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

Authors: *Leo Blondel*[1], *Tamsin E. M. Jones*[2,3] *and Cassandra G. Extavour*[1,2*]

Affiliations:

1. Department of Molecular and Cellular Biology, Harvard University, 16 Divinity Avenue, Cambridge MA, USA
2. Department of Organismic and Evolutionary Biology, Harvard University, 16 Divinity Avenue, Cambridge MA, USA
3. Current address: European Bioinformatics Institute, EMBL-EBI, Wellcome Genome Campus, Hinxton, Cambridgeshire, UK

* Correspondence to extavour@oeb.harvard.edu

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 can facilitate evolution of novel developmental mechanisms.

One Sentence Summary: Our research shows that gene origin processes often considered highly unusual, including HGT and de novo coding region evolution, can give rise to novel genes that can both participate in pre-existing gene regulatory networks, and also facilitate the evolution of novel developmental mechanisms.

29 **Main Text:** Heritable variation is the raw material of evolutionary change. Genetic variation can
30 arise from mutation and gene duplication of existing genes (*1*), or through *de novo* processes (*2*),
31 but the extent to which such novel, or "orphan" genes participate significantly in the
32 evolutionary process is unclear. Mutation of existing cis-regulatory (*3*) or protein coding regions
33 (*4*) can drive evolutionary change in developmental processes. However, recent studies in
34 animals and fungi suggest that novel genes can also drive phenotypic change (*5*). Although
35 counterintuitive, novel genes may be integrating continuously into otherwise conserved gene
36 networks, with a higher rate of partner acquisition than subtler variations on preexisting genes
37 (*6*). Moreover, in humans and fruit flies, a large proportion of novel genes are expressed in the
38 brain, suggesting their participation in the evolution of major organ systems (*7, 8*). However,
39 while next generation sequencing has improved their discovery, the developmental and
40 evolutionary significance of novel genes remains understudied.
41
42 The mechanism of formation of a novel gene may have implications for its function. Novel genes
43 that arise by duplication, thus possessing the same biophysical properties as their parent genes,
44 have innate potential to participate in preexisting cellular and molecular mechanisms (*1*).
45 However, orphan genes lacking sequence similarity to existing genes must form novel functional
46 molecular relationships with extant genes, in order to persist in the genome. When such genes
47 arise by introduction of foreign DNA into a host genome through horizontal gene transfer
48 (HGT), they may introduce novel, already functional sequence information into a genome.
49 Whether genes created by HGT show a greater propensity to contribute to or enable novel
50 processes is unclear. Endosymbionts in the host germ line cytoplasm (germ line symbionts) could
51 increase the occurrence of evolutionarily relevant HGT events, as foreign DNA integrated into

52 the germ line genome is transferred to the next generation. HGT from bacterial endosymbionts
53 into insect genomes appears widespread, involving transfer of metabolic genes or even larger
54 genomic fragments to the host genome (see for example *9, 10-12*).
55
56 Here we examined the evolutionary origins of the *oskar* (*osk*) gene, long considered a novel gene
57 that evolved to be indispensable for insect reproduction (*13*). First discovered in *Drosophila*
58 *melanogaster* (*14*), *osk* is necessary and sufficient for assembly of germ plasm, a cytoplasmic
59 determinant that specifies the germ line in the embryo. Germ plasm-based germ line
60 specification appears derived within insects, confined to insects that undergo metamorphosis
61 (Holometabola) (*15, 16*). Initially thought exclusive to Diptera (flies and mosquitoes), its
62 discovery in a wasp, another holometabolous insect with germ plasm (*17*), led to the hypothesis
63 that *oskar* originated as a novel gene at the base of the Holometabola approximately 300 Mya,
64 facilitating the evolution of insect germ plasm as a novel developmental mechanism (*17*).
65 However, its subsequent discovery in a cricket (*15*), a hemimetabolous insect without germ
66 plasm (*18*), implied that *osk* was instead at least 50 My older, and that its germ plasm role was
67 derived rather than ancestral (*19*). Despite its orphan gene status, *osk* plays major developmental
68 roles, interacting with the products of many genes highly conserved across animals (*13, 20, 21*).
69 *osk* thus represents an example of a novel gene that not only functions within pre-existing gene
70 networks in the nervous system (*15*), but has also evolved into the only animal gene known to be
71 both necessary and sufficient for germ line specification (*22, 23*).
72
73 The evolutionary origins of this remarkable gene are unknown. Osk contains two biophysically
74 conserved domains, an N-terminal LOTUS domain and a C-terminal hydrolase-like domain

75 called OSK (*20, 24*) (Figure 1a). An initial BLASTp search using the full-length *D.*
76 *melanogaster osk* sequence as a query yielded either other holometabolous insect *osk* genes, or
77 partial hits for the LOTUS or OSK domains (E-value < 0.01; Supplementary files: BLAST
78 search results). This suggested that full length *osk* was unlikely to be a duplication of any other
79 known gene. This prompted us to perform two more BLASTp searches, one using each of the
80 two conserved Osk protein domains individually as query sequences. Strikingly, in this BLASTp
81 search, although we recovered several eukaryotic hits for the LOTUS domain, we recovered no
82 eukaryotic sequences that resembled the OSK domain, even with very low E-value stringency
83 (E-value < 10; see Methods section "*BLAST searches of oskar"* for an explanation of E-value
84 threshold choices; Supplementary files: BLAST search results).
85
86 To understand this anomaly, we built an alignment of 95 Oskar sequences (Supplementary files:
87 Alignments>OSKAR_MUSCLE_FINAL.fasta; Supplementary Table S2) and used a custom
88 iterative HMMER sliding window search tool to compare each domain with protein sequences
89 from all domains of life. Sequences most similar to the LOTUS domain were almost exclusively
90 eukaryotic sequences (Supplementary Table 3). In contrast, those most similar to the OSK
91 domain were bacterial, specifically sequences similar to SGNH-like hydrolases (*20, 24*) (Pfam
92 Clan: SGNH_hydrolase - CL0264; Supp. Table 4; Figure 1b). To visualize their relationships, we
93 graphed the sequence similarity network for the sequences of these domains and their closest
94 hits. We observed that the majority of LOTUS domain sequences clustered within eukaryotic
95 sequences (Figure 1c). In contrast, OSK domain sequences formed an isolated cluster, a small
96 subset of which formed a connection to bacterial sequences (Figure 1d). These data are
97 consistent with a previous suggestion, based on BLAST results (*17*), that HGT from a bacterium

98 into an ancestral insect genome may have contributed to the evolution of *osk*. However, this
99 possibility was not adequately addressed by previous analyses, which were based on alignments
100 of full length Osk containing only eukaryotic sequences as outgroups (*15*). To rigorously test this
101 hypothesis, we therefore performed phylogenetic analyses of the two domains independently. A
102 finding that LOTUS sequences were nested within eukaryotes, while OSK sequences were
103 nested within bacteria, would provide support for the HGT hypothesis.
104
105 Both Maximum likelihood and Bayesian approaches confirmed this prediction (Figure 2a, Figure
106 2–figure supplements 1 and 2), and these results were robust to changes in the methods of
107 sequence alignment (Figure 2-figure supplements 6-10). As expected, LOTUS sequences from
108 Osk proteins were related to other eukaryotic LOTUS domains, to the exclusion of the only three
109 bacterial sequences that met our E-value cutoff for inclusion in the analyses (Figs. 2a, Figure 2–
110 figure supplements 1 and 2; see Methods and Supplemental Text). LOTUS sequences from non-
111 Oskar proteins were almost exclusively eukaryotic. (Supplementary Table 3); only three bacterial
112 sequences matched the LOTUS domain with an E-value < 0.01. Osk LOTUS domains clustered
113 into two distinct clades, one comprising all Dipteran sequences, and the other comprising all
114 other Osk LOTUS domains examined from both holometabolous and hemimetabolous orders
115 (Figure 2a). Dipteran Osk LOTUS sequences formed a monophyletic group that branched sister
116 to a clade of LOTUS domains from Tud5 family proteins of non-arthropod animals (NAA). NAA
117 LOTUS domains from Tud7 family members were polyphyletic, but most of them formed a
118 clade branching sister to (Osk LOTUS + NAA Tud5 LOTUS). Non-Dipteran Osk LOTUS
119 domains formed a monophyletic group that was related in a polytomy to the aforementioned

120 (NAA Tud7 LOTUS + (Dipteran Osk LOTUS + NAA Tud5 LOTUS)) clade, and to various
121 arthropod Tud7 family LOTUS domains.
122
123 The fact that Tud7 LOTUS domains are polyphyletic suggests that arthropod domains in this
124 family may have evolved differently than their homologues in other animals. The relationships of
125 Dipteran LOTUS sequences were consistent with the current hypothesis for interrelationships
126 between Dipteran species (*25*) Similarly, among the non-Dipteran Osk LOTUS sequences, the
127 hymenopteran sequences form a clade to the exclusion of the single hemimetabolous sequence
128 (from the cricket *Gryllus bimaculatus*), consistent with the monophyly of Hymenoptera (*26*). It is
129 unclear why Dipteran Osk LOTUS domains cluster separately from those of other insect Osk
130 proteins. We speculate that the evolution of the Long Oskar domain (*27, 28*), which appears to be
131 a novelty within Diptera (Supplementary Files: Alignments>OSKAR_MUSCLE_FINAL.fasta),
132 may have influenced the evolution of the Osk LOTUS domain in at least some of these insects.
133 Consistent with this hypothesis, of the 17 Dipteran *oskar* genes we examined, the seven *oskar*
134 genes possessing a Long Osk domain clustered into two clades based on the sequences of their
135 LOTUS domain. One of these clades comprised five Drosophila species (*D. willistoni*, *D.*
136 *mojavensis*, *D. virilis*, *D. grimshawi* and *D. immigrans*), and the second was composed of two
137 calyptrate flies from different superfamilies, *Musca domestica* (Muscoidea) and *Lucilia cuprina*
138 (Oestroidea).
139
140 In summary, the LOTUS domain of Osk proteins is most closely related to a number of other
141 LOTUS domains found in eukaryotic proteins, as would be expected for a gene of animal origin,

142 and the phylogenetic interrelationships of these sequences is largely consistent with the current
143 species or family level trees for the corresponding insects.
144
145 In contrast, OSK domain sequences were nested within bacterial sequences (Figure 2b, Figure 2–
146 figure supplements 3, 4). This bacterial, rather than eukaryotic, affinity of the OSK domain was
147 recovered even when different sequence alignment methods were used (Figure 2- figure
148 supplements 7-11). The only eukaryotic proteins emerging from the iterative HMMER search for
149 OSK domain sequences that had an E-value < 0.01 were all from fungi. All five of these
150 sequences were annotated as Carbohydrate Active Enzyme 3 (CAZ3), and all CAZ3 sequences
151 formed a clade that was sister to a clade of primarily Firmicutes. Most bacterial sequences used
152 in this analysis were annotated as lipases and hydrolases, with a high representation of GDSL-
153 like hydrolases (Supplementary Table S4). OSK sequences formed a monophyletic group but did
154 not branch sister to the other eukaryotic sequences in the analysis. Within this OSK clade, the
155 topology of sequence relationships was largely concordant with the species tree for insects (*29*),
156 as we recovered monophyletic Diptera to the exclusion of other insect species. However, the
157 single orthopteran OSK sequence (from the cricket *Gryllus bimaculatus*) grouped within the
158 Hymenoptera, rather than branching as sister to all other insect sequences in the tree, as would be
159 expected for this hemimetabolous sequence (*29*).
160
161 Importantly, OSK sequences did not simply form an outgroup to bacterial sequences. To formally
162 reject the possibility that the eukaryotic OSK clade has a sister group relationship to all bacterial
163 sequences in the analysis, we performed topology constraint analyses using the Swofford–Olsen–
164 Waddell–Hillis (SOWH) test, which assigns statistical support to alternative phylogenetic

165 topologies (*30*). We used the SOWHAT tool (*31*) to compare the HGT-supporting topology to
166 two alternative topologies with constraints more consistent with vertical inheritance. The first
167 was constrained by domain of life, disallowing paraphyletic relationships between sequences
168 from the same domain of life (Figure 2–figure supplement 5a). The second required monophyly
169 of Eukaryota, but allowed paraphyletic relationships between bacterial and archaeal sequences
170 (Figure 2–figure supplement 5b). We found that the topologies of both of these constrained trees
171 were significantly worse than the result we had recovered with our phylogenetic analysis (Figure
172 2–figure supplement 5), namely that the closest relatives of the OSK domain were bacterial
173 rather than eukaryotic sequences (Figure 2b, Figure 2–figure supplements 3 and 4).
174
175 OSK sequences formed a well-supported clade nested within bacterial GDSL-like lipase
176 sequences. The majority of these bacterial sequences were from the Firmicutes, a bacterial
177 phylum known to include insect germline symbionts (*32, 33*). All other sequences from classified
178 bacterial species, including a clade branching as sister to all other sequences, belonged either to
179 the Bacteroidetes or to the Proteobacteria. Members of both of these phyla are also known
180 germline symbionts of insects (*9, 34*) and other arthropods (*35*). In sum, the distinct phylogenetic
181 relationships of the two domains of Oskar are consistent with a bacterial origin for the OSK
182 domain. Further, the specific bacterial clades close to OSK suggest that an ancient arthropod
183 germ line endosymbiont could have been the source of a GDSL-like sequence that was

184 transferred into an ancestral insect genome, and ultimately gave rise to the OSK domain of
185 *oskar*.
186
187 We then asked if two additional sequence characteristics, GC3 content and codon use, were
188 consistent with distinct domain of life origins for the two Oskar domains (*36*). Under our
189 hypothesis, the HGT event that contributed to *oskar*'s formation would have occurred at least
190 480 Mya, in a common insect ancestor (*29*). This means that we would be highly unlikely to be
191 able to definitively identify any extant bacterial species from which the OSK sequence were
192 putatively derived, as has been possible, for example, with much more recent HGT events on the
193 order of a few million years ago (see for example *9*). However, we reasoned that if evolutionary
194 time had not completely erased the original GC3 content and codon use signatures from the
195 putative bacterially donated sequence (OSK), we might detect some differences in these
196 parameters both from the LOTUS domain, and from the host genome. Thus, we performed a
197 parametric analysis of these parameters for 17 well annotated insect genomes (Supplementary
198 Table 5).
199
200 To quantify the null hypothesis, we calculated an "Intra-Gene distribution" for all genes in the
201 genome, which assessed the correlation between codon use in the 5' and 3' halves of a given
202 gene. To determine whether the AT3/GC3 content correlation between the OSK and LOTUS
203 domains of *oskar* was different from that between the 5' and 3' ends of an average gene in the
204 genome, we then calculated the residuals of the Intra-Gene distribution and the LOTUS-OSK
205 domain pair. Pooling the residuals together revealed that the GC3 content was significantly

206 different between the LOTUS and OSK domains, compared to what would be expected within an
207 average gene in that genome (Figure 3a-c).
208
209 To ask if codon use was different between the OSK domain and other genes in the genome, we
210 used a modified "codon adaptation index" (CAI) approach (37). We calculated the frequency of
211 codon use for every gene in the available transcriptomes, and calculated a "global CAI" for all
212 genes in each of these species (rather than only for highly expressed genes, since we cannot be
213 confident which ones those are based on the available data for these species). We compared the
214 genome-wide CAI distribution with the CAI values of the full length oskar mRNA, the LOTUS
215 domain and the OSK domain, and found that the latter three CAI values fell well within the
216 ranges of the genome-wide CAI distribution, for all 17 genomes studied (Figure 3 Supplement
217 6A). Similarly, the differences between the LOTUS and OSK domain CAI values were not
218 significantly different from those between the 5' and 3' ends of each of the other genes in the
219 genome (Figure 3 Supplement 6B), and the OSK and LOTUS CAI values were similar to median
220 values for the 5' and 3' ends of other genes in the same genome (Figure 3 Supplement 6B). In
221 sum, while the overall codon use of the OSK and LOTUS domains of *oskar* appears similar to
222 each other and to the codon use of other genes in the same genome, the GC3 content is
223 significantly different between the two domains, and from that of an average gene in the genome.
224
225 While the results of our GC3 content analyses are consistent with the hypothesis of an HGT
226 origin for *oskar*, we cannot exclude the possibility that other mechanisms might impact this
227 codon use difference between two different regions of *oskar*. For example, it is possible that a
228 form of translational selection might be acting to promote differential codon use in distinct

229 regions of *oskar* (*38*), as a consequence of different folding requirements for different regions of
230 the protein (*39, 40*). Codon use may influence differential translation of ordered versus
231 disordered regions of the same protein (*41*), and Oskar contains regions of high predicted
232 disorder in addition to the LOTUS and OSK domains (*20, 42*). However, in the present study we
233 have examined only the latter two domains, and to our knowledge there is no evidence of
234 differential translational regulation of these two well-folded domains of Oskar.
235
236 An alternative parameter that could impact codon use across the *oskar* gene is a possible
237 requirement for specific structures of the *oskar* mRNA. Given that the *oskar* transcript plays a
238 translation-independent role in the *Drosophila* germ line (*43, 44*), it is conceivable that structural
239 constraints might be different on different regions of its mRNA. Moreover, specific stem-loop
240 structures within the oskar 3' UTR are required for its localization within the oocyte (*45-47*),
241 consistent with structural constraints on *oskar* sequence evolution. However, we are unaware of
242 any evidence suggesting specific roles for the structure of the coding region of the *oskar*
243 transcript. Thus, taken together with the sequence similarity and strong phylogenetic evidence
244 presented above, we propose that the results of these analyses support an HGT origin for the
245 OSK domain (Figure 4).
246
247 While multiple mechanisms can give rise to novel genes, HGT is arguably among the least well
248 understood, as it involves multiple genomes and ancient biotic interactions between donor and
249 host organisms that are often difficult to reconstruct. In the case of *oskar*, however, the fact that
250 both germline symbionts (*48*) and HGT events (*9*) are widespread in insects, provides a plausible
251 biological mechanism consistent with our hypothesis that fusion of eukaryotic and bacterial

252 domain sequences led to the birth of this novel gene. Under this hypothesis, this fusion would
253 have taken place before the major diversification of insects, nearly 500 million years ago (*29*).
254
255 Once arisen, novel genes might be expected to disappear rapidly, given that pre-existing gene
256 regulatory networks operated successfully without them (*1*). However, it is clear that novel genes
257 can evolve functional connections with existing networks, become essential (*49*), and in some
258 cases lead to new functions (*50*) and contribute to phenotypic diversity (*5*). Even given the
259 growing number of convincing examples of HGT from both prokaryotic and eukaryotic origins
260 (see for example *51, 52-54*), some authors suspect that the contribution of horizontal gene
261 transfer to the acquisition of novel traits has been underestimated across animals (*55*). Moreover,
262 the functional contribution of genes horizontally transferred specifically from bacteria to insects
263 has been documented for a range of adaptive phenotypes (see for example *56, 57, 58*), including
264 digestive metabolism (*10, 11, 59*), glycolysis (*60*) complex symbiosis (*12*) and endosymbiont
265 cell wall construction (*61*). *oskar* plays multiple critical roles in insect development, from neural
266 patterning (*15, 62*) to oogenesis (*43*). In the Holometabola, a clade of nearly one million extant
267 species (*63*), *oskar*'s co-option to become necessary and sufficient for germ plasm assembly is
268 likely the cell biological mechanism underlying the evolution of this derived mode of insect
269 germ line specification (*15, 17, 19*). Our study thus provides evidence that HGT can not only
270 introduce functional genes into a host genome, but also, by contributing sequences of individual
271 domains, generate genes with entirely novel domain structures that may facilitate the evolution
272 of novel developmental mechanisms.
273
274

275 **Methods**

276

277 ***BLAST searches of Oskar***

278 All BLAST (*64*) searches were performed using the NCBI BLASTp tool suite on the non-
279 redundant (nr) database. Amino Acid (AA) sequences of *D. melanogaster* full length Oskar
280 (EMBL ID AAF54306.1), as well as the AA sequences for the *D. melanogaster* Oskar LOTUS
281 (AA 139-238) and OSK (AA 414-606) domains were used for the BLAST searches. We used the
282 default NCBI cut-off parameters (E-value cut-off of 10) for searches using OSK and LOTUS as
283 queries, and a more stringent E-value threshhold of 0.01 for the search using full length *D.*
284 *melanogaster* Oskar as a query. We chose an E-value threshold of 10 for LOTUS and OSK to
285 capture potentially highly divergent homologs of the two domains, especially for the OSK
286 domain, where we were looking for any viable candidate for a homologous eukaryotic domain.
287 All BLAST searches results are included in the Supplementary files: BLAST search results.

288

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

290 101 1KITE transcriptomes (*65*) (Supplementary Table 1) were downloaded and searched using
291 the local BLAST program (BLAST+) using the tblastn algorithm with default parameters, with
292 Oskar protein sequences of *Drosophila melanogaster, Aedes aegypti, Nasonia vitripennis* and
293 *Gryllus bimaculatus* as queries (EntrezIDs: NP_731295.1, ABC41128.1, NP_001234884.1 and
294 AFV31610.1 respectively). For all of these 1KITE transcriptome searches, predicted protein
295 sequences from transcript data were obtained by in silico translation using the online ExPASy
296 translate tool (https://web.expasy.org/translate/), taking the longest open reading frame. Publicly
297 available sequences in the non-redundant (nr), TSA databases at NCBI, and a then-unpublished

298 transcriptome (*66*) (kind gift of Matthew Benton and Siegfried Roth, University of Cologne)
299 were subsequently searched using the web-based BLAST tool hosted at NCBI, using the tblastn
300 algorithm with default parameters. Sequences used for queries were the four Oskar proteins
301 described above, and newfound *oskar* sequences from the 1KITE transcriptomes of *Baetis*
302 *pumilis, Cryptocercus wright,* and *Frankliniella cephalica*. For both searches, *oskar* orthologs
303 were identified by the presence of BLAST hits on the same transcript to both the LOTUS (N-
304 terminal) and OSK (C-terminal) regions of any of the query *oskar* sequences, regardless of E-
305 values. The sequences found were aligned using MUSCLE (8 iterations) (*67*) into a 46-sequence
306 alignment (Supplementary files: Alignments>OSKAR_MUSCLE_INITIAL.fasta). From this
307 alignment, the LOTUS and OSK domains were extracted (Supplementary files:
308 Alignments>LOTUS_MUSCLE_INITIAL.fasta and
309 Alignments>OSK_MUSCLE_INITIAL.fasta) to define the initial Hidden Markov Models
310 (HMM) using the hmmbuild tool from the HMMER tool suite with default parameters (*68*). 126
311 insect genomes and 128 insect transcriptomes (from the Transcriptome Shotgun Assembly TSA
312 database: https://www.ncbi.nlm.nih.gov/Traces/wgs/?view=TSA) were subsequently
313 downloaded from NCBI (download date September 29, 2015 ; Supplementary table 1). Genomes
314 were submitted to Augustus v2.5.5 (*69*) (using the *D. melanogaster* exon HMM predictor) and
315 SNAP v2006-07-28 (*70*) (using the default 'fly' HMM) for gene discovery. The resulting
316 nucleotide sequence database comprising all 309 downloaded and annotated genomes and
317 transcriptomes, was then translated in six frames to generate a non-redundant amino acid
318 database (where all sequences with the same amino acid content are merged into one). This
319 process was automated using a series of custom scripts available here:
320 https://github.com/Xqua/Genomes. The non-redundant amino acid database was searched using

321 the HMMER v3.1 tool suite (*68*) and the HMM for the LOTUS and OSK domains described
322 above. A hit was considered positive if it consisted of a contiguous sequence containing both a
323 LOTUS domain and an OSK domain, with the two domains separated by an inter-domain
324 sequence. We imposed no length, alignment or conservation criteria on the inter-domain
325 sequence, as this is a rapidly-evolving region of Oskar protein with predicted high disorder (*20,*
326 *24, 71*). Positive hits were manually curated and added to the main alignment, and the search was
327 performed iteratively until no more new sequences meeting the above criteria were discovered.
328 This resulted in a total of 95 Oskar protein sequences, (see Supplementary Table 2 for the
329 complete list). Using the final resulting alignment (Supplementary Files:
330 Alignments>OSKAR_MUSCLE_FINAL.fasta), the LOTUS and OSK domains were extracted
331 from these sequences (Supplementary Files: Alignments>LOTUS_MUSCLE_FINAL.fasta and
332 Alignments>OSK_MUSCLE_FINAL.fasta), and the final three HMM (for full-length Oskar,
333 OSK, and LOTUS domains) used in subsequent analyses were created using hmmbuild with
334 default parameters (Supplementary files: HMM>OSK.hmm, HMM>LOTUS.hmm and
335 HMM>OSKAR.hmm).
336
337 ***Iterative HMMER search of OSK and LOTUS domains***
338 A reduced version of TrEMBL (*72*) (v2016-06) was created by concatenating all hits (regardless
339 of E-value) for sequences of the LOTUS domain, the OSK domain and full-length Oskar, using
340 hmmsearch with default parameters and the HMM models created above from the final
341 alignment. This reduced database was created to reduce potential false positive results that might
342 result from the limited size of the sliding window used in the search approach described here.
343 The full-length Oskar alignment of 1133 amino acids (Supplementary files:

344 Alignments>OSKAR_MUSCLE_FINAL.fasta) was split into 934 sub-alignments of 60 amino
345 acids each using a sliding window of one amino acid. Each alignment was converted into a
346 HMM using hmmbuild, and searched against the reduced TrEMBL database using hmmsearch
347 using default parameters. Domain of life origin of every hit sequence at each position was
348 recorded. Eukaryotic sequences were further classified as Oskar/Non-Oskar and Arthropod/Non-
349 Arthropod. Finally, for the whole alignment, the counts for each category were saved and plotted
350 in a stack plot representing the proportion of sequences from each category to create Figure 1b.
351 The python code used for this search is available at https://github.com/Xqua/Iterative-HMMER.
352

353 ***Sequence Similarity Networks***

354 LOTUS and OSK domain sequences from the final alignment obtained as described above (see
355 "*Hidden Markov Model (HMM) generation and alignments of the OSK and LOTUS domains*";
356 Supplementary files: Alignments>LOTUS_MUSCLE_FINAL.fasta and
357 Alignments>OSK_MUSCLE_FINAL.fasta) were searched against TrEMBL (*72*) (v2016-06)
358 using HMMER. All hits with E-value < 0.01 were consolidated into a fasta file that was then
359 entered into the EFI-EST tool (*73*) using default parameters to generate a sequence similarity
360 network. An alignment score corresponding to 30% sequence identity was chosen for the
361 generation of the final sequence similarity network. Finally, the network was graphed using
362 Cytoscape 3 (*74*).
363

364 ***Phylogenetic Analysis Based on MUSCLE Alignment***

365 For both the LOTUS and OSK domains, in cases where more than one sequence from the same
366 organism was retrieved by the search described above in "*Iterative HMMER Search of OSK and*

367 *LOTUS domains*", only the sequence with the lowest E-value was used for phylogenetic analysis.
368 For the LOTUS domain, the first 97 best hits (lowest E-value) were selected, and the only three
369 bacterial sequences that satisfied an E-value < 0.01 were manually added. For *oskar* sequences,
370 if more than one sequence per species was obtained by the search, only the single sequence per
371 species with the lowest E-value was kept for analysis, generating a set of 100 sequences for the
372 LOTUS domain, and 87 sequences for the OSK domain. Unique identifiers for all sequences
373 used to generate alignments for phylogenetic analysis are available in Supplementary Tables S3,
374 S4. For both datasets, the sequences were then aligned using MUSCLE (*67*) (8 iterations) and
375 trimmed using trimAl (*75*) with 70% occupancy. The resulting alignments that were subject to
376 phylogenetic analysis are available in Supplementary Files:
377 Alignments>LOTUS_MUSCLE_TREE.fasta and Alignments>OSK_MUSCLE_TREE.fasta. For
378 the maximum likelihood tree, we used RaxML v8.2.4 (*76*) with 1000 bootstraps, and the models
379 were selected using the automatic RaxML model selection tool. The substitution model chosen
380 for both domains was LGF. For the Bayesian tree inference, we used MrBayes V3.2.6 (*77*) with
381 a Mixed model (prset aamodel=Mixed) and a gamma distribution (lset rates=Gamma). We ran
382 the MonteCarlo for 4 million generations (std < 0.01) for the OSK domain, and for 3 million
383 generations (std < 0.01) for the LOTUS domain. For the tree comparisons (Fig2 supplementary
384 figures 8-9), the RaxML best tree output from the MUSCLE and PRANK alignments were
385 compared using the tool Phylo.io (*78*).
386
387 ***Phylogenetic analysis based on PRANK alignment***
388 For the OSK domain, the raw full length sequences obtained from the HMMER search were
389 aligned to each other using HMMER HMM based alignment tool: hmmalign, with the same

390 HMM used to do the search, namely OSK.hmm (supplementary data: Data/HMM/OSK.hmm).
391 Starting from this base alignment, we used the default alignment method option offered by
392 PRANK (version: v.170427) ((*79*). We then used PRANK to realign those sequences, which in
393 turn led to a usable alignment for phylogenetic analysis. This alignment was trimmed using the
394 same parameters as described in *Hidden Markov Model (HMM) generation and alignments of*
395 *the OSK and LOTUS domains* above. The final alignment is available in supplementary data:
396 Alignment/OSK_prank_aligned.fasta. We then performed a phylogenetic analysis of this
397 alignment using RAXML with the same parameters described in *Phylogenetic Analysis Based on*
398 *MUSCLE Alignment* above. The resulting tree is presented in Figure 2 Supplementary figures 7
399 and 8.
400
401 For the LOTUS domain, the raw full length sequences obtained from the HMMER search were
402 aligned to each other using the HMMER HMM based alignment tool: hmmalign, with the same
403 HMM used to do the search, namely LOTUS.hmm (Supplementary data:
404 Data/HMM/LOTUS.hmm). Starting from this base alignment, we then used PRANK with
405 default options to realign those sequences. This alignment was trimmed using the same
406 parameters as described in the *Hidden Markov Model (HMM) generation and alignments of the*
407 *OSK and LOTUS domains*. The final alignment is available in supplementary data:
408 Alignments/LOTUS_prank_aligned.fasta. We then performed a phylogenetic analysis using
409 RAXML with the same parameters described above in *Phylogenetic Analysis Based on MUSCLE*
410 *alignment*. The resulting trees are presented in Figure 2 Supplementary figures 6 and 9.

411 ***Phylogenetic Analysis Based on T Coffee alignment***

412 For the LOTUS and OSK domain, the raw full length sequences obtained from the HMMER
413 search were aligned to each other using T-Coffee with its default parameters (*80*). This
414 alignment was trimmed using the same parameters as described in *Hidden Markov Model*
415 *(HMM) generation and alignments of the OSK and LOTUS domains* above. The final alignment
416 is available in supplementary data: Alignment/LOTUS_tcoffee_aligned.fasta
417 Alignment/OSK_tcoffee_aligned.fasta. We then performed a phylogenetic analysis of this
418 alignment using RAXML with the same parameters described in *Phylogenetic Analysis Based on*
419 *MUSCLE Alignment* above. The resulting trees are presented in Figure 2 Supplementary figures
420 10 and 11.

421

422 ***Visual Comparison of Phylogenetic Trees***

423 To compare the trees obtained with different alignment tools, we used Phylo.io (*78*). The trees
424 were imported in Newick format, and the Phylo.io tool generated the mirrored and aligned
425 versions of the trees represented in Figure 2 Supplementary figures 8, 9, 12 and 13. The color of
426 the branches is the tree similarity score, where lighter colors represent a higher number of
427 topological differences. Exactly, it is a custom implementation of the Jacard Index by Phylo.io.

