Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Long-read RNA-seq delineates temporal transcriptional dynamics in multiplexed and sexed single medfly embryos

**Anthony Bayega**
McGill University
**Spyros Oikonomopoulos**
McGill University
**Dimitris Rallis**
University of Thessaly
**Dionisia Mavritsakis**
Royal College of Surgeons in Ireland
**Antonia Spanomitrou**
University of Thessaly
**Kostas D Mathiopoulos**
University of Thessaly
**Jiannis Ragoussis**
ioannis.ragoussis@mcgill.ca

McGill University

**Research Article**

**Keywords:** Medfly, single embryo development, Nanopore RNA sequencing, Ceratitis capitata, absolute normalization

**Posted Date:** August 14th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-9751782/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** Competing interest reported. JR was a member of the MinION Access Program (MAP) and has received free-of-charge flow cells and sequencing kits from Oxford Nanopore Technologies for other projects. JR has had no other financial support from ONT. AB and JR have

received reimbursement for travel costs associated with attending the Nanopore Community meetings organized by Oxford Nanopore Technologies. The rest of the authors do not have competing interests.

# Abstract

Long-read transcriptome sequencing provides an opportunity to characterize transcript structure and gene expression dynamics simultaneously, particularly in organisms whose developmental programs remain incompletely understood. The Mediterranean fruit fly (*Ceratitis capitata*) is one of the most economically important agricultural pests worldwide, yet the molecular events that govern its earliest stages of embryogenesis remain poorly resolved. Here, we combined long-read RNA sequencing, absolute transcript quantification, embryo sexing, and pseudotemporal trajectory analysis to investigate transcriptional dynamics during the first fifteen hours of medfly development. Single embryos were collected at hourly intervals, and both RNA and genomic DNA were isolated from each embryo. External RNA standards enabled estimation of transcript abundance on an absolute scale, while embryo sex was determined using Y-specific molecular markers.

Genome-guided transcriptome reconstruction identified 22,875 transcripts, including 3,879 previously unannotated genes absent from current reference gene models. Transcriptome-based clustering and lineage reconstruction recapitulated the continuous developmental trajectory and resolved developmental heterogeneity that was not apparent from collection time alone. Analysis of transcriptional activity identified successive waves of zygotic genome activation and revealed extensive remodeling of maternally supplied transcripts during the transition from maternal to zygotic developmental control. We further identified coordinated modules of temporally regulated genes associated with distinct developmental processes and found evidence of early dosage compensation among X-linked genes.

Together, these findings provide a high-resolution view of early medfly embryogenesis and establish a framework for investigating developmental regulation, sex determination, and chromosome biology in an agriculturally important insect species.


# Introduction

The Mediterranean fruit fly (*Ceratitis capitata*) is among the most destructive agricultural insect pests worldwide, infesting hundreds of cultivated fruit and vegetable species across tropical, subtropical, and temperate regions [1, 2]. Its remarkable ecological adaptability, broad host range, and invasive potential have made it a major target of integrated pest management programs. Among the most successful control approaches is the Sterile Insect Technique (SIT), in which mass-reared sterile males are released to suppress wild populations. The continued development of genetic and molecular strategies to enhance SIT efficacy depends upon a detailed understanding of the developmental processes that govern embryogenesis, sex determination, and early gene regulation [3–5]. Although the medfly is the most extensively studied “true” fruit fly at genomic and molecular levels [6] with a published whole genome assembly that is both of good quality and annotated [7], much remains to be understood about its embryonic development.

Embryonic development is characterized by extensive and highly coordinated changes in gene expression. Shortly after fertilization, development is initially directed by transcripts and proteins deposited in the oocyte during oogenesis. As embryogenesis proceeds, control gradually shifts from maternally supplied products to transcripts synthesized from the embryonic genome, a process commonly referred to as the maternal-to-zygotic transition (MZT) [8]. During this period, large-scale transcript degradation, de novo transcription, and developmental patterning occur simultaneously, making early embryos one of the most dynamic biological systems for studying gene regulation.

Despite the economic importance of *C. capitata*, many aspects of its earliest developmental stages remain incompletely characterized. Previous investigations have identified key components of sex determination pathways and have generated valuable genomic resources, yet the transcriptional events that accompany early embryonic development have not been resolved at high temporal resolution. In particular, the timing and progression of zygotic genome activation (ZGA), the dynamics of maternal transcript turnover, and the establishment of sex-specific developmental programs remain poorly understood.

Transcriptomic approaches have become indispensable tools for investigating developmental biology. However, conventional short-read RNA sequencing provides only partial information about transcript structure because individual reads generally cover only small portions of full-length transcripts [9, 10]. Reconstruction of complete isoforms therefore relies on computational assembly, which can be challenging in organisms with incomplete genome annotations or complex alternative splicing patterns. Long-read sequencing technologies overcome many of these limitations by generating reads that span entire transcript molecules, thereby enabling direct characterization of transcript isoforms, alternative splicing events, [11, 12] and previously unannotated genes [13–16]. These capabilities have proven particularly valuable in non-model organisms, where transcriptome annotations are often incomplete.

Recent advances in long-read sequencing have also enabled transcriptomic analyses at increasingly fine developmental resolution. Rather than averaging expression profiles across large populations of embryos, it is now possible to investigate transcriptional variation among individual embryos. This distinction is particularly important during embryogenesis because developmental progression frequently varies among embryos collected at the same nominal time point. Consequently, although chronological age has been a standard practice to perform time-course experiments [17–19], transcriptome-based approaches can provide a more accurate representation of developmental state [18].

In addition to developmental heterogeneity, sex-specific gene regulation represents another major source of biological variation during embryogenesis. In tephritid fruit flies, sex determination pathways have been studied extensively because of their relevance to genetic control strategies and male-only release programs. The discovery of the Y-linked male-determining factor *Maleness-on-the-Y* (MoY) [4] has provided new opportunities to investigate the earliest molecular events that distinguish male and female developmental trajectories. However, relatively little is known about how sex-specific transcriptional

programs emerge during embryogenesis or how mechanisms such as dosage compensation are established during early development.

Accurate quantification of developmental transcript dynamics presents additional analytical challenges. During embryogenesis, total RNA content changes substantially as maternal transcripts are degraded and zygotic transcription increases. Under these conditions, interpretation of expression data based solely on relative normalization can be difficult because apparent changes in transcript abundance may reflect global shifts in transcriptome composition rather than true biological differences [20]. Absolute quantification approaches that incorporate external RNA standards [21] provide an alternative framework for measuring transcript abundance and facilitate direct comparisons across developmental stages characterized by large-scale changes in RNA content.

In the present study, we combined long-read RNA sequencing, embryo sexing, absolute transcript quantification, and pseudotemporal trajectory analysis to investigate early embryogenesis in *C. capitata*. Single embryos were collected across the first fifteen hours of development, enabling reconstruction of developmental progression at high temporal resolution. This period encompasses 3 key embryo development stages: the maternal-to-zygotic transition (4–5 hours AEL [17]), cellular blastoderm (~ 8.5 hours AEL [19, 22]), and the beginning of gastrulation (9.5–10 hours AEL [19]). Using a genome-guided transcriptome assembly strategy, we identified thousands of previously unannotated genes and transcripts, characterized transcriptional dynamics associated with the maternal-to-zygotic transition, and examined the emergence of sex-specific expression patterns. Furthermore, we evaluated evidence for early dosage compensation and reconstructed developmental trajectories that capture embryo-to-embryo variation beyond chronological age. Together, these analyses provide a comprehensive view of early medfly development and establish a foundation for future investigations of developmental regulation, sex determination, and genetic control strategies in this globally important pest species.

# Materials and methods

### Mediterranean fruit fly rearing

Mediterranean fruit flies (*Ceratitis capitata*)'Wiederman' strain, that were used in this study are a lab strain. Flies were reared in appropriate holding cages at 25 ± 1 °C, 60 ± 10% relative humidity and 14 L:10D cycles according to the conditions previously described [23].

### Embryo collection, RNA extraction and quality control

A schematic of the workflow is shown in **Figure 1A (Supplementary Figure S1)**. To obtain synchronized samples of *Ceratitis capitata* embryos across early development, eggs were collected at hourly intervals from 0 to 15 hours after egg laying (AEL). For each time-point, 20-30 eggs were collected within a narrow 5-minute window to ensure maximal developmental synchrony. Immediately following collection, the 0-hour (time-0) embryos were flash-frozen in liquid nitrogen to halt development. For subsequent time-points, freshly laid eggs were incubated at 25 ± 1 °C for the desired duration and then flash-frozen in

liquid nitrogen at the end of each hour. This process was repeated hourly until the 15-hour time-point. Following freezing, samples were transferred to 2 mL tubes containing RNASafeguard reagent (Dominique Dutscher SAS, Bernolsheim, France) to preserve RNA integrity during storage and downstream processing. From each of the 16 time points, we randomly picked 15 embryos which were separately sheared using a pellet pestle cordless motor (Thermo Fisher Scientific) in a 0.5 mL tube. Once homogenized, 2 µL of ERCC [21] Spike-In Mix 1 (Thermo Fischer Scientific) diluted 1740 times in water were added to each sheared embryo and mixed. We initially compared the performance of two kits: Qiagen Allprep spin columns (Qiagen, Germany) and Nucleospin RNA/DNA XS kit (Macherey Nagel, Germany). We chose the Qiagen Allprep spin columns due to higher recovery rate of RNA (90% versus 10%, **Supplementary Figure S2**). From each single embryo we simultaneously extracted total RNA and DNA using the Qiagen Allprep spin columns according to manufacturer instructions. Eluted DNA was used for sexing the embryos. We only analyzed the profile of genomic DNA for embryos at 15 hours AEL. The quantity of the extracted RNA was determined using a Qubit RNA HS Assay Kit (Thermo Fischer Scientific, USA). The quality of the isolated RNA was assessed using an Agilent TapeStation instrument and Agilent RNA ScreenTape kit as per manufacturer's instruction (Agilent, CA, USA). See RNA metric for each sample in**Supplementary Table S1, Supplementary Figure S3, Supplementary Figure S4, Supplementary Figure S5, Supplementary Figure S6.**

**Embryo sexing**

We obtained five sets of primers; CcY [17], internal transcribed spacer 1 (ITS1 [24] ), Y114 [25], maleness on the Y (MoY, [4]), and glucose-6-phosphate dehydrogenase (G6PD [26]). We optimized the PCR conditions of the primers and also evaluated their suitability to determine the sex of medfly embryos (G6PD was used as a loading control). In our optimized PCR conditions, 12.5 µL of reaction consisting of 1X LongAmp Taq 2X Master Mix (NEB, USA), 0.25 µM of each of the forward and reverse primer, and DNA template were run with initial denaturation at 94 °C for 2 minutes followed by 35 cycles of denaturation at 94 °C for 1 minute, annealing at 58 °C for 1 minute, extension at 72 °C for 2 minutes and a final extension at 72 °C for 5 minutes (**Supplementary Figure S7**). In the end we selected two primer sets for sexing embryos: CcY and Y114. CcY was selected because it yields a different band for males and females. Y114 was selected because females yield no product while males yield a product. These two primer sets also had high sensitivities (**Supplementary Figure S8**). Further, we estimated that enough DNA could be extracted from single embryos to enable sexing by our optimized PCR starting from three hours after egg laying (**Supplementary Figure S9**). We also confirmed a previously reported restriction fragment length polymorphism (RFLP) in medflies [24] (**Supplementary Figure S10**). An example of embryo sexing by PCR is shown in **Supplementary Figure S11**.

