ARTICLE

Evolution of a cytoplasmic determinant: evidence for the biochemical basis of functional evolution of the novel germ line regulator Oskar

Leo Blondel[1], Savandara Besse[1,2] , Emily L. Rivard[1], Guillem Ylla[3], and Cassandra G. Extavour[1,3]

1. Department of Molecular and Cellular Biology, Harvard University, Cambridge MA, USA
2. Current address: Department of Biochemistry and Molecular Medicine, Université de Montréal, Montréal Québec, Canada
3. Department of Organismic and Evolutionary Biology, Harvard University, Cambridge MA, USA

Corresponding author: Cassandra G. Extavour extavour@oeb.harvard.edu

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

© The Author(s) 2021. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Germ line specification is essential in sexually reproducing organisms. Despite their critical role, the evolutionary history of the genes that specify animal germ cells is heterogeneous and dynamic. In many insects, the gene *oskar* is required for the specification of the germ line. However, the germ line role of *oskar* is thought to be a derived role resulting from co-option from an ancestral somatic role. To address how evolutionary changes in protein sequence could have led to changes in the function of Oskar protein that enabled it to regulate germ line specification, we searched for *oskar* orthologs in 1565 publicly available insect genomic and transcriptomic datasets. The earliest-diverging lineage in which we identified an *oskar* ortholog was the order Zygentoma (silverfish and firebrats), suggesting that *oskar* originated before the origin of winged insects. We noted some order-specific trends in *oskar* sequence evolution, including whole gene duplications, clade-specific losses, and rapid divergence. An alignment of all known 379 Oskar sequences revealed new highly conserved residues as candidates that promote dimerization of the LOTUS domain. Moreover, we identified regions of the OSK domain with conserved predicted RNA binding potential. Furthermore, we show that despite a low overall amino acid conservation, the LOTUS domain shows higher conservation of predicted secondary structure than the OSK domain. Finally, we suggest new key amino acids in the LOTUS domain that may be involved in the previously reported Oskar-Vasa physical interaction that is required for its germ line role.

Keywords: *oskar*, *vasa*, Drosophila, germ plasm, germ cell, LOTUS domain, RNA binding, Hidden Markov Models, Hymenoptera, Lepidoptera, Zygentoma

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

Introduction

With the evolution of obligate multicellularity, many organisms faced a challenge considered a major evolutionary transition: allocating only some cells (germ line) to pass on their genetic material to the next generation, relegating the remainder (soma) to death upon death of the organism (reviewed in (Kirk 2005)). While there are multiple mechanisms of germ cell specification, they can be grouped into two broad categories, induction or inheritance (reviewed in (Extavour and Akam 2003)). Under induction, cells respond to an external signal by adopting germ cell fate. Under the inheritance mechanism, maternally synthesized cytoplasmic molecules, located within a specialized cytoplasm called germ plasm, are deposited in the oocyte and “inherited” by a subset of cells during early embryonic divisions. Cells inheriting these molecules commit to a germ line fate (reviewed in (Extavour and Akam 2003)).

The inheritance mechanism in insects that undergo metamorphosis (Holometabola) appears to have evolved by co-option of a key gene, *oskar*. *oskar* was first identified in forward genetic screens for axial patterning mutants in *Drosophila melanogaster* (Lehmann and Nüsslein-Volhard 1986). For the first 20 years following its discovery, *oskar* appeared to be restricted to Drosophilids (Clark, et al. 2007). Its later discovery in the mosquitoes *Aedes aegypti*, *Anopheles gambiae* and *Culex quinquefasciatus* (Juhn and James 2006; Juhn, et al. 2008) and the wasp *Nasonia vitripennis* (Lynch, et al. 2011) suggested the hypothesis that *oskar* emerged at the base of the Holometabola, and facilitated the evolution of germ plasm in these insects (Lynch, et al. 2011). However, our subsequent identification of *oskar* homologs in the cricket *Gryllus bimaculatus* (Ewen-Campen, et al. 2012), and in many additional hemimetabolous insect species (Blondel, et al. 2020), demonstrated that *oskar* predates the Holometabola, and must be at least as old as the major radiation of insects (Misof, et al. 2014). Two secondary losses of *oskar* from insect genomes have also been reported, in the beetle *Tribolium castaneum* (Lynch, et al. 2011) and the honeybee *Apis mellifera* (Dearden, et al. 2006), and neither of these insects appear to use germ plasm to establish their germ lines (Nelson 1915; Nagy, et al. 1994; Dearden 2006; Schroder 2006). Whether *oskar* is ubiquitous across all insect orders, whether it is truly unique to insects, the evidence for or against potential losses or duplications of the *oskar* locus across insects, and the evolutionary dynamics of the locus, remain unknown.

oskar remains, to our knowledge, the only gene that has been experimentally demonstrated to be both necessary and sufficient to induce the formation of functional primordial

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

germ cells (called pole cells in *Drosophila*) (Kim-Ha, et al. 1991; Ephrussi and Lehmann 1992). Thus, in *D. melanogaster* (Lehmann and Nüsslein-Volhard 1986; Kim-Ha, et al. 1991; Ephrussi and Lehmann 1992) and potentially more broadly in holometabolous insects with germ plasm (Lynch, et al. 2011; Rafiqi, et al. 2020), *oskar* plays an essential germ line role. However, it is clear that *oskar*'s germ line function can evolve rapidly, as even within the genus *Drosophila*, *oskar* homologs from different species cannot always substitute for each other (Webster, et al. 1994; Jones and Macdonald 2007). Moreover, the ancestral function of this gene may have been in the nervous system rather than the germ line (Ewen-Campen, et al. 2012). The current hypothesis is therefore that it was co-opted to play a key role in the acquisition of an inheritance-based germ line specification mechanism approximately 300 million years ago (Misof, et al. 2014), in the lineage leading to the Holometabola (Ewen-Campen, et al. 2012). Thus, the case of *oskar* offers an opportunity to study the evolution of protein function at multiple levels of biological organization, from the genesis of a novel protein, through to potential co-option events and the evolution of functional variation.

Neofunctionalization often correlates with a change in the fitness landscape of the protein sequence caused by novel biochemical constraints imposed by amino acid sequence changes (Sikosek, et al. 2012; Sikosek and Chan 2014). Such potential constraints may be revealed by analyzing the conservation of amino acids, their chemical properties, or structure at the secondary, tertiary or quaternary levels (Sikosek and Chan 2014). Oskar has two well-structured domains conserved across identified homologs to date (Blondel, et al. 2020): an N-terminal Helix Turn Helix (HTH) domain termed LOTUS with potential RNA binding properties (Anantharaman, et al. 2010; Jeske, et al. 2015; Yang, et al. 2015; Jeske, et al. 2017), and a C-terminal GDSL-lipase-like domain called OSK (Jeske, et al. 2015; Yang, et al. 2015) (Figure 1). These two domains are linked by an unstructured highly variable interdomain sequence (Ahuja and Extavour 2014; Jeske, et al. 2015; Yang, et al. 2015). We previously showed that this domain structure is likely the result of a horizontal transfer event of a bacterial GDSL-lipase-like domain, followed by the fusion of this domain with a LOTUS domain in the host genome (Blondel, et al. 2020). Biochemical assays of the properties of the LOTUS and OSK domains provide some clues as to the molecular mechanisms that Oskar uses to assemble germ plasm in *D. melanogaster*. The LOTUS domain is capable of homodimerization (Jeske, et al. 2015; Jeske, et al. 2017), and directly binds and enhances the helicase activity of the ATP-dependent DEAD

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

box helicase Vasa, a germ plasm component (Jeske, et al. 2017). The OSK domain resembles GDSL lipases in sequence (Jeske, et al. 2015; Yang, et al. 2015; Blondel, et al. 2020), but is predicted to lack enzymatic activity, as the conserved amino acid triad (S200 D202 H205) that defines the active site of these lipases is not conserved in OSK (Anantharaman, et al. 2010; Jeske, et al. 2015; Yang, et al. 2015). Instead, co-purification experiments suggest that OSK has RNA binding properties, consistent with its predicted basic surface residues (Jeske, et al. 2015; Yang, et al. 2015). Whether or how changes in the primary sequence of Oskar can explain the evolution of its molecular mechanism or tissue-specific function, remain unknown.

To date, sequences of approximately 100 *oskar* homologs have been reported (Lynch, et al. 2011; Jeske, et al. 2015; Quan and Lynch 2016; Blondel, et al. 2020). However, the vast majority of these are from the Holometabola, and it is thus unclear whether analysis of these sequences alone would have sufficient power to allow extrapolation of conservation and divergence of putative biochemical properties across insects broadly speaking. Multiple hypotheses as to the molecular mechanistic function of particular amino acids in the LOTUS and OSK domains in *D. melanogaster* have been proposed (Jeske, et al. 2015; Yang, et al. 2015; Jeske, et al. 2017), but without sufficient taxon sampling, the potential relevance of these mechanisms to *oskar*'s evolution and function in other insects is unclear.

Here we address these outstanding questions by applying a rigorous bioinformatic pipeline to generate the most complete collection of *oskar* sequences to date. By analyzing 1862 Pancrustacean genomes and transcriptomes, we show that *oskar* likely first arose at least 400 million years ago, before the advent of winged insects (Pterygota). We find that the *oskar* locus has been lost independently in some insect orders, including near-total absence from the order Hemiptera, and clarify that the absence of *oskar* from the *Bombyx mori* and *Tribolium castaneum* genomes (discussed in Quan and Lynch 2016) does not reflect a general absence of *oskar* from Lepidoptera or Coleoptera. By comparing Oskar sequences in a phylogenetic context, we reveal that distinct biophysical properties of Oskar are associated with Hemimetabola and Holometabola. We use these observations to propose testable hypotheses regarding the putative biochemical basis of evolutionary change in Oskar function across insects.

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

Results

HMM-based discovery pipeline yields hundreds of novel oskar *homologs*

We wished to study the evolution of the *oskar* gene sequence as comprehensively as possible across all insects. To expand our previous collection of nearly 100 homologous sequences (Blondel, et al. 2020), we designed a new bioinformatics pipeline to scan and search for *oskar* homologs across all 1565 NCBI insect transcriptomes and genomes that were publicly available at the time of analysis (Supplementary Table S1; Figure 2; see Methods: *Genome and transcriptome pre-processing* for NCBI accession numbers and additional information). First, we used the HMMER tool suite to build HMM models for each of the LOTUS and OSK domains, using our previously generated multiple sequence alignments (MSA) (Blondel, et al. 2020). We subjected genomes to *in silico* gene model inference using Augustus (Stanke, et al. 2006). We translated the resulting predicted transcripts, as well as the predicted transcripts from RNA-seq datasets, in all six frames. We then scanned the resulting protein sequences for the presence of LOTUS and OSK domains using the aforementioned HMM models. Sequences were designated as *oskar* homologs based on the same criteria as in our previous study (Blondel, et al. 2020), namely, sequences containing both a LOTUS and an OSK domain (Jeske, et al. 2015), separated by a variable interdomain region. We then aligned all sequences using *hmmalign* and the HMM derived from our previously published full length Oskar alignment (Blondel, et al. 2020). The first iteration of the alignment was manually curated as previously described (Blondel, et al. 2020), and sequence duplicates and sequences that did not align correctly were removed. All subsequent iterations were automatically curated following the process described in *Methods: Identification of oskar homologs.*

With these methods, we recovered a total of 379 unique *oskar* sequences from 350 unique species. To our knowledge, this comprises the largest collection of *oskar* homologs described to date. To determine if *oskar* homologs might predate Insecta, we applied the discovery pipeline to all 31 genomes and 266 transcriptomes of non-insect pancrustaceans available at the time of analysis (see Methods: *Genomes and transcriptomes preprocessing* for complete list). However, we did not recover any non-insect sequences meeting our criteria for *oskar* homologs (Figure 3), strongly suggesting that *oskar* is restricted to the insect lineage (Lynch, et al. 2011; Ahuja and Extavour 2014).

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

We found that 58.65% of RefSeq genomes (78/133), 30.42% of GenBank genomes (94/309), and 21.19% of transcriptomes (238/1123) analyzed contained predicted *oskar* homologs (Supplementary Table S1 and Supplementary Figure S1a). Given that detection of putative homologs is highly dependent on the quality of the genome assembly and annotation, we asked whether there were differences in the assembly statistics of genomes with and without predicted *oskar* homologs. We observed a significant difference in N50, L50, number of contigs and number of scaffolds between genomes lacking *oskar* hits and those where *oskar* was identified (Mann-Whitney U test p-value < 0.05). Genomes where we did not find *oskar* showed a significantly higher mean/median contig and scaffold count, smaller contig and scaffold N50 length, larger contig and scaffold L50, and more contigs or scaffolds per genome length, than genomes where we detected an *oskar* homolog (Mann Whitney U test p<0.05; Supplementary Figure S2; Supplementary Table S2). We interpret this to mean that *oskar* may appear to be absent from these datasets due to potentially incomplete sequencing, suggesting that deeper sequencing in these lineages could potentially reveal additional new *oskar* homologs in future studies. However, we note that we believe that our analysis provides strong evidence for true *oskar* loss in some lineages, given their very deeply sequenced and well-annotated genomes (e.g. *Apis mellifera*, *Tribolium castaneum*).

oskar *predates the divergence of Ametabola and other insects*

We identified *oskar* homologs in 15 of the 29 generally recognized insect orders (Misof, et al. 2014), including eight holometabolous orders, six hemimetabolous orders, and one ametabolous order (Figure 3). This result is consistent with our previous proposals that *oskar* predates the origins of the Holometabola (Ewen-Campen, et al. 2012; Blondel, et al. 2020). The novel finding of an *oskar* homolog from the silverfish *Atelura formicaria* (Zygentoma) allows us to date back the origin of *oskar* further than previous analyses, to at least 420 million years ago (Misof, et al. 2014), before the divergence of Ametabola from the remaining insect lineages.

We then explored the distribution of *oskar* sequences across insect phylogeny. Interestingly, we identified multiple lineages where *oskar* appeared to have been lost independently, including confirming the previously reported (Lynch, et al. 2011) losses from the genomes of the red flour beetle *Tribolium castaneum*, the honeybee *Apis mellifera*, and the silk moth *Bombyx mori* (Figure 3). Notably, within Lepidoptera we identified *oskar* homologs in only

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

three species, despite the fact that we searched 232 available lepidopteran sequence datasets, including 17 well-annotated RefSeq genomes and 135 transcriptomes (Figure 3 and Supplementary Figure S3). In principle, this apparent widespread absence of *oskar* in Lepidoptera could be due to unusually rapid evolution of the *oskar* sequence in this lineage, which might render lepidopteran *oskar* homologs undetectable by our methods. However, we note that the only four lepidopteran homologs we detected all belonged to species of the basally branching *Adelidae* and *Palaephatidae* families. We therefore favor the interpretation that *oskar* was lost from a last common ancestor of *Meessiidae* and *Palaphaetidae*, approximately 180 million years ago, with the consequence that the majority of extant lepidopteran lineages lack an *oskar* homolog (Supplementary Figure S3) (Mitter, et al. 2017; Kawahara, et al. 2019).

The Hemiptera also appear to have lost *oskar*, based on our analysis of the 222 datasets available for this clade, including 12 RefSeq genomes and 192 transcriptomes. However, we did identify an *oskar* homolog in the Thysanoptera, which is a hemipteran sister group (Misof, et al. 2014). Finally, we identified *oskar* homologs in only four of the 11 orders of the Polyneoptera for which data were available. With the exception of Mantodea (13 transcriptomes), the four orders with detectable *oskar* sequences all had more than ten available sequence datasets (Plecoptera: three genomes and eight transcriptomes; Orthoptera: three genomes and 28 transcriptomes; Phasmatodea: 13 genomes and 31 transcriptomes; Blattodea: five genomes and 51 transcriptomes). The remaining orders had fewer than eight datasets each available for analysis (Figure 3; Supplementary Table S1), which could account for the apparent paucity of *oskar* genes in this group. However, we cannot rule out the possibility that *oskar* in the Polyneoptera may have diverged beyond our ability to detect it, or that it may have been lost multiple times, as observed for multiple holometabolous orders.

As well as multiple convergent losses of *oskar*, we also uncovered evidence for independent instances of duplication of the *oskar* locus. We defined a putative duplication instance as two or more *oskar* sequences (possessing both a LOTUS and OSK domain as per our definition) in the same species that shared less than 80% sequence similarity. All of these events were detected within the Hymenoptera. We therefore performed a phylogenetic analysis of the hymenopteran sequences to test the hypothesis that these were the result of duplication events (Figure 4; Supplementary Figure S4). Our analysis of hymenopteran *oskar* sequences recovered previously published hymenopteran phylogenetic relationships (Peters, et al. 2017).

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

We found that *oskar* was duplicated in the four Figitidae species studied, a family of parasitoid wasps. Moreover, one out of ten examined Cynipidae species, as well as the only Ceraphronidae species examined, also harbored a duplicated *oskar* sequence. Multiple *oskar* duplications were also identified in the Chalcidoid wasps, notably in the Mymaridae (all three species studied), the Eupelmidae (two out of three species), the Aphelinidae (both species) and the Pteromalidae (one out of 17 species). Finally, we identified two additional apparently independent duplication events in the Aculeata, one in the wasp *Polistes fuscatus* (of 29 Vespidae, including three additional *Polistes* species, two with RefSeq genomes (*P. canadensis* and *P. dominula*) in which *oskar* was identified in single copy), and one in the red imported fire ant *Solenopsis invicta* (of 41 Formicidae species, including the congeneric *S. fugax*, with a GenBank genome in which *oskar* was identified in single copy).

Evidence for oskar *expression in multiple somatic tissues*

In studied insects to date, *oskar* is expressed and required in one or both of the germ line (Juhn and James 2006; Juhn, et al. 2008; Lynch, et al. 2011; Lehmann 2016) or the nervous system (Ewen-Campen, et al. 2012; Xu, et al. 2013). We asked whether these expression patterns could be detected in the insects studied here. To this end, we downloaded all available metadata for the transcriptomes analyzed here, to obtain information on the source tissues and developmental stages. We obtained these data for 371 out of the 1123 transcriptomes in our analysis, including both holometabolous and hemimetabolous orders (see Methods: *TSA metadata parsing and curation*). To first explore the distribution of *oskar* expression in the brain and the germ line, we binned the different tissues reported in the metadata into two categories, brain or germ line. This was done independently of the developmental stage (if that information was included in the metadata) by creating a mapping table and checking the extracted tissues against this table (Supplementary Table S3 at GitHub repository ***TableS3_germline_brain_table.csv***). We then cross referenced our homology detection with these metadata. We found evidence for *oskar* expression in the germ line of four orders (Phasmatodea, Hymenoptera, Coleoptera and Diptera), and in the brain of five orders (Orthoptera, Blattodea, Hymenoptera, Coleoptera, Diptera) (see Methods: *TSA metadata parsing and curation* for details on keyword extractions). For the vast majority of the datasets examined, transcriptomes were not generated with comparable methods for different organ

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

systems from the same species, such that we cannot make strong statements about the relative expression levels of *oskar* in the reproductive and nervous systems. However, we did perform a limited assessment of this question using previously published transcriptomes from the mosquito *Aedes aegypti* (Diptera) (Matthews, et al. 2016) (Supplementary Figure S5) and the cricket *Gryllus bimaculatus* (Orthoptera) (Whittle, Kulkarni, Chung, et al. 2021; Whittle, Kulkarni and Extavour 2021) (Supplementary Figure S6), and RT-PCR on isolated gonads and heads from *Drosophila melanogaster* (Diptera), the weevil *Callosobruchus maculatus* (Coleoptera) and the stick insect *Aretaon asperrimus* (Phasmatodea) (Supplementary Materials). For *D. melanogaster*, significant expression was detected only in female gonads (Supplementary Figure S7). For the remaining four species, while *oskar* transcripts were detected in both gonads and heads, levels appeared higher in gonads than in brains (Supplementary Figure S5-S7).

In addition, we found evidence of *oskar* expression in several somatic tissues not previously implicated in studies of *oskar* expression and function. These tissues included the midgut (*Polistes fuscatus, Sitophilus oryzae*), fat body (*Polistes fuscatus, Arachnocampa luminosa*), salivary gland (*Culex tarsalis, Anopheles aquasalis, Leptinotarsa decemlineata*), venom gland (*Culicoides sonorensis, Fopius arisanus*), and silk gland (*Bactrocera cucurbitae*) (Supplementary Figure S8). In terms of developmental stage, we detected expression of *oskar* during embryonic, larval or nymphal stages only in holometabolous insects, and for most hemimetabolous insects, *oskar* was detected in transcriptomes derived from adults (Figure 3). However, it is important to note that for most species, transcriptomes were available only from adult tissues, rather than from a full range of developmental stages (Supplementary Figure S5). We therefore cannot rule out the possibility that *oskar* expression at pre-adult stages is also a feature of multiple Hemimetabola. Indeed, we previously reported that *oskar* is expressed and required in the embryonic nervous system of a cricket, a hemimetabolous insect (Ewen-Campen, et al. 2012).

The Long Oskar domain is an evolutionary novelty specific to a subset of Diptera

D. melanogaster has two isoforms of Oskar (Markussen, et al. 1995): Short Oskar, containing the LOTUS, OSK and interdomain regions, and Long Oskar, containing all three domains of Short Oskar as well as an additional 5’ domain (Supplementary Figure S9). It was previously

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

reported that Long Oskar was absent from *N. vitripennis*, *C. pipiens* and *G. bimaculatus* (Lynch, et al. 2011; Ewen-Campen, et al. 2012), and within our alignment of Oskar sequences we could only detect the Long Oskar isoform within Diptera. Therefore, using our dataset, we asked when these two isoforms had evolved. We selected the dipteran sequences from our Oskar alignment and then grouped the sequences by family. We plotted the amino acid occupancy at each alignment position (Supplementary Figure S9), and found that Long Oskar predates the Drosophilids, being identified as early as the *Pinpunculidae* (Supplementary Figure S9). Moreover, following the evolution of the Long Oskar isoform, the Long Oskar domain was retained in all families except for the *Glossinidae* and *Scathophagidae*. However, given that we identified only eight and two Oskar sequences for these families respectively, we cannot eliminate the possibility that apparent absence of the Long Oskar domain in these groups reflects our small sample size, rather than true evolutionary loss.

The LOTUS and OSK domains evolved differently between hemimetabolous and holometabolous insects

The fact that an *oskar*-dependent germ plasm mode of germ line specification mechanism has been identified only in holometabolous insects suggests that *oskar* may have been co-opted in this clade for this function (Ewen-Campen, et al. 2012). Under this hypothesis, evolution of the *oskar* sequence in the lineage leading to the Holometabola may have changed the physico-chemical properties of Oskar protein, such that it acquired germ plasm nucleation abilities in these insects. To test this hypothesis, we asked whether there were particular sequence features associated with Oskar proteins from holometabolous insects, in which Oskar can assemble germ plasm, and hemimetabolous insects, which appear to lack *oskar*-dependent germ plasm. In particular, we assessed the differential conservation of amino acids at particular positions across Oskar and asked if these might be predicted to change the physico-chemical properties of Oskar in specific ways that could potentially be relevant to germ plasm nucleation. We used the Valdar score (Valdar 2002) as the main conservation indicator for this study (see GitHub file ***scores.csv***), as this metric accounts not only for transition probabilities, stereochemical properties and amino acid frequency gaps, but also for the availability of sequence diversity in the dataset. It computes a weighted score, where sequences from less well-represented clades contribute proportionally more to the score than sequences from overrepresented clades. Due

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

317 to the highly unbalanced availability of genomic and transcriptomic data between
318 hemimetabolous and holometabolous sequences (Supplementary Table S1; Figure 3) the
319 choice of a weighted score was necessary to avoid biasing the results towards insect orders
320 such as Diptera or Hymenoptera. To study the difference between hemimetabolous and
321 holometabolous sequences, we did not use the Valdar score directly, but instead computed the
322 conservation ratio between both groups for each position, which we call the Conservation bias
323 (See Methods: Computation of the Conservation Bias). We plotted the conservation bias on the
324 solved three-dimensional crystal structure of the *D. melanogaster* LOTUS and OSK domains
325 (Jeske, et al. 2015; Yang, et al. 2015) to ask whether specific functionally relevant structures
326 showed phylogenetic or other patterns of residue conservation (Figure 5).
327 First, we asked if the conservation score at the scale of domains was different between
328 holometabolous and hemimetabolous sequences. We observed that the conservation bias for
329 the LOTUS domain was centered around a mean of 1.00, indicating that both Holometabola and
330 Hemimetabola displayed a similar conservation of the LOTUS domain (Figure 5a). For the OSK
331 domain however, the conservation bias was centered around 0.84, indicating that the
332 hemimetabolous sequences displayed a higher level of conservation compared to
333 holometabolous sequences (Figure 5a). To interrogate specific biochemical hypotheses, we
334 then examined the degree of conservation bias in different regions of the protein structure. We
335 asked if the amino acids of the β sheets of the LOTUS domain thought to be involved in
336 dimerization of the protein (Jeske, et al. 2015; Yang, et al. 2015) displayed conservation bias.
337 Both β sheets had an overall even bias (mean: 1.03 and 1.05 for β1 and β2 respectively) between
338 both groups (Figure 5b). Second, as we had observed that hemimetabolous OSK was more
339 conserved overall than holometabolous OSK, we asked if there were any clear patterns of
340 conservation bias in specific regions of the OSK domain (Figure 5a and b). We found that some
341 of the secondary structures within the OSK domain showed a differential conservation (α2: 0.54,
342 α6: 0.42, β2: 0.52), whereas other structures were within less than 0.1 of the median value for
343 OSK. Moreover, we observed a large pocket of amino acids showing a conservation bias
344 towards hemimetabolous sequences located on the surface of OSK (Figure 5c). This particular
345 area contains the previously reported important amino acids for the RNA binding function of OSK
346 (Jeske, et al. 2015; Yang, et al. 2015) namely, R442, R436 and R576. The electrostatic
347 properties at those positions were conserved in the holometabolous sequences R436: 0.36,

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

R442: 0.29 and R576: 0.81 (Figure 5d), but not in hemimetabolous sequences. In other words, these specific amino acid residues are outliers in that they are more specifically conserved in holometabolous OSK sequences, but are located within a domain that overall is more conserved in Hemimetabola.

To gain further insight into the differences in conservation across insects, we reduced the multiple sequence alignment dimensionality using a Multiple Correspondence Analysis (MCA), an equivalent of PCA for categorical variables (Lebart, et al. 1984). We performed the dimensionality reduction for the full-length Oskar sequence alignment as well as for the LOTUS and OSK alignments (Supplementary Figure S10. Interestingly, we found that most of the variance in sequence space was due to dipterans and hymenopterans (Supplementary Figure S10). When we considered the OSK domain only, we identified clusters of *Drosophilidae*, *Culicidae* and *Formicidae* sequences (Supplementary Figure S10). This clustering is also reflected for the LOTUS domain, where the *Drosophilidae* and *Culicidae* contribute to a high amount of variance in the first MCA dimension. However, for the LOTUS domain, the *Formicidae* sequences do not cluster away from other Oskar sequences (Supplementary Figure S10). This suggests that the LOTUS domain of Diptera diverged in sequence between *Drosophilidae* and *Culicidae*.

