

Telomere-to-telomere genome assembly of the Dipteran *Bactrocera dorsalis* from a single individual

Corresponding Author: Professor Guirong Wang

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)

The manuscript by Liu et al describes the first T2T assembly for Diptera, improving over existing assemblies for crop pest tephritid fly *Bactrocera dorsalis*. The paper presents a considerable advancement for insect genomics and includes appropriate and quite detailed exploration of genomic regions that were previously missing from the chromosomal assembly. The paper is generally well written (with some minor obscurities) and generously illustrated. My main concern for the present version is that much of technical details are lacking, especially in sequencing libraries preparation methods - arguably, the section that would be one of the most interesting for the readers interested in T2T assembly generation from small and challenging to maintain organisms. Computational methods description would also benefit from command listings being deposited, however I understand this might not be feasible at this point.

More detailed comments in order of appearance in text:

l.90 - please provide statistics for pooled Illumina sequencing

l.94-95 - "average read length" - is it mean, median or N50?

l.101-102 - statistics of ONT and Hi-C sequencing are not given anywhere. For ONT, I would be keen to see the read length distribution profile, especially given that length filtering was extensively used during the assembly

Paragraph on assembly strategy (starting at l.104) looks like it belongs to methods - and it is also somewhat more clearly structured. Also, its understanding would benefit a lot from sharing the code implementation. In results, I would expect to see justifications for the methods chosen as well as some statistics

l.108-109 - "broken up and chopped with 10kb" - unclear what this means

l.110 - no parametrisation is given for coverage-based sex chromosome identification

l.116 - "network graph generated from unitigs" - I would just say unitig graph

l.127 - histone cluster gap fill done with identical copies? Or some sequence variation was evaluated and accounted for?

l.129 - I would appreciate more elaboration on scaffold validation: based on Fig 1e, several large-scale interventions were done, is it possible to provide statistics on these changes - and ideally some example juicer plots comparing before and after intervention?

l.130 - again, some polishing statistics would be appreciated. Also, the method did not look like a standard one-tool solution to me, so justifications or references on the selected approach would be appreciated

l.142 - other genome assemblies for the same species, right?

l.157 - I'd suggest a paragraph break between repeat and gene annotations

Paragraph starting at l.164 stylistically belong to Discussion. Also, TE-associated centromeres and telomeres of *Drosophila* seem to be not very representative for Diptera in general, where both telomeres (<https://doi.org/10.1093/biolinnean/blad068>) and centromeres (<https://doi.org/10.1007/s00412-011-0320-2>, <https://doi.org/10.1371/journal.pone.0079393>) often seem associated with satellite DNA.

l.210-211 - unclear what "48% of the ONT read estimation" means and how this is connected to peaks Fig S4a (and what is the interpretation of these peaks)

l.217-219 - this seems to belong to the next paragraph discussing rDNA-sex chromosome link in more detail.

l.242 - phylogenetic clustering of NUMTS remains undiscussed. Do you think this reflects common origin, e.g. via duplications involving NUMTs? Or this could be some technical artefact?

l.258 - "dosage compensation" remains undiscussed, next sentence is focused on heterochromatin vs euchromatin activity differences, not whole-chromosome X-autosomal expression balance compensation mechanism as I was expecting.

l.280 - was MoY detected in the same exact species?

l.292 - could it be a R1-dependent retrogene?

l.313 onwards - synteny patterns might be moved to discussion. As this paragraph deals with large-scale evolutionary patterns, I would expect some citations for Muller elements concept. Also, for *Drosophila* and mosquitoes, within-arm large-scale inversions are one of main drivers of chromosomal evolution - indeed, these can be driven by recombination between TE copies. Re ref 53, I would argue that HT of TEs would mostly affects the repeat landscape, i.e. which repeats are being active in a given species, but not synteny specifically. The statement at l.321-322 looks a bit speculative for results section. Discussions of sex chromosome transitions in Diptera might benefit from incorporating insights from this ref <https://doi.org/10.1371/journal.pbio.1002078>

l.356 - here, 90% aa sequence identity is mentioned, later in text (l.697) similar estimates appear in nucleotide context

l.358-363 - I'm not too familiar with the field, but is it possible that chimeric transcripts in case of these very close tandem genes might result from transcription not terminating for the first gene, and proceeding to second and third? Then, alternative splicing could combine transcripts from different genes.

l.429 - "polyphagous host range" - consider rephrasing

l.436 - other sequencing technologies omitted

Sample collection, extraction and sequencing methods subsection does not include any details on DNA extractions, very brief detail on library prep. While this might be ok for Illumina, where the process is highly standardised, PacBio and ONT protocols details are very important to understand the specifics of the data produced.

- For PacBio, which version of LI protocol was used? what were the size selection parameters?

- For ONT, which DNA extraction protocol you were able to use to yield long DNA fragments? How many samples you used and how much DNA was yielded? What was the pre-size-selection fragment length profile? What were the size selection parameters? How much DNA was left after size selection?

- For Hi-C, what kit (and version) were used? What were the QC procedures?

- For RNAseq, no indication of DNA extraction, RT, QC, library prep, or sequencing platforms are given. Sequencing statistics were not presented or discussed either.

l.482 - "QV of 0.99" - this does not seem right, QV is $\gg 1$ for CCS

l.492 - hifiasm v334 - possibly r334 (v0.15.1)?

l.495 - refseq version does not seem right

Paragraph starting at l.495 is hard to interpret in terms of implementation - was the contamination removal run three times (reads, assemblies, read overlaps)? If so, how do contigs with $>40\%$ ref overlap appear if all reads with $>40\%$ ref overlap were removed? If all filtering was done at once, why does the last sentence say only clean reads were used for assemblies, while filtering mentioned contigs?

l.507 - please insert paragraph break after genomescope section, it is separate from long read assembly

Starting from l.509, this seems like ONT read mapping and read ends uniformity exploration corresponding to l105-109 in results - both of those were not very clear to me, so please consider rephrasing. Also, what was used to map ONT reads?

Paragraph starting at l.517 is a mixture of: pacbio read mapping, coverage based sex chromosome identification, ont alignment, purging, scaffolding, assembly curation. None of those are described completely and the logic of including all of these procedures into one paragraph was not clear. Some of the data are seemingly reused in later paragraphs in the section dealing with gap closing and polishing. Overall, the whole process descriptions leaves many open questions, many of which would be easy to answer with code listing available:

- what issues are addressed by read end uniformity testing and subsequent breaks?

- what is the benefit of generating alignments with three tools and were all of these tools used to both pacbio and ont reads?

- why sex chromosome linkage is estimated early in the process? Does this affect any of the downstream processes?

- how do you use information on low-complexity regions definition in gap closing with ont? The process as it is described seems to be agnostic towards the underlying graph structure

- What exactly happens during alternate contig removal? Hi-C data is mapped to linear genomes, so reference to graphs is not relevant.

- For ONT you mention the problem of structural variation coming from pooled DNA sample and the workaround of manual investigation of called SVs. What was the decision making process? What would you do if SV overlapped region not resolved by pacbio data?

l.549 - why the centromere annotation cannot be done on final T2T linear version of the assembly?

l.592 - references to NCBI datasets for proteins are not given.

l.595 - what is the logic for 5kbp intron length cutoff? This does not seem long enough to me

l.597 - "manually integrated" - probably not completely manually? Also, what were the rules for integration of genes missing in one or two methods?

l.626 - what was the query for the NUMT search? If this mt assembly was generated as part of your assembly, what were the processes for its finishing and QC to ensure that mitochondrial contig is correctly represented?

l.631 - table s-mt - incorrect reference

l.642 - methods for differential expression section seemingly missing (or mis-placed). Please also check if any additional methods are needed regarding OR genes repertoire investigation (anything for Fig6a-d)?

l.667 - consider adding more details about target sequences (or ref to supplementary)

l.708-711 - insufficient level of detail, but assume it follows ref 4 (mentioned in l.679)

Overall, main figures were way too small when printed - assume web version will still work

l.776 (fig2) - what are "hifiasm units"?

Fig 6c - legend on tissues for expression heatmap is needed

Supp Fig 2 - are all three assemblies from the same species? If so, something should be said about the extent of large-scale structural variation between the assemblies. If those are different species, these should be indicated together with some notion of relative divergence

Supp Fig 4 eluded my understanding - what do the peaks in sf4a represent? Three rRNAs and their lengths - or maybe something else? What does the subplot b demonstrate? Do you see a canonical structure and some deviations from that?

I.859 (SF5) "multiplex species" - maybe just gene copies?

SF6 - in printed version, I don't see any signal in rows 2 and 3. What are horizontal and vertical lines? If those delineate chromosomes, why horizontals are different if you have the same genome on y axis?

SF7 extends Fig6 e-f - maybe include a reference?

Reviewer #2

(Remarks to the Author)

This manuscript presents a telomere-to-telomere (T2T) assembly of the *Bactrocera dorsalis* genome using long-read sequencing. Based on this assembly, the authors provide comprehensive analyses of centromeres, telomeres, nuclear-mitochondrial DNA segments (NUMTs), sex chromosomes, and tandemly duplicated gene families, particularly olfactory receptors (ORs). Through CRISPR knockout (KO) experiments, they demonstrate that one OR locus (BdorOR88a12) is involved in detecting methyl eugenol (ME), a compound used for pest control. This study represents a substantial contribution to both agricultural pest management and the understanding of species diversity. However, several aspects of the manuscript require improvement.

Major Concerns

1. Figure Layout and Legends

- o The font size in figures is too small and should be increased for better readability.

- o Several figure legends are incomplete or unclear. For example:

Clarify the meaning of green and blue clades in Figure 4e and define GR/OR in Figure 6a.

Correct missing or inaccurate legends in Figures 4g and 5b.

- o Figure order errors (e.g., Figure 1b appears before Figure 1a in Line 93) should be addressed.

- o Specific parts of Figures 1a and 1b lack clarity. For instance:

In Step 2 of Figure 1a, the disconnected core element appears identical across situations.

The long strings in Steps 2 and 3 seem to have inconsistent meanings. These diagrams and their legends need clearer explanations.

2. T2T Assembly Quality

- o Include details on sequencing read depth and coverage for methods such as HiFi, ONT, NGS, and Hi-C to enhance reproducibility.

- o Consider adding GCI as a metric to assess assembly quality (e.g., <https://academic.oup.com/bioinformatics/article/40/11/btae633/7829147>).

- o Compare this assembly with previous versions by specifying the number of novel genes or exons newly annotated.

3. BdorOR88a12 and KO Experiments (Figure 6)

- o Explore whether different transcript isoforms can be distinguished and quantified. Consider qPCR with transcript-specific primers to determine if the gene's function relies on a single or chimeric transcript.

- o Clarify which isoform was knocked out and elaborate on the functional consequences of the KO experiments.

- o Verify sgRNA design for repetitive loci like ORs, ensuring mappability and minimizing off-target effects.

Minor Concerns

1. The manuscript extensively discusses telomeres on sex chromosomes but provides limited information on autosomal telomeres. Elaborate on the significance of the 528 bp feature and the presence or absence of TTAGGG repeats.

2. Clarify why only 459 single-copy orthologous genes were used, given the more extensive annotation in related species like *D. melanogaster*.

3. When analyzing centromeres, the authors relied on TRF (v4.09.1) and two candidate sequences for FISH. This approach is insufficient for centromere validation, requiring ChIP-seq or CUT&Tag. Terms like "candidate centromere" should be used to temper claims.

4. Address the potential conflict between the two software tools mentioned in the "Raw data preprocessing" section (Lines 481–502).

5. Improve NUMT mutation analysis:

- o Explain why only three NUMTs are shown in Figure 3a, despite identifying 11 NUMTs in Figure 3b. Are these the largest?

- o The genomic average GC content is 40%, while the average GC content of MT DNA is 28%. Consistent with the authors' hypothesis that NUMT GC content evolved from 28% toward 40%, older NUMT copies should exhibit a stronger AT → GC mutation bias compared to the MT reference genome. This appears to be the case as in the second column of Figure 3b (as shown in the bottom branch including MT13 or MT5). I suggest that authors include this point in the main text.

6. Correct minor errors:

- o Line 636: Fix the GitHub link to "<https://github.com/soybase/crossmap-workflow>."

- o Line 679: Resolve the citation format error.

- o Clarify which Hifiasm version (v334 or 0.18.5) was used.

Reviewer #3

(Remarks to the Author)

In this paper, Liu et al. generate a T2T genome assembly of the pest species *Bactrocera dorsalis*. This insect is of economic importance and the author's approach illustrates the process to be followed to characterize the genome of similar pests for which amenable approaches such as those typically used for *Drosophila* species are of limited use due to e.g. high levels of polymorphism, difficulty to generate inbred lines, etc... First, combining HiFi and ON sequencing, the authors generate a genome assembly, which is highly contiguous, with complex repetitive regions resolved, and with telomeres, centromeres, and the Y chromosome represented. Key foci in this manuscript whose organization is uncovered include centromeric and

telomeric regions as well NUMT insertion loci. Further, the new assembly includes the sex chromosomes, and the authors addressed several aspects including dosage compensation, sex-bias differences in expression, and the characterization of a Y-linked gene that are relevant for sex-determination during early development including their paralogs in other chromosomes. At the comparative genomic level, the authors examine the level of maintenance of the ancestral chromosomal synteny at different phylogenetic distances, as well as the relationship of the X chromosome of *B. dorsalis* with those in other species. In the last section of the results, the authors explore different aspects of the OR gene families, including chromosome and cluster distribution, expression and formation of chimeric transcripts for one of the clusters, and phenotypic effects resulting from knocking out two adjacent OR paralogs.

Issues and concerns:

1. Improvements in relation to previously published genome assemblies and gene annotations in *B. dorsalis*. Although the authors devote several sections to highlight key regions (centromere, telomere, NUMT loci) that are now well represented in the T2T genome assembly, it is not clear whether the euchromatic fraction has significantly improved. For example, the authors provide a comparison of the T2T assembly with other published assemblies (Sup. Fig. 2, line 142). Are there differences in the orientation of blocks of adjacent genes? What about the number of copies of some multigene families arranged in tandem? And relative to repetitive sequences?
2. Something similar to point 1 applies to the gene annotation. There are no clear indications about how the new gene annotation differs from previous ones in NCBI for this same species.
Put simply (1 & 2), the putative improvement in the resources generated by the authors relative to previous ones are just superficially presented.
3. Sex-expression differences. First, the authors should clarify whether some of the differences in gene expression are impacted by allometric differences between the sexes rather than differences in transcription rate. Second, it is not fully clear to me Fig. 4c. Specifically, what do the authors mean with DEGs in both Female and Male? Perhaps that a gene is male-biased in one anatomical part but female-biased in a different one. This should be clarified.
4. Synteny conservation interpretation (line 315). Synteny conservation can also be explained by the lack of interchromosomal rearrangements, which have a higher detrimental effect than intrachromosomal rearrangements (e.g. inversions) during meiosis. It is also unclear the explanation given in line 311 to the outcome of the phylogenomic tree. The argument seems a bit speculative or not explained properly.
5. Synteny conservation analysis (paragraph starting in line 313). The analysis does not seem easily reproducible. To start with, the authors should provide the list of 459 single-copy genes with their names and chromosomal and molecular coordinates across all the species considered [by the way, in line 645 there is a reference to a Supplementary Table but it is not clear to which one]. Besides, the result (lack of gene order conservation is expected given the prevalence of paracentric inversions in this type of insects). Lastly, the last bit on transposons must be rewritten. The authors should become familiar with how TEs can facilitate, through ectopic recombination events, the generation of chromosomal inversions in *Drosophila* (the literature is abundant; see for example some key papers from Michael Ashburner and Alfredo Ruiz). Said this, it is not unusual that chromosomal inversions do not get generated by TEs in the genus *Drosophila*.
6. Interpretation of the difference in OR copy number between *D. melanogaster*/mosquitoes and *B. dorsalis* (lines 343-346). What do the authors mean by *B. dorsalis* has evolved over the past few million years, therefore a higher copy number in OR genes might have facilitated its adaptation to diverse environments? For example, the most recent ancestor of *D. melanogaster* with the *simulans* clade goes back to 1.4 million years ago only. In my view, the authors are overinterpreting a simple difference in copy number across clades, which might relate to a differential presence of a genomic feature yet to be identified, e.g. a particular sequence feature able to mediate duplications in tandem; a sequence feature that could be more abundant in *B. dorsalis*. This issue is found again in the Discussion section (paragraph lines 418-433), in which the comparison drawn in relation to plant genomes is by the way unnecessary.
7. Sentence in lines 403-405. Please clarify how the genome resources generated here actually provide new insights into chromosome structure polymorphism. It is unclear to which results the authors are referring.
8. Last paragraph in the Discussion. In the way the authors provide some of the results, I do not think that some of the resources generated in this article are necessary useful to the research community. For example, the authors could provide the list of orthogroups and their constituent genes between *B. dorsalis* and the rest of species used in comparative analyses (Supplementary table 11), being the one between *B. dorsalis* and *D. melanogaster* extremely useful to many colleagues.

