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.

Large-scale reconstructions of Drosophila transcriptome identify ten thousands of new transcripts and transcription readthrough events

Haonan Duanmu
Meizhen Li
Zihan Zhou
Xinyan Li
Hao Chen
Kang He
Fei Li
lifei18@zju.edu.cn

Research Article

Keywords: Drosophila, transcriptome, network, super-housekeeping genes, transcription readthrough

Posted Date: February 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8670304/v1

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

Additional Declarations: No competing interests reported.

Abstract

Comprehensive transcriptome annotation is critical for understanding gene regulation, developmental plasticity, and evolutionary constraint. Here, we reconstruct a high-resolution transcriptome atlas of *Drosophila melanogaster* by integrating 33,186 high-quality RNA-seq libraries encompassing multiple developmental stages, tissues, and environmental conditions. The final assembly contains 398,168 transcripts, and includes 136,828 protein-coding mRNAs and 125,131 long non-coding RNAs (lncRNAs), identifying 4,515 previously unannotated protein-coding transcripts and 28,528 novel lncRNAs. Notably, we identify 49 transcripts which are highly expressed in more than 90% samples, named as super-housekeeping transcripts. Surprisingly, among them, 22 were annotated as uncharacterized transcripts. We then carried out structure-based annotation by comparing their structures with protein PDB100 database, showing that most of these super-housekeeping transcripts share highly conserved structure with proteins associate with ribosome production, mitochondrial respiration, and nucleic acid processing and 40 of them serve as hub genes across multiple co-expression network modules. In addition, we also identify 890,750 events of transcription readthrough in *Drosophila*, which are specially enriched in pupal metamorphosis and tended to be induced by irradiation, starvation, and high-fat diet. Together, these findings indicate that present understanding of *Drosophila* transcripts is still a tip of an iceberg. The functional complexity of the *Drosophila* transcription is worthy of further attentions.

Introduction

Next-generation RNA sequencing (RNA-seq) has revolutionized transcriptomics by uncovering pervasive transcription, extensive isoform diversity, and numerous functional elements beyond conventional gene models (Djebali et al., 2012; ENCODE Project Consortium, 2012). In humans, landmark efforts such as the GENCODE catalog of long noncoding RNAs (Derrien et al., 2012), the CHESS database derived from thousands of RNA-seq datasets (Pertea et al., 2018), and the multi-tissue Tabula Sapiens single-cell atlas (Tabula Sapiens Consortium, 2022) have established benchmarks for comprehensive transcriptome annotation and functional discovery. Collectively, these resources refined gene models and revealed previously unappreciated regulatory principles underlying development, homeostasis, and disease.

Drosophila melanogaster, a cornerstone model organism for genetics, development, and evolution, has similarly been at the forefront of genome annotation since its initial sequencing, which identified approximately 13,600 protein-coding genes (Adams et al., 2000). Subsequent community-driven curation efforts and systematic computational analyses rapidly refined this annotation, with successive genome releases incrementally improving gene models and incorporating diverse classes of noncoding RNAs, including microRNAs and small nucleolar RNAs (Misra et al., 2002; Smith et al., 2007). A transformative leap was achieved through the modENCODE project, which integrated more than 700 datasets encompassing RNA-seq, ChIP-seq, and DNase-seq across multiple developmental stages and cell types. This large-scale effort tripled the fraction of the genome associated with biochemical function, identified 1,938 previously unannotated genes, and mapped 52,914 novel or revised exons, ultimately establishing a comprehensive view of transcriptional complexity and regulatory networks that spans approximately 80% of the genome (Gerstein et al., 2010; Roy et al., 2010; Brown et al., 2014).

In the years following the modENCODE project, several dedicated resources have further advanced our understanding of context-dependent transcriptional programs in Drosophila melanogaster. FlyAtlas provided the first systematic tissue-level expression profiles across 14 adult tissues using microarray-based measurements (Chintapalli et al., 2007). This resource was later extended as FlyAtlas 2, which incorporated RNA-seq data to improve sensitivity and enable sex-specific analyses (Leader et al., 2018). In parallel, developmental time-course studies characterized dynamic changes in gene expression and alternative splicing from embryonic to adult stages (Graveley et al., 2011; Brown et al., 2014). Population-scale analyses further quantified inter-individual transcriptional variation using the *Drosophila* Genetic Reference Panel (Lin et al., 2016). More recently, Qian et al. (2023) compiled a transcriptome atlas from 12,049 high-quality public RNA-seq datasets, enabling stratification of expression patterns by anatomical context, developmental stage, and sex.

Despite these advances, most existing resources remain largely anchored to previously annotated gene models (Qian et al., 2023; Burghardt et al., 2023; Sakaguchi et al., 2023; Meyer et al., 2024). Although recent studies have expanded the scale of

analysis, the full potential of the rapidly growing public RNA-seq archives has not yet been integrated into a unified and quantitative framework.

To address this limitation, we generated a comprehensive, high-resolution transcriptome resource for *D. melanogaster* by integrating 33,186 publicly available RNA-seq datasets. This collection spans embryonic, larval, pupal, and adult stages, multiple tissues, both sexes, and a broad range of environmental perturbations. Together, this resource provides a unified view of the *Drosophila* transcriptome and supports systematic analyses beyond fragmented gene centric annotations..

Results

Comprehensive reconstruction and characterization of the D. melanogaster transcriptome

We assembled a comprehensive transcriptomic dataset comprising 45,233 publicly available *D.melanogaster* RNA-seq libraries retrieved from the NCBI Sequence Read Archive (SRA; accessed June 30, 2022). To ensure analytical robustness, libraries with incomplete metadata, insufficient sequencing depth, or low mapping quality were excluded. From this curated collection, we reconstructed the *D. melanogaster* transcriptome using 33,186 high-quality RNA-seq libraries that met a Hisat2 alignment rate threshold of > 50%. This final dataset includes samples spanning seven tissues, four developmental stages, and six experimental conditions, providing broad coverage of transcriptomic diversity (Tables 1–2). RNA-seq reads were aligned to the *D.melanogaster* reference genome (FlyBase r6.45), resulting in the identification of 398,168 transcripts and covering more than 99% of the genome.

We next assessed the coding potential of all 398,168 transcripts. Among these, 136,828 transcripts were classified as protein-coding mRNAs, 125,131 as long non-coding RNAs (lncRNAs), and the remaining 136,109 as unclassified or other RNA types (Table 3). In comparison, FlyBase r6.45 annotates 30,799 protein-coding transcripts and 3,051 non-coding RNAs. Our assembly therefore substantially expands the current annotation, including 4,515 previously unannotated protein-coding transcripts and 28,528 novel lncRNAs that do not overlap with any known FlyBase gene models (Fig. S1).

Together, the identification of this large number of previously unannotated coding and non-coding transcripts markedly extends the scope and resolution of the *Drosophila* genome annotation. This resource provides a comprehensive and high-quality transcriptomic framework that supports systematic analyses of gene expression and transcript diversity in *D.melanogaster*.

Table 1
RNA-seq sample counts across developmental stages and tissue in *D.melanogaster*

Stage	Tissue	Number
Embryo	Whole Body	1030
Larva	Neuron	487
	Wing	332
	Reproductive System	14
	Fat Body	71
	Digestive	121
	Whole Body	431
Pupa	Head + Thorax	12
	Neuron	176
	Reproductive System	39
	Digestive	10
	Whole Body	102
Adult	Head + Thorax	1992
	Neuron	925
	Wing	8
	Reproductive System	838
	Fat Body	66
	Digestive	203
	Whole Body	3072

Table 2
Summary of RNA-seq samples under different experimental treatments

Treatment		Number
Temperature	High Temperature (> 25℃)	641
	Low Temperature (< 25℃)	128
Sucrose	High Sucrose (> 0.15M)	47
Sugar	High Sugar	50
Fat	High Fat	104
Starvation	Starvation	110
Radiation	Radiation	58
Toxic Treatment	Toxic Treatment	50

Table 3
Summary of transcriptome assembly and annotation statistics for *D.melanogaster*.