Evidence for evolution of stronger dimerization potential of the Oskar LOTUS domain in Holometabola

The LOTUS domain dimerizes *in vitro* through electrostatic and hydrophobic contacts of Arg215 of the β2 sheet and Thr195, Asp197 and Leu200 of the α2 helix (Jeske, et al. 2015; Yang, et al. 2015). To date, however, the biological significance of Oskar dimerization remains unknown. Moreover, the dimerization of the LOTUS domain does not appear to be conserved across all Oskar sequences (Jeske, et al. 2015). Specifically, ten LOTUS domains from non-drosophilid species were tested for dimerization, and only LOTUS domains from *Drosophilidae*, *Tephritidae* and *Pteromalidae* formed homodimers (Jeske, et al. 2015). The other sequences tested, from *Culicidae*, *Formicidae* and *Gryllidae*, remained monomeric under the tested conditions (Jeske, et al. 2015). We selected the LOTUS sequences in our alignment from those six families and placed them into one of two groups, dimeric and monomeric LOTUS, under the assumption that any sequence from that family would conserve the dimerization (or absence thereof) properties

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

previously reported (Jeske, et al. 2015). We asked whether we could detect any evolutionary changes between the two groups in properties of known important dimerization interfaces and residues in our sequence alignment (Jeske, et al. 2015).

In the *D. melanogaster* structure, two key amino acids, D197 and R215, are predicted to form hydrogen bonds that stabilize the dimer (Jeske, et al. 2015). We found that in the dimer group, the electrostatic properties of these two amino acids are highly conserved (-0.75 for D197 and 0.81 for R215), while in the monomer group the electrostatic interaction is not conserved (0.03 for D197 and -0.11 for R215) (Figure 6e). Given the differential conservation between the two groups, our results support the previous finding that disrupting this interaction prevents dimerization (Jeske, et al. 2015). L200 was previously hypothesized to stabilize the interface via hydrophobic forces (Jeske, et al. 2015). We observed that the hydrophobicity of this residue is highly conserved in the dimer group (L200: 0.89), but that in the monomer group this residue is hydrophilic (L200: 2.33) (Figure 6f). In sum, our analyses show that key amino acids in the LOTUS domain evolved differently in distinct insect lineages, in a way that may explain why some insect LOTUS domains dimerize and some do not.

Conservation of the Oskar-Vasa interaction interface

Next, we asked whether we could detect differential conservation of regions or residues within the LOTUS-Vasa interface. It was previously reported that the LOTUS domain of Oskar acts as an interaction domain with Vasa (Jeske, et al. 2017), a key protein with a conserved role in the establishment of the animal germ line (Hay, et al. 1990; Lasko 2013). The interaction between *D. melanogaster* Oskar's LOTUS domain and Vasa is through an interaction surface situated in the pocket formed by the helices α2 and α5 of the LOTUS domain (Figure 6a b and c). Due to the essential role that *vasa* plays in germ line determination (reviewed in Raz 2000; Noce, et al. 2001; Extavour and Akam 2003; Ewen-Campen, et al. 2010; Lasko 2013), and the potential co-option of *oskar* to the germ line determination mechanism in Holometabola (Ewen-Campen, et al. 2012), we hypothesized that evolutionary conservation of the residues of this interface might be detectable. First, we observed that the residues of the LOTUS domain α2 and α5 helices, which directly contact Vasa (Jeske, et al. 2017) were highly conserved overall (α2 average Valdar score 0.49; α5 Valdar score 0.56) (Figure 6b). Specifically, we observed that the previously *in vitro*-confirmed Vasa interacting amino acids A162 and L228 of the LOTUS domain

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

were highly conserved (Valdar score: 0.64 for both residues) (Jeske, et al. 2017). We also noted that Q235 and H227 of the LOTUS domain α5 helix are also highly conserved, suggesting them as putative novel important interaction partners (Valdar score: 0.90 and 0.90 for both residues) (Figure 6b). Moreover, facing the LOTUS domain H227 is Vasa M540, which may act as a proton donor to form a hydrogen bond between the histidine ring and the sulfur atom of the methionine (Pal and Chakrabarti 2001) (Figure 6b and b'). The LOTUS domain α2 helix is overall slightly less conserved than the LOTUS domain α5 helix (Valdar score: 0.49 vs 0.56) (Figure 6a, b'', c''), but hydrophobic properties are conserved on one side of the α2 helix (Figure 6c, c') forming a motif of conserved amino acid properties (Figure 6c'').

Previous reports have hypothesized that the *D. melanogaster* LOTUS domain could act as a dsRNA binding domain (Anantharaman, et al. 2010; Callebaut and Mornon 2010). However, in *D. melanogaster*, it was later reported that the LOTUS domain did not bind to nucleotides (Jeske, et al. 2015). Therefore, using our dataset we assessed the potential RNA binding properties of LOTUS domains to test the conservation of this prediction. We used the RNABindR algorithm (Terribilini, et al. 2007) to predict potential RNA binding sites of the LOTUS domain, and computed a conservation score for each position (Terribilini, et al. 2007). We found that the α5 helix is the location in the LOTUS domain that has the most conserved prediction for RNA binding (Figure 6d). We therefore suggest that the possibility that LOTUS binds RNA directly warrants further experimental examination.

Finally, we asked whether the secondary structure of the LOTUS domain might be conserved. Secondary structures are often indicative of the tertiary structure of a domain. Therefore, we reasoned that the secondary structure might be conserved even if the sequence varies. We submitted the LOTUS sequences from all identified Oskar homologs to the Jpred4 servers (Drozdetskiy, et al. 2015) for secondary structure prediction and mapped the results onto the Oskar alignment we obtained. We found that the secondary structure of LOTUS is highly conserved throughout Oskar homologs, with the exception of the α1 helix (Supplementary Figure S11) which displays a low conservation score of 0.19 (Figure 6a).

The core of the OSK domain is conserved

We asked whether the OSK domain showed any differential conservation across the different parts of the domain. We found that the OSK domain of Oskar showed an overall conservation

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

441 across all insects, similar to the LOTUS domain (Valdar score: 0.51) (Figure 7a). However, the
442 conservation pattern is higher in the core amino acids (Valdar score average of core amino acid:
443 0.54) when compared to the residues at the surface (Valdar score average for surface amino
444 acid: 0.23) (Figure 7a). Despite the overall low conservation of the residues at the surface of the
445 OSK domain, we found that the electrostatic properties are conserved overall (electrostatic
446 conservation score >0; conserved) in the previously reported putative RNA binding pocket
447 (Yang, et al. 2015). However, as previously mentioned, this conservation is stronger in
448 holometabolous sequences (Figure 5d). These results are in accordance with the potential role
449 of OSK as an RNA Binding domain in the context of germ plasm assembly (Jeske, et al. 2015;
450 Yang, et al. 2015). We also submitted the OSK sequences to the same secondary structure
451 analysis performed on LOTUS. We found that, as for the LOTUS domain, the secondary
452 structure of OSK is highly conserved throughout all insect sequences analyzed (Supplementary
453 Figure S11).
454 We then asked if the conservation patterns observed at the core of OSK were clustered
455 in sequence motifs. When we looked at the location of the highly conserved amino acids, we
456 found that the conservation was driven by four well-defined sequence motifs (Figure 7c, c', c'',
457 c'''). Given that *oskar* plays different roles in Holometabola and Hemimetabola, we asked
458 whether the conserved OSK motifs showed any difference in conservation between these two
459 groups. Of the four highly conserved OSK core motifs (Figure 7c, c', c'', c'''), two of them (Figure
460 7c: Valdar average score: 0.80 and c'' Valdar average score: 0.71) were conserved across all
461 insects, but the other two showed differential conservation between the holometabolous and
462 hemimetabolous sequences (Figure 7c': Valdar score average Holometabola: 0.78;
463 Hemimetabola: 0.58 c'':' Valdar score average Holometabola: 0.70, Hemimetabola: 0.55).
464 Finally, we noted that only one of the affected OSK domain residues in known loss of function
465 *oskar* alleles affecting posterior patterning in *D. melanogaster*, S457, is conserved across all
466 insects (Valdar score: 0.86). This suggests that the role of the other previously reported
467 important amino acids in the function of *D. melanogaster* OSK (Yang, et al. 2015) might not be
468 conserved in other insects (red positions in Figure 7c, c', c'', c''').

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

Discussion

An expanded collection of oskar *homologs*

oskar provides a powerful case study of functional evolution of a gene with an unusual genesis (Blondel, et al. 2020). Here, we gathered the most extensive set of homologous *oskar* sequences to date. However, most insect genomic and transcriptomic data have been generated from only a few orders, and the vast majority from the Holometabola. Diptera, Lepidoptera, Coleoptera, Hymenoptera and Hemiptera represent 82% of the datasets available at the time of this analysis. We emphasize that expanded taxon sampling, particularly for the Hemimetabola, will be critical for further studies of the evolution of protein function across insects. Moreover, only a small proportion (27% for tissue type, 26% for organism stage, and 14% for sex) of the TSA datasets contained usable metadata regarding the stage and tissue type sampled. Future standardization of the nature and format of transcriptomic metadata would also be a worthwhile endeavor that could increase the efficiency and efficacy of future work.

Convergent losses and duplications of oskar *in insect evolution*

A previous report suggested that *oskar* had been lost from the genome of the silk moth *B. mori* (Lynch, et al. 2011). Our analysis of 232 datasets across 44 of the 126 described lepidopteran families (Kawahara, et al. 2019) strongly suggests that the loss of *oskar* in the Lepidoptera (butterflies and moths) is not unique to the silk moth, but rather occurred early and repeatedly in lepidopteran evolution. The fact that *oskar* is a component of the oosome at the posterior of the oocyte (the wasp germ plasm analog (Quan, et al. 2019)) and required for germ cell formation in the wasp *Nasonia vitripennis* (Lynch, et al. 2011) implies that a common ancestor of Holometabola had already established an *oskar*-dependent inheritance mode of germ line specification. Therefore, the apparent subsequent loss in nearly all Lepidoptera examined of a gene responsible for the establishment of the germ plasm in other Holometabola might seem unexpected. Few studies have directly addressed the molecular mechanisms of germ cell specification in Lepidoptera. In *B. mori* (Bombicidae), *vasa* mRNA (Nakao 1999) and protein (Nakao, et al. 2006), and the transcripts of one of four *nanos* homologs (*nanos-O*) (Nakao, et al. 2008), have been detected in a region of ventral cortical cytoplasm in pre-blastoderm stage embryos. As putative primordial germ cells form in this location at later stages (Miya 1958), some

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

500 authors have speculated that a germ plasm, located ventrally rather than posteriorly, may specify
501 germ cells in this moth (Toshiki, et al. 2000; Nakao, et al. 2008). However, recent knockdown
502 experiments showed that maternal *nanos-O* is dispensable for germ cell formation (Nakao and
503 Takasu 2019), consistent with a zygotic, inductive mechanism. In the butterfly *Pararge aegeria*
504 (Nymphalidae), no *oskar* homolog has been identified in the genome (Carter, et al. 2013), but
505 the transcripts of one of four identified *nanos* homologs (*nanos-O*) have been detected in a small
506 region of ventral cortical ooplasm, again prompting speculation that this lepidopteran may also
507 deploy a germ plasm (Carter, et al. 2015). We suggest that if these or other Lepidoptera do
508 indeed rely on germ plasm to specify their germ line, they may do so using a germ plasm
509 nucleator other than Oskar. For most studied Lepidoptera, however, classical embryological
510 studies report the first appearance of primordial germ cells at post-blastoderm stages, either
511 from the ventral midline of the cellular blastoderm or early germ band (Woodworth 1889; Tomaya
512 1902; Sehl 1931; Miya 1953, 1958, 1975; Tanaka 1987), from the coelomic sac mesoderm of
513 the abdomen (Johannsen 1929; Eastham 1930; Saito 1937; Presser and Rutschky 1957;
514 Kobayashi and Ando 1984), or from the primary ectoderm of the caudal germ band (Schwangart
515 1905; Lautenschlager 1932; Ando and Tanaka 1979; Tanaka 1987; Guelin 1994)
516 (Supplementary Figure S3). Taken together, these data suggest that an inductive mechanism
517 may operate to specify germ cells in most moths and butterflies. We speculate that the loss of
518 *oskar* from most lepidopteran genomes may have facilitated or necessitated secondary
519 reversion to the hypothesized ancestral inductive mechanism for germ line specification.
520 Another order with apparent near-total absence of *oskar* homologs is the Hemiptera (true
521 bugs), whose sister group Thysanoptera (thrips) nevertheless possesses *oskar*. This secondary
522 loss of *oskar* from a last common hemipteran ancestor correlates with the reported post-
523 blastoderm appearance of primordial germ cells in the embryo. Classical studies on most
524 hemipteran species describe germ cell formation as occurring after cellular blastoderm
525 formation, on the inner (yolk-facing) side of the posterior blastoderm surface (Metschnikoff 1866;
526 Witlaczil 1884; Will 1888; Mellanby 1935; Butt 1949; Kelly and Huebner 1989; Heming and
527 Huebner 1994). A notable exception to this is the parthenogenetic pea aphid *Acyrthosiphon*
528 *pisum*, for which strong gene expression and morphological evidence supports a germ plasm-
529 driven germ cell specification mechanism in both sexual and asexual modes (Miura, et al. 2003;
530 Chang, et al. 2006; Lin, et al. 2014). In contrast, studies of the aphids *Aphis plantoides, A. rosea*

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

and *A. pelargonii* describe no germ plasm, and post-blastoderm germ cell formation (Metschnikoff 1866; Witlaczil 1884; Will 1888). However, the genomes of all aphids studied here, including *A. pisum* and three *Aphis* species, appear to lack *oskar*. This suggests that germ plasm assembly in *A. pisum* either does not require a nucleator molecule or uses a novel non-Oskar nucleator.

In the Hymenoptera (ants, bees, wasps and sawflies), our results strongly suggest that *oskar* was lost from the genome of the last common ancestor of bees and spheroid wasps (Supplementary Figure S12). Our analysis further suggests multiple additional independent losses in as many as 25 other hymenopteran lineages, including some for which good quality RefSeq genomes were available (e.g. the slender twig ant *Pseudomyrmex gracilis* or the wheat stem sawfly *Cephus cinctus* (Supplementary Figure S12). However, it would be premature to draw strong conclusions about the number of independent losses given the predominance of transcriptome data in the Hymenoptera.

In addition to convergent losses of *oskar*, we also found evidence for clade-specific duplications of *oskar* in the Hymenoptera. Seven of the nine families containing these putative duplications are families of parasitoid wasps; the remaining two families are ants (Formicidae) and the group of yellowjackets, hornets, and paper wasps (Vespidae) (Figure 4). The phylogenetic relationships of these groups make it highly unlikely that a duplication occurred only once in their last common ancestor, which would be the last common ancestor of all wasps, bees and ants (i.e. Apocrita, all hymenopterans except sawflies) (Supplementary Figure S12). We suggest that the most parsimonious hypothesis is one of three to five independent duplications of *oskar*, followed by at least nine to 14 independent reversions to a single copy, or total loss of the locus (Supplementary Figure 9).

No notable life history characteristics appear to unite those species with multiple *oskar* homologs: they include eusocial and solitary, sting-bearing and stingless, parasitoid and non-parasitic insects. To our knowledge, neither is there anything unique about the germ line specification process in Hymenoptera with one or more than one *oskar* homolog. Most Hymenoptera appear to use a germ plasm-driven mechanism to specify germ cells in early blastoderm stage embryos (Supplementary Figure S12 and references therein), and we identified *oskar* homologs for all such species described in the embryological literature (Supplementary Figure S12). In the notable example of the honeybee *Apis mellifera*, in which

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

562 cytological and molecular evidence suggests germ cell arise from abdominal mesoderm
563 (Bütschli 1870; Nelson 1915; Fleig and Sander 1985, 1986; Zissler 1992; Gutzeit, et al. 1993;
564 Dearden 2006), we identified no *oskar* homolog in its well-annotated genome (Supplementary
565 Figure S12), as noted previously by other authors (Lynch, et al. 2011). However, no major
566 differences in germ plasm or pole cell formation have been reported in species or families of
567 ants or wasps with duplicated *oskar* loci, compared with close relatives that possess *oskar* in
568 single copy (e.g. compare the ants *Solenopsis invicta* (at least 2 *oskars*) and *Aphaenogaster*
569 *rudis* (1 *oskar*) (Khila and Abouheif 2008), or the pteromalid wasps *Nasonia vitripennis* (1 *oskar*)
570 (Lynch and Desplan 2010; Lynch, et al. 2011; Quan, et al. 2019) and *Otitesella tsamvi* (2 *oskars*).
571 Thus, future studies that independently abrogate the functions of each paralog individually, will
572 be needed to determine the biological significance, if any, of these *oskar* duplications.
573
574 *Evolution of the Long Oskar domain*
575 We showed that the Long Oskar domain is an evolutionary novelty confined to a subset of
576 Diptera. This raises the question of whether the evolution of this domain led to any novel
577 functional properties of *oskar* in these Diptera, relative to its functions in other insects. The only
578 data available on the specific functions of the Long Oskar domain are from studies on *D.*
579 *melanogaster*. The Long Oskar (606 amino acids: possessing the Long Oskar domain) and Short
580 Oskar (468 amino acids: lacking the Long Oskar domain) isoforms are generated by translation
581 of *oskar* mRNA from alternate initiation codons within the same transcript (Markussen, et al.
582 1995). Short Oskar alone cannot maintain *oskar* mRNA or either protein isoform at the posterior
583 pole of the oocyte or embryo (Vanzo and Ephrussi 2002). However, Short Oskar alone is able
584 to promote the formation of pole cells, albeit many fewer than wild type (Markussen, et al. 1995).
585 In contrast, Long Oskar alone can anchor *oskar* mRNA, Oskar protein, and mitochondria at the
586 posterior pole, but cannot promote pole cell formation (Rongo, et al. 1997; Vanzo and Ephrussi
587 2002; Hurd, et al. 2016). *In vitro*, Short Oskar has a higher affinity for germ plasm components
588 than Long Oskar (Breitwieser, et al. 1996; Babu, et al. 2004; Anne and Mechler 2005; Megosh,
589 et al. 2006; Suyama, et al. 2009; Anne 2010). Furthermore, Short Oskar associates with the
590 cytoplasmic germ granules themselves, while Long Oskar instead associates with endosomal
591 membranes (Vanzo, et al. 2007). These observations have led to the model that Long Oskar's
592 main role is to recruit and anchor Short Oskar to the posterior, where Short Oskar is responsible

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

for germ plasm assembly *per se* (Markussen, et al. 1995; Vanzo and Ephrussi 2002; Tanaka and Nakamura 2008; Tanaka, et al. 2011; Hurd, et al. 2016).

The molecular basis for the apparently distinct roles of these two isoforms remains largely unclear, and are unlikely to reside entirely within the Long Oskar domain. *In vivo* assessments of the 139-amino acid Long Oskar domain alone show that it is necessary and sufficient to maintain mitochondria at the oocyte cortex (Hurd, et al. 2016). This Long Oskar domain-mediated mitochondrial maintenance requires an intact F-actin cortical cytoskeleton, which is modified by the presence of the Long Oskar domain (Tanaka and Nakamura 2008; Tanaka, et al. 2011; Hurd, et al. 2016). Compared to controls, *long oskar* null mutant flies (possessing only Short Oskar) generate fewer PGCs with fewer mitochondria, and their ovaries lack germ cells more often than controls (Hurd, et al. 2016).

Although the Long Oskar isoform thus appears to play important and unique roles in functional germ plasm assembly in *D. melanogaster*, these roles appear to be performed perfectly well by the single isoform possessed by nearly all other insects, which in terms of sequence is essentially equivalent to Short Oskar. One or more of posterior *oskar* and germ plasm localization, posterior pole cell formation, and mitochondrial enrichment within germ plasm have been reported for species of ants, bees, wasps, beetles, mosquitoes and flies that all lack a Long Oskar isoform (Nardon 1971; Jaglarz, et al. 2003; Goltsev, et al. 2004; Zhurov, et al. 2004; Juhn and James 2006; Nardon 2006; Juhn, et al. 2008; Lynch, et al. 2011; Yoon, et al. 2019; Rafiqi, et al. 2020). We note, however, that many of these species are reported to possess an oosome, which is a single, morphologically distinct discrete non-membrane-bound organelle that houses germ plasm components (Meng 1968; Nardon 1971; Klag and Bilinski 1993; Jaglarz, et al. 2003; Zhurov, et al. 2004; Nardon 2006; Lynch, et al. 2011; Quan, et al. 2019). This is distinct from most *Drosophila* species for which data are available, whose germ plasm is in the form of multiple smaller granules loosely clustered near the posterior cortex (Mahowald 1962; Mahowald 1968; Mahowald, et al. 1976). We therefore speculate that the evolution of the Long Oskar domain may have enabled tight cortical anchoring of germ plasm components via interaction with endosomes and/or the F-actin cytoskeleton, eliminating the need for an oosome to ensure integrity or local concentration of germ plasm.

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

Re-examination of potential interactions between the LOTUS domain and RNA

Proteins with a LOTUS domain commonly participate in nucleic acid binding (Williams, et al. 1993; Gajiwala and Burley 2000; Liu, et al. 2001; Aravind, et al. 2005; Lachke, et al. 2011; Cui, et al. 2013; Harami, et al. 2013; Mukherjee, et al. 2014). LOTUS domain-containing proteins, particularly RNA-binding proteins (Cui, et al. 2013), are often enriched in germ plasm (Anantharaman, et al. 2010; Callebaut and Mornon 2010), as are specific RNAs (Ephrussi, et al. 1991; Wang and Lehmann 1991; Jongens, et al. 1992; Smith, et al. 1992; Kobayashi, et al. 1995; Nakamura, et al. 1996; Mahowald 2001; Vanzo and Ephrussi 2002; Ewen-Campen, et al. 2010). However, to date there is no direct evidence that Oskar's LOTUS domain interacts directly with RNA. We were therefore intrigued to find that our bioinformatic analysis suggested that the LOTUS helix α5 might have binding RNA ability (Figure 6d). Consistent with the possibility that Oskar's LOTUS domain might somehow interact with RNA *in vivo*, we have observed that a loss of function *oskar* allele lacking the entire LOTUS domain (*oskar[ΔLOTUS]*), is unable to direct accumulation of Nanos protein in the germ plasm (Extavour lab, unpublished observation). If the OSK domain, which unlike the LOTUS domain, binds *nanos* mRNA *in vitro* (Jeske, et al. 2015; Yang, et al. 2015), were sufficient to ensure Nanos protein localization *via nanos* mRNA recruitment, then germ plasm in *oskar[ΔLOTUS]* flies should contain Nanos protein. Our opposite result could indicate that LOTUS plays a role in RNA binding and/or local translation of *nanos* mRNA. In principle, this could be indirect, for example aided by LOTUS-mediated oligomerization (Jeske, et al. 2015; Yang, et al. 2015), or it could be via direct LOTUS-RNA contacts that have not yet been detected in biochemical studies. Further, we note that LOTUS-RNA interactions have, to our knowledge, been probed biochemically and genetically only in *D. melanogaster*, which does not rule out the existence of such binding interactions in other insects.

Functional implications of differential conservation of regions of the LOTUS and OSK domains

We have identified novel conserved amino acid positions that we hypothesize are important for the Vasa binding properties of the LOTUS domain and the RNA properties binding of the OSK domain (Figures 6 and 7). Our observation of the conservation of the LOTUS domain α2 helix is consistent with its previously reported importance in LOTUS-Vasa binding (Jeske, et al. 2017).

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

652 In the α2 helix, we also observed high conservation of H227 and Q235. The positions of these
653 residues suggest they may contribute to the interaction between Vasa and LOTUS, but they
654 have not, to our knowledge, yet been implicated functionally *in vitro* or *in vivo*. We suggest they
655 should therefore be the target of future mutational studies. Moreover, evolution at the interface
656 between two proteins involve amino acids on both sides of the surface. Therefore, further studies
657 looking at potential co-evolution between Oskar and Vasa could shed light on whether the
658 conserved amino acids that we identified in the LOTUS domain interact with similarly conserved
659 Vasa residues, or whether evolutionary variations in Oskar-Vasa interactions may be explained
660 by co-evolution of specific residues at their interaction surfaces (Andreani, et al. 2020).

661 We also uncovered an interesting new conservation pattern within the OSK domain. The
662 conserved amino acids were more abundant in the core of the domain than on the surface. This
663 differential conservation might be relevant to the acquisition of a germ plasm nucleator role of
664 *oskar* in the Holometabla (Figure 5). We noted that the basic properties of surface residues
665 previously reported for *D. melanogaster* (Yang, et al. 2015) are conserved across insects, which
666 might indicate that the RNA binding properties of OSK observed in *D. melanogaster* (Jeske, et
667 al. 2015; Yang, et al. 2015) are also conserved throughout holometabolous insects. We
668 speculate that the comparatively low amino acid conservation of the surface residues in
669 Holometabolous OSK domains, which nevertheless display highly conserved basic properties,
670 could have allowed greater flexibility in the co-evolution of specific RNA binding partners for the
671 OSK domains of different lineages.
672
673 *OSK evolved differentially between holometabolous and hemimetabolous insects*
674 Finally, we observed a differential conservation of the OSK domain between hemimetabolous
675 and holometabolous insects. Specifically, we found that the OSK sequence was less conserved
676 across the Holometabola than across the Hemimetabola. This observation raises two potential
677 hypotheses regarding the role of the OSK domain in the functional evolution of Oskar. First,
678 perhaps the apparently relaxed purifying selection experienced by OSK in the Holometabola
679 was necessary for the co-option of *oskar* to a germ plasm nucleation role. Second, Oskar might
680 have a function in the hemimetabolous insects that requires strong conservation of OSK. More
681 studies on the roles and biochemical properties of OSK in hemimetabolous insects will be

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

required to test these hypotheses and further our understanding of the biological relevance of this differential conservation.

In conclusion, analysis of the large dataset of novel Oskar sequences presented here provides multiple new testable hypotheses concerning the molecular mechanisms and functional evolution of *oskar*, that will inform future studies on the contribution of this unusual gene to the evolution of animal germ cell specification.

Materials and Methods

Lead contact and materials availability

This study did not generate new unique reagents. This study generated new python3 code and supplementary files referred to below, all of which are available https://github.com/extavourlab/Oskar_Evolution. Requests for further information and requests for resources and reagents should be directed to and will be fulfilled by Cassandra G. Extavour (extavour@oeb.harvard.edu).

Experimental model and subject details

This study used no cell culture lines. This study used live samples of *Drosophila melanogaster* and *Callosobruchus maculatus* and ethanol-preserved samples of *Aretaon asperrimus*. The study also used previously generated genomic and transcriptomic datasets. All the information regarding how those datasets were generated can be found on their respective NCBI pages. The list of all the datasets used in this study can be found in the following files: **genome_insect_database.csv**, **transcriptome_insect_database.csv**, **genome_crustacean_database.csv**, and **transcriptome_crustacean_database.csv**.

Genome and transcriptome preprocessing

We collected all available genome and transcriptome datasets from the NCBI repository registered in September 2019 (Figure 2). NCBI maintains two tiers of genomic data: RefSeq, which contains curated and annotated genomes, and GenBank, which contains non-annotated assembled genomic sequences. Transcriptomes are stored in the Transcriptome Shotgun

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

Assembly (TSA) database, with metadata including details on their origin. Among the registered datasets, five genomes were not yet available, and 40 transcriptomes were only available in the NCBI Trace repository. As they did not comply with the TSA database standards, they were excluded from the analysis. To search for *oskar* homologs in datasets retrieved from GenBank, we needed to generate *in silico* gene model predictions. We used the genome annotation tool Augustus (Stanke et al. 2006), which requires a Hidden Markov Model (HMM) gene model. To use HMMs producing gene models that would be as accurate as possible for non-annotated genomes, we selected the most closely related species (species with the most recent last common ancestor) that possessed an annotated RefSeq genome. We then used the Augustus training tool to build an HMM gene model for each genome.