428

429 ***Statistical Analysis of Tree Topology***

430 To statistically evaluate our best-supported topology of the OSK and LOTUS trees, we compared
431 constrained topologies to the highest likelihood trees using the SOWHAT tool (*31*). SOWHAT
432 automates the stringent SOWH phylogenetic topology test (*30*), and compares the log likelihood
433 between generated trees. We defined three constrained trees to test our results, one requiring

434 monophyly of all domains of life, a second requiring only eukaryotic monophyly, and the last
435 one requiring monophyly of the oskar LOTUS domain (Supplementary Files:
436 Data>Trees>constrained_kingdom_tree.tre, constrained_eukmono_tree.tre &
437 constrained_lotus_mono_tree.tre). We then ran SOWHAT using its default parameters, 1000
438 bootstraps, and the two constrained trees against the OSK or LOTUS alignment used to generate
439 the phylogenetic trees (Supplementary Files: Alignments>OSK_MUSCLE_TREE.fasta &
440 LOTUS_MUSCLE_TREE.fasta). All best trees generated by SOWHAT are available in
441 (Supplementary Files: Data>Trees>SOWHAT_*_test.tre).
442
443 ***Selection of Sequences for Codon Use Analysis***
444 To study the codon use of the OSK and LOTUS domains, we chose 17 well-annotated (defined
445 as possessing at least 8,000 annotated genes) insect genomes that included a confidently
446 annotated *oskar* orthologue from the NCBI nucleotide database. The complete list and accession
447 numbers of the sequences used for this analysis is in Supplementary Table 5. This list contains
448 *oskar* sequences from genomes that were either added to the databases after the first *oskar*
449 sequence search or re-annotated after said search. Therefore the sequences coming from the
450 following organisms are not represented in the final *oskar* alignment: *Harpegnathos saltator,*
451 *Fopius arisanus, Athalia rosae, Orussus abietinus, Stomoxys calcitrans, Bactrocera oleae,*
452 *Neodiprion lecontei*.
453
454 ***Calculation of Codon Adaptation Index (CAI)***
455 For each of the 17 genomes indicated in Table SX, a codon use index was calculated by
456 measuring the Relative Synonymous Codon Use (RCSU) of all genes in each genome. Using this

457 codon use index, the CAI for all genes, the 3' and 5' cuts of the intra-gene distribution, Oskar,
458 OSK and LOTUS was calculated for each genome. All calculations were performed using the
459 Biopython library (https://biopython.org/ version 1.70). The Delta CAI was computed by taking
460 the CAI difference between the 5' and 3' or LOTUS and OSK. The Z score of the intra-gene and
461 LOTUS OSK were computed against the distribution of CAI of all genes in a genome.
462
463 ***Generation of Intra-Gene Distribution of Ccodon Use***
464 We wished to determine whether *oskar* differed from the null hypothesis that a given gene would
465 follow similar codon use throughout its sequence. To generate a distribution of codon use
466 similarity across a gene for all genes in the genomes studied, we generated what we named the
467 "Intra-Gene" sequence distribution (Figure 3 Supplement 5). Each gene was cut into two
468 fragments at a random position "x" following the rule: 351 < x < Length_gene - 351, x modulo 3
469 = 0 (Corresponding Jupyter notebook file: Scripts>notebook>Codon Analysis AT3 GC3 and A3
470 T3 G3 C3 Section: 4). In other words, for this analysis we split each transcript into two parts,
471 requiring that each part be a minimum of 351 bp (117 codons) long. This approach allowed us to
472 sample at least 117 codons for each of the 3' and 5' parts of a transcript.
473
474 ***Fitting a Linear Model of Codon Use***
475 Using the Intra-Gene null distribution generated above, we fitted a linear model of codon use
476 frequencies per gene for the wobble position and AT3 GC3 content. To do so, we measured the
477 different frequencies of A3, T3, G3 and C3 (any codon ending in A was counted as A3) and AT3
478 GC3. Then, we fitted a linear model to the pairs of 5' and 3' regional codon use values for within
479 each gene, obtained from the Intra-Gene distribution described above (conserving the 3'/5'

480 position information), and for the OSK and LOTUS domains, for each of the 17 genomes
481 analyzed (Supp Table 3). We then calculated the residuals of the Intra-Gene distribution and the
482 LOTUS-OSK distribution. Finally, we determined the Pearson correlation coefficient for all
483 genomes pooled together, and all *oskar* genes pooled together (Corresponding Jupyter notebook
484 file: Scripts>notebook>Codon Analysis AT3 GC3 and A3 T3 G3 C3 Section: 7 and 8).
485
486 ***Calculation and Analysis of the Codon Use Z_score***
487 For each genome, the codon use frequency for AT3/GC3 and A3/T3/G3/C3 was calculated as
488 described above. Then, Z scores for each sequence from the Intra-Gene, OSK or LOTUS domain
489 sequences were calculated against the corresponding genome frequency distribution. The Z
490 scores were then used to generate the analysis of Pearson correlation coefficients shown in
491 Figures 3, Figure 3–figure supplement 1 and 2 (Corresponding Jupyter notebook file:
492 Scripts>notebook>Codon Analysis AT3 GC3 and A3 T3 G3 C3 Section: 3, 5 and 6).
493
494 ***Data Availability***
495 All sequences discovered using the automatic annotation pipeline described in (M&M HMM and
496 oskar search) are annotated as such in Supplementary Table S2.
497
498 ***Code Availability***
499 All custom code generated for this study is available in the GitHub repository
500 https://github.com/extavourlab/Oskar_HGT, commit ID
501 2565e2f313db9eb310f907ef37ab644985fba565.
502

503 **References**

504

505 1. J. S. Taylor, J. Raes, Duplication and divergence: the evolution of new genes and old
506 ideas. *Annual review of genetics* **38**, 615-643 (2004).
507 2. D. Tautz, T. Domazet-Loso, The evolutionary origin of orphan genes. *Nat. Rev. Genet.*
508 **12**, 692-702 (2011).
509 3. P. J. Wittkopp, G. Kalay, Cis-regulatory elements: molecular mechanisms and
510 evolutionary processes underlying divergence. *Nat. Rev. Genet.* **13**, 59-69 (2011).
511 4. H. E. Hoekstra, J. A. Coyne, The locus of evolution: evo devo and the genetics of
512 adaptation. *Evolution* **61**, 995-1016 (2007).
513 5. S. Chen, B. H. Krinsky, M. Long, New genes as drivers of phenotypic evolution. *Nat.*
514 *Rev. Genet.* **14**, 645-660 (2013).
515 6. W. Zhang, P. Landback, A. R. Gschwend, B. Shen, M. Long, New genes drive the
516 evolution of gene interaction networks in the human and mouse genomes. *Genome Biol.*
517 **16**, 202 (2015).
518 7. Y. E. Zhang, P. Landback, M. Vibranovski, M. Long, New genes expressed in human
519 brains: implications for annotating evolving genomes. *BioEssays* **34**, 982-991 (2012).
520 8. S. Chen *et al.*, Frequent recent origination of brain genes shaped the evolution of foraging
521 behavior in Drosophila. *Cell Reports* **1**, 118-132 (2012).
522 9. J. C. Dunning Hotopp *et al.*, Widespread lateral gene transfer from intracellular bacteria
523 to multicellular eukaryotes. *Science (New York, NY)* **317**, 1753-1756 (2007).
524 10. R. Acuna *et al.*, Adaptive horizontal transfer of a bacterial gene to an invasive insect pest
525 of coffee. *Proc Natl Acad Sci U S A* **109**, 4197-4202 (2012).
526 11. D. B. Sloan *et al.*, Parallel histories of horizontal gene transfer facilitated extreme
527 reduction of endosymbiont genomes in sap-feeding insects. *Mol. Biol. Evol.* **31**, 857-871
528 (2014).
529 12. F. Husnik *et al.*, Horizontal gene transfer from diverse bacteria to an insect genome
530 enables a tripartite nested mealybug symbiosis. *Cell* **153**, 1567-1578 (2013).
531 13. R. Lehmann, Germ Plasm Biogenesis--An Oskar-Centric Perspective. *Curr. Top. Dev.*
532 *Biol.* **116**, 679-707 (2016).
533 14. R. Lehmann, C. Nüsslein-Volhard, Abdominal Segmentation, Pole Cell Formation, and
534 Embryonic Polarity Require the Localized Activity of *oskar*, a Maternal Gene in
535 *Drosophila*. *Cell* **47**, 144-152 (1986).
536 15. B. Ewen-Campen, J. R. Srouji, E. E. Schwager, C. G. Extavour, *oskar* Predates the
537 Evolution of Germ Plasm in Insects. *Curr. Biol.* **22**, 2278-2283 (2012).
538 16. C. G. Extavour, M. E. Akam, Mechanisms of germ cell specification across the
539 metazoans: epigenesis and preformation. *Development* **130**, 5869-5884 (2003).
540 17. J. A. Lynch *et al.*, The Phylogenetic Origin of *oskar* Coincided with the Origin of
541 Maternally Provisioned Germ Plasm and Pole Cells at the Base of the Holometabola.
542 *PLoS Genetics* **7**, e1002029 (2011).
543 18. B. Ewen-Campen, S. Donoughe, D. N. Clarke, C. G. Extavour, Germ cell specification
544 requires zygotic mechanisms rather than germ plasm in a basally branching insect. *Curr.*
545 *Biol.* **23**, 835-842 (2013).
546 19. E. Abouheif, Evolution: oskar Reveals Missing Link in Co-optive Evolution. *Curr. Biol.*
547 **23**, R24-R25 (2012).

548 20. M. Jeske *et al.*, The Crystal Structure of the *Drosophila* Germline Inducer Oskar
549 Identifies Two Domains with Distinct Vasa Helicase- and RNA-Binding Activities. *Cell*
550 *Reports* **12**, 587-598 (2015).
551 21. M. Jeske, C. W. Muller, A. Ephrussi, The LOTUS domain is a conserved DEAD-box
552 RNA helicase regulator essential for the recruitment of Vasa to the germ plasm and
553 nuage. *Genes Dev.* **31**, 939-952 (2017).
554 22. A. Ephrussi, R. Lehmann, Induction of germ cell formation by *oskar*. *Nature* **358**, 387-
555 392 (1992).
556 23. J. Kim-Ha, J. L. Smith, P. M. Macdonald, *oskar* mRNA is localized to the posterior pole
557 of the *Drosophila* oocyte. *Cell* **66**, 23-35 (1991).
558 24. N. Yang *et al.*, Structure of *Drosophila* Oskar reveals a novel RNA binding protein.
559 *Proc. Natl. Acad. Sci. USA* **112**, 11541-11546 (2015).
560 25. A. H. Kirk-Spriggs, B. J. Sinclair, Eds., *Manual of Afrotropical Diptera*, (South African
561 National Biodiversity Institute, Pretoria, South Africa, 2017), vol. 1.
562 26. R. S. Peters *et al.*, Evolutionary History of the Hymenoptera. *Curr. Biol.* **27**, 1013-1018
563 (2017).
564 27. N. F. Vanzo, A. Ephrussi, Oskar anchoring restricts pole plasm formation to the posterior
565 of the *Drosophila* oocyte. *Development* **129**, 3705-3714 (2002).
566 28. T. R. Hurd *et al.*, Long Oskar Controls Mitochondrial Inheritance in *Drosophila*
567 *melanogaster*. *Dev. Cell* **39**, 560-571 (2016).
568 29. B. Misof *et al.*, Phylogenomics resolves the timing and pattern of insect evolution.
569 *Science* **346**, 763-767 (2014).
570 30. D. L. Swofford, G. J. Olsen, P. J. Waddell, D. M. Hillis, in *Molecular Systematics 2nd*
571 *Edition,* D. M. Hillis, C. Moritz, B. K. Mable, Eds. (Sinauer Associates, Inc., Sinauer,
572 MA, 1996), chap. 11, pp. 407-453.
573 31. S. H. Church, J. F. Ryan, C. W. Dunn, Automation and Evaluation of the SOWH Test
574 with SOWHAT. *Syst. Biol.* **64**, 1048-1058 (2015).
575 32. D. Wheeler, A. J. Redding, J. H. Werren, Characterization of an ancient lepidopteran
576 lateral gene transfer. *PLoS ONE* **8**, e59262 (2013).
577 33. S. T. Chepkemoi *et al.*, Identification of Spiroplasmainsolitum symbionts in Anopheles
578 gambiae. *Wellcome Open Research* **2**, 90 (2017).
579 34. E. Zchori-Fein, S. J. Perlman, S. E. Kelly, N. Katzir, M. S. Hunter, Characterization of a
580 'Bacteroidetes' symbiont in Encarsia wasps (Hymenoptera: Aphelinidae): proposal of
581 'Candidatus Cardinium hertigii'. *International Journal of Systematic and Evolutionary*
582 *Microbiology* **54**, 961-968 (2004).
583 35. E. Zchori-Fein, S. J. Perlman, Distribution of the bacterial symbiont Cardinium in
584 arthropods. *Mol. Ecol.* **13**, 2009-2016 (2004).
585 36. T. Tuller, Codon bias, tRNA pools and horizontal gene transfer. *Mobile Genetic Elements*
586 **1**, 75-77 (2011).
587 37. P. M. Sharp, W. H. Li, The codon Adaptation Index--a measure of directional
588 synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15**,
589 1281-1295 (1987).
590 38. J. J. Li, G. L. Chew, M. D. Biggin, Quantitative principles of cis-translational control by
591 general mRNA sequence features in eukaryotes. *Genome Biol.* **20**, 162 (2019).
592 39. C. H. Yu *et al.*, Codon Usage Influences the Local Rate of Translation Elongation to
593 Regulate Co-translational Protein Folding. *Mol Cell* **59**, 744-754 (2015).

594 40. S. Pechmann, J. Frydman, Evolutionary conservation of codon optimality reveals hidden
595 signatures of cotranslational folding. *Nat Struct Mol Biol* **20**, 237-243 (2013).
596 41. M. Zhou, T. Wang, J. Fu, G. Xiao, Y. Liu, Nonoptimal codon usage influences protein
597 structure in intrinsically disordered regions. *Mol. Microbiol.* **97**, 974-987 (2015).
598 42. P. Krishnakumar *et al.*, Functional equivalence of germ plasm organizers. *PLoS Genet*
599 **14**, e1007696 (2018).
600 43. A. Jenny *et al.*, A translation-independent role of oskar RNA in early Drosophila
601 oogenesis. *Development* **133**, 2827-2833 (2006).
602 44. M. Kanke *et al.*, oskar RNA plays multiple noncoding roles to support oogenesis and
603 maintain integrity of the germline/soma distinction. *RNA* **21**, 1096-1109 (2015).
604 45. H. Jambor, C. Brunel, A. Ephrussi, Dimerization of oskar 3' UTRs promotes hitchhiking
605 for RNA localization in the Drosophila oocyte. *RNA* **17**, 2049-2057 (2011).
606 46. H. Jambor, S. Mueller, S. L. Bullock, A. Ephrussi, A stem-loop structure directs oskar
607 mRNA to microtubule minus ends. *RNA* **20**, 429-439 (2014).
608 47. M. Zubradt *et al.*, DMS-MaPseq for genome-wide or targeted RNA structure probing in
609 vivo. *Nat Methods* **14**, 75-82 (2017).
610 48. K. Bourtzis, T. A. Miller, Eds., *Insect Symbiosis*, (CRC Press, Boca Raton, FL, 2006),
611 vol. 3, pp. 304.
612 49. S. Chen, Y. E. Zhang, M. Long, New genes in *Drosophila* quickly become essential.
613 *Science* **330**, 1682-1685 (2010).
614 50. G. Cornelis *et al.*, Ancestral capture of syncytin-Car1, a fusogenic endogenous retroviral
615 envelope gene involved in placentation and conserved in Carnivora. *Proc. Natl. Acad.*
616 *Sci. USA* **109**, E432-441 (2012).
617 51. F. Husnik, J. P. McCutcheon, Functional horizontal gene transfer from bacteria to
618 eukaryotes. *Nat Rev Microbiol* **16**, 67-79 (2018).
619 52. I. Di Lelio *et al.*, Evolution of an insect immune barrier through horizontal gene transfer
620 mediated by a parasitic wasp. *PLoS Genet* **15**, e1007998 (2019).
621 53. N. Wybouw, Y. Pauchet, D. G. Heckel, T. Van Leeuwen, Horizontal Gene Transfer
622 Contributes to the Evolution of Arthropod Herbivory. *Genome Biol. Evol.* **8**, 1785-1801
623 (2016).
624 54. D. G. Quispe-Huamanquispe, G. Gheysen, J. F. Kreuze, Horizontal Gene Transfer
625 Contributes to Plant Evolution: The Case of Agrobacterium T-DNAs. *Front Plant Sci* **8**,
626 2015 (2017).
627 55. L. Boto, Horizontal gene transfer in the acquisition of novel traits by metazoans. *Proc*
628 *Biol Sci* **281**, 20132450 (2014).
629 56. A. C. Wilson, R. P. Duncan, Signatures of host/symbiont genome coevolution in insect
630 nutritional endosymbioses. *Proc Natl Acad Sci U S A* **112**, 10255-10261 (2015).
631 57. S. Lopez-Madrigal, R. Gil, Et tu, Brute? Not Even Intracellular Mutualistic Symbionts
632 Escape Horizontal Gene Transfer. *Genes (Basel)* **8**, (2017).
633 58. N. A. Provorov, O. P. Onischuk, Microbial Symbionts of Insects: Genetic Organization,
634 Adaptive Role, and Evolution. *Mikrobiologiya* **87**, 151-163 (2018).
635 59. M. Shelomi *et al.*, Horizontal Gene Transfer of Pectinases from Bacteria Preceded the
636 Diversification of Stick and Leaf Insects. *Sci Rep* **6**, 26388 (2016).
637 60. Z. Zeng *et al.*, Bacterial endosymbiont Cardinium cSfur genome sequence provides
638 insights for understanding the symbiotic relationship in Sogatella furcifera host. *BMC*
639 *Genomics* **19**, 688 (2018).

640 61. D. C. Bublitz *et al.*, Peptidoglycan Production by an Insect-Bacterial Mosaic. *Cell* **179**,
641 703-712 e707 (2019).
642 62. X. Xu, J. L. Brechbiel, E. R. Gavis, Dynein-Dependent Transport of *nanos* RNA in
643 *Drosophila* Sensory Neurons Requires Rumpelstiltskin and the Germ Plasm Organizer
644 Oskar. *J. Neurosci.* **33**, 14791-14800 (2013).
645 63. J. A. Rees, K. Cranston, Automated assembly of a reference taxonomy for phylogenetic
646 data synthesis. *Biodiversity Data Journal* **5**, e12581 (2017).
647 64. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment
648 search tool. *J. Mol. Biol.* **215**, 403-410 (1990).
649 65. H. Aspöck *et al.* (2018).
650 66. M. A. Benton, N. J. Kenny, K. H. Conrads, S. Roth, J. A. Lynch, Deep, Staged
651 Transcriptomic Resources for the Novel Coleopteran Models Atrachya menetriesi and
652 Callosobruchus maculatus. *PLoS ONE* **11**, e0167431 (2016).
653 67. R. C. Edgar, MUSCLE: a multiple sequence alignment method with reduced time and
654 space complexity. *BMC Bioinformatics* **5**, 113 (2004).
655 68. S. R. Eddy. (2007).
656 69. M. Stanke, R. Steinkamp, S. Waack, B. Morgenstern, AUGUSTUS: a web server for
657 gene finding in eukaryotes. *Nucleic Acids Res.* **32**, W309-312 (2004).
658 70. I. Korf, Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
659 71. A. Ahuja, C. G. Extavour, Patterns of molecular evolution of the germ line specification
660 gene *oskar* suggest that a novel domain may contribute to functional divergence in
661 *Drosophila*. *Dev. Genes Evol.* **222**, 65-77 (2014).
662 72. U. Consortium, The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res.* **37**,
663 D169-174 (2009).
664 73. J. A. Gerlt *et al.*, Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): A web
665 tool for generating protein sequence similarity networks. *Biochimica et Biophysica Acta*
666 *(BBA) - Proteins and Proteomics* **1854**, 1019-1037 (2015).
667 74. P. Shannon *et al.*, Cytoscape: a software environment for integrated models of
668 biomolecular interaction networks. *Genome Res.* **13**, 2498-2504 (2003).
669 75. S. Capella-Gutierrez, J. M. Silla-Martinez, T. Gabaldon, trimAl: a tool for automated
670 alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972-1973
671 (2009).
672 76. A. Stamatakis, RAxML version 8: a tool for phylogenetic analysis and post-analysis of
673 large phylogenies. *Bioinformatics* **30**, 1312-1313 (2014).
674 77. J. P. Huelsenbeck, F. Ronquist, MRBAYES: Bayesian inference of phylogenetic trees.
675 *Bioinformatics* **17**, 754-755 (2001).
676 78. O. Robinson, D. Dylus, C. Dessimoz, Phylo.io: Interactive Viewing and Comparison of
677 Large Phylogenetic Trees on the Web. *Mol. Biol. Evol.* **33**, 2163-2166 (2016).
678 79. A. Loytynoja, Phylogeny-aware alignment with PRANK. *Methods Mol. Biol.* **1079**, 155-
679 170 (2014).
680 80. C. Notredame, D. G. Higgins, J. Heringa, T-Coffee: A novel method for fast and accurate
681 multiple sequence alignment. *J. Mol. Biol.* **302**, 205-217 (2000).
682

Acknowledgments:

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

Funding: This work was supported by funds from Harvard University.

Author contributions: CGME conceived of the project and overall experimental design. TEMJ collected initial transcriptome datasets and identified *oskar* orthologues therein. LB built the HMM model, identified additional orthologues, and performed sequence, phylogenetic, cluster and codon use analyses. LB and CGME interpreted data and wrote the manuscript.

Competing interests: The authors declare no competing interests.

Data and materials availability: All data is available in the main text or the supplementary materials.

Supplementary Information 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:

BLONDEL_JONES_EXTAVOUR_HGT_SUPPLEMENTARY_MATERIALS_ALL

(THIS PDF DOCUMENT)

1. Supplementary Tables
    a. Table S1: List of genomes and transcriptomes used for automated *oskar* search.
    b. Table S2: List of *oskar* sequences used in the final alignment.
    c. Table S3: List of sequences used for phylogenetic analysis of the LOTUS domain.
    d. Table S4: List of sequences used for phylogenetic analysis of the OSK domain.
    e. Table S5: List of genomes analyzed for codon use.

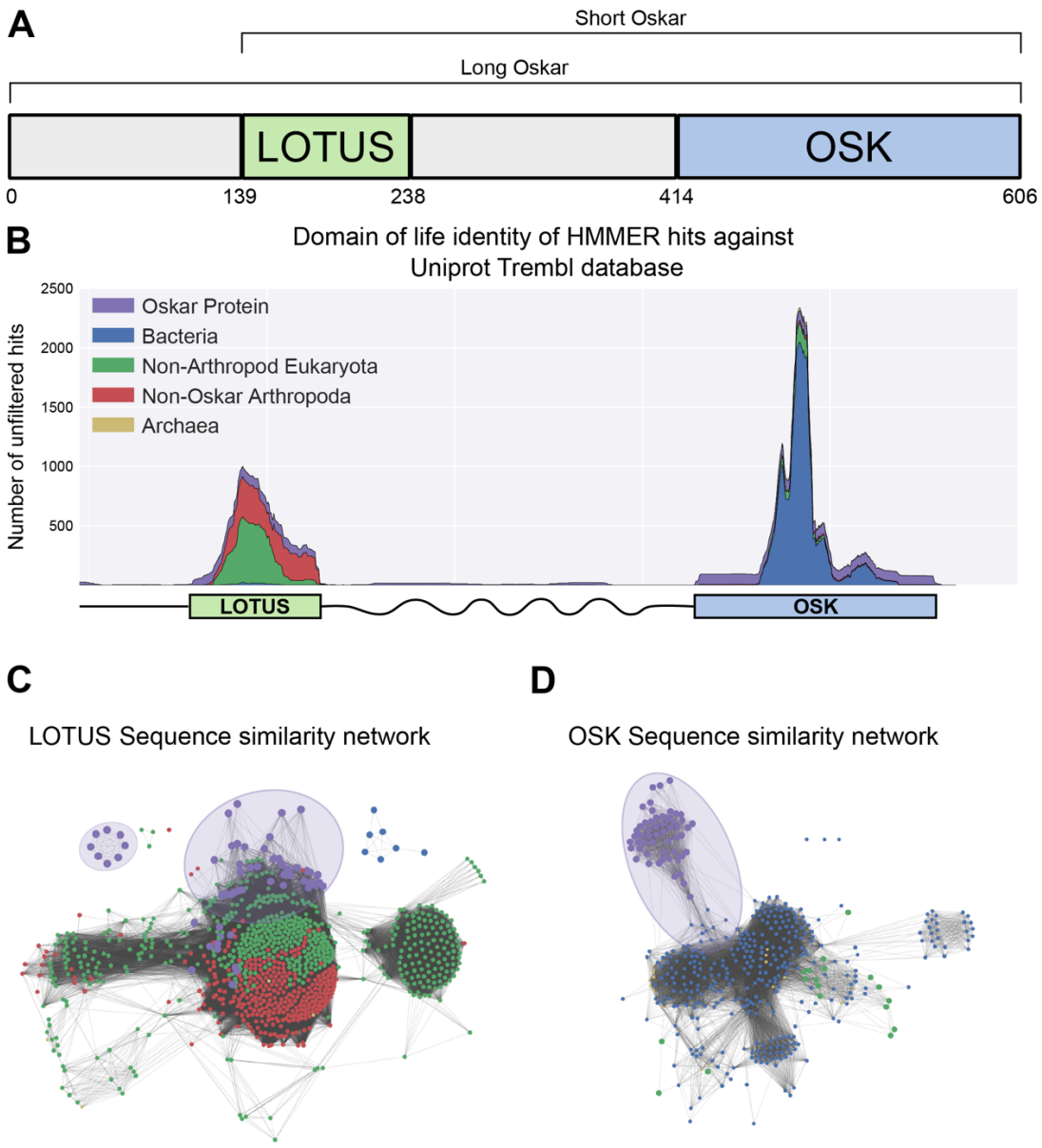
All scripts used for this analysis are hosted on GitHub at
https://github.com/extavourlab/Oskar_HGT

OSKAR HGT SUPPLEMENTARY INFORMATION FILES
(FOLDER)

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.

Please download Oskar HGT SUPPLEMENTARY INFORMATION FILES Folder here:

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



741

742 **Figure 1**. **Sequence analysis of the Oskar gene. a**, Schematic representation of the Oskar gene. The LOTUS and
743 OSK hydrolase-like domains are separated by a poorly conserved region of predicted high disorder and variable
744 length between species. In some dipterans, a region 5' to the LOTUS domain is translated to yield a second isoform,
745 called Long Oskar. Residue numbers correspond to the *D. melanogaster* Osk sequence. **b**, Stackplot of domain of
746 life identity of HMMER hits across the protein sequence. For a sliding window of 60 Amino Acids across the
747 protein sequence (X axis), the number of hits in the Trembl (UniProt) database (Y axis) is represented and color
748 coded by domain of life origin (see Methods: Iterative HMMER search of OSK and LOTUS domains), stacked on
749 top of each other. **c, d** EFI-EST-generated graphs of the sequence similarity network of the LOTUS (**c**) and OSK (**d**)
750 domains of Oskar (*73*). Sequences were obtained using HMMER against the UniProtKB database. Most Oskar
751 LOTUS sequences cluster within eukaryotes and arthropods. In contrast, Oskar OSK sequences cluster most
752 strongly with a small subset of bacterial sequences.

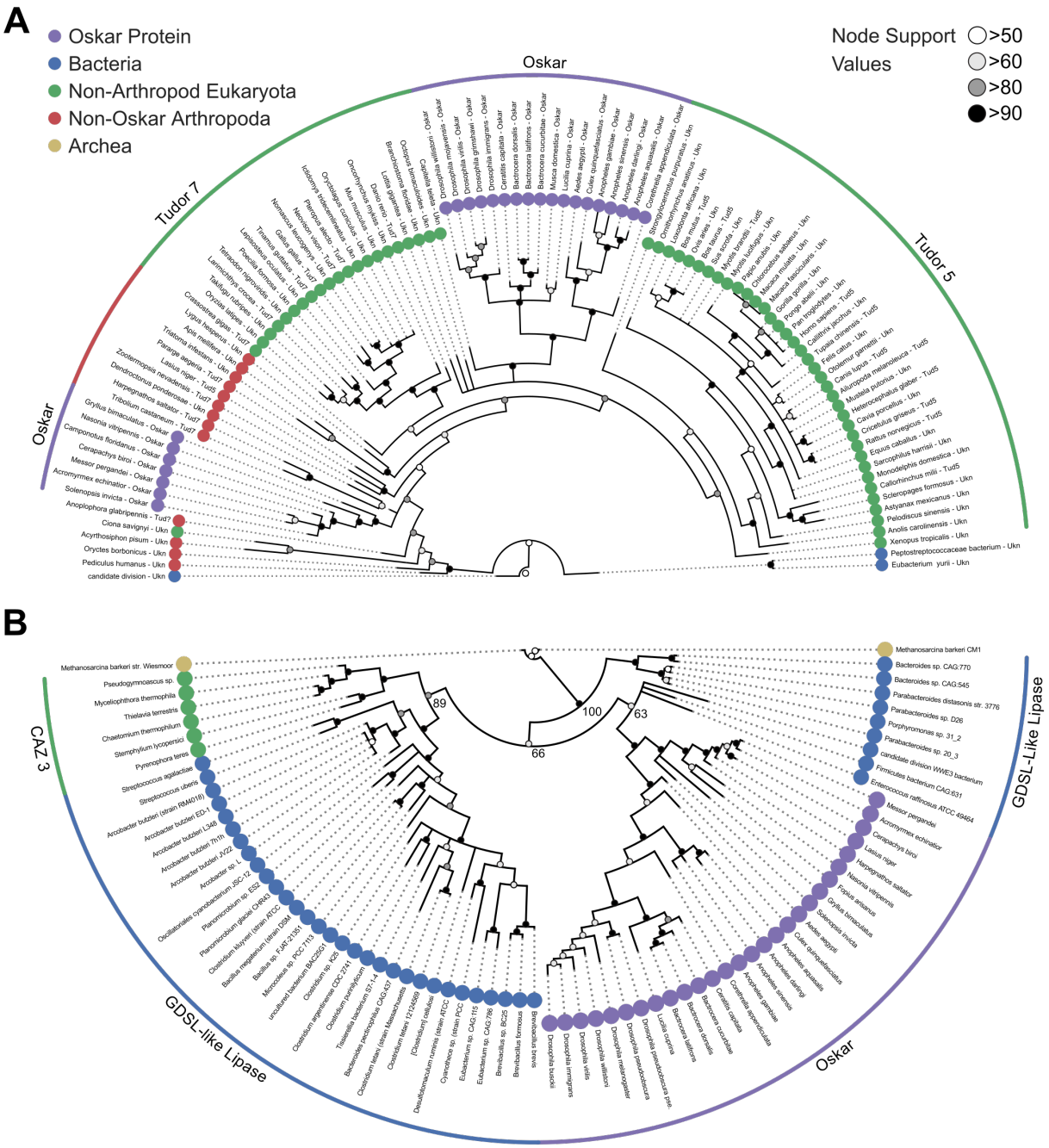


Figure 2. **Phylogenetic analysis of the LOTUS and OSK domains. a**, Bayesian consensus tree for the LOTUS domain. Three major LOTUS-containing protein families are represented within the tree: Tudor 5, Tudor 7, and Oskar. Oskar LOTUS domains form two clades, one containing only dipterans and one containing 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 to contain germ line symbionts in insects. The ten non-Oskar eukaryotic sequences in the analysis form one 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 supplement 1-4.

