# 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 new genes. New 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 Oskar sequences suggests that Oskar arose through a novel gene formation history involving fusion of eukaryotic and prokaryotic sequences. This work shows that highly unusual gene origin processes can birth novel genes that can facilitate evolution of novel developmental mechanisms.

**Main Text:**

**Introduction:** Heritable variation is the raw material of evolutionary change. Genetic variation can arise from mutation and gene duplication of existing genes (*1*), or through *de novo* processes (*2*), but the extent to which such novel, or “orphan” genes participate significantly in the evolutionary process is unclear. Mutation of existing cis-regulatory (*3*) or protein coding regions (*4*) can drive evolutionary change in developmental processes. However, recent studies in animals and fungi suggest that new genes can also drive phenotypic change (*5*). Although counterintuitive, novel genes may be integrating continuously into otherwise conserved gene networks, with a higher rate of partner acquisition than subtler variations on preexisting genes (*6*). Moreover, in humans and fruit flies, a large proportion of new genes are expressed in the brain, suggesting their participation in the evolution of major organ systems (*7, 8*). However, while next generation sequencing has improved their discovery, the developmental and evolutionary significance of new genes remains understudied.

The mechanism of formation of a new gene may have implications for its function. New genes that arise by duplication, thus possessing the same biophysical properties as their parent genes, have innate potential to participate in preexisting cellular and molecular mechanisms (*1*). However, orphan genes lacking sequence similarity to existing genes must form novel functional molecular relationships with extant genes, in order to persist in the genome. When such genes arise by introduction of foreign DNA into a host genome through horizontal gene transfer (HGT), they may introduce novel, already functional sequence information into a genome. Whether genes created by HGT show a greater propensity to contribute to or enable novel processes is unclear. Endosymbionts in the host germ line cytoplasm (germ line symbionts) could increase the occurrence of evolutionarily relevant HGT events, as foreign DNA integrated into the germ line genome is transferred to the next generation. HGT from bacterial endosymbionts into insect genomes appears widespread, involving transfer of metabolic genes or even larger genomic fragments to the host genome (*9*).

Here we examined the evolutionary origins of the *oskar* (*osk*) gene, long considered a novel gene that evolved to be indispensable for insect reproduction (*10*). First discovered in *Drosophila melanogaster* (*11*), *osk* is necessary and sufficient for assembly of germ plasm, a cytoplasmic determinant that specifies the germ line in the embryo. Germ plasm-based germ line specification appears derived within insects, confined to insects that undergo metamorphosis (Holometabola) (*12, 13*). Initially thought exclusive to Diptera (flies and mosquitoes), its discovery in a wasp, another holometabolous insect with germ plasm (*14*), led to the hypothesis that *oskar* originated as a novel gene at the base of the Holometabola approximately 300 Mya, facilitating the evolution of insect germ plasm as a novel developmental mechanism (*14*). However, its subsequent discovery in a cricket (*12*), a basally branching insect without germ plasm (*15*), implied that *osk* was instead at least 50 My older, and that its germ plasm role was derived rather than ancestral (*16*). Despite its orphan gene status, *osk* plays major developmental roles, interacting with the products of many genes highly conserved across animals (*10, 17, 18*). *osk* thus represents an example of a new gene that not only functions within pre-existing gene networks in the nervous system (*12*), but has also evolved into the only animal gene known to be both necessary and sufficient for germ line specification (*19, 20*).

The evolutionary origins of this remarkable gene are unknown. Osk contains two biophysically conserved domains, an N-terminal LOTUS domain and a C-terminal hydrolase-like domain called OSK (*17, 21*) (Fig. 1a). A BLASTp search using the full-length *D. melanogaster osk* sequence as a query yielded only other holometabolous *osk* genes (E-value < 0.01), or hits for the LOTUS or OSK domains (E-value <10) (Supplementary files: BLAST search results). This suggested that full length *osk* was unlikely to be a duplication of any other known gene, prompting us to perform a BLASTp search on each conserved Osk protein domain individually. Strikingly, in our BLASTp search, we recovered no eukaryotic sequences that resembled the OSK domain (E-value < 10) (Supplementary files: BLAST search results).

**Results:** To understand this anomaly, we built an alignment of 95 Oskar sequences (Supplementary files: Alignments>OSKAR_FINAL.fasta) and used a custom iterative HMMER sliding window search tool to compare each domain with protein sequences from all domains of life. Sequences most similar to the LOTUS domain were almost exclusively eukaryotic sequences (Supplementary Table 3). In contrast, those most similar to the OSK domain were bacterial, specifically sequences similar to SGNH-like hydrolases (*17, 21*) (Pfam Clan: SGNH_hydrolase - CL0264; Supp. Table 4; Fig. 1b). To visualize their relationships, we graphed the sequence similarity network for the sequences of these domains and their closest hits. We observed that the majority of LOTUS domain sequences clustered within eukaryotic sequences (Fig. 1c). In contrast, OSK domain sequences formed an isolated cluster, a small subset of which formed a connection to bacterial sequences (Fig. 1d). These data are consistent with a previous suggestion, based on BLAST results (*14*), that HGT from a bacterium into an ancestral insect genome may have contributed to the evolution of *osk*. However, this possibility was not adequately addressed by previous analyses, which were based on alignments of full length Osk containing only eukaryotic sequences as outgroups (*12*). To rigorously test this hypothesis, we therefore performed phylogenetic analyses of the two domains independently.

A finding that LOTUS sequences branch within eukaryotes, while OSK sequences branch within bacteria, would provide support for the HGT hypothesis.

Both Maximum likelihood and Bayesian approaches confirmed this prediction (Fig. 2). As expected, LOTUS sequences from Osk proteins were related to other eukaryotic LOTUS domains, to the exclusion of the only three bacterial sequences with sufficient similarity to include in the analyses (Figs. 2a, S1, S2; see Methods and Supplemental Text). In contrast, OSK domain sequences branched within bacterial sequences (Fig. 2b, S3, S4). Importantly, OSK sequences did not simply form an outgroup to bacterial sequences. Instead, they formed a well-supported clade nested within bacterial GDSL-like lipase sequences. The majority of these bacterial sequences were from the Firmicutes, a bacterial phylum known to include insect germline symbionts (*22, 23*). All other sequences from classified bacterial species, including a clade branching basally to all other sequences, belonged either to the Bacteroidetes or to the Proteobacteria. Members of both of these phyla are also known germline symbionts of insects (*9, 24*) and other arthropods (*25*). In sum, the distinct phylogenetic relationships of the two domains of Oskar are consistent with a bacterial origin for the OSK domain. Further, the specific bacterial clades close to OSK suggest that an ancient arthropod germ line endosymbiont could have been the source of a GDSL-like sequence that was transferred into an ancestral insect genome, and ultimately gave rise to the OSK domain of *oskar*.

We then asked if two additional sequence characteristics, GC3 content and codon use, were consistent with distinct domain of life origins for the two Oskar domains (*26*). Under our hypothesis, the HGT event that contributed to *oskar*'s formation would have occurred at least 480 Mya, in a common insect ancestor (*27*). We reasoned that if evolutionary time had not completely erased such signatures from the putative bacterially donated sequence (OSK), we might detect differences from the LOTUS domain, and from the host genome. Thus, we performed a parametric analysis of these parameters for 17 well annotated insect genomes (Supplementary Table 5). To quantify the null hypothesis, we calculated an "Intra-Gene distribution" for all genes in the genome, which showed a linear correlation between codon use in the 5' and 3' halves of a given gene. In contrast, the codon use between the LOTUS and OSK domains did not follow this correlation for nearly all measures of codon use (Fig. 3a, 3b, S5). For each genome, we then calculated the residuals of the Intra-Gene distribution and the LOTUS-OSK pair. Pooling the residuals together revealed that the GC3 content was drastically different between the LOTUS and OSK domains, compared to what would be expected within an average gene in that genome (Fig. 3c). Finally, to quantify the codon use difference, we compared the cosine distance in codon use between the LOTUS and OSK domains, with that of the Inter-Gene and Intra-Gene distributions. We found that the LOTUS-OSK distance was closer to that measured between two different, random genes, than between two parts of the same gene (Inter-Gene and Intra-Gene distributions, respectively; Fig. 3d). In sum, whereas most genes have similar codon use across all regions of their coding sequence, the OSK and LOTUS domains of *oskar* use codons in different ways. Together with the phylogenetic and sequence similarity evidence presented above, these analyses are consistent with an HGT origin for the OSK domain (Fig. 4).

**Discussion:** While multiple mechanisms can give rise to new genes, HGT is arguably among the least well understood, as it involves multiple genomes and ancient biotic interactions between donor and host organisms that are often difficult to reconstruct. In the case of *oskar*, however, the fact that both germline symbionts (*28*) and HGT events (*9*) are widespread in insects, provides a plausible biological mechanism consistent with our hypothesis that fusion of eukaryotic and bacterial domain sequences led to the birth of this novel gene.

Once arisen, novel genes might be expected to disappear rapidly, given that pre-existing gene regulatory networks operated successfully without them (*1*). However, it is clear that new genes can evolve functional connections with existing networks, become essential (*29*), and in some cases lead to new functions (*30*) and contribute to phenotypic diversity (*5*). *oskar* plays multiple critical roles in insect development, from neural patterning (*12, 31*) to oogenesis (*32*). In the Holometabola, a clade of nearly one million extant species (*33*), *oskar*'s co-option to become necessary and sufficient for germ plasm assembly is likely the cell biological mechanism underlying the evolution of this derived mode of insect germ line specification (*12, 14, 16*). Our study thus provides evidence that HGT can not only introduce functional genes into a host genome, but also, by contributing sequences of individual domains, generate genes with entirely novel domain structures that may facilitate the evolution of novel developmental mechanisms.