We automated this process by creating a series of python scripts that performed the following tasks:

1) ***1.1_insect_database_builder.py***: This script collects the NCBI metadata regarding genomes and transcriptomes. Using the NCBI Entrez API, it collects the most up to date information on RefSeq, GenBank, and TSA to generate two CSV files: *genome_insect_database.csv* and *transcriptome_insect_database.csv*.
2) ***1.2_data_downloader.py***: This is a python wrapper around the *rsync* tool that downloads the sequence datasets present in the tables created by (1). It automatically downloads all the available information into a local folder.
3) ***1.3_run_augustus_training.py***: This is a python wrapper around the Augustus training tool. It uses the metadata gathered using (1) and the sequence information gathered using (2) to build HMM gene models of all RefSeq datasets. It outputs sbatch scripts that can be run either locally, or on a SLURM-managed cluster. Those scripts will create unique HMM gene models per species.

At the time of this analysis (September 2019), 133 insect genomes were collected from the RefSeq database, 309 genomes from the GenBank database, and 1123 transcriptomes from the TSA database. All the accession numbers and metadata are available in the two tables (***genome_insect_database.csv*** and ***transcriptome_insect_database.csv***) provided in the supplementary files. This pipeline was repeated for crustaceans and the information can be

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

found in the following two files: ***genome_crustacean_database.csv*** and ***transcriptome_crustacean_database.csv***.

Creation of protein sequence databases

The classical approach for homology detection compares protein sequences to amino acid HMM corresponding to the gene of interest. Since we used three different NCBI databases, we performed the following preprocessing actions:

1) RefSeq: well-annotated genomes from NCBI contain gene model translation; no extra processing was required.
2) GenBank: Using the HMMs created from the RefSeq databases, we created gene models for each GenBank genome using Augustus and a custom HMM gene model. To choose which HMM gene model to use, we selected the one for each insect order that had the highest training accuracy. In the case where an insect order did not have any member in the RefSeq database, we used the model of the most closely related order. We then translated the inferred coding sequences to create a protein database for each genome. The assignment of the models used to infer the proteins of each GenBank genome is available in the ***Table_S4_models.csv*** available through the GitHub repository for this study at https://github.com/extavourlab/Oskar_Evolution. To automate the process, we created a custom python script available in the file **1.4_run_augustus.py**.
3) TSA: Transcriptomes were translated using the emboss tool Transeq (Madeira, et al. 2019). We used this tool with the default parameters, except for the six-frame translation, trim and clean flags. This generated amino acid sequences for each transcript and each potential reading frame.

Identification of oskar homologs

The *oskar* gene is composed of two conserved domains, LOTUS and OSK, separated by a highly variable interdomain linker sequence (Ahuja and Extavour 2014; Jeske, et al. 2015; Yang, et al. 2015). To our knowledge, no other gene reported in any domain of life possesses this domain composition (Blondel, et al. 2020). Therefore, here we use the same definition of *oskar* homology as in our previous work: a sequence possessing a LOTUS domain followed by an

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

interdomain region, and then an OSK domain (Blondel, et al. 2020). To maximize the number of potential homologs, we searched each sequence with the previously generated HMM for the LOTUS and OSK domains (Blondel, et al. 2020). The presence and order of each domain were then verified for each potential hit and only sequences with the previously defined Oskar structure were kept for further processing. We used the HMMER 3.1 tool suite to build the domain HMM (*hmmbuild* with default parameters), and then searched the generated protein databases (see *Creation of protein sequence databases* above) using those models (*hmmsearch* with default parameters). Hits with an E-value >= 0.05 were discarded. A summary of all searches performed is compiled in ***Table_S5_searches.csv*** in the GitHub repository for this study at https://github.com/extavourlab/Oskar_Evolution.

All the hits were then aligned with *hmmalign* with default parameters and the HMM of the full-length Oskar alignment previously generated (Blondel, et al. 2020). The resulting sequences were automatically processed to remove assembly artifacts, and potential isoforms. This filtration step was automated and went as follows: First, the sequences were grouped by taxon. Then each group of sequences was aligned using MUSCLE (Edgar 2004) with default parameters. The Hamming distance (Hamming 1950), a metric that computes the number of different letters between two strings, between each sequence in the alignment, was computed. If any group of sequences had a Hamming distance of >80%, then we only kept the sequence with the lowest E-value match. This created a set of sequences containing multiple *oskar* homologs per species only if they were the likely product of a gene duplication event. We then used the resulting new alignment to generate a new domain HMM and a new full-length Oskar HMM (using *hmmbuild* with default parameters) and ran further iterations of this detection pipeline until we could detect no new *oskar* homologs in the available sequence datasets. We called this final set the **filtered set** of sequences and used it in all subsequent homology analyses unless otherwise specified.

The Oskar sequences obtained are available in the following supplementary files: ***Oskar_filtered.aligned.fasta***, ***Oskar_filtered.fasta*** and ***Oskar_consensus.hmm***. The domain definitions for the LOTUS and OSK domains are available in the following supplementary files: ***Oskar_filtered.aligned.LOTUS_domain.fasta***, ***LOTUS_consensus.hmm***, ***Oskar_filtered.aligned.OSK_domain.fasta***, ***OSK_consensus.hmm*** (see ***1.5_Oskar_tracker.ipynb***).

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

Correlative analysis of assembly quality and absence of oskar

Using the metadata gathered previously from NCBI databases (see *Genomes and transcriptomes preprocessing* above) we created two pools of source data: genomes where we identified an *oskar* sequence, and genomes where we failed to find a sequence that met our homology criteria. We then compared the two distributions for each of the 8 available assembly statistics: (1) Contig and (2) Scaffold N50, (3) Contig and (4) Scaffold L50, (5) Contig and (6) Scaffold counts, and (7) Number of Contigs and (8) Scaffolds per genome length. Finally, we performed a Mann-Whitney U statistical analysis to compare the means of the two distributions (see ***2.1_Oskar_discovery_quality.ipynb***).

TSA metadata parsing and curation

Datasets in the TSA database are associated with a biosample object that contains all the metadata surrounding the RNA sequencing acquisitions. These metadata can include information about one or both the tissue of origin and the organism's developmental stage. We first automated the retrieval of these metadata using a custom python script that used the NCBI Entrez API (see ***2.3_Oskar_tissues_stages.ipynb***). However, the metadata proved to be complex to parse for the following reasons: (1) not all projects had the data entered in the corresponding tag, (2) some data contained typographical errors, and (3) multiple synonyms were used to describe the same thing with different words in different datasets. We therefore created a custom parsing and cleaning pipeline that corrected mistakes and aggregated them into a cohesive set of unique terms that we thought would be most informative to interpret the presence or absence of *oskar* homologs (see ***2.3_Oskar_tissues_stages.ipynb*** to see the mapping table). This strategy sacrificed some of the fine-grained information contained in custom metadata (for example "right leg" became "leg") but allowed us to analyze the expression of *oskar* using consistent criteria throughout all the datasets. This pipeline generated, for all available datasets, a table of tissues and developmental stages including *oskar* presence or absence in the dataset (see ***Oskar_all_tissues_stages.csv***).

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

Dimensionality reduction of Oskar alignment sequence space

The Oskar alignment was subjected to a Multiple Correspondence Analysis (MCA). Similar to a PCA, dimension vectors were first computed to maximize the spread of the underlying data in the new dimensions, except that instead of a continuous dataset, each variable (here an amino acid at a given position) contributes to the continuous value on that dimension. Once the projection vectors are computed, each sequence was then mapped onto the dimensions. Each amino acid position (column) in the alignment was considered a dimension with a possible value set of 21 (20 amino acids and gap). We first removed the columns of low information (columns that had less than 30% amino acid occupancy) using trimal (Capella-Gutierrez, et al. 2009) with a cutoff parameter set at 0.3. Then, the alignment was decomposed into its eigenvectors, and projected to the first three components. To perform this decomposition, we implemented a previously developed preprocessing method (Rausell, et al. 2010) in a python script (see ***MCA.py*** and ***2.8_Oskar_MCA_Analysis.ipynb***) and performed the eigenvector decomposition with the previously developed MCA python library (see *Key Resource Table*). We ran the same algorithm on the LOTUS domain, OSK domain, and full-length Oskar alignments obtained above (see *Identification of* oskar *homologs* above).

Phylogenetic inference of Oskar sequences in the Hymenoptera

We aligned all hymenopteran Oskar sequences using PRANK (Loytynoja 2014) with default parameters. We then manually annotated duplicated sequences by considering two sequences from the same species that had < 80% amino acid identity, as within-species duplications of *oskar*. We trimmed this alignment to remove all columns with less than 50% occupancy using trimal with the cutoff parameter set at 0.5. To reconstruct the phylogeny of these sequences, we used the maximum likelihood inference software RAxML (Stamatakis 2014) with a gamma-distributed protein model, and activated the flag for auto model selection. We ran 100 bootstraps and then visualized and annotated the obtained tree with Ete3 (Huerta-Cepas, et al. 2016) in a custom ipython notebook (see ***2.7_Oskar_duplication.ipynb***).

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

Calculation of Oskar conservation scores

Using the large set of homologous Oskar sequences obtained as described above, we computed different conservation scores for each amino acid position. This methodology relies on the hypotheses that if an amino acid, or its associated chemical properties at a particular position in the sequence are important for the structure and/or function of the protein, they will be conserved across evolution. We considered multiple conservation metrics, each highlighting a particular aspect of the protein's properties as described in the followng sections. The scores can be found in the supplementary file ***scores.csv***.

Computation of the Valdar score

The Valdar score (Valdar 2002) attempts to account for transition probabilities, stereochemical properties, amino acid frequency gaps, and, particularly essential for this study, sequence weighting. Due to the heterogeneity of sequence dataset availability, most Oskar sequences occupy only a small portion of insect diversity, primarily Hymenoptera, and Diptera. Sequence weighting allows for the normalization of the influence of each sequence on the score based on how many similar sequences are present in the alignment (Valdar 2002). We implemented the algorithm described in (Valdar 2002) in a python script (see ***besse_blondel_conservation_scores.py***), then calculated the conservation scores for the Oskar alignment we generated above.

Computation of the Jensen-Shannon Divergence score

Jensen-Shannon Divergence (JSD) (Lin 1991; Capra and Singh 2007) uses the amino acid and stereochemical properties to infer the "amount" of evolutionary pressure an amino acid position may be subject to. This score uses an information theory approach by measuring how much information (in bits) any position in the alignment brings to the overall alignment (Capra and Singh 2007). This score also takes into account neighboring amino acids in calculating the importance of each amino acid. We used the previously published python code to calculate the JSD of our previously generated Oskar alignment (Capra and Singh 2007) (see ***score_conservation.py***).

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

Computation of the Conservation Bias

The measure of differences in conservation between the holometabolous and hemimetabolous Oskar sequences presented in the results was done as follows: we first split the alignment into two groups containing the sequences from each clade (see ***2.4_Oskar_pgc_specification.ipynb***). Due to the high heterogeneity in taxon sampling between hemimetabolous and holometabolous insects, we ran a bootstrapped approximation of the conservation scores on holometabolous sequences. We randomly selected N sequences (N = the number of hemimetabolous sequences), computed the Valdar conservation score (see *Computation of the Valdar score* above), and stored it. After 1000 iterations, we computed the mean conservation score for each position for holometabolous sequences. For hemimetabolous sequences, we directly calculated the Valdar score using the method as described above (see *Computation of the Valdar score* above). For each position, we then computed what we refer to as the "conservation bias" between Holometabola and Hemimetabola by taking the ratio of the log of the conservation score Holometabola and Hemimetabola. Conservation Bias = $\log(\text{Valdar}_{holo}) / \log(\text{Valdar}_{hemi})$ for each position (see ***3.4_LogRatio_Bootstrap.ipynb***).

Computation of the electrostatic conservation score

To study the conservation of electrostatic properties of the Oskar protein we computed our own implementation of an electrostatic conservation score (see ***besse_blondel_conservation_scores.py***). Aspartic acid and Glutamic acid were given a score of -1, Arginine and Lysine a score of 1, and Histidine a score of 0.5. All other amino acids were given a score of 0. Then, we summed the electrostatic score for each sequence at each position and divided this raw score by the total number of sequences in the alignment. This computation assigns a score between -1 and 1 at each position, -1 being a negative charge conserved across all sequences, and 1 a positive charge.

Computation of the hydrophobic conservation score

To study the conservation of hydrophobic properties of the Oskar protein we implemented our own hydrophobic conservation score (see ***besse_blondel_conservation_scores.py***). At each position, each amino acid was given a hydrophobic score taken from a previously published

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

scoring table (Moon and Fleming 2011). (This table is implemented in the ***besse_blondel_conservation_score.py*** file for simplicity.) Scores at each position were then averaged across all sequences. This metric allowed us to measure the hydrophobicity conservation of each position in the alignment and is bounded between 5.39 and -2.20.

Computation of the RNA binding affinity score

RNA binding sites are defined as areas with positively charged residues and hydrophobic residues. To estimate the conservation of RNA binding sites in *oskar* homologs, we used RNABindR v2.0 (Terribilini, et al. 2007), an algorithm predicting putative RNA binding sites based on sequence information only. We automated the calculation for each sequence by writing a python script that submitted a request to the RNABindR web service (see ***RNABindR_run_predictions.py***). We then aggregated all results into a scoring matrix, and averaged the score obtained for each position. We call this score the RNABindR score and hypothesize that it reflects the conservation of RNA binding properties of the protein. Importantly, this score was obtained in 2017 for only a subset of 219 proteins used in this study (indicated in the supplementary files at: 03_Oskar_scores_generation/RNABindR_raw_sources). Since then, the RNABindR server has been defunct and we could not repeat those measurements as the source code for this software is unavailable.

Computation of secondary structure conservation

Due to the overall low conservation of the LOTUS domain, we decided to see whether the secondary structure was conserved. To this end, we used the secondary structure prediction algorithm JPred 4 (Drozdetskiy, et al. 2015). Given an amino acid sequence, this tool returns a positional prediction for α-helix, β-sheet or unstructured. We used the JPred4 web servers to compute the predictions and processed them into a secondary structure alignment (see ***2.6_Oskar_lotus_osk_structures.ipynb***). We then used WebLogo (Crooks, et al. 2004) to visualize the conservation of the secondary structure.

Visualization of conservation scores

We used PyMOL (DeLano 2002) to map the computed conservation scores onto the solved structures of LOTUS and OSK (Jeske, et al. 2015; Jeske, et al. 2017). At the time of writing, no

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

952 full-length Oskar protein structure had been reported. With the caveat that all visualization was
953 done on the structure of the *Drosophila melanogaster* protein domains, we created a custom
954 python script that augments PyMOL with automatic display and coloring capacities. This script
955 is available as ***Oskar_pymol_visualization.py***, and contains a manual at the beginning of the
956 file. For the OSK domain, we used the structure PDBID: 5A4A, and for the LOTUS domain,
957 PDBID: 5NT7 (Jeske, et al. 2015; Jeske, et al. 2017). The LOTUS structure we used is in
958 complex with Vasa, and in a dimeric form (Jeske, et al. 2017), allowing for easy interpretation of
959 the different conservation scores. For the OSK structure, we removed the residues 399-401 and
960 604-606 from the PDB file as those amino acids did not align across all sequences and therefore
961 showed highly biased conservation scores.
962
963 *Statistical analysis*
964 All statistical analyses were performed using the scipy stats module (https://www.scipy.org/).
965 Significance thresholds for p-values were set at 0.05. Statistical tests and p-values are reported
966 in the figure legends. All statistical tests can be found in the ipython notebooks mentioned below.
967
968 *Data and code availability*
969 The study generated a series of python 3 script and python 3 ipython notebook files that perform
970 the entire analysis. All the results presented in this paper can be reproduced by running the
971 aforementioned python 3 code. The primary data, oskar homologs, Oskar alignments, trees, and
972 conservation statistics as well as the code created and used are available as supplementary
973 information. For ease of access, legibility, and reproducibility, the code and datasets have been
974 deposited in a GitHub repository available at https://github.com/extavourlab/Oskar_Evolution
975 (commit ID 4eaaa5b11352277e43da72b98bbad397663293fe).
976
977 *Software and libraries*
978 All software and libraries used in this study are published under open source libre licenses and
979 are therefore available to any researcher.

Type	Name	Version	Source
Software	HMMER	3.1.b2	http://hmmer.org/

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

Software	PyMOL	1.8.x	https://pymol.org
Software	rsync	3.1.2	http://rsync.samba.org/
Software	Python 3	3.7	https://www.python.org/
Software	Mrbayes	3.2.6	http://nbisweden.github.io/MrBayes/
Software	trimal	1.2rev59	http://trimal.cgenomics.org/
Software	transeq	6.6.0.0	http://emboss.sourceforge.net/apps/cvs/emboss/apps/transeq.html
Software	augustus	2.5.5	http://augustus.gobics.de/
Software	JPred4	4.0	http://www.compbio.dundee.ac.uk/jpred/
Software	RNABindR	2.0	http://ailab1.ist.psu.edu/RNABindR/
Software	Inkscape	0.92.3	https://inkscape.org/
Library	jupyter	4.4.0	https://jupyter.org/
Library	ete3	3.3.1	http://etetoolkit.org
Library	pandas	0.25.1	https://pandas.pydata.org/
Library	mca	1.0.3	https://pypi.org/project/mca/
Library	fuzzywuzzy	0.17.0	https://github.com/seatgeek/fuzzywuzzy
Library	BeautifulSoup4	4.6.3	https://pypi.org/project/beautifulsoup4/
Library	biopython	1.74	https://pypi.org/project/biopython/
Library	numpy	1.16.2	https://www.numpy.org/
Library	seaborn	0.9.0	https://seaborn.pydata.org/
Library	matplotlib	3.0.0	https://matplotlib.org/
Library	scipy	1.1.0	https://www.scipy.org/
Library	progressbar	3.38.0	https://github.com/niltonvolpato/python-progressbar

Acknowledgements

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

This work was supported by funds from Harvard University, support to SB from the Master's in Bioinformatics Program of the University of Bordeaux, and support to ER from the NSF-Simons center for Mathematical and Statistical Analysis of Biology at Harvard (award number #1764269), the Harvard Quantitative Biology Initiative, and the Herchel Smith Graduate Fellowship. We thank members of the Extavour lab for discussion.

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

Figure Legends

Figure 1: Overview of Oskar protein structure. The most common isoform of the Oskar protein, Short Oskar, is composed of two well-folded domains, LOTUS and OSK, separated by an interdomain sequence. A second isoform of the protein called Long Oskar is present in some Dipteran insects, which contains a 5' domain as well as the three domains of Short Oskar. Below the schematic representation is a rendering of the previously reported solved structures for the LOTUS (PDBID: 5NT7) and OSK (PDBID: 5A4A) domains (Jeske, et al. 2015; Yang, et al. 2015) with a speculative rendering of the unfolded interdomain region shown with a dashed line

Figure 2: Schematic presentation of the *oskar* homolog detection pipeline. Sequences were collected automatically from the three NCBI databases, GenBank (GCA), RefSeq (GCF) and Transcriptome Shotgun Assembly Database (TSA). RefSeq genomes were used to generate Augustus gene model HMMs, which were used to annotate and predict proteins in the non-annotated genomes obtained from GenBank. Transcripts from the TSA database were 6-frame translated using TRANSEQ. Amino acid sequences were consolidated into three protein databases. *hmmsearch* from the HMMER tool suite was used to search for LOTUS and OSK hits in those sequences. Sequences with hits for both the LOTUS and OSK domains with an E-value <0.05 were annotated as *oskar* sequences. Sequences were then cleaned to remove duplicates (sequences with <80% sequence similarity coming from the same organism). The resulting sequences were aligned using *hmmalign*, and the process was repeated until no new sequences were identified. Finally, the sequences were consolidated with the dataset metadata into the *oskar* homolog database that was used for all subsequent analyses.

Figure 3: Summary of *oskar* distribution and expression in insects. Phylogeny from (Misof, et al. 2014). Symbols in order from left to right: (i) vertical rectangles: grey: no *oskar* homolog was identified in this order. Color (unique for each order): at least one *oskar* homolog was identified in this order. (ii) number of datasets searched. (iii) horizontal rectangles: proportion of searched datasets in which an *oskar* homolog was identified. (iv) pie chart: proportion of *oskar* sequences identified in RefSeq (GCF) datasets. (v) pie chart: proportion of *oskar* sequences identified in GenBank (GCA) datasets. (vi) pie chart: proportion of *oskar* sequences identified in Transcriptome Shotgun Assembly Database (TSA) datasets. (vii) *oskar* sequences identified in tissue related to germ line (transcriptomes derived from reproductive organs, eggs or embryos). (viii) *oskar* sequences identified in tissue related to the brain (transcriptomes derived from brain or head). (ix) *oskar* sequences identified in an egg stage transcriptome. (x) *oskar* sequences identified in a larval stage transcriptome. (xi) *oskar* sequences identified in a pupal stage transcriptome. (xii) *oskar* sequences identified in a nymphal or juvenile stage transcriptome. (xiii) *oskar* sequences identified in an adult transcriptome. All numbers represented graphically here are in Supplementary Table 1. No datasets were available for Protura, Diplura or Isoptera at the time of analysis.

Figure 4: Phylogenetic reconstruction of hymenopteran Oskar sequences. Phylogenetic tree inferred using RaxML with 100 bootstraps. Each leaf represents an Oskar homolog. Gray circles: only one Oskar sequence was identified. Red circles: putatively duplicated Oskar sequences identified (sequence similarity <80%). Only families which contained a putative duplication are shown here; see Supplementary Figure S4 for the results of our *oskar* search in the context of a more complete hymenopteran phylogeny.

Figure 5: Differential conservation of amino acids between hemimetabolous and holometabolous Oskar sequences. (a) Box plot showing the conservation bias for each of the LOTUS and OSK domains between hemimetabolous and holometabolous Oskar sequences. Statistical difference was tested using a Mann Whitney U test ($p < 0.05$). **(b)** Ribbon diagram of LOTUS (PDBID: 5NT7) and OSK (PDBID: 5A4A) domain structures, where each amino acid is colored by conservation bias on the color scale shown in

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

(a). **(c, d)** Protein surface representation of the OSK domain (PDBID: 5A4A) from two different angles. Black dashed lines indicate the three amino acids reported previously to be necessary for OSK binding to RNA in *D. melanogaster* (Jeske, et al. 2015; Yang, et al. 2015). **(c)** Amino acids colored by conservation bias on the color scale shown in **(a)**. Cyan: amino acids more highly conserved in hemimetabolous sequences; magenta: amino acids more highly conserved in holometabolous sequences. **(d)** Amino acids colored by electrostatic conservation score. Left: hemimetabolous sequences; right: holometabolous sequences.

Figure 6: Conservation analysis of the LOTUS domain. (a) Ribbon diagram of a LOTUS domain dimer (cyan/magenta) in complex with two Vasa molecules (yellow) (PDBID: 5NT7) from two different angles. Each LOTUS amino acid is colored based on its Valdar conservation score. **(b, c)** Sequence Logo of the α5 and α2/α3 helices respectively generated with WebLogo (Crooks, et al. 2004). Black: hydrophobic residues; blue: charged residues; green: polar residues. **(b', b")** Ribbon diagram of the conserved α5 helix, with key amino acids displayed as sticks and colored by Valdar conservation score. Two potential novel Vasa-LOTUS contacts (H227 and Q235) are highlighted with dashed lines. **(c')** Ribbon diagram of the conserved α2 helix, with key amino acids displayed as sticks and colored by hydrophobicity/hydrophily conservation score. **(c")** Ribbon diagram of the conserved α2 helix, with key amino acids displayed as sticks and colored by Valdar conservation score. **(d)** Surface mesh rendering colored with the RNABindR RNA binding conservation score. **(e, f)** Ribbon diagram of the LOTUS β sheet dimerization interface. Left: conservation of monomeric LOTUS domains; right: dimeric LOTUS domains. **(e)** Amino acids colored by electrostatic conservation score. Dashed lines indicate the key electrostatic interaction thought to stabilize the dimerization. **(f)** Amino acids colored by hydrophobicity/hydrophily conservation score. Dashed lines indicate the key hydrophobic pocket thought to stabilize the dimerization.

Figure 7: Conservation analysis of the OSK domain. (a) Ribbon diagram of the OSK domain (PDBID: 5A4A) from two different angles. Each amino acid is colored based on its Valdar conservation score. **(b)** Protein surface representation of the OSK domain colored by Valdar conservation, electrostatic conservation and hydrophobicity/hydrophily conservation score. **(c, c', c'', c''')** Ribbon diagram of newly detected conserved motifs of the OSK domain, showing sequence Logo (bottom row) residues as sticks. Each amino acid is colored with Valdar conservation scores of holometabolous (top row) and hemimetabolous (middle row) OSK sequences. Bottom row: sequence Logos of each conserved motif generated with WebLogo (Crooks, et al. 2004). Black: hydrophobic residues; blue: charged residues; green: polar residues. Red numbers: amino acid locations of *D. melanogaster* loss of function *oskar* alleles leading to the loss of *oskar* localization to the posterior pole during embryogenesis (P425S = *osk[8]* (Kim-Ha, et al. 1991); S452L = *osk[255]* = *osk[7]* (Lehmann and Nüsslein-Volhard 1986; Kim-Ha, et al. 1991); S457F = *osk[6B10]* (Breitwieser, et al. 1996)) or to reduced RNA-binding affinity of the OSK domain (R436E (Yang, et al. 2015)).

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

References

Ahuja A, Extavour CG. 2014. Patterns of molecular evolution of the germ line specification gene *oskar* suggest that a novel domain may contribute to functional divergence in *Drosophila*. Development Genes and Evolution 222:65-77.

Anantharaman V, Zhang D, Aravind L. (p15752 co-authors). 2010. OST-HTH: a novel predicted RNA-binding domain. Biology direct 5:13.

Ando H, Tanaka M. 1979. Early embryonic devekopment of the primitive moths, *Enduclyta signifer* Walker and *E. excrescens* Butler (Lepidoptera: Hepialidae). International Journal of Insect Morphology and Embryology 9:67-77.

Andreani J, Quignot C, Guerois R. 2020. Structural prediction of protein interactions and docking using conservation and coevolution. Wiley Interdisciplinary Reviews: Computational Molecular Science 10.

Anne J. (p17767 co-authors). 2010. Targeting and Anchoring Tudor in the Pole Plasm of the *Drosophila* Oocyte. PLoS ONE 5:e14362.

Anne J, Mechler BM. 2005. Valois, a component of the nuage and pole plasm, is involved in assembly of these structures, and binds to Tudor and the methyltransferase Capsuleen. Development 132:2167-2177.

Aravind L, Anantharaman V, Balaji S, Babu MM, Iyer LM. 2005. The many faces of the helix-turn-helix domain: transcription regulation and beyond. FEMS Microbiology Reviews 29:231-262.

Babu K, Cai Y, Bahri S, Yang X, Chia W. (p01098 co-authors). 2004. Roles of Bifocal, Homer, and F-actin in anchoring Oskar to the posterior cortex of *Drosophila* oocytes. Genes and Development 18:138-143.

Blondel L, Jones TEM, Extavour CG. 2020. Bacterial contribution to genesis of the novel germ line determinant *oskar*. Elife 9:e45539.