**cDNA library synthesis**

All extracted total RNA was used to separately synthesize cDNA libraries for 15 embryos per time point following our recently published Panhandle protocol [27]. To amplify the cDNA, only 12 of 15 embryos per time point were selected based on their Tapestation profiles and yield. We determined the effect of

PCR cycles on cDNA profile (**Supplementary Figure S12**) and used 17 cycles of cDNA amplification for each of the 12 embryos. The quality/profile of each PCR product was assessed using the Agilent 2200 TapeStation system and the Agilent High Sensitivity D5000 ScreenTape. Quantity was estimated using Qubit dsDNA HS Assay kit (Thermo Fischer Scientific, USA).

**Library preparation, sequencing, and basecalling**

Ten embryos per time point were selected based on their total RNA profile, cDNA profile, cDNA concentration, and sex, aiming to include equal numbers of male and female embryos. Each sample was barcoded using ONT's native barcoding kit EXP-NBD196 following manufacturer instructions for the SQK-LSK108 protocol. Samples were then pooled into three batches; time points 0 – 5 hours AEL, 6 – 10 hours AEL, and 11 – 15 hours AEL. A final cleanup of pooled samples was performed using 0.7X AMPure XP beads (Beckman Coulter). **Supplementary Figure S13 s**hows the final Tapestation profile of ready-to-sequence libraries. Sequencing was performed using R9.4 flow cells on the PromethION run by MinKNOW version (3.6.5). All basecalling was performed during the sequencing run using Guppy version 3.2.10. Basecalling included classification of reads as Fail (quality score of less than 9) or Pass (quality score of 9 and above) and assignment to the embryo of origin depending on the barcode sequence added (demultiplexing). Reads whose barcodes could not be reliably identified were grouped as 'Unclassified'. See **Supplementary Table S2** for read statistics.

**Data processing**

A schematic of the data analysis workflow is shown in **Supplementary Figure S14.** Following basecalling, all Pass reads were processed with Porechop to remove sequencing adapters and barcodes followed by Pychopper to classify reads as full-length (containing adapters on both ends), rescued (chimeric reads with barcodes inside the sequence), or unclassified (single or no barcode detected). The resultant reads were combined and processed with Cutadapt [28] to remove their poly(A) tails. Reads were aligned to the NCBI *Ceratitis capitata* annotation release 103 gene models and genome assembly (GenBank GCF_000347755.3, assembly name of Ccap_2.1 [7]) supplemented with ERCC spike sequences. Off-assembly gene sequences were identified by mapping Ccap_2.1-unaligned reads onto the more recent long-read-based assembly EGII-3.2.1 (GenBank accession GCA_905071925.1 [29]). The EGII-3.2.1 assembly is a more contiguous assembly at a scaffold level with scaffold N50 of 77.38 Mb compared to 1.66 Mb for the Ccap_2.1 assembly. However, the Ccap_2.1 assembly is currently the official and annotated release and is more contiguous than EGII-3.2.1 at a contig level with contig N50 of 845 kb compared to 616 kb for EGII-3.2.1.

**Genome guided long-read transcriptome assembly**

To obtain a long-read transcriptome assembly of the medfly representative of our samples we aligned adapter and poly(A) trimmed reads from all time points reads to the Ccap_2.1 genome. The resultant aligned reads were processed by FLAIR [30] and Stringtie2 [31]. For FLAIR, samples were processed in 3 batches and their bed files combined before the final 'collapse' step. For Stringtie2, all samples were

processed in one batch. Publicly available medfly embryo short-read RNA datasets were obtained from NCBI short read archive and aligned to the genome using STAR [32] to obtain splice junction coordinates. Such junctions with over 3 supporting short reads were supplied to FLAIR to increase splice-site accuracy of the long-reads. The resultant transcripts were further collapsed using ToFU [33] in order to filter out transcripts with less than 90% alignment identity and those with less than 80% alignment coverage and to group transcripts into genes and isoforms based on their alignment coordinates. The resultant transcriptome was analyzed using SQANTI [34]. We filtered the transcripts to remove those with less than 30 long and short reads and those with over 60% adenines downstream of their transcription end site unless they possessed one or more exact splice junction matches to NCBI gene models. ORFinder was used to identify coding sequences (CDS) among transcripts in our assembly. We used custom scripts to either label these transcripts as isoforms of NCBI annotated gene models or label them as independent 'novel' genes. Reads that did not align on the Ccap_2.1 assembly were aligned to the EGII-3.2.1 assembly. We repeated the transcriptome assembly process described above. Indeed, some new genes were identified this way and here referred to as off-assembly genes. We generated a new transcriptome assembly and an updated GTF file that included the off-assembly genes, proteome, and transcriptome.

**Pseudo-time analysis**

We derived cell clusters using RaceID [35]. The RaceID algorithm was designed for single cell RNA-seq data with molecular counts derived using Unique Molecular Identifiers (UMIs). Since our data lacked UMIs, we instead used absolute gene counts derived using ERCC external standards that were sequenced as part of each sample. Firstly, we removed all genes that were not expressed in at least one embryo. We then filtered out embryos with less than 50,000 total transcript counts. We also removed all genes with less than 1000 transcript in at least one embryo. We then performed lineage tracing using FateID [35].

**Dosage compensation analysis**

To assess the chromosome homologies between *C. capitata* and *D. melanogaster*, the gene nucleotide sequences for each assembly were obtained from NCBI. Homologous gene sequences between the two organisms were found with BLASTn [36], filtering for evalue < 5e-4 and percent identity > 60. Syntenic collinear block were then detected with MCScanX (https://github.com/wyp1125/MCScanX) and visualized with SynVisio (https://github.com/kiranbandi/synvisio). To detect dosage compensation on X genes, the EGII_3.2.1 genome was split to its contig component, as this version is known to contain the X and Y sequences in the same scaffold [37]. We then mapped separate male and female short-read datasets available from SRA (ERR4026336, ERR4026337) using bwa-mem [38], selected only unique mapping reads and converted the alignment files to bedGraph using bedtools [39]. We then obtained BigWig files using bedGraphToBigWig (https://github.com/ucscGenomeBrowser/kent). The BigWig files with male and female coverage values were used as input to R-CQ with contig CAJHJT010000034.1.84 as an autosomal normalizer. Contigs having an lnMedian(R-CQ) < -0.4 were selected as X-linked.

Embryonic genes were then mapped on the contig sequences using splice-aware minimap2 [40]. We filtered only the primary alignments and selected the genes that mapped on the X-linked contigs. Similarly the genes that mapped on the contigs that comprise scaffold_5 were marked as chromosome 5-linked based on previous chromosomal assignments of the EGII_3.2.1 genome [29]. The normalized embryo counts were assigned to each X and chromosome 5 gene and we excluded those that contained genes of the rRNA cluster that is known to reside both on the X and Y chromosomes of Medfly and genes of the Ceratitis mitochondrion that were found miss-assembled with sequences of the Medfly X. Genes were then filtered based on the consistency of their expression across at least 60% of samples of the same sex that belong to a single developmental cluster. The calculation of expression ratios was performed by dividing the normalized expression values of each X gene with the average expression of chromosome 5 genes.

# Results

### Drosophila melanogaster homolog assignment

The medfly genome is diploid, consisting of six pairs of chromosomes which include a pair of heterochromatic sex chromosomes with the male being the heterogametic sex [41]. The medfly has two draft genome assemblies: Ccap_2.1 (GenBank accession GCF_000347755.3) and EGII-3.2.1 (GenBank accession GCA_905071925.1). The Ccap_2.1 assembly was generated primarily using short-reads [7] while the EGII-3.2.1 was generated primarily using long-reads [29]. Structural and functional genome annotation has been performed for the Ccap_2.1 using the NCBI Eukaryotic Genome annotation pipeline to yield the Annotation release 103. The annotation contains 14,236 genes of which 109 are pseudogenes, 12,563 are protein coding genes while 1,564 are non-protein coding. A total of 4,328 genes have isoforms yielding a total of 22,936 mRNAs. We assigned a *D. melanogaster* homologue to 11,006 (87.6 ) out of the 12,563 protein coding genes (E-value ≤ 1e-3, **Supplementary Table S3**). Of these, 18.7 were identified in the UniProtKB/Swiss-Prot database, which comprises high quality manually annotated and non-redundant proteins, while the remaining 81.3 were identified in the UniProtKB/TrEMBL database which contains high quality computationally annotated and classified proteins. Therefore, except for maleness on Y (MoY), all protein names used in reference to medfly are the derived homologs of *D. melanogaster*.

### Medfly de novo genome-guided transcriptome assembly identifies new genes and isoforms

We sequenced cDNA libraries from single medfly embryos collected at hourly intervals for the first 15 hours of development including zero time point taken immediately following egg laying. We used our optimized and published Panhandle protocol [27] aimed at capturing poly(A) + RNA (Fig. 1A). At each time point we selected 10 embryos for processing. A total of 219 million reads were generated of which 132 million (56%) were ‘Pass’ and ‘classified’ reads that were used for transcriptome assembly (see **Supplementary Table S2** for sequencing and alignment statistics). Two cDNA libraries separately generated from pooled and mixed-sex embryos at five and six hours after egg laying were also included

for transcriptome assembly. We used 3 long-read transcriptome assembly tools; Stringtie2, FLAIR, and Cupcake ToFU to derive our final transcriptome assembly. The *de novo* assembled transcriptome was analyzed using SQANTI [34].

The transcriptome assembly contained a total of 10,740 genes of which 6,861 (63.9%) matched the NCBI annotated genes, while 3,879 genes (36.1%) were new, hereafter referred to as novel genes (**Supplementary Table S4**). The inclusion criteria for novel isoforms and genes is included in Supplementary Materials. The total transcripts identified were 22,857 of which 19,329 were novel, thus, expanding the medfly transcriptome by 85% over the current NCBI annotation. Structural comparison showed that most of the transcripts in the long-read *de novo* transcriptome assembly contained new splice junctions not seen among NCBI annotated genes (NNC, 34.97%). This was followed by transcripts that contained a new combination of splice junctions already exhibited by NCBI annotated genes (NIC, 26.33%, **Supplementary Table S4**, **Fig. 1B**). Among the 3,879 novel genes that were identified, 1617 (43%) were predicted to contain an open reading frame suggesting that the rest of the novel genes were noncoding genes. Among the novel protein coding genes, only 14.6% (236) were assigned a *D. melanogaster* homolog. Further, 75% of novel genes were mono-exon compared to only 12% among NCBI annotated genes (Fig. 1C). Novel genes also contained less number of isoforms (Fig. 1D) but showed comparable expression compared to annotated genes (Fig. 1E). This suggested that genes that are non-coding, and/or are mono exon might be more likely to be missed in computational gene prediction pipelines. Noteworthy, the recently discovered mono-exon *Ceratitis capitata maleness on Y* gene (CcMoY [4]), which encodes a key maleness determining protein in medfly was identified among our mono-exon protein-coding novel genes. We further identified 371 genes that were miss-annotated as two or more separate genes but we could find single transcript reads covering these genes (**Supplementary Table S4**, **Supplementary Figure S15**). A more detailed exploration of the long-read derived transcriptome is shown in Supplementary material. We also provide an improved gene model for Doublesex showing that in the medfly, *Dsx* is a 126,100 bp gene composed of 6 exons (**Supplementary Figure S16**). Exon 5 is the key sex-differentiating exon being present in females and spliced out in males leading to a 315 and 394 amino acid protein, respectively, as reported previously [42].

