ARTICLE

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

Leo Blondel[1], Savandara Besse[1,2] 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

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

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)). This is soma-germ line differentiation, where only cells from the germ line will create the next generation (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, collectively 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* orthologs 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 germ cells (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* orthologs 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 orthologs 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 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* orthologs 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.

Results

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

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 orthologous sequences (Blondel, et al. 2020), we designed a new bioinformatics pipeline to scan and search for *oskar* orthologs 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* orthologs 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), and manually curated sequence duplicates and sequences that did not align correctly.

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* orthologs described to date. To determine if *oskar* orthologs 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* orthologs (Figure 3), strongly suggesting that *oskar* is restricted to the insect lineage (Lynch, et al. 2011; Ahuja and Extavour 2014).

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* orthologs (Supplementary Table S1 and Supplementary Figure S1a). Given that detection of putative orthologs 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* orthologs. 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* ortholog (Mann Whitney U test p<0.05; Supplementary Figure S2; Supplementary Table S2).

oskar *predates the divergence of Ametabola and other insects*

We identified *oskar* orthologs in 15 of the 29 generally recognized (Misof, et al. 2014) insect orders, including eight holometabolous orders, six hemimetabolous orders, and one ametabolous order (Figure 3). This result is consistent with our previous proposals

198 that *oskar* predates the origins of the Holometabola (Ewen-Campen, et al. 2012; Blondel,
199 et al. 2020). The novel finding of an *oskar* ortholog from the silverfish *Atelura formicaria*
200 (Zygentoma) allows us to date back the origin of *oskar* further than previous analyses, to
201 at least 420 million years ago (Misof, et al. 2014), before the divergence of Ametabola
202 from the remaining insect lineages.
203
204 We then explored the distribution of *oskar* sequences across insect phylogeny.
205 Interestingly, we identified multiple lineages where *oskar* appeared to have been lost
206 independently, including confirming the previously reported (Lynch, et al. 2011) losses
207 from the genomes of the red flour beetle *Tribolium castaneum*, the honeybee *Apis*
208 *mellifera*, and the silk moth *Bombyx mori* (Figure 3). Notably, within Lepidoptera we
209 identified *oskar* orthologs in only four species, despite the fact that we searched 232
210 available lepidopteran sequence datasets, including 17 well-annotated RefSeq genomes
211 and 135 transcriptomes (Figure 3 and Supplementary Figure S3). In principle, this
212 apparent widespread absence of *oskar* in Lepidoptera could be due to unusually rapid
213 evolution of the *oskar* sequence in this lineage, which might render lepidopteran *oskar*
214 orthologs undetectable by our methods. However, we note that the only four lepidopteran
215 orthologs we detected all belonged to species of the basally branching *Adelidae* and
216 *Palaephatidae* families. We therefore favor the interpretation that *oskar* was lost from a
217 last common ancestor of *Meessiidae* and *Palaphaetidae*, approximately 180 million years
218 ago, with the consequence that the majority of extant lepidopteran lineages lack an *oskar*
219 ortholog (Supplementary Figure S3) (Mitter, et al. 2017; Kawahara, et al. 2019).
220
221 The Hemiptera also appear to have lost *oskar*, based on our analysis of the 222 datasets
222 available for this clade, including 12 RefSeq genomes and 192 transcriptomes. However,
223 we did identify an *oskar* ortholog in the Thysanoptera, which is a hemipteran sister group
224 (Misof, et al. 2014). Finally, we identified *oskar* orthologs in only four of the 11 orders of
225 the Polyneoptera for which data were available. With the exception of Mantodea (13
226 transcriptomes), the four orders with detectable *oskar* sequences all had more than ten
227 available sequence datasets (Plecoptera: three genomes and eight transcriptomes;
228 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). 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 orthology 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). 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 S5). In terms of developmental stage, only holometabolous insects appeared to express *oskar* during embryonic, larval or nymphal stages; for all other 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

289 *oskar* is expressed and required in the embryonic nervous system of a cricket, a
290 hemimetabolous insect (Ewen-Campen, et al. 2012).
291

292 *The Long Oskar domain is an evolutionary novelty specific to a subset of Diptera*
293 *D. melanogaster* has two isoforms of Oskar (Markussen, et al. 1995): Short Oskar,
294 containing the LOTUS, OSK and interdomain regions, and Long Oskar, containing all
295 three domains of Short Oskar as well as an additional 5' domain (Supplementary Figure
296 S7). It was previously reported that Long Oskar was absent from *N. vitripennis*, *C. pipiens*
297 and *G. bimaculatus* (Lynch, et al. 2011; Ewen-Campen, et al. 2012), and within our
298 alignment of Oskar sequences we could only detect the Long Oskar isoform within
299 Diptera. Therefore, using our dataset, we asked when these two isoforms had evolved.
300 We selected the dipteran sequences from our Oskar alignment and then grouped the
301 sequences by family. We plotted the amino acid occupancy at each alignment position
302 (Supplementary Figure S7), and found that Long Oskar predates the Drosophilids, being
303 identified as early as the *Pinpunculidae* (Supplementary Figure S7). Moreover, following
304 the evolution of the Long Oskar isoform, the Long Oskar domain was retained in all
305 families except for the *Glossinidae* and *Scathophagidae*. However, given that we
306 identified only eight and two Oskar sequences for these families respectively, we cannot
307 eliminate the possibility that apparent absence of the Long Oskar domain in these groups
308 reflects our small sample size, rather than true evolutionary loss.

309

310 *The LOTUS and OSK domains evolved differently between hemimetabolous and*
311 *holometabolous insects*
312 The fact that an *oskar*-dependent germ plasm mode of germ line specification mechanism
313 has been identified only in holometabolous insects suggests that *oskar* may have been
314 co-opted in this clade for this function (Ewen-Campen, et al. 2012). Under this hypothesis,
315 evolution of the *oskar* sequence in the lineage leading to the Holometabola may have
316 changed the physico-chemical properties of Oskar protein, such that it acquired germ
317 plasm nucleation abilities in these insects. To test this hypothesis, we asked whether
318 there were particular sequence features associated with Oskar proteins from

holometabolous insects, in which Oskar can assemble germ plasm, and hemimetabolous insects, which lack 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 to the highly unbalanced availability of genomic and transcriptomic data between hemimetabolous and holometabolous sequences (Supplementary Table S1; Figure 3) the choice of a weighted score was necessary to avoid biasing the results towards insect orders such as Diptera or Hymenoptera. To study the difference between hemimetabolous and holometabolous sequences, we did not use the Valdar score directly, but instead computed the conservation ratio between both groups for each position, which we call the Conservation bias (See Methods: Computation of the Conservation Bias). We plotted the conservation bias on the solved three-dimensional crystal structure of the *D. melanogaster* LOTUS and OSK domains (Jeske, et al. 2015; Yang, et al. 2015) to ask whether specific functionally relevant structures showed phylogenetic or other patterns of residue conservation (Figure 5).

First, we asked if the overall conservation score of the domains was different between holometabolous and hemimetabolous sequences. We observed that the conservation bias for the LOTUS domain was centered around a mean of 1.00, indicating that both Holometabola and Hemimetabola displayed a similar conservation of the LOTUS domain (Figure 5a). For the OSK domain however, the conservation bias was centered around 0.84, indicating that the hemimetabolous sequences displayed a higher level of conservation compared to holometabolous sequences (Figure 5a). We then looked at the conservation bias scores *in-situ* on the LOTUS domain structure. We asked if the amino acids of the β sheets of the LOTUS domain thought to be involved in dimerization of the

protein (Jeske, et al. 2015; Yang, et al. 2015) displayed conservation bias. Both β sheets had an overall even bias (mean: 1.03 and 1.05 for β1 and β2 respectively) between both groups (Figure 5b). Second, as we had observed that hemimetabolous OSK was more conserved overall than holometabolous OSK, we asked if there were any clear patterns of conservation bias in specific regions of the structure (Figure 5a and b). We found that some of the secondary structures within OSK showed a differential conservation (α2: 0.54, α6: 0.42, β2: 0.52), whereas other structures were within less than 0.1 of the median value for OSK. Moreover, we observed a large pocket of amino acids showing a conservation bias towards hemimetabolous sequences located on the surface of OSK (Figure 5c). This particular area contains the previously reported important amino acids for the RNA binding function of OSK (Jeske, et al. 2015; Yang, et al. 2015) namely, R442, R436 and R576. The electrostatic properties at those positions were conserved in the holometabolous sequences R436: 0.36, R442: 0.29 and R576: 0.81 (Figure 5d), but not in hemimetabolous sequences.

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 S7). Interestingly, we found that most of the variance in sequence space was due to dipterans and hymenopterans (Supplementary Figure S7). When we considered the OSK domain only, we identified clusters of *Drosophilidae*, *Culicidae* and *Formicidae* sequences (Supplementary Figure S7). 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 S7). 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 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 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 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 changes in the conservation of the residues of this interface might be detectable between Holometabola and Hemimetabola. 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 reported Vasa interacting amino acids A162 and L228 of the LOTUS domain 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 likely to be important interaction partners due to their high conservation (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).

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 orthologs 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 orthologs, with the exception of the α1 helix (Supplementary Figure S8) 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 across all insects, similar to the LOTUS domain (Valdar score: 0.51) (Figure 7a). However, the conservation pattern is higher in the core amino acids (Valdar score average of core amino acid: 0.54) when compared to the residues at the surface (Valdar score average for surface amino acid: 0.23) (Figure 7a). Despite the overall low conservation of the residues at the surface of the OSK domain, we found that the electrostatic properties are conserved overall (electrostatic conservation score >0; conserved) in the previously reported putative RNA binding pocket (Yang, et al. 2015). However, as previously mentioned, this conservation is stronger in holometabolous

sequences (Figure 5d). These results are in accordance with the potential role of OSK as an RNA Binding domain in the context of germ plasm assembly (Jeske, et al. 2015; Yang, et al. 2015). We also submitted the OSK sequences to the same secondary structure analysis performed on LOTUS. We found that, as for the LOTUS domain, the secondary structure of OSK is highly conserved throughout all insect sequences analyzed (Supplementary Figure S8).

We then asked if the conservation patterns observed at the core of OSK were clustered in sequence motifs. When we looked at the location of the highly conserved amino acids, we found that the conservation was driven by four well-defined sequence motifs (Figure 7c, c', c'', c'''). Given that *oskar* plays different roles in Holometabola and Hemimetabola, we asked whether the conserved OSK motifs showed any difference in conservation between these two groups. Of the four highly conserved OSK core motifs (Figure 7c, c', c'', c'''), two of them (Figure 7c: Valdar average score: 0.80 and c'' Valdar average score: 0.71) were conserved across all insects, but the other two showed differential conservation between the holometabolous and hemimetabolous sequences (Figure 7c': Valdar score average Holometabola: 0.78; Hemimetabola: 0.58 c'':' Valdar score average Holometabola: 0.70, Hemimetabola: 0.55). Finally, we noted that only one of the affected OSK domain residues in known loss of function *oskar* alleles affecting posterior patterning in *D. melanogaster*, S457, is conserved across all insects (Valdar score: 0.86). This suggests that the role of the other previously reported important amino acids in the function of *D. melanogaster* OSK (Yang, et al. 2015) might not be conserved in other insects (red positions in Figure 7c, c', c'', c''').

Discussion

An expanded collection of oskar orthologs

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 orthologous *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* orthologs (*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 authors have speculated that a germ plasm, located ventrally rather than posteriorly, may specify germ cells in this moth (Toshiki, et al. 2000; Nakao, et al. 2008). However, recent knockdown experiments showed that maternal *nanos-O* is dispensable for germ cell formation (Nakao and Takasu 2019), consistent with a zygotic, inductive mechanism. In the butterfly *Pararge argeria* (Nymphalidae), no *oskar* ortholog has been identified in the genome (Carter, et al. 2013), but the transcripts of one of four identified

nanos orthologs (*nanos-O*) have been detected in a small region of ventral cortical ooplasm, again prompting speculation that this lepidopteran may also deploy a germ plasm (Carter, et al. 2015). We suggest that if these or other Lepidoptera do indeed rely on germ plasm to specify their germ line, they may do so using a germ plasm nucleator other than Oskar. For most studied Lepidoptera, however, classical embryological studies report the first appearance of primordial germ cells at post-blastoderm stages, either from the ventral midline of the cellular blastoderm or early germ band (Woodworth 1889; Tomaya 1902; Sehl 1931; Miya 1953, 1958, 1975; Tanaka 1987), from the coelomic sac mesoderm of the abdomen (Johannsen 1929; Eastham 1930; Saito 1937; Presser and Rutschky 1957; Kobayashi and Ando 1984), or from the primary ectoderm of the caudal germ band (Schwangart 1905; Lautenschlager 1932; Ando and Tanaka 1979; Tanaka 1987; Guelin 1994) (Figure S6). Taken together, these data suggest that an inductive mechanism may operate to specify germ cells in most moths and butterflies. We speculate that the loss of *oskar* from most lepidopteran genomes may have facilitated or necessitated secondary reversion to the hypothesized ancestral inductive mechanism for germ line specification.

