

Near complete assembly of *Drosophila melanogaster* Canton S strain genome

Received: 9 December 2024

Accepted: 14 November 2025

Published online: 03 December 2025

Yan-Nan Liu^{1,2,3,10}, Jian-Jun Gao^{1,2,3,10}, Xiao-Lin Zhuang⁴,
Dong-Dong Wu^{4,5,6,7,8,9}✉ & Yan-Bo Sun^{1,2,3,5}✉

The fruit fly *Drosophila melanogaster* is a pivotal model organism, yet its reference genome (ISO-1 strain) retains unresolved gaps in complex regions. Here, we present a near-complete telomere-to-telomere genome assembly (Dm.nT2T) of the Canton S strain (male), generated using an integrative approach combining PacBio HiFi, Oxford Nanopore ultra-long reads, and Hi-C data. This assembly spans 161.63 Mb, closes 93.28% of gaps in the current reference genome, and improves contiguity (contig N50 of 21.93 Mb). We identify centromeric and telomeric regions, yielding a more complete genome map. Comparative analysis with the ISO-1 reference identifies 7989 structural variants, including large insertions/deletions enriched in telomeric and centromeric regions. Notably, we detect SINE transposable elements, and identify 92 genes via multi-strategy annotation. Functional validation of two genes shows that *Chr3L.2449*, expressed in olfactory tissues, is critical for olfactory sensitivity. Dm.nT2T provides a high-quality resource for exploring genomic complexity, strain diversity, and gene function in *Drosophila*.

Drosophila melanogaster (fruit fly) is a small insect found worldwide, exhibiting significant genetic diversity across numerous strains¹. As a cornerstone model organism in biological research, it is widely utilized across genetics, neuroscience, species evolution, and developmental biology due to its short life cycle, ease of cultivation, and straightforward genetic manipulation^{2,3}. Its compact genome and the availability of advanced genetic tools have yielded crucial insights into gene function and regulation, and mechanisms of complex traits and diseases^{4,5}. Moreover, owing to its genetic similarities to humans, it serves as an invaluable tool for understanding human disease mechanisms^{6,7}. However, despite ongoing studies seeking to improve genome completeness of this classic model organism^{8,9}, its genome

still contains the structurally complex regions^{10,11}, constraining our capacity to fully decipher the genome's functional landscape.

The evolution of *D. melanogaster* genome assemblies mirrors advancements in sequencing technologies. Early efforts relied on Sanger sequencing and clone-based mapping, culminating in the shotgun assembly (Release 1, ~120 Mb)¹². This release contained gaps and misassemblies but provided the first comprehensive eukaryotic genome reference beyond yeast. Substantial refinements (Release 2–6) improved gap closure and extended coverage into heterochromatin, but technical limitations, particularly with short-read sequencing, left ~250 unresolved gaps in the current reference genome Release 6 (R6), predominantly in structurally complex regions^{13–15}.

¹State Key Laboratory of Vegetation Structure, Function and Construction (VegLab), Yunnan University, Kunming, China. ²Yunnan Key Laboratory of Biological Adaptation, Conservation and Utilization, School of Ecology and Environmental Science, Yunnan University, Kunming, China. ³Ministry of Education Key Laboratory for Transboundary Ecoscience of Southwest China, Yunnan University, Kunming, China. ⁴State Key Laboratory of Genetic Evolution & Animal Models, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China. ⁵Southwest United Graduate School, Kunming, China. ⁶Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming, China. ⁷National Resource Center for Non-Human Primates, Kunming Primate Research Center, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China. ⁸National Research Facility for Phenotypic & Genetic Analysis of Model Animals (Primate Facility), Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China. ⁹KIZ-CUHK Joint Laboratory of Bioresources and Molecular Research in Common Diseases, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China. ¹⁰These authors contributed equally: Yan-Nan Liu, Jian-Jun Gao. ✉e-mail: wudongdong@mail.kiz.ac.cn; sunyanbo@ynu.edu.cn

These regions predominantly localize within highly repetitive sequences and centromeric/telomeric areas¹⁵, which are essential for genome stability, exemplified by centromeres that ensure accurate chromosome segregation and telomeres that protect chromosomal ends^{16,17}. It has been established that telomeres in *D. melanogaster* are maintained not by telomerase, but through the targeted transposition of specialized retrotransposons, primarily *HeT-A*, *TART*, and *TAHRE* (collectively termed HTT elements)^{18,19}, yet the telomere length dynamics remain incompletely understood. Similarly, centromeres in *D. melanogaster*, composed of tandem satellite repeats²⁰, have eluded base-pair resolution mapping, limiting insights into their epigenetic regulation and role in genome stability.

Recent breakthroughs in long-read sequencing platforms (PacBio HiFi, Oxford Nanopore) and sophisticated assembly algorithms now enable the construction of telomere-to-telomere (T2T) assemblies for complex genomes²¹. These technologies bypass the limitations of short reads by spanning repetitive regions and scaffolding contigs into chromosomes. Landmark achievements include the T2T assembly of the human genome (CHM13), resolving inaccessible regions such as centromeres, telomeres, and segmental duplications, revealing unprecedented insights into genome structure, variation, and function²². This T2T paradigm has rapidly expanded to other key model organisms, demonstrating its transformative potential for comprehensive genomic analysis^{23,24}. Here, we present a near-

complete genome assembly (Dm.nT2T) of the *Canton S* strain (a classical wild-type line) using an integrative approach combining PacBio HiFi, ONT ultra-long reads, Illumina short reads, and Hi-C data. This assembly resolves critical gaps in the current reference, identifies the centromeric and telomeric regions, and uncovers additional genomic elements. By comparing the Dm.nT2T with the *ISO-1* reference strain, we further highlight intraspecific genomic variation, providing a resource to enhance understanding of *D. melanogaster* biology, evolution, and complex trait genetics.

Results

A near-complete genome assembly of *D. melanogaster*

To construct a higher-quality genome assembly of *D. melanogaster*, we utilized an integrative approach that combined deeply sequenced data from multiple sequencing platforms. These included 17.36 Gb (~107.39×) of PacBio HiFi Circular Consensus Sequencing (CCS) reads, 156.25 Gb (~966.74×) of Oxford Nanopore Technology (ONT) ultra-long (UL) reads, 30.84 Gb (~149.37×) of short paired-end reads, and 24.14 Gb (~190.8×) of High-throughput Chromatin conformation capture (Hi-C) sequencing reads (Fig. 1a and Supplementary Table 1). The quality and length distribution of the raw long-read sequencing data confirmed the overall reliability of the input reads, with HiFi mean read length was 7141.6 bp and ONT ultra-long mean read length was 26,487.9 bp (Supplementary Fig. 1). For HiFi reads, the mean Phred

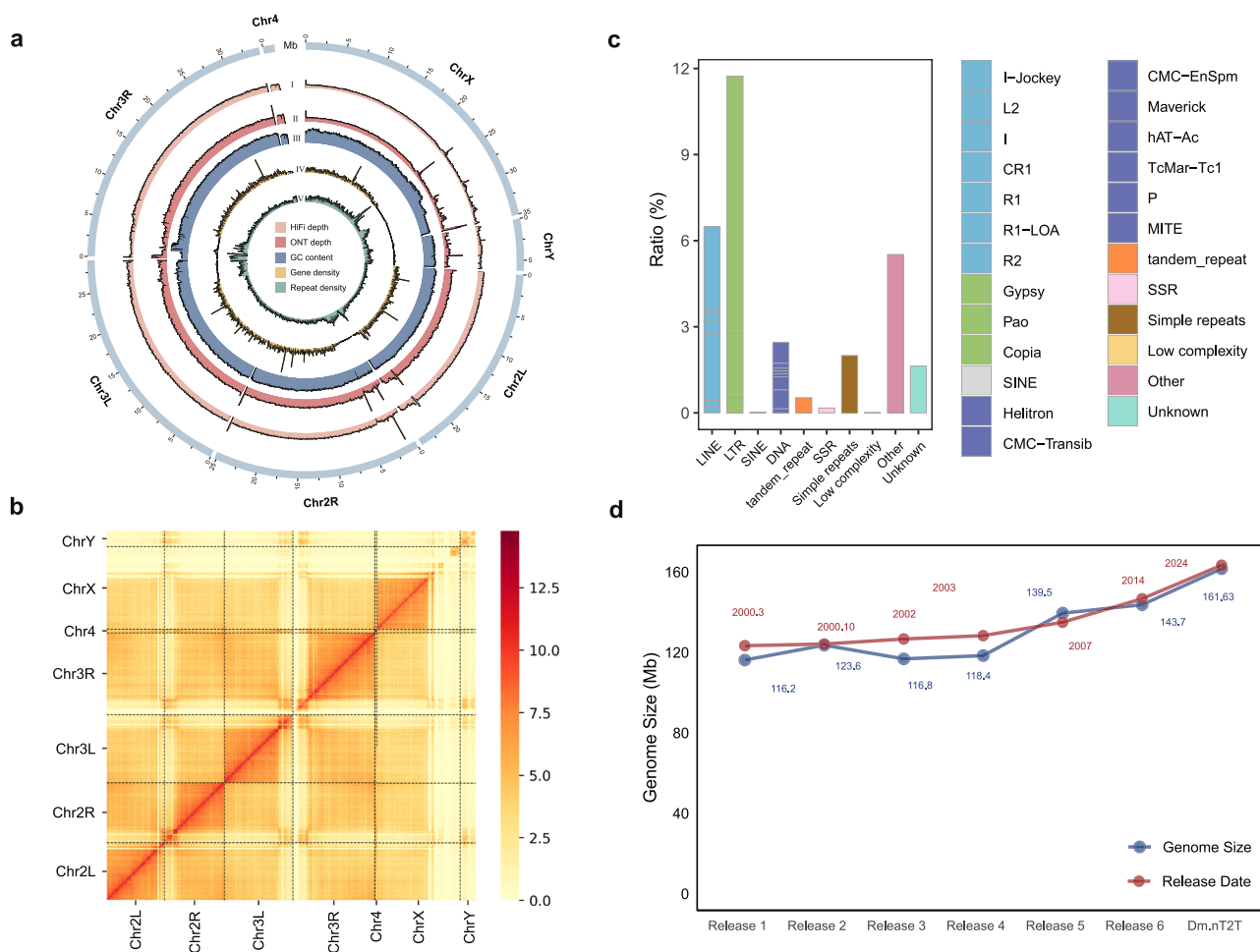


Fig. 1 | Overview of the Dm.nT2T genome assembly and annotation. a Circos plot depicting genome features at 50-kb resolution across the seven assembled chromosomes. From outer to inner ring: (I) PacBio HiFi read depths, (II) ONT UL read depths, (III) GC content, (IV) gene density, and (V) repeat density. **b** Hi-C chromatin interaction map of the Dm.nT2T genome assembly. Point colors

represent the logarithmic value of interaction strength between corresponding genomic bin pairs, with intensity increasing from yellow to black. **c** Distribution and relative abundance of repetitive sequence types within the assembly. **d** Historical release timeline and corresponding genome sizes for *D. melanogaster* reference genome versions. Source data are provided as a Source Data file.

Table 1 | Statistical comparison between the R6 and Dm.nT2T genome assembly

	Release 6 plus ISO1 MT	Dm.nT2T
Size (Mb)	143.7	161.63
Num. chromosomes	7	7
Contig N50 (Mb)	21.5	21.93
Gap number	268	18
Num. genes	17,894	17,898

quality score was Q28.9 (equivalent to a base error rate of 0.13%), while the median quality score reached Q37.1 (0.02% error rate), demonstrating high accuracy for the majority of sequencing reads. The ONT reads displayed an N50 value of 48,181 bp, indicating a significant fraction of ultra-long reads. The average read quality score was Q14.4 (corresponding to an estimated error rate of approximately 3.6%), and the median quality score was Q15.1 (error rate of approximately 3.1%), consistent with the typical performance range of current ONT technology.