Category		Count	% of total	New
Input data	RNA-seq datasets	33,186	—	—
Transcripts	Assembled transcripts	398,168	—	—
	Protein-coding transcripts (mRNAs)	156,290	34.4%	9,704
	Long non-coding RNA transcripts (lncRNAs)	125,131	31.4%	108,564
Annotation recovery vs FlyBase	lncRNA genes recovered	2,430 / 3,052	79.6%	—
	Protein-coding genes recovered	13,229 / 13,986	94.6%	—

Super-Housekeeping transcripts with potential essential function

To evaluate the expression profiles of the newly identified transcripts in *D.melanogaster*, we quantified expression levels for 136,828 protein-coding transcripts and 125,131 lncRNAs using fragments per kilobase of transcript per million mapped reads (FPKM). The majority of these transcripts showed low expression, with approximately 90% displaying FPKM values below 1 in more than 90% of samples (Fig. 1A). Notably, over 95% of lncRNAs were detected in only a single tissue or developmental stage, indicating highly tissue- or stage-specific expression patterns (Fig. 1B). In contrast, more than 65% of protein-coding mRNAs were expressed across two or more tissues or developmental stages (Fig. 1C). Consistent with these observations, heatmap analyses across tissues and developmental stages revealed pronounced heterogeneity among non-coding transcripts, whereas protein-coding transcripts exhibited more uniform expression dominated by a core set of broadly expressed genes (Fig. 1B–C).

Among all transcripts, we identified 49 that were highly expressed in more than 90% of samples and defined these as super-housekeeping transcripts (Fig. 1D). These transcripts were distributed across four chromosomes and six scaffolds, with notable enrichment in ribosomal DNA (rDNA) regions and the mitochondrial genome (Fig. 2A). Functional annotation based on BLAST searches against the Swiss-Prot database indicated that this group includes ten ribosomal RNAs, four mitochondrial respiratory chain subunits, and nine retrotransposon-derived transcripts. In addition, the set contains one non-coding RNA, one transcript encoding a zinc finger nucleic acid–binding protein, and two pseudogenes. Among these, three transcripts (FBtr0346873, FBtr0346885, and FBtr0100861) displayed particularly stable expression levels, as reflected by low coefficients of variation (CV < 2.0), suggesting their potential utility as reference genes for qRT–PCR normalization. In contrast, several transcripts, including MSTRG.1134.1, MSTRG.1129.1, and FBtr0307366, showed substantially higher variability (CV > 4.0) (Table 4).

Notably, 22 of the super-housekeeping transcripts were annotated as encoding unknown proteins using conventional sequence-based annotation pipelines, owing to limited sequence similarity to characterized proteins. We therefore explored whether structural information could provide additional functional insight. Protein structures for these 22 transcripts were predicted using AlphaFold3 and subsequently compared against the PDB100 structural database. Structure-based alignments revealed that six previously uncharacterized transcripts exhibit high structural similarity to functionally annotated proteins, including components associated with translation (such as ribosomal and RNA-binding domains), energy metabolism (mitochondrial-related structures), and nucleic acid processing enzymes with endonuclease- or reverse transcriptase like folds (Fig. 2B–C; Figs. S2–S3; Tables 5–6).

Table 4
Longest ORFs per transcript with GC content and CAI

Transcript ID	Transcript len	ORF len	Pep length	GC content	CAI
MSTRG.844.1	1060	435	145	48.28	0.843
MSTRG.842.1	1963	264	87	30.68	0.915
MSTRG.841.1	1963	264	87	30.68	0.915
MSTRG.839.1	1813	276	91	46.38	0.851
MSTRG.838.1	1080	171	57	50.88	0.839
MSTRG.837.1	1125	432	143	49.77	0.822
MSTRG.694.1	1134	270	89	38.52	0.884
MSTRG.670.1	2904	267	88	29.96	0.935
MSTRG.3270.14	10660	324	107	43.83	0.842
MSTRG.3256.9	11309	126	41	11.11	0.907
MSTRG.246.1	1047	186	61	48.92	0.827
MSTRG.181.1	2511	270	89	38.52	0.884
MSTRG.180.1	2511	270	89	38.52	0.884
MSTRG.1437.1	1078	264	87	31.06	0.915
MSTRG.1416.1	2168	291	96	39.86	0.863
MSTRG.1328.1	1134	141	46	43.26	0.796
MSTRG.1327.1	1134	141	46	43.26	0.796
MSTRG.1208.1	1136	288	95	30.56	0.914
MSTRG.1134.1	1012	159	52	22.01	0.896
MSTRG.1129.1	1002	174	57	40.23	0.833
MSTRG.1128.1	1002	174	57	40.23	0.833
MSTRG.1102.1	3076	1014	337	54.54	0.831

Table 5
Best structural matches for candidate ORFs using afdb-swissprot

ORF ID	Scaffold	Top Target	Prob.	Description	Scientific Name	Category
MSTRG.670.1	211000022278724	AF-A5IJJ8-F1-model_v4	0.47	DNA polymerase III PolC-type	*Thermotoga petrophila*	Cell cycle
MSTRG.838.1	211000022278878	AF-P16423-F1-model_v4	0.92	Retrovirus-related Pol polyprotein	*Drosophila melanogaster*	Retrotransposon
MSTRG.844.1	211000022278881	AF-P16423-F1-model_v4	1	Retrovirus-related Pol polyprotein	*Drosophila melanogaster*	Retrotransposon
MSTRG.1102.1	211000022279109	AF-P16425-F1-model_v4	1	Putative 115 kDa retrotransposon protein	*Drosophila melanogaster*	Retrotransposon
MSTRG.1208.1	211000022279222	AF-A7FD46-F1-model_v4	1	ppGpp pyrophosphatase	*Yersinia pseudotuberculosis*	Signal regulation
MSTRG.1327.1	211000022279343	AF-Q5BIS9-F1-model_v4	0.77	AMP-activated protein kinase subunit beta-1	*Bos taurus*	Signal regulation
MSTRG.1328.1	211000022279343	AF-Q5BIS9-F1-model_v4	0.77	AMP-activated protein kinase subunit beta-1	*Bos taurus*	Signal regulation

Table 6
Best structural matches for candidate ORFs using PDB100

ORF ID	Scaffold	Top Target	Prob.	Description	Scientific Name	Category
MSTRG.838.1	211000022278878	2rv0-assembly1	0.95	Solution structure of *Homo sapiens* zinc finger protein	*Homo sapiens*	Zinc finger
MSTRG.844.1	211000022278881	8gh6-assembly1	1	*Bombyx mori* R2 retrotransposon endonuclease domain	*Bombyx mori*	Retrotransposon
MSTRG.1102.1	211000022279109	1wdu-assembly1	1	Endonuclease domain from *Bombyx mori* R2 retrotransposon	*Bombyx mori*	Retrotransposon
MSTRG.1208.1	211000022279222	5xiy-assembly1	0.57	Small subunit ribosomal protein	*Leishmania major*	Ribosome
MSTRG.1327.1	211000022279343	4eaj-assembly1	0.69	Co-crystal structure of human AMPK with AMP	*Homo sapiens*	AMPK/Metabolism
MSTRG.1328.1	211000022279343	4eaj-assembly1	0.69	Co-crystal structure of human AMPK with AMP	*Homo sapiens*	AMPK/Metabolism

Co-expression network modules across tissues and developmental stages

To characterize spatiotemporal regulatory networks in *D.melanogaster*, we constructed weighted gene co-expression networks (WGCNA) separately for each developmental stage and each tissue. Each network was partitioned into 25 to 337 co-expression modules, collectively capturing the hierarchical organization of transcriptional regulation across diverse biological contexts (Table S1). Functional enrichment analyses indicated that these modules were predominantly associated with ribosome biogenesis, RNA processing, oxidative phosphorylation, chromatin remodeling, and signal transduction, reflecting a balance between core cellular maintenance and context-specific regulatory programs.

