eLife

SHORT REPORT

Figures and figure supplements

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

Leo Blondel *et al*

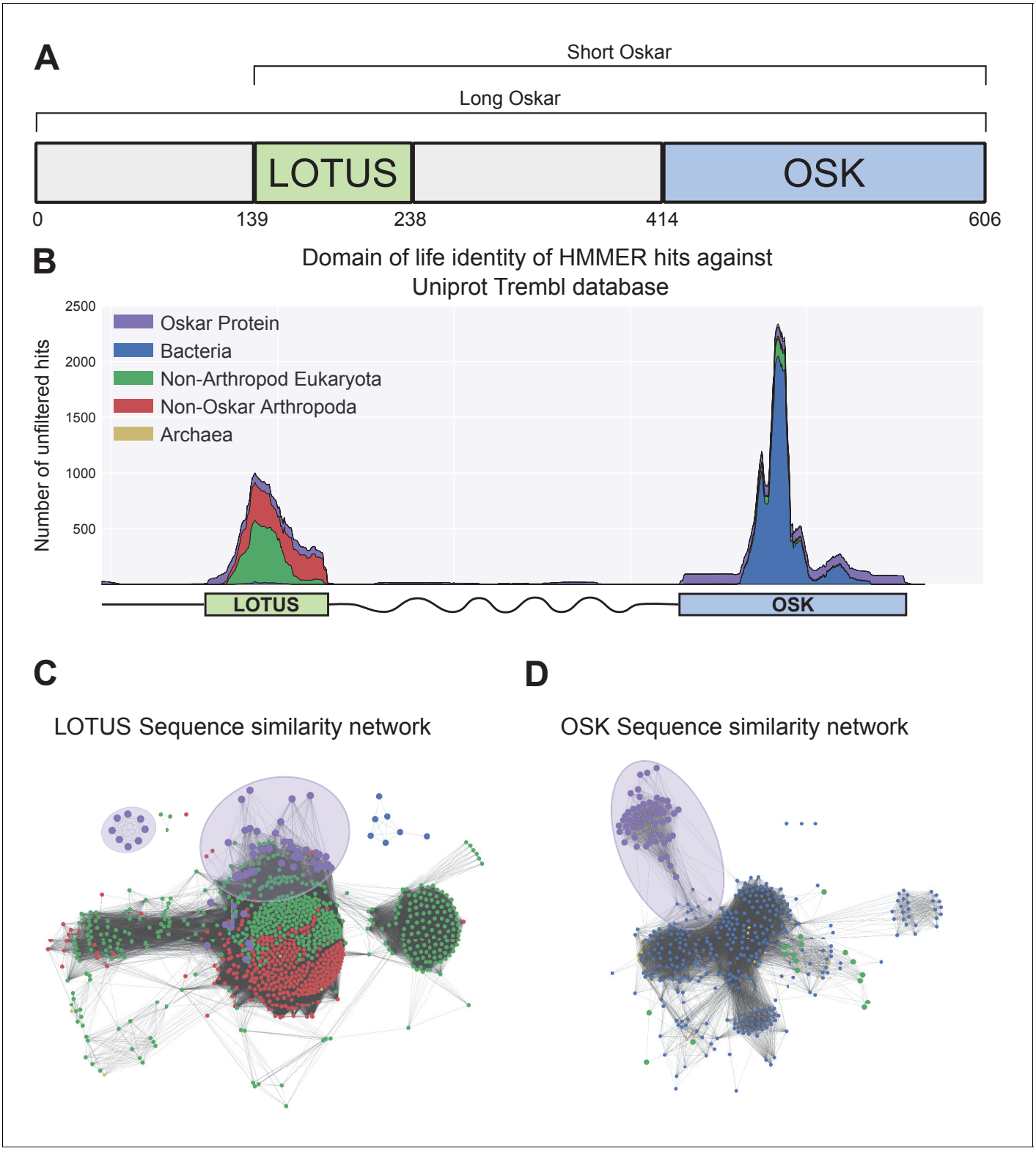

Figure 1. Sequence analysis of the Oskar gene. (**a**) Schematic representation of the Oskar gene. The LOTUS and OSK hydrolase-like domains are separated by a poorly conserved region of predicted high disorder and variable length between species. In some dipterans, a region 5' to the LOTUS domain is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

Figure 1 continued on next page

Figure 1 continued

HMMER search of OSK and LOTUS domains), stacked on top of each other. (**c, d**) EFI-EST-generated graphs of the sequence similarity network of the LOTUS (**c**) and OSK (**d**) domains of Oskar (***Gerlt et al., 2015***). Sequences were obtained using HMMER against the UniProtKB database. Most Oskar LOTUS sequences cluster within eukaryotes and arthropods. In contrast, Oskar OSK sequences cluster most strongly with a small subset of bacterial sequences.

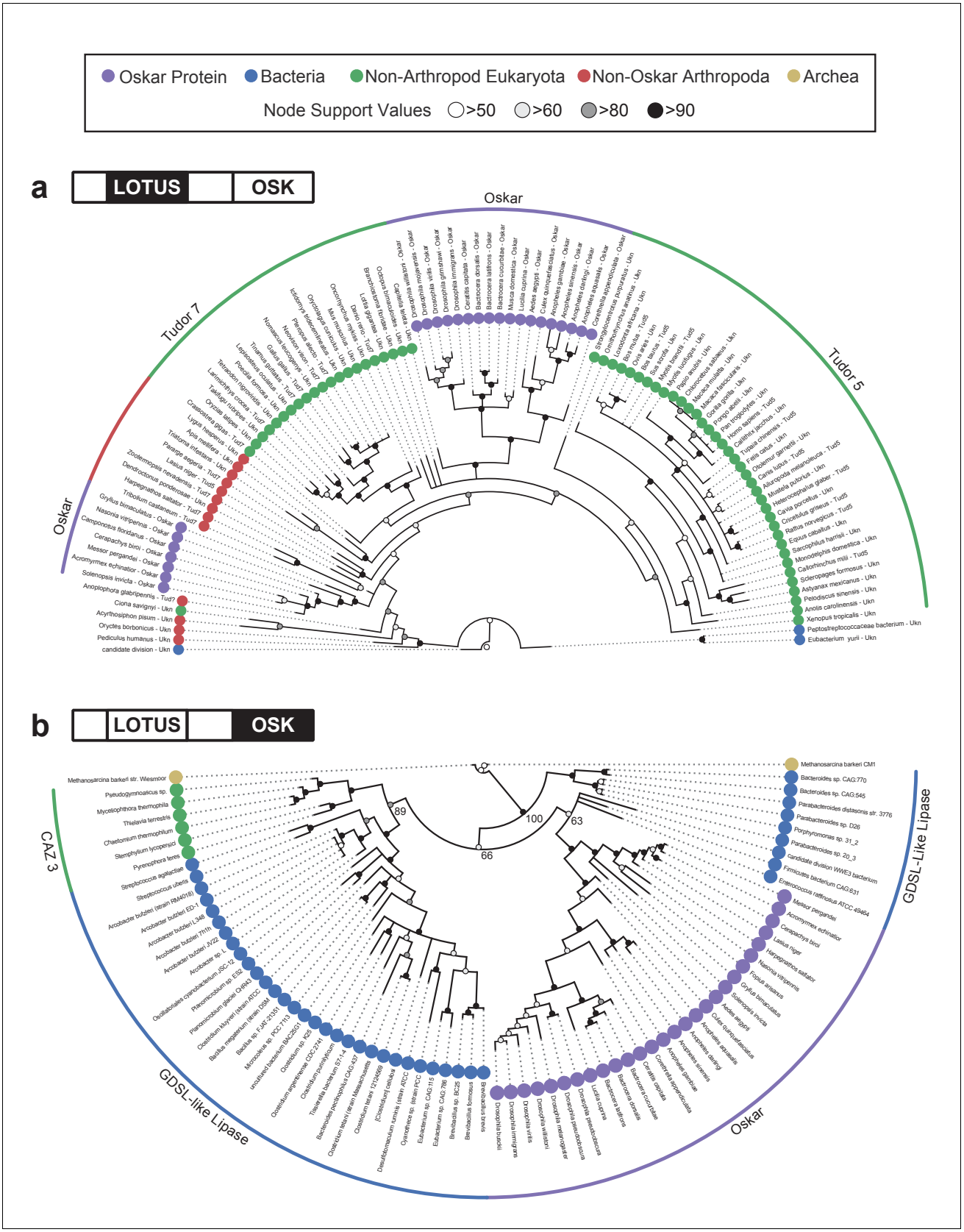

Figure 2. Phylogenetic analysis of the LOTUS and OSK domains. (**a**) Bayesian consensus tree for the LOTUS domain. Three major LOTUS-containing protein families are represented within the tree: Tudor 5, Tudor 7, and Oskar. Oskar LOTUS domains form two clades, one containing only dipterans

Figure 2 continued on next page

Figure 2 continued

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 a single clade comprising fungal Carbohydrate Active Enzyme 3 (CAZ3) proteins. For Bayesian and RaxML trees with all accession numbers and node support values see ***Figure 2—figure supplements 1–4***.

