Biodiversity Bioinformatics: from large-scale phylogenomics to gene families and functions

DAY 2: August 27th 2024

Robert M. Waterhouse

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

UNIL | Université de Lausanne

Swiss Institute of Bioinformatics

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

robert.waterhouse@sib.swiss @rmwaterhouse www.rmwaterhouse.org

Instructor Biography-Introduction

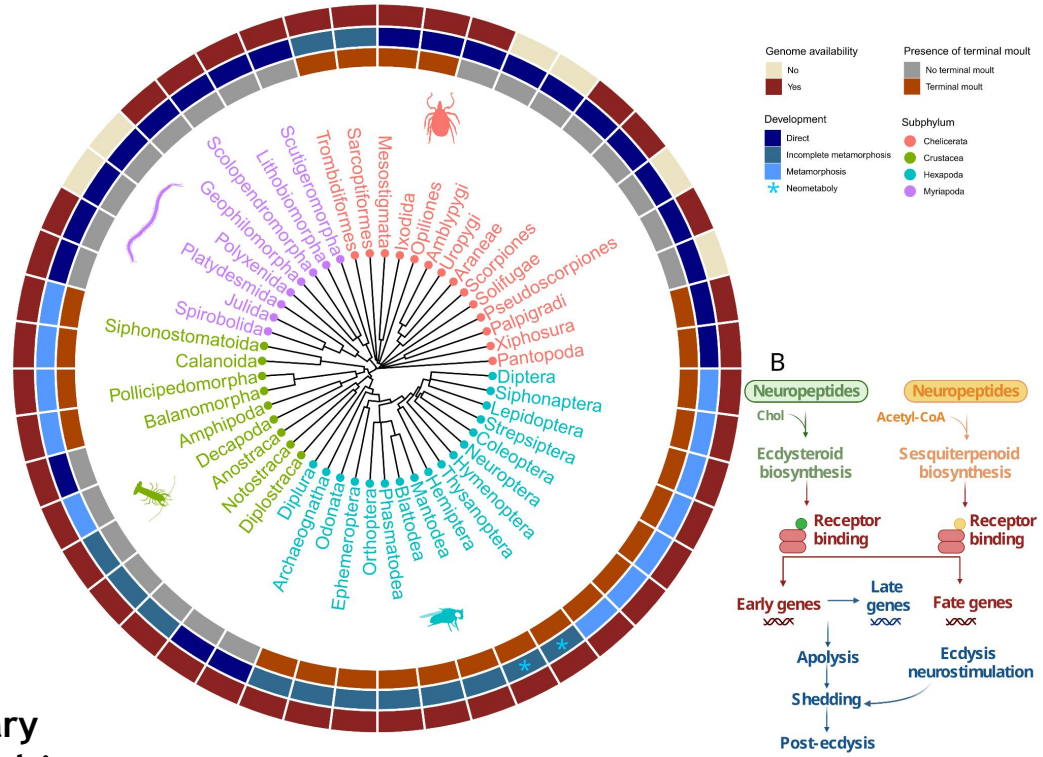
2023-	Director, Environmental Bioinformatics Group SIB Swiss Institute of Bioinformatics	
2017-23	SNF Assistant Professor University of Lausanne	
2015-16	Marie Curie Fellow & Maître assistant University of Geneva *ZDOBNOV*	
2013-14	Marie Curie Outgoing Fellow Massachusetts Institute of Technology *KELLIS*	
2009-12	Postdoctoral Researcher University of Geneva *ZDOBNOV*	
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	

Teaching Assistant

Giulia
Campli

Comparative evolutionary genomics of arthropod moulting

A



B

Goals for Today's Workshop

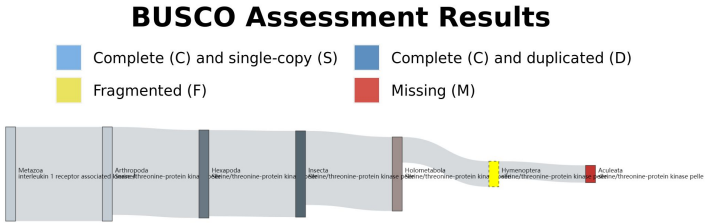
- ❑ Understand the principles of graph-based orthology delineation using OrthoDB as an example
- ❑ Learn how to browse and query OrthoDB
- ❑ Learn how to use BUSCO to assess genomics data
- ❑ Learn how collect input data from an orthology database to build a phylogenetic tree
- ❑ Learn how to build gene and species trees and perform gene-tree-species-tree reconciliation
- ❑ Learn how to use orthologue gene counts to estimate ancestral gene content
- ❑ Critically compare results from tree reconciliations and ancestral state/count reconstructions

OrthoDB

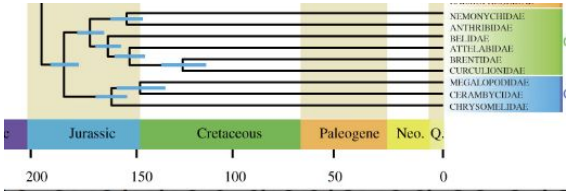
BUSCO

Comparative Genomics Hands-On: Concepts and Applications

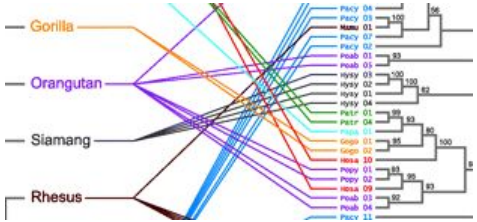
OrthoDB orthology and BUSCO quality



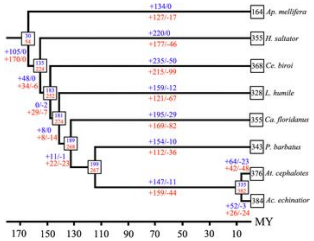
Species & gene tree estimation



Gene-tree-species-tree reconciliation



Gene ancestral state reconstruction



Comparative Genomics Hands-On: Databases and Analysis Software

OrthoDB - orthology database

BUSCO - Benchmarking Universal Single-Copy Orthologues

MAFFT - multiple sequence alignment tool

TrimAl - alignment filtering tool

RAxML - maximum likelihood phylogenies

Treerecs - tree reconciliation software

CAFE - computational analysis of gene family evolution

iTOL - online phylogenetic tree viewer

Chronos function of the ape package in R

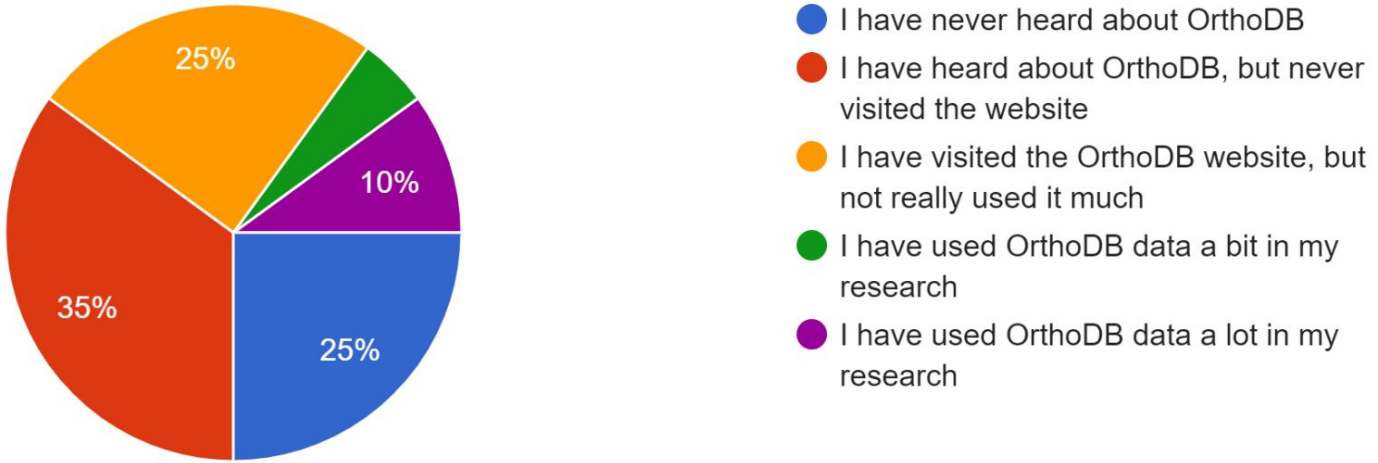
Quick Quiz

https://forms.gle/9FHVpcK4FZ4riZx6A

SwissOrthology

How familiar are you with OrthoDB, the hierarchical catalogue of orthologues?

20 responses



How familiar are you with BUSCO, the Benchmarking Universal Single-Copy Orthologues?

20 responses

Orthology Delineation

What is orthology?

How do we delineate orthologs?

And why do we need to?
(species/gene trees/copy-number)

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, Paralogs, and Evolutionary Genomics[1]

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, Paralogs, and Evolutionary Genomics[1]

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[1]

Eugene V. Koonin

Annu. Rev. Genet.
2005. 39:309–38

Orthology – what is it?

Homologs

Common Ancestor

Orthologs

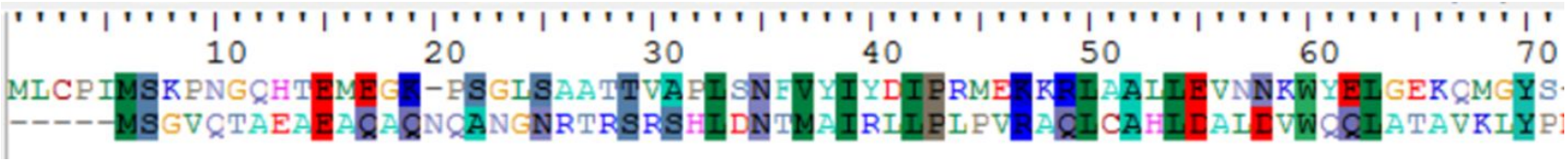
Speciation Event

Paralogs

Duplication Event

Sequence Homology – what is it?

Homology between protein or DNA sequences is typically inferred from their sequence similarity



Sequence homology search tools, e.g. BLAST,
attempt to detect '**excess**' similarity
i.e. greater similarity or identity than expected by chance
=> statistically significant similarity

Sequence Homology – what is it?

"the link between **similarity** and **homology** is often misunderstood"

An Introduction to Sequence Similarity ("Homology") Searching

William R. Pearson[1]
[1]University of Virginia School of Medicine, Charlottesville, VA

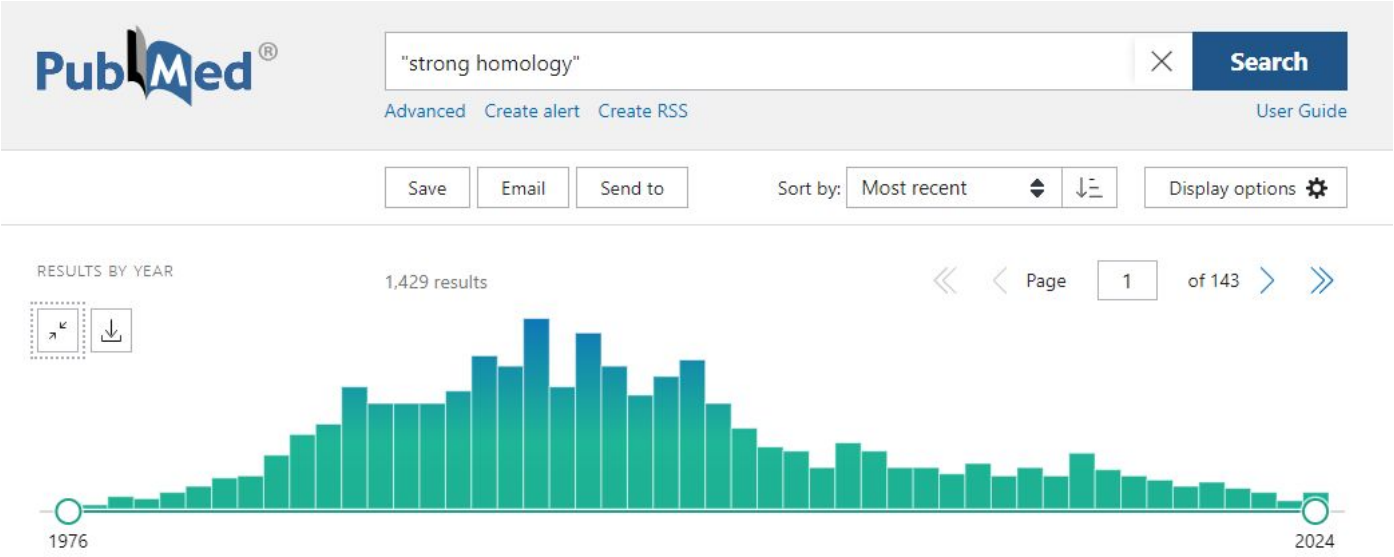
A pair of sequences can have **high** or **low** sequence similarity

But this does not translate to **strong** or **weak** homology!