## References

1. J. S. Taylor, J. Raes, Duplication and divergence: the evolution of new genes and old ideas. *Annual review of genetics* **38**, 615-643 (2004).
2. D. Tautz, T. Domazet-Loso, The evolutionary origin of orphan genes. *Nat. Rev. Genet.* **12**, 692-702 (2011).
3. P. J. Wittkopp, G. Kalay, Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat. Rev. Genet.* **13**, 59-69 (2011).
4. H. E. Hoekstra, J. A. Coyne, The locus of evolution: evo devo and the genetics of adaptation. *Evolution* **61**, 995-1016 (2007).
5. S. Chen, B. H. Krinsky, M. Long, New genes as drivers of phenotypic evolution. *Nat. Rev. Genet.* **14**, 645-660 (2013).
6. W. Zhang, P. Landback, A. R. Gschwend, B. Shen, M. Long, New genes drive the evolution of gene interaction networks in the human and mouse genomes. *Genome Biol.* **16**, 202 (2015).
7. Y. E. Zhang, P. Landback, M. Vibranovski, M. Long, New genes expressed in human brains: implications for annotating evolving genomes. *BioEssays* **34**, 982-991 (2012).
8. S. Chen *et al.*, Frequent recent origination of brain genes shaped the evolution of foraging behavior in Drosophila. *Cell Reports* **1**, 118-132 (2012).
9. J. C. Dunning Hotopp *et al.*, Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. *Science (New York, NY)* **317**, 1753-1756 (2007).
10. R. Lehmann, Germ Plasm Biogenesis--An Oskar-Centric Perspective. *Curr. Top. Dev. Biol.* **116**, 679-707 (2016).
11. R. Lehmann, C. Nüsslein-Volhard, Abdominal Segmentation, Pole Cell Formation, and Embryonic Polarity Require the Localized Activity of *oskar*, a Maternal Gene in *Drosophila*. *Cell* **47**, 144-152 (1986).
12. B. Ewen-Campen, J. R. Srouji, E. E. Schwager, C. G. Extavour, *oskar* Predates the Evolution of Germ Plasm in Insects. *Curr. Biol.* **22**, 2278-2283 (2012).
13. C. G. Extavour, M. E. Akam, Mechanisms of germ cell specification across the metazoans: epigenesis and preformation. *Development* **130**, 5869-5884 (2003).
14. J. A. Lynch *et al.*, The Phylogenetic Origin of *oskar* Coincided with the Origin of Maternally Provisioned Germ Plasm and Pole Cells at the Base of the Holometabola. *PLoS Genetics* **7**, e1002029 (2011).
15. B. Ewen-Campen, S. Donoughe, D. N. Clarke, C. G. Extavour, Germ cell specification requires zygotic mechanisms rather than germ plasm in a basally branching insect. *Curr. Biol.* **23**, 835-842 (2013).
16. E. Abouheif, Evolution: oskar Reveals Missing Link in Co-optive Evolution. *Curr. Biol.* **23**, R24-R25 (2012).
17. M. Jeske *et al.*, The Crystal Structure of the Drosophila Germline Inducer Oskar Identifies Two Domains with Distinct Vasa Helicase- and RNA-Binding Activities. *Cell Reports* **12**, 587-598 (2015).
18. M. Jeske, C. W. Muller, A. Ephrussi, The LOTUS domain is a conserved DEAD-box RNA helicase regulator essential for the recruitment of Vasa to the germ plasm and nuage. *Genes Dev.* **31**, 939-952 (2017).

19. A. Ephrussi, R. Lehmann, Induction of germ cell formation by *oskar*. *Nature* **358**, 387-392 (1992).

20. J. Kim-Ha, J. L. Smith, P. M. Macdonald, *oskar* mRNA is localized to the posterior pole of the *Drosophila* oocyte. *Cell* **66**, 23-35 (1991).

21. N. Yang *et al.*, Structure of Drosophila Oskar reveals a novel RNA binding protein. *Proc. Natl. Acad. Sci. USA* **112**, 11541-11546 (2015).

22. D. Wheeler, A. J. Redding, J. H. Werren, Characterization of an ancient lepidopteran lateral gene transfer. *PLoS ONE* **8**, e59262 (2013).

23. S. T. Chepkemoi *et al.*, Identification of Spiroplasmainsolitum symbionts in Anopheles gambiae. *Wellcome Open Research* **2**, 90 (2017).

24. E. Zchori-Fein, S. J. Perlman, S. E. Kelly, N. Katzir, M. S. Hunter, Characterization of a 'Bacteroidetes' symbiont in Encarsia wasps (Hymenoptera: Aphelinidae): proposal of 'Candidatus Cardinium hertigii'. *International Journal of Systematic and Evolutionary Microbiology* **54**, 961-968 (2004).

25. E. Zchori-Fein, S. J. Perlman, Distribution of the bacterial symbiont Cardinium in arthropods. *Mol. Ecol.* **13**, 2009-2016 (2004).

26. T. Tuller, Codon bias, tRNA pools and horizontal gene transfer. *Mobile Genetic Elements* **1**, 75-77 (2011).

27. B. Misof *et al.*, Phylogenomics resolves the timing and pattern of insect evolution. *Science* **346**, 763-767 (2014).

28. K. Bourtzis, T. A. Miller, Eds., *Insect Symbiosis*, (CRC Press, Boca Raton, FL, 2006), vol. 3, pp. 304.

29. S. Chen, Y. E. Zhang, M. Long, New genes in *Drosophila* quickly become essential. *Science* **330**, 1682-1685 (2010).

30. G. Cornelis *et al.*, Ancestral capture of syncytin-Car1, a fusogenic endogenous retroviral envelope gene involved in placentation and conserved in Carnivora. *Proc. Natl. Acad. Sci. USA* **109**, E432-441 (2012).

31. X. Xu, J. L. Brechbiel, E. R. Gavis, Dynein-Dependent Transport of nanos RNA in Drosophila Sensory Neurons Requires Rumpelstiltskin and the Germ Plasm Organizer Oskar. *The Journal of Neuroscience* **33**, 14791-14800 (2013).

32. A. Jenny *et al.*, A translation-independent role of oskar RNA in early Drosophila oogenesis. *Development* **133**, 2827-2833 (2006).

33. J. A. Rees, K. Cranston, Automated assembly of a reference taxonomy for phylogenetic data synthesis. *Biodiversity Data Journal* **5**, e12581 (2017).

34. J. A. Gerlt *et al.*, Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): A web tool for generating protein sequence similarity networks. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* **1854**, 1019-1037 (2015).

**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 Materials:**

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

Supplementary figures

- Figure S1: LOTUS Domain RaxML Tree.
- Figure S2: LOTUS Domain Bayesian Tree.
- Figure S3: OSK Domain RaxML Tree.
- Figure S4: OSK Domain Bayesian Tree.
- Figure S5: AT3/GC3 correlations between the LOTUS and OSK domains.
- Figure S6: A3/T3/G3/C3 correlations between the LOTUS and OSK domains.

Supplementary tables

- Table S1: List of genomes and transcriptomes used for automated *oskar* search.
- Table S2: List of *oskar* sequences used in the final alignment.
- Table S3: List of sequences used for phylogenetic analysis of the LOTUS domain.
- Table S4: List of sequences used for phylogenetic analysis of the OSK domain.
- Table S5: List of genomes analyzed for codon use.

1. Supplementary Discussion (Blondel_Jones_Extavour_HGT_HGT_Paper_SuppInfo_V4_181108.docx)
2. Supplementary References (Blondel_Jones_Extavour_HGT_HGT_Paper_SuppInfo_V4_181108.docx)
3. Folder titled "Supplementary Information Files" containing the following sub-folders
    a. Supplementary Information Files>Alignments
        i. *All sequences identified and analyzed in this study, in FASTA format and with corresponding Alignments*
    b. Supplementary Information Files>BLAST search results
        i. *Results of BLASTP searches with full length Oskar, OSK or LOTUS domains as queries*
    c. Supplementary Information Files>Data

        i. *Necessary files for running the different ipython notebooks:*

            1. *Taxonomy: Conversion table for UniProt ID to taxon information. (uniprot_ID_taxa.tsv )*
            2. *Codon_Genes: Contains the measured codon frequency for the different genomes studied as .csv or .tsv files (organism_name.csv/ tsv), along with the DNA sequences of LOTUS and OSK domains used in the codon use analysis (LOTUS_Seqeuences.gb and SGNH_Seqeuences.gb)*
            3. *Trees: Contains the tree files obtained from RaxML and MrBayes phylogenetic analyses of the OSK and LOTUS domains.*

    d. Supplementary Information Files>HMM

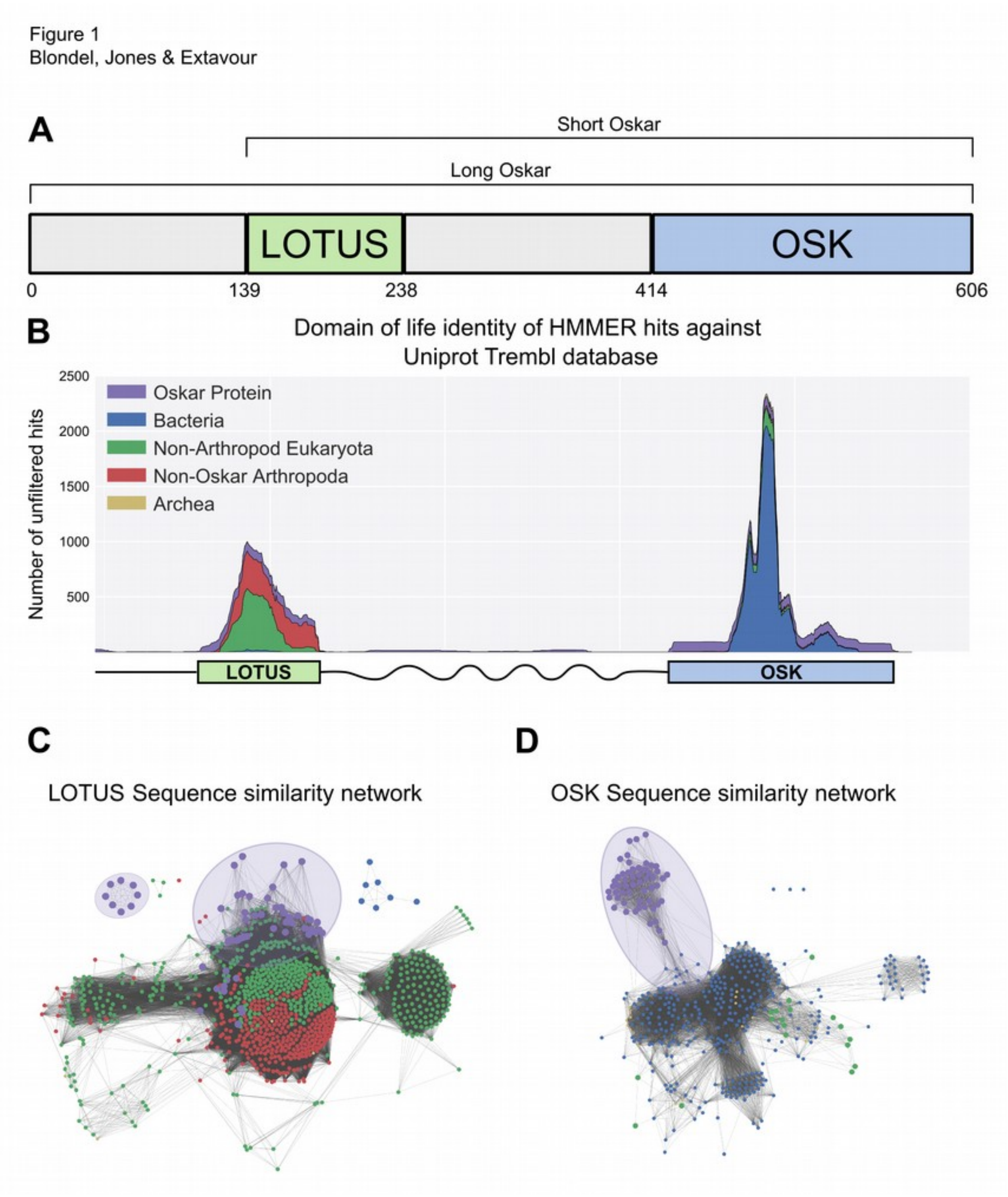
        i. *HMM models used for iterative searching for sequences similar to full-length Oskar, LOTUS and OSK domains*

