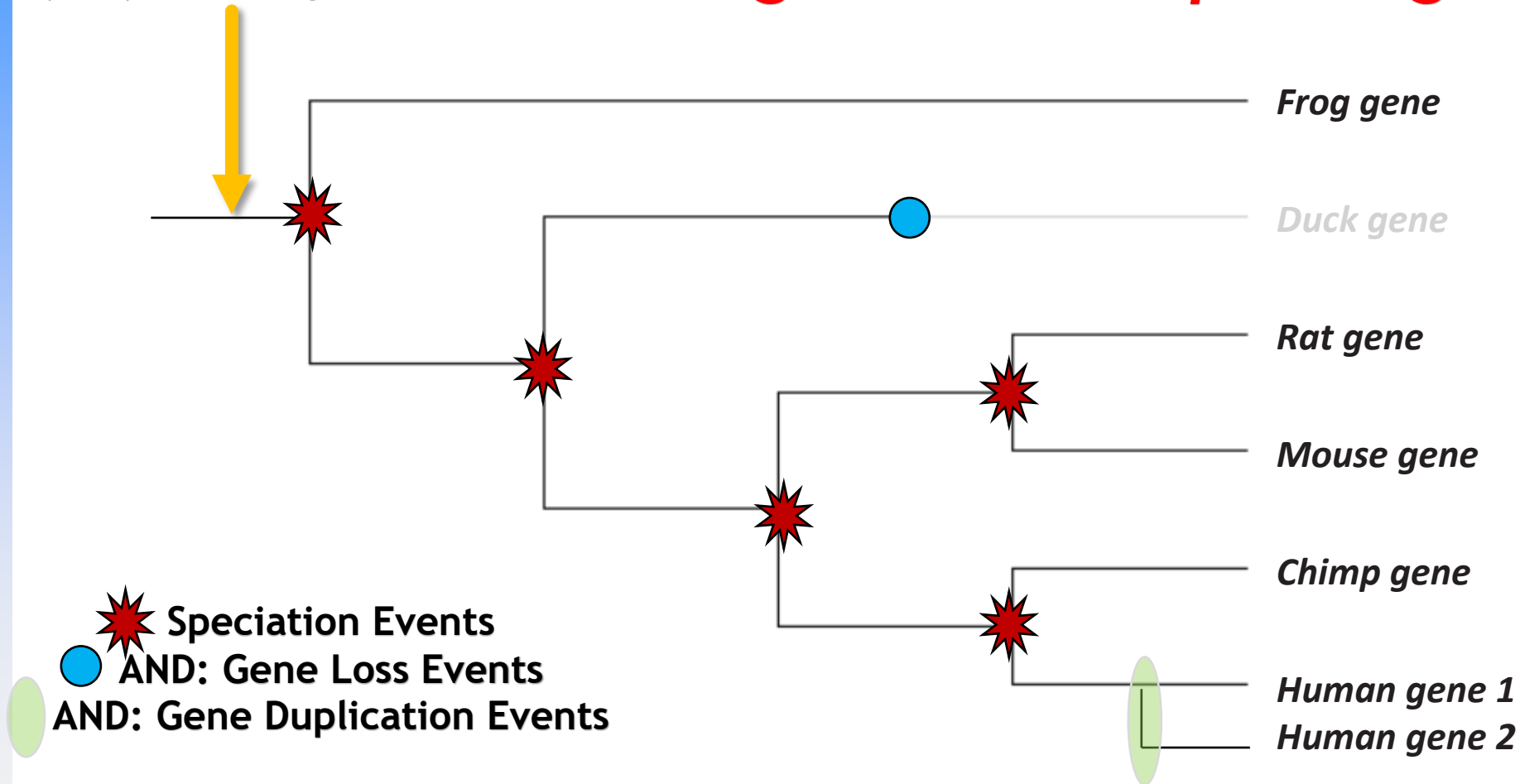


## Single-Copy Orthologs with Losses

# Evolution ≠ simple

Last Common Ancestor  
(LCA) of all 6 species

**Human gene 1 & 2 = paralogs**



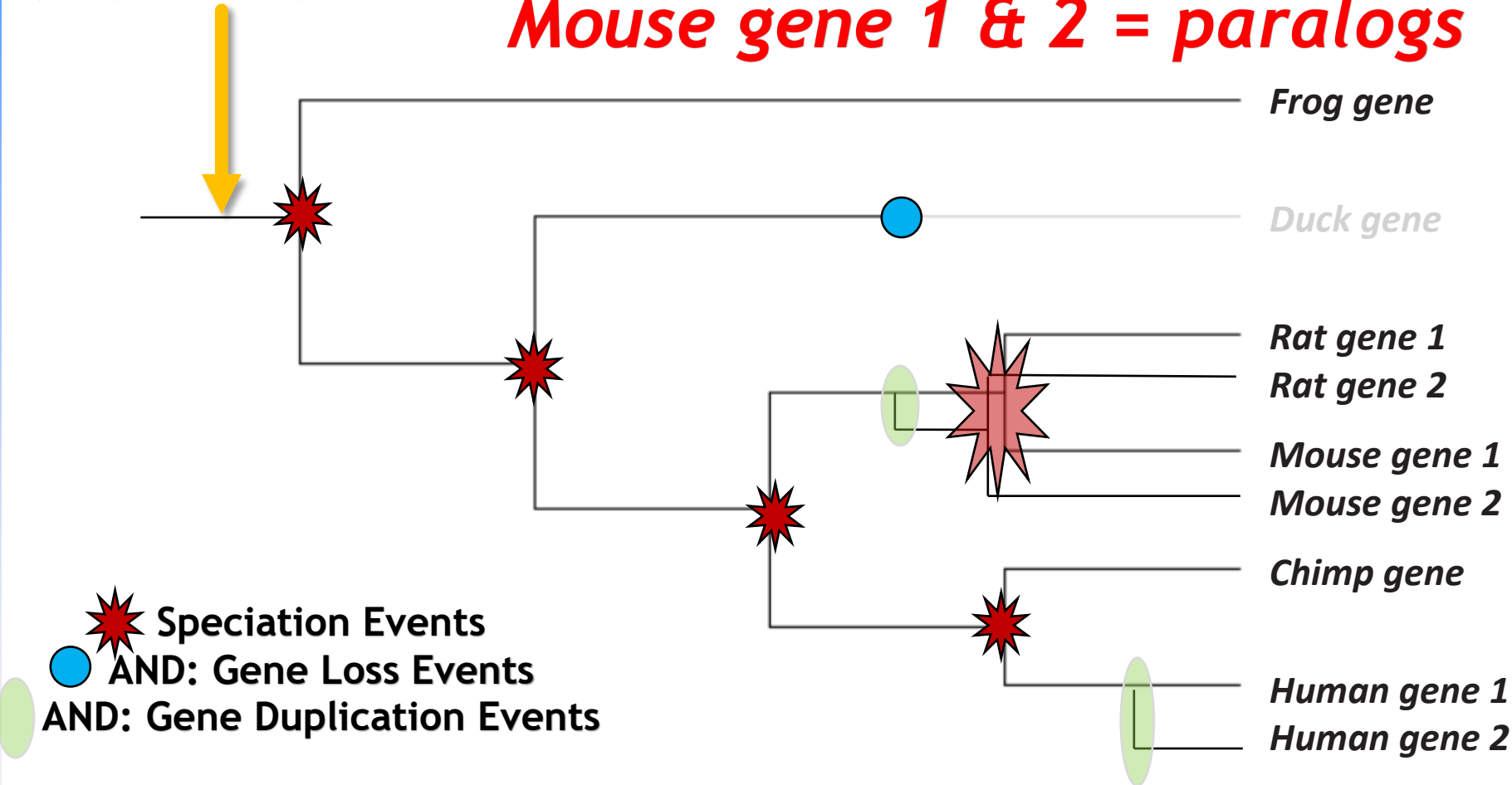
**Single-Copy Orthologs with Gains**



# Evolution ≠ simple

**Rat gene 1 & 2 = paralogs**  
**Mouse gene 1 & 2 = paralogs**

Last Common Ancestor  
