BMC Genomics

RESEARCH

Open Access

A high-quality de novo genome assembly based on nanopore sequencing of a wild-caught coconut rhinoceros beetle (*Oryctes rhinoceros*)

Igor Filipović[1,2*], Gordana Rašić[2], James Hereward[1], Maria Gharuka[3], Gregor J. Devine[2], Michael J. Furlong[1] and Kayvan Etebari[1]

Abstract

Background: An optimal starting point for relating genome function to organismal biology is a high-quality nuclear genome assembly, and long-read sequencing is revolutionizing the production of this genomic resource in insects. Despite this, nuclear genome assemblies have been under-represented for agricultural insect pests, particularly from the order Coleoptera. Here we present a de novo genome assembly and structural annotation for the coconut rhinoceros beetle, *Oryctes rhinoceros* (Coleoptera: Scarabaeidae), based on Oxford Nanopore Technologies (ONT) long-read data generated from a wild-caught female, as well as the assembly process that also led to the recovery of the complete circular genome assemblies of the beetle's mitochondrial genome and that of the biocontrol agent, Oryctes rhinoceros nudivirus (OrNV). As an invasive pest of palm trees, *O. rhinoceros* is undergoing an expansion in its range across the Pacific Islands, requiring new approaches to management that may include strategies facilitated by genome assembly and annotation.

Results: High-quality DNA isolated from an adult female was used to create four ONT libraries that were sequenced using four MinION flow cells, producing a total of 27.2 Gb of high-quality long-read sequences. We employed an iterative assembly process and polishing with one lane of high-accuracy Illumina reads, obtaining a final size of the assembly of 377.36 Mb that had high contiguity (fragment N50 length = 12 Mb) and accuracy, as evidenced by the exceptionally high completeness of the benchmarked set of conserved single-copy orthologous genes (BUSCO completeness = 99.1%). These quality metrics place our assembly ahead of the published Coleopteran genomes, including that of an insect model, the red flour beetle (*Tribolium castaneum*). The structural annotation of the nuclear genome assembly contained a highly-accurate set of 16,371 protein-coding genes, with only 2.8% missing BUSCOs, and the expected number of non-coding RNAs. The number and structure of paralogous genes in a gene family like Sigma GST is lower than in another scarab beetle (*Onthophagus taurus*), but higher than in the red flour beetle (*Tribolium castaneum*), which suggests expansion of this GST class in Scarabaeidae. The quality of our gene models was also confirmed with the correct placement of *O. rhinoceros* among other members of the rhinoceros beetles (subfamily Dynastinae) in a phylogeny based on the sequences of 95 protein-coding genes in 373 beetle species from all major

*Correspondence: i.filipovic@uq.edu.au; filipovic.igor@gmail.com

[1] School of Biological Sciences, The University of Queensland, St. Lucia, Australia

Full list of author information is available at the end of the article

lineages of Coleoptera. Finally, we provide a list of 30 candidate dsRNA targets whose orthologs have been experimentally validated as highly effective targets for RNAi-based control of several beetles.

Conclusions: The genomic resources produced in this study form a foundation for further functional genetic research and management programs that may inform the control and surveillance of *O. rhinoceros* populations, and we demonstrate the efficacy of de novo genome assembly using long-read ONT data from a single field-caught insect.

Keywords: Genome assembly, Genome annotation, Single insect nanopore sequencing, *Oryctes rhinoceros*, Coleoptera

Background

Adult coconut rhinoceros beetles, *Oryctes rhinoceros* L. (Coleoptera: Scarabaeidae), feed by boring into the crown of coconut palms. This damages growing tissue and significantly reduces coconut yields and can lead to the death of trees. Native to southeast Asia, this pest was accidentally introduced into Samoa in 1909 [1], and it has since spread across the tropical Pacific, bringing a significant threat to the livelihoods of the peoples of Pacific island nations for whom the coconut palm ('the tree of life') is an important source of food, fibre and timber. Invasive populations of *O. rhinoceros* have been suppressed over the past 60 years through management approaches that included the release of a biocontrol agent, Oryctes rhinoceros nudivirus (OrNV) [2]. However, a highly damaging infestation by *O. rhinoceros* in Guam in 2007 was not controlled with OrNV, and the beetle's subsequent expansion to other Pacific Islands including Papua New Guinea, Hawaii, Solomon Islands, and most recently Vanuatu and New Caledonia [3–6], suggests potential changes in this biological system [7] that require new approaches to management, including the isolation and deployment of highly virulent OrNV strains for specific *O. rhinoceros* genotypes [8].

Genome sequencing has enabled better understanding of population outbreaks, invasion and adaptation mechanisms in insect pests [9, 10]. Functional and comparative genomics studies are identifying new targets for control and the implementation of integrated pest management strategies [11]. Draft genome assembly is generally a good starting point for relating genome function to organismal biology, but the production of this genomic resource for agricultural pests has lagged behind that of some other insects [11, 12]. A recent project aiming to tackle this lag is the Ag100Pest Initiative, led by the United States Department of Agriculture's Agricultural Research Service (USDA-ARS), that is set to produce reference quality genome assemblies for the top 100 arthropod agricultural pests in the USA, with nearly one third of species belonging to Coleoptera [13].

Draft genome assemblies are very useful for population genomic analyses, enabling the design of, for example, optimal protocols for reduced genome representation sequencing [14]. However, draft genome assemblies that are highly fragmented, incomplete or misassembled have limited use for functional genomic studies. Transcriptome assemblies are useful for studying functionally and sufficiently transcribed parts of the genome, but only complete and accurate genome assemblies provide information on non-transcribed regions (e.g. promoters, enhancers) that can have important influences on gene expression and, ultimately, economically-important phenotypes [13]. In addition, different types of non-translated RNAs (e.g. microRNAs, lncRNAs) are often not detected in transcriptome studies but are included in complete and accurate genome assemblies. These can help us understand how insect pests interact and respond to their hosts, pathogens, the environment and they can reveal new targets for novel genetic control measures (e.g. RNAi [15], gene drives [16, 17]).

Obtaining high-quality genome assemblies is often challenging in insects [12], particularly from short-read sequencing data (e.g. Illumina) for species with high levels of DNA polymorphism and repetitive genomic elements [18]. These issues are further compounded for insects of small physical size or for partial specimens, as they may require whole genome amplification or the pooling of several individuals to obtain enough DNA for library preparations. Different methods of whole genome amplification vary in their ability to preserve specific genetic variation and can be biased against regions with high GC-content, smaller and low-abundance DNA fragments [19]. They can also create chimeric fragments and amplify contaminating DNA that can be erroneously integrated into the target assembly. Pooling of individuals is preferably done with individuals from a line that has undergone inbreeding to reduce genetic variation, but many pest species cannot be colonised in the laboratory. Moreover, for those insects that can be lab-reared, intensive inbreeding procedures such as full-sib mating for tens of generations may not reduce heterozygosity in all parts of the genome (e.g. [20]). The pooling of wild-caught samples is particularly problematic given the possibility of combining cryptic species or biotypes,

which would impact assembly quality and lead to spurious biological conclusions. When presented with a highly heterozygous genome or a pool of diverse haplotypes, the standard assembly process tends to report a heterozygous region as alternative contigs (instead of collapsing them into a single haplo-contig) and is unable to resolve multiple paths between homo- and heterozygotic regions, producing a highly fragmented assembly with an erroneously inflated total size [21]. Such assemblies cause problems in genome annotation and downstream analyses, giving fragmented gene models, wrong gene copy numbers, and broken synteny. They also preclude linkage mapping and genome-wide association studies.

The development of long-read sequencing technologies is revolutionizing the production of contiguous and complete insect genome assemblies [18], but their requirement for large quantities of input DNA have complicated their application to single-insect assemblies. However, new low-input protocols were recently demonstrated for Pacific Biosciences (PacBio) long-read sequencing, producing high-quality single-insect genome assemblies for the mosquito *Anopheles coluzzii* [22] and spotted lanternfly *Lycorma delicatula* [23]. A chromosome-level assembly was recently reported for a single outbred *Drosophila melanogaster* generated using a combination of long-read sequences from Oxford Nanopore Technologies (ONT), Illumina short-read sequences and Hi-C data [24]. However, the small size of this insect necessitated genome amplification to prepare the sequencing libraries, and the final assembly was ~20% smaller than the canonical reference genome for *D. melanogaster* [24].

Here, we present a high-quality de novo genome assembly based on ONT long-read data from a single wild-caught adult female of the coconut rhinoceros beetle (*O. rhinoceros*, NCBI:txid72550). The amount of DNA extracted from this large insect was sufficient to prepare multiple ONT libraries without genome amplification. Data from just one flow cell were enough to produce a high-quality draft assembly of the beetle's nuclear genome, and data from four MinION flow cells enabled the assembly that is among the most accurate and complete of the published Coleopteran genomes, as well as the assembly of its mitochondrial genome [25], and the genome of the biocontrol agent Oryctes rhinoceros nudivirus (OrNV) [26] that had infected the individual we analysed.