    e. Supplementary Information Files>Scripts

        i. *All custom scripts used to implement the analysis pipelines described.*

2. Supplementary Information Files>Tables

    **a.** Supplementary Tables S1-S5 describing databases searched/analyzed and all search results; Legends in Blondel_Jones_Extavour_HGT_HGT_Paper_SuppInfo_V4_181108.docx



299

**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 3' 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 Methods: Iterative HMMER search of OSK and LOTUS domains), stacked on top of each other. **c, d** EFI-EST[34]-generated graphs of the sequence similarity network of the LOTUS (**c**) and OSK (**d**) domains of Oskar. 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.

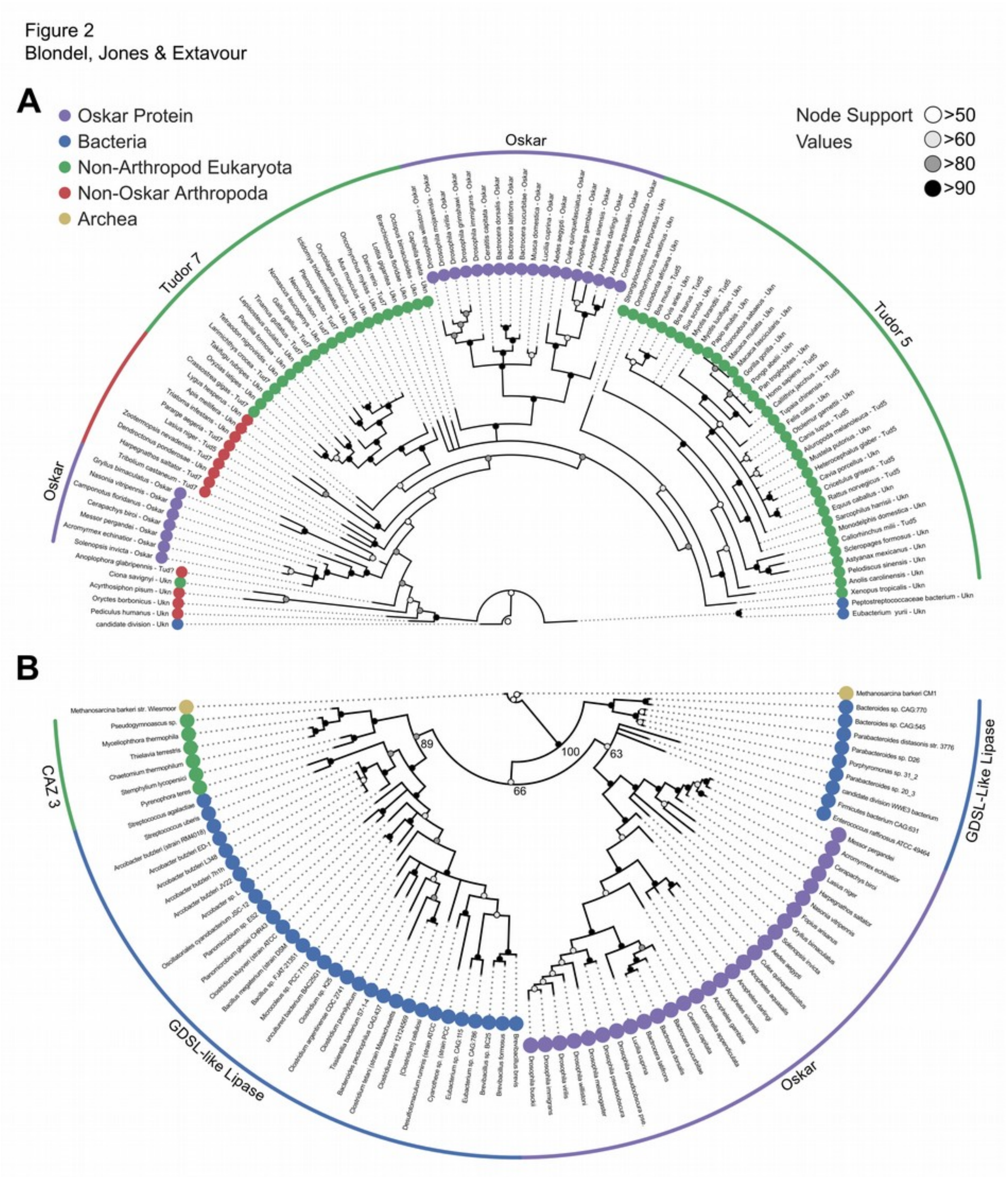


**Figure 2**. **Phylogenetic analysis of the LOTUS and OSK domains. a**, Bayesian consensus tree for the LOTUS domain. Three major LOTUS-containing protein families are represented within the tree: Tudor 5, Tudor 7, and Oskar. Oskar LOTUS domains form two clades, one containing only dipterans and one 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 Extended Data Figures S1-4.

Figure 3
Blondel, Jones & Extavour

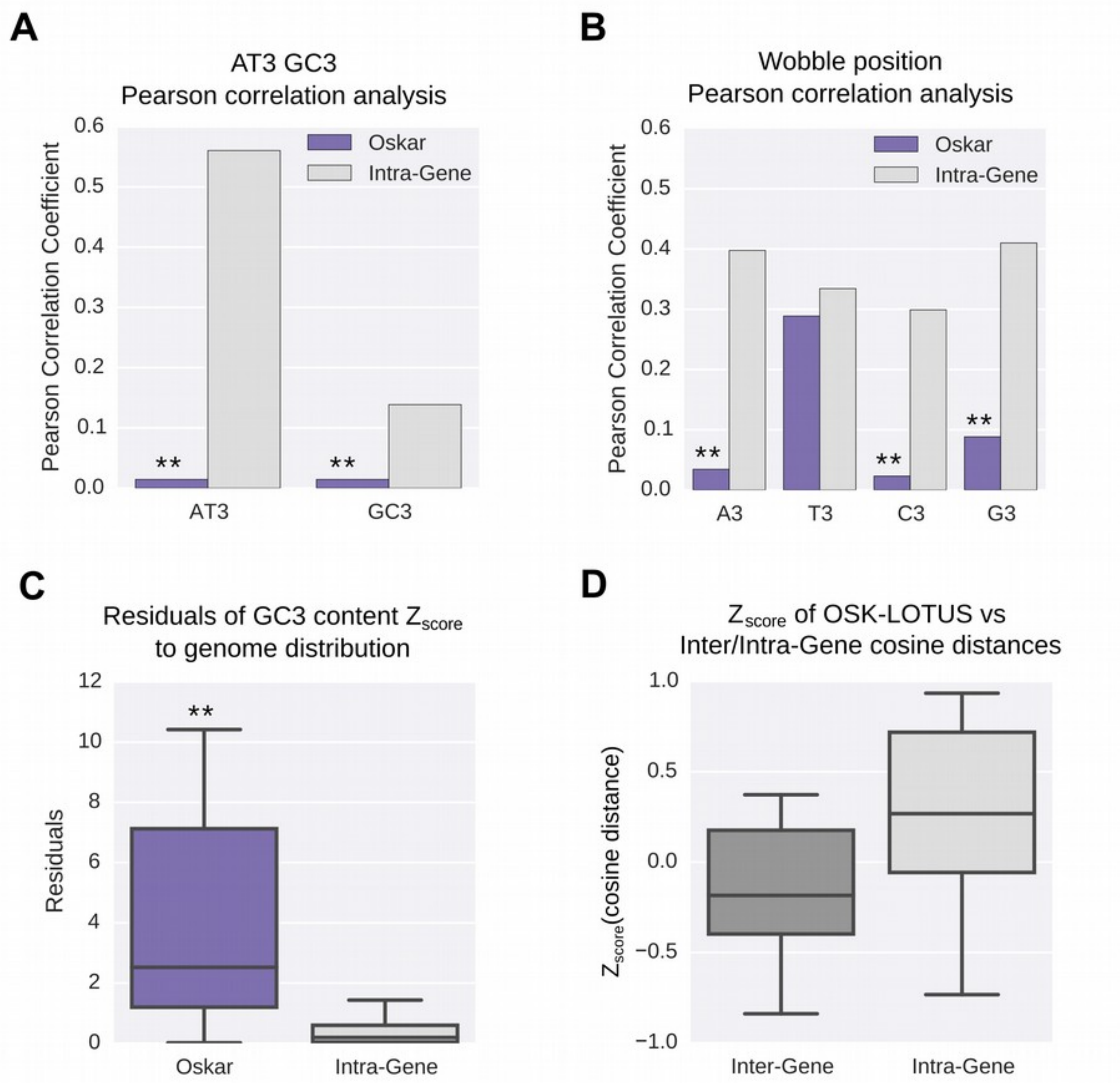


**Figure 3. Parametric analysis of codon use for the LOTUS and OSK domains. a**, Pearson correlation analysis of AT3 and GC3 content for Oskar vs other genes. AT3 and GC3 content are correlated across the sequence of a gene for all genes in a given genome (grey), but not between the LOTUS and OSK domains of Oskar (purple). (**: Pearson correlation p-value > 0.1) **b**, Pearson correlation analysis of wobble position identity for the Oskar gene vs other genes. Wobble position identity content is correlated across the sequence 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) **c**, Analysis of GC3 content. Measure of the residuals of Z scores for Oskar gene GC3 content (LOTUS vs OSK) and the Intra-Gene GC3 content. 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 within across the sequences of other genes within a given genome (grey). (** : Mann-Whitney U test p-value $< 10^{-5}$) **d**, Cosine distance analysis of codon frequencies. The distance distribution in codon use between the LOTUS and OSK domain is less than the measured null distribution distance in codon use between any two unrelated genes (Inter-Gene; dark grey), but greater than the expected distance within a gene (Intra-Gene; light grey).

