



Bacterial contribution to genesis of the novel germ line determinant *oskar*

Leo Blondel¹, Tamsin EM Jones^{2†}, Cassandra G Extavour^{1,2*}

¹Department of Molecular and Cellular Biology, Harvard University, Cambridge, United States; ²Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, United States

Abstract New cellular functions and developmental processes can evolve by modifying existing genes or creating novel genes. Novel genes can arise not only via duplication or mutation but also by acquiring foreign DNA, also called horizontal gene transfer (HGT). Here we show that HGT likely contributed to the creation of a novel gene indispensable for reproduction in some insects. Long considered a novel gene with unknown origin, *oskar* has evolved to fulfil a crucial role in insect germ cell formation. Our analysis of over 100 insect Oskar sequences suggests that *oskar* arose *de novo* via fusion of eukaryotic and prokaryotic sequences. This work shows that highly unusual gene origin processes can give rise to novel genes that may facilitate evolution of novel developmental mechanisms.

Introduction

***For correspondence:**
extavour@oeb.harvard.edu

Present address: [†]European Bioinformatics Institute, EMBL-EBI, Wellcome Genome Campus, Hinxton, United Kingdom

Competing interests: The authors declare that no competing interests exist.

Funding: See page 11

Received: 29 January 2019

Accepted: 23 February 2020

Published: 24 February 2020

Reviewing editor: Antonis Rokas, Vanderbilt University, United States

© Copyright Blondel et al. This article is distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use and redistribution provided that the original author and source are credited.

Heritable variation is the raw material of evolutionary change. Genetic variation can arise from mutation and gene duplication of existing genes ([Taylor and Raes, 2004](#)), or through *de novo* processes ([Tautz and Domazet-Lošo, 2011](#)), but the extent to which such novel, or 'orphan' genes participate significantly in the evolutionary process is unclear. Mutation of existing cis-regulatory ([Wittkopp and Kalay, 2012](#)) or protein coding regions ([Hoekstra and Coyne, 2007](#)) can drive evolutionary change in developmental processes. However, recent studies in animals and fungi suggest that novel genes can also drive phenotypic change ([Chen et al., 2013](#)). Although counterintuitive, novel genes may be integrating continuously into otherwise conserved gene networks, with a higher rate of partner acquisition than subtler variations on preexisting genes ([Zhang et al., 2015](#)). Moreover, in humans and fruit flies, a large proportion of novel genes are expressed in the brain, suggesting their participation in the evolution of major organ systems ([Zhang et al., 2012; Chen et al., 2012](#)). However, while next generation sequencing has improved their discovery, the developmental and evolutionary significance of novel genes remains understudied.

The mechanism of formation of a novel gene may have implications for its function. Novel genes that arise by duplication, thus possessing the same biophysical properties as their parent genes, have innate potential to participate in preexisting cellular and molecular mechanisms ([Taylor and Raes, 2004](#)). However, orphan genes lacking sequence similarity to existing genes must form novel functional molecular relationships with extant genes, in order to persist in the genome. When such genes arise by introduction of foreign DNA into a host genome through horizontal gene transfer (HGT), they may introduce novel, already functional sequence information into a genome. Whether genes created by HGT show a greater propensity to contribute to or enable novel processes is unclear. Endosymbionts in the host germ line cytoplasm (germ line symbionts) could increase the occurrence of evolutionarily relevant HGT events, as foreign DNA integrated into the germ line genome is transferred to the next generation. HGT from bacterial endosymbionts into insect genomes appears widespread, involving transfer of metabolic genes or even larger genomic

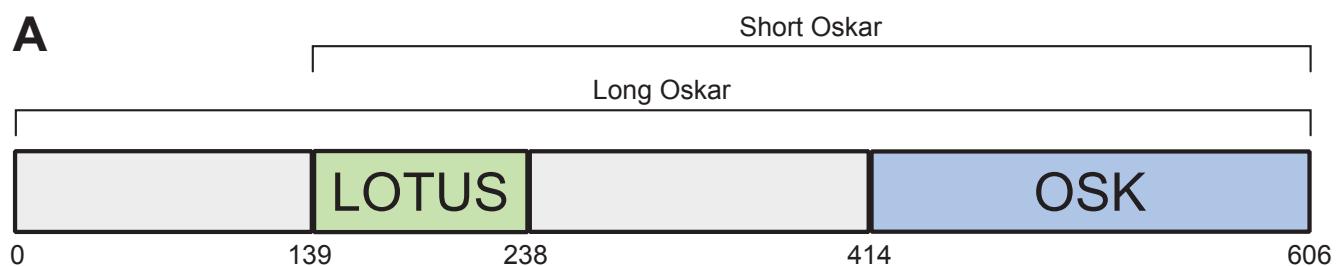
fragments to the host genome (see for example [Dunning Hotopp et al., 2007](#); [Acuna et al., 2012](#); [Sloan et al., 2014](#); [Husnik et al., 2013](#)).

Here we examined the evolutionary origins of the *oskar* (*osk*) gene, long considered a novel gene that evolved to be indispensable for insect reproduction ([Lehmann, 2016](#)). First discovered in *Drosophila melanogaster* ([Lehmann and Nüsslein-Volhard, 1986](#)), *osk* is necessary and sufficient for assembly of germ plasm, a cytoplasmic determinant that specifies the germ line in the embryo. Germ plasm-based germ line specification appears derived within insects, confined to insects that undergo metamorphosis (Holometabola) ([Ewen-Campen et al., 2012](#); [Extavour and Akam, 2003](#)). Initially thought exclusive to Diptera (flies and mosquitoes), its discovery in a wasp, another holometabolous insect with germ plasm ([Lynch et al., 2011](#)), led to the hypothesis that *oskar* originated as a novel gene at the base of the Holometabola approximately 300 Mya, facilitating the evolution of insect germ plasm as a novel developmental mechanism ([Lynch et al., 2011](#)). However, its subsequent discovery in a cricket ([Ewen-Campen et al., 2012](#)), a hemimetabolous insect without germ plasm ([Ewen-Campen et al., 2013](#)), implied that *osk* was instead at least 50 My older, and that its germ plasm role was derived rather than ancestral ([Abouheif, 2013](#)). Despite its orphan gene status, *osk* plays major developmental roles, interacting with the products of many genes highly conserved across animals ([Lehmann, 2016](#); [Jeske et al., 2015](#); [Jeske et al., 2017](#)). *osk* thus represents an example of a novel gene that not only functions within pre-existing gene networks in the nervous system ([Ewen-Campen et al., 2012](#)), but has also evolved into the only animal gene that has been experimentally demonstrated to be both necessary and sufficient to specify functional primordial germ line cells ([Ephrussi and Lehmann, 1992](#); [Kim-Ha et al., 1991](#)).

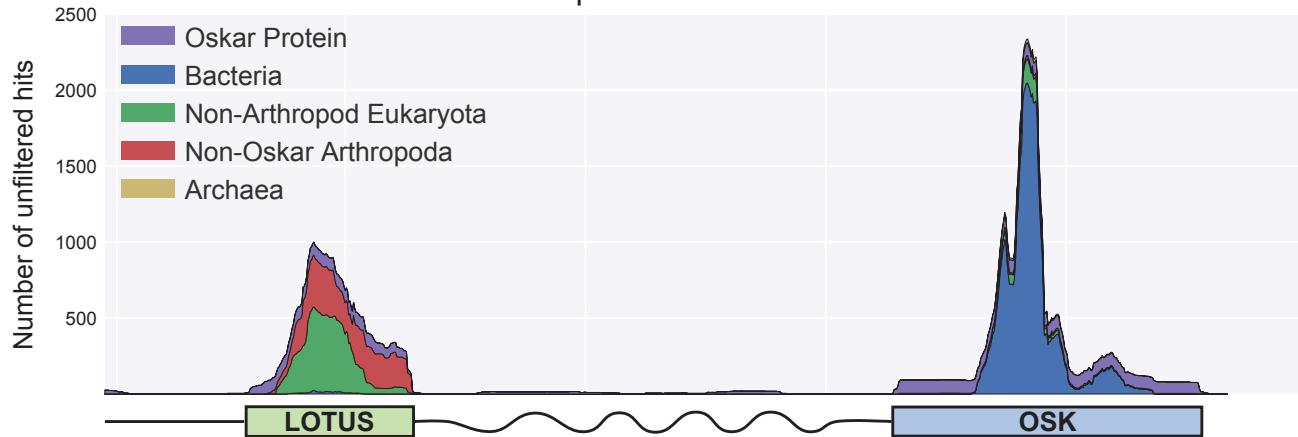
The evolutionary origins of this remarkable gene are unknown. Osk contains two biophysically conserved domains, an N-terminal LOTUS domain and a C-terminal hydrolase-like domain called OSK ([Jeske et al., 2015](#); [Yang et al., 2015](#); [Figure 1a](#)). An initial BLASTp search using the full-length *D. melanogaster* *osk* sequence as a query yielded either other holometabolous insect *osk* genes, or partial hits for the LOTUS or OSK domains (E-value < 0.01; [Source data 1](#): BLAST search results). This suggested that full length *osk* was unlikely to be a duplication of any other known gene. This prompted us to perform two more BLASTp searches, one using each of the two conserved Osk protein domains individually as query sequences. Strikingly, in this BLASTp search, although we recovered several eukaryotic hits for the LOTUS domain, we recovered no eukaryotic sequences that resembled the OSK domain, even with very low E-value stringency (E-value < 10; see Materials and methods section “BLAST searches of *oskar*” for an explanation of E-value threshold choices; [Source data 1](#): BLAST search results).

To understand this anomaly, we built an alignment of 95 Oskar sequences ([Source data 1 Alignments>OSKAR_MUSCLE_FINAL.fasta](#); [Supplementary file 1A and B](#)) and used a custom iterative HMMER sliding window search tool to compare each domain with protein sequences from all domains of life. Sequences most similar to the LOTUS domain were almost exclusively eukaryotic sequences ([Supplementary file 1C](#)). In contrast, those most similar to the OSK domain were bacterial, specifically sequences similar to SGNH-like hydrolases ([Jeske et al., 2015](#); [Yang et al., 2015](#)) (Pfam Clan: SGNH_hydrolase - CL0264; [Supplementary file 1D](#); [Figure 1b](#)). To visualize their relationships, we graphed the sequence similarity network for the sequences of these domains and their closest hits. We observed that the majority of LOTUS domain sequences clustered within eukaryotic sequences ([Figure 1c](#)). In contrast, OSK domain sequences formed an isolated cluster, a small subset of which formed a connection to bacterial sequences ([Figure 1d](#)). These data are consistent with a previous suggestion, based on BLAST results ([Lynch et al., 2011](#)), that HGT from a bacterium into an ancestral insect genome may have contributed to the evolution of *osk*. However, this possibility was not formally addressed by previous analyses, which were based on alignments of full length Osk containing only eukaryotic sequences as outgroups ([Ewen-Campen et al., 2012](#)). To rigorously test this hypothesis, we therefore performed phylogenetic analyses of the two domains independently. A finding that LOTUS sequences were nested within eukaryotes, while OSK sequences were nested within bacteria, would provide support for the HGT hypothesis.

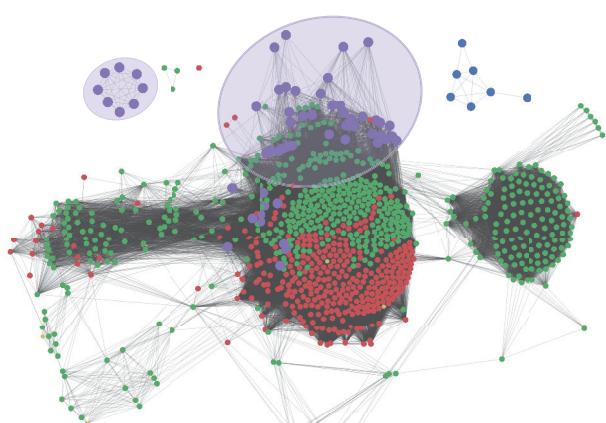
Both Maximum likelihood and Bayesian approaches confirmed this prediction ([Figure 2a](#), [Figure 2—figure supplements 1 and 2](#)), and these results were robust to changes in the methods of sequence alignment ([Figure 2—figure supplements 6, 7, 8, 9, 10](#)). As expected, LOTUS sequences from Osk proteins were related to other eukaryotic LOTUS domains, to the exclusion of the only three bacterial sequences that met our E-value cutoff for inclusion in the analyses ([Figure 2a](#),



B Domain of life identity of HMMER hits against Uniprot Trembl database

**C**

LOTUS Sequence similarity network

**D**

OSK Sequence similarity network

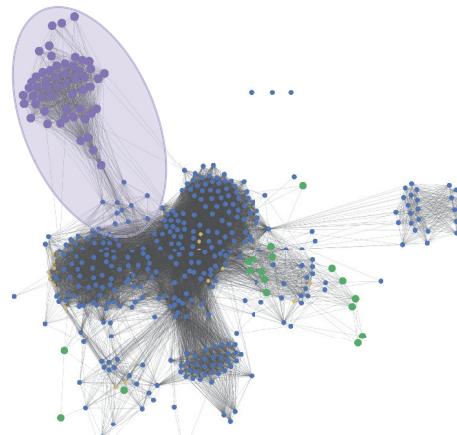


Figure 1. Sequence analysis of the Oskar gene. (a) Schematic representation of the Oskar gene. The LOTUS and OSK hydrolase-like domains are separated by a poorly conserved region of predicted high disorder and variable length between species. In some dipterans, a region 5' to the LOTUS domain is translated to yield a second isoform, called Long Oskar. Residue numbers correspond to the *D. melanogaster* Osk sequence. (b) Stackplot of domain of life identity of HMMER hits across the protein sequence. For a sliding window of 60 Amino Acids across the protein sequence (X axis), the number of hits in the Trembl (UniProt) database (Y axis) is represented and color coded by domain of life origin (see Materials and methods: Iterative HMMER search of OSK and LOTUS domains), stacked on top of each other. (c, d) EFI-EST-generated graphs of the sequence similarity network of the

Figure 1 continued on next page

Figure 1 continued

LOTUS (c) and OSK (d) domains of Oskar (**Gerlt et al., 2015**). Sequences were obtained using HMMER against the UniProtKB database. Most Oskar LOTUS sequences cluster within eukaryotes and arthropods. In contrast, Oskar OSK sequences cluster most strongly with a small subset of bacterial sequences.

Figure 2—figure supplements 1 and 2; see Materials and methods and Supplemental Text). LOTUS sequences from non-Oskar proteins were almost exclusively eukaryotic. (**Supplementary file 1**); only three bacterial sequences matched the LOTUS domain with an E-value < 0.01. Osk LOTUS domains clustered into two distinct clades, one comprising all Dipteran sequences, and the other comprising all other Osk LOTUS domains examined from both holometabolous and hemimetabolous orders (**Figure 2a**). Dipteran Osk LOTUS sequences formed a monophyletic group that branched sister to a clade of LOTUS domains from Tud5 family proteins of non-arthropod animals (NAA). NAA LOTUS domains from Tud7 family members were polyphyletic, but most of them formed a clade branching sister to (Osk LOTUS + NAA Tud5 LOTUS). Non-Dipteran Osk LOTUS domains formed a monophyletic group that was related in a polytomy to the aforementioned (NAA Tud7 LOTUS + (Dipteran Osk LOTUS + NAA Tud5 LOTUS)) clade, and to various arthropod Tud7 family LOTUS domains.

The fact that Tud7 LOTUS domains are polyphyletic suggests that arthropod domains in this family may have evolved differently than their homologues in other animals. The relationships of Dipteran LOTUS sequences were consistent with the current hypothesis for interrelationships between Dipteran species (**Kirk-Spriggs and Sinclair, 2017**). Similarly, among the non-Dipteran Osk LOTUS sequences, the hymenopteran sequences form a clade to the exclusion of the single hemimetabolous sequence (from the cricket *Gryllus bimaculatus*), consistent with the monophyly of Hymenoptera (**Peters et al., 2017**). It is unclear why Dipteran Osk LOTUS domains cluster separately from those of other insect Osk proteins. We speculate that the evolution of the Long Oskar domain (**Vanzo and Ephrussi, 2002; Hurd et al., 2016**), which appears to be a novelty within Diptera (**Source data 1: Alignments>OSKAR_MUSCLE_FINAL.fasta**), may have influenced the evolution of the Osk LOTUS domain in at least some of these insects. Consistent with this hypothesis, of the 17 Dipteran *oskar* genes we examined, the seven *oskar* genes possessing a Long Osk domain clustered into two clades based on the sequences of their LOTUS domain. One of these clades comprised five *Drosophila* species (*D. willistoni*, *D. mojavensis*, *D. virilis*, *D. grimshawi* and *D. immigrans*), and the second was composed of two calyptrate flies from different superfamilies, *Musca domestica* (Muscoidae) and *Lucilia cuprina* (Oestroidea).

In summary, the LOTUS domain of Osk proteins is most closely related to a number of other LOTUS domains found in eukaryotic proteins, as would be expected for a gene of animal origin, and the phylogenetic interrelationships of these sequences are largely consistent with the current species or family level trees for the corresponding insects.

In contrast, OSK domain sequences were nested within bacterial sequences (**Figure 2b, Figure 2—figure supplements 3 and 4**). This bacterial, rather than eukaryotic, affinity of the OSK domain was recovered even when different sequence alignment methods were used (**Figure 2—figure supplements 7, 8, 9, 10 and 11**). The only eukaryotic proteins emerging from the iterative HMMER search for OSK domain sequences that had an E-value < 0.01 were all from fungi. All five of these sequences were annotated as Carbohydrate Active Enzyme 3 (CAZ3), and all CAZ3 sequences formed a clade that was sister to a clade of primarily Firmicutes. Most bacterial sequences used in this analysis were annotated as lipases and hydrolases, with a high representation of GDSL-like hydrolases (**Supplementary file 1D**). OSK sequences formed a monophyletic group but did not branch sister to the other eukaryotic sequences in the analysis. Within this OSK clade, the topology of sequence relationships was largely concordant with the species tree for insects (**Misof et al., 2014**), as we recovered monophyletic Diptera to the exclusion of other insect species. However, the single orthopteran OSK sequence (from the cricket *G. bimaculatus*) grouped within the Hymenoptera, rather than branching as sister to all other insect sequences in the tree, as would be expected for this hemimetabolous sequence (**Misof et al., 2014**).

Importantly, OSK sequences did not simply form an outgroup to bacterial sequences. To formally reject the possibility that the eukaryotic OSK clade has a sister group relationship to all bacterial sequences in the analysis, we performed topology constraint analyses using the Swofford–Olsen–Waddell–Hillis (SOWH) test, which assigns statistical support to alternative phylogenetic topologies

● Oskar Protein ● Bacteria ● Non-Arthropod Eukaryota ● Non-Oskar Arthropoda ● Archaea
 Node Support Values ○ >50 ○ >60 ● >80 ● >90

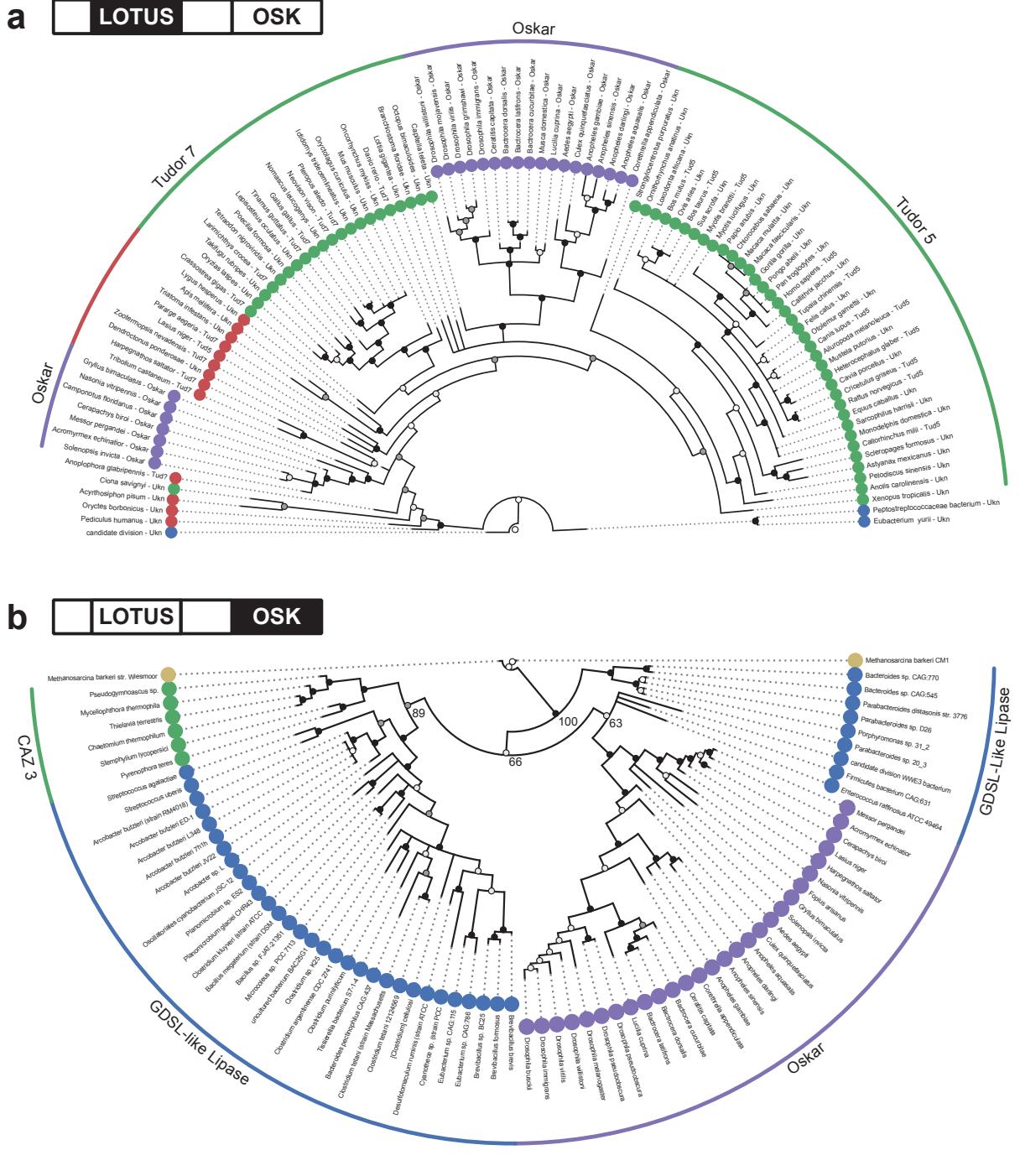


Figure 2. Phylogenetic analysis of the LOTUS and OSK domains. (a) Bayesian consensus tree for the LOTUS domain. Three major LOTUS-containing protein families are represented within the tree: Tudor 5, Tudor 7, and Oskar. Oskar LOTUS domains form two clades, one containing only dipterans and one containing all other represented insects (hymenopterans and orthopterans). The tree was rooted to the three bacterial sequences added in the dataset. (b) Bayesian consensus tree for the OSK domain. The OSK domain is nested within GDSL-like domains of bacterial species from phyla known

Figure 2 continued

to contain germ line symbionts in insects. The ten non-Oskar eukaryotic sequences in the analysis form a single clade comprising fungal Carbohydrate Active Enzyme 3 (CAZ3) proteins. For Bayesian and RaxML trees with all accession numbers and node support values see **Figure 2—figure supplements 1–4**.

The online version of this article includes the following figure supplement(s) for figure 2:

Figure supplement 1. LOTUS Domain RaxML MUSCLE Tree.

Figure supplement 2. LOTUS Domain Bayesian MUSCLE Tree.

Figure supplement 3. OSK Domain RaxML MUSCLE Tree.

Figure supplement 4. OSK Domain Bayesian MUSCLE Tree.

Figure supplement 5. SOWHAT constrained trees and results.

Figure supplement 6. LOTUS Domain RaxML PRANK Tree.

Figure supplement 7. OSK Domain RaxML PRANK Tree.

Figure supplement 8. OSK Tree PRANK Comparison.

Figure supplement 9. LOTUS Tree PRANK Comparison.

Figure supplement 10. LOTUS Domain RaxML T-Coffee Tree.

Figure supplement 11. OSK Domain RaxML T-Coffee Tree.

Figure supplement 12. OSK Tree T-Coffee Comparison.

Figure supplement 13. LOTUS Tree T-Coffee Comparison.

([Swofford et al., 1996](#)). We used the SOWHAT tool ([Church et al., 2015](#)) to compare the HGT-supporting topology to two alternative topologies with constraints more consistent with vertical inheritance. The first was constrained by domain of life, disallowing paraphyletic relationships between sequences from the same domain of life (**Figure 2—figure supplement 5a**). The second required monophyly of Eukaryota but allowed paraphyletic relationships between bacterial and archaeal sequences (**Figure 2—figure supplement 5b**). We found that the topologies of both of these constrained trees were significantly worse than the result we had recovered with our phylogenetic analysis (**Figure 2—figure supplement 5**), namely that the closest relatives of the OSK domain were bacterial rather than eukaryotic sequences **Figure 2b**, **Figure 2—figure supplements 3 and 4**.

OSK sequences formed a well-supported clade nested within bacterial GDSL-like lipase sequences. The majority of these bacterial sequences were from the Firmicutes, a bacterial phylum known to include insect germ line symbionts ([Wheeler et al., 2013](#); [Chepkemoi et al., 2017](#)). All other sequences from classified bacterial species, including a clade branching as sister to all other sequences, belonged either to the Bacteroidetes or to the Proteobacteria. Members of both of these phyla are also known germ line symbionts of insects ([Dunning Hotopp et al., 2007](#); [Zchori-Fein et al., 2004](#)) and other arthropods ([Zchori-Fein and Perlman, 2004](#)). In sum, the distinct phylogenetic relationships of the two domains of Oskar are consistent with a bacterial origin for the OSK domain. Further, the specific bacterial clades close to OSK suggest that an ancient arthropod germ line endosymbiont could have been the source of a GDSL-like sequence that was transferred into an ancestral insect genome, and ultimately gave rise to the OSK domain of *oskar* (**Figure 3**).

While multiple mechanisms can give rise to novel genes, HGT is arguably among the least well understood, as it involves multiple genomes and ancient biotic interactions between donor and host organisms that are often difficult to reconstruct. In the case of *oskar*, however, the fact that both germ line symbionts ([Bourtzis and Miller, 2006](#)) and HGT events ([Dunning Hotopp et al., 2007](#)) are widespread in insects, provides a plausible biological mechanism consistent with our hypothesis that fusion of eukaryotic and bacterial domain sequences led to the birth of this novel gene. Under this hypothesis, this fusion would have taken place before the major diversification of insects, nearly 500 million years ago ([Misof et al., 2014](#)).

