

ORTHOLOGY PREDICTION

21.11.2014



- **INTRODUCTION**
- **ORTHOLOGY PREDICTION METHODS**
- **EXAMPLE**
- **EXERCISE**

INTRODUCTION

What is orthology?

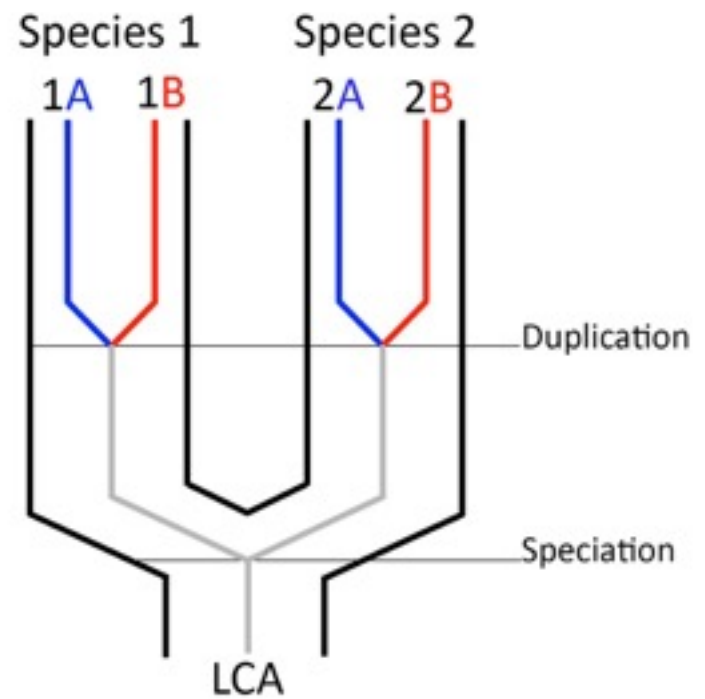
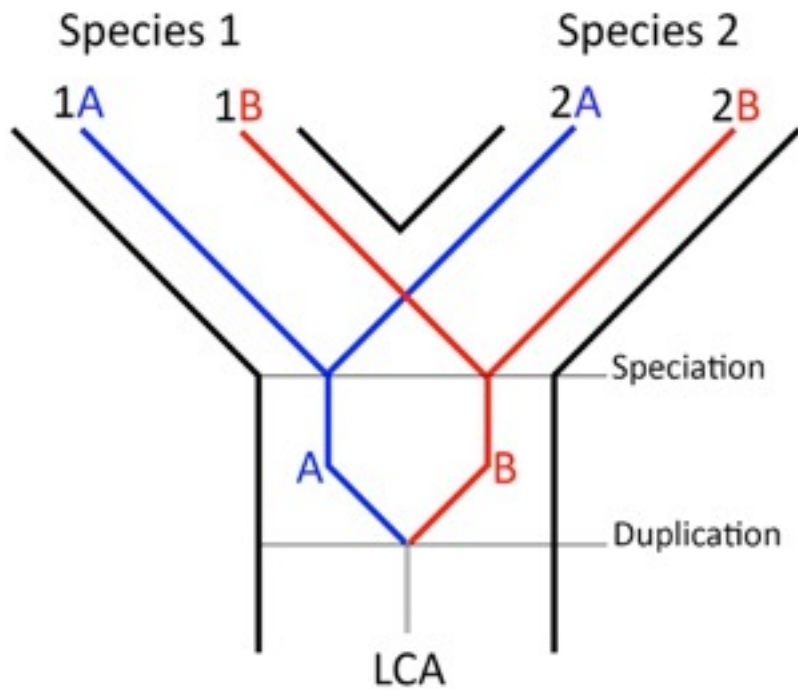
Homology = share ancestry



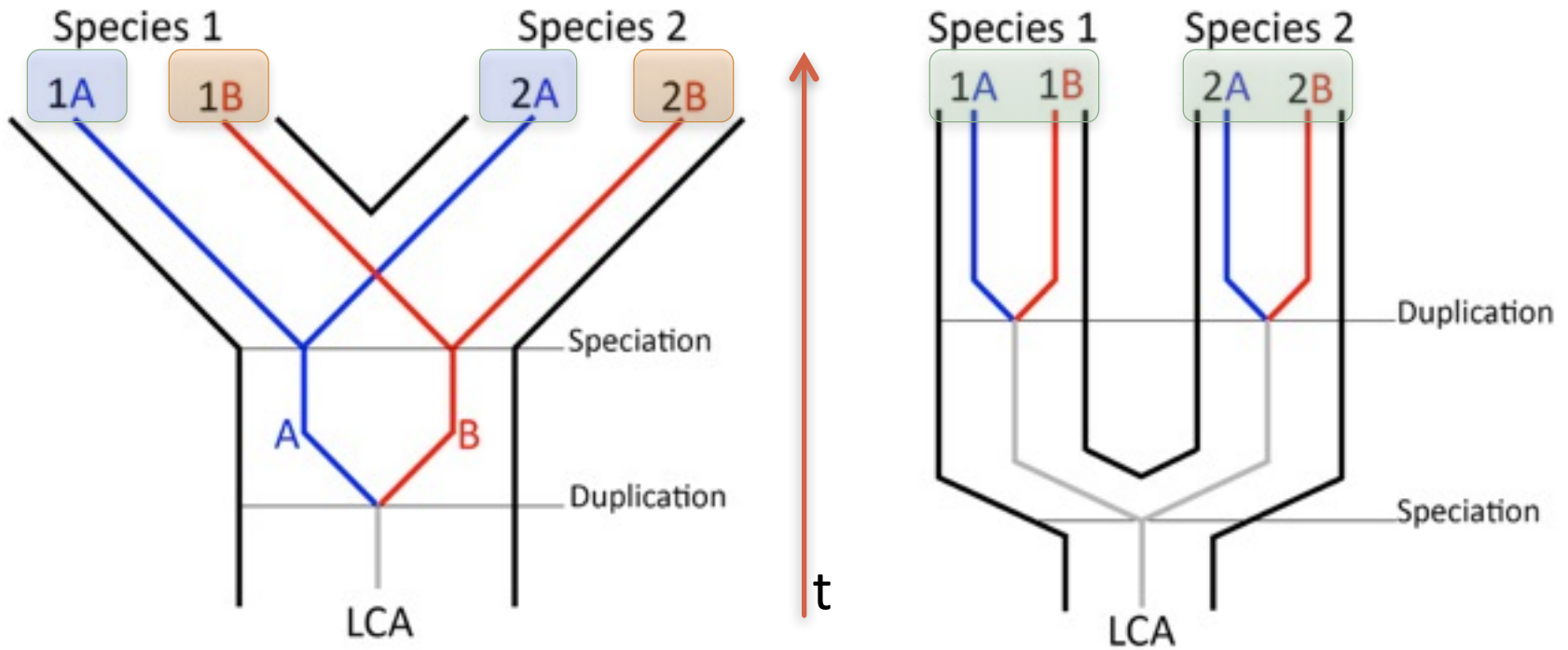
Ortholog = separated by **speciation** event

Paralog = separated by **duplication** event

What is orthology?

