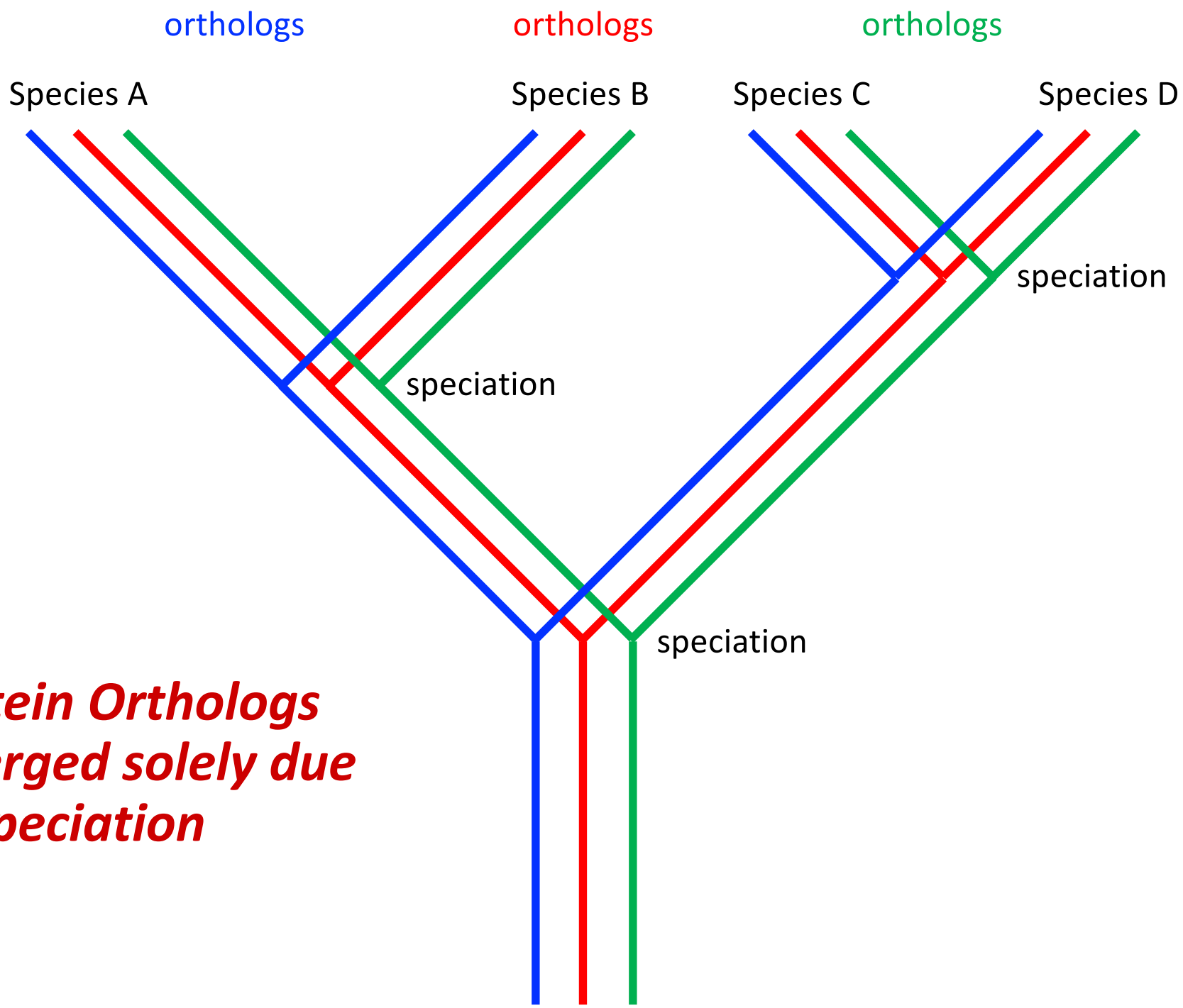
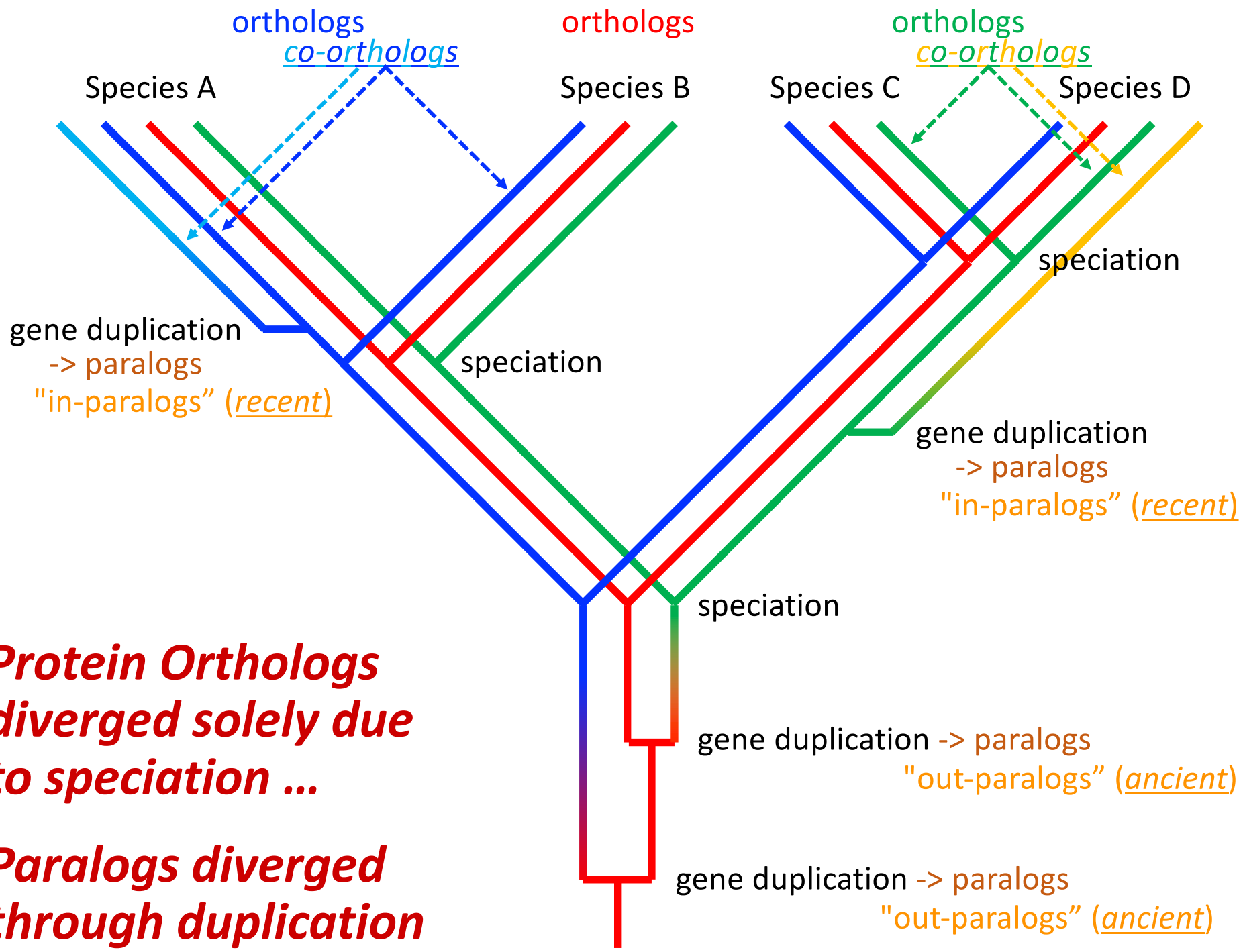
Protein Orthology for Comparative Genomics:

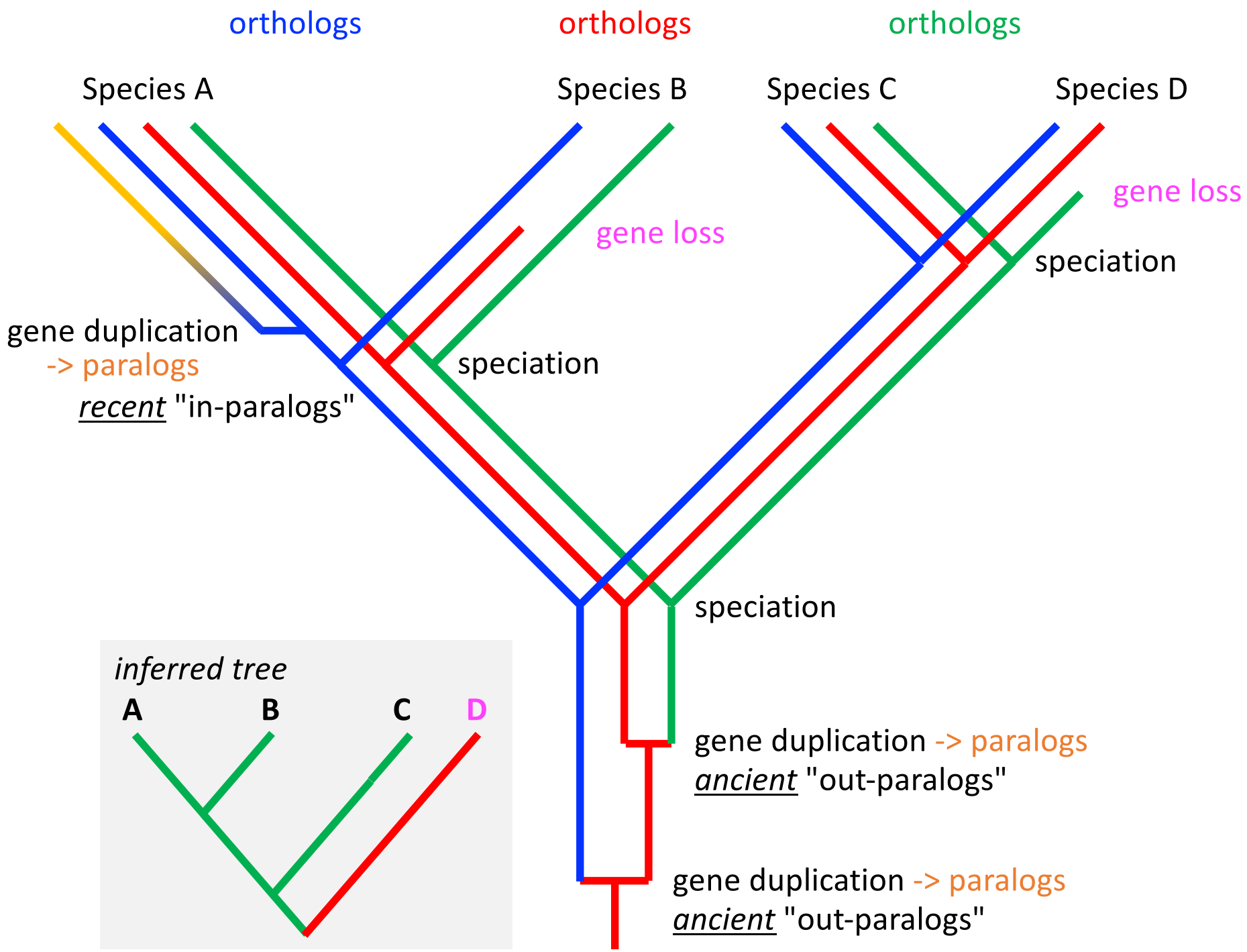
- ***Genome Comparisons** (synteny)*
- ***Functional Comparisons** (particularly for proteins)*
  - ***how make reliable inferences … based on functional data in other species?***
- ***Protein Homology: Orthology & Paralogy***
- ***Methods based on:***
  - ***functional characterization** (typically unavailable at genomic scale)*
  - ***structure** (ditto … and not necessarily = function)*
  - ***expert curation** (labor-intensive, especially for eukaryotic taxa)*
  - ***synteny** (common for closely-related taxa, but decays rapidly)*
  - ***protein sequence similarity** – e.g. OrthoMCL, OrthoFinder (FungiDB, Mycocosm)*
  - ***domain structure** – e.g. EggNOG, Compara (Ensembl)*
  - ***phylogeny** – e.g. OMA, OrthoFinder2*