Breitwieser W, Markussen F-H, Horstmann H, Ephrussi A. 1996. Oskar protein interaction with Vasa represents an essential step in polar granule assembly. Genes and Development 10:2179-2188.

Bütschli O. 1870. Zur Entwicklungsgeschichte der Biene. Zeitschrift für Wissenschaftliche Zoologie 20:519-564.

Butt FH. 1949. Embryology of the Milkweed Bug, *Oncopeltus fasciatus* (Hemiptera). Cornell Experiment Station Memoir 283:2-43.

Callebaut I, Mornon J-P. (p19296 co-authors). 2010. LOTUS, a new domain associated with small RNA pathways in the germline. Bioinformatics 26:1140-1144.

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

1111 Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. 2009. trimAl: a tool for automated alignment
1112 trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972-1973.

1113 Capra JA, Singh M. 2007. Predicting functionally important residues from sequence conservation.
1114 Bioinformatics 23:1875-1882.

1115 Carter J-M, Baker SC, Pink R, Carter DRF, Collins A, Tomlin J, Gibbs M, Breuker CJ. 2013.
1116 Unscrambling butterfly oogenesis. BioMedCentral Genomics 14:283-283.

1117 Carter JM, Gibbs M, Breuker CJ. 2015. Divergent RNA Localisation Patterns of Maternal Genes
1118 Regulating Embryonic Patterning in the Butterfly Pararge aegeria. PLoS ONE 10:e0144471.

1119 Chang CC, Lee WC, Cook CE, Lin GW, Chang T. 2006. Germ-plasm specification and germline
1120 development in the parthenogenetic pea aphid *Acyrthosiphon pisum*: Vasa and Nanos as markers.
1121 International Journal of Developmental Biology 50:413-421.

1122 Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, Kaufman TC, Kellis M, Gelbart W,
1123 Iyer VN, et al. (p04187 co-authors). 2007. Evolution of genes and genomes on the *Drosophila*
1124 phylogeny. Nature 450:203-218.

1125 Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence logo generator. Genome
1126 Research 14:1188-1190.

1127 Cui G, Botuyan MV, Mer G. 2013. (1)H, (15)N and (13)C resonance assignments for the three LOTUS
1128 RNA binding domains of Tudor domain-containing protein TDRD7. Biomol NMR Assign 7:79-83.

1129 Dearden PK. 2006. Germ cell development in the Honeybee (*Apis mellifera*); *vasa* and *nanos*
1130 expression. BMC Developmental Biology 6:6.

1131 Dearden PK, Wilson MJ, Sablan L, Osborne PW, Havler M, McNaughton E, Kimura K, Milshina NV,
1132 Hasselmann M, Gempe T, et al. 2006. Patterns of conservation and change in honey bee
1133 developmental genes. Genome Research 16:1376-1384.

1134 DeLano WL. 2002. Pymol: An Open-Source Molecular Graphics Tool. CCP4 Newsletter on Protein
1135 Crystallography 40:82-92.

1136 Drozdetskiy A, Cole C, Procter J, Barton GJ. 2015. JPred4: a protein secondary structure prediction
1137 server. Nucleic Acids Research 43:W389-394.

1138 Eastham LES. 1930. The embryology of *Pieris rapae* - Organogeny. Philosophical Transactions of the
1139 Royal Society of London. Series B: Biological Sciences 219:1-50.

1140 Edgar RC. (r20416 co-authors). 2004. MUSCLE: a multiple sequence alignment method with reduced
1141 time and space complexity. BMC Bioinformatics 5:113.

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

1142 Ephrussi A, Dickinson LK, Lehmann R. (p13297 co-authors). 1991. Oskar organizes the germ plasm
1143 and directs localization of the posterior determinant *nanos*. Cell 66:37-50.

1144 Ephrussi A, Lehmann R. 1992. Induction of germ cell formation by *oskar*. Nature 358:387-392.

1145 Ewen-Campen B, Schwager EE, Extavour CG. 2010. The molecular machinery of germ line
1146 specification. Molecular Reproduction and Development 77:3-18.

1147 Ewen-Campen B, Srouji JR, Schwager EE, Extavour CG. 2012. *oskar* Predates the Evolution of Germ
1148 Plasm in Insects. Current Biology 22:2278-2283.

1149 Extavour CG, Akam ME. 2003. Mechanisms of germ cell specification across the metazoans:
1150 epigenesis and preformation. Development 130:5869-5884.

1151 Fleig R, Sander K. 1985. Blastoderm development in honey bee embryogenesis as seen in the
1152 scanning electron microscope. International Journal of Invertebrate Reproduction and Development
1153 8:279-286.

1154 Fleig R, Sander K. 1986. Embryogenesis of the Honeybee *Apis mellifera* L (Hymenoptera, Apidae) - an
1155 SEM Study. International Journal of Insect Morphology and Embryology 15:449-462.

1156 Gajiwala KS, Burley SK. 2000. Winged helix proteins. Current Opinion in Structural Biology 10:110-116.

1157 Goltsev Y, Hsiong W, Lanzaro G, Levine M. 2004. Different combinations of gap repressors for
1158 common stripes in *Anopheles* and *Drosophila* embryos. Developmental Biology 275:435-446.

1159 Guelin M. 1994. [Activity of W-sex heterochromatin and accumulation of the nuage in nurse cells of the
1160 lepidopteran *Ephestia*]. C. R. Acad. Sci. Paris. Ser. III 317:54-61.

1161 Gutzeit HO, Zissler D, Fleig R. 1993. Oogenesis in the Honeybee *Apis mellifera* - Cytological
1162 Observations on the Formation and Differentiation of Previtellogenic Ovarian Follicles. Rouxs Archives
1163 of Developmental Biology 202:181-191.

1164 Hamming RW. 1950. Error Detecting and Error Correcting Codes. The Bell System Technical Journal
1165 29:147-160.

1166 Harami GM, Gyimesi M, Kovacs M. 2013. From keys to bulldozers: expanding roles for winged helix
1167 domains in nucleic-acid-binding proteins. Trends in Biochemical Sciences 38:364-371.

1168 Hay B, Jan LY, Jan YN. 1990. Localization of *vasa*, a component of *Drosophila* polar granules, in
1169 maternal-effect mutants that alter embryonic anteroposterior polarity. Development 109:425-433.

1170 Heming BS, Huebner E. 1994. Development of the Germ Cells and Reproductive Primordia in Male and
1171 Female Embryos of *Rhodnius prolixus* Stal (Hemiptera, Reduviidae). Canadian Journal of Zoology
1172 72:1100-1119.

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

1173 Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: Reconstruction, Analysis, and Visualization of
1174 Phylogenomic Data. Molecular Biology and Evolution 33:1635-1638.

1175 Hurd TR, Herrmann B, Sauerwald J, Sanny J, Grosch M, Lehmann R. 2016. Long Oskar Controls
1176 Mitochondrial Inheritance in *Drosophila melanogaster*. Developmental Cell 39:560-571.

1177 Jaglarz MK, Nowak Z, Bilinski SM. 2003. The Balbiani body and generation of early asymmetry in the
1178 oocyte of a tiger beetle. Differentiation 71:142-151.

1179 Jeske M, Bordi M, Glatt S, Muller S, Rybin V, Muller CW, Ephrussi A. 2015. The Crystal Structure of the
1180 *Drosophila* Germline Inducer Oskar Identifies Two Domains with Distinct Vasa Helicase- and RNA-
1181 Binding Activities. Cell Reports 12:587-598.

1182 Jeske M, Muller CW, Ephrussi A. 2017. The LOTUS domain is a conserved DEAD-box RNA helicase
1183 regulator essential for the recruitment of Vasa to the germ plasm and nuage. Genes and Development
1184 31:939-952.

1185 Johannsen OA. 1929. Some phases in the embryonic development of *Diacrisia virginica* Fabr.
1186 (Lepidoptera). J. Morphol. Physiol. 2:493-541.

1187 Jones JR, Macdonald PM. 2007. Oskar controls morphology of polar granules and nuclear bodies in
1188 *Drosophila*. Development 134:233-236.

1189 Jongens TA, Hay B, Jan LY, Jan YN. 1992. The *germ cell-less* Gene Product: A Posteriorly Localized
1190 Component Necessary for Germ Cell Development in *Drosophila*. Cell 70:569-584.

1191 Juhn J, James AA. 2006. *oskar* gene expression in the vector mosquitoes, *Anopheles gambiae* and
1192 *Aedes aegypti*. Insect Molecular Biology 15:363-372.

1193 Juhn J, Marinotti O, Calvo E, James AA. 2008. Gene structure and expression of *nanos* (*nos*) and
1194 *oskar* (*osk*) orthologues of the vector mosquito, *Culex quinquefasciatus*. Insect Molecular Biology
1195 17:545-552.

1196 Kawahara AY, Plotkin D, Espeland M, Meusemann K, Toussaint EFA, Donath A, Gimnich F, Frandsen
1197 PB, Zwick A, Dos Reis M, et al. 2019. Phylogenomics reveals the evolutionary timing and pattern of
1198 butterflies and moths. Proceedings of the National Academy of Sciences of the United States of
1199 America 116:22657-22663.

1200 Kelly GM, Huebner E. 1989. Embryonic development of the hemipteran insect *Rhodnius prolixus*.
1201 Journal of Morphology 199:175-196.

1202 Khila A, Abouheif E. 2008. Reproductive constraint is a developmental mechanism that maintains social
1203 harmony in advanced ant societies. Proceedings of the National Academy of Sciences of the United
1204 States of America 105:17884-17889.

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

1205 Kim-Ha J, Smith JL, Macdonald PM. 1991. *oskar* mRNA is localized to the posterior pole of the
1206 *Drosophila* oocyte. Cell 66:23-35.

1207 Kirk DL. 2005. A twelve-step program for evolving multicellularity and a division of labor. Bioessays
1208 27:299-310.

1209 Klag J, Bilinski S. 1993. Oosome Formation in 2 Ichneumonid Wasps. Tissue & Cell 25:121-128.

1210 Kobayashi S, Amikura R, Nakamura A, Saito H, Okada M. 1995. Mislocalization of *oskar* Product in the
1211 Anterior Pole Results in Ectopic Localization of Mitochondrial Large Ribosomal RNA in *Drosophila*
1212 Embryos. Developmental Biology 169:384-386.

1213 Kobayashi Y, Ando H. 1984. Mesodermal Organogenesis in the Embryo of the Primitive Moth,
1214 *Neomicropteryx nipponensis* Issiki (Lepidoptera, Micropterygidae). Journal of Morphology 181:29-47.

1215 Lachke SA, Alkuraya FS, Kneeland SC, Ohn T, Aboukhalil A, Howell GR, Saadi I, Cavallesco R, Yue Y,
1216 Tsai AC, et al. 2011. Mutations in the RNA granule component TDRD7 cause cataract and glaucoma.
1217 Science 331:1571-1576.

1218 Lasko P. 2013. The DEAD-box helicase Vasa: evidence for a multiplicity of functions in RNA processes
1219 and developmental biology. Biochimica et Biophysica Acta 1829:810-816.

1220 Lautenschlager F. 1932. Die Embryonalentwicklung der weiblichen Keimdruse bei der Psychide
1221 *Solenobia triquetella*. Zool. Jarh. 56:121-162.

1222 Lebart L, Morineau A, Warwick KM. 1984. Multivariate Descriptive Statistical Analysis: Correspondence
1223 Analysis and Related Techniques for Large Matrices. Chichester, UK: John Wiley & Sons.

1224 Lehmann R. 2016. Germ Plasm Biogenesis--An Oskar-Centric Perspective. Current Topics in
1225 Developmental Biology 116:679-707.

1226 Lehmann R, Nüsslein-Volhard C. 1986. Abdominal Segmentation, Pole Cell Formation, and Embryonic
1227 Polarity Require the Localized Activity of *oskar*, a Maternal Gene in *Drosophila*. Cell 47:144-152.

1228 Lin GW, Cook CE, Miura T, Chang CC. 2014. Posterior localization of ApVas1 positions the preformed
1229 germ plasm in the sexual oviparous pea aphid Acyrthosiphon pisum. EvoDevo 5:18.

1230 Lin J. 1991. Divergence Measures Based on the Shannon Entropy. IEEE Transactions on Information
1231 Theory / Professional Technical Group on Information Theory 37:145-151.

1232 Liu Y, Manna A, Li R, Martin WE, Murphy RC, Cheung AL, Zhang G. 2001. Crystal structure of the
1233 SarR protein from Staphylococcus aureus. Proc Natl Acad Sci U S A 98:6877-6882.

1234 Loytynoja A. 2014. Phylogeny-aware alignment with PRANK. Methods in Molecular Biology 1079:155-
1235 170.

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

1236 Lynch JA, Desplan C. (p16123 co-authors). 2010. Novel modes of localization and function of *nanos* in
1237 the wasp *Nasonia*. Development 137:3813-3821.

1238 Lynch JA, Özüak O, Khila A, Abouheif E, Desplan C, Roth S. 2011. The Phylogenetic Origin of *oskar*
1239 Coincided with the Origin of Maternally Provisioned Germ Plasm and Pole Cells at the Base of the
1240 Holometabola. PLoS Genetics 7:e1002029.

1241 Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, Basutkar P, Tivey ARN, Potter SC,
1242 Finn RD, et al. 2019. The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids
1243 Research 47:W636-W641.

1244 Mahowald AP. 2001. Assembly of the *Drosophila* germ plasm. Int Rev Cytol 203:187-213.

1245 Mahowald AP. 1962. Fine Structure of Pole Cells and Polar Granules in *Drosophila melanogaster*.
1246 Journal of Experimental Zoology:201-215.

1247 Mahowald AP. (p13340 co-authors). 1968. Polar granules of *Drosophila*. II. Ultrastructural changes
1248 during early embryogenesis. Journal of Experimental Zoology 167:237-261.

1249 Mahowald AP, Illmensee K, Turner FR. (p09066 co-authors). 1976. Interspecific transplantation of polar
1250 plasm between *Drosophila* embryos. Journal of Cell Biology 70:358-373.

1251 Markussen FH, Michon AM, Breitwieser W, Ephrussi A. 1995. Translational control of *oskar* generates
1252 short OSK, the isoform that induces pole plasm assembly. Development 121:3723-3732.

1253 Matthews BJ, McBride CS, DeGennaro M, Despo O, Vosshall LB. 2016. The neurotranscriptome of the
1254 Aedes aegypti mosquito. BioMedCentral Genomics 17:32.

1255 Megosh HB, Cox DN, Campbell C, Lin H. 2006. The role of PIWI and the miRNA machinery in
1256 *Drosophila* germline determination. Current Biology 16:1884-1894.

1257 Mellanby H. 1935. The early embryonic development of *Rhodnius prolixus* (Hemiptera, Heteroptera).
1258 Quarterly Journal of Microscopical Science 78:71-90.

1259 Meng C. 1968. Strukturwandel und histochimie Befunde iunbesondere am Oosom wahrend der
1260 Oogenese und nach der Ablage des Eies von *Pimpla turionellae* L. (Hymenoptera, Ichenumonidae).
1261 Wilhelm Roux' Archiv. 161:162-208.

1262 Metschnikoff E. 1866. Embryologische Studien an Insekten. Zeit. f. wiss Zool. 16:389-500.

1263 Misof B, Liu S, Meusemann K, Peters RS, Donath A, Mayer C, Frandsen PB, Ware J, Flouri T, Beutel
1264 RG, et al. 2014. Phylogenomics resolves the timing and pattern of insect evolution. Science 346:763-
1265 767.

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

1266 Mitter C, Davis DR, Cummings MP. 2017. Phylogeny and Evolution of Lepidoptera. Annual Review of
1267 Entomology 62:265-283.

1268 Miura T, Braendle C, Shingleton A, Sisk G, Kambhampati S, Stern DL. 2003. A comparison of
1269 parthenogenetic and sexual embryogenesis of the pea aphid *Acyrthosiphon pisum* (Hemiptera :
1270 Aphidoidea). Journal of Experimental Zoology Part B-Molecular and Developmental Evolution 295B:59-
1271 81.

1272 Miya K. 1953. The presumptive genital region at the blastoderm stage of the silkworm egg. Journal of
1273 the Faculty of Agriculture of Iwate University:223-227.

1274 Miya K. 1958. Studies on the embryonic development of the gonad in the silkworm, *Bombyx mori* L.
1275 Part I. Differentiation of germ cells. Journal of the Faculty of Agriculture of Iwate University 3:436-467.

1276 Miya K. 1975. Ultrastructural changes of embryonic cells during organogenesis in the silkworm,
1277 *Bombyx mori*. I. The Gonad. Journal of the Faculty of Agriculture of Iwate University 12:329-338.

1278 Moon CP, Fleming KG. 2011. Side-chain hydrophobicity scale derived from transmembrane protein
1279 folding into lipid bilayers. Proceedings of the National Academy of Sciences of the United States of
1280 America 108:10174-10177.

1281 Mukherjee D, Datta AB, Chakrabarti P. 2014. Crystal structure of HlyU, the hemolysin gene
1282 transcription activator, from Vibrio cholerae N16961 and functional implications. Biochimica et
1283 Biophysica Acta 1844:2346-2354.

1284 Nagy L, Riddiford L, Kiguchi K. 1994. Morphogenesis in the Early Embryo of the Lepidopteran *Bobyx*
1285 *mori*. Developmental Biology:137-151.

1286 Nakamura A, Amikura R, Mukai M, Kobayashi S, Lasko PF. 1996. Requirement for a noncoding RNA in
1287 *Drosophila* polar granules for germ cell establishment. Science 274:2075-2079.

1288 Nakao H. 1999. Isolation and characterization of a *Bombyx vasa*-like gene. Development Genes and
1289 Evolution 209:312-316.

1290 Nakao H, Hatakeyama M, Lee JM, Shimoda M, Kanda T. 2006. Expression pattern of *Bombyx* vasa-
1291 like (BmVLG) protein and its implications in germ cell development. Development Genes and Evolution
1292 216:94-99.

1293 Nakao H, Matsumoto T, Oba Y, Niimi T, Yaginuma T. 2008. Germ cell specification and early
1294 embryonic patterning in Bombyx mori as revealed by nanos orthologues. Evolution and Development
1295 10:546-554.

1296 Nakao H, Takasu Y. 2019. Complexities in Bombyx germ cell formation process revealed by Bm-nosO
1297 (a Bombyx homolog of nanos) knockout. Developmental Biology 445:29-36.

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

1298 Nardon P. 1971. Contribution a l'étude des symbiotes ovariens de Sitophilus sasakii: localisation,
1299 histochimie et ultrastructure chez la femelle adulte. Comptes Rendus de l Academie des Sciences.
1300 Serie III, Sciences de la Vie 272D:2975-2978.

1301 Nardon P. 2006. Ovogenese et transmission des bactéries symbiotiques chez le charancon Sitophilus
1302 oryzae L. (Coleoptera: Curculionoidea). Annales de la Société Entomologique de France 42:129-164.

1303 Nelson JA. 1915. The embryology of the honey bee. Princeton: Princeton University Press.

1304 Noce T, Okamoto-Ito S, Tsunekawa N. 2001. *Vasa* homolog genes in mammalian germ cell
1305 development. Cell Structure and Function 26:131-136.

1306 Pal D, Chakrabarti P. 2001. Non-hydrogen bond interactions involving the methionine sulfur atom.
1307 Journal of Biomolecular Structure and Dynamics 19:115-128.

1308 Peters RS, Krogmann L, Mayer C, Donath A, Gunkel S, Meusemann K, Kozlov A, Podsiadlowski L,
1309 Petersen M, Lanfear R, et al. 2017. Evolutionary History of the Hymenoptera. Current Biology 27:1013-
1310 1018.

1311 Presser BD, Rutschky CW. 1957. The embryonic development of the corn earworm, *Heliothis zea*
1312 (Boddie) (Lepidoptea, Phalaenidae). Annals of the Entomological Society of America 50:133-164.

1313 Quan H, Arsala D, Lynch JA. 2019. Transcriptomic and functional analysis of the oosome, a unique
1314 form of germ plasm in the wasp *Nasonia vitripennis*. BMC Biology 17:78.

1315 Quan H, Lynch JA. 2016. The evolution of insect germline specification strategies. Current Opinion in
1316 Insect Science 13:99-105.

1317 Rafiqi AM, Rajakumar A, Abouheif E. 2020. Origin and elaboration of a major evolutionary transition in
1318 individuality. Nature 585:239-244.

1319 Rausell A, Juan D, Pazos F, Valencia A. 2010. Protein interactions and ligand binding: from protein
1320 subfamilies to functional specificity. Proceedings of the National Academy of Sciences of the United
1321 States of America 107:1995-2000.

1322 Raz E. 2000. The function and regulation of *vasa*-like genes in germ-cell development. Genome
1323 Biology 1:1-6.

1324 Rongo C, Broihier HT, Moore L, Van Doren M, Forbes A, Lehmann R. 1997. Germ Plasm Assembly
1325 and Germ Cell Migration in *Drosophila*. Cold Spring Harbour Symposia on Quantitative Biology LXII:1-
1326 11.

1327 Saito. 1937. On the development of the Tusser, *Antheraea pernyi* Guerin-Meneville, with special
1328 reference to the comparative embryology of insects. Journal of the Faculty of Agriculture of Hokkaido
1329 Imperial University 40:35-109.

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

1330 Schroder R. 2006. *vasa* mRNA accumulates at the posterior pole during blastoderm formation in the
1331 flour beetle *Tribolium castaneum*. Development, Genes and Evolution 216:277-283.

1332 Schwangart F. 1905. Zur Entwickslungsgeschichte der Lepidopteren. Biol. Centralbl. 25:777-789.

1333 Sehl A. 1931. Furchung und Bildung der Keimanlage bei der Mehlmotte *Ephestia kuehniella*. Zell. Zeit.
1334 Morph. U. Okol. 1:429-506.

1335 Sikosek T, Chan HS. 2014. Biophysics of protein evolution and evolutionary protein biophysics. Journal
1336 of the Royal Society, Interface / the Royal Society 11:20140419.

1337 Sikosek T, Chan HS, Bornberg-Bauer E. 2012. Escape from Adaptive Conflict follows from weak
1338 functional trade-offs and mutational robustness. Proceedings of the National Academy of Sciences of
1339 the United States of America 109:14888-14893.

1340 Smith JL, Wilson JE, Macdonald PM. 1992. Overexpression of *oskar* directs ectopic activation of *nanos*
1341 and presumptive pole cell formation in *Drosophila* embryos. Cell 70:849-859.

1342 Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large
1343 phylogenies. Bioinformatics 30:1312-1313.

1344 Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. 2006. AUGUSTUS: ab initio
1345 prediction of alternative transcripts. Nucleic Acids Research 34:W435-439.

1346 Suyama R, Jenny A, Curado S, Pellis-van Berkel W, Ephrussi A. (p07861 co-authors). 2009. The actin-
1347 binding protein Lasp promotes Oskar accumulation at the posterior pole of the *Drosophila* embryo.
1348 Development 136:95-105.

1349 Tanaka M. 1987. Differentiation and behaviour of Primordial Germ Cells during the Early Embryonic
1350 Development of *Parnassius glacialis* Butler, *Luehdorfia japonica* Leech and *Byasa (Atrophaneura)*
1351 *alcinous alcinous* Klug (Lepidoptera: Papilionidae). In: Ando H, Jura C, editors. Recent Advances in
1352 Insect Embryology in Japan and Poland. Tsukuba: Arthropod. Embryol. Soc. Jpn.
1353 ISEBU Co. Ltd. p. 255-266.

1354 Tanaka T, Kato Y, Matsuda K, Hanyu-Nakamura K, Nakamura A. 2011. *Drosophila* Mon2 couples
1355 Oskar-induced endocytosis with actin remodeling for cortical anchorage of the germ plasm.
1356 Development 138:2523-2532.

1357 Tanaka T, Nakamura A. (p07886 co-authors). 2008. The endocytic pathway acts downstream of Oskar
1358 in *Drosophila* germ plasm assembly. Development 135:1107-1117.

1359 Terribilini M, Sander JD, Lee JH, Zaback P, Jernigan RL, Honavar V, Dobbs D. 2007. RNABindR: a
1360 server for analyzing and predicting RNA-binding sites in proteins. Nucleic Acids Research 35:W578-
1361 584.

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

1362 Tomaya K. 1902. On the embryology of the silkworm. Bulletin of the College of Agriculture, Tokyo 5:73-
1363 111.

1364 Toshiki T, Chantal C, R., Toshio K, Eappen A, Mari K, Natuo K, Jean-Luc T, Bernard M, Gérard C, Paul
1365 S, et al. 2000. Germline transformation of the silkworm *Bombyx mori* L. using a piggyBac transposon-
1366 derived vector. Nature Biotech.:81-84.

1367 Valdar WS. 2002. Scoring residue conservation. Proteins 48:227-241.

1368 Vanzo N, Oprins A, Xanthakis D, Ephrussi A, Rabouille C. 2007. Stimulation of endocytosis and actin
1369 dynamics by Oskar polarizes the *Drosophila* oocyte. Developmental Cell 12:543-555.

1370 Vanzo NF, Ephrussi A. 2002. Oskar anchoring restricts pole plasm formation to the posterior of the
1371 *Drosophila* oocyte. Development 129:3705-3714.

1372 Wang C, Lehmann R. 1991. Nanos is the localized posterior determinant in *Drosophila*. Cell 66:637-
1373 647.

1374 Webster PJ, Suen J, Macdonald PM. 1994. *Drosophila virilis oskar* transgenes direct body patterning
1375 but not pole cell formation or maintenance of mRNA localization in *D. melanogaster*. Development
1376 120:2027-2037.

1377 Whittle CA, Kulkarni A, Chung N, Extavour CG. 2021. Adaptation of codon and amino acid use for
1378 translational functions in highly expressed cricket genes. BioMedCentral Genomics 22.

1379 Whittle CA, Kulkarni A, Extavour CG. 2021. Evolutionary dynamics of sex-biased genes expressed in
1380 cricket brains and gonads. Journal of Evolutioary Biology doi:10.1111/jeb.13889.

1381 Will L. 1888. Entwicklungsgescshichte der viviparen Aphiden. Zool. Jarh. 3:201-280.

1382 Williams SG, Attridge SR, Manning PA. 1993. The transcriptional activator HlyU of Vibrio cholerae:
1383 nucleotide sequence and role in virulence gene expression. Molecular Microbiology 9:751-760.

1384 Witlaczil E. 1884. Entwicklungsgeschichte der Aphiden. Zeit. f. wiss Zool. 40:559-690.

1385 Woodworth CW. 1889. Studies on the embryological development of *Euvanessa antiopa*. In: Scudder,
1386 editor. Butterflies of Eastern United States and Canada. p. 102.

1387 Xu X, Brechbiel JL, Gavis ER. 2013. Dynein-Dependent Transport of *nanos* RNA in *Drosophila* Sensory
1388 Neurons Requires Rumpelstiltskin and the Germ Plasm Organizer Oskar. Journal of Neuroscience
1389 33:14791-14800.

1390 Yang N, Yu Z, Hu M, Wang M, Lehmann R, Xu RM. 2015. Structure of *Drosophila* Oskar reveals a
1391 novel RNA binding protein. Proceedings of the National Academy of Sciences of the United States of
1392 America 112:11541-11546.