Homology is the **conclusion**, i.e. given the level of similarity the sequences are likely to have arisen from a common ancestor

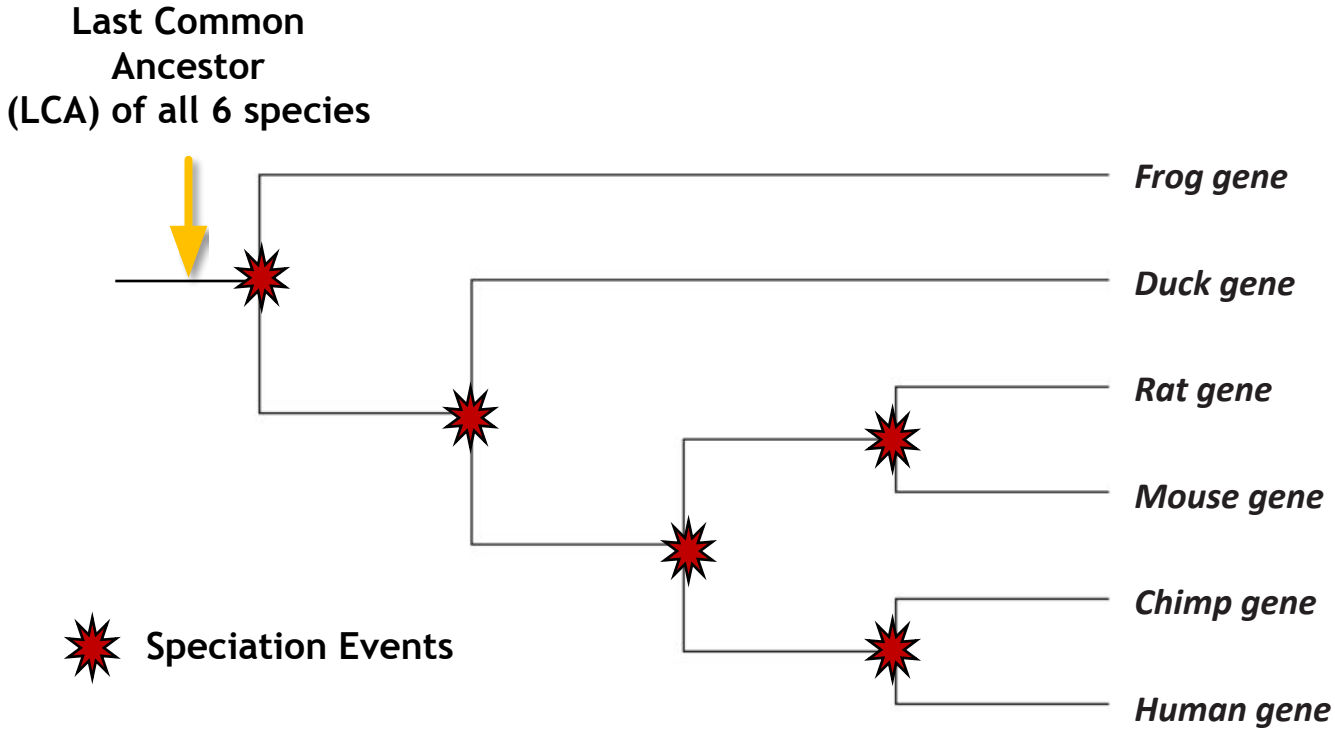
Sequence Homology – what is it?

"the link between **similarity** and **homology** is often misunderstood"



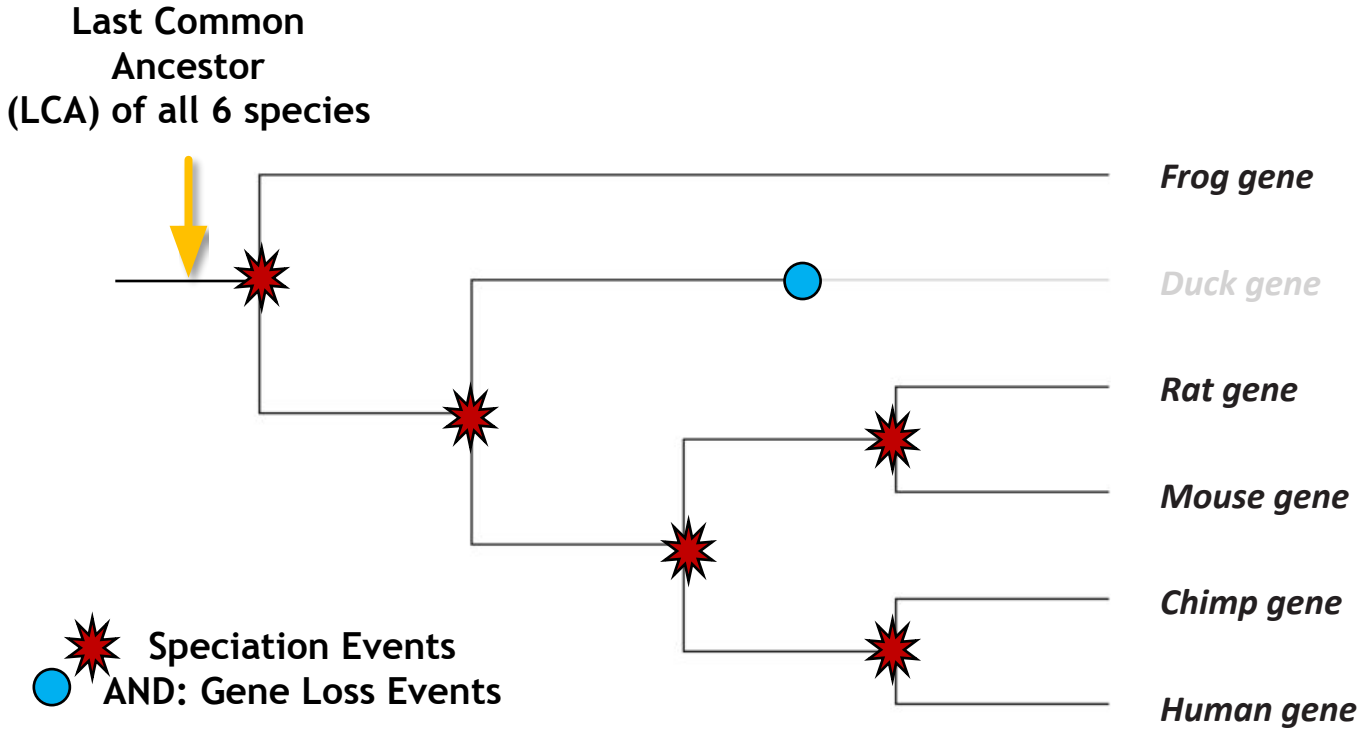
It is still worth pointing this out in 2024!