Figure 4
Blondel, Jones & Extavour

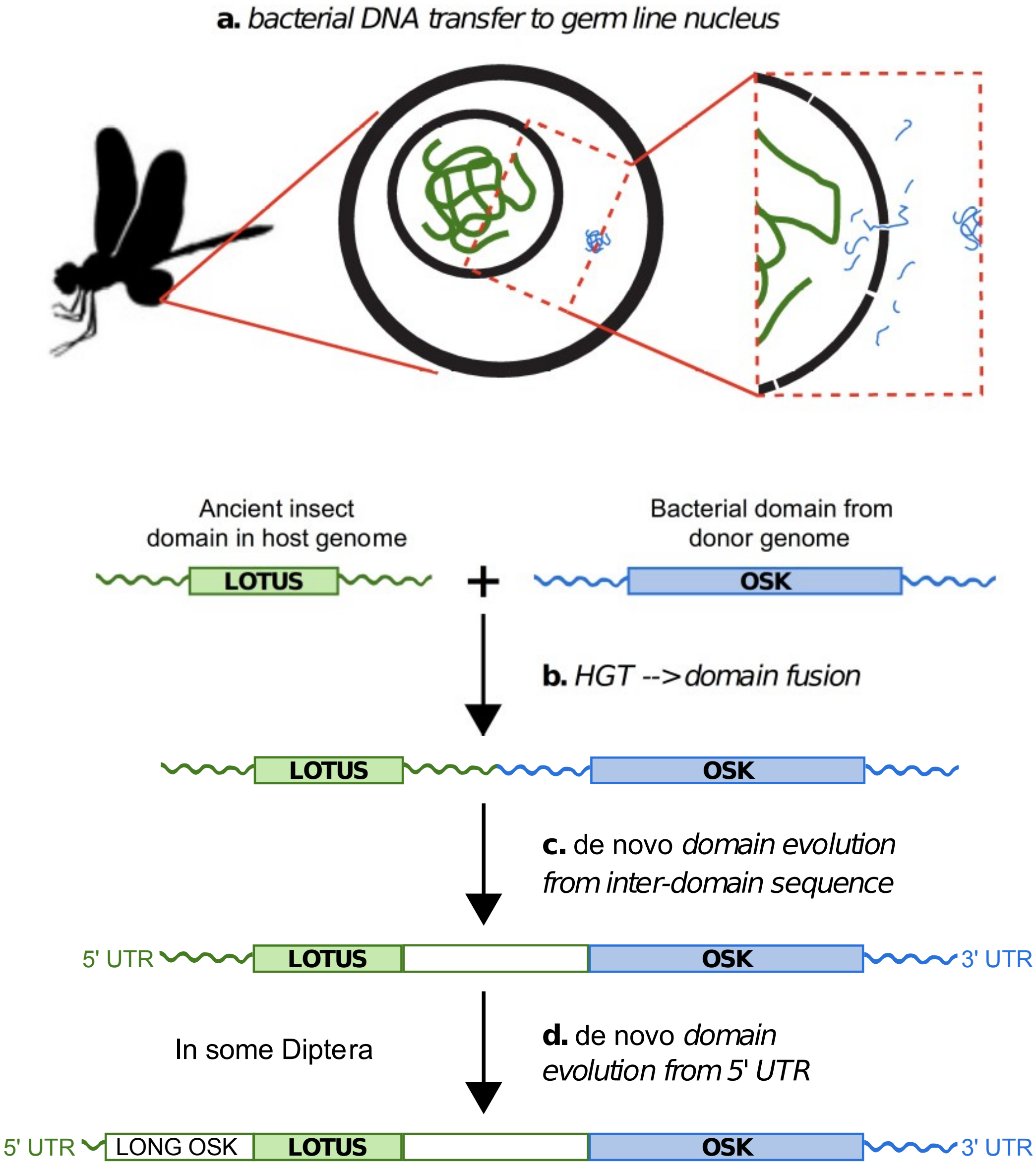

a. bacterial DNA transfer to germ line nucleus
Ancient insect domain in host genome
LOTUS
+
Bacterial domain from donor genome
OSK
b. HGT --> domain fusion
LOTUS
OSK
c. de novo domain evolution from inter-domain sequence
5' UTR
LOTUS
OSK
3' UTR
In some Diptera
d. de novo domain evolution from 5' UTR
5' UTR
LONG OSK
LOTUS
OSK
3' UTR

**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* undergoes *de novo* coding evolution to form the Long Oskar domain.

## Materials and Methods

### *BLAST searches of oskar*

All BLAST[1] searches were performed using the NCBI BLASTp tool suite on the non-redundant (nr) database. Amino Acid (AA) sequences of *D. melanogaster* full length Oskar (EMBL ID AAF54306.1), as well as the AA sequences for the LOTUS (AA 139-238) and OSK (AA 414-606) domains were used for the BLAST searches, using the default NCBI cut-off parameters. As per NCBI defaults, the E-value cut-off was set at 10. All BLAST searches results are included in the Supplementary files: BLAST search results.

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

101 1KITE transcriptomes[2] (Supplementary Table 1) were downloaded and searched using the local BLAST program (BLAST+) using the tblastn algorithm with default parameters, with Oskar protein sequences of *Drosophila melanogaster, Aedes aegypti, Nasonia vitripennis* and *Gryllus bimaculatus* as queries (EntrezIDs: NP_731295.1, ABC41128.1, NP_001234884.1 and AFV31610.1 respectively). For all of these 1KITE transcriptome searches, predicted protein sequences from transcript data were obtained by in silico translation using the online ExPASy translate tool (https://web.expasy.org/translate/), taking the longest open reading frame. Publicly available sequences in the non-redundant (nr), TSA databases at NCBI, and a then-unpublished transcriptome[3] (kind gift of Matthew Benton and Siegfried Roth, University of Cologne) were subsequently searched using the web-based BLAST tool hosted at NCBI, using the tblastn algorithm with default parameters. Sequences used for queries were the four Oskar proteins described above, and newfound *oskar* sequences from the 1KITE transcriptomes of

*Baetis pumilis, Cryptocercus wright,* and *Frankliniella cephalica*. For both searches, *oskar* orthologs were identified by the presence of BLAST hits on the same transcript to both the LOTUS (N-terminal) and OSK (C-terminal) regions of any of the query *oskar* sequences, regardless of E-values. The sequences found were aligned using MUSCLE (8 iterations)[4] into a 46-sequence alignment (Supplementary files: Alignments>OSKAR_INITIAL.fasta). From this alignment, the LOTUS and OSK domains were extracted (Supplementary files: Alignments>LOTUS_INITIAL.fasta and Alignments>OSK_INITIAL.fasta) to define the initial Hidden Markov Models (HMM) using the hmmbuild tool from the HMMER tool suite with default parameters[5]. 126 insect genomes and 128 insect transcriptomes (from the Transcriptome Shotgun Assembly TSA database: https://www.ncbi.nlm.nih.gov/Traces/wgs/?view=TSA) were subsequently downloaded from NCBI (download date September 29, 2015 ; Supplementary table 1). Genomes were submitted to Augustus v2.5.5[6] (using the *D. melanogaster* exon HMM predictor) and SNAP v2006-07-28[7] (using the default 'fly' HMM) for gene discovery. The resulting nucleotide sequence database comprising all 309 downloaded and annotated genomes and transcriptomes, was then translated in six frames to generate a non-redundant amino acid database (where all sequences with the same amino acid content are merged into one). This process was automated using a series of custom scripts available here: https://github.com/Xqua/Genomes. The non-redundant amino acid database was searched using the HMMER v3.1 tool suite[5] and the HMM for the LOTUS and OSK domains described above. A hit was considered positive if it consisted of a contiguous sequence containing both a LOTUS domain and an OSK domain, with the two domains separated by an inter-domain sequence. We imposed no length, alignment or conservation criteria on the inter-domain sequence, as this is a rapidly-evolving region of Oskar protein with predicted high disorder[8-10].

Positive hits were manually curated and added to the main alignment, and the search was performed iteratively until no more new sequences meeting the above criteria were discovered. This resulted in a total of 95 Oskar protein sequences, (see Supplementary Table 2 for the complete list). Using the final resulting alignment (Supplementary Files: Alignments>OSKAR_FINAL.fasta), the LOTUS and OSK domains were extracted from these sequences (Supplementary Files: Alignments>LOTUS_FINAL.fasta and Alignments>OSK_FINAL.fasta), and the final three HMM (for full-length Oskar, OSK, and LOTUS domains) used in subsequent analyses were created using hmmbuild with default parameters (Supplementary files: HMM>OSK.hmm, HMM>LOTUS.hmm and HMM>OSKAR.hmm).

### *Iterative HMMER search of OSK and LOTUS domains*

A reduced version of TrEMBL[11] (v2016-06) was created by concatenating all hits (regardless of E-value) for sequences of the LOTUS domain, the OSK domain and full-length Oskar, using hmmsearch with default parameters and the HMM models created above from the final alignment. This reduced database was created to reduce potential false positive results that might result from the limited size of the sliding window used in the search approach described here. The full-length Oskar alignment of 1133 amino acids (Supplementary files: Alignments>OSKAR_FINAL.fasta) was split into 934 sub-alignments of 60 amino acids each using a sliding window of one amino acid. Each alignment was converted into a HMM using hmmbuild, and searched against the reduced TrEMBL database using hmmsearch using default parameters. Domain of life origin of every hit sequence at each position was recorded. Eukaryotic sequences were further classified as Oskar/Non-Oskar and Arthropod/Non-

Arthropod. Finally, for the whole alignment, the counts for each category were saved and plotted in a stack plot representing the proportion of sequences from each category to create Fig. 1b. The python code used for this search is available at https://github.com/Xqua/Iterative-HMMER.

