OrthoDB

An evolutionary perspective to interpreting genomics data

The i5k Webinar Series
February 1st 2017

Prof. Evgeny M. Zdobnov & Dr Robert M. Waterhouse
University of Geneva &
Swiss Institute of Bioinformatics
Geneva, Switzerland

UNIVERSITÉ DE GENÈVE
FACULTÉ DE MÉDECINE
Department of Genetic Medicine and Development

SIB
Swiss Institute of Bioinformatics

<u>What</u> is orthology?

Understanding the definitions

<u>How</u> does one delineate orthology?

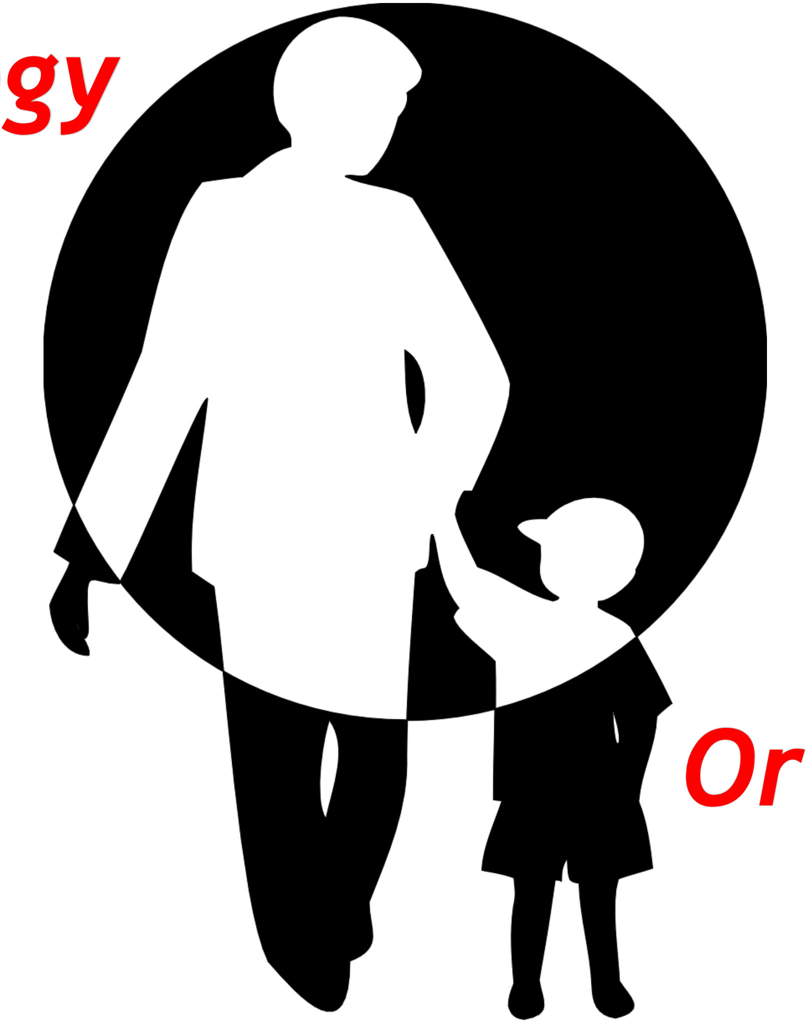
Getting to grips with the methodologies

<u>What</u> does OrthoDB offer?

Using orthology in your research

What is orthology?

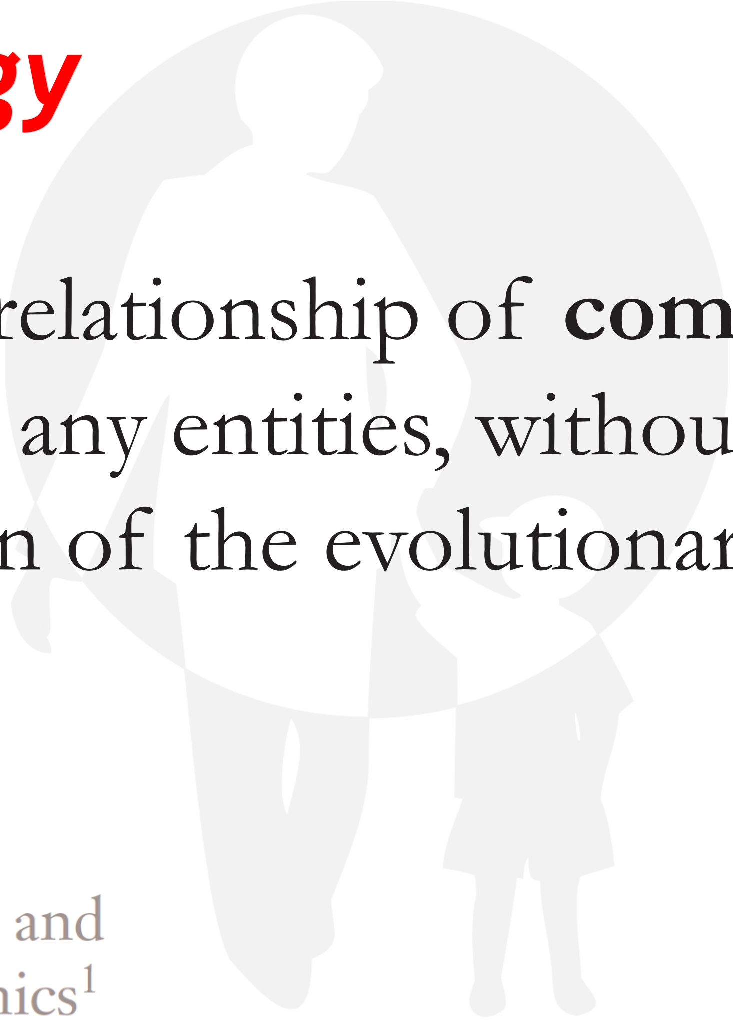
Homology



Orthology

What is ~~orthology~~ homology?

Homology



“designates a relationship of **common descent** between any entities, without further specification of the evolutionary scenario”

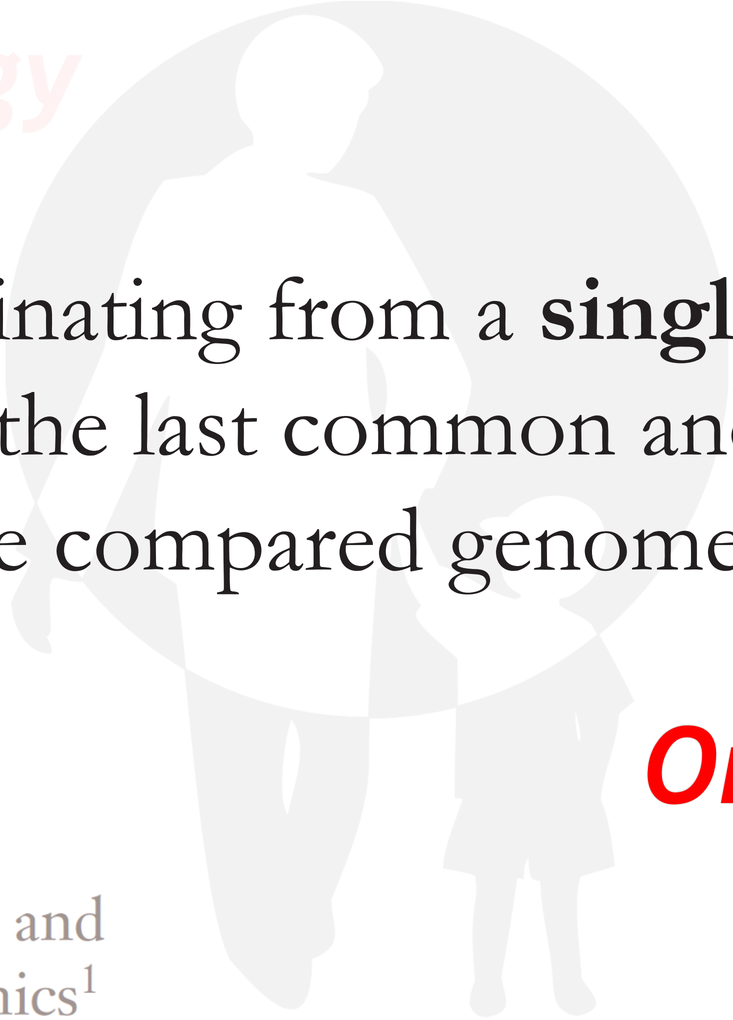
Orthologs, Paralogs, and Evolutionary Genomics[1]

Eugene V. Koonin

Annu. Rev. Genet. 2005. 39:309–38

What is orthology?

Homology



“genes originating from a **single ancestral gene** in the last common ancestor of the compared genomes”

Orthology

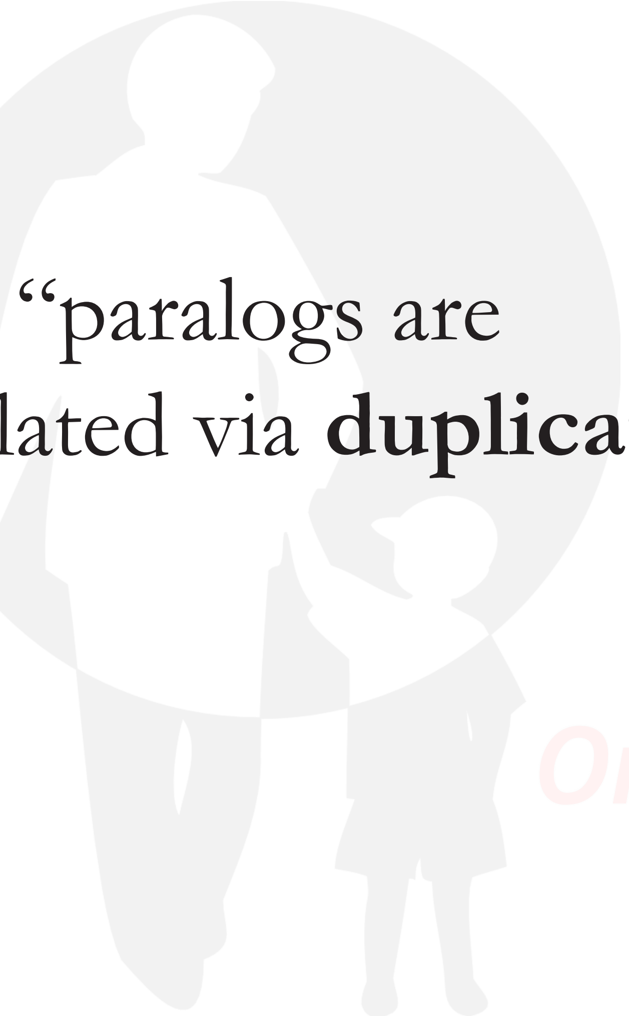
Orthologs, Paralogs, and Evolutionary Genomics[1]

Eugene V. Koonin

Annu. Rev. Genet.
2005. 39:309–38

What is paralogy?

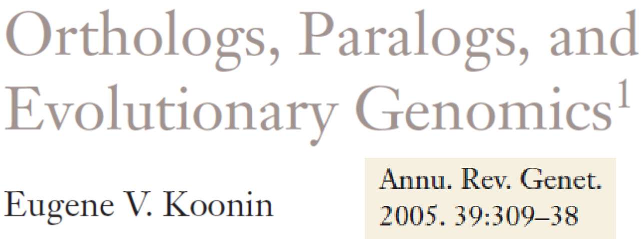
Homology



“paralogs are
genes related via **duplication**”

Paralogy

Orthology



Orthologs, Paralogs, and Evolutionary Genomics[1]

Eugene V. Koonin

Annu. Rev. Genet.
2005. 39:309–38

What are homologs, orthologs, paralogs?

Homologs

Common Ancestor

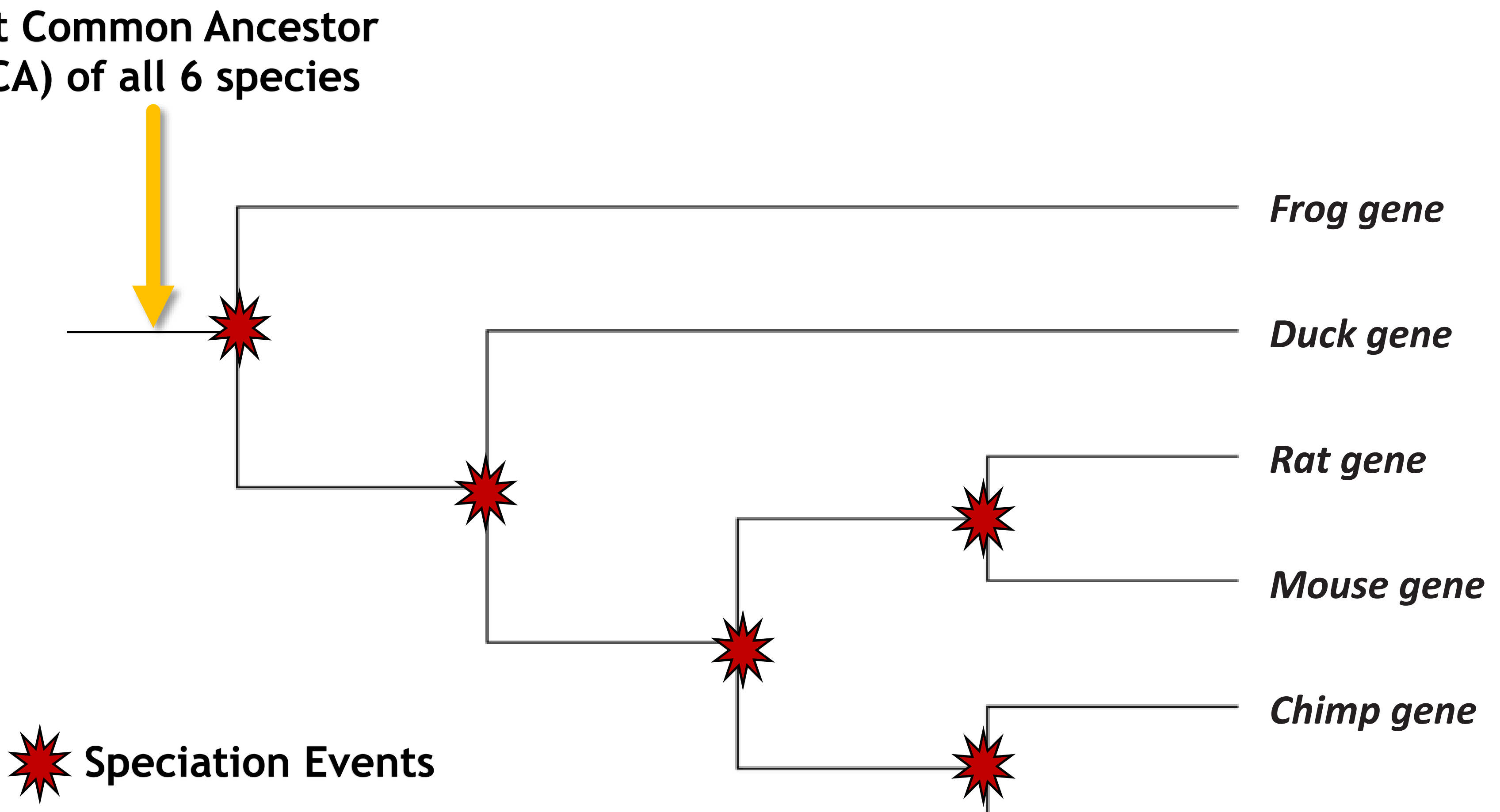
Paralogs

Duplication
Event

Orthologs

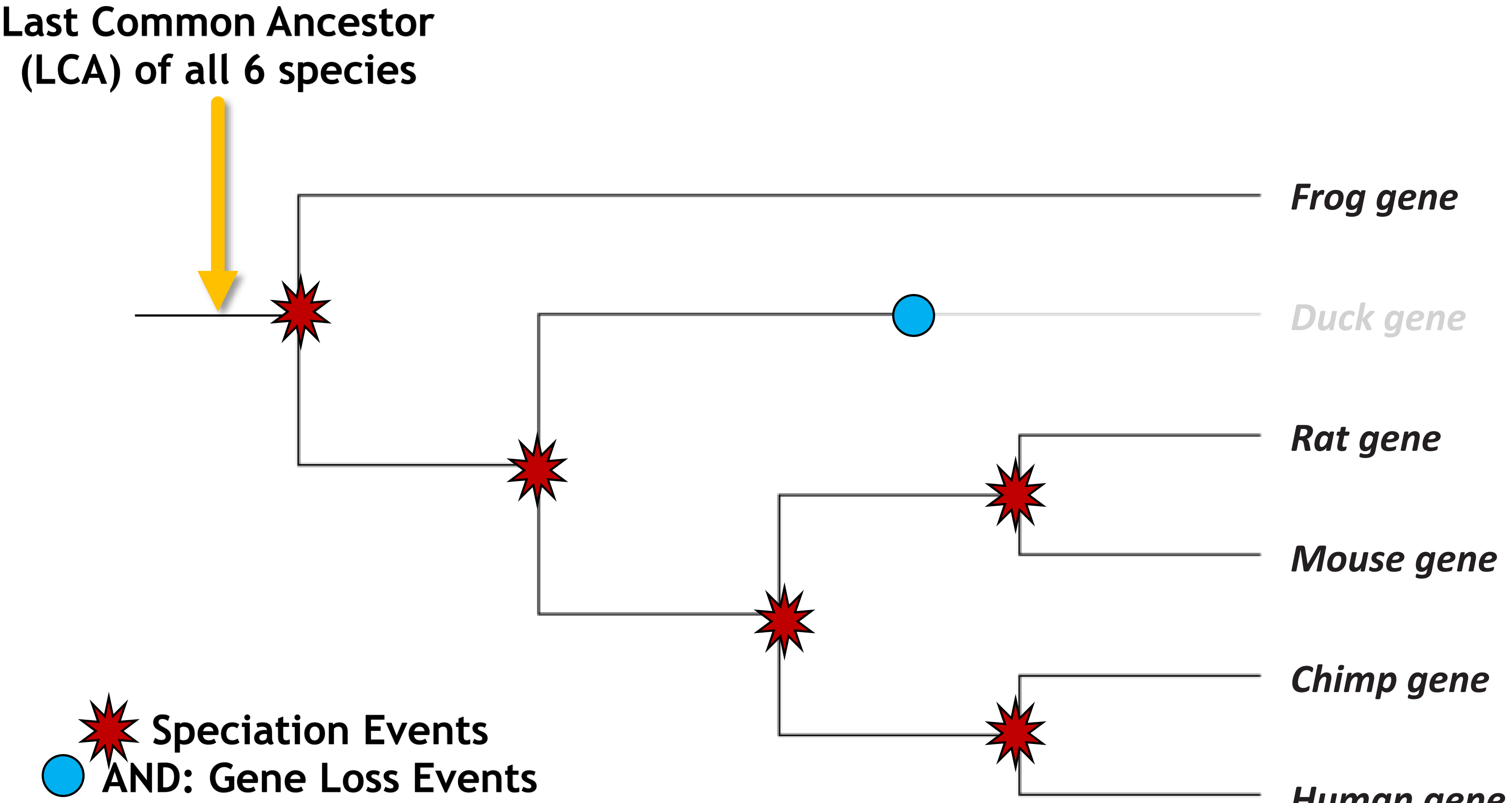
Speciation
Event

Orthology: a simple scenario



Single-Copy Orthologs

Orthology: evolution ≠ a simple scenario

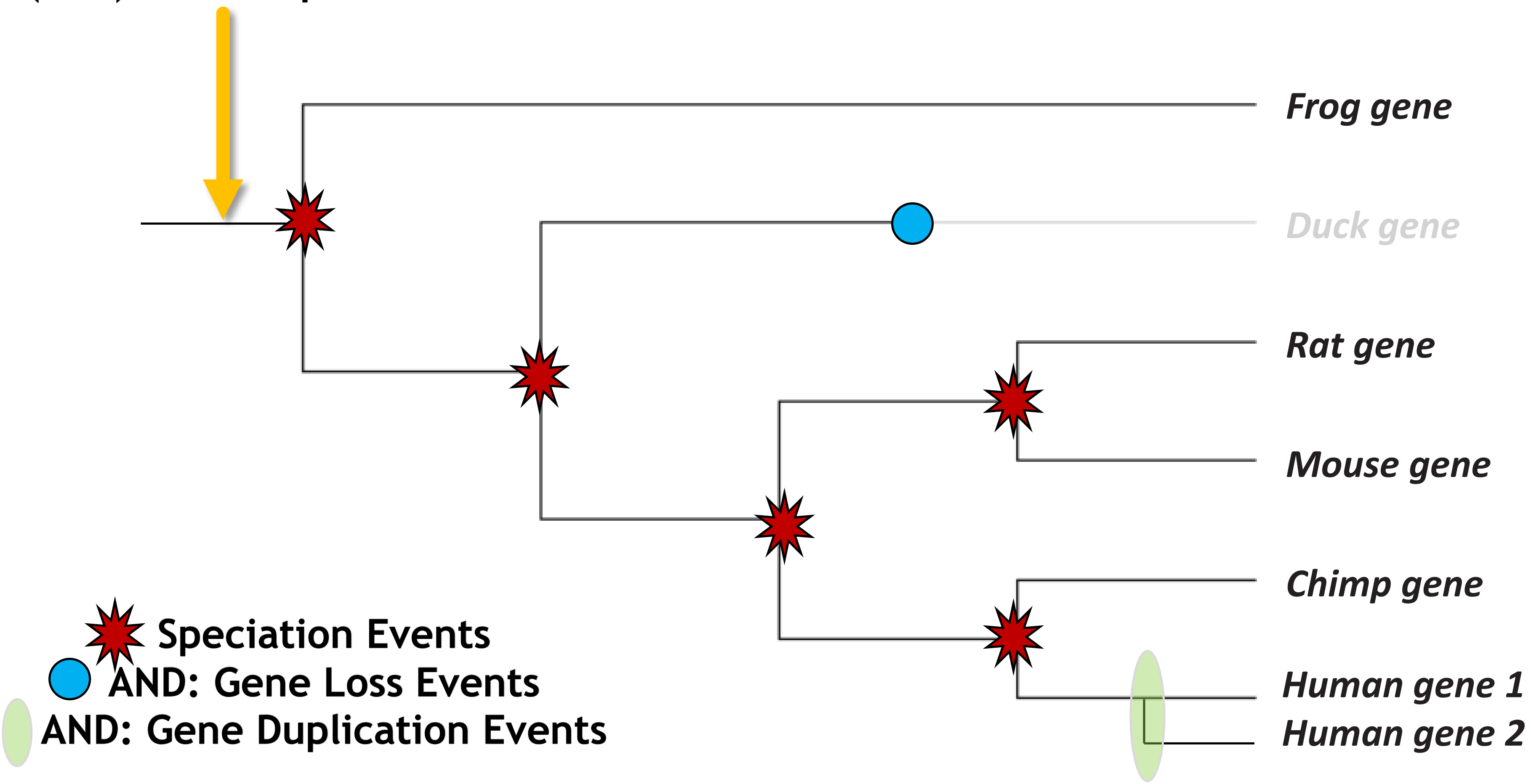


Single-Copy Orthologs with Losses

Orthology: evolution ≠ a simple scenario

Last Common Ancestor (LCA) of all 6 species

Human gene 1 & 2 = paralogs



Frog gene

Duck gene

Rat gene

Mouse gene

Chimp gene

Human gene 1

Human gene 2

✷ Speciation Events

● AND: Gene Loss Events

● AND: Gene Duplication Events

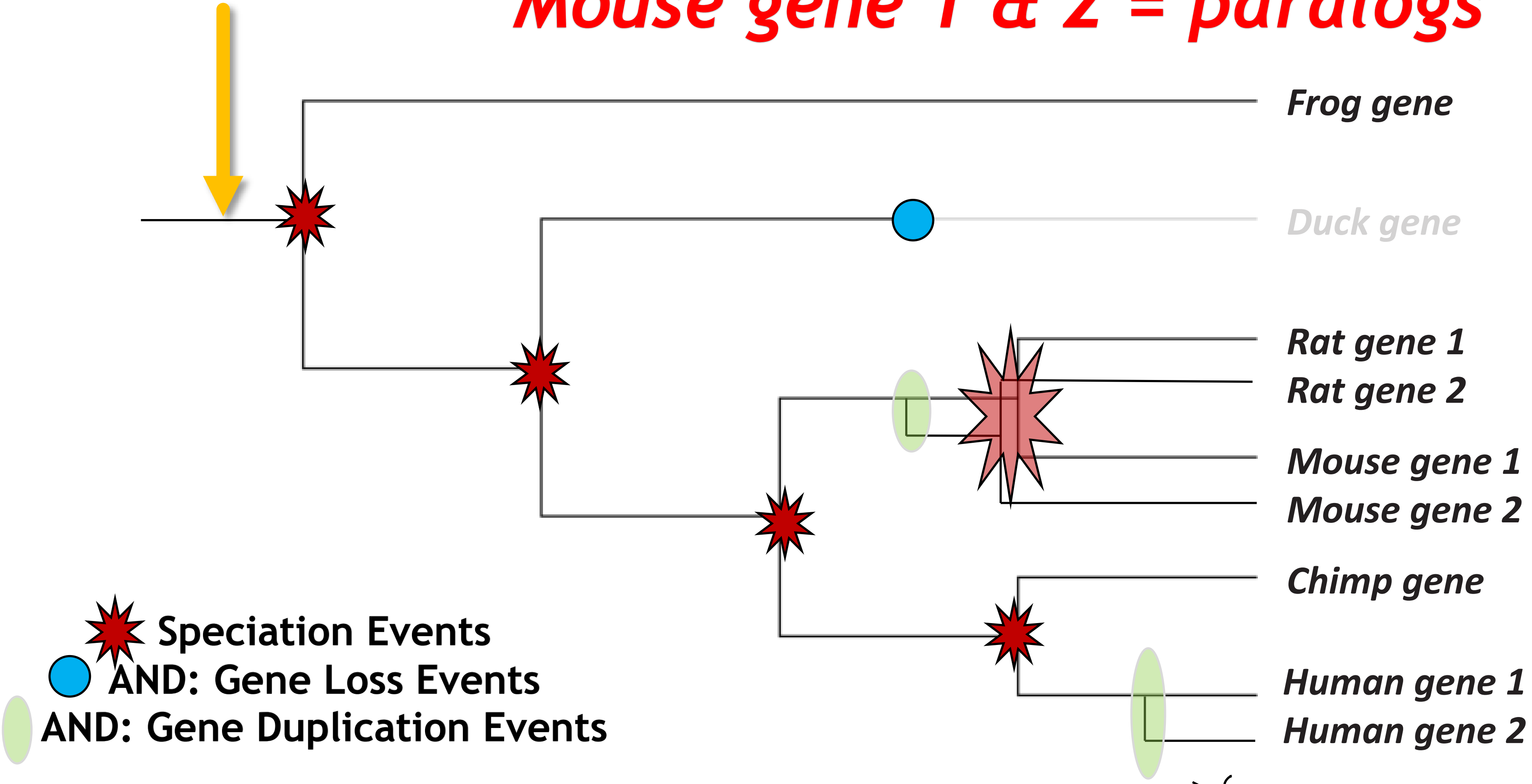
Single-Copy Orthologs with Gains

Orthology: evolution ≠ a simple scenario

Last Common Ancestor (LCA) of all 6 species

Rat gene 1 & 2 = paralogs

Mouse gene 1 & 2 = paralogs



Single-Copy Orthologs with old Gains

What is orthology?

Homology

Recognizing similarities as evidence of shared ancestry

Orthology

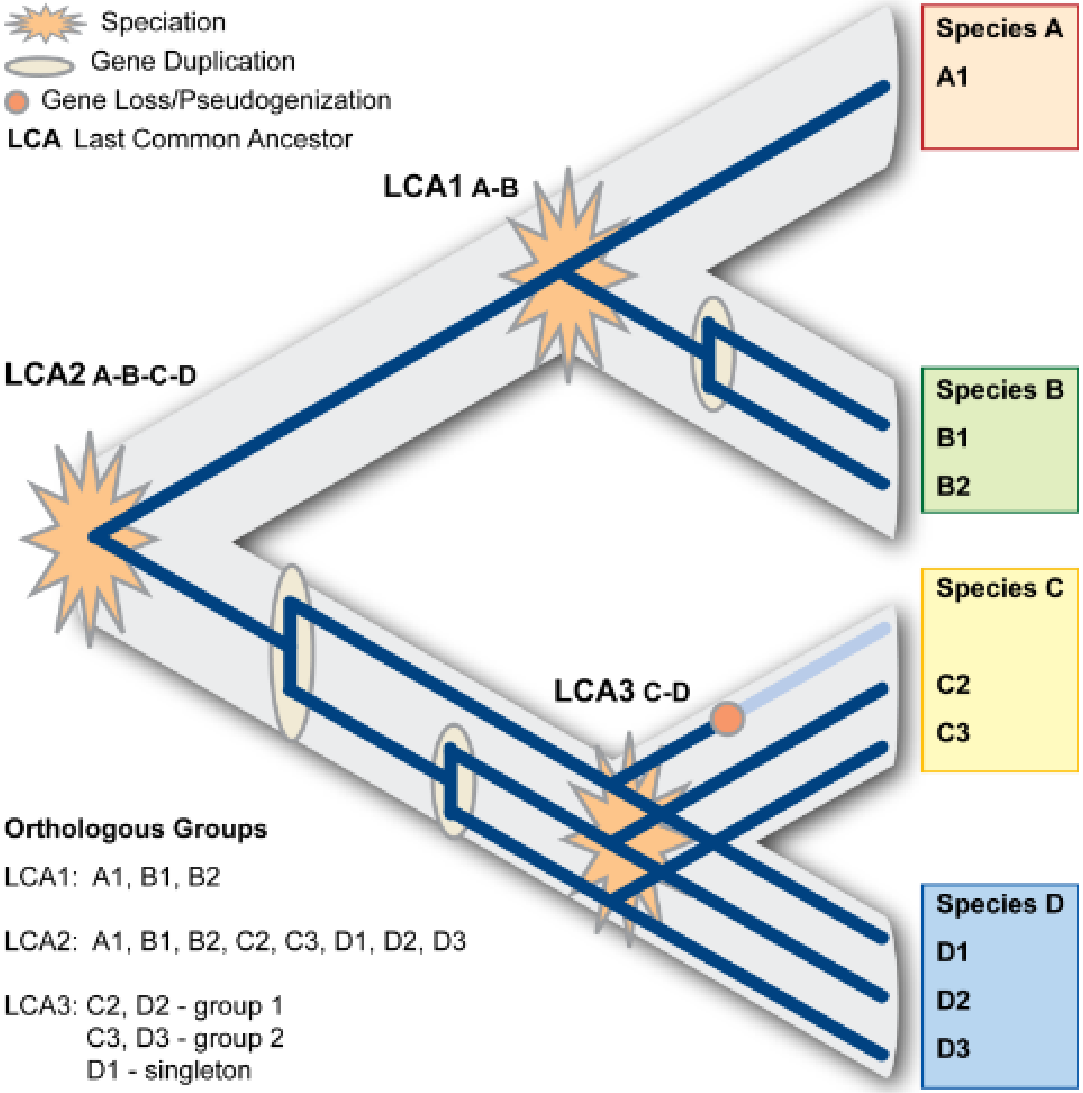
Orthologues arise by vertical descent from a single gene of the last common ancestor

Hierarchy

Orthology is relative to the species radiation under consideration

Orthologous Groups

All genes descended from single gene of last common ancestor



Orthologous Groups

LCA1: A1, B1, B2

LCA2: A1, B1, B2, C2, C3, D1, D2, D3

LCA3: C2, D2 - group 1
C3, D3 - group 2
D1 - singleton

What orthology is NOT!

Orthology & Paralogy

… are concepts defined by **evolutionary** scenarios …

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

Orthology ≠ Function

… nevertheless …

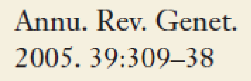
Homology refers to **common decent**, and so generally:

just as the sequences themselves are **inherited**

so too can the **biological functions** of the encoded proteins

Orthology ≈ Function

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



Annu. Rev. Genet.
2005. 39:309–38

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

What is orthology?

Understanding the definitions

How does one delineate orthology?

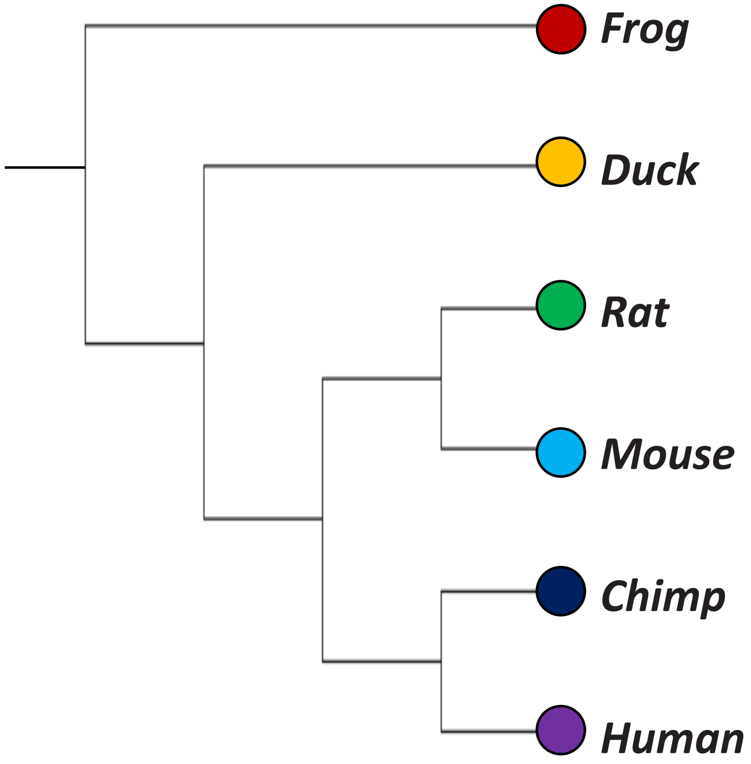
Getting to grips with the methodologies

What does OrthoDB offer?