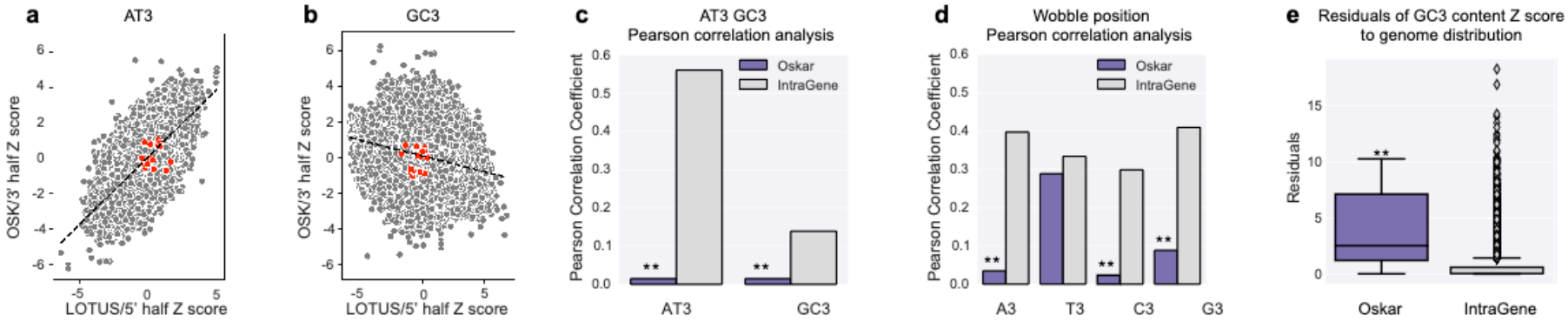


Figure 3. Parametric analysis of codon use for the LOTUS and OSK domains. a, Scatter plot representing the full distribution of correlations from Pearson correlation analysis for AT3 use in 3' and 5' halves of all non-*oskar* genes in the 17 genomes analyzed (grey), compared to the correlation for the OSK and LOTUS domains of *oskar* (red). The Z score ranges of correlations for the *oskar* domains are within the distribution for all genes in the genome, although correlations of use across the gene are lower for *oskar* than for non-*oskar* genes (see also **c**). **b**, Scatter plot representing the full distribution of correlations from Pearson correlation analysis for GC3 use in 3' and 5' halves of all non-*oskar* genes in the 17 genomes analyzed (grey), compared to the correlation for the OSK and LOTUS domains of *oskar* (red). The Z score ranges of correlations for the *oskar* domains are within the distribution for all genes in the genome, although correlations of use across the gene are lower for *oskar* than for non-*oskar* genes (see also **c**). **c**, Pearson correlation analysis of AT3 and GC3 content for Oskar vs other genes. AT3 and GC3 content are correlated across the 5' and 3' halves of a gene for all genes in a given genome (grey), but not between the LOTUS and OSK domains of Oskar (purple). (**: Pearson correlation p-value > 0.1, AT3 r=0.014 p=0.646 and GC3 r=0.014 and p=0.646) **d**, Pearson correlation analysis of wobble position identity for the Oskar gene vs other genes. Wobble position identity content is correlated across the 5' and 3' halves of a gene for all genes in a given genome (grey) but not between the LOTUS and OSK domains of Oskar (purple), with the exception of T3. (**: Pearson correlation p-value > 0.1, A3 r=0.034 p=0.476, T3 r=0.289 p=0.026, C3 r=0.023 p=0.560 and G3 r=0.088 p=0.247) **e**, Analysis of GC3 content. Measure of the residuals of Z scores for Oskar gene GC3 content (LOTUS vs OSK: purple) and the IntraGene GC3 content (grey). The GC3 content of the LOTUS and OSK domains does not follow a linear relationship, and the residuals are significantly higher (purple) than those observed across the 5' and 3' halves of other genes within a given genome (grey). (** : Mann-Whitney U test p-value < 10^{-5} p=$1.927.10^{-7}$).

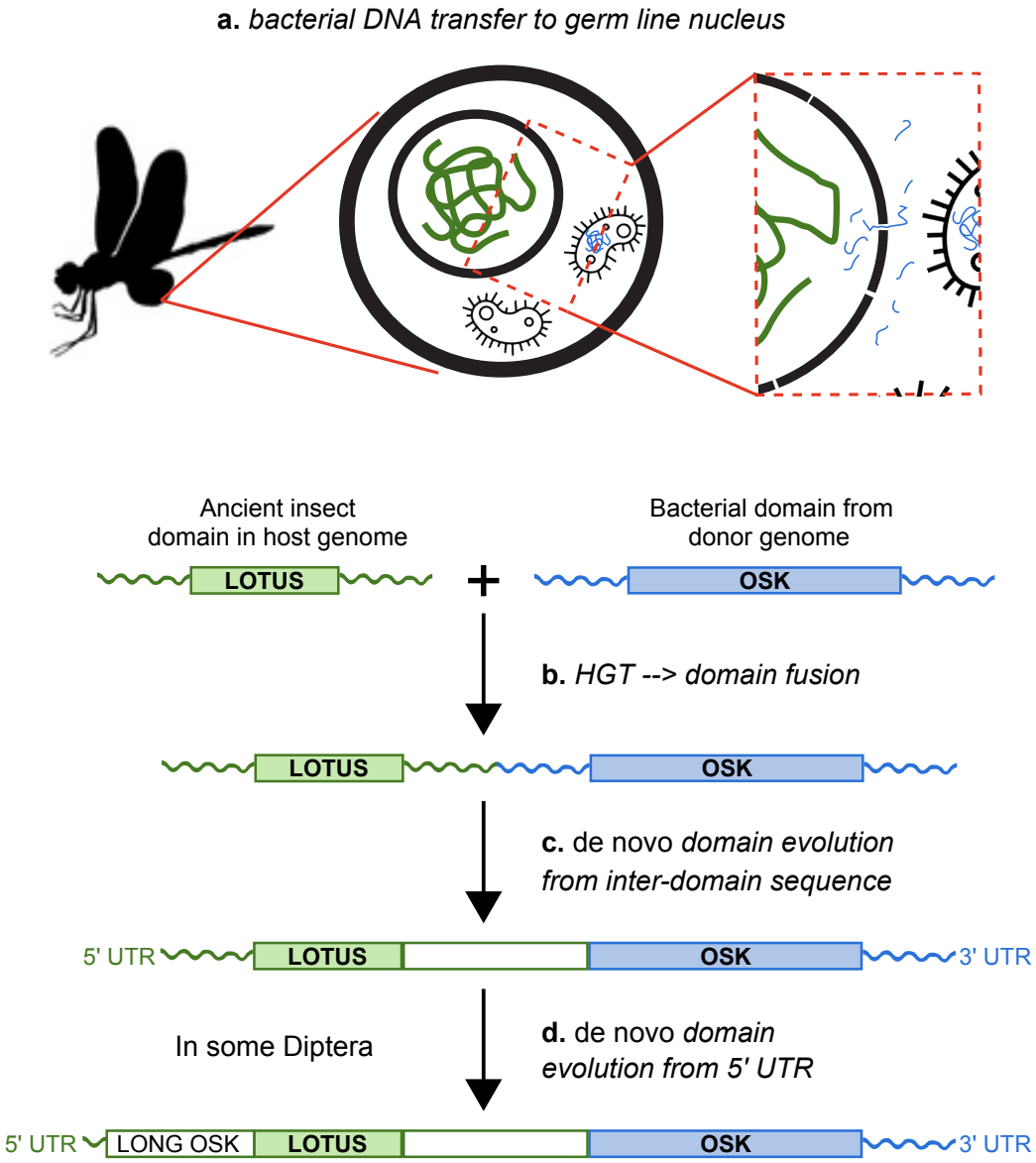


Figure 4. Hypothesis for the origin of *oskar*. Integration of the OSK domain close to a LOTUS domain in an ancestral insect genome. **a**, DNA containing a GDSL-like domain from an endosymbiotic germ line bacterium is transferred to the nucleus of a germ cell in an insect common ancestor. **b**, DNA 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.

Supplementary Figure Legends

Figure 2 Supplements

Figure 2–figure supplement 1: LOTUS Domain RaxML MUSCLE Tree. Phylogenetic tree of the HMMER sequences retrieved from the UniProt database using the LOTUS alignment HMM model. The top 97 hits were selected for phylogenetic analysis, and the only three bacterial sequences found to be a match were added to the alignment manually. The resulting 100 sequences were aligned using MUSCLE with default settings. The sequences were filtered to contain only one sequence per species (best E-value kept) yielding 100 sequences for analysis. Finally, the tree was created using RaxML v8.2.4, using 1000 bootstraps and model selection performed by the RaxML automatic model selection tool. See "Phylogenetic Analysis" in Methods for further detail. Sequences are color-coded as follows: Purple = Oskar; Red = Non-Oskar Arthropod; Green = Non-Arthropod Eukaryote; Blue = Bacteria. Names following leaves display the UniProt accession number followed by the species name and the UniProt protein name.

Figure 2–figure supplement 2: LOTUS Domain Bayesian MUSCLE Tree. Phylogenetic tree of the HMMER sequences retrieved from the UniProt database using the LOTUS alignment HMM model. 100 sequences were chosen for analysis as described for Supplementary Figure 1. The tree was created using Mr Bayes V3.2.6 using a Mixed model (prset aamodel=Mixed) and a gamma distribution (lset rates=Gamma). The algorithm was allowed to run for 3 million generations to achieve a std < 0.01. See "Phylogenetic Analysis" in Methods for further detail. Sequences are color-coded as follows: Purple = Oskar; Red = Non-Oskar Arthropod; Green = Non-Arthropod Eukaryote; Blue = Bacteria. Names following leaves display the UniProt accession number followed by the species name and the UniProt protein name.

Figure 2–figure supplement 3: OSK Domain RaxML MUSCLE Tree. Phylogenetic tree of the HMMER sequences retrieved from the UniProt database using the OSK alignment HMM model. The top 95 hits were selected for phylogenetic analysis, and the only five non-Oskar eukaryotic sequences found to be a match were added to the alignment manually. The resulting 100 sequences were aligned using MUSCLE with default settings. The sequences were filtered to contain only one sequence per species (best E-value kept), yielding 87 sequences for analysis. Finally, the tree was created using RaxML v8.2.4, using 1000 bootstraps and model selection performed by the RaxML automatic model selection tool. See "Phylogenetic Analysis" in Methods for further detail. Sequences are color-coded as follows: Purple = Oskar; Red = Non-Oskar Arthropod; Green = Non-Arthropod Eukaryote; Blue = Bacteria. Names following leaves display the UniProt accession number followed by the species name and the UniProt protein name.

Figure 2–figure supplement 4: OSK Domain Bayesian MUSCLE Tree. Phylogenetic tree of the HMMER sequences hit on the UniProt database using the OSK alignment HMM model. 87 sequences were chosen for analysis as described for Supplementary Figure 3.The tree was created using Mr Bayes V3.2.6 using a Mixed model (prset aamodel=Mixed) and a gamma distribution (lset rates=Gamma). The algorithm was allowed to run for 4 million generations to achieve a std < 0.01. See "Phylogenetic Analysis" in Methods for further detail. Sequences are color-coded as follows: Purple = Oskar; Red = Non-Oskar Arthropod; Green = Non-Arthropod Eukaryote; Blue

834 = Bacteria. Names following leaves display the UniProt accession number followed by the species
835 name and the UniProt protein name.
836
837 **Figure 2–figure supplement 5: SOWHAT constrained trees and results.** Two trees constrained
838 by alternative relationships that would be expected under vertical transmission of sequences were
839 designed and tested against our result supporting a putative HGT event of the OSK domain. (**a**)
840 The first tree (right) is constrained by domain of life, requiring bacterial and eukaryotic sequences
841 to be monophyletic, and disallowing sister group relationships of subsets of eukaryotic sequences
842 and bacterial sequences. Our unconstrained tree topology (left) outperformed this topology with a
843 p-value of 0.002 (95% confidence interval upper: 0.007 lower: 0.0002). (**b**) The second tree
844 requires monophyly of Eukaryota. Our unconstrained tree topology (left) outperformed this
845 topology with a p-value of 0.009 (95% confidence interval upper: 0.017 lower: 0.004). (**c**) The
846 third tree tested whether the LOTUS domain split observed in the tree generated with the MUSCLE
847 alignment was significantly different from a tree where the LOTUS sequences formed a
848 monophyly. The unconstrained tree (left) outperformed this topology with a p-value of 0.037 (95%
849 confidence interval upper: 0.05 lower: 0.026).
850

851 **Figure 2–figure supplement 6: LOTUS Domain RaxML PRANK Tree.** Phylogenetic tree of
852 the same sequences used for the previous LOTUS trees. The sequences were aligned using
853 PRANK and the tree generated with RaxML as described in ***Phylogenetic Analysis Based on***
854 ***PRANK alignment***. Sequences are color-coded as follows: Purple = Oskar; Red = Non-Oskar
855 Arthropod; Green = Non-Arthropod Eukaryote; Blue = Bacteria. Names following leaves display
856 the UniProt accession number followed by the species name and the UniProt protein name.
857

858 **Figure 2–figure supplement 7: OSK Domain RaxML PRANK Tree.** Phylogenetic tree of the
859 same sequences used for the previous OSK trees. The sequences were aligned using PRANK and
860 the tree generated with RaxML as described in ***Phylogenetic Analysis Based on PRANK***
861 ***alignment***. Sequences are color-coded as follows: Purple = Oskar; Red = Non-Oskar Arthropod;
862 Green = Non-Arthropod Eukaryote; Blue = Bacteria. Names following leaves display the
863 UniProt accession number followed by the species name and the UniProt protein name.
864
865 **Figure 2–figure supplement 8: OSK Tree PRANK Comparison.** Comparison of the tree
866 obtained with RaxML starting from the MUSCLE alignment (left) versus the PRANK alignment
867 (right) for the OSK domain. Similarity scores for the branching events are color coded from
868 yellow to blue (see figure color bar legend). The OSK (purple) clade and CAZ3 (green) clade
869 have been colored and compacted for readability as they do not have any internal branching
870 changes. Node color is blue if the leaf is a sequence, and red if this is a compacted group of
871 sequences.
872
873 **Figure 2–figure supplement 9: LOTUS Tree PRANK Comparison.** Comparison of the tree
874 obtained with RaxML starting from the MUSCLE alignment (left) versus the PRANK alignment
875 (right) for the LOTUS domain. Similarity scores for the branching events are color coded from
876 yellow to blue (see figure color bar legend). The LOTUS (purple) clades have been colored and

877 compacted for readability as they do not have any internal branching changes. Node color is blue
878 if the leaf is a sequence, and red if this is a compacted group of sequences.
879

880 **Figure 2–figure supplement 10: LOTUS Domain RaxML T-Coffee Tree.** Phylogenetic tree
881 of the same sequences used for the previous LOTUS trees. The sequences were aligned using T-
882 Coffee and the tree generated with RaxML as described in ***Phylogenetic Analysis Based on* T-**
883 **Coffee *alignment***. Sequences are color-coded as follows: Purple = Oskar; Red = Non-Oskar
884 Arthropod; Green = Non-Arthropod Eukaryote; Blue = Bacteria. Names following leaves display
885 the UniProt accession number followed by the species name and the UniProt protein name.

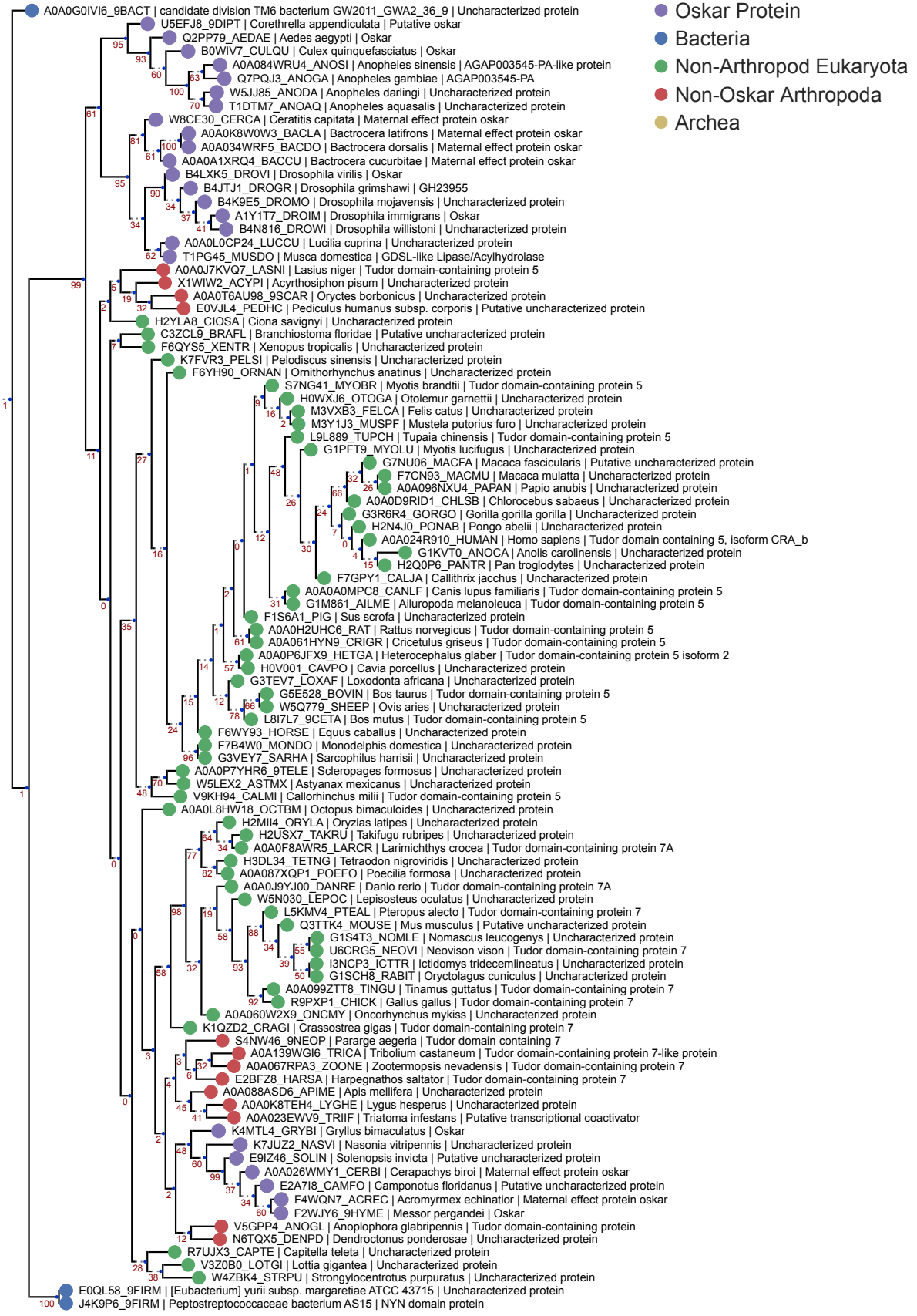
886

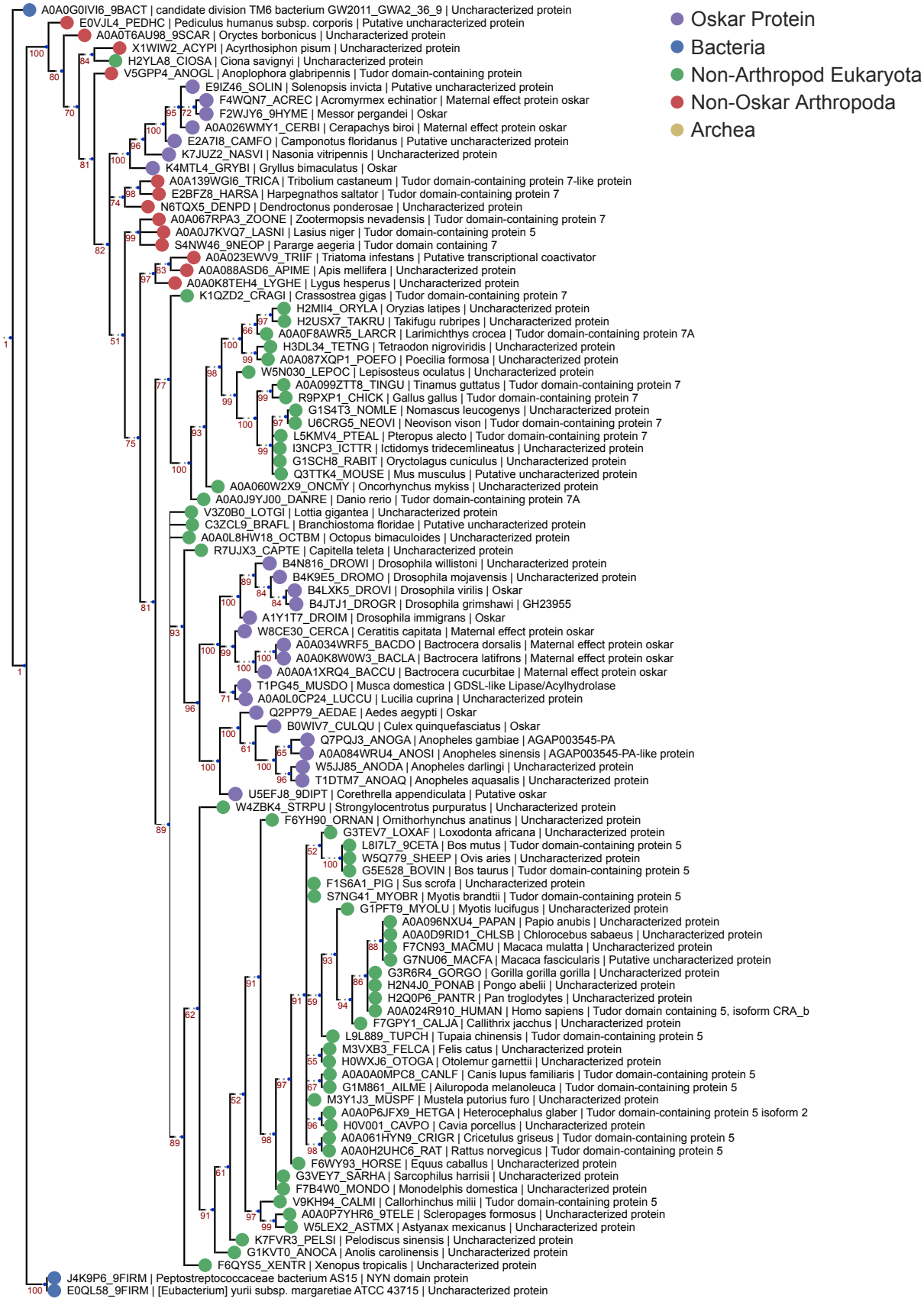
887 **Figure 2–figure supplement 11: OSK Domain RaxML T-Coffee Tree.** Phylogenetic tree of
888 the same sequences used for the previous OSK trees. The sequences were aligned using T-Coffee
889 and the tree generated with RaxML as described in ***Phylogenetic Analysis Based on* T-Coffee**
890 ***alignment***. Sequences are color-coded as follows: Purple = Oskar; Red = Non-Oskar Arthropod;
891 Green = Non-Arthropod Eukaryote; Blue = Bacteria. Names following leaves display the
892 UniProt accession number followed by the species name and the UniProt protein name.

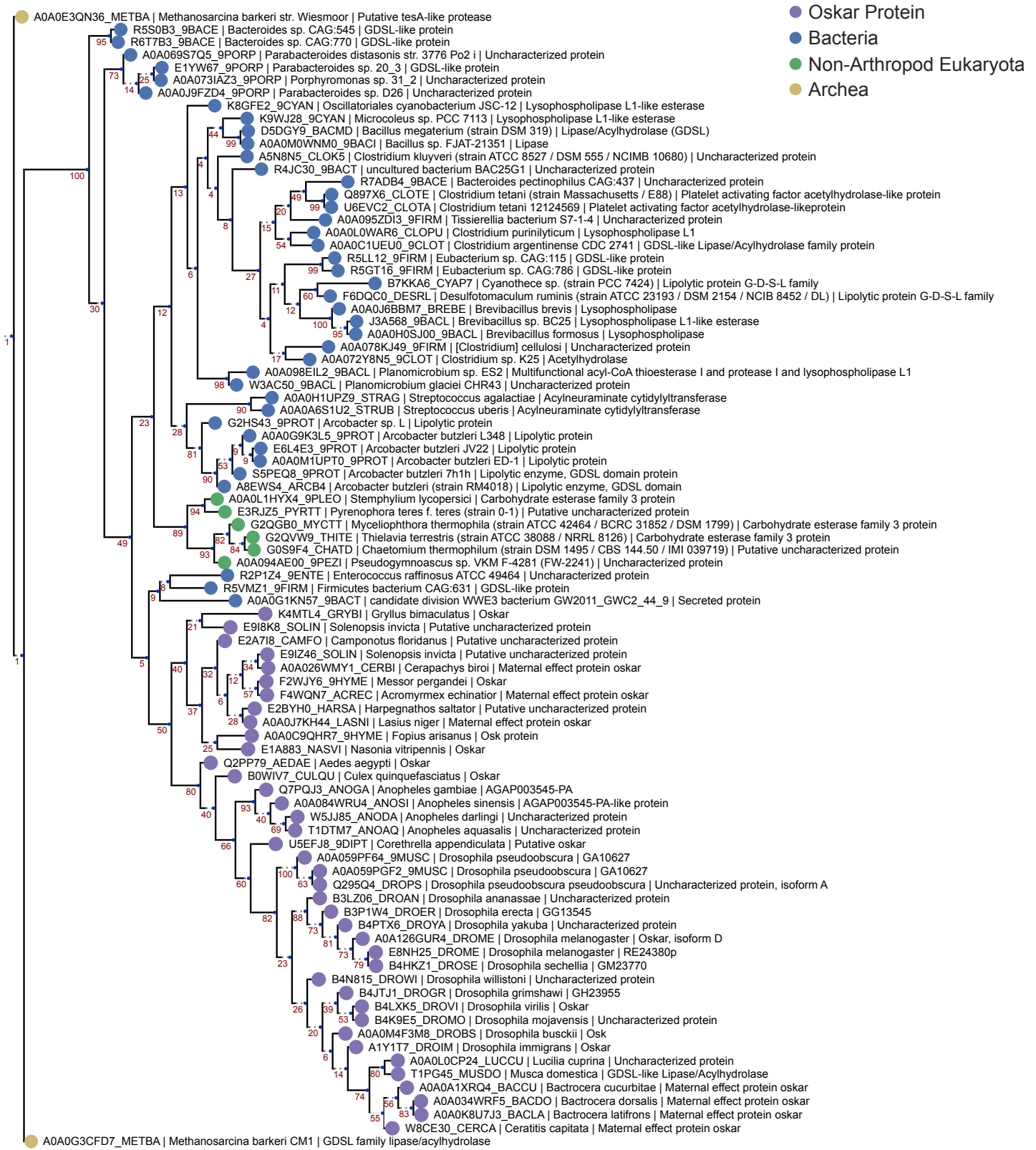
893

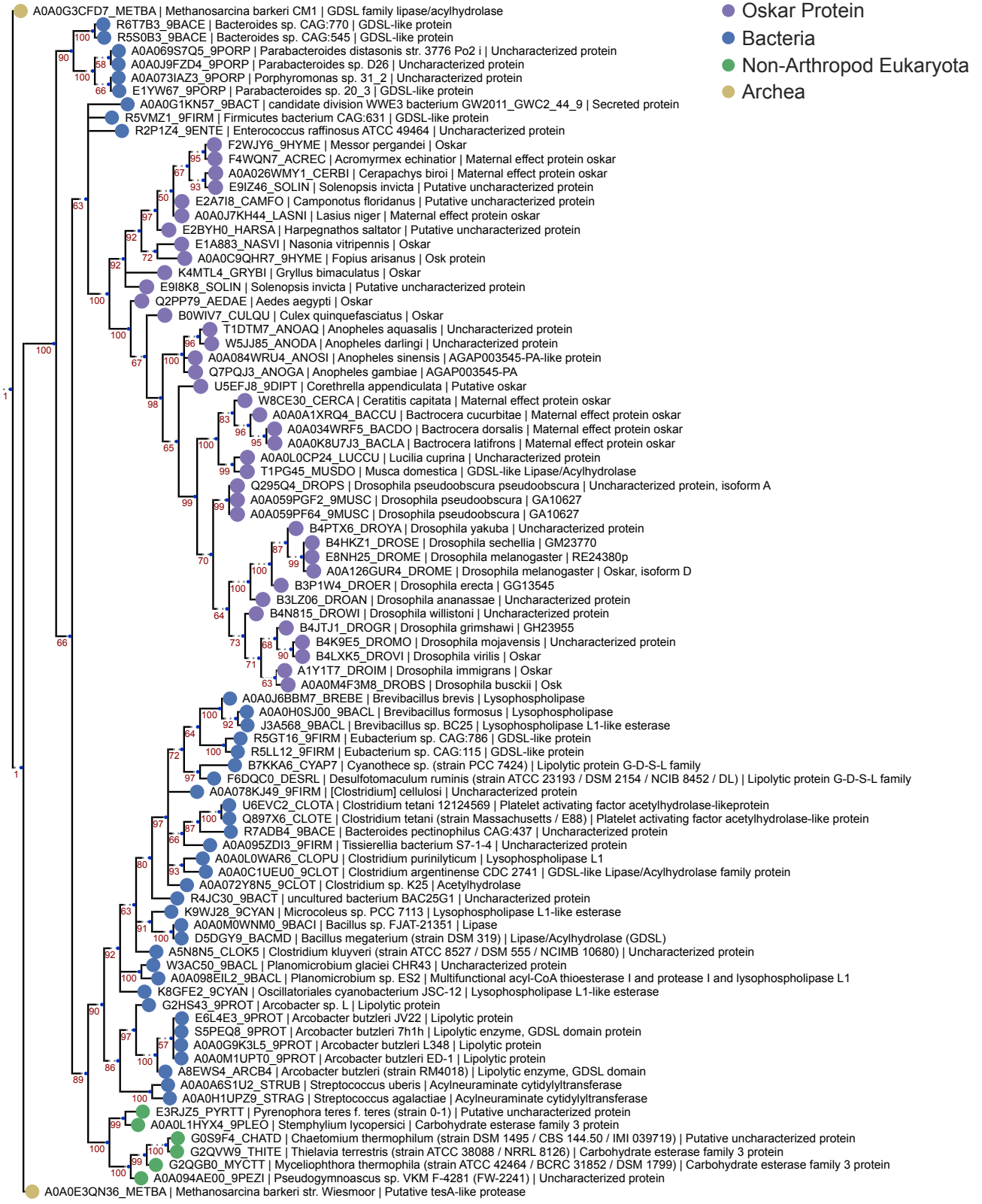
894 **Figure 2–figure supplement 12: OSK Tree T-Coffee Comparison.** Comparison of the tree
895 obtained with RaxML starting from the MUSCLE alignment (left) versus the T-Coffee alignment
896 (right) for the OSK domain. Similarity scores for the branching events are color coded from
897 yellow to blue (see figure color bar legend). The OSK (purple) clade and CAZ3 (green) clade
898 have been colored and compacted for readability as they do not have any internal branching
899 changes. Node color is blue if the leaf is a sequence, and red if this is a compacted group of
900 sequences.
901

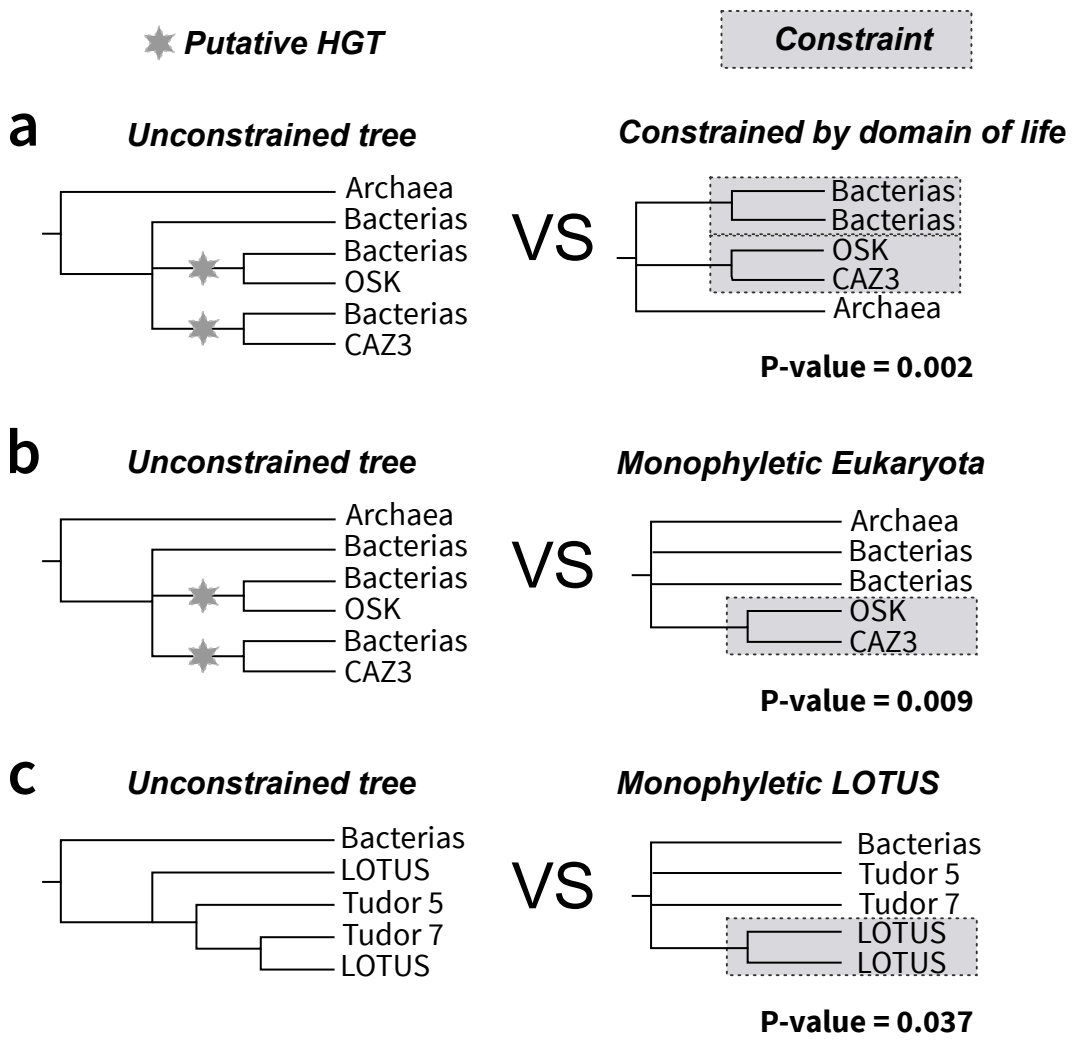
902 **Figure 2–figure supplement 13: LOTUS Tree T-Coffee Comparison.** Comparison of the tree
903 obtained with RaxML starting from the MUSCLE alignment (left) versus the T-Coffee alignment
904 (right) for the LOTUS domain. Similarity scores for the branching events are color coded from
905 yellow to blue (see figure color bar legend). The LOTUS (purple) clades have been colored and
906 compacted for readability as they do not have any internal branching changes. Node color is blue
907 if the leaf is a sequence, and red if this is a compacted group of sequences.
908
909
910

