RESPONSE LETTER TO THE COMMENTS OF THE RECOMMENDER AND REVIEWERS

Resubmission MS Title:

Draft genome and transcriptomic sequence data of three invasive insect species

Available at: https://doi.org/10.1101/2024.09.02.610743

Authors:

Eric Lombaert, Christophe Klopp, Aurélie Blin, Gwenolah Annonay, Carole Iampietro, Jérôme Lluch, Marine Sallaberry, Sophie Valière, Riccardo Poloni, Mathieu Joron & Emeline Deleury

Submitted to: PCI Genomics

Dear Recommender,

We found the reviewer's comments and criticisms helpful and cogent. We have incorporated most of their suggestions and feel our manuscript is much improved. We hope that you will find the MS appropriate for recommendation in PCI Genomics.

Please find below our detailed responses (in italic characters) to your comments as well as to those of the two reviewers (in bold characters).

Best wishes,

Eric Lombaert on behalf of all authors

Recommender's comments:

Dear Eric Lombaert and colleagues,

Your manuscript entitled "Draft genome and transcriptomic sequence data of three invasive insect species" has been reviewed by two colleagues. Globally the reviews are positive but there are substantial criticisms that you need to address before I can finally decide on whether this preprint can be recommended or not.

In particular, the quality of the assembly of L occidentalis should be improved and discussed as there seems to be issues related to the incorrect handling of the high heterozygosity, possibly in conjunction with the high repeat content of this genome, as well as the potential presence of contaminants. These issues may be addressed using more appropriate methods (purge_dups, pretext). Given the high repeat content of this genome, it may be that some BUSCO genes indeed belong to duplicated regions. A detailed analysis of the BUSCO genes flagged as duplicates in Table 3 (and Table 4) could enable to clarify this.

The quality of the assembly may also certainly be improved using higher HiFi coverage.

Additionally, I find it intriguing that only 72.5% of RNAseq short reads map to the assembly (and even less than 50% for some samples). This suggests that either there is some level of contamination in your data, or that the un-assembled regions of the genome (likely repeats) are transcribed. I think this point is particularly interesting to clarify and discuss.

Finally, although this paper is a data paper and its main contribution is to give access to this data for the community, I think that it indeed would greatly benefit from providing more context on the biology of these species, and discuss the pecularities of their genomes as well as the associated bioinformatics challenges.

Thank you for your feedback and the detailed comments. We have carefully addressed the concerns raised in the reviews, as outlined in our responses to the reviewers (see below). In brief:

1. *Quality of* Leptoglossus occidentalis *assembly:*

   We have reviewed and corrected the initial errors in Table 3, including the BUSCO values, which led to the misinterpretation of duplication rates. We sincerely apologize for this mistake and have re-verified all relevant values in the manuscript. Our responses to the reviewers further clarify the steps taken to address concerns regarding allelic duplications and the methods used to assess assembly quality.

2. *HiFi coverage:*

   Regarding the sequencing depth for Leptoglossus occidentalis*, we acknowledge that higher sequencing depth could have been beneficial. However, as detailed in our response to the reviewers (see below), our assembly quality was initially underestimated due to an incorrect N50*

value in the manuscript. The correct value (147.7 Mb) indicates a high level of contiguity for this assembly, despite the lower sequencing depth compared to the two other species. We have revised the manuscript accordingly and carefully verified all reported values.

3. *RNAseq mapping:*

As noted, our data do indeed exhibit variable mapping rates, with some individuals showing low values. To explicitly highlight this observation, we have now added a sentence in the manuscript (Lines 81-84): "Mapping the reads from each library onto the genome assemblies built from these same data resulted in good alignment rates for long reads (nearly 100%) and Hi-C sequences (close to 97%). For RNA-seq, mapping rates were more variable, with an average of 72.5% and some individuals exhibiting low values (see Table 2 for details)."

To further explore the data quality, we performed additional analyses on the two datasets exhibiting the lowest mapping rates: SRA accession numbers SRR30002757 (the Tecia solanivora *individual with a 36.92% mapping rate) and SRR30002765 (the* Leptoglossus occidentalis *individual with a 58.32% mapping rate). We first used Kraken2 (v2.1.3; Wood et al. 2019) to explore the possibility of contamination contributing to the unaligned reads. Taxonomic classification of RNAseq reads, summarized in Table A below, confirms that most sequences are eukaryotic and primarily of insect origin. Bacterial reads are present, but their proportion is relatively low (e.g., below 3.5% in the most affected sample). While some level of contamination is likely, it does not appear to be the main factor driving the lower mapping rates. Instead, other factors, such as the relatively high level of heterozygosity, may contribute to these reduced mapping rates by complicating alignment. Indeed, we observed significantly higher alignment rates when using HiSat2 (Kim et al. 2019) instead of STAR, as HiSat2 applies less stringent criteria and allows more mismatches (results not shown).*

	SRR30002757 *% of total reads*	*SRR30002765* *% of total reads*
Unclassified	*6.25%*	*13.19%*
Insecta	*47.12%*	*44.08%*
Other arthropods	*4.11%*	*3.46%*
Other Eucaryotes	*18.99%*	*24.29%*
Bacteria	*3.43%*	*1.07%*
Archaea	*0.01%*	*0.04%*
Viruses	*0.05%*	*0.09%*
Other	*20.03%*	*13.79%*

Table A: Taxonomic classification of RNAseq reads for the two datasets with the lowest mapping rates.