Assembly with NextDenovo²⁵, followed by polishing with minimap2²⁶ and Nextpolish²⁷, yielded 30 contigs without any manual filtering. Hi-C scaffolding further clustered 25 contigs into seven chromosomes (Chr2L, Chr2R, Chr3L, Chr3R, Chr4, ChrX, and ChrY) with LACHESIS²⁸, supported by distinct diagonal interactions in Hi-C heatmaps (Fig. 1b). Following gap filling with Nextagap²⁵, the final near-complete assembly of *D. melanogaster* (termed Dm.nT2T) spans 161.63 Mb, with a contig N50 of 21.93 Mb (Table 1). Quality evaluation showed a k-mer completeness of 98.0% and a QV of 40.24 for the assembly using Merqury²⁹ (Supplementary Fig. 2). BUSCO³⁰ analysis further demonstrated a high level of completeness of genome assembly, achieving 98.8% completeness (Supplementary Table 2), showing a slight enhancement over the current reference genome (Release 6 plus ISO1 MT, R6). Additionally, we performed GC-depth analysis, which demonstrated a GC content distribution ranging from 30% to 60% with the sequencing depth concentrated around 1000×. This profile exhibited a unimodal distribution without anomalous deviations (Supplementary Fig. 3). These results indicate the absence of sequence contamination and confirm the high completeness and accuracy of the Dm.nT2T genome assembly. A comparison with previous *D. melanogaster* assemblies using long-read technologies further illustrates the progressive improvements over the R6 reference and additional advances achieved by the Dm.nT2T assembly (Supplementary Table 3).

To further evaluate assembly quality, we analyzed ribosomal DNA (rDNA) arrays and histone gene clusters, which represent known problematic regions on the X and Y chromosomes (ChrX and ChrY). Canonical rDNA sequences (18S, 5.8S, 2S, and 28S) from the R6 reference were employed as queries for BLASTn³¹ searches against the Dm.nT2T assembly (*E*-value ≤ 1e-20). This analysis showed that within Dm.nT2T, the 18S, 5.8S, and 28S rDNA sequences can be precisely localized to ChrX (Supplementary Data 1). Notably, we also detected high-confidence matches for these rDNA units on ChrY, where the presence of an rDNA cluster has long been recognized^{32–34}, but was not annotated in the R6 genome assembly, indicating a significant enhancement in rDNA resolution. Furthermore, the 18S and 28S rDNA sequences, which are restricted to unplaced contigs in the R6 genome, were successfully anchored to both ChrX and ChrY in the Dm.nT2T assembly (Supplementary Data 1), further demonstrating an improvement in rDNA array assembly, particularly concerning the sex chromosomes. Similarly, to evaluate the resolution of the histone gene clusters, 111 canonical histone gene sequences were retrieved from R6, encompassing all five major histone families (H1, H2A, H2B, H3, and H4). A stringent BLASTn search (identity ≥ 95%, *E*-value ≤ 1e-20) was performed against the Dm.nT2T assembly. Our analysis showed that all

histone genes exhibited >95% sequence identity and mapped contiguously to an extended tandem repeat region on Chr2L (coordinates: 21.4 Mb to 22.1 Mb; Supplementary Data 2). In contrast to the restricted interval of 21.4–21.5 Mb in the R6 assembly, this 0.7 Mb expansion demonstrates successful resolution of the repetitive region in the Dm.nT2T assembly. Furthermore, several histone H2B sequences originally annotated on Chr2L in R6 were found to align to a region on Chr2R (3.087–3.088 Mb) in our assembly, providing additional evidence that unassembled histone sequences have been accurately resolved in Dm.nT2T.

The repetitive sequence content in the Dm.nT2T genome assembly was estimated as 30.57%, with transposable elements (TEs) accounting for 22.79% (Supplementary Table 4). This percentage is slightly higher than that observed in the R6 genome, which has 21.7% TEs. The predominant TE classes identified comprise DNA transposons (accounting for 3.98%), Long Terminal Repeats (LTRs; 11.91%), Long Interspersed Nuclear Elements (LINEs; 6.87%), and Short Interspersed Nuclear Elements (SINEs; 0.02%) (Fig. 1c). Notably, the presence of SINE elements in our assembly is intriguing since prior research and literature do not report their existence in *D. melanogaster*³⁵.

Structural comparisons with the current reference genome

The reference genome of *D. melanogaster* has undergone continuous refinement over the past several decades, with the assembled genome size gradually increasing alongside advances in sequencing technologies. In comparison, Dm.nT2T is 17.93 Mb larger than the R6 genome (Fig. 1d and Supplementary Table 5) and closed 93.28% of R6 gaps (Table 1), representing the largest chromosomal-level genome published to date for *D. melanogaster*. Syntenic alignment revealed conserved collinearity between the Dm.nT2T and R6 genome assemblies (Fig. 2a and Supplementary Fig. 4).

Within the alignment results of the two assemblies, we observed numerous small structural variations (SVs). To more precisely detect them, we mapped HiFi reads to the R6 genome and utilized four different tools to identify SVs, including Sniffles2³⁶, SVIM³⁷, PBSV, and CuteSV³⁸. After initial results filtered, a total of 7989 SVs were finally identified, including 2862 insertions (INs), 3208 deletions (DELs), 242 duplications (DUPS), 203 inversions (INVs), and 1474 translocations (TRAs). The distribution of SVs exhibited enrichment at the telomeric regions of all autosomes, one terminus of the X chromosome (ChrX), and ubiquitously throughout the Y chromosome (ChrY). These regions were predominantly comprised of insertions and deletions (INDELs) and TRAs, with pronounced enrichment observed at the termini of Chr2R, Chr3R, and ChrX (Fig. 2b). INDELs accounted for the vast majority of these variants, with sizes ranging from 51 bp to 92,641 bp, and 79.57% of them exceed 100 bp in length (Fig. 2c). Interestingly, these SVs are most enriched in the putative centromeric regions. RepeatMasker (<http://repeatmasker.org>) analysis showed only 1% of SVs overlapped repetitive sequences, suggesting the vast majority of variations are not attributable to repetitive sequences. We also investigated their spatial relationship to assembly gaps in Dm.nT2T to assess whether these variants are indicative of potential scaffolding errors. We conducted a comparative structural variant analysis between Dm.nT2T and the R6 using SyRI³⁹. Our analysis revealed that none of the identified INVs or TRAs overlapped or even located within 10 kb of any gap region, providing robust support for the structural integrity of our scaffolding and the validity of the detected SVs.

Notably, our analysis revealed an interesting observation: the assembled Y chromosome exhibits extensive structural variations compared to the reference genome R6, with a length extension to 6.27 Mb. Furthermore, considering that Chang et al. reported a 14.6 Mb Y chromosome assembly⁴⁰, this discrepancy raises critical questions regarding the accuracy of our Y chromosome assembly. To address this issue, we further compared our ChrY (Y_Dm.nT2T) with the Chang

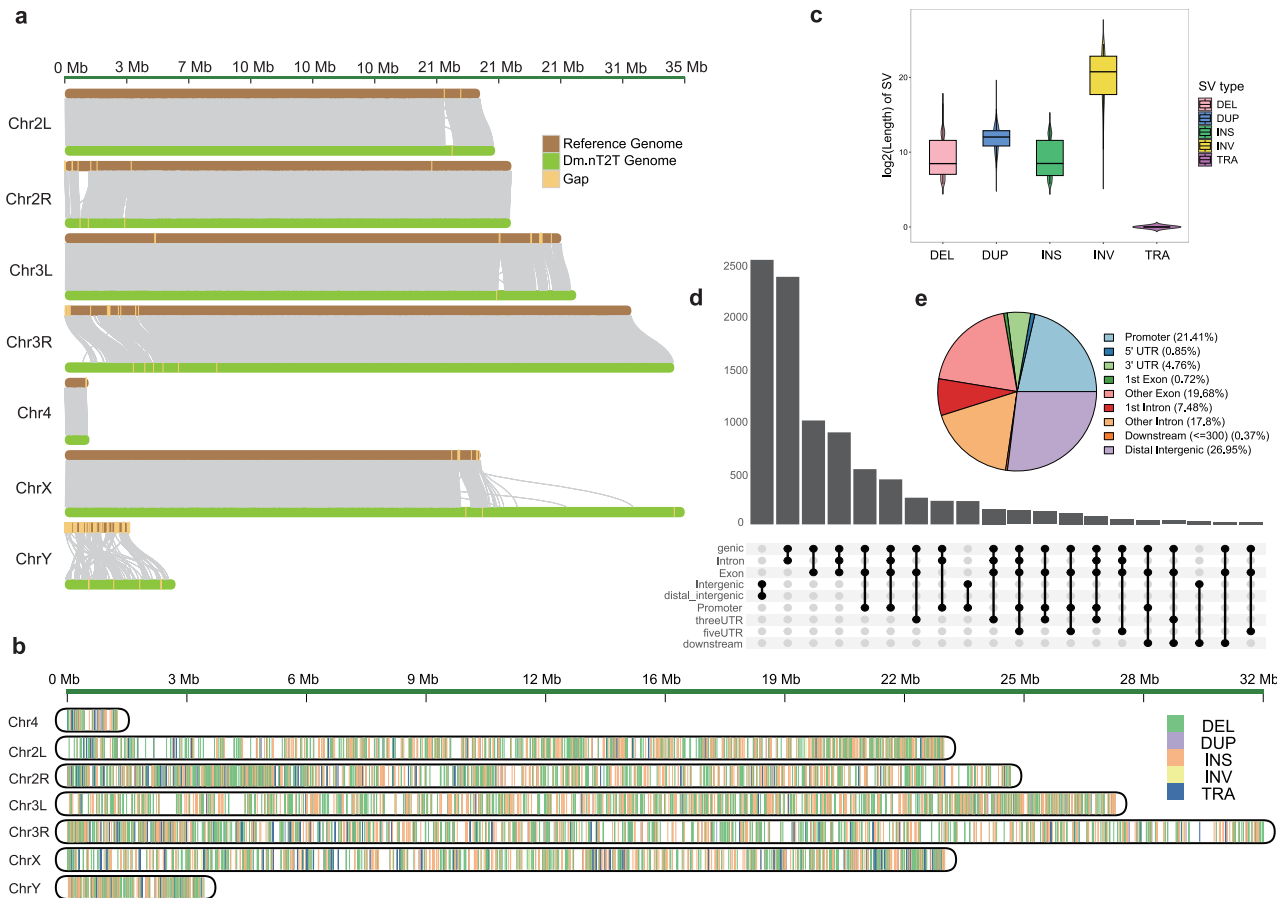


Fig. 2 | Genomic comparison and structural variants between the R6 and the Dm.nT2T genome assemblies. **a** Syntentic alignment between the R6 and Dm.nT2T genome assemblies. Gray regions denote syntentic blocks, while yellow segments indicate assembly gaps on divergent chromosomes. **b** Ideogram of density distribution of different types of SVs on chromosomes of the Dm.nT2T genome assembly (DEL Deletion, DUP Duplication, INS Insertion, INV Inversion, TRA

Translocation). **c** The logarithmic scale size of different types of SVs (DEL, $n = 3209$; DUP, $n = 242$; INS, $n = 2862$; INV, $n = 203$; TRA, $n = 1473$). **d** The Upset plot delineating the annotation of genomic features. **e** The Pie chart represents the functional annotation percentage of genomic features. Source data are provided as a Source Data file.

Table 2 | Statistical annotation of structural variants identified by ANNOVAR

Type	Number	Ratio (%)
Intronic	2838	38.20678514
Intergenic	3472	46.74205708
Exonic	310	4.173397954
UTR	239	3.217555197
Upstream	287	3.863758751
Downstream	282	3.79644588

assembly (Y_Chang). Results showed that the Y_Dm.nT2T exhibits superior contiguity: Y_Dm.nT2T is a single, fully contiguous contig, with N50/N90/L50 values equal to its full length (~6.27 Mb) (Supplementary Table 6). In contrast, the fragmentation of the Y_Chang assembly (55 contigs) and its elevated ambiguity rate (N-base frequency per 100 kbp: 171.19 compared with 6.38) reflect the inherent trade-off associated with prioritizing repetitive sequence recovery at the expense of structural coherence.

We further used ANNOVAR⁴¹ to annotate and evaluate the potential functional impacts of these SVs. The results indicated that most SVs occur in intergenic and intronic regions, accounting for 46.7% and 38.2%, respectively (Table 2). Only 4.2% of SVs are located in

the exonic region, 3.2% in the untranslated regions (UTRs), 3.9% upstream, and 3.8% downstream (within 1 kb of a gene). Given that the intronic and intergenic regions frequently harbor functionally important regulatory elements⁴², the presence of SVs in these areas might have significant implications for gene expression regulation⁴³. Gene Ontology (GO) enrichment analysis showed that genes ($n = 4587$) within 5 kb of the SVs were significantly enriched in processes related to response to hormone, mRNA metabolic process, tube morphogenesis, and protein maturation (Supplementary Fig. 5).

