nature portfolio

# Peer Review File

# Near Complete Assembly of Drosophila melanogaster Canton S strain genome

Corresponding Author: Professor Yan-Bo Sun

**This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.**

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
Review of, “Near Complete Genome Assembly of Drosophila Unveils Novel genes and Complex Heterochromatin Architecture” for Nature Communications

Liu et al., describe a new genome assembly of a strain of Drosophila melanogaster, Canton S collected in Beijing. The sequence is derived from the male progeny of a single pair mating. The authors used a new hybrid strategy called telomere-to-telomere (T2T) which includes PacBio, Oxtford Nanopore, Illumina paired-end sequencing and HiC and two assemblers Canu and NextDenovo followed by genome polishing with minimpa2 and Nextpolish. They generated 30 contigs and used Hi-C data to place 25 of the scaffolds onto the four pairs of chromosomes; three autosomes and one pair of sex chromosomes. The second and third chromosomes are metacentric with left and right arms, the fourth is a dot acrocentric chromosome, the X is a large acrocentric and the Y is submetacentric. Liu et al., have expanded the genome by 17.93 Mb. In addition to the genome assembly, they used existing RNA-seq (cited as being from the ENA database) to generate an experimentally derived transcriptome. The genome sequence and transcriptome were then used to produce a genome annotation that was compared to the D. melanogaster Release 6 ISO1 reference strain sequence to identify new genes. According to their abstract, they identified 78 novel protein-coding genes. Two of these were studied genetically using the CRISPR/Cas9 system to generate mutants. One of these was highlighted in the abstract as they determined it to be a new olfactory gene which when mutated has a reduced sensitivity to smell. With the major and minor concerns properly addressed, the manuscript should be re-reviewed to determine whether it will be accepted for publication or not.

The paper will be valuable to a number of communities. Major and minor concerns are described below.

A major concern is that Liu et al., did their T2T studies with a variant of Canton S, native to Beijing and not the sequenced reference ISO1 strain. It is well known from the earliest cytological (PMID: 4207116) and sequence studies (e.g. PMID: 17737996; PMID: 29255259) that strain differences result in very different genome sequence which affects gene numbers and heterochromatin size. The strains and sequence differences merit discussion in the introduction.

Also, the strain should be given an identification number and deposited with a number of stock centers (e.g. Bloomington) so these flies are available to the community so these studies can be repeated.

For their RNA studies they reference ENA which is the European Nucleotide Database and as the source of the data, yet in the Supplemental Table S7 it appears that the sequences are from NCBI and have SRR numbers. It is important for these studies to include the strains as an additional column in Supplemental Table S7 since the RNA samples appear to come from many different studies. Furthermore, all the numbers in the manuscript need to be reviewed and reconciled, especially those that describe new genes (e.g. abstract 78; Fig 4. 96; other instances 217; 114; 105 and in the discussion is reduced to “several”) and functionally validated genes (e.g. Figure 4, 3; line 425, 2 and abstract only one). The sequence of these new genes should be included as a Supplemental Figure. They could also include information on paralogs and orthologs.

Finally, the BioProject, ID PRJCA033379, is not yet public and according to the website https://ngdc.cncb.ac.cn/search/specific?db=bioproject&q=PRJCA033379 will not be available until December 7th of 2026. It should be released for review and before publication.

Minor concerns to be addressed before publication are listed below.

Minor Issues

1. Line 124. There are not “7 major chromosomes” in Drosophila. There are 4 pairs (described above). Please correct all instances. The chromosome numbers are not 0 padded please correct this as well (e.g. Chr2, not Chr02L)

2. Lines 135-136. The figure legend for Figure 1 needs to be corrected –Fig 1C defines Fig 1D and Fig1D defines Fig1C.

3. Line 135. Error types (1- 4) should be defined the first time they are used. Currently described lines 388-390.

4. Line 143. Report the genes not recalled in a Supplemental Table. Describe why they weren’t recalled.

5. Lines 147- 152. Figure 2 needs to be corrected, A-D misordered. What is labeled B is described in D. What is labeled D is described in B.

6. Line 148. Change “highly” to “high”.

7. Line 167-177. What QC was performed to make sure no contaminating sequences were introduced that could account for the SINE elements?

8. Line 187 Change “(Figure 1B)” to (Figure 1B; Sup Table 2). To Supplemental Table 2 add the Rel 6 numbers and the differences as two separate columns.

9. Line 204-208. Change “B) Ideogram of density distribution of different type of SVs on chromosome of the genome assembly. (C) The size of different type of SVs processed by logarithm.” to “(B) Ideogram of density distribution of different types of SVs on chromosomes of the genome assembly. (C) The logarithmic scale size of different types of SVs.”.

10. Line 211. When they discuss structural variants, they should mention strain differences as the possibility for the variation.

11. Lines 230-231. The reference sequence was assembled and compared with restriction digests of individual BAC clones mapped to the assembly. In addition, the reference was compared to an optical map. These two strategies confirmed the reference assembly order and orientation. What have the authors done to independently confirm their assembly? It would be wise to be careful about suggesting Rel6 has misassemblies without specific cases being confirmed in the proper reference strain.

12. Line 275. IGV should be spelled out - Integrated Genome Viewer. Please define this acronym the first time it is used.

13. Line 327. Change “12,340.5 copies” to “an average of 12,340 copies”. Not clear what .5 copies represents.

14. Lines 328-329. Since the authors are discussing the centromeric end of the X chromosome, the sentence “This repeat is located at the right end of the X chromosome.” is redundant and could be deleted.

15. Line 346-347 Change “the 4 chromosome” to “the 4th chromosome”.

16. Line 381-383. Change “We identified four main types of mis-annotations:” to “In our automated computationally derived annotation, we identified four main types of misannotations:”.

17. Line 390-398. Are the 114 genes homologs (truly missing from the reference genome) or paralogs (where there are many copies of which one is missing)?

18. Line 500. add a reference to the sentence “However, Drosophila telomeres are composed of three different transposons: HeT-A, TART and TAHRE, all of which belong to the retrotransposon class”, such as reviewed in Pardue and DeBaryshe 1999. In general, many background citations are missing and should be added when they describe work that is not their own.

19. Line 503. Clarify the sentence “We also found that this telomere pattern is absent in the Drosophila genus”. Needs a reference. Not clear what “this telomere pattern” refers to the human type or the Drosophila type. Also there are no experiments in this study where you describe other members of the Drosophila genus which has 1600 species.

20. Line 507. Change “that The P-element” to “that the P-element”.

21. Lines 497 – 512. The discussion about telomeres as it stands is not merited and should be deleted. How much have they improved the sequence? This should be described.

22. Line 531. Materials and Methods. Make sure to include all software versions when software is mentioned, also cite the primary references or the URL. Examples of software with missing citations are listed below but is not comprehensive.

23. Lines 544. Change “After the sample was qualified, size-select of long DNA fragments were performed using the BluePippin system” to “After the sample was quantified, sizeselection of long DNA fragments was performed using the BluePippin system”.

24. Line 562. Change “fragments were screened by the BluePippin” to “fragments were screened using the BluePippin (Sage Science, USA) system”.

25. Line 570. What version of Guppy was used? What Adapter trimming parameters were used? Were reads filtered for size?

26. Line 585. Incomplete sentence “And Illumina short reads were (what belongs here?) using Nextpolish”.

27. Line 590. Change “controlling” to “control”

28. Line 591. Clarify what you mean by “as former research” and add a reference.

29. Line 599. Delete the second instance of scaffolds. From - “The scaffolds were further clustered, ordered, and oriented scaffolds” to “The scaffolds were further clustered, ordered, and oriented”.

30. Line 611. Replace (ref) with the actual citation.

31. Line 616 -618. Please clarify the following sentence “. Besids, base accuracy of the assembly was calculated with bcftools.”

32. Line 618. Add a reference for the LTR Assembly Index calculation.

33. Line 624. Add a reference for the GMATA and Tandem Repeats Finder software.

34. Line 626-7. Correct spelling and italicize ab initio.

35. Line 627. Correct A to a in the sentence starting “Briefly”.

36. Line 628. Add references for MITE-hunter and RepeatModeler.

37. Line 642. Add a reference for the GeMoMa software.

38. Line 643. Add a reference for the Liftoff software.

39. Line 646. Add a reference for the STAR software.

40. Line 650. Add a reference for the AUGUSTUS software.

41. Line 652. Add a reference for the EVM software.

42. Line 653. Add the version and the reference for InterProScan

43. Line 667. Add the version and the reference for tRNAscan-SE

44. Line 668-9. Add the versions and references for Rfam database and RNAmmer

45. Line 679. Replace “using TRF2GFF (reference)” with the actual citation.

46. Line 687. Replace “MACS2 (reference, version)” with the citation and version.

47. Line 713. Include the accession numbers for the ATAC-seq datasets or the Sup Table.

48. Line 749. How was the stable mutant Chr02L.722 homozygous line validated?

49. Line 755-6. How was the stable mutant Chr03R.2449 homozygous line validated?

50. Line 760. We selected 5-day-old Chr03R.2449 mutant Drosophila (experimental group) and non-knockout Drosophila (control group). Were these males, females, larvae, adults? What does the 5-day-old refer to?

51. Line 762. “Based on previously published studies” needs a reference.

52. Line 764-5. What are the units of the OCT solution?

53. Line 781. “To compare the differences in the Index values between the CK and gene2 groups.” Define these groups

Reviewer #2

(Remarks to the Author)
Major Comments
1. Genome assembly is not publicly available and has not been deposited in an appropriate database.
This study emphasizes the superior quality of the newly assembled genome compared to the previous reference genome. The raw sequencing data have been deposited in the National Genomics Data Center (NGDC) of China, allowing researchers to access the data to validate or reproduce this study. However, the most crucial dataset—the genome FASTA file and annotation files—is missing from the database. The authors should publicly release these files to ensure the reproducibility and accessibility of their findings.

Additionally, since Nature Communications is part of the Nature portfolio, it is highly recommended to follow their data-sharing policies. Relevant guidelines can be found here:

- "Reporting standards and availability of data, materials, code and protocols"
https://www.nature.com/nature-portfolio/editorial-policies/reporting-standards#availability-of-data
This page states:
"... the preferred approach is to make data available in repositories. Scientific Data, a Nature Portfolio journal, maintains a list of approved and recommended data repositories to support researchers seeking suitable repositories for their data."

- "Data Repository Guidance"
https://www.nature.com/sdata/policies/repositories
This document specifies that "Raw sequencing data (reads or traces), Genome assemblies, Annotated sequences, and Sample metadata" must be publicly available in recognized repositories such as INSDC repositories or the Genome Sequence Archive (GSA). Currently, some of these datasets remain inaccessible, making it impossible to evaluate their quality.

Furthermore, since the raw sequencing data have been uploaded to NGDC, it is necessary to clarify whether NGDC is part of the International Nucleotide Sequence Database Collaboration (INSDC). If not, the dataset must be made available in NCBI GenBank (for genome assembly) and NCBI SRA (for raw sequencing reads) or another recognized repository.
https://www.insdc.org/

2. Issues in updating the genome
It is unclear how the final genome assembly will be utilized.

To illustrate the importance of maintaining previous reference genomes, an analogy can be drawn from human genome research. The human reference genome, GRCh38, originally produced using Sanger sequencing, remains in use despite the advent of long-read sequencing technologies. A vast body of research and extensive annotation data have accumulated based on GRCh38. Consequently, even though large-scale projects now employ long-read sequencing to refine the human genome or construct a human pangenome, GRCh38 cannot simply be discarded. Instead, it continues to serve as the backbone while additional long-read genome assemblies are incorporated into a graph-based reference genome.

A similar approach seems necessary for Drosophila melanogaster. The existing annotations cannot be discarded and must be retained. Given this, the present study should not merely provide a new genome assembly and annotations for the same Canton-S strain, but rather integrate:

- the existing Canton-S strain reference genome,
- the newly generated Canton-S strain genome assembly, and
- previously available genome assemblies from other strains

into a graph-based genome reference. This would ensure that the updated reference genome maintains compatibility with prior annotations.

If full integration is too complex (although I do not believe it is), the study must at least transfer all previous gene annotations and other functional annotations to the new genome using a liftover approach. Without either of these efforts, it is questionable whether the dataset produced in this study will be widely adopted by the research community.

3. Data and assembly quality should be analyzed much more thoroughly.
Currently, it is difficult to assess the quality of the dataset, raising concerns about its suitability for publication in Scientific Data. The following additional analyses should be included:

(1) Read-level quality assessment
- Read length and base quality distributions should be provided. For example, a 2D contour plot with read length on the x-axis and mean read quality on the y-axis should be included.
- For HiFi and ONT, these data should be included.
- Since the sequencing platform is not clearly stated, it is difficult to determine whether quality adjustments are necessary. However, if a PacBio platform prior to Revio was used, phred scores should be corrected—otherwise, an excessive number of values exceeding 40 will appear. The corrected values should be plotted so that all scores remain ≤40.
(2) Assembly validation and error checking

- Scaffolding errors introduced during de novo assembly must be addressed before gap filling. It is well known that Hi-C alone frequently introduces scaffolding errors.
- Example: In Figure 1A, read depth distributions for HiFi and ONT do not appear evenly distributed. If this is accurate, it suggests poor genome quality in many regions. I suspect that telomeric and/or centromeric regions were not properly assembled. The authors should thoroughly analyze high- and low-depth regions.
(3) Structural variant validation
- A comparative structural variant analysis should be conducted between the previous reference genome and the new genome using SyRi or similar tools.
- Example: Translocations and inversions detected by SyRi should be examined. If scaffolding errors are present, structural variants are likely to be detected near contig junctions (i.e., gap regions). A quantitative analysis should be provided to assess the number of clear scaffolding errors in both the existing reference genome and the newly generated genome.
(4) Base-level assembly accuracy
- Quality value (QV) and k-mer completeness should be assessed.
- Merqury can be used to compare the k-mer distribution of short-read WGS data against the genome assembly, allowing QV and k-mer completeness to be calculated.
- Lines 140–141: "Quality evaluation using Merqury indicated robust completeness and high quality (QV) for the assembly (Supplementary Figure 2)." → The exact values of k-mer-based completeness and QV must be explicitly provided in the manuscript. Supplementary Figure 2 only contains a visual plot without numerical values.
4. Additional analyses are required.
With long-read sequencing, the authors can analyze methylation profiles of this fly sample. Additionally, segmental duplications (SDs) and other complex genomic rearrangements should be thoroughly analyzed. Key questions:
- How many SDs exist in this genome and across Drosophila populations?
- What are the SD lengths?
- Are there differences in intra- vs. inter-chromosomal SDs?
- Are SDs prone to mutations?

Human genome studies can provide insights into these aspects:
Segmental duplications and their variation in a complete human genome
https://www.science.org/doi/10.1126/science.abj6965
Increased mutation and gene conversion within human segmental duplications
https://www.nature.com/articles/s41586-023-05895-y
Structural polymorphism and diversity of human segmental duplications
https://www.nature.com/articles/s41588-024-02051-8

5. Many critical typos and conceptual errors
There are numerous typos and conceptual errors, including in the title. The term heterochromatin architecture is used, yet no evidence is provided for heterochromatin features such as chromatin accessibility or histone profiles. The title should be corrected.

A detailed list of errors and suggested corrections follows. (See original document for specifics.)

This manuscript requires intensive revision.

5. Many critical typos and conceptual errors
This manuscript contains numerous typos and conceptual errors. For example, the title includes "heterochromatin architecture", yet the authors do not provide any supporting evidence, such as chromatin accessibility data or histone profile results. Furthermore, it is unclear which genomic regions the authors emphasize under this term.

Is the focus on centromeric regions, or does it refer to other previously uncharacterized regions within the genome?
If the authors aim to highlight telomeric, subtelomeric, or centromeric regions, terms such as highly repetitive regions, complex regions, or similar descriptors would be more appropriate than heterochromatic regions.
Therefore, the title should be revised to accurately reflect the content of the manuscript.

Additionally, the manuscript contains an unacceptable number of typos and misleading concepts. Some errors are so severe that they cast doubt on whether the manuscript has undergone proper proofreading. These must be corrected before submission.

(1) Terminology issues
- The term "telomere" refers to the telomeric repeat sequence + telomeric binding proteins. In most cases within this manuscript, "telomeric region" would be a more appropriate term than "telomere".
- Similarly, "centromere" should be replaced with "centromeric regions".