(LCA) of all 6 species

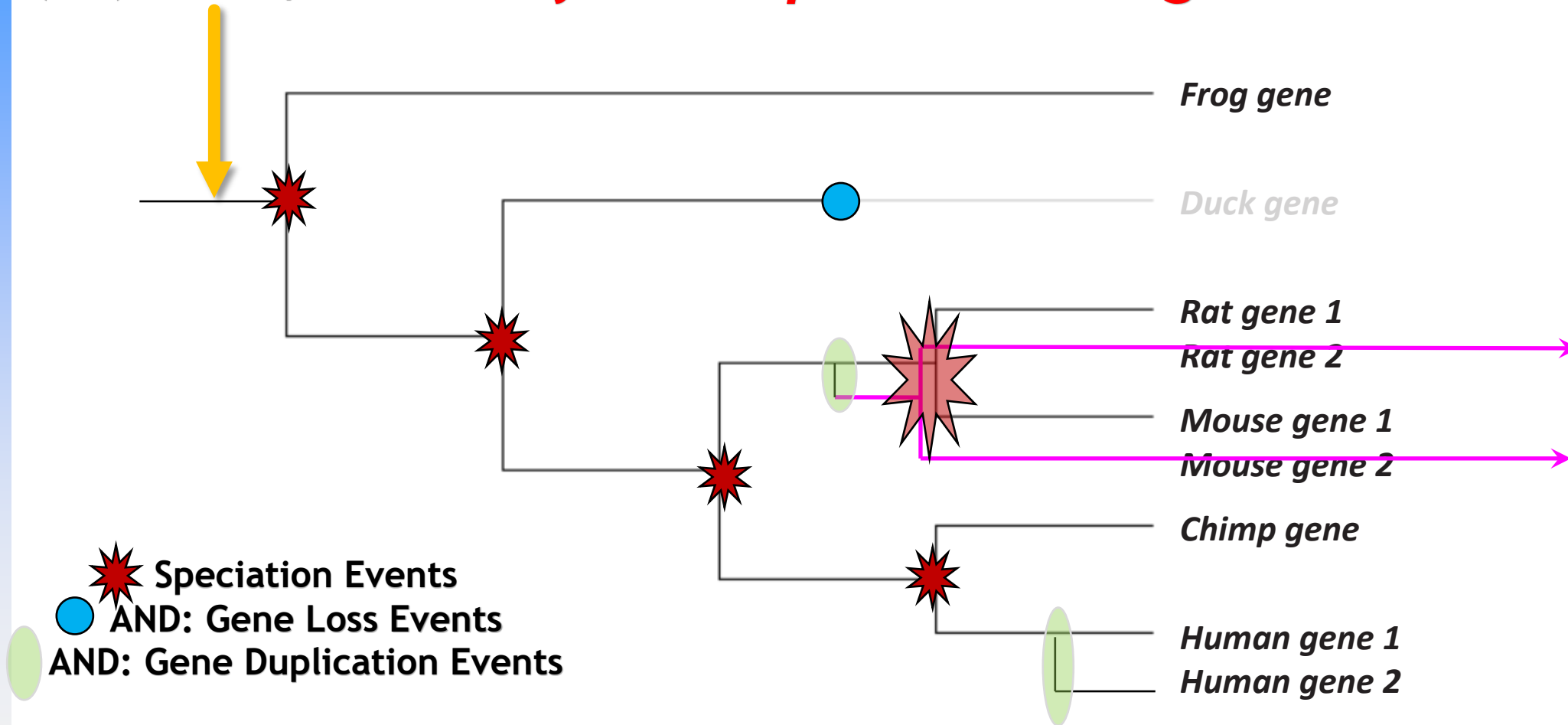


**Single-Copy Orthologs with Gains**

# Evolution ≠ simple

+ *fast sequence divergence*

Last Common Ancestor  
(LCA) of all 6 species



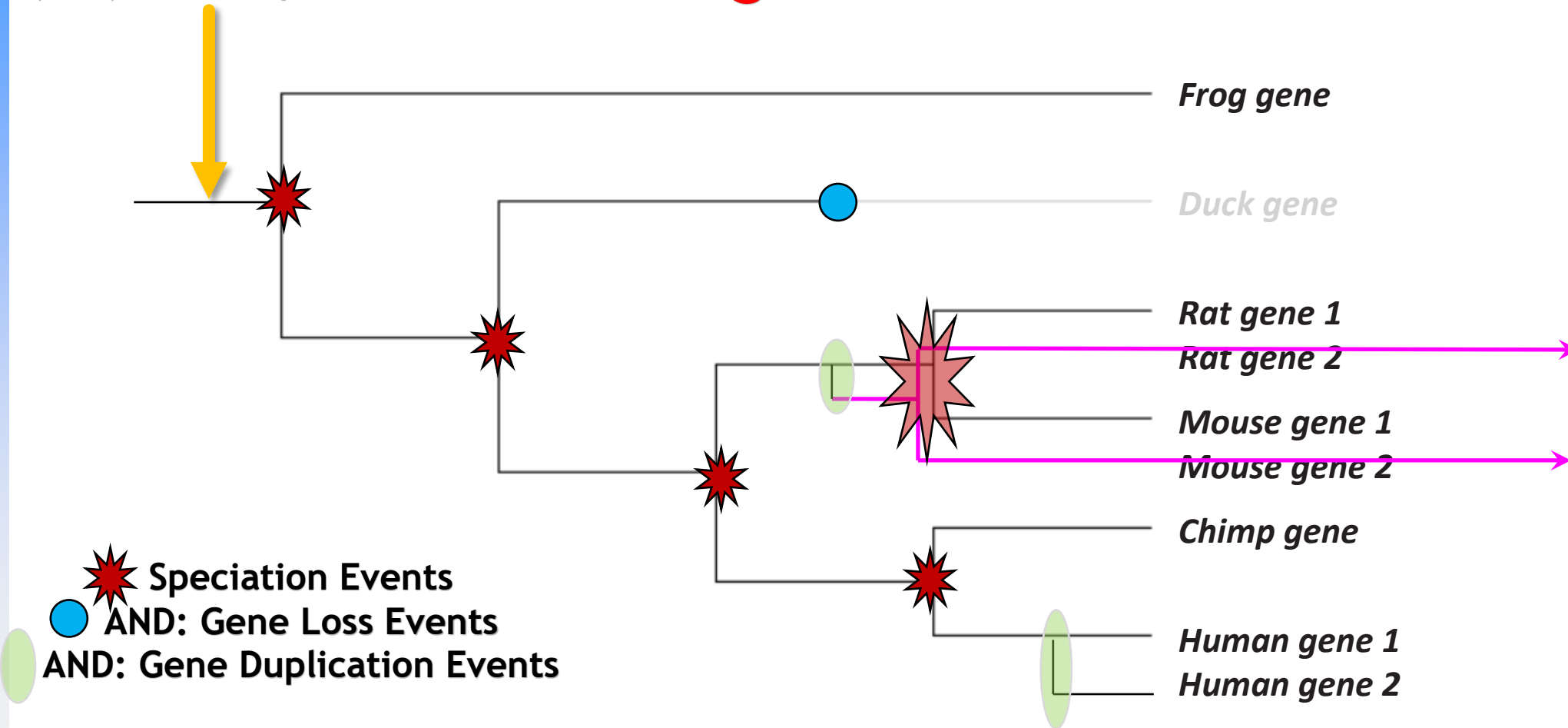
★ Speciation Events  
● AND: Gene Loss Events  
○ AND: Gene Duplication Events