Across networks, we consistently identified three major module classes enriched for ribosomal protein genes, mitochondrial respiratory genes, and RNA-binding factors. These modules were recurrently detected across multiple developmental stages, including embryos, larvae, pupae, and adults, as well as in several tissues such as neurons, ovaries, and testes, suggesting the presence of conserved transcriptional programs supporting essential cellular functions.

We defined hub transcripts as those ranking within the top 5% of intramodular connectivity (Kwithin) in each module, yielding a total of 22,387 hub transcripts. Among these, 40 of the previously defined super-housekeeping transcripts functioned as high-connectivity hubs across multiple networks (Table S2), with their recurrence across tissues or developmental stages ranging from 1 to 15. Notably, these hub transcripts and their associated modules showed strong functional concordance with the biological roles of their respective tissues and developmental stages. For example, FBtr0307366 served as a hub transcript in the adult head–enriched “turquoise” module, which was significantly enriched for Gene Ontology terms related to translation initiation, ribosome biogenesis, and proteostasis. Similarly, FBtr0346878 anchored the adult fat body–enriched “turquoise” module associated with rRNA processing, telomere maintenance, and cytokine signaling, whereas FBtr0346900 emerged as a hub transcript in the larval wing–enriched “black” module linked to Toll-like receptor signaling, Wnt pathway regulation, and mitochondrial protein targeting (Fig. S3).

To quantitatively assess the preservation of these co-expression modules across developmental stages, we performed NetRep-based permutation testing. Core modules associated with metabolic processes and RNA processing exhibited high preservation across all developmental stages ($p < 0.01$), indicating that transcriptional programs underlying translation and energy metabolism remain stable throughout ontogeny (Fig. 3A–B). In contrast, modules related to transcriptional regulation, chromatin organization, and developmental signaling showed low or nonsignificant preservation during stage transitions, consistent with extensive network rewiring accompanying developmental progression (Fig. 3C).

Together, these results support a hierarchically organized transcriptome network architecture in *Drosophila*, comprising a conserved core of metabolic and housekeeping-related modules alongside tissue- and stage-specific regulatory modules that undergo dynamic reconfiguration. This organization enables the maintenance of essential cellular functions while permitting flexible transcriptional responses required for precise developmental transitions, thereby balancing transcriptional stability and regulatory plasticity across development..

Extensive transcription readthrough in Drosophila

Unexpectedly, we observed that transcripts from some adjacent genes were merged into a single transcript. Further analysis indicated that this phenomenon resulted from transcriptional readthrough, in which transcription extends beyond canonical termination sites. This observation prompted us to systematically investigate the prevalence of transcriptional readthrough in *D.melanogaster*. We found that transcriptional readthrough was detectable across nearly all developmental stages and physiological contexts. In total, 890,750 readthrough events were identified, with an average length of 14.8 kb, indicating that transcriptional termination escape is a widespread yet quantitatively variable feature of the *Drosophila* transcriptome.

The frequency of transcriptional readthrough differed markedly across developmental stages. Pupal samples exhibited the highest incidence of readthrough events (median = 28, mean = 45.43, $p < 0.001$), followed by larval samples (median = 15, mean = 31.84), whereas adult samples showed substantially lower levels (median = 7, mean = 20.96) (Fig. 4A). Across all

stages, 9,366 genes showed evidence of transcriptional readthrough; however, only 1,035 genes (11.4%) were shared by at least two stages. This limited overlap indicates that transcriptional readthrough is developmentally regulated rather than constitutive.

Tissue-specific analyses further showed that fewer than 10% of expressed genes in each tissue exhibited transcriptional readthrough, with the strongest enrichment observed in adult neurons (1,783 genes) and embryos (1,582 genes) (Fig. 4B). Readthrough loci were non-randomly distributed across the genome, with chromosomes 2L and 3R together accounting for nearly half (49%) of all events. In contrast, the Y chromosome, ribosomal DNA (rDNA) clusters, and mitochondrial loci were largely devoid of transcriptional readthrough (Fig. 4C).

Environmental perturbations further modulated transcriptional readthrough. Among the conditions examined, irradiation induced the strongest increase in readthrough frequency (2.8-fold, $p < 10^{-5}$), followed by high-fat diet (1.9-fold, $p < 0.01$) and starvation (1.7-fold, $p < 0.05$). In contrast, high-sucrose feeding reduced transcriptional readthrough by approximately 45%, while heat and cold stress had only modest or negligible effects (≤ 1.2-fold) (Fig. 4D). Pairwise correlation analyses revealed limited overlap among readthrough-positive gene sets across conditions (mean Jaccard index = 0.18), indicating that transcriptional readthrough is highly context-specific and shaped by developmental and physiological states rather than representing a uniform background process.

Integrated analyses of both short- and long-read transcriptomic datasets identified a conserved subset of seven transcriptional readthrough–prone genes that were consistently detected in both third-instar larval (L3) and pupal transcriptomes (Fig. 5A). Within this group, *FBgn0024804* emerged as the most prominent locus, being observed in more than 15,000 transcriptomic samples and exhibiting transcriptional extension exceeding 10 kb beyond the annotated termination site (Fig. 5B–C). Other recurrent loci, including *FBgn0026096*, *FBgn0029176*, and *FBgn0031098*, were each detected in over 1,000 samples. The extended transcription at these loci was reproducibly observed across biological replicates and experimental conditions, indicating that these readthrough events represent stable and regulated transcriptional features rather than sporadic transcriptional noise.

Together, these results provide evidence that transcriptional readthrough beyond canonical gene boundaries is a widespread yet incompletely characterized phenomenon in *D.melanogaster*. This process is developmentally dynamic, tissue- and environment-specific, and involves a small but conserved set of high-frequency readthrough genes, such as *FBgn0024804*, which may reflect fundamental mechanisms of stress-responsive regulation and post-termination transcriptional control.

Discussion

Here, we reconstructed the *D.melanogaster* transcriptome by integrating more than 33,000 high-quality RNA-seq libraries, yielding a transcriptional landscape that is substantially more complex and dynamic than previously described. Our results portray the fly transcriptome as a highly regulated and condition-responsive system, in which non-canonical features—including transcriptional readthrough, pervasive lncRNA expression, and extensive isoform diversity—represent integral components of gene regulation rather than transcriptional noise (Brown et al., 2014; Graveley et al., 2011).

A key insight from this study is that sequence conservation alone is insufficient to delineate functional genomic elements. In particular, our analyses highlight the limitations of conventional sequence-based annotation pipelines, which systematically overlook a subset of essential transcripts. This includes 22 stably expressed “super-housekeeping” transcripts that were previously classified as hypothetical or uncharacterized due to limited sequence similarity (Jungreis et al., 2021). By applying AlphaFold3-based *de novo* structure prediction and structure-based comparisons against the PDB100 database, we found that six of these transcripts adopt protein folds closely resembling those of functionally characterized proteins, including ribosomal components, RNA-binding domains, mitochondrial metabolic enzymes, and retrotransposon-associated nucleases. These structural similarities were detected despite minimal sequence-level homology, indicating that functional conservation can be maintained at the level of three-dimensional structure even under rapid sequence divergence.

Together, these findings suggest that structure-level conservation represents an underappreciated dimension of genome function that is largely inaccessible to homology-based annotation strategies (Oates et al., 2023; Varadi et al., 2022). Incorporating structural information into transcriptome annotation pipelines may therefore provide a complementary framework for identifying functionally relevant elements, particularly in cases where evolutionary distance limits sequence-based inference. Such structure-informed approaches are likely to be broadly applicable across eukaryotic genomes and may help refine functional annotation in both model and non-model organisms (Kozlov et al., 2023)..