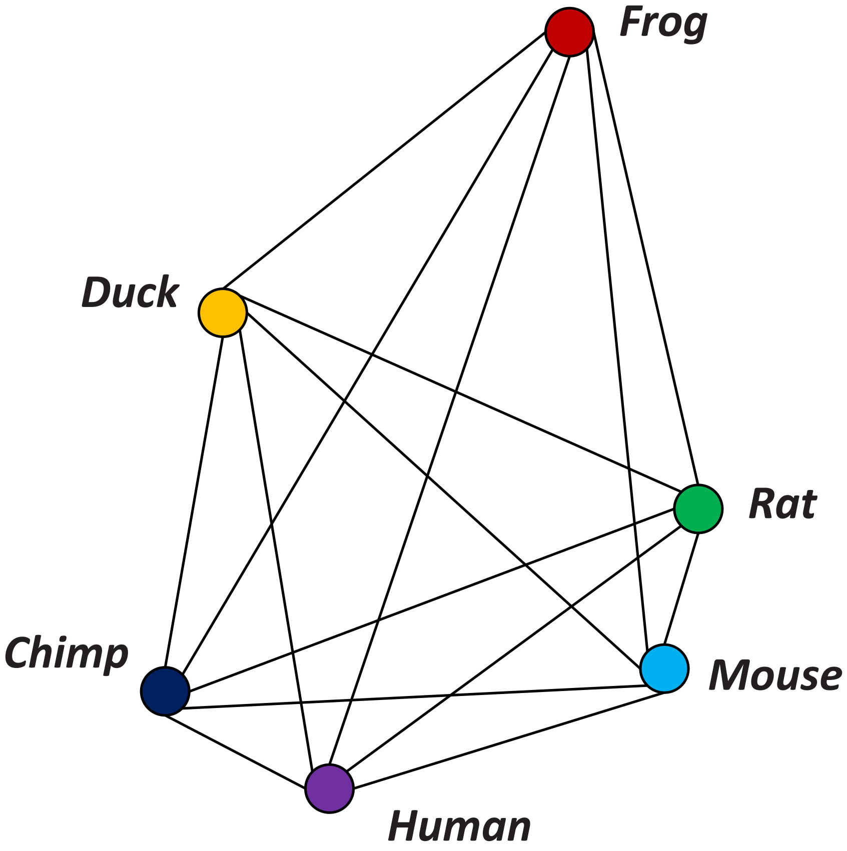
Using orthology in your research

How does one delineate orthology?

tree-based approaches



graph-based approaches

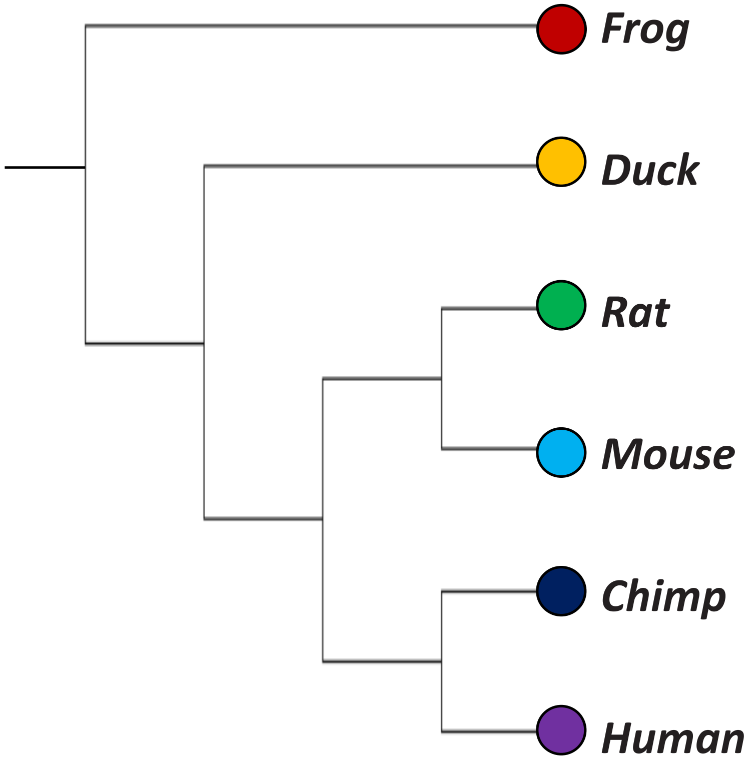


Inferring Orthology and Paralogy Methods Mol Biol. 2012;855:259-79

How does one delineate orthology?

tree-based approaches

Tree Reconciliation



Incongruences between gene and species trees can be explained in terms of speciation, duplication, and loss events on the gene tree

Most methods rely on parsimony - the most likely reconciliation is the one which requires the least number of gene duplications and losses

Inferring Orthology and Paralogy Methods Mol Biol. 2012;855:259-79

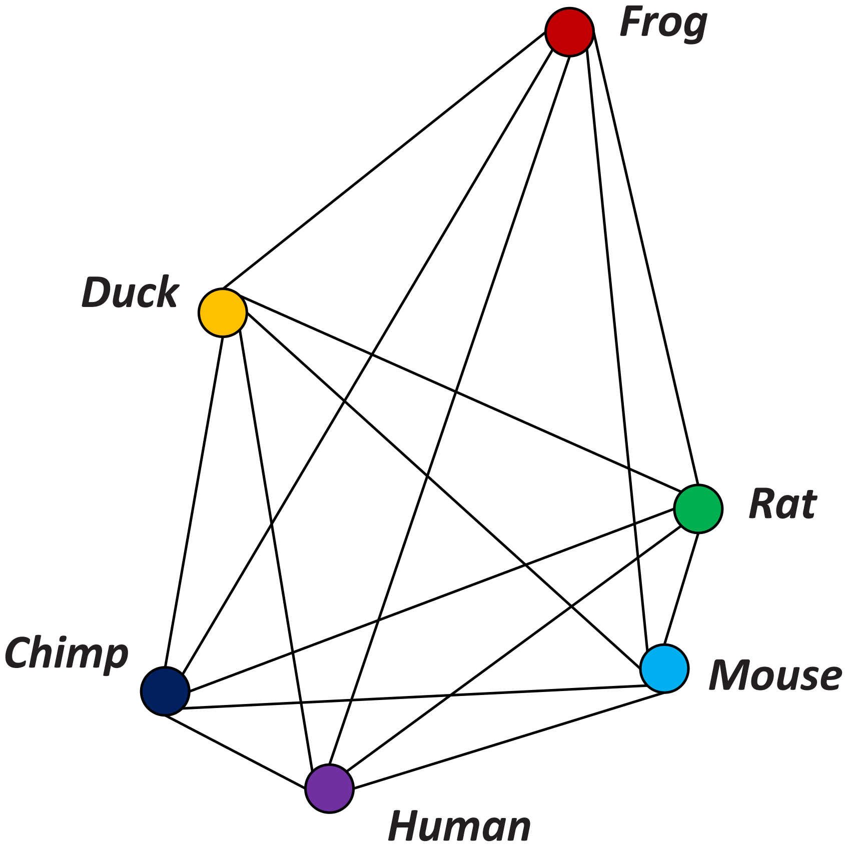
How does one delineate orthology?

Graph Clustering graph-based approaches

Graph construction by mapping all pairwise gene homologies

Genes are nodes on the graph connected by edges reflecting their ‘evolutionary distances’

Clustering then considers all pairwise information to build orthologous groups

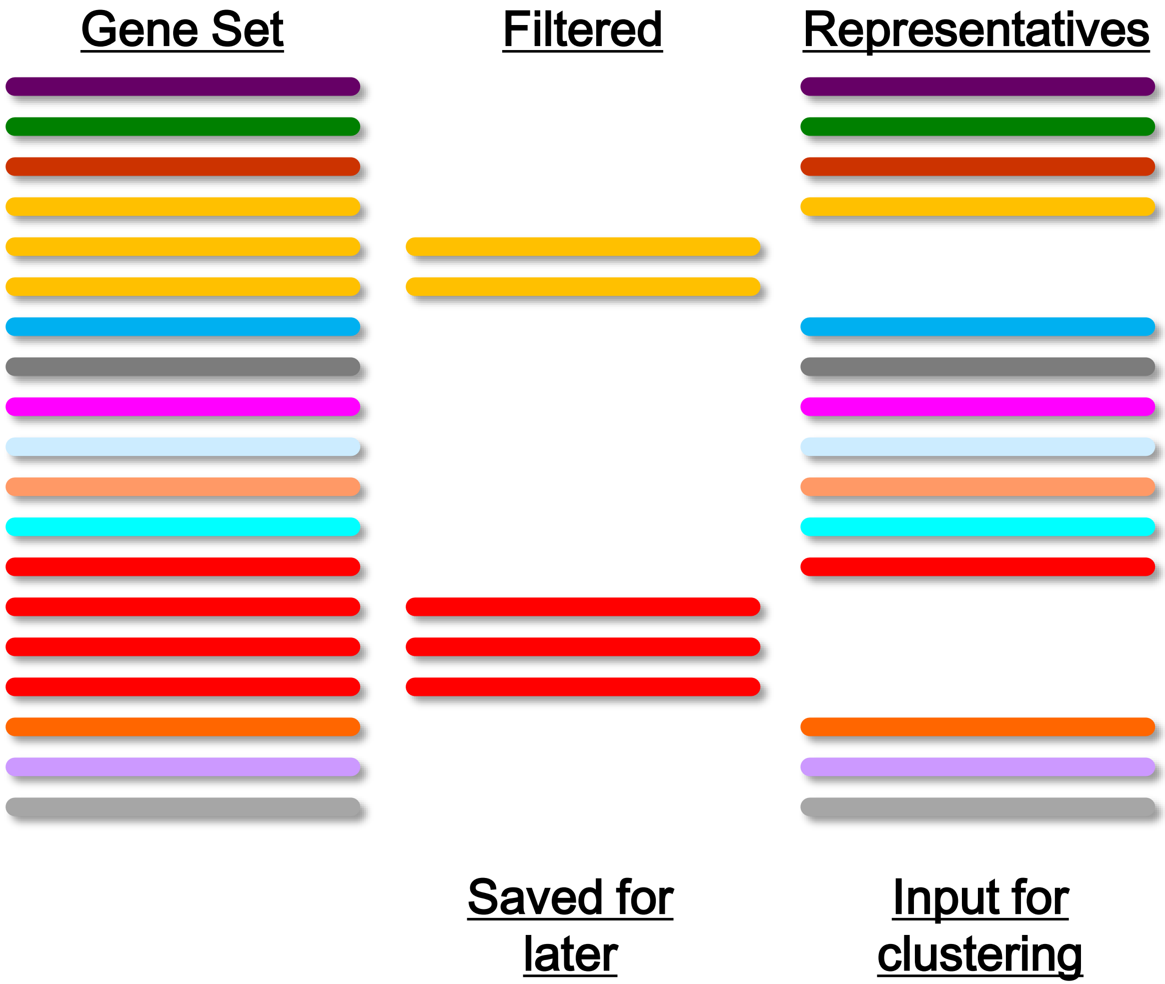


Inferring Orthology and Paralogy Methods Mol Biol. 2012;855:259-79

How does *OrthoDB* delineate orthology?

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

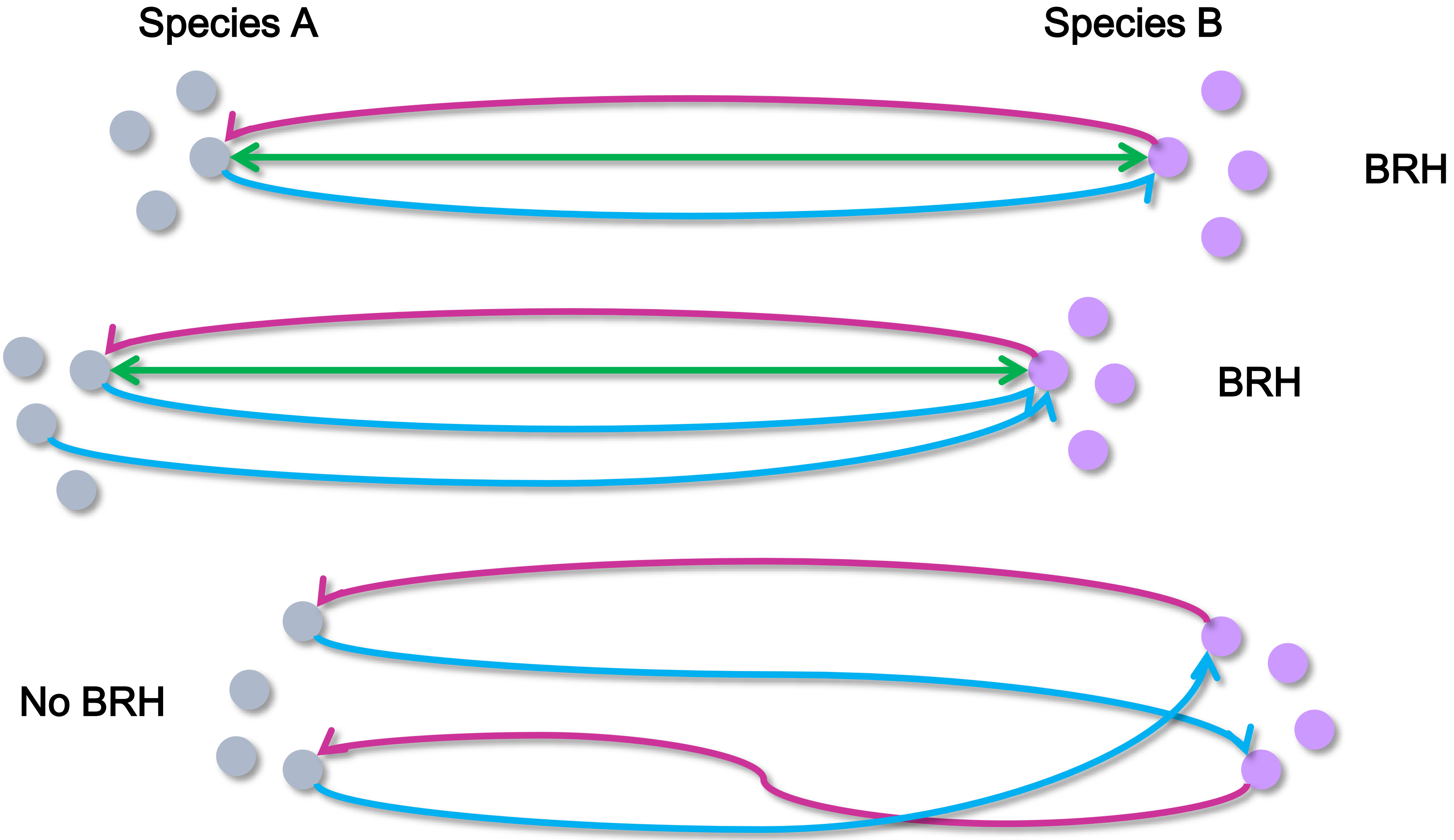
B) Remove near-identical proteins from each gene set (97% identity)



How does *OrthoDB* delineate orthology?

A) All-against-all Smith-Waterman pairwise alignments

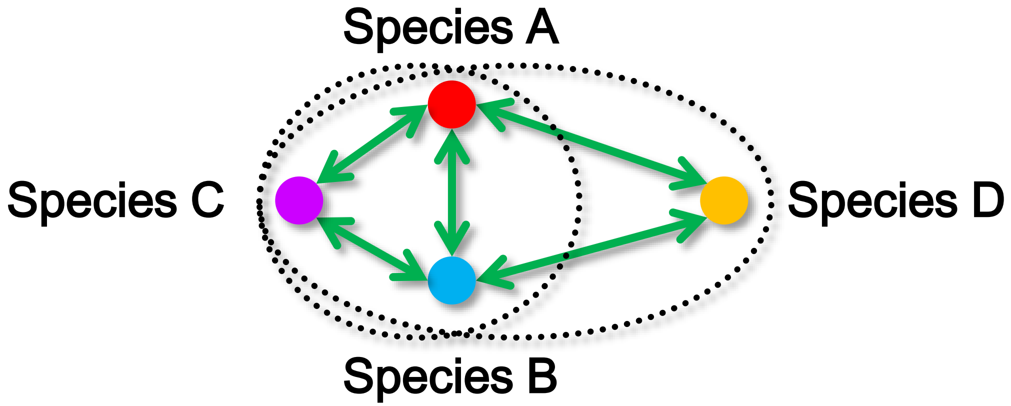
B) Define Best-Reciprocal-Hits BRHs: between proteins from species A & B

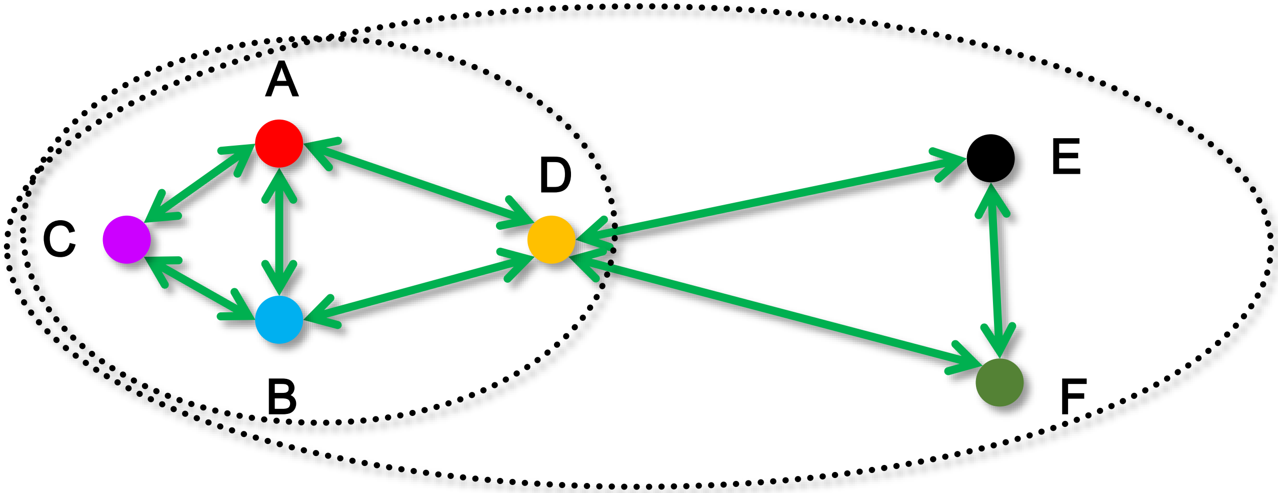


How does OrthoDB delineate orthology?

A) Build BRH TRAINGLES: start with highest-scoring BRHs and move down the list

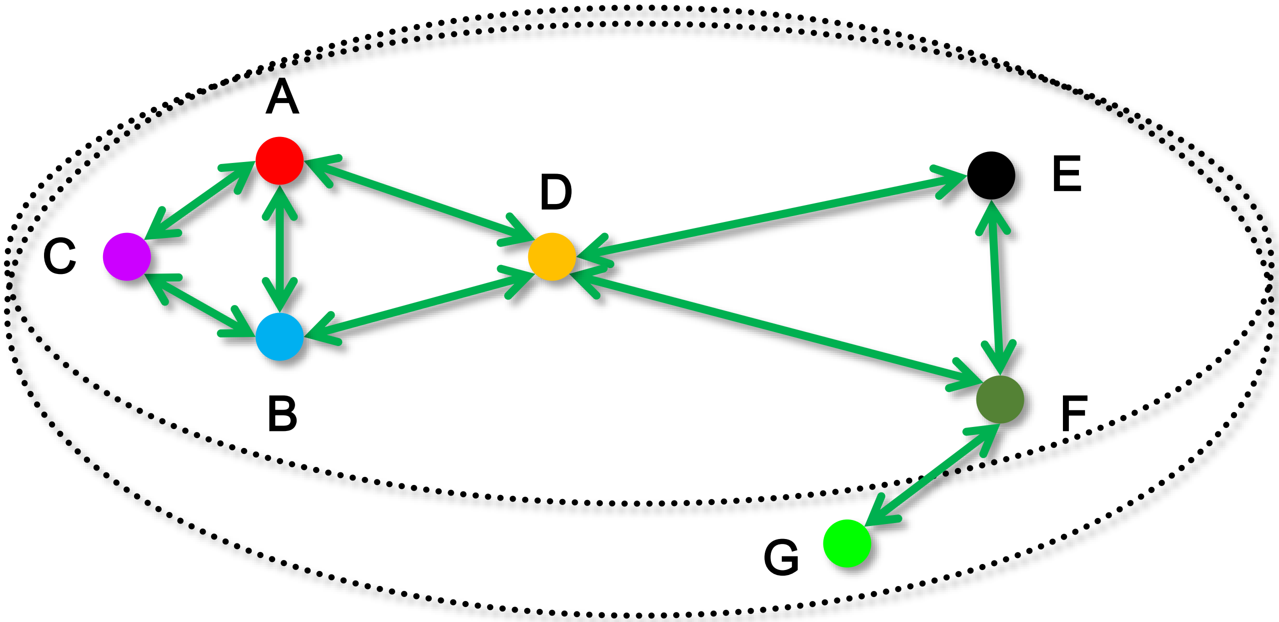
B) BRH TRIANGLES at e<1e-3 cut-off & >20aa alignment overlap





How does OrthoDB delineate orthology?

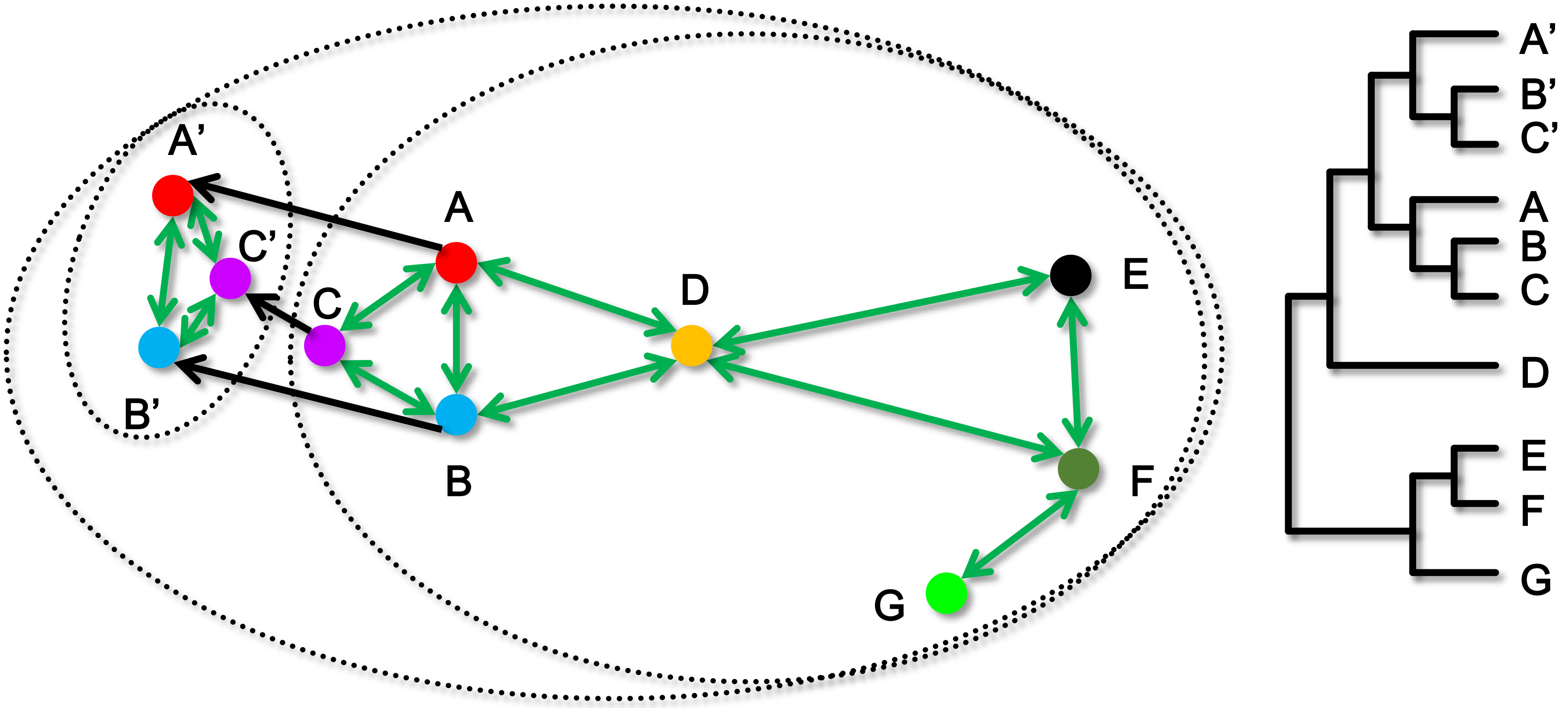
BRHs connected to triangles, but which don’t form triangles themselves
=> join clusters with e<1e-6 cut-off & >20aa alignment overlap



How does *OrthoDB* delineate orthology?

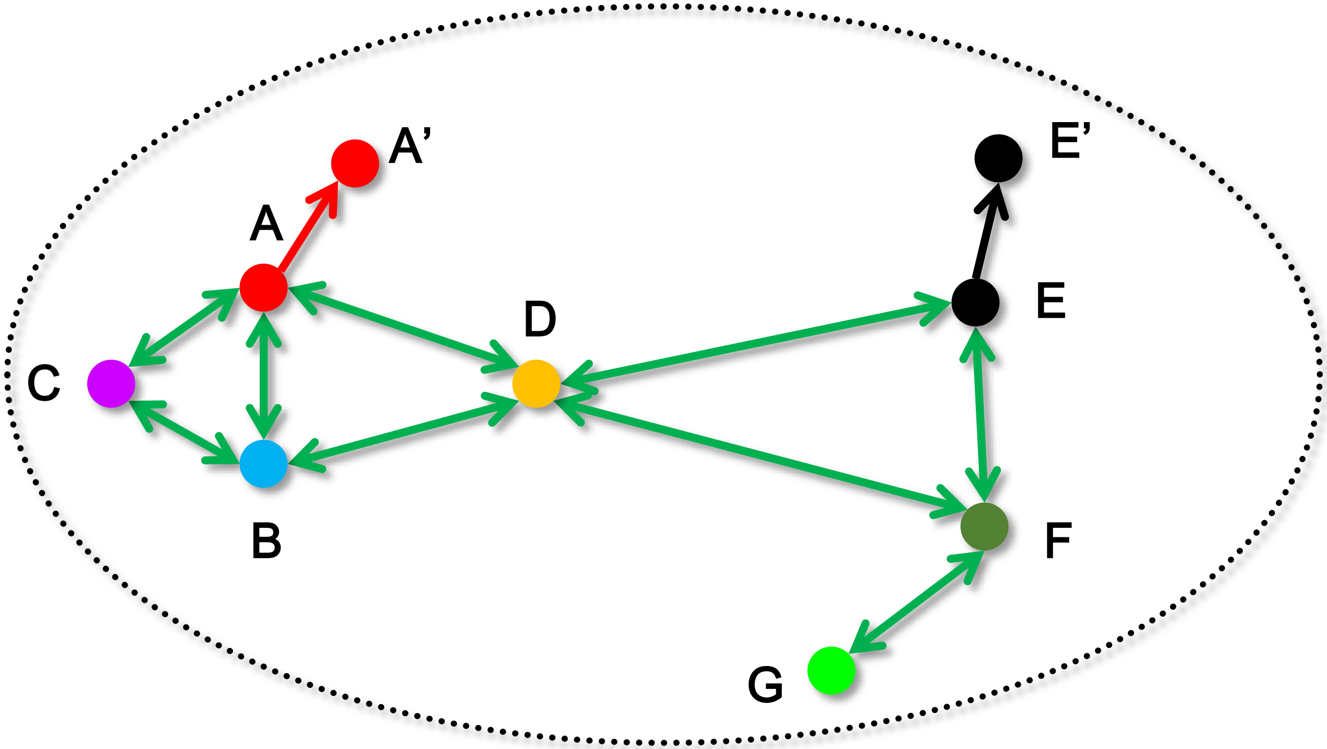
A) PARALOGOUS GROUPS: within-species homologs in different clusters

B) If the within-species homolog scores are better than the within-cluster BRH scores, the paralogous cluster can be merged into the main cluster



How does *OrthoDB* delineate orthology?

A) Consider within-species homologs that DID NOT get clustered (singletons)
B) If the within-species homolog score is better than within-cluster BRH scores, the singleton is added to the cluster as a paralog
C) Also the near-identical proteins that were initially excluded from clustering



How does OrthoDB delineate orthology?

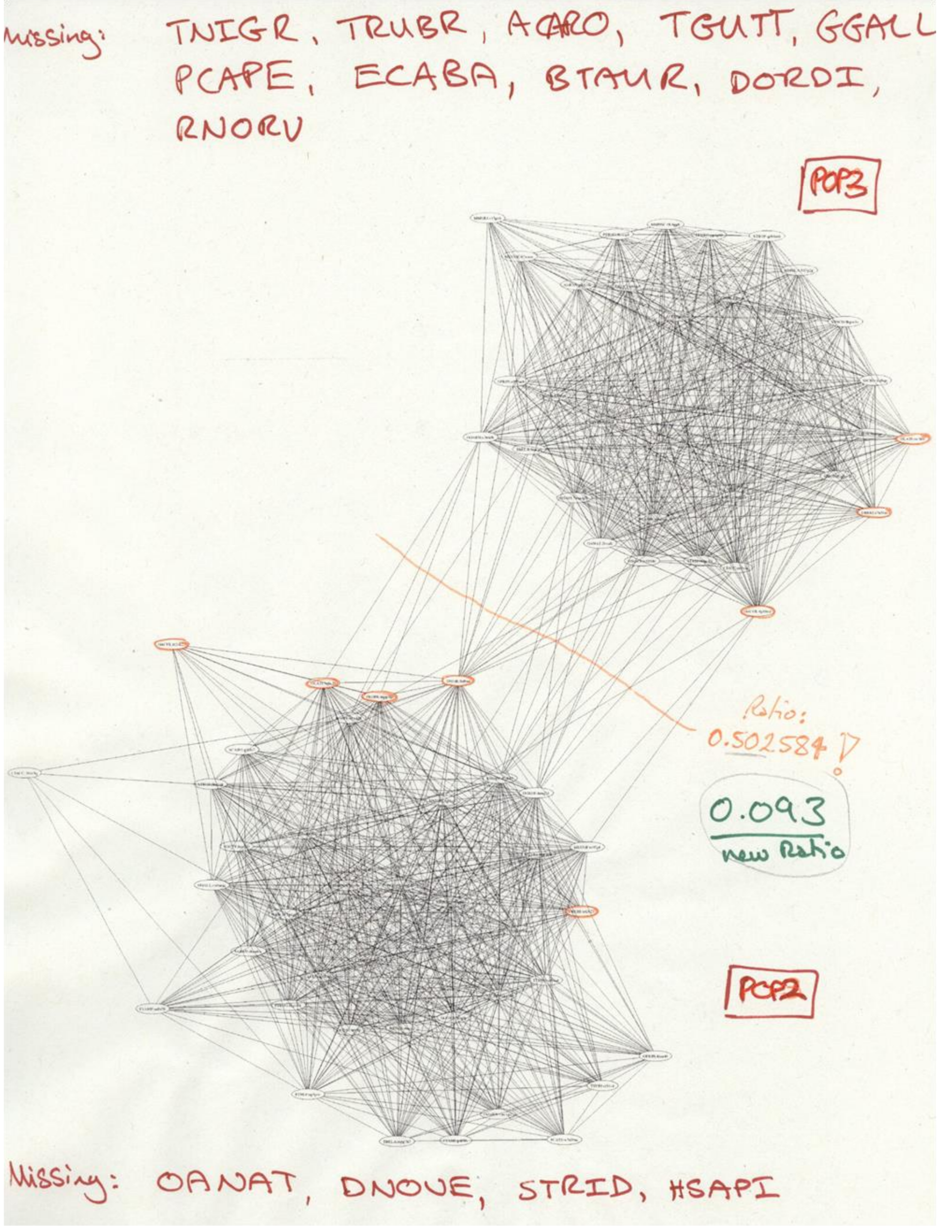
Real data are COMPLEX, e.g. cases of DIFFERENTIAL GENE LOSS

Prevent cluster merges where WITHIN-cluster connectivity is much stronger than BETWEEN-cluster connectivity

Real example:

POP3 missing from 10 vertebrates

POP2 missing from 4 vertebrates


Missing: TNIGR, TRUBR, ACARO, TGUTT, GGALL
PCAPE, ECABA, BTAUR, DORDI,
RNORV
POP3
Ratio:
0.502584 ?!
0.093
new Ratio
POP2
Missing: OANAT, DNOVE, STRID, HSAPI

How does OrthoDB delineate orthology?

Pairwise Relationships

❖ All-Against-All Alignments

❖ Delineate Best-Reciprocal-Hits

↓

Core Clusters

❖ Progressive BRH Triangulation

↓

Extended Clusters

❖ Add Pair-Only BRHs

❖ Add Paralogous Groups

❖ Add Paralogs

What is orthology?