*Single-Copy Orthologs with Gains*

# Evolution ≠ simple

**Paralogs  $R1+R2$   $M1+M2$   $H1+H2$**

Last Common Ancestor  
(LCA) of all 6 species



**Orthologs  $F+R1+R2+M1+M2+C+H1+H2$**



# Orthology – what is it?

## Homology

Recognizing similarities as evidence of shared ancestry

## Orthology

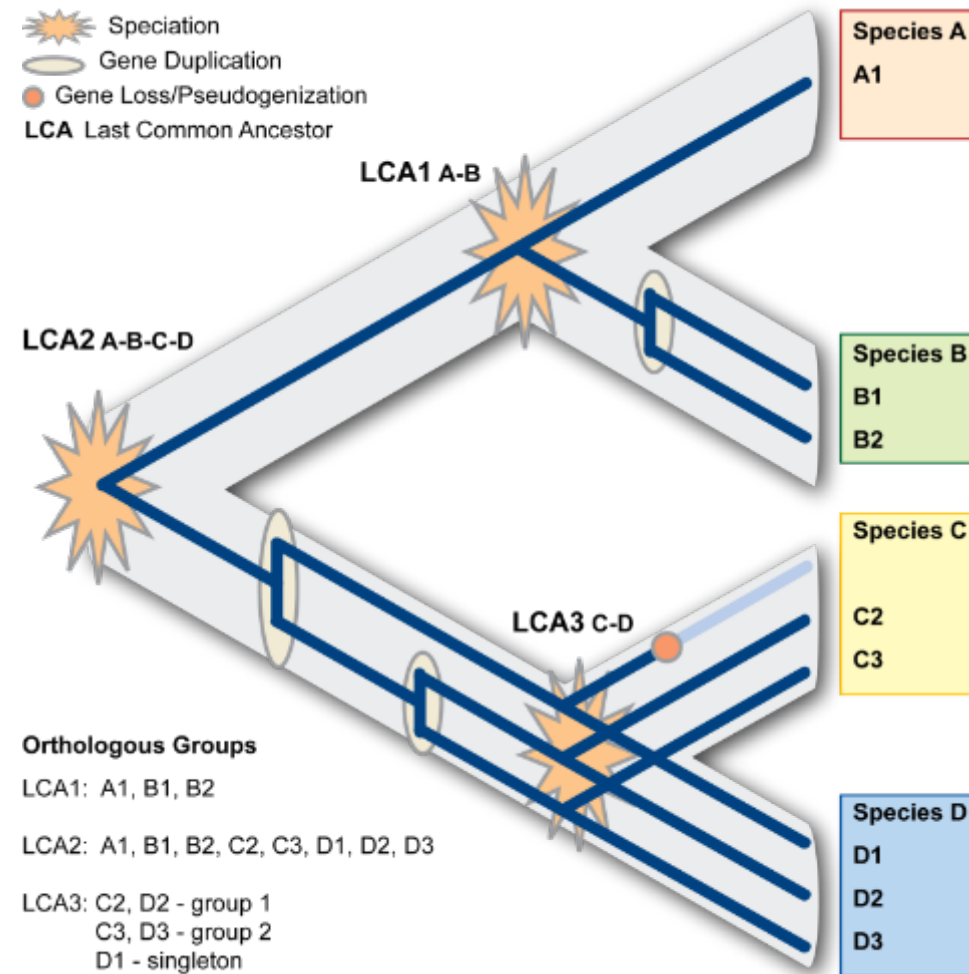
Orthologues arise by vertical descent from a single gene of the last common ancestor

## Hierarchy

Orthology is relative to the species radiation under consideration

## Orthologous Groups

All genes descended from a single gene of the last common ancestor



OrthoDB: the hierarchical catalog of eukaryotic orthologs in 2011

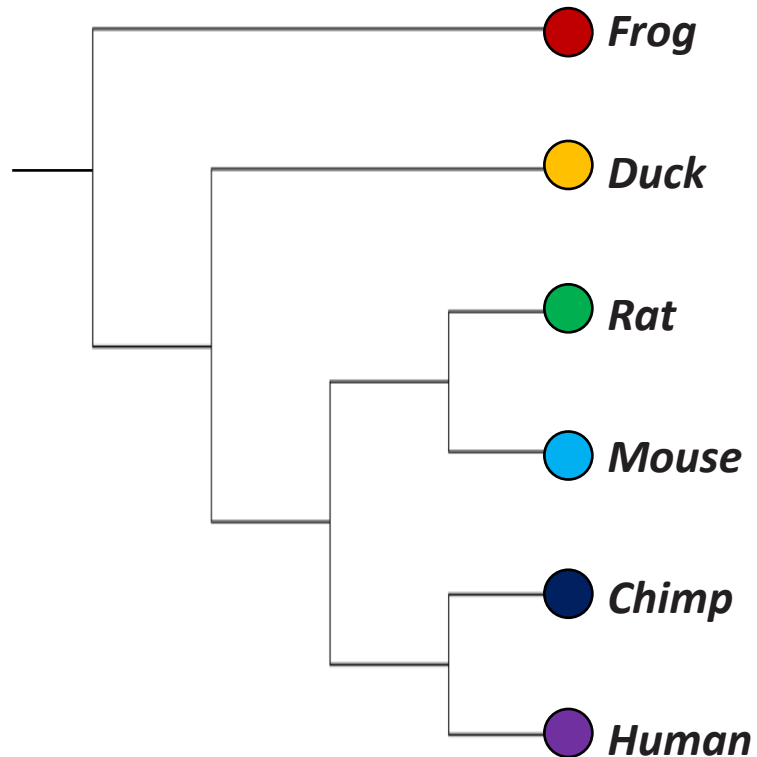
Robert M. Waterhouse<sup>1,2</sup>, Evgeny M. Zdobnov<sup>1,2,3</sup>, Fredrik Tegenfeldt<sup>1,2</sup>, Jia Li<sup>1,2</sup> and Evgenia V. Kriventseva<sup>1,2,\*</sup>

