RESEARCH ARTICLE

REVISED Chromosome-scale assembly of the yellow mealworm genome

[version 3; peer review: 2 approved]

A high-quality reference genome for *Tenebrio molitor* breeding and sustainable production

Evangelia Eleftheriou[1], Jean-Marc Aury[1], Benoît Vacherie[2], Benjamin Istace[1], Caroline Belser[1], Benjamin Noel[1], Yannick Moret[3], Thierry Rigaud[3], Fabrice Berro[4], Sona Gasparian[4], Karine Labadie-Bretheau[2], Thomas Lefebvre[4], Mohammed-Amin Madoui[1,3,5]

[1]Génomique Métabolique, Genoscope, Institut François Jacob, Commissariat à l'Energie Atomique (CEA), CNRS, Univ Evry, Université Paris-Saclay, Université Paris-Saclay, Evry, 91057, France
[2]Genoscope, Institut de biologie François Jacob, CEA, Université Paris-Saclay, Evry, 91057, France
[3]Équipe Écologie Évolutive, UMR CNRS 6282 BioGéoSciences, Université de Bourgogne Franche-Comté, Dijon, 21000, France
[4]Ynsect, Evry, 91000, France
[5]Service d'Etude des Prions et des Infections Atypiques (SEPIA), Institut François Jacob, Commissariat à l'Energie Atomique et aux Energies Alternatives (CEA), Université Paris Saclay, Fontenay-aux-Roses, France

v3 **First published:** 17 Aug 2021, **1**:94
https://doi.org/10.12688/openreseurope.13987.1
Second version: 25 Feb 2022, **1**:94
https://doi.org/10.12688/openreseurope.13987.2
Latest published: 05 Sep 2022, **1**:94
https://doi.org/10.12688/openreseurope.13987.3

Abstract

Background: The yellow mealworm beetle, *Tenebrio molitor*, is a promising alternative protein source for animal and human nutrition and its farming involves relatively low environmental costs. For these reasons, its industrial scale production started this century. However, to optimize and breed sustainable new *T. molitor* lines, the access to its genome remains essential.
Methods: By combining Oxford Nanopore and Illumina Hi-C data, we constructed a high-quality chromosome-scale assembly of *T. molitor*. Then, we combined RNA-seq data and available coleoptera proteomes for gene prediction with GMOVE.
Results: We produced a high-quality genome with a N50 = 21.9Mb with a completeness of 99.5% and predicted 21,435 genes with a median size of 1,780 bp. Gene orthology between *T. molitor* and *Tribolium castaneum* showed a highly conserved synteny between the two coleoptera and paralogs search revealed an expansion of histones in the *T. molitor* genome.
Conclusions: The present genome will greatly help fundamental and applied research such as genetic breeding and will contribute to the sustainable production of the yellow mealworm.

Open Peer Review

Approval Status ✓ ✓

	1	2
version 3 (revision) 05 Sep 2022		✓ view
version 2 (revision) 25 Feb 2022	✓ view	? view
version 1 17 Aug 2021	? view	

1. **Barbara Feldmeyer**, Senckenberg Biodiversity and Climate Research Centre, Frankfurt am Main, Germany

2. **Gregor Bucher**, University of Göttingen, Göttingen, Germany

Any reports and responses or comments on the

Keywords
Yellow Mealworm, Tenebrio molitor, genomics, chromosome-scale assembly

article can be found at the end of the article.

This article is included in the Horizon 2020 gateway.

This article is included in the Genetics and Genomics gateway.

Corresponding author: Mohammed-Amin Madoui (amadoui@genoscope.cns.fr)

Author roles: Eleftheriou E: Data Curation, Formal Analysis, Investigation, Methodology, Validation; **Aury JM**: Methodology, Resources, Software, Supervision; **Vacherie B**: Formal Analysis, Investigation; **Istace B**: Methodology, Resources, Software; **Belser C**: Methodology, Resources, Software; **Noel B**: Methodology, Resources, Software; **Moret Y**: Resources; **Rigaud T**: Resources; **Berro F**: Funding Acquisition; **Gasparian S**: Funding Acquisition; **Labadie-Bretheau K**: Formal Analysis, Investigation, Methodology, Supervision; **Lefebvre T**: Funding Acquisition, Project Administration; **Madoui MA**: Conceptualization, Funding Acquisition, Project Administration, Supervision, Validation, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: Fabrice Berro, Sona Gasparian and Thomas Lefebvre work for Ynsect, a private company breeding yellow mealworms. Ynsect is also the coordinator of the H2020 FARMYNG project that supports this study. No competing interests were disclosed for the other co-others.

Grant information: This research was financially supported by the European Union's Horizon 2020 research and innovation programme under the grant agreement No 837750 (project FARMYNG). This study is the first deliverable of the work package 5 devoted to yellow mealworm breeding strategy.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Eleftheriou E, Aury JM, Vacherie B *et al.* **Chromosome-scale assembly of the yellow mealworm genome [version 3; peer review: 2 approved]** Open Research Europe 2022, **1**:94 https://doi.org/10.12688/openreseurope.13987.3

First published: 17 Aug 2021, **1**:94 https://doi.org/10.12688/openreseurope.13987.1

REVISED **Amendments from Version 2**

In the current version of the manuscript, we took into account all issues raised by the second reviewer. More specifically, we included a novel analysis of Tenebrio-specific duplicated genes and tried to provide some functional classification of the most expended gene families. This presented now on the Table 3. Also, we described the location of these duplicated genes to better support their presence and to avoid misinterpretation of the genome annotation that could be due to genome assembly problems. We also added a new paragraph that tends to explain the difference of genome size between *Tenebrio molitor* and *Tribolium castaneum*. Furthermore, we also corrected some minor typos.

Any further responses from the reviewers can be found at the end of the article

Plain language summary

We provide the genome sequence of the yellow mealworm, *Tenebrio molitor*, by combining high-throughput sequencing technologies to obtain a genome assembly that well represents the 10 mealworm chromosomes. We also identified the *Tenebrio molitor* gene set and compared its organisation to that of the red flour beetle. This new genomic resource will help breeders to develop new mealworm lines to face future global human nutrition problems by providing protein-rich and ecologically friendly mealworm production systems.

Introduction

The global human population is estimated to reach approximately nine billion people by 2050, thus the demand for animal protein is expected to increase by 76%[1]. Such an increase questions the sustainability of our conventional food and feed production systems. At the same time, we also need to reduce the impact of agriculture on our environment[2]. Today, insect production is considered a sustainable alternative for food and feed production for several reasons. First, the suitable nutritional composition of edible insects[3] and second, the relatively low environmental impact its production involves compared to other conventional livestock production systems[4,5].

In this context, the yellow mealworm beetle *Tenebrio molitor* has been described as a promising alternative protein source for animal and even human nutrition[6]. For these reasons several companies have pioneered the production of *T. molitor* at industrial scale. However, despite being promising for sustainable food security, mass production of *T. molitor* remains relatively primitive and challenging[7].

The genetic improvement of *T. molitor* is one of these challenges. Indeed, several quantitative traits of industrial importance such as growth rate, fertility, protein rate or susceptibility to pathogens need to be mapped to allow the development of molecular-based breeding programs to speed up the development of new lines with improved agronomic traits. However, suitable genomic resources on *T. molitor* are needed to accelerate such genetic programs.

Previous efforts to produce *T. molitor* transcriptomes and more recently the draft genome using 10X genomics technology have been published[8]. While this latter technology was promising on diploid and heterozygous insects[9–12], its application to *T. molitor* produced a fragmented assembly with a 90% of BUSCO completeness and no genome annotation. This particular effort motivated the development of a new genome assembly that would allow deeper genomic analyses such as quantitative trait locus mapping or genomic estimated breeding values analysis.

Here, we present a *T. molitor* genome assembly based on the combination of long, short reads and Hi-C data. The genome assembly and annotation quality are analysed and a comparison to the red flour beetle (*Tribolium castaneum*) genome is described to show how the current genome can be an asset for academic research and breeding.

Methods

Biological material and insect rearing

Tenebrio molitor samples were provided by Ynsect and bred at CEA-Genoscope (Evry, France). The individuals were fed with bran and apple and kept at room temperature and humidity. For the genome sequencing, male pupae which possess XY chromosomes were selected, starved for three days and used for DNA extraction. For mRNA extraction, embryos, larva, pupae, adult males and females were isolated without specific diet. Embryos were collected within a week after egg-laying.

DNA extraction

Genomic DNA (gDNA) was extracted from a single pupa male to generate both Illumina PCR-free, PromethION and Dovetail Hi-C libraries. In order to generate long reads on the Oxford Nanopore Technologies devices, high-quality and high-molecular-weight (HMW) DNA was needed. For this purpose, DNA was isolated following the protocol provided by Oxford Nanopore Technologies, Oxford, UK (ONT), "High molecular weight gDNA extraction from plant leaves" provided by the ONT Community in March, 2019 (CTAB-Genomic-tip). This protocol involves a conventional CTAB extraction followed by purification using commercial Qiagen Genomic tips (QIAGEN, MD, USA). DNA fragment size selection was performed using the Short Read Eliminator (Circulomics, MD, USA) instead of AMPpure XP beads. A single pupa male weighing 170mg was cryoground in liquid nitrogen. The fine powder was divided in one-third for the Hi-C library and two-thirds for both Illumina PCR-free and PromethION libraries. The two-thirds of the powder was transferred to a lysis Carlson buffer supplemented with RNase A. After 1h-incubation, proteins were removed with chloroform extraction and DNA was precipitated with isopropanol and centrifugation. The pellet was then purified using the Qiagen Genomic tip 100/G, following the manufacturer's instructions. DNA was quantified by a dsDNA-specific fluorimetric quantitation method using Qubit dsDNA HS Assays (Catalog #Q32851, ThermoFisher Scientific, Waltham, MA). HMW gDNA quality was checked on a 2200 TapeStation

automated electrophoresis system (Agilent, CA, USA) and the length of the DNA molecules was estimated to be over 60Kb.

PromethION library preparation and sequencing

HMW gDNA was size-selected using the Short Read Eliminator kit (SKU SS-100-101-01, Circulomics, MD, USA). The ONT library was prepared with the Oxford Nanopore SQK-LSK109 kit, according to the following protocol. Genomic DNA fragments (3μg) were repaired and 3'-adenylated with the NEBNext FFPE DNA Repair Mix (Catalog#M6630, New England Biolabs, Ipswich, MA, USA) and the NEBNext® Ultra™ II End Repair/dA-Tailing Module (Catalog#E7546, NEB). Sequencing adapters provided by ONT were ligated using the NEBNext Quick Ligation Module (Catalog#E6056, NEB). After purification with AMPure XP beads (Beckmann Coulter, Brea, CA, USA), half of the library was mixed with the Sequencing Buffer (ONT) and the Loading Bead (ONT) and loaded on a PromethION R9.4.1 flow cell. The second half of the library was loaded on the flow cell after a Nuclease Flush using the Flow Cell Wash Kit (Catalog#EXPWSH003, ONT) according to the ONT protocol. After 48h of the sequencing run, a second Nuclease Flush was performed and a third library was loaded on the flow cell. Nucleotide bases were called using Guppy version 4.0.1[13] and the raw reads were used for genome assembly.

