G3, 2023, 13(6), jkad079
https://doi.org/10.1093/g3journal/jkad079
Advance Access Publication Date: 1 April 2023
Genome Report

Whole genome assemblies of *Zophobas morio* and *Tenebrio molitor*

Sabhjeet Kaur, Sydnie A. Stinson, George C. diCenzo*

Department of Biology, Queen's University, 116 Barrie Street, Kingston, Ontario K7L 3N6, Canada

*Corresponding author. Email: george.dicenzo@queensu.ca

Abstract

Zophobas morio (=*Zophobas atratus*) and *Tenebrio molitor* are darkling beetles with industrial importance due to their use as feeder insects and their apparent ability to biodegrade plastics. High quality genome assemblies were recently reported for both species. Here, we report additional independent *Z. morio* and *T. molitor* genome assemblies generated from Nanopore and Illumina data. Following scaffolding against the published genomes, haploid assemblies of 462 Mb (scaffold N90 of 16.8 Mb) and 258 Mb (scaffold N90 of 5.9 Mb) were produced for *Z. morio* and *T. molitor*, respectively. Gene prediction led to the prediction of 28,544 and 19,830 genes for *Z. morio* and *T. molitor*, respectively. Benchmarking Universal Single Copy Orthologs (BUSCO) analyses suggested that both assemblies have a high level of completeness; 91.5 and 89.0% of the BUSCO endopterygota marker genes were complete in the *Z. morio* assembly and proteome, respectively, while 99.1 and 92.8% were complete in the *T. molitor* assembly and proteome, respectively. Phylogenomic analyses of four genera from the family Tenebrionidae yielded phylogenies consistent with those previously constructed based on mitochondrial genomes. Synteny analyses revealed large stretches of macrosynteny across the family Tenebrionidae, as well as numerous within-chromosome rearrangements. Finally, orthogroup analysis identified ~28,000 gene families across the family Tenebrionidae, of which 8,185 were identified in all five of the analyzed species, and 10,837 were conserved between *Z. morio* and *T. molitor*. We expect that the availability of multiple whole genome sequences for *Z. morio* and *T. molitor* will facilitate population genetics studies to identify genetic variation associated with industrially relevant phenotypes.

Keywords: *Tenebrio molitor*, *Zophobas morio*, *Zophobas atratus*, mealworm, superworm, darkling beetle, Tenebrionidae, whole genome sequence

Introduction

The order Coleoptera is the largest order of the animal kingdom, accounting for 25% of known animal species and 40% of known insect species. Better known as beetles, the order Coleoptera consists of over 360,000 known species (Audisio *et al.* 2015), with some researchers estimating the true number of beetle species to be 1.5 million (Stork *et al.* 2015). Despite the taxonomic richness of this order, beetles are under-represented in genome sequencing projects. A recent meta-analysis noted that the phylum Arthopoda is the most under-sequenced animal phylum, with the order Coleoptera being the most under-sequenced order within the phylum (Hotaling, Kelley, *et al.* 2021). A separate study found that the number of publicly available genome sequences for beetle species is only 20% the expected amount given the total number of available insect genomes (Hotaling, Sproul, *et al.* 2021). Significant efforts are therefore required to increase the number of available beetle genome sequences if the Earth BioGenome Project is to succeed (Lewin *et al.* 2018, 2022).

Tenebrionidae, commonly known as darkling beetles, is a family within the order Coleoptera that is estimated to contain ~20,000 species (Bouchard *et al.* 2017). To date, the genomes of seven species within the family Tenebrionidae have been sequenced: *Tribolium castaneum* (Tribolium Genome Sequencing Consortium 2008; Herndon *et al.* 2020), *Tribolium madens*, *Tribolium freemani*, *Tribolium confusum*, *Asbolus verrucosus*, *Tenebrio molitor* (Eriksson *et al.* 2020; Eleftheriou *et al.* 2022), and *Zophobas morio* (=*Zophobas atratus*). The latter two species are of particular interest due to their industrial relevance. Commonly known as mealworms and superworms, respectively, the larvae of *T. molitor* and *Z. morio* are routinely used as feeder insects for pet reptiles and fish. They are also consumed by humans in some cultures (Ramos-Elorduy 2009), with additional societies becoming increasingly interested in their use in aquafeed and human food products as an alternative to other animal products (van Huis 2013; Ribeiro *et al.* 2018; Rumbos and Athanassiou 2021). Interestingly, several recent studies have provided evidence that mealworms and superworms have the potential to biodegrade various types of plastics (Yang *et al.* 2015a; Brandon *et al.* 2018; Peng *et al.* 2020; Yang *et al.* 2020). While the mealworm and superworm gut microbiota appear to play an important role in the breakdown of plastic polymers (Yang *et al.* 2015b; Peng *et al.* 2020), insect-encoded enzymes may also contribute to this process (Yang *et al.* 2021).

As the current study was in progress, high quality genome assemblies were made publicly available for *T. molitor* (Eleftheriou *et al.* 2022) and *Z. morio*, and annotations were provided for the

Received: January 15, 2023. Accepted: March 29, 2023

T. molitor genome assembly. Here, we report independent genome assemblies and genome annotations for both *T. molitor* and *Z. morio*. The availability of genome annotations for both organisms will support future efforts to understand the mechanisms underlying plastic degradation by these insects. Likewise, the presence of multiple genome assemblies will facilitate population genetic studies that may be of value to breeders.

Materials and methods

Insect rearing and tissue collection

T. molitor and *Z. morio* larvae were purchased from a local pet store in Kingston, ON, Canada. The insect larvae were fed a diet of oatmeal, wheat bran, and carrot until harvest. For isolation of DNA, larvae were flash frozen in liquid nitrogen, and the head and legs of a single larvae per species were isolated for genomic DNA isolation. The collected tissues were crushed using sterile pestles in 1.5-mL tubes. For isolation of RNA, larvae were placed in 70% ethanol, following which their exoskeleton was cut open, and the digestive tract removed. Digestive tracts were immediately flash frozen and stored at −80°C until use.

DNA isolation and whole genome sequencing

For both *Z. morio* and *T. molitor*, crushed tissue from the head and legs of a single larvae was used for DNA extraction with Monarch Genomic DNA Purification Kits (New England Biolabs) following the manufacturer's instructions. Briefly, insect tissue was resuspended in 10-μL Proteinase K, 20-μL EDTA, and 200 μL of the tissue lysis buffer supplied with the kit. The samples were incubated for 3 hours at 56°C with brief vortexing every 15 minutes, followed by treatment with RNase A for 5 minutes. Following binding of the genomic DNA to the silica membrane in the column, DNA was eluted with 100 μL of elution buffer and stored at −20°C until further processing.

Prior to Nanopore sequencing, the *Z. morio* DNA sample was size selected using a BluePippin instrument and a 0.75% agarose gel with the high pass protocol. The sample collected from the BluePippin instrument was cleaned-up and concentrated using an equal volume of AMPure XP Beads (Beckman Coulter), with the DNA eluted in in 60-μL DNAse and RNAse free water. Size selection was not performed for the *T. molitor* DNA sample. DNA samples were adjusted to a concentration of 20 ng/μL, following which library preparation was performed using a Ligation Sequencing kit (SQK-LSK100; Oxford Nanopore Technologies) following the manufacturer's instructions. Sequencing was performed using a minION with R9.4.1 flow cells and the MinKNOW software. Basecalling was performed using GPU-enabled Guppy version 5.011+ 2b6dbffa5 and the high accuracy model (Oxford Nanopore Technologies).

Illumina sequencing and library preparation were performed at Génome Québec (Montréal, QC, Canada). Library preparation was performed using NxSeq AmpFREE Low DNA Fragment Library Kits (Lucigen) following the manufacturer's instructions. Samples were then sequenced with 150-bp paired-end technology using a S4 flow cell on an Illumina NovaSeq 6000 instrument. Illumina reads were trimmed using platanus_trim version 1.0.7 (Kajitani *et al.* 2014).

RNA isolation and sequencing

Total insect and microbial RNA was isolated from the digestive tracts of either one *Z. morio* or two *T. molitor* larvae using ZymoBIOMICS RNA miniprep kits (Zymo Research), following the manufacturer's instructions.

RNA depletion, library preparation, and Illumina sequencing were performed at Génome Québec (Montréal, QC, Canada). Depletion of rRNA was performed using FastSelect kits (Qiagen) with probes from both the -rRNA fly and 5S/16S/23S rRNA kits, following the manufacturer's instructions. However, the fly (*Drosophila melanogaster*) rRNA probes were not efficient at removal of *Z. morio* or *T. molitor* rRNA as indicated by most of the sequencing reads mapping to beetle rRNA loci. Library preparation was completed using NEBNext Ultra II Directional RNA Library Prep kits (New England Biolabs), following the manufacturer's instructions. Samples were then sequenced with 150-bp paired-end technology using a S4 flow cell on an Illumina NovaSeq 6000 instrument. Illumina reads were filtered using BBduk version 38.96 (Bushnell 2014) and then trimmed using trimmomatic version 0.39 (Bolger *et al.* 2014) with the following parameters: LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36.

