



The landscape and evolutionary dynamics of derived lncRNAs in insects



Dong Jing¹, Yang Mei¹, Hang Zhou¹, Chenghao Huang¹, Zuoqi Wang¹, Shuo Jin¹, Hao Chen¹, Haonan Duanmu¹, Ying Liu^{1,2}✉, Fei Li¹ & Kang He¹✉

Long noncoding RNAs (lncRNAs) function as crucial regulatory elements in eukaryotes, yet have remained largely unexplored across diverse insect lineages. We present a comprehensive analysis of lncRNA atlases across 115 insect species, identifying 647,691 multi-exonic lncRNAs with substantial variation (1119 to 40,850 per species). Our analysis reveals that lncRNA abundance correlates positively with both genome size and transposable element (TE) content. Approximately 500 diverse transcriptomes were found to represent the minimum threshold for comprehensive lncRNA identification. While most lncRNAs show limited sequence conservation, we identified 4806 derived lncRNAs that maintain syntenic relationships with protein-coding genes. These derived lncRNAs, which are significantly shaped by TE insertions, exhibit higher expression levels, greater regulatory complexity, and stronger functional conservation compared to conventional lncRNAs. They are primarily involved in regulating development and insect behavior. One such example is *Msex*-*lnc-001918*, which is derived from a conserved Lepidoptera orthogroup and retains ancestral functions in regulating spermatogenesis. This study provides fundamental insights into insect lncRNA evolution and demonstrates how the transformation of protein-coding genes can drive novel regulatory mechanisms in insect genomes.

Insects represent one of the largest and most diverse groups of organisms, with over a million described species adapted to various environments. They play important roles in ecosystems¹, agriculture², human health³, and the economy⁴. Their remarkable adaptability and diversity are driven by complex regulatory networks, among which long non-coding RNAs (lncRNAs) have emerged as key players. These regulatory molecules, typically exceeding 200 nucleotides in length, participate in critical biological processes such as development^{5,6}, dosage compensation⁷, sex determination⁸, insecticide resistance^{9–11}, and immunity¹², despite their non-coding nature. Advances in high-throughput sequencing technologies have facilitated extensive characterization of insect lncRNAs, revealing their diverse functions and evolutionary dynamics. This has led to increasing interest in understanding their evolution and employing comparative genomics to study their functional determinants¹³.

Recent evolutionary studies show that lncRNAs can originate through multiple mechanisms. One particularly interesting pathway involves the transformation of protein-coding genes^{14,15}, through which the resulting derived lncRNAs often gain novel functions¹⁶. A classic example is the *Xist*

noncoding RNA in *Drosophila melanogaster*, which is evolved from a protein-coding gene and located in the X inactivation center (*Xic*). It initiates X chromosome inactivation, a process essential for dosage compensation in female mammals¹⁷. Similarly, *JPX*, a lncRNA derived from *Usp1*, plays a critical role in regulating *Xist*. The observation that deletion of the *JPX* locus impairs X-inactivation¹⁸ demonstrates that these derived lncRNAs can be integrated into complex regulatory networks.

While numerous studies of derived lncRNAs have made significant progress in vertebrates^{14,16}, the systematic identification and characterization remain largely unexplored in insects. Derived lncRNAs might play crucial yet undiscovered regulatory roles. This knowledge gap is particularly significant, as experimental validation of lncRNA functions remains technically challenging in insects. The absence of large-scale studies on insect derived lncRNAs represents a major gap in our understanding of transcriptome evolution, particularly given the ecological importance of insects. Therefore, investigating derived lncRNAs in insects not only addresses a fundamental gap in evolutionary biology but also provides a strategic approach to identifying functionally relevant

¹State Key Laboratory of Rice Biology and Breeding & Ministry of Agricultural and Rural Affairs Key Laboratory of Molecular Biology of Crop Pathogens and Insect Pests, Institute of Insect Sciences, Zhejiang University, Hangzhou, China. ²Key Laboratory of Green Prevention and Control of Agricultural Transboundary Pests of Yunnan Province/Agricultural Environment and Resource Research Institute, Yunnan Academy of Agricultural Sciences, Kunming, China.

✉ e-mail: liuying@yaas.org.cn; hekang@zju.edu.cn

lncRNAs from the vast pool of non-coding transcripts. In this study, we present a comprehensive analysis of insect lncRNAs across 115 species, utilizing an extensive dataset of RNA-seq samples obtained from Insectbase2.0¹⁹ and the National Center for Biotechnology Information (NCBI)²⁰. Through the previously synteny-based approaches^{14,16}, we identified 4806 derived lncRNAs, and clarified the potential role that TEs play in the evolutionary process. By comparing the expression and regulatory features of derived lncRNAs with other lncRNAs, we explored the potential functions of derived lncRNAs and demonstrated the preservation of the functions of syntenic genes during their transformation into lncRNAs. Our research contributes to the fundamental knowledge of insect lncRNA biology and provides insights into how derived lncRNAs serve as sources of important regulatory functions.

Results

Comprehensive characterization of lncRNA atlases across 115 insect species

To construct a comprehensive insect lncRNA atlas, our study integrated extensive genomic and transcriptomic data from 115 insect species. This dataset included high-quality reference genomes (BUSCO completeness > 90%) and 30,840 RNA-seq samples (totaling ~147 T, as of June 5, 2023), sourced from InsectBase 2.0¹⁹ and NCBI (Supplementary Data 3). Each species was represented by at least 50 samples from diverse developmental stages or tissues. The transcriptome coverage varied markedly across species, ranging from *Aedes aegypti* (2260 samples, ~12.4 T) to *Ephemera danica* (50 samples, ~39.2 G).

Using this comprehensive dataset, we performed genome-guided transcript assemblies and merged all samples for each species. Initial candidate lncRNAs were identified using the FeeLnc pipeline²¹, and potential protein-coding genes were subsequently filtered using CPC2²² tool (Fig. S1A). We thus identified a range of 1119 (*Cochliomyia hominivorax*) to 40,850 (*A. aegypti*) multi-exonic candidate lncRNAs per species, for a total of 647,691 lncRNAs (Fig. 1A). These lncRNAs were systematically classified into intergenic (including divergent, convergent, and same strand) and genic (including overlapping, containing, and nested)²¹, based on their genomic loci relative to protein-coding genes (Fig. S1A). Notably, nested lncRNAs were the most abundant, while containing lncRNAs were the least abundant across all orders (Fig. S1B).

We observed a substantial variation (~36-fold) in lncRNA numbers across species, with Lepidoptera and Hemiptera showing higher average numbers (Fig. S1C). This variability necessitated careful consideration of the effects of data quality and quantity. To address this, we performed iterative lncRNA predictions using incrementally increasing sample numbers (in 50-sample intervals). This analysis revealed that the number of identified lncRNAs increased rapidly with sample size and plateaued when a species was represented by ~500 transcriptome datasets covering diverse developmental stages and tissue types (Fig. 1B). However, we emphasize that this is not an absolute threshold, as lncRNA discovery also depends on sequencing depth, library quality, and transcriptomic complexity. Thus, the “500-sample” point should be interpreted as a general benchmark observed in our dataset, rather than a strict threshold.

Based on these findings, we focused our subsequent analyses on species with more than 500 samples to explore factors influencing lncRNA abundance. This refined dataset included 13 species from five orders: Coleoptera (*Tribolium castaneum*), Diptera (*Drosophila pseudoobscura*, *Anopheles gambiae*, *Drosophila simulans*, *A. aegypti*), Hemiptera (*Acyrthosiphon pisum*), Hymenoptera (*Nasonia vitripennis*, *Bombus terrestris*, *Apis mellifera*), and Lepidoptera (*Helicoverpa armigera*, *Spodoptera frugiperda*, *Melitaea cinxia*, *Bombyx mori*). Among these species, genome size varied by ~54-fold, and we observed a strong positive correlation between lncRNA numbers and genome size ($R = 0.927$, $P < 0.001$, Fig. 1C), indicating that larger genomes tend to harbor more lncRNAs. Additionally, lncRNA abundance showed a significant correlation with genomic transposon content ($R = 0.510$, $P < 0.05$, Fig. 1D), supporting previous observations that link lncRNA origins to transposons^{13,23,24}.

Evolutionary analysis of lncRNA orthologous families

To enable direct cross-species comparisons, we reconstructed lncRNA orthologous families (orthogroups) based on sequence similarity using OrthoFinder^{25,26}. This analysis successfully identified 60,177 lncRNA orthogroups, with a striking observation that over 50% were species-specific (Fig. 2A, Fig. S2). This high proportion of species-specific lncRNAs highlights their rapid evolutionary turnover, resulting in poor sequence conservation. Notably, we also observed a strong positive correlation ($R = 0.91$, $P < 0.001$, Fig. S1D) between the total number of lncRNAs and the count of species-specific lncRNAs, demonstrating that the expansion of lncRNAs is predominantly driven by the emergence of species-specific sequences.

To better understand the evolutionary dynamics of lncRNAs, we performed a parallel analysis on protein-coding gene orthogroups (Fig. 2A, Fig. S2). Comparative analysis between lncRNAs and protein-coding genes revealed distinct evolutionary patterns. Although there were no 1:1 protein-coding orthologs among 115 insect species, 170 orthogroups were conserved across all species (Fig. 2A, Fig. S2). In contrast, lncRNAs showed markedly lower conservation. The most widespread lncRNA orthogroup was present in only 81 species (Fig. 2A, Fig. S2). Furthermore, more than 98.5% of lncRNA orthogroups were restricted to fewer than 10% of the species (Fig. 2A, Fig. S2). These results underscore the profoundly lower sequence conservation of lncRNAs compared to protein-coding genes, highlighting the rapid evolutionary turnover of lncRNA sequences. Given that poor sequence conservation is a common feature of lncRNAs in both mammals and plants^{27,28}, we next assessed their positional conservation using established syntenic relationships²⁹. Positional conservation occurs when lncRNAs in different species are found flanking orthologous genes within a certain distance and have the same relative orientation. Despite the predominance of species-specific lncRNAs, we identified 21 lncRNA orthogroups with conserved syntenic relationships (Fig. 2A, Fig. S2), indicating that some lncRNAs, while lacking detectable primary-sequence conservation, retain positional conservation.

