UNIVERSITÉ DE GENÈVE
FACULTÉ DE MÉDECINE

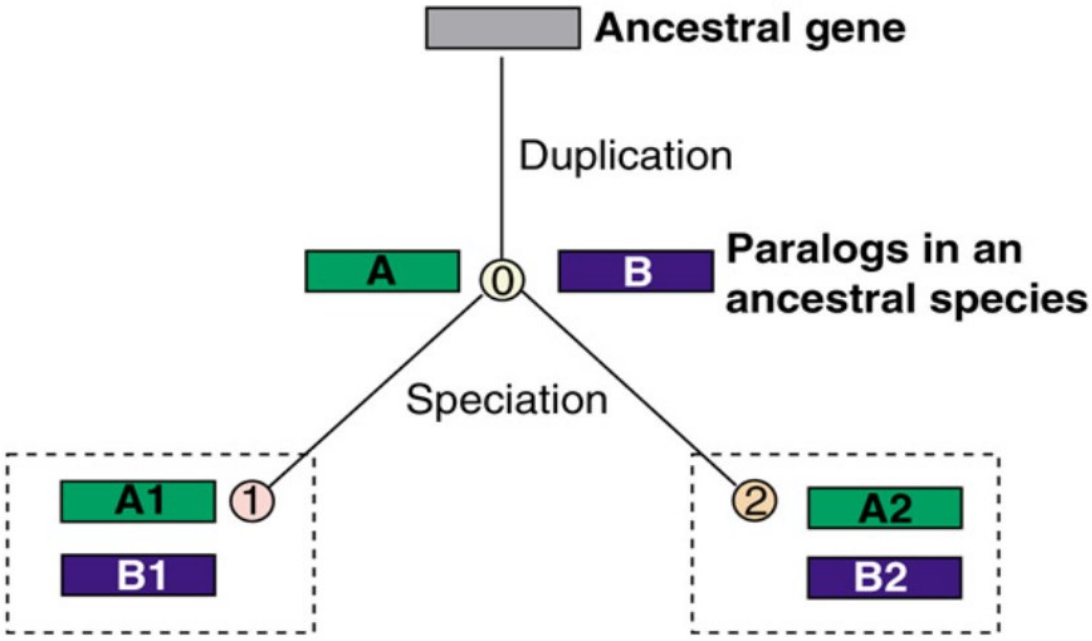
SIB Swiss Institute of Bioinformatics

OrthoDB

Querying SIB resources with SPARQL

Orthologues are

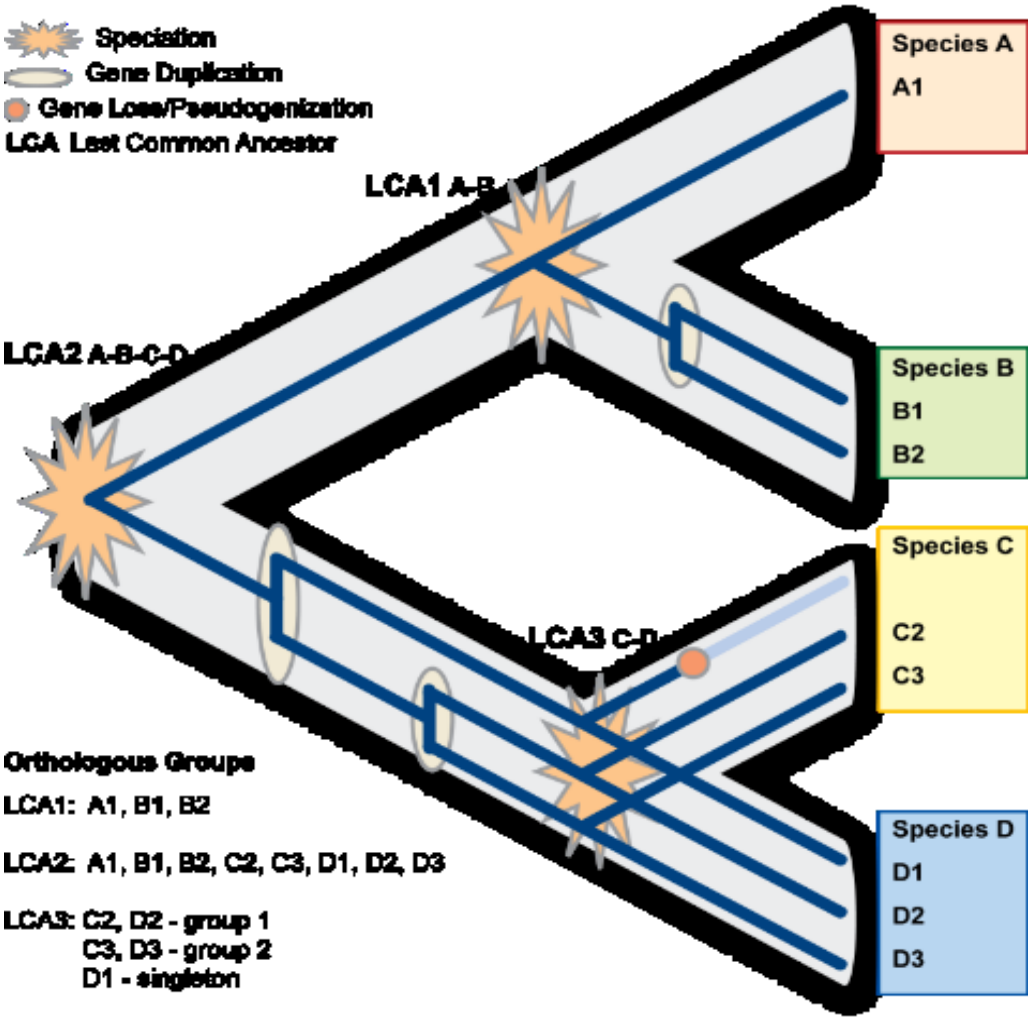
Two **homologous genes** in two different species that derive from a **single gene in the last common ancestor** of the species



A1 and A2 are orthologues B1 and B2 are orthologues

Eugene V Koonin "An apology for orthologs - or brave new memes", Genome Biol. 2001; 2(4)