Illumina PCR-free library preparation and sequencing

The PCR-free library was prepared using the Kapa Hyper Prep Kit (Catalog#KK8505, KapaBiosystems, Wilmington, MA, USA), following the manufacturer's recommendations. Briefly, qDNA (1.5μg) was sonicated to a 100–1,500-bp size range using a Covaris E220 sonicator (Covaris, Woburn, MA, USA). The fragments were end-repaired, then 3'-adenylated and Illumina adapters were added. The ligation products were purified with AMPure XP beads (Beckmann Coulter Genomics, Danvers, MA, USA). The library was quantified by qPCR using the KAPA Library Quantification Kit for Illumina Libraries (Catalog#07960140001, KapaBiosystems), and the library profiles were assessed on an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). The libraries were sequenced on an Illumina HiSeq4000 instrument (Illumina, San Diego, CA, USA) using 150 bp read chemistry in paired-end mode. After the Illumina sequencing, an in-house quality control process was applied to the reads that passed the Illumina quality filters, as described by Alberti and colleagues[14].

Dovetail Hi-C library preparation and sequencing

Another third of the cryoground powder (from the DNA extraction section) was used to generate a Hi-C library using the Dovetail Hi-C preparation kit (Dovetail Genomics, Scotts Valley, CA, USA), according to the manufacturer's protocol (manual version 1.03). After the cross-linking of animal tissues, the chromatin was normalized and then immobilized on capture beads before enzyme restriction digestion. The digested DNA ends were marked with biotin and ligated to create chimeric molecules. After reversal cross-linking, DNA was purified and then followed by library generation. The Dovetail Hi-C library quality was checked as described above and sequenced on an Illumina HiSeq4000 instrument (Illumina, San Diego, CA, USA) using 150 base-length read chemistry in paired-end mode.

RNA extraction

Embryos, larva, pupae, adult males and females were collected for later mRNA extraction. Tissue samples were mechanically homogenized using ZR Bashing Bead Lysis tube (ZymoResearch, CA, USA) with the FastPrep-24™ 5G Instrument (MP Biomedicals, Santa Ana, CA, USA). Nucleic acids were then extracted from homogenized suspension using the ZR-Duet DNA/RNA MiniPrep Plus kit (Catalog # D7003, ZymoResearch, CA, USA). Extracted RNA was quantified with RNA-specific fluorometric quantitation on a Qubit 2.0 Fluorometer using Qubit RNA HS Assay (Thermo Fisher Scientific, Waltham, MA, USA). Integrity of total RNA was assessed on an Agilent Bioanalyzer, using the RNA 6,000 Pico LabChip kit (Catalog # 5067-1513, Agilent Technologies, Santa Clara, CA).

RNA library preparation and sequencing

RNA-seq library preparations were carried out from 500ng total RNA using the TruSeq Stranded mRNA kit (Catalog #20020595, Illumina, San Diego, CA, USA), which allows mRNA strand orientation, i.e sequence reads occur in antisense orientation only. Poly(A)+ RNA was selected with oligo(dT) beads, chemically fragmented and converted into single-stranded cDNA using random hexamer priming. Then, the second strand was generated to create double-stranded cDNA. cDNA was then 3'-adenylated, and Illumina adapters were added. Ligation products were PCR-amplified. Ready-to-sequence Illumina libraries were then quantified by qPCR using the KAPA Library Quantification Kit for Illumina Libraries (Catalog #KK4824, KapaBiosystems, Wilmington, MA, USA), and library profiles evaluated with an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). Each library was sequenced using 151bp paired end reads chemistry on a NovaSeq 6000 Illumina sequencer.

Genome assembly

The *T. molitor* genome size was estimated using GenomeScope[15] (GenomeScope, RRID:SCR_017014) v1 with Illumina reads (Table S1, *Extended data*) and a *k*-mer value of 31. We applied YACRD[16] (version 0.6.0) to the raw nanopore reads (Table S1, *Extended data*) to detect potential chimeras. Both "all-vs-all alignment" and "yacrd scrubbing" steps were performed with the recommended parameters and removed 109,066 chimeric reads. The 2,372,861 non-chimeric reads were corrected using NECAT[17] with parameters GENOME_SIZE, PREP_OUTPUT_COVERAGE and CNS_OUTPUT_COVERAGE set to 310,000,000, 60 and 40, respectively, to first correct the longest 60x reads and afterwards, the longest 40x corrected reads were extracted to assemble 250,277 reads.

Because nanopore reads contain systematic errors in homopolymeric regions, the output assembly was polished three times using Racon[18] (Racon, RRID:SCR_017642) with default

parameters with the nanopore reads and two times Hapo-G[19] with the Illumina reads. The assembly was merged into a single haplotype genome assembly using HaploMerger2[20] (Figure S2, *Extended data*) and polished using two rounds of Hapo-G with the Illumina reads.

To increase the contiguity of the assembly to a chromosome-scale level (Table S1, *Extended data*), we aligned Hi-C paired-end reads to the polished haploid assembly with *bwa – mem*[21] (BWA, RRID:SCR_010910). Because Hi-C captures conformation via proximity-ligated fragments, paired-end reads are first mapped independently (as single-end reads) and subsequently paired in a later step. Hi-C reads and alignments contain experimental artifacts so the alignments need some additional processing. We use alignment filtering method using Arima Genomics pipeline (https://github.com/ArimaGenomics/mapping_pipeline) and applied the script "filter l" to each bam file (Read1 and Read2) and afterwards paired the filtered single-end Hi-C reads using "two_read_bam_combiner.pl". Then, with Picard tools (https://broadinstitute.github.io/picard/), we added read groups to the combined BAM file (with the command AddOrReplaceReadGroups) and discarded any PCR duplicates present in the paired-end BAM file (with the command MarkDuplicates). We scaffolded the assembly with the Hi-C data using SALSA2[22] and obtained 138 scaffolds. Only scaffolds larger than 35kb were kept resulting in a final assembly of 112 scaffolds (Table S3 and Figure S10, *Extended data*). The largest scaffolds were manually checked for missassembly using the sequencing information and the synteny with the ten *Tribolium castaneum* chromosomes (cf. comparative genomics section).

Transcriptome assembly

RNA-seq reads from six transcriptomes derived from different developmental stages (embryos, larvae, pupae, and adults), sexes (females and males) and public data of two RNA-seq samples generated from pooled bacterial infected *T. molitor* (PRJNA646689, *Underlying data*) were assembled using Velvet[23] (Velvet, RRID:SCR_010755) version 1.2.07 and Oases[24] (Oases, RRID:SCR_011896) version 0.2.08 with *k*-mer size set to 81 and 63 for the in-house and public RNA-seq reads, respectively (Table S2, *Extended data*). The first five bases of contigs 5' and 3' ends were removed. The sequences were masked for low-complexity using DustMasker (version 1.0.0 from the BLAST 2.10.0 package) and only contigs larger than 150bp with more than 75% of unmasked bases were kept. To address the problem of merged chimeric contigs, a post-processing of Oases contigs has been done. Assembly tools often erroneously merge sequences into one single contig and Oases is prone to this behaviour. To address this problem, we used an in-house script that splits chimeric contigs. Splitting a contig into regions where different ORFs appear, or regions where abrupt shifts in read coverage occur, could streamline the gene-prediction process. Based on combined resources such as the pileup-coverage, the research of ORFs (TransDecoder https://github.com/TransDecoder/TransDecoder/releases) and domains, this tool aims to split contigs sequences with different functional sites form different contigs. Reads were mapped to the contigs with BWA-mem and the consistent paired-end reads were selected. Chimeric contigs were identified and split (uncovered regions) based on coverage information from consistent paired-end reads. Moreover, open reading frames (ORF) and domains were searched using respectively TransDecoder and CDDsearch (Conserved Domain Database, RRID:SCR_002077). We only allowed breaks outside ORF and domains. Finally, the read strand information was used to correctly orient the RNA-seq contigs.

Genome annotation

Repeated sequence masking. Low complexity regions of the assembly were masked with the DustMasker[25] algorithms (version 1.0.0 from the BLAST 2.10.0 package). Transposable elements (TEs) and other repeats were annotated and masked using RepeatMasker[26] (RepeatMasker, RRID:SCR_012954) version open-4.0.5 with rmblastn[26] version 2.10.0+. The assembly was compared to classified sequences of the RepeatMasker complete database 20150807. We set the custom library RepeatMasker.lib of version 4.0.5 to the -lib parameter[27].

Transcriptome and proteome alignments. mRNA contigs from the eight samples were aligned to the assembly in a two-step strategy. First, BLAT[28] (BLAT, RRID:SCR_011919) (version 36 with default parameters) was used for fast localizing genomic regions and the best match of each contig was kept. A second local alignment was performed with Est2Genome[29] (version 5.2 with default parameters). Aligned contigs with overlap higher than 80% and more than 95% identity were retained. Additionally, proteomes of four other Coleoptera (*T. castaneum* (Herndon *et al.*, 2020), *Ontophagus taurus*, *Asbolus verrucosus*, *Dendroctonus ponderosae*) and *T. molitor* proteins from UniProt[30] database were aligned to the genome in a two-step strategy. First, using BLAT (version 36 with default parameter) matches with score higher than 90% of the best match score were retained. Second, alignments were refined using Genewise[31] (version 2.2.0 default parameters) and proteins with more than 50% of their length aligned onto the assembly were kept.

Gene predictions. To identify the gene structure, the transcriptomic and protein alignments were combined using Gmove[32] (Gmove, RRID:SCR_019132) (Note S2, *Extended data*). Protein alignments from the five coleoptera were merged into a single file and provided to Gmove (–prot parameter). We also set transcriptomic alignments from eight different samples (Table S2, *Extended data*) to the –rna parameter and activated the –score option to keep the gene model with the highest score. Based on *T. castaneum* gene features, we set the maximal size of intron and minimal size of exons to 150,000bp and 3bp using the -m and -e parameters, respectively. To prevent false positive gene predictions due to a large number of single-exon transcripts, sample-specific single-exon transcripts were removed before running Gmove.

Several criteria were applied sequentially to filter the gene predictions. We used HMMER[33] (Hmmer, RRID:SCR_005305) (version 3.2.1, June 2018) to find pfam domains, DIAMOND[34]

(DIAMOND, RRID:SCR_016071) version 0.9.24 for protein searches against ncbi-nr database, RepeatModeler[35] (RepeatModeler, RRID:SCR_015027) version 2.0.1 for *ab initio* repeats screening and TransposonPSI[36] to compare predicted models with transposable elements. Gmove initially predicted 24,870 genes, 27% of which were intronless. While we are more confident in multi-exon gene predictions, we were cautious with intronless genes corresponding potentially to transposable elements or false positive predictions caused by the fragmented alignments of transcripts. To solve this problem, we launched HMMER (version 3.2.1 with e-value set to 10e-5) for detecting known pfam domains and a DIAMOND analysis against ncbi-nr database for protein hits (version 0.9.24 with – evalue 10e-5, –unal 0). Furthermore, Repeat-Modeler version 2.0.1 was used for screening *ab initio* repeats in the *T. molitor* assembly and 46.77% of the genome was masked. Then, we focused on overlaps between the predicted genes and repeats, using commands from BEDtools. Genes with exons highly covered by repeats (>90%) were automatically classified as repeats. In parallel, we used transposonPSI.pl to align the virtual cDNA proteins of the 24,870 predictions against the TransposonPSI_08222010 library. We selected the single best transposonPSI match for each protein (from file proteins.fasta.TPSI.topHits) and tagged the corresponding genes as transposable elements. At this point, we excluded genes (single and multi-exon) that were either highly covered by repeats (RepeatModeler) or TE tagged (TransposonPSI) without any blastp/pfam hit. We also excluded intronless genes that were predicted only by RNA-seq evidence (not any Coleoptera protein overlap) and at the same time composed of >80% untranslated regions (ratio UTR/(UTR+CDS)) without any pfam/blastp hit.

