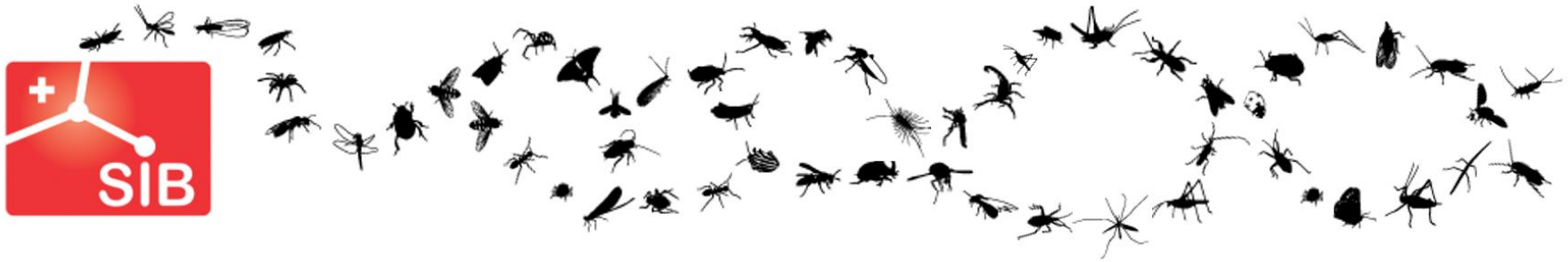


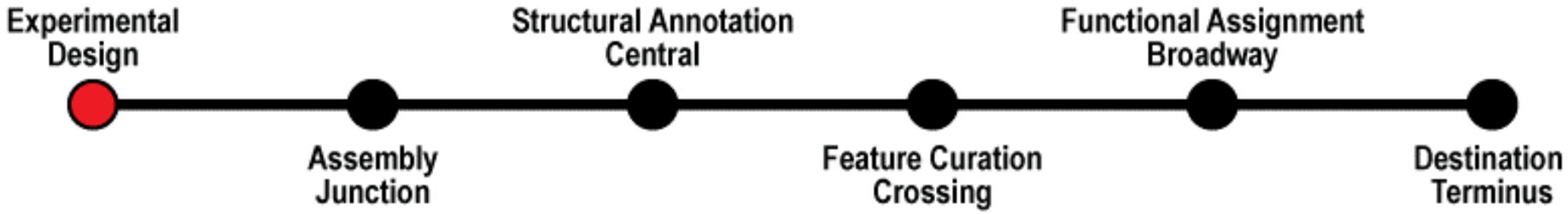
UniL SIB 3-5 September 2018

Orthology: concepts & methods

✉ robert.waterhouse@gmail.com @rmwaterhouse www.rmwaterhouse.org

The ‘bigger picture’ of a genome project

- Assembly
- Annotation
- Analysis



By the end of this lecture …

What is orthology?

How do we delineate orthologs?

And why do we need to?

✉ robert.waterhouse@gmail.com 🐦 @rmwaterhouse 🌐 www.rmwaterhouse.org

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

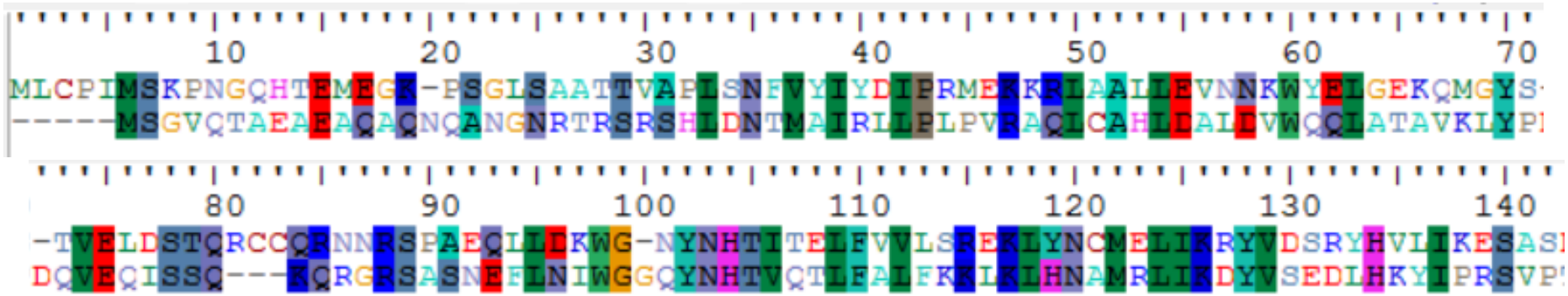
Speciation Event

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 (hence associated expectation value) to have arisen from a common ancestor

Orthology - what is it?

Homologs

Orthologs *Paralogs*

'The term homolog was introduced by **Richard Owen** in 1843 to designate "the same organ in different animals under every variety of form and function."'

'**Darwin** himself never used the term homology, but less than a year after the publication of the Origin, **Huxley**, in his review of Darwin's work, invoked homology as evidence of evolution.'

Orthology - what is it?

Homologs

Orthologs ***Paralogs***

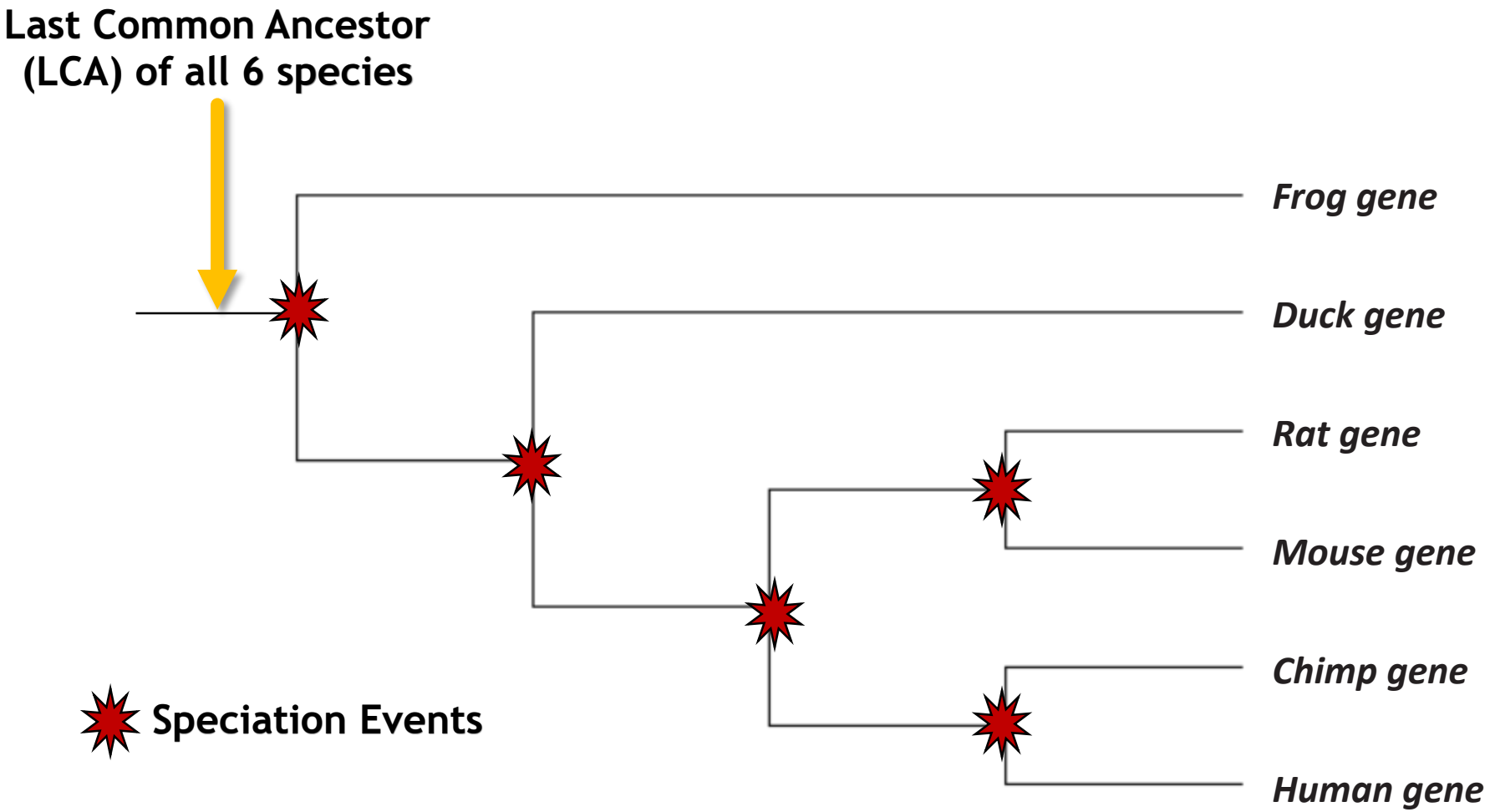
… the distinction between orthologs and paralogs and the terms themselves were introduced by **Walter Fitch** in 1970 in a now classic paper:

Fitch WM. *Syst. Zool.* 19:99–106. 1970.