In addition, the systematic detection of nearly 900,000 transcriptional readthrough events challenges the long-standing view of readthrough as a stochastic by-product of imperfect transcription termination. Instead, our analyses indicate that transcriptional readthrough follows defined developmental dynamics, peaking during pupal metamorphosis, and is robustly induced by environmental stressors such as irradiation and nutrient deprivation (Hertz et al., 2013). Mechanistically, stress-associated perturbations of transcription termination provide a parsimonious explanation, including compromised cleavage and polyadenylation, dissociation of the Integrator complex, or altered control of transcription elongation (Vilborg et al., 2015; Rosa-Mercado et al., 2021). The recurrent occurrence of readthrough at evolutionarily conserved, high-frequency loci (for example, *FBgn0024804*), together with its reported dependence on regulatory factors such as Nab2 (Alpert et al., 2020), highlights these sites as tractable regulatory nodes. Such loci may contribute to the generation of chimeric transcripts or modulate mRNA stability in response to physiological demands (D'Erchia et al., 2015). Collectively, these observations support a model in which transcriptional readthrough represents a regulated layer of gene expression rather than a transcriptional error, contributing to phenotypic plasticity during development and stress responses (Fuchs et al., 2022).

The hierarchical organization revealed by co-expression network analyses further reinforces this dual regulatory architecture. A conserved core of housekeeping-related modules supports fundamental cellular functions and maintains transcriptional stability (Gillis et al., 2014; Oldham et al., 2008), whereas a subset of modules undergoes dynamic rewiring in response to developmental transitions or environmental perturbations (Liu et al., 2020). Notably, many super-housekeeping transcripts occupy central hub positions within these networks, linking transcriptional stability with regulatory flexibility. Their combined features of constitutive expression and structural conservation suggest that they reside at key nodes of the cellular machinery, subject to strong purifying selection while enabling adaptive responses that support organismal robustness (Eisenberg & Levanon, 2013; Zhang & Li, 2021).

The scale and depth of the reconstructed *D. melanogaster* transcriptome represent not merely an incremental update to existing annotations, but a substantial expansion of known transcriptional complexity in a canonical model organism (Graveley et al., 2011). By capturing rare, condition-specific, and developmentally transient isoforms across a broad range of biological contexts (Li et al., 2020), this resource supports a shift from gene-centric views toward a dynamic, systems-level understanding of the transcriptome (Gingeras, 2009). Importantly, the integration of large-scale co-expression modeling (Langfelder & Horvath, 2008), systematic detection of non-canonical transcription events (Tilgner et al., 2012), and structure-based functional inference reveals layers of functional organization that remain inaccessible to sequence-based approaches alone.

In summary, this transcriptomic resource provides a foundation for future studies of gene regulation, transcriptome evolution, and systems biology, not only in *Drosophila* but also as a general framework for dissecting complex transcriptomes across eukaryotes.

Method

RNA-seq Dataset Collection and Preprocessing

We retrieved 45,233 publicly available *D.melanogaster* RNA-seq datasets from the NCBI Sequence Read Archive (SRA; accessed June 30, 2022) (Leinonen et al., 2011), spanning multiple tissues, developmental stages, and experimental conditions (Tables 1–2). Raw sequencing reads were extracted using SRA Toolkit v2.10.9 (fastq-dump --split-3) (SRA Toolkit, 2020). Sequencing layout (single-end or paired-end) was inferred from file structure. Reads were aligned to the D.

melanogaster reference genome (FlyBase r6.45) (Thurmond et al., 2019) using HISAT2 v2.2.1 with default parameters and the --dta and -t options enabled (Kim et al., 2019). Paired-end and single-end reads were mapped using the – 1/-2 and -U options, respectively. Alignment outputs were converted to sorted and indexed BAM files using SAMtools v1.13 (Li et al., 2009). FASTQ files were discarded after successful alignment to minimize storage requirements.

Transcriptome assembly

Transcriptome assembly was performed using StringTie v2.1.7 in reference-guided mode with the FlyBase annotation as a guide (-G dmel_flybase.gtf), enabling estimation of transcript abundances (-B) and gene-level expression (-A) (Pertea et al., 2015). Assembled GTF files from all samples were merged using stringtie --merge to generate a unified transcriptome annotation encompassing all tissues, developmental stages, and experimental conditions.

Identification of lncRNAs and novel protein-coding transcripts

Candidate long non-coding RNAs (lncRNAs) were identified using CPAT v3.0.4 with a Drosophila-specific hexamer table, applying a coding probability cutoff of < 0.39 (Wang et al., 2013). Transcripts showing significant similarity to Swiss-Prot proteins were excluded based on DIAMOND BLASTX v2.1.8 searches (E-value < 1e–5) (Buchfink et al., 2015). To further eliminate false positives, open reading frames were scanned using getorf (EMBOSS v6.6.0) (Rice et al., 2000), and conserved protein domains were identified using Pfam-A (pfam_scan.pl) (Finn et al., 2016). Structured RNAs were removed using Infernal v1.1.4 against the Rfam covariance model database (Rfam.cm; score ≥ 40) (Nawrocki & Eddy, 2013). Transcripts passing all filters were classified as high-confidence lncRNAs.

Protein-coding transcripts were identified using TransDecoder v5.5.0 (Haas et al., 2013). Transcripts with CPAT coding probabilities ≥ 0.39 and significant matches to Swiss-Prot Insecta proteins (sequence identity > 80%) were retained as protein-coding. Only transcripts with FPKM ≥ 1 in at least one RNA-seq dataset were considered high-confidence novel mRNAs.

Structure prediction using AlphaFold 3

The top-ranked model for each protein was selected based on the composite ranking_score, defined as:

ranking_score = 0.8×ipTM + 0.2×pTM + 0.5×disorder − 100×has_clash

The highest-scoring model was retained and exported in mmCIF format. To ensure compatibility with downstream structural comparison tools, structures were converted to PDB format using Gemmi (Global Phasing Ltd., 2023). Terminal OXT atoms, which may interfere with structural alignment, were removed using a custom script, yielding refined structure files. These models were subjected to structure-based homology searches against two structural databases: the PDB100 non-redundant set of experimentally determined protein structures (Holm, 2023) and the AlphaFold Protein Structure Database Swiss-Prot subset (AFDB-SwissProt) (Varadi et al., 2022). Structural comparisons were performed using Foldseek (van Kempen et al., 2023), enabling sensitive detection of structural homologs in the absence of significant sequence similarity.

Detection of transcription readthrough events

StringTie-assembled transcripts were converted to BED format and compared with FlyBase annotations (Thurmond et al., 2019). Transcripts extending ≥ 5 kb downstream of annotated termination sites were classified as readthrough transcripts, consistent with previous definitions of downstream-of-gene transcription (Vilborg et al., 2015; Rosa-Mercado & Steitz, 2022). Overlapping transcription readthrough events were collapsed into the longest representative per locus.

Co-expression network construction and preservation analysis

FPKM values were grouped by tissue or developmental stage. Low-expression genes were filtered (FPKM > 0.1 in ≥ 5 samples; SD > 0.2). Co-expression networks were constructed using WGCNA v1.71 (Langfelder & Horvath, 2008), with soft-thresholding

powers determined by pickSoftThreshold. Networks were built from adjacency and TOM matrices, and modules were defined via dynamic tree cutting (minClusterSize = 30, deepSplit = 2).

Module preservation was assessed using NetRep v1.2.7 with 1,000 permutations (Ritchie et al., 2016). Embryo networks served as the discovery set, and preservation was tested in other stages and tissues. Modules were considered preserved if most NetRep statistics yielded empirical $P < 0.05$.

Functional enrichment analysis

Functional enrichment analysis was performed to identify over-represented Gene Ontology (GO) terms and KEGG pathways among gene sets of interest (e.g., differentially expressed genes or co-expression modules). Protein sequences were annotated using eggNOG-mapper v2 (http://eggnog-mapper.embl.de/) to assign Gene Ontology (GO) terms and KEGG Orthology (KO) identifiers (Huerta-Cepas et al., 2017; Cantalapiedra et al., 2021). The resulting annotations were processed to generate TERM2GENE and TERM2NAME mapping files. Gene identifiers were converted to Entrez Gene IDs using bioDBnet (https://biodbnet-abcc.ncifcrf.gov/) to ensure compatibility with downstream analysis. Enrichment analysis was conducted in R (v4.x) using the clusterProfiler package (Yu et al., 2012), which implements Over-Representation Analysis (ORA) via the enricher() function, with all annotated genes in the genome as the background. A term was considered significantly enriched if it met: adjusted p-value (Benjamini–Hochberg) < 0.05, minimum gene set size ≥ 5, maximum gene set size ≤ 500. GO analysis covered the three canonical ontologies: Biological Process (BP), Molecular Function (MF), and Cellular Component (CC) (Ashburner et al., 2000; Gene Ontology Consortium, 2023). KEGG pathway enrichment was based on KO assignments from eggNOG. Visualization and interpretation of results were supported by the enrichplot package (Yu et al., 2012).