Orthology – simple scenario



Single-Copy Orthologs

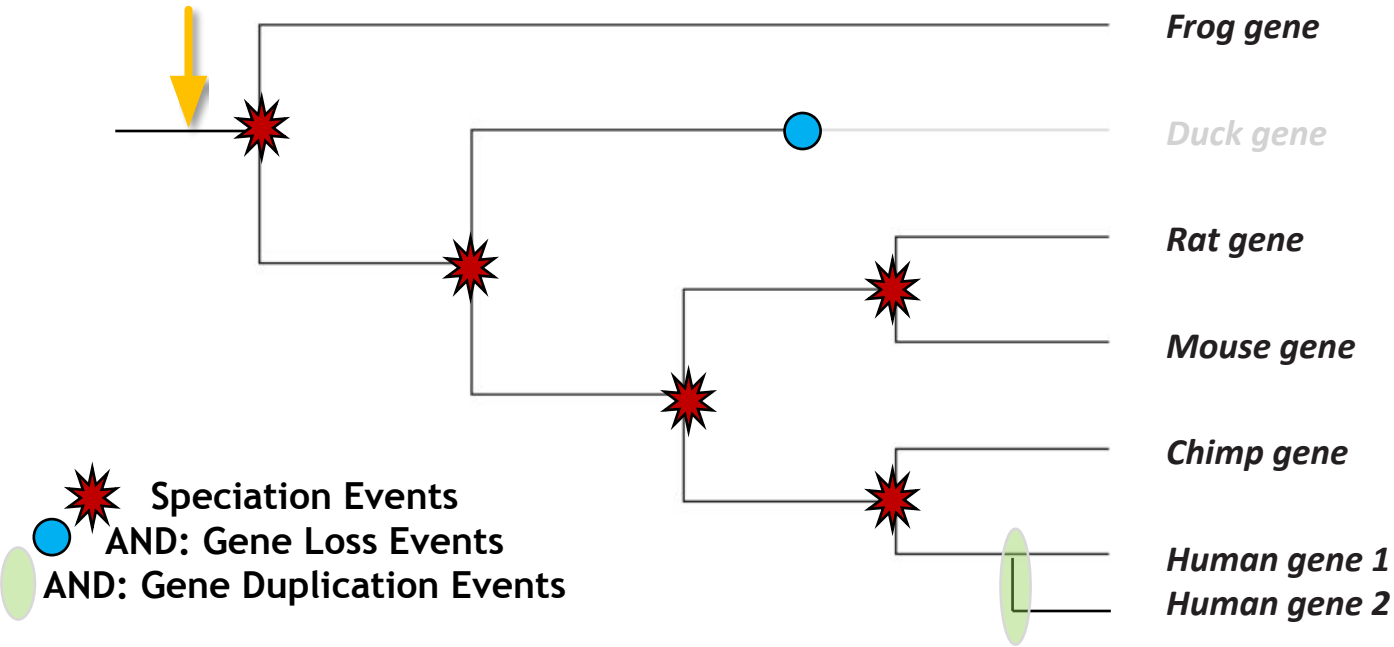
Evolution ≠ simple



Single-Copy Orthologs with Losses

Evolution ≠ simple

Human gene 1 & 2 = paralogs



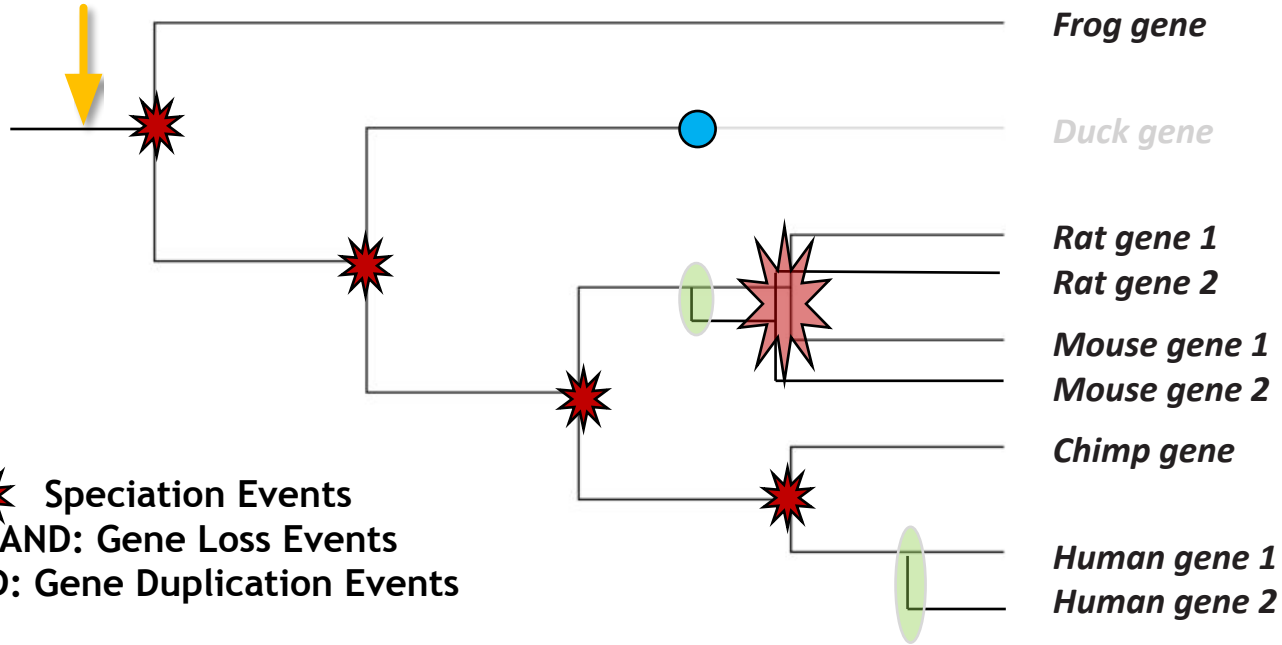
Single-Copy Orthologs with Gains

Evolution ≠ simple

Rat gene 1 & 2 = paralogs

Mouse gene 1 & 2 = paralogs

Last Common Ancestor (LCA) of all 6 species

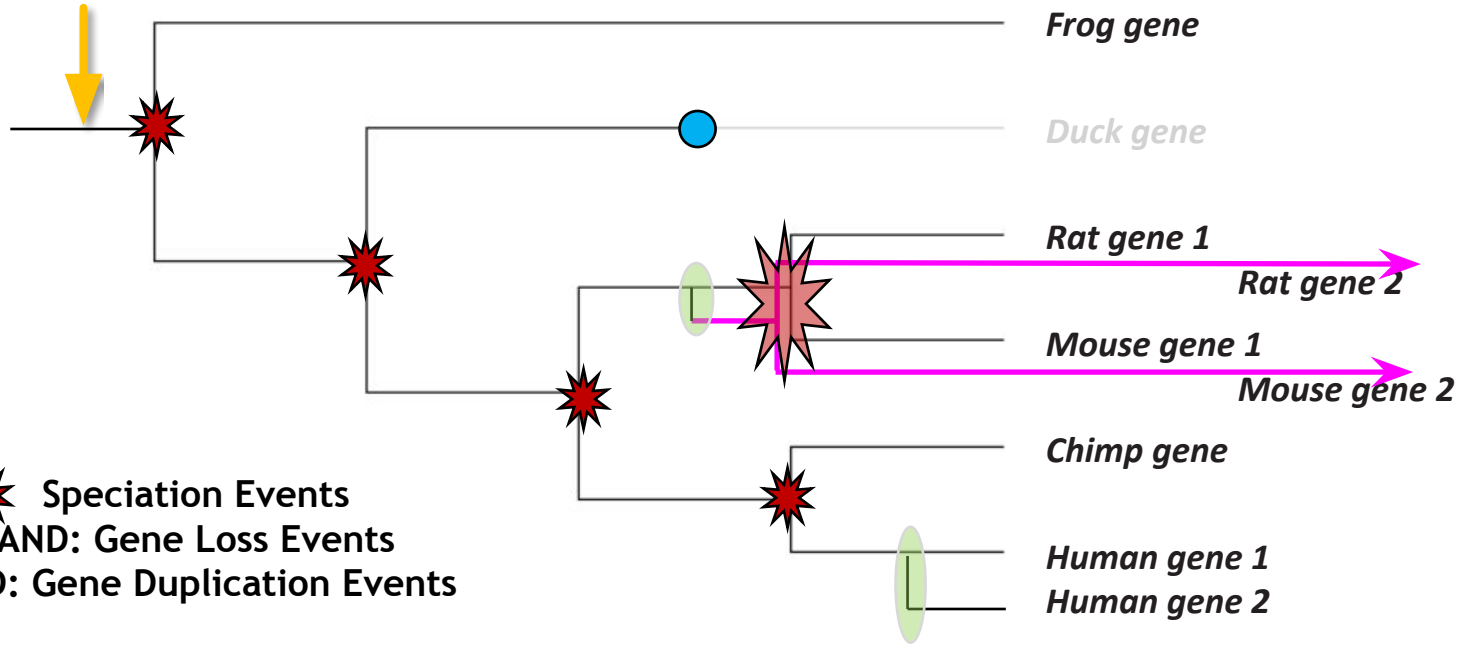


Speciation Events
AND: Gene Loss Events
AND: Gene Duplication Events

Single-Copy Orthologs with Gains

Evolution ≠ simple

+ fast sequence divergence

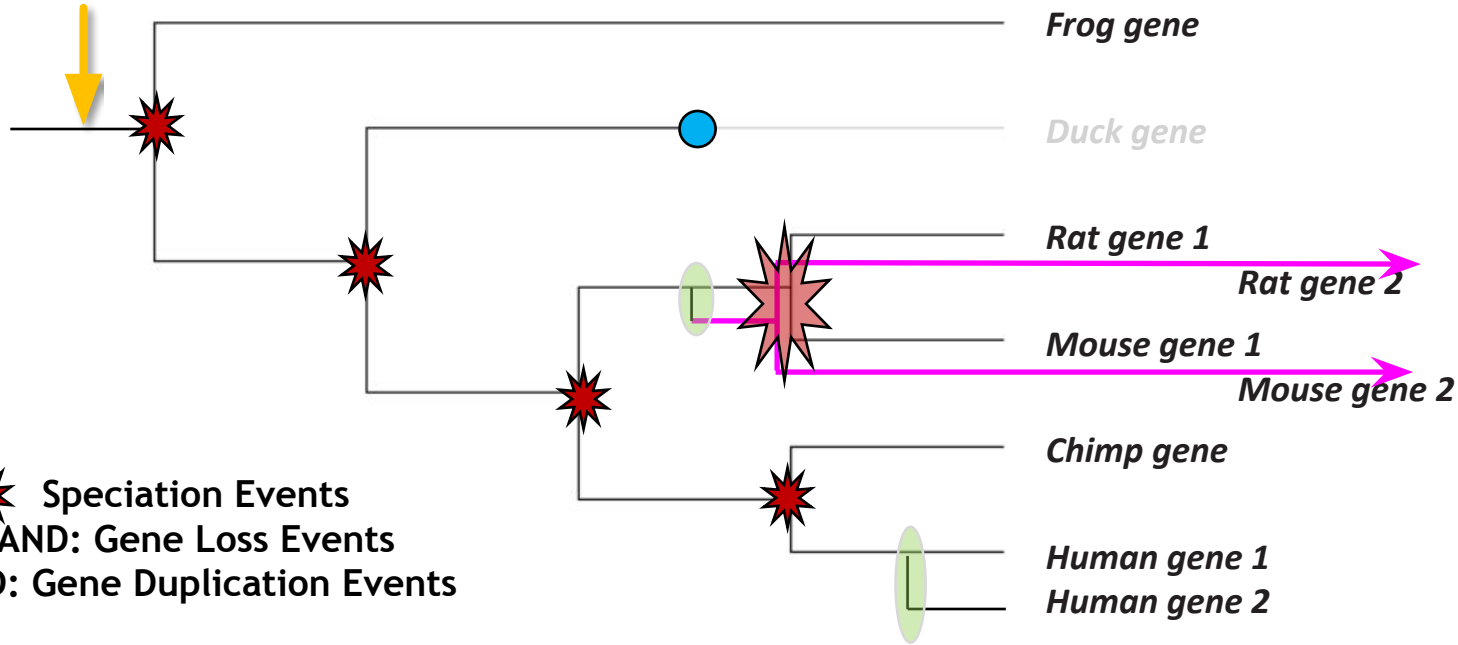


Single-Copy Orthologs with Gains

Evolution ≠ simple

Last Common Ancestor (LCA) of all 6 species

Paralogs R1+R2 M1+M2 H1+H2



Speciation Events
AND: Gene Loss Events
AND: Gene Duplication Events

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

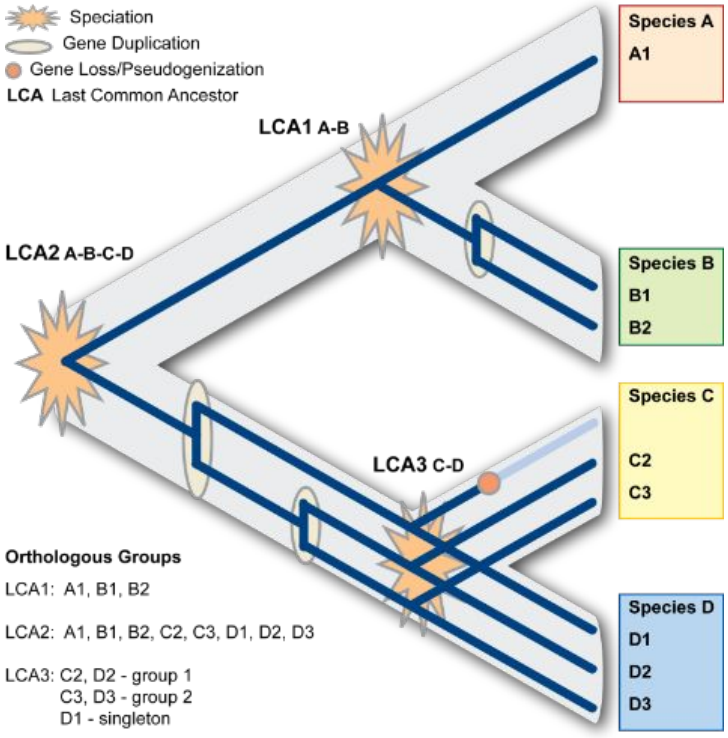
Orthology – what is it?

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



Nucleic Acids Research

OrthoDB: the hierarchical catalog of eukaryotic orthologs in 2011

Robert M. Waterhouse[1,2], Evgeny M. Zdobnov[1,2,3], Fredrik Tegenfeldt[1,2], Jia Li[1,2] and Evgenia V. Kriventseva[1,2,*]

Orthology Delineation

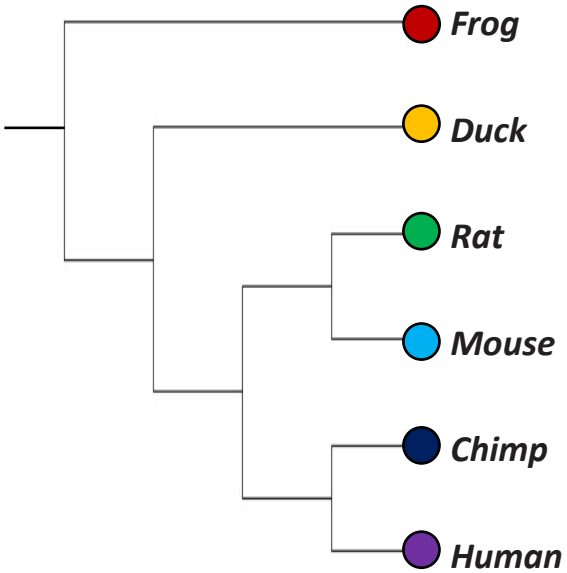
What is orthology?

How do we delineate orthologs?