DISTINGUISHING HOMOLOGOUS FROM ANALOGOUS PROTEINS

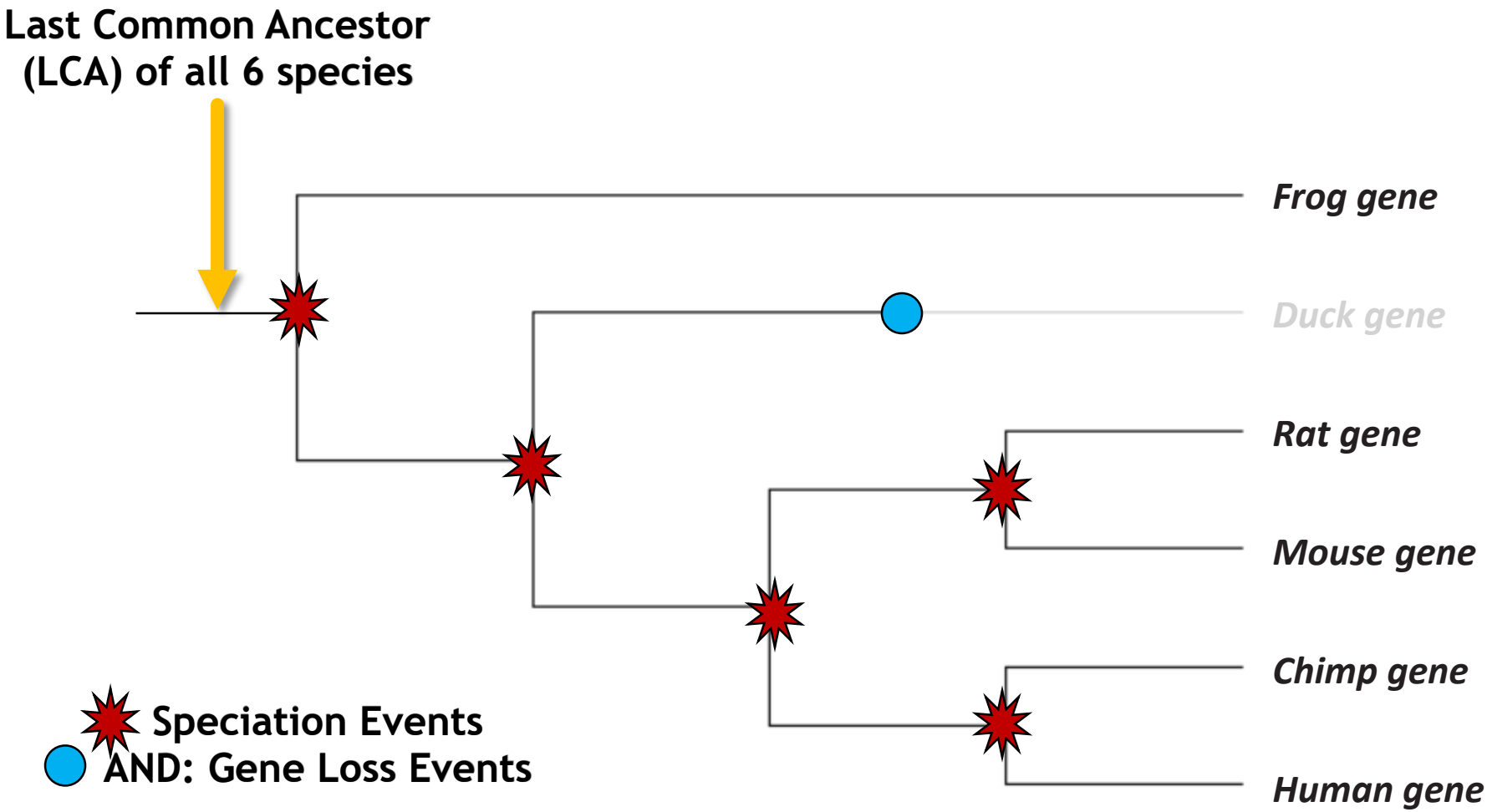
WALTER M. FITCH

Orthology - simple scenario



Single-Copy Orthologs

Evolution ≠ simple

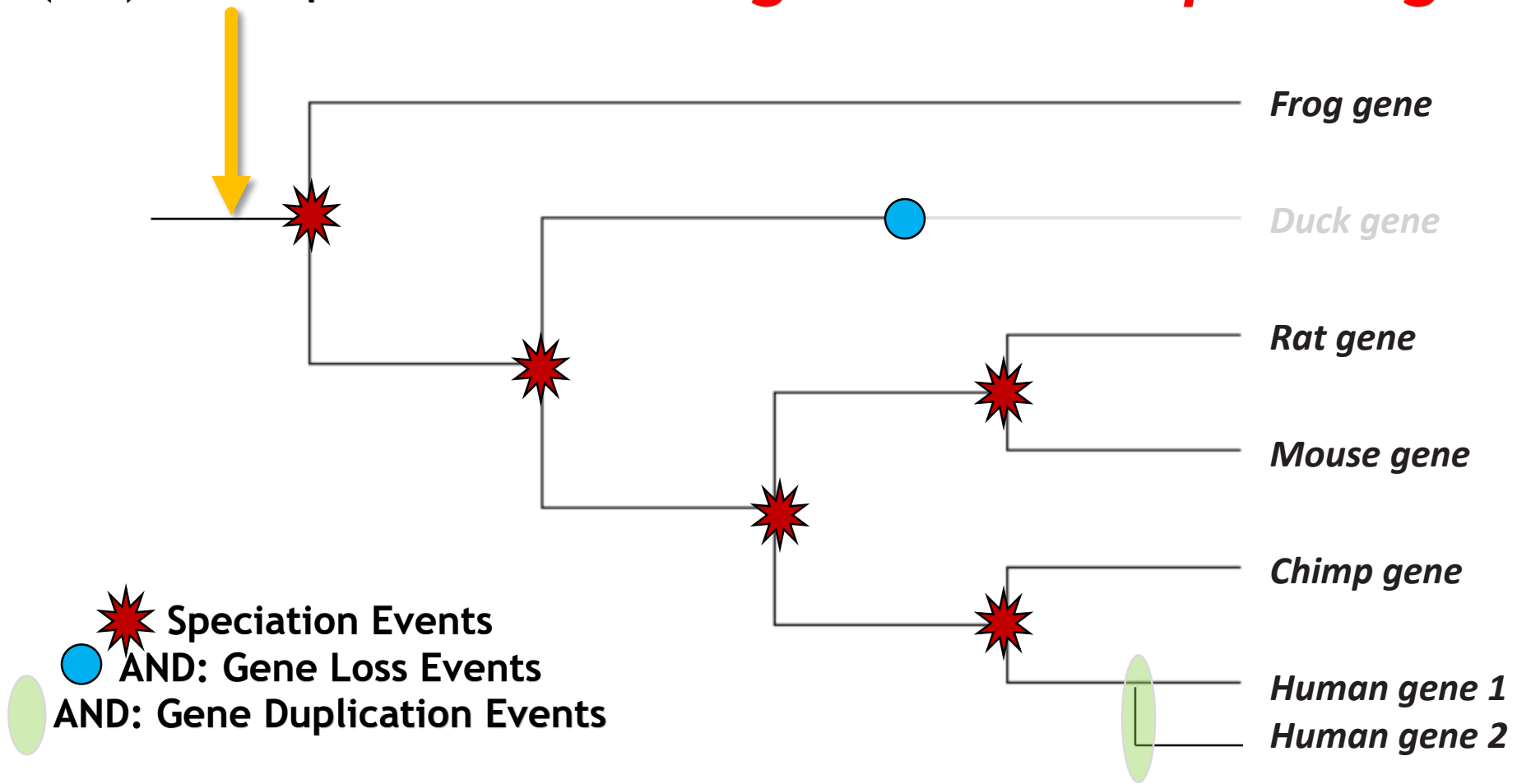


Single-Copy Orthologs with Losses

Evolution ≠ simple

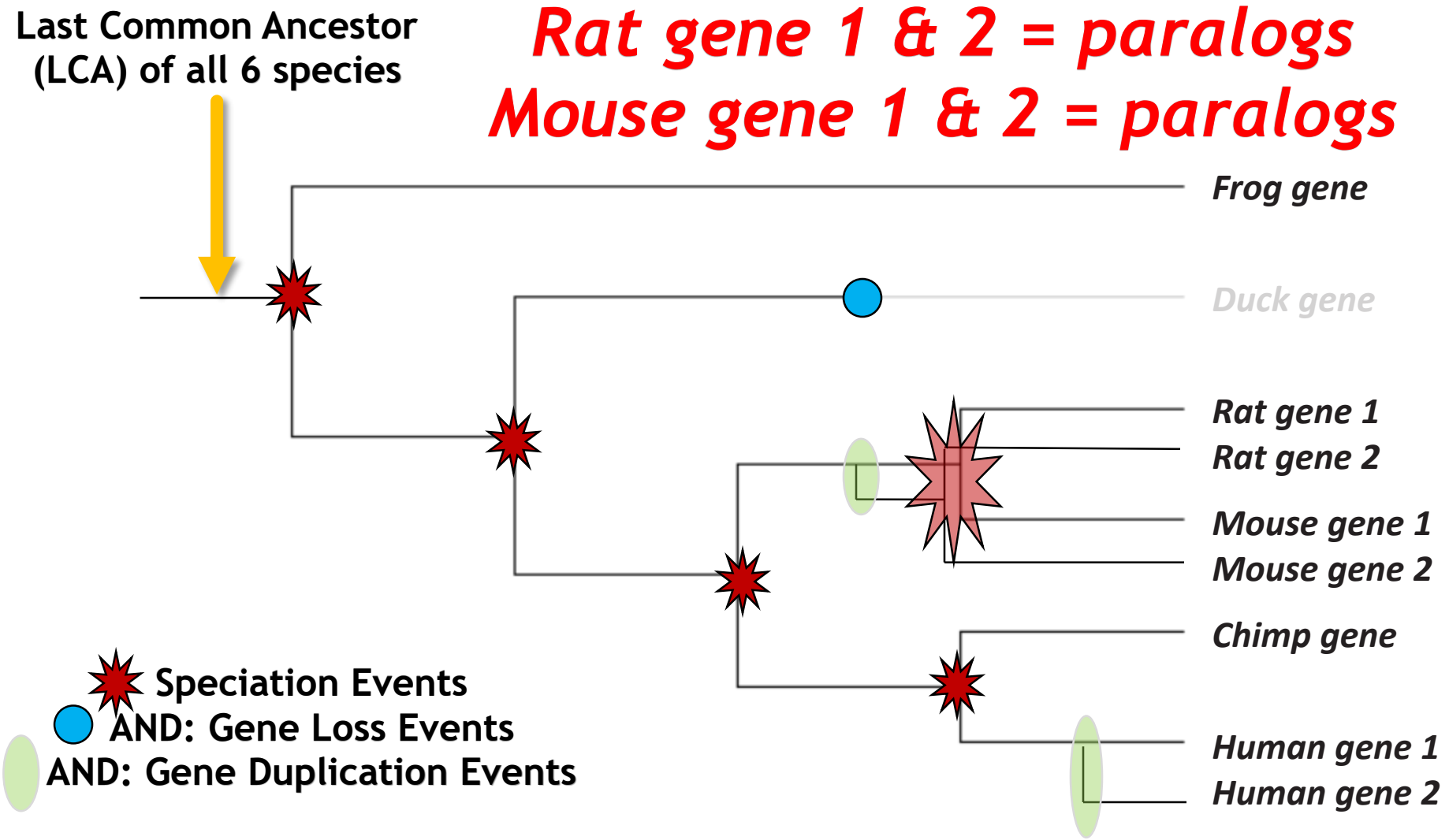
Last Common Ancestor (LCA) of all 6 species