Despite the lower mapping rates observed in certain samples, we believe that these RNAseq datasets remain valuable for the community. To illustrate their utility, we quantified the number of reads assigned to annotated features using featureCounts from the Subread package (v2.0.8; Liao et al. 2014). We found that 57.3% and 48.8% of aligned reads were assigned to features (exons in genes) in SRR30002757 and SRR30002765, respectively. Furthermore, the

number of genes supported by at least 10 reads remains substantial (11,779 in SRR30002757 and 9,977 in SRR30002765). High-coverage genes are also present, with some features accumulating hundreds of thousands of mapped reads. The top 10 mapped features for dataset SRR30002757:

Feature identifier	*Feature length*	*Number of mapped reads*
_ptg000071l_000004	*1452*	*837418*
_ptg000001l_000178	*2461*	*565962*
_ptg000062l_000012	*1036*	*391855*
_ptg000001l_000177	*2462*	*367495*
_ptg000015l_000246	*724*	*276669*
_ptg000015l_000247	*626*	*234098*
_ptg000008l_000629	*1168*	*234053*
_ptg000015l_000254	*957*	*217870*
_ptg000008l_000500	*3106*	*188081*
_ptg000015l_000457	*891*	*177958*

The top 10 mapped features for dataset SRR30002765:

Feature identifier	*Feature length*	*Number of mapped reads*
_HiC_scaffold_4_000095	*1394*	*584798*
_HiC_scaffold_3_001063	*7273*	*492805*
_HiC_scaffold_11_000409	*1393*	*361915*
_HiC_scaffold_6_001549	*2453*	*252857*
_HiC_scaffold_2_000576	*923*	*202557*
_HiC_scaffold_4_001567	*912*	*196786*
_HiC_scaffold_6_000039	*2483*	*166270*
_HiC_scaffold_5_001909	*2194*	*160340*
_HiC_scaffold_6_000040	*2473*	*157224*
_HiC_scaffold_4_000416	*2078*	*150933*

These results shows that, despite the suboptimal mapping rates, these RNAseq datasets retain significant information usable for gene annotation and expression studies.

In summary, while we acknowledge the limitations in mapping rates (which we now explicitly mention in the revised manuscript), the data remain useful for the community, particularly for gene annotation efforts. We thank the recommender for raising this point, as it allowed us to investigate potential sources of unmapped reads and reinforce the relevance of these datasets.

4. *<u>Context and biological relevance:</u>*

   We have substantially expanded the introduction (Background section), more than doubling its length, to provide additional context on the biology of the studied species and the specific challenges associated with their genome assemblies. This revised section now offers a more comprehensive discussion of their genomic characteristics and bioinformatics challenges.

For further details, please refer to our responses to the reviewers below. We hope these revisions address your concerns and further highlight the quality and relevance of our manuscript.

Review #1 by Jean-Marc Aury

The article presents a thorough description of the sequencing, assembly, and annotation of the genome of three invasive insect species. The sequencing methods used, as well as the tools for assembly and annotation, appear appropriate. I congratulate the authors for their contribution to obtaining reference genomes for biodiversity.

I have some concerns regarding the workflow, specifically the failure to eliminate potential allele duplications (typically done using tools such as purgedup). This is especially important considering the high heterozygosity rate observed in all three species. The genome assembly of Cydalima perspectalis is 5% larger than the available assembly (which included the use of purgedup), raising questions about potential duplications or errors. For Leptoglossus occidentalis, the BUSCO duplication rate is extremely high, indicating that there may be an issue with assembly accuracy. Furthermore, no mention is made of the detection of potential contaminants (e.g., bacterial sequences), which is a crucial step in ensuring the correctness of public databases, especially in terms of taxonomic assignments.