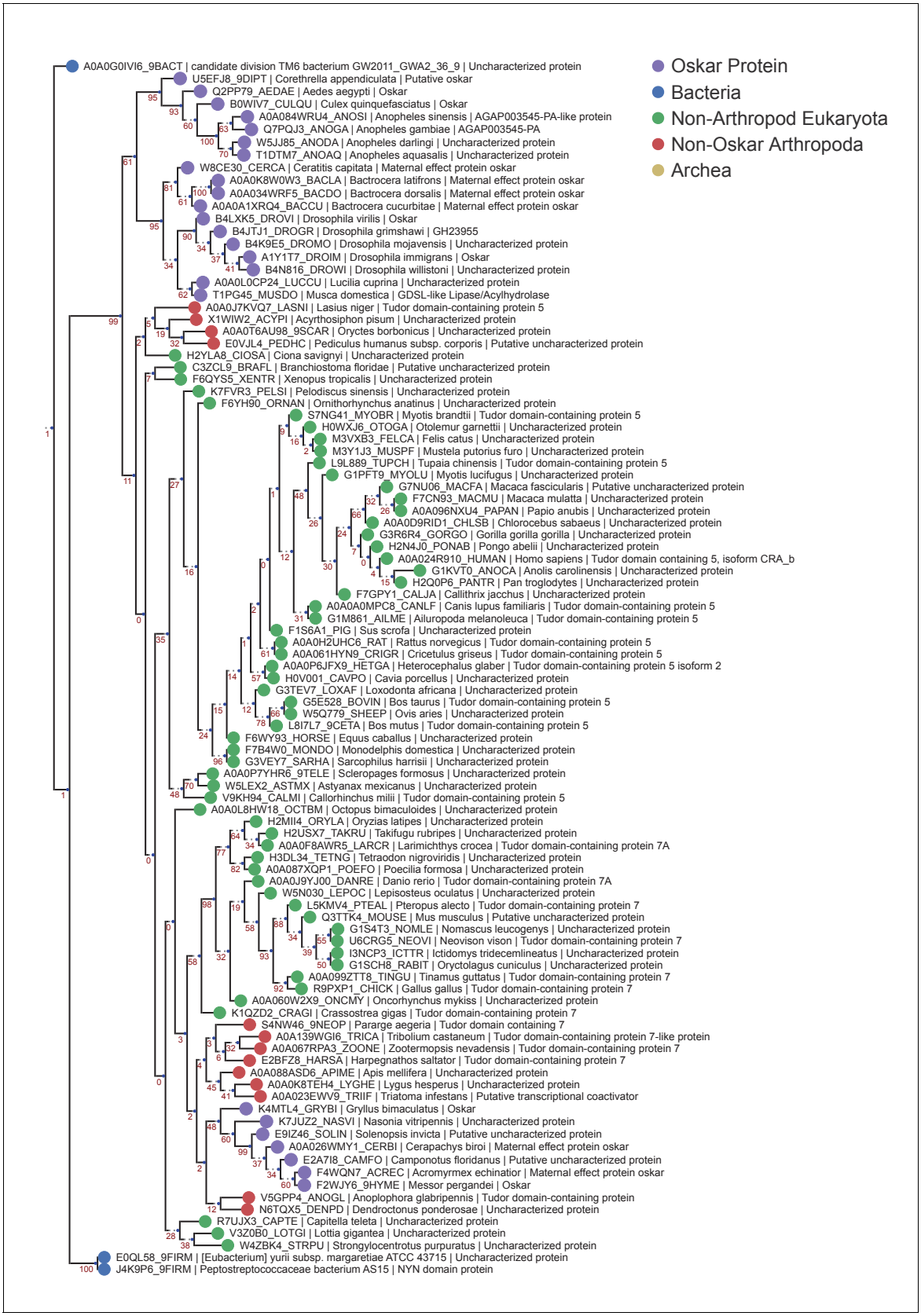

Figure 2—figure supplement 1. LOTUS Domain RaxML MUSCLE Tree. Phylogenetic tree of the HMMER sequences retrieved from the UniProt database using the LOTUS alignment HMM model. The top 97 hits were selected for phylogenetic analysis, and the only three bacterial sequences

Figure 2—figure supplement 1 continued on next page

Figure 2—figure supplement 1 continued

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 Materials and 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.

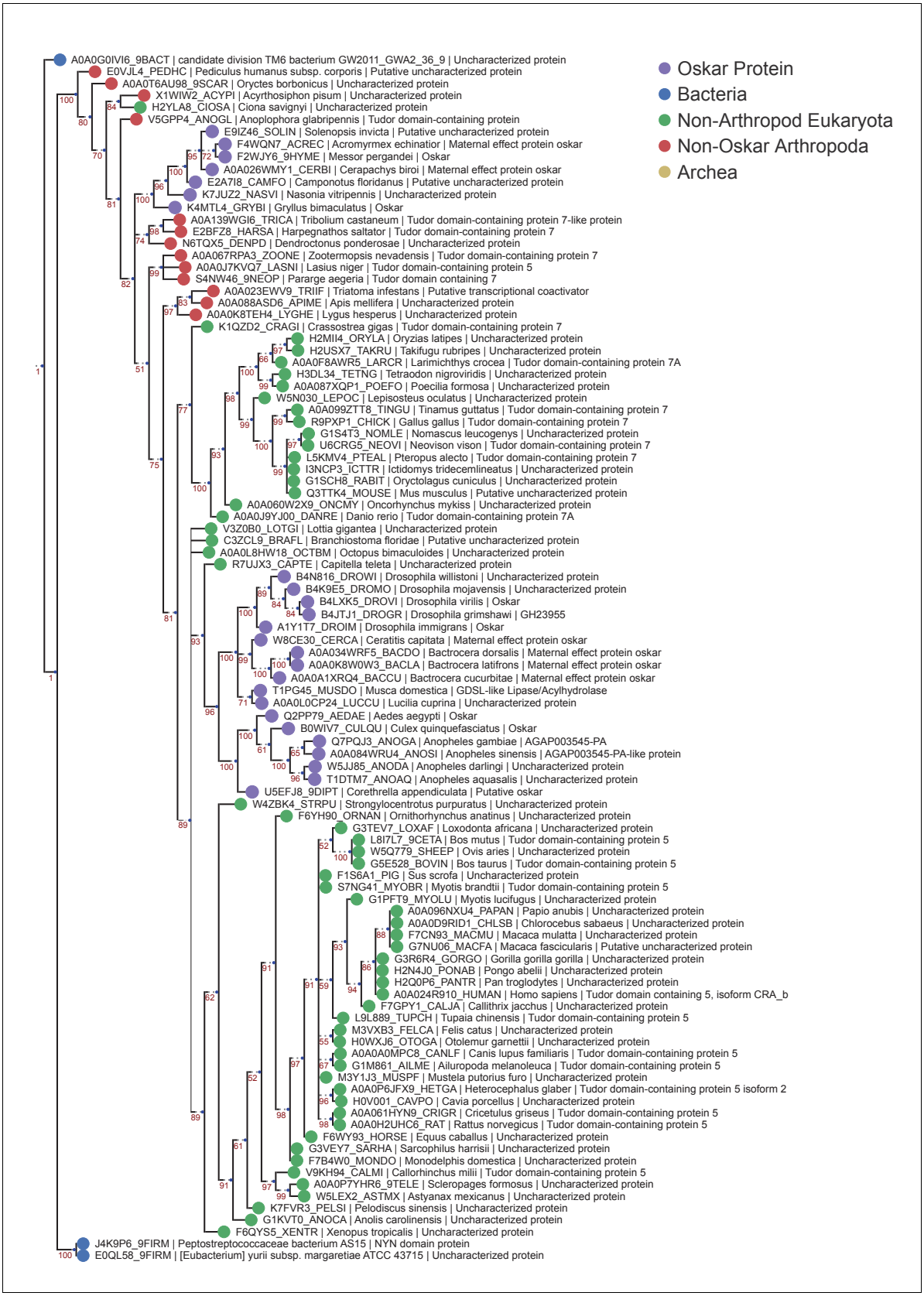

Figure 2—figure supplement 2. LOTUS Domain Bayesian MUSCLE Tree. Phylogenetic tree of the HMMER sequences retrieved from the UniProt database using the LOTUS alignment HMM model. 100 sequences were chosen for analysis as described for ***Figure 2—figure supplement 1***. The tree

Figure 2—figure supplement 2 continued on next page

Figure 2—figure supplement 2 continued

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 Materials and 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.