Another order with apparent near-total absence of *oskar* orthologs is the Hemiptera (true bugs), whose sister group Thysanoptera (thrips) nevertheless possesses *oskar*. This secondary loss of *oskar* from a last common hemipteran ancestor correlates with the reported post-blastoderm appearance of primordial germ cells in the embryo. Classical studies on most hemipteran species describe germ cell formation as occurring after cellular blastoderm formation, on the inner (yolk-facing) side of the posterior blastoderm surface (Metschnikoff 1866; Witlaczil 1884; Will 1888; Mellanby 1935; Butt 1949; Kelly and Huebner 1989; Heming and Huebner 1994). A notable exception to this is the parthenogenetic pea aphid *Acyrthosiphon pisum*, for which strong gene expression and morphological evidence supports a germ plasm-driven germ cell specification mechanism in both sexual and asexual modes (Miura, et al. 2003; Chang, et al. 2006; Lin, et al. 2014). In contrast, studies of the aphids *Aphis plantoides, A. rosea* 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*

562 and three *Aphis* species, appear to lack *oskar*. This suggests that germ plasm assembly
563 in *A. pisum* either does not require a nucleator molecule or uses a novel non-Oskar
564 nucleator.

565 In the Hymenoptera (ants, bees, wasps and sawflies), our results strongly suggest that
566 *oskar* was lost from the genome of the last common ancestor of bees and spheroid wasps
567 (Supplementary Figure S9). Our analysis further suggests multiple additional
568 independent losses in as many as 25 other hymenopteran lineages, including some for
569 which good quality RefSeq genomes were available (e.g. the slender twig ant
570 *Pseudomyrmex gracilis* or the wheat stem sawfly *Cephus cinctus* (Supplementary Figure
571 S9). However, it would be premature to draw strong conclusions about the number of
572 independent losses given the predominance of transcriptome data in the Hymenoptera.

573 In addition to convergent losses of *oskar*, we also found evidence for clade-specific
574 duplications of *oskar* in the Hymenoptera. Seven of the nine families containing these
575 putative duplications are families of parasitoid wasps; the remaining two families are ants
576 (Formicidae) and the group of yellowjackets, hornets, and paper wasps (Vespidae)
577 (Figure 4). The phylogenetic relationships of these groups make it highly unlikely that a
578 duplication occurred only once in their last common ancestor, which would be the last
579 common ancestor of all wasps, bees and ants (i.e. Apocrita, all hymenopterans except
580 sawflies) (Supplementary Figure S9). We suggest that the most parsimonious hypothesis
581 is one of three to five independent duplications of *oskar*, followed by at least nine to 14
582 independent reversions to a single copy, or total loss of the locus (Supplementary Figure
583 9).

584 No notable life history characteristics appear to unite those species with multiple *oskar*
585 orthologs: they include eusocial and solitary, sting-bearing and stingless, parasitoid and
586 non-parasitic insects. To our knowledge, neither is there anything unique about the germ
587 line specification process in Hymenoptera with one or more than one *oskar* ortholog. Most
588 Hymenoptera appear to use a germ plasm-driven mechanism to specify germ cells in
589 early blastoderm stage embryos (Supplementary Figure S9 and references therein), and
590 we identified *oskar* orthologs for all such species described in the embryological literature
591 (Supplementary Figure S9). In the notable example of the honeybee *Apis mellifera*, in

592 which cytological and molecular evidence suggests germ cell arise from abdominal
593 mesoderm (Bütschli 1870; Nelson 1915; Fleig and Sander 1985, 1986; Zissler 1992;
594 Gutzeit, et al. 1993; Dearden 2006), we identified no *oskar* ortholog in its well-annotated
595 genome (Supplementary Figure S9), as noted previously by other authors (Lynch, et al.
596 2011). However, no major differences in germ plasm or pole cell formation have been
597 reported in species or families of ants or wasps with duplicated *oskar* loci, compared with
598 close relatives that possess *oskar* in single copy (e.g. compare the ants *Solenopsis invicta*
599 (at least 2 *oskars*) and *Aphaenogaster rudis* (1 *oskar*) (Khila and Abouheif 2008), or the
600 pteromalid wasps *Nasonia vitripennis* (1 *oskar*) (Lynch and Desplan 2010; Lynch, et al.
601 2011; Quan, et al. 2019) and *Otitesella tsamvi* (2 *oskars*). Thus, future studies that
602 independently abrogate the functions of each paralog individually, will be needed to
603 determine the biological significance, if any, of these *oskar* duplications.

604

605 *Functional implications of differential conservation of the LOTUS and OSK domains*
606 We have identified novel conserved amino acid positions that we hypothesize are
607 important for the Vasa binding properties of the LOTUS domain and the RNA properties
608 binding of the OSK domain (Figure 6 and 7). Our observation of the conservation of the
609 LOTUS domain α2 helix is consistent with its previously reported importance LOTUS-
610 Vasa binding (Jeske, Müller, and Ephrussi 2017). In the α2 helix, we also observed high
611 conservation of H227 and Q235. The positions of these residues suggest they may
612 contribute to the interaction between Vasa and LOTUS. We suggest they should therefore
613 be the target of future mutational studies.
614
615 We also uncovered an interesting new conservation pattern within the OSK domain. The
616 conserved amino acids were more abundant in the core of the domain than on the
617 surface. This differential conservation might be relevant to the acquisition of a germ plasm
618 nucleator role of *oskar* in the Holometabla (Figure 5). We noted that the basic properties
619 of surface residues previously reported for *D. melanogaster* (Yang, et al. 2015) are
620 conserved across insects, which might indicate that the RNA binding properties of OSK
621 observed in *D. melanogaster* (Jeske, et al. 2015; Yang, et al. 2015) are also conserved

throughout holometabolous insects. We speculate that the comparatively low amino acid conservation of the surface residues in Holometabolous OSK domains, which nevertheless display highly conserved basic properties, could have allowed greater flexibility in the co-evolution of specific RNA binding partners for the OSK domains of different lineages.

OSK evolved differentially between holometabolous and hemimetabolous insects

Finally, we observed a differential conservation of the OSK domain between hemimetabolous and holometabolous insects. Specifically, we found that the OSK sequence was less conserved across the Holometabola than across the Hemimetabola. This observation raises two potential hypotheses regarding the role of the OSK domain in the functional evolution of Oskar. First, perhaps the apparently relaxed purifying selection experienced by OSK in the Holometabola was necessary for the co-option of *oskar* to a germ plasm nucleation role. Second, Oskar might have a function in the hemimetabolous insects that requires strong conservation of OSK. More studies on the roles and biochemical properties of OSK in hemimetabolous insects will be 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 animal model, nor any cell culture lines. However, it 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 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* orthologs 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 found in the following two files: ***genome_crustacean_database.csv*** and ***transcriptome_crustacean_database.csv***.

Creation of protein sequence databases

The classical approach for orthology 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 orthologs

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* orthology as in our previous work: a sequence possessing a LOTUS domain followed by an interdomain region, and then an OSK domain (Blondel, et al. 2020). To maximize the number of potential orthologs, 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* orthologs 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* orthologs in the available sequence datasets. We called this final set the **filtered set** of sequences and used it in all subsequent orthology 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***).

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 orthology 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* orthologs (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***).

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 *orthologs* 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***).

Calculation of Oskar conservation scores

Using the large set of orthologous 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***).

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($Valdar_{holo}$) / log($Valdar_{hemi}$) 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 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* orthologs, 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 full-length Oskar protein structure had been reported. With the caveat that all visualization was done on the structure of the *Drosophila melanogaster* protein domains, we created a custom python script that augments PyMOL with automatic display and coloring capacities. This script is available as ***Oskar_pymol_visualization.py***, and contains a manual at the beginning of the file. For the OSK domain, we used the structure PDBID: 5A4A, and for the LOTUS domain, PDBID: 5NT7 (Jeske, et al. 2015; Jeske, et al. 2017). The LOTUS structure we used is in complex with Vasa, and in a dimeric form (Jeske, et al. 2017), allowing for easy interpretation of the different conservation scores. For the OSK structure, we removed the residues 399-401 and 604-606 from the PDB file as those amino acids did not align across all sequences and therefore showed highly biased conservation scores.

Statistical analysis

All statistical analyses were performed using the scipy stats module (https://www.scipy.org/). Significance thresholds for p-values were set at 0.05. Statistical tests and p-values are reported in the figure legends. All statistical tests can be found in the ipython notebooks mentioned below.

Data and code availability

The study generated a series of python 3 script and python 3 ipython notebook files that perform the entire analysis. All the results presented in this paper can be reproduced by running the aforementioned python 3 code. The primary data, *oskar* orthologs, Oskar alignments, trees, and conservation statistics as well as the code created and used are available as supplementary information. For ease of access, legibility, and reproducibility, the code and datasets have been deposited in a GitHub repository available at https://github.com/extavourlab/Oskar_Evolution.

Software and libraries

All software and libraries used in this study are published under open source libre licenses and are therefore available to any researcher.

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

This work was supported by funds from Harvard University, and support to SB from the Master's in Bioinformatics Program of the University of Bordeaux. We thank members of the Extavour lab for discussion.

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* ortholog 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* ortholog 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* ortholog was identified in this order. Color (unique for each order): 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. (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 ortholog. 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 **(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)).

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.

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

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

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.

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:2179-2188.

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 1:1-23.

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.

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

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

1096 Carter J-M, Baker SC, Pink R, Carter DRF, Collins A, Tomlin J, Gibbs M, Breuker CJ. 2013.
1097 Unscrambling butterfly oogenesis. BioMedCentral Genomics 14:283-283.

1098 Carter JM, Gibbs M, Breuker CJ. 2015. Divergent RNA Localisation Patterns of Maternal Genes
1099 Regulating Embryonic Patterning in the Butterfly Pararge aegeria. PLoS ONE 10:e0144471.

1100 Chang CC, Lee WC, Cook CE, Lin GW, Chang T. 2006. Germ-plasm specification and germline
1101 development in the parthenogenetic pea aphid *Acyrthosiphon pisum*: Vasa and Nanos as
1102 markers. International Journal of Developmental Biology 50:413-421.

1103 Chen X-x, Achterberg Cv. 2018. Systematics, Phylogeny, and Evolution of Braconid Wasps: 30
1104 Years of Progress. Annual Review of Entomology 64:1-24.

1105 Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, Kaufman TC, Kellis M,
1106 Gelbart W, Iyer VN, et al. (p04187 co-authors). 2007. Evolution of genes and genomes on the
1107 *Drosophila* phylogeny. Nature 450:203-218.

1108 Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence logo generator.
1109 Genome Research 14:1188-1190.

1110 Dearden PK. 2006. Germ cell development in the Honeybee (*Apis mellifera*); *vasa* and *nanos*
1111 expression. BMC Developmental Biology 6:6.

1112 Dearden PK, Wilson MJ, Sablan L, Osborne PW, Havler M, McNaughton E, Kimura K, Milshina
1113 NV, Hasselmann M, Gempe T, et al. 2006. Patterns of conservation and change in honey bee
1114 developmental genes. Genome Research 16:1376-1384.

1115 DeLano WL. 2002. Pymol: An Open-Source Molecular Graphics Tool. CCP4 Newsletter on
1116 Protein Crystallography 40:82-92.

1117 Donnell DM, Corley LS, Chen G, Strand MR. 2004. Caste determination in a polyembryonic
1118 wasp involves inheritance of germ cells. Proceedings of the National Academy of Sciences of
1119 the United States of America 101:10095-10100.

1120 Drozdetskiy A, Cole C, Procter J, Barton GJ. 2015. JPred4: a protein secondary structure
1121 prediction server. Nucleic Acids Research 43:W389-394.

1122 Eastham LES. 1930. The embryology of *Pieris rapae* - Organogeny. Philosophical Transactions
1123 of the Royal Society of London. Series B: Biological Sciences 219:1-50.

1124 Edgar RC. (r20416 co-authors). 2004. MUSCLE: a multiple sequence alignment method with
1125 reduced time and space complexity. BMC Bioinformatics 5:113.

1126 Ephrussi A, Lehmann R. 1992. Induction of germ cell formation by *oskar*. Nature 358:387-392.

1127 Ewen-Campen B, Schwager EE, Extavour CG. 2010. The molecular machinery of germ line
1128 specification. Molecular Reproduction and Development 77:3-18.

1129 Ewen-Campen B, Srouji JR, Schwager EE, Extavour CG. 2012. *oskar* Predates the Evolution of
1130 Germ Plasm in Insects. Current Biology 22:2278-2283.

1131 Extavour CG, Akam ME. 2003. Mechanisms of germ cell specification across the metazoans:
1132 epigenesis and preformation. Development 130:5869-5884.

1133 Field J, Ohl M, Kennedy M. 2011. A molecular phylogeny for digger wasps in the tribe
1134 Ammophilini (Hymenoptera, Apoidea, Sphecidae). Systematic Entomology 36:732-740.

1135 Fleig R, Sander K. 1985. Blastoderm development in honey bee embryogenesis as seen in the
1136 scanning electron microscope. International Journal of Invertebrate Reproduction and
1137 Development 8:279-286.

1138 Fleig R, Sander K. 1986. Embryogenesis of the Honeybee *Apis mellifera* L (Hymenoptera,
1139 Apidae) - an SEM Study. International Journal of Insect Morphology and Embryology 15:449-
1140 462.