# Unsupervised clustering of embryos traces developmental trajectory

To enable absolute quantification of gene expression, we added equal amounts of ERCC external RNA control mixtures, containing spike-in RNAs of known copy number, to each embryo sample as previously described [20, 43] (**Supplementary Table S5** and **Supplementary Table S6**). This approach, known as absolute quantification or direct absolute normalization, allowed us to estimate transcript copy numbers per gene per embryo. As shown in **Supplementary Figure S17**, this method outperforms conventional relative quantification, especially in transcriptionally dynamic systems like developing embryos. This method was further validated by empirically measured amounts of cDNA generated per embryo (**Supplementary Figure S18**). We further examined pairwise correlation of embryos and show that

indeed, adjacent time points had higher Spearman correlation than distant time points, except for the 8 hour time point which showed some anomalies (**Supplementary Figure S19**, **Supplementary Table S7**). Across all our samples, the median limit of detection was 19,000 transcripts per embryo (**Supplementary Table S8**). We then used RaceID and FateID algorithms developed for unsupervised clustering and lineage tracing in single-cell RNA-seq, respectively, to achieve the same objective with our single embryos. A similar application in Drosophila has demonstrated the feasibility and utility of this approach [18]. Although we grouped our embryos into 16 categories based on one-hour sampling intervals following egg laying, RaceID returned 10 clusters (Fig. 2A). Eleven embryos were excluded from downstream analysis based on the ectopic early expression of previously validated late-expressed genes, such as *sisA* and *slam* [17], as well as other Drosophila-validated markers [44]. These outlier embryos were the sole expressers of these genes within their clusters, suggesting developmental misclassification. Clustering was then repeated, yielding largely consistent cluster assignments and supporting the stability of the grouping (Fig. 2B). Expression analysis of housekeeping genes confirmed robust transcriptional activity across all embryos (**Supplementary Figure S20**). Embryo sexing showed that as expected, embryos in cluster 1 were mostly not sexed and sexing improved as development progressed (**Supplementary Figure S21**). Lineage tracing showed a single trajectory starting from the least developed to the most advanced embryos (Fig. 2C). A comparison of real-time and pseudo-temporal grouping showed largely consistent results (Fig. 2D). For example, cluster 1 contained embryos from 0–3 hours AEL and cluster 9 contained embryos from 11–13 and 15 hours AEL. Nonetheless, certain deviations highlighted the limitations of real-time-based staging. Notably, cluster 10, which was projected as containing the most advanced embryos, contained only embryos collected 14 hours AEL. In this case, the embryos collected 15 hours AEL would have been expected to be the most advanced. These findings highlight the additional resolution gained by transcriptome-based staging over conventional time after-egg-laying classification.

# Medfly embryos undergo successive waves of zygotic genome activation

In Drosophila, embryos undergo two waves of zygotic genome activation (ZGA): a minor wave and a major wave. The minor wave, which lasts from about 1 to 2 hours after fertilization [8], is characterized by transcription of about 20 genes by nuclear cycle 9 (NC9) and another 43 by NC10 [45, 46]. The major wave which starts around 2.5 hours after fertilization during NC14 is characterized by transcription of about 3540 genes [45]. We probed the expression of *C. capitata* homologs of these *D. melanogaster* ZGA genes. Out of the 20 early minor ZGA *D. melanogaster* genes, eight had medfly homologs all of which had early expression data (**Supplementary Table S9**). Among these 8 genes, 4 (*sisA*, *CG13712*, *CG6885*, *amos*) were indeed first expressed by embryos in cluster 5 (Fig. 3A), 3 (*E(spl)m4-BFM*, *l(1)sc*, and *sc*) were first expressed by embryos in cluster 6, while *tsg* was clearly maternally deposited, being expressed by embryos in cluster 1 (**Supplementary Figure S22**).

Among the 43 Drosophila late minor ZGA genes, we had 22 *C. capitata* homologs of which 19 had expression. Among these 19 genes, 9 (47%) were first expressed by embryos in cluster 5 while 6 (32%)

were first expressed by embryos in cluster 6. Two were expressed in cluster 1 while 2 had sporadic expression in only a few embryos. We therefore, conclude that cluster 5 marks the first wave of zygotic genome activation in our medfly embryos.

Out of the 3542 *D. melanogaster* major ZGA genes, we had *C. capitata* homologs for 2535 genes (72%). Among these homologs, we had expression for 1477 genes of which 918 were expressed in cluster 1 while 79 were part of the minor wave of ZGA and 147 were indeed, expressed in cluster 6 (Fig. 3B). We therefore, conclude that cluster 6 marks the second wave of zygotic genome activation in our medfly embryos

We then aggregated all genes that were used to mark the two waves of ZGA in medfly and probed their expression on a pseudotime scale. Indeed, genes involved in the first wave of ZGA commenced expression earlier than genes in the second wave of ZGA (pseudotime 40 versus 60, **Fig. 3C** and **Fig. 3D**, respectively). We further performed qPCR in a new set of embryos collected at different time points to validate the expression of 5 genes known to play a role in early embryo development. qPCR results indeed, validated the profiles we observed using pseudotemporal analysis (Fig. 3E-F).

We also identified other 334 genes in Cluster 5 that had evidence of being zygotic which could now be new markers of the first wave of ZGA in medfly (**Supplementary Table S9, Supplementary Figure S23**). Cluster 5 was mostly made up of embryos taken at 4 hours AEL. In total, we find 930 genes to be expressed in Cluster 5 and 4 and thus are part of the first ZGA wave in the medfly. Among genes active in the medfly second wave of ZGA, we identified a total of 919 genes expressed in Cluster 6 and 7 that had evidence of being zygotic which we also argue are part of the medfly second wave of zygotic genome activation (**Supplementary Table S9**). These data suggest that medfly embryos undergo two waves of zygotic genome activation clearly marked by clusters 5 and 6, respectively. Clusters 1, 2, and 3 represent embryos that only contain maternally deposited transcripts prior to zygotic genome activation.

Further, we observed that the maternally deposited genes had longer gene lengths compared to zygotic genes (median of 1.8 kb versus 1.5 kb, respectively) although this difference did not reach statistical significance (Wilcoxon p-value, 0.05, **Supplementary Figure S24**). There was no significant difference between the transcript lengths of maternally deposited genes and zygotic genes (Wilcoxon p-value, 0.75). However, the number of exons and total intron lengths of zygotic genes were significantly lower compared to maternally deposited genes (Wilcoxon p-values < 2.2e-16 for both). These data suggested that zygotic genes tend to be short and have less exons.

# Pseudo-temporal clustering highlights relevant biological processes

We performed differential gene expression to determine significantly up or down regulated genes in clusters 1 through 9 (Fig. 4A). We combined all embryos in clusters 1 and 3 and compared them to embryos in clusters 5 and 4. We then compared embryos in clusters 5 and 4 against cluster 6 (Fig. 4B). Finally, we compared embryos in cluster 6 against cluster 7 (Fig. 4C), and cluster 7 against cluster 9

(Fig. 4D). We then used the significantly down or upregulated genes to perform gene set enrichment analysis using *D. melanogaster* as the reference organism. Relative to cluster 5/4, cluster 1/3 was enriched in metabolic process, mitochondrial gene expression, cellular processes and translation. In contrast, cluster 5/4, where we detected the first expression of Drosophila minor ZGA genes, was enriched in regionalization, pattern specification, multicellular organism development, anatomical structure development, and system development processes.

Cluster 6, where we first detected the Drosophila major ZGA genes, was enriched in gene expression, RNA metabolic processes, cellular organization, nucleic acid metabolism, and biosynthetic processes. Cluster 7 was enriched in animal organ development, system development, organ morphogenesis, multicellular development, and cellular development process. Cluster 9 was enriched in nervous system development, cell membrane projection organization, anatomical structure development, system development, and cell differentiation.

Cluster 5 was very interesting as it both showed the first wave of ZGA and was enriched in critical developmental and differentiation processes. We thus probed factors that might be triggering this activation. In Drosophila, the transcription factor Zelda is known to be a major activator of the minor wave of ZGA. The promoter regions of genes transcribed during the Drosophila minor wave of ZGA are enriched in the motif CAGGTAG which is bound by Zelda to mediate the expression of these genes. In the olive fruit fly, we reported a similar finding [43]. However, probing the promoters of genes expressed in the medfly minor wave of ZGA did not return Zelda motifs. However, it is known that many genes do not require Zelda for their regulation during ZGA. Ibarra-Morales et al., [44] showed that in Drosophila, the transcription start sites of up to 65% of zygotic genes have histone 2A Z (H2A.Z) enrichment. They further showed that knockdown of H2A.Z causes down regulation of a host of genes and compromises 3D chromatic structure. We obtained *D. melanogaster* genes annotated as to whether they contain Zelda motif or are enriched with H2A.Z or not (H2A.Z + and H2A,Z-, respectively), or have discordant results [44, 47]. For each of the significantly down or upregulated genes in our medfly clusters, we assigned them to their corresponding *D. melanogaster* annotation; Zelda motif, H2A.Z+, H2A.Z-, or discordant. There was an 8% increase in the Zelda motif enriched genes among cluster 5/4 upregulated genes compared to cluster 1/3 while the number of H2A.Z enriched genes decreased from 80% to 47% (**Supplementary Figure S25**). This suggested that neither Zelda nor H2A.Z could explain the mechanism of zygotic genome activation in the medfly. Further into development there was neither clear enrichment of Zelda nor H2A.Z among upregulated genes further suggesting that these mechanisms probably do not play a significant role in medfly zygotic genome activation. The mechanism behind zygotic genome activation in medfly embryos, therefore, remains elusive.

We also used the real-time temporal data and performed temporal gene expressions using DPGP [20, 48] which jointly models data clusters with a Dirichlet process and temporal dependencies with Gaussian processes. We identified 200 gene clusters with differing profiles suggesting specific roles for these clusters during defined developmental periods (**Supplementary Table S10, Supplementary Table S11, Supplementary Table S12**, Supplementary material). We then assigned the clusters to 4 groups

(**Supplementary Figure S26**) and found that indeed zygotic genes recapitulate the critical developmental biological processes as seen with the pseudotemporal clustering. Taken together, these results show that cluster 5 marks the switching of key developmental process in our medfly embryos and that the mechanism of this trigger is different in medflies compared to Drosophila with respect to Zelda and H2A.Z.

# Maternal to zygotic transition in medfly shows dramatic reorganization of maternal transcripts

During egg development, the female insect deposits large amounts of proteins and mRNA and other resources into the oocyte. The zygote depends on these maternal transcripts for its early development. The current medfly NCBI gene annotation contains 12,563 protein coding genes. Among these protein coding genes, 2713 (21.6%) were detected at 0 hours AEL (detection limit of 1000 TPE). Among these, the most abundant genes we could assign *D. melanogaster* homologs to were gene7923 and gene6477 which encode Ubiquitin-63E and ribosomal protein L39 and accounted for 1.6% and 1.1% of total transcripts per embryo at 0 hours AEL, respectively. We then examined the change in the number of poly(A) transcripts in the developing medfly embryo during its first 15 hours of development. We noticed that the number of transcripts per embryo averaged $0.68 \times 10^9$ (± 0.14) at 0 hour AEL and then increased to $1.6 \times 10^9$ (± 0.49) at 1 hour AEL before dropping to $0.46 \times 10^9$ (± 0.09) at 2 hours AEL (Fig. 5A). Between 2 hours AEL and 3 hours AEL poly(A) transcripts increased to $1 \times 10^9$ (± 0.52). We also mathematically converted the number of transcripts per embryo to nanograms of mRNA per embryo and observed a similar trend as would be expected (**Supplementary Figure S27**). By comparison, *D. melanogaster* embryos contain ~ $3 \times 10^9$ copies of maternal poly(A) transcripts [49] during early development while the closely related *B. oleae* embryos contains $1.5 \times 10^9$ poly(A) RNA transcripts at 1 hour AEL [43].

