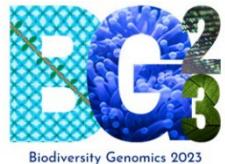


Biodiversity Genomics Academy 2023



BUSCO - from QC to gene prediction and phylogenomics



September 20th 2023

Robert M. Waterhouse

SIB Swiss Institute of Bioinformatics, Switzerland



Swiss Institute of
Bioinformatics

✉ robert.waterhouse@gmail.com

Ⓜ️ Ⓛ @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



Swiss Institute of
Bioinformatics



UNIL | Université de Lausanne



Imperial College
London

wellcome trust



Goals for Today's Workshop

- Learn how to use BUSCO to assess genomics data
- Understand what BUSCO is attempting to achieve
- Learn how the BUSCO datasets are constructed
- Understand how OrthoDB delineates orthologous groups that are used as the main input for BUSCO

OrthoDB
BUSCO



Quick Quiz

The screenshot shows a Google Forms quiz titled "Biodiversity Bioinformatics: Quick Quiz". The quiz has one question: "How familiar are you with BUSCO, the Benchmarking Universal Single-Copy Orthologues?". The response scale ranges from 1 to 5, with 1 being "I have never heard about BUSCO" and 5 being "I have used the BUSCO assessment tools a lot in my research". The quiz has 29 responses.

SwissOrthology

Biodiversity Bioinformatics: Quick Quiz

robert.waterhouse@gmail.com Switch accounts

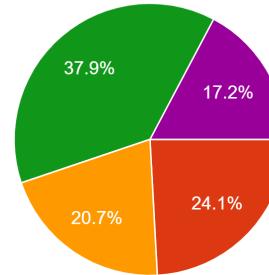
Not shared

* Indicates required question

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

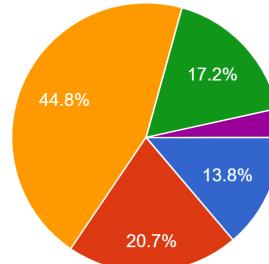
BUSCO

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



- I have never heard about BUSCO
- I have heard about BUSCO, but never visited the website
- I have visited the BUSCO website, but not really used the tool much
- I have used the BUSCO assessment tools a bit in my research
- I have used the BUSCO assessment tools a lot in my research

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



- I have never heard about OrthoDB
- I have heard about OrthoDB, but never visited the website
- I have visited the OrthoDB website, but not really used it much
- I have used OrthoDB data a bit in my research
- I have used OrthoDB data a lot in my research



Quick Quiz

The screenshot shows a Google Forms quiz titled "Biodiversity Bioinformatics: Quick Quiz". The quiz has one question: "How familiar are you with BUSCO, the Benchmarking Universal Single-Copy Orthologues?". The response scale ranges from 1 to 5, with 1 being "I have never heard about BUSCO" and 5 being "I have used the BUSCO assessment tools a lot in my research". The quiz is shared with "robert.waterhouse@gmail.com" and is not shared publicly. It has 23 responses.

robert.waterhouse@gmail.com Switch accounts

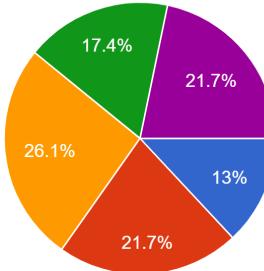
Not shared

* Indicates required question

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

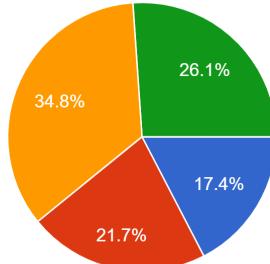
BUSCO

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



- I have never heard about BUSCO
- I have heard about BUSCO, but never visited the website
- I have visited the BUSCO website, but not really used the tool much
- I have used the BUSCO assessment tools a bit in my research
- I have used the BUSCO assessment tools a lot in my research

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



- I have never heard about OrthoDB
- I have heard about OrthoDB, but never visited the website
- I have visited the OrthoDB website, but not really used it much
- I have used OrthoDB data a bit in my research
- I have used OrthoDB data a lot in my research



BUSCO: Benchmarking Universal Single -Copy Orthologues

*BUSCO relies on Orthology data
from OrthoDB*

*Therefore we will start
with an introduction to Orthology
and OrthoDB*



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 Genomics¹

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 Genomics¹

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 Genomics¹

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. Pearson¹

¹University of Virginia School of Medicine, Charlottesville, VA

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

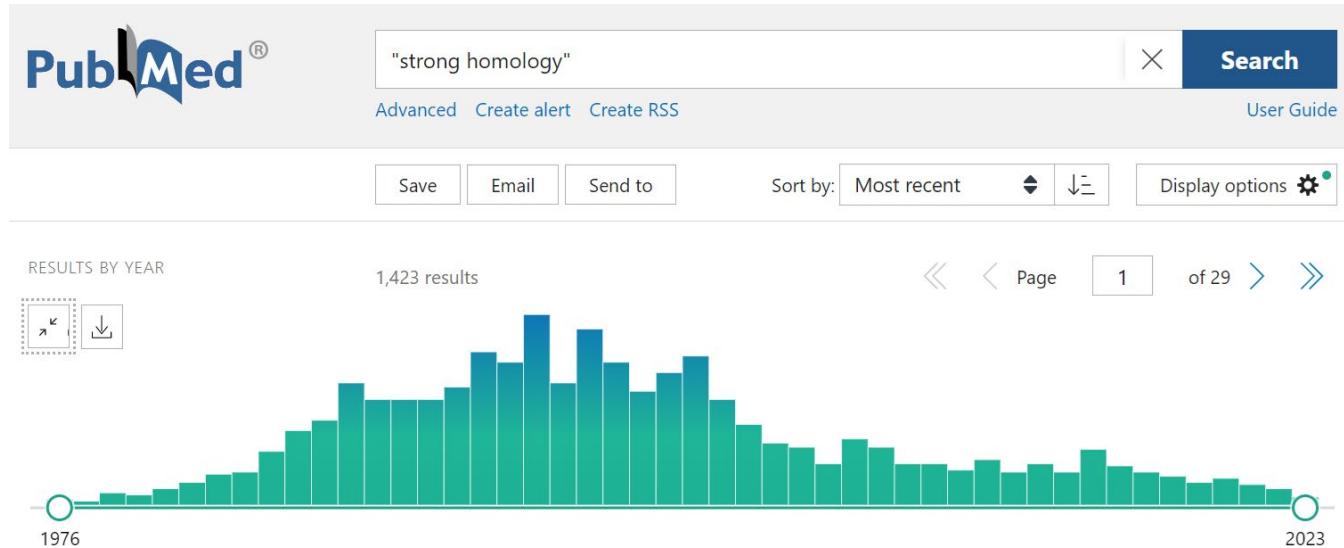
But this does not translate to **strong** or **weak** homology!

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?

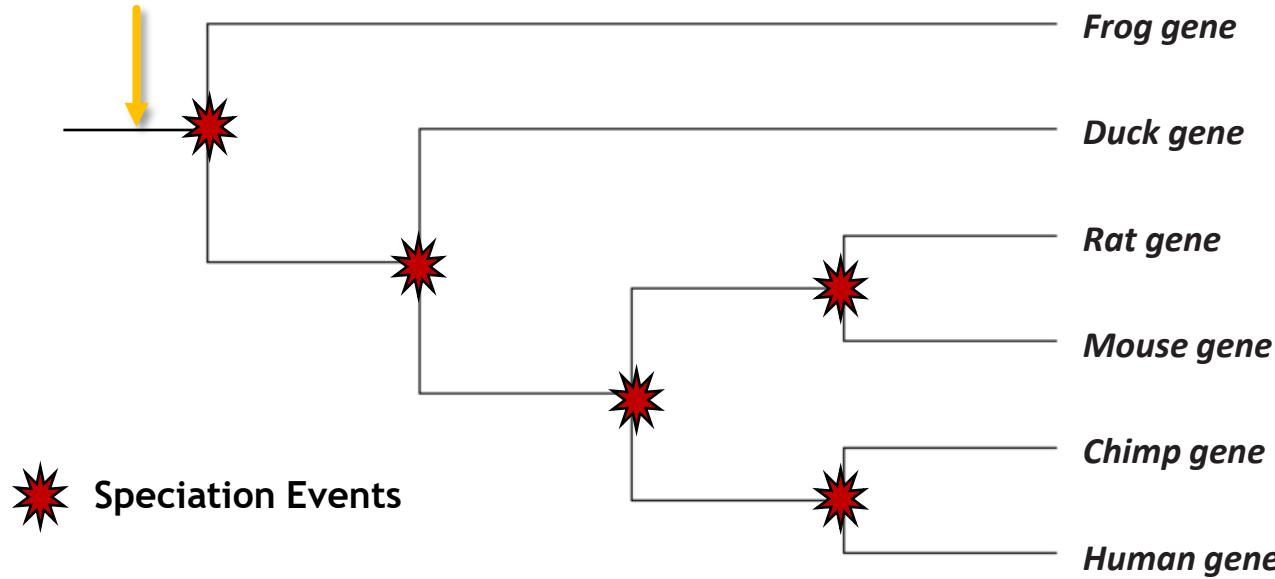
“the link between similarity and homology
is often misunderstood”



It is still worth pointing this out in 2023!

Orthology – simple scenario

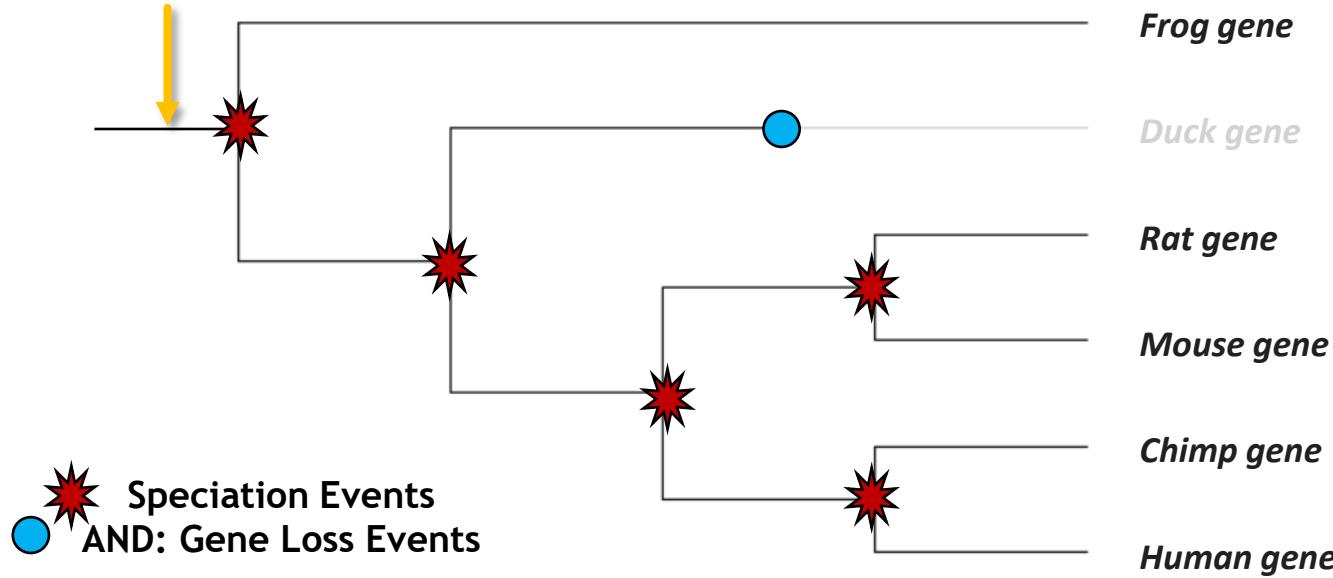
Last Common
Ancestor
(LCA) of all 6 species