Origins and evolution of syntenic lncRNAs from protein-coding genes

Through systematic synteny analysis, we found that a subset of lncRNAs maintains syntenic relationships with protein-coding genes. Previous studies have proposed that these lncRNAs, referred to as derived lncRNAs, may have evolved from protein-coding gene loci that have lost their coding capacity^{14,16}. To investigate lncRNAs potentially derived from protein-coding genes, we applied a previously described approach¹⁴ to examine this evolutionary process.

First, we employed MUMmer³⁰ for whole-genome alignment between pairs of species to identify the genomic alignment chains. We found low genome alignment rates even among closely related species within the same family, highlighting the rapid evolution of insect genomes (Fig. S3). Homologous protein-lncRNA pairs (X and Y) were identified based on the following criteria: (i) X and Y represent the flanking genes of the syntenic protein-coding gene and the lncRNA, respectively, located within 100 kb and showing confirmed orthology; (ii) the relative orientation between the syntenic gene and X, and between the lncRNA and Y, is consistent (Fig. S4A); and (iii) both gene pairs are located on the same strand (Fig. S4B). Finally, we considered pairs of lncRNA and protein-coding genes when there was a homologous protein-coding gene pair on one side and a genomic alignment chain on the other (Fig. S4C). These identified lncRNAs were defined as derived lncRNAs.

This approach identified 4806 derived lncRNAs across six orders, with *D. melanogaster* having 366 derived lncRNAs, making it the species with the highest number (Fig. 2B). Since Blattodea and Coleoptera have a small number of derived lncRNAs, we therefore only retained data from Hemiptera, Hymenoptera, Lepidoptera, and Diptera for further analysis. Among these four orders, the number of derived lncRNAs in Hymenoptera was significantly higher than in Hemiptera and Lepidoptera ($**P < 0.01$, Wilcoxon's test), while there were no significant differences compared to Diptera (Fig. S5A). Furthermore, an analysis of genome alignment rates

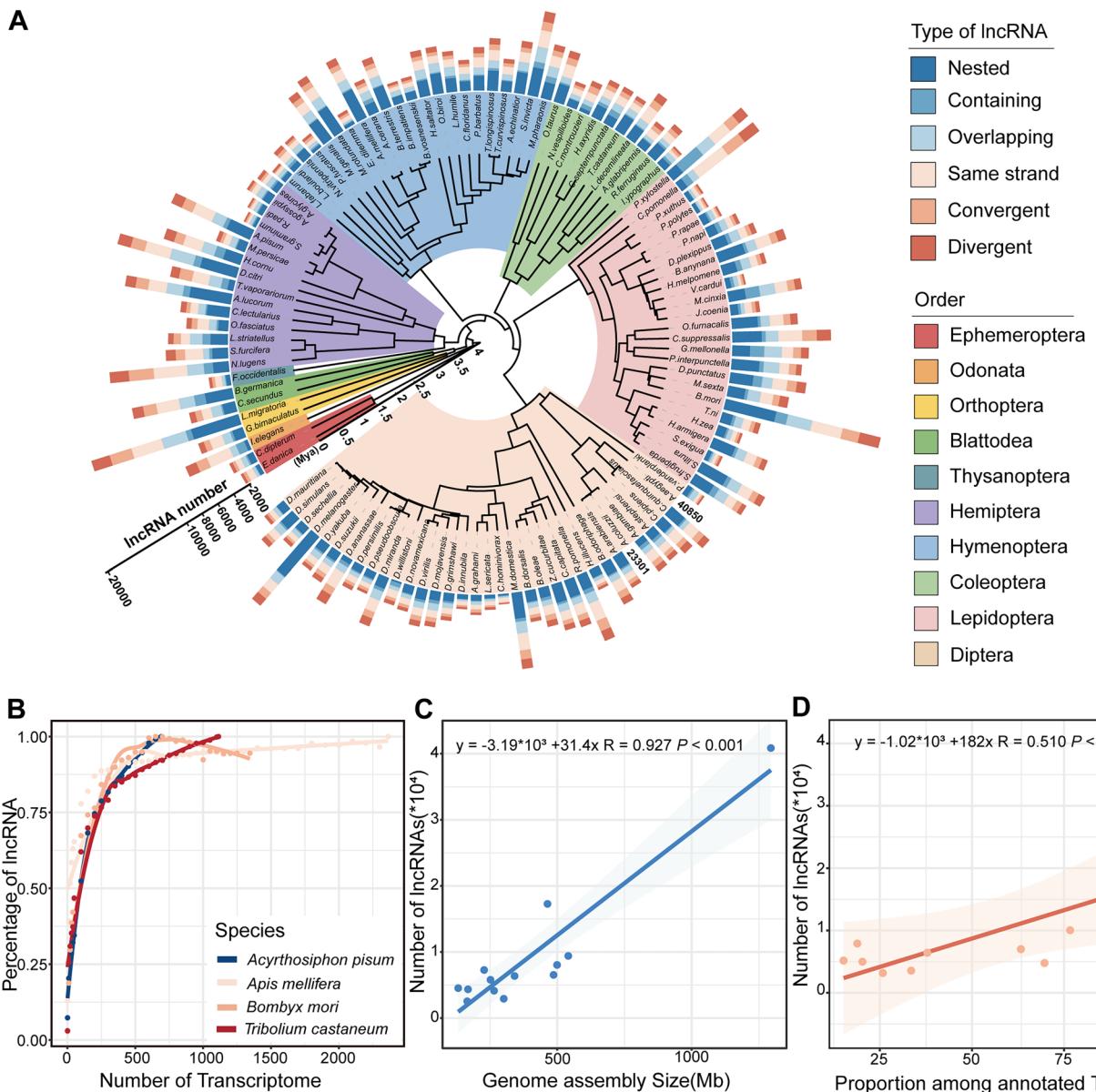


Fig. 1 | Identification of insect lncRNAs and factors influencing lncRNA quantity. **A** Phylogenetic distribution of lncRNA types across 115 insect species representing 10 orders. Bar colors represent different lncRNA types, and the height indicates the total count of identified lncRNAs per species. The tree indicates phylogenetic relationships among the species. **B** Saturation curves showing cumulative lncRNA numbers plotted against increasing transcriptome sample sizes (increments of 50) for four representative species: *A. pisum* (pea aphid, n = 694), *A. mellifera* (honey bee, n = 2362), *B. mori* (silkworm, n = 1344), and *T. castaneum* (red flour beetle, n = 1113). **C** Strong positive correlation between genome size and lncRNA

abundance ($R = 0.927, P < 0.001$) among species with >500 transcriptomes. Scatter plot displays genome size (x-axis, in Mb) versus total lncRNA count (y-axis). Each point represents a species, with regression line and confidence interval shown.

D Significant positive correlation between lncRNA abundance and genomic TE content ($R = 0.510, P < 0.05$) among species with >500 transcriptomes. Scatter plot shows the percentage of genome covered by TEs (x-axis) versus total lncRNA count (y-axis). Each point represents one species, with regression line and confidence interval shown.

between different species revealed that Hymenoptera insects exhibited higher alignment rates than the other orders ($***P < 0.001$, $****P < 2.2 \times 10^{-16}$, Wilcoxon's test) (Fig. S5B). This suggests a correlation between the number of derived lncRNAs and genome alignment rates, highlighting the potential influence of genomic conservation on lncRNA evolution across different insect lineages.

Evolutionary transitions in sequence architecture

To further investigate the origin of these derived lncRNAs, we assessed whether derived lncRNAs differ from their syntenic genes in terms of sequence features. Detailed sequence analysis revealed substantial

divergence between derived lncRNAs and their syntenic genes, with similarity ranging from 0.6% to 85.1% and an average value of 26.7% (Fig. 3A). Characteristically, syntenic genes were generally more compact than typical protein-coding genes ($P < 2.2 \times 10^{-16}$, Wilcoxon's test) (Fig. 3B), with notably shorter intronic regions (Fig. S6A, B). During the transition from protein-coding genes to derived lncRNAs, derived lncRNAs contained significantly fewer exons and introns than their syntenic genes ($****P < 2.2 \times 10^{-16}$, Wilcoxon's test) (Fig. S6A, B). Moreover, the average length of both exons and introns was markedly shorter in derived lncRNAs ($****P < 2.2 \times 10^{-16}$, Wilcoxon's test) (Fig. S6C, D), consistent with the overall reduction in transcript size (Fig. S3B). This contrasts with prior