Orthologues genes and Last Common Ancestors



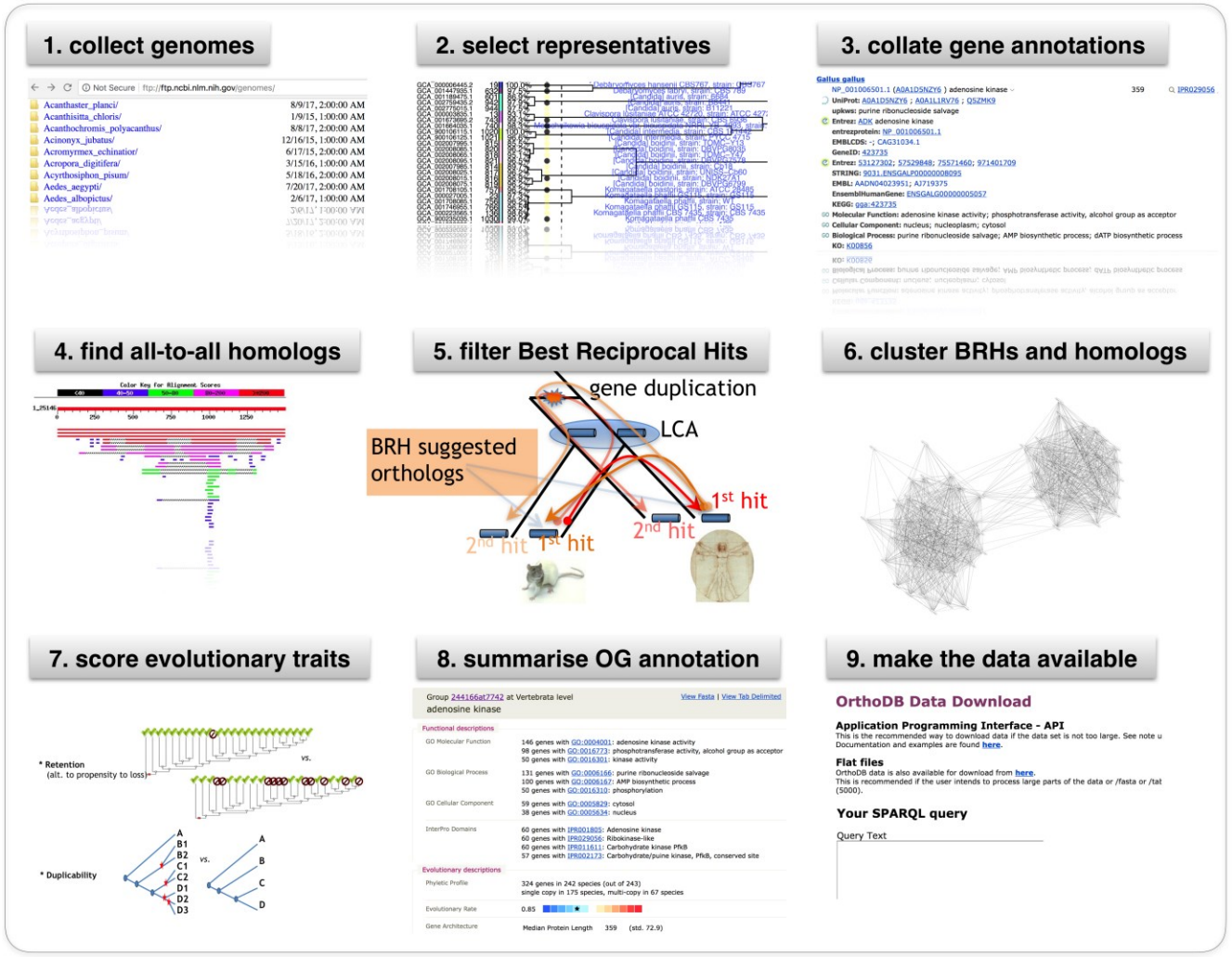
Orthology is a proxy for gene function

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

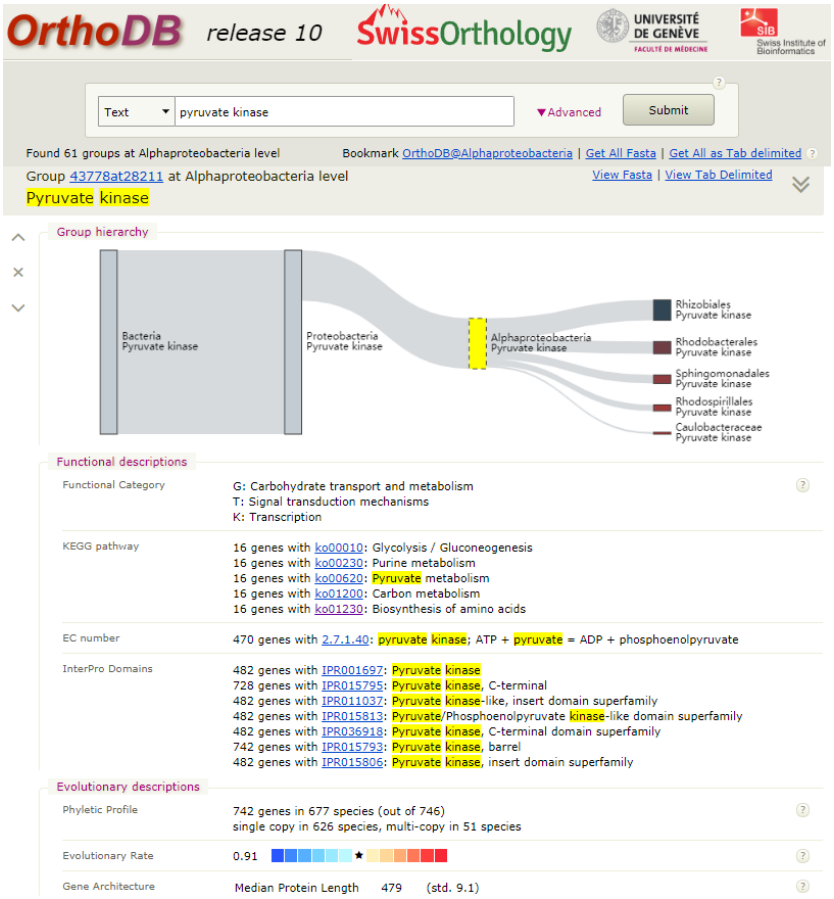
Orthologs, Paralogs, and Evolutionary Genomics[1]

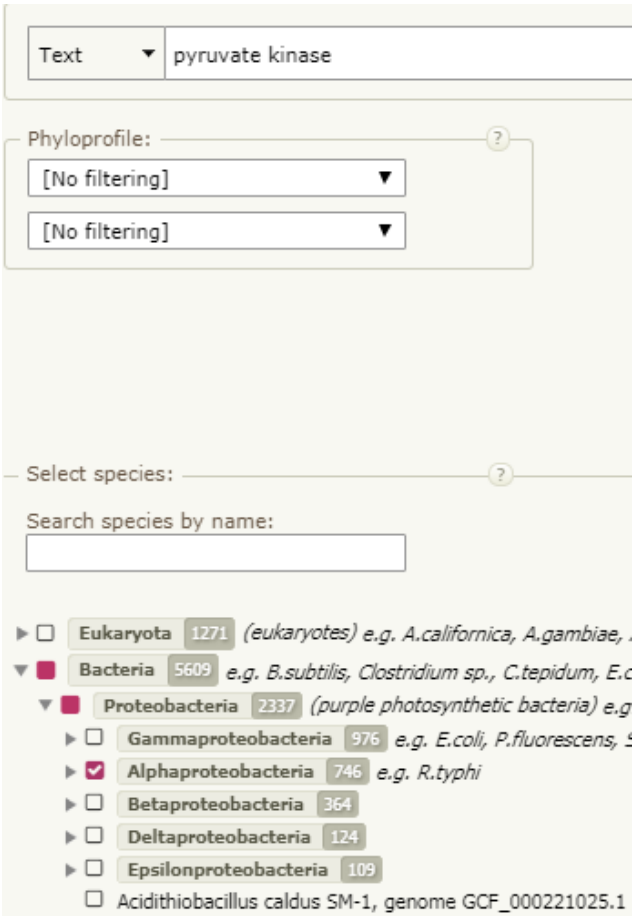
Eugene V. Koonin

OrthoDB analysis flow



OrthoDB – a catalog of hierarchical orthologous groups





http://orthodb.org

OrthoDB – a catalog of hierarchical orthologous groups

- 15K organisms
  - 1271 Eukaryota
  - 5609 Bacteria
  - 404 Archaea
  - 7963 Viruses
- 40M genes
- 1005 taxonomic levels
- 9M orthologous groups