Results and discussion

ONT library preparation and sequencing

We used a customized Solid-phase Reversible Immobilization (SPRI) bead-based protocol to extract high molecular weight (HMW) DNA from an *O. rhinoceros* female (see Materials and methods, Fig. 1A-C). Given the large size of the insect, we achieved high quantity (~10 µg) and quality HMW DNA (Supplemental Fig. 1), that we size-selected with the Circulomics XS kit (Fig. 1C), and prepared four standard ligation-based ONT libraries. Each library was sequenced on a MinION Flow Cell (Fig. 1D), yielding between 896,000 and 1.48 million raw reads. After basecalling with Guppy v.3.2.4, we obtained a total 29.4 Gb of sequence data with 89.8% passing the QC filtering (Phred score ⩾8). 26.4 Gb of high-quality data with the read length N50 of ~11.3 kb were used for downstream analyses (Supplemental Table 1), and the longest recorded read that passed the QC filtering was 143.6 kb. For the second round of analyses, we used the newer base-caller version, Guppy v4.2.2, which improved the yield of high-quality reads (a total of 29.5 Gb of data, 92.1% passing the QC filtering, Phred score ⩾8). These 27.2 Gb of high-quality reads had a length N50 of ~11.2 kb and were used for the main downstream analyses (Supplemental Table 1). The longest read that passed the QC filtering in this dataset version was ~148.4 kb.

Genome assembly and quality assessment

Because we expected the long-read data (LR) to contain some percentage of mitochondrial, bacterial and other contaminant DNA reads, we first ran the long-read assembler Flye version 2.5 (Fig. 2A) in metagenome mode that accommodates a highly non-uniform coverage of genomic fragments and is sensitive to under-represented sequences [27]. The initial draft assembly graph (S4-i-v1-g, Fig. 2B) consisted of 512 nodes with N50 length of 7.9 Mb and total assembly size of 370.4 Mb. This initial draft assembly graph was then screened for the mitochondrial genome sequence, expecting a circular node 11 kb to 22 kb in size (based on a typical mitogenome size in insects [28]), and a disproportionately high depth of coverage (given that there are tens/hundreds of copies of the mitochondrial genome per nuclear genome copy in each cell). We identified one node with such characteristics: edge_110 (Fig. 2D) was 21,039 bp in length and had a median coverage of 10,292X, showing the NCBI 'blastn' match with the mitochondrial genome assembly sequences (complete or partial) of beetles and other insects. Another circular node (edge_371) (Fig. 2C) with a high depth of coverage (1196X) was 126,204 bp in length, which we identified through the NCBI 'blastn' search as the Oryctes rhinoceros nudivirus (OrNV), a double-stranded DNA virus used as a biocontrol agent against *O. rhinoceros* [29]. Both nodes were removed from the draft assembly graph and analysed separately (Fig. 2E-F), and their detailed characterization is described elsewhere [25, 26].

Given the potential for ONT basecalling to introduce systemic indel errors in the homopolymer regions

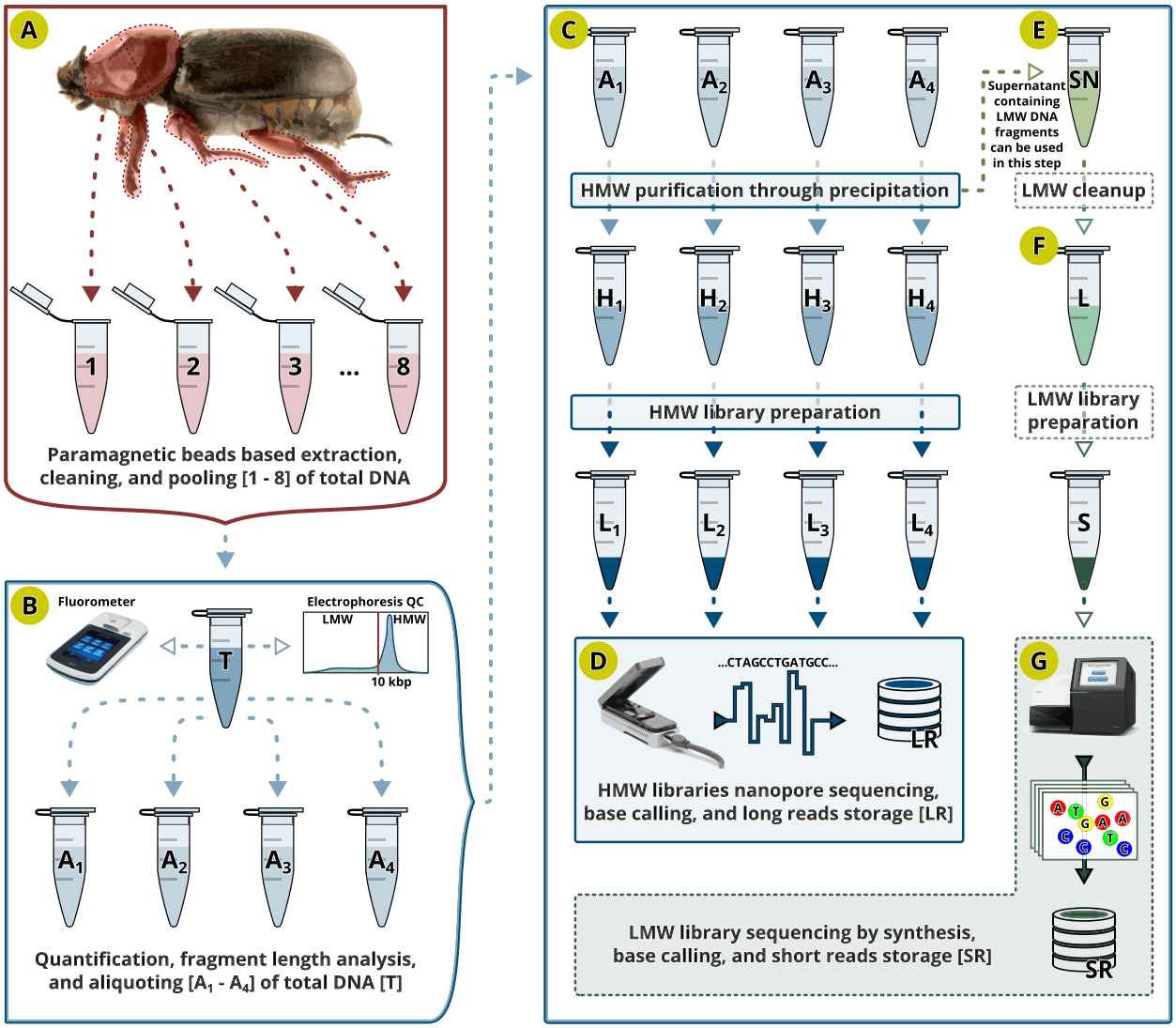

Fig. 1 Optimal library preparation and sequencing. (**A**) Total DNA is extracted from thorax and legs and all 8 extracts are pooled (**B**) and assessed for DNA quality and quantity. DNA extract is aliquoted [A1-A4] and (**C**) high molecular weight [HMW] DNA is precipitated [H1-H4] and used for ONT library preparations. (**D**) Each library [L1-L4] is sequenced on one Oxford Nanopore MinION flow cell. High accuracy base caller transforms the raw nanopore data into long reads [LR]. (**E**) Supernatant [SN] containing low molecular weight DNA [LMW] can also be cleaned [L] and used for the preparation of libraries [S] for short read sequencing [SR]

of the ONT-based assemblies [30], we used Pilon [31], BWA-MEM aligner [32] and more accurate Illumina Whole Genome Sequence data to remove small indels in the initial linearised draft assembly (Fig. 2G-I). We used the previously generated Illumina short reads from a whole-genome sequencing library that we prepared using the NebNext Ultra DNA II Kit (New England Biolabs, USA) with DNA extracted from another *O. rhinoceros* female that was collected from the same geographic location. The short-fragment Illumina library (Fig. 1F-G) contained ~ 39.4 Gb of 150 bp paired end read data. We point out that Illumina sequencing library intended for the polishing of an ONT-based assembly would ideally be prepared from the same individual that was used to generate the long-read data. This would allow not only the correction of indels but also the correction of SNPs in the assembly consensus sequences. For the experiments with small-bodied insects that yield limited amounts of DNA, we recommend using the Low Molecular Weight (LMW) DNA found in the supernatant of the ONT library preparation mix (LMW depletion step, Fig. 1E).