Additionally, we searched for overlaps between predicted intronless genes and CDS of protein or mRNA evidence. Then, we discarded any intronless gene accomplishing none of the following conditions: (i) A gene that is predicted from at least one mRNA and one protein evidence. (ii) A gene that is predicted from mRNA transcripts of at least two different samples and (iii) A gene that is predicted from at least a *T. molitor* protein (from Uniprot). If none of the above criteria was met and a gene did not have any pfam/blastp hit either, then it was removed. After this filtering process the different annotation supports were combined to obtain a final set of 21,435 gene predictions (see Figure S11, *Extended data* for the genome annotation workflow).

Comparative genomics

Homology search between the 21,435 *T. molitor* predicted genes and the 22,610 *T. castaneum* protein isoforms was performed. We used blastp[37] (NCBI BLAST, RRID:SCR_004870) v.2.10.0+ with a maximum e-value set to 1e-10 and found 10,495 reciprocal best hits between the two species. Using NUCmer from the MUMmer4.0beta[38] (MUMmer, RRID:SCR_018171) package, we plotted the alignments between the 16 longest *T. molitor* scaffolds and the 10 *T. castaneum* chromosomes. To observe the synteny between the two beetle genomes, we combined the associations inferred from the MUMmer plot with the localization of the orthologous genes and constructed a Circos plot[39] (Circos, RRID:SCR_011798). Finally, 9,760 reciprocal best matches out of the total best hits (10,495) corresponded to orthologous genes between the 16 *T. molitor* scaffolds and the 10 *T. castaneum* chromosomes.

Results and discussion

Tenebrio molitor chromosome-scale genome assembly

The *T. molitor* genome size was estimated around 310 Mb with a heterozygosity rate of 1.43% (Figure S1, *Extended data*). By combining long, short reads and Hi-C data, we obtained a final genome assembly of 287.9 Mb (Table 1) representing a single haplotype of the *T. molitor* diploid genome (2n=20)[40] with a BUSCO[41] (BUSCO, RRID:SCR_015008) completeness of 99.5% (using version 5.0.0 with Insecta database odb10) (Figure S1, *Extended data*). The assembly presents a N50 of 21.9Mb, which is higher than *T. castaneum*'s one[42] and much higher than the previously published *T. molitor* genome N50 (24.1kb). In our assembly, the largest 16 scaffolds represent 90% of the total assembly, leading to a chromosome-scale assembly which provides high-quality support for gene annotation.

Nearly 6% of the assembly was masked for repeated elements with a majority of simple DNA repeats (49,992) and transposons (29,182). The next most abundant repeats were long interspersed nuclear elements (11,417) followed by long terminal repeats (7,950). Overall, these four types of repeats account for 5.31% of the masked genome assembly (Table S4, *Extended data*).

Several studies pointed out the presence of a 142bp satellite highly present in the *T. molitor* genome[8,43,44]. RepeatMasker detected 406 instances of the satellite repeats across 26 scaffolds covering up to 248,412 bp (or 0.08% of the assembly). Additionally, we performed a BLAST analysis with more stringent alignment parameters (BLASTn overlap >80%, identity ≥90%) and the satellite was newly detected in 17 scaffolds. We also found two variant sequences of this satellite (blastn evalue ≤10e-5, word_size=10) highly represented in scaffold 23. The longest form covers approximately 89% of the satellite (126-129bp), with average identity score 77%, while the shorter one, which is more abundant, covers about 44% of satellite (62-66bp) with mean sequence similarity of 85% (Figure S5, *Extended data*).

The mitochondrial genome was detected in two scaffolds. More precisely, the genbank mitochondrial genome of *T. molitor* (15,785 bp) was aligned to our assembly using Minimap2[45] and detected three times in scaffold 94 with a nucleotide identity of 85–89% (Figure S6, *Extended data*) but also in several other regions of the same scaffold with a lower nucleotide identity (53–73%) (Table S5, *Extended data*). The high copy number of mitochondrial DNA (mtDNA) per cell leads to relatively high depth of coverage, which causes misassemblies. NECAT constructed initially one single contig presenting three supplicated mitochondrial genomes. To resolve this misassembly, we re-assembled long reads that aligned to scaffold 94, using Flye version 2.9 (Flye, RRID:SCR_017016) with genome size

Table 1. Assembly and BUSCO Metrics for *T. molitor* (versions 2020, 2021) and *T. castaneum*.

Assembly statistics	Tenebrio 2021	Tenebrio 2020	Tribolium
# Contigs	112 (110 nuclear + 2 mitochondrial)	31,390	2,082
Cumulative size	287,931,689	280,780,514	165,944,485
Max contig length	33,042,542	271,822	31,381,287
Mean contig length	2,570,819	8,945	79,704
N50 (L50)	21,885,684 (6)	24,131 (3,180)	15,265,516 (5)
N90 (L90)	5,674,206 (16)	3,289 (16,525)	885,624 (12)
auN	18,643,178	30,387	15,592,941
GC%	36.72%	36.03%	33.86%
Number of N	28,500 (0.01%)	0 (0.00%)	13,515,130 (8.14%)
BUSCO on genome (N = 1,367)			
Complete	1,360 (99.5%)	1,213 (88.7%)	1,357 (99.2%)
Duplicated	7 (0.5%)	52 (3.8%)	6 (0.4%)
Fragmented	3 (0.2%)	67 (4.9%)	5 (0.4%)
Missing	4 (0.3%)	87 (6.4%)	5 (0.4%)

parameter set to 15k. Subsequently, the mitogenome was polished using Racon and Hapo-G with short reads (with the same methods used for the whole genome assembly) and obtained one single contig of 15,724 bp. The latter aligns with 98.39% identity to the *T. molitor* genbank mitogenome (15,785 bp) (Figure S8). Mitochondrial DNA was also detected in scaffold 65 (Figure S7, Extended data). However, due to its low ANI (50–75%) and the small fraction it occupies in the scaffold, we considered this alignment as a probable insertion of mtDNA in the nuclear genome. In view of the above considerations, we kept scaffold 65 in the current nuclear genome assembly and removed scaffold 94 as the mitochondrial genome.

Tenebrio molitor genome annotation

By combining RNA-seq and Coleoptera proteomes, we predicted a total of 21,435 genes which is higher than the number observed in *T. castaneum*. Beside this difference, other metrics are very comparable (Table 2). Quality of the gene prediction was assessed using BUSCO version 5.0.0 with Insecta database odb10 which contains 1,367 genes and showed a gene completeness of 96.5%. The published gene prediction based on the *T. castaneum* genome has fewer genes but a higher BUSCO score, which reflects the completeness of the gene prediction while the BUSCO score on the genome assembly reflects the completeness of the genome assembly. The tools and resources (transcriptomes, proteomes) used for gene prediction are different for the two beetles, so we can expect different gene completion between the two predictions. However, the observed difference in the number of predicted genes may rather refer to gene evolution, for example through gene duplication as further explained.

Not surprisingly, the two species share similar characteristics in terms of CDS lengths and number of exons (Figure S3, *Extended data*) as illustrated in a linear regression model with R^2=0.931 (Figure 1).

After stringent gene prediction filtering (see Methods section), the gene structure patterns remained enriched in single-exon genes 22% (compared to 7% of *T. castaneum*) (Table 2). Interestingly, 85% of them have a pfam or BlastP hit (evalue=10e-5), suggesting that they are *bona fide* gene predictions. Preliminary results show the existence of paralogous genes among them.

Genes and repeats evolution in the *Tenebrio molitor* genome

Paralogs search in the *T. molitor* and *T. castaneum* proteomes revealed a higher proportion of paralogs in *T. molitor* (Figure 2A). Their functional analyses through Pfam domain annotation showed the overabundance of histone-coding genes organized in blocks located in 25 different scaffolds (Figure 2B). As an example, one of these scaffolds (scaffold_25 ~545Kb) contains 108 genes coding for histones over its 162 predicted genes. Moreover, the global analysis of histone-coding genes in *T. molitor* showed that 806 genes coding for histones were mono-exonic. Several other protein families were overabundant in *T. molitor* compared to *T. castaneum* (Table 3). Among them, we can highlight the olfactory receptors containing the 7tm protein domain and two families of proteins involved in the developmental processes, the juvenile hormone binding proteins and the ecdysone kinases. Further investigations of the gene expression will greatly help to understand the function role in the *T. molitor* biology. Most of the duplicated genes

Table 2. Annotation and BUSCO metrics for *T. molitor* 2021 and *T. castaneum*.

Annotation Statistics	Tenebrio 2021	Tribolium
Number of genes (without isoforms)	21,435	14,503
Number of intronless genes	4,898	1,109
Gene length (mean : median)	7,590 : 1,779	8,032 : 2,364
Gene length without UTR (mean : median)	5,785 : 1,147	7,900 : 2,341
Number of exons per gene (mean : median)	4.15 : 3	5.19 : 4
Number of exons per gene (mean : median) Restricted to multi-exon genes	5.08 : 4	5.54 : 4
CDSs length (mean : median)	1,177 : 783	1,839 : 1,454
CDSs length (mean : median) Restricted to multi-exon genes	1,356 : 1,071	1,921 : 1,548
Cumulative size of coding sequences (%)	25,230,147 (8.8%)	26,681,223 (16.1%)
Number of introns	67,414	60,774
Intron length (mean : median)	1,465 : 55	1,446 : 53
Percentage of contigs with >= 1 gene (% in bases)	82.9% (99.2%)	18.5% (97.6%)
BUSCO with Insecta database (N = 1,367)		
Complete	1,319 (96.5%)	1,361 (99.6%)
Duplicated	8 (0.6%)	339 (24.8%)
Fragmented	12 (0.9%)	3 (0.2%)
Missing	36 (2.6%)	3 (0.2%)

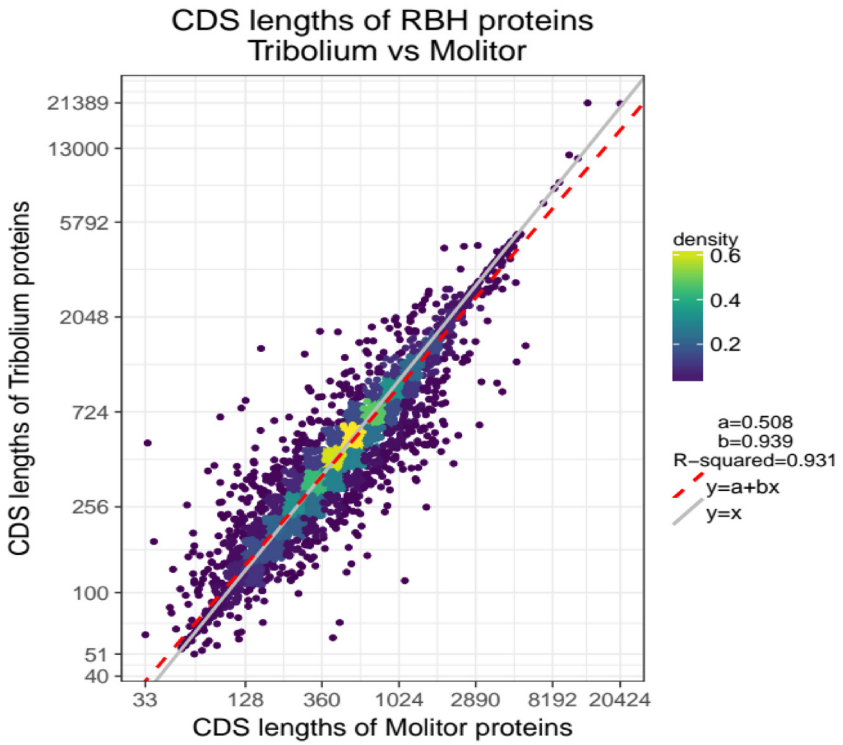

Figure 1. CDS length association for 10,495 orthologous genes of *T. molitor* and *T. castaneum*. Comparison plot with CDS lengths of *T. molitor* on x-axis and CDS lengths of *T. castaneum* on y-axis. Lengths (points) are log-scaled and coloured based on their density (highest density= yellow, lowest density=dark violet). The linear regression model best fitting the data is represented by the red-dashed line $y = a + bx$ with parameters a=0.508 and b=0.939. Higher densities are observed in the central part of the cloud and along the red-dashed fitted regression line.