OrthoDB – two APIs

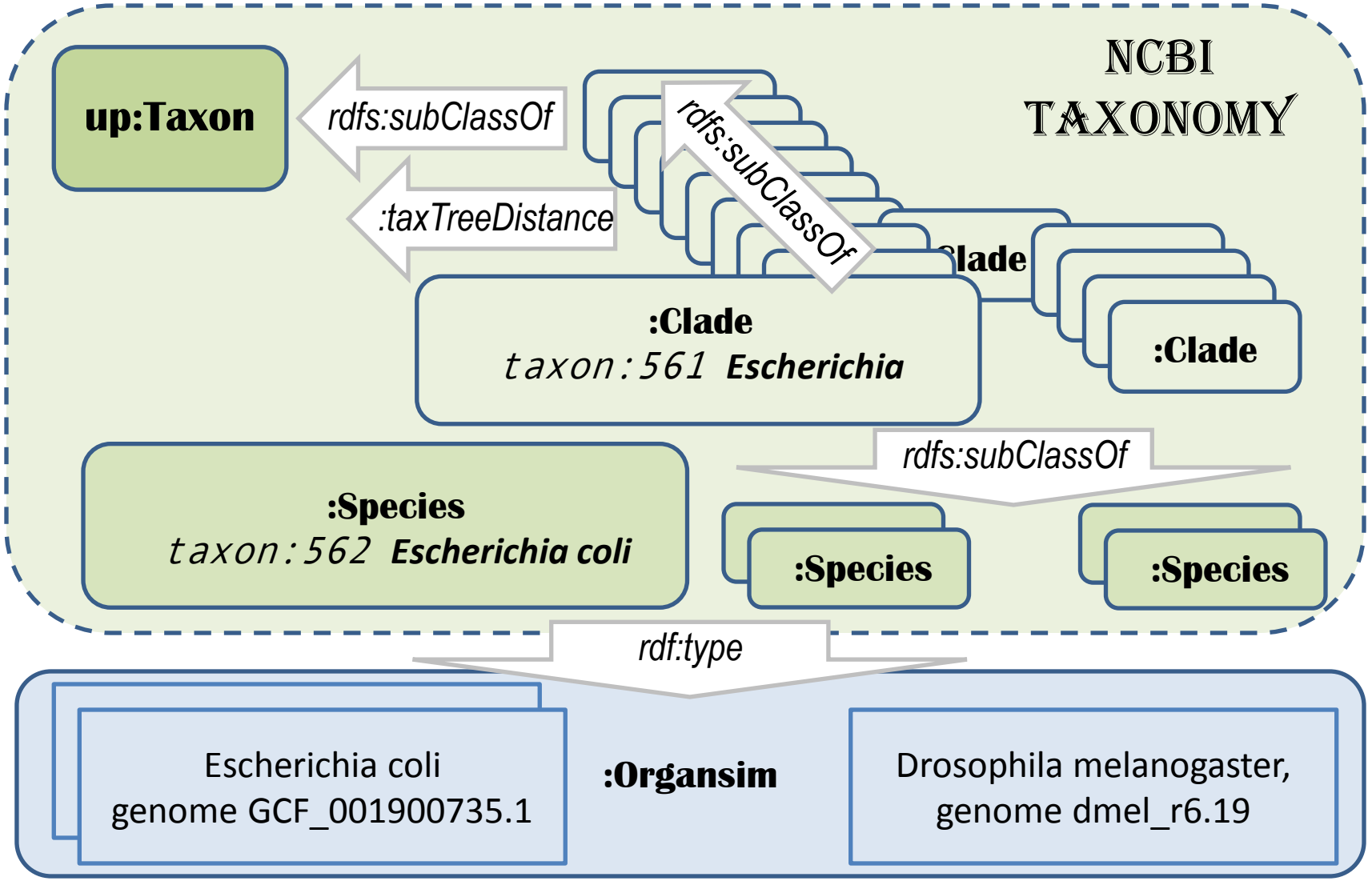
- Queries via programmable URLs
- SPARQL endpoint, http://sparql.orthodb.org

About

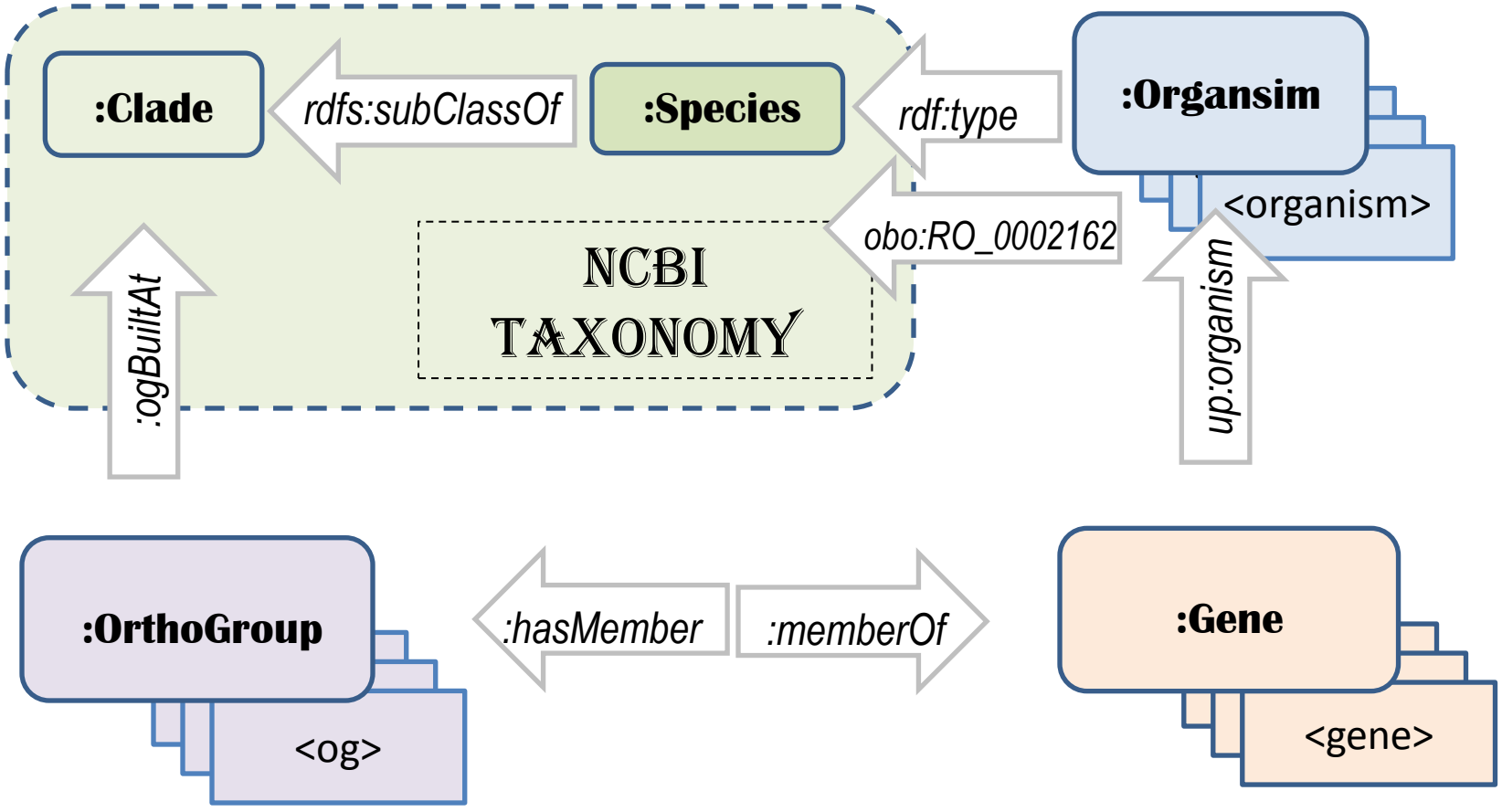
This SPARQL 1.1 endpoint serves OrthoDB data. The OrthoDB release 10.1 consists of 2'246'378'105 RDF triples describing evolutionary and functional properties of 40'614'194 genes from 15247 organisms clustered in 8'952'780 orthologous groups on 1004 taxonomic levels.