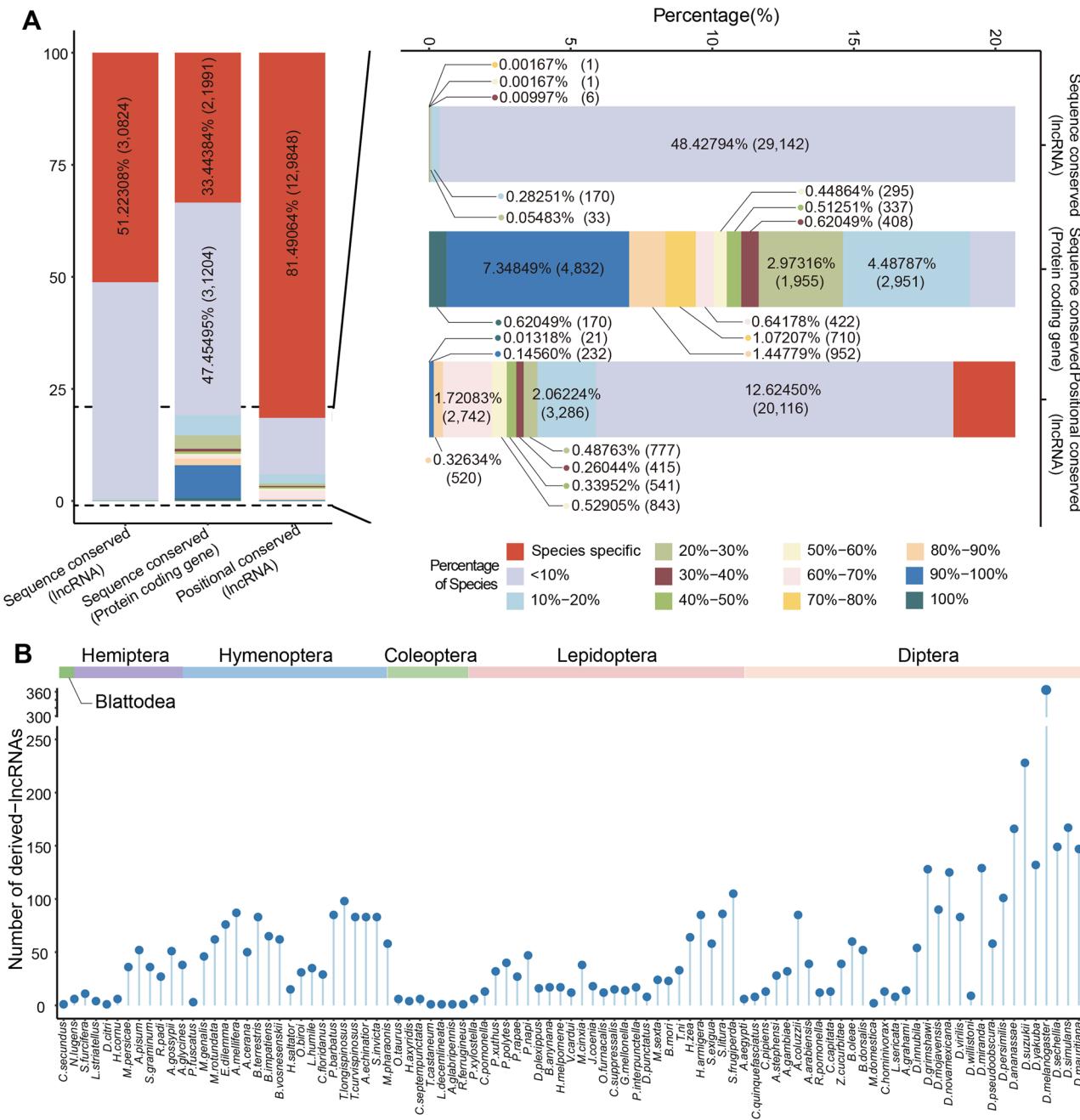


Fig. 2 | Evolutionary analysis of lncRNA orthogroups and identification of derived lncRNAs. A Comparative analysis of orthogroup distribution patterns among primary sequence conserved lncRNAs, protein-coding genes, and positional conserved lncRNAs. The stacked bar chart illustrates the relative proportions of

orthogroups shared across different numbers of species. Colors indicate the percentage of species sharing each orthogroup category. **B** Lollipop plot shows the number of derived lncRNAs detected in each species, arranged by taxonomic relationships.

findings in mammals, where noncoding genes often exhibit larger transcripts and longer introns than protein-coding genes³¹. Our results suggest that in insects, the structural reduction of derived lncRNAs during the transition from protein-coding genes may follow a distinct evolutionary trajectory.

Analysis of TE content revealed distinct patterns. Other lncRNAs exhibited a significantly higher fraction of TE-overlapping sequences than derived lncRNAs ($P < 2.2 \times 10^{-16}$, Wilcoxon's test). Similarly, a smaller fraction of syntenic gene sequences overlapped with TEs than other protein-coding genes ($P < 2.2 \times 10^{-16}$, Wilcoxon's test). However, we observed a significant increase in TE content during the transition from protein-coding genes to lncRNAs ($P < 2.2 \times 10^{-16}$, Wilcoxon's test) (Fig. 3C). TEs

preferentially overlapped with intronic regions ($***P < 0.001$, $****P < 2.2 \times 10^{-16}$, Wilcoxon's test), whereas the exonic regions of protein-coding genes exhibited low TE overlap (Fig. 3D). Strikingly, the overlap in exonic regions increased substantially when protein-coding genes evolved into lncRNAs ($P < 2.2 \times 10^{-16}$, Wilcoxon's test) (Fig. 3D).

The TE landscape in these sequences was dominated by DNA and LTR transposons, with LINE elements showing a unique pattern. Despite constituting only 12.28% of TEs in derived lncRNAs, LINEs showed distinctive enrichment in exonic regions during the transition from protein-coding genes to lncRNAs (Fig. 3E). This pattern suggests that specific TE classes may play a role in the evolutionary transformation from protein-coding genes to lncRNAs.

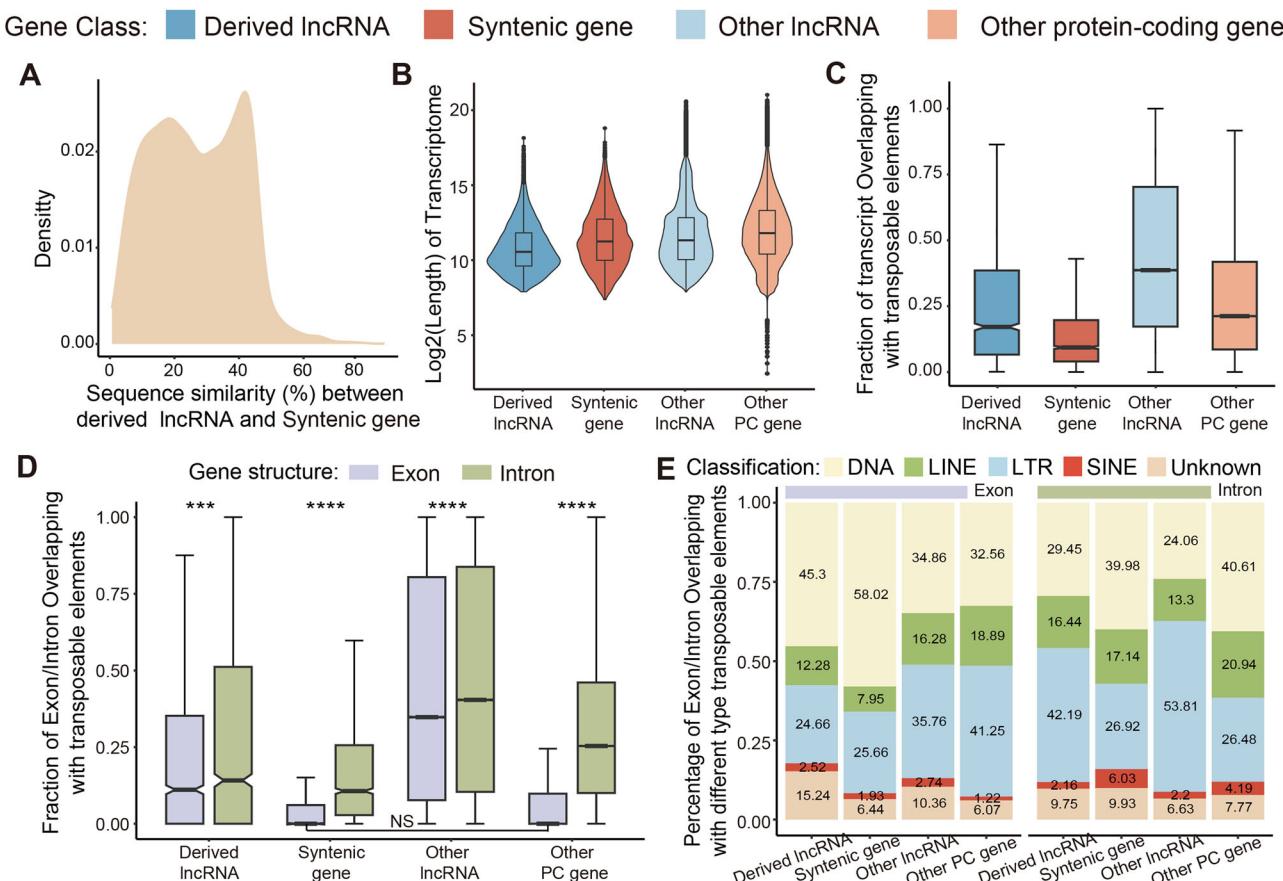


Fig. 3 | Comparative analysis of sequence features between derived lncRNAs and their syntenic genes. A Distribution of the sequence similarity between derived lncRNAs and their corresponding syntenic protein-coding genes. B Comparative analysis of transcript lengths (\log_2 -transformed) across four gene categories, statistical significance ($P < 0.0001$, Wilcoxon's test) indicated for pairwise comparisons, unless indicated otherwise. C TE overlap fractions in transcripts of the four gene categories, statistical significance ($P < 0.0001$, Wilcoxon's test) indicated for pairwise comparisons, unless indicated otherwise. D Detailed analysis of TE distribution in exonic and intronic regions. Paired box plots compare TE fractions in exons versus

introns within each gene type (*** $P < 0.001$ and **** $P < 0.0001$, Wilcoxon's test), with statistical significance shown for both within-type and between-type comparisons ($P < 0.0001$, Wilcoxon's test). E Compositional analysis of TE types (DNA transposons, LTRs, LINEs, etc.) in each gene category. Stacked bar plot shows the relative abundance of different TE classes within each gene type. Blue = derived lncRNAs ($n = 4806$), red = syntenic protein-coding genes ($n = 6290$), light blue = other lncRNAs ($n = 642,885$), orange = other protein-coding genes ($n = 2,030,166$). Error bars represent standard deviation unless otherwise indicated.