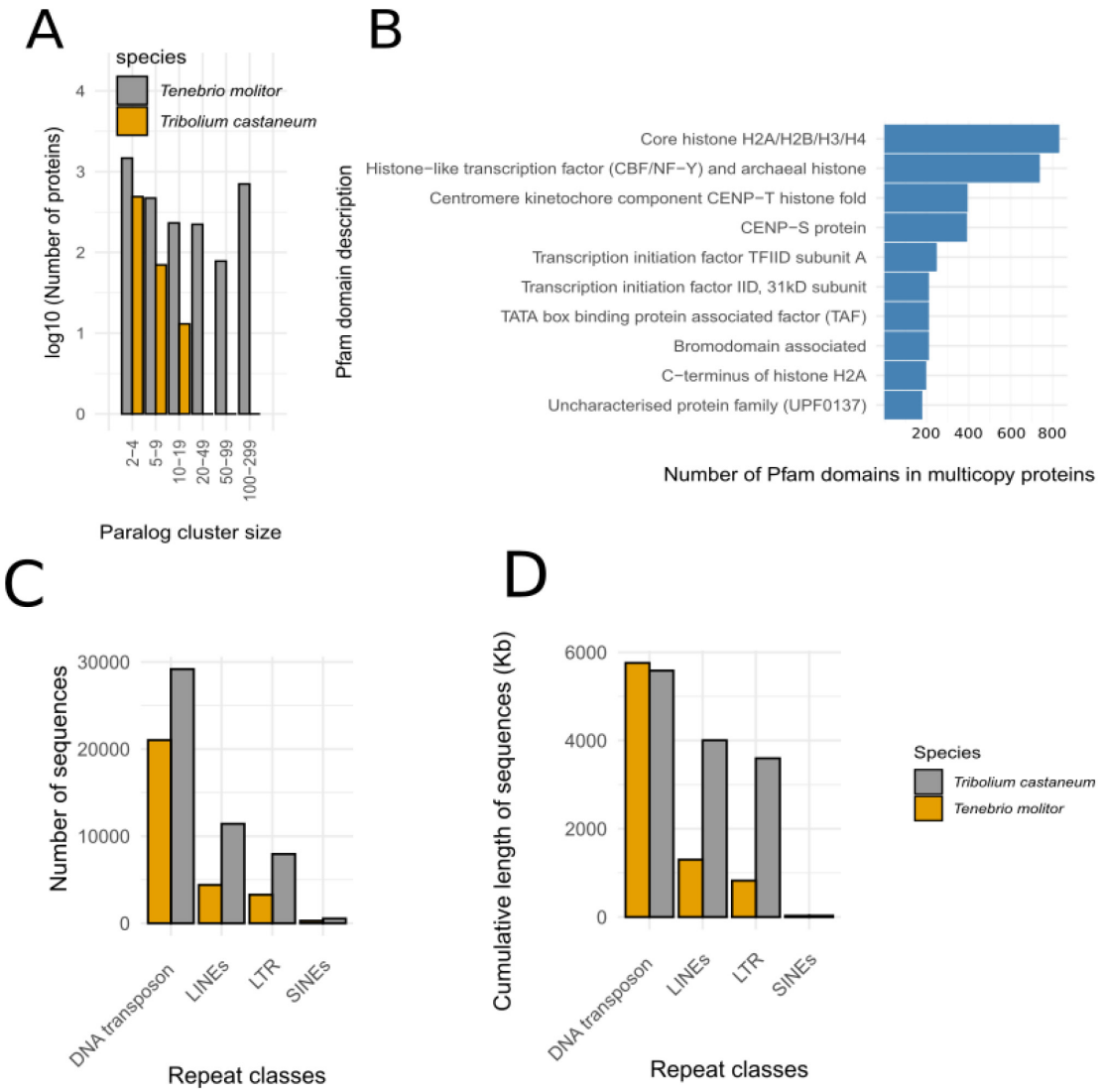

Figure 2. Paralog and repeated sequences analysis between *T. molitor* and *T. castaneum*. A. Number of paralogs (log scale) found in the *T. molitor* and *T. castaneum* paralog clusters. **B**. Functional annotation of the *T. molitor* paralogs found in the top 10 largest clusters. **C**. Number of major transposons in the *T. molitor* and *T. castaneum* genome. **D**. Cumulative length of major transposons the *T. molitor* and *T. castaneum* genome.

presented above are organized in small clusters randomly distributed in large scaffolds of the genome but for the antifreeze proteins that are all localized in a single large cluster. In addition, the duplicated gene clusters are located in gene rich regions of large scaffolds which supports bona fide gene duplications rather genome annotation bias.

Taken together, these results showed that the genome of *T. molitor* has experienced several duplications of mono-exonic histone-coding genes that may also explain the genome size difference between *T. molitor* and *T. castaneum*. However, as the technologies and methods used to produce and annotate the genome of *T. castaneum* were not the same, we cannot ensure that this expansion of histone genes is specific to *T. molitor* or shared among Tenebrionidae.

While DNA transposons are about 50% more abundant in *T. molitor*, they represent about a similar cumulative length (Figure 2C and B, Supplementary Table 4). On the opposite, the LINEs, SINEs and LTR transposons are about two to three times more abundant in *T. molitor* and their cumulative size is correlated to their abundance. However, in both genomes, the total length of repeated elements represents only 5 to 6% of the genome assembly (Supplementary Table 4).

Macrosynteny between the *Tenebrio molitor* and *Tribolium castaneum* genomes

The macrosynteny between the *T. molitor* scaffolds and *T. castaneum* chromosomes (Figure 3) showed a strong conservation of the genome. The current *T. molitor* assembly lacks the integration of genetic data and linkage groups to

Table 3. Overview of overabundant protein families in *T. molitor*.

Protein family	Tenebrio	Tribolium
Histone	1103	46
Ankyrin domain protein	238	159
Leucin rich repeat protein	213	178
Odorant receptor (7tm)	208	134
ABC transporter	159	87
Myb/SANT-like DNA-binding domain protein	122	30
Juvenile hormone binding protein	110	43
Ecdysone kinase	92	39
Antifreeze protein	42	0

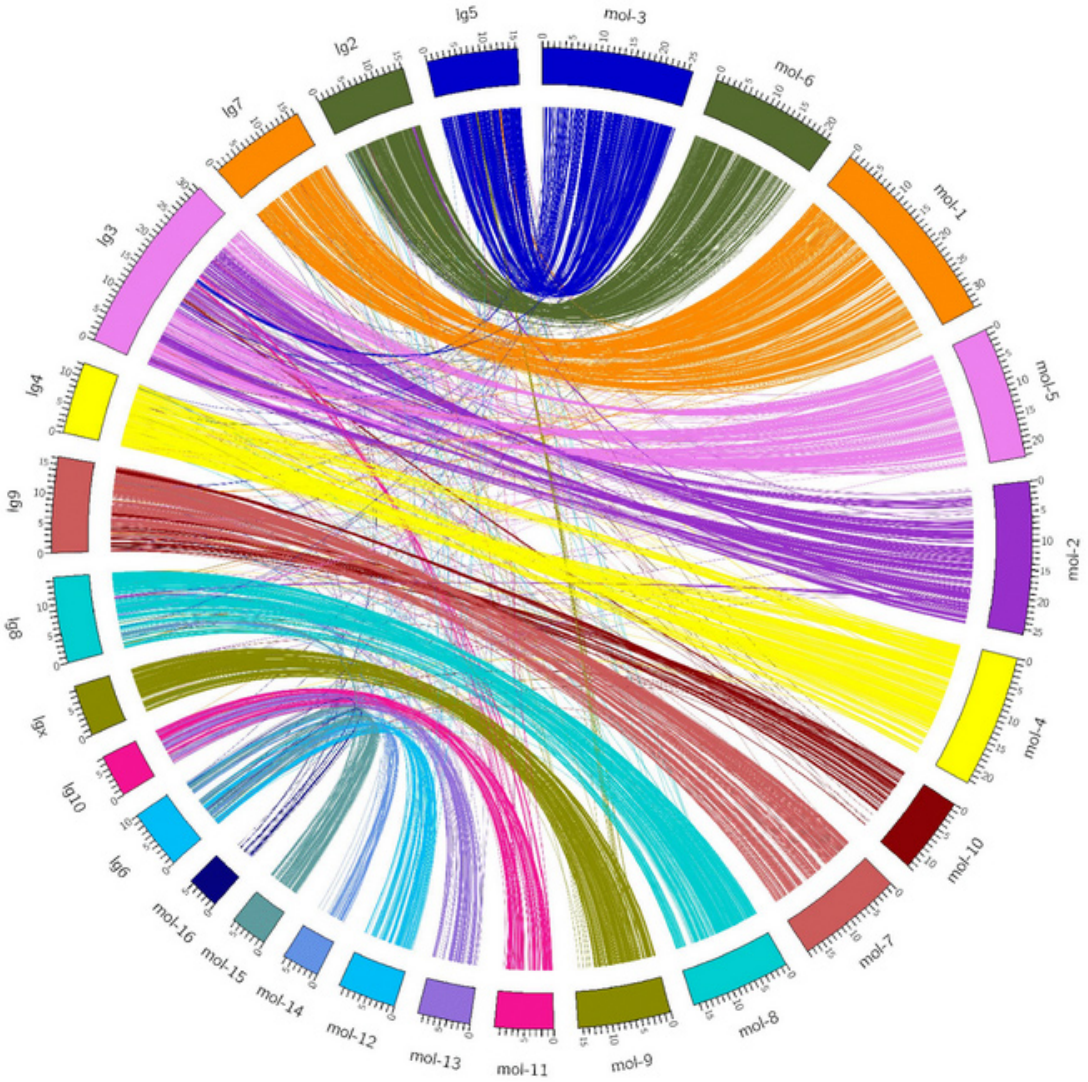

Figure 3. Synteny between *T. molitor* and *T. castaneum*. In the right semi-circle, the longest 16 *T. molitor* scaffolds are represented by orthogonal curved blocks placed next to each other. They are followed by the 10 *T. castaneum* chromosomes (left semi-circle). The unit length of the tick spacing of the blocks is 1Mb so that each block is proportional to the real size of a scaffold/chromosome. The 9,760 protein reciprocal best matches are drawn with colourful arches linking the orthologous regions between the two species.

reconstruct the entire chromosomes, and the assembly remains fragmented which makes the detection of possible chromosome rearrangements impossible. Future genetic works on *T. molitor* leading to the construction of a high-density map will greatly help to anchor the current assembly on linkage groups to obtain the complete two-dimensional chromosome organisation.