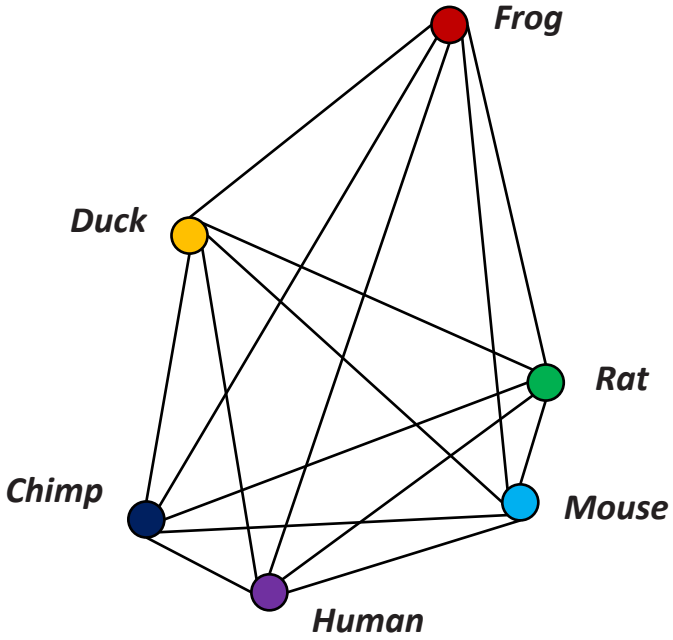
And why do we need to?
(species/gene trees/copy-number)

How do we delineate Orthology?

tree-based approaches

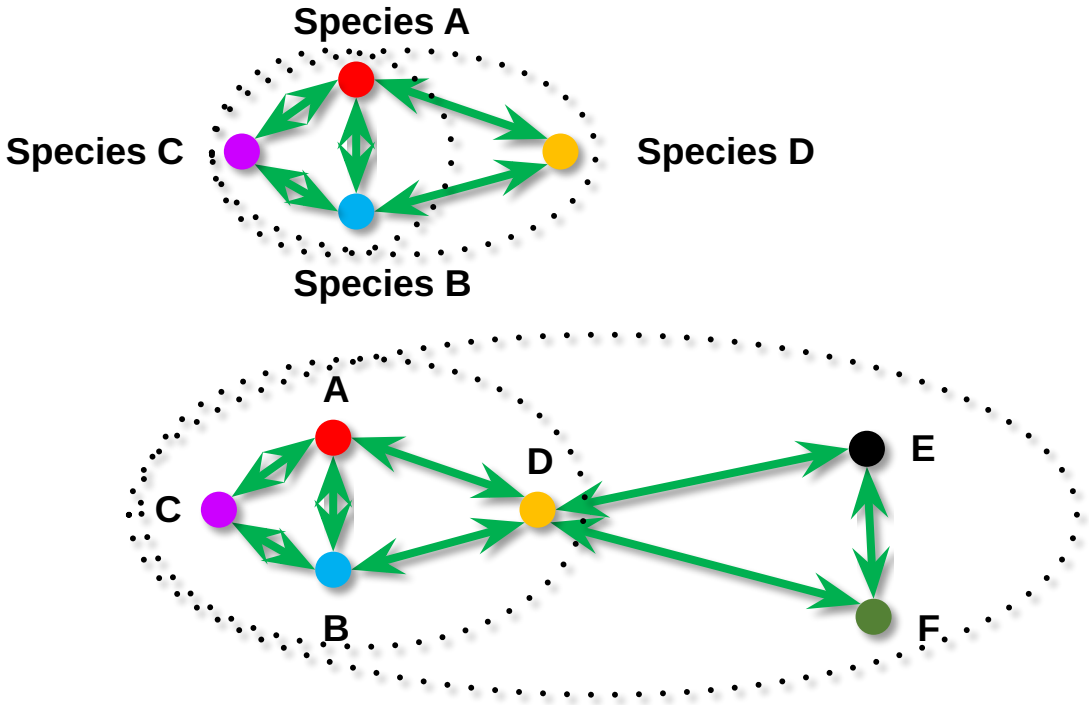


graph-based approaches

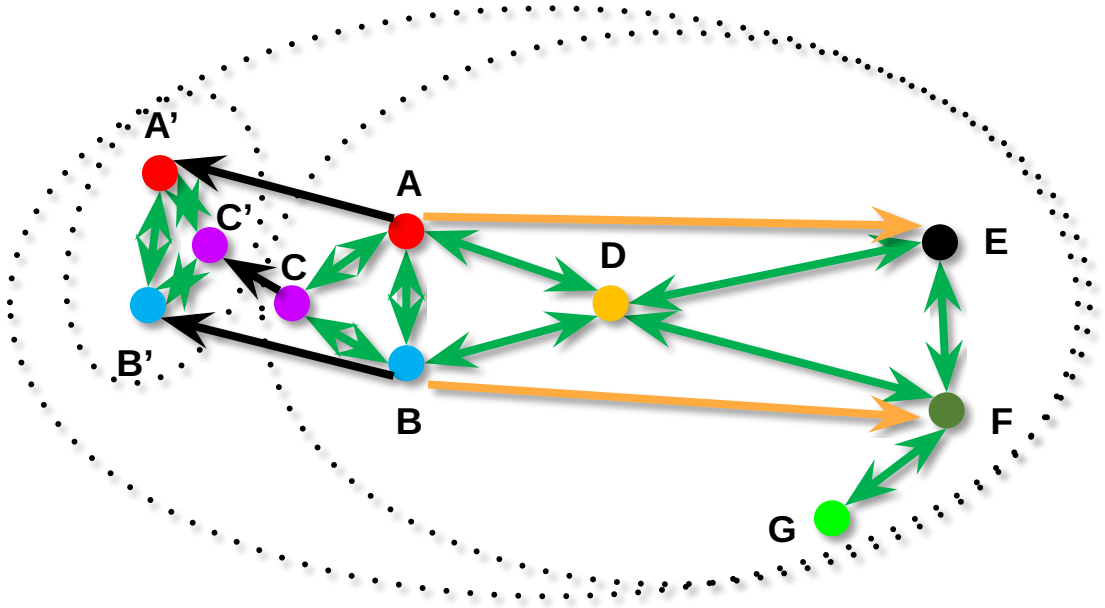


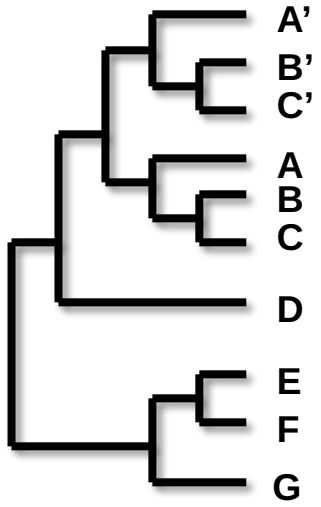
Single-Copy Orthologs

Graph-based best-reciprocal-hits



Within-clade duplications





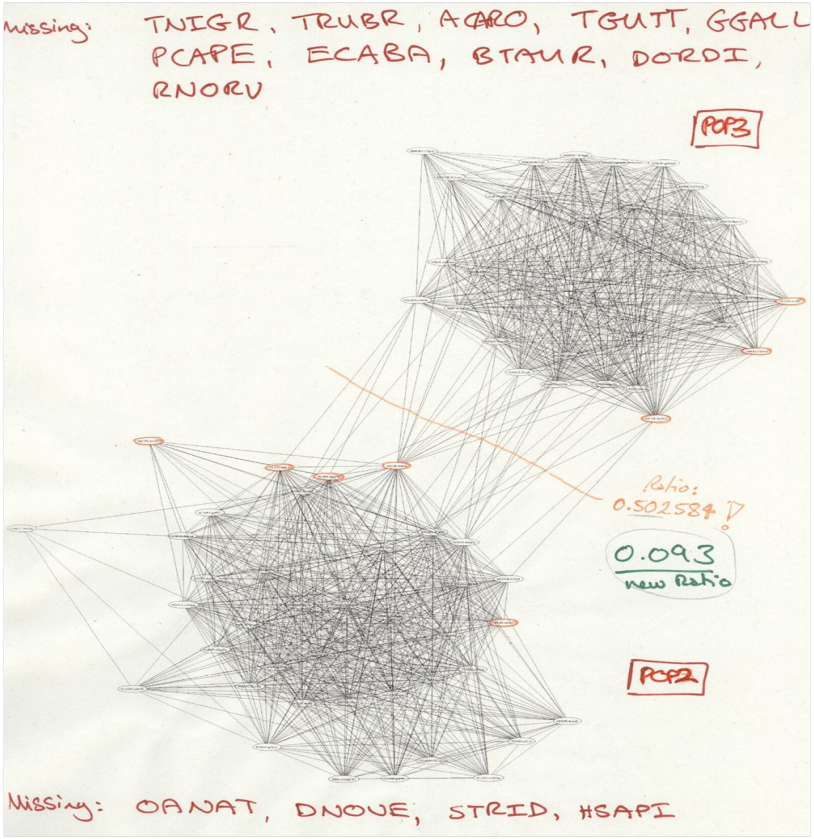
Real-world data can be messy!

Real example:

POP3 missing from 10 vertebrates

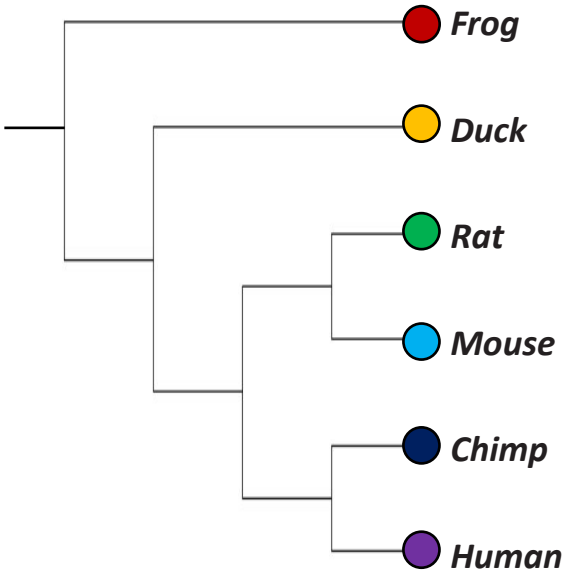
POP2 missing from 4 vertebrates

Two orthologous groups start to merge into one

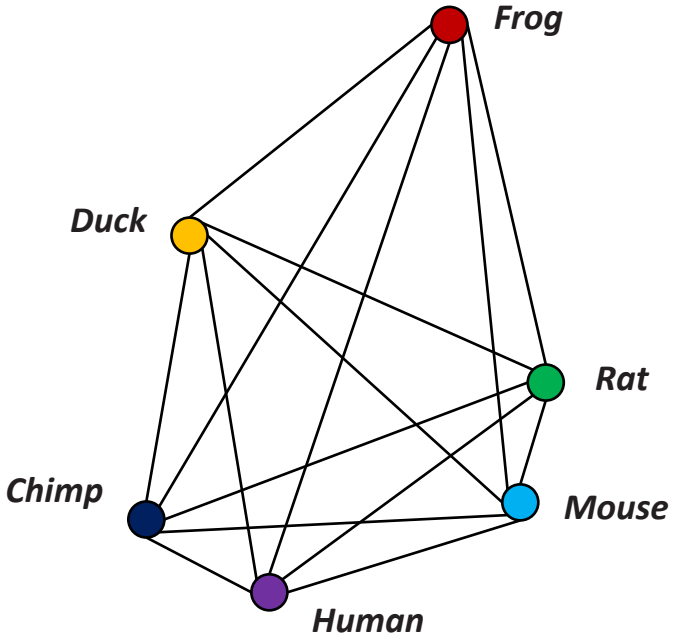


How do we delineate Orthology?

tree-based approaches



graph-based approaches



Single-Copy Orthologs

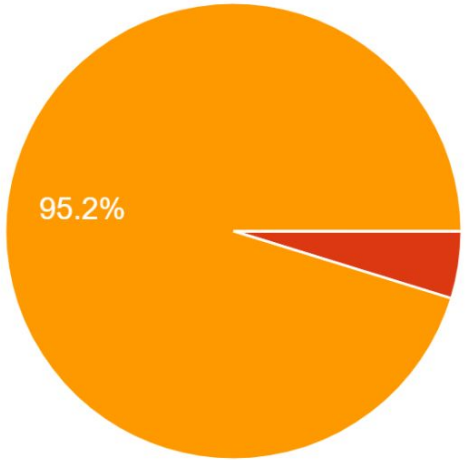
Quick Quiz

https://forms.gle/CZbu8hiLo2Qv4Bj49

SwissOrthology

Which description best describes your understanding of orthology? Orthologues are genes in different species ...

21 responses



... that evolved from an ancestral gene without duplications or losses

... that perform the same specific biological function

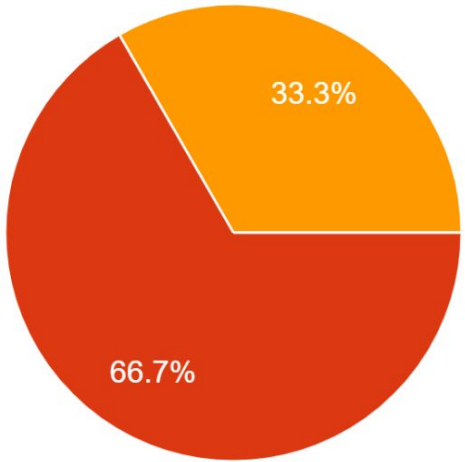
... that evolved from a single gene in the last common ancestor

... that have the highest significant sequence homology

... that produce a gene tree that matches the species phylogeny

Which description best describes your understanding of how OrthoDB delineates orthology?

21 responses



- Gene trees are reconciled with the known species tree to define speciations and duplications
- Best-reciprocal-hits determine how genes are progressively added to form orthologous groups
- The full all-against-all best-reciprocal-hit graph is progressively split to define groups of orthologues

Orthology Delineation

What is orthology?

