Genomic Resources for the Scuttle Fly *Megaselia abdita*: A Model Organism for Comparative Developmental Studies in Flies

Running title: Genomic resources for *M. abdita*

Authors:

Ayse Tenger-Trolander[1*] <ayse.trolander@gmail.com> (ORCiD 0000-0001-8319-0273)
Ezra Amiri[1] (ORCiD 0000-0002-6241-6656)
Valentino Gantz[2,3] (ORCiD 0000-0003-2453-0711)
Chun Wai Kwan[1,4] (ORCiD 0000-0003-3974-2979)
Sheri A Sanders[5] (ORCiD 0000-0003-2969-2541)
Urs Schmidt-Ott[1*] <uschmid@uchicago.edu> (ORCiD 0000-0002-1351-9472)
* Authors for correspondence

Affiliations:

1. University of Chicago, Dept. of Organismal Biology and Anatomy, 1027 East 57th Street, Chicago, Illinois 60637, USA
2. Section of Cell and Developmental Biology, University of California San Diego, La Jolla, CA 92093, USA.
3. Pattern Biosciences, Inc. 681 Gateway Blvd, South San Francisco, CA 94080
4. Laboratory for Epithelial Morphogenesis, RIKEN Center for Biosystems Dynamics Research, 2-2-3 Minatojima-minamimachi, Chuo-ku, Kobe, Hyogo 650-0047, Japan
5. Notre Dame University, 252 Galvin Life Science Center/Freimann Life Science Center, Notre Dame, Indiana 46556, USA

Key words: Genome assembly, genome annotation, transcriptome, non-traditional model organism, evolutionary development, synteny analysis, orphan genes

Summary statement

We report a chromosome-level genome assembly and annotation and transcriptomes for an emerging developmental model organism, the phorid fly Megaselia abdita, which is phylogenetically intermediate between Drosophila and mosquitoes.

Abstract

The order Diptera (true flies) holds promise as a model taxon in evolutionary developmental biology due to the inclusion of the model organism, *Drosophila melanogaster*, and the ability to cost-effectively rear many species in laboratories. One of them, the scuttle fly *Megaselia abdita* (Phoridae) has been used in evolutionary developmental biology for 30 years and is an excellent phylogenetic intermediate between fruit flies and mosquitoes but remains underdeveloped in genomic resources. Here, we present a *de novo* chromosome-level assembly and annotation of *M. abdita* and transcriptomes of 9 embryonic and 4 postembryonic stages. We also compare 9 stage-matched embryonic transcriptomes between *M. abdita* and *D. melanogaster.* Our analysis of these resources reveals extensive chromosomal synteny with *D. melanogaster,* 28 orphan genes with embryo-specific expression including a novel F-box LRR

gene in *M. abdita*, and conserved and diverged features of gene expression dynamics between *M. abdita* and *D. melanogaster*. Collectively, our results provide a new reference for studying the diversification of developmental processes in flies.

Introduction

Comparing related species is a powerful approach to understanding how mechanisms of development evolve and diversify. Their natural diversity and their phylogenetic history can aid in revealing core principles and inform our understanding of evolutionary transitions. Additionally, careful comparisons of developmental mechanisms between multiple species in "model taxa" are critical for determining the directionality of change and identifying mutations responsible for important evolutionary shifts in developmental processes. Such mutations may not necessarily be adaptive and their significance as drivers of evolutionary change might be overlooked, given that complex developmental gene networks can enhance a population's permissiveness for the passive fixation of mutational variants that open novel paths of adaptive evolution (Kimura and Ohta, 1974; Lynch, 2007a, 2007b).

While it is impractical to adapt a large set of closely related vertebrate model organisms for laboratory studies, insects, in particular Diptera (true flies), offer a cost-effective alternative (Grimaldi and Engel, 2005; Schmidt-Ott and Lynch, 2016; Wiegmann et al., 2011). Flies are particularly appealing for the comparative study of developmental mechanisms because they include a leading model organism in developmental biology, *Drosophila melanogaster*, and many more species that are relatively easily cultured in laboratories. Developmental biologists have introduced several new dipteran model organisms in recent years, including the humpbacked fly, *Megaselia abdita*, which has been particularly useful for studying the evolution of developmental mechanisms in dipteran embryos because of its technical advantages and phylogenetic position (Rafiqi et al., 2011). It belongs to the large family Phoridae, also known as scuttle flies (Disney, 1994; Li et al., 2024), and represents a lineage that separated from the *Drosophila* lineage ca. 145 million years ago at the beginning of the Cyclorrhapha radiation, roughly 100 million years into the dipteran radiation (Grimaldi and Engel, 2005). Developmental biologists started using *M. abdita* as an experimental system in the 1990s to study the evolution of axial pattern formation (Bullock et al., 2004; Crombach and Jaeger, 2021; Crombach et al., 2016; Liu et al., 2018; Rohr et al., 1999; Schmidt-Ott et al., 1994; Stauber et al., 1999; Stauber et al., 2000; Stauber et al., 2002; Wotton et al., 2015a; Wotton et al., 2015b; Wotton et al., 2015c; Yoder and Carroll, 2006), and subsequently the evolution of extraembryonic tissue (Caroti et al., 2018; Fraire-Zamora et al., 2018; Horn et al., 2015; Kwan et al., 2016; Rafiqi et al., 2008; Rafiqi et al., 2010; Rafiqi et al., 2012; Schmidt-Ott and Kwan, 2022; Stauber et al., 1999; Wotton, 2014; Wotton et al., 2014), and other aspects of embryo development (Caroti et al., 2015; Dey et al., 2023; Tanaka et al., 2015; Vicoso and Bachtrog, 2015). However, the limited availability of genomic resources in *M. abdita* (Jimenez-Guri et al., 2013; Vicoso and Bachtrog, 2015) and the Phoridae in general (Feng et al., 2020; Rasmussen and Noor, 2009; Zhong et al., 2016) has limited the potential of this model organism by precluding genome-wide and epigenetic experimental approaches (Figure 1).

Here we provide a *de novo* assembled and annotated chromosome-level genome for *M. abdita*, alongside stage-specific transcriptomes across its life cycle. These resources provide an excellent basis for genome-wide and epigenetic experimental approaches for an understudied but phylogenetically important branch of the Diptera. They also establish synteny relationships with the chromosomes of *D. melanogaster*, highlight conserved and divergent features of Hox gene clusters, and provide insights into embryonic gene expression dynamics and orphan genes. Collectively, our results will help to establish dipterans as a model taxon to study the evolution of developmental mechanisms from gene regulation to neural networks and behavior.

Results and Discussion

Genome assembly

We generated a *de novo* reference genome for *Megaselia abdita* using combined long read and chromatin conformation capture methods obtained from several hundred embryos of a 10-generation inbred line. Long-read, high-fidelity sequences were generated by PacBio (HiFi PacBio reads), and chromatin conformation capture sequences were generated by Dovetail Genomics' Omni-C method. After removing 33 scaffolds identified as contamination, the initial draft assembly spanned 592.8 megabases (Mb) contained in 89 scaffolds with an N50 of 12.6 Mb. Using HiRise, a software designed to scaffold genome assemblies with proximity ligation data (Putnam et al., 2016), the draft assembly was refined using the Omni-C reads (Figure S1).

The final assembly is 592.8 Mb contained in 15 scaffolds with an N50 of 212.8 Mb (Table 1). The genome size is comparable to a previous estimate of 562.7 Mb based on flow cytometry data (Picard et al. 2012). We estimated heterozygosity at 0% - 0.24% (Figure S2A). We masked 67.9% of the genome, constituting repetitive elements (Figure S2B). The largest three scaffolds correspond to the three chromosomes of *M. abdita* (Table S1). We aligned the remaining 12 scaffolds to the largest three and found that these scaffolds contain repetitive sequences that cannot be correctly assembled into chromosome-level scaffolds.

Genome annotation

We annotated the reference genome of *M. abdita* using evidence from RNAseq transcripts and protein sequence databases, and a robust genome annotation pipeline (Figure S3). The process involved mapping RNAseq data and assembling the transcriptome, mapping protein sequences to the reference genome, and generating gene models using various prediction software. We then created consensus gene models using EVidenceModeler, updated the models to include UTRs and alternative isoforms, and filtered out transposable element models before functionally annotating the genes with Eggnog-mapper (Haas et al., 2008; Huerta-Cepas et al., 2019; Cantalapiedra et al., 2021). In the Materials and Methods section, we provide a detailed walkthrough of this pipeline.

M. abdita's genome contains 11,934 protein-coding genes (Table 2). We assessed the quality of the annotation with Benchmarking Universal Single-Copy Orthologs (BUSCO) (Simão et al.,

2015; Manni et al., 2021). BUSCO evaluates annotation completeness by looking for the presence of highly conserved 'single-copy orthologs' across specific taxonomic groups. For example, the BUSCO Eukaryota database expects 255 orthologs, while the Diptera database expects 3,285. For the *M. abdita* genome annotation we found 93% eukaryotic and 88% dipteran 'complete single copy orthologs', indicating a high-quality assembly (Table 2).

The genome size of *M. abdita* is significantly larger than that of *D. melanogaster* (592.8 Mb vs 139.5 Mb). We also found the mean gene length of protein-coding genes in *M. abdita* to be significantly longer than that of *D. melanogaster*, ~19 kilobases (kb) vs ~6.9 kb respectively. However, the lengths of transcripts and exons are remarkably consistent between the two species indicating that *M. abdita's* larger genome size is partially attributable to longer introns (Table S2).

Genome browser and genomic analysis tools

To improve accessibility and usability of the genomic data hosted on NCBI, we developed a genome browser ecosystem centered on JBrowse2 (Diesh et al. 2023). This ecosystem is available as a cloud image on NSF's Jetstream2 platform (image: *Megaselia abdita Genome Resources*, Hancock et al. 2021, Boerner et al. 2023), with a portable Docker image currently in development. The browser integrates tools for comprehensive genomic analysis, including BLAST search (SequenceServer2.0, Priyam et al. 2019), CRISPR guide RNA design (modified crisprDesigner, Beeber and Chain 2020), differential gene expression (DGE) analysis via R Shiny (freecount, Brooks et al. 2024), and synteny mapping (ShinySyn, Xiao and Lam 2022). Future updates will include expanded sgRNA profiling, Docker support for deployment on commercial cloud platforms, and continuing optimization of workflows, ensuring this resource remains a powerful and accessible tool for genomic research and education.

Synteny analysis reveals significant collinearity between *M. abdita* and *D. melanogaster* genomes, including HOX gene cluster arrangement

Analysis of chromosomal synteny can aid in the identification of orthologs and highlight genomic regions with conserved regulatory potential. To investigate genome-wide synteny between *M. abdita* and *D. melanogaster*, we performed a synteny analysis, identifying significant collinearity between their genomes, including the arrangement of the split HOX gene cluster. We compared the synteny and collinearity of *M. abdita* scaffolds with the *D. melanogaster* genome and identified 387 collinear blocks encompassing 2,396 genes (Figure 2A). Large portions of *D. melanogaster* chromosomes are syntenic with single scaffolds in *M. abdita*. Scaffold 1 of *M. abdita* largely corresponds to chromosome arm 3L, chromosome 4, and chromosome arm 2R of *D. melanogaster*, while scaffold 2 aligns with chromosome arm 2L and the distal portion of chromosome arm 3R, and scaffold 3 with the proximal portion of chromosome arm 3R and the X chromosome. The arrangement of the HOX genes in *M. abdita* mirrored that of *D. melanogaster*, with distinct Antennapedia and Ultrabithorax complexes separated by 53,306 kb in *M. abdita* (9,978 kb in *D. melanogaster*). Both complexes are located on Scaffold 2 of *M. abdita's* genome and maintain the same gene order seen in *D. melanogaster*, including the cuticle gene complex (Figure 2B). Additionally, the genes *zerknüllt* (*zen*) and *amalgam* have

158 undergone duplication in *M. abdita* (Figure 2B). *Zen* has experienced many duplications in
159 Diptera (Mulhair and Holland 2024). Since only the ~60 amino acid homeodomain of *zen* is
160 comparable between species, it is difficult to establish the relatedness of *M. abdita's zen-like* to
161 other *zen* genes and duplications by sequence alone.