Genome assembly

Genome assemblies were constructed using a multistage process as represented visually in Fig. 1. First, Nanopore reads were assembled into draft assemblies using Flye version 2.9-b1768 (Kolmogorov *et al.* 2019) (Assembly 1 in Fig. 1). Flye assemblies were polished once using Racon version 1.4.22 (Vaser *et al.* 2017) with Nanopore reads aligned to the draft assemblies with minimap2 version 2.20-r1061 (Li 2018). Assemblies were further polished using Medaka version 1.4.1 (Oxford Nanopore Technologies) and the Nanopore reads. The assemblies were then polished using Pilon version 1.24 (Walker *et al.* 2014) with the trimmed Illumina reads mapped to the assemblies using bowtie2 version 2.4.4 (Langmead and Salzberg 2012) and processed with samtools version 1.12 (Li *et al.* 2009). A final round of polishing was performed using Hapo-G version 1.2 (Aury and Istace 2021) with the trimmed Illumina reads mapped to the assemblies using bowtie2.

In parallel, hybrid assemblies were constructed using MaSuRCA version 4.0.3 (Zimin *et al.* 2017) with the Nanopore and trimmed Illumina reads (Assembly 2 in Fig. 1). The MaSuRCA assemblies were polished once using Pilon with the trimmed Illumina reads mapped to the assemblies with bowtie2. A second round of polishing was performed using Hapo-G with the trimmed Illumina reads mapped to the assemblies with bowtie2.

The Flye and MaSuRCA assemblies were merged using quickmerge version 0.3 (Chakraborty *et al.* 2016) and NUCmer version 4.0.0 rc1 (Kurtz *et al.* 2004). Quickmerge was run using a minimum alignment length of 10,000 nucleotides and a length cutoff for anchor contigs approximately equal to the N50 of the self assembly. In addition, quickmerge was run twice for each insect species; once with the Flye assembly as the "self" assembly (Assembly 3 in Fig. 1) and once with the Flye assembly as the "hybrid" assembly (Assembly 4 in Fig. 1). The Flye and MaSuRCA assemblies were also merged using RagTag patch version 2.1.0 (Alonge *et al.* 2022) and NUCmer. As with quickmerge, RagTag patch was run twice for each insect species: once with the Flye assembly as the "query" sequence (Assembly 5 in Fig. 1) and once with the Flye assembly as the "target" sequence (Assembly 6 in Fig. 1).

Haplotigs were purged from the Flye and MaSuRCA assemblies using purge_dups version 1.2.5 (Guan *et al.* 2020), with the self–self alignment performed using minimap2 version 2.18-r1015 (Assemblies 11 and 12 in Fig. 1). The Flye and MaSuRCA assemblies purged of haplotigs were then merged using quickmerge and RagTag patch as described above (Assemblies 13–16 in Fig. 1). Purge_dups was then used to remove haplotigs from all

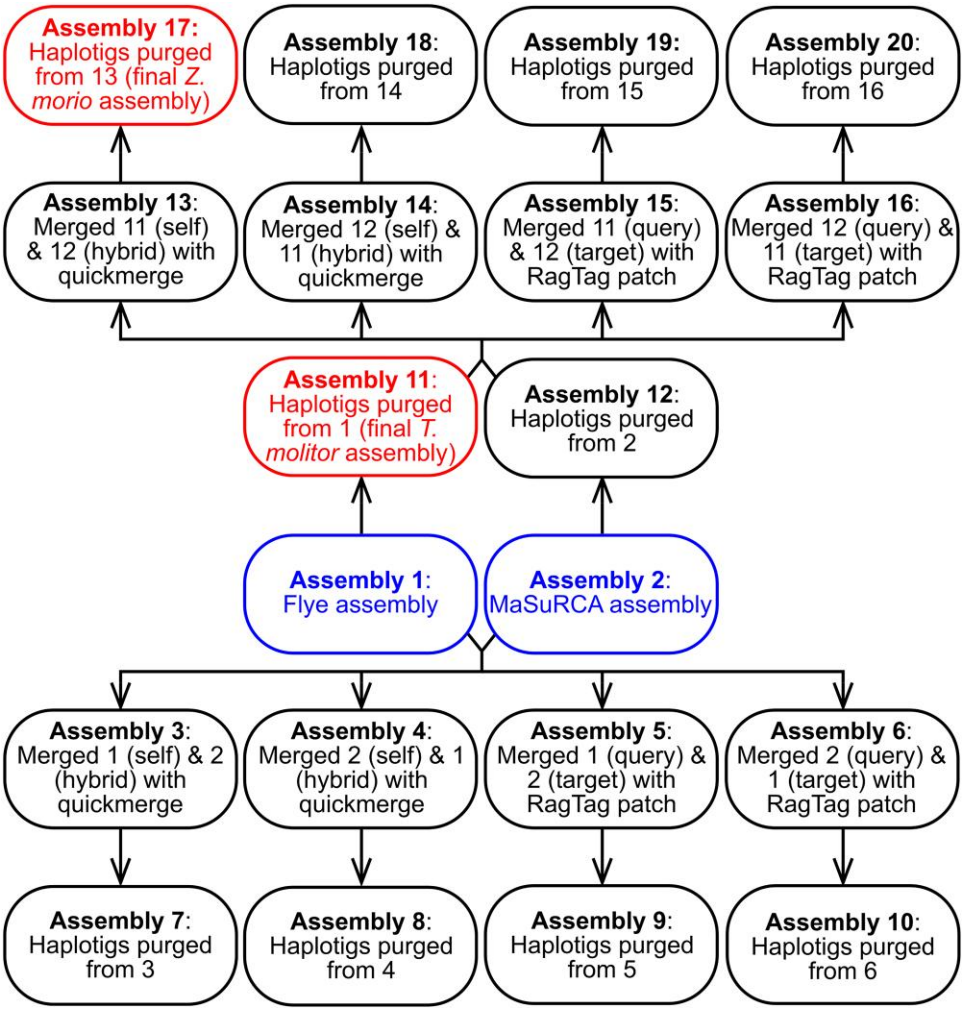

Fig. 1. Genome assembly workflow. A flowchart is provided depicting the overall genome assembly strategy of this study. Boxes represent the 20 assemblies created per species. Arrows indicate the flow of assemblies from one step to the next; for example, Assembly 3 was created from a combination of Assembly 1 and Assembly 2. The initial assemblies (Assemblies 1 and 2) are shown in blue. The assemblies chosen for annotation and downstream analyses (Assemblies 11 and 17) are shown in red.

merged assemblies created from either the full or purged assemblies (Assemblies 7–10 and 17–20 in Fig. 1). Finally, all *T. molitor* assemblies were scaffolded using RagTag scaffold and a previously published *T. molitor* assembly [European Nucleotide Archive (ENA) accession PRJEB44755] (Eleftheriou *et al.* 2022), whereas the *Z. morio* assemblies were scaffolded against a publicly available *Z. atratus* assembly (GenBank accession GCA_022388445.1).

Assembly statistics and measures of genome completeness and redundancy were performed for all assemblies as described below, and the results used to select one assembly to move forward. For *T. molitor*, the purged Flye assembly (Assembly 11 in Fig. 1) was selected for annotation. For *Z. morio*, the assembly chosen for annotation was the one created by using purge_dups on the assembly produced by running quickmerge with the purged MaSuRCA and the purged Flye assemblies as the hybrid and self assemblies, respectively (Assembly 17 in Fig. 1).

The mitochondrial genomes of the *Z. morio* and *T. molitor* assemblies were identified by querying the assemblies using BLASTn version 2.10.1+ and previously published *Z. morio* (GenBank accession: MK140669.1) (Bai *et al.* 2019) or *T. molitor* (GenBank accession: KF418153.1) (Liu and Wang 2014) mitochondrial sequences, respectively. The searches revealed that scaffolding led to the *Z. morio* mitochondrion being merged with a nuclear DNA contig; the merged contig was therefore broken to isolate the mitochondrion as a single contig, and one copy of the overlap between the two ends of the mitochondrial contig was removed.

Submission of the genome assemblies to National Center for Biotechnology (NCBI) led to the identification of contamination in the assemblies, including the presence of adapter and microbial sequences. Contaminating sequences were removed from the *T. molitor* and *Z. morio* assemblies, and scaffolds were split at sites of contamination. Following decontamination, the nuclear genome was again scaffolded against the appropriate reference assembly as described above.

