

Orthology Inference

Christophe Dessimoz
cdessimoz@inf.ethz.ch



*Bioinformatics and Comparative Genome Analysis
Institut Pasteur Paris, July 2010*

ETH Zürich



Outline

- **Motivation and Definition**
- **Orthology Inference**
 - Pairwise methods
 - Tree reconciliation methods
- **Verification**
- **Limitations and Future Directions**

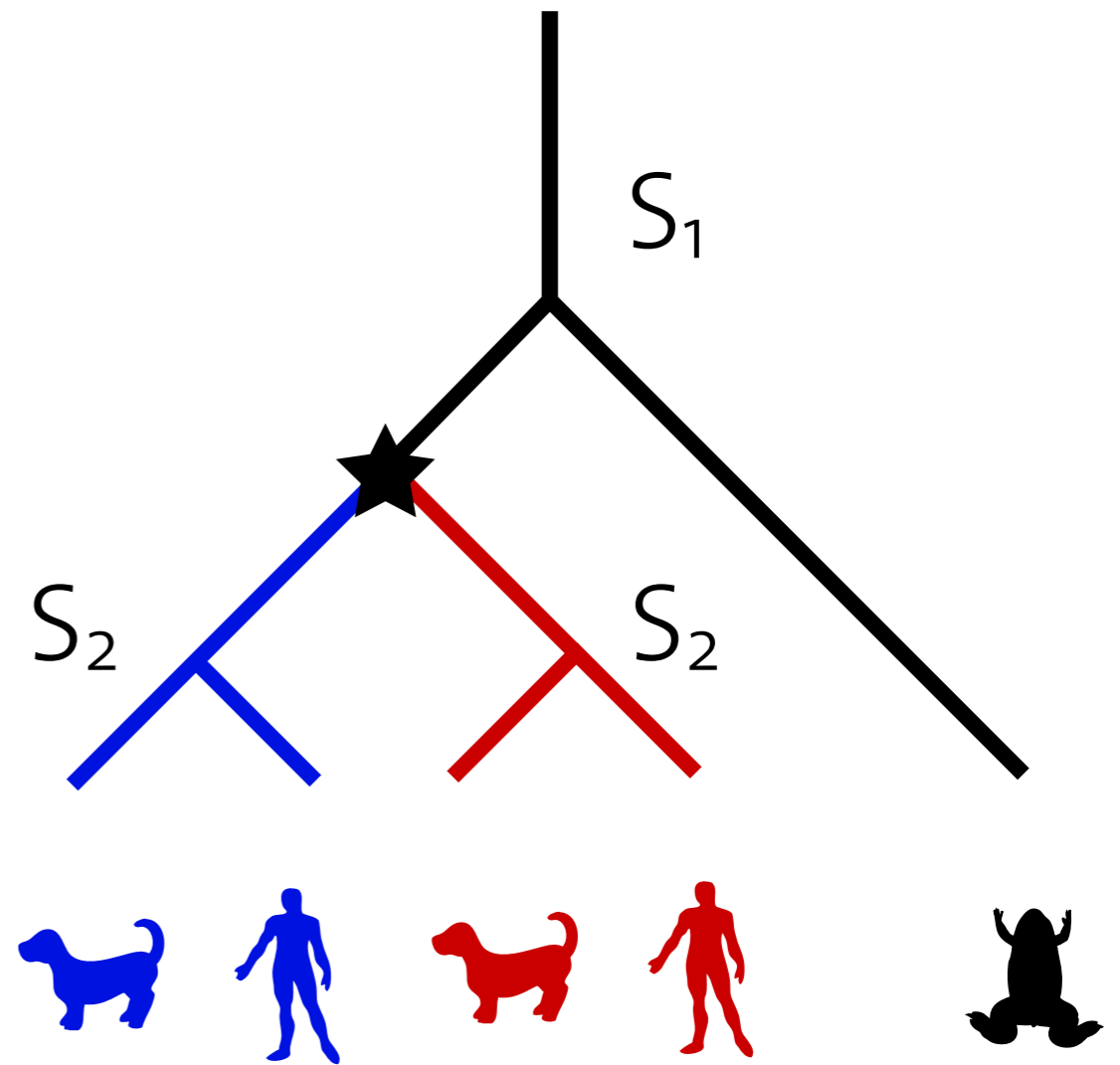
Motivation and Definition

Definition

DISTINGUISHING HOMOLOGOUS FROM ANALOGOUS PROTEINS

WALTER M. FITCH

the mammals (Fitch and Margoliash 1967). Therefore, there should be two subclasses of homology. Where the homology is the result of gene duplication so that both copies have descended side by side during the history of an organism, (for example, α and β hemoglobin) the genes should be called *paralogous* (para = in parallel). Where the homology is the result of speciation so that the history of the gene reflects the history of the species (for example α hemoglobin in man and mouse) the genes should be called *orthologous* (ortho = exact). Phylogenies require orthologous, not paralogous, genes. Note that the present method does not permit



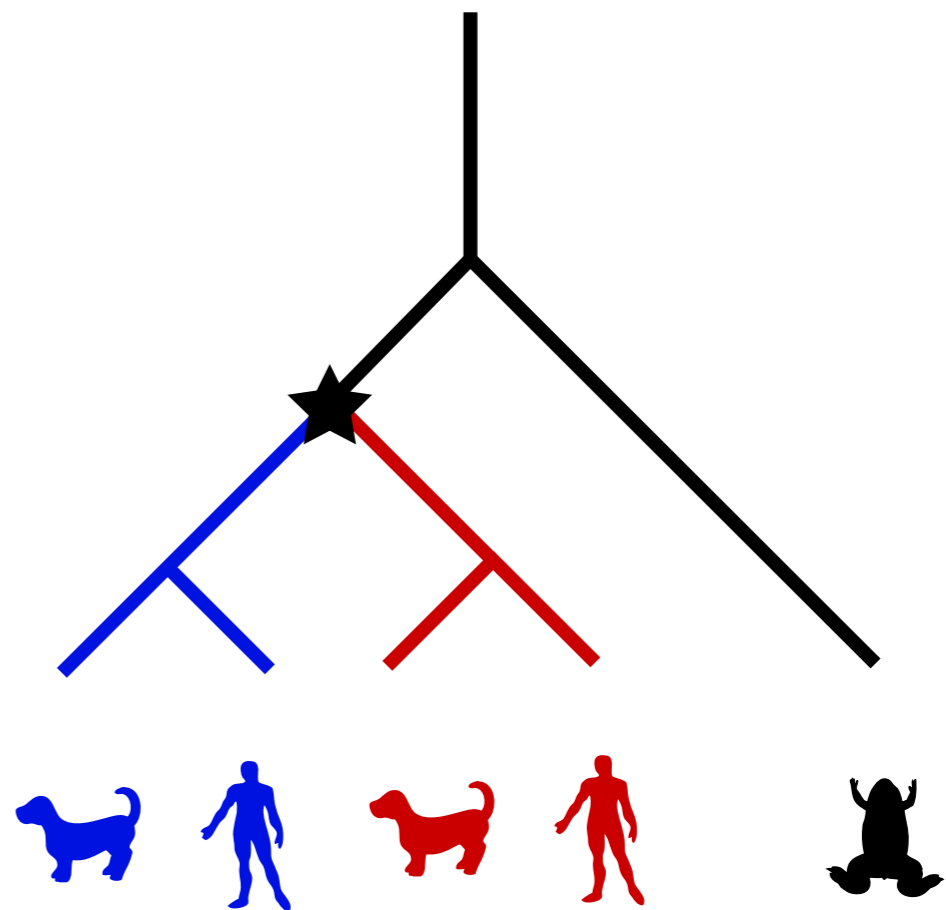
Systematic Zoology, Vol. 19, No. 2 (Jun., 1970), pp. 99-113

Formally

Two homologous genes (x,y) are orthologs if they started diverging through a speciation event.

Observations

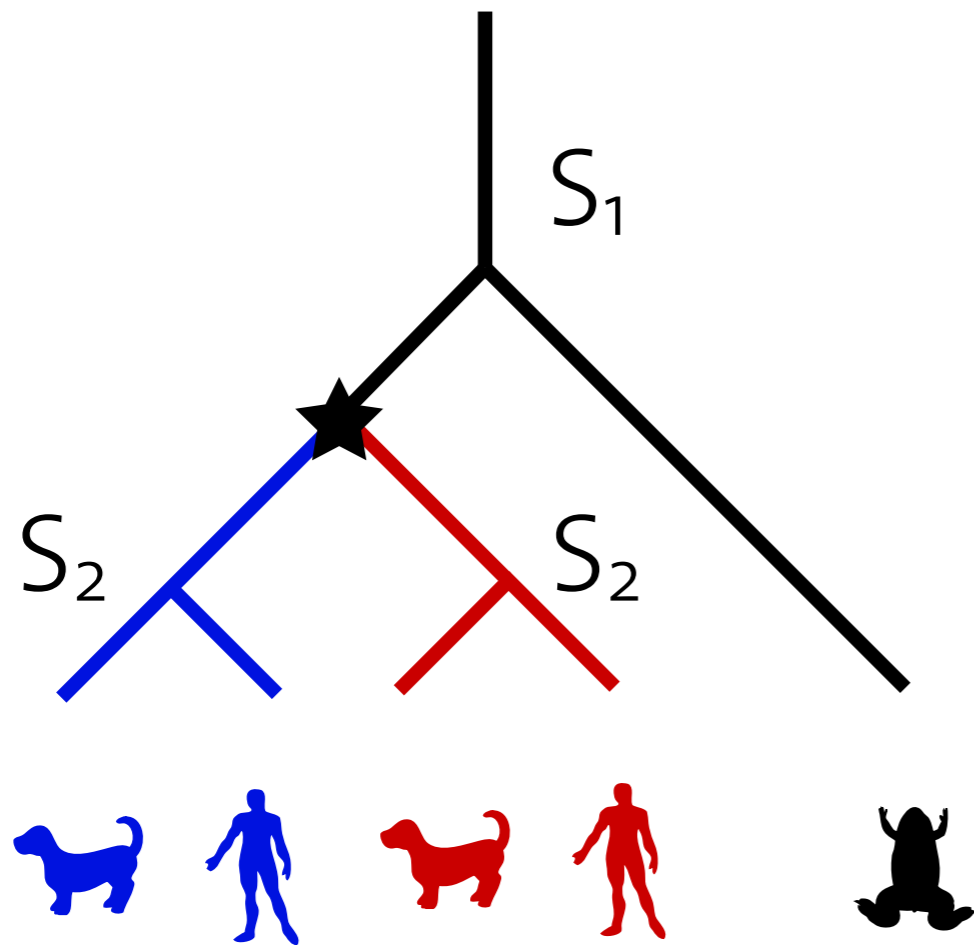
- relation defined on a *pair* of genes
- non-transitive





Why useful?

- **Species Tree Reconstruction**
- **Functional annotation**
Prevalent model:
 - orthologs share similar function
 - paralogs have different functions
- **Physical Mapping between genomes**
- ...