162 *M. abdita*'s *zen* gene is expressed in the serosa and has been characterized in detail (Caroti et
163 al., 2018; Kwan et al., 2016; Rafiqi et al., 2008; Rafiqi et al., 2010; Rafiqi et al., 2012; Stauber et
164 al., 1999). Consistent with these studies, we detected the *M. abdita's zen* transcript in
165 embryonic stages 5 - 15, coinciding with the time when the serosa is specified and maintained
166 (Figure S4). In contrast, the newly identified *zen-like* gene's expression was only detected at
167 stage 5 (Figure S4). Our findings underscore how a well-annotated genome can facilitate
168 precise comparisons of chromosomal architecture which will enable the identification of both
169 conserved regulatory landscapes and structural variation implicated in phenotypic diversity.

170 **Major transition in transcriptional expression profile during germband retraction**

171 The extent to which gene expression is similar or different between stages and species provides
172 a basis for identifying developmental windows of accelerated change, heterochronic shifts, and
173 evolutionary divergence. To identify conserved features and differences between embryonic
174 stages of *M. abdita* and *D. melanogaster*, we performed RNA-seq on single embryos from
175 stages 1, 5, 8, 9, 10, 12, 13, 15, 16, and 17 for both species (Figure S5). We staged embryos
176 based on morphology, corresponding to established staging schemes for each species
177 (Campos-Ortega & Hartenstein, 1997; Wotton et al., 2014). We later excluded the *M. abdita*
178 stage 16 embryo due to failed sequencing. Despite similar embryo sizes and uniform library
179 preparation and sequencing conditions, *M. abdita* embryos had roughly twice the number of
180 reads per embryo compared to *D. melanogaster* (Figure S6A) though the total number of genes
181 with reads assigned was comparable between species with 10,373 in *D. melanogaster* to 9,999
182 in *M. abdita*. The number of genes expressed was similar across embryonic stages between
183 species as well (Figure S6B). The number of reads mapping to genes doubled in *M. abdita*
184 (8,941 vs. 3,863 reads/gene), which was consistent with its 2x read count. We used normalized
185 read counts (transcripts per million or TPM) to account for the global expression level
186 differences, sample-to-sample variation, and differences in transcript length between genes.
187 Our RNAseq data showed a major shift in transcriptional expression during germband retraction
188 in both *M. abdita* and *D. melanogaster*. A multidimensional scaling (MDS) plot, which plotted
189 samples based on the similarity of their top 500 most differentially expressed genes, revealed
190 clustering of embryo transcriptomes before and after germband retraction (Figure 3A).

191 To more closely look at the gene networks in both species, we identified the differentially
192 expressed genes (DEGs) between developmental stages. Due to the absence of biological
193 replicates for each embryo, we used a k-mean clustering approach to group samples and
194 calculate a global dispersion estimate ($\sigma = 0.36$ for both *M. abdita* and *D. melanogaster*). Both
195 species have high dispersion estimates but variance in expression was very similar in both
196 datasets (Figure S7 & S8). Despite the high dispersion in data, we had sufficient power to detect
197 DEGs with fold differences ±3 and p-value < 0.001.

198 We performed pairwise comparisons of each sequential stage: the zygote, containing maternal-
199 deposited transcripts (Stage 1); cellularization, where maternal to zygotic transition (MZT) of
200 gene expression occurs (Stage 5); gastrulation, germband extension & retraction (Stages 8-12);
201 dorsal closure (Stages 13-16), and the final embryonic stage (stage 17). We detected 1,567 and
202 2,398 DEGs in *M. abdita* and *D.* in *melanogaster* respectively (Figure 3B). *M. abdita* exhibited
203 more dynamic gene expression than *D. melanogaster* between cellularization and gastrulation
204 (stages 5-8), with 418 DEGs in *M. abdita* compared to 123 in *D. melanogaster* (Figure 3B).
205 Between stages 8 and 12, corresponding with the period of germband extension after
206 gastrulation and germband retraction, there were markedly fewer DEGs in both species than in
207 the following ~1 hour of developmental time between stages 12 (germband retraction) and 13
208 (beginning of dorsal closure), when both species showed a strikingly dynamic shift in gene
209 expression (Figure 3B), albeit to a different extent. Between stages 12 and 13, *D. melanogaster*
210 had 1,463 DEGs compared to *M. abdita's* 564 DEGs (Figure 3B), indicating sharper increases
211 and decreases in gene expression with the onset of dorsal closure in *D. melanogaster* than in
212 *M. abdita*. The increased turning on and off of genes between stages 12 and 13 in both species
213 also coincides with the phylotypic stage of development when evolutionary divergence between
214 species is most constrained (Sander 1983; Kalinka et al., 2010), suggesting that dynamic
215 changes in gene expression impose evolutionary constraints on gene regulatory networks.

216 To analyze the expression patterns of DEGs, we clustered them using DEGreports and plotted
217 the expression of each cluster (Figure S9) (Pantano 2024). The clusters confirmed that the most
218 dynamic transcriptional shift occurred during the transitions between stages 12 to 13 for both
219 species (Figure 3B). We then assessed whether the genes expressed in each cluster were
220 significantly enriched for any Biological Process Gene Ontology (BP GO) terms using the
221 Search Tool for the Retrieval of Interacting Genes/Proteins (String) (Szklarczyk et al. 2023)
222 (Figure S9). *M. abdita*'s clusters 4 and 9 and *D. melanogaster's* clusters 6 and 10 were
223 significantly enriched for embryonic, body plan, and systems development terms (Figure 4A &
224 S10A) and contained many well-characterized developmental genes in Drosophila (Figure
225 S10B&C). Genes found in *M. abdita* cluster 1 and *D. melanogaster* clusters 2 and 11, which
226 exhibited a pronounced increase in expression at stage 13, were enriched for terms related to
227 nervous system, muscle, and cuticle development (Figure 4B). In addition to these similarities,
228 *D. melanogaster's* more extreme expression changes between stages 12 and 13 (Figure 4 A-C)
229 could reflect the fact that Drosophila employs a specialized tissue (amnioserosa) and
230 evolutionarily novel mechanisms for dorsal closure (Schmidt-Ott and Kwan, 2022).

231 **Investigation of 'orphan' genes exclusively expressed during embryogenesis in *M. abdita***
232 **reveals novel F-Box LRR gene**

233 As the evolution of new genes within lineages and species is an important mechanism of
234 diversification of developmental mechanisms (Chen et al., 2010; Chen et al. 2013), we searched
235 *M. abdita's* genome for 'orphan' genes. Orphan genes are either highly diverged from known
236 sequences or represent newly evolved genes (Vakirlis et al., 2020; Xia et al., in press). In *M.*
237 *abdita's* genome annotation, eggnog-mapper was unable to assign orthologs to 1,049 gene
238 models. Further searches using *blastn* on the coding sequences and *blastp* on the predicted

239 protein sequences did not reveal any sequence similarity to other Dipterans. Approximately
240 8.7% of *M. abdita's* genes were 'orphan' genes.

241 We found that approximately 10% (109/1,049) of these orphan genes were expressed in
242 embryos, including 28 that were expressed *exclusively* during embryogenesis (Figure 5A). Most
243 of these genes (24) exhibited sharp expression peaks, with 14 peaking at stage 5
244 (cellularization), 4 at stage 12 (germband retraction), and smaller groups peaking at stages 13,
245 15, and 17. Four genes showed broader expression patterns, with two highly expressed during
246 stages 8-10 (germband elongation) and two across stages 5-12 (cellularization to germband
247 retraction).

248 To further characterize these 28 embryo-specific genes, we examined their open reading
249 frames and protein sequences (Table S3). We classified 21 as likely coding and of those we
250 further classified 9 as likely encoding stable proteins (blue highlighted rows in Table S3).
251 Additional searches of InterProScan's database revealed no known protein domains or any
252 domains consistent with known transposases (Jones et al., 2014). We then used Alphafold2 to
253 generate protein structure models (Jumper et al., 2021). Two of these genes resulted in
254 confident structures (Table S3, pTM > 0.5). One of them could not be related to any known
255 protein (Figure 5B); however, the second showed significant structural similarity to F-box
256 leucine-rich repeat (LRR) proteins (Figure 5C).

257 We then compared our F-box LRR orphan protein sequence to known F-box-LRR genes in *D.*
258 *melanogaster* but found no obvious ortholog. We identified orthologs to *D. melanogaster* F-box
259 LRR genes *Ppa*, *Kdm2*, *Fbxl4*, *Fbxl7*, *Fbl6*, CG32085, CG9003, and CG8272. In total, we
260 identified 15 F-box LRR genes in *M. abdita (*8 with clear *D. melanogaster* orthologs, 6 with
261 orthologs in other dipteran species, and our orphan). One of the dipteran F-box LRR orthologs
262 in *M. abdita* (Scaffold 3, geneID #3336) had an identical expression profile (stage 5) to the
263 orphan F-box LRR.

264 F-box LRRs are components of the 'E3 ubiquitin ligase SCF complex' which ubiquitinates
265 targeted proteins for later degradation by the cell. Specifically, the F-box domain binds to *Skp1*
266 and the LRR domain binds to the target, the molecule slated for ubiquitination and degradation.
267 F-box LRRs are known to ubiquinate important developmental signaling molecules in *D.*
268 *melanogaster* including the pair-rule gene *paired* (*prd*) which is bound by the F-box LRR protein
269 *Partner of paired* (*Ppa*) (Raj et al., 2000). *Ppa* is unusual in that its expression is patterned
270 rather than uniform as most F-Box LRR genes seem to be in *D. melanogaster* embryogenesis
271 (Das et al., 2002). Given that in *D. melanogaster*, the 12 best-known F-box LRR proteins (*Skp2*,
272 *Ppa*, *Kdm2*, *FipoQ*, *Fbxl4*, *Fbxl7*, *Fbl6*, CG32085, CG13766, CG12402, CG9003, CG8272) are
273 expressed throughout embryogenesis according to both our RNAseq data and ENCODE gene
274 expression data, the stage-restricted expression of *M. abdita*'s orphan F-box LRR gene and 13
275 other orphan genes with expression peaking before gastrulation might reflect previously
276 overlooked developmental differences between *M. abdita* and *D. melanogaster* at the
277 blastoderm stage.

278

Conclusions

The scuttle fly *M. abdita* is an important non-traditional model organism with hitherto very limited genomic resources. We have filled this gap by providing genomic and transcriptomic resources. By assembling a chromosome-level genome and annotating it, we revealed substantial chromosomal synteny with *Drosophila melanogaster* while uncovering many orphan genes. Additionally, our comparative transcriptome analysis across embryogenesis highlights conserved and divergent regulatory dynamics. The discovery of a novel F-box LRR gene, expressed exclusively during embryogenesis, underscores the potential of *M. abdita* to reveal new insights into the evolution of developmental gene networks. Ultimately, we hope the addition of these resources will further comparative research of developmental mechanisms in Diptera and continue the development of Diptera as a model taxon more broadly.