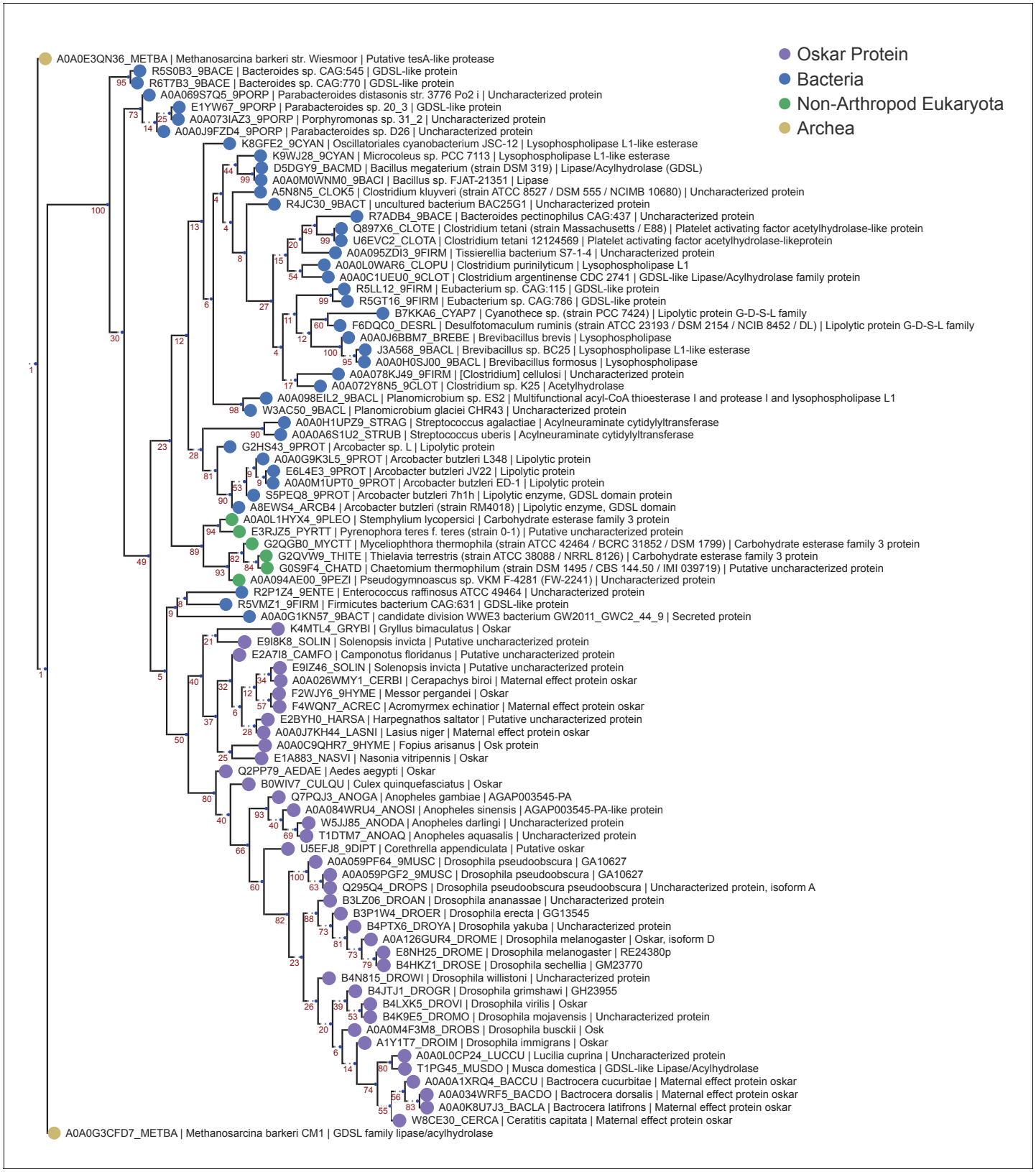

Figure 2—figure supplement 3. OSK Domain RaxML MUSCLE Tree. Phylogenetic tree of the HMMER sequences retrieved from the UniProt database using the OSK alignment HMM model. The top 95 hits were selected for phylogenetic analysis, and the only five non-Oskar eukaryotic sequences found to be a match were added to the alignment manually. The resulting 100 sequences were aligned using MUSCLE with default settings. The sequences

Figure 2—figure supplement 3 continued on next page

Figure 2—figure supplement 3 continued

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 Materials and 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.

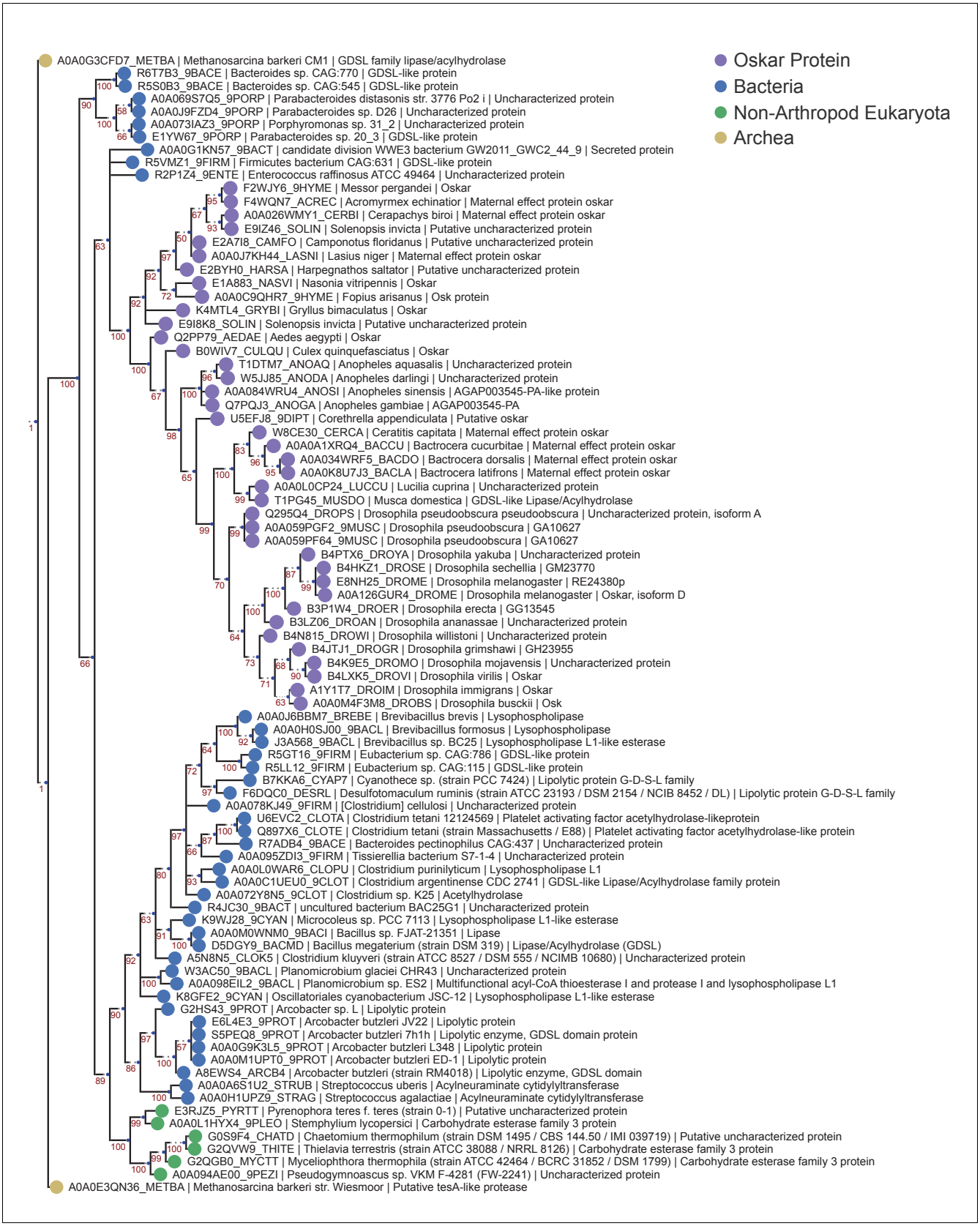

Figure 2—figure supplement 4. OSK Domain Bayesian MUSCLE Tree. Phylogenetic tree of the HMMER sequences hit on the UniProt database using the OSK alignment HMM model. 87 sequences were chosen for analysis as described for ***Figure 2—figure supplement 3***.The tree was created using

Figure 2—figure supplement 4 continued on next page

Figure 2—figure supplement 4 continued

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 Materials and 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.