This endpoint cooperates with UniProt, NextProt and Ensembl endpoints due to adopting URIs of UniProt proteins (23'096'350), NextProt proteins (18'746) and Ensembl genes (684'451). The dataset also provides a number of clickable links to NCBI genes (13'587'519) and proteins (36'283'239), Ensembl Genomes (6'056'193), Interpro (35'285) and GO (21'975) resources.

OrthoDB taxonomic tree



Genes and orthologous groups



Clades, Species and Organisms

Clades

?clade a :Clade; rdfs:subClassOf+ taxon:2; up:scientificName ?name.

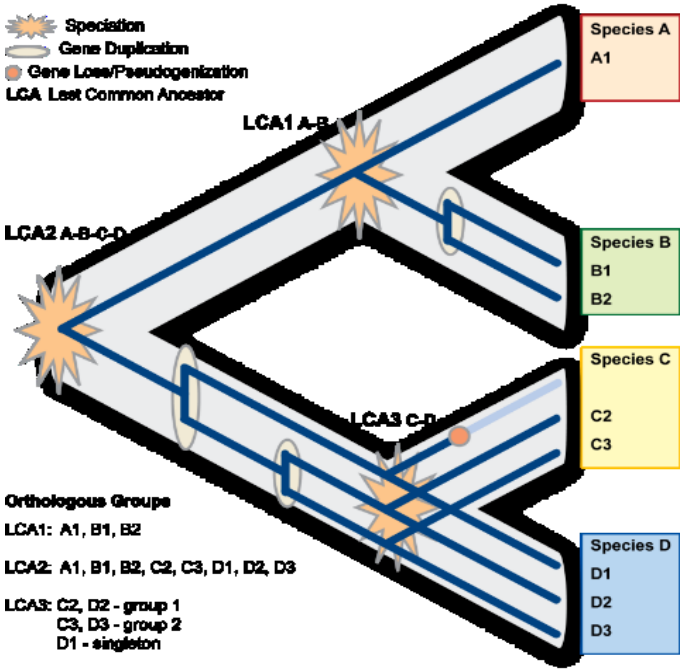
Species, Organsims

?clade a :Clade; up:scientificName "Escherichia".
?taxon a :Species; up:scientificName ?**tx_name**; rdfs:subClassOf+ ?clade.
?org a :Organism,?taxon; up:scientificName ?**org_name**.

tx_name	org_name
"Escherichia albertii"	"Escherichia albertii, genome GCF_001549955.1"
"Escherichia coli Nissle 1917"	"Escherichia coli Nissle 1917, genome GCF_000714595.1"
"Escherichia coli"	"Escherichia coli, genome GCF_001617565.1"
"Escherichia coli"	"Escherichia coli, genome GCF_001900655.1"
"Escherichia coli"	"Escherichia coli, genome GCF_001900735.1"
"Escherichia coli"	"Escherichia coli, genome GCF_001900945.1"
"Escherichia fergusonii ATCC 35469"	"Escherichia fergusonii ATCC 35469, genome GCF_000026225.1"
"Escherichia coli O157:H16"	"Escherichia coli O157:H16, genome GCF_000827105.1"

Result IRIs for clade, taxon, org are functional URLs pointing to OrthoDB pages

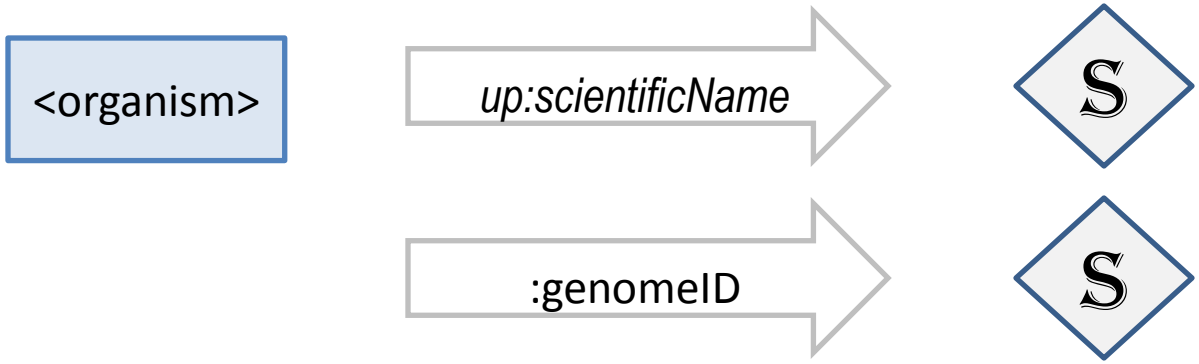
Find Last Common Ancestor (LCA) for fruit fly and honey bee




```
select *
where {
        ?lca a :Clade ; up:scientificName ?lcaname .
        taxon:7227  rdfs:subClassOf* ?lca .
        taxon:7460 rdfs:subClassOf* ?lca .
        filter (not exists {
                ?xca a :Clade ; rdfs:subClassOf ?lca .
                taxon:7227    rdfs:subClassOf* ?xca .
                taxon:7460    rdfs:subClassOf* ?xca .
        })
}
```

lca	lcaname
http://purl.uniprot.org/taxonomy/33392	"Holometabola"