Declarations

Corresponding authors

Correspondence to Fei Li and Kang He.

Competing interests

The authors declare no competing interests.

Author Contribution

Haonan Duanmu: Conceptualization, Formal analysis, Visualization, Data curation, Writing – original draft. Meizhen Li: Formal analysis, Methodology, Investigation, Resources. Zihan Zhou: Resources, Investigation. Xinyan Li: Resources, Investigation. Hao Chen: Resources, Investigation. Kang He: Conceptualization, Writing – review & editing, Supervision, Funding acquisition. Fei Li: Conceptualization, Writing – review & editing, Supervision, Funding acquisition.

Acknowledgements

This work was supported by 2024 Yangtze River Delta Science and Technology Innovation Community Joint Research (Basic Research) Project (2024CSJZN0900), the Zhejiang Provincial Agricultural Project (2025SNJF046), the Zhejiang Provincial Natural Science Foundation of China (LZ23C140002) and the Fundamental Research Funds for the Central Universities (226-2025-00207)

References

1. Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 630, 493–500 (2024). https://doi.org/10.1038/s41586-024-07487-w
2. Bennett, C. J. et al. A re-analysis of an existing *Drosophila melanogaster* RNA-seq dataset reveals expression variability across conditions. bioRxiv (2024). https://doi.org/10.1101/2024.04.10.588867
3. Brown, J. B. et al. Diversity and dynamics of the *Drosophila* transcriptome. Nature 512, 393–399 (2014). https://doi.org/10.1038/nature12962
4. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. Nat. Methods 12, 59–60 (2015). https://doi.org/10.1038/nmeth.3176
5. Burghardt, E. et al. Transcriptome analysis reveals temporally regulated genetic networks during border cell migration. BMC Genomics 24, 9839 (2023). https://doi.org/10.1186/s12864-023-09839-8
6. Caldas, J. et al. Transcription readthrough is prevalent in healthy human tissues and associated with inherent genomic features. Genome Res. 34, 123–135 (2024). https://doi.org/10.1101/gr.277903.123
7. Derrien, T. et al. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res. 22, 1775–1789 (2012). https://doi.org/10.1101/gr.132159.111
8. D'Erchia, A. M. et al. The pervasive transcription of our genome: the challenge of functional characterization. Nucleic Acids Res. 43, 6807–6817 (2015). https://doi.org/10.1093/nar/gkv650
9. Djebali, S. et al. Landscape of transcription in human cells. Nature 489, 101–108 (2012). https://doi.org/10.1038/nature11233
10. Eisenberg, E. & Levanon, E. Y. Human housekeeping genes, revisited. Trends Biochem. Sci. 38, 565–567 (2013). https://doi.org/10.1016/j.tibs.2013.07.010
11. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012). https://doi.org/10.1038/nature11247
12. Finn, R. D. et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 44, D279–D285 (2016). https://doi.org/10.1093/nar/gkv1344
13. Fuchs, R. T. et al. Transcriptional readthrough is a regulated process with functional consequences. RNA 28, 1421–1435 (2022). https://doi.org/10.1261/rna.079143.122
14. Gingeras, T. R. Implications of the pervasive transcription of the human genome. Science 326, 405–408 (2009). https://doi.org/10.1126/science.1175953
15. Graveley, B. R. et al. The developmental transcriptome of *Drosophila melanogaster.* Nature 471, 473–479 (2011). https://doi.org/10.1038/nature09715
16. Grosso, A. R. et al. Pervasive transcription read-through promotes aberrant expression of oncogenes and chimeric RNAs in renal carcinoma. eLife 4, e09214 (2015). https://doi.org/10.7554/eLife.09214
17. Haas, B. J. et al. De novo transcript sequence reconstruction from RNA-seq: reference generation and analysis with Trinity. Nat. Protoc. 8, 1494–1512 (2013). https://doi.org/10.1038/nprot.2013.084
18. He, Y. et al. Transcriptional readthrough RNAs reflect the complexity of transcriptome regulation. RNA Biol. 15, 1671–1683 (2018). https://doi.org/10.1080/15476286.2018.1501729
19. Hertz, G. Z. et al. Genome-wide analysis of transcription readthrough in *Drosophila*. Genome Res. 23, 2042–2051 (2013). https://doi.org/10.1101/gr.159165.113
20. Jungreis, I. et al. Conservation of small open reading frames in *Drosophila* reveals widespread translation of novel peptides. eLife 10, e69619 (2021). https://doi.org/10.7554/eLife.69619
21. Kim, D. et al. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat. Biotechnol. 37, 907–915 (2019). https://doi.org/10.1038/s41587-019-0201-4
22. Kindgren, P. et al. Transcriptional read-through of the long noncoding RNA SVALKA governs plant cold acclimation. Nat. Commun. 9, 4561 (2018). https://doi.org/10.1038/s41467-018-07010-6

23. Kozlov, A. M. et al. Structure-guided functional annotation improves gene models in non-model organisms. Genome Res. 33, 1473–1485 (2023). https://doi.org/10.1101/gr.277758.122
24. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9, 559 (2008). https://doi.org/10.1186/1471-2105-9-559
25. Leinonen, R. et al. The Sequence Read Archive. Nucleic Acids Res. 39, D19–D21 (2011). https://doi.org/10.1093/nar/gkq1019
26. Li, H. et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079 (2009). https://doi.org/10.1093/bioinformatics/btp352
27. Li, Y. I. et al. RNA sequencing of 893 diverse human tissues reveals widespread isoform diversity. Cell 182, 1282–1296.e20 (2020). https://doi.org/10.1016/j.cell.2020.07.017
28. Ling, J. et al. Stress-induced transcription readthrough generates novel chimeric transcripts in mammalian cells. Cell Rep. 30, 3785–3797.e6 (2020). https://doi.org/10.1016/j.celrep.2020.02.075
29. Liu, Y. et al. Dynamic rewiring of gene regulatory networks during development and stress. Nat. Commun. 11, 3381 (2020). https://doi.org/10.1038/s41467-020-17183-4
30. Meyer, H. et al. Combined transcriptome and proteome profiling reveal cell-type-specific functions of *Drosophila* nephrocytes. Commun. Biol. 7, 1424 (2024).
31. Morgan, J. T. et al. Downstream-of-gene transcripts are a conserved feature of stress-induced transcription reprogramming. Mol. Cell 82, 2379–2394 (2022). https://doi.org/10.1016/j.molcel.2022.05.010
32. Muniz, L. et al. Control of gene expression in senescence by transcription read-through. Cell Rep. 21, 2433–2446 (2017). https://doi.org/10.1016/j.celrep.2017.11.009
33. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 29, 2933–2935 (2013). https://doi.org/10.1093/bioinformatics/btt509
34. Oates, M. E. et al. Protein structure reveals function beyond sequence homology in uncharacterized human proteins. Cell 186, 2635–2650.e18 (2023). https://doi.org/10.1016/j.cell.2023.04.017
35. Oldham, M. C. et al. Functional organization of the transcriptome in human brain. Proc. Natl. Acad. Sci. USA 105, 5076–5081 (2008). https://doi.org/10.1073/pnas.0800375105
36. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. 33, 290–295 (2015). https://doi.org/10.1038/nbt.3122
37. Pertea, M. et al. CHESS: a new human gene catalog curated from thousands of large-scale RNA-seq experiments. Genome Biol. 19, 208 (2018). https://doi.org/10.1186/s13059-018-1590-2
38. Pintarelli, R. et al. Read-through transcripts in normal human lung tissue: transcription overlaps with genes involved in cancer. Oncotarget 7, 26441–26456 (2016). https://doi.org/10.18632/oncotarget.8472
39. Qian, S. H. et al. Integrating massive RNA-seq data to elucidate transcriptome dynamics across development and tissues in *Drosophila*. BMC Genomics 24, 1000 (2023).
40. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 16, 276–277 (2000). https://doi.org/10.1016/S0168-9525(00)02024-2
41. Ritchie, S. C. et al. A scalable permutation approach reveals replication and preservation patterns of network modules in large datasets. Cell Syst. 3, 71–82 (2016). https://doi.org/10.1016/j.cels.2016.06.012
42. Rosa-Mercado, N. A. & Steitz, J. A. Who let the DoGs out? Biogenesis of stress-induced readthrough transcripts. Trends Genet. 38, 993–1006 (2022). https://doi.org/10.1016/j.tig.2022.07.007
43. Rutkowski, A. J. et al. Widespread disruption of host transcription termination in HSV-1 infection. Cell 160, 983–995 (2015). https://doi.org/10.1016/j.cell.2015.01.025
44. Sakaguchi, S. et al. Single-cell transcriptome atlas of *Drosophila* gastrula 2.0. Cell Rep. 40, 111234 (2023).
45. Schier, A. F. Transcriptional readthrough: a noisy nuisance or a regulatory feature? Trends Genet. 36, 795–798 (2020). https://doi.org/10.1016/j.tig.2020.07.009