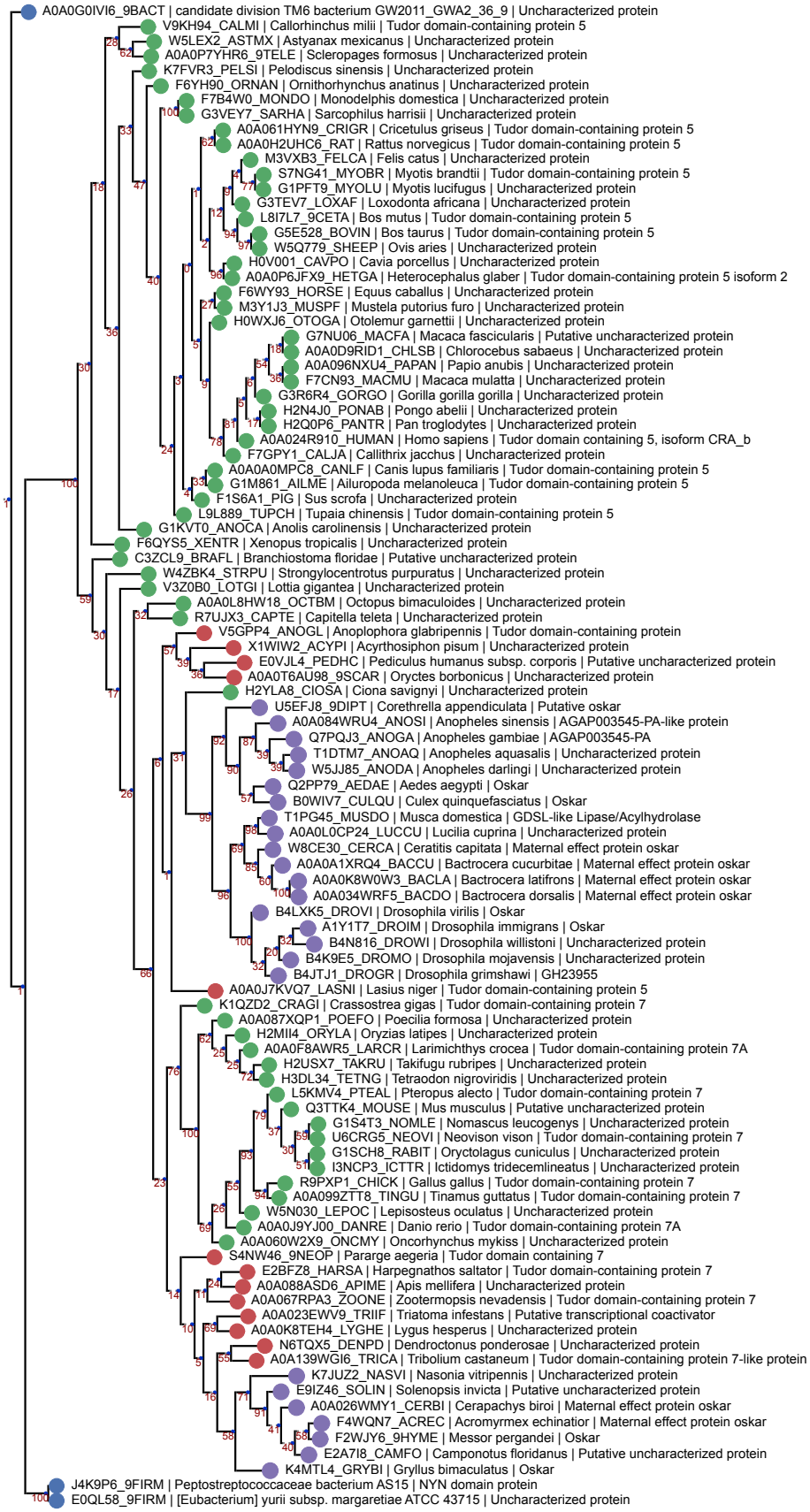

Oskar Protein
Bacteria
Non-Arthropod Eukaryota
Non-Oskar Arthropoda
Archea
A0A0G0IVI6_9BACT | candidate division TM6 bacterium GW2011_GWA2_36_9 | Uncharacterized protein
U5EFJ8_9DIPT | Corethrella appendiculata | Putative oskar
Q2PP79_AEDAE | Aedes aegypti | Oskar
B0WIV7_CULQU | Culex quinquefasciatus | Oskar
A0A084WRU4_ANOSI | Anopheles sinensis | AGAP003545-PA-like protein
Q7PQJ3_ANOGA | Anopheles gambiae | AGAP003545-PA
W5JJ85_ANODA | Anopheles darlingi | Uncharacterized protein
T1DTM7_ANOAQ | Anopheles aquasalis | Uncharacterized protein
W8CE30_CERCA | Ceratitis capitata | Maternal effect protein oskar
A0A0K8W0W3_BACLA | Bactrocera latifrons | Maternal effect protein oskar
A0A034WRF5_BACDO | Bactrocera dorsalis | Maternal effect protein oskar
A0A0A1XRQ4_BACCU | Bactrocera cucurbitae | Maternal effect protein oskar
B4LXK5_DROVI | Drosophila virilis | Oskar
B4JTJ1_DROGR | Drosophila grimshawi | GH23955
B4K9E5_DROMO | Drosophila mojavensis | Uncharacterized protein
A1Y1T7_DROIM | Drosophila immigrans | Oskar
B4N816_DROWI | Drosophila willistoni | Uncharacterized protein
A0A0L0CP24_LUCCU | Lucilia cuprina | Uncharacterized protein
T1PG45_MUSDO | Musca domestica | GDSL-like Lipase/Acylhydrolase
A0A0J7KVQ7_LASNI | Lasius niger | Tudor domain-containing protein 5
X1WIW2_ACYPI | Acyrthosiphon pisum | Uncharacterized protein
A0A0T6AU98_9SCAR | Oryctes borbonicus | Uncharacterized protein
E0VJL4_PEDHC | Pediculus humanus subsp. corporis | Putative uncharacterized protein
H2YLA8_CIOSA | Ciona savignyi | Uncharacterized protein
C3ZCL9_BRAFL | Branchiostoma floridae | Putative uncharacterized protein
F6QYS5_XENTR | Xenopus tropicalis | Uncharacterized protein
K7FVR3_PELSI | Pelodiscus sinensis | Uncharacterized protein
F6YH90_ORNAN | Ornithorhynchus anatinus | Uncharacterized protein
S7NG41_MYOBR | Myotis brandtii | Tudor domain-containing protein 5
H0WXJ6_OTOGA | Otolemur garnettii | Uncharacterized protein
M3VXB3_FELCA | Felis catus | Uncharacterized protein
M3Y1J3_MUSPF | Mustela putorius furo | Uncharacterized protein
L9L889_TUPCH | Tupaia chinensis | Tudor domain-containing protein 5
G1PFT9_MYOLU | Myotis lucifugus | Uncharacterized protein
G7NU06_MACFA | Macaca fascicularis | Putative uncharacterized protein
F7CN93_MACMU | Macaca mulatta | Uncharacterized protein
A0A096NXU4_PAPAN | Papio anubis | Uncharacterized protein
A0A0D9RID1_CHLSB | Chlorocebus sabaeus | Uncharacterized protein
G3R6R4_GORGO | Gorilla gorilla gorilla | Uncharacterized protein
H2N4J0_PONAB | Pongo abelii | Uncharacterized protein
A0A024R910_HUMAN | Homo sapiens | Tudor domain containing 5, isoform CRA_b
G1KVT0_ANOCA | Anolis carolinensis | Uncharacterized protein
H2Q0P6_PANTR | Pan troglodytes | Uncharacterized protein
F7GPY1_CALJA | Callithrix jacchus | Uncharacterized protein
A0A0A0MPC8_CANLF | Canis lupus familiaris | Tudor domain-containing protein 5
G1M861_AILME | Ailuropoda melanoleuca | Tudor domain-containing protein 5
F1S6A1_PIG | Sus scrofa | Uncharacterized protein
A0A0H2UHC6_RAT | Rattus norvegicus | Tudor domain-containing protein 5
A0A061HYN9_CRIGR | Cricetulus griseus | Tudor domain-containing protein 5
A0A0P6JFX9_HETGA | Heterocephalus glaber | Tudor domain-containing protein 5 isoform 2
H0V001_CAVPO | Cavia porcellus | Uncharacterized protein
G3TEV7_LOXAF | Loxodonta africana | Uncharacterized protein
G5E528_BOVIN | Bos taurus | Tudor domain-containing protein 5
W5Q779_SHEEP | Ovis aries | Uncharacterized protein
L8I7L7_9CETA | Bos mutus | Tudor domain-containing protein 5
F6WY93_HORSE | Equus caballus | Uncharacterized protein
F7B4W0_MONDO | Monodelphis domestica | Uncharacterized protein
G3VEY7_SARHA | Sarcophilus harrisii | Uncharacterized protein
A0A0P7YHR6_9TELE | Scleropages formosus | Uncharacterized protein
W5LEX2_ASTMX | Astyanax mexicanus | Uncharacterized protein
V9KH94_CALMI | Callorhinchus milii | Tudor domain-containing protein 5
A0A0L8HW18_OCTBM | Octopus bimaculoides | Uncharacterized protein
H2MII4_ORYLA | Oryzias latipes | Uncharacterized protein
H2USX7_TAKRU | Takifugu rubripes | Uncharacterized protein
A0A0F8AWR5_LARCR | Larimichthys crocea | Tudor domain-containing protein 7A
H3DL34_TETNG | Tetraodon nigroviridis | Uncharacterized protein
A0A087XQP1_POEFO | Poecilia formosa | Uncharacterized protein
A0A0J9YJ00_DANRE | Danio rerio | Tudor domain-containing protein 7A
W5N030_LEPOC | Lepisosteus oculatus | Uncharacterized protein
L5KMV4_PTEAL | Pteropus alecto | Tudor domain-containing protein 7
Q3TTK4_MOUSE | Mus musculus | Putative uncharacterized protein
G1S4T3_NOMLE | Nomascus leucogenys | Uncharacterized protein
U6CRG5_NEOVI | Neovison vison | Tudor domain-containing protein 7
I3NCP3_ICTTR | Ictidomys tridecemlineatus | Uncharacterized protein
G1SCH8_RABIT | Oryctolagus cuniculus | Uncharacterized protein
A0A099ZTT8_TINGU | Tinamus guttatus | Tudor domain-containing protein 7
R9PXP1_CHICK | Gallus gallus | Tudor domain-containing protein 7
A0A060W2X9_ONCMY | Oncorhynchus mykiss | Uncharacterized protein
K1QZD2_CRAGI | Crassostrea gigas | Tudor domain-containing protein 7
S4NW46_9NEOP | Pararge aegeria | Tudor domain containing 7
A0A139WGI6_TRICA | Tribolium castaneum | Tudor domain-containing protein 7-like protein
A0A067RPA3_ZOONE | Zootermopsis nevadensis | Tudor domain-containing protein 7
E2BFZ8_HARSA | Harpegnathos saltator | Tudor domain-containing protein 7
A0A088ASD6_APIME | Apis mellifera | Uncharacterized protein
A0A0K8TEH4_LYGHE | Lygus hesperus | Uncharacterized protein
A0A023EWV9_TRIIF | Triatoma infestans | Putative transcriptional coactivator
K4MTL4_GRYBI | Gryllus bimaculatus | Oskar
K7JUZ2_NASVI | Nasonia vitripennis | Uncharacterized protein
E9IZ46_SOLIN | Solenopsis invicta | Putative uncharacterized protein
A0A026WMY1_CERBI | Cerapachys biroi | Maternal effect protein oskar
E2A7I8_CAMFO | Camponotus floridanus | Putative uncharacterized protein
F4WQN7_ACREC | Acromyrmex echinatior | Maternal effect protein oskar
F2WJY6_9HYME | Messor pergandei | Oskar
V5GPP4_ANOGL | Anoplophora glabripennis | Tudor domain-containing protein
N6TQX5_DENPD | Dendroctonus ponderosae | Uncharacterized protein
R7UJX3_CAPTE | Capitella teleta | Uncharacterized protein
V3Z0B0_LOTGI | Lottia gigantea | Uncharacterized protein
W4ZBK4_STRPU | Strongylocentrotus purpuratus | Uncharacterized protein
E0QL58_9FIRM | [Eubacterium] yurii subsp. margaretiae ATCC 43715 | Uncharacterized protein
J4K9P6_9FIRM | Peptostreptococcaceae bacterium AS15 | NYN domain protein

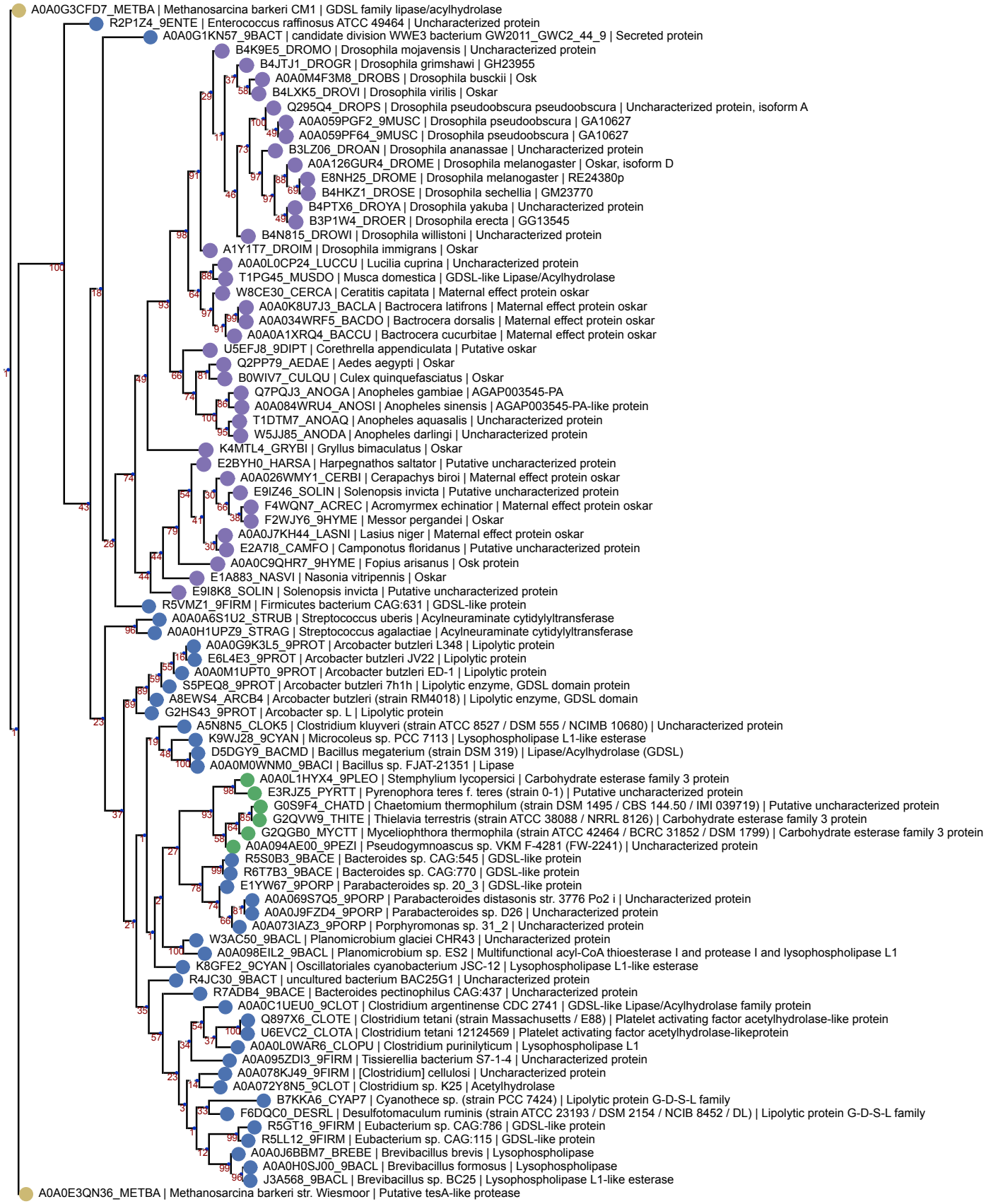

Oskar Protein
Bacteria
Non-Arthropod Eukaryota
Non-Oskar Arthropoda
Archea
A0A0G0IVI6_9BACT | candidate division TM6 bacterium GW2011_GWA2_36_9 | Uncharacterized protein
E0VJL4_PEDHC | Pediculus humanus subsp. corporis | Putative uncharacterized protein
A0A0T6AU98_9SCAR | Oryctes borbonicus | Uncharacterized protein
X1WIW2_ACYPI | Acyrthosiphon pisum | Uncharacterized protein
H2YLA8_CIOSA | Ciona savignyi | Uncharacterized protein
V5GPP4_ANOGL | Anoplophora glabripennis | Tudor domain-containing protein
E9IZ46_SOLIN | Solenopsis invicta | Putative uncharacterized protein
F4WQN7_ACREC | Acromyrmex echinatior | Maternal effect protein oskar
F2WJY6_9HYME | Messor pergandei | Oskar
A0A026WMY1_CERBI | Cerapachys biroi | Maternal effect protein oskar
E2A7I8_CAMFO | Camponotus floridanus | Putative uncharacterized protein
K7JUZ2_NASVI | Nasonia vitripennis | Uncharacterized protein
K4MTL4_GRYBI | Gryllus bimaculatus | Oskar
A0A139WGI6_TRICA | Tribolium castaneum | Tudor domain-containing protein 7-like protein
E2BFZ8_HARSA | Harpegnathos saltator | Tudor domain-containing protein 7
N6TQX5_DENPD | Dendroctonus ponderosae | Uncharacterized protein
A0A067RPA3_ZOONE | Zootermopsis nevadensis | Tudor domain-containing protein 7
A0A0J7KVQ7_LASNI | Lasius niger | Tudor domain-containing protein 5
S4NW46_9NEOP | Pararge aegeria | Tudor domain containing 7
A0A023EWV9_TRIIF | Triatoma infestans | Putative transcriptional coactivator
A0A088ASD6_APIME | Apis mellifera | Uncharacterized protein
A0A0K8TEH4_LYGHE | Lygus hesperus | Uncharacterized protein
K1QZD2_CRAGI | Crassostrea gigas | Tudor domain-containing protein 7
H2MII4_ORYLA | Oryzias latipes | Uncharacterized protein
H2USX7_TAKRU | Takifugu rubripes | Uncharacterized protein
A0A0F8AWR5_LARCR | Larimichthys crocea | Tudor domain-containing protein 7A
H3DL34_TETNG | Tetraodon nigroviridis | Uncharacterized protein
A0A087XQP1_POEFO | Poecilia formosa | Uncharacterized protein
W5N030_LEPOC | Lepisosteus oculatus | Uncharacterized protein
A0A099ZTT8_TINGU | Tinamus guttatus | Tudor domain-containing protein 7
R9PXP1_CHICK | Gallus gallus | Tudor domain-containing protein 7
G1S4T3_NOMLE | Nomascus leucogenys | Uncharacterized protein
U6CRG5_NEOVI | Neovison vison | Tudor domain-containing protein 7
L5KMV4_PTEAL | Pteropus alecto | Tudor domain-containing protein 7
I3NCP3_ICTTR | Ictidomys tridecemlineatus | Uncharacterized protein
G1SCH8_RABIT | Oryctolagus cuniculus | Uncharacterized protein
Q3TTK4_MOUSE | Mus musculus | Putative uncharacterized protein
A0A060W2X9_ONCMY | Oncorhynchus mykiss | Uncharacterized protein
A0A0J9YJ00_DANRE | Danio rerio | Tudor domain-containing protein 7A
V3Z0B0_LOTGI | Lottia gigantea | Uncharacterized protein
C3ZCL9_BRAFL | Branchiostoma floridae | Putative uncharacterized protein
A0A0L8HW18_OCTBM | Octopus bimaculoides | Uncharacterized protein
R7UJX3_CAPTE | Capitella teleta | Uncharacterized protein
B4N816_DROWI | Drosophila willistoni | Uncharacterized protein
B4K9E5_DROMO | Drosophila mojavensis | Uncharacterized protein
B4LXK5_DROVI | Drosophila virilis | Oskar
B4JTJ1_DROGR | Drosophila grimshawi | GH23955
A1Y1T7_DROIM | Drosophila immigrans | Oskar
W8CE30_CERCA | Ceratitis capitata | Maternal effect protein oskar
A0A034WRF5_BACDO | Bactrocera dorsalis | Maternal effect protein oskar
A0A0K8W0W3_BACLA | Bactrocera latifrons | Maternal effect protein oskar
A0A0A1XRQ4_BACCU | Bactrocera cucurbitae | Maternal effect protein oskar
T1PG45_MUSDO | Musca domestica | GDSL-like Lipase/Acylhydrolase
A0A0L0CP24_LUCCU | Lucilia cuprina | Uncharacterized protein
Q2PP79_AEDAE | Aedes aegypti | Oskar
B0WIV7_CULQU | Culex quinquefasciatus | Oskar
Q7PQJ3_ANOGA | Anopheles gambiae | AGAP003545-PA
A0A084WRU4_ANOSI | Anopheles sinensis | AGAP003545-PA-like protein
W5JJ85_ANODA | Anopheles darlingi | Uncharacterized protein
T1DTM7_ANOAQ | Anopheles aquasalis | Uncharacterized protein
U5EFJ8_9DIPT | Corethrella appendiculata | Putative oskar
W4ZBK4_STRPU | Strongylocentrotus purpuratus | Uncharacterized protein
F6YH90_ORNAN | Ornithorhynchus anatinus | Uncharacterized protein
G3TEV7_LOXAF | Loxodonta africana | Uncharacterized protein
L8I7L7_9CETA | Bos mutus | Tudor domain-containing protein 5
W5Q779_SHEEP | Ovis aries | Uncharacterized protein
G5E528_BOVIN | Bos taurus | Tudor domain-containing protein 5
F1S6A1_PIG | Sus scrofa | Uncharacterized protein
S7NG41_MYOBR | Myotis brandtii | Tudor domain-containing protein 5
G1PFT9_MYOLU | Myotis lucifugus | Uncharacterized protein
A0A096NXU4_PAPAN | Papio anubis | Uncharacterized protein
A0A0D9RID1_CHLSB | Chlorocebus sabaeus | Uncharacterized protein
F7CN93_MACMU | Macaca mulatta | Uncharacterized protein
G7NU06_MACFA | Macaca fascicularis | Putative uncharacterized protein
G3R6R4_GORGO | Gorilla gorilla gorilla | Uncharacterized protein
H2N4J0_PONAB | Pongo abelii | Uncharacterized protein
H2Q0P6_PANTR | Pan troglodytes | Uncharacterized protein
A0A024R910_HUMAN | Homo sapiens | Tudor domain containing 5, isoform CRA_b
F7GPY1_CALJA | Callithrix jacchus | Uncharacterized protein
L9L889_TUPCH | Tupaia chinensis | Tudor domain-containing protein 5
M3VXB3_FELCA | Felis catus | Uncharacterized protein
H0WXJ6_OTOGA | Otolemur garnettii | Uncharacterized protein
A0A0A0MPC8_CANLF | Canis lupus familiaris | Tudor domain-containing protein 5
G1M861_AILME | Ailuropoda melanoleuca | Tudor domain-containing protein 5
M3Y1J3_MUSPF | Mustela putorius furo | Uncharacterized protein
A0A0P6JFX9_HETGA | Heterocephalus glaber | Tudor domain-containing protein 5 isoform 2
H0V001_CAVPO | Cavia porcellus | Uncharacterized protein
A0A061HYN9_CRIGR | Cricetulus griseus | Tudor domain-containing protein 5
A0A0H2UHC6_RAT | Rattus norvegicus | Tudor domain-containing protein 5
F6WY93_HORSE | Equus caballus | Uncharacterized protein
G3VEY7_SARHA | Sarcophilus harrisii | Uncharacterized protein
F7B4W0_MONDO | Monodelphis domestica | Uncharacterized protein
V9KH94_CALMI | Callorhinchus milii | Tudor domain-containing protein 5
A0A0P7YHR6_9TELE | Scleropages formosus | Uncharacterized protein
W5LEX2_ASTMX | Astyanax mexicanus | Uncharacterized protein
K7FVR3_PELSI | Pelodiscus sinensis | Uncharacterized protein
G1KVT0_ANOCA | Anolis carolinensis | Uncharacterized protein
F6QYS5_XENTR | Xenopus tropicalis | Uncharacterized protein
J4K9P6_9FIRM | Peptostreptococcaceae bacterium AS15 | NYN domain protein
E0QL58_9FIRM | [Eubacterium] yurii subsp. margaretiae ATCC 43715 | Uncharacterized protein


Oskar Protein
Bacteria
Non-Arthropod Eukaryota
Archea
A0A0E3QN36_METBA | Methanosarcina barkeri str. Wiesmoor | Putative tesA-like protease
R5S0B3_9BACE | Bacteroides sp. CAG:545 | GDSL-like protein
R6T7B3_9BACE | Bacteroides sp. CAG:770 | GDSL-like protein
A0A069S7Q5_9PORP | Parabacteroides distasonis str. 3776 Po2 i | Uncharacterized protein
E1YW67_9PORP | Parabacteroides sp. 20_3 | GDSL-like protein
A0A073IAZ3_9PORP | Porphyromonas sp. 31_2 | Uncharacterized protein
A0A0J9FZD4_9PORP | Parabacteroides sp. D26 | Uncharacterized protein
K8GFE2_9CYAN | Oscillatoriales cyanobacterium JSC-12 | Lysophospholipase L1-like esterase
K9WJ28_9CYAN | Microcoleus sp. PCC 7113 | Lysophospholipase L1-like esterase
D5DGY9_BACMD | Bacillus megaterium (strain DSM 319) | Lipase/Acylhydrolase (GDSL)
A0A0M0WNM0_9BACI | Bacillus sp. FJAT-21351 | Lipase
A5N8N5_CLOK5 | Clostridium kluyveri (strain ATCC 8527 / DSM 555 / NCIMB 10680) | Uncharacterized protein
R4JC30_9BACT | uncultured bacterium BAC25G1 | Uncharacterized protein
R7ADB4_9BACE | Bacteroides pectinophilus CAG:437 | Uncharacterized protein
Q897X6_CLOTE | Clostridium tetani (strain Massachusetts / E88) | Platelet activating factor acetylhydrolase-like protein
U6EVC2_CLOTA | Clostridium tetani 12124569 | Platelet activating factor acetylhydrolase-likeprotein
A0A095ZDI3_9FIRM | Tissierella bacterium S7-1-4 | Uncharacterized protein
A0A0L0WAR6_CLOPU | Clostridium purinilyticum | Lysophospholipase L1
A0A0C1UEU0_9CLOT | Clostridium argentinense CDC 2741 | GDSL-like Lipase/Acylhydrolase family protein
R5LL12_9FIRM | Eubacterium sp. CAG:115 | GDSL-like protein
R5GT16_9FIRM | Eubacterium sp. CAG:786 | GDSL-like protein
B7KKA6_CYAP7 | Cyanothece sp. (strain PCC 7424) | Lipolytic protein G-D-S-L family
F6DQC0_DESRL | Desulfotomaculum ruminis (strain ATCC 23193 / DSM 2154 / NCIB 8452 / DL) | Lipolytic protein G-D-S-L family
A0A0J6BBM7_BREBE | Brevibacillus brevis | Lysophospholipase
J3A568_9BACL | Brevibacillus sp. BC25 | Lysophospholipase L1-like esterase
A0A0H0SJ00_9BACL | Brevibacillus formosus | Lysophospholipase
A0A078KJ49_9FIRM | [Clostridium] cellulosi | Uncharacterized protein
A0A072Y8N5_9CLOT | Clostridium sp. K25 | Acetylhydrolase
A0A098EIL2_9BACL | Planomicrobium sp. ES2 | Multifunctional acyl-CoA thioesterase I and protease I and lysophospholipase L1
W3AC50_9BACL | Planomicrobium glaciei CHR43 | Uncharacterized protein
A0A0H1UPZ9_STRAG | Streptococcus agalactiae | Acylneuraminate cytidylyltransferase
A0A0A6S1U2_STRUB | Streptococcus uberis | Acylneuraminate cytidylyltransferase
G2HS43_9PROT | Arcobacter sp. L | Lipolytic protein
A0A0G9K3L5_9PROT | Arcobacter butzleri L348 | Lipolytic protein
E6L4E3_9PROT | Arcobacter butzleri JV22 | Lipolytic protein
A0A0M1UPT0_9PROT | Arcobacter butzleri ED-1 | Lipolytic protein
S5PEQ8_9PROT | Arcobacter butzleri 7h1h | Lipolytic enzyme, GDSL domain protein
A8EWS4_ARCB4 | Arcobacter butzleri (strain RM4018) | Lipolytic enzyme, GDSL domain
A0A0L1HYX4_9PLEO | Stemphylium lycopersici | Carbohydrate esterase family 3 protein
E3RJZ5_PYRTT | Pyrenophora teres f. teres (strain 0-1) | Putative uncharacterized protein
G2QGB0_MYCTT | Myceliophthora thermophila (strain ATCC 42464 / BCRC 31852 / DSM 1799) | Carbohydrate esterase family 3 protein
G2QVW9_THITE | Thielavia terrestris (strain ATCC 38088 / NRRL 8126) | Carbohydrate esterase family 3 protein
G0S9F4_CHATD | Chaetomium thermophilum (strain DSM 1495 / CBS 144.50 / IMI 039719) | Putative uncharacterized protein
A0A094AE00_9PEZI | Pseudogymnoascus sp. VKM F-4281 (FW-2241) | Uncharacterized protein
R2P1Z4_9ENTE | Enterococcus raffinosus ATCC 49464 | Uncharacterized protein
R5VMZ1_9FIRM | Firmicutes bacterium CAG:631 | GDSL-like protein
A0A0G1KN57_9BACT | candidate division WWE3 bacterium GW2011_GWC2_44_9 | Secreted protein
K4MTL4_GRYBI | Gryllus bimaculatus | Oskar
E9I8K8_SOLIN | Solenopsis invicta | Putative uncharacterized protein
E2A7I8_CAMFO | Camponotus floridanus | Putative uncharacterized protein
E9IZ46_SOLIN | Solenopsis invicta | Putative uncharacterized protein
A0A026WMY1_CERBI | Cerapachys biroi | Maternal effect protein oskar
F2WJY6_9HYME | Messor pergandei | Oskar
F4WQN7_ACREC | Acromyrmex echinatior | Maternal effect protein oskar
E2BYH0_HARSA | Harpegnathos saltator | Putative uncharacterized protein
A0A0J7KH44_LASNI | Lasius niger | Maternal effect protein oskar
A0A0C9QHR7_9HYME | Fopius arisanus | Osk protein
E1A883_NASVI | Nasonia vitripennis | Oskar
Q2PP79_AEDAE | Aedes aegypti | Oskar
B0WIV7_CULQU | Culex quinquefasciatus | Oskar
Q7PQJ3_ANOGA | Anopheles gambiae | AGAP003545-PA
A0A084WRU4_ANOSI | Anopheles sinensis | AGAP003545-PA-like protein
W5JJ85_ANODA | Anopheles darlingi | Uncharacterized protein
T1DTM7_ANOAQ | Anopheles aquasalis | Uncharacterized protein
U5EFJ8_9DIPT | Corethrella appendiculata | Putative oskar
A0A059PF64_9MUSC | Drosophila pseudoobscura | GA10627
A0A059PGF2_9MUSC | Drosophila pseudoobscura | GA10627
Q295Q4_DROPS | Drosophila pseudoobscura pseudoobscura | Uncharacterized protein, isoform A
B3LZ06_DROAN | Drosophila ananassae | Uncharacterized protein
B3P1W4_DROER | Drosophila erecta | GG13545
B4PTX6_DROYA | Drosophila yakuba | Uncharacterized protein
A0A126GUR4_DROME | Drosophila melanogaster | Oskar, isoform D
E8NH25_DROME | Drosophila melanogaster | RE24380p
B4HKZ1_DROSE | Drosophila sechellia | GM23770
B4N815_DROWI | Drosophila willistoni | Uncharacterized protein
B4JTJ1_DROGR | Drosophila grimshawi | GH23955
B4LXK5_DROVI | Drosophila virilis | Oskar
B4K9E5_DROMO | Drosophila mojavensis | Uncharacterized protein
A0A0M4F3M8_DROBS | Drosophila busckii | Osk
A1Y1T7_DROIM | Drosophila immigrans | Oskar
A0A0L0CP24_LUCCU | Lucilia cuprina | Uncharacterized protein
T1PG45_MUSDO | Musca domestica | GDSL-like Lipase/Acylhydrolase
A0A0A1XRQ4_BACCU | Bactrocera cucurbitae | Maternal effect protein oskar
A0A034WRF5_BACDO | Bactrocera dorsalis | Maternal effect protein oskar
A0A0K8U7J3_BACLA | Bactrocera latifrons | Maternal effect protein oskar
W8CE30_CERCA | Ceratitis capitata | Maternal effect protein oskar
A0A0G3CFD7_METBA | Methanosarcina barkeri CM1 | GDSL family lipase/acylhydrolase