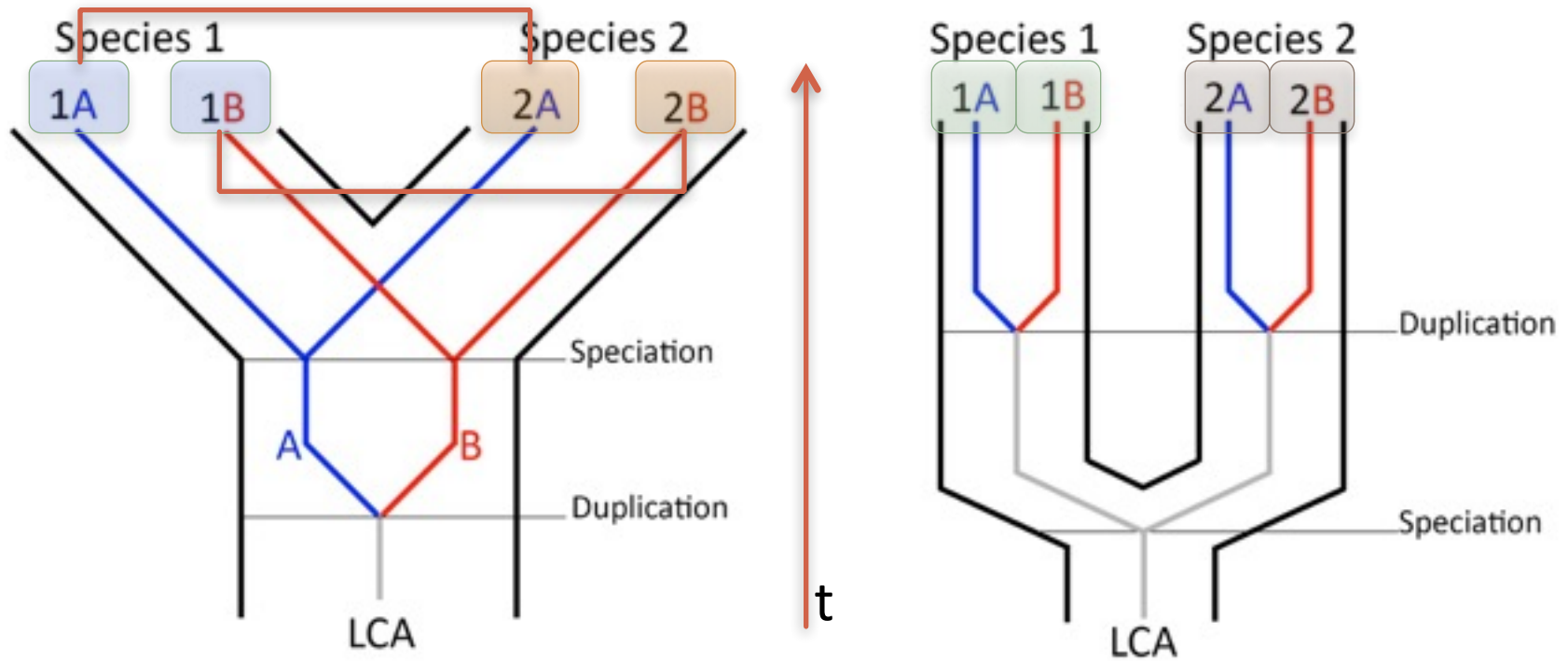


What is orthology?



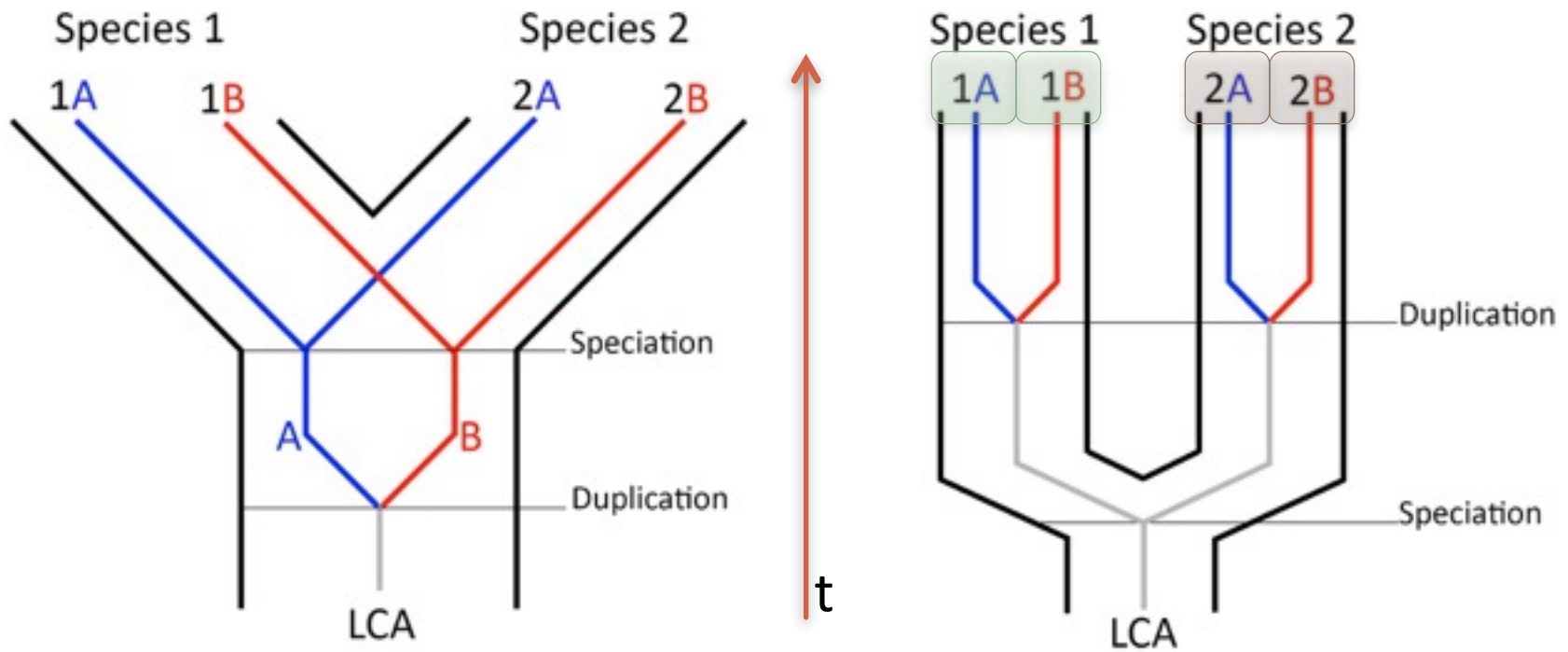
speciation -> ORTHOLOG

What is orthology?



