

# Supplementary Information

## NCBI Orthologs: Public Resource and Scalable Method for Computing High-Precision Orthologs Across Eukaryotic Genomes

Dong-Ha Oh, Alexander Astashyn, Barbara Robbertse, Nuala A O'leary, W Ray Anderson, Laurie Breen, Eric Cox, Olga Ermolaeva, Robert Falk, Vichet Hem, J Bradley Holmes, Patrick Masterson, Kelly M McGarvey, Eyal Mozes, John Torcivia, Mirian T N Tsuchiya, Craig Wallin, Francoise Thibaud-Nissen, Terence D. Murphy, Vamsi K. Kodali\*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA.

\* To whom correspondence should be addressed: Email:

[kodalivk@ncbi.nlm.nih.gov](mailto:kodalivk@ncbi.nlm.nih.gov)

**Supplementary Table 1** Orthologs calculated and submitted to the Quest for Orthologs (QfO) benchmarking service for evaluation

| Anchor tax_id | Anchor organism name                       | Query tax_id | Query organism name                                      | # Ortholog pairs |
|---------------|--------------------------------------------|--------------|----------------------------------------------------------|------------------|
| 10090         | <i>Mus musculus</i> (Mouse)                | 10116        | <i>Rattus norvegicus</i> (Rat)                           | 16,841           |
| 9606          | <i>Homo sapiens</i> (Human)                | 9598         | <i>Pan troglodytes</i> (Chimpanzee)                      | 17,124           |
| 9606          | <i>Homo sapiens</i> (Human)                | 9595         | <i>Gorilla gorilla gorilla</i> (Western lowland gorilla) | 16,809           |
| 9606          | <i>Homo sapiens</i> (Human)                | 10090        | <i>Mus musculus</i> (Mouse)                              | 16,096           |
| 9606          | <i>Homo sapiens</i> (Human)                | 9913         | <i>Bos taurus</i> (Bovine)                               | 15,783           |
| 9606          | <i>Homo sapiens</i> (Human)                | 10116        | <i>Rattus norvegicus</i> (Rat)                           | 15,493           |
| 9606          | <i>Homo sapiens</i> (Human)                | 9615         | <i>Canis lupus familiaris</i> (Dog)                      | 15,098           |
| 9606          | <i>Homo sapiens</i> (Human)                | 13616        | <i>Monodelphis domestica</i> (Gray short-tailed opossum) | 13,964           |
| 9606          | <i>Homo sapiens</i> (Human)                | 9031         | <i>Gallus gallus</i> (Chicken)                           | 12,329           |
| 9606          | <i>Homo sapiens</i> (Human)                | 8364         | <i>Xenopus tropicalis</i> (Western clawed frog)          | 12,262           |
| 7955          | <i>Danio rerio</i> (Zebrafish)             | 8090         | <i>Oryzias latipes</i> (Japanese rice fish)              | 13,860           |
| 7955          | <i>Danio rerio</i> (Zebrafish)             | 7918         | <i>Lepisosteus oculatus</i> (Spotted gar)                | 12,289           |
| 7227          | <i>Drosophila melanogaster</i> (Fruit fly) | 7165         | <i>Anopheles gambiae</i> (African malaria mosquito)      | 6,216            |
| 7227          | <i>Drosophila melanogaster</i> (Fruit fly) | 7070         | <i>Tribolium castaneum</i> (Red flour beetle)            | 5,685            |
| 7227          | <i>Drosophila melanogaster</i> (Fruit fly) | 6945         | <i>Ixodes scapularis</i> (Black-legged tick)             | 3,458            |

**Supplementary Table 2** Examples of orthologs gained by transitive anchor, together with the number of homologous neighbors (N).

| RefSeq accession | Species name                | Common name           | # Protein-coding genes | # Orthologs vs. fruit fly (%) | Mean N vs. fruit fly | # Orthologs vs. honey bee (%) | Mean N vs. honey bee |
|------------------|-----------------------------|-----------------------|------------------------|-------------------------------|----------------------|-------------------------------|----------------------|
| GCF_003254395.2  | <i>Apis mellifera</i>       | honey bee             | 9,935                  | 5,547 (55.8)                  | 0.95                 | anchor                        | 20.00                |
| GCF_001442555.1  | <i>Apis cerana</i>          | Asiatic honeybee      | 10,719                 | 5,255 (49.0)                  | 0.92                 | <b>9,086 (84.8)</b>           | <b>13.05</b>         |
| GCF_910591885.1  | <i>Bombus terrestris</i>    | buff-tailed bumblebee | 10,310                 | 5,612 (54.4)                  | 0.95                 | <b>8,844 (85.8)</b>           | <b>12.43</b>         |
| GCF_024542735.1  | <i>Bombus huntii</i>        |                       | 11,088                 | 5,590 (50.4)                  | 0.94                 | <b>8,814 (79.5)</b>           | <b>12.42</b>         |
| GCF_001652005.2  | <i>Ceratina calcarata</i>   |                       | 11,173                 | 5,285 (47.3)                  | 0.89                 | <b>8,632 (77.3)</b>           | <b>6.49</b>          |
| GCF_910589235.1  | <i>Vespa crabro</i>         | European hornet       | 10,204                 | 5,512 (54.0)                  | 0.92                 | <b>8,437 (82.7)</b>           | <b>4.72</b>          |
| GCF_905475345.1  | <i>Vespula vulgaris</i>     |                       | 9,996                  | 5,530 (55.3)                  | 0.92                 | <b>8,474 (84.8)</b>           | <b>4.69</b>          |
| GCF_016802725.1  | <i>Solenopsis invicta</i>   | red fire ant          | 14,790                 | 5,514 (37.3)                  | 0.92                 | <b>8,394 (56.8)</b>           | <b>4.54</b>          |
| GCF_013373865.1  | <i>Monomorium pharaonis</i> | pharaoh ant           | 14,019                 | 5,362 (38.2)                  | 0.91                 | <b>8,367 (59.7)</b>           | <b>4.47</b>          |
| GCF_003672135.1  | <i>Ooceraea biroi</i>       | clonal raider ant     | 11,927                 | 5,359 (44.9)                  | 0.91                 | <b>8,312 (69.7)</b>           | <b>4.36</b>          |
| GCF_009193385.2  | <i>Nasonia vitripennis</i>  | jewel wasp            | 13,602                 | 5,306 (39.0)                  | 0.88                 | <b>7,730 (56.8)</b>           | <b>2.91</b>          |
| GCF_013357705.1  | <i>Chelonus insularis</i>   |                       | 10,548                 | 5,458 (51.7)                  | 0.92                 | <b>7,846 (74.4)</b>           | <b>2.68</b>          |