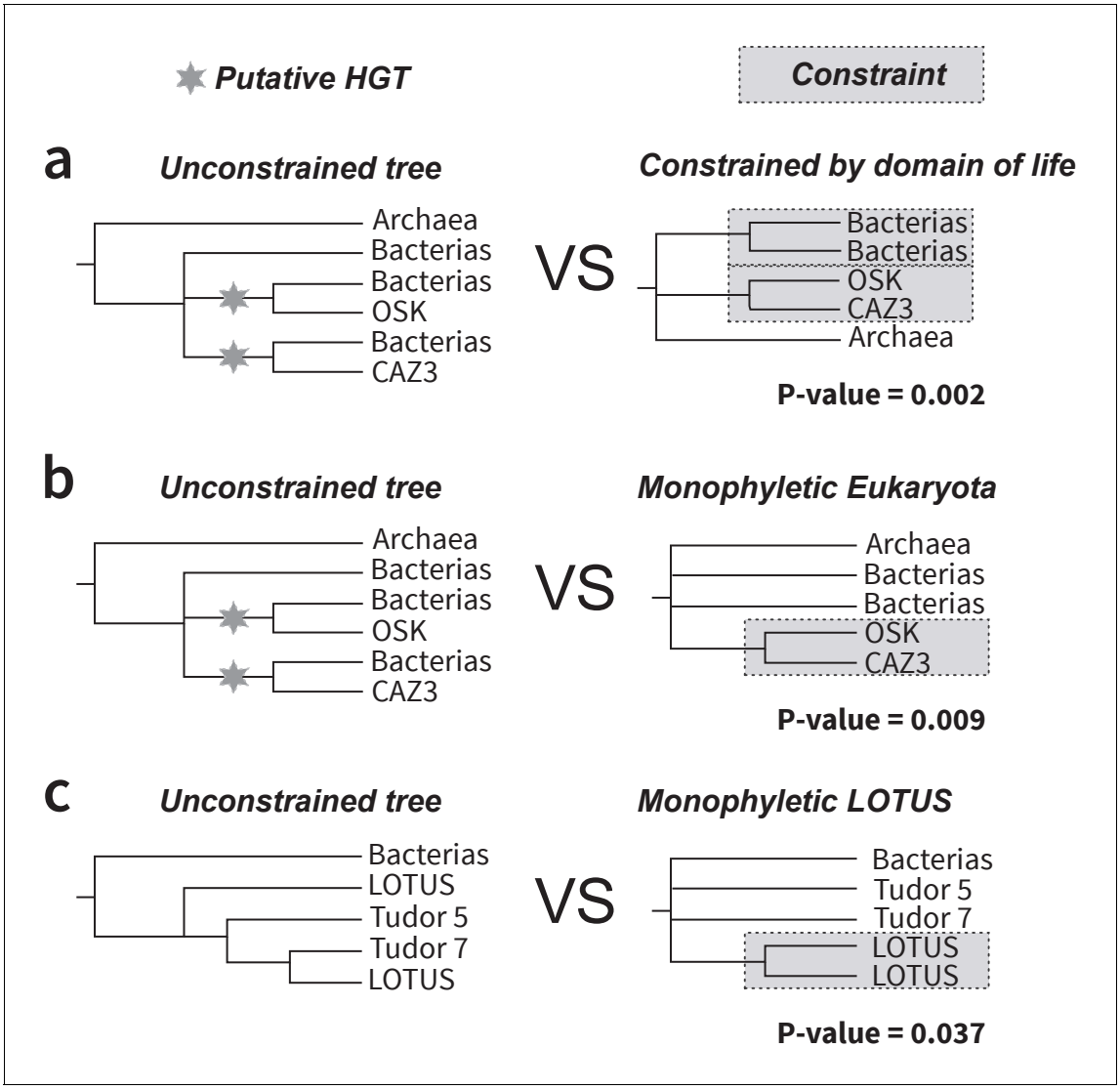

Figure 2—figure supplement 5. SOWHAT constrained trees and results. Two trees constrained by alternative relationships that would be expected under vertical transmission of sequences were designed and tested against our result supporting a putative HGT event of the OSK domain. (**a**) The first tree (right) is constrained by domain of life, requiring bacterial and eukaryotic sequences to be monophyletic, and disallowing sister group relationships of subsets of eukaryotic sequences and bacterial sequences. Our unconstrained tree topology (left) outperformed this topology with a p-value of 0.002 (95% confidence interval upper: 0.007 lower: 0.0002). (**b**) The second tree requires monophyly of Eukaryota. Our unconstrained tree topology (left) outperformed this topology with a p-value of 0.009 (95% confidence interval upper: 0.017 lower: 0.004). (**c**) The third tree tested whether the LOTUS domain split observed in the tree generated with the MUSCLE alignment was significantly different from a tree where the LOTUS sequences formed a monophyly. The unconstrained tree (left) outperformed this topology with a p-value of 0.037 (95% confidence interval upper: 0.05 lower: 0.026).