We hypothesized that the embryo must be changing the distribution of poly(A) transcripts prior to their maternal-to-zygotic degradation. To test this, we compared the abundance of transcripts at 1 hour AEL relative to the abundance of the same transcripts at 3 hours AEL. We then sorted the transcripts from the most abundant to the least abundant between 1 and 2 hours AEL (Fig. 5B). We noticed that indeed, there is a re-organization of maternally deposited transcripts such that the least abundant transcripts at 2 hours AEL have a higher abundancy at 3 hours relative to their abundance at 1 hour AEL. We think that this is a deliberate re-organization of maternal transcripts prior to their degradation. Noteworthy, our RNA-seq protocol utilized an oligo(dT) primer for cDNA synthesis and thus, the dynamics observed in the first 3 hours of medfly embryo development relate to poly(A) transcripts. Starting from 3 hours AEL onwards, we observed a gradual reduction in abundance of transcripts per embryo and mRNA from $1 \times 10^9$ (± 0.52) at 3 hours AEL to $0.1 \times 10^9$ (± 0.04) at 15 hours AEL, a 90% drop (Fig. 5A). We believe that this is a reflection of the degradation of maternal transcripts as part of maternal-to-zygotic transition (MZT). In Drosophila, MZT involves destabilization of up to 35% of maternally supplied transcripts by

maternally encoded proteins by the end of 3 hours after fertilization [8] while in *B. oleae*, we previously detected a 36% drop in transcripts per embryo by 6 hours AEL.

To elucidate the mechanism of MZT, we returned to the clustering of embryos we performed based on their gene expression. First, we performed a differential gene expression between clusters 1 and 5 and identified significantly down regulated genes. We then ranked the expression pattern of these genes in cluster 1 and put them into four quartiles, Q1, Q2, Q3, Q4, respectively. We then summed the number of transcripts per embryo for all genes in each quartile across all clusters. We noted that the abundance of genes gradually reduced from cluster 1 through cluster 6 (Fig. 5C). We believe that this is a reflection of the gradual degradation of maternal transcripts as observed in other organisms during maternal to zygotic transition. We then determined the ratio of the total transcripts in each quartile in cluster 1 to that in the respective quartile of Cluster 6, to determine the rate of degradation (Fig. 5D). Here, we noted a rate (ratio) of 4, 6, 11, and 18 for Q1, Q2, Q3, Q4, respectively. This suggested that the rate of degradation of maternal transcripts was proportional to the initial abundance of transcripts in Cluster 1. Thus, transcripts in quartile 4, which contained the most abundant transcripts in cluster 1 had a 4 fold higher rate of degradation compared to transcripts in quartile 1. On the contrary, transcripts that had shown no significant differential expression showed similar degradation rates regardless of the quartile (Fig. 5D).

Further, we investigated the presence of X chromosome dosage compensation in our data. We identified scaffolds and contigs in the genome assembly that correspond to X chromosome and autosomes, from which we detected 762 transcripts that derive from the medfly X. We then used the sexing information and pseudotime clustering of embryos to filter X transcripts for consistency of expression across samples of the same sex that belong in the same developmental cluster. This resulted in 380 X genes for which we calculated the per-cluster sex-specific expression and compared it to the expression levels of genes from the autosomal chromosome 5. The results indicate that expression from the X initiates during the minor wave of ZGA at Cluster 5 & 4 and stably increases across the developmental trajectory of the embryos. A comparable ratio of X:Chr5 expression was observed between male and female samples from Cluster 4 and onwards (Fig. 6, Supplementary Fig. 28), suggesting the presence of dosage compensation on X-linked genes, even from the early onset of transcription activation. No sexing information was available for samples of Cluster 5, thus we cannot exclude a sex biased expression during the initiation of the minor wave of ZGA. Assuming that the chromosome 5 transcript levels represent the diploid expression state, the X gene expression peak at Cluster 9 and 10 indicates that dosage compensation might act towards the direction of the hemizygous state. During Cluster 9, the median X gene expression is almost half that of chromosome 5. During that time-point, females express X genes at ~ 52% of the chromosome 5 levels, with the respective male X expression being at ~ 55%. Similarly for Cluster 10, the female X expression is at ~ 72% with males expressing the X at ~ 62% relative to chromosome 5 levels (Fig. 6).

.

Taken together, these results suggest that there's a dramatic reorganization of maternal transcripts in the first 3 hours after egg laying, prior to MZT. Further, MZT happens gradually at a rate commensurate with abundance of transcripts. Finally, the expression of X genes appears to follow the MZT and is compensated early on ZGA between male and female embryos.

# Discussion

We provide a high-resolution view of early embryogenesis in *Ceratitis capitata* through the integration of long-read RNA sequencing, absolute transcript quantification, embryo sexing, and transcriptome-based developmental trajectory reconstruction. By profiling individual embryos across the first fifteen hours of development, we resolved developmental heterogeneity that is obscured when embryos are grouped solely according to collection time. The resulting transcriptomic framework enabled characterization of transcriptional dynamics associated with the maternal-to-zygotic transition, identification of previously unannotated genes and transcript isoforms, and investigation of sex-specific developmental processes.

A major advantage of the present study is the use of single embryo transcriptomes to reconstruct developmental progression using unsupervised clustering [18, 19]. Embryonic development is inherently asynchronous, and embryos collected at identical time points frequently differ in developmental state. Our pseudotime analyses revealed a continuous developmental trajectory that captured transcriptional progression more effectively than chronological age alone. We identified 10 distinct clusters. We excluded 11 embryos based on their aberrant expression of *sisA*, *slam*, *tup*, and *NetA*; genes known to be expressed later in embryo development. Gabrieli et al., showed previously that *sisA* and *slam* are not supplied maternally but are encoded by the zygote [17] and thus could be used to filter out embryos. Similar observations have been reported in other developmental systems, where transcriptome-derived trajectories provide improved resolution of developmental transitions compared with conventional time-course analyses [18]. These findings highlight the value of single-embryo approaches for studying dynamic developmental processes and suggest that transcriptome-based staging may complement or, in some contexts, surpass morphology-based classification methods.

The long-read sequencing strategy substantially expanded existing transcriptomic resources for *C. capitata*. Genome-guided transcriptome reconstruction identified 22,857 transcripts and 3879 previously unannotated genes absent from current reference annotations. This is in line with our previous olive fruit fly work where we identified 10,840 genes [43] and in Drosophila 0-3 hour embryos where 9,777 genes were reported [18]. Improved transcript models were particularly valuable for genes exhibiting complex exon structures and alternative splicing patterns. The ability of long-read technologies to capture full-length transcript structures directly has transformed transcriptome characterization in both model and non-model organisms, reducing ambiguity associated with transcript reconstruction from short-read data and improving the accuracy of gene annotations [13, 18]. Our findings demonstrate that substantial transcriptomic complexity remains to be characterized especially in species without well-established genomic resources.

The developmental time course captured extensive transcriptome remodeling during the maternal-to-zygotic transition; the clearance of a large proportion of maternal transcripts and proteins originally deposited into the oocyte during oogenesis, and the initiation of zygotic transcription [8]. Past studies have shown that maternal transcripts represent up to 75 % of total protein-coding genes in *Drosophila melanogaster* [46] and Danio relio (zebrafish) [50], and 62% in the olive fruit fly [43]. In the current work, we detected a lower number of protein coding genes of only 21.6% at 0 hours AEL, likely due to insufficient sequencing depth to capture many of the low expressed genes. In the current work, we measured a 90% reduction in transcripts per embryo between 3 and 15 hours AEL, in contrast to the previously reported amounts of cleared transcripts in Drosophila (35%, [8]) and *B. oleae* (36%, [43]). It is likely that the shorter timeframe we studied in olive fruit flies and Drosophila explains the difference.

Absolute transcript quantification revealed pronounced changes in transcript abundance throughout early embryogenesis, reflecting coordinated degradation of maternally deposited transcripts and progressive activation of zygotic transcription. Very similarly to our previous work with the olive fruit fly [43], we noticed dramatic and rapid fluctuations in mRNA content per embryo within the first 3 hours of embryo development. These findings represent, to our knowledge, the first report of such dramatic mRNA content fluctuations during early embryogenesis in insects. The appearance of this phenomenon in two tephritid species, occurring at the same developmental time post oviposition, suggests that it is a conserved feature warranting further molecular investigation to delineate its mechanism and biological significance. We believe that this dramatic shift in embryo mRNA content precedes MZT initiation and occurs through modulation of the poly(A) tails of transcripts by poly-adenylation and de-adenylation processes as reported in Drosophila [51-57]. Following this dramatic event, we noticed a more gradual drop in embryonic mRNA content from 4 hours AEL till the end of our sampling at 15 hours AEL. We believe that this is a reflection of the MZT process initiating at 4 hours AEL. We further explored the manner of maternal transcript degradation by comparing the rate of degradation as a function of initial gene expression level. Our data indicate a gradual degradation process, with the most highly expressed genes having the fastest rate of degradation. This contrasts a report in *D. melanogaster* where the rate of degradation was similar across highly and lowly expressed genes [44]. Our findings suggests that the degradation machinery may act non-specifically, disproportionally affecting the most abundant transcripts, possibly to reduce the levels of the most highly expressed genes as the embryo prepares for cellularization. These observations are consistent with the MZT described in numerous metazoan species and support the view that maternal transcript clearance and zygotic genome activation represent overlapping rather than strictly sequential processes [8, 49]. Further, our work suggests that transcriptome remodeling begins prior to the onset of widespread zygotic transcription.

Several clusters of genes displayed coordinated temporal expression patterns corresponding to distinct developmental intervals. Genes enriched during early stages were predominantly associated with maternally supplied functions, whereas genes activated later were linked to developmental patterning, cellular differentiation, and embryonic morphogenesis suggesting zygotic genome activation (ZGA); the initiation of transcription from the genome of the zygote. In *D. melanogaster*, ZGA start around 1-2 hours after oviposition [8] and occurs in two successive waves involving at leat 3560 genes [45]. We provide

strong evidence that medfly embryos also undergo two waves of zygotic genome activation. Gabrieli et al. [17] reached the same conclusion although they found the first wave to start before 4 hours after oviposition while the second wave started 5 after hours after oviposition coinciding with the migration of nuclei to the posterior of the embryo and establishment of pole cells [18, 58]. Although Zelda is known to play a key role in ZGA in drosophila [59, 60] and the olive fruit fly [43], we did not find the well-known Zelda motif “CAGGTAG” among the medfly early zygotic genes. Gabrieli et al. [17] have also reported lack of TAGteam sites in *Ceratitis capitata* based on *sry-α*. Further, we also found that many well-known *D. melanogaster* early zygotic genes such as *sna*, *kni*, *nullo*, *dpp*, *gt*, *odd*, and *nrt* are maternally supplied in *C. capitata*. Our data suggests that although *Zelda* activates early zygotic genes in drosophilids and *B. oleae* [43], it does not in *C. capitata*. The temporal structure observed here provides a framework for future functional studies aimed at identifying regulators that govern early developmental transitions in tephritid fruit flies.

One notable outcome of the present analysis was the ability to investigate sex-specific developmental processes during embryogenesis. By combining transcriptomic profiling with embryo sex determination, we identified transcriptional differences between male and female embryos and tracked the developmental emergence of sex-associated gene expression patterns. The detection of *Maleness-on-the-Y* (MoY) [4] expression within developing embryos further validates the utility of the approach for studying sex determination pathways. Given the central role of sexing technologies in sterile insect technique programs, improved understanding of early sex determination mechanisms may contribute to future genetic control strategies.

Beyond its biological significance, improved understanding of early embryonic transcriptional regulation has practical implications for pest management. The Mediterranean fruit fly remains one of the most economically damaging agricultural pests worldwide, and continued development of genetic control approaches depends upon detailed knowledge of developmental pathways. Early embryogenesis represents a particularly attractive target because many processes critical for viability, sex determination, and reproductive development are established during this period. The transcriptomic resources generated in this study therefore provide a foundation for future investigations aimed at identifying candidate genes and regulatory pathways that may be exploited for genetic control strategies, including improvements to sterile insect technique programs and the development of next-generation molecular interventions.