Minor comments:

1. More informative than the average read length is the equivalent of N50 applied to the length distribution of sequencing reads (lines 94-95).
2. It would be nice to see the % that the different contaminants represented for both initial sequencing outputs (to be added to Sup. Table 1).

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

The updated manuscript features addresses the issues raised in the previous round of review. In particular, much more methodology details are given, which allows for evaluation of the assembly process and definitely adds to the appreciation of amount of manual adjustments still needed to finalise T2T assembly even for relatively compact genome of a dipteran insect. I also appreciated that most of the figures in the response were also added as supplementary figures and believe that

the remaining ones used for demonstrations for the assembly process (Hi-C plot of haplotypic duplication, coverage breaks in misjoins) can be also added to the supplementary material.

Reviewer #2

(Remarks to the Author)

I appreciate the authors' significant efforts in addressing my previous concerns. Most of my comments have been satisfactorily resolved, resulting in a substantial improvement in the manuscript's quality. With the following minor issues addressed, the manuscript will be ready for publication:

1. "Assembly continuity achieves a GCI score of 31.5 for the whole genome, and 51.4 and 57.0 for X and Y chromosomes separately."
Typically, the Y chromosome is the most challenging to assemble due to its overrepresentation of repetitive sequences, which usually results in a lower GCI score. However, the authors report the highest score for the Y chromosome, which is unexpected. Since GCI software is relatively new, it may not perform optimally on certain genomes. Please verify whether the software failed in this specific case. If so, consider omitting this information.
2. "However, due to the high sequence similarity (90%) of the CDS regions among the three copies, current qPCR methods cannot precisely distinguish or quantify individual copies."
Please explicitly state this limitation to caution readers.
3. "The results showed that the expression pattern of the three copies remained unchanged after the knockout."
This statement is unclear. If these copies were knocked out, how could the expression pattern remain unchanged? Please clarify the writing, particularly in the main text.

Additionally, the link to the China National GeneBank Database (CNCBdb) in the Data Availability section is currently inaccessible. Please ensure the resource is made available upon formal acceptance of the manuscript.

Reviewer #3

(Remarks to the Author)

The new submission ameliorates some important limitations (transparency and level of detail) identified in the first version. But a few aspects could still benefit from adding further details and accuracy:

1. Reply to point 5, reviewer 3, new text. The sentence 'The mechanistic basis of this evolutionary stasis appears multifaceted. In *Drosophila*, phylogenetic reconstruction demonstrates that large-scale intra-arm inversions dominate chromosomal evolution while preserving global gene order through compensatory rearrangements' is a bit speculative. Specifically, the part underlined. What do the authors mean by "compensatory rearrangements"? What is the evidence for that statement?
2. Like above, I am not sure what is the point of the underlined text in the following sentence: 'Studies from *Drosophila* genus have shown this stability is maintained through large-scale within-arm rearrangements caused by inversion, although the molecular drivers remain enigmatic'. We know from Ashburner and Ruiz's publications that transposable elements can mediate the origin of some chromosomal inversions while for others it has not been possible to find evidence of such TE involvement, and instead duplications of flanking regions at the inversion breakpoints have been documented. This led to the proposed staggered-breakpoints model (ref. 68). In sum, there is either involvement of TEs or not. We have examples of both types of scenarios. Therefore, what are the authors referring to with "enigmatic"?
3. Data availability section. It is indicated that the chromosome assembly 'will be' submitted to the China National GeneBank Database. Shouldn't it be already submitted and available?
4. Code availability. It seems unnecessary providing the code for built-in functions (e.g. in R) used in conventional statistical analyses, but I see it necessary in the case of the analytical framework used to generate the genome assembly of *B. dorsalis* (and even the one for the differential expression analysis). It would dispel doubts, enhance the transparency of the work done, and increase the visibility of the paper.
5. In the supplementary file with the orthogroups composition, tracking the identification of the genes from different species is difficult. Additionally, for *D. melanogaster*, the FlyBase identifier (e.g. FBgn0038979) is preferred.
6. Relative to the analyses presented in Fig. 5. The choice of species is strange in some cases. For example, *Anopheles coluzzii* is included but not an emblematic species such as *Anopheles gambiae*, which in theory has a much better assembly and annotation. Or in the genus *Drosophila*, the authors choose *Drosophila mauritiana* along with *Drosophila melanogaster*, which gives the impression of conservation of gene order and strong chromosomal stasis. These two are sister species, having diverged only a few million years ago, and differentiated by one single inversion. The authors could have chosen *Drosophila pseudoobscura* or *Drosophila virilis*, much more distantly related to *Drosophila melanogaster* to make the fair point that despite the large intra-chromosomal repatterning, there is no exchange between different Muller's elements. In sum, the rationale for choosing some species is far from clear and justified, and in fact a bit misleading.
7. Reply to point 5, reviewer 3. The authors could be more explicit, indicating the name of the file, which I believe is `Bdo_phaseBlks.csv_6spe`. They can do that in the Methods.

Version 2:

Reviewer comments:

Reviewer #2

(Remarks to the Author)

My remaining minor comments have been addressed well, and I have no further comments.

Reviewer #3

(Remarks to the Author)

I appreciate the authors' effort for improving unclear aspects of the manuscript. Nevertheless, there are still some minimal inaccuracies that should be fixed. In reply to this reviewer's comments:

Point 1. The new text is correct and the citation (67) to Prof. Schaeffer's review is pertinent.

Point 2. The text in lines 436-443 is unnecessarily long with some errors. Reference 67 does not talk specifically about the comparison between *D. simulans* and *D. yakuba*. That comparison, which involves the analysis of sequence features at the breakpoint regions of chromosomal inversions, is in fact part of the content in reference 68. Please read again the paragraph that begins with "Nucleotide sequences at inversion breakpoints have provided clues about how ..." in reference 67.

Points 3, 4, 5, and 7. Great improvements.

Point 6. I appreciate the inclusion of *D. pseudoobscura*. One thing that surprised me though is that now the phylogenetic analysis includes 1914 single-copy shared orthologs while in previous versions of the manuscript this number was 459. I think it is positive that the analyses had been repeated with a larger number of genes, but I am surprised that the authors do not indicate in their reply why this dramatic disparity. Perhaps I am missing something obvious but what has changed?

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

The manuscript by Liu et al describes the first T2T assembly for Diptera, improving over existing assemblies for crop pest tephritid fly Bactrocera dorsalis. The paper presents a considerable advancement for insect genomics and includes appropriate and quite detailed exploration of genomic regions that were previously missing from the chromosomal assembly. The paper is generally well written (with some minor obscurities) and generously illustrated. My main concern for the present version is that much of technical details are lacking, especially in sequencing libraries preparation methods - arguably, the section that would be one of the most interesting for the readers interested in T2T assembly generation from small and challenging to maintain organisms. Computational methods description would also benefit from command listings being deposited, however I understand this might not be feasible at this point.

Authors' response:

Thank you very much for your valuable suggestions and detailed review. Your efforts and insightful comments have greatly improved our manuscript. We deeply appreciate your time and assistance.

More detailed comments in order of appearance in text:

l.90 - please provide statistics for pooled Illumina sequencing

Authors' response:

We appreciate your comment.

We have added the statistics for Illumina sequencing to the main text as follows:

“The genetic background was assessed with 23.06 Gb Illumina sequencing data of pooled DNA from five individuals using GenomeScope, which revealed an average heterozygosity of 1.81%.” Additionally, we have included a new supplementary table (Supplementary Table 1-1) with detailed information.”

Supplementary Table 1 - Statistics of sequencing data

(1) Illumina sequencing of pooled DNA from five individuals

Metrics	Read1		Read2	
	raw data	after filtering	raw data	after filtering
Total Reads	77,113,009	77,112,955	77,113,009	77,112,955
Total Bases	11,531,065,295	11,531,051,099	11,529,467,949	11,529,451,867
Q20 Bases (%)	11,256,480,024	11,256,469,684	11,190,612,026	11,190,600,513
	(97.62%)	(97.62%)	(97.06%)	(97.06%)
Q30 Bases (%)	10,735,850,583	10,735,841,554	10,548,788,094	10,548,779,317
	(93.10%)	(93.10%)	(91.49%)	(91.49%)
Mapping rate	98.36%			

l.94-95 - "average read length" - is it mean, median or N50?

Authors' response:

Thank you for your comment.

Yes, this refers to the mean read length. The corresponding statistics have been provided in Supplementary Table 1-2.

Supplementary Table 1 - Statistics of sequencing data					
(2) Reads of third generation sequencing					
	HiFi-CCS female	for Filtered HiFi-CCS female	for HiFi-CCS male	Filtered for male	HiFi-CCS ONT reads
Mean read length	11,500	11,550	12,335	12,287	37,792
Mean read quality	30	30	28	28	13
Median read length	11,089	11,139	11,906	11,858	30,941
Median read quality	38	38	33	33	14
Number of reads	2,593,480	2,222,421	1,585,639	1,551,322	2,508,117
Read length N50	11,416	11,470	12,194	12,150	59,916
STDEV read length	1,943	1,948	1,814	1,817	32,783
Total bases	29,824,224,634	25,667,967,376	19,558,761,563	19,060,311,110	94,787,558,719

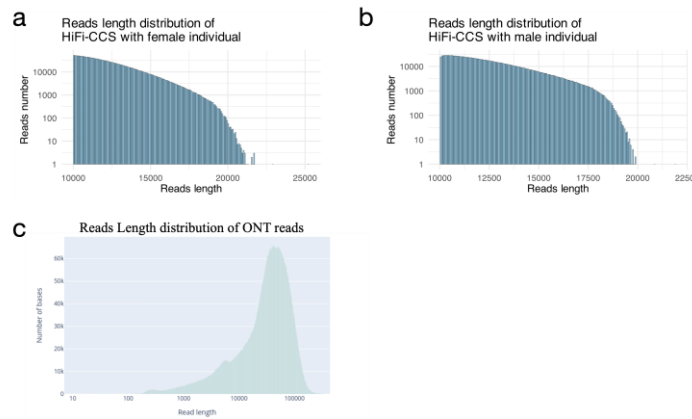
l.101-102 - statistics of ONT and Hi-C sequencing are not given anywhere. For ONT, I would be keen to see the read length distribution profile, especially given that length filtering was extensively used during the assembly

Authors' response:

Thank you for your comment.

We have provided detailed statistics of ONT and Hi-C sequencing in Supplementary Table 1-3. The reads length distribution profile was shown in Supplementary Figure 2.

Supplementary Fig. 2 The reads length distribution profile of HiFi-CCS and ONT reads



Paragraph on assembly strategy (starting at l.104) looks like it belongs to methods - and it is also somewhat more clearly structured. Also, its understanding would benefit a lot from sharing the code implementation. In results, I would expect to see justifications for the methods chosen as well as some statistics.

Authors' response:

Thank you for your comment.

To minimize redundancy with the methods section, we removed the two paragraphs from main text and refined a detailed description of method choices and parameter statistics is included in the methods part.

Refined method section:

“Genome assembly

(1) General assembly procedure with Hifiasm, purged_dups and Juicer pipeline

Initial assembly were assembled with HiFi CCS reads by using hifiasm (r603) with parameters (`--hg-size 540 m--primary -r 5 -D 6 -m 1000000 --b-cov 2 -l 1 -O 2`). The primary superscaffold (p_utg) produced by Hifiasm underwent haplotypic duplication removal using Purge_Dups (v1.2.6). The Juicer84 (v1.6) and 3D-DNA pipelines⁸⁵ (v2.20) were used to anchor contigs to pseudochromosomes with Hi-C data. The results were further manually refined in Juicebox⁸⁶ (v2.20).

This finalized assembly was employed to facilitate the identification of candidate regions corresponding to X chromosome fragments, telomeric sequences, and centromeric sequences.

(2) Unitigs validation:

ONT reads were mapped to hifiasm unitig contigs with minimap2 (v2.26), retaining only those reads with a mapping length of at least 30 kb and a mapping length coverage above 95%. For each 200 bp genomic bin, the relative distance between the mapped reads and genome loci (calculated as $(\text{read_center} - \text{genome_loci}) / \text{read_length}$) was computed. This value is expected to follow a uniform distribution within the range of $[-0.5, 0.5]$. The Anderson–Darling and Shapiro–Wilk normality tests were applied to evaluate this distribution, excluding terminal 8 kb regions of contigs from calculation. The 200 loci with the lowest p-values were manually inspected, and the 12 suspicious loci with disproportionate read distributions were resolved by breaking the corresponding contigs. Additionally, approximately 10 kb were trimmed from the newly generated contig ends to avoid potential erroneous extensions caused by individual HiFi CCS reads.

(3) Sex chromosome assembling

We first aimed to assemble the Y chromosome, considering its small size, abundant repetitive sequences, and weak Hi-C interactions.

HiFi CCS reads from male and female samples were aligned to initial hifiasm unitigs and contigs using Winnowmap²⁸⁷ (v2.03), chosen for its superior performance in repetitive regions. Sex chromosome candidate fragments were identified based on differences in read coverage, assuming that the coverage of autosomes should follow a binomial distribution. Differential fragment coverage was analyzed using DEGseq⁸⁸, applying a stringent cutoff of p-value < 0.001 to select candidate sex chromosomes contigs. Candidate contigs rich in repetitive elements, such as histone or rDNA sequences, were excluded. Several validated Y chromosome markers, including the GIGYF family protein (MT978097) and Moy protein, previously

identified as transcribed Y-specific markers of *B.dorsalis*, were used to confirm contigs belonging to the Y chromosome.

For the X chromosome, the primary assembly relied on the p_utg contigs. Structural continuity was manually confirmed by assessing the alignment consistency of unitigs and HiFi CCS reads. Mapping result of Nanopore long reads (>80 kb) was extracted and converted into a connected graph of unitigs. Relationships among contigs were manually verified and extended based on maximal read overlaps and minimal mismatches.”

(4) Identification of genome structural features

The terminal ends of the original hifiasm- derived unitigs and contigs were extended using ONT long reads to characterize repetitive centromeric and telomeric elements. The “Hifiasm+purged_dups+Hi-C” based genome assembly was generated to facilitate the candidate selection. Given no centromeric repeat units had been previously reported for *B.dorsalis*, we hypothesized that they might be located near the ends of the assembled super-contigs.

For monomer sequences of centromere and telomere, the detailed analysis was shown in the following corresponding sections.

(5) Scaffolding

Telomeric, centromeric, histone, and rDNA arrays in assembled unitigs (r_utg) were annotated using RepeatMasker and visualized as “yarn balls” in Bandage89 (v0.8.1). Based on repetitive element composition, unitig graph was divided into low-complexity, repeat-rich regions (primarily derived from heterochromatic and scattered regions) and high-complexity, gene-rich regions.