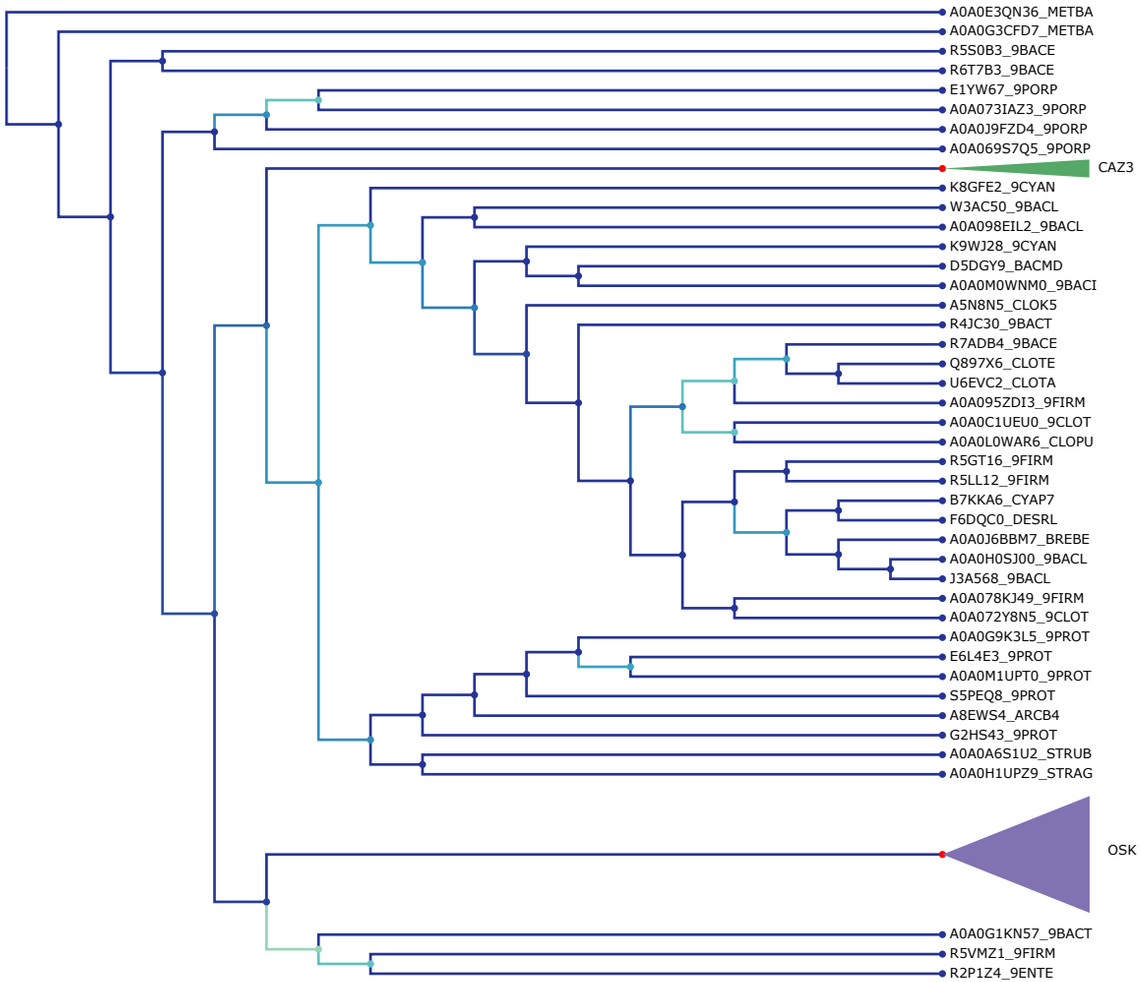

Oskar Protein
Bacteria
Non-Arthropod Eukaryota
Archea
A0A0G3CFD7_METBA | Methanosarcina barkeri CM1 | GDSL family lipase/acylhydrolase
R6T7B3_9BACE | Bacteroides sp. CAG:770 | GDSL-like protein
R5S0B3_9BACE | Bacteroides sp. CAG:545 | GDSL-like protein
A0A069S7Q5_9PORP | Parabacteroides distasonis str. 3776 Po2 i | Uncharacterized protein
A0A0J9FZD4_9PORP | Parabacteroides sp. D26 | Uncharacterized protein
A0A073IAZ3_9PORP | Porphyromonas sp. 31_2 | Uncharacterized protein
E1YW67_9PORP | Parabacteroides sp. 20_3 | GDSL-like protein
A0A0G1KN57_9BACT | candidate division WWE3 bacterium GW2011_GWC2_44_9 | Secreted protein
R5VMZ1_9FIRM | Firmicutes bacterium CAG:631 | GDSL-like protein
R2P1Z4_9ENTE | Enterococcus raffinosus ATCC 49464 | Uncharacterized protein
F2WJY6_9HYME | Messor pergandei | Oskar
F4WQN7_ACREC | Acromyrmex echinatior | Maternal effect protein oskar
A0A026WMY1_CERBI | Cerapachys biroi | Maternal effect protein oskar
E9IZ46_SOLIN | Solenopsis invicta | Putative uncharacterized protein
E2A7I8_CAMFO | Camponotus floridanus | Putative uncharacterized protein
A0A0J7KH44_LASNI | Lasius niger | Maternal effect protein oskar
E2BYH0_HARSA | Harpegnathos saltator | Putative uncharacterized protein
E1A883_NASVI | Nasonia vitripennis | Oskar
A0A0C9QHR7_9HYME | Fopius arisanus | Osk protein
K4MTL4_GRYBI | Gryllus bimaculatus | Oskar
E9I8K8_SOLIN | Solenopsis invicta | Putative uncharacterized protein
Q2PP79_AEDAE | Aedes aegypti | Oskar
B0WIV7_CULQU | Culex quinquefasciatus | Oskar
T1DTM7_ANOAQ | Anopheles aquasalis | Uncharacterized protein
W5JJ85_ANODA | Anopheles darlingi | Uncharacterized protein
A0A084WRU4_ANOSI | Anopheles sinensis | AGAP003545-PA-like protein
Q7PQJ3_ANOGA | Anopheles gambiae | AGAP003545-PA
U5EFJ8_9DIPT | Corethrella appendiculata | Putative oskar
W8CE30_CERCA | Ceratitis capitata | Maternal effect protein oskar
A0A0A1XRQ4_BACCU | Bactrocera cucurbitae | Maternal effect protein oskar
A0A034WRF5_BACDO | Bactrocera dorsalis | Maternal effect protein oskar
A0A0K8U7J3_BACLA | Bactrocera latifrons | Maternal effect protein oskar
A0A0L0CP24_LUCCU | Lucilia cuprina | Uncharacterized protein
T1PG45_MUSDO | Musca domestica | GDSL-like Lipase/Acylhydrolase
Q295Q4_DROPS | Drosophila pseudoobscura pseudoobscura | Uncharacterized protein, isoform A
A0A059PGF2_9MUSC | Drosophila pseudoobscura | GA10627
A0A059PF64_9MUSC | Drosophila pseudoobscura | GA10627
B4PTX6_DROYA | Drosophila yakuba | Uncharacterized protein
B4HKZ1_DROSE | Drosophila sechellia | GM23770
E8NH25_DROME | Drosophila melanogaster | RE24380p
A0A126GUR4_DROME | Drosophila melanogaster | Oskar, isoform D
B3P1W4_DROER | Drosophila erecta | GG13545
B3LZ06_DROAN | Drosophila ananassae | Uncharacterized protein
B4N815_DROWI | Drosophila willistoni | Uncharacterized protein
B4JTJ1_DROGR | Drosophila grimshawi | GH23955
B4K9E5_DROMO | Drosophila mojavensis | Uncharacterized protein
B4LXK5_DROVI | Drosophila virilis | Oskar
A1Y1T7_DROIM | Drosophila immigrans | Oskar
A0A0M4F3M8_DROBS | Drosophila busckii | Osk
A0A0J6BBM7_BREBE | Brevibacillus brevis | Lysophospholipase
A0A0H0SJ00_9BACL | Brevibacillus formosus | Lysophospholipase
J3A568_9BACL | Brevibacillus sp. BC25 | Lysophospholipase L1-like esterase
R5GT16_9FIRM | Eubacterium sp. CAG:786 | GDSL-like protein
R5LL12_9FIRM | Eubacterium sp. CAG:115 | GDSL-like protein
B7KKA6_CYAP7 | Cyanothece sp. (strain PCC 7424) | Lipolytic protein G-D-S-L family
F6DQC0_DESRL | Desulfotomaculum ruminis (strain ATCC 23193 / DSM 2154 / NCIB 8452 / DL) | Lipolytic protein G-D-S-L family
A0A078KJ49_9FIRM | [Clostridium] cellulosi | Uncharacterized protein
U6EVC2_CLOTA | Clostridium tetani 12124569 | Platelet activating factor acetylhydrolase-likeprotein
Q897X6_CLOTE | Clostridium tetani (strain Massachusetts / E88) | Platelet activating factor acetylhydrolase-like protein
R7ADB4_9BACE | Bacteroides pectinophilus CAG:437 | Uncharacterized protein
A0A095ZDI3_9FIRM | Tissierella bacterium S7-1-4 | Uncharacterized protein
A0A0L0WAR6_CLOPU | Clostridium purinilyticum | Lysophospholipase L1
A0A0C1UEU0_9CLOT | Clostridium argentinense CDC 2741 | GDSL-like Lipase/Acylhydrolase family protein
A0A072Y8N5_9CLOT | Clostridium sp. K25 | Acetylhydrolase
R4JC30_9BACT | uncultured bacterium BAC25G1 | Uncharacterized protein
K9WJ28_9CYAN | Microcoleus sp. PCC 7113 | Lysophospholipase L1-like esterase
A0A0M0WNM0_9BACI | Bacillus sp. FJAT-21351 | Lipase
D5DGY9_BACMD | Bacillus megaterium (strain DSM 319) | Lipase/Acylhydrolase (GDSL)
A5N8N5_CLOK5 | Clostridium kluyveri (strain ATCC 8527 / DSM 555 / NCIMB 10680) | Uncharacterized protein
W3AC50_9BACL | Planomicrobium glaciei CHR43 | Uncharacterized protein
A0A098EIL2_9BACL | Planomicrobium sp. ES2 | Multifunctional acyl-CoA thioesterase I and protease I and lysophospholipase L1
K8GFE2_9CYAN | Oscillatoriales cyanobacterium JSC-12 | Lysophospholipase L1-like esterase
G2HS43_9PROT | Arcobacter sp. L | Lipolytic protein
E6L4E3_9PROT | Arcobacter butzleri JV22 | Lipolytic protein
S5PEQ8_9PROT | Arcobacter butzleri 7h1h | Lipolytic enzyme, GDSL domain protein
A0A0G9K3L5_9PROT | Arcobacter butzleri L348 | Lipolytic protein
A0A0M1UPT0_9PROT | Arcobacter butzleri ED-1 | Lipolytic protein
A8EWS4_ARCB4 | Arcobacter butzleri (strain RM4018) | Lipolytic enzyme, GDSL domain
A0A0A6S1U2_STRUB | Streptococcus uberis | Acylneuraminate cytidylyltransferase
A0A0H1UPZ9_STRAG | Streptococcus agalactiae | Acylneuraminate cytidylyltransferase
E3RJZ5_PYRTT | Pyrenophora teres f. teres (strain 0-1) | Putative uncharacterized protein
A0A0L1HYX4_9PLEO | Stemphylium lycopersici | Carbohydrate esterase family 3 protein
G0S9F4_CHATD | Chaetomium thermophilum (strain DSM 1495 / CBS 144.50 / IMI 039719) | Putative uncharacterized protein
G2QVW9_THITE | Thielavia terrestris (strain ATCC 38088 / NRRL 8126) | Carbohydrate esterase family 3 protein
G2QGB0_MYCTT | Myceliophthora thermophila (strain ATCC 42464 / BCRC 31852 / DSM 1799) | Carbohydrate esterase family 3 protein
A0A094AE00_9PEZI | Pseudogymnoascus sp. VKM F-4281 (FW-2241) | Uncharacterized protein
A0A0E3QN36_METBA | Methanosarcina barkeri str. Wiesmoor | Putative tesA-like protease


Putative HGT
Constraint
a
Unconstrained tree
Archaea
Bacterias
Bacterias
OSK
Bacterias
CAZ3
VS
Constrained by domain of life
Bacterias
Bacterias
OSK
CAZ3
Archaea
P-value = 0.002
b
Unconstrained tree
Archaea
Bacterias
Bacterias
OSK
Bacterias
CAZ3
VS
Monophyletic Eukaryota
Archaea
Bacterias
Bacterias
OSK
CAZ3
P-value = 0.009
c
Unconstrained tree
Bacterias
LOTUS
Tudor 5
Tudor 7
LOTUS
VS
Monophyletic LOTUS
Bacterias
Tudor 5
Tudor 7
LOTUS
LOTUS
P-value = 0.037

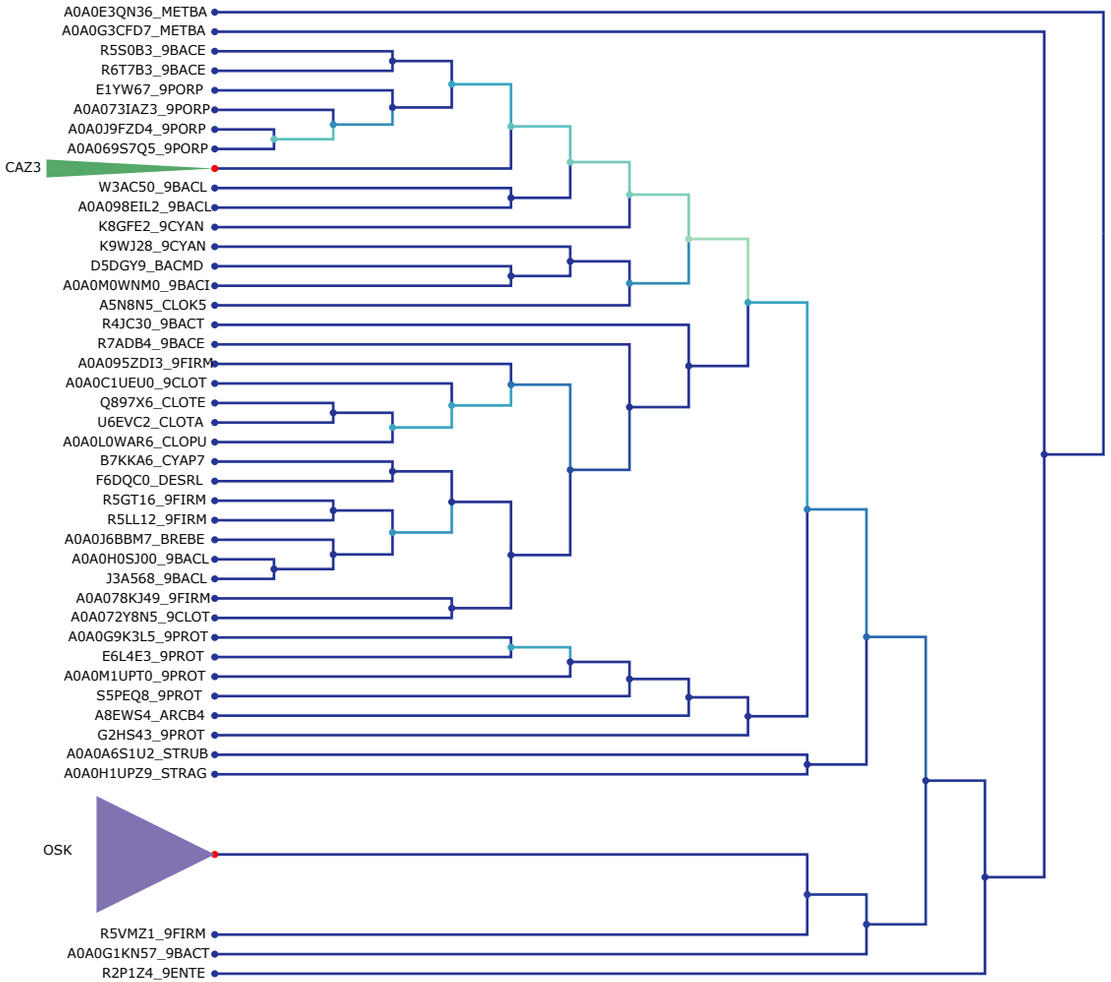

A0A0G0IVI6_9BACT | candidate division TM6 bacterium GW2011_GWA2_36_9 | Uncharacterized protein
V9KH94_CALMI | Callorhinchus milii | Tudor domain-containing protein 5
W5LEX2_ASTMX | Astyanax mexicanus | Uncharacterized protein
A0A0P7YHR6_9TELE | Scleropages formosus | Uncharacterized protein
K7FVR3_PELSI | Pelodiscus sinensis | Uncharacterized protein
F6YH90_ORNAN | Ornithorhynchus anatinus | Uncharacterized protein
F7B4W0_MONDO | Monodelphis domestica | Uncharacterized protein
G3VEY7_SARHA | Sarcophilus harrisii | Uncharacterized protein
A0A061HYN9_CRIGR | Cricetulus griseus | Tudor domain-containing protein 5
A0A0H2UHC6_RAT | Rattus norvegicus | Tudor domain-containing protein 5
M3VXB3_FELCA | Felis catus | Uncharacterized protein
S7NG41_MYOBR | Myotis brandtii | Tudor domain-containing protein 5
G1PFT9_MYOLU | Myotis lucifugus | Uncharacterized protein
G3TEV7_LOXAF | Loxodonta africana | Uncharacterized protein
L8I7L7_9CETA | Bos mutus | Tudor domain-containing protein 5
G5E528_BOVIN | Bos taurus | Tudor domain-containing protein 5
W5Q779_SHEEP | Ovis aries | Uncharacterized protein
H0V001_CAVPO | Cavia porcellus | Uncharacterized protein
A0A0P6JFX9_HETGA | Heterocephalus glaber | Tudor domain-containing protein 5 isoform 2
F6WY93_HORSE | Equus caballus | Uncharacterized protein
M3Y1J3_MUSPF | Mustela putorius furo | Uncharacterized protein
H0WXJ6_OTOGA | Otolemur garnettii | Uncharacterized protein
G7NU06_MACFA | Macaca fascicularis | Putative uncharacterized protein
A0A0D9RID1_CHLSB | Chlorocebus sabaeus | Uncharacterized protein
A0A096NXU4_PAPAN | Papio anubis | Uncharacterized protein
F7CN93_MACMU | Macaca mulatta | Uncharacterized protein
G3R6R4_GORGO | Gorilla gorilla gorilla | Uncharacterized protein
H2N4J0_PONAB | Pongo abelii | Uncharacterized protein
H2Q0P6_PANTR | Pan troglodytes | Uncharacterized protein
A0A024R910_HUMAN | Homo sapiens | Tudor domain containing 5, isoform CRA_b
F7GPY1_CALJA | Callithrix jacchus | Uncharacterized protein
A0A0A0MPC8_CANLF | Canis lupus familiaris | Tudor domain-containing protein 5
G1M861_AILME | Ailuropoda melanoleuca | Tudor domain-containing protein 5
F1S6A1_PIG | Sus scrofa | Uncharacterized protein
L9L889_TUPCH | Tupaia chinensis | Tudor domain-containing protein 5
G1KVT0_ANOCA | Anolis carolinensis | Uncharacterized protein
F6QYS5_XENTR | Xenopus tropicalis | Uncharacterized protein
C3ZCL9_BRAFL | Branchiostoma floridae | Putative uncharacterized protein
W4ZBK4_STRPU | Strongylocentrotus purpuratus | Uncharacterized protein
V3Z0B0_LOTGI | Lottia gigantea | Uncharacterized protein
A0A0L8HW18_OCTBM | Octopus bimaculoides | Uncharacterized protein
R7UJX3_CAPTE | Capitella teleta | Uncharacterized protein
V5GPP4_ANOGL | Anoplophora glabripennis | Tudor domain-containing protein
X1WIW2_ACYPI | Acyrthosiphon pisum | Uncharacterized protein
E0VJL4_PEDHC | Pediculus humanus subsp. corporis | Putative uncharacterized protein
A0A0T6AU98_9SCAR | Oryctes borbonicus | Uncharacterized protein
H2YLA8_CIOSA | Ciona savignyi | Uncharacterized protein
U5EFJ8_9DIPT | Corethrella appendiculata | Putative oskar
A0A084WRU4_ANOSI | Anopheles sinensis | AGAP003545-PA-like protein
Q7PQJ3_ANOGA | Anopheles gambiae | AGAP003545-PA
T1DTM7_ANOAQ | Anopheles aquasalis | Uncharacterized protein
W5JJ85_ANODA | Anopheles darlingi | Uncharacterized protein
Q2PP79_AEDAE | Aedes aegypti | Oskar
B0WIV7_CULQU | Culex quinquefasciatus | Oskar
T1PG45_MUSDO | Musca domestica | GDSL-like Lipase/Acylhydrolase
A0A0L0CP24_LUCCU | Lucilia cuprina | Uncharacterized protein
W8CE30_CERCA | Ceratitis capitata | Maternal effect protein oskar
A0A0A1XRQ4_BACCU | Bactrocera cucurbitae | Maternal effect protein oskar
A0A0K8W0W3_BACLA | Bactrocera latifrons | Maternal effect protein oskar
A0A034WRF5_BACDO | Bactrocera dorsalis | Maternal effect protein oskar
B4LXK5_DROVI | Drosophila virilis | Oskar
A1Y1T7_DROIM | Drosophila immigrans | Oskar
B4N816_DROWI | Drosophila willistoni | Uncharacterized protein
B4K9E5_DROMO | Drosophila mojavensis | Uncharacterized protein
B4JTJ1_DROGR | Drosophila grimshawi | GH23955
A0A0J7KVQ7_LASNI | Lasius niger | Tudor domain-containing protein 5
K1QZD2_CRAGI | Crassostrea gigas | Tudor domain-containing protein 7
A0A087XQP1_POEFO | Poecilia formosa | Uncharacterized protein
H2MII4_ORYLA | Oryzias latipes | Uncharacterized protein
A0A0F8AWR5_LARCR | Larimichthys crocea | Tudor domain-containing protein 7A
H2USX7_TAKRU | Takifugu rubripes | Uncharacterized protein
H3DL34_TETNG | Tetraodon nigroviridis | Uncharacterized protein
L5KMV4_PTEAL | Pteropus alecto | Tudor domain-containing protein 7
Q3TTK4_MOUSE | Mus musculus | Putative uncharacterized protein
G1S4T3_NOMLE | Nomascus leucogenys | Uncharacterized protein
U6CRG5_NEOVI | Neovison vison | Tudor domain-containing protein 7
G1SCH8_RABIT | Oryctolagus cuniculus | Uncharacterized protein
I3NCP3_ICTTR | Ictidomys tridecemlineatus | Uncharacterized protein
R9PXP1_CHICK | Gallus gallus | Tudor domain-containing protein 7
A0A099ZTT8_TINGU | Tinamus guttatus | Tudor domain-containing protein 7
W5N030_LEPOC | Lepisosteus oculatus | Uncharacterized protein
A0A0J9YJ00_DANRE | Danio rerio | Tudor domain-containing protein 7A
A0A060W2X9_ONCMY | Oncorhynchus mykiss | Uncharacterized protein
S4NW46_9NEOP | Parargе aegeria | Tudor domain containing 7
E2BFZ8_HARSA | Harpegnathos saltator | Tudor domain-containing protein 7
A0A088ASD6_APIME | Apis mellifera | Uncharacterized protein
A0A067RPA3_ZOONE | Zootermopsis nevadensis | Tudor domain-containing protein 7
A0A023EWV9_TRIIF | Triatoma infestans | Putative transcriptional coactivator
A0A0K8TEH4_LYGHE | Lygus hesperus | Uncharacterized protein
N6TQX5_DENPD | Dendroctonus ponderosae | Uncharacterized protein
A0A139WGI6_TRICA | Tribolium castaneum | Tudor domain-containing protein 7-like protein
K7JUZ2_NASVI | Nasonia vitripennis | Uncharacterized protein
E9IZ46_SOLIN | Solenopsis invicta | Putative uncharacterized protein
A0A026WMY1_CERBI | Cerapachys biroi | Maternal effect protein oskar
F4WQN7_ACREC | Acromyrmex echinatior | Maternal effect protein oskar
F2WJY6_9HYME | Messor pergandei | Oskar
E2A7I8_CAMFO | Camponotus floridanus | Putative uncharacterized protein
K4MTL4_GRYBI | Gryllus bimaculatus | Oskar
J4K9P6_9FIRM | Peptostreptococcaceae bacterium AS15 | NYN domain protein
E0QL58_9FIRM | [Eubacterium] yurii subsp. margaretiae ATCC 43715 | Uncharacterized protein

