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]
Chun Wai Kwan[1,4]
Sheri A Sanders[5]
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. The naturally occurring diversity of mechanisms and their phylogenetic history can aid in revealing core principles and inform our understanding of evolutionary transitions. Additionally, careful comparisons of multiple species in “model taxa” are critical for determining the directionality of change and identifying mutations responsible for important evolutionary shifts in developmental processes. 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), such mutations may not be adaptive and their significance as drivers of evolutionary change might be overlooked.

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), 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.

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 1). 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).

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

M. abdita's *zen* gene is expressed in the serosa and has been characterized in detail (Caroti et al., 2018; Kwan et al., 2016; Rafiqi et al., 2008; Rafiqi et al., 2010; Rafiqi et al., 2012; Stauber et al., 1999). Consistent with these studies, we detected the *M. abdita's zen* transcript in embryonic stages 5 - 15, coinciding with the time when the serosa is specified and maintained (Figure S3). In contrast, the newly identified *zen-like* gene's expression was only detected at stage 5 (Figure S3). Our findings underscore how a well-annotated genome can facilitate precise comparisons of chromosomal architecture which will enable the identification of both conserved regulatory landscapes and structural variation implicated in phenotypic diversity.

Major transition in transcriptional expression profile during germband extension and retraction

The extent to which gene expression is similar or different between stages and species provides a basis for identifying developmental windows of accelerated change, heterochronic shifts, and evolutionary divergence. To identify conserved features and differences between embryonic stages of *Megaselia abdita* and *Drosophila melanogaster*, we performed RNA-seq on single embryos from stages 1, 5, 8, 9, 10, 12, 13, 15, 16, and 17 for both species (Figure S4). We staged embryos based on morphology, corresponding to established staging schemes for each species (Campos-Ortega & Hartenstein, 1997; Wotton et al., 2014). We later excluded the *M. abdita* stage 16 embryo due to failed sequencing. Despite similar embryo sizes and uniform library preparation and sequencing conditions, *M. abdita* embryos had roughly twice the number of reads per embryo compared to *D. melanogaster* (Figure S5A) though the total number of genes with reads assigned was comparable between species with 10,373 in *D. melanogaster* to 9,999 in *M. abdita*. The number of genes expressed was similar across embryonic stages between species as well (Figure S5B). The number of reads mapping to genes doubled in *M. abdita* (8,941 vs. 3,863 reads/gene) consistent with its 2x read count. We used normalized read counts (transcript per million or TPM) to account for the global expression level differences, sample-to-sample variation, and differences in transcript length between genes. Our RNAseq data showed a major shift in transcriptional expression during germband retraction in both *M. abdita* and *D. melanogaster*. A multidimensional scaling (MDS) plot, which plotted samples based on the similarity of their top 500 most differentially expressed genes, revealed clustering of embryo transcriptomes before and after germband retraction (Figure 3A).

The transition from germband retraction to dorsal closure marks a shift from system-wide body plan development to nervous system-specific development in both *M. abdita* and *D. melanogaster*

To more closely look at the gene networks in both species, we identified the differentially expressed genes (DEGs) between developmental stages. Due to the absence of biological replicates for each embryo, we used a k-mean clustering approach to group samples and calculate a global dispersion estimate (σ = 0.36 for both *M. abdita* and *D. melanogaster*). Both species have high dispersion estimates but variance in expression was very similar in both datasets (Figure S6 & S7). Despite the high dispersion in data, we had sufficient power to detect DEGs with fold differences ±3 and p-value < 0.001.

We performed pairwise comparisons of each sequential stage: the zygote, containing maternal-deposited transcripts (Stage 1); cellularization, where maternal to zygotic transition (MZT) of gene expression occurs (Stage 5); gastrulation, germband extension & retraction (Stages 8-12); dorsal closure (Stages 13-16), and the final embryonic stage (stage 17). We detected 1,567 and 2,398 DEGs in *M. abdita* and *D.* in *melanogaster* respectively (Figure 3B). In approximately 1 hour of developmental time between stages 12 (germband retraction) and 13 (dorsal closure), both species showed a strikingly dynamic shift in gene expression. However, *D. melanogaster* had 1,463 DEGs compared to *M. abdita's* 564 DEGs at this same time point (Figure 3B). Conversely, *M. abdita* exhibited more dynamic gene expression earlier, between cellularization and gastrulation (stages 5-8), with 418 DEGs compared to 123 in *D. melanogaster* (Figure 3B). Between stages 8 and 12, there are markedly fewer DEGs in both species, corresponding with the period of gastrulation and germband retraction (Figure 3B). This pause in the turning on and off of genes is coincident with the phylotypic stage of development (Kalinka et al., 2010). It is important to reiterate that this analysis highlights the most dynamic (largest changes in expression) genes between sequential developmental stages – the total number of expressed genes at any given stage was very similar between species (Figure S5B) - suggesting that *D. melanogaster* exhibits sharper increases and decreases in expression than *M. abdita* after the phylotypic phase ends while *M. abdita* seems to have evolved more dynamic expression before the phylotypic phase begins.