### *Sequence Similarity Networks*

LOTUS and OSK domain sequences from the final alignment obtained as described above (see "*Hidden Markov Model (HMM) generation and alignments of the OSK and LOTUS domains*"; Supplementary files: Alignments>LOTUS_FINAL.fasta and Alignments>OSK_FINAL.fasta) were searched against TrEMBL[11] (v2016-06) using HMMER. All hits with E-value < 0.01 were consolidated into a fasta file that was then entered into the EFI-EST tool[12] using default parameters to generate a sequence similarity network. An alignment score corresponding to 30% sequence identity was chosen for the generation of the final sequence similarity network. Finally, the network was graphed using Cytoscape 3[13].

### *Phylogenetic Analysis*

For both the LOTUS and OSK domains, in cases where more than one sequence from the same organism was retrieved by the search described above in "*Iterative HMMER Search of OSK and LOTUS domains*", only the sequence with the lowest E-value was used for phylogenetic analysis. For the LOTUS domain, the first 97 best hits (lowest E-value) were selected, and the only three bacterial sequences that satisfied an E-value < 0.01 were manually added. For the OSK domain, the first 95 best hits (lowest E-value) were selected, and the only five eukaryotic sequences that satisfied an E-value < 0.01 were manually added. The sequences were filtered to contain only one sequence per species (best E-value kept) generating a set of 100 sequences for the LOTUS domain, and 87 for the OSK domain. Unique identifiers for all sequences used to generate alignments for phylogenetic analysis are available in Supplementary Tables S3, S4. For both datasets, the sequences were then aligned using MUSCLE[4] (8 iterations) and trimmed using trimAl[14] with 70% occupancy. The resulting alignments that were subject to phylogenetic analysis are available in Supplementary Files: Alignments>LOTUS_TREE.fasta and Alignments>OSK_TREE.fasta. For the maximum likelihood tree, we used RaxML v8.2.4[15] with 1000 bootstraps, and the models were selected using the automatic RaxML model selection tool. The substitution model chosen for both domains was LGF. For the Bayesian tree inference, we used MrBayes V3.2.6[16] with a Mixed model (prset aamodel=Mixed) and a gamma distribution (lset rates=Gamma). We ran the MonteCarlo for 4 million generations (std < 0.01) for the OSK domain, and for 3 million generations (std < 0.01) for the LOTUS domain.

***Selection of sequences for codon use analysis***

To study the codon use of the OSK and LOTUS domains, we chose 17 well-annotated (defined as possessing at least 8,000 annotated genes) insect genomes that included a confidently annotated *oskar* orthologue from the NCBI nucleotide database. The complete list and accession numbers of the sequences used for this analysis is in Supplementary Table 5. This list contains *oskar* sequences from genomes that were either added to the databases after the first *oskar* sequence search or re-annotated after said search. Therefore the sequences coming from the following organisms are not represented in the final *oskar* alignment: *Harpegnathos saltator, Fopius arisanus, Athalia rosae, Orussus abietinus, Stomoxys calcitrans, Bactrocera oleae, Neodiprion lecontei*.

### *Generation of Intra-Gene distribution of codon use*

We wished to determine whether *oskar* differed from the null hypothesis that a given gene would follow similar codon use throughout its sequence. To generate a distribution of codon use similarity across a gene for all genes in the genomes studied, we generated what we named the "Intra-Gene" sequence distribution. Each gene was cut into two fragments at a random position "x" following the rule: 384 < x < Length_gene - 384, x modulo 3 = 0 (Corresponding Jupyter notebook file: Scripts>notebook>Codon Analysis AT3 GC3 and A3 T3 G3 C3 Section: 4). Thus, we sampled each codon at least twice, preserving the coding frame.

### *Fitting a linear model of codon use*

Using the Intra-Gene null distribution generated above, we fitted a linear model of codon use frequencies per gene for the wobble position and AT3 GC3 content. To do so, we measured the different frequencies of A3, T3, G3 and C3 (any codon ending in A was counted as A3) and AT3 GC3. Then, we fitted a linear model to the pairs of 5' and 3' regional codon use values for within each gene, obtained from the Intra-Gene distribution described above (conserving the 3'/5' position information), and for the OSK and LOTUS domains, for each of the 17 genomes analyzed (Supp Table 3). We then calculated the residuals of the Intra-Gene distribution and the LOTUS-OSK distribution. Finally, we determined the Pearson correlation coefficient for all genomes pooled together, and all *oskar* genes pooled together (Corresponding Jupyter notebook file: Scripts>notebook>Codon Analysis AT3 GC3 and A3 T3 G3 C3 Section: 7 and 8).

### *Calculation of cosine distance*

For a given sequence S, we assigned a vector C of dimension 64 (one for each codon). Because the sum of all codon frequencies is 1, C is normalized; we thus used the cosine similarity distance between a given pair of vectors as a metric to quantify the distance in codon use between two sequences. We measured this distance distribution between all the genes in a given genome to create the Inter-Gene distance distribution. Then, we repeated the process but measured the distance between all pairs of genes in the Intra-Gene sequence set per genome. Next, we measured the distance between the LOTUS and OSK domains for each genome. Finally, we determined the Z score of the distance between the LOTUS and OSK domains, and the Inter-Gene and Intra-Gene distance distributions (Corresponding Jupyter notebook file: Scripts>notebook>Cosine Distance Analysis).

### *Calculation and analysis of the codon use Z_score*

For each genome, the codon use frequency for AT3/GC3 and A3/T3/G3/C3 was calculated as described above. Then, Z scores for each sequence from the Intra-Gene, OSK or LOTUS domain sequences were calculated against the corresponding genome frequency distribution. The Z scores were then used to generate the analysis of Pearson correlation coefficients shown in Figures 3, S5 and S6 (Corresponding Jupyter notebook file: Scripts>notebook>Codon Analysis AT3 GC3 and A3 T3 G3 C3 Section: 3, 5 and 6).

### *Data availability*

All sequences discovered using the automatic annotation pipeline described in (M&M HMM and oskar search) are annotated as such in Supplementary Table S2.

***Code availability***

All custom code generated for this study is available in Supplementary Information>Scripts.

## Methods References

1 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410, doi:10.1016/S0022-2836(05)80360-2 (1990).

2 Aspöck, H. *et al. 1KITE - 1K Insect Transcriptome Evolution*, <http://www.1kite.org/> (2018).

3 Benton, M. A., Kenny, N. J., Conrads, K. H., Roth, S. & Lynch, J. A. Deep, Staged Transcriptomic Resources for the Novel Coleopteran Models Atrachya menetriesi and Callosobruchus maculatus. *PLoS ONE* **11**, e0167431, doi:10.1371/journal.pone.0167431 (2016).

4 Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113 (2004).

5 Eddy, S. R. *HMMER: biosequence analysis using profile hidden Markov models*, <http://hmmer.org/> (2007).

6 Stanke, M., Steinkamp, R., Waack, S. & Morgenstern, B. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res.* **32**, W309-312, doi:10.1093/nar/gkh379 (2004).

7 Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59, doi:10.1186/1471-2105-5-59 (2004).

8 Jeske, M. *et al.* The Crystal Structure of the Drosophila Germline Inducer Oskar Identifies Two Domains with Distinct Vasa Helicase- and RNA-Binding Activities. *Cell Rep* **12**, 587-598, doi:10.1016/j.celrep.2015.06.055 (2015).

9 Ahuja, A. & Extavour, C. G. Patterns of molecular evolution of the germ line specification gene *oskar* suggest that a novel domain may contribute to functional divergence in *Drosophila*. *Dev. Genes Evol.* **222**, 65-77 (2014).

10 Yang, N. *et al.* Structure of Drosophila Oskar reveals a novel RNA binding protein. *Proc Natl Acad Sci U S A* **112**, 11541-11546, doi:10.1073/pnas.1515568112 (2015).

11 Consortium, U. The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res.* **37**, D169-174, doi:10.1093/nar/gkn664 (2009).

12 Gerlt, J. A. *et al.* Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): A web tool for generating protein sequence similarity networks. *Biochim. Biophys. Acta* **1854**, 1019-1037, doi:10.1016/j.bbapap.2015.04.015 (2015).

13 Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498-2504, doi:10.1101/gr.1239303 (2003).

14 Capella-Gutierrez, S., Silla-Martinez, J. M. & Gabaldon, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972-1973, doi:10.1093/bioinformatics/btp348 (2009).

15 Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313, doi:10.1093/bioinformatics/btu033 (2014).

16 Huelsenbeck, J. P. & Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754-755 (2001).

## 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:

1. Supplementary Discussion (this document)
2. Supplementary References (this document)
3. Folder titled “Supplementary Information Files” containing the following sub-folders
    - *a.* Supplementary Information Files>Alignments
        - i. *All sequences identified and analyzed in this study, in FASTA format and with corresponding Alignments*
    - *b.* Supplementary Information Files>BLAST search results
        - i. *Results of BLASTP searches with full length Oskar, OSK or LOTUS domains as queries*
    - *c.* Supplementary Information Files>Data
        - i. *Necessary files for running the different ipython notebooks:*
            1. *Taxonomy: Conversion table for UniProt ID to taxon information. (uniprot_ID_taxa.tsv )*
            2. *Codon_Genes: Contains the measured codon frequency for the different genomes studied as .csv or .tsv files (organism_name.csv/tsv), along with the DNA sequences of LOTUS and OSK domains used in the codon use analysis (LOTUS_Seqeuences.gb and SGNH_Seqeuences.gb)*
            3. *Trees: Contains the tree files obtained from RaxML and MrBayes phylogenetic analyses of the OSK and LOTUS domains.*
    - *d.* Supplementary Information Files>HMM
        - i. *HMM models used for iterative searching for sequences similar to full-length Oskar, LOTUS and OSK domains*
    - *e.* Supplementary Information Files>Scripts
        - i. *All custom scripts used to implement the analysis pipelines described.*
    - *f.* Supplementary Information Files>Tables
        - i. *Supplementary Tables S1-S5 describing databases searched/analyzed and all search results; Legends in this document*

Please download Supplementary Information Files here: https://www.dropbox.com/s/q4sd5rty24gxprg/Blondel_Jones_Extavour_HGT_Supplementary%20Information%20Files.zip?dl=0

## Supplementary Discussion

### *Phylogenetic relationships of the Oskar LOTUS domain*

LOTUS sequences from non-Oskar proteins that were sufficiently similar to the Osk LOTUS domain to be included in an alignment for phylogenetic analysis, were almost exclusively eukaryotic. (Supplementary Table 3). Only three bacterial sequences matched the LOTUS domain with an E-value < 0.01, and were included in the alignment (Supplementary Table 3). Osk LOTUS domains clustered into two distinct clades, one comprising all Dipteran sequences, and the other comprising all other Osk LOTUS domains examined from both holometabolous and hemimetabolous orders (Fig. 2a). Dipteran Osk LOTUS sequences formed a monophyletic group that branched sister to a clade of LOTUS domains from Tud5 family proteins of non-arthropod animals (NAA). NAA LOTUS domains from Tud7 family members were polyphyletic, but most of them formed a clade branching sister to (Osk LOTUS + NAA Tud5 LOTUS). Non-Dipteran Osk LOTUS domains formed a monophyletic group that was related in a polytomy to the aforementioned (NAA Tud7 LOTUS + (Dipteran Osk LOTUS + NAA Tud5 LOTUS)) clade, and to various arthropod Tud7 family LOTUS domains.