(2) Typo and error list
Line 31: Despite → Despite of
Line 37: telomeres and centromeres regions → telomeric and centromeric regions
Line 62: release 2, 3, 4, 5were → releases 2, 3, 4, and 5 were
Line 67: species evoluation, developmental biology → species evolution and developmental biology
Line 83: quality of genome sequencing → quality of the genome (This sentence refers to assembled genomes, not

sequencing methods.)
Line 85: Caenorhabditis elegans should be included as an important reference.
Reference: Recompleting the Caenorhabditis elegans genome
https://pubmed.ncbi.nlm.nih.gov/31123080/
Line 88: repeat sequence families → repetitive sequence families
Line 92: Amphibalanus Amphitrite → Amphibalanus amphitrite (Incorrect capitalization; a major taxonomic error.)
Line 93: barnacles. → Missing reference
- Additionally, in this context, discussing barnacle genomes seems unnecessary. Instead of this example, the authors should discuss newly identified genes from the human T2T genome.
- Since Amphibalanus amphitrite was not assembled using a T2T approach on a single individual, it is not an appropriate example for this discussion.
- Similarly, as done in human genome research, the study should analyze genes identified through segmental duplications (SDs) or other newly assembled genomic regions.
(3) Methodology issues
Line 119: Canu → HiCanu? (Specify if HiCanu was used, as this affects the assembly process.)
Line 122: Manually curation → Manual curation
- What were the criteria for filtering contigs?
- Which aspects were considered for manual curation? The manuscript must provide detailed information on the filtering process.
Line 122: Hi-C contact map should be included in the main figure, rather than in the supplementary materials.
(4) Issues in figures and data representation
Figure 1A: Each circle should be numbered to improve readability.
Line 131: 50kb → 50-kb
Line 133:
PacBio HiFi data sequencing depth and ONT data sequencing depth;
→ PacBio HiFi and ONT UL read depths. (The original phrasing is redundant.)
Figure 1B:
- The y-axis should start at zero.
- Alternatively, normalizing by initial genome size would improve the visualization.
Line 135: melanogaster; → melanogaster. (Unnecessary semicolon)
Figure 1C:
- The figure does not represent "type I errors" as stated in the legend.
- Instead, it shows the composition of repetitive sequences.
- The legend of Figure 1C should be replaced with the legend for Figure 1D.
- The figure should compare the current genome with previous versions of the fruit fly genome.
Figure 1D:
- The figure lacks a legend and explanation.
- The purpose of the figure is unclear. A detailed description should be added.
Line 147: ventromeric → centromeric (The term "ventromeric" does not exist in genomics.)

Final remarks
The sheer number of typos and conceptual errors in this manuscript is concerning. While minor typographical errors are expected in any draft, the extent of errors here suggests a lack of careful proofreading and insufficient attention to detail.

At this stage, the manuscript requires extensive revision before it can be considered for publication.

Reviewer #3

(Remarks to the Author)
The goal of this study is to generate a near telomere-to-telomere (T2T) genome assembly for Drosophila melanogaster and to update the regulatory and genic content associated with the new assembly. To this end, the authors combine long-read sequencing data (Oxford Nanopore and PacBio), Illumina short reads, and Hi-C chromatin conformation data. The proposed assembly adds approximately 18 Mb of new sequence and resolves ~250 gaps relative to the D. melanogaster reference genome assembly (Release 6, or R6, generated in 2014). Most of the new sequence corresponds to the (peri)centromeric region of the X chromosome (+11.4 Mb), with additional sequences added to chromosome 4 (+2.1 Mb) and the pericentromeric region of chromosome arm 3R (+2.5 Mb). Notably, the new Y chromosome assembly proposes numerous segmental inversions relative to the R6 reference.

With this updated assembly, the authors investigate the distribution of repetitive elements (transposable elements and short repeats), describe general properties of telomeric and centromeric regions, and present an updated gene annotation. This gene annotation identifies 217 new or refined gene structures, including 105 genes with no known orthologs outside D. melanogaster. The authors also analyze ATAC-seq data, suggesting the presence of previously uncharacterized regulatory regions. Finally, they use CRISPR/Cas9 editing to validate the biological relevance of two newly annotated genes.

Overall, the most significant contribution of this study is the presentation of a near-T2T assembly for D. melanogaster, with expanded genic and repeat annotations. The new assembly successfully resolves many gaps present in the R6 reference assembly and adds a substantial amount of (peri)centromeric sequence. However, the total size of the assembly (~161 Mb) falls short of a true T2T or near-gapless genome, which is estimated to be ~215 Mb in males—suggesting that ~50 Mb

remains unassembled. Additionally, while comparisons are consistently made against R6, the study omits comparison with several more recent long-read-based assemblies, which are known improvements over R6. Furthermore, the Hi-C data used to scaffold contigs in repetitive regions may be compromised due to limitations in the mapping strategy, raising concerns about the reliability of the added (peri)centromeric sequences.

Major Concerns:
1. Assembly Quality and Comparisons:
The manuscript highlights that the new assembly is larger than R6, but fails to reference or compare it against other high-quality long-read assemblies (e.g., Miller et al., 2018; Chang and Larracuente, 2019; Ellison et al., 2020; Kim et al., 2021; Bologa et al., 2022). These studies already include improvements such as gap closure, expanded (peri)centromeric regions, and enhanced Y chromosome assemblies. In particular, the study by Chang and Larracuente is directly relevant because it also provides new gene and repeat annotations, and it should be used as a comparison point.

2. Assembly Size vs. Expected Genome Size:
The expected haploid genome size is ~215 Mb in males (~175 Mb in females) while the proposed new assembly is ~161 Mb (for males), indicating that large regions remain unassembled.

3. Hi-C Data and Reliability of Scaffolding:
The manuscript reports using Hi-C data to scaffold contigs in repeat-rich regions. However, uniquely mapped read pairs are essential for reliable Hi-C scaffolding in these areas. The authors claim to use a Bowtie2 command (L595) to retain uniquely mapped reads, but Bowtie2 does not offer such a parameter; custom filtering is required. The absence of uniquely mapped reads likely undermines the reliability of the Hi-C scaffolding, particularly for heterochromatic regions. Indeed, the Hi-C contact map (Supplementary Figure 1) lacks robust diagonal signals for centromeres and the Y chromosome, suggesting the scaffolding is unreliable.

4. Y Chromosome Size and Quality:
While the R6 genome includes ~4 Mb of assembled Y chromosome sequence, this study extends it modestly (~7 Mb). However, Chang and Larracuente (2019) previously assembled ~14.6 Mb of the Y chromosome using a targeted male-specific sequencing strategy that is more sensitive to repetitive elements. The present study does not represent an improvement over this prior work.

5. Y Chromosome Structure Discrepancies:
The new Y chromosome assembly proposes a substantial reorganization relative to R6. However, Chang and Larracuente (2019) provided a Y chromosome assembly for ISO-1 that aligns well with both the R6 reference and cytological data. The authors should reassess their findings in light of this prior work. Furthermore, concerns about Hi-C reliability call into question the accuracy of the Y chromosome assembly.

Additional Concerns:

6.Gene Annotations:
6.1) “Release 6 plus ISO1” (L369) refers to the genome assembly, not a specific annotation release. The current gene annotation for R6 is r6.62 (FB2025_01), and newer annotations have added genes and transcripts. The authors must specify which version they used and compare against the most recent release.
6.2) When discussing new gene annotations, it would be helpful to know which ones are associated with newly assembled sequences (e.g., pericentromeric X and 3R) and which ones refine previously assembled regions.
6.3) Table S5 lists 72 newly defined genes (text mentions 71). A supplementary table should also list genes with updated structures.
6.4) The authors validate two genes using CRISPR/Cas9. For the edited gene Chr02R.722, changes in transcriptional profiles are shown. For the edited gene Chr03L.2449, behavioral assays are presented, but transcriptional effects are not. Please include data on whether editing affects this gene’s expression or causes genome-wide transcriptional changes.

7. Resolution of Known Gaps:
It remains unclear whether known problematic regions (e.g., the histone gene cluster on 2L, rDNA arrays on the X and Y) have been resolved. If so, these regions should be described.

8. Preliminary Assemblies:
The manuscript mentions using Canu and NextDenovo for preliminary assemblies but does not clarify whether these tools produced consistent results or how decisions were made on which contigs to retain. Please include this information.

9. Pre- and Post-Hi-C Assembly Comparison:
The study should present a comparative view of the assembly before and after Hi-C scaffolding, especially for the pericentromeric region of the X and the Y chromosome.

10. Telomeric Elements:
Canonical telomeric elements (HET-A, TART, TAHRE) are included in R6. The current study adds limited new information on these sequences due to a lack of detailed analyses.

11. Telomeric Sequences in R6:
Contrary to L256–259, R6 does include telomeric sequences. The authors should correct this statement and expand their

analysis of telomeres, including potential differences in length and composition across chromosome ends or between assemblies.

12. Discussion Section:
The Discussion section is particularly underdeveloped and contains several factual errors. Telomeric repeats (L498-499) are not AATAT (animals) or CCATA (plants); rather, common repeats in animals include TTAGG (TTAGGG in humans) whereas TTTAGGG is common on plants. Retrotransposons at telomeres are not unique to D. melanogaster (as mentioned in L502-509), and retrotransposon-based telomere structures were already known and present in R6. The study does not "resolve" centromeres or telomeres (L481)—these regions are expanded but remain incomplete.

13. Cross-Species Gene Annotation:
Provide version numbers, accession IDs, and links for gene annotations used from D. yakuba, D. simulans, D. sechellia, D. erecta, and D. ananassae.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
Liu et al., in their revised version of “Near Complete Genome Assembly of Drosophila melanogaster Unveils Novel Genes and Complex Genomic Regions” have successfully addressed all my concerns and questions. The manuscript is now worthy of publication.

Reviewer #2

(Remarks to the Author)
The authors have addressed most of my questions (though not all) adequately. In my view, the manuscript has been improved and is now more clearly written. I believe it can make a valuable contribution to the field. I hope that future genome assembly studies based on this work will further serve as a resource for the scientific community.

Reviewer #3

(Remarks to the Author)
I would like to thank the authors for their efforts improving the manuscript. I have, however, a few comments and concerns (see below) that I believe should be addressed before final acceptance.

#1. Previous D. melanogaster assemblies based on long-read sequences were not mentioned.
The authors have a detailed and sensible response to my concern, “… However, we acknowledge that meaningful comparisons with other state-of-the-art long-read assemblies are essential to underscore the unique value of our work. To address your concern, we carefully examined and compared our genome assembly with previously published long-read D. melanogaster assemblies”. Indeed, many of these assemblies had already improved the R6 assembly used here as reference. An important aspect of the authors’ response is a table showing the main properties of previous long-read-based assemblies, and those from this nT2T. This allows a reader to appreciate the advances made relative to R6 before this study, and the additional advances of the proposed nT2T presented here.

This informative Table, however, is only shown in the ‘response to reviewers’, but not included in the revised version of this manuscript. I believe it should be included as Table or as Supplemental data. Perhaps the best place to describe previous assemblies and this table would be after line 85, where the authors indicate that long-reads, sometimes with Hi-C, have been used to generate close-to-T2T in several organisms.

#2. rDNA clusters.
The authors describe a cluster localized on the ChrX (lines 133-134) and then write (line 135), “Notably, we also detected high-confidence matches for these rDNA units on ChrY, a region where the R6 genome lacks such annotations, indicating a significant enhancement in rDNA resolution.” As written, a reader could infer that the presence of an rDNA cluster in the Y chromosome was an unexpected finding. Although it is true that R6 does not have rDNA on its short Y chromosome sequence, the presence (and some key characteristics) of an rDNA cluster in the Y chromosome has been known for many years, with a predicted size of 2.2Mb (Bianciardi et al. 2012; Ritossa 1976; Polanco et al. 1998). Adding these previous studies would add context and help a reader to appreciate better the new data (we knew the cluster there and its size, now we know its sequence…).

#3. Telomeric sequences and variation between strains.
The authors indicate (lines 295-196) that the telomeric regions of this nT2T (297kb) are markedly larger than those in R6 (74.3 kb), arguing that many contigs that look like telomeric in R6 “remain in unplaced contigs”. If the authors argument is correct, one cannot trust the telomeric sequences presented in R6 and the order of their components is unreliable. As a consequence, the authors’ proposal (lines 577-581) that the telomere structural differences “more likely reflect genuine polymorphisms between strains:.”, is unwarranted at this point and should be reassessed, or removed. Either R6 is reliable (and comparisons to nT2T can reveal variation between strains and polymorphism), or it is not (in which case apparent

differences cannot infer true polymorphism).

#4. SINE elements in D. melanogaster.
The authors write (line 478), “A striking discovery was the identification of SINE transposable elements (0.02% of the genome), previously unreported in D. melanogaster.” This result can be important but I would like to request the authors to provide some key information. First, please report the consensus sequence of this D. melanogaster SINE to allow the community to expand this report in other strains and species (as it is now it is not possible). Second, SINE-like elements have been reported in D. melanogaster in the past, but without much follow-up. Please discuss possible overlap of the proposed SINE with DINE-1 (Kaminker et al. 2002) and/or suffix (Tchurikov and Kretova 2007) elements.

#5. Annotation.
The authors have included the new genome assembly (fasta) and annotation (gff) files (accessed via figshare). Unless I missed it, I could not find annotations (location and sequence) of any transposable element (including SINEs) across the new nT2T genome sequences. Please, include the annotation of the transposable elements reported in this study.

#6. Hi-C and repetitive sequences.
Please, include the command lines used to identify uniquely mapped Hi-C reads, which is essential for reliable Hi-C analyses of repetitive sequences (centromeres and the Y chromosome). As written, the methods do not describe how this was achieved.

Other:
- Line 55. “harbors” might be the wrong word. “However, despite ongoing studies seeking to improve genome completeness of this classic model organism, its genome still harbors the structurally complex regions”
- The Canton-S strain used in this study is said to come from the Bloomington Drosophila Stock Center (BDSC_64349). A minor issue is that this strain was replaced on Feb 2016. Please indicate when the strain was obtained from BDSC.
- Line 108. “… the average read length was 26,487.9 bp, with..” is already mentioned in line 104 (“…ONT ultra-long read length was 26,487.9 bp.”)

## Reviewer #1:

*Liu et al., describe a new genome assembly of a strain of Drosophila melanogaster, Canton S collected in Beijing. The sequence is derived from the male progeny of a single pair mating. The authors used a new hybrid strategy called telomere-to-telomere (T2T) which includes PacBio, Oxtford Nanopore, Illumina paired-end sequencing and HiC and two assemblers Canu and NextDenovo followed by genome polishing with minimpa2 and Nextpolish. They generated 30 contigs and used Hi-C data to place 25 of the scaffolds onto the four pairs of chromosomes; three autosomes and one pair of sex chromosomes. The second and third chromosomes are metacentric with left and right arms, the fourth is a dot acrocentric chromosome, the X is a large acrocentric and the Y is submetacentric. Liu et al., have expanded the genome by 17.93 Mb. In addition to the genome assembly, they used existing RNA-seq (cited as being from the ENA database) to generate an experimentally derived transcriptome. The genome sequence and transcriptome were then used to produce a genome annotation that was compared to the D. melanogaster Release 6 ISO1 reference strain sequence to identify new genes. According to their abstract, they identified 78 novel protein-coding genes. Two of these were studied genetically using the CRISPR/Cas9 system to generate mutants. One of these was highlighted in the abstract as they determined it to be a new olfactory gene which when mutated has a reduced sensitivity to smell. With the major and minor concerns properly addressed, the manuscript should be re-reviewed to determine whether it will be accepted for publication or not.*

**Response:** We sincerely appreciate the reviewer's insightful summary and constructive feedback on our work, as your comments proved instrumental in elevating its quality. We have carefully addressed these points and incorporated the revisions into the manuscript. A detailed summary of these changes is provided below.

### Major Issues

***Q1. A major concern is that Liu et al., did their T2T studies with a variant of Canton S, native to Beijing and not the sequenced reference ISO1 strain. It is well known from the earliest cytological (PMID: 4207116) and sequence studies (e.g. PMID: 17737996; PMID: 29255259) that strain differences result in very different genome sequence which affects gene numbers and heterochromatin size. The strains and sequence differences merit discussion in the introduction.***

**Response:** We thank the reviewer for raising the critical issue of genomic differences among *Drosophila* strains. In our study, we utilized the *Canton S* strain. The *Canton S*

strain of *D. melanogaster* also holds significant importance in biological research as a classical wild-type strain. Characterized by its normal physiological and behavioral traits, it serves as a foundational experimental reference standard, frequently employed as a control group to investigate the characteristics of mutant strains or to validate the effects of various experimental manipulations. Owing to its well-defined wild-type phenotype, this strain has been instrumental in advancing our understanding of genetic, developmental, and molecular mechanisms using *Drosophila* as a model organism.

Indeed, as the reviewer noted, significant genomic sequence differences exist between strains. For instance, the study cited by the reviewer (PMID: 17737996) compared the DNA sequences of the bithorax region between *Canton S* and *Oregon R* strains, revealing approximately 1% sequence divergence in that region, reflecting natural DNA polymorphism among strains. Consequently, our work provides an additional high-quality genome assembly representing a distinct genotype, complementing the *ISO-1* reference and enhancing our understanding of genomic and phenotypic diversity.

We fully concur with the reviewer's suggestion that strain differences should be discussed in the manuscript. Accordingly, we have incorporated a discussion section (Lines 569-573) in the revised manuscript addressing the distinctions between the strain we studied and the *ISO-1* reference strain. Additionally, we emphasize the value of investigating natural genetic diversity through multiple wild-type backgrounds and enhancing our comprehension of heterochromatic variability. We appreciate the reviewer's valuable contribution to improving our manuscript.

***Q2. Also, the strain should be given an identification number and deposited with a number of stock centers (e.g. Bloomington) so these flies are available to the community so these studies can be repeated.***

**Response:** We appreciate the reviewer's comment. The *D. melanogaster* strain utilized in this study possesses the identification number RRID: BDSC_64349 at the Bloomington Drosophila Stock Center. This essential information, inadvertently omitted previously, has been incorporated into the "Materials and Methods" section of the revised manuscript (Lines 620-621).