Several observations from this study also raise new questions regarding the regulation of embryonic gene expression. The dynamic transcript abundance changes observed during the earliest developmental intervals suggest that extensive transcriptome remodeling occurs before widespread zygotic transcription becomes established. Similar patterns have been reported in other developmental systems and may reflect coordinated degradation, processing, or post-transcriptional regulation of maternally deposited transcripts [53, 57]. Further investigation of these early events may reveal previously unrecognized regulatory mechanisms that contribute to developmental competence and successful transition to embryonic genome control.

Finally, we also provide a profiling of the X-linked Medfly gene expression and the first evidence for the presence of X chromosomes dosage compensation between male and female embryos. Although dosage compensation has been extensively described in *D. melanogaster*, the same principles do not apply in medfly, or Tephritids in general, as different Muller elements segregate as X chromosome between the two species (Supplementary Figure 7). Our results indicate that dosage compensation in Medfly acts early in development during the minor wave of ZGA (Figure 6, Supplementary Figure 8), in contrast to the progressive activation in *D. melanogaster* [18]. However, due to limitations in the sexing information we cannot support the compensated expression of the few X genes expressed during cluster 5 (Supplementary Figure 9). In addition, we compared the medfly X gene expression to that of chromosome 5 that is homologues to *D. melanogaster*'s X, for which dosage compensation acts towards the diploid state. We show that in our datasets the expression levels from Medfly's X reach almost half the level of chromosome 5, suggesting a direction of dosage compensation toward the hemizygous state (Figure 6). Nevertheless, since we observe an increasing X expression trend towards cluster 10, we cannot exclude the further increase of X gene expression up to the dosage compensation at the diploid level.

Although the present study substantially expands our understanding of medfly embryogenesis, several limitations should be acknowledged. While hourly sampling provided relatively high temporal resolution, developmental transitions occurring on shorter timescales may not have been fully captured. Additionally, transcriptomic analyses identify associations between gene expression patterns and developmental processes but do not directly establish causality. Functional validation through gene perturbation, genome editing, or transgenic approaches will be necessary to determine the specific roles of candidate regulatory genes identified here. Future integration of transcriptomic, epigenomic, proteomic, and spatial expression datasets may provide a more comprehensive understanding of the molecular events governing early development.

In summary, this study establishes a comprehensive transcriptomic framework for investigating early embryogenesis in *Ceratitis capitata*. By integrating long-read sequencing, absolute transcript quantification, embryo sex determination, and developmental trajectory reconstruction, we generated a detailed view of transcriptional regulation during the maternal-to-zygotic transition. The identification of thousands of previously unannotated genes, characterization of developmental expression programs, evidence for early dosage compensation, and resolution of sex-specific transcriptional dynamics collectively advance our understanding of medfly developmental biology. These findings provide both a valuable genomic resource and a foundation for future studies of embryogenesis, sex determination, and genetic control in this globally important agricultural pest.

# Declarations

### Conflict of interest

JR was a member of the MinION Access Program (MAP) and has received free-of-charge flow cells and sequencing kits from Oxford Nanopore Technologies for other projects. JR has had no other financial support from ONT. AB and JR have received reimbursement for travel costs associated with attending the Nanopore Community meetings organized by Oxford Nanopore Technologies. The rest of the authors do not have competing interests.

**Acknowledgements**

This study benefited from discussions at International Atomic Energy Agency funded meetings for the Coordinated Research Project “Generic approach for the development of genetic sexing strains for SIT applications”, IAEA (CRP No.: D4.40.03) Vienna, Austria. Particularly, we would like to thank Kostas Bourtzis for discussions and encouragement.

**Author contribution statement**

AB, SO, DR participated in study design, conducted the experiments, participated in the data analysis and drafting of the manuscript. DM helped in nucleic acid extraction. AS was responsible for colony rearing and synchronized embryo collections. KDM and JR conceived the study, sourced funding, and oversaw the project execution. All authors read, corrected, and approved the final version of this article.

**Ethics statements**

**Studies involving animal subjects:** No ethical approval was required for this study, as the experimental procedures involved invertebrates (*Ceratitis capitata*), which are not covered by animal ethics legislation.

**Studies involving human subjects:** No human studies are presented in this manuscript.

**Inclusion of identifiable human data:** No potentially identifiable human images or data is presented in this study.

**Data availability statement**

The datasets presented in this study have been deposited to NCBI’s SRA with project number PRJNA1328737. The individual SRR numbers for each dataset can be found in Supplementary Table 14.

**Funding**

This study was supported by a Genome Canada Technical Development grant and by the International Atomic Energy Authority Program of Coordinated Activities CRP 044003, grant number 23358 to JR, and grant number 23378 awarded to KM. JR is funded by the Canada Foundation for Innovation (CFI) and the CFI Leaders Opportunity Fund (32557), Compute Canada Resource Allocation Project (WST-164-AB) and Genome Innovation Node (244819). This study was further supported by the postgraduate program of the Department of Biochemistry and Biotechnology of the University of Thessaly “Advanced

Experimental and Computational Biosciences” and the European Union’s Horizon Europe Research and Innovation Programme REACT (Grant agreement 101059523).

# References

1. Malacrida AR, Gomulski LM, Bonizzoni M, Bertin S, Gasperi G and Guglielmino CR. Globalization and fruitfly invasion and expansion: the medfly paradigm. Genetica. 2007;131 1:1-9. doi:10.1007/s10709-006-9117-2.
2. Szyniszewska AM and Tatem AJ. Global assessment of seasonal potential distribution of Mediterranean fruit fly, Ceratitis capitata (Diptera: Tephritidae). PloS one. 2014;9 11:e111582. doi:10.1371/journal.pone.0111582.
3. Caceres C. Mass rearing of temperature sensitive genetic sexing strains in the Mediterranean fruit fly (Ceratitis capitata). Genetica. 2002;116 1:107-16. doi:10.1023/a:1020967810703.
4. Meccariello A, Salvemini M, Primo P, Hall B, Koskinioti P, Dalíková M, et al. Maleness-on-the-Y (MoY) orchestrates male sex determination in major agricultural fruit fly pests. Science (New York, NY). 2019;365 6460:1457-60. doi:10.1126/science.aax1318.
5. Davydova S, Yu D and Meccariello A. Genetic engineering for SIT application: a fruit fly-focused review. Insect Sci. 2025; doi:10.1111/1744-7917.70038.
6. Scolari F, Gomulski LM, Gabrieli P, Manni M, Savini G, Gasperi G, et al. How functional genomics will impact fruit fly pest control: the example of the Mediterranean fruit fly, Ceratitis capitata. BMC genetics. 2014;15 Suppl 2:S11. doi:10.1186/1471-2156-15-s2-s11.
7. Papanicolaou A, Schetelig MF, Arensburger P, Atkinson PW, Benoit JB, Bourtzis K, et al. The whole genome sequence of the Mediterranean fruit fly, Ceratitis capitata (Wiedemann), reveals insights into the biology and adaptive evolution of a highly invasive pest species. Genome biology. 2016;17 1:192. doi:10.1186/s13059-016-1049-2.
8. Tadros W and Lipshitz HD. The maternal-to-zygotic transition: a play in two acts. Development. 2009;136 18:3033-42. doi:10.1242/dev.033183.
9. Angelini C, De Canditiis D and De Feis I. Computational approaches for isoform detection and estimation: good and bad news. BMC bioinformatics. 2014;15:135. doi:10.1186/1471-2105-15-135.
10. Steijger T, Abril JF, Engstrom PG, Kokocinski F, Hubbard TJ, Guigo R, et al. Assessment of transcript reconstruction methods for RNA-seq. Nature methods. 2013;10 12:1177-84. doi:10.1038/nmeth.2714.
11. Oikonomopoulos S, Wang YC, Djambazian H, Badescu D and Ragoussis J. Benchmarking of the Oxford Nanopore MinION sequencing for quantitative and qualitative assessment of cDNA populations. Scientific reports. 2016;6:31602. doi:10.1038/srep31602.
12. Weirather JL, de Cesare M, Wang Y, Piazza P, Sebastiano V, Wang XJ, et al. Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. F1000Research. 2017;6:100. doi:10.12688/f1000research.10571.2.

13. Byrne A, Beaudin AE, Olsen HE, Jain M, Cole C, Palmer T, et al. Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. Nature communications. 2017;8:16027. doi:10.1038/ncomms16027.
14. Clark M, Wrzesinski T, Garcia-Bea A, Kleinman J, Hyde T, Weinberger D, et al. Long-read sequencing reveals the splicing profile of the calcium channel gene CACNA1C in human brain. Mol Psychiatry. 2020;25 1:37. doi:10.1038/s41380-019-0583-1.
15. Liu H, Smith TPL, Nonneman DJ, Dekkers JCM and Tuggle CK. A high-quality annotated transcriptome of swine peripheral blood. BMC genomics. 2017;18 1:479. doi:10.1186/s12864-017-3863-7.
16. Singh N, Sahu DK, Chowdhry R, Mishra A, Goel MM, Faheem M, et al. IsoSeq analysis and functional annotation of the infratentorial ependymoma tumor tissue on PacBio RSII platform. Meta gene. 2016;7:70-5. doi:10.1016/j.mgene.2015.11.004.
17. Gabrieli P, Falaguerra A, Siciliano P, Gomulski LM, Scolari F, Zacharopoulou A, et al. Sex and the single embryo: early deveiopment in the Mediterranean fruit fly, Ceratitis capitata. BMC developmental biology. 2010;10:12. doi:10.1186/1471-213x-10-12.
18. Pérez-Mojica JE, Enders L, Walsh J, Lau KH and Lempradl A. Continuous transcriptome analysis reveals novel patterns of early gene expression in Drosophila embryos. Cell Genom. 2023;3 3:100265. doi:10.1016/j.xgen.2023.100265.
19. Stefani R. N. D. SD, and Perondini A. L. P. Early developmental stages of Ceratitis capitata embryos. In: N. BB, editor. Proceedings of the 6th International symposium on fruit flies of economic importance 6-10 May 2002. Stellenbosch, South Africa: Isteg Scientific Publications, Irene, South Africa; 2004. p. 55-8.
20. Owens NDL, Blitz IL, Lane MA, Patrushev I, Overton JD, Gilchrist MJ, et al. Measuring Absolute RNA Copy Numbers at High Temporal Resolution Reveals Transcriptome Kinetics in Development. Cell reports. 2016;14 3:632-47. doi:10.1016/j.celrep.2015.12.050.
21. Baker SC, Bauer SR, Beyer RP, Brenton JD, Bromley B, Burrill J, et al. The External RNA Controls Consortium: a progress report. Nature methods. 2005;2 10:731-4. doi:10.1038/nmeth1005-731.
22. Schetelig MF, Horn C, Handler AM and Wimmer EA. Development of an Embryonic Lethality System in Mediterranean Fruit Fly Ceratitis capitata. In: *Area-Wide Control of Insect Pests* (eds Vreysen MJB, Robinson AS and Hendrichs J), Dordrecht, 2007// 2007, pp.85-93. Springer Netherlands.
23. Mintzas AC, Chrysanthis G, Christodoulou C and Marmaras VJ. Translation of the mRNAs coding for the major hemolymph proteins of Ceratitis capitata in cell-free system: Comparison of the translatable mRNA levels to the respective biosynthetic levels of the proteins in the fat body during development. Developmental Biology. 1983;95 2:492-6. doi:https://doi.org/10.1016/0012-1606(83)90051-9.
24. Douglas LJ, Untalan PM and Haymer DS. Molecular sexing in the Mediterranean fruit fly, Ceratitis capitata. Insect biochemistry and molecular biology. 2004;34 2:159-65. doi:10.1016/j.ibmb.2003.08.006.