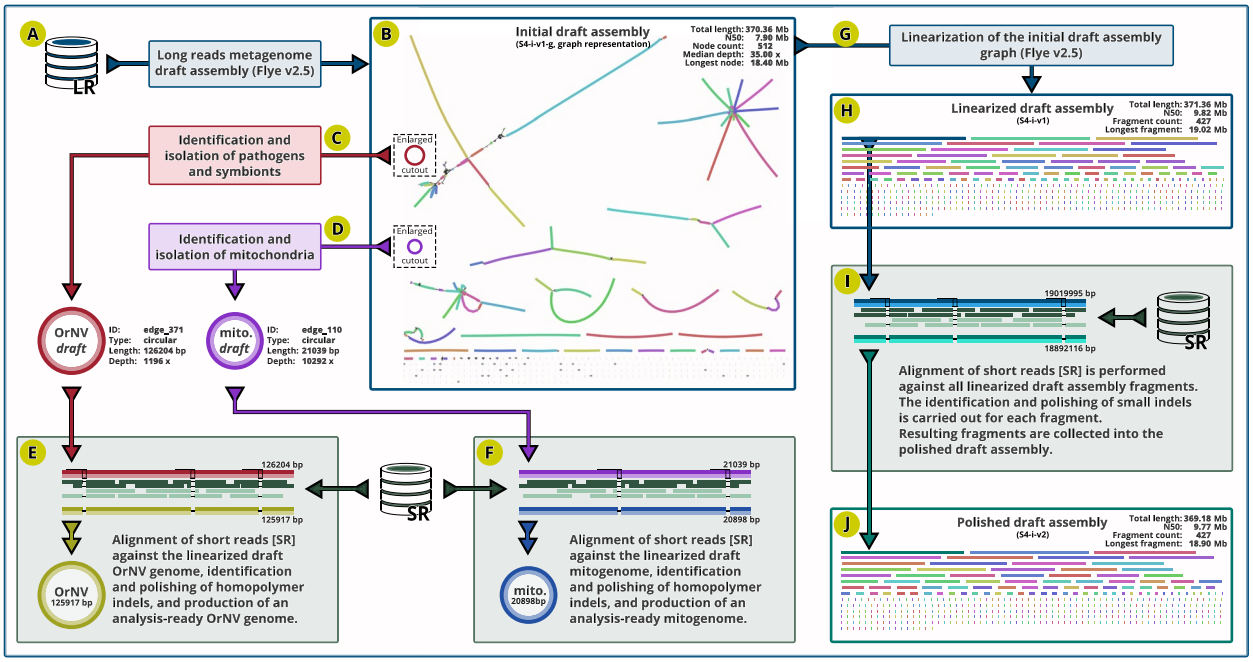

Fig. 2 Identification and elimination of organelle, pathogen and other contaminant genomes. (**A**) Long-read data [LR] are used to generate the initial draft assembly (**B**), circular assembly for OrNV (**C**) and mitochondria (**D**) are identified and removed. Short-read data [SR] were used to remove erroneous indels in homopolymers (**E**, **F**) to produce analysis-ready assemblies [25, 26]. The remainder of the draft assembly (**G**) is linearized (**H**) and short reads [SR] are used to remove erroneous indels (**I**) in each scaffold, producing an initial polished nuclear genome assembly for *Oryctes rhinoceros* (**J**)

It is also worth noting that the indel error correction with the Illumina short reads has limitations in repetitive regions of the assembly, where short reads cannot be accurately aligned. For polishing, we used 92.4% of the Illumina reads that aligned to the initial genome assembly (S4-i-v1). Of the remaining reads, 6.1% aligned to the mitogenome and 0.2% to the OrNV genome, leaving 1.3% of the short-reads unaligned. The resulting polished initial genome assembly version S4-i-v2 consisted of 427 fragments (6 scaffolds and 421 contigs, Fig. 2J), with the fragment N50 length of 9.8 Mb, the longest fragment of 18.9 Mb, and a total assembly size of 369.2 Mb (34.9% GC content).

A quantitative assessment of the initial assembly's accuracy and completeness was done through the benchmarking analysis of conserved genes, as implemented in BUSCO [33]. Using the BUSCO collection of 2124 genes from the endopterygota database (endopterygota_odb10), we found that the initial polished assembly (S4-i-v2) contained 97.9% complete genes, with 97.2% occurring as single copies and only 0.9% missing. In comparison, BUSCO analysis of the unpolished assembly version (S4-i-v1) recovered only 65.1% genes as complete and 19.6% as missing, revealing the substantial impact of the uncorrected indel errors on gene prediction and detection (Supplemental Table 2).

To further improve the assembly quality, we used the latest available version of the base-caller Guppy (v4.2.2) in high accuracy mode, and the latest available version of the long-read assembler Flye (v2.8.2) to generate multiple draft assemblies (Fig. 3A-B) by increasing the minimum read overlap parameter for each assembly from 5 kb to 10 kb in increments of 500 bases. The Illumina short-reads were aligned against each draft assembly using BWA-MEM (Fig. 3C), and the resulting alignments were further utilised to polish indels within each draft assembly (Fig. 3D). This iterative process produced a collection of 11 polished draft assemblies (Fig. 3E), and each was assessed for contiguity (assembly-stats "https://github.com/sanger-pathogens/assembly-stats") and completeness (BUSCO) (Fig. 3F) (Supplemental Table 2). The best overall assembly (S4-7k-1v2) was produced with a minimal read overlap of 7 kb, and this parameter value was used to repeat the assembly, polishing and assessment two additional times (producing S4-7k-2v2 and S4-7k-3v2). The best of these three versions (S4-7k-2v2) was selected for further processing. We then removed the OrNV and mitochondrial sequences from the assembly (published previously [25, 26]), and this version

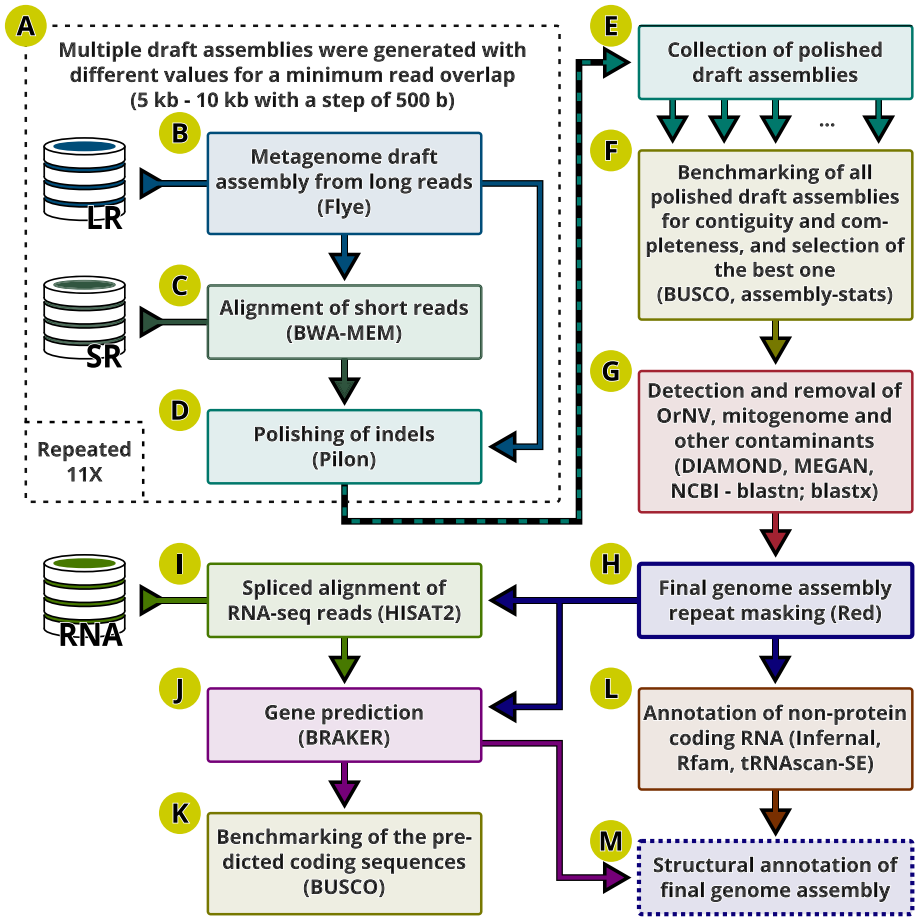

Fig. 3 Iterative assembly refinement. (**A**) Multiple polished draft assemblies are generated (**B**-**D**), collected (**E**) and benchmarked for completeness and contiguity (**F**) to determine the optimal read overlap for the long leads [LR]. (**G**) Optimal draft assembly is screened for potential contaminants. (**H**) The repeats are detected and soft masked. The splice-aware alignments (**I**) of the RNAseq datasets [RNA] are used for gene prediction (**J**), and then assessed for completeness (**K**). Annotations of the non-protein coding RNAs (**L**) are added to form the final structural annotation (**M**)

(S4-7k-2v3) was further analysed with DIAMOND [34] and MEGAN [35, 36] in order to identify potential contaminant fragments. All assembly sequences that were not classified within Arthropoda in this pipeline were additionally checked against the NCBI's online databases of nucleotide (nt/nr) and non-redundant protein sequences (nr) to identify the origin of a putative contaminant sequence (Fig. 3G). Given that none of the analysed sequences had a significant BLAST hit to a taxon other than Coleoptera, we did not consider them as contaminants and did not remove them from the final genome assembly (S4–74-2v3, Fig. 3H). This final assembly consisted of 1013 fragments (6 scaffolds and 1007 contigs), with the fragment length N50 of 10.7 Mb and the longest fragment (contig_6) of 32.7 Mb (Table 1, Supplemental Table 2).

The size of our final *O. rhinoceros* nuclear genome assembly (S4-7k-2v3, GenBank assembly accession: GCA_020654165.1) was 377.4 Mb, which is very similar to the latest assembly for the congeneric beetle *O. borbonicus* (371.60 Mb in ungapped length, NCBI accession:

Table 1 Assembly statistics for *Oryctes rhinoceros*

Total sequence length	377,356,435
Total ungapped length	377,355,735
Gaps between scaffolds	0
Number of scaffolds	1013
Scaffold N50	10,697,081
Scaffold L50	12
Number of contigs	1020
Contig N50	10,534,518
Contig L50	12

General statistics information includes total sequence (gapped and ungapped) length, scaffold and contig number as well as their N50 and L50

GCA_902654985.1). The quality of our *O. rhinoceros* assembly, however, is superior to that of *O. borbonicus*, both in terms of contiguity (contig L50: *O. rhinoceros* vs. *O. borbonicus* = 12 vs. 571 (Supplemental Table 3)) and completeness (BUSCOs: *O. rhinoceros* = 99.1% complete, 0.5% missing; *O. borbonicus* = 96.1% complete, 3.5% missing) (Supplemental Table 4). Of note is that the original assembly for *O. borbonicus*, generated with the short-read Illumina technology, was first reported to be 518 Mb [37], but refinement with the 10X Genomics data led to a 28% reduction in size (removal of more than 140 Mb). The inflated size of the initial assembly was explained as a consequence of an incorrect haploidization of the assembly i.e., divergent haplotypes were assembled separately across many parts of the genome [38]. This exemplifies the difficulties of the assembly process based on the short-read sequencing of samples that have high genome-wide variability. Conversely, our *O. rhinoceros* assembly indicates that the correct haploidization is not problematic for long-read assemblers like Flye [27], particularly when the long-read data are generated from a single insect.