To investigate whether the Dm.nT2T assembly permits identification of additional genetic features that are present in the *Drosophila* genome but not represented in the reference, we performed further analysis of ATAC-seq data derived from 12 distinct *D. melanogaster* tissues, downloaded from the ENA database (Supplementary Data 3). These data were aligned to both the R6 genome and the Dm.nT2T genome assembly. Following peak calling, open chromatin regions were identified based on consensus peaks across biological replicates, yielding a total of 212,111 peaks across all 12 tissues. After filtering out peaks overlapping with those identified in the R6 assembly, we identified 9447 peaks unique to the Dm.nT2T assembly. Annotation of these open chromatin regions demonstrated that distal intergenic regions and intergenic regions constitute the primary sites of peak enrichment (Fig. 2d). Specifically, peaks within distal intergenic regions accounted for 26.95% of the total, followed by promoter regions at 21.41% (Fig. 2e).

Table 3 | The location of the telomeric region in the Dm.nT2T genome assembly

Chromosome	Left start	End	Length (bp)	Right start	End	Length (bp)
Chr2L	1	40,088	40,087	–	–	–
Chr2R	–	–	–	25,182,000	25,256,150	74,150
Chr3L	–	–	–	–	–	–
Chr3R	–	–	–	34,458,346	34,507,726	49,380
Chr4	–	–	–	–	–	–
ChrX	12	76,250	76,238	–	–	–
ChrY	2,715,543	2,759,968	44,425	5,769,889	5,784,058	14,169

Characterization of segmental duplications and population analysis

Segmental duplications (SDs), defined as genomic regions longer than 1 kb with sequence identity exceeding 90%⁴⁴, were systematically analyzed in the Dm.nT2T genome assembly. Our analysis revealed a total SD content of 15.7 Mb (~9.8% of the genome) with distinct distribution patterns: intra-chromosomal SDs dominated (14.3 Mb, 91.1%), followed by inter-chromosomal duplications (3.5 Mb, 22.3%), and complex overlapping events spanning both categories (2 Mb, 12.7%). Spatial analysis demonstrated striking centromeric enrichment of SDs (Supplementary Fig. 6), mirroring evolutionary conservation patterns observed across multiple T2T-sequenced species^{45,46}.

To elucidate the evolutionary dynamics of SDs across natural populations, we analyzed genome-wide variation patterns using resequencing data from six globally distributed *D. melanogaster* populations (including Asia, Europe & Northern Africa, Oceania & North America, Ethiopia, Western & Eastern Africa, and Southern Africa; Supplementary Data 4)¹. After rigorous quality control, 1.84 million high-confidence biallelic SNPs were retained, with 17,568 SNPs (15.79 Mb) localized to SD regions and 1.83 million SNPs (140.07 Mb) in non-SD regions. SD regions exhibited a 12-fold reduction in SNP density (1.11 SNPs/kb vs. 13.03 SNPs/kb in non-SD regions), supporting constrained mutagenesis or enhanced repair fidelity in duplicated loci. Allele frequency spectra further revealed population-specific signatures: Ethiopian and sub-Saharan African populations (excluding North Africa) showed higher SD-region allele frequencies compared to derived populations (Supplementary Fig. 7), reflecting either higher ancestral diversity retention due to larger effective population sizes or stronger purifying selection on African-lineage SDs to potentially maintain centromere-proximal repeat integrity.

Identification of telomeric and centromeric complex regions

Since telomeres function to maintain chromosomal stability and prevent DNA damage in eukaryotes⁴⁷, comprehensive mapping of telomeric regions therefore holds implications extending beyond enhancing genome completeness, facilitating a deeper understanding of telomere function. Studies have demonstrated that the telomeric regions of *D. melanogaster* are composed of three families of retrotransposons: *HeT-A*, *TART*, and *TAHRE*^{48,49}. This organization contrasts with the tandem repeats of TG-rich microsatellite sequences observed in most other species^{50,51}. Furthermore, it exhibits unique features distinguishing it even from other species within the genus⁵². To identify telomeric regions, we employed lastz (<https://github.com/lastz/lastz>) to align the transposons *HETA* (6081 bp), *TART-A* (15576 bp), *TART-B1* (10654 bp), and *TAHRE* (10463 bp) to the Dm.nT2T genome assembly. Analysis using the Integrative Genomics Viewer (IGV)⁵³ identified six telomeric regions. No telomeric regions were detected on Chr3L and Chr4. Among the identified telomeric regions, the longest is the ChrX left-arm telomere (76.24 kb), while the shortest is the ChrY right-arm telomere (14.17 kb; Table 3). We further systematically compared the average length, TE composition, and genomic

location of telomeric regions in Dm.nT2T with those in the R6 reference, using the same method for telomeric region identification as applied in Dm.nT2T. This analysis highlights substantial differences in telomere representation between the two assemblies. The total telomeric length in Dm.nT2T reaches 297 kb, markedly exceeding the 74.3 kb observed in R6, whose sequences remain unplaced contigs. On average, telomeric regions in Dm.nT2T span 49.7 kb, whereas those in R6 measure only 1.6 kb. These differences primarily reflect the more complete and accurate reconstruction of telomeric regions in the Dm.nT2T genome assembly. However, we cannot fully exclude the possibility that some variation arises from genuine strain-specific polymorphisms, which will require further comparative analyses across multiple *D. melanogaster* strains. Moreover, TE composition analysis reveals that Dm.nT2T telomeres comprise multiple TE families, including *HET-A*, *TART*, *TAHRE*, and *HETA_DS1*. In contrast, R6 telomeres contain only *HET-A* elements.

To detect centromeric regions of *D. melanogaster*, we employed a two-step method. Firstly, due to the high content of simple repeats in centromeric regions, we used Tandem Repeat Finder (TRF)⁵⁴ to identify genome regions rich in repeat sequences. Regions exhibiting a relatively condensed and tightly packed array of repeat units were deemed potential centromeric candidates, with the constituent repeat units designated as centromeric motifs. Based on reported lengths of *Drosophila* centromeric motifs⁵⁵, we selected a range of 5–500 bp to scan for tandem repeat sequences across the genome, following the pipeline described in the grape T2T genome⁵⁶. Our assembly revealed that 5-bp repeats are the most abundant monomeric unit, with 290,576 copies constituting approximately 0.9% of the total genome (Supplementary Table 7). These repeats were distributed across almost all chromosomes. Additional prevalent repeat lengths encompassed 10 bp (corresponding to the *Drosophila* *Prod* satellite), 11 bp, 8 bp, 12 bp (corresponding to the *Dodeca* satellite), 7, 6, and 359 bp (Fig. 3a). Consequently, regions enriched with these TRF-identified repeat motifs were designated as candidate centromeric regions, with 5–12 bp simple repeat sequences constituting the primary constituents of *D. melanogaster*'s candidate centromeres, wherein 5-bp units exhibit an overwhelming dominance (Fig. 3b). Next, we performed a self-alignment of the Dm.nT2T genome assembly using StainedGlass⁵⁷ to display sequence repetitiveness and alignment patterns, thereby pinpointing highly repetitive centromeric regions. This analysis allowed us to visualize candidate centromeric regions on each chromosome (Supplementary Fig. 8). By integrating the two results, we conclusively defined the centromeric regions of the Dm.nT2T genome assembly, ranging from 35.3 kb on Chr4 to 2.6 Mb on ChrX, with an average length of 920 kb (Table 4).

To further validate the accuracy of the candidate centromeric regions, we reanalyzed publicly available ChIP data for CENP-A (Supplementary Data 3), a histone variant that serves as an epigenetic marker for centromeres¹⁶. CENP-A binding regions represent the definitive benchmark for centromere identification. ChIP-seq analysis confirmed the accuracy of our candidate centromere regions. Specifically, CENP-A exhibited significant peak enrichment predominantly within

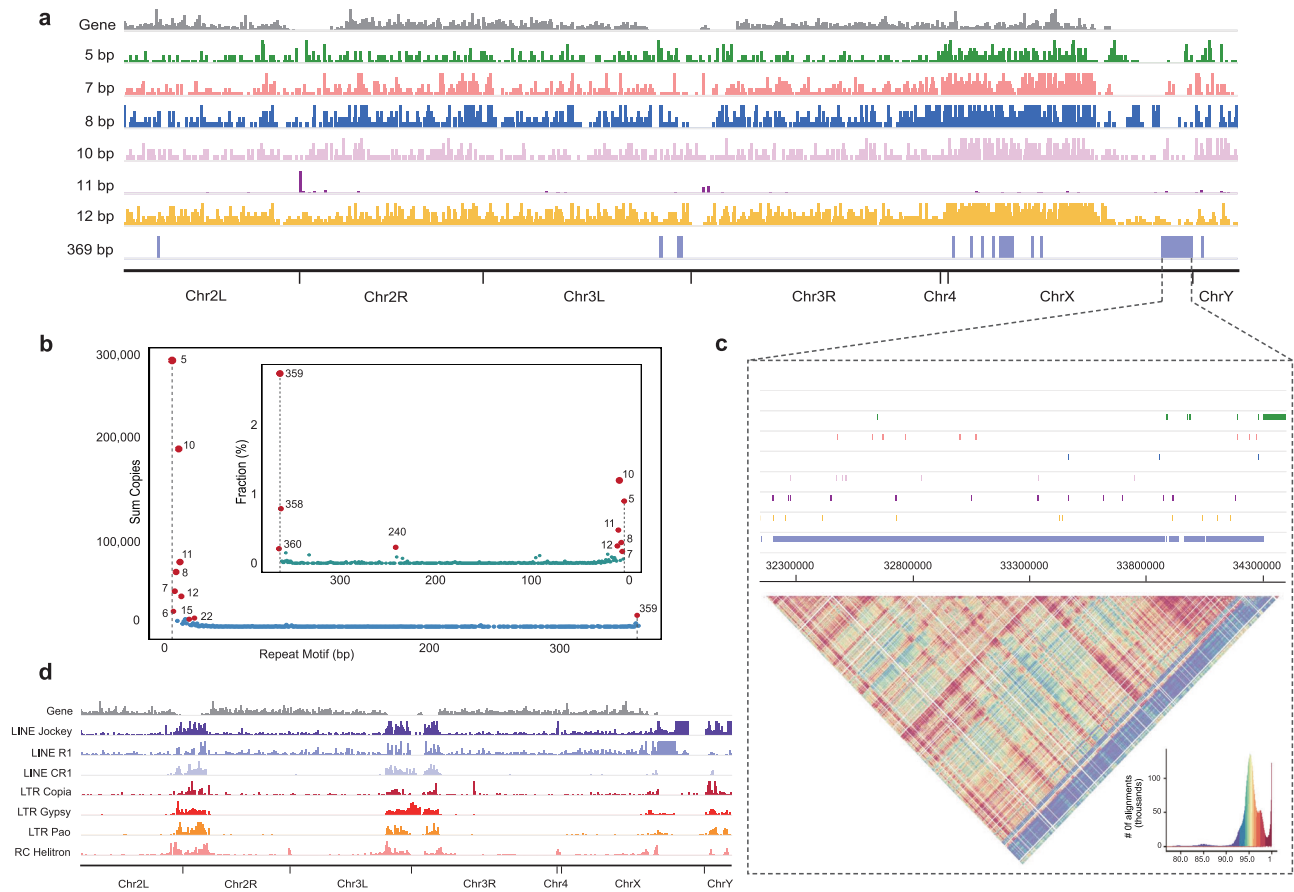


Fig. 3 | Identification and visualization of centromeric repeat motifs of the Dm.nT2T genome assembly. **a** Distribution of the top 7 repeat motifs across chromosomes. **b** Distribution of high-copy-number simple repeat motifs within the

genome. **c** Heatmap of sequence similarity in the centromeric regions of the X chromosome, with colors representing identity. **d** Distribution of different types of transposons across chromosomes. Source data are provided as a Source Data file.