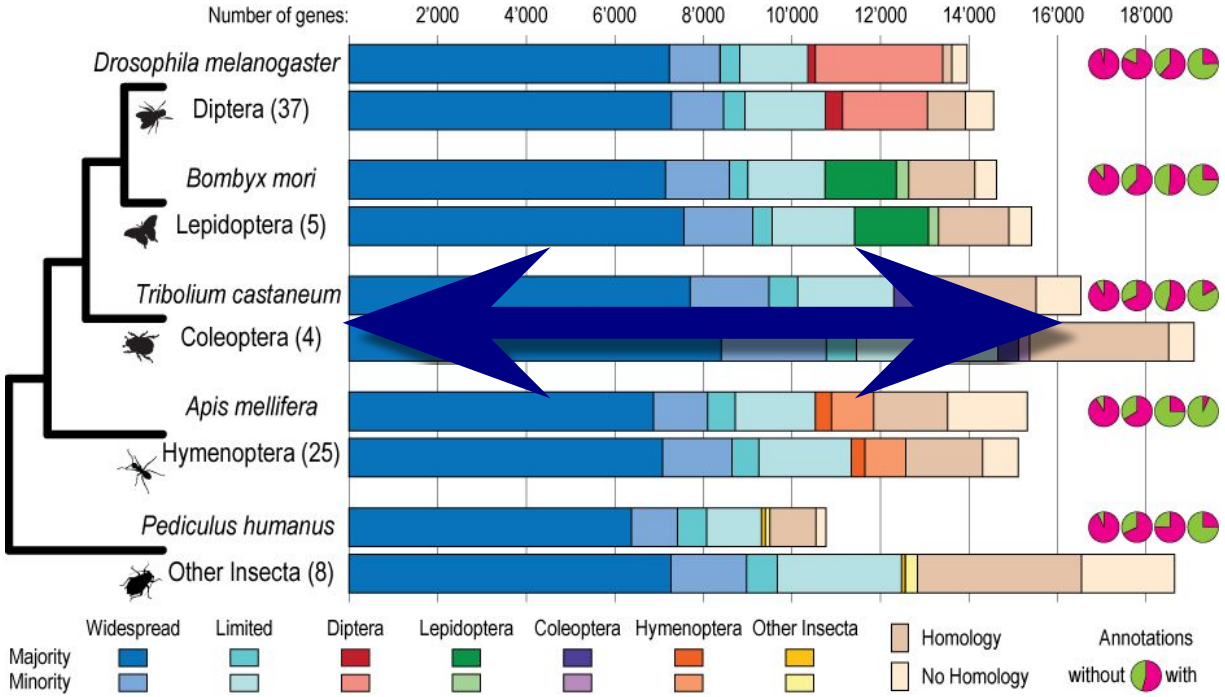
How do we delineate orthologs?

And why do we need to?
(species/gene trees/copy-number)

Orthology – why do we need it?

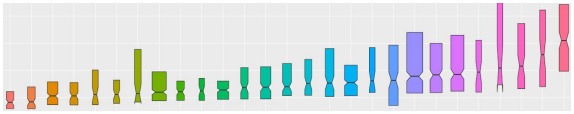
1) Tracing the **Evolutionary Histories** of all genes in extant species

2) Building **Hypotheses on Gene Function** informed by evolution

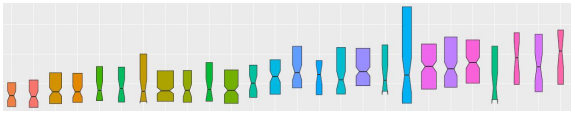


CompGeno: gene family evolutionary dynamics

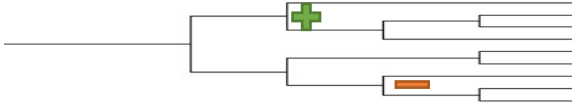
Protein Sequence Divergence



DNA Selection/Constraint



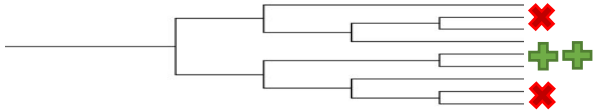
Gene Gain/Loss Rates



Intron Gain/Loss Rates

Stop-Codon Readthrough

Copy-Number/Universality



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

Orthology ≠ Function … BUT …

By tracing the **Evolutionary Histories** of all genes in extant species

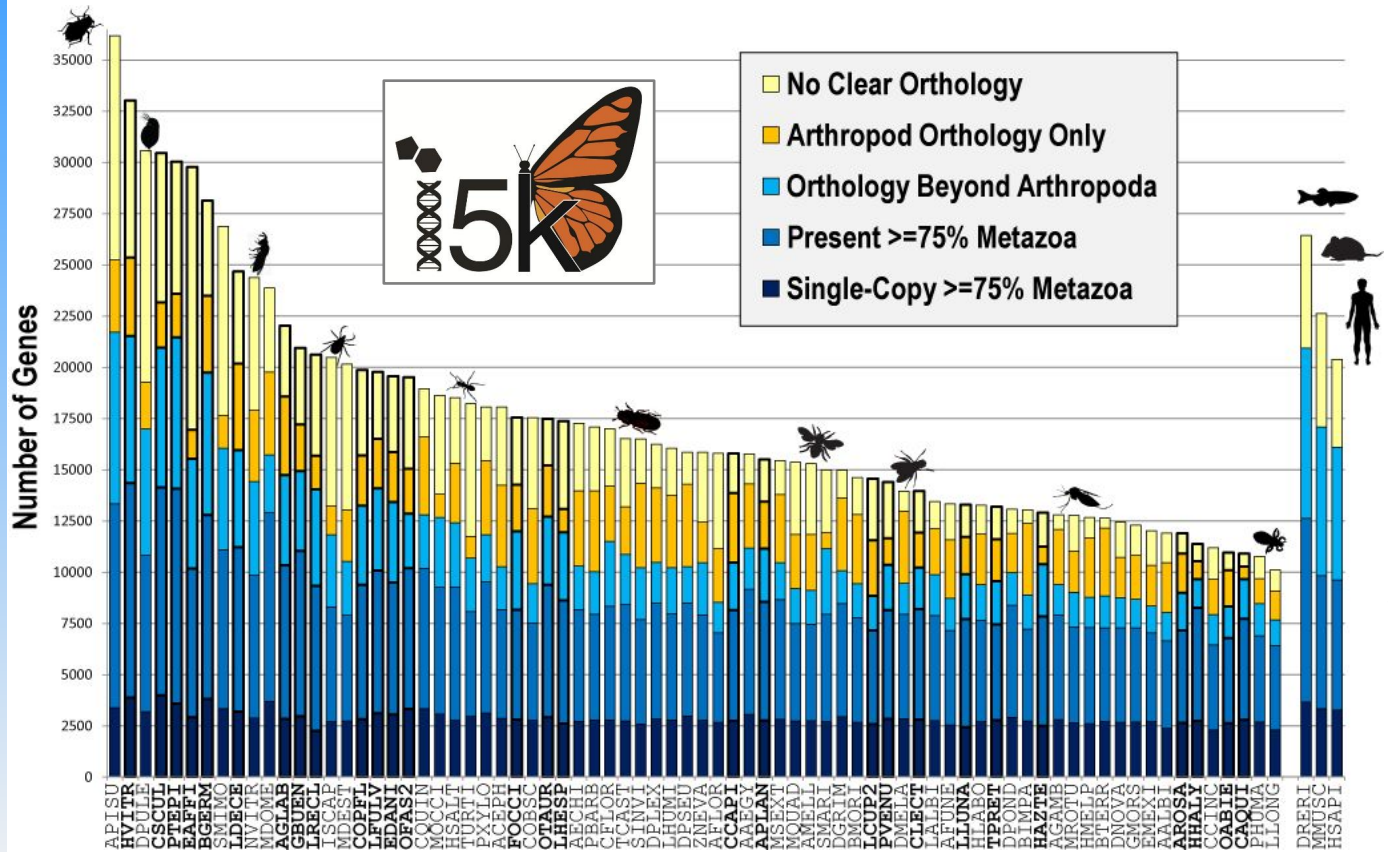
We can build **Hypotheses on Gene Function** informed by evolution

"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."