To analyze the expression patterns of DEGs, we clustered them using DEGreports and plotted the expression of each cluster (Figure S8) (Pantano 2024). The clusters confirmed that the most dynamic transcriptional shifts occur during the transitions between stages 1 to 5, 5 to 8, and 12 to 13 for both species (Figure 3B). We then assessed whether the genes expressed in each cluster were significantly enriched for any Biological Process Gene Ontology (BP GO) terms using the Search Tool for the Retrieval of Interacting Genes/Proteins (String) (Szklarczyk et al. 2023) (Figure S8). *M. abdita*'s clusters 4 and 9 and *D. melanogaster's* clusters 6 and 10 were significantly enriched for embryonic, body plan, and systems development terms (Figure 4A & S9A) and contained many well-characterized developmental genes in Drosophila (Figure S9B&C). Genes found in *M. abdita* cluster 1 and *D. melanogaster* clusters 2 and 11 exhibited a pronounced increase in expression at stage 13 and were enriched for terms related to nervous system, muscle, and cuticle development (Figure 4B). These findings highlight a conserved shift toward nervous system development at the end of germband retraction in both species, while also suggesting subtle differences in the timing and clustering of body plan-related gene expression. In addition to their similarities, *D. melanogaster's* more extreme expression changes between stages 12 and 13 (Figure 4 A-C) are notable because at these stages *Drosophila* employs an evolutionarily novel mechanism involving a new tissue called the amnioserosa for dorsal closure (Schmidt-Ott and Kwan, 2022) which may be contributing to the distinct transcriptional dynamics observed between the species.

Our RNAseq analysis identified major transcriptional transitions between developmental stages 1 and 5, 5 and 8, and 12 and 13, revealing conserved and diverged patterns of gene expression between *M. abdita* and *D. melanogaster*. This work was made possible by mapping the stage-specific transcriptomes to the annotated genome, which allows us to link expression patterns to specific genomic features.

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

As the evolution of new genes within lineages and species is an important mechanism of diversification of developmental mechanisms (Chen et al., 2010; Chen et al. 2013), we searched *M. abdita's* genome for 'orphan' genes. Orphan genes are either highly diverged from known sequences or represent newly evolved genes (Vakirlis et al., 2020; Xia et al., in press). In *M. abdita's* genome annotation, eggnog-mapper was unable to assign orthologs to 1,049 gene models. Further searches using *blastn* on the coding sequences and *blastp* on the predicted protein sequences did not reveal any sequence similarity to other Dipterans. Approximately 8.7% of *M. abdita's* genes were 'orphan' genes.

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

To further characterize these 28 embryo-specific genes, we examined their open reading frames and protein sequences (Table S3). We classified 21 as likely coding and of those we further classified 9 as likely stable proteins (blue highlighted rows in Table S3). Additional searches of InterProScan's database revealed no known protein domains or any domains consistent with known transposases (Jones et al., 2014). We then used Alphafold2 to generate protein structure models (Jumper et al., 2021). Two of these genes resulted in confident structures (Table S3, pTM > 0.5). One of them could not be related to any known protein (Figure 5B); however, the second showed significant structural similarity to F-box leucine-rich repeat (LRR) proteins (Figure 5C). F-box LRRs are components of the 'E3 ubiquitin ligase SCF complex' which ubiquitinates targeted proteins for later degradation by the cell. Specifically, the F-box domain binds to *Skp1* and the LRR domain binds to the target, the molecule slated for ubiquitination and degradation.

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

F-box LRRs are known to ubiquinate important developmental signaling molecules in *D. melanogaster* including the pair-rule gene *paired* (*prd*) which is bound by the F-box LRR protein *Partner of paired* (*Ppa*) (Raj et al., 2000). *Ppa* is unusual in that its expression is patterned rather than uniform as most F-Box LRR genes seem to be in *D. melanogaster* embryogenesis (Das et al., 2002). Given that in *D. melanogaster*, the 12 best-known F-box LRR proteins (*Skp2*, *Ppa*, *Kdm2*, *FipoQ*, *Fbxl4*, *Fbxl7*, *Fbl6*, CG32085, CG13766, CG12402, CG9003, CG8272) are expressed throughout embryogenesis according to our both our RNAseq data and ENCODE gene expression data, the stage-restricted expression of *M. abdita*'s orphan F-box LRR gene and 13 other orphan genes with expression peaking before gastrulation might reflect previously overlooked developmental differences between *M. abdita* and *D. melanogaster* at the blastoderm stage.

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.