Genome annotation

Low complexity regions of the genome were masked in a multistep fashion. First, RepeatMasker version 4.1.2-p1 (Tarailo-Graovac and Chen 2009) was run on each assembly using the rmblast version 2.11.0 search engine (Tarailo-Graovac and Chen 2009), Tandem Repeats Finder version 4.0.9 (Benson 1999), species designation of "Insecta", and the Dfam version 3.2 (Hubley *et al.* 2016) and RepBase RepeatMasker edition 20181016 (Bao *et al.* 2015) databases. Then, sdust version 0.1-r2 (Li 2018) was run on each assembly. Finally, for each assembly, regions

masked by sdust but not masked by RepeatMaster were soft-masked in the fasta file returned by RepeatMasker using BEDtools version 2.26.0 (Quinlan and Hall 2010).

Following masking, the RNAseq reads were mapped to the genome assemblies using STAR version 2.7.10a (Dobin *et al.* 2013) and a two-pass procedure (Veeneman *et al.* 2016). A multistep gene prediction was then performed using BRAKER version 2.1.6 (Hoff *et al.* 2016, 2019; Brůna *et al.* 2021). Gene prediction was first performed using BRAKER with the softmasking option and the RNAseq alignment file produced by STAR. A second, independent gene prediction was then performed using BRAKER with the softmasking option and a protein database consisting of (1) the single-copy and multi-copy complete Endopterygota Benchmarking Universal Single-Copy Orthologs (BUSCO) genes identified in the assembly to be annotated, (2) proteins from the *T. castaneum* genome annotation (ENA accession PRJNA12540) (Tribolium Genome Sequencing Consortium 2008), and (3) proteins from a previously published *T. molitor* genome annotation (ENA accession PRJEB44755) (Eleftheriou *et al.* 2022). Subsequently the two gene prediction files were combined and filtered using TSEBRA version 1.0.3 (Gabriel *et al.* 2021) with default settings except for the intron_support parameter being set to 0.2. BRAKER dependencies included: samtools version 1.15-8-gbdc5bb8, bamtools version 2.5.2 (Barnett *et al.* 2011), the GeneMark suite version 4.69 (Lomsadze 2005, 2014; Brůna *et al.* 2020), DIAMOND version 2.0.13 (Buchfink *et al.* 2015), AUGUSTUS version 3.4.0 (Stanke *et al.* 2006, 2008), and Spaln version 2.3.3d (Gotoh 2008; Iwata and Gotoh 2012).

Following TSERBA, coding regions fully contained within another coding region on the same DNA strand were removed from the GFF annotation files. Likewise, the smaller of two overlapping genes on the same strand were removed. The GFF files were sorted and tidied using GenomeTools version 1.5.10 (Gremme *et al.* 2013) and converted to GenBank format with table2asn (NCBI). For both genome assemblies, protein fasta files containing all predicted isoforms were created by extracting the protein sequences from the GenBank files.

Quality assessment of the genome assemblies

Genome assembly statistics were determined using the stats.sh function of BBmap version 38.90 (Bushnell 2014). Genome completeness metrics were calculated using BUSCO version 5.1.2 (Manni *et al.* 2021) with the eukaryota_odb10 and endopterygota_odb10 datasets with MetaEuk version 4-a0f584d (Levy Karin *et al.* 2020), HMMER version 3.2.1 (Eddy 2009), and BLAST+ version 2.12.0 (Camacho *et al.* 2009). Kmer QV was estimated using Yak version 4bdd51d (github.com/lh3/yak).

Estimation of genome heterozygosity

The fastq files containing the forward and reverse Illumina sequencing reads were concatenated as a single file and used as input for Jellyfish version 2.3.0 (Marçais and Kingsford 2011) to count canonical kmer sizes with a size of 21 and create a kmer count histogram. The output of Jellyfish was passed to the GenomeScope webserver (qb.cshl.edu/genomescope/) (Vurture *et al.* 2017) to estimate genome heterozygosity.

Comparative genomics

Orthofinder version 2.5.4 (Emms and Kelly 2015; Emms and Kelly 2019) was used to group Tenebrionidae proteins into orthogroups with the BLAST search engine and the -y option. As input, Orthofinder was provided proteins annotated in the two genome assemblies produced in this study (i.e. *T. molitor* and *Z. morio*), as well as the proteins annotated in the genome assemblies for *A. verrucosus* (GenBank accession GCA_004193795.1), *T. castaneum* (RefSeq accession GCF_000002335.3), *T. madens* (RefSeq accession GCF_015345945.1), and a previously published *T. molitor* genome (GenBank accession GCA_907166875.3). The other Tenebrionidae genomes available through NCBI Genome database were not included as they lacked annotations. Orthofinder dependencies included: BLAST version 2.10.1+, DendroBLAST (Kelly and Maini 2013), fastME version 2.1.4 (Lefort *et al.* 2015), MCL clustering (Van Dongen 2000), and the ETE tree library (Huerta-Cepas *et al.* 2016). A distance matrix of the proteomes was then computed using Jaccard distances and the "distance" function of the R package "philentropy" (Drost 2018), following which a dendogram was constructed using the "bionj" function of the R package "ape" (Popescu *et al.* 2012).

Synteny between Tenebrionidae genomes was detected and visualized using the D-Genies version 1.3.0 webserver (dgenies.toulouse.inra.fr) (Cabanettes and Klopp 2018). First, scaffolds larger than 1 Mb were extracted from each assembly of interest using pullseq version 1.0.2 (github.com/bcthomas/pullseq) and then sorted by length in descending order using seqkit version 2.2.0 (Shen *et al.* 2016). Pairwise analyses were then performed with D-Genies using the minimap2 version 2.24 aligner and the "many repeats" option. Dot plots were constructed for the *Z. morio* and *T. molitor* genome assemblies produced in the current work, previously published *Z. morio* (GenBank accession GCA_022388445.1) and *T. molitor* (GCA_907166875.3) genome assemblies, as well as genome assemblies for *T. madens* (GCA_015345945.1), *T. castaneum* (GCA_000002335.3), *T. freemani* (GCA_022388455.1), and *T. confusum* (GCA_019155225.1). The published genome assembly of *A. verrucosus* was excluded as all contigs were shorter than 1 Mb.

To detect SNPs between the *T. molitor* and *Z. morio* assemblies produced in this and previous studies, the assemblies were first aligned using NUCmer with the options "–minmatch 100 –mincluster 1000 –diagfactor 10 –banded –diagdiff 5". The output of NUCmer was then fed into the show-snps function of the MUMmer package, run with the "-Clr" options.

Phylogenomic analysis

A phylogenomic tree was constructed to show the evolutionary relationships between *T. molitor* (this study and GenBank accessions GCA_907166875.3 and GCA_014282415.2), *Z. morio* (this study and GCA_022388445.1), *T. madens* (GCA_015345945.1), *T. castaneum* (GCA_000002335.3), *T. freemani* (GCA_022388455.1), *T. confusum* (GCA_019155225.1), and *A. verrucosus* (GCA_004193795.1). First, highly conserved genes were detected in all genomes using BUSCO version 5.3.0 with the eukaryota_odb10 database. A set of 155 single copy orthologs present in all 10 genome assemblies were identified, following which each set of orthologs was individually aligned using MAFFT version 7.310 with the localpair option (Katoh and Standley 2013) and trimmed using trimAl version 1.4rev22 and the automated1 option (Capella-Gutiérrez *et al.* 2009). Trimmed alignments were then concatenated and used as input to construct a maximum likelihood phylogeny with RAxML version 8.2.12 with the GAMMA model of rate heterogeneity and the JTT amino acid substitution model with empirical amino acid frequencies. The JTT model with empirical frequencies was chosen based on the results of a preliminary run of RAxML using automatic model selection. The final tree represents the bootstrap best true following 100 bootstrap replicates and was visualized using the iTol webserver (Letunic and Bork 2016).

Results and discussion

Z. morio and *T. molitor* novel genome assemblies

High quality genome assemblies were recently made publicly available for *T. molitor* (Eleftheriou *et al.* 2022) and *Z. morio* (GenBank accession GCA_022388445.1). Here, we independently performed whole genome sequencing and assembly of an additional individual for each species, obtained from a separate source. Genomic DNA isolated from single *Z. morio* and *T. molitor* individuals was sequenced using a Nanopore minION (*Z. morio*: 2.79 Gb with a N50 read length of 14,531 nt; *T. molitor*: 8.84 Gb with a N50 read length of 4,113 nt) and an Illumina NovaSeq to generate 150 bp paired-end reads (*Z. morio*: 220 Gb; *T. molitor*: 167 Gb). Multiple approaches to genome assembly were taken to increase the likelihood of producing a good-quality assembly (see *Materials and methods*). Variations to the assembly pipeline included: (1) performing either Nanopore-only assembly using Flye or hybrid assembly using MaSuRCA, (2) optionally merging the Flye and MaSuRCA assemblies using quickmerge or RagTag, and (3) optionally purging haplotigs using purge_dups. Following scaffolding against existing genome assemblies (see *Materials and methods*), a single assembly to move forward to annotation was chosen for each species based on the genome size, contig and scaffold N50, and BUSCO scores as a measure of assembly completeness and redundancy (Supplementary Table 1 in File 2).