46. Tabula Sapiens Consortium. The Tabula Sapiens: a multiple-organ, single-cell transcriptomic atlas of humans. Science 376, eabl4896 (2022). https://doi.org/10.1126/science.abl4896
47. Thurmond, J. et al. FlyBase 2.0: the next generation. Nucleic Acids Res. 47, D759–D765 (2019). https://doi.org/10.1093/nar/gky1003
48. Tilgner, H. et al. Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcription in the human genome but inefficient for lncRNAs. Nature 489, 574–578 (2012). https://doi.org/10.1038/nature11478
49. Tress, M. L. et al. The implications of alternative splicing in the ENCODE tissue transcriptomes. Mol. Syst. Biol. 13, 914 (2017). https://doi.org/10.15252/msb.20167372
50. Ulitsky, I. Evolution to the rescue: using comparative genomics to understand long non-coding RNAs. Nat. Rev. Genet. 17, 601–614 (2016). https://doi.org/10.1038/nrg.2016.85
51. Varadi, M. et al. The AlphaFold Protein Structure Database in 2022: expanding coverage to enable large-scale structural systems biology. Nucleic Acids Res. 51, D487–D494 (2022). https://doi.org/10.1093/nar/gkac1052
52. Vilborg, A. et al. Widespread inducible transcription downstream of human genes. Mol. Cell 59, 449–461 (2015). https://doi.org/10.1016/j.molcel.2015.06.016
53. Zhang, L. & Li, W.-H. Strong purifying selection on hub genes in protein–protein interaction networks. Mol. Biol. Evol. 38, 1058–1068 (2021). https://doi.org/10.1093/molbev/msaa277

Figures

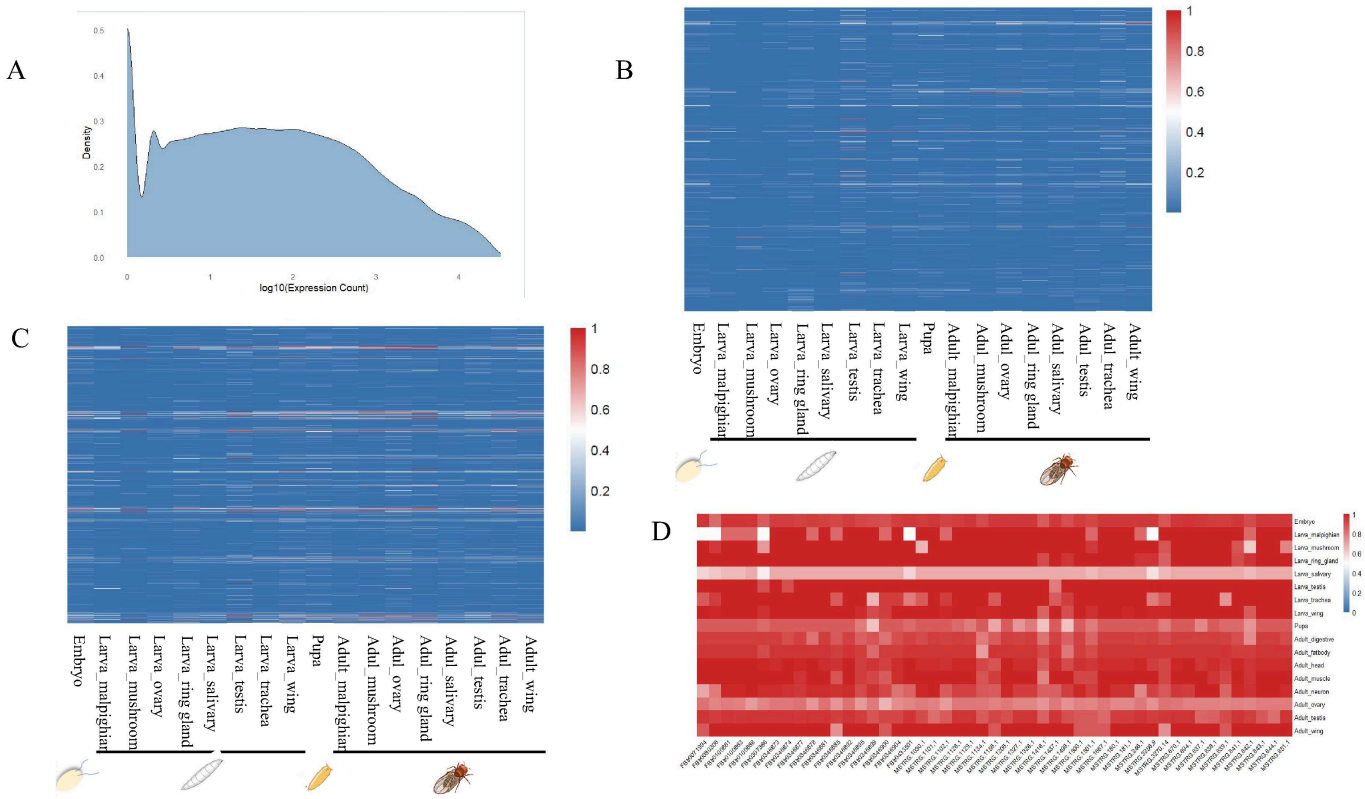


Figure 1

Classification and expression landscape of coding and noncoding transcripts in *D.melanogaster*. The reconstructed transcriptome reveals distinct lncRNA classes, a broad range of expression levels, and contrasting patterns of tissue- and stage-specific expression between coding and noncoding genes. (A) Density plot of transcript expression levels (log10 counts). (B) Heatmap of lncRNA expression across developmental stages and adult tissues. (C) Heatmap of protein-coding

transcript expression across the same contexts, highlighting a small subset of ubiquitously expressed “super-housekeeping” transcripts. (D) Heatmap showing expression profiles of candidate super-housekeeping transcripts across developmental stages and tissues. These transcripts were consistently detected in >90% of samples, highlighting their ubiquitous expression.