Human gene 1 & 2 = paralogs



Single-Copy Orthologs with Gains

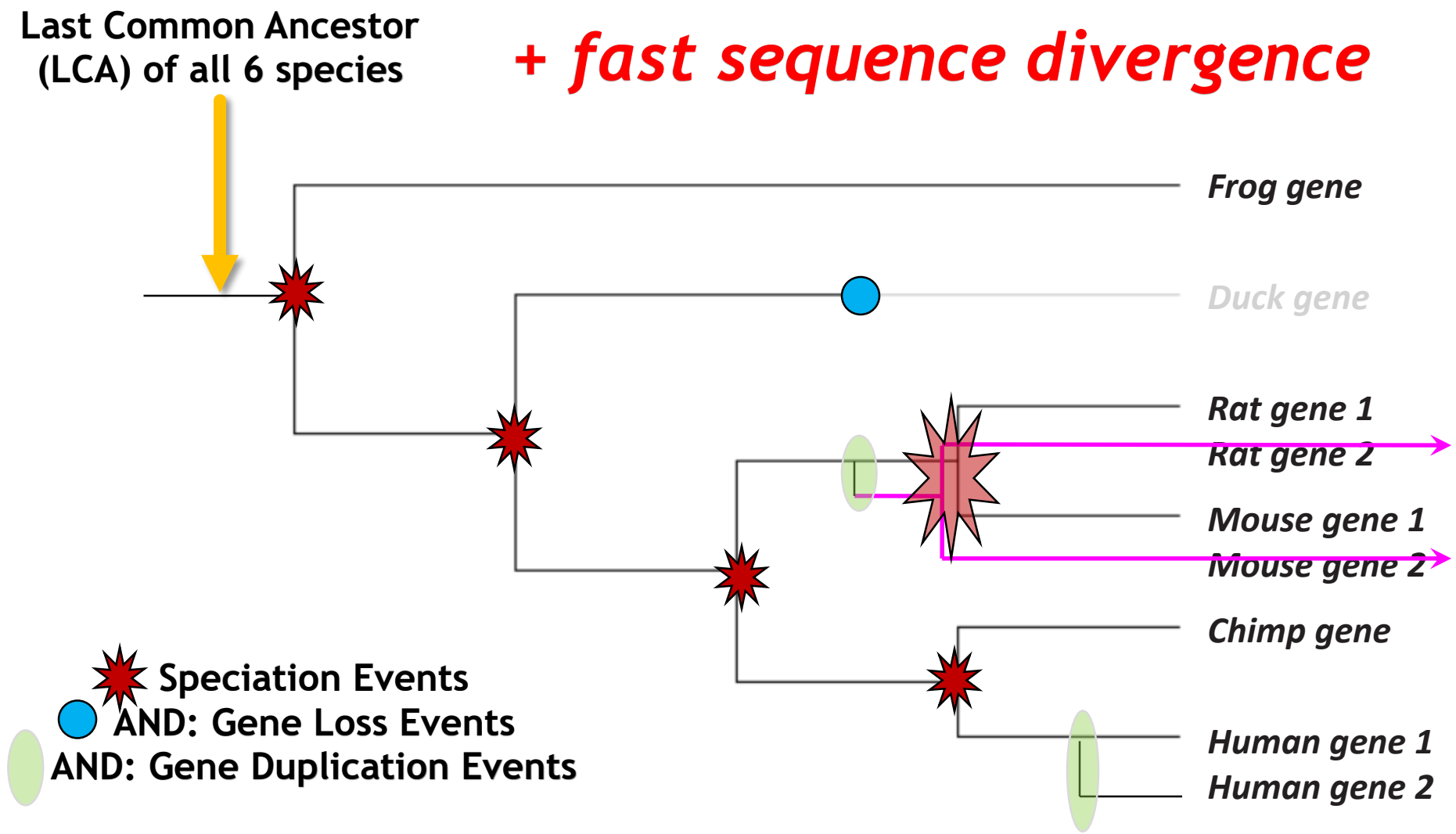
Evolution ≠ simple



Single-Copy Orthologs with Gains

old

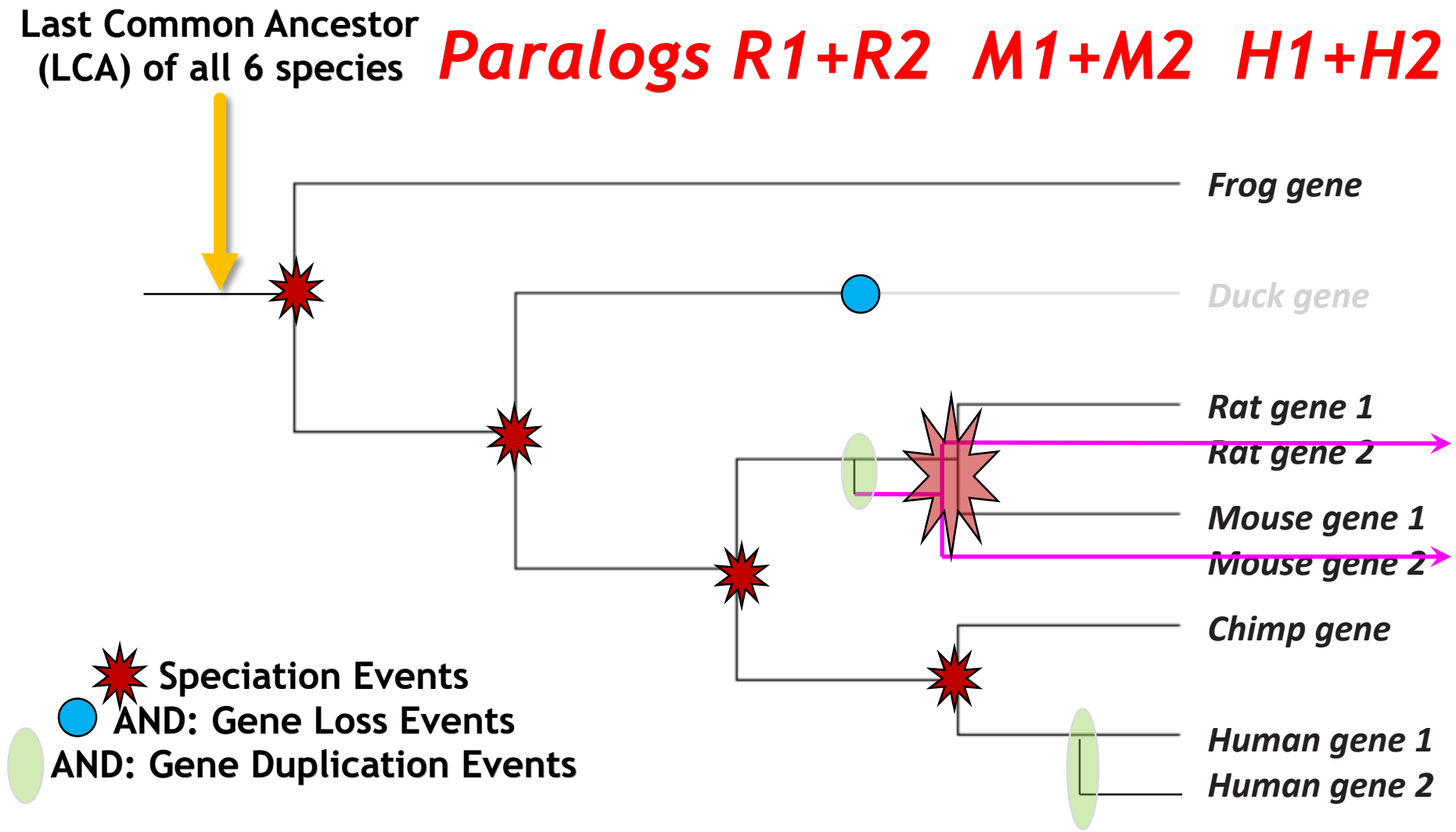
Evolution ≠ simple



Single-Copy Orthologs with Gains

old

Evolution ≠ simple



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

Orthology - what is it?

Homology

Recognizing similarities as evidence of shared ancestry

Orthology

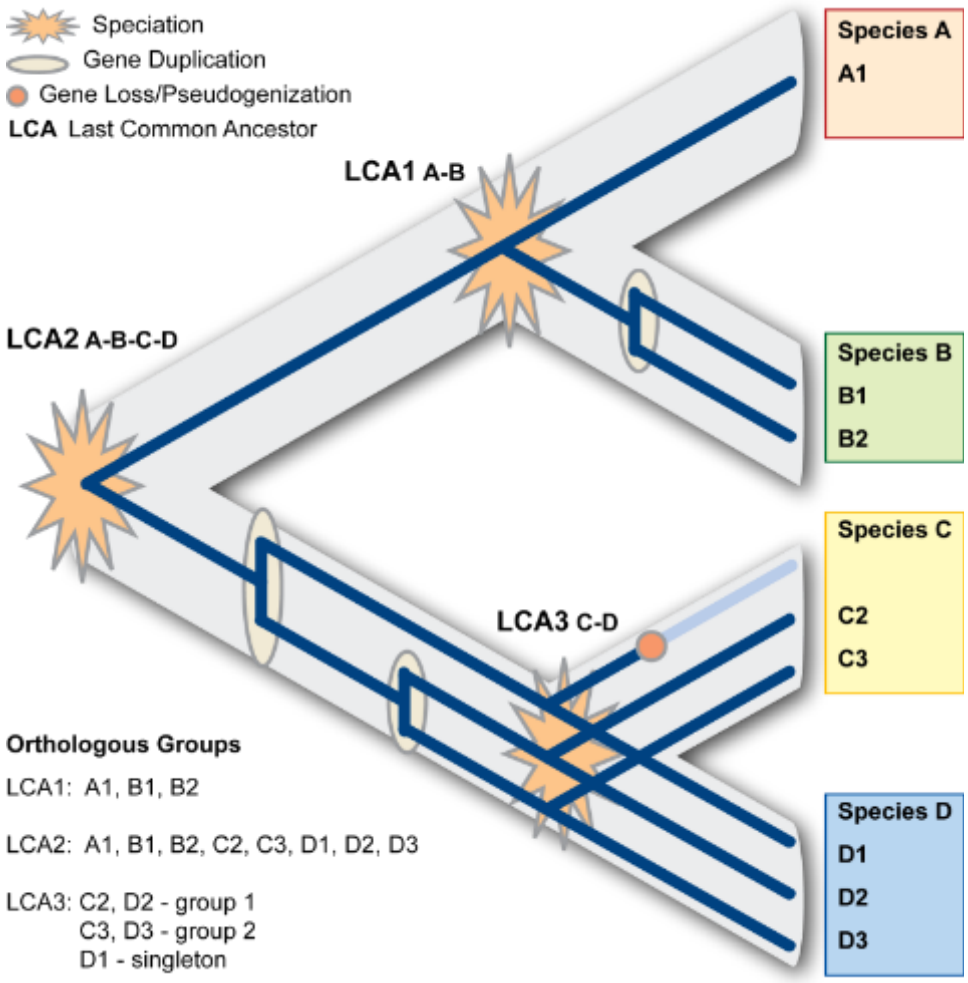
Orthologues arise by vertical descent from a single gene of the last common ancestor

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,*]

How do we delineate Orthology?