Once arisen, novel genes might be expected to disappear rapidly, given that pre-existing gene regulatory networks operated successfully without them ([Taylor and Raes, 2004](#)). However, it is clear that novel genes can evolve functional connections with existing networks, become essential ([Chen et al., 2010](#)), and in some cases lead to new functions ([Cornelis et al., 2012](#)) and contribute to phenotypic diversity ([Chen et al., 2013](#)). Even given the growing number of convincing examples of HGT from both prokaryotic and eukaryotic origins (see for example [Husnik and McCutcheon, 2018](#); [Di Lelio et al., 2019](#); [Wybouw et al., 2016](#); [Quispe-Humanquise et al., 2017](#)), some authors suspect that the contribution of horizontal gene transfer to the acquisition of novel traits has

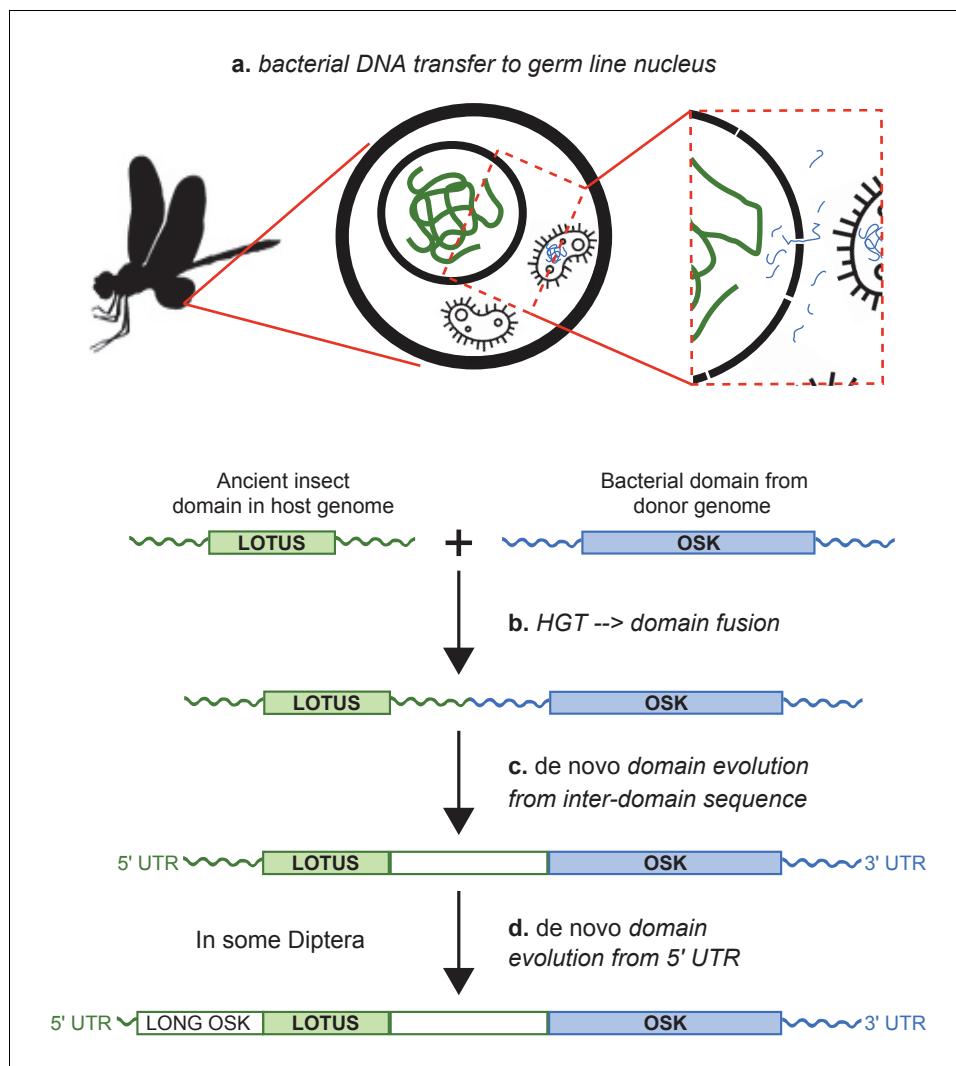


Figure 3. Hypothesis for the origin of *oskar*. Integration of the OSK domain close to a LOTUS domain in an ancestral insect genome. (a) DNA containing a GDSL-like domain from an endosymbiotic germ line bacterium is transferred to the nucleus of a germ cell in an insect common ancestor. (b) DNA damage or transposable element activity induces an integration event in the host genome, close to a pre-existing LOTUS-like domain. (c) The region between the two domains undergoes *de novo* coding evolution, creating an open reading frame with a unique, chimeric domain structure. (d) In some Diptera, including *D. melanogaster*, part of the 5' UTR of *oskar* has undergone *de novo* coding evolution to form the Long Oskar domain.

been underestimated across animals (Boto, 2014). Moreover, the functional contribution of genes horizontally transferred specifically from bacteria to insects has been documented for a range of adaptive phenotypes (see for example Wilson and Duncan, 2015; López-Madrigal and Gil, 2017; Provorov and Onishchuk, 2018), including digestive metabolism (Acuna et al., 2012; Sloan et al., 2014; Shelomi et al., 2016), glycolysis (Zeng et al., 2018) complex symbiosis (Husnik et al., 2013) and endosymbiont cell wall construction (Bublitz et al., 2019). *oskar* plays multiple critical roles in insect development, from neural patterning (Ewen-Campen et al., 2012; Xu et al., 2013) to oogenesis (Jenny et al., 2006). In the Holometabola, a clade of nearly one million extant species (Rees and Cranston, 2017), *oskar*'s co-option to become necessary and sufficient for germ plasm assembly is likely the cell biological mechanism underlying the evolution of this derived mode of insect germ line specification (Ewen-Campen et al., 2012; Lynch et al., 2011; Abouheif, 2013). Our study thus provides evidence that HGT can not only introduce functional genes into a host

genome, but also, by contributing sequences of individual domains, generate genes with entirely novel domain structures that may facilitate the evolution of novel developmental mechanisms.

Materials and methods

BLAST searches of Oskar

All BLAST (*Altschul et al., 1990*) searches were performed using the NCBI BLASTp tool suite on the non-redundant (nr) database. Amino Acid (AA) sequences of *D. melanogaster* full length Oskar (EMBL ID AAF54306.1), as well as the AA sequences for the *D. melanogaster* Oskar LOTUS (AA 139–238) and OSK (AA 414–606) domains were used for the BLAST searches. We used the default NCBI cut-off parameters (E-value cut-off of 10) for searches using OSK and LOTUS as queries, and a more stringent E-value threshold of 0.01 for the search using full length *D. melanogaster* Oskar as a query. We chose an E-value threshold of 10 for LOTUS and OSK to capture potentially highly divergent homologs of the two domains, especially for the OSK domain, where we were looking for any viable candidate for a homologous eukaryotic domain. All BLAST searches results are included in the **Source data 1: BLAST search results**.

Hidden Markov Model (HMM) generation and alignments of the OSK and LOTUS domains

101 1KITE transcriptomes (*Misof et al., 2014; Supplementary file 1A*) were downloaded and searched using the local BLAST program (BLAST+) using the tblastn algorithm with default parameters, with Oskar protein sequences of *Drosophila melanogaster*, *Aedes aegypti*, *Nasonia vitripennis* and *Gryllus bimaculatus* as queries (EntrezIDs: NP_731295.1, ABC41128.1, NP_001234884.1 and AFV31610.1 respectively). For all of these 1KITE transcriptome searches, predicted protein sequences from transcript data were obtained by in silico translation using the online ExPASy translate tool (<https://web.expasy.org/translate/>), taking the longest open reading frame. Publicly available sequences in the non-redundant (nr), TSA databases at NCBI, and a then-unpublished transcriptome (*Benton et al., 2016*) (kind gift of Matthew Benton and Siegfried Roth, University of Cologne) were subsequently searched using the web-based BLAST tool hosted at NCBI, using the tblastn algorithm with default parameters. Sequences used for queries were the four Oskar proteins described above, and newfound oskar sequences from the 1KITE transcriptomes of *Baetis pumilis*, *Cryptocercus wright*, and *Frankliniella cephalica*. For both searches, oskar orthologs were identified by the presence of BLAST hits on the same transcript to both the LOTUS (N-terminal) and OSK (C-terminal) regions of any of the query oskar sequences, regardless of E-values. The sequences found were aligned using MUSCLE (eight iterations) (*Edgar, 2004*) into a 46-sequence alignment (**Source data 1: Alignments > OSKAR_MUSCLE_INITIAL.fasta**). From this alignment, the LOTUS and OSK domains were extracted (**Source data 1: Alignments > LOTUS_MUSCLE_INITIAL.fasta** and **Alignments > OSK_MUSCLE_INITIAL.fasta**) to define the initial Hidden Markov Models (HMM) using the hmmbuild tool from the HMMER tool suite with default parameters (<http://hmmer.org/>; *Eddy, 2011*). 126 insect genomes and 128 insect transcriptomes (from the Transcriptome Shotgun Assembly TSA database: <https://www.ncbi.nlm.nih.gov/Traces/wgs/?view=TSAs>) were subsequently downloaded from NCBI (download date September 29, 2015; **Supplementary file 1A**). Genomes were submitted to Augustus v2.5.5 (*Stanke et al., 2004*) (using the *D. melanogaster* exon HMM predictor) and SNAP v2006-07-28 (*Korf, 2004*) (using the default 'fly' HMM) for gene discovery. The resulting nucleotide sequence database comprising all 309 downloaded and annotated genomes and transcriptomes, was then translated in six frames to generate a non-redundant amino acid database (where all sequences with the same amino acid content are merged into one). This process was automated using a series of custom scripts available here: <https://github.com/Xqua/Genomes>. The non-redundant amino acid database was searched using the HMMER v3.1 tool suite (*Eddy, 2011*) and the HMM for the LOTUS and OSK domains described above. A hit was considered positive if it consisted of a contiguous sequence containing both a LOTUS domain and an OSK domain, with the two domains separated by an inter-domain sequence. We imposed no length, alignment or conservation criteria on the inter-domain sequence, as this is a rapidly-evolving region of Oskar protein with predicted high disorder (*Jeske et al., 2015; Yang et al., 2015; Ahuja and Extavour, 2014*). Positive hits were manually curated and added to the main alignment, and the search was performed

iteratively until no more new sequences meeting the above criteria were discovered. This resulted in a total of 95 Oskar protein sequences, (see *Supplementary file 1B* for the complete list). Using the final resulting alignment (*Source data 1: Alignments > OSKAR_MUSCLE_FINAL.fasta*), the LOTUS and OSK domains were extracted from these sequences (*Source data 1: Alignments > LOTUS_MUSCLE_FINAL.fasta* and *Alignments > OSK_MUSCLE_FINAL.fasta*), and the final three HMM (for full-length Oskar, OSK, and LOTUS domains) used in subsequent analyses were created using hmmbuild with default parameters (*Source data 1: HMM >OSK.hmm, HMM >LOTUS.hmm and HMM >OSKAR.hmm*).

Iterative HMMER search of OSK and LOTUS domains

A reduced version of TrEMBL (*U Consortium, 2005*) (v2016-06) was created by concatenating all hits (regardless of E-value) for sequences of the LOTUS domain, the OSK domain and full-length Oskar, using hmmsearch with default parameters and the HMM models created above from the final alignment. This reduced database was created to reduce potential false positive results that might result from the limited size of the sliding window used in the search approach described here. The full-length Oskar alignment of 1133 amino acids (*Source data 1: Alignments > OSKAR_MUSCLE_FINAL.fasta*) was split into 934 sub-alignments of 60 amino acids each using a sliding window of one amino acid. Each alignment was converted into a HMM using hmmbuild, and searched against the reduced TrEMBL database using hmmsearch using default parameters. Domain of life origin of every hit sequence at each position was recorded. Eukaryotic sequences were further classified as Oskar/Non-Oskar and Arthropod/Non-Arthropod. Finally, for the whole alignment, the counts for each category were saved and plotted in a stack plot representing the proportion of sequences from each category to create *Figure 1b*. The python code used for this search is available at <https://github.com/Xqua/Iterative-HMMER>.

Sequence similarity networks

LOTUS and OSK domain sequences from the final alignment obtained as described above (see 'Hidden Markov Model (HMM) generation and alignments of the OSK and LOTUS domains'; *Source data 1: Alignments > LOTUS_MUSCLE_FINAL.fasta* and *Alignments > OSK_MUSCLE_FINAL.fasta*) were searched against TrEMBL (*U Consortium, 2005*) (v2016-06) using HMMER. All hits with E-value <0.01 were consolidated into a fasta file that was then entered into the EFI-EST tool (*Gerlt et al., 2015*) using default parameters to generate a sequence similarity network. An alignment score corresponding to 30% sequence identity was chosen for the generation of the final sequence similarity network. Finally, the network was graphed using Cytoscape 3 (*Shannon et al., 2003*).

Phylogenetic analysis based on MUSCLE alignment

For both the LOTUS and OSK domains, in cases where more than one sequence from the same organism was retrieved by the search described above in 'Iterative HMMER Search of OSK and LOTUS domains', only the sequence with the lowest E-value was used for phylogenetic analysis. For the LOTUS domain, the first 97 best hits (lowest E-value) were selected, and the only three bacterial sequences that satisfied an E-value <0.01 were manually added. For oskar sequences, if more than one sequence per species was obtained by the search, only the single sequence per species with the lowest E-value was kept for analysis, generating a set of 100 sequences for the LOTUS domain, and 87 sequences for the OSK domain. Unique identifiers for all sequences used to generate alignments for phylogenetic analysis are available in *Supplementary files 1C, 1D*. For both datasets, the sequences were then aligned using MUSCLE (*Edgar, 2004*) (eight iterations) and trimmed using trimAI (*Capella-Gutiérrez et al., 2009*) with 70% occupancy. The resulting alignments that were subject to phylogenetic analysis are available in *Source data 1: Alignments > LOTUS_MUSCLE_TREE.fasta* and *Alignments > OSK_MUSCLE_TREE.fasta*. For the maximum likelihood tree, we used RaxML v8.2.4 (*Stamatakis, 2014*) with 1000 bootstraps, and the models were selected using the automatic RaxML model selection tool. The substitution model chosen for both domains was LGF. For the Bayesian tree inference, we used MrBayes V3.2.6 (*Huelsenbeck and Ronquist, 2001*) with a Mixed model (prset aamodel = Mixed) and a gamma distribution (lset rates = Gamma). We ran the Monte-Carlo for 4 million generations (std <0.01) for the OSK domain, and for 3 million generations

(std <0.01) for the LOTUS domain. For the tree comparisons (**Figure 2—figure supplements 8, 9**), the RaxML best tree output from the MUSCLE and PRANK alignments were compared using the tool [Phylo.io](#) (Robinson et al., 2016).

Phylogenetic analysis based on PRANK alignment

For the OSK domain, the raw full length sequences obtained from the HMMER search were aligned to each other using the HMMER HMM-based alignment tool: hmalign, with the same HMM used to do the search, namely OSK.hmm (supplementary data: Data/HMM/OSK.hmm). Starting from this base alignment, we used the default alignment method option offered by PRANK (version: v.170427) (Löytynoja, 2014). We then used PRANK to realign those sequences, which in turn led to a usable alignment for phylogenetic analysis. This alignment was trimmed using the same parameters as described in *Hidden Markov Model (HMM) generation and alignments of the OSK and LOTUS domains* above. The final alignment is available in supplementary data: Alignment/OSK_prank_aligned.fasta. We then performed a phylogenetic analysis of this alignment using RAXML with the same parameters described in *Phylogenetic Analysis Based on MUSCLE Alignment* above. The resulting tree is presented in **Figure 2—figure supplements 7 and 8**.

For the LOTUS domain, the raw full length sequences obtained from the HMMER search were aligned to each other using the HMMER HMM-based alignment tool: hmalign, with the same HMM used to do the search, namely LOTUS.hmm (Supplementary data: Data/HMM/LOTUS.hmm). Starting from this base alignment, we then used PRANK with default options to realign those sequences. This alignment was trimmed using the same parameters as described in the *Hidden Markov Model (HMM) generation and alignments of the OSK and LOTUS domains*. The final alignment is available in supplementary data: Alignments/LOTUS_prank_aligned.fasta. We then performed a phylogenetic analysis using RAXML with the same parameters described above in *Phylogenetic Analysis Based on MUSCLE alignment*. The resulting trees are presented in **Figure 2—figure supplements 6 and 9**.

Phylogenetic analysis based on T coffee alignment

For the LOTUS and OSK domains, the raw full length sequences obtained from the HMMER search were aligned to each other using T-Coffee with its default parameters (Notredame et al., 2000). This alignment was trimmed using the same parameters as described in *Hidden Markov Model (HMM) generation and alignments of the OSK and LOTUS domains* above. The final alignment is available in supplementary data: Alignment/LOTUS_tcoffee_aligned.fasta Alignment/OSK_tcoffee_aligned.fasta. We then performed a phylogenetic analysis of this alignment using RAXML with the same parameters described in *Phylogenetic Analysis Based on MUSCLE Alignment* above. The resulting trees are presented in **Figure 2—figure supplements 10 and 11**.

Visual comparison of phylogenetic trees

To compare the trees obtained with different alignment tools, we used [Phylo.io](#) (Robinson et al., 2016). The trees were imported in Newick format, and the [Phylo.io](#) tool generated the mirrored and aligned versions of the trees represented in **Figure 2—figure supplements 8, 9, 12 and 13**. The color of the branches is the tree similarity score, where lighter colors represent a higher number of topological differences. It is a custom implementation of the Jaccard Index by [Phylo.io](#).

Statistical analysis of tree topology

To statistically evaluate our best-supported topology of the OSK and LOTUS trees, we compared constrained topologies to the highest likelihood trees using the SOWHAT tool (Church et al., 2015). SOWHAT automates the stringent SOWH phylogenetic topology test (Swofford et al., 1996), and compares the log likelihood between generated trees. We defined three constrained trees to test our results, one requiring monophyly of all domains of life, a second requiring only eukaryotic monophyly, and the last one requiring monophyly of the Oskar LOTUS domain (**Source data 1**: Data > Trees > constrained_kingdom_tree.tre, constrained_eukmono_tree.tre and constrained_lotus_mono_tree.tre). We then ran SOWHAT using its default parameters, 1000 bootstraps, and the two constrained trees against the OSK or LOTUS alignment used to generate the phylogenetic trees

(**Source data 1:** Alignments > OSK_MUSCLE_TREE.fasta and LOTUS_MUSCLE_TREE.fasta). All best trees generated by SOWHAT are available in (**Source data 1:** Data > Trees > SOWHAT_*_test.tre).

Code availability

All custom code generated for this study is available in the GitHub repository https://github.com/extavourlab/Oskar_HGT, commit ID 6f6c4c50dfb9391567d70f9eea922f3876a4e153 (**Blondel et al., 2020**; copy archived at https://github.com/elife sciences-publications/Oskar_HGT).

Scripts

All scripts used herein are hosted on GitHub at https://github.com/extavourlab/Oskar_HGT.

Acknowledgements

We thank Sean Eddy, Chuck Davis, and Extavour lab members for discussion.

Additional information

Funding

Funder

Harvard University

Author

Leo Blondel
Cassandra G Extavour
Tamsin E M Jones

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

Leo Blondel, Data curation, Formal analysis, Validation, Visualization, Methodology, Writing - original draft, Writing - review and editing; Tamsin EM Jones, Data curation, Writing - review and editing; Cassandra G Extavour, Conceptualization, Supervision, Funding acquisition, Writing - original draft, Project administration, Writing - review and editing

Author ORCIDs

Leo Blondel  <http://orcid.org/0000-0003-2276-4821>

Tamsin EM Jones  <https://orcid.org/0000-0002-0027-0858>

Cassandra G Extavour  <https://orcid.org/0000-0003-2922-5855>

Decision letter and Author response

Decision letter <https://doi.org/10.7554/eLife.45539.sa1>

Author response <https://doi.org/10.7554/eLife.45539.sa2>

Additional files

Supplementary files

- Source data 1. Alignment and Sequence Classification Tools & Data. **Subfolder "Alignments":** All sequences identified and analyzed in this study, in FASTA format and with corresponding Alignments. Subfolder BLAST search results: Results of BLASTP searches with full length Oskar, OSK or LOTUS domains as queries. **Subfolder "Data":** Necessary files for running the different IPython notebooks: **a. Subfolder "HMM":** HMM models used for iterative searching for sequences similar to full-length Oskar, LOTUS and OSK domains; **b. Subfolder "Taxonomy":** Conversion table for UniProt ID to taxon information (uniprot_ID_taxa.tsv); **c. Subfolder "Trees":** Contains the tree files obtained from i. RaxML phylogenetic analyses of the OSK and LOTUS domains aligned with MUSCLE, T-Coffee or PRANK; ii. MrBayes phylogenetic analyses of the OSK and LOTUS domains aligned with MUSCLE; iii. SOWHAT analyses.

- Supplementary file 1. Supplementary tables. (A) List of genomes and transcriptomes used for automated oskar search. (B) List of Oskar sequences used in the final alignment. (C) List of sequences used for phylogenetic analysis of the LOTUS domain. (D) List of sequences used for phylogenetic analysis of the OSK domain.
- Transparent reporting form

Data availability

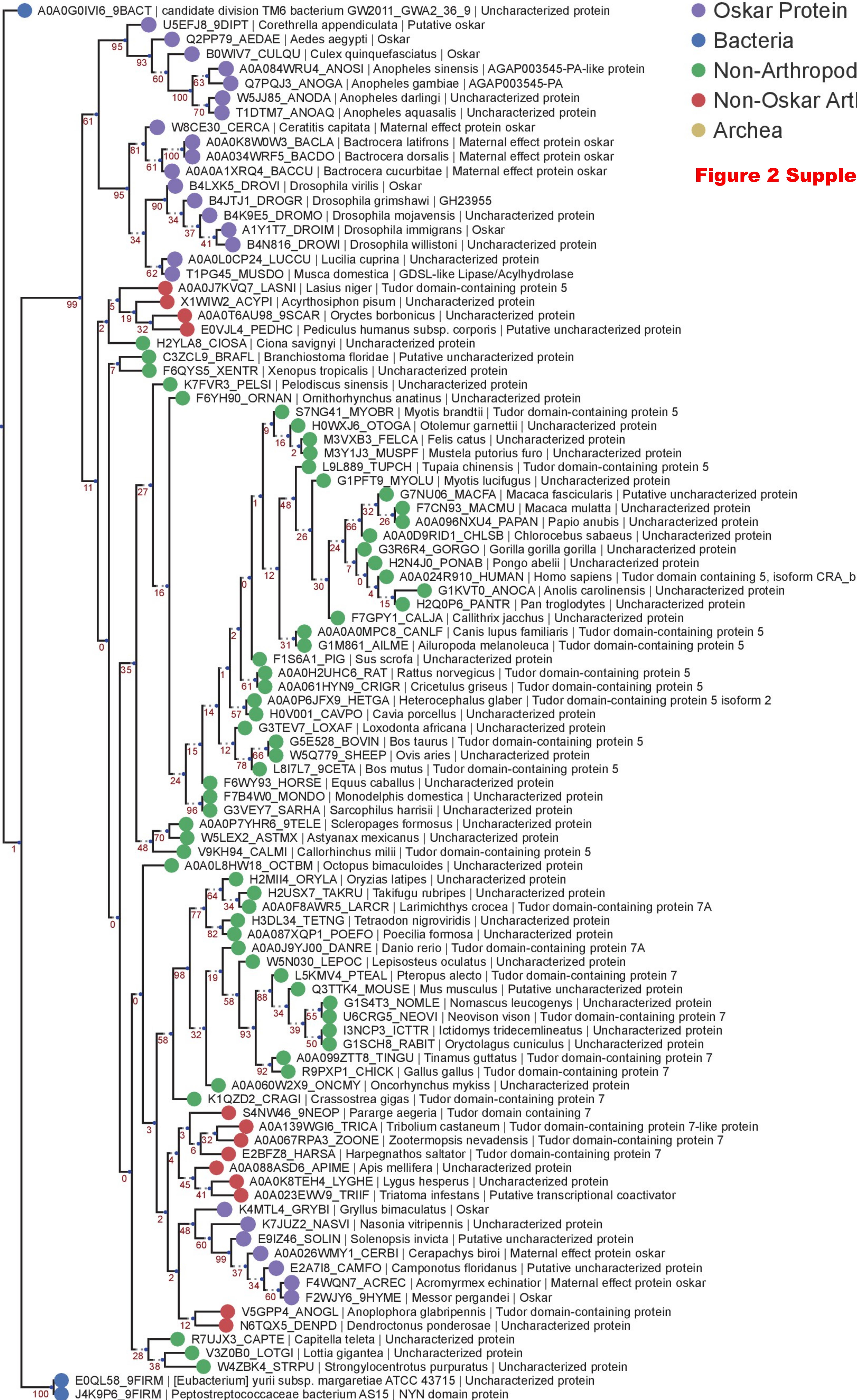
All data are available in the main text or the supplementary materials.