Comparison with other available nuclear genome assemblies in Coleoptera

A recent 'state of the field' overview of insect genome assemblies [18] reports that this biological resource has been significantly underrepresented in Coleoptera (i.e. few genome assemblies are produced relative to the species richness), but that long-read sequencing is revolutionizing the creation of high-quality assemblies across insect groups [18]. We analysed 39 representative nuclear genome assemblies in the Coleoptera (out of 41 accessed from NCBI's GenBank in October 2020) and found that one third were generated with data that included long-read sequences (nine assemblies with PacBio, four with ONT). For a total set of 39 analysed assemblies (Fig. 4), the mean fragment N50 was 6.9 Mb (median: 298.9 kb, SD: 19.9 Mb) and the mean BUSCO completeness was 88.4% (median: 92.4%, SD: 14.3%). These quality metrics are above the average for a set of 601 assemblies from 20 insect orders (N50: 1.1 Mb, BUSCO completeness: 87.5%, [18]).

Our *O. rhinoceros* assembly had the highest assembly accuracy and completeness among 39 benchmarked Coleopteran genomes, having only 0.5% missing BUSCOs (10 out of 2124 core genes) and 0.4% fragmented BUSCOs (9 out of 2124 core genes) (Fig. 4). A genome assembly from another member of the family Scarabaeidae, *Onthophagus taurus*, had the same number of missing BUSCOs but twice as many duplicated genes (2.7%), and a substantially lower assembly contiguity, with scaffold (fragment) L50 of 160 versus 12 in *O. rhinoceros* (Supplemental Table 4).

Inspection of other beetle genomes that have also been assembled using ONT data confirms that this technology facilitates production of assemblies with high contiguity and completeness (Supplemental Tables 3 and 4). However, two examples in true weevils, *Rhynchophorus ferrugineus* and *Listronotus bonariensis* (Curculionidae), reveal that low completeness and accuracy (evident as low number of single copy BUSCOs) exist in ONT-based assemblies that have both low and high contiguity (Supplemental Tables 3 and 4).

Until recently, ONT's requirement for large amount of input DNA precluded full utilization of this technology in small-bodied insects, but a recent example of the ONT-based assembly from a single *Drosophila* [39] indicates that genome assemblies of high completeness (96.9% complete BUSCOs), albeit partial genome length (85%), could be achieved with this technology even for very small insects. Considering that the limiting factor of ONT technology is the density of available nanopores per flow cell, the sequencing yield could be improved by having more shorter DNA fragments rather than fewer long ones when the amount of input DNA is small. We also note that the improvements in the later versions of both the ONT basecaller Guppy and the long-read assembler Flye are reflected in a substantially better draft assembly prior to any indel polishing (see Fig. 4: S4-7k-2v1 versus the equivalent non-polished assembly S4-i-v1 that was produced with the older software versions).

Structural annotation and quality assessment

To delineate protein-coding genes, we used the BRAKER pipeline (Fig. 3J) which enables an automated training of the gene prediction tools (GeneMark-EX and AUGUSTUS) with the extrinsic evidence from the RNA-Seq experiments [40–46]. We used the publicly-available RNA-seq data that cover different life stages of *O. rhinoceros*, from early instar larva, late instar larva, pupa, and the adult stage (NCBI accession: PRJNA486419; [47]), which is expected to maximize the probability of capturing the sequences of the entire set of expressed genes in this organism. To check data quality from these RNA-seq samples, we first aligned the reads against our genome assembly with the splice-aware aligner HISAT2 [48], and used these alignments to produce a genome-guided transcriptome with Trinity [49]. The assembled transcriptome had a very high BUSCO completeness (97.5%), indicating that the source RNA-seq dataset provides an excellent training set for gene prediction. Along with these aligned RNA-seq reads, the BRAKER pipeline was supplied with the final genome assembly (S4-7k-2v3) that had the repetitive regions (transposons and simple

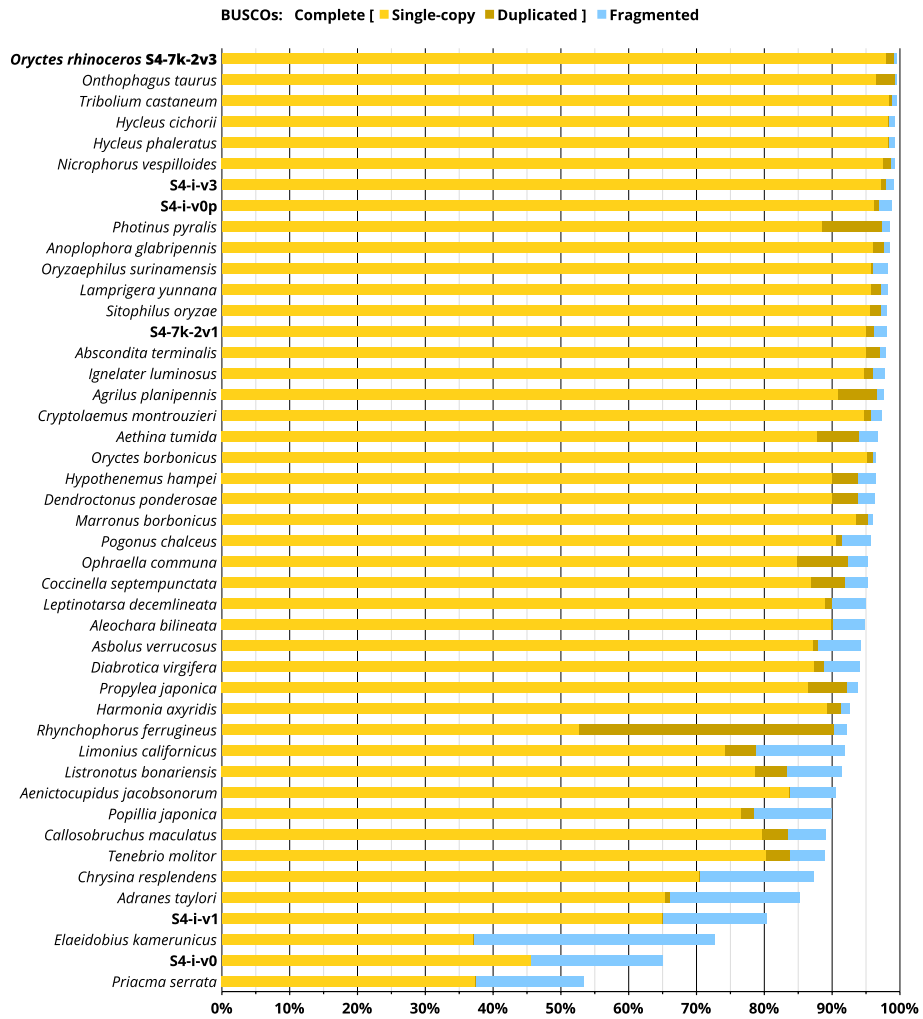

Fig. 4 BUSCO comparison. Genome assemblies of *O. rhinoceros* and other Coleoptera are ranked by the number of detected BUSCOs. Assemblies of the highest quality have a very high percentage of complete single copy BUSCOs (yellow bar), and a small percentage of duplicated BUSCOs (brown bar) as well as fragmented BUSCOs (blue bar). *Oryctes rhinoceros* assemblies are: S4-7k-2v3 (final polished version), S4-i-v3 (initial polished version), S4-i-v0p (draft polished version from a single flow cell of long-read data), S4-7k-2v1 (final unpolished version), S4-i-v1 (initial unpolished version), S4-i-v0 (draft unpolished version from a single flow cell of long-read data)

repeats) soft-masked on 32.7% of the assembly sequences (using the repeat detector Red [50]). The gene prediction algorithm produced a set of 16,375 protein-coding genes with a total of 20,072 transcripts. Our results match the available data for other members of Coleoptera; for example, 16,538 genes were reported for the bull-headed dung beetle *Onthophagus taurus* (Scarabaeidae) [51, 52], and the latest reference annotation for the red flour beetle *Tribolium castaneum* (Tenebrionidae) reports 16,593 genes with a total of 18,536 transcripts [53].

We also delineated the non-protein-coding RNAs, using tRNAscan-SE [54] and Infernal [55] with the Rfam database [56, 57] (Fig. 3L). The annotation produced predictions for 18 tRNA-like pseudogenes, one selenocysteine tRNA gene, and 13 unknown isotypes. The number of tRNA genes predicted in *O. rhinoceros* (392) is highly congruent with another scarab beetle, *O. taurus*, that has 395 predicted tRNA genes [51, 52]. Our annotation with all predicted protein-coding genes, as well as non-protein-coding genes (including rRNA, miRNA) and other features is provided as a gff3 file (Fig. 3M) (Additional file 1).