Nucleic Acids Research

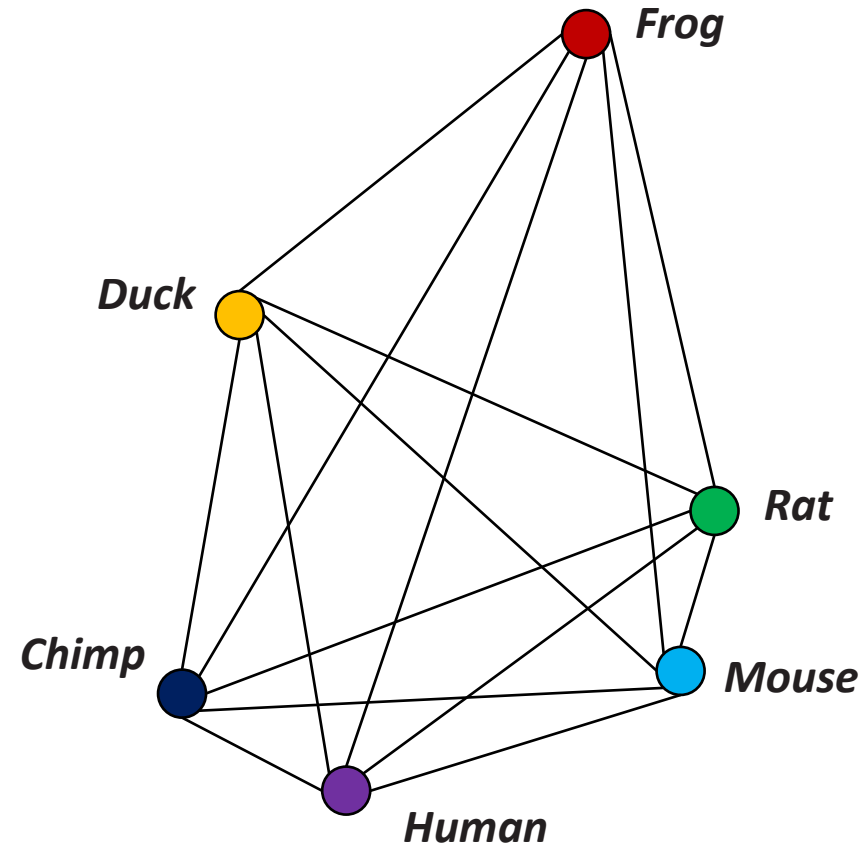


# How do we delineate Orthology?

tree-based approaches



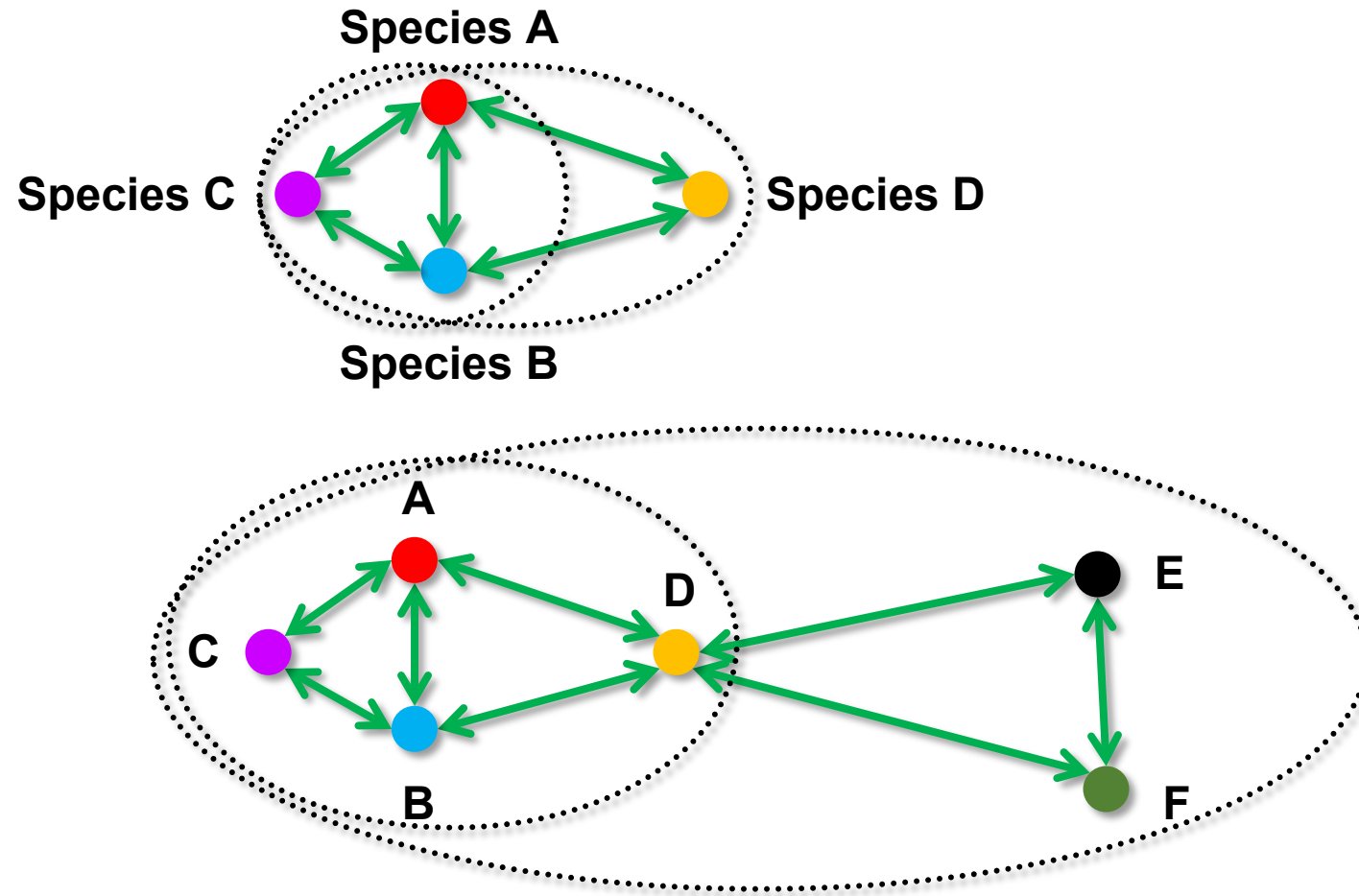
graph-based approaches



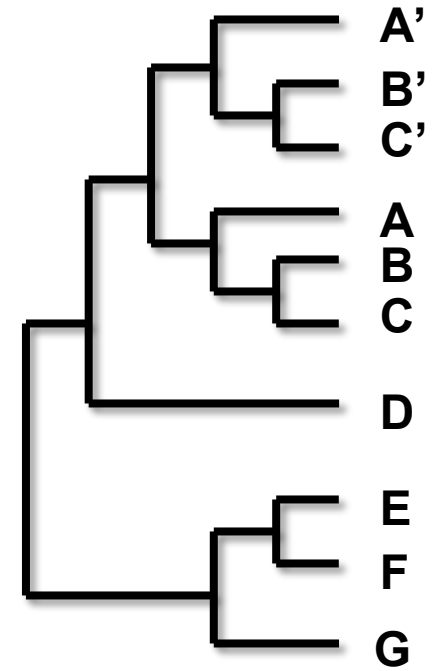
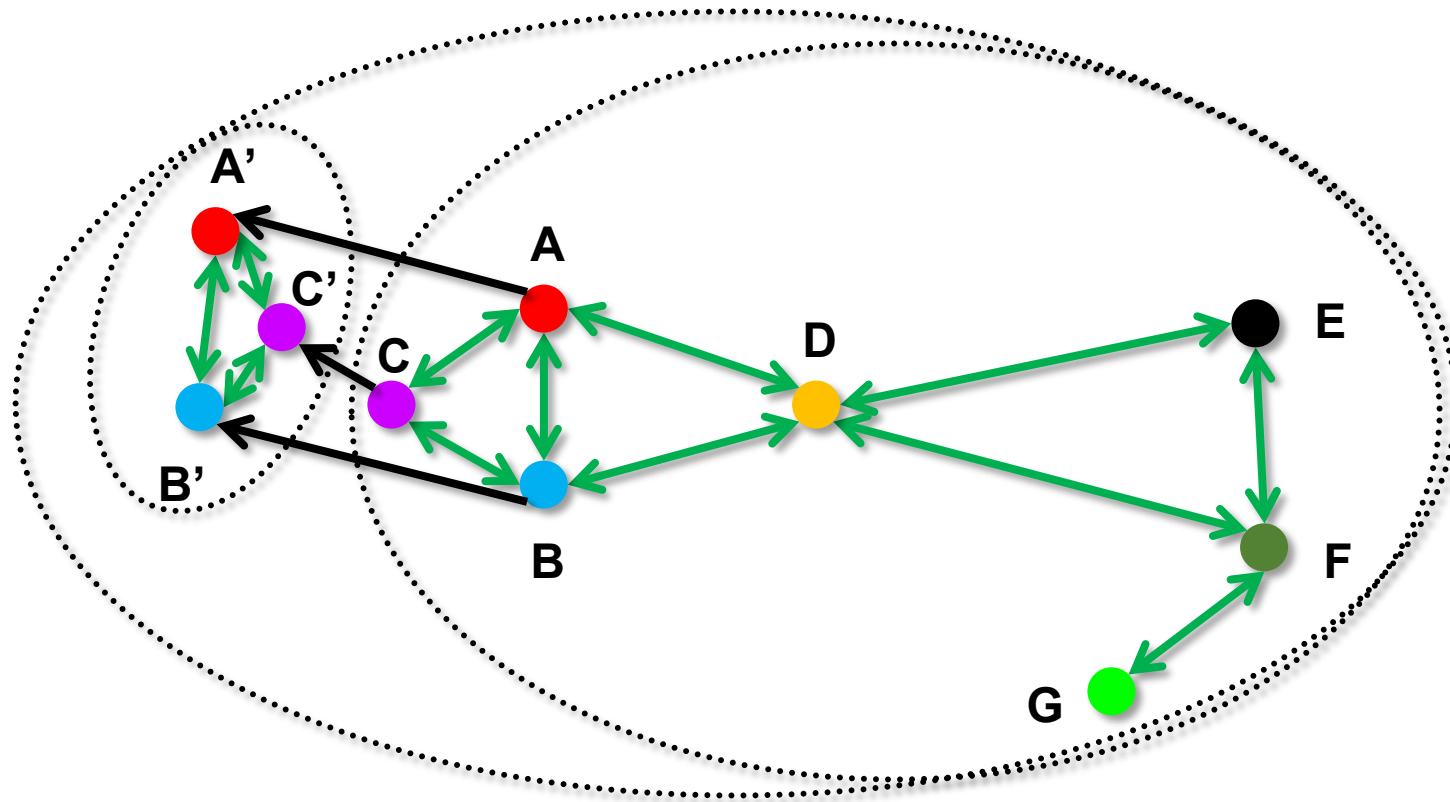
**Single-Copy Orthologs**