Inferring Orthology and Paralogy

Adrian M. Altenhoff and Christophe Dessimoz

Methods Mol Biol. 2012;855:259-79

Abstract

The distinction between orthologs and paralogs, genes that started diverging by speciation versus duplication, is relevant in a wide range of contexts, most notably phylogenetic tree inference and protein function annotation. In this chapter, we provide an overview of the methods used to infer orthology and paralogy. We survey both graph-based approaches (and their various grouping strategies) and tree-based approaches, which solve the more general problem of gene/species tree reconciliation. We discuss conceptual differences among the various orthology inference methods and databases, and examine the difficult issue of verifying and benchmarking orthology predictions. Finally, we review typical applications of orthologous genes, groups, and reconciled trees and conclude with thoughts on future methodological developments.

graph-based approaches

tree-based approaches

How do we delineate Orthology?

tree-based approaches

Table 2
Overview of gene/species tree reconciliation methods and their main properties

Method	Species tree[a]	Rooting[b]	Gene tree uncertainty[c]	Framework[d]	Available Algo/DB	Reference
SDI	Fully resolved	n.a.	None	MP	X/–	(30)
RIO	Fully resolved	min dupl	Bootstrap	MP	–/X[5]	(37)
OrthoStrapper	Fully resolved	min dupl	Bootstrap	MP	X/–	(39)
GSR	Fully resolved	n.a.	n.a.	Probabilistic	X/–	(54, 57)
HOGENOM	Partially resolved	Min dupl	Multifurcate	MP	X/X	(50, 79)
Softparsmap	Partially resolved	Min dupl + min loss	Multifurcate	MP	X/–	(38)
Ensembl/ TreeBeST	Partially resolved	Min dupl + min loss	None	MP	–/X	(31, 32)
LOFT	Species overlap	Min dupl	None	MP	X/–	(33)
PhylomeDB	Species overlap	Outgroup	None	MP	–/X	(34)
BranchClust	Species overlap	Min number of clusters	None	n.a.	–/X	(35)

How do we delineate Orthology?

graph-based approaches

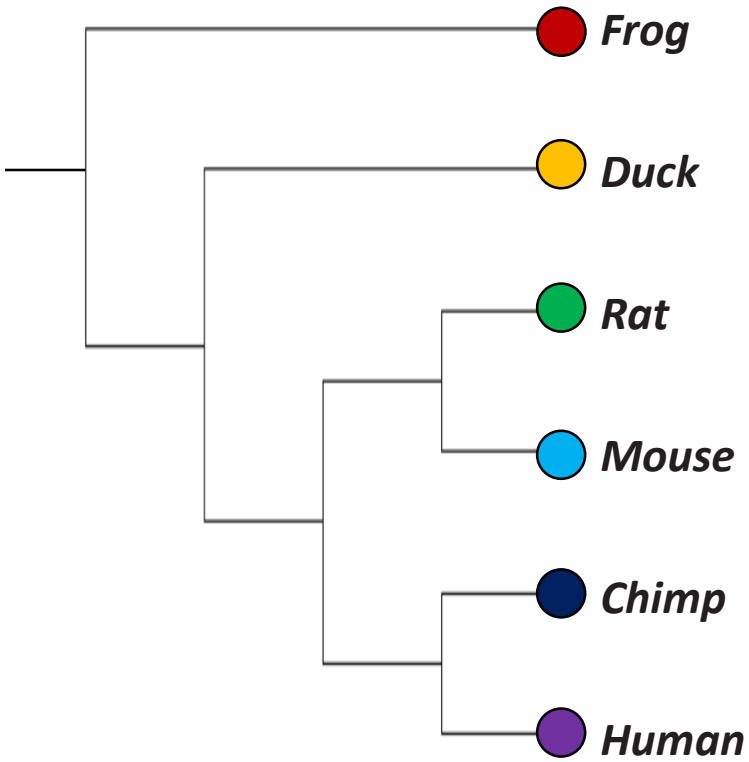
Table 1
Overview of graph-based orthology inference methods and their main properties

Method	In-paralogs	Based on	Grouping strategy	Database	Extra	Available Algo/DB	Reference
COG	Yes	BLAST scores	Merged adjacent triangles of BeTs	COG/KOG		X/X	(6)
BBH	No	BLAST scores	n.a.	n.a.		−/−	(7)
Inparanoid	Yes	BLAST scores	Only between pairs of species	Inparanoid		X/X	(10, 73)
RSD	No	ML distance estimates	n.a.	RoundUp		X/X	(13, 74)
OMA	Yes	ML distance estimates	Every pair is ortholog	OMA Browser	Detects differential gene loss	−/X	(11, 75)
OrthoMCL	Yes	BLAST scores	MCL clusters	OrthoMCL-DB		X/X	(18, 76)
EggNOG	Yes	BLAST scores	Merged adjacent triangles of BeTs	EggNOG	Computed at several levels of taxonomic tree	−/X	(21, 77)
OrthoDB	Yes	Smith Waterman scores	Merged adjacent triangles of BeTs	OrthoDB	Computed at any level of taxonomic tree	−/X	(22)
COCO-CL	Yes	MSA-induced scores	Hierarchical clusters	n.a.		X/−	(23)
OrthoInspector	Yes	BLAST scores	Only between pairs of species	OrthoInspector		X/X	(78)

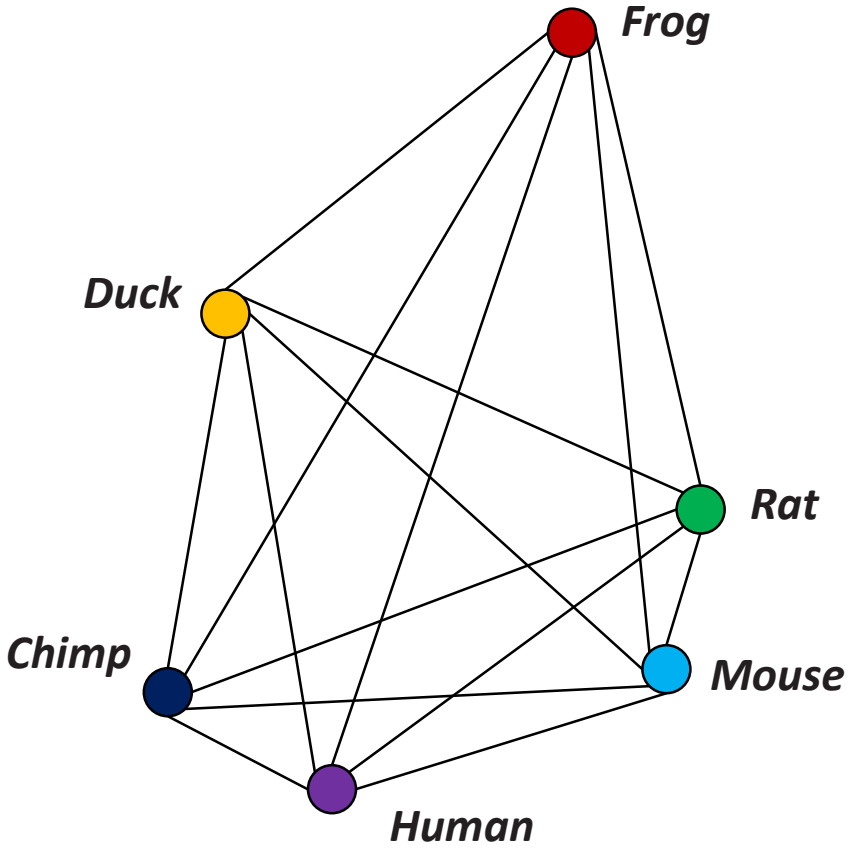
n.a. not applicable

How do we delineate Orthology?

tree-based approaches



graph-based approaches



Single-Copy Orthologs

So which approaches are best?

Standardized benchmarking in the quest for orthologs

Adrian M Altenhoff[1,2], Brigitte Boeckmann[3], Salvador Capella-Gutierrez[4–6], Daniel A Dalquen[7], Todd DeLuca[8], Kristoffer Forslund[9], Jaime Huerta-Cepas[9], Benjamin Linard[10], Cécile Pereira[11,12], Leszek P Pryszcz[4], Fabian Schreiber[13], Alan Sousa da Silva[13], Damian Szklarczyk[14,15], Clément-Marie Train[1], Peer Bork[9,16,17], Odile Lecompte[18], Christian von Mering[14,15], Ioannis Xenarios[3,19,20], Kimmen Sjölander[21], Lars Juhl Jensen[22], Maria J Martin[13], Matthieu Muffato[13], Quest for Orthologs consortium[23], Toni Gabaldón[4,5,24], Suzanna E Lewis[25], Paul D Thomas[26], Erik Sonnhammer[27] & Christophe Dessimoz[7,20,28–30]

OPEN ACCESS Freely available online

PLOS ONE

A Phylogeny-Based Benchmarking Test for Orthology Inference Reveals the Limitations of Function-Based Validation

Kalliopi Trachana[4,9], Kristoffer Forslund[1,9], Tomas Larsson[1,2], Sean Powell[1], Tobias Doerks[1], Christian von Mering[5], Peer Bork[1,3,9]*

So which approaches are best?

Orthology prediction methods: A quality assessment using curated protein families

*Kalliopi Trachana[1)], Tomas A. Larsson[1)2)], Sean Powell[1)], Wei-Hua Chen[1)], Tobias Doerks[1)], Jean Muller[3)4)] and Peer Bork[1)5)]**

Conceptual framework and pilot study to benchmark phylogenomic databases based on reference gene trees

Brigitte Boeckmann, Marc Robinson-Rechavi, Ioannis Xenarios and Christophe Dessimoz

How does OrthoDB delineate Orthology?

OrthoDB

The Hierarchical Catalog of Orthologs ***v9.1***

OrthoDB is a comprehensive catalog of orthologs, i.e. genes inherited by extant species from their last common ancestor. Arising from a single ancestral gene, orthologs form the cornerstone for comparative studies and allow for the generation of hypotheses about the inheritance of gene functions. Each phylogenetic clade or subclade of species has a distinct common ancestor, making the concept of orthology inherently hierarchical. From its conception, OrthoDB explicitly addressed this hierarchy by delineating orthologs at each major species radiation of the species phylogeny. The more closely related the species, the more finely-resolved the gene orthologies.

Read more or cite
"OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs."
Zdobnov EM et al, NAR, Nov 2016, PMID:27899580

Examples of how you can query OrthoDB
Cytochrome P450, protease | peptidase, kinase -serine, FBgn0036816, GO:0006950, immune response, stress response, breast cancer, diabetes.

Help, **Video Presentation** and **Email**: support[at]orthodb.org

Data downloads Protein sequences and orthologous group annotations for major clades.
OrthoDB software Can be used to compute orthologs on custom data.
BUSCO.v3 Assessing completeness of genome assembly and annotation with single-copy genes.

OrthoDB-News Join the mailing list to keep abreast of the latest developments.

Previous OrthoDB Releases

- OrthoDB9 2015: 172 vertebrates, 133 arthropods, 227 fungi, 25 basal metazoans, 3663 bacteria and 31 plants
- OrthoDB8 2014: 61 vertebrates, 87 arthropods, 227 fungi, 12 basal metazoans, and 2627 bacteria
- OrthoDB7 2013: 64 vertebrates, 57 arthropods, 175 fungi, 14 basal metazoans, and 1115 bacteria
- OrthoDB6 2012: 52 vertebrates, 45 arthropods, 142 fungi, 13 basal metazoans, and 1115 bacteria
- OrthoDB5 2011: 48 vertebrates, 33 arthropods, 73 fungi, and 12 basal metazoans

Implementation 1
Input Data

How does OrthoDB delineate orthology?