Table 4 | The location of the centromeric region in the Dm.nT2T genome assembly

Chromosome	Centromere	Start	End	Length (bp)
Chr2L	CEN1	23,947,953	24,206,470	258,517
Chr2R	CEN2	859,024	929,547	70,523
Chr3L	CEN3	28,146,212	28,652,792	506,580
Chr3R	CEN4	404,492	1,805,165	1,400,673
Chr4	CEN5	824,137	859,453	35,316
ChrX	CEN6	32,308,449	34,952,072	2,643,623
ChrY	CEN7	1,314,078	2,838,305	1,524,227

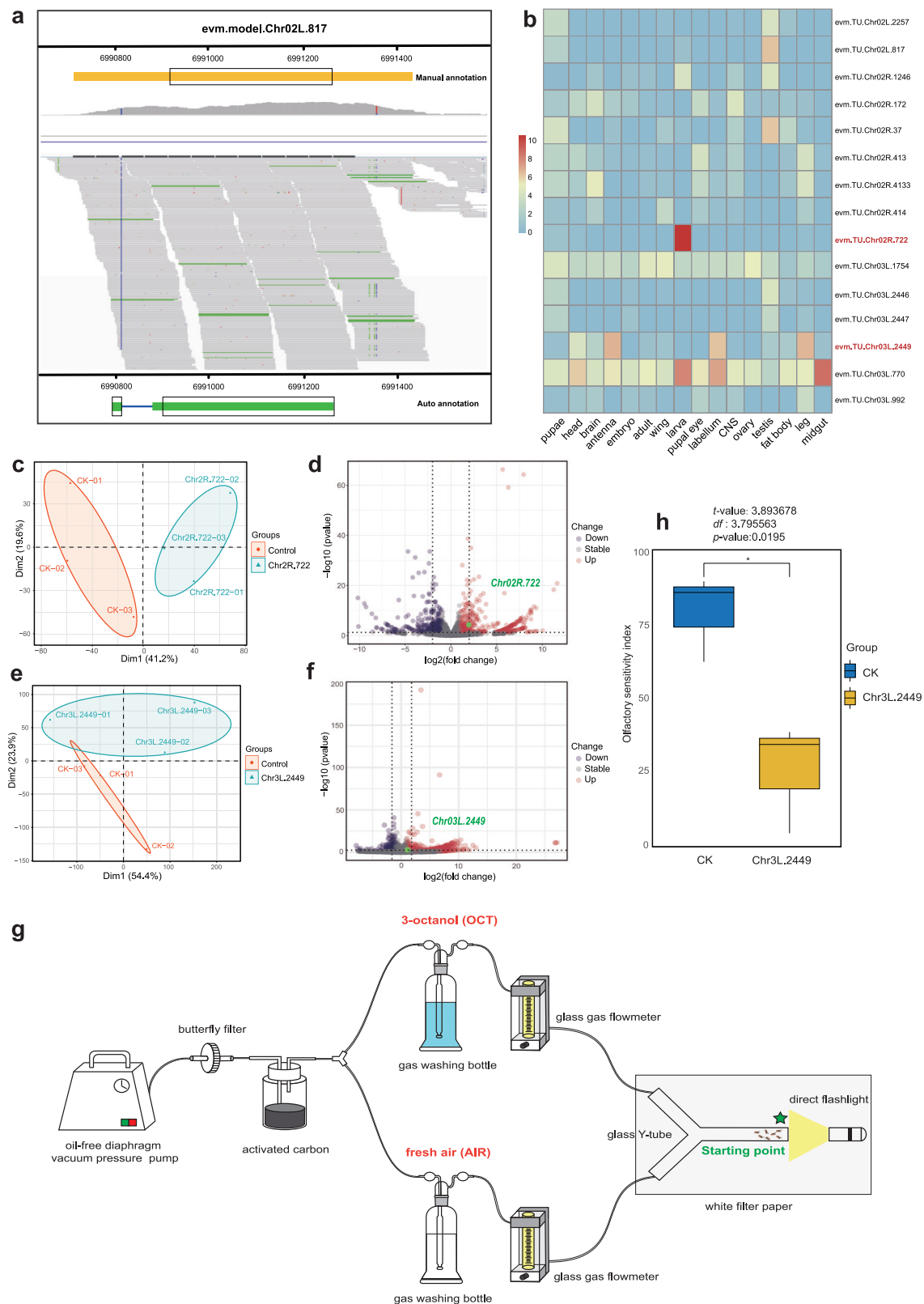
the mid-regions of autosomes, the distal right arm of ChrX, and the proximal left arm of ChrY relative to input controls (Supplementary Fig. 9). Notably, the centromeric region of ChrX primarily consists of a 359 bp repeat motif (average 12,340 copies, constituting 2.74% of the genome; Fig. 3c), consistent with previous reports⁵⁸.

We also analyzed the transposon distribution pattern across the genome to further verify the precise localization and composition of the centromeric region. Manual curation was performed on non-LTR and LTR transposons derived from the TE annotation data. We found that the *D. melanogaster* centromeric region mainly consists of seven types of transposons: *Jockey*, *CRI*, *RI*, *Copia*, *Gypsy*, *Pao*, and *Helitron*, mainly belonging to the LINE, LTR, and RC (Rolling-Circle) families. Notably, the centromeric regions of most chromosomes were enriched with the *Jockey* family, which is consistent with earlier reports²⁰ (Fig. 3d).

Refinement of gene annotation and gene discovery
To refine the gene annotations, we employed two complementary approaches. Initially, we utilized multi-evidence approaches, including homologous sequences, transcriptome sequences, and de novo predictions, which were submitted to EvidenceModeler (EVM)⁵⁹ to annotate the assembly. Then, all genes were lifted over from the current reference genome of *D. melanogaster* (r6.54) using Liftoff⁶⁰. By filtering overlapping genes between these two parts, we obtained the primary annotation set comprising 17,898 genes, among which there were 213 potential genes.

To investigate the 213 candidate genes, we conducted reciprocal best-hit (RBH) BLASTP³¹ alignments between their protein sequences derived from the Dm.nT2T and R6 genome assemblies. This analysis identified 80 homologous genes. The remaining 133 genes were subjected to cross-validation against transcript assemblies from multiple tissues to confirm transcriptional activity. This validation finally yielded 92 genes (Supplementary Data 5), among which, 90 genes reside in unannotated genomic intervals, while 2 were reconstructed from unassembled regions via TBLASTN³¹ interrogation. Coding potential assessment using CPC2⁶¹ predicted 74 genes as protein-coding. Interestingly, these genes exhibit non-random chromosomal distribution: Chr3R harbors ~63% of loci, contrasting with minimal representation on sex chromosomes.

To enhance the precision of gene annotation in the Dm.nT2T genome, we implemented a hybrid curation pipeline integrating computational predictions with manual verification⁶². Using the specialized genome browser IGV-GSaman (<https://gitee.com/CJchen/IGV-SRNA>), a tool optimized for resolving complex gene architectures



through integrative visualization of multi-omics evidence, we systematically rectified annotation errors based on cross-tissue transcriptomic alignments. This manual curation targeted three prevalent misannotation types endemic to automated annotation systems: (i) fragmented gene models, (ii) gene fission artifacts, and (iii) gene fusion errors. Through iterative refinement of exon boundaries and splicing isoforms (exemplified in Fig. 4a for type I), we corrected structural

defects in 35/92 (38%) genes, with fragmented gene models constituting the dominant error class (32/35 cases, 94.3%; Supplementary Data 6).

Functional validation of genes

To validate the biological functions of these genes, we retrieved 48 publicly accessible RNA-seq datasets encompassing 16 distinct tissues

Fig. 4 | Comprehensive analysis of gene identification and functional validation. **a** An example of type I error (fragmented gene models) corrected in structural refinement of misannotated genes. **b** Expression heatmap of genes in 16 tissues, with red representing the two genes selected for functional validation (Only genes with TPM > 5 are shown). **c** Transcriptomic consequences of Chr2R.722 knockout with PCA analysis results of transcriptomic data (mutant versus control) (Chr2R.722: Chr2R.722 knockout). **d** Volcano plot of significantly upregulated genes in the Chr2R.722 mutant strain flies ($p\text{-adj} = 0.0022$). **e** Transcriptomic consequences of Chr3L.2449 knockout with PCA analysis results of transcriptomic data

(mutant versus control) (Chr3L.2449: Chr3L.2449 knockout). **f** Volcano plot of significantly upregulated genes in the Chr3L.2449 mutant strain flies ($p\text{-adj} = 0.0227$). **g** Diagram of the odor preference experimental setup. **h** Comparison of olfactory sensitivity indices between Chr3L.2449 mutant strain flies and non-mutant flies (CK Control group, Chr3L.2449 Chr3L.2449 mutant). Each group consisted of three biological replicates, with 100 five-day-old flies per replicate ($n = 3$ for each group). A two-tailed independent samples t -test was used, with $p < 0.001$ indicated as ***, $p < 0.01$ as **, and $p < 0.05$ as *. Source data are provided as a Source Data file.

from the ENA database, with three biological replicates per tissue (Supplementary Data 7), enabling the derivation of their expression profiles. Analysis revealed that a distinct subset of these genes displayed statistically significant tissue-specific expression patterns (Fig. 4b). Among them, we selected two representative genes, Chr2R.722 and Chr3L.2449, for functional validation via multiple experimental approaches. Notably, Chr2R.722 exhibited predominant expression during larval stages, whereas Chr3L.2449 showed marked enrichment in legs, labellum, and antenna.

Using CRISPR/Cas9 technology, a frameshift mutation was introduced within the open reading frame (ORF) of Chr2R.722, anticipated to disrupt its functional integrity. To validate the mutation, transcriptome sequencing was conducted on homozygous mutant flies and wild-type controls using larval samples. Principal component analysis (PCA) of the transcriptome data revealed clear segregation between mutant and control groups (Fig. 4c). Differential expression analysis confirmed significant downregulation of Chr2R.722 expression in mutant flies compared to controls, verifying the efficacy of the genetic manipulation and establishing the mutation (Fig. 4d and Supplementary Data 8). Nevertheless, homozygous mutant flies exhibited no observable phenotypic abnormalities, including feeding, locomotion, or reproductive behaviors. Gene Ontology (GO) functional enrichment analysis of differentially expressed genes following Chr2R.722 mutation suggested its potential implication in biological processes, including chitin-based cuticle sclerotization, response to monoamine, and *Toll* and *Imd* signaling pathways (Supplementary Fig. 10). These processes are related to the development of the exoskeleton, neural transmission, and immune responses in fruit fly larvae. Given that the larval stage is crucial for the growth and development of fruit flies, and larvae are particularly susceptible to pathogen invasion⁶³, the participation of Chr2R.722 in *Toll* and *Imd* signaling pathways—key immune response mechanisms in fruit flies—implies a potential role in bolstering the larvae's resistance to pathogens.

Similarly, for Chr3L.2449, a frameshift mutation was introduced into its ORF. PCA analysis revealed distinct separation between groups, and differential expression analysis demonstrated significant downregulation of Chr3L.2449 in mutant flies relative to controls, thereby supporting the validity of the experimental findings and confirming the mutation (Fig. 4e, f and Supplementary Data 9). Since Chr3L.2449 is highly expressed in tissues harboring nearly all olfactory receptors and specific taste receptors in *D. melanogaster*, this observation suggests its potential impact on olfactory function. To investigate this, we conducted olfactory choice behavioral assays comparing mutant flies deficient in this gene with wild-type control flies (Fig. 4g). In each trial, 100 mutant flies (5 days old adult) and 100 control flies (5 days old adult) were introduced into a Y-tube, with one channel infused with the aversive odorant 3-Octanol (OCT), which is commonly used in odor experiments due to its repellent effect on fruit flies⁶⁴, while the other channel received fresh air. Through six sets of repeated experiments (three for mutants and three for controls), we assessed the olfactory sensitivity of both mutant and control flies. Our results revealed that the olfactory ability of the mutant flies showed a significant degradation compared to the control flies (Fig. 4h). Even when the odors in the two channels were switched, the difference in olfactory sensitivity still existed, although it was not statistically significant (Supplementary

Fig. 11). This suggests that the mutation of Chr3L.2449 caused a general impairment in the mutant flies' olfactory system, leading to their diminished sense of smell.

Discussion

The present study reports a near-complete T2T genome assembly of *Drosophila melanogaster* (Dm.nT2T), representing a significant advancement in resolving the genomic complexity of this classic model organism. By integrating state-of-the-art sequencing technologies and computational approaches, we addressed the majority of longstanding gaps in the reference genome, uncovered additional genomic elements, and provided functional insights into unannotated genes.

Multi-technology integration resolves complex genomic regions

The generation of Dm.nT2T represents a methodological breakthrough enabled by the strategic combination of PacBio HiFi, ONT UL reads, and Hi-C data. This integrative approach overcame technical limitations in resolving repetitive and structurally complex regions. PacBio HiFi reads (17.36 Gb, ~107.4×) provided high base accuracy (>99.9%), critical for validating small-scale variations and repetitive element boundaries, while ONT UL reads (156.25 Gb, ~966.7×) with mean lengths of 26.49 kb spanned large repetitive arrays (e.g., centromeric satellites and telomeric retrotransposons) that confounded shorter reads. Hi-C data further facilitated chromosome-scale scaffolding, anchoring 25 contigs into seven major chromosomes with robust diagonal interactions in chromatin contact maps (Fig. 1b), confirming assembly accuracy. This pipeline achieved a contig N50 of 21.93 Mb and closed 93.28% of gaps in the R6 genome. Using quality assessment metrics (QV: 40.24; BUSCO: 98.8%), we demonstrated that multi-technology integration not only improved genome completeness but also enhanced accuracy. The Dm.nT2T assembly identified six telomeric regions (e.g., ChrX left-arm telomere, 76.25 kb) and defined centromeric regions enriched in 5–12 bp satellite repeats (Fig. 3a–c). These advancements highlight the utility of multi-technology integration for resolving complex eukaryotic genomes, serving as a model for future T2T efforts in non-model organisms.