Organism predicates



Genes

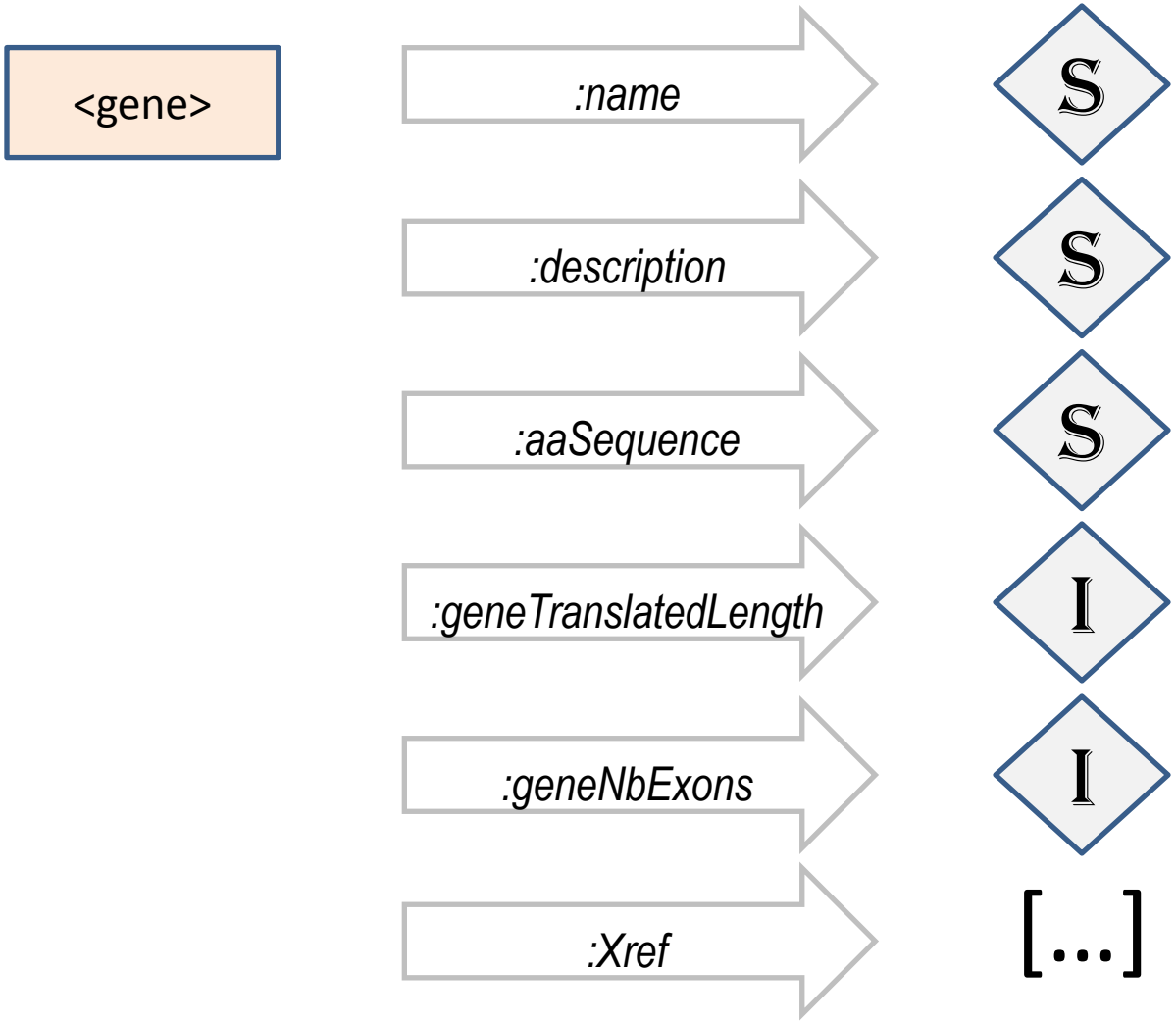
Genes

?org a :Organism; up:scientificName "Escherichia coli, genome GCF_001617565.1".
?gene a :Gene; **up:organism** ?org; :name ?**gene_name**; :description ?**description**.

gene_name	description
"fucA"	"L-fuculose phosphate aldolase"
"alaS"	"Alanine--tRNA ligase"
"hemG"	"Molybdopterin-guanine dinucleotide biosynthesis protein B"
"SY51_RS02910"	"Lipoprotein"
"rbsA"	"Ribose import ATP-binding protein RbsA"
"WM90_RS12030"	"HCP oxidoreductase"
"gltB"	"Glutamate synthase"
"smpA"	"Outer membrane protein assembly factor BamE"
"WM90_RS24495"	"Addiction module toxin RelE"

Result IRIs for org, gene are functional URLs pointing to OrthoDB pages

Gene predicates



Orthologous groups (OGs)

Orthogroups

```
select ?og ?og_name ?distance ?clade_name  (count(1) as ?cnt)
where {
?org a :Organism; up:scientificName "Escherichia coli, genome GCF_001617565.1".
?gene a :Gene; up:organism ?org; :name "alaS"; :memberOf ?og.
?og :ogBuiltAt [up:scientificName ?clade_name; :taxTreeDistance ?distance];
:name ?og_name; :hasMember ?gene2
} group by ?clade_name ?og ?og_name ?distance order by ?distance
```

og	og_name	distance	clade_name	cnt
http://purl.orthodb.org/odbgroup/91428at2	"Alanine--tRNA ligase"	0	"Bacteria"	5604
http://purl.orthodb.org/odbgroup/13856at1224	"Alanine--tRNA ligase"	1	"Proteobacteria"	2294
http://purl.orthodb.org/odbgroup/26372at1236	"Alanine--tRNA ligase"	2	"Gammaproteobacteria"	956
http://purl.orthodb.org/odbgroup/7453at91347	"Alanine--tRNA ligase"	3	"Enterobacterales"	211
http://purl.orthodb.org/odbgroup/1091at543	"Alanine--tRNA ligase"	4	"Enterobacteriaceae"	98
http://purl.orthodb.org/odbgroup/132at561	"Alanine--tRNA ligase"	5	"Escherichia"	7

Result IRIs for org, gene are functional URLs pointing to OrthoDB pages

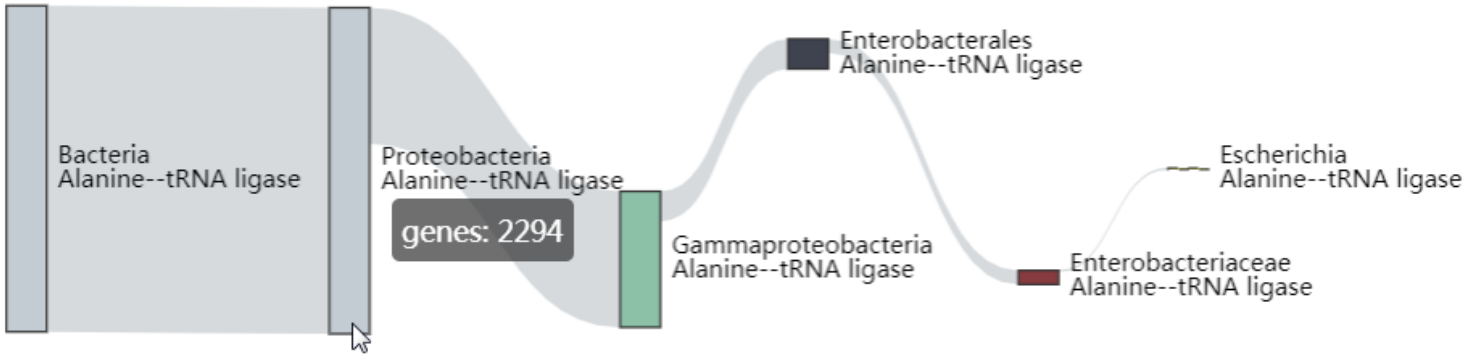
Hierarchy of OGs is a paraphrase of the gene in evolutionary retrospective