Single-Copy Orthologs

Evolution ≠ simple

Last Common
Ancestor
(LCA) of all 6 species

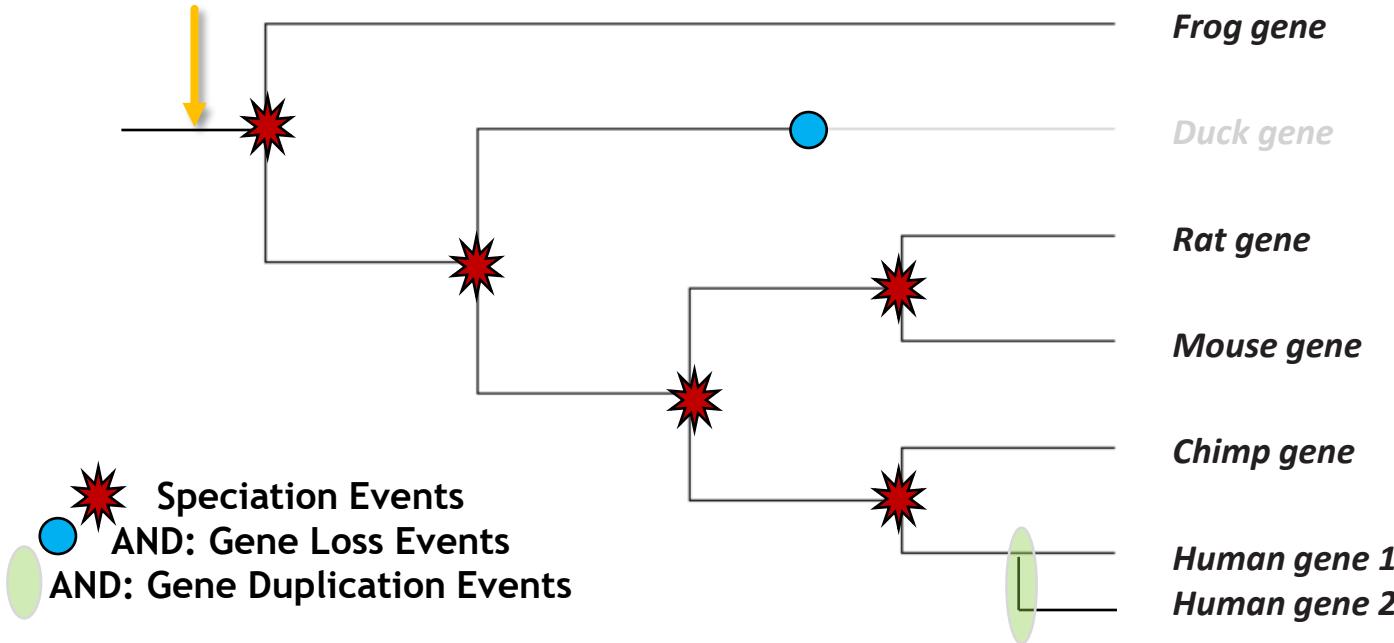


Single-Copy Orthologs with Losses

Evolution ≠ simple

Last Common
Ancestor
(LCA) of all 6 species

Human gene 1 & 2 = paralogs

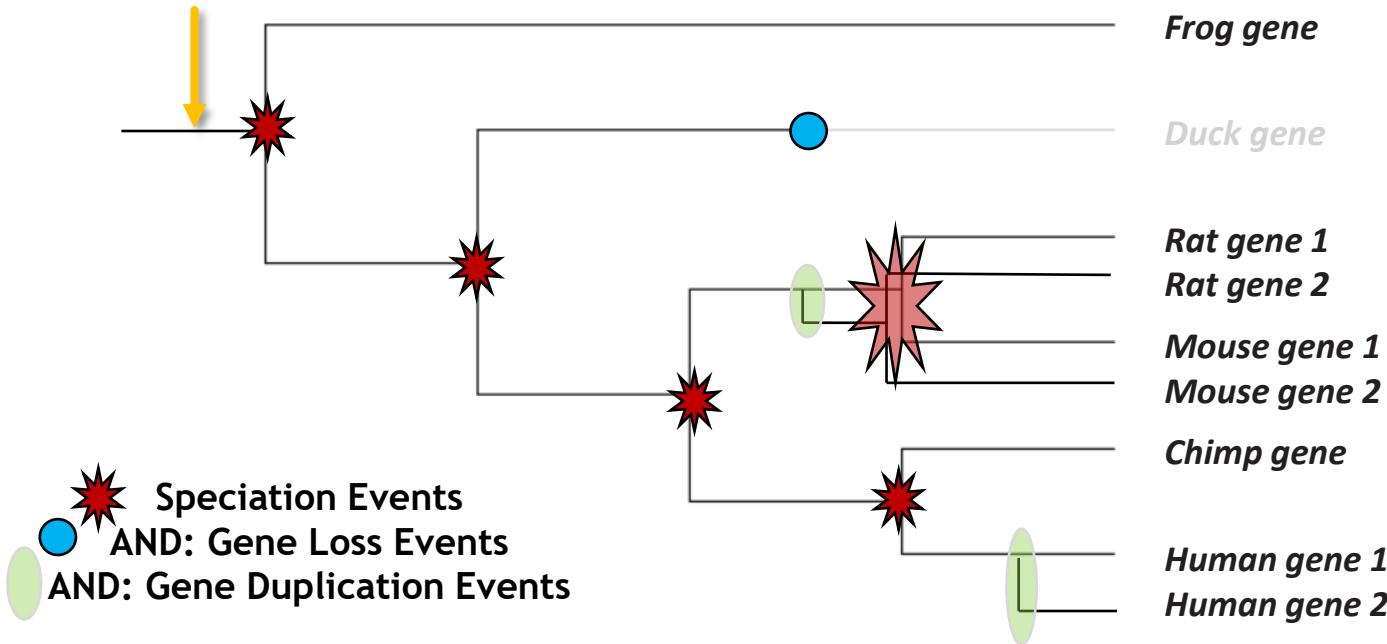


Single-Copy Orthologs with Gains

Evolution ≠ simple

Last Common
Ancestor
(LCA) of all 6 species

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

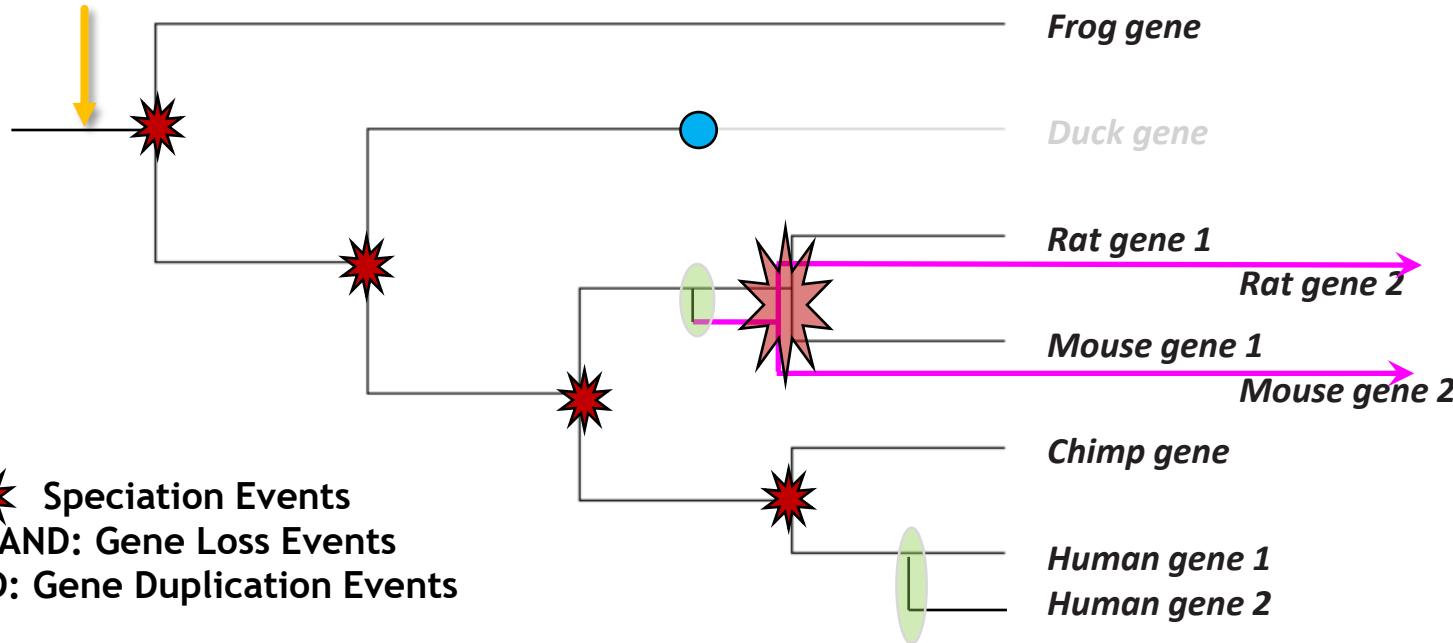


Single-Copy Orthologs with Gains

Evolution ≠ simple

+ fast sequence divergence

Last Common
Ancestor
(LCA) of all 6 species

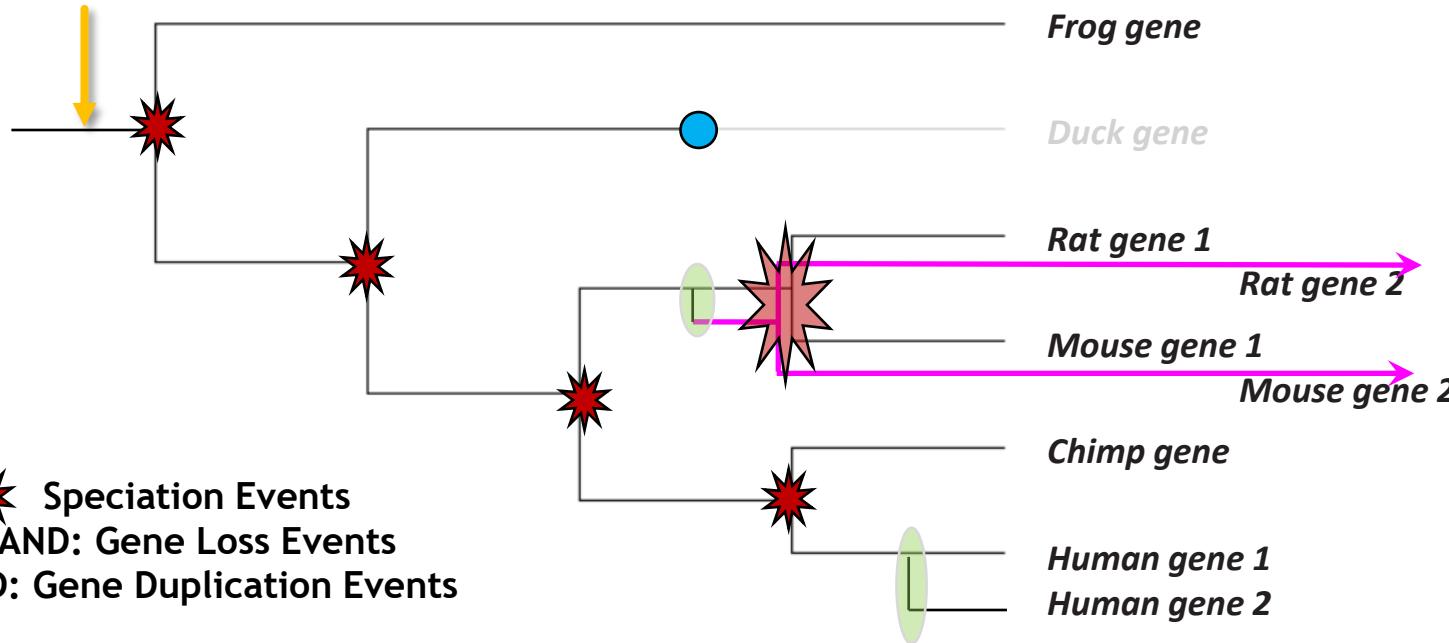


Single-Copy Orthologs with Gains

Evolution ≠ simple

Last Common
Ancestor
(LCA) of all 6 species

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



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

Orthology – what is it?

Homology

Recognizing similarities as evidence of shared ancestry

Orthology

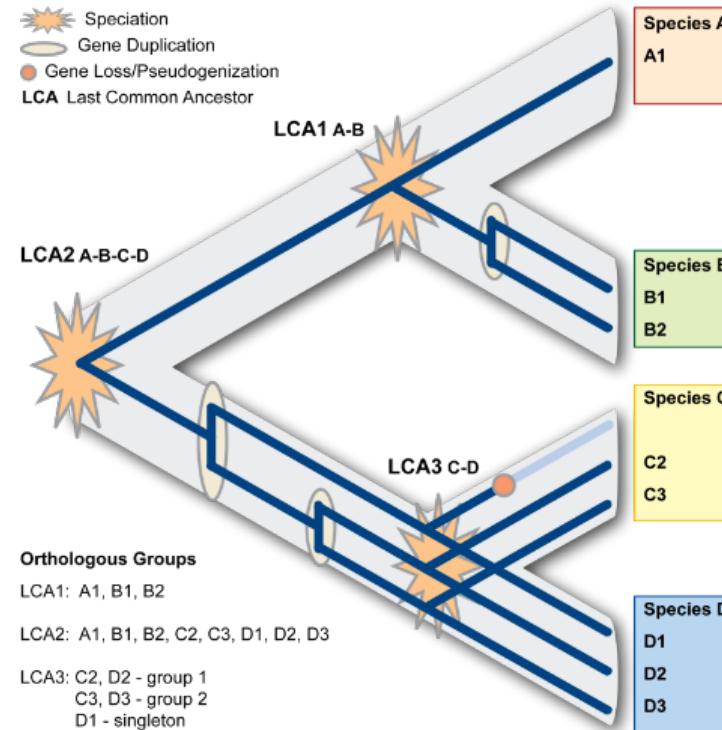
Orthologues arise by vertical descent from a single gene of the last common ancestor

Hierarchy

Orthology is relative to the species radiation under consideration

Orthologous Groups

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



Orthology Delineation

What is orthology?

How do we delineate orthologs?

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



Orthology Delineation

OrthoDB v11



About Documentation

SparQL

API

Data

Soft

Charts

Upload

Login

Text

e.g. hsp70, sex-lethal, "cytochrome c", kinase -serine

► Advanced

Submit



The hierarchical catalog of orthologs
mapping genomics to functional data

Eukaryotes
1,952

Prokaryotes
18,158

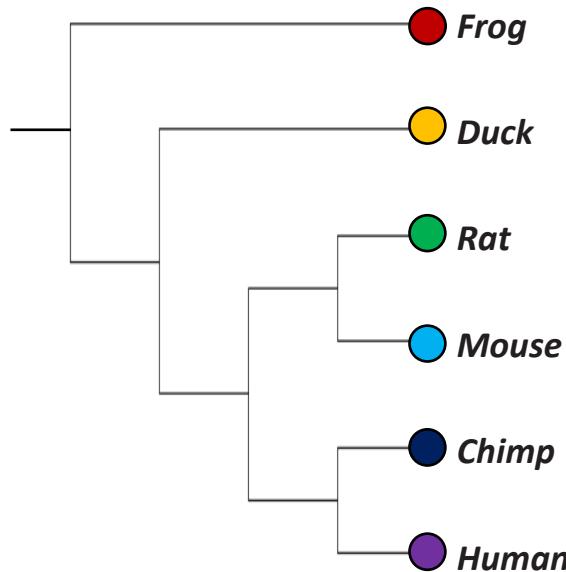
Viruses
7,962

Genes
100M

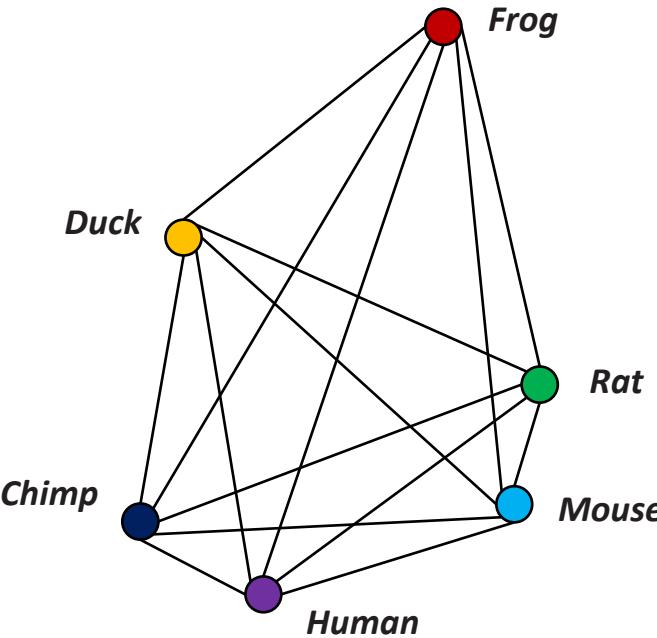


How do we delineate Orthology?