Materials and Methods

Generation of inbred *M. abdita* line

To generate single crosses of *M. abdita*, we collected pupae at the end of the pupal stage, when pupae darken approximately 1–2 days before eclosion, and transferred them to 35 x 10 mm petri dishes (Fisher Scientific, Cat. No. 50-820-644) until hatching. We monitored the plates every morning and every two hours during the day to collect virgin females. We paired each virgin female with a single male fly and allowed them to mate for two days in 35 x 10 mm petri dishes containing a gel solution made from 2% agar in water. On the third day, we prepared egg-laying vials by boiling 0.8 g of a fish food mixture—composed of 1 part spirulina flakes (Aquatic Eco-Systems Inc., Cat. No. ZSF5) and 2 parts sinking powder (Aquatic Eco-Systems Inc., Cat. No. F1A)—in 10 mL of a 0.8% agar solution (EMD Millipore, Cat. No. 1.01614). We added approximately 1 mL of this solution to 10 x 75 mm culture tubes (Fisher Scientific, Cat. No. 14-961-25) and allowed it to solidify. After cooling, we added 0.1 g of fish food on top using weighing paper, followed by 200 µL of water. We used a cotton swab to compact the food and clean any excess moisture from the sides of the tubes. We then transferred the mating pairs from the agar plates to the culture tubes and plugged the tubes with rayon balls (TIDI, Cat. No. 969162). See supplemental figure S11 for visual diagram of this process. We established multiple single crosses and tracked them using a progressive hierarchical code to identify the lineage. We conducted nine generations of sibling × sibling single-pair matings across three separate parallel lineages (A, B, and C). At generation six, we generated pooled crosses within each lineage to allow for the mixing of potentially lethal recessive alleles that may have accumulated during the single-cross procedure. Following this, we maintained the B lineage, as it exhibited higher overall fertility and health.

1.1 – 1.3 Genome Assembly

1.1 de novo library preparation, sequencing, and assembly

We generated a *de novo* reference genome for *M. abdita* with sequencing data from HiFi PacBio reads and Dovetail's OmniC libraries (Cantata Bio). For HiFi PacBio sequencing, we collected and snap-froze in liquid nitrogen ~600 dechorionated embryos (mostly stages 16 and

318 17) from a 10-generation inbred line of a previously established laboratory culture of *M. abdita*
319 Schmitz, 1959 (Schmidt-Ott et al., 1994). We sent these samples to Dovetail Genomics for HiFi
320 PacBio library preparation and sequencing. Library preparation circularizes fragments so they
321 can be read many times to generate a high-fidelity consensus sequence. Sequencing on the
322 SMRT (Single Molecule, Real-Time) nanofluidic chip generates the long reads inherent to
323 PacBio. PacBio generated 25.4 gigabase-pairs reads (42x coverage) from which Dovetail
324 generated a haplotype-resolved draft assembly using the Hifiasm assembler (Hifiasm1 v0.15.4-
325 r347 with default parameters) (Cheng et al. 2021). We used blobtools v1.1.1 to identify possible
326 contamination and removed 33 scaffolds from the assembly (Challis et al., 2020; Laetsch and
327 Blaxter, 2017). After filtering out haplotigs and contig overlaps with purge_dups v1.2.5, 89
328 scaffolds remained (Guan et al., 2020).

329 *1.2 Omni-C library preparation and sequencing*

330 We collected and snap-froze an additional ~600-700 dechorionated embryos and ~60 young
331 pupae in liquid nitrogen from the same inbred line of *M. abdita* for Omni-C Library preparation
332 and sequencing. Omni-C is a proprietary technology for long-range proximity ligation and
333 sequencing of genomic libraries which captures spatial information within the genome through
334 chromatin fixation and sequencing. Omni-C differs from other Hi-C preparations in that it digests
335 the chromatin with a sequence-independent endonuclease. This eliminates biases inherent to
336 competitive restriction enzyme-based approaches. Samples were treated with formaldehyde to
337 fix the chromatin and then digested with DNAse I. The resulting ends were then repaired, and
338 biotinylated bridge adapters were ligated to the ends. Subsequent steps involved proximity
339 ligation of adapter-ligated ends, reversal of formaldehyde-induced crosslinks, and DNA
340 purification. Non-internally ligated biotin residues were removed from the purified DNA.
341 Sequencing libraries were prepared with NEBNext Ultra enzymes and Illumina-compatible
342 adapters. Before PCR enrichment, biotin-containing fragments were isolated using streptavidin
343 beads. Sequencing was performed on the Illumina HiSeqX platform.

344 *1.3 Scaffolding assembly with HiRise*

345 Both the Hifiasm draft assembly and Dovetail OmniC sequencing reads were used as input for
346 HiRise, a software tailored for scaffolding genome assemblies with proximity ligation data
347 (Putnam et al., 2016). Based on spatial data from the chromatin conformation capture (Figure
348 S1), HiRise pinpoints regions where contigs are joined incorrectly (misjoins) in the initial
349 assembly and utilizes this spatial information to re-orient contigs and construct larger scaffolds.
350 The OmniC library sequences were aligned to the draft assembly using the Burrows-Wheeler
351 Aligner (Li and Durbin, 2009). HiRise-analyzed read pairs are mapped to the draft scaffolds to
352 develop a genomic distance likelihood model. This model is then used to identify and break
353 potential misjoins, score prospective joins, and execute joins surpassing a set threshold. These
354 scaffolds consist of sequentially arranged contigs separated by gaps. We used QUAST to
355 calculate %GC, N50, L50, and Ns per 100 kbp (Gurevich et al., 2014). We then repeat-masked
356 the genome with RepeatMasker (Smit et al., 2013). To estimate heterozygosity and sequencing
357 error rates, we calculated the frequency spectrum of canonical 21-mers using jellyfish, and input
358 the resulting histograms into GenomeScope (Marçais and Kingsford, 2011; Vurture et al., 2017).

Using NUCmer (mummer v3.23 software package), we aligned the non-chromosome size scaffolds (4 – 15) back to the three chromosome-size scaffolds (1 - 3) and found repetitive sequences present in scaffolds 4 – 15 (Kurtz et al., 2004; Marçais et al., 2018).

2.1 – 2.3 Genome Annotation

2.1 Repeat masking the genome

To generate a custom library for repeat masking, we ran RepeatModeler v2.0.4 on the genome to find/model potential repeats (Smit and Hubley, 2008). We then used RepeatMasker's script fambd.py v0.4.3 to extract Arthropoda records from the dfam database and combined these sequences with the RepeatModeler output to use as the repeat library. We used RepeatMasker v4.1.5 to soft-masked the genome with our custom library of repetitive low-complexity DNA and transposable elements (Smit et al., 2013).

2.2 Data used as evidence of genome features

We downloaded available Illumina RNAseq data for *M. abdita* from NCBI which included paired end reads from 3 adults and pooled embryos (Table S4). We generated RNAseq data for 9 precisely staged embryos, first and third instar larval stages, and a 1-day-old pupal stage (Table S4). For protein evidence, we downloaded all Dipteran protein sequences from NCBI's RefSeq database which included 2,122,027 sequences from 617 species (Sayer et al., 2022). We also downloaded Uniprot's complete protein sequence file (UniProt Release 2023_04) which contains 570,157 sequences from 14,509 species (The UniProt Consortium, 2023).

2.3 Annotation pipeline

We annotated the reference genome of *M. abdita* using evidence from all RNAseq transcripts (Table S4), protein sequences, and gene prediction software (Figure S3). The choices of software used in this pipeline are based on the methods section of VanKuren's 'Draft *Papilio alphenor* assembly and annotation' (VanKuren 2023). First, we assembled RNA transcripts using two transcript assemblers: Stringtie v2.2.1 and Trinity v2.15.1 (Pertea et al, 2015; Grabherr et al., 2011; Haas et al., 2013). Trinity performs both genome-guided and *de novo* assembly; we assembled transcripts with both methods. For Trinity and Stringtie's genome-guided assemblies, we first mapped reads to the genome with 'Spliced Transcripts Alignment to a Reference' software (STAR v2.7.10b) (Dobin et al., 2013). Next, we used the 'Program to Assemble Spliced Alignments' (PASA v2.5.3) to identify gene structures from all three assemblies (Haas et al., 2003). We predicted gene models directly from the PASA assemblies using the PASA plug-in TransDecoder v5.7.1 which identifies candidate coding regions from Trinity and StringTie assemblies. To create protein alignments, we used the software Exonerate v2.2.0 which maps protein sequences to the genome (Slater and Birney, 2005). We used the BRAKER3 pipeline (braker.pl v3.0.6) to predict gene models from mapped RNAseq reads and protein data (Gabriel et al., 2023). BRAKER3 relies on the software Augustus and GeneMark to predict gene models (Stanke et al., 2006; Brůna et al., 2020). We also generated our own *ab initio* gene structure predictions, using GlimmerHMM v3.0.4 (Majoros et al., 2004). First, we collected "hints" for training the *ab initio* predictors by extracting protein-coding hints from the

398 protein alignments using Augustus' exonerate2hints function, intron hints from mapped RNA
399 seq reads using Augustus' bam2hints function, and exon/intron hints from the PASA assemblies
400 using Augustus' bam2exonhints function. We additionally used the coding predictions from
401 Transdecoder to create training models and further refined those models using the
402 lib.selectTrainingModels function from Funannotate (Palmer and Stajich, 2019). This training
403 data was used to run GlimmerHMM with hints as guidance (Majoros et al., 2004). We then
404 provided the PASA assemblies, mapped protein data, TransDecoder predictions, *ab initio*
405 predictions, BRAKER3 predictions, and a file weighting each line of evidence to the software
406 Evidence Modeler v2.1.0. Evidence Modeler constructed the consensus gene structures which
407 were updated by PASA to add UTRs and identify alternative transcripts (Haas et al., 2008). We
408 removed gene structures that overlapped with RepeatMasker output using a Funannotate
409 function called "RemoveBadModels." We then used Funannotate v1.8.1 to identify annotations
410 that match known transposable elements (TEs) and repeat proteins using BLAST and updated
411 the annotation to remove remaining TEs. We used AGAT v1.4.1 to remove genes with an open
412 reading frame (ORF) < 100 amino acids in length and any associated gene structures (Dainat,
413 2022). We assigned gene names using eggNOG-mapper v2.1.12 which relies on orthology
414 predictions to functionally annotate genes (Cantalapiedra et al., 2021; Huerta-Cepas et al.,
415 2019).

416

417 **Synteny Analysis**

418 We used MCScanX (primary release) to compare the synteny of *M. abdita* and *D. melanogaster*
419 genomes (Wang et al., 2012). MCScanX identifies syntenic blocks based on a score given to
420 each gene pair. We set the match_score = 50 (default), match size = 5 (number of genes
421 required to constitute a syntenic block), gap_pentaly = 0 (no penalty for gaps), and max_gaps =
422 100. We used the output of this run to generate our synteny map. We used the circlize R
423 package to plot the results (Gu et al., 2014).

424

425 **Embryo Staging for RNAseq**

426 For precise embryo staging, embryos of the appropriate age were mounted on a microscope
427 slide under halocarbon 27 oil and observed in a Zeiss Axiophot compound microscope
428 equipped with a 10x objective and DIC (differential interference contrast) optics until they
429 reached the desired stage. They were photographed and immediately processed for RNA
430 extraction as previously described (Lott et al., 2014). We collected *M. abdita* and stage-matched
431 *D. melanogaster* embryos at stages 1, 5, 8, 9, 10, 12, 13, 15, 16, and 17. Photos of the
432 sequenced embryos can be found in Figure S5.

433

434 **RNA isolation and Sequencing**

435 We incubated each sample for 5-10 minutes at room temperature in TRIzol. We then froze each
436 sample in TRIzol at -80° C. We extracted total RNA using the TRIzol/Phenol-chloroform protocol
437 detailed in the appendix (Protocol 10) of 'Functional evolution of a morphogenetic gradient' by
438 Chun Wai Kwan. The University of Chicago genomics core facility constructed cDNA libraries
439 using the TruSeq kit with PCR (Illumina, CA, USA). The cDNA libraries were barcoded and
440 multiplexed for 100bp paired-end sequencing on one lane of a HiSeq Illumina 2000 sequencer.

Differential Gene Expression Analysis

M. abdita had ~2x the reads for each sample compared to *D. melanogaster* (Figure S6A). We chose not to down sample *M. abdita* reads to match *D. melanogaster* to avoid losing power to detect changes between genes within *M. abdita* developmental stages. We justified this by looking for any evidence that the higher number of reads skewed the relationship between gene expression and variance, but found that the relationship between mean expression and variance in expression is similar in both species without subsetting the data (Figure S7). Additionally, the average, median, and variance in gene expression (TPM) were very similar across development in both species (Figure S8).