The fact that Tud7 LOTUS domains are polyphyletic suggests that arthropod domains in this family may have undergone heterogeneous evolutionary processes relative to their homologues in other animals. The relationships of Dipteran LOTUS sequences were consistent with the current hypothesis for interrelationships between Dipteran species[1] Similarly, among the non-Dipteran Osk LOTUS sequences, the hymenopteran sequences form a clade to the exclusion of the single hemimetabolous sequence (from the cricket *Gryllus bimaculatus*), consistent with the monophyly of Hymenoptera[2]. It is unclear why Dipteran Osk LOTUS domains cluster separately from those of other insect Osk proteins. We speculate that the evolution of the Long Oskar domain[3,4], which appears to be a novelty within Diptera (Supplementary Files: Alignments>OSKAR_FINAL.fasta), may have influenced the evolution of the Osk LOTUS domain in at least some of these insects. Consistent with this hypothesis, of the 17 Dipteran *oskar* genes we examined, the seven *oskar* genes possessing a Long Osk domain clustered into two clades based on the sequences of their LOTUS domain. One of these clades comprised five Drosophila species (*D. willistoni*, *D. mojavensis*, *D. virilis*, *D. grimshawi* and *D. immigrans*), and the second was composed of two calyptrate flies from different superfamilies, *Musca domestica* (Muscoidea) and *Lucilia cuprina* (Oestroidea).

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

### *Phylogenetic relationships of the Oskar OSK domain*

The only eukaryotic proteins emerging from the iterative HMMER search for OSK domain sequences that had an E-value < 0.01 were all from fungi. All five of these sequences were annotated as  Carbohydrate Active Enzyme 3 (CAZ3). Most bacterial sequences used in this analysis were annotated as lipases and hydrolases, with a high representation of GDSL-like hydrolases (Supplementary Table S4). OSK sequences formed a monophyletic group but did not branch sister to the other eukaryotic sequences in the analysis. Instead, all CAZ3 sequences formed a clade that was sister to a clade of primarily Firmicutes. We recovered a monophyletic group of Proteobacteria nested within that Firmicutes clade. All Bacteroidetes sequences also formed a monophyletic group, which branched sister to all other sequences except for the two Archaeal sequences in the analysis. Within the OSK clade, the topology of sequence relationships was largely concordant with the species tree for insects [5], as we recovered monophyletic Diptera to the exclusion of other insect species. However, the single orthopteran OSK sequence (from the cricket *Gryllus bimaculatus*) grouped within the Hymenoptera, rather than branching basally to all insect sequences as would be expected for this hemimetabolous sequence.

## Supplementary Table Legends

(see Supplementary Information Files>Tables>Supp TableX)

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

**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.

**Supplementary Table S3: List of sequences 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.

**Supplementary Table S4: List of sequences 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.

**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*).

## Supplementary References

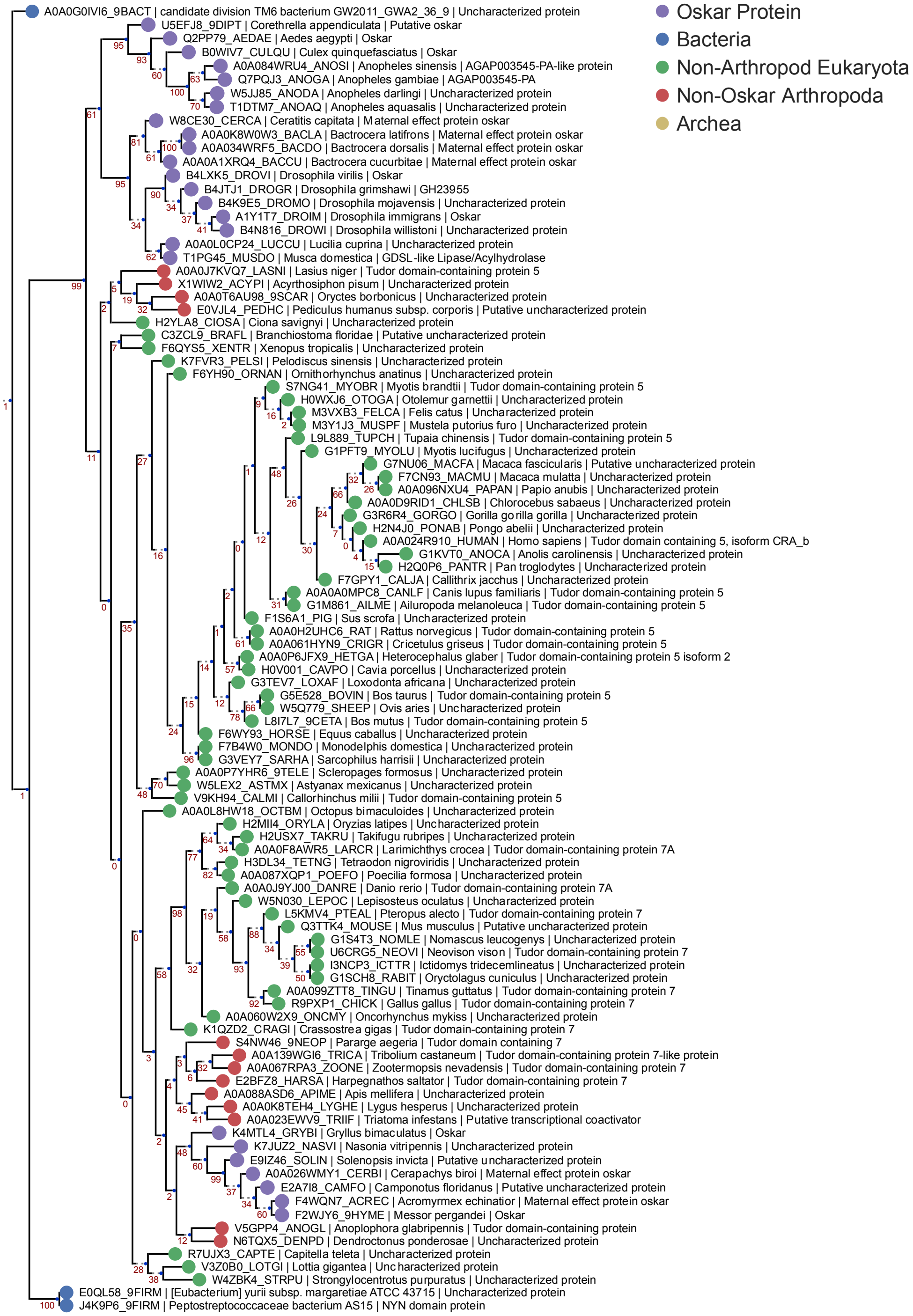
1 Kirk-Spriggs, A. H. & Sinclair, B. J. Vol. 1 (South African National Biodiversity Institute, Pretoria, South Africa, 2017).

2 Peters, R. S. *et al.* Evolutionary History of the Hymenoptera. *Curr. Biol.* **27**, 1013-1018, doi:10.1016/j.cub.2017.01.027 (2017).

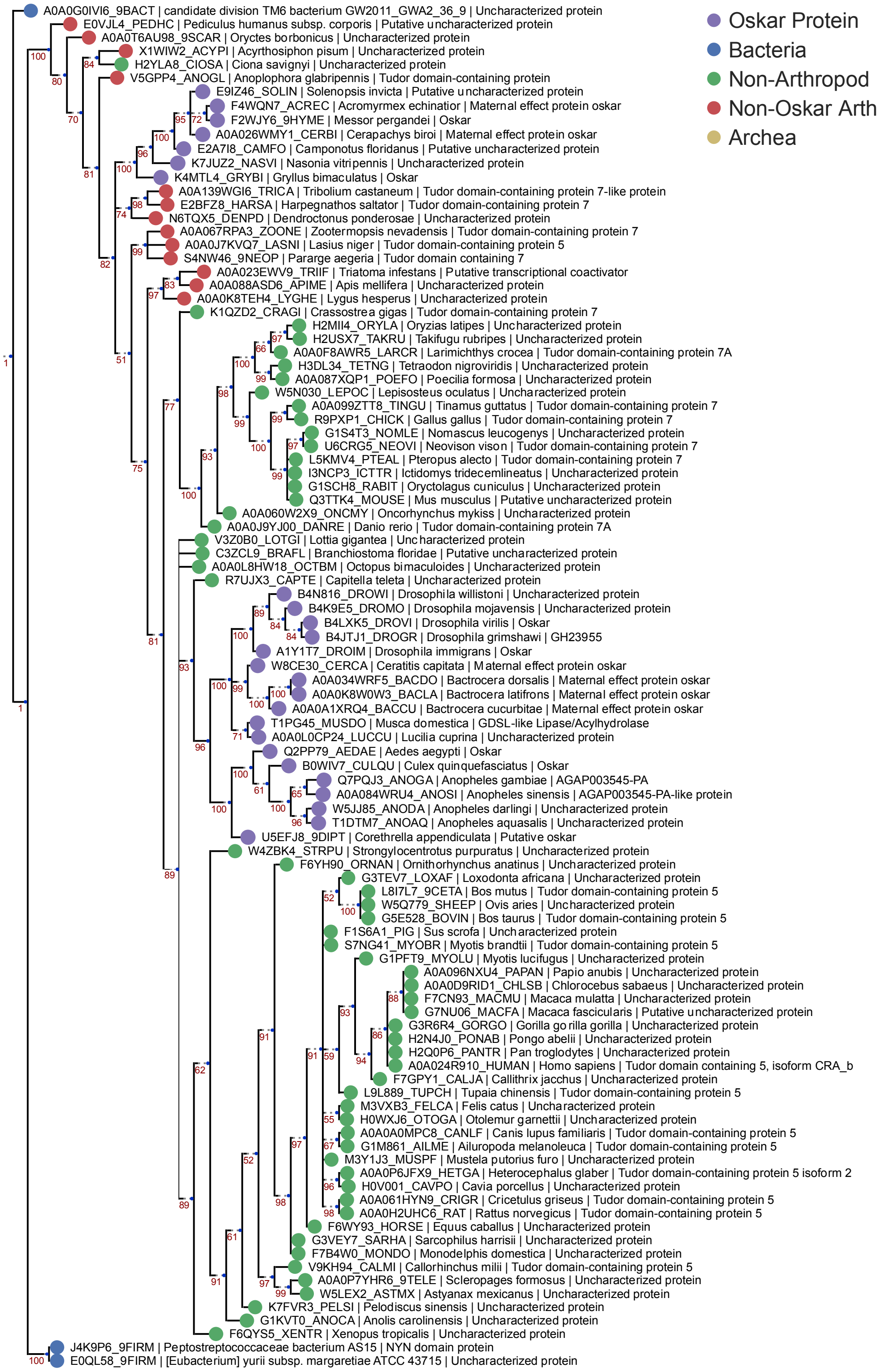
3 Vanzo, N. F. & Ephrussi, A. Oskar anchoring restricts pole plasm formation to the posterior of the *Drosophila* oocyte. *Development* **129**, 3705-3714 (2002).

