G3, 2024, 14(6), jkae084
https://doi.org/10.1093/g3journal/jkae084
Advance Access Publication Date: 17 April 2024
Genome Report

A reference quality genome assembly for the jewel scarab *Chrysina gloriosa*

Terrence Sylvester,[1,2] Zachary Hoover,[1,3] Carl E. Hjelmen,[1,4] Michelle M. Jonika,[1,5] Leslie T. Blackmon,[1] James M. Alfieri,[6,7] J. Spencer Johnston,[8] Sean Chien,[1] Tahmineh Esfandani,[1] Heath Blackmon[1,5,6,*]

[1]Department of Biology, Texas A&M University, College Station, TX 77843, USA
[2]Department of Biology, University of Memphis, Memphis, TN 38111, USA
[3]Department of Biochemistry and Biophysics, Texas A&M University, College Station, TX 77843, USA
[4]Department of Biology, Utah Valley University, Orem, UT 84058, USA
[5]Interdisciplinary Program in Genetics and Genomics, Texas A&M University, College Station, TX 77843, USA
[6]Interdisciplinary Program in Ecology and Evolutionary Biology, Texas A&M University, College Station, TX 77843, USA
[7]Department of Molecular Biosciences, University of Texas at Austin, Austin, TX 78712, USA
[8]Department of Entomology, Texas A&M University, College Station, TX 77843, USA

*Corresponding author: Heath Blackmon, Department of Biology, Texas A&M University, 3258 TAMU, College Station, TX 77843, USA. Email: coleoguy@gmail.com

The jewel scarab *Chrysina gloriosa* is one of the most charismatic beetles in the United States and is found from the mountains of West Texas to the Southeastern Arizona sky islands. This species is highly sought by professional and amateur collectors worldwide due to its gleaming metallic coloration. However, the impact of the large-scale collection of this beetle on its populations is unknown, and there is a limited amount of genetic information available to make informed decisions about its conservation. As a first step, we present the genome of *C. gloriosa*, which we reconstructed using a single female specimen sampled from our ongoing effort to document population connectivity and the demographic history of this beetle. Using a combination of long-read sequencing and Omni-C data, we reconstructed the *C. gloriosa* genome at a near-chromosome level. Our genome assembly consisted of 454 scaffolds spanning 642 MB, with the 10 largest scaffolds capturing 98% of the genome. The scaffold N50 was 72 MB, and the BUSCO score was 95.5%. This genome assembly will be an essential tool to accelerate understanding *C. gloriosa* biology and help make informed decisions for the conservation of *Chrysina* and other species with similar distributions in this region. This genome assembly will further serve as a community resource for comparative genomic analysis.

Keywords: Coleoptera genomes; sex chromosomes; sky islands; beetles

Introduction

The scarab beetle *Chrysina gloriosa* LeConte 1854 (previously known as *Plusiotis gloriosa*) is a charismatic beetle found in the continental United States and 1 of the 4 beetles in the genus *Chrysina* with a range that extends into the United States (LeConte 1854; Cazier 1951; Young 1957; Hawks 2001). Commonly known as the glorious beetle or glorious scarab, *C. gloriosa* has a metallic green body with silver stripes and blue eyes. Adult *C. gloriosa* depend on juniper trees as a food source, while larval forms depend on decaying logs (Ritcher 1966). In the United States, *C. gloriosa* is currently limited to higher elevation mountains from West Texas to the mountains of Southeastern Arizona, an area commonly known as the sky islands. High-elevation regions of these sky islands act as refugia for *C. gloriosa* and 3 other *Chrysina* species reaching the United States and are thought to represent remnants of a more widespread distribution that occurred during the Pleistocene epoch (Young 1957). These species of the genus *Chrysina* are a relic of the cooler and wetter era of the Pleistocene epoch (Young 1957).

Due to its colorful nature, *C. gloriosa* is highly sought by professional and amateur collectors worldwide (Genoways and Baker 1979). With the current trend of rising global temperatures and the aggressive collection of these beetles, extreme care may be justified to preserve the genetic diversity of these beetles in each mountain range. However, the absence of biological data sets and a lack of published biological literature of not only *C. gloriosa* but other beetles in this genus makes it difficult to make informed decisions about the conservation of these beetles. Much of the published literature on *Chrysina* species, including *C. gloriosa*, focuses on physical aspects of the coloration of the beetle, such as the polarizing properties of the cuticle (e.g. Sharma *et al.* 2009; Brady and Cummings 2010).