We aligned *M. abdita* RNAseq reads to the reference genome generated in this publication and *D. melanogaster* RNAseq reads to *D. melanogaster*'s genome (Accession: GCF_000001215.4). We used Subread v2.0.5 to align reads (Liao et al., 2013). 90-99% of reads mapped to the genome for each sample. We input the aligned bam files into Subread's featureCount function which assigns the reads to a genomic feature from the annotation file (gff). At this point, we performed the analysis in Rstudio using the free and open source statistical language R and various R packages including EdgeR, tidyr, dplyr, and ggplot2 (R Core Team, 2023; Robinson et al., 2009; Posit Team, 2024; Wickham et al. 2016; Wickham et al., 2023; Wickham et al., 2024). We used EdgeR to assess gene expression differences between samples. We filtered out lowly expressed genes and normalized data by library size (TMM normalization) for both species. We calculated counts per million (CPM) to normalize the difference in raw reads between samples and then calculated transcript per million (TPM) to account for differences in gene length. We calculated gene length as the coding sequence length for each gene's longest isoform. All TPM expression data are plotted as $\log_2(\text{TPM} + 1)$.

As our RNA-seq samples do not include biological replicates, we estimated the squared coefficient of variation (BCV) using k-means clustering of the samples. We began by calculating a distance matrix for the samples and extracting the first four eigenvectors, which together explained >95% of the variance. To determine the optimal number of clusters (k), we calculated the within-cluster sum of squared errors (WSS) and Silhouette scores for k values ranging from 1 to 8, selecting k = 5 based on these metrics. For *M. abdita*, clustering with k = 5 grouped the embryonic stages as follows: Group 1 (stage 1), Group 2 (stage 5), Group 3 (stages 8–12), Group 4 (stages 13–15), and Group 5 (stage 17). For *D. melanogaster*, the clusters were: Group 1 (stage 1), Group 2 (stages 5 and 8), Group 3 (stages 9–12), Group 4 (stages 13–16), and Group 5 (stage 17). Using these groupings, we estimated the dispersion as $\sigma=0.36$ for both *M. abdita* and *D. melanogaster.* These estimates were applied globally. Differential expression analysis was performed using EdgeR's exactTest() function, comparing gene expression between pairwise embryonic stages rather than the k-means groups. Genes were considered differentially expressed if they met the criteria: fold change < -3 or > 3 and p-value < 0.001. We then took all differentially expressed genes (DEGs) and input that expression data into DEGreports pattern() function. DEGreports clusters DEGs based on expression profile similarity (Pantano 2024). We used these clusters for gene ontology enrichment analysis described below.

484 **Gene Ontology Enrichment Analysis**
485 We used the STRING database v12.0 (Search Tool for the Retrieval of Interacting
486 Genes/Proteins) to perform enrichment analysis of Biological Process Gene Ontology (GO)
487 terms for our clustered DEG lists (Szklarczyk et al., 2023). STRING compares input gene sets
488 to a reference genome to identify networks of interacting genes and enrichments in biological
489 processes. Since *M. abdita* is not available in STRING, we first used NCBI's Blast tool to select
490 the top *D. melanogaster* ortholog match/hit for each *M. abdita* gene (Altschul et al., 1990). Using
491 FlyBase's batch download tool, we retrieved the corresponding FlyBase IDs, which STRING
492 accepts as input (Öztürk-Çolak et al., 2024). STRING performed the enrichment analysis
493 identifying the Biological Process GO terms significantly associated with each developmental
494 stage. STRING measures enrichment based on the strength of enrichment
495 (Log_{10}(observed/expected)), false discovery rate (p-values corrected for multiple testing with
496 Benjamini-Hochberg), and the signal (weighted harmonic mean between observed/expected
497 ratio and -log(FDR).
498
499 **'Orphan' gene identification**
500 We identified genes from *M. abdita*'s annotation file for which eggNOG-mapper could not assign
501 an ortholog. Next, we assessed the expression of these genes in our RNA-seq data. To identify
502 genes with exclusively embryonic expression, we focused on those with a CPM > 1 in at least
503 one of the nine embryonic stages and a CPM < 1 in pupal, larval, and adult stages. We
504 validated these genes further by analyzing their ORFs using NCBI's 'Open reading frame
505 finder'. We used CPC2 to evaluate the nucleic acid sequences to assess coding potential (Kang
506 et al., 2017). Finally, we performed sequence similarity searches using NCBI's blastn and blastp
507 tools, querying both nucleotide and protein sequences against the entire NCBI database as well
508 as against Dipteran-specific sequences (Altschul et al., 1990; Sayer et al., 2022).
509
510 **Protein structure prediction and structural similarity search**
511 We used Protparam to assess protein stability, aliphatic index, and hydropathicity of predicted
512 proteins (Gasteiger et al., 2005). We then searched all predicted protein sequences against
513 InterProScan to look for any missed domains, specifically to look for evidence of transposable
514 elements that were not discovered with blast searches (Jones et al., 2014). We used the
515 AlphaFold2.ipynb provided by ColabFold v1.5.5 to predict protein structures (Jumper et al.,
516 2021; Mirdita et al., 2022;). If the predicted protein had a predicted template modeling (pTM)
517 score > 0.5 we then uploaded the structure to FoldSeek's website and searched the available
518 databases (AlphaFold/Proteome, AlphaFold/Swiss-Prot, AlphaFold/UniProt50, BFMD) for
519 proteins with similar structure (van Kempen et al., 2024; Varadi et al., 2022; Varadi et al., 2024).
520 We used Protein Imager to generate publication-quality images of protein structures (Tomasello
521 et al., 2020).

522 **Acknowledgments**

523 We thank Lily Shiue and Qianyu Jin at Cantata Bio for assembling the genome. We thank
524 Nicholas VanKuren for his notes on genome annotation. We thank Yaikhomba Mutum for the
525 conversation on protein structure prediction and structural similarity searches. We thank Xiang-
526 Ru (Shannon) Xu for optimization of the *M. abdita* single cross procedure and the drawings in

Figure S11. We thank the University of Chicago's Center for Research Informatics and Research Computing Center for hosting and maintaining the high-performance computing clusters used during this work.

JETSTREAM2: This work used Jetstream2 at Indiana University through allocation BIO220075 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296.

GBCF: Bioinformatic work was supported in part by the Notre Dame University Genomics and Bioinformatics Core Facility.

Competing interests

The authors declare no competing interests.

Funding

Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R01GM127366 and the University of Chicago. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data and Resource Availability

Genome: The annotated *Megaselia abdita* genome can be found in NCBI's Genome database under BioProject Accession PRJNA1164289.
RNASeq data: All fastq files containing the raw sequencing reads for each sample have been uploaded to NCBI's Sequence Read Archive (SRA) and can be found under the BioProject Accession PRJNA1200075. Individual BioSample and SRA accession numbers can be found in supplementary table S4 including those samples that were not generated in this study but used as evidence for the annotation of the genome.
Genome Browser: Jetstream2 at Indiana University is a resource provider for NSF's ACCESS program which aims to broaden access to super computing resources at no cost to researchers. To access the *Megaselia abdita* genome browser and related tools, create an ACCESS ID, then use this ID create an account and login to Jetstream2. You will apply for an 'allocation' of credits which can be used on Jetstream2. Detailed instructions on the use of Jetstream2 can be found at https://jetstream-cloud.org/get-started/index.html

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology* 215, 403–410.

Beeber, D. and Chain, F. J. (2020). crispRdesignR: A Versatile Guide RNA Design Package in R for CRISPR/Cas9 Applications. *J Genomics* **8**, 62–70.

Boerner, T. J., Deems, S., Furlani, T. R., Knuth, S. L. and Towns, J. (2023). ACCESS: Advancing Innovation: NSF's Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support. In *Practice and Experience in Advanced Research Computing 2023: Computing for the Common Good*, pp. 173–176. New York, NY, USA: Association for Computing Machinery.

Brooks, E. M., Sanders, S. A. and Pfrender, M. E. (2024). freeCount: A Coding Free Framework for Guided Count Data Visualization and Analysis. In *Practice and Experience in Advanced Research Computing 2024: Human Powered Computing*, pp. 1–4. New York, NY, USA: Association for Computing Machinery.

Brůna, T., Lomsadze, A. and Borodovsky, M. (2020). GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. *NAR Genomics and Bioinformatics* **2**, lqaa026.

Bullock, S. L., Stauber, M., Prell, A., Hughes, J. R., Ish-Horowicz, D. and Schmidt-Ott, U. (2004). Differential cytoplasmic mRNA localisation adjusts pair-rule transcription factor activity to cytoarchitecture in dipteran evolution. *Development* **131**, 4251–4261.

Campos-Ortega, J. A. and Hartenstein, V. (1997). *The Embryonic Development of Drosophila melanogaster*. Berlin, Heidelberg: Springer.

Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. and Huerta-Cepas, J. (2021). eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular Biology and Evolution* **38**, 5825–5829.

Caroti, F., Urbansky, S., Wosch, M. and Lemke, S. (2015). Germ line transformation and in vivo labeling of nuclei in Diptera: report on *Megaselia abdita* (Phoridae) and *Chironomus riparius* (Chironomidae). *Development Genes and Evolution* **225**, 179.

Caroti, F., González Avalos, E., Noeske, V., González Avalos, P., Kromm, D., Wosch, M., Schütz, L., Hufnagel, L. and Lemke, S. (2018). Decoupling from yolk sac is required for extraembryonic tissue spreading in the scuttle fly *Megaselia abdita*. *eLife* **7**, e34616.

591 **Challis, R., Richards, E., Rajan, J., Cochrane, G. and Blaxter, M.** (2020). BlobToolKit –
592 Interactive Quality Assessment of Genome Assemblies. *G3: Genes|Genomes|Genetics*
593 **10**, 1361.

594 **Chen, S., Zhang, Y. E. and Long, M.** (2010). New Genes in Drosophila Quickly Become
595 Essential. *Science* **330**, 1682–1685.

596 **Chen, S., Krinsky, B. H. and Long, M.** (2013). New genes as drivers of phenotypic
597 evolution. *Nat Rev Genet* **14**, 645–660.

598 **Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. and Li, H.** (2021). Haplotype-resolved
599 de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* **18**, 170–
600 175.

601 **Crombach, A. and Jaeger, J.** (2012). Life's attractors: understanding developmental
602 systems through reverse engineering and in silico evolution. *Adv Exp Med Biol* **751**,
603 93–119.

604 **Crombach, A., Wotton, K. R., Jiménez-Guri, E. and Jaeger, J.** (2016). Gap Gene
605 Regulatory Dynamics Evolve along a Genotype Network. *Molecular Biology and*
606 *Evolution* **33**, 1293–1307.

607 **Dainat J. 2022.** Another Gtf/Gff Analysis Toolkit (AGAT): Resolve interoperability issues
608 and accomplish more with your annotations. Plant and Animal Genome XXIX
609 Conference. https://github.com/NBISweden/AGAT.

610 **Das, T., Purkayastha-Mukherjee, C., D'Angelo, J. and Weir, M.** (2002). A conserved F-
611 box gene with unusual transcript localization. Dev Genes Evol 212, 134–140.

612 **Diesh, C., Stevens, G. J., Xie, P., De Jesus Martinez, T., Hershberg, E. A., Leung, A.,**
613 **Guo, E., Dider, S., Zhang, J., Bridge, C., et al.** (2023). JBrowse 2: a modular genome
614 browser with views of synteny and structural variation. *Genome Biology* **24**, 74.

615 **Dey, B., Kaul, V., Kale, G., Scorcelletti, M., Takeda, M., Wang, Y.-C. and Lemke, S.**
616 (2023). Divergent evolutionary strategies preempt tissue collision in fly gastrulation.
617 2023.10.09.561568.

618 **Disney, R. H. L.** (1994). *Scuttle Flies: The Phoridae*. Dordrecht: Springer Netherlands.

619 **Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P.,**
620 **Chaisson, M. and Gingeras, T. R.** (2013). STAR: ultrafast universal RNA-seq aligner.
621 *Bioinformatics* **29**, 15–21.