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

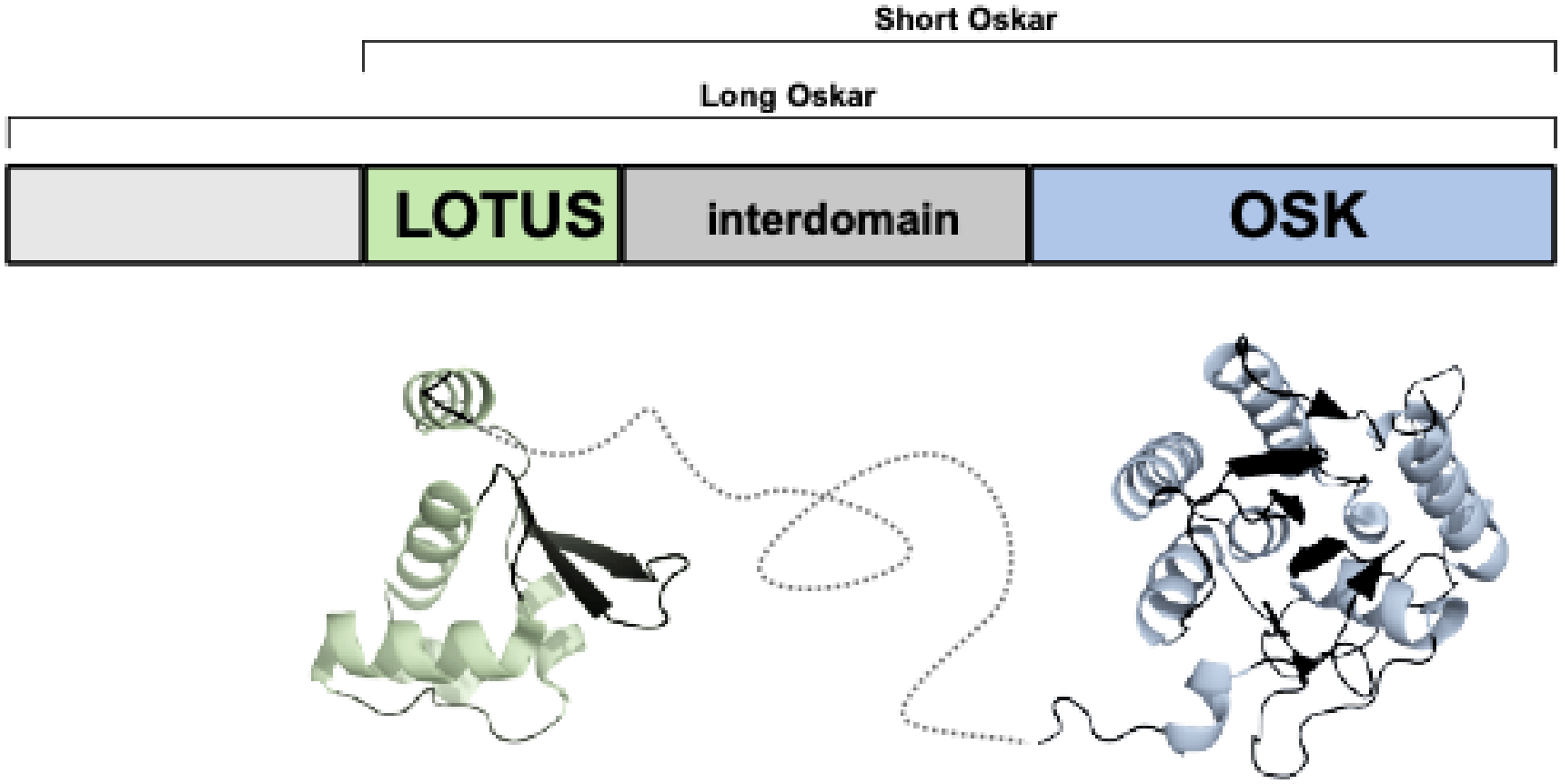
Yoon Y, Klomp J, Martin-Martin I, Criscione F, Calvo E, Ribeiro J, Schmidt-Ott U. 2019. Embryo polarity in moth flies and mosquitoes relies on distinct old genes with localized transcript isoforms. Elife 8.

Zhurov V, Terzin T, Grbic M. 2004. Early blastomere determines embryo proliferation and caste fate in a polyembryonic wasp. Nature 432:764-769.

Zissler D. 1992. From egg to pole cells: ultrastructural aspects of early cleavage and germ cell determination in insects. Micr. Res. and Tech.:49-74.

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

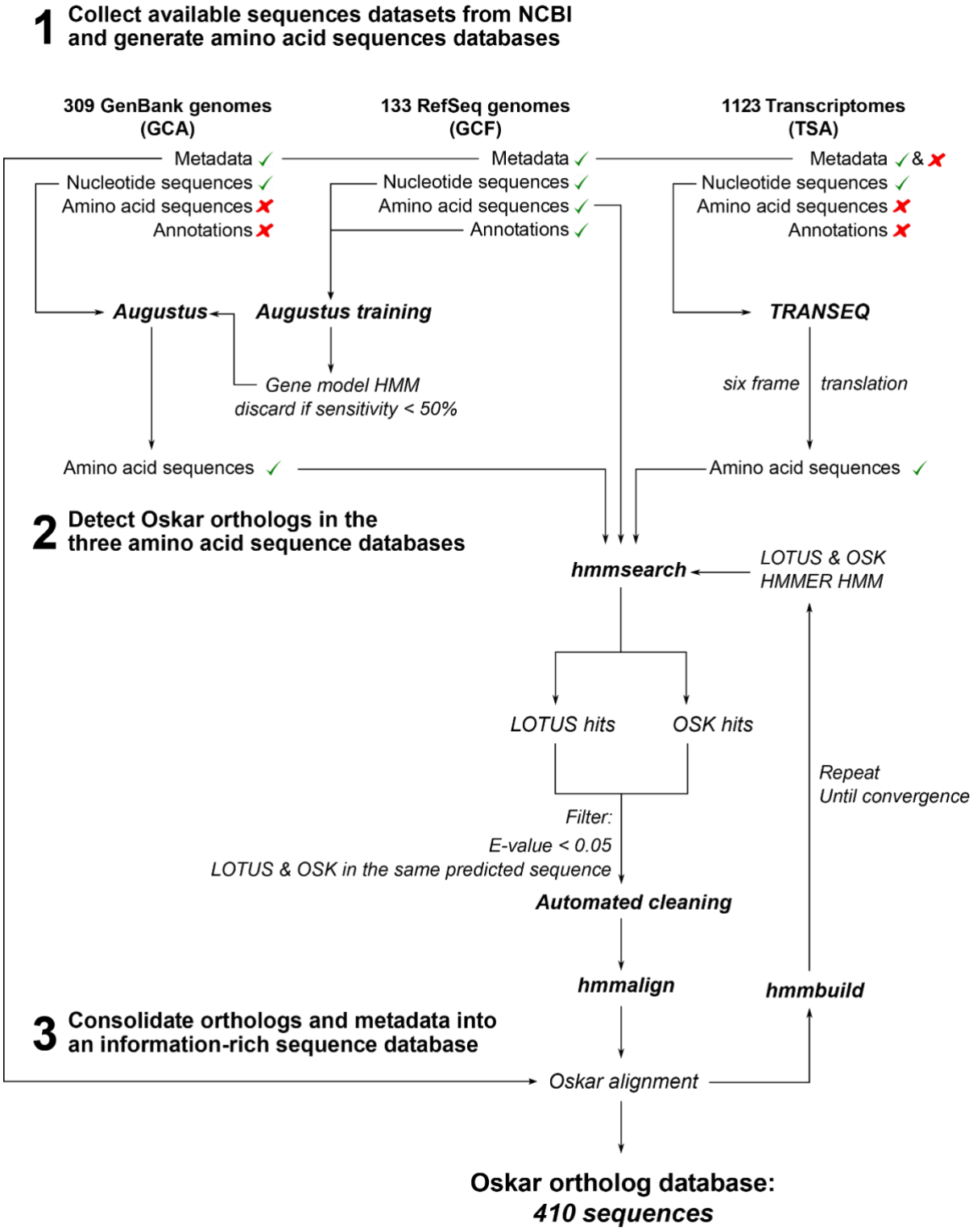
Figure 1



.

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

1404 **Figure 2**
1405



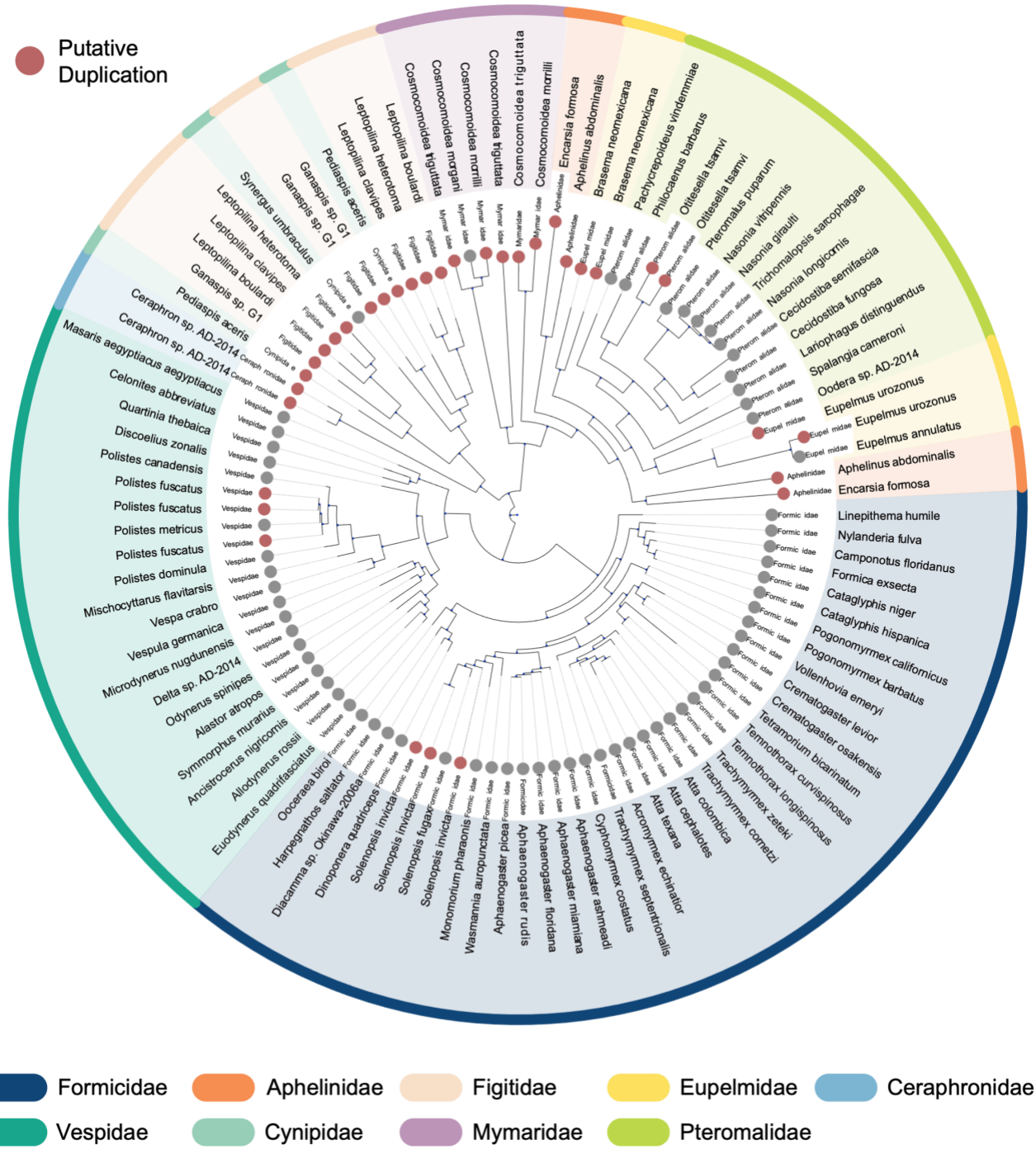
1406

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

Figure 3

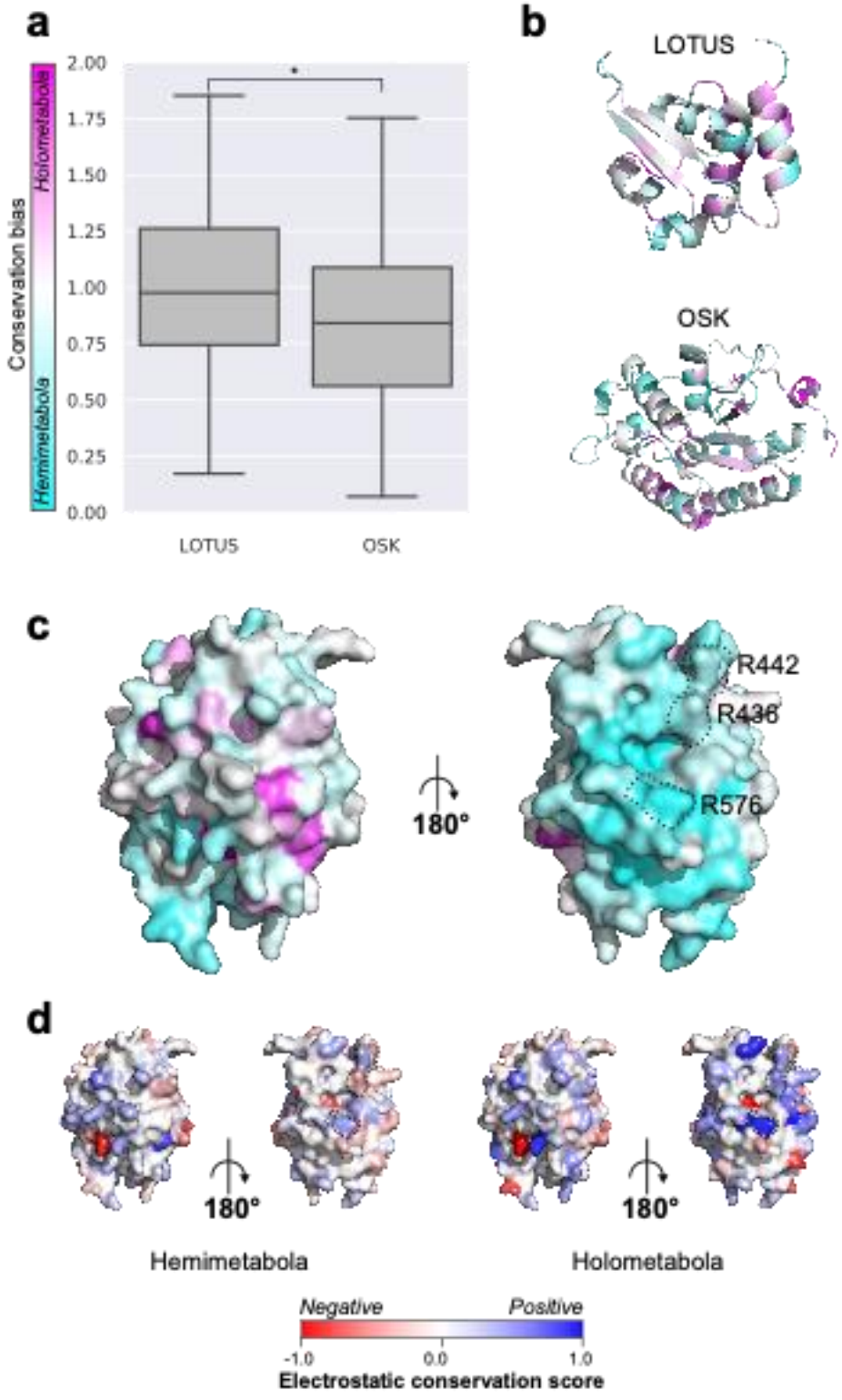
Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

Figure 4



Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

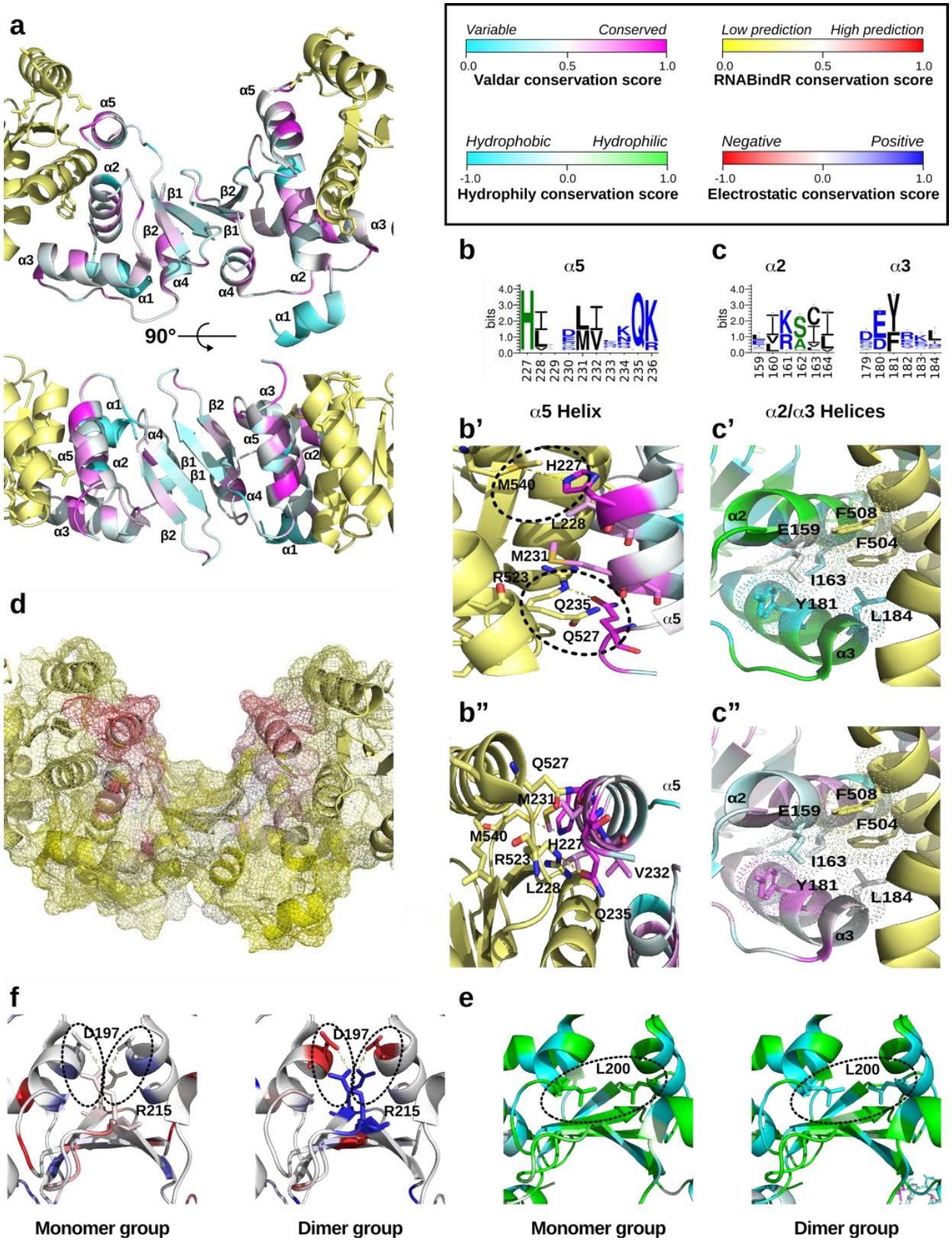
Figure 5



Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

1416 **Figure 6**

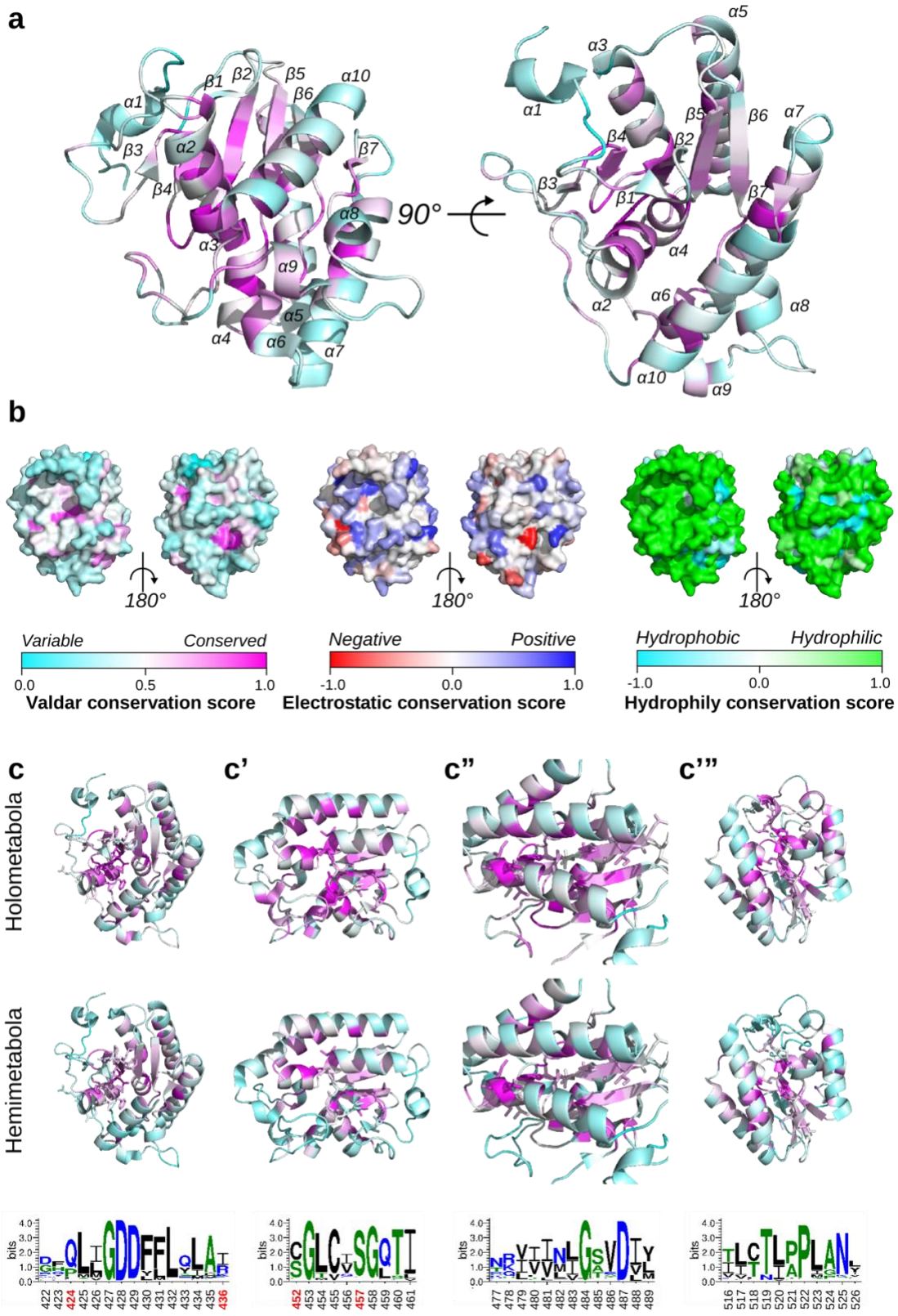
1417



1418

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

Figure 7



Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab284/6373905 by guest on 27 September 2021

Supplementary Materials

Evolution of a cytoplasmic determinant: evidence for the biochemical basis of functional evolution of the novel germ line regulator Oskar

Leo Blondel, Savandara Besse, Emily Rivard, Guillem Ylla, and Cassandra G. Extavour

These Supplementary Materials contain the following:

- Supplementary Text
- Supplementary Methods
- Supplementary References
- Legends for Supplementary Figures S1 through S12 (this document)
- Supplementary Figures S1 through S12 (this document)
- Legends for Supplementary Tables S1 through S5 (this document)
- Supplementary Tables S1 and S2 (this document)
- Supplementary Table S3 through S5 are provided in the GitHub repository for this study at https://github.com/extavourlab/Oskar_Evolution

Supplementary Text

oskar *expression levels in tissue-specific transcriptomes from a mosquito*

We examined all the TSA transcriptomes included in our original analysis, to determine which of them had a bioproject containing SRA sequences for more than one tissue type processed in the same manner, thus useful to derive relative *oskar* expression levels in distinct tissue types. Only one species, the mosquito *Aedes aegypti* (Diptera), satisfied this criterion. We examined the transcript quantification results provided by the authors of the original bioproject (Matthews, et al. 2016; their Supplementary Table 5). We also generated our own quantification analysis using the tool kallisto (Bray, et al. 2016)), with the TSA transcriptome (GFNA01) as an index, and all the SRA reads from the reported bioproject. We found that both quantification analyses yielded similar results (Supplementary Figure S5). Specifically, *oskar* transcripts were detectable in the brain, the ovaries, and what Matthews and colleagues (2016) describe as the "abdominal tip" of female mosquitoes (Supplementary Figure S5A, C). The latter tissue type is defined as the three posterior abdominal segments, which include the external and internal genitalia and ovipositor, but not the ovaries (2016). Expression in the abdominal tip was detected only in blood-fed female mosquitoes 96 hours after feeding, but not in non-blood-fed females (Supplementary Figure S5B, D). Given that female mosquito eggs mature after being blood-fed (Laurence 1977), we hypothesize that the abdominal tip *oskar* expression comes from matured eggs that have moved from the ovary into the bursa within the abdominal tip. In addition, in this dataset, expression levels of *oskar* are higher in the ovaries than in the brain (Supplementary Figure S5A, C). This could reflect a difference in transcriptional activity between the two tissue types, or it could reflect the possibility that the number of cells expressing *oskar* is lower in the brain than in the ovaries. *In situ* hybridization or a similar spatial expression approach will be needed to better understand the specific nature and number of cells that express *oskar* in these two mosquito organ systems.

oskar *expression levels in organ system-specific transcriptomes from a cricket*

We assessed the levels of *oskar* transcript in a methodologically comparable dataset of transcriptomes of 13 reproductive and nervous system tissues from males and females of the cricket *Gryllus bimaculatus* (Orthoptera) recently generated in our laboratory (Whittle, Kulkarni, Chung, et al. 2021; Whittle, Kulkarni, et al. 2021a). We detected the highest *oskar* expression levels in the female ovaries (mean value 114.10 transcripts per million (TPM)), followed by male testes and mixed-sex embryos (mean values 20.2 and 20.18 TPM respectively) (Supplementary Figure S6). Very low levels of *oskar* (mean 1.57 to 5.60 TPM) were detected in all other analyzed tissues (Supplementary Figure S6). The Table below shows the transcripts per million (TPMs) of *oskar* detected in each RNA-seq library of *G. bimaculatus* tissues obtained by Whittle, Kulkarni, et al. (2021b). For each tissue, two or three biological replicates were sequenced.