The benchmarking analysis (Fig. 3K) indicated that our structural annotation of protein-coding genes in *O. rhinoceros* assembly is of high quality, with only 2.8% of BUSCOs missing. Somewhat higher missingness obtained for the annotated gene set when compared to the assembly (2.8% vs. 0.5% missing BUSCOs) can be

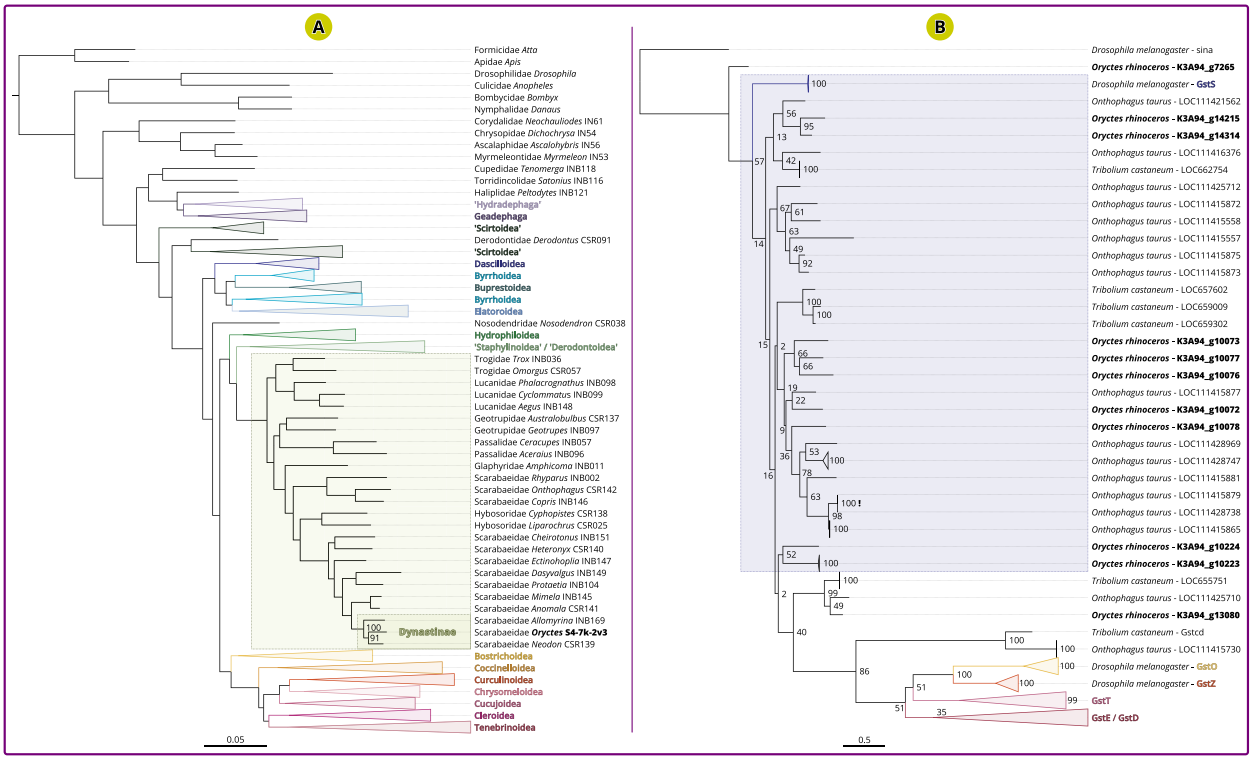

Fig. 5 RAxML phylogeny with single-copy orthologs and sigma GST genes. (**A**) RAxML-ng tree generated with concatenated amino acid alignments from 95 genes in 384 taxa (374 Coleoptera, 10 outgroup taxa), including *O. rhinoceros*. Phylogeny is well resolved, with all branches having > 65% support (all major lineages have > 90% support). *Oryctes rhinoceros* is correctly placed within Dynastinae, along with two other members of this subfamily (branch support 100%). Superfamily Scarabaeoidea is shaded in green. (**B**) RAxML tree generated with nucleotide sequences of GST genes identified in *O. rhinoceros, O. taurus, T. castaeum* and *D. melanogaster*. Nine putative sigma GST paralogues are detected in *O. rhinoceros*, while *O. taurus* contains 12 genes (with one technical duplication (LOC111415879 and LOC111428738) that resulted from the terminal position of the predicted gene on two separate scaffolds). Although seven sigma GST genes were previously reported in *T. castaneum* [37], only five were detected in its current genome annotation [53]. Sigma GST class is shaded in blue. Other GST classes (omega, delta, theta, epsilon, zeta) are found as divergent and highly-supported branches

explained by the similar completeness of RNAseq dataset (1.8% missing BUSCOs) that was used by the annotation pipeline to guide gene predictions. It could also indicate that the annotation pipeline, which uses multiple sources of evidence, has generated slightly inferior gene models for a set of single-copy orthologs than the single-predictor approach that BUSCO takes when working directly on the assembly sequences [58]. Such differences have been reported, for example, in the BUSCO assessment of 15 *Anopheles* mosquito genomes and their annotated gene sets [58].

Predicted gene models recover the correct phylogenetic placement of O. rhinoceros

In addition to the BUSCO metrics, we assessed the quality of our predictions of single-copy orthologous genes through a phylogenetic analysis. We used the largest data source for beetle phylogenetics to date, generated by Zhang et al. [59], that includes partial sequences of 95 nuclear protein-coding genes from 373 beetle species and 10 outgroup taxa. Out of 95 genes used by Zhang et al. [59], 94 genes were identified in our structural annotation, and one gene that was missing from the annotation was identified in our assembly. The concatenated alignment supermatrix consisted of 24,542 amino acids (Additional file 2) and the phylogeny was estimated using the maximum likelihood method in RAxML-ng [60] (Additional file 3). The resulting phylogeny was well resolved, and *O. rhinoceros* was correctly grouped with two other members of the Dynastinae subfamily (branch support = 100%, Fig. 5A), further confirming the high quality of our predictions of single-copy genes.

Possible expansion of sigma GST genes in Scarabaeidae

We wanted to check if paralogous genes are also correctly predicted in our annotation. Sigma Glutathione-S-Transferase genes (Sigma GSTs) belong to an ancient gene family and one of six classes of cytosolic GSTs in

insects, that were previously reported to have undergone *Oryctes*-specific expansion [37]. Meyer and colleagues found 12 Sigma GST paralogs in their *O. borbonicus* assembly, while the genomes of four other insects they analysed, including two beetles (*Tribolium castaneum* and *Dendroctonus ponderosae*), did not have more than seven paralogues in this GST class [37]. Based on this pattern, they hypothesized that the expansion of Sigma GST genes occurred specifically in the beetle lineage containing *Oryctes* species. Assuming that initial *O. borbonicus* assembly contained divergent haplotypes that were not correctly haploidized [38], the likelihood of an erroneous inference of gene duplications in this taxon is high. Our *O. rhinoceros* assembly and annotation recovered nine Sigma GST genes grouped on two contigs (Fig. 5B, Additional files 4 and 5). We then analysed genome annotations of two other beetles whose assemblies also showed very high BUSCO completeness (> 98%), *O. taurus* and *T. castaneum,* as well as the annotation of *Drosophila melanogaster* that is considered a gold standard for this genomic resource in insects. Sigma GST is found in only one copy in *D. melanogaster*, while five paralogs were detected in *T. castaneum* and 12 in *O. taurus* (Fig. 5B), Based on this limited taxon sampling, there is an indication that sigma GST family expansion occurred in the Scarabaidae lineage, as both *O. taurus* and *O. rhinoceros* (Scarabaeidae) contain more sigma GST genes than *T. castaneum* (Tenebrionidae), and these sigma duplications might have an important role in eliminating the by-products of oxidative stress [61]. However, more genome assemblies and annotations of very high accuracy and completeness are needed across Coleoptera to be able to confidently infer evolutionary expansion of gene families in this insect order.

Application of genomic resources for *O. rhinoceros* management

RNAi target discovery

RNA interference (RNAi) is a promising new approach for insect pest control, particularly for beetles that exhibit a robust environmental RNAi response [62, 63]. RNAi is a highly-specific gene-silencing mechanism in which double-stranded RNA (dsRNA) directs cleavage of complementary endogenous mRNA. When targeting essential insect genes, RNAi causes rapid mortality and could be developed into a control tool that is integrated with other pest management tactics.

Through the mining of our *O. rhinoceros* assembly and annotation, we identified orthologs of all 30 genes (Supplemental Table 5) that were experimentally validated as effective RNAi targets in *T. castaneum*, ten of which were also validated in *Diabrotica v. virgifera* and four in *Brassicogethes aeneus* [64]. The strongest candidates for initial testing in *O. rhinoceros* are orthologs of *D. melanogaster*'s *Prp19, Spt5* and *RPII-215* (Supplemental Table 5), as they exhibited > 79% mortality upon injection or feeding with dsRNA in at least two of the three tested beetles [65].