Thank you for your comments regarding potential allele duplications and contamination detection. Below, we address each of your concerns:

1. *Allele Duplication and Assembly Accuracy for* Leptoglossus occidentalis*:*

   Due to an oversight when finalizing the submitted manuscript, the BUSCO values initially reported for Leptoglossus occidentalis *in Table 3 were those obtained before applying* purge_dups*. We confirm that* purge_dups *was indeed applied prior to scaffolding, reducing the total assembly size from 2.098 Gb to 1.769 Gb and the number of contigs from 7,377 to 4,976. This information, which was mistakenly omitted in the previous version, has now been added to the text (Lines 226-229). The updated BUSCO scores (using version 5.2.2, as in the submitted manuscript) show a significant decrease in duplication (from 16.2% to 2.4%), confirming that allele duplications were effectively handled. We apologize for this omission and for any confusion it may have caused.*

2. *Updated BUSCO Analyses:*

   As part of this review process, we took the opportunity to recalculate BUSCO scores for all three species using the latest available version of the BUSCO software package (v5.7.1) and have incorporated the updated results into Table 3. The updated values are:

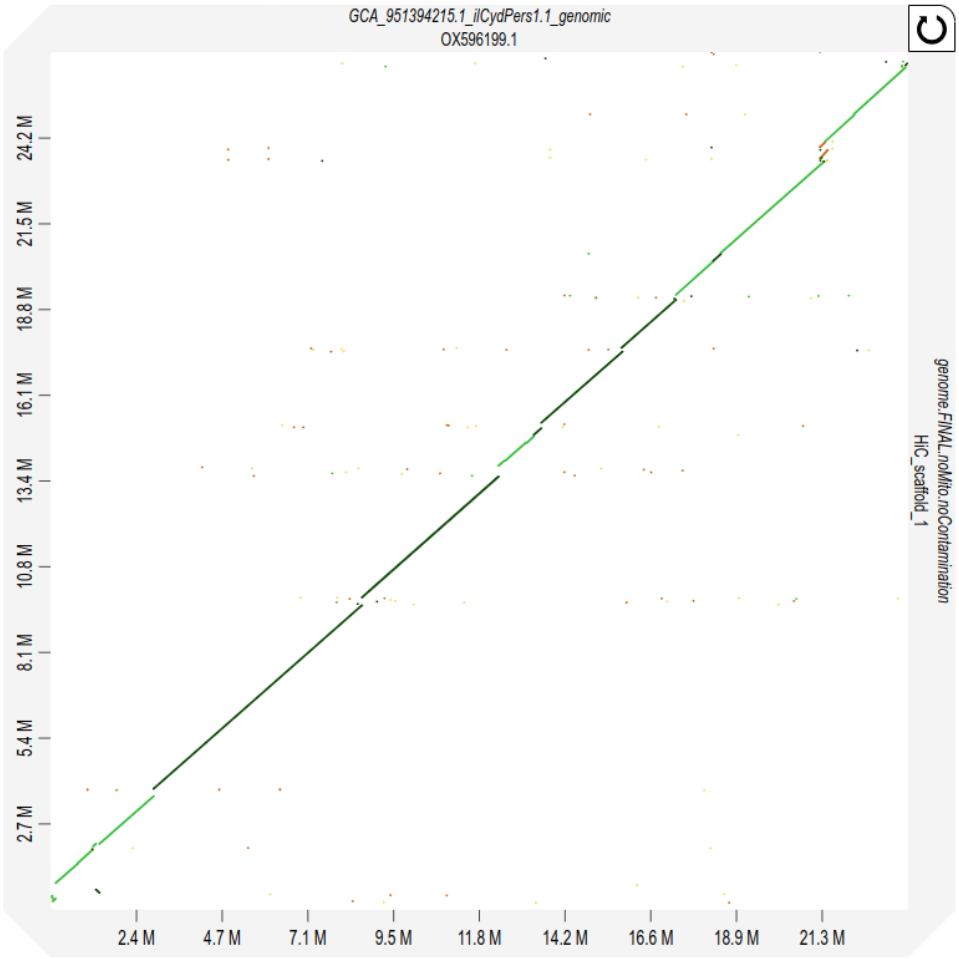
   - Leptoglossus occidentalis*: C:98.9%[S:96.6%,D:2.3%], F:0.4%, M:0.7% (n=1367)*
   - Cydalima perspectalis*: C:99.7%[S:99.5%,D:0.2%], F:0.1%, M:0.2% (n=1367)*
   - Tecia solanivora*: C:99.5%[S:98.8%,D:0.7%], F:0.4%, M:0.1% (n=1367)*

These results confirm the high completeness and low duplication rates of our assemblies. Additionally, to further assess the reliability of the Leptoglossus occidentalis *assembly, we compared its BUSCO scores to those of* Leptoglossus phyllopus*, the closest available related species with a reference genome. The similarities (*Leptoglossus phyllopus*: C:99.6%[S:97.6%,D:2.0%],F:0.1%,M:0.3%) show that our assembly scores correspond to those of other published assemblies.*

3. *<u>Genome Size Differences:</u>*

   Regarding the genome size discrepancy observed for Cydalima perspectalis*, we acknowledge that our assembly is 3.4% larger than the previously published version. Such differences can arise due to methodological variations, including sequencing depth, assembly parameters, and repeat resolution. A detailed comparison reveals that all chromosomes in our assembly are larger, with size differences ranging from +0.24% to +11.9%.*

   An examination of the Z chromosome, which shows the highest relative size increase (+11.9%), indicates that the additional sequences are primarily located in peri-telomeric regions and in previously unassembled blocks, as evidenced by gaps in the other assembly (Figure A).