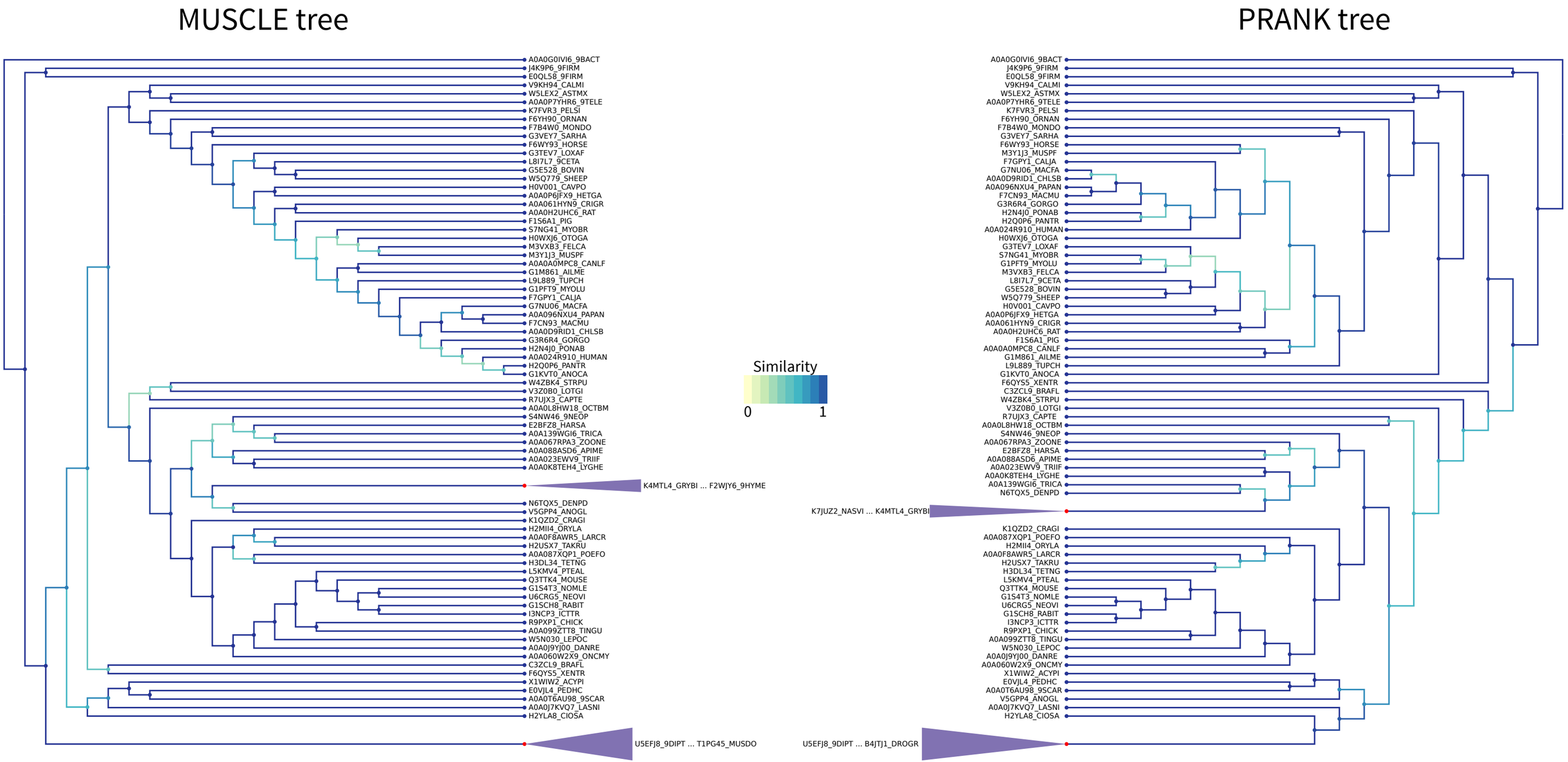

A0A0G3CFD7_METBA | Methanosarcina barkeri CM1 | GDSL family lipase/acylhydrolase
R2P1Z4_9ENTE | Enterococcus raffinosus ATCC 49464 | Uncharacterized protein
A0A0G1KN57_9BACT | candidate division WWE3 bacterium GW2011_GWC2_44_9 | Secreted protein
B4K9E5_DROMO | Drosophila mojavensis | Uncharacterized protein
B4JTJ1_DROGR | Drosophila grimshawi | GH23955
A0A0M4F3M8_DROBS | Drosophila busckii | Osk
B4LXK5_DROVI | Drosophila virilis | Oskar
Q295Q4_DROPS | Drosophila pseudoobscura pseudoobscura | Uncharacterized protein, isoform A
A0A059PGF2_9MUSC | Drosophila pseudoobscura | GA10627
A0A059PF64_9MUSC | Drosophila pseudoobscura | GA10627
B3LZ06_DROAN | Drosophila ananassae | Uncharacterized protein
A0A126GUR4_DROME | Drosophila melanogaster | Oskar, isoform D
E8NH25_DROME | Drosophila melanogaster | RE24380p
B4HKZ1_DROSE | Drosophila sechellia | GM23770
B4PTX6_DROYA | Drosophila yakuba | Uncharacterized protein
B3P1W4_DROER | Drosophila erecta | GG13545
B4N815_DROWI | Drosophila willistoni | Uncharacterized protein
A1Y1T7_DROIM | Drosophila immigrans | Oskar
A0A0L0CP24_LUCCU | Lucilia cuprina | Uncharacterized protein
T1PG45_MUSDO | Musca domestica | GDSL-like Lipase/Acylhydrolase
W8CE30_CERCA | Ceratitis capitata | Maternal effect protein oskar
A0A0K8U7J3_BACLA | Bactrocera latifrons | Maternal effect protein oskar
A0A034WRF5_BACDO | Bactrocera dorsalis | Maternal effect protein oskar
A0A0A1XRQ4_BACCU | Bactrocera cucurbitae | Maternal effect protein oskar
U5EFJ8_9DIPT | Corethrella appendiculata | Putative oskar
Q2PP79_AEDAE | Aedes aegypti | Oskar
B0WIV7_CULQU | Culex quinquefasciatus | Oskar
Q7PQJ3_ANOGA | Anopheles gambiae | AGAP003545-PA
A0A084WRU4_ANOSI | Anopheles sinensis | AGAP003545-PA-like protein
T1DTM7_ANOAQ | Anopheles aquasalis | Uncharacterized protein
W5JJ85_ANODA | Anopheles darlingi | Uncharacterized protein
K4MTL4_GRYBI | Gryllus bimaculatus | Oskar
E2BYH0_HARSA | Harpegnathos saltator | Putative uncharacterized protein
A0A026WMY1_CERBI | Cerapachys biroi | Maternal effect protein oskar
E9IZ46_SOLIN | Solenopsis invicta | Putative uncharacterized protein
F4WQN7_ACREC | Acromyrmex echinatior | Maternal effect protein oskar
F2WJY6_9HYME | Messor pergandei | Oskar
A0A0J7KH44_LASNI | Lasius niger | Maternal effect protein oskar
E2A7I8_CAMFO | Camponotus floridanus | Putative uncharacterized protein
A0A0C9QHR7_9HYME | Fopius arisanus | Osk protein
E1A883_NASVI | Nasonia vitripennis | Oskar
E9I8K8_SOLIN | Solenopsis invicta | Putative uncharacterized protein
R5VMZ1_9FIRM | Firmicutes bacterium CAG:631 | GDSL-like protein
A0A0A6S1U2_STRUB | Streptococcus uberis | Acylneuraminate cytidylyltransferase
A0A0H1UPZ9_STRAG | Streptococcus agalactiae | Acylneuraminate cytidylyltransferase
A0A0G9K3L5_9PROT | Arcobacter butzleri L348 | Lipolytic protein
E6L4E3_9PROT | Arcobacter butzleri JV22 | Lipolytic protein
A0A0M1UPT0_9PROT | Arcobacter butzleri ED-1 | Lipolytic protein
S5PEQ8_9PROT | Arcobacter butzleri 7h1h | Lipolytic enzyme, GDSL domain protein
A8EWS4_ARCB4 | Arcobacter butzleri (strain RM4018) | Lipolytic enzyme, GDSL domain
G2HS43_9PROT | Arcobacter sp. L | Lipolytic protein
A5N8N5_CLOK5 | Clostridium kluyveri (strain ATCC 8527 / DSM 555 / NCIMB 10680) | Uncharacterized protein
K9WJ28_9CYAN | Microcoleus sp. PCC 7113 | Lysophospholipase L1-like esterase
D5DGY9_BACMD | Bacillus megaterium (strain DSM 319) | Lipase/Acylhydrolase (GDSL)
A0A0M0WNM0_9BACI | Bacillus sp. FJAT-21351 | Lipase
A0A0L1HYX4_9PLEO | Stemphylium lycopersici | Carbohydrate esterase family 3 protein
E3RJZ5_PYRTT | Pyrenophora teres f. teres (strain 0-1) | Putative uncharacterized protein
G0S9F4_CHATD | Chaetomium thermophilum (strain DSM 1495 / CBS 144.50 / IMI 039719) | Putative uncharacterized protein
G2QVW9_THITE | Thielavia terrestris (strain ATCC 38088 / NRRL 8126) | Carbohydrate esterase family 3 protein
G2QGB0_MYCTT | Myceliophthora thermophila (strain ATCC 42464 / BCRC 31852 / DSM 1799) | Carbohydrate esterase family 3 protein
A0A094AE00_9PEZI | Pseudogymnoascus sp. VKM F-4281 (FW-2241) | Uncharacterized protein
R5S0B3_9BACE | Bacteroides sp. CAG:545 | GDSL-like protein
R6T7B3_9BACE | Bacteroides sp. CAG:770 | GDSL-like protein
E1YW67_9PORP | Parabacteroides sp. 20_3 | GDSL-like protein
A0A069S7Q5_9PORP | Parabacteroides distasonis str. 3776 Po2 i | Uncharacterized protein
A0A0J9FZD4_9PORP | Parabacteroides sp. D26 | Uncharacterized protein
A0A073IAZ3_9PORP | Porphyromonas sp. 31_2 | Uncharacterized protein
W3AC50_9BACL | Planomicrobium glaciei CHR43 | Uncharacterized protein
A0A098EIL2_9BACL | Planomicrobium sp. ES2 | Multifunctional acyl-CoA thioesterase I and protease I and lysophospholipase L1
K8GFE2_9CYAN | Oscillatoriales cyanobacterium JSC-12 | Lysophospholipase L1-like esterase
R4JC30_9BACT | uncultured bacterium BAC25G1 | Uncharacterized protein
R7ADB4_9BACE | Bacteroides pectinophilus CAG:437 | Uncharacterized protein
A0A0C1UEU0_9CLOT | Clostridium argentinense CDC 2741 | GDSL-like Lipase/Acylhydrolase family protein
Q897X6_CLOTE | Clostridium tetani (strain Massachusetts / E88) | Platelet activating factor acetylhydrolase-like protein
U6EVC2_CLOTA | Clostridium tetani 12124569 | Platelet activating factor acetylhydrolase-likeprotein
A0A0L0WAR6_CLOPU | Clostridium purinilyticum | Lysophospholipase L1
A0A095ZDI3_9FIRM | Tissierellia bacterium S7-1-4 | Uncharacterized protein
A0A078KJ49_9FIRM | [Clostridium] cellulosi | Uncharacterized protein
A0A072Y8N5_9CLOT | Clostridium sp. K25 | Acetylhydrolase
B7KKA6_CYAP7 | Cyanothece sp. (strain PCC 7424) | Lipolytic protein G-D-S-L family
F6DQC0_DESRL | Desulfotomaculum ruminis (strain ATCC 23193 / DSM 2154 / NCIB 8452 / DL) | Lipolytic protein G-D-S-L family
R5GT16_9FIRM | Eubacterium sp. CAG:786 | GDSL-like protein
R5LL12_9FIRM | Eubacterium sp. CAG:115 | GDSL-like protein
A0A0J6BBM7_BREBE | Brevibacillus brevis | Lysophospholipase
A0A0H0SJ00_9BACL | Brevibacillus formosus | Lysophospholipase
J3A568_9BACL | Brevibacillus sp. BC25 | Lysophospholipase L1-like esterase
A0A0E3QN36_METBA | Methanosarcina barkeri str. Wiesmoor | Putative tesA-like protease

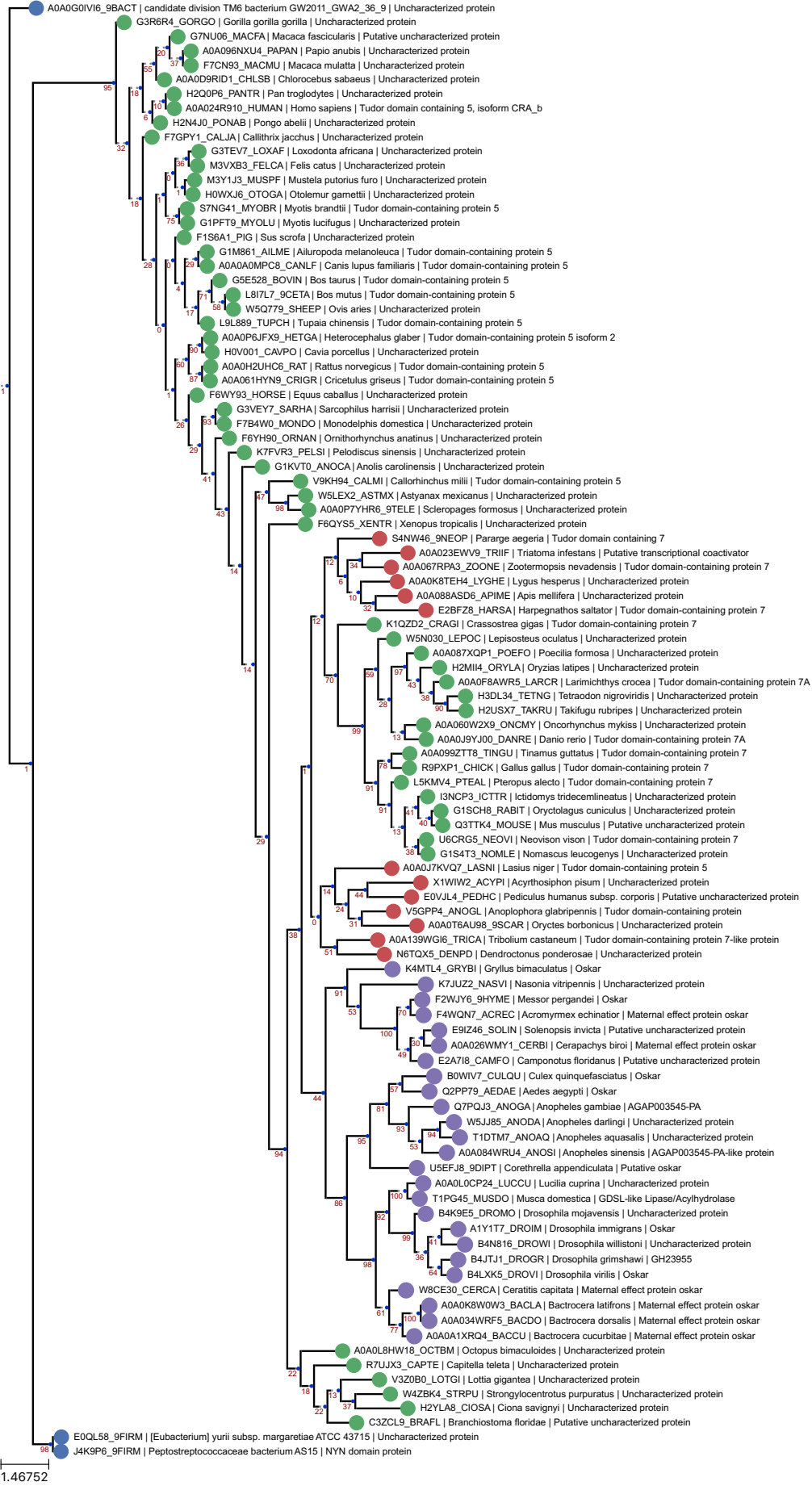
MUSCLE tree

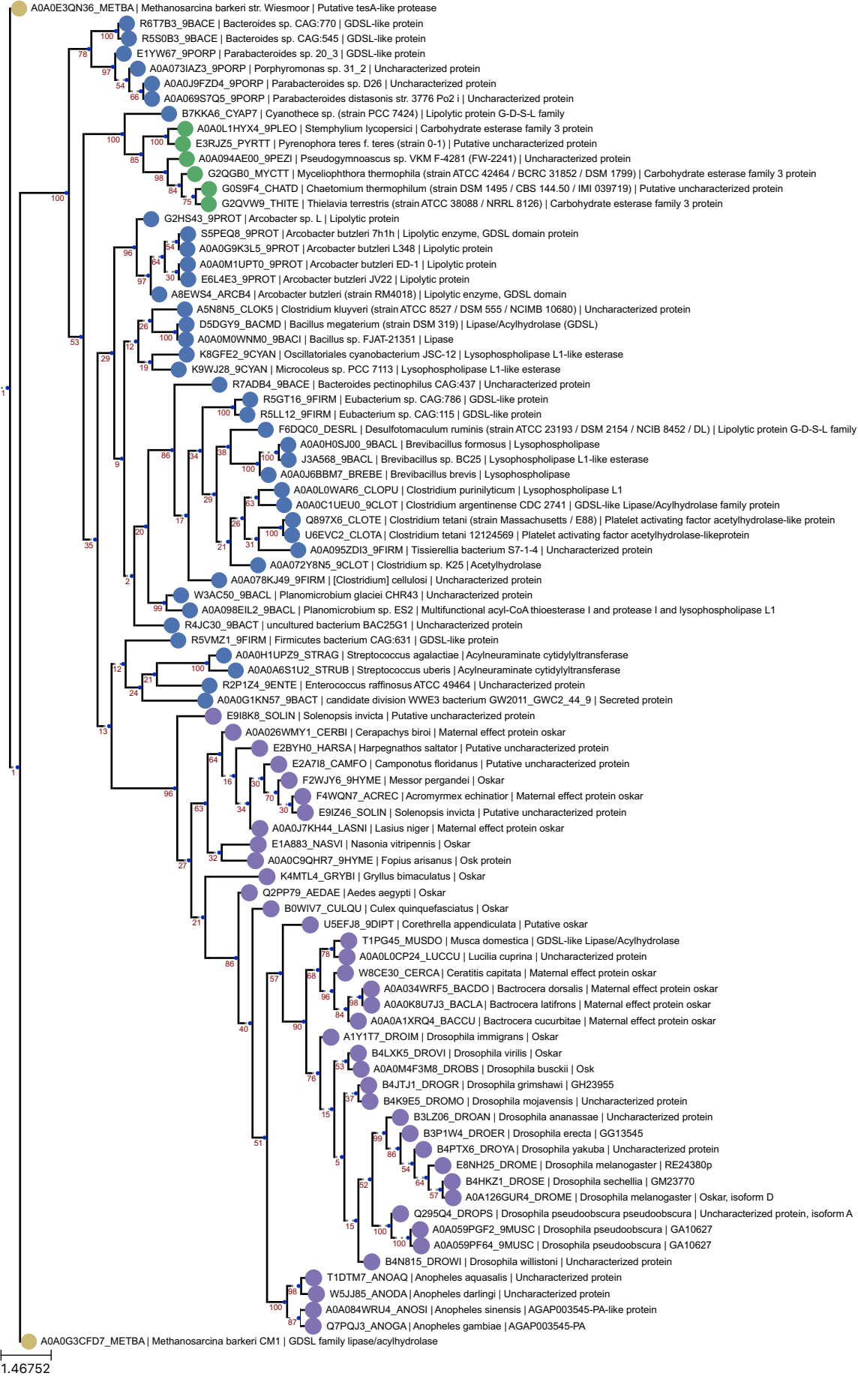


PRANK tree

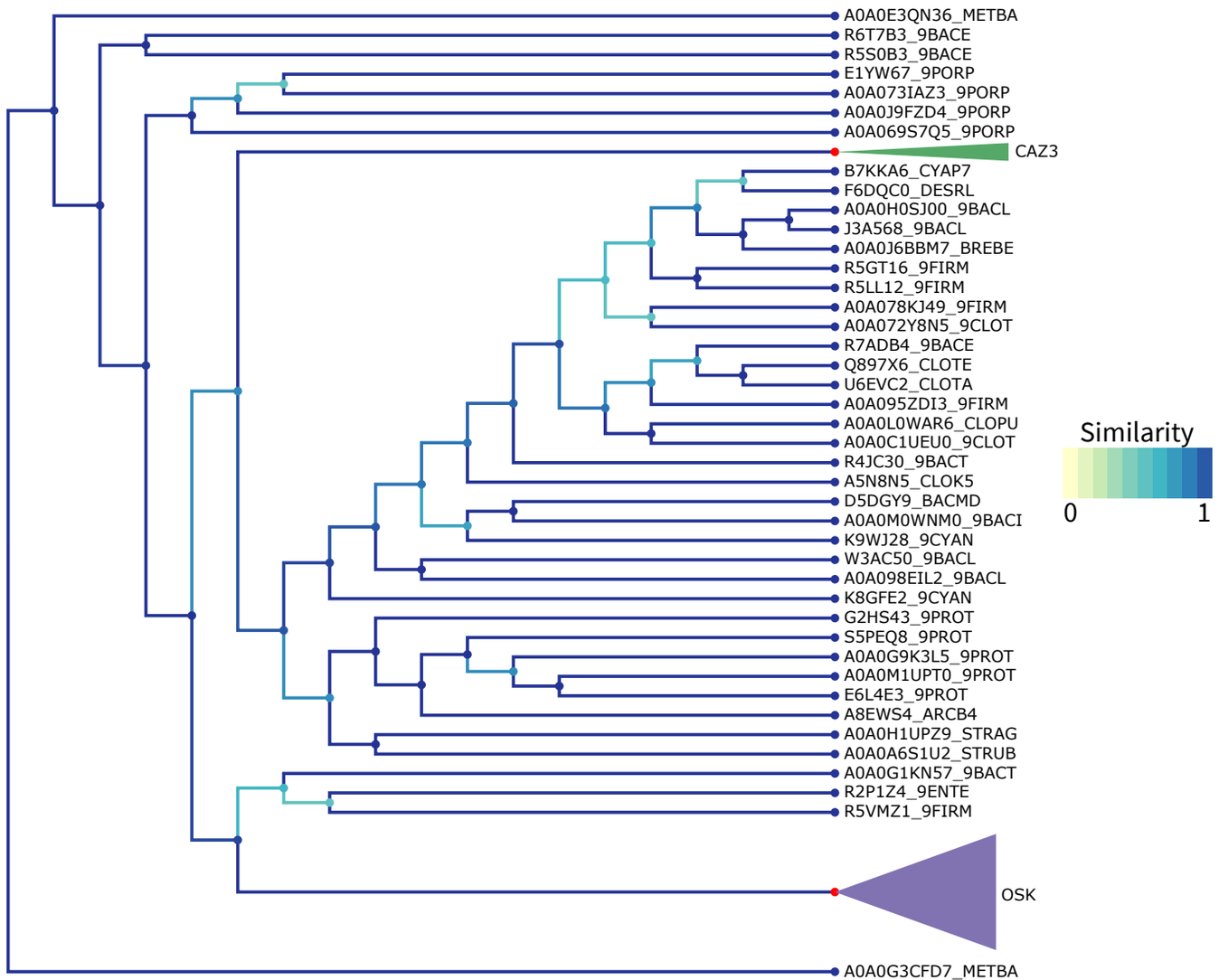



MUSCLE tree
PRANK tree
A0A0G0IVI6_9BACT
J4K9P6_9FIRM
E0QL58_9FIRM
V9KH94_CALMI
W5LEX2_ASTMX
A0A0P7YHR6_9TELE
K7FVR3_PELSI
F6YH90_ORNAN
F7B4W0_MONDO
G3VEY7_SARHA
F6WY93_HORSE
G3TEV7_LOXAF
L8I7L7_9CETA
G5E528_BOVIN
W5Q779_SHEEP
H0V001_CAVPO
A0A0P6JFX9_HETGA
A0A061HYN9_CRIGR
A0A0H2UHC6_RAT
F1S6A1_PIG
S7NG41_MYOBR
H0WXJ6_OTOGA
M3VXB3_FELCA
M3Y1J3_MUSPF
A0A0A0MPC8_CANLF
G1M861_AILME
L9L889_TUPCH
G1PFT9_MYOLU
F7GPY1_CALJA
G7NU06_MACFA
A0A096NXU4_PAPAN
F7CN93_MACMU
A0A0D9RID1_CHLSB
G3R6R4_GORGO
H2N4J0_PONAB
A0A024R910_HUMAN
H2Q0P6_PANTR
G1KVT0_ANOCA
W4ZBK4_STRPU
V3Z0B0_LOTGI
R7UJX3_CAPTE
A0A0L8HW18_OCTBM
S4NW46_9NEOP
E2BFZ8_HARSA
A0A139WGI6_TRICA
A0A067RPA3_ZOONE
A0A088ASD6_APIME
A0A023EWV9_TRIIF
A0A0K8TEH4_LYGHE
K4MTL4_GRYBI ... F2WJY6_9HYME
N6TQX5_DENPD
V5GPP4_ANOGL
K1QZD2_CRAGI
H2MII4_ORYLA
A0A0F8AWR5_LARCR
H2USX7_TAKRU
A0A087XQP1_POEFO
H3DL34_TETNG
L5KMV4_PTEAL
Q3TTK4_MOUSE
G1S4T3_NOMLE
U6CRG5_NEOVI
G1SCH8_RABIT
I3NCP3_ICTTR
R9PXP1_CHICK
A0A099ZTT8_TINGU
W5N030_LEPOC
A0A0J9YJ00_DANRE
A0A060W2X9_ONCMY
C3ZCL9_BRAFL
F6QYS5_XENTR
X1WIW2_ACYPI
E0VJL4_PEDHC
A0A0T6AU98_9SCAR
A0A0J7KVQ7_LASNI
H2YLA8_CIOSA
U5EFJ8_9DIPT ... T1PG45_MUSDO
Similarity
0
1
A0A0G0IVI6_9BACT
J4K9P6_9FIRM
E0QL58_9FIRM
V9KH94_CALMI
W5LEX2_ASTMX
A0A0P7YHR6_9TELE
K7FVR3_PELSI
F6YH90_ORNAN
F7B4W0_MONDO
G3VEY7_SARHA
F6WY93_HORSE
M3Y1J3_MUSPF
F7GPY1_CALJA
G7NU06_MACFA
A0A0D9RID1_CHLSB
A0A096NXU4_PAPAN
F7CN93_MACMU
G3R6R4_GORGO
H2N4J0_PONAB
H2Q0P6_PANTR
A0A024R910_HUMAN
H0WXJ6_OTOGA
G3TEV7_LOXAF
S7NG41_MYOBR
G1PFT9_MYOLU
M3VXB3_FELCA
L8I7L7_9CETA
G5E528_BOVIN
W5Q779_SHEEP
H0V001_CAVPO
A0A0P6JFX9_HETGA
A0A061HYN9_CRIGR
A0A0H2UHC6_RAT
F1S6A1_PIG
A0A0A0MPC8_CANLF
G1M861_AILME
L9L889_TUPCH
G1KVT0_ANOCA
F6QYS5_XENTR
C3ZCL9_BRAFL
W4ZBK4_STRPU
V3Z0B0_LOTGI
R7UJX3_CAPTE
A0A0L8HW18_OCTBM
S4NW46_9NEOP
A0A067RPA3_ZOONE
E2BFZ8_HARSA
A0A088ASD6_APIME
A0A023EWV9_TRIIF
A0A0K8TEH4_LYGHE
A0A139WGI6_TRICA
N6TQX5_DENPD
K7JUZ2_NASVI ... K4MTL4_GRYBI
K1QZD2_CRAGI
A0A087XQP1_POEFO
H2MII4_ORYLA
A0A0F8AWR5_LARCR
H2USX7_TAKRU
H3DL34_TETNG
L5KMV4_PTEAL
Q3TTK4_MOUSE
G1S4T3_NOMLE
U6CRG5_NEOVI
G1SCH8_RABIT
I3NCP3_ICTTR
R9PXP1_CHICK
A0A099ZTT8_TINGU
W5N030_LEPOC
A0A0J9YJ00_DANRE
A0A060W2X9_ONCMY
X1WIW2_ACYPI
E0VJL4_PEDHC
A0A0T6AU98_9SCAR
V5GPP4_ANOGL
A0A0J7KVQ7_LASNI
H2YLA8_CIOSA
U5EFJ8_9DIPT ... B4JTJ1_DROGR

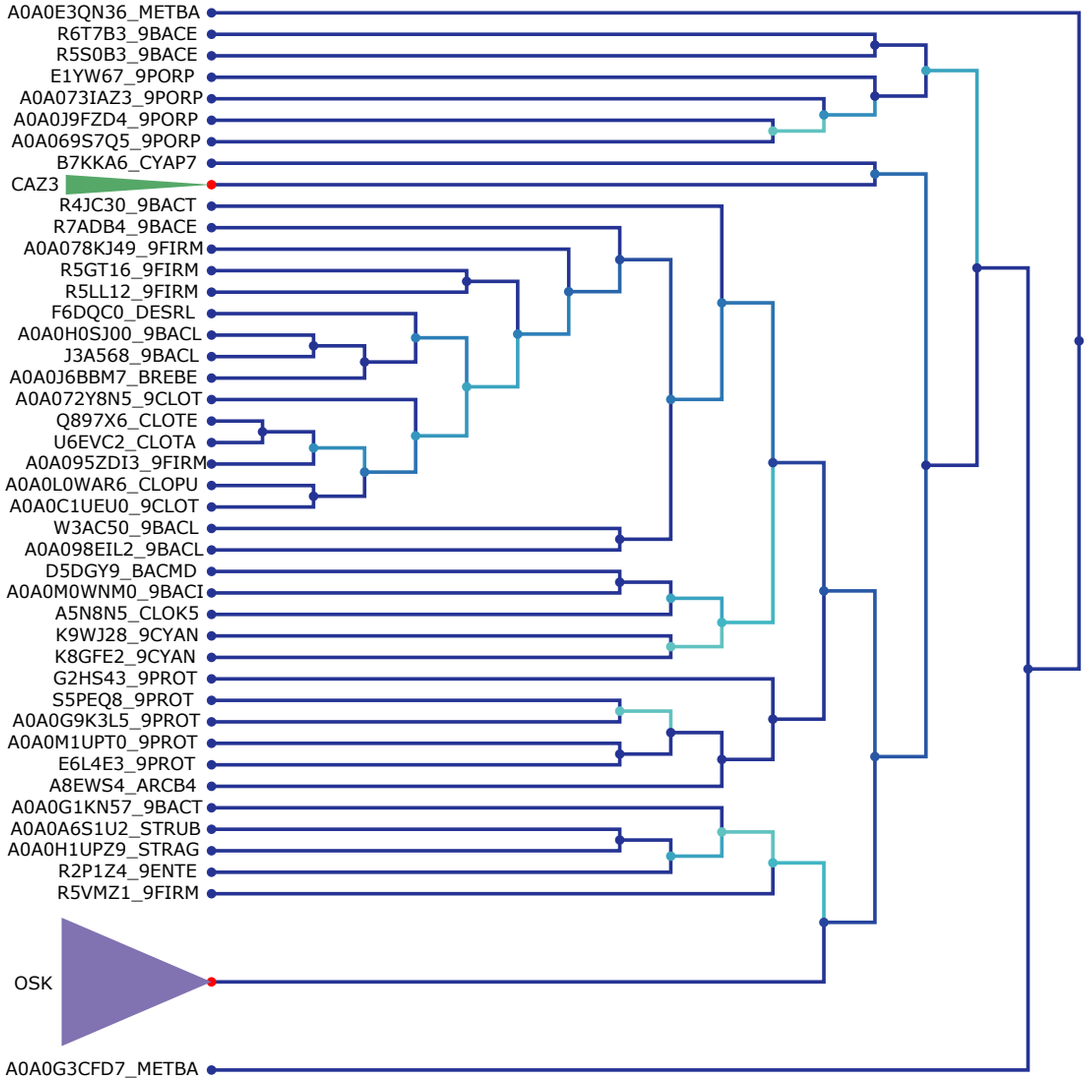

A0A0G0IVI6_9BACT | candidate division TM6 bacterium GW2011_GWA2_36_9 | Uncharacterized protein
G3R6R4_GORGO | Gorilla gorilla gorilla | Uncharacterized protein
G7NU06_MACFA | Macaca fascicularis | Putative uncharacterized protein
A0A096NXU4_PAPAN | Papio anubis | Uncharacterized protein
F7CN93_MACMU | Macaca mulatta | Uncharacterized protein
A0A0D9RID1_CHLSB | Chlorocebus sabaeus | Uncharacterized protein
H2Q0P6_PANTR | Pan troglodytes | Uncharacterized protein
A0A024R910_HUMAN | Homo sapiens | Tudor domain containing 5, isoform CRA_b
H2N4J0_PONAB | Pongo abelii | Uncharacterized protein
F7GPY1_CALJA | Callithrix jacchus | Uncharacterized protein
G3TEV7_LOXAF | Loxodonta africana | Uncharacterized protein
M3VXB3_FELCA | Felis catus | Uncharacterized protein
M3Y1J3_MUSPF | Mustela putorius furo | Uncharacterized protein
H0WXJ6_OTOGA | Otolemur garnettii | Uncharacterized protein
S7NG41_MYOBR | Myotis brandtii | Tudor domain-containing protein 5
G1PFT9_MYOLU | Myotis lucifugus | Uncharacterized protein
F1S6A1_PIG | Sus scrofa | Uncharacterized protein
G1M861_AILME | Ailuropoda melanoleuca | Tudor domain-containing protein 5
A0A0A0MPC8_CANLF | Canis lupus familiaris | Tudor domain-containing protein 5
G5E528_BOVIN | Bos taurus | Tudor domain-containing protein 5
L8I7L7_9CETA | Bos mutus | Tudor domain-containing protein 5
W5Q779_SHEEP | Ovis aries | Uncharacterized protein
L9L889_TUPCH | Tupaia chinensis | Tudor domain-containing protein 5
A0A0P6JFX9_HETGA | Heterocephalus glaber | Tudor domain-containing protein 5 isoform 2
H0V001_CAVPO | Cavia porcellus | Uncharacterized protein
A0A0H2UHC6_RAT | Rattus norvegicus | Tudor domain-containing protein 5
A0A061HYN9_CRIGR | Cricetulus griseus | Tudor domain-containing protein 5
F6WY93_HORSE | Equus caballus | Uncharacterized protein
G3VEY7_SARHA | Sarcophilus harrisii | Uncharacterized protein
F7B4W0_MONDO | Monodelphis domestica | Uncharacterized protein
F6YH90_ORNAN | Ornithorhynchus anatinus | Uncharacterized protein
K7FVR3_PELSI | Pelodiscus sinensis | Uncharacterized protein
G1KVT0_ANOCA | Anolis carolinensis | Uncharacterized protein
V9KH94_CALMI | Callorhinchus milii | Tudor domain-containing protein 5
W5LEX2_ASTMX | Astyanax mexicanus | Uncharacterized protein
A0A0P7YHR6_9TELE | Scleropages formosus | Uncharacterized protein
F6QYS5_XENTR | Xenopus tropicalis | Uncharacterized protein
S4NW46_9NEOP | Pararge aegeria | Tudor domain containing 7
A0A023EWV9_TRIIF | Triatoma infestans | Putative transcriptional coactivator
A0A067RPA3_ZOONE | Zootermopsis nevadensis | Tudor domain-containing protein 7
A0A0K8TEH4_LYGHE | Lygus hesperus | Uncharacterized protein
A0A088ASD6_APIME | Apis mellifera | Uncharacterized protein
E2BFZ8_HARSA | Harpegnathos saltator | Tudor domain-containing protein 7
K1QZD2_CRAGI | Crassostrea gigas | Tudor domain-containing protein 7
W5N030_LEPOC | Lepisosteus oculatus | Uncharacterized protein
A0A087XQP1_POEFO | Poecilia formosa | Uncharacterized protein
H2MII4_ORYLA | Oryzias latipes | Uncharacterized protein
A0A0F8AWR5_LARCR | Larimichthys crocea | Tudor domain-containing protein 7A
H3DL34_TETNG | Tetraodon nigroviridis | Uncharacterized protein
H2USX7_TAKRU | Takifugu rubripes | Uncharacterized protein
A0A060W2X9_ONCMY | Oncorhynchus mykiss | Uncharacterized protein
A0A0J9YJ00_DANRE | Danio rerio | Tudor domain-containing protein 7A
A0A099ZTT8_TINGU | Tinamus guttatus | Tudor domain-containing protein 7
R9PXP1_CHICK | Gallus gallus | Tudor domain-containing protein 7
L5KMV4_PTEAL | Pteropus alecto | Tudor domain-containing protein 7
I3NCP3_ICTTR | Ictidomys tridecemlineatus | Uncharacterized protein
G1SCH8_RABIT | Oryctolagus cuniculus | Uncharacterized protein
Q3TTK4_MOUSE | Mus musculus | Putative uncharacterized protein
U6CRG5_NEOVI | Neovison vison | Tudor domain-containing protein 7
G1S4T3_NOMLE | Nomascus leucogenys | Uncharacterized protein
A0A0J7KVQ7_LASNI | Lasius niger | Tudor domain-containing protein 5
X1WIW2_ACYPI | Acyrthosiphon pisum | Uncharacterized protein
E0VJL4_PEDHC | Pediculus humanus subsp. corporis | Putative uncharacterized protein
V5GPP4_ANOGL | Anoplophora glabripennis | Tudor domain-containing protein
A0A0T6AU98_9SCAR | Oryctes borbonicus | Uncharacterized protein
A0A139WGI6_TRICA | Tribolium castaneum | Tudor domain-containing protein 7-like protein
N6TQX5_DENPD | Dendroctonus ponderosae | Uncharacterized protein
K4MTL4_GRYBI | Gryllus bimaculatus | Oskar
K7JUZ2_NASVI | Nasonia vitripennis | Uncharacterized protein
F2WJY6_9HYME | Messor pergandei | Oskar
F4WQN7_ACREC | Acromyrmex echinatior | Maternal effect protein oskar
E9IZ46_SOLIN | Solenopsis invicta | Putative uncharacterized protein
A0A026WMY1_CERBI | Cerapachys biroi | Maternal effect protein oskar
E2A7I8_CAMFO | Camponotus floridanus | Putative uncharacterized protein
B0WIV7_CULQU | Culex quinquefasciatus | Oskar
Q2PP79_AEDAE | Aedes aegypti | Oskar
Q7PQJ3_ANOGA | Anopheles gambiae | AGAP003545-PA
W5JJ85_ANODA | Anopheles darlingi | Uncharacterized protein
T1DTM7_ANOAQ | Anopheles aquasalis | Uncharacterized protein
A0A084WRU4_ANOSI | Anopheles sinensis | AGAP003545-PA-like protein
U5EFJ8_9DIPT | Corethrella appendiculata | Putative oskar
A0A0L0CP24_LUCCU | Lucilia cuprina | Uncharacterized protein
T1PG45_MUSDO | Musca domestica | GDSL-like Lipase/Acylhydrolase
B4K9E5_DROMO | Drosophila mojavensis | Uncharacterized protein
A1Y1T7_DROIM | Drosophila immigrans | Oskar
B4N816_DROWI | Drosophila willistoni | Uncharacterized protein
B4JTJ1_DROGR | Drosophila grimshawi | GH23955
B4LXK5_DROVI | Drosophila virilis | Oskar
W8CE30_CERCA | Ceratitis capitata | Maternal effect protein oskar
A0A0K8W0W3_BACLA | Bactrocera latifrons | Maternal effect protein oskar
A0A034WRF5_BACDO | Bactrocera dorsalis | Maternal effect protein oskar
A0A0A1XRQ4_BACCU | Bactrocera cucurbitae | Maternal effect protein oskar
A0A0L8HW18_OCTBM | Octopus bimaculoides | Uncharacterized protein
R7UJX3_CAPTE | Capitella teleta | Uncharacterized protein
V3Z0B0_LOTGI | Lottia gigantea | Uncharacterized protein
W4ZBK4_STRPU | Strongylocentrotus purpuratus | Uncharacterized protein
H2YLA8_CIOSA | Ciona savignyi | Uncharacterized protein
C3ZCL9_BRAFL | Branchiostoma floridae | Putative uncharacterized protein
E0QL58_9FIRM | [Eubacterium] yurii subsp. margaretiae ATCC 43715 | Uncharacterized protein
J4K9P6_9FIRM | Peptostreptococcaceae bacterium AS15 | NYN domain protein
1.46752