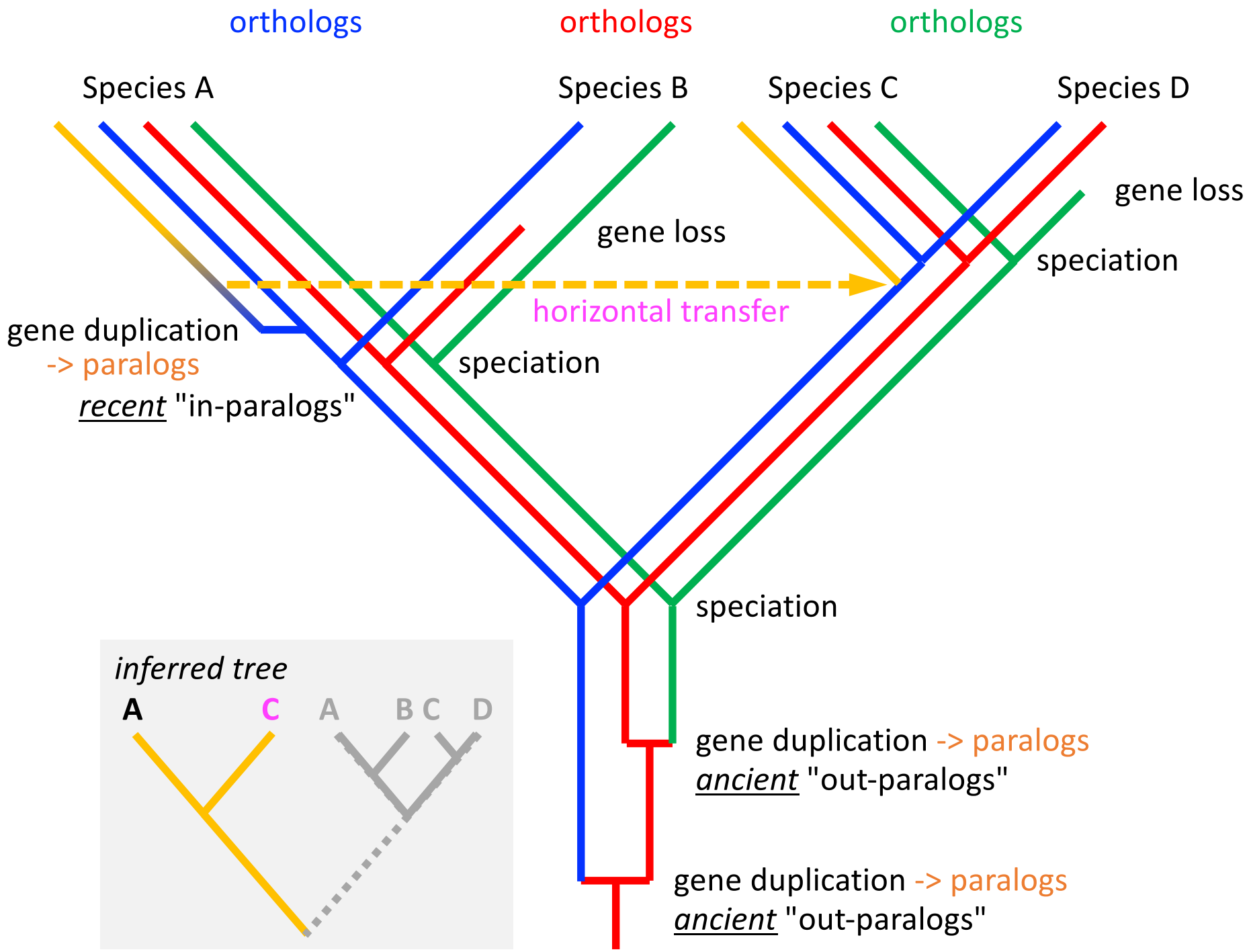
  Note: Benchmarking is difficult without a gold standard (Q4O)!
- ***Orthology applications:***
  - ***Validation of genome assemblies & completeness (e.g. BUSCO)***
  - ***Gene gains, losses, amplification, horizontal transfer (phyletic pattern profiling)***
  - ***Functional inference for understudied taxa (GO terms, etc)***



Protein Orthologs diverged solely due to speciation


orthologs
co-orthologs
orthologs
orthologs
co-orthologs
Species A
Species B
Species C
Species D
speciation
gene duplication
-> paralogs
"in-paralogs" (recent)
speciation
gene duplication
-> paralogs
"in-paralogs" (recent)
speciation
gene duplication -> paralogs
"out-paralogs" (ancient)
gene duplication -> paralogs
"out-paralogs" (ancient)
Protein Orthologs diverged solely due to speciation …
Paralogs diverged through duplication


orthologs
orthologs
orthologs
Species A
Species B
Species C
Species D
gene loss
gene loss
speciation
gene duplication
-> paralogs
recent "in-paralogs"
speciation
speciation
inferred tree
A
B
C
D
gene duplication -> paralogs
ancient "out-paralogs"
gene duplication -> paralogs
ancient "out-paralogs"


orthologs
orthologs
orthologs
Species A
Species B
Species C
Species D
gene loss
gene loss
speciation
horizontal transfer
gene duplication
-> paralogs
recent "in-paralogs"
speciation
speciation
inferred tree
A
C
A
B C
D
gene duplication -> paralogs
ancient "out-paralogs"
gene duplication -> paralogs
ancient "out-paralogs"

Protein Orthology for Comparative Genomics:

- ***Genome Comparisons** (synteny)*
- ***Functional Comparisons** (particularly for proteins)*
  - ***how make reliable inferences … based on functional data in other species?***
- ***Protein Homology: Orthology & Paralogy***
- ***Methods based on:***
  - ***functional characterization** (typically unavailable at genomic scale)*
  - ***structure** (ditto … and not necessarily = function)*
  - ***expert curation** (labor-intensive, especially for eukaryotic taxa)*
  - ***synteny** (common for closely-related taxa, but decays rapidly)*
  - ***protein sequence similarity** – e.g. OrthoMCL, OrthoFinder (FungiDB, Mycocosm)*
  - ***domain structure** – e.g. EggNOG, Compara (Ensembl)*
  - ***phylogeny** – e.g. OMA, OrthoFinder2*

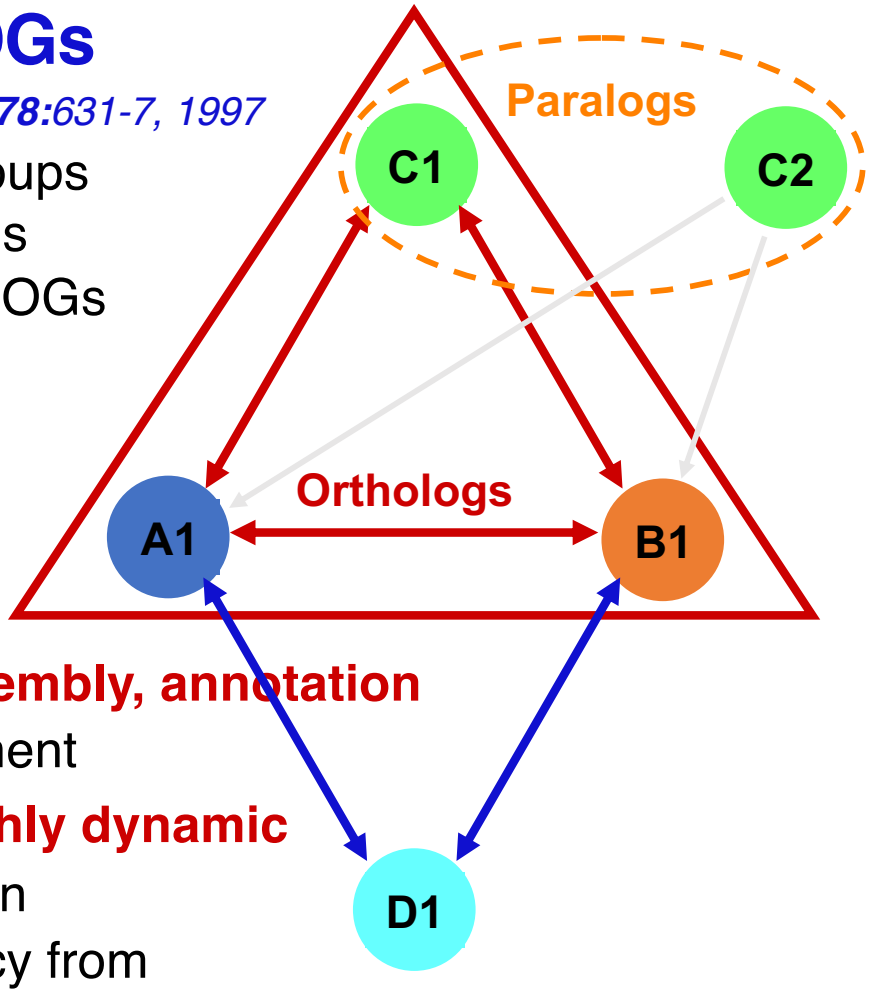
  Note: Benchmarking is difficult without a gold standard (Q4O)!
- ***Orthology applications:***
  - ***Validation of genome assemblies & completeness (e.g. BUSCO)***
  - ***Gene gains, losses, amplification, horizontal transfer (phyletic pattern profiling)***
  - ***Functional inference for understudied taxa (GO terms, etc)***

Genomic-scale identification of orthologous groups

Prokaryotic ortholog groups: COGs

*Tatusov et al. Science **278:**631-7, 1997*

Major resource for prokaryotic orthologous groups
Triangles of reciprocal best hits from 3 lineages
Merge triangles with a common side to form COGs
One-way best hits added as paralogs
Extensive manual curation



Challenges in identifying eukaryotic ortholog groups:

Incomplete (meta)genome sequences, assembly, annotation
- Difficult to ensure correct ortholog assignment