Investigating interactions with a biocontrol agent

The assembled and annotated genome of *O. rhinoceros* provides an excellent opportunity to get genome-wide insight into the interaction between this insect pest and its control agent, Oryctes rhinoceros nudivirus (OrNV). For example, differences in the pattern on genome-wide expression can be traced between insects that have been experimentally infected with OrNV and the control group (non-infected) via transcriptome analysis. This approach for identifying putative infection-responsive genes has been used to study the interaction between one of the most important crop pests, the diamond-back moth *Plutella xylostella*, and the fungal insect pathogens, *Beauveria bassiana* and *Metarhizium anisopliae*, that have been widely used as insecticides [66]. For this type of a study, having access to a high-quality genome annotation is very important, as it has been shown that quality of a genome annotation strongly influences the inference of gene expression [67]. Identifying key *O. rhinoceros* genes that respond to OrNV infection could narrow a search for the causal genomic changes underlying the suspected attenuation of OrNV pathogenicity against this beetle. Namely, the resurgence and spread of *O. rhinoceros* over the last decade is hypothesized to be driven by the emergence of the virus-tolerant beetle populations and/or less virulent OrNV strains. The molecular basis for this suspected change in the beetle-OrNV interaction could reside in the regulation of small interfering RNAs (siRNAs) that are a known part of the insect immune response to viral infections [8, 68]. Our annotation contains the predictions for various non-protein-coding RNAs, laying a good foundation for further in-depth characterization of these regulatory genomic elements in *O. rhinoceros*.

Conclusions

We provide a highly contiguous and accurate nuclear genome assembly and structural annotation for an important invasive pest of palm trees, the scarab beetle *O. rhinoceros.* The assembly is based on the ONT sequencing of a single wild female, further demonstrating the utility of long-reads (and ONT sequencing in particular) in generating high-quality de novo genome assemblies from field specimens. Along with our structural annotation, this genomic resource opens up avenues for further biological discoveries aiming to

improve the management of this pest, from the functional studies of interactions with the existing biocontrol agents, to the development of new control solutions via RNAi tools.

Materials and methods

Field collection and DNA isolation

Oryctes rhinoceros adults were collected from a pheromone trap (Oryctalure, P046-Lure, ChemTica Internacional, S. A., Heredia Costa Rica) on Guadalcanal, Solomon Islands in January 2019 and preserved in 95% ethanol. High-molecular weight (HMW) DNA was extracted from a single female using a customized paramagnetic (SPRI) bead-based protocol. Specifically, we dissected pieces of tissue from four legs and the thorax, avoiding the abdomen to minimize the proportion of gut microbiota in the total DNA extract (Fig. 1A). We incubated approximately 50 mm^3 of tissue in each of the eight 1.7 mL eppendorf tubes with 360 µL ATL buffer, 40 µL of proteinase K (Qiagen Blood and Tissue DNA extraction kit) for 3 h at room temperature, while rotating end-over-end at 1 rpm. Four hundred microliters of AL buffer were added to each reaction and incubated for 10 min at room temperature, followed by the addition of 8 µL of RNase A and incubation for 5 minutes at room temperature. To remove the tissue debris, each tube was spun down for 1 min at 16,000 rcf and 600 µL of homogenate was transferred to a fresh tube. Six hundred microliters of the SPRI bead solution were added to each homogenate and incubated for 30 min while rotating at end-over-end at 1 rpm. After two washes with 75% ethanol, DNA in each tube was eluted in 50 µL of TE buffer. All eight elutions were combined and DNA quality was assessed on the 4200 Tapestation system (Agilent) and with the Qubit broad-range DNA kit (Fig. 1B). Finally, we used the Circulomics Short Read Eliminator XS kit to enrich the DNA elution with fragments longer than 10 kb (i.e. High Molecular Weight, HMW, DNA, Fig. 1C).

ONT library preparation and sequencing

One microgram of the size-selected HMW DNA was used as the starting material for the preparation of each ONT library, following the manufacturer's guidelines for the Ligation Sequencing Kit SQK-LSK109 (Oxford Nanopore Technologies, Cambridge UK). Four libraries were sequenced on four R9.4.1 flow cells using the MinION sequencing device and the ONT MinKNOW Software (Oxford Nanopore Technologies, Cambridge UK) (Fig. 1C).

Genome assembly

High accuracy base-calling from the raw ONT data was computed with Guppy v3.2.4 (for the initial assembly) and Guppy v4.2.2 (for the final assembly). The initial genome assembly (S4-i-v1) was produced with Flye version 2.5 [27] using the following input parameters: the approximate genome size (--genome-size) of 430 Mb, based on the size of an initial genome assembly in a related species *O. borbonicus* [37] two iterations of polishing (--iterations 2), aimed at correcting a small number of extra errors based on the improvements on how reads align to the corrected assembly; a minimum overlap between two reads (--min-overlap 5000) of 5000 bp; and a metagenome mode (--meta) to allow for the recovery of mitochondrial, symbiont, pathogen and other "contaminant" genomes, given that this mode is sensitive to highly variable coverage and under-represented sequences [27]. Flye version 2.8.2 was used during the iterative process for the final genome assembly (S4-7k-2v), with the parameter '--min-overlap' ranging from 5000 bp to 10,000 bp in 500 bp increments while keeping other parameters (--genome-size, --iterations, --meta) unchanged.

Identification of pathogens, symbionts, contaminants

Screening of the circular nodes with a disproportionately high coverage in the initial genome assembly graph identified the OrNV and mitogenome, and they were removed from further analyses. A linearized set of the remaining putative genome assembly sequences (contigs and scaffolds) were locally compared against the NCBI non-redundant protein (nr) database using DIAMOND [34] version 0.9.24 in 'blastx' mode. The NCBI database was downloaded from ftp.ncbi.nih.gov/blast/db/FASTA/. The results obtained with DIAMOND were analysed with the metagenome analyser tool MEGAN [36]. Any sequence not classified within Arthropoda was also checked against the NCBI's online database of nucleotide (nt/nr) and non-redundant protein sequences (nr) to identify the origin of a suspected contaminant sequence.

Polishing of the genome assembly with Illumina reads

Indel errors in the homopolymer regions represent inherent basecalling errors of the ONT platform [30]. To remove putative indel errors in the draft assembly, we used the genome polishing program Pilon version 1.23 [31] that was supplied with the spliced-aware alignments of the Illumina reads from one whole-genome sequencing library. DNA for this Illumina library originates from a female beetle collected in the same location as the female used for the ONT sequencing. Because Illumina

and ONT data did not come from the same individual, we only performed indel polishing. The Illumina sequences were produced on a HiSeq X10 platform by Novogene (Beijing, China) using the 150 bp paired-end chemistry, and were processed in Trimmomatic [69] to remove Illumina adapters, and trim and filter each read based on the minimum phred score of 20.

Evaluation of genome assemblies

The completeness of the initial genome assembly (S4-i-v3) was evaluated using: (a) alignment of DNA-seq data, (b) alignment of RNA-seq data, and (c) the recovery of the benchmarking universal single copy orthologs (BUSCOs) [33]. We used the BWA-MEM aligner with default settings and recorded the percentage of mapped Illumina reads from the whole-genome sequencing dataset (Illumina DNA library described above) and four independently-generated RNA-seq datasets from the beetle's four life stages [47] (NCBI SRA Accession: PRJNA486419) that were combined prior to alignment with the beetle genome assembly. The number of recovered universal single-copy orthologs (SCOs) was obtained using the "genome autolineage" mode in BUSCO version 4.0.6, that first searched the databases 'eukaryota_odb10' (7 species, 255 SCOs), and 'endopterygota_odb10' (56 species, 2124 SCOs). To perform the comparative benchmarking, the same BUSCO analysis was done for 39 representative assemblies in the Coleoptera out of 41 that were available in the NCBI's GenBank in October 2020 (Supplemental Table 4). Two Coleoptera genomes (for *Protaetia brevitarsis* GCA_004143645.1, and *Alaus oculatus* GCA_009852465.1) were excluded due to a persistent BUSCO analysis failure with their assembly files.

Structural annotation

To perform the structural annotation of the final genome assembly, we used the independently-generated RNA-seq datasets from the beetle's four different life stages (NCBI SRA Accession: PRJNA486419) [47]. The RNA-seq reads were pruned of the Illumina adapters and aligned against our genome assembly with the splice-aware aligner HISAT2 (Fig. 3I). The quality and completeness of these RNA-seq data were assessed through the transcriptome assembly in Trinity version 2.10.0 [49, 70], using the default settings in two modes: de novo and genome-guided assembly. To avoid incorporating the extraneous RNA sequences into the de novo transcriptome assembly, we used only those reads that were mapped with HISAT2 [48] to our S4-i-v3 genome assembly. The completeness of each transcriptome assembly was evaluated with BUSCO, using the 'auto-lineage' mode. The final genome assembly (S4-7k-2v3) and the splice-aware alignments (from HISAT2) were used for the genome-guided transcriptome assembly using the BRAKER pipeline version 2.1.4 (https://github.com/Gaius-Augustus/BRAKER/releases/tag/v2.1.4). Annotation of the non-coding RNA genes was done with tRNAscan-SE version 2.0.6 [54, 71] and Infernal version 1.1.3 [55] against the Rfam database v14.2 [56, 57] that was available on Sep 72,020 (ftp://ftp.ebi.ac.uk/pub/databases/Rfam/14.2/).