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

1.2 Omni-C library preparation and sequencing

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

1.3 Scaffolding assembly with HiRise

Both the Hifiasm draft assembly and Dovetail OmniC sequencing reads were used as input for HiRise, a software tailored for scaffolding genome assemblies with proximity ligation data (Putnam et al., 2016). Based on spatial data from the chromatin conformation capture (Figure S1), HiRise pinpoints regions where contigs are joined incorrectly (misjoins) in the initial assembly and utilizes this spatial information to re-orient contigs and construct larger scaffolds. The OmniC library sequences were aligned to the draft assembly using the Burrows-Wheeler Aligner (Li and Durbin, 2009). HiRise-analyzed read pairs are mapped to the draft scaffolds to develop a genomic distance likelihood model. This model is then used to identify and break potential misjoins, score prospective joins, and execute joins surpassing a set threshold. These scaffolds consist of sequentially arranged contigs separated by gaps. We used QUAST to calculate %GC, N50, L50, and Ns per 100 kbp (Gurevich et al., 2014). We then repeat-masked the genome with RepeatMasker (Smit et al., 2013). To estimate heterozygosity and sequencing error rates, we calculated the frequency spectrum of canonical 21-mers using jellyfish, and input 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 1). 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 protein alignments using Augustus' exonerate2hints function, intron hints from mapped RNA seq reads using Augustus' bam2hints function, and exon/intron hints from the PASA assemblies using Augustus' bam2exonhints function. We additionally used the coding predictions from Transdecoder to create training models and further refined those models using the lib.selectTrainingModels function from Funannotate (Palmer and Stajich, 2019). This training data was used to run GlimmerHMM with hints as guidance (Majoros et al., 2004). We then provided the PASA assemblies, mapped protein data, TransDecoder predictions, *ab initio* predictions, BRAKER3 predictions, and a file weighting each line of evidence to the software Evidence Modeler v2.1.0. Evidence Modeler constructed the consensus gene structures which were updated by PASA to add UTRs and identify alternative transcripts (Haas et al., 2008). We removed gene structures that overlapped with RepeatMasker output using a Funannotate function called "RemoveBadModels." We then used Funannotate v1.8.1 to identify annotations that match known transposable elements (TEs) and repeat proteins using BLAST and updated the annotation to remove remaining TEs. We used AGAT v1.4.1 to remove genes with an open reading frame (ORF) < 100 amino acids in length and any associated gene structures (Dainat, 2022). We assigned gene names using eggNOG-mapper v2.1.12 which relies on orthology predictions to functionally annotate genes (Cantalapiedra et al., 2021; Huerta-Cepas et al., 2019).

Synteny Analysis

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

Embryo Staging for RNAseq

For precise embryo staging, embryos of the appropriate age were mounted on a microscope slide under halocarbon 27 oil and observed in a Zeiss Axiophot compound microscope equipped with a 10x objective and DIC (differential interference contrast) optics until they reached the desired stage. They were photographed and immediately processed for RNA extraction as previously described (Lott et al., 2014). We collected *M. abdita* and stage-matched *D. melanogaster* embryos at stages 1, 5, 8, 9, 10, 12, 13, 15, 16, and 17. Photos of the sequenced embryos can be found in Figure S4.

RNA isolation and Sequencing

We incubated each sample for 5-10 minutes at room temperature in TRIzol. We then froze each sample in TRIzol at -80° C. We extracted total RNA using the TRIzol/Phenol-chloroform protocol detailed in the appendix (Protocol 10) of ‘Functional evolution of a morphogenetic gradient’ by Chun Wai Kwan. The University of Chicago genomics core facility constructed cDNA libraries using the TruSeq kit with PCR (Illumina, CA, USA). The cDNA libraries were barcoded and 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 S5A). 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 S6). Additionally, the average, median, and variance in gene expression (TPM) were very similar across development in both species (Figure S7).

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 $\log_2$(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 σ=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.

Gene Ontology Enrichment Analysis

We used the STRING database v12.0 (Search Tool for the Retrieval of Interacting Genes/Proteins) to perform enrichment analysis of Biological Process Gene Ontology (GO) terms for our clustered DEG lists (Szklarczyk et al., 2023). STRING compares input gene sets to a reference genome to identify networks of interacting genes and enrichments in biological processes. Since *M. abdita* is not available in STRING, we first used NCBI's Blast tool to select the top *D. melanogaster* ortholog match/hit for each *M. abdita* gene (Altschul et al., 1990). Using FlyBase's batch download tool, we retrieved the corresponding FlyBase IDs, which STRING accepts as input (Öztürk-Çolak et al., 2024). STRING performed the enrichment analysis identifying the Biological Process GO terms significantly associated with each developmental stage. STRING measures enrichment based on the strength of enrichment (Log_{10}(observed/expected)), false discovery rate (p-values corrected for multiple testing with Benjamini-Hochberg), and the signal (weighted harmonic mean between observed/expected ratio and -log(FDR).

‘Orphan’ gene identification

We identified genes from *M. abdita*'s annotation file for which EggNOG-mapper could not assign an ortholog. Next, we assessed the expression of these genes in our RNA-seq data. To identify genes with exclusively embryonic expression, we focused on those with a CPM > 1 in at least one of the nine embryonic stages and a CPM < 1 in pupal, larval, and adult stages. We validated these genes further by analyzing their ORFs using NCBI's ‘Open reading frame finder’. We used CPC2 to evaluate the nucleic acid sequences to assess coding potential (Kang et al., 2017). Finally, we performed sequence similarity searches using NCBI's blastn and blastp tools, querying both nucleotide and protein sequences against the entire NCBI database as well as against Dipteran-specific sequences (Altschul et al., 1990; Sayer et al., 2022).

Protein structure prediction and structural similarity search

We used Protparam to assess protein stability, aliphatic index, and hydropathicity of predicted proteins (Gasteiger et al., 2005). We then searched all predicted protein sequences against InterProScan to look for any missed domains, specifically to look for evidence of transposable elements that were not discovered with blast searches (Jones et al., 2014). We used the AlphaFold2.ipynb provided by ColabFold v1.5.5 to predict protein structures (Jumper et al., 2021; Mirdita et al., 2022;). If the predicted protein had a predicted template modeling (pTM) score > 0.5 we then uploaded the structure to FoldSeek's website and searched the available databases (AlphaFold/Proteome, AlphaFold/Swiss-Prot, AlphaFold/UniProt50, BFMD) for proteins with similar structure (van Kempen et al., 2024; Varadi et al., 2022; Varadi et al., 2024). We used Protein Imager to generate publication-quality images of protein structures (Tomasello et al., 2020).