tree-based approaches



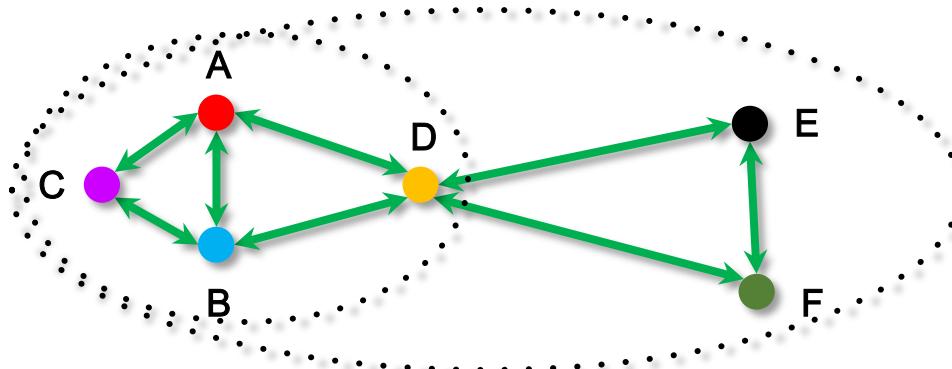
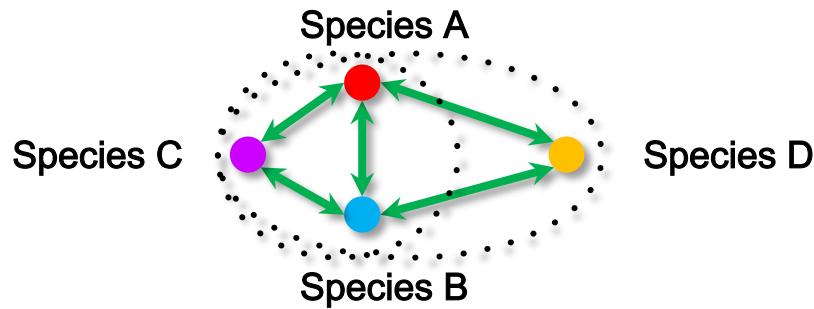
graph-based approaches



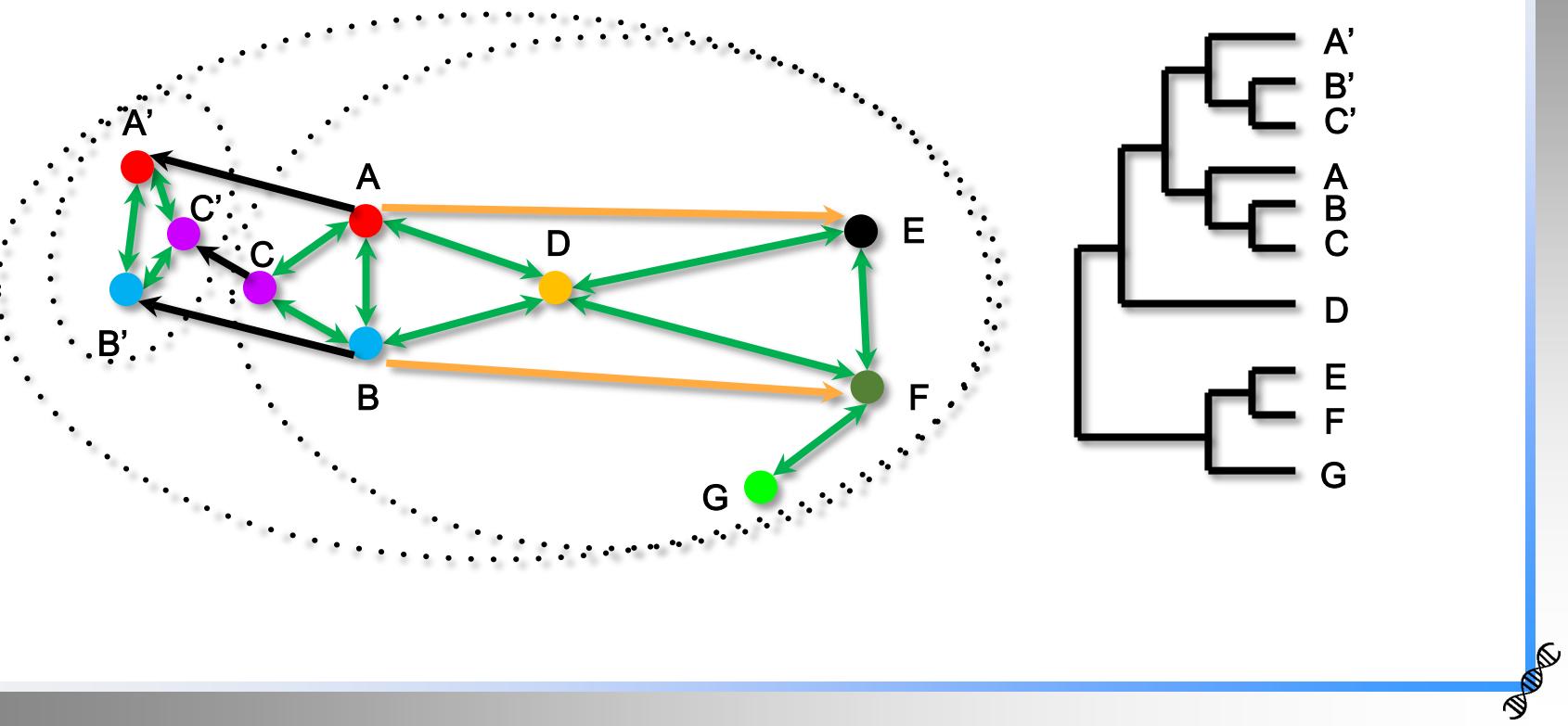
Single-Copy Orthologs



Graph-based best -reciprocal -hits



Within -clade duplications



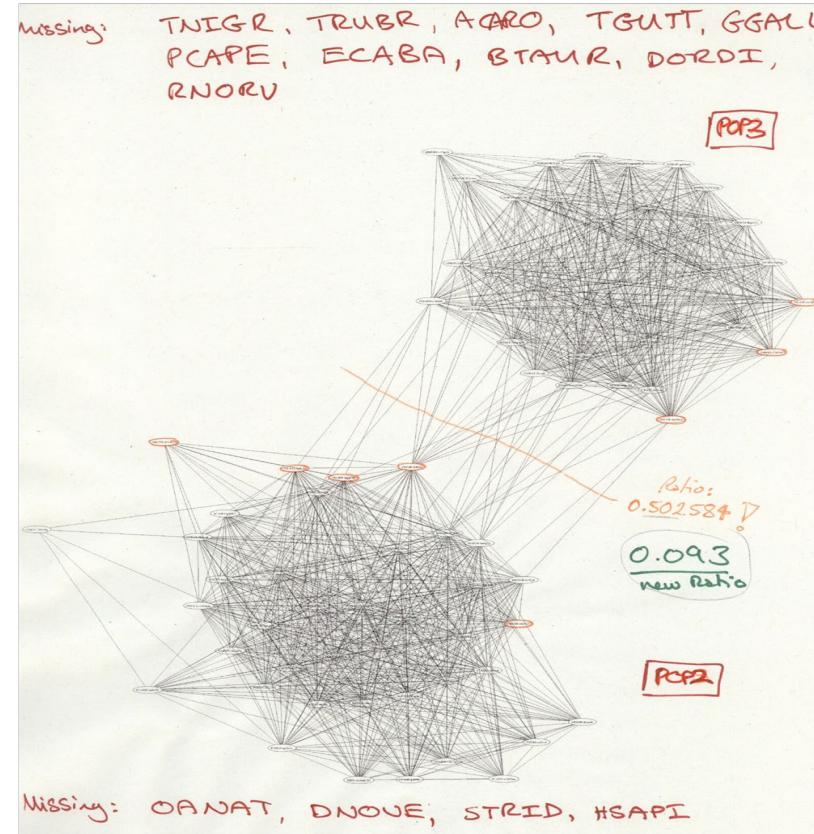
Real-world data can be messy!

Real example:

POP3 missing from 10 vertebrates

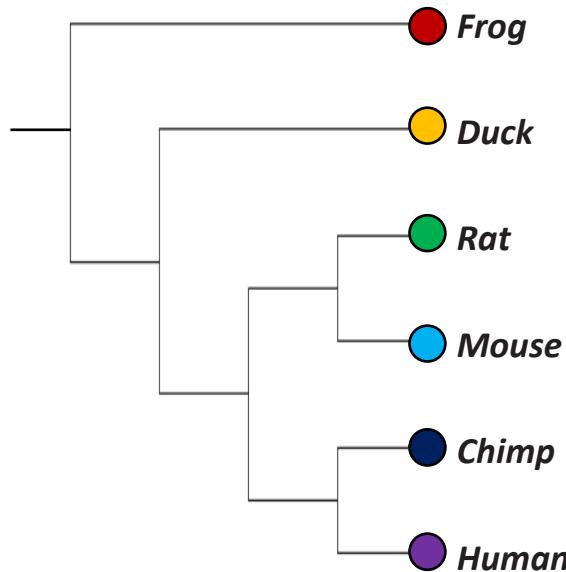
POP2 missing from 4 vertebrates

Two orthologous groups start to merge into one

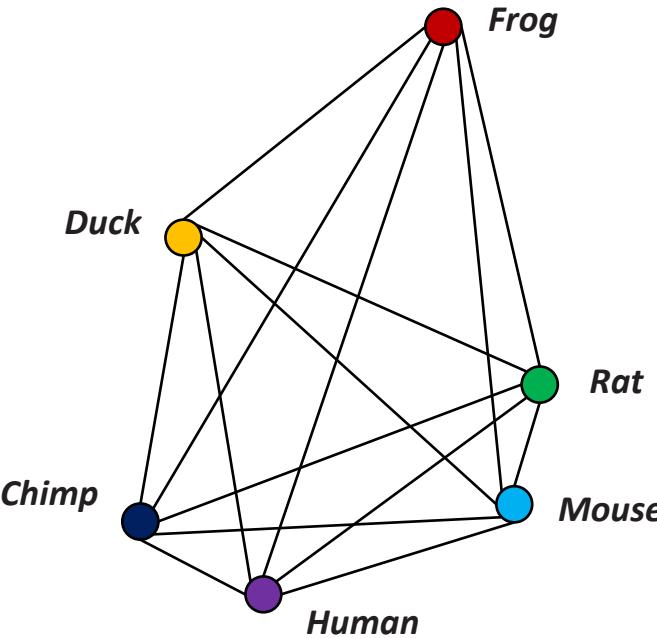


How do we delineate Orthology?

tree-based approaches



graph-based approaches



Single-Copy Orthologs



Orthology Delineation

What is orthology?

How do we delineate orthologs?

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



Quick Quiz

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

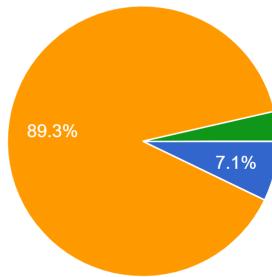
28 responses



Biodiversity Bioinformatics: Quick Quiz

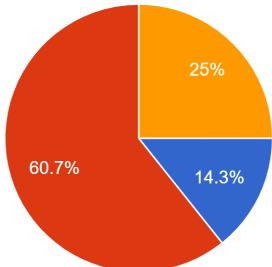
Orthology – What?

Orthology – How?



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

28 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

- 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



Quick Quiz

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

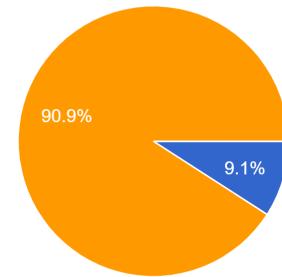
22 responses



Biodiversity Bioinformatics: Quick Quiz

Orthology – What?

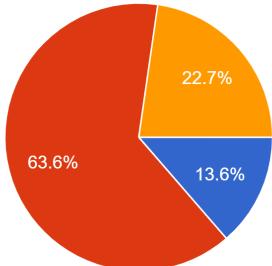
Orthology – How?



- ... 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?

