Peer Community In Genomics

Three more reference genomes of invasive insect species

Vincent Lacroix *based on peer reviews by* ***Jean-Marc Aury*** *and* ***Nicolas Parisot***

Eric Lombaert, Christophe Klopp, Aurélie Blin, Gwenolah Annonay, Carole Iampietro, Jérôme Lluch, Marine Sallaberry, Sophie Valière, Riccardo Poloni, Mathieu Joron, Emeline Deleury (2025) Draft genome and transcriptomic sequence data of three invasive insect species. bioRxiv, ver. 2, peer-reviewed and recommended by Peer Community in Genomics. https://doi.org/10.1101/2024.12.02.626401

Submitted: 09 December 2024, Recommended: 20 May 2025

Cite this recommendation as:
Lacroix, V. (2025) Three more reference genomes of invasive insect species. *Peer Community in Genomics*, 100425. https://doi.org/10.24072/pci.genomics.100425

Published: 20 May 2025

The number and prevalence of invasive species have risen in the last decades together with international trade (Hulme 2021). As invasive species, insects have received less attention than plants. In particular, there are fewer reference genomes currently available, although genomic resources can be useful to investigate invasion dynamics and develop more effective management strategies.

Lombaeart et al. (2025) sequenced and assembled the genomes of three species: *Cydalima perspectalis* (the box tree moth), *Leptoglossus occidentalis* (the western conifer seed bug), and *Tecia solanivora* (the Guatemalan tuber moth), which have in common their rapid spread and severe impact on their respective host plants (boxwoods, conifers and potatoes). The authors generated PacBio HiFi reads, together with Hi-C data to obtain assemblies meeting international quality standards. They also generated short-read RNA-seq data, which they used to provide initial structural annotations of the genes. The resulting reference genomes are still in a draft state because repeats and heterozygosity are notoriously hard to handle. The most challenging genome to assemble was *L. occidentalis*, with an estimated size of 1.5 Gb, an estimated repeat content of 58%, and an estimated heterozygosity of 1.8%. The raw data produced can still be analysed more in depth to characterise further the repeat content and the heterozygosity of these species.

These reference genomes can readily be used for identifying genetic markers of interest for a variety of applications. In a general context where there is a growing awareness that data production is associated with a significant part of the carbon footprint of research (De Paepe et al. 2024), this dataset has high chances to be extensively reused and analysed by the community.

References:

De Paepe M, Jeanneau L, Mariette J, Aumont O, Estevez-Torres A (2024) Purchases dominate the carbon footprint of research laboratories. PLOS Sustainability and Transformation, 3, e0000116. https://doi.org/10.1371/journal.pstr.0000116

Hulme, P E (2021) Unwelcome exchange: international trade as a direct and indirect driver of biological invasions worldwide. One Earth 4, 666–679. https://doi.org/10.1016/j.oneear.2021.04.015

Lombaert E, Klopp C, Blin A, Annonay G, Iampietro C, Lluch J, Sallaberry M, Valière S, Poloni R, Joron M, Deleury E (2025) Draft genome and transcriptomic sequence data of three invasive insect species. bioRxiv, ver. 2 peer-reviewed and recommended by PCI Genomics. https://doi.org/10.1101/2024.12.02.626401

Reviews

Evaluation round #2

Reviewed by Jean-Marc Aury, 10 April 2025

I have read the revised manuscript and reviewed the authors' responses. They have addressed all of my concerns.

However, I would like a minor clarification regarding the assembly contiguity of L. occidentalis. The contig N50 is below 1 Mb (EBP standard), and obviously, the scaffold N50 is higher. You mentioned, "The correct N50 after scaffolding is 147.7 Mb, not 0.55 Mb as previously stated in the text." However, I am referring to the contig N50, which remains at 0.55 Mb.

I suggest that the authors be more transparent and state for example: "N50 values indicate a high level of contiguity for all three assemblies, exceeding 15 Mb at the contig level for C. perspectalis and T. solanivora, and less than 1 Mb for L. occidentalis. However, Hi-C data allows for the generation of a chromosome-scale assembly with a scaffold N50 of 147.7 Mb."

Reviewed by Nicolas Parisot, 24 March 2025

The authors have responded adequately to my suggestions.

Evaluation round #1

DOI or URL of the preprint: https://doi.org/10.1101/2024.12.02.626401
Version of the preprint: 1

Authors' reply, 18 March 2025

Download author's reply

Decision by Vincent Lacroix, posted 17 January 2025, validated 18 January 2025

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 adress 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.

With my best wishes,

Vincent Lacroix

Reviewed by Jean-Marc Aury, 10 January 2025

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.

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.

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 *Teci 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.

Here are few other comments:

- 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?
- "Merqucy" should be corrected to "Merqury" in 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 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 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.

Reviewed by Nicolas Parisot, 15 January 2025

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 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.

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.

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

- **Title and abstract**
  - Does the title clearly reflect the content of the article? **Yes**
  - Does the abstract present the main findings of the study? **Yes**
- **Introduction**
  - Are the research questions/hypotheses/predictions clearly presented? **Yes**
  - Does the introduction build on relevant research in the field? **Yes, but the manuscript would benefit from a more comprehensive presentation of the biological context.**

- **Materials and methods**
  - Are the methods and analyses sufficiently detailed to allow replication by other researchers? **Yes**
  - Are the methods and statistical analyses appropriate and well described? **Yes**

- **Results**
  - In the case of negative results, is there a statistical power analysis (or an adequate Bayesian analysis or equivalence testing)? **Not applicable**
  - Are the results described and interpreted correctly? **Yes**

- **Discussion**
  - Have the authors appropriately emphasized the strengths and limitations of their study/theory/methods/argument? **No, the manuscript would benefit from a more comprehensive discussion on data quality.**
  - Are the conclusions adequately supported by the results (without overstating the implications of the findings)? **Yes**