More terminology

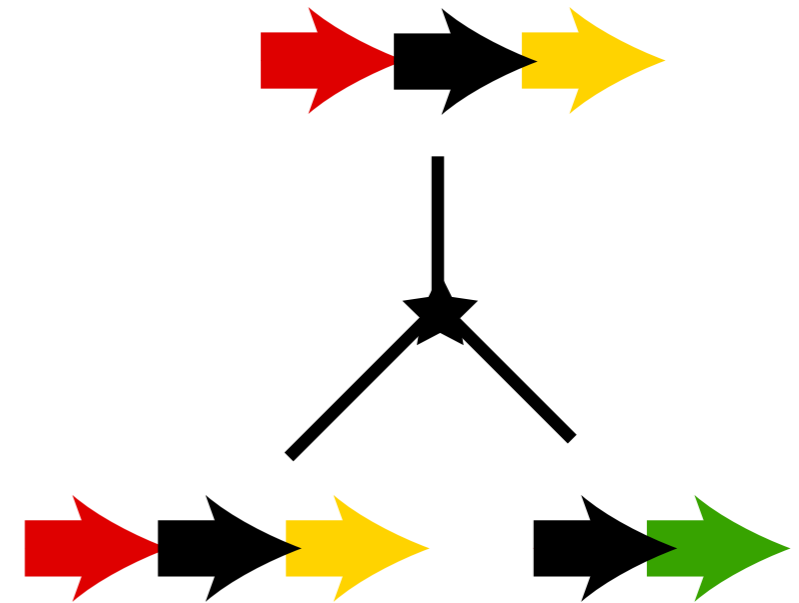


in-paralogs( ,  , S_1)

out-paralogs( ,  , S_2)

Other usages of orthology*

- Genes with the same *function*
 - “isofunctional homologs”
 - “equivalogs”
- Homologs that are in the same genomic context
 - “positional ortholog”



*not recommended

Inference

Homology

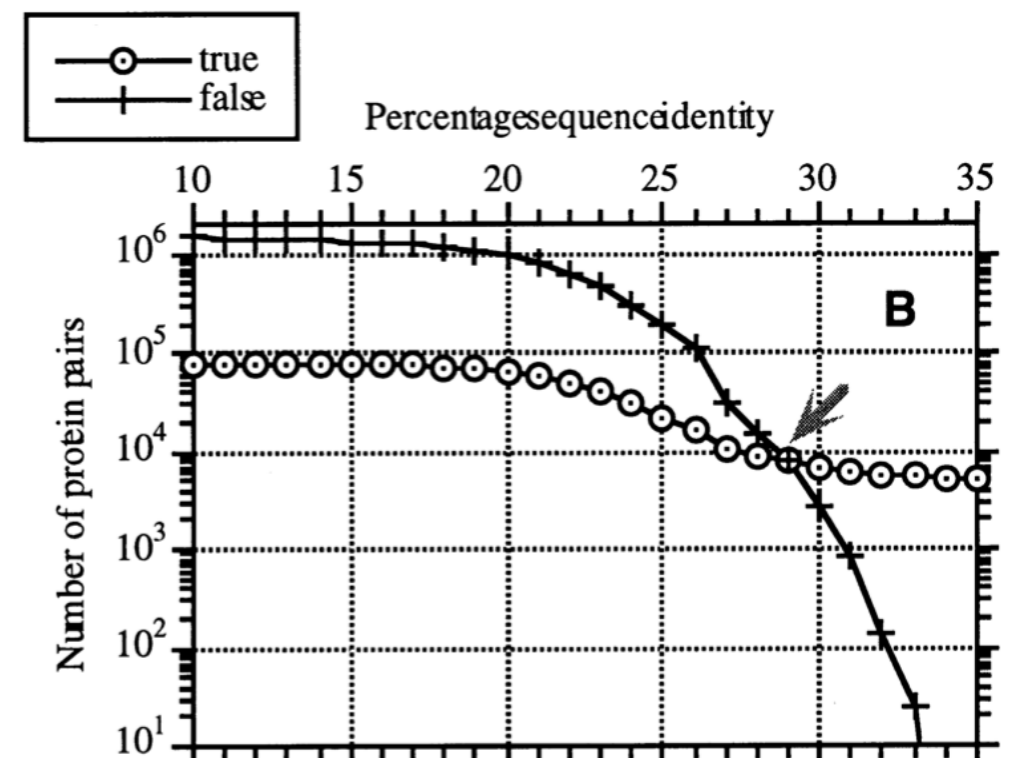
- Commonly inferred by sequence similarity:
 - “All-against-all”
 - Profile-based search
- At low similarity (20-30% identity, “twilight zone”), **protein structure** tends to be better conserved, *but*



structure often unknown



only for conserved regions



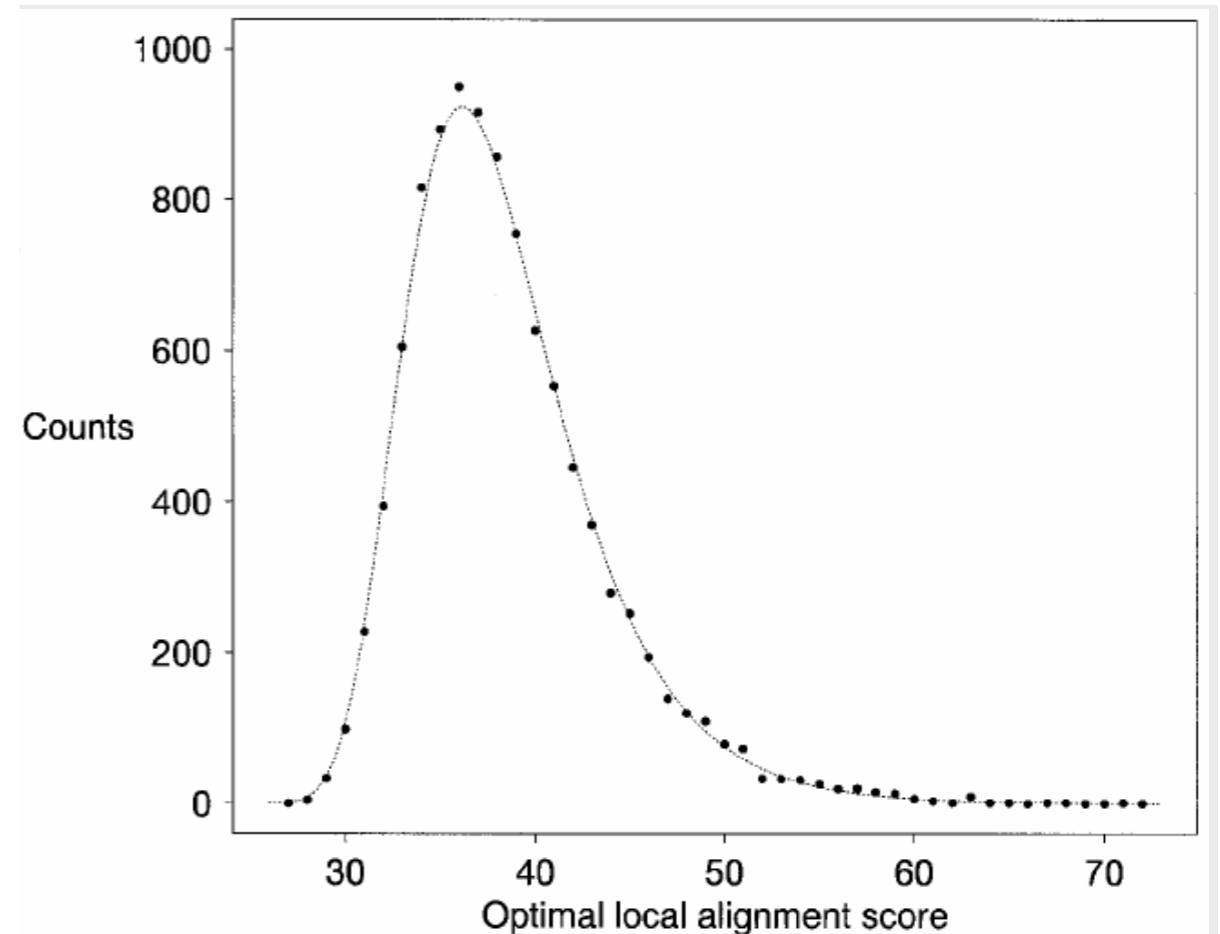
B Rost, *Protein Engineering* vol.12 no.2 pp.85-94, 1999

Pairwise alignments statistics

Alignment score of 2 unrelated sequences are distributed according to Gumbel distribution (2 parameters, fat tail)

Params are estimated from seq lengths and scoring matrix (Karlin-Altschul theory)

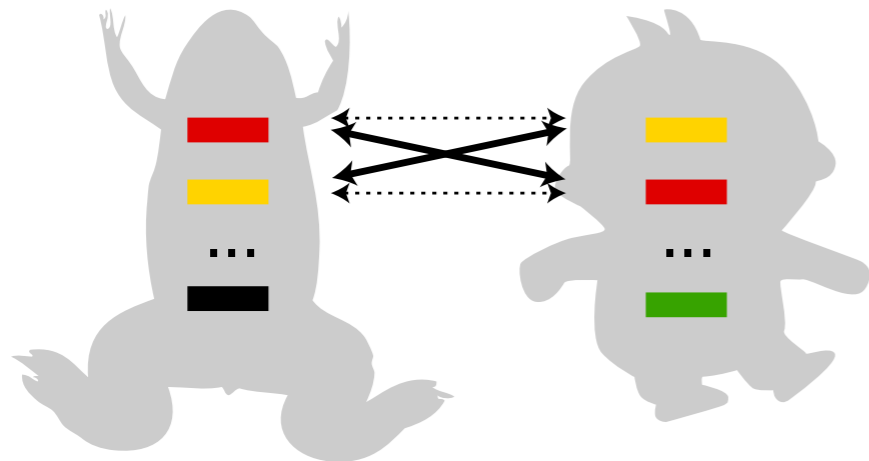
Significance is assessed by its **E-value** (expected # spurious matches with score equal or higher), computed from probability density



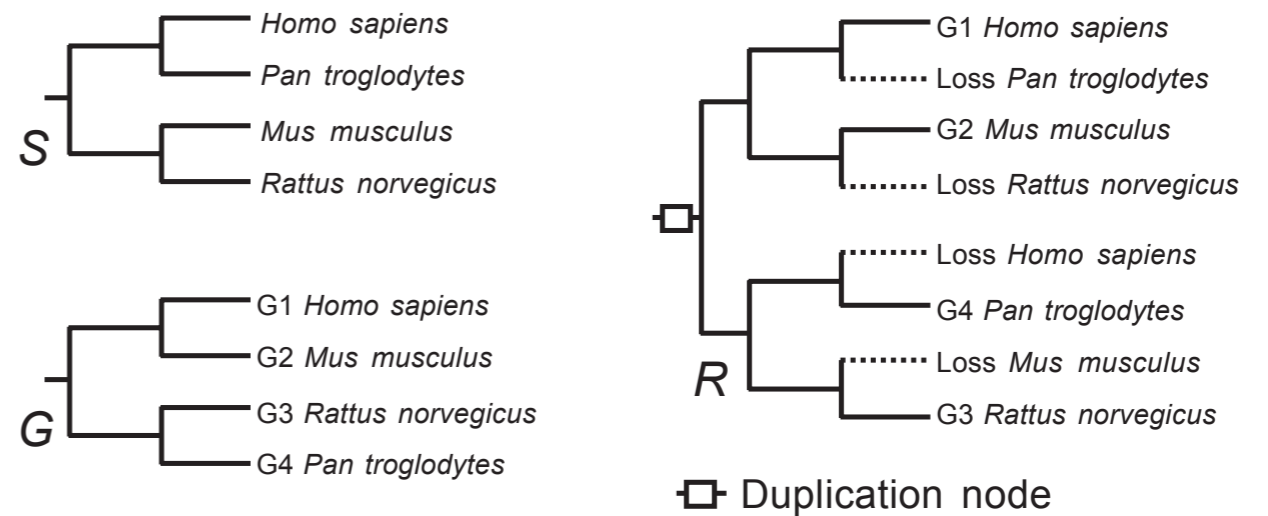
http://www.math.ku.dk/~richard/courses/binf_project/Stinus-BLAST.pdf