22 responses

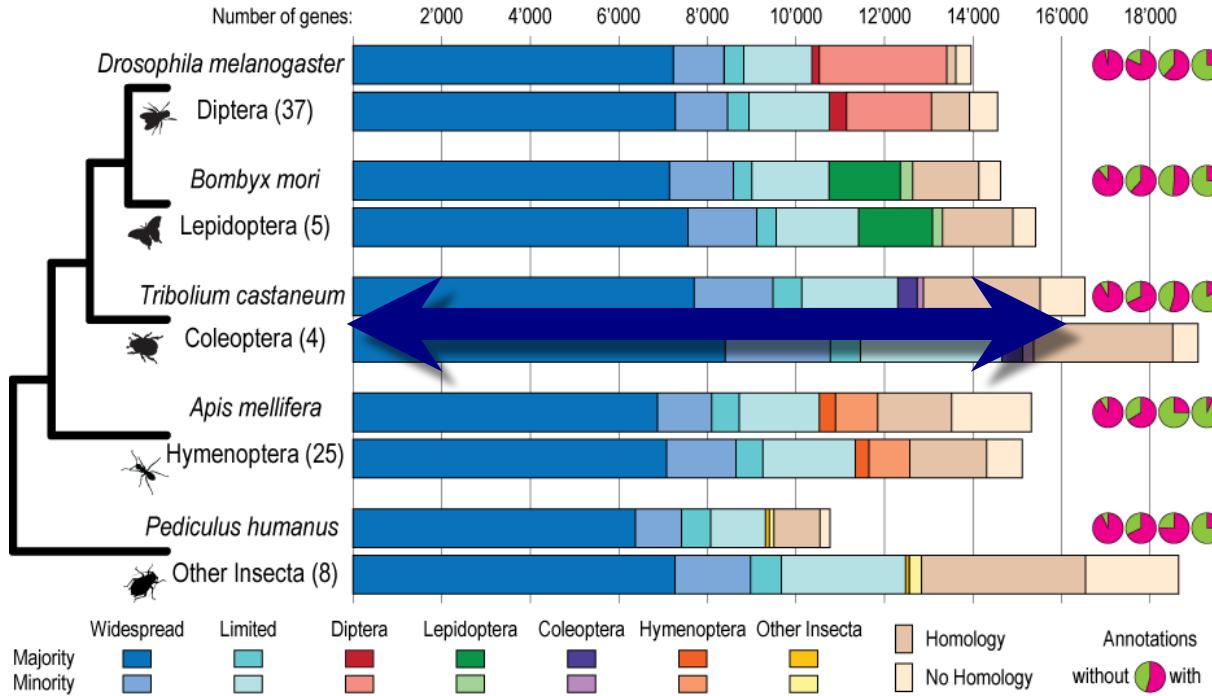


- Gene trees are reconciled with the known species tree to define speciations and duplications
- Best-reciprocal-hits determine how genes are progressively added to form orthologous groups
- The full all-against-all best-reciprocal-hit graph is progressively split to define groups of orthologues



Orthology – 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

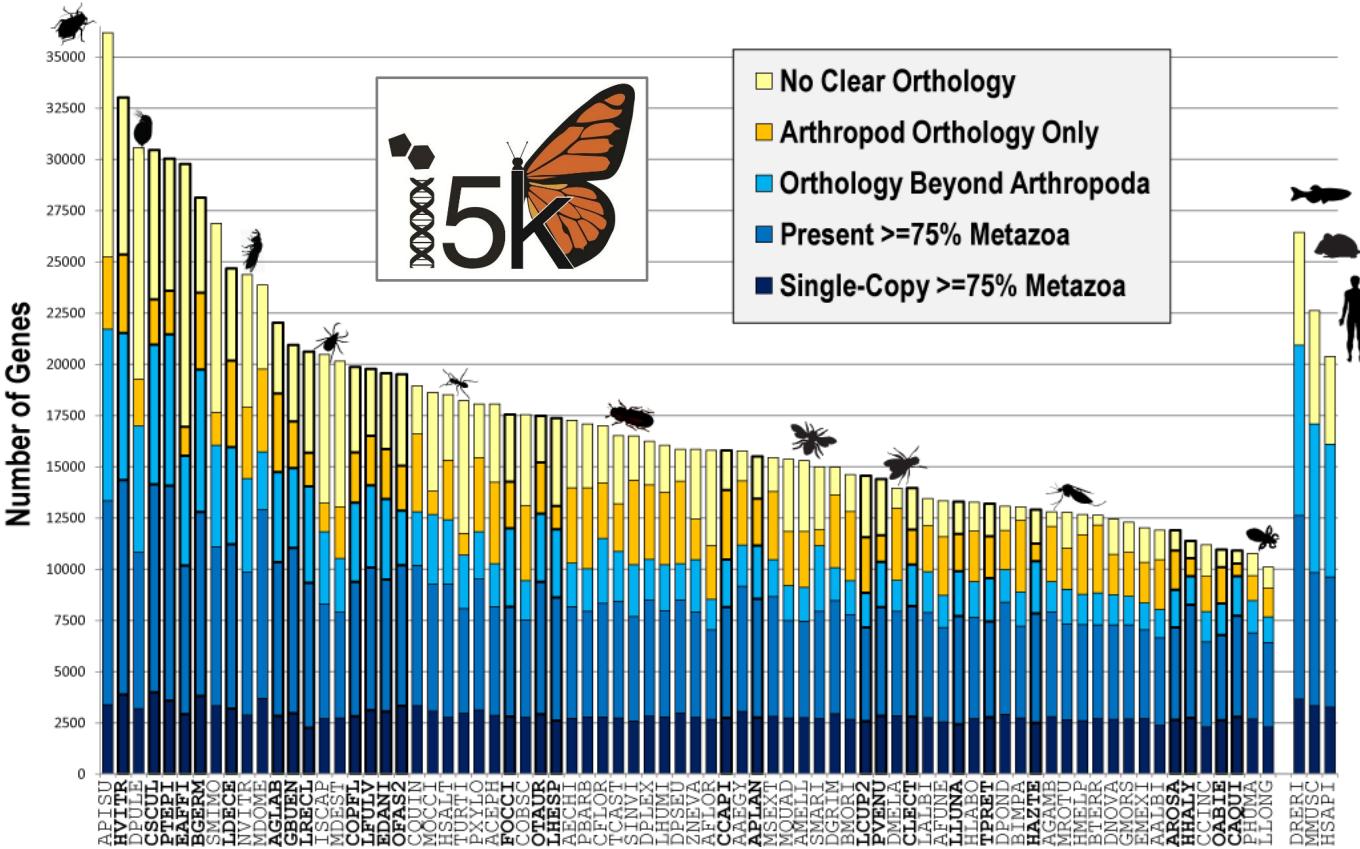
“validity of the conjecture on **functional equivalency** of orthologs is crucial for reliable annotation of newly sequenced genomes and, more generally, for the progress of functional genomics.

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.”

Annu. Rev. Genet.
2005. 39:309–38

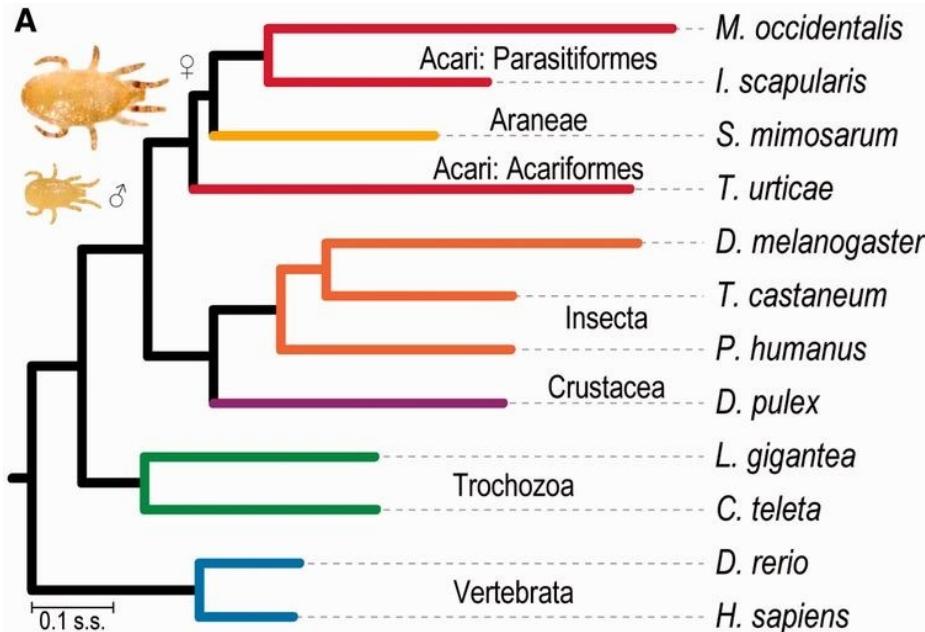


Evolutionary histories: classes



Unique
Variable
Common

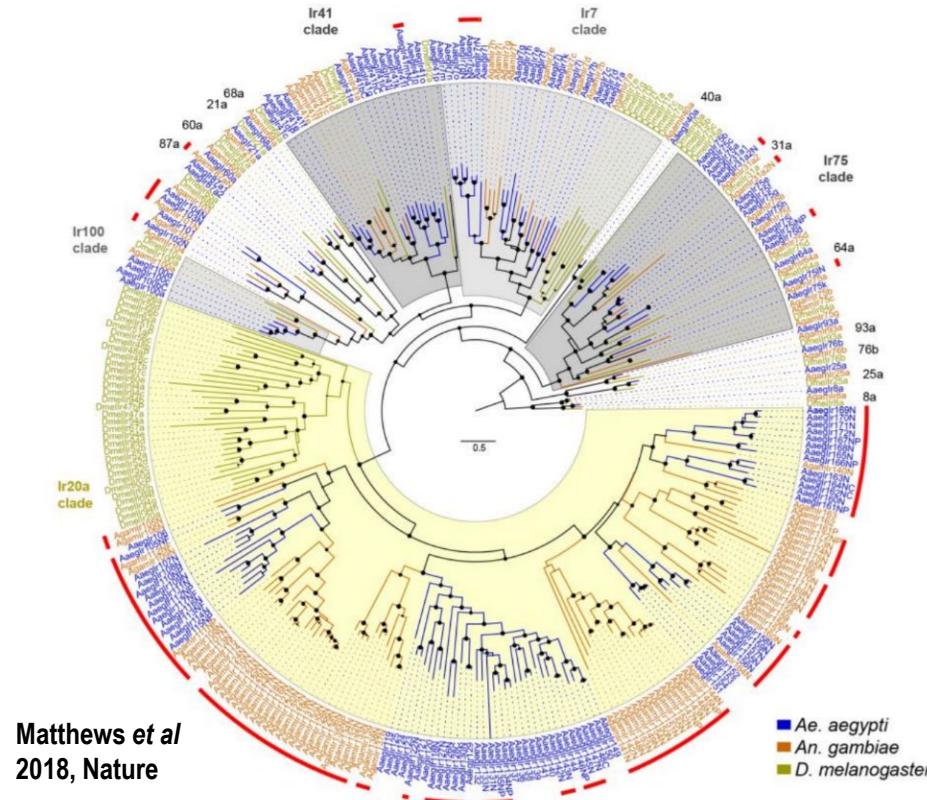
Species Tree Estimation



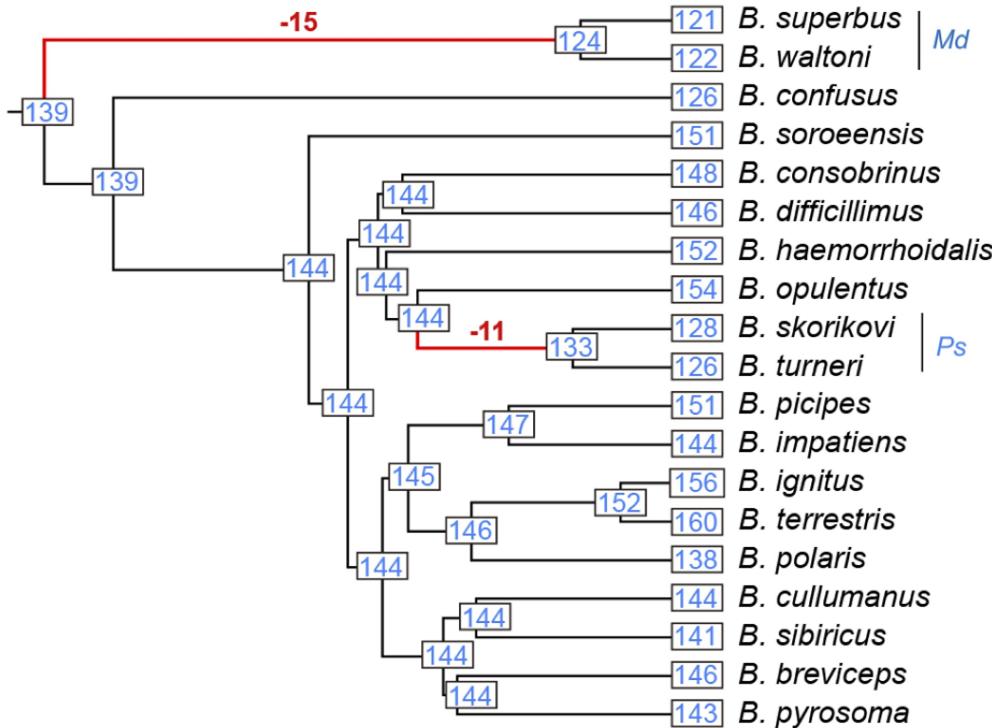
Phylogenomics with single -copy orthologues

Gene Family Tree Building

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



Ancestral Copy -Number Reconstruction



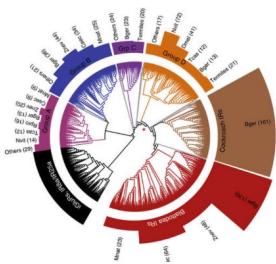
Bumblebee Odorant Receptors : two major gene loss events



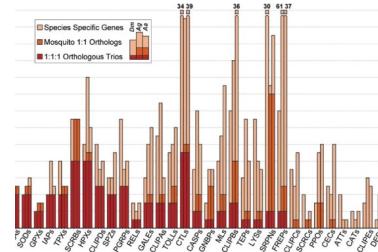
Dynamically evolving families

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

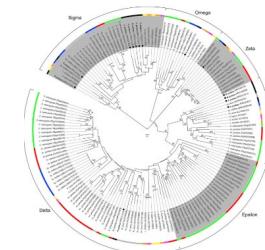
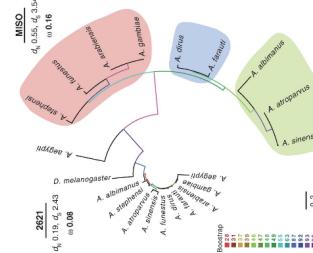
IMMUNITY



CHEMOSENSATION



DETOXIFICATION



Goals for Today's Workshop

- Learn how to use BUSCO to assess genomics data
- Understand what BUSCO is attempting to achieve
- Learn how the BUSCO datasets are constructed
- Understand how OrthoDB delineates orthologous groups that are used as the main input for BUSCO