For low-complexity regions (high repetitive in sequence), contigs along with their neighboring regions were isolated from assemble graph. Contig path surrounding the centromere and telomere was resolved by using the primary unitig graph and was further simplified through graph alignment and gap closing with ONT reads:

ONT reads were mapped to hifiasm r_utg.gfa using GraphAligner90 (v1.0.17) was used to map ONT reads to hifiasm “r_utg.gfa”. Only reads meeting the following criteria were retained: (1) 95% of the length aligned to the genome (to simplify the assembly graph and construct the longest contig connection), and (2) at least one terminal alignment longer than 20 kb and an NM value/mapping length < 0.05. If the ends covered multiple contigs, the alignment had to support the existing r_utg contig connection. New connection relationships were established using mutual best alignments inferred from Minimap2, Winnowmap2, and GraphAligner.

For high-complexity regions, hifiasm outputs were verified and redundancies removed based on Hi-C interactions and ONT alignments to remove additional alternate contigs in the diploid bubble structure.

(6) Genome polish

Due to the high heterozygosity in the population, short-read or ONT reads-based polishing would introduce more structural diversity. Thus, we limited our approach to using unitigs and super contigs to replace shorter paths within bubble regions, effectively creating a hifiasm primary-like assembly. We first called SNVs with ONT reads (>40kb) using SVision91. Homozygous deletions or insertions > 3 kb, supported by more than 10 reads, were manually verified at 8 loci by aligning ONT/HiFi CCS reads. Suspected errors at 5 loci were corrected using assembled contig sequences.

(7) Gap filling for histone region

The only remaining gap in the assembly was a histone cluster located in the pericentromeric region of chromosome 2, which could not be directly closed using long ONT reads or string graphs. The copy number of the histone gene cluster was estimated to be approximately 170 copies based on ONT reads and approximately 150 copies based on HiFi CCS reads. Due to high sequence conservation (Supplementary Figure XX), the final assembly filled this gap with 150 copies of the histone gene, comprising approximately 802 kb. The heterogeneous contigs and scaffolding in the Hifiasm results were manually validated using Hi-C information. Finally, the assembly was polished with primary unitigs and contigs.”

l.108-109 - "broken up and chopped with 10kb" - unclear what this means

Authors' response:

Thank you for your comment.

The suspicious loci with disproportionate read distributions were resolved by breaking up the contigs, and approximately 10 kb from the new termini of these contigs were trimmed to avoid potential erroneous extensions caused by individual HiFi CCS reads (average length ~15 kb). We have updated the corresponding description in the Methods section.

l.110 - no parametrisation is given for coverage-based sex chromosome identification

Authors' response:

Thank you for your comment.

We have updated the corresponding description in the Methods section as previous response in section of “Sex chromosome assembling”.

l.116 - "network graph generated from unitigs" - I would just say unitig graph

Authors' response:

Thank you for your comment.

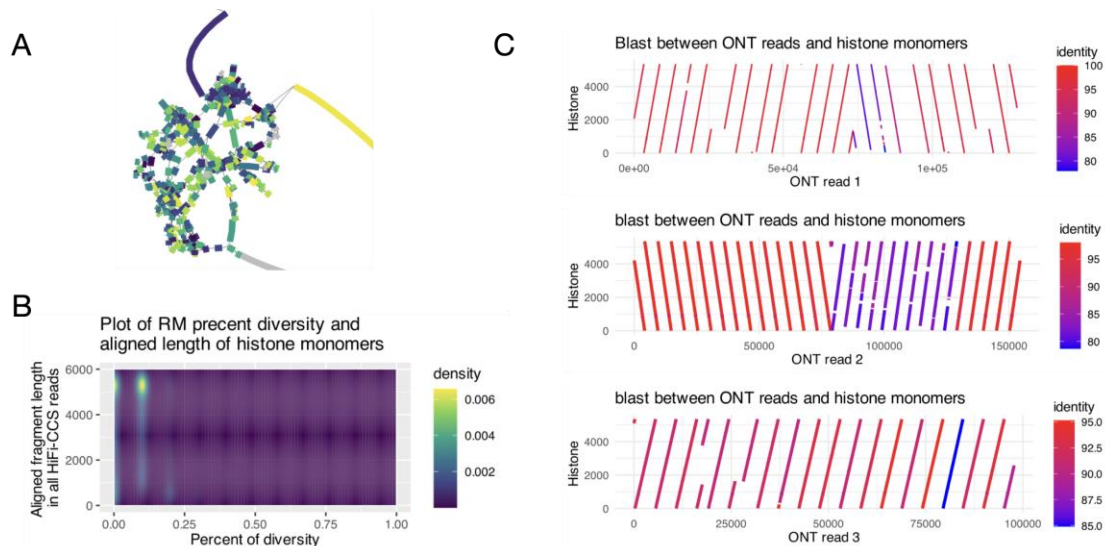
The phrase has been simplified to "unitig graph" as suggested.

l.127 - histone cluster gap fill done with identical copies? Or some sequence variation was evaluated and accounted for?

Authors' response:

Thank you for your comment.

The histone array regions were highly complex in the initial assembly, making it difficult to resolve them solely based on read overlaps (A). Therefore, we assessed the histone-associated segments in the original HiFi CCS reads annotated by RepeatMasker. Most annotated regions showed high sequence conservation (B). Additionally, we randomly selected three longer histone-associated reads from the ONT dataset, further confirming that the histone units are highly conserved in both sequence length and sequence similarity (C).

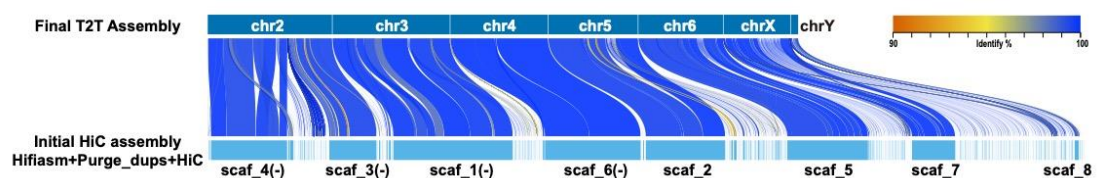


l.129 - I would appreciate more elaboration on scaffold validation: based on Fig 1e, several large-scale interventions were done, is it possible to provide statistics on these changes - and ideally some example juicer plots comparing before and after intervention?

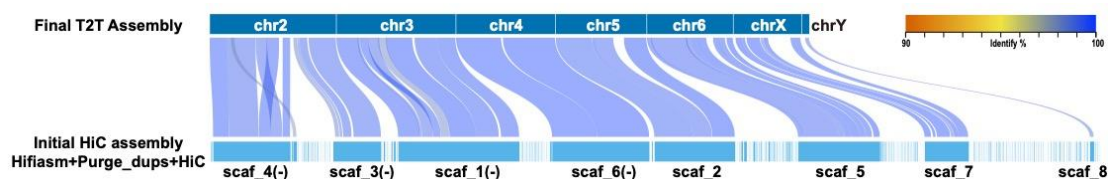
Authors' response:

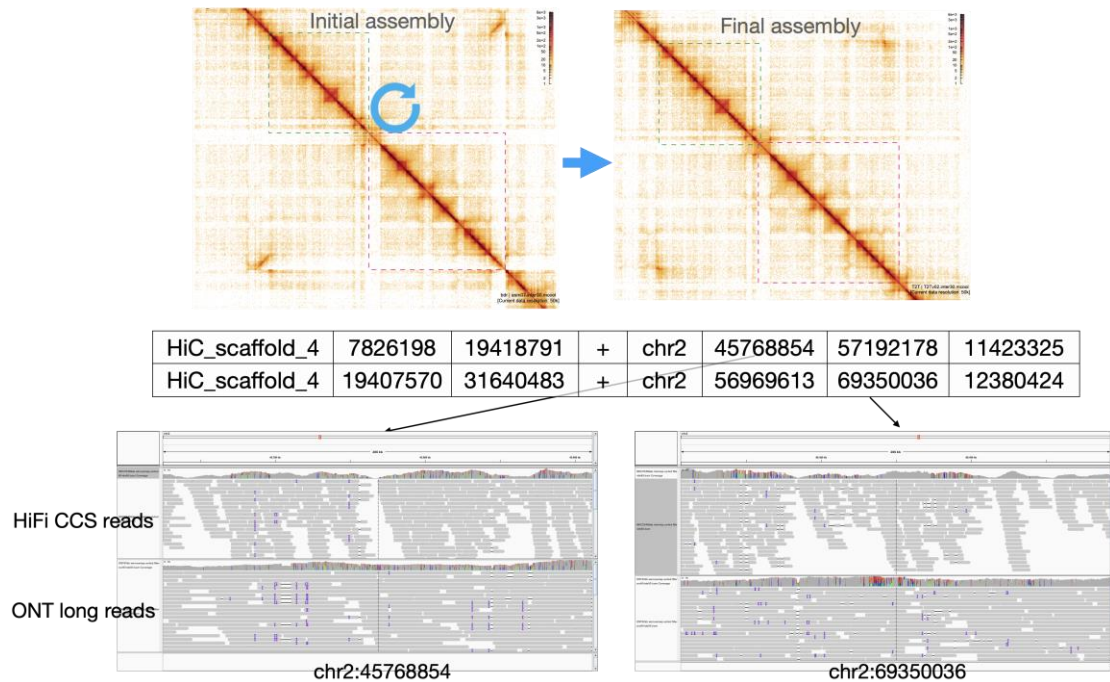
Thank you for your comment.

The most significant large-scale rearrangements were due to chromosome directions of these genomes were not consistent with each other, thus showing as unnecessary inversions between the two assemblies. We have adjusted the orientation of most chromosomes/scaffolds so that they consistently point in the same direction. Only a few large-scale inversions remain, as clearly indicated in the updated Fig 1e of the revised submission.



Here, we highlighted differences based on alignment results filtered with a minimum of 90% identity and a 2 Mb block length. Only two major inversions remain after filtering. We further emphasized supporting evidence from Hi-C data, demonstrating the superior quality of the T2T assembly.





l.130 - again, some polishing statistics would be appreciated. Also, the method did not look like a standard one-tool solution to me, so justifications or references on the selected approach would be appreciated

Authors' response:

Thank you for your comment.

The polishing procedure for a highly heterozygous single individual's genome assembly was based on two main considerations:

First, the average coverage of HiFi CCS reads exceeded 30x in homozygous regions of the diploid genome. This depth of coverage allows accurate and straightforward assembly with hifiasm without the need for additional polishing using sequencing data from other individuals. Second, standard polishing pipelines were unsuitable due to high heterozygosity, which introduced haplotype mis-assembly and significant noise resulting from alignment ambiguities. This caused issues in conventional polishing methods, such as NextPolish, which frequently generated warnings like:

“Warning, Too many (12.477%) fragment mappings in chr6, please polish the genome with other reads first, or adjust the mapping parameters to tolerate more errors, such as use asm20/map-pb instead of asm5 for minimap2, continue anyway...”

Therefore, we limited our approach to using unitigs and super-contigs to replace shorter paths within bubble regions, effectively creating a hifiasm primary-like assembly. Additionally, ONT reads longer than 40 kb were employed for structural variant (SV) calling using SVision. SV regions longer than 3 kb and supported by more than 10 reads (8 location) were manually reviewed and corrected (5 location), taking p-utg and r-utg contig alignment results into account. We have updated the corresponding description in the Methods section as previous response.

l.142 - other genome assemblies for the same species, right?

Authors' response:

Thank you for your comment.

Yes, we modified the original sentence to “other published genome assemblies of *B. dorsalis*”

l.157 - I'd suggest a paragraph break between repeat and gene annotations

Authors' response:

Thank you for your comment.

We have broken the two parts.

Paragraph starting at l.164 stylistically belong to Discussion. Also, TE-associated centromeres and telomeres of Drosophila seem to be not very representative for Diptera in general, where both telomeres (<https://doi.org/10.1093/biolinnean/blad068>) and centromeres (<https://doi.org/10.1007/s00412-011-0320-2>, <https://doi.org/10.1371/journal.pone.0079393>) often seem associated with satellite DNA.

Authors' response:

Thank you for your comment.

We have shortened the introductory text of this paragraph in the main text and moved part of the content to the “Discussion section”. For centromeres, genomic screening detected no significant sequence homology to known dipteran centromeric satellites candidates - specifically the BoR300 satellite of *Bactrocera oleae*¹ or the AT-rich centromeric repeats of *Sciara coprophila*².

Main text:

“Both centromeres and telomeres play key roles in maintaining genome stability and mediate the precisely equal distribution of genetic material during cell division. Unlike other insects, species within the order Diptera exhibit high diversity in their telomeric and centromeric motifs.”

Discussion of telomere sequence:

“Notably, Dipteran species have evolved telomerase-independent mechanisms for chromosomal end maintenance, with two distinct telomeric architectures identified to date³. The telomeric structure of *D.melanogaster* can extend through several kilobases and consists of a mixed repetitive array of HetA/TART/TAHRE elements, along with telomere-associated sequences^{4,5}. The other type, found in lower Diptera, consists of long tandemly repeated sequences³. The telomeric element sequence of *B.dorsalis* differs substantially from other seven species of the suborder Brachycera, all of which have the long repetitive element in the end of chromosome.”

l.210-211 - unclear what "48% of the ONT read estimation" means and how this is connected to peaks Fig S4a (and what is the interpretation of these peaks)

Authors' response:

Thank you for your comment.

We have updated the figure legend for Fig. S4a.

“All large subunit rDNA (LSU-rDNA) related ONT long reads (>80 kb) were identified using srfutils.js ‘reads enlong’ and aligned with minimap2 using the parameters (-c -N1000000 -f1000 -r100,100). Aligned LSU-rDNA blocks were filtered with a minimum coverage and identity threshold of 0.2. If a read contains insertions within LSU-rDNA element, the aligned block was fragmented into two regions of different lengths. In the raw ONT long reads, the LSU-rDNA homologous regions displayed three distinct length peaks at 1,247 bp, 2,697 bp, and 3,939 bp (the complete unit), indicated that approximately 48% of LSU-rDNA units were fragmented into two parts rather than existing as intact units.”

l.217-219 - this seems to belong to the next paragraph discussing rDNA-sex chromosome link in more detail.

Authors’ response:

Thank you for your comment. We have revised the order of this paragraph.

l.242 – phylogenetic clustering of NUMTS remains undiscussed. Do you think this reflects common origin, e.g. via duplications involving NUMTs? Or this could be some technical artefact?

Authors’ response:

Thank you for your comment.

We hypothesize the presence of NUMTs in the nuclear genome represent recurrent independent integration events of mitochondrial genome sequences into the nuclear genome over evolutionary time. throughout evolutionary history.

Phylogenetic reconstruction rules out NUMTs duplication mechanisms, as ancestral insertions show progressive sequence erosion while contemporary NUMTs display intact structure of mitochondrial genome. Additionally, technical artifacts are unlikely, given the consistent coverage by HiFi CCS and ONT reads both within and at the boundaries of these regions, with the existing structural arrangement supported by long ONT reads spanning entire NUMTs loci.

l.258 - “dosage compensation” remains undiscussed, next sentence is focused on heterochromatin vs euchromatin activity differences, not whole-chromosome X-autosomal expression balance compensation mechanism as I was expecting.

Authors’ response:

Thank you for your comment. We have revised the main text as following:

In *Drosophila*, dosage compensation is achieved by hyperactivating the single male X chromosome. At the male X:A ratio (0.5), autosomal Dpn repressors outweigh X-encoded Sis activators, silencing the X-encoded *Sxl* gene. The male-specific MSL2 protein (which suppressed by *sxl*) assembles the MSL/DCC complex to bind the X and acetylates histone H4K16 to loosen chromatin dosage compensation. However, in *Bactrocera* species, sex determination mechanism does not involve *Sxl*⁶. Whether a similar dosage compensation mechanism exists remains unclear and warrants further investigation.

l.280 - was MoY detected in the same exact species?

Authors’ response:

Thank you for your comment.

Yes, several studies have described presence of Moy gene ortholog in Tephritidae family⁶, although its upstream and downstream functions remain unresolved, and with previous studies consistently reporting a single-copy genomic configuration. In contrast, our results show that there are three Moy paralogs on the Y chromosome, of which at least two copies significantly expressed during early stage of embryonic development.

l.292 - could it be a R1-dependent retrogene?

Authors' response:

Thank you for your comment.