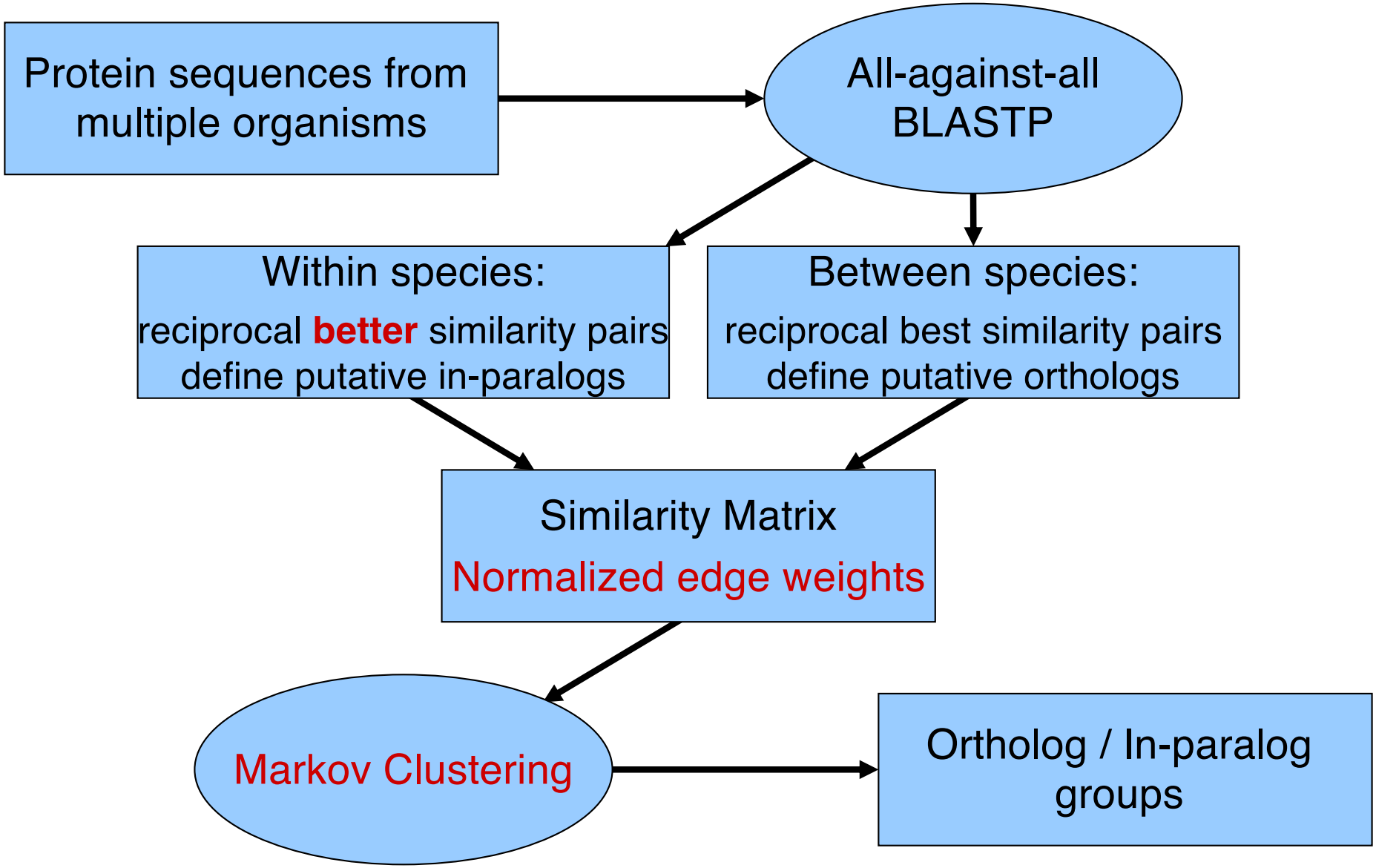
Large genomes, extensive duplication, highly dynamic
- Places a premium on automated annotation
- Difficult to distinguish functional redundancy from functional divergence (in- *vs* out-paralogs)

Complicated protein domain structure
- Difficult to distinguish proteins with shared domains but distinct functions

Manual curation generally not feasible

*Li et al, Genome Res **13:**2178 (2003)*
*Chaudhary & Roos, Nature BT **23:**1089 (2005)*
*Chen et al, NAR **34:**363 (2006)*
*Chen et al, PLoS1 **2:**e383 (2007)*
*Morrison et al, Science **317:**1921 (2007)*
*Peterson et al, Prot Sci **18:**1306 (2009)*
*Coyne et al, Genome Biol **12:**R100 (2011)*
*Fischer et al, Curr Prot Bioinf **6:**1 (2011)*

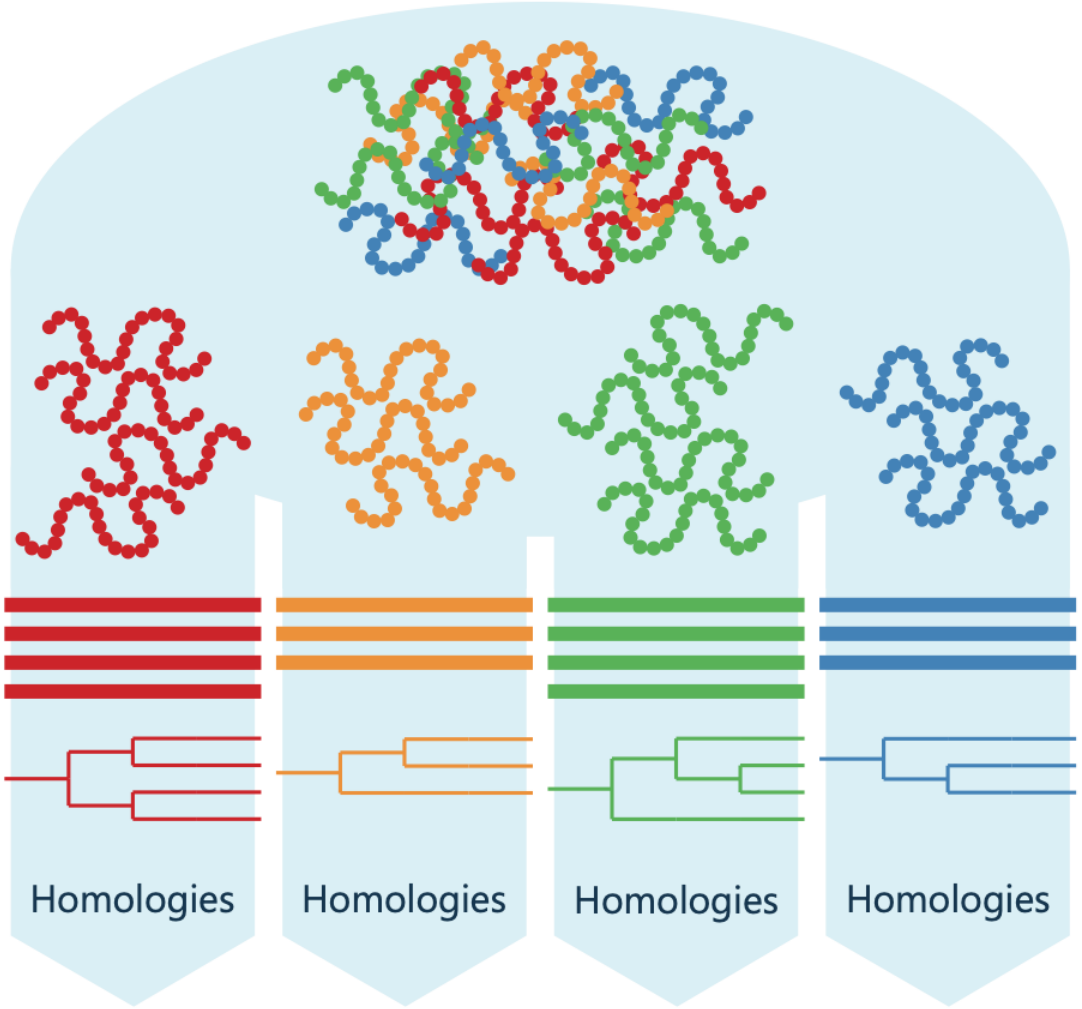
Identification of ortholog groups using OrthoMCL



Li et al, Genome Research ***13:****2178-90 (2003)*

Gene/protein trees

1. Representative translation of each gene from all species
2. All-vs-all HMM search to classify into families or clustering
3. Multiple protein alignment
4. Phylogenetic tree for each aligned cluster and reconciliation against NCBI taxonomy
5. Ortho-/paralogue inference



OPEN ACCESS Freely available online

PLoS one

Assessing Performance of Orthology Detection Strategies Applied to Eukaryotic Genomes

Feng Chen[1,3], Aaron J. Mackey[2,3]*, Jeroen K. Vermunt[4], David S. Roos[2,3]*

1 Department of Chemistry, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America, 2 Department of Biology, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America, 3 Genomics Institute, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America, 4 Department of Methodology and Statistics, Tilburg University, The Netherlands

Orthology detection is critically important for accurate functional annotation, and has been widely used to facilitate studies on comparative and evolutionary genomics. Although various methods are now available, there has been no comprehensive analysis of performance, due to the lack of a genomic-scale 'gold standard' orthology dataset. Even in the absence of such datasets, the comparison of results from alternative methodologies contains useful information, as agreement enhances confidence and disagreement indicates possible errors. Latent Class Analysis (LCA) is a statistical technique that can exploit this information to reasonably infer sensitivities and specificities, and is applied here to evaluate the performance of various orthology detection methods on a eukaryotic dataset. Overall, we observe a trade-off between sensitivity and specificity in orthology detection, with BLAST-based methods characterized by high sensitivity, and tree-based methods by high specificity. Two algorithms exhibit the best overall balance, with both sensitivity and specificity>80%: INPARANOID identifies orthologs across two species while OrthoMCL clusters orthologs from multiple species. Among methods that permit clustering of ortholog groups spanning multiple genomes, the (automated) OrthoMCL algorithm exhibits better within-group consistency with respect to protein function and domain architecture than the (manually curated) KOG database, and the homolog clustering algorithm TribeMCL as well. By way of using LCA, we are also able to comprehensively assess similarities and statistical dependence between various strategies, and evaluate the effects of parameter settings on performance. In summary, we present a comprehensive evaluation of orthology detection on a divergent set of eukaryotic genomes, thus providing insights and guides for method selection, tuning and development for different applications. Many biological questions have been addressed by multiple tests yielding binary (yes/no) outcomes but no clear definition of truth, making LCA an attractive approach for computational biology.

Citation: Chen F, Mackey AJ, Vermunt JK, Roos DS (2007) Assessing Performance of Orthology Detection Strategies Applied to Eukaryotic Genomes. PLoS ONE 2(4): e383. doi:10.1371/journal.pone.0000383

INTRODUCTION

The rapid growth in the availability of genome sequence data, from an ever-increasing range of relatively obscure species, places a premium on the automated identification of orthologs to facilitate functional annotation, and studies on comparative and evolutionary genomics. Homologous proteins share a common ancestry, and may be characterized as either orthologs (which evolve by speciation only) or paralogs (which arise by gene duplication) [1,2]. Orthologs typically retain similar domain architecture and occupy the same functional niche following speciation, while (functionally redundant) paralogs are likely to diverge with new functions through point mutations and domain recombinations [3,4].

The concepts of orthology and paralogy are well-established in classical and molecular systematics [1], and have been extended to describe more complicated situations associated with extensive gene duplications commonly observed in eukaryotic species [4–6]. In- and out-paralogs are analogous to the phylogenetic concepts in- and out-groups, denoting genes duplicated subsequent or prior to speciation, respectively. Recent duplications yield in-paralogs that may exhibit a many-to-one or many-to-many ortholog relationship with genes in the other species (termed co-orthologs).