OrthoDB
BUSCO



Assessing genomics data quality: BUSCO

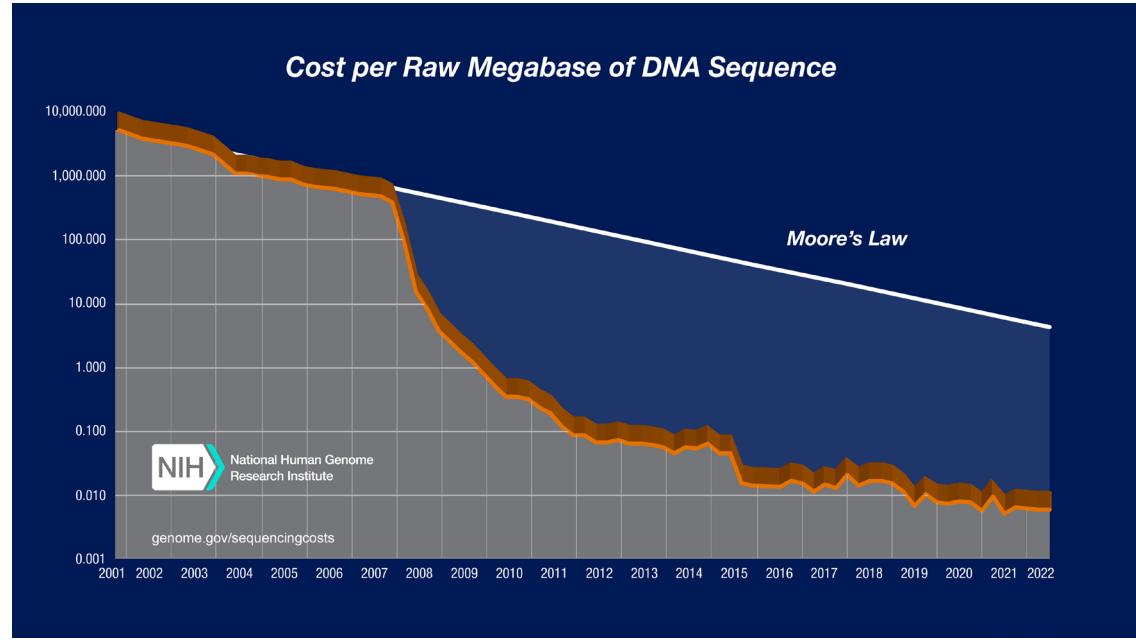
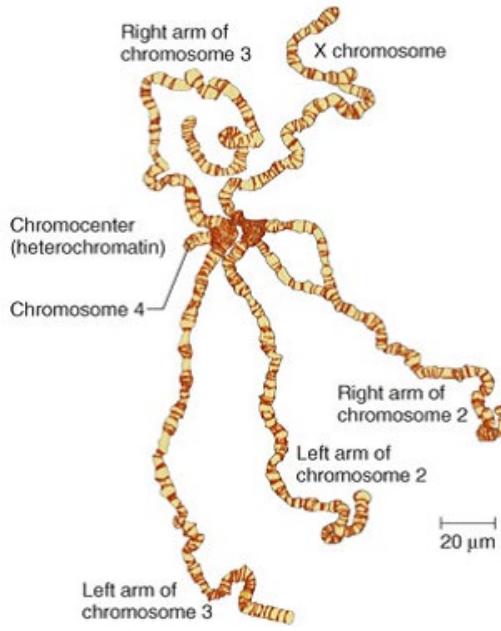
What is BUSCO?

How does BUSCO work?

*Why do we need BUSCO?
(BUSCO in action)*



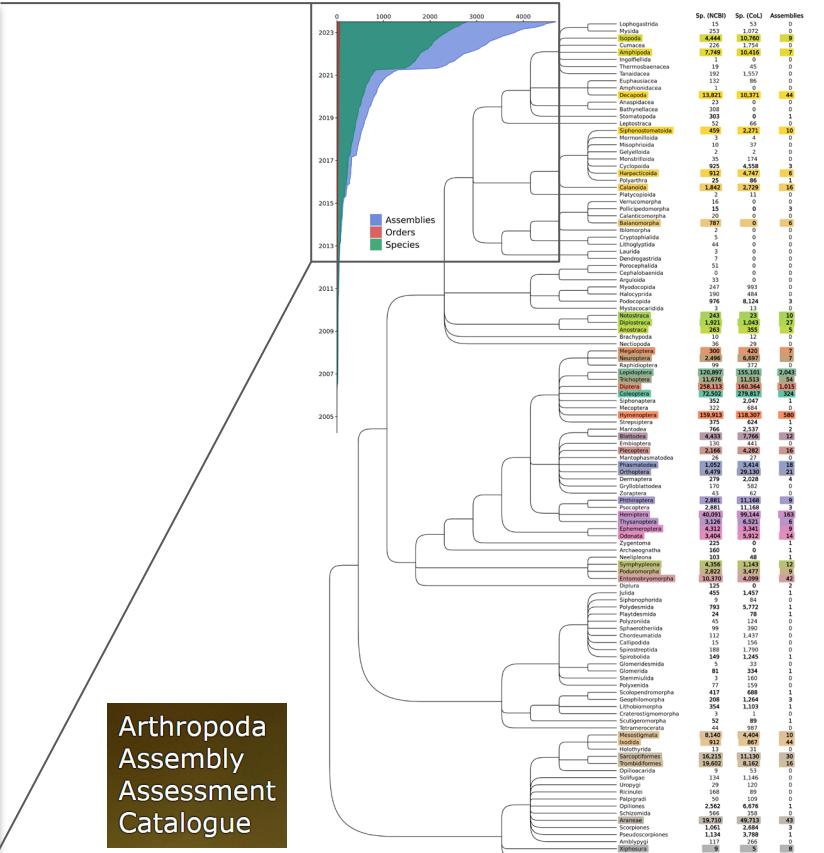
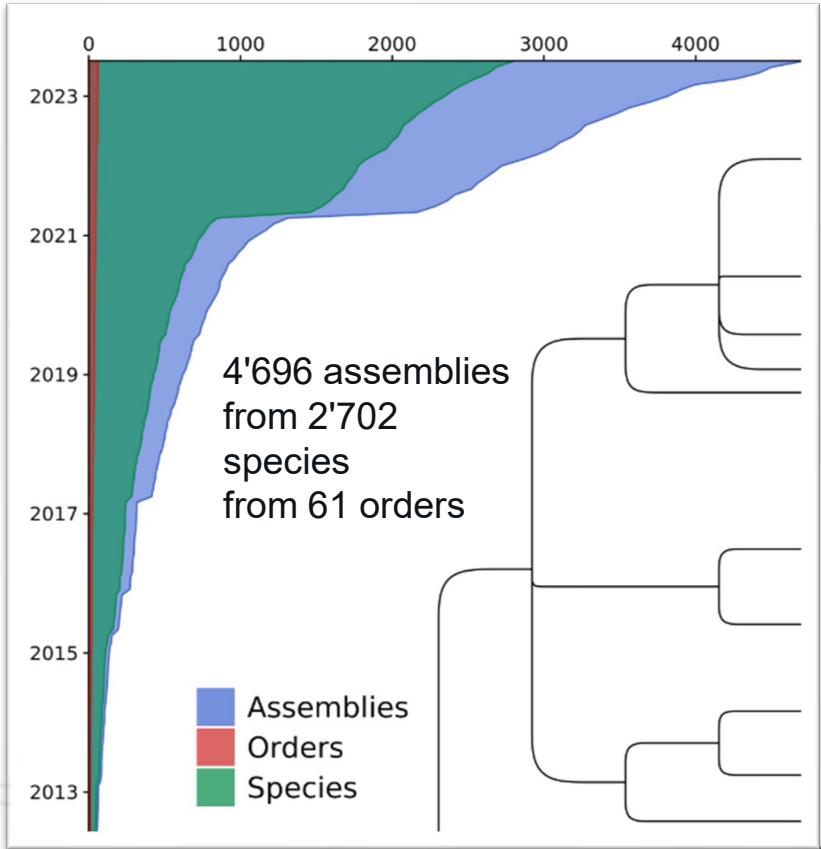
Genomics for Everyone!



ATAATAACGCCGAGATCCCAGAAATGACTCCTCTTACTTAGATTGAAGCTTCATAACGCCATGCGTCTGATCAAAGAT
TACGTTAGCGAGGATCTGCACAAGTACATACCGAGGAGCGTGGCCACCATCAGCGAGCTGCGCGCTGCTCCCGATTCCAG
TGCCAAAGTAAACATGCGCCGCTTACCCATGCGCTTACCAACGAGCTACACATTAAGCACCAGCACACGG
CAACGGAGGAGATACCCAGCTGGAGTCCCTGGGCAATATACACATTAAGCACCAGCACAGCGGGCAGGCCAATCCTTGCTG
GAGATCGATTATGCGGAGCTAGAAAACGCCACGGACGGCTGGAGTCCGGATAATCGACTGGGACAGGGCGGATTGGAGA

Nucleotide-level resolution

Rapidly increasing numbers of cool genomes

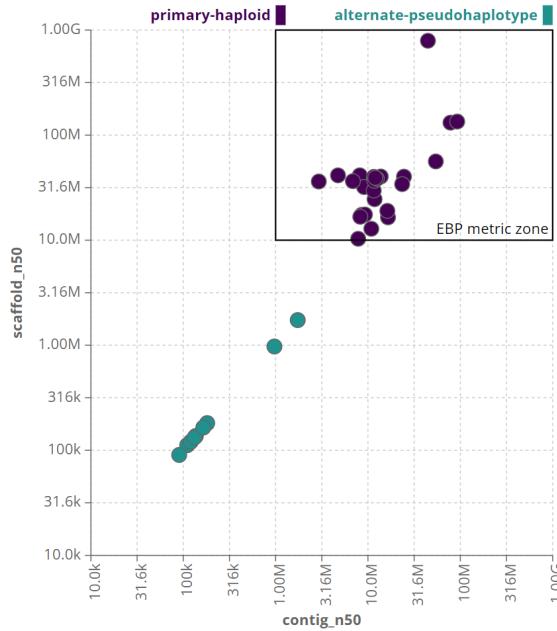


BioGenome Projects producing new data



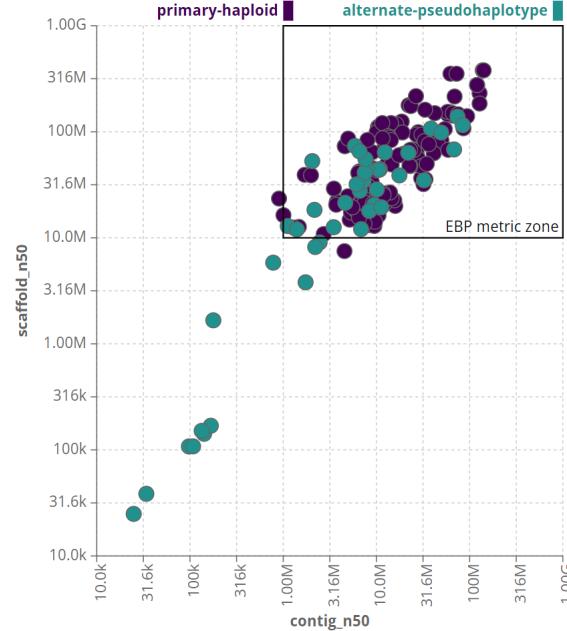
USDA
United States Department of Agriculture
Agricultural Research Service

Ag100Pest Initiative (Ag100Pest)

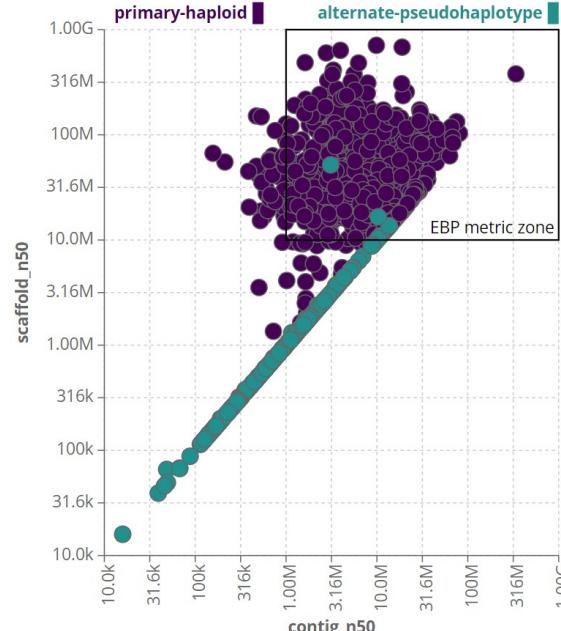


CALIFORNIA
CONSERVATION
GENOMICS
PROJECT

The California Conservation Genomics Project (CCGP)



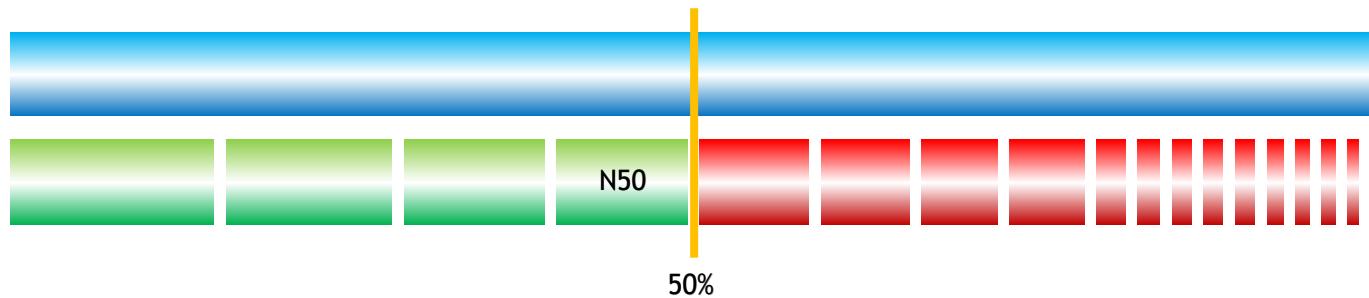
European Reference Genome Atlas (ERGA)



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: evolutionarily expected genes

Widespread genes in extant species from a given taxa
should be present in any newly sequenced species



Features in common:
6 legs
2 compound eyes
1 pair of wings
Etc.