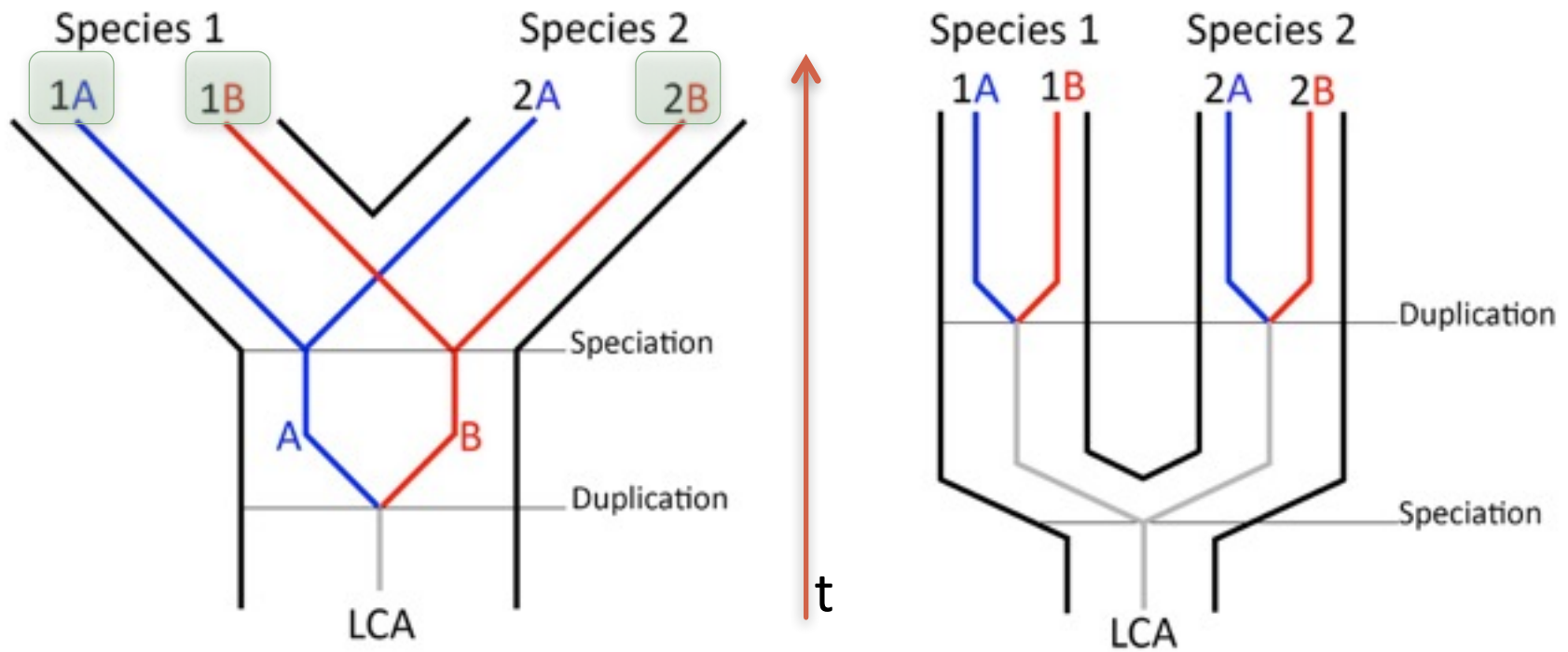
duplication -> PARALOG

What is orthology?



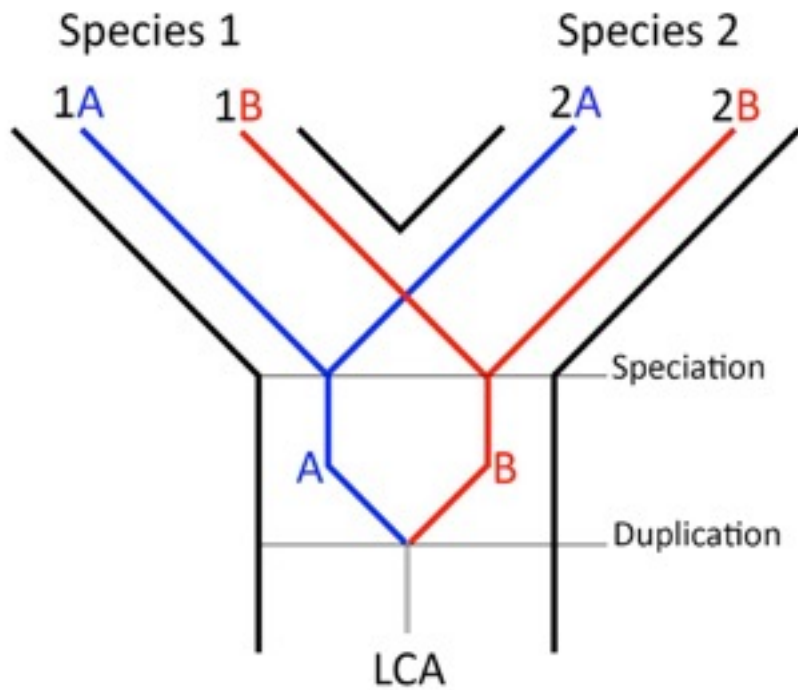
in-paralogs
(duplication occurred after speciation)

What is orthology?

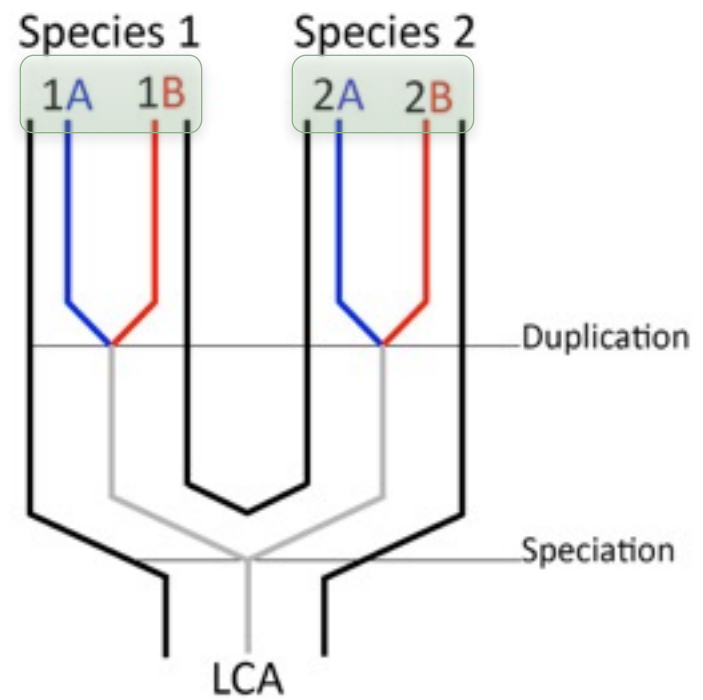


out-paralogs
(duplication occurred before speciation)

What is orthology?