This paper presents the reconstructed genome of *C. gloriosa* as an initial step to aid in making informed conservation decisions for this species. Furthermore, our genome assembly will be the first chromosome-level assembly constructed for this charismatic beetle genus. A single female specimen collected in 2019 was used to perform Nanopore long-read sequencing for genome construction. In addition, Omni-C data were used to scaffold the genome at the chromosome level. The *C. gloriosa* genome assembly has a total length of 642 MB and a scaffold N50 of 72 MB. Our genome assembly will serve as a reference for mapping short-read data and

Received on 23 January 2024; accepted on 29 March 2024

variant discovery in an ongoing study of the population genetics of *C. gloriosa* and serve as a helpful resource in comparative genomic analysis.

Methods

Sample collection, DNA extraction, and sequencing

In an ongoing effort to document the population connectivity, *C. gloriosa* samples were collected from West Texas and Southeast Arizona during July and August from 2017 to 2019. We collected a total of 82 samples across 5 mountain ranges (Davis Mountains, Texas; Huachuca Mountains, Arizona; Chiricahua Mountains, Arizona; Madera Canyon, Arizona; and Piloncillo Mountains, Arizona). All specimens were collected at night between 7.00 PM and 11.00 PM (i.e. the first 4 h after sundown) using a combination of black light and Mercury vapor light. A single female specimen collected from Ida Canyon in the Huachuca Mountains (31.3807°N, −110.3298°W) on 2019 July 28 was used for genome construction.

To determine the biological sex of our specimen, the abdomen was dissected and examined for reproductive structures. DNA extraction and sequencing were done at the Texas A&M Institute for Genome Sciences and Society Core facility. Leg muscle tissue was dissected, and DNA was extracted using the Nanobind insect BIG DNA kit v 0.18 (Circulomics) following the Circulomics high-molecular-weight insect DNA extraction protocol. Extracted DNA integrity was assessed using a Genomic DNA ScreenTape on a TapeStation (Agilent). The Nanopore sequencing platform was used to generate long-read sequencing, and sequencing libraries were prepared following the manufacturer's protocol using the SQK-RAD004 rapid sequencing library. Six R9.4.1 MinION flow cells generated 45.56 GB of sequencing data at an estimated 53× coverage (the estimated genome size [next section] was 850 MB using flow cytometry).

For all other 81 samples, a single muscle tissue was dissected from 1 hind leg of each beetle, and DNA extractions were performed using the QIAGEN blood and tissue DNA extraction kit (QIAGEN) following the manufacturer's protocol. A NanoDrop One (Thermo Fisher) was used to examine the DNA extraction quality, while the Quantus Fluorometer (Promega) was used to quantify DNA. DNA was sequenced through the Texas A&M AgriLife Genomics and Bioinformatics Service center (https://www.txgen.tamu.edu/) using the Illumina short-read sequencing platform and 2 NovaSeq 6000 sequencing lanes. All specimens were sequenced using the 2 × 150 bp paired-end method at 1–2× coverage.

Estimation of genome size

Flow cytometric methods following Johnston *et al.* (2019) were used to determine the *C. gloriosa* genome size. Neural tissue from individual frozen samples of *C. gloriosa* was dissected and deposited into 1 mL of Galbraith buffer. All samples were coprepared with a standard (lab stock of *Drosophila virilis*, genome size = 328 Mbp). Samples were gently ground with a Kontes "A" pestle ∼15 times to release nuclei. After passing samples through 41 mm mesh filters, samples were stained with 25 µL of 1 mg/µL propidium iodide and incubated in the dark. Samples were run on a Beckman Coulter CytoFlex flow cytometer with a 488 nm blue laser. Means of 2C nuclei fluorescence peaks were measured for both sample and standard using gating methods supplied within the instrument's software before calculating the estimated genome size.