1141 Fleischmann VG. 1975. Origin and embryonic development of fertile gonads with and without
1142 pole cells of *Pimpla turionellae* L. (Hymenoptera, Ichneumonidae). Zool. Jb. Anat. Bd. 94:375-
1143 411.

1144 Gatenby JB. 1920. The Cytoplasmic Inclusions of the Germ Cells. Part VI. On the origin and
1145 probable constitution of the germ-cell determinant of *Apanteles glomeratus*, with a note on the
1146 secondary nuclei. Quarterly Journal of Microscopical Science 64:133-153.

1147 Gatenby JB. 1917a. The embryonic development of *Trichogramma evanescens* Westw.,
1148 monoembryonic egg parasite of *Donacia simplex*. Quarterly Journal of Microscopical Science
1149 62:149-187.

1150 Gatenby JB. 1918. The segregation of germ cells in *Trichogramma evanescens*. Quarterly
1151 Journal of Microscopical Science 63:161-173.

1152 Gatenby JB. 1917b. The segregation of the germ-cells in *Trichogramma evanescens*. Quarterly
1153 Journal of Microscopical Science 62:149-187.

1154 Grbic' M. 2000. "Alien" wasps and evolution of development. Bioessays 22:920-932.

1155 Grbic' M. 2003. Polyembryony in parasitic wasps: evolution of a novel mode of development.
1156 International Journal of Developmental Biology 47:633-642.

1157 Grbic' M, Nagy LM, Carroll SB, Strand M. 1996. Polyembryonic development: insect pattern
1158 formation in a cellularised environment. Development:795-804.

1159 Guelin M. 1994. [Activity of W-sex heterochromatin and accumulation of the nuage in nurse
1160 cells of the 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
1163 Archives of Developmental Biology 202:181-191.

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

1166 Hay B, Jan LY, Jan YN. 1990. Localization of *vasa*, a component of *Drosophila* polar granules,
1167 in maternal-effect mutants that alter embryonic anteroposterior polarity. Development 109:425-
1168 433.

1169 Hegner RW. 1914. Studies on germ cells. III. The origin of the Keimbahn-determinants in a
1170 parasitic Hymenopteran, *Copidosoma*. Anatomischer Anzeiger 3-4:51-69.

1171 Heming BS, Huebner E. 1994. Development of the Germ Cells and Reproductive Primordia in
1172 Male and Female Embryos of *Rhodnius prolixus* Stal (Hemiptera, Reduviidae). Canadian
1173 Journal of Zoology 72:1100-1119.

1174 Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: Reconstruction, Analysis, and Visualization of
1175 Phylogenomic Data. Molecular Biology and Evolution 33:1635-1638.

1176 Jeske M, Bordi M, Glatt S, Muller S, Rybin V, Muller CW, Ephrussi A. 2015. The Crystal
1177 Structure of the *Drosophila* Germline Inducer Oskar Identifies Two Domains with Distinct Vasa
1178 Helicase- and RNA-Binding Activities. Cell Reports 12:587-598.

1179 Jeske M, Muller CW, Ephrussi A. 2017. The LOTUS domain is a conserved DEAD-box RNA
1180 helicase regulator essential for the recruitment of Vasa to the germ plasm and nuage. Genes
1181 and Development 31:939-952.

1182 Johannsen OA. 1929. Some phases in the embryonic development of *Diacrisia virginica* Fabr.
1183 (Lepidoptera). J. Morphol. Physiol. 2:493-541.

1184 Jones JR, Macdonald PM. 2007. Oskar controls morphology of polar granules and nuclear
1185 bodies in *Drosophila*. Development 134:233-236.

1186 Juhn J, James AA. 2006. *oskar* gene expression in the vector mosquitoes, *Anopheles gambiae*
1187 and *Aedes aegypti*. Insect Molecular Biology 15:363-372.

1188 Juhn J, Marinotti O, Calvo E, James AA. 2008. Gene structure and expression of *nanos* (*nos*)
1189 and *oskar* (*osk*) orthologues of the vector mosquito, *Culex quinquefasciatus*. Insect Molecular
1190 Biology 17:545-552.

1191 Kawahara AY, Plotkin D, Espeland M, Meusemann K, Toussaint EFA, Donath A, Gimnich F,
1192 Frandsen PB, Zwick A, Dos Reis M, et al. 2019. Phylogenomics reveals the evolutionary timing
1193 and pattern of butterflies and moths. Proceedings of the National Academy of Sciences of the
1194 United States of America 116:22657-22663.

1195 Kelly GM, Huebner E. 1989. Embryonic development of the hemipteran insect *Rhodnius*
1196 *prolixus*. Journal of Morphology 199:175-196.

1197 Khila A, Abouheif E. 2008. Reproductive constraint is a developmental mechanism that
1198 maintains social harmony in advanced ant societies. Proceedings of the National Academy of
1199 Sciences of the United States of America 105:17884-17889.

1200 Kim-Ha J, Smith JL, Macdonald PM. 1991. *oskar* mRNA is localized to the posterior pole of the
1201 *Drosophila* oocyte. Cell 66:23-35.

1202 Kirk DL. 2005. A twelve-step program for evolving multicellularity and a division of labor.
1203 Bioessays 27:299-310.

1204 Kobayashi Y, Ando H. 1984. Mesodermal Organogenesis in the Embryo of the Primitive Moth,
1205 *Neomicropteryx nipponensis* Issiki (Lepidoptera, Micropterygidae). Journal of Morphology
1206 181:29-47.

1207 Koscielska MK, Koscielski B. 1987. Early embryonic development of *Tritneptis diprionis*
1208 (Chalcidoidea, Hymenoptera). In: Ando H, Jura C, editors. Recent Advances in Insect
1209 Embryology in Japan and Poland. Tsukuba: Arthropod. Embryol. Soc. Jpn.

1210 ISEBU Co. Ltd. p. 207-214.

1211 Lasko P. 2013. The DEAD-box helicase Vasa: evidence for a multiplicity of functions in RNA
1212 processes and developmental biology. Biochimica et Biophysica Acta 1829:810-816.

1213 Lautenschlager F. 1932. Die Embryonalentwicklung der weiblichen Keimdruse bei der Psychide
1214 *Solenobia triquetella*. Zool. Jarh. 56:121-162.

1215 Lebart L, Morineau A, Warwick KM. 1984. Multivariate Descriptive Statistical Analysis:
1216 Correspondence Analysis and Related Techniques for Large Matrices. Chichester, UK: John
1217 Wiley & Sons.

1218 Lehmann R. 2016. Germ Plasm Biogenesis--An Oskar-Centric Perspective. Current Topics in
1219 Developmental Biology 116:679-707.

1220 Lehmann R, Nüsslein-Volhard C. 1986. Abdominal Segmentation, Pole Cell Formation, and
1221 Embryonic Polarity Require the Localized Activity of *oskar*, a Maternal Gene in *Drosophila*. Cell
1222 47:144-152.

1223 Lin GW, Cook CE, Miura T, Chang CC. 2014. Posterior localization of ApVas1 positions the
1224 preformed germ plasm in the sexual oviparous pea aphid Acyrthosiphon pisum. EvoDevo 5:18.

1225 Lin J. 1991. Divergence Measures Based on the Shannon Entropy. IEEE Transactions on
1226 Information Theory / Professional Technical Group on Information Theory 37:145-151.

1227 Loytynoja A. 2014. Phylogeny-aware alignment with PRANK. Methods in Molecular Biology
1228 1079:155-170.

1229 Lynch JA, Desplan C. (p16123 co-authors). 2010. Novel modes of localization and function of
1230 *nanos* in the wasp *Nasonia*. Development 137:3813-3821.

1231 Lynch JA, Özüak O, Khila A, Abouheif E, Desplan C, Roth S. 2011. The Phylogenetic Origin of
1232 *oskar* Coincided with the Origin of Maternally Provisioned Germ Plasm and Pole Cells at the
1233 Base of the Holometabola. PLoS Genetics 7:e1002029.

1234 Maddison DR, Schultz K-S, Maddison WP. 2007. The Tree of Life Web Project. Zootaxa
1235 1668:19-40.

1236 Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, Basutkar P, Tivey ARN, Potter
1237 SC, Finn RD, et al. 2019. The EMBL-EBI search and sequence analysis tools APIs in 2019.
1238 Nucleic Acids Research 47:W636-W641.

1239 Malm T, Nyman T. 2015. Phylogeny of the symphytan grade of Hymenoptera: new pieces into
1240 the old jigsaw(fly) puzzle. Cladistics 31:1-17.

1241 Markussen FH, Michon AM, Breitwieser W, Ephrussi A. 1995. Translational control of *oskar*
1242 generates short OSK, the isoform that induces pole plasm assembly. Development 121:3723-
1243 3732.

1244 Mellanby H. 1935. The early embryonic development of *Rhodnius prolixus* (Hemiptera,
1245 Heteroptera). Quarterly Journal of Microscopical Science 78:71-90.

1246 Metschnikoff E. 1866. Embryologische Studien an Insekten. Zeit. f. wiss Zool. 16:389-500.

1247 Misof B, Liu S, Meusemann K, Peters RS, Donath A, Mayer C, Frandsen PB, Ware J, Flouri T,
1248 Beutel RG, et al. 2014. Phylogenomics resolves the timing and pattern of insect evolution.
1249 Science 346:763-767.

1250 Mitter C, Davis DR, Cummings MP. 2017. Phylogeny and Evolution of Lepidoptera. Annual
1251 Review of Entomology 62:265-283.

1252 Miura T, Braendle C, Shingleton A, Sisk G, Kambhampati S, Stern DL. 2003. A comparison of
1253 parthenogenetic and sexual embryogenesis of the pea aphid *Acyrthosiphon pisum* (Hemiptera :
1254 Aphidoidea). Journal of Experimental Zoology Part B-Molecular and Developmental Evolution
1255 295B:59-81.

1256 Miya K. 1953. The presumptive genital region at the blastoderm stage of the silkworm egg.
1257 Journal of the Faculty of Agriculture of Iwate University:223-227.

1258 Miya K. 1958. Studies on the embryonic development of the gonad in the silkworm, *Bombyx*
1259 *mori* L. Part I. Differentiation of germ cells. Journal of the Faculty of Agriculture of Iwate
1260 University 3:436-467.

1261 Miya K. 1975. Ultrastructural changes of embryonic cells during organogenesis in the silkworm,
1262 *Bombyx mori*. I. The Gonad. Journal of the Faculty of Agriculture of Iwate University 12:329-
1263 338.

1264 Moon CP, Fleming KG. 2011. Side-chain hydrophobicity scale derived from transmembrane
1265 protein folding into lipid bilayers. Proceedings of the National Academy of Sciences of the
1266 United States of America 108:10174-10177.

1267 Nagy L, Riddiford L, Kiguchi K. 1994. Morphogenesis in the Early Embryo of the Lepidopteran
1268 *Bobyx mori*. Developmental Biology:137-151.

1269 Nakao H. 1999. Isolation and characterization of a *Bombyx vasa*-like gene. Development Genes
1270 and Evolution 209:312-316.

1271 Nakao H, Hatakeyama M, Lee JM, Shimoda M, Kanda T. 2006. Expression pattern of *Bombyx*
1272 vasa-like (BmVLG) protein and its implications in germ cell development. Development Genes
1273 and Evolution 216:94-99.

1274 Nakao H, Matsumoto T, Oba Y, Niimi T, Yaginuma T. 2008. Germ cell specification and early
1275 embryonic patterning in Bombyx mori as revealed by nanos orthologues. Evolution and
1276 Development 10:546-554.

1277 Nakao H, Takasu Y. 2019. Complexities in Bombyx germ cell formation process revealed by
1278 Bm-nosO (a Bombyx homolog of nanos) knockout. Developmental Biology 445:29-36.

1279 Nelson JA. 1915. The embryology of the honey bee. Princeton: Princeton University Press.

1280 Noce T, Okamoto-Ito S, Tsunekawa N. 2001. *Vasa* homolog genes in mammalian germ cell
1281 development. Cell Structure and Function 26:131-136.

1282 Nyman T, Zinovjev AG, Vikberg V, Farrell BD. 2006. Molecular phylogeny of the sawfly
1283 subfamily Nematinae (Hymenoptera: Tenthredinidae). Systematic Entomology 31:569-583.

1284 Pal D, Chakrabarti P. 2001. Non-hydrogen bond interactions involving the methionine sulfur
1285 atom. Journal of Biomolecular Structure and Dynamics 19:115-128.

1286 Peters RS, Krogmann L, Mayer C, Donath A, Gunkel S, Meusemann K, Kozlov A,
1287 Podsiadlowski L, Petersen M, Lanfear R, et al. 2017. Evolutionary History of the Hymenoptera.
1288 Current Biology 27:1013-1018.

1289 Peters RS, Niehuis O, Gunkel S, Bläser M, Mayer C, Podsiadlowski L, Kozlov A, Donath A,
1290 Noort Sv, Liu S, et al. 2018. Transcriptome sequence-based phylogeny of chalcidoid wasps
1291 (Hymenoptera: Chalcidoidea) reveals a history of rapid radiations, convergence, and
1292 evolutionary success. Molecular Phylogenetics and Evolution 120:286-296.