622 **Feng, D., Li, J. and Liu, G.** (2020). The complete mitochondrial genomes of two scuttle
623 flies, *Megaselia spiracularis* and *Dohrniphora cornuta* (Diptera: Phoridae).
624 *Mitochondrial DNA B Resour* **5**, 1208–1209.

625 **Fraire-Zamora, J. J., Jaeger, J. and Solon, J.** (2018). Two consecutive microtubule-
626 based epithelial seaming events mediate dorsal closure in the scuttle fly *Megaselia*
627 *abdita*. *eLife* **7**, e33807.

628 **Gabriel, L., Brůna, T., Hoff, K. J., Ebel, M., Lomsadze, A., Borodovsky, M. and Stanke,**
629 **M.** (2023). BRAKER3: Fully automated genome annotation using RNA-Seq and protein
630 evidence with GeneMark-ETP, AUGUSTUS and TSEBRA. *bioRxiv*
631 2023.06.10.544449.

632 **Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M., Appel, R. and**
633 **Bairoch, A.** (2005). Protein Identification and Analysis Tools on the Expasy Server. In
634 *The Proteomics Protocols Handbook*, pp. 571–607.

635 **Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I.,**
636 **Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al.** (2011). Full-length
637 transcriptome assembly from RNA-Seq data without a reference genome. *Nat*
638 *Biotechnol* **29**, 644–652.

639 **Grimaldi, D. and Engel, M.** (2005). *Evolution of Insects*. Cambridge University Press.

640 **Gu, Z., Gu, L., Eils, R., Schlesner, M., and Brors, B.** (2014)**.** Circlize implements and
641 enhances circular visualization in R. *Bioinformatics*, **30**, 2811-2812**.**

642 **Guan, D., McCarthy, S. A., Wood, J., Howe, K., Wang, Y. and Durbin, R.** (2020).
643 Identifying and removing haplotypic duplication in primary genome assemblies.
644 *Bioinformatics* **36**, 2896–2898.

645 **Gurevich, A., Saveliev, V., Vyahhi, N. and Tesler, G.** (2013). QUAST: quality assessment
646 tool for genome assemblies. *Bioinformatics* **29**, 1072–1075.

647 **Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith Jr, R. K., Hannick, L. I.,**
648 **Maiti, R., Ronning, C. M., Rusch, D. B., Town, C. D., et al.** (2003). Improving the

649 Arabidopsis genome annotation using maximal transcript alignment assemblies.
650 *Nucleic Acids Research* **31**, 5654–5666.

651 **Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., White, O., Buell,**
652 **C. R. and Wortman, J. R.** (2008). Automated eukaryotic gene structure annotation
653 using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome*
654 *Biology* **9**, R7.

655 **Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J.,**
656 **Couger, M. B., Eccles, D., Li, B., Lieber, M., et al.** (2013). De novo transcript
657 sequence reconstruction from RNA-Seq: reference generation and analysis with
658 Trinity. *Nat Protoc* **8**, 10.1038/nprot.2013.084.

659 **Hancock, D. Y., Fischer, J., Lowe, J. M., Snapp-Childs, W., Pierce, M., Marru, S.,**
660 **Coulter, J. E., Vaughn, M., Beck, B., Merchant, N., et al.** (2021). Jetstream2:
661 Accelerating cloud computing via Jetstream. In *Practice and Experience in Advanced*
662 *Research Computing 2021: Evolution Across All Dimensions*, pp. 1–8. New York, NY,
663 USA: Association for Computing Machinery.

664 **Horn, T., Hilbrant, M. and Panfilio, K. A.** (2015). Evolution of epithelial morphogenesis:
665 phenotypic integration across multiple levels of biological organization. *Frontiers in*
666 *Genetics* **6**, 303.

667 **Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K.,**
668 **Cook, H., Mende, D. R., Letunic, I., Rattei, T., Jensen, L. J., et al.** (2019). eggNOG
669 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource
670 based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* **47**, D309–D314.

671 **Jiménez-Guri, E., Huerta-Cepas, J., Cozzuto, L., Wotton, K. R., Kang, H.,**
672 **Himmelbauer, H., Roma, G., Gabaldón, T. and Jaeger, J.** (2013). Comparative
673 transcriptomics of early dipteran development. *BMC Genomics* **14**, 123.

674 **Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H.,**
675 **Maslen, J., Mitchell, A., Nuka, G., et al.** (2014). InterProScan 5: genome-scale
676 protein function classification. *Bioinformatics* **30**, 1236–1240.

677 **Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O.,**
678 **Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al.** (2021). Highly
679 accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589.

680 **Kalinka, A. T., Varga, K. M., Gerrard, D. T., Preibisch, S., Corcoran, D. L., Jarrells, J.,**
681 **Ohler, U., Bergman, C. M. and Tomancak, P.** (2010). Gene expression divergence
682 recapitulates the developmental hourglass model. *Nature* **468**, 811–814.

683 **Kang, Y.-J., Yang, D.-C., Kong, L., Hou, M., Meng, Y.-Q., Wei, L. and Gao, G.** (2017).
684 CPC2: a fast and accurate coding potential calculator based on sequence intrinsic
685 features. *Nucleic Acids Research* **45**, W12–W16.

686 **Kimura, M. and Ohta, T.** (1974). On Some Principles Governing Molecular Evolution.
687 *Proceedings of the National Academy of Sciences of the United States of America* **71**,
688 2848.

689 **Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C. and**
690 **Salzberg, S. L.** (2004). Versatile and open software for comparing large genomes.
691 *Genome Biol* **5**, R12.

692 **Kwan, C. W., Gavin-Smyth, J., Ferguson, E. L. and Schmidt-Ott, U.** (2016). Functional
693 evolution of a morphogenetic gradient. *eLife* **5**, e20894.

694 **Kwan, C.W.** (2017). Functional evolution of a morphogenetic gradient. [Doctoral
695 dissertation,The University of Chicago]. https://doi.org/10.6082/M1H41PH7

696 **Li, H. and Durbin, R.** (2009). Fast and accurate short read alignment with Burrows–
697 Wheeler transform. *Bioinformatics* **25**, 1754–1760.

698 **Li, X., Hash, J. M., Hartop, E., Yang, D., Smith, P. T. and Brown, B. V.** (2024) A
699 molecular phylogeny of scuttle flies (Diptera: Phoridae) unveils extensive concordance
700 but intriguing divergences from morphological results. *Systematic Entomology*.

701 **Liao, Y., Smyth, G. K. and Shi, W.** (2013). The Subread aligner: fast, accurate and
702 scalable read mapping by seed-and-vote. *Nucleic Acids Research* **41**, e108.

703 **Liu, Q., Onal, P., Datta, R. R., Rogers, J. M., Schmidt-Ott, U., Bulyk, M. L., Small, S.**
704 **and Thornton, J. W.** (2018). Ancient mechanisms for the evolution of the bicoid
705 homeodomain's function in fly development. *eLife* **7**, e34594.

706 **Lomsadze, A., Burns, P. D. and Borodovsky, M.** (2014). Integration of mapped RNA-Seq
707 reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Res*
708 **42**, e119.

Lott, S. E., Villalta, J. E., Zhou, Q., Bachtrog, D. and Eisen, M. B. (2014). Sex-Specific Embryonic Gene Expression in Species with Newly Evolved Sex Chromosomes. *PLOS Genetics* **10**, e1004159.

Lynch, M. (2007a). The frailty of adaptive hypotheses for the origins of organismal complexity. *Proceedings of the National Academy of Sciences* **104**, 8597–8604.

Lynch, M. (2007b). *The origins of genome architecture*. Sunderland, MA: Sinauer Associates, Inc.

Majoros, W. H., Pertea, M. and Salzberg, S. L. (2004). TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879.

Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. and Zdobnov, E. M. (2021). BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol Biol Evol* **38**, 4647–4654.

Marçais, G. and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770.

Marçais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S. L. and Zimin, A. (2018). MUMmer4: A fast and versatile genome alignment system. *PLOS Computational Biology* **14**, e1005944.

Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S. and Steinegger, M. (2022). ColabFold: making protein folding accessible to all. *Nat Methods* **19**, 679–682.

Mulhair, P. O. and Holland, P. W. H. (2024). Evolution of the insect Hox gene cluster: Comparative analysis across 243 species. *Seminars in Cell & Developmental Biology* **152–153**, 4–15.

Öztürk-Çolak, A., Marygold, S. J., Antonazzo, G., Attrill, H., Goutte-Gattat, D., Jenkins, V. K., Matthews, B. B., Millburn, G., dos Santos, G., Tabone, C. J., et al. (2024). FlyBase: updates to the Drosophila genes and genomes database. *Genetics* **227**, iyad211.

Palmer, J. and Stajich, J. (2019). nextgenusfs/funannotate: funannotate v1.5.3.

Pantano, L. (2024). *DEGreport: Report of DEG analysis*. R package version 1.42.0, http://lpantano.github.io/DEGreport/.

739 **Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T. and Salzberg,**
740 **S. L.** (2015). StringTie enables improved reconstruction of a transcriptome from RNA-
741 seq reads. *Nat Biotechnol* **33**, 290–295.

742 **Picard, C. J., Johnston, J. S. and Tarone, A. M.** (2012). Genome Sizes of Forensically
743 Relevant Diptera. *Journal of Medical Entomology* **49**, 192–197.

744 **Posit team**. (2024). RStudio: Integrated development environment for R. Posit Software,
745 PBC, Boston, MA. <http://www.posit.co/>

746 **Priyam, A., Woodcroft, B. J., Rai, V., Moghul, I., Munagala, A., Ter, F., Chowdhary, H.,**
747 **Pieniak, I., Maynard, L. J., Gibbins, M. A., et al.** (2019). Sequenceserver: A Modern
748 Graphical User Interface for Custom BLAST Databases. *Mol Biol Evol* **36**, 2922–2924.

749 **Putnam, N. H., O'Connell, B. L., Stites, J. C., Rice, B. J., Blanchette, M., Calef, R.,**
750 **Troll, C. J., Fields, A., Hartley, P. D., Sugnet, C. W., et al.** (2016). Chromosome-
751 scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res*
752 **26**, 342–350.

753 **R Core Team.** (2023) R: A Language and Environment for Statistical Computing. R
754 Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

755 **Rafiqi, Ab. M., Lemke, S., Ferguson, S., Stauber, M. and Schmidt-Ott, U.** (2008).
756 Evolutionary origin of the amnioserosa in cyclorrhaphan flies correlates with spatial and
757 temporal expression changes of zen. *Proceedings of the National Academy of*
758 *Sciences* **105**, 234–239.

759 **Rafiqi, A. M., Lemke, S. and Schmidt-Ott, U.** (2010). Postgastrular zen expression is
760 required to develop distinct amniotic and serosal epithelia in the scuttle fly *Megaselia*.
761 *Dev Biol* **341**, 282–290.

762 **Rafiqi, A. M., Lemke, S. and Schmidt-Ott, U.** (2011). The scuttle fly *Megaselia abdita*
763 (Phoridae): a link between Drosophila and Mosquito development. *Cold Spring Harb*
764 *Protoc* **2011**, pdb.emo143.

765 **Rafiqi, Ab. M., Park, C.-H., Kwan, C. W., Lemke, S. and Schmidt-Ott, U.** (2012). BMP-
766 dependent serosa and amnion specification in the scuttle fly *Megaselia abdita*.
767 *Development* **139**, 3373–3382.

768 **Raj, L., Vivekanand, P., Das, T. K., Badam, E., Fernandes, M., Jr, R. L. F., Brent, R.,**
769 **Appel, L. F., Hanes, S. D. and Weir, M.** (2000). Targeted localized degradation of
770 Paired protein in Drosophila development. *Current Biology* **10**, 1265–1272.

771 **Rasmussen, D. A. and Noor, M. A.** (2009). What can you do with 0.1× genome coverage?
772 A case study based on a genome survey of the scuttle fly *Megaselia scalaris*
773 (Phoridae). *BMC Genomics* **10**, 382.

774 **Robinson, M. D., McCarthy, D. J. and Smyth, G. K.** (2009). edgeR: a Bioconductor
775 package for differential expression analysis of digital gene expression data.
776 *Bioinformatics* **26**, 139.