Several strategies have been employed to distinguish probable (co-)orthologs from paralogs, as summarized in Table 1: phylogeny-based methods include RIO (Resampled Inference of Orthology) [7] and Orthostrapper/HOPS (Hierarchical grouping of Orthologous and Paralogous Sequences) [8,9]; methods based on evolutionary distance metrics include RSD (Reciprocal Smallest Distance) [10,11]; BLAST-based methods include Reciprocal Best Hit (RBH), COG (Cluster of Orthologous Groups) [12–15]/KOG (euKaryotic Orthologous Groups) [15], and Inparanoid [5,16]. The problem of orthology detection is particularly acute for eukaryotic genomes, because of their large size, the difficulty of defining accurate gene models, the complexity of protein domain architectures, and rampant gene duplications [3,17]. To address these difficulties, we previously developed the OrthoMCL algorithm [18], which improves on RBH by (*i*) recognizing co-ortholog relationships (Figure 1), (*ii*) using a normalization step to correct for systematic biases when comparing specific pairs of genomes, and (*iii*) using a Markov graph clustering (MCL) algorithm [19] to define ortholog groups. OrthoMCL and Inparanoid exhibit similar performance when comparing two species, but the former is extensible to cluster orthologs across

Academic Editor: Cedric Feschotte, Pasteur Institute, France

Received March 6, 2007; Accepted March 13, 2007; Published April 18, 2007

Funding: This research was supported by grants from the NIH, and implementation by a Bioinformatics Resource Center contract from NIAID; DSR is a Senior Scholar in Global Infectious Diseases of the Ellison Medical Foundation.

Competing Interests: The authors have declared that no competing interests exist.

* To whom correspondence should be addressed. E-mail: droos@sas.upenn.edu

¤ Current address: Informatics, GlaxoSmithKline, Collegeville, Pennsylvania, United States of America

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]

Achieving high accuracy in orthology inference is essential for many comparative, evolutionary and functional genomic analyses, yet the true evolutionary history of genes is generally unknown and orthologs are used for very different applications across phyla, requiring different precision–recall trade-offs. As a result, it is difficult to assess the performance of orthology inference methods. Here, we present a community effort to establish standards and an automated web-based service to facilitate orthology benchmarking. Using this service, we characterize 15 well-established inference methods and resources on a battery of 20 different benchmarks. Standardized benchmarking provides a way for users to identify the most effective methods for the problem at hand, sets a minimum requirement for new tools and resources, and guides the development of more accurate orthology inference methods.

Evolutionarily related genes (homologs) across different species are often divided into gene pairs that originated through speciation events (orthologs) and gene pairs that originated through duplication events (paralogs)[1]. This distinction is useful in a broad range of contexts, including phylogenetic tree inference, genome annotation, comparative genomics and gene function prediction[2–4]. Accordingly, dozens of methods[5] and resources[6–8] for orthology inference have been developed.

Because the true evolutionary history of genes is typically unknown, assessing the performance of these orthology inference methods is not straightforward. Several indirect approaches have been proposed. Based on the notion that orthologs tend to be functionally more similar than paralogs (a notion now referred to as the ortholog conjecture[9–12]), Hulsen *et al.*[13] used several measures of functional conservation (coexpression levels, protein–protein interactions and protein domain conservation) to benchmark orthology inference methods. Chen *et al.*[14] proposed an unsupervised learning approach based on consensus among different orthology methods. Altenhoff and Dessimoz[15] introduced a phylogenetic benchmark measuring the concordance between gene trees reconstructed from putative orthologs and undisputed species trees. More recently, several 'gold standard' reference sets, either manually curated[16,17] or derived from trusted resources[18], have been used as benchmarks. Finally, Dalquen *et al.*[19] used simulated genomes to assess orthology inference in the presence of varying amounts of duplication, lateral gene transfer and sequencing artifacts.

This wide array of benchmarking approaches poses considerable challenges to developers and users of orthology methods. Conceptually, the choice of an appropriate benchmark strongly depends on the application at hand. Practically, most methods are not available as stand-alone programs and thus cannot easily be

[1]Department of Computer Science, ETH Zurich, Zurich, Switzerland. [2]Computational Biochemistry Research Group, Swiss Institute of Bioinformatics, Zurich, Switzerland. [3]Swiss-Prot Group, Swiss Institute of Bioinformatics, Geneva, Switzerland. [4]Bioinformatics and Genomics Programme, Centre for Genomic Regulation, Barcelona Institute of Science and Technology, Barcelona, Spain. [5]Universitat Pompeu Fabra, Barcelona, Spain. [6]Yeast and Basidiomycete Research Group, CBS-Fungal Biodiversity Centre, Utrecht, the Netherlands. [7]Department of Genetics, Evolution, and Environment, University College London, London, UK. [8]Center for Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, USA. [9]Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany. [10]Department of Life Sciences, Natural History Museum, London, UK. [11]Université Paris-Sud, Laboratoire de Recherche en Informatique, Orsay, France. [12]Université Paris-Sud, Institute for Integrative Biology of the Cell, Orsay, France. [13]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, UK. [14]Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland. [15]Bioinformatics Systems Biology Group, Swiss Institute of Bioinformatics, Zurich, Switzerland. [16]Germany Molecular Medicine Partnership Unit, University Hospital Heidelberg and European Molecular Biology Laboratory, Heidelberg, Germany. [17]Max Delbrück Centre for Molecular Medicine, Berlin, Germany. [18]LIGC, Computer Science Department, ICube, University of Strasbourg, Strasbourg, France. [19]Vital-IT, Swiss Institute of Bioinformatics, Lausanne, Switzerland. [20]Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland. [21]Department of Bioengineering, University of California, Berkeley, California, USA. [22]The Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [23]A list of authors who are consortium members is provided at the end of the manuscript. [24]Institució Catalana de Recerca I Estudis Avançats, Barcelona, Spain. [25]Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, California, USA. [26]Division of Bioinformatics, Department of Preventive Medicine, University of Southern California, Los Angeles, California, USA. [27]Stockholm Bioinformatics Center, Department of Biochemistry and Biophysics, Stockholm University, Science for Life Laboratory, Solna, Sweden. [28]Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland. [29]Department of Computer Science, University College London, London, UK. [30]Swiss Institute of Bioinformatics, Biophore Building, Lausanne, Switzerland. Correspondence should be addressed to C.D. (christophe.dessimoz@unil.ch).

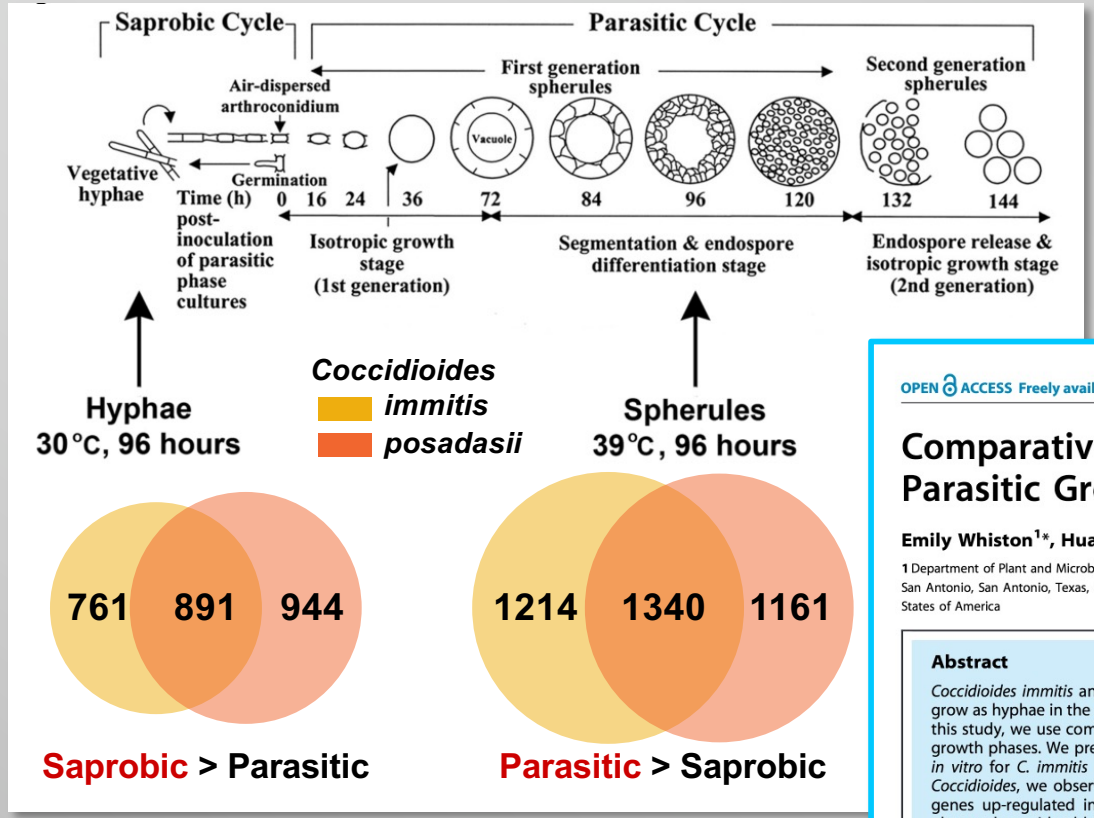
RECEIVED 28 JANUARY; ACCEPTED 9 MARCH; PUBLISHED ONLINE 4 APRIL 2016: DOI:10.1038/NMETH.3830

Protein Orthology for Comparative Genomics:

- ***Genome Comparisons** (synteny)*
- ***Functional Comparisons** (particularly for proteins)*
  - ***how make reliable inferences … based on functional data in other species?***
- ***Protein Homology: Orthology & Paralogy***
- ***Methods based on:***
  - ***functional characterization** (typically unavailable at genomic scale)*
  - ***structure** (ditto … and not necessarily = function)*
  - ***expert curation** (labor-intensive, especially for eukaryotic taxa)*
  - ***synteny** (common for closely-related taxa, but decays rapidly)*
  - ***protein sequence similarity** – e.g. OrthoMCL, OrthoFinder (FungiDB, Mycocosm)*
  - ***domain structure** – e.g. EggNOG, Compara (Ensembl)*
  - ***phylogeny** – e.g. OMA, OrthoFinder2*

  Note: Benchmarking is difficult without a gold standard (Q4O)!
- ***Orthology applications:***
  - ***Validation of genome assemblies & completeness (e.g. BUSCO)***
  - ***Gene gains, losses, amplification, horizontal transfer (phyletic pattern profiling)***
  - ***Functional inference for understudied taxa (GO terms, etc)***

Transcriptional differences between Parasitic & Saprobic stage C. immitis & C. posadasii



OPEN ACCESS Freely available online

PLoS one

Comparative Transcriptomics of the Saprobic and Parasitic Growth Phases in *Coccidioides* spp

Emily Whiston[1]*, Hua Zhang Wise[2], Thomas J. Sharpton[3], Ginger Jui[1], Garry T. Cole[2], John W. Taylor[1]

1 Department of Plant and Microbial Biology, University of California, Berkeley, California, United States of America, **2** Department of Biology, The University of Texas at San Antonio, San Antonio, Texas, United States of America, **3** The J. David Gladstone Institutes, University of California San Francisco, San Francisco, California, United States of America

Abstract

Coccidioides immitis and *C. posadasii*, the causative agents of coccidioidomycosis, are dimorphic fungal pathogens, which grow as hyphae in the saprobic phase in the environment and as spherules in the parasitic phase in the mammalian host. In this study, we use comparative transcriptomics to identify gene expression differences between the saprobic and parasitic growth phases. We prepared Illumina mRNA sequencing libraries for saprobic-phase hyphae and parasitic-phase spherules *in vitro* for *C. immitis* isolate RS and *C. posadasii* isolate C735 in biological triplicate. Of 9,910 total predicted genes in *Coccidioides*, we observed 1,298 genes up-regulated in the saprobic phase of both *C. immitis* and *C. posadasii* and 1,880 genes up-regulated in the parasitic phase of both species. Comparing the saprobic and parasitic growth phases, we observed considerable differential expression of cell surface-associated genes, particularly chitin-related genes. We also observed

Table 1. Top 15 genes with significantly higher expression (up-regulated) in the saprobic phase.