Yes. Regions annotated as “non-LTR retrotransposon R1” are indicated as locations with “reverse transcriptase-like” features in Fig4. a and b.

We wish to clarify a correction in the original manuscript: The genomic locus referenced in this section was erroneously labeled and has been updated to BdoYG000019 in the revised version.

l.313 onwards - synteny patterns might be moved to discussion. As this paragraph deals with large-scale evolutionary patterns, I would expect some citations for Muller elements concept. Also, for Drosophila and mosquitoes, within-arm large-scale inversions are one of main drivers of chromosomal evolution - indeed, these can be driven by recombination between TE copies. Re ref 53, I would argue that HT of TEs would mostly affects the repeat landscape, i.e. which repeats are being active in a given species, but not synteny specifically. The statement at l.321-322 looks a bit speculative for results section.

Authors' response:

We sincerely appreciate the constructive feedback and have implemented the following revisions to address these concerns:

- (1) This synteny pattern discussion part has been moved to Discussion section.
- (2) Combined with your valuable suggestion and the constructive feedback from Reviewer 3, we emphasized the Muller elements concept to explain the structural evolution of the genome. We also removed references to DNA transposons and revised that section as follows

Main text revised in discussion as following:

“The conservation of ancestral chromosomal synteny demonstrates extraordinary evolutionary persistence across insect lineages. In species such as lepidopterans and other insects, ancestral chromosomal synteny is often retained even after millions of years of divergence, suggesting that gene content conservation plays a crucial role in maintaining functional stability⁷. Our analysis revealed that gene content of Muller elements remains stable across the species we examined. in suborder Brachycera, and this conservation can even be traced back to the *Anopheles gambiae* genome⁸. The mechanistic basis of this evolutionary stasis appears multifaceted. In *Drosophila*, phylogenetic reconstruction demonstrates that large-scale intra-arm inversions dominate chromosomal evolution while preserving global gene order through compensatory rearrangements.

Studies from *Drosophila* genus have shown this stability is maintained through large-scale within-arm rearrangements caused by inversion, although the molecular drivers remain enigmatic⁹. Comparison among *Drosophila melanogaster* and two closely related species,

Drosophila simulans and *Drosophila yakuba*, suggest the mechanism of staggered breaks, forming inverted duplication sequence, primarily drive the generation of inversions¹⁰. ”

Discussions of sex chromosome transitions in Diptera might benefit from incorporating insights from this ref <https://doi.org/10.1371/journal.pbio.1002078>

Authors' response:

Thank you for your comment.

We have included this section in the main text:

“This observation aligns with the paradigm-shifting findings of Vicoso and Bachtrog¹¹, who demonstrated that Diptera have undergone an independent sex-chromosome transition, challenging the traditional view of stable XY systems in this order. Specifically, the syntenic conservation between the *B. dorsalis* X chromosome and *D. melanogaster* chromosome 4 (Muller F element) suggests ancestral Diptera pattern, where the dot chromosome (Muller F) typically served as the proto-sex chromosome.”

l.356 - here, 90% aa sequence identity is mentioned, later in text (l.697) similar estimates appear in nucleotide context

Authors' response:

Thank you for your comment. We have revised the description for consistency.

The identity of both amino acid and nucleotide sequences is highly similar across the three copies of BdorOR88a:

Amino acids: 1 vs 2 = 90.58%; 1 vs 3 = 93.43%; 2 vs 3 = 94.93%.

Nucleotides: 1 vs 2 = 92.93%; 1 vs 3 = 93.77%; 2 vs 3 = 96.39%.

l.358-363 - I'm not too familiar with the field, but is it possible that chimeric transcripts in case of these very close tandem genes might result from transcription not terminating for the first gene, and proceeding to second and third? Then, alternative splicing could combine transcripts from different genes.

Authors' response:

Thank you for your comment.

Chimeric transcripts are typically formed through two mechanisms: trans-splicing and cis-splicing. Trans-splicing is common in cases where two genes are located at different positions in the genome, such as the classic example of JAZF1-JJAZ1 in endometrial stromal sarcomas (ESSs) in humans (Li et al. 2008). Cis-splicing involved the read through (i.e., cis-splicing between adjacent genes) occurs more frequently between neighboring genes, exemplified by SLC45A3-ELK4, which is closely associated with prostate cancer (Zhang et al. 2012). In this study, the three 88a gene copies involved are indeed adjacent genes. However, we have not detected transcription signals from the intergenic regions in the transcriptome. Therefore, we hypothesize that these chimeric transcripts might still be formed through trans-splicing. Nevertheless, further evidence is needed to elucidate the exact mechanism.

Li H, Wang JL, Mor G, Sklar J 2008. A neoplastic gene fusion mimics trans-splicing of RNAs in normal Human cells. *Science*. 321: 1357-1361.

Zhang YM, Gong M, Yuan HL, Park HG, Frierson HF, Li H 2012. Chimeric transcript generated by cis-splicing of adjacent genes regulates prostate cancer cell proliferation. *Cancer Discovery*. 2(7): 598-607.

l.429 - "polyphagous host range" - consider rephrasing

Authors' response:

We revised them as "potentially supporting its broad host range".

l.436 - other sequencing technologies omitted

Sample collection, extraction and sequencing methods subsection does not include any details on DNA extractions, very brief detail on library prep. While this might be ok for Illumina, where the process is highly standardised, PacBio and ONT protocols details are very important to understand the specifics of the data produced.

Authors' response:

Thank you for your comment. Additional details of the experimental methodology are provided below:

DNA-extraction

Tissue was ground into a fine powder with liquid nitrogen and lysed in SDS lysis buffer at the appropriate temperature. After centrifugation, the supernatant was subjected to phenol-chloroform extraction, and nucleic acids were precipitated with isopropanol. The precipitates were washed twice with ethanol, air-dried, and dissolved in TE buffer overnight to obtain high-quality DNA.

RNA-extraction

Total RNA was isolated from tissues using the TRIzol reagent (Invitrogen, Carlsbad, CA, United States) according to the manufacturer's instructions. RNA integrity was assessed using the Fragment Analyzer 5400 (Agilent Technologies, CA, USA).

- For PacBio, which version of LI protocol was used? what were the size selection parameters?

Authors' response:

The HiFi library was prepared using the SMRTbell® Express Template Prep Kit 2.0. Starting with 1.7 µg of high molecular weight (HMW) genomic DNA (most fragments >30 kb), and the DNA were sheared and an average size of ~15kb was retained.

- For ONT, which DNA extraction protocol you were able to use to yield long DNA fragments? How many samples you used and how much DNA was yielded? What was the pre-size-selection fragment length profile? What were the size selection parameters? How much DNA was left after size selection?

Authors' response:

The procedure included DNA fragmentation with diluted FRA, rapid adapter ligation, and DNA precipitation with PPT buffer. All steps were performed with care to preserve long DNA fragments, utilizing wide-bore tips to avoid mechanical shearing. For the Oxford Nanopore ultra-long library preparation, 78.3 µg of long DNA, extracted from a mixture of five larvae and exhibiting a prominent band around 100 kb, was size-selected and processed using the

Ligation Sequencing Kit (SQK-ULK001; Oxford Nanopore Technologies, Oxford, UK) following the manufacturer's instructions. Sequencing was subsequently performed on the PromethION sequencer with an R9.4 chip (Oxford Nanopore Technologies, Oxford, UK).

- For Hi-C, what kit (and version) were used? What were the QC procedures?

Authors' response:

Library construction was carried out following the standard protocol described by Belton et al. (2012), with some modifications. The sample was ground in liquid nitrogen and crosslinked with 4% formaldehyde under vacuum at room temperature for 30 minutes. The reaction was quenched with 2.5 M glycine. After centrifugation and washing with buffer, the sample pellet was resuspended in lysis buffer for nuclei isolation. The nuclei were solubilized with diluted SDS, incubated at 65°C for 10 minutes, and neutralized with Triton X-100. DNA was digested with the 4-cutter restriction enzyme MboI, followed by biotin labeling and blunt-end ligation. Crosslinks were reversed through proteinase K treatment, and DNA was purified using phenol-chloroform extraction. Unligated biotin was removed using T4 DNA polymerase. DNA fragments were then sheared by sonication, repaired, and enriched using streptavidin-coated magnetic beads to isolate biotin-labeled fragments. Finally, A-tails were added to the DNA ends, and Illumina paired-end sequencing adapters were ligated. The Hi-C library was amplified by PCR and sequenced using the Illumina NovaSeq 6000 platform with 150 bp paired-end reads.

Raw data was filtered to remove adapter sequences, reads with no base information, and low-quality reads, resulting in clean reads. Clean reads were further processed to remove duplicates. The quality control criteria included Q30 > 80%, an error rate of less than 1%, and a GC content distribution without significant separation for individual reads.

Belton, J.-M., McCord, R. P., Gibcus, J. H., Naumova, N., Zhan, Y. & Dekker, J. (2012) Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods*, 58, 268-276.

- For RNAseq, no indication of DNA extraction, RT, QC, library prep, or sequencing platforms are given. Sequencing statistics were not presented or discussed either.

Authors' response:

The library were prepared by using NEBNext Ultra™ RNA Library Prep Kit for Illumina (NEB, USA). Briefly, mRNA was purified using poly-T oligo-attached magnetic beads and fragmented using divalent cations under elevated temperature. first strand cDNA was synthesized with random hexamer primers and M-MuLV Reverse Transcriptase, followed by second strand cDNA synthesis using DNA Polymerase I and RNase H. Overhangs were converted to blunt ends, and 3' adenylation was performed before ligating NEBNext Adaptor with hairpin loop structure. Size selection of ~250–300 bp cDNA fragments was performed using AMPure XP beads, and USER Enzyme was applied to process adaptor-ligated fragments. PCR amplification was carried out with Phusion High-Fidelity DNA Polymerase and indexed primers. The library was purified and assessed on an Agilent Bioanalyzer 2100 system. Finally, clustering was performed on an Illumina cBot system using the TruSeq PE Cluster Kit v3, and sequencing was conducted on the Illumina NovaSeq 6000 platform to generate 150 bp paired-end reads.

Adapter sequences, reads with uncertain bases (>10%), and low-quality reads were removed from raw data, resulting in clean reads with Q30 > 90% and an error rate of less than 1%. All quality control steps were performed using Fastp v0.19.7. The details of statistical description of QC were added into Supplementary Table 20 and 21.

l.482 - "QV of 0.99" - this does not seem right, QV is >> 1 for CCS

Authors' response:

Thank you for your comment.

Yes, we mistakenly referred to the `rq:f` cutoff in raw HiFi CCS data as a QV score. In fact, `rq = 0.99` corresponds to approximately QV 20. We have corrected this error in the main text.

l.492 – hifiasm v334 – possibly r334 (v0.15.1)?

Authors' response:

Thank you for your comment.

Yes, we corrected in main text.

l.495 – refseq version does not seem right

Authors' response:

Thank you for your comment.

Yes, we revised it as “Refseq (release205)”

Paragraph starting at l.495 is hard to interpret in terms of implementation – was the contamination removal run three times (reads, assemblies, read overlaps)? If so, how do contigs with >40% ref overlap appear if all reads with >40% ref overlap were removed? If all filtering was done at once, why does the last sentence say only clean reads were used for assemblies, while filtering mentioned contigs?

Authors' response:

Thank you for your comment.

Removing contamination reads from the raw data is challenging for several reasons:

- (1) The source of contamination is unknown. If we rely on nt database comparisons, the time cost is unacceptable. It is also difficult to determine whether a read originates from a bacterial genome or from homologous sequences in non-insect organisms
- (2) If we restrict the BLAST reference to a specific genome, some regions of bacterial genomes may be overlooked due to the complexity and diversity of bacterial genomic content.
- (3) We are uncertain whether contamination sequences might affect the assembly results from different software. For most other species, the sequencing coverage is sufficient to rule out interference from some contamination sequences

To address these issues, we first align reads to bacterial RefSeq genomes and record the corresponding aligned regions in HiFi CCS reads. Meanwhile, we perform pre-assembly of all raw data using both hifiasm and hicanu to assemble contigs of bacterial and fungal origin as completely as possible. By analyzing GFA (Graphical Fragment Assembly) information, we identify bacterial homologous sequences within the assembled fragments (unitigs, contigs, and super-contigs). If a significant portion of a contig shows homology to a bacterial genome, we consider the entire contig to be contamination.

In our approach, a cutoff value of 0.4 is applied, validated by comparing the entire fragment with the NT database. If a contig is identified as contamination, all reads comprising that contig are excluded from hifiasm assembly. Additionally, any reads that overlap more than 70% with these identified contamination reads are also removed from subsequent assemblies.

We have revised the description above to ensure clarity and avoid potential misunderstandings.

l.507 - please insert paragraph break after genomescope section, it is separate from long read assembly

Authors' response:

Thank you for your comment. We have revised the content and isolated this paragraph.

Starting from l.509, this seems like ONT read mapping and read ends uniformity exploration corresponding to l105-109 in results – both of those were not very clear to me, so please consider rephrasing. Also, what was used to map ONT reads?

Authors' response:

Thanks for your comment.

As indicated in our previous response, we have revised the main text and restructured the Methods section according to your suggestions, ensuring improved clarity and logical organization. Additionally, we have supplemented the method description with details about ONT read mapping using minimap2.

Paragraph starting at l.517 is a mixture of: pacbio read mapping, coverage based sex chromosome identification, ont alignment, purging, scaffolding, assembly curation. None of those are described completely and the logic of including all of these procedures into one paragraph was not clear. Some of the data are seemingly reused in later paragraphs in the section dealing with gap closing and polishing. Overall, the whole process descriptions leaves many open questions, many of which would be easy to answer with code listing available:

Authors' response:

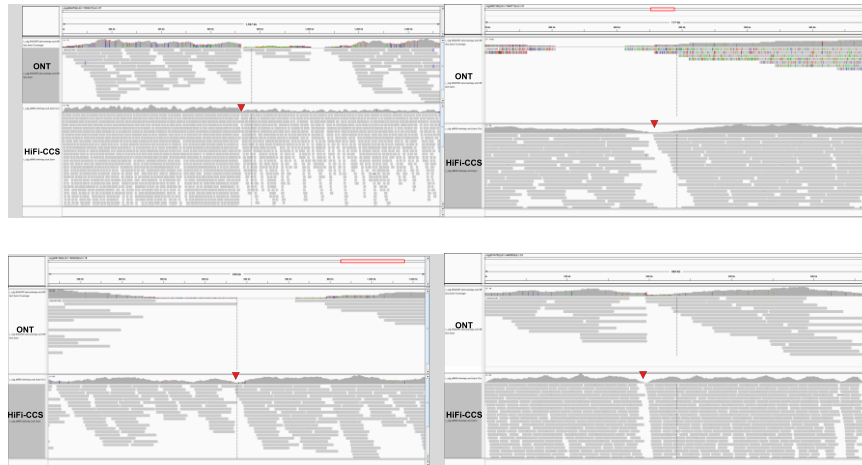
Thanks for your comment.

Our current computational approach places significant constraints on the pipeline, involving numerous manual adjustments and iterative optimizations. Additionally, most custom scripts utilized in this process were developed in Perl. We are working on creating a semi-automated workflow based on our assembly experience, aiming to streamline and automate the existing process for assembling genomes of similarly low-coverage and highly heterozygous small insects. However, due to time limitations, we have not yet systematically organized and documented the extensive number of scripts.

- what issues are addressed by read end uniformity testing and subsequent breaks?

Authors' response:

Here, we provide several examples illustrating potential errors in certain regions of the hifiasm unitigs, which can be identified based on ONT read distribution patterns and supporting evidence.



- what is the benefit of generating alignments with three tools and were all of these tools used to both pacbio and ont reads?

Authors' response:

We utilized minimap2, Winnowmap2, and GraphAligner for different genomic regions. Since hifiasm internally employs minimap2, it offers superior performance and speed advantages for most autosomal regions, making it our preferred tool for processing HiFi CCS data. Winnowmap2, however, demonstrates greater accuracy in assembling highly repetitive regions, and thus we applied it for ONT reads and HiFi CCS data alignment in repetitive areas. For regions characterized by complex graphical relationships, relying solely on Winnowmap2 is insufficient to accurately determine haplotype relationships among multiple contigs. Therefore, we additionally employed GraphAligner results to support and clarify contig path relationships.