Acknowledgments

We thank Lily Shiue and Qianyu Jin at Cantata Bio for assembling the genome. We thank Nicholas VanKuren for his notes on genome annotation. We thank Yaikhomba Mutum for the conversation on protein structure prediction and structural similarity searches. We thank Xiang-Ru (Shannon) Xu for optimization of the *M. abdita* single cross procedure and the drawings in figure S10.

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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith Jr, R. K., Hannick, L. I., Maiti, R., Ronning, C. M., Rusch, D. B., Town, C. D., et al. (2003). Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research* **31**, 5654–5666.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Slater, G. S. C. and Birney, E. (2005). Automated generation of heuristics for biological 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.

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

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

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

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

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

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

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

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

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

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

Wiegmann, B. M., Trautwein, M. D., Winkler, I. S., Barr, N. B., Kim, J.-W., Lambkin, C., Bertone, M. A., Cassel, B. K., Bayless, K. M., Heimberg, A. M., et al. (2011). Episodic radiations in the fly tree of life. *Proceedings of the National Academy of Sciences* **108**, 5690–5695.

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

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

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

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

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

Xia, S., Chen, J., Arsala, D., Emerson, J. and Long, M. (In press). The origin of new genes is a general evolutionary process of functional innovation. *Nature Genetics*.

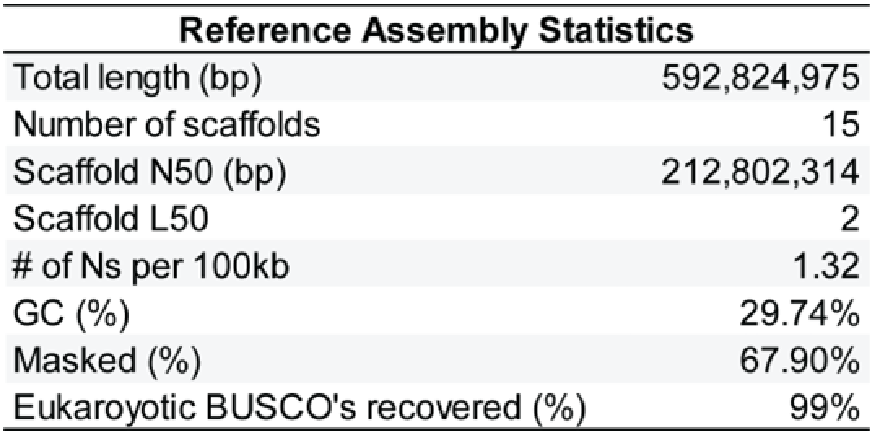
Xiao, Z. and Lam, H.-M. (2022). ShinySyn: a Shiny/R application for the interactive visualization and integration of macro- and micro-synteny data. *Bioinformatics* **38**, 4406–4408.