Group 132at561 at Escherichia level

Alanine--tRNA ligase

View Fasta | View Tab Delimited

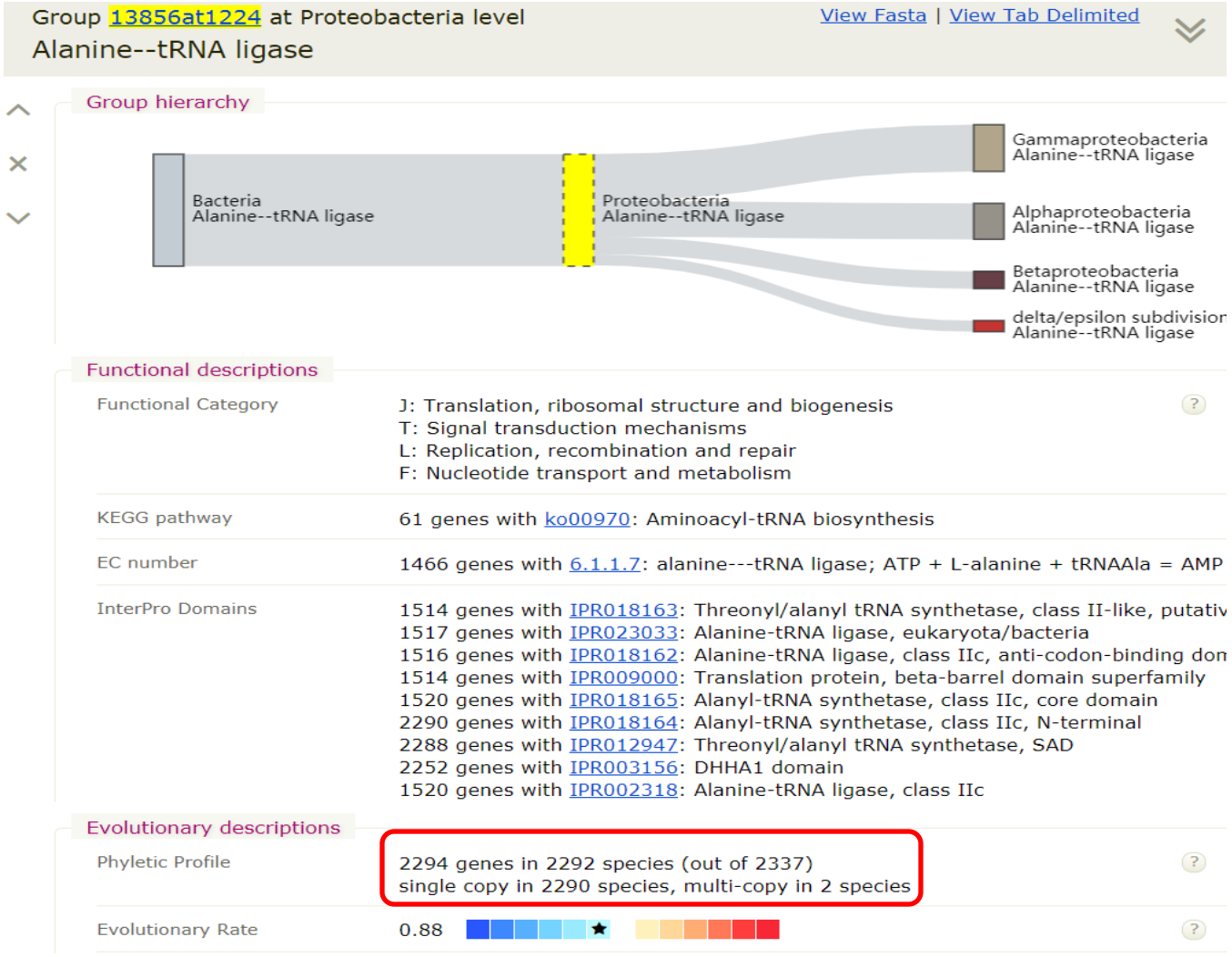
Group hierarchy



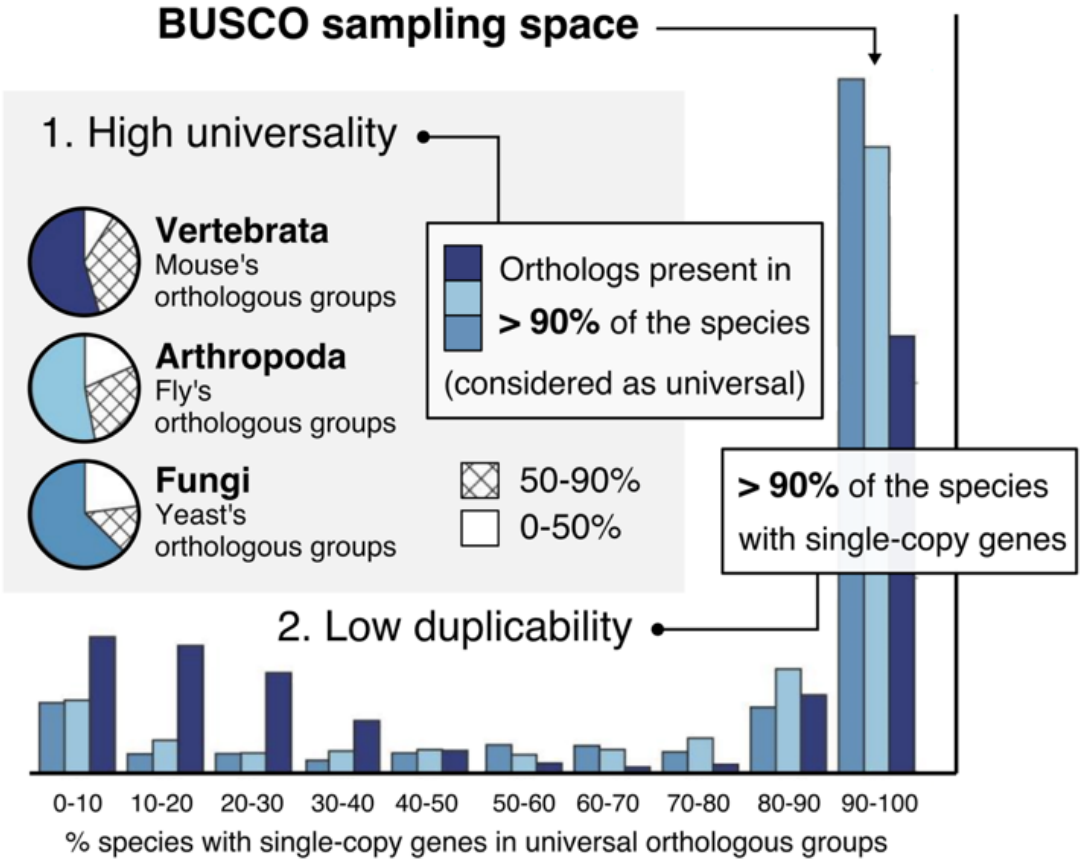
Functional descriptions

Functional Category	J: Translation, ribosomal structure and biogenesis L: Replication, recombination and repair F: Nucleotide transport and metabolism T: Signal transduction mechanisms
EC number	5 genes with 6.1.1.7: alanine---tRNA ligase; ATP + L-alanine + tRNAAla = AMP + diphospł
InterPro Domains	6 genes with IPR002318: Alanine-tRNA ligase, class IIc 7 genes with IPR003156: DHHA1 domain 7 genes with IPR012947: Threonyl/alanyl tRNA synthetase, SAD