25. Zhou Q, Untalan PM and Haymer DS. Repetitive A-T rich DNA sequences from the Y chromosome of the Mediterranean fruit fly, Ceratitis capitata. Genome. 2000;43 3:434-8. doi:10.1139/g99-133.
26. He M and Haymer DS. Genetic relationships of populations and the origins of new infestations of the Mediterranean fruit fly. Molecular ecology. 1999;8 8:1247-57. doi:10.1046/j.1365-294x.1999.00685.x.
27. Bayega A, Oikonomopoulos S, Wang YC and Ragoussis J. Improved Nanopore full-length cDNA sequencing by PCR-suppression. Front Genet. 2022;13:1031355. doi:10.3389/fgene.2022.1031355.
28. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. 2011. 2011;17 1:3. doi:10.14806/ej.17.1.200.
29. Ward CM, Aumann RA, Whitehead MA, Nikolouli K, Leveque G, Gouvi G, et al. White pupae phenotype of tephritids is caused by parallel mutations of a MFS transporter. Nature communications. 2021;12 1:491. doi:10.1038/s41467-020-20680-5.
30. Tang AD, Soulette CM, van Baren MJ, Hart K, Hrabeta-Robinson E, Wu CJ, et al. Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. Nature communications. 2020;11 1:1438. doi:10.1038/s41467-020-15171-6.
31. Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL and Pertea M. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome biology. 2019;20 1:278. doi:10.1186/s13059-019-1910-1.
32. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics (Oxford, England). 2013;29 1:15-21. doi:10.1093/bioinformatics/bts635.
33. Elizabeth T: ToFu: Identifying unique isoforms from long-read RNA-seq data. https://github.com/Magdoll/cDNA_Cupcake/wiki/ Accessed 14 Aug 2023.
34. Tardaguila M, de la Fuente L, Marti C, Pereira C, Pardo-Palacios FJ, Del Risco H, et al. SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. Genome research. 2018;28 3:396. doi:10.1101/gr.222976.117.
35. Herman JS, Sagar and Grün D. FateID infers cell fate bias in multipotent progenitors from single-cell RNA-seq data. Nature methods. 2018;15 5:379-86. doi:10.1038/nmeth.4662.
36. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC bioinformatics. 2009;10:421. doi:10.1186/1471-2105-10-421.
37. Rallis D, Tsoumani KT, Krsticevic F, Papathanos PA, Gouvi G, Meccariello A, et al. Detection of sex chromosomes in Tephritid pests using R-CQ and KAMY, two computational methods to enable species specific control. bioRxiv. 2025:2023.10.27.564325. doi:10.1101/2023.10.27.564325.
38. Li H and Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England). 2009;25 14:1754-60. doi:10.1093/bioinformatics/btp324.
39. Quinlan AR and Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics (Oxford, England). 2010;26 6:841-2. doi:10.1093/bioinformatics/btq033.

40. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics (Oxford, England). 2018;34 18:3094-100. doi:10.1093/bioinformatics/bty191.
41. Zacharopoulou A. Polytene chromosome maps in the medfly Ceratitis capitata. Genome. 1990;33 2:184-97.
42. Saccone G, Salvemini M, Pane A and Polito LC. Masculinization of XX Drosophila transgenic flies expressing the Ceratitis capitata DoublesexM isoform. Int J Dev Biol. 2008;52 8:1051-7. doi:10.1387/ijdb.082657gs.
43. Bayega A, Oikonomopoulos S, Gregoriou ME, Tsoumani KT, Giakountis A, Wang YC, et al. Nanopore long-read RNA-seq and absolute quantification delineate transcription dynamics in early embryo development of an insect pest. Scientific reports. 2021;11 1:7878. doi:10.1038/s41598-021-86753-7.
44. Ibarra-Morales D, Rauer M, Quarato P, Rabbani L, Zenk F, Schulte-Sasse M, et al. Histone variant H2A.Z regulates zygotic genome activation. Nature communications. 2021;12 1:7002. doi:10.1038/s41467-021-27125-7.
45. Kwasnieski JC, Orr-Weaver TL and Bartel DP. Early genome activation in Drosophila is extensive with an initial tendency for aborted transcripts and retained introns. Genome research. 2019;29 7:1188-97. doi:10.1101/gr.242164.118.
46. De Renzis S, Elemento O, Tavazoie S and Wieschaus EF. Unmasking activation of the zygotic genome using chromosomal deletions in the Drosophila embryo. PLoS Biol. 2007;5 5:e117. doi:10.1371/journal.pbio.0050117.
47. Blythe SA and Wieschaus EF. Zygotic genome activation triggers the DNA replication checkpoint at the midblastula transition. Cell. 2015;160 6:1169-81. doi:10.1016/j.cell.2015.01.050.
48. McDowell IC, Manandhar D, Vockley CM, Schmid AK, Reddy TE and Engelhardt BE. Clustering gene expression time series data using an infinite Gaussian process mixture model. PLoS computational biology. 2018;14 1:e1005896. doi:10.1371/journal.pcbi.1005896.
49. Davidson EH. II - The Nature and Function of Maternal Transcripts. Gene Activity in Early Development (Third Edition). San Diego: Academic Press; 1986. p. 45-123.
50. Harvey SA, Sealy I, Kettleborough R, Fenyes F, White R, Stemple D, et al. Identification of the zebrafish maternal and paternal transcriptomes. Development. 2013;140 13:2703-10. doi:10.1242/dev.095091.
51. Benoit P, Papin C, Kwak JE, Wickens M and Simonelig M. PAP- and GLD-2-type poly(A) polymerases are required sequentially in cytoplasmic polyadenylation and oogenesis in Drosophila. Development. 2008;135 11:1969-79. doi:10.1242/dev.021444.
52. Cui J, Sackton KL, Horner VL, Kumar KE and Wolfner MF. Wispy, the Drosophila homolog of GLD-2, is required during oogenesis and egg activation. Genetics. 2008;178 4:2017-29. doi:10.1534/genetics.107.084558.
53. Cui J, Sartain CV, Pleiss JA and Wolfner MF. Cytoplasmic polyadenylation is a major mRNA regulator during oogenesis and egg activation in Drosophila. Dev Biol. 2013;383 1:121-31.

doi:10.1016/j.ydbio.2013.08.013.

54. Semotok JL, Cooperstock RL, Pinder BD, Vari HK, Lipshitz HD and Smibert CA. Smaug recruits the CCR4/POP2/NOT deadenylase complex to trigger maternal transcript localization in the early Drosophila embryo. Current biology : CB. 2005;15 4:284-94. doi:10.1016/j.cub.2005.01.048.
55. Vastenhouw NL, Cao WX and Lipshitz HD. The maternal-to-zygotic transition revisited. Development. 2019;146 11 doi:10.1242/dev.161471.
56. Newton FG, Harris RE, Sutcliffe C and Ashe HL. Coordinate post-transcriptional repression of Dpp-dependent transcription factors attenuates signal range during development. Development. 2015;142 19:3362. doi:10.1242/dev.123273.
57. Weidmann CA, Raynard NA, Blewett NH, Van Etten J and Goldstrohm AC. The RNA binding domain of Pumilio antagonizes poly-adenosine binding protein and accelerates deadenylation. RNA (New York, NY). 2014;20 8:1298-319. doi:10.1261/rna.046029.114.
58. Foe VE and Alberts BM. Studies of nuclear and cytoplasmic behaviour during the five mitotic cycles that precede gastrulation in Drosophila embryogenesis. J Cell Sci. 1983;61:31-70. doi:10.1242/jcs.61.1.31.
59. ten Bosch JR, Benavides JA and Cline TW. The TAGteam DNA motif controls the timing of Drosophila pre-blastoderm transcription. Development. 2006;133 10:1967-77. doi:10.1242/dev.02373.
60. Liang HL, Nien CY, Liu HY, Metzstein MM, Kirov N and Rushlow C. The zinc-finger protein Zelda is a key activator of the early zygotic genome in Drosophila. Nature. 2008;456 7220:400-3. doi:10.1038/nature07388.

# Figures

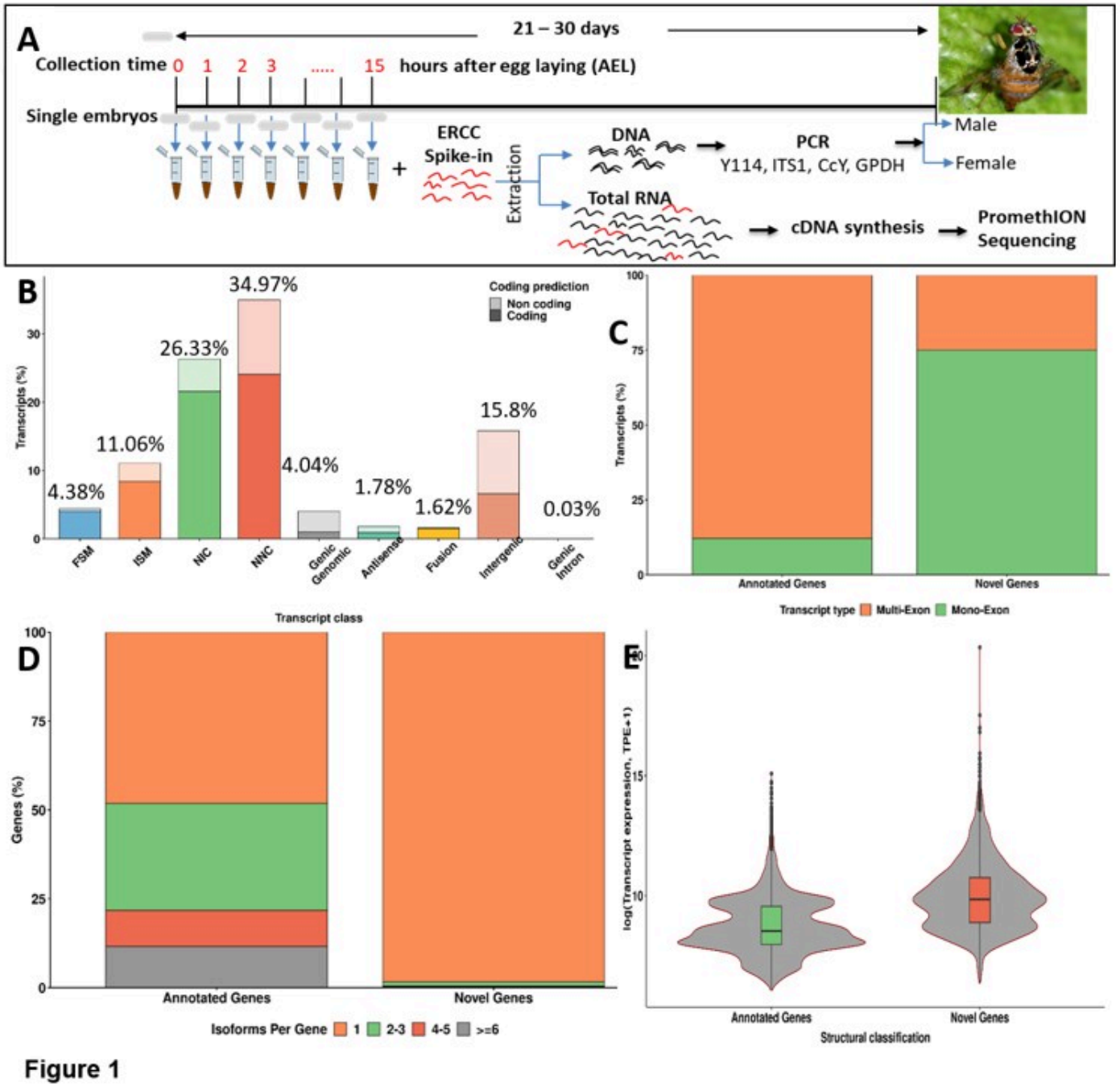



**Figure 1**

**Figure 1**

**Summary of the workflow and transcriptome assembly**. A) Schematic of the library preparation workflow. Single embryos were collected at hourly intervals for the first 15 hours after egg laying (AEL). From each time point, 10 embryos were randomly selected and individually crushed using a pestle and mortar. Equal amounts of ERCC (RNA external standards) were added to each sample followed by DNA and total RNA extraction from each embryo. The DNA was used to sex the embryos via PCR while the RNA was used for cDNA library preparation and long-read RNA-seq. B) Comparison of the long-read derived transcriptome assembly to the NCBI predicted transcriptome. We used SQANTI **[35]** to align, match, and compare gene models in our long-read assembly to the NCBI transcriptome. The results are

categorized as. FSM; full splice match, ISM; incomplete splice match, NIC; novel in catalogue, NNC; novel not in catalogue (see **Supplementary Table S1** for explanation). C) Same as B but comparing exon number between annotated and novel genes. D) Same as B but comparing number of isoforms between annotated and novel genes. E) Comparison of gene expression (log transcripts per embryo, TPE) between annotated and novel genes.