Sample	Replicate #	TPM
embryos	1	13.10
embryos	2	27.26
female_accessory gland	1	2.00
female_accessory gland	2	5.49
female_brain	1	3.74
female_brain	2	3.89
female_carcass	1	5.82

female_carcass	2	7.27
female_carcass	3	0.00
female_ovary	1	132.49
female_ovary	2	92.23
female_ovary	3	117.59
female_somatic_gonad	1	5.41
female_somatic_gonad	2	4.23
female_somatic_gonad	3	7.16
female_ventral_cord	1	3.16
female_ventral_cord	2	0.68
female_ventral_cord	3	5.83
male_accessory_gland	1	1.49
male_accessory_gland	2	1.73
male_brain	1	3.82
male_brain	2	1.63
male_carcass	1	3.74
male_carcass	2	0.00
male_somatic_gonad	1	2.55
male_somatic_gonad	2	7.81
male_testes	1	26.31
male_testes	2	14.10
male_ventral_cord	1	1.15
male_ventral_cord	2	1.98

Semi-quantitative RT-PCR assessment of* oskar *expression levels in a fly, a weevil, and a stick insect

We used semi-quantitative RT-PCR to assess relative tissue-level *oskar* expression levels in male and female gonads and heads in three insect species: the fruit fly *Drosophila melanogaster* (Diptera; wild type strain Oregon R), the weevil *Callosobruchus maculatus* (Coleoptera), and the stick insect *Aretaon asperrimus* (Phasmatodea). In *D. melanogaster*, strong *oskar* expression was detected in female ovaries but none was detected in female heads (Supplementary Figure S7A). Barely detectable expression was also evident in male gonads and heads (Supplementary Figure S7A). This observation is consistent with transcriptome based-reports from FlyAtlas 2 (Leader, et al. 2017), but to our knowledge, no roles for oskar have been reported in these tissues in this fly.

In *C. maculatus*, expression of oskar was detected in male and female gonads and heads, as well as in embryos (Supplementary Figure 7B). Expression levels were highest in ovaries, followed by heads of both sexes, and lowest but still easily detectable in male gonads (Supplementary Figure S7B). The function of *oskar* in this weevil remains unknown, but these expression data suggest that it could function in one or both of the brain or gonads in both sexes.

In *A. asperrimus*, where we examined only female specimens, expression levels were higher in ovaries than in heads (Supplementary Figure S7C). As for the weevil, functional roles of *oskar* remain untested in this stick insect, but these expression data suggest that it may function in the female nervous and/or reproductive systems.

Supplementary Methods

Analysis of Aedes aegypti oskar transcript levels

To quantify the expression level of *oskar* in published transcriptomes of *A. aegypti* (Matthews, et al. 2016), we used the published tool kallisto (Bray, et al. 2016). We first built an index using the *index* command on the complete transcriptome (TSA ID GFNA01). We then downloaded each set of reads from the NCBI SRA repository for the bioproject (PRJNA236239) and applied the quantification *quant* command to each individual dataset. Finally, we aggregated the results with the bioproject metadata to generate the final results. We also compared our results to the published quantification of transcripts done by the authors of the original study (Matthews, et al. 2016).

Analysis of *Gryllus bimaculatus* oskar transcript levels

The RNA-seq libraries of different tissues of *Gryllus bimaculatus* tissues were recently generated in our laboratory (Whittle, Kulkarni, et al. 2021b) and are available at NCBI (PRJNA564136). Cutadapt v3.4 (Martin 2011) was used to remove adapters and reads shorter than 20 nucleotides from the original fastq files. The gene expression in each library was quantified in transcripts per million (TPM) with RSEM v1.2.29 (Li and Dewey 2011), mapping the reads with STAR v2.7.0e1 (Dobin, et al. 2013) against the *G. bimaculatus* genome (Ylla, et al. 2021).

The *G. bimaculatus* gene GBI_01840 was identified as the putative locus coding for the previously reported *G. bimaculatus* Oskar protein (AFV31610.1) (Ewen-Campen, et al. 2012). Therefore, the TPM corresponding to the gene GBI_01840 in each library were considered to represent *oskar* mRNA expression in each of the sequenced libraries.

Semi-quantitative RT-PCR assessment of tissue-level *oskar* expression

Study species were as follows: live *Drosophila melanogaster* (Oregon R; Bloomington Stock Center #5); live *Callosobruchus maculatus* (Carolina Biological #144180); 100% ethanol-preserved *Aretaon asperrimus* (kind gift from Thies Büscher and Stanislav Gorb, Kiel University). We manually dissected ovaries, testes (including seminal vesicles), and whole heads from each specimen in 1X PBS. For *C. maculatus*, we also collected mixed-stage embryos for analysis as follows: adult female weevils were placed on a plate of fresh mung beans (Rani) and allowed to lay eggs for 24 hours. Embryos were removed from the beans by shaking beans with a wash solution (0.06M NaCl (VWR), 0.15% Triton X-100 (VWR), 50% bleach (Clorox) in distilled water) for 10 minutes. Embryos were collected in a basket and then rinsed with distilled water several times. Eggs and dissected tissues were transferred to TRIzol (Invitrogen) for RNA extraction. Total RNA was extracted using the manufacturer's protocol, including treatment of the RNA with DNase (Ambion) for 30 minutes at 37°C to remove genomic DNA. cDNA was synthesized using SuperScript III (Invitrogen). The amount of RNA used for DNase treatment and cDNA synthesis was standardized for tissues of a given species using a Nanodrop spectrophotometer. PCR was conducted with the resulting cDNA templates using Phusion polymerase (New England Biolabs) and the following PCR program: 98°C for 3 min; 35 cycles of 98°C for 30 sec, 70°C for 30 sec, 72°C for 30 sec (*D. melanogaster*, *C. maculatus*) or 1 min (*A. asperrimus*); and 72°C for 10 min. Primers were designed for both *oskar* and a housekeeping control gene for each species (see sequences of primers in table below). A pair of primers for each species were designed to span multiple exons, when possible, to test for genomic DNA contamination; no such contamination was detected. Reactions were run on 1.25% agarose gels

with a 1Kb plus DNA ladder (Invitrogen), and DNA products were visualized with Apex Safe DNA Gel Stain (Genesee Scientific).

Gene	Forward Primer	Reverse Primer	Product Size
Dmel oskar 1	ATGACGCCCACGCCAACGATTT	GCAATCAAATCGCACCACGCCC	534 bp
Dmel oskar 2	TTGCTGAGCCACGCCCAGAATG	GGCGGTTTTCAGTCGGTTCGGT	628 bp
Dmel RpL32	CACCAGTCGGATCGATATGC	CGATCCGTAACCGATGTTG	120 bp
Cmac oskar 1	GCTCTCAAAAGTGGCCAAAAACGA	ACTGCCAGCACAAATGACTTCA	700 bp
Cmac oskar 2	TGTGCTGGAAAAGTGCCACATGT	TAGGCAGCTTGGGAGACGGTGG	505 bp
Cmac RpL32	CAACTGCTGAGCACGTTCCACA	GGGTGAGAAGGCGCTTCAAGGG	224 bp
Aasp oskar 1	CGTCTCTTGCGTGGGCCATCAG	GCCGATGACTTGCCCGTTCCTC	380 bp
Aasp oskar 2	GAGGAACGGGCAAGTCATCGGC	ACTTCTTCACGCGGTGCAAGCA	294 bp
Aasp oskar 3	GGGAGGAGTTCCATTAGCACGGA	AGGCCTGGAAACTTTTGGCGGT	532 bp
Aasp RpL32	GCCATCTGTGGGTTACGGCAGC	GCTACGCAGGCGAGCATTTGCA	229 bp

Supplementary Figure Legends

Supplementary Figure S1: Summary statistics of the search for *oskar* orthologs. (a) Summary of searches and results for each of the three sources of data searched, from left to right: (i) The total number of datasets searched from all three sources (TSA: Transcriptome Shotgun Assembly Database; GCA: GenBank; GCF: RefSeq); (ii) the number of filtered *oskar* sequences identified in each of those datasets; and (iii) the proportion of filtered *oskar* sequences identified in each of the three sources. **(b)** Summary statistics broken down by insect orders. Only orders where an *oskar* sequence was identified are shown. From left to right: (iv) The number of *oskar* sequences identified in each of the three data sources; (v) the total number of filtered *oskar* sequences identified per order; (vi) the proportion of all searched datasets per order where an *oskar* sequences was identified. See also Supplementary Table 1

Supplementary Figure S2: Genome and transcriptome quality correlation to *oskar* identification. Shown are box plots of the distribution of *oskar* orthologs identified (ortholog identified or not identified) with respect to multiple genome and transcriptome quality metrics. For each metric, the means of both distributions were tested for significant differences using a Mann Whitney U test. A bar with an * is displayed if the p-value was less than 0.05. Mean and median values presented in Supplementary Table S2.

Supplementary Figure S3: Evidence for loss of *oskar* in Lepidoptera. Phylogeny of the Lepidoptera as per (Kawahara, et al. 2019). Next to each lepidopteran family are shown summary data regarding the status of *oskar* identification in our searches. Symbols with column labels in order from left to right: (i) vertical rectangles: grey: no *oskar* ortholog was identified in this family; range: at least one *oskar* ortholog was identified in this order. (ii) number of datasets searched. (iii) horizontal rectangles: proportion of searched datasets in which an *oskar* ortholog was identified; colors as in (i); numbers and proportions at right. (iv) pie chart: proportion of *oskar* sequences identified in RefSeq (GCF) datasets; numbers and proportions at right. (v) pie chart: proportion of *oskar* sequences identified in GenBank (GCA) datasets; numbers and proportions at right. (vi) pie chart: proportion of *oskar* sequences identified in Transcriptome Shotgun Assembly Database (TSA) datasets; numbers and proportions at right. Circles to the right of some family names indicate that there is literature evidence for involvement of germ plasm (black) or no germ plasm (white) in germ cell specification. Numbers to the left of the circles indicate references to the primary literature as follows: [1]: (Kobayashi and Ando 1984); [2] (Ando and Tanaka 1979); [3] (Lautenschlager 1932); [4] (Anderson and Wood 1968); [5] (Tanaka 1987); [6] (Woodworth 1889); [7] (Eastham 1930); [8] (Berg and Gassner 1978); [9-10] (Sehl 1931; Guelin 1994); [11] (Johannsen 1929); [12] (Presser and Rutschky 1957); [13-23]: (Tomaya 1902; Schwangart 1905; Saito 1937; Miya 1953, 1958, 1975; Nakao 1999; Toshiki, et al. 2000; Nakao, et al. 2006; Nakao, et al. 2008; Nakao and Takasu 2019). No datasets were available for Urudidea, Sesidea, Alucitidea, Callidulidea, Mimallonidea, Drepanidea or Lasiocampidea at the time of analysis.

Supplementary Figure S4: Evidence for duplication of *oskar* in Hymenoptera. Phylogenetic tree of all hymenopteran Oskar sequences inferred using RaxML with 100 bootstraps. Branch length normalized to show only the topology. Each leaf is an Oskar ortholog. Gray: only one Oskar sequence was identified in this species. Red: putatively duplicated Oskar sequences (sequence

similarity < 80%; see Methods). Families containing *oskar* duplications are highlighted as per Figure 4.

Supplementary Figure S5: Tissue-level *oskar* expression in a mosquito. *oskar* transcripts per million (TPM) in different mosquito tissues based on data reported in REF. (a, b) Quantified from the raw reported data using kallisto (Bray, et al. 2016) and (c, d) using the original study's quantification results (Matthews, et al. 2016). Error bars show the standard error of at least three independent measurements. All tissues reported in the study that were not brain, ovaries or abdominal tips were placed in the "other tissues" category. As Matthews and colleagues (2016) performed RNA sequencing experiments on female mosquitoes before and after a blood meal, (b) and (d) show the impact of the two different feedings conditions on *oskar* expression in the abdominal tip and ovaries.

Supplementary Figure S6: Tissue-level *oskar* expression in a cricket. Transcripts per million (TPM) of *oskar* in each tissue of *G. bimaculatus* for which RNA-seq was generated by Whittle, Kulkarni, et al. (2021b). For each tissue type, either two or three biological samples were sequenced as replicates. Each dot represents the expression level of *oskar* in a single biological sample.

Supplementary Figure S7: Tissue-level *oskar* expression in a fly, a weevil and a stick insect. Three study species were selected to evaluate tissue-level *oskar* expression patterns utilizing RT-PCR: *Drosophila melanogaster* (a), *Callosobruchus maculatus* (b), and *Aretaon asperrimus* (c). The top of each figure shows a schematic of the *oskar* transcript with exons marked if known (not drawn to scale). Approximate primer locations for PCR products are marked with arrows below each schematic. Gel images show RT-PCR products for the amplified *oskar* regions and a housekeeping control gene. Templates for these reactions included cDNA synthesized from male (m) and female (f) adult head and gonads, as well as from 0-24 hour embryos for *C .maculatus*. Water was used as a negative control.

Supplementary Figure S8: Tissue and developmental stage metadata analysis of *oskar* identification in transcriptome datasets. (a) Proportion of analyzed datasets that were sequenced from the developmental stages indicated on the Y axis. **(b)** Proportion of analyzed datasets per developmental stage in which an *oskar* ortholog was identified (red). **(c)** Proportion of analyzed datasets that were sequenced from the tissue type indicated on the Y axis. **(d)** Proportion of analyzed datasets per tissue type in which an *oskar* ortholog was identified (red)

Supplementary Figure S9: Evolution of the structure of Oskar in Diptera. Left: dipteran phylogeny from (Maddison, et al. 2007; Wiegmann, et al. 2011). Top: schematic representation of Oskar domain structure. Blue: heatmap showing the overall occupancy of an amino acid position in the Oskar alignment trimmed for at least 10% overall occupancy at a given position. For each dipteran family, occupancy at a given position is defined as (number of non-gap amino acids / number of sequences in that family). If a 3' or 5' extension (defined as a coding sequence unbroken by stop codons, 5' of the first residue of the LOTUS domain, or 3' of the last residues of the OSK domain but 5' to a predicted poly-A tail) was detected in a family, a black box outlines the putative domain. Any such identified 5' domains were designated as putative "Long Oskar" domains.

Supplementary Figure S10: Multiple Correspondence Analysis (MCA) of full-length Oskar, the OSK domain and the LOTUS domain. MCA analysis of trimmed (30% occupancy) alignments for (a) full-length Oskar, (b) the OSK domain and (c) the LOTUS domain colored by insect order (see legend at right). The alignment was projected onto the first three main MCA dimensions (1, 2 and 3). Each dot corresponds to one sequence. Dotted line outlines specific families of interest as discussed in the text

Supplementary Figure S11: Oskar domains secondary structure conservation. Sequence Logo of Jpred4 predictions for LOTUS and OSK domains showing the conservation of secondary structures, computed with WebLogo (Crooks, et al. 2004). The height of each letter represents that state's (X, H or B) conservation throughout the alignment in bits. X (black): unfolded amino acids; H (red): α helices; E (blue): β sheets. **(a)** Prediction for the LOTUS domain. **(b)** Prediction for the OSK domain.

Supplementary Figure S12: Duplications and losses of *oskar* in Hymenoptera. Absence (magenta) or presence of *oskar* orthologs detected in single copy (cyan) or multiple copies (yellow) in the genomic or transcriptomic datasets examined in this study. Genera shown in italics indicate individual species searched and are abbreviated simply for space reasons. Genera shown in regular type (not italics) indicate a summary of the results from multiple congeneric species, which were nearly always consistent within genera; in all cases where intrageneric results for *oskar* presence or absence were inconsistent, we gave precedence for the finding obtained from a genome sequence (GCF or GCA) over findings obtained from a transcriptome (TSA)), o for Hymenoptera species or genera. For some species, germ cell specification via germ plasm (black circles) or differentiation from mesoderm (no germ plasm; white circles) has been reported in the literature, with primary data references indicated by numbers as follows: [1-6]: (Bütschli 1870; Fleig and Sander 1985, 1986; Zissler 1992; Gutzeit, et al. 1993; Dearden 2006); [7]: (Khila and Abouheif 2008); [8-10]: (Bull 1982; Lynch and Desplan 2010; Lynch, et al. 2011); [11]: (Koscielska and Koscielski 1987); [12-13]: (Silvestri 1906, 1908); [12, 14-21]: (Silvestri 1906; Hegner 1914; Grbic', et al. 1996; Strand and Grbic' 1997; Grbic' 2000, 2003; Donnell, et al. 2004; Zhurov, et al. 2004); [22-25]: (Gatenby 1917a; Gatenby 1917b; Gatenby 1918; Gatenby 1920); [24]: (Amy 1961); [27-28]: (Gatenby 1920; Tawfik 1957); [29-30]: (Bronskill 1959; Fleischmann 1975); [31]: (Shafiq 1954); [32]: (Sumitani, et al. 2003). Phylogenetic relationships as per (Nyman, et al. 2006; Field, et al. 2011; Schmidt 2013; Prous, et al. 2014; Ward 2014; Malm and Nyman 2015; Vilhelmsen 2015; Ward, et al. 2016; Peters, et al. 2017; Chen and Achterberg 2018; Peters, et al. 2018; Sharanowski, et al. 2021). Evolution of major hymenopteran life history characteristics (eusociality, pollen collecting, stinger, parasitoidism) as per (Peters, et al. 2017).

Supplementary References

Amy RL. 1961. The embryology of *Habobracon juglandis* (Ashmead). Journal of Morphology 109:199-217.

Anderson DT, Wood EC. 1968. The morphological basis of embryonic movements in the light brown apple moth, *Epiphyas postvittana* (Walk.) (Lepidoptera, Tortricidae). Australian Journal of Zoology 16:763-793.

Ando H, Tanaka M. 1979. Early embryonic devekopment of the primitive moths, *Enduclyta signifer* Walker and *E. excrescens* Butler (Lepidoptera: Hepialidae). International Journal of Insect Morphology and Embryology 9:67-77.

Berg GJ, Gassner G. 1978. Fine structure of the blastoderm embryo of the pink bollworm, *Pectinophora gossypiella* (Saunders) (Lepidoptera: gelechiidae). International Journal of Insect Morphology and Embryology 1:81+105.

Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. Nature Biotechnology 34:525-527.

Bronskill JF. 1959. Embryology of *Pimpla turionellae* (L.) (Hymenoptera: Ichneumonidae). Canadian Journal of Zoology 37:655-688.

Bull AL. 1982. Stages of living embryos in the jewel wasp *Mormoniella (Nasonia) vitripennis* (Walker) (Hymenoptera: Pteromalidae). International Journal of Insect Morphology and Embryology 11:1-23.

Bütschli O. 1870. Zur Entwicklungsgeschichte der Biene. Zeitschrift für Wissenschaftliche Zoologie 20:519-564.

Chen X-x, Achterberg Cv. 2018. Systematics, Phylogeny, and Evolution of Braconid Wasps: 30 Years of Progress. Annual Review of Entomology 64:1-24.

Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence logo generator. Genome Research 14:1188-1190.

Dearden PK. 2006. Germ cell development in the Honeybee (*Apis mellifera*); *vasa* and *nanos* expression. BMC Developmental Biology 6:6.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29:15-21.

Donnell DM, Corley LS, Chen G, Strand MR. 2004. Caste determination in a polyembryonic wasp involves inheritance of germ cells. Proceedings of the National Academy of Sciences of the United States of America 101:10095-10100.

Eastham LES. 1930. The embryology of *Pieris rapae* - Organogeny. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 219:1-50.

293 Ewen-Campen B, Srouji JR, Schwager EE, Extavour CG. 2012. Oskar predates the evolution of
294 germ plasm in insects. Curr Biol 22:2278-2283.

295 Field J, Ohl M, Kennedy M. 2011. A molecular phylogeny for digger wasps in the tribe
296 Ammophilini (Hymenoptera, Apoidea, Sphecidae). Systematic Entomology 36:732-740.

297 Fleig R, Sander K. 1985. Blastoderm development in honey bee embryogenesis as seen in the
298 scanning electron microscope. International Journal of Invertebrate Reproduction and
299 Development 8:279-286.

300 Fleig R, Sander K. 1986. Embryogenesis of the Honeybee *Apis mellifera* L (Hymenoptera,
301 Apidae) - an SEM Study. International Journal of Insect Morphology and Embryology 15:449-
302 462.

303 Fleischmann VG. 1975. Origin and embryonic development of fertile gonads with and without
304 pole cells of *Pimpla turionellae* L. (Hymenoptera, Ichneumonidae). Zool. Jb. Anat. Bd. 94:375-
305 411.

306 Gatenby JB. 1920. The Cytoplasmic Inclusions of the Germ Cells. Part VI. On the origin and
307 probable constitution of the germ-cell determinant of *Apanteles glomeratus*, with a note on the
308 secondary nuclei. Quarterly Journal of Microscopical Science 64:133-153.

309 Gatenby JB. 1917a. The embryonic development of *Trichogramma evanescens* Westw.,
310 monoembryonic egg parasite of *Donacia simplex*. Quarterly Journal of Microscopical Science
311 62:149-187.

312 Gatenby JB. 1918. The segregation of germ cells in *Trichogramma evanescens*. Quarterly
313 Journal of Microscopical Science 63:161-173.

314 Gatenby JB. 1917b. The segregation of the germ-cells in *Trichogramma evanescens*. Quarterly
315 Journal of Microscopical Science 62:149-187.

316 Grbic' M. 2000. "Alien" wasps and evolution of development. Bioessays 22:920-932.

317 Grbic' M. 2003. Polyembryony in parasitic wasps: evolution of a novel mode of development.
318 International Journal of Developmental Biology 47:633-642.

319 Grbic' M, Nagy LM, Carroll SB, Strand M. 1996. Polyembryonic development: insect pattern
320 formation in a cellularised environment. Development:795-804.

321 Guelin M. 1994. [Activity of W-sex heterochromatin and accumulation of the nuage in nurse
322 cells of the lepidopteran *Ephestia*]. C. R. Acad. Sci. Paris. Ser. III 317:54-61.

323 Gutzeit HO, Zissler D, Fleig R. 1993. Oogenesis in the Honeybee *Apis mellifera* - Cytological
324 Observations on the Formation and Differentiation of Previtellogenic Ovarian Follicles. Rouxs
325 Archives of Developmental Biology 202:181-191.

Hegner RW. 1914. Studies on germ cells. III. The origin of the Keimbahn-determinants in a parasitic Hymenopteran, *Copidosoma*. Anatomischer Anzeiger 3-4:51-69.

Johannsen OA. 1929. Some phases in the embryonic development of *Diacrisia virginica* Fabr. (Lepidoptera). J. Morphol. Physiol. 2:493-541.

Kawahara AY, Plotkin D, Espeland M, Meusemann K, Toussaint EFA, Donath A, Gimnich F, Frandsen PB, Zwick A, Dos Reis M, et al. 2019. Phylogenomics reveals the evolutionary timing and pattern of butterflies and moths. Proceedings of the National Academy of Sciences of the United States of America 116:22657-22663.

Khila A, Abouheif E. 2008. Reproductive constraint is a developmental mechanism that maintains social harmony in advanced ant societies. Proceedings of the National Academy of Sciences of the United States of America 105:17884-17889.

Kobayashi Y, Ando H. 1984. Mesodermal Organogenesis in the Embryo of the Primitive Moth, *Neomicropteryx nipponensis* Issiki (Lepidoptera, Micropterygidae). Journal of Morphology 181:29-47.

Koscielska MK, Koscielski B. 1987. Early embryonic development of *Tritneptis diprionis* (Chalcidoidea, Hymenoptera). In: Ando H, Jura C, editors. Recent Advances in Insect Embryology in Japan and Poland. Tsukuba: Arthropod. Embryol. Soc. Jpn.

ISEBU Co. Ltd. p. 207-214.

Laurence BR. 1977. Ovary development in mosquitoes: a review. Adv. Invertebr. Repr. 1:154-165.

Lautenschlager F. 1932. Die Embryonalentwicklung der weiblichen Keimdruse bei der Psychide *Solenobia triquetella*. Zool. Jarh. 56:121-162.

Leader DP, Krause SA, Pandit A, Davies SA, Dow JAT. 2017. FlyAtlas 2: a new version of the Drosophila melanogaster expression atlas with RNA-Seq, miRNA-Seq and sex-specific data. Nucleic Acids Research 46:gkx976-.

Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12:323.

Lynch JA, Desplan C. (p16123 co-authors). 2010. Novel modes of localization and function of *nanos* in the wasp *Nasonia*. Development 137:3813-3821.

Lynch JA, Özüak O, Khila A, Abouheif E, Desplan C, Roth S. 2011. The Phylogenetic Origin of *oskar* Coincided with the Origin of Maternally Provisioned Germ Plasm and Pole Cells at the Base of the Holometabola. PLoS Genetics 7:e1002029.

Maddison DR, Schultz K-S, Maddison WP. 2007. The Tree of Life Web Project. Zootaxa 1668:19-40.

Malm T, Nyman T. 2015. Phylogeny of the symphytan grade of Hymenoptera: new pieces into the old jigsaw(fly) puzzle. Cladistics 31:1-17.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. journal 17:10-12.

Matthews BJ, McBride CS, DeGennaro M, Despo O, Vosshall LB. 2016. The neurotranscriptome of the Aedes aegypti mosquito. BioMedCentral Genomics 17:32.

Miya K. 1953. The presumptive genital region at the blastoderm stage of the silkworm egg. Journal of the Faculty of Agriculture of Iwate University:223-227.

Miya K. 1958. Studies on the embryonic development of the gonad in the silkworm, *Bombyx mori* L. Part I. Differentiation of germ cells. Journal of the Faculty of Agriculture of Iwate University 3:436-467.

Miya K. 1975. Ultrastructural changes of embryonic cells during organogenesis in the silkworm, *Bombyx mori*. I. The Gonad. Journal of the Faculty of Agriculture of Iwate University 12:329-338.

Nakao H. 1999. Isolation and characterization of a *Bombyx vasa*-like gene. Development Genes and Evolution 209:312-316.

Nakao H, Hatakeyama M, Lee JM, Shimoda M, Kanda T. 2006. Expression pattern of *Bombyx* vasa-like (BmVLG) protein and its implications in germ cell development. Development Genes and Evolution 216:94-99.

Nakao H, Matsumoto T, Oba Y, Niimi T, Yaginuma T. 2008. Germ cell specification and early embryonic patterning in Bombyx mori as revealed by nanos orthologues. Evolution and Development 10:546-554.

Nakao H, Takasu Y. 2019. Complexities in Bombyx germ cell formation process revealed by Bm-nosO (a Bombyx homolog of nanos) knockout. Developmental Biology 445:29-36.

Nyman T, Zinovjev AG, Vikberg V, Farrell BD. 2006. Molecular phylogeny of the sawfly subfamily Nematinae (Hymenoptera: Tenthredinidae). Systematic Entomology 31:569-583.

Peters RS, Krogmann L, Mayer C, Donath A, Gunkel S, Meusemann K, Kozlov A, Podsiadlowski L, Petersen M, Lanfear R, et al. 2017. Evolutionary History of the Hymenoptera. Current Biology 27:1013-1018.

Peters RS, Niehuis O, Gunkel S, Bläser M, Mayer C, Podsiadlowski L, Kozlov A, Donath A, Noort Sv, Liu S, et al. 2018. Transcriptome sequence-based phylogeny of chalcidoid wasps (Hymenoptera: Chalcidoidea) reveals a history of rapid radiations, convergence, and evolutionary success. Molecular Phylogenetics and Evolution 120:286-296.

Presser BD, Rutschky CW. 1957. The embryonic development of the corn earworm, *Heliothis zea* (Boddie) (Lepidoptea, Phalaenidae). Annals of the Entomological Society of America 50:133-164.

Prous M, Blank SM, Goulet H, Heibo E, Liston A, Malm T, Nyman T, Schmidt S, Smith DR, Vårdal H, et al. 2014. The genera of Nematinae (Hymenoptera, Tenthredinidae). Journal of Hymenoptera Research 40:1-69.

Saito. 1937. On the development of the Tusser, *Antheraea pernyi* Guerin-Meneville, with special reference to the comparative embryology of insects. Journal of the Faculty of Agriculture of Hokkaido Imperial University 40:35-109.