# ***Graph-based best-reciprocal-hits***



# ***Within-clade duplications***



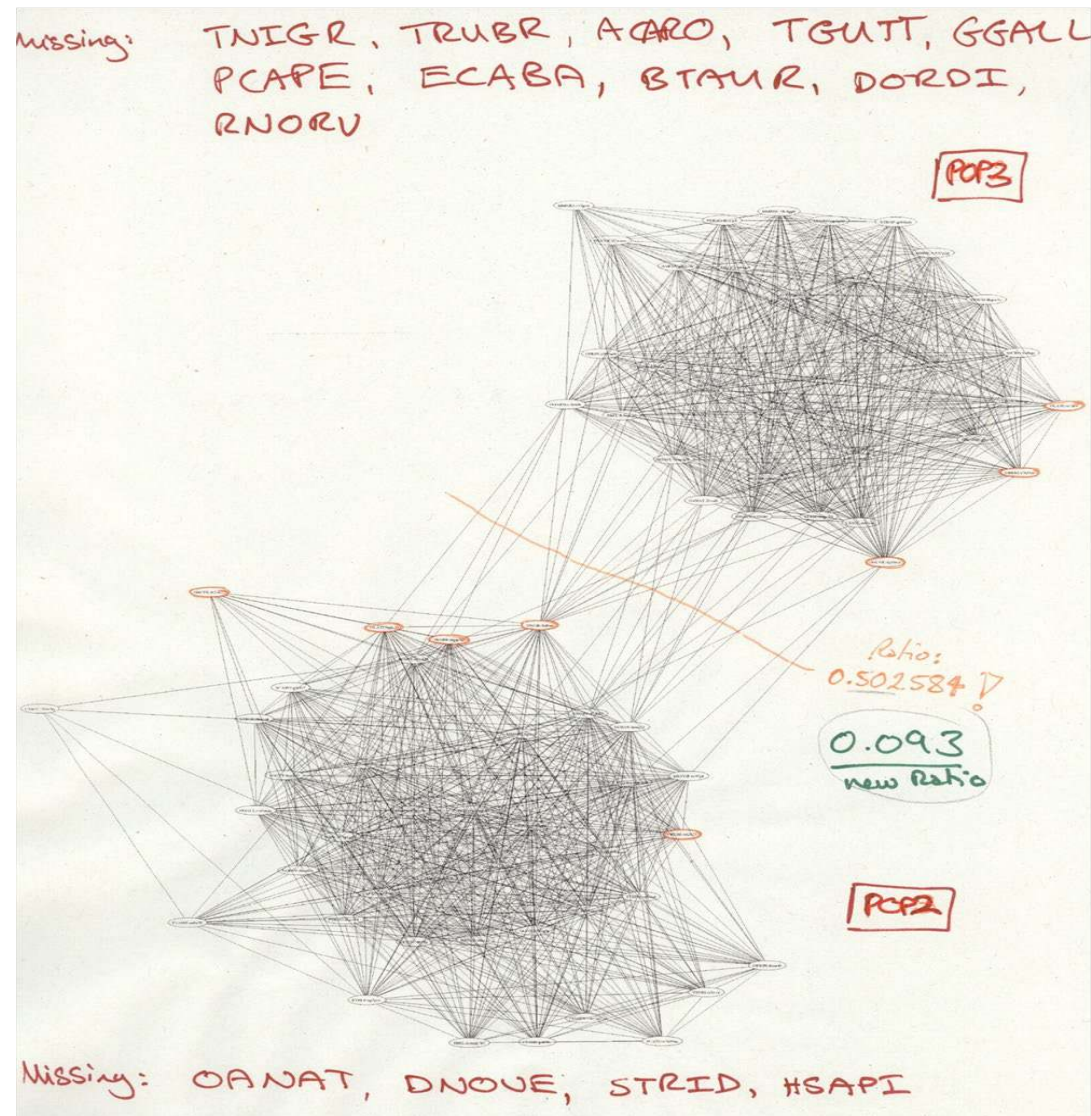
# Real-world data can be messy!

Real example:

POP3 missing from 10 vertebrates

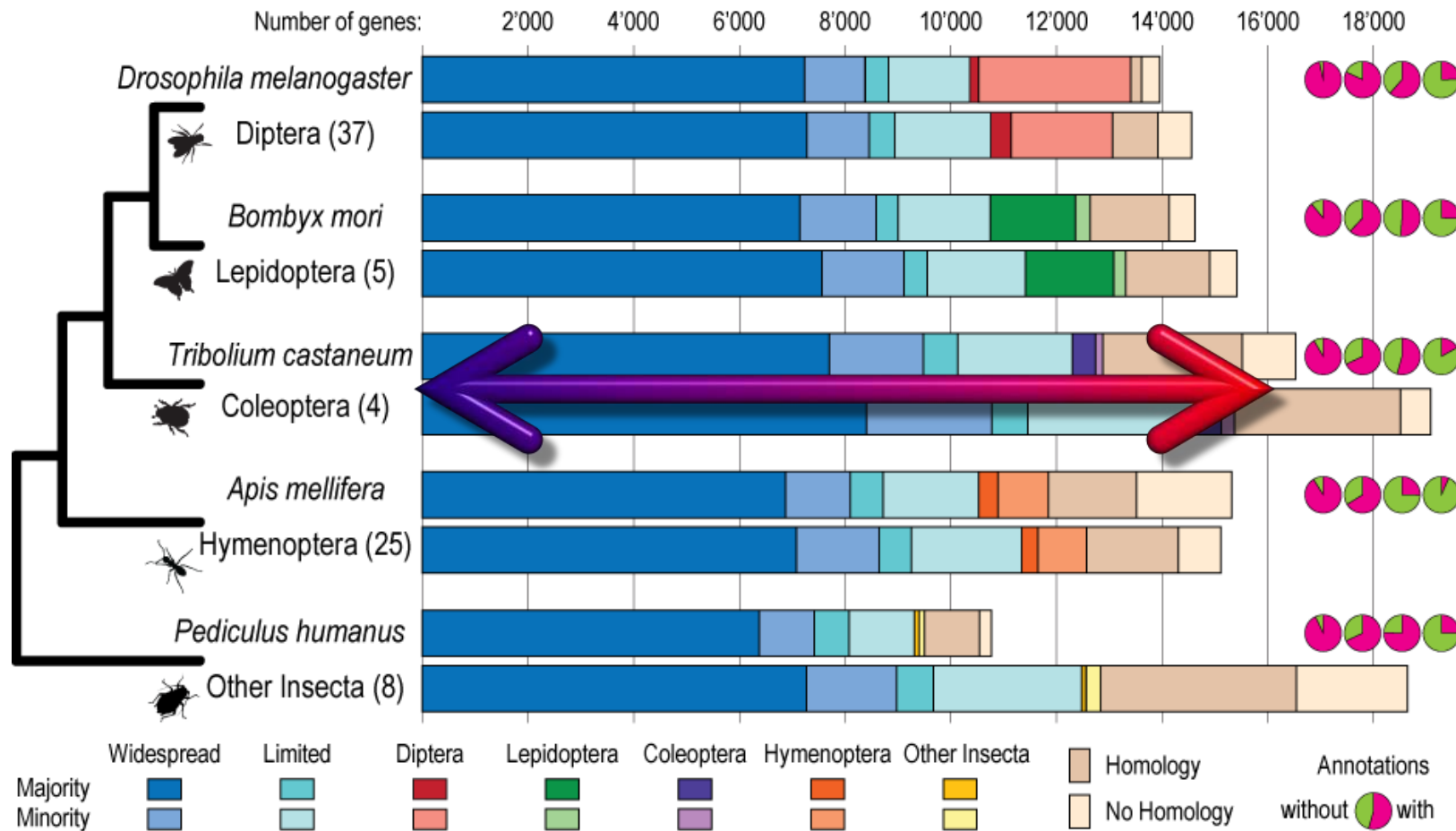
POP2 missing from 4 vertebrates

Two orthologous groups start to merge into one



# Orthology – why do we need it?

- 1) Tracing the **Evolutionary Histories** of all genes in extant species
- 2) Building **Hypotheses on Gene Function** informed by evolution