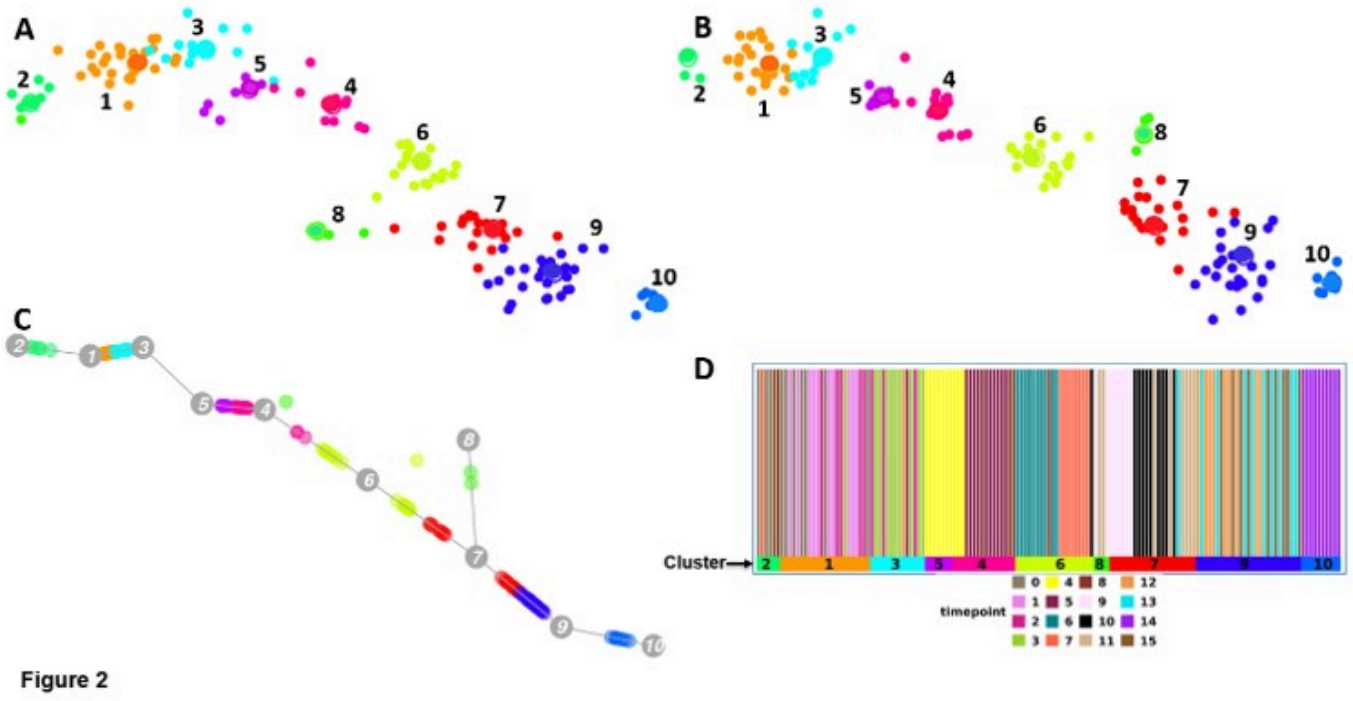



Figure 2

## Figure 2

**Unsupervised clustering of embryos**. A) Unsupervised clustering of embryos using FateID [35]. Clustering was based on normalized absolute number of transcripts per embryo. B) Embryos in 'A' were filtered to remove those that had aberrant expression of genes with a previously validated expression profile such as *sisA* and *slam slam* [17], as well as other Drosophila markers [44]. In total 11 embryos were filtered out and the clusters regenerated as in 'A'. C) Lineage tracing was performed on the clusters in 'B' using StemID [35]. The developmental trajectory shows a linear profile from the most immature embryos in cluster 2 to the most mature embryos in cluster 10. D) A schematic comparing pseudo time staging of embryos with staging based on hours after egg-laying (AEL). The embryos are grouped according to their pseudo time clustering with the bottom horizontal bars demarcating the clusters while the vertical bars, which represent each embryo, are colored based on the real-time of each embryo AEL.

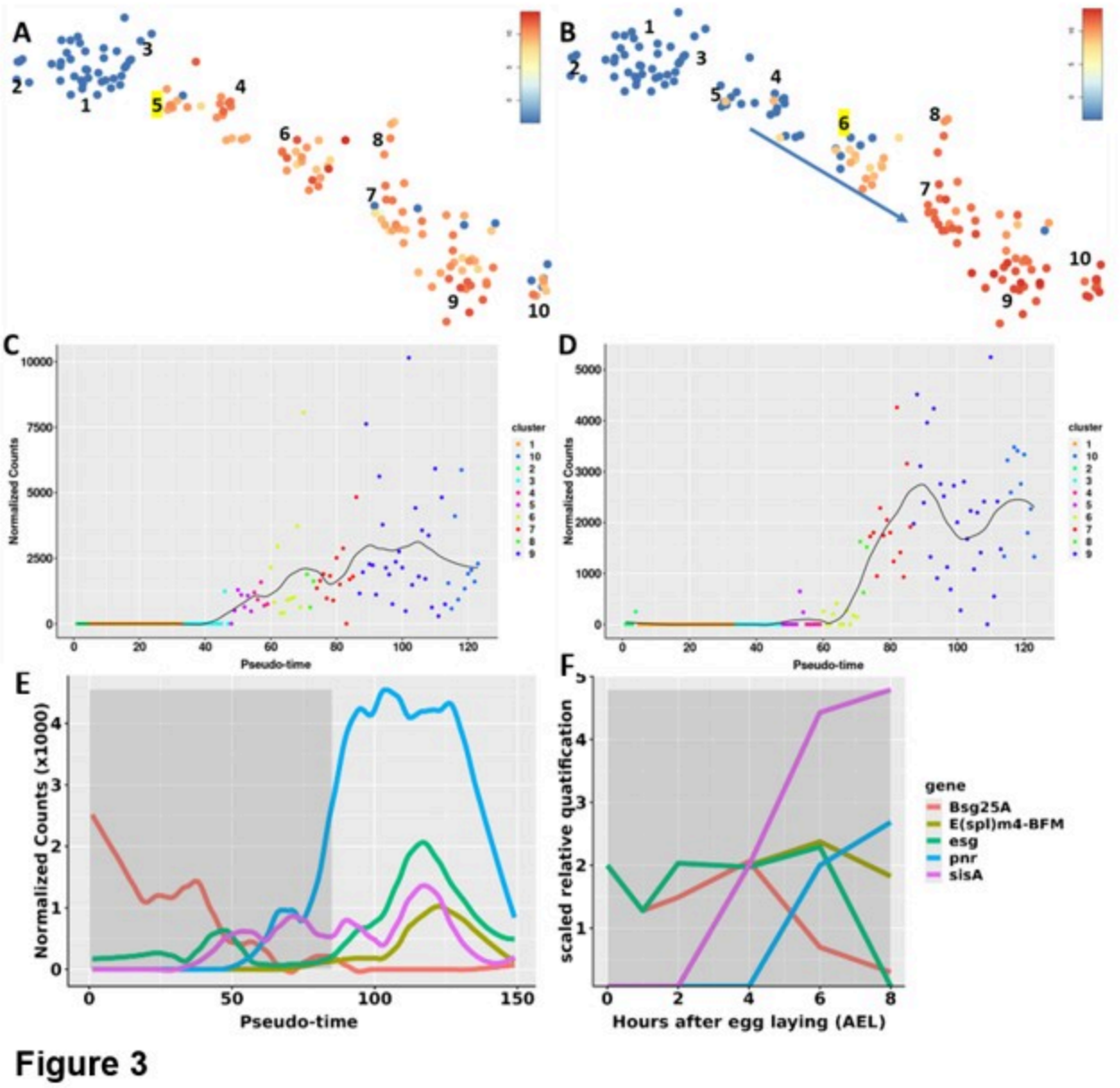



Figure 3

**Figure 3**

**Zygotic genome activation (ZGA)**. A) We probed the expression of genes previously confirmed to mark the minor wave of ZGA in *Drosophila melanogaster* including *sisA*, *amos*, *CG13712*, and *CG6885* [47]. Cluster 5 showed the first expression of these genes potentially making cluster 5 as a marker of the first wave of ZGA in our medfly embryos. B) Same as 'A' except we probed for *D. melanogaster* major wave of ZGA genes including *grh*, *l(1)sc*, *sc*, and *NetA* [45]. Cluster 6 showed the first main expression of these genes potentially marking cluster 6 as a marker of the second wave of ZGA in our medfly embryos. C) Aggregated expression profile of 4 genes used to mark first wave of medfly ZGA (*sisA*, *amos*, *CG13712*, and *CG6885*). D) Aggregated expression profile of 4 genes used to mark the second wave of medfly ZGA (*grh*, *l(1)sc*, *sc*, and *NetA*). E) Expression profile of select genes on a pseudo time scale. F) qPCR

validation of the expression of genes in ‘E’. We collected new embryos from 0 to 8 hours after egg-laying and performed qPCR targeting 5 genes (*Bsg25A*, *E(spl)m4-BFM*, *esg*, *pnr*, *sisA*). The grey shaded region shows a similar sampling time between pseudo time and real-time staging.