Orthology

Bidirectional Best-Hit



Gene/Species Tree Reconciliation



Dufayard et al., Bioinformatics, 2005

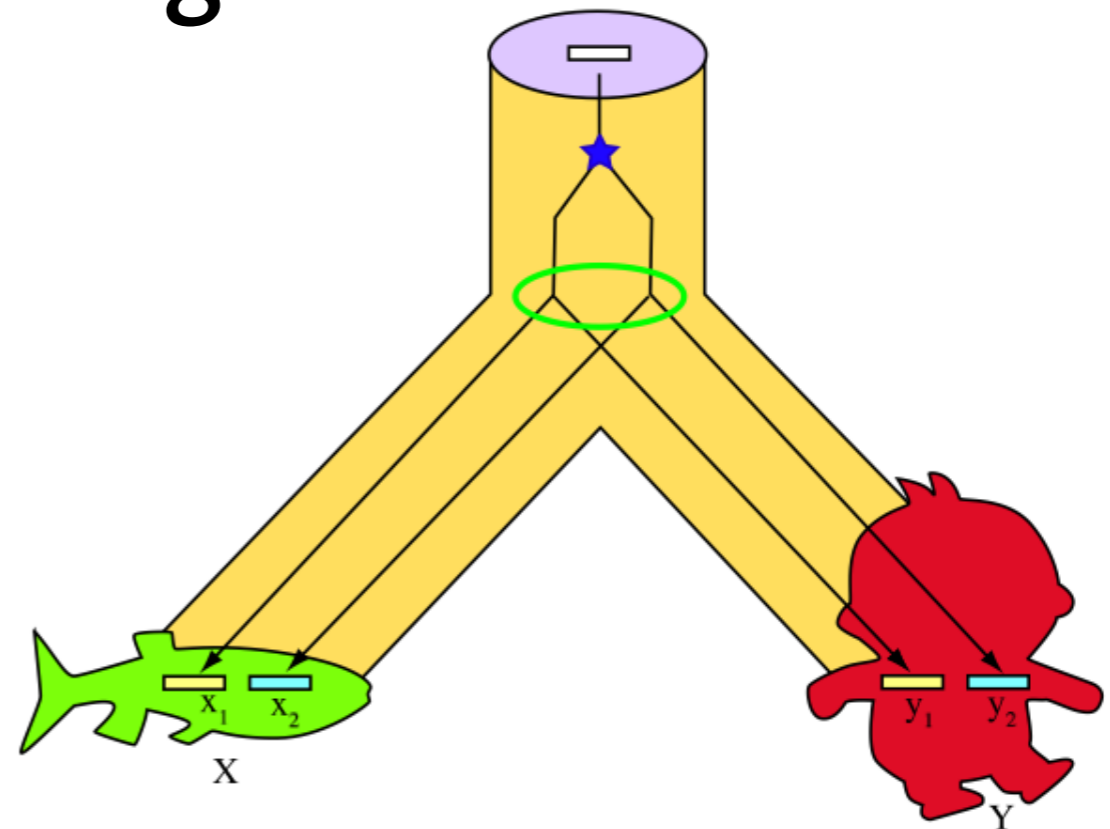
Pairwise Methods

The basic idea

Between two species, orthologs are closer than paralogs.

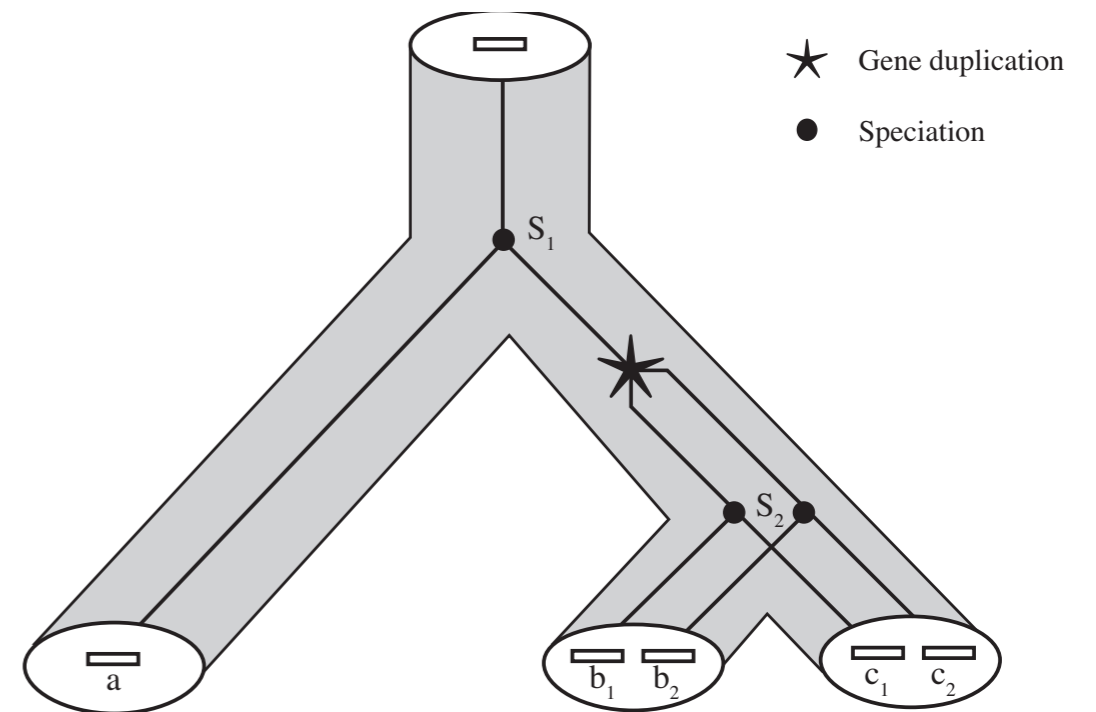
- Closer genes usually have higher alignment scores
→ species-specific top scoring hit is likely to be an ortholog
- Corresponding ortholog might be missing
→ require symmetry

“Bidirectional best hit” (BBH)



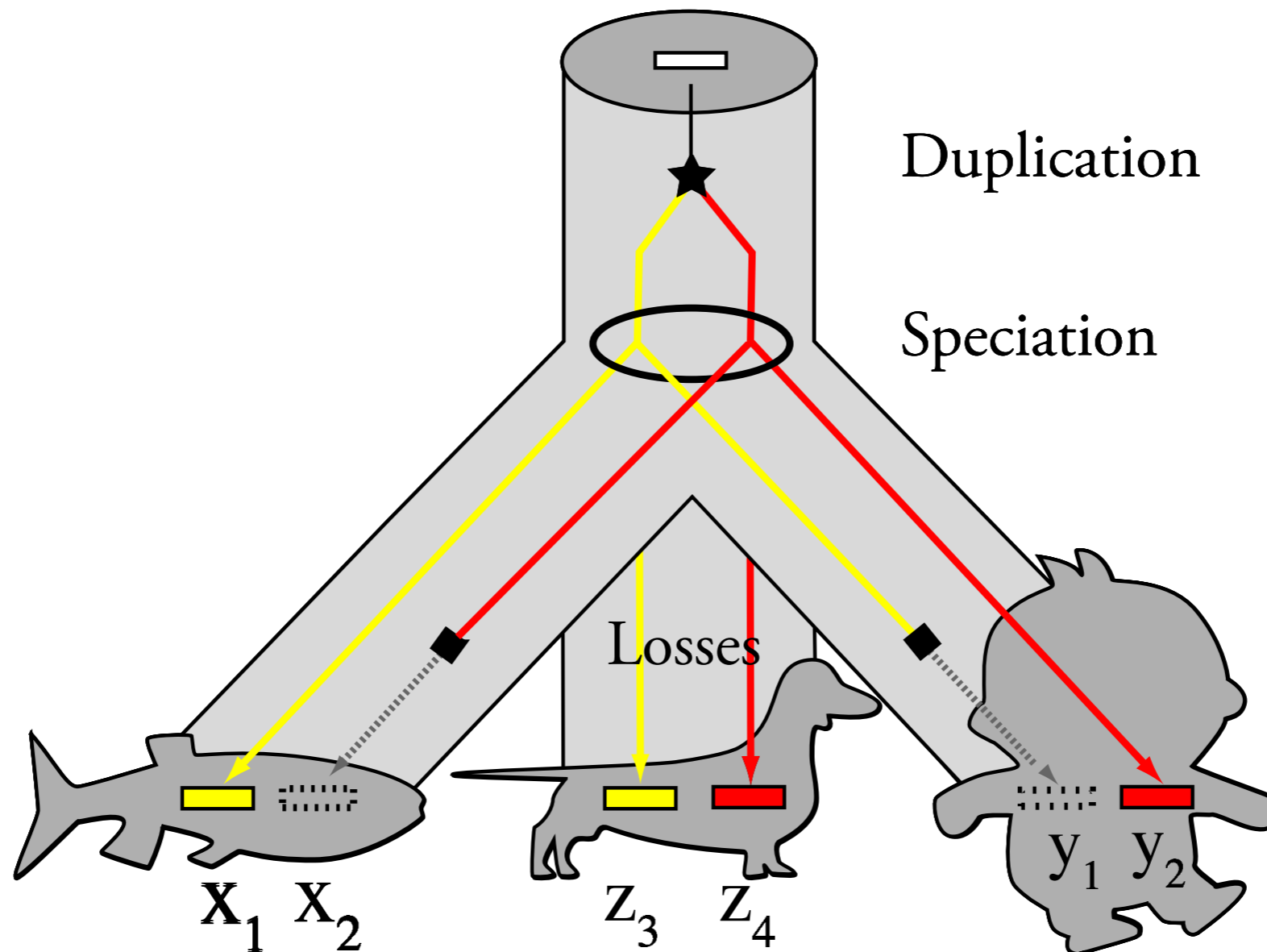
Refinements

- Instead of score, use evolutionary distance
“Reciprocal smallest distance” (RSD)
- Relax the top/smallest requirement to include more than one orthologs (e.g. 1:many orthology)
- Take into account statistical uncertainty of distance estimates
- Detect differential gene losses