777 **Rohr, K. B., Tautz, D. and Sander, K.** (1999). Segmentation gene expression in the
778 mothmidge Clogmia albipunctata (Diptera, Psychodidae) and other primitive dipterans.
779 *Dev Gene Evol* **209**, 145–154.

780 **Sander, K. (1983).** The evolution of patterning mechanisms: gleanings from insect
781 embryogenesis and spermatogenesis. In Development and Evolution (Eds. B. C.
782 Goodwin, N. Holder, C. C. Wylie), Cambridge University Press, Cambridge, 137-159.

783 **Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C.,**
784 **Connor, R., Funk, K., Kelly, C., Kim, S., et al.** (2021). Database resources of the
785 National Center for Biotechnology Information. *Nucleic Acids Res* **50**, D20–D26.

786 **Schmidt-Ott, U., Sander, K. and Technau, G. M.** (1994). Expression of engrailed in
787 embryos of a beetle and five dipteran species with special reference to the terminal
788 regions. *Rouxs Arch Dev Biol* **203**, 298–303.

789 **Schmidt-Ott, U. and Lynch, J. A.** (2016). Emerging developmental genetic model systems
790 in holometabolous insects. *Curr Opin Genet Dev* **39**, 116–128.

791 **Schmidt-Ott, U. and Kwan, C. W.** (2022). How two extraembryonic epithelia became one:
792 serosa and amnion features and functions of Drosophila's amnioserosa. *Philosophical*
793 *Transactions of the Royal Society B: Biological Sciences* **377**, 20210265.

794 **Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. and Zdobnov, E. M.**
795 (2015). BUSCO: assessing genome assembly and annotation completeness with
796 single-copy orthologs. *Bioinformatics* **31**, 3210–3212.

797 **Slater, G. S. C. and Birney, E.** (2005). Automated generation of heuristics for biological
798 sequence comparison. *BMC Bioinformatics* **6**, 31.

Smit, A. and Hubley, R. (2008). RepeatModeler.

Smit, A., Hubley, R. and Green, P. (2013). RepeatMasker.

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S. and Morgenstern, B. (2006). AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Research* **34**, W435–W439.

Stauber, M., Jäckle, H. and Schmidt-Ott, U. (1999). The anterior determinant bicoid of Drosophila is a derived Hox class 3 gene. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 3786.

Stauber, M., Taubert, H. and Schmidt-Ott, U. (2000). Function of bicoid and hunchback homologs in the basal cyclorrhaphan fly *Megaselia* (Phoridae). *Proceedings of the National Academy of Sciences* **97**, 10844–10849.

Stauber, M., Prell, A. and Schmidt-Ott, U. (2002). A single Hox3 gene with composite bicoid and zerknüllt expression characteristics in non-Cyclorrhaphan flies. *Proceedings of the National Academy of Sciences* **99**, 274–279.

Szklarczyk, D., Kirsch, R., Koutrouli, M., Nastou, K., Mehryary, F., Hachilif, R., Gable, A. L., Fang, T., Doncheva, N. T., Pyysalo, S., et al. (2023). The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res* **51**, D638–D646.

Tanaka, K., Diekmann, Y., Hazbun, A., Hijazi, A., Vreede, B., Roch, F. and Sucena, É. (2015). Multispecies Analysis of Expression Pattern Diversification in the Recently Expanded Insect Ly6 Gene Family. *Mol Biol Evol* **32**, 1730–1747.

Tomasello, G., Armenia, I. and Molla, G. (2020). The Protein Imager: a full-featured online molecular viewer interface with server-side HQ-rendering capabilities. *Bioinformatics* **36**, 2909–2911.

The UniProt Consortium (2023). UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research* **51**, D523–D531.

Vakirlis, N., Carvunis, A.-R. and McLysaght, A. (2020). Synteny-based analyses indicate that sequence divergence is not the main source of orphan genes. *Elife* **9**, e53500.

827 **van Kempen, M., Kim, S. S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C. L. M.,**
828 **Söding, J. and Steinegger, M.** (2024). Fast and accurate protein structure search with
829 Foldseek. *Nat Biotechnol* **42**, 243–246.

830 **VanKuren, N.** (2023)**.** Draft *Papilio alphenor* assembly and annotation. Dryad.
831 https://doi.org/10.5061/dryad.n2z34tn2x

832 **Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan,**
833 **D., Stroe, O., Wood, G., Laydon, A., et al.** (2022). AlphaFold Protein Structure
834 Database: massively expanding the structural coverage of protein-sequence space
835 with high-accuracy models. *Nucleic Acids Research* **50**, D439–D444.

836 **Varadi, M., Bertoni, D., Magana, P., Paramval, U., Pidruchna, I., Radhakrishnan, M.,**
837 **Tsenkov, M., Nair, S., Mirdita, M., Yeo, J., et al.** (2024). AlphaFold Protein Structure
838 Database in 2024: providing structure coverage for over 214 million protein sequences.
839 *Nucleic Acids Research* **52**, D368–D375.

840 **Vicoso, B. and Bachtrog, D.** (2015). Numerous transitions of sex chromosomes in
841 Diptera. *PLoS Biol* **13**, e1002078.

842 **Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski,**
843 **J. and Schatz, M. C.** (2017). GenomeScope: fast reference-free genome profiling from
844 short reads. *Bioinformatics* **33**, 2202–2204.

845 **Wang, Y., Tang, H., DeBarry, J. D., Tan, X., Li, J., Wang, X., Lee, T., Jin, H., Marler, B.,**
846 **Guo, H., et al.** (2012). MCScanX: a toolkit for detection and evolutionary analysis of
847 gene synteny and collinearity. *Nucleic Acids Research* **40**, e49.

848 **Wickham H.** (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New
849 York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.

850 **Wickham. H., François, R., Henry, L., Müller, K., Vaughan, D.** (2023). *dplyr: A Grammar*
851 *of Data Manipulation*. R package version 1.1.4, https://github.com/tidyverse/dplyr,
852 https://dplyr.tidyverse.org.

853 **Wickham, H., Vaughan. D., and Girlich, M.** (2024). *tidyr: Tidy Messy Data*. R package
854 version 1.3.1, https://github.com/tidyverse/tidyr, https://tidyr.tidyverse.org.

855 **Wiegmann, B. M., Trautwein, M. D., Winkler, I. S., Barr, N. B., Kim, J.-W., Lambkin, C.,**
856 **Bertone, M. A., Cassel, B. K., Bayless, K. M., Heimberg, A. M., et al.** (2011).

857 Episodic radiations in the fly tree of life. *Proceedings of the National Academy of*
858 *Sciences* **108**, 5690–5695.

859 **Wotton, K. R.** (2014). Heterochronic shifts in germband movements contribute to the rapid
860 embryonic development of the coffin fly *Megaselia scalaris*. *Arthropod Struct Dev* **43**,
861 589–594.

862 **Wotton, K. R., Jiménez-Guri, E., Matheu, B. G. and Jaeger, J.** (2014). A Staging
863 Scheme for the Development of the Scuttle Fly *Megaselia abdita*. *PLOS ONE* **9**,
864 e84421.

865 **Wotton, K. R., Jiménez-Guri, E., Crombach, A., Janssens, H., Alcaine-Colet, A.,**
866 **Lemke, S., Schmidt-Ott, U. and Jaeger, J.** (2015a). Quantitative system drift
867 compensates for altered maternal inputs to the gap gene network of the scuttle fly
868 *Megaselia abdita*. *eLife* **4**, e04785.

869 **Wotton, K. R., Jiménez-Guri, E., Crombach, A., Cicin-Sain, D. and Jaeger, J.** (2015b).
870 High-resolution gene expression data from blastoderm embryos of the scuttle fly
871 *Megaselia abdita*. *Sci Data* **2**, 150005.

872 **Wotton, K. R., Jiménez-Guri, E. and Jaeger, J.** (2015c). Maternal Co-ordinate Gene
873 Regulation and Axis Polarity in the Scuttle Fly *Megaselia abdita*. *PLOS Genetics* **11**,
874 e1005042.

875 **Xia, S., Chen, J., Arsala, D., Emerson, J. J. and Long, M.** (2025). Functional innovation
876 through new genes as a general evolutionary process. *Nat Genet* 1–15.

877 **Xiao, Z. and Lam, H.-M.** (2022). ShinySyn: a Shiny/R application for the interactive
878 visualization and integration of macro- and micro-synteny data. *Bioinformatics* **38**,
879 4406–4408.

880 **Yoder, J. H. and Carroll, S. B.** (2006). The evolution of abdominal reduction and the
881 recent origin of distinct Abdominal-B transcript classes in Diptera. *Evol Dev* **8**, 241–
882 251.

883 **Zhong, M., Wang, X., Liu, Q., Luo, B., Wu, C. and Wen, J.** (2016). The complete
884 mitochondrial genome of the scuttle fly, *Megaselia scalaris* (Diptera: Phoridae).
885 *Mitochondrial DNA A DNA Mapp Seq Anal* **27**, 182–184.

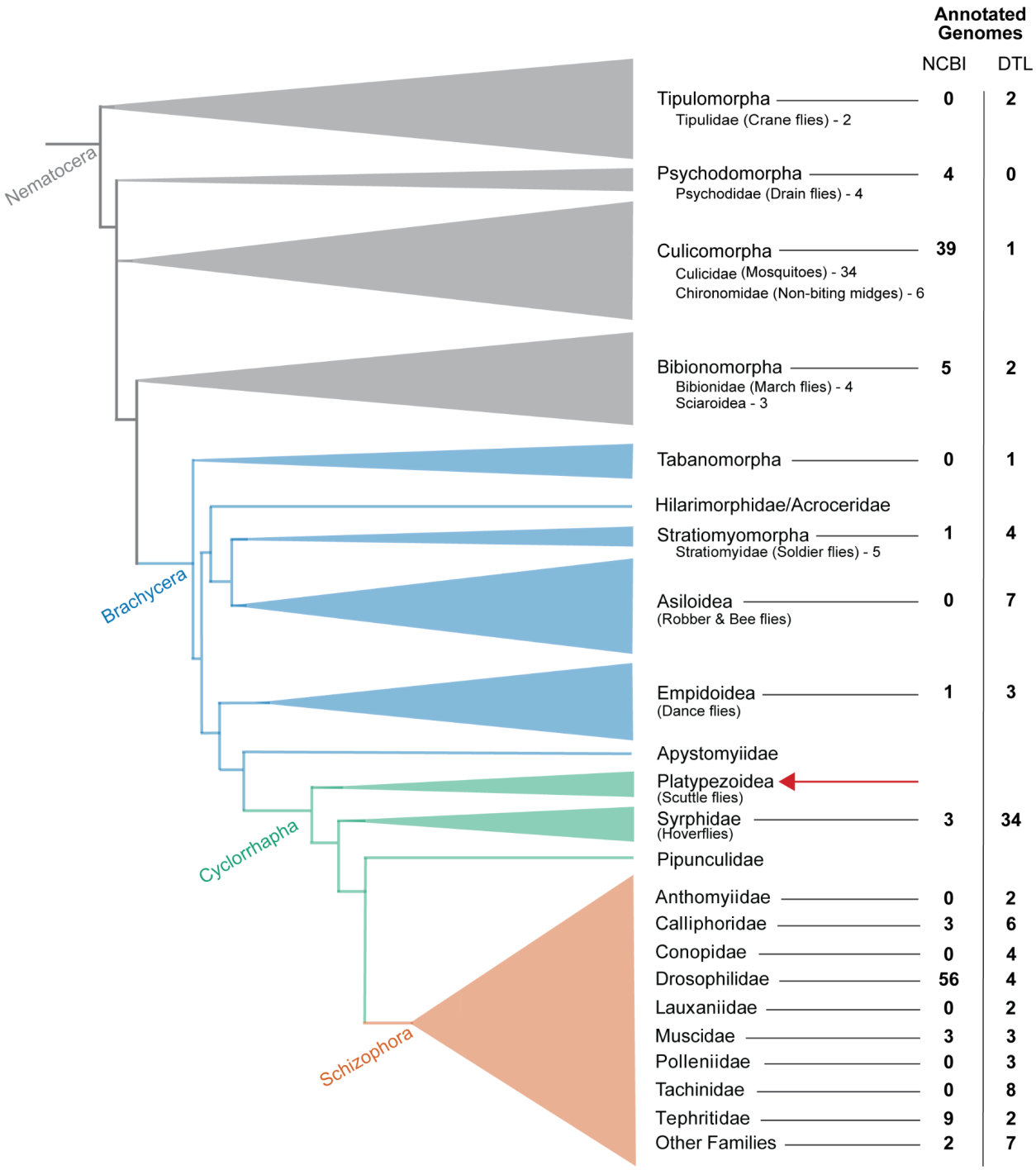


Figure 1. Phylogeny of Diptera with available genome annotations across lineages. The phylogeny is based on Wiegmann et al. (2011) and Jimenez-Guri et al. (2013). The width of each triangle reflects species richness within each group. The number of annotated genomes available for each clade in the genome database at National Center for Biotechnology Information (NCBI) or in the *Darwin Tree of Life* (DTL) project that have not been uploaded to NCBI Genomes as of Feb. 4th 2025 are indicated. The red arrow highlights the phylogenetic position of *Megaselia abdita*.