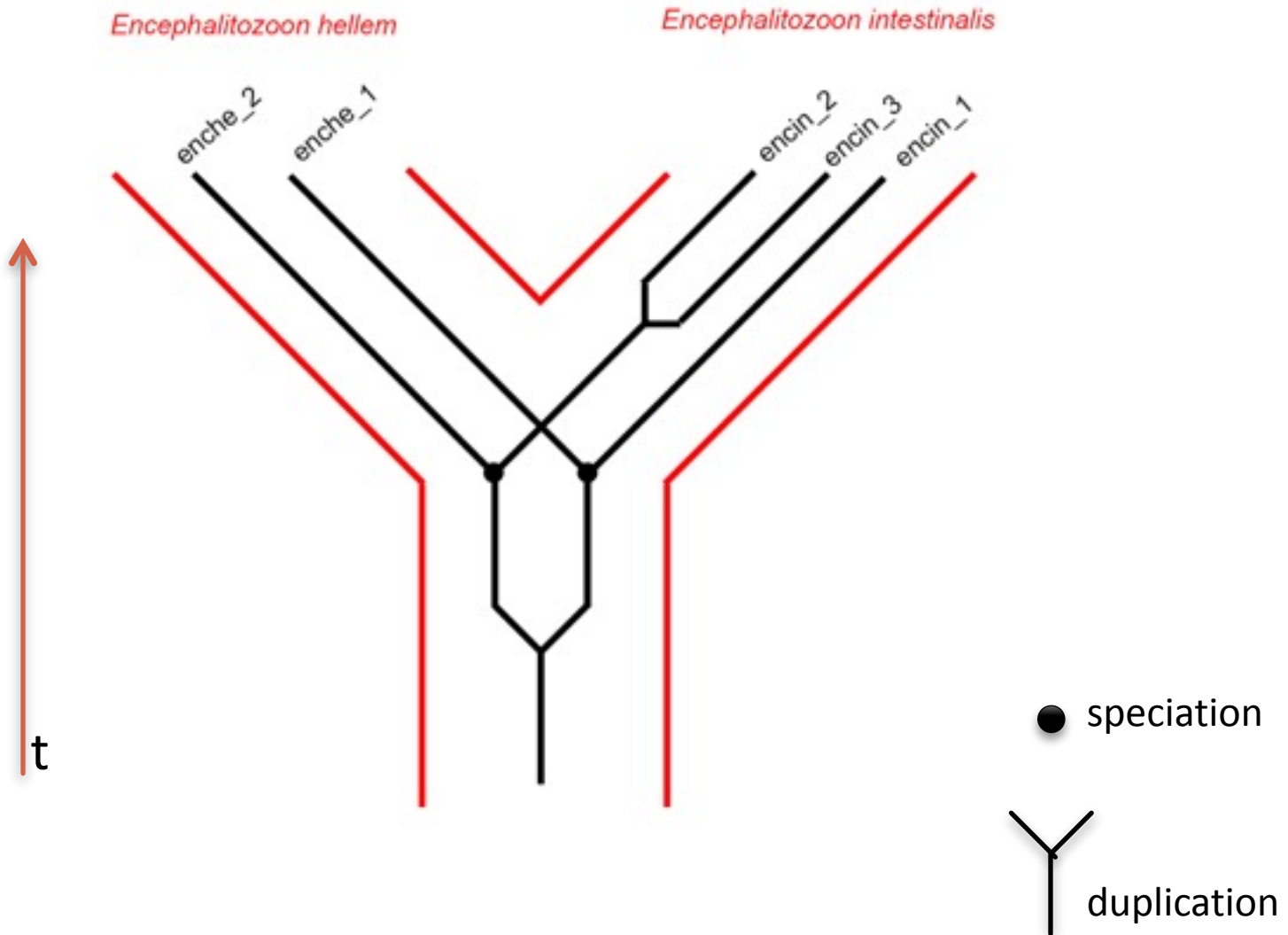


t ↑

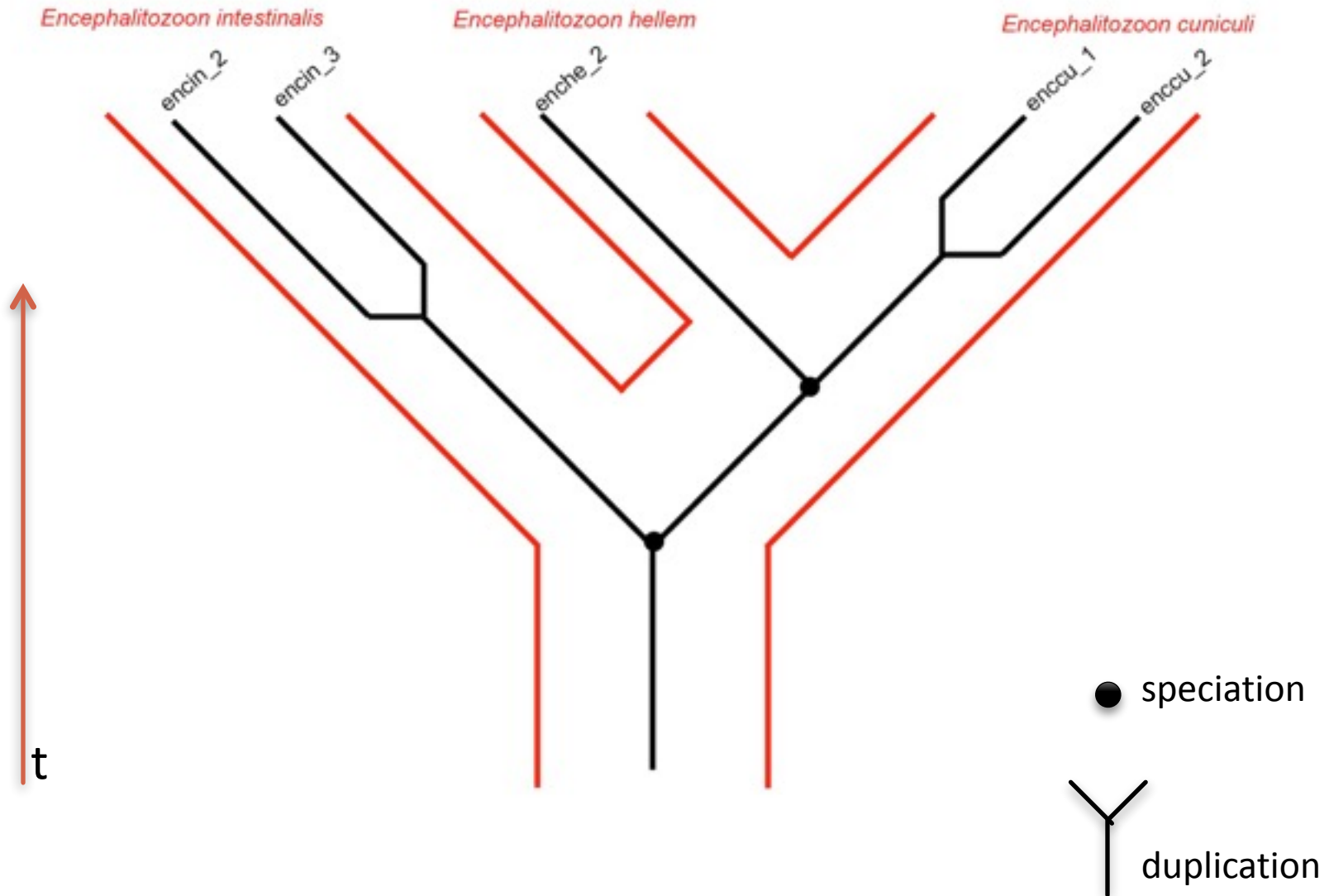


co-orthologs

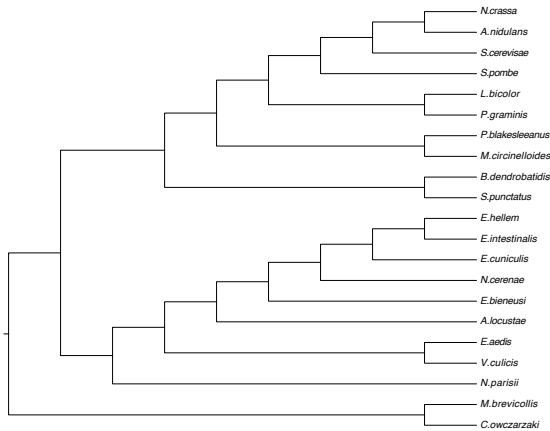
What is orthology?



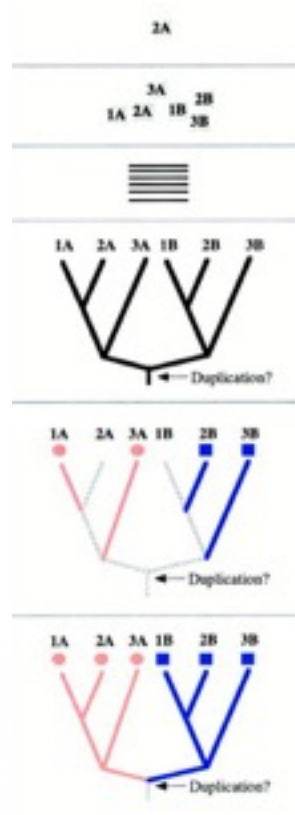
What is orthology?



Why orthology important?



tree reconstruction



CHOOSE GENE(S) OF INTEREST

IDENTIFY HOMOLOGS

ALIGN SEQUENCES

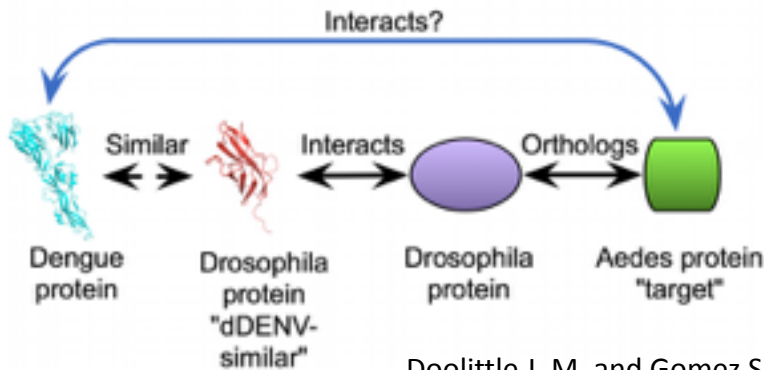
CALCULATE GENE TREE

OVERLAY KNOWN FUNCTIONS ONTO TREE

INFER LIKELY FUNCTION OF GENE(S) OF INTEREST