Up in SAPROBIC stage *Coccidioides immitis* vs *C.posadasii*

Fold difference*	Annotation	Gene ID
185	Conserved protein (insect antifreeze protein repeat domain, predicted secreted)	CIMG_00925
166	Acetyltransferase	CIMG_07556
106	Acetamidase	CIMG_02374
101	Conserved hypothetical protein (predicted secreted)	CIMG_03870
94	Fungal hydrophobin (predicted secreted)	CIMG_06615
75	Conserved protein (PAN domain, predicted secreted)	CIMG_09824
53	Conserved protein (zinc-finger domain)	CIMG_00099
53	Conserved hypothetical protein	CIMG_06344
43	Putative serine proteinase	CIMG_09304
41	Cell wall synthesis protein (beta-glucosidase domain, SUN family, predicted secreted)	CIMG_05254
34	Hypothetical protein (predicted secreted)	CIMG_07839
31	Hypothetical protein	CIMG_13374
31	Helix-loop-helix transcription factor	CIMG_02390
29	Conserved hypothetical protein (pyridine nucleotide-disulphide oxidoreductase domain, predicted secreted)	CIMG_07557
24	Prp4 (CRoW domain-containing protein, predicted secreted)	CIMG_07303

Table 2. Top 15 genes with significantly higher expression (up-regulated) in the parasitic phase.

Up in PARASITIC stage *Coccidioides immitis* & *C.posadasii*

Fold difference*	Annotation	Gene ID
125	Conserved hypothetical protein (DUF 536)	CIMG_09539
68	Heat shock protein 30 (Hsp20/alpha-crystallin domain)	CIMG_01749
31	Conserved hypothetical protein	CIMG_12822
30	Conserved hypothetical protein (YCII-related domain)	CIMG_07089
29	Conserved hypothetical protein	CIMG_13084
28	Hypothetical protein	CIMG_11522
28	Hypothetical protein	CIMG_05235
26	Polysaccharide deacetylase (Arp2/3 complex subunit Arc16)	CIMG_02628
24	Conserved hypothetical protein (predicted secreted)	CIMG_00509
23	Hypothetical protein	CIMG_11203
19	Spherule outer-wall glycoprotein (SOWgp, predicted secreted)	CIMG_04613
18	Conserved hypothetical protein	CIMG_10488
18	Hypothetical protein	CIMG_10670
17	Sphingosine hydroxylase	CIMG_01209
17	Conserved hypothetical protein	CIMG_04740

Download Data Files BETA makes downloading genome-scale files such as genome.fasta or GFF files easy and quick. Please check out the new Beta tool design and Contact Us with comments.

Search for...

expand all | collapse all

Filter the searches below...

- Genes
  - Annotation, curation and identifiers
  - Epigenomics
  - Function prediction
  - Gene models
  - Genetic variation
  - Genomic Location
  - Immunology
  - Orthology and synteny
  - Pathways and interactions
  - Phenotype
  - Protein features and properties
  - Protein targeting and localization
  - Proteomics
  - Sequence analysis
  - Structure analysis
  - Taxonomy
  - Text
  - Transcriptomics
- Organisms
- Popset Isolate Sequences
- Genomic Sequences
- Genomic Segments
- SNPs
- FSTs

Overview of Resources and Tools

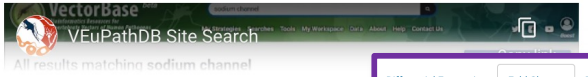
Take a Tour | Getting Started | Search Strategies | Genome Browser | Transcriptomic Resources | Phenotypic Data | Analyze My Data | Downloads | How to Submit Data | Curation and Annotation

Getting Started

VEuPathDB is packed with data, tools and visualizations that can help answer your research questions. We gather data from many sources, analyze according to standard workflows, and present the results for you to mine in a point and click interface. Here's how to get started:

SITE SEARCH: Explore the site; find what you need

Enter a term or ID in the site search box at the top of any page. The site search finds documents and records that contain your term and returns a summary of categorized matches. Its easy to find genes, pathways, searches, data sets and more with the site search.



Read More

Tutorials and Exercises

Apollo: Manual gene annotation

Structural and functional community curation with Apollo, a real time collaborative genome annotation and curation platform

Gene Pages

Explore the data and images on gene pages

Gene Pages

Genetic Variation

SNPs and CNV based on wh genome sequencing

Genetic Variation

News »

FungiDB 63 Released
WED MAY 03 2023

FungiDB 62 Released
THU MAR 09 2023

FungiDB 61 Released
THU DEC 15 2022

FungiDB 60 Released

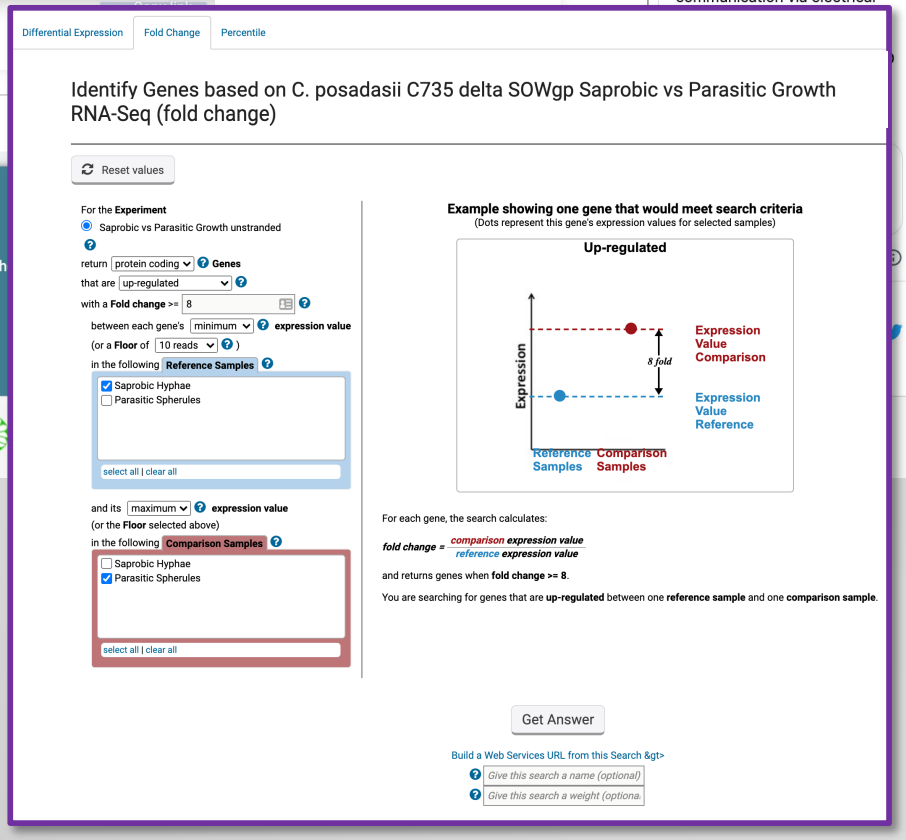
See all news

Tweets from @fungidb

FungiDB Retweeted

Francis M. Ma... @fmarti... · May 6

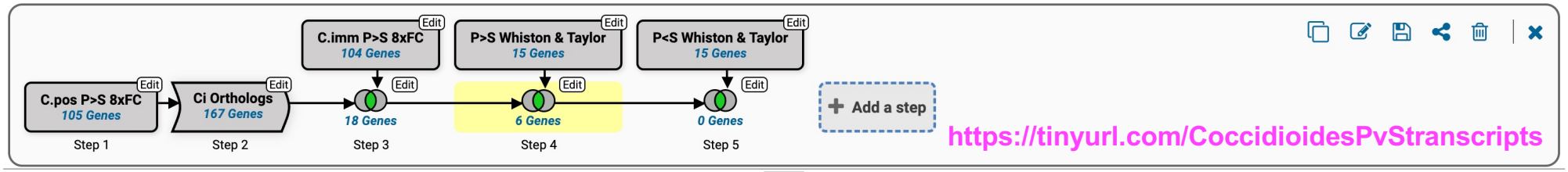
The fungal grid: Fungal communication via electrical



BRC NIAID Bioinformatics Resource Centers | GBC GLOBAL CORE BIODATA RESOURCE

Opened (1) All (340) Public (39) Help

*Coccidioides Parasitic > Saprobic FC (Whiston & Taylor 2012) **

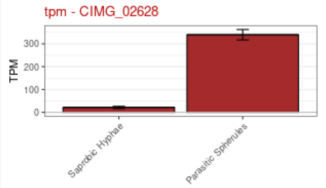
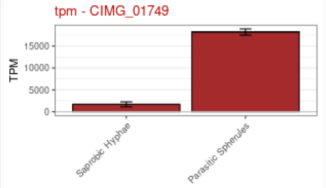
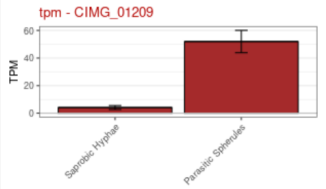
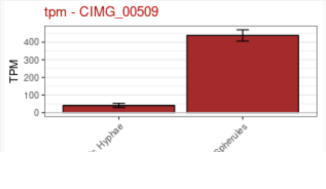