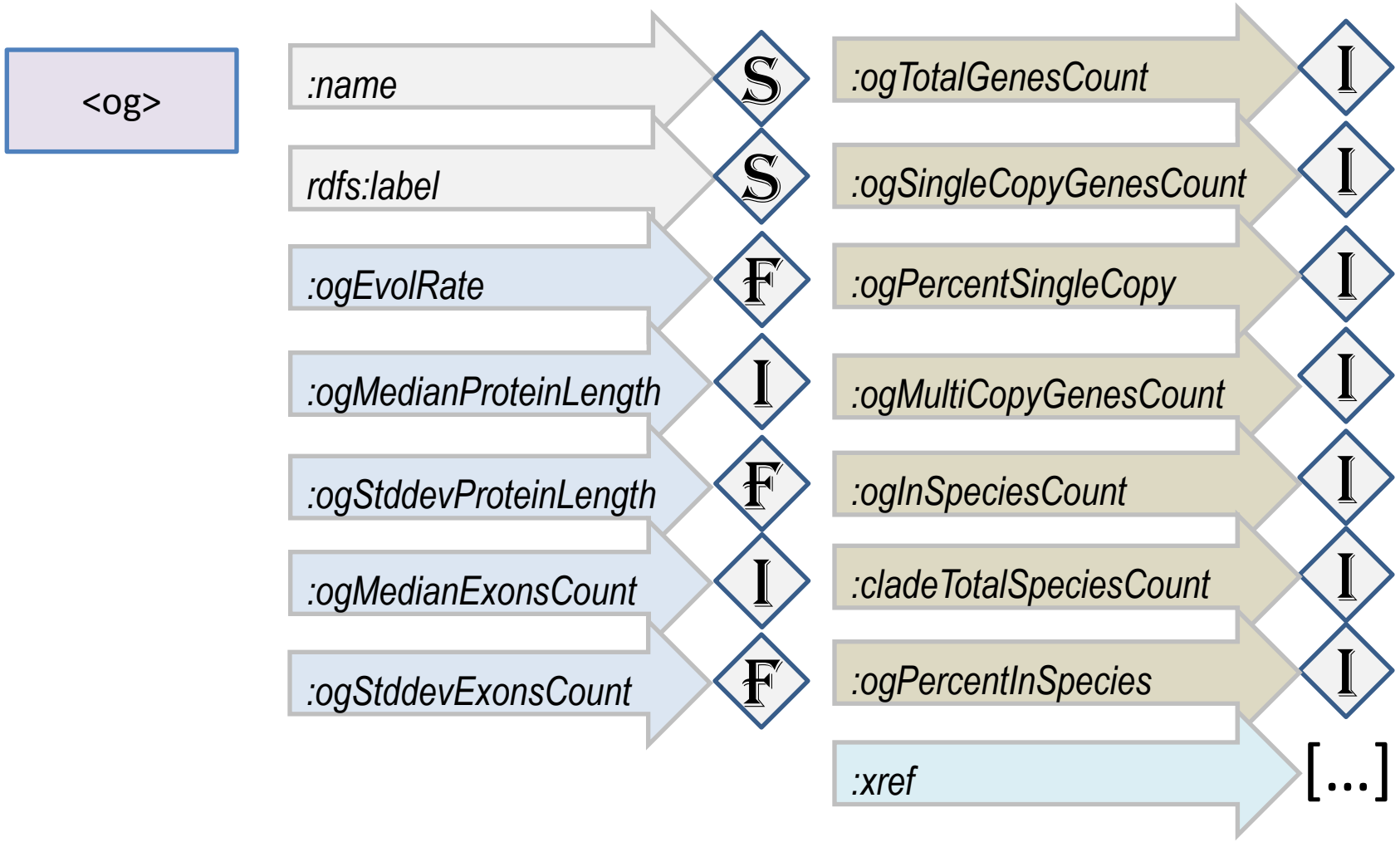
Evolutionary stable genes persist through hierarchy of clades



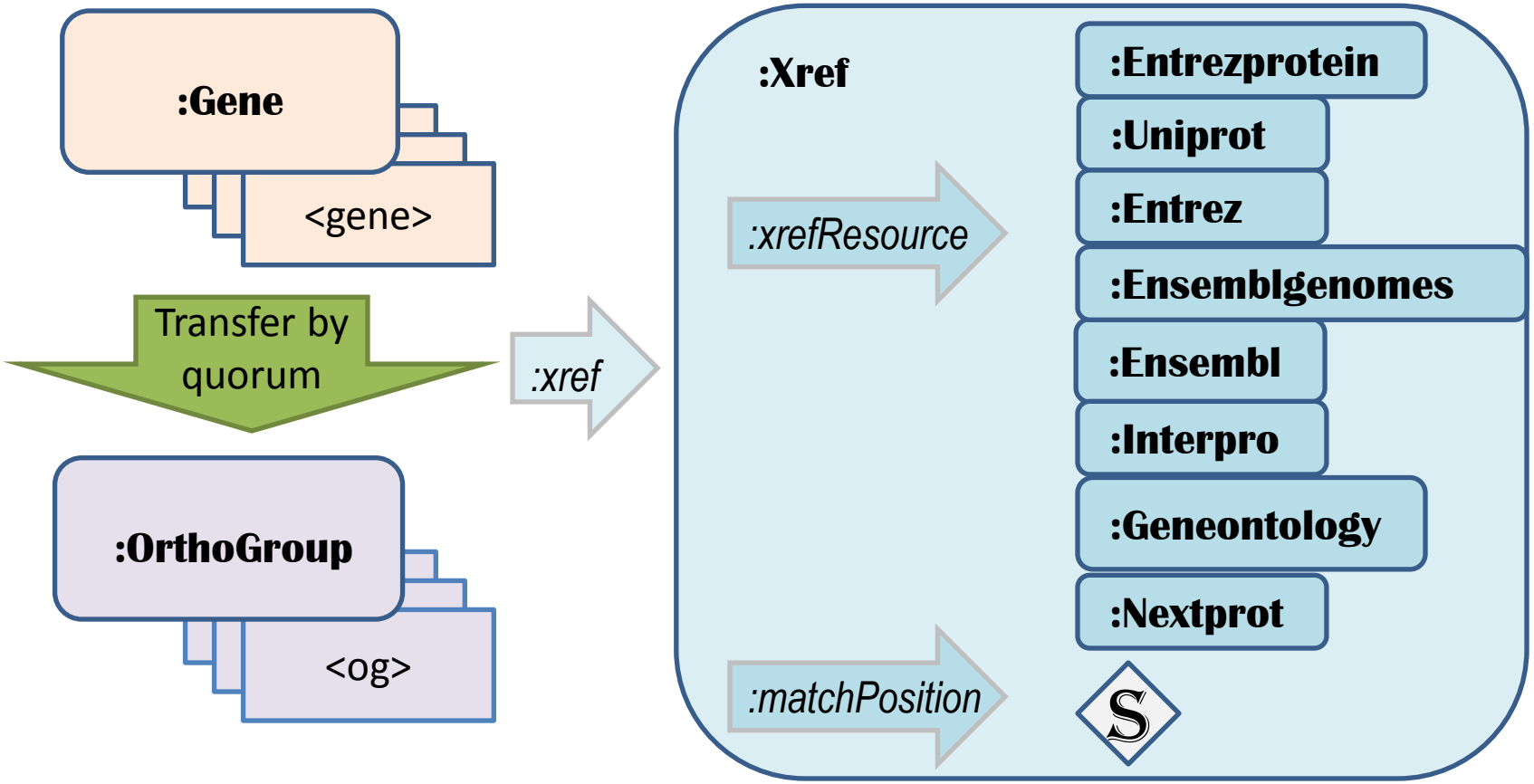
BUSCO- **B**enchmarking **S**ets of **U**niversal **S**ingle-**C**opy **O**rthologs



Orthologous group predicates



External references, aka xrefs



Human and zebra-fish immune system orthologous genes: use of GO-terms

```
select ?gene_h_name ?gene_zf_name ?og_description ?go_id ?go_name
where {
?gene_h a :Gene; :name ?gene_h_name; :memberOf ?og. # both genes in same OG
?gene_h up:organism/a [up:scientificName "Homo sapiens"]
?gene_zf a :Gene; :name ?gene_zf_name; :memberOf ?og. # both genes in same OG
?gene_zf up:organism/a [up:scientificName "Danio rerio"].
.
?og a :OrthoGroup; :ogBuiltAt [up:scientificName "Vertebrata"];
:name ?og_description; :ogEvolRate ?evolrate; :xref [a :Xref; :xrefResource ?xref]
.
?xref a :Geneontology; rdfs:label ?go_id;  :name ?go_name.
filter (contains(?go_name,'immune response'))
} order by desc(?evolrate) limit 6
```

Genes from fastest-evolving OGs

gene_h_name	gene_zf_name	og_description
"CD4"	"cd4-1"	"Immunoglobulin"
"CD4"	"cd4-1"	"Immunoglobulin"
"IL12RB1"	"si:dkey-13m1.2"	"Fibronectin type III"
"CD226"	"cd226"	"Immunoglobulin-like domain"
"IRF7"	"irf7"	"Interferon regulatory factor-3"
"IRF7"	"irf7"	"Interferon regulatory factor-3"

go_id	go_name
"GO:0035397"	"helper T cell enhancement of adaptive immune response"
"GO:0006955"	"immune response"
"GO:0002827"	"positive regulation of T-helper 1 type immune response"
"GO:0002891"	"positive regulation of immunoglobulin mediated immune response"
"GO:0002819"	"regulation of adaptive immune response"
"GO:0016064"	"immunoglobulin mediated immune response"