(a)

```
"""
Simplified decision logic for classifying a pair of putative orthologs.
```

Input:

```
hp: A candidate homolog pair with similarity scores and coverage metrics
top: A synthetic homolog pair constructed using maximum scores for each metric
    observed among all homolog pairs involving one of query or subject genes,
    except the homolog pair under consideration
```

Output:

Boolean value indicating if the pair should be classified as orthologs

Algorithm:

```
"""
@dataclass
class HomologPair:
    msy_sc      : int      # microsynteny score
    nuc_sc      : float    # nucleotide (exome-space) alignment score
    prot_sc     : float    # pairwise protein sequence similarity score
    prot_aln_len : int      # length of the protein alignment
    q_cov       : float    # prot_aln_len / query protein length
    s_cov       : float    # prot_aln_len / subject protein length
    is_top_scoring_for_query_or_subject: bool
                                     # whether prot_sc is maximal over other protein
                                     # alignments involving these query and subject

def is_orthologs_pair(
    hp : HomologPair, # Candidate pair of potentially orthologous genes.
    top : HomologPair # Synthetic pair of orthologs with maximum possible
                      # scores for each metric among all homolog pairs
                      # involving either query or subject,
                      # excluding the current pair under consideration
) -> bool :
    is_nuc_maximal    = hp.nuc_sc >= top.nuc_sc
    is_prot_maximal   = hp.prot_sc >= top.prot_sc
    has_nuc_support    = 0.95 * hp.nuc_sc > top.nuc_sc
    has_prot_support   = 0.95 * hp.prot_sc > top.prot_sc \
                        and hp.is_top_scoring_for_query_or_subject
    is_high_coverage = (
        min(hp.q_cov, hp.s_cov) > 0.5
        and max(hp.q_cov, hp.s_cov) > 0.9
    )

    has_microsynteny_support = (
        (hp.msy_sc > 0 and top.msy_sc == 0 and is_prot_maximal)
        or (hp.msy_sc > 1 and top.msy_sc == 0)
        or (hp.msy_sc > top.msy_sc + 1 and is_nuc_maximal)
        or (hp.msy_sc > top.msy_sc and is_nuc_maximal and has_prot_support)
    )

    # stricter criteria for a case without microsynteny support
    is_well_supported_outlier = (
        not hp.query.is_possible_pseudogene
        and not hp.subject.is_possible_pseudogene
        and top.msy_sc == 0
        and is_high_coverage
        and has_nuc_support
        and has_prot_support
    )

    return has_microsynteny_support or is_well_supported_outlier
```

(b)

|                           | Homolog pair 1         | Homolog pair 2         | Homolog pair 3         | Homolog pair 4      |
|---------------------------|------------------------|------------------------|------------------------|---------------------|
| q_org                     | <i>Mus musculus</i>    | <i>Mus musculus</i>    | <i>Mus musculus</i>    | <i>Mus musculus</i> |
| s_org                     | <i>Homo sapiens</i>    | <i>Homo sapiens</i>    | <i>Homo sapiens</i>    | <i>Homo sapiens</i> |
| q_geneid<br>(Gene symbol) | 100504234<br>(Ccgc170) | 100504234<br>(Ccgc170) | 100504234<br>(Ccgc170) | 80384<br>(Tex21)    |
| s_geneid<br>(Gene symbol) | 80129<br>(CCDC170)     | 728047<br>(GOLGA80)    | 643699<br>(GOLGA8N)    | 80129<br>(CCDC170)  |
| q_multi                   | 3                      | 3                      | 3                      | 1                   |
| s_multi                   | 2                      | 5                      | 5                      | 2                   |
| q_prot                    | NP_001357921.1         | NP_001357922.1         | NP_001357922.1         | XP_006516456.1      |
| s_prot                    | NP_079335.2            | XP_047288962.1         | NP_001269423.1         | XP_047275328.1      |
| q_cov                     | 0.9612                 | 0.4789                 | 0.5331                 | 0.929               |
| s_cov                     | 0.9706                 | 0.5137                 | 0.5601                 | 0.7469              |
| prot_aln_len              | 694                    | 318                    | 354                    | 484                 |
| prot_sc                   | 0.506                  | 0.0306                 | 0.0325                 | 0.0749              |
| nuc_sc                    | 0.1275                 | 0                      | 0                      | 0                   |
| msy_sc                    | 14                     | 0                      | 0                      | 0                   |
| ortholog                  | T                      | F                      | F                      | F                   |