SINE transposons and strain-specific structural variation

SINE (0.02% of the genome) were detected in the Dm.nT2T assembly of *D. melanogaster*, which were not identified in the R6 reference genome. SINEs, which rely on other retrotransposons for mobility, are known to influence genome structure and function in mammals (e.g., human *Alu* elements)^{65,66}. To evaluate whether these elements correspond to previously described SINE-like families, we compared our two SINE consensus sequences (Supplementary Table 8) with DINE-1⁶⁷ and the suffix element (<https://www.ncbi.nlm.nih.gov/nucore/AF363625.1/>). The two consensus sequences showed high similarity to each other (~76%) but very low similarity to DINE-1 (~16–20%) and suffix (~25–30%), indicating that they are evolutionarily distinct. These findings suggest that the SINE elements identified here represent an independent lineage in *D. melanogaster*. Their presence in Dm.nT2T suggests unrecognized roles in *Drosophila* genome evolution, potentially driving adaptive changes through insertional mutagenesis or regulatory rewiring. Future studies should investigate their origin (e.g.,

horizontal transfer or ancestral retention) and functional impact via comparative genomics with other *Drosophila* species.

Systematic comparison of the Dm.nT2T assembly with the R6 genome enabled comprehensive identification of numerous SVs, encompassing INDELs, inversions, and translocations. INDELs were the predominant type, with over 79% of them exceeding 100 bp. The majority of such large SVs exceed the detection capacity of short-read sequencing technologies, again underscoring the critical importance of high-quality reference genomes for resolving structural variation.

As a major component of genomic variation, the SVs' functional impact frequently exceeds that of single-nucleotide polymorphisms (SNPs), positioning them as significant drivers of phenotypic diversity⁶⁸. However, it should be emphasized that these SVs likely represent substantial genetic divergence between their respective strains (*Canton S* versus *ISO-1*) rather than assembly artifacts in the current reference genome. This interpretation is supported by existing literature demonstrating that inter-strain genetic variations can indeed reach such magnitudes of divergence. For instance, Chakraborty et al. detected 1890 INDELs >100 bp between the *A4* and *ISO-1* strains, impacting over 7 Mb of genomic sequence and exhibiting strong associations with complex traits⁶⁹. Notably, structural variation also manifests within individual strains. Courret and Larracunte, employing cytogenetic approaches, documented extensive structural rearrangements proximal to the X centromere in *D. simulans*⁷⁰. This region displayed polymorphisms not only between strains but also among single-isolate subcultures derived from different laboratories, and even within individual isolates. Collectively, the observed inter- and intra-strain SV heterogeneity necessitates future investigations integrating long-read sequencing across multiple individuals and strains, coupled with transcriptomic and phenotypic analyses, to elucidate the functional and evolutionary significance of these SVs.

Functional validation of genes: insights into olfactory biology

Leveraging the improved assembly, this study identified 92 genes, the majority of which had been overlooked in previous genome annotations. In this study, we validated two genes (Chr2R.722 and Chr3L.2449) using targeted gene knockout methodologies. Although both genes showed no significant morphological changes in homozygous mutants, a detailed functional assessment revealed that the knockout of Chr3L.2449 results in a substantial impairment in olfactory capability (Fig. 4h and Supplementary Fig. 11), implicating this gene in olfactory signaling. In the R6 version, this gene (Chr3L.2449) is located in an unassembled gap and is highly expressed in olfactory tissues, showing significant coding potential. These findings thus highlight the critical role of complete, gap-free genome assemblies in uncovering functionally relevant genomic elements previously obscured by assembly limitations.

Previous research has extensively documented that olfactory functions in *Drosophila* are predominantly governed by specialized olfactory receptor neurons (ORNs) within the antennal lobe regions of the brain, where nearly each ORN forms a precise synaptic connection with a corresponding projection neuron (PN) to facilitate odor discrimination and signal transmission^{71,72}. Based on these established neural pathways, we postulate that the Chr3L.2449 gene may function analogously, potentially acting as a key component in a similar receptor-to-neuron mapping system. Further functional validation, such as electrophysiological recordings or neuron-specific expression assays, will be essential to clarify its exact interactions and integration within the brain's olfactory circuitry.

Evolutionary and comparative genomic implications

Dm.nT2T also provides the insights into *Drosophila* genome evolution. In most eukaryotic organisms, including plants and animals, centromeres are characterized by extensive arrays of simple tandem repeats spanning megabases, typically comprising ~150 bp repeat

units, satellite DNA sequences, and transposable elements^{73,74}. Primate centromeres further exhibit higher-order repeats (HORs), formed by the specific organization of distinct repeat monomers into larger repeat units following a specific pattern⁷⁵. In contrast, the centromeres of *D. melanogaster* lack HORs and instead harbor various combinations of repetitive satellite sequences ranging from 5 to 12 bp in length⁷⁶. This study further quantified the copy numbers of individual satellite repeats (e.g., 5 bp repeat occurs 290,576 copies). The short centromeric satellite sequences are also present in other *Drosophila* species, albeit with species-specific motifs. This sequence brevity suggests rapid diversification of centromeric motifs within the *Drosophila* genus, which warrants comparative analyses across additional species.

Regarding the telomeric structures of *D. melanogaster*, they exhibit distinct characteristics compared to those of other eukaryotic organisms. In most animals, telomeres are maintained by the conserved telomeric repeat motif TTAGG, whereas the consensus sequence in plants is TTTAGGG. In contrast, the telomeres of *D. melanogaster* are elongated primarily through the targeted transposition of three specialized retrotransposons: *HeT-A*, *TART*, and *TAHRE*⁴⁹. The distinctive telomere structure maintained by retrotransposons appears to be largely absent in the vast majority of other *Drosophila* species, except in *D. virilis* and *D. yakuba*^{77,78}. Prior research into the dynamics of insect transposons provides relevant context; it demonstrated the remarkably rapid invasion and spread of *P* elements into natural populations of Japanese *D. simulans* through horizontal transfer⁷⁹. This phenomenon, while rare across the genus, suggests the possibility of independent evolutionary events or shared ancestral mechanisms in these lineages. This finding highlights the potential for mobile genetic elements to dramatically alter genome architecture quickly. Consequently, we hypothesize that the unique, retrotransposon-dependent telomere pattern of *D. melanogaster* may itself originate from similar parasitic or invasion events involving the capture and domestication of TE sequences during its evolutionary history. However, given that the genus *Drosophila* encompasses over 1600 extant species⁸⁰, with the majority remaining uncharacterized in terms of telomere architecture, the evolutionary diversity of telomere structures within this genus represents a critical area for future investigation.

We additionally identified length variations in the telomeric regions compared to the R6 genome, likely reflecting natural variation in telomere organization among different *Drosophila* strains. It should be noted that our study employed the *Canton S* strain, which holds a long history of utilization in genetic research^{81,82}, whereas the established reference genome is based on the *ISO-1* strain. Previous studies have demonstrated genomic variation among *D. melanogaster* strains. For example, *Canton S* and *Oregon R* strains exhibit ~1% sequence divergence in the *bithorax* region⁸³. It has also been demonstrated that different *D. melanogaster* strain possesses distinct telomeric structures, which are frequently correlated with variations in the copy number, arrangement, and insertion sites of telomere-associated retrotransposons, highlighting the highly dynamic and plastic nature of telomeric regions⁴⁸. Therefore, we propose that the observed structural discrepancies are more likely to reflect genuine polymorphisms between strains rather than technical artifacts of genome assembly. This underscores the critical importance of accounting for strain background in telomere studies. Therefore, the Dm.nT2T genome serves as a valuable complement to the reference genome, providing a broader view of intraspecific genomic variation and facilitating the discovery of structural variants potentially missed by a single reference.

Limitations and future directions

However, it is noteworthy that despite the ultra-high-coverage (~1000×) achieved in the Dm.nT2T assembly and incremental improvements in reconstructing complex regions, persistent gaps

remain, reflecting persistent technical challenges in resolving ultra-long tandem repeats. Regarding genome assembly size, although our assembly achieves a 17.93 Mb increase (161 versus 143 Mb) in size compared to the current reference genome (R6), the total assembly length still falls short of previously estimated *C*-values for *D. melanogaster*, which range between 117 Mb and 205 Mb (<http://genomesize.com/results.php?page=1>). The remaining discrepancy of approximately 54 Mb (relative to the 215 Mb estimate) likely comprises: (1) Heterochromatic regions (pericentromeric and Y chromosome satellite DNA); and/or (2) Strain-specific structural variation: the *Canton S* strain may harbor deletions or rearrangements relative to the strains used in *C*-value measurements, contributing to size differences. Future efforts could employ ultra-long Strand-seq or targeted CRISPR-based scaffolding to achieve a complete characterization of chromosome mechanics by closing these residual gaps.

Additionally, functional validation of genes in this study is also preliminary. Expanded phenotyping analyses (e.g., stress resistance, lifespan assays) and molecular investigations (e.g., ChIP-seq for regulatory element mapping) are warranted to elucidate their biological roles. Further, the long-term integration of the Dm.nT2T assembly with multi-omics datasets (e.g., single-cell RNA-seq, epigenomic profiles) will facilitate systems-level analyses of gene regulatory networks. Comparative genomics across *Drosophila* strains and species could further reveal the evolutionary dynamics of SINEs, centromeres, and telomeres, enhancing our understanding of genome plasticity.

Methods

Ethics

Experiments in this study used *Drosophila melanogaster* (fruit fly) samples. According to widely accepted guidelines, research involving fruit flies does not require formal ethical approval.

Fly sampling

Adult males of *D. melanogaster* (*Canton S* strain, stock number 64349, the Bloomington Stock Center) were used for genome sequencing obtained in 2018. For this, adult males were sampled from each of ten temporary lines (each established with a pair of virgin male and female maintained with cornmeal medium under ~20 °C, 60% RH, and continuous illumination). For each of these lines, newly emerged *F*₁ males were collected daily and preserved at -80 °C. The line yielding the maximal total number of *F*₁ males was subjected to DNA extraction and genome sequencing.

DNA extraction

DNA sample was extracted for line #14 (for which a total of 236 *F*₁ males were sampled) by the SDS method and purified with QIAGEN® Genomic kit (Cat#13343, QIAGEN) according to the standard operating procedure provided by the manufacturer. DNA degradation and contamination of the extracted DNA were monitored on 1% agarose gels. DNA purity was then detected using a NanoDrop™ One UV-Vis spectrophotometer (Thermo Fisher Scientific, USA), with OD_{260/280} ranging from 1.8 to 2.0 and OD_{260/230} ranging from 2.0 to 2.2. At last, DNA concentration was further measured by Qubit® 3.0 Fluorometer (Invitrogen, USA).

ONT library preparation

A total amount of 2 µg DNA per sample was used as input material for the ONT library preparations. After the sample was quantified, size selection of long DNA fragments was performed using the BluePippin system (Sage Science, USA). Next, the ends of DNA fragments were repaired, and ligation reactions were conducted with NEBNext Ultra II End Repair/da-tailing Kit (Cat# E7546). The adapter in the LSK109 kit was used for further ligation reaction, and Qubit® 3.0 Fluorometer (Invitrogen, USA) was used to quantify the size of Library fragments.

PacBio library preparation

SMRTbell target size libraries were constructed for sequencing according to PacBio's standard protocol (Pacific Biosciences, CA, USA) using 15 kb preparation solutions. The main steps for library preparation are: (1) gDNA shearing, (2) DNA damage repair, (3) blunt-end ligation with hairpin adapters from the SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences), (4) size selection, and (5) binding to polymerase. Briefly, a total amount of 2 µg DNA per sample was used for the DNA library preparations. The genomic DNA sample was sheared by g-TUBEs (Covaris, USA) according to the expected size of the fragments for the library. Single-strand overhangs were then removed, and DNA fragments were damaged, repaired, and polished, and ligated with the stem-loop adaptor for PacBio sequencing. Link-failed fragments were further removed by exonuclease, and target fragments were screened using the BluePippin (Sage Science, USA) system. The SMRTbell library was then purified using AMPure PB, and an Agilent 2100 Bioanalyzer (Agilent Technologies, USA) was used to detect the size of Library fragments. All sequencing was performed on a PromethION sequencer (Oxford Nanopore Technologies, UK) instrument and the PacBio Sequel II instrument with Sequel II Sequencing Kit 2.0 for ONT and PacBio on Nextomics, respectively.