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

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

Figures and Tables

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%

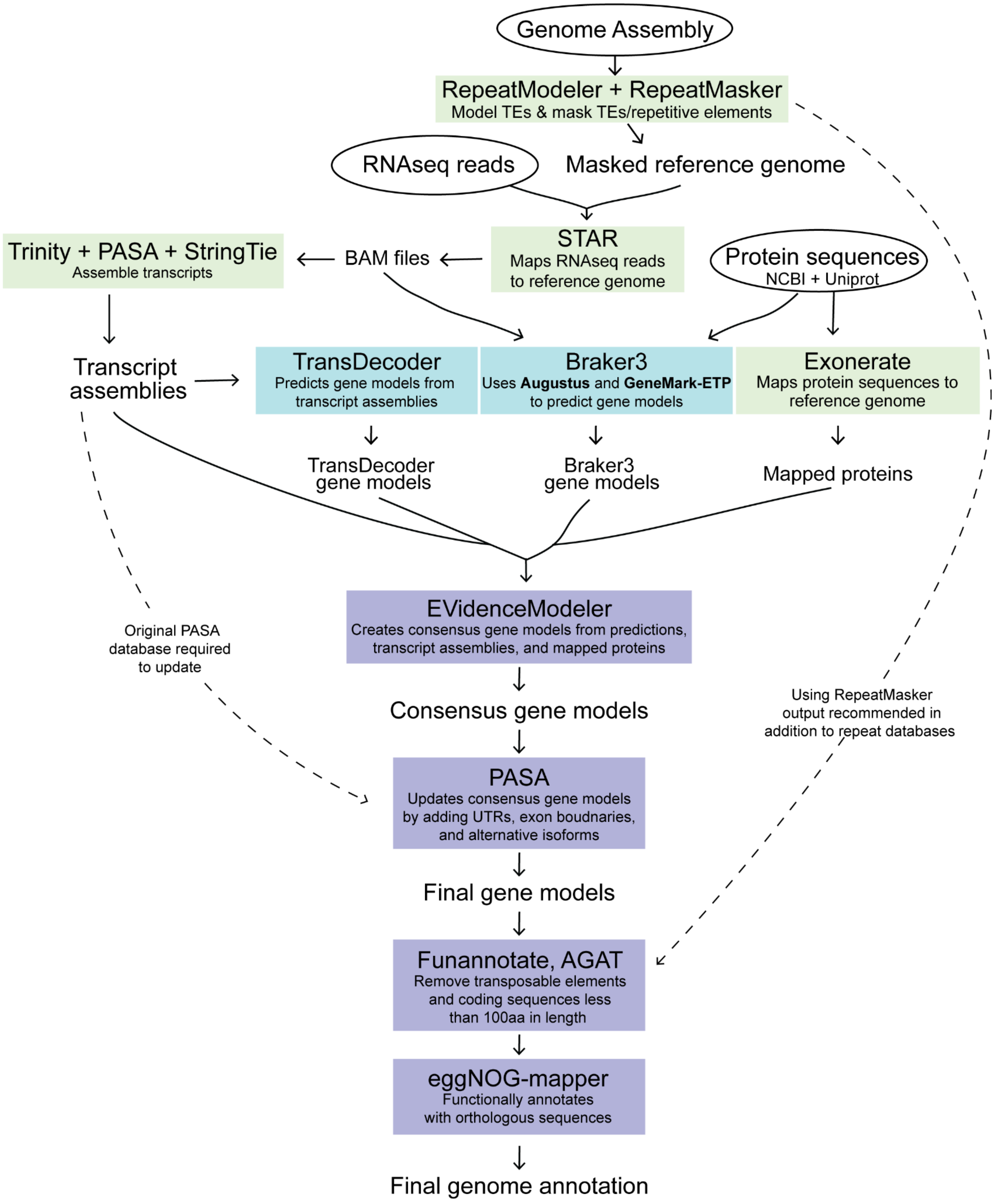
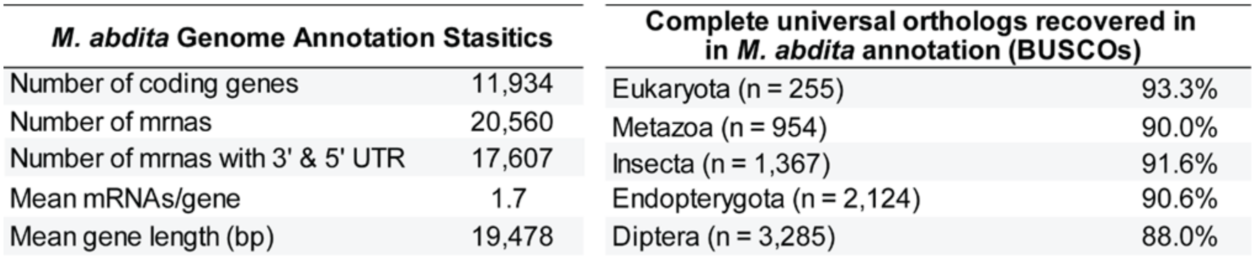

Genome Assembly
RepeatModeler + RepeatMasker
Model TEs & mask TEs/repetitive elements
RNAseq reads
Masked reference genome
STAR
Maps RNAseq reads to reference genome
BAM files
Trinity + PASA + StringTie
Assemble transcripts
Protein sequences
NCBI + Uniprot
Transcript assemblies
TransDecoder
Predicts gene models from transcript assemblies
Braker3
Uses **Augustus** and **GeneMark-ETP** to predict gene models
Exonerate
Maps protein sequences to reference genome
TransDecoder gene models
Braker3 gene models
Mapped proteins
EVidenceModeler
Creates consensus gene models from predictions, transcript assemblies, and mapped proteins
Original PASA database required to update
Consensus gene models
Using RepeatMasker output recommended in addition to repeat databases
PASA
Updates consensus gene models by adding UTRs, exon boudnaries, and alternative isoforms
Final gene models
Funannotate, AGAT
Remove transposable elements and coding sequences less than 100aa in length
eggNOG-mapper
Functionally annotates with orthologous sequences
Final genome annotation

Figure 1. Genome annotation pipeline. The starting input files are shown in black ovals and include the genome assembly along with two lines of evidence: RNAseq reads (fastq format) and protein sequences obtained from NCBI and UniProt. Software tools are represented in colored boxes: green indicates mapping software, blue indicates gene model generation

software, and purple indicates post-gene model processing and functional annotation tools. Arrows pointing from a software box to unboxed text represent the output files generated by the software. Arrows leading from unboxed text to a software box indicate input files used by the software.