Using the above approach, haploid genome representations of 462 Mb (scaffold N90: 16.8 Mb) and 258 Mb (scaffold N90: 5.9 Mb) were produced for *Z. morio* and *T. molitor*, respectively (Table 1). By comparison, the existing *Z. morio* and *T. molitor* genome assemblies are 478 Mb and 288 Mb, respectively, suggesting that some repetitive regions were collapsed in our assemblies. Alignment of the published *Z. morio* and *T. molitor* mitochondrial sequences (GenBank accessions MK140669.1 and KF418153.1) identified full-length mitochondrial genomes as single contigs in both assemblies.

Estimates of genome quality indicated that both the *Z. morio* and *T. molitor* assemblies are of high quality. Using Yak and the Illumina data, the accuracies of the haploid assemblies were estimated at QV36 and QV31 for the *Z. morio* and *T. molitor* assemblies, respectively. Additionally, 98.1% of the BUSCO endopterygota marker genes were identified as single-copy, complete genes in the *T. molitor* assembly (Table 1), which is comparable to the value of 98.4% in the existing *T. molitor* assembly. Likewise, 90.1% of the BUSCO endopterygota marker genes were identified as single-copy, complete genes in the *Z. morio* assembly (Table 1), which is slightly lower than the value of 94.5% in the existing *Z. morio* assembly.

Z. morio and *T. molitor* genome annotations

Gene prediction was performed using the BRAKER pipeline (Hoff *et al.* 2016; Hoff *et al.* 2019; Brůna *et al.* 2021) and a combination of gut RNA-seq data and a protein database (see *Materials and methods*). This process resulted in 28,544 and 19,830 predicted genes for *Z. morio* and *T. molitor*, respectively (Table 2). In comparison, a previous study predicted 21,435 genes in the *T. molitor* genome (Eleftheriou *et al.* 2022); to date, there are no other reports of *Z. morio* gene predictions. BUSCO analyses of the proteomes indicated that the gene predictions are good quality; 89 and 93% of the endopterygota marker genes were identified in the *Z. morio* and *T. molitor* proteomes, respectively, compared to 96% in the previous *T. molitor* annotation (Table 2). Likewise, 94 and 95% of the eukaryota marker genes were identified in the *Z. morio* and *T. molitor* proteomes, respectively, compared to 96% in the previous *T. molitor* annotation (Supplementary Table 2 in File 2). The higher frequency of multi-copy BUSCO marker proteins in our proteomes likely reflects the inclusion of alternate isoforms, whereas alternate isoforms were not included in the previously-reported *T. molitor* annotation.

Gene summary statistics for *Z. morio* and *T. molitor* revealed several similarities. The median exon length, median intron length, median coding sequence (CDS) length, and median exon count per gene were similar in both species (Table 2). Likewise, the number of alternate transcripts per gene is comparable in the two species: ~1.07 transcripts per gene in *Z. morio* compared to ~1.08 transcripts per gene in *T. molitor*. On the other hand, the median gene length in our *T. molitor* assembly (1,319) is ~21% longer than the median gene length (1,087) in our *Z. morio* assembly; however, this pattern does not hold true when comparing the median *Z. morio* gene length to the median gene length (1,147) in the previous *T. molitor* annotation. Additionally, whereas *Z. morio* encodes more genes than does *T. molitor*, the coding density of *T. molitor* (~77 genes per Mb in our assembly) is higher than that of *Z. morio* (~62 genes per Mb). Lastly, the percentage of genes lacking introns is higher in *Z. morio* (~41%) relative to *T. morio* (between 23 and 29%, depending on the *T. morio* annotation) (Table 2).

Repetitive DNA and low complexity DNA were annotated in the *Z. morio* and *T. molitor* genome assemblies using RepeatMasker

Table 1. *Z. morio* and *T. molitor* genome assembly statistics.

	Zophobas morio (This study)	*Zophobas morio* (GCA_022388445.1)	*Tenebrio molitor* (This study)	*Tenebrio molitor* (GCA_907166875.3)
Cumulative scaffold length (bp)	461,985,688	478,145,070	258,289,333	287,839,991
Number of scaffolds	4,179	562	1,987	111
Number of contigs	8,884	937	6,142	167
G + C content (%)	33.34	33.44	36.17	36.72
Number of Ns	470,362	37,038	416,000	28,500
Largest scaffold (Mb)	74.58	78.287	32.201	33.043
Scaffold N50 (Mb)	48.007	53.051	20.827	21.886
Scaffold L50	4	4	6	6
Largest contig (Mb)	1.58	37.237	2.345	20.298
Contig N50 (Mb)	0.228	6.738	0.205	6.327
Contig L50	535	15	300	13
BUSCO complete single-copy[a]	1914	2008	2083	2101
BUSCO complete multi-copy[a]	30	26	22	10
BUSCO fragmented[a]	39	28	5	3
BUSCO missing[a]	141	62	14	10

[a] Determined using the BUSCO endopterygota_odb10 dataset.

Table 2. *T. molitor* and *Z. morio* gene prediction statistics.

	Zophobas morio (This study)	*Tenebrio molitor* (This study)	*Tenebrio molitor* (GCA_907166875.3)[a]
Number of genes	28,544	19,830	21,435
Number of transcripts	30,408	21,405	Not reported
Cumulative length of CDSs (bp)[b]	34,142,077	26,175,824	25,230,147
Median gene length (bp)[c]	1,087	1,319	1,147
Median CDS length (bp)	813	876	783
Median exon length (bp)	193	190	NR
Median intron length (bp)	54	54	55
Number of intronless genes	11,785	5,788	4,898
Median number of exons per gene	2	3	3
Median number of exons per multi-exon gene	5	5	4
BUSCO complete single-copy[d]	1781	1897	2019
BUSCO complete multi-copy[d]	108	75	15
BUSCO fragmented[d]	61	47	21
BUSCO missing[d]	174	105	69

[a] Values in this column were taken from Eleftheriou *et al.* (2022), except for the BUSCO scores.
[b] The cumulative CDS length includes the length of all isoforms for the two genomes reported in this study.
[c] Does not include 5′ or 3′ untranslated regions (UTRs).
[d] Determined using the BUSCO endopterygota_odb10 dataset.

Table 3. Repetitive elements identified in the *Z. morio* and *T. molitor* genome assemblies.

	Zophobas morio (This study)		*Zophobas morio* (GCA_022388445.1)		*Tenebrio molitor* (This study)		*Tenebrio molitor* (GCA_907166875.3)	
	Number	Length in bp (% genome)	Number	Length in bp (% genome)	Number	Length in bp (% genome)	Number	Length in bp (% genome)
1. Retroelements	29,549	10,528,828 (2.28)	34,685	13,407,425 (2.80)	12,374	4,520,998 (1.75)	17,663	8,414,714 (2.92)
1.1. SINEs	68	3,773 (0.00)	59	3,124 (0.00)	656	38,522 (0.01)	698	41,673 (0.01)
1.2. Penelope	369	63,623 (0.01)	435	79,243 (0.02)	97	22,953 (0.01)	117	37,207 (0.01)
1.3. LINEs	22,178	6,512,833 (1.41)	25,852	8,034,913 (1.68)	8,261	2,607,205 (1.01)	12,012	4,754,646 (1.65)
1.4. LTR elements	7,303	4,012,222 (0.87)	8,774	5,369,388 (1.12)	3,457	1,875,271 (0.73)	4,953	3,618,395 (1.26)
2. DNA transposons[a]	30,175	7,921,967 (1.71)	36,209	9,174,416 (1.92)	29,324	5,065,342 (1.96)	33,229	6,333,033 (2.20)
2.1. hobo-Activator	5,783	1,218,848 (0.26)	7,094	1,480,673 (0.31)	2,383	424,087 (0.16)	3,040	615,254 (0.21)
2.2. Tc1-IS630-Pogo	10,134	1,962,296 (0.42)	11,406	2,244,033 (0.47)	14,715	2,439,817 (0.94)	15,910	2,694,117 (0.94)
2.3. PiggyBac	321	107,250 (0.02)	363	127,592 (0.03)	211	56,619 (0.02)	268	71,450 (0.02)
2.4. Tourist/ Harbinger	121	27,527 (0.01)	152	36,067 (0.01)	221	58,904 (0.02)	254	66,996 (0.02)
2.5. Other	346	67,694 (001)	497	100,900 (0.02)	339	58,038 (0.02)	368	67,930 (0.02)
3. Other interspersed repeats	14,360	1,783,621 (0.39)	16,697	2,138,909 (0.45)	3,602	445,645 (0.17)	4,114	511,359 (0.18)
3.1. Helitrons	12,205	1,607,354 (0.35)	14,289	1,939,181 (0.41)	3,022	383,582 (0.15)	3,463	437,499 (0.15)
3.2. Unclassified	2,155	176,267 (0.04)	2,408	199,728 (0.04)	580	62,063 (0.02)	651	73,860 (0.03)
4. Simple DNA	126,697	6,570,028 (1.42)	129,837	6,989,302 (1.46)	58,729	2,909,186 (1.13)	62,276	2,924,775 (1.02)
4.1. Satellites	94	50,568 (0.01)	102	22,789 (0.00)	667	388,913 (0.15)	424	262,889 (0.09)
4.2. Simple repeats	103,420	5,081,071 (1.10)	106,366	5,434,824 (1.14)	46,170	1,900,043 (0.74)	49,952	2,089,046 (0.73)
4.3. Low complexity	23,183	1,438,389 (0.31)	23,369	1,531,689 (0.32)	11,892	620,230 (0.24)	11,900	572,840 (0.20)
5. Low complexity—sdust[b]	695,028	44,147,235 (9.56)	709,088	45,874,388 (9.59)	240,473	4,105,278 (1.59)	257,144	4,330,299 (1.50)
6. Small RNA	279	67,945 (0.01)	284	80,965 (0.02)	226	25,318 (0.01)	2,353	903,298 (0.31)