A0A0E3QN36_METBA | Methanosarcina barkeri str. Wiesmoor | Putative tesA-like protease
R6T7B3_9BACE | Bacteroides sp. CAG:770 | GDSL-like protein
R5S0B3_9BACE | Bacteroides sp. CAG:545 | GDSL-like protein
E1YW67_9PORP | Parabacteroides sp. 20_3 | GDSL-like protein
A0A073IAZ3_9PORP | Porphyromonas sp. 31_2 | Uncharacterized protein
A0A0J9FZD4_9PORP | Parabacteroides sp. D26 | Uncharacterized protein
A0A069S7Q5_9PORP | Parabacteroides distasonis str. 3776 Po2 i | Uncharacterized protein
B7KKA6_CYAP7 | Cyanothece sp. (strain PCC 7424) | Lipolytic protein G-D-S-L family
A0A0L1HYX4_9PLEO | Stemphylium lycopersici | Carbohydrate esterase family 3 protein
E3RJZ5_PYRTT | Pyrenophora teres f. teres (strain 0-1) | Putative uncharacterized protein
A0A094AE00_9PEZI | Pseudogymnoascus sp. VKM F-4281 (FW-2241) | Uncharacterized protein
G2QGB0_MYCTT | Myceliophthora thermophila (strain ATCC 42464 / BCRC 31852 / DSM 1799) | Carbohydrate esterase family 3 protein
G0S9F4_CHATD | Chaetomium thermophilum (strain DSM 1495 / CBS 144.50 / IMI 039719) | Putative uncharacterized protein
G2QVW9_THITE | Thielavia terrestris (strain ATCC 38088 / NRRL 8126) | Carbohydrate esterase family 3 protein
G2HS43_9PROT | Arcobacter sp. L | Lipolytic protein
S5PEQ8_9PROT | Arcobacter butzleri 7h1h | Lipolytic enzyme, GDSL domain protein
A0A0G9K3L5_9PROT | Arcobacter butzleri L348 | Lipolytic protein
A0A0M1UPT0_9PROT | Arcobacter butzleri ED-1 | Lipolytic protein
E6L4E3_9PROT | Arcobacter butzleri JV22 | Lipolytic protein
A8EWS4_ARCB4 | Arcobacter butzleri (strain RM4018) | Lipolytic enzyme, GDSL domain
A5N8N5_CLOK5 | Clostridium kluyveri (strain ATCC 8527 / DSM 555 / NCIMB 10680) | Uncharacterized protein
D5DGY9_BACMD | Bacillus megaterium (strain DSM 319) | Lipase/Acylhydrolase (GDSL)
A0A0M0WNM0_9BACI | Bacillus sp. FJAT-21351 | Lipase
K8GFE2_9CYAN | Oscillatoriales cyanobacterium JSC-12 | Lysophospholipase L1-like esterase
K9WJ28_9CYAN | Microcoleus sp. PCC 7113 | Lysophospholipase L1-like esterase
R7ADB4_9BACE | Bacteroides pectinophilus CAG:437 | Uncharacterized protein
R5GT16_9FIRM | Eubacterium sp. CAG:786 | GDSL-like protein
R5LL12_9FIRM | Eubacterium sp. CAG:115 | GDSL-like protein
F6DQC0_DESRL | Desulfotomaculum ruminis (strain ATCC 23193 / DSM 2154 / NCIB 8452 / DL) | Lipolytic protein G-D-S-L family
A0A0H0SJ00_9BACL | Brevibacillus formosus | Lysophospholipase
J3A568_9BACL | Brevibacillus sp. BC25 | Lysophospholipase L1-like esterase
A0A0J6BBM7_BREBE | Brevibacillus brevis | Lysophospholipase
A0A0L0WAR6_CLOPU | Clostridium purinilyticum | Lysophospholipase L1
A0A0C1UEU0_9CLOT | Clostridium argentinense CDC 2741 | GDSL-like Lipase/Acylhydrolase family protein
Q897X6_CLOTE | Clostridium tetani (strain Massachusetts / E88) | Platelet activating factor acetylhydrolase-like protein
U6EVC2_CLOTA | Clostridium tetani 12124569 | Platelet activating factor acetylhydrolase-likeprotein
A0A095ZDI3_9FIRM | Tissierellia bacterium S7-1-4 | Uncharacterized protein
A0A072Y8N5_9CLOT | Clostridium sp. K25 | Acetylhydrolase
A0A078KJ49_9FIRM | [Clostridium] cellulosi | Uncharacterized protein
W3AC50_9BACL | Planomicrobium glaciei CHR43 | Uncharacterized protein
A0A098EIL2_9BACL | Planomicrobium sp. ES2 | Multifunctional acyl-CoA thioesterase I and protease I and lysophospholipase L1
R4JC30_9BACT | uncultured bacterium BAC25G1 | Uncharacterized protein
R5VMZ1_9FIRM | Firmicutes bacterium CAG:631 | GDSL-like protein
A0A0H1UPZ9_STRAG | Streptococcus agalactiae | Acylneuraminate cytidylyltransferase
A0A0A6S1U2_STRUB | Streptococcus uberis | Acylneuraminate cytidylyltransferase
R2P1Z4_9ENTE | Enterococcus raffinosus ATCC 49464 | Uncharacterized protein
A0A0G1KN57_9BACT | candidate division WWE3 bacterium GW2011_GWC2_44_9 | Secreted protein
E9I8K8_SOLIN | Solenopsis invicta | Putative uncharacterized protein
A0A026WMY1_CERBI | Cerapachys biroi | Maternal effect protein oskar
E2BYH0_HARSA | Harpegnathos saltator | Putative uncharacterized protein
E2A7I8_CAMFO | Camponotus floridanus | Putative uncharacterized protein
F2WJY6_9HYME | Messor pergandei | Oskar
F4WQN7_ACREC | Acromyrmex echinatior | Maternal effect protein oskar
E9IZ46_SOLIN | Solenopsis invicta | Putative uncharacterized protein
A0A0J7KH44_LASNI | Lasius niger | Maternal effect protein oskar
E1A883_NASVI | Nasonia vitripennis | Oskar
A0A0C9QHR7_9HYME | Fopius arisanus | Osk protein
K4MTL4_GRYBI | Gryllus bimaculatus | Oskar
Q2PP79_AEDAE | Aedes aegypti | Oskar
B0WIV7_CULQU | Culex quinquefasciatus | Oskar
U5EFJ8_9DIPT | Corethrella appendiculata | Putative oskar
T1PG45_MUSDO | Musca domestica | GDSL-like Lipase/Acylhydrolase
A0A0L0CP24_LUCCU | Lucilia cuprina | Uncharacterized protein
W8CE30_CERCA | Ceratitis capitata | Maternal effect protein oskar
A0A034WRF5_BACDO | Bactrocera dorsalis | Maternal effect protein oskar
A0A0K8U7J3_BACLA | Bactrocera latifrons | Maternal effect protein oskar
A0A0A1XRQ4_BACCU | Bactrocera cucurbitae | Maternal effect protein oskar
A1Y1T7_DROIM | Drosophila immigrans | Oskar
B4LXK5_DROVI | Drosophila virilis | Oskar
A0A0M4F3M8_DROBS | Drosophila busckii | Osk
B4JTJ1_DROGR | Drosophila grimshawi | GH23955
B4K9E5_DROMO | Drosophila mojavensis | Uncharacterized protein
B3LZ06_DROAN | Drosophila ananassae | Uncharacterized protein
B3P1W4_DROER | Drosophila erecta | GG13545
B4PTX6_DROYA | Drosophila yakuba | Uncharacterized protein
E8NH25_DROME | Drosophila melanogaster | RE24380p
B4HKZ1_DROSE | Drosophila sechellia | GM23770
A0A126GUR4_DROME | Drosophila melanogaster | Oskar, isoform D
Q295Q4_DROPS | Drosophila pseudoobscura pseudoobscura | Uncharacterized protein, isoform A
A0A059PGF2_9MUSC | Drosophila pseudoobscura | GA10627
A0A059PF64_9MUSC | Drosophila pseudoobscura | GA10627
B4N815_DROWI | Drosophila willistoni | Uncharacterized protein
T1DTM7_ANOAQ | Anopheles aquasalis | Uncharacterized protein
W5JJ85_ANODA | Anopheles darlingi | Uncharacterized protein
A0A084WRU4_ANOSI | Anopheles sinensis | AGAP003545-PA-like protein
Q7PQJ3_ANOGA | Anopheles gambiae | AGAP003545-PA
A0A0G3CFD7_METBA | Methanosarcina barkeri CM1 | GDSL family lipase/acylhydrolase
1.46752

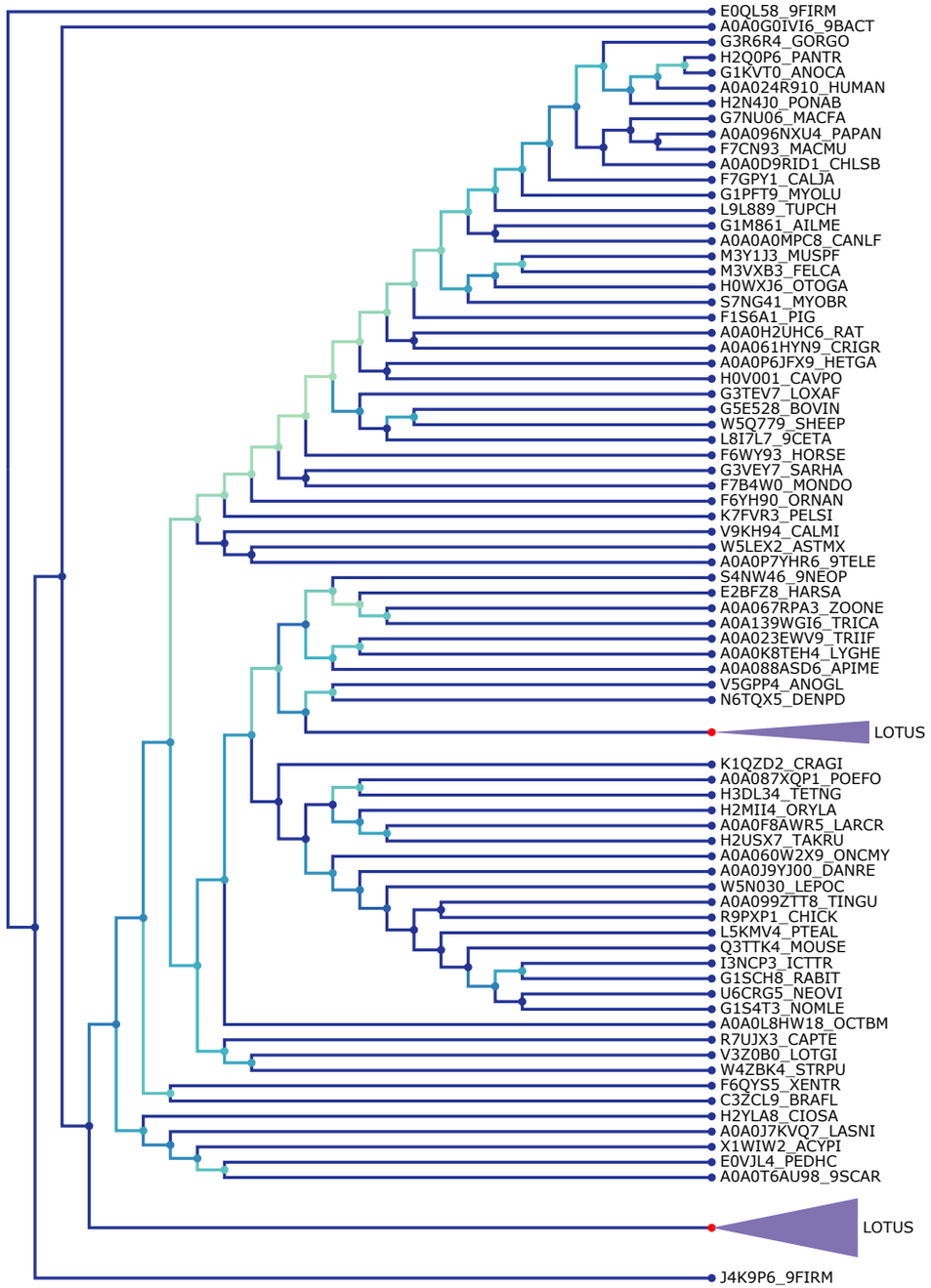
MUSCLE tree



T COFFEE tree



MUSCLE tree



T COFFEE tree

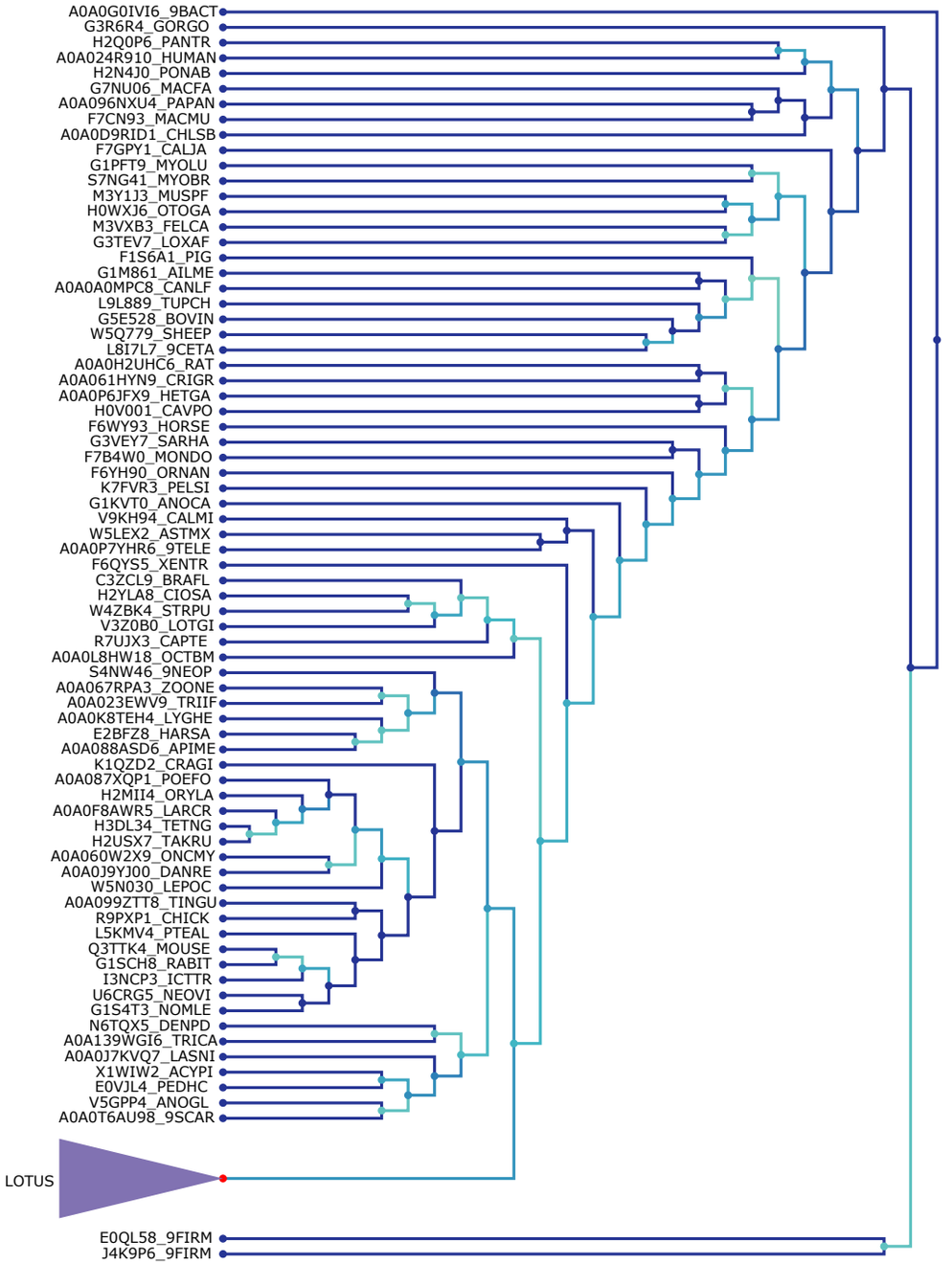


Figure 3 Supplements

Figure 3–figure supplement 1: GC3 correlations between the LOTUS and OSK domains. IntraGene distribution scatter plot for the coding sequences of the 17 genomes analyzed. Sequences were cut into two parts as per the description in Methods "Generation of IntraGene distribution of codon use". Z-scores were calculated for the 5' and 3' ends against the genome distribution. Finally, the 5' and 3' "domain" values were plotted against each other and a linear regression was performed. The Oskar sequence was then plotted in red. A kernel density estimation is displayed as black contour lines, increasing in shades of blue as the density increases.

Figure 3–figure supplement 2: AT3 correlations between the LOTUS and OSK domains. Intra-Gene distribution scatter plot for the coding sequences of the 17 genomes analyzed. Sequences were cut into two parts as per the description in Methods "Generation of intra-gene distribution of codon use". Z-scores were calculated for the 5' and 3' ends against the genome distribution. Finally, the 5' and 3' "domain" values were plotted against each other and a linear regression was performed. The Oskar sequence was then plotted in red. A kernel density estimation is displayed as black contour lines, increasing in shades of blue as the density increases.

Figure 3–figure supplement 3: A3/T3/G3/C3 correlations between the LOTUS and OSK domains. (**a**) Intra-Gene distribution scatter plot for the coding sequences of the 17 genomes analyzed. Sequences were cut into two parts as per the description in Methods "Generation of intra-gene distribution of codon use". The A3, T3, G3 and C3 codon use was measured, and Z-score calculations, value plots and linear regression were performed as described for Supplementary Figure 5. The A3, T3 G3 and C3 content is generally similar in the 5' and 3'regions of all genes across the genome (A3: $r^2 = 0.40$, $p = 0$; T3: $r^2 = 0.34$, $p = 0$; G3: $r^2 = 0.40$, $p = 0$; C3: $r^2 = 0.30$, $p = 0$). (**b**) OSK vs LOTUS A3, T3, G3 and C3 use across the 17 genomes analyzed. The A3, T3, G3 and C3 content Z-score were calculated against the genome distribution. The A3, G3 and C3 content of the two domains of the Oskar gene are not correlated with each other. However, the T3 distribution follows a linear correlation similar to the one found across the Intra-Gene distribution (A3: $r^2 = -0.04$, $p = 0.48$; T3: $r^2 = 0.29$, $p = 0.026$; G3: $r^2 = -0.09$, $p = 0.25$; C3: $r^2 = 0.02$, $p = 0.59$).

Figure 3–figure supplement 4: OSK-LOTUS VS Random Cut method. Comparison of Wobble position use of codon between OSK and LOTUS domains versus the Oskar gene cut at random following the same procedure as the intra-gene distribution. All analyses were performed as described in Figure 3.

Figure 3–figure supplement 5: IntraGene Schematic. Schematic representation of the intra-gene distribution. For each gene, a random cut preserving the frame was chosen. The position was forced to be over 351bp and less than the length of the sequence minus 351bp. Each part of the sequence was then split into two bins, one for the 5' and one for the 3'. Moreover, the gene identity was preserved to compare the 5' and 3' end of the same sequence in all following analysis. For more information see *Generation of Intra-Gene distribution of codon use.*

957 **Figure 3–figure supplement 6: Genome and Oskar CAI.** This Figure presents the results of
958 the Codon Adaptation Index (CAI) analysis of *oskar*. The CAI were computed against a codon
959 index measured for each genome (see *Calculation of Codon Adaptation Index (CAI)*). (A) The
960 distribution of CAI for each gene in the genome (grey box whisker plot). Colored dots represent
961 are the CAI values for full length Oskar (green), OSK domain (orange) and LOTUS domain
962 (blue) sequences. (B) The distribution of the Delta CAI, or the difference in CAI between two
963 halves of the same gene sequence. Grey = intra-gene distribution, where for each gene, the 5'
964 and 3' CAI were computed and the difference calculated. Red = Delta CAI between the LOTUS
965 and OSK domains of Oskar (C). The Z score distribution of the CAI scores for the 3' (dark grey)
966 and 5' (light grey) halves of a gene. Colored dots represent the CAI scores for the LOTUS
967 domain (blue) and OSK domain (orange) of Oskar.
968

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

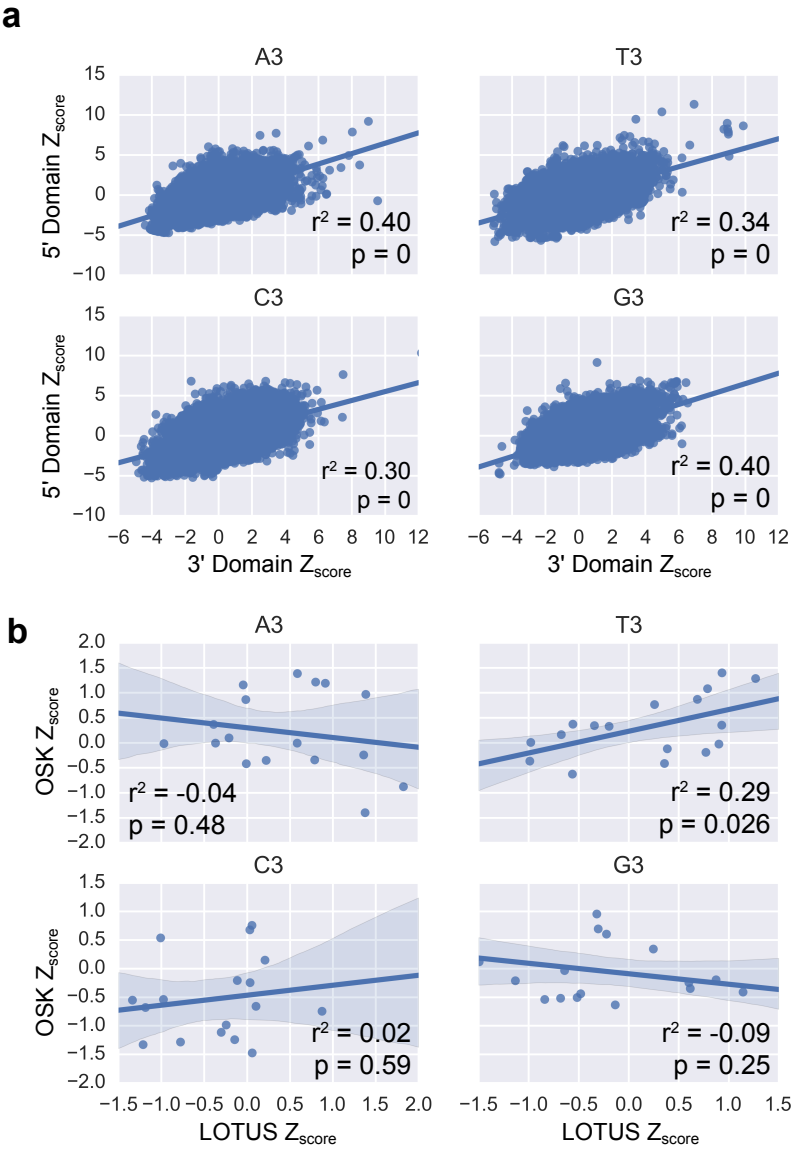
Leo Blondel, Tamsin E. M. Jones and Cassandra G. Extavour

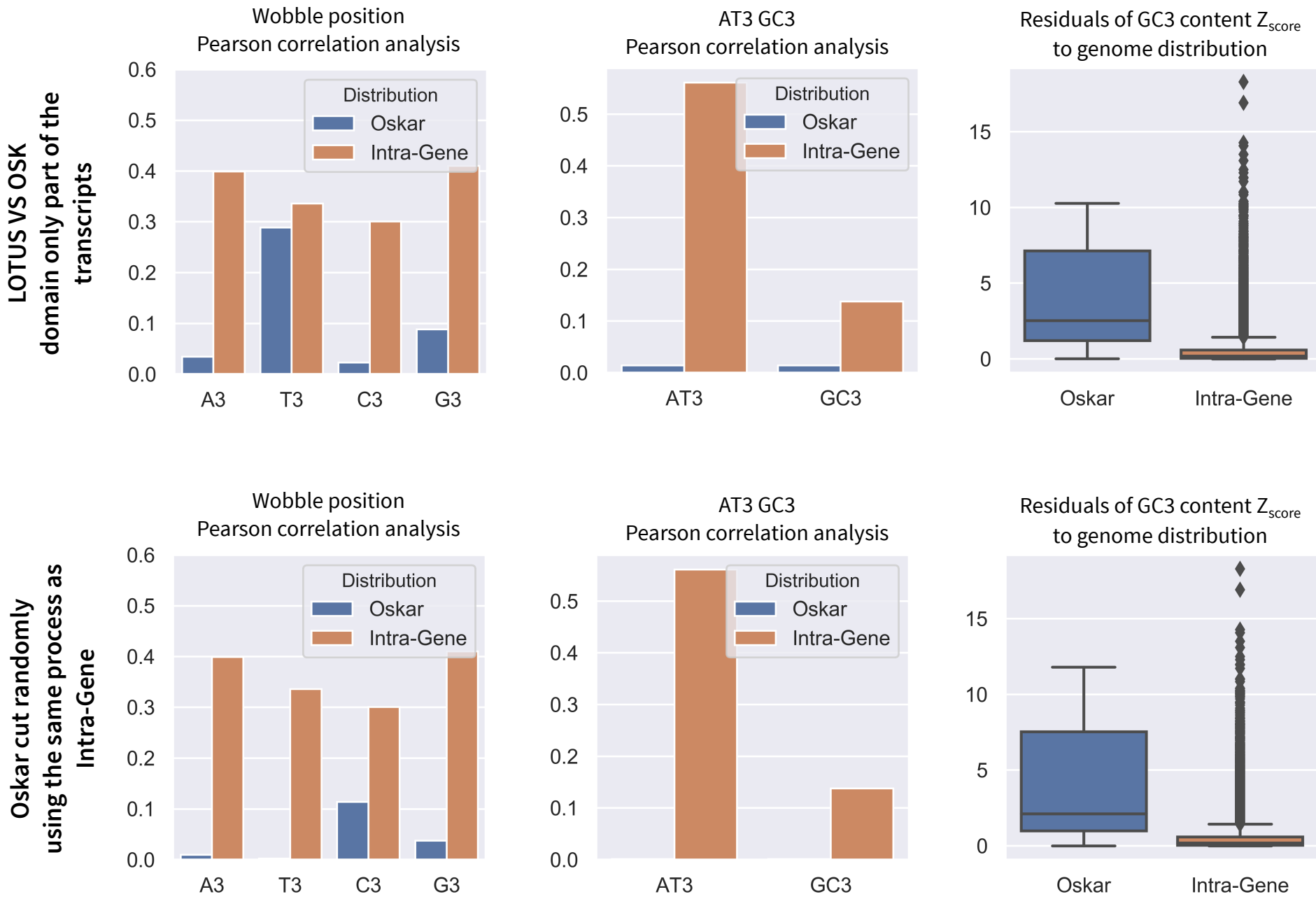
Figure 3–figure supplement 1: GC3 correlations between the LOTUS and OSK domains. IntraGene distribution scatter plot for the coding sequences of the 17 genomes analyzed. Sequences were cut into two parts as per the description in Methods "Generation of IntraGene distribution of codon use". Z-scores were calculated for the 5' and 3' ends against the genome distribution. Finally, the 5' and 3' "domain" values were plotted against each other and a linear regression was performed. The Oskar sequence was then plotted in red. A kernel density estimation is displayed as black contour lines, increasing in shades of blue as the density increases.