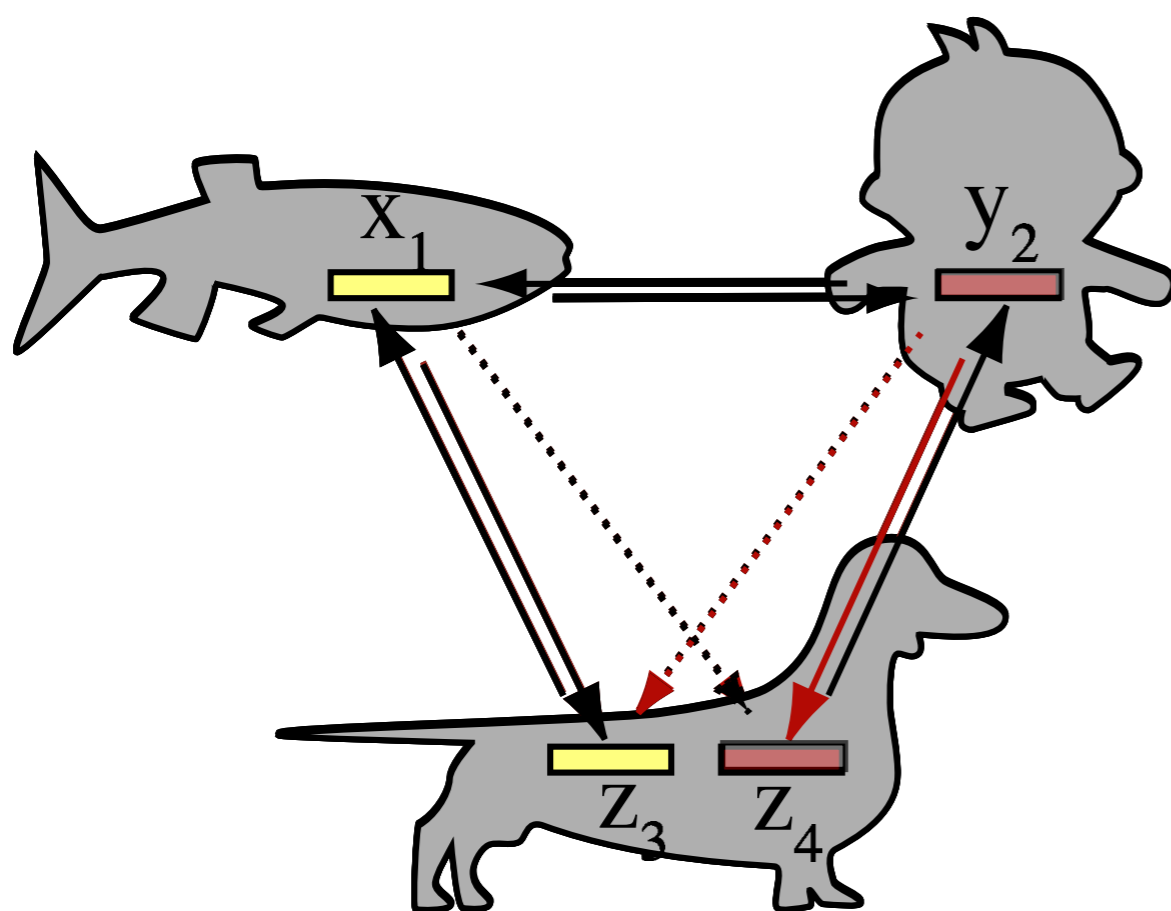


Wall et al., Bioinformatics, 2003
Dessimoz et al., RECOMB CG Dublin, 2005
Fulton et al., BMC Bioinformatics, 2006
Dessimoz et al., Nucleic Acids Res, 2006
Roth et al., BMC Bioinformatics, 2008

Verification of stable pairs



Verification of stable pairs

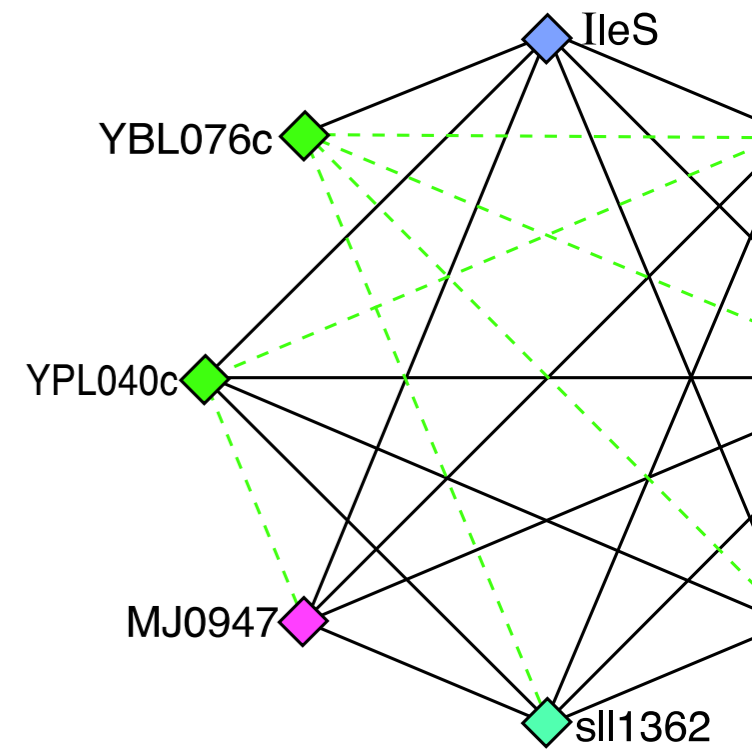


- (x_1, z_3) and (y_2, z_4) are stable pairs
- (x_1, z_3) signif. closer than (x_1, z_4)
- (y_1, z_4) signif. closer than (y_2, z_3)
- (x_1, z_4) and (y_2, z_3) not signif. different

Dessimoz, Boeckmann, et al., Nucl Acid Res, 2006

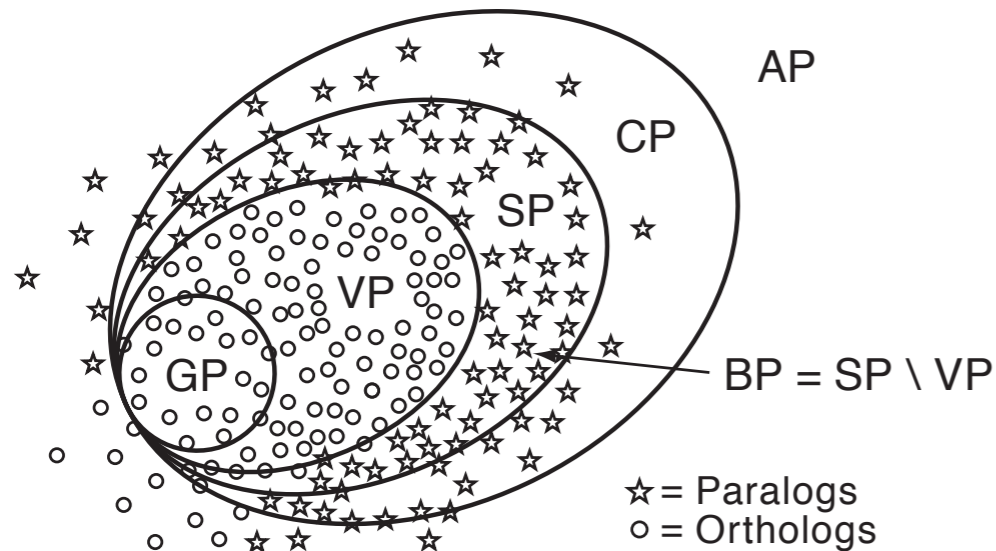
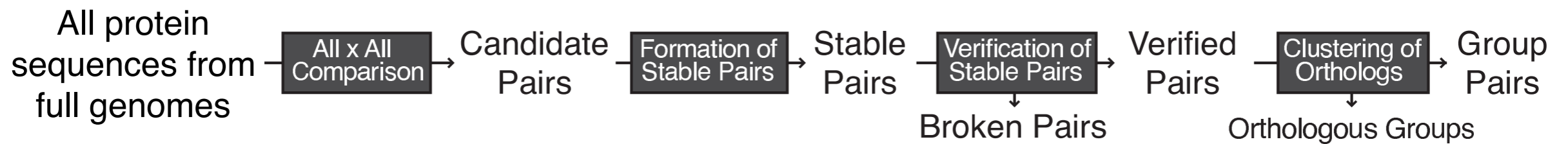
How to group orthologs?

- **If interested in particular gene x:**
 - group all genes orthologous to x
- **COGs database:**
 - group “triangles” of orthologs
 - merge triangles with common face
- **InParanoid (on pairs of genomes):**
 - start with a pair of orthologs
 - add in-paralogs (w.r.t. the only speciation)
- **OMA**
 - all pairs in a given group are orthologs



Tatusov et al. Science 1997

OMA



Pairs	Evolutionary Relation
All Pairs (AP)	Any
Candidate Pairs (CP)	Homologs
Stable Pairs (SP)	Orthologs, Pseudo-Orthologs
Broken Pairs (BP)	Paralogs
Verified Pairs (VP)	Orthologs
Group Pairs (GP)	Close Orthologs

Roth et al., BMC Bioinformatics, 2008

http://omabrowser.org



OMA Browser

http://omabrowser.org/cgi-bin/gateway.pl

OMA browser

Dataset: [more info]

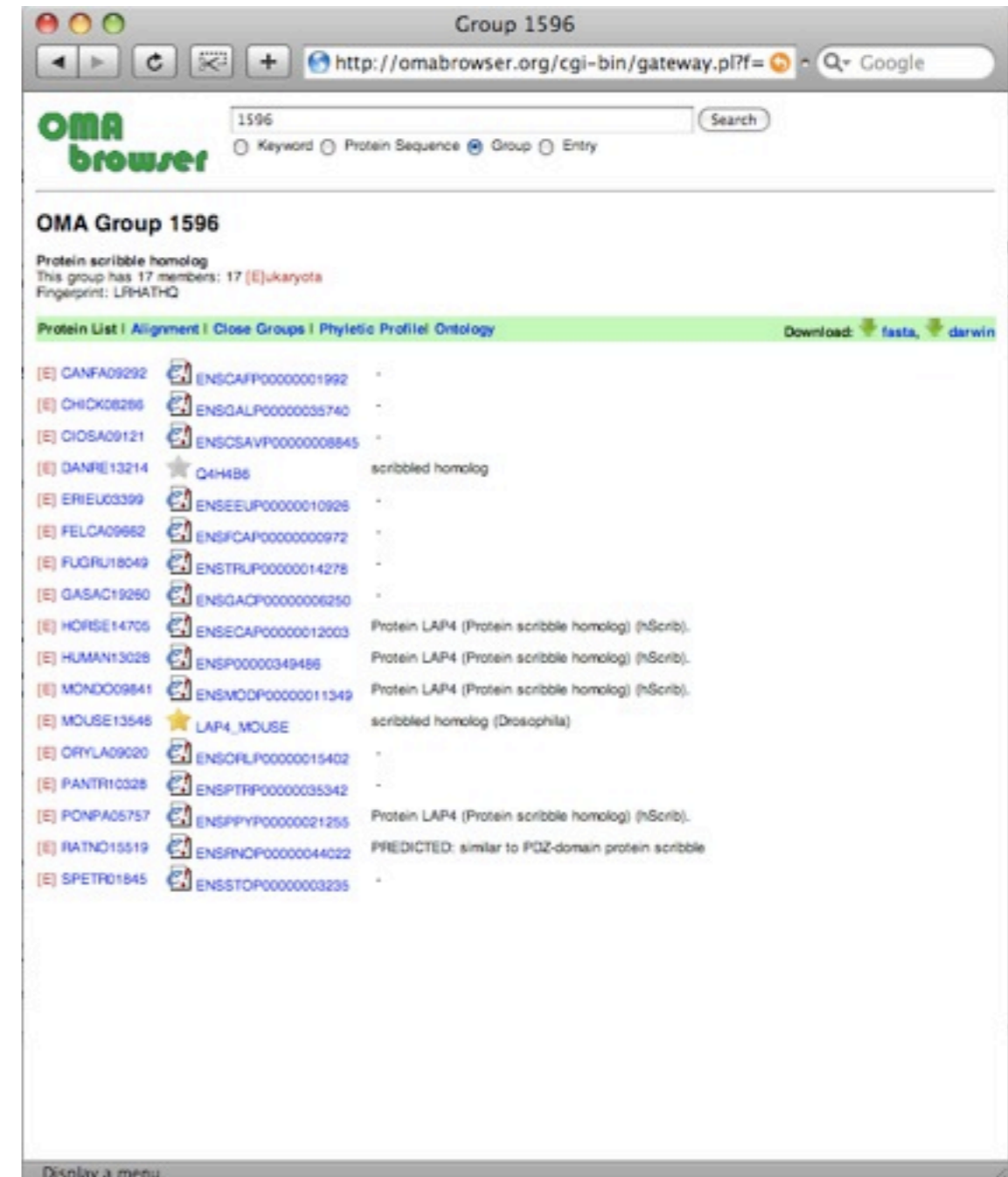
Keyword Protein Sequence Group Entry

Examples: [Search for "human" and "insulin"](#) - [OMA Group 1596](#) - [Entry ECOLI1190](#)

Data export: [Download](#) | [API](#)

About: [OMA Browser](#) | [The OMA Algorithm](#) | [This dataset](#)

©2007, CBRG, ETH Zurich.