protein function prediction

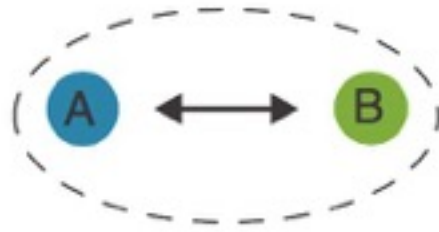
Eisen J. A. (1998)



Doolittle J. M. and Gomez S. M. (2011)

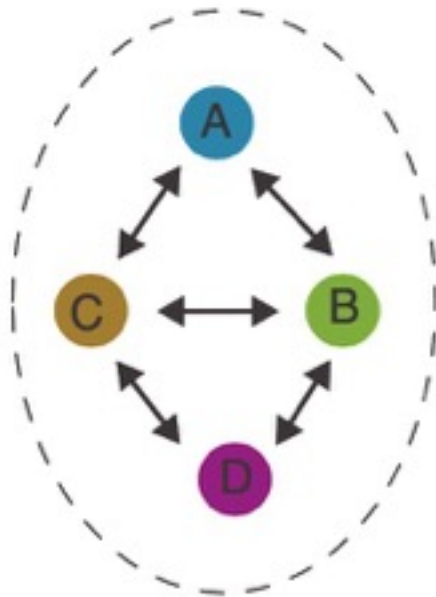
protein-protein interaction analysis

OTHOLOGY PREDICTION METHODS



Pairwise species
comparison

Reciprocal Best Hit
InParanoid



Multi-species
comparison

OrthoMCL

COG (Cluster of Orthologous Groups)

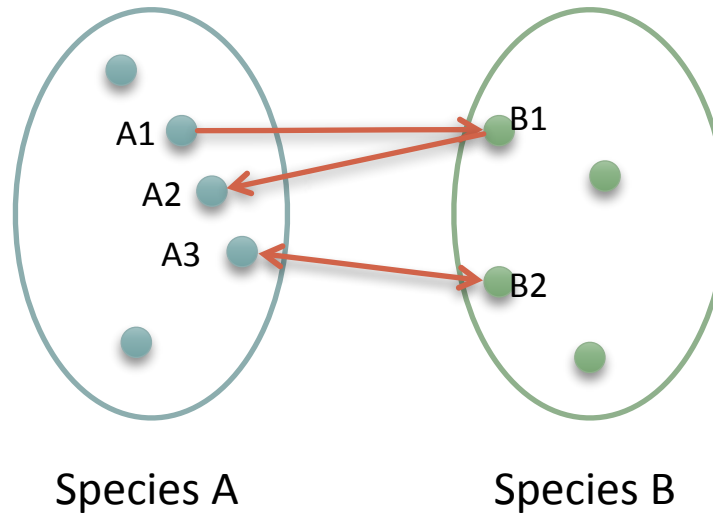
eggNOG (evolutionary genealogy of genes: Non-supervised Orthologous Groups)