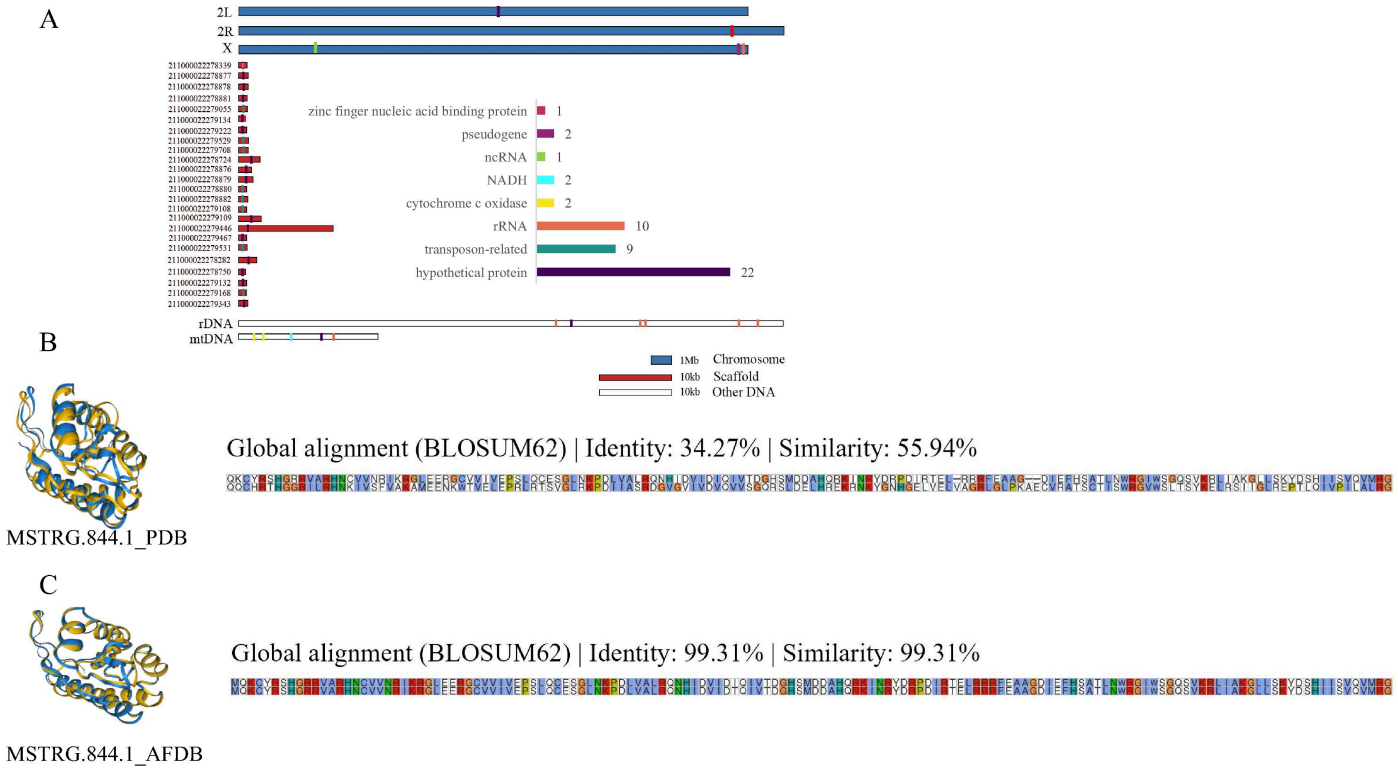


Figure 2

Identification and characterization of super-housekeeping transcripts in *D.melanogaster*. (A) Chromosomal distribution and functional classification of the 49 identified super-housekeeping transcripts. Categories include ribosomal RNAs (rRNA), mitochondrial respiratory chain subunits (NADH, cytochrome c oxidase), transposon-related transcripts, and 22 hypothetical proteins. (B) Structural match of the same transcript (MSTRG.844.1) in PDB100, showing perfect alignment with a *Bombyx mori* R2 retrotransposon endonuclease domain (probability = 1.0), highlighting conservation of its core functional domain across insects. (C) Structural match of candidate transcript MSTRG.844.1 in AlphaFold DB, showing perfect alignment with a *D.melanogaster* retrovirus-related Pol polyprotein (probability = 1.0).

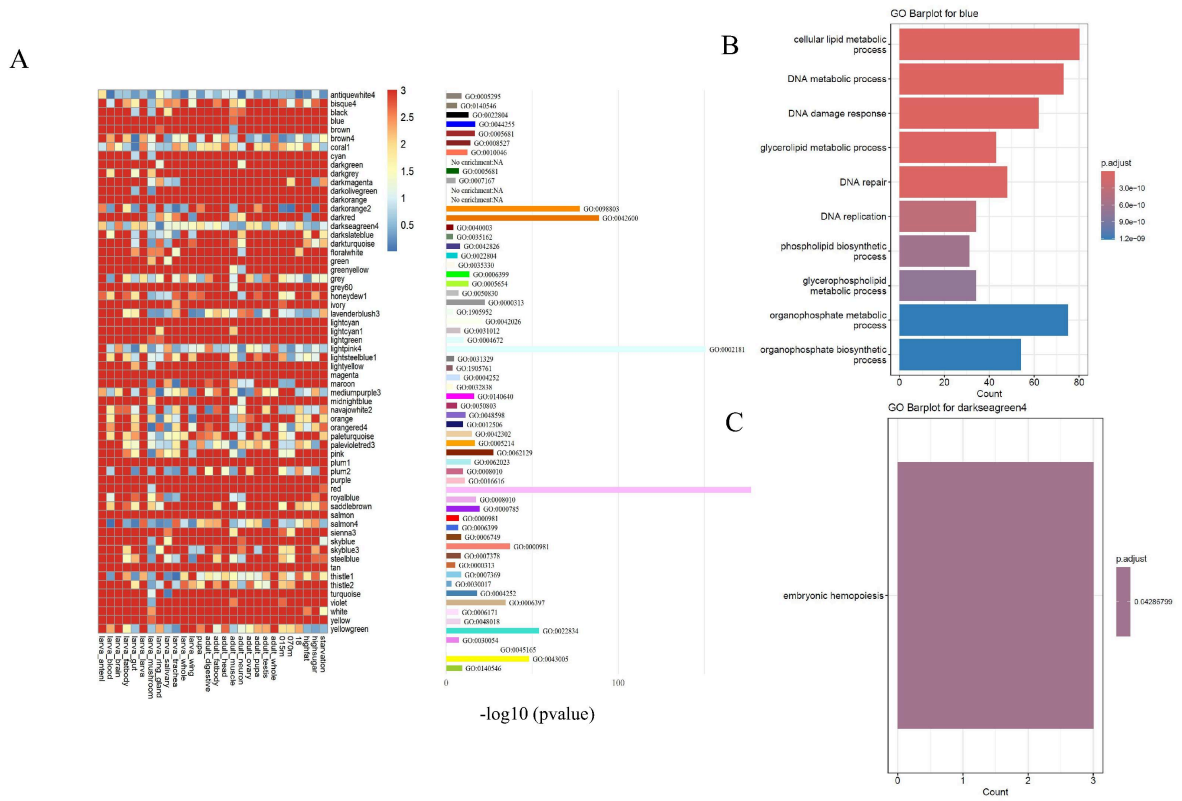


Figure 3

Co-expression network modules and their developmental preservation in *D.melanogaster*. (A) Heatmap of module–trait correlations across developmental stages and tissues. Color scale indicates correlation strength, with red representing positive associations and blue representing negative associations. (B) Gene Ontology (GO) enrichment analysis of representative modules. Each bar indicates significantly enriched biological processes, with bar length representing the number of genes and color indicating adjusted *p*-value. (C) Example of a stage-specific module (darkseagreen4), significantly enriched for embryonic hemopoiesis, illustrating functional specialization of peripheral modules.