1293 Presser BD, Rutschky CW. 1957. The embryonic development of the corn earworm, *Heliothis*
1294 *zea* (Boddie) (Lepidoptea, Phalaenidae). Annals of the Entomological Society of America
1295 50:133-164.

1296 Prous M, Blank SM, Goulet H, Heibo E, Liston A, Malm T, Nyman T, Schmidt S, Smith DR,
1297 Vårdal H, et al. 2014. The genera of Nematinae (Hymenoptera, Tenthredinidae). Journal of
1298 Hymenoptera Research 40:1-69.

1299 Quan H, Arsala D, Lynch JA. 2019. Transcriptomic and functional analysis of the oosome, a
1300 unique form of germ plasm in the wasp *Nasonia vitripennis*. BMC Biology 17:78.

1301 Quan H, Lynch JA. 2016. The evolution of insect germline specification strategies. Current
1302 Opinion in Insect Science 13:99-105.

1303 Rafiqi AM, Rajakumar A, Abouheif E. 2020. Origin and elaboration of a major evolutionary
1304 transition in individuality. Nature 585:239-244.

1305 Rausell A, Juan D, Pazos F, Valencia A. 2010. Protein interactions and ligand binding: from
1306 protein subfamilies to functional specificity. Proceedings of the National Academy of Sciences of
1307 the United States of America 107:1995-2000.

1308 Raz E. 2000. The function and regulation of *vasa*-like genes in germ-cell development. Genome
1309 Biology 1:1-6.

1310 Saito. 1937. On the development of the Tusser, *Antheraea pernyi* Guerin-Meneville, with special
1311 reference to the comparative embryology of insects. Journal of the Faculty of Agriculture of
1312 Hokkaido Imperial University 40:35-109.

1313 Schmidt C. 2013. Molecular phylogenetics of ponerine ants (Hymenoptera: Formicidae:
1314 Ponerinae). Zootaxa 3647:201-250.

1315 Schroder R. 2006. *vasa* mRNA accumulates at the posterior pole during blastoderm formation in
1316 the flour beetle *Tribolium castaneum*. Development, Genes and Evolution 216:277-283.

1317 Schwangart F. 1905. Zur Entwickslungsgeschichte der Lepidopteren. Biol. Centralbl. 25:777-
1318 789.

1319 Sehl A. 1931. Furchung und Bildung der Keimanlage bei der Mehlmotte *Ephestia kuehniella*.
1320 Zell. Zeit. Morph. U. Okol. 1:429-506.

1321 Shafiq SA. 1954. A study of the embryonic development of the Gooseberry Sawfly, *Pteronidea*
1322 *ribesii*. Quarterly Journal of Microscopical Science 95:93-114.

1323 Sharanowski BJ, Ridenbaugh RD, Piekarski PK, Broad GR, Burke GR, Deans AR, Lemmon AR,
1324 Lemmon ECM, Diehl GJ, Whitfield JB, et al. 2021. Phylogenomics of Ichneumonoidea
1325 (Hymenoptera) and implications for evolution of mode of parasitism and viral endogenization.
1326 Molecular Phylogenetics and Evolution 156:107023.

1327 Sikosek T, Chan HS. 2014. Biophysics of protein evolution and evolutionary protein biophysics.
1328 Journal of the Royal Society, Interface / the Royal Society 11:20140419.

1329 Sikosek T, Chan HS, Bornberg-Bauer E. 2012. Escape from Adaptive Conflict follows from
1330 weak functional trade-offs and mutational robustness. Proceedings of the National Academy of
1331 Sciences of the United States of America 109:14888-14893.

1332 Silvestri F. 1906. Contribuzioni alla conoscenza biologica degli Imenotteri parassiti. I. Biologia
1333 del *Litomastix truncellatus* Dalm. Annali della r. Scuola Superiore di Agricoltura in Portici 6:3-51.

1334 Silvestri F. 1908. Contribuzioni alla conoscenza degli Imenotteri parassiti. Bollettino del
1335 Laboratorio di Zoologia Generale e Agraria della r. Scuola Superiore d'Agricoltura (AFTW.
1336 Facoltà Agraria) in Portici 3:29-84.

1337 Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of
1338 large phylogenies. Bioinformatics 30:1312-1313.

1339 Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. 2006. AUGUSTUS: ab initio
1340 prediction of alternative transcripts. Nucleic Acids Research 34:W435-439.

1341 Strand MR, Grbic' M. 1997. The Development and Evolution of Polyembryonic Insects. Current
1342 Topics in Developmental Biology 35:121-159.

1343 Sumitani M, Yamamoto DS, Oishi K, Lee JM, Hatakeyama M. 2003. Germline transformation of
1344 the sawfly, *Athalia rosae* (Hymenoptera: Symphyta), mediated by a piggyBac-derived vector.
1345 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.

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

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.

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

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.

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

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.

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

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

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

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

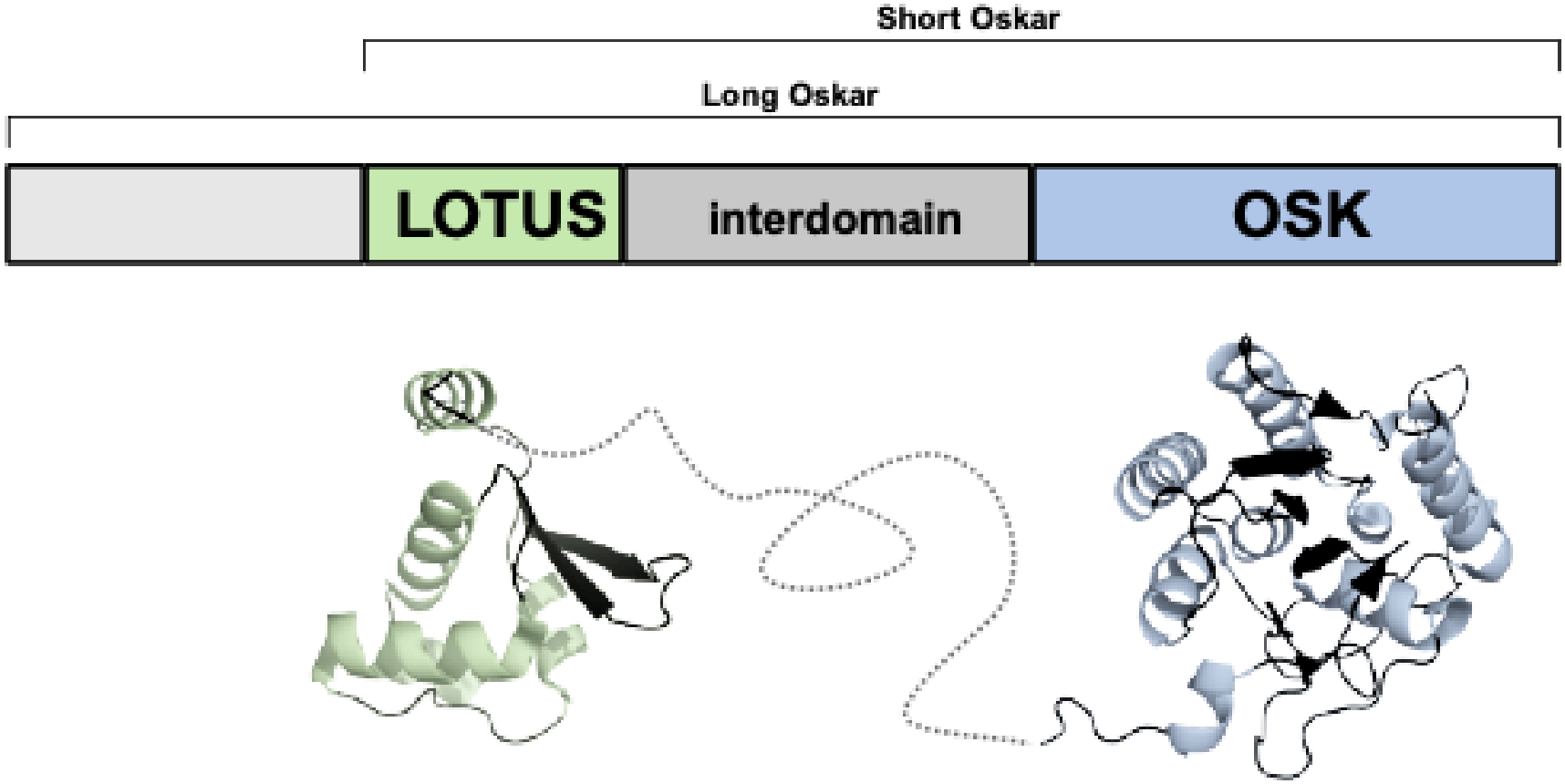
1383 Yang N, Yu Z, Hu M, Wang M, Lehmann R, Xu RM. 2015. Structure of *Drosophila* Oskar
1384 reveals a novel RNA binding protein. Proceedings of the National Academy of Sciences of the
1385 United States of America 112:11541-11546.

1386 Zhurov V, Terzin T, Grbic M. 2004. Early blastomere determines embryo proliferation and caste
1387 fate in a polyembryonic wasp. Nature 432:764-769.

1388 Zissler D. 1992. From egg to pole cells: ultrastructural aspects of early cleavage and germ cell
1389 determination in insects. Micr. Res. and Tech.:49-74.

1390

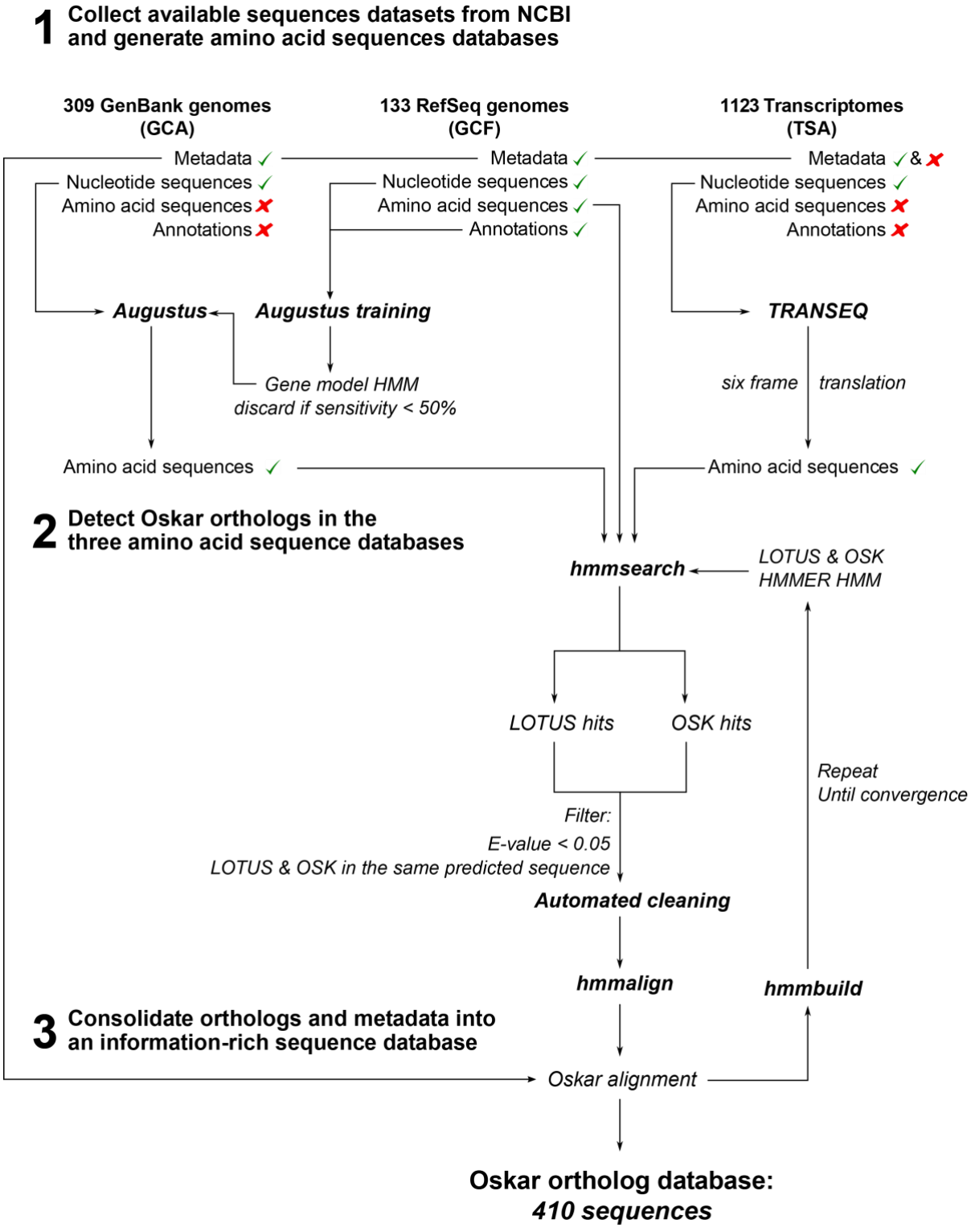
Figure 1



.