References

- Abouheif E. 2013. Evolution: oskar reveals missing link in co-optive evolution. *Current Biology* **23**:R24–R25. DOI: <https://doi.org/10.1016/j.cub.2012.11.028>, PMID: 23305666
- Acuna R, Padilla BE, Florez-Ramos CP, Rubio JD, Herrera JC, Benavides P, Lee S-J, Yeats TH, Egan AN, Doyle JJ, Rose JKC. 2012. Adaptive horizontal transfer of a bacterial gene to an invasive insect pest of coffee. *PNAS* **109**:4197–4202. DOI: <https://doi.org/10.1073/pnas.1121190109>
- Ahuja A, Extavour CG. 2014. Patterns of molecular evolution of the germ line specification gene oskar suggest that a novel domain may contribute to functional divergence in *Drosophila*. *Development Genes and Evolution* **224**:65–77. DOI: <https://doi.org/10.1007/s00427-013-0463-7>, PMID: 24407548
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of Molecular Biology* **215**:403–410. DOI: [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2), PMID: 2231712
- Benton MA, Kenny NJ, Conrads KH, Roth S, Lynch JA. 2016. Deep, staged transcriptomic resources for the novel coleopteran models *Atrachya menetriesii* and *Callosobruchus maculatus*. *PLOS ONE* **11**:e0167431. DOI: <https://doi.org/10.1371/journal.pone.0167431>, PMID: 27907180
- Blondel L, Jones TEM, Extavour CG. 2020. Supporting scripts for Bacterial contribution to the genesis of the novel germ line determinant oskar. GitHub. 370f62a. https://github.com/extavourlab/Oskar_HGT
- Boto L. 2014. Horizontal gene transfer in the acquisition of novel traits by metazoans. *Proceedings of the Royal Society B: Biological Sciences* **281**:20132450. DOI: <https://doi.org/10.1098/rspb.2013.2450>
- Bourtzis K, Miller TA. 2006. *Insect Symbiosis*. Boca Raton FL: CRC Press. DOI: [https://doi.org/10.1653/0015-4040\(2003\)086\[0493:BR\]2.0.CO;2](https://doi.org/10.1653/0015-4040(2003)086[0493:BR]2.0.CO;2)
- Bublitz DC, Chadwick GL, Magyar JS, Sandoz KM, Brooks DM, Mesnage S, Ladinsky MS, Garber AI, Bjorkman PJ, Orphan VJ, McCutcheon JP. 2019. Peptidoglycan production by an Insect-Bacterial mosaic. *Cell* **179**:703–712. DOI: <https://doi.org/10.1016/j.cell.2019.08.054>
- Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**:1972–1973. DOI: <https://doi.org/10.1093/bioinformatics/btp348>, PMID: 19505945
- Chen S, Zhang YE, Long M. 2010. New genes in *Drosophila* quickly become essential. *Science* **330**:1682–1685. DOI: <https://doi.org/10.1126/science.1196380>, PMID: 21164016
- Chen S, Spletter M, Ni X, White KP, Luo L, Long M. 2012. Frequent recent origination of brain genes shaped the evolution of foraging behavior in *Drosophila*. *Cell Reports* **1**:118–132. DOI: <https://doi.org/10.1016/j.celrep.2011.12.010>, PMID: 22832161
- Chen S, Krinsky BH, Long M. 2013. New genes as drivers of phenotypic evolution. *Nature Reviews Genetics* **14**:645–660. DOI: <https://doi.org/10.1038/nrg3521>, PMID: 23949544
- Chepkemoi ST, Mararo E, Butungi H, Paredes J, Masiga D, Sinkins SP, Herren JK. 2017. Identification of *Spiroplasmainsolitum* symbionts in *Anopheles gambiae*. *Wellcome Open Research* **2**:90. DOI: <https://doi.org/10.12688/wellcomeopenres.12468.1>, PMID: 29152597
- Church SH, Ryan JF, Dunn CW. 2015. Automation and evaluation of the SOWH test with SOWHAT. *Systematic Biology* **64**:1048–1058. DOI: <https://doi.org/10.1093/sysbio/syv055>
- Cornelis G, Heidmann O, Bernard-Stoecklin S, Reynaud K, Véron G, Mulot B, Dupressoir A, Heidmann T. 2012. Ancestral capture of syncytin-Car1, a fusogenic endogenous retroviral envelope gene involved in Placentation and conserved in carnivorans. *PNAS* **109**:E432–E441. DOI: <https://doi.org/10.1073/pnas.1115346109>, PMID: 22308384
- Di Lelio I, Illiano A, Astarita F, Gianfranceschi L, Horner D, Varriacchio P, Amoresano A, Pucci P, Pennacchio F, Caccia S. 2019. Evolution of an insect immune barrier through horizontal gene transfer mediated by a parasitic wasp. *PLOS Genetics* **15**:e1007998. DOI: <https://doi.org/10.1371/journal.pgen.1007998>, PMID: 30835731
- Dunning Hotopp JC, Clark ME, Oliveira DC, Foster JM, Fischer P, Muñoz Torres MC, Giebel JD, Kumar N, Ishmael N, Wang S, Ingram J, Nene RV, Shepard J, Tomkins J, Richards S, Spiro DJ, Ghedin E, Slatko BE, Tettelin H, Werren JH. 2007. Widespread lateral gene transfer from intracellular Bacteria to multicellular eukaryotes. *Science* **317**:1753–1756. DOI: <https://doi.org/10.1126/science.1142490>, PMID: 17761848
- Eddy SR. 2011. Accelerated profile HMM searches. *PLOS Computational Biology* **7**:e1002195. DOI: <https://doi.org/10.1371/journal.pcbi.1002195>, PMID: 22039361
- Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**:113. DOI: <https://doi.org/10.1186/1471-2105-5-113>, PMID: 15318951

- Ephrussi A**, Lehmann R. 1992. Induction of germ cell formation by oskar. *Nature* **358**:387–392. DOI: <https://doi.org/10.1038/358387a0>, PMID: 1641021
- Ewen-Campen B**, Srouji JR, Schwager EE, Extavour CG. 2012. Oskar predates the evolution of germ plasm in insects. *Current Biology* **22**:2278–2283. DOI: <https://doi.org/10.1016/j.cub.2012.10.019>, PMID: 23122849
- Ewen-Campen B**, Donoughe S, Clarke DN, Extavour CG. 2013. Germ cell specification requires zygotic mechanisms rather than germ plasm in a basally branching insect. *Current Biology* **23**:835–842. DOI: <https://doi.org/10.1016/j.cub.2013.03.063>, PMID: 23623552
- Extavour CG**, Akam M. 2003. Mechanisms of germ cell specification across the metazoans: epigenesis and preformation. *Development* **130**:5869–5884. DOI: <https://doi.org/10.1242/dev.00804>, PMID: 14597570
- Gerlt JA**, Bouvier JT, Davidson DB, Imker HJ, Sadkin B, Slater DR, Whalen KL. 2015. Enzyme function Initiative-Enzyme similarity tool (EFI-EST): A web tool for generating protein sequence similarity networks. *Biochimica Et Biophysica Acta (BBA) - Proteins and Proteomics* **1854**:1019–1037. DOI: <https://doi.org/10.1016/j.bbapap.2015.04.015>
- Hoekstra HE**, Coyne JA. 2007. The locus of evolution: evo-devo and the genetics of adaptation. *Evolution* **61**: 995–1016. DOI: <https://doi.org/10.1111/j.1558-5646.2007.00105.x>
- Huelsenbeck JP**, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**:754–755. DOI: <https://doi.org/10.1093/bioinformatics/17.8.754>, PMID: 11524383
- Hurd TR**, Herrmann B, Sauerwald J, Sanny J, Grosch M, Lehmann R. 2016. Long Oskar controls mitochondrial inheritance in *Drosophila melanogaster*. *Developmental Cell* **39**:560–571. DOI: <https://doi.org/10.1016/j.devcel.2016.11.004>, PMID: 27923120
- Husnik F**, Nikoh N, Koga R, Ross L, Duncan RP, Fujie M, Tanaka M, Satoh N, Bachtrog D, Wilson AC, von Dohlen CD, Fukatsu T, McCutcheon JP. 2013. Horizontal gene transfer from diverse Bacteria to an insect genome enables a tripartite nested mealybug symbiosis. *Cell* **153**:1567–1578. DOI: <https://doi.org/10.1016/j.cell.2013.05.040>, PMID: 23791183
- Husnik F**, McCutcheon JP. 2018. Functional horizontal gene transfer from Bacteria to eukaryotes. *Nature Reviews Microbiology* **16**:67–79. DOI: <https://doi.org/10.1038/nrmicro.2017.137>, PMID: 29176581
- Jenny A**, Hachet O, Závorszky P, Cyrlaff A, Weston MD, Johnston DS, Erdélyi M, Ephrussi A. 2006. A translation-independent role of oskar RNA in early *Drosophila* oogenesis. *Development* **133**:2827–2833. DOI: <https://doi.org/10.1242/dev.02456>, PMID: 16835436
- Jeske M**, Bordi M, Glatt S, Müller S, Rybin V, Müller CW, Ephrussi A. 2015. The crystal structure of the *Drosophila* germline inducer Oskar identifies two domains with distinct Vasa helicase- and RNA-Binding activities. *Cell Reports* **12**:587–598. DOI: <https://doi.org/10.1016/j.celrep.2015.06.055>, PMID: 26190108
- Jeske M**, Müller CW, Ephrussi A. 2017. The LOTUS domain is a conserved DEAD-box RNA helicase regulator essential for the recruitment of Vasa to the germ plasm and nuage. *Genes & Development* **31**:939–952. DOI: <https://doi.org/10.1101/gad.297051.117>, PMID: 28536148
- Kim-Ha J**, Smith JL, Macdonald PM. 1991. Oskar mRNA is localized to the posterior pole of the *Drosophila* oocyte. *Cell* **66**:23–35. DOI: [https://doi.org/10.1016/0092-8674\(91\)90136-M](https://doi.org/10.1016/0092-8674(91)90136-M), PMID: 2070416
- Kirk-Spriggs AH**, Sinclair BJ. 2017. *Manual of Afrotropical Diptera; Ptychopteridae*. Pretoria, South Africa: South African National Biodiversity Institute.
- Korf I**. 2004. Gene finding in novel genomes. *BMC Bioinformatics* **5**:59. DOI: <https://doi.org/10.1186/1471-2105-5-59>, PMID: 15144565
- Lehmann R**. 2016. Germ plasm biogenesis—an Oskar-Centric perspective. *Current Topics in Developmental Biology* **116**:679–707. DOI: <https://doi.org/10.1016/bs.ctdb.2015.11.024>, PMID: 26970648
- Lehmann R**, Nüsslein-Volhard C. 1986. Abdominal segmentation, pole cell formation, and embryonic polarity require the localized activity of Oskar, a maternal gene in *Drosophila*. *Cell* **47**:141–152. DOI: [https://doi.org/10.1016/0092-8674\(86\)90375-2](https://doi.org/10.1016/0092-8674(86)90375-2), PMID: 3093084
- López-Madrigal S**, Gil R. 2017. Et tu, brute? not even intracellular mutualistic symbionts escape horizontal gene transfer. *Genes* **8**:247. DOI: <https://doi.org/10.3390/genes8100247>
- Löytynoja A**. 2014. Phylogeny-aware alignment with PRANK. *Methods in Molecular Biology* **1079**:155–170. DOI: https://doi.org/10.1007/978-1-62703-646-7_10, PMID: 24170401
- Lynch JA**, Ozük O, Khila A, Abouheif E, Desplan C, Roth S. 2011. The Phylogenetic Origin of Oskar Coincided With the Origin of Maternally Provisioned Germ Plasm and Pole Cells at the Base of the Holometabola. *PLOS Genetics* **7**:e1002029. DOI: <https://doi.org/10.1371/journal.pgen.1002029>, PMID: 21552321
- Misof B**, Liu S, Meusemann K, Peters RS, Donath A, Mayer C, Frandsen PB, Ware J, Flouri T, Beutel RG, Niehuis O, Petersen M, Izquierdo-Carrasco F, Wappler T, Rust J, Aberer AJ, Aspöck U, Aspöck H, Bartel D, Blanke A, et al. 2014. Phylogenomics resolves the timing and pattern of insect evolution. *Science* **346**:763–767. DOI: <https://doi.org/10.1126/science.1257570>, PMID: 25378627
- Notredame C**, Higgins DG, Heringa J. 2000. T-Coffee: a novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology* **302**:205–217. DOI: <https://doi.org/10.1006/jmbi.2000.4042>, PMID: 10964570
- Peters RS**, Krogmann L, Mayer C, Donath A, Gunkel S, Meusemann K, Kozlov A, Podsiadlowski L, Petersen M, Lanfear R, Diez PA, Heraty J, Kjer KM, Klopstein S, Meier R, Polidori C, Schmitt T, Liu S, Zhou X, Wappler T, et al. 2017. Evolutionary history of the Hymenoptera. *Current Biology* **27**:1013–1018. DOI: <https://doi.org/10.1016/j.cub.2017.01.027>, PMID: 28343967
- Provorov NA**, Onishchuk OP. 2018. Microbial symbionts of insects: genetic organization, adaptive role, and evolution. *Microbiology* **87**:151–163. DOI: <https://doi.org/10.1134/S002626171802011X>

- Quispe-Huamanquispe DG**, Gheysen G, Kreuze JF. 2017. Horizontal gene transfer contributes to plant evolution: the case of *Agrobacterium* T-DNAs. *Frontiers in Plant Science* **8**:2015. DOI: <https://doi.org/10.3389/fpls.2017.02015>, PMID: 29225610
- Rees J**, Cranston K. 2017. Automated assembly of a reference taxonomy for phylogenetic data synthesis. *Biodiversity Data Journal* **5**:e12581. DOI: <https://doi.org/10.3897/BDJ.5.e12581>
- Robinson O**, Dylus D, Dessimoz C. 2016. Phylo.io: interactive viewing and comparison of large phylogenetic trees on the web. *Molecular Biology and Evolution* **33**:2163–2166. DOI: <https://doi.org/10.1093/molbev/msw080>, PMID: 27189561
- Shannon P**, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research* **13**:2498–2504. DOI: <https://doi.org/10.1101/gr.1239303>, PMID: 14597658
- Shelomi M**, Danchin EG, Heckel D, Wipfler B, Bradler S, Zhou X, Pauchet Y. 2016. Horizontal gene transfer of pectinases from Bacteria preceded the diversification of stick and leaf insects. *Scientific Reports* **6**:26388. DOI: <https://doi.org/10.1038/srep26388>, PMID: 27210832
- Sloan DB**, Nakabachi A, Richards S, Qu J, Murali SC, Gibbs RA, Moran NA. 2014. Parallel histories of horizontal gene transfer facilitated extreme reduction of endosymbiont genomes in sap-feeding insects. *Molecular Biology and Evolution* **31**:857–871. DOI: <https://doi.org/10.1093/molbev/msu004>, PMID: 24398322
- Stamatakis A**. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**:1312–1313. DOI: <https://doi.org/10.1093/bioinformatics/btu033>, PMID: 24451623
- Stanke M**, Steinkamp R, Waack S, Morgenstern B. 2004. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Research* **32**:W309–W312. DOI: <https://doi.org/10.1093/nar/gkh379>, PMID: 15215400
- Swofford DL**, Olsen GJ, Waddell PJ. 1996. Phylogenetic inference. In: Moritz C, Hillis DM, Mable BK (Eds). *Molecular Systematics*. 2nd Edition. Sinauer, MA: Sinauer Associates, Inc. p. 407–453.
- Tautz D**, Domazet-Lošo T. 2011. The evolutionary origin of orphan genes. *Nature Reviews Genetics* **12**:692–702. DOI: <https://doi.org/10.1038/nrg3053>, PMID: 21878963
- Taylor JS**, Raes J. 2004. Duplication and divergence: the evolution of new genes and old ideas. *Annual Review of Genetics* **38**:615–643. DOI: <https://doi.org/10.1146/annurev.genet.38.072902.092831>, PMID: 15568988
- U Consortium**. 2005. The universal protein resource (UniProt). *Nucleic Acids Research* **2009**:169–174. DOI: <https://doi.org/10.1093/nar/gkl929>
- Vanvo NF**, Ephrussi A. 2002. Oskar anchoring restricts pole plasm formation to the posterior of the *Drosophila* oocyte. *Development* **129**:3705–3714. PMID: 12117819
- Wheeler D**, Redding AJ, Werren JH. 2013. Characterization of an ancient lepidopteran lateral gene transfer. *PLOS ONE* **8**:e59262. DOI: <https://doi.org/10.1371/journal.pone.0059262>, PMID: 23533610
- Wilson AC**, Duncan RP. 2015. Signatures of host/symbiont genome coevolution in insect nutritional endosymbioses. *PNAS* **112**:10255–10261. DOI: <https://doi.org/10.1073/pnas.1423305112>, PMID: 26039986
- Wittkopp PJ**, Kalay G. 2012. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nature Reviews Genetics* **13**:59–69. DOI: <https://doi.org/10.1038/nrg3095>
- Wybouw N**, Pauchet Y, Heckel DG, Van Leeuwen T. 2016. Horizontal gene transfer contributes to the evolution of arthropod herbivory. *Genome Biology and Evolution* **8**:1785–1801. DOI: <https://doi.org/10.1093/gbe/evw119>, PMID: 27307274
- Xu X**, Brechbiel JL, Gavis ER. 2013. Dynein-Dependent Transport of nanos RNA in *Drosophila* Sensory Neurons Requires Rumpelstiltskin and the Germ Plasm Organizer Oskar. *Journal of Neuroscience* **33**:14791–14800. DOI: <https://doi.org/10.1523/JNEUROSCI.5864-12.2013>, PMID: 24027279
- Yang N**, Yu Z, Hu M, Wang M, Lehmann R, Xu RM. 2015. Structure of *Drosophila* Oskar reveals a novel RNA binding protein. *PNAS* **112**:11541–11546. DOI: <https://doi.org/10.1073/pnas.1515568112>, PMID: 26324911
- Zchori-Fein E**, Perlman SJ, Kelly SE, Katzir N, Hunter MS. 2004. Characterization of a 'Bacteroidetes' symbiont in Encarsia wasps (Hymenoptera: Aphelinidae): proposal of 'Candidatus Cardinium hertigii'. *International Journal of Systematic and Evolutionary Microbiology* **54**:961–968. DOI: <https://doi.org/10.1099/ijs.0.02957-0>, PMID: 15143050
- Zchori-Fein E**, Perlman SJ. 2004. Distribution of the bacterial symbiont *Cardinium* in arthropods. *Molecular Ecology* **13**:2009–2016. DOI: <https://doi.org/10.1111/j.1365-294X.2004.02203.x>, PMID: 15189221
- Zeng Z**, Fu Y, Guo D, Wu Y, Ajayi OE, Wu Q. 2018. Bacterial endosymbiont *Cardinium cSfur* genome sequence provides insights for understanding the symbiotic relationship in *Sogatella furcifera* host. *BMC Genomics* **19**:688. DOI: <https://doi.org/10.1186/s12864-018-5078-y>, PMID: 30231855
- Zhang YE**, Landback P, Vibranovski M, Long M. 2012. New genes expressed in human brains: implications for annotating evolving genomes. *BioEssays* **34**:982–991. DOI: <https://doi.org/10.1002/bies.201200008>, PMID: 23001763
- Zhang W**, Landback P, Gschwend AR, Shen B, Long M. 2015. New genes drive the evolution of gene interaction networks in the human and mouse genomes. *Genome Biology* **16**:202. DOI: <https://doi.org/10.1186/s13059-015-0772-4>, PMID: 26424194



- Oskar Protein
- Bacteria
- Non-Arthropod Eukaryota
- Non-Oskar Arthropoda
- Archaea

Figure 2 Supplement 1

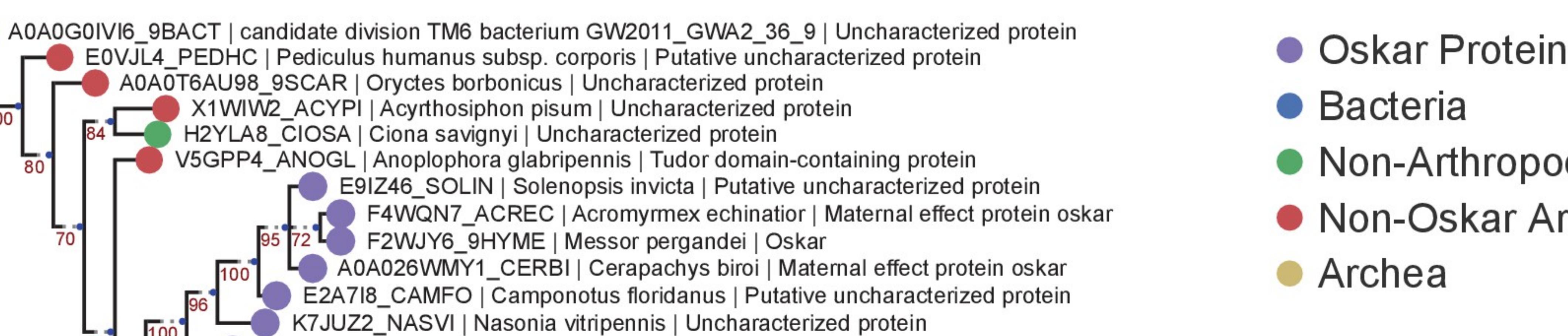
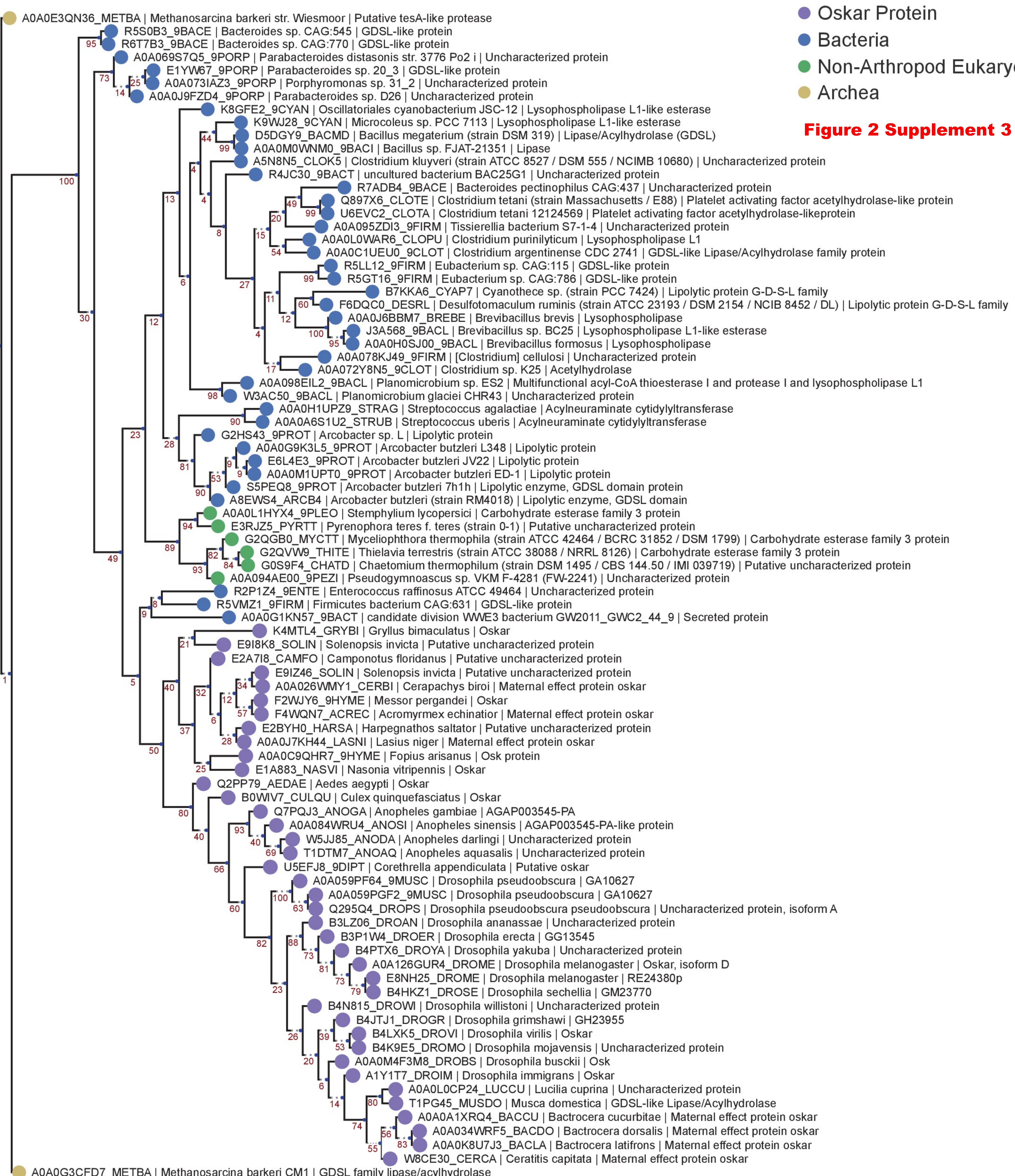