Conclusions

Our sequencing and assembly strategy to build the heterozygous genome of *T. molitor* by combining long read and Hi-C showed its efficiency and provided a high-quality genome assembly and the first genome annotation with a high completeness. Thanks to this new genomic resource, future work focusing on population, quantitative and functional genomics of genes of interest will be facilitated and will greatly improve our knowledge on the molecular basis of the *T. molitor* biology. Duplication of histones has been well described in many genomes, but here the number of duplications might be one of the highest described The presence of a relatively large number of monoexonic histone-coding genes supported by transcripts and conserved protein domains constitutes a field of investigation to understand the biological role and the evolution of these genes. The comparison of *T. molitor* to other available Coleoptera genomes will also aid better understanding of Coleoptera evolution and diversification. Additionally, thanks to the availability of the *T. molitor* genome and genes, new breeding programs can take advantage of this resource to improve and optimize mealworm production at the industrial scale through the combination of phenotypes and whole-genome genotypes to perform genome-wide association studies, quantitative trait locus analyses and genome estimation breeding values analysis.

Data availability

Underlying data

European Nucleotide Archive: Chromosome-scale assembly of the yellow mealworm genome. Accession number PRJEB44684.

European Nucleotide Archive: Chromosome-scale assembly of the yellow mealworm genome. Accession number PRJEB44703.

European Nucleotide Archive: Chromosome-scale assembly of the yellow mealworm genome. Accession number PRJEB44755.

NCBI BioProject: Mater immunity, reference transcriptome of *Tenebrio molitor*. Accession number PRJNA646689.

Other underlying data for the tenebrio genome are available on GitHub and Zenodo.

Zenodo: madoui/Tenebrio_Genome: updated supp data. https://doi.org/10.5281/zenodo.5499691[46].

This project contains the following underlying data:

- Supplementary_Data.pdf / Supplementary Table 6: Samples' accession numbers.
- Data / monoexonic (BED file with coordinates of monoexonic genes)
- Data / repeat (BED file with coordinates of the repeats)

Extended data

All extended data are available on GitHub and Zenodo.

Zenodo : madoui/Tenebrio_Genome: updated supp data. https://doi.org/10.5281/zenodo.5499691[46].

This project contains the following extended data within the file 'Supplementary_Data.pdf':

- Supplementary Table 1: Genomic data
- Supplementary Table 2: Transcriptomic data
- Supplementary Table 3: Metrics for long reads, contigs and scaffolds through different steps
- Supplementary Table 4: Repeats
- Supplementary Note 2: Gmove
- Supplementary Figure 1: GenomeScope Profile for *T. molitor*
- Supplementary Figure 2: K-mer plot before and after Haplomerger
- Supplementary Figure 3: Comparison of CDS lengths and number of exons of orthologous genes between *T. molitor* and *T. castaneum*
- Supplementary Figure 4: Aligning *T. molitor* to *T. castaneum*
- Supplementary Figure 5: Position of the 142 bp satellite (TMSATE1) on scaffolds 16, 58, 99, 23 and their coverage by Illumina Reads
- Supplementary Figure 6: Presence of mitochondrial genome on scaffold 94
- Supplementary Table 5: Alignment between the mitochondrial genome and the scaffold 94
- Supplementary Figure 7: Presence of mitochondrial genome on scaff 65
- Supplementary Figure 8: Alignment of scaffolds 94, 65
- Supplementary Figure 9: Coverage of scaffolds 65, 94 by Illumina mitochondrial reads
- Supplementary Figure 10: Assembly workflow
- Supplementary Figure 11: Annotation workflow

This project also contains the following extended data:

- assembly_workflow.pdf (details of the genome assembly method)
- annotation_workflow.html (details of the genome annotation method)

Acknowledgements

We acknowledge the Bio-based Industry for the organisational support and evaluation of the H2020 FARMYNG project.

References

1. Alexandratos N, Bruinsma J: **World agriculture towards 2030/2050: the 2012 revision.** 2012.
Reference Source
2. Steinfeld H, Gerber P, Wassenaar T, *et al.*: **Livestock's long shadow.** 2006.
Reference Source
3. Nowak V, Persijn D, Rittenschober D, *et al.*: **Review of food composition data for edible insects.** *Food Chem.* 2016; **193**: 39–46.
PubMed Abstract | Publisher Full Text
4. Oonincx DG, de Boer IJ: **Environmental Impact of the Production of Mealworms as a Protein Source for Humans - A Life Cycle Assessment.** *PLoS One.* 2012; **7**(12): e51145.
PubMed Abstract | Publisher Full Text | Free Full Text
5. van Huis A: **Potential of Insects as Food and Feed in Assuring Food Security.** *Annu Rev Entomol.* 2013; **58**(1): 563–583.
PubMed Abstract | Publisher Full Text
6. Cortes Ortiz JA, Ruiz AT, Morales-Ramos JA, *et al.*: **Insect Mass Production Technologies.** In: *Insects as Sustainable Food Ingredients.* Elsevier. 2016; 153–201.
Publisher Full Text
7. Morales-Ramos JA, Kelstrup HC, Rojas MG, *et al.*: **Body mass increase induced by eight years of artificial selection in the yellow mealworm (Coleoptera: Tenebrionidae) and life history trade-offs.** *J Insect Sci.* 2019; **19**(2): 4.
PubMed Abstract | Publisher Full Text | Free Full Text
8. Eriksson T, Andere AA, Kelstrup H, *et al.*: **The yellow mealworm (*Tenebrio molitor*) genome: a resource for the emerging insects as food and feed industry.** *J Insects Food Feed.* 2020; **6**(5): 445–455.
Publisher Full Text
9. de la Paz Celorio-Mancera M, Rastas P, Steward RA, *et al.*: **Chromosome Level Assembly of the Comma Butterfly (*Polygonia c-album*).** *Genome Biol Evol.* 2021; **13**(5): evab054.
PubMed Abstract | Publisher Full Text | Free Full Text
10. Dias GB, Altammami MA, El-Shafie HAF, *et al.*: **Haplotype-resolved genome assembly enables gene discovery in the red palm weevil *Rhynchophorus ferrugineus*.** *Sci Rep.* 2021; **11**(1): 9987.
PubMed Abstract | Publisher Full Text | Free Full Text
11. Wang X, Kelkar YD, Xiong X, *et al.*: **Genome report: Whole genome sequence and annotation of the parasitoid jewel wasp *Nasonia giraulti* laboratory strain RV2X[u].** *G3 (Bethesda).* 2020; **10**(8): 2565–2572.
PubMed Abstract | Publisher Full Text | Free Full Text
12. Biello R, Singh A, Godfrey CJ, *et al.*: **A chromosome-level genome assembly of the woolly apple aphid, *Eriosoma lanigerum* Hausmann (Hemiptera: Aphididae).** *Mol Ecol Resour.* 2021; **21**(1): 316–326.
PubMed Abstract | Publisher Full Text
13. Wick RR, Judd LM, Holt KE: **Performance of neural network basecalling tools for Oxford Nanopore sequencing.** *Genome Biol.* 2019; **20**(1): 129.
PubMed Abstract | Publisher Full Text | Free Full Text
14. Alberti A, Poulain J, Engelen S, *et al.*: **Viral to metazoan marine plankton nucleotide sequences from the *Tara* Oceans expedition.** *Sci Data.* 2017; **4**: 170093.
PubMed Abstract | Publisher Full Text | Free Full Text
15. Vurture GW, Sedlazeck FJ, Nattestad M, *et al.*: **GenomeScope: Fast reference-free genome profiling from short reads.** *Bioinformatics.* 2017; **33**(14): 2202–2204.
PubMed Abstract | Publisher Full Text | Free Full Text
16. Marijon P, Chikhi R, Varré JS: **Yacrd and fpa: Upstream tools for long-read genome assembly.** *Bioinformatics.* 2020; **36**(12): 3894–3896.
PubMed Abstract | Publisher Full Text
17. Chen Y, Nie F, Xie SQ, *et al.*: **Efficient assembly of nanopore reads via highly accurate and intact error correction.** *Nat Commun.* 2021; **12**(1): 60.
PubMed Abstract | Publisher Full Text | Free Full Text
18. Vaser R, Sović I, Nagarajan N, *et al.*: **Fast and accurate de novo genome assembly from long uncorrected reads.** *Genome Res.* 2017; **27**(5): 737–746.
PubMed Abstract | Publisher Full Text | Free Full Text
19. Aury JM, Istace B: **Hapo-G, haplotype-aware polishing of genome assemblies with accurate reads.** *NAR Genom Bioinform.* 2021; **3**(2): lqab034.
PubMed Abstract | Publisher Full Text | Free Full Text
20. Huang S, Kang M, Xu A: **HaploMerger2: rebuilding both haploid sub-assemblies from high-heterozygosity diploid genome assembly.** *Bioinformatics.* 2017; **33**(16): 2577–2579.
PubMed Abstract | Publisher Full Text | Free Full Text
21. Li H: **Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.** 2013.
Reference Source
22. Ghurye J, Rhie A, Walenz BP, *et al.*: **Integrating Hi-C links with assembly graphs for chromosome-scale assembly.** *PLoS Comput Biol.* 2019; **15**(8): e1007273.
PubMed Abstract | Publisher Full Text | Free Full Text
23. Zerbino DR, Birney E: **Velvet: Algorithms for de novo short read assembly using de Bruijn graphs.** *Genome Res.* 2008; **18**(5): 821–829.
PubMed Abstract | Publisher Full Text | Free Full Text
24. Schulz MH, Zerbino DR, Vingron M, *et al.*: ***Oases*: Robust *de novo* RNA-seq assembly across the dynamic range of expression levels.** *Bioinformatics.* 2012; **28**(8): 1086–1092.
PubMed Abstract | Publisher Full Text | Free Full Text
25. Morgulis A, Gertz EM, Schäffer AA, *et al.*: **A fast and symmetric DUST implementation to mask low-complexity DNA sequences.** *J Comput Biol.* 2006; **13**(5): 1028–1040.
PubMed Abstract | Publisher Full Text
26. Tarailo-Graovac M, Chen N: **Using RepeatMasker to identify repetitive elements in genomic sequences.** *Curr Protoc Bioinformatics.* 2009; Chapter 4: Unit 4.10.
PubMed Abstract | Publisher Full Text
27. Bao W, Kojima KK, Kohany O: **Repbase Update, a database of repetitive elements in eukaryotic genomes.** *Mob DNA.* 2015; **6**(1): 11.
PubMed Abstract | Publisher Full Text | Free Full Text
28. Kent WJ: **BLAT--the BLAST-like alignment tool.** *Genome Res.* 2002; **12**(4): 656–664.
PubMed Abstract | Publisher Full Text | Free Full Text
29. Mott R: **EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA.** *Comput Appl Biosci.* 1997; **13**(4): 477–478.
PubMed Abstract | Publisher Full Text
30. UniProt Consortium: **UniProt: The universal protein knowledgebase in 2021.** *Nucleic Acids Res.* 2021; **49**(D1): D480–D489.
PubMed Abstract | Publisher Full Text | Free Full Text
31. Birney E, Clamp M, Durbin R: **GeneWise and Genomewise.** *Genome Res.* 2004; **14**(5): 988–995.
PubMed Abstract | Publisher Full Text | Free Full Text
32. Dubarry M, Noel B, Rukwavu T, *et al.*: **Gmove a tool for eukaryotic gene predictions using various evidences.** *F1000Res.* 2016; **5**.
Publisher Full Text
33. Eddy SR: **Accelerated profile HMM searches.** *PLoS Comput Biol.* 2011; **7**(10): 1002195.
PubMed Abstract | Publisher Full Text | Free Full Text
34. Buchfink B, Reuter K, Drost HG: **Sensitive protein alignments at tree-of-life scale using DIAMOND.** *Nat Methods.* 2021; **18**(4): 366–368.
PubMed Abstract | Publisher Full Text | Free Full Text
35. Flynn JM, Hubley R, Goubert C, *et al.*: **RepeatModeler2: Automated genomic discovery of transposable element families.** *biorxiv*. 2019.
Publisher Full Text
36. Haas BJ: **TransposonPSI: An Application of PSI-Blast to Mine (Retro-)Transposon ORF Homologies.** 2011.
Reference Source
37. Altschul SF, Gish W, Miller W, *et al.*: **Basic local alignment search tool.** *J Mol Biol.* 1990; **215**(3): 403–410.
PubMed Abstract | Publisher Full Text
38. Marçais G, Delcher AL, Phillippy AM, *et al.*: **MUMmer4: A fast and versatile genome alignment system.** *PLoS Comput Biol.* 2018; **14**(1): e1005944.
PubMed Abstract | Publisher Full Text | Free Full Text
39. Krzywinski M, Schein J, Birol I, *et al.*: **Circos: An information aesthetic for comparative genomics.** *Genome Res.* 2009; **19**(9): 1639–1645.
PubMed Abstract | Publisher Full Text | Free Full Text