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

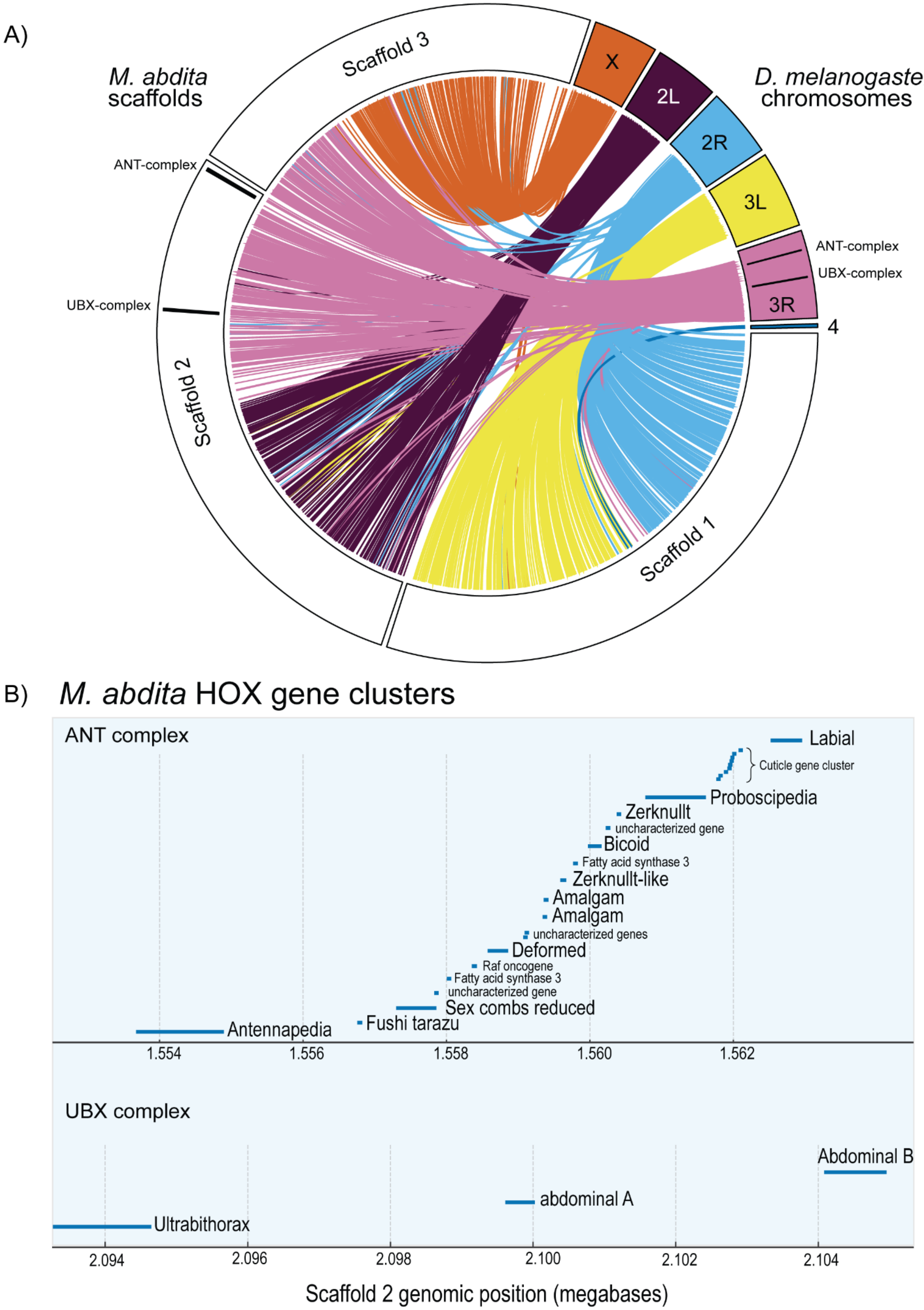

A)
M. abdita scaffolds
D. melanogaster chromosomes
Scaffold 3
Scaffold 2
Scaffold 1
X
2L
2R
3L
3R
4
ANT-complex
UBX-complex
ANT-complex
UBX-complex
B)
M. abdita HOX gene clusters
ANT complex
Labial
Cuticle gene cluster
Proboscipedia
Zerknullt
uncharacterized gene
Bicoid
Fatty acid synthase 3
Zerknullt-like
Amalgam
Amalgam
uncharacterized genes
Deformed
Raf oncogene
Fatty acid synthase 3
uncharacterized gene
Sex combs reduced
Fushi tarazu
Antennapedia
1.554
1.556
1.558
1.560
1.562
UBX complex
Abdominal B
abdominal A
Ultrabithorax
2.094
2.096
2.098
2.100
2.102
2.104
Scaffold 2 genomic position (megabases)