# ***Orthology ≠ Function ... BUT ...***

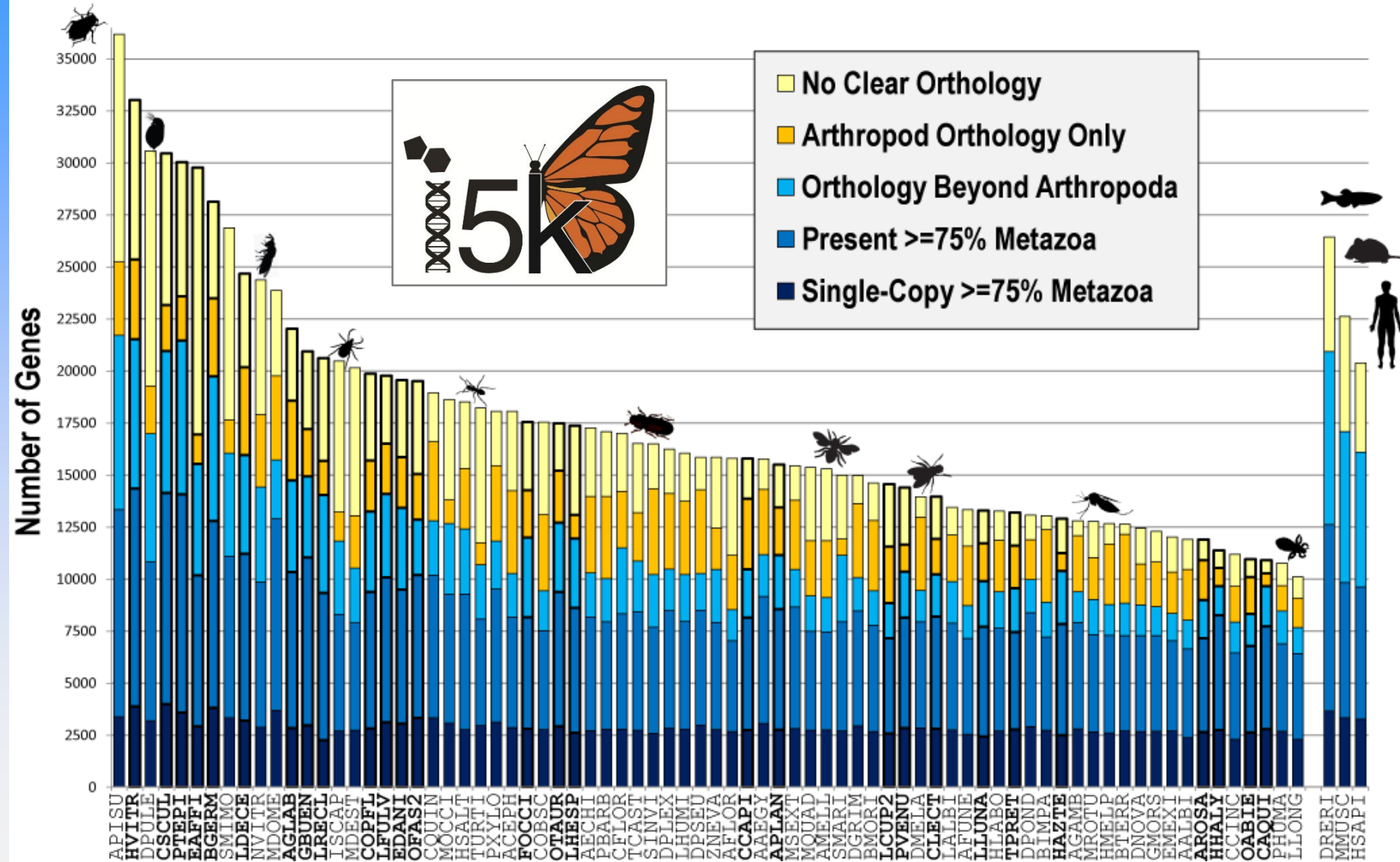
By tracing the **Evolutionary Histories** of all genes in extant species  
We can build **Hypotheses on Gene Function** informed by evolution

“validity of the conjecture on **functional equivalency** of orthologs is crucial for reliable annotation of newly sequenced genomes and, more generally, for the progress of functional genomics.

The huge majority of genes in the sequenced genomes will **never be studied experimentally**, so for most genomes **transfer of functional information** between orthologs is the only means of detailed functional characterization.”