Group 1596

OMA browser

1596

Keyword Protein Sequence Group Entry

OMA Group 1596

Protein scribble homolog
This group has 17 members: 17 [E]ukaryota
Fingerprint: LPHATHQ

[Protein List](#) | [Alignment](#) | [Close Groups](#) | [Phyetic Profile](#) | [Ontology](#) Download: [fasta](#), [darwin](#)

[E] CANFA09292	ENSCAIP0000001992	-
[E] CHICK08286	ENSQALP00000035740	-
[E] GIOSA09121	ENSOSAVP00000008845	-
[E] DANRE13214	Q4H4B6	scribbled homolog
[E] ERIEU03399	ENSEEUP00000010926	-
[E] FELCA09662	ENSIFCAP0000000972	-
[E] FUGRU18049	ENSTRUP00000014278	-
[E] GASAC19290	ENSGACP00000006250	-
[E] HORSE14705	ENSECAP00000012003	Protein LAP4 (Protein scribble homolog) (hScrib).
[E] HUMAN13028	ENSP00000349486	Protein LAP4 (Protein scribble homolog) (hScrib).
[E] MONDO09841	ENSMODP00000011349	Protein LAP4 (Protein scribble homolog) (hScrib).
[E] MOUSE13546	LAP4_MOUSE	scribbled homolog (Drosophila)
[E] ORYLA09020	ENSOFRLP00000015402	-
[E] PANTR10328	ENSPTRP00000035342	-
[E] PONPA05757	ENSPPYP00000021255	Protein LAP4 (Protein scribble homolog) (hScrib).
[E] RATND15519	ENSFNCP00000044022	PREDICTED: similar to PGZ-domain protein scribble
[E] SPETRD1845	ENSSTOP00000003235	-

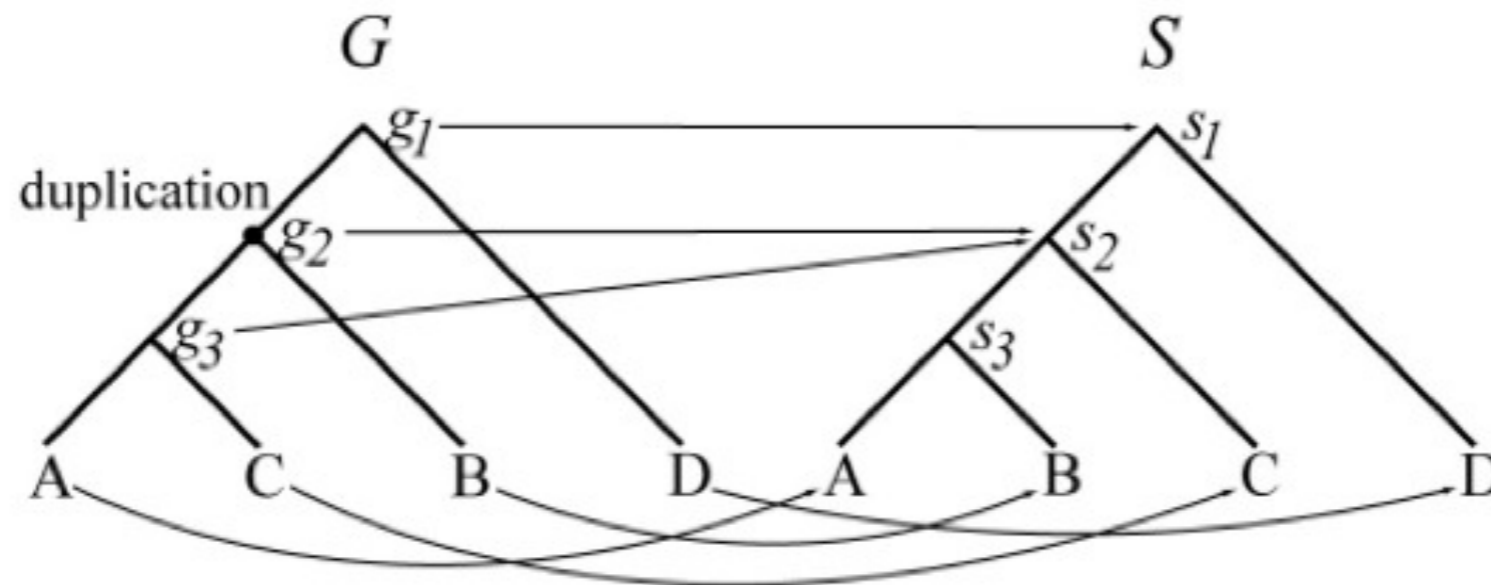
Display a menu

Tree Reconciliation

Maximum Parsimony
Reconciliation

Map between G and S

*Introduced implicitly: Goodman et al 1979
explicitly: Page 1994*



Zmasek & Eddy 2001

For any $g \in G$, let $M(g) \in S$ be the smallest (lowest) node in S satisfying $\gamma(g) \subseteq \sigma(M(g))$. That is, $M(g)$ points to the ancestral species in S that (we infer) harbored ancestral gene g .

Map \leftrightarrow Reconciliation

- Mirkin, Muchnik, Smith (1996) conjecture that the map cost function coincide with number of gene duplication and losses.
- Zhang (1997) and Eulenstein (1997) independently prove it, and identify efficient algorithms to compute map: in $O(n)$ and $O(n * \alpha(n))$ respectively.

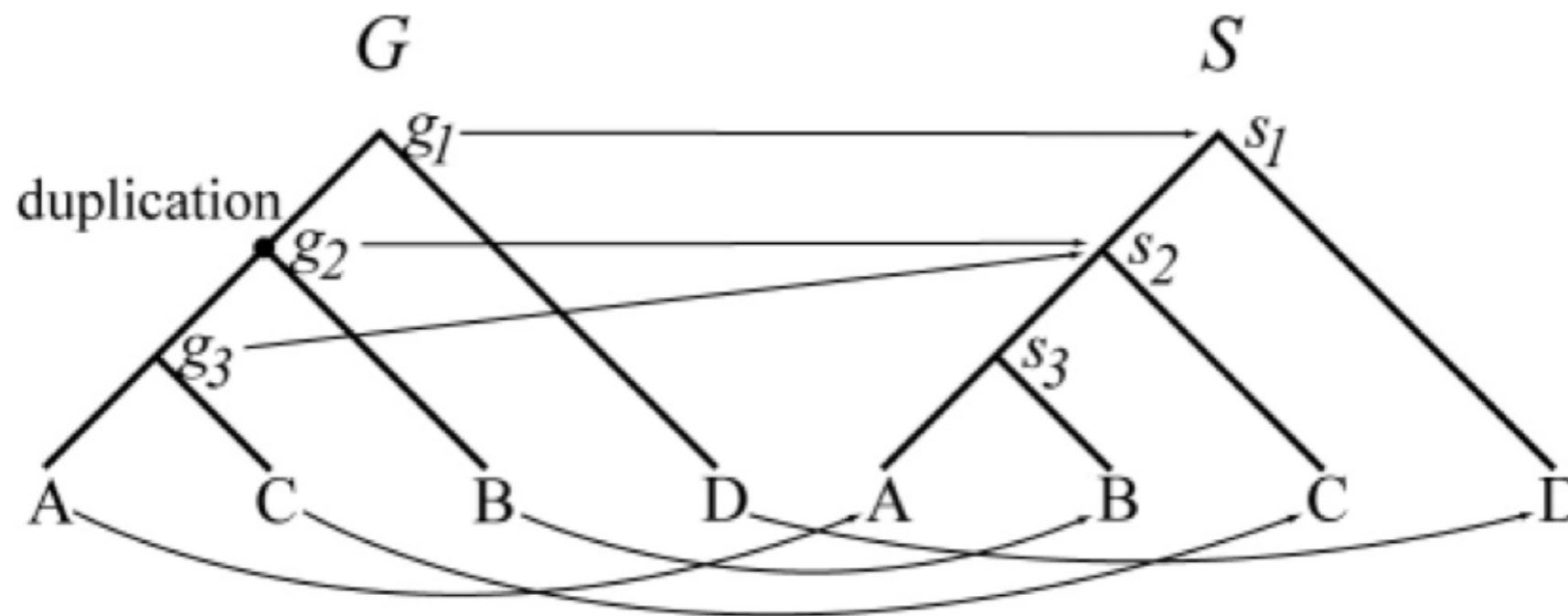


A simple algorithm to infer gene duplication and speciation events on a gene tree

Christian M. Zmasek and Sean R. Eddy

Howard Hughes Medical Institute, Department of Genetics, Washington University
School of Medicine, St Louis, MO 63110, USA

Received on January 24, 2001; revised on April 5, 2001; accepted on April 6, 2001