[a] En-Spm or MuDR-IS905 DNA transposons were not identified in any of the genomes.
[b] All repetitive elements were identified using RepeatMaser, except for this row that was determined using sdust, which is a reimplementation of the symmetric DUST algorithm that gives the same output as NCBI's dustmasker.

(Tarailo-Graovac and Chen 2009) and sdust (Li 2018), respectively (Table 3). Interspersed repeats accounted for 5.17 and 5.30% of the previously reported *Z. morio* and *T. molitor* genomes, respectively. The values were somewhat lower for our *Z. morio* and *T. molitor* assemblies at 4.38 and 3.88%, respectively, further indicating that some repetitive regions were collapsed in our assemblies. This likely reflects differences in the lengths of the long reads used in our study and the previous studies, with our data lacking enough ultralong reads to span longer repeats. Here, the N50 read length for the Nanopore was 4,113 nt for *T. molitor* and 14,531 nt for *Z. morio*; in contrast, Eleftheriou *et al.* (2022) produced Nanopore reads with a N50 value of 34,818 nt.

In general, the abundance of DNA transposons, long terminal repeat (LTR) elements, and long interspersed nuclear elements (LINEs) is similar between the two species although the precise make-up varies modestly (Table 3 and Supplementary Table 3 in File 2). For example, when normalized by assembly length, *Tc1-IS630-Pogo* DNA transposons are ~100% abundant in *T. molitor*, whereas L2/CR1/Rex LINEs are ~300% more abundant in *Z. morio*. Likewise, the abundance of helitrons is modestly different between species, with them being ~170% more abundant in *Z. morio*. In contrast, the abundance of short interspersed nuclear elements (SINEs) differs dramatically and are >18-times more abundant in *T. molitor* when normalized by assembly length (Table 3). Likewise,

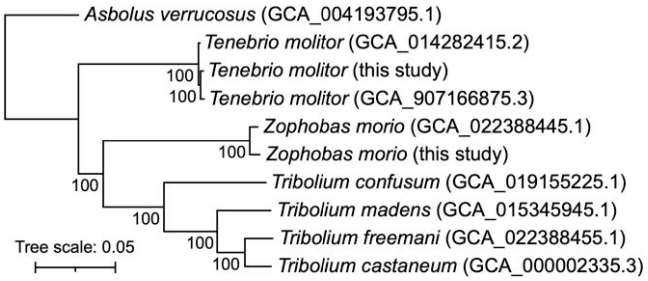

Fig. 2. Phylogeny of the family Tenebrionidae. An unrooted maximum likelihood phylogeny of the family Tenebrionidae is shown. The phylogeny is drawn with *Asbolus verrucosus* as the outgroup based on the rooted species tree returned by OrthoFinder. The phylogeny represents the bootstrap best tree following 100 bootstrap replicates, which was prepared using RAxML with the concatenated alignment of 155 single-copy orthologs encoded by all 10 genomes. Values at the nodes represent the bootstrap support, while the scale bar represents the mean number of amino acid substitutions per site.

satellite sequences are >12-times more abundant in *T. molitor* when normalized by assembly length (Table 3). The difference in satellite sequence abundance is driven by the prevalence of the 142 bp TMSATE1 satellite sequence in *T. molitor* (Petitpierre *et al.* 1988; Davis and Wyatt 1989), which we detected 505 times in our *T. molitor* assembly and 415 times in the existing *T. molitor* assembly. On the other hand, low complexity DNA, as identified with sdust, is ~540% more abundant in *Z. morio* compared to *T. molitor* (Table 3).

Z. morio and *T. molitor* genome sequence variation

Using the k-mer-based approach of Genomescope together with Illumina data generated from a single individual per species, genome-wide heterozygosity was estimated to be 0.93 and 0.94% in *Z. morio* and *T. molitor*, respectively. By comparison, a higher heterozygosity of 1.43% was previously reported for *T. molitor* (Eleftheriou *et al.* 2022), which could be a result of the individuals coming from different populations. In comparing our *T. molitor* assembly with that of Eleftheriou *et al.* (2022), we identified ~2.68 million SNPs, corresponding to ~1.0% of the genomes. The between individual sequence variability was lower for the two *Z. morio* genomes; ~2.39 million SNPs were identified, corresponding to ~0.5% of the genomes.

Comparative genomics of the family Tenebrionidae

A set of 155 single-copy orthologs was used to construct a maximum likelihood phylogeny to examine the evolutionary relationships between the seven Tenebrionidae species with sequenced genomes, representing four genera. The analysis suggests that the genera *Zophobas* and *Trilobium* are more closely related to each other than either are to the genus *Tenebrio*, and that the genus *Asbolus* is the most distantly related genus (Fig. 2). These observations are consistent with a previously presented phylogeny constructed from 13 mitochondrial genes (Bai *et al.* 2019).

Dot plot analyses revealed large stretches of macrosynteny across the *Z. morio* and *T. molitor* genome assemblies (Fig. 3). While many within-chromosome rearrangements were evident, no translocations were detected. Likewise, macrosynteny and within-chromosome rearrangement were observed when comparing the *Z. morio* or *T. molitor* assemblies to the genomes of four *Tribolium* spp. (Supplementary Fig. 1–4 in File 1). In addition, the previously reported whole-chromosome translocation within