Data quality control

Nanopore sequencers output FAST5 files containing signal data. Base calling was first performed to convert the FAST5 files to FASTQ format using Guppy (v3.2.2) (<https://nanoporetech.com/zh/software/other/guppy>) with the configuration file dna_r9.4.1_450bps_fast.cfg and built-in adapter trimming parameter (-trim_adapters yes). The raw reads of fastq format with mean_qscore_template <7 were then filtered, resulting in pass reads, and were directly used for subsequent genome assembly. Besides, the PacBio raw subreads were converted into circular consensus sequence (CCS) reads using the official CCS tool provided by PacBio.

Genome de novo assembly

We used multi-step strategies for genome assembly primarily based on ONT ultra-long reads. First, raw ONT UL reads were self-corrected using the NextCorrect module of NextDenovo²⁵ (v2.3.1) with parameters: reads_cutoff:1k, seed_cutoff:143k, generating 5.9 Gb of consistent sequences (CNS reads). These CNS reads were assembled using the NextGraph module based on sequence correlation, producing a preliminary genome assembly with a total size of 161.61 Mb and a contig N50 of 21.92 Mb. To improve the accuracy of the assembly, we performed multiple rounds of polishing. First, ONT long reads were used for three rounds of polishing with Nextpolish²⁷ (v1.3.0) using default parameters. Then, PacBio HiFi reads were applied for another three rounds of polishing, followed by four rounds of polishing using BGI short reads. After these steps, the final polished genome assembly was obtained.

We employed Hi-C data to facilitate genome assembly. A total of 161,299,842 paired-end raw reads were generated from the libraries. Subsequently, quality control of Hi-C raw data was conducted using fastp⁸⁴ (v0.12.6) with default parameters, consistent with established methodologies⁸⁵. Firstly, low-quality sequences (quality scores < 20), adaptor sequences, and sequences shorter than 30 bp were filtered out. Then, the clean paired-end reads were mapped to the draft assembled sequence using Bowtie2⁸⁶ (v2.3.2; with parameters: -end-to-end -very-sensitive -L 30 -score-min L, -0.6, -0.2). These parameters were selected to enhance mapping specificity and reduce multi-mapping, as adopted in previous telomere-to-telomere genome assembly study⁸⁷. Unmapped reads were examined to identify potential junction sites (resulting from ligation at restriction sites). These reads were subsequently trimmed and realigned using Bowtie2. The results from the two rounds of alignment were combined, yielding a total of 53,198,850 high-confidence mapped paired-end reads. Only valid

interaction pairs were retained from the mapped paired-end reads for further analysis. The scaffolds were further clustered, ordered, and oriented onto chromosomes using LACHESIS²⁸ (<https://github.com/shendurelab/LACHESIS>) with the following parameters: CLUSTER_MIN_RE_SITES = 100, CLUSTER_MAX_LINK_DENSITY = 2.5, CLUSTER_NONINFORMATIVE_RATIO = 1.4, ORDER_MIN_N_RES_IN_TRUNK = 60, ORDER_MIN_N_RES_IN_SHREDS = 60. Finally, placement and orientation errors exhibiting obvious discrete chromatin interaction patterns were manually adjusted.

To evaluate the completeness of the final genome assembly, multiple methods were used to assess the quality of the genome assembly. BUSCO³⁰ (v4.0.5) was used for the assessment with default parameters. To evaluate the accuracy of the assembly, all the short paired-end reads were mapped to the assembled genome using BWA⁸⁸ (v0.7.12) with default parameters. The coverage of expressed genes of the assembly was examined by aligning all the RNA-seq reads against the assembly using HISAT2⁸⁹ (v2.1.0) with default parameters. Minimap2²⁶ (v2.24) software with parameters “-a -x map-ont -t 30” was used to map the long reads. We used the read re-mapping ratio and genome coverage of sequencing reads as indicators by using samtools⁹⁰ (v1.4). Besides, to assess sequence consistency, the base-level accuracy of the assembly was estimated by aligning short reads to the assembly using BWA⁸⁸ (mem mode with default parameters), followed by variant calling using samtools and bcftools (v1.8.0)⁹⁰, both with default parameters. The homozygous single-nucleotide variant (SNV) rate was calculated from the resulting VCF file and used as the estimated error rate of the assembly. To avoid including mitochondrial sequences in the assembly, the draft genome assembly was submitted to the NT library, and aligned sequences were eliminated.

Repeat sequence annotation

First, we identified the simple repeat sequences (SSRs) by using software GMATA⁹¹ (v2.2), and TRF⁵⁴ (v4.07b) was used to identify all tandem repeat elements in the Dm.nT2T genome assembly. And then, TEs were identified using a combination of ab initio and homology-based methods. Briefly, a repeat library for the Dm.nT2T was predicted using MITE-hunter⁹² and RepeatModeler⁹³ (open-1.0.11) with default parameters to identify and cluster repetitive elements. Results from LTR_FINDER⁹⁴ (v1.07) and LTRharvest⁹⁵ (v1.6.2) were integrated, and false positives were removed using the LTR_retriever⁹⁶ (v2.9.0) pipeline. The obtained library was then aligned to TEclass Repbase to classify the type of each repeat family. Finally, we used RepeatMasker to search for TEs by mapping sequences against the de novo repeat library and Repbase TE library. All overlapping TEs belonging to the same repeat class were collated and combined. We classified these LTRs as either intact or non-intact LTRs.

Gene annotation

Before gene prediction, the assembled genome was hard and soft-masked using RepeatMasker. We used three independent strategies, including de novo prediction, homology-based search, and reference-guided transcriptome assembly for gene prediction in a repeat-masked genome. Specifically, genome assemblies and annotation files from closely related species (*D. yakuba*, *D. simulans*, *D. sechellia*, *D. erecta*, and *D. ananassae*) were downloaded from the NCBI Assembly database to perform homology-based gene prediction with GeMoMa⁹⁷ (v1.6.1). The corresponding assembly accession numbers and download links are provided in Supplementary Table 9. Then we used the software Liftoff⁶⁰ (v1.6.3, parameters: -chroms chrom.txt -polish -copies) to annotate protein-coding genes of the Dm.nT2T genome assembly based on the R6 genome. For transcriptome prediction, we used the pipeline to align filtered RNA-seq data to the Dm.nT2T genome assembly using STAR⁹⁸ (v2.7.3a) with default parameters. The transcripts were then assembled using StringTie⁹⁹ (v1.3.4 d), and ORFs were predicted using PASA⁵⁹ (v2.5.2). For the de novo prediction, RNA-

seq reads were de novo assembled using StringTie and analyzed with PASA to produce a training set. Then we used AUGUSTUS¹⁰⁰ (v3.3.1) with default parameters for gene prediction. Finally, gene models from these three methods were integrated into a non-redundant set of high-confidence gene models. EVM⁵⁹ (v1.1.1) was used to produce an integrated gene set from which genes with TE were removed using the TransposonPSI (<https://transposonpsi.sourceforge.net>), and the mis-coded genes were further filtered. Untranslated regions (UTRs) and alternative splicing regions were determined using PASA based on RNA-seq assemblies. We retained the longest transcripts for each locus, and regions outside of the ORFs were designated UTRs.

Two methods were used to predict the functions of protein-coding genes. First, Blastp³¹ (v2.7.1) was used to search against protein sequences at the NCBI nonredundant protein database (NR), KEGG, KOG, and the Swiss-Prot database with an *E*-value cutoff of $1e \times 05$. Second, protein domain and gene ontology term annotations were performed using InterProScan¹⁰¹ (5.33–72.0) with default parameters. Results from the five database searches were concatenated. To obtain the ncRNA (non-coding RNA), two strategies were used: searching against the database and prediction with the model. Transfer RNAs (tRNAs) were predicted using tRNAscan-SE¹⁰² (v2.0) with eukaryote parameters. MicroRNA, rRNA, small nuclear RNA, and small nucleolar RNA were detected using Infernal cmscan to search the Rfam¹⁰³ (14.0) database. The rRNAs and their subunits were predicted using RNAmmer¹⁰⁴ (v1.2).

Identification of telomeric and centromeric regions

For telomeres, we used RepeatMasker to annotate the repetitive sequences in the Dm.nT2T genome assembly. Based on the literature, we retrieved four types of *Drosophila* telomere-specific transposons from the RMRBMeta.embl database. We performed a lastz search using these four transposons and utilized a custom script to calculate the total alignment length of the telomeric sequences for each chromosome. For the centromeric region, we first conducted tandem repeat annotation of the entire genome using TRF⁵⁴ (v4.07b) (parameters: 2 7 80 10 50 500 -f -d -m). We then merged all annotation results using TRF2GFF (v0.0.4, <https://github.com/Adamtaranto/TRF2GFF>). Following the filtering methods from the grape genome pipeline, we identified several candidate centromeric repeat motifs. We also retrieved previously published characteristic transposons in centromeric regions. By visualizing the distribution of these repeat motifs and transposons in IGV, we initially delineated the centromeric regions. To more accurately validate the regions we identified, we reanalyzed published CENP-A ChIP data. The ChIP-seq paired-end reads of CENP-A were downloaded from the ENA database and mapped to the Dm.nT2T genome assembly using BWA⁸⁷ (v0.7.12). We then used MACS2¹⁰⁵ (v2.2.9.1) for peak calling (parameters: -t replicate. bam -c control. bam -f BAMPE -broad -B -g dm -outdir macs2_result -n CENPA_rep1 2 > CENPA_rep1.macs2.log &). The peak calling results from the three replicates were merged. Finally, we utilized a custom R script to visualize the distribution of peaks across the genome.

Structural variation and synteny analysis

We performed pairwise genome alignments between the Dm.nT2T genome assembly and the R6 genome by using the MUMmer4¹⁰⁶ (v4.0.0) with parameters: -maxmatch -c 250 -l 50 -g 100. Only one-to-one alignment was kept in subsequent analysis. Then we visualize the results with RectChr (v1.36, <https://github.com/hewm2008/RectChr>). We mapped PaBio HiFi reads (sequence depth: 107×) to the R6 by pbmm2 (v1.13.1, <https://github.com/PacificBiosciences/pbmm2>) with default parameters. Then, four different structural variation detection software, including Sniffles2³⁶ (v2.2), SVIM³⁷ (v2.0.0), PBSV (v2.9.0), and CuteSV³⁸ (v2.1.0), were used to detect SVs. For each software result, we filtered and retained the variants with the “PASS” tag. And then SURVIVOR¹⁰⁷ (v1.0.7) was performed to merge all filtered SVs

from these methods with parameters (1000 2 1 1 0 50). The final SV results retained greater than 50 bp and were supported by at least two SV detection software. The SVs annotation was performed by ANNOVAR⁴¹ with the default parameters. The repeat sequence annotation of Indels was used with RepeatMasker with parameters: -nolow -no_is -norna -species "drosophila melanogaster". We performed GO functional enrichment analysis using Metascape¹⁰⁸ (v3.5.20240901), selecting only the significantly enriched GO terms.

ATAC-seq mapping and peak calling

Twelve ATAC-seq datasets were downloaded from the ENA database. The ATAC-seq reads for each tissue were mapped to both the reference genome and the Dm.nT2T genome assembly using Bowtie2⁸⁶ (v2.5.4). Peak calling was performed for each tissue using Genrich (v0.6.1, available at <https://github.com/jsh58/Genrich>) with parameters -j, -r, and -p 0.01. The peak calling results from all tissues were then merged. Using the reference genome coordinates as a standard, we converted the Dm.nT2T coordinates and filtered out peaks that overlapped with the reference genome by more than 1 bp. ChIPseeker¹⁰⁹ (v1.30.3) was utilized to annotate the peaks obtained after filtering and define the TSS region as ± 1 kb for peak profile visualization. Finally, functional annotation of peaks was performed using ChIPseeker's built-in functions.

Analysis of SDs in the genome and population

To investigate the distribution and variation of SDs in the Dm.nT2T genome assembly and across different populations, we conducted both genome-level and population-level SD identification analysis. We calculated the sequence length (>1 kb) and alignment identity (>90% identity) to identify the SDs of Dm.nT2T. We first identified SDs using Biser¹¹⁰ (v1.4). Obtained SDs were further classified into intra-chromosomal, inter-chromosomal, and overlapping categories. The genomic distribution of SDs was visualized using Circos¹¹¹ (0.69–9).