Gene Sets

UniProt

e! NCBI

JGI DOE JOINT GENOME INSTITUTE

BeetleBase
Tribolium Castaneum

wFleaBase

Annotations

GENEONTOLOGY
Unifying Biology

OMIM®
Online Mendelian Inheritance in Man®
Online Catalog of Human Genes and Genetic Disorders

SGD Saccharomyces GENOME DATABASE

ZFIN
The Zebrafish Model Organism Database

UniProt

MGI

WormBase

InterPro
Protein sequence analysis & classification

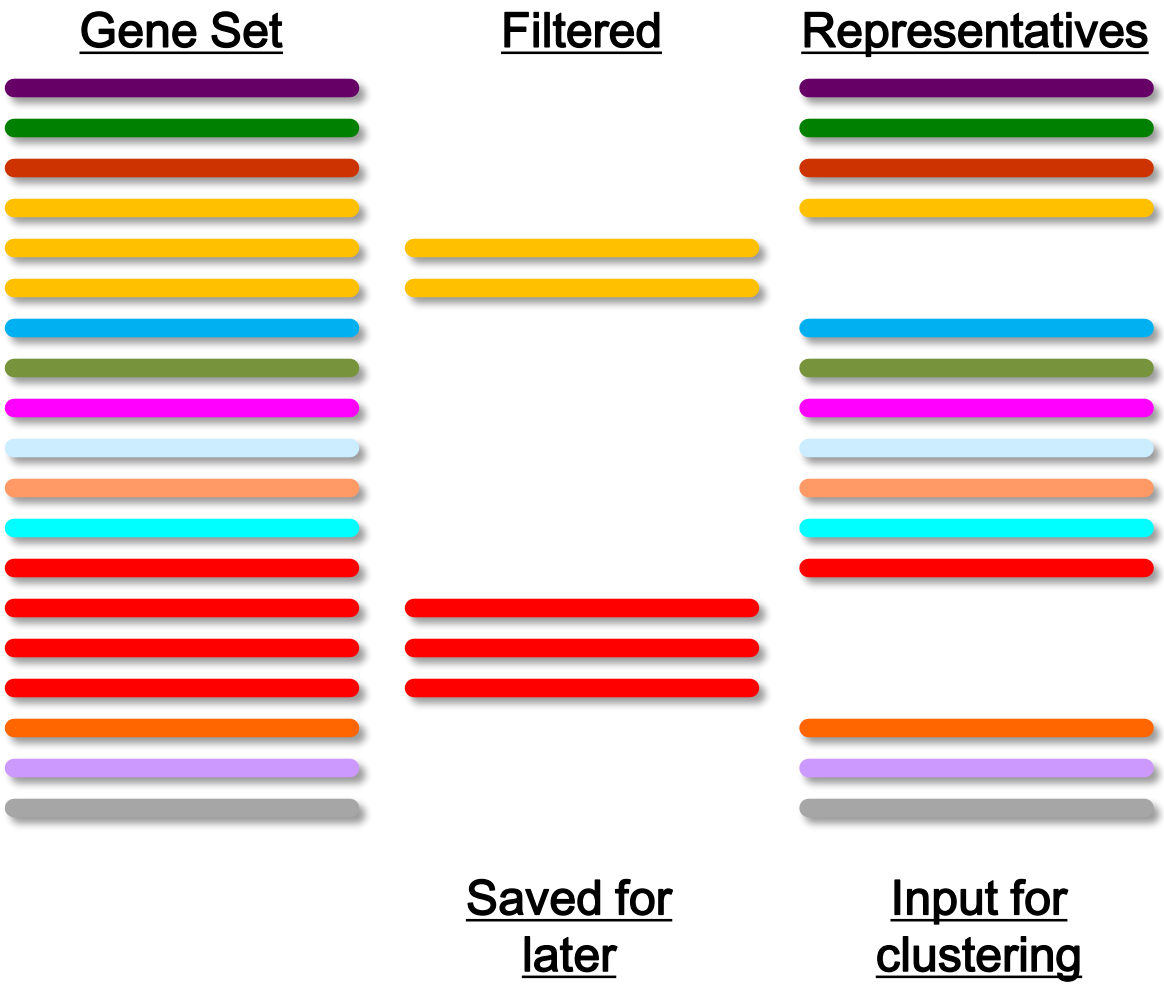
DEG Database of Essential Genes

Implementation 2 Preparation

How does OrthoDB delineate orthology?

A) Select longest protein-coding transcript from any genes with alternative transcripts

B) Remove near-identical proteins from each gene set (97% identity)

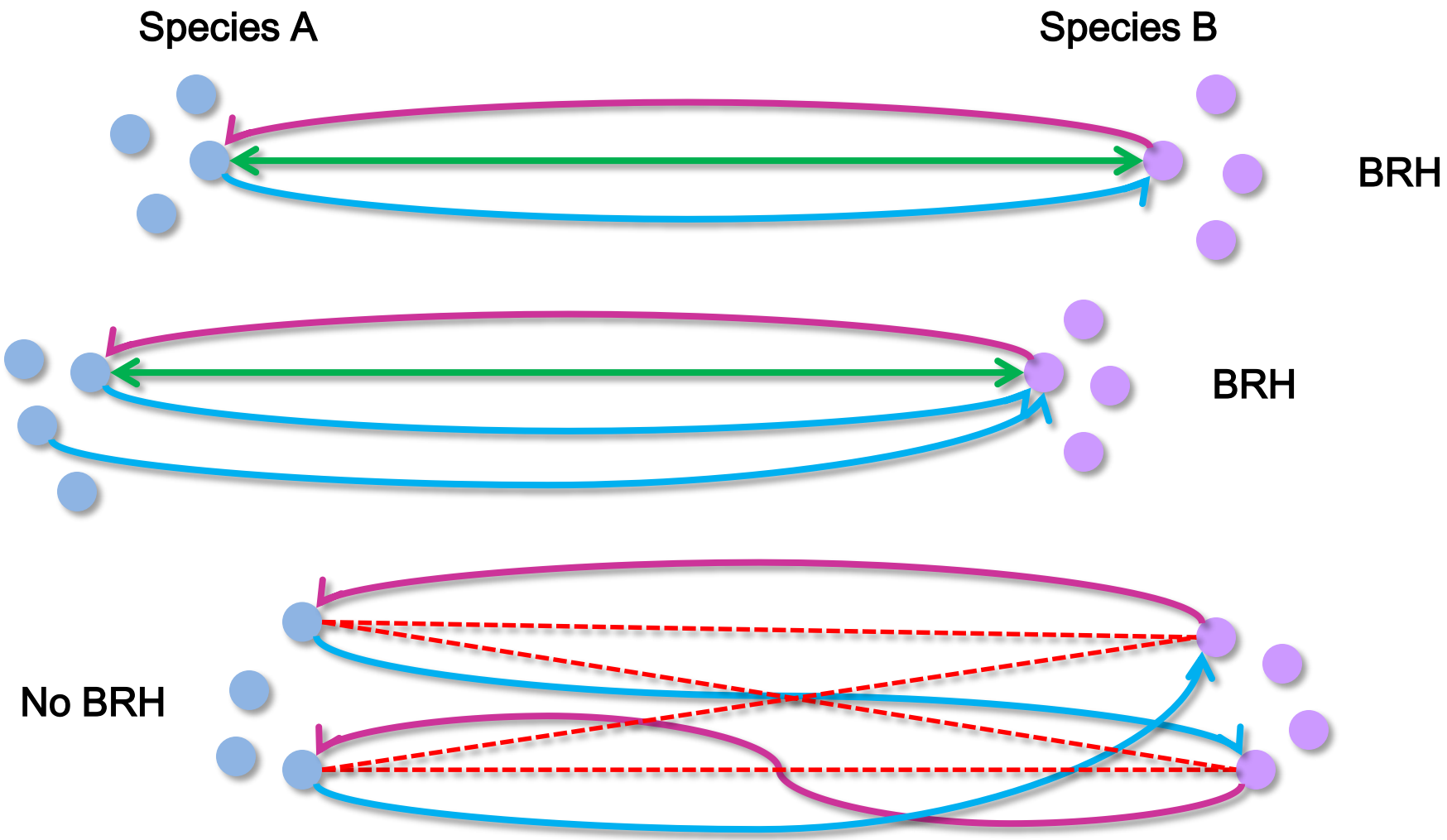


Implementation 3
Best-Reciprocal Hits - BRHs

How does OrthoDB delineate orthology?

A) All-against-all Smith-Waterman pairwise alignments: SWIPE – Rognes 2011.

B) Is best-scoring hit from species A protein to species B protein reciprocal?