Regulatory landscape and expression patterns of derived lncRNAs

To investigate the regulatory characteristics that distinguish derived lncRNAs from other lncRNAs, we systematically analyzed their genomic positioning and associated regulatory elements. Derived lncRNAs showed distinct spatial distribution patterns and regulatory features relative to other lncRNAs. Our investigation revealed that derived lncRNAs were, on average, associated with 20 adjacent protein-coding genes within 100 kb, exceeding the number observed for other lncRNAs ($P < 2.2 \times 10^{-16}$, Wilcoxon's test) (Fig. 4A). Additionally, conservation analysis revealed that 10% of derived lncRNAs were retained across five or more species, showing higher conservation compared to other lncRNAs ($P < 2.2 \times 10^{-16}$, Wilcoxon's test). In contrast, their corresponding syntenic protein-coding genes were less conserved than other protein-coding genes ($P < 2.2 \times 10^{-16}$, Wilcoxon's test) (Fig. 4B).

Investigation of regulatory elements revealed enhanced regulatory potential in derived lncRNAs. We used miRanda³² to predict the miRNA binding sites, showing that derived lncRNAs have an average of 11.4 putative miRNA binding sites, significantly more than other lncRNAs ($P < 2.2 \times 10^{-16}$, Wilcoxon's test) (Fig. 4C). Additionally, derived lncRNA promoter regions showed enhanced transcription factor binding compared to other lncRNAs (Fig. 4D), suggesting more complex transcriptional regulation.

Because ncRNA expression can correlate with distal (*trans*) protein-coding genes³³, we performed co-expression analysis (Pearson's $r > 0.5$). Derived lncRNAs had significantly more co-expressed protein-coding partners than other lncRNAs ($P < 2.2 \times 10^{-16}$, Wilcoxon's test) (Fig. 4E). Moreover, derived lncRNAs exhibited higher expression levels than both other lncRNAs and protein-coding genes (**** $P < 2.2 \times 10^{-16}$, Wilcoxon's test) (Fig. 4F). Both derived lncRNAs and their syntenic genes showed similar patterns of tissue specificity (not significant, Wilcoxon's test), with broader expression profiles than other lncRNAs and protein-coding genes (Fig. 4G).

Co-expression between lncRNAs and adjacent protein-coding genes

We next assessed the potential *cis*-regulatory functions of derived lncRNAs by analyzing their co-expression patterns with neighboring protein-coding genes within 100 kb. We employed a guilt-by-association approach, using co-expression with protein-coding genes to infer the putative functions of derived lncRNAs. Most expression correlations between derived lncRNAs and their adjacent protein-coding genes were below $|r| = 0.5$, and genomic distance did not significantly affect correlation strength ($R = 0.22$, $P > 0.05$). However, we observed a trend in which increasing distance was associated with decreased correlation (Fig. 5A).

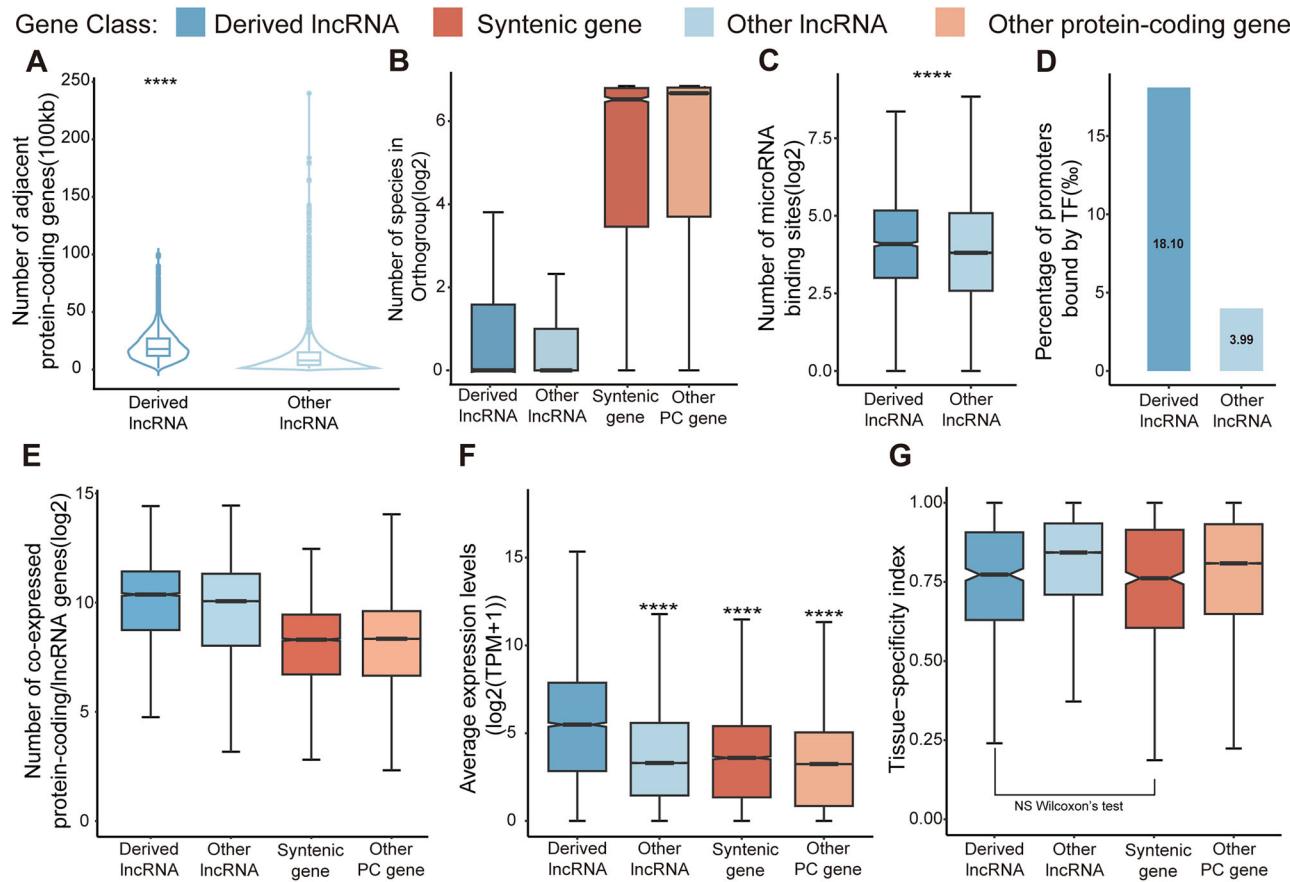


Fig. 4 | Comprehensive characterization of derived lncRNAs regulation and expression patterns. A Violin plot the number of adjacent protein-coding genes, comparing derived lncRNAs and other lncRNAs across all species ($****P < 0.001$, Wilcoxon's test). B Evolutionary conservation measured as the average number of species (\log_2 -transformed) per orthogroup across the four gene categories, with statistical significance ($P < 0.01$, Wilcoxon's test) indicated for pairwise comparisons unless otherwise noted. C Abundance of putative miRNA binding sites (\log_2 -transformed) in derived versus other lncRNAs ($P < 0.0001$, Wilcoxon's test). D Proportion of lncRNAs with transcription factor binding sites (TFBS) in their promoter regions, comparing derived lncRNAs and other lncRNAs. E Number of co-expressed mRNA/lncRNA pairs (\log_2 -transformed), with statistical significance

($P < 0.01$, Wilcoxon's test) indicated for pairwise comparisons unless otherwise noted. F Average expression levels [$\log_2(\text{TPM} + 1)$] across the four gene categories, showing consistently higher expression in derived lncRNAs compared to other categories ($****P < 0.001$, Wilcoxon's test). G Tissue-specificity indices (0 = broad expression, 1 = specific expression) across the four gene categories, with statistical significance ($P < 0.01$, Wilcoxon's test) indicated for pairwise comparisons unless otherwise noted. Derived lncRNAs ($n = 508$), syntenic protein-coding genes ($n = 626$), other lncRNAs ($n = 102,653$), other protein-coding genes ($n = 193,614$), blue = derived lncRNAs ($n = 4806$), red = syntenic protein-coding genes ($n = 6290$), light blue = other lncRNAs ($n = 642,885$), orange = other protein-coding genes ($n = 2,030,166$). Error bars represent standard deviation unless otherwise indicated.

In total, we identified 13,424 lncRNA-mRNA co-expression pairs across species, with *Drosophila yakuba* showing the highest number, reaching 904 pairs (Fig. S7A). We then conducted Gene Ontology (GO) analysis on the lncRNA-mRNA pairs for each species and calculated the similarity among GO terms, followed by a statistical assessment of the functional categories across different orders. The results indicated that a specific class of derived lncRNAs in insects was significantly enriched in developmental and metabolic processes, with 155 and 153 associated terms, respectively. Additionally, the regulation of biological processes (111 terms) highlighted the important role of derived lncRNAs in regulating insect biological processes. Further, we identified a diverse range of functions, including cell cycle processes, cellular localization, immunity, and responses to stimuli. Notably, the process of insect behavior (21 terms) was relatively sparse. However, it is involved in male courtship, circadian rhythm, feeding and other important insect activities (Fig. 5B, Supplementary Data 4).