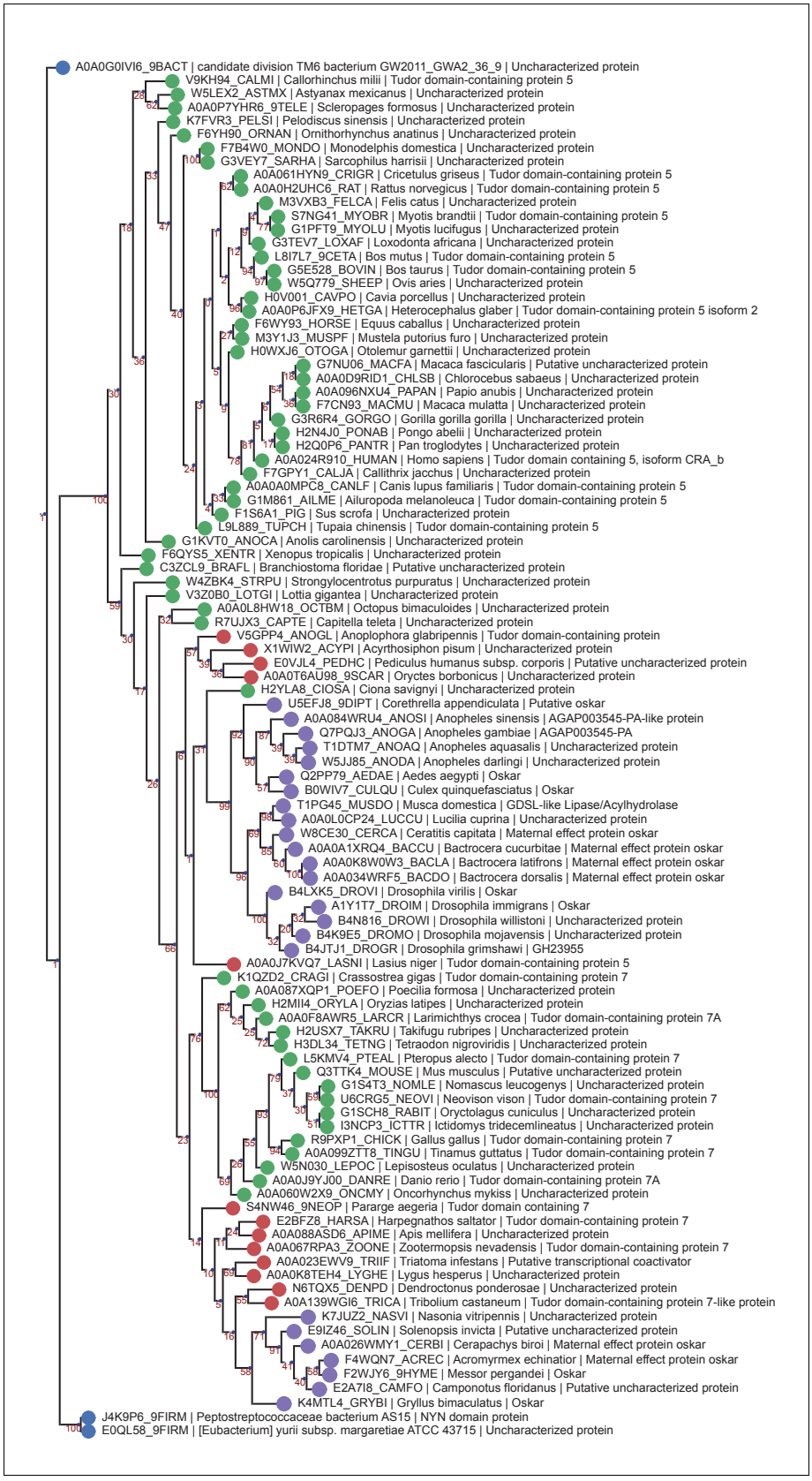

Figure 2—figure supplement 6. LOTUS Domain RaxML PRANK Tree. Phylogenetic tree of the same sequences used for the previous LOTUS trees. The sequences were aligned using PRANK and the tree generated with RaxML

Figure 2—figure supplement 6 continued on next page

Figure 2—figure supplement 6 continued

as described in *Phylogenetic Analysis Based on PRANK alignment*. 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.

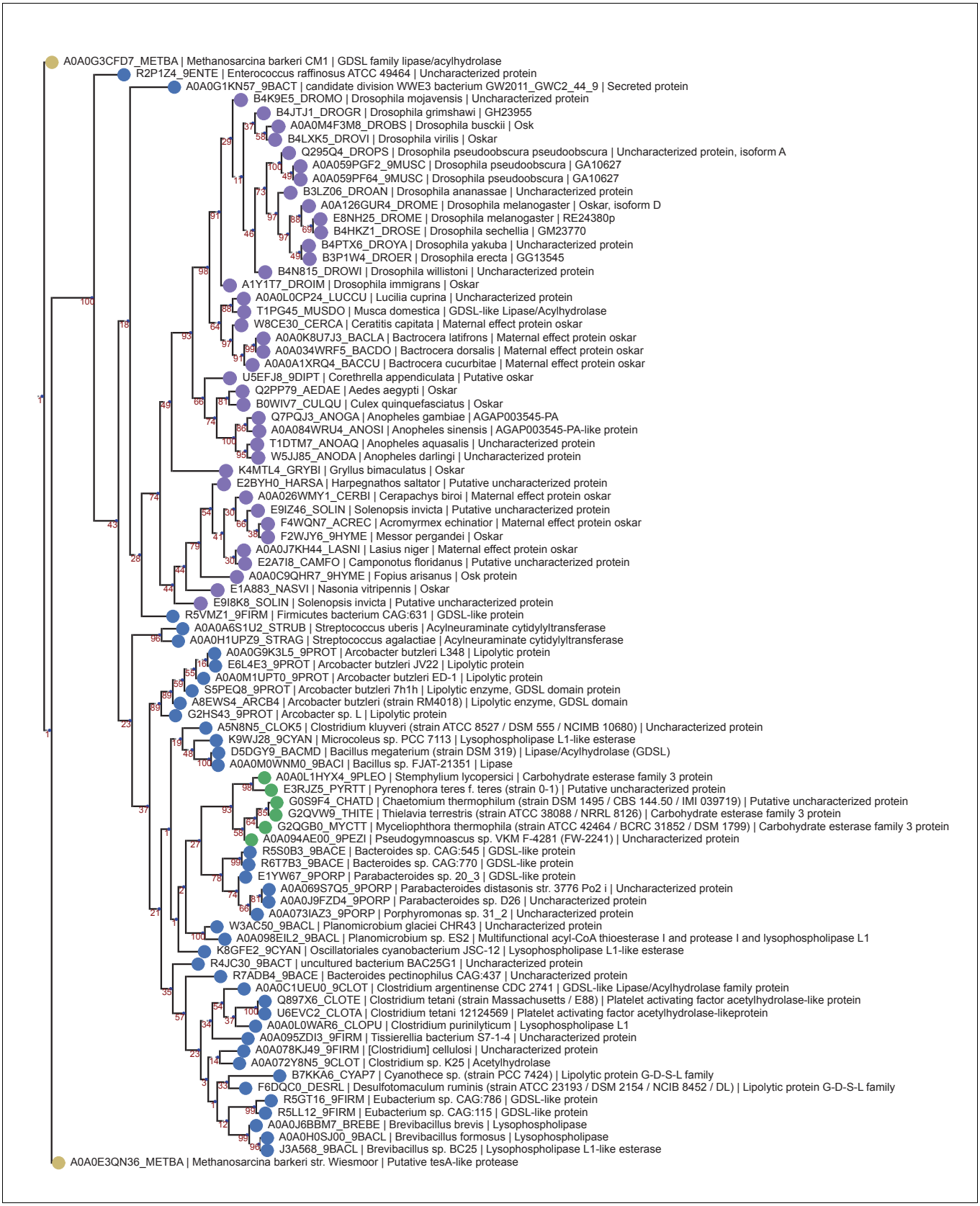

Figure 2—figure supplement 7. OSK Domain RaxML PRANK Tree. Phylogenetic tree of the same sequences used for the previous OSK trees. The sequences were aligned using PRANK and the tree generated with RaxML as described in *Phylogenetic Analysis Based on PRANK alignment*.

Figure 2—figure supplement 7 continued on next page

Figure 2—figure supplement 7 continued

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.

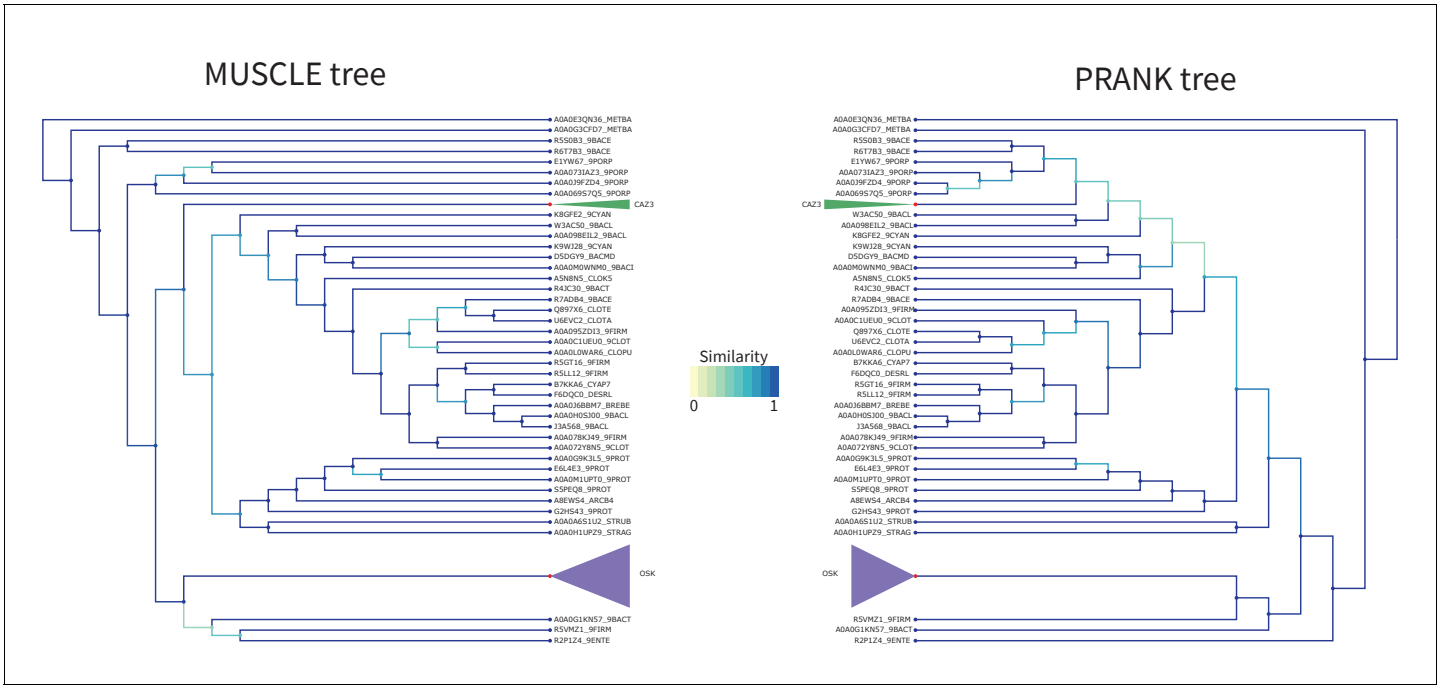

Figure 2—figure supplement 8. OSK Tree PRANK Comparison. Comparison of the tree obtained with RaxML starting from the MUSCLE alignment (left) versus the PRANK alignment (right) for the OSK domain. Similarity scores for the branching events are color coded from yellow to blue (see figure color bar legend). The OSK (purple) clade and CAZ3 (green) clade have been colored and compacted for readability as they do not have any internal branching changes. Node color is blue if the leaf is a sequence, and red if this is a compacted group of sequences.