Understanding the definitions

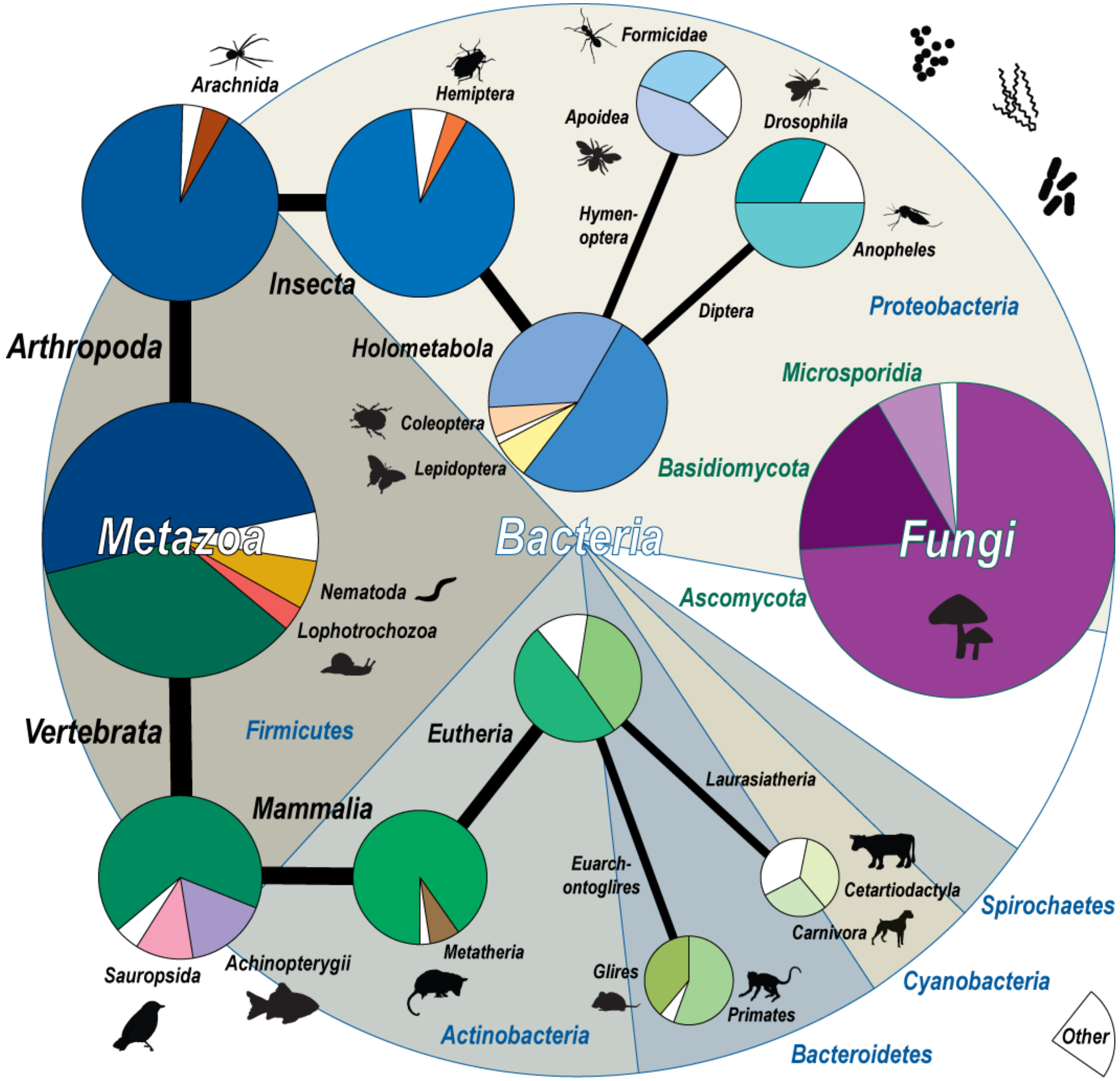
How does one delineate orthology?

Getting to grips with the methodologies

What does OrthoDB offer?

Using orthology in your research

Orthology @ OrthoDB



Species Coverage:

- 3663 Bacteria
- 435 Archaea
- 3139 Viruses
- 588 Eukaryota

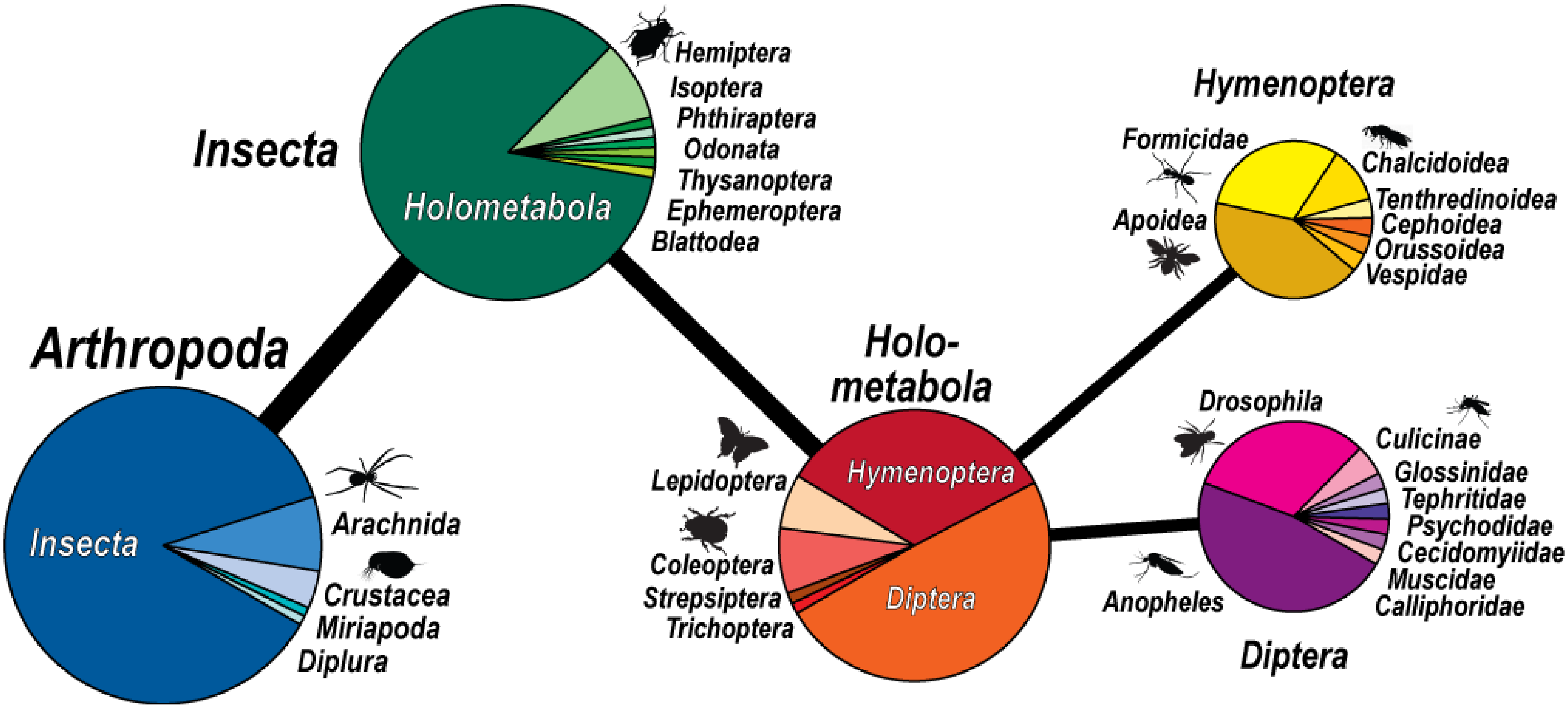
- 227 Fungi
- 31 Plants
- 330 Metazoa

- 133 Arthropoda
- 172 Vertebrata

Access:

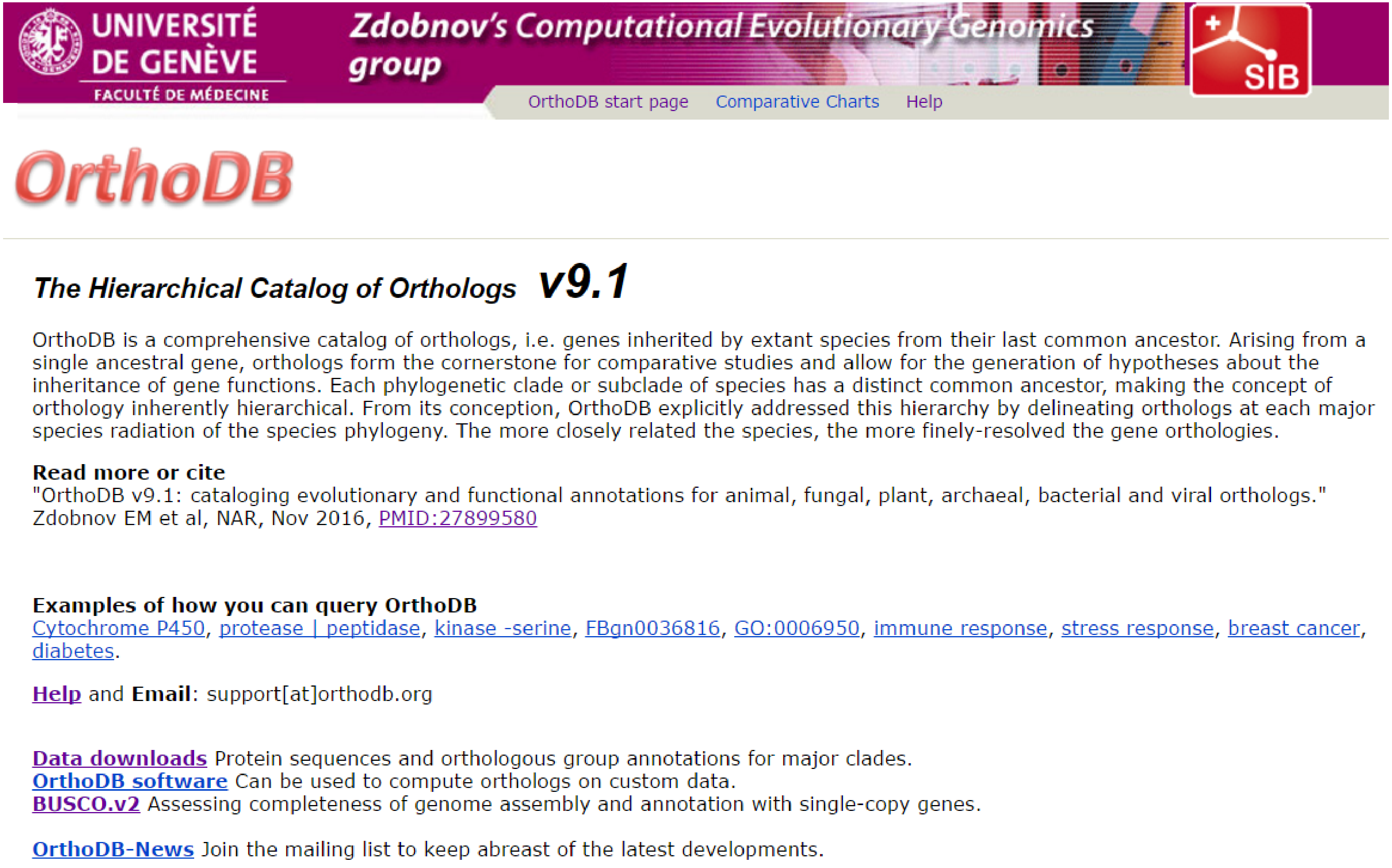
- ✓ Web browser
- ✓ JSON API
- ✓ Data downloads
- ✓ Software package

Orthology @ OrthoDB



i5K species adding to the diversity of sampled lineages!

Using OrthoDB in your research



The Hierarchical Catalog of Orthologs v9.1

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

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

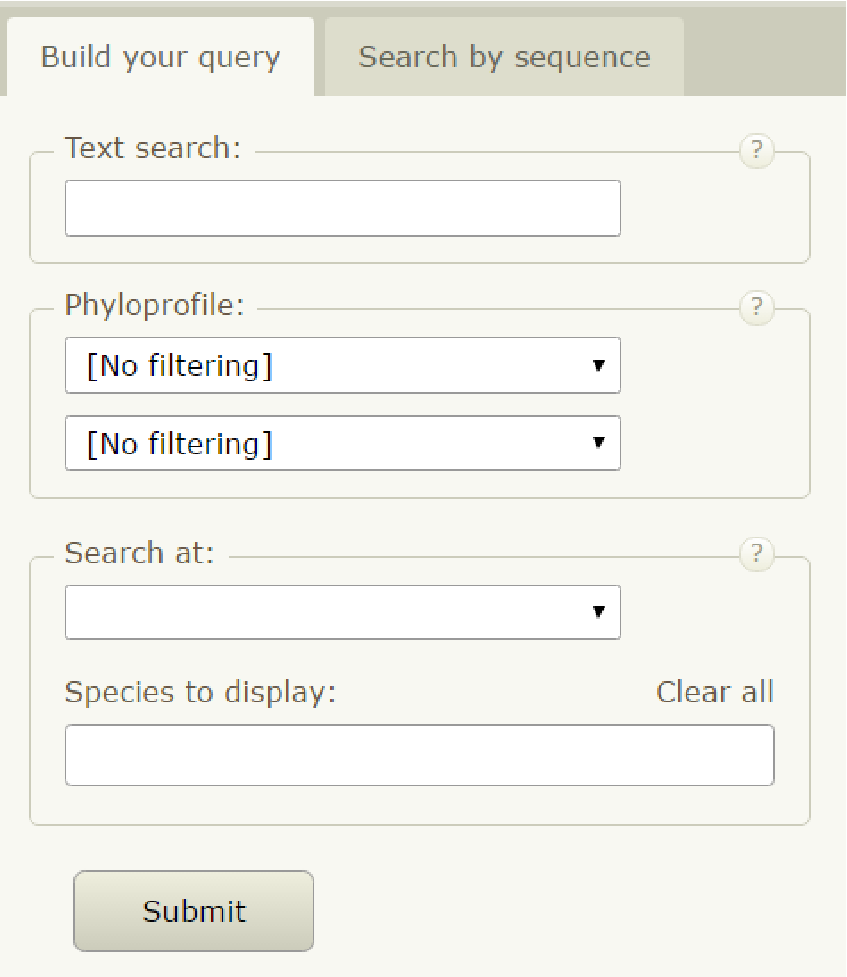
Examples of how you can query OrthoDB
Cytochrome P450, protease | peptidase, kinase -serine, FBgn0036816, GO:0006950, immune response, stress response, breast cancer, diabetes.

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

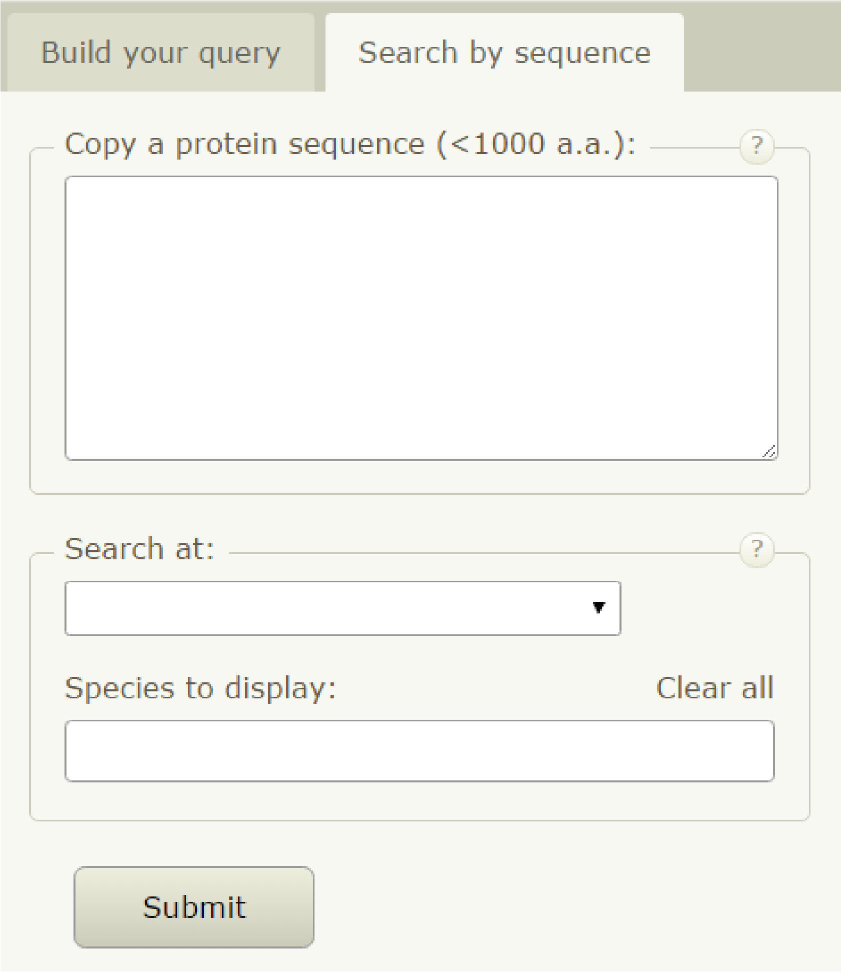
Data downloads Protein sequences and orthologous group annotations for major clades.
OrthoDB software Can be used to compute orthologs on custom data.
BUSCO.v2 Assessing completeness of genome assembly and annotation with single-copy genes.

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

Using OrthoDB in your research



Phyloprofile: [No filtering] [No filtering]

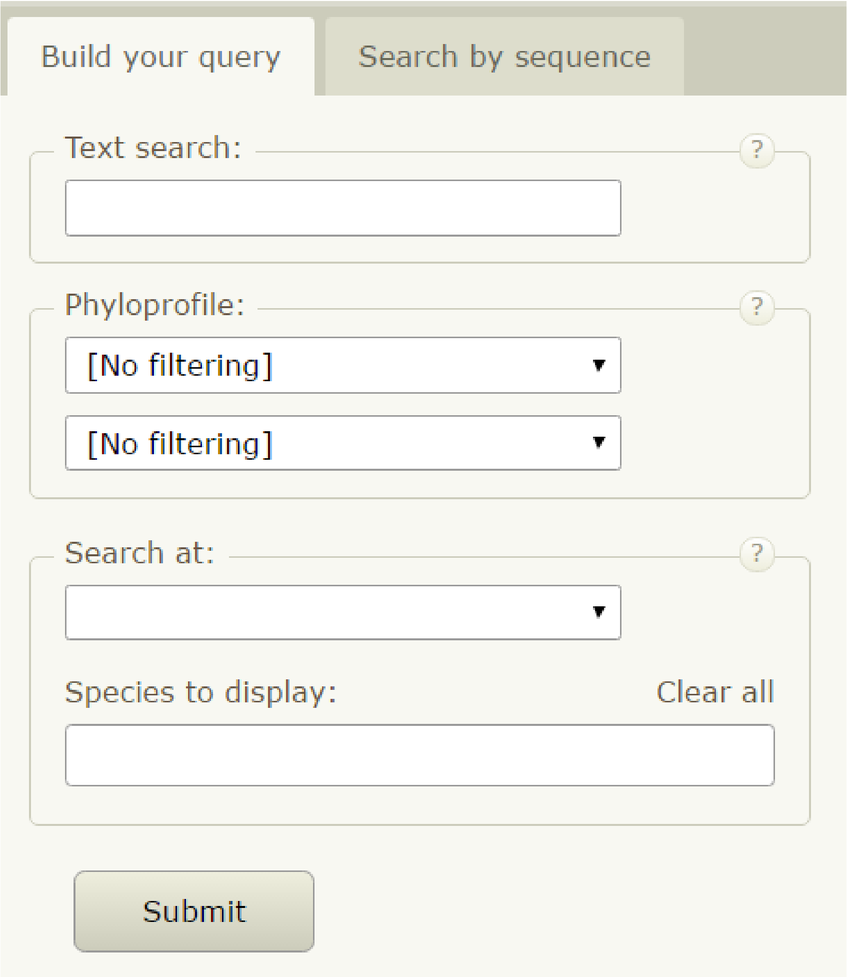


Main entry points for browsing orthology data:

TEXT SEARCH

SEQUENCE SEARCH

Using *OrthoDB* in your research

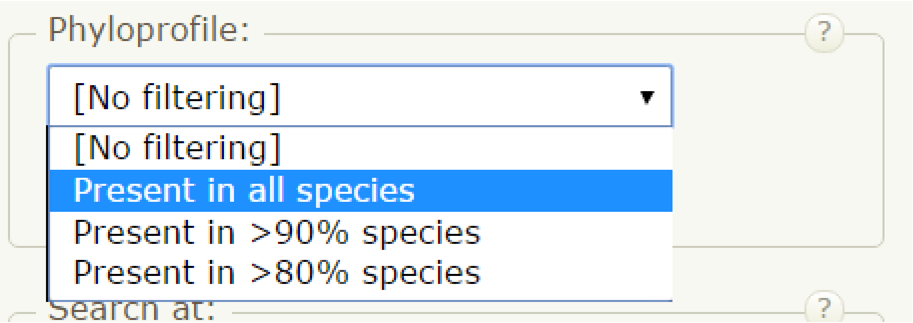


Text search:

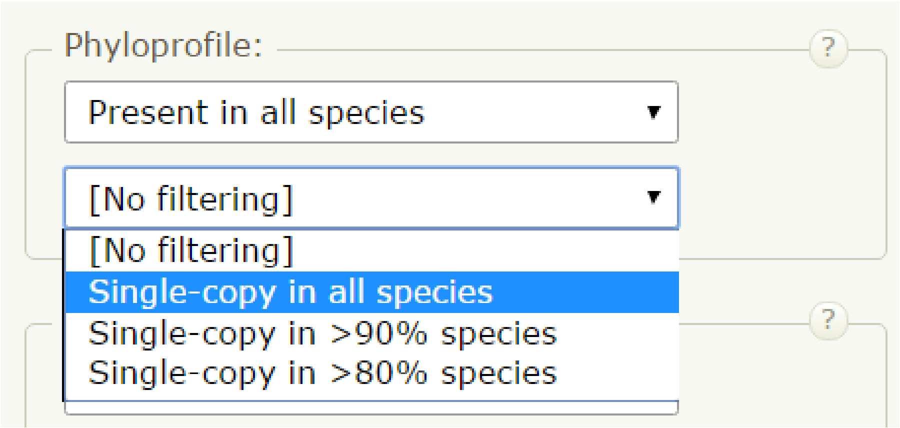
Phyloprofile:
[No filtering]
[No filtering]

Search at:

Species to display:



Phyloprofile:
[No filtering]


?
[No filtering]
Single-copy in all species
Single-copy in >90% species
Single-copy in >80% species

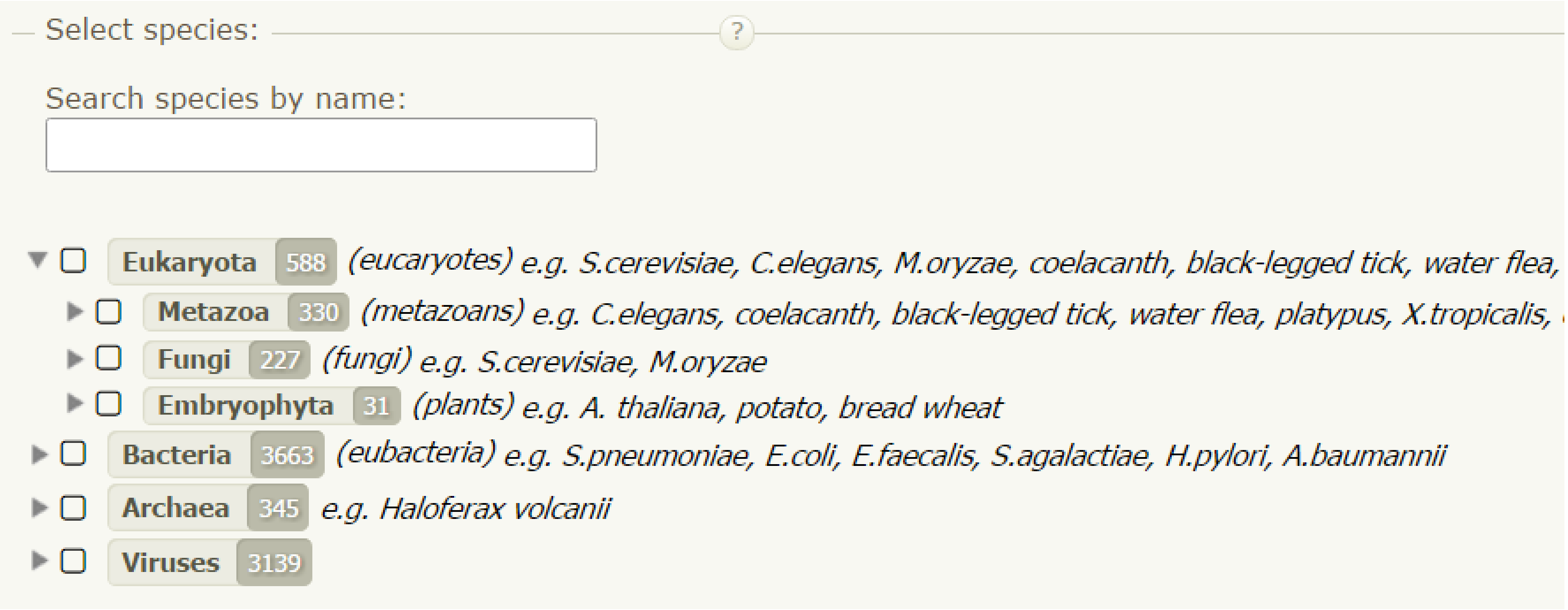
Phyloprofile:
Present in all species
[No filtering]

Phyloprofile filtering of text-search results:

PRESENCE

SINGLE-COPYNESS

Using OrthoDB in your research



Select species: ?

Search species by name:

▼ □ Eukaryota 588 (eucaryotes) e.g. S.cerevisiae, C.elegans, M.oryzae, coelacanth, black-legged tick, water flea,

▶ □ Metazoa 330 (metazoans) e.g. C.elegans, coelacanth, black-legged tick, water flea, platypus, X.tropicalis,

▶ □ Fungi 227 (fungi) e.g. S.cerevisiae, M.oryzae

▶ □ Embryophyta 31 (plants) e.g. A. thaliana, potato, bread wheat

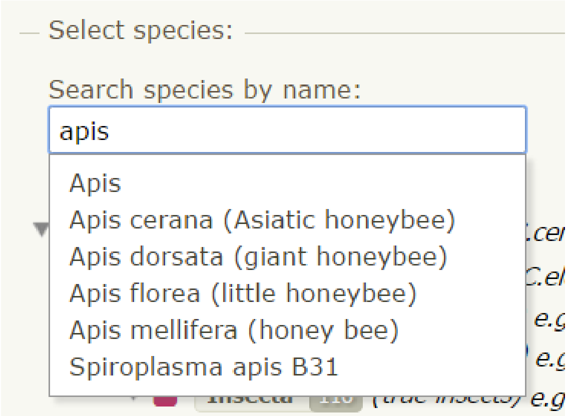
▶ □ Bacteria 3663 (eubacteria) e.g. S.pneumoniae, E.coli, E.faecalis, S.agalactiae, H.pylori, A.baumannii

▶ □ Archaea 345 e.g. Haloferax volcanii

▶ □ Viruses 3139

Select species of interest …

1) *Select species from the tree*
2) *Select nodes from the tree*
3) *Search species names to select*



Select species:

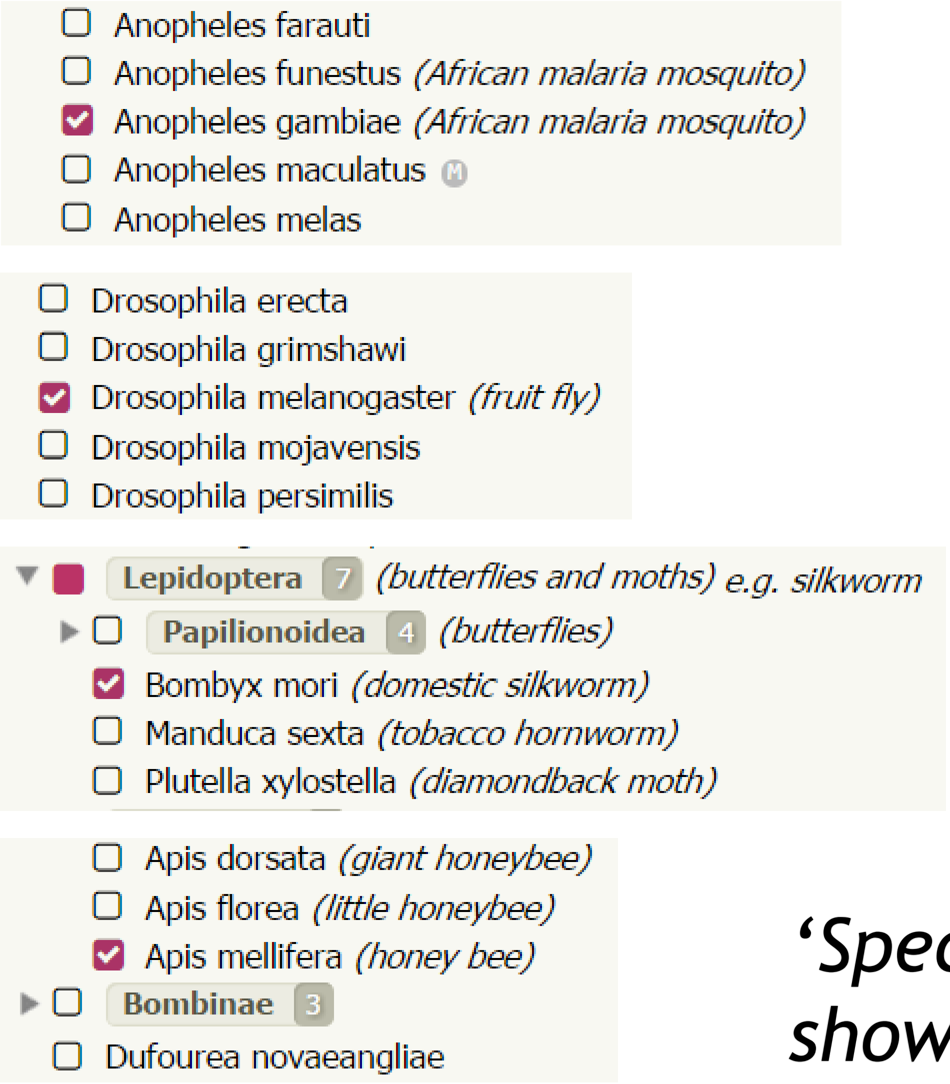
Search species by name:

apis

Apis
Apis cerana (Asiatic honeybee)
Apis dorsata (giant honeybee)
Apis florea (little honeybee)
Apis mellifera (honey bee)
Spiroplasma apis B31

Using OrthoDB in your research

Selection tree expands and selected species marked

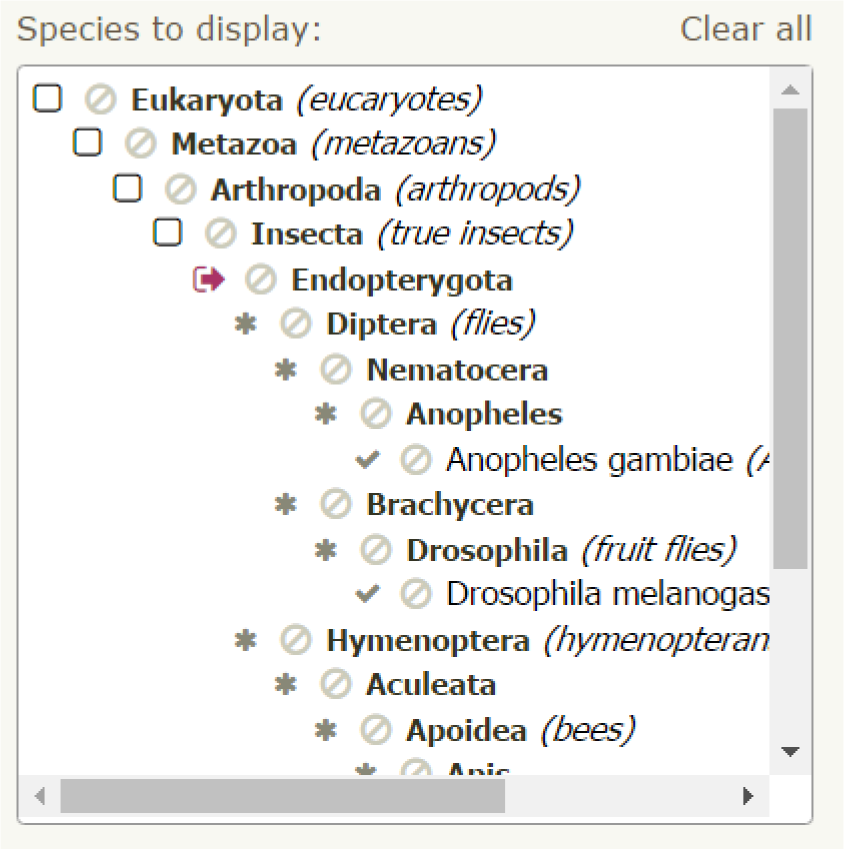


□ Anopheles farauti
□ Anopheles funestus *(African malaria mosquito)*
☑ Anopheles gambiae *(African malaria mosquito)*
□ Anopheles maculatus Ⓜ
□ Anopheles melas

□ Drosophila erecta
□ Drosophila grimshawi
☑ Drosophila melanogaster *(fruit fly)*
□ Drosophila mojavensis
□ Drosophila persimilis

▼ ■ **Lepidoptera** 7 *(butterflies and moths) e.g. silkworm*
▶ □ **Papilionoidea** 4 *(butterflies)*
☑ Bombyx mori *(domestic silkworm)*
□ Manduca sexta *(tobacco hornworm)*
□ Plutella xylostella *(diamondback moth)*

□ Apis dorsata *(giant honeybee)*
□ Apis florea *(little honeybee)*
☑ Apis mellifera *(honey bee)*
▶ □ **Bombinae** 3
□ Dufourea novaeangliae



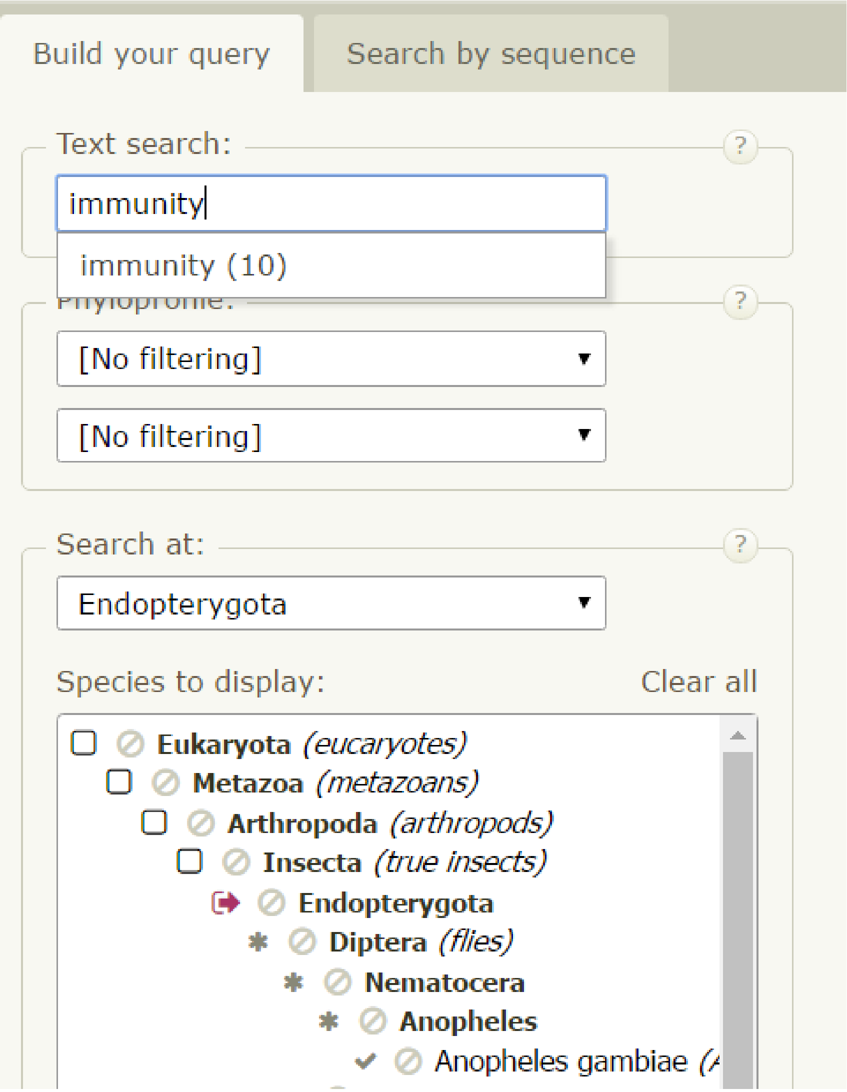
Species to display: Clear all

□ ⊘ **Eukaryota** *(eucaryotes)*
□ ⊘ **Metazoa** *(metazoans)*
□ ⊘ **Arthropoda** *(arthropods)*
□ ⊘ **Insecta** *(true insects)*
➪ ⊘ **Endopterygota**
✱ ⊘ **Diptera** *(flies)*
✱ ⊘ **Nematocera**
✱ ⊘ **Anopheles**
✓ ⊘ Anopheles gambiae (
✱ ⊘ **Brachycera**
✱ ⊘ **Drosophila** *(fruit flies)*
✓ ⊘ Drosophila melanogas
✱ ⊘ **Hymenoptera** *(hymenopteran*
✱ ⊘ **Aculeata**
✱ ⊘ **Apoidea** *(bees)*
✱ ⊘ **Apis**

‘Species to display’ panel now shows only selected species

Using *OrthoDB* in your research

Search for a term, e.g. immunity



Build your query | Search by sequence

Text search: ?

immunity

immunity (10)

Phyloprofile: ?