***Q3. For their RNA studies they reference ENA which is the European Nucleotide Database and as the source of the data, yet in the Supplemental Table S7 it appears that the sequences are from NCBI and have SRR numbers. It is important for these studies to include the strains as an additional column in Supplemental Table S7 since the RNA samples appear to come from many different studies.***

**Response:** We thank the reviewer for highlighting the necessity to clarify the data sources and associated sample metadata. While the datasets were downloaded from the

ENA, they were obtained via mirroring from the NCBI Sequence Read Archive (SRA), hence retaining their SRR accession numbers. We apologize for the potential confusion arising from this discrepancy and have revised both the legend of Supplementary Table 13 (originally Supplementary Table 7) and the relevant description in the Materials and Methods section to explicitly address this point. Furthermore, in accordance with the reviewer's recommendation, we have incorporated an additional column into Supplementary Table 13 detailing the specific *D. melanogaster* strain utilized for each dataset.

***Q4. Furthermore, all the numbers in the manuscript need to be reviewed and reconciled, especially those that describe new genes (e.g. abstract 78; Fig 4. 96; other instances 217; 114; 105 and in the discussion is reduced to "several") and functionally validated genes (e.g. Figure 4, 3; line 425, 2 and abstract only one). The sequence of these new genes should be included as a Supplemental Figure. They could also include information on paralogs and orthologs.***

**Response**: We sincerely thank the reviewer for identifying the inconsistencies in the reported counts of newly identified genes and functionally validated genes. The variations observed between the abstract (78), main text (e.g., 217, 114, 105), and discussion ("several") resulted from the application of divergent classification criteria at different manuscript sections in the original version. All instances reporting gene quantities within the manuscript have been meticulously revised to ensure strict numerical consistency and clarity.

Notably, in response to the reviewer's other comment (#17), we have refined and standardized the gene identification workflow. Consequently, there is a slight change in the total number of novel genes (n = 92). Information about the novel genes has been updated in Supplemental Table 11. The updated numbers are explicitly detailed within the revised manuscript (Lines 363-373). Regarding the number of functionally validated genes, there were only 2 novel genes were selected for functional validation.

***Q5. Finally, the BioProject, ID PRJCA033379, is not yet public and according to the website https://ngdc.cncb.ac.cn/search/specific?db=bioproject&q=PRJCA033379 will not be available until December 7th of 2026. It should be released for review and before publication.***

**Response:** Thanks very much for your comments. We have now deposited the genome assembly and raw sequencing data to the NCBI database under BioProject accession number PRJNA1237537. Additionally, the genome assembly and annotation files can be accessed via figshare at https://figshare.com/articles/dataset/The_near_complete_genome_assembly_of_Drosophila_melanogaster/28642964. The "Data Availability Statement" section of the

revised manuscript has been updated accordingly, providing the relevant accession numbers to ensure full public access to all essential datasets (Lines 932-939).

**Minor Issues**

***1. Line 124. There are not "7 major chromosomes" in Drosophila. There are 4 pairs (described above). Please correct all instances. The chromosome numbers are not 0 padded please correct this as well (e.g. Chr2, not Chr02L)***

**Response:** We appreciate your identification of this issue. We agree that *Drosophila melanogaster* possesses 4 pairs of chromosomes (X/Y, 2, 3, and 4) in its karyotype. However, in the context of genome assembly and annotation, it is widely referred to as the seven assembled chromosomes or the seven major scaffolds. This terminology is consistent with prior assemblies, such as R6 of the *D. melanogaster* reference genome. To avoid confusion, we have now corrected "7 major chromosomes" to "seven assembled chromosomes" in the revised manuscript (Lines 165-166). Furthermore, zero-padding has been removed, and chromosome labels have been revised accordingly (Lines 115-116).

***2. Lines 135-136. The figure legend for Figure 1 needs to be corrected –Fig 1C defines Fig 1D and Fig1D defines Fig1C.***

**Response:** We apologize for this error. In the revised manuscript, the legend for Figure 1C has been corrected accordingly. Due to structural modifications in the manuscript, the original Fig. 1D, depicting an example of an error type, has been renumbered as Fig. 4a (Lines 388-390).

***3. Line 135. Error types (1- 4) should be defined the first time they are used. Currently described lines 388-390.***

**Response:** Thanks very much for your suggestion. We have revised it in the revised manuscript (Lines 380-382).

***4. Line 143. Report the genes not recalled in a Supplemental Table. Describe why they weren't recalled.***

**Response:** We appreciate your suggestion. In the revised manuscript, we performed a complementary assessment using Compleasm (Li et al., 2022), a tool leveraging the latest BUSCO database, which has been validated in multiple recent studies as exhibiting greater accuracy and robustness compared to traditional BUSCO workflows (Olagunju et al., 2024; Porubsky et al., 2025), to further evaluate our genome completeness. The Compleasm results indicated a genome completeness of 99.93%, with only two genes reported as missing: *uncharacterized protein isoform A* and *Kelch repeat type 1*. This demonstrates that the genome assembly represents a highly complete

reconstruction. Nevertheless, a minor fraction of genes remains undetected, which may be due to (1) technical limitations in resolving highly repetitive regions where gene-coding sequences may be embedded within assembly uncertainties, and/or (2) annotation biases against non-canonical gene structures (e.g., long non-coding RNAs, circular RNAs) lacking conserved open reading frames or promoter signatures. It is noteworthy that such rare gene absences are common even in high-quality assemblies and do not substantially impact the reliability of downstream analyses (as below references).

**Reference:**

Olagunju TA, Rosen BD, Neibergs HL, et al. Telomere-to-telomere assemblies of cattle and sheep Y-chromosomes uncover divergent structure and gene content. *Nat Commun* **15**, 8277 (2024).

Porubsky D, Dashnow H, Sasani TA, et al. Human de novo mutation rates from a four-generation pedigree reference. *Nature* (2025).

***5. Lines 147- 152. Figure 2 needs to be corrected, A-D misordered. What is labeled B is described in D. What is labeled D is described in B.***

**Response:** We appreciate your careful review. Due to structural modifications in the manuscript, the original Fig. 2 has been renumbered as Fig. 3. We have accordingly corrected panels B and D to ensure consistency in the revised manuscript (Lines 347-352).

***6. Line 148. Change "highly" to "high".***

**Response:** Revised (Line 349).

***7. Line 167-177. What QC was performed to make sure no contaminating sequences were introduced that could account for the SINE elements?***

**Response:** We appreciate the reviewer for raising this critical point. To assess potential contamination that could explain the SINE elements and mitigate its effects, the genome assembly was aligned against the NT database using BLASTN. Sequences exceeding 1 Mb in length were partitioned into 50 kb bins for analysis. For unsegmented sequences, taxonomic assignment was determined based on the best alignment hit. For sequences analyzed in bins, the final classification was assigned to the species with the highest number of matching bins. This process enabled the taxonomic classification of each sequence to ascertain its likely origin.

Additionally, we conducted GC-depth analysis to further evaluate the presence of exogenous contamination within the assembled genome. The results revealed a GC content distribution ranging from 30% to 60% and a sequencing depth concentrated around 1000×, exhibiting a unimodal distribution without anomalous deviations (newly added Supplementary Fig. 3). Collectively, these analyses indicate that the assembly is

devoid of significant exogenous contamination and that the identified SINE elements are likely endogenous to the target genome, a very interesting finding.

***8. Line 187 Change "(Figure 1B)" to (Figure 1B; Sup Table 2). To Supplemental Table 2 add the Rel 6 numbers and the differences as two separate columns.***

**Response:** We appreciate the suggestion. To enhance clarity in the revised manuscript, the main text has been updated to cite both Fig. 1d (originally Fig. 1b) and Supplementary Table 6 (originally Supplementary Table 2). Furthermore, Supplementary Table 6 has been augmented to include the GC content of the Dm.nT2T genome assembly, alongside the chromosome lengths and GC content derived from the R6 genome. Please refer to the supplementary materials for details.

***9. Line 204-208. Change "B) Ideogram of density distribution of different type of SVs on chromosome of the genome assembly. (C) The size of different type of SVs processed by logarithm." to "(B) Ideogram of density distribution of different types of SVs on chromosomes of the genome assembly. (C) The logarithmic scale size of different types of SVs.".***

**Response:** Thanks a lot for your suggestion. Revised (Lines 244-246).

***10. Line 211. When they discuss structural variants, they should mention strain differences as the possibility for the variation.***

**Response:** We appreciate the reviewer's insightful comment. In the revised manuscript (Lines 492-509), we have expanded the Discussion to explicitly emphasize that structural variants (SVs) may reflect inherent strain differences, a perspective substantiated by multiple studies. We now highlight that SVs are not only prevalent across diverse *Drosophila* strains but also exhibit significant polymorphism within individual strains. This underscores the importance of analyzing multiple individuals and strains in future research to validate the stability and functional relevance of these variants.

***11. Lines 230-231. The reference sequence was assembled and compared with restriction digests of individual BAC clones mapped to the assembly. In addition, the reference was compared to an optical map. These two strategies confirmed the reference assembly order and orientation. What have the authors done to independently confirm their assembly? It would be wise to be careful about suggesting Rel6 has misassemblies without specific cases being confirmed in the proper reference strain.***

**Response:** We recognize the methodological rigor of the validation approaches employed during the assembly of the R6 reference genome, including BAC clone

mapping and optical map comparison, which provide robust evidence supporting its assembly accuracy. In our original statement, our intent was not to assert specific misassemblies within the R6 reference genome, but rather to propose alternative interpretations for the structural variations observed between our newly assembled genome and the reference that could not be explained by repetitive sequences. We acknowledge that these discrepancies may equally reflect inherent structural polymorphisms between strains, rather than errors within the reference assembly itself. To avoid any unsupported implications, we have removed the reference to potential misassemblies in R6 accordingly in the manuscript.

***12. Line 275. IGV should be spelled out - Integrated Genome Viewer. Please define this acronym the first time it is used.***

**Response:** Thanks very much for your suggestion. Revised (Lines 287-288).

***13. Line 327. Change "12,340.5 copies" to "an average of 12,340 copies". Not clear what .5 copies represents.***

**Response:** We acknowledge the reviewer's concern. The value of 12,340.5 copies was derived using the computational method described by Shi et al. (2023), employing the TRF (Tandem Repeats Finder) software. This number represents the average copy number across the genome assembly. We have revised as suggested to enhance clarity and accuracy (Line 335).

***14. Lines 328-329. Since the authors are discussing the centromeric end of the X chromosome, the sentence "This repeat is located at the right end of the X chromosome." is redundant and could be deleted.***

**Response:** Thanks for your suggestion. We have thoroughly rephrased the original sentence to improve clarity and accuracy. The revised sentence is in Lines 334-336.

***15. Line 346-347 Change "the 4 chromosome" to "the 4th chromosome".***

**Response:** We have revised it to "Chr4" (Lines 325-326).

***16. Line 381-383. Change "We identified four main types of mis-annotations:" to "In our automated computationally derived annotation, we identified four main types of misannotations:".***

**Response:** Thank you for the suggestion. We have thoroughly rephrased the original sentence to improve clarity and accuracy. The revised sentence is in Lines 380-382.

***17. Line 390-398. Are the 114 genes homologs (truly missing from the reference genome) or paralogs (where there are many copies of which one is missing)?***

**Response:** .

**Response:** We sincerely thank the reviewer for this comment. In the revised manuscript, we adopted a new and more stringent pipeline to identify novel genes. First, we filtered overlapping genes between the EVM-version annotation and the Liftoff-version annotation, and we obtained the primary annotation set comprising 17,898 genes, among which there were 213 potential novel genes. To determine whether the 213 genes are novel, we performed a reciprocal best-hit (RBH) BLASTP alignment between the protein sequences of these 213 genes from the Dm.nT2T genome assembly and the R6 genome. A total of 63 orthologous genes and 17 paralogous genes were identified. The remaining 133 genes were cross-validated with transcript assemblies from multiple tissues to confirm their transcriptional activity, resulting in the identification of 92 novel genes (Supplementary Table 11), of which 90 originated from unannotated regions and 2 from unassembled regions using TBLASTN. Finally, we used CPC2 to predict coding potential and identified a total of 74 genes with coding potential. The above method and results for identifying novel genes have been updated in the revised manuscript in Lines 861-877.

***18. Line 500. add a reference to the sentence "However, Drosophila telomeres are composed of three different transposons: HeT-A, TART and TAHRE, all of which belong to the retrotransposon class", such as reviewed in Pardue and DeBaryshe 1999. In general, many background citations are missing and should be added when they describe work that is not their own.***

**Response:** Thanks for your suggestion. We have added the reference accordingly (Lines 69-72). Furthermore, we conducted a thorough review of the entire manuscript and incorporated additional background citations as necessary to acknowledge prior research and to enhance the clarity and scientific rigor of the revised manuscript.

***19. Line 503. Clarify the sentence "We also found that this telomere pattern is absent in the Drosophila genus". Needs a reference. Not clear what "this telomere pattern" refers to the human type or the Drosophila type. Also there are no experiments in this study where you describe other members of the Drosophila genus which has 1600 species.***

**Response:** We apologize for the confusing and arbitrary expression. Here, "this telomere pattern" specifically denotes the *D. melanogaster* type telomeres composed of retrotransposons. As evidenced by previous studies, the telomere architecture in *D. melanogaster* exhibits distinctive features that differentiate it from most other *Drosophila* species (http://cfb.ceitec.muni.cz/telobase). However, given that *Drosophila* encompasses over 1,600 extant species, 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 have modified this sentence in the revised manuscript (Lines 551-566).

***20. Line 507. Change "that The P-element" to "that the P-element".***

**Response:** We have rephrased the original sentence; the revised sentence is in Lines 554-556.

***21. Lines 497 – 512. The discussion about telomeres as it stands is not merited and should be deleted. How much have they improved the sequence? This should be described.***

**Response:** We thank the reviewer for the valuable comments. Regarding the suggestion to remove the section on telomeres, we respectfully propose its retention. As presented in Table 1, the quantitative improvements of our assembly encompass enhanced contiguity and gap closure, thereby facilitating an opportunity to investigate telomeric regions, the discussed in assembly improvement in Lines 458-476. We have revised the text to clarify that our observations on telomeres extend beyond mere assembly extension; they also address documented biological differences in telomeric structure across *Drosophila* strains (Lines 567-583). Retaining this section offers additional biological context within the broader framework of *Drosophila* telomere biology.

***22. Line 531. Materials and Methods. Make sure to include all software versions when software is mentioned, also cite the primary references or the URL. Examples of software with missing citations are listed below but is not comprehensive.***

**Response:** Thank you for your valuable comments. We have addressed the previously omitted software version numbers and references.

***23. Lines 544. Change "After the sample was qualified, size-select of long DNA fragments were performed using the BluePippin system" to "After the sample was quantified, sizeselection of long DNA fragments was performed using the BluePippin system".***

**Response:** Thanks a lot for your suggestion. We have revised accordingly (Lines 638-639).

***24. Line 562. Change "fragments were screened by the BluePippin" to "fragments were screened using the BluePippin (Sage Science, USA) system".***

**Response:** Thanks a lot for your suggestion. We have revised accordingly (Line 656).

***25. Line 570. What version of Guppy was used? What Adapter trimming parameters were used? Were reads filtered for size?***

**Response:** Thank you to the reviewer for the important comment. We used Guppy (v3.2.2) for basecalling with the configuration file dna_r9.4.1_450bps_fast.cfg. During the basecalling process, Guppy automatically detects and trims adapter sequences at the start of reads (SQK-LSK109 adapters) using its built-in algorithm (--trim_adapters yes). Subsequently, quality filtering was conducted by retaining only sequences exhibiting a `mean_qscore_template` ≥ 7. The reads shorter than 1 kb were excluded using NextDenovo. These high-quality pass reads were directly utilized for downstream genome assembly. The Guppy version and parameters have been incorporated into the revised manuscript (Lines 664-666).

***26. Line 585. Incomplete sentence "And Illumina short reads were (what belongs here?) using Nextpolish".***

**Response:** We apologize for the oversight. In our study, we used BGI reads, which refer to short-read sequencing data generated by the DNBSEQ platform developed by BGI (Shenzhen, China). We have provided a more detailed description of the method in the revised manuscript (Lines 679-682).

***27. Line 590. Change "controlling" to "control"***

**Response:** Revised accordingly (Line 684).

***28. Line 591. Clarify what you mean by "as former research" and add a reference.***

**Response:** Thank you for pointing out the missing reference. We have now clarified this sentence in the revised manuscript (Line 686) and added the corresponding reference.

***29. Line 599. Delete the second instance of scaffolds. From - "The scaffolds were further clustered, ordered, and oriented scaffolds" to "The scaffolds were further clustered, ordered, and oriented".***

**Response:** Thanks a lot. We have revised it accordingly (Lines 695-696).

***30. Line 611. Replace (ref) with the actual citation.***

**Response:** Revised accordingly (Line 705).

***31. Line 616 -618. Please clarify the following sentence ". Besids, base accuracy of the assembly was calculated with bcftools."***

**Response:** Thanks. We have corrected the spelling errors and expanded this section by providing additional computational details to clarify its meaning. A more detailed exposition of this analytical procedure and an updated reference list have been incorporated into the revised manuscript (Lines 712-714).

***32. Line 618. Add a reference for the LTR Assembly Index calculation.***

**Response:** Thanks very much for your comment. It is noteworthy that we have removed this section of analysis from the revised manuscript. Our reasons are as follows:

We initially performed LAI calculations following the method described by Ou et al. (2018) to assess the assembly quality of our genome:

$$\text{Raw LAI}=\frac{\text{Intact LTR retrotransposon length}}{\text{Total LTR sequence length}}\times 100 \quad (1)$$

$$\text{LAI}=\text{raw LAI}+2.8138\times(94-\text{whole genome LTR identity}) \quad (2)$$

This method evaluates the completeness and continuity of LTR-retrotransposons (LTR-RTs), which makes it particularly suitable for complex plant genomes that are rich in LTR-RTs--for example, maize and wheat (Sabot et al., 2005; Schnable et al., 2009), in which LTRs can constitute over 70% of the genome. However, this method proves unsuitable for fly genomes: our assembled genome yielded an LAI of 6.34, while the R6 exhibited an LAI of 11.87—both values falling considerably below the established "gold standard" threshold of LAI > 20 (Ou et al., 2018). Consequently, LAI may not be a sensitive indicator of genome quality in this context.