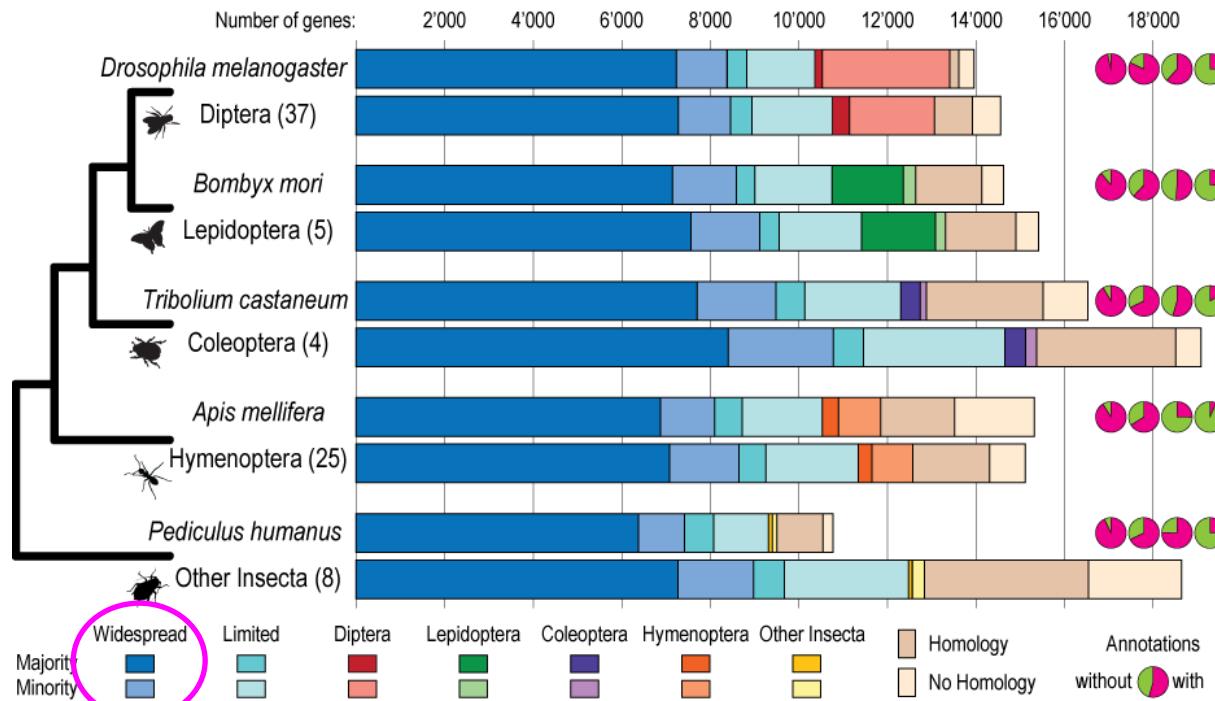


Therefore we EXPECT:
6 legs
2 compound eyes
1 pair of wings
Etc.

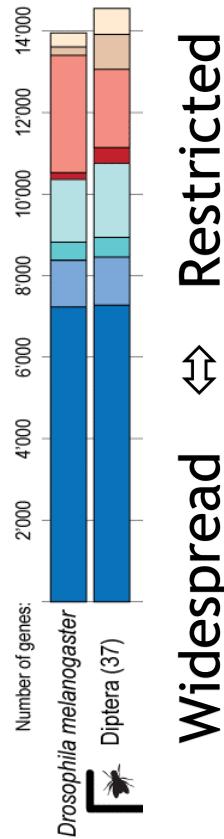


BUSCO: evolutionarily expected genes

Widespread genes in extant species from a given taxa
should be present in any newly sequenced species



BUSCO: looking for widespread & unique genes



Widespread ⇔ Restricted

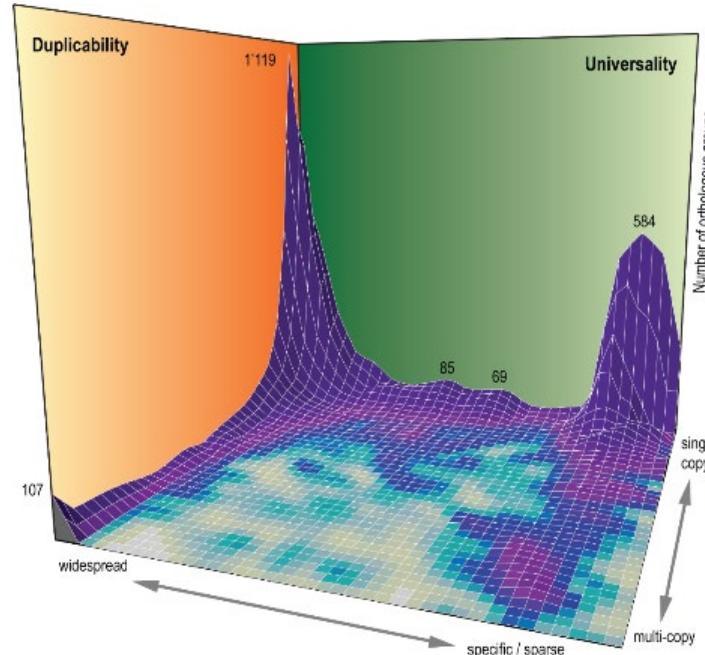
Drosophila melanogaster

orthology with genes from 80 insects

From mostly single-copy to mostly multi-copy

From present universally to present sparsely

Orthology Landscape



BUSCO: looking for widespread & unique genes



Benchmarking Universal
Single-Copy Orthologues

QUEST FOR QUALITY

“BUSCO CALIDAD”

“BUSCO QUALIDADE”

Genome analysis

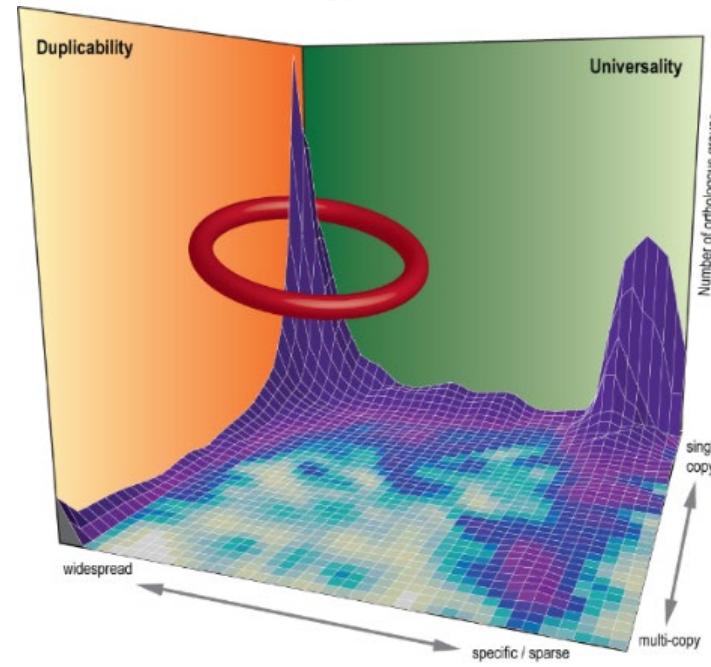
Applications Note

Bioinformatics

BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs

Felipe A. Simão^a, Robert M. Waterhouse^a, Panagiotis Ioannidis, Evgenia V. Kriventseva, Evgeny M. Zdobnov^b

Orthology Landscape



<http://busco.ezlab.org>

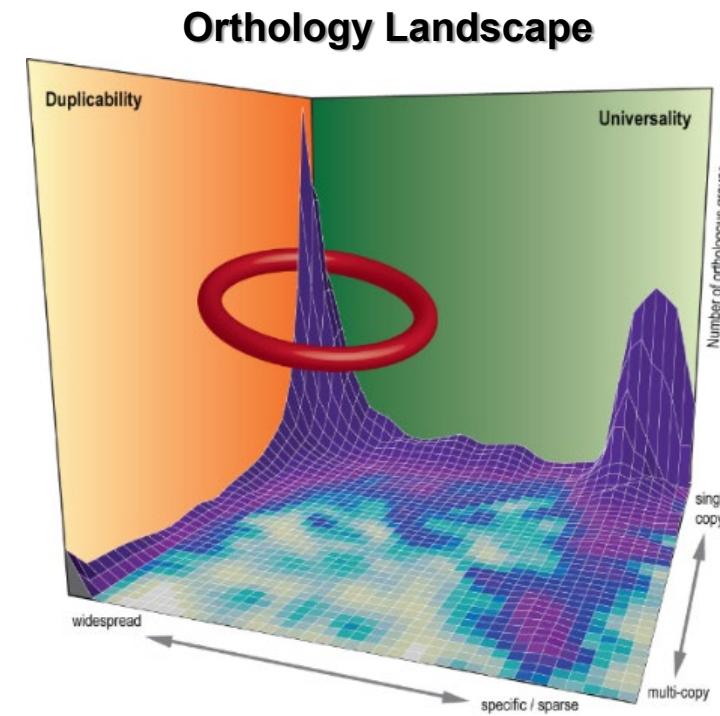
BUSCO: looking for widespread & unique genes

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

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

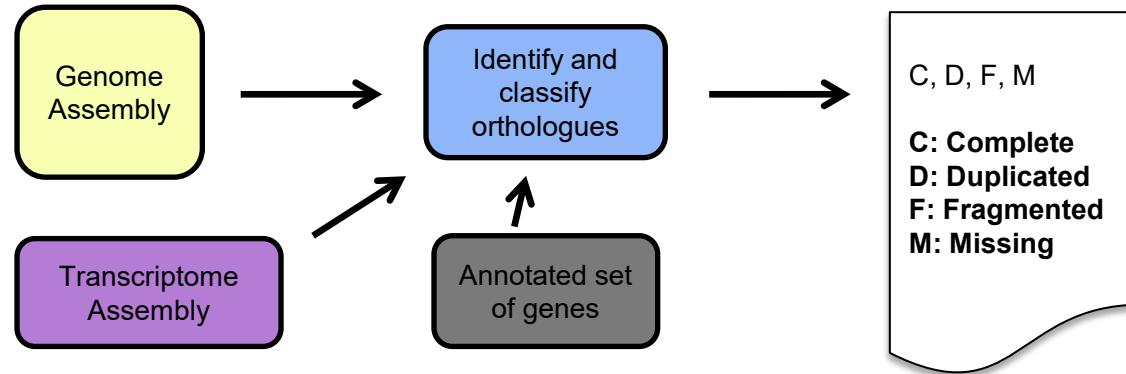
Implemented Assessments
Gene Content Completeness
genome assemblies
annotated gene sets
assembled transcriptomes

Bonus Features
genes for phylogenomics
gene predictor training

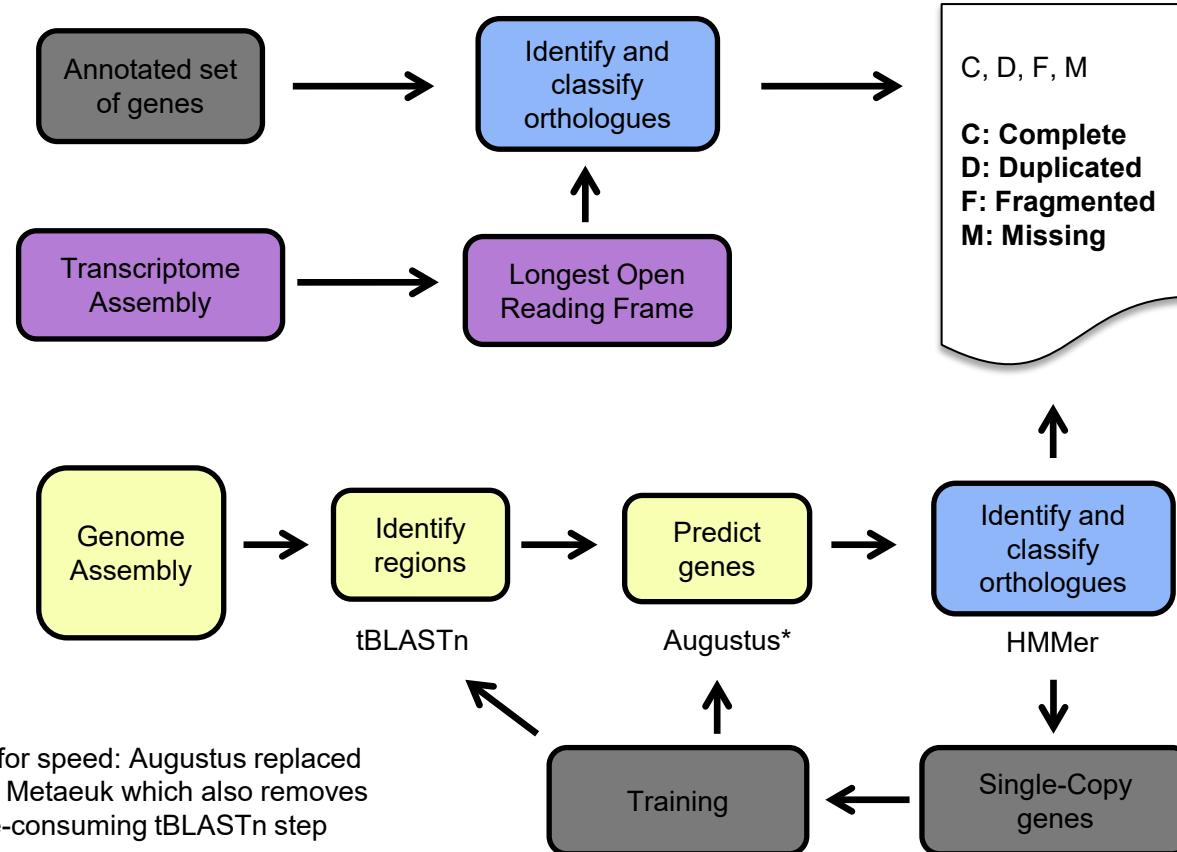


<http://busco.ezlab.org>

BUSCO completeness assessments



BUSCO completeness assessments



*v5 for speed: Augustus replaced with Metaeuk which also removes time-consuming tBLASTn step

Building BUSCO lineage datasets

For each clade/lineage ...