6 Genes (6 ortholog groups)

Gene Results | Genome View | Analyze Results

Rows per page: 100 | Download | Send to... | Add Columns

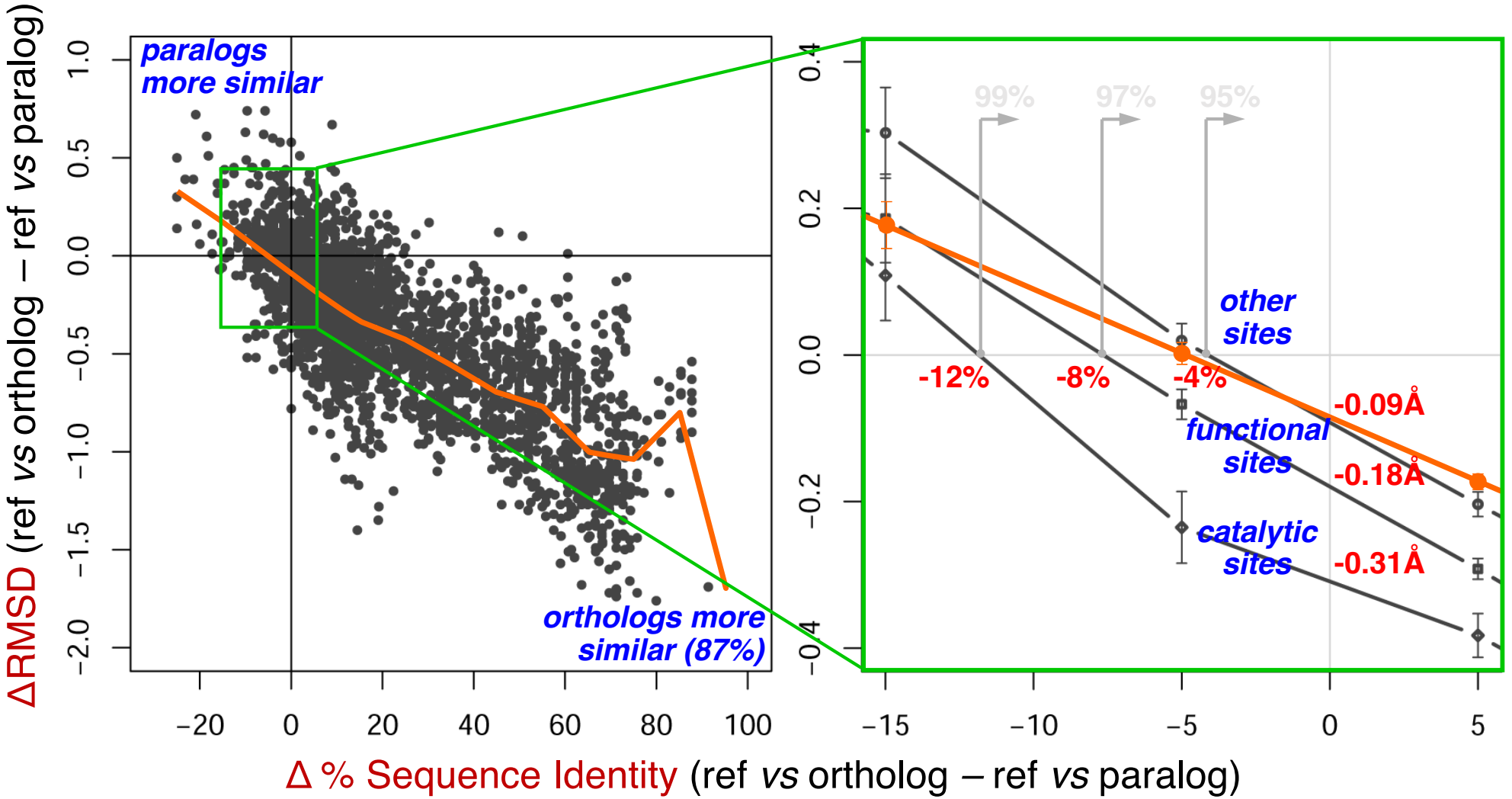
Show Organism Filter

Gene ID	Transcript ID	Genomic Location (Gene)	Product Description	Computed GO Functions	Computed GO Processes	C.i. Saprobic/Parasitic RNA-Seq - tpm Graph
CIMG_02628	CIMG_02628-t26_1	GG704911:1,361,297..1,362,421(-)	Arp2/3 complex subunit Arc16	catalytic activity;hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds	carbohydrate metabolic process	
CIMG_01749	CIMG_01749-t26_1	GG704911:3,854,425..3,855,421(+)	Heat shock protein 30	N/A	N/A	
CIMG_01209	CIMG_01209-t26_1	GG704911:5,319,073..5,320,322(-)	C4-hydroxylase	iron ion binding;oxidoreductase activity	lipid biosynthetic process;obsolete oxidation-reduction process	
CIMG_00509	CIMG_00509-t26_1	GG704911:7,147,867..7,148,818(+)	hypothetical protein	N/A	N/A	


FungiDB
Fungal & Oomycete Informatics Resources
Site search, e.g. NCU06658 or *reductase or "binding protein"
My Strategies Searches Tools My Workspace Data About Help Contact Us
David
Genome Track View Help
Coccidioides immitis RS
Share
Select tracks
GG704911
GG704911:7131721..7164740 (33.02 KI
Go
Annotated Transcripts (UTRs in White when available)
CIMG_00513 histone H1/5
CIMG_12867 hypothetical protein
CIMG_15505 rRNA
CIMG_00508 metalloproteinase 4
CIMG_00506 hypothetical protein
CIMG_00504 flap endonuclease 1
CIMG_00502 glutaryl-CoA dehydrogenase
CIMG_00514 hosphatidylinositol-4-phosphate 5-kinase its3
CIMG_12868 Kinetochore Sim4 complex subunit FTA2, putative
CIMG_00507 Protein kinase-like domain-containing protein
CIMG_00503 glyoxalase/bleomycin resistance protein/dioxygenas
CIMG_00512 hypothetical protein
CIMG_00509 hypothetical protein
CIMG_00505 Family A G protein-coupled receptor-like superfamily protein, putative
Community annotations from Apollo
RNA-Seq Evidence for Introns
Matches Annotation
Unannotated (Strong Evidence)
transcription up-regulated in parasitic spherules (in both C. immitis & C. posadasii
Saprobic vs Parasitic Growth - Saprobic Hyphae (unique) Coverage
Saprobic vs Parasitic Growth - Parasitic Spherules (unique) Coverage
Transcriptional analysis of Coccidioides immitis mycelia and spherules - Mycelia (unique forward) Coverage
Transcriptional analysis of Coccidioides immitis mycelia and spherules - Mycelia (unique reverse) Coverage
Transcriptional analysis of Coccidioides immitis mycelia and spherules - Young spherules (unique forward) Coverage
Transcriptional analysis of Coccidioides immitis mycelia and spherules - Young spherules (unique reverse) Coverage
Transcriptional analysis of Coccidioides immitis mycelia and spherules - Mature spherules (unique forward) Coverage
Transcriptional analysis of Coccidioides immitis mycelia and spherules - Mature spherules (unique reverse) Coverage
Syntenic Sequences and Genes (Shaded by Orthology)
Gene cluster in pathogenic (not nonpath) Coccidioides
… also pathogenic paraocccidioides (but not Aspergillus
cimmRS gene
cimmRS span
cposC735deltSOWgp gene
cposC735deltSOWgp span
cposSilveira gene
cposSilveira span
cposSilveira2022 gene
cposSilveira2022 span
uree1704 gene
uree1704 span
pbraPb03 gene
pbraPb03 span
pbraPb18 gene
pbraPb18 span
plutPb01 gene
plutPb01 span
sbra5110 gene
sbra5110 span
anidFGSCA4 gene
anidFGSCA4 span
COMMUNITY CHAT

Orthologs display greater structural conservation than paralogs ... particularly at functional sites

Comparison of Reference / Ortholog / Paralog (AOP) triplets
499 catalytic sites (CSA), 4,041 functional sites (PDB)



*M.E. Peterson et al., Protein Science **18:**1306-15 (2009)*

Feng Chen et al. (2011)

Welcome to OrthoMCL where you will find the newest versions of our interface, features, tools and data. Since OrthoMCL went through a major algorithm update, we are keeping the old release 5 available in case needed. Here is a form for sending your feedback about the new site.

Search for...

expand all | collapse all

Filter the searches below...

Ortholog Groups

- % Pairs w/ Similarity
- All Groups
- Avg % Homology
- Avg % Identity
- Avg % Match Length
- Avg E-Value
- EC Number
- Group ID(s)
- Group or Sequence ID
- Number of Sequences
- Number of Taxa
- PFam ID or Keyword
- Phyletic Pattern
- Text Terms

Proteins

- All Proteins
- BLAST
- EC Number

Overview of Resources and Tools

OrthoMCL FAQ | About OrthoMCL | Types of Searches in OrthoMCL | Understanding Group Search Results | Search Strategies | Phyletic Pattern Search | Transforming Results | Assign Proteins to Groups | Downloads

About OrthoMCL

OrthoMCL is a genome-scale algorithm for grouping orthologous protein sequences. Such orthologous sequences not only share evolutionary history, but also share function. Thus, ortholog prediction is important in predicting the function of newly identified proteins. Indeed, detection of orthologs has become more widespread with the rapid progress in genome sequencing and the discovery of protein sequences (Glover et al. 2019). Importantly, proteins in OrthoMCL groups have been shown to display a high degree of functional conservation (e.g., a group's proteins have consistent EC numbers) (Li et al. 2003), highlighting that OrthoMCL is useful for functional annotation of newly sequenced genomes.

OrthoMCL not only identifies groups shared by proteins from two or more species, but also groups representing species-specific gene expansion families. To achieve this, the OrthoMCL algorithm starts with reciprocal best BLAST hits within each proteome as potential in-

Read More

Tutorials and Exercises

Grid view

Orthology Phyletic Patterns	OrthoMCL tutorial	Search Strategies (Basics)	Search Strategies (Advanced)
	Learn standard searches and		

News and Tweets

COMMUNITY CHAT

saur	cper	bant	lmon	spne	cbot	bmal	bpse	rsol	yent	sent	cbur	vcho	ypes	sfle	ftul	ecol	cjej	wsuc	rpro	wend	bsui	atum	rtyp	gsul	cpne	mtub	drad	deth	ctep	tmar	mlep	syne	rbal	tpal	aaeo	nmar	hbut	smar	ssol	msed
1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	0	0	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0	0

ihos	cmaq	ckor	nequ	halo	tvol	mmar	hwal	mjan	aful	msmi	lbra	tbru	lmex	tviv	tcon	tbrg	lmaj	linf	tcru	einv	edis	ddis	ehis	gthe	rcom	atha	csat	micr	ppat	otau	crei	vcar	tpse	cmer	tthe	pviv	pfal	pber	pyoe	pkno
0	1	1	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	2	1	1	1	1	1	1	1	0	1	1	1	1	1

pcha	tpar	tann	bbov	cmur	tgon	ncan	cpar	chom	aory	ylip	spom	psti	ncra	scer	egos	cimm	cpos	calb	mgri	klac	dhan	anid	afum	gzea	cgla	ecun	eint	ebie	pchr	lbic	cneg	cneo	isca	dmel	aaeg	bmor	amel	cpip	phum	apis
1	0	0	0	1	1	1	0	0	1	1	1	1	1	1	1	1	1	2	0	1	1	1	1	1	1	0	0	0	1	1	1	1	0	0	0	0	0	0	0

agam	nvec	tadh	drer	trub	tnig	cint	oana	mor	hsap	mmus	mdom	mmul	clup	ptro	ecab	ggal	cele	bmaa	cbri	sman	mbre	tvag	glae	glab	pram	glam
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0