Given this, retaining this metric may potentially mislead readers or raise unnecessary concerns, and so we have removed this result from the revised manuscript. Instead, we rely on other widely accepted metrics, such as BUSCO completeness, Quality Value (QV), and N50 length, which collectively provide a robust, multidimensional evaluation of assembly quality.

***33. Line 624. Add a reference for the GMATA and Tandem Repeats Finder software.***

**Response:** Revised accordingly (Line 721).

***34. Line 626-7. Correct spelling and italicize ab initio.***

**Response:** Revised accordingly (Line 723).

***35. Line 627. Correct A to a in the sentence starting "Briefly".***

**Response:** Revised accordingly (Line 723).

***36. Line 628. Add references for MITE-hunter and RepeatModeler.***

**Response:** Revised accordingly (Line 724).

***37. Line 642. Add a reference for the GeMoMa software.***

**Response:** Revised accordingly (Line 740).

***38. Line 643. Add a reference for the Liftoff software.***

**Response:** Revised accordingly (Line 742).

***39. Line 646. Add a reference for the STAR software.***

**Response:** Revised accordingly (Line 745).

***40. Line 650. Add a reference for the AUGUSTUS software.***

**Response:** Revised accordingly (Line 749).

***41. Line 652. Add a reference for the EVM software.***

**Response:** Revised accordingly (Line 751).

***42. Line 653. Add the version and the reference for InterProScan***

**Response:** Revised accordingly (Line 761).

***43. Line 667. Add the version and the reference for tRNAscan-SE***

**Response:** Revised accordingly (Line 764).

***44. Line 668-9. Add the versions and references for Rfam database and RNAmmer***

**Response:** Revised accordingly (Lines 766-767).

***45. Line 679. Replace "using TRF2GFF (reference)" with the actual citation.***

**Response:** We deeply apologize for this oversight. We have consequently added a comprehensive citation to its GitHub repository within the revised manuscript (Line 777).

***46. Line 687. Replace "MACS2 (reference, version)" with the citation and version.***

**Response:** We sincerely apologize once more for this oversight. We have now incorporated the software versions and reference details (Line 785).

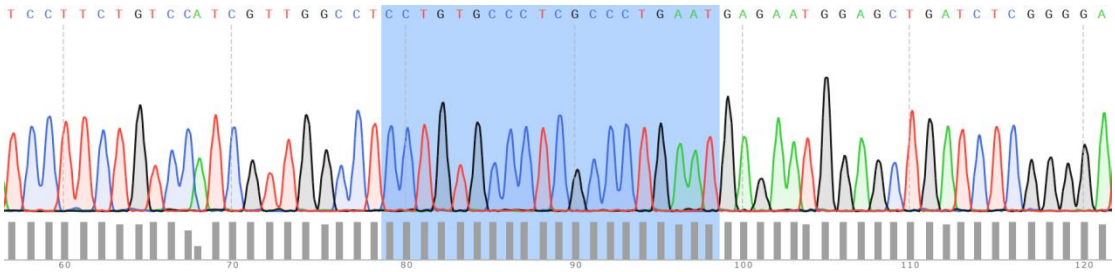
***47. Line 713. Include the accession numbers for the ATAC-seq datasets or the Sup Table.***

**Response:** We acknowledge this suggestion. The accession numbers for all ATAC-seq datasets, along with the corresponding sample tissues, source publications, or project IDs as applicable, have been incorporated into Supplementary Table 8.

***48. Line 749. How was the stable mutant Chr02L.722 homozygous line validated?***

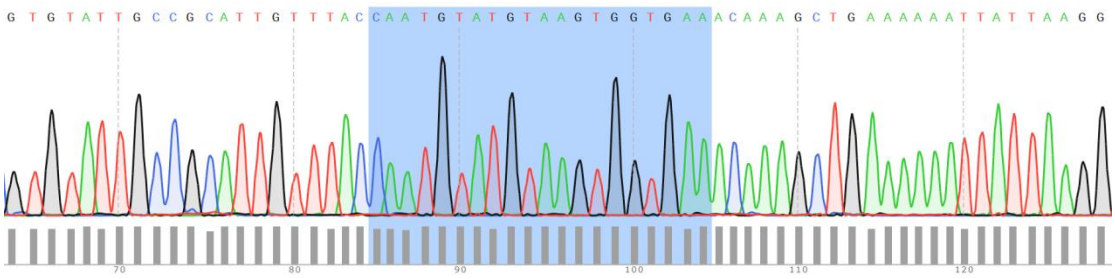
**Response:** This is one crucial question. In our study, the stable homozygous mutant line *Chr2L.722* was established through multiple generations of genetic screening and

molecular validation. The detailed process is as follows: a) When the injected P0 embryos grew into adults, they were crossed with TM3/TM6B. The genomic DNA of the P0 and F1 flies was extracted. PCR was performed using primers for validation of Indel; b) Verified F1 heterozygotes were then crossed again with TM6B to maintain and propagate the mutant allele in the F2 generation; c) F2 individuals were intercrossed to generate the F3 generation, from which homozygous mutants were selected; d) To confirm homozygosity, we performed PCR and Sanger sequencing on the target locus in F3 individuals. The sequencing chromatogram at the mutation site was carefully examined: a single peak indicates a homozygous mutant, while overlapping double peaks would suggest a heterozygote. As shown in the chromatogram below, the region surrounding the mutation site (± 10 bp, highlighted in blue) displays clean single peaks, confirming that the F3 individual is homozygous. This multi-generational breeding and molecular validation process ensures that *Chr2L.722* is a genetically stable homozygous mutant line.



***49. Line 755-6. How was the stable mutant Chr03R.2449 homozygous line validated?***

**Response:** The validation process for the stable homozygous mutant line *Chr3R.2449* adhered to that previously detailed for *Chr2L.722*. Similarly, PCR amplification coupled with Sanger sequencing was performed on F3 individuals. Subsequent chromatogram analysis focused specifically on the target mutation locus. The observation of a single sequencing peak at the mutation site, demarcated by the blue-highlighted region (±10 bp) in the figure below, definitively confirmed the homozygous status of the mutant line. We have added this information and the answer to comment (#48) into the revised manuscript (Lines 895-899).



***50. Line 760. We selected 5-day-old Chr03R.2449 mutant Drosophila (experimental group) and non-knockout Drosophila (control group). Were these males, females, larvae, adults? What does the 5-day-old refer to?***

**Response:** This is another crucial question. The term "5-day-old *Drosophila*" specifically refers to adult flies collected and utilized five days after eclosion (Shuai et al., 2010 *Cell*; Li et al., 2023 *Science*). As these two key studies demonstrate, flies within this precise age range are well-suited for stable transcriptomic and behavioral investigations. Importantly, sex-specific distinctions were not employed in this study, as our core objective was to delve into general transcriptional and phenotypic effects unbiased by sex. We have added the above reference information in the revised manuscript (Lines 902-904).

***51. Line 762. "Based on previously published studies" needs a reference.***

**Response:** Thanks a lot for your suggestion. We have added relevant references (Lines 904-905).

***52. Line 764-5. What are the units of the OCT solution?***

**Response:** In our experiments, the OCT solution was prepared through dilution, specifically by adding 3-octanol to mineral oil at a precise volume ratio of 1.5:1000. This means that for every 1000 units of mineral oil, 1.5 units of OCT were incorporated, yielding a final solution with a 3-octanol volume fraction of precisely 0.0015 (unitless). This clarification has been integrated into the revised manuscript, as detailed in Lines 908-909.

***53. Line 781. "To compare the differences in the Index values between the CK and gene2 groups." Define these groups***

**Response:** We have revised the text accordingly (Lines 926-927). The "gene2" group specifically denotes adult flies carrying *Chr3R.2449* mutations, whereas the CK group represents adult flies without mutations in this gene.

## Reviewer #2:

### Major Comments

***Q1. Genome assembly is not publicly available and has not been deposited in an appropriate database.***

***a) This study emphasizes the superior quality of the newly assembled genome compared to the previous reference genome. The raw sequencing data have been deposited in the National Genomics Data Center (NGDC) of China, allowing researchers to access the data to validate or reproduce this study. However, the most crucial dataset—the genome FASTA file and annotation files—is missing from the database. The authors should publicly release these files to ensure the reproducibility and accessibility of their findings.***

**Response**: We thank the reviewer for highlighting the issue regarding the availability of the genome assembly and annotation data. We sincerely apologize for the oversight in not submitting these files to the designated repository promptly. We fully concur that access to these files is essential for ensuring transparency and reproducibility. Now, all the data generated in this study have been deposited in the National Center for Biotechnology Information (NCBI) under BioProject accession number PRJNA1237537. Additionally, the genome assembly and annotation files can be accessed via figshare at https://figshare.com/articles/dataset/The_near_complete_genome_assembly_of_Drosophila_melanogaster/28642964. This information has been updated accordingly in the "Data Availability Statement" section of the revised manuscript (Lines 932-939).

***b) Additionally, since Nature Communications is part of the Nature portfolio, it is highly recommended to follow their data-sharing policies. Relevant guidelines can be found here:***

*- "Reporting standards and availability of data, materials, code and protocols"*
*https://www.nature.com/nature-portfolio/editorial-policies/reporting-standards#availability-of-data*
*This page states:*
*"... the preferred approach is to make data available in repositories. Scientific Data, a Nature Portfolio journal, maintains a list of approved and recommended data repositories to support researchers seeking suitable repositories for their data."*
*- "Data Repository Guidance"*
*https://www.nature.com/sdata/policies/repositories*
*This document specifies that "Raw sequencing data (reads or traces), Genome*

*assemblies, Annotated sequences, and Sample metadata" must be publicly available in recognized repositories such as INSDC repositories or the Genome Sequence Archive (GSA). Currently, some of these datasets remain inaccessible, making it impossible to evaluate their quality.*

**Response:** We sincerely appreciate the reviewer's suggestion. We acknowledge that certain data were not publicly accessible in INSDC member databases or other recognized repositories at the initial submission; we regret this oversight. We have now successfully submitted the genome assembly and annotation (as mentioned above).

***c) Furthermore, since the raw sequencing data have been uploaded to NGDC, it is necessary to clarify whether NGDC is part of the International Nucleotide Sequence Database Collaboration (INSDC). If not, the dataset must be made available in NCBI GenBank (for genome assembly) and NCBI SRA (for raw sequencing reads) or another recognized repository.*** *https://www.insdc.org/*

**Response:** As mentioned above, we have submitted the raw sequencing data to the NCBI database under the BioProject accession number PRJNA1237537. By the way, NGDC is a distinguished national-level life science data center established by the Beijing Institute of Genomics, Chinese Academy of Sciences. Its Genome Sequence Archive (GSA) has also been highly recommended as a data storage and sharing platform by *Springer Nature* since 2021 and has also supported prominent genome studies, such as those on cotton and tobacco. However, NGDC is not currently an official member of the INSDC.

***Q2. Issues in updating the genome***

***It is unclear how the final genome assembly will be utilized.***

***To illustrate the importance of maintaining previous reference genomes, an analogy can be drawn from human genome research. The human reference genome, GRCh38, originally produced using Sanger sequencing, remains in use despite the advent of long-read sequencing technologies. A vast body of research and extensive annotation data have accumulated based on GRCh38. Consequently, even though large-scale projects now employ long-read sequencing to refine the human genome or construct a human pangenome, GRCh38 cannot simply be discarded. Instead, it continues to serve as the backbone while additional long-read genome assemblies are incorporated into a graph-based reference genome.***

***A similar approach seems necessary for Drosophila melanogaster. The existing annotations cannot be discarded and must be retained. Given this, the present study should not merely provide a new genome assembly and annotations for the same Canton-S strain, but rather integrate:***

***- the existing Canton-S strain reference genome,***

***- the newly generated Canton-S strain genome assembly, and***
***- previously available genome assemblies from other strains***
***into a graph-based genome reference. This would ensure that the updated reference genome maintains compatibility with prior annotations.***
***If full integration is too complex (although I do not believe it is), the study must at least transfer all previous gene annotations and other functional annotations to the new genome using a liftover approach. Without either of these efforts, it is questionable whether the dataset produced in this study will be widely adopted by the research community.***

**Response:** Thank you for your thoughtful and constructive feedback, which has greatly guided our reflection on the study's alignment with community needs and the practical utility of our genome assembly. We sincerely appreciate the opportunity to clarify our approach based on your insights, and we address your concerns as follows:

**1. Value of the existing reference genome**

We fully concur with your emphasis on preserving existing reference genomes as valuable resources. For *Drosophila melanogaster*, genetic diversity across strains is well-documented, and we explicitly acknowledge that our high-quality genome assembly of *Canton S* does not aim to replace the current reference genome. Instead, it is intended to serve as a complementary, high-quality resource for future research-particularly for studies focused on this specific *Canton S* strain or requiring comparisons across multiple genomic sequences. By providing an additional assembly, we aim to enrich the *Drosophila* genomic toolkit while ensuring compatibility with, rather than undermining, existing references.

**2. Integration into a graph-based reference genome**

We recognize the importance of graph-based reference frameworks, as exemplified by your analogy to human pangenome efforts. Currently, we are conducting population genomic and pan-genome analyses across multiple *Drosophila* strains, which will lay critical groundwork for such integration. However, due to the complexity and scope of this endeavor, we plan to focus on these advanced integrative analyses in a separate, dedicated paper. This allows us to thoroughly address the technical and biological nuances of graph-based reference construction, ensuring robust methodology and comprehensive discussion elements we believe are essential for meaningful community adoption.

**3. Integration of gene annotations from other strains**

We share your concern about maintaining annotation continuity. For our assembly, we first performed de novo annotation by integrating *ab initio* gene prediction, homology-based evidence, and RNA-seq data. Subsequently, we used Liftoff (v1.6.3, Shumate & Salzberg, 2021) to transfer the gene annotations from the current *D. melanogaster* reference genome (r6.54) onto our assembled genome. The final

annotations, obtained by integrating the transferred and newly generated annotations, were used for subsequent downstream analyses. This strategy ensures backward compatibility with existing annotations while fully leveraging the advantages of our improved genome assembly.

We share your concern about maintaining annotation continuity. To explore this, we downloaded genome assemblies from seven distinct *Drosophila* strains and applied a *de novo* annotation pipeline to each (as these assemblies did not come with annotation files). Unfortunately, the annotation results were of poor quality, with high rates of fragmented or misassigned gene models (see the table below), which may reflect differences in assembly quality, completeness of annotation, or genuine biological variation between strains. Given these limitations, we concluded that integrating these low-confidence annotations would risk introducing errors into downstream analyses. Instead, we will focus our efforts on refining annotations for our new *Canton S* assembly to ensure high accuracy. When the improved annotations are available for these other strains in the future, we will gladly revisit the integration strategy.

By clarifying these points, we aim to enhance the transparency of our study's goals and limitations while aligning our efforts with the research community's needs for reliable, usable genomic resources. We will revise the manuscript to explicitly state these intentions, including discussions of the existing reference genome's value, our plans for future graph-based integration, and the rationale behind our annotation decisions.

Thank you again for your invaluable comment, which has been instrumental in strengthening the study's rigor and relevance.

| GenBank | strain | Level | Release Date | No. protein genes |
|---|---|---|---|---|
| r6.54 | ISO-1 | Chromosome | 2014 | 13,962 |
| GCA_003401735.1 | Canton-S | Chromosome | 2018 | 8,675 |
| GCA_004798075.2 | DGRP-732 | Chromosome | 2019 | 8,711 |
| GCA_002300595.1 | A4 | Chromosome | 2017 | 9,025 |
| GCA_042606445.1 | ISO-1 | Chromosome | 2024 | 9,236 |
| GCA_020141675.1 | RAL375 | Chromosome | 2021 | 8,603 |
| GCA_020141495.1 | SE_Sto_11_22 | Chromosome | 2021 | 8,733 |
| GCA_024500395.1 | ORw1118 | Chromosome | 2022 | 10,065 |

***Q3. Data and assembly quality should be analyzed much more thoroughly. Currently, it is difficult to assess the quality of the dataset, raising concerns about its suitability for publication in Scientific Data. The following additional analyses should be included:***

***(1) Read-level quality assessment***

***a) Read length and base quality distributions should be provided. For example, a 2D***

***contour plot with read length on the x-axis and mean read quality on the y-axis should be included.***

***- For HiFi and ONT, these data should be included.***

**Response:** We sincerely thank the reviewer for highlighting the need for enhanced data quality documentation. As suggested, we have now included two-dimensional contour plots with marginal histograms in Supplemental Fig. 1, comparing read length (x-axis) versus average read quality (y-axis) for: a) PacBio HiFi data; and b) ONT ultra-long reads. Evaluation of the sequencing read quality shows that the HiFi data exhibit high base accuracy and favorable read length distribution. A total of 2,430,495 HiFi reads were generated, comprising approximately 17.36 Gb of sequence data. The mean read length was 7,141.6 bp, and the read length N50 reached 13,979 bp, indicating a high proportion of long reads suitable for high-quality genome assembly. The average Phred quality score was Q28.9 (corresponding to a base error rate of approximately 0.13%), and the median quality score was Q37.1 (approximately 0.02% error rate), suggesting that the majority of reads are highly accurate. These characteristics provide a solid foundation for *de novo* genome assembly, structural variation detection, and the resolution of complex genomic regions.