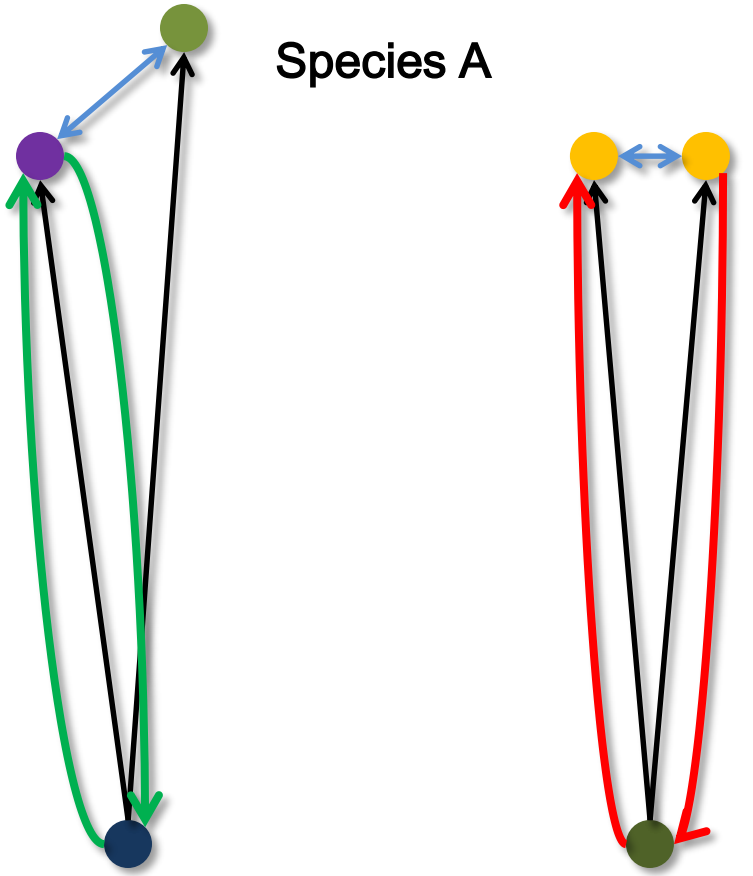


Implementation 4

Reason for filtering near-identical proteins

A) Remove cases of very similar scores

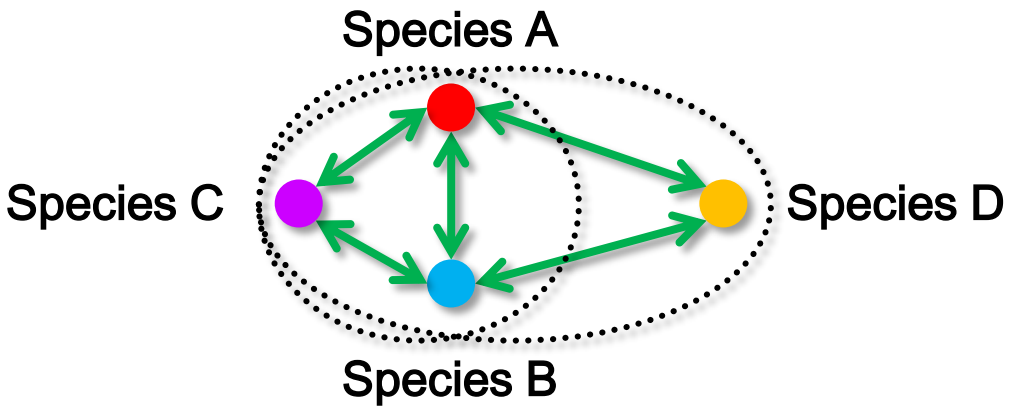
B) Improve BRH recall

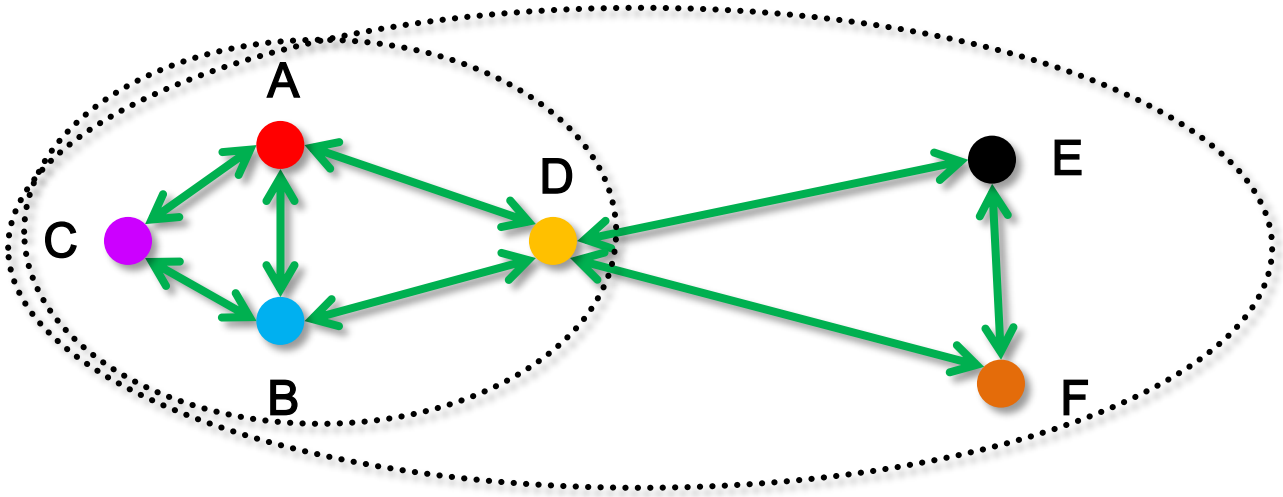


Implementation 5
BRH Triangles

A) Starting with highest-scoring BRHs and moving down the list

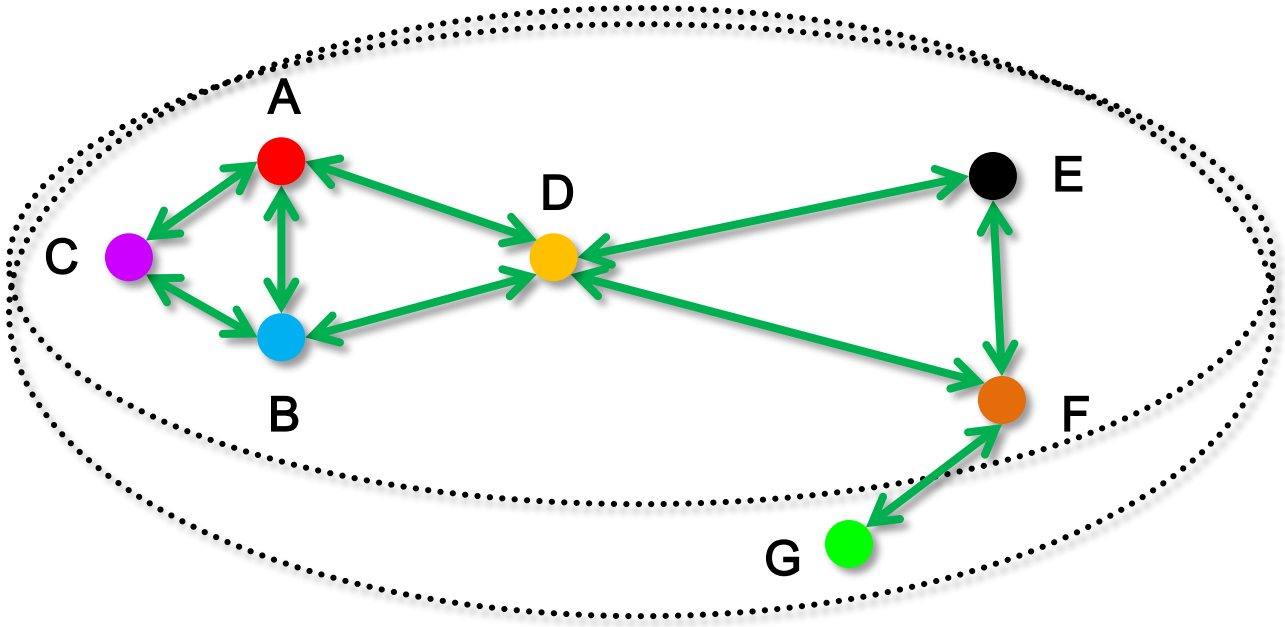
B) BRH Triangles at e<1e-3 cut-off & >20aa alignment overlap





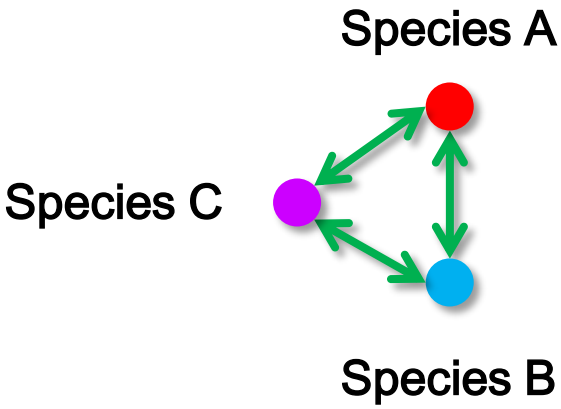
Implementation 6
BRH Pairs

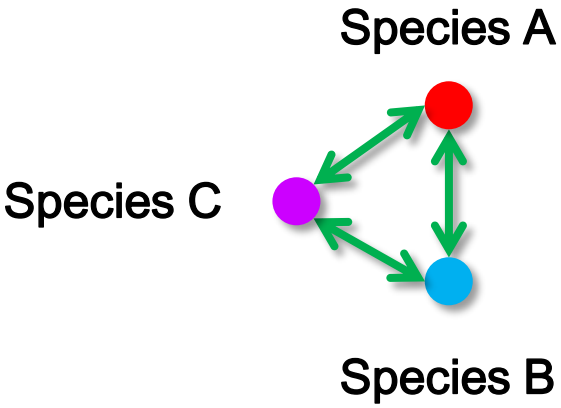
BRHs connected to triangles, but which don't form triangles themselves
=> join clusters with e<1e-6 cut-off & >20aa alignment overlap

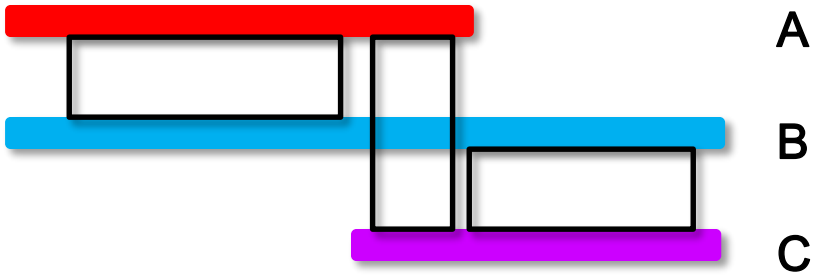


Implementation 7
Alignment Overlap Requirement

>20aa alignment overlap to avoid domain-walking



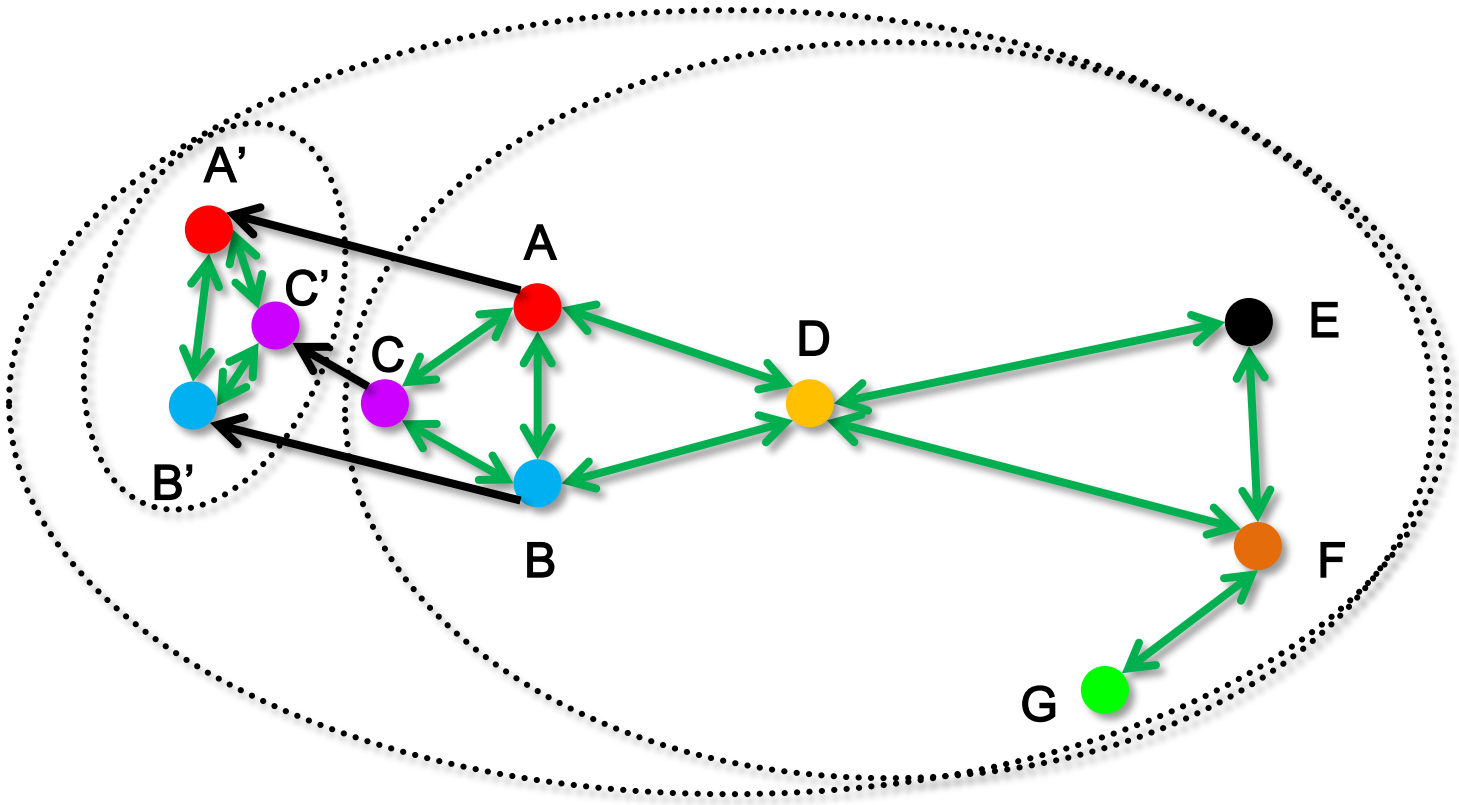


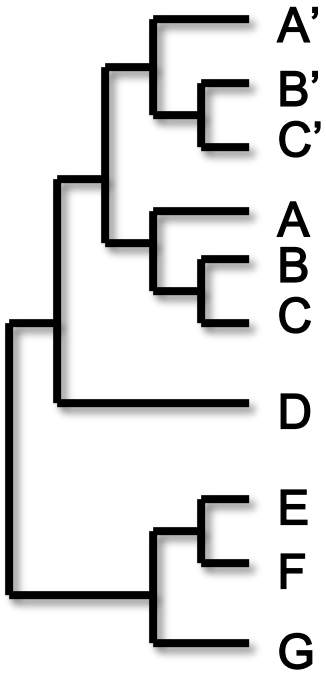


Implementation 8
Inparalogous Groups

How does OrthoDB delineate orthology?