For population-level analysis, we downloaded resequencing data from six *Drosophila melanogaster* population groups from NCBI under accession number PRJNA924845. Raw reads were subjected to quality control using FastQC (v0.12.1, <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) and fastp⁸⁴ (v0.23.2), followed by alignment to the Dm.nT2T using BWA⁸³ (v0.7.19). The alignment files were processed with samtools⁸⁹ (v1.17) to convert SAM to BAM format and sort the data. SNP calling was performed using GATK¹¹² (v4.6.2), and raw SNPs were filtered using GATK's VariantFiltration module with the following parameters: QD < 2.0||QUAL < 30.0||SOR > 3.0||FS > 60.0||MQ < 40.0||MQRankSum < -12.5||ReadPosRankSum < -8.0. We used VCFtools¹¹³ (v0.1.17) to retain high-quality biallelic SNPs, filtering out sites with more than 10% missing genotypes and minor allele frequency (MAF) below 0.05. The filtered SNPs were intersected with SD and non-SD regions using bedtools¹¹⁴ (v2.31.1), and the total lengths of SD and non-SD regions were calculated from BED coordinates. SNPs within SD regions were extracted into separate VCF files. Population-specific SNP subsets within SDs were generated based on population metadata using bcftools⁸⁹ (v1.21). We annotated allele frequencies (AF) using the bcftools +fill-tags plugin and labeled each SNP dataset with the corresponding population group identifiers. All labeled datasets were then merged into a comprehensive file for downstream analysis of allele frequency distributions across populations.

Gene identification and function validation

Transcriptomic data from sixteen tissues (three biological replicates per tissue/stage) were downloaded from the ENA database for gene expression analysis. These selected reads underwent quality control and were then removed of adapters and filtered for low-quality bases using fastp⁸⁴ (v0.23.2). Only high-quality reads were retained for subsequent analysis. The cleaned reads were mapped to the Dm.nT2T genome assembly using HISAT2⁸⁸ (v2.1.0), and the gene expression

profile was created using the StringTie⁹⁸ (v1.3.4d) pipeline. The TPM (Transcripts Per Million) value was used to plot a heatmap using an R script. Differential expression analysis was performed using DESeq2¹¹⁵ (v1.34.0), with differentially expressed genes selected based on $p < 0.05$ and $\log_2|FC| > 1$.

We defined genes not annotated in the R6 genome as those that are: (a) absent in R6, (b) lacking significant BLASTP hits to R6 proteins, and (c) supported by transcriptional evidence. First, we used the lift-over method to identify 213 potential genes whose coordinates do not overlap with the R6 genome. Then, we conducted an RBH analysis using BLASTP³¹ (v2.14.1) with an *E*-value threshold of $1e-5$, comparing the protein sequences of the 213 genes from the Dm.nT2T genome assembly to the annotated protein set of the R6 genome. This allowed us to differentiate putative orthologs from paralogs. Genes that lacked significant hits in the R6 protein database were designated as candidate genes. Finally, to validate their transcriptional activity, we cross-referenced the corresponding GFF annotations with transcriptome assemblies derived from multiple tissues. Protein sequences of candidate genes with evidence of transcription were subsequently aligned to the R6 genome assembly using TBLASTN³¹ (v2.14.1) with the same *E*-value threshold ($1e-5$) to evaluate whether these sequences map to unannotated genomic regions or potentially reflect loci absent from the reference assembly. Finally, we assessed the protein-coding potential of the confirmed candidates using CPC2⁶² (standalone-1.0.1), thereby refining our identification of bona fide protein-coding genes.

We focused on two tissue-specific genes: Chr2R.722 and Chr3L.2449 to validate their biological functions. We used CRISPR/Cas9 to perform a frame shift mutation on Chr2R.722 and a knockout on Chr3L.2449. We used the CRISPR-NHEJ technique, which introduces small insertions or deletions, leading to frameshift mutations in the coding region of the gene. Based on the DNA sequence of Chr2R.722, we designed PCR primers containing the frameshift mutation site (CTA TCCACGAAATCACTGACC and GTAGTTGCCATGTCCGTAGTTGCC) and gRNA. After amplification, we purified the mutated fragment and cloned it into a Cas9-DNA plasmid. We selected the *w1118* strain of *D. melanogaster* for microinjection. Subsequently, we constructed a stable homozygous line for the gene, and after multiple generations of screening, we finally obtained a stable homozygous line of the gene. For Chr3L.2449, we used CRISPR technology to induce a double-strand break at a specific location in the genome, and by utilizing the NHEJ repair mechanism, we deleted a specific sequence to achieve gene knockout. Based on the DNA sequence of Chr3L.2449, we designed PCR primers (TGAAATACAGACATTCTTCTATAC and ATGTATTAAC-TACTTTGCCATTTA) and gRNA. Similarly, we constructed the vector using a Cas9-DNA plasmid and selected the *w1118* strain of *D. melanogaster* for microinjection. After cross-identification and constructing a stable line. The homozygous mutant line was established through three generations of genetic crosses with balancer strains, followed by PCR and Sanger sequencing to confirm homozygosity at the target locus based on chromatogram analysis to obtain a stable homozygous line for this gene. To further investigate the transcriptional effects of these mutations, RNA-seq was performed on both mutant and non-mutant flies. For Chr2R.722, third-instar larvae were collected, with three biological replicates for both mutant and non-mutant groups, each replicate consisting of 30 larvae. For Chr3L.2449, leg tissues from 5-day-old adult flies were used, with three biological replicates for each group, each replicate containing 60 legs. RNA sequencing was performed on the Illumina NovaSeq 6000 platform using the PE150 sequencing strategy.

Olfactory choice experiment

The experiment was conducted in a quiet room at a temperature of $(25 \pm 1)^\circ\text{C}$, with uniform lighting and no odors present in the room. We selected adults that were 5-day-old post-eclosion for both the Chr3L.2449 mutant (experimental group) and non-knockout flies

(control group), without sexing. The choice of 5-day-old adult flies was based on previous studies^{116,117}, which utilized flies within comparable age ranges. This age group was chosen because these individuals exhibit optimal robustness and sensitivity for our experimental paradigm. OCT (3-octanol) is generally regarded as a neutral to mildly aversive odor for *Drosophila*, and it has been commonly used in olfactory-related behavioral assays^{118,119}. Therefore, 40 mL of OCT solution was prepared by mixing 60 μ L of OCT with 39.94 mL of mineral oil, resulting in a 1.5×10^{-3} (unitless) concentration. This solution was then connected to the experimental apparatus. A tube of experimental group flies (~100) was anesthetized by placing it in a freezer at -15°C for 15 s.

The anesthetized flies were then placed at the entrance of the straight arm of the Y-tube (starting point), and the entrance was sealed with gauze. Once the flies woke up, the air pump was activated, and OCT was introduced into one arm of the Y-tube, while fresh air was introduced into the other arm. After a 5-min exposure, timing began, and the flies were allowed 5 min to choose. Once most flies had made a stable choice, the airflow was stopped, and the number of flies remaining in each of the two arms was counted. This process was repeated for three trials. After completing the experiment with the experimental group, the Y-tube was cleaned with ethanol and dried. The airflow to the two arms was swapped, and the process was repeated three more times. The control group of flies underwent the same procedure. By counting the number of flies entering each of the two arms across three replicates, the average number was used to calculate the flies' olfactory perception score (Score). The Score = (number of flies entering the aversive odor arm)/(total number of flies). A higher score indicates poorer olfactory perception ability. To compare the differences in the Index values between the Chr3L2449 group (adults carrying the Chr3L2449 mutation) and the CK group (wild-type adults without the mutation), we performed an independent samples *t*-test using the *t*-test() function from the *rstatix* package. Based on the *p*-values, $p < 0.001$ was marked as ***, $p < 0.01$ as **, $p < 0.05$ as *, and other cases were labeled as "ns" (not significant). All statistical analyses were performed using R (v4.1.0).

Data availability

The genome assembly and the raw sequencing data generated in this study, including ONT ultra-long reads, PacBio HiFi reads, Hi-C reads, RNA-seq reads, and full-length transcriptome reads, have been deposited in the NCBI Sequence Read Archive (SRA) under BioProject accession number [PRJNA1237537](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1237537). The gene and transposable element annotation are available via figshare [<https://doi.org/10.6084/m9.figshare.28642964.v3>]. Source data are provided with this paper.

References

- Chen, J. et al. From sub-Saharan Africa to China: evolutionary history and adaptation of *Drosophila melanogaster* revealed by population genomics. *Sci. Adv.* **10**, eadh3425 (2024).
- Prüßing, K., Voigt, A. & Schulz, J. B. *Drosophila melanogaster* as a model organism for Alzheimer's disease. *Mol. Neurodegener.* **8**, 35 (2013).
- Ong, C., Yung, L. Y., Cai, Y., Bay, B. H. & Baeg, G. H. *Drosophila melanogaster* as a model organism to study nanotoxicity. *Nanotoxicology* **9**, 396–403 (2015).
- Pletcher, S. D. et al. Genome-wide transcript profiles in aging and calorically restricted *Drosophila melanogaster*. *Curr. Biol.* **12**, 712–723 (2002).
- Ries, A. S., Hermanns, T., Poeck, B. & Strauss, R. Serotonin modulates a depression-like state in *Drosophila* responsive to lithium treatment. *Nat. Commun.* **8**, 15738 (2017).
- Lloyd, T. E. & Taylor, J. P. Flightless flies: *Drosophila* models of neuromuscular disease. *Ann. N. Y. Acad. Sci.* **1184**, e1–e20 (2010).
- Yamamura, R., Ooshio, T. & Sonoshita, M. Tiny *Drosophila* makes giant strides in cancer research. *Cancer Sci.* **112**, 505–514 (2021).
- Ellison, C. E. & Cao, W. Nanopore sequencing and Hi-C scaffolding provide insight into the evolutionary dynamics of transposable elements and piRNA production in wild strains of *Drosophila melanogaster*. *Nucleic Acids Res.* **48**, 290–303 (2020).
- Kim, B. Y. et al. Highly contiguous assemblies of 101 drosophilid genomes. *Elife* **10**, e66405 (2021).
- Khost, D. E., Eickbush, D. G. & Larracuente, A. M. Single-molecule sequencing resolves the detailed structure of complex satellite DNA loci in *Drosophila melanogaster*. *Genome Res.* **27**, 709–721 (2017).
- Brown, E. J., Nguyen, A. H. & Bachtrog, D. The *Drosophila* Y chromosome affects heterochromatin integrity genome-wide. *Mol. Biol. Evol.* **37**, 2808–2824 (2020).
- Adams, M. D. et al. The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195 (2000).
- Celniker, S. E. et al. Finishing a whole-genome shotgun: release 3 of the *Drosophila melanogaster* euchromatic genome sequence. *Genome Biol.* **3**, research0079 (2002).
- Hoskins, R. A. et al. Sequence finishing and mapping of *Drosophila melanogaster* heterochromatin. *Science* **316**, 1625–1628 (2007).
- Hoskins, R. A. et al. The release 6 reference sequence of the *Drosophila melanogaster* genome. *Genome Res.* **25**, 445–458 (2015).
- Mendiburo, M. J., Padeken, J., Fülöp, S., Schepers, A. & Heun, P. *Drosophila* CENH3 is sufficient for centromere formation. *Science* **334**, 686–690 (2011).
- Lu, X. & Liu, L. Genome stability from the perspective of telomere length. *Trends Genet.* **40**, 175–186 (2024).
- Mason, J. M. & Biessmann, H. The unusual telomeres of *Drosophila*. *Trends Genet.* **11**, 58–62 (1995).
- Pardue, M. L. & DeBaryshe, P. G. *Drosophila* telomeres: two transposable elements with important roles in chromosomes. *Genetica* **107**, 189–196 (1999).
- Chang, C. H. et al. Islands of retroelements are major components of *Drosophila* centromeres. *PLoS Biol.* **17**, e3000241 (2019).
- Li, H. & Durbin, R. Genome assembly in the telomere-to-telomere era. *Nat. Rev. Genet.* **25**, 658–670 (2024).
- Nurk, S. et al. The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
- Yoshimura, J. et al. Recombining the *Caenorhabditis elegans* genome. *Genome Res.* **29**, 1009–1022 (2019).
- Huang, Z. et al. Evolutionary analysis of a complete chicken genome. *Proc. Natl. Acad. Sci. USA* **120**, e2216641120 (2023).
- Hu, J. et al. NextDenovo: an efficient error correction and accurate assembly tool for noisy long reads. *Genome Biol.* **25**, 107 (2024).
- Li, H. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37**, 4572–4574 (2021).
- Hu, J., Fan, J., Sun, Z. & Liu, S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* **36**, 2253–2255 (2020).
- Burton, J. N. et al. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **31**, 1119–1125 (2013).
- Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).
- Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.* **38**, 4647–4654 (2021).

31. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
32. Ritossa, F., Malva, C., Boncinelli, E., Graziani, F. & Polito, L. The first steps of magnification of DNA complementary to ribosomal RNA in *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. USA* **68**, 1580–1584 (1971).
33. Polanco, C., Ana, I. G., Álvaro de la, F. & Dover, G. A. Multigene family of ribosomal DNA in *Drosophila melanogaster* reveals contrasting patterns of homogenization for IGS and ITS spacer regions: a possible mechanism to resolve this paradox. *Genetics* **149**, 243–256 (1998).
34. Bianciardi, A., Boschi, M., Swanson, E. E., Belloni, M. & Robbins, L. G. Ribosomal DNA organization before and after magnification in *Drosophila melanogaster*. *Genetics* **191**, 703–723 (2012).
35. Kaminker, J. S. et al. The transposable elements of the *Drosophila melanogaster* euchromatin: a genomics perspective. *Genome Biol.* **3**, research0084 (2002).
36. Smolka, M. et al. Detection of mosaic and population-level structural variants with Sniffles2. *Nat. Biotechnol.* **42**, 1571–1580 (2024).
37. Heller, D. & Vingron, M. SVIM: structural variant identification using mapped long reads. *Bioinformatics* **35**, 2907–2915 (2019).
38. Jiang, T. et al. Long-read-based human genomic structural variation detection with cuteSV. *Genome Biol.* **21**, 189 (2020).
39. Goel, M., Sun, H., Jiao, W. B. & Schneeberger, K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol.* **20**, 277 (2019).
40. Chang, C. H. & Larracuente, A. M. Heterochromatin-enriched assemblies reveal the sequence and organization of the *Drosophila melanogaster* Y chromosome. *Genetics* **211**, 333–348 (2019).
41. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
42. Long, H. K., Prescott, S. L. & Wysocka, J. Ever-changing landscapes: transcriptional enhancers in development and evolution. *Cell* **167**, 1170–1187 (2016).
43. Dunham, I. et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
44. Bailey, J. A., Yavor, A. M., Massa, H. F., Trask, B. J. & Eichler, E. E. Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res.* **11**, 1005–1017 (2001).
45. Vollger, M. R. et al. Segmental duplications and their variation in a complete human genome. *Science* **376**, eabj6965 (2022).
46. Liu, J. et al. The complete telomere-to-telomere sequence of a mouse genome. *Science* **386**, 1141–1146 (2024).
47. Chakravarti, D., LaBella, K. A. & DePinho, R. A. Telomeres: history, health, and hallmarks of aging. *Cell* **184**, 306–322 (2021).
48. Abad, J. P. et al. *TAHRE*, a novel telomeric retrotransposon from *Drosophila melanogaster*, reveals the origin of *Drosophila* telomeres. *Mol. Biol. Evol.* **21**, 1620–1624 (2004).
49. McCullers, T. J. & Steiniger, M. Transposable elements in *Drosophila*. *Mob. Genet. Elem.* **7**, 1–18 (2017).
50. Meyne, J., Ratliff, R. L. & Moyzis, R. K. Conservation of the human telomere sequence (TTAGGG)_n among vertebrates. *Proc. Natl. Acad. Sci. USA* **86**, 7049–7053 (1989).
51. Fajkus, J., Sýkorová, E. & Leitch, A. R. Telomeres in evolution and evolution of telomeres. *Chromosome Res.* **13**, 469–479 (2005).
52. Lyčka, M. et al. TeloBase: a community-curated database of telomere sequences across the tree of life. *Nucleic Acids Res.* **52**, D311–D321 (2024).
53. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
54. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
55. Carmena, M., Abad, J. P., Villasante, A. & Gonzalez, C. The *Drosophila melanogaster* dodecasatellite sequence is closely linked to the centromere and can form connections between sister chromatids during mitosis. *J. Cell Sci.* **105**, 41–50 (1993).
56. Shi, X. et al. The complete reference genome for grapevine (*Vitis vinifera* L.) genetics and breeding. *Hortic. Res.* **10**, uhad061 (2023).
57. Vollger, M. R., Kerpedjiev, P., Phillippy, A. M. & Eichler, E. E. StainedGlass: interactive visualization of massive tandem repeat structures with identity heatmaps. *Bioinformatics* **38**, 2049–2051 (2022).
58. Carlson, M. & Brutlag, D. Different regions of a complex satellite DNA vary in size and sequence of the repeating unit. *J. Mol. Biol.* **135**, 483–500 (1979).
59. Haas, B. J. et al. Automated eukaryotic gene structure annotation using EvidenceModeler and the program to assemble spliced alignments. *Genome Biol.* **9**, R7 (2008).
60. Shumate, A. & Salzberg, S. L. Liftoff: accurate mapping of gene annotations. *Bioinformatics* **37**, 1639–1643 (2021).
61. 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).
62. Feng, J. et al. The pineapple reference genome: telomere-to-telomere assembly, manually curated annotation, and comparative analysis. *J. Integr. Plant Biol.* **66**, 2208–2225 (2024).
63. Kapila, R., Kashyap, M., Poddar, S., Gangwal, S. & Prasad, N. G. G. Evolution of pathogen-specific improved survivorship post-infection in populations of *Drosophila melanogaster* adapted to larval crowding. *PLoS ONE* **16**, e0250055 (2021).
64. Gao, Y. et al. Genetic dissection of active forgetting in labile and consolidated memories in *Drosophila*. *Proc. Natl. Acad. Sci. USA* **116**, 21191–21197 (2019).
65. Bourque, G. et al. Ten things you should know about transposable elements. *Genome Biol.* **19**, 199 (2018).
66. Batzer, M. A. & Deininger, P. L. Alu repeats and human genomic diversity. *Nat. Rev. Genet.* **3**, 370–379 (2002).
67. Yang, H. P. & Barbash, D. A. Abundant and species-specific DINE-1 transposable elements in 12 *Drosophila* genomes. *Genome Biol.* **9**, R39 (2008).
68. Chakraborty, M. et al. Hidden genetic variation shapes the structure of functional elements in *Drosophila*. *Nat. Genet.* **50**, 20–25 (2018).
69. Chakraborty, M., Emerson, J. J., Macdonald, S. J. & Long, A. D. Structural variants exhibit widespread allelic heterogeneity and shape variation in complex traits. *Nat. Commun.* **10**, 4872 (2019).
70. Courret, C. & Larracuente, A. M. High levels of intra-strain structural variation in *Drosophila simulans* X pericentric heterochromatin. *Genetics* **225**, iyad176 (2023).
71. Hallem, E. A., Ho, M. G. & Carlson, J. R. The molecular basis of odor coding in the *Drosophila* antenna. *Cell* **117**, 965–979 (2004).
72. Masse, N. Y., Turner, G. C. & Jefferis, G. S. Olfactory information processing in *Drosophila*. *Curr. Biol.* **19**, R700–R713 (2009).
73. Kursel, L. E. & Malik, H. S. Centromeres. *Curr. Biol.* **26**, R487–R490 (2016).
74. Melters, D. P. et al. Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biol.* **14**, R10 (2013).
75. Altemose, N. et al. Complete genomic and epigenetic maps of human centromeres. *Science* **376**, eabl4178 (2022).
76. Lohe, A. R., Hilliker, A. J. & Roberts, P. A. Mapping simple repeated DNA sequences in heterochromatin of *Drosophila melanogaster*. *Genetics* **134**, 1149–1174 (1993).

77. Casacuberta, E. & Pardue, M. L. Transposon telomeres are widely distributed in the *Drosophila* genus: TART elements in the virilis group. *Proc. Natl. Acad. Sci. USA* **100**, 3363–3368 (2003).
78. Casacuberta, E. & Pardue, M. L. HeT-A and TART, two *Drosophila* retrotransposons with a bona fide role in chromosome structure for more than 60 million years. *Cytogenet Genome Res.* **110**, 152–159 (2005).
79. Yoshitake, Y., Inomata, N., Sano, M., Kato, Y. & Itoh, M. The P element invaded rapidly and caused hybrid dysgenesis in natural populations of *Drosophila simulans* in Japan. *Ecol. Evol.* **8**, 9590–9599 (2018).
80. O'Grady, P. M. & DeSalle, R. Phylogeny of the genus *Drosophila*. *Genetics* **209**, 1–25 (2018).
81. Horn, M. et al. The circadian clock improves fitness in the fruit fly, *Drosophila melanogaster*. *Front. Physiol.* **10**, 1374 (2019).
82. Capek, M. et al. Evolution of temperature preference in flies of the genus *Drosophila*. *Nature* **641**, 447–455 (2025).
83. Bender, W. et al. Molecular genetics of the bithorax complex in *Drosophila melanogaster*. *Science* **221**, 23–29 (1983).
84. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
85. Wu, H. et al. Telomere-to-telomere genome assembly of a male goat reveals variants associated with cashmere traits. *Nat. Commun.* **15**, 10041 (2024).
86. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
87. Hu, G. et al. A telomere-to-telomere genome assembly of cotton provides insights into centromere evolution and short-season adaptation. *Nat. Genet.* **57**, 1031–1043 (2025).
88. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
89. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
90. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Giga-science* **10**, giab008 (2021).
91. Wang, X. & Wang, L. GMATA: an integrated software package for genome-scale SSR mining, marker development and viewing. *Front. Plant Sci.* **7**, 1350 (2016).
92. Han, Y. & Wessler, S. R. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res.* **38**, e199 (2010).
93. Flynn, J. M. et al. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA* **117**, 9451–9457 (2020).
94. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268 (2007).
95. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinforma.* **9**, 18 (2008).
96. Ou, S. & Jiang, N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* **176**, 1410–1422 (2018).
97. Keilwagen, J., Hartung, F. & Grau, J. GeMoMa: homology-based gene prediction utilizing intron position conservation and RNA-seq data. *Methods Mol. Biol.* **1962**, 161–177 (2019).
98. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
99. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
100. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637–644 (2008).
101. Zdobnov, E. M. & Apweiler, R. InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**, 847–848 (2001).
102. Lowe, T. M. & Chan, P. P. tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Res.* **44**, W54–W57 (2016).
103. Griffiths-Jones, S. et al. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.* **33**, D121–D124 (2005).
104. Lagesen, K. et al. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* **35**, 3100–3108 (2007).
105. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
106. Marçais, G. et al. MUMmer4: a fast and versatile genome alignment system. *PLoS Comput. Biol.* **14**, e1005944 (2018).
107. Jeffares, D. C. et al. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat. Commun.* **8**, 14061 (2017).
108. Zhou, Y. et al. Metascope provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat. Commun.* **10**, 1523 (2019).
109. Yu, G., Wang, L. G. & He, Q. Y. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* **31**, 2382–2383 (2015).
110. Išerić, H., Alkan, C., Hach, F. & Numanagić, I. Fast characterization of segmental duplication structure in multiple genome assemblies. *Algorithms Mol. Biol.* **17**, 4 (2022).
111. Krzywinski, M. et al. Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).
112. McKenna, A. et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
113. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
114. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
115. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
116. Lu, T. C. et al. Aging fly cell atlas identifies exhaustive aging features at cellular resolution. *Science* **380**, eadg0934 (2023).
117. Shuai, Y. et al. Forgetting is regulated through Rac activity in *Drosophila*. *Cell* **140**, 579–589 (2010).
118. Tully, T. & Quinn, W. G. Classical conditioning and retention in normal and mutant *Drosophila melanogaster*. *J. Comp. Physiol. A* **157**, 263–277 (1985).
119. Beshel, J. & Zhong, Y. Graded encoding of food odor value in the *Drosophila* brain. *J. Neurosci.* **33**, 15693–15704 (2013).

Acknowledgements

The authors thank Dr. Yi Zhong for the fly stock, Dr. Qian Li, Dr. Bin-Yan Lu, and Dr. Yi-Bo Luo for advice, Zhifan Guo, Jinxia Luo, and Bin Zuo for their technical support. The project was supported by Yunnan Fundamental Research Projects (202401BC070011) and the National Key Research Development Program of China (2022YFF0802300).

Author contributions

Y.B.S. and D.D.W. conceived the study and developed the overall research plan. Y.N.L., Y.B.S., and D.D.W. designed the experiments and the methodology. Y.N.L. and X.L.Z. performed data analysis. J.J.G. collected samples. Y.N.L. and J.J.G. performed the experiments. Y.N.L. and Y.B.S. wrote the manuscript. Y.B.S., D.D.W., Y.N.L. and X.L.Z. revised the manuscript. All authors have read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-025-67031-w>.

Correspondence and requests for materials should be addressed to Dong-Dong Wu or Yan-Bo Sun.

Peer review information *Nature Communications* thanks Jun Kim, Susan Celniker and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. 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