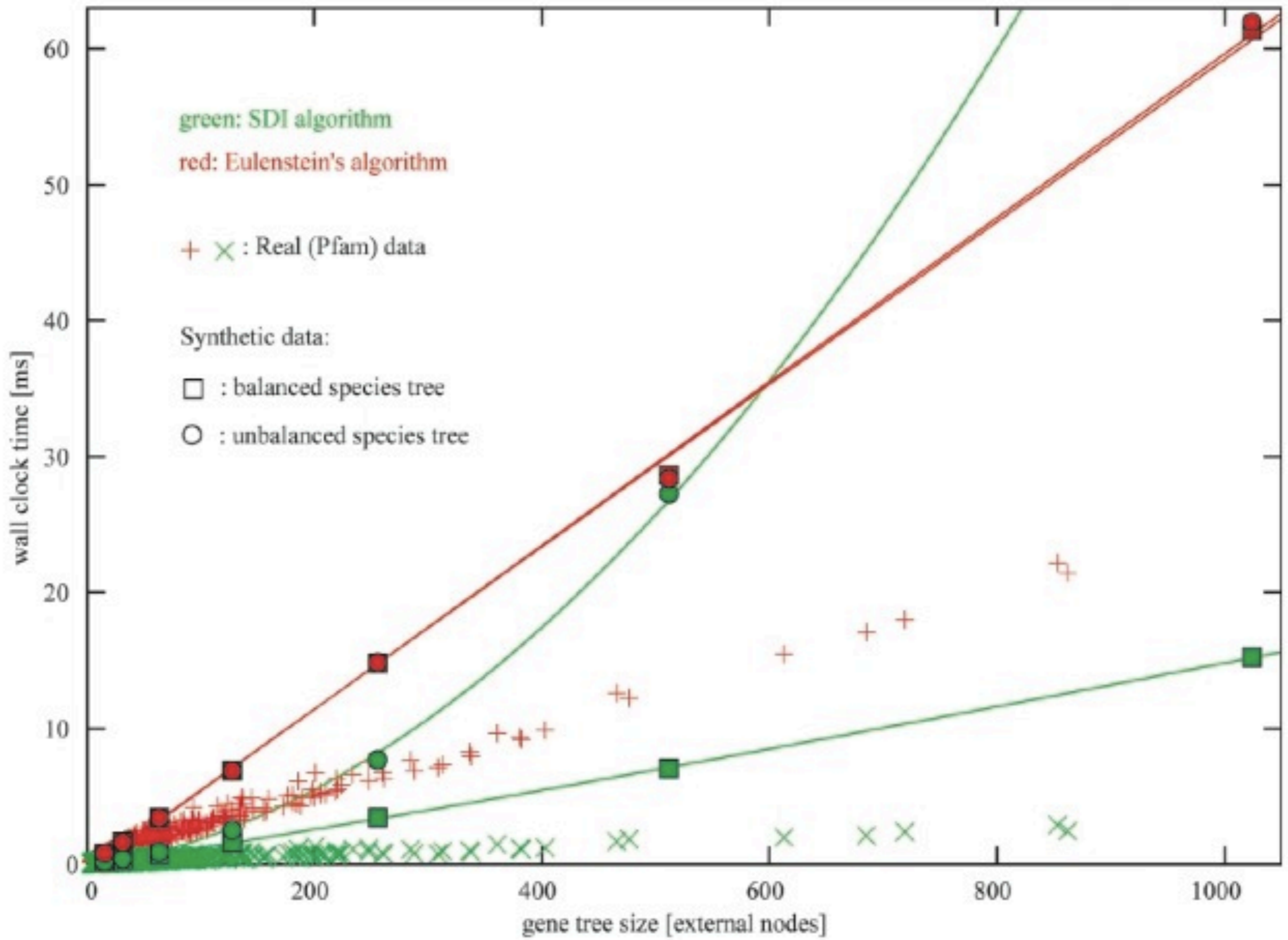


Mapping

- Map leaves in G to their species in S
- Map inner node g_i in postfix order (from leaves to root):
 - Map G_i to the lowest node s_i such that the species below g_i are all included below s_i

Duplication node assignment

- If g_i maps in S to the same node as one of its children, g_i is a duplication node



Tree Rooting?

- **Center of gravity**
 - Storm & Sonnhammer 2002
- **Min # of duplications**
 - Hallett & Lagergren 2000
 - Zmasek & Eddy 2002
(min height to break ties)
- **Outgroup (tricky)**
 - Huerta-Cepas et al. 2007

Tree Inference Errors?

- **Bootstrapping**
 - Storm & Sonnhammer 2002
 - Zmasek & Eddy 2002
- **Multifurcation (unresolved branches)**
 - Dufayard et al. 2005
 - Berglund-Sonnhammer et al. 2006
 - Durand et al. 2006

High Duplication/Loss Rates?

- Maximum parsimony criterion may be inappropriate

BIOINFORMATICS

Vol. 19 Suppl. 1 2003, pages i7–i15
DOI: 10.1093/bioinformatics/btg1000



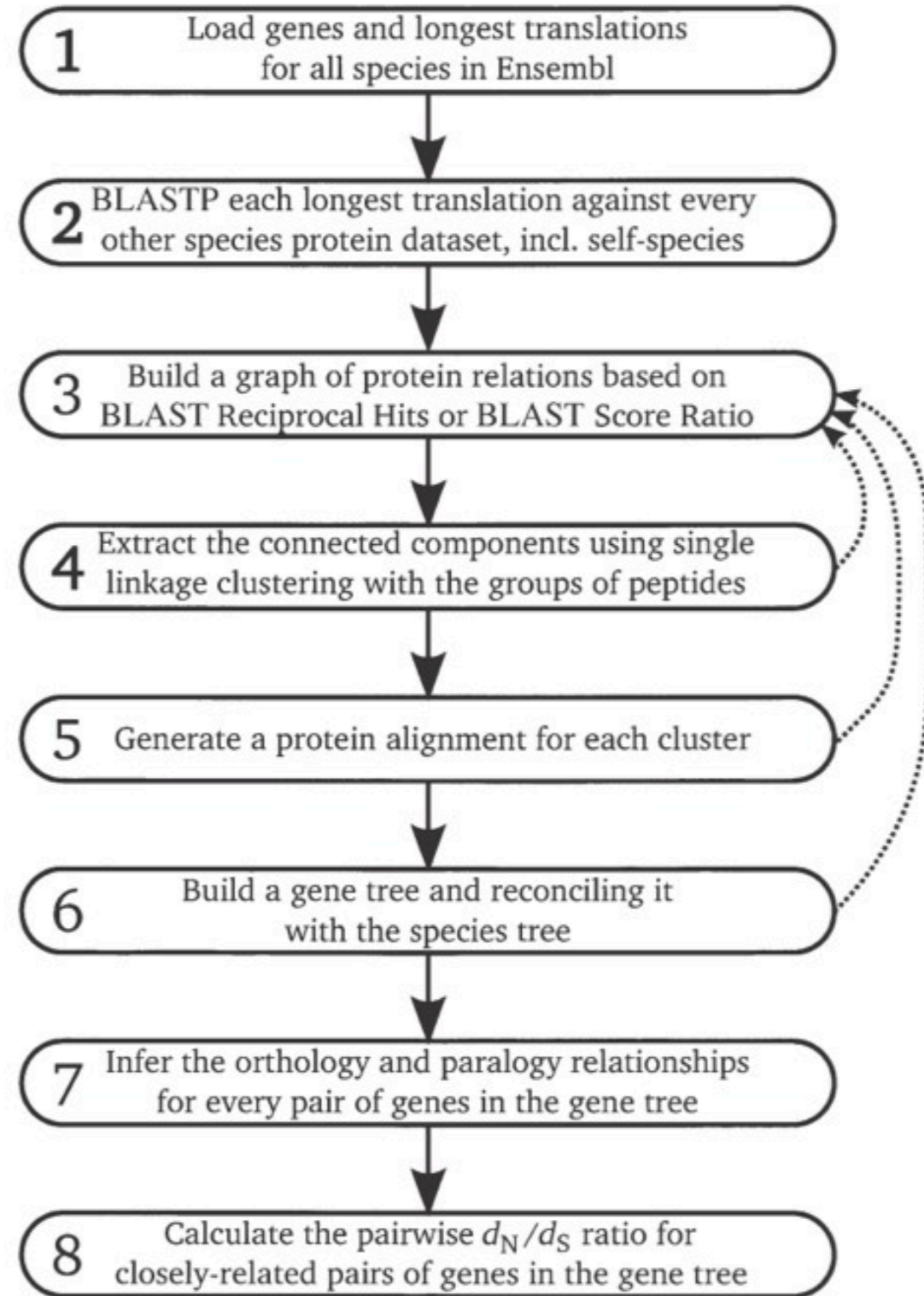
Bayesian gene/species tree reconciliation and orthology analysis using MCMC

Lars Arvestad^{2,*}, Ann-Charlotte Berglund¹, Jens Lagergren¹ and Bengt Sennblad²

¹SBC and Department of Numerical Analysis and Computing Science, KTH, SE-100 44, Stockholm and ²SBC and Center for Genomics and Bioinformatics, Karolinska Institutet, SE-171 77, Stockholm, Sweden

Case study: Ensembl Compara

Vilella et al. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res (2009) vol. 19 (2) pp. 327-35



Evaluating Orthology Predictions

Altenhoff and Dessimoz, PLoS Comput Biol 2009

Previous work

Research

Benchmarking ortholog identification methods using functional genomics data

Tim Hulsen*, Martijn A Huynen*, Jacob de Vlieg**† and Peter MA Groenen†

Genome Biology 2006, **7**:R31

Assessing Performance of Orthology Detection Strategies Applied to Eukaryotic Genomes

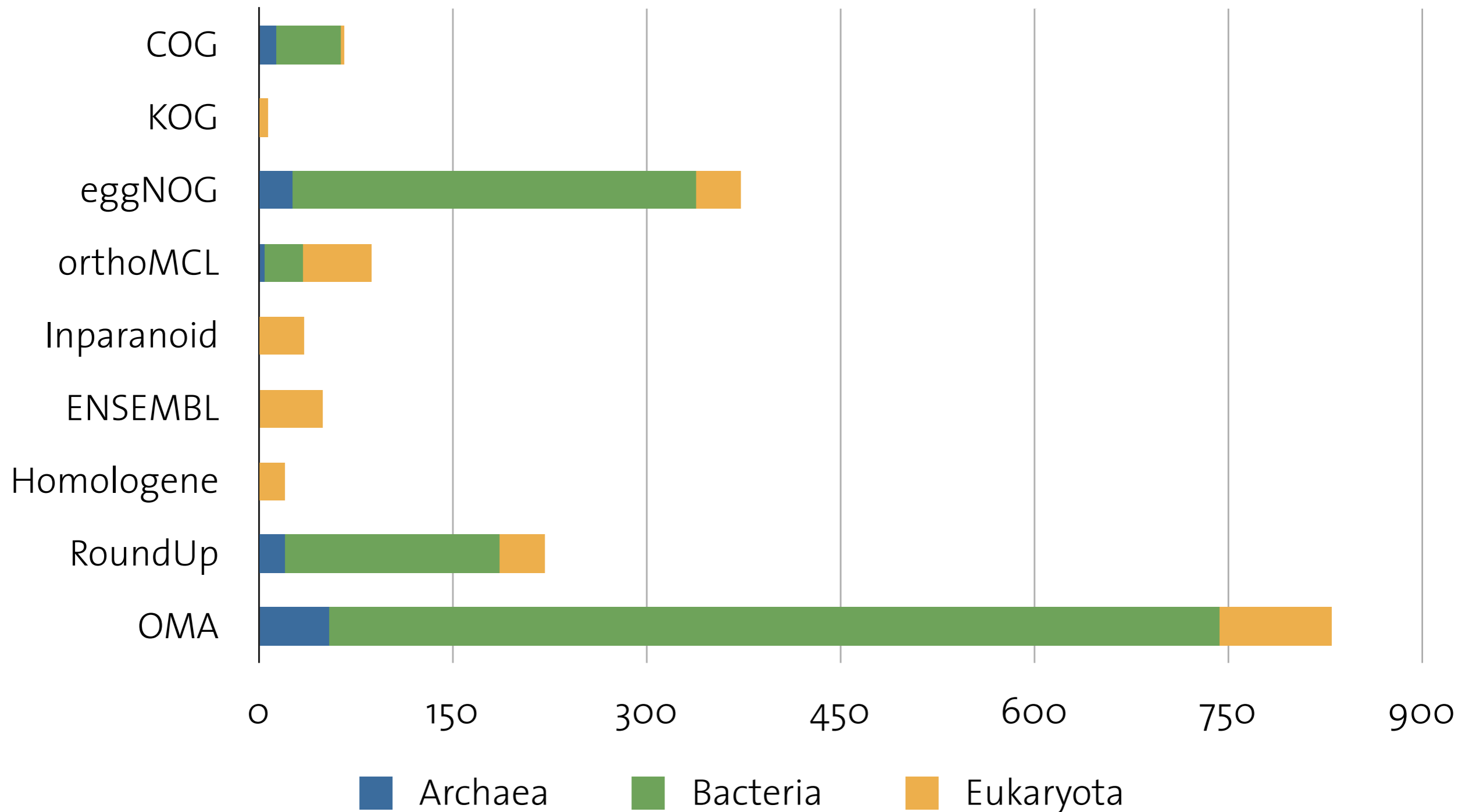
Feng Chen^{1,3}, Aaron J. Mackey^{2,3#}, Jeroen K. Vermunt⁴, David S. Roos^{2,3*}

April 2007 | Issue 4 | e383



Projects under Scrutiny

May 2009



Comparison Approach

- **For each project, map sequences to OMA**
7.16 million sequences in total
329.2 million orthology relations
- **Intersection over all projects: \emptyset !**
 - ➔ “pairwise” tests with OMA
 - ➔ “intersection” tests with subset