The usage and application of the sequence alignment tools are similar to those in the Verkko software package.

- why sex chromosome linkage is estimated early in the process? Does this affect any of the downstream processes?

Authors' response:

Given the high repeat content and small fragment size of the Y chromosome, combined with limited Hi-C support, there was concern about potential compression or misassembly of sex chromosome fragments or sex-autosome fragment misassemblies during autosomal assembly. Therefore, we aimed to accurately assemble sex chromosome regions by integrating scaffold relationship, ONT-supported haplotype relationships, and alignment results from non-repetitive regions in candidate contigs.

- how do you use information on low-complexity regions definition in gap closing with ont? The process as it is described seems to be agnostic towards the underlying graph structure

Authors' response:

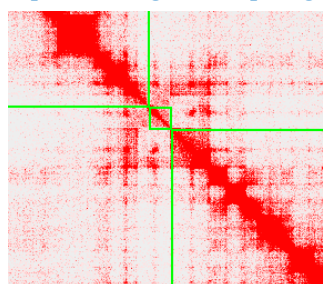
We apologize for the confusion caused by the previous version of the Methods description. We have now refined the entire genome assembly section. As previously described, for low-complexity regions, we typically use GraphAligner to simplify contig graph paths. For terminal

regions in low-complexity regions, we consider the alignment results from Winnowmap2 to select the ONT long reads with supporting the connection of terminal adjacent contigs and with the minimal mismatch identity.

- *What exactly happens during alternate contig removal? Hi-C data is mapped to linear genomes, so reference to graphs is not relevant.*

Authors' response:

The assembly of these regions primarily relied on hifiasm's p_utg assembly relationships, preserving the longest haplotype relationships in the final diploid assembly as super-contigs. Although hifiasm addresses haplotypic relationships in regions with high coverage, our sample's highly heterozygous genetic background resulted in coverage significantly lower than 30x for each haplotype path in heterozygous regions. Thus, ONT data were essential for validating phase switches in super-contigs and bridging assembly gaps. Furthermore, Hi-C density plots efficiently identified unique redundant contigs within high heterozygosity or repetitive regions in p_utg assembly bubbles (as illustrated in the accompanying figure)



- *For ONT you mention the problem of structural variation coming from pooled DNA sample and the workaround of manual investigation of called SVs. What was the decision making process? What would you do if SV overlapped region not resolved by pacbio data?*

Authors' response:

Please refer to our previous response for 1.130 and modifications in the revision.

l.549 – why the centromere annotation cannot be done on final T2T linear version of the assembly?

Authors' response:

Thank you for your comment.

The Hi-C–based genome assembly, generated using HiFi-CCS reads, exhibited fragmentation at contig ends and centromeric regions, which is a limitation of assembler when handling highly repetitive sequences in diploid genome. As no centromeric repeat units had been previously reported for this species, we hypothesized that they might be located near the ends of the assembled super-contigs. To test this, we leveraged accurately extended ONT long reads to search for candidate regions and experimentally validated their locations, ultimately identifying the centromere-associated sequences.

l.592 - references to NCBI datasets for proteins are not given.

Authors' response:

Thank you for your comment.

We used the protein sequences of *Bactrocera cucurbitae* (GCF_000806345.2), *Bactrocera oleae* (GCF_001188975.3), *Bactrocera tryoni* (GCF_016617805.1), and *Bactrocera latifrons* (GCF_001853355.1), which were obtained from NCBI RefSeq. This information will be included in the manuscript.

l.595 – what is the logic for 5kb intron length cutoff? This does not seem long enough to me

Authors' response:

Thank you for your comment.

Sorry for the confusion. We applied a 5 kb intron length cutoff during ab initio gene annotation using Augustus to minimize potential errors caused by the incorrect merging of adjacent gene models.

We have revised this section.

l.597 - "manually integrated" - probably not completely manually? Also, what were the rules for integration of genes missing in one or two methods?

Authors' response:

Thank you for your comment.

We add the additional information into our manuscript.

“The CDS annotation files generated with the three approaches were manually integrated. First, extract the CDS information from the annotation files of the three methods separately and merge them using GffRead107 (v0.12.7). If different prediction methods generate conflicting CDS structures at the same gene locus, use the following priority to determine the final structure: transcript-based evidence > homology-based prediction > ab initio annotation. For CDS structures obtained from ab initio annotation, if multiple short CDS structures are integrated into the same gene locus, this indicates that the gene may be a fragmented gene. For Augustus genes containing introns longer than 5,000 bp, prioritize retaining transcript structures supported by ISO-Seq data. If no ISO-Seq data is available, retain only structures with a complete Open Reading Frame (ORF).”

l.626 - what was the query for the NUMT search? If this mt assembly was generated as part of your assembly, what were the processes for its finishing and QC to ensure that mitochondrial contig is correctly represented?

Authors' response:

The complete mitochondrial genome sequence was used to identify NUMTs. Among the hifiasm unitigs, utg003105c contains a full-length mitochondrial genome with a coverage of 433× and a length of 15,917 bp. This sequence shows highly identity to the *B.dorsalis* mitochondrial reference sequence in RefSeq.

l.631 - table s-mt - incorrect reference

Authors' response:

Sorry for the typo and thank you for your correction.

Supplementary Table 10 was added to replace the incorrect type.

l.642 - methods for differential expression section seemingly missing (or mis-placed). Please also check if any additional methods are needed regarding OR genes repertoire investigation (anything for Fig6a-d)?

Authors' response:

The detail method for differential expression genes analysis and OR annotation are revised as follows:

DEGs for RNA-seq data

The RNA-Seq reads obtained from *B. dorsalis* peripheral nervous tissue were quality-filtered as description in “Raw data preprocessing”. The filtered reads were aligned to the *B. dorsalis* genome using HISAT2¹² (v2.2.1) with default parameters. Based on the alignment results, featureCounts (v2.0.1) was employed to quantify annotated genes in the genome, compiling a raw gene expression matrix and a TPM (Transcripts Per Million) matrix for the peripheral nervous tissue. The DESeq2¹³ was applied to analyze sex-specific differentially expressed genes (DEGs) in the peripheral nervous tissue, with thresholds set at fold change > 1.5 and adjusted p-value (p.adjust) < 0.05. DEGs between sexes across different tissues were subsequently extracted for further analysis. Heatmaps were visualized using the pheatmap package in R.

Refine the structure of OR genes

Protein sequences of olfactory receptor (OR) genes from dipteran insects published in NCBI were collected, and incomplete sequences were filtered out. Using the hmmbuild function in HMMER, a Hidden Markov Model (HMM) was constructed based on these protein sequences. This model, along with the 7tm_6 odorant receptor HMM profile (PF02949) from the Pfam database, was applied via the hmmsearch function to screen all *B. dorsalis* protein sequences, retaining hits with e-value < 1e-3 as candidate OR genes supported by domain evidence. Separately, OR gene protein sequences of *D.melanogaster* from the Swiss-Prot database were collected and aligned against all *B. dorsalis* protein sequences using PSI-BLAST. Hits with e-value < 1e-3 were selected as candidate OR genes supported by homology evidence. The intersection of the two candidate sets was taken to obtain the final candidate OR gene list. The protein sequences of the final candidate genes were analyzed with InterProScan (v5.55-88.0) against the Pfam, TMHMM, Phobius, and PANTHER databases to validate domains and perform functional predictions. Genes predicted as ORs or containing 7-transmembrane domains (7tm_6) were retained as the final candidate OR genes.

Publicly available OR and gustatory receptor (GR) protein sequences of *D.melanogaster*, *B.correcta*, *B.cucurbitae*, *B.tau*, *B.minax*, as well as OR protein sequences of *E.balteatus* and *E.corollae*, were searched and downloaded online. These sequences, along with the amino acid sequences of OR genes annotated in this study from *B. dorsalis*, were subjected to multiple sequence alignment using MAFFT (v7.515). A phylogenetic tree was then constructed by the maximum likelihood method with IQ-TREE (v2.2.0.3) using parameters: -bb 1000 -nt 20 -m MFP -bnni.

OR-related expression data were extracted from the peripheral nervous tissue expression profiles and differential expression analysis results mentioned in the previous sections.

l.667 - consider adding more details about target sequences (or ref to supplementary)

Authors' response:

We have added the following target sequence information of DNA probes into the Supplementary tables 22.

l.708-711 - insufficient level of detail, but assume it follows ref 4 (mentioned in l.679)

Authors' response:

We have revised the reference index for our previous work.

Yuan et al. 2022 Protocols for CRISPR/Cas9 Mutagenesis of the Oriental Fruit Fly *Bactrocera dorsalis*. Journal of visualized Experiment. (187). e64195. <https://doi.org/10.3791/64195> (2022).

Overall, main figures were way too small when printed - assume web version will still work

Authors' response:

Thank you for your comment. We have resized font size in figures and optimize the layout of the diagram.

l.776 (fig2) - what are "hifiasm units"?

Authors' response:

This should be "Hifiasm unitigs" and we revised this in the legend for fig.2.

Fig 6c - legend on tissues for expression heatmap is needed

Authors' response:

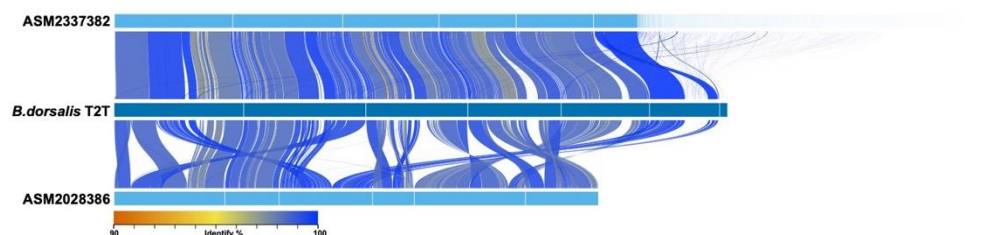
Revised as below:

A: Antennae, M: Mouthpart, F-L: Fore leg, M-L: Mid leg, H-L: Hind leg; G-L: Genitalia.

Supp Fig 2 - are all three assemblies from the same species? If so, something should be said about the extent of large-scale structural variation between the assemblies. If those are different species, these should be indicated together with some notion of relative divergence

Authors' response:

The two extra assemblies are both created with third generation sequences data and ASM2337382 currently was the RefSeq reference for *B. dorsalis*. We modified the direction of contig to show the consistence. For ASM2028386, their original Hi-C plot is too weak to distinct the block of chromosome.



Supp Fig 4 eluded my understanding - what do the peaks in sf4a represent? Three rRNAs and their lengths - or maybe something else? What does the subplot b demonstrate? Do you see a canonical structure and some deviations from that?

Authors' response:

- (1) We have updated the figure legend of Fig. S4 to explain the peak. In ONT reads, LSU rDNA homologous regions displayed three distinct length peaks at 1,247 bp, 2,697 bp, and 3,939 bp (the complete LSU rDNA repeat unit), suggesting some LSU rDNA units were fragmented into two parts rather than existing as complete units.
- (2) Subplot b illustrates that certain sequences in our assembly were involved in the rDNA insertion process. Additionally, specific repeat sequences were found downstream of the entire rDNA large subunit and following the rDNA repeat unit. Since there is limited research on this topic in the genome of the oriental fruit fly or other Dipteran species (excluding *Drosophila*), the function of these sequences remains unclear. However, it is evident that rDNA large subunit insertions are not uncommon in Diptera and have even been observed in fungi¹⁴.

l.859 (SF5) "multiplex species" - maybe just gene copies?

Authors' response:

It means the consensus hexapeptide KXNSRT in N-terminal region of MoY homologs in Tephritidae family⁶. we have revised the content in the legend part.

SF6 - in printed version, I don't see any signal in rows 2 and 3. What are horizontal and vertical lines? If those delineate chromosomes, why horizontals are different if you have the same genome on y axis?

Authors' response:

The plots illustrate genome-wide synteny between *B. dorsalis* (y-axis) and other species. Conserved Muller element structures are observed only in closely related species (row 1), where synteny with the *B. dorsalis* genome is largely preserved. In more distantly related species (rows 2 and 3), extensive genomic inversions have disrupted gene-based synteny patterns. However, upon closer inspection, homologous genes still tend to cluster within corresponding chromosomal structures, reflecting a degree of structural conservation.

We have updated the figure legend to reflect this description.

SF7 extends Fig6 e-f - maybe include a reference?

Authors' response:

We have revised content related with Supplementary Fig.8.

Reviewer #2 (Remarks to the Author):

*This manuscript presents a telomere-to-telomere (T2T) assembly of the *Bactrocera dorsalis* genome using long-read sequencing. Based on this assembly, the authors provide comprehensive analyses of centromeres, telomeres, nuclear-mitochondrial DNA segments (NUMTs), sex chromosomes, and tandemly duplicated gene families, particularly olfactory receptors (ORs). Through CRISPR knockout (KO) experiments, they demonstrate that one OR locus (*BdorOR88a12*) is involved in detecting methyl eugenol (ME), a compound used for pest control. This study represents a substantial contribution to both agricultural pest management and the understanding of species diversity. However, several aspects of the manuscript require improvement.*

Authors' response:

Thank you for your time and insightful comments on our manuscript. We have carefully revised the manuscript according to the suggestions. We sincerely thank you for your valuable time and constructive comments. We have thoroughly revised the manuscript based on your suggestions

Major Concerns

1. Figure Layout and Legends

o The font size in figures is too small and should be increased for better readability.

Authors' response:

Thank you for your comment. We have resized font size in figures and optimize the layout of the diagram.

o Several figure legends are incomplete or unclear. For example:

Clarify the meaning of green and blue clades in Figure 4e and define GR/OR in Figure 6a.

Correct missing or inaccurate legends in Figures 4g and 5b.

Authors' response:

As per your suggestion, we have reexamined and revised the legend of all figures, while also providing explanations in some legend part.

o Figure order errors (e.g., Figure 1b appears before Figure 1a in Line 93) should be addressed.

Authors' response:

Thank you for your comment. We have revised the Figure order issue and reexamined manuscripts.

o Specific parts of Figures 1a and 1b lack clarity. For instance:

In Step 2 of Figure 1a, the disconnected core element appears identical across situations.

The long strings in Steps 2 and 3 seem to have inconsistent meanings. These diagrams and their legends need clearer explanations.

Authors' response:

Thank you for your comment.

To minimize redundancy with the methods section, we removed the two paragraphs from main text and refined a detailed description of method choices and parameter statistics is included in the methods part.

2. T2T Assembly Quality

o Include details on sequencing read depth and coverage for methods such as HiFi, ONT, NGS, and Hi-C to enhance reproducibility.

Authors' response:

Thank you for your comment.

(1) For illumina sequencing:

We have added the statistics for Illumina sequencing to the main text as follows:

“The genetic background was assessed with 23.06 Gb Illumina sequencing data of pooled DNA from five individuals using GenomeScope. The results revealed a mapping rate of 98.36% and an average heterozygosity of 1.81%.”

Additionally, we have included a new supplementary table (Supplementary Table 1-1) with detailed information.

METRICS	READ1		READ2	
	raw data	after filtering	raw data	after filtering
TOTAL READS	77,113,009	77,112,955	77,113,009	77,112,955
TOTAL BASES	11,531,065,295	11,531,051,099	11,529,467,949	11,529,451,867
Q20 BASES (%)	11,256,480,024 (97.62%)	11,256,469,684 (97.62%)	11,190,612,026 (97.06%)	11,190,600,513 (97.06%)
Q30 BASES (%)	10,735,850,583 (93.10%)	10,735,841,554 (93.10%)	10,548,788,094 (91.49%)	10,548,779,317 (91.49%)

(2) for ONT and HiFi CCS reads, the corresponding statistics have been provided in Supplementary Table 1-2.