4 Hurd, T. R. *et al.* Long Oskar Controls Mitochondrial Inheritance in Drosophila melanogaster. *Dev Cell* **39**, 560-571, doi:10.1016/j.devcel.2016.11.004 (2016).

5 Misof, B. *et al.* Phylogenomics resolves the timing and pattern of insect evolution. *Science* **346**, 763-767, doi:10.1126/science.1257570 (2014).

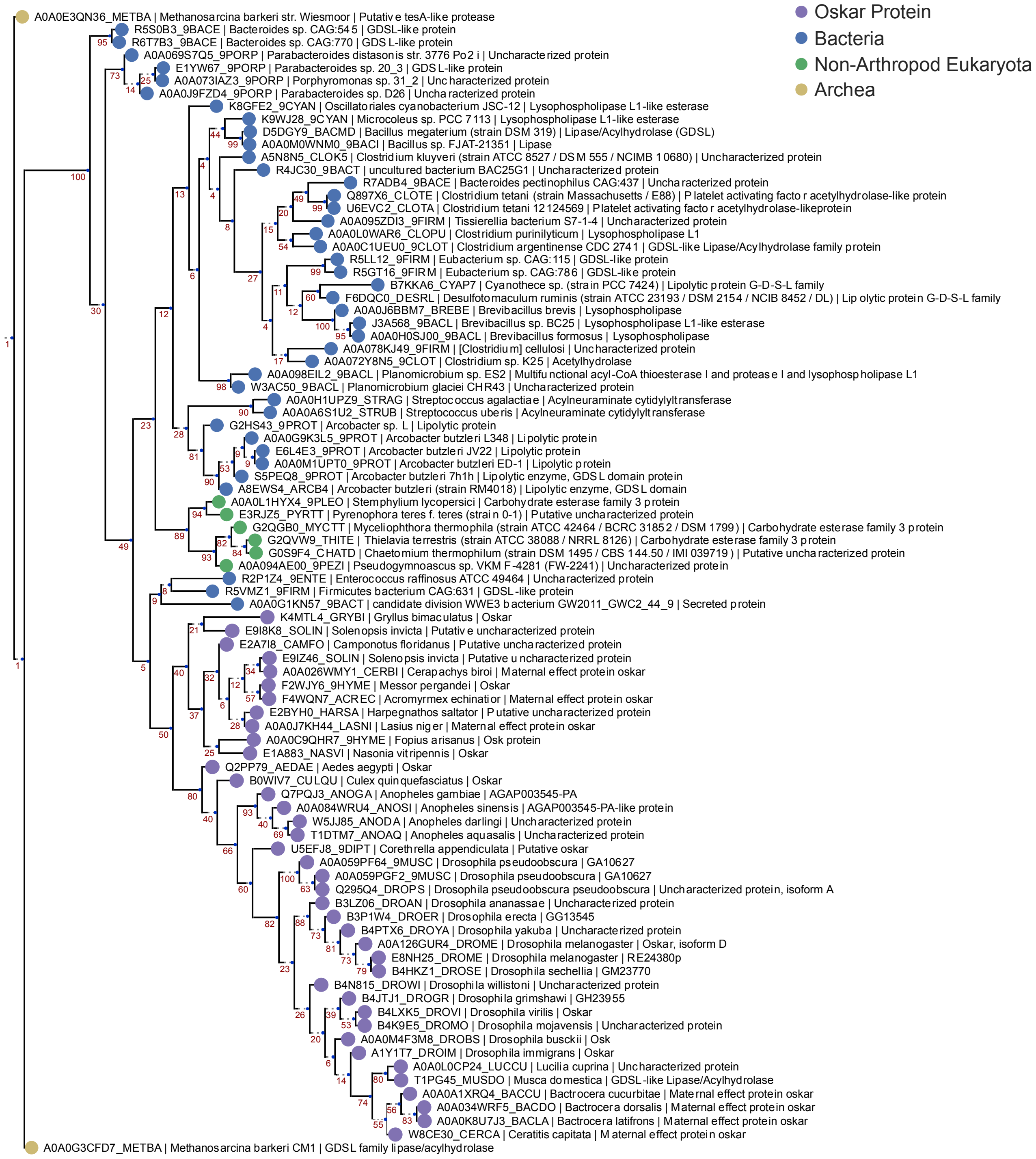

Oskar Protein
Bacteria
Non-Arthropod Eukaryota
Non-Oskar Arthropoda
Archea
A0A0G0IVI6_9BACT | candidate division TM6 bacterium GW2011_GWA2_36_9 | Uncharacterized protein
U5EFJ8_9DIPT | Corethrella appendiculata | Putative oskar
Q2PP79_AEDAE | Aedes aegypti | Oskar
B0WIV7_CULQU | Culex quinquefasciatus | 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

**Extended Data Figure S1: LOTUS Domain RaxML 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.

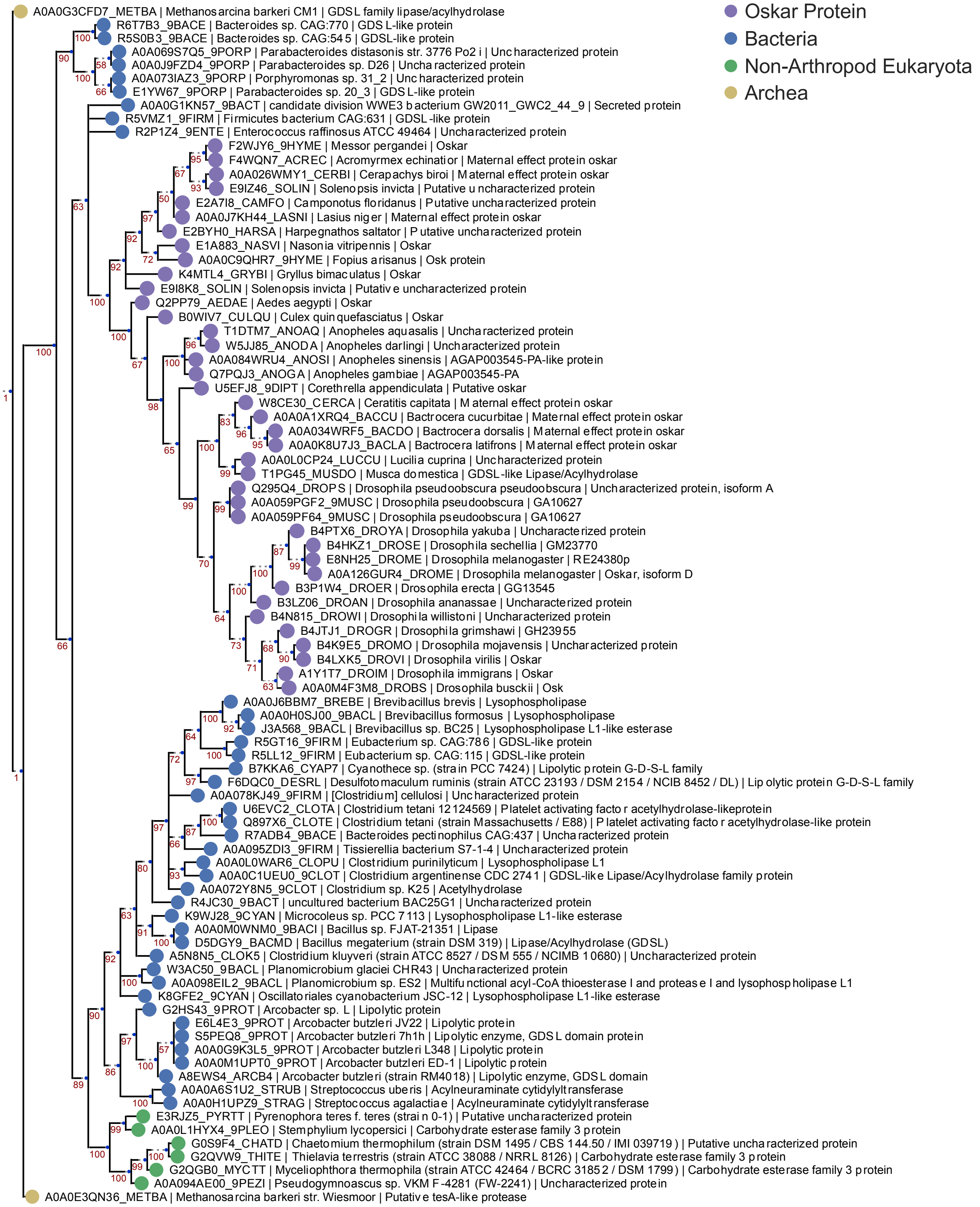


● Oskar Protein
● Bacteria
● Non-Arthropod Eukaryota
● Non-Oskar Arthropoda
● Archea
A0A0G0IVI6_9BACT | candidate division TM6 bacterium GW2011_GWA2_36_9 | Uncharacterized protein
E0VJL4_PEDHC | Pediculus humanus subsp. corporis | Putative uncharacterized protein
A0A0T6AU98_9SCAR | Oryctes borbonicus | 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
100
80
84
70
81
95
72
100
96
100
98
74
82
99
83
97
1
51
97
66
100
99
98
77
99
99
100
97
93
99
75
100
1
81
89
84
84
100
93
100
99
100
100
71
96
100
61
65
100
100
96
89
52
100
62
91
93
88
86
91
59
94
91
55
97
67
52
96
98
98
61
97
91
99
89
100