(c)

|                           | Homolog pair 1      | Homolog pair 2      | Homolog pair 3      | Homolog pair 4      |
|---------------------------|---------------------|---------------------|---------------------|---------------------|
| q_org                     | <i>Bos taurus</i>   | <i>Bos taurus</i>   | <i>Bos taurus</i>   | <i>Bos taurus</i>   |
| s_org                     | <i>Homo sapiens</i> | <i>Homo sapiens</i> | <i>Homo sapiens</i> | <i>Homo sapiens</i> |
| q_geneid<br>(Gene symbol) | 504248<br>(PRPS1L1) | 504248<br>(PRPS1L1) | 781227<br>(PRPS1)   | 781227<br>(PRPS1)   |
| s_geneid<br>(Gene symbol) | 221823<br>(PRPS1L1) | 5631<br>(PRPS1)     | 221823<br>(PRPS1L1) | 5631<br>(PRPS1)     |
| q_multi                   | 3                   | 3                   | 3                   | 3                   |
| s_multi                   | 3                   | 3                   | 3                   | 3                   |
| q_prot                    | NP_001095953.1      | NP_001095953.1      | NP_001039654.1      | NP_001039654.1      |
| s_prot                    | NP_787082.1         | NP_002755.1         | NP_787082.1         | NP_002755.1         |
| q_cov                     | 0.9635              | 0.9635              | 1.0000              | 1.0000              |
| s_cov                     | 0.9969              | 0.9969              | 1.0000              | 1.0000              |
| prot_aln_len              | 317                 | 317                 | 318                 | 318                 |
| prot_sc                   | 0.7304              | 0.8046              | 0.8936              | 1.0000              |
| nuc_sc                    | 0.1950              | 0.0948              | 0.0945              | 0.6113              |
| msy_sc                    | 15                  | 0                   | 0                   | 11                  |
| ortholog                  | T                   | F                   | F                   | T                   |

(d)