Trachana K. et al. (2011)



Pairwise species
comparison

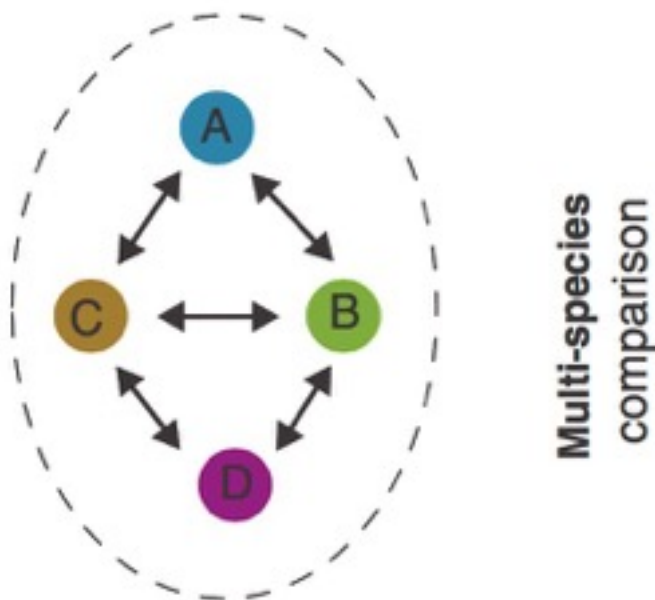
Reciprocal Best Hit



BLAST-based approaches

OrthoMCL: use All-vs-all BLASTp and cluster BBHs using Markov model

COG, eggNOG: identify three-way BBHs in 3 different species and merge triangles that share a common side.

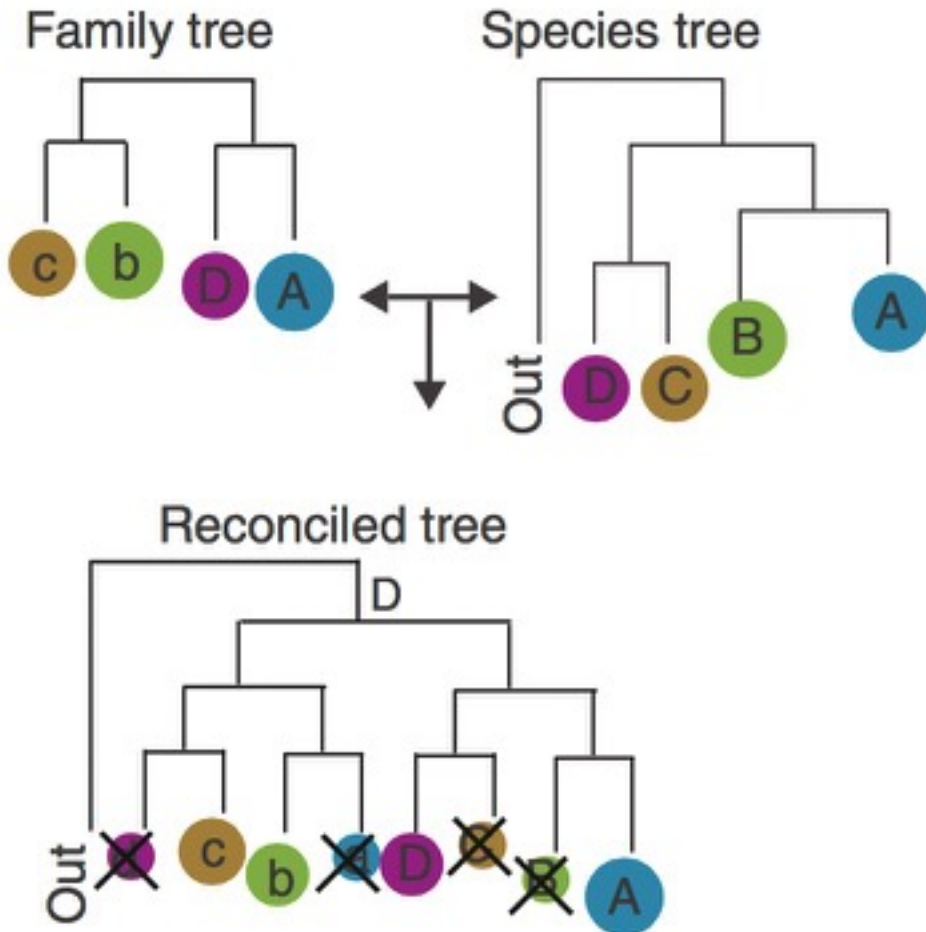


OrthoMCL

COG (Cluster of Orthologous Groups)

eggNOG (evolutionary genealogy of genes: Non-supervised Orthologous Groups)

Tree-based approaches



Trachana K. et al. (2011)

TreeFam
PhylomeDB
EnsemblCompara

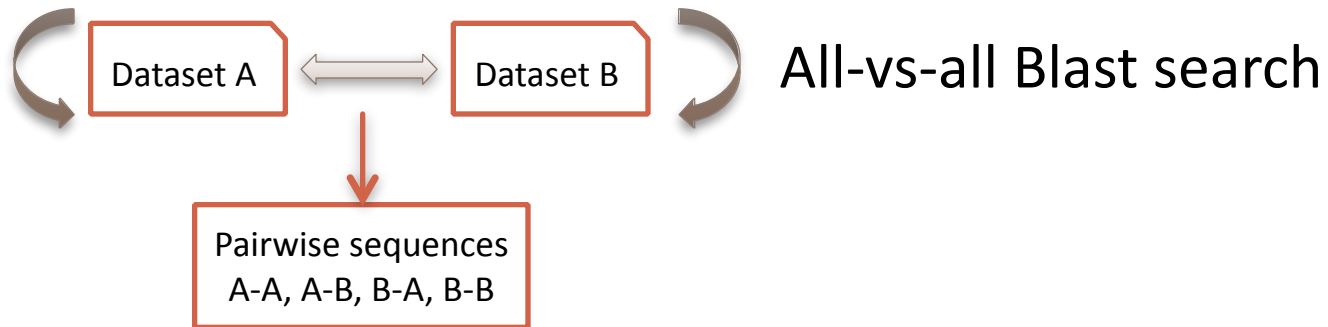
EXAMPLE

Orthology prediction using InParanoid

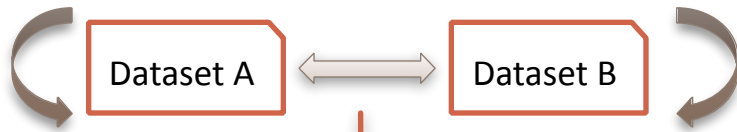
Dataset A

Dataset B

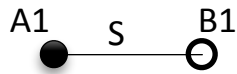
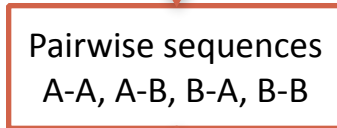
Orthology prediction using InParanoid