Figure 2 Supplement 2

- Oskar Protein
- Bacteria
- Non-Arthropod Eukaryota
- Archea

Figure 2 Supplement 3



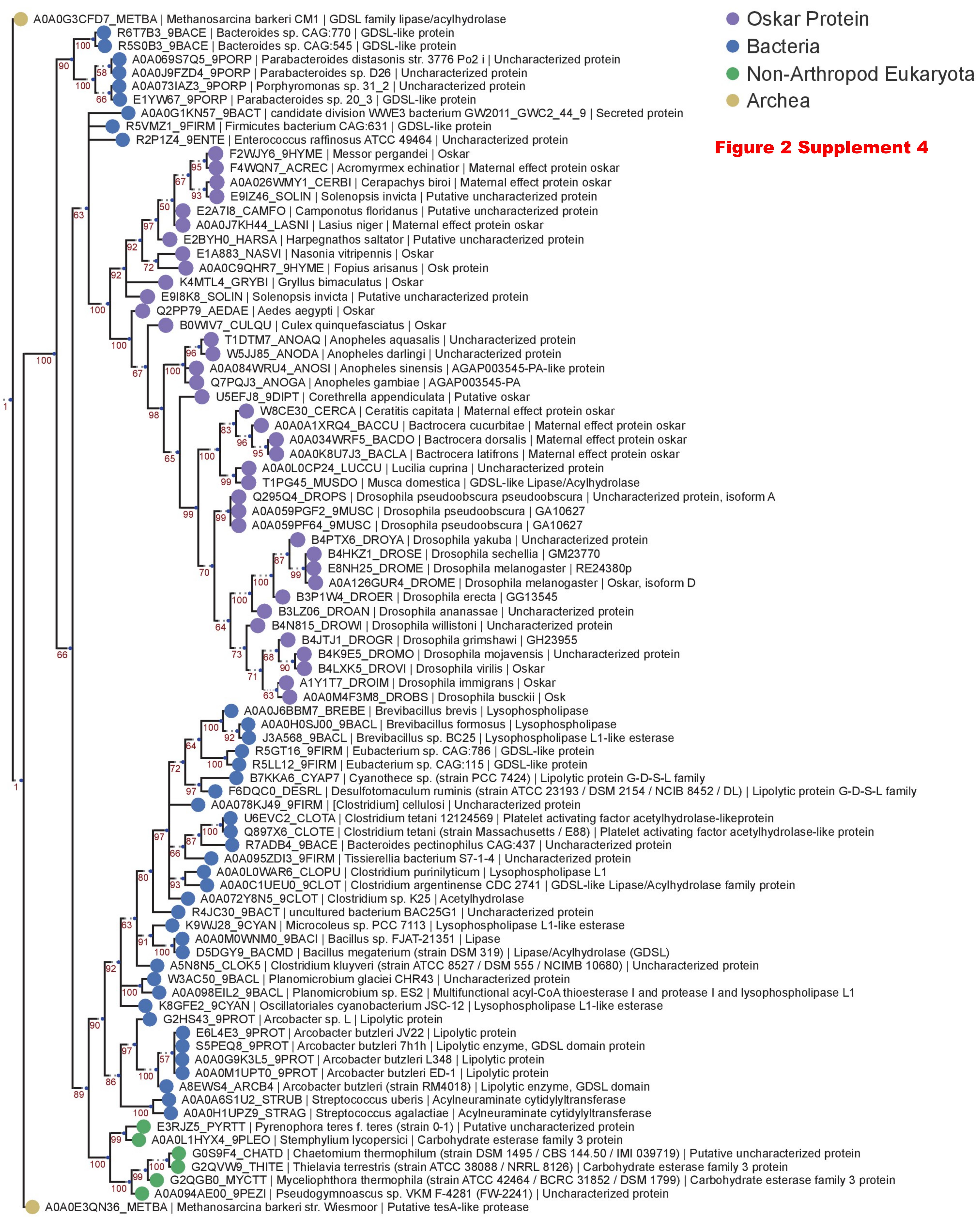
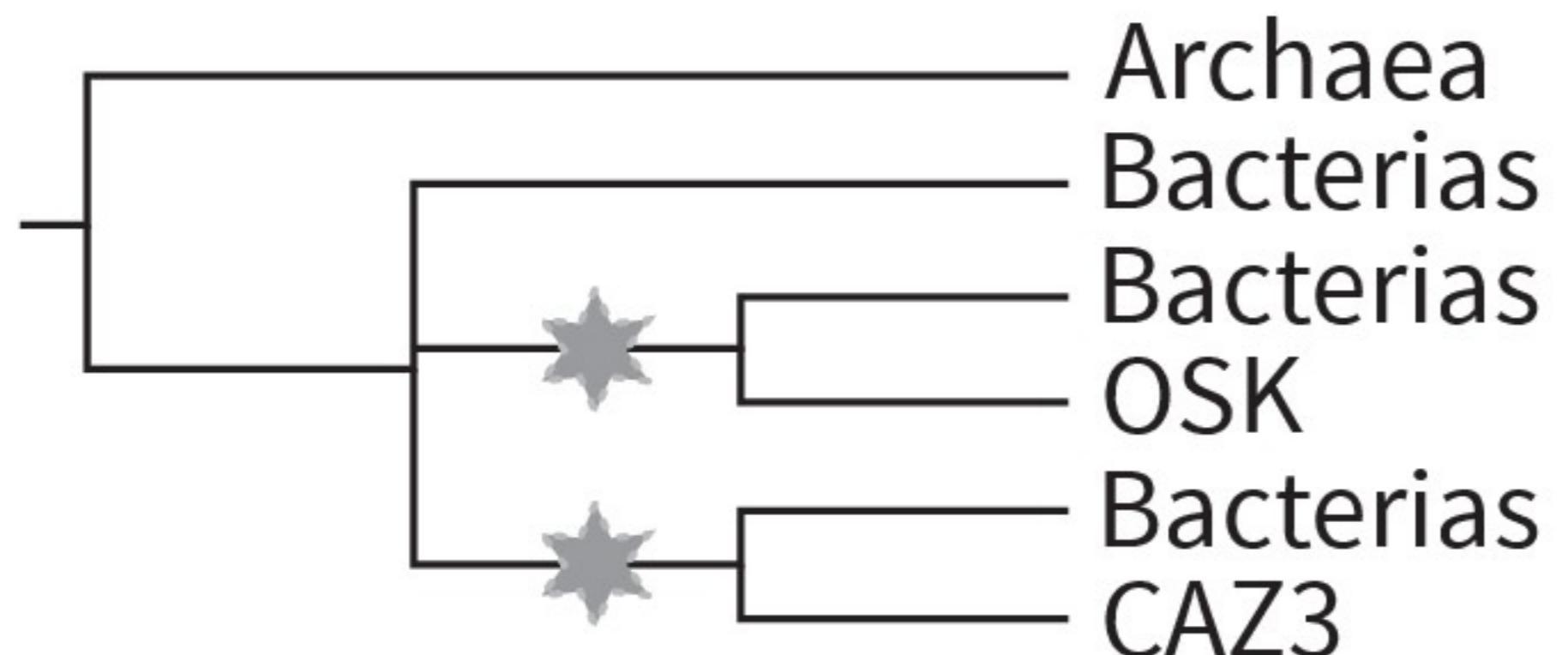
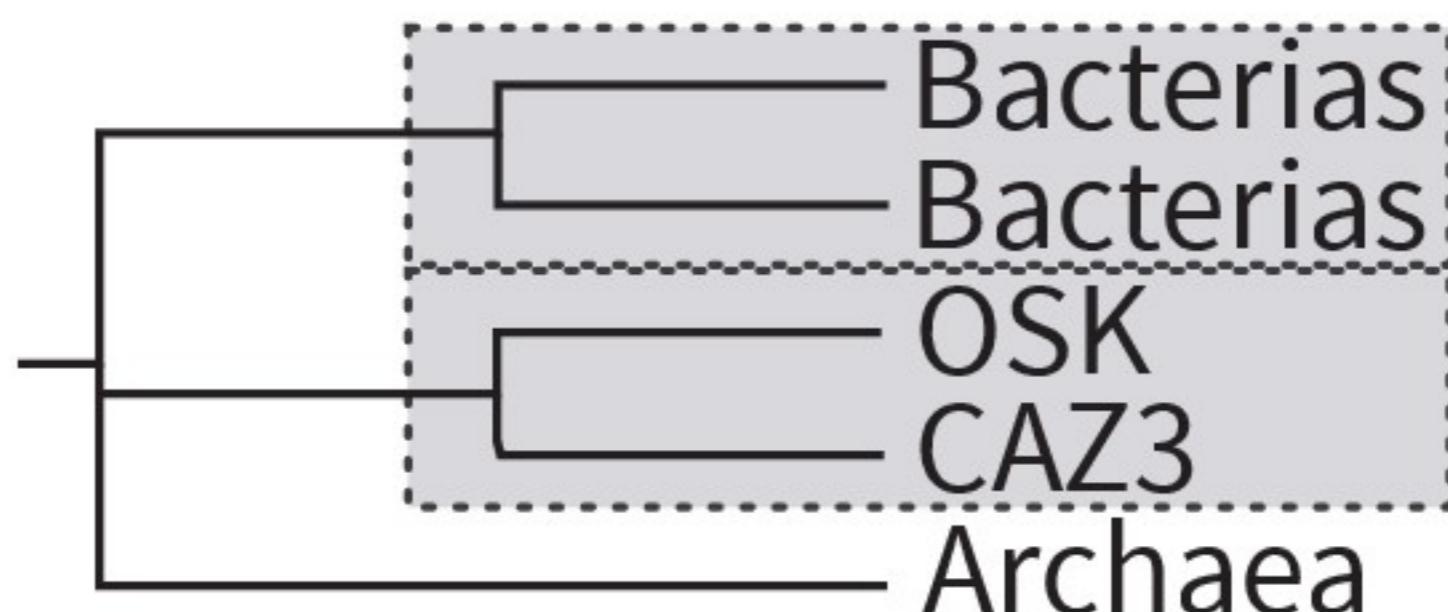
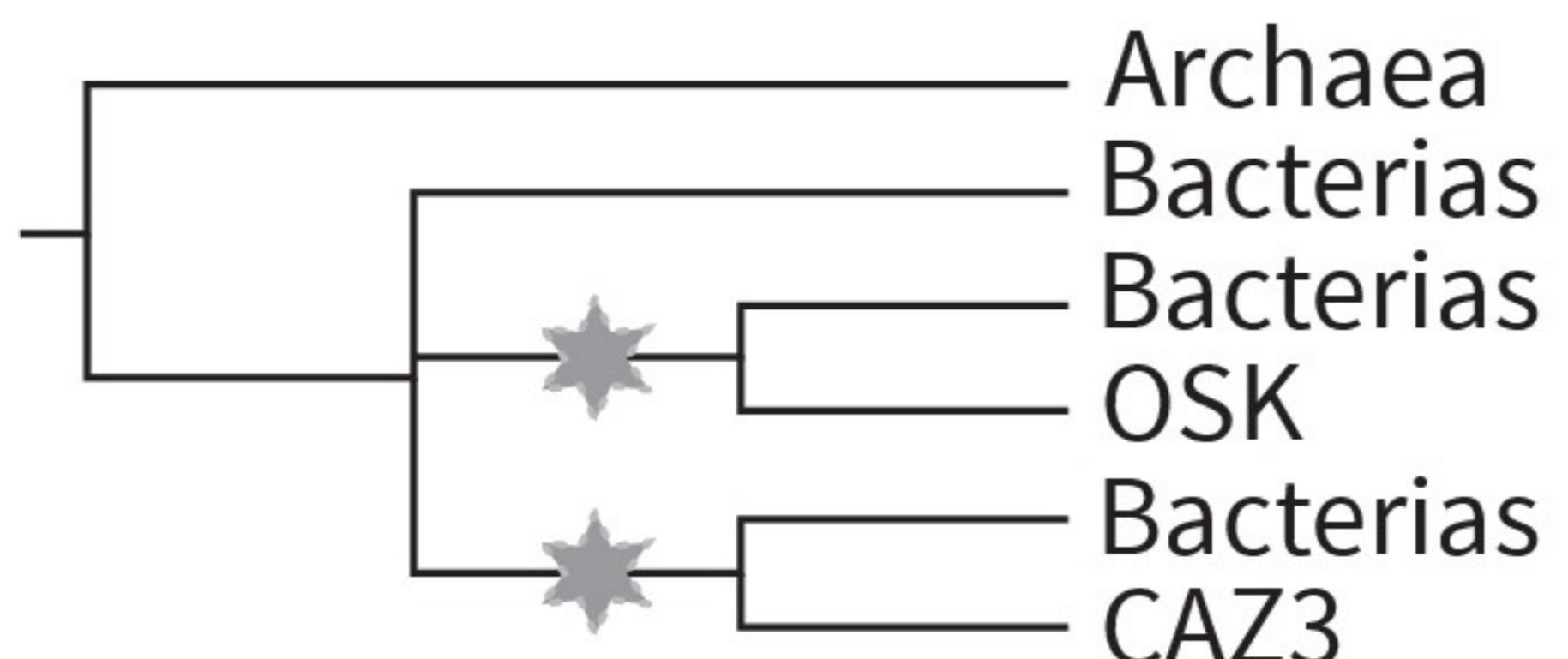
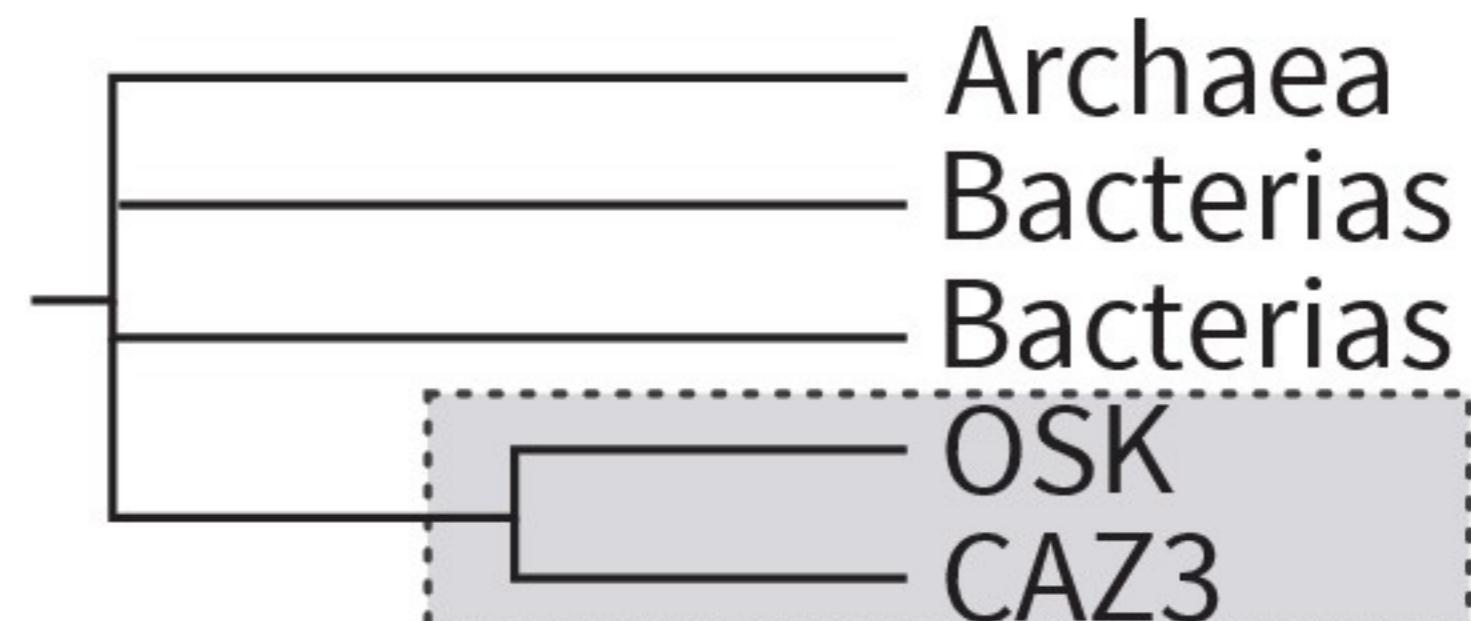
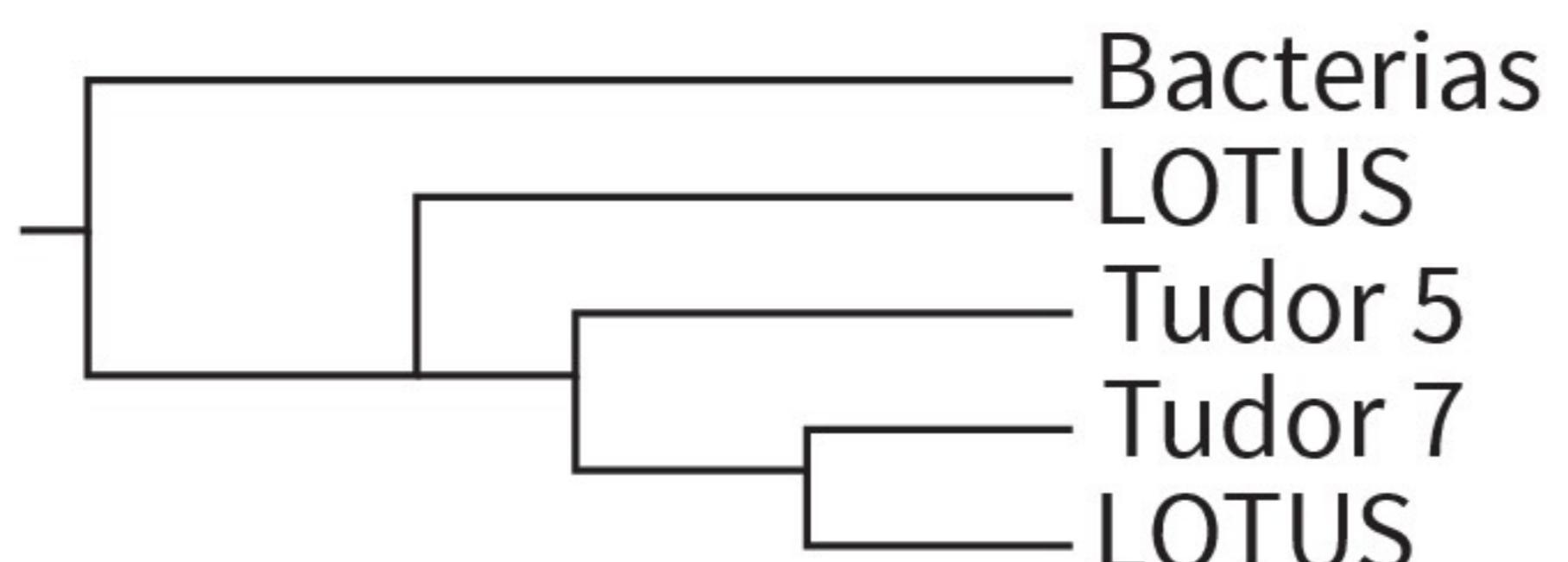
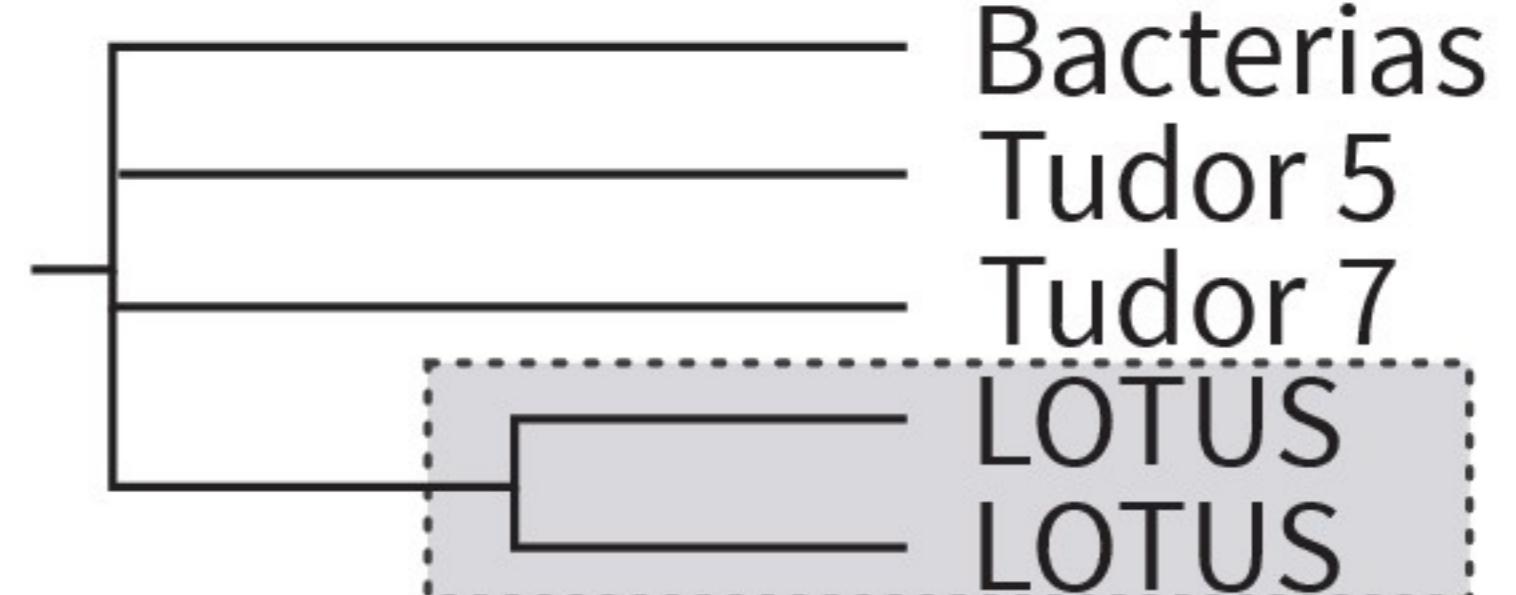


Figure 2 Supplement 4

a*Unconstrained tree***VS***Constrained by domain of life***P-value = 0.002****b***Unconstrained tree***VS***Monophyletic Eukaryota***P-value = 0.009****c***Unconstrained tree***VS***Monophyletic LOTUS***P-value = 0.037**

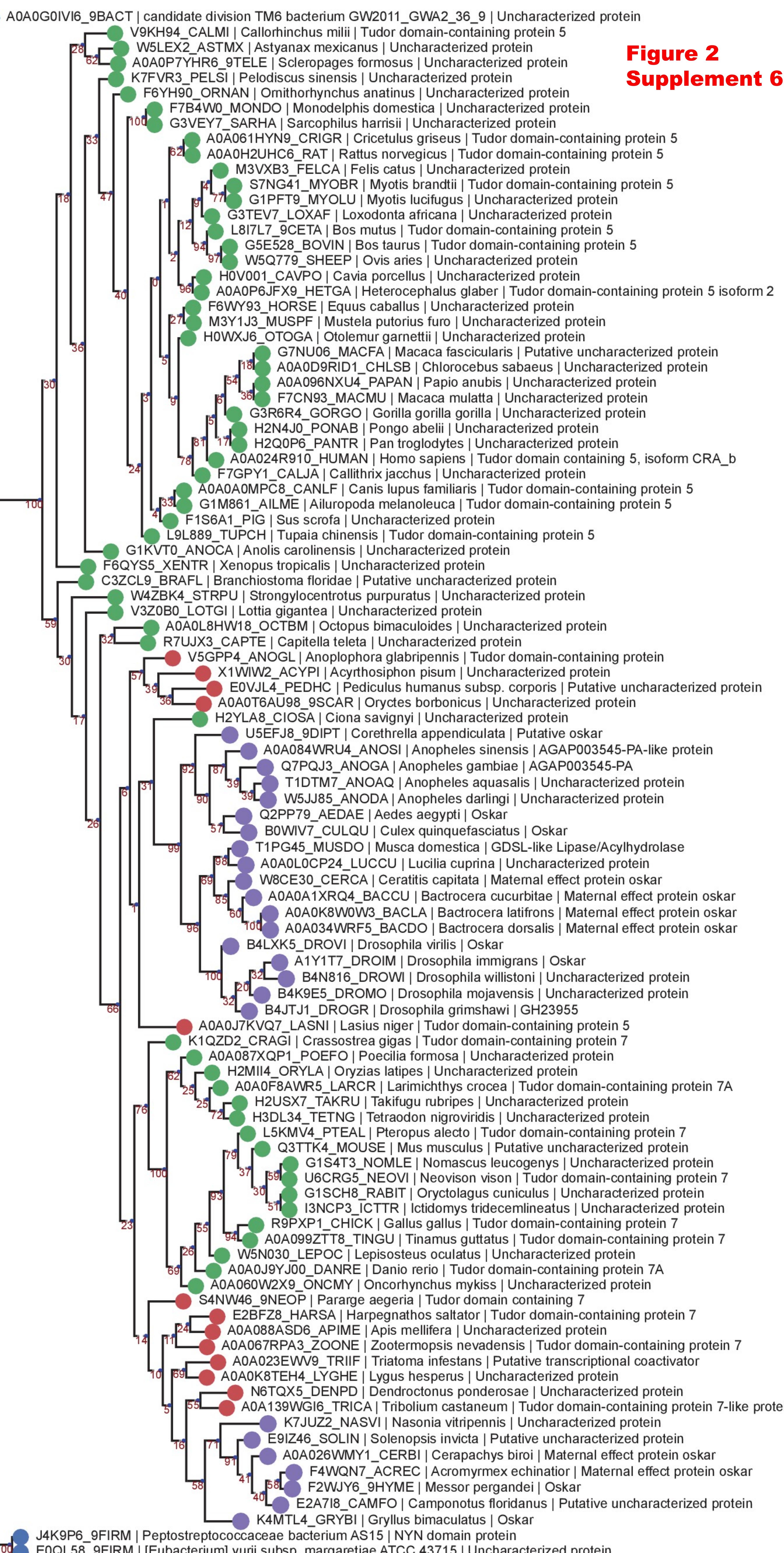


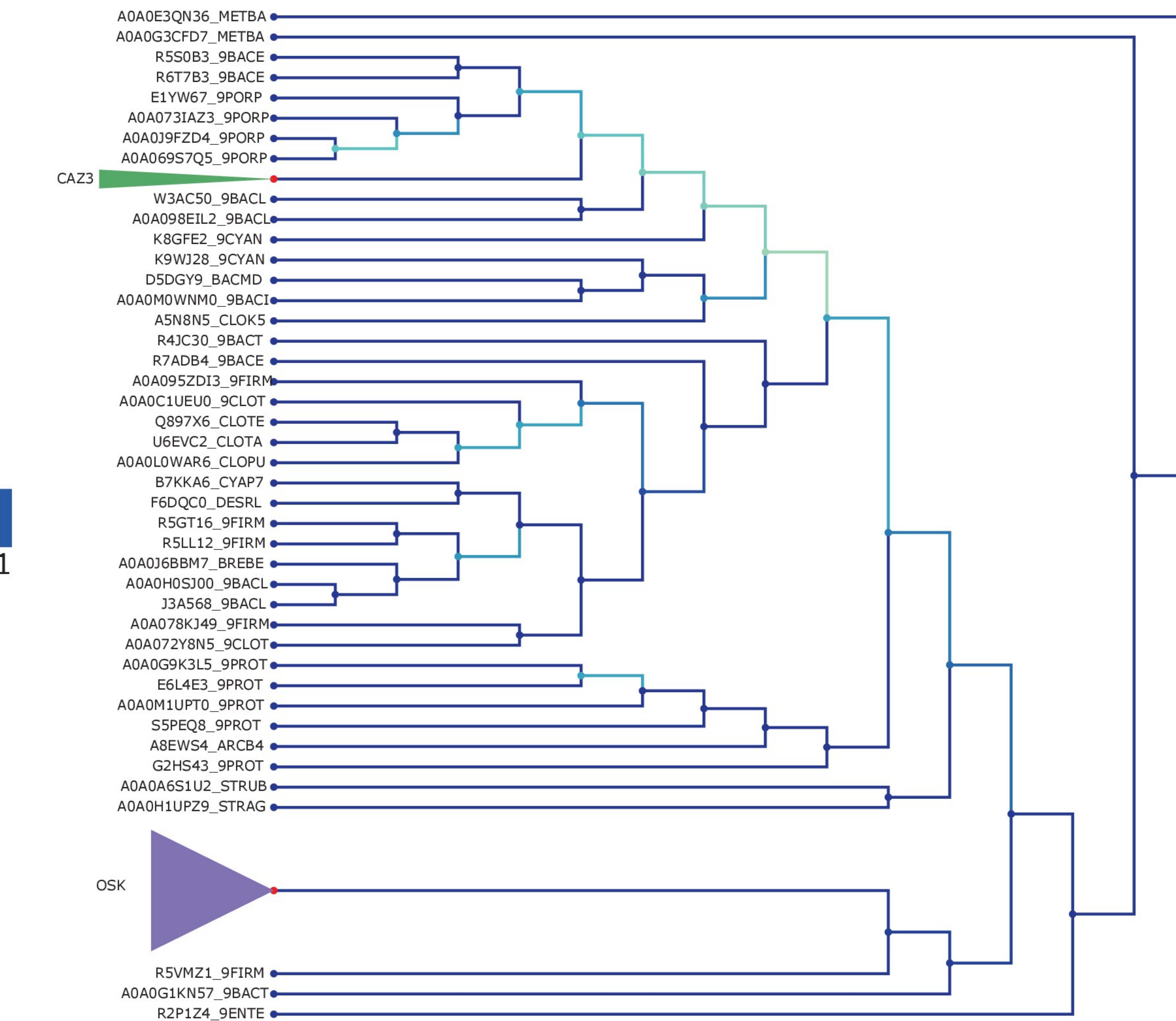
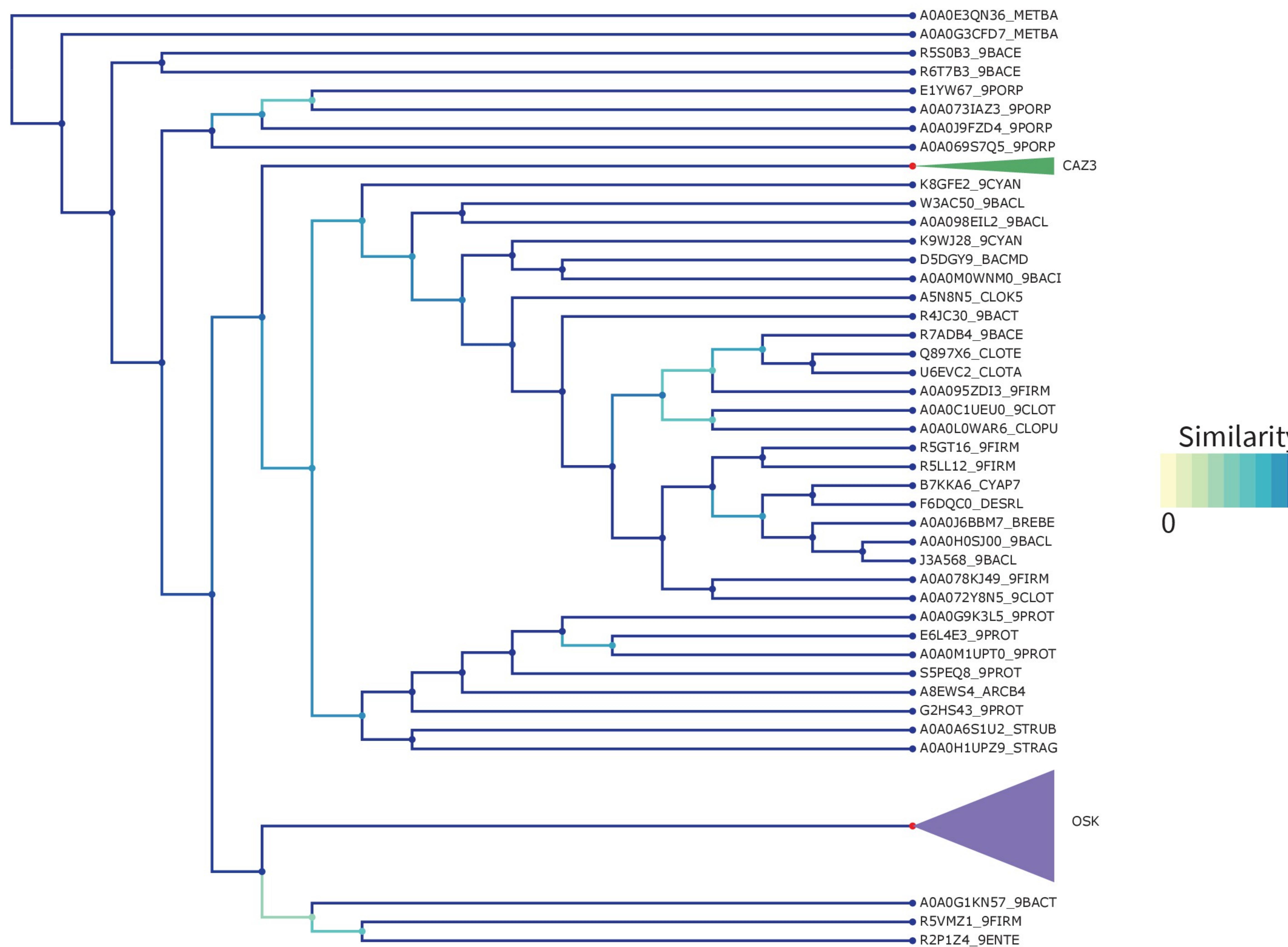
Figure 2
Supplement 6