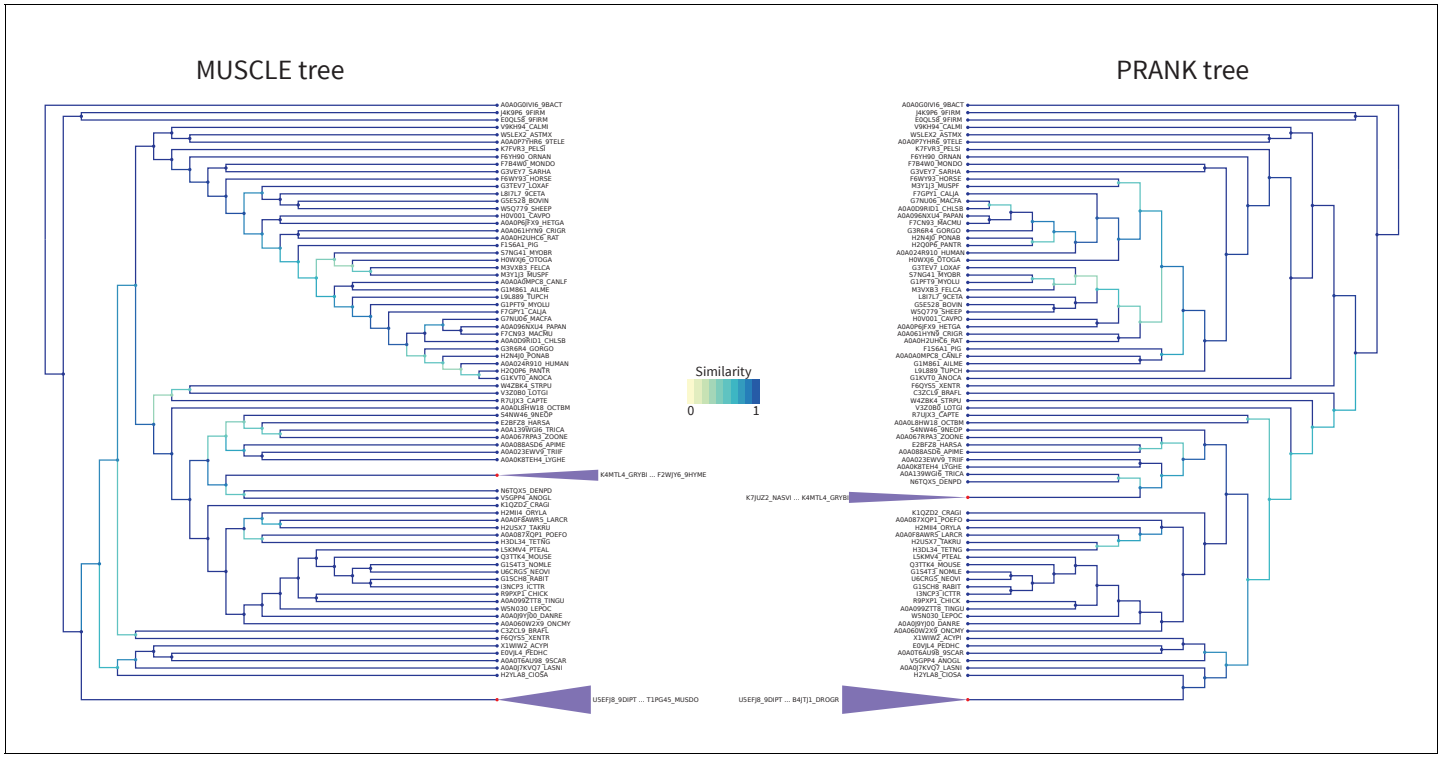

Figure 2—figure supplement 9. LOTUS Tree PRANK Comparison. Comparison of the tree obtained with RaxML starting from the MUSCLE alignment (left) versus the PRANK alignment (right) for the LOTUS domain. Similarity scores for the branching events are color coded from yellow to blue (see figure color bar legend). The LOTUS (purple) clades have been colored and compacted for readability as they do not have any internal branching changes. Node color is blue if the leaf is a sequence, and red if this is a compacted group of sequences.

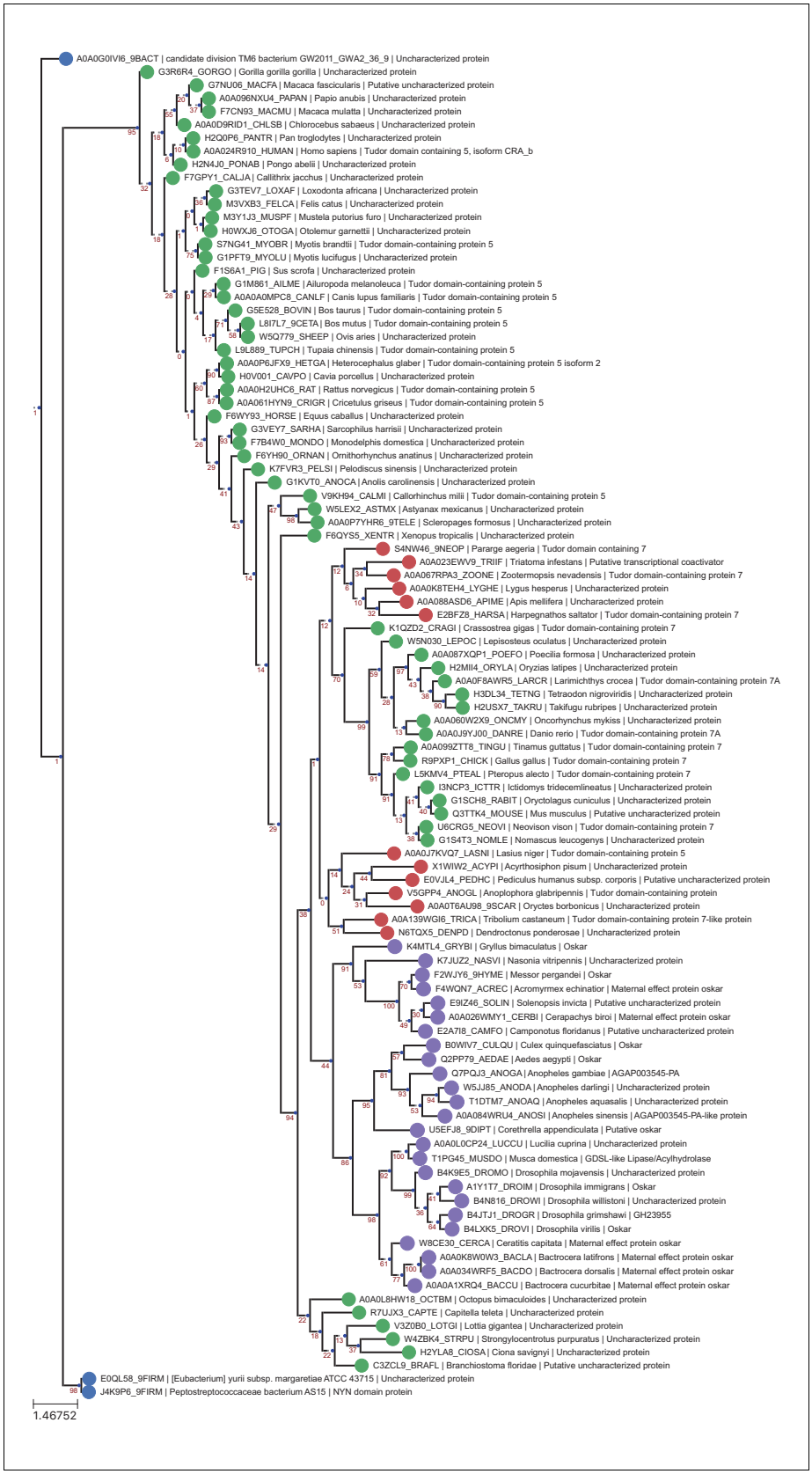

Figure 2—figure supplement 10. LOTUS Domain RaxML T-Coffee Tree. Phylogenetic tree of the same sequences used for the previous LOTUS trees. The sequences were aligned using T-Coffee and the tree

Figure 2—figure supplement 10 continued on next page

Figure 2—figure supplement 10 continued

generated with RaxML as described in *Phylogenetic Analysis Based on* T-Coffee *alignment*. 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.