<u>Figure A:</u> Dot-plots showing Z Chromosome alignment visualized with d-genies (Cabanettes and Klopp 2018) of our Cydalima perspectalis assembled genome (Y-axis) to previously publicly available genome of the same species (GCA_951394215.1; X-axis).

Telomere identification analyses across all scaffolds reveal a significantly higher number of properly placed telomeric sequences in our assembly, strongly suggesting improved completeness in these regions (see Figure B for a detailed view of the 32 chromosomes).

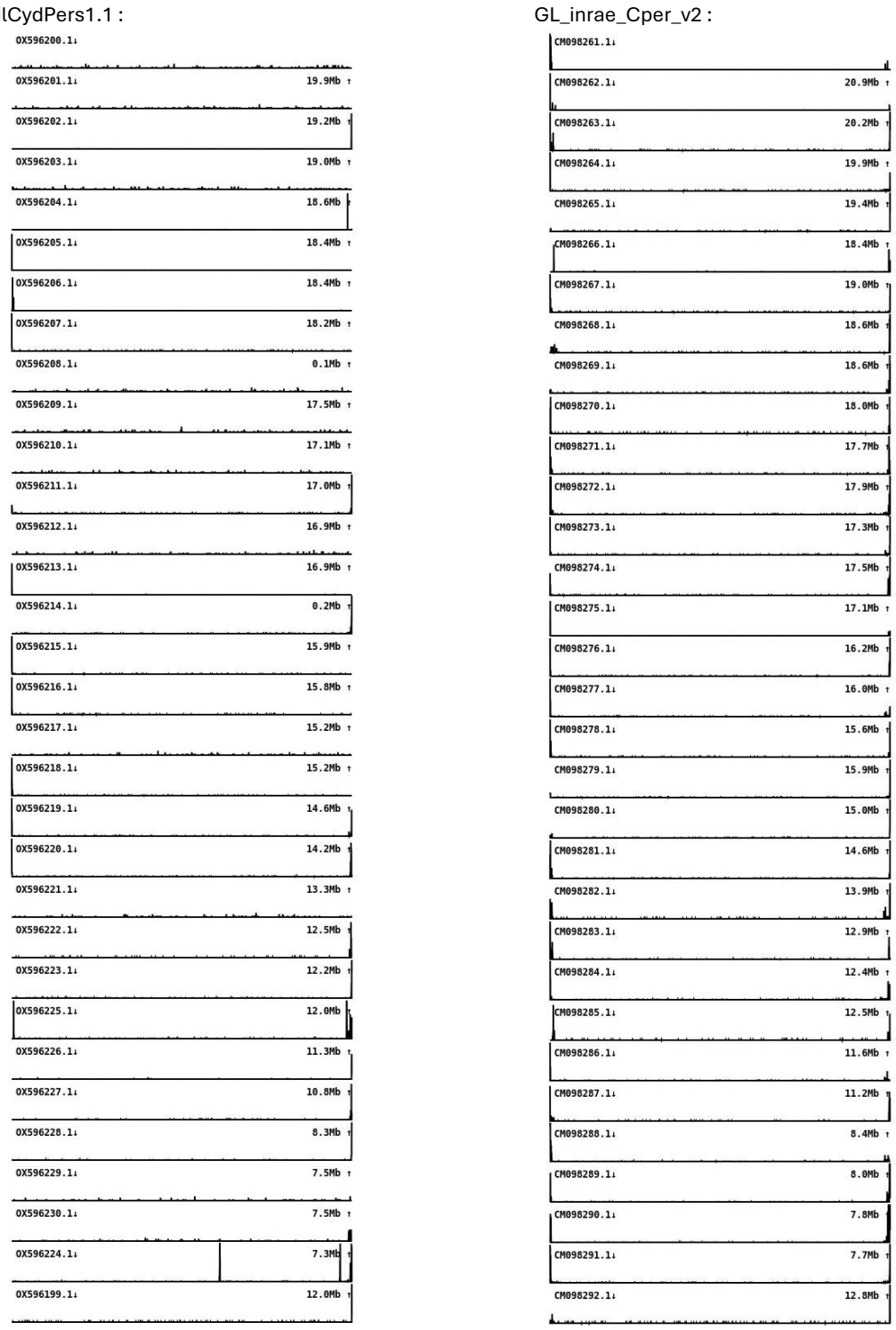


Figure B: Occurrence of telomeric repeats across the 32 assembled chromosomes of Cydalima perspectalis*. Analysis was performed using tidk (Brown et al. 2025). Peaks at chromosome ends indicate successfully assembled telomeres, while their absence suggests missing sequences. The y-axis represents the frequency of the telomeric repeats, while the x-axis represents the position along each chromosome. Telomeric repeats are found much more frequently in the assembly of this study (right panel) than in the previously published assembly (left panel, GCA_951394215.1).*