Dovetail Omni-C library preparation and sequencing

Dovetail Genomics performed Omni-C library preparation and sequencing. For each Dovetail Omni-C library, chromatin was fixed in the nucleus with formaldehyde, digested with DNaseI, and extracted. The chromatin ends were repaired and ligated to a biotinylated bridge adapter, followed by proximity ligation of adapter-containing ends. After proximity ligation, crosslinks were reversed, and DNA was purified. Purified DNA was treated to remove biotin not internal to ligated fragments. Sequencing libraries were generated using NEBNext Ultra enzymes and Illumina-compatible adapters. Biotin-containing fragments were isolated using streptavidin beads before PCR enrichment of each library. The library was sequenced on an Illumina HiSeqX platform to produce ∼30× genome coverage.

Genome assembly

Reads from each flow cell were concatenated into a single fastq file and reads below 1,000 bp were filtered using the program Filtlong v0.2.1 (https://github.com/rrwick/Filtlong). The genome was assembled using NextDenovo v2.5.0 (Hu *et al.* 2023) with default parameters setting the genome size to 850 MB based on the genome size estimate. Finally, the genome was polished using NextPolish v1.4.0 with default parameters (Hu *et al.* 2020).

Contaminant screening was performed using BlobTools v1.1.1 (Laetsch and Blaxter 2017). To generate the BlobTools map file, raw reads were mapped against the assembled genome using minimap2 v2.24 (Li 2018). The resulting sam file was converted to bam format using the samtools (v1.12) view module and sorted and indexed using samtools sort and index modules (Li *et al.* 2009; Danecek *et al.* 2021). To generate the BlobTools hits file, a preprocessed nucleotide BLAST database was downloaded from the National Center for Biotechnology Information (NCBI) (downloaded on 2021 October 25). The BLASTN tool of NCBI BLAST v2.12.0 was used to BLAST the contigs against the preformatted BLAST database to generate the hits file (Camacho *et al.* 2009). Using the map and hits file, BlobTools was used to create a blobplot showing the contamination level in the genome assembly.

Assembly and annotation of mitochondria

The mitochondrial assembly of *Tribolium castaneum* was downloaded through the NCBI genome browser (accessed date 2022 May 25) and queried against the *C. gloriosa* assembly using NCBI BLAST v2.12.0 to filter mitochondrial contigs (Camacho *et al.* 2009). A dot plot was built using LAST v1045, comparing the *C. gloriosa* mitochondrial contig with the *T. castaneum* mitochondrial assembly for further confirmation. The LAST plots indicated that *C. gloriosa* mitochondria were assembled multiple times (back-to-back assembly). Therefore, *C. gloriosa* mitochondria were reassembled using a circularization tool preventing back-to-back assembly. We used contigs that completely covered the *T. castaneum* mitochondrial genome for the secondary assembly and circularization approach to avoid incorporation of nuclear mitochondrial DNA. All confirmed mitochondrial contigs were removed from the nuclear assembly.

All reads were mapped to the mitochondrial contig using minimap2 v2.24 and filtered aligned reads using the SAMtools view module (Li *et al.* 2009; Li 2018; Danecek *et al.* 2021). Spurious alignments were removed by filtering all mapped reads with mapping quality below 30. The resulting reads were assembled using Unicycler v0.4.9 under default settings to generate the mitochondrial assembly (Wick *et al.* 2017). The *C. gloriosa* mitochondrial contig was annotated using the MITOS 2 web server (Bernt *et al.* 2013) using RefSeq 89 Metazoa as the reference sequence and invertebrate as the genetic code. The maximum overlap parameter was set to 100, and "sensitive" was selected under the ncRNA tab. All other parameters were kept as default. Finally, the mitochondrial annotation was visualized using OG-DRAW v1.3.1 (Greiner *et al.* 2019).