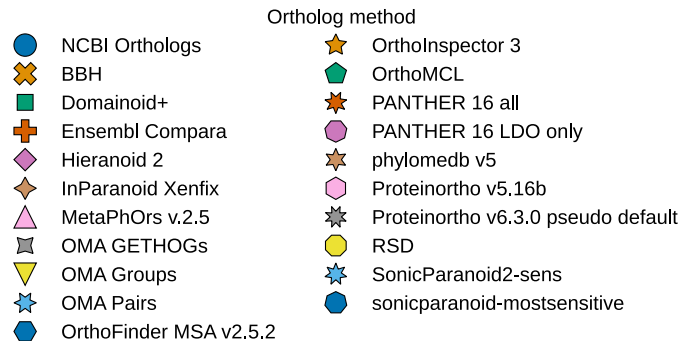
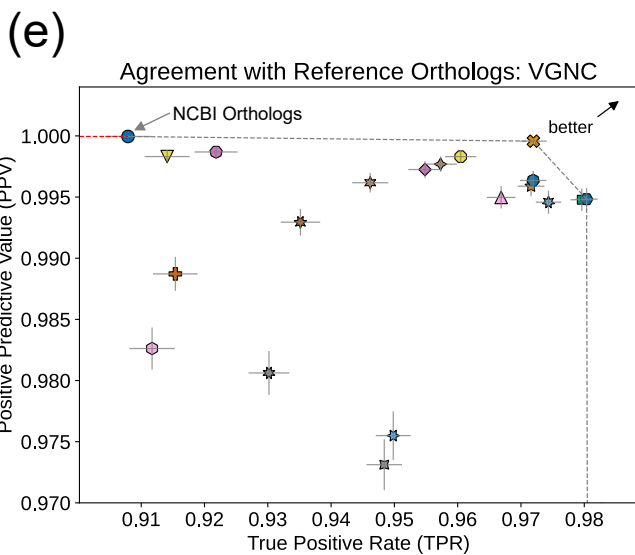
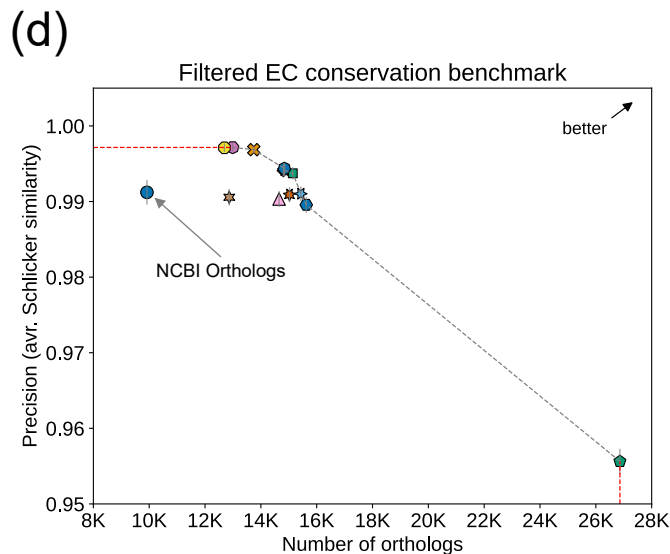
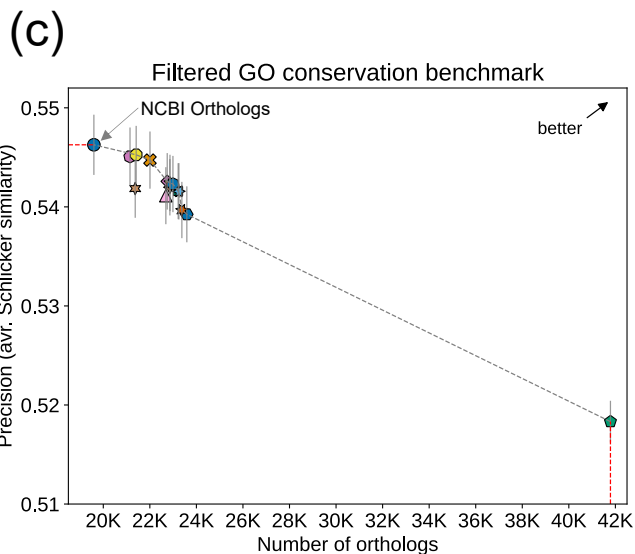
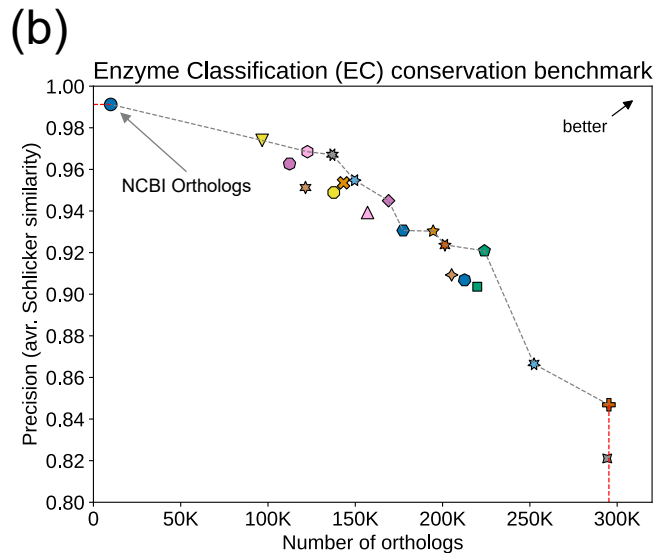
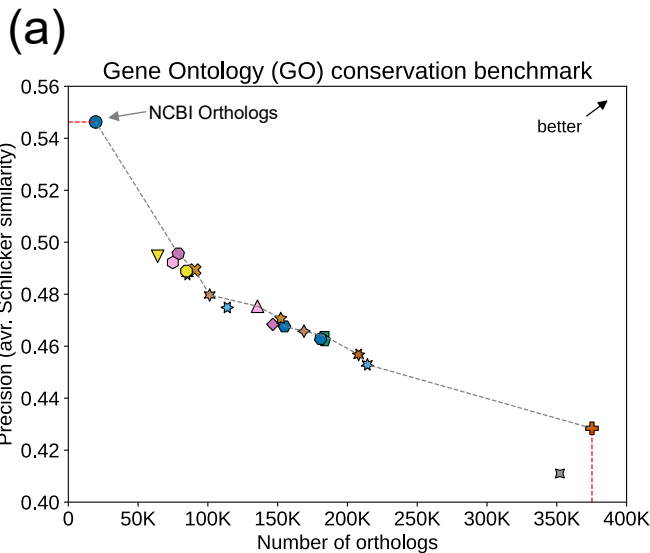
|                           | Homolog pair 1        | Homolog pair 2        | Homolog pair 3        | Homolog pair 4      |
|---------------------------|-----------------------|-----------------------|-----------------------|---------------------|
| q_org                     | <i>Bos taurus</i>     | <i>Bos taurus</i>     | <i>Bos taurus</i>     | <i>Bos taurus</i>   |
| s_org                     | <i>Homo sapiens</i>   | <i>Homo sapiens</i>   | <i>Homo sapiens</i>   | <i>Homo sapiens</i> |
| q_geneid<br>(Gene symbol) | 617475<br>(LOC617475) | 617475<br>(LOC617475) | 617475<br>(LOC617475) | 101902099<br>(F8A1) |
| s_geneid<br>(Gene symbol) | 8263<br>(F8A1)        | 474383<br>(F8A2)      | 474384<br>(F8A3)      | 8263<br>(F8A1)      |
| q_multi                   | 3                     | 3                     | 3                     | 3                   |
| s_multi                   | 2                     | 2                     | 2                     | 2                   |
| q_prot                    | XP_002699745.1        | XP_002699745.1        | XP_002699745.1        | XP_024844326.2      |
| s_prot                    | NP_036283.2           | NP_001007524.1        | NP_001007525.1        | NP_036283.2         |
| q_cov                     | 0.9946                | 0.9946                | 0.9946                | 0.9423              |
| s_cov                     | 0.9892                | 0.9892                | 0.9892                | 0.9677              |
| prot_aln_len              | 367                   | 367                   | 367                   | 359                 |
| prot_sc                   | 0.7463                | 0.7463                | 0.7463                | 0.7099              |
| nuc_sc                    | 0.1144                | 0.1144                | 0.1144                | 0.3802              |
| msy_sc                    | 6                     | 7                     | 7                     | 9                   |
| ortholog                  | F                     | F                     | F                     | T                   |