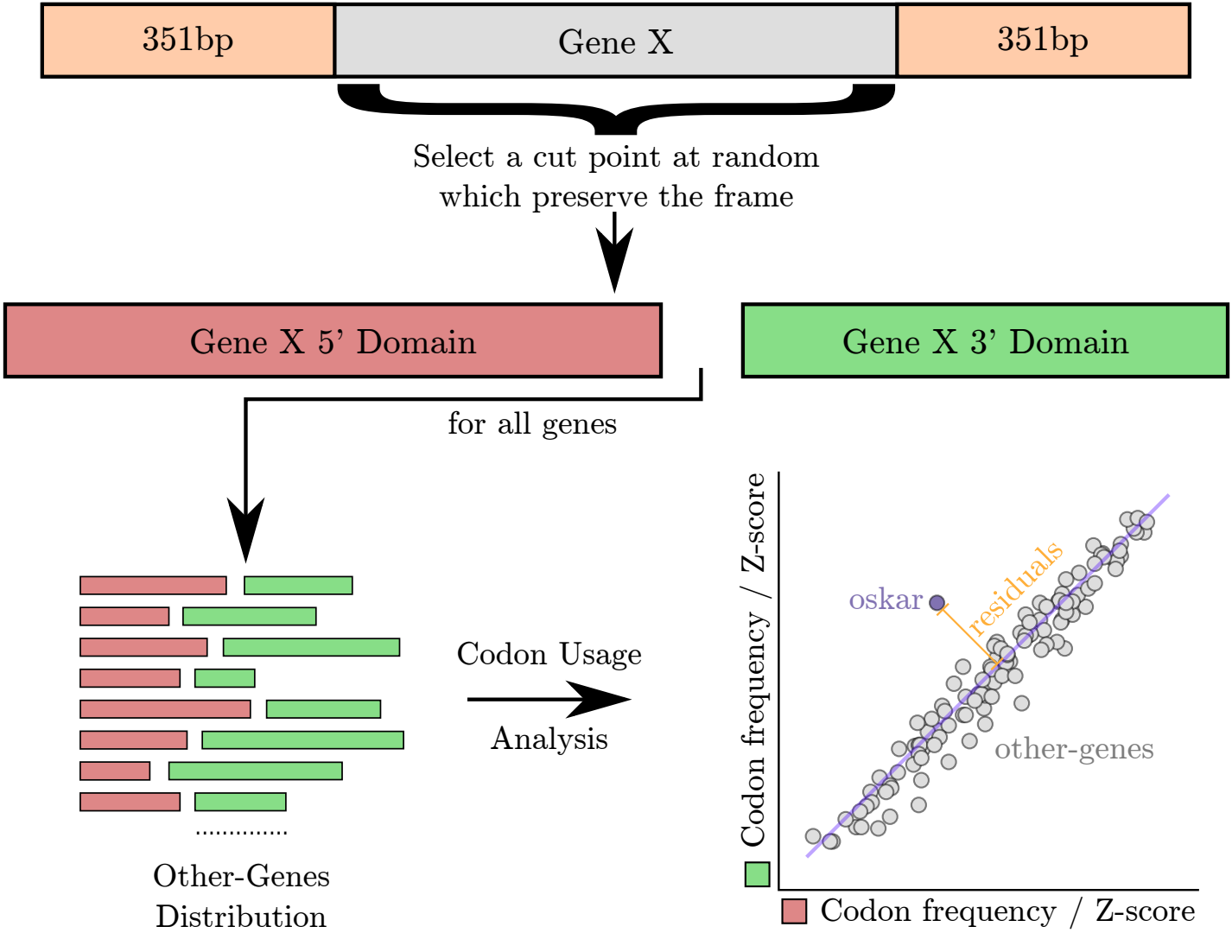
Figure download link:
https://www.dropbox.com/s/kxjypf064kano9u/Blondel_Jones_Extavour_HGT_Figure%203%E2%80%93figure%20supplement%201%3A%20GC3%20correlations%20between%20the%20LOTUS%20and%20OSK%20domains.pdf?dl=0

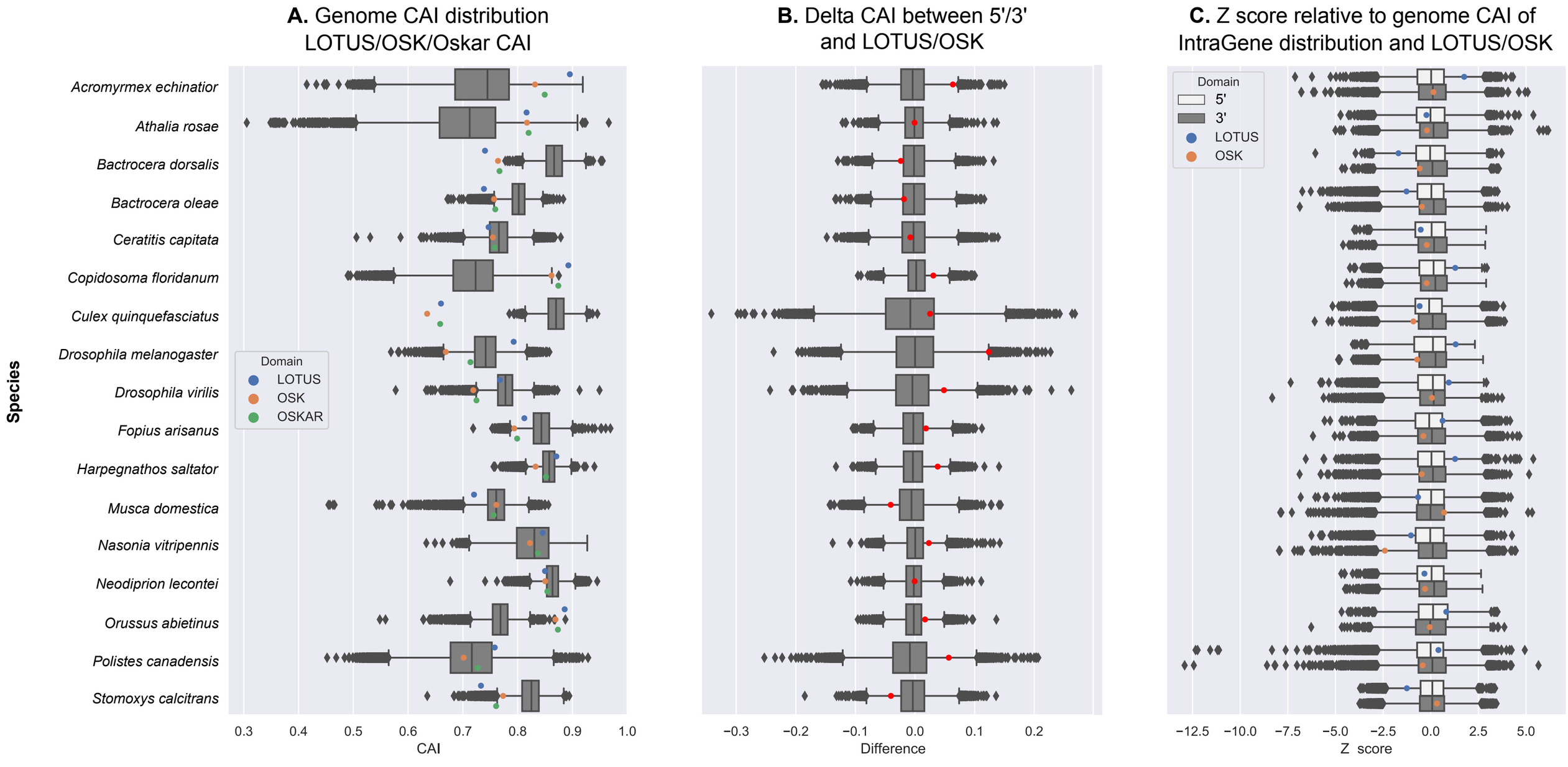
Figure 3–figure supplement 2: AT3 correlations between the LOTUS and OSK domains. Intra-Gene distribution scatter plot for the coding sequences of the 17 genomes analyzed. Sequences were cut into two parts as per the description in Methods "Generation of intra-gene distribution of codon use". Z-scores were calculated for the 5' and 3' ends against the genome distribution. Finally, the 5' and 3' "domain" values were plotted against each other and a linear regression was performed. The Oskar sequence was then plotted in red. A kernel density estimation is displayed as black contour lines, increasing in shades of blue as the density increases.

Figure download link:
https://www.dropbox.com/s/1gt5xzu6eigikd5/Blondel_Jones_Extavour_HGT_Figure%203%E2%80%93figure%20supplement%202%3A%20AT3%20correlations%20between%20the%20LOTUS%20and%20OSK%20domains.pdf?dl=0


a
A3
T3
C3
G3
5' Domain Z_score
3' Domain Z_score
r² = 0.40
p = 0
r² = 0.34
p = 0
r² = 0.30
p = 0
r² = 0.40
p = 0
b
OSK Z_score
LOTUS Z_score
r² = -0.04
p = 0.48
r² = 0.29
p = 0.026
r² = 0.02
p = 0.59
r² = -0.09
p = 0.25


LOTUS VS OSK
domain only part of the transcripts
Wobble position
Pearson correlation analysis
Distribution
Oskar
Intra-Gene
A3
T3
C3
G3
AT3 GC3
Pearson correlation analysis
AT3
GC3
Residuals of GC3 content Z_{score}
to genome distribution
Oskar
Intra-Gene
Oskar cut randomly
using the same process as
Intra-Gene


351bp
Gene X
351bp
Select a cut point at random
which preserve the frame
Gene X 5' Domain
Gene X 3' Domain
for all genes
Other-Genes
Distribution
Codon Usage
Analysis
oskar
residuals
other-genes
Codon frequency / Z-score
Codon frequency / Z-score


A. Genome CAI distribution LOTUS/OSK/Oskar CAI
B. Delta CAI between 5'/3' and LOTUS/OSK
C. Z score relative to genome CAI of IntraGene distribution and LOTUS/OSK
Species
Acromyrmex echinatior
Athalia rosae
Bactrocera dorsalis
Bactrocera oleae
Ceratitis capitata
Copidosoma floridanum
Culex quinquefasciatus
Drosophila melanogaster
Drosophila virilis
Fopius arisanus
Harpegnathos saltator
Musca domestica
Nasonia vitripennis
Neodiprion lecontei
Orussus abietinus
Polistes canadensis
Stomoxys calcitrans
Domain
LOTUS
OSK
OSKAR
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
CAI
−0.3
−0.2
−0.1
0.0
0.1
0.2
Difference
Domain
5'
3'
LOTUS
OSK
−12.5
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
Z score

Supplementary Information 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:

BLONDEL_JONES_EXTAVOUR_HGT_SUPPLEMENTARY_MATERIALS_ALL
(THIS PDF DOCUMENT)

1. Supplementary Tables
    a. Table S1: List of genomes and transcriptomes used for automated *oskar* search.
    b. Table S2: List of *oskar* sequences used in the final alignment.
    c. Table S3: List of sequences used for phylogenetic analysis of the LOTUS domain.
    d. Table S4: List of sequences used for phylogenetic analysis of the OSK domain.
    e. Table S5: List of genomes analyzed for codon use.

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

OSKAR HGT SUPPLEMENTARY INFORMATION FILES
(FOLDER)

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.

Please download Oskar HGT SUPPLEMENTARY INFORMATION FILES Folder here:

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

42
43 **Supplementary Table S1: List of genomes and transcriptomes used for automated *oskar* search.**
44 List of genomes and transcriptomes that were downloaded, annotated, and searched for *oskar*
45 sequences (*see "Hidden Markov Model (HMM) generation and alignments of the OSK and LOTUS*
46 *domains*" in Methods). The table reports the database provenance (NCBI genome or TSA, or 1KITE
47 database) and the accession number. The TSA accession ID can be searched using the NCBI TSA
48 browser here: https://www.ncbi.nlm.nih.gov/Traces/wgs/?view=TSA.
49

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

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

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

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

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

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

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

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

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

Supplementary Table S2: 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	*Aretaon asperrimus*	Heteropterygidae	Phasmatodea	P	HMMER
TSA	GAKJ01010751.1	*Sitodiplosis mosellana*	Cecidomyiidae	Diptera	P	HMMER
TSA	GATI01010233.1	*Bombylius major*	Bombyliidae	Diptera	P	HMMER
TSA	GAWW01000144.1	*Tenthredo koehleri*	Tenthredinidae	Hymenoptera	P	HMMER
TSA	GAIW01009539.1	*Ganaspis sp.*	Figitidae	Hymenoptera	P	HMMER
TSA	GATY01008637.1	*Chrysis viridula*	Chrysididae	Hymenoptera	P	HMMER
TSA	GAXM01030263.1	*Mischocyttarus flavitarsis*	Vespidae	Hymenoptera	P	HMMER
TSA	GAFR01040300.1	*Polistes canadensis*	Vespidae	Hymenoptera	P	HMMER
TSA	GAZD01106195.1	*Lipara lucens*	Chloropidae	Diptera	P	HMMER
TSA	GAVA01002196.1	*Triarthria setipennis*	Tachinidae	Diptera	P	HMMER
TSA	GACV01001831.1	*Philopotamus ludificatus*	Philopotaminae	Trichoptera	P	HMMER
TSA	GAPE01019095.1	*Brassicogethes aeneus*	Nitidulidae	Coleoptera	P	HMMER
TSA	GBCX01024638.1_7.1	*Dastarcus helophoroides**	Bothrideridae	Coleoptera	P	BLAST 1
TSA	GAKJ01010751.1	*Sitodiplosis mosellana*	Cecidomyiidae	Diptera	P	HMMER
TSA	GAWM01006639.1	*Culicoides sonorensis*	Ceratopogonidae	Diptera	P	HMMER
TSA	GAMD01000859.1	*Anopheles aquasalis*	Culicidae	Diptera	P	HMMER
TSA	GBEO01001325.1	*Anopheles sinensis*	Culicidae	Diptera	P	HMMER
TSA	GAKP01002609.1	*Bactrocera dorsalis*	Tephritidae	Diptera	P	HMMER
TSA	GAXM01030263.1	*Mischocyttarus flavitarsis*	Vespidae	Hymenoptera	P	HMMER
TSA	GAFR01040300.1	*Polistes canadensis canadensis*	Vespidae	Hymenoptera	P	HMMER
TSA	GBGV01010610.1	*Polistes metricus*	Vespidae	Hymenoptera	P	HMMER
TSA	GAIW01011550.1	*Ganaspis sp. G1*	Figitidae	Hymenoptera	P	HMMER
TSA	GAJC01011221.1	*Leptopilina heterotoma*	Figitidae	Hymenoptera	P	HMMER
TSA	GAIW01009539.1	*Ganaspis sp. G1*	Figitidae	Hymenoptera	P	HMMER

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

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

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

58

Supplementary Table S3: List of sequences and their BLAST results used for phylogenetic analysis of the LOTUS domain. The sequences were obtained by searching the TrEMBL database using hmmsearch and the final HMM generated for LOTUS (Supplementary files: HMM>LOTUS.hmm). Reported are the UniProtID (*Accession Number*), the Domain and Phylum origin of the sequence, the E-value, score and bias given by hmmsearch, and the description of the target from UniProt. To obtain sequences for each entry, either search UniProt directly (https://www.uniprot.org/) or consult the final alignment in Supplementary Files: Alignments>LOTUS_TREE.fasta. Phylum abbreviations: A = Arthropoda; An = Annelida; E = Echinodermata; F = Firmicutes; M = Mollusca; T = Tunicata; V = Vertebrata; ? = unclassified

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
B4N816_DROWI	Eukarya	A	8.50E-25	98.3	9.6	Uncharacterized protein OS=Drosophila willistoni GN=Dwil\GK11116 PE=4 SV=2
B4K9E5_DROMO	Eukarya	A	8.70E-25	98.2	0.6	Uncharacterized protein OS=Drosophila mojavensis GN=Dmoj\GI10055 PE=4 SV=2
A0A026WMY1_CERBI	Eukarya	A	2.20E-24	97	0.1	Maternal effect protein oskar OS=Cerapachys biroi GN=X777_01612 PE=4 SV=1
C3ZCL9_BRAFL	Eukarya	A	5.40E-24	95.7	0	Putative uncharacterized protein OS=Branchiostoma floridae 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
B4JTJ1_DROGR	Eukarya	A	8.30E-23	91.9	1.4	GH23955 OS=Drosophila grimshawi GN=Dgri\GH23955 PE=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
T1PG45_MUSDO	Eukarya	A	1.60E-21	87.8	1.2	GDSL-like Lipase/Acylhydrolase OS=Musca domestica PE=2 SV=1
A0A0L8HW18_OCTBM	Eukarya	M	1.70E-21	87.7	7.5	Uncharacterized protein OS=Octopus bimaculoides GN=OCBIM_22005378mg PE=4 SV=1
F6QYS5_XENTR	Eukarya	V	6.10E-20	82.7	0.2	Uncharacterized protein (Fragment) OS=Xenopus tropicalis PE=4 SV=1
B0WIV7_CULQU	Eukarya	A	8.90E-20	82.2	0.1	Oskar OS=Culex quinquefasciatus GN=CpipJ_CPIJ007471 PE=2 SV=1
A0A034WRF5_BACDO	Eukarya	A	1.30E-19	81.6	13.1	Maternal effect protein oskar OS=Bactrocera dorsalis GN=OSKA 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
X1WIW2_ACYPI	Eukarya	A	1.70E-19	81.3	12.3	Uncharacterized protein OS=Acyrthosiphon pisum GN=LOC100162069 PE=4 SV=1
A0A023EWV9_TRIIF	Eukarya	A	1.80E-19	81.2	0.2	Putative transcriptional coactivator (Fragment) OS=Triatoma infestans PE=2 SV=1
A0A0L0CP24_LUCCU	Eukarya	A	2.40E-19	80.8	0.5	Uncharacterized protein OS=Lucilia cuprina GN=FF38_12727 PE=4 SV=1
A0A0K8W0W3_BACLA	Eukarya	A	5.00E-19	79.8	0	Maternal effect protein oskar OS=Bactrocera latifrons GN=osk_1 PE=4 SV=1
A0A0A1XRQ4_BACCU	Eukarya	A	7.20E-19	79.3	6.4	Maternal effect protein oskar OS=Bactrocera cucurbitae GN=osk PE=4 SV=1
A0A088ASD6_APIME	Eukarya	A	2.90E-18	77.3	0.2	Uncharacterized protein OS=Apis mellifera GN=LOC726241 PE=4 SV=1
K1QZD2_CRAGI	Eukarya	M	1.60E-17	75	0	Tudor domain-containing protein 7 OS=Crassostrea gigas GN=CGI_10018436 PE=4 SV=1
A0A139WGI6_TRICA	Eukarya	A	1.70E-17	74.9	0.1	Tudor domain-containing protein 7-like protein OS=Tribolium castaneum GN=TcasGA2_TC034722 PE=4 SV=1
F6YH90_ORNAN	Eukarya	V	1.80E-17	74.8	0	Uncharacterized protein OS=Ornithorhynchus anatinus GN=TDRD5 PE=4 SV=1
W5JJ85_ANODA	Eukarya	A	2.10E-17	74.6	0	Uncharacterized protein OS=Anopheles darlingi GN=AND_005442 PE=4 SV=1
A0A084WRU4_ANOSI	Eukarya	A	2.40E-17	74.4	0.2	AGAP003545-PA-like protein OS=Anopheles sinensis GN=ZHAS_00021239 PE=4 SV=1
H2YLA8_CIOSA	Eukarya	T	2.60E-17	74.3	0.1	Uncharacterized protein OS=Ciona savignyi GN=Csa.10307 PE=4 SV=1
G1KVT0_ANOCA	Eukarya	V	2.70E-17	74.2	0.1	Uncharacterized protein OS=Anolis carolinensis GN=TDRD5 PE=4 SV=1
G3TEV7_LOXAF	Eukarya	V	4.50E-17	73.5	0	Uncharacterized protein OS=Loxodonta africana GN=TDRD5 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
M3VXB3_FELCA	Eukarya	V	1.30E-14	65.6	0	Uncharacterized protein OS=Felis catus GN=TDRD5 PE=4 SV=1
A0A099ZTT8_TINGU	Eukarya	V	1.40E-14	65.6	0	Tudor domain-containing protein 7 OS=Tinamus guttatus GN=N309_12928 PE=4 SV=1
G3R6R4_GORGO	Eukarya	V	1.50E-14	65.5	0	Uncharacterized protein OS=Gorilla gorilla gorilla GN=TDRD5 PE=4 SV=1
L9L889_TUPCH	Eukarya	V	1.60E-14	65.4	0.3	Tudor domain-containing protein 5 OS=Tupaia chinensis GN=TREES_T100015801 PE=4 SV=1
A0A0J7KVQ7_LASNI	Eukarya	A	1.70E-14	65.3	0	Tudor domain-containing protein 5 (Fragment) OS=Lasius niger GN=RF55_5458 PE=4 SV=1
E0VJL4_PEDHC	Eukarya	A	1.70E-14	65.3	0.9	Putative uncharacterized protein OS=Pediculus humanus subsp. corporis GN=Phum_PHUM247930 PE=4 SV=1
A0A0D9RID1_CHLSB	Eukarya	V	1.70E-14	65.2	0	Uncharacterized protein OS=Chlorocebus sabaeus GN=TDRD5 PE=4 SV=1
E2BFZ8_HARSA	Eukarya	A	2.00E-14	65	0.9	Tudor domain-containing protein 7 OS=Harpegnathos saltator GN=EAI_14615 PE=4 SV=1
A0A0T6AU98_9SCAR	Eukarya	A	2.10E-14	65	3.9	Uncharacterized protein (Fragment) OS=Oryctes borbonicus GN=AMK59_7658 PE=4 SV=1
R9PXP1_CHICK	Eukarya	V	2.30E-14	64.8	0	Tudor domain-containing protein 7 OS=Gallus gallus GN=TDRD7 PE=4 SV=1
F7CN93_MACMU	Eukarya	V	2.70E-14	64.6	0	Uncharacterized protein OS=Macaca mulatta GN=TDRD5 PE=4 SV=1
A0A0A0MPC8_CANLF	Eukarya	V	2.70E-14	64.6	0	Tudor domain-containing protein 5 OS=Canis lupus familiaris GN=TDRD5 PE=4 SV=1
M3Y1J3_MUSPF	Eukarya	V	2.90E-14	64.5	0	Uncharacterized protein OS=Mustela putorius furo GN=TDRD5 PE=4 SV=1
H2N4J0_PONAB	Eukarya	V	3.00E-14	64.5	0	Uncharacterized protein OS=Pongo abelii GN=TDRD5 PE=4 SV=1
H2Q0P6_PANTR	Eukarya	V	3.20E-14	64.4	0	Uncharacterized protein OS=Pan troglodytes GN=TDRD5 PE=4 SV=1
W5LEX2_ASTMX	Eukarya	V	3.50E-14	64.3	0	Uncharacterized protein OS=Astyanax mexicanus PE=4 SV=1
A0A024R910_HUMAN	Eukarya	V	4.00E-14	64.1	0	Tudor domain containing 5, isoform CRA_b OS=Homo sapiens GN=TDRD5 PE=4 SV=1
A0A0P6JFX9_HETGA	Eukarya	V	4.10E-14	64.1	0	Tudor domain-containing protein 5 isoform 2 OS=Heterocephalus glaber GN=TDRD5 PE=4 SV=1
H0WXJ6_OTOGA	Eukarya	V	4.70E-14	63.8	0	Uncharacterized protein OS=Otolemur garnettii GN=TDRD5 PE=4 SV=1
H0V001_CAVPO	Eukarya	V	5.20E-14	63.7	0	Uncharacterized protein OS=Cavia porcellus GN=TDRD5 PE=4 SV=1
G7NU06_MACFA	Eukarya	V	5.80E-14	63.6	0	Putative uncharacterized protein OS=Macaca fascicularis 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
S4NW46_9NEOP	Eukarya	A	1.20E-13	62.5	0	Tudor domain containing 7 (Fragment) OS=Pararge aegeria PE=4 SV=1
L5KMV4_PTEAL	Eukarya	V	2.10E-13	61.8	0	Tudor domain-containing protein 7 OS=Pteropus alecto GN=PAL_GLEAN10008027 PE=4 SV=1
I3NCP3_ICTTR	Eukarya	V	2.40E-13	61.6	0	Uncharacterized protein OS=Ictidomys tridecemlineatus GN=TDRD7 PE=4 SV=1
G1SCH8_RABIT	Eukarya	V	2.50E-13	61.5	0	Uncharacterized protein OS=Oryctolagus cuniculus GN=TDRD7 PE=4 SV=1
Q3TTK4_MOUSE	Eukarya	V	2.50E-13	61.5	0	Putative uncharacterized protein (Fragment) OS=Mus musculus GN=Tdrd7 PE=2 SV=1
A0A0H2UHC6_RAT	Eukarya	V	3.90E-13	60.9	0	Tudor domain-containing protein 5 OS=Rattus norvegicus GN=Tdrd5 PE=4 SV=1
U6CRG5_NEOVI	Eukarya	V	3.90E-13	60.9	0	Tudor domain-containing protein 7 OS=Neovison vison GN=TDRD7 PE=2 SV=1
K7FVR3_PELSI	Eukarya	V	5.10E-13	60.5	0	Uncharacterized protein OS=Pelodiscus sinensis GN=TDRD5 PE=4 SV=1
J4K9P6_9FIRM	Bacteria	F	2.90E-07	42.1	31.8	NYN domain protein OS=Peptostreptococcaceae bacterium AS15 GN=HMPREF1142_1162 PE=4 SV=1
E0QL58_9FIRM	Bacteria	F	0.00014	33.5	29.1	Uncharacterized protein OS=[Eubacterium] yurii subsp. margaretiae ATCC 43715 GN=HMPREF0379_1756 PE=4 SV=1
A0A0G0IVI6_9BACT	Bacteria	?	0.0025	29.5	1.9	Uncharacterized protein OS=candidate division TM6 bacterium GW2011_GWA2_36_9 GN=US32_C0002G0021 PE=4 SV=1

Supplementary Table S4: List of sequences and their BLAST results used for phylogenetic analysis of the OSK domain. The sequences were obtained by searching the TrEMBL database using hmmsearch and the final HMM generated for OSK (Supplementary files: HMM>OSK.hmm). Reported parameters are as described for Supplementary Table S3. To obtain sequences for each entry, either search UniProt directly (https://www.uniprot.org/) or consult the final alignment in Supplementary Files: Alignments>OSK_TREE.fasta. Phylum Abbreviations: A = Arthropoda; Ar = Archaea; As = Ascomycota; B = Bacteroidetes; C = Cyanobacteria; Eu = Euryarchaeota; F = Firmicutes; Fu = Fungi; P = Proteobacteria

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=Dgri\GH23955 PE=4 SV=1
F4WQN7_ACREC	Eukarya	A	8.10E-34	127	0.9	Maternal effect protein oskar OS=Acromyrmex echinatior GN=G5I_08127 PE=4 SV=1
E2A7I8_CAMFO	Eukarya	A	2.80E-33	125.2	0.4	Putative uncharacterized protein OS=Camponotus floridanus GN=EAG_03874 PE=4 SV=1
W8CE30_CERCA	Eukarya	A	5.80E-33	124.2	0.3	Maternal effect protein oskar OS=Ceratitis capitata GN=OSKA PE=2 SV=1
A0A0A1XRQ4_BACCU	Eukarya	A	7.50E-33	123.9	0.4	Maternal effect protein oskar OS=Bactrocera cucurbitae GN=osk PE=4 SV=1
B4N815_DROWI	Eukarya	A	8.10E-33	123.8	0.3	Uncharacterized protein OS=Drosophila willistoni GN=Dwil\GK11117 PE=4 SV=2
A0A026WMY1_CERBI	Eukarya	A	2.40E-32	122.3	0.7	Maternal effect protein oskar OS=Cerapachys biroi GN=X777_01612 PE=4 SV=1
E9IZ46_SOLIN	Eukarya	A	2.50E-32	122.2	0.1	Putative uncharacterized protein (Fragment) OS=Solenopsis invicta GN=SINV_01516 PE=4 SV=1
A0A034WRF5_BACDO	Eukarya	A	2.70E-32	122.1	0.5	Maternal effect protein oskar OS=Bactrocera dorsalis GN=OSKA PE=4 SV=1
A0A0K8U7J3_BACLA	Eukarya	A	5.70E-32	121	0.5	Maternal effect protein oskar OS=Bactrocera latifrons GN=osk_2 PE=4 SV=1
Q2PP79_AEDAE	Eukarya	A	6.30E-32	120.9	0.1	Oskar OS=Aedes aegypti PE=4 SV=1
B4K9E5_DROMO	Eukarya	A	8.60E-32	120.5	0.1	Uncharacterized protein OS=Drosophila mojavensis GN=Dmoj\GI10055 PE=4 SV=2
B4LXK5_DROVI	Eukarya	A	1.20E-31	120	0.1	Oskar OS=Drosophila virilis GN=osk PE=4 SV=1
A0A0M4F3M8_DROBS	Eukarya	A	1.80E-31	119.4	0.4	Osk OS=Drosophila busckii GN=Dbus_chr3Rg607 PE=4 SV=1
A0A0L0CP24_LUCCU	Eukarya	A	2.90E-31	118.8	1.5	Uncharacterized protein OS=Lucilia cuprina GN=FF38_12727 PE=4 SV=1
B4PTX6_DROYA	Eukarya	A	1.00E-30	117	0.5	Uncharacterized protein OS=Drosophila yakuba GN=Dyak\GE25914 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
A0A084WRU4_ANOSI	Eukarya	A	1.30E-27	107.1	0.1	AGAP003545-PA-like protein OS=Anopheles sinensis GN=ZHAS_00021239 PE=4 SV=1
W5JJ85_ANODA	Eukarya	A	4.50E-27	105.3	0	Uncharacterized protein OS=Anopheles darlingi GN=AND_005442 PE=4 SV=1
Q7PQJ3_ANOGA	Eukarya	A	6.00E-27	104.9	0	AGAP003545-PA OS=Anopheles gambiae GN=osk PE=4 SV=3
B0WIV7_CULQU	Eukarya	A	1.30E-26	103.9	0.4	Oskar OS=Culex quinquefasciatus GN=CpipJ_CPIJ007471 PE=2 SV=1
U5EFJ8_9DIPT	Eukarya	A	4.00E-25	99.1	0	Putative oskar (Fragment) OS=Corethrella appendiculata PE=2 SV=1
A0A126GUR4_DROME	Eukarya	A	3.10E-24	96.2	0.3	Oskar, isoform D OS=Drosophila melanogaster GN=osk PE=4 SV=1
A0A059PGF2_9MUSC	Eukarya	A	3.10E-24	96.2	0.5	GA10627 (Fragment) OS=Drosophila pseudoobscura GN=GA10627 PE=4 SV=1
K4MTL4_GRYBI	Eukarya	A	1.20E-23	94.2	0.2	Oskar OS=Gryllus bimaculatus PE=2 SV=1
A0A0C9QHR7_9HYME	Eukarya	A	5.40E-22	89	0	Osk protein OS=Fopius arisanus GN=osk PE=4 SV=1
A0A059PF64_9MUSC	Eukarya	A	1.90E-19	80.8	0.7	GA10627 (Fragment) OS=Drosophila pseudoobscura GN=GA10627 PE=4 SV=1
A0A0J7KH44_LASNI	Eukarya	A	1.40E-13	61.9	0	Maternal effect protein oskar (Fragment) OS=Lasius niger GN=RF55_10783 PE=4 SV=1
E2BYH0_HARSA	Eukarya	A	4.00E-12	57.3	0	Putative uncharacterized protein OS=Harpegnathos saltator GN=EAI_08923 PE=4 SV=1
A0A0J6BBM7_BREBE	Bacteria	F	3.10E-10	51.2	0	Lysophospholipase OS=Brevibacillus brevis GN=AB432_04505 PE=4 SV=1
A0A0H0SJ00_9BACL	Bacteria	F	9.60E-10	49.6	0	Lysophospholipase OS=Brevibacillus formosus GN=AA984_15375 PE=4 SV=1
E9I8K8_SOLIN	Bacteria	A	5.90E-09	47.1	1	Putative uncharacterized protein (Fragment) OS=Solenopsis invicta GN=SINV_16199 PE=4 SV=1
J3A568_9BACL	Bacteria	F	5.80E-08	43.9	0	Lysophospholipase L1-like esterase OS=Brevibacillus sp. BC25 GN=PMI05_03395 PE=4 SV=1

G2HS43_9PROT	Bacteria	P	1.10E-07	43	1	Lipolytic protein OS=Arcobacter sp. L GN=ABLL_2651 PE=4 SV=1
E6L4E3_9PROT	Bacteria	P	1.70E-07	42.4	0.7	Lipolytic protein OS=Arcobacter butzleri JV22 GN=HMPREF9401_1319 PE=4 SV=1
R5GT16_9FIRM	Bacteria	F	2.60E-07	41.8	0	GDSL-like protein OS=Eubacterium sp. CAG:786 GN=BN782_00012 PE=4 SV=1
A0A078KJ49_9FIRM	Bacteria	F	3.40E-07	41.4	0	Uncharacterized protein OS=[Clostridium] cellulosi GN=CCDG5_0508 PE=4 SV=1
A8EWS4_ARCB4	Bacteria	P	3.80E-07	41.3	0.7	Lipolytic enzyme, GDSL domain OS=Arcobacter butzleri (strain RM4018) GN=Abu_2183 PE=4 SV=1
S5PEQ8_9PROT	Bacteria	P	4.00E-07	41.2	0.4	Lipolytic enzyme, GDSL domain protein OS=Arcobacter butzleri 7h1h GN=A7H1H_2115 PE=4 SV=1
A0A0G9K3L5_9PROT	Bacteria	P	4.00E-07	41.2	0.6	Lipolytic protein OS=Arcobacter butzleri L348 GN=AA20_04280 PE=4 SV=1
R6T7B3_9BACE	Bacteria	B	4.80E-07	41	0	GDSL-like protein OS=Bacteroides sp. CAG:770 GN=BN777_00744 PE=4 SV=1
A0A0A6S1U2_STRUB	Bacteria	F	4.90E-07	40.9	0.2	Acylneuraminate cytidylyltransferase OS=Streptococcus uberis GN=NC01_08240 PE=4 SV=1
W3AC50_9BACL	Bacteria	F	5.50E-07	40.8	0.3	Uncharacterized protein OS=Planomicrobium glaciei CHR43 GN=G159_16940 PE=4 SV=1
A0A0M1UPT0_9PROT	Bacteria	P	5.50E-07	40.8	0.4	Lipolytic protein OS=Arcobacter butzleri ED-1 GN=ABED_1978 PE=4 SV=1
U6EVC2_CLOTA	Bacteria	F	6.30E-07	40.6	0.7	Platelet activating factor acetylhydrolase-likeprotein OS=Clostridium tetani 12124569 GN=BN906_00617 PE=4 SV=1
A0A098EIL2_9BACL	Bacteria	F	7.90E-07	40.3	0.1	Multifunctional acyl-CoA thioesterase I and protease I and 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
B7KKA6_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=Oscillatoriales 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 cytidylyltransferase 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

76

Supplementary Table S5: List of genomes analyzed for codon use.

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

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