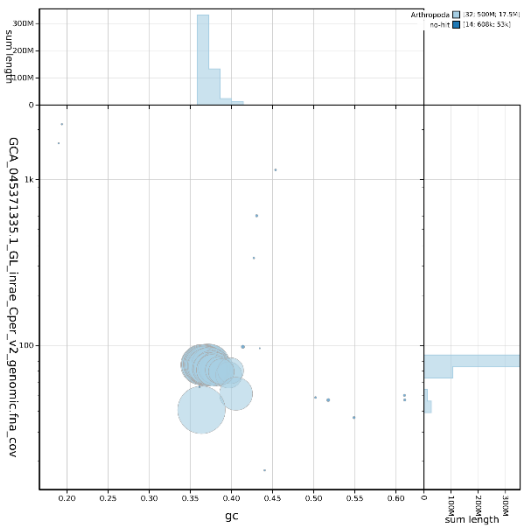
We did not find specific methodological details in Broad et al. (2024) that could fully explain these differences. However, our assembly exhibits higher contiguity, greater depth, and improved telomer presence, which can explain why the NCBI chose it as reference.

4. *<u>Contaminant Detection:</u>*

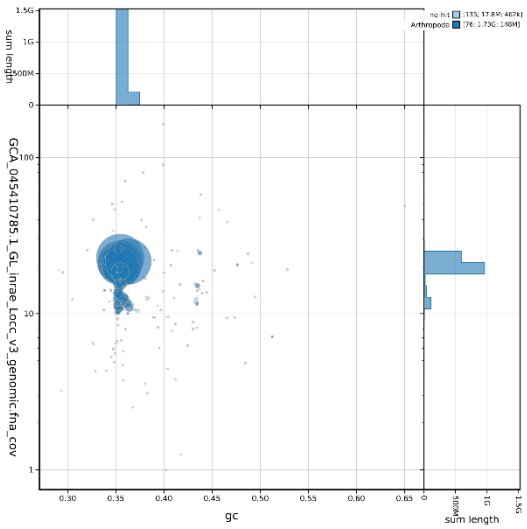
We acknowledge the importance of ensuring contamination-free genome assemblies. During the initial submission process, NCBI's automated screening identified a number of scaffolds as potential contaminants. These scaffolds were systematically removed before finalizing the assemblies.

To further validate the absence of contamination in the final assemblies presented in this study (i.e., the versions currently available on NCBI), we conducted an additional analysis using BlobToolkit. The results confirm that no significant contamination remains in any of the three assemblies (see Figure C).

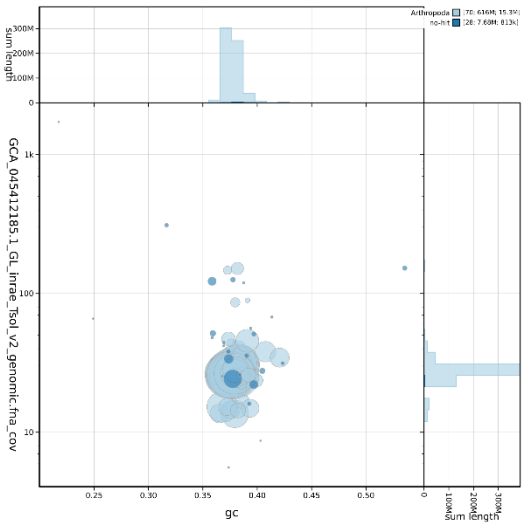
Cydalima perspectalis



Leptoglossus occidentalis



Tecia solanivora



<u>Figure C</u>: BlobTools GC-coverage plots for each of the three species (Laetsch and Blaxter 2017). Each circle represents a scaffold, with size proportional to sequence length. The x-axis indicates GC content, while the y-axis represents sequencing depth. Colors correspond to taxonomic assignments. Histograms display the distribution of sequence length sums along each axis. No significant contamination is detected in any of the assemblies.

Another concern I have is regarding the sequencing of a genome that is already available at the chromosome level. It would be beneficial for the authors to explain the motivation behind this duplication effort, particularly since this genome has the highest coverage in the study. On the other hand, the genome of Leptoglossus occidentalis would benefit from higher coverage to meet the minimal N50 contig standard of 1Mb, as the current coverage seems insufficient. And the genome of Tecia solanivora would also benefit from Hi-C data to get a reference-quality genome.