Derived lncRNAs exhibit functional conservation with syntenic protein genes

To understand the functional implications of derived lncRNAs, we analyzed cases in which species possessed derived lncRNAs but lacked their syntenic gene homologs (Fig. S7B). Comparative genomic analysis in Lepidoptera

revealed that *Msex-lnc-001918* is a derived lncRNA in *Manduca sexta*, whereas its syntenic regions in other species retain protein-coding genes (*Hmel006382.1*, *Harm008025.1*, and *Prap015593.1*). These syntenic genes belong to the OG0013937 orthogroup, which exhibits order-specific conservation in Lepidoptera with 1:1:1 orthologous relationships and characteristically low expression levels (Fig. S8A).

Gene structural analysis revealed distinct architectural features of *Msex-lnc-001918*. Its first exon (~3600 bp in length) contained four TEs (two LINEs, one SINE, and one unclassified) but shows no sequence homology with syntenic genes. The subsequent two exons demonstrated marked sequence similarity to syntenic protein-coding genes. In contrast, none of the syntenic genes exhibited TE overlaps in exons, with only *H. armigera* containing a LINE element in its intronic region (Fig. S8B).

We then analyzed the expression patterns of *Msex-lnc-001918* across different tissues and found it to be tissue-specific, with the highest expression in the testis ($\tau = 0.88$, $*P < 0.05$, $**P < 0.01$, Wilcoxon's test) (Fig. 6A). Remarkably, *Msex-lnc-001918* exhibited higher expression levels compared to its syntenic genes ($****P < 2.2 \times 10^{-16}$, Wilcoxon's test) (Fig. 6B).

In order to infer the potential functions of these genes, we employed Weighted Gene Co-expression Network Analysis (WGCNA) to analyze gene co-expression networks of lncRNAs and protein-coding genes in each

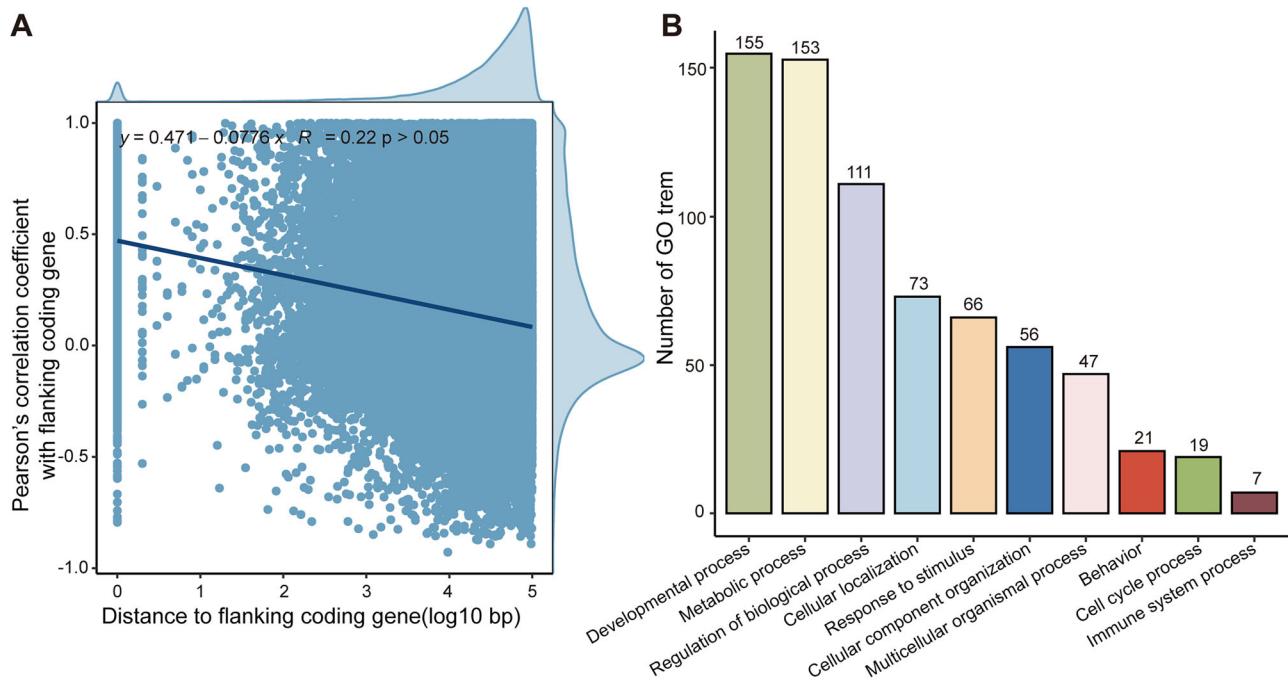


Fig. 5 | Co-expression of derived lncRNAs with adjacent protein-coding genes. A Distance-correlation analysis between lncRNA-mRNA pairs ($n = 83,615$), showing relationship between physical genomic distance (x-axis) and expression correlation ($R = 0.22$, $P > 0.05$, y-axis). Scatter plot includes trend line and

confidence intervals. B Comprehensive biological process enrichment analysis (derived lncRNAs with their adjacent protein-coding genes). Different colors represent the number of biological processes.

species. This analysis grouped genes with similar expression patterns into distinct modules: midnightblue in *M. sexta* (*Msex-lnc-001918*, $n = 540$), yellow in *Heliconius melpomene* (*Hmel006382.1*, $n = 1580$), black in *H. armigera* (*Harm008025.1*, $n = 1303$) and skyblue in *Pieris rapae* (*Prap015593.1*, $n = 326$). Functional enrichment analysis of these modules revealed that they are involved in energy metabolism and organelle organization, mainly related to spermatogenesis, including cilium assembly, sperm axoneme assembly, and sperm motility (Fig. 6C–F). The consistent enrichment patterns across different species suggest that *Msex-lnc-001918* may participate in similar biological processes as its syntenic genes, despite substantial structural modifications.

Discussion

Transcriptome studies have revealed pervasive transcription of complex genomes, and lncRNAs occupy a large fraction of these genomes, although they were previously considered “junk” transcripts³⁴. Increasing numbers of functional studies on lncRNAs^{35,36} are revealing their importance, however, limited data are available for insects. Our comprehensive analysis across 115 insect species offers insights into the genomic features, evolutionary dynamics, and potential functional implications of insect lncRNAs, with a particular focus on those derived from protein-coding genes.

To ensure annotation accuracy and biological relevance across diverse taxa, we limited our analysis to multi-exonic lncRNAs. While single-exonic lncRNAs are numerically abundant, they are especially prone to transcriptional noise, incomplete splicing, or read-through artifacts, making them difficult to distinguish from nonfunctional transcripts^{13,37}. Multi-exonic lncRNAs, by contrast, are more likely to be genuine regulatory elements, as their splicing structure provides stronger transcriptional evidence. Nevertheless, we acknowledge that some functional single-exonic lncRNAs may have been excluded by this criterion, and future studies using strand-specific or long-read sequencing technologies should revisit this question more thoroughly.

The number of identified lncRNAs varied markedly among species and was strongly influenced by genome size and sample size. Our saturation analysis indicated that around 500 transcriptomes are required to identify

most lncRNAs within a species. However, we emphasize that this is an empirical estimate derived from high-quality species, and that sample diversity, depth, and quality collectively influence lncRNA detection. One key insight is that the number of lncRNAs is strongly influenced by reference genome size, suggesting that larger genomes may harbor more complex regulatory networks involving lncRNAs. In humans, TEs overlap more than two thirds of mature lncRNA transcripts and account for 30% of their sequence length. Consistently, we found a significant correlation between lncRNA counts and transposon overlap. The correlation between lncRNA abundance and TE content is consistent with findings in mammals²³, indicating conserved mechanisms of lncRNA evolution across diverse taxa.

The reconstruction of orthologous families revealed that over 50% of lncRNA orthogroups were species-specific, compared with only 33% of protein-coding gene orthogroups. This is consistent with previous studies showing that lncRNAs lack sequence conservations^{38–41} and reinforces the view that lncRNAs undergo rapid evolutionary turnover^{42,43}. However, our analysis also demonstrated that some lncRNAs maintain conserved synteny relationships with protein-coding genes, despite lacking detectable primary-sequence conservation. This finding suggests that synteny may serve as a crucial framework for understanding the evolutionary trajectories of lncRNAs, particularly in insects where their deep evolutionary history (>400 million years) and low genomic synteny present unique challenges.

Derived lncRNAs are likely to have evolved from ancestral protein-coding genes. This aligns with emerging studies showing that pseudogenization of coding genes can serve as a substrate for regulatory innovation through lncRNAs^{16,17,44}. In this context, derived lncRNAs may represent a functional subset of processed pseudogene transcripts. Analysis of the differences between derived lncRNAs and their syntenic protein-coding genes reveals significant structural and sequence evolution. While some derived lncRNAs retain considerable sequence homology to their syntenic counterparts, many have diverged significantly, reflecting the evolutionary pressures and functional shifts that accompany the transition from coding to noncoding roles. As these syntenic genes evolve into derived lncRNAs, a consistent decrease in transcript length is observed. This trend may be indicative of the functional shift that occurs when a gene loses its coding