[No filtering] ▾

[No filtering] ▾

Search at: ?

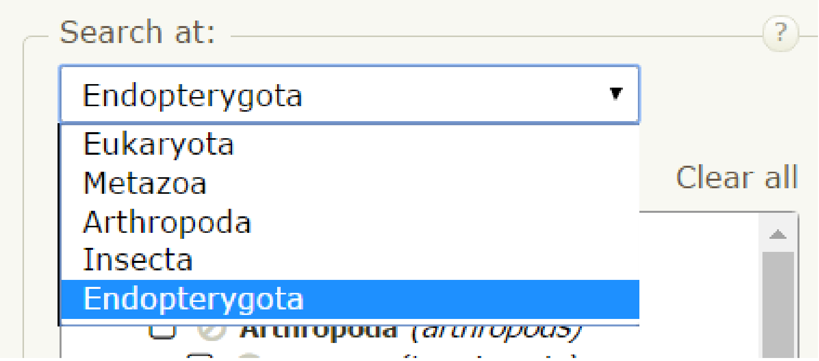
Endopterygota ▾

Species to display: Clear all

- □ Eukaryota *(eucaryotes)*
- □ Metazoa *(metazoans)*
- □ Arthropoda *(arthropods)*
- □ Insecta *(true insects)*
- Endopterygota
- ✱ Diptera *(flies)*
- ✱ Nematocera
- ✱ Anopheles
- ✓ Anopheles gambiae (

Autocomplete with counts of cached terms shown

NB: 'Search at' is now set automatically to the last common ancestor level of all the species you selected You can choose older one:



Search at: ?

Endopterygota ▾

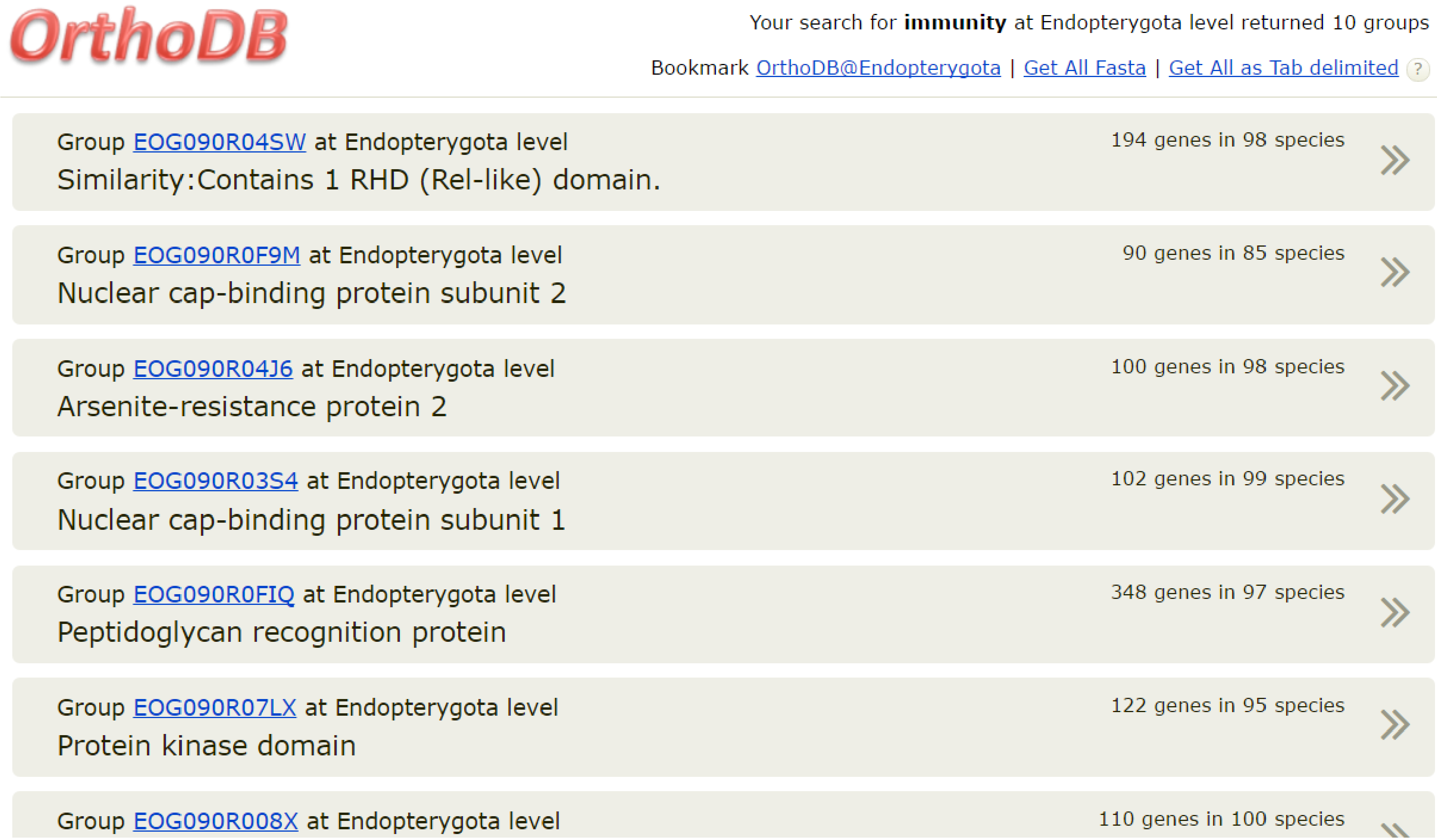
- Eukaryota
- Metazoa
- Arthropoda
- Insecta
- Endopterygota

Clear all

□ Arthropoda *(arthropods)*

Using OrthoDB in your research

10 orthologous groups returned



Your search for **immunity** at Endopterygota level returned 10 groups

Bookmark OrthoDB@Endopterygota | Get All Fasta | Get All as Tab delimited ?

Group EOG090R04SW at Endopterygota level — 194 genes in 98 species
Similarity:Contains 1 RHD (Rel-like) domain.

Group EOG090R0F9M at Endopterygota level — 90 genes in 85 species
Nuclear cap-binding protein subunit 2

Group EOG090R04J6 at Endopterygota level — 100 genes in 98 species
Arsenite-resistance protein 2

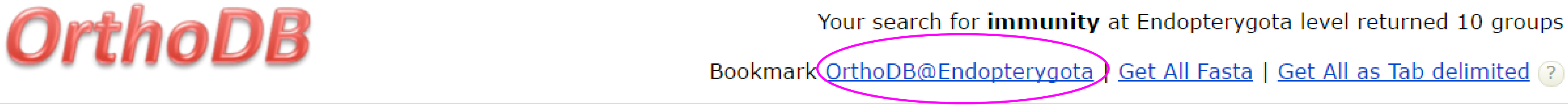
Group EOG090R03S4 at Endopterygota level — 102 genes in 99 species
Nuclear cap-binding protein subunit 1

Group EOG090R0FIQ at Endopterygota level — 348 genes in 97 species
Peptidoglycan recognition protein

Group EOG090R07LX at Endopterygota level — 122 genes in 95 species
Protein kinase domain

Group EOG090R008X at Endopterygota level — 110 genes in 100 species

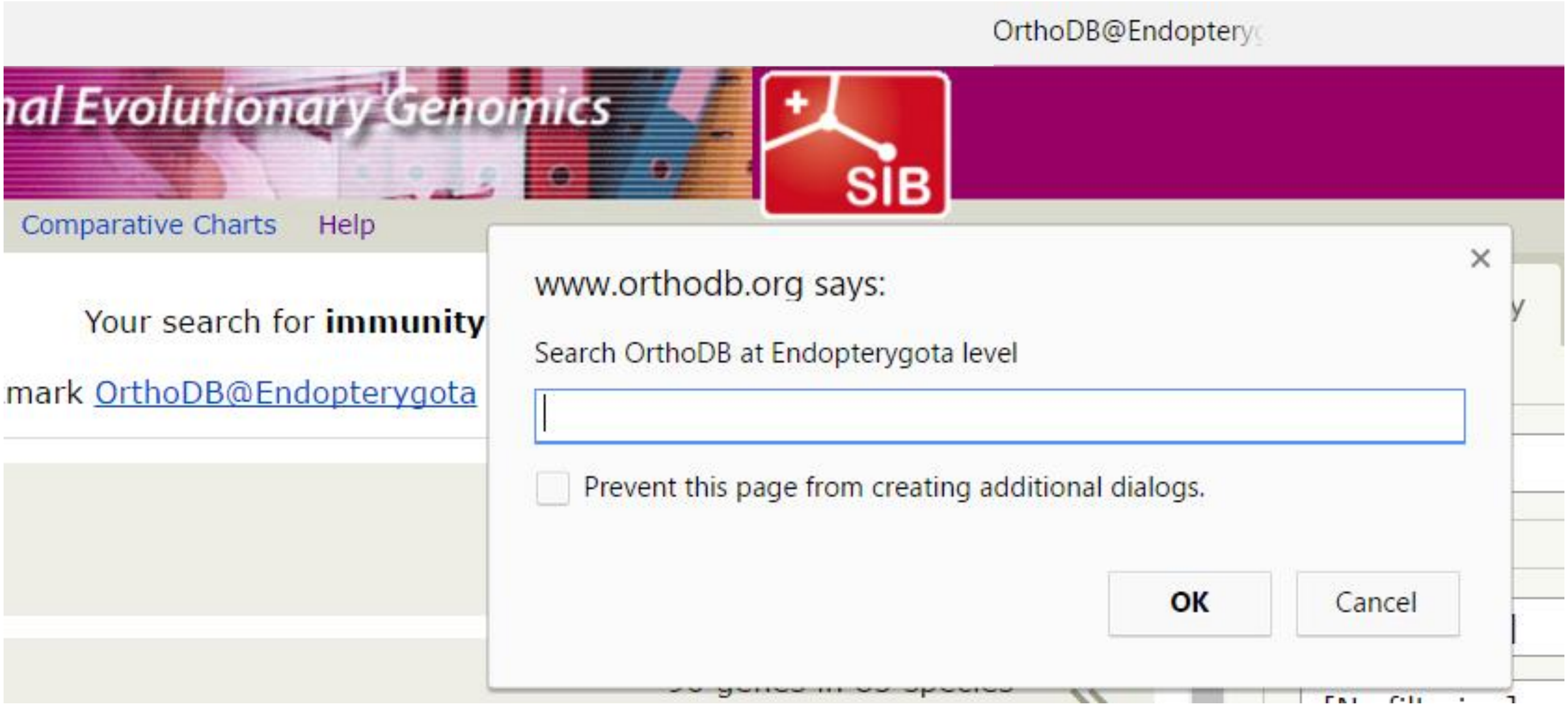
Using *OrthoDB* in your research



Your search for **immunity** at Endopterygota level returned 10 groups

Bookmark OrthoDB@Endopterygota | Get All Fasta | Get All as Tab delimited ?

1. Drag the Bookmarklet link to your toolbar to be able to quickly and easily conduct future searches at this level



Your search for **immunity**

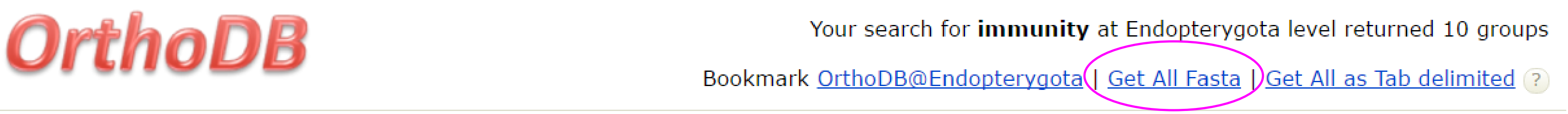
mark OrthoDB@Endopterygota

www.orthodb.org says:

Search OrthoDB at Endopterygota level

□ Prevent this page from creating additional dialogs.

Using OrthoDB in your research

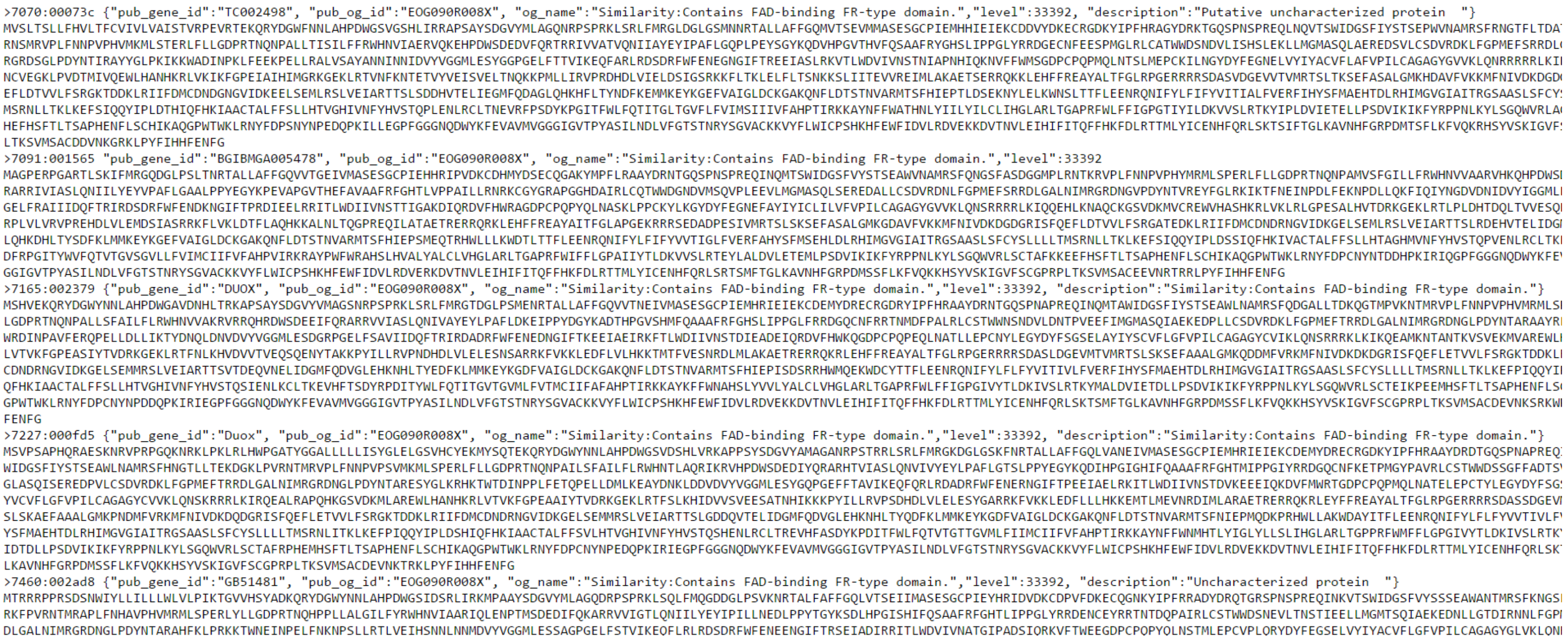


OrthoDB

Your search for **immunity** at Endopterygota level returned 10 groups

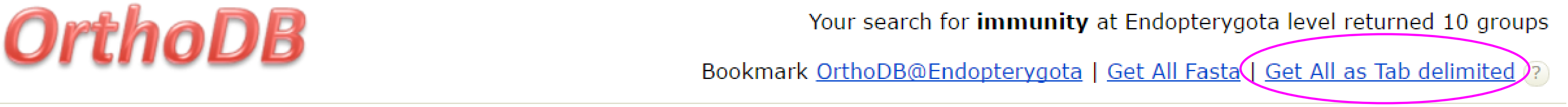
Bookmark OrthoDB@Endopterygota | Get All Fasta | Get All as Tab delimited ?

2. Get ALL protein sequences (FASTA format) from the selected species for ALL 10 of the search result orthologous groups



```
>7070:00073c {"pub_gene_id":"TC002498", "pub_og_id":"EOG090R008X", "og_name":"Similarity:Contains FAD-binding FR-type domain.","level":33392, "description":"Putative uncharacterized protein  "}
MVSLTSLLFHVLTFCVIVLVAISTVRPEVRTEKQRYDGWFNNLAHPDWGSVGSHLIRRAPSAYSDGVYMLAGQNRPSPRKLSRLFMRGLDGLGSMNNRTALLAFFGQMVTSEVMMASESGCPIEMHHIEIEKCDDVYDKECRGDKYIPFHRAGYDRKTGQSPNSPREQLNQVTSWIDGSFIYSTSEPWVNAMRSFRNGTFLTDA
RNSMRVPLFNNPVPHVMKMLSTERLFLLGDPRTNQNPALLTISILFFRWHNVIAERVQKEHPDWSDEDVFQRTRRIVVATVQNIIAYEYIPAFLGQPLPEYSGYKQDVHPGVTHVFQSAAFRYGHSLIPPGLYRRDGECNFEESPMGLRLCATWWDSNDVLISHSLEKLLMGMASQLAEREDSVLCSDVRDKLFGPMEFSRRDL
RGRDSGLPDYNTIRAYYGLPKIKKWADINPKLFEEKPELLRALVSAYANNINNIDVYVGGMLESYGGPGELFTTVIKEQFARLRDSDRFWFENEGNGIFTREEIASLRKVTLWDVIVNSTNIAPNHIQKNVFFWMSGDPCPQPMQLNTSLMEPCKILNGYDYFEGNELVYIYACVFLAFVPILCAGAGYGVVKLQNRRRRRLKII
NCVEGKLPVDTMIVQEWLHANHKRLVKIKFGPEIAIHIMGRKGEKLRTVNFKNTETVYYVEISVELTNQKKPMLLIRVPRDHDLVIELDSIGSRKKFLTKLELFLTSNKKSLIITEVVREIMLAKAETSERRQKKLEHFFREAYALTFGLRPGERRRRSDASVDGEVVTVMRTSLTKSEFASALGMKHDAVFVKKMFNIVDKDGD
EFLDTVVLFSRGKTDDKLRIIFDMCDNDGNGVIDKEELSEMLRSLVEIARTTSLSDDHVTELIEGMFQDAGLQHKHFLTYNDFKEMMKEYKGEFVAIGLDCKGAKQNFLDTSTNVARMTSFHIEPTLDSEKNYLELKWNSLTTFLEENRQNIFYLFIFYVITIALFVERFIHYSFMAEHTDLRHIMGVGIAITRGSAASLSFCY
MSRNLLTKLKEFSIQQYIPLDTHIQFHKIAACTALFFSLLHTVGHIVNFYHVSTQPLENLRCLTNEVRFPSDYKPGITFWLFQTITGLTGVFLFVIMSIIIVFAHPTIRKKAYNFFWATHNLYIILYILCLIHGLARLTGAPRFWLFFIGPGTIYILDKVVSLRTKYIPLDVIETELLPSDVIKIKFYRPPNLKYLSGQWVRLA
HEFHSFTLTSAPHENFLSCHIKAQGPWTWKLRNYFDPSNYNPEDQPKILLEGPFGGGNQDWYKFEVAVMVGGGIGVTPYASILNDLVFGTSTNRYSGVACKKVYFLWICPSHKHFEWFIDVLRDVEKKDVTNVLEIHIFITQFFHKFDLRTTMLYICENHFQRLSKTSIFTGLKAVNHFGRPDMTSFLKFVQKRHSYVSKIGVF
LTKSVMSACDDVNKGRKLPYFIHHFENFG
>7091:001565 "pub_gene_id":"BGIBMGA005478", "pub_og_id":"EOG090R008X", "og_name":"Similarity:Contains FAD-binding FR-type domain.","level":33392
MAGPERPGARTLSKIFMRGQDGLPSLTNRTALLAFFGQVVTGEIVMASESGCPIEHHRIPVDKCDHMYDSECQGAKYMPFLRAAYDRNTGQSPNSPREQINQMTSWIDGSFVYSTSEAWVNAMRSFQNGSFASDGGMPLRNTKRVPLFNNPVPHYMRMLSPERLFLLGDPRTNQNPAMVSFGILLFRWHNVVAARVHKQHPDWS
RARRIVIASLQNIILYEYVPAFLGAALPPYEGYKPEVAPGVTHEFAVAAFRFGHTLVPPAILLRNRKCGYGRAPGGHDAIRLCQTWWDGNDVMSQVPLEEVLMGMASQLSEREDALLCSDVRDNLFGPMEFSRRDLGALNIMRGRDNGVPDYNTVREYFGLRKIKTFNEINPDLFEKNPDLLQKFIQIYNGDVDNIDVYIGGML
GELFRAIIIDQFTRIRDSDRFWFENDKNGIFTPRDIEELRRITLWDIIVNSTTIGAKDIQRDVFHWRAGDPCPQPYQLNASKLPPCKYLKGYDYFEGNEFAYIYICLILVFVPILCAGAGYGVVKLQNSRRRRLKIQQEHLKNAQCKGSVDKMVCREWVHASHKRLVKLRLGPESALHVTDRKGEKLRTLPLDHTDQLTVVESQ
RPLVLVRVPREHDLVLEMDSIASRRKFLVKLDTFLAQHKKALNLTQGPREQILATAETRERRQRKLEHFFREAYAITFGLAPGEKRRRSEDADPESIVMRTSLSKSEFASALGMKGDAVFVKKMFNIVDKDGDGRISFQEFLDTVVLFSRGATEDKLRIIFDMCDNDRNGVIDKGELSEMLRSLVEIARTTSLRDEHVTELIDG
LQHKDHLTYSDFKLMMKEYKGEFVAIGLDCKGAKQNFLDTSTNVARMTSFHIEPSMEQTRHWLLLKWDTLTTFLEENRQNIFYLFIFYVVTIGLFVERFAHYSFMSEHLDLRHIMGVGIAITRGSAASLSFCYSLLLLTMSRNLLTKLKEFSIQQYIPLDSSIQFHKIVACTALFFSLLHTAGHMVNFYHVSTQPVENLRCLTK
DFRPGITYWVFQTVTGVSGVLLFVIMCIIFVFAHPVIRKRAYPWFWRAHSLHVALYALCLVHGLARLTGAPRFWIFFLGPAIIYTLDKVVSLRTEYLALDVLETEMLPSDVIKIKFYRPPNLKYLSGQWVRLSCTAFKKEEFHSFTLTSAPHENFLSCHIKAQGPWTWKLRNYFDPCNYNTDDHPKIRIQGPFGGGNQDWYKFE
GGIGVTPYASILNDLVFGTSTNRYSGVACKKVYFLWICPSHKHFEWFIDVLRDVERKDVTNVLEIHIFITQFFHKFDLRTTMLYICENHFQRLSRTSMFTGLKAVNHFGRPDMSSFLKFVQKKHSYVSKIGVFSCGPRPLTKSVMSACEEVNRTRRLPYFIHHFENFG
>7165:002379 {"pub_gene_id":"DUOX", "pub_og_id":"EOG090R008X", "og_name":"Similarity:Contains FAD-binding FR-type domain.","level":33392, "description":"Similarity:Contains FAD-binding FR-type domain."}
MSHVEKQRYDGWYNNLAHPDWGAVDNHLTRKAPSAYSDGVYVMAGSNRPSPRKLSRLFMRGTDGLPSMENRTALLAFFGQVVTNEIVMASESGCPIEMHRIEIEKCDEMYDRECRGDRYIPFHRAAYDRNTGQSPNAPREQINQMTAWIDGSFIYSTSEAWLNAMRSFQDGALLTDKQGTMPVKNTMRVPLFNNPVPHVMRMLS
LGDPRTNQNPALLSFAILFLRWHNVVAKRVRRQHRDWSDEEIFQRARRVVIASLQNIVAYEYLPAFLDKEIPPYDGYKADTHPGVSHMFQAAAFRFGHSLIPPGLFRRDGQCNFRRTNMDFPALRLCSTWWNSNDVLDNTPVEEFIMGMASQIAEKEDPLLCSDVRDKLFGPMEFTRRDLGALNIMRGRDNGLPDYNTARAAYR
WRDINPAVFERQPELLDLLIKTYDNQLDNVDVYVGGMLESDGRPGELFSAVIIDQFTRIRDADRFWFENEDNGIFTKEEIAEIRKFTLWDIIVNSTDIEADEIQRDVFHWKQGDPCPQPEQLNATLLEPCNYLEGYDYFSGSELAYIYSCVFLGFVPILCAGAGYCVIKLQNSRRRKLKIKQEAMKNTANTKVSVEKMVAREWL
LVTVKFGPEASIYTVDRKGEKLRTFNLKHVDVVTVEQSQENYTAKKPYILLRVPNDHDLVLELESNSARRKFVKKLEDFLVLHKKTMTFVESNRDLMLAKAETRERRQKRLEHFFREAYALTFGLRPGERRRRSDASLDGEVMTVMRTSLSKSEFAAALGMKQDDMFVRKMFNIVDKDKDGRISFQEFLETVVLFSRGKTDDKL
CDNDRNGVIDKGELSEMMRSLVEIARTTSVTDEQVNELIDGMFQDVGLEHKNHLTYEDFKLMMKEYKGDFVAIGLDCKGAKQNFLDTSTNVARMTSFHIEPISDSRRHWMQEKWDCYTTFLEENRQNIFYLFLFYVITIVLFVERFIHYSFMAEHTDLRHIMGVGIAITRGSAASLSFCYSLLLLTMSRNLLTKLKEFPIQQYI
QFHKIAACTALFFSLLHTVGHIVNFYHVSTQSIENLKCLTKEVHFTSDYRPDITYWLFQTITGVTGVMLFVTMCIIFAFAHPTIRKKAYKFFWNAHSLYVVLYALCLVHGLARLTGAPRFWLFFIGPGIVYTLDKIVSLRTKYMALDVIETDLLPSDVIKIKFYRPPNLKYLSGQWVRLSCTEIKPEEMHSFTLTSAPHENFLS
GPWTWKLRNYFDPCNYNPDDQPKIRIEGPFGGGNQDWYKFEVAVMVGGGIGVTPYASILNDLVFGTSTNRYSGVACKKVYFLWICPSHKHFEWFIDVLRDVEKKDVTNVLEIHIFITQFFHKFDLRTTMLYICENHFQRLSKTSMFTGLKAVNHFGRPDMSSFLKFVQKKHSYVSKIGVFSCGPRPLTKSVMSACDEVNKSRKW
FENFG
>7227:000fd5 {"pub_gene_id":"Duox", "pub_og_id":"EOG090R008X", "og_name":"Similarity:Contains FAD-binding FR-type domain.","level":33392, "description":"Similarity:Contains FAD-binding FR-type domain."}
MSVPSAPHQRAESKNRVPRPGQKNRKLPKLRLHWPGATYGGALLLLLISYGLELGSVHCYEKMYSQTEKQRYDGWYNNLAHPDWGSVDSHLVRKAPPSYSDGVYAMAGANRPSTRRLSRLFMRGKDGLGSKFNRTALLAFFGQLVANEIVMASESGCPIEMHRIEIEKCDEMYDRECRGDKYIPFHRAAYDRDTGQSPNAPREQ
WIDGSFIYSTSEAWLNAMRSFHNGTLLTEKDGKLPVRNTMRVPLFNNPVPSVMKMLSPERLFLLGDPRTNQNPAILSFAILFLRWHNTLAQRIKRVHPDWSDEDIYQRARHTVIASLQNVIVYEYLPAFLGTSLPPYEGYKQDIHPGIGHIFQAAAFRFGHTMIPPGIYRRDGQCNFKETPMGYPAVRLCSTWWDSSGFFADTS
GLASQISEREDPVLCSDVRDKLFGPMEFTRRDLGALNIMRGRDNGLPDYNTARESYGLKRHKTWTDINPPLFETQPELLDMLKEAYDNKLDDVDVYVGGMLESYGQPGEFFTAVIKEQFQRLRDADRFWFENERNGIFTPEEIAELRKITLWDIIVNSTDVKEEEIQKDVFMWRTGDPCPQPMQLNATELEPCTYLEGYDYFSG
YVCVFLGFVPILCAGAGYCVVKLQNSKRRRLKIRQEALRAPQHKGSVDKMLAREWLHANHKRLVTVKFGPEAAIYTVDRKGEKLRTFSLKHIDVVSVEESATNHIKKKPYILLRVPSDHDLVLELESYGARRKFVKKLEDFLLLHKKEMTLMEVNRDIMLARAETRERRQKRLEYFFREAYALTFGLRPGERRRRSDASSDGEV
SLSKAEFAAALGMKPNDMFVRKMFNIVDKDQDGRISFQEFLETVVLFSRGKTDDKLRIIFDMCDNDRNGVIDKGELSEMMRSLVEIARTTSLGDDQVTELIDGMFQDVGLEHKNHLTYQDFKLMMKEYKGDFVAIGLDCKGAKQNFLDTSTNVARMTSFNIEPMQDKPRHWLLAKWDAYITFLEENRQNIFYLFLFYVVTIVLF
YSFMAEHTDLRHIMGVGIAITRGSAASLSFCYSLLLLTMSRNLITKLKEFPIQQYIPLDSHIQFHKIAACTALFFSVLHTVGHIVNFYHVSTQSHENLRCLTREVHFASDYKPDITFWLFQTVTGTTGVMLFIIMCIIFVFAHPTIRKKAYNFFWNMHTLYIGLYLLSLIHGLARLTGPPRFWMFFLGPGIVYTLDKIVSLRTK
IDTDLLPSDVIKIKFYRPPNLKYLSGQWVRLSCTAFRPHEMHSFTLTSAPHENFLSCHIKAQGPWTWKLRNYFDPCNYNPEDQPKIRIEGPFGGGNQDWYKFEVAVMVGGGIGVTPYASILNDLVFGTSTNRYSGVACKKVYFLWICPSHKHFEWFIDVLRDVEKKDVTNVLEIHIFITQFFHKFDLRTTMLYICENHFQRLSK
LKAVNHFGRPDMSSFLKFVQKKHSYVSKIGVFSCGPRPLTKSVMSACDEVNKTRKLPYFIHHFENFG
>7460:002ad8 {"pub_gene_id":"GB51481", "pub_og_id":"EOG090R008X", "og_name":"Similarity:Contains FAD-binding FR-type domain.","level":33392, "description":"Uncharacterized protein  "}
MTRRRPPRSDSNWIYLLILLLWLVLPIKTGVVHSYADKQRYDGWYNNLAHPDWGSIDSRLIRKMPAAYSDGVYMLAGQDRPSPRKLSQLFMQGDDGLPSVKNRTALFAFFGQLVTSEIIMASESGCPIEYHRIDVDKCDPVFDKECQGNKYIPFRRADYDRQTGRSPNSPREQINKVTSWIDGSFVYSSSEAWANTMRSFKNGS
RKFPVRNTMRAPLFNHAVPHVMRMLSPERLYLLGDPRTNQHPPLLALGILFYRWHNVIAARIQLENPTMSDEDIFQKARRVVIGTLQNIILYEYIPILLNEDLPPYTGYKSDLHPGISHIFQSAAFRFGHTLIPPGLYRRDENCEYRRTNTDQPAIRLCSTWWDSNEVLTNSTIEELLMGMTSQIAEKEDNLLGTDIRNNLFGP
DLGALNIMRGRDNGLPDYNTARAHFKLPRKKTWNEINPELFNKNPSLLRTLVEIHSNNLNNMDVYVGGMLESSAGPGELFSTVIKEQFLRLRDSDRFWFENEENGIFTRSEIADIRRITLWDVIVNATGIPADSIQRKVFTWEEGDPCPQPYQLNSTMLEPCVPLQRYDYFEGSELVYIYACVFLGFVPILCAGAGYGLVKLQN
```