Figure 2 Supplement 7

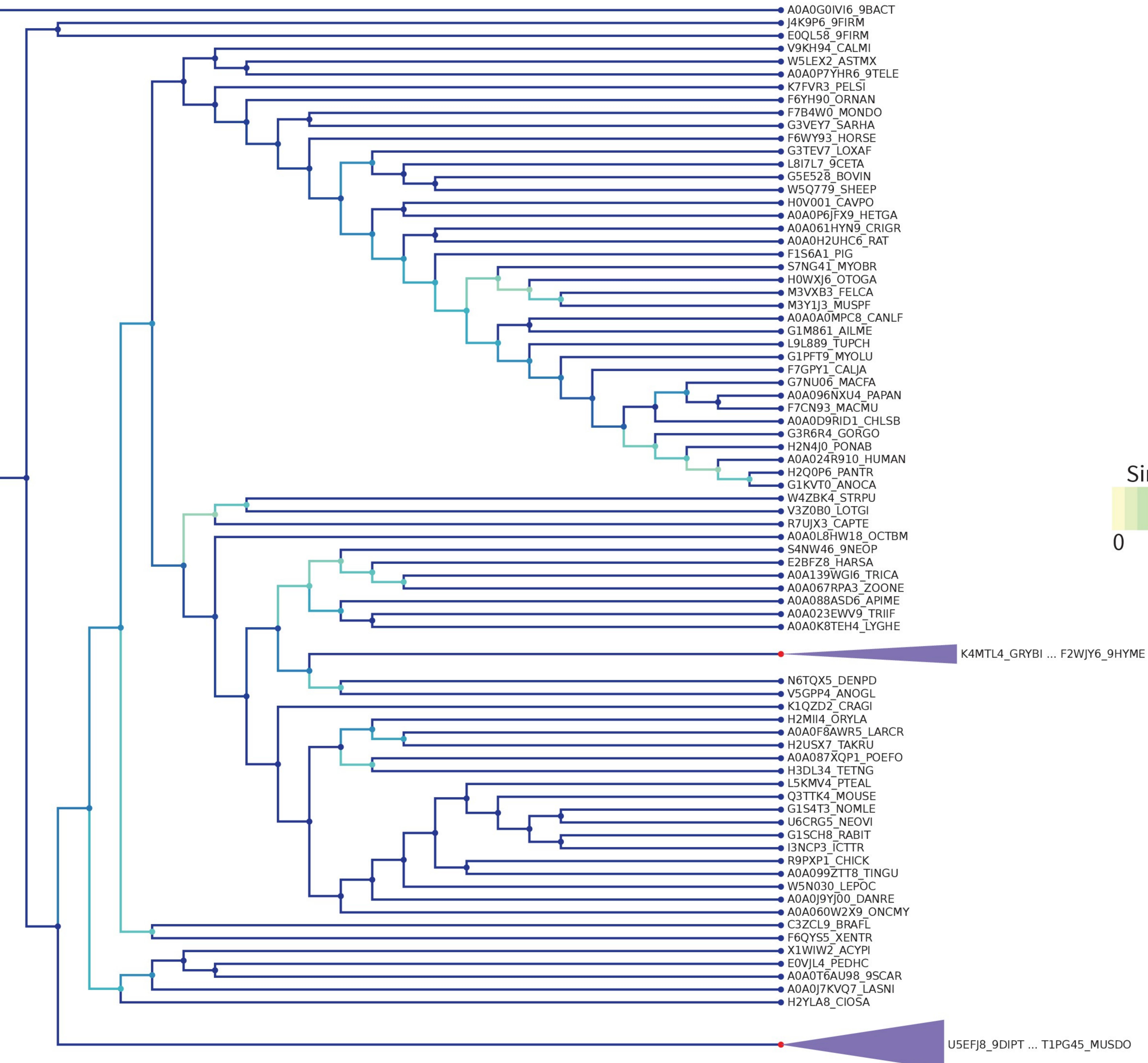
MUSCLE tree

Figure 2 Supplement 8

PRANK tree



MUSCLE tree



PRANK tree

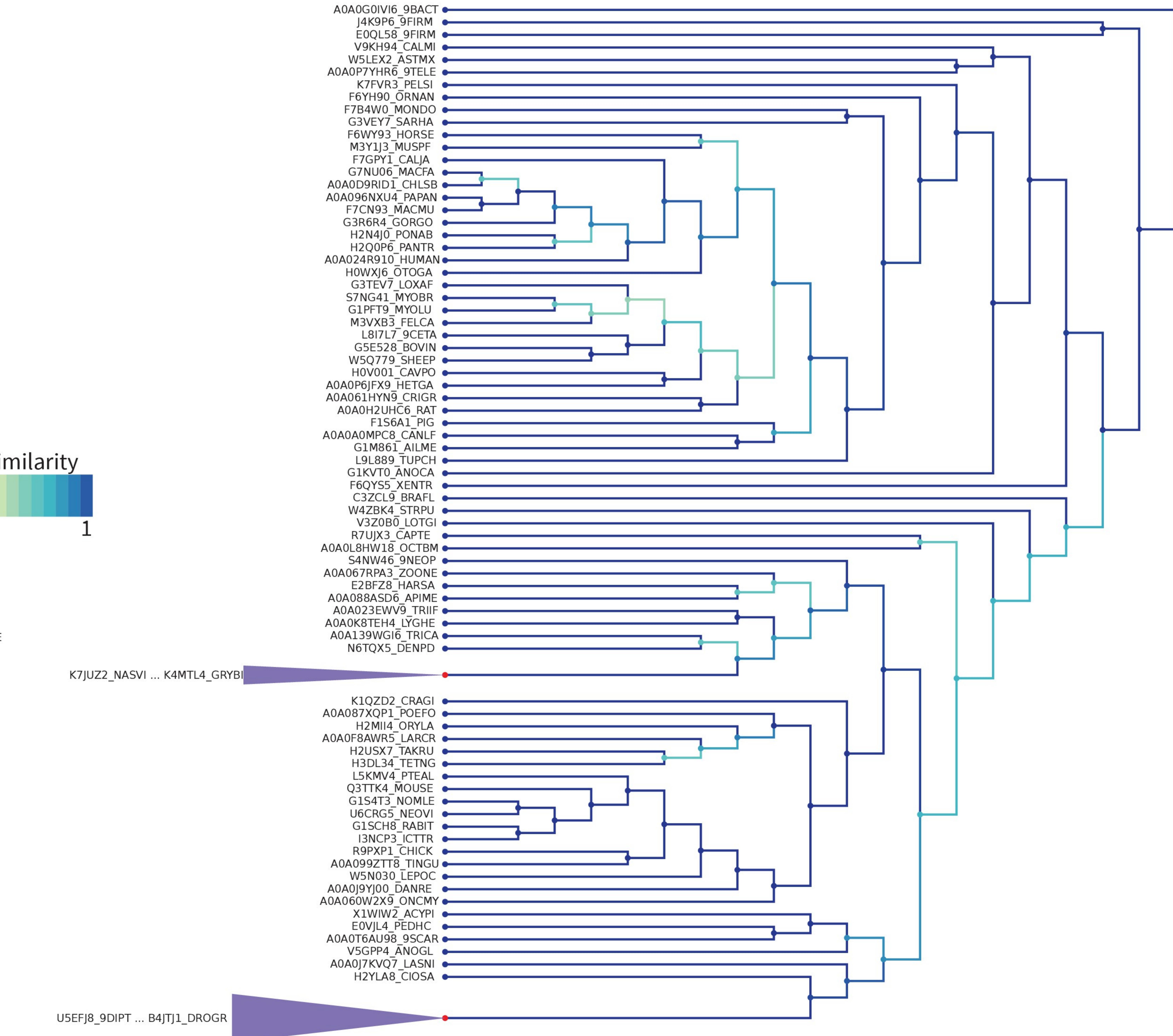


Figure 2 Supplement 10

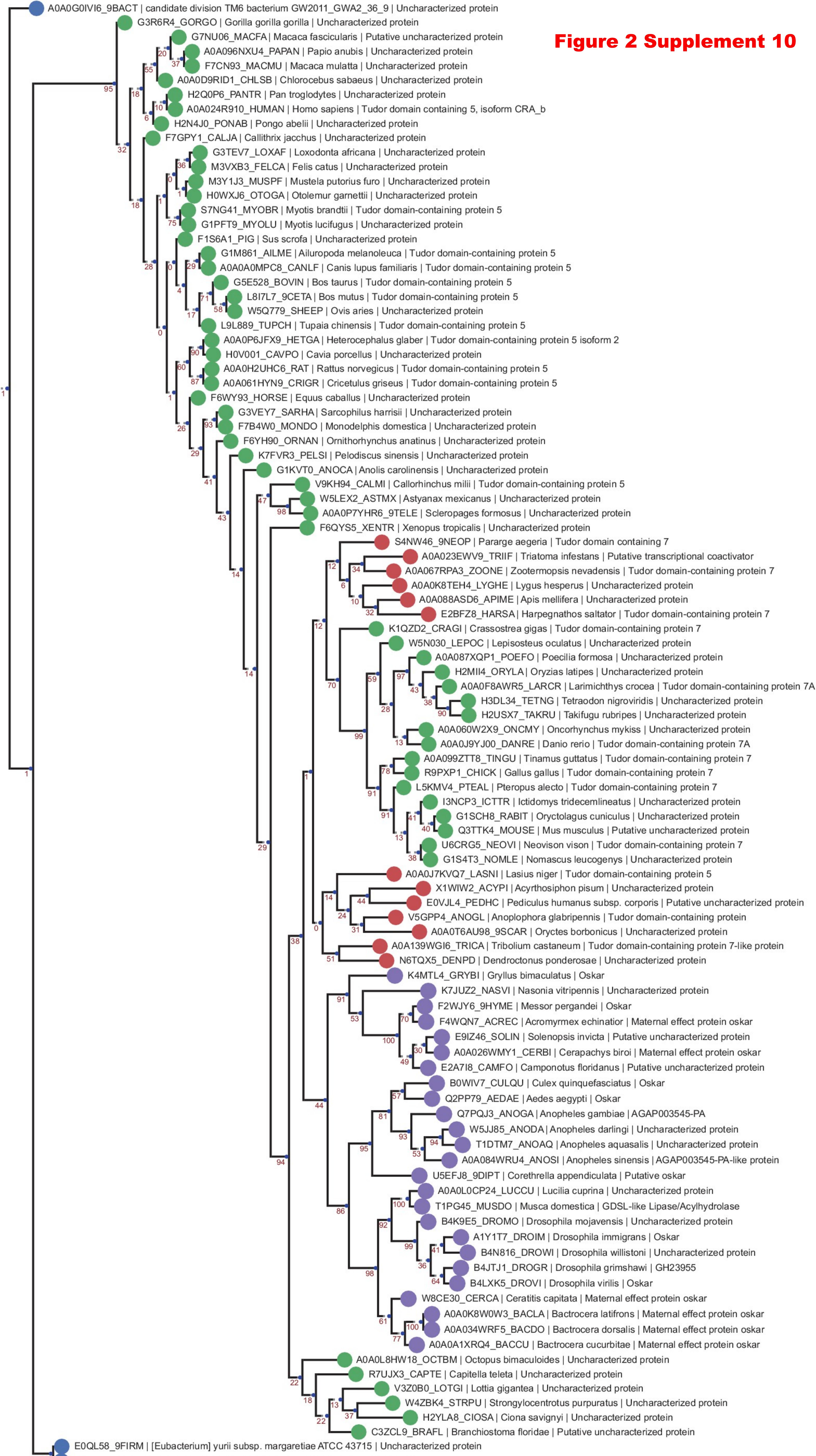
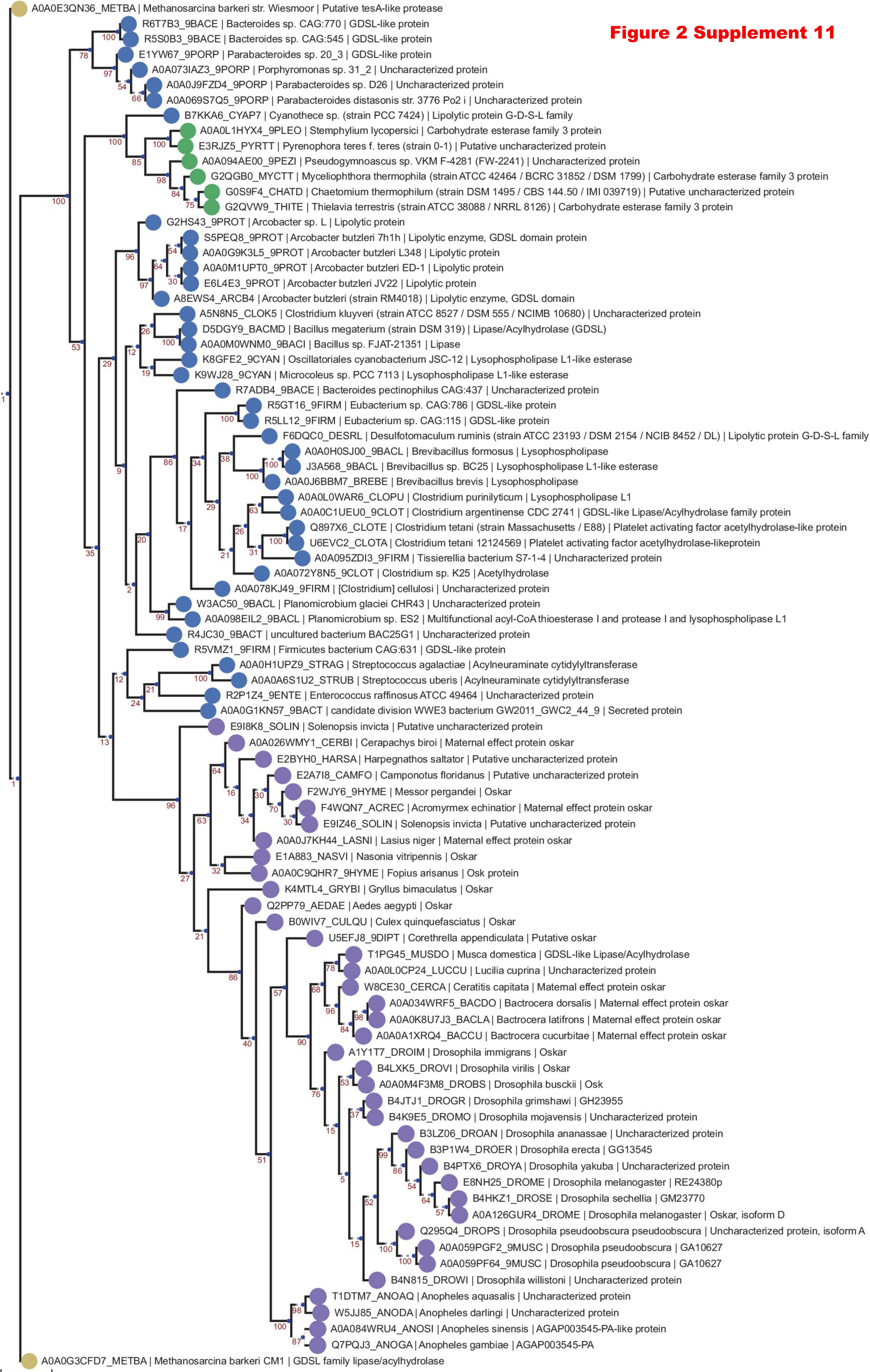


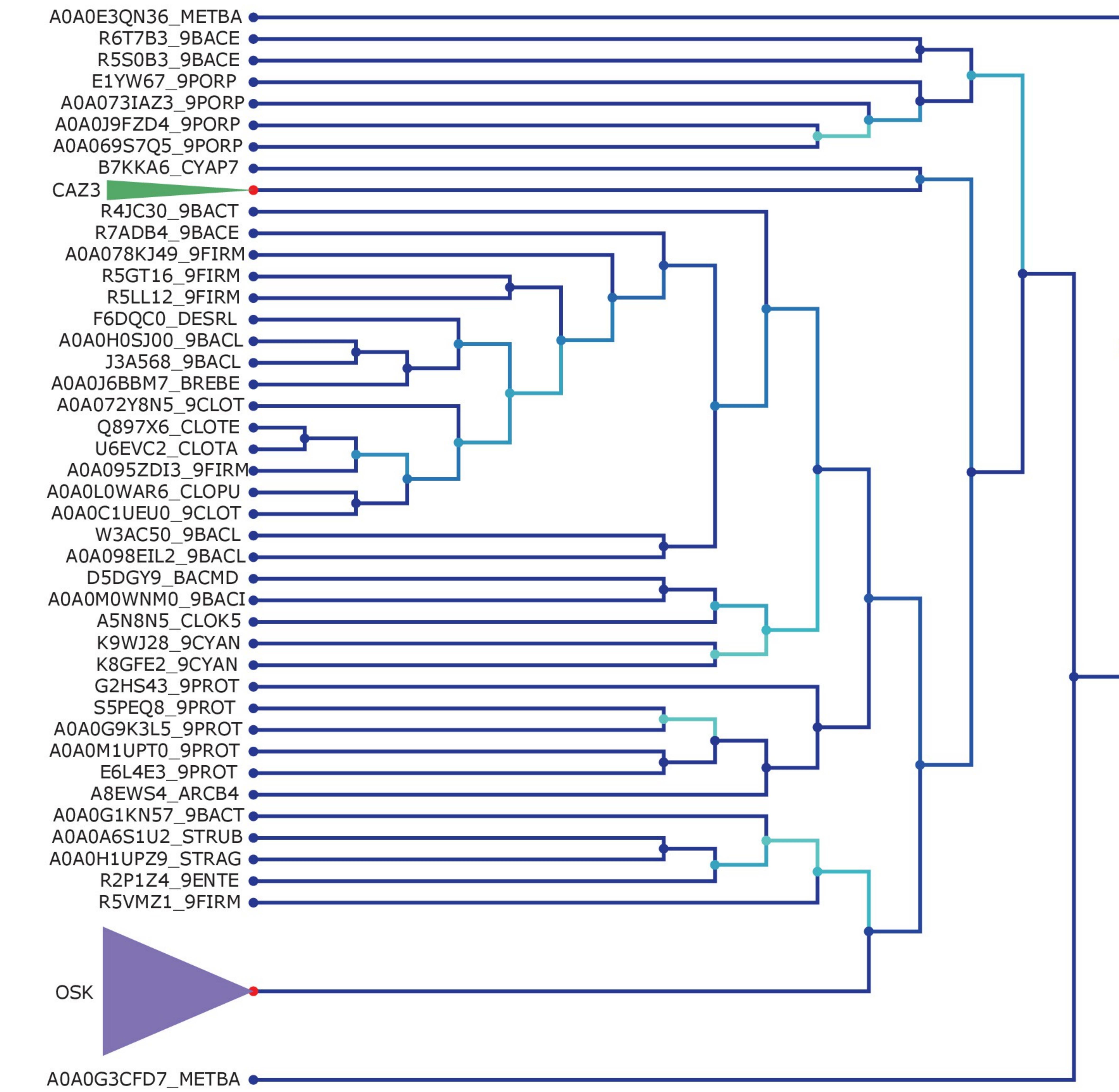
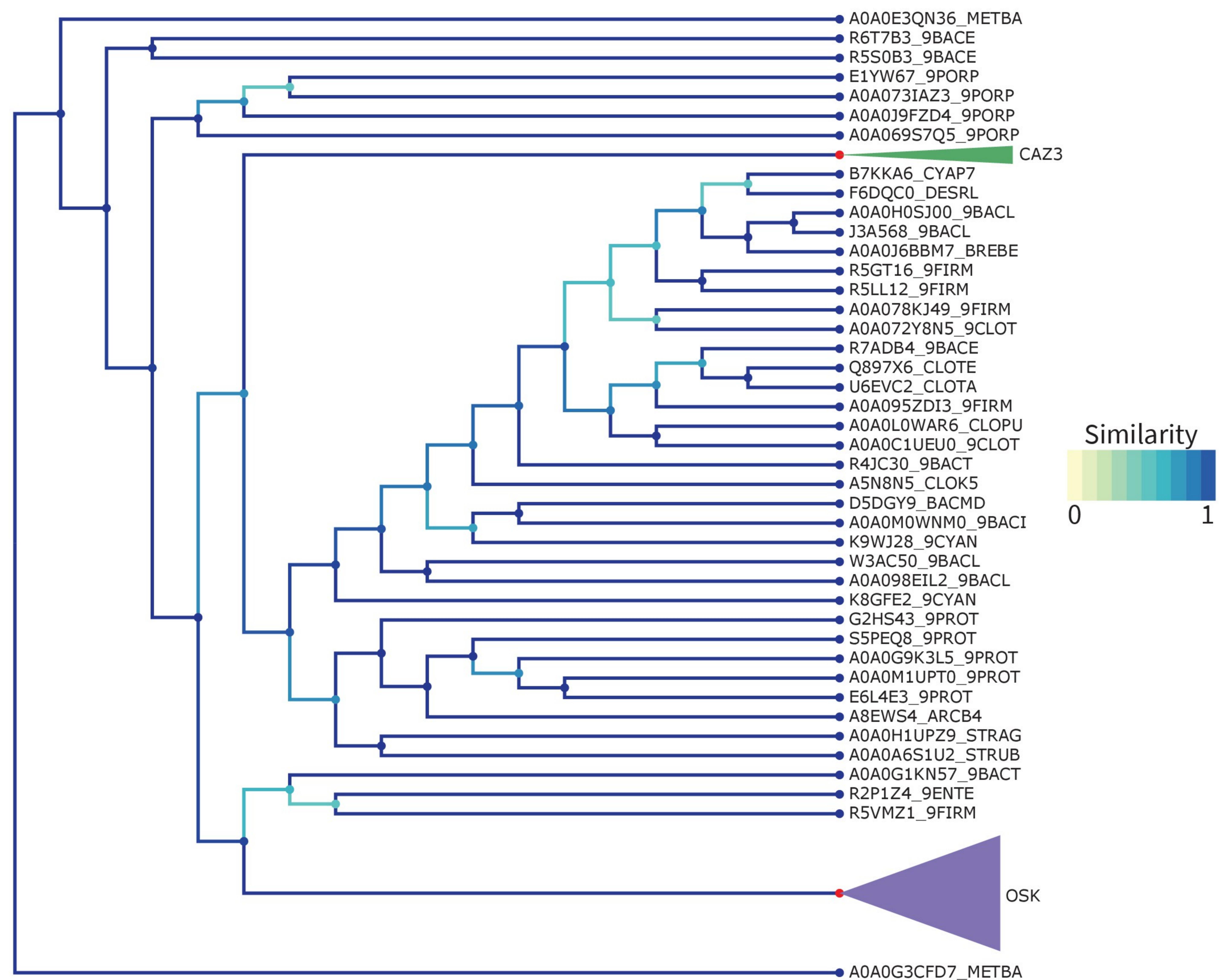
Figure 2 Supplement 11



MUSCLE tree

Figure 2 Supplement 12

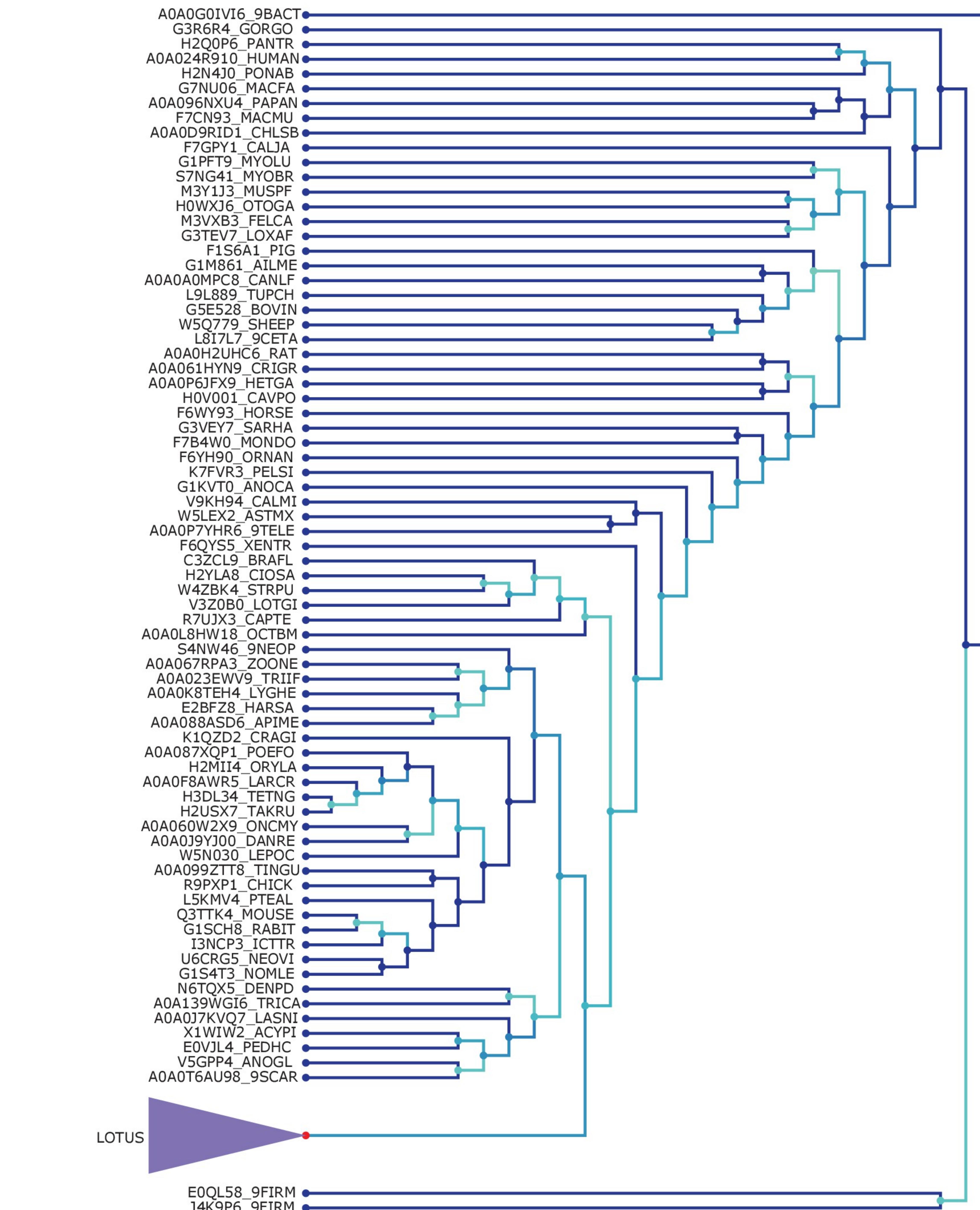
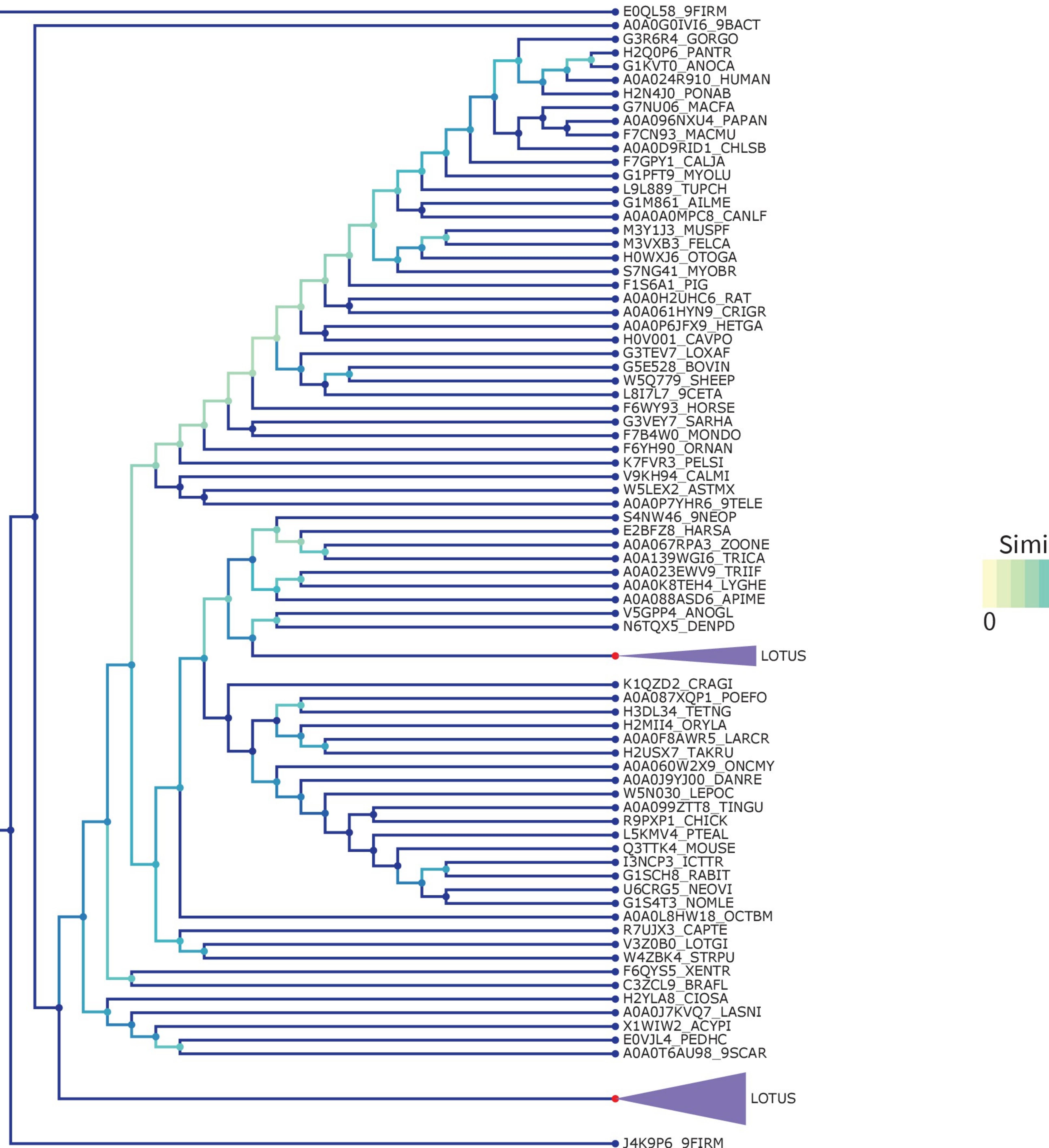
T COFFEE tree



MUSCLE tree

T COFFEE tree

Figure 2 Supplement 13



Supplementary File 1 for

Bacterial contribution to genesis of the novel germ line determinant *oskar*

Leo Blondel, Tamsin E. M. Jones and Cassandra G. Extavour

The Supplementary Information for this paper consists of the following elements:

Supplementary File 1

THIS PDF DOCUMENT: FILE NAME BLONDEL_JONES_EXTAVOUR_HGT_SUPPLEMENTARY_FILE_1

1. Supplementary Tables

- a. Supplementary File 1A: List of genomes and transcriptomes used for automated *oskar* search.
 - b. Supplementary File 1B: List of *oskar* sequences used in the final alignment.
 - c. Supplementary File 1C: List of sequences used for phylogenetic analysis of the LOTUS domain.
 - d. Supplementary File 1D: List of sequences used for phylogenetic analysis of the OSK domain.
 - e. Supplementary File 1E: List of genomes analyzed for codon use.