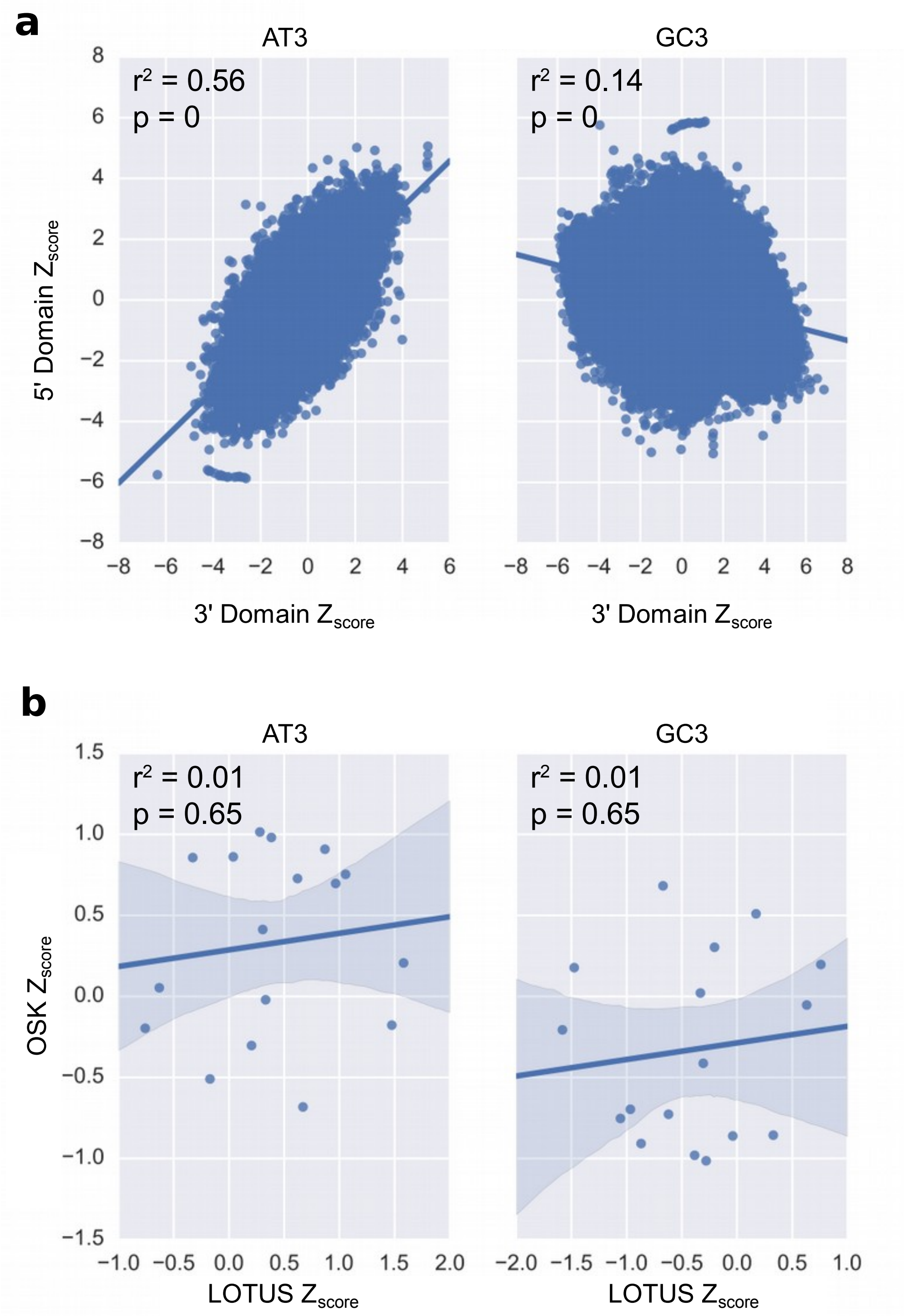
**Extended Data Figure S2: LOTUS Domain Bayesian Tree.** Phylogenetic tree of the HMMER sequences retrieved from the UniProt database using the LOTUS alignment HMM model. 100 sequences were chosen for analysis as described for Supplementary Figure 1. The tree was created using Mr Bayes V3.2.6 using a 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.


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 | Tissierellia 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
95 73 25 14 100 44 99 4 4 13 49 99 20 8 15 54 6 27 99 11 60 12 30 12 100 4 95 17 1 98 90 23 28 81 9 53 9 90 94 89 82 84 93 49 8 9 1 21 5 40 32 34 12 57 6 37 28 50 25 80 40 93 40 69 66 60 100 63 82 88 73 81 73 79 23 26 39 53 20 6 14 80 74 56 83 55

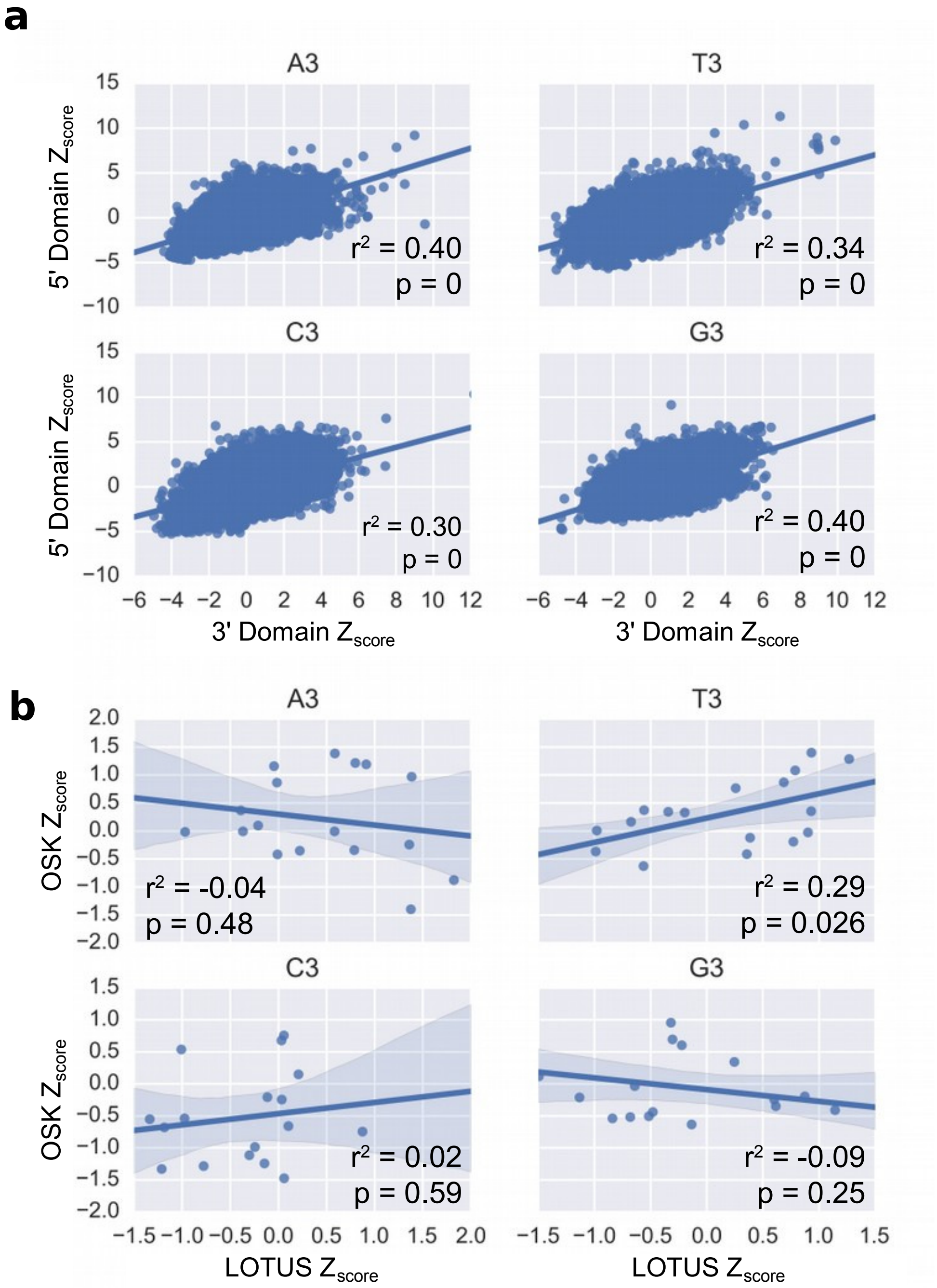
**Extended Data Figure S3: OSK Domain RaxML Tree.** Phylogenetic tree of the HMMER sequences retrieved from the UniProt database using the OSK alignment HMM model. The top 95 hits were selected for phylogenetic analysis, and the only five non-Oskar eukaryotic sequences found to be a match were added 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.


● 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 | Tissierellia 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

**Extended Data Figure S4: OSK Domain Bayesian 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 = Bacteria. Names following leaves display the UniProt accession number followed by the species name and the UniProt protein name.



**Extended Data Figure S5: AT3/GC3 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 AT3 and GC3 codon use was measured and a Z-score was calculated against the genome distribution. Finally, the 5' and 3' “domain” values were plotted against each other and a linear regression was . The AT3 and GC3 content is generally similar in the 5’ and 3’regions of all genes across the genome (AT3: $r^2 = 0.56$, p = 0; GC3: $r^2 = 0.14$, p = 0). (**b**) OSK vs LOTUS AT3 and GC3 use across the 17 genomes analyzed. The AT3 and GC3 content Z-scores were calculated against the genome distribution. The AT3 and GC3 content of the two domains of the Oskar gene are not correlated with each other. (AT3: $r^2 = 0.01$, p = 0.65; GC3: $r^2 = 0.01$, p = 0.65).



**Extended Data Figure S6: 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).