In addition, a total of 6,092,727 ultra-long reads were obtained using Oxford Nanopore Technologies (ONT), yielding 156.25 Gb of sequencing data. The average read length was 26,487.9 bp, and the N50 was 48,181 bp, indicating a substantial proportion of ultra-long reads. The average read quality score was Q14.4 (corresponding to an estimated error rate of ~3.6%), and the median quality score was Q15.1 (error rate of ~3.1%), which falls within the acceptable range for current ONT technology. Overall, the ONT dataset is of good quality and offers strong support for *de novo* genome assembly and comprehensive genome reconstruction.

***b) Since the sequencing platform is not clearly stated, it is difficult to determine whether quality adjustments are necessary. However, if a PacBio platform prior to Revio was used, phred scores should be corrected—otherwise, an excessive number of values exceeding 40 will appear. The corrected values should be plotted so that all scores remain ≤40.***

**Response:** We appreciate the reviewer's meticulous attention to data quality control. The sequencing was performed on PacBio Sequel II instruments using the HiFi sequencing mode. The data generated comprised high-fidelity (HiFi) reads produced by converting raw subreads into circular consensus sequence (CCS) reads via the official ccs tool from PacBio.

Given that Phred scores in HiFi reads are recalculated by CCS consensus algorithms to reflect empirical base-calling accuracy, the known inflation issue in raw subreads does not propagate to final HiFi data. Thus, for Sequel II-derived HiFi reads

generated using official tools CCS, no additional Phred score correction is empirically warranted. We have incorporated the data processing information into the "Materials and Methods" section of the revised manuscript (Lines: 658-661).

***(2) Assembly validation and error checking***

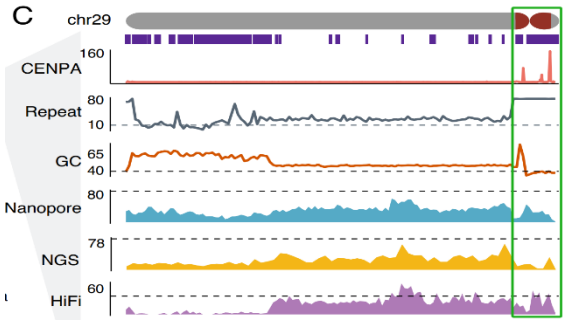
***- Scaffolding errors introduced during de novo assembly must be addressed before gap filling. It is well known that Hi-C alone frequently introduces scaffolding errors.***

***- Example: In Figure 1A, read depth distributions for HiFi and ONT do not appear evenly distributed. If this is accurate, it suggests poor genome quality in many regions. I suspect that telomeric and/or centromeric regions were not properly assembled. The authors should thoroughly analyze high- and low-depth regions.***

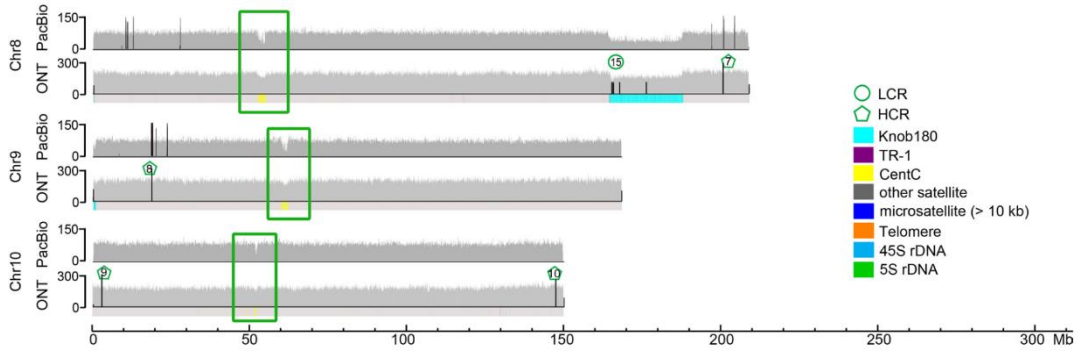
**Response:** We thank the reviewer for raising this crucial point. In this study, we implemented multiple rigorous post-assembly validation steps. Specifically, we mapped both short- and long-read sequencing data back to the assembled genome, achieving high alignment rates of 95.94% and 98.66%, respectively, indicating a remarkable level of assembly completeness and consistency. Furthermore, BUSCO assessment demonstrated 98.8% completeness, and the QV score of 40.24 robustly supports the accuracy and structural stability of the assembly across the vast majority of regions.

Regarding the read depth distribution presented in Figure 1A, we acknowledge that certain regions exhibit notably uneven depth, particularly evident in the ONT and HiFi data. Upon detailed inspection, these regions largely correspond to centromeric regions characterized by high repetitive sequence content (see updated Fig. 1a). Similar read depth patterns have also been documented in high-quality T2T assemblies of other species, such as chicken (Huang et al. 2023 *PNAS*) and maize (Chen et al., 2023 *Nature*), as illustrated in the following figures.

It should be noted that our sequencing depth has reached an exceptionally high level (~1000X), which indicates that certain regions in the genome indeed pose resolution difficulties. We fully agree that further refinements incorporating additional long-range data or leveraging advanced graph-based assembly tools are necessary to specifically resolve these low-depth regions in future work. We deeply appreciate the reviewer's valuable insights.




*Evolutionary analysis of a complete chicken genome*
https://www.pnas.org/doi/abs/10.1073/pnas.2216641120




*A complete telomere-to-telomere assembly of the maize genome*
https://www.nature.com/articles/s41588-023-01419-6

***(3) Structural variant validation***

***- A comparative structural variant analysis should be conducted between the previous reference genome and the new genome using SyRi or similar tools.***

***- Example: Translocations and inversions detected by SyRi should be examined. If scaffolding errors are present, structural variants are likely to be detected near contig junctions (i.e., gap regions). A quantitative analysis should be provided to assess the number of clear scaffolding errors in both the existing reference genome and the newly generated genome.***

**Response:** We sincerely thank the reviewer for their constructive suggestion concerning structural variant validation. Following this guidance, we conducted a comparative structural variant analysis between Dm.nT2T and the reference genome (R6) using SyRI, focusing specifically on identifying large-scale structural variants like inversions and translocations. To assess whether these variants are indicative of potential scaffolding errors, we investigated their spatial relationship to assembly gaps in Dm.nT2T. Using bedtools intersect, our analysis revealed that none of the SyRI-identified inversions or translocations overlapped with gap regions, indicating these structural variants are unlikely to stem from scaffolding artifacts. Furthermore, using bedtools closest, we investigated the proximity of structural variants to gaps. The results

demonstrated that no inversions or translocations were located within 10 kb of any gap region, providing further robust support for the structural integrity of our scaffolding and the validity of the detected SVs. These findings collectively strongly indicate that the structural rearrangements identified by SyRI represent true biological differences between strains, rather than assembly or scaffolding errors.

***(4) Base-level assembly accuracy***

***- Quality value (QV) and k-mer completeness should be assessed.***

***- Merqury can be used to compare the k-mer distribution of short-read WGS data against the genome assembly, allowing QV and k-mer completeness to be calculated.***

***- Lines 140–141: "Quality evaluation using Merqury indicated robust completeness and high quality (QV) for the assembly (Supplementary Figure 2)." → The exact values of k-mer-based completeness and QV must be explicitly provided in the manuscript. Supplementary Figure 2 only contains a visual plot without numerical values.***

**Response:** We deeply regret the oversight of omitting the k-mer-based completeness and QV values and sincerely thank the reviewer for highlighting this gap. In the revised manuscript, we have promptly incorporated the exact k-mer-based completeness and QV values derived from the Merqury analysis directly into the main text (Lines 119-120), as suggested. Furthermore, we have added these crucial metrics to the caption of Supplementary Fig. 2 to fully represent the figure information.

***4. Additional analyses are required.***

***With long-read sequencing, the authors can analyze methylation profiles of this fly sample. Additionally, segmental duplications (SDs) and other complex genomic rearrangements should be thoroughly analyzed. Key questions:***

***- How many SDs exist in this genome and across Drosophila populations?***

***- What are the SD lengths?***

***- Are there differences in intra- vs. inter-chromosomal SDs?***

***- Are SDs prone to mutations?***

*Human genome studies can provide insights into these aspects:*

*Segmental duplications and their variation in a complete human genome*

*https://www.science.org/doi/10.1126/science.abj6965*

*Increased mutation and gene conversion within human segmental duplications*

*https://www.nature.com/articles/s41586-023-05895-y*

*Structural polymorphism and diversity of human segmental duplications*

*https://www.nature.com/articles/s41588-024-02051-8*

**Response:** We sincerely thank the reviewer for these constructive suggestions. We

address each point sequentially:

**a) Methylation analysis considerations:**

We acknowledge the potential of long-read sequencing technologies (Oxford Nanopore and PacBio) for detecting DNA base modifications. However, previous studies have shown that *Drosophila melanogaster* lacks canonical DNA methylation, with extremely low levels of 5mC, and there is currently no strong evidence supporting its functional relevance in this species (Lyko et al., 2011). Consequently, methylation analysis in this context would yield limited biological insights. Furthermore, the absence of raw signal files (FAST5 format) from Oxford Nanopore sequencing precludes accurate methylation calling using current computational tools.

**Reference:** Lyko F, Ramsahoye B & Jaenisch R. DNA methylation in *Drosophila melanogaster*. *Nature* **408**, 538-540 (2000).

**b) SD characterization analysis**

Following the reviewer's suggestion, we performed additional SD characterization. Using Biser, we identified approximately 15.7 Mb of SDs in the Dm.nT2T genome assembly (161.3 Mb), representing ~9.8% of the genome. Classification revealed 14.3 Mb intra-chromosomal SDs, 3.5 Mb inter-chromosomal SDs, and 2 Mb SDs involved in both event types. A new figure (Supplementary Figure 6) visualizing SD genomic distribution demonstrates predominant enrichment in pericentromeric and centromeric regions.

**c) Population-level SD variation analysis**

To address SD variation across populations, we analyzed resequencing data from six populations (n=120 individuals; Supplementary Table 9). After GATK-based high-quality SNP filtering, we identified 1,843,389 SNPs. Among these, 17,568 SNPs resided within SD regions (total 15,786,211 bp; density: 1.11 SNP/kb), while 1,825,821 SNPs occurred in non-SD regions (140,068,244 bp; density: 13.03 SNP/kb). This indicates a significantly reduced mutation rate within SD regions versus non-SD regions, suggesting greater structural stability and/or stronger purifying selection. Allele frequency analysis of SD-located SNPs per population (Supplementary Fig. 7) revealed elevated frequencies in Ethiopian and other African populations (excluding North Africa). This pattern potentially reflects stronger purifying selection, enhanced structural conservation, and retention of ancestral alleles within these populations, possibly attributable to reduced genetic drift or bottleneck effects compared to derived populations. These SD analyses have been incorporated into the revised manuscript's Results (Lines 250-274) and Materials and Methods (Lines 822-849) sections. We thank the reviewer for comments that enhanced this work.

***5. Many critical typos and conceptual errors***

***This manuscript contains numerous typos and conceptual errors. For example, the***

***title includes "heterochromatin architecture", yet the authors do not provide any supporting evidence, such as chromatin accessibility data or histone profile results. Furthermore, it is unclear which genomic regions the authors emphasize under this term. Is the focus on centromeric regions, or does it refer to other previously uncharacterized regions within the genome? If the authors aim to highlight telomeric, subtelomeric, or centromeric regions, terms such as highly repetitive regions, complex regions, or similar descriptors would be more appropriate than heterochromatic regions. Therefore, the title should be revised to accurately reflect the content of the manuscript.***

**Response:** We sincerely appreciate the reviewer for highlighting the inappropriate terminology in the title. We fully agree with the reviewer that the term "heterochromatin" remains ambiguous without chromatin-level evidence. Consequently, we have updated the title to: "Near Complete Genome Assembly of *Drosophila melanogaster* Unveils Novel Genes and Complex Genomic Regions", and meticulously clarified the terminology throughout the manuscript, replacing "heterochromatin" with "complex genomic regions" where appropriate.

***Additionally, the manuscript contains an unacceptable number of typos and misleading concepts. Some errors are so severe that they cast doubt on whether the manuscript has undergone proper proofreading. These must be corrected before submission.***

**Response:** We are sincerely grateful for the reviewer's careful reading and for raising these crucial questions and valuable suggestions. We apologize for the oversight and acknowledge that the manuscript contained a number of typos and unclear or misleading statements. We have now meticulously proofread and revised the entire manuscript to correct these issues, enhance clarity, and ensure conceptual accuracy. We have carefully addressed all these issues within the revised manuscript.

***(1) Terminology issues***

***- The term "telomere" refers to the telomeric repeat sequence + telomeric binding proteins. In most cases within this manuscript, "telomeric region" would be a more appropriate term than "telomere".***

***- Similarly, "centromere" should be replaced with "centromeric regions".***

**Response:** We thank the reviewer for raising this point. We have carefully reviewed the entire manuscript and revised relevant descriptions of telomeres and centromeres in all instances.

***(2) Typo and error list***

***Line 31: Despite → Despite of***

***Line 37: telomeres and centromeres regions → telomeric and centromeric regions***
***Line 62: release 2, 3, 4, 5were → releases 2, 3, 4, and 5 were***
***Line 67: species evoluation, developmental biology → species evolution and developmental biology***
***Line 83: quality of genome sequencing → quality of the genome (This sentence refers to assembled genomes, not sequencing methods.)***
**Response:** We appreciate the reviewer for carefully reviewing our work and giving helpful suggestions on language and grammar mistakes. We have thoroughly rephrased these original sentences to improve clarity and accuracy in the "Abstract" section.

***Line 85: Caenorhabditis elegans should be included as an important reference.***
***Reference: Recompleting the Caenorhabditis elegans genome***
*https://pubmed.ncbi.nlm.nih.gov/31123080/*
**Response:** We appreciate your valuable recommendation. We have incorporated *Caenorhabditis elegans* as a reference genome within the model organism comparisons section in the revised manuscript (Line 85).

***Line 88: repeat sequence families → repetitive sequence families***
**Response:** Thanks for your suggestion. We have rephrased the "Introduction" section.

***Line 92: Amphibalanus Amphitrite → Amphibalanus amphitrite (Incorrect capitalization; a major taxonomic error.)***
**Response:** We have removed this example according to the next comment.

***Line 93: barnacles. → Missing reference***
***- Additionally, in this context, discussing barnacle genomes seems unnecessary. Instead of this example, the authors should discuss newly identified genes from the human T2T genome.***
***- Since Amphibalanus amphitrite was not assembled using a T2T approach on a single individual, it is not an appropriate example for this discussion.***
***- Similarly, as done in human genome research, the study should analyze genes identified through segmental duplications (SDs) or other newly assembled genomic regions.***
**Response:** We appreciate the reviewer's suggestion. Following the reviewer's recommendation, this example has been removed from the manuscript. Instead, we now reference findings from the human T2T genome project (Lines 80-83). The reference list has been updated accordingly in the revised manuscript. Furthermore, in response to the reviewer's Comment #4, analyses of segmental duplications (SDs) at both the *D. melanogaster* T2T genome level and the population level have been added to the

revised manuscript (Lines 251-274).

***(3) Methodology issues***

***Line 119: Canu → HiCanu? (Specify if HiCanu was used, as this affects the assembly process.)***

**Response:** We sincerely apologize for the lack of clarity and any resultant confusion. In our preliminary analyses, we evaluated both Canu and NextDenovo for initial genome assembly. However, the assembly generated by NextDenovo demonstrated markedly superior results; consequently, it was selected for all subsequent downstream analyses. The Canu-derived assembly was not incorporated into the final version. We have accordingly revised the Methods and Materials section to explicitly clarify this point (Line 113,673).

***Line 122: Manually curation → Manual curation***

***- What were the criteria for filtering contigs?***

***- Which aspects were considered for manual curation? The manuscript must provide detailed information on the filtering process.***

**Response**: We appreciate the reviewer's comment on this detail. We apologize for the inaccurate description in the original manuscript. The 30 contigs were derived directly from the preliminary assembly generated by NextDenovo, without any manual filtering or selection. Manual curation was performed after Hi-C-based scaffolding to order and orient the contigs. The manual adjustments were guided by the principles that inter-chromosomal Hi-C signals are generally weaker than intra-chromosomal signals, and that within a chromosome, loci that are physically closer tend to exhibit stronger Hi-C interactions. We have revised the relevant section to accurately reflect the workflow (Lines 114 and Lines 672-675).

***Line 122: Hi-C contact map should be included in the main figure, rather than in the supplementary materials.***

**Response:** Thanks for your suggestion. We have now included the Hi-C contact map in Fig. 1b of the revised manuscript (Lines 162-170).

***(4) Issues in figures and data representation***

***Figure 1A: Each circle should be numbered to improve readability.***

***Line 131: 50kb → 50-kb***

***Line 133: PacBio HiFi data sequencing depth and ONT data sequencing depth;***

***→ PacBio HiFi and ONT UL read depths. (The original phrasing is redundant.)***

**Response:** Thank you so much for your helpful suggestion. We've updated Fig. 1a and its legend accordingly as suggested, clearly numbering each track to significantly

enhance readability (Lines 162-167).