Source Data 1

ZIPPED FOLDER: DOWNLOAD HERE

https://www.dropbox.com/sh/zqf6kpo0kzav7xp/AAC5WPVm9lrDrZHqZg_RlsiTa?dl=0

1. Subfolder **Alignments**: All sequences identified and analyzed in this study, in FASTA format and with corresponding Alignments
 2. Subfolder **BLAST search results**: Results of BLASTP searches with full length Oskar, OSK or LOTUS domains as queries
 3. Subfolder **Data**: Necessary files for running the different IPython notebooks:
 - a. Subfolder **HMM**: HMM models used for iterative searching for sequences similar to full-length Oskar, LOTUS and OSK domains
 - b. Subfolder **Taxonomy**: Conversion table for UniProt ID to taxon information.
(uniprot_ID_taxa.tsv)
 - c. Subfolder **Trees**: Contains the tree files obtained from
 - i. RaxML phylogenetic analyses of the OSK and LOTUS domains aligned with MUSCLE, T-Coffee or PRANK
 - ii. MrBayes phylogenetic analyses of the OSK and LOTUS domains aligned with MUSCLE
 - iii. SOWHAT analyses.

Scripts

All scripts used herein are hosted on GitHub at https://github.com/extavourlab/Oskar_HGT

44
45
46
47
48
49
50
51

Supplementary File 1A: List of genomes and transcriptomes used for automated *oskar* search.
List of genomes and transcriptomes that were downloaded, annotated, and searched for *oskar* sequences (see “*Hidden Markov Model (HMM) generation and alignments of the OSK and LOTUS domains*” in Methods). The table reports the database provenance (NCBI genome or TSA, or 1KITE database) and the accession number. The TSA accession ID can be searched using the NCBI TSA browser here: <https://www.ncbi.nlm.nih.gov/Traces/wgs/?view=TSAs>.

Type	Accession ID	Organism Name
Genome	GCA_000001765.2	<i>Drosophila pseudoobscura pseudoobscura</i>
Genome	GCA_000002325.2	<i>Nasonia vitripennis</i>
Genome	GCA_000002335.2	<i>Tribolium castaneum</i>
Genome	GCA_000004775.1	<i>Nasonia giraulti</i>
Genome	GCA_000004795.1	<i>Nasonia longicornis</i>
Genome	GCA_000005115.1	<i>Drosophila ananassae</i>
Genome	GCA_000005135.1	<i>Drosophila erecta</i>
Genome	GCA_000005195.1	<i>Drosophila persimilis</i>
Genome	GCA_000005245.1	<i>Drosophila virilis</i>
Genome	GCA_000005925.1	<i>Drosophila willistoni</i>
Genome	GCA_000005975.1	<i>Drosophila yakuba</i>
Genome	GCA_000006295.1	<i>Pediculus humanus corporis</i>
Genome	GCA_000143395.2	<i>Atta cephalotes</i>
Genome	GCA_000147175.1	<i>Camponotus floridanus</i>
Genome	GCA_000147195.1	<i>Harpegnathos saltator</i>
Genome	GCA_000149185.1	<i>Mayetiola destructor</i>
Genome	GCA_000151625.1	<i>Bombyx mori</i>
Genome	GCA_000151715.1	<i>Bombyx mori</i>
Genome	GCA_000181055.3	<i>Rhodnius prolixus</i>
Genome	GCA_000184785.1	<i>Apis florea</i>
Genome	GCA_000187875.1	<i>Daphnia pulex</i>
Genome	GCA_000187915.1	<i>Pogonomyrmex barbatus</i>
Genome	GCA_000188075.1	<i>Solenopsis invicta</i>
Genome	GCA_000209185.1	<i>Culex quinquefasciatus</i>
Genome	GCA_000211455.3	<i>Anopheles darlingi</i>
Genome	GCA_000214255.1	<i>Bombus terrestris</i>
Genome	GCA_000217595.1	<i>Linepithema humile</i>
Genome	GCA_000220665.2	<i>Drosophila ficusphila</i>
Genome	GCA_000220905.1	<i>Megachile rotundata</i>
Genome	GCA_000224215.2	<i>Drosophila kikkawai</i>
Genome	GCA_000224235.2	<i>Drosophila takahashii</i>
Genome	GCA_000233415.2	<i>Drosophila biarmipes</i>
Genome	GCA_000235995.1	<i>Danaus plexippus plexippus</i>
Genome	GCA_000236305.2	<i>Drosophila rhopaloa</i>

Genome	GCA_000239435.1	<i>Tetranychus urticae</i>
Genome	GCA_000239455.1	<i>Strigamia maritima</i>
Genome	GCA_000259055.1	<i>Drosophila simulans</i>
Genome	GCA_000262585.1	<i>Manduca sexta</i>
Genome	GCA_000262795.1	<i>Phlebotomus papatasii</i>
Genome	GCA_000265325.1	<i>Lutzomyia longipalpis</i>
Genome	GCA_000269505.2	<i>Drosophila miranda</i>
Genome	GCA_000281935.1	<i>Mengenilla moldrzyki</i>
Genome	GCA_000298335.1	<i>Drosophila albomicans</i>
Genome	GCA_000325945.1	<i>Plutella xylostella</i>
Genome	GCA_000330985.1	<i>Plutella xylostella</i>
Genome	GCA_000341935.1	<i>Cephus cinctus</i>
Genome	GCA_000344095.1	<i>Athalia rosae</i>
Genome	GCA_000347755.1	<i>Ceratitis capitata</i>
Genome	GCA_000349025.1	<i>Anopheles minimus</i>
Genome	GCA_000349045.1	<i>Anopheles stephensi</i>
Genome	GCA_000349065.1	<i>Anopheles quadriannulatus</i>
Genome	GCA_000349085.1	<i>Anopheles funestus</i>
Genome	GCA_000349105.1	<i>Anopheles epiroticus</i>
Genome	GCA_000349145.1	<i>Anopheles dirus</i>
Genome	GCA_000349165.1	<i>Anopheles christyi</i>
Genome	GCA_000349185.1	<i>Anopheles arabiensis</i>
Genome	GCA_000371365.1	<i>Musca domestica</i>
Genome	GCA_000439205.1	<i>Anopheles nili</i>
Genome	GCA_000441895.2	<i>Anopheles sinensis</i>
Genome	GCA_000469605.1	<i>Apis dorsata</i>
Genome	GCA_000472105.1	<i>Drosophila suzukii</i>
Genome	GCA_000473185.1	<i>Anopheles maculatus</i>
Genome	GCA_000473375.1	<i>Anopheles culicifacies</i>
Genome	GCA_000473445.2	<i>Anopheles farauti</i>
Genome	GCA_000473505.1	<i>Anopheles atroparvus</i>
Genome	GCA_000473525.2	<i>Anopheles melas</i>
Genome	GCA_000473845.2	<i>Anopheles merus</i>
Genome	GCA_000475195.1	<i>Diaphorina citri</i>
Genome	GCA_000500325.1	<i>Leptinotarsa decemlineata</i>
Genome	GCA_000503995.1	<i>Ceratosolen solmsi marchali</i>
Genome	GCA_000507165.1	<i>Ephemera danica</i>
Genome	GCA_000516895.1	<i>Locusta migratoria</i>
Genome	GCA_000599845.1	<i>Trichogramma pretiosum</i>
Genome	GCA_000611835.1	<i>Ooceraea biroi</i>
Genome	GCA_000648655.1	<i>Copidosoma floridanum</i>
Genome	GCA_000648675.1	<i>Cimex lectularius</i>

Genome	GCA_000648695.1	<i>Onthophagus taurus</i>
Genome	GCA_000648945.1	<i>Limnephilus lunatus</i>
Genome	GCA_000671735.1	<i>Glossina fuscipes fuscipes</i>
Genome	GCA_000688715.1	<i>Glossina pallidipes</i>
Genome	GCA_000688735.1	<i>Glossina austeni</i>
Genome	GCA_000695345.1	<i>Bactrocera tryoni</i>
Genome	GCA_000695645.1	<i>Pachypsylia venusta</i>
Genome	GCA_000696155.1	<i>Zootermopsis nevadensis</i>
Genome	GCA_000696205.1	<i>Oncopeltus fasciatus</i>
Genome	GCA_000696795.1	<i>Halyomorpha halys</i>
Genome	GCA_000696855.1	<i>Homalodisca vitripennis</i>
Genome	GCA_000699065.1	<i>Lucilia cuprina</i>
Genome	GCA_000762945.1	<i>Blattella germanica</i>
Genome	GCA_000775305.1	<i>Belgica antarctica</i>
Genome	GCA_000786065.1	<i>Piezodorus guildinii</i>
Genome	GCA_000786525.1	<i>Chironomus tentans</i>
Genome	GCA_000789215.2	<i>Bactrocera dorsalis</i>
Genome	GCA_000818775.1	<i>Glossina palpalis gambiensis</i>
Genome	GCA_000836235.1	<i>Papilio xuthus</i>
Genome	GCA_000934665.1	<i>Catajapyx aquilonaris</i>
Genome	GCA_000956155.1	<i>Cotesia vestalis</i>
Genome	GCA_000956235.1	<i>Wasmannia auropunctata</i>
Genome	GCA_000956255.1	<i>Anopheles punctulatus</i>
Genome	GCA_000956275.1	<i>Anopheles koliensis</i>
Genome	GCA_001014415.1	<i>Phortica variegata</i>
Genome	GCA_001014435.1	<i>Mayetiola destructor</i>
Genome	GCA_001014445.1	<i>Scaptodrosophila lebanonensis</i>
Genome	GCA_001014505.1	<i>Chironomus riparius</i>
Genome	GCA_001014625.1	<i>Bactrocera oleae</i>
Genome	GCA_001014835.1	<i>Lucilia sericata</i>
Genome	GCA_001014935.1	<i>Liriomyza trifolii</i>
Genome	GCA_001014945.1	<i>Clogmia albipunctata</i>
Genome	GCA_001015115.1	<i>Eutreta diana</i>
Genome	GCA_001015145.1	<i>Eristalis dimidiata</i>
Genome	GCA_001015175.1	<i>Megaselia abdita</i>
Genome	GCA_001015215.1	<i>Holcocephala fusca</i>
Genome	GCA_001015235.1	<i>Sphyracephala brevicornis</i>
Genome	GCA_001015335.1	<i>Stomoxys calcitrans</i>
Genome	GCA_001017275.1	<i>Calliphora vicina</i>
Genome	GCA_001017455.1	<i>Neobellieria bullata</i>
Genome	GCA_001017515.1	<i>Tephritis californica</i>
Genome	GCA_001017525.1	<i>Teleopsis dalmanni</i>

Genome	GCA_001017535.1	<i>Tipula oleracea</i>
Genome	GCA_001077435.1	<i>Glossina morsitans morsitans</i>
Genome	GCA_001186385.1	<i>Diuraphis noxia</i>
Genome	GCA_001188975.2	<i>Bactrocera oleae</i>
Genome	GCA_001272555.1	<i>Dufourea novaeangliae</i>
Genome	GCF_000142985.2	<i>Acyrtosiphon pisum</i>
Genome	GCF_000208615.1	<i>Ixodes scapularis</i>
Genome	GCF_000004015.3	<i>Aedes aegypti</i>
TSA	GBKU	<i>Trichoplusia ni</i>
TSA	GBRL	<i>Sitodiplosis mosellana</i>
TSA	GAAX	<i>Forficula auricularia</i>
TSA	GACV	<i>Philopotamus ludificatus</i>
TSA	GADM	<i>Drosophila malerkotliana malerkotliana</i>
TSA	GAEO	<i>Pissodes strobi</i>
TSA	GAFI	<i>Dendroctonus frontalis</i>
TSA	GAFR	<i>Polistes canadensis</i>
TSA	GAHN	<i>Drosophila serrata</i>
TSA	GAHP	<i>Ostrinia nubilalis</i>
TSA	GAHQ	<i>Ostrinia scapulalis</i>
TSA	GAIW	<i>Ganaspis sp. G1</i>
TSA	GAJA	<i>Leptopilina boulardi</i>
TSA	GAJC	<i>Leptopilina heterotoma</i>
TSA	GAKJ	<i>Sitodiplosis mosellana</i>
TSA	GAMD	<i>Anopheles aquasalis</i>
TSA	GANH	<i>Microplitis demolitor</i>
TSA	GANO	<i>Corethrella appendiculata</i>
TSA	GAPE	<i>Brassicogethes aeneus</i>
TSA	GAPT	<i>Ectopsocus briggsi</i>
TSA	GASE	<i>Prorhinotermes simplex</i>
TSA	GASG	<i>Yponomeuta evonymellus</i>
TSA	GASN	<i>Thermobia domestica</i>
TSA	GASO	<i>Tricholepidion gertschi</i>
TSA	GASQ	<i>Tetrix subulata</i>
TSA	GASS	<i>Platycentropus radiatus</i>
TSA	GAST	<i>Polyommatus icarus</i>
TSA	GASV	<i>Notostira elongata</i>
TSA	GASW	<i>Mantis religiosa</i>
TSA	GASX	<i>Folsomia candida</i>
TSA	GASY	<i>Dysseriocrania subpurpurella</i>
TSA	GATA	<i>Meloe violaceus</i>
TSA	GATB	<i>Metallyticus splendidus</i>
TSA	GATC	<i>Nemophora degeerella</i>

TSA	GATD	<i>Pogonognathellus</i> sp. AD-2013
TSA	GATG	<i>Corydalus cornutus</i>
TSA	GATH	<i>Bittacus pilicornis</i>
TSA	GATI	<i>Bombylius major</i>
TSA	GATJ	<i>Bibio marci</i>
TSA	GATU	<i>Baetis</i> sp. AD-2013
TSA	GATV	<i>Perla marginata</i>
TSA	GATW	<i>Aleochara curtula</i>
TSA	GATY	<i>Chrysis viridula</i>
TSA	GATZ	<i>Sminthurus viridis</i>
TSA	GAUE	<i>Anurida maritima</i>
TSA	GAUF	<i>Leuctra</i> sp. AD-2013
TSA	GAUG	<i>Meinertellus cundinamarcensis</i>
TSA	GAUH	<i>Panorpa vulgaris</i>
TSA	GAUI	<i>Subilla</i> sp. AD-2014
TSA	GAUM	<i>Machilis hrabei</i>
TSA	GAUN	<i>Cercopis vulnerata</i>
TSA	GAUO	<i>Velia caprai</i>
TSA	GAUV	<i>Acanthosoma haemorrhoidale</i>
TSA	GAUW	<i>Apachyus charteceus</i>
TSA	GAUX	<i>Ceuthophilus</i> sp. AD-2013
TSA	GAUY	<i>Gyrinus marinus</i>
TSA	GAUZ	<i>Stenobothrus lineatus</i>
TSA	GAVA	<i>Triarthria setipennis</i>
TSA	GAVB	<i>Triodia sylvina</i>
TSA	GAVM	<i>Hydroptila</i> sp. AD-2013
TSA	GAVV	<i>Pseudomallada prasinus</i>
TSA	GAVW	<i>Epiophlebia superstes</i>
TSA	GAVX	<i>Timema cristinae</i>
TSA	GAWC	<i>Aretaon asperimus</i>
TSA	GAWD	<i>Medauroidea extradentata</i>
TSA	GAWE	<i>Ramulus artemis</i>
TSA	GAWF	<i>Sipyloidea sipylus</i>
TSA	GAWG	<i>Extatosoma tiaratum</i>
TSA	GAWK	<i>Ceratophyllus gallinae</i>
TSA	GAWM	<i>Culicoides sonorensis</i>
TSA	GAWP	<i>Grylloblatta bifratrilecta</i>
TSA	GAWQ	<i>Okanagana villosa</i>
TSA	GAWR	<i>Menopon gallinae</i>
TSA	GAWS	<i>Periplaneta americana</i>
TSA	GAWT	<i>Empusa pennata</i>
TSA	GAWU	<i>Aposthonia japonica</i>

TSA	GAWW	<i>Tenthredo koehleri</i>
TSA	GAWX	<i>Trialeurodes vaporariorum</i>
TSA	GAWZ	<i>Gryllotalpa sp. AD-2013</i>
TSA	GAXA	<i>Isonychia bicolor</i>
TSA	GAXB	<i>Tanzaniophasma sp. AD-2013</i>
TSA	GAXC	<i>Thrips palmi</i>
TSA	GAXE	<i>Acerentomon sp. AD-2013</i>
TSA	GAXF	<i>Planococcus citri</i>
TSA	GAXG	<i>Gynaikothrips ficorum</i>
TSA	GAXH	<i>Parides eurimedes</i>
TSA	GAXI	<i>Tetrodontophora bielanensis</i>
TSA	GAXM	<i>Mischocyttarus flavitarsis</i>
TSA	GAXO	<i>Argochrysis armilla</i>
TSA	GAXS	<i>Pepsis grossa</i>
TSA	GAXT	<i>Crioscolia alcione</i>
TSA	GAXW	<i>Euroleon nostras</i>
TSA	GAXX	<i>Rhyacophila fasciata</i>
TSA	GAXZ	<i>Trichocera saltator</i>
TSA	GAYA	<i>Zorotypus caudelli</i>
TSA	GAYB	<i>Zygaena fausta</i>
TSA	GAYC	<i>Osmylus fulvicephalus</i>
TSA	GAYD	<i>Blaberus atropos</i>
TSA	GAYE	<i>Frankliniella cephalica</i>
TSA	GAYH	<i>Conwentzia psociformis</i>
TSA	GAYI	<i>Xenophysella greensladeae</i>
TSA	GAYJ	<i>Atelura formicaria</i>
TSA	GAYK	<i>Boreus hyemalis</i>
TSA	GAYL	<i>Cosmioperla kuna</i>
TSA	GAYM	<i>Calopteryx splendens</i>
TSA	GAYN	<i>Campodea augens</i>
TSA	GAYO	<i>Cordulegaster boltonii</i>
TSA	GAYP	<i>Ctenocephalides felis</i>
TSA	GAYQ	<i>Forficula auricularia</i>
TSA	GAYV	<i>Liposcelis bostrychophila</i>
TSA	GAYY	<i>Acanthocasuarina muellerianae</i>
TSA	GAYZ	<i>Ranatra linearis</i>
TSA	GAZA	<i>Haploembia palaui</i>
TSA	GAZB	<i>Lepicerus sp. AD-2013</i>
TSA	GAZD	<i>Lipara lucens</i>
TSA	GAZE	<i>Mastotermes darwiniensis</i>
TSA	GAZF	<i>Essigella californica</i>
TSA	GAZG	<i>Eurylophella sp. AD-2013</i>

TSA	GAZH	<i>Inocellia crassicornis</i>
TSA	GAZM	<i>Stylops melittae</i>
TSA	GAZN	<i>Cryptocercus wrighti</i>
TSA	GAZQ	<i>Aretaon asperrimus</i>
TSA	GAZT	<i>Prosatirthia teretrirostris</i>
TSA	GBAB	<i>Musca domestica</i>
TSA	GBBP	<i>Teleopsis dalmanni</i>
TSA	GBCX	<i>Dastarcus helophoroides</i>
TSA	GBDM	<i>Helicoverpa armigera</i>
TSA	GBGV	<i>Polistes metricus</i>
1KITE	INSnfrTAJRAAPEI	<i>Machilis hrabei</i>
1KITE	INSnfrTAFRAAPEI	<i>Meinertellus cundinamarcensis</i>
1KITE	INSfrgTAVRAAPEI	<i>Blaberus atropos</i>
1KITE	INSfrgTAARAAPEI	<i>Periplaneta americana</i>
1KITE	INSytvTCDRAAPEI	<i>Cryptocercus wrighti</i>
1KITE	INSbttTCRAAPEI	<i>Arrhenodes minutus</i>
1KITE	INSnfrTBERAAPEI	<i>Gyrinus marinus</i>
1KITE	INSytvTAJRAAPEI	<i>Lepicerus sp.</i>
1KITE	INShauTAYRAAPEI	<i>Meloe violaceus</i>
1KITE	INShauTBERAAPEI	<i>Aleochara curtula</i>
1KITE	INSbttTIRRAAPEI	<i>Folsomia candida</i>
1KITE	INSnfrTAIRRAAPEI	<i>Anurida maritima</i>
1KITE	INSjdsTAKRAAPEI	<i>Tetrodontophora bielanensis</i>
1KITE	INShauTAFRAAPEI	<i>Sminthurus viridis/nigromaculatus</i>
1KITE	INShauTAJRAAPEI	<i>Pogonognathellus longicornis/flavescens</i>
1KITE	INSfrgTALRAAPEI	<i>Apachyus charteceus</i>
1KITE	INSjdsTBNRAAPEI	<i>Forficula auricularia</i>
1KITE	INStmbTABRAAPEI	<i>Campodea augens</i>
1KITE	INSjdsTANRAAPEI	<i>Occasjapyx japonicus</i>
1KITE	INSbusTBKRAAPEI	<i>Bombylius major</i>
1KITE	INSytvTBWRAAPEI	<i>Lipara lucens</i>
1KITE	INSnfrTBFRRAAPEI	<i>Triarthria setipennis</i>
1KITE	INSbusTBCRABPEI	<i>Bibio marci</i>
1KITE	INSjdsTBERAAPEI	<i>Trichocera saltator</i>
1KITE	INSfrgTAZRAAPEI	<i>Aposthonia japonica</i>
1KITE	INSytvTAHRAAPEI	<i>Haploembia palaui</i>
1KITE	INShauTAKRAAPEI	<i>Baetis sp.</i>
1KITE	INSytvTCERAAPEI	<i>Eurylophella sp.</i>
1KITE	INSnfrTAKRAAPEI	<i>Ephemera danica</i>
1KITE	INSjdsTAGRAAPEI	<i>Isonychia bicolor</i>
1KITE	INSfrgTAKRAAPEI	<i>Galloisiana yuasai</i>
1KITE	INSnfrTBKRAAPEI	<i>Grylloblatta bifratrilecta</i>

1KITE	INSnfrTANRAAPEI	<i>Cercopis vulnerata</i>
1KITE	INSnfrTBLRAAPEI	<i>Okanagana villosa</i>
1KITE	INSfrgTBCRAAPEI	<i>Nilaparvata lugens</i>
1KITE	INSjdsTARRAAPEI	<i>Xenophysella greensladeae</i>
1KITE	INSnfrTAPRAAPEI	<i>Acanthosoma haemorrhoidale</i>
1KITE	INShauTAPRAAPEI	<i>Notostira elongata</i>
1KITE	INSytvTANRAAPEI	<i>Ranatra linearis</i>
1KITE	INSnfrTAORAAPEI	<i>Velia caprai</i>
1KITE	INSfrgTAPRAAPEI	<i>Trialeurodes vaporariorum</i>
1KITE	INSytvTBHRAAPEI	<i>Essigella californica</i>
1KITE	INSjdsTAIRAAPEI	<i>Planococcus citri</i>
1KITE	INSytvTALRAAPEI	<i>Acanthocasuarina muellerianae</i>
1KITE	INSnfrTAQRAAPEI	<i>Cotesia vestalis</i>
1KITE	INShauTAQRABPEI	<i>Chrysis viridula</i>
1KITE	INSjdsTAURAAPEI	<i>Leptopilina clavipes</i>
1KITE	INSnfrTAARAAPEI	<i>Orussus abietinus</i>
1KITE	INSfrgTATRAAPEI	<i>Tenthredo koehleri</i>
1KITE	INStmbTBPRAAPEI	<i>Mastotermes darwiniensis</i>
1KITE	INSbusTBMRAAPEI	<i>Prorhinotermes simplex</i>
1KITE	INShauTABRAAPEI	<i>Nemophora degeerella</i>
1KITE	INSbusTBDRAAPEI	<i>Dysserocrania subpurpurella</i>
1KITE	INSnfrTAVRAAPEI	<i>Triodia sylvina</i>
1KITE	INShauTBGRAAPEI	<i>Polyommatus icarus</i>
1KITE	INSjdsTAJRAAPEI	<i>Parides eurimedes</i>
1KITE	INShauTBFRAAPEI	<i>Yponomeuta evonymellus</i>
1KITE	INSjdsTAWRAAPEI	<i>Zygaena fausta</i>
1KITE	INSfrgTASRAAPEI	<i>Empusa pennata</i>
1KITE	INShauTAARAAPEI	<i>Mantis religiosa</i>
1KITE	INShauTAMRAAPEI	<i>Metallyticus splendidus</i>
1KITE	INSfrgTBBRAAPEI	<i>Tanzaniophasma sp.</i>
1KITE	INSbtTARAAPEI	<i>Bittacus pilicornis</i>
1KITE	INStmbTAWRAAPEI	<i>Boreus hyemalis</i>
1KITE	INShauTACRAAPEI	<i>Panorpa vulgaris</i>
1KITE	INSbtTKRAAPEI	<i>Corydalus cornutus</i>
1KITE	INSnfrTARRAAPEI	<i>Pseudomallada prasinus</i>
1KITE	INSjdsTBQRAAPEI	<i>Conwentzia psociformis</i>
1KITE	INSjdsTATRAAPEI	<i>Euroleon nostras</i>
1KITE	INSjdsTBJRAAPEI	<i>Osmylus fulvicephalus</i>
1KITE	INSjdsTBHRAAPEI	<i>Cordulegaster boltonii</i>
1KITE	INSfrgTAHRAAPEI	<i>Epiophlebia superstes</i>
1KITE	INStmbTAARAAPEI	<i>Calopteryx splendens</i>
1KITE	INSnfrTAMRAAPEI	<i>Stenobothrus lineatus</i>

1KITE	INStmbTBCRBAPEI	<i>Prosarthria teretirostris</i>
1KITE	INShauTANRAAPEI	<i>Tetrix subulata</i>
1KITE	INSfrgTAXRABPEI	<i>Gryllotalpa sp.</i>
1KITE	INSnfrTBIRAAPEI	<i>Ceuthophilus sp.</i>
1KITE	INStmbTBERAAPEI	<i>Aretaon asperimus</i>
1KITE	INSfrgTAORAAPEI	<i>Peruphasma schultei</i>
1KITE	INSnfrTBPRAAPEI	<i>Timema cristinae</i>
1KITE	INStmbTBFRRAAPEI	<i>Cosmioperla kuna</i>
1KITE	INSnfrTALRAAPEI	<i>Leuctra sp.</i>
1KITE	INShauTALRAAPEI	<i>Perla marginata</i>
1KITE	INSjdsTAHRAAPEI	<i>Acerentomon sp.</i>
1KITE	INSfrgTAFRRAAPEI	<i>Menopon gallinae</i>
1KITE	INSytvTCFRAAPEI	<i>Ectopsocus briggsi</i>
1KITE	INStmbTBGRAAPEI	<i>Liposcelis bostrychophila</i>
1KITE	INSnfrTAGRAAPEI	<i>Xanthostigma xanthostigma</i>
1KITE	INSnfrTBARAAPEI	<i>Ceratophyllus gallinae</i>
1KITE	INStmbTAYRAAPEI	<i>Ctenocephalides felis</i>
1KITE	INSytvTBKRAAPEI	<i>Stylops melittae</i>
1KITE	INSjdsTADRAAPEI	<i>Gynaikothrips ficorum</i>
1KITE	INSjdsTABRAAPEI	<i>Frankliniella cephalica</i>
1KITE	INSjdsTACRAAPEI	<i>Thrips palmi</i>
1KITE	INSnfrTBJRAAPEI	<i>Hydroptila actia/argosa</i>
1KITE	INSjdsTBSRAAPEI	<i>Rhyacophila fasciata</i>
1KITE	INSjdsTAQRAAPEI	<i>Zorotypus caudelli</i>
1KITE	INSjdsTAVRAAPEI	<i>Atelura formicaria</i>
1KITE	INSbttTJRAAPEI	<i>Tricholepidion gertschi</i>
1KITE	INSbttTSRAAPEI	<i>Thermobia domestica</i>

53
54
55
56
57
58
59**Supplementary File 1B: List of *oskar* sequences used in the final alignment.**

List of accession numbers and database provenance of the sequences used in the final alignments of Oskar analysed herein. The table contains the database provenance (*Type*), the database accession number (*ID*), the species, family and order, and extraction notes. In the “Annotation” Column, P = homolog identified by pipeline; DB = homolog identified by database annotation. *Sequence recomposed from two transcripts: GBCX01024638.1 and GBCX01024637.