896 **Table 1.** *M. abdita* reference assembly statistics and quality metrics.

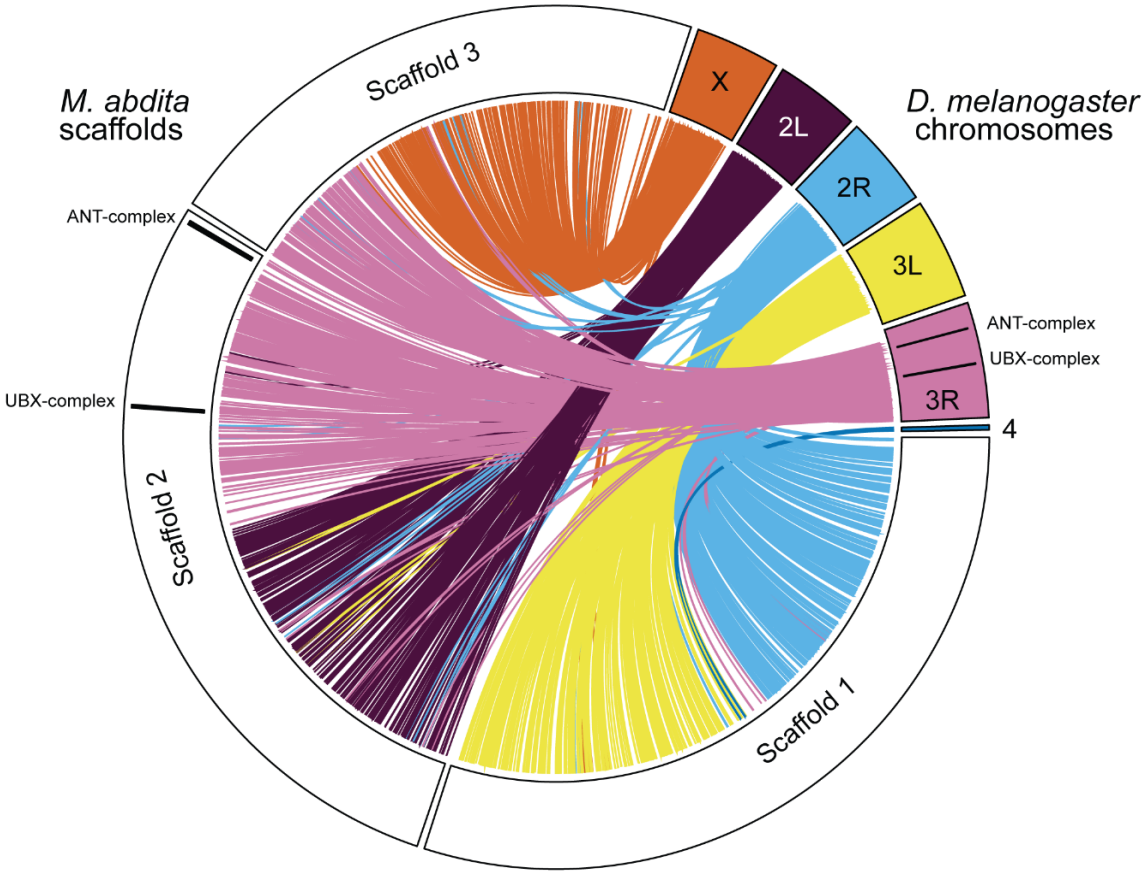
Reference Assembly Statistics	
Total length (bp)	592,824,975
Number of scaffolds	15
Scaffold N50 (bp)	212,802,314
Scaffold L50	2
# of Ns per 100kb	1.32
GC (%)	29.74%
Masked (%)	67.90%
Eukaroyotic BUSCO's recovered (%)	99%

897 **Table 2.** Genome annotation statistics (left) and quality metrics (right) for *M. abdita*. The quality
898 metrics include the percentage of complete universal orthologs identified in the annotation
899 across five BUSCO datasets. The number of genes in each lineage-specific BUSCO database
900 is shown in parentheses.

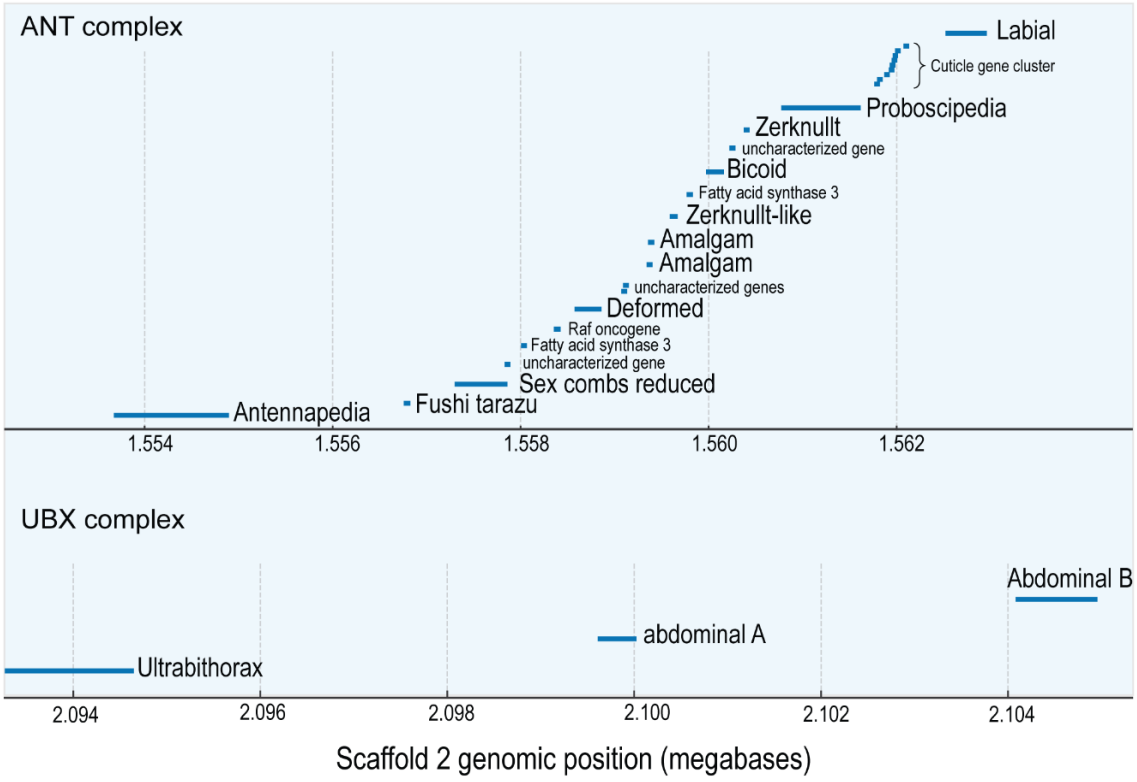
***M. abdita* Genome Annotation Stasitics**	
Number of coding genes	11,934
Number of mrnas	20,560
Number of mrnas with 3' & 5' UTR	17,607
Mean mRNAs/gene	1.7
Mean gene length (bp)	19,478

Complete universal orthologs recovered in in *M. abdita* annotation (BUSCOs)	
Eukaryota (n = 255)	93.3%
Metazoa (n = 954)	90.0%
Insecta (n = 1,367)	91.6%
Endopterygota (n = 2,124)	90.6%
Diptera (n = 3,285)	88.0%

A)



B) *M. abdita* HOX gene clusters



901

902 **Figure 2.** Synteny and Hox gene organization between *D. melanogaster* and *M. abdita*.
903 A) Synteny analysis between *D. melanogaster* chromosomes and *M. abdita* scaffolds. Lines
904 represent groups of collinear genes, connecting their positions between *D. melanogaster*
905 chromosomes and *M. abdita* chromosome-sized scaffolds. Colors correspond to *D.*
906 *melanogaster* chromosomes. The ANT and UBX Hox gene complexes are highlighted by black
907 lines within the chromosomes.
908 B) Visualization of *M. abdita* Hox gene clusters. The Antennapedia (top) and Ultrabithorax
909 (bottom) complexes are both located on Scaffold 2. Each blue line indicates the genomic
910 position and length of a gene.