We understand the reviewer's concern regarding the sequencing of Cydalima perspectalis*, given that another chromosome-level genome is available. However, at the time we initiated our work, no such assembly was publicly accessible. The reference genome became available in public databases in June 2023, after we had already started sequencing and assembling. Despite this, we decided to proceed with our analysis due to the high quality of our data and the added value of sequencing an individual of the* fusca *morph, which has a distinctive dark phenotype. This unique aspect of our dataset, already noted in Table 5, is now better highlighted in the manuscript (Lines 62-66): "Although a genome assembly has been previously published for* C. perspectalis *(Broad et al. 2024), it was generated from an individual of the typica morph (light-colored). Here, we provide a new draft genome for this species based on an individual of the fusca morph (dark-colored), allowing for comparative genomic studies between these morphs."*

Regarding Leptoglossus occidentalis*, we fully acknowledge that higher sequencing depth would be beneficial. However, we would like to clarify that our assembly is of higher quality than initially suggested due to an error in the reported N50 value in the manuscript. The correct N50 after scaffolding is 147.7 Mb, not 0.55 Mb as previously stated in the text. This value was already correctly reported in Table 3 and on NCBI. With this correction, the assembly quality of* L. occidentalis *is in fact comparable to that of the other species, despite the lower sequencing depth. We have now updated the manuscript accordingly to reflect this correction (Lines 91-100): "For each of the three species, C. perspectalis, L. occidentalis and T. solanivora, draft de novo genomes were assembled. Hi-C sequencing enhanced scaffolding for C. perspectalis (public read set: ERR11217097) and L. occidentalis (read set from this study). N50 values indicate a high level of contiguity for all three assemblies, exceeding 15 Mb in each case. C. perspectalis had the least fragmented assembly, with a total length of 469.1 Mb, only 52 scaffolds and a high sequencing depth of 75X (Table 3). Despite its larger genome size (1.77 Gb) and lower sequencing depth (22.5X), L. occidentalis exhibits strong contiguity, as reflected by its high N50 value of 147.7 Mb (Table 3). Additional quality indicators, including BUSCO scores and Mercury QV metrics, further validated the overall high quality of all three assemblies (Table 3)."*

Finally, we completely agree with the reviewer's suggestion that Tecia solanivora *would benefit from Hi-C data to achieve a reference-quality genome. We indeed attempted Hi-C sequencing on two*

separate occasions but were unsuccessful, likely due to the freshness constraints of our available specimens. Given that T. solanivora *is restricted to Central and South America, as well as the Canary Islands, obtaining fresh individuals suitable for Hi-C experiments is particularly challenging. Nonetheless, we acknowledge the importance of this approach and hope that future efforts will allow for a chromosome-level assembly.*

In short, I believe the authors are not far from achieving genomes that meet current quality standards. In my opinion, they should make the effort to remove potential allelic duplications (using purgedup as well as during the manual curation step; Pretext is a useful tool for this, as it allows plotting coverage on the Hi-C map to easily detect remaining allelic duplications) and potential contaminants. Additionally, I think they should aim for higher HiFi coverage for Leptoglossus occidentalis and generate a Hi-C library for Tecia solanivora. Indeed, these genome assemblies will provide a solid foundation for future analyses, and having reference genomes at the standard quality will ensure that their work is used for decades to come.

Thank you for your comments. We are confident that the concerns regarding potential allelic duplications have been adequately addressed. As mentioned earlier, we performed purge_dups, and the BUSCO results and k-mer spectra do not show evidence of significant duplications. Regarding Leptoglossus occidentalis*, we corrected the initial BUSCO values in Table 3, which were previously misinterpreted as indicating high duplication rates. We apologize once again for this oversight.*

Regarding the HiFi depth for Leptoglossus occidentalis and the Hi-C for Tecia solanivora, we have already discussed the current limitations in our previous responses. Despite these limitations, we are confident that the assemblies represent a solid foundation for future work.

<u>Here are few other comments:</u>

- The assembly size of Cydalima perspectalis is stated as 469.1 Mb, which does not match the size of the assembly available on NCBI (500.4 Mb). Could the authors clarify which value is correct?

Thank you for bringing this to our attention and we apologize for the oversight. The correct assembly size for Cydalima perspectalis *is 500.4 Mb, as reported on NCBI. This discrepancy was due to an error in the table, which has now been corrected. We have also taken the opportunity to thoroughly check all values in all tables and have updated them accordingly.*

- "Merqucy" should be corrected to "Merqury" in Table 3.

This has been corrected (Table 3).

- The statement "N50 values indicate good assembly quality" is inaccurate. N50 is not a quality value; it only reflects the contiguity of the assembly. I suggest rephrasing this sentence.

The text has been changed (Lines 93-94): "N50 values indicate a high level of contiguity for all three assemblies, exceeding 15 Mb in each case."

- The sentence "52 scaffolds at 75X" requires clarification. Does this mean that several assemblies were performed with varying coverages? If not, the sentence should be rephrased for clarity.

The assembly was performed with a single sequencing depth. With thus rephrased the sentence (Lines 94-96): "C. perspectalis had the least fragmented assembly, with a total length of 469.1 Mb, only 52 scaffolds and a high sequencing depth of 75X (Table 3)."

- The sample of Tecia solanivora was collected in Colombia, but the article does not specify whether the necessary agreements for acquiring this sample were respected. This should be clearly stated.