Annu. Rev. Genet.
2005. 39:309–38

Evolutionary histories: classes



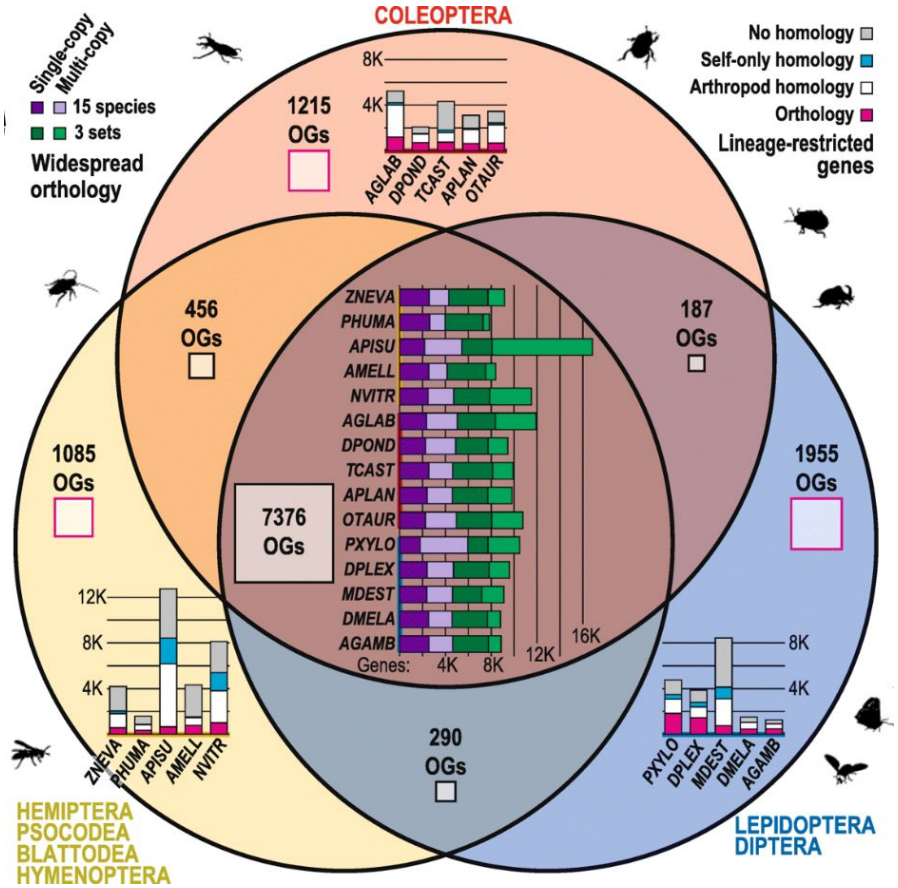
Unique

Variable

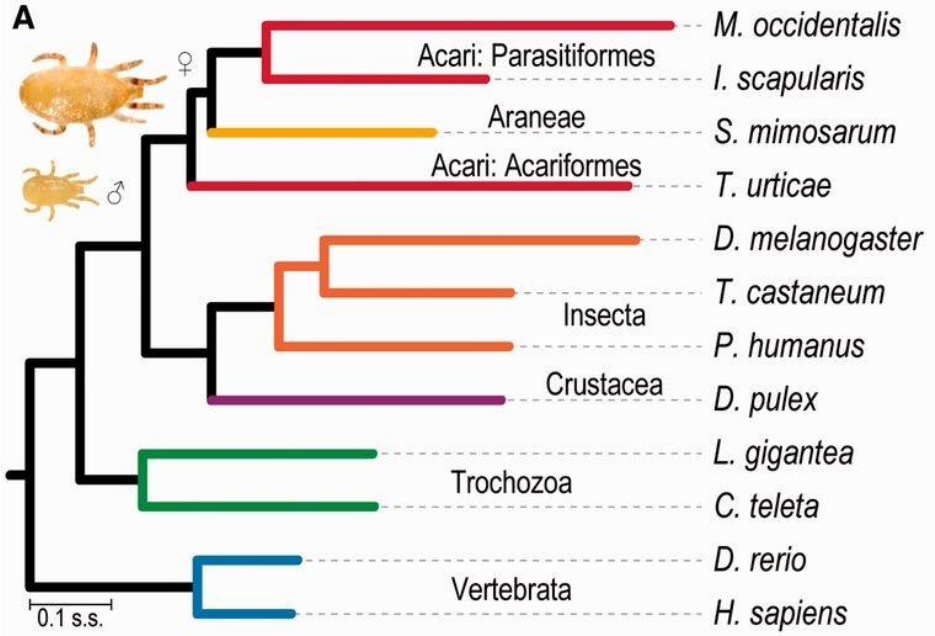
Common

Evolutionary histories: classes

Clade-specific
&
variable-count
orthologues



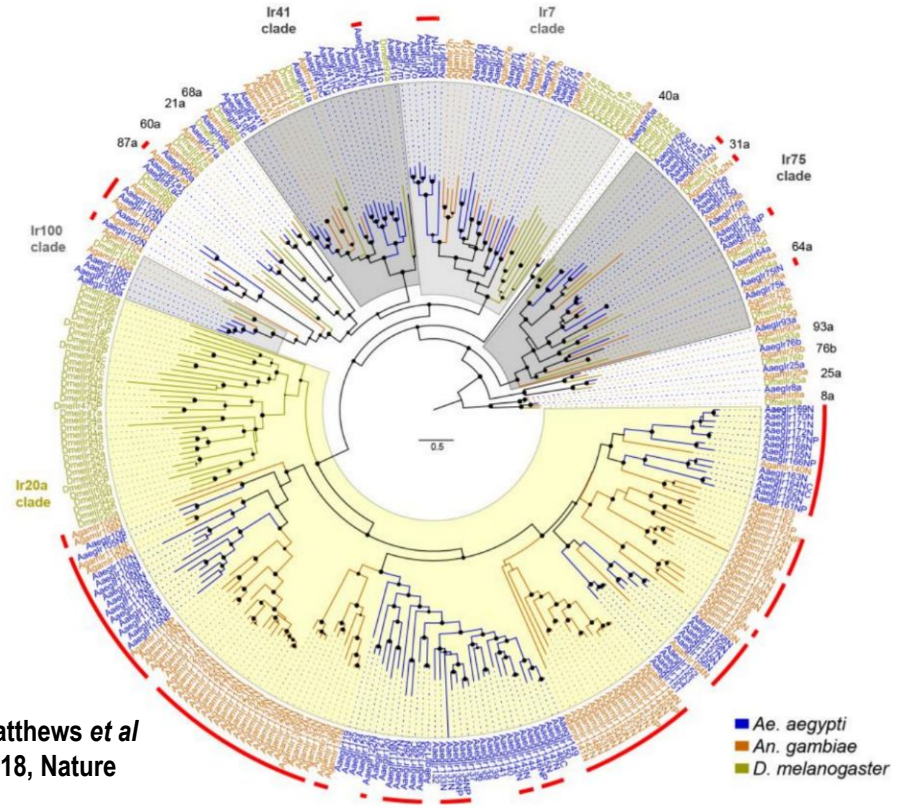
Species Tree Estimation



Phylogenomics with single-copy orthologues

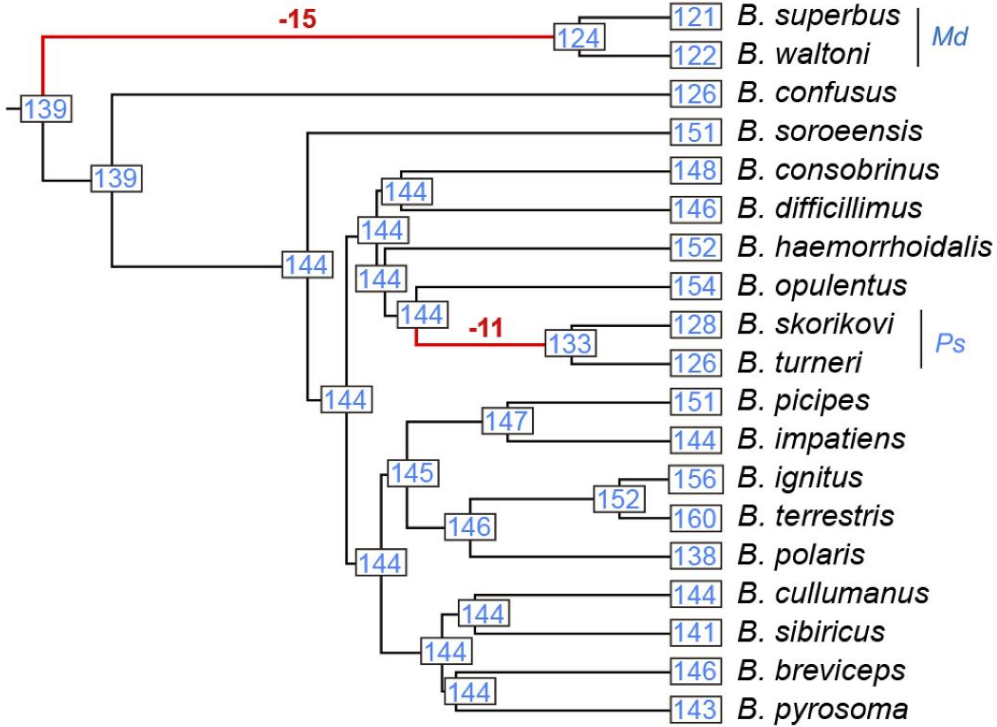
Gene Family Tree Building

All Ionotropic Receptors OrthoGroups in three species: conserved and dynamic IR OGs



Matthews *et al*
2018, Nature

Ancestral Copy-Number Reconstruction

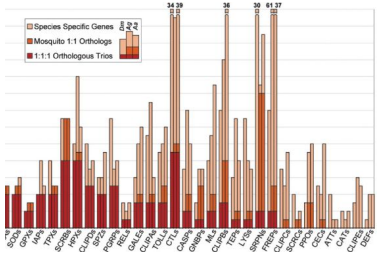


Bumblebee Odorant Receptors : two major gene loss events

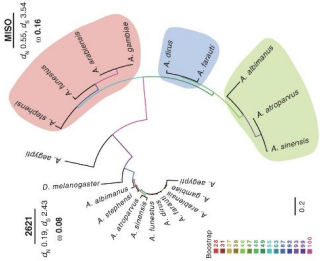
Dynamically evolving families

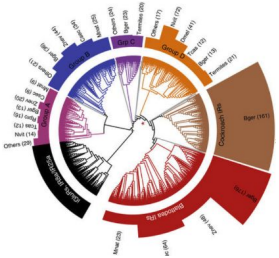
Many of the most biologically interesting genes and gene families show highly dynamic evolutionary histories

IMMUNITY



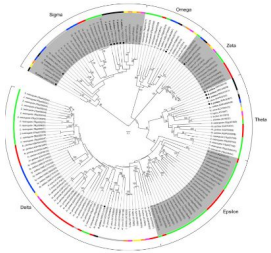
REPRODUCTION





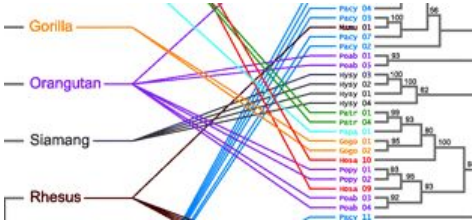
CHEMOSENSATION

DETOXIFICATION

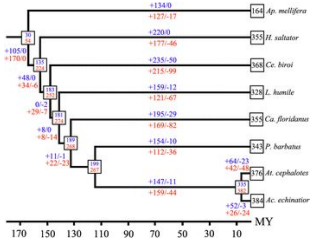


Inferring gene evolutionary histories

Gene-tree-species-tree reconciliation



Gene ancestral state reconstruction



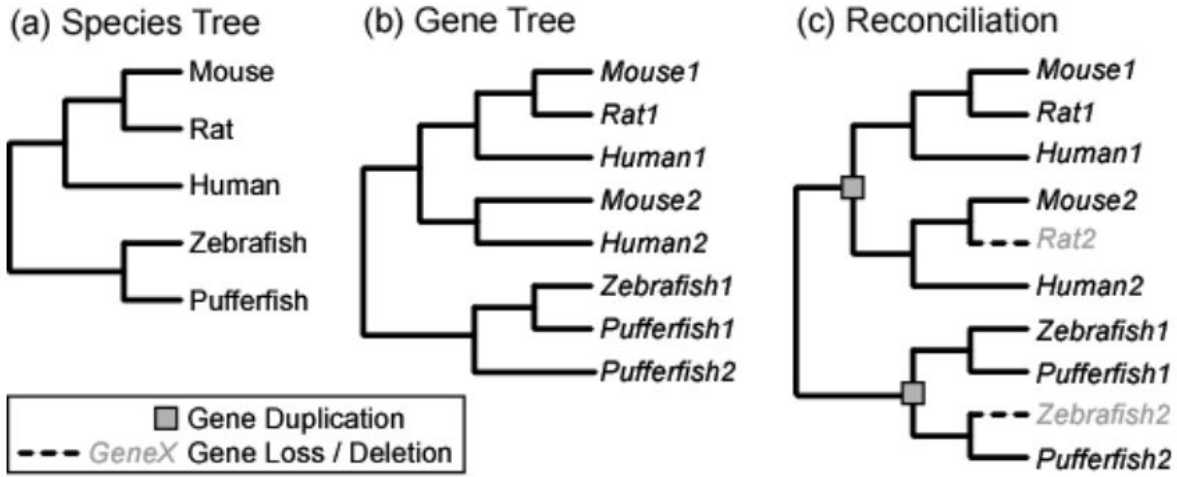
INPUTS

- 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

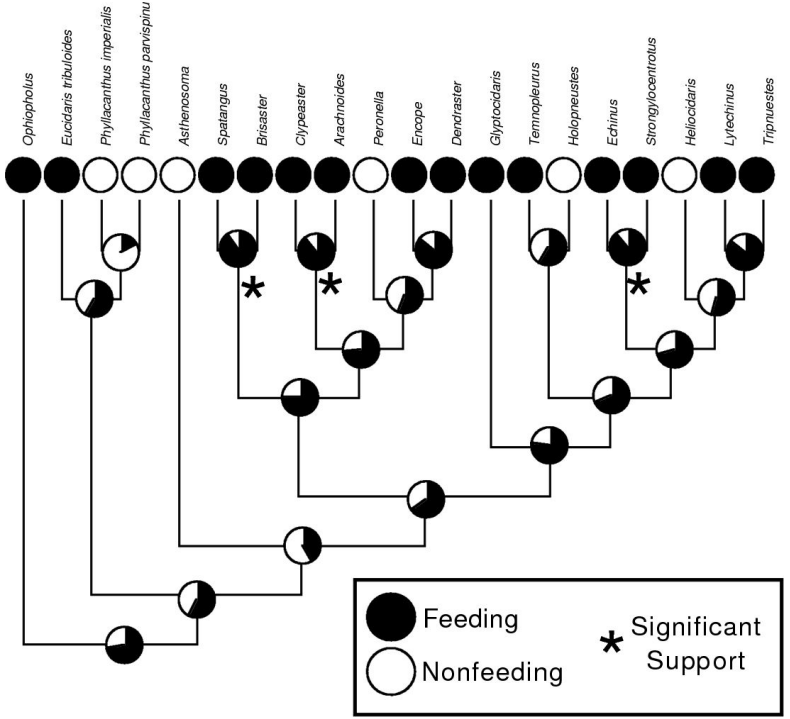
- Compute the minimum "cost" resolution
- Impacted by assumptions on costs for duplication and losses

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 events 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?

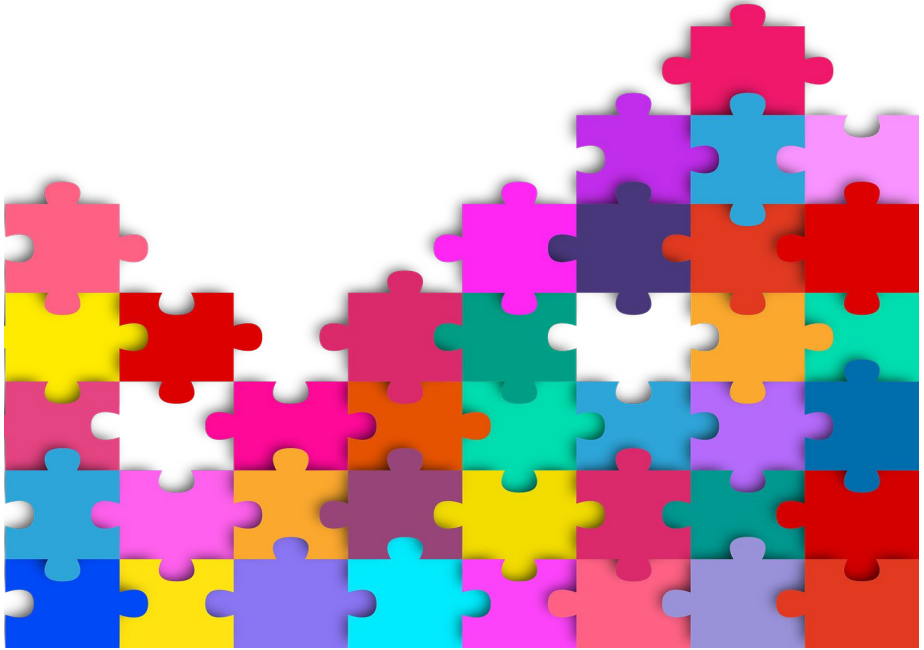
And why do we need to?
(species/gene trees/copy-number)

Orthology Delineation

What could possibly go wrong?