# Evolutionary histories: classes

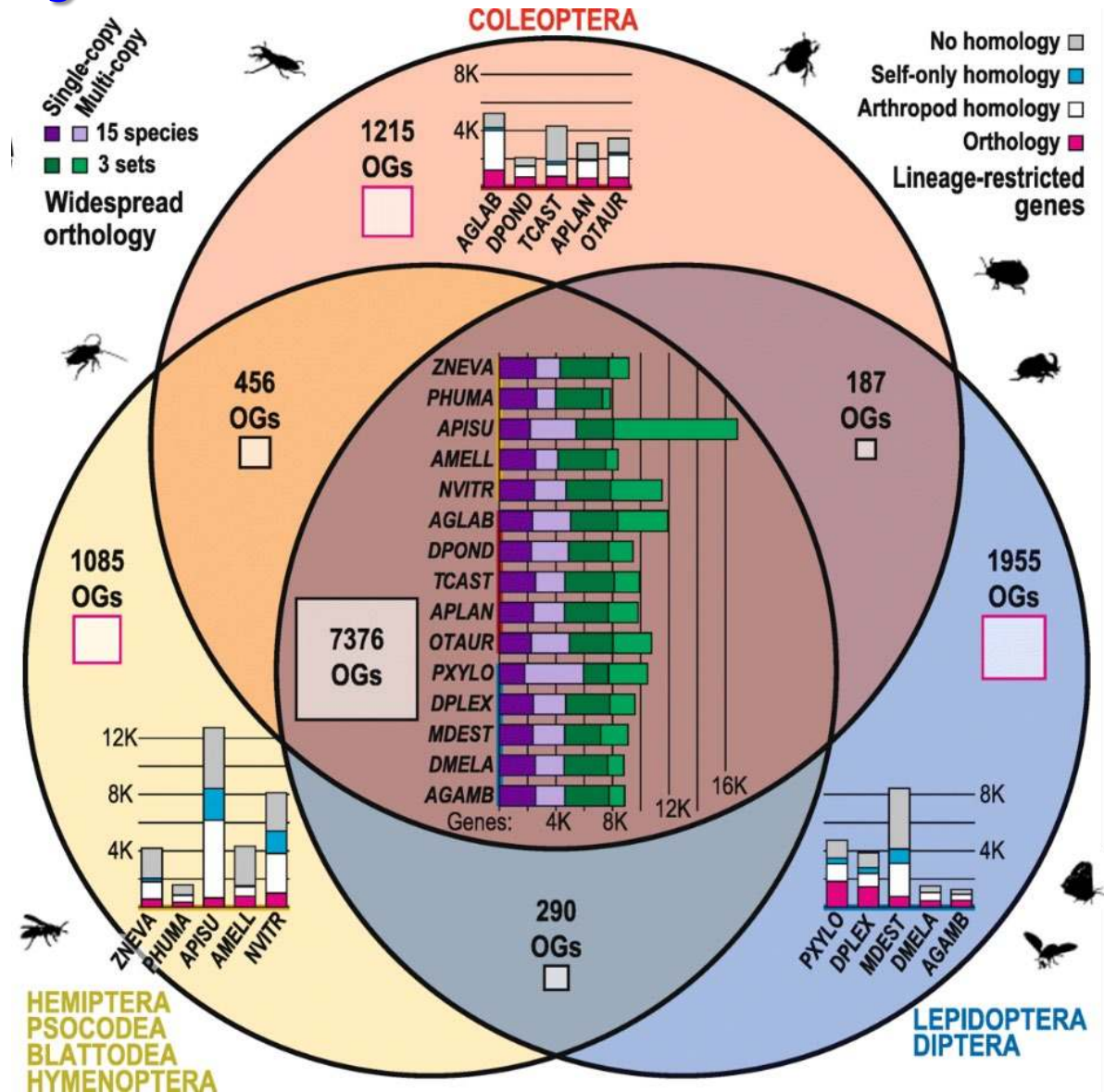


Unique  
Variable  
Common

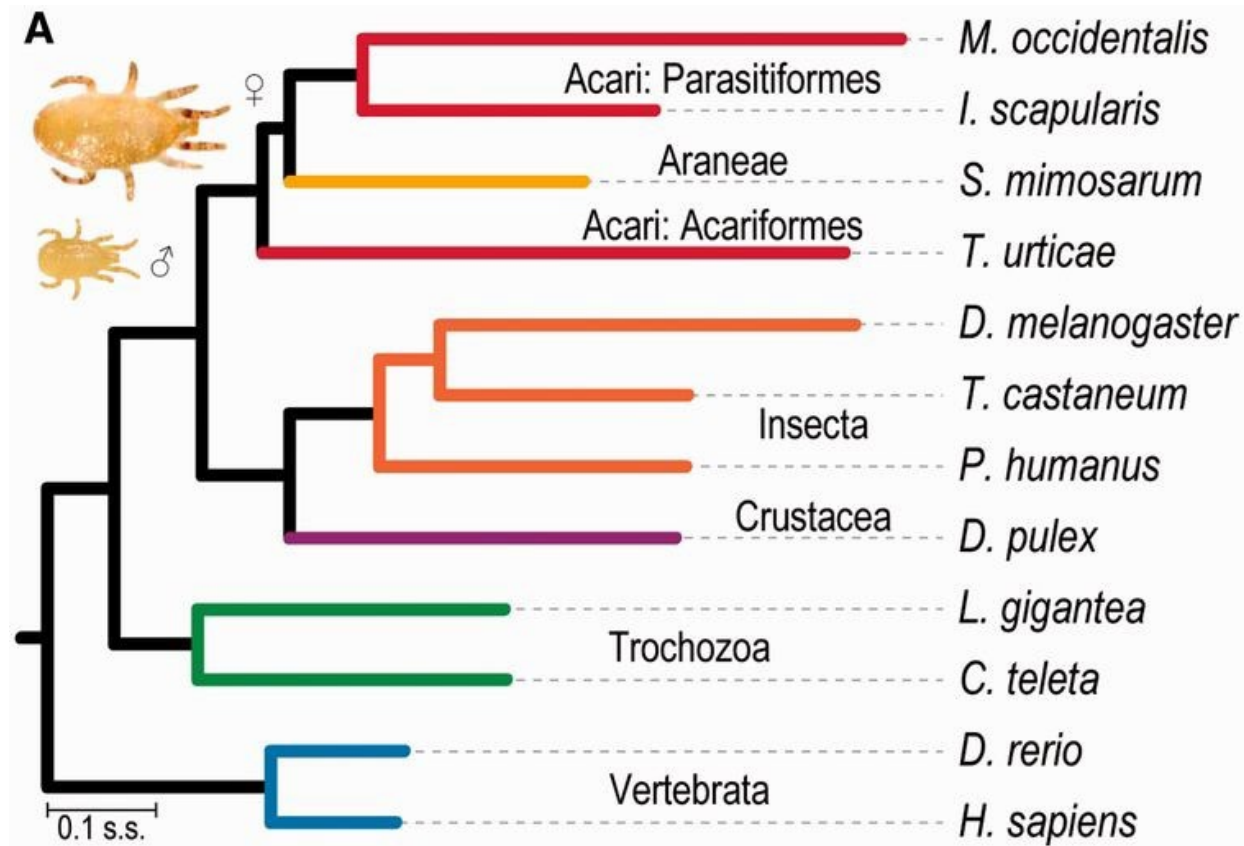


# Evolutionary histories: classes

## Clade-specific & variable-count orthologues



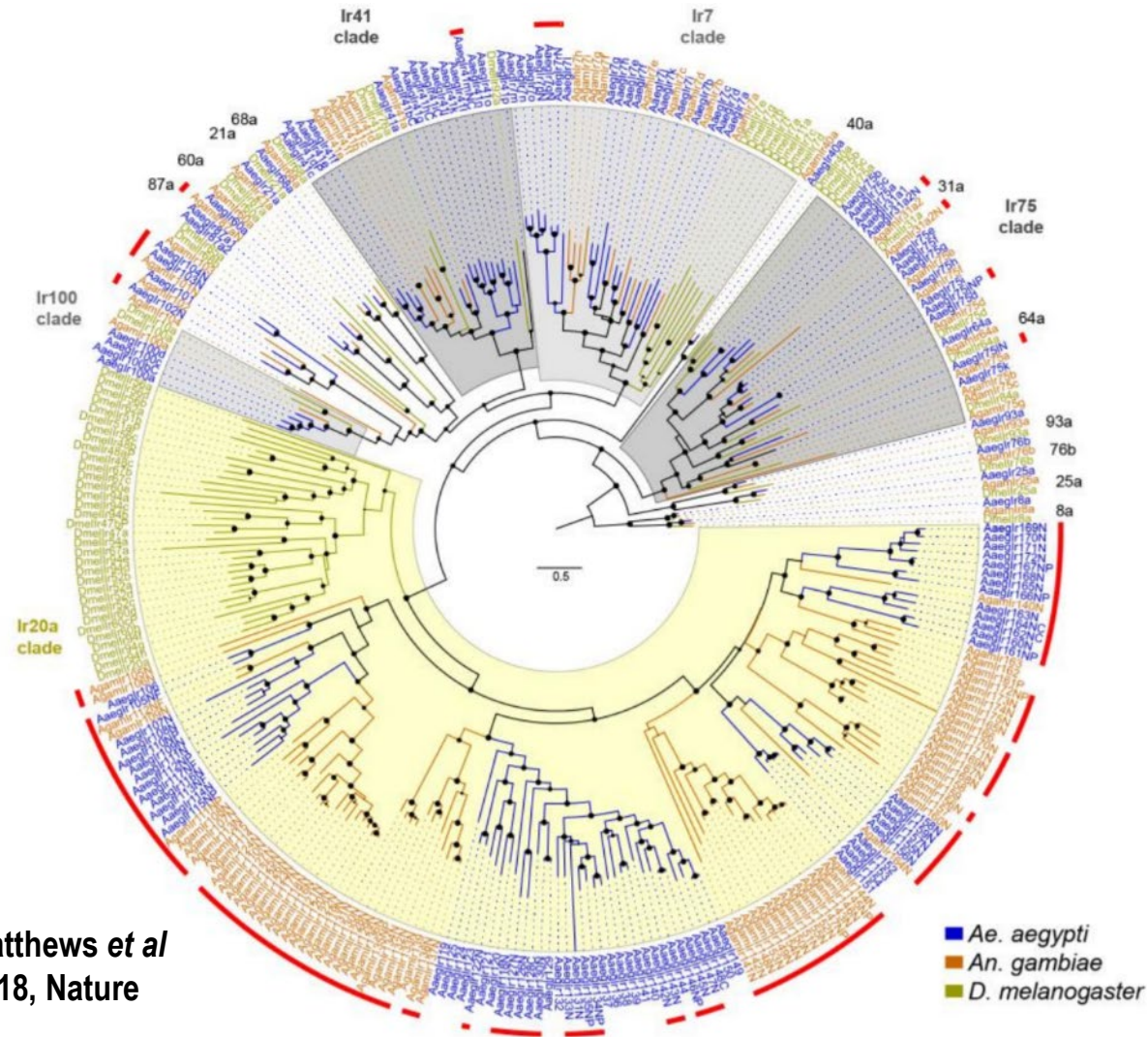
# Species Tree Estimation



Phylogenomics with single-copy orthologues



# Gene Family Tree Building



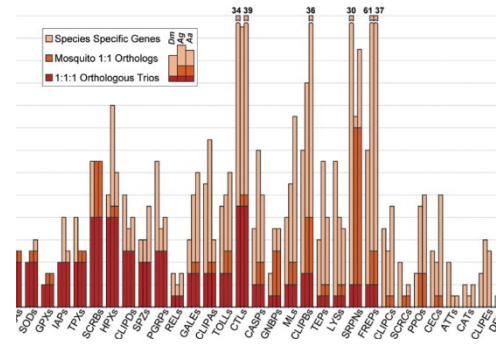
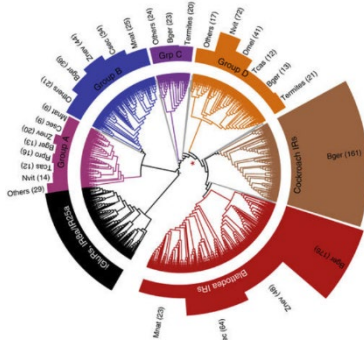
Matthews et al  
2018, Nature



# Dynamically evolving families

Many of the most biologically interesting genes and gene families show highly dynamic evolutionary histories

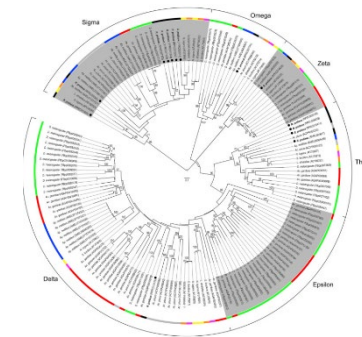
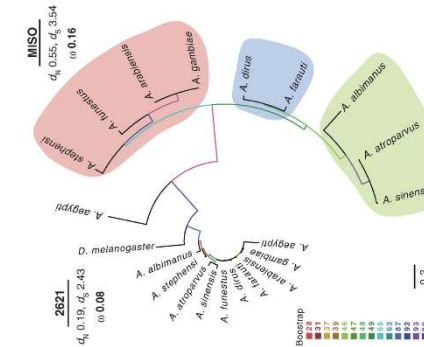
IMMUNITY