Using OrthoDB in your research

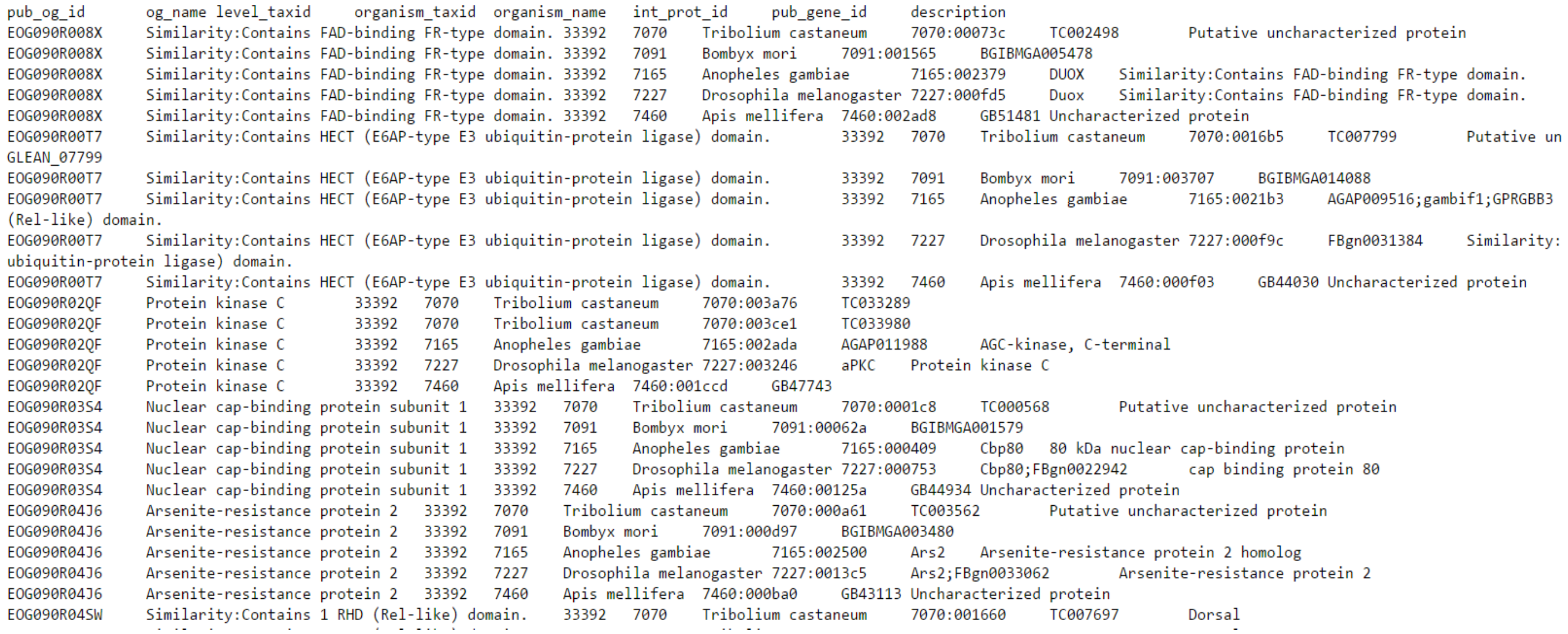


OrthoDB

Your search for **immunity** at Endopterygota level returned 10 groups

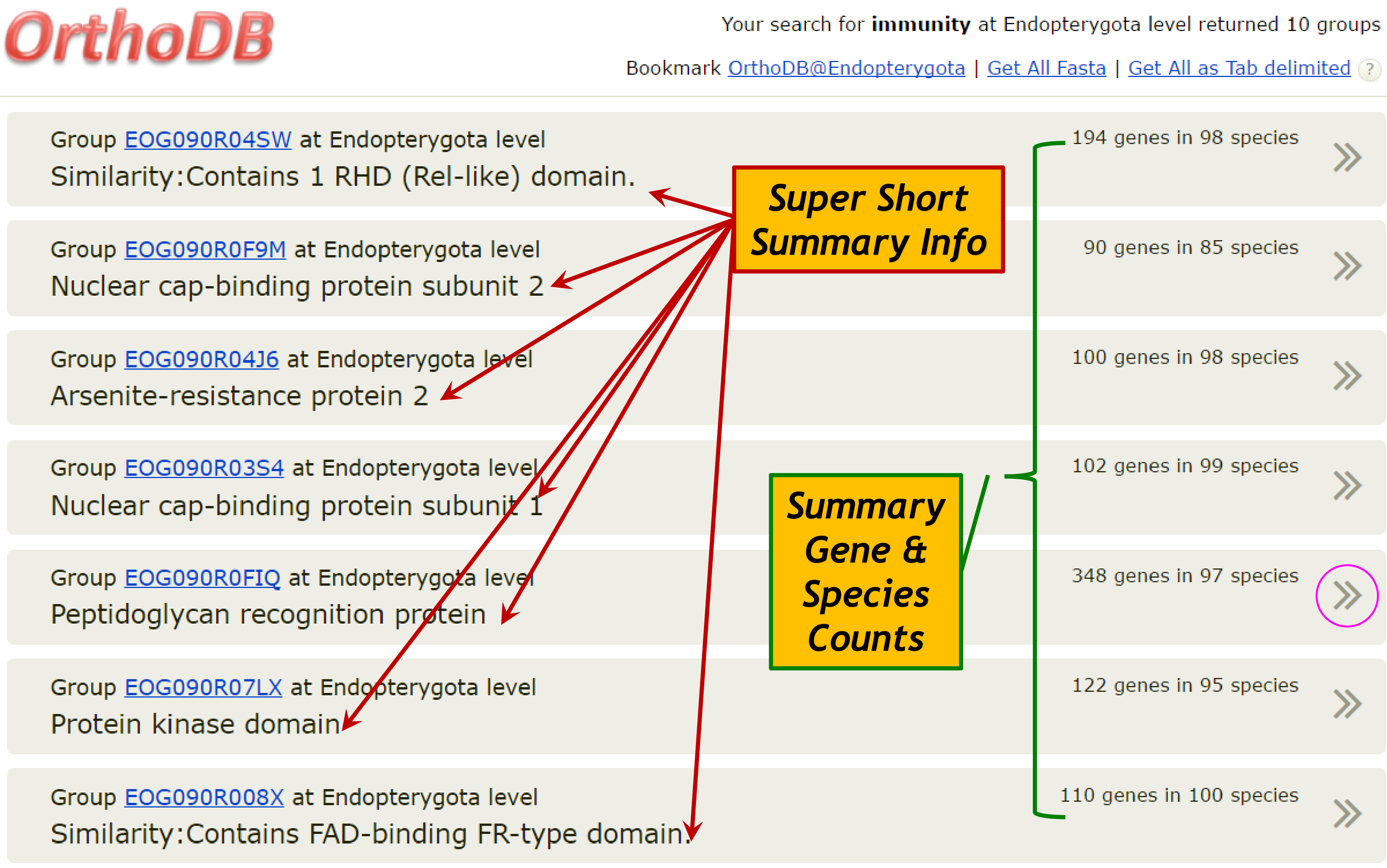
Bookmark OrthoDB@Endopterygota | Get All Fasta | Get All as Tab delimited ?

3. Get ALL gene/protein information of the genes from the selected species for ALL 10 of the search result orthologous groups



```
pub_og_id       og_name level_taxid     organism_taxid  organism_name   int_prot_id     pub_gene_id     description
EOG090R008X     Similarity:Contains FAD-binding FR-type domain. 33392   7070    Tribolium castaneum     7070:00073c     TC002498        Putative uncharacterized protein
EOG090R008X     Similarity:Contains FAD-binding FR-type domain. 33392   7091    Bombyx mori     7091:001565     BGIBMGA005478
EOG090R008X     Similarity:Contains FAD-binding FR-type domain. 33392   7165    Anopheles gambiae       7165:002379     DUOX    Similarity:Contains FAD-binding FR-type domain.
EOG090R008X     Similarity:Contains FAD-binding FR-type domain. 33392   7227    Drosophila melanogaster 7227:000fd5     Duox    Similarity:Contains FAD-binding FR-type domain.
EOG090R008X     Similarity:Contains FAD-binding FR-type domain. 33392   7460    Apis mellifera  7460:002ad8     GB51481 Uncharacterized protein
EOG090R00T7     Similarity:Contains HECT (E6AP-type E3 ubiquitin-protein ligase) domain.        33392   7070    Tribolium castaneum     7070:0016b5     TC007799        Putative un
GLEAN_07799
EOG090R00T7     Similarity:Contains HECT (E6AP-type E3 ubiquitin-protein ligase) domain.        33392   7091    Bombyx mori     7091:003707     BGIBMGA014088
EOG090R00T7     Similarity:Contains HECT (E6AP-type E3 ubiquitin-protein ligase) domain.        33392   7165    Anopheles gambiae       7165:0021b3     AGAP009516;gambif1;GPRGBB3
(Rel-like) domain.
EOG090R00T7     Similarity:Contains HECT (E6AP-type E3 ubiquitin-protein ligase) domain.        33392   7227    Drosophila melanogaster 7227:000f9c     FBgn0031384     Similarity:
ubiquitin-protein ligase) domain.
EOG090R00T7     Similarity:Contains HECT (E6AP-type E3 ubiquitin-protein ligase) domain.        33392   7460    Apis mellifera  7460:000f03     GB44030 Uncharacterized protein
EOG090R02QF     Protein kinase C        33392   7070    Tribolium castaneum     7070:003a76     TC033289
EOG090R02QF     Protein kinase C        33392   7070    Tribolium castaneum     7070:003ce1     TC033980
EOG090R02QF     Protein kinase C        33392   7165    Anopheles gambiae       7165:002ada     AGAP011988      AGC-kinase, C-terminal
EOG090R02QF     Protein kinase C        33392   7227    Drosophila melanogaster 7227:003246     aPKC    Protein kinase C
EOG090R02QF     Protein kinase C        33392   7460    Apis mellifera  7460:001ccd     GB47743
EOG090R03S4     Nuclear cap-binding protein subunit 1   33392   7070    Tribolium castaneum     7070:0001c8     TC000568        Putative uncharacterized protein
EOG090R03S4     Nuclear cap-binding protein subunit 1   33392   7091    Bombyx mori     7091:00062a     BGIBMGA001579
EOG090R03S4     Nuclear cap-binding protein subunit 1   33392   7165    Anopheles gambiae       7165:000409     Cbp80   80 kDa nuclear cap-binding protein
EOG090R03S4     Nuclear cap-binding protein subunit 1   33392   7227    Drosophila melanogaster 7227:000753     Cbp80;FBgn0022942       cap binding protein 80
EOG090R03S4     Nuclear cap-binding protein subunit 1   33392   7460    Apis mellifera  7460:00125a     GB44934 Uncharacterized protein
EOG090R04J6     Arsenite-resistance protein 2   33392   7070    Tribolium castaneum     7070:000a61     TC003562        Putative uncharacterized protein
EOG090R04J6     Arsenite-resistance protein 2   33392   7091    Bombyx mori     7091:000d97     BGIBMGA003480
EOG090R04J6     Arsenite-resistance protein 2   33392   7165    Anopheles gambiae       7165:002500     Ars2    Arsenite-resistance protein 2 homolog
EOG090R04J6     Arsenite-resistance protein 2   33392   7227    Drosophila melanogaster 7227:0013c5     Ars2;FBgn0033062        Arsenite-resistance protein 2
EOG090R04J6     Arsenite-resistance protein 2   33392   7460    Apis mellifera  7460:000ba0     GB43113 Uncharacterized protein
EOG090R04SW     Similarity:Contains 1 RHD (Rel-like) domain.    33392   7070    Tribolium castaneum     7070:001660     TC007697        Dorsal
```

Using OrthoDB in your research



OrthoDB

Your search for **immunity** at Endopterygota level returned 10 groups

Bookmark OrthoDB@Endopterygota | Get All Fasta | Get All as Tab delimited ?

Group EOG090R04SW at Endopterygota level
Similarity:Contains 1 RHD (Rel-like) domain.
194 genes in 98 species »

Group EOG090R0F9M at Endopterygota level
Nuclear cap-binding protein subunit 2
90 genes in 85 species »

Group EOG090R04J6 at Endopterygota level
Arsenite-resistance protein 2
100 genes in 98 species »

Group EOG090R03S4 at Endopterygota level
Nuclear cap-binding protein subunit 1
102 genes in 99 species »

Group EOG090R0FIQ at Endopterygota level
Peptidoglycan recognition protein
348 genes in 97 species »

Group EOG090R07LX at Endopterygota level
Protein kinase domain
122 genes in 95 species »

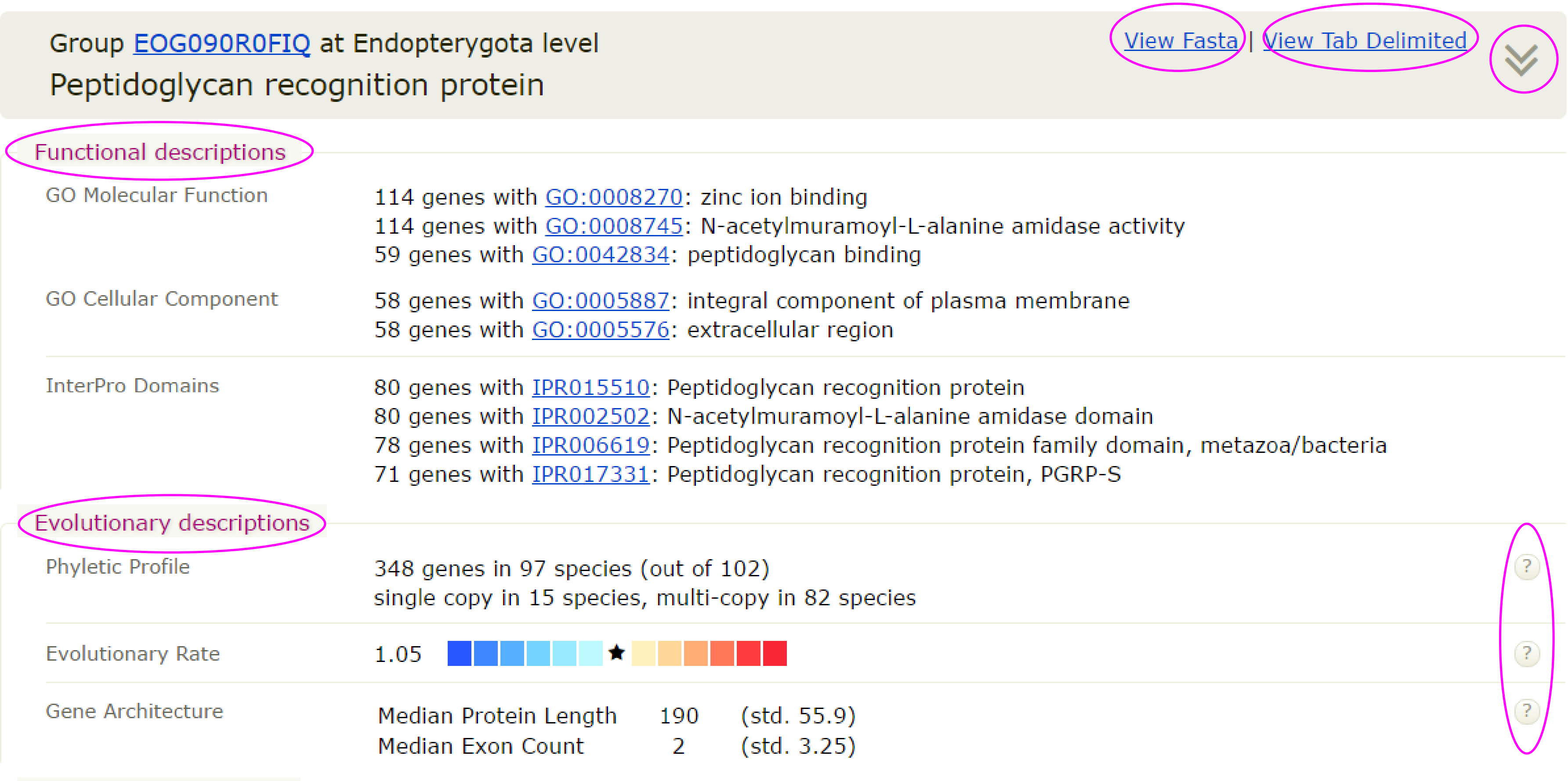
Group EOG090R008X at Endopterygota level
Similarity:Contains FAD-binding FR-type domain.
110 genes in 100 species »

Super Short Summary Info

Summary Gene & Species Counts

Using *OrthoDB* in your research

Expanded PGRP orthologous group



Group EOG090R0FIQ at Endopterygota level
Peptidoglycan recognition protein

View Fasta | View Tab Delimited

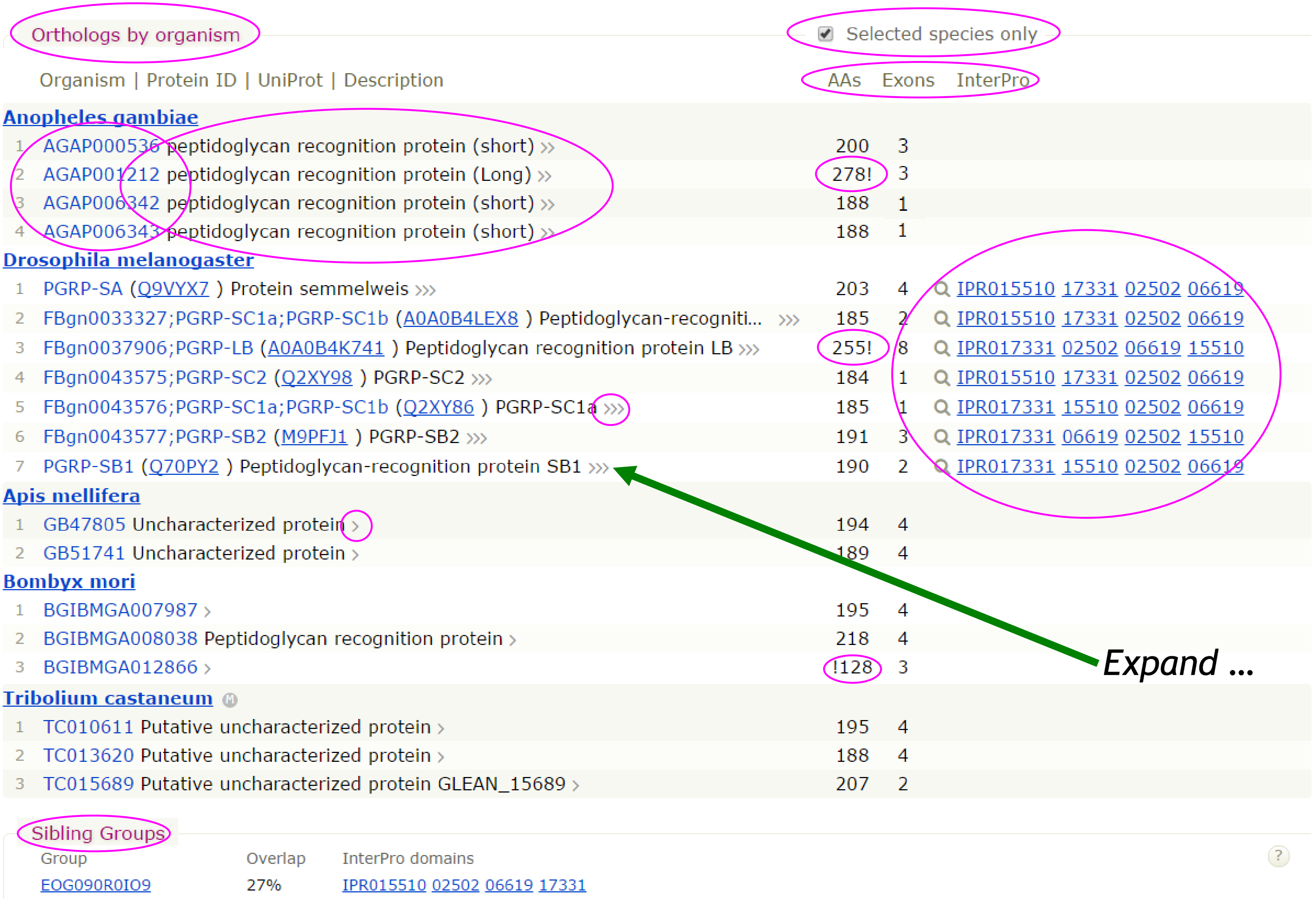
Functional descriptions

GO Molecular Function	114 genes with GO:0008270: zinc ion binding 114 genes with GO:0008745: N-acetylmuramoyl-L-alanine amidase activity 59 genes with GO:0042834: peptidoglycan binding
GO Cellular Component	58 genes with GO:0005887: integral component of plasma membrane 58 genes with GO:0005576: extracellular region
InterPro Domains	80 genes with IPR015510: Peptidoglycan recognition protein 80 genes with IPR002502: N-acetylmuramoyl-L-alanine amidase domain 78 genes with IPR006619: Peptidoglycan recognition protein family domain, metazoa/bacteria 71 genes with IPR017331: Peptidoglycan recognition protein, PGRP-S

Evolutionary descriptions

Phyletic Profile	348 genes in 97 species (out of 102) single copy in 15 species, multi-copy in 82 species
Evolutionary Rate	1.05 ★
Gene Architecture	Median Protein Length 190 (std. 55.9) Median Exon Count 2 (std. 3.25)

Using OrthoDB in your research



Orthologs by organism

☑ Selected species only

	Organism | Protein ID | UniProt | Description	AAs	Exons	InterPro
	Anopheles gambiae			
1	AGAP000536 peptidoglycan recognition protein (short) »	200	3	
2	AGAP001212 peptidoglycan recognition protein (Long) »	278!	3	
3	AGAP006342 peptidoglycan recognition protein (short) »	188	1	
4	AGAP006343 peptidoglycan recognition protein (short) »	188	1	
	Drosophila melanogaster			
1	PGRP-SA (Q9VYX7) Protein semmelweis ›››	203	4	IPR015510 17331 02502 06619
2	FBgn0033327;PGRP-SC1a;PGRP-SC1b (A0A0B4LEX8) Peptidoglycan-recogniti... ›››	185	2	IPR015510 17331 02502 06619
3	FBgn0037906;PGRP-LB (A0A0B4K741) Peptidoglycan recognition protein LB ›››	255!	8	IPR017331 02502 06619 15510
4	FBgn0043575;PGRP-SC2 (Q2XY98) PGRP-SC2 ›››	184	1	IPR015510 17331 02502 06619
5	FBgn0043576;PGRP-SC1a;PGRP-SC1b (Q2XY86) PGRP-SC1a ›››	185	1	IPR017331 15510 02502 06619
6	FBgn0043577;PGRP-SB2 (M9PFJ1) PGRP-SB2 ›››	191	3	IPR017331 06619 02502 15510
7	PGRP-SB1 (Q70PY2) Peptidoglycan-recognition protein SB1 ›››	190	2	IPR017331 15510 02502 06619
	Apis mellifera			
1	GB47805 Uncharacterized protein ›	194	4	
2	GB51741 Uncharacterized protein ›	189	4	
	Bombyx mori			
1	BGIBMGA007987 ›	195	4	
2	BGIBMGA008038 Peptidoglycan recognition protein ›	218	4	
3	BGIBMGA012866 ›	!128	3	
	Tribolium castaneum			
1	TC010611 Putative uncharacterized protein ›	195	4	
2	TC013620 Putative uncharacterized protein ›	188	4	
3	TC015689 Putative uncharacterized protein GLEAN_15689 ›	207	2	

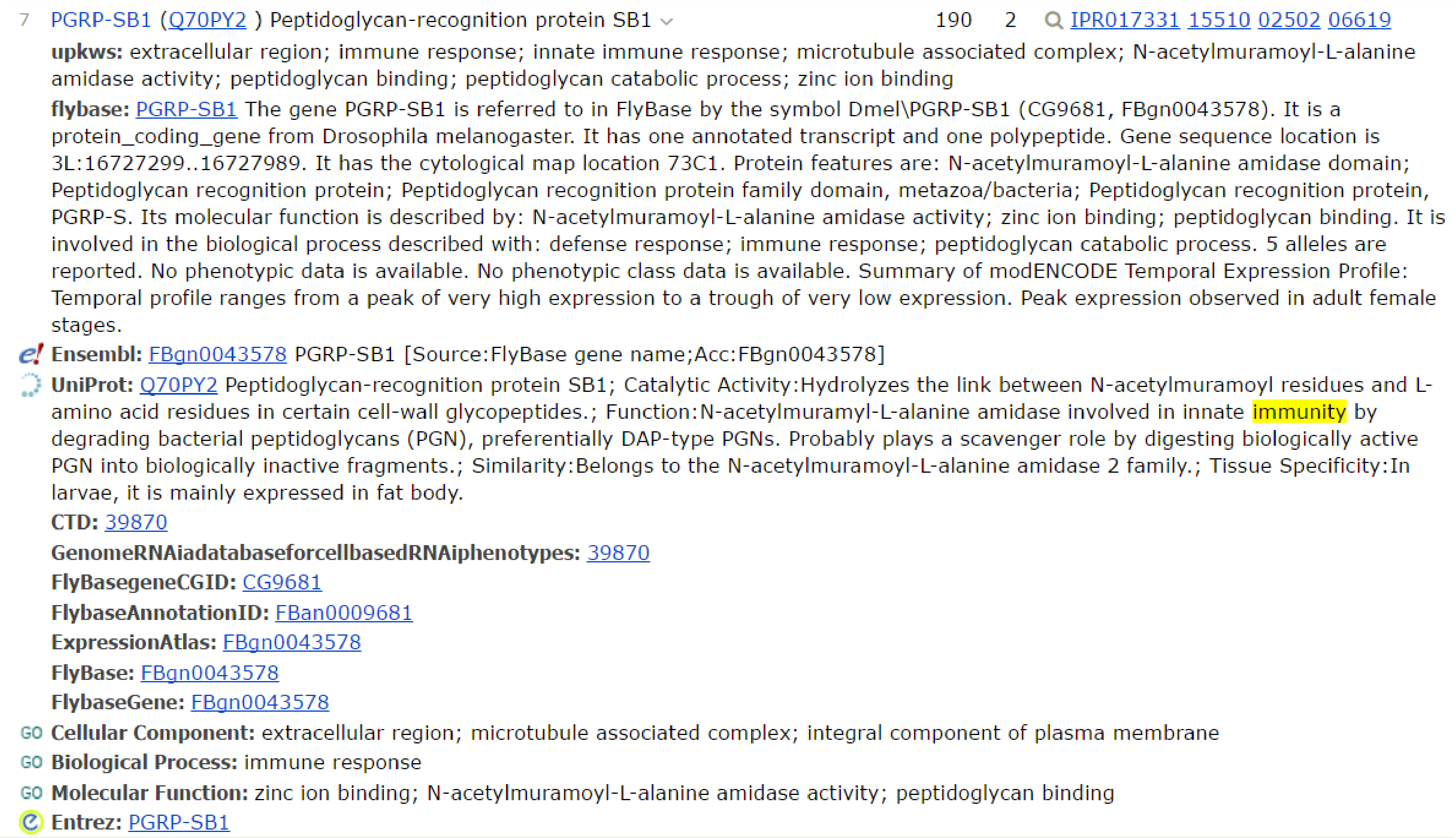
Expand ...

Sibling Groups

Group	Overlap	InterPro domains
EOG090R0IO9	27%	IPR015510 02502 06619 17331

Using OrthoDB in your research

Expanded gene annotation (incl. search term)



7 PGRP-SB1 (Q70PY2) Peptidoglycan-recognition protein SB1 190 2 IPR017331 15510 02502 06619

upkws: extracellular region; immune response; innate immune response; microtubule associated complex; N-acetylmuramoyl-L-alanine amidase activity; peptidoglycan binding; peptidoglycan catabolic process; zinc ion binding

flybase: PGRP-SB1 The gene PGRP-SB1 is referred to in FlyBase by the symbol Dmel\PGRP-SB1 (CG9681, FBgn0043578). It is a protein_coding_gene from Drosophila melanogaster. It has one annotated transcript and one polypeptide. Gene sequence location is 3L:16727299..16727989. It has the cytological map location 73C1. Protein features are: N-acetylmuramoyl-L-alanine amidase domain; Peptidoglycan recognition protein; Peptidoglycan recognition protein family domain, metazoa/bacteria; Peptidoglycan recognition protein, PGRP-S. Its molecular function is described by: N-acetylmuramoyl-L-alanine amidase activity; zinc ion binding; peptidoglycan binding. It is involved in the biological process described with: defense response; immune response; peptidoglycan catabolic process. 5 alleles are reported. No phenotypic data is available. No phenotypic class data is available. Summary of modENCODE Temporal Expression Profile: Temporal profile ranges from a peak of very high expression to a trough of very low expression. Peak expression observed in adult female stages.

Ensembl: FBgn0043578 PGRP-SB1 [Source:FlyBase gene name;Acc:FBgn0043578]

UniProt: Q70PY2 Peptidoglycan-recognition protein SB1; Catalytic Activity:Hydrolyzes the link between N-acetylmuramoyl residues and L-amino acid residues in certain cell-wall glycopeptides.; Function:N-acetylmuramyl-L-alanine amidase involved in innate immunity by degrading bacterial peptidoglycans (PGN), preferentially DAP-type PGNs. Probably plays a scavenger role by digesting biologically active PGN into biologically inactive fragments.; Similarity:Belongs to the N-acetylmuramoyl-L-alanine amidase 2 family.; Tissue Specificity:In larvae, it is mainly expressed in fat body.

CTD: 39870

GenomeRNAiadatabaseforcellbasedRNAiphenotypes: 39870

FlyBasegeneCGID: CG9681

FlybaseAnnotationID: FBan0009681

ExpressionAtlas: FBgn0043578

FlyBase: FBgn0043578

FlybaseGene: FBgn0043578

GO **Cellular Component:** extracellular region; microtubule associated complex; integral component of plasma membrane

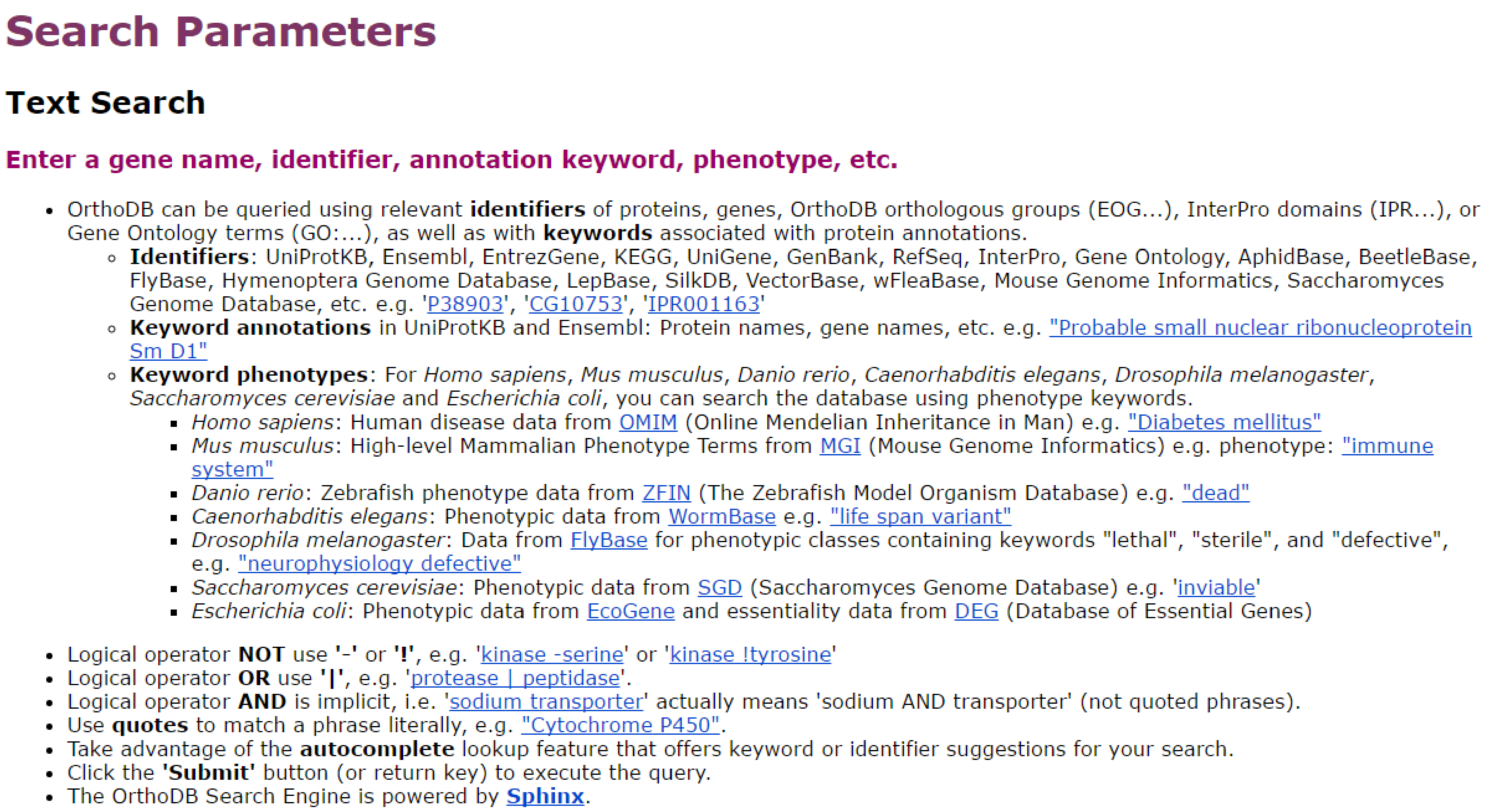
GO **Biological Process:** immune response

GO **Molecular Function:** zinc ion binding; N-acetylmuramoyl-L-alanine amidase activity; peptidoglycan binding

Entrez: PGRP-SB1

Using *OrthoDB* in your research

Use HELP page to learn about OrthoDB features



Search Parameters

Text Search

Enter a gene name, identifier, annotation keyword, phenotype, etc.