**Supplementary Fig. 1** Pseudocode to identify orthologs among homolog pairs and examples of metrics produced by the pipeline to illustrate the choice of orthologs. (a) shows the decision process to determine if a homolog pair is an ortholog. The homolog pair must meet the criteria for either the “has\_microsynteny\_support” or “is\_well\_supported\_outlier”, to qualify as an ortholog pair.

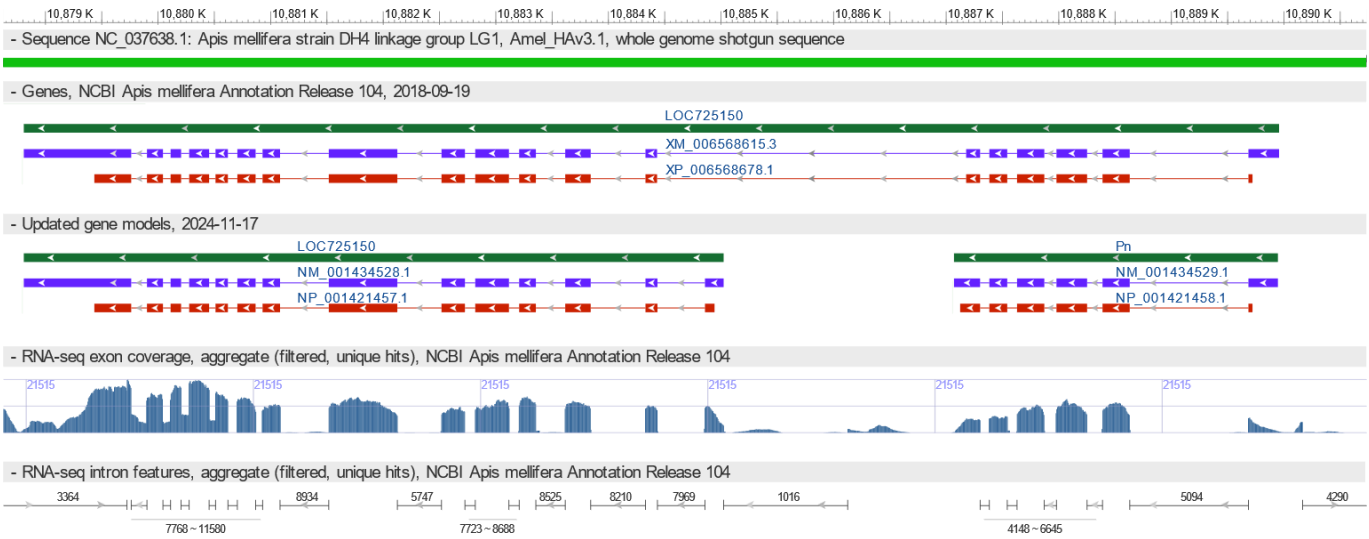
(b) This figure illustrates the evaluation of a mouse-human homolog pair (“Homolog pair 1”) for orthology. The mouse gene (q\_geneid = 100504234) had three homolog pairs (q\_multi = 3) with human genes 80129, 728047, and 643699. Conversely, the human gene (s\_geneid = 80129) was part of two homolog pairs (s\_multi = 2) with mouse genes 100504234 and 80384. Homolog pair 1 (100504234, 80129) was assessed against all other homolog pairs involving either gene. Homolog pair 1 exhibited superior (bold and in blue color) protein similarity score (prot\_sc), nucleotide sequence alignment score (nuc\_sc), and microsynteny score (msy\_sc) compared to the competing homolog pairs. Note that homologous gene pairs can be represented by different protein isoforms (q\_prot and s\_prot), which scored the highest protein similarity among all isoform pairs. Gene symbols were shown in parentheses together with the locus IDs. For homolog pairs sharing a member with the Homolog pair 1, scores superior and inferior compared to Homolog pair 1 were marked blue and red, respectively.

(c) The same table as (b) for a case where Homolog pair 1 (504248, 221823) between human and cattle was identified as an ortholog pair based on nucleotide sequence alignment score and strong microsynteny conservation despite lower protein similarity score compared to Homolog pair 2 (504248, 5631). Human gene 5631 (PRPS1) had another cattle homolog (781227), with the identical protein sequences and strong microsynteny signal, declared as an ortholog.