Type	ID	Species	Family	Order	Annotation	Note
TSA	GAWC01068734.1	<i>Aretaon asperrimus</i>	Heteropterygidae	Phasmatodea	P	HMMER
TSA	GAKJ01010751.1	<i>Sitodiplosis mosellana</i>	Cecidomyiidae	Diptera	P	HMMER
TSA	GATI01010233.1	<i>Bombylius major</i>	Bombyliidae	Diptera	P	HMMER
TSA	GAWW01000144.1	<i>Tenthredo koehleri</i>	Tenthredinidae	Hymenoptera	P	HMMER
TSA	GAIW01009539.1	<i>Ganaspis sp.</i>	Figitidae	Hymenoptera	P	HMMER
TSA	GATY01008637.1	<i>Chrysis viridula</i>	Chrysididae	Hymenoptera	P	HMMER
TSA	GAXM01030263.1	<i>Mischocyttarus flavitarsis</i>	Vespidae	Hymenoptera	P	HMMER
TSA	GAFR01040300.1	<i>Polistes canadensis</i>	Vespidae	Hymenoptera	P	HMMER
TSA	GAZD01106195.1	<i>Lipara lucens</i>	Chloropidae	Diptera	P	HMMER
TSA	GAVA01002196.1	<i>Triarthria setipennis</i>	Tachinidae	Diptera	P	HMMER
TSA	GACV01001831.1	<i>Philopotamus ludificatus</i>	Philopotaminae	Trichoptera	P	HMMER
TSA	GAPE01019095.1	<i>Brassicogethes aeneus</i>	Nitidulidae	Coleoptera	P	HMMER
TSA	GBCX01024638.1_7.1	<i>Dastarcus helophoroides</i> *	Bothrideridae	Coleoptera	P	BLAST 1
TSA	GAKJ01010751.1	<i>Sitodiplosis mosellana</i>	Cecidomyiidae	Diptera	P	HMMER
TSA	GAWM01006639.1	<i>Culicoides sonorensis</i>	Ceratopogonidae	Diptera	P	HMMER
TSA	GAMD01000859.1	<i>Anopheles aquasalis</i>	Culicidae	Diptera	P	HMMER
TSA	GBEO01001325.1	<i>Anopheles sinensis</i>	Culicidae	Diptera	P	HMMER
TSA	GAKP01002609.1	<i>Bactrocera dorsalis</i>	Tephritidae	Diptera	P	HMMER
TSA	GAXM01030263.1	<i>Mischocyttarus flavitarsis</i>	Vespidae	Hymenoptera	P	HMMER
TSA	GAFR01040300.1	<i>Polistes canadensis canadensis</i>	Vespidae	Hymenoptera	P	HMMER
TSA	GBGV01010610.1	<i>Polistes metricus</i>	Vespidae	Hymenoptera	P	HMMER
TSA	GAIW01011550.1	<i>Ganaspis sp. G1</i>	Figitidae	Hymenoptera	P	HMMER
TSA	GAJC01011221.1	<i>Leptopilina heterotoma</i>	Figitidae	Hymenoptera	P	HMMER
TSA	GAIW01009539.1	<i>Ganaspis sp. G1</i>	Figitidae	Hymenoptera	P	HMMER

TSA	GAJA01020544.1	<i>Leptopilina boulardi</i>	Figitidae	Hymenoptera	P	HMMER
TSA	GAJC01009625.1	<i>Leptopilina heterotoma</i>	Figitidae	Hymenoptera	P	HMMER
TSA	GAXO01016630.1	<i>Argochrysis armilla</i>	Chrysidae	Hymenoptera	P	HMMER
TSA	GAEO01004319.1	<i>Pissodes strobi</i>	Curculionidae	Coleoptera	P	HMMER
TSA	GBBP01080309.1	<i>Teleopsis dalmanni</i>	Diopsidae	Diptera	P	HMMER
TSA	GAVM01000124.1	<i>Hydroptila actia/argosa</i>	Hydroptilidae	Trichoptera	P	HMMER
TSA	GAWC01068728	<i>Aretaon asperrimus</i>	Heteropterygidae	Phasmatodea	P	HMMER
TSA	GBEO01001325.1	<i>Anopheles sinensis</i>	Culicidae	Diptera	P	HMMER
Genome	GCA_000349125.1	<i>Anopheles albimanus</i>	Culicidae	Diptera	P	Snap
Genome	GCA_000349065.1	<i>Anopheles quadriannulatus</i>	Culicidae	Diptera	P	Augustus
Genome	GCA_000349185.1	<i>Anopheles arabiensis</i>	Culicidae	Diptera	P	Augustus
Genome	GCA_000473845.2	<i>Anopheles merus</i>	Culicidae	Diptera	P	Augustus
Genome	GCA_000473375.1	<i>Anopheles culicifacies</i>	Culicidae	Diptera	P	Augustus
Genome	GCA_000349085.1	<i>Anopheles funestus</i>	Culicidae	Diptera	P	Augustus
Genome	GCA_000209185.1	<i>Culex quinquefasciatus</i>	Culicidae	Diptera	P	Snap
Genome	GCF_000004015.3	<i>Aedes aegypti</i>	Culicidae	Diptera	P	Augustus
Genome	GCA_001014445.1	<i>Scaptodrosophila lebanonensis</i>	Drosophilidae	Diptera	P	Snap
Genome	GCA_000005925.1	<i>Drosophila willistoni</i>	Drosophilidae	Diptera	P	Augustus
Genome	GCA_000005115.1	<i>Drosophila ananassae</i>	Drosophilidae	Diptera	P	Snap
Genome	GCA_000236285.2	<i>Drosophila bipectinata</i>	Drosophilidae	Diptera	P	Augustus
Genome	GCA_000224215.2	<i>Drosophila kikkawai</i>	Drosophilidae	Diptera	P	Snap
Genome	GCA_000236325.2	<i>Drosophila eugracilis</i>	Drosophilidae	Diptera	P	Augustus
Genome	GCA_000001215.4	<i>Drosophila melanogaster</i>	Drosophilidae	Diptera	P	Augustus
Genome	GCA_000259055.1	<i>Drosophila simulans</i>	Drosophilidae	Diptera	P	Snap
Genome	GCA_000005215.1	<i>Drosophila sechellia</i>	Drosophilidae	Diptera	P	Augustus
Genome	GCA_000005135.1	<i>Drosophila erecta</i>	Drosophilidae	Diptera	P	Augustus
Genome	GCA_000005975.1	<i>Drosophila yakuba</i>	Drosophilidae	Diptera	P	Augustus
Genome	GCA_000220665.2	<i>Drosophila ficusphila</i>	Drosophilidae	Diptera	P	Snap
Genome	GCA_000224195.2	<i>Drosophila elegans</i>	Drosophilidae	Diptera	P	Snap
Genome	GCA_000236305.2	<i>Drosophila rhopaloa</i>	Drosophilidae	Diptera	P	Augustus
Genome	GCA_000233415.2	<i>Drosophila biarmipes</i>	Drosophilidae	Diptera	P	Snap

Genome	GCA_000224235.2	<i>Drosophila takahashii</i>	Drosophilidae	Diptera	P	Augustus
Genome	GCA_000269505.2	<i>Drosophila miranda</i>	Drosophilidae	Diptera	P	Augustus
Genome	GCA_000001765.2	<i>Drosophila pseudoobscura</i>	Drosophilidae	Diptera	P	Snap
Genome	GCA_000005195.1	<i>Drosophila persimilis</i>	Drosophilidae	Diptera	P	Snap
Genome	GCA_000298335.1	<i>Drosophila albomicans</i>	Drosophilidae	Diptera	P	Snap
Genome	GCA_000005155.1	<i>Drosophila grimshawi</i>	Drosophilidae	Diptera	P	Snap
Genome	GCA_000005175.1	<i>Drosophila mojavensis</i>	Drosophilidae	Diptera	P	Snap
Genome	GCA_000005245.1	<i>Drosophila virilis</i>	Drosophilidae	Diptera	P	Snap
Genome	GCA_000371365.1	<i>Musca domestica</i>	Muscidae	Diptera	P	Augustus
Genome	GCA_000648655.1	<i>Copidosoma floridanum</i>	Encyrtidae	Hymenoptera	P	Blast
Protein	ABC54566.1	<i>Anopheles gambiae</i>	Culicidae	Diptera	DB	-
Protein	KYN09041	<i>Trachymyrmex cornetzi</i>	Formicidae	Hymenoptera	DB	-
Protein	KYM88541	<i>Atta colombica</i>	Formicidae	Hymenoptera	DB	-
Protein	XP_012060266	<i>Atta cephalotes</i>	Formicidae	Hymenoptera	DB	-
Protein	XP_011057669	<i>Acromyrmex echinatior</i>	Formicidae	Hymenoptera	DB	-
Protein	ADM07366	<i>Messor pergandei</i>	Formicidae	Hymenoptera	DB	-
Protein	XP_008556449	<i>Microplitis demolitor</i>	Microgastrinae	Hymenoptera	DB	-
Protein	XP_012229836	<i>Linepithema humile</i>	Dolichoderinae	Hymenoptera	DB	-
Protein	ABC54566	<i>Anopheles gambiae</i>	Culicidae	Diptera	DB	-
Protein	ACB20969	<i>Culex quinquefasciatus</i>	Culicidae	Diptera	DB	-
Protein	ABC41128	<i>Aedes aegypti</i>	Culicidae	Diptera	DB	-
Protein	XP_004529162	<i>Ceratitis capitata</i>	Tephritidae	Diptera	DB	-
Protein	NP_001234884	<i>Nasonia vitripennis</i>	Pteromalidae	Hymenoptera	DB	-
Protein	XM_011167600	<i>Solenopsis invicta</i>	Formicidae	Hymenoptera	DB	-
Protein	JR477371.1	<i>Rhynchophorus ferrugineus</i>	Curculionidae	Coleoptera	DB	-
Nucleotide	JO902149	<i>Aedes albopictus</i>	Aedes	Diptera	DB	-
Nucleotide	JO874052	<i>Aedes albopictus</i>	Aedes	Diptera	DB	-
Nucleotide	JO885398	<i>Aedes albopictus</i>	Aedes	Diptera	DB	-
1KITE	INStmbTBGRAAPEI-33	<i>Liposcelis bostrychophila</i>	Liposcelididae	Psocodea	P	Blast
1KITE	INSjdsTABRAAPEI-20	<i>Frankliniella cephalica</i>	Thripidae	Thysanoptera	P	Blast
1KITE	INSjdsTACRAAPEI-21	<i>Thrips palmi</i>	Thripidae	Thysanoptera	P	Blast

1KITE	INShauTABRAAPEI-93	<i>Nemophora degeerella</i>	Adelidae	Lepidoptera	P	Blast
1KITE	INSbttTARAAPEI-9	<i>Platycentropus radiatus</i>	Limnephilidae	Trichoptera	P	Blast
1KITE	INSnfrTALRAAPEI-31	<i>Leuctra sp.</i>	Leuctridae	Plecoptera	P	Blast
1KITE	INShauTAKRAAPEI-90	<i>Baetis pumilus</i>	Baetidae	Ephemeroptera	P	Blast
1KITE	INSjdsTADRAAPEI-22	<i>Gynaikothrips ficorum</i>	Phlaeothripidae	Thysanoptera	P	Blast
1KITE	INStmbTAWRAAPEI-13	<i>Boreus hyemalis</i>	Boreidae	Mecoptera	P	Blast
1KITE	INSytvTCDRAAPEI-35	<i>Cryptocercus wrighti</i>	Cryptocercidae	Blattodea	P	Blast
1KITE	INSnfrTAQRAAPEI-37	<i>Cotesia vestalis</i>	Braconidae	Hymenoptera	P	Blast
1KITE	INSjdsTAURAAPEI-62	<i>Leptopilina clavipes</i>	Figitidae	Hymenoptera	P	Blast

61 **Supplementary File 1C: List of sequences and their BLAST results used for phylogenetic analysis of the LOTUS domain.**

62 The sequences were obtained by searching the TrEMBL database using hmmsearch and the final HMM generated for LOTUS
 63 (Supplementary files: HMM>LOTUS.hmm). Reported are the UniProtID (*Accession Number*), the Domain and Phylum origin of the
 64 sequence, the E-value, score and bias given by hmmsearch, and the description of the target from UniProt. To obtain sequences for each
 65 entry, either search UniProt directly (<https://www.uniprot.org/>) or consult the final alignment in Supplementary Files:
 66 Alignments>LOTUS_TREE.fasta. Phylum abbreviations: A = Arthropoda; An = Annelida; E = Echinodermata; F = Firmicutes; M =
 67 Mollusca; T = Tunicata; V = Vertebrata; ? = unclassified

68

Accession ID	Domain	Phylum	E-value	score	bias	Description of Target
V3Z0B0_LOTGI	Eukarya	M	8.30E-32	120.8	0.1	Uncharacterized protein OS=Lottia gigantea GN=LOTGIDRAFT_236389 PE=4 SV=1
R7UJX3_CAPTE	Eukarya	An	4.00E-30	115.3	0	Uncharacterized protein OS=Capitella teleta GN=CAPTEDRAFT_218952 PE=4 SV=1
E9IZ46_SOLIN	Eukarya	A	2.80E-27	106.2	0.1	Putative uncharacterized protein (Fragment) OS=Solenopsis invicta GN=SINV_01516 PE=4 SV=1
K7JUZ2_NASVI	Eukarya	A	9.00E-27	104.6	0.2	Uncharacterized protein OS=Nasonia vitripennis GN=oskar PE=4 SV=1
F4WQN7_ACREC	Eukarya	A	1.30E-25	100.9	0.1	Maternal effect protein oskar OS=Acromyrmex echinatior GN=G5I_08127 PE=4 SV=1
W4ZBK4_STRPU	Eukarya	E	1.70E-25	100.5	0.1	Uncharacterized protein OS=Strongylocentrotus purpuratus GN=Sp-Tdrd5 PE=4 SV=1
E2A7I8_CAMFO	Eukarya	A	1.90E-25	100.4	2.1	Putative uncharacterized protein OS=Camponotus floridanus GN=EAG_03874 PE=4 SV=1
F2WJY6_9HYME	Eukarya	A	3.40E-25	99.6	0	Oskar (Fragment) OS=Messor pergandei PE=2 SV=1 Uncharacterized protein OS=Drosophila willistoni
B4N816_DROWI	Eukarya	A	8.50E-25	98.3	9.6	GN=Dwil\GK11116 PE=4 SV=2 Uncharacterized protein OS=Drosophila mojavensis
B4K9E5_DROMO	Eukarya	A	8.70E-25	98.2	0.6	Maternal effect protein oskar OS=Cerapachys biroi GN=Dmoj\GI10055 PE=4 SV=2
A0A026WMY1_CERBI	Eukarya	A	2.20E-24	97	0.1	GN=X777_01612 PE=4 SV=1 Putative uncharacterized protein OS=Branchiostoma floridae
C3ZCL9_BRAFL	Eukarya	A	5.40E-24	95.7	0	GN=BRAFLDRAFT_64001 PE=4 SV=1
B4LXK5_DROVI	Eukarya	A	8.00E-24	95.2	0.9	Oskar OS=Drosophila virilis GN=osk PE=4 SV=1
K4MTL4_GRYBI	Eukarya	A	6.70E-23	92.2	0	Oskar OS=Gryllus bimaculatus PE=2 SV=1 GH23955 OS=Drosophila grimshawi GN=Dgr\GH23955 PE=4
B4JTJ1_DROGR	Eukarya	A	8.30E-23	91.9	1.4	SV=1
A1Y1T7_DROIM	Eukarya	A	8.60E-23	91.9	0.8	Oskar OS=Drosophila immigrans GN=osk PE=4 SV=1
Q2PP79_AEDAE	Eukarya	A	1.70E-22	90.9	0	Oskar OS=Aedes aegypti PE=4 SV=1

W8CE30_CERCA	Eukarya	A	2.00E-22	90.7	0.3	Maternal effect protein oskar OS=Ceratitis capitata GN=OSKA PE=2 SV=1 GDSL-like Lipase/Acylhydrolase OS=Musca domestica PE=2
T1PG45_MUSDO	Eukarya	A	1.60E-21	87.8	1.2	SV=1 Uncharacterized protein OS=Octopus bimaculoides GN=OCBIM_22005378mg PE=4 SV=1
A0A0L8HW18_OCTBM	Eukarya	M	1.70E-21	87.7	7.5	Uncharacterized protein (Fragment) OS=Xenopus tropicalis PE=4 SV=1
F6QYS5_XENTR	Eukarya	V	6.10E-20	82.7	0.2	Oskar OS=Culex quinquefasciatus GN=CpipJ_CPIJ007471
B0WIV7_CULQU	Eukarya	A	8.90E-20	82.2	0.1	PE=2 SV=1 Maternal effect protein oskar OS=Bactrocera dorsalis GN=OSKA
A0A034WRF5_BACDO	Eukarya	A	1.30E-19	81.6	13.1	PE=4 SV=1
Q7PQJ3_ANOGA	Eukarya	A	1.40E-19	81.5	0	AGAP003545-PA OS=Anopheles gambiae GN=osk PE=4 SV=3 Uncharacterized protein OS=Acyrthosiphon pisum GN=LOC100162069 PE=4 SV=1
X1WIW2_ACYPI	Eukarya	A	1.70E-19	81.3	12.3	Putative transcriptional coactivator (Fragment) OS=Triatoma infestans PE=2 SV=1
A0A023EWV9_TRIIF	Eukarya	A	1.80E-19	81.2	0.2	Uncharacterized protein OS=Lucilia cuprina GN=FF38_12727
A0A0L0CP24_LUCCU	Eukarya	A	2.40E-19	80.8	0.5	PE=4 SV=1 Maternal effect protein oskar OS=Bactrocera latifrons GN=osk_1
A0A0K8W0W3_BACLA	Eukarya	A	5.00E-19	79.8	0	PE=4 SV=1 Maternal effect protein oskar OS=Bactrocera cucurbitae GN=osk
A0A0A1XRQ4_BACCU	Eukarya	A	7.20E-19	79.3	6.4	PE=4 SV=1 Uncharacterized protein OS=Apis mellifera GN=LOC726241
A0A088ASD6_APIME	Eukarya	A	2.90E-18	77.3	0.2	PE=4 SV=1 Tudor domain-containing protein 7 OS=Crassostrea gigas GN=CGI_10018436 PE=4 SV=1
K1QZD2_CRAGI	Eukarya	M	1.60E-17	75	0	Tudor domain-containing protein 7-like protein OS=Tribolium castaneum GN=TcasGA2_TC034722 PE=4 SV=1
A0A139WGI6_TRICA	Eukarya	A	1.70E-17	74.9	0.1	Uncharacterized protein OS=Ornithorhynchus anatinus GN=TDRD5 PE=4 SV=1
F6YH90_ORNAN	Eukarya	V	1.80E-17	74.8	0	Uncharacterized protein OS=Anopheles darlingi GN=AND_005442 PE=4 SV=1
W5JJ85_ANODA	Eukarya	A	2.10E-17	74.6	0	AGAP003545-PA-like protein OS=Anopheles sinensis GN=ZHAS_00021239 PE=4 SV=1
A0A084WRU4_ANOSI	Eukarya	A	2.40E-17	74.4	0.2	Uncharacterized protein OS=Ciona savignyi GN=Csa.10307 PE=4 SV=1
H2YLA8_CIOSA	Eukarya	T	2.60E-17	74.3	0.1	Uncharacterized protein OS=Anolis carolinensis GN=TDRD5
G1KVT0_ANOCA	Eukarya	V	2.70E-17	74.2	0.1	PE=4 SV=1 Uncharacterized protein OS=Loxodonta africana GN=TDRD5
G3TEV7_LOXAF	Eukarya	V	4.50E-17	73.5	0	PE=4 SV=1
A0A0K8TEH4_LYGHE	Eukarya	A	6.40E-17	73	0	Uncharacterized protein OS=Lygus hesperus PE=4 SV=1

A0A067RPA3_ZOONE	Eukarya	A	7.10E-17	72.9	0.1	Tudor domain-containing protein 7 OS=Zootermopsis nevadensis GN=L798_01728 PE=4 SV=1
G3VEY7_SARHA	Eukarya	V	1.30E-16	72.1	0	Uncharacterized protein OS=Sarcophilus harrisii GN=TDRD5 PE=4 SV=1
L8I7L7_9CETA	Eukarya	V	1.40E-16	72	0.1	Tudor domain-containing protein 5 (Fragment) OS=Bos mutus GN=M91_03486 PE=4 SV=1
W5Q779_SHEEP	Eukarya	V	1.40E-16	71.9	0.5	Uncharacterized protein OS=Ovis aries GN=TDRD5 PE=4 SV=1
G5E528_BOVIN	Eukarya	V	2.30E-16	71.3	0	Tudor domain-containing protein 5 OS=Bos taurus GN=TDRD5 PE=4 SV=1
H2MII4_ORYLA	Eukarya	V	2.90E-16	71	0	Uncharacterized protein OS=Oryzias latipes GN=TDRD7 (1 to many) PE=4 SV=1
N6TQX5_DENPD	Eukarya	A	3.20E-16	70.8	0.1	Uncharacterized protein (Fragment) OS=Dendroctonus ponderosae GN=YQE_11709 PE=4 SV=1
F6WY93_HORSE	Eukarya	V	5.80E-16	70	0	Uncharacterized protein OS=Equus caballus GN=TDRD5 PE=4 SV=1
A0A0J9YJ00_DANRE	Eukarya	V	5.90E-16	69.9	0	Tudor domain-containing protein 7A (Fragment) OS=Danio rerio GN=tdrd7a PE=1 SV=2
U5EFJ8_9DIPT	Eukarya	A	9.50E-16	69.3	0.9	Putative oskar (Fragment) OS=Corethrella appendiculata PE=2 SV=1
H2USX7_TAKRU	Eukarya	V	9.60E-16	69.3	0	Uncharacterized protein OS=Takifugu rubripes PE=4 SV=1
F1S6A1_PIG	Eukarya	V	1.10E-15	69.1	0	Uncharacterized protein OS=Sus scrofa GN=TDRD5 PE=4 SV=2
T1DTM7_ANOAQ	Eukarya	A	1.10E-15	69	0	Uncharacterized protein (Fragment) OS=Anopheles aquasalis PE=2 SV=1
A0A060W2X9_ONCMY	Eukarya	V	2.20E-15	68.1	0	Uncharacterized protein OS=Oncorhynchus mykiss GN=GSONMT00078733001 PE=4 SV=1
S7NG41_MYOBR	Eukarya	V	2.60E-15	67.9	0.1	Tudor domain-containing protein 5 OS=Myotis brandtii GN=D623_10022817 PE=4 SV=1
F7B4W0_MONDO	Eukarya	V	2.70E-15	67.8	0	Uncharacterized protein OS=Monodelphis domestica GN=TDRD5 PE=4 SV=2
V9KH94_CALMI	Eukarya	V	3.50E-15	67.5	0	Tudor domain-containing protein 5 OS=Callorhinchus milii PE=2 SV=1
A0A0P7YHR6_9TELE	Eukarya	V	4.70E-15	67	0	Uncharacterized protein OS=Scleropages formosus GN=Z043_114704 PE=4 SV=1
G1PFT9_MYOLU	Eukarya	V	5.70E-15	66.8	0.1	Uncharacterized protein OS=Myotis lucifugus GN=TDRD5 PE=4 SV=1
H3DL34_TETNG	Eukarya	V	7.70E-15	66.4	0	Uncharacterized protein OS=Tetraodon nigroviridis PE=4 SV=1
A0A096NXU4_PAPAN	Eukarya	V	8.70E-15	66.2	0	Uncharacterized protein OS=Papio anubis GN=TDRD5 PE=4 SV=1
F7GPY1_CALJA	Eukarya	V	1.10E-14	65.9	0	Uncharacterized protein OS=Callithrix jacchus GN=TDRD5 PE=4 SV=1