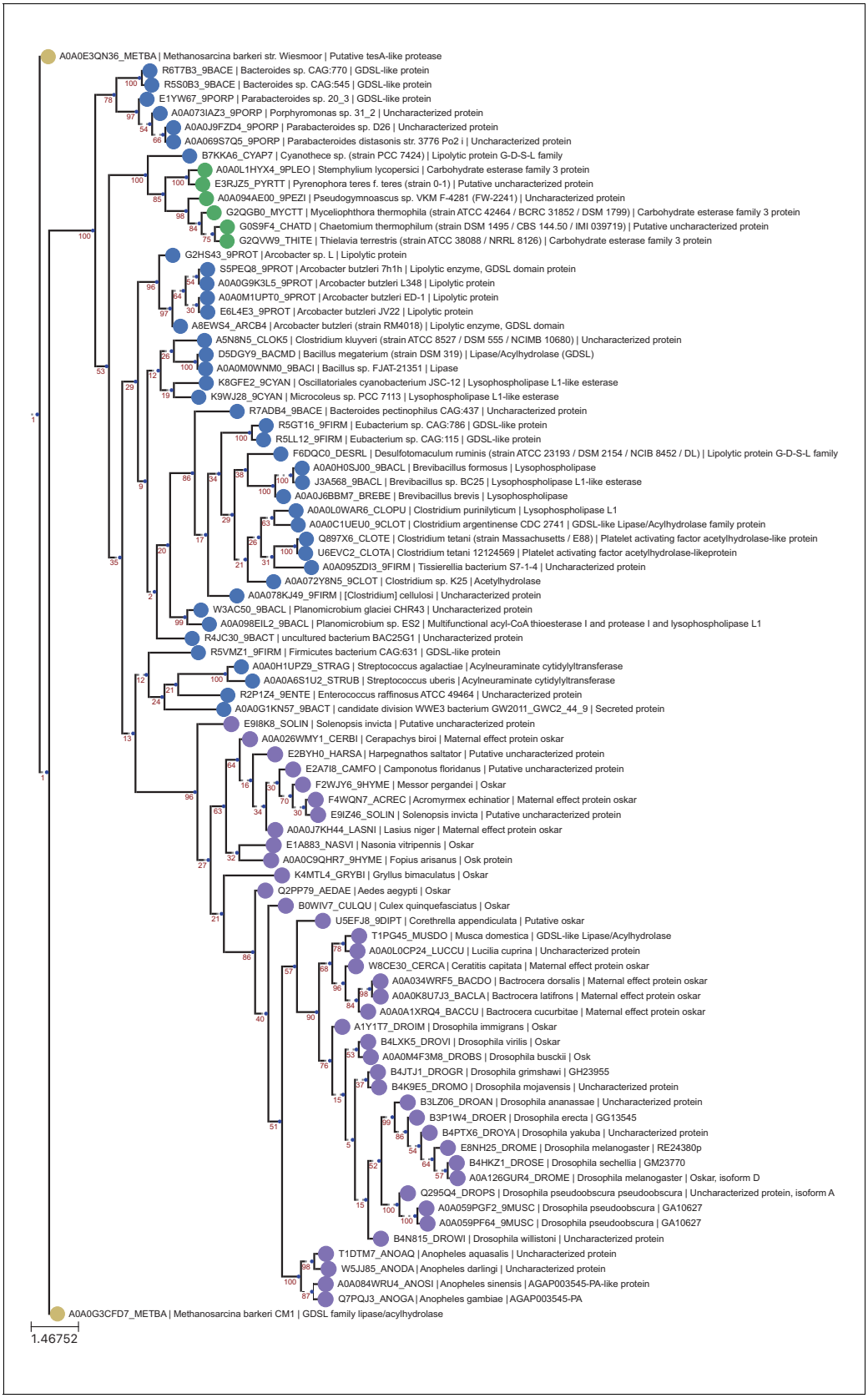

Figure 2—figure supplement 11. OSK Domain RaxML T-Coffee Tree. Phylogenetic tree of the same sequences used for the previous OSK trees. The sequences were aligned using T-Coffee and the tree generated with RaxML as described in *Phylogenetic Analysis Based on* T-Coffee *alignment.*

Figure 2—figure supplement 11 continued on next page

Figure 2—figure supplement 11 continued

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.

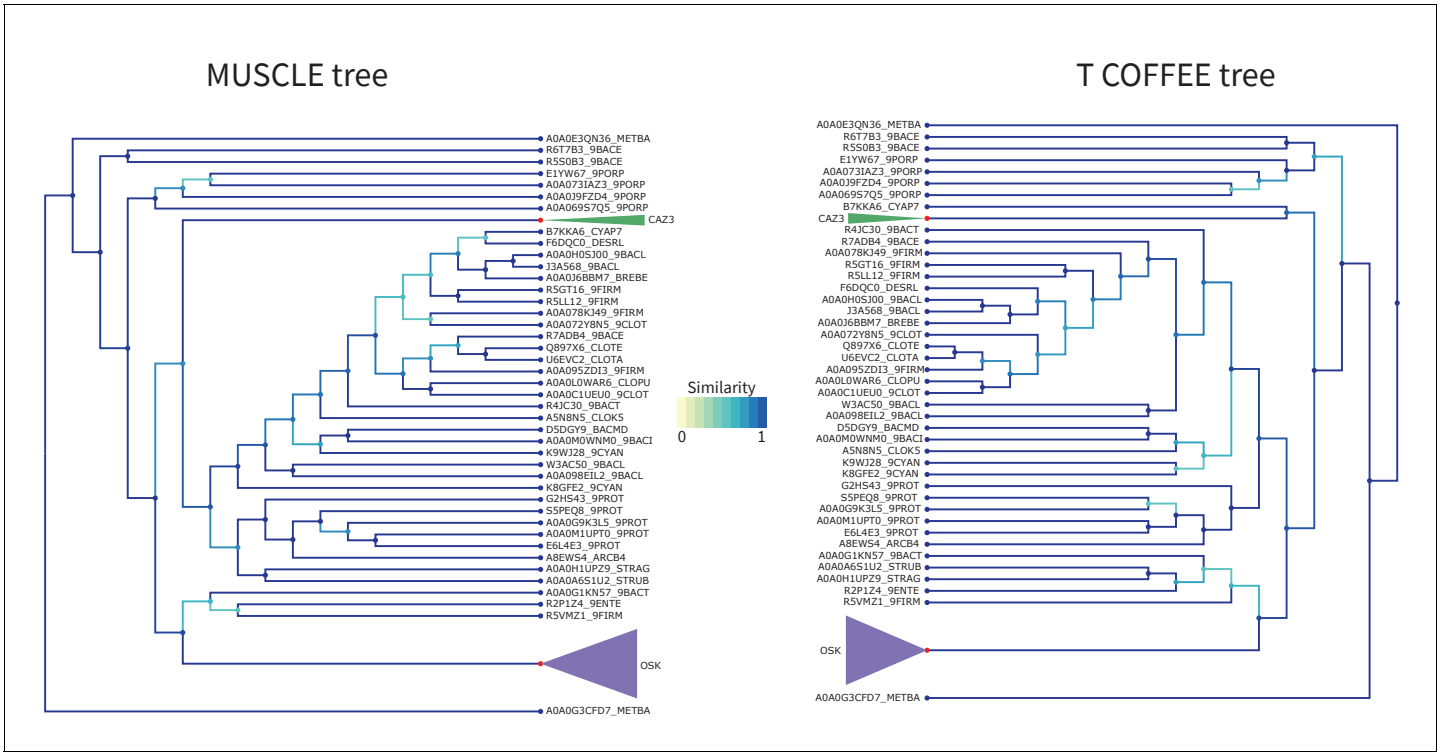

Figure 2—figure supplement 12. OSK Tree T-Coffee Comparison. Comparison of the tree obtained with RaxML starting from the MUSCLE alignment (left) versus the T-Coffee alignment (right) for the OSK domain. Similarity scores for the branching events are color coded from yellow to blue (see figure color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