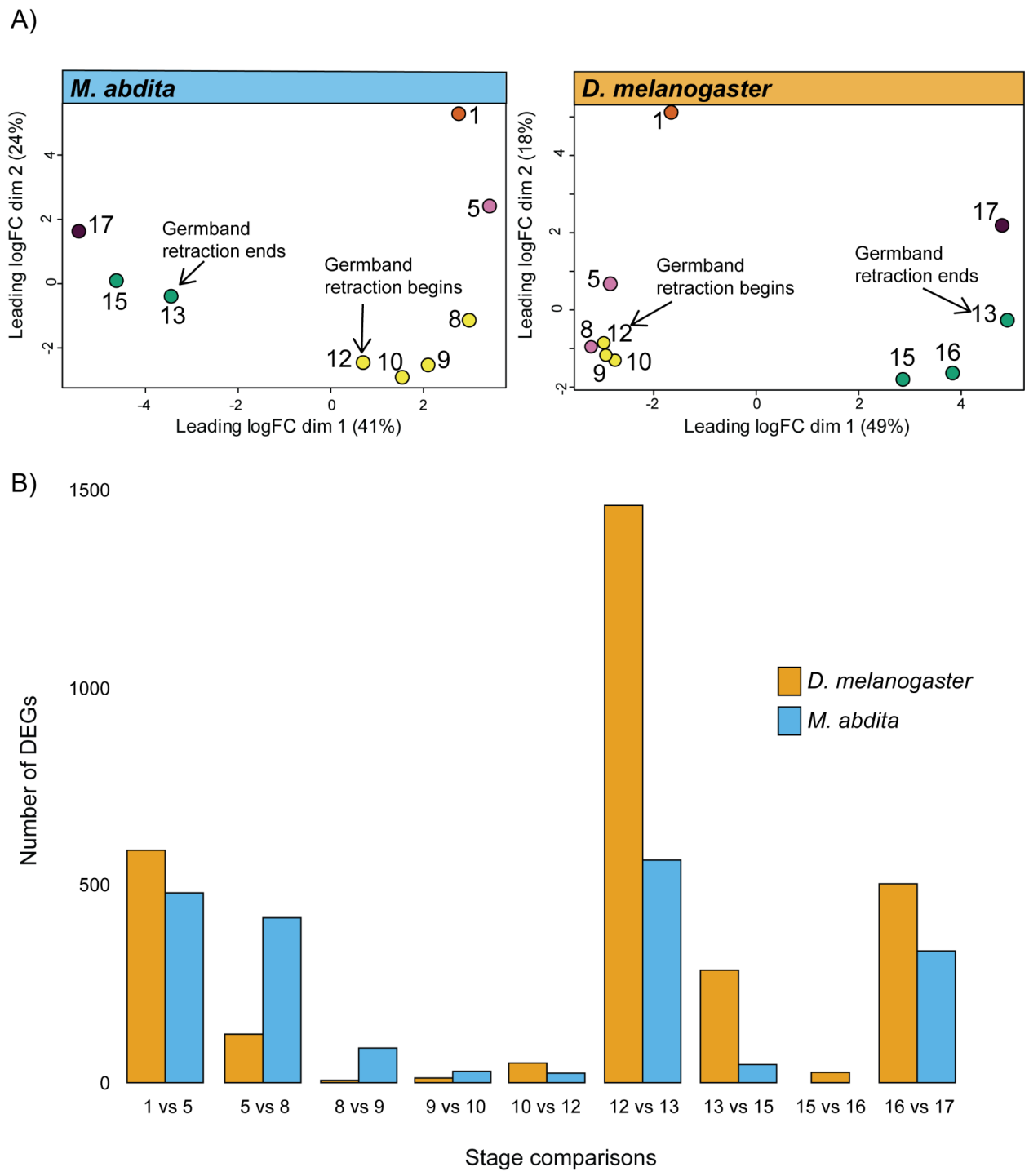
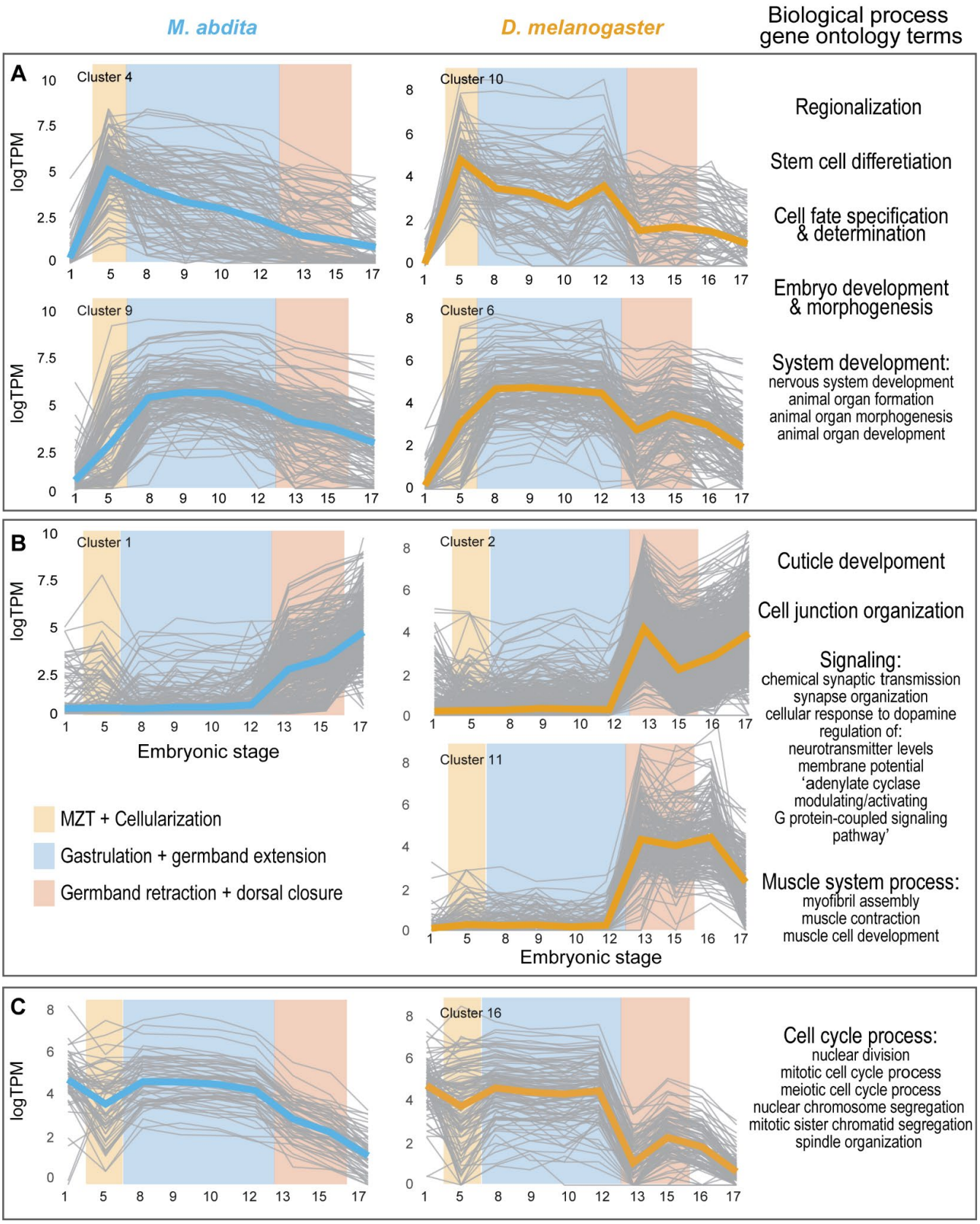


Figure 3. MDS clustering and differential gene expression across developmental stages in *M. abdita* and *D. melanogaster*.
A) Multidimensional scaling (MDS) plots based on the top 500 most differentially expressed genes from single-embryo RNA-seq samples. The *M. abdita* data are shown on the left (blue) and the *D. melanogaster* data on the right (yellow). Points represent individual samples, and their colors correspond to groupings identified by *k*-means clustering. Arrows indicate key developmental events: germband retraction begins and ends.

B) Number of differentially expressed genes (DEGs) between sequential developmental stages. Comparisons (e.g., "1 vs 5") indicate the number of DEGs identified between stage 1 and stage 5. Yellow bars represent *D. melanogaster*, and blue bars represent *M. abdita*.


M. abdita
D. melanogaster
Biological process gene ontology terms
A
Cluster 4
Cluster 10
Cluster 9
Cluster 6
logTPM
Regionalization
Stem cell differetiation
Cell fate specification & determination
Embryo development & morphogenesis
System development:
nervous system development
animal organ formation
animal organ morphogenesis
animal organ development
B
Cluster 1
Cluster 2
Cluster 11
Embryonic stage
Cuticle develpoment
Cell junction organization
Signaling:
chemical synaptic transmission
synapse organization
cellular response to dopamine
regulation of:
neurotransmitter levels
membrane potential
'adenylate cyclase modulating/activating G protein-coupled signaling pathway'
Muscle system process:
myofibril assembly
muscle contraction
muscle cell development
MZT + Cellularization
Gastrulation + germband extension
Germband retraction + dorsal closure
C
Cluster 16
Cell cycle process:
nuclear division
mitotic cell cycle process
meiotic cell cycle process
nuclear chromosome segregation
mitotic sister chromatid segregation
spindle organization

921

Figure 4. Expression dynamics of gene clusters and enriched GO terms during embryogenesis in *M. abdita* and *D. melanogaster*.
A-C) Expression profiles of differentially expressed gene clusters during embryogenesis and their enriched Biological Process Gene Ontology (BP GO) terms for *M. abdita* (left - blue) and *D. melanogaster* (right - yellow). Grey lines show the expression profile of the individual genes within the cluster and bolded lines represent the average expression profile of the entire cluster. DEGreports generated cluster names (e.g., "Cluster 4") which are arbitrary but retained here for continuity. Key embryonic developmental events: MZT + cellularization, gastrulation + germband extension, and germband retraction + dorsal closure are highlighted in yellow, blue, and red, respectively. Shared enriched BP GO terms for the boxed clusters are listed on the right.

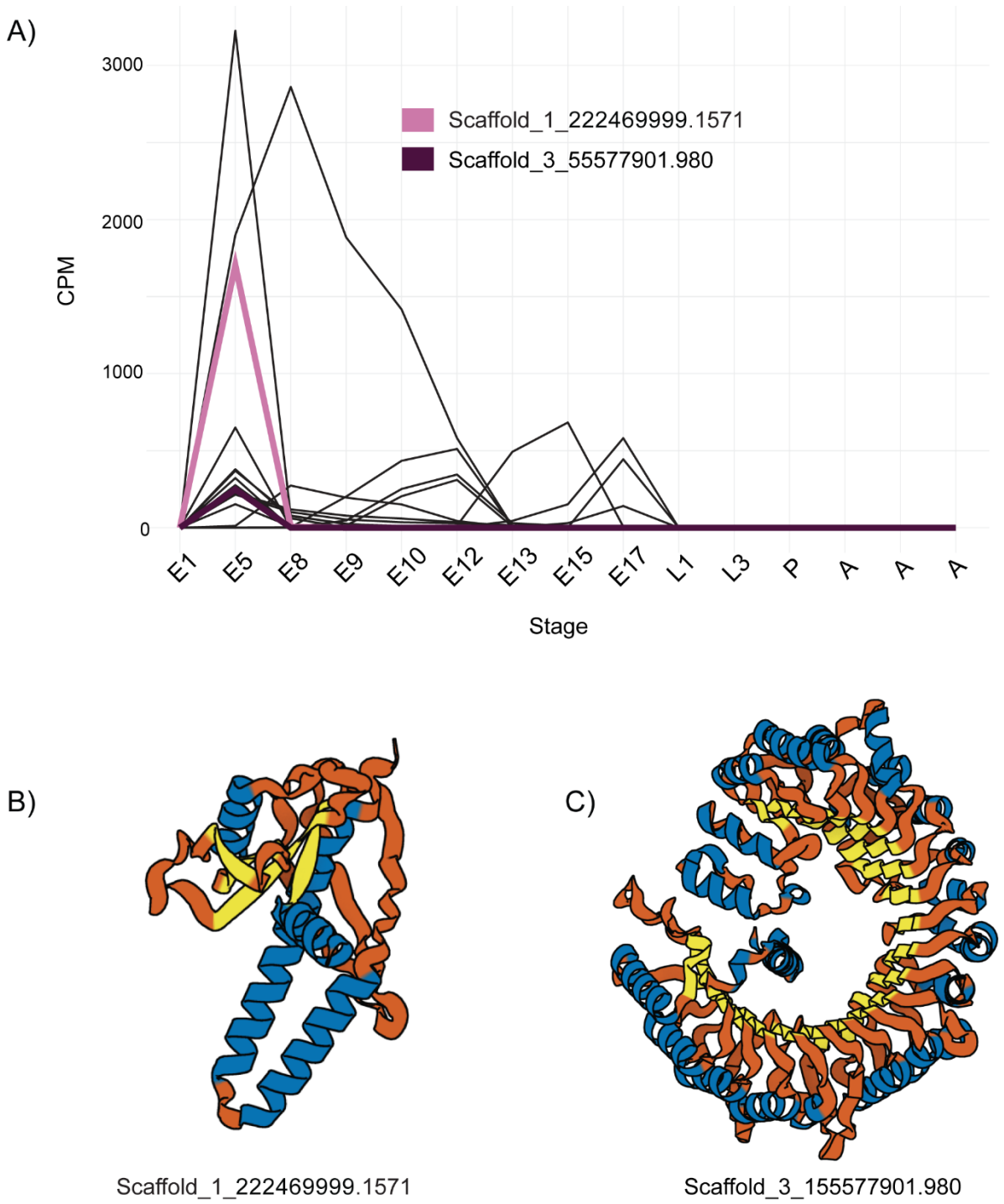


Figure 5. Expression and predicted structures of embryonically expressed orphan genes in *M. abdita*.
A) Expression profiles of orphan genes with expression limited to embryonic stages. E indicates embryonic, L indicates larval, P indicates pupal, and A indicates adult (multiple samples).

Highlighted in pink and dark purple are two genes with high-confidence protein structure predictions. All other ‘orphan’ gene expression profiles are shown in black.
B) AlphaFold-predicted protein structure for Scaffold_1_222469999.1571.
C) AlphaFold-predicted protein structure for Scaffold_3_155577901.980. Sequence similarity searches suggest this gene encodes a novel F-box-LRR protein. In the structures, alpha helices are highlighted in blue, and beta sheets are highlighted in yellow.