***Figure 1B:***

***- The y-axis should start at zero.***

***- Alternatively, normalizing by initial genome size would improve the visualization.***

***Line 135: melanogaster; → melanogaster. (Unnecessary semicolon)***

**Response:** Thanks very much. Following your suggestion, we have revised to start the y-axis at zero in Fig. 1d, ensuring an accurate visual representation as requested. Additionally, we duly corrected the semicolon to a period in the revised manuscript (Lines 171-172).

***Figure 1C:***

***- The figure does not represent "type I errors" as stated in the legend.***

***- Instead, it shows the composition of repetitive sequences.***

***- The legend of Figure 1C should be replaced with the legend for Figure 1D.***

***- The figure should compare the current genome with previous versions of the fruit fly genome.***

**Response:** Thanks for your valuable suggestion. In the revised manuscript, we've implemented the necessary corrections to the Fig. 1c and 1d legends accordingly. And Fig. 4a corresponds to the original Figure 1D in Lines 386-390.

***Figure 1D:***

***- The figure lacks a legend and explanation.***

***- The purpose of the figure is unclear. A detailed description should be added.***

**Response:** Thanks very much for your helpful comment. We have restructured the figure accordingly. In the revised manuscript, Fig. 4a corresponds to the original Figure 1D, allowing for a clearer presentation of the annotation and expression profiles of the novel genes. We have added the legend and explanation in the revised manuscript (Lines 386-390).

***Line 147: ventromeric → centromeric (The term "ventromeric" does not exist in genomics.)***

**Response:** We apologize for the oversight. Revised accordingly (Line 347).

***Final remarks***

***The sheer number of typos and conceptual errors in this manuscript is concerning. While minor typographical errors are expected in any draft, the extent of errors here suggests a lack of careful proofreading and insufficient attention to detail.***

***At this stage, the manuscript requires extensive revision before it can be considered***

***for publication.***

**Response:** We sincerely appreciate the reviewer's detailed feedback and fully acknowledge the concerns regarding the typographical and conceptual errors in our initial manuscript. We take full responsibility for these shortcomings and have undertaken a comprehensive revision to address them rigorously. To resolve typographical issues, we have engaged a professional editing service to refine the language and conducted multiple rounds of proofreading with all co-authors cross-verifying the text. We implemented rigorous validation of all technical content and numerical data. We would be grateful for any further guidance to improve its quality.

## Reviewer #3 (Remarks to the Author):

***The goal of this study is to generate a near telomere-to-telomere (T2T) genome assembly for Drosophila melanogaster and to update the regulatory and genic content associated with the new assembly. To this end, the authors combine long-read sequencing data (Oxford Nanopore and PacBio), Illumina short reads, and Hi-C chromatin conformation data. The proposed assembly adds approximately 18 Mb of new sequence and resolves ~250 gaps relative to the D. melanogaster reference genome assembly (Release 6, or R6, generated in 2014). Most of the new sequence corresponds to the (peri)centromeric region of the X chromosome (+11.4 Mb), with additional sequences added to chromosome 4 (+2.1 Mb) and the pericentromeric region of chromosome arm 3R (+2.5 Mb). Notably, the new Y chromosome assembly proposes numerous segmental inversions relative to the R6 reference.***

***With this updated assembly, the authors investigate the distribution of repetitive elements (transposable elements and short repeats), describe general properties of telomeric and centromeric regions, and present an updated gene annotation. This gene annotation identifies 217 new or refined gene structures, including 105 genes with no known orthologs outside D. melanogaster. The authors also analyze ATAC-seq data, suggesting the presence of previously uncharacterized regulatory regions. Finally, they use CRISPR/Cas9 editing to validate the biological relevance of two newly annotated genes.***

***Overall, the most significant contribution of this study is the presentation of a near-T2T assembly for D. melanogaster, with expanded genic and repeat annotations. The new assembly successfully resolves many gaps present in the R6 reference assembly and adds a substantial amount of (peri)centromeric sequence. However, the total size of the assembly (~161 Mb) falls short of a true T2T or near-gapless genome, which is estimated to be ~215 Mb in males—suggesting that ~50 Mb remains unassembled. Additionally, while comparisons are consistently made against R6, the study omits comparison with several more recent long-read-based assemblies, which are known improvements over R6. Furthermore, the Hi-C data used to scaffold contigs in repetitive regions may be compromised due to limitations in the mapping strategy, raising concerns about the reliability of the added (peri)centromeric sequences.***

**Response:** We sincerely thank the reviewer for the meticulous summary of our work and for the insightful and constructive feedback aimed at enhancing it. We have meticulously addressed each point in detail below.

### Major Concerns:

***1. Assembly Quality and Comparisons:***

***The manuscript highlights that the new assembly is larger than R6, but fails to reference or compare it against other high-quality long-read assemblies (e.g., Miller et al., 2018; Chang and Larracuente, 2019; Ellison et al., 2020; Kim et al., 2021; Bologa et al., 2022). These studies already include improvements such as gap closure, expanded (peri)centromeric regions, and enhanced Y chromosome assemblies. In particular, the study by Chang and Larracuente is directly relevant because it also provides new gene and repeat annotations, and it should be used as a comparison point.***

**Response:** Thank you for drawing attention to this critical oversight; your feedback has been instrumental in strengthening the manuscript's rigor and contextualization within existing literature. We fully acknowledge that our initial submission failed to adequately reference or compare our new assembly against key long-read studies, including those by Miller et al. (2018), Chang and Larracuente (2019), Ellison et al. (2020), Kim et al. (2021), and Bologa et al. (2022). We sincerely appreciate your guidance and have taken immediate steps to address this.

In our initial investigation, we prioritized emphasizing enhancements relative to the widely used R6 reference genome, given its predominance in *Drosophila* research. However, we acknowledge that meaningful comparisons with other state-of-the-art long-read assemblies are essential to underscore the unique value of our work. To address your concern, we carefully examined and compared our genome assembly with previously published long-read *D. melanogaster* assemblies. The key assembly statistics are summarized in the table below:

| | Genome coverage | Genome size / Mp | No.contig | N50 / Mb |
|---|---|---|---|---|
| Solares et al., 2018 | 30.2× | 142.8 | 231 | 21.31 |
| Chang & Larracuente, 2019 | 11× | 155.6 | 200 | 21.69 |
| Ellison and Cao, 2020 | 54× | 134.7 | 113 | 6.6 |
| | 71× | 139.6 | 179 | 5.44 |
| Kim et al., 2021 | 52× | 144.7 | 303 | 22.5 |
| Bologa et al., 2022 | 25× | 164.4 | 3586 | 0.12 |
| | 30× | 138.9 | 1531 | 0.49 |
| **This study** | **967×** | **161.6** | **30** | **21.93** |

The above comparison demonstrates that our methodology achieves superior performance across critical genome assembly metrics, particularly in sequencing depth (ONT coverage: 967×) and assembly continuity (30 contigs, N50: 21.93 Mb). These results highlight the efficacy of our integrated multi-platform strategy, incorporating PacBio HiFi reads, ONT ultra-long reads, and Hi-C data. This approach significantly enhanced base-level accuracy while substantially improving chromosome-level assembly completeness, especially within complex genomic regions.

Furthermore, our assembly employed a distinct *D. melanogaster* strain (*Canton S*). Genetic diversity across *D. melanogaster* strains is extensively documented. We explicitly acknowledge that our novel *Canton S* genome assembly does not aim to replace the current reference genome. Instead, it serves as a complementary, high-quality resource for future research, particularly investigations focusing on this specific *Canton S* strain or requiring comparative analyses across multiple genomic sequences. Consequently, our work not only complements existing genome references both methodologically and qualitatively but also establishes a crucial foundation for developing a multi-strain *D. melanogaster* reference atlas and advancing population-scale genomic research.

***2. Assembly Size vs. Expected Genome Size:***

***The expected haploid genome size is ~215 Mb in males (~175 Mb in females) while the proposed new assembly is ~161 Mb (for males), indicating that large regions remain unassembled.***

**Response:** We acknowledge your thorough assessment of our genome assembly size. The current *D. melanogaster* reference genome spans approximately 143 Mb. Experimental measurements based on C-values estimate the haploid genome size to range between 117 Mb and 205 Mb (http://genomesize.com/results.php?page=1). Our assembled genome size of 161 Mb falls within this range, occupying an intermediate position between the reference assembly and the upper boundary. The remaining ~54 Mb discrepancy (relative to the 215 Mb C-value estimate) likely comprises: 1) Heterochromatic regions (pericentromeric and Y chromosome satellite DNA); and/or 2) Strain-specific structural variation: *Canton S* strain may harbor deletions or rearrangements relative to the strains used in C-value measurements, contributing to size differences.

Furthermore, we surveyed assemblies deposited in the NCBI database. Except for one outlier assembly annotated as “genome length too large” (~190 Mb total length), the longest reliable assembly identified was 177.2 Mb. This suggests our assembly achieves an equilibrium between completeness and accuracy, avoiding artificial inflation from misassembled repeats, a common feature in larger outliers. Despite its size, our assembly demonstrates robust quality metrics (e.g., contig number, N50, QV, and BUSCO scores), validating its utility as a high-quality resource for functional and comparative genomics.

In summary, while our assembly does not attain the full C-value estimate, it still represents a significant improvement in key regions critical for most *Drosophila* research. Its size is biologically plausible, technically justified, and balanced against accuracy qualities that enhance its utility to the community.

***3. Hi-C Data and Reliability of Scaffolding:***

***The manuscript reports using Hi-C data to scaffold contigs in repeat-rich regions. However, uniquely mapped read pairs are essential for reliable Hi-C scaffolding in these areas. The authors claim to use a Bowtie2 command (L595) to retain uniquely mapped reads, but Bowtie2 does not offer such a parameter; custom filtering is required. The absence of uniquely mapped reads likely undermines the reliability of the Hi-C scaffolding, particularly for heterochromatic regions. Indeed, the Hi-C contact map (Supplementary Figure 1) lacks robust diagonal signals for centromeres and the Y chromosome, suggesting the scaffolding is unreliable.***

**Response:** We appreciate the reviewer's expertise in highlighting this technical nuance. Our analysis pipeline explicitly retained uniquely mapped reads through **a two-step filtering approach** (Bowtie2 alignment + Post-alignment filtering). As correctly noted, Bowtie2 lacks a direct parameter for retaining uniquely mapped reads, which was first employed with the parameters `--very-sensitive -L 30` to independently map Read1 and Read2 in single-end mode. The `-L 30` setting enhances sensitivity and helps mitigate multi-mapping by favoring longer, more specific seed matches. Subsequently, for reads failing initial mapping, junction fragments at *DpnII* restriction sites were extracted, trimmed, and subjected to a second round of mapping. The alignment results from both rounds were merged, retaining exclusively read pairs where both ends exhibited unique genomic mapping. Non-conforming ligation products were removed, and only valid pairs were utilized for scaffolding via the LACHESIS pipeline. The manuscript (Lines 686-695) has been revised to clarify this filtering methodology and our operational definition of uniquely mapped read pairs.

Regarding the observed weak diagonal signals within the Hi-C contact map, particularly concerning the Y chromosome, this represents a well-documented limitation inherent to the Hi-C methodology when applied to repetitive regions characterized by low mappability. Such genomic segments inherently exhibit reduced sequence alignment rates and consequently yield sparse interaction signals. While we acknowledge that these signals are suboptimal, this constitutes an expected technical constraint inherent to the approach and does not invalidate the overall scaffolding strategy, during which we had employed stringent parameters coupled with multiple alignment-filtering steps to maximize data accuracy and reliability. Therefore, despite the inherent challenges posed by repetitive regions, we assert that our final scaffolds are technically robust and reliable within the current technological constraints.

***4. Y Chromosome Size and Quality:***

***While the R6 genome includes ~4 Mb of assembled Y chromosome sequence, this study extends it modestly (~7 Mb). However, Chang and Larracuente (2019) previously assembled ~14.6 Mb of the Y chromosome using a targeted male-specific***

***sequencing strategy that is more sensitive to repetitive elements. The present study does not represent an improvement over this prior work.***

**Response:** We sincerely appreciate your insightful comment on our Y chromosome assembly and acknowledge the critical context provided by the work of Chang and Larracuente (2019). Their pioneering use of male-specific, repeat-sensitive sequencing strategies to target Y chromosome elements has indeed set a high standard for *Drosophila* Y-linked genomic research, and we thank the reviewer for highlighting this important comparison. Below, we clarify the context, technical strengths, and unique contributions of our Y chromosome assembly:

**1) Acknowledgment of Chang's landmark achievement**

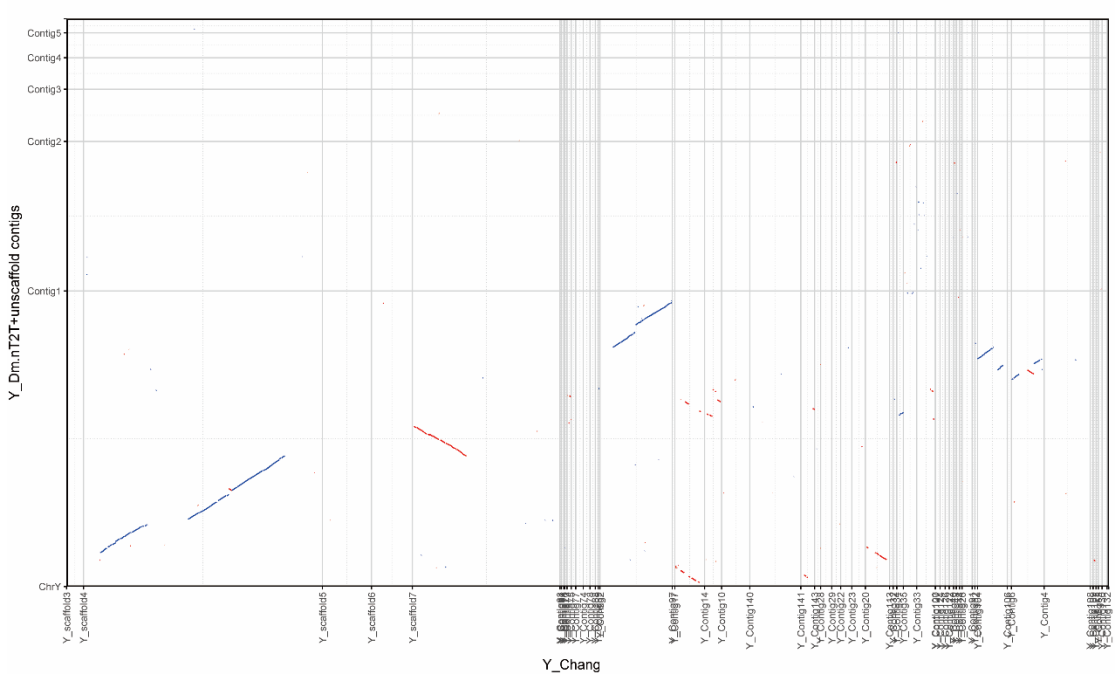
We fully recognize that Chang and Larracuente (2019) achieved a more extensive Y chromosome assembly (~14.6 Mb) through a targeted, male-specific sequencing approach optimized for repetitive regions. Their work remains a cornerstone of Y chromosome genomics, and we have explicitly cited and discussed their findings in the revised manuscript to contextualize our efforts.

**2) Quantitative comparison of Y chromosome assembly metrics**

To address your comment, we directly compared our Y chromosome assembly (Y_Dm.nT2T) with the Chang and Larracuente assembly (Y_Chang). As shown in the Supplementary Table 7, while Y_Chang captures a broader portion of the Y chromosome (larger total length), **our assembly 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). In contrast, Y_Chang's fragmentation (55 contigs) and higher ambiguity rate (No. of N bases per 100 kbp: 171.19 vs. 6.38) reflect the trade-off of prioritizing repetitive sequence recovery over structural coherence. Notably, our largest contig even exceeds the size of any individual contig in Y_Chang.

**3) Validation of Y-linked sequence confidence**

To validate the reliability of our Y chromosome assembly, we conducted collinearity analysis between the Y_Dm.nT2T+unscaffold contigs (n = 5 of 12) and the Y_Chang reference sequence. These results demonstrate that the Y chromosome has an extensive and unambiguous alignment, but no significant collinearity between the unscaffold contigs and Y_Chang (illustrated in the figure below), indicating their likely exclusion from the Y chromosome. This exclusion reinforces the precision of our assembly, ensuring that only authentic Y-linked sequences are included. Collectively, these findings support **Y_Dm.nT2T as a high-confidence Y-linked sequence, despite its smaller total length**.




**4) Strain specificity and functional utility**

Moreover, considering that Y_Dm.nT2T and Y_Chang were derived from different *D. melanogaster* strains (*Canton S* versus *ISO-1*), the genetic divergence may partly explain the observed length differences. It has been noted that genome sizes can vary significantly among strains (e.g., A4 (GCA_002300595.1): 144.1 Mb, GCA_002300595.1; ORw1118 (GCA_024500395.1): 177.2 Mb), and this could also contribute to the differences in Y chromosome size.

**5) Summary**

In summary, our study placed primary emphasis on generating a high-quality, strain-specific genome assembly for *Drosophila*, rather than specializing in Y chromosome-centric sequencing. This aim broadens its utility for diverse genetic investigations—such as gene expression analysis and variant mapping—which demand balanced chromosomal representation. Although our Y chromosome assembly does not exceed the size of Y_Chang, it exhibits distinct advantages: superior contiguity and significantly enhanced functional annotation quality. These features render it ideally suited for chromosomal anchoring and comprehensive downstream analyses. We view this work not as a direct advancement over Y_Chang, but as an essential complementary resource, thus enriching the *Drosophila* genomic toolkit with high-fidelity, strain-specific references.