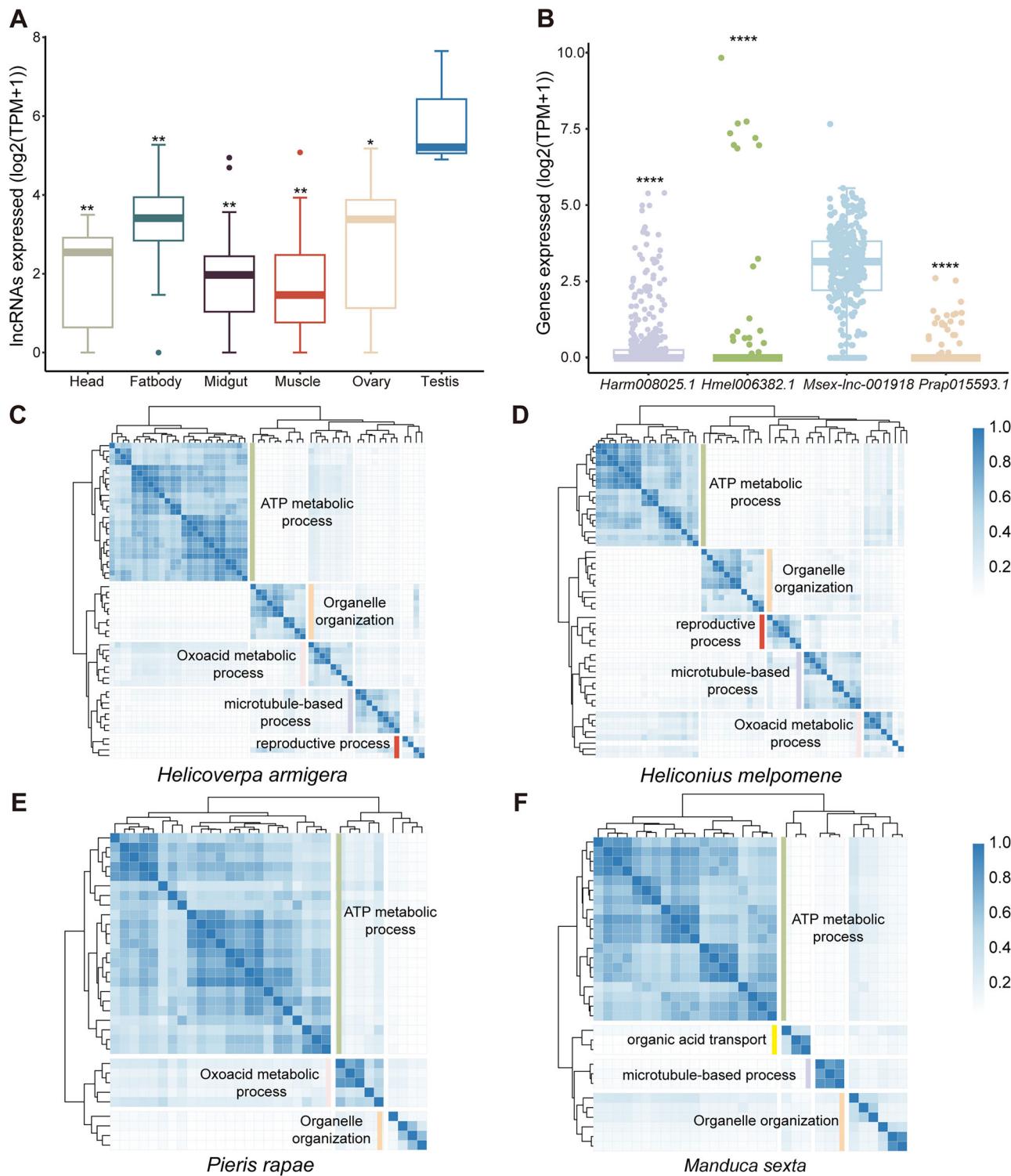


Fig. 6 | Detailed functional analysis of the lepidopteran-specific derived lncRNA Msex-lnc-001918. **A** Tissue-specific expression profile ($\log_2(\text{TPM} + 1)$) of *Msex-lnc-001918* across multiple tissues. The expression level in testis was significantly higher than other tissues. (Brain = 19, fatbody = 193, muscle = 28, midgut = 19, ovary = 15, testes = 3, * $P < 0.05$, ** $P < 0.01$, and *** $P < 0.001$, Wilcoxon's test). **B** Average expression levels ($\log_2(\text{TPM} + 1)$) for *Harm008025.1* ($n = 526$), *Hmel006382.1* ($n = 151$), *Msex-lnc-001918* ($n = 334$), and *Prap015593.1* ($n = 148$). Box plots show

distribution of expression levels across samples, with *Msex-lnc-001918* displaying significantly higher expression ($P < 0.0001$, Wilcoxon's test). **C–F** Hierarchical clustering analysis of enriched biological processes across related genes. Heatmap shows process similarity scores, with blue indicating high similarity and white indicating low similarity. Dendrogram demonstrates clustering of functionally related processes. Error bars represent standard deviation unless otherwise indicated.

potential. Beyond overall transcript shortening, our analyses revealed that derived lncRNAs also contain fewer exons and introns, and these structural components are markedly shorter than those of their syntenic genes. This pattern suggests a pronounced simplification of gene structure during the

transition from coding to noncoding states. We propose that the loss of protein-coding function reduces selective pressure to maintain complex exon–intron structures, allowing sequence deletions and rearrangements to accumulate. At the same time, the resulting shorter and simpler transcripts

may be more transcriptionally efficient and better suited for regulatory roles. This structural simplification leads to an increased tolerance for TE insertions, which can provide regulatory motifs and promote the functional diversification of derived lncRNAs.

TE insertions can lead protein-coding genes to lose their coding potential, thereby causing loss of function¹³. Only a small fraction of syntenic gene sequences overlapped with TEs, but this fraction increased significantly as these sequences transitioned into lncRNAs. These patterns suggest that TEs may be associated with the regulatory remodeling of lncRNAs by introducing novel sequence elements that could potentially influence gene expression. The preferential overlap of TEs with intronic regions in syntenic protein-coding genes, followed by an increase in exonic TE content in derived lncRNAs, reflects a possible structural transition during evolution. While these observations indicate a potential link between TE insertion and regulatory complexity, further experimental evidence is needed to clarify the direct functional impact of TEs on lncRNA regulation. Interestingly, while DNA and LTR TEs showed a high percentage of overlap across all gene types, LINEs exhibited a distinct increase in their overlap with exons in derived lncRNAs compared to syntenic genes. This suggests that LINEs may have a unique role in shaping the structure and function of derived lncRNAs, despite comprising only a small proportion (12.28%) of all TEs in these transcripts. The increased presence of LINEs in derived lncRNAs could be indicative of their involvement in regulatory processes, possibly through the modulation of gene expression or the introduction of novel regulatory sequences. These findings highlight the dynamic relationship between TEs and derived lncRNAs during evolution, particularly in the context of regulatory element acquisition and functional adaptation.

The functional prediction of derived lncRNAs revealed their potential roles in *cis* regulation^{33,45,46}, influencing the expression of neighboring protein-coding genes. Moreover, lncRNAs often interact with TFs and miRNAs to participate in regulatory networks^{47–50}. Our findings showed that derived lncRNAs are associated with more adjacent protein-coding genes, increased miRNA-binding sites, and a higher percentage of overlap with TFs (Fig. 4A, C, D), which suggests that derived lncRNAs may play critical roles in coordinating the expression of neighboring genes and be involved in post-transcriptional regulation, thereby influencing broader regulatory networks. While traditional lncRNAs show low expression levels^{27,51–53}, derived lncRNAs exhibited significantly higher expression than both conventional lncRNAs and protein-coding genes. Their elevated expression, coupled with broader tissue distribution, suggests an evolutionary trend toward active regulatory roles. The investigation into the regulatory roles of derived lncRNAs highlights their capacity to act in *cis*, influencing the expression of neighboring protein-coding genes³³. Interestingly, an inverse relationship between regulatory influence and genomic distance aligns with known distance-dependent regulatory mechanisms, while suggesting potential long-range effects beyond traditional proximity constraints³⁷. GO analysis of lncRNA-mRNA pairs highlighted a range of biological processes, with notable enrichment in development, biological regulation, and metabolic terms. The identification of these functional categories provides a valuable framework for future investigations. These characteristics indicate that derived lncRNAs may compensate for or complement the functions of their syntenic genes through diverse regulatory mechanisms, including both proximal and distal regulation.

Focusing on the specific derived lncRNA *Msex-lnc-001918* in *Manduca sexta*, our analysis revealed a testis-specific expression pattern, with higher expression than its syntenic protein-coding homologs. Its structural features, including TE insertions, and its involvement in spermatogenesis-related co-expression modules, suggest that *Msex-lnc-001918* may retain ancestral regulatory functions while adopting tissue-specific roles. The conservation of associated biological processes across Lepidoptera species further supports functional continuity despite structural divergence. However, the functional role of *Msex-lnc-001918* remains to be experimentally validated. In future work, we aim to collaborate with international laboratories to facilitate the *in vivo* validation of derived lncRNA functions.

In summary, this study provides a comprehensive analysis of lncRNA evolution and function across 115 insect species. We identified 647,691 lncRNAs and demonstrated that lncRNA abundance is significantly correlated with genome size and TE content. Through synteny analysis, we discovered 4806 derived lncRNAs that are likely to have evolved from protein-coding genes. These derived lncRNAs exhibit distinct features including higher expression levels, broader tissue distribution, and increased regulatory complexity compared with conventional lncRNAs. The case study of *Msex-lnc-001918* in *M. sexta* demonstrates that derived lncRNAs can retain functional relationships with their ancestral genes. However, these findings should be interpreted with caution, given the limitations of RNA-seq data quality and the need for experimental validation. Future studies that integrate computational approaches with functional experiments will further advance our understanding of the biological significance of derived lncRNAs in insect evolution and adaptation.