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

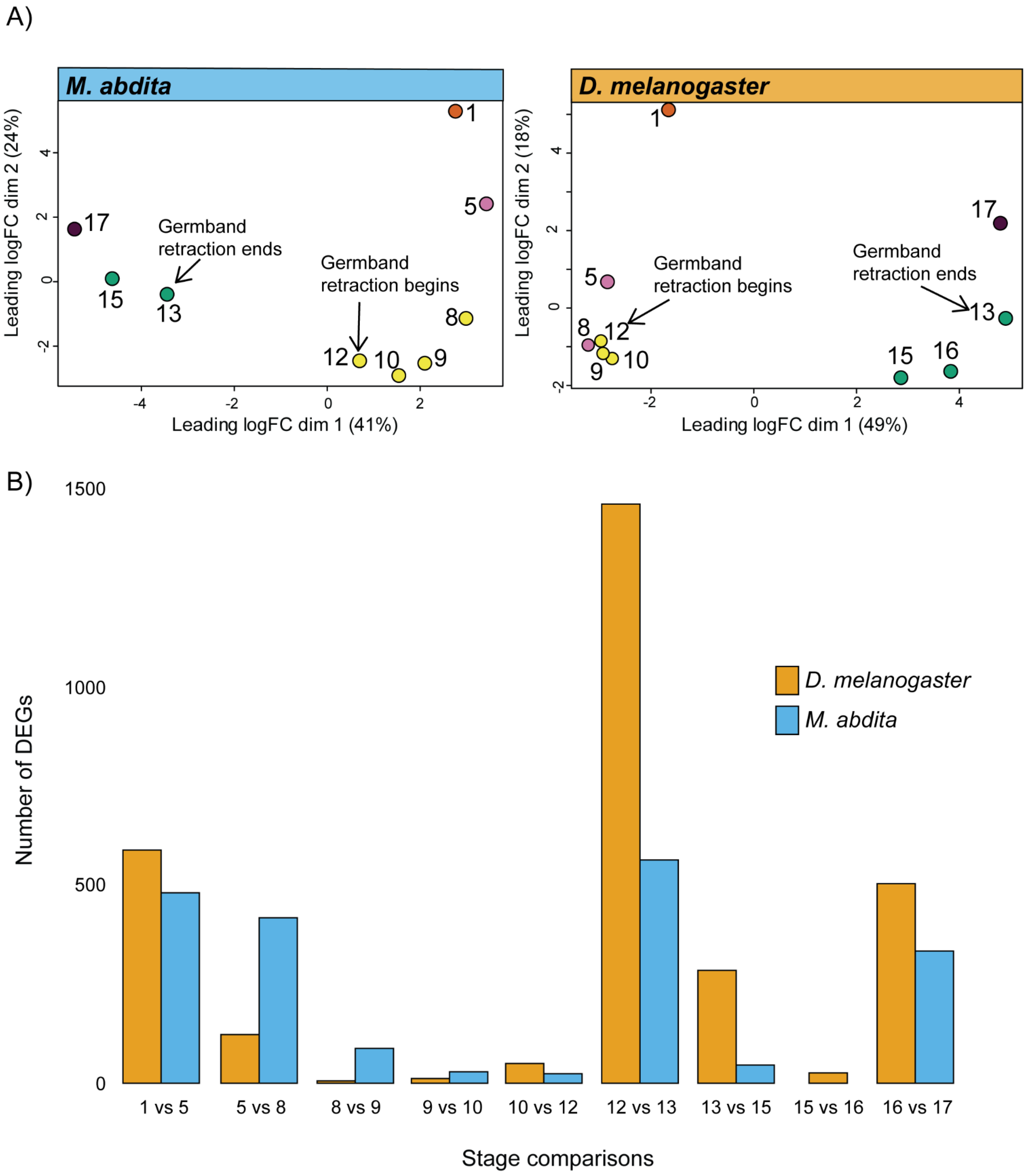


Figure 3. A) Multidimensional scaling (MDS) plots based on the top 500 most differentially expressed genes from single-embryo RNA-seq samples. The *M. abdita* data are shown on the left (blue) and the *D. melanogaster* data on the right (yellow). Points represent individual samples, and their colors 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*.