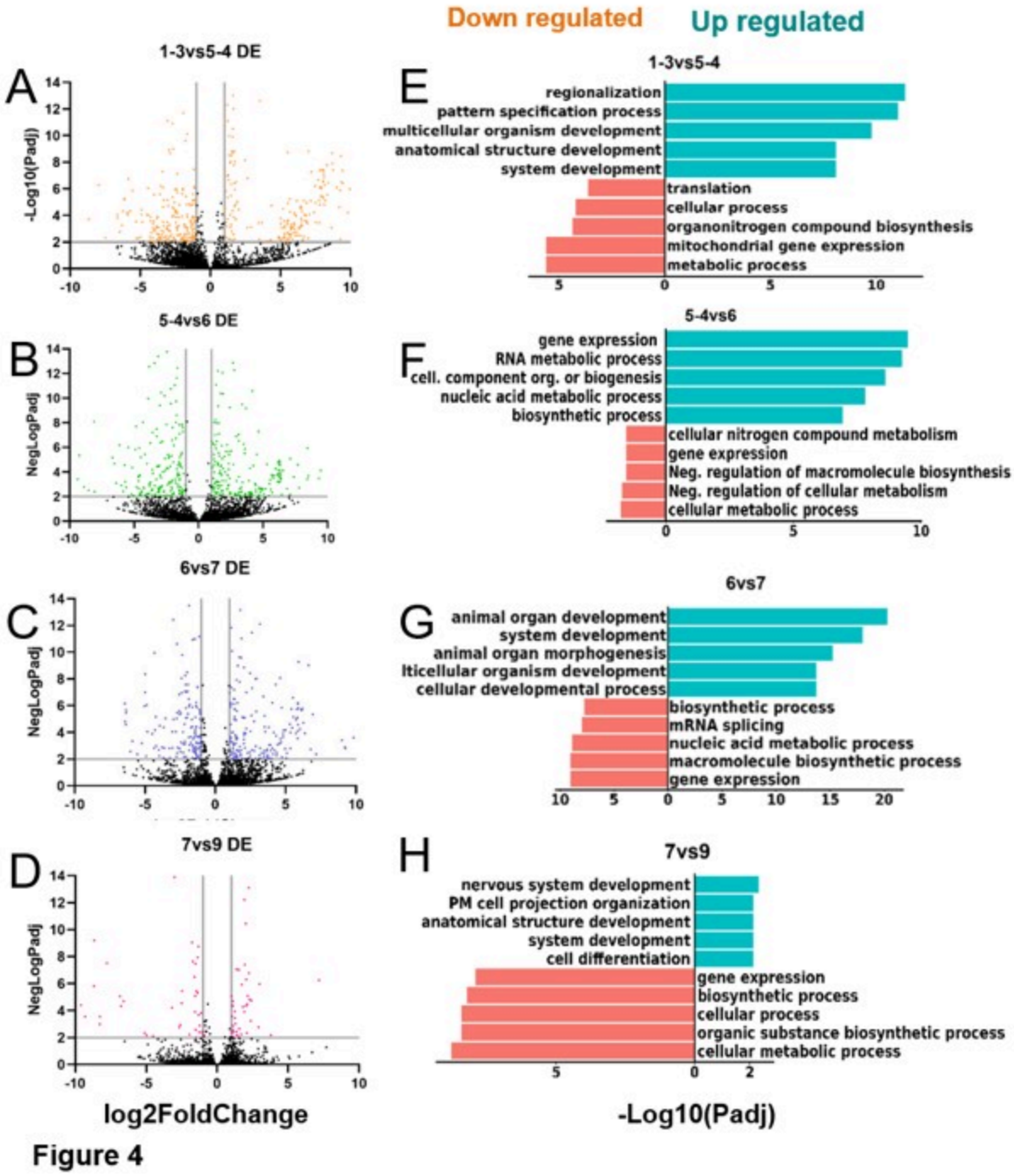
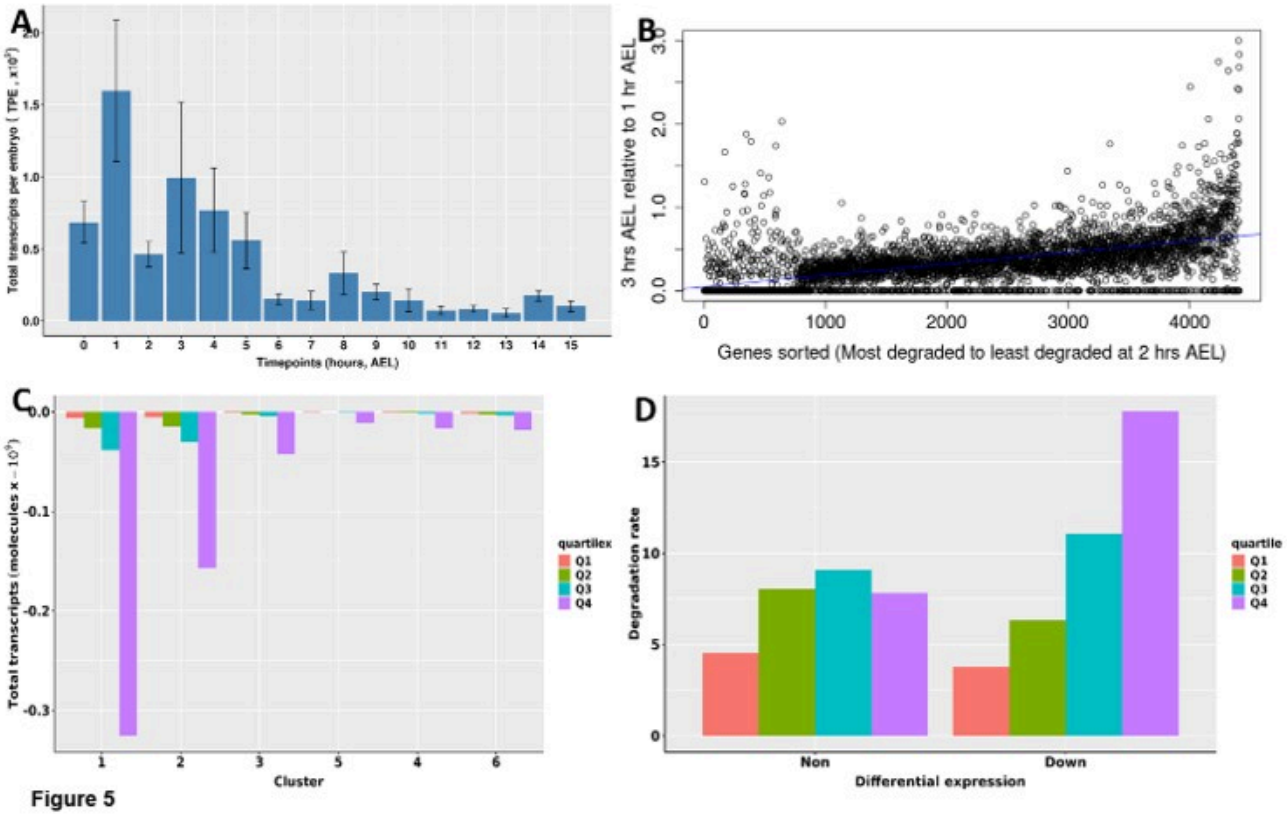



**Figure 4**

**Enriched biological processes in the different clusters**. A-D) Differential gene expression was performed in successive clusters. We combined all embryos in clusters 1 and 3. We also combined clusters 5 and

4. We then determined significantly differentially expressed genes in clusters 1-3 versus 5-4 (A), 5-4 versus 6 (B), 6 versus 7 (C), and 7 versus 9 (D). Key genes are shown in each figure. E-H) We used the significantly differentially expressed genes to perform gene set enrichment analysis and sorted the biological process by negative log10 of the adjusted p-value. We picked the top 5 biological processes in each of the clusters. Blue bars to the right indicate significantly upregulated processes while the orange bars to the left are significantly downregulated processes.




## Figure 5

**Reorganization of maternal transcripts during maternal-to-zygotic transition (MZT)**. A) Profile of total transcripts per embryo across time points. We performed absolute normalization to obtain the total number of transcripts per gene per embryo and then summed this. The bars indicate the mean for all embryos at that time point while error bars indicate standard deviation of the mean. B) The ratio of gene expression of embryos at 3 hours after egg laying (AEL) and 1 hour AEL were determined. Genes were sorted based on their gene expression levels at 2 hours AEL, from least abundant (most degraded) to most abundant (least degraded). C) We obtained differentially downregulated genes in cluster 1 relative to cluster 6. These genes were sorted according to their expression levels in cluster 1 and put in 4 quartiles: quartile 1 (Q1) contained the least abundant genes while Q4 contained the most abundant genes. The expression levels of all genes within each quartile were summed across clusters 1 to 6. D)

We obtained both differentially downregulated (Down) and non-differentially expressed genes (Non) in cluster 1 relative to cluster 6. These genes were sorted according to their expression levels at in cluster 1 and put in 4 quartiles as explained in ‘C’. We then determine the ratio of expression in cluster 1 relative to cluster 6 to determine the ‘degradation rate’.

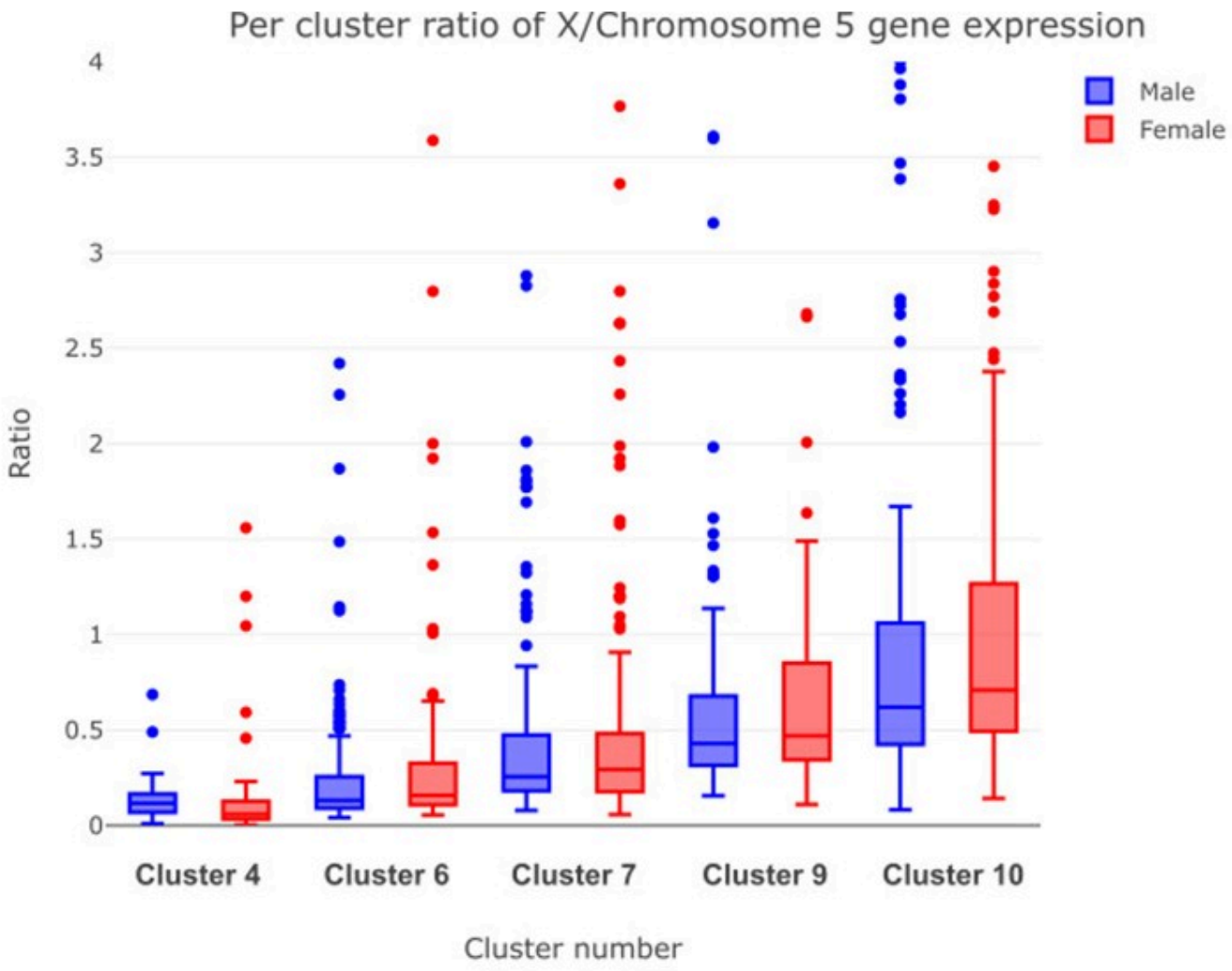



**Figure 6**

**Figure 6**

**Assessment of dosage compensation in medfly embryos.** First, we identified the Medfly early zygotic genes on X chromosome and chromosome 5. We calculated the male and female average zygotic gene expression from chromosome 5 genes across the different clusters. Then, we derived the X:chromosome 5 ratio for each X chromosome gene individually by calculating the average male and female expression of that X gene across different clusters, and dividing it with the respective male or female chromosome 5 average expression.

# Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementarymaterialsV5Flat.docx
- SupplemetarymaterialTablesv3.zip