Phylogenetic analysis using 95 genes across all major lineages of Coleoptera

The reported nucleotide alignment supermatrix with the sequences from 95 genes in 373 Coleoptera and 10 outgroup taxa (from Zhang et al. [59]) was partitioned into 95 separate alignments, each of which was then translated into amino acid sequences. Blastp was used to find their orthologs in *O. rhinoceros* annotation, identifying 94 genes. One remaining ortholog was found in *O. rhinoceros* assembly using blastx. Each of the 95 gene transcript sequences in *O. rhinoceros* was then aligned against the original amino acid alignment (from Zhang et al) using CLUSTAL Omega and all 95 separate alignments were then concatenated into the resulting alignment matrix with 24,542 amino acids. Maximum likelihood tree was inferred using RAxML-ng version 1.0.2 [60] with parameters: --model Blosum62 --opt-branches on --opt-model on --tree pars{10}, rand{10} --all --bs-trees autoMRE{200} --bs-cutoff 0.03 on the unpartitioned alignment (given that Zhang et al. [59] report high congruency between partitioned and non-partitioned datasets). The final nexus tree file is available in the Supplementary Data (Additional file 4). The tree visualization was done in FigTree [72].

Analysis of the sigma GST gene family

Genes from the Sigma Glutathione-S-Transferase family in *O. rhinoceros* were identified using blastp match (E-value << e-5) between the protein translated coding DNA sequences (CDS) of *Drosophila's* Gluthatione-S-transferase S1 gene (*GstS1*) and all of the protein translated CDS derived from our annotation. The protein sequences of the identified genes were then searched in the *O. rhinoceros* assembly using blastn, in case some sigma GST genes are missing from our annotation. We also extracted all CDS from genes that had a GST term in the annotation of two Coleoptera (*O. taurus* (Scarabaidae) and *T. castaneum* (Tenebrionidae) with the highest BUSCO score for genome assembly and annotation) and *D. melanogaster,* which cover other classes of GSTs (omega, delta, epsilon, theta, zeta). Nucleotide sequences identified and extracted across all four taxa (total of 137 sequences including *D. melanogaster* sina gene (*sina*) as an outgroup) were aligned using Clustal Omega [73], and

maximum likelihood tree was inferred using RAxML version 8.2.11 [74] with parameters: -m GTRGAMMAI -f a -x 1 -N 500 -p 10 on the unpartitioned alignment.

Identification of putative RNAi targets

Identification of orthologs of 30 RNAi targets previously validated in *T. castaneum* (subset of which was also validated in *D. v. virgifera*, and *B. aeneus* [64]) was done using blastx match between reported dsRNA and CDS translation derived from *O. rhinoceros* annotation. Sequence alignment between the identified target *O. rhinoceros* ortholog and dsRNA sequences from *T. castaneum* was used to determine dsRNA sequence for each of 30 putative RNAi targets in *O. rhinoceros*.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-022-08628-z.

Additional file 1. ZIP archive for the structural genome annotation for *Oryctes rhinoceros* in general feature format (.gff).

Additional file 2. ZIP archive with an alignment supermatrix in PHYLIP format (.phy) for 95 single-copy orthologous genes used to construct the phylogenetic tree of all major lineages of Coleoptera (Fig. 5A).

Additional file 3. ZIP archive with the ML phylogenetic tree in NEXUS format (.nex) depicted in Fig. 5A.

Additional file 4. ZIP archive of an alignment supermatrix in PHYLIP format (.phy), used to construct the sigma GST genes phylogenetic tree depicted in Fig. 5B.

Additional file 5. ZIP archive of the ML tree in NEXUS format (.nex) for the sigma GST genes, depicted in Fig. 5B.

Additional file 6: Supplemental Figure 1. TapeStation® report. Genomic DNA quality control analysis of size, concentration and integrity.

Additional file 7: Supplemental Table 1. Libraray statistics. Libraray statistics for each ONT library used in the assembly process.

Additional file 8: Supplemental Table 2. Genome assembly statistics for all genome assembly versions (intermediate and final).

Additional file 9: Supplemental Table 3. Metadata and assembly statistics for the Coleoptera genomes deposited to NCBI used in the comparative analysis for BUSCO statistics.

Additional file 10: Supplemental Table 4. Comparative BUSCO statistics for Coleoptera genome assemblies and the assemblies generated in this study.

Additional file 11: Supplemental Table 5. Putative RNAi targets in *Oryctes rhinoceros*.

Acknowledgements

I.F. was supported by The University of Queensland Graduate School Research Training Program Tuition Fee Offset and Research Training Program Stipend scholarship. The authors would like to thank three anonymous reviewers for their comments and suggestions that have improved the final manuscript.

Authors' contributions

I.F., K.E. and M.J.F. conceptualised the study. I.F. and G.R. devised methodology. I.F. conducted investigation; performed formal analysis and data visualisation; implemented software and wrote the original draft manuscript. I.F. and J.H. curated the data. M.J.F., K.E, G.R., G.J.D. and M.G. provided resources. M.J.F., G.J.D., K.E. and G.R. provided supervision/mentoring. M.J.F. and K.E. obtained funding. M.J.F. administrated the project. All authors reviewed, edited and approved the manuscript.

Funding

This project was supported by the Australian Centre for International Agricultural Research project HORT/2016/185, the University of Queensland (UQECR2057321) and by core funds from the Mosquito Control Laboratory at QIMR Berghofer MRI.

Availability of data and materials

The *Oryctes rhinoceros* genome assembly S4-7k-2v3 is available in NCBI GenBank [GCA_020654165.1]. Functional annotation and additional supporting data are available as supplementary files and tables. Raw long reads used in this study are available for download via NCBI [Bioproject: PRJNA752921].

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]School of Biological Sciences, The University of Queensland, St. Lucia, Australia. [2]Mosquito Control Laboratory, QIMR Berghofer Medical Research Institute, Brisbane, QLD, Australia. [3]Research Division, Ministry of Agriculture and Livestock, Honiara, Solomon Islands.

Received: 27 September 2021 Accepted: 3 May 2022
Published online: 07 June 2022

References

1. Friederichs K. Über den gegenwärtigen Stand der Bekämpfung des Nashornkäfers (Oryctes rhinoceros L.) in Samoa. Tropenpflanzer. 1913;17:538–56.
2. Huger AM. The Oryctes virus: its detection, identification, and implementation in biological control of the coconut palm rhinoceros beetle, Oryctes rhinoceros (Coleoptera: Scarabaeidae). J Invertebr Pathol. 2005;89(1):78–84. https://doi.org/10.1016/j.jip.2005.02.010.
3. Tsatsia F, et al. The status of coconut rhinoceros beetle, Oryctes rhinoceros (L) Scarabaeidae: Dynastinae, in Solomon Islands. 2018. Available: https://www.semanticscholar.org/paper/The-status-of-Coconut-Rhinoceros-Beetle-%2C-Oryctes-(-Tsatsia-Wratten/34458fb95bcb217674efefca264ef70b74765404. Accessed 10 Sept 2021.
4. Ero M, Sar S, Kawi A, Tenakanai D, Gende P, Bonneau L. Detection of the Guam biotype (CRB-G) Oryctes rhinoceros Linneaus (Coleoptera: Scarabaeidae) in Port Moresby, Papua New Guinea. Planter. 2016; Available: https://www.semanticscholar.org/paper/Detection-of-the-Guam-biotype-(CRB-G)-Oryctes-in-Ero-Sar/95d6ad2d790e2b1a5bc604fd895e025a40f01dfb. Accessed 10 Sept 2021.
5. Reil JB, San Jose M, Rubinoff D. Low variation in nuclear and mitochondrial DNA inhibits resolution of invasion pathways across the Pacific for the coconut rhinoceros beetle (Scarabeidae: Oryctes rhinoceros). Proc Hawaii Entomol Soc. 2016;48:57–69.
6. Etebari K, et al. Examination of population genetics of the Coconut Rhinoceros Beetle (Oryctes rhinoceros) and the incidence of its biocontrol agent (Oryctes rhinoceros nudivirus) in the South Pacific Islands. Curr Res Insect Sci. 2021;1:100015. https://doi.org/10.1016/j.cris.2021.100015.
7. Bedford GO. Possibility of evolution in culture of the *Oryctes* Nudivirus of the coconut rhinoceros beetle *Oryctes rhinoceros* (Coleoptera: Scarabaeidae: Dynastinae). Adv Entomol. 2018;06(01):27–33. https://doi.org/10.4236/ae.2018.61004.