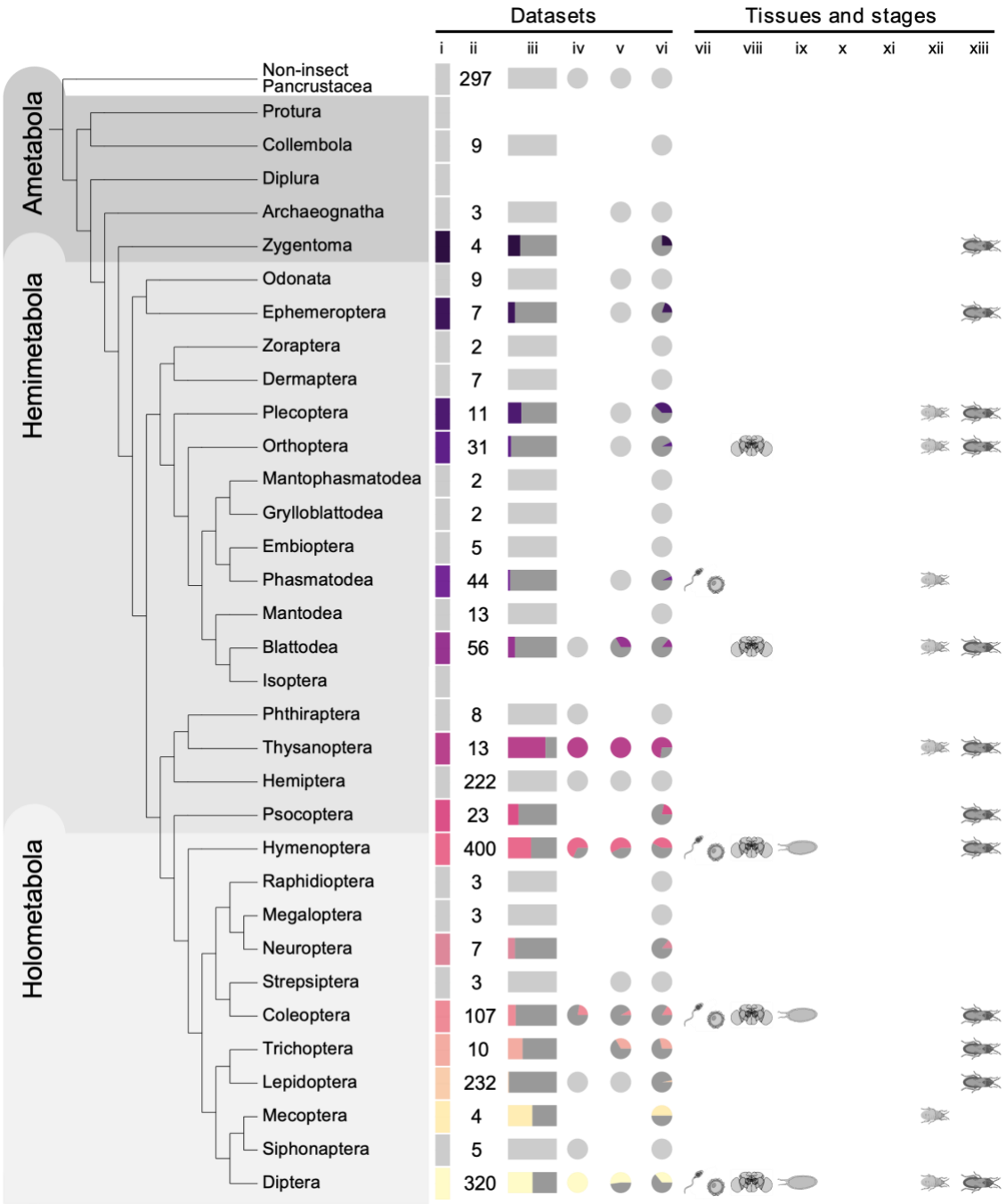
1395 **Figure 2**

1396

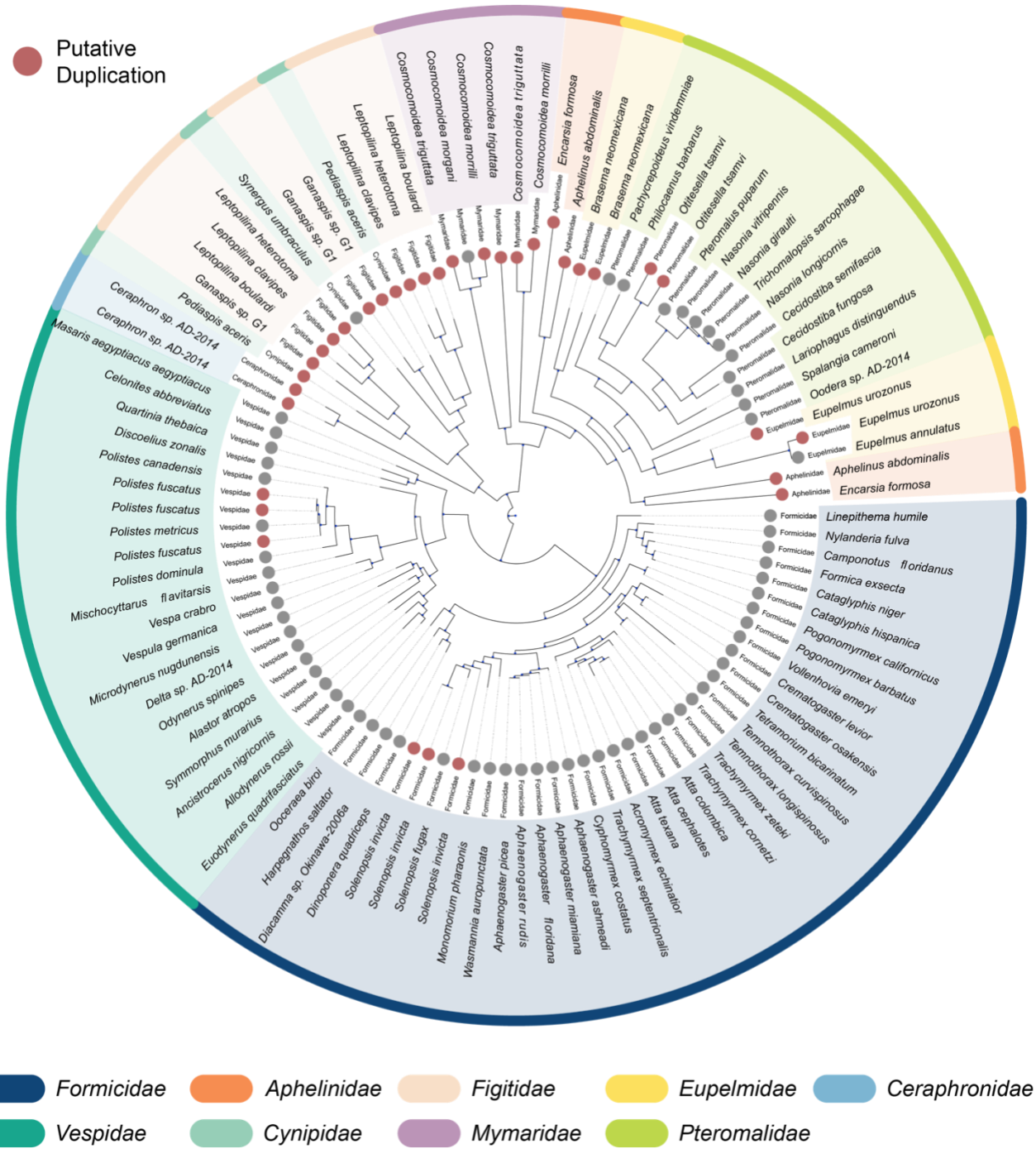


1397

Figure 3

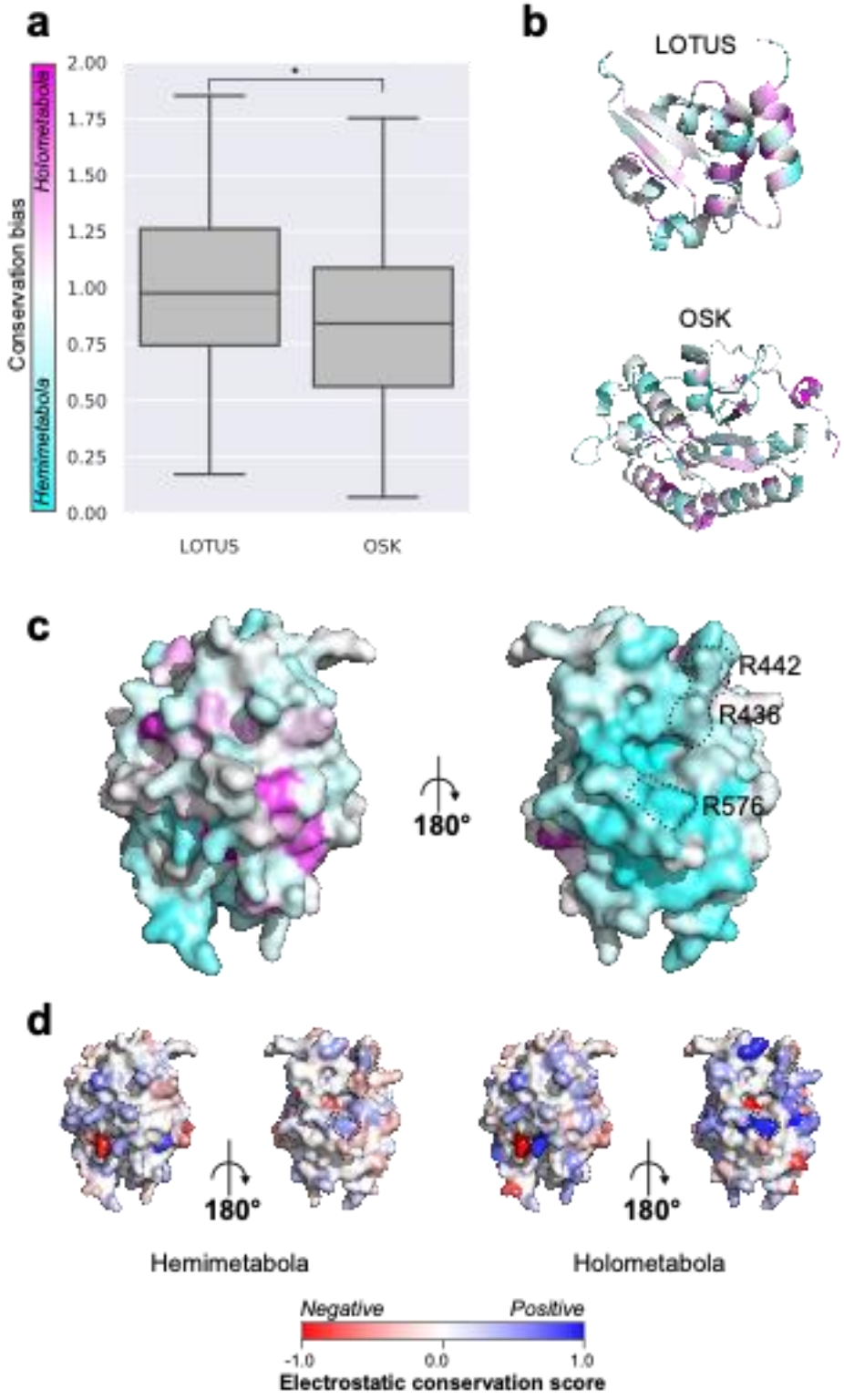


1401 **Figure 4**



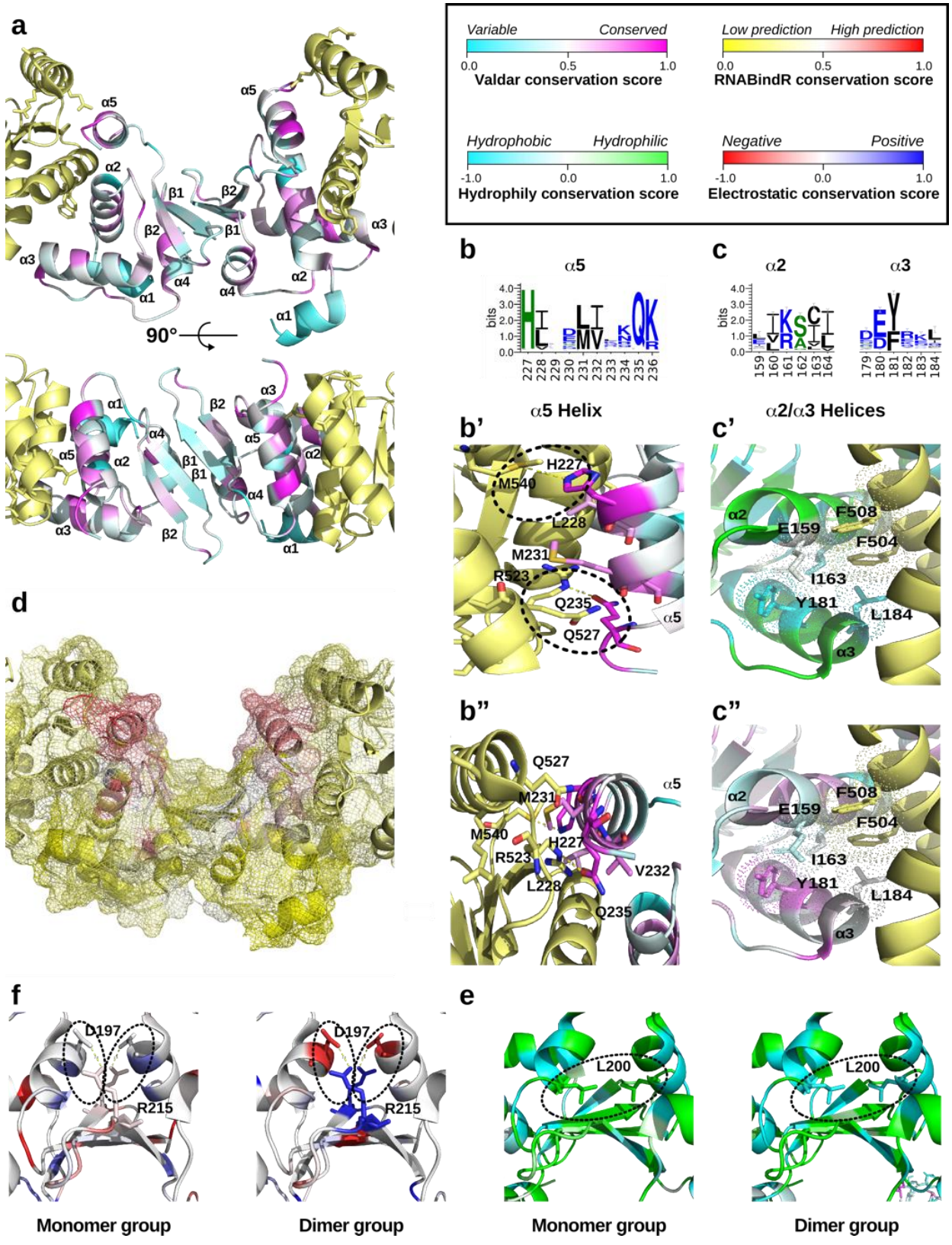
1402

1403 **Figure 5**



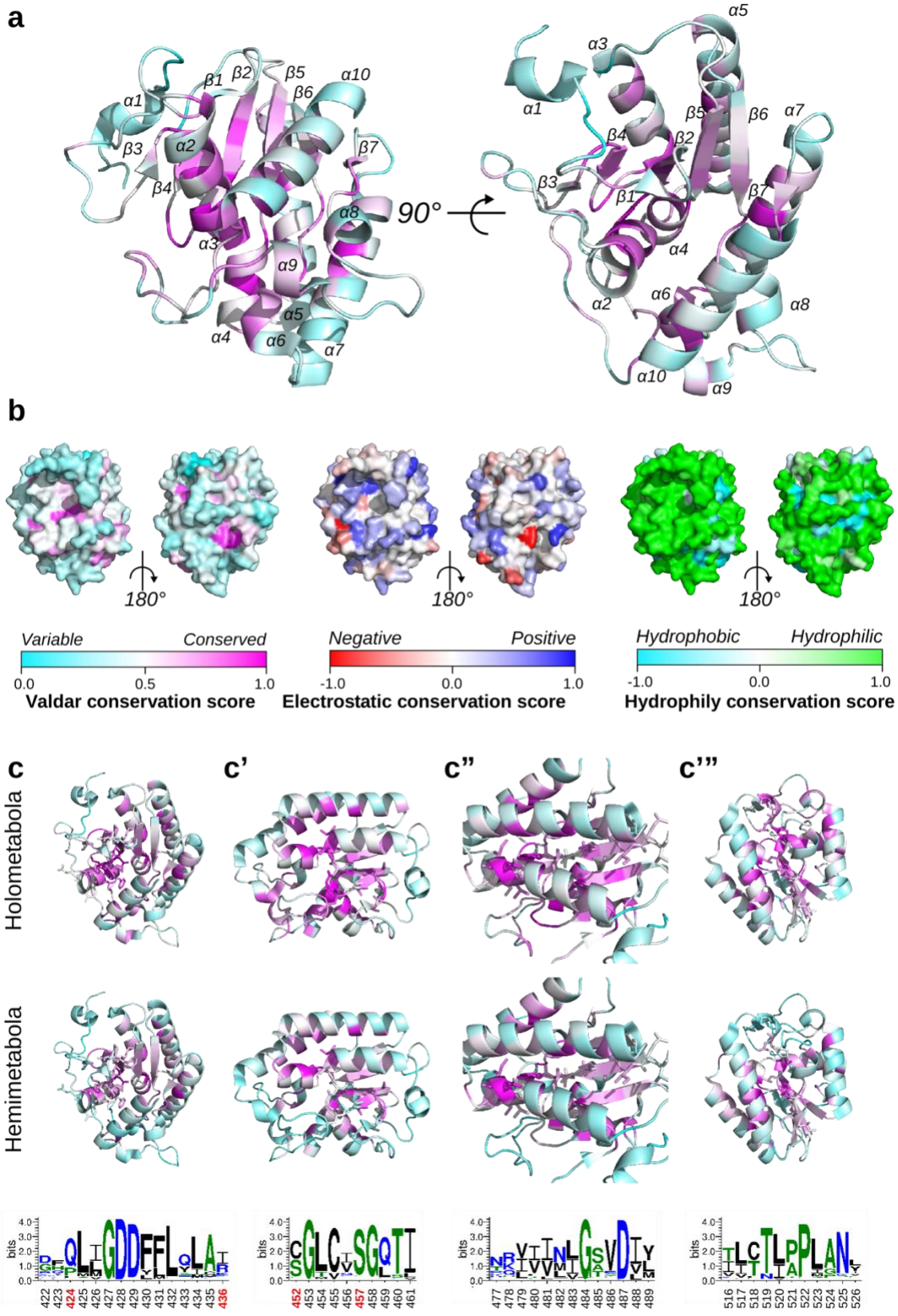
1404

1405 **Figure 6**



1406

1407 **Figure 7**



1408

Supplementary Materials

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

Leo Blondel, Savandara Besse and Cassandra G. Extavour

These Supplementary Materials contain the following:

- Supplementary Tables S1 through S5
- Supplementary Figures S1 through S9

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.

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 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 S6: 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

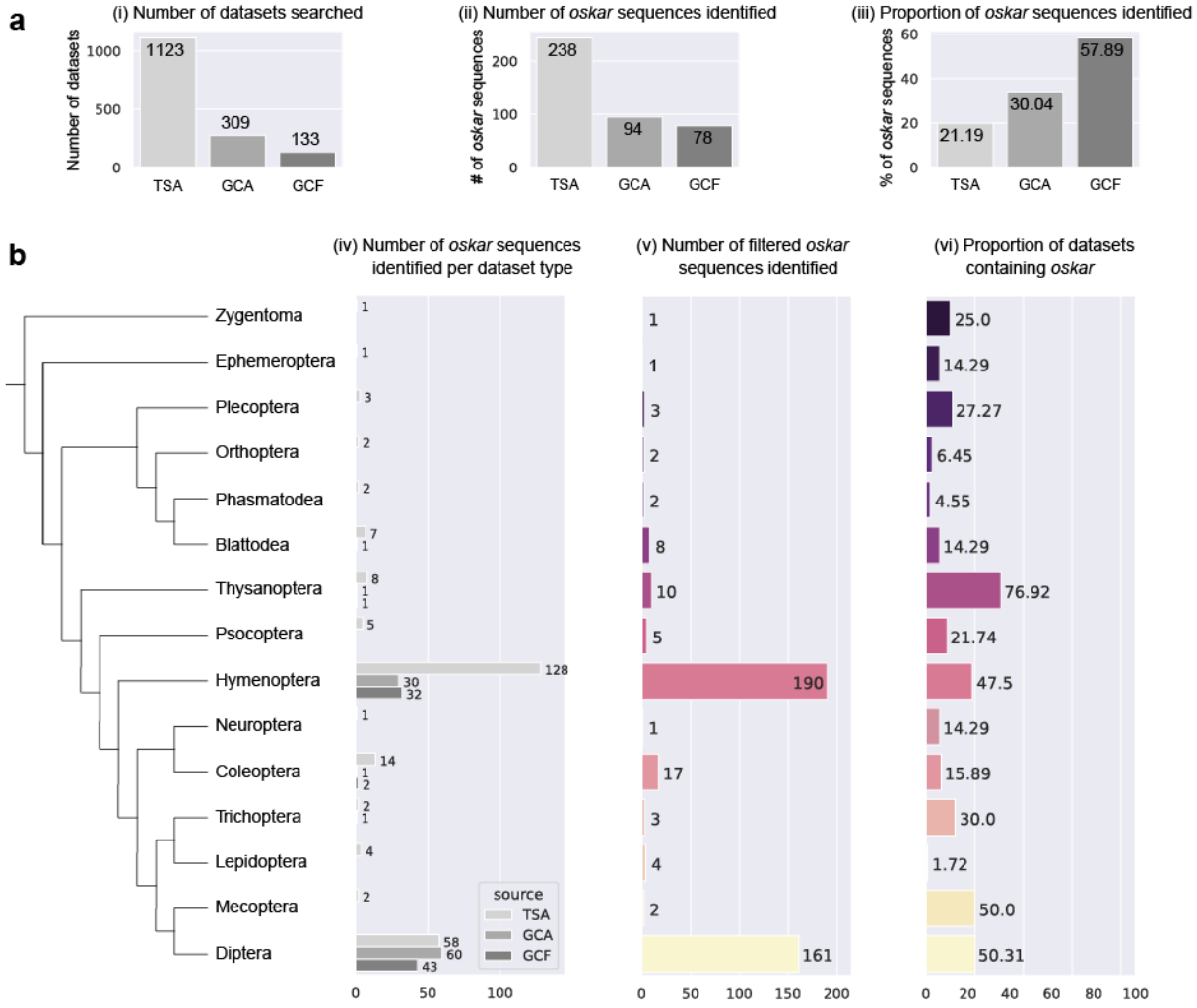