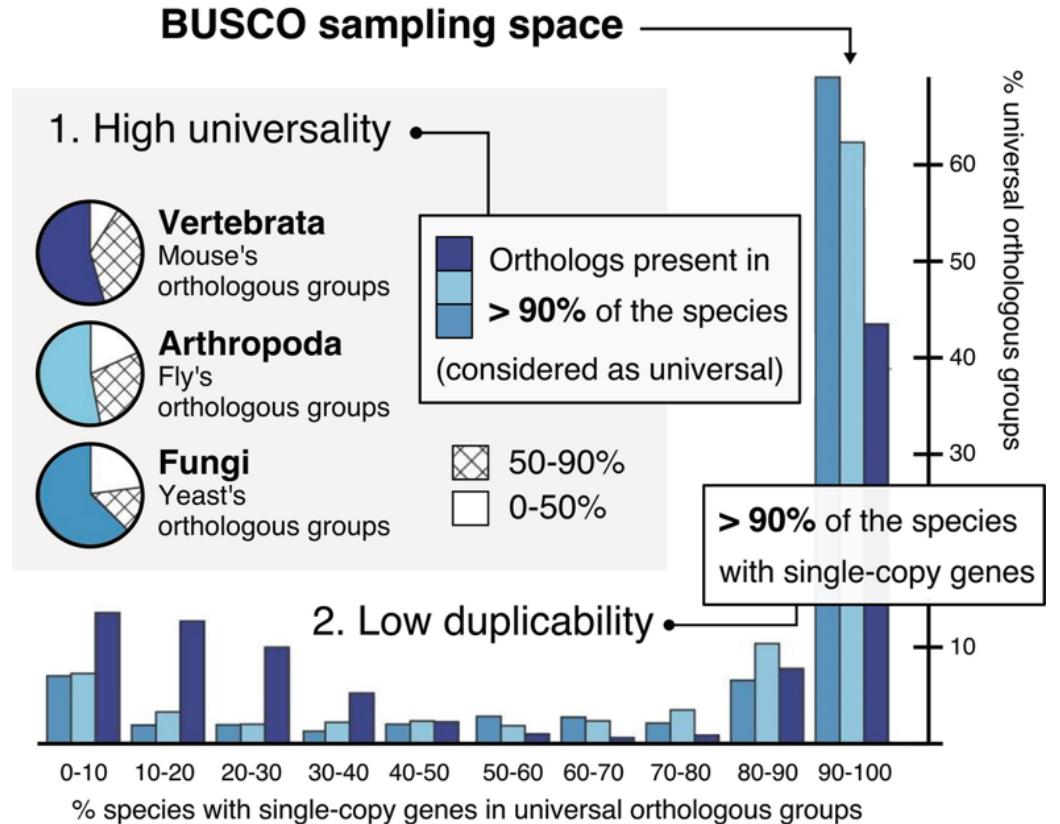
E.g. vertebrates, arthropods, or fungi

Filter OrthoGroups to retain those:

- Present in >90% species
- Single-Copy in >90% species

To obtain lineage datasets of

Benchmarking Universal
Single-Copy Orthologues



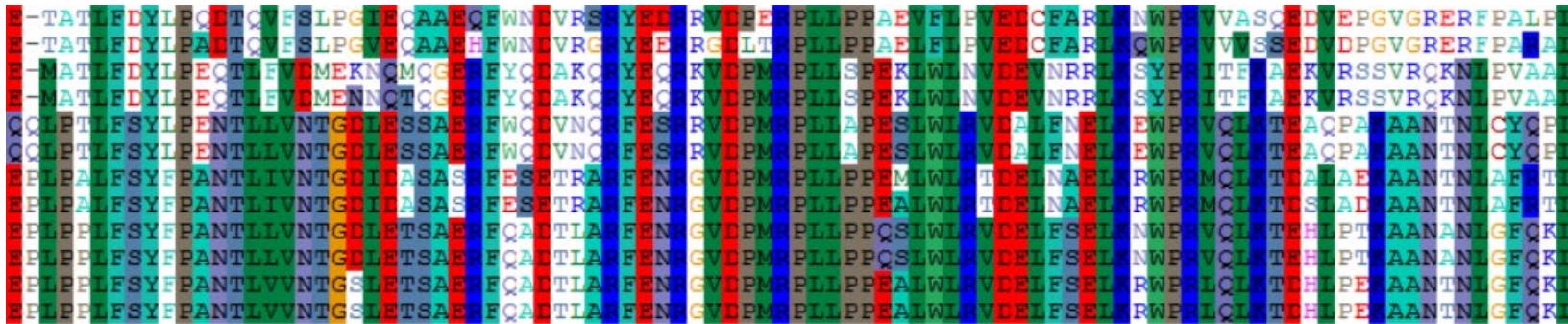
Building BUSCO lineage datasets



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*



Building BUSCO lineage datasets

Filtering of initial BUSCO sets

*HMM profiles run against all proteins
from all input species*

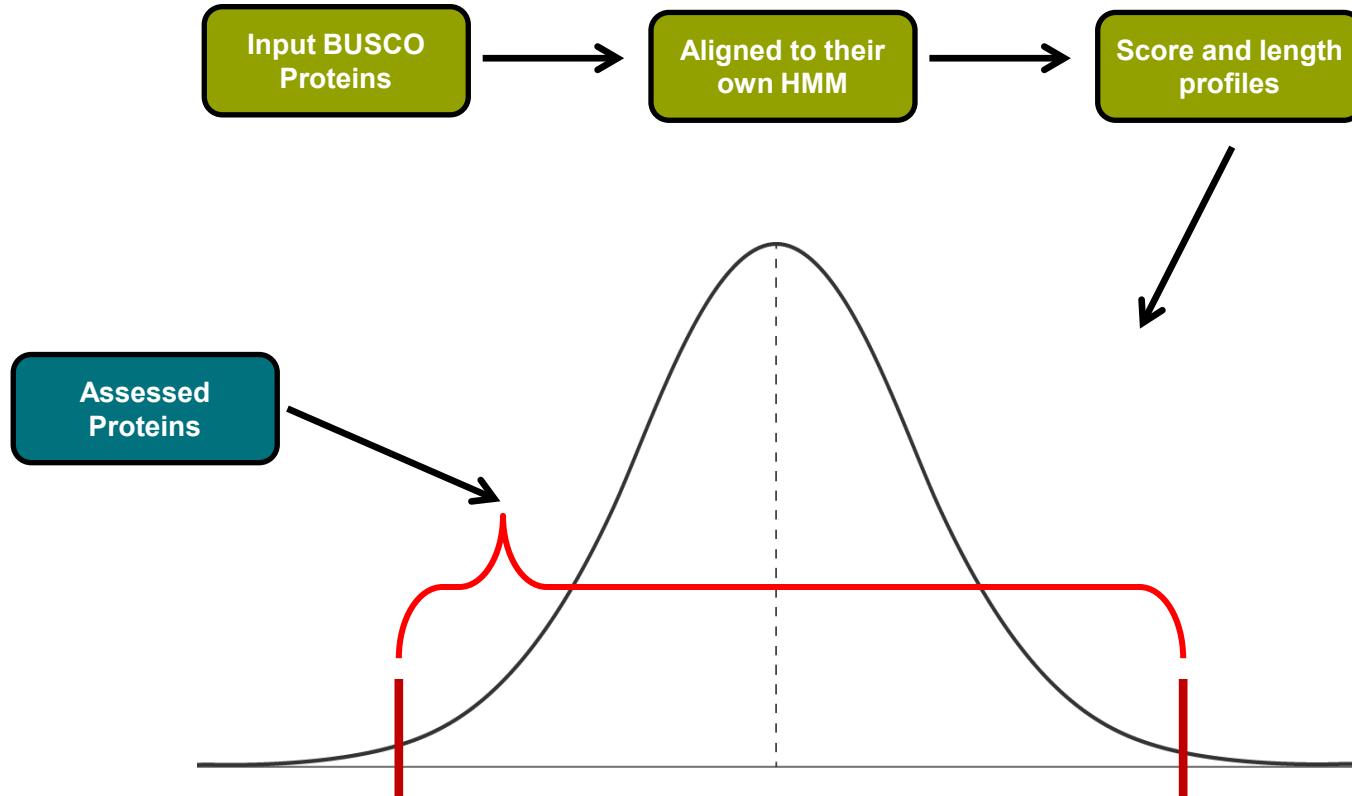
*Score and length cut-offs fine-tuned for each BUSCO
to maximise sensitivity and specificity*

*Only BUSCO profiles with high sensitivity
and specificity are kept*

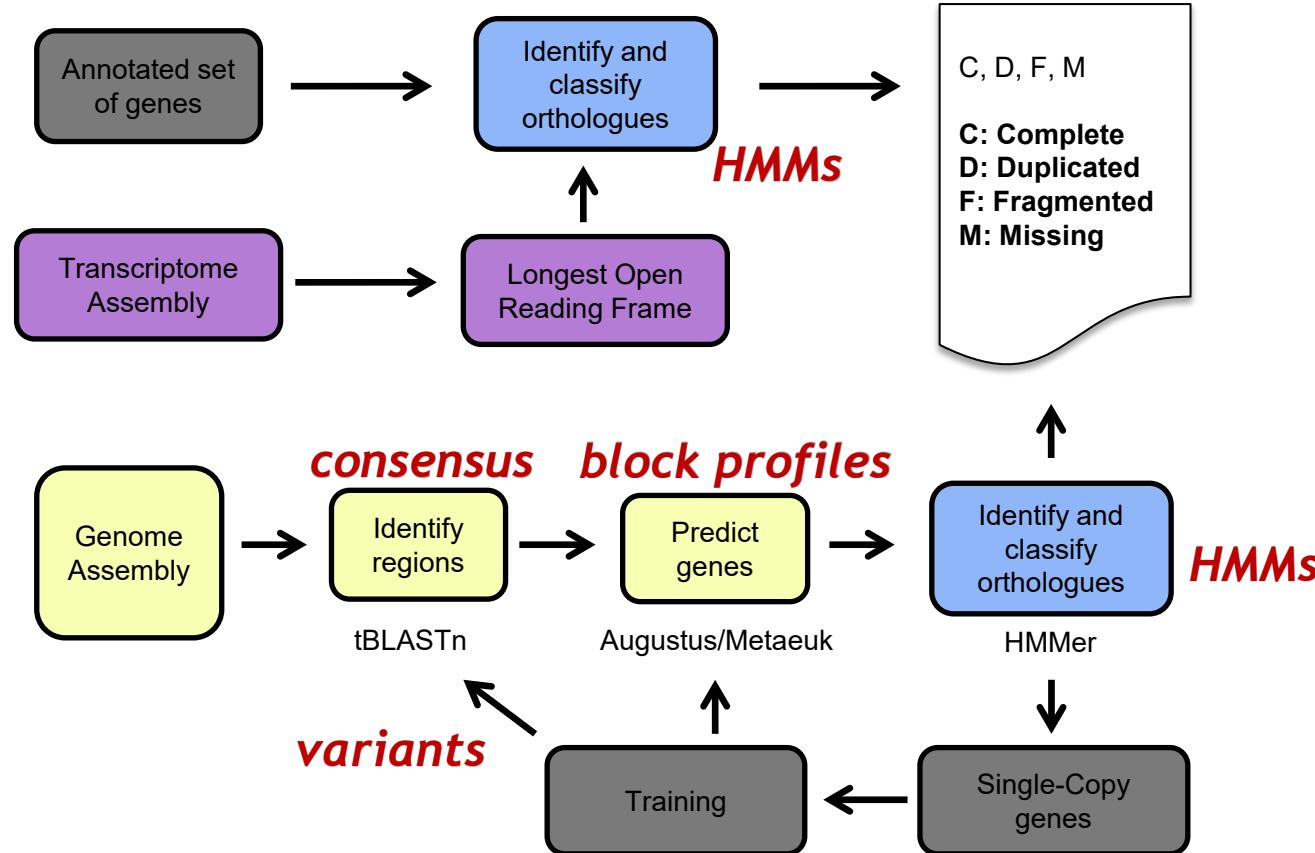
*Testing of BUSCO profiles on non-input species
remove BUSCOs whose Augustus/MetaEuk step fails*



Identify and classify orthologues – HOW?



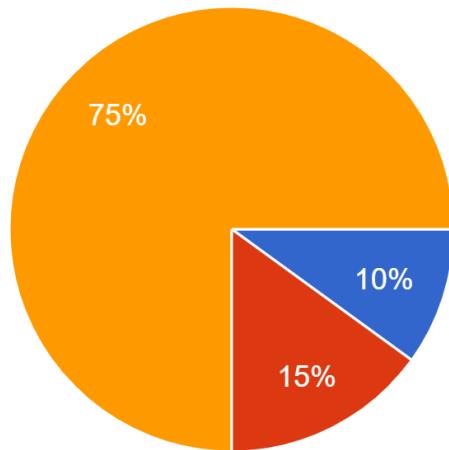
BUSCO completeness assessments



Quick Quiz

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

20 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



Assessing genomics data quality: BUSCO

What is BUSCO?

How does BUSCO work?

*Why do we need BUSCO?
(BUSCO in action)*

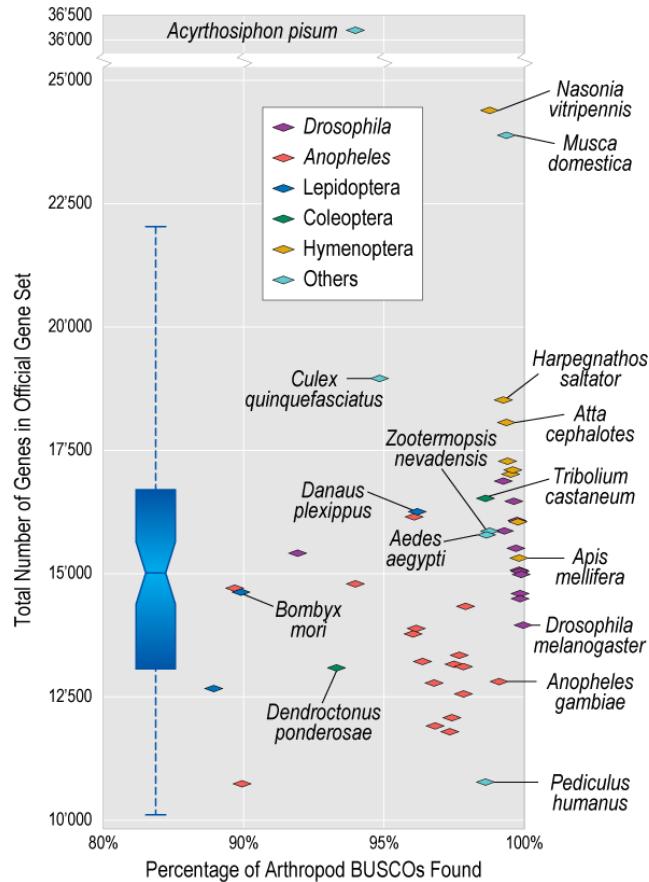


BUSCO research applications

Many genomics studies now using BUSCO - what for?