Orthology Delineation

What could possibly go wrong?



Assessing genomics data quality: BUSCO

What is BUSCO ?

How does BUSCO work ?

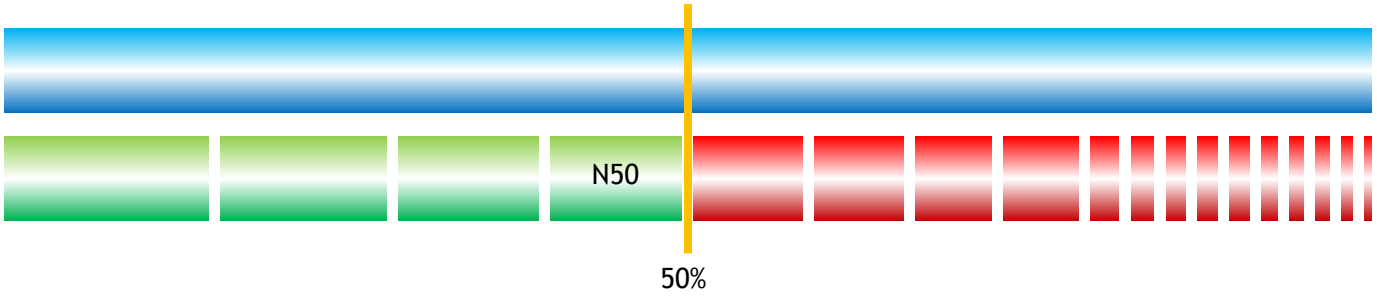
Benchmarking Universal Single-Copy Orthologues

How can we gauge the quality of these resources?

1) Does the assembly size match the expected genome size?

2) How fragmented is the assembly?

Assembly contig or scaffold N50 size:
half the assembly is found on contigs/scaffolds of length N50 or greater



3) How ‘gappy’ is the assembly?

4) Does the assembly contain all the genes it is expected to?
How much of a multi-life-stage transcriptome maps back to the assembly?
How many of the ‘expected’ genes are actually in the assembly?

BUSCO: looking for widespread & unique genes

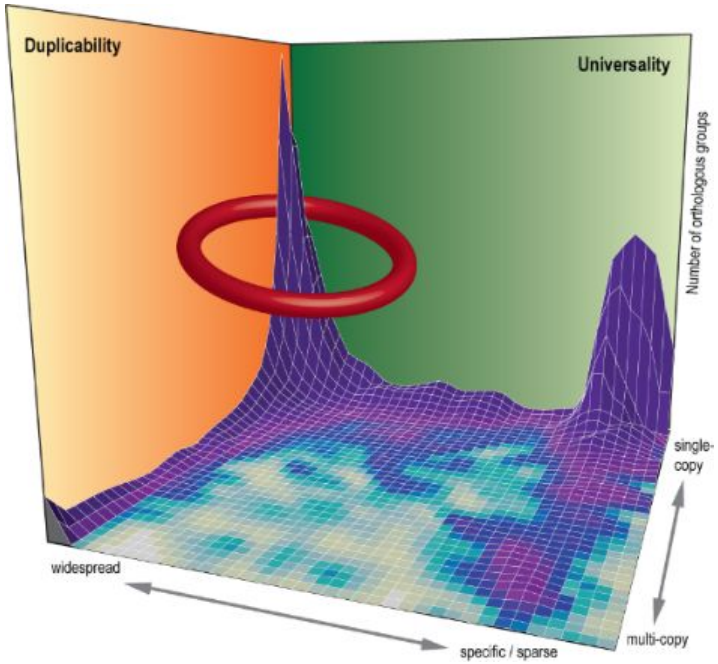
Ortho-Groups with genes
found in the majority of species
as single-copy orthologues

Evolutionary Expectation - 90 x 90
for them to be found in any
newly-sequenced genome

Implemented Assessments
Gene Content Completeness
genome assemblies
annotated gene sets
assembled transcriptomes

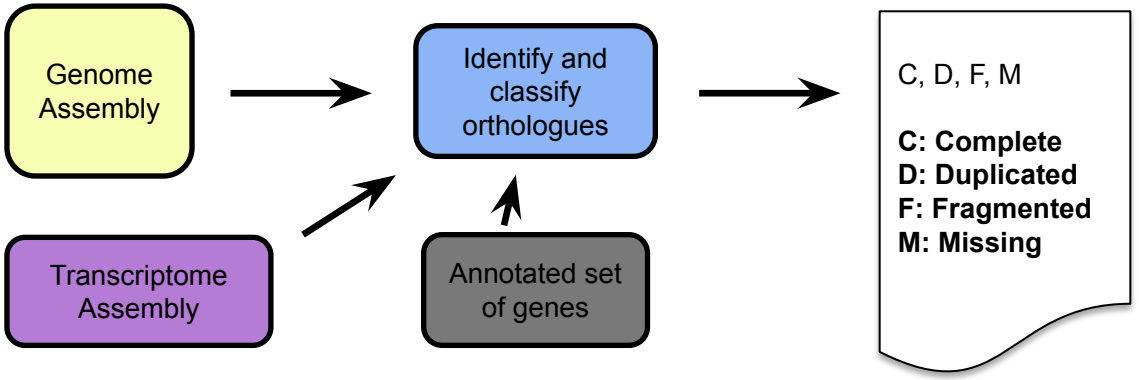
Bonus Features
genes for phylogenomics
gene predictor training

Orthology Landscape



http://busco.ezlab.org

BUSCO completeness assessments



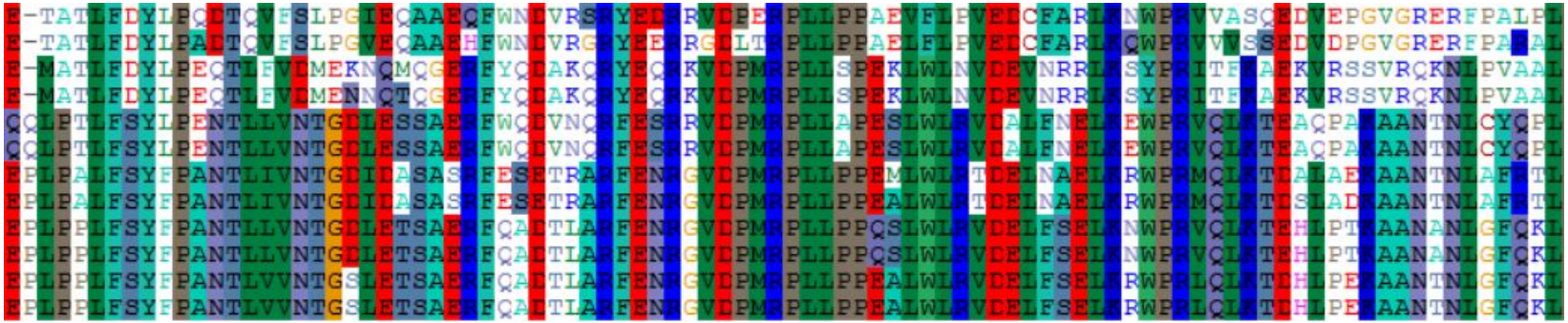
Building BUSCO lineage datasets

Species filtering to select best representatives from each clade

- Avoiding biasing the alignments with closely-related species

Building BUSCO lineage datasets

1) Multiple protein sequence alignments for each orthologous group



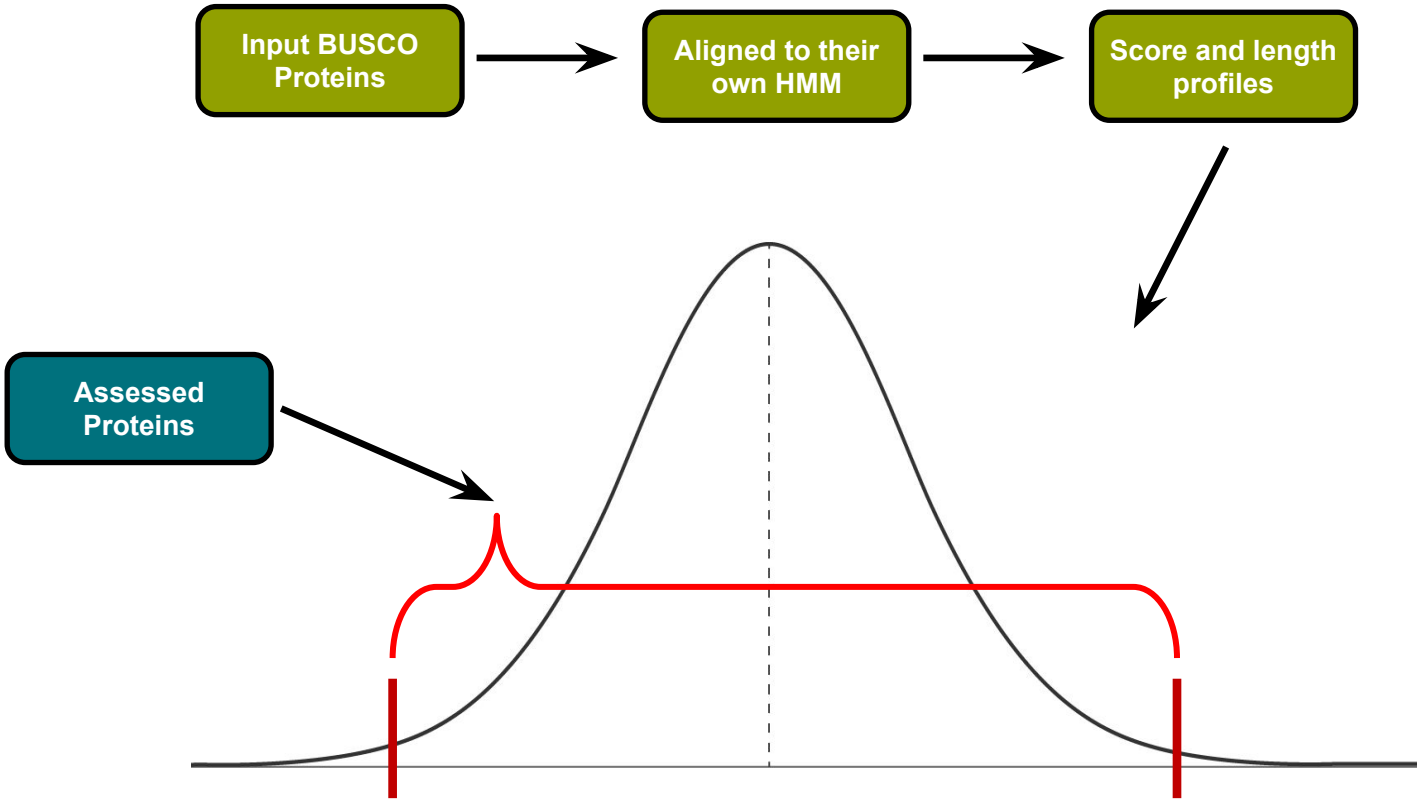
2) HMM profiles from alignments for searching protein sequences