Supplementary Figure S7: Evolution of the structure of Oskar in Diptera. Left: dipteran phylogeny from (Maddison, et al. 2007; Wiegmann, et al. 2011). Top: schematic representation of Oskar domain structure. Blue: heatmap showing the overall occupancy of an amino acid position in the Oskar alignment 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 S8: 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 S9. 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

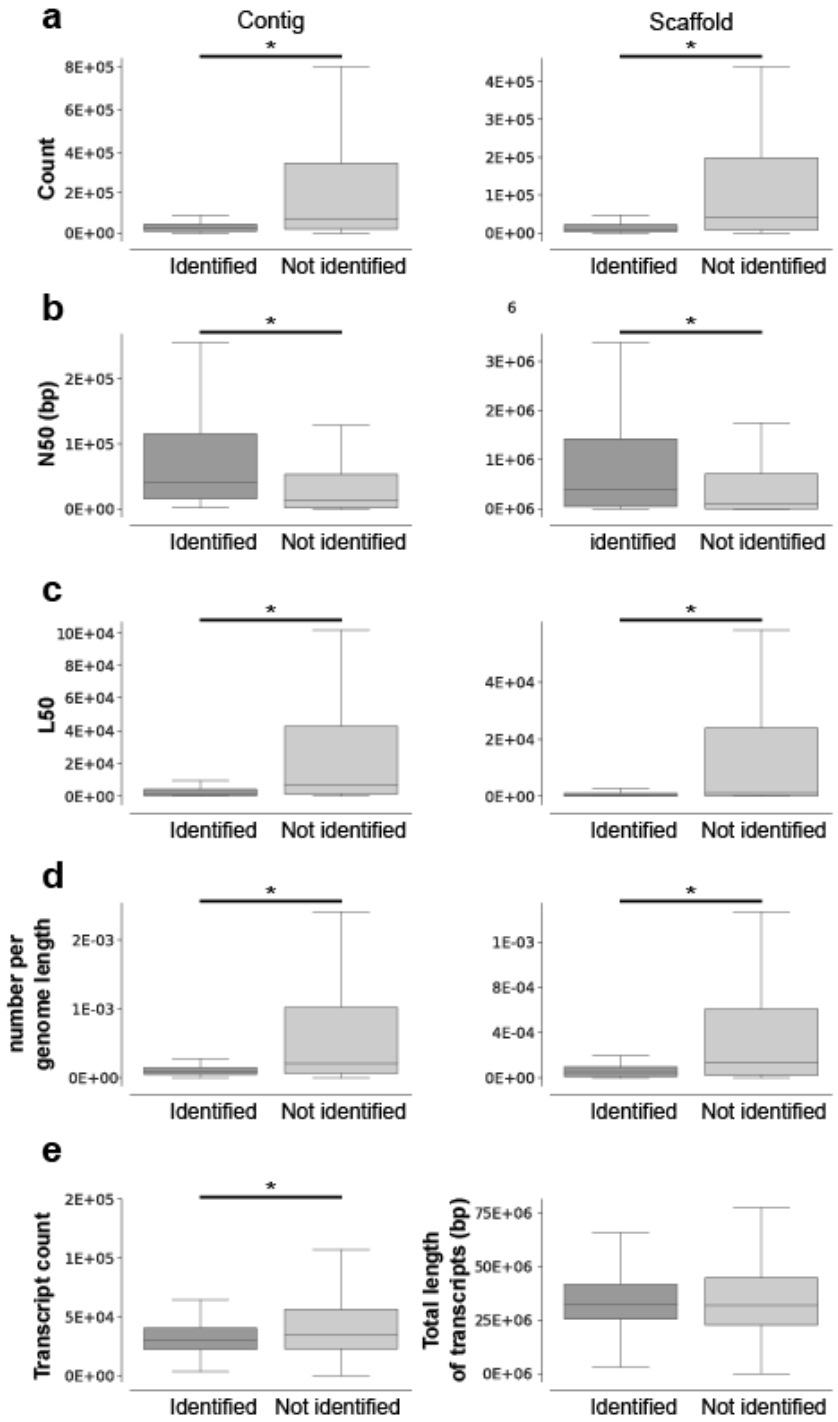
findings obtained from a transcriptome (TSA)), o for Hymenoptera species or genera. For some species, germ cell specification via germ plasm (black circles) or differentiation from mesoderm (no germ plasm; white circles) has been reported in the literature, with primary data references indicated by numbers 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 Figure 1