Assessment of Orthologs

Phylogeny

Conserved function

Species Tree Discordance

Gene Ontology

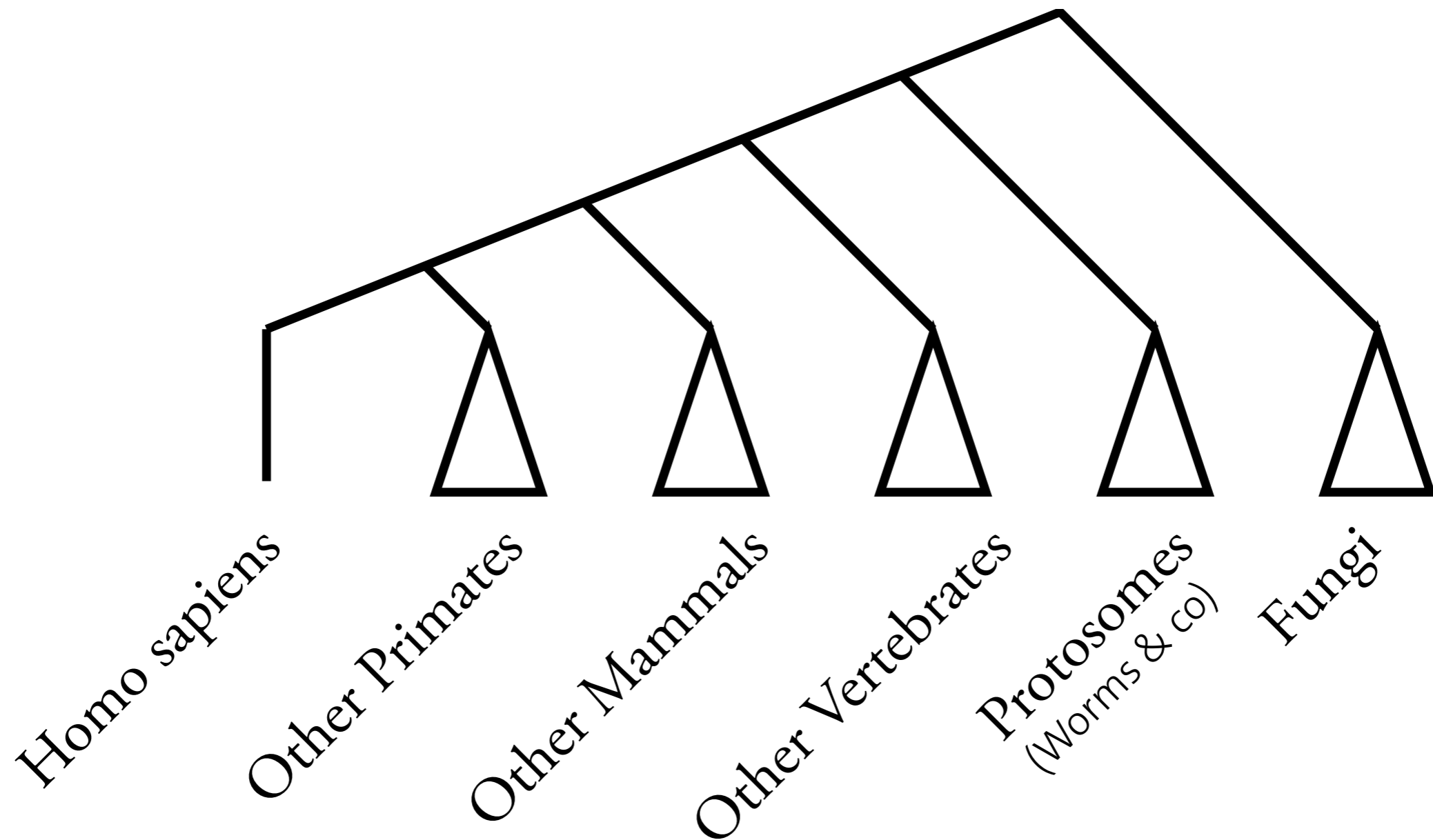
**Phylogenetic Analyses
from Literature**