	HiFi-CCS female	for Filtered for female	HiFi-CCS male	HiFi-CCS for Filtered for male	HiFi-CCS ONT reads
Mean read length	11,500	11,550	12,335	12,287	37,792
Mean read quality	30	30	28	28	13
Median read length	11,089	11,139	11,906	11,858	30,941
Median read quality	38	38	33	33	14
Number of reads	2,593,480	2,222,421	1,585,639	1,551,322	2,508,117
Read length N50	11,416	11,470	12,194	12,150	59,916
STDEV read length	1,943	1,948	1,814	1,817	32,783
Total bases	29,824,224,634	25,667,967,376	19,558,761,563	19,060,311,110	94,787,558,719

(3) for Hi-C data, the corresponding statistics have been provided in Supplementary Table 1-3

Sequenced Read Pairs	111,350,858
Normal Paired	42,480,310 (38.15%)
Chimeric Paired	44,954,976 (40.37%)
Chimeric Ambiguous	19,753,031 (17.74%)
Unmapped	4,162,541 (3.74%)
Ligation Motif Present	62,185,923 (55.85%)

Alignable (Normal+Chimeric Paired)	87,435,286 (78.52%)
Unique Reads	82,267,858 (73.88%)
PCR Duplicates	4,952,138 (4.45%)
Optical Duplicates	215,290 (0.19%)
Library Complexity Estimate	738,728,077
Intra-fragment Reads	3,703,603 (3.33% / 4.50%)
Below MAPQ Threshold	29,016,981 (26.06% / 35.27%)
Hi-C Contacts	49,547,274 (44.50% / 60.23%)
Inter-chromosomal	6,191,962 (5.56% / 7.53%)
Intra-chromosomal	43,355,312 (38.94% / 52.70%)
Short Range (<20Kb)	17,825,941 (16.01% / 21.67%)
Long Range (>20Kb)	25,529,148 (22.93% / 31.03%)

o Consider adding GCI as a metric to assess assembly quality (e.g., <https://academic.oup.com/bioinformatics/article/40/11/btae633/7829147>).

Authors' response:

We have calculated GCI score for our assembly and added this content in main text and method section.

“Assembly continuity achieves a GCI score of 31.5 for whole genome, and 51.4 and 57.0 for X and Y chromosomes separately. “

o Compare this assembly with previous versions by specifying the number of novel genes or exons newly annotated.

Authors' response:

Thank you for your comment.

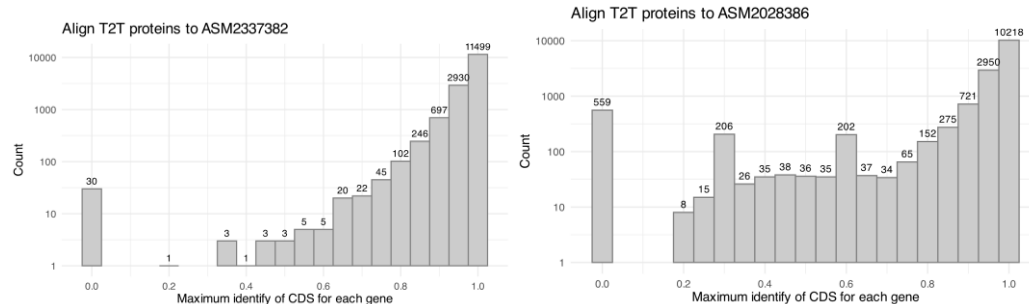
A comparative analysis between the telomere-to-telomere (T2T) assembly and the RefSeq reference genome (ASM2337382) revealed 199 uniquely identified genes in the T2T assembly and 8 exclusively annotated genes in the RefSeq genome. To address potential biases arising from heterogeneous sample origins and divergent annotation pipelines across assemblies, we performed systematic proteome-level alignments and retaining only the isoform with the highest sequence identity for each gene. This comparison can identify members of gene family that were absolutely omitted in assembly. The full catalog of assembly-specific genes is provided in Supplementary Table 5.

In the result, at least 30 genes completely missed in RefSeq reference genome (ASM2337382), including: Three Y-linked paralogs of the *Moy* genes and nine genes with identified KEGG annotation.

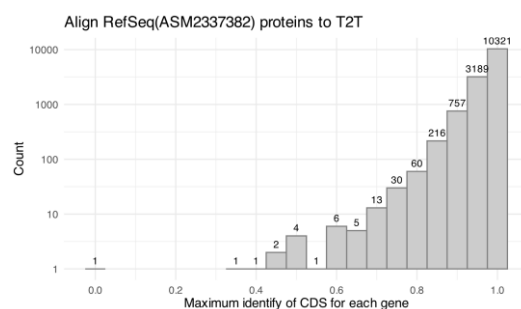
Gene KEGG

Bdo2G002439	cyclin-dependent kinase-like [EC:2.7.11.22]
Bdo2G002522	DNA replication licensing factor MCM6 [EC:5.6.2.3]
Bdo3G000475	prolyl 3-hydroxylase /prolyl 3,4-dihydroxylase [EC:1.14.11.-]
Bdo3G000568	acid-sensing ion channel, other
Bdo3G000894	TELO2-interacting protein 1

Bdo3G000895 bifunctional polynucleotide phosphatase/kinase [EC:3.1.3.32 2.7.1.78]
 Bdo3G000898 protein SMG6 [EC:3.1.-.-]
 Bdo3G000899 WD repeat-containing protein 91
 Bdo3G000900 nuclear receptor subfamily 1 group I



For the gene families from Refseq genome, one gene (gene-LOC109579598) was not found in out T2T assembly. This gene was ab initio annotated with Gnomon and function as uncharacterized. This gene can be found in the original HiFiasm untigs but unselected in final assembly, which is a bubble contig with half of the average coverage (utg000937l; LN:i:769505|rd:i:16).



3. BdorOR88a12 and KO Experiments (Figure 6)

o Explore whether different transcript isoforms can be distinguished and quantified. Consider qPCR with transcript-specific primers to determine if the gene's function relies on a single or chimeric transcript.

Authors' response:

With current technology, we are able to distinguish between different gene copies in omics data and perform quantitative analysis. Transcriptome data show that the expression levels of copies 1 and 2 in the wild type, as measured by TPM, account for about 80-90% of the total expression of the three copies. This suggests that the primary functions are likely mediated by copies 1 and 2. However, due to the high sequence similarity (90%) of the CDS regions among the three copies, current qPCR methods cannot precisely distinguish or quantify individual copies.

[illegible]

o Clarify which isoform was knocked out and elaborate on the functional consequences of the KO experiments.

Authors' response:

Genotype analysis revealed that both highly expressed copies, BdorOR88a1 and BdorOR88a2, were successfully knocked out in the mutant, while BdorOR88a3 remained intact. However, if the knockout cause expression level of BdorOR88a3 changed, eg. dramatically increased following the knockout of the two highly expressed copies, it could potentially interfere with the functional experiments. To address this concern, we performed additional RNA-seq of the antennae in the mutant lacking BdorOR88a1 and BdorOR88a2. The results showed that the expression pattern of the three copies remained unchanged after the knockout, with BdorOR88a1 and BdorOR88a2 showing high expression levels and BdorOR88a3 remaining lowly expressed. Moreover, all detected chimeric transcripts carried premature termination mutations. Therefore, we concluded that the functionally significant contributions of BdorOR88a1 and BdorOR88a2, as well as their chimeric transcripts, were successfully eliminated in the mutant. We added this information to FigS6 and revised corresponding part of manuscript (including RNA-seq quality control results, normalized expression profiles of the three gene copies, and relevant chimeric transcripts.).

o Verify sgRNA design for repetitive loci like ORs, ensuring mappability and minimizing off-target effects.

Authors' response:

To investigate the biological functions of the predominantly expressed genes BdorOR88a-v1 and BdorOR88a-v2, we aimed to generate double-knockout mutants for these two gene copies. However, the high sequence similarity among the BdorOR88a1-3 gene copies poses a significant challenge in designing specific sgRNAs that exclusively target individual copies. An alternative strategy is to design sgRNAs targeting the shared regions of these genes.

However, this approach is also limited, as a single sgRNA targeting the shared region often fails to efficiently induce simultaneous mutations at two distinct genomic locations, making it challenging to achieve successful double mutants.

To overcome this limitation, we optimized our approach by designing multiple sgRNAs that target the shared regions and co-injecting these sgRNAs. This multi-sgRNA strategy is intended to create an additive effect, increasing the likelihood of inducing successful double mutations. Specifically: sgRNAs 1, 3, and 4 were designed to target the shared regions of all three copies (BdorOR88a1-3), thus broadening the mutation coverage. sgRNA 2 was specifically designed to target the shared region of BdorOR88a-v1 and BdorOR88a-v2, further enhancing editing efficiency and specificity at these two locations.

Analysis of these target sites showed that all four sgRNAs align accurately with the desired genomic regions (identity = 100%, e-value < 1×10^{-4}). Potential off-target effects were minimal: most non-target regions lacked NGG PAM sites, and the few regions containing PAM sites had low guide sequence matching scores (~65%). This information was added into the Supplementary Table 23.

Minor Concerns

1. The manuscript extensively discusses telomeres on sex chromosomes but provides limited information on autosomal telomeres. Elaborate on the significance of the 528 bp feature and the presence or absence of TTAGGG repeats.

Authors' response:

Thank you for your comment.

We have added an extra paragraph concerned about new telomere in the Discussion section.

“Notably, Dipteran species have evolved telomerase-independent mechanisms for chromosomal end maintenance, with two distinct telomeric architectures identified to date³. The telomeric structure of *D.melanogaster* can extend through several kilobases and consists of a mixed repetitive array of HetA/TART/TAHRE elements, along with telomere-associated sequences^{4,5}. The other type, found in lower Diptera, consists of long tandemly repeated sequences³. The telomeric element sequence of *B.dorsalis* differs substantially from other seven species of the suborder Brachycera, all of which have the long repetitive element in the end of chromosome.”

2. Clarify why only 459 single-copy orthologous genes were used, given the more extensive annotation in related species like *D. melanogaster*.

Authors' response:

Thank you for your comment.

When we performed gene family analysis, we used the processed data originally prepared for synteny analysis. This led to the exclusion of genes located on scaffolds shorter than 100 kb. In the assemblies of *Teleopsis dalmanni* (GCA_002237135.2, 3 chromosomes + 5644 unplaced scaffolds) and *Bradysia coprophila* (GCF_014529535.1, 4 chromosomes + 320 unplaced scaffolds), a large number of genes were not assigned to scaffolds longer than 100 kb. As a result, the number of single-copy orthologous genes obtained was relatively low.

We re-analysis the data with whole gene datasets, and finally yielding 2,054 single-copy orthologous genes. The topology of the phylogenetic tree remained unchanged. We have corrected the results and revised the main text and corresponding figures.

3. *When analyzing centromeres, the authors relied on TRF (v4.09.1) and two candidate sequences for FISH. This approach is insufficient for centromere validation, requiring ChIP-seq or CUT&Tag. Terms like “candidate centromere” should be used to temper claims.*

Authors’ response:

Thank you for your comment.

We agree with you that insufficiency for centromere validation and revised the claim with “candidate centromere” in main text.

4. *Address the potential conflict between the two software tools mentioned in the “Raw data preprocessing” section (Lines 481–502).*

Authors’ response:

Thank you for your comment.

We used Hicanu and Hifiasm to pre-assemble all raw reads, aiming to cluster potential contamination sequences from microbes or fungi. By working with reads rather than contigs, we streamlined the process of removing the corresponding sequences when necessary. The above description has been refined to improve clarity and minimize potential confusion.

5. *Improve NUMT mutation analysis:*

o Explain why only three NUMTs are shown in Figure 3a, despite identifying 11 NUMTs in Figure 3b. Are these the largest?

Authors’ response:

Thank you for your comment.

We randomly selected representative cases from the three phylogenetically distinct NUMT clusters to demonstrate the progressive erosion of homologous syntenic architecture over evolutionary timescales. The figure legend has been updated accordingly, and plots for each chromosome have been included in the Supplementary fig.10.

o The genomic average GC content is 40%, while the average GC content of MT DNA is 28%. Consistent with the authors’ hypothesis that NUMT GC content evolved from 28% toward 40%, older NUMT copies should exhibit a stronger AT -> GC mutation bias compared to the MT reference genome. This appears to be the case as in the second column of Figure 3b (as shown in the bottom branch including MT13 or MT5). I suggest that authors include this point in the main text.

Authors’ response:

We sincerely appreciate your suggestion, and we have incorporated a description of this evolutionary characteristic of NUMTs into the main text.

“The expression levels of mitochondrial gene regions within the 11 NUMTs are extremely low. When we examined the sequence composition, we observed that with increasing time since insertion, the inserted mitochondrial sequences presented a reduced transition-to-transversion (Ti/Tv) ratio, along with increasing GC content, gradually increasing from the average level of

28% in the mitochondria to 40% in the nuclear genome (Fig 3b). Older NUMT copies are expected to exhibit a stronger AT-to-GC mutation bias compared to the mitochondrial reference genome, as highlighted by the base composition differences observed between chr2MT5 and chr5MT1.”

6. *Correct minor errors:*

o Line 636: Fix the GitHub link to “<https://github.com/soybase/crossmap-workflow>.”

Authors’ response:

Thank you for your correction.

Corrected!

o Line 679: Resolve the citation format error.

Authors’ response:

Thank you for your correction.

Corrected!

o Clarify which Hifiasm version (v334 or 0.18.5) was used.

Authors’ response:

Thank you for your correction.

Corrected!

Reviewer #3 (Remarks to the Author):

*In this paper, Liu et al. generate a T2T genome assembly of the pest species *Bactrocera dorsalis*. This insect is of economic importance and the author's approach illustrates the process to be followed to characterize the genome of similar pests for which amenable approaches such as those typically used for *Drosophila* species are of limited use due to e.g. high levels of polymorphism, difficulty to generate inbred lines, etc... First, combining HiFi and ON sequencing, the authors generate a genome assembly, which is highly contiguous, with complex repetitive regions resolved, and with telomeres, centromeres, and the Y chromosome represented. Key foci in this manuscript whose organization is uncovered include centromeric and telomeric regions as well NUMT insertion loci. Further, the new assembly includes the sex chromosomes, and the authors addressed several aspects including dosage compensation, sex-bias differences in expression, and the characterization of a Y-linked gene that are relevant for sex-determination during early development including their paralogs in other chromosomes. At the comparative genomic level, the authors examine the level of maintenance of the ancestral chromosomal synteny at different phylogenetic distances, as well as the relationship of the X chromosome of *B. dorsalis* with those in other species. In the last section of the results, the authors explore different aspects of the OR gene families, including chromosome and cluster distribution, expression and formation of chimeric transcripts for one of the clusters, and phenotypic effects resulting from knocking out two adjacent OR paralogs.*

Authors' response:

We sincerely thank you for your valuable time and constructive comments. We have thoroughly revised the manuscript based on your suggestions.

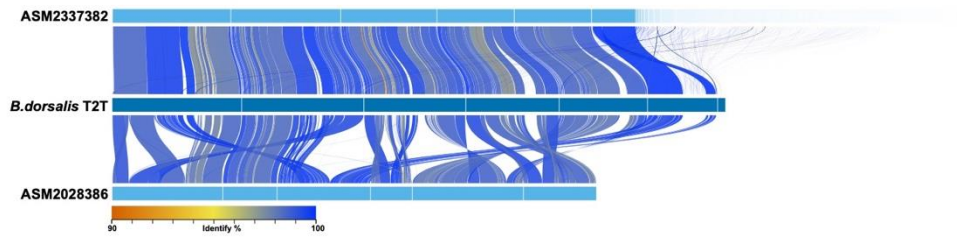
Issues and concerns:

1. Improvements in relation to previously published genome assemblies and gene annotations in *B. dorsalis*. Although the authors devote several sections to highlight key regions (centromere, telomere, NUMT loci) that are now well represented in the T2T genome assembly, it is not clear whether the euchromatic fraction has significantly improved. For example, the authors provide a comparison of the T2T assembly with other published assemblies (Sup. Fig. 2, line 142). Are there differences in the orientation of blocks of adjacent genes? What about the number of copies of some multigene families arranged in tandem? And relative to repetitive sequences?