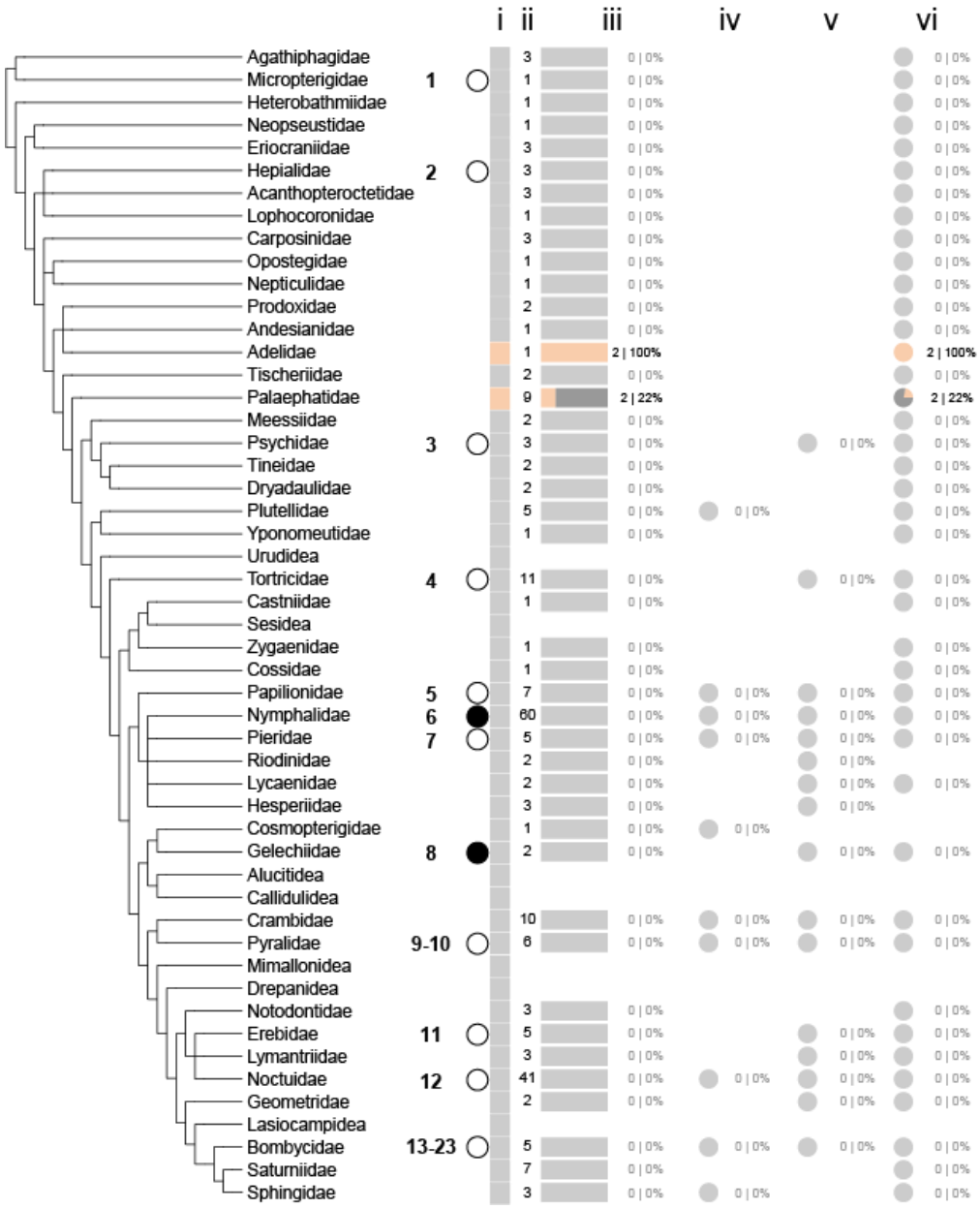


.