(d) In this example, the human genome included three paralogs F8A1/2/3 (coagulation factor VIII-associated 1-like 1/2/3), with F8A2 and F8A3 located adjacent to each other and F8A1 separated from F8A2/3 by several loci. All three encoded proteins with identical sequences. The cattle gene 617475 cannot be unambiguously paired with a human gene locus due to identical scores with F8A2 and F8A3 and lack of stronger microsynteny signal with F8A1. On the other hand, its paralog 101902099 was paired to human F8A1 as orthologs based on well conserved genes in the neighborhood (Homolog pair 4).

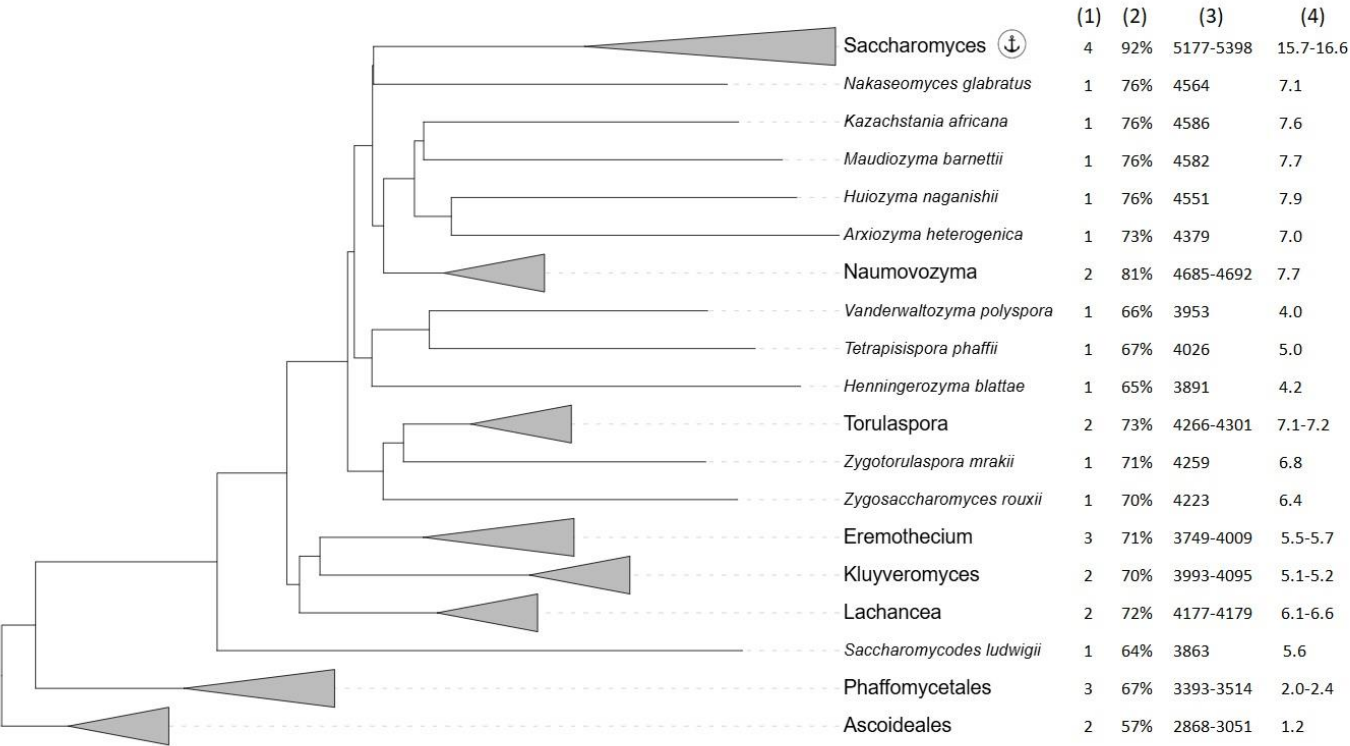


**Supplementary Fig. 2** We evaluated NCBI Orthologs method using the Orthology Benchmarking service provided by the Quest for Orthologs (QfO) consortium (Altenhoff, Nevers, et al. 2024). Based on the April 2020 reference proteome datasets, we calculated orthologs between fishes versus zebrafish, other vertebrates versus human, and insects versus fruit fly as detailed in Methods and Supplementary Table 1. Orthologs of the same anchor gene were transitively inferred as orthologs among themselves. Publicly available results for twenty orthology methods were plotted together with NCBI Orthologs for the Gene Ontology (GO) conservation (**a**, **c**), Enzyme Classification (EC) conservation (**b**, **d**), and Agreement with Reference Orthologs: VGNC benchmark (**e**). Pareto fronts are indicated by dashed lines. In (**c**) and (**d**), GO and EC conservation benchmarks were run on a filtered set of predicted ortholog pairs. This set comprised data from all QfO 2020 methods whose ortholog pairs are available in simple tabular format. These ortholog pairs were then further subsetting to include only species pairs evaluated by our pipeline. This ensures a fair comparison and mitigates any bias from functional conservation varying with evolutionary distance. In (**e**), the Pareto front extended to the OrthoMCL data point ( $x=0.9828$ ,  $y=0.4751$ ) plotted below the y-axis range. See Nevers, et al. 2022 for details on how precision (y-axes), recall (x-axes), and their standard errors (error bars) were determined for each test. The raw QfO test results for the three tests plotted in (a—e), as well as the Agreement with Reference Gene Phylogenies: SwissTree test, are described in Supplementary Data 1.