Genes from slowest-evolving OGs

gene_h_name	gene_zf_name	og_description
"RBPJ"	"rbpjb"	"RBP-J/Cbf11/Cbf12, DNA binding"
"RBPJ"	"rbpja"	"RBP-J/Cbf11/Cbf12, DNA binding"
"RBPJ"	"rbpja"	"RBP-J/Cbf11/Cbf12, DNA binding"
"RBPJ"	"rbpjb"	"RBP-J/Cbf11/Cbf12, DNA binding"
"DKFZp686D10173;POLR3B"	"polr3b"	"RNA polymerase Rpb2, domain 2"
"DKFZp686D10173;POLR3B"	"polr3b"	"RNA polymerase Rpb2, domain 2"
"ILF2"	"ilf2"	"interleukin enhancer-binding factor 2"
"RPS6"	"rps6"	"Ribosomal protein S6e"

go_id	go_name
"GO:0006959"	"humoral immune response"
"GO:0006959"	"humoral immune response"
"GO:0006959"	"humoral immune response"
"GO:0006959"	"humoral immune response"
"GO:0045089"	"positive regulation of innate immune response"
"GO:0045089"	"positive regulation of innate immune response"
"GO:0006955"	"immune response"
"GO:0002309"	"T cell proliferation involved in immune response"

Federation OrthoDB - NextProt

```
prefix : <http://purl.orthodb.org/>
PREFIX np: <http://nextprot.org/rdf#>

select distinct ?gene_h_name ?gene_zf_name ?disease
where {
        service <https://sparql.nextprot.org/> {
        select ?entry ?disease WHERE {
        ?entry np:isoform / np:disease / rdfs:comment ?disease
} limit 99
}
?gene rdfs:seeAlso ?entry; :name ?gene_h_name;  :memberOf /:hasMember ?gene2.
?gene2 :name ?gene_zf_name; up:organism/a [up:scientificName "Danio rerio"].
} limit 20
```

ZF orthologues of human genes implicated in a disease

gene_h_name	gene_zf_name	
"ERG"	"fev"	"Ewing sarcoma (ES) [MIM:612219]: A highly malignant, metastatic,
"MYLK"	"mylk5"	"Megacystis-microcolon-intestinal hypoperistalsis syndrome"^^<htt
"MYLK"	"mylk5"	"Aortic aneurysm, familial thoracic 7 (AAT7) [MIM:613780]: A dise
"MYLK"	"mylk5"	"Familial thoracic aortic aneurysm and aortic dissection"^^<http:
"MYLK"	"mylkb"	"Megacystis-microcolon-intestinal hypoperistalsis syndrome"^^<htt
"MYLK"	"mylkb"	"Aortic aneurysm, familial thoracic 7 (AAT7) [MIM:613780]: A dise
"MYLK"	"mylkb"	"Familial thoracic aortic aneurysm and aortic dissection"^^<http:
"MYLK"	"mylka"	"Megacystis-microcolon-intestinal hypoperistalsis syndrome"^^<htt
"MYLK"	"mylka"	"Aortic aneurysm, familial thoracic 7 (AAT7) [MIM:613780]: A dise
"MYLK"	"mylka"	"Familial thoracic aortic aneurysm and aortic dissection"^^<http:
"ERG"	"erg"	"Ewing sarcoma (ES) [MIM:612219]: A highly malignant, metastatic,
"ERG"	"fli1a"	"Ewing sarcoma (ES) [MIM:612219]: A highly malignant, metastatic,
"ERG"	"fli1b"	"Ewing sarcoma (ES) [MIM:612219]: A highly malignant, metastatic,
"AIRE"	"phf12a"	"Familial isolated hypoparathyroidism due to impaired PTH secreti
"AIRE"	"phf12a"	"Autoimmune polyendocrinopathy type 1"^^<http://www.w3.org/2001/X
"AIRE"	"phf12a"	"Autoimmune polyendocrine syndrome 1, with or without reversible
"PATL2"	"si:ch211-103b1.2"	"Female infertility due to oocyte meiotic arrest"^^<http://www.w3
"PATL2"	"si:ch211-103b1.2"	"Oocyte maturation defect 4 (OOMD4) [MIM:617743]: An infertility

Federation OrthoDB - OMA

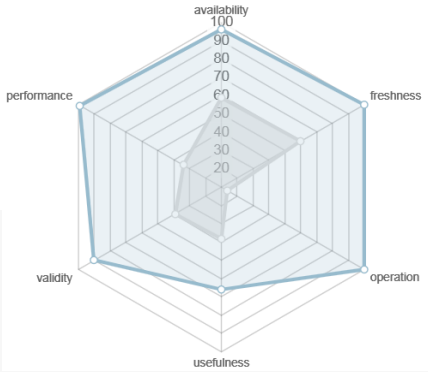
```
select * where {
?gene_odb a :Gene; :name ?gene_name_odb; rdfs:seeAlso ?up;
:memberOf/:name ?og_name.
filter (?up = uniprot:P12345)

service <https://sparql.omabrowser.org/> {
?gene_oma a orth:Protein; rdfs:label ?gene_name_oma; lscr:xrefUniprot ?up.
?node orth:hasHomologousMember* ?gene_oma .
}} limit 3
```

gene_odb	gene_name_odb	up	og_name
http://purl.orthodb.org/odbgene/9986_0_0010c7	"GOT2"	http://purl.uniprot.org/uniprot/P12345	"Pyridoxal phosphate-dependent transferase"
http://purl.orthodb.org/odbgene/9986_0_0010c7	"GOT2"	http://purl.uniprot.org/uniprot/P12345	"Pyridoxal phosphate-dependent transferase"
http://purl.orthodb.org/odbgene/9986_0_0010c7	"GOT2"	http://purl.uniprot.org/uniprot/P12345	"Pyridoxal phosphate-dependent transferase"

gene_oma	gene_name_oma	node
https://omabrowser.org/oma/info/RABIT10926	"GOT2"	http://omabrowser.org/ontology/oma#PARALOG_GROUP_685335_462
https://omabrowser.org/oma/info/RABIT10926	"GOT2"	http://omabrowser.org/ontology/oma#GROUP_685335_Theria_483
https://omabrowser.org/oma/info/RABIT10926	"GOT2"	http://omabrowser.org/ontology/oma#GROUP_685335_Metazoa_408

OrthoDB rank @ YummyData



Score Ranking

Name	URL	
Life Science Dictionary	http://lsd.dbcls.jp/sparql	95
Colil	http://colil.dbcls.jp/sparql	93
OrthoDB @ sib.swiss	http://sparql.orthodb.org/sparql	91
Rhea	https://sparql.rhea-db.org/sparql	89
Allie	http://data.allie.dbcls.jp/sparql	89
Bio2RDF	http://bio2rdf.org/sparql	86
Agronomic Linked Data (AgroLD)	http://sparql.southgreen.fr/	83
WikiPathways	http://sparql.wikipathways.org	82
Linked Open Data platform for EBI data	http://www.ebi.ac.uk/rdf/services/sparql	81

https://yummydata.org/endpoints

Computational evolutionary genomics group (Evgeny Zdobnov)

UNIVERSITÉ DE GENÈVE
FACULTÉ DE MÉDECINE

SIB Swiss Institute of Bioinformatics