40. Juan C, Petitpierre E: **C-banding and DNA content in seven species of Tenebrionidae (Coleoptera).** *Genome.* 1989; **32**(5): 834–839.
Publisher Full Text

41. Simão FA, Waterhouse RM, Ioannidis P, *et al.*: **BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.** *Bioinformatics.* 2015; **31**(19): 3210–3212.
PubMed Abstract | **Publisher Full Text**

42. Lorenzen MD, Doyungan Z, Savard J, *et al.*: **Genetic linkage maps of the red flour beetle, *Tribolium castaneum*, based on bacterial artificial chromosomes and expressed sequence tags.** *Genetics.* 2005; **170**(2): 741–747.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

43. Petitpierre E, Gatewood JM, Schmid CW: **Satellite DNA from the beetle *Tenebrio molitor*.** *Experientia.* 1988; **44**(6): 498–499.
Publisher Full Text

44. Davis CA, Wyatt GR: **Distribution and sequence homogeneity of an abundant satellite DNA in the beetle, *Tenebrio Molitor*.** *Nucleic Acids Res.* 1989; **17**(14): 5579–5586.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

45. Li H: **Minimap2: pairwise alignment for nucleotide sequences.** *Bioinformatics.* 2018; **34**(18): 3094–3100.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

46. Madoui A: **madoui/Tenebrio_Genome: updated supp data (v0.4).** *Zenodo*. 2021.
http://www.doi.org/10.5281/zenodo.5499691

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 3

Reviewer Report 05 September 2022

https://doi.org/10.21956/openreseurope.16289.r30028

✔ **Gregor Bucher**
Johann Friedrich Blumenbach Institute, GZMB, University of Göttingen, Göttingen, Germany

The authors have addressed all my concerns and I congratulate them for this paper!

I have to confess that "chromosome scale" still seems a bit of an overstatement to me but that may be a matter of personal opinion.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Gene function studies in beetles; evolution and development of the insect head and brain; RNAi in pest control

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 2

Reviewer Report 13 April 2022

https://doi.org/10.21956/openreseurope.15728.r28846

? **Gregor Bucher**
[1] Johann Friedrich Blumenbach Institute, GZMB, University of Göttingen, Göttingen, Germany
[2]

Johann Friedrich Blumenbach Institute, GZMB, University of Göttingen, Göttingen, Germany

Eleftheriou et al. present their work on enhancing the genomic sequence and annotation of the yellow mealworm *Tenebrio molitor*, an insect species potentially providing an alternative source for feed and food. Based on extensive sequencing of DNA and RNA and by using long- and short read technology and Hi-C data they provide a very much-improved version of the genomic sequence and gene set of this species (e.g. dramatically increasing N50). Finally, they assess aspects of genome assembly and gene set by comparing with another well-sequenced beetle, *Tribolium castaneum*.

Questions with respect to the analyses:

1. You compare your gene set and genome to *T. castaneum*. What assembly and gene set did you use for that purpose? Richards et al[1]., Herndon et al[2]., NCBI, ...? For example, I could e.g. not match your N50 values (table 1) or gene numbers (table 2) to those published in the latest genome paper (Herndon et al[2]). Please provide references and explain, where you got your numbers from.

2. The higher number of annotated genes in *T. molitor* compared to *T. castaneum* is potentially interesting but-in my opinion-not yet sufficiently supported. It could be interesting biology or annotation artifact (some numbers in table 2 would be in line with the latter assumption: shorter transcripts with less exons but more single exon genesin *T. molitor*. It could for instance have happened during the processing step to split erroneously assembled transcripts). Some suggestions on alternative approaches for annotation/splitting transcripts in order to test, whether the number really is that different:
   - If I understood correctly, you first assembled the transcriptome from RNA-seq data only and used algorithms to split potentially fused transcripts. While I find the given criteria convincing, the number of splits still depends on the parameters that you choose. Why not use your new and excellent genome sequence information for that purpose? If two parts of a predicted transcript map in close proximity to each other and in the right orientation on the same scaffold, it should probably not be split.

   - A more fundamental question: Why not first map the RNA-seq reads onto your genome assembly and then apply annotation tools?

   - A very good criterion to fuse or to split predictions are “split reads” (i.e. RNA-seq reads that span intron boundaries and therefore map to different exon boundaries on the genome thereby precisely mapping the intron).

3. Strongly increased paralog cluster size. This finding seems unexpected and interesting but again I would like to see some further analyses to confirm this. The increase is specifically striking with respect to the large cluster sizes (> 20 copies) where *T. castaneum* does not seem to have any cluster but *T. molitor* many. Your analysis in Fig. 2B seems based on all clusters - it would be nice to manually follow-up some of the intriguing large clusters in order to confirm that they are real and to understand the underlying pattern. Focused on the large clusters questions could be: What protein families are heavily expanded (apart from histones)? Where in the genome are the paralogs of selected expanded gene families located (clustered or evenly distributed, do they map to the chromosome parts that seem ambiguous in their syntenty to *Tribolium*?). If they are enriched in ambiguous genome regions – would that indicate assembly problems or localized genome changes? The results should either confirm this intriguing finding and allow for

hypotheses on the underlying patterns or they might reveal that this reflects annotation bias.

4. Genome size difference: The overall genome size is different and I would have appreciated some words on what the explanation is. More ORFs? Paralog expansion? Repetitive sequences (you mention that in both genomes that is around 5-6%). You mention that the histone expansion may explain this – how much (in %) does this gene family contribute to the expansion – base on that number: is it one of many drivers or one of the key drivers? What are the other drivers?

5. BUSCO: you provide two different BUSCO analyses in table 1 vs. table 2. Either you remove the one that you consider less precise or you explain what the difference in the analyses is and what it means.

6. Synteny: for most scaffolds, it seems quite clear to what chromosome of *Tribolium* they correspond – either individually or fused with other scaffolds (based on Fig. 3). LG6 seems to be split into several – any explanation for that? Do you expect to have assembled the Y chromosome (given that you used a male) – how does that compare with *Tribolium*? Given that you know the number of expected chromosomes in your species, some thoughts on which scaffolds probably belong to one chromosome would be nice.

Typos and minor comments:
Title:
While the authors provide an excellent new assembly I agree with the first reviewer that the term "chromosome-scale" may seem a bit strong given that based on latest technology, some of the currently emerging insect genomes reflect the expected number of chromosomes almost completely.

Plain language summary:
"red flour beetle"

Methods:
The metamorphic stage should be called pupa (the term nymph seems to be reserved for the pre-adult stages in hemimetabolous insects) – please replace (3 times at least).

I suggest using "embryos" or "embryonic stages" instead of "eggs"
"… with chloroform centrifugation…." Did you mean extraction?

RNA extraction: Please specify temperature and the time window of egg collection as this determines the embryonic stages that are covered.
"… occur in the same orientation as antisense RNA." - maybe more clear "…occur in antisense orientation only."
"cDNA was then 3'-adenylated…"
Transcriptome assembly: "…developmental stages (embryos, larvae, pupae adults)"
"…and only contigs larger than 150…" – did you mean "reads" here? Same issue two lines below.
"… this tool aims to split contigs with different functional sites into different contigs." – is this what you meant?

Results:
" 7% of *T. Castaneum*…"

"... we cannot ensure that this expansion of histone genes is specific to ..."
There are more DNA transposons but the overall length is similar – does that mean that they are shorter?

References

1. Tribolium Genome Sequencing Consortium, Richards S, Gibbs RA, Weinstock GM, et al.: The genome of the model beetle and pest Tribolium castaneum.*Nature*. 2008; **452** (7190): 949-55 PubMed Abstract | Publisher Full Text
2. Herndon N, Shelton J, Gerischer L, Ioannidis P, et al.: Enhanced genome assembly and a new official gene set for Tribolium castaneum.*BMC Genomics*. 2020; **21** (1): 47 PubMed Abstract | Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and does the work have academic merit?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Gene function studies in beetles; evolution and development of the insect head and brain; RNAi in pest control

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 24 Aug 2022

Mohammed-Amin Madoui

Eleftheriou et al. present their work on enhancing the genomic sequence and annotation of the yellow mealworm Tenebrio molitor, an insect species potentially providing an alternative source for feed and food. Based on extensive sequencing of DNA and RNA and by using long- and short read technology and Hi-C data they provide a very much-improved

version of the genomic sequence and gene set of this species (e.g. dramatically increasing N50). Finally, they assess aspects of genome assembly and gene set by comparing with another well-sequenced beetle, Tribolium castaneum.

Questions with respect to the analyses:
Question 1: You compare your gene set and genome to *T. castaneum*. What assembly and gene set did you use for that purpose? Richards et al1., Herndon et al2., NCBI, ...? For example, I could e.g. not match your N50 values (table 1) or gene numbers (table 2) to those published in the latest genome paper (Herndon et al2). Please provide references and explain, where you got your numbers from.
Answer 1: Indeed the reference of the genome version used for comparative genomics was missing. We used the Herndon et al version. We added the reference in the manuscript. The genome assembly and gene annotation metrics were calculated directly from the data downloaded from NCBI https://www.ncbi.nlm.nih.gov/genome/216.

Question 2.1: The higher number of annotated genes in *T. molitor* compared to *T. castaneum* is potentially interesting but in my opinion-not yet sufficiently supported. It could be interesting biology or annotation artifact (some numbers in table 2 would be in line with the latter assumption: shorter transcripts with less exons but more single exon genesin *T. molitor*. It could for instance have happened during the processing step to split erroneously assembled transcripts). Some suggestions on alternative approaches for annotation/splitting transcripts in order to test, whether the number really is that different:

• If I understood correctly, you first assembled the transcriptome from RNA-seq data only and used algorithms to split potentially fused transcripts. While I find the given criteria convincing, the number of splits still depends on the parameters that you choose. Why not use your new and excellent genome sequence information for that purpose?

If two parts of a predicted transcript map in close proximity to each other and in the right orientation on the same scaffold, it should probably not be split.