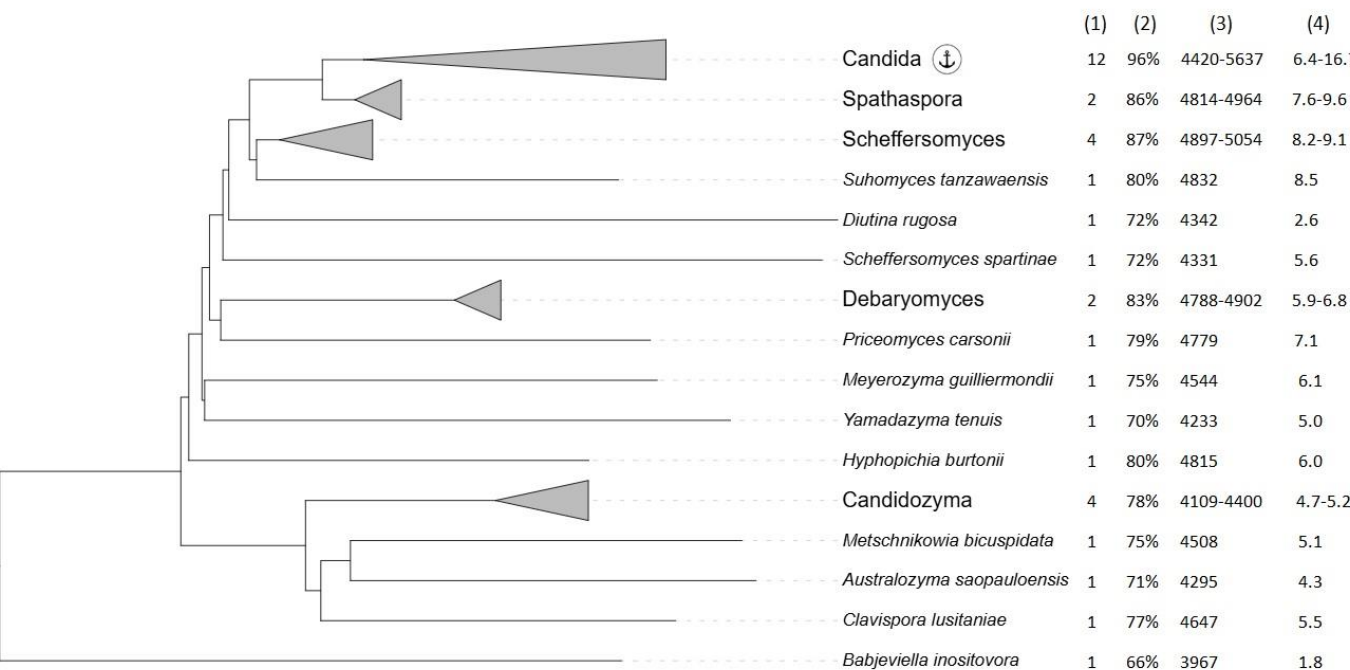


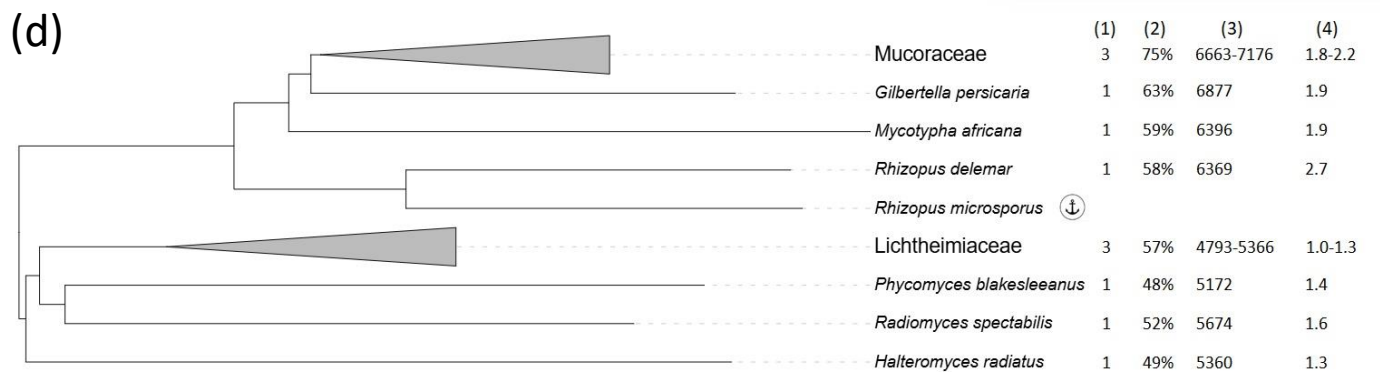
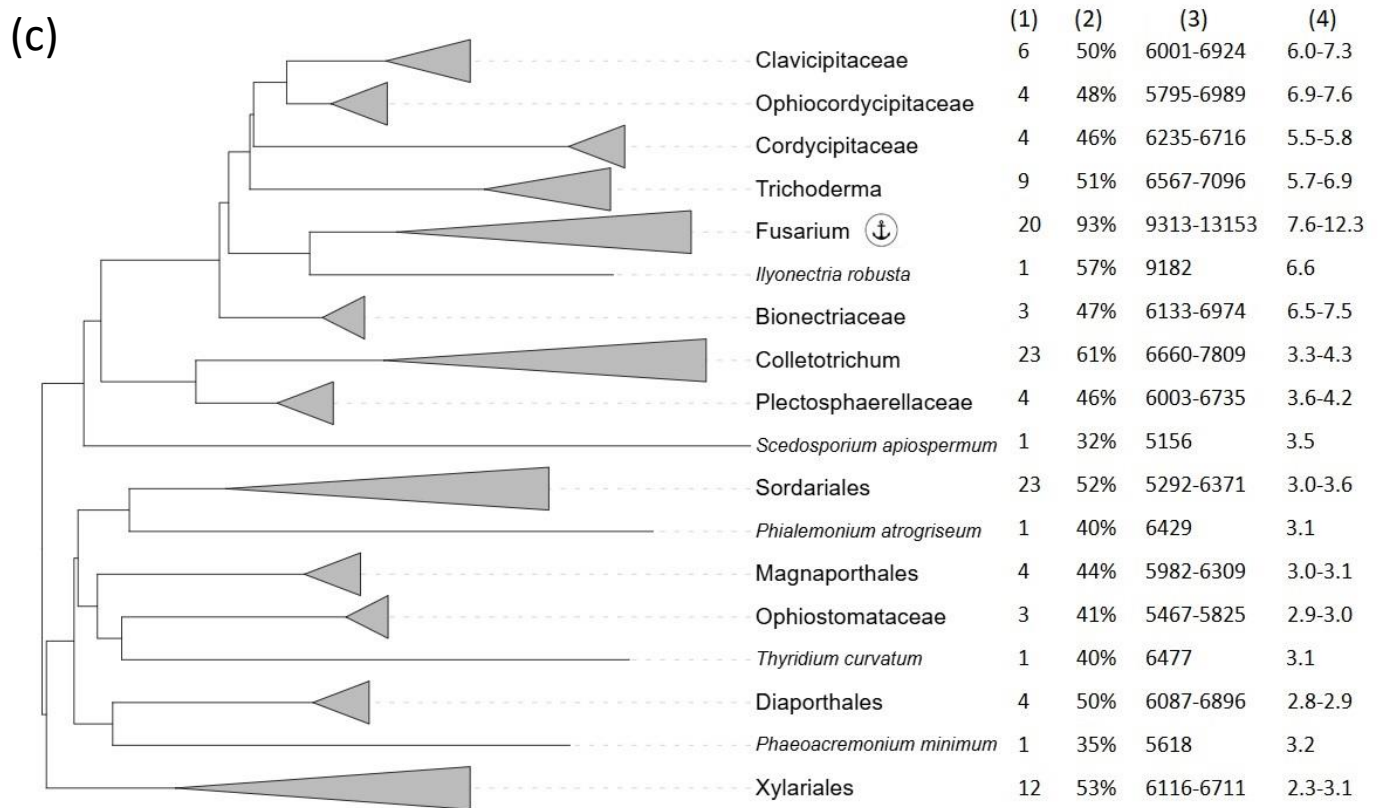
**Supplementary Fig. 3** Curation of chimeric gene model in *Apis mellifera* annotation. Genome Workbench view of *Apis mellifera* chromosome LG1 (assembly GCF\_003254395.2) comparing the original Annotation Release 104 annotation of gene LOC725150 with its currently public, updated version. Annotation Release 104 included a chimeric gene model for LOC725150, encompassing an adjacent gene. The RefSeq curation team resolved this by: (1) suppressing the existing chimeric transcript (XM\_006568615.3); (2) creating a new transcript (NM\_001434528.1) to accurately represent LOC725150; and (3) establishing a new Pn gene with its transcript (NM\_001434529.1). Bottom tracks display RNA-seq evidence, including exon coverage data and reads spanning adjacent introns, supporting the corrected annotation.

(a)



(b)





**Supplementary Fig. 4** Excerpts from the Fungi genome-based protein distance tree showing relationships and ortholog statistics: (1) number of assemblies compared; (2) percentage of anchor proteins with orthologs; (3) ortholog count range (min—max) and (4) average orthologous neighbor count range (min—max), for the comparison anchor and RefSeq data in each taxonomic group: **(a)** *Saccharomyces cerevisiae* strain S288C proteins (16,202) on assembly GCF\_000146045.2 compared within the *Saccharomycetes* (budding yeasts & others); **(b)** *Candida albicans* strain SC5314 proteins (6,030) on assembly GCF\_000182965.3 compared within the *Serinales* (CUG-Ser1 clade in the class Pichiomycetes); **(c)** *Fusarium oxysporum* strain Fo47 proteins (16,202) on assembly GCF\_013085055.1 compared within the *Sordariomycetes* (*a filamentous ascomycetes class*); **(d)** *Rhizopus microsporus* strain ATCC 52813 proteins (10,888) on assembly GCF\_002708625.1 compared within the *Mucorales* (*pin molds*).