***5. Y Chromosome Structure Discrepancies:***

***The new Y chromosome assembly proposes a substantial reorganization relative to R6. However, Chang and Larracuente (2019) provided a Y chromosome assembly for***

***ISO-1 that aligns well with both the R6 reference and cytological data. The authors should reassess their findings in light of this prior work. Furthermore, concerns about Hi-C reliability call into question the accuracy of the Y chromosome assembly.***

**Response:** As discussed above, we fully acknowledge the Y chromosome assembly (Y_Chang) generated by Chang and Larracuente (2019). This seminal work constitutes a crucial reference for contemporary investigations of the *Drosophila* Y chromosome. In the present study, genome assembly was achieved through whole-genome long-read sequencing integrated with Hi-C-based scaffolding strategies. To further evaluate the accuracy of our Y chromosome assembly (Y_Dm.nT2T), synteny analyses were further performed comparing Y_Dm.nT2T versus the Y chromosome from the R6 reference (Y_dm6), and Y_Chang versus Y_dm6. The results demonstrate that, compared with Y_Chang (as shown in Fig. A below), despite instances of rearrangement and potential deletion, Y_Dm.nT2T exhibits synteny with Y_dm6 across several major structural units (as shown in Fig. B below), thereby supporting the overall reliability of our assembly. We postulate that these discrepancies may reflect authentic structural polymorphisms among distinct *Drosophila* strains, consistent with observations documented in prior studies (Lines 492-509).

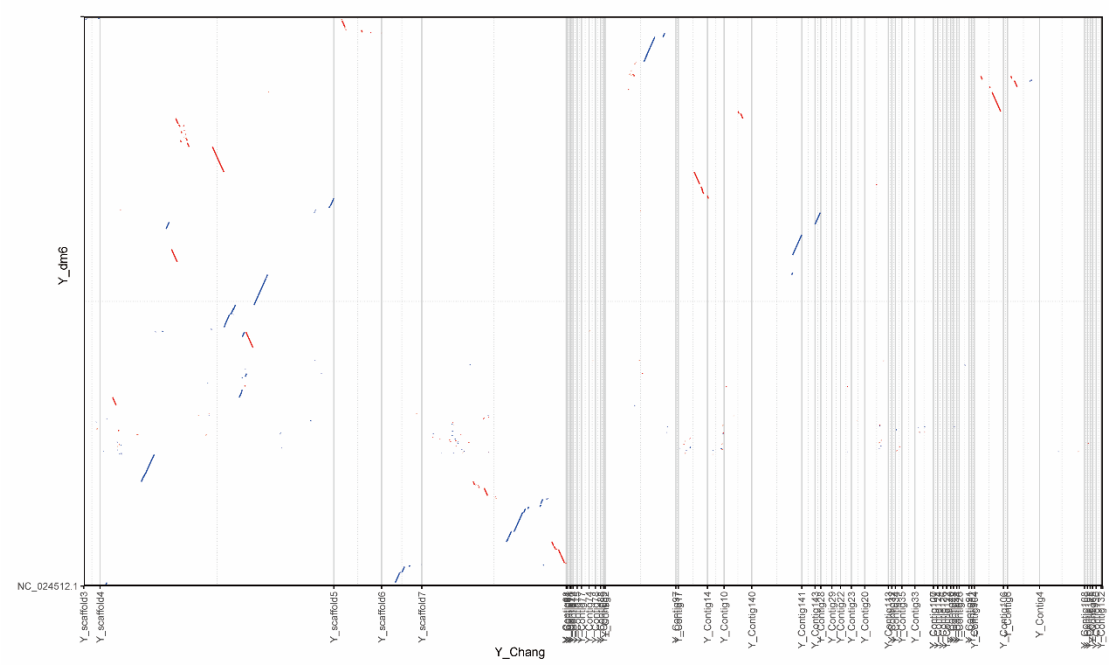



**Fig. A.** Synteny comparison of the Y chromosome between the dm6 and the Chang et al. genome assembly.

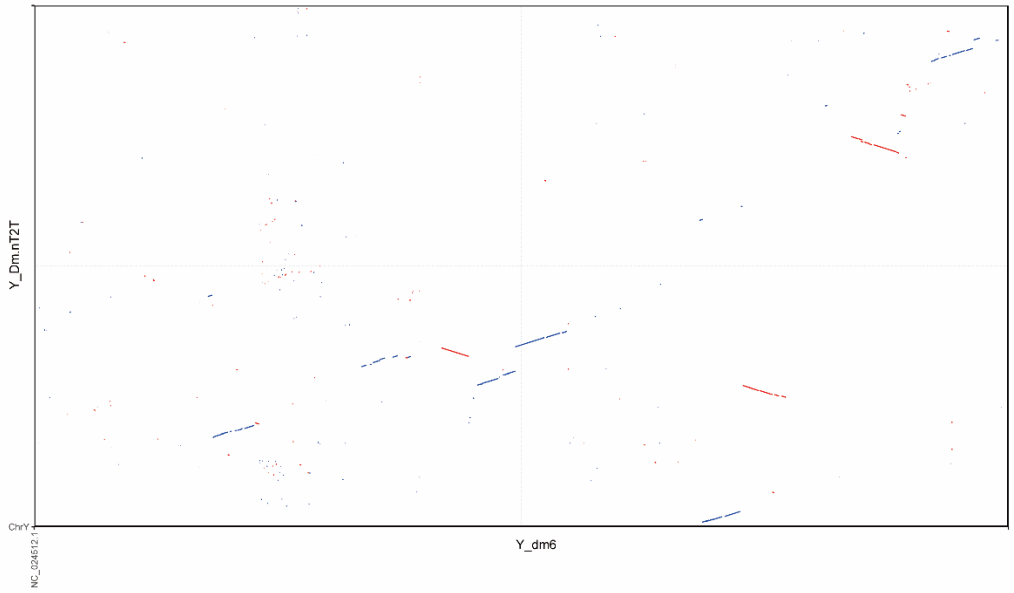



**Fig. B.** Synteny comparison of the Y chromosome between the dm6 and the Dm.nT2T genome assembly.

***Additional Concerns:***

***6.Gene Annotations:***

***6.1) "Release 6 plus ISO1" (L369) refers to the genome assembly, not a specific annotation release. The current gene annotation for R6 is r6.62 (FB2025_01), and newer annotations have added genes and transcripts. The authors must specify which version they used and compare against the most recent release.***

**Response:** We sincerely thank the reviewer for highlighting the ambiguity concerning the specific annotation version employed. In this study, we utilized the annotation obtained from NCBI, corresponding to FlyBase r6.54 (FB2023_05). We apologize for this oversight and have now explicitly stated the annotation version in the revised manuscript (Line 359). Following the reviewer's suggestion, we performed a comprehensive comparison of key annotation statistics between the r6.54 release and the latest r6.63 (FB2025_02) release. The findings are concisely summarized below. As detailed in the table below, both the gene count and overall transcript structure demonstrate remarkable consistency between these two versions. Critically, r6.63 introduced no new genes relative to r6.54. Consequently, we are confident that downstream analyses conducted using r6.54 continue to hold validity and are unlikely to be substantially impacted by differences in annotation versions. We will implement systematic updates by conducting regular audits of R6's gene annotation revisions and synchronizing these updates into our genomic annotation database through an automated pipeline, maintaining persistent version alignment between both systems.

| | r6.54 | r6.63 |
|---|---|---|
| Number of genes | 13,986 | 13,986 |
| Number of mDNA | 30,802 | 30,836 |
| mean mRNA per gene | 2.2 | 2.2 |
| Total gene length (bp) | 97,299,686 | 97,304,375 |
| Total mRNA length (bp) | 336,308,280 | 336,477,351 |
| mean gene length (bp) | 6,956 | 6,957 |
| mean mRNA length (bp) | 10,918 | 10,911 |

***6.2) When discussing new gene annotations, it would be helpful to know which ones are associated with newly assembled sequences (e.g., pericentromeric X and 3R) and which ones refine previously assembled regions.***

**Response:** We thank the reviewer for this insightful question. In our annotation pipeline, we defined novel genes as those that are absent in the R6 genome, lacking significant BLASTP hits to R6 proteins, and exhibiting transcriptional evidence. In the revised manuscript, we have refined the novel gene identification workflow—encompassing liftover analysis, blastp search, and coding potential calculation—leading to a modestly refined total count of novel genes. We identified 92 novel genes (74 protein-coding and 18 noncoding RNA genes), among which 90 exhibited corresponding DNA sequences in the R6 assembly, indicating their prior oversight in annotation. The remaining two novel genes (newly assembled genes), uniquely present in our assembly. Them was subjected to gene editing experiments within this study. These changes are now explicitly documented in the revised text, as detailed in Lines 363-371.

***6.3) Table S5 lists 72 newly defined genes (text mentions 71). A supplementary table should also list genes with updated structures.***

**Response:** Thanks for your suggestion, we have included a Supplementary Table 11 detailing all novel genes.

***6.4) The authors validate two genes using CRISPR/Cas9. For the edited gene Chr02R.722, changes in transcriptional profiles are shown. For the edited gene Chr03L.2449, behavioral assays are presented, but transcriptional effects are not. Please include data on whether editing affects this gene's expression or causes genome-wide transcriptional changes.***

**Response:** We appreciate the reviewer's inquiry regarding gene *Chr03L.2449*. In response, we additionally performed RNA-seq analysis (*Chr03L.2449* mutant flies versus wild-type controls, with three biological replicates per group) to evaluate the transcriptional impact of the CRISPR/Cas9-mediated mutation. Differential expression

analysis revealed that *Chr03L.2449* was significantly downregulated in mutant flies compared to wild-type, demonstrating a direct effect of the editing event on the gene's expression. These findings further support the functional relevance of the CRISPR/Cas9 editing. We have now incorporated these data and the corresponding description in the revised manuscript (Lines 430-438), along with two new panels (Fig. 4e,4f).

***7. Resolution of Known Gaps:***

***It remains unclear whether known problematic regions (e.g., the histone gene cluster on 2L, rDNA arrays on the X and Y) have been resolved. If so, these regions should be described.***

**Response:** This is an important suggestion. In response, we conducted a detailed assessment of the assembly quality for both the rDNA arrays and the histone gene clusters. We first evaluated the rDNA regions. Canonical rDNA sequences (18S, 5.8S, 2S, and 28S) were obtained from the R6 (GCF_000001215.4) and used as queries in BLASTn searches against our newly assembled Dm.nT2T genome (E-value ≤ 1e−20). In the Dm.nT2T genome assembly, the 18S, 5.8S, and 28S rDNA sequences were accurately located on the X chromosome, compared with the R6 genome. Notably, we also identified high-confidence matches for these sequences on the Y chromosome--whereas in the R6 reference, these rDNA units were only located on the X chromosome. This indicates a substantial improvement in rDNA resolution on the Y chromosome. In addition, the 18S and 28S rDNA sequences, which are confined to unplaced contigs in R6, were successfully anchored to both ChrX and ChrY in the Dm.nT2T genome assembly. These results demonstrate a marked improvement in the assembly of rDNA arrays, particularly on the sex chromosomes. Detailed alignment results are provided in Supplementary Table 3.

To assess the resolution of the histone gene cluster, we retrieved 111 canonical histone gene sequences from the reference genome, covering all five major histone families (H1, H2A, H2B, H3, and H4), and performed a stringent BLASTn search (≥95% identity, E-value ≤ 1e−20) against the Dm.nT2T genome assembly. All histone genes showed >95% sequence identity and mapped to a contiguous tandem repeat region on Chr2L, spanning approximately 21.4-22.1 Mb. Notably, in the R6 genome, the histone gene cluster is confined to the region 21.4-21.5 Mb on Chr2L. Therefore, our results resolve additional histone sequences that were previously missing from the region 21.5-22.1 Mb in R6. Furthermore, we found that several histone H2B sequences from Chr2L in R6 aligned well to a region on Chr2R (3.087-3.088 Mb) in our Dm.nT2T genome assembly. This indicates that, compared to the R6 genome, we have resolved previously unassembled histone sequences on both Chr2L and Chr2R. The structure of the Chr2L region is highly consistent with the known organization of the *Drosophila* histone gene

cluster, demonstrating that this previously challenging repetitive region has been successfully resolved in the Dm.nT2T assembly. Detailed alignment results are provided in Supplementary Table 4.

Taken together, these results demonstrate that the Dm.nT2T genome exhibits significantly improved assembly quality in these previously problematic regions and provides a more complete representation of the genome than previous references.

***8. Preliminary Assemblies:***

***The manuscript mentions using Canu and NextDenovo for preliminary assemblies but does not clarify whether these tools produced consistent results or how decisions were made on which contigs to retain. Please include this information.***

**Response:** Thank you very much for your valuable comment. In our initial analyses, we evaluated both Canu and NextDenovo for preliminary genome assembly. However, the assembly generated by NextDenovo demonstrated significantly superior results in terms of contiguity and accuracy, leading us to select it exclusively for all downstream analysis. Consequently, the results from Canu were excluded from the final assembly version. Regrettably, this crucial detail was not explicitly stated in the manuscript, and we inadvertently retained a reference to Canu without proper context. We have now revised the Methods and Materials section to clearly explain this decision (Lines 672-677).

***9. Pre- and Post-Hi-C Assembly Comparison:***

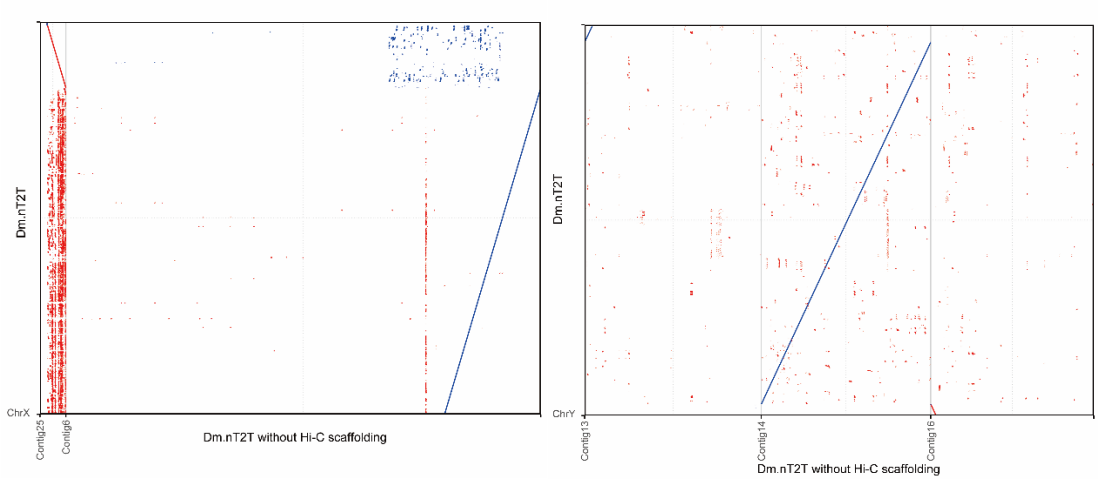
***The study should present a comparative view of the assembly before and after Hi-C scaffolding, especially for the pericentromeric region of the X and the Y chromosome.***

**Response:** We thank the reviewer for highlighting this important issue regarding the pre- and post-Hi-C assembly comparison. We performed a detailed comparison of the genome assembly before and after Hi-C scaffolding, focusing on continuity and structural integrity. As summarized in the table below, the pre-scaffolding assembly ("Dm.nT2T without Hi-C scaffolding") comprised 30 contigs, with the largest contig measuring ~25.9 Mb and an N50 of ~21.9 Mb. Remarkably, after applying Hi-C-based scaffolding ("Dm.nT2T"), the assembly was dramatically organized into 12 chromosome-scale scaffolds. This restructuring markedly enhanced contiguity (N50 = ~28.96 Mb; L50 reduced from 4 to 3) and increased the maximum contig size to ~35.1 Mb.

| | Dm.nT2T | Dm.nT2T without Hi-C scaffolding |
|---|---|---|
| # contigs (>= 0 bp) | 12 | 30 |
| # contigs (>= 1000 bp) | 12 | 30 |
| # contigs (>= 5000 bp) | 12 | 30 |
| # contigs (>= 10000 bp) | 12 | 30 |
| # contigs (>= 25000 bp) | 12 | 30 |
| # contigs (>= 50000 bp) | 12 | 30 |
| Total length (>= 0 bp) | 161631083 | 161629283 |
| Total length (>= 1000 bp) | 161631083 | 161629283 |
| Total length (>= 5000 bp) | 161631083 | 161629283 |
| Total length (>= 10000 bp) | 161631083 | 161629283 |
| Total length (>= 25000 bp) | 161631083 | 161629283 |
| Total length (>= 50000 bp) | 161631083 | 161629283 |
| # contigs | 12 | 30 |
| Largest contig | 35101060 | 25900893 |
| Total length | 161631083 | 161629283 |
| GC (%) | 41.00 | 41.00 |
| N50 | 28956753 | 21927606 |
| N90 | 24366070 | 2060620 |
| auN | 28127745.4 | 18173902.8 |
| L50 | 3 | 4 |
| L90 | 5 | 12 |
| # N's per 100 kbp | 1.11 | 0.00 |

Furthermore, we conducted a detailed comparative analysis of the pericentromeric regions of the X and Y chromosomes before and after Hi-C scaffolding. Our analyses revealed that the pre- and post-Hi-C assemblies for both chromosomes were highly collinear, indicating that Hi-C scaffolding preserved the structural integrity of the pericentromeric regions. Specifically, for the X chromosome, pericentromeric sequences initially dispersed across contig6 and contig25 were precisely integrated while maintaining strong collinearity following Hi-C scaffolding. Similarly, the Y chromosome pericentromeric region, originally spanning contig13, contig14, and contig16, also demonstrated consistent collinearity after scaffolding.