Methods

Data collection and identification of lncRNAs

We used a dataset consisting of 30,840 RNA-seq libraries (excluding *D. melanogaster*), representing 115 species from 10 orders, covering 14 major tissues and 10 developmental stages. All the data for each sample are downloaded from NCBI²⁰. LncRNAs were annotated by previous methods³⁹. Sequencing raw data of each sample were filtered using fastp⁵⁴ (v0.23.2); then transcripts were assembled with StringTie³⁵ (v2.1.4), based on reads alignments with HISAT2⁵⁶ (v2.2.1). Further, the assemblies of same species were merged with StringTie (v2.1.4) using the command: stringtie --merge -p 12 -G <species. gff3> -o <merge. gtf> <merged. list>. Next, we removed transcripts from each species which overlapped with protein-coding genes on the same strand and were less than 200 nucleotides in length. Furthermore, the coding potential of the candidate transcripts was estimated using FeeLnc²¹ (v0.2.1) and CPC2²² (v1.0.1). For FeeLnc, due to the lack of a known lncRNA datasets, we used the protein-coding genes as training dataset with the “shuffle” mode. CPC2 was then used to screen FeeLnc’s results, retaining the transcripts labeled as “noncoding”. Finally, we used blastn⁵⁷ (v2.11.0) to compare the transcripts against known insect rRNAs in the NCBI database, removing those with E ≤ 10 – 5 and identity ≥ 70%. The remaining genes were defined as lncRNAs (Fig. S1A; Supplementary Data 3). These lncRNAs were be classified into six types based on their location relative to adjacent protein-coding genes, including overlapping, containing, nested, divergent, convergent, and same strand.

Construction of homologous lncRNA families and positional conservation analysis of lncRNAs

We used Orthofinder²⁶ (v2.5.4) to investigate homologous based on sequence similarity: orthofinder -f <all_species_lncRNA.gtf> -t 2 -S blast -d. Positional conservation analysis were carried out using previously established methods²⁹. In brief, the adjacent protein-coding genes within 100 kb of lncRNAs were used to assess synteny among the species. If there was at least one conserved protein-coding neighbor (identified by Orthofinder) with the same orientation in the two analyzed species, we considered these lncRNA pairs as positionally conserved due to the syntenic evidence.

Identification of syntenic genes and derived lncRNAs pairs

Firstly, we identified syntenic modules by performing whole-genome alignment between different species’ genome assemblies was performed using the nucmer script from MUMmer³⁰ (v4.0.0): nucmer -maxgap = 500 -mincluster = 100 -prefix = <sp1sp2> <sp1. genome. fa> <sp2. genome. fa> -t 4. Then, we identified the adjacent (100 kb) protein-coding genes and the genome alignment areas for lncRNAs and protein-coding genes. We named these groups “Upstream” and “Downstream” according to their locations. Next, homologous gene pairs were identified if they meet three conditions: (i) same orthogroup (identified by orthofinder) (ii) same relative orientation and (iii) same strand. Finally, for a protein-coding gene and lncRNA potentially syntenic if there were homologous genes on one side

and syntenic genome alignment chain on the other side. LncRNA was defined derived lncRNA and the syntenic protein-coding gene were referred to as syntenic gene (Fig. S4).

Analysis of insect TEs

Repeat calling was performed using Extensive de-novo TE Annotator⁵⁸ (EDTA, v2.0.0) pipeline, which combines LTR_FINDER⁵⁹, LTR_retriever⁶⁰, and RepeatModeler⁶¹ among others, for each species. The query genome fasta file was used as input for EDTA, as well as a cds fasta file. In the output file, these transposons were classified into following types: DNA, LTR, LINE/SINE, and unknown. We then used bedtools⁶² intersect to calculate the proportion of each lncRNA type and protein-coding genes that overlap with different classes of repeats.

Estimation of miRNA binding sites and TF binding at promoters

Mature miRNAs and miRNA precursors of each species were download from miRBase database⁶³. We then utilized the miRanda³² (v3.3) software to predict potential miRNA target sites within the studied lncRNA. Specifically, we input the mature miRNA sequences along with the lncRNA sequences into the miRanda program. Based on the output file, we calculate the number of binding sites between lncRNAs and miRNAs.

For transcription factor (TF) annotation, we used the AnimalTFDB 4.0 database⁶⁴. First, we downloaded the Hidden Markov Model (HMM) profiles for the TF families of the fruit fly. Second, we employed the hmmsearch program from the HMMER package to search all protein sequences of each species against the downloaded HMM profiles in order to predict the TFs.

We defined the promoter regions as 2000 bp upstream to 1000 bp downstream of the start of the first exon³⁷. Subsequently, bedtools⁶² (v.2.31.1) intersect was used to intersect predicted TF-binding sites with our defined promoter regions, confirming the presence of TF-binding sites within these areas.

Gene expression estimation and tissue-specific index

We estimated Transcripts Per Million (TPM) values using StringTie⁵⁵ (v2.1.4) to quantify the expression levels of each lncRNA and protein-coding gene. To evaluate tissue specificity, we calculated the specificity metric τ (Tau), which measures the degree of tissue-specific gene expression[65]^{54,65}. The metric was derived from previously established methods. TPM values were first averaged across the same tissues. We then considered the maximum expression value across all tissues (\exp_{\max}) for each gene, the expression value in each specific tissue (\exp_i), and the total number of tissues analyzed (n). Specifically, the Tissue-Tau for a given gene was calculated using the formula:

$$\frac{\sum_{i=1}^n \left(1 - \frac{\exp_i}{\exp_{\max}}\right)}{n - 1}$$

This value ranges from 0 for housekeeping genes, which exhibit broad expression across tissues, to 1 for tissue-specific genes, indicating a high degree of tissue specificity.

Co-expression with adjacent coding genes

Bedtools⁶² (v.2.31.1) window was used to identify adjacent protein-coding gene within 100 kb. Next, we estimated Pearson's expression correlation between lncRNA-mRNA pairs using all the RNA-seq libraries in our dataset. Finally, we identified candidate *cis*-coexpressed lncRNA-mRNA pairs by selecting those with correlation coefficients greater than 0.5 that were close to the lncRNA.

Reconstruction the co-expression network

Weighted gene co-expression networks were constructed to cluster genes with similar expression patterns using the WGCNA⁶⁶ R package. Initially, we determined the appropriate soft power using the powerEstimate function. We then set minModuleSize = 30 and deepSplit = 2 to construct cutree. Next, we merged similar modules with a cutHeight = 0.25. Finally, protein-

coding genes and lncRNAs with strong connectivity (weight score > 0.1) from each module were filtered.

Gene ontology enrichment analysis

For the lncRNA-mRNA pairs or identified modules in each species, the clusterProfiler⁶⁷ (v.4.6.2) R package was employed for all Gene Ontology (GO) enrichments analysis for its associated protein-coding genes. We focused on "Biological Process" category for further analysis. We utilized the GOSemSim⁶⁸ (2.24.0) package to assess the semantic similarity among the enriched GO terms, which computes GO term similarities based on *org.Dm.eg.db* reference database. Additionally, we visualized the results using the pheatmap (v.1.0.12) package, which facilitated the classification and annotation of the functional roles of the derived lncRNAs in each species.

Statistical analysis and visualization

All statistical analyses were performed using the R packages, specifically utilizing the dplyr (v.1.1.4), stringr (v.1.5.1), and tidyverse (v.1.3.0) packages. All plots were generated in R using the ggplot2 (v.3.5.0), ggpunbr (v.0.6.0), reshape2 (v.1.4.4), and pheatmap (v.1.0.12). Data were analyzed using non-parametric tests due to non-normal distribution patterns of most variables. Differences between groups were evaluated using the Wilcoxon's test. Correlation coefficients were calculated using Pearson's rank correlation method. All analyses were based on biologically independent samples, with sample size (n) defined as the number of independent biological replicates or distinct species examined, as detailed in the figure legends. Each statistical test, sample size, and measure of variation are indicated in the respective figure legends.

Data availability

All transcriptome datasets used in this study were obtained from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA). The accession numbers for all datasets are listed in the run_data section of Supplementary Data 3. No new sequencing data were generated in this work. Large processed data generated during this study, including source data of Fig. 3B-D, Fig. 4A, B, E and F and Fig. S6A-D have been deposited in Figshare and are publicly available at: <https://doi.org/10.6084/m9.figshare.30464195>. Source Data can be found in Supplementary Data 1, 2. All other data supporting the findings of this study are available within the article and its Supplementary Information files.

Received: 28 April 2025; Accepted: 27 November 2025;

Published online: 11 December 2025

References

1. Peña-Kairath, C. et al. Insect pollination in deep time. *Trends Ecol. Evol.* **38**, 749–759 (2023).
2. Outhwaite, C. L., McCann, P. & Newbold, T. Agriculture and climate change are reshaping insect biodiversity worldwide. *Nature* **605**, 97–102 (2022).
3. Shaw, W. R. & Catteruccia, F. Vector biology meets disease control: using basic research to fight vector-borne diseases. *Nat. Microbiol.* **4**, 20–34 (2019).
4. Cerritos, R. Insects as food: an ecological, social and economical approach. *CABI Rev* **2009**, 1–10 (2009).
5. Wen, K. J. et al. Critical roles of long noncoding RNAs in *Drosophila* spermatogenesis. *Genome Res.* **26**, 1233–1244 (2016).
6. Maeda, R. K. et al. The lncRNA male-specific abdominal plays a critical role in *Drosophila* accessory gland development and Male fertility. *PLoS Genet.* **14**, e1007519 (2018).
7. Deng, X. X. & Meller, V. H. roX RNAs are required for increased expression of X-linked genes in *Drosophila melanogaster* males. *Genetics* **174**, 1859–1866 (2006).
8. Mulvey, B. B., Olcese, U., Cabrera, J. R. & Horabin, J. I. An interactive network of long non-coding RNAs facilitates the *Drosophila* sex

determination decision. *Biochim. Biophys. Acta* **1839**, 773–784 (2014).