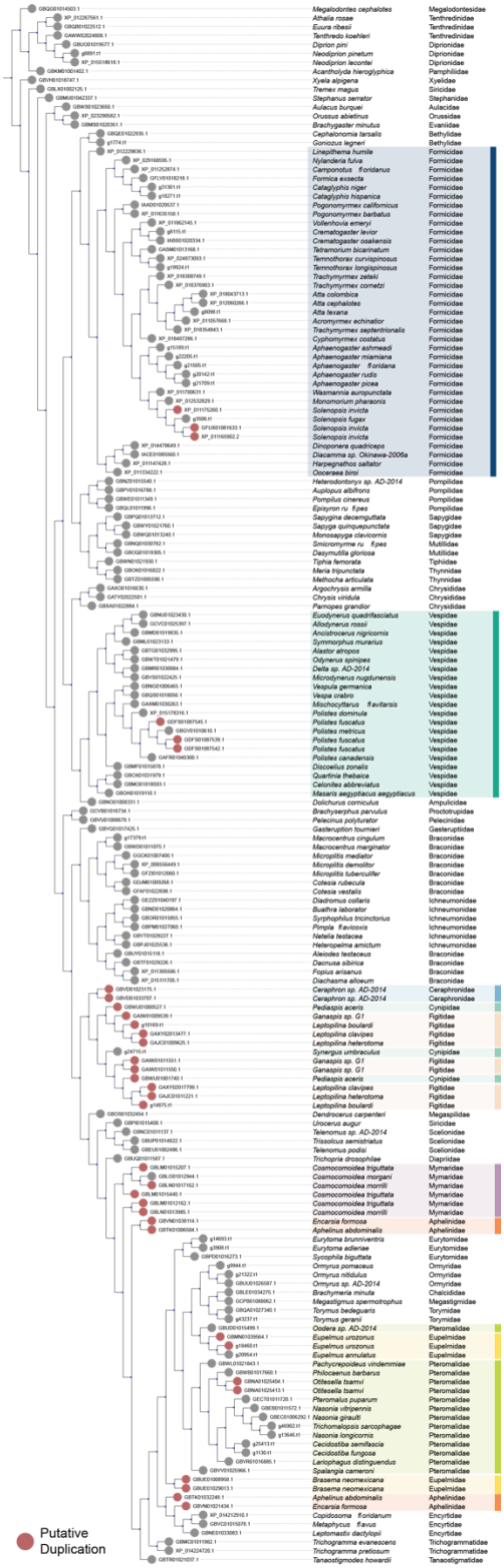
Supplementary Figure 2



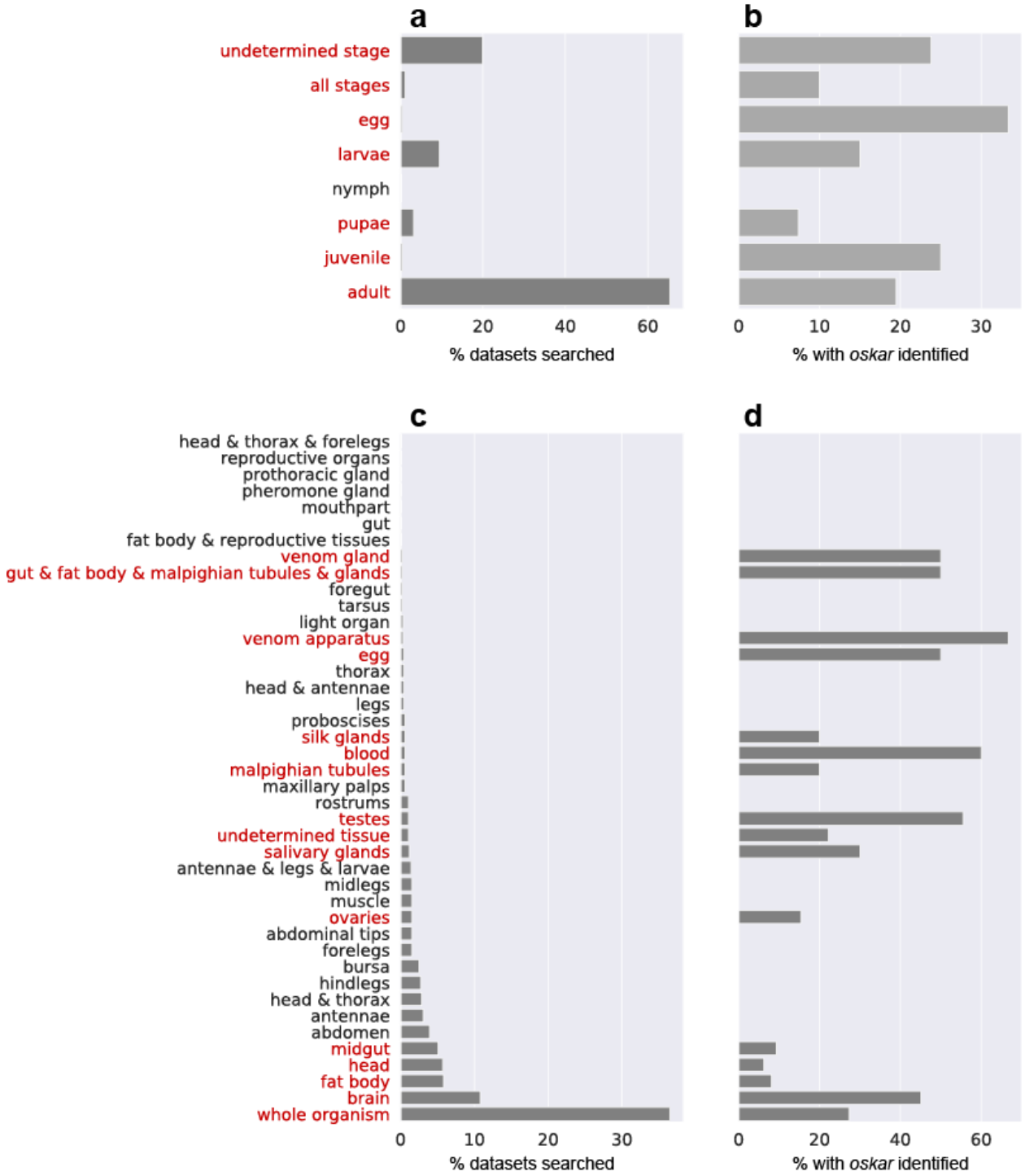
Supplementary Figure 3



Supplementary Figure 4

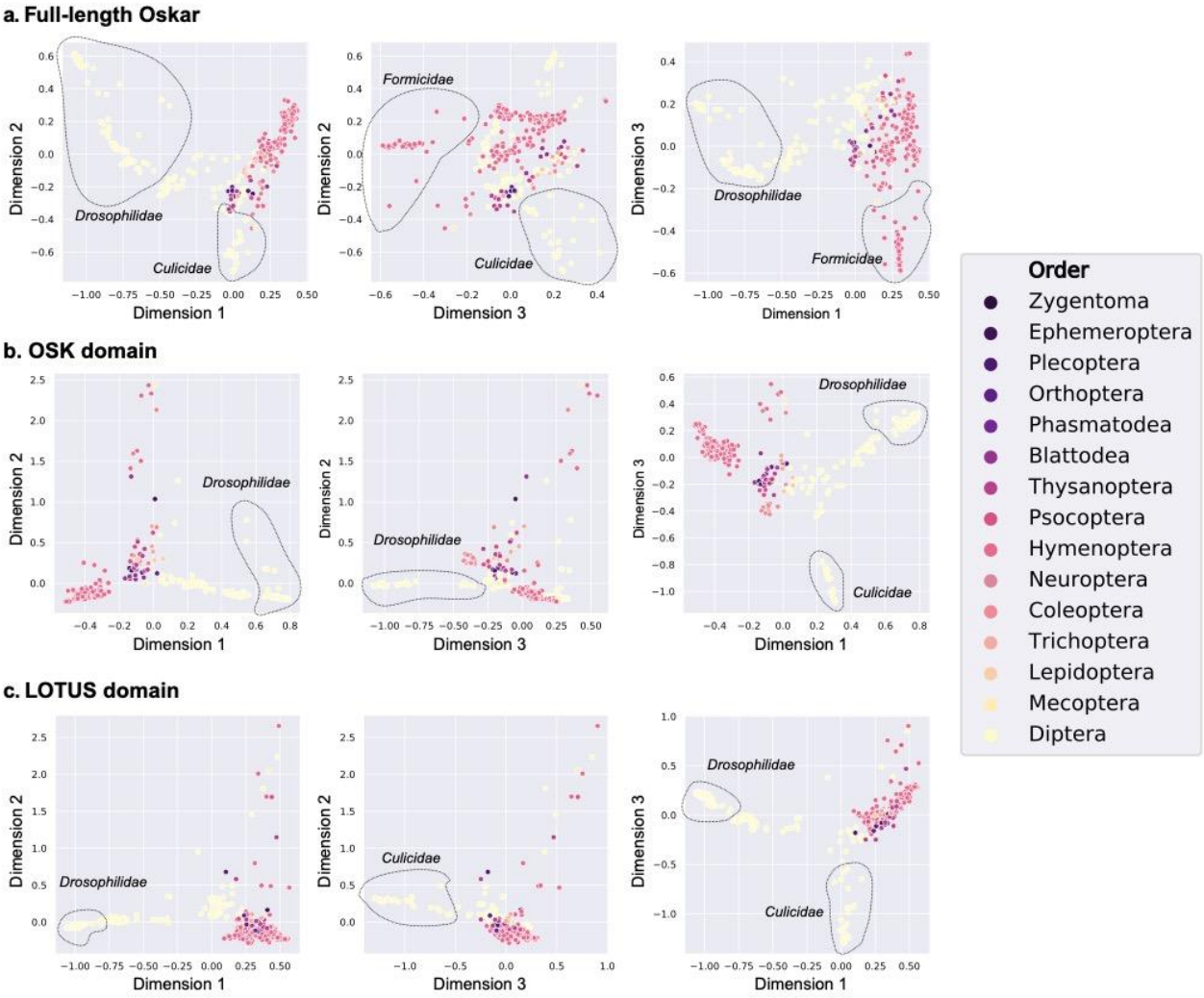


Supplementary Figure 5



1594 **Supplementary Figure 6**

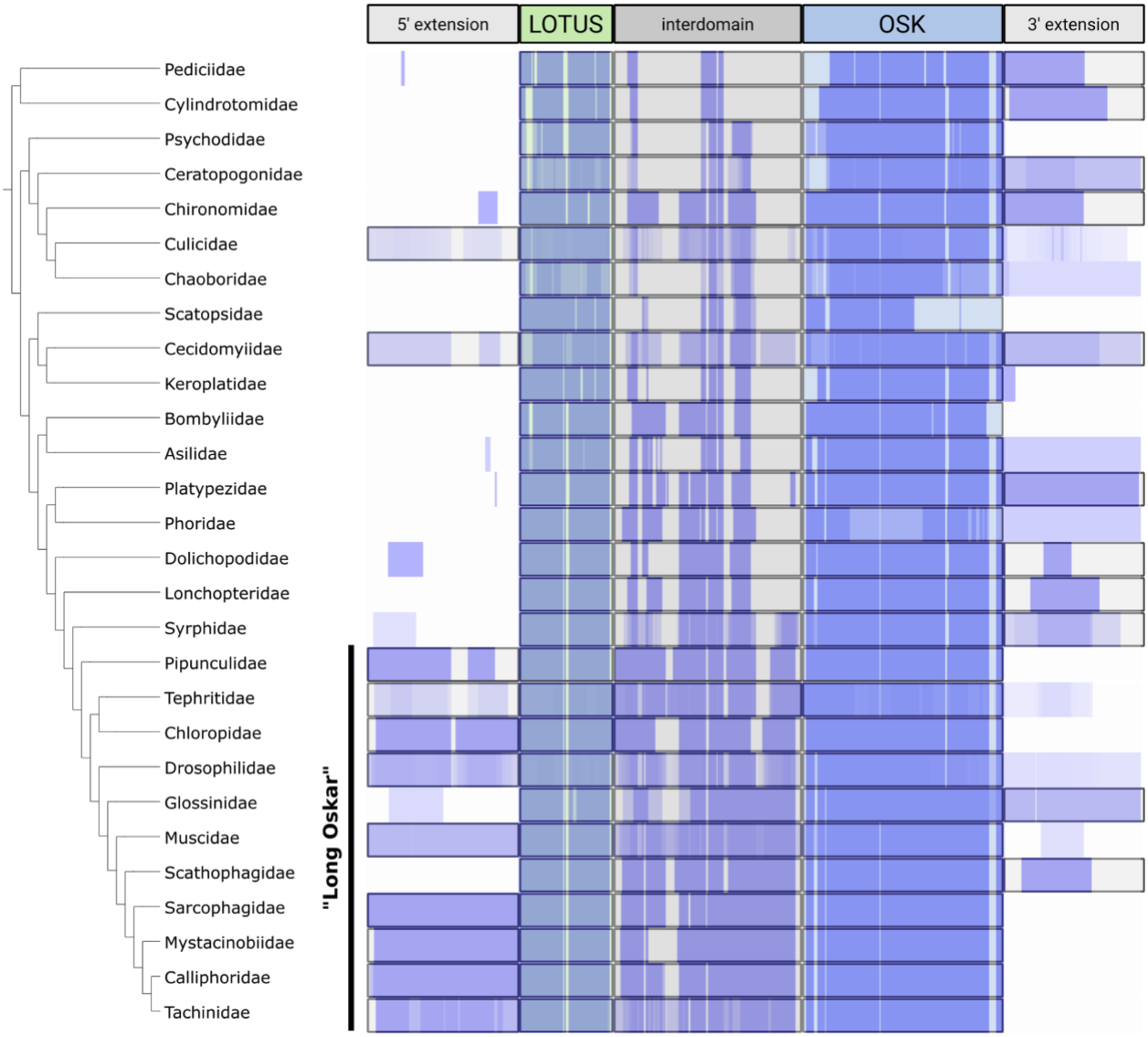
1595



1596

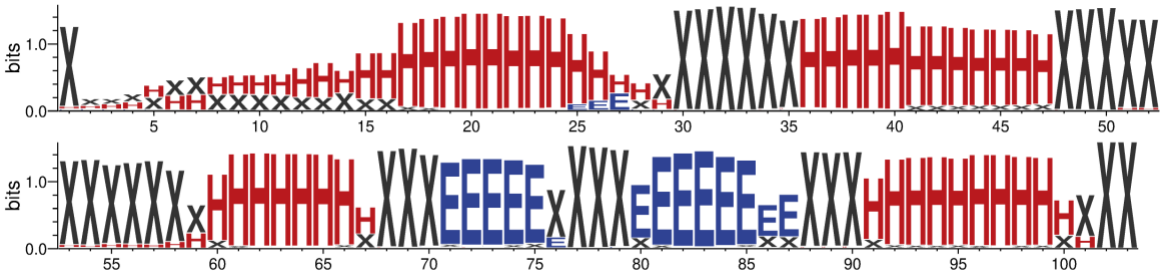
1597 .

Supplementary Figure 7

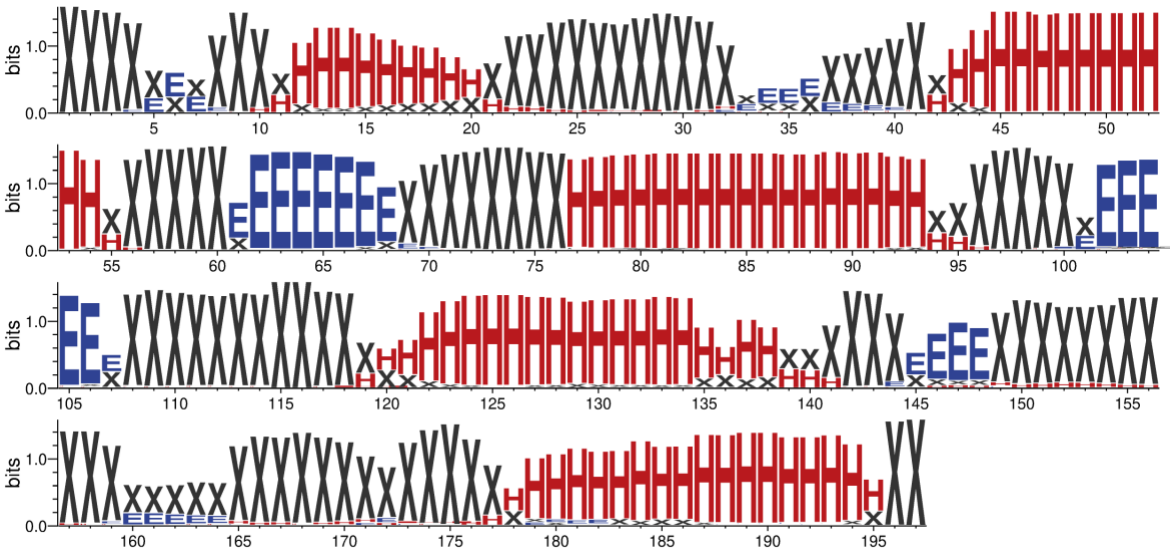


1602 **Supplementary Figure 8**

a. LOTUS secondary structure conservation

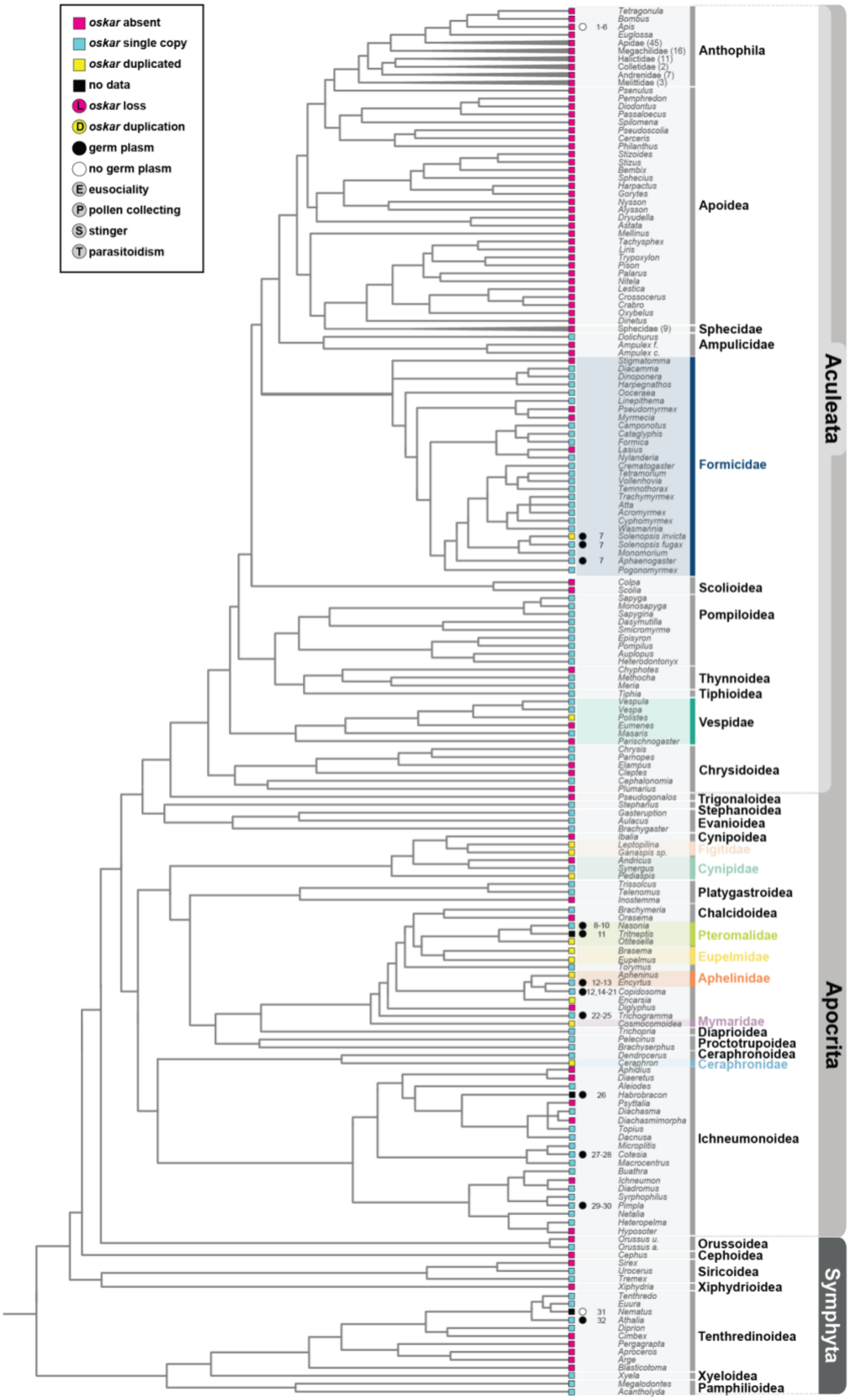


b. OSK secondary structure conservation



1603
1604

Supplementary Figure 9



1607 **Supplementary Table S1**

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

1608
1609
1610

1611 **Supplementary Table S2**

1612

	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

1613

1614