Schmidt C. 2013. Molecular phylogenetics of ponerine ants (Hymenoptera: Formicidae: Ponerinae). Zootaxa 3647:201-250.

Schwangart F. 1905. Zur Entwickslungsgeschichte der Lepidopteren. Biol. Centralbl. 25:777-789.

Sehl A. 1931. Furchung und Bildung der Keimanlage bei der Mehlmotte *Ephestia kuehniella*. Zell. Zeit. Morph. U. Okol. 1:429-506.

Shafiq SA. 1954. A study of the embryonic development of the Gooseberry Sawfly, *Pteronidea ribesii*. Quarterly Journal of Microscopical Science 95:93-114.

Sharanowski BJ, Ridenbaugh RD, Piekarski PK, Broad GR, Burke GR, Deans AR, Lemmon AR, Lemmon ECM, Diehl GJ, Whitfield JB, et al. 2021. Phylogenomics of Ichneumonoidea (Hymenoptera) and implications for evolution of mode of parasitism and viral endogenization. Molecular Phylogenetics and Evolution 156:107023.

Silvestri F. 1906. Contribuzioni alla conoscenza biologica degli Imenotteri parassiti. I. Biologia del *Litomastix truncellatus* Dalm. Annali della r. Scuola Superiore di Agricoltura in Portici 6:3-51.

Silvestri F. 1908. Contribuzioni alla conoscenza degli Imenotteri parassiti. Bollettino del Laboratorio di Zoologia Generale e Agraria della r. Scuola Superiore d'Agricoltura (AFTW. Facoltà Agraria) in Portici 3:29-84.

Strand MR, Grbic' M. 1997. The Development and Evolution of Polyembryonic Insects. Current Topics in Developmental Biology 35:121-159.

Sumitani M, Yamamoto DS, Oishi K, Lee JM, Hatakeyama M. 2003. Germline transformation of the sawfly, *Athalia rosae* (Hymenoptera: Symphyta), mediated by a piggyBac-derived vector. Insect Biochem Mol Biol 33:449-458.

Tanaka M. 1987. Differentiation and behaviour of Primordial Germ Cells during the Early Embryonic Development of *Parnassius glacialis* Butler, *Luehdorfia japonica* Leech and *Byasa (Atrophaneura) alcinous alcinous* Klug (Lepidoptera: Papilionidae). In: Ando H, Jura C, editors.

Recent Advances in Insect Embryology in Japan and Poland. Tsukuba: Arthropod. Embryol. Soc. Jpn.

ISEBU Co. Ltd. p. 255-266.

Tawfik MFS. 1957. Alkaline phosphatase in the germ-cell determinant of the egg of *Apanteles*. Journal of Insect Physiology 1:286-291.

Tomaya K. 1902. On the embryology of the silkworm. Bulletin of the College of Agriculture, Tokyo 5:73-111.

Toshiki T, Chantal C, R., Toshio K, Eappen A, Mari K, Natuo K, Jean-Luc T, Bernard M, Gérard C, Paul S, et al. 2000. Germline transformation of the silkworm *Bombyx mori* L. using a piggyBac transposon-derived vector. Nature Biotech.:81-84.

Vilhelmsen L. 2015. Morphological phylogenetics of the Tenthredinidae (Insecta:Hymenoptera). Invertebrate Systematics 29:164-190.

Ward PS. 2014. The Phylogeny and Evolution of Ants. Annual Review of Ecology, Evolution, and Systematics 45:23-43.

Ward PS, Blaimer BB, Fisher BL. 2016. A revised phylogenetic classification of the ant subfamily Formicinae (Hymenoptera: Formicidae), with resurrection of the genera Colobopsis and Dinomyrmex. Zootaxa 4072:343-357.

Whittle CA, Kulkarni A, Chung N, Extavour CG. 2021. Adaptation of codon and amino acid use for translational functions in highly expressed cricket genes. BioMedCentral Genomics 22.

Whittle CA, Kulkarni A, Extavour CG. 2021a. Evolutionary dynamics of sex-biased genes expressed in cricket brains and gonads. Journal of Evolutioary Biology doi:10.1111/jeb.13889.

Whittle CA, Kulkarni A, Extavour CG. 2021b. Evolutionary dynamics of sex-biased genes expressed in cricket brains and gonads. J Evol Biol 34:1188-1211.

Wiegmann BM, Trautwein MD, Winkler IS, Barr NB, Kim J-W, Lambkin C, Bertone MA, Cassel BK, Bayless KM, Heimberg AM, et al. (r40919 co-authors). 2011. Episodic radiations in the fly tree of life. Proceedings of the National Academy of Sciences 108:5690-5695.

Woodworth CW. 1889. Studies on the embryological development of *Euvanessa antiopa*. In: Scudder, editor. Butterflies of Eastern United States and Canada. p. 102.

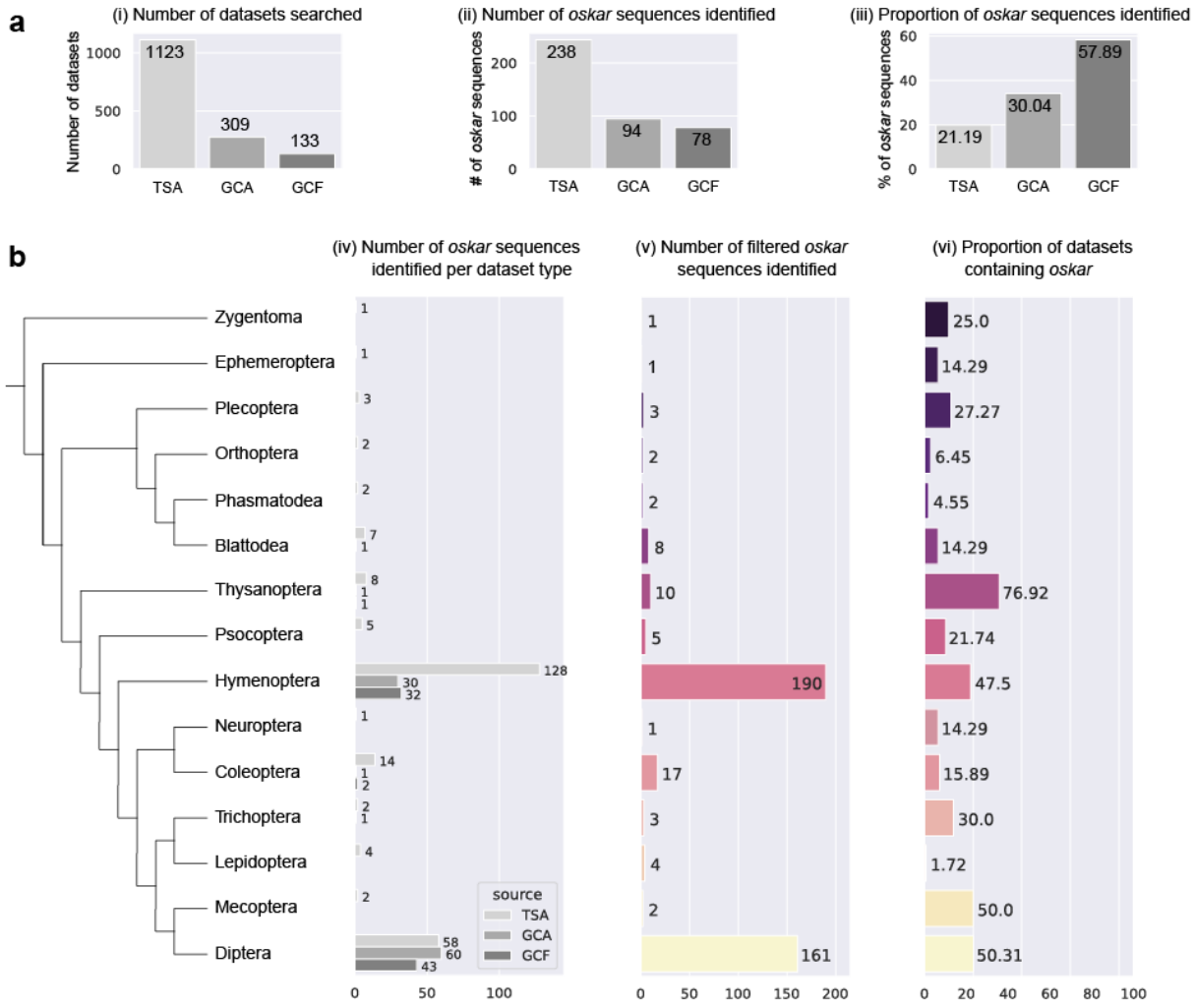
Ylla G, Nakamura T, Itoh T, Kajitani R, Toyoda A, Tomonari S, Bando T, Ishimaru Y, Watanabe T, Fuketa M, et al. 2021. Insights into the genomic evolution of insects from cricket genomes. Commun Biol 4:733.

Zhurov V, Terzin T, Grbic M. 2004. Early blastomere determines embryo proliferation and caste fate in a polyembryonic wasp. Nature 432:764-769.

461 Zissler D. 1992. From egg to pole cells: ultrastructural aspects of early cleavage and germ cell
462 determination in insects. Micr. Res. and Tech.:49-74.

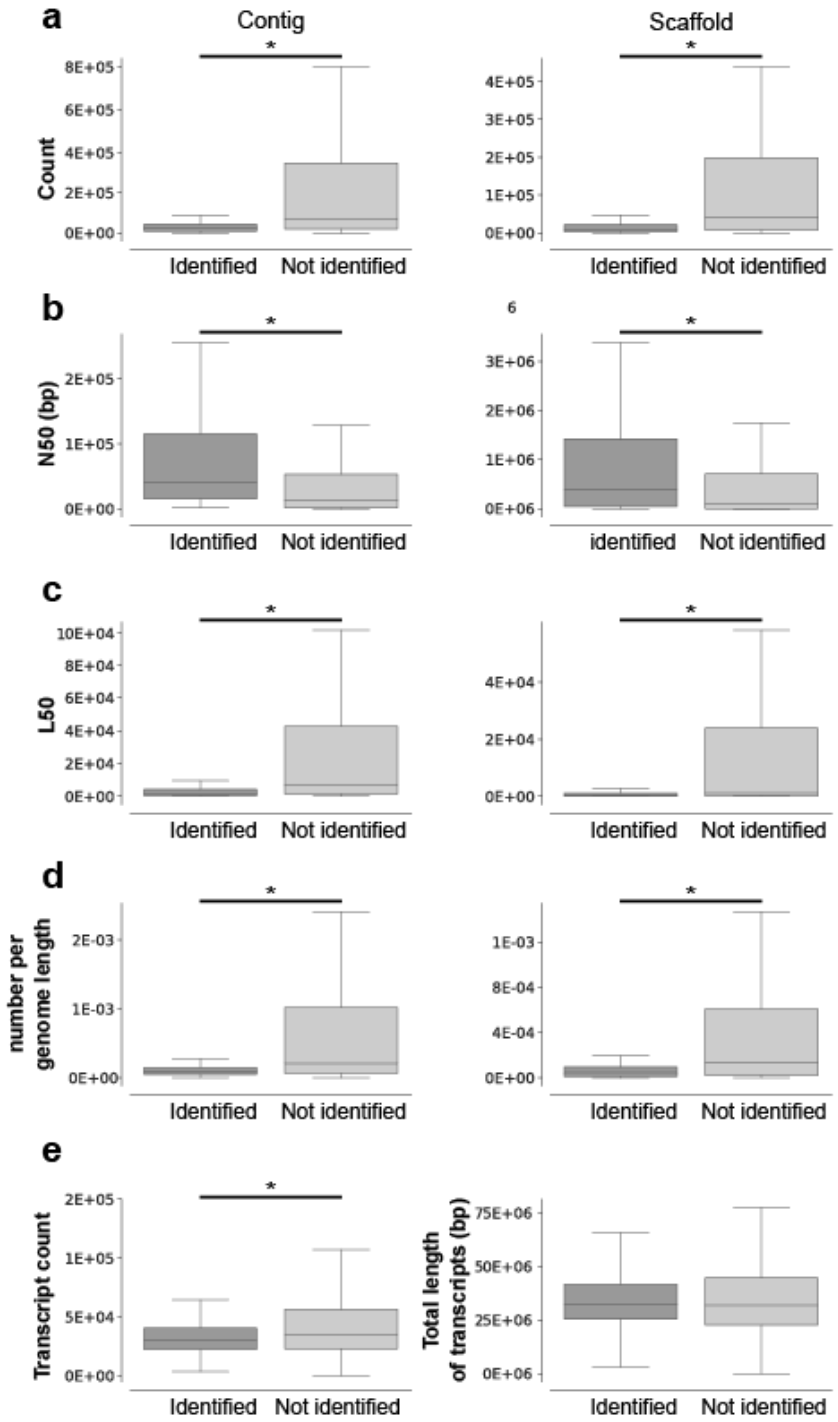
463

Supplementary Figure S1



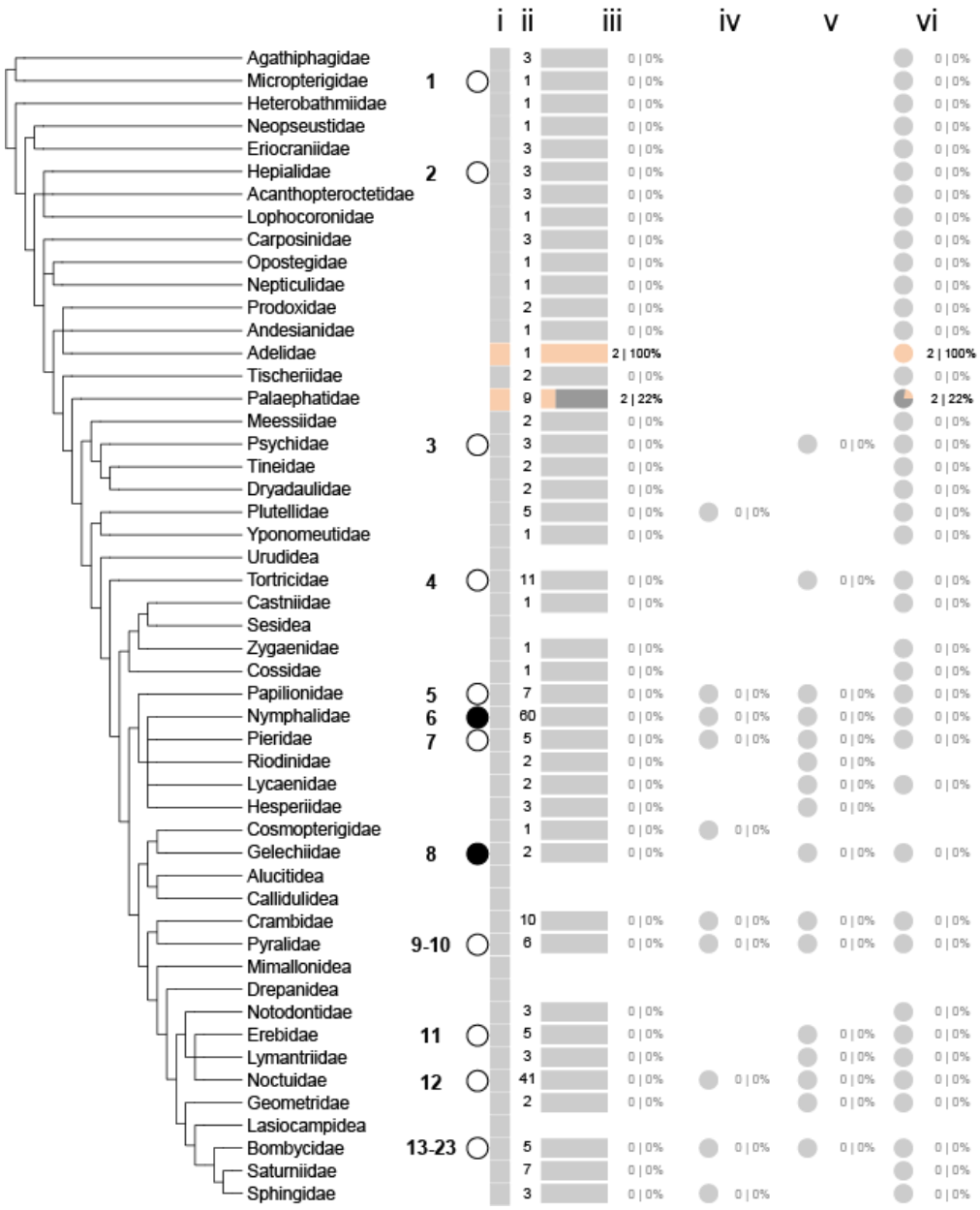
.