- Chen, Y. B., Dougan, K. E., Nguyen, Q., Bhattacharya, D. & Chan, C. X. Genome-wide transcriptome analysis reveals the diversity and function of long non-coding RNAs in dinoflagellates. *NAR Genom. Bioinf.* **6**, lqae016 (2024).
- Lw, M. et al. Genome-wide identification of long non-coding RNAs (lncRNAs) associated with malathion resistance in *Bactrocera dorsalis*. *Pest Manag. Sci.* **77**, 2292–2301 (2021).
- T, P. et al. Functional investigation of lncRNAs and target cytochrome P450 genes related to spirotetramat resistance in *Aphis gossypii* glover. *Pest Manag. Sci.* **78**, 1982–1991 (2022).
- H, Z. et al. *Drosophila* relish activating lncRNA-CR33942 transcription facilitates antimicrobial peptide expression in *imd* innate immune response. *Front. Immunol.* **13**, 905899 (2022).
- Ulitsky, I. Evolution to the rescue: using comparative genomics to understand long non-coding RNAs. *Nat. Rev. Genet.* **17**, 601–614 (2016).
- Hezroni, H. et al. A subset of conserved mammalian long non-coding RNAs are fossils of ancestral protein-coding genes. *Genome Biol.* **18**, 162 (2017).
- Agostini, F., Zagalak, J., Attig, J., Ule, J. & Luscombe, N. M. Intergenic RNA mainly derives from nascent transcripts of known genes. *Genome Biol.* **22**, 136 (2021).
- Wen, Z.-Y., Kang, Y.-J., Ke, L., Yang, D.-C. & Gao, G. Genome-wide identification of gene loss events suggests loss relics as a potential source of functional lncRNAs in humans. *Mol. Biol. Evol.* **40**, msad103 (2023).
- Duret, L., Chureau, C., Samain, S., Weissenbach, J. & Avner, P. The *xist* RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. *Science* **312**, 1653–1655 (2006).
- Romito, A. & Rougeulle, C. Origin and evolution of the long non-coding genes in the X-inactivation center. *Biochimie* **93**, 1935–1942 (2011).
- Mei, Y. et al. InsectBase 2.0: a comprehensive gene resource for insects. *Nucleic Acids Res.* **50**, D1040–D1045 (2022).
- Sayers, E. W. et al. Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **50**, D20–D26 (2021).
- Wucher, V. et al. FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome. *Nucleic Acids Res.* **45**, e57 (2017).
- Kang, Y.-J. et al. CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Res.* **45**, W12–W16 (2017).
- Kapusta, A. et al. Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS Genet.* **9**, e1003470 (2013).
- Zhao, T. et al. LncRNAs in polyploid cotton interspecific hybrids are derived from transposon neofunctionalization. *Genome Biol.* **19**, 195 (2018).
- Di Marsico, M., Paytuví Gallart, A., Sanseverino, W. & Aiese Cigliano, R. Greene 2.0: a comprehensive database of plant long non-coding RNAs. *Nucleic Acids Res.* **50**, D1442–D1447 (2022).
- Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
- Deng, P. C., Liu, S., Nie, X. J., Weining, S. & Wu, L. Conservation analysis of long non-coding RNAs in plants. *Sci. China Life Sci.* **61**, 190–198 (2018).
- Necsulea, A. et al. The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* **505**, 635–640 (2014).
- Ulitsky, I., Shkumatava, A., Jan, C. H., Sive, H. & Bartel, D. P. Conserved function of lncRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell* **147**, 1537–1550 (2011).
- Marçais, G. et al. MUMmer4: a fast and versatile genome alignment system. *PLoS Comput. Biol.* **14**, e1005944 (2018).
- Zhao, Y. et al. Identification and analysis of mouse non-coding RNA using transcriptome data. *Sci. China Life Sci.* **59**, 589–603 (2016).
- Enright, A. et al. MicroRNA targets in *Drosophila*. *Genome Biol.* **4**, P8 (2003).
- Yan, P. X., Luo, S., Lu, J. Y. & Shen, X. H. *Cis*- and *trans*-acting lncRNAs in pluripotency and reprogramming. *Curr. Opin. Genet. Dev.* **46**, 170–178 (2017).
- Palazzo, A. F. & Koonin, E. V. Functional long non-coding RNAs evolve from junk transcripts. *Cell* **183**, 1151–1161 (2020).
- Carrieri, C. et al. Long non-coding antisense RNA controls *Uchl1* translation through an embedded SINEB2 repeat. *Nature* **491**, 454–457 (2012).
- Chen, L.-L. & Kim, V. N. Small and long non-coding RNAs: past, present, and future. *Cell* **187**, 6451–6485 (2024).
- Sarropoulos, I., Marin, R., Cardoso-Moreira, M. & Kaessmann, H. Developmental dynamics of lncRNAs across mammalian organs and species. *Nature* **571**, 510–514 (2019).
- Johnsson, P., Lipovich, L., Grandér, D. & Morris, K. V. Evolutionary conservation of long non-coding RNAs; sequence, structure, function. *Biochim. Biophys. Acta Gen. Subj.* **1840**, 1063–1071 (2014).
- Hovhannisan, H. & Gabaldón, T. The long non-coding RNA landscape of candida yeast pathogens. *Nat. Commun.* **12**, 7317 (2021).
- Li, T. et al. Long non-coding RNA derived from lncRNA–mRNA Co-expression networks modulates the locust phase change. *Genom. Proteom. Bioinform.* **18**, 664–678 (2020).
- Washietl, S., Kellis, M. & Garber, M. Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. *Genome Res.* **24**, 616–628 (2014).
- Quinn, J. J. et al. Rapid evolutionary turnover underlies conserved lncRNA–genome interactions. *Genes Dev.* **30**, 191–207 (2016).
- Paralkar, V. R. et al. Lineage and species-specific long noncoding RNAs during erythro-megakaryocytic development. *Blood* **123**, 1927–1937 (2014).
- Milligan, M. J. & Lipovich, L. Pseudogene-derived lncRNAs: emerging regulators of gene expression. *Front. Genet.* **5**, 476 (2014).
- Wu, Z. M. et al. Regulation of lncRNA expression. *Cell. Mol. Biol. Lett.* **19**, 561–575 (2014).
- Gil, N. & Ulitsky, I. Regulation of gene expression by *cis*-acting long non-coding RNAs. *Nat. Rev. Genet.* **21**, 102–117 (2020).
- Ye, S. et al. Bioinformatics method to predict two regulation mechanism: TF-miRNA–mRNA and lncRNA–miRNA–mRNA in pancreatic cancer. *Cell Biochem. Biophys.* **70**, 1849–1858 (2014).
- Hu, B. et al. The mRNA-miRNA-lncRNA regulatory network and factors associated with prognosis prediction of hepatocellular carcinoma. *Genom. Proteom. Bioinform.* **19**, 913–925 (2021).
- Zhu, B., Li, L. H., Wei, R., Liang, P. & Gao, X. W. Regulation of GSTu1-mediated insecticide resistance in *Plutella xylostella* by miRNA and lncRNA. *PLoS Genet.* **17**, e1009888 (2021).
- Chen, X. et al. Integration of lncRNA–miRNA–mRNA reveals novel insights into oviposition regulation in honey bees. *PeerJ* **5**, e3881 (2017).
- Lopez-Ezquerro, A., Harrison, M. C. & Bornberg-Bauer, E. Comparative analysis of lincRNA in insect species. *BMC Evol. Biol.* **17**, 155 (2017).
- Darbellay, F. & Necsulea, A. Comparative transcriptomics analyses across species, organs, and developmental stages reveal functionally constrained lncRNAs. *Mol. Biol. Evol.* **37**, 240–259 (2020).
- Sang, S. Y. et al. Data integration and evolutionary analysis of long non-coding RNAs in 25 flowering plants. *BMC Genom.* **22**, 739 (2021).
- Chen, S. F., Zhou, Y. Q., Chen, Y. R. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
- Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).

56. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
57. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinform.* **10**, 421 (2009).
58. Ou, S. J. et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275 (2019).
59. Ou, S. J. & Jiang, N. LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. *Mobile DNA* **10**, 48 (2019).
60. Ou, S. J. & Jiang, N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* **176**, 1410–1422 (2018).
61. Flynn, J. M. et al. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA* **117**, 9451–9457 (2020).
62. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
63. Kozomara, A., Birgaoanu, M. & Griffiths-Jones, S. miRBase: from microRNA sequences to function. *Nucleic Acids Res.* **47**, D155–D162 (2019).
64. Shen, W.-K. et al. AnimalTFDB 4.0: a comprehensive animal transcription factor database updated with variation and expression annotations. *Nucleic Acids Res.* **51**, D39–D45 (2023).
65. Yanai, I. et al. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21**, 650–659 (2005).
66. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinform.* **9**, 559 (2008).
67. Yu, G. C., Wang, L.-G., Han, Y. Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).
68. Yu, G. C. et al. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* **26**, 976–978 (2010).

Acknowledgements

This work was supported by National Key Research and Development Program of China (2021YFD1400100, 2022YFD1401600), Yunnan Province Agricultural Basic Research Joint Project (202301BD070001-045), Provincial Natural Science Foundation of China (LZ23C140002) and 2024 Yangtze River Delta Science and Technology Innovation Community Joint Research (Basic Research) Project (2024CSJZN0900).

Author contributions

D.J. contributed in study set-up, data analysis and wrote the manuscript. Y.M. and H.Z. assisted with data analysis and manuscript revision. Z.W.,

S.J., and H.C. were responsible for data collection and downloading. H.D. and H.C. revised the manuscript. F.L. and K.H. contributed to experimental design, provided funding support, and revised the manuscript. Y.L. provided funding support and revised the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s42003-025-09347-3>.

Correspondence and requests for materials should be addressed to Ying Liu or Kang He.

Peer review information *Communications Biology* thanks Wanfei Liu and the other, anonymous, reviewer for their contribution to the peer review of this work. Primary Handling Editors: John Mulley, Aylin Bircan and Christina Karlsson Rosenthal. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025