Authors' response:

Thank you for your comment.

(1) The two extra assemblies are both created with third generation sequences data and ASM2337382 currently was the RefSeq reference for females *B. dorsalis*. The most significant large-scale rearrangements were due to chromosome directions of these genomes were not consistent with each other, thus showing as unnecessary inversions between the assemblies. We have adjusted the orientation of most chromosomes/scaffolds so that they consistently point in the same direction. For ASM2028386, their original Hi-C plot is too weak to distinct the block of chromosome.



(2) In the following response, we compared differences between identified genes. Our assembly shows a highly consistent structural order and provides more complete sequences.

2. Something similar to point 1 applies to the gene annotation. There are no clear indications about how the new gene annotation differs from previous ones in NCBI for this same species. Put simply (1 & 2), the putative improvement in the resources generated by the authors relative to previous ones are just superficially presented.

Authors' response:

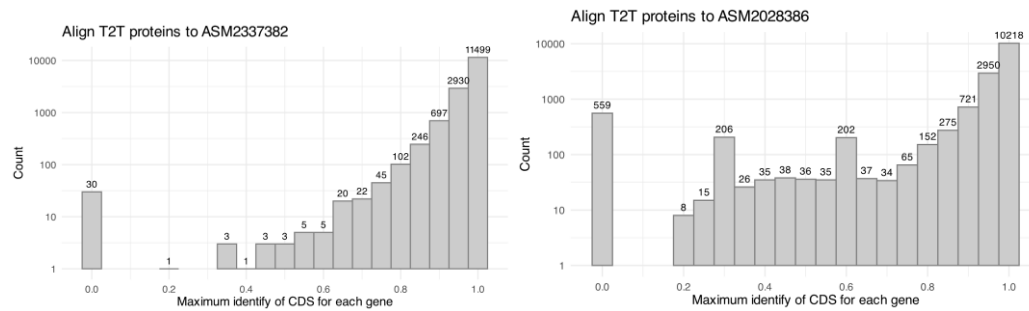
Thank you for your comment.

A comparative analysis between the telomere-to-telomere (T2T) assembly and the RefSeq reference genome (ASM2337382) revealed 199 uniquely identified genes in the T2T assembly and 8 exclusively annotated genes in the RefSeq genome.

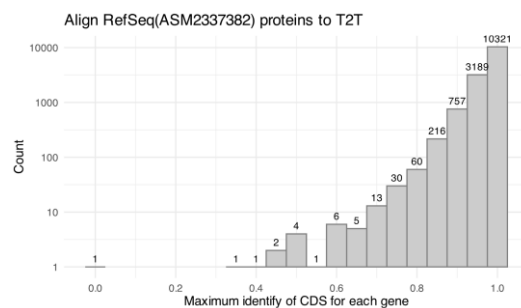
To address potential biases arising from heterogeneous sample origins and divergent annotation pipelines across assemblies, we performed systematic proteome-level alignments and retaining only the isoform with the highest sequence identity for each gene. This comparison can identify members of gene family that were absolutely omitted in assembly. The full catalog of assembly-specific genes is provided in Supplementary Table 5.

In the result, at least 30 genes completely missed in RefSeq reference genome (ASM2337382), including: Three Y-linked paralogs of the *Moy* genes and nine genes with identified KEGG annotation.

Gene	KEGG
Bdo2G002439	cyclin-dependent kinase-like [EC:2.7.11.22]
Bdo2G002522	DNA replication licensing factor MCM6 [EC:5.6.2.3]
Bdo3G000475	prolyl 3-hydroxylase /prolyl 3,4-dihydroxylase [EC:1.14.11.-]
Bdo3G000568	acid-sensing ion channel, other
Bdo3G000894	TELO2-interacting protein 1
Bdo3G000895	bifunctional polynucleotide phosphatase/kinase [EC:3.1.3.32 2.7.1.78]
Bdo3G000898	protein SMG6 [EC:3.1.-.-]
Bdo3G000899	WD repeat-containing protein 91
Bdo3G000900	nuclear receptor subfamily 1 group I



For the gene families from Refseq genome, one gene (gene-LOC109579598) was not found in out T2T assembly. This gene was ab initio annotated with Gnomon and function as uncharacterized. This gene can be found in the original HiFiasm untigs but unselected in final assembly, which is a bubble contig with half of the average coverage (utg000937l; LN:i:769505|rd:i:16).



3. Sex-expression differences. First, the authors should clarify whether some of the differences in gene expression are impacted by allometric differences between the sexes rather than differences in transcription rate. Second, it is not fully clear to me Fig. 4c. Specifically, what do the authors mean with DEGs in both Female and Male? Perhaps that a gene is male-biased in one anatomical part but female-biased in a different one. This should be clarified.

Authors' response:

Thank you for your comment.

(a) All tissues were dissected from peripheral tissues of mature adults for male and female, which fed under identical physiological conditions.

(b) No significant sexual dimorphism was detected in peripheral tissue size/volume.

(c) For Fig.4c, the DEGs selected based on the following criterion: a gene showed significant expression differences between male and female samples in any of the peripheral tissues. As a result, some genes may exhibit sex-biased expression in opposite directions across different tissues. We have revised the legend of fig.4c to clarify the statement.

We agree with the reviewer that RNA-seq data alone cannot precisely distinguish whether differential gene expression arises from differences in male and female cell-type composition (i.e., allometric differences) or from variations in transcriptional activity within the same cell type. To accurately differentiate these underlying causes, complementary approaches such as single-cell RNA sequencing or flow cytometry-based cell sorting followed by targeted quantification would indeed be necessary.

However, the primary objective of our study is to infer the potential functions of XY-specific genes. The rationale for identifying and focusing on DEGs is primarily to facilitate functional enrichment analysis and to predict broader functional networks to which these genes may belong. Therefore, determining the precise cellular origin of these DEGs is beyond the scope of our current investigation. For instance, our analysis identified male-biased DEGs enriched in pathways related to energy metabolism.

4. Synteny conservation interpretation (line 315). Synteny conservation can also be explained by the lack of interchromosomal rearrangements, which have a higher detrimental effect than intrachromosomal rearrangements (e.g. inversions) during meiosis. It is also unclear the explanation given in line 311 to the outcome of the phylogenomic tree. The argument seems a bit speculative or not explained properly.

Authors' response:

Thank you for your comment.

(1) Concerns about synteny is refer to our following response.

(2) We agree with you that our former description of phylogenomic tree is not clear. In the revision, this sentence has been rephrased, and a couple of sentences were supplemented to make our descriptions clear and logical.

“With 2,054 concatenated single-copy orthologous genes, we inferred the phylogenomic tree and indicated species within the genus *Bactrocera* represent a relatively late-diverging lineage within the Tephritidae family of the broader Brachycera clade(Fig.5a). Their recent adaptive radiation has been collectively driven by global geographic expansion and rapid host-associated adaptive coevolution¹⁵”

5. Synteny conservation analysis (paragraph starting in line 313). The analysis does not seem easily reproducible. To start with, the authors should provide the list of 459 single-copy genes with their names and chromosomal and molecular coordinates across all the species considered [by the way, in line 645 there is a reference to a Supplementary Table but it is not clear to which one]. Besides, the result (lack of gene order conservation is expected given the prevalence of paracentric inversions in this type of insects). Lastly, the last bit on transposons must be rewritten. The authors should become familiar with how TEs can facilitate, through ectopic recombination events, the generation of chromosomal inversions in *Drosophila* (the literature is abundant; see for example some key papers from Michael Ashburner and Alfredo Ruiz). Said this, it is not unusual that chromosomal inversions do not get generated by TEs in the genus *Drosophila*.

Authors' response:

Thank you for your comment.

We provided the input file of synteny plot in Supplementary Dataset 1. This file can used as the customized file for riparian plot in GENESPACE, with the information of the species name, gene name, location and relationship.

Combined with your valuable suggestion and the constructive feedback from Reviewer 1, we emphasized the Muller elements concept to explain the structural evolution of the genome. We also removed references to DNA transposons and revised that section as follows

Main text revised in discussion as following:

“The conservation of ancestral chromosomal synteny demonstrates extraordinary evolutionary persistence across insect lineages. In species such as lepidopterans and other insects, ancestral chromosomal synteny is often retained even after millions of years of divergence, suggesting that gene content conservation plays a crucial role in maintaining functional stability⁷. Our analysis revealed that gene content of Muller elements remains stable across the species we examined. in suborder Brachycera, and this conservation can even be traced back to the *Anopheles gambiae* genome⁸. The mechanistic basis of this evolutionary stasis appears multifaceted. In *Drosophila*, phylogenetic reconstruction demonstrates that large-scale intra-arm inversions dominate chromosomal evolution while preserving global gene order through compensatory rearrangements.

Studies from *Drosophila* genus have shown this stability is maintained through large-scale within-arm rearrangements caused by inversion, although the molecular drivers remain enigmatic⁹. Comparison among *D.melanogaster* and two closely related species, *Drosophila simulans* and *Drosophila yakuba*, suggest the mechanism of staggered breaks, forming inverted duplication sequence, primarily drive the generation of inversions¹⁰. ”

6. Interpretation of the difference in OR copy number between *D. melanogaster*/mosquitoes and *B. dorsalis* (lines 343-346). What do the authors mean by *B. dorsalis* has evolved over the past few million years, therefore a higher copy number in OR genes might have facilitated its adaptation to diverse environments? For example, the most recent ancestor of *D. melanogaster* with the *simulans* claded goes back to 1.4 million years ago only. In my view, the authors are overinterpreting a simple difference in copy number across clades, which might relate to a differential presence of a genomic feature yet to be identified, e.g. a particular sequence feature able to mediate duplications in tandem; a sequence feature that could be more abundant in *B. dorsalis*. This issue is found again in the Discussion section (paragraph lines 418-433), in which the comparison drawn in relation to plant genomes is by the way unnecessary.

Authors' response:

Thank you for your comment.

We hypothesize that olfactory genes generated through duplication, if retaining their function and not degraded into pseudogenes, might be associated with the organism's adaptation to different ecological environments. These duplicated genes may allow the detection of a more diverse range of odors, which could play a significant role in the survival and evolution of the organism. Similar hypotheses have been proposed in previous studies. For instance, the expansion of ORs in Lepidoptera (butterflies and moths) has been hypothesized to correlate with the diversification of angiosperms¹⁶, while the expansion of ORs in mammals is thought to be associated with the evolution of herbivory¹⁷. In the case of *B.dorsalis*, its diverse host range suggests that the large tandem duplications of OR genes in its genome may be associated with the need to detect a wide variety of host-derived semiochemicals.

7. Sentence in lines 403-405. Please clarify how the genome resources generated here actually provide new insights into chromosome structure polymorphism. It is unclear to which results the authors are referring.

Authors' response:

Thank you for your comment.

We have revised this part into:

“Our genome provides a valuable example of a complete genome assembly within Diptera species. The continuous generation of new high-quality genomes not only advances comparative genomics research but also enhances our understanding of genome evolution in pest species that are in constant conflict with humans.”

8. Last paragraph in the Discussion. In the way the authors provide some of the results, I do not think that some of the resources generated in this article are necessary useful to the research community. For example, the authors could provide the list of orthogroups and their constituent genes between *B. dorsalis* and the rest of species used in comparative analyses (Supplementary table 11), being the one between *B. dorsalis* and *D. melanogaster* extremely useful to many colleagues.

Authors' response:

Thank you for your suggestion. We have deposited the results of the orthogroup in orthofinders in Supplementary Dataset 1. In addition to the Supplementary Table provided in the appendix, the data used for figure generation will also be made available in accordance with the Nature Communications Source Data requirements.

Minor comments:

1. More informative than the average read length is the equivalent of N50 applied to the length distribution of sequencing reads (lines 94-95).

Authors' response:

Thank you for your comment.

We have added the statistics for Illumina sequencing, HiFi CCS reads and ONT reads in the Supplementary Table 1

2. It would be nice to see the % that the different contaminants represented for both initial sequencing outputs (to be added to Sup. Table 1).

Authors' response:

Thank you for your comment.

This information has been provided in this revised main text:

“As a result, 14.31% of reads from the female individual and 2.1% from the male individual were excluded.”

Reference

- 1 Tsoumani, K. T., Drosopoulou, E., Mavragani-Tsipidou, P. & Mathiopoulos, K. D. Molecular characterization and chromosomal distribution of a species-specific transcribed centromeric satellite repeat from the olive fruit fly, *Bactrocera oleae*. *PLoS One* **8**, e79393 (2013). <https://doi.org:10.1371/journal.pone.0079393>
- 2 Escriba, M. C. *et al.* Molecular and cytological characterization of repetitive DNA sequences from the centromeric heterochromatin of *Sciara coprophila*. *Chromosoma* **120**, 387-397 (2011). <https://doi.org:10.1007/s00412-011-0320-2>
- 3 Lukhtanov, V. A. & Pazhenkova, E. A. Diversity and evolution of telomeric motifs and telomere DNA organization in insects. *Biol J Linn Soc* **140**, 536-555 (2023). <https://doi.org:10.1093/biolinnean/blad068>
- 4 de Lima, L. G., Svartman, M. & Kuhn, G. C. S. Dissecting the Satellite DNA Landscape in Three Cactophilic Sequenced Genomes. *G3-Genes Genom Genet* **7**, 2831-2843 (2017). <https://doi.org:10.1534/g3.117.042093>
- 5 Mason, J. M., Frydrychova, R. C. & Biessmann, H. *Drosophila* telomeres: an exception providing new insights. *Bioessays* **30**, 25-37 (2008). <https://doi.org:10.1002/bies.20688>
- 6 Meccariello, A. *et al.* Maleness-on-the-Y (MoY) orchestrates male sex determination in major agricultural fruit fly pests. *Science* **365**, 1457-1460 (2019). <https://doi.org:10.1126/science.aax1318>
- 7 Wright, C. J., Stevens, L., Mackintosh, A., Lawniczak, M. & Blaxter, M. Comparative genomics reveals the dynamics of chromosome evolution in Lepidoptera. *Nat Ecol Evol* **8** (2024). <https://doi.org:10.1038/s41559-024-02329-4>
- 8 Coluzzi, M., Sabatini, A., della Torre, A., Di Deco, M. A. & Petrarca, V. A polytene chromosome analysis of the *Anopheles gambiae* species complex. *Science* **298**, 1415-1418 (2002). <https://doi.org:10.1126/science.1077769>
- 9 Schaeffer, S. W. Muller "Elements" in *Drosophila*: How the Search for the Genetic Basis for Speciation Led to the Birth of Comparative Genomics. *Genetics* **210**, 3-13 (2018). <https://doi.org:10.1534/genetics.118.301084>
- 10 Ranz, J. M. *et al.* Principles of genome evolution in the *Drosophila melanogaster* species group. *PLoS Biol* **5**, e152 (2007). <https://doi.org:10.1371/journal.pbio.0050152>
- 11 Vicoso, B. & Bachtrog, D. Numerous transitions of sex chromosomes in Diptera. *PLoS Biol* **13**, e1002078 (2015). <https://doi.org:10.1371/journal.pbio.1002078>
- 12 Perte, M., Kim, D., Perte, G. M., Leek, J. T. & Salzberg, S. L. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat Protoc* **11**, 1650-1667 (2016). <https://doi.org:10.1038/nprot.2016.095>
- 13 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014). <https://doi.org:10.1186/s13059-014-0550-8>
- 14 Wang, C., Li, Z., Typas, M. A. & Butt, T. M. Nuclear large subunit rDNA group I intron distribution in a population of *Beauveria bassiana* strains: phylogenetic implications. *Mycol Res* **107**, 1189-1200 (2003). <https://doi.org:10.1017/s0953756203008505>
- 15 Zhao, Z., Carey, J. R. & Li, Z. The Global Epidemic of *Bactrocera* Pests: Mixed-Species Invasions and Risk Assessment. *Annu Rev Entomol* **69**, 219-237 (2024). <https://doi.org:10.1146/annurev-ento-012723-102658>

- 16 Engsontia, P., Sangket, U., Chotigeat, W. & Satasook, C. Molecular evolution of the odorant and gustatory receptor genes in lepidopteran insects: implications for their adaptation and speciation. *J Mol Evol* **79**, 21-39 (2014).
<https://doi.org/10.1007/s00239-014-9633-0>
- 17 Hughes, G. M. *et al.* The Birth and Death of Olfactory Receptor Gene Families in Mammalian Niche Adaptation. *Mol Biol Evol* **35**, 1390-1406 (2018).
<https://doi.org/10.1093/molbev/msy028>

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

The updated manuscript features addresses the issues raised in the previous round of review. In particular, much more methodology details are given, which allows for evaluation of the assembly process and definitely adds to the appreciation of amount of manual adjustments still needed to finalise T2T assembly even for relatively compact genome of a dipteran insect. I also appreciated that most of the figures in the response were also added as supplementary figures and believe that the remaining ones used for demonstrations for the assembly process (Hi-C plot of haplotypic duplication, coverage breaks in misjoins) can be also added to the supplementary material.