CHEMOSENSATION

DETOXIFICATION

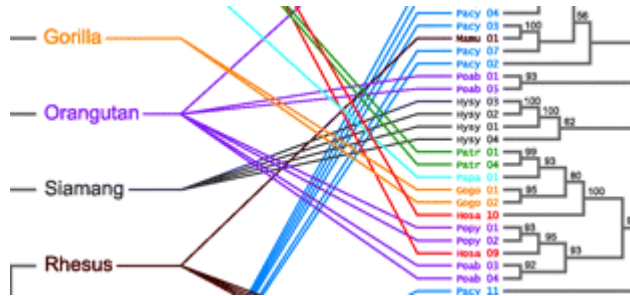
REPRODUCTION



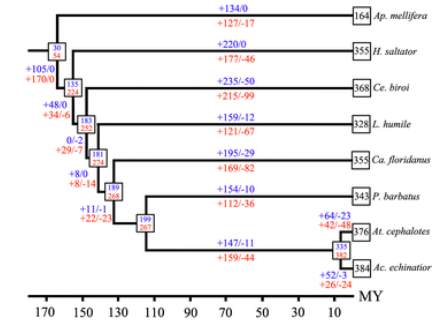
# Inferring gene evolutionary histories



## Gene-tree-species-tree reconciliation



## Gene ancestral state reconstruction



## INPUTS

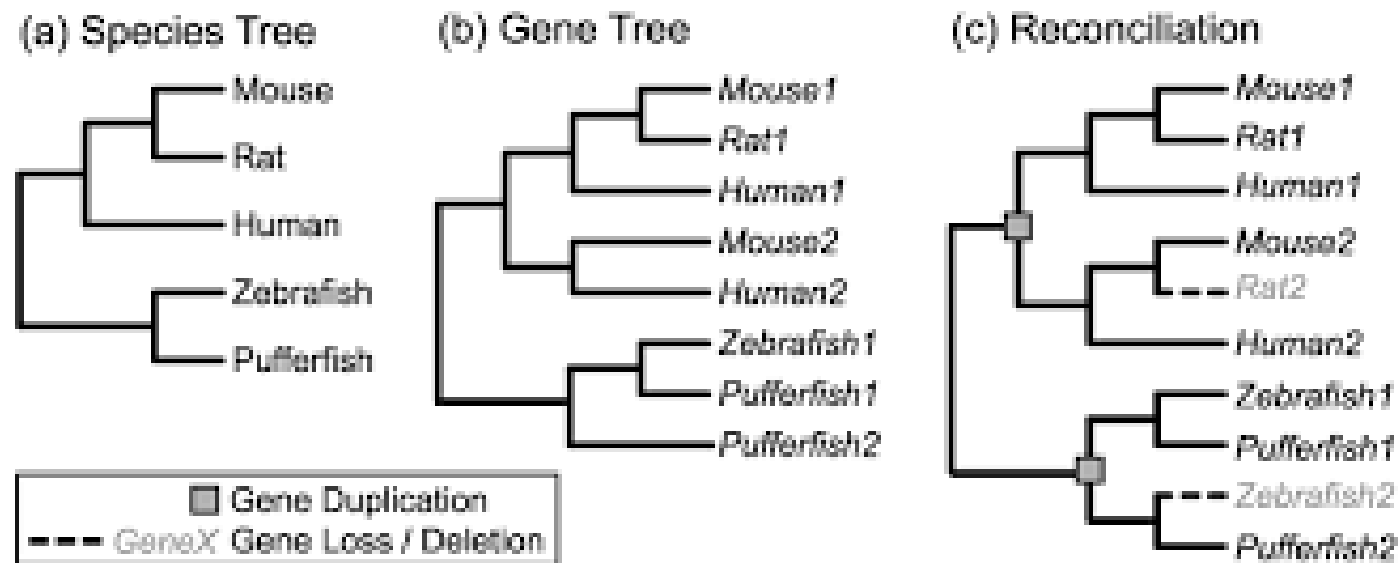
- Confident species phylogeny
- Individual gene trees
  - From orthologous groups
  - From homologous gene families

- Confident species phylogeny
- Counts of orthologues per species
  - From orthologous groups
  - From homologous gene families



# Gene-tree-species-tree reconciliation

A gene tree-species tree reconciliation explains the evolution of a gene tree within the species tree given a model of gene-family evolution



- Given all possible duplication and loss events that are compatible
- Compute the minimum “cost” resolution
  - Impacted by assumptions on costs for duplication and losses

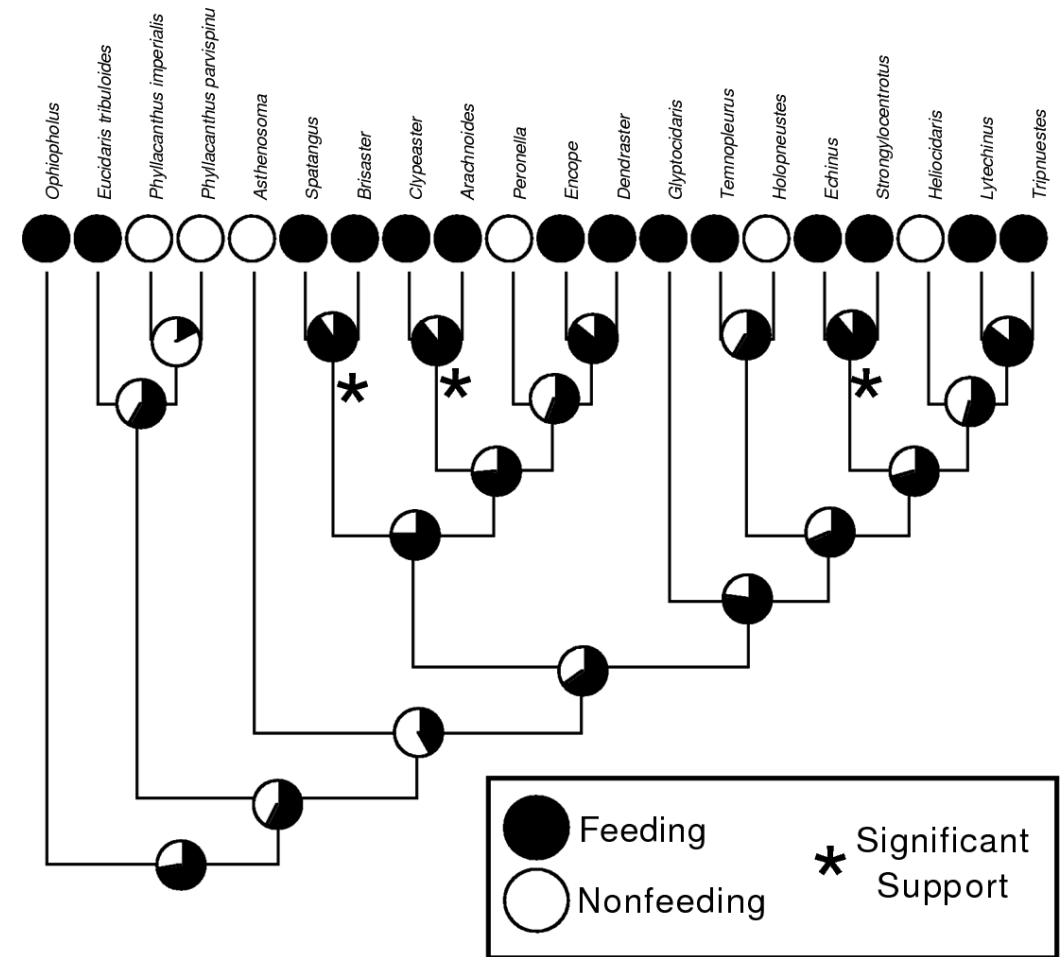


# Gene ancestral state reconstruction

Ancestral state reconstruction in general is the extrapolation back in time from measured characteristics of individuals to their common ancestors

Ancestral gene content reconstruction follows the same principles:

- Extant characters are gene counts
- A gene **birth and death** process is used to model gene gain and loss across a user-specified phylogenetic tree



# ***Inferring gene evolutionary histories***

Reconciliation or Reconstruction is used to map evolutionary events  
– gene gains and losses onto a species phylogeny

Understanding how these event relate to the biology and evolution of the organisms being studied then requires additional data

- phenotype data like organismal traits (metabolism, ecology, etc)
- functional genomics data like expression / functional annotations

The first steps of data analysis are almost universal:

- [1] delineation of families and/or orthologous groups
- [2] building a robust species phylogeny



# ***Orthology Delineation***

***What is orthology?***

***How do we delineate orthologs?***

***And why do we need to?  
(species/gene trees)***



# ***Break time***