Enzyme Classification

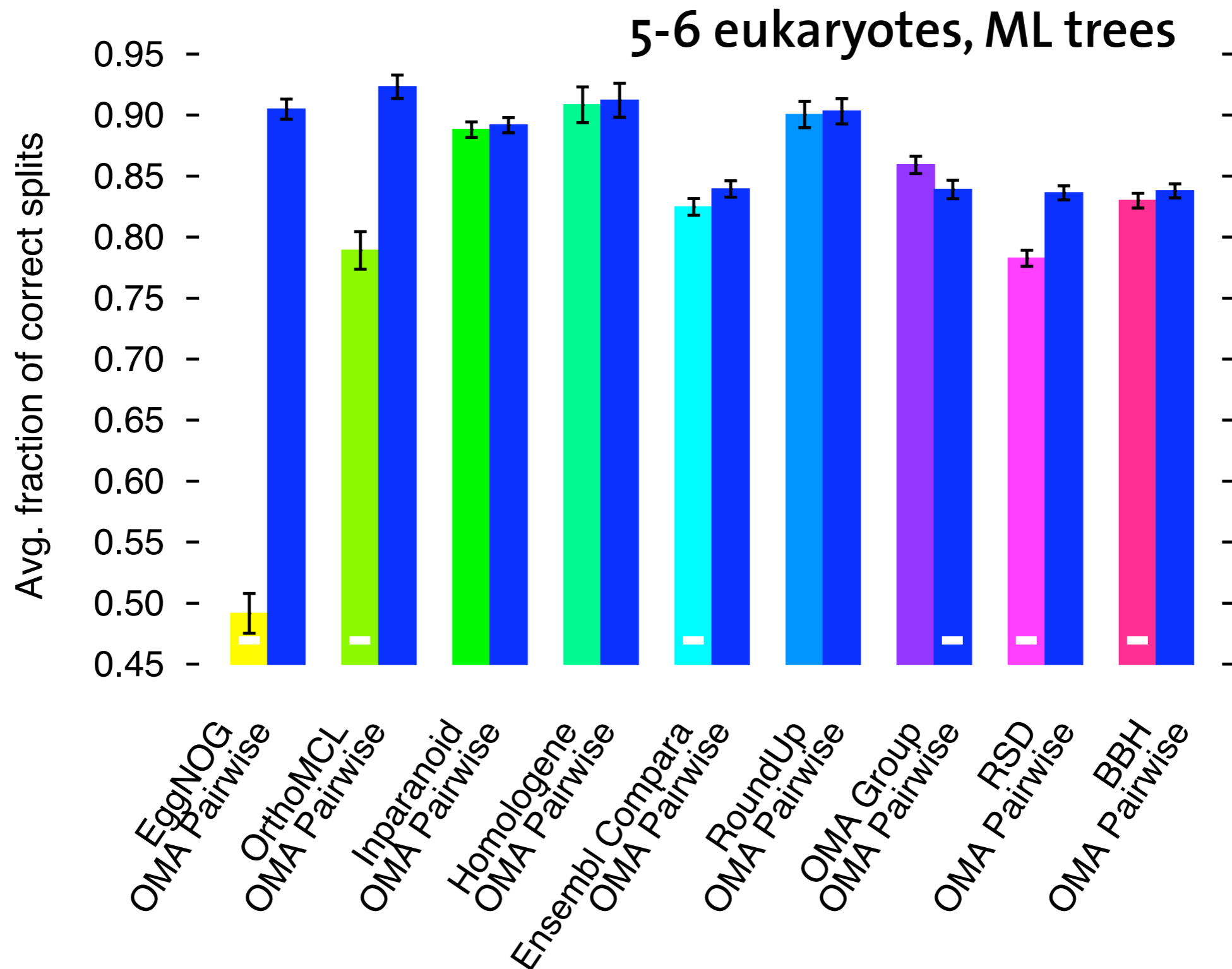
Gene Expression

Genomic Context

Species Tree Discordance

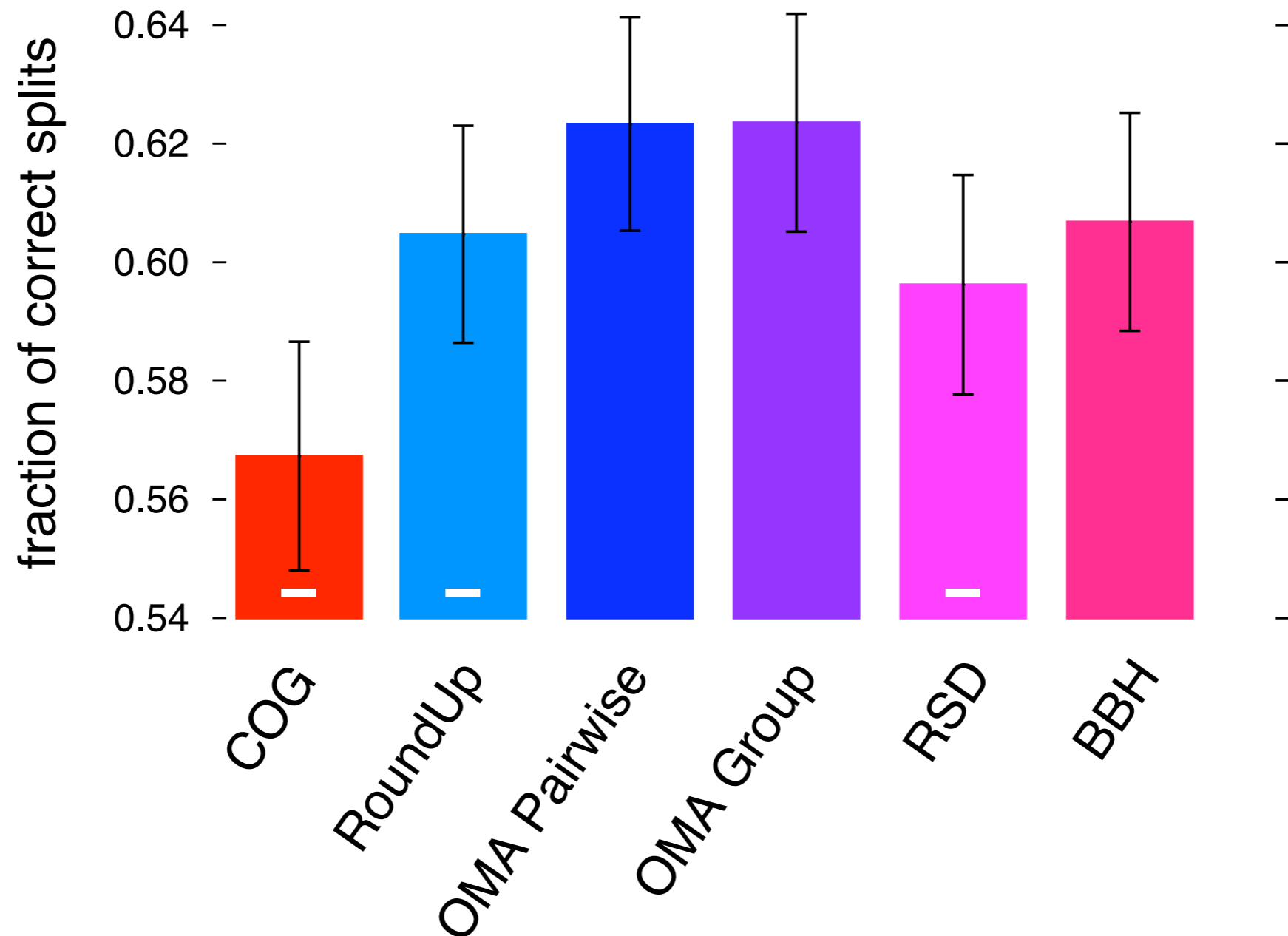


Species Tree Discordance

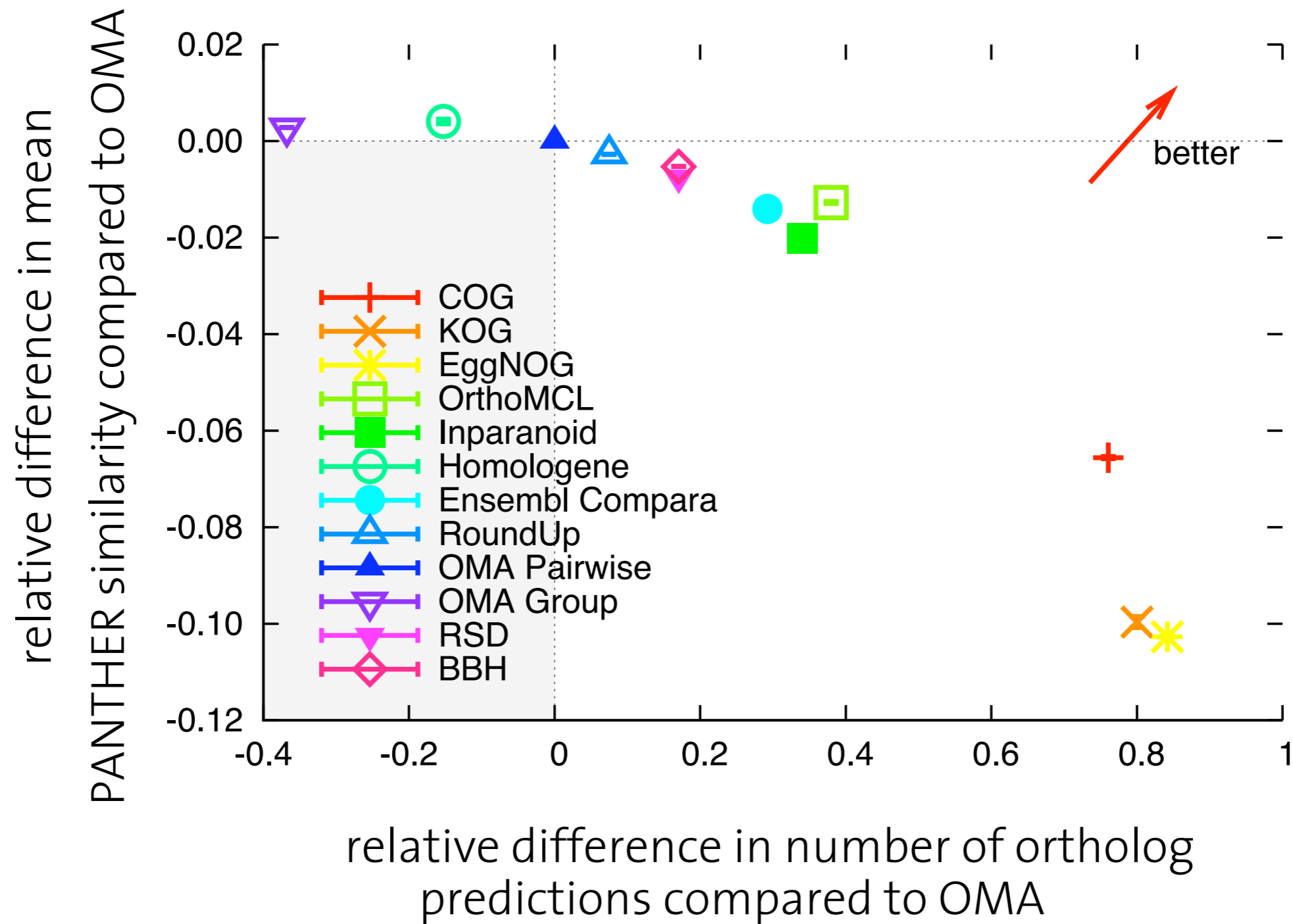


Species Tree Discordance

6-7 bacteria, distance trees



Panther Ontology Conservation



Limitations and Future Directions

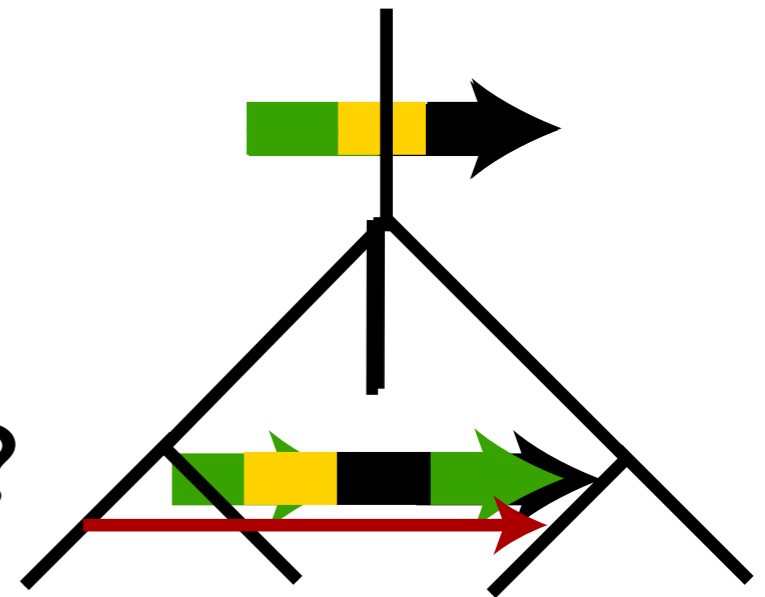
Limits of Pairwise Relations

- Useful if focused on a particular gene
- But several shortcomings:
 - Evolutionary distance?
 - Function conservation?
 - Grouping strategy?



Limits of Model

- Lateral gene transfer?
- Gene fusion/fission?
- Domain shuffling?
- Heterogeneous population?
- Hybridization?



Dessimoz et al., RECOMB 2008

Limits of Orthology for function inference

Opinion

Cell
PRESS

How confident can we be that orthologs are similar, but paralogs differ?

Romain A. Studer and Marc Robinson-Rechavi

Department of Ecology and Evolution, Biophore, Lausanne University, CH-1015 Lausanne, Switzerland and Swiss Institute of Bioinformatics, CH-1015 Lausanne, Switzerland

Homologous genes are classified into orthologs and paralogs, depending on whether they arose by speciation or duplication. It is widely assumed that orthologs share similar functions, whereas paralogs are expected to diverge more from each other. But does this assumption hold up on further examination? We present evidence that orthologs and paralogs are not so different in either their evolutionary rates or their mechanisms of divergence. We emphasize the importance of appropriately designed studies to test models of gene evolution between orthologs and between paralogs. Thus, functional change between orthologs might be as common as between paralogs, and future studies should be designed to test the impact of duplication against this alternative model.

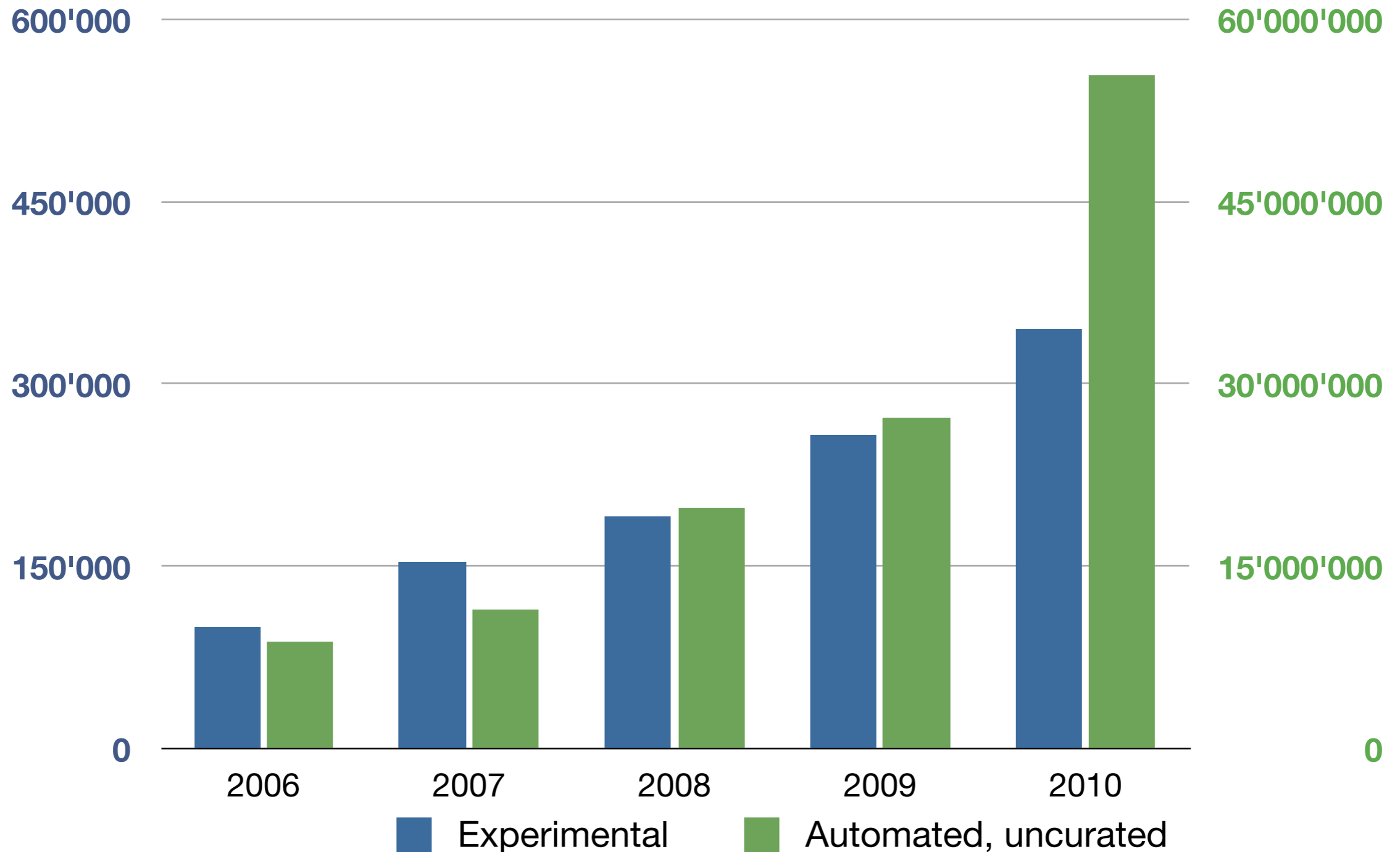
tion. But the assumption that changes in function are commonly associated with duplication has rarely been explicitly tested. Although there have been many studies of comparative genomics focused on the role of duplication (for a review, see Ref. [1]), few have compared the evolution of paralogs with the evolution of orthologs. However, these studies repeatedly find little, if any, specific impact of duplication. This pattern is surprising if the standard model is correct.

This 'standard model' makes two predictions. First, paralogs are expected to diverge more per unit of time than orthologs. Second, paralogs are expected to diverge frequently in ways that are rarely observed between orthologs; for example, different substrate specificities. Divergence can concern different aspects of gene function [3],

Trends Genet (2009) vol. 25 (5) pp. 210-216

Limits of Computational Inference

Growth of GO Annotations



Questions?