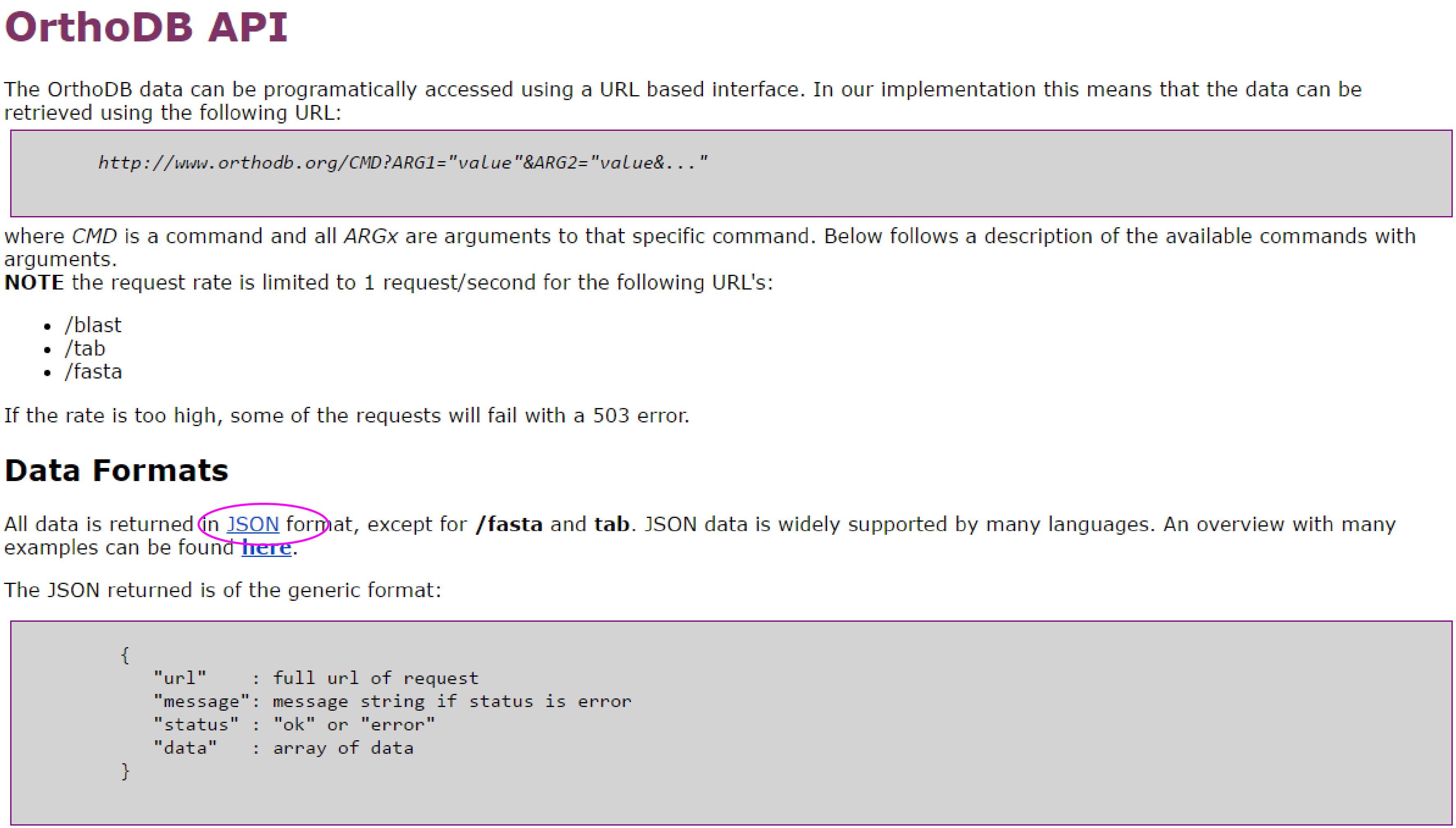
- OrthoDB can be queried using relevant **identifiers** of proteins, genes, OrthoDB orthologous groups (EOG...), InterPro domains (IPR...), or Gene Ontology terms (GO:...), as well as with **keywords** associated with protein annotations.
  - **Identifiers**: UniProtKB, Ensembl, EntrezGene, KEGG, UniGene, GenBank, RefSeq, InterPro, Gene Ontology, AphidBase, BeetleBase, FlyBase, Hymenoptera Genome Database, LepBase, SilkDB, VectorBase, wFleaBase, Mouse Genome Informatics, Saccharomyces Genome Database, etc. e.g. 'P38903', 'CG10753', 'IPR001163'
  - **Keyword annotations** in UniProtKB and Ensembl: Protein names, gene names, etc. e.g. "Probable small nuclear ribonucleoprotein Sm D1"
  - **Keyword phenotypes**: For *Homo sapiens*, *Mus musculus*, *Danio rerio*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Escherichia coli*, you can search the database using phenotype keywords.
    - *Homo sapiens*: Human disease data from OMIM (Online Mendelian Inheritance in Man) e.g. "Diabetes mellitus"
    - *Mus musculus*: High-level Mammalian Phenotype Terms from MGI (Mouse Genome Informatics) e.g. phenotype: "immune system"
    - *Danio rerio*: Zebrafish phenotype data from ZFIN (The Zebrafish Model Organism Database) e.g. "dead"
    - *Caenorhabditis elegans*: Phenotypic data from WormBase e.g. "life span variant"
    - *Drosophila melanogaster*: Data from FlyBase for phenotypic classes containing keywords "lethal", "sterile", and "defective", e.g. "neurophysiology defective"
    - *Saccharomyces cerevisiae*: Phenotypic data from SGD (Saccharomyces Genome Database) e.g. 'inviable'
    - *Escherichia coli*: Phenotypic data from EcoGene and essentiality data from DEG (Database of Essential Genes)

- Logical operator **NOT** use '-' or '**!**', e.g. 'kinase -serine' or 'kinase !tyrosine'
- Logical operator **OR** use '**|**', e.g. 'protease | peptidase'.
- Logical operator **AND** is implicit, i.e. 'sodium transporter' actually means 'sodium AND transporter' (not quoted phrases).
- Use **quotes** to match a phrase literally, e.g. "Cytochrome P450".
- Take advantage of the **autocomplete** lookup feature that offers keyword or identifier suggestions for your search.
- Click the **'Submit'** button (or return key) to execute the query.
- The OrthoDB Search Engine is powered by **Sphinx**.

Using OrthoDB in your research

www.orthodb.org/?page=api

Programmatic data access: using the API



OrthoDB API

The OrthoDB data can be programatically accessed using a URL based interface. In our implementation this means that the data can be retrieved using the following URL:

```
        http://www.orthodb.org/CMD?ARG1="value"&ARG2="value&..."
```

where *CMD* is a command and all *ARGx* are arguments to that specific command. Below follows a description of the available commands with arguments.
NOTE the request rate is limited to 1 request/second for the following URL's:

- /blast
- /tab
- /fasta

If the rate is too high, some of the requests will fail with a 503 error.

Data Formats

All data is returned in JSON format, except for **/fasta** and **tab**. JSON data is widely supported by many languages. An overview with many examples can be found **here**.

The JSON returned is of the generic format:

```
          {
             "url"    : full url of request
             "message": message string if status is error
             "status" : "ok" or "error"
             "data"   : array of data
          }
```

Using OrthoDB in your research

Programmatic data access: using the API

```
wget -O myogs.txt "http://www.orthodb.org/v9.1/search?
level=33392&species=7165,7227,7460,7091,7070
&query=immunity"
```

Web-get into output file ‘myogs.txt’

Use OrthoDB’s search mode

Which level to search and which species to return?

What query term or phrase to search for?

Using OrthoDB in your research

Programmatic data access: using the API

```
wget -O myogs.txt "http://www.orthodb.org/v9.1/search?

level=33392&species=7165,7227,7460,7091,7070

&query=immunity"
```

```
Resolving www.orthodb.org (www.orthodb.org)... 129.194.231.60

Connecting to www.orthodb.org (www.orthodb.org)|129.194.231.60|:80... connected.

HTTP request sent, awaiting response... 200 OK
Length: 429 [application/json]
Saving to: 'myogs.txt'

myogs.txt    100%[=======================================>]     429  --.-KB/s  in 0s

 (52.3 MB/s) - 'myogs.txt' saved [429/429]
```

Using OrthoDB in your research

Programmatic data access: using the API

```
wget -O myogs.txt "http://www.orthodb.org/v9.1/search?
level=33392&species=7165,7227,7460,7091,7070
&query=immunity"
```

```
more myogs.txt
{"status": "ok", "message": null, "data": ["EOG090R04SW",
"EOG090R0F9M", "EOG090R04J6", "EOG090R03S4", "EOG090R0FIQ",
"EOG090R07LX", "EOG090R008X", "EOG090R00T7", "EOG090R02QF",
"EOG090R0AXC"], "count": 10, "skip": 0, "limit": 1000, "query":
"immunity", "level": 33392, "url":
"http://www.orthodb.org/v9.1/search?level=33392&species=7165,7227,
7460,7091,7070&query=immunity", "universal": null, "singlecopy":
null, "inclusive": 1}
```

Using OrthoDB in your research

Programmatic data access: using the API

```
perl -e '@ogs=`cat myogs.txt`=~/(EOG\S{8})/g; foreach
$og (@ogs) { $gp="$og\.txt"; `wget -O $gp
"http://www.orthodb.org/v9.1/tab?id=$og&species=7165,
7227,7460,7091,7070&long=1"`; }'
```

Loop through groups (here using Perl)

Web-get for each group

This time a ‘tab’ search, i.e. get gene annotations

Long option to get sequences as well

Using OrthoDB in your research

Programmatic data access: using the API

```
perl -e '@ogs=`cat myogs.txt`=~/(EOG\S{8})/g; foreach
$og (@ogs) { $gp="$og\.txt"; `wget -O $gp
"http://www.orthodb.org/v9.1/tab?id=$og&species=7165,
7227,7460,7091,7070&long=1"`; }'
```

```
Resolving www.orthodb.org (www.orthodb.org)... 129.194.231.60
Connecting to www.orthodb.org (www.orthodb.org)|129.194.231.60|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 490 [text/html]
Saving to: 'EOG090R02QF.txt'

EOG090R02QF.txt   100%[================================>]     490  --.-KB/s   in 0s

 (48.3 MB/s) - 'EOG090R02QF.txt' saved [490/490]
```

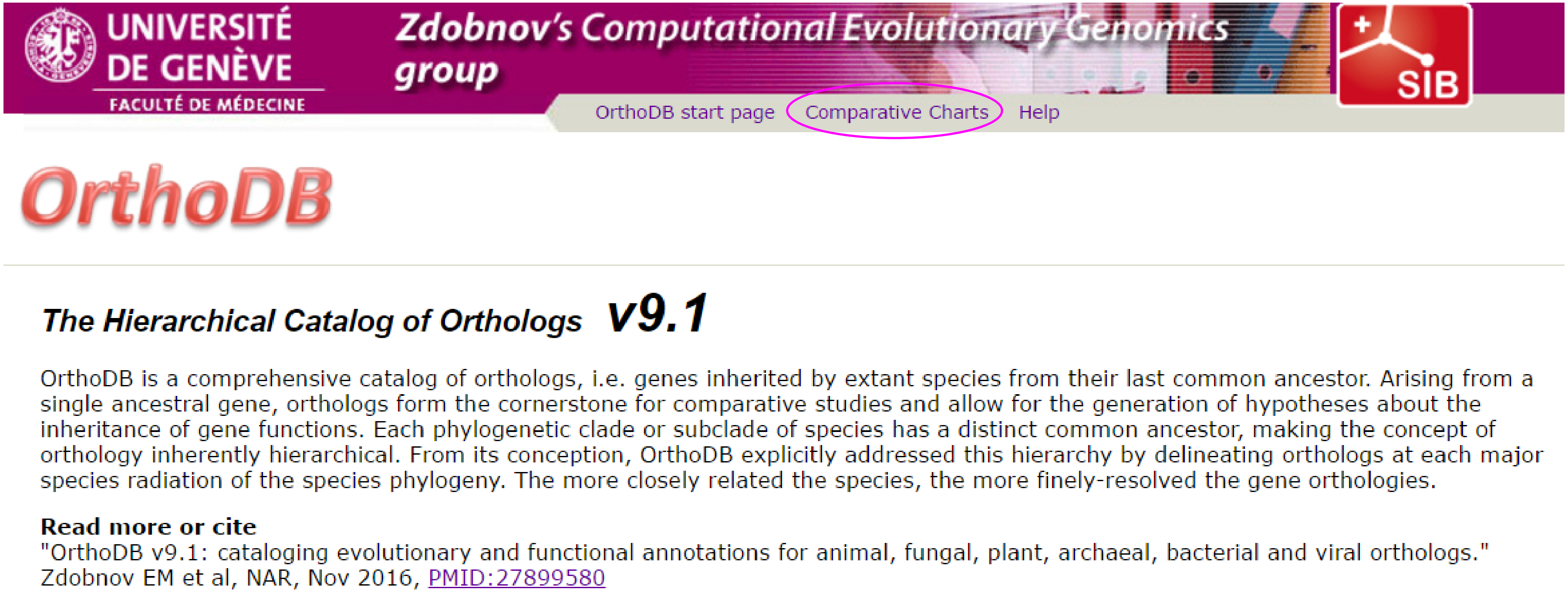
Using OrthoDB in your research

Programmatic data access: using the API

```
more EOG090R0AXC.txt

pub_og_id       og_name level_taxid     organism_taxid  organism_name
int_prot_id     pub_gene_id     description     sequence

EOG090R0AXC     nucleoporin, p88, putative      33392   7070    Tribolium castaneum
7070:002c17     TC014994        Putative uncharacterized protein GLEAN_14994
MDSTDYLGLSKIKILKNVREAVPEKLKKSINLLAVKYGVLFTWDFANNCVLTLNIKAARSNDGDNVTHQNLFPVLPVMFQPELLLVNDT
GTLLLVAGPSGIIVMELPAMHLLYGADSRDVVFCRTHTLDERLLICSDVVQVRQVRFHPGSPRNTHIVALTSDNTLRLYNIENRSAVSV
SKVTIGETPIGVFPGTKTSFLAAFGEVGVDFDFGQPEITKSPTNDETQELQWPVFVLRGDGSVYSVTVPLEPKAKWAVKGPLPQNTPEG
NPRMEACAIICLNTNPEVVCIANSNGTILHSIVLPLDHETRELLCFE

EOG090R0AXC     nucleoporin, p88, putative      33392   7091    Bombyx mori
7091:0033a6     BGIBMGA013223
MTYVAIIKYEYISLVILFCLVENPYEKSKGMIIRSTTYIYIKIFILVEITGRPCMIPTRSYSLDEKFLYTTGEIRRVHWHPISLSHVLV
LVSNNAIRLYNVTLKTGPKLVKTYSIGPKPTSLLAGKTILDSLGDTAVDFTPTPDAEHILILRGDGEIYMMDCDLTNKSPLQPKLVGPL
AIYPPADDNYGSDSCCILCMGGSDIPPLVVIATSSAALYHCLLLPNSEKEESDRDGYALYVVETVELDVVPEPDAEPYPVQLIKCTDDT
YACVHAAGAHTVALPVLAALRHYARAPDGNHPPLGRLYGHTTLTVHSPLCRL
```

Using OrthoDB in your research

Comparative analysis with your own data



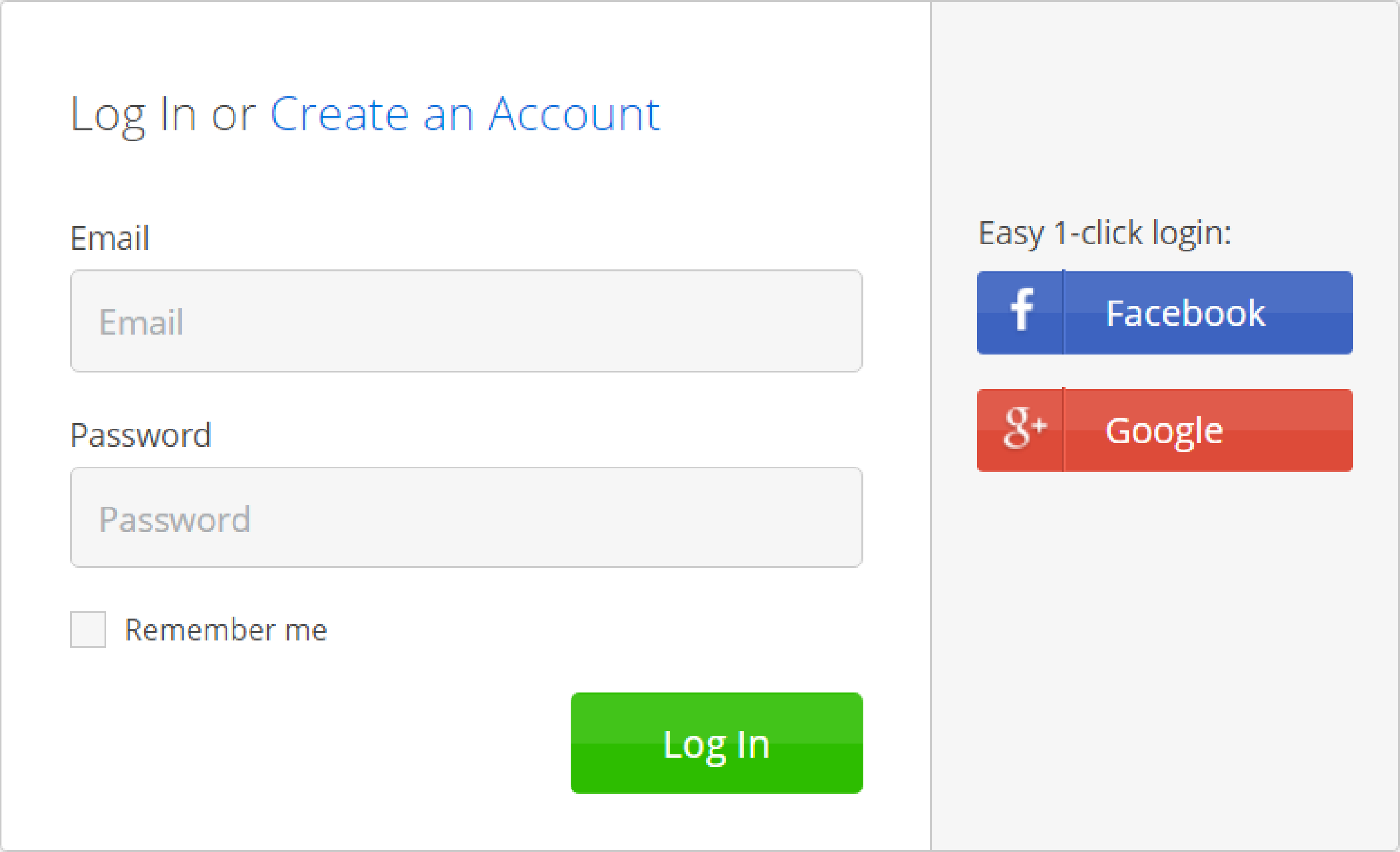
***The Hierarchical Catalog of Orthologs* v9.1**

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

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

Using *OrthoDB* in your research

Comparative analysis with your own data



Log In or Create an Account

Email

Email

Password

Password

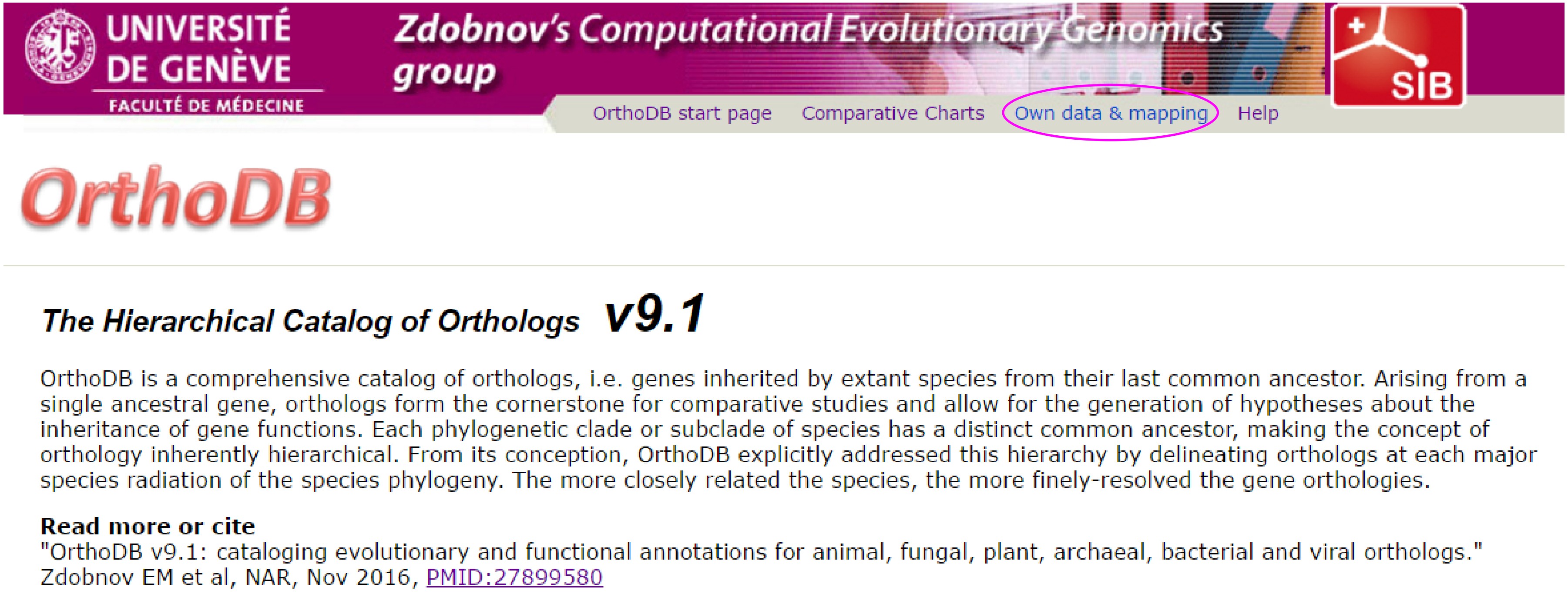
□ Remember me

Easy 1-click login:

Forgot Password?

Using OrthoDB in your research

Comparative analysis with your own data



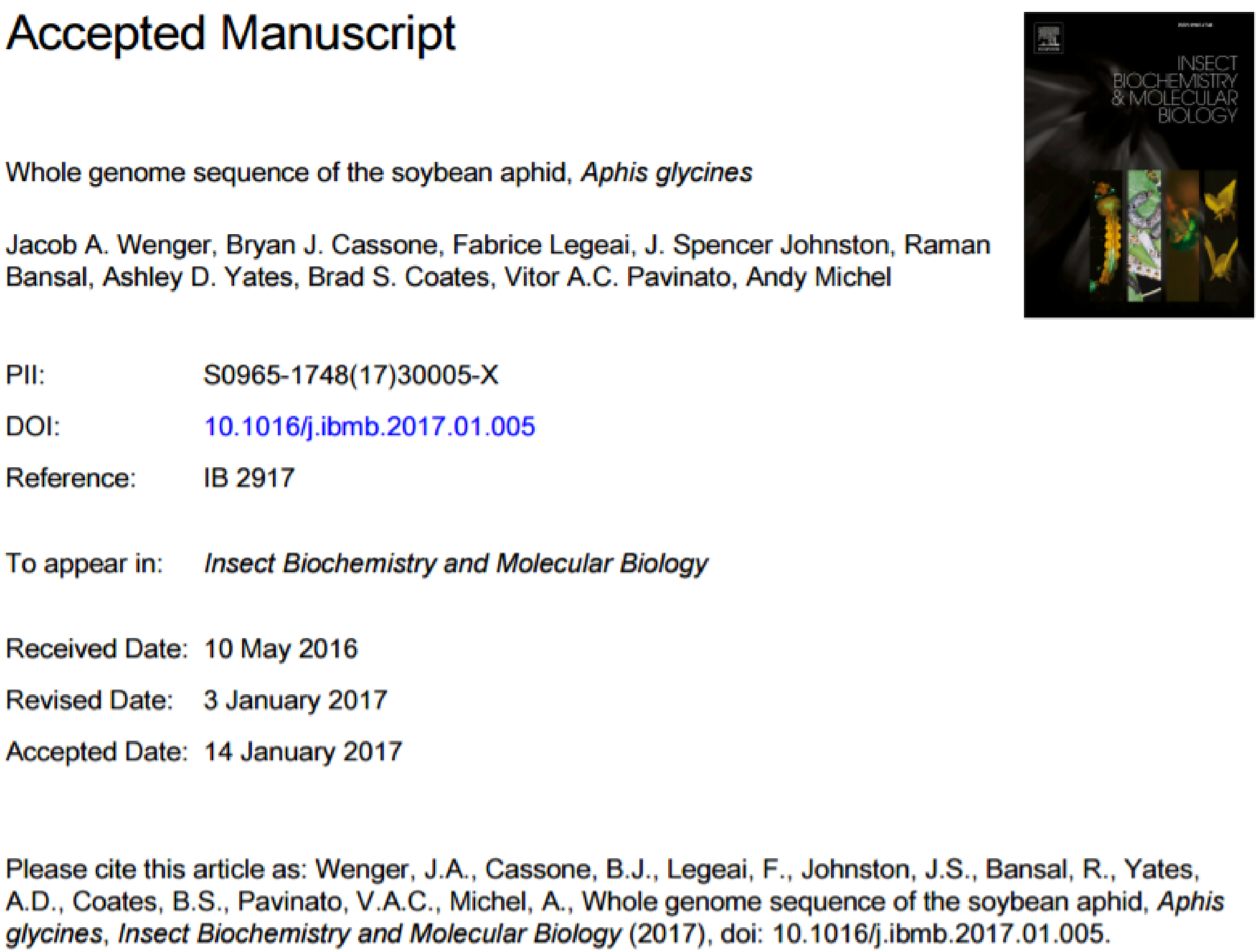
The Hierarchical Catalog of Orthologs **v9.1**

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

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

Using OrthoDB in your research

Example: newly-published aphid genome



Accepted Manuscript

Whole genome sequence of the soybean aphid, *Aphis glycines*

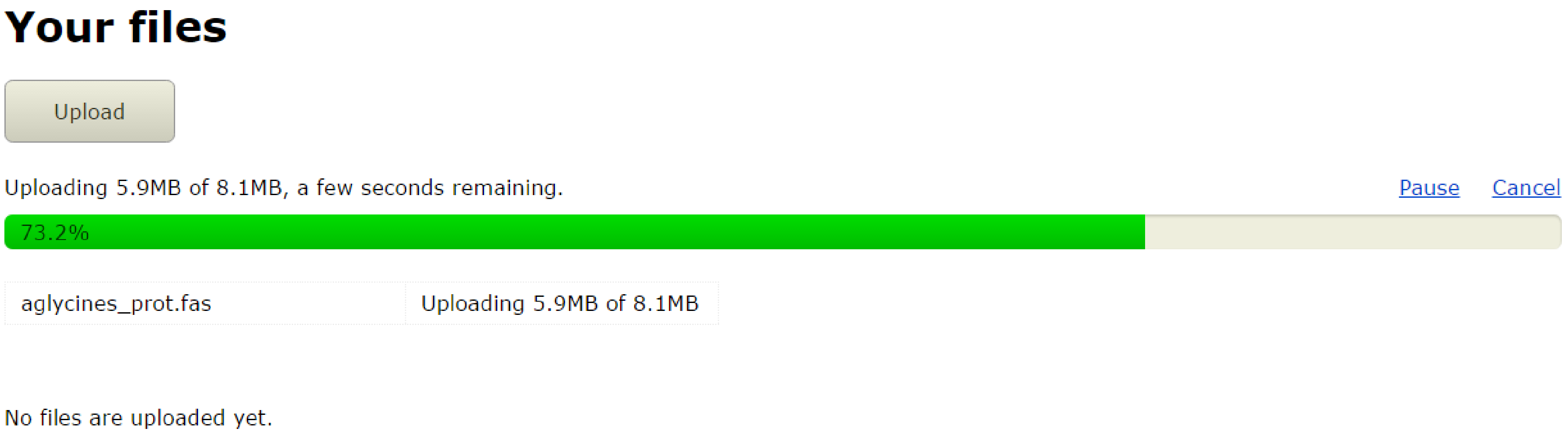
Jacob A. Wenger, Bryan J. Cassone, Fabrice Legeai, J. Spencer Johnston, Raman Bansal, Ashley D. Yates, Brad S. Coates, Vitor A.C. Pavinato, Andy Michel

PII:	S0965-1748(17)30005-X
DOI:	10.1016/j.ibmb.2017.01.005
Reference:	IB 2917
To appear in:	*Insect Biochemistry and Molecular Biology*
Received Date:	10 May 2016
Revised Date:	3 January 2017
Accepted Date:	14 January 2017

Please cite this article as: Wenger, J.A., Cassone, B.J., Legeai, F., Johnston, J.S., Bansal, R., Yates, A.D., Coates, B.S., Pavinato, V.A.C., Michel, A., Whole genome sequence of the soybean aphid, *Aphis glycines*, *Insect Biochemistry and Molecular Biology* (2017), doi: 10.1016/j.ibmb.2017.01.005.

Using OrthoDB in your research

Comparative analysis with your own data



Your files

Uploading 5.9MB of 8.1MB, a few seconds remaining.

aglycines_prot.fas	Uploading 5.9MB of 8.1MB

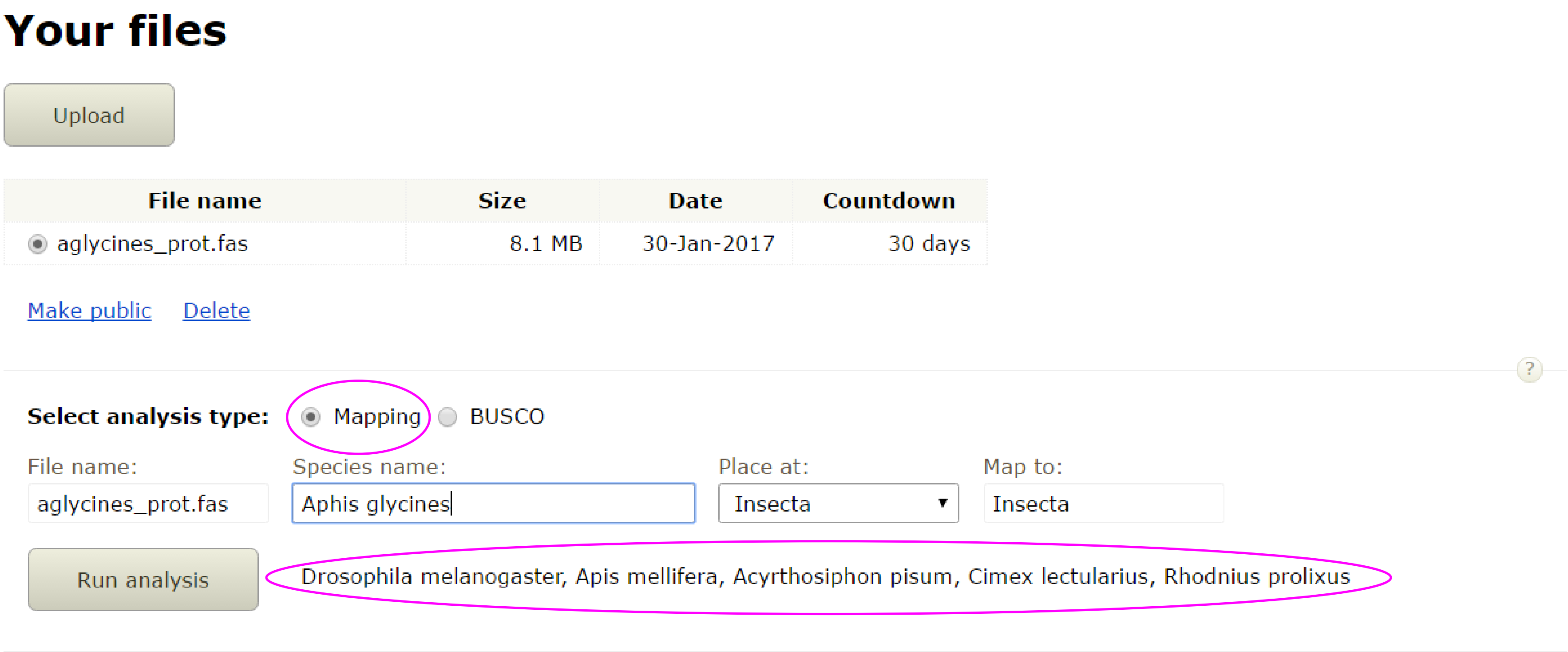
No files are uploaded yet.