Description

Edge Options

Edge Type

- [x] Ortholog
- [x] Coortholog
- [x] Inparalog
- [ ] Other Similarities

E-Value Cutoff

Max E-Value: 1E-35

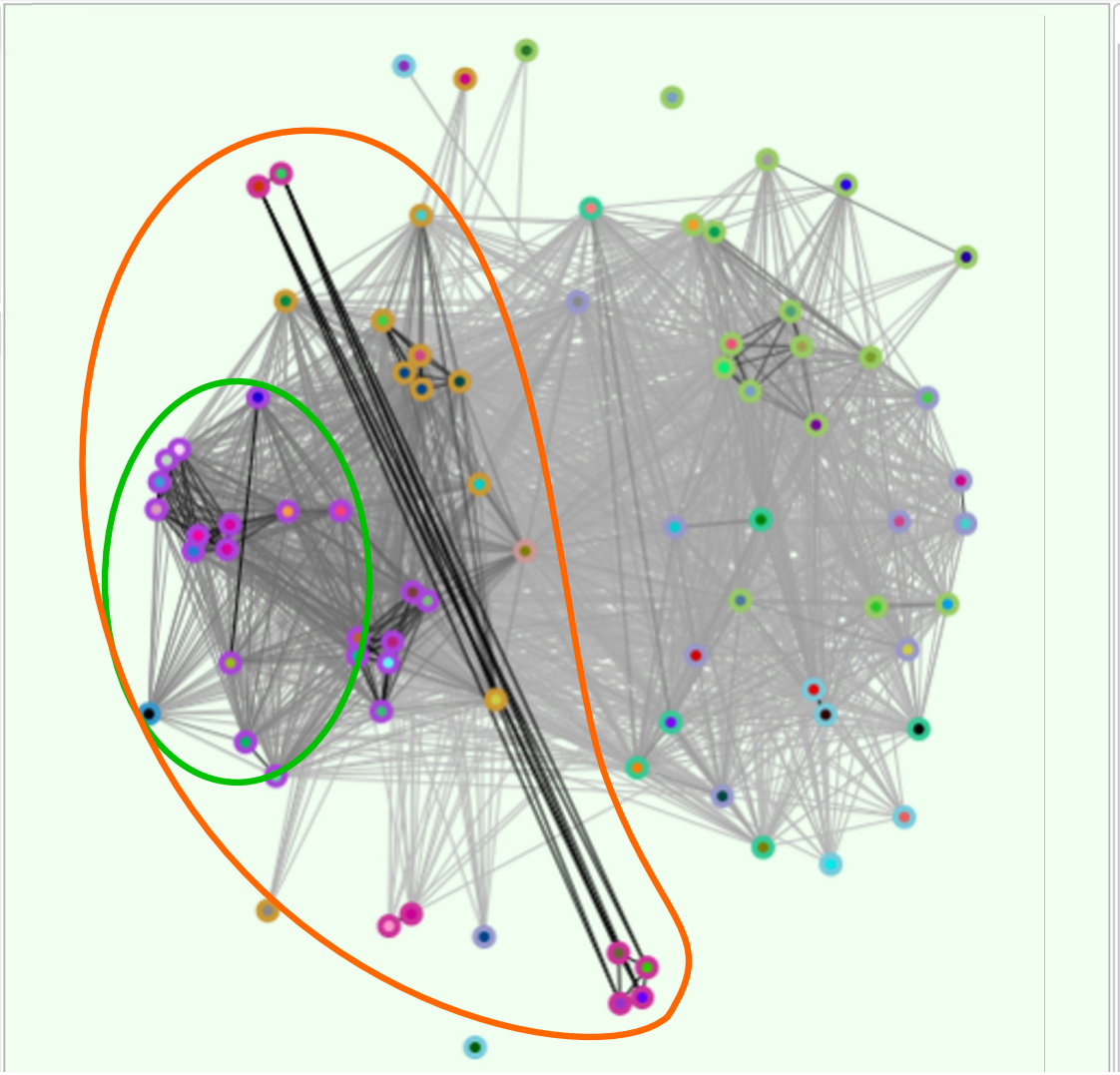
Node Options

Show Nodes By

(•) Taxa ( ) EC Numbers ( ) PFam Domains

Mouse over a taxon legend to highlight sequences of that taxon.

bant (1)	cbot (1)	cper (1)
lmon (1)	saur (1)	spne (1)
atum (1)	bsui (1)	bmal (1)
bpse (1)	cjej (1)	cbur (1)
ecol (1)	ftul (1)	gsul (1)
rsol (1)	sent (2)	sfle (1)
vcho (1)	wsuc (1)	yent (1)
ypes (1)	aaeo (1)	cpne (1)
ctep (1)	deth (1)	drad (1)
mtub (1)	mlep (1)	rbal (1)
syne (1)	tmar (1)	cmaq (1)
ckor (1)	halo (1)	hwal (1)
nmar (1)	tvol (1)	cmur (1)
ncan (1)	pber (1)	pcha (1)

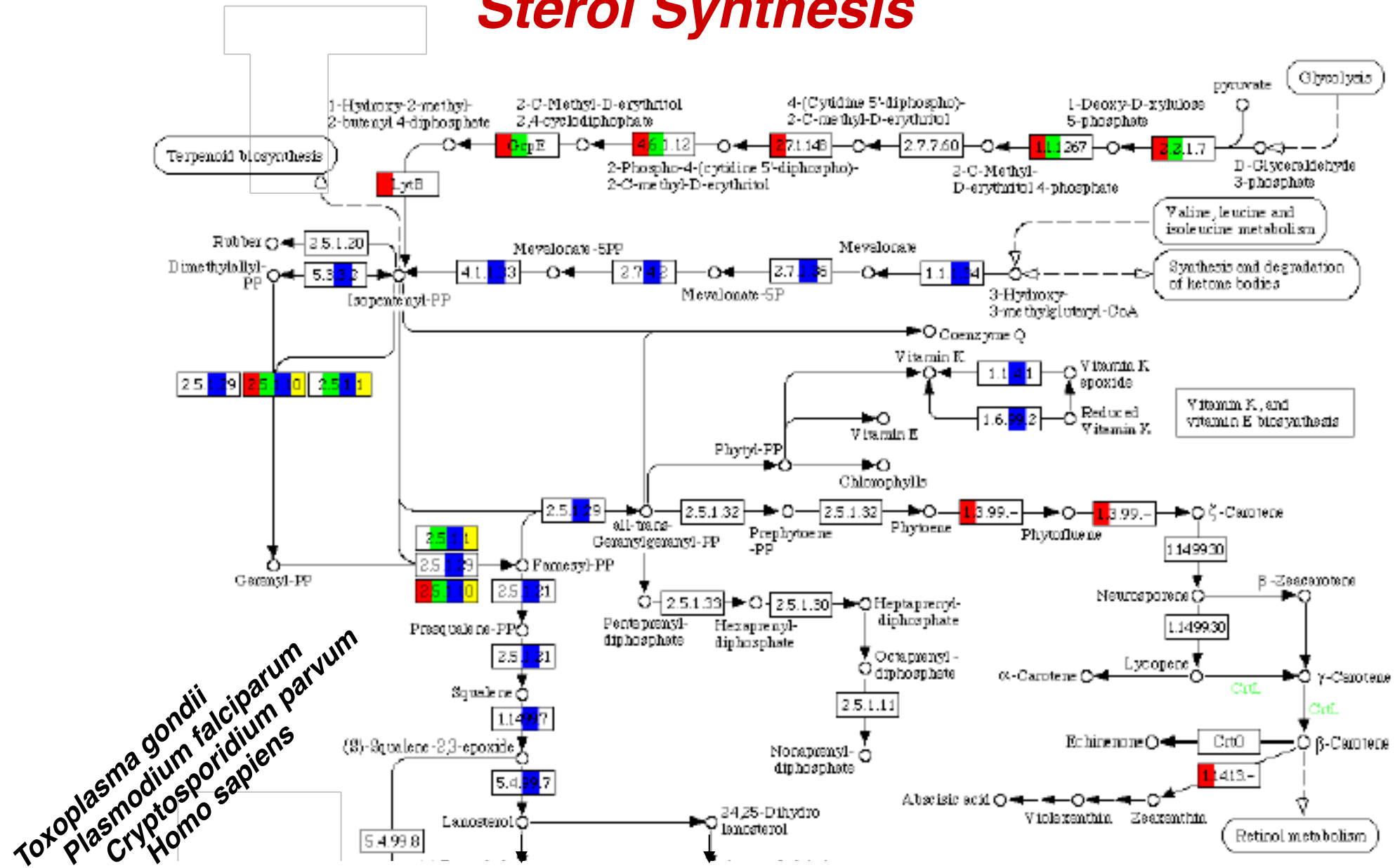


Sequence List | Node Detail

Search:

Accession	Taxon	L
aaeo|NP_214013	aaeo	399
afum|Afu2g09840	afum	463
anid|AN6032	anid	451
aory|AO090011000651	aory	464
atha|NP_194729	atha	554
atha|NP_564953	atha	484
atum|NP_354359	atum	261
bant|YP_016674	bant	280
bmal|YP_102541	bmal	296
bpse|YP_333020.1	bpse	296
bsui|NP_698035	bsui	279
calb|calb_sc5314__orf19.579	calb	789
calb|CAWT_04470	calb	789
cbot|YP_001786179.1	cbot	395
cbur|NP_820340	cbur	297
cgla|XP_448052	cgla	807
cimm|CIMG_07826T0	cimm	463
cjej|YP_002344015	cjej	380
ckor|YP_001737263	ckor	403
cmaq|YP_001540023	cmaq	303
cmer|CMQ278C	cmer	500
cmur|CMU_032400	cmur	217
cneg|CNAG_02786T0	cneg	736
cneo|CNBG_3441T0	cneo	651
cper|YP_695722	cper	269

Painting Orthologs onto KEGG Pathways: Sterol Synthesis



Toxoplasma gondii
Plasmodium falciparum
Cryptosporidium parvum
Homo sapiens

to cholesterol

http://EuPathDB.org

Orthology-based inference

- Eukaryota (EUKA):
  - Alveolates (ALVE):
    - Ciliates (CILI): tthe
    - Apicomplexa (APIC):
      - Coccidia (COCC): chom cmu ✗cpar ncan ✓tgon
      - Aconoidasida (ACON):
        - Haemosporida (HAEM): pber pcha ✓pfal pkno pviv pyoe
        - Piroplasmida (PIRO): bbov tann ✓tpar
  - Amoebozoa (AMOE): ddis ehis edis einv
  - Euglenozoa (EUGL): lbra linf lmaj lmex tbru tbrg tcon tcru tviv
  - ✓Viridiplantae (VIRI):
    - Streptophyta (STRE): atha osat ppat rcom
    - Chlorophyta (CHLO): crei otau
    - Rhodophyta (RHOD): cmer
    - Cryptophyta (CRYP): gthe
    - Bacillariophyta (BACI): tpse
  - Fungi (FUNG):
    - Microsporidia (MICR): ecun
    - Basidiomycota (BASI): cneo lbic pchr
    - Ascomycota (ASCO): afum aory cgla cimm cpos dhan egos gzea klac
  - ✗Metazoa (META):
    - ✗Platyhelminthes (PLAT): sman
    - ✗Nematodes (NEMA): bmaa cbri cele

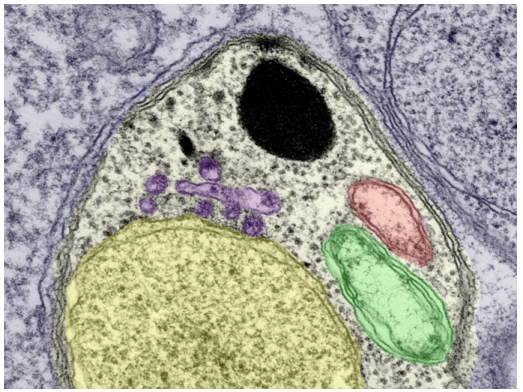
Apicoplast genes should be …

present in:

- *Toxoplasma*
- *Plasmodium*
- *Theileria*
- *Plants*

… but not in:

- *Cryptosporidium*
- *Mammals*



Horizontal gene transfer facilitated the evolution of plant parasitic mechanisms in the oomycetes

Thomas A. Richards[a,b,1], Darren M. Soanes[a], Meredith D. M. Jones[a,b], Olga Vasieva[c], Guy Leonard[a,b], Konrad Paszkiewicz[a], Peter G. Foster[b], Neil Hall[c], and Nicholas J. Talbot[a]

[a]Biosciences, University of Exeter, Exeter EX4 4QD, United Kingdom; [b]Department of Zoology, Natural History Museum, London SW7 5BD, United Kingdom; and [c]School of Biological Sciences, University of Liverpool, Liverpool L69 7ZB, United Kingdom

Edited by W. Ford Doolittle, Dalhousie University, Halifax, Canada, and approved July 27, 2011 (received for review March 31, 2011)

Horizontal gene transfer (HGT) can radically alter the genomes of microorganisms, providing the capacity to adapt to new lifestyles, environments, and hosts. However, the extent of HGT between eukaryotes is unclear. Using whole-genome, gene-by-gene phylogenetic analysis we demonstrate an extensive pattern of cross-kingdom HGT between fungi and oomycetes. Comparative genomics, including the de novo genome sequence of *Hyphochytrium catenoides*, a free-living sister of the oomycetes, shows that these transfers largely converge within the radiation of oomycetes that colonize plant tissues. The repertoire of HGTs includes a large number of putatively secreted proteins; for example, 7.6% of the secreted proteome of the sudden oak death parasite *Phytophthora ramorum* has been acquired from fungi by HGT. Transfers include gene products with the capacity to break down plant cell walls and acquire sugars, nucleic acids, nitrogen, and phosphate sources from the environment. Predicted HGTs also include proteins implicated in resisting plant defense mechanisms and effector proteins for attacking plant cells. These data are consistent with the hypothesis that some oomycetes became successful plant parasites by multiple acquisitions of genes from fungi.