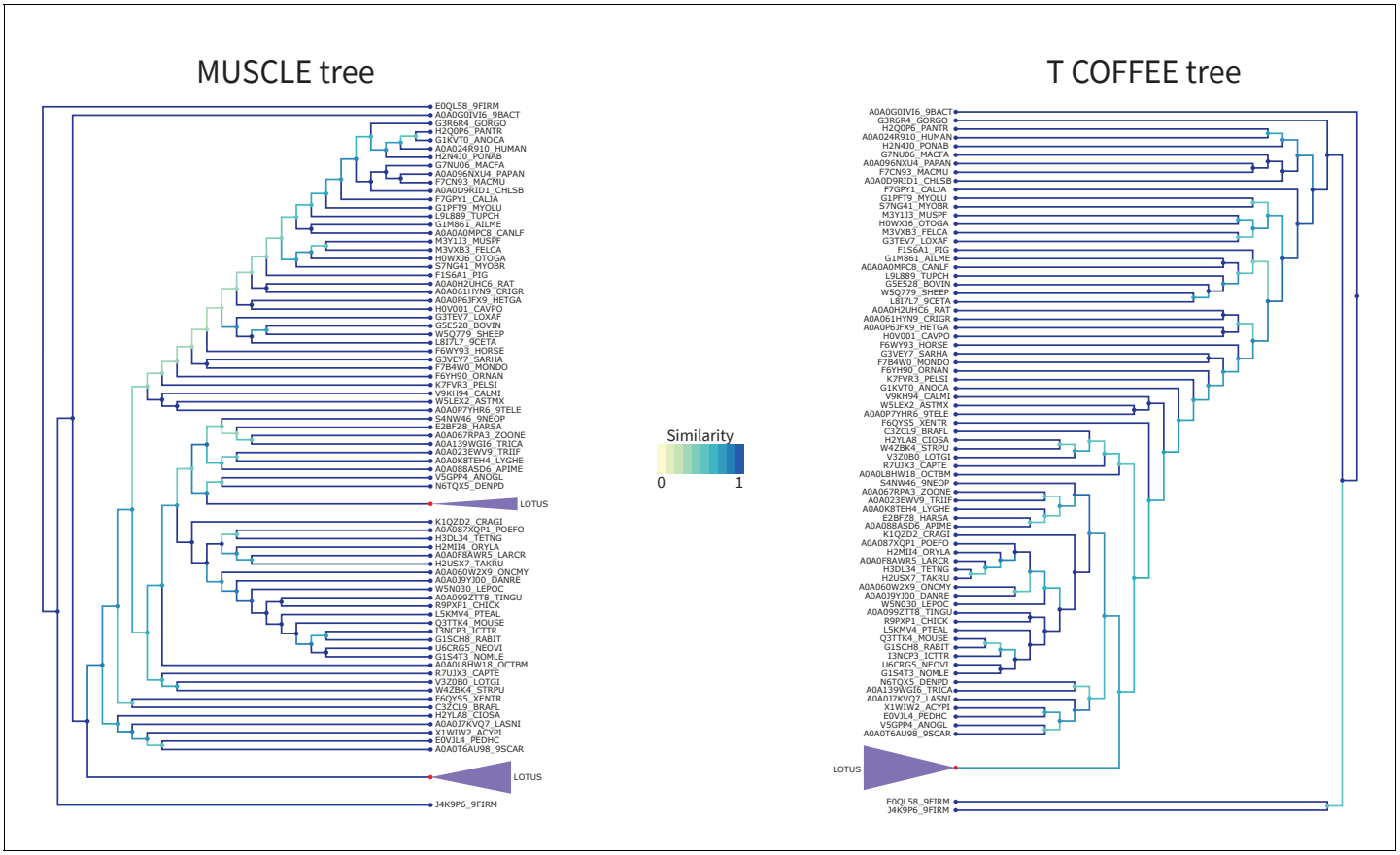

Figure 2—figure supplement 13. LOTUS Tree T-Coffee Comparison. Comparison of the tree obtained with RaxML starting from the MUSCLE alignment (left) versus the T-Coffee alignment (right) for the LOTUS domain. Similarity scores for the branching events are color coded from yellow to blue (see figure color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

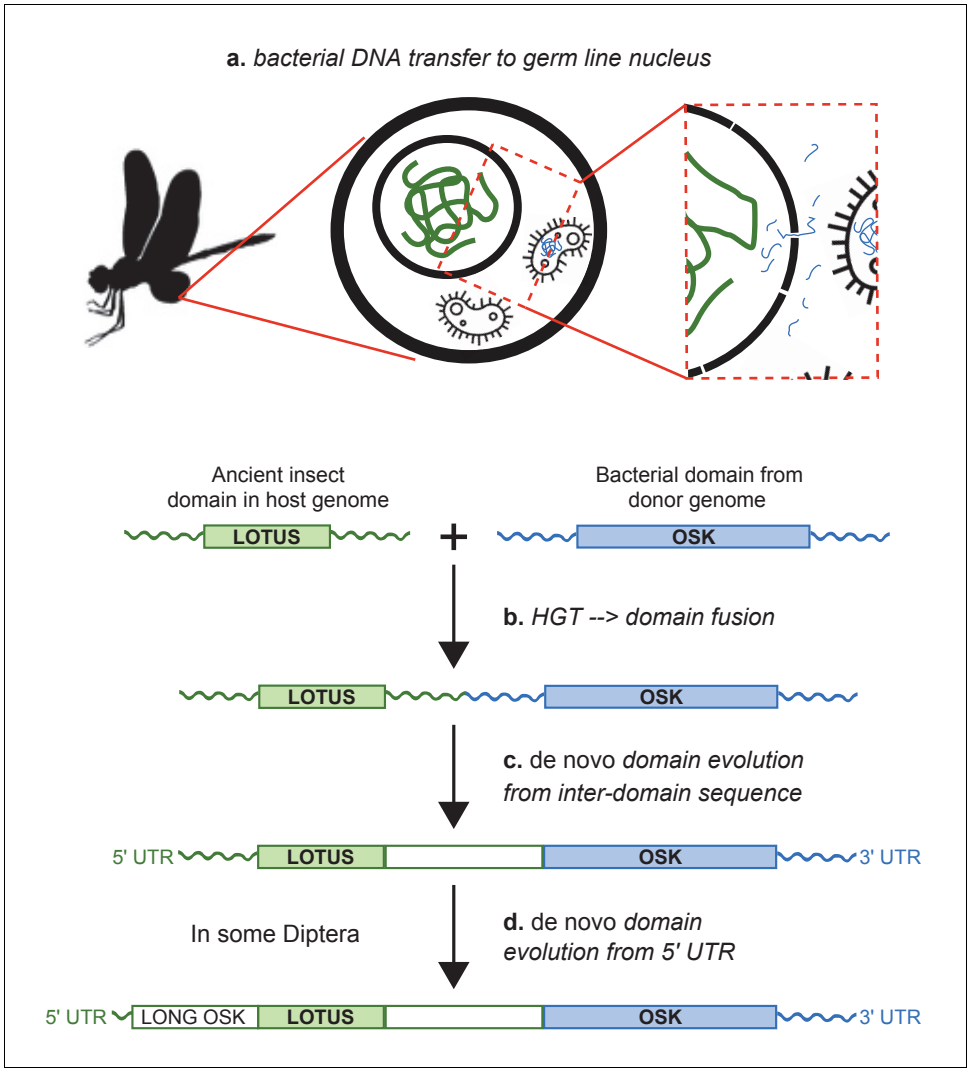

Figure 3. Hypothesis for the origin of *oskar*. Integration of the OSK domain close to a LOTUS domain in an ancestral insect genome. (**a**) DNA containing a GDSL-like domain from an endosymbiotic germ line bacterium is transferred to the nucleus of a germ cell in an insect common ancestor. (**b**) DNA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