Make sure your protein sequences are really in a proper FASTA format!

Make sure to select just one protein per gene in the case of alternative transcripts

Using OrthoDB in your research

Comparative analysis with your own data



Your files

Upload

File name	Size	Date	Countdown
● aglycines_prot.fas	8.1 MB	30-Jan-2017	30 days

Make public   Delete

Select analysis type: ● Mapping ○ BUSCO

File name: aglycines_prot.fas

Species name: Aphis glycines

Place at: Insecta ▾

Map to: Insecta

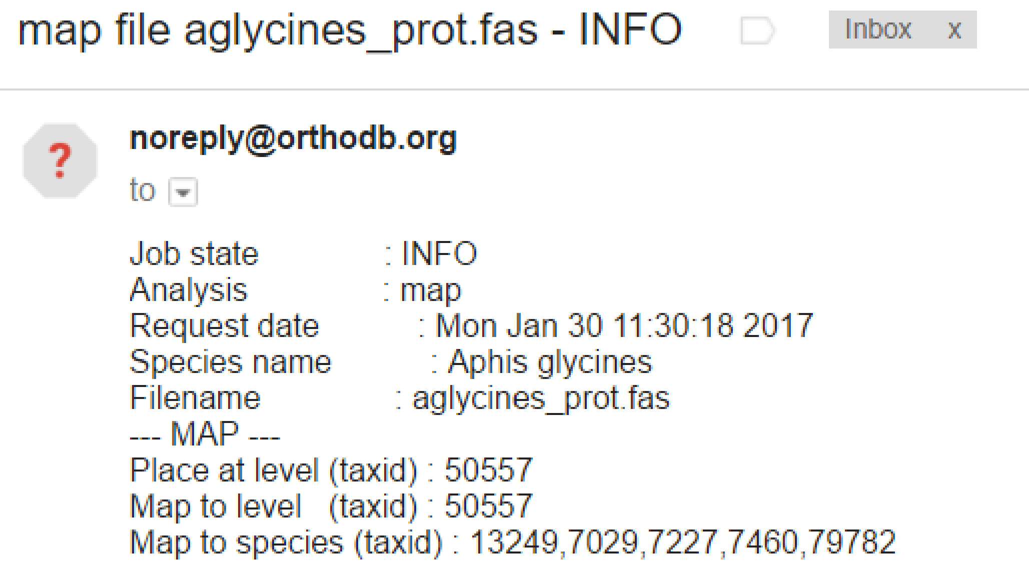
Run analysis

Drosophila melanogaster, Apis mellifera, Acyrthosiphon pisum, Cimex lectularius, Rhodnius prolixus

Use species selector to choose 1-5 compara species
Note automatic selection of LCA

Using *OrthoDB* in your research

Comparative analysis with your own data

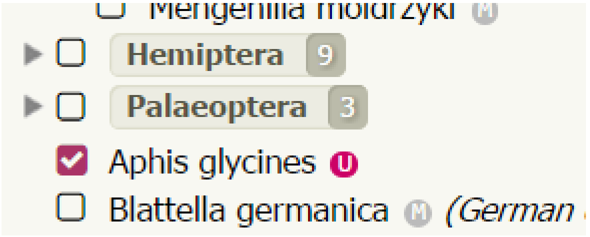


map file aglycines_prot.fas - INFO   Inbox x

noreply@orthodb.org

to

```
Job state          : INFO
Analysis           : map
Request date          : Mon Jan 30 11:30:18 2017
Species name          : Aphis glycines
Filename            : aglycines_prot.fas
--- MAP ---
Place at level (taxid) : 50557
Map to level   (taxid) : 50557
Map to species (taxid) : 13249,7029,7227,7460,79782
```



□ Mengenilla moldrzyki M

▶ □ Hemiptera 9

▶ □ Palaeoptera 3

☑ Aphis glycines U

□ Blattella germanica M (German

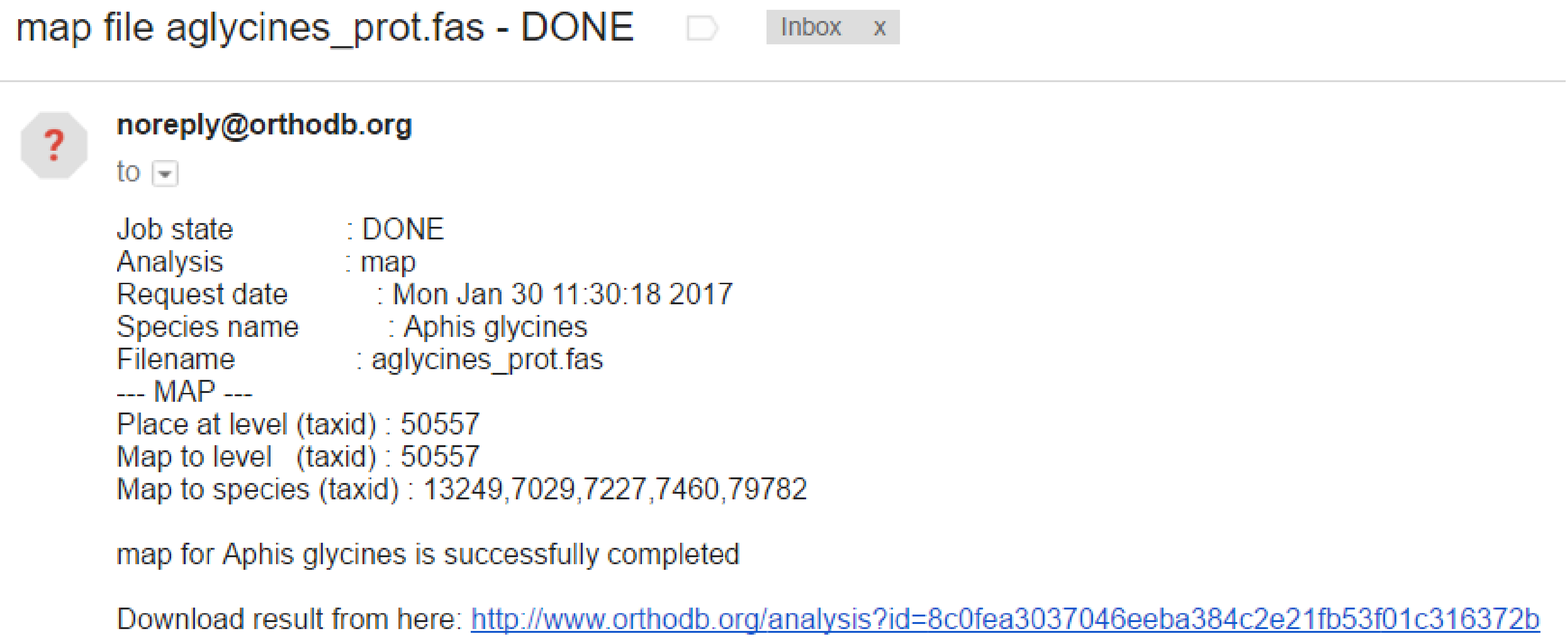
Will appear on the tree

Mapping will take quite some time!

You should receive updates on the progress of your mapping job by email

Using OrthoDB in your research

Comparative analysis with your own data



map file aglycines_prot.fas - DONE

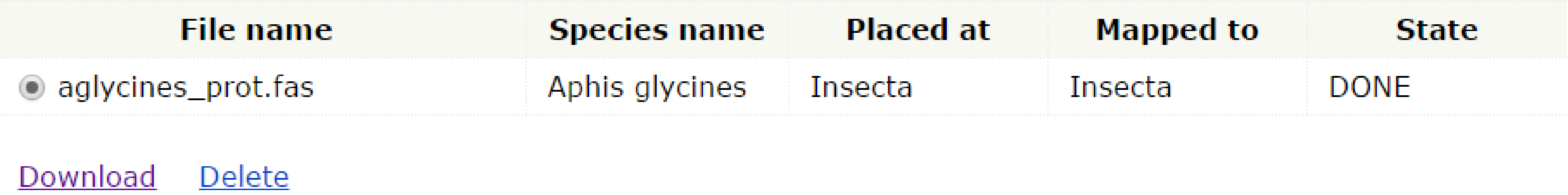
noreply@orthodb.org

```
Job state              : DONE
Analysis               : map
Request date           : Mon Jan 30 11:30:18 2017
Species name           : Aphis glycines
Filename               : aglycines_prot.fas
--- MAP ---
Place at level (taxid) : 50557
Map to level   (taxid) : 50557
Map to species (taxid) : 13249,7029,7227,7460,79782

map for Aphis glycines is successfully completed

Download result from here: http://www.orthodb.org/analysis?id=8c0fea3037046eeba384c2e21fb53f01c316372b
```

1) *Download results of your gene IDs mapped to OrthoDB orthologous group IDs*
2) *Browse OrthoDB with your species included*

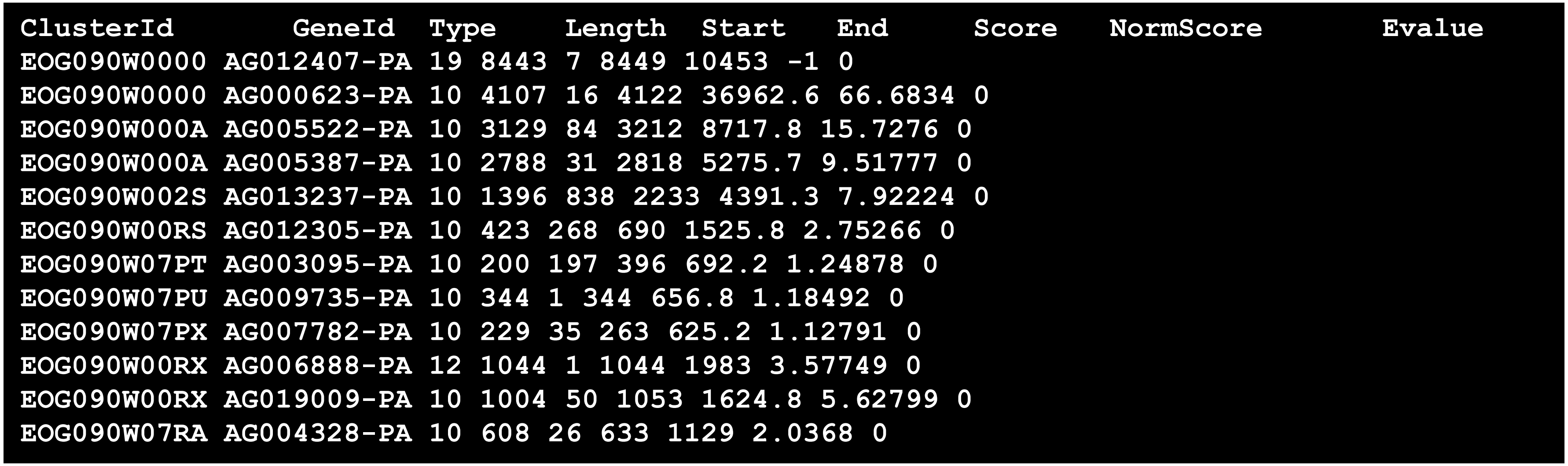
Using OrthoDB in your research

Comparative analysis with your own data



File name	Species name	Placed at	Mapped to	State
◉ aglycines_prot.fas	Aphis glycines	Insecta	Insecta	DONE

Download Delete

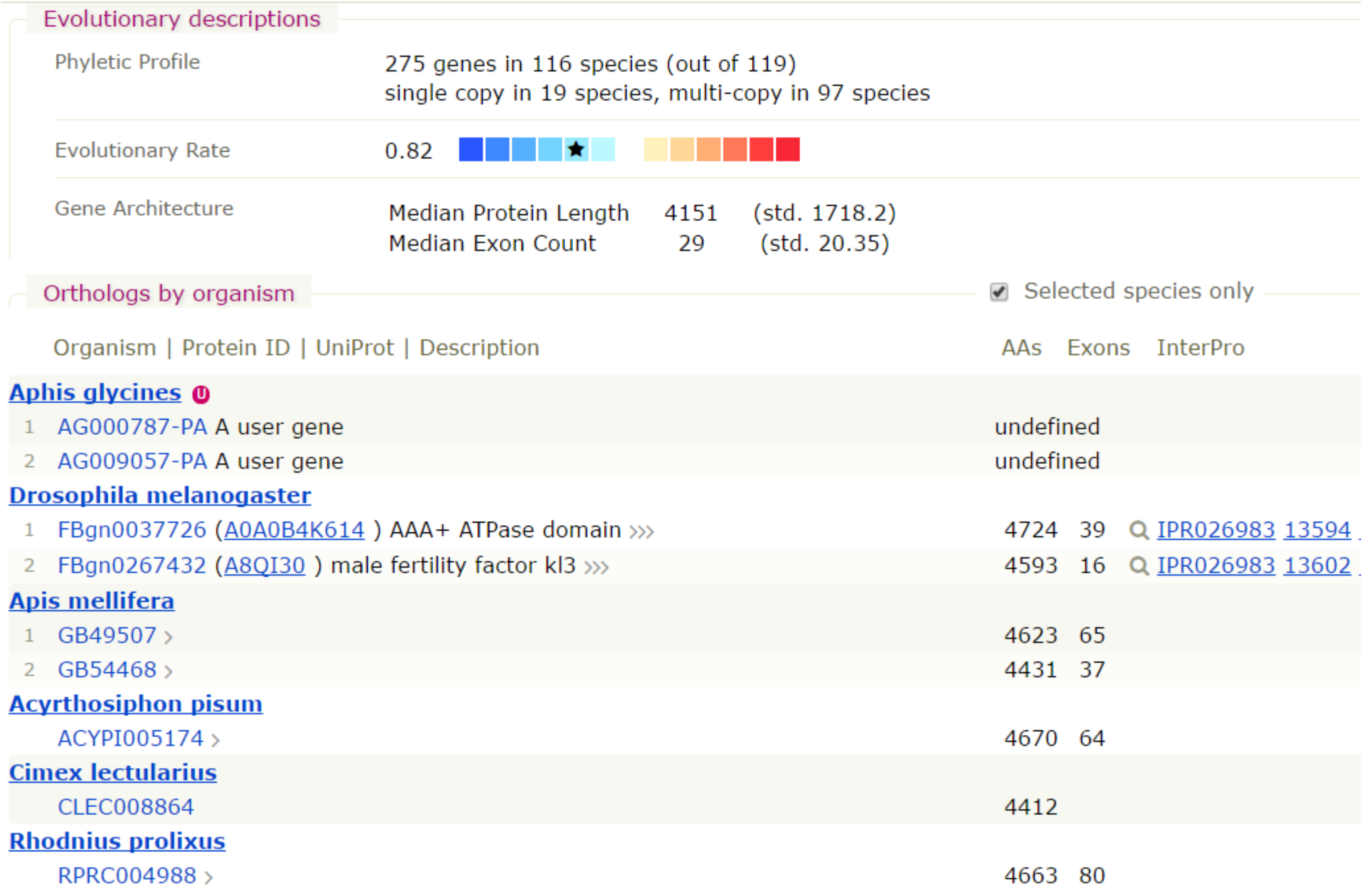


```
ClusterId        GeneId  Type    Length  Start   End     Score    NormScore      Evalue
EOG090W0000 AG012407-PA 19 8443 7 8449 10453 -1 0
EOG090W0000 AG000623-PA 10 4107 16 4122 36962.6 66.6834 0
EOG090W000A AG005522-PA 10 3129 84 3212 8717.8 15.7276 0
EOG090W000A AG005387-PA 10 2788 31 2818 5275.7 9.51777 0
EOG090W002S AG013237-PA 10 1396 838 2233 4391.3 7.92224 0
EOG090W00RS AG012305-PA 10 423 268 690 1525.8 2.75266 0
EOG090W07PT AG003095-PA 10 200 197 396 692.2 1.24878 0
EOG090W07PU AG009735-PA 10 344 1 344 656.8 1.18492 0
EOG090W07PX AG007782-PA 10 229 35 263 625.2 1.12791 0
EOG090W00RX AG006888-PA 12 1044 1 1044 1983 3.57749 0
EOG090W00RX AG019009-PA 10 1004 50 1053 1624.8 5.62799 0
EOG090W07RA AG004328-PA 10 608 26 633 1129 2.0368 0
```

1) Download results of your gene IDs mapped to OrthoDB orthologous group IDs

Using OrthoDB in your research

Comparative analysis with your own data



Evolutionary descriptions

Phyletic Profile	275 genes in 116 species (out of 119) single copy in 19 species, multi-copy in 97 species
Evolutionary Rate	0.82 🟦🟦🟦🟦★🟦 🟨🟧🟧🟧🟥🟥
Gene Architecture	Median Protein Length 4151 (std. 1718.2) Median Exon Count 29 (std. 20.35)

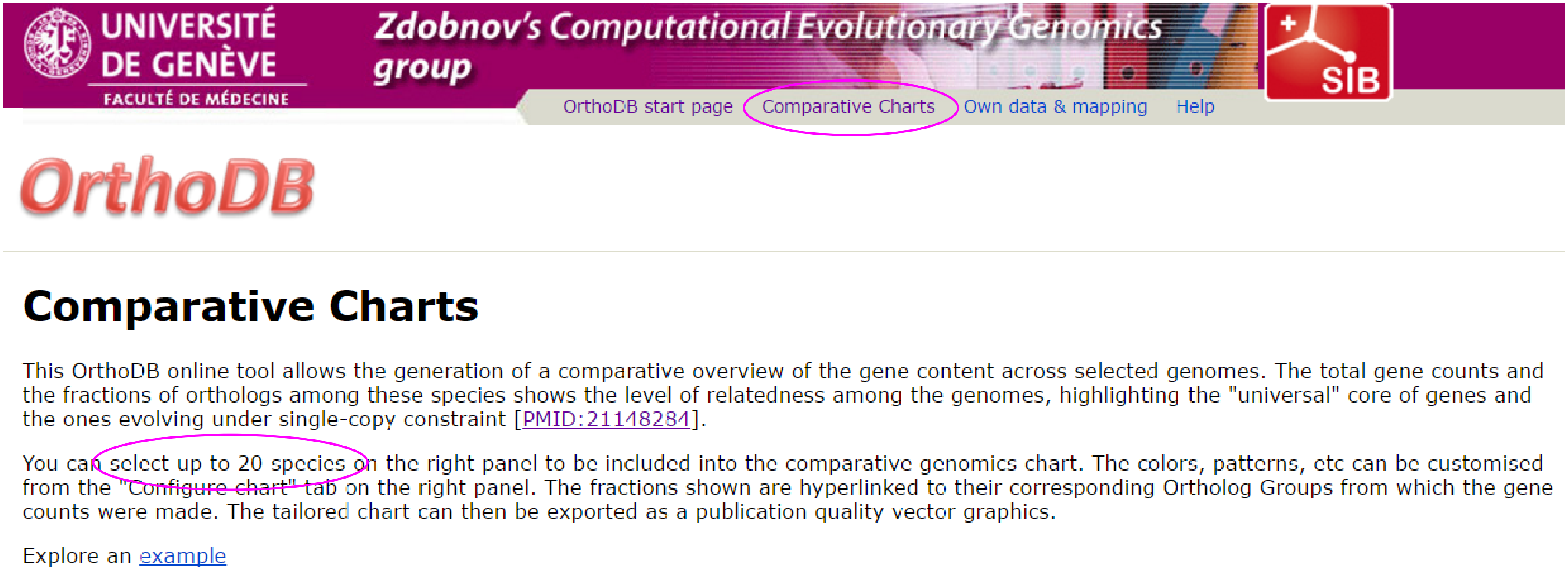
Orthologs by organism ☑ Selected species only

	Organism | Protein ID | UniProt | Description	AAs	Exons	InterPro
	Aphis glycines U			
1	AG000787-PA A user gene	undefined		
2	AG009057-PA A user gene	undefined		
	Drosophila melanogaster			
1	FBgn0037726 (A0A0B4K614) AAA+ ATPase domain ›››	4724	39	IPR026983 13594
2	FBgn0267432 (A8QI30) male fertility factor kl3 ›››	4593	16	IPR026983 13602
	Apis mellifera			
1	GB49507 ›	4623	65	
2	GB54468 ›	4431	37	
	Acyrthosiphon pisum			
	ACYPI005174 ›	4670	64	
	Cimex lectularius			
	CLEC008864	4412		
	Rhodnius prolixus			
	RPRC004988 ›	4663	80	

2) Browse OrthoDB with your species included

Using *OrthoDB* in your research

Comparative analysis



Comparative Charts

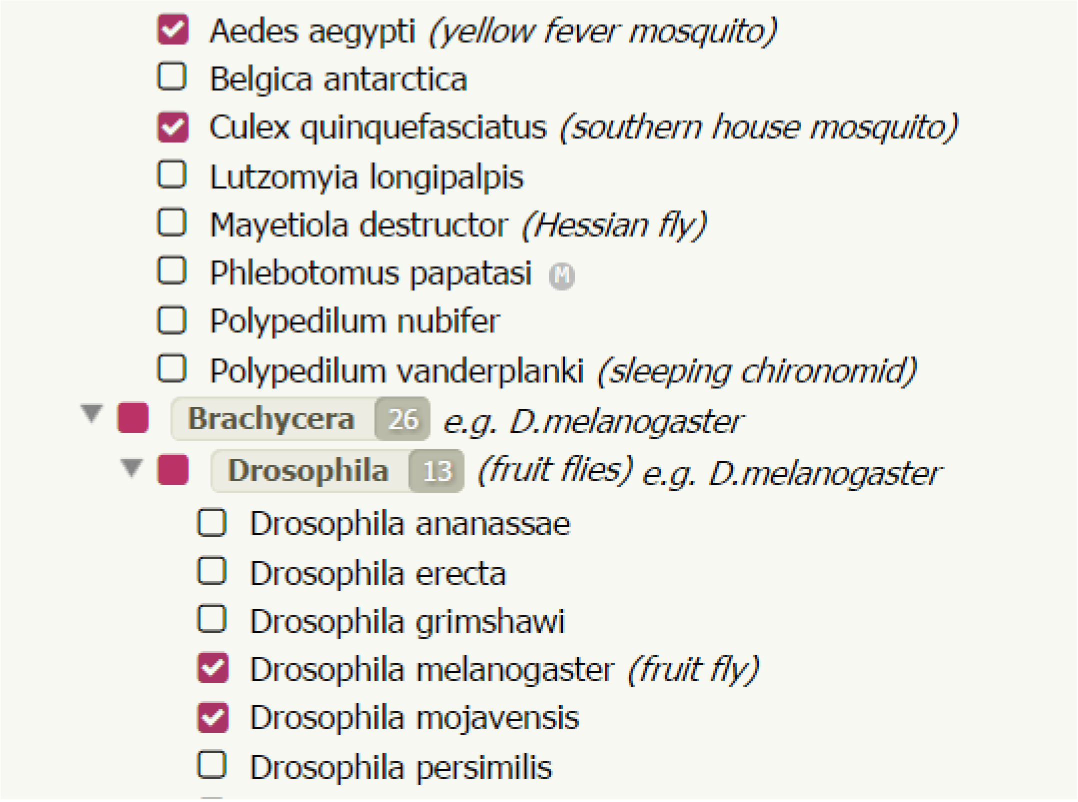
This OrthoDB online tool allows the generation of a comparative overview of the gene content across selected genomes. The total gene counts and the fractions of orthologs among these species shows the level of relatedness among the genomes, highlighting the "universal" core of genes and the ones evolving under single-copy constraint [PMID:21148284].

You can select up to 20 species on the right panel to be included into the comparative genomics chart. The colors, patterns, etc can be customised from the "Configure chart" tab on the right panel. The fractions shown are hyperlinked to their corresponding Ortholog Groups from which the gene counts were made. The tailored chart can then be exported as a publication quality vector graphics.

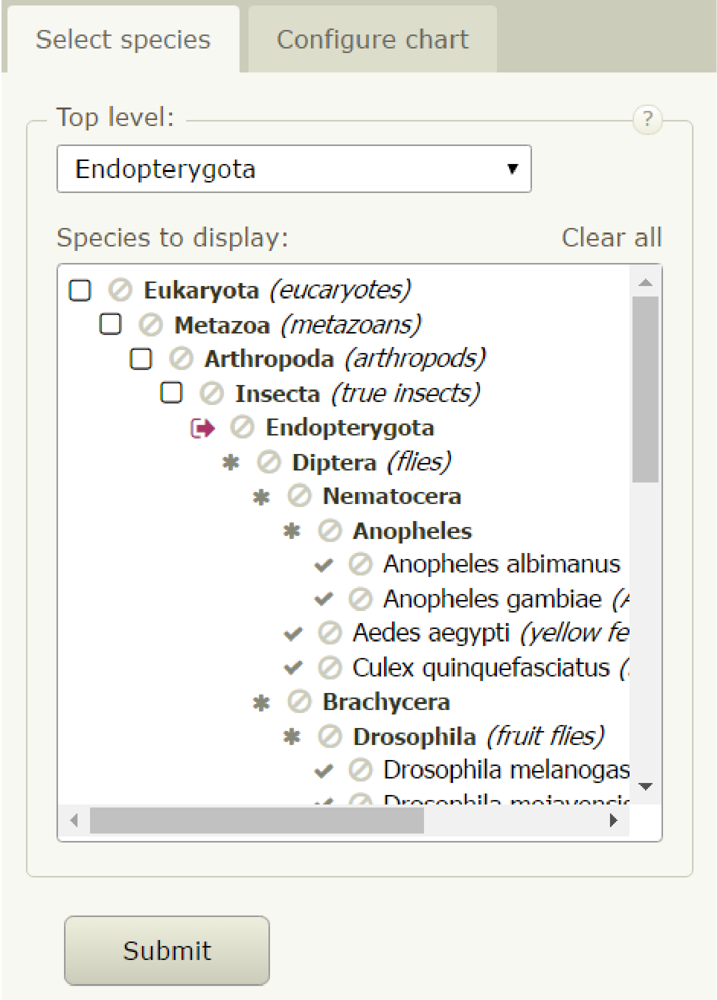
Explore an example

Using *OrthoDB* in your research

Comparative analysis



☑ Aedes aegypti *(yellow fever mosquito)*
□ Belgica antarctica
☑ Culex quinquefasciatus *(southern house mosquito)*
□ Lutzomyia longipalpis
□ Mayetiola destructor *(Hessian fly)*
□ Phlebotomus papatasi Ⓜ
□ Polypedilum nubifer
□ Polypedilum vanderplanki *(sleeping chironomid)*
▼ ■ Brachycera 26 *e.g. D.melanogaster*
▼ ■ Drosophila 13 *(fruit flies) e.g. D.melanogaster*
□ Drosophila ananassae
□ Drosophila erecta
□ Drosophila grimshawi
☑ Drosophila melanogaster *(fruit fly)*
☑ Drosophila mojavensis
□ Drosophila persimilis



Top level: Endopterygota ▾ ?

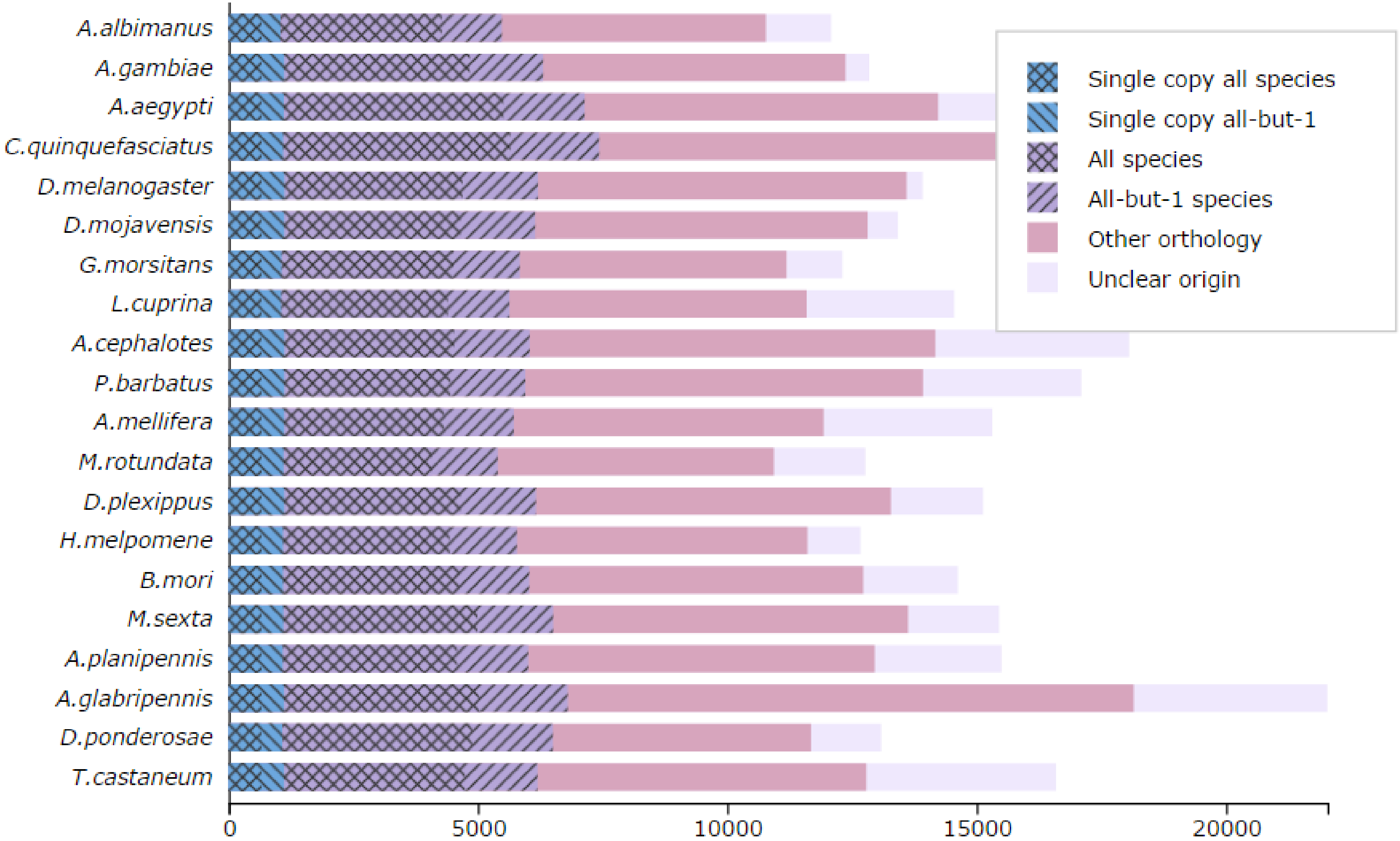
Species to display: Clear all

□ ⊘ Eukaryota *(eucaryotes)*
□ ⊘ Metazoa *(metazoans)*
□ ⊘ Arthropoda *(arthropods)*
□ ⊘ Insecta *(true insects)*
⊘ Endopterygota
✱ ⊘ Diptera *(flies)*
✱ ⊘ Nematocera
✱ ⊘ Anopheles
✓ ⊘ Anopheles albimanus
✓ ⊘ Anopheles gambiae *(A*
✓ ⊘ Aedes aegypti *(yellow fe*
✓ ⊘ Culex quinquefasciatus *(*
✱ ⊘ Brachycera
✱ ⊘ Drosophila *(fruit flies)*
✓ ⊘ Drosophila melanogas
✓ ⊘ Drosophila mojavensis

Select up to 20 species, automatic selection last common ancestor: Submit!