Answer 2.1: Although the splitting method is based on several parameters, it splits only 5% of the contigs. And 60% of the contigs split in two were oppositely oriented (orientation is inferred by RNA-reads alignment). The rest (40% of split contigs) with the same orientation were split because of a significant difference in coverage (based on the mpileup of RNA reads) between the two sides of the break point. Considering the proportion of split transcript contigs, we think that the high number of single-exon genes in *T. molitor* is not produced during the splitting step, but reveals interesting biology in the *Tenebrio molitor* genome (now presented in the manuscript). Here is a typical example of a contig split into two parts: See figure 1 here.

• The track, named “Gmove Final 21435 genes”, shows two different genes. They are two co-oriented tandem duplicated genes.

• All RNA-seq contigs (purple tracks) show two distinct genes.

• What strengthens the necessity of splitting are the protein mappings (green tracks). None

of these alignments covers the two loci with all the exons. • For *T. castaneum* the same protein aligns twice in a row («esterase FE4 isoform X1 on the left and «esterase E4» on the right.
• Also, Gmove predicts two genes spanning all exons based on the split contigs.

Here is another example of a tandem single-exon gene is the leucine-rich repeat-containing protein of the figure 2 here.

• At first, the transcripts are assembled with Oases. However, Oases does not take into account the orientation (as other assemblers do, eg Trinity). So, it makes sense to look for RNA contigs erroneously assembled.

•Subsequently, reads that are not properly paired are eliminated. Then, depending on the depth of coverage, potential breakpoints are defined.

•We finally validate the breakpoints based on ORFs (Transdecoder) and domains (CDDsearch) information.

Question 2.2: A more fundamental question: Why not first map the RNA-seq reads onto your genome assembly and then apply annotation tools?
Answer 2.2: This is, indeed, another way to perform the annotation, although it does not guaranty more robust results. We could have used, for instance, hisat2 combined with StringTie in order to align RNA-seq reads onto the genome and then assemble. However, it was not a priority to test every possible methods of annotation.

Question 2.3: A very good criterion to fuse or to split predictions are “split reads” (i.e. RNA-seq reads that span intron boundaries and therefore map to different exon boundaries on the genome thereby precisely mapping the intron).
Answer 2.3: Yes, split reads are indeed good predictors to fuse or split predictions when you work with read alignments but in our case we work with contigs alignment so we use “split contigs” to define intron boundaries, which are also good predictors and similar to the ”split reads” approach.

Question 3.1: Strongly increased paralog cluster size. This finding seems unexpected and interesting but again I would like to see some further analyses to confirm this. The increase is specifically striking with respect to the large cluster sizes (> 20 copies) where *T. castaneum* does not seem to have any cluster but *T. molitor* many. Your analysis in Fig. 2B seems based on all clusters - it would be nice to manually follow-up some of the intriguing large clusters in order to confirm that they are real and to understand the underlying pattern. Focused on the large clusters, questions could be: What protein families are heavily expanded (apart from histones)?
Answer 3.1: It is a very interested question. To address it, we added the following table in the main manuscript (now table 3) that illustrates the biology behind the gene duplications. See table 1.

Question 3.2: Where in the genome are the paralogs of selected expanded gene families located (clustered or evenly distributed, do they map to the chromosome parts that seem

ambiguous in their syntenty to Tribolium?).
Answer 3.2: As for histones, most of the duplicated genes are organized in small clusters randomly distributed in large scaffolds of the genome but for the antifreeze proteins that are all located in a single large cluster. This information is now mentioned in the manuscript.

Question 3.3: If they are enriched in ambiguous genome regions – would that indicate assembly problems or localized genome changes? The results should either confirm this intriguing finding and allow for hypotheses on the underlying patterns or they might reveal that this reflects annotation bias.
Answer 3.3: Manual curation showed that the duplicated genes clusters are located in gene rich regions of large scaffolds which support the presence of many duplication events in Tenebrio. Except for the antifreeze proteins that are located in a single cluster on the scaffold 7. This information is now mentioned in the manuscript.

Question 4: Genome size difference: The overall genome size is different and I would have appreciated some words on what the explanation is. More ORFs? Paralog expansion? Repetitive sequences (you mention that in both genomes that is around 5-6%). You mention that the histone expansion may explain this – how much (in %) does this gene family contribute to the expansion – base on that number: is it one of many drivers or one of the key drivers? What are the other drivers?
Answer 4: This is a difficult question to address and we have tried to give some new elements. We identified one repeat element of 77kb present 318 times in the *Tenebrio molitor* genome and absent in Tribolium. The cumulative size of this element leads to 20Mb of sequences. On the other side, the histone gene sequences correspond to 7.4Mb and the cumulative size of the gene famlies listed in the table 3 lead to 16.3Mb. By taking into account these numbers, we see that both gene duplication and repeated element participated to the genome expansion. Also by looking at the circus plot (Figure 3), we see that the large scaffolds anchoring on Tribolium linkage groups are larger than the Tribolium chromosomes but for the LG8. This shows that the expansion is quite homogeneous in the different Tenebrio chromosomes and involved both genes and repeats expansion.

Question 5: BUSCO: you provide two different BUSCO analyses in table 1 vs. table 2. Either you remove the one that you consider less precise or you explain what the difference in the analyses is and what it means.
Answer 5: Indeed, the BUSCO analysis on the genome assembly reflects the completeness of the genome assembly while the BUSCO analysis on the genome annotation reflects the completeness of the gene prediction. This has been mentioned as follow "The published gene prediction based on the *T. castaneum* genome has fewer genes but a higher BUSCO score, which reflects the completeness of the gene prediction while the BUSCO score on the genome assembly reflects the completeness of the genome assembly. »

Question 6.1: Synteny: for most scaffolds, it seems quite clear to what chromosome of Tribolium they correspond – either individually or fused with other scaffolds (based on Fig. 3). LG6 seems to be split into several – any explanation for that?
Answer 6.1: We can explain the presence of two large scaffolds anchoring on the LG6 of Tribolium by a lack of data allowing joining the two scaffolds. More long-range information

provided by long reads, Hi-C or optical maps and the use of the Tenebrio genetic map will help to mind this gap.

Question 6.2: Do you expect to have assembled the Y chromosome (given that you used a male) – how does that compare with Tribolium?
Answer 6.2: Indeed, we expect to have assembled the Y chromosome but the investigation of the male specific genes is still undergoing. The Tribolium sequence of the Y chromosome was not provided in our data. A high-density genetic map of Tenebrio will be soon constructed and will help to identify the scaffolds corresponding to the Y chromosomes. Also, future work on male specific gene expression ill also help to identify Y chromosomes genes.

Question 6.3: Given that you know the number of expected chromosomes in your species, some thoughts on which scaffolds probably belong to one chromosome would be nice.
Answer 6.3: We are currently building a genetic map for *T. molitor* that will allow us to properly anchor our genome assembly on the chromosomes. This will also help to characterize the Y chromosome. We plan to release the pseudo-molecules after the complete analysis of the genetic data.

Typos and minor comments:
Comment 1: Title: While the authors provide an excellent new assembly I agree with the first reviewer that the term “chromosome-scale” may seem a bit strong given that based on latest technology, some of the currently emerging insect genomes reflect the expected number of chromosomes almost completely.
Answer: We understand your opinion concerning the title. We obtained one or two very large scaffolds for each chromosome (according to the synteny with *T. castaneum*) except for one chromosome corresponding to the linkage group 6 in *T. castaneum* where we obtained four large scaffolds. Thus, we used the term “chromosome-scale” because the scaffolds have the chromosome length scale i.e several megabases. We distinguish the term “chromosome-scale”, commonly used when you obtain very large scaffolds reaching the chromosome size, to the “telomere-to-telomere” term which is used when you obtain the complete chromosome sequences without gaps.

Comment 2: Plain language summary: “red flour beetle”
Answer: This has been corrected

Comment 3: Methods: The metamorphic stage should be called pupa (the term nymph seems to be reserved for the pre-adult stages in hemimetabolous insects) – please replace (3 times at least).
Answer: all the occurrences has been corrected

Comment 4: I suggest using “embryos” or “embryonic stages” instead of “eggs”
Answer: This has been modified

Comment 5: “… with chloroform centrifugation….” Did you mean extraction?
Answer: Indeed, we meant extraction, this has been changed

Comment 6: RNA extraction: Please specify temperature and the time window of egg collection as this determines the embryonic stages that are covered.
Answer: The eggs were collected at RT and collected within a week after laying. This is now mentioned in the manuscript.

Comment 7 "... occur in the same orientation as antisense RNA." - maybe more clear "...occur in antisense orientation only."
Answer: This has been modified according to your advice

Comment 8: "cDNA was then 3'-adenylated..."
Answer: This has been corrected

Comment 9: Transcriptome assembly: "...developmental stages (embryos, larvae, pupae adults)"
Answer: This has been corrected

Comment 10: "...and only contigs larger than 150..." – did you mean "reads" here? Same issue two lines below.
Answer: We meant "contigs"

Comment 11: "... this tool aims to split contigs with different functional sites into different contigs." – is this what you meant?
Answer: This has been corrected

Comment 12: Results: " 7% of T. Castaneum..."
Answer: This has been corrected

Comment 13: "... we cannot ensure that this expansion of histone genes is specific to ..."
Answer: The sentence has been modified according to your advice

Comment 14: There are more DNA transposons but the overall length is similar – does that mean that they are shorter?
Answer: It seems that DNA transposons are possibly shorter in Tribolium compared to Tenebrio. This observation goes in the sense of a genome expansion in Tenebrio compared to Tribolium in which DNA transposons may have participated.

We gratefully thank the reviewer for its very helpful comments and suggestions.

Competing Interests: No competing interests were disclosed.

Reviewer Report 04 April 2022

https://doi.org/10.21956/openreseurope.15728.r28658

Barbara Feldmeyer
[1] Senckenberg Biodiversity and Climate Research Centre, Frankfurt am Main, Germany
[2] Senckenberg Biodiversity and Climate Research Centre, Frankfurt am Main, Germany

I have read the revised version of the manuscript, and response to reviewer. The authors did a good a job in revising the manuscript, and I have no further comments.

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and does the work have academic merit?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Evolutionary biology, genomics, transcriptomics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 24 November 2021

https://doi.org/10.21956/openreseurope.15073.r27973

Barbara Feldmeyer
[1] Senckenberg Biodiversity and Climate Research Centre, Frankfurt am Main, Germany
[2] Senckenberg Biodiversity and Climate Research Centre, Frankfurt am Main, Germany

The authors construct a new “chromosome” assembly of the T. molitor genome using current state of the art technology. While the manuscript is well written and clear in most parts there are three major aspects which should be addressed before indexing of the manuscript.

General major comments

1. I find the title = chromosome scale somewhat misleading given the information later in the text. The authors obtain 112 scaffolds using the HiC data but then discuss 16 scaffolds representing 90% of the genome. Do you know/expect T.molitor to have 16 chromosomes? I could not find any information on that matter. Is the conclusion of having a chromosome scale assembly justified and why? Please add this information in the discussion section for example.

2. The comparative genomic section consists of a Blastp of both protein sets and a synteny plot. I understand that genome re-arrangement analyses are not feasible, but what about the gene content. Which genes or gene families are found or maybe enriched in the +5000 genes in T.molitor? How do repeat contents differ between species?

3. The mitochondrial sequence seems to be attached to 2 (3?) scaffolds, where it should be a single circular sequence only. This indicates some extent of misassembly and should at least be manually curated.