Before obtaining Tecia solanivora *samples from Colombia, we contacted the competent national authority, the Ministry of Environment and Sustainable Development, on June 29, 2021, to inquire about the necessary authorizations. Their response, received on August 6, 2021, confirmed that no specific authorization was required due to the invasive status of the species in Colombia. They stated:*

"Thank you for contacting us regarding the prior consent procedures, the benefit-sharing terms applicable and the competent national authorities to request authorization and negotiate the benefit-sharing terms for accessing genetic resources of Tecia solanivora, we inform you that the provisions on access to genetic resources and their by-products in Colombia apply only for the obtention and use of DNA, RNA and / or metabolite molecules from native Colombian species for bioprospecting, commercial or industrial purposes. In this regard, given that Tecia solanivora is not native to Colombia and is native to Guatemala to access its genetic resources and / or its by-products, there are no legal obligations with Colombia, even if the samples of T. solanivora that are to be used have been obtained from the country."

Review #2 by Nicolas Parisot

In the manuscript entitled "Draft genome and transcriptomic sequence data of three invasive insect species", Lombaert et al. report on the generation of high-quality genomic and transcriptomic data for three invasive insect species: Cydalima perspectalis (box tree moth), Leptoglossus occidentalis (western conifer seed bug), and Tecia solanivora (Guatemalan tuber moth). The Authors used whole-genome sequencing, RNA-seq, and Hi-C scaffolding to produce critical resources studying the genetic mechanisms underpinning biological invasions and the development of pest management strategies.

Despite the fact that the resources provided in this work represent a valuable addition to the field of insect pests, I have however several minor suggestions to enhance biological context, clarity and reproducibility.

Even though this work represents a data paper, the manuscript would benefit from i) a more comprehensive presentation of the biological context and ii) a thorough discussion on data quality. For instance, the introduction could give a broader presentation of the biology of the three insects studied.

The introduction has been revised to provide a more detailed biological context for the three invasive insect species studied. It now includes additional information on their host plants, ecological impacts, and invasion histories, emphasizing their importance as invasive species and the relevance of generating genomic resources for these taxa (Lines 37-70).

We have also included a short statement highlighting the high quality of the assemblies (Lines 136-139).

The heterozygosity and repeat content of the genomes are intriguing. A deeper exploration and discussion of how these genomic features impact genome assembly and annotation quality would enhance the manuscript.

Genome size and repeat content are generally correlated, with larger genomes tending to have a higher proportion of repetitive elements. Our data follow this expected pattern, with L. occidentalis exhibiting both the largest genome and the highest repeat content. Heterozygosity levels, on the other hand, remain within the typical range observed in insects. While high repeat content and heterozygosity can complicate genome assembly by increasing fragmentation risk and making haplotype resolution more challenging, our assemblies remain of high quality, as indicated by their strong continuity metrics (e.g.,

N50 values exceeding 15 Mb in all cases) and additional quality assessments (BUSCO scores, Mercury QV metrics).

We have added a brief statement highlighting that our assemblies maintain high contiguity and completeness despite the challenges posed by heterozygosity and repeat content (Lines 136-139): "Overall, our assemblies show high contiguity and completeness, despite the presence of heterozygosity and repeat content (Table 1). The use of HiFi long reads, combined with Hi-C scaffolding where available, allowed us to mitigate these challenges and produce high-quality genomic resources."

The figures and tables, particularly Figure 1 (k-mer spectra), are informative but could benefit from more detailed legends to aid interpretation. For instance, explaining the significance of specific patterns observed in the k-mer spectra would be helpful. Table 3 provides comprehensive genome metrics, but a supplementary table comparing these data to other available insect genomes would contextualize the results.

We have revised the legend of Figure 1 to provide a clearer explanation of the k-mer spectra patterns and their significance. The updated legend now describes the axes and the interpretation of the main peaks (Lines 113-117).

Thank you for your suggestion regarding a comparative table with other insect genomes. While we understand the interest in providing broader context, we believe that including such a table in the manuscript would not be entirely relevant. Indeed, the choice of species to compare is subjective, and genome assembly metrics are highly dependent on factors such as genome size and sequencing strategies. Moreover, sequencing technologies and assembly methods evolve rapidly, meaning that current comparisons may become outdated rather quickly.

However, we would like to emphasize that our assemblies meet established quality standards. In particular, our metrics align well with the criteria proposed by Rhie et al. (2021) and the Earth BioGenome Project report on assembly standards (September 2024; https://www.earthbiogenome.org/report-on-assembly-standards). For example, a Complete BUSCO score above 90% is considered a marker of high-quality assembly, which our species exceed. Similarly, an N50 above 10 Mb is often used as a benchmark, and our assemblies also meet this criterion.

To further illustrate this point, we provide three tables below (Tables B, C and D) which show key assembly quality metrics for our species alongside three relatively close species with publicly available reference assemblies from NCBI. We hope that this additional information will reassure you regarding the quality of the assemblies.