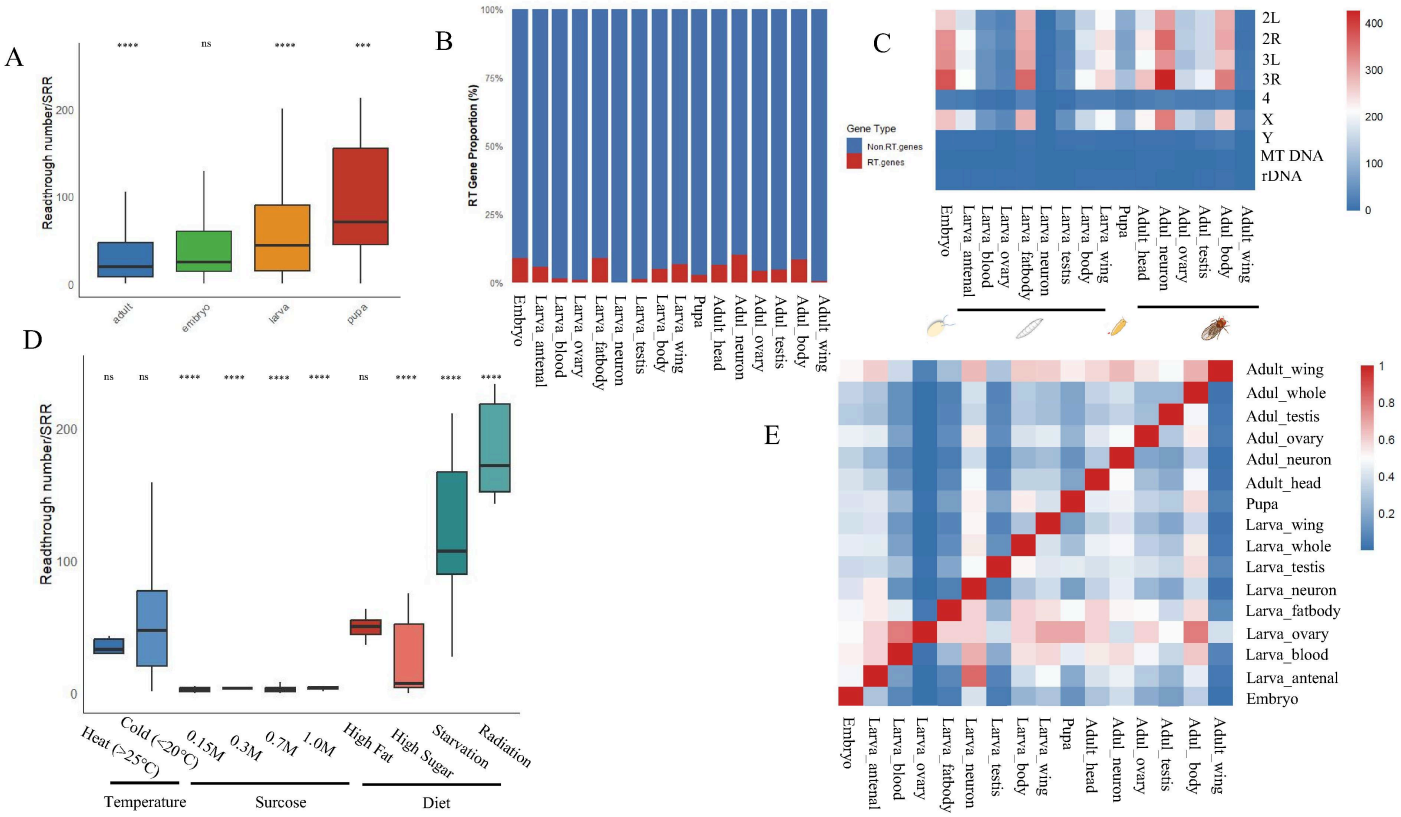


Figure 4

Developmental, tissue-specific, and stress-responsive patterns of transcriptional readthrough (RT) in *D. melanogaster*. (A) Boxplot showing the distribution of RT frequency across developmental stages. Pupae exhibited the highest incidence of RT, followed by larvae, whereas embryos and adults showed lower frequencies. (B) Proportion of RT versus non-RT genes across tissues and stages. RT events were observed in fewer than 10% of expressed genes per tissue. (C) Chromosomal distribution of RT genes across different developmental stages and tissues. RT was enriched in discrete autosomal regions but largely absent from the Y chromosome, rDNA, and mitochondrial loci. (D) Boxplot showing the effects of environmental perturbations on RT frequency. High-fat diet, starvation, and irradiation significantly increased RT incidence, with irradiation having the strongest effect, whereas high-sucrose treatment suppressed RT. Heat and cold stress produced only modest or nonsignificant effects. (E) Heatmap of pairwise overlap of RT genes across tissues and developmental stages, demonstrating strong context specificity.

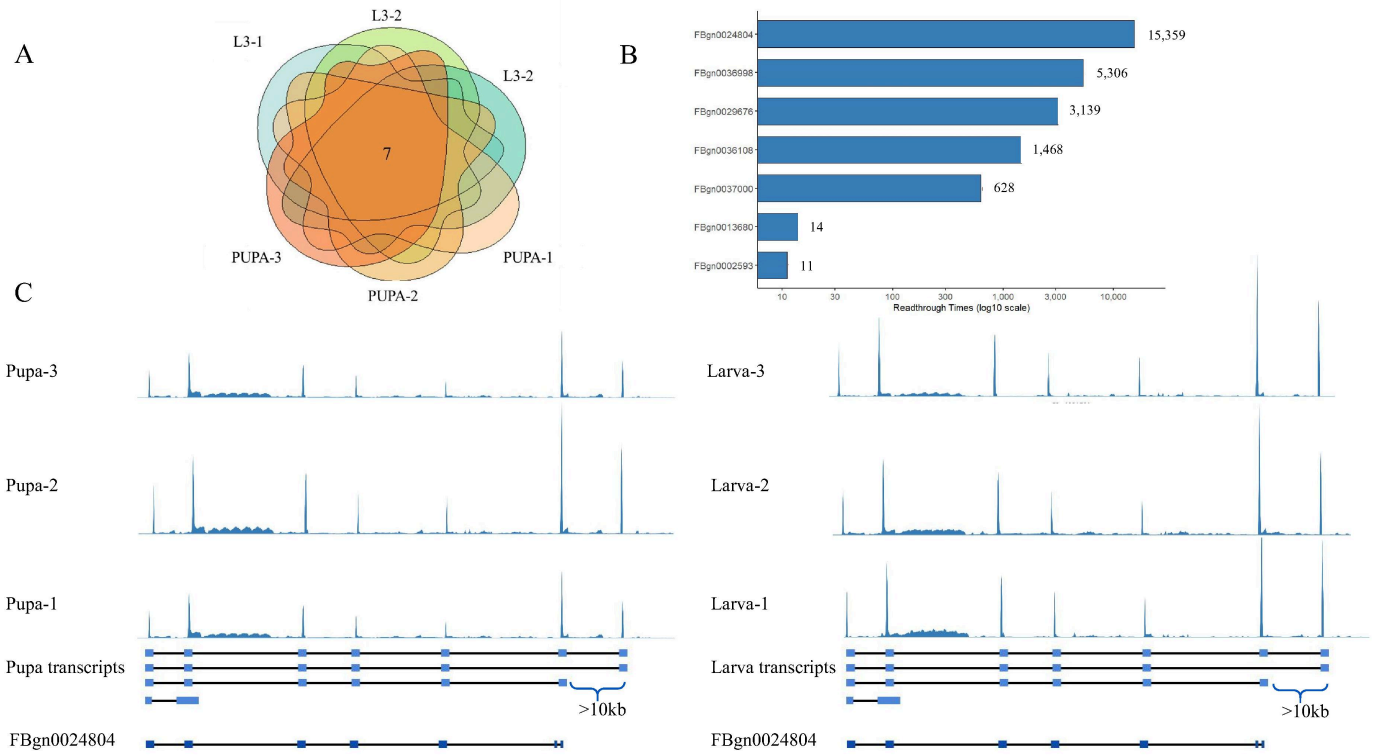


Figure 5

Conserved and high-frequency transcriptional readthrough (RT) genes in *D.melanogaster*. (A) Venn diagram showing overlap of RT genes detected across independent larval (L3) and pupal (PUPA) datasets, highlighting a conserved set of seven shared RT genes. (B) Bar plot of the most frequently detected RT genes. (C) Genome browser views of representative high-frequency RT loci. RNA-seq coverage tracks illustrate extended transcription signals beyond annotated termination sites across multiple datasets. Expanded locus diagrams show transcription extending >10 kb downstream of canonical gene boundaries (example: *FBgn0024804*). Blue boxes indicate annotated exons, and black lines denote extended RT regions.

Supplementary Files

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

- SuppleFig.docx
- SuppleTable.xlsx