BUSCO in action: insect gene sets



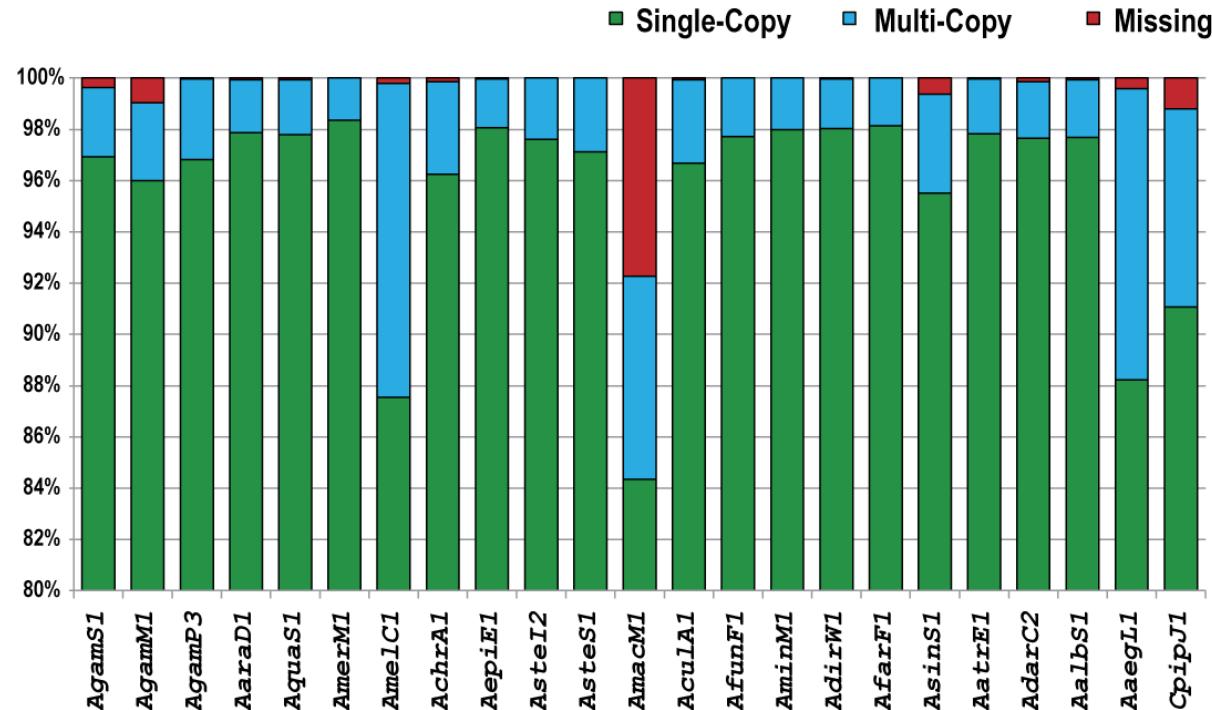
Large gene sets are
not necessarily complete

Current Opinion in
Insect Science
2015, 7: 15–23
A maturing understanding of the composition
of the insect gene repertoire
Robert M Waterhouse^{1,2,3,4}

Small gene sets are
not necessarily incomplete



BUSCO in action: Anopheles genomes



- Most remarkably complete in terms of genes
- Missing: *An. maculatus* - fragmented assembly
- *An. christyi*, also fragmented, but few missing
- Duplicates: *An. melas* - assembly haplotypes (fixed)

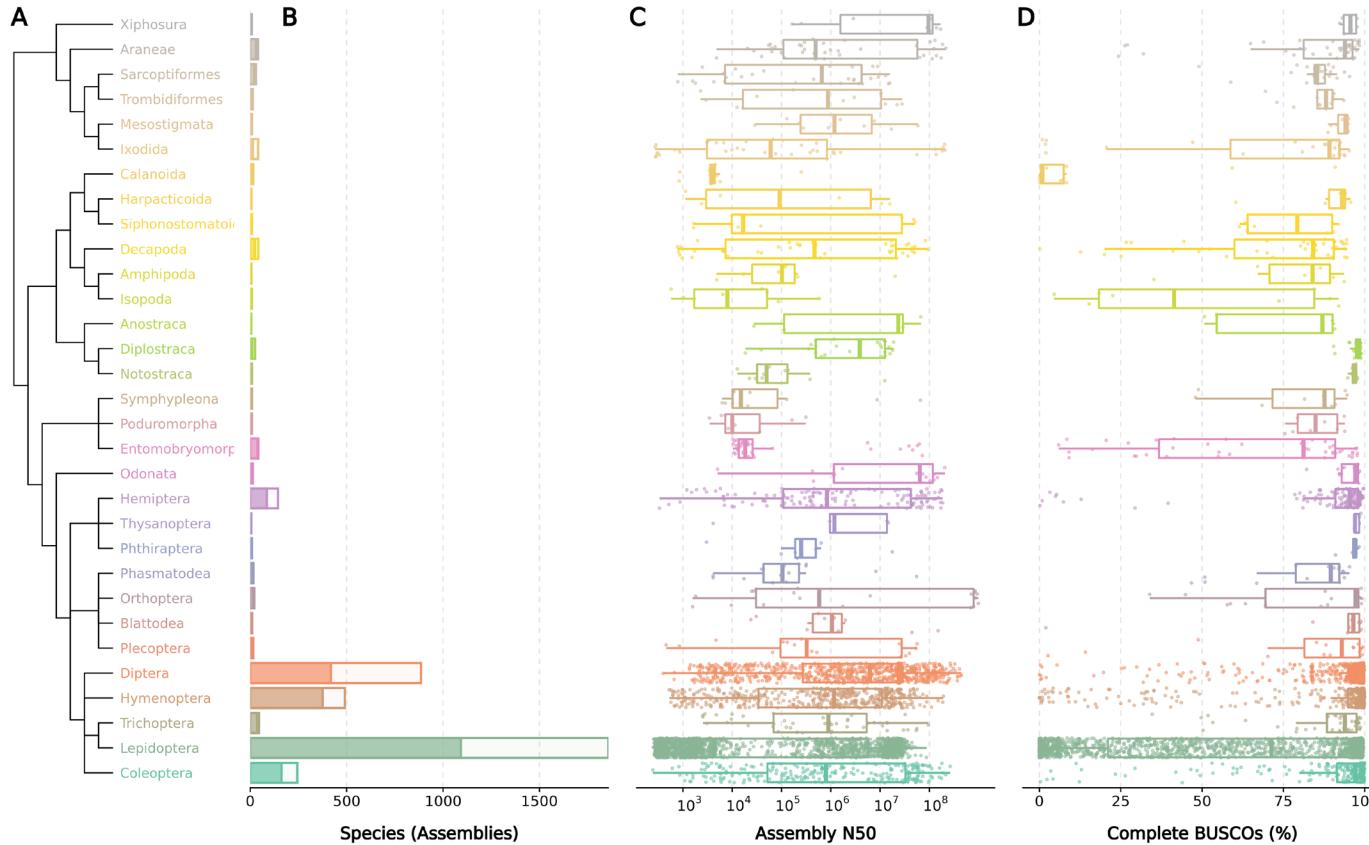


BUSCO utilities beyond Quality Control

- Comparative Genomics
- Gene Predictor Training
- Phylogenomic Analyses



BUSCO in Comparative Genomics

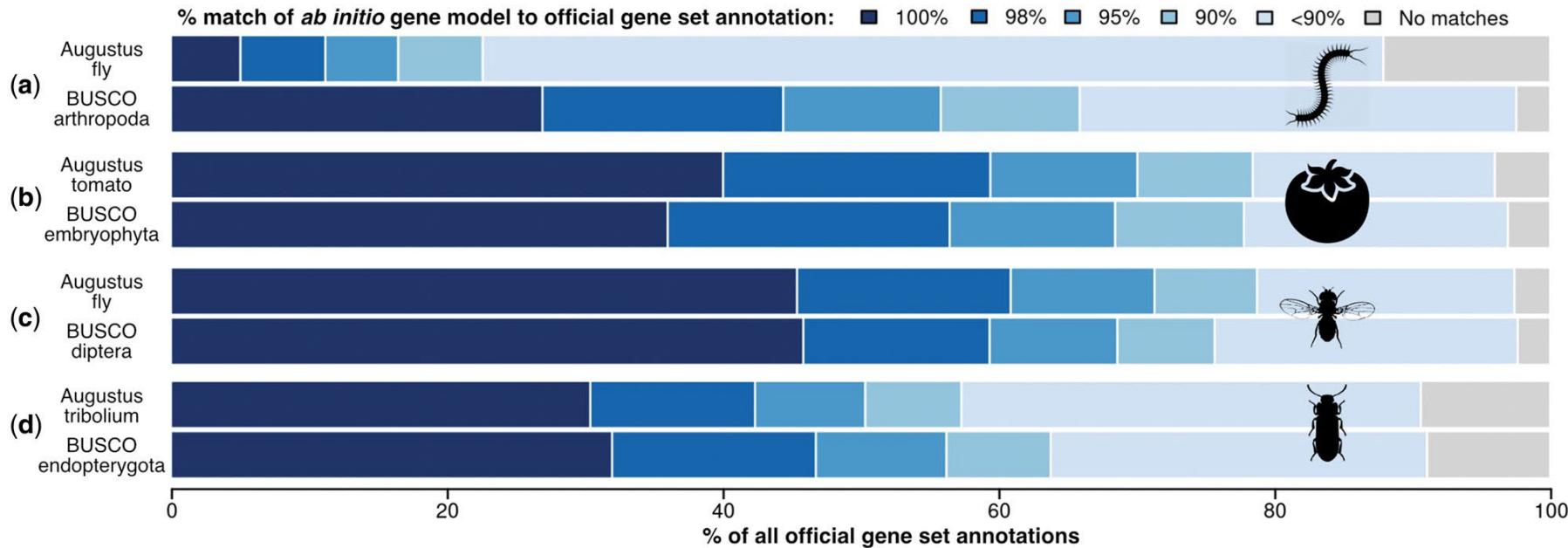


**Analyses sensitive
to incomplete
datasets**

Quantitative assessments with BUSCO offer logical selection criteria to help focus on the most complete genomic resources available



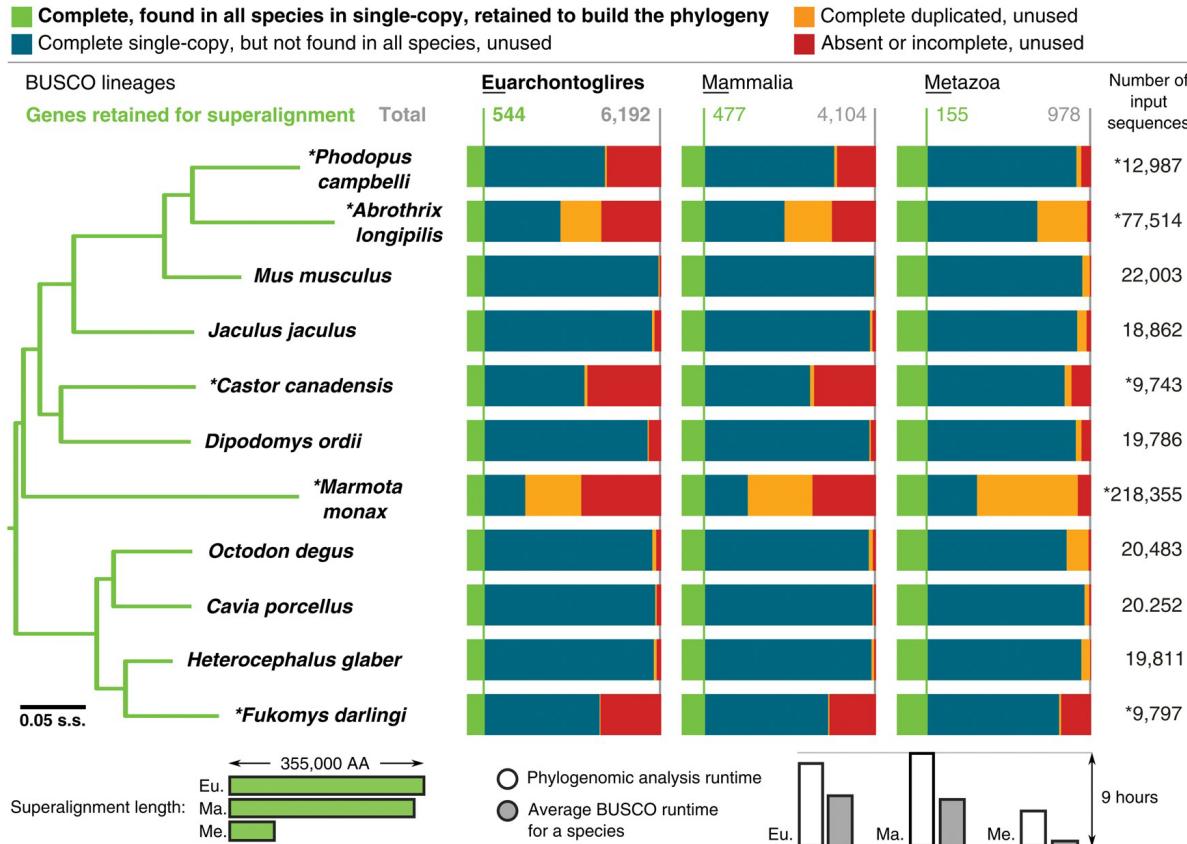
BUSCO in Gene Predictor Training



BUSCOs, being generally widely- and well-conserved genes, offer ideal predefined sets for such training procedures, even without the need to first perform RNAseq



BUSCO in Phylogenomic Analyses



Estimating true phylogenetic relationships among organisms is a prerequisite to almost any evolutionary study

BUSCOs represent predefined sets of reliable markers where assessments of genomes, annotated gene sets, and/or transcriptomes can identify shared subsets from different types of genomic data for phylogenomics studies



BUSCO utilities – BUSCO in action

- Quality Control
- Comparative Genomics
- Gene Predictor Training
- Phylogenomic Analyses



Goals for Today's Workshop

- Learn how to use BUSCO to assess genomics data
- Understand what BUSCO is attempting to achieve
- Learn how the BUSCO datasets are constructed
- Understand how OrthoDB delineates orthologous groups that are used as the main input for BUSCO

OrthoDB
BUSCO