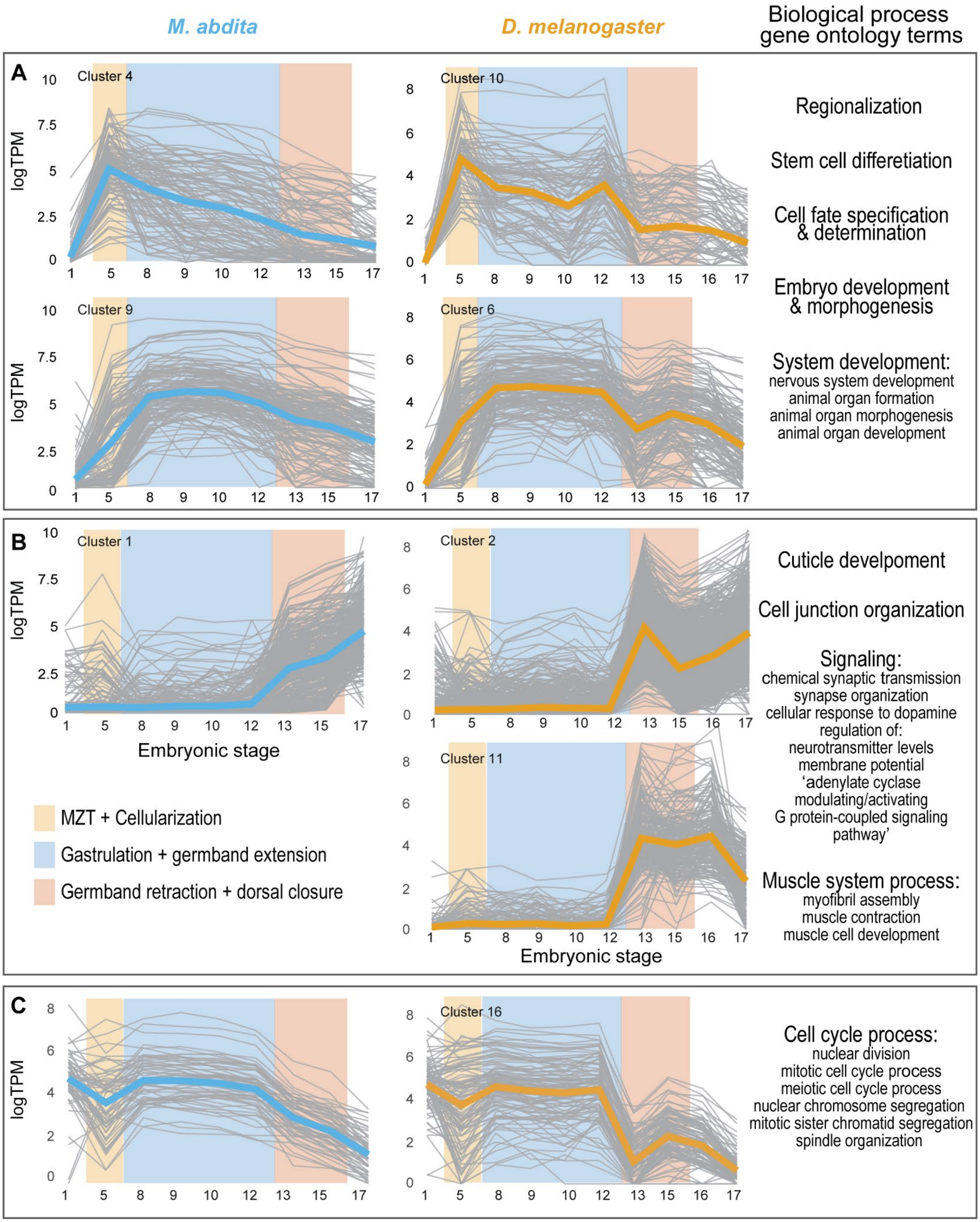

M. abdita
D. melanogaster
Biological process gene ontology terms
A
Cluster 4
Cluster 10
Cluster 9
Cluster 6
logTPM
Regionalization
Stem cell differetiation
Cell fate specification & determination
Embryo development & morphogenesis
System development:
nervous system development
animal organ formation
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
1 5 8 9 10 12 13 15 16 17
0 2 4 6 8
0 2.5 5 7.5 10

Figure 4. 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.

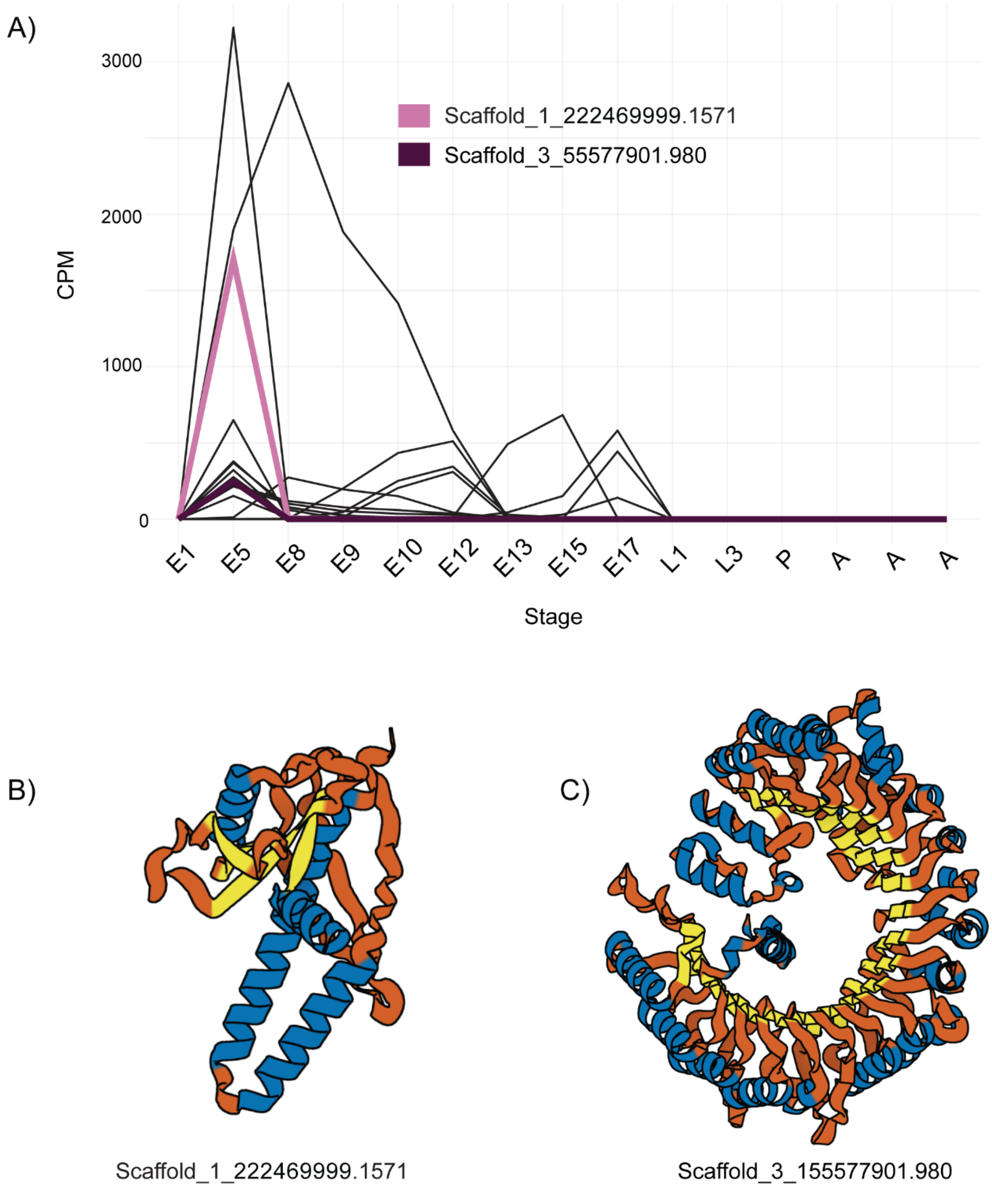


Figure 5. A) Expression profiles of orphan genes with expression limited to embryonic stages. E indicates embryonic, L indicates larval, P indicates pupal, and A indicates adult (multiple samples). Highlighted in pink and dark purple are two genes with high-confidence protein structure predictions. All 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.