Introduction

- The global population: specify that you are talking about the HUMAN not beetle population.

- “However, prior genomic resources on *T. molitor* are needed to accelerate such genetic programs.”: Not sure that “prior” is the right wording here. Maybe “suitable”, or “ a good quality genomes”...

- Change “Here, we propose a *T. molitor* genome assembly based” to “Here, we present”

- “kept at room temperature and humidity”: % humidity information is missing.

- Change “put on a starvation for three days” to “ were starved for three days”

Methods

- “For mRNA extraction, eggs, larva, nymphs, adult males and females were isolated without specific diet”: More information needed. Did you make one extraction from all samples, did you isolate each sample individually, did you sequence one pool or each sample individually, which kit did you use for extraction, did you Illumina sequence and which depth? => after reading further I realize that this information is given later. Maybe rephrase the sentence to “Eggs, larva, nymphs, adult males and females were collected for later mRNA extraction.”

- “HMW gDNA was size-selected using Short Read Eliminator” change to “...using the Short...”

- “following protocol, using the Oxford Nanopore” remove “using”.
- Dovetail “The one-third of the cryoground” change to “Another third of the ….”
- “obtained 138 scaffolds. Only scaffolds larger than 35kb were kept resulting in a final assembly of 112 scaffolds” why >35kb?
- Transcriptome “The contigs 5’ and 3’ends were cleaned.” Using which tool and parameters?
- Why did you use Velvet and Oases especially since they are prone to create chimeras? Why not use Trinity for example?
- Change “the research of ORFs” to “the identification of ORFs”.
- Change “aims to split contigs sequences” either to “aims to split contigs with different…” or “aims to split contig sequences”
- Do you make the contig splitting tool available somewhere? Is the code accessible on github or in the supplement? If so add reference.
- “BLAT (version 36 with default parameter) matches with score” change to “with a score”

Results and Discussion

- Table 1 compares the newest version of Tmolitor with the older and Tcastaneum but on contig scale. Why contig and not scaffold, or chromosome scale?
- “of 99.5% (using version 5.0.0 with Insecta database odb10) while the predicted genome size was 310 Mb (Figure S1, *Extended data*)” remove “while the predicted genome size was 310 Mb (Figure S1, *Extended data*)” from this sentence. You give this information at the beginning of the paragraph already.
- "Furthermore, mitochondrial genome“ change to “ “the mitochondrial genome”.
- “More precisely, the mitochondrial genome of *T. molitor* (15,785 bp) was aligned to the assembly using Minimap245.”: This whole mt paragraph is confusing. Do you mean you aligned an existing mt genome, from the previous assembly? From which other study? => add reference and/or accession number. And you aligned it to identify the mt genome in your assembly. Why is the mt genome split across scaffolds? It should be on single contig only? => It sounds like misassembly and should be corrected. I.e. the mt parts should be assembled into a separate mt genome and should be removed from the scaffolds.
- “a total of 21,435 genes which is higher than the number observed in *T. castaneum*”: could you discuss why you think you have more genes but lower Busco (Table2)?

Conclusions

- Change “future works” to “future work”.

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and does the work have academic merit?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Evolutionary biology, genomics, transcriptomics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Comments on this article

Version 1

Author Response 21 Feb 2022

Mohammed-Amin Madoui

Reviewer comments in italics.

- *I find the title = chromosome scale somewhat misleading given the information later in the text. The authors obtain 112 scaffolds using the HiC data but then discuss 16 scaffolds representing 90% of the genome. Do you know/expect T.molitor to have 16 chromosomes? I could not find any information on that matter. Is the conclusion of having a chromosome scale assembly justified and why? Please add this information in the discussion section for example.*

Response: We understand your opinion concerning the title. As *Tribolium castaneum*, *Tenebrio molitor* have 10 chromosomes (Juan, Carlos *et al.* "Improving beetle karyotype analysis: restriction endonuclease banding of Tenebrio molitor chromosomes." *Heredity* 65 (1990): 157-162.). We obtained one or two very large scaffolds for each chromosome (according to the synteny with *T. castaneum*) except for one chromosome corresponding to the linkage group 6 in *T. castaneum*

where we obtained four large scaffolds. Thus, we used the term “chromosome-scale” because the scaffold has the chromosome length scale i.e., several megabases. We may distinguish the term “chromosome-scale“ commonly used when you obtain very large scaffolds to the “telomere-to-telomere” term which is used when you obtain the complete chromosome sequences without gaps.

- *The comparative genomic section consists of a Blastp of both protein sets and a synteny plot. I understand that genome re-arrangement analyses are not feasible, but what about the gene content. Which genes or gene families are found or maybe enriched in the +5000 genes in T.molitor? How do repeat contents differ between species?*

Response: This is a very interesting point, to address this question we searched for paralogs in *T. molito*r and *T. castananeum*. We found a higher number of paralogs in *T. molitor* compared to *T. castaneum*. Among these paralogs, we found that 860 genes were histones. In the new manuscript version, we illustrated these results in Figure 1.A and Figure 1.B. The repeat analysis showed that they represent 5 to 6% of the total genome size. A detailed distribution of repeat distribution in the two genomes is now illustrated in Figure 1.C and Figure 1.D.

- *The mitochondrial sequence seems to be attached to 2 (3?) scaffolds, where it should be a single circular sequence only. This indicates some extent of misassembly and should at least be manually curated.*

Response: Indeed, the mitochondrial genome was reassembled separately and a new paragraph explains the issues we found.

- *The global population: specify that you are talking about the HUMAN not beetle population.*

Response: This is now specified.

- *“However, prior genomic resources on T. molitor are needed to accelerate such genetic programs.”: Not sure that “prior” is the right wording here. Maybe “suitable”, or “ a good quality genomes”…*

Response: We changed “prior” by “suitable”.

- *Change “Here, we propose a T. molitor genome assembly based” to “Here, we present”*

Response: We changed “propose” by “present”.

- *“kept at room temperature and humidity”: % humidity information is missing.*

Response: Unfortunately, the humidity was not monitored.

- *Change “put on a starvation for three days” to “were starved for three days”*

Response: We changed the sentence according to the reviewer's recommendation.

- “*For mRNA extraction, eggs, larva, nymphs, adult males and females were isolated without specific diet”: More information needed. Did you make one extraction from all samples, did you isolate*

each sample individually, did you sequence one pool or each sample individually, which kit did you use for extraction, did you Illumina sequence and which depth? => after reading further I realize that this information is given later. Maybe rephrase the sentence to "Eggs, larva, nymphs, adult males and females were collected for later mRNA extraction."

Response: We changed the sentence according to the reviewer's suggestion.

- *"HMW gDNA was size-selected using Short Read Eliminator" change to "...using the Short..."*

Response: We changed the sentence according to the reviewer's suggestion.

- *"following protocol, using the Oxford Nanopore" remove "using".*

Response: We changed the sentence by "The ONT library was prepared with the Oxford Nanopore SQK-LSK109 kit, according to the following protocol".

- *Dovetail "The one-third of the cryoground" change to "Another third of the ...."*

Response: We changed the sentence according to the reviewer's suggestion.

- *"obtained 138 scaffolds. Only scaffolds larger than 35kb were kept resulting in a final assembly of 112 scaffolds" why >35kb?*

Response: It corresponds to the size of the shortest Nanopore read supported by the NECAT assembler.

- *Transcriptome "The contigs 5' and 3'ends were cleaned." Using which tool and parameters?*

Response: We did not use a specific tool to clean the contigs. Now, we precise the way the contigs are cleaned as follow "The first five bases of the contigs 5' and 3' ends containing N's were removed"

- *Why did you use Velvet and Oases especially since they are prone to create chimeras? Why not use Trinity for example?*

Response: Trinity was tested, it produced the shortest contigs and the gene prediction using the Trinity contigs had a lower BUSCO score.

- *Change "the research of ORFs" to "the identification of ORFs".*

Response: We changed the sentence according to the reviewer's suggestion

- *Change "aims to split contigs sequences" either to "aims to split contigs with different..." or "aims to split contig sequences"*

Response: We changed the sentence according to the second reviewer's suggestion.

- *Do you make the contig splitting tool available somewhere? Is the code accessible on github or in*

the supplement? If so, add reference.

Response: To perform this task, we used a custom in-house perl script that is not portable on other platforms. However, for reproducibility of our results, we provide the details of the method: Reads were mapped to the contigs with BWA-mem (Li *et al*., 2009) and the consistent paired-end reads were selected. Chimeric contigs were identified and split (uncovered regions) based on coverage information from consistent paired-end reads. Moreover, open reading frames (ORF) and domains were searched using respectively TransDecoder (Haas *et al.*, 2013) and CDDsearch (Marchler-Bauer *et al.*, 2011). We only allowed breaks outside ORF and domains. Finally, the read strand information was used to correctly orient the RNA-seq contigs. This method is now clearly explained as part of the Materials and Methods section.

- *"BLAT (version 36 with default parameter) matches with score" change to "with a score"*

Response: We changed the sentence according to the second reviewer's suggestion Results and Discussion.

- *Table 1 compares the newest version of Tmolitor with the older and Tcastaneum but on contig scale. Why contig and not scaffold, or chromosome scale?*

Response: We compared the scaffolds in both cases, we replaced the word "contig" by "scaffold" in table 1.

- *"of 99.5% (using version 5.0.0 with Insecta database odb10) while the predicted genome size was 310 Mb (Figure S1, Extended data)" remove "while the predicted genome size was 310 Mb (Figure S1, Extended data)" from this sentence. You give this information at the beginning of the paragraph already.*

Response: This element has been removed

- *"Furthermore, mitochondrial genome" change to " "the mitochondrial genome".*

Response: This has been changed.

- *"More precisely, the mitochondrial genome of T. molitor (15,785 bp) was aligned to the assembly using Minimap245.": This whole mt paragraph is confusing. Do you mean you aligned an existing mt genome, from the previous assembly? From which other study? => add reference and/or accession number. And you aligned it to identify the mt genome in your assembly. Why is the mt genome split across scaffolds? It should be on single contig only? => It sounds like misassembly and should be corrected. I.e. the mt parts should be assembled into a separate mt genome and should be removed from the scaffolds.*

Response: We aligned this mitochondrial genome https://www.ncbi.nlm.nih.gov/nuccore/KF418153 to our assembly (in the original article, "mitochondrial genome" is a link leading to the ncbi page). The mitochondrial genome is found in scaffold_94 three times. In general, mitochondrial genomes are more covered by reads than other regions. So, the assembler may have estimated multiple copy. As for scaffold_65, it is not a

completely mitochondrial sequence. It looks more like an insertion of the mitochondrial genome has occurred in the nuclear genome which is commonly observed. Thus, the scaffold_65 was kept in the genome assembly. The misassembly on scaffold 94 could be resolved by choosing one single copy (the one in the middle of the scaffold having the best match-length and identity percentage to the mitochondrial genome of NCBI). However, in that way, we would not be able to assemble it with the part of 7kb. So, we prefer to simply remove this scaffold from the assembly.

- *"a total of 21,435 genes which is higher than the number observed in T. castaneum": could you discuss why you think you have more genes but lower Busco (Table2)?*

Response: Indeed, we observe more genes but a lower busco score in *T. molitor*. By analysing the missing busco genes in Tenebrio, we found that they were found in Tribolum isoforms. As we do not propose isoforms in the gene prediction, it is possible that we oversee these genes.

- *Change "future works" to "future work".*

Response: This has been modified

Competing Interests: No competing interests were disclosed.