3) Consensus sequences for searching genome assemblies

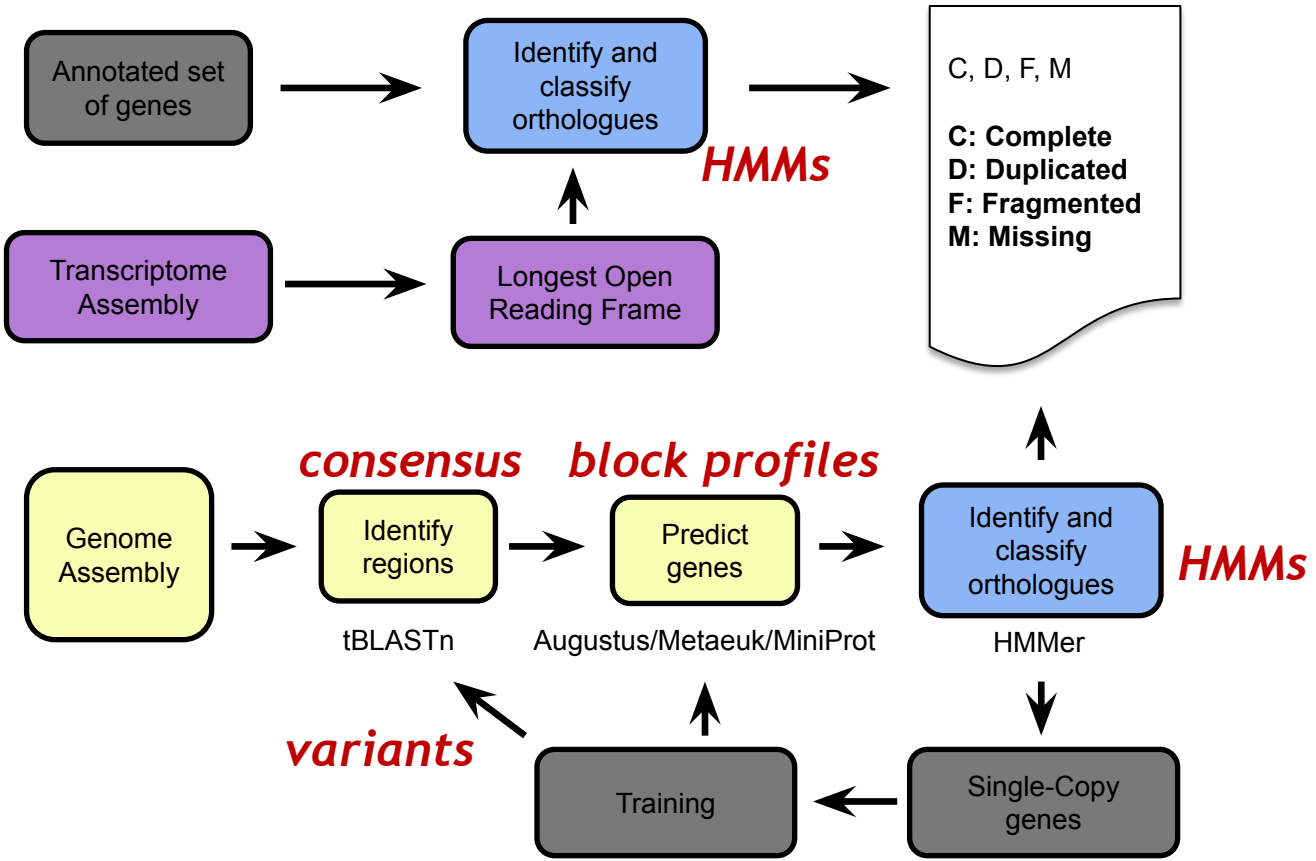
4) Consensus sequence variants for searching genome assemblies

5) Augustus block profiles for predicting gene models

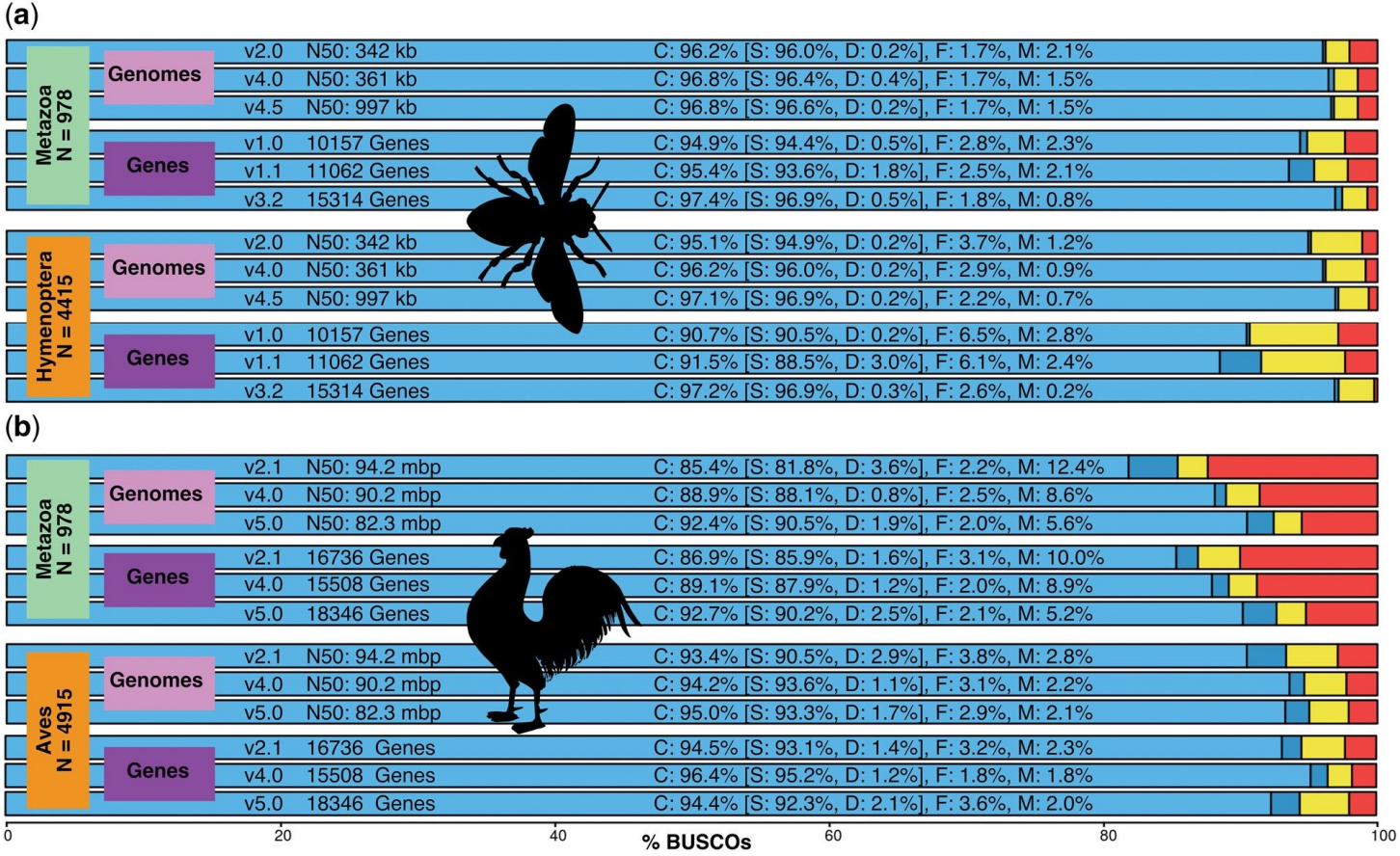
Identify and classify orthologues – HOW ?



BUSCO completeness assessments



Enabling standardised quality assessments



Complete:
C, blues

Complete & Single:
S, light blue

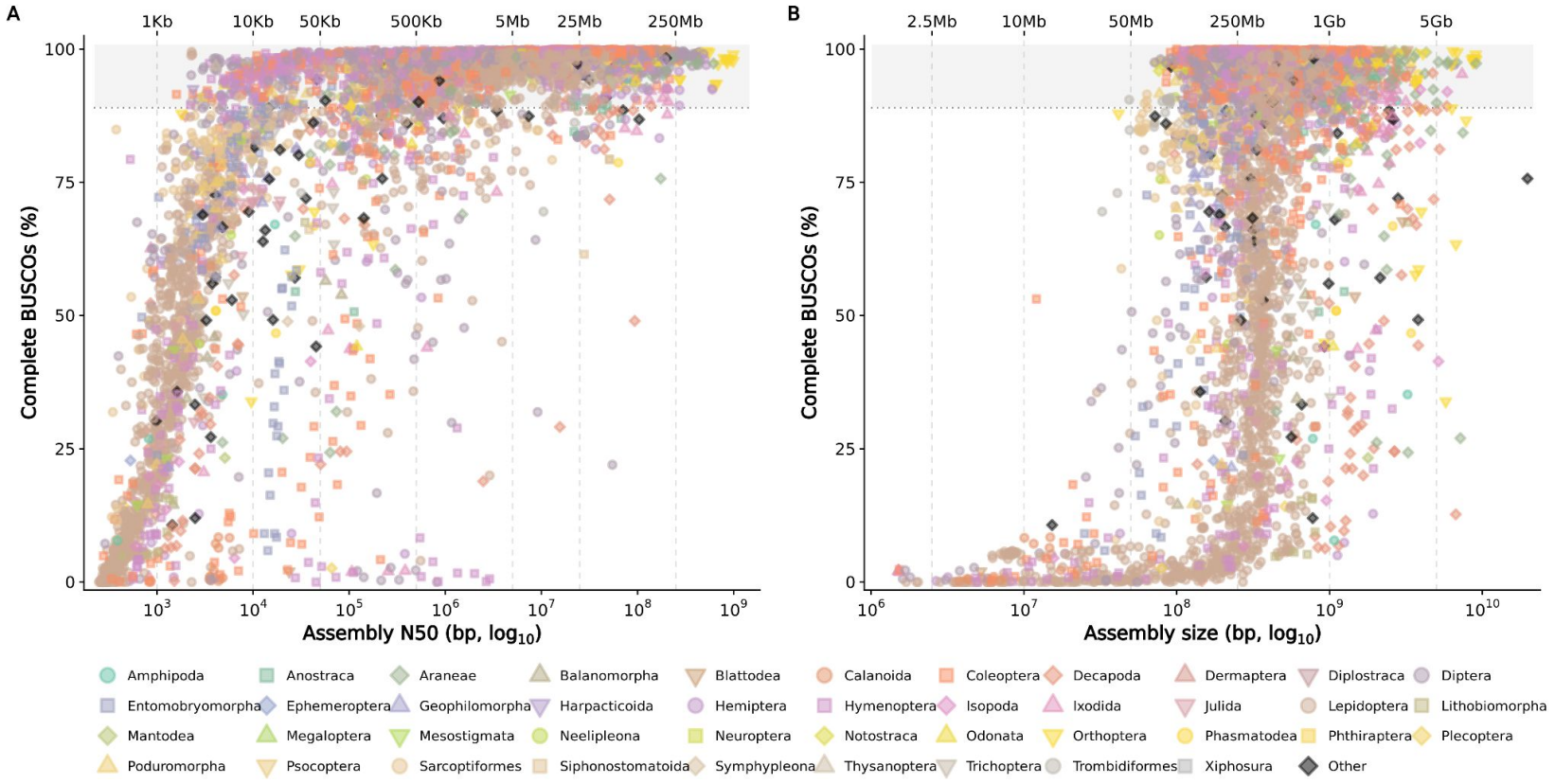
Complete Duplicated:
D, dark blue

Fragmented:
F, yellow

Missing:
M, red

For **Genomes** (pink) and **Genes** (purple)

Garbage in, Garbage out



Orthology Delineation

What is orthology?

How do we delineate orthologs?

And why do we need to?
(species/gene trees/copy-number)

What is BUSCO?

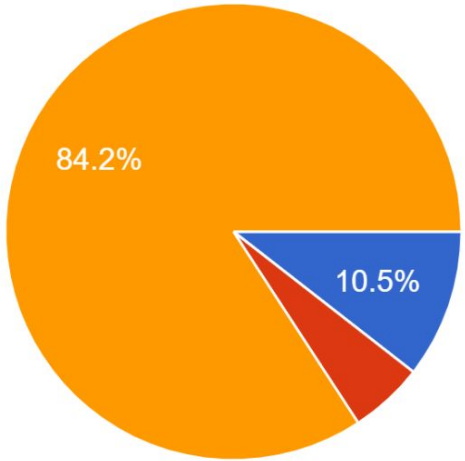
How does BUSCO work?

Quick Quiz

https://forms.gle/uVHA4WoYRCb8W2fS6

Which description best describes your understanding of what BUSCO aims to achieve?

19 responses



- To assess the sequencing quality of genomic data including genomes, gene sets, and transcriptomes
- To identify and score all highly conserved genes in a newly sequenced and annotated genome or transcriptome
- To estimate completeness of genomic data including genomes, gene sets, and transcriptomes in terms of expected gene content

The European Reference Genome Atlas ERGA

EUROPEAN REFERENCE GENOME ATLAS

Coordinating the sequencing and analysis of European eukaryote species

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

ERGA Knowledge Hub
https://knowledge.erga-biodiversity.eu/

A diverse group of researchers in Europe
Join the community: www.erga-biodiversity.eu