W5N030_LEPOC	Eukarya	V	1.20E-14	65.8	0.1	Uncharacterized protein OS=Lepisosteus oculatus PE=4 SV=1 Uncharacterized protein OS=Felis catus GN=TDRD5 PE=4
M3VXB3_FELCA	Eukarya	V	1.30E-14	65.6	0	SV=1 Tudor domain-containing protein 7 OS=Tinamus guttatus
A0A099ZTT8_TINGU	Eukarya	V	1.40E-14	65.6	0	GN=N309_12928 PE=4 SV=1 Uncharacterized protein OS=Gorilla gorilla gorilla GN=TDRD5
G3R6R4_GORGO	Eukarya	V	1.50E-14	65.5	0	PE=4 SV=1 Tudor domain-containing protein 5 OS=Tupaia chinensis
L9L889_TUPCH	Eukarya	V	1.60E-14	65.4	0.3	GN=TREES_T100015801 PE=4 SV=1 Tudor domain-containing protein 5 (Fragment) OS=Lasius niger
A0A0J7KVQ7_LASNI	Eukarya	A	1.70E-14	65.3	0	GN=RF55_5458 PE=4 SV=1 Putative uncharacterized protein OS=Pediculus humanus subsp.
E0VJL4_PEDHC	Eukarya	A	1.70E-14	65.3	0.9	corporis GN=Phum_PHUM247930 PE=4 SV=1 Uncharacterized protein OS=Chlorocebus sabaeus GN=TDRD5
A0A0D9RID1_CHLSB	Eukarya	V	1.70E-14	65.2	0	PE=4 SV=1 Tudor domain-containing protein 7 OS=Harpegnathos saltator
E2BFZ8_HARSA	Eukarya	A	2.00E-14	65	0.9	GN=EAI_14615 PE=4 SV=1 Uncharacterized protein (Fragment) OS=Oryctes borbonicus
A0A0T6AU98_9SCAR	Eukarya	A	2.10E-14	65	3.9	GN=AMK59_7658 PE=4 SV=1 Tudor domain-containing protein 7 OS=Gallus gallus
R9PXP1_CHICK	Eukarya	V	2.30E-14	64.8	0	GN=TDRD7 PE=4 SV=1 Uncharacterized protein OS=Macaca mulatta GN=TDRD5 PE=4
F7CN93_MACMU	Eukarya	V	2.70E-14	64.6	0	SV=1 Tudor domain-containing protein 5 OS=Canis lupus familiaris
A0A0A0MPC8_CANLF	Eukarya	V	2.70E-14	64.6	0	GN=TDRD5 PE=4 SV=1 Uncharacterized protein OS=Mustela putorius furo GN=TDRD5
M3Y1J3_MUSPF	Eukarya	V	2.90E-14	64.5	0	PE=4 SV=1 Uncharacterized protein OS=Pongo abelii GN=TDRD5 PE=4
H2N4J0_PONAB	Eukarya	V	3.00E-14	64.5	0	SV=1 Uncharacterized protein OS=Pan troglodytes GN=TDRD5 PE=4
H2Q0P6_PANTR	Eukarya	V	3.20E-14	64.4	0	SV=1 Uncharacterized protein OS=Astyanax mexicanus PE=4 SV=1
W5LEX2_ASTMX	Eukarya	V	3.50E-14	64.3	0	Tudor domain containing 5, isoform CRA_b OS=Homo sapiens
A0A024R910_HUMAN	Eukarya	V	4.00E-14	64.1	0	GN=TDRD5 PE=4 SV=1 Tudor domain-containing protein 5 isoform 2
A0A0P6JFX9_HETGA	Eukarya	V	4.10E-14	64.1	0	OS=Heterocephalus glaber GN=TDRD5 PE=4 SV=1 Uncharacterized protein OS=Otolemur garnettii GN=TDRD5
H0WXJ6_OTOGA	Eukarya	V	4.70E-14	63.8	0	PE=4 SV=1 Uncharacterized protein OS=Cavia porcellus GN=TDRD5 PE=4
H0V001_CAVPO	Eukarya	V	5.20E-14	63.7	0	SV=1 Putative uncharacterized protein OS=Macaca fascicularis
G7NU06_MACFA	Eukarya	V	5.80E-14	63.6	0	GN=EGM_01633 PE=4 SV=1

G1S4T3_NOMLE	Eukarya	V	6.40E-14	63.4	0	Uncharacterized protein OS=Nomascus leucogenys GN=TDRD7 PE=4 SV=1
G1M861_AILME	Eukarya	V	6.80E-14	63.3	0	Tudor domain-containing protein 5 OS=Ailuropoda melanoleuca GN=TDRD5 PE=4 SV=1
A0A061HYN9_CRIGR	Eukarya	V	7.10E-14	63.3	0	Tudor domain-containing protein 5 OS=Cricetulus griseus GN=H671_5g14992 PE=4 SV=1
A0A0F8AWR5_LARCR	Eukarya	V	1.00E-13	62.7	0	Tudor domain-containing protein 7A OS=Larimichthys crocea GN=EH28_10800 PE=4 SV=1
V5GPP4_ANOGL	Eukarya	A	1.10E-13	62.6	1.4	Tudor domain-containing protein (Fragment) OS=Anoplophora glabripennis GN=TDRD7 PE=4 SV=1
A0A087XQP1_POEFO	Eukarya	V	1.10E-13	62.6	0	Uncharacterized protein OS=Poecilia formosa PE=4 SV=2 Tudor domain containing 7 (Fragment) OS=Pararge aegeria
S4NW46_9NEOP	Eukarya	A	1.20E-13	62.5	0	PE=4 SV=1 Tudor domain-containing protein 7 OS=Pteropus alecto
L5KMF4_PTEAL	Eukarya	V	2.10E-13	61.8	0	GN=PAL_GLEAN10008027 PE=4 SV=1 Uncharacterized protein OS=Ictidomys tridecemlineatus
I3NCP3_ICTTR	Eukarya	V	2.40E-13	61.6	0	GN=TDRD7 PE=4 SV=1 Uncharacterized protein OS=Oryctolagus cuniculus GN=TDRD7
G1SCH8_RABIT	Eukarya	V	2.50E-13	61.5	0	PE=4 SV=1 Putative uncharacterized protein (Fragment) OS=Mus musculus
Q3TTK4_MOUSE	Eukarya	V	2.50E-13	61.5	0	GN=Tdrd7 PE=2 SV=1 Tudor domain-containing protein 5 OS=Rattus norvegicus
A0A0H2UHC6_RAT	Eukarya	V	3.90E-13	60.9	0	GN=Tdrd5 PE=4 SV=1 Tudor domain-containing protein 7 OS=Neovison vison
U6CRG5_NEovi	Eukarya	V	3.90E-13	60.9	0	GN=TDRD7 PE=2 SV=1 Uncharacterized protein OS=Pelodiscus sinensis GN=TDRD5
K7FVR3_PELSI	Eukarya	V	5.10E-13	60.5	0	PE=4 SV=1 NYN domain protein OS=Peptostreptococcaceae bacterium
J4K9P6_9FIRM	Bacteria	F	2.90E-07	42.1	31.8	AS15 GN=HMPREF1142_1162 PE=4 SV=1 Uncharacterized protein OS=[Eubacterium] yurii subsp.
E0QL58_9FIRM	Bacteria	F	0.00014	33.5	29.1	margaretiae ATCC 43715 GN=HMPREF0379_1756 PE=4 SV=1 Uncharacterized protein OS=candidate division TM6 bacterium
A0A0G0IVI6_9BACT	Bacteria	?	0.0025	29.5	1.9	GW2011_GWA2_36_9 GN=US32_C0002G0021 PE=4 SV=1

71 **Supplementary File 1D: List of sequences and their BLAST results used for phylogenetic analysis of the OSK domain.**
 72 The sequences were obtained by searching the TrEMBL database using hmmsearch and the final HMM generated for OSK (Supplementary
 73 files: HMM>OSK.hmm). Reported parameters are as described for Supplementary Table S3. To obtain sequences for each entry, either
 74 search UniProt directly (<https://www.uniprot.org/>) or consult the final alignment in Supplementary Files: Alignments>OSK_TREE.fasta.
 75 Phylum Abbreviations: A = Arthropoda; Ar = Archaea; As = Ascomycota; B = Bacteroidetes; C = Cyanobacteria; Eu = Euryarchaeota; F =
 76 Firmicutes; Fu = Fungi; P = Proteobacteria
 77

Accession Number	Domain	Phylum	E-value	Score	Bias	Description of target
A1Y1T7_DROIM	Eukarya	A	2.90E-34	128.4	0.6	Oskar OS=Drosophila immigrans GN=osk PE=4 SV=1
F2WJY6_9HYME	Eukarya	A	3.50E-34	128.2	0.9	Oskar (Fragment) OS=Messor pergandei PE=2 SV=1
B4JTJ1_DROGR	Eukarya	A	5.00E-34	127.6	0.1	GH23955 OS=Drosophila grimshawi GN=Dgrl\GH23955 PE=4 SV=1 Maternal effect protein oskar OS=Acromyrmex echinatior
F4WQN7_ACREC	Eukarya	A	8.10E-34	127	0.9	GN=G5I_08127 PE=4 SV=1 Putative uncharacterized protein OS=Camponotus floridanus
E2A7I8_CAMFO	Eukarya	A	2.80E-33	125.2	0.4	GN=EAG_03874 PE=4 SV=1 Maternal effect protein oskar OS=Ceratitis capitata GN=OSKA PE=2
W8CE30_CERCA	Eukarya	A	5.80E-33	124.2	0.3	SV=1 Maternal effect protein oskar OS=Bactrocera cucurbitae GN=osk
A0A0A1XRQ4_BACCU	Eukarya	A	7.50E-33	123.9	0.4	PE=4 SV=1 Uncharacterized protein OS=Drosophila willistoni GN=Dwil\GK11117
B4N815_DROWI	Eukarya	A	8.10E-33	123.8	0.3	PE=4 SV=2 Maternal effect protein oskar OS=Cerapachys biroi GN=X777_01612
A0A026WMY1_CERBI	Eukarya	A	2.40E-32	122.3	0.7	PE=4 SV=1 Putative uncharacterized protein (Fragment) OS=Solenopsis invicta
E9IZ46_SOLIN	Eukarya	A	2.50E-32	122.2	0.1	GN=SINV_01516 PE=4 SV=1 Maternal effect protein oskar OS=Bactrocera dorsalis GN=OSKA
A0A034WRF5_BACDO	Eukarya	A	2.70E-32	122.1	0.5	PE=4 SV=1 Maternal effect protein oskar OS=Bactrocera latifrons GN=osk_2
A0A0K8U7J3_BACLA	Eukarya	A	5.70E-32	121	0.5	PE=4 SV=1 Maternal effect protein oskar OS=Aedes aegypti PE=4 SV=1
Q2PP79_AEDAE	Eukarya	A	6.30E-32	120.9	0.1	Uncharacterized protein OS=Drosophila mojavensis
B4K9E5_DROMO	Eukarya	A	8.60E-32	120.5	0.1	GN=Dmoj\GI10055 PE=4 SV=2 Oskar OS=Drosophila virilis GN=osk PE=4 SV=1
B4LXK5_DROVI	Eukarya	A	1.20E-31	120	0.1	Oskar OS=Drosophila busckii GN=DBus_chr3Rg607 PE=4 SV=1
A0A0M4F3M8_DROBS	Eukarya	A	1.80E-31	119.4	0.4	Uncharacterized protein OS=Lucilia cuprina GN=FF38_12727 PE=4
A0A0L0CP24_LUCCU	Eukarya	A	2.90E-31	118.8	1.5	SV=1 Uncharacterized protein OS=Drosophila yakuba GN=Dyak\GE25914
B4PTX6_DROYA	Eukarya	A	1.00E-30	117	0.5	PE=4 SV=1

B3P1W4_DROER	Eukarya	A	2.70E-30	115.6	0.5	GG13545 OS=Drosophila erecta GN=Dere\GG13545 PE=4 SV=1
T1PG45_MUSDO	Eukarya	A	4.20E-30	115.1	0.7	GDSL-like Lipase/Acylhydrolase OS=Musca domestica PE=2 SV=1
B4HKZ1_DROSE	Eukarya	A	5.20E-30	114.8	0.2	GM23770 OS=Drosophila sechellia GN=Dsec\GM23770 PE=4 SV=1
Q295Q4_DROPS	Eukarya	A	6.40E-30	114.5	0.2	Uncharacterized protein, isoform A OS=Drosophila pseudoobscura pseudoobscura GN=Dpse\GA10627 PE=4 SV=2
E8NH25_DROME	Eukarya	A	7.30E-30	114.3	0.2	RE24380p (Fragment) OS=Drosophila melanogaster GN=osk-RA PE=2 SV=1
T1DTM7_ANOAQ	Eukarya	A	1.00E-29	113.8	0	Uncharacterized protein (Fragment) OS=Anopheles aquasalis PE=2 SV=1
B3LZ06_DROAN	Eukarya	A	2.20E-28	109.5	0.1	Uncharacterized protein OS=Drosophila ananassae GN=Dana\GF17692 PE=4 SV=1
E1A883_NASVI	Eukarya	A	4.20E-28	108.6	0.2	Oskar OS=Nasonia vitripennis PE=2 SV=1 AGAP003545-PA-like protein OS=Anopheles sinensis GN=ZHAS_00021239 PE=4 SV=1
A0A084WRU4_ANOSI	Eukarya	A	1.30E-27	107.1	0.1	Uncharacterized protein OS=Anopheles darlingi GN=AND_005442 PE=4 SV=1
W5JJ85_ANODA	Eukarya	A	4.50E-27	105.3	0	Putative oskar (Fragment) OS=Corethrella appendiculata PE=2 SV=1
Q7PQJ3_ANOGA	Eukarya	A	6.00E-27	104.9	0	AGAP003545-PA OS=Anopheles gambiae GN=osk PE=4 SV=3 Oskar OS=Culex quinquefasciatus GN=CpipJ_CPIJ007471 PE=2 SV=1
B0WIV7_CULQU	Eukarya	A	1.30E-26	103.9	0.4	Oskar, isoform D OS=Drosophila melanogaster GN=osk PE=4 SV=1 GA10627 (Fragment) OS=Drosophila pseudoobscura GN=GA10627 PE=4 SV=1
U5EFJ8_9DIPT	Eukarya	A	4.00E-25	99.1	0	Oskar OS=Gryllus bimaculatus PE=2 SV=1
A0A126GUR4_DROME	Eukarya	A	3.10E-24	96.2	0.3	Oskar protein OS=Fopius arisanus GN=osk PE=4 SV=1 GA10627 (Fragment) OS=Drosophila pseudoobscura GN=GA10627 PE=4 SV=1
A0A059PGF2_9MUSC	Eukarya	A	3.10E-24	96.2	0.5	Maternal effect protein oskar (Fragment) OS=Lasius niger GN=RF55_10783 PE=4 SV=1
K4MTL4_GRYBI	Eukarya	A	1.20E-23	94.2	0.2	Putative uncharacterized protein OS=Harpegnathos saltator GN=EAI_08923 PE=4 SV=1
A0A0C9QHR7_9HYME	Eukarya	A	5.40E-22	89	0	Lysophospholipase OS=Brevibacillus brevis GN=AB432_04505 PE=4 SV=1
A0A059PF64_9MUSC	Eukarya	A	1.90E-19	80.8	0.7	Lysophospholipase OS=Brevibacillus formosus GN=AA984_15375 PE=4 SV=1
A0A0J7KH44_LASNI	Eukarya	A	1.40E-13	61.9	0	Putative uncharacterized protein (Fragment) OS=Solenopsis invicta GN=SINV_16199 PE=4 SV=1
E2BYH0_HARSA	Eukarya	A	4.00E-12	57.3	0	Lysophospholipase L1-like esterase OS=Brevibacillus sp. BC25 GN=PMI05_03395 PE=4 SV=1
A0A0J6BBM7_BREBE	Bacteria	F	3.10E-10	51.2	0	
A0A0H0SJ00_9BACL	Bacteria	F	9.60E-10	49.6	0	
E9I8K8_SOLIN	Bacteria	A	5.90E-09	47.1	1	
J3A568_9BACL	Bacteria	F	5.80E-08	43.9	0	

G2HS43_9PROT	Bacteria	P	1.10E-07	43	1	Lipolytic protein OS=Arcobacter sp. L GN=ABLL_2651 PE=4 SV=1 Lipolytic protein OS=Arcobacter butzleri JV22
E6L4E3_9PROT	Bacteria	P	1.70E-07	42.4	0.7	GN=HMPREF9401_1319 PE=4 SV=1 GDSL-like protein OS=Eubacterium sp. CAG:786 GN=BN782_00012
R5GT16_9FIRM	Bacteria	F	2.60E-07	41.8	0	PE=4 SV=1 Uncharacterized protein OS=[Clostridium] cellulosi
A0A078KJ49_9FIRM	Bacteria	F	3.40E-07	41.4	0	GN=CCDG5_0508 PE=4 SV=1 Lipolytic enzyme, GDSL domain OS=Arcobacter butzleri (strain
A8EWS4_ARCB4	Bacteria	P	3.80E-07	41.3	0.7	RM4018) GN=Abu_2183 PE=4 SV=1 Lipolytic enzyme, GDSL domain protein OS=Arcobacter butzleri 7h1h
S5PEQ8_9PROT	Bacteria	P	4.00E-07	41.2	0.4	GN=A7H1H_2115 PE=4 SV=1 Lipolytic protein OS=Arcobacter butzleri L348 GN=AA20_04280
A0A0G9K3L5_9PROT	Bacteria	P	4.00E-07	41.2	0.6	PE=4 SV=1 GDSL-like protein OS=Bacteroides sp. CAG:770 GN=BN777_00744
R6T7B3_9BACE	Bacteria	B	4.80E-07	41	0	PE=4 SV=1 Acylneuraminate cytidyltransferase OS=Streptococcus uberis
A0A0A6S1U2_STRUB	Bacteria	F	4.90E-07	40.9	0.2	GN=NC01_08240 PE=4 SV=1 Uncharacterized protein OS=Planomicrobium glaciei CHR43
W3AC50_9BACL	Bacteria	F	5.50E-07	40.8	0.3	GN=G159_16940 PE=4 SV=1 Lipolytic protein OS=Arcobacter butzleri ED-1 GN=ABED_1978 PE=4
A0A0M1UPT0_9PROT	Bacteria	P	5.50E-07	40.8	0.4	SV=1 Platelet activating factor acetylhydrolase-likeprotein OS=Clostridium
U6EVC2_CLOTA	Bacteria	F	6.30E-07	40.6	0.7	tetani 12124569 GN=BN906_00617 PE=4 SV=1 Multifunctional acyl-CoA thioesterase I and protease I and
A0A098EIL2_9BACL	Bacteria	F	7.90E-07	40.3	0.1	lysophospholipase L1 OS=Planomicrobium sp. ES2 GN=BN1080_01055 PE=4 SV=1
A0A069S7Q5_9PORP	Bacteria	B	1.10E-06	39.9	0	Uncharacterized protein OS=Parabacteroides distasonis str. 3776 Po2 i GN=M090_4091 PE=4 SV=1
Q897X6_CLOTE	Bacteria	F	1.20E-06	39.7	0.8	Platelet activating factor acetylhydrolase-like protein OS=Clostridium tetani (strain Massachusetts / E88) GN=CTC_00594 PE=4 SV=1
A0A0J9FZD4_9PORP	Bacteria	B	1.30E-06	39.6	0	Uncharacterized protein OS=Parabacteroides sp. D26 GN=HMPREF1000_00856 PE=4 SV=1
K9WJ28_9CYAN	Bacteria	C	1.60E-06	39.3	1.2	Lysophospholipase L1-like esterase OS=Microcoleus sp. PCC 7113 GN=Mic7113_4114 PE=4 SV=1
A0A0G1KN57_9BACT	Bacteria	?	3.70E-06	38.1	0	Secreted protein OS=candidate division WWE3 bacterium GW2011_GWC2_44_9 GN=UW82_C0006G0015 PE=4 SV=1
R5VMZ1_9FIRM	Bacteria	F	5.00E-06	37.7	0.1	GDSL-like protein OS=Firmicutes bacterium CAG:631 GN=BN742_01282 PE=4 SV=1
A0A0L0WAR6_CLOPU	Bacteria	F	7.20E-06	37.2	0.4	Lysophospholipase L1 OS=Clostridium purinilyticum GN=CLPU_6c00720 PE=4 SV=1
R5S0B3_9BACE	Bacteria	B	7.30E-06	37.2	0	GDSL-like protein OS=Bacteroides sp. CAG:545 GN=BN702_00435 PE=4 SV=1

A0A073IAZ3_9PORP	Bacteria	B	7.60E-06	37.1	0	Uncharacterized protein OS=Porphyromonas sp. 31_2 GN=HMPREF1002_01104 PE=4 SV=1
R4JC30_9BACT	Bacteria	?	7.90E-06	37.1	0.1	Uncharacterized protein OS=uncultured bacterium BAC25G1 GN=metaSSY_00600 PE=4 SV=1
A0A0C1UEU0_9CLOT	Bacteria	F	7.90E-06	37	0.8	GDSL-like Lipase/Acylhydrolase family protein OS=Clostridium argentinense CDC 2741 GN=U732_2423 PE=4 SV=1
R5LL12_9FIRM	Bacteria	F	1.00E-05	36.7	0	GDSL-like protein OS=Eubacterium sp. CAG:115 GN=BN470_02036 PE=4 SV=1
A0A0E3QN36_METBA	Archaea	Eu	1.20E-05	36.5	0.2	Putative tesA-like protease OS=Methanosarcina barkeri str. Wiesmoor GN=MSBRW_2234 PE=4 SV=1
B7KKAA6_CYAP7	Bacteria	C	1.60E-05	36.1	0.3	Lipolytic protein G-D-S-L family OS=Cyanothece sp. (strain PCC 7424) GN=PCC7424_2577 PE=4 SV=1
R2P1Z4_9ENTE	Bacteria	F	1.60E-05	36	0.1	Uncharacterized protein OS=Enterococcus raffinosus ATCC 49464 GN=UAK_02837 PE=4 SV=1
A0A072Y8N5_9CLOT	Bacteria	F	1.80E-05	35.9	1.5	Acetylhydrolase OS=Clostridium sp. K25 GN=Z957_08245 PE=4 SV=1
K8GFE2_9CYAN	Bacteria	C	2.00E-05	35.8	0.1	Lysophospholipase L1-like esterase OS=Oscillatioriales cyanobacterium JSC-12 GN=OsccyDRAFT_3941 PE=4 SV=1
A0A095ZDI3_9FIRM	Bacteria	F	2.00E-05	35.8	0.3	Uncharacterized protein OS=Tissierellia bacterium S7-1-4 GN=HMPREF1634_08565 PE=4 SV=1
E1YW67_9PORP	Bacteria	B	2.20E-05	35.6	0	GDSL-like protein OS=Parabacteroides sp. 20_3 GN=HMPREF9008_00759 PE=4 SV=1
A0A0M0WNM0_9BACI	Bacteria	F	2.60E-05	35.4	0.4	Lipase OS=Bacillus sp. FJAT-21351 GN=AMS61_13120 PE=4 SV=1
A0A0G3CFD7_METBA	Archaea	Eu	2.80E-05	35.3	0.1	GDSL family lipase/acylhydrolase OS=Methanosarcina barkeri CM1 GN=MCM1_0752 PE=4 SV=1
F6DQC0_DESRL	Bacteria	F	3.40E-05	35	0	Lipolytic protein G-D-S-L family OS=Desulfotomaculum ruminis (strain ATCC 23193 / DSM 2154 / NCIB 8452 / DL) GN=Desru_1430 PE=4 SV=1
D5DGY9_BACMD	Bacteria	F	3.50E-05	35	0.5	Lipase/Acylhydrolase (GDSL) OS=Bacillus megaterium (strain DSM 319) GN=BMD_3140 PE=4 SV=1
A5N8N5_CLOK5	Bacteria	F	3.70E-05	34.9	1.6	Uncharacterized protein OS=Clostridium kluyveri (strain ATCC 8527 / DSM 555 / NCIMB 10680) GN=CKL_1624 PE=4 SV=1
R7ADB4_9BACE	Bacteria	F	3.90E-05	34.8	0.2	Uncharacterized protein OS=Bacteroides pectinophilus CAG:437 GN=BN656_00903 PE=4 SV=1
A0A0H1UPZ9_STRAG	Bacteria	F	4.30E-05	34.7	0.2	Acylneuraminate cytidyltransferase OS=Streptococcus agalactiae GN=WA03_09270 PE=4 SV=1
A0A094AE00_9PEZI	Eukarya	As	0.00036	31.7	0.1	Uncharacterized protein (Fragment) OS=Pseudogymnoascus sp. VKM F-4281 (FW-2241) GN=V493_03380 PE=4 SV=1
A0A0L1HYX4_9PLEO	Eukarya	As	0.00066	30.9	0	Carbohydrate esterase family 3 protein OS=Stemphylium lycopersici GN=TW65_91054 PE=4 SV=1

G2QGB0_MYCTT	Eukarya	As	0.00074	30.7	0.1	Carbohydrate esterase family 3 protein OS=Myceliophthora thermophila (strain ATCC 42464 / BCRC 31852 / DSM 1799) GN=MYCTH_53698 PE=4 SV=1
E3RJZ5_PYRTT	Eukarya	As	0.0014	29.8	0	Putative uncharacterized protein OS=Pyrenophora teres f. teres (strain 0-1) GN=PTT_08513 PE=4 SV=1
G2QVW9_THITE	Eukarya	As	0.0066	27.7	0.1	Carbohydrate esterase family 3 protein OS=Thielavia terrestris (strain ATCC 38088 / NRRL 8126) GN=THITE_2042744 PE=4 SV=1
G0S9F4_CHATD	Eukarya	As	0.01	27.1	0	Putative uncharacterized protein OS=Chaetomium thermophilum (strain DSM 1495 / CBS 144.50 / IMI 039719) GN=CTHT_0045680 PE=4 SV=1

79 **Supplementary File 1E: List of genomes analyzed for codon use.**

80 This table lists the 17 genomes that were downloaded and analyzed for codon use as described in
 81 “Selection of sequences for codon use analysis” in Methods. All genomes can be downloaded from
 82 <https://www.ncbi.nlm.nih.gov/genome/browse#!/overview/>. The table lists the species name (*Species*),
 83 family (*Family*) and Order (*Order*), NCBI genome accession number (*Genome ID*), and the *oskar*
 84 NCBI Nucleotide accession number (*oskar Nucleotide ID*).
 85

Species	Family	Order	Genome ID	Oskar Nucleotide ID†
<i>Drosophila melanogaster</i>	Drosophilidae	Diptera	GCA_001014345.1	NM_169248.4
<i>Nasonia vitripennis</i>	Pteromalidae	Hymenoptera	GCA_000002325.2	HM535628.1
<i>Culex quinquefasciatus</i>	Culicidae	Diptera	GCA_000209185.1	EU517695.1
<i>Drosophila virilis</i>	Drosophilidae	Diptera	GCA_000005245.1	L22556.1
<i>Ceratitis capitata</i>	Tephritidae	Diptera	GCA_000347755.4	LOC101450245
<i>Musca domestica</i>	Muscidae	Diptera	GCA_000371365.1	LOC101890691
<i>Acromyrmex echinatior</i>	Formicidae	Hymenoptera	GCA_000204515.1	LOC105147973
<i>Harpegnathos saltator</i>	Formicidae	Hymenoptera	GCA_000147195.1	LOC105187957
<i>Bactrocera dorsalis</i>	Tephritidae	Diptera	GCA_000789215.2	LOC105232054
<i>Fopius arisanus</i>	Braconidae	Hymenoptera	GCA_000806365.1	LOC105267990
<i>Athalia rosae</i>	Tenthredinidae	Hymenoptera	GCA_000344095.2	LOC105692731
<i>Orussus abietinus</i>	Orussidae	Hymenoptera	GCA_000612105.2	LOC105696794
<i>Stomoxys calcitrans</i>	Muscidae	Diptera	GCA_001015335.1	LOC106086381
<i>Bactrocera oleae</i>	Tephritidae	Diptera	GCA_001188975.2	LOC106622417
<i>Copidosoma floridanum</i>	Encyrtidae	Hymenoptera	GCA_000648655.2	LOC106642594
<i>Polistes canadensis</i>	Vespidae	Hymenoptera	GCA_001313835.1	LOC106790143
<i>Neodiprion lecontei</i>	Diprionidae	Hymenoptera	GCF_001263575.1	LOC107223453

86