Orthology prediction using InParanoid

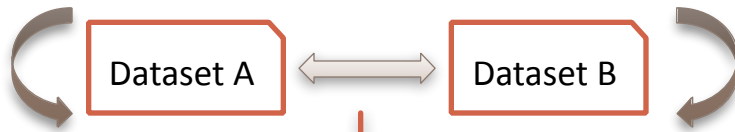


All-vs-all Blast search



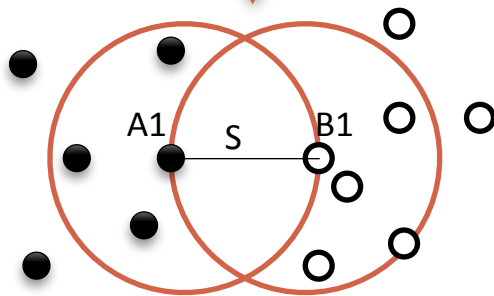
Find potential seed orthologs

Orthology prediction using InParanoid



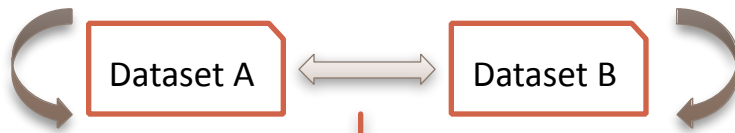
All-vs-all Blast search

Pairwise sequences
A-A, A-B, B-A, B-B

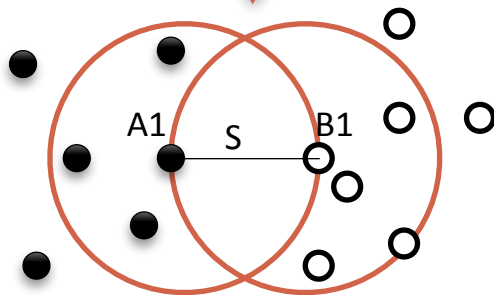
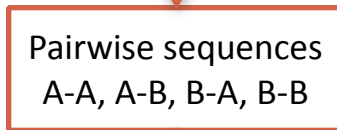


Find potential seed orthologs
& add in-paralogs

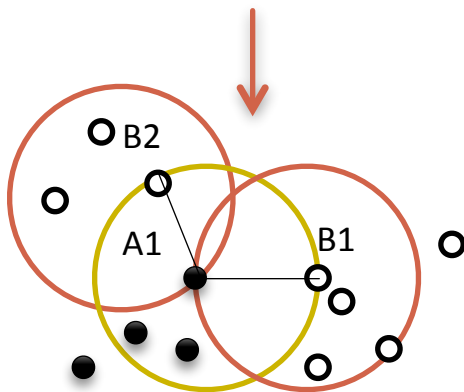
Orthology prediction using InParanoid



All-vs-all Blast search



Find potential seed orthologs
& add in-paralogs



Resolve overlapping groups



InParanoid: ortholog groups with inparalogs

273 organisms: 3718323 sequences

Version 8.0, Updated December 2013 ([release notes](#))

- BROWSE** the database - Select two species and view all their orthologs
- SEARCH BY SEQUENCE IDs** - View orthologs of a specific gene or protein
- TEXT SEARCH** - Query InParanoid by keywords
- BLAST SEARCH** - Find orthologs in InParanoid similar to your protein sequence
- DOWNLOAD DATA** - Obtain tables, html, orthoXML, sequences and core data
- SUMMARY OF INPARANOID** - Statistics of the database and genomes used
- ORTHOPHYLOGRAM** - Phylogenetic tree based on the average fraction of InParanoid orthologs between species.

Stand-alone InParanoid Program

InParanoid Version 4.1 is available [here](#)



Stockholm Bioinformatics Centre 2013, Supported by BILS



<http://inparanoid.sbc.su.se/cgi-bin/index.cgi>



InParanoid: ortholog groups with inparalogs

273 organisms: 3718323 sequences

Version 8.0, Updated December 2013 ([release notes](#))

- BROWSE** the database - Select two species and view all their orthologs
- SEARCH BY SEQUENCE IDs** - View orthologs of a specific gene or protein
- TEXT SEARCH** - Query InParanoid by keywords
- BLAST SEARCH** - Find orthologs in InParanoid similar to your protein sequence
- DOWNLOAD DATA** - Obtain tables, html, orthoXML, sequences and core data
- SUMMARY OF INPARANOID** - Statistics of the database and genomes used
- ORTHOPHYLOGRAM** - Phylogenetic tree based on the average fraction of InParanoid orthologs between species.

Stand-alone InParanoid Program

InParanoid Version 4.1 is available [here](#)



Stockholm Bioinformatics Centre 2013, Supported by BILS



Orthology search for particular gene

Orthology prediction using InParanoid

The screenshot displays the InParanoid8 web interface. At the top left, the UniProt logo is visible. Below it, a navigation bar includes links for BLAST, Align, and Retrieve/ID Mapping. The main header features the InParanoid8 logo and the tagline "Ortholog Groups with Inparalogs". A secondary navigation bar contains links for Home, Browse, Gene search, Text search, Blast, Downloads, Previous version, Summary, FAQ, and Help.

On the left side, a sidebar shows details for the query: Q5T4S7 - UBR4_HUMAN. It lists the protein as E3 ubiquitin-protein, the gene as UBR4, the organism as Homo sapiens (Human), and the status as Reviewed.

The main content area is titled "Gene search". It contains a text input field with the value "Q5T4S7". Below the input field is a dropdown menu for "Select ID type" with options "all", "geneid", and "proteinid". There are two radio buttons: "Search all species" (selected) and "Limit the search to the following species:". Below these options is a scrollable list of species names, including Acromyrmex echinator, Acyrthosiphon pisum, Aedes aegypti, Agaricus bisporus var. burnettii, Ailuropoda melanoleuca, Ajellomyces capsulata, Amphimedon queenslandica, Anolis carolinensis, Anopheles darlingi, Anopheles gambiae, Apis mellifera, Aquifex aeolicus, Arabidopsis thaliana, Arthrobotrys oligospora, Arthroderma gypseum, Ashbya gossypii, Aspergillus kawachii, Atta cephalotes, Aureococcus anophagefferens, and Auricularia delicata.

At the bottom of the search area, there is a field for "Exclude inparalogs scoring below:" with the value "0.05" and a "Senden" button.