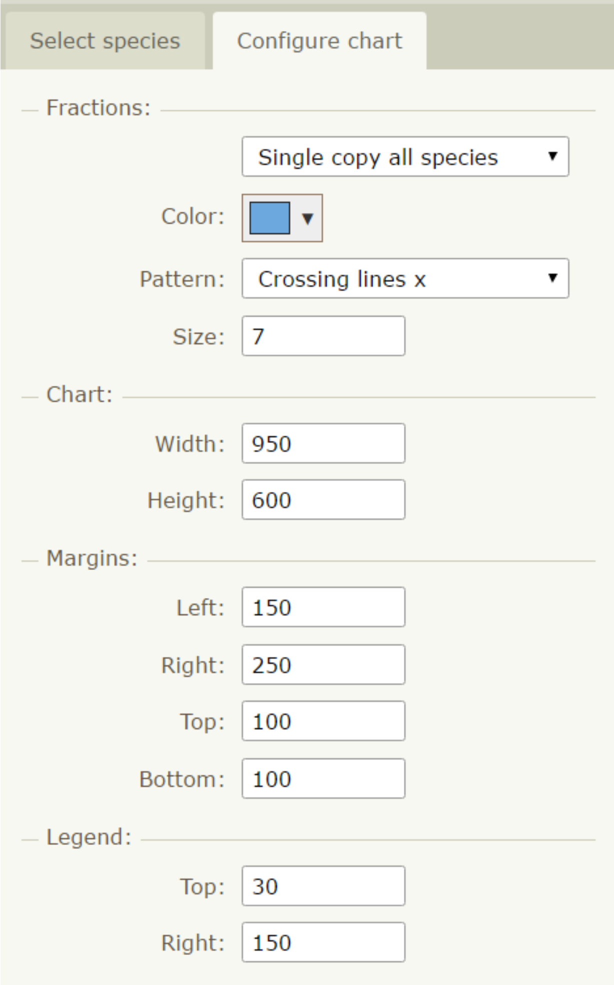
Using OrthoDB in your research

Comparative analysis



Using OrthoDB in your research

Comparative analysis



Select species | Configure chart

Fractions:

Single copy all species ▾

Color: ▾

Pattern: Crossing lines x ▾

Size: 7

Chart:

Width: 950

Height: 600

Margins:

Left: 150

Right: 250

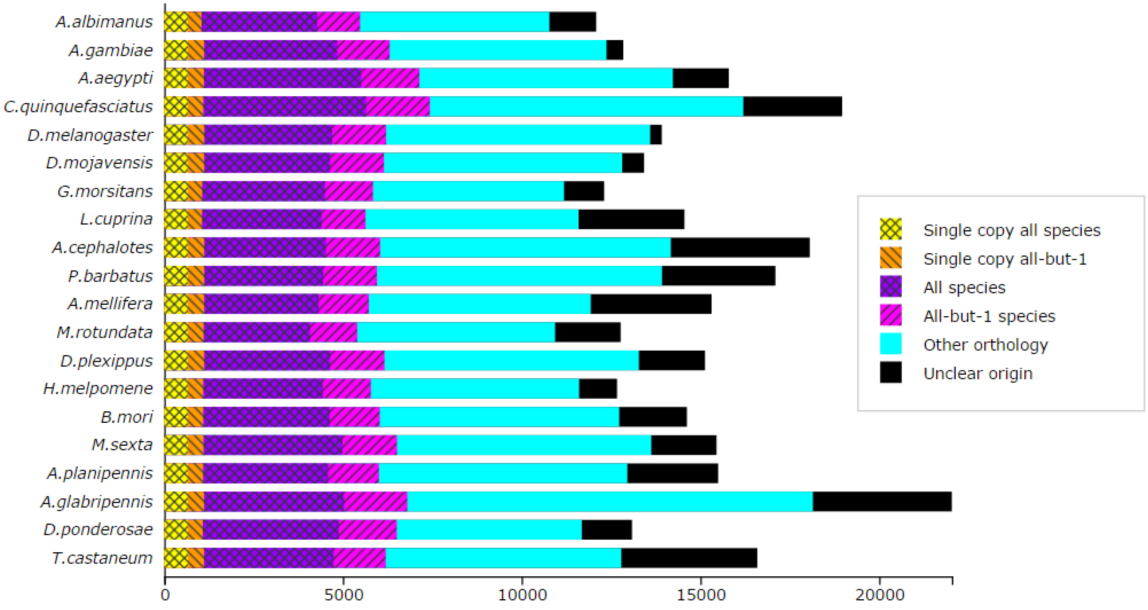
Top: 100

Bottom: 100

Legend:

Top: 30

Right: 150

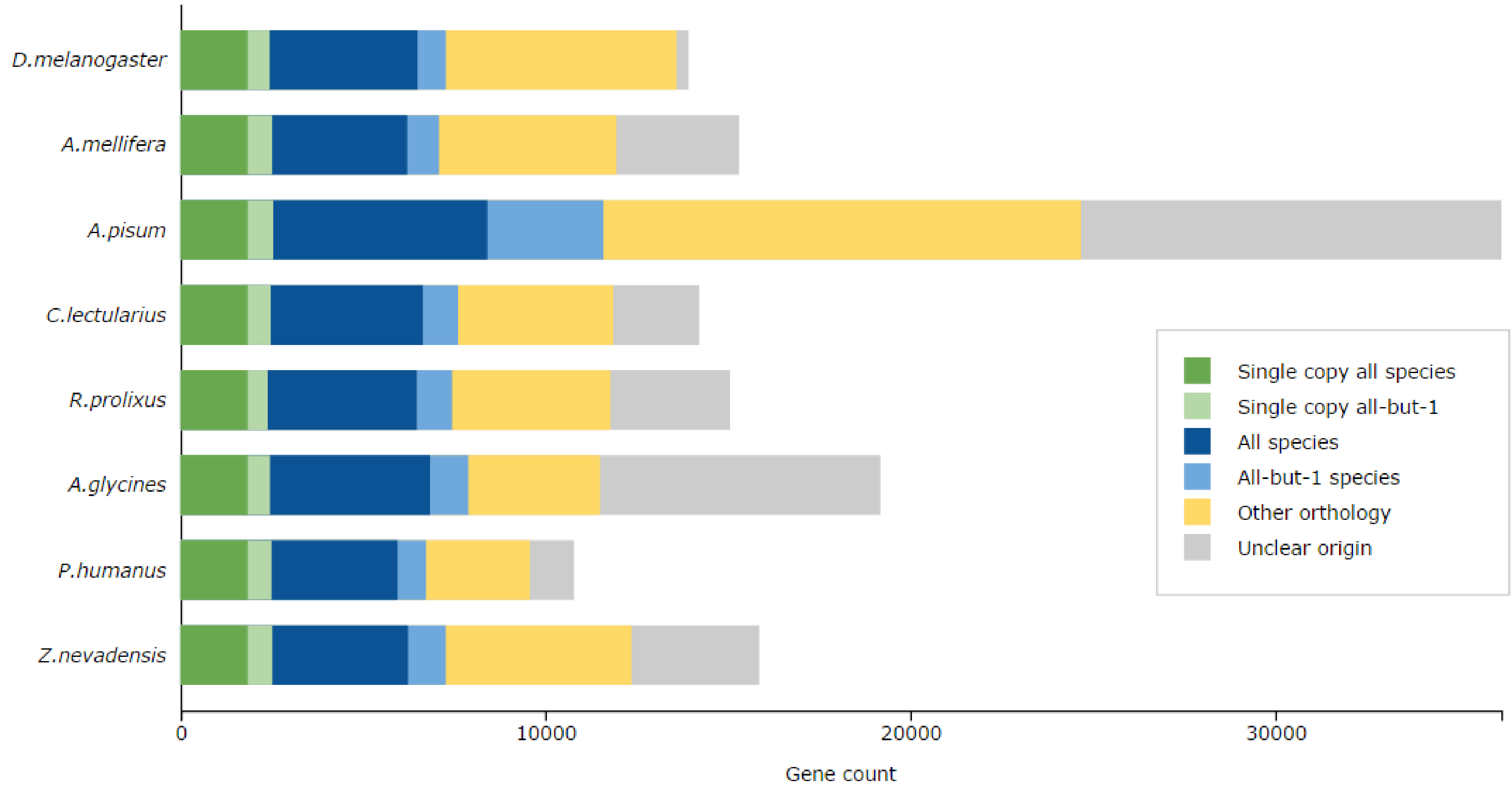


Configure the chart size, colours, margins, position of legend etc.
Then export as an image for your manuscript!

Save as vector image | Save as bitmap

Using *OrthoDB* in your research

Comparative analysis with your own data

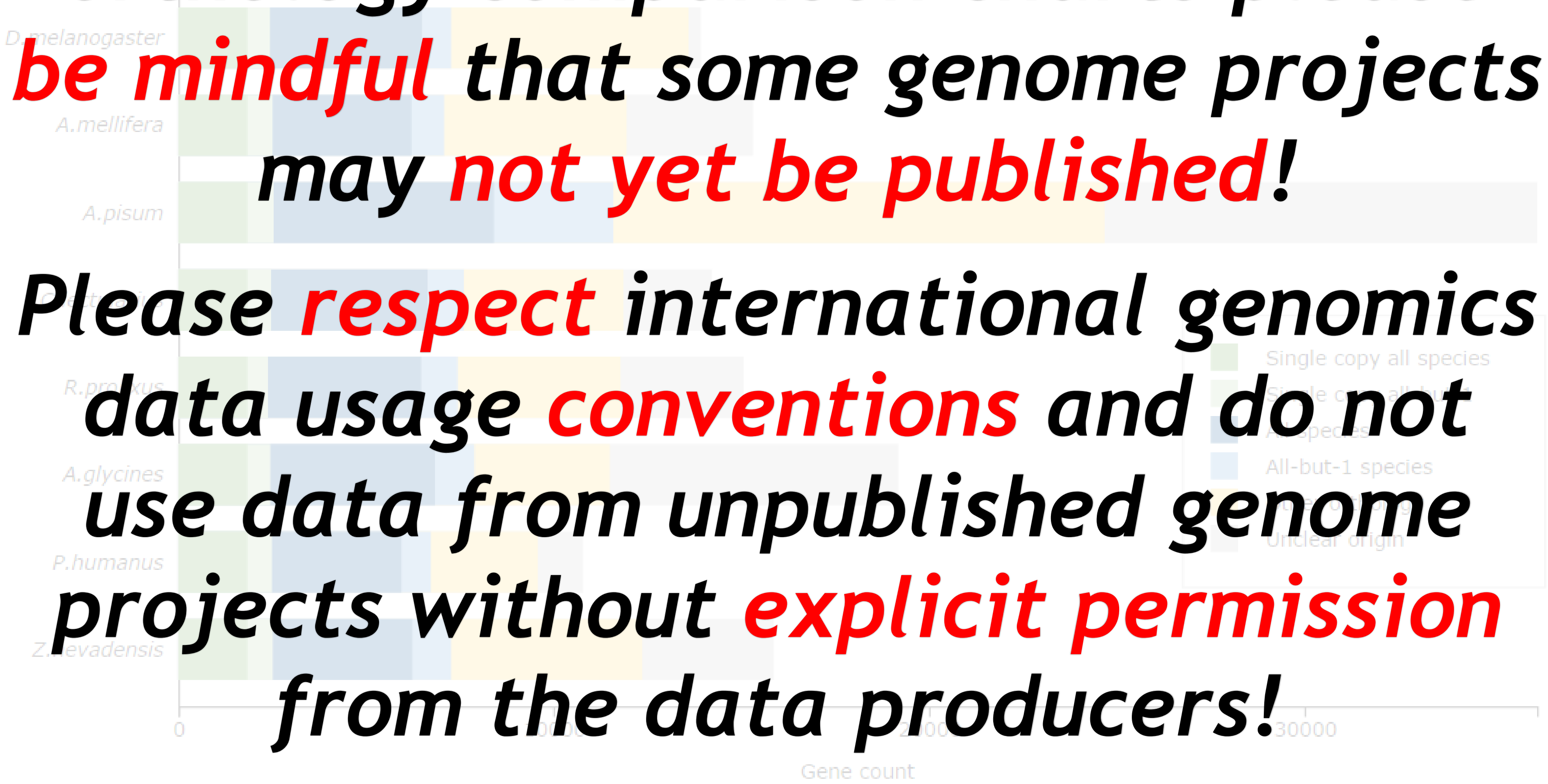


Using OrthoDB in your research

Comparative analysis with your own data

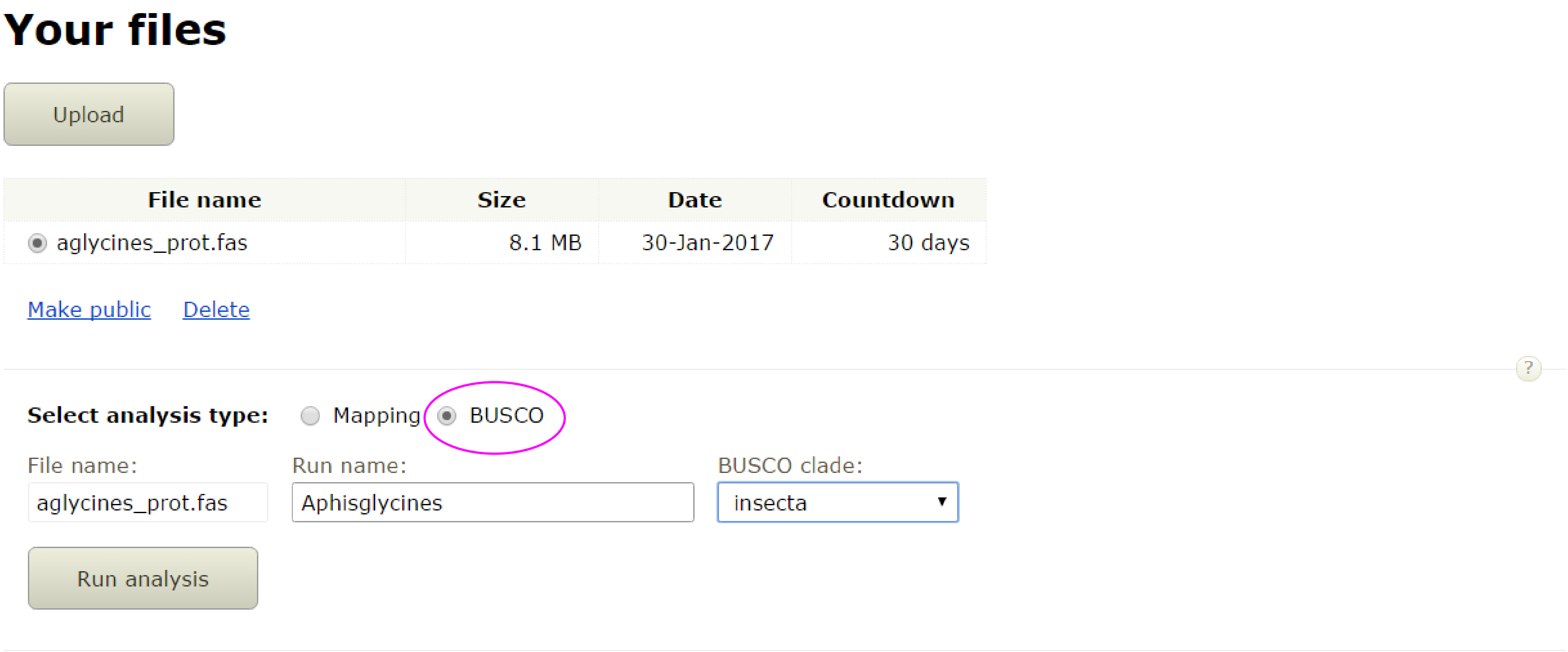
When selecting species for orthology comparison charts please be mindful that some genome projects may not yet be published!

Please respect international genomics data usage conventions and do not use data from unpublished genome projects without explicit permission from the data producers!



Using *OrthoDB* in your research

BUSCO protein set assessments



Your files

	File name	Size	Date	Countdown
●	aglycines_prot.fas	8.1 MB	30-Jan-2017	30 days

Select analysis type: ○ Mapping ● BUSCO

File name: aglycines_prot.fas

Run name: Aphisglycines

BUSCO clade: insecta

Gene annotation set quality control with
Benchmarking Universal Single-Copy Orthologs

Using BUSCO in your research

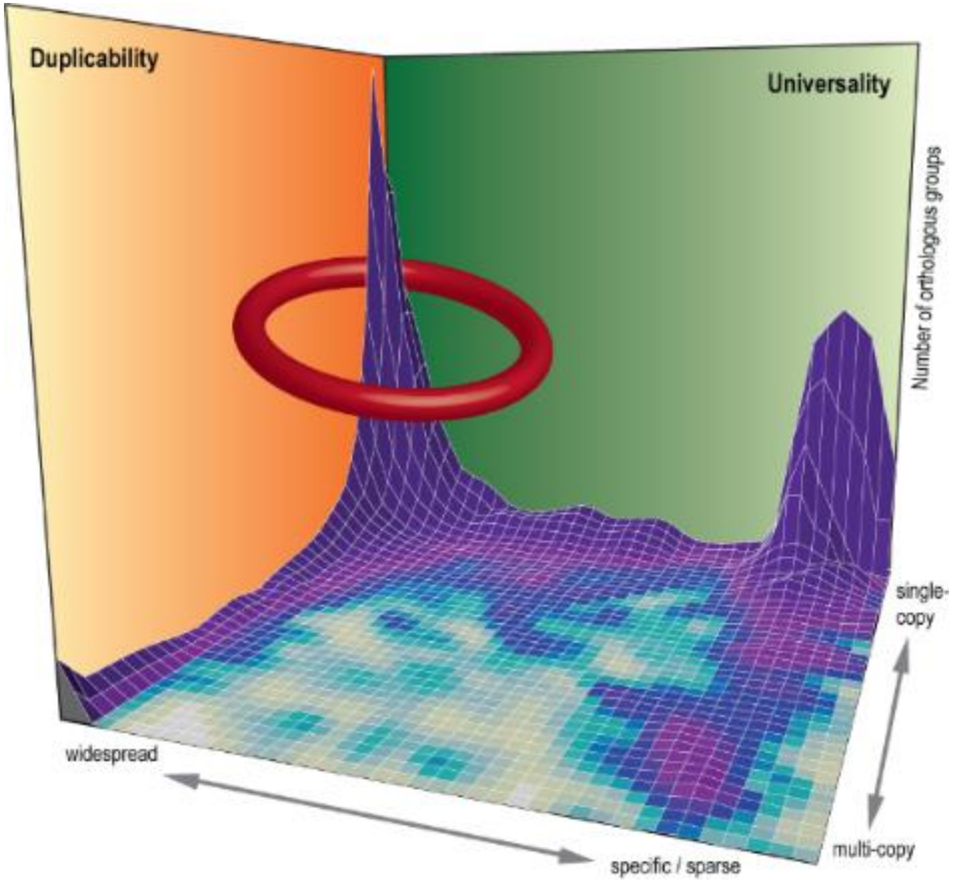
Ortho-Groups with genes found in the majority of species as single-copy orthologues

Evolutionary Expectation for them to be found in any newly-sequenced genome

Implemented Assessments
Gene Content Completeness
# genome assemblies
# annotated gene sets
# assembled transcriptomes

Bonus Features
# genes for phylogenomics
# gene predictor training

Orthology Landscape

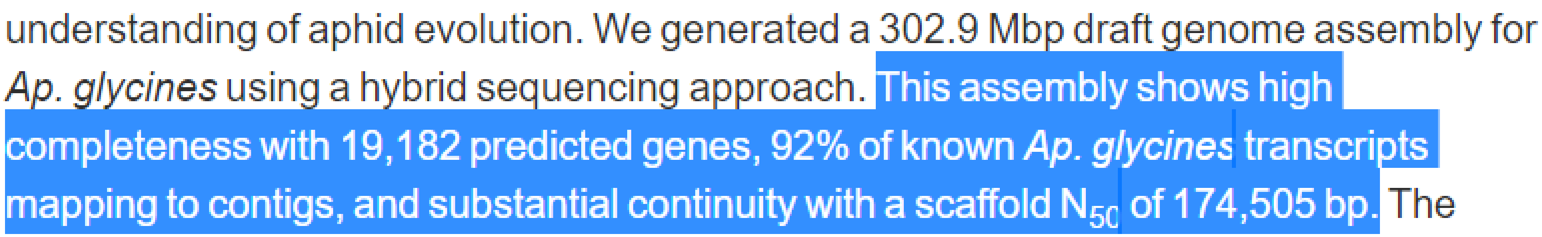


http://busco.ezlab.org

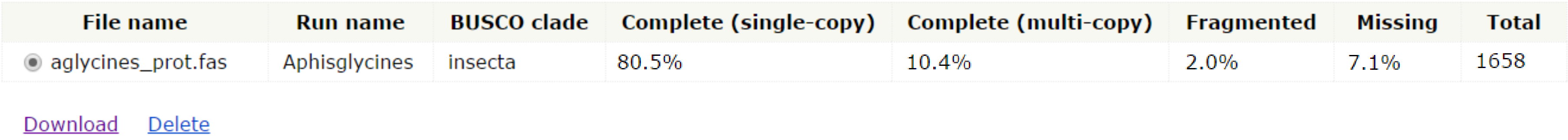
Using *OrthoDB* in your research

BUSCO protein set assessments

Wenger *et al*, 2017



understanding of aphid evolution. We generated a 302.9 Mbp draft genome assembly for *Ap. glycines* using a hybrid sequencing approach. This assembly shows high completeness with 19,182 predicted genes, 92% of known *Ap. glycines* transcripts mapping to contigs, and substantial continuity with a scaffold N_{50} of 174,505 bp. The



File name	Run name	BUSCO clade	Complete (single-copy)	Complete (multi-copy)	Fragmented	Missing	Total
● aglycines_prot.fas	Aphisglycines	insecta	80.5%	10.4%	2.0%	7.1%	1658

```
C:90.9%[S:80.5%,D:10.4%],F:2.0%,M:7.1%,n:1658

        1508    Complete BUSCOs (C)
        1335    Complete and single-copy BUSCOs (S)
        173     Complete and duplicated BUSCOs (D)
        33      Fragmented BUSCOs (F)
        117     Missing BUSCOs (M)
        1658    Total BUSCO groups searched
```

What is orthology?

Understanding the definitions

How does one delineate orthology?

Getting to grips with the methodologies

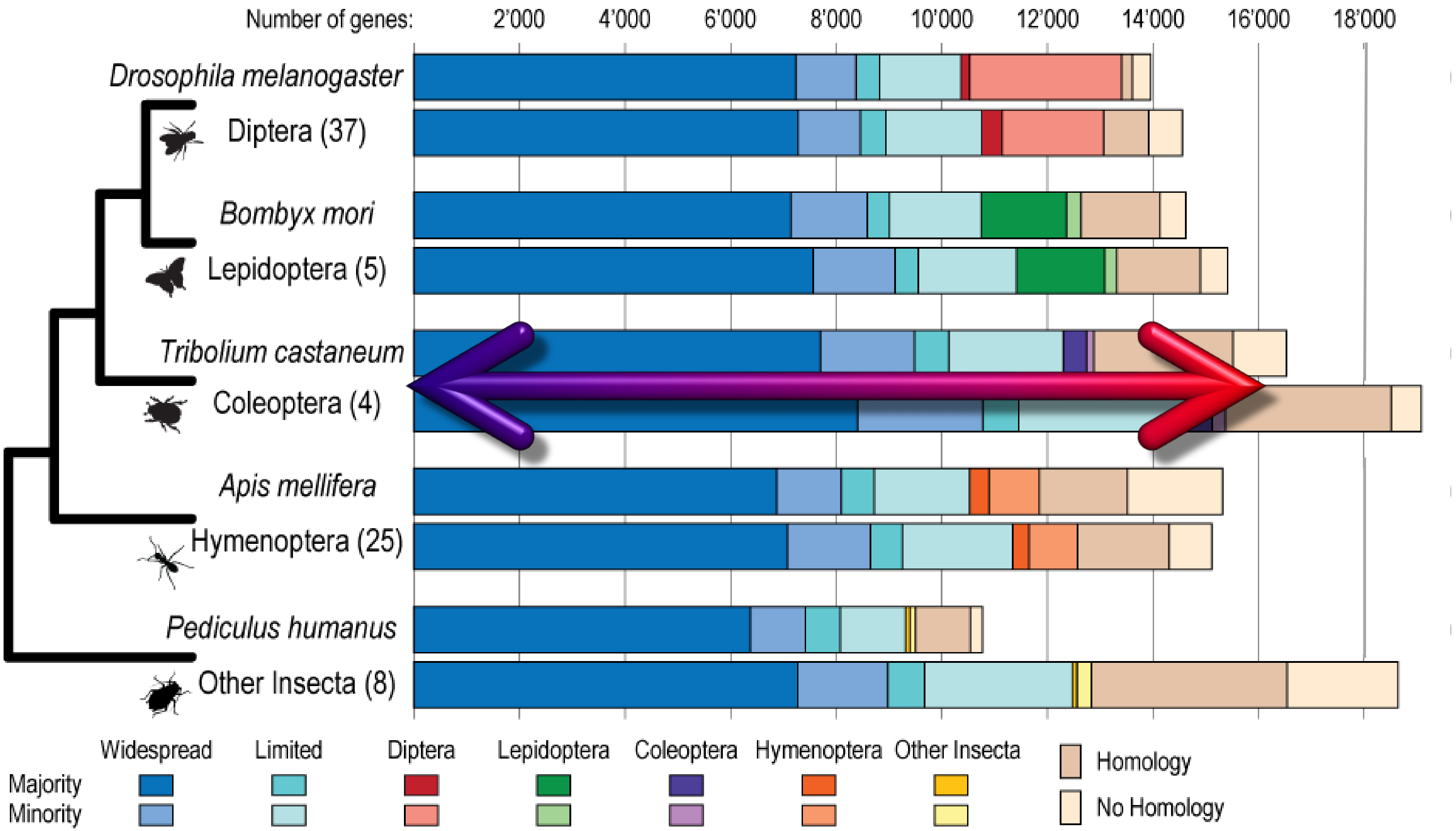
What does OrthoDB offer?

Using orthology in your research

Why does one need to delineate orthology?

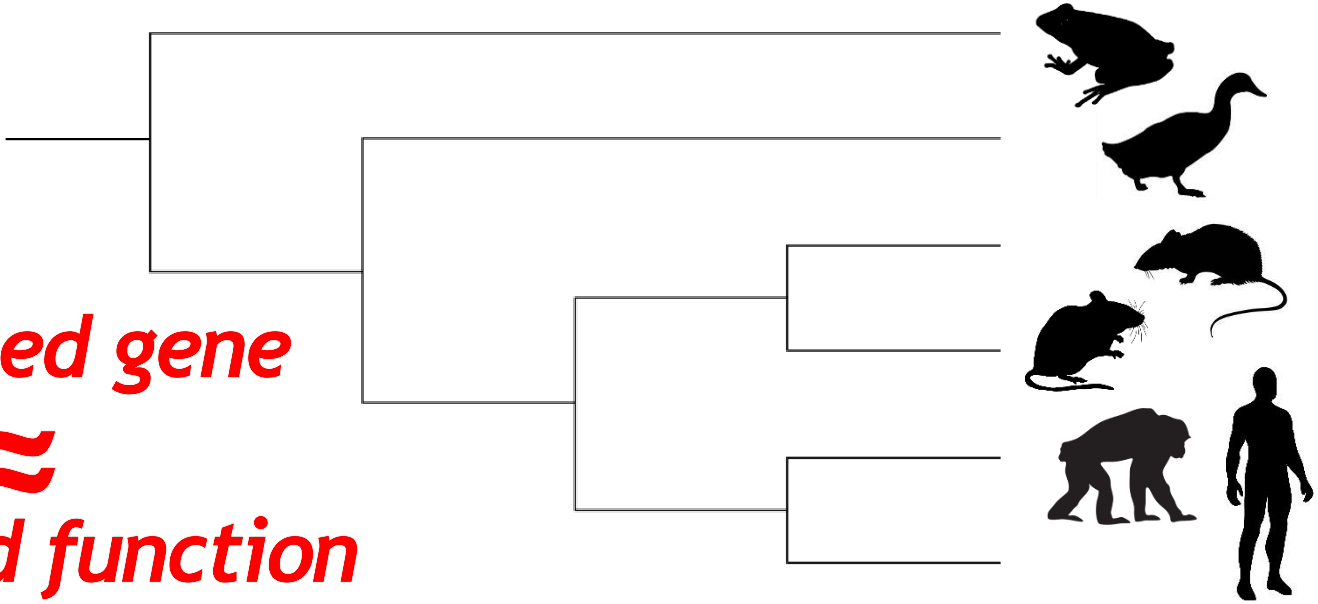
1) Tracing the **Evolutionary Histories** of all genes in extant species

2) Building **Hypotheses on Gene Function** informed by evolution



Why does one need to delineate orthology?

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



Inherited gene

≈

Inherited function

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

Why does one need to delineate orthology?

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

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

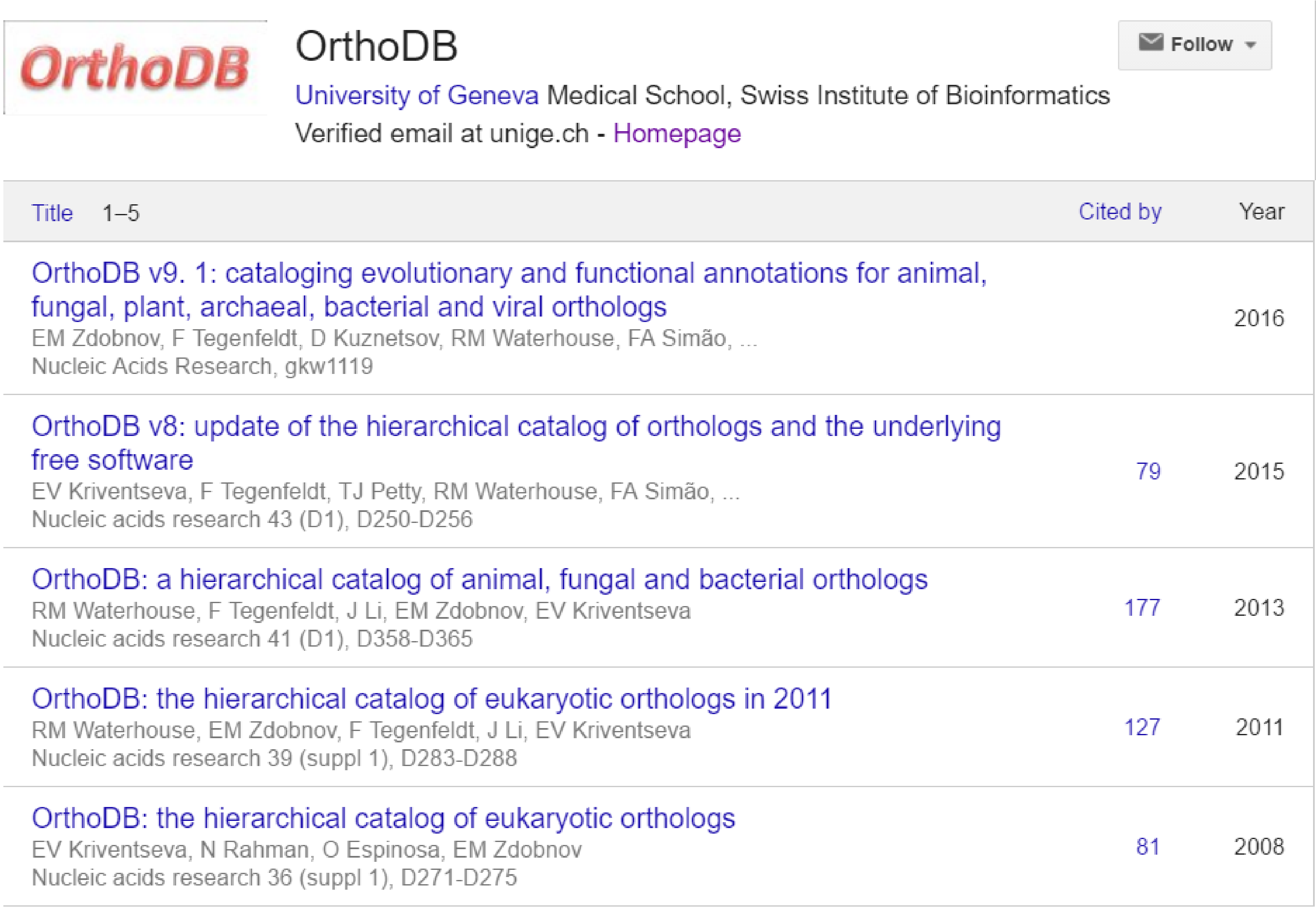
“The validity of the conjecture on functional equivalency of orthologs is crucial for reliable annotation of newly sequenced genomes and, more generally, for the progress of functional genomics.

The huge majority of genes in the sequenced genomes will never be studied experimentally, so for most genomes transfer of functional information between orthologs is the only means of detailed functional characterization.”

Annu. Rev. Genet.
2005. 39:309–38

Have you or will you use OrthoDB orthology?

Please cite relevant Nucleic Acids Research database issue publication



OrthoDB

University of Geneva Medical School, Swiss Institute of Bioinformatics
Verified email at unige.ch - Homepage

Title 1–5	Cited by	Year
OrthoDB v9. 1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs EM Zdobnov, F Tegenfeldt, D Kuznetsov, RM Waterhouse, FA Simão, ... Nucleic Acids Research, gkw1119		2016
OrthoDB v8: update of the hierarchical catalog of orthologs and the underlying free software EV Kriventseva, F Tegenfeldt, TJ Petty, RM Waterhouse, FA Simão, ... Nucleic acids research 43 (D1), D250-D256	79	2015
OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs RM Waterhouse, F Tegenfeldt, J Li, EM Zdobnov, EV Kriventseva Nucleic acids research 41 (D1), D358-D365	177	2013
OrthoDB: the hierarchical catalog of eukaryotic orthologs in 2011 RM Waterhouse, EM Zdobnov, F Tegenfeldt, J Li, EV Kriventseva Nucleic acids research 39 (suppl 1), D283-D288	127	2011
OrthoDB: the hierarchical catalog of eukaryotic orthologs EV Kriventseva, N Rahman, O Espinosa, EM Zdobnov Nucleic acids research 36 (suppl 1), D271-D275	81	2008

Questions? Write to us: support@orthodb.org

Prof. Evgeny Zdobnov
Professor, Uni. Geneva
Group Leader, SIB

Dr Evgenia Kriventseva
Research Associate
- OrthoDB Management

Dmitry Kuznetsov
Scientific Programmer
- OrthoDB Development

Dr Fredrik Tegenfeldt
Scientific Programmer
- OrthoDB Development

Dr Panos Ioannidis
Postdoctoral Fellow
- Arthropod Genomics

Dr Robert Waterhouse
Postdoctoral Fellow
- Arthropod Genomics

Mathieu Seppey
PhD Student
- BUSCO Assessments

Felipe Simão
PhD Student
- BUSCO Assessments

& all other members of the Zdobnov Computational Evolutionary Genomics Group