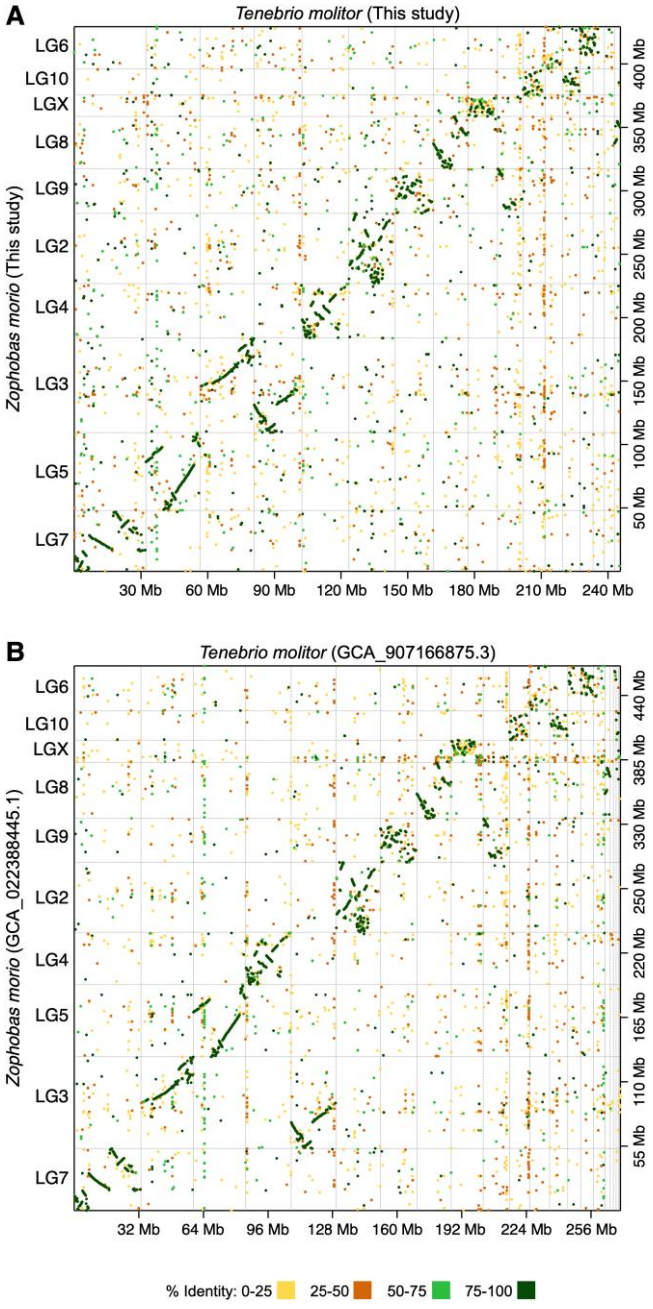

Fig. 3. Macrosynteny between the *Zophobas morio* and *Tenebrio molitor* genomes. Dot plots are shown comparing the *Z. morio* and *T. molitor* genomes of a) this study, or b) previously published sequences (GCA_022388445.1 and GCA_907166875.3). Dot plots were created using D-Genies with the minimap2 aligner. Prior to the dot plot analyses, genomes were filtered to remove scaffolds less than 1 Mb in length for visualization purposes. Dashed grey lines delineate scaffolds. The dot colors indicate the average percent identity of the match. Linkage groups have been added to the *Z. morio* scaffolds according to the previously published *Z. morio* genome assembly (GCA_022388465.1).

the *T. confusum* lineage was observed (Smith 1952; Samollow *et al.* 1983) (Supplementary Fig. 1–4 in File 1).

To explore genome evolution within the family Tenebrionidae, OrthoFinder was used to group proteins from the five sequenced and annotated species of the genera *Asbolus*, *Tenebrio*, *Zophobas*, and *Tribolium*. This analysis identified a core set of 7,738 gene families present in all of the annotated genomes (Fig. 4 and Supplementary Fig. 5 in File 1); this core set increases to 8,185

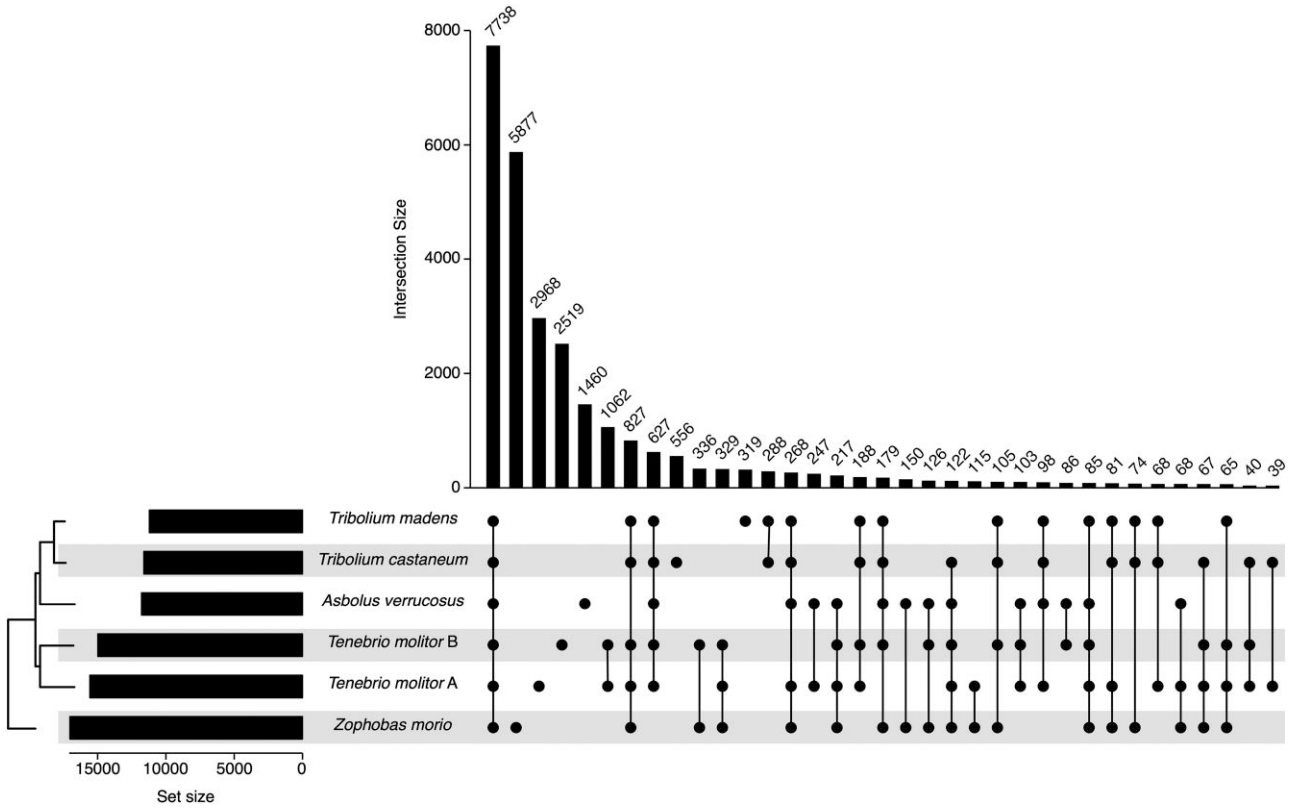

Fig. 4. Conservation of gene families across the family Tenebrionidae. Orthofinder was used to group the annotated proteins of *A. verrucosus* (GenBank accession GCA_004193795.1), *T. castaneum* (RefSeq accession GCF_000002335.3), *T. madens* (RefSeq accession GCF_015345945.1), *T. molitor* A (GenBank accession GCA_907166875.3), *T. molitor* B (this study), and *Z. morio* (this study) into gene families. Gene family conservation was summarized using UpSetR (Conway *et al.* 2017), and the 35 most abundant intersections are shown (see Supplementary Fig. 5 in File 1 for a version with all intersections). The set size shows the total number of gene families in a given proteome, while the intersect size shows the number of gene families conserved across the indicated proteomes. The midpoint rooted dendogram presented in the bottom left represents the clustering of the proteomes creating using a neighbor-joining approach and a distance matrix constructed from the gene family presence/absence data and Jaccard distances.

when including gene families present in at least one of the two *T. molitor* genome assemblies. The remaining 19,692 gene families were variably present in each genome, of which 14,761 gene-families (53% of all gene families) were species specific. In addition, a total of 10,837 gene families were conserved across *T. molitor* and *Z. morio*, when accounting for genes found in at least one *T. molitor* genome. Finally, a neighbor-joining tree constructed from a distance matrix of gene family presence/absence indicated that genome size was better correlated with proteome similarity than was phylogenetic relatedness (Fig. 4).

Conclusions

We report whole genome sequences for *Z. morio* (462 Mb; scaffold N90: 16.8 Mb) and *T. molitor* (258 Mb; scaffold N90: 5.9 Mb). The *Z. morio* genome is 80% larger than the *T. molitor* genome, in part due to an increase in interspersed repeats (20–25 Mb vs 10–15 Mb) and low complexity DNA (~45 Mb vs ~4 Mb), but also due to an increase in protein coding genes (28,544 vs 19,830 based on our gene predictions). Although many genomic rearrangements were detected, macrosynteny was observed between the *Z. morio* and *T. molitor* genomes, and 10,837 gene families were identified in both *Z. morio* and *T. molitor*.

We expect that the availability of multiple whole genome sequences for *Z. morio* and *T. molitor* will help facilitate future population genetics studies to identify genetic variation associated with industrially relevant phenotypes. In addition, these genome sequences will support studies of the microbiomes of darkling beetles by facilitating the removal of contaminating host DNA during metagenomic studies.

Data availability

Files 1 and 2 are available through FigShare at doi.org/10.6084/m9.figshare.21779096. All sequencing data generated in this study have been deposited to the NCBI under BioProject PRJNA820846. The Nanopore and Illumina DNA sequencing reads are available through the Sequence Read Archive (SRA) with the accession numbers SRR18507645, SRR18507646, SRR18507647, and SRR18507648. The Illumina RNA sequencing reads are available through the SRA with the accession numbers SRR18735291 and SRR18735292. Genome assemblies and annotations are available through the NCBI GenBank database with the accession numbers: JALNTZ000000000 and JALNUA00 0000000. Scripts to repeat the computational work reported in this manuscript are available at github.com/diCenzo-Lab/006_2022_Tenebrionidae_genomes.