We also examined repeat content in these regions. Pericentromeric regions in X and Y chromosomes showed >98% repeat content consistently before and after Hi-C scaffolding, indicating retention of high repetitiveness without structural alterations. Together, these results demonstrate that the Hi-C data significantly improved scaffold continuity and chromosomal anchoring, especially in the structurally complex pericentromeric regions, and provide strong confidence in the large-scale integrity of the final assembly.




***10. Telomeric Elements:***

***Canonical telomeric elements (HET-A, TART, TAHRE) are included in R6. The current study adds limited new information on these sequences due to a lack of detailed analyses.***

**Response:** We sincerely appreciate your insightful comment on our analysis of canonical telomeric elements (*HET-A*, *TART*, *TAHRE*) and acknowledge that this aspect of the study requires deeper elaboration. To address your concern, we systematically compared the average length, TE composition, and genomic location of telomeric elements in the Dm.nT2T with those in the R6 reference, using the same method for telomere identification as applied in the Dm.nT2T genome assembly.

Comparative analysis highlights substantial differences in telomere representation. The total telomeric length in the Dm.nT2T genome assembly reaches 297 kb, markedly exceeding the 74.3 kb observed in R6, whose sequences remain unplaced contigs. On average, telomeres in Dm.nT2T span 49.7 kb, whereas those in R6 measure only 1.6 kb. Moreover, TE composition analysis reveals that Dm.nT2T telomeres comprise multiple TE families, including *HET-A*, *TART*, *TAHRE*, and *HETA_DSi*. In contrast, R6 telomeres contain only *HET-A* elements.

Together, these findings demonstrate that the Dm.nT2T assembly achieves a more complete and accurate resolution of telomeric regions. The extended telomeric lengths and the increased diversity of telomere-associated TEs suggest improved assembly continuity and enhanced biological fidelity compared to the previous reference.

While our study was not primarily focused on telomere biology, these additional analyses demonstrate that our *Canton S* assembly provides novel, strain-specific insights into canonical telomeric elements, complementing R6 by resolving their structure more contiguously and linking their sequence variation to functional phenotypes. Thank you again for your constructive feedback, which has been instrumental in strengthening the depth and completeness of our study.

***11. Telomeric Sequences in R6:***

***Contrary to L256–259, R6 does include telomeric sequences. The authors should correct this statement and expand their analysis of telomeres, including potential differences in length and composition across chromosome ends or between assemblies.***

**Response:** We thank the reviewer for pointing out the inaccurate statement regarding the presence of telomeric sequences in the R6 assembly. We have corrected the relevant description in the revised manuscript (Lines 300-302).

Following the suggestion, we assessed telomere lengths in the Dm.nT2T and R6 genome assemblies. We cited Hoskins et al. (2015), which utilized whole-genome optical mapping to detect and excise fragments at chromosomal ends in R6, thereby pinpointing potential telomeric regions. These findings were contrasted with regions in Dm.nT2T identified through telomere-associated retrotransposons, including *HeT-A*, *TART*, and *TAHRE*. The telomere length in R6 is based on the prediction of potential regions, whereas our method utilizes transposon alignment information to determine telomere positions and lengths, making it methodologically more accurate. As shown in the table below, both assemblies identified telomeres across six chromosomes, with Dm.nT2T uncovering telomeres for the first time on chromosomes YL and YR (undetected in the R6 version). Regarding telomere lengths: Dm.nT2T exhibited substantially longer telomeres than R6 on chromosome 2R (74.0 kb vs. 7.3 kb) and chromosome XL (76.0 kb vs. 19.6 kb), while being shorter than R6 on chromosome 2L (40.0 kb vs. 78.1 kb) and chromosome 3R (49.0 kb vs. 138.4 kb).

| | **Dm.nT2T** | **R6 (Predicted)** |
|---|---|---|
| Chr2L | 40 kb | 78.1 kb |
| Chr2R | 74 kb | 7.3 kb |
| Chr3L | - | 38.4 kb |
| Chr3R | 49 kb | 138.4 kb |
| Chr4L | - | - |
| Chr4R | - | 54.3 kb |
| ChrXL | 76 kb | 19.6 kb |
| ChrXR | - | - |
| ChrYL | 44 kb | - |
| ChrYR | 14 kb | - |

It is worth noting that the telomere lengths in the R6 assembly were predicted by Hoskins et al. (2015). However, our analysis suggests that these telomere lengths may have been overestimated. By aligning telomeric marker sequences (*HET-A*, *TART*, *TAHRE*, *TART_B1*) to the R6 genome using lastz (v1.04.52), we found that all identified telomeric regions were located on contigs that are not assembled onto chromosomes.

Based on contig IDs and the NCBI definition, these contigs can be classified into three types: Y chromosome-derived contigs, unplaced contigs, and contigs from either the X or Y chromosome. These three types of sequences are approximately 36.2 kb, 33 kb, and 5 kb in length, respectively, and all are composed solely of *HET-A* elements. In contrast, the telomeric regions in Dm.nT2T show diverse TE compositions: Chr02L (*HETA*, *TAHRE*, *HETA_DSi*), Chr02R (*TART_B1*, *TAHRE*, *HETA*, *HETA_DSi*), Chr03R (*TAHRE*, *HETA*, *HETA_DSi*), ChrX (*TAHRE*, *HETA*, *HETA_DSi*), and ChrY (*TAHRE*, *HETA_DSi*, *HETA*). These results indicate that the Dm.nT2T genome assembly provides more complete and chromosome-resolved telomeric structures.

***12. Discussion Section:***

***The Discussion section is particularly underdeveloped and contains several factual errors. Telomeric repeats (L498-499) are not AATAT (animals) or CCATA (plants); rather, common repeats in animals include TTAGG (TTAGGG in humans) whereas TTTAGGG is common on plants. Retrotransposons at telomeres are not unique to D. melanogaster (as mentioned in L502-509), and retrotransposon-based telomere structures were already known and present in R6. The study does not "resolve" centromeres or telomeres (L481)—these regions are expanded but remain incomplete.***

**Response:** Thanks very much for your insightful comment. Accordingly, we have expanded the discussion section in the revised manuscript. For instance, having identified a considerable number of SVs, we now include a dedicated discussion delving into the differences in SVs observed among various *Drosophila* strains (Lines 492-509). We have also included a discussion on the variation in telomeres across these different strains (Lines 567-581).

Furthermore, we are grateful to the reviewer for identifying our significant error in describing the Telomeric repeats. We deeply apologize for this mistake. The reviewer is correct that TTAGG represents a common telomeric repeat motif in many animals, whereas TTTAGGG is indeed typical of plants. We have modified this description accordingly in the revised manuscript (Lines 548-549).

We have also corrected this sentence: "*We also found that this telomere pattern is absent in the Drosophila genus and even in other insects, such as butterflies and dragonflies, where other species still adhere to the general patterns*" (Lines 551-566). Indeed, similar retrotransposon-based telomere mechanisms have been documented in specific species like *D. virilis* and *D. yakuba* (Casacuberta & Pardue, 2003; 2005).

Finally, we acknowledge that our assembly has not fully resolved centromeric and telomeric regions, as some gaps remain in our genome despite achieving a sequencing depth of ~1000×. We have refined the text to more accurately convey that our study has partially enhanced and extended the assembly in these regions, rather than claiming full resolution. However, these gaps do not significantly impact our overall findings. Future

efforts will focus on closing these gaps to enhance the completeness of the genome assembly, ensuring a more comprehensive understanding of telomere dynamics in these species. We appreciate the reviewer for highlighting this critical detail.

***13. Cross-Species Gene Annotation:***

***Provide version numbers, accession IDs, and links for gene annotations used from D. yakuba, D. simulans, D. sechellia, D. erecta, and D. ananassae.***

**Response:** We appreciate the reviewer's suggestion. The gene annotations used in our study were all downloaded from the NCBI Genome database. These annotation files (in GFF3 format) were downloaded directly from the NCBI Assembly portal (https://www.ncbi.nlm.nih.gov/assembly/). Below are the accession numbers and corresponding download links for each *Drosophila* species. This relevant information has also been added to Supplementary Table 14 and described in the revised manuscript in the Materials and Methods section (Lines 738-742).

| Species | Accession number | Source link |
|---|---|---|
| *D. yakuba* | GCF_016746365.2 | https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/016/746/365/GCF_016746365.2_Prin_Dyak_Tai18E2_2.1/GCF_016746365.2_Prin_Dyak_Tai18E2_2.1_genomic.gff.gz |
| *D. simulans* | GCF_016746395.2 | https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/016/746/395/GCF_016746395.2_Prin_Dsim_3.1/GCF_016746395.2_Prin_Dsim_3.1_genomic.gff.gz |
| *D. sechellia* | GCA_004382195.2 | https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/004/382/195/GCF_004382195.2_ASM438219v2/GCF_004382195.2_ASM438219v2_genomic.gff.gz |
| *D. erecta* | GCF_003286155.1 | https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/003/286/155/GCF_003286155.1_DereRS2/GCF_003286155.1_DereRS2_genomic.gff.gz |
| *D. ananassae* | GCF_017639315.1 | https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/017/639/315/GCF_017639315.1_ASM1763931v2/GCF_017639315.1_ASM1763931v2_genomic.gff.gz |

**Reviewer #1 (Remarks to the Author):**

***Liu et al., in their revised version of "Near Complete Genome Assembly of Drosophila melanogaster Unveils Novel Genes and Complex Genomic Regions" have successfully addressed all my concerns and questions. The manuscript is now worthy of publication.***

**Response:** Thanks very much for your positive evaluation and careful reading. We are delighted that our revisions addressed your concerns and you find the manuscript suitable for publication. Your thoughtful comments and encouragement are greatly appreciated.

**Reviewer #2 (Remarks to the Author):**

***The authors have addressed most of my questions (though not all) adequately. In my view, the manuscript has been improved and is now more clearly written. I believe it can make a valuable contribution to the field. I hope that future genome assembly studies based on this work will further serve as a resource for the scientific community.***

**Response:** Thanks for your careful reading and constructive comments. We're pleased revisions improved clarity and you see it as a valuable contribution. We appreciate your encouragement and hope our work serves as a resource for future genome assembly studies.

**Reviewer #3 (Remarks to the Author):**

***#1. Previous D. melanogaster assemblies based on long-read sequences were not mentioned.***
***The authors have a detailed and sensible response to my concern, "…However, we acknowledge that meaningful comparisons with other state-of-the-art long-read assemblies are essential to underscore the unique value of our work. To address your concern, we carefully examined and compared our genome assembly with previously published long-read D. melanogaster assemblies". Indeed, many of these assemblies had already improved the R6 assembly used here as reference. An important aspect of the authors' response is a table showing the main properties of previous long-read-based assemblies, and those from this nT2T. This allows a reader to appreciate the advances made relative to R6 before this study, and the additional advances of the proposed nT2T presented here.***

***This informative Table, however, is only shown in the 'response to reviewers', but not***

***included in the revised version of this manuscript. I believe it should be included as Table or as Supplemental data. Perhaps the best place to describe previous assemblies and this table would be after line 85, where the authors indicate that long-reads, sometimes with Hi-C, have been used to generate close-to-T2T in several organisms.***

**Response:** Thanks for this suggestion. In the revised manuscript, we have incorporated the comparative table as Supplementary Table 3 (Lines 125-128).

***#2. rDNA clusters.***

***The authors describe a cluster localized on the ChrX (lines 133-134) and then write (line 135), "Notably, we also detected high-confidence matches for these rDNA units on ChrY, a region where the R6 genome lacks such annotations, indicating a significant enhancement in rDNA resolution." As written, a reader could infer that the presence of an rDNA cluster in the Y chromosome was an unexpected finding. Although it is true that R6 does not have rDNA on its short Y chromosome sequence, the presence (and some key characteristics) of an rDNA cluster in the Y chromosome has been known for many years, with a predicted size of 2.2Mb (Bianciardi et al. 2012; Ritossa 1976; Polanco et al. 1998). Adding these previous studies would add context and help a reader to appreciate better the new data (we knew the cluster there and its size, now we know its sequence…).***

**Response:** We appreciate the reviewer for this clarification. In the revised manuscript, we have cited these foundational studies and clarified that our finding does not represent novel detection of this cluster. Instead, it provides the first complete, sequence-resolved characterization of this locus (lines 136-137).

***#3. Telomeric sequences and variation between strains.***

***The authors indicate (lines 295-196) that the telomeric regions of this nT2T (297kb) are markedly larger than those in R6 (74.3 kb), arguing that many contigs that look like telomeric in R6 "remain in unplaced contigs". If the authors argument is correct, one cannot trust the telomeric sequences presented in R6 and the order of their components is unreliable. As a consequence, the authors' proposal (lines 577-581) that the telomere structural differences "more likely reflect genuine polymorphisms between strains:.", is unwarranted at this point and should be reassessed, or removed. Either R6 is reliable (and comparisons to nT2T can reveal variation between strains and polymorphism), or it is not (in which case apparent differences cannot infer true polymorphism).***

**Response:** We sincerely thank the reviewer for this insightful comment. We agree that the incomplete representation of telomeric regions in R6 compromises our ability to definitively attribute the observed differences to true polymorphisms between strains.

In the revised manuscript, we have carefully revised this statement to clarify that the major differences likely stem from the more complete reconstruction of telomeric regions in the Dm.nT2T assembly (Lines 277-281).

***#4. SINE elements in D. melanogaster.***

***The authors write (line 478), "A striking discovery was the identification of SINE transposable elements (0.02% of the genome), previously unreported in D. melanogaster." This result can be important but I would like to request the authors to provide some key information. First, please report the consensus sequence of this D. melanogaster SINE to allow the community to expand this report in other strains and species (as it is now it is not possible). Second, SINE-like elements have been reported in D. melanogaster in the past, but without much follow-up. Please discuss possible overlap of the proposed SINE with DINE-1 (Kaminker et al. 2002) and/or suffix (Tchurikov and Kretova 2007) elements.***

**Response:** Thanks to the reviewer for these comments. We have now added the consensus sequences of the SINE element in Supplementary Table 8 (Line 441). We also evaluated the potential relationship between the newly identified SINE elements and previously characterized SINE-like elements through pairwise sequence comparisons. Our analysis showed that the SINE elements characterized in this study may constitute a novel and independent SINE family within the *D. melanogaster* genome. A brief discussion of these results has been incorporated into the revised manuscript (Lines 439-446).

***#5. Annotation.***

***The authors have included the new genome assembly (fasta) and annotation (gff) files (accessed via figshare). Unless I missed it, I could not find annotations (location and sequence) of any transposable element (including SINEs) across the new nT2T genome sequences. Please, include the annotation of the transposable elements reported in this study.***

**Response:** Thanks a lot for your comments. We have now added the TE annotation to the figshare at https://figshare.com/articles/dataset/The_near_complete_genome_assembly_of_Drosophila_melanogaster/28642964. The figshare link has been updated accordingly in the revised manuscript (Lines 905-906).

***#6. Hi-C and repetitive sequences.***

***Please, include the command lines used to identify uniquely mapped Hi-C reads, which is essential for reliable Hi-C analyses of repetitive sequences (centromeres and the Y chromosome). As written, the methods do not describe how this was achieved.***

**Response:** We thank the reviewer for the insightful comment regarding this technical detail. We used the following Bowtie2 command for Hi-C read alignment: *bowtie2 --very-sensitive -L 30 --score-min L, -0.6, -0.2*. These parameters were selected to maximize mapping specificity and minimize multi-mapping. Although this approach does not strictly enforce unique mapped reads at the software level, the use of a longer seed length (-L 30) and a stringent alignment scoring threshold (--score-min L, -0.6, -0.2) effectively increases mapping specificity and substantially reduces multi-mapped reads. The same software and parameters were adopted by Wang et al. (*Nature Genetics*, 2025) for Hi-C data processing in repeat-rich genomes. We have revised the Methods section (Lines 650-656) to include the exact command line and to clarify our description in the revised manuscript.

**ref:** 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).

***Other:***

***- Line 55. "harbors" might be the wrong word. "However, despite ongoing studies seeking to improve genome completeness of this classic model organism, its genome still harbors the structurally complex regions"***

**Response:** Revised. Thanks (Line 52).

***- The Canton-S strain used in this study is said to come from the Bloomington Drosophila Stock Center (BDSC_64349). A minor issue is that this strain was replaced on Feb 2016. Please indicate when the strain was obtained from BDSC.***

**Response:** Thank the reviewer for pointing this detail. The *Canton-S* strain (BDSC_64349) used in this study was obtained from the Bloomington Drosophila Stock Center in 2018. This information has been added to the revised manuscript (Line 581).

***- Line 108. "… the average read length was 26,487.9 bp, with.." is already mentioned in line 104 ("…ONT ultra-long read length was 26,487.9 bp.")***

**Response:** Thank you for pointing this out. We have removed the redundant mention of the ONT average read length and revised the sentence in Lines 108-110.