	Cydalima perspectalis	Cydalima perspectalis *previsous genome*	Pyrausta nigrata	Ostrinia nubilalis
Total-length	*500.44*	*483.69*	*538.9*	*495.5*
No. of scaffold	*46*	*200*	*34*	*52*
N50 scaffold length (Mb)	*17.46*	*16.86*	*19*	*16.46*
L50 scaffold count	*13*	*14*	*13*	*14*
Final GC%	*37.2*	*37*	*37.5*	*37.5*
Mean depth	*75*	*57*	*45*	*44*
Busco complete	*99.7*	*94.9*	*98.8*	*98.5*
Busco single	*99.5*	*94.7*	*98.7*	*98.2*
Busco duplicates	*0.2*	*0.2*	*0.1*	*0.3*
Busco fragmented	*0.1*	*0.9*	*0.4*	*0.5*
Busco missing	*0.2*	*4.2*	*0.8*	*1*

Table B: Main quality assembly metrics of Cydalima perspectalis *and three related species.*

	Leptoglossus occidentalis	Leptoglossus phyllopus	Gonocerus acuteangulatus	Nezara viridula
Total-length	*1745.64*	*1665.57*	*1106.2*	*1185.13*
No. of scaffold	*211*	*141*	*173*	*84*
N50 scaffold length (Mb)	*147.7*	*155.07*	*121.35*	*181.51*
L50 scaffold count	*5*	*4*	*4*	*3*
Final GC%	*35.52*	*35.5*	*34.5*	*32*
Mean depth	*22.5*	*84*	*28*	*100*
Busco complete	*98.9*	*99.6*	*99.2*	*97.7*
Busco single	*96.6*	*97.6*	*97.6*	*70.6*
Busco duplicates	*2.3*	*2*	*1.6*	*27.1*
Busco fragmented	*0.4*	*0.1*	*0.4*	*0.9*
Busco missing	*0.7*	*0.3*	*0.4*	*1.4*

Table C: Main quality assembly metrics of Leptoglossus occidentalis *and three related species.*

	Tecia solanivora	Anarsia innoxiella	Scrobipalpa costella	Carpatolechia fugitivella
Total-length	*623.3 (scaffold)*	*302.93*	*603.18*	*493.08*
No. of scaffold	*NA*	*32*	*46*	*142*
N50 scaffold length (Mb)	*NA*	*10.42*	*22.16*	*17.19*
L50 scaffold count	*NA*	*13*	*12*	*13*
Final GC%	*37.81*	*36*	*38.5*	*37*
Mean depth	*23.4*	*78*	*39*	*41*
Busco complete	*99.5*	*98*	*98.2*	*97.9*
Busco single	*98.8*	*97.4*	*97.4*	*97.1*
Busco duplicates	*0.7*	*0.5*	*0.8*	*0.8*
Busco fragmented	*0.4*	*0.5*	*0.5*	*0.5*
Busco missing	*0.1*	*1.6*	*1.3*	*1.6*

Table D: Main quality assembly metrics of Tecia solanivora *and three related species.*

There are occasional typographical and grammatical errors. A thorough proofreading is recommended.

We have thoroughly proofread and corrected the manuscript to address any typographical and grammatical errors.

References

Broad GR, Boyes D, Poloni R (2024) The genome sequence of the Box-tree Moth , *Cydalima perspectalis* (Walker , 1859) Darwin Tree of Life Barcoding collective , Wellcome Sanger Institute Tree of Life Management , Samples and Laboratory Wellcome Sanger Institute Scientific Operations. Wellcome Open Res 9:272. https://doi.org/10.12688/wellcomeopenres.21678.1

Brown MR, Gonzalez P, Rosa D La, Blaxter M (2025) tidk : a toolkit to rapidly identify telomeric repeats from genomic datasets. Bioinformatics btaf049. https://doi.org/10.1093/bioinformatics/btaf049

Cabanettes F, Klopp C (2018) D-GENIES: Dot plot large genomes in an interactive, efficient and simple way. PeerJ 2018:. https://doi.org/10.7717/peerj.4958

Kim D, Paggi JM, Park C, et al (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol 37:907–915. https://doi.org/10.1038/s41587-019-0201-4

Laetsch DR, Blaxter ML (2017) BlobTools: Interrogation of genome assemblies. F1000Research 6:1287. https://doi.org/10.12688/f1000research.12232.1

Liao Y, Smyth GK, Shi W (2014) FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30:923–930. https://doi.org/10.1093/bioinformatics/btt656

Rhie A, McCarthy SA, Fedrigo O, et al (2021) Towards complete and error-free genome assemblies of all vertebrate species. Nature 592:737–746. https://doi.org/10.1038/s41586-021-03451-0

Wood DE, Lu J, Langmead B (2019) Improved metagenomic analysis with Kraken 2. Genome Biol 20:1–13. https://doi.org/10.1186/s13059-019-1891-0