Acknowledgments

We thank Zhengxin Sun for kindly performing BluePippin size selection of the genomic DNA isolated from *Z. morio*. We also thank Philippe Daoust and other personnel at Génome Québec for the advice on sequencing strategies and for performing all the

Illumina sequencing reported in this study. This research was enabled, in part, through computational resources provided by Compute Ontario (computeontario.ca) and Digital Research Alliance of Canada (alliancecan.ca).

Funding

This research was supported by the "Optimizing a microbial platform to break down and valorize waste plastic" project funded by the Government of Canada through Genome Canada and Ontario Genomics (OGI-207), the Government of Ontario through an Ontario Research Fund (ORF)—Large Scale Applied Research Project (LSARP) grant (File 18414), and the Imperial Oil Limited through a University Research Award.

Conflicts of interest statement

The author(s) declare no conflict of interest.

Literature cited

Alonge M, Lebeigle L, Kirsche M, Jenike K, Ou S, Aganezov S, Wang X, Lippman ZB, Schatz MC, Soyk S. 2022. Automated assembly scaffolding elevates a new tomato system for high-throughput genome editing. Genome Biol. 23:258. doi:10.1186/s13059-022-02823-7.

Audisio P, Alonso Zarazaga M-A, Slipinski A, Nilsson A, Jelínek J, Taglianti A, Turco F, Otero C, Canepari C, Kral D, *et al.* 2015. Fauna Europaea: Coleoptera 2 (excl. series Elateriformia, Scarabaeiformia, Staphyliniformia and superfamily Curculionoidea). BDJ. 3:e4750. doi:10.3897/BDJ.3.e4750.

Aury J-M, Istace B. 2021. Hapo-G, haplotype-aware polishing of genome assemblies with accurate reads. NAR Genom Bioinform. 3(2): lqab034. doi:10.1093/nargab/lqab034.

Bai Y, Wang H, Li G, Luo J, Liang S, Li C. 2019. Complete mitochondrial genome of the super mealworm *Zophobas atratus* (Fab.) (Insecta: Coleoptera: Tenebrionidae). Mitochondrial DNA B Resour. 4(1): 1300–1301. doi:10.1080/23802359.2019.1591237.

Bao W, Kojima KK, Kohany O. 2015. Repbase update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 6(1):11. doi: 10.1186/s13100-015-0041-9.

Barnett DW, Garrison EK, Quinlan AR, Stromberg MP, Marth GT. 2011. Bamtools: a C++ API and toolkit for analyzing and managing BAM files. Bioinformatics. 27(12):1691–1692. doi:10.1093/bioinformatics/btr174.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27(2):573–580. doi:10.1093/nar/27.2.573.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 30(15):2114–2120. doi:10.1093/bioinformatics/btu170.

Bouchard P, Smith ABT, Douglas H, Gimmel ML, Brunke AJ, Kanda K. 2017. Biodiversity of Coleoptera. In: Foottit RG, Adler PH, editors. Insect Biodiversity. Chichester, UK: John Wiley & Sons, Ltd. p. 337–417. [accessed 2022 Apr 10]. https://onlinelibrary.wiley.com/doi/10.1002/9781118945568.ch11.

Brandon AM, Gao S-H, Tian R, Ning D, Yang S-S, Zhou J, Wu W-M, Criddle CS. 2018. Biodegradation of polyethylene and plastic mixtures in mealworms (larvae of *Tenebrio molitor*) and effects on the gut microbiome. Environ Sci Technol. 52(11):6526–6533. doi:10.1021/acs.est.8b02301.

Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. 2021. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. NAR Genom Bioinform. 3(1):lqaa108. doi:10.1093/nargab/lqaa108.

Brůna T, Lomsadze A, Borodovsky M. 2020. GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. NAR Genom Bioinform. 2(2):lqaa026. doi:10.1093/nargab/lqaa026.

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 12(1):59–60. doi:10.1038/nmeth.3176.

Bushnell B. 2014. BBMap: A Fast, Accurate, Splice-Aware Aligner. https://www.osti.gov/biblio/1241166.

Cabanettes F, Klopp C. 2018. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. PeerJ. 6:e4958. doi:10.7717/peerj.4958.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinform. 10(1):421. doi:10.1186/1471-2105-10-421.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. Trimal: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 25(15):1972–1973. doi:10.1093/bioinformatics/btp348.

Chakraborty M, Baldwin-Brown JG, Long AD, Emerson JJ. 2016. Contiguous and accurate *de novo* assembly of metazoan genomes with modest long read coverage. Nucleic Acids Res. 44(19):e147. doi:10.1093/nar/gkw654.

Conway JR, Lex A, Gehlenborg N. 2017. Upsetr: an R package for the visualization of intersecting sets and their properties. Bioinformatics. 33(18):2938–2940. doi:10.1093/bioinformatics/btx364.

Davis CA, Wyatt GR. 1989. Distribution and sequence homogeneity of an abundant satellite DNA in the beetle, *Tenebrio Molitor*. Nucl Acids Res. 17(14):5579–5586. doi:10.1093/nar/17.14.5579.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 29(1):15–21. doi:10.1093/bioinformatics/bts635.

Drost H-G. 2018. Philentropy: information theory and distance quantification with R. J Open Source Softw. 3(26):765. doi:10.21105/joss.00765.

Eddy SR. 2009. A new generation of homology search tools based on probabilistic inference. Genome Inform. 23(1):205–211. doi:10.1142/9781848165632_0019.

Eleftheriou E, Aury J-M, Vacherie B, Istace B, Belser C, Noel B, Moret Y, Rigaud T, Berro F, Gasparian S, *et al.* 2022. Chromosome-scale assembly of the yellow mealworm genome. Open Res Europe. 1:94. doi:10.12688/openreseurope.13987.2.

Emms DM, Kelly S. 2015. Orthofinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 16(1):157. doi:10.1186/s13059-015-0721-2.

Emms DM, Kelly S. 2019. Orthofinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 20(1):238. doi:10.1186/s13059-019-1832-y.

Eriksson T, Andere AA, Kelstrup H, Emery VJ, Picard CJ. 2020. The yellow mealworm (*Tenebrio molitor*) genome: a resource for the emerging insects as food and feed industry. J Insects Food Feed. 6(5): 445–455. doi:10.3920/JIFF2019.0057.

Gabriel L, Hoff KJ, Brůna T, Borodovsky M, Stanke M. 2021. TSEBRA: transcript selector for BRAKER. BMC Bioinformatics. 22(1):566. doi:10.1186/s12859-021-04482-0.

Gotoh O. 2008. A space-efficient and accurate method for mapping and aligning cDNA sequences onto genomic sequence. Nucl Acids Res. 36(8):2630–2638. doi:10.1093/nar/gkn105.

Gremme G, Steinbiss S, Kurtz S. 2013. Genometools: a comprehensive software library for efficient processing of structured genome annotations. IEEE/ACM Trans Comput Biol Bioinf. 10(3): 645–656. doi:10.1109/TCBB.2013.68.

Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R. 2020. Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics. 36(9):2896–2898. doi:10.1093/bioinformatics/btaa025.

Herndon N, Shelton J, Gerischer L, Ioannidis P, Ninova M, Dönitz J, Waterhouse RM, Liang C, Damm C, Siemanowski J, *et al.* 2020. Enhanced genome assembly and a new official gene set for *Tribolium castaneum*. BMC Genomics. 21(1):47. doi:10.1186/s12864-019-6394-6.

Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. 2016. BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. Bioinformatics. 32(5): 767–769. doi:10.1093/bioinformatics/btv661.

Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. 2019. Whole-Genome annotation with BRAKER. In: Kollmar M, editor. Gene Prediction. Vol. 1962. New York (NY): Springer. (Methods in Molecular Biology). p. 65–95. [accessed 2022 Apr 7]. http://link.springer.com/10.1007/978-1-4939-9173-0_5.

Hotaling S, Kelley JL, Frandsen PB. 2021. Toward a genome sequence for every animal: where are we now? Proc Natl Acad Sci U S A. 118(52):e2109019118. doi:10.1073/pnas.2109019118.

Hotaling S, Sproul JS, Heckenhauer J, Powell A, Larracuente AM, Pauls SU, Kelley JL, Frandsen PB. 2021. Long reads are revolutionizing 20 years of insect genome sequencing. Genome Biol Evol. 13(8):evab138. doi:10.1093/gbe/evab138.

Hubley R, Finn RD, Clements J, Eddy SR, Jones TA, Bao W, Smit AFA, Wheeler TJ. 2016. The Dfam database of repetitive DNA families. Nucleic Acids Res. 44(D1):D81–D89. doi:10.1093/nar/gkv1272.

Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. Mol Biol Evol. 33(6): 1635–1638. doi:10.1093/molbev/msw046.

Iwata H, Gotoh O. 2012. Benchmarking spliced alignment programs including Spaln2, an extended version of Spaln that incorporates additional species-specific features. Nucleic Acids Res. 40(20): e161. doi:10.1093/nar/gks708.

Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, Yabana M, Harada M, Nagayasu E, Maruyama H, *et al.* 2014. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Res. 24(8): 1384–1395. doi:10.1101/gr.170720.113.

Katoh K, Standley DM. 2013. MAFFT Multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 30(4):772–780. doi:10.1093/molbev/mst010.

Kelly S, Maini PK. 2013. DendroBLAST: approximate phylogenetic trees in the absence of multiple sequence alignments. PLoS One. 8(3):e58537. doi:10.1371/journal.pone.0058537.

Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 37(5): 540–546. doi:10.1038/s41587-019-0072-8.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. Genome Biol. 5(2):R12. doi:10.1186/gb-2004-5-2-r12.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods. 9(4):357–359. doi:10.1038/nmeth.1923.

Lefort V, Desper R, Gascuel O. 2015. FastME 2.0: a comprehensive, accurate, and fast distance-based phylogeny inference program. Mol Biol Evol. 32(10):2798–2800. doi:10.1093/molbev/msv150.

Letunic I, Bork P. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. 44(W1):W242–W245. doi:10.1093/nar/gkw290.

Levy Karin E, Mirdita M, Söding J. 2020. Metaeuk—sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics. Microbiome. 8(1):48. doi:10.1186/s40168-020-00808-x.

Lewin HA, Richards S, Lieberman Aiden E, Allende ML, Archibald JM, Bálint M, Barker KB, Baumgartner B, Belov K, Bertorelle G, *et al.* 2022. The Earth BioGenome Project 2020: starting the clock. Proc Natl Acad Sci U S A. 119(4):e2115635118. doi:10.1073/pnas.2115635118.

Lewin HA, Robinson GE, Kress WJ, Baker WJ, Coddington J, Crandall KA, Durbin R, Edwards SV, Forest F, Gilbert MTP, *et al.* 2018. Earth BioGenome Project: sequencing life for the future of life. Proc Natl Acad Sci U S A. 115(17):4325–4333. doi:10.1073/pnas.1720115115.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 34(18):3094–3100. doi:10.1093/bioinformatics/bty191.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map format and SAMtools. Bioinformatics. 25(16):2078–2079. doi:10.1093/bioinformatics/btp352.

Liu L-N, Wang C-Y. 2014. Complete mitochondrial genome of yellow meal worm (*Tenebrio molitor*). Dongwuxue Yanjiu. 35(6):537–545. doi:10.13918/j.issn.2095-8137.2014.6.537.

Lomsadze A. 2005. Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res. 33(20):6494–6506. doi:10.1093/nar/gki937.

Lomsadze A, Burns PD, Borodovsky M. 2014. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. Nucl Acids Res. 42(15):e119. doi:10.1093/nar/gku557.

Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. 2021. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol Biol Evol. 38(10): 4647–4654. doi:10.1093/molbev/msab199.

Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 27(6): 764–770. doi:10.1093/bioinformatics/btr011.

Peng B-Y, Li Y, Fan R, Chen Z, Chen J, Brandon AM, Criddle CS, Zhang Y, Wu W-M. 2020. Biodegradation of low-density polyethylene and polystyrene in superworms, larvae of *Zophobas atratus* (Coleoptera: Tenebrionidae): broad and limited extent depolymerization. Environ Pollut. 266:(Part 1):115206. doi:10.1016/j.envpol.2020.115206.

Petitpierre E, Gatewood JM, Schmid CW. 1988. Satellite DNA from the beetle *Tenebrio molitor*. Experientia. 44(6):498–499. doi:10.1007/BF01958925.

Popescu A-A, Huber KT, Paradis E. 2012. ape 3.0: new tools for distance-based phylogenetics and evolutionary analysis in R. Bioinformatics. 28(11):1536–1537. doi:10.1093/bioinformatics/bts184.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 26(6):841–842. doi:10.1093/bioinformatics/btq033.

Ramos-Elorduy J. 2009. Anthropo-entomophagy: cultures, evolution and sustainability. Entomol Res. 39(5):271–288. doi:10.1111/j.1748-5967.2009.00238.x.

Ribeiro N, Abelho M, Costa R. 2018. A review of the scientific literature for optimal conditions for mass rearing *Tenebrio molitor* (Coleoptera: Tenebrionidae). J Entomol Sci. 53(4):434–454. doi:10.18474/JES17-67.1.

Rumbos CI, Athanassiou CG. 2021. The superworm, *Zophobas morio* (Coleoptera:Tenebrionidae): a "Sleeping Giant" in nutrient sources. J Insect Sci. 21(2):13. doi:10.1093/jisesa/ieab014.

Samollow PB, Dawson PS, Riddle RA. 1983. X-linked and autosomal inheritance patterns of homologous genes in two species of *Tribolium*. Biochem Genet. 21(1–2):167–176. doi:10.1007/BF02395401.

Shen W, Le S, Li Y, Hu F. 2016. Seqkit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PLoS One. 11(10):e0163962. doi:10.1371/journal.pone.0163962.

Smith SG. 1952. The evolution of heterochromatin in the genus *Tribolium* (Tenebrionidae: Coleoptera). Chromosoma. 4(1):585–610. doi:10.1007/BF00325793.

Stanke M, Diekhans M, Baertsch R, Haussler D. 2008. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics. 24(5):637–644. doi:10.1093/bioinformatics/btn013.

Stanke M, Schöffmann O, Morgenstern B, Waack S. 2006. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinform. 7(1):62. doi:10.1186/1471-2105-7-62.

Stork NE, McBroom J, Gely C, Hamilton AJ. 2015. New approaches narrow global species estimates for beetles, insects, and terrestrial arthropods. Proc Natl Acad Sci U S A. 112(24):7519–7523. doi:10.1073/pnas.1502408112.

Tarailo-Graovac M, Chen N. 2009. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics. 25(1):4.10.1–4.10.14. doi:10.1002/0471250953.bi0410s25. [accessed 2022 Apr 4]. https://onlinelibrary.wiley.com/doi/10.1002/0471250953.bi0410s25.

Tribolium Genome Sequencing Consortium. 2008. The genome of the model beetle and pest *Tribolium castaneum*. Nature. 452-(7190):949–955. doi:10.1038/nature06784.

Van Dongen S. 2000. Graph clustering by flow simulation [PhD Thesis]. The Netherlands: University of Utrecht.

Van Huis A. 2013. Potential of insects as food and feed in assuring food security. Annu Rev Entomol. 58(1):563–583. doi:10.1146/annurev-ento-120811-153704.

Vaser R, Sović I, Nagarajan N, Šikić M. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 27(5):737–746. doi:10.1101/gr.214270.116.

Veeneman BA, Shukla S, Dhanasekaran SM, Chinnaiyan AM, Nesvizhskii AI. 2016. Two-pass alignment improves novel splice junction quantification. Bioinformatics. 32(1):43–49. doi:10.1093/bioinformatics/btv642.

Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, Schatz MC. 2017. Genomescope: fast reference-free genome profiling from short reads. Bioinformatics. 33(14):2202–2204. doi:10.1093/bioinformatics/btx153.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, *et al.* 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 9(11):e112963. doi:10.1371/journal.pone.0112963.

Yang L, Gao J, Liu Y, Zhuang G, Peng X, Wu W-M, Zhuang X. 2021. Biodegradation of expanded polystyrene and low-density polyethylene foams in larvae of *Tenebrio molitor* Linnaeus (Coleoptera: Tenebrionidae): broad versus limited extent depolymerization and microbe-dependence versus independence. Chemosphere. 262:127818. doi:10.1016/j.chemosphere.2020.127818.

Yang Y, Wang J, Xia M. 2020. Biodegradation and mineralization of polystyrene by plastic-eating superworms *Zophobas atratus*. Sci Total Environ. 708:135233. doi:10.1016/j.scitotenv.2019.135233.

Yang Y, Yang J, Wu W-M, Zhao J, Song Y, Gao L, Yang R, Jiang L. 2015a. Biodegradation and mineralization of polystyrene by plastic-eating mealworms: part 1. Chemical and physical characterization and isotopic tests. Environ Sci Technol. 49(20):12080–-12086. doi:10.1021/acs.est.5b02661.

Yang Y, Yang J, Wu W-M, Zhao J, Song Y, Gao L, Yang R, Jiang L. 2015b. Biodegradation and mineralization of polystyrene by plastic-eating mealworms: part 2. Role of gut microorganisms. Environ Sci Technol. 49(20):12087–12093. doi:10.1021/acs.est.5b02663.

Zimin AV, Puiu D, Luo M-C, Zhu T, Koren S, Marçais G, Yorke JA, Dvořák J, Salzberg SL. 2017. Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. Genome Res. 27(5):787–792. doi:10.1101/gr.213405.116.

Editor: J. Comeron