Orthology search
for particular gene

Orthology prediction using InParanoid



InParanoid8
Ortholog Groups with Inparalogs

Home | Browse | Gene search | Text search | Blast | Downloads | Previous version | Summary | FAQ | Help

Searching all species for the proteinid **Q5T457** excluding inparalogs scoring below 0.05

[xml](#)

Inparalog and Orthologs cluster for **Homo sapiens** and **Cricetulus griseus**

Cluster 11					
Protein ID	Species	Score	Bootstrap	Description	Alternative ID
Q5T457	Homo sapiens	1	100%	E3 ubiquitin-protein ligase UBR4	UBR4_HUMAN (UniProt)
G31905	Cricetulus griseus	1	100%	E3 ubiquitin-protein ligase UBR4	G31905_CRIGR (UniProt)

Inparalog and Orthologs cluster for **Homo sapiens** and **Heterocephalus glaber**

Cluster 4					
Protein ID	Species	Score	Bootstrap	Description	Alternative ID
Q5T457	Homo sapiens	1	100%	E3 ubiquitin-protein ligase UBR4	UBR4_HUMAN (UniProt)
G5B2M0	Heterocephalus glaber	1	100%	E3 ubiquitin-protein ligase UBR4	G5B2M0_HETGA (UniProt)

Inparalog and Orthologs cluster for **Homo sapiens** and **Acromyrmex echinator**

Cluster 5					
Protein ID	Species	Score	Bootstrap	Description	Alternative ID
Q5T457	Homo sapiens	1	100%	E3 ubiquitin-protein ligase UBR4	UBR4_HUMAN (UniProt)
F4WKS3	Acromyrmex echinator	1	100%	Protein purity of essence	F4WKS3_ACREC (UniProt)

Inparalog and Orthologs cluster for **Homo sapiens** and **Camponotus floridanus**

Cluster 7					
Protein ID	Species	Score	Bootstrap	Description	Alternative ID
Q5T457	Homo sapiens	1	100%	E3 ubiquitin-protein ligase UBR4	UBR4_HUMAN (UniProt)
E2AZ34	Camponotus floridanus	1	100%	Protein purity of essence	E2AZ34_CAMFO (UniProt)

Orthology search for particular gene



InParanoid: ortholog groups with inparalogs

273 organisms: 3718323 sequences

Version 8.0, Updated December 2013 ([release notes](#))

BROWSE the database - Select two species and view all their orthologs
SEARCH BY SEQUENCE IDs - View orthologs of a specific gene or protein
TEXT SEARCH - Query InParanoid by keywords
BLAST SEARCH - Find orthologs in InParanoid similar to your protein sequence
DOWNLOAD DATA - Obtain tables, html, orthoXML, sequences and core data
SUMMARY OF INPARANOID - Statistics of the database and genomes used
ORTHOPHYLOGRAM - Phylogenetic tree based on the average fraction of InParanoid orthologs between species.

Stand-alone InParanoid Program

InParanoid Version 4.1 is available [here](#)



Stockholm Bioinformatics Centre 2013, Supported by BILS



Browse for all ortholog groups between 2 species

Orthology prediction using InParanoid



Browse the database

Choose two species



Stockholm Bioinformatics Centre 2013, Supported by BILS



Browse for all ortholog groups between 2 species

Orthology prediction using InParanoid



Inparalog and Orthologs clusters for **Homo sapiens** and **Mus musculus** (in total 16785)

< previous - next >

Home Browse

Cluster 51					
Protein ID	Species	Score	Bootstrap	Description	Alternative ID
Q8WXX0	Homo sapiens	1	100%	Dynein heavy chain 7, axonemal	DYH7_HUMAN (UniProt)
L7N1Y0	Mus musculus	1	100%	Protein Dnahc7b	L7N1Y0_MOUSE (UniProt)
E9Q0T8	Mus musculus	0.993		Protein Dnahc7a	E9Q0T8_MOUSE (UniProt)

Cluster 52					
Protein ID	Species	Score	Bootstrap	Description	Alternative ID
Q9NYQ8	Homo sapiens	1	100%	Protocadherin Fat 2	FAT2_HUMAN (UniProt)
Q5F226	Mus musculus	1	100%	Protocadherin Fat 2	FAT2_MOUSE (UniProt)

Cluster 53					
Protein ID	Species	Score	Bootstrap	Description	Alternative ID
Q86W11	Homo sapiens	1	100%	Fibrocystin-L	PKHL1_HUMAN (UniProt)
Q80ZA4	Mus musculus	1	100%	Fibrocystin-L	PKHL1_MOUSE (UniProt)

Cluster 54					
Protein ID	Species	Score	Bootstrap	Description	Alternative ID
P58107	Homo sapiens	1	100%	Epiplakin	EPIPL_HUMAN (UniProt)
Q8R0W0	Mus musculus	1	100%	Epiplakin	EPIPL_MOUSE (UniProt)

Cluster 55					
Protein ID	Species	Score	Bootstrap	Description	Alternative ID
Q03164	Homo sapiens	1	100%	Histone-lysine N-methyltransferase HLL	HLL1_HUMAN (UniProt)
P55200	Mus musculus	1	100%	Histone-lysine N-methyltransferase HLL	HLL1_MOUSE (UniProt)

Browse for all ortholog groups between 2 species

Orthology prediction using InParanoid



Inparalog and Orthologs clusters for **Homo sapiens** and **Mus musculus** (in total 16785)

< previous - next >

Home | Browse

Protein ID	Species	Score	Bootstrap	Description	Alternative ID
Q8WXX0	Homo sapiens	1	100%	Dynein heavy chain 7, axonemal	DYH7_HUMAN (UniProt)
L7N1Y0	Mus musculus	1	100%	Protein Dnahc7b	L7N1Y0_MOUSE (UniProt)
E9Q0T8	Mus musculus	0.993		Protein Dnahc7a	E9Q0T8_MOUSE (UniProt)

in-paralog to L7N1Y0

Protein ID	Species	Score	Bootstrap	Description	Alternative ID
Q9NRY2	Homo sapiens	1	100%	Fat2	FAT2_HUMAN (UniProt)
Q5F226	Mus musculus	1	100%	Fat2	FAT2_MOUSE (UniProt)

Protein ID	Species	Score	Bootstrap	Description	Alternative ID
Q86W11	Homo sapiens	1	100%	Fibrocystin-1	PKHL1_HUMAN (UniProt)
Q80ZA4	Mus musculus	1	100%	Fibrocystin-1	PKHL1_MOUSE (UniProt)

Protein ID	Species	Score	Bootstrap	Description	Alternative ID
P58107	Homo sapiens	1	100%	Epiplakin	EPIPL_HUMAN (UniProt)
Q8R0W0	Mus musculus	1	100%	Epiplakin	EPIPL_MOUSE (UniProt)

Protein ID	Species	Score	Bootstrap	Description	Alternative ID
Q03164	Homo sapiens	1	100%	Histone-lysine N-methyltransferase HLL	HLL1_HUMAN (UniProt)
P55200	Mus musculus	1	100%	Histone-lysine N-methyltransferase HLL	HLL1_MOUSE (UniProt)

Browse for all ortholog groups between 2 species

EXERCISE

Cladonia grayi

vs

Saccharomyces cerevisiae