8. Etebari K, Parry R, Beltran MJB, Furlong MJ. Transcription profile and genomic variations of Oryctes rhinoceros nudivirus in coconut rhinoceros beetles. J Virol. 2020;94(22). https://doi.org/10.1128/JVI.01097-20.
9. Kirk H, Dorn S, Mazzi D. Molecular genetics and genomics generate new insights into invertebrate pest invasions. Evol Appl. 2013;6(5):842–56.
10. Rius M, Bourne S, Hornsby HG, Chapman MA. Applications of next-generation sequencing to the study of biological invasions. Curr Zool. 2015;61(3):488–504. https://doi.org/10.1093/czoolo/61.3.488.
11. Grilli S, Galizi R, Taxiarchi C. Genetic technologies for sustainable management of insect pests and disease vectors. Sustain Sci Pract Policy. 2021;13(10):5653.
12. Li F, et al. Insect genomes: progress and challenges. Insect Mol Biol. 2019;28(6):739–58.
13. Childers AK, et al. The USDA-ARS Ag100Pest initiative: high-quality genome assemblies for agricultural pest arthropod research. Insects. 2021;12(7):626. https://doi.org/10.3390/insects12070626.
14. Rašić G, Filipović I, Weeks AR, Hoffmann AA. Genome-wide SNPs lead to strong signals of geographic structure and relatedness patterns in the major arbovirus vector, Aedes aegypti. BMC Genomics. 2014;15:275.
15. Vogel E, Santos D, Mingels L, Verdonckt T-W, Broeck JV. RNA interference in insects: protecting beneficials and controlling pests. Front Physiol. 2018;9:1912.
16. Buchman A, Marshall JM, Ostrovski D, Yang T, Akbari OS. Synthetically engineered Medea gene drive system in the worldwide crop pest Drosophila suzukii. Proc Natl Acad Sci. 2018;115(18):4725–30. https://doi.org/10.1073/pnas.1713139115.
17. Carballar-Lejarazú R, et al. Next-generation gene drive for population modification of the malaria vector mosquito, Anopheles gambiae. Proc Natl Acad Sci. 2020;117(37):22805–14. https://doi.org/10.1073/pnas.2010214117.
18. Hotaling S, et al. Long reads are revolutionizing 20 years of insect genome sequencing. Genome Biol Evol. 2021;13(8). https://doi.org/10.1093/gbe/evab138.
19. Sabina J, Leamon JH. Bias in whole genome amplification: causes and considerations. Methods Mol Biol. 2015;1347:15–41.
20. Turissini DA, Gamez S, White BJ. Genome-wide patterns of polymorphism in an inbred line of the African malaria mosquito Anopheles gambiae. Genome Biol Evol. 2014;6(11):3094–104. https://doi.org/10.1093/gbe/evu243.
21. Asalone KC, et al. Regional sequence expansion or collapse in heterozygous genome assemblies. PLoS Comput Biol. 2020;16(7):e1008104.
22. Kingan SB, et al. A high-quality de novo genome assembly from a single mosquito using PacBio sequencing. Genes. 2019;10(1):62. https://doi.org/10.3390/genes10010062.
23. Kingan SB, et al. A high-quality genome assembly from a single, field-collected spotted lanternfly (Lycorma delicatula) using the PacBio Sequel II system. Gigascience. 2019;8(10). https://doi.org/10.1093/gigascience/giz122.
24. Adams M, et al. One fly-one genome: chromosome-scale genome assembly of a single outbred Drosophila melanogaster. Nucleic Acids Res. 2020;48(13):e75.
25. Filipović I, Hereward JP, Rašić G, Devine GJ, Furlong MJ, Etebari K. The complete mitochondrial genome sequence of (Coleoptera: Scarabaeidae) based on long-read nanopore sequencing. PeerJ. 2021;9:e10552.
26. Etebari K, Filipović I, Rašić G, Devine GJ, Tsatsia H, Furlong MJ. Complete genome sequence of Oryctes rhinoceros nudivirus isolated from the coconut rhinoceros beetle in Solomon Islands. Virus Res. 2020;278:197864.
27. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019;37(5):540–6.
28. Cameron SL. How to sequence and annotate insect mitochondrial genomes for systematic and comparative genomics research: sequencing insect mt genomes. Syst Entomol. 2014;39(3):400–11.
29. Marschall KJ. Introduction of a new virus disease of the coconut rhinoceros beetle in Western Samoa. Nature. 1970;225(5229):288–9.
30. Mikheyev AS, Tin MMY. A first look at the Oxford Nanopore MinION sequencer. Mol Ecol Resour. 2014;14(6):1097–102.
31. Walker BJ, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 2014;9(11):e112963.
32. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM: arXiv [q-bio.GN]; 2013. Available: http://arxiv.org/abs/1303.3997.
33. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210–2.
34. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12(1):59–60.
35. Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic data. Genome Res. 2007;17(3):377–86.
36. Huson DH, et al. MEGAN community edition - interactive exploration and analysis of large-scale microbiome sequencing data. PLoS Computat Biol. 2016;12(6):e1004957. https://doi.org/10.1371/journal.pcbi.1004957.
37. Meyer JM, Markov GV, Baskaran P, Herrmann M, Sommer RJ, Rödelsperger C. Draft genome of the scarab beetle Oryctes borbonicus on La Réunion Island. Genome Biol Evol. 2016;8(7):2093–105. https://doi.org/10.1093/gbe/evw133.
38. Latorre SM, et al. Museum phylogenomics of extinct Oryctes beetles from the Mascarene Islands: Cold Spring Harbor Laboratory; 2020. https://doi.org/10.1101/2020.02.19.954339.
39. Heavens D, et al. How low can you go? Driving down the DNA input requirements for nanopore sequencing. bioRxiv. 2021. https://doi.org/10.1101/2021.10.15.464554.
40. Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. Bioinformatics. 2016;32(5):767–9.
41. Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. Whole-genome annotation with BRAKER. Methods Mol Biol. 2019;1962:65–95.
42. Li H, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.
43. Barnett DW, Garrison EK, Quinlan AR, Stromberg MP, Marth GT. BamTools: a C API and toolkit for analyzing and managing BAM files. Bioinformatics. 2011;27(12):1691–2. https://doi.org/10.1093/bioinformatics/btr174.
44. Lomsadze A, Burns PD, Borodovsky M. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. Nucleic Acids Res. 2014;42(15):e119.
45. Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics. 2008;24(5):637–44.
46. Stanke M, Schöffmann O, Morgenstern B, Waack S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics. 2006;7:62.
47. Arvind K, Rajesh MK, Josephrajkumar A, Grace T. Dataset of de novo assembly and functional annotation of the transcriptome of certain developmental stages of coconut rhinoceros beetle, Oryctes rhinoceros L. Data Brief. 2020;28:105036. https://doi.org/10.1016/j.dib.2019.105036.
48. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37(8):907–15.
49. Haas BJ, et al. De novo transcript sequence reconstruction from RNA-seq using the trinity platform for reference generation and analysis. Nat Protoc. 2013;8(8):1494–512.
50. Girgis HZ. Red: an intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale. BMC Bioinformatics. 2015;16(1):1–19.
51. i5K Consortium. The i5K initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. J Hered. 2013;104(5):595–600.
52. *Onthophagus taurus* annotation report. https://web.archive.org/web/20201202145121/https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Onthophagus_taurus/100/. Accessed 8 Apr 2021.
53. Herndon N, et al. Enhanced genome assembly and a new official gene set for Tribolium castaneum. BMC Genomics. 2020;21(1):47.
54. Chan PP, Lowe TM. tRNAscan-SE: searching for tRNA genes in genomic sequences. Methods Mol Biol. 2019;1962:1–14.
55. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics. 2013;29(22):2933–5.
56. Kalvari I, et al. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. Nucleic Acids Res. 2020. https://doi.org/10.1093/nar/gkaa1047.
57. Kalvari I, et al. Non-coding RNA analysis using the Rfam database. Curr Protoc Bioinformatics. 2018;62(1):e51.

58. Waterhouse RM, Seppey M, Simão FA, Zdobnov EM. Using BUSCO to assess insect genomic resources. Methods Mol Biol. 2019;1858:59–74.
59. Zhang S-Q, et al. Evolutionary history of Coleoptera revealed by extensive sampling of genes and species. Nat Commun. 2018;9(1):205.
60. Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics. 2019;35(21):4453–5.
61. Shi H, et al. Glutathione S-transferase (GST) genes in the red flour beetle, Tribolium castaneum, and comparative analysis with five additional insects. Genomics. 2012;100(5):327–35.
62. Tomoyasu Y, Miller SC, Tomita S, Schoppmeier M, Grossmann D, Bucher G. Exploring systemic RNA interference in insects: a genome-wide survey for RNAi genes in Tribolium. Genome Biol. 2008;9(1):R10.
63. Watanabe S, Adams B-L, Kong A, Masang N, Vowell T, Melzer M. Identification of genes that result in high mortality of Oryctes rhinoceros (Scarabaeidae: Coleoptera) when targeted using an RNA interference approach: implications for large invasive insects. Ann Entomol Soc Am. 2020;113(4):310–7.
64. Knorr E, et al. Knockdown of genes involved in transcription and splicing reveals novel RNAi targets for pest control. Front Agron. 2021;3. https://doi.org/10.3389/fagro.2021.715823.
65. Knorr E, et al. Gene silencing in Tribolium castaneum as a tool for the targeted identification of candidate RNAi targets in crop pests. Sci Rep. 2018;8(1):2061.
66. Chu Z-J, Wang Y-J, Ying S-H, Wang X-W, Feng M-G. Genome-wide host-pathogen interaction unveiled by transcriptomic response of diamondback moth to fungal infection. PLoS One. 2016;11(4):e0152908.
67. Zhao S, Zhang B. Impact of gene annotation on RNA-seq data analysis. In: Next generation sequencing - advances, applications and challenges; 2016. https://doi.org/10.5772/61197.
68. Williams T, Virto C, Murillo R, Caballero P. Covert infection of insects by Baculoviruses. Front Microbiol. 2017;8:1337.
69. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20.
70. Grabherr MG, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29(7):644–52. https://doi.org/10.1038/nbt.1883.
71. Chan PP, Lin BY, Mak AJ, Lowe TM. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. Nucleic Acids Res. 2021. https://doi.org/10.1093/nar/gkab688.
72. Rambaut A. FigTree. 2014. Available: http://tree.bio.ed.ac.uk/software/figtree/.
73. Sievers F, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal omega. Mol Syst Biol. 2011;7:539.
74. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9):1312–3.

Publisher's Note