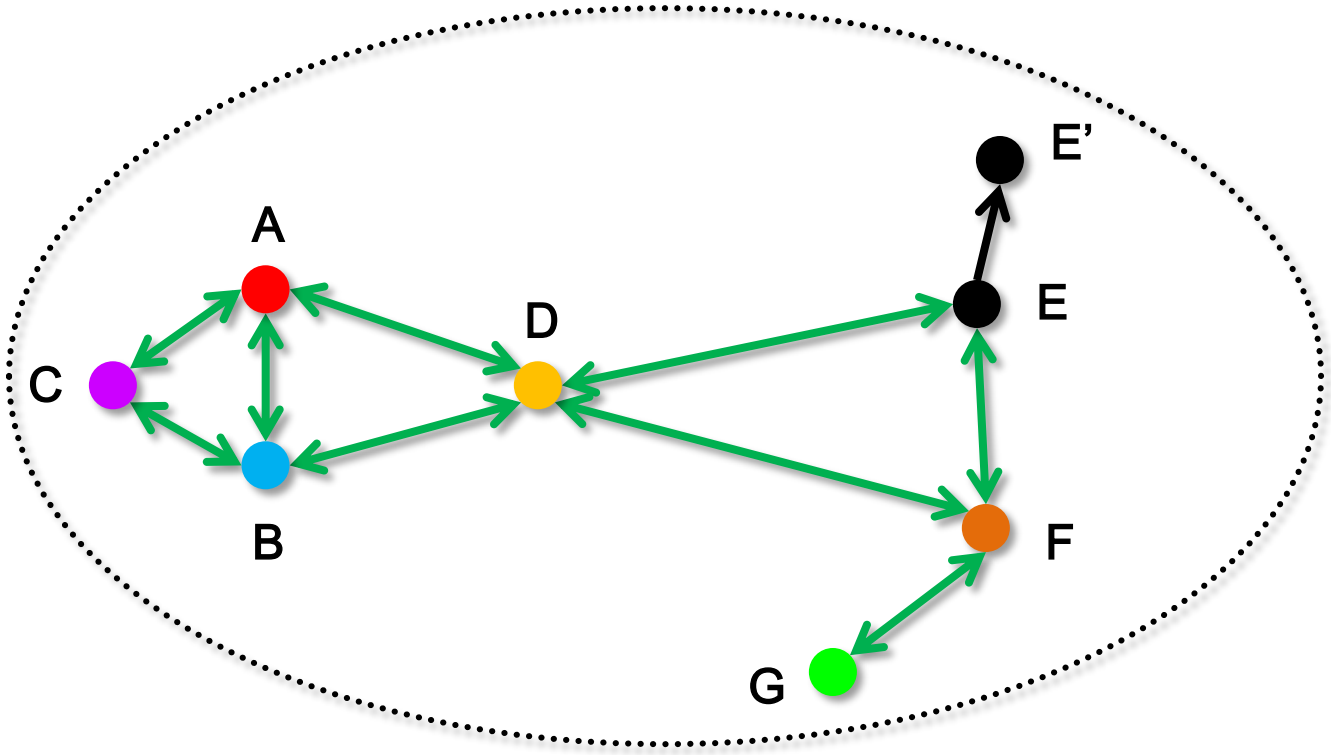
A) Consider within-species homologs in different clusters
B) If the within-species homolog score is better than any within-cluster BRH score, the inparalogous cluster is merged





Implementation 9
Inparalogues

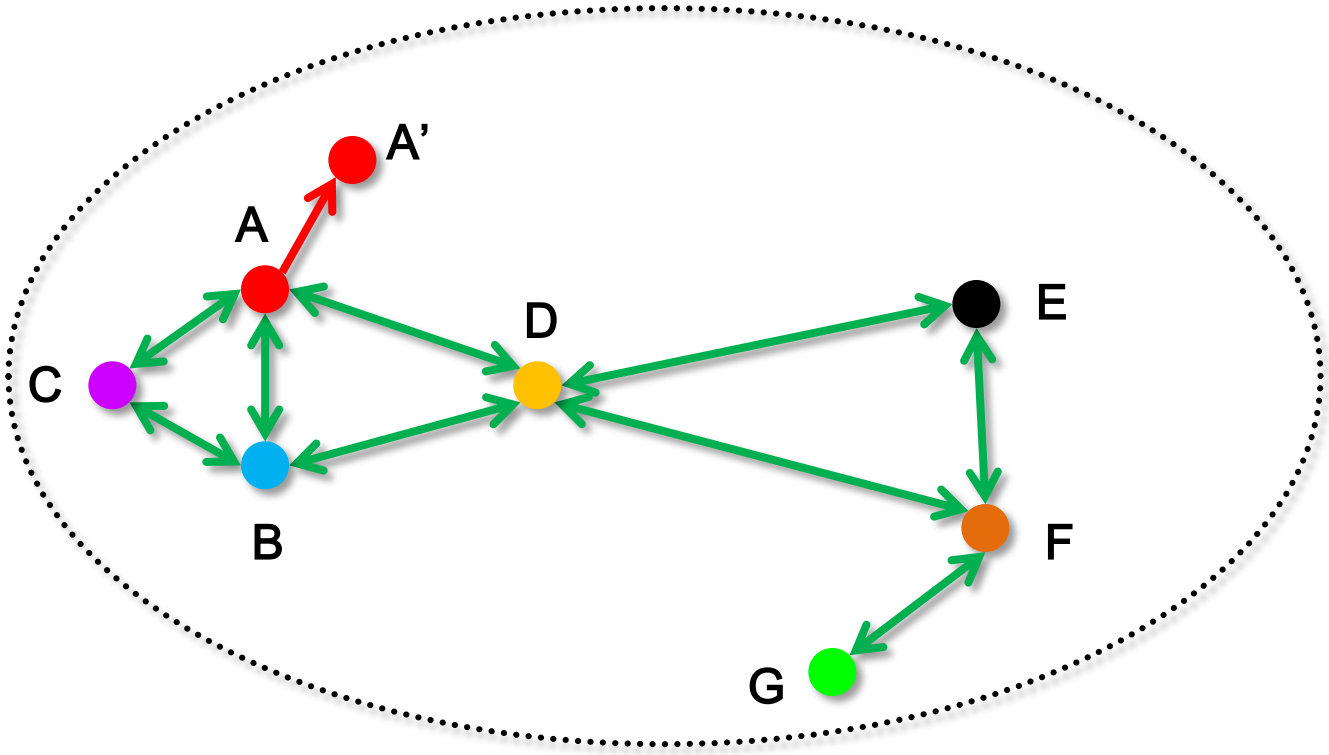
A) Consider within-species homologs that DID NOT get clustered (singletons)
B) If the within-species homolog score is better than any within-cluster BRH score, the singleton is added to the cluster as an inparalogue



Implementation 10
97% paralogues

How does OrthoDB delineate orthology?

A) Add near-identical proteins that were excluded from the clustering
B) If the representative was clustered add the 97% identical proteins to the cluster



Implementation 11
Differential Losses

OrthoDB

A) Rules for complex cases
B) Example: 1 differential loss each

Lost gene B2

OG2 genes will now 'see' gene B1 as their Best Hit & for species D it will be the BRH

OG1 – OG2
Homologous groups

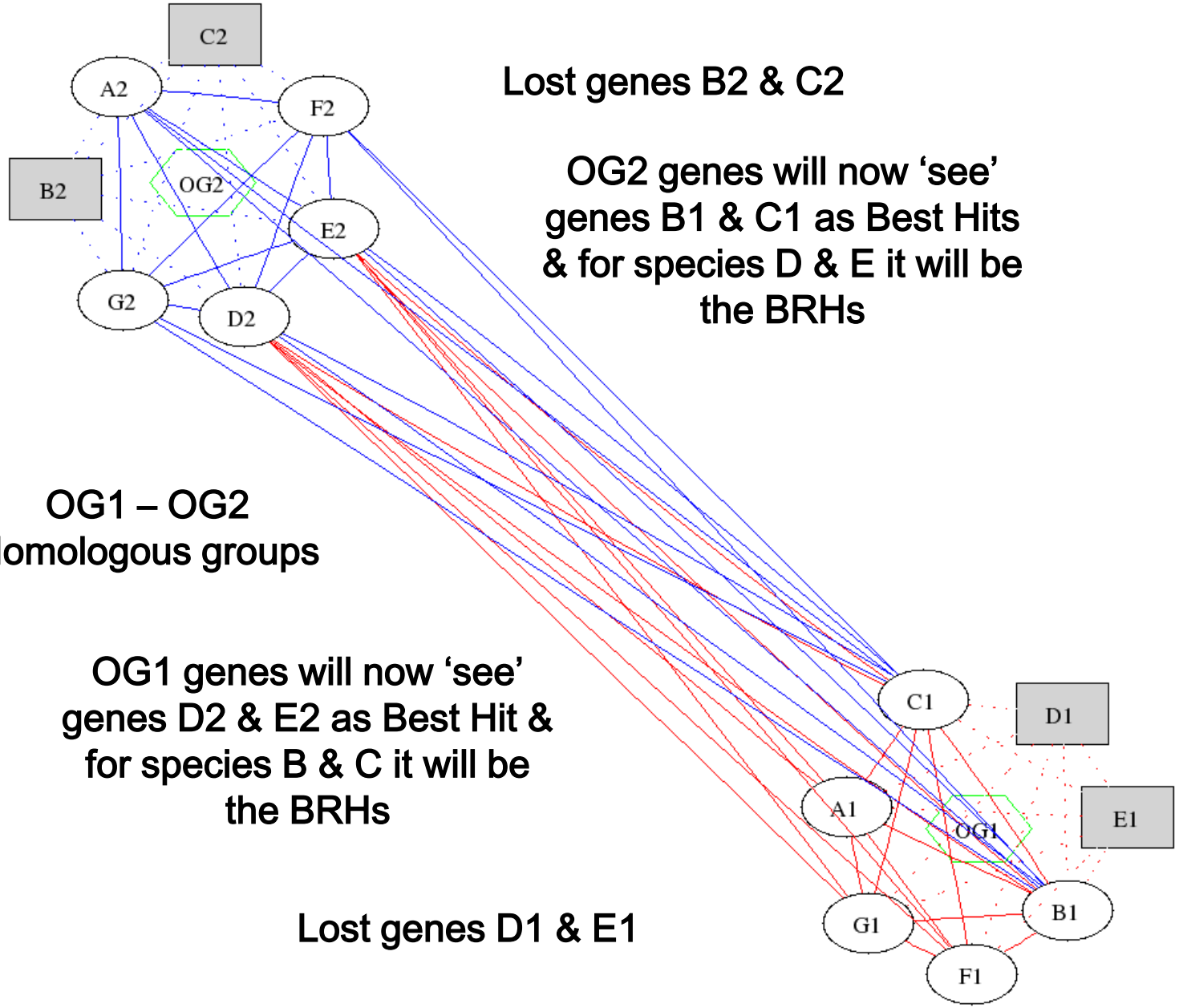
Lost gene D1

OG1 genes will now 'see' gene D2 as their Best Hit & for species B it will be the BRH