468 **Supplementary Figure S2**
469

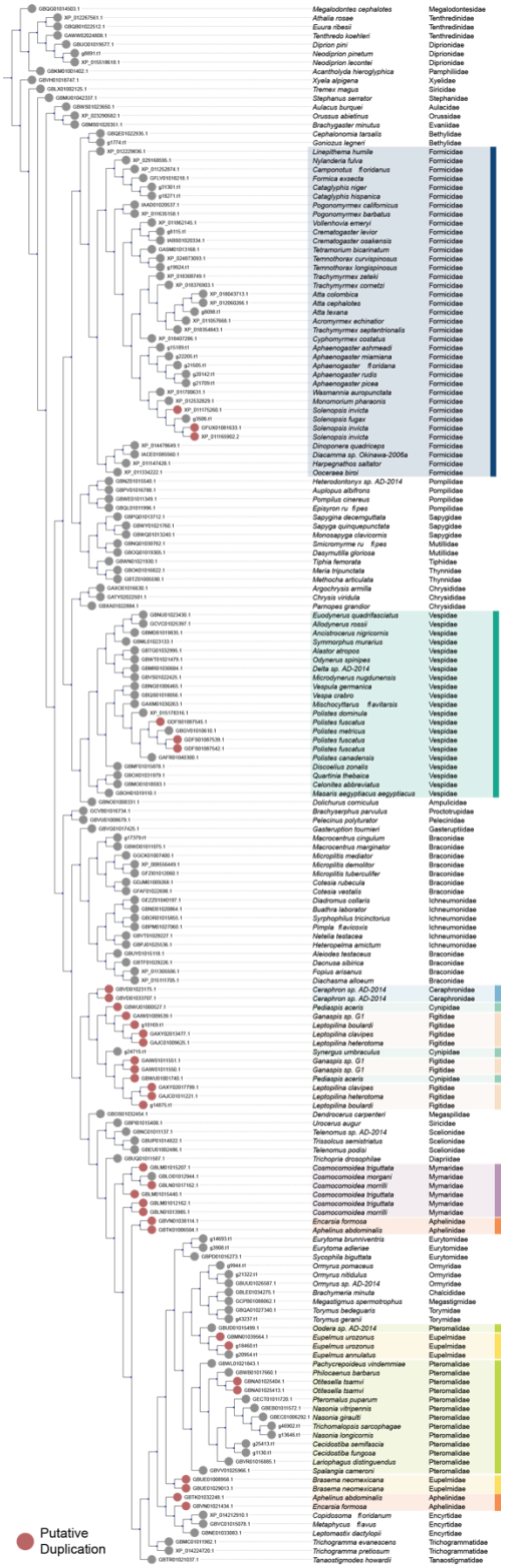


470
471
472

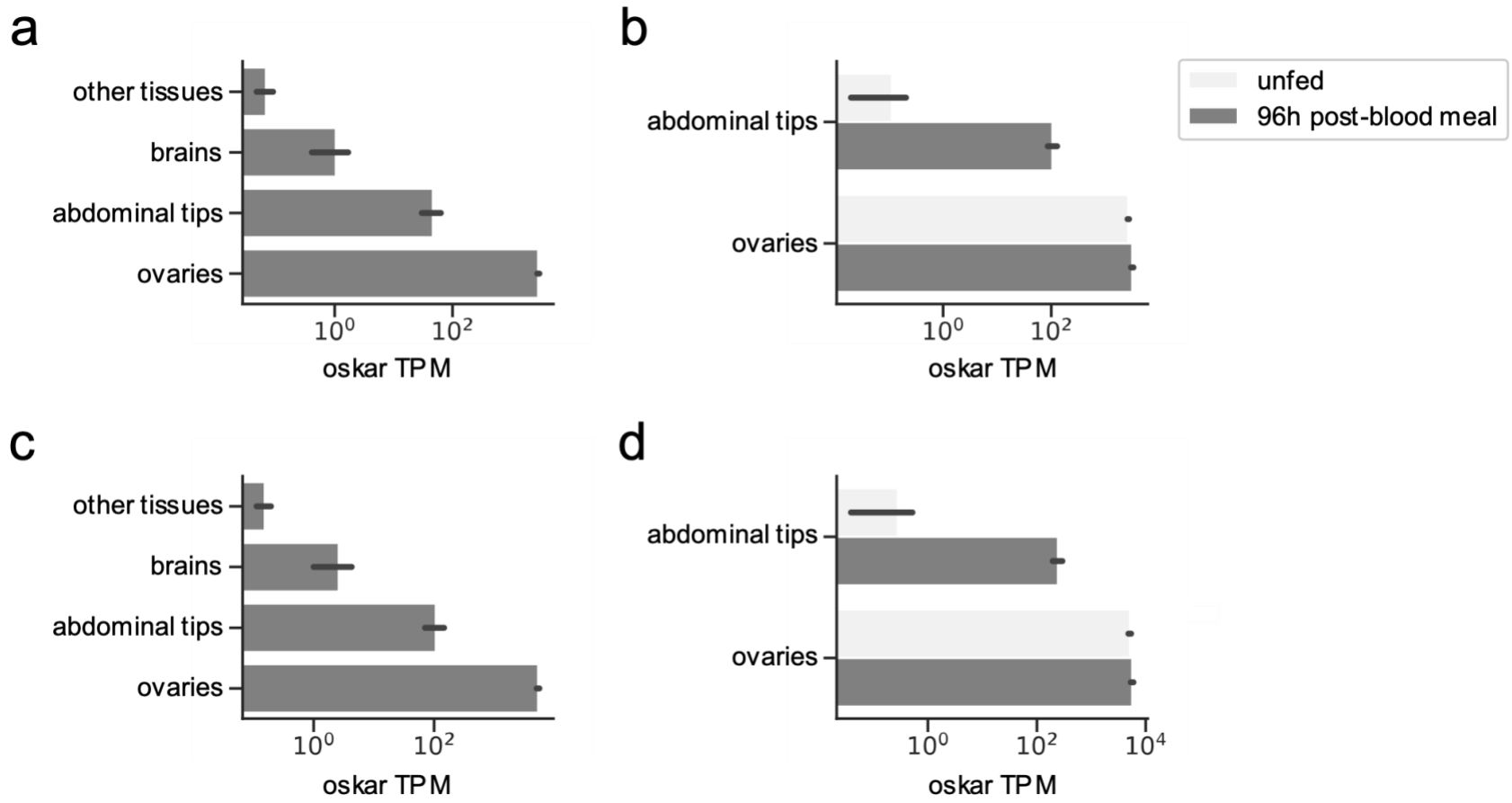
Supplementary Figure S3



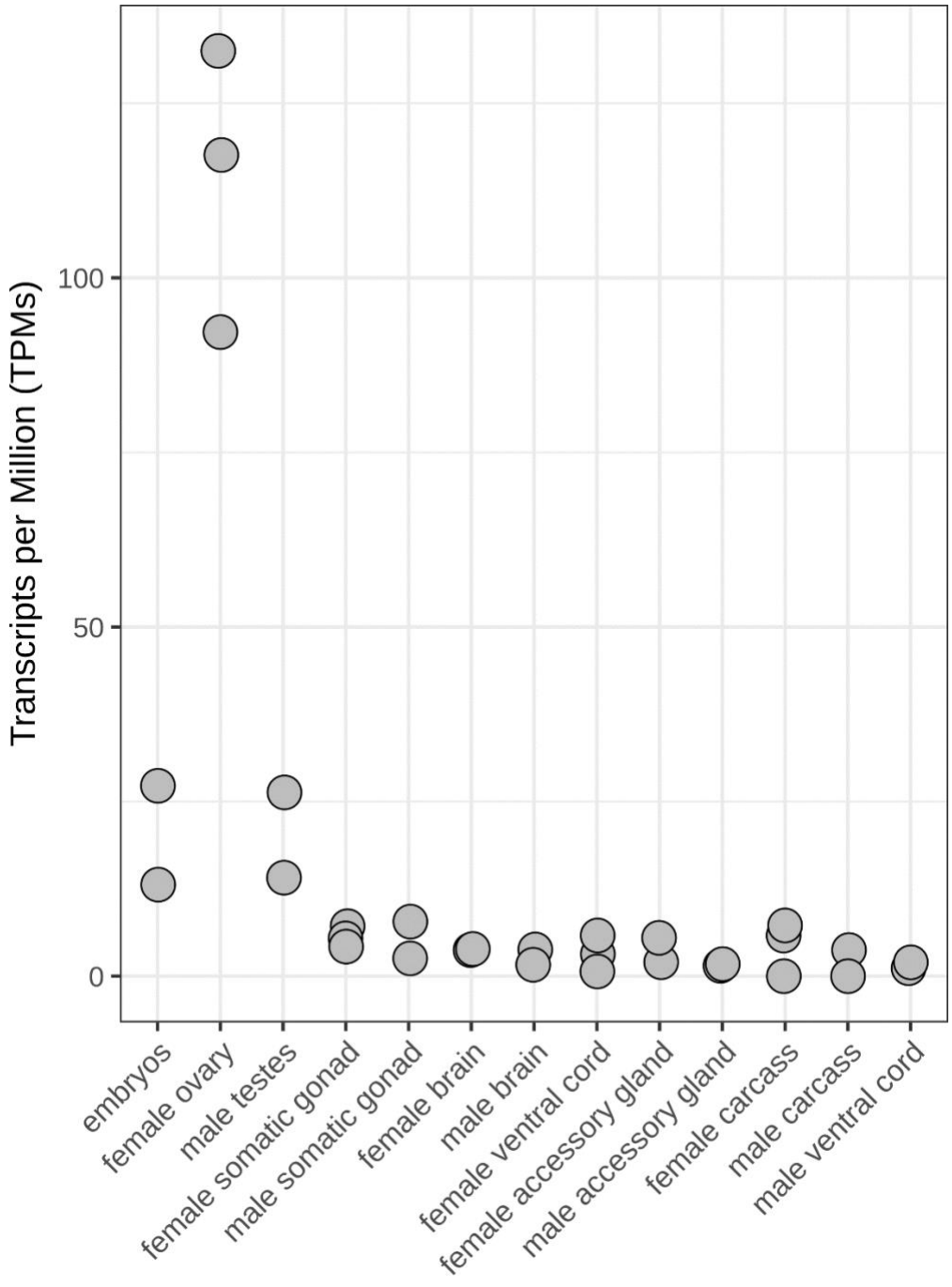
Supplementary Figure S4



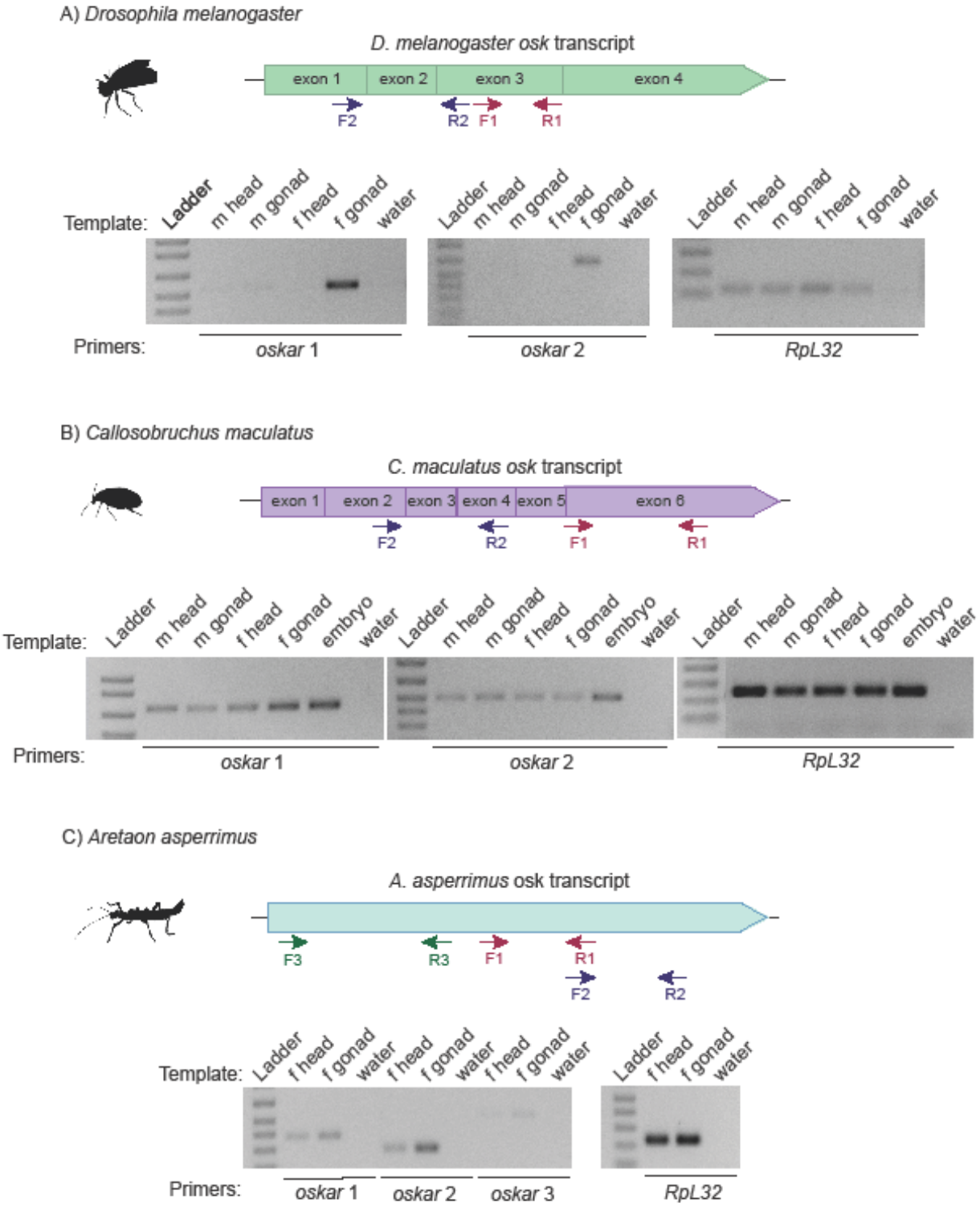
Supplementary Figure S5



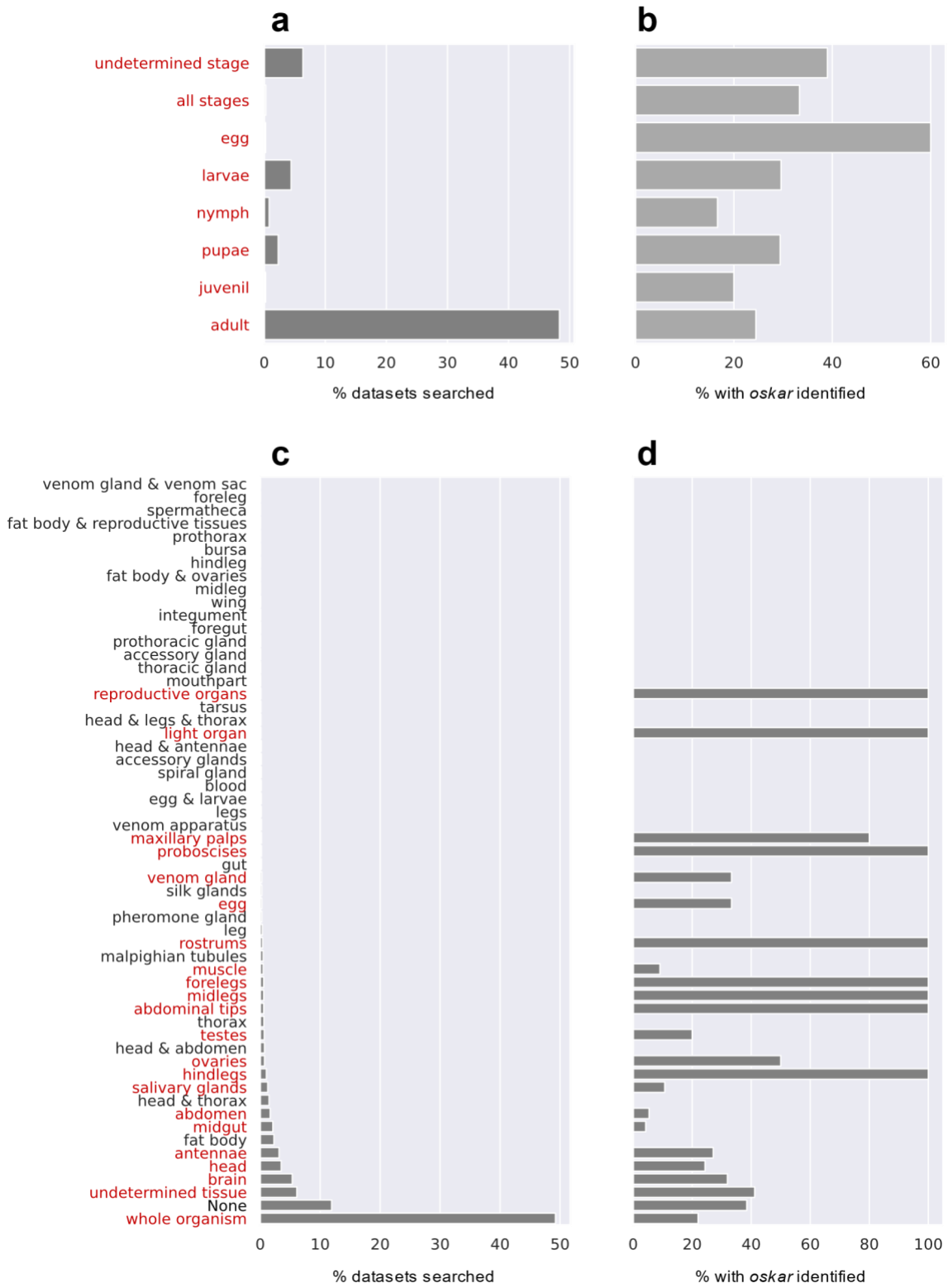
Supplementary Figure S6



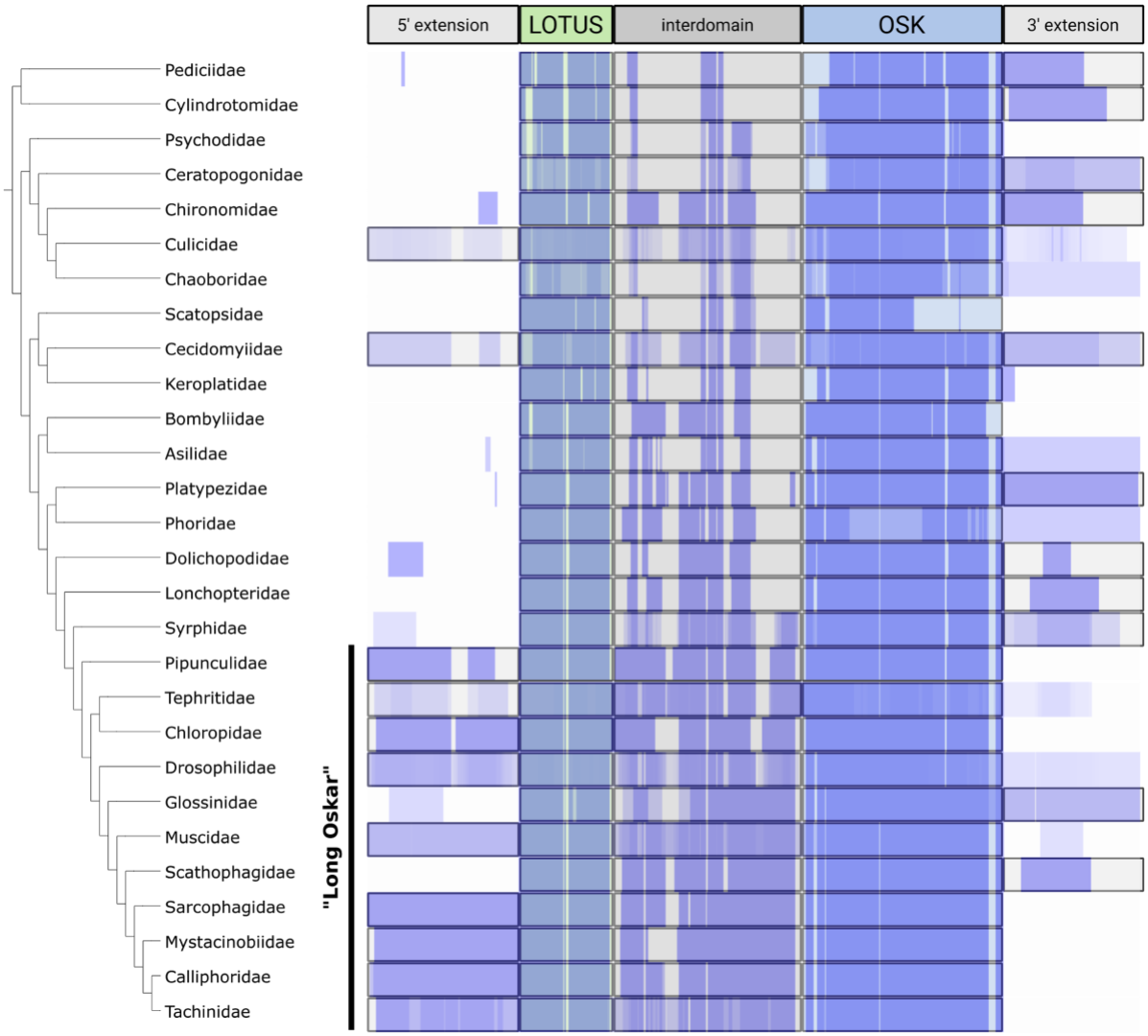
Supplementary Figure S7



Supplementary Figure S8

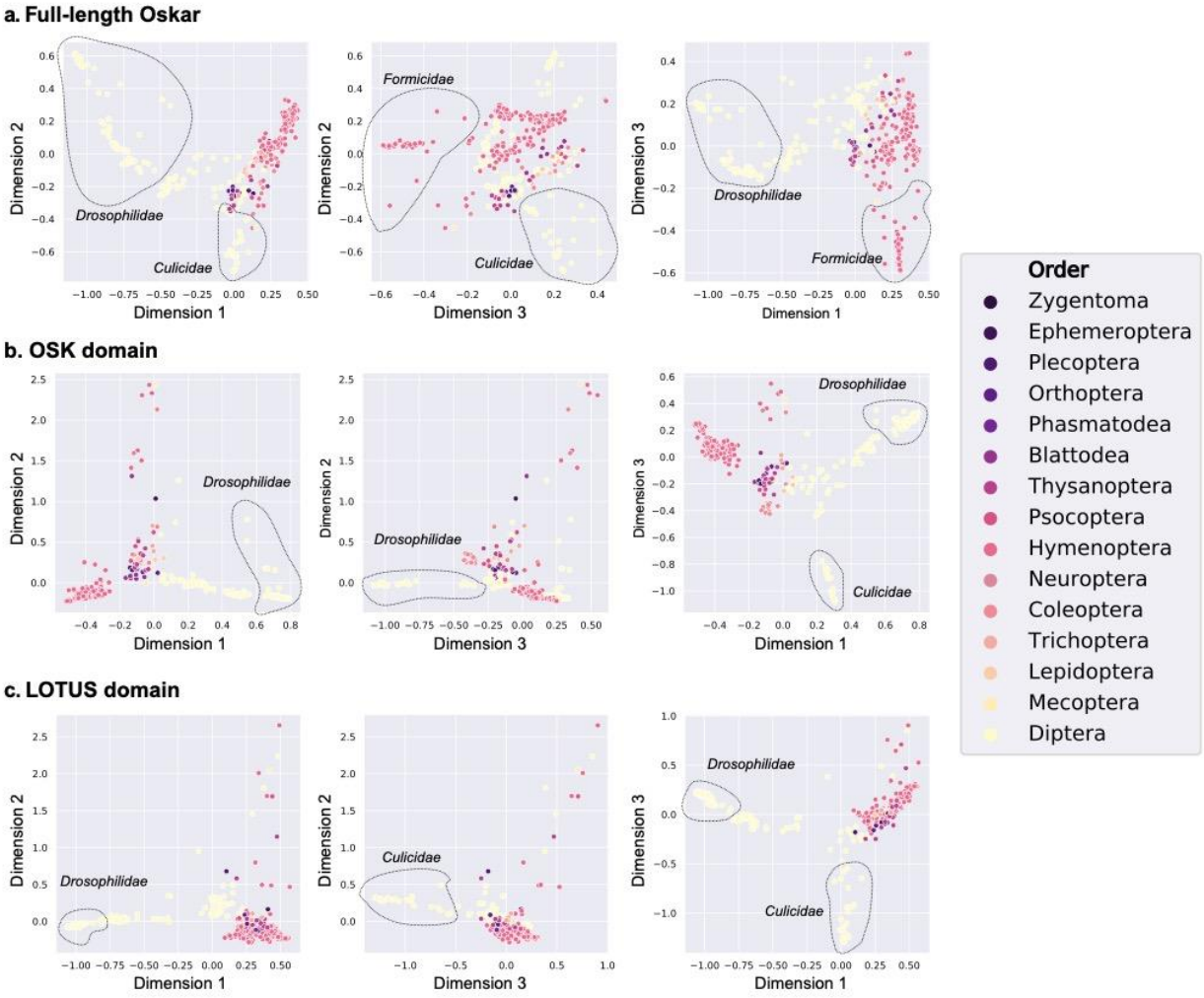


Supplementary Figure S9



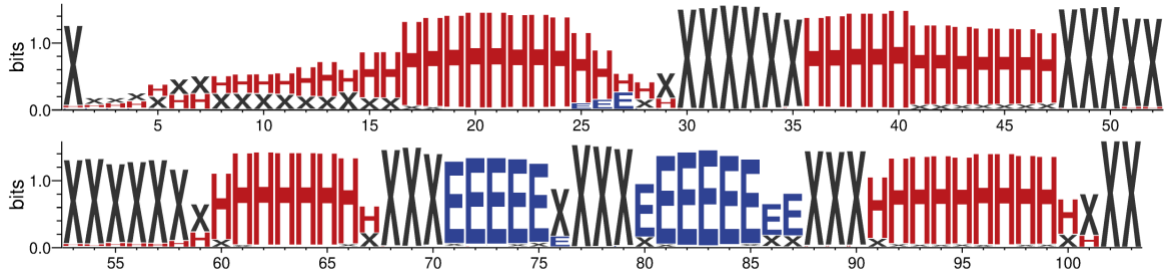
.

Supplementary Figure S10

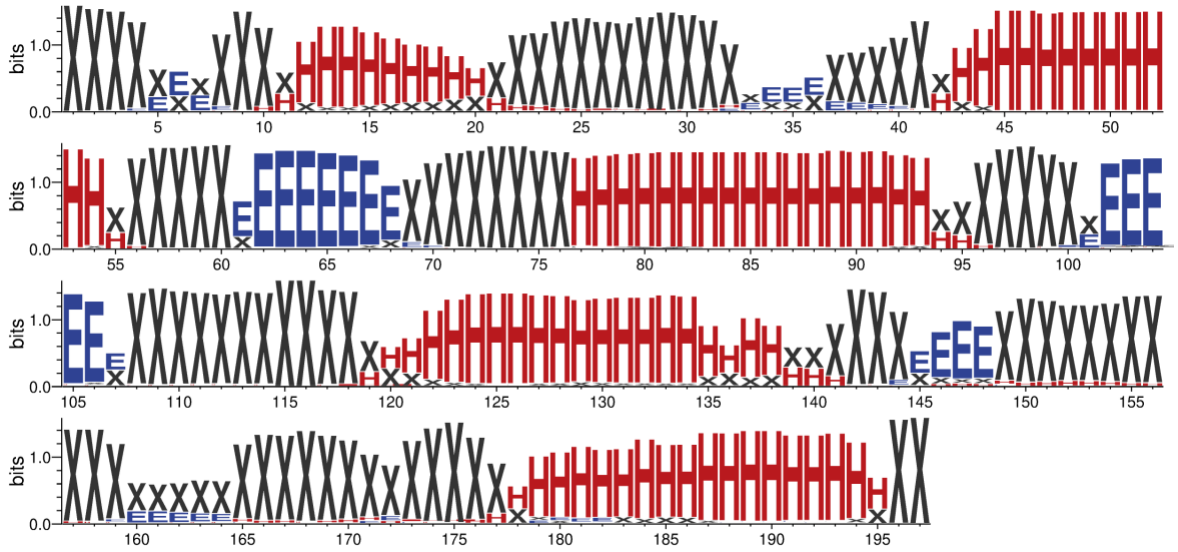


Supplementary Figure S11

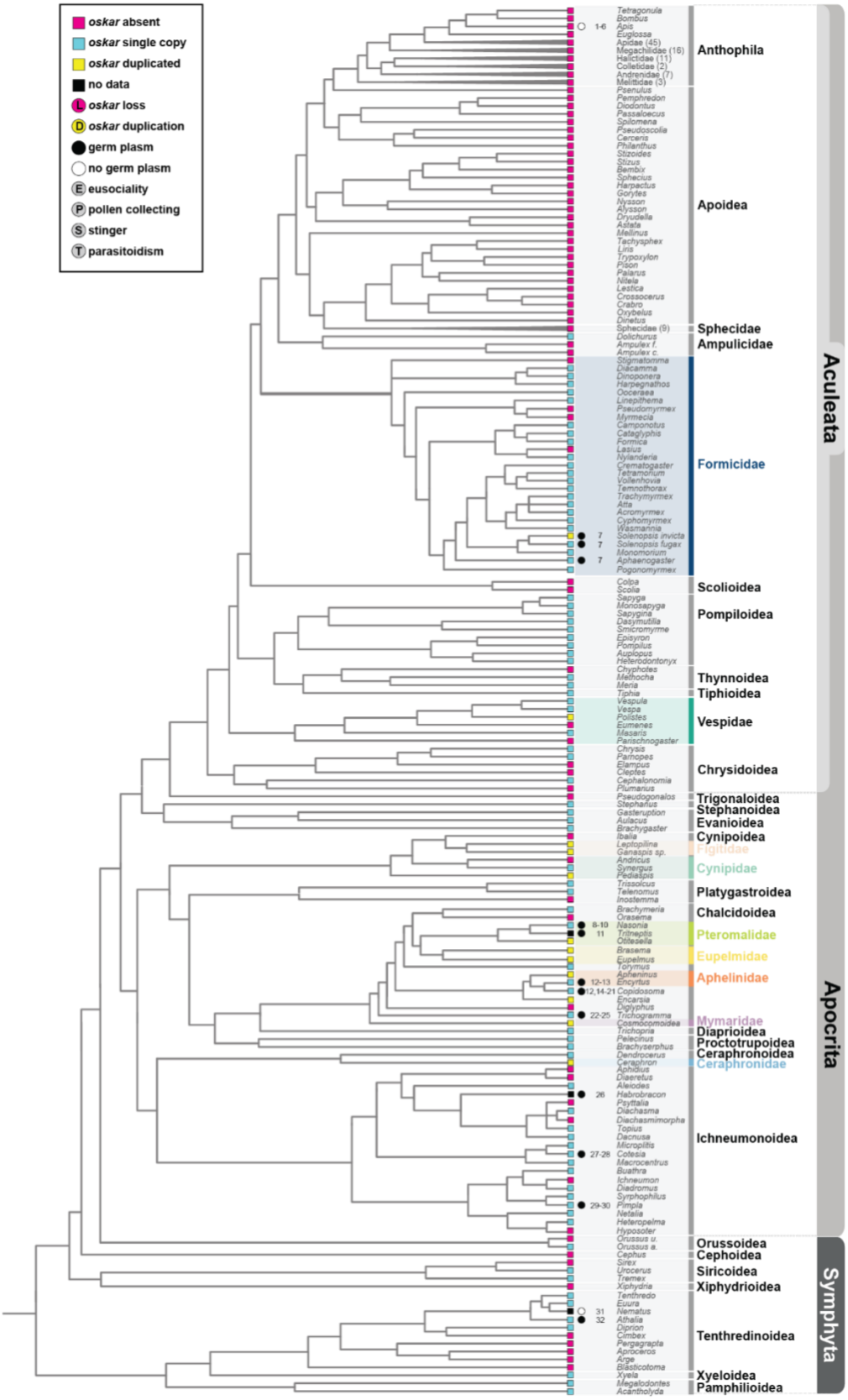
a. LOTUS secondary structure conservation



b. OSK secondary structure conservation



Supplementary Figure S12



Supplementary Table Legends

Supplementary Table S1: Number of *oskar* sequences identified per order and per data source. Each row corresponds to an order and a data source: GCF: RefSeq; GCA: GenBank, TSA: Transcriptome Shotgun Assembly Database. "Filtered hits" column indicates the number of hits after the filtration algorithm described in the Methods is applied. Rightmost column defines the proportion of *oskar* sequences identified, as the number of datasets with a filtered hit divided by the total number of datasets searched.

Supplementary Table S2: Genome quality correlation to *oskar* identification. Mean and median values for the distributions of each indicated genome quality parameter, in which *oskar* was (a) or was not (b) identified. The means of both distributions are significantly different for all metrics (Mann Whitney U test, $p<0.05$). See Supplementary Figure S2 of graphical representation of distributions.

Supplementary Table S3. Assignment of metadata to germ line or brain categories. This table is found in ***Data>02_oskar_analyses/2/3/TableS3_germline_brain_table.csv*** at the GitHub repository https://github.com/extavourlab/Oskar_Evolution.

Supplementary Table S4. Models used to create protein sequence databases. This table shows which models were used to run the *ab initio* gene detection algorithm Augustus as described in Methods and Materials. Column order corresponds to any GCA dataset of an organism from this order. "Family" column is only used if a member of this order but of a different family was used. Finally, "augustus_model" shows which GCF dataset or premade augustus model, was used to run the gene prediction. This table is found in ***Data>Tables>TableS4_models.csv*** at the GitHub repository https://github.com/extavourlab/Oskar_Evolution.

Supplementary Table S5. *oskar* search results master table. This table summarizes all results of the *oskar* search performed on each dataset. Each row corresponds to a dataset. Columns are as follows: Id: the dataset NCBI identifier; Species: the organism's species name; Family_name: the organism's family name; Order_name: the organism's order name; Hits: the number of sequences in the dataset found that satisfy our criteria for *oskar* orthology; Source: the NCBI database from which this dataset was downloaded; Filtered_hits: the number of *oskar* sequins in remaining the dataset after the filtration process was applied to all identified *oskar* sequences. For more information on the criteria used for *oskar* orthology and the filtration process, please see the Materials and Methods "*Identification of oskar orthologs"*. This table is found in ***Data>Tables>TableS5_models.csv*** at the GitHub repository https://github.com/extavourlab/Oskar_Evolution.

549
550 **Supplementary Table S1**
551

Insect Order	Source	Number of datasets searched	Total hits	Filtered hits	% of datasets with *oskar* identified
Archaeognatha	GCA	1	0	0	0
Archaeognatha	TSA	2	0	0	0
Blattodea	GCA	3	1	1	33.33
Blattodea	GCF	2	0	0	0
Blattodea	TSA	51	7	7	13.73
Coleoptera	GCA	12	1	1	8.33
Coleoptera	GCF	9	3	2	22.22
Coleoptera	TSA	86	31	14	16.28
Collembola	TSA	9	0	0	0
Dermaptera	TSA	7	0	0	0
Diptera	GCA	115	63	60	52.17
Diptera	GCF	43	58	43	100
Diptera	TSA	162	72	58	35.8
Embioptera	TSA	5	0	0	0
Ephemeroptera	GCA	2	0	0	0
Ephemeroptera	TSA	5	1	1	20
Grylloblattodea	TSA	2	0	0	0
Hemiptera	GCA	18	0	0	0
Hemiptera	GCF	12	0	0	0
Hemiptera	TSA	192	1	0	0
Hymenoptera	GCA	52	32	30	57.69
Hymenoptera	GCF	47	36	32	68.09
Hymenoptera	TSA	301	157	128	42.52
Lepidoptera	GCA	80	0	0	0
Lepidoptera	GCF	17	0	0	0

Lepidoptera	TSA	135	24	4	2.96
Mantodea	TSA	13	0	0	0
Mantophasmatodea	TSA	2	0	0	0
Mecoptera	TSA	4	2	2	50
Megaloptera	TSA	3	0	0	0
Neuroptera	TSA	7	1	1	14.29
Odonata	GCA	2	0	0	0
Odonata	TSA	7	0	0	0
Orthoptera	GCA	3	0	0	0
Orthoptera	TSA	28	2	2	7.14
Phasmatodea	GCA	13	0	0	0
Phasmatodea	TSA	31	6	2	6.45
Phthiraptera	GCF	1	0	0	0
Phthiraptera	TSA	7	0	0	0
Plecoptera	GCA	3	0	0	0
Plecoptera	TSA	8	3	3	37.5
Psocoptera	TSA	23	5	5	21.74
Raphidioptera	TSA	3	0	0	0
Siphonaptera	GCF	1	0	0	0
Siphonaptera	TSA	4	0	0	0
Strepsiptera	GCA	1	0	0	0
Strepsiptera	TSA	2	0	0	0
Thysanoptera	GCA	1	1	1	100
Thysanoptera	GCF	1	1	1	100
Thysanoptera	TSA	11	10	8	72.73
Trichoptera	GCA	3	1	1	33.33
Trichoptera	TSA	7	2	2	28.57
Zoraptera	TSA	2	0	0	0
Zygentoma	TSA	4	1	1	25

Crustacea	TSA	168	0	0	0
Crustacea	GCF	1	0	0	0
Crustacea	GCA	11	0	0	0

552
553
554

Supplementary Table S2

	Genome parameter	(a) *oskar* Identified	(b) *oskar* not identified	ratio (a):(b)
# contigs	mean	255,015	43,280	5.89
	median	69,255	20,653	3.35
# scaffolds	mean	182,706	23,596	7.74
	median	40,960	9,398	4.36
contig N50 (bp)	mean	324,036	726,696	0.45
	median	14,052	40,079	0.35
scaffold N50	mean	2,636,825	5,695,299	0.46
	median	96,730	385,460	0.25
contig L50	mean	40,955	3,701	11.07
	median	6,868	1,300	5.28
scaffold L50	mean	27,269	1,500	18.18
	median	1,131	191	59.53
# contigs per genome length	mean	0.00060	0.00017	3.53
	median	0.00021	0.00009	2.33
# scaffolds per genome length	mean	0.00045	0.00009	5.00
	median	0.00013	0.00005	2.60