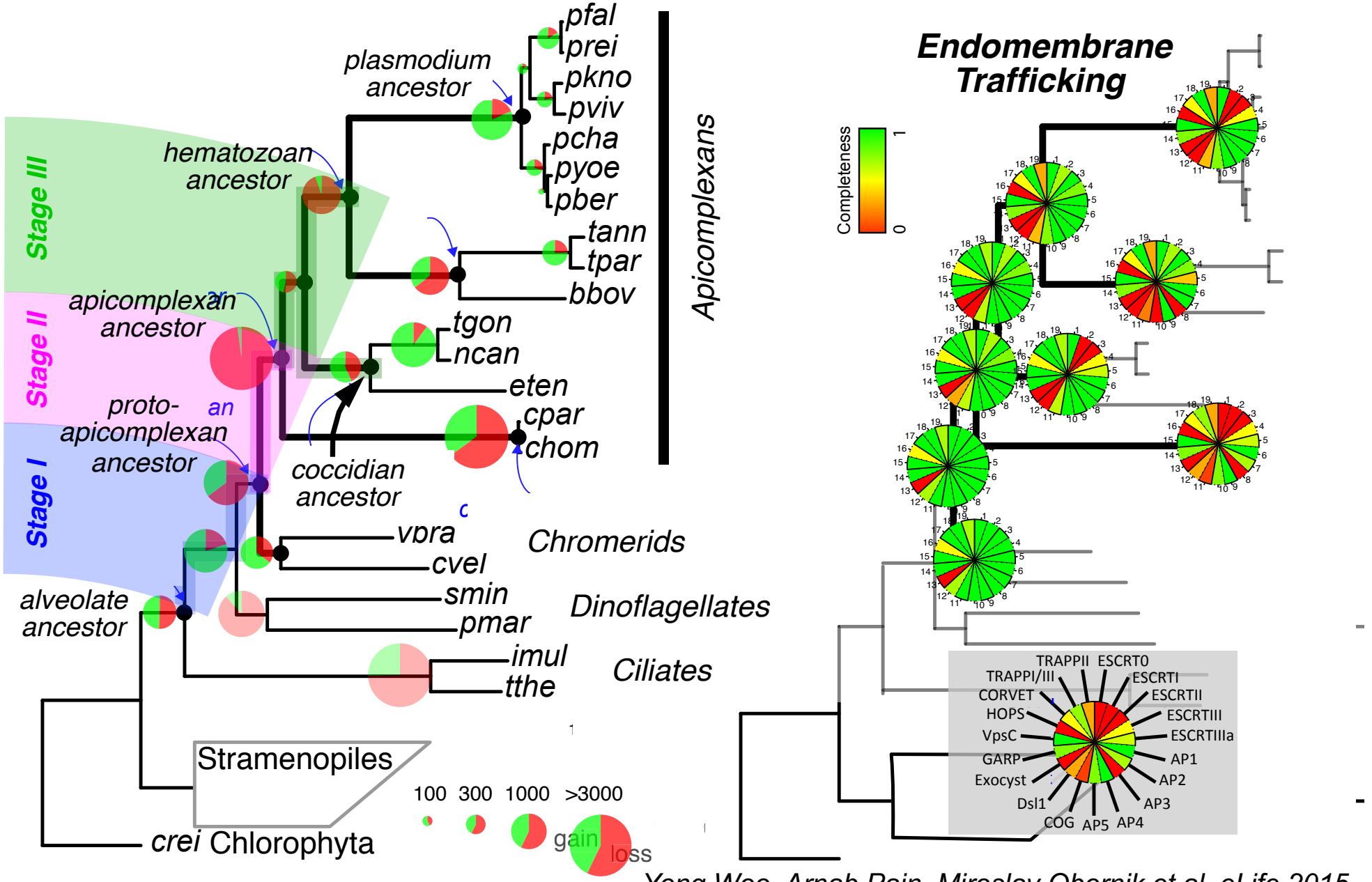
osmotrophy | pseudofungi | lateral gene transfer

Horizontal gene transfer (HGT) involves the transfer of genetic material between reproductively isolated lineages (1, 2) and has been an important factor in prokaryotic evolution (3–5). By contrast, the role of HGT between eukaryotic genomes is *ramorum*, for example, whereas the Irish potato famine of the 19th century was caused by the late blight parasite *Phytophthora infestans*. Important crop diseases caused by fungi include the devastating rice blast disease caused by *M. oryzae* and the rusts, smuts, and mildews that affect wheat, barley, and maize. In this study we report that HGT between fungi and oomycetes has occurred to a far greater degree than hitherto recognized (19). Our previous analysis suggested four strongly supported cases of HGT, but by using a whole-genome, gene-by-gene phylogenetic analysis we now reveal a pattern of 34 transfers and propose that these transfers have been fundamental to the evolution of plant parasitic traits within the oomycetes.

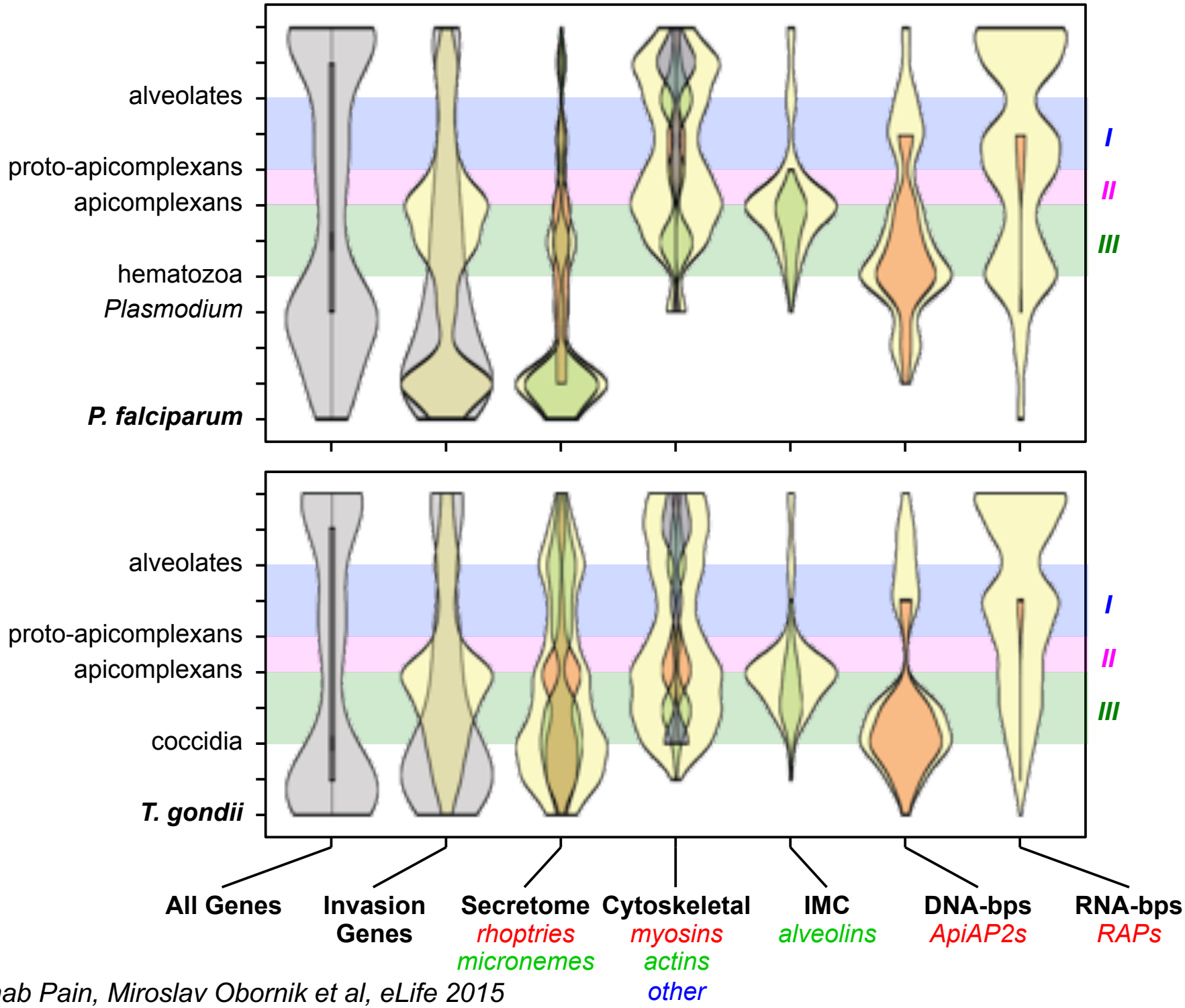
Results and Discussion

Identifying and Testing the Pattern of HGT Between Fungi and Oomycetes. Among the best methods for identifying HGT is to identify a gene phylogeny that places a taxonomic group (recipient) within the branches of a distantly related group (donor) in direct contradiction to the known phylogenetic relationships of the respective taxa (2, 6). To identify all potential gene transfers between fungi and oomycetes, we selected the predicted proteomes of the oomycete species *Phytophthora ramorum*, *Phytophthora sojae*, *Phytophthora infestans*, and *Hyaloperonospora parasitica* (also named *Hyaloperonospora arabidopsidis*) (21–23). We processed each proteome separately, first excluding all putative transposable elements (*SI Materials and Methods*). Then

What happened when? Ortholog group gains and losses during apicomplexan evolution



Yong Woo, Arnab Pain, Miroslav Obornik et al, eLife 2015



Yong Woo, Arnab Pain, Miroslav Obornik et al, eLife 2015

Protein Orthology for Comparative Genomics:

- ***Genome Comparisons** (synteny)*
- ***Functional Comparisons** (particularly for proteins)*
  - ***how make reliable inferences … based on functional data in other species?***
- ***Protein Homology: Orthology & Paralogy***
- ***Methods based on:***
  - ***functional characterization** (typically unavailable at genomic scale)*
  - ***structure** (ditto … and not necessarily = function)*
  - ***expert curation** (labor-intensive, especially for eukaryotic taxa)*
  - ***synteny** (common for closely-related taxa, but decays rapidly)*
  - ***protein sequence similarity** – e.g. OrthoMCL, OrthoFinder (FungiDB, Mycocosm)*
  - ***domain structure** – e.g. EggNOG, Compara (Ensembl)*
  - ***phylogeny** – e.g. OMA, OrthoFinder2*

  Note: Benchmarking is difficult without a gold standard (Q4O)!
- ***Orthology applications:***
  - ***Validation of genome assemblies & completeness (e.g. BUSCO)***
  - ***Gene gains, losses, amplification, horizontal transfer (phyletic pattern profiling)***
  - ***Functional inference for understudied taxa (GO terms, etc)***