Implementation 12
Differential Losses

A) Rules for complex cases
B) Example: 2 differential losses each

OrthoDB



Lost genes B2 & C2

OG2 genes will now 'see' genes B1 & C1 as Best Hits & for species D & E it will be the BRHs

OG1 – OG2
Homologous groups

OG1 genes will now 'see' genes D2 & E2 as Best Hit & for species B & C it will be the BRHs

Lost genes D1 & E1

Implementation 13
Differential Losses

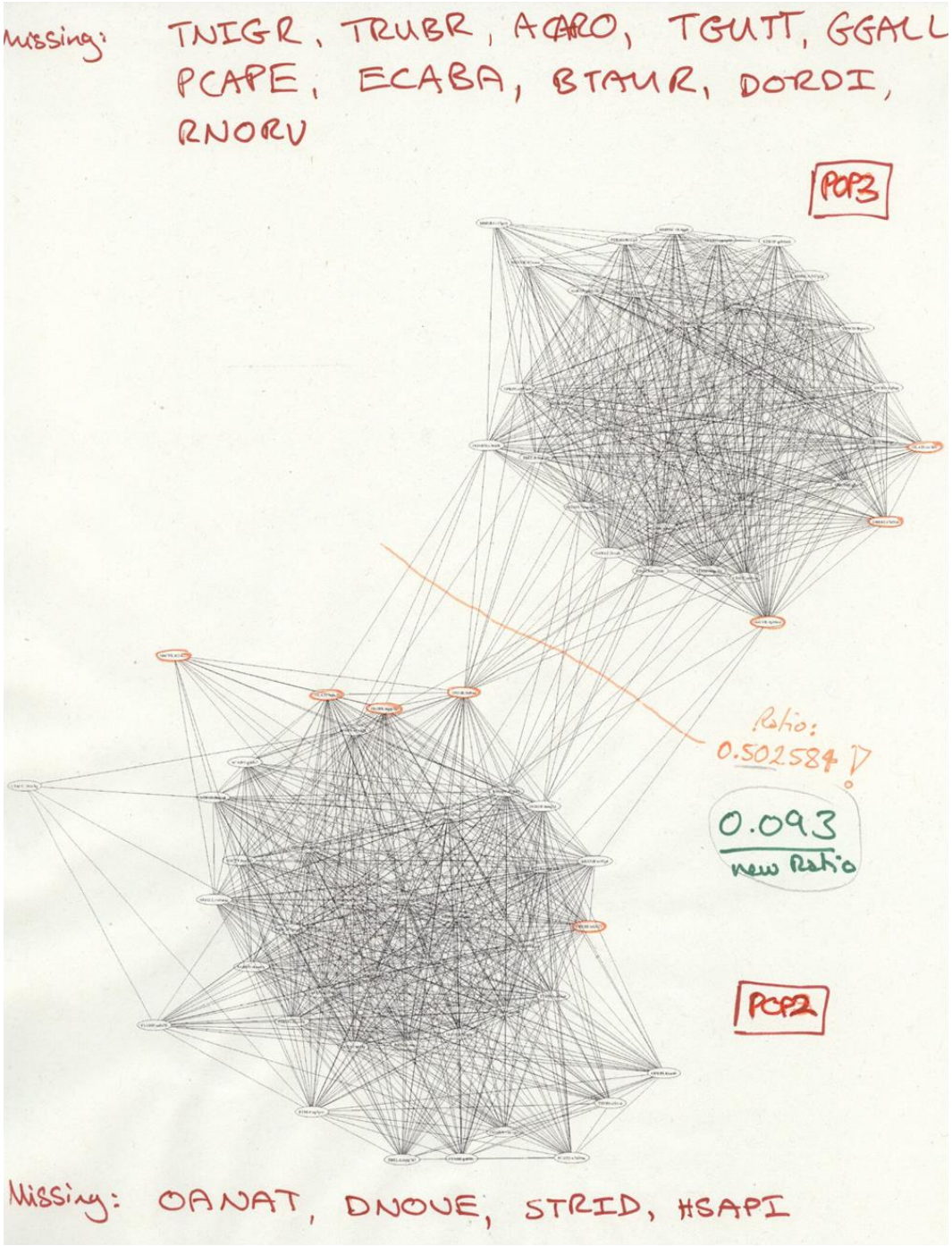
OrthoDB

Real example:

POP3 missing from
10 vertebrates

POP2 missing from
4 vertebrates

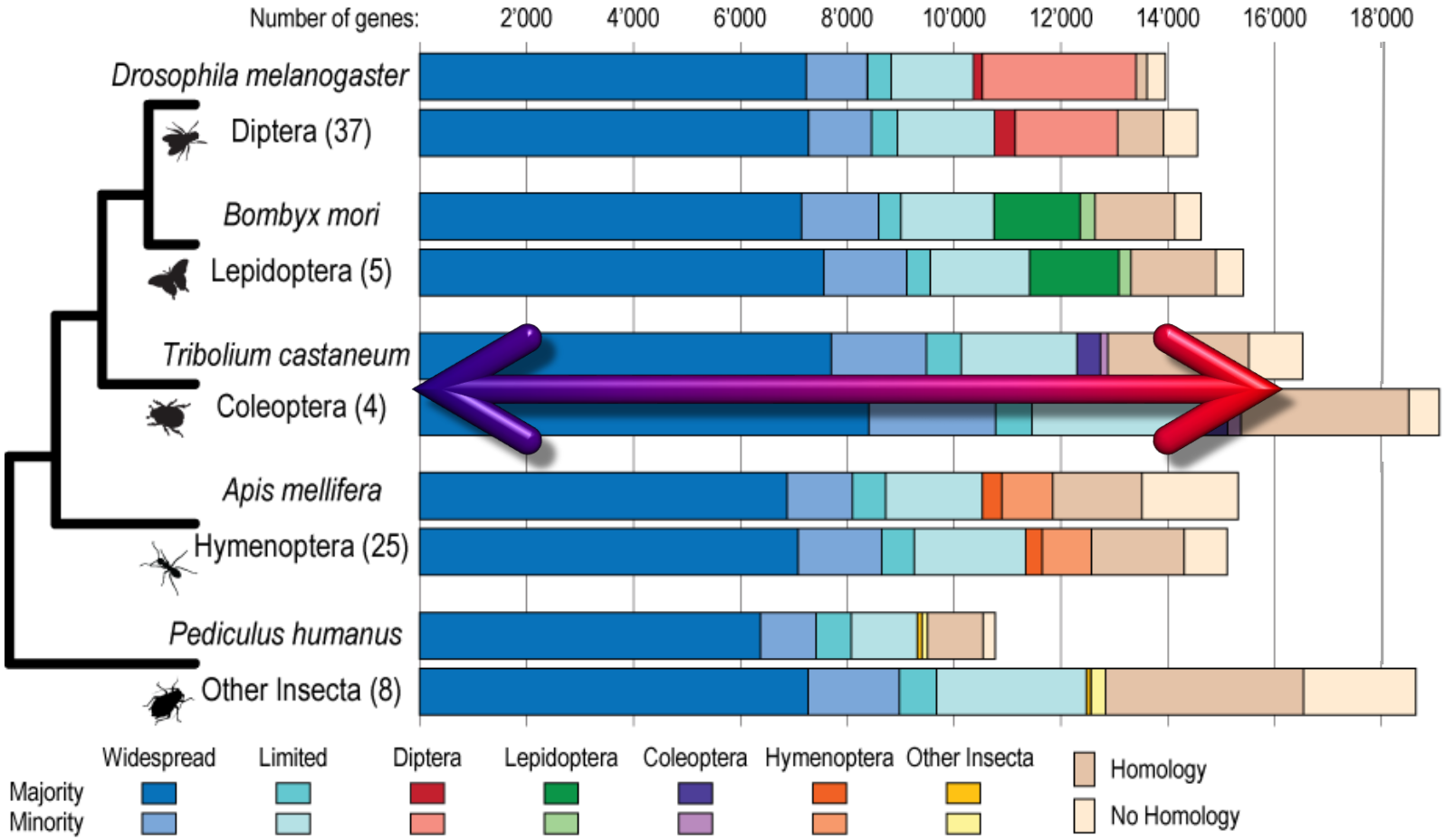
Prevent cluster merges where within-cluster connectivity is much stronger than between cluster connectivity



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

Orthology & Paralogy

… are concepts defined by evolutionary scenarios …

there is nothing in this definition that refers to gene function!

… nevertheless …

Homology refers to common decent, and so generally:

just as the sequences themselves are inherited

so too can the biological functions of the encoded proteins

Orthology ≠ Function

As orthologs share a common ancestry … they can be considered to be "equivalent" genes in different species

Thus, any hypothesis that they share a common function is a relatively reasonable "best guess" assumption

"a crucial property of orthologs, which is both theoretically plausible and empirically supported, is that they typically perform equivalent functions in the respective organisms"

Annu. Rev. Genet.
2005. 39:309–38

Orthology, Paralogy ≠ Function

"As in the case of orthology, the definition of paralogy does not refer to biological function, but there are major functional connotations. Generally, paralogs perform biologically distinct, even if mechanistically related, functions."

Annu. Rev. Genet.
2005. 39:309–38

Resolving the Ortholog Conjecture: Orthologs Tend to Be Weakly, but Significantly, More Similar in Function than Paralogs

Adrian M. Altenhoff[1,2], Romain A. Studer[2,3,4], Marc Robinson-Rechavi[2,3], Christophe Dessimoz[1,2,5]*

'As gene duplication is considered an important source of functional innovation, the "standard model" posits that orthologs tend to have a conserved function, whereas paralogs tend to diverge in function'

Orthology ≠ Function … BUT …

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

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

"The 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

Orthology: What? How? Why?

What is orthology?

How do we delineate orthologs?

And why do we need to?

robert.waterhouse@gmail.com @rmwaterhouse www.rmwaterhouse.org