Authors' response:

We sincerely thank the reviewer for the encouraging and insightful comments. We are especially grateful for the reviewer's valuable suggestion and have included the relevant information in the supplementary material as recommended.

Reviewer #2 (Remarks to the Author):

I appreciate the authors' significant efforts in addressing my previous concerns. Most of my comments have been satisfactorily resolved, resulting in a substantial improvement in the manuscript's quality. With the following minor issues addressed, the manuscript will be ready for publication:

Authors' response:

We sincerely thank the reviewer for the positive and encouraging comments. We provide point-by-point responses to the remaining minor issues as following:

1. *“Assembly continuity achieves a GCI score of 31.5 for the whole genome, and 51.4 and 57.0 for X and Y chromosomes separately.”*

Typically, the Y chromosome is the most challenging to assemble due to its overrepresentation of repetitive sequences, which usually results in a lower GCI score. However, the authors report the highest score for the Y chromosome, which is unexpected. Since GCI software is relatively new, it may not perform optimally on certain genomes. Please verify whether the software failed in this specific case. If so, consider omitting this information.

Authors' response:

GCI was originally designed for genomes with low heterozygosity and has been tested on human (CHM13), rice (MH63), Arabidopsis (Col-CEN), and chicken assemblies. In our case, two key factors may have affected the result of GCI.

First, the genome exhibits relatively high heterozygosity in the autosomes, whereas the sex chromosomes show much lower heterozygosity. This is likely due to reduced recombination rates and a more compact chromatin structure in sex chromosomes, especially in a limited population.

Second, the sequencing coverage of HiFi-CCS reads in our dataset is lower compared to the other reference samples, primarily due to the limited biomass available from individual insect. As a result, the combined GCI score was more heavily influenced by ONT long reads, which were obtained from pooled samples. This further reduces the GCI values, particularly for the autosomes.

2. *“However, due to the high sequence similarity (90%) of the CDS regions among the three copies, current qPCR methods cannot precisely distinguish or quantify individual copies.”*

Please explicitly state this limitation to caution readers.

Authors' response:

We sincerely thank the reviewer for the valuable suggestion. As advised, we have now explicitly stated this limitation in the discussion of revised manuscript to caution readers.

“Here, our current study has certain limitations. Specifically, we were unable to distinguish the functions of individual transcript isoforms, as the high sequence similarity (>90%) among the tandemly duplicated copies of *BdorOR88a* presents a significant challenge for accurate expression quantification. To overcome these limitations, future studies could adopt more sensitive techniques, such as RNAscope, and integrate ligand identification, mutant analyses,

and electrophysiological approaches to uncover how *BdorOR88a* contributes to behavioral adaptation and environmental responses in *B. dorsalis*.”

3. “The results showed that the expression pattern of the three copies remained unchanged after the knockout.”

This statement is unclear. If these copies were knocked out, how could the expression pattern remain unchanged? Please clarify the writing, particularly in the main text.

Authors’ response:

Thank you for your comment.

CRISPR/Cas9 editing caused the loss of *BdorOR88a1* and *BdorOR88a2* function by introducing premature stop codons, leading to truncated, nonfunctional proteins. However, transcription was not eliminated, and mRNA of the targeted genes remained detectable.

Transcriptomic profiling was conducted to investigate whether the typically low-expressed *BdorOR88a3* might exhibit upregulated transcription as a functional compensation following the knockout of *BdorOR88a1* and *BdorOR88a2*, which could potentially obscure the functional roles of the highly expressed gene copies.

Therefore, we have added the following clarification to the main text:

“To ensure that the attenuation of ME-directed chemotaxis behavior was not affected by potential functional compensation from the typically low-expressed *BdorOR88a3*, we conducted transcriptomic analysis of the mutants. This step is important because low-expressed genes, like *BdorOR88a3*, may sometimes exhibit upregulated expression or compensatory effects when other homologous gene copies (such as *BdorOR88a1* and *BdorOR88a2*) are knocked out. Such compensation could lead to potential misinterpretations of the phenotypic effects of the highly expressed gene copies. The results revealed that the transcriptional expression pattern of *BdorOR88a3* remained unchanged after the knockout of *BdorOR88a1* and *BdorOR88a2*. This finding rules out the possibility of functional compensation from *BdorOR88a3*, confirming that the observed behavioral changes are primarily attributable to the loss of function of the two highly expressed gene copies (Supplementary Fig. 7g-i).”

Additionally, the link to the China National GeneBank Database (CNGBdb) in the Data Availability section is currently inaccessible. Please ensure the resource is made available upon formal acceptance of the manuscript.

Authors’ response:

Thank you for your kind reminder. The genome assembly was already submitted prior to the manuscript’s submission. The reviewer access has now been updated and is available at the following address:

<https://ngdc.cncb.ac.cn/gwh/Assembly/reviewersPage/ekdsPpIDVivUIaeWhCjsBGBUBwVuzsffCpyqRNhEFAUyhJTrQSdILPFUPatTmQeH>

According to the policy of CNGBdb, the release of all associated data will be triggered with the public release of the genome data. In addition, we have provided the running title and the name of the first author on the same page. Once the publication is indexed in PubMed, the data release will be automatically triggered.

Reviewer #3 (Remarks to the Author):

The new submission ameliorates some important limitations (transparency and level of detail) identified in the first version. But a few aspects could still benefit from adding further details and accuracy:

Authors' response:

We sincerely thank the reviewer for the constructive comments. We appreciate the suggestions for further refinement and have carefully addressed the remaining issues by adding more details and improving the accuracy as requested. Please find our point-by-point responses below.

1. Reply to point 5, reviewer 3, new text. The sentence 'The mechanistic basis of this evolutionary stasis appears multifaceted. In Drosophila, phylogenetic reconstruction demonstrates that large-scale intra-arm inversions dominate chromosomal evolution while preserving global gene order through compensatory rearrangements' is a bit speculative. Specifically, the part underlined. What do the authors mean by "compensatory rearrangements"? What is the evidence for that statement?

Authors' response:

Thank you for your insightful comments. We agree that the original summary regarding *Drosophila* was overly speculative, and the use of the term "compensatory rearrangements" was not appropriate in this context.

Here, we previously used "compensatory" to describe the long-term stable chromosome arms across *Drosophila*, rather than a direct corrective relationship between individual rearrangements.

This pattern has been discussed in the context of:

1. The dominance of intra-arm inversions that preserve chromosome arm integrity, as summarized in the literature

Ref: S. W. Schaeffer. Muller "Elements" in *Drosophila*: How the Search for the Genetic Basis for Speciation Led to the Birth of Comparative Genomics. Genetics 2018.

2. One of recent theoretical model suggest that the inversion polymorphisms in *Drosophila* are necessary. That can facilitate the accumulation of sexually antagonistic alleles, which in turn reduces the fitness cost of recombination and helps stabilize these polymorphisms

Ref: C. S. McAllester and J. E. Pool. The potential of inversions to accumulate balanced sexual antagonism is supported by simulations and *Drosophila* experiments. Elife 2025)

In light of your comment, we have revised the sentence to more accurately reflect the evidence: "In *Drosophila*, comparative genomic study and phylogenetic reconstruction demonstrates that large-scale intra-arm inversions dominate chromosomal evolution."

2. Like above, I am not sure what is the point of the underlined text in the following sentence: 'Studies from Drosophila genus have shown this stability is maintained through large-scale within-arm rearrangements caused by inversion, although the molecular drivers remain enigmatic'. We know from Ashburner and Ruiz's publications that transposable elements can mediate the origin of some chromosomal inversions while for others it has not been possible to

find evidence of such TE involvement, and instead duplications of flanking regions at the inversion breakpoints have been documented. This led to the proposed staggered-breakpoints model (ref. 68). In sum, there is either involvement of TEs or not. We have examples of both types of scenarios. Therefore, what are the authors referring to with “enigmatic”?

Authors’ response:

The term “Enigmatic” refers to the description in Stephen W. Schaeffer’s review, which states “Nucleotide sequences at inversion breakpoints have provided clues about how inversions are generated in populations, although there does not appear to be a universal mechanism for the breakage and rearrangement of chromosomal segments.”

According to your suggestion, we have revised the sentence in the manuscript to:

“At the nucleotide sequence level, comparison between *D. simulans* and *D. yakuba* show that most inversion breakpoints are consistent with the staggered-breakpoint model, while only a few are associated with repetitive sequences”

3. Data availability section. It is indicated that the chromosome assembly ‘will be’ submitted to the China National GeneBank Database. Shouldn’t it be already submitted and available?

Authors’ response:

Thank you for your kind reminder. The genome assembly was already submitted prior to the manuscript’s submission. The reviewer access has now been updated and is available at the following address:

<https://ngdc.cncb.ac.cn/gwh/Assembly/reviewersPage/ekdsPpIDVivUIaeWhCjsBGBUBwVuzsffCpyqRNhEFAUyhJTrQSdILPFUPatTmQeH>

According to the policy of CNGBdb, the release of all associated data will be triggered with the public release of the genome data. In addition, we have provided the running title and the name of the first author on the same page. Once the publication is indexed in PubMed, the data release will be automatically triggered.

4. Code availability. It seems unnecessary providing the code for built-in functions (e.g. in R) used in conventional statistical analyses, but I see it necessary in the case of the analytical framework used to generate the genome assembly of B. dorsalis (and even the one for the differential expression analysis). It would dispel doubts, enhance the transparency of the work done, and increase the visibility of the paper.

Authors’ response:

We appreciate the reviewer’s thoughtful suggestion. We agree that providing the analytical framework used for genome assembly and differential expression analysis would enhance the transparency and reproducibility of our work. In response, we have now made the relevant custom workflow descriptions and other dataset publicly available on <https://github.com/linqiangATcaas/BdoT2T>

As noted in the manuscript, all software tools used in our analyses have already been listed. We did not develop any new algorithms during the assembly process; instead, our customizations were primarily limited to read filtering and contig connection strategies.

5. In the supplementary file with the orthogroups composition, tracking the identification of the genes from different species is difficult. Additionally, for *D. melanogaster*, the FlyBase identifier (e.g. FBgn0038979) is preferred.

Authors' response:

We have reorganized this part of the dataset and updated the accompanying documentation (ReadMe.txt in Supplementary Dataset 1). The revised version now includes additional gene-related information, such as Gene Name, Gene ID, and the full name of the protein product. For *D. melanogaster*, we have further added the FlyBase identifiers. We hope these improvements will facilitate easier retrieval and use of the data by readers.

6. Relative to the analyses presented in Fig. 5. The choice of species is strange in some cases. For example, *Anopheles coluzzii* is included but not an emblematic species such as *Anopheles gambiae*, which in theory has a much better assembly and annotation. Or in the genus *Drosophila*, the authors choose *Drosophila mauritiana* along with *Drosophila melanogaster*, which gives the impression of conservation of gene order and strong chromosomal stasis. These two are sister species, having diverged only a few million years ago, and differentiated by one single inversion. The authors could have chosen *Drosophila pseudoobscura* or *Drosophila virilis*, much more distantly related to *Drosophila melanogaster* to make the fair point that despite the large intra-chromosomal repatterning, there is no exchange between different Muller's elements. In sum, the rationale for choosing some species is far from clear and justified, and in fact a bit misleading.

Authors' response:

Thank you for your valuable comments.

The species included in our analysis were primarily selected based on the completeness of their reference genome assemblies, with a preference for assemblies derived from male individuals and those showing better BUSCO completeness scores. In addition, we prioritized assemblies that have an official RefSeq Annotation Release, as these tend to offer higher-quality gene annotations with better traceability, data sources, and annotation consistency compared to GCF assemblies without formal annotation releases.

However, we acknowledge that we did not initially consider the phylogenetic representativeness of the selected species for visualizing genome synteny. In response to your suggestion, we have recalculated and updated **Figure 5** accordingly.

In the revised Figure 5, the neo-X chromosome in *D. pseudoobscura* clearly reflects a fusion event between the ancestral X chromosome and chromosome arm 3L of *Drosophila melanogaster*.

7. Reply to point 5, reviewer 3. The authors could be more explicit, indicating the name of the file, which I believe is *Bdo_phaseBlks.csv_6spe*. They can do that in the Methods.

Authors' response:

Thank you for the suggestion.

As advised, we have now explicitly included the file (Supplementary Dataset 2) contain *Bdo_phaseBlks.csv* and *Gene_coordinate* in the Methods section to help readers more easily

locate and interpret the corresponding data.

Both “Supplementary Dataset 1” and “Supplementary Dataset 2” have been moved to the method section.

“The orthogroup composition of OrthoFinder result and the genomic coordinate information used for the synteny analysis are provided in Supplementary Dataset 1 and Supplementary Dataset 2, respectively.”

REVIEWERS' COMMENTS

Reviewer #2 (Remarks to the Author):

My remaining minor comments have been addressed well, and I have no further comments.

Authors' response:

Thank you very much for your valuable suggestions. We deeply appreciate your time and assistance.

Reviewer #3 (Remarks to the Author):

I appreciate the authors' effort for improving unclear aspects of the manuscript.

Authors' response:

We sincerely thank you for your valuable time and constructive comments.

Nevertheless, there are still some minimal inaccuracies that should be fixed. In reply to this reviewer's comments:

Point 1. The new text is correct and the citation (67) to Prof. Schaeffer's review is pertinent.

Authors' response:

Thank you for your comment.

*Point 2. The text in lines 436-443 is unnecessarily long with some errors. Reference 67 does not talk specifically about the comparison between *D. simulans* and *D. yakuba*. That comparison, which involves the analysis of sequence features at the breakpoint regions of chromosomal inversions, is in fact part of the content in reference 68. Please read again the paragraph that begins with "Nucleotide sequences at inversion breakpoints have provided clues about how" in reference 67.*

Authors' response:

Thank you for your comment. We replaced the Ref 68 with 67 and remove the redundant content.

Points 3, 4, 5, and 7. Great improvements.

Authors' response:

Thank you for your help.

*Point 6. I appreciate the inclusion of *D. pseudoobscura*. One thing that surprised me though is that now the phylogenetic analysis includes 1914 single-copy shared orthologs while in previous versions of the manuscript this number was 459. I think it is positive that the analyses had been repeated with a larger number of genes, but I am surprised that the authors do not indicate in their reply why this dramatic disparity. Perhaps I am missing something obvious but what has changed?*

Authors' response:

Thank you for your comment.

This issue originated in the initial submission, during which we used stricter filtering criteria

for the dataset. When we performed gene family analysis, we used the processed data originally prepared for synteny analysis. This led to the exclusion of genes located on scaffolds shorter than 100 kb. In the assemblies of *Teleopsis dalmanni* (GCA_002237135.2, 3 chromosomes + 5644 unplaced scaffolds) and *Bradysia coprophila* (GCF_014529535.1, 4 chromosomes + 320 unplaced scaffolds), a large number of genes were not assigned to scaffolds longer than 100 kb. As a result, the number of single-copy orthologous genes obtained was relatively low. We re-analysis the data with whole gene datasets and finally yielding 1,914 single-copy orthologous genes. The topology of the phylogenetic tree remained unchanged. We have corrected the results and revised the main text and corresponding figures.