Scaffolding nuclear assembly

The initial contact map was created using Juicer v2.0 with the --assembly flag to generate input files necessary for 3d-DNA (Durand *et al.* 2016; Dudchenko *et al.* 2017). The scaffolding program 3d-DNA was run with 5 rounds of miss-join correction and used sites with a mapping quality of 30 or higher for scaffolding and visualization. Next, the assembly tools module in Juicebox v1.11 (JBAT) (Dudchenko *et al.* 2018) was used to curate the assembly output from 3d-DNA, following the recommendations of Howe *et al.* (2021). However, only large-scale, easily noticeable miss-assemblies (e.g. combining separate contigs into scaffolds) were curated. The software HiC-Hiker v1.0.0 (Nakabayashi and Morishita 2020) was used to further correct fine-scale misassembles present within scaffolds using a probabilistic approach to determine the possible orientation for a given set of contigs. The maximum distance for the probability of observing a contact between 2 loci (-K flag) was set to calculate automatically. Finally, the completeness of the *C. gloriosa* genome was assessed using BUSCO v5.2.2 (Simão *et al.* 2015). In addition, scaffolds were run separately as a query to a *T. castaneum* reference genome (Tcas5.2: GCF_000002335.3) and the scarab beetle *Trypoxylus dichotomus* genome (GCA_023509865.1) with minimap2 v2.24 (Li 2018). The resulting pairwise alignment file was used to generate Circos plots between the *C. gloriosa* and *T. castaneum* and *T. dichotomus* genomes using Circos v0.69-9 (Krzywinski *et al.* 2009).

Sex chromosome identification

Both read depth and BLAST-based approaches were used to identify X chromosome scaffolds. For the read depth-based approach, filtered Illumina short reads from population-level sequencing were used. These short reads were filtered for adapter sequences and low-quality regions. Forward and reverse reads from each lane were combined to generate a single forward and a single reverse read file for each specimen. Reads were quality-checked using fastQC v0.11.9, and all reports were combined using multiQC v1.11 for analysis (Andrews 2010; Ewels *et al.* 2016). Then, fastp was used to remove adapter sequences and trim reads based on quality (Chen *et al.* 2018). Reads with a score of <20 on the Phred scale, and all unpaired reads were discarded, and only paired reads were used for mapping and sex chromosome identification.

The nuclear and mitochondrial genomes were combined into a single fasta file with both organelle and nuclear genomic information. The *C. gloriosa* genome was indexed using SAMtools faidx and BWA (v0.7.17) index modules (Li *et al.* 2009; Li and Durbin 2009; Li 2013; Danecek *et al.* 2021). All reads were mapped using the BWA-MEM algorithm, and the resulting SAM files were converted to BAM format using the SAMtools view module. Next, the mapped files were sorted by name and coordinate order using the SAMtools sort module. The SAMtools fixmate module was used to correct errors on read-pairing due to the alignment program, and the SAMtools markdup module was used to remove duplicate reads. Finally, mapped reads were extracted using the SAMtools view module. Only sites with a mapping quality >30 were kept to remove spurious mappings.

The mean read depth per scaffold was generated using the SAMtools depth command, and the mean depth of the 10 longest scaffolds was extracted, representing the 10 chromosomes of a typical scarab karyotype. In different specimens, the smallest of the 10 scaffolds had a similar or half the coverage compared with the rest of the examined scaffolds. Each specimen had varying coverage from 1× to 2× for the longest 9 scaffolds. Therefore, coverage across all 10 scaffolds was normalized to compare the coverage between specimens. For a given sample, the coverage of all scaffolds was divided by the mean coverage of the 9 longest scaffolds. From there, the coverage for scaffold 10 was compared with the rest of the scaffolds and assigned a biological sex based on the average coverage of scaffold 10.

To further validate our findings, the *C. gloriosa* scaffolds were queried against the *T. castaneum* genome and the matching scaffolds of *C. gloriosa* were compared with the *T. castaneum* X chromosome.

Genome annotation

A de novo repeat library was created using RepeatModeler v2.0.4 to annotate the repeats in the genome assembly (Flynn *et al.* 2020). The species-specific de novo repeat library was combined with the Coleoptera-specific repeat library extracted from the RepeatMasker v4.1.5 repeat database (http://www.repeatmasker.org/RepeatMasker/). RepeatMasker v4.1.5 repeat database included a curated repeat library from Dfam v3.7 and RepBase v20181026 databases (Bao *et al.* 2015; Hubley *et al.* 2016). The utility script famdb.py was used to extract Coleoptera-specific repeat sequences. Then, RepeatMasker v4.1.5 was used with the custom repeat library to generate a soft-masked genome for genome annotation. We used the utility scripts calcDivergenceFromAlign.pl and createRepeatLandscape.pl of RepeatMasker v4.1.5 to calculate the amount of divergence within repeat sequence classes.

The software BRAKER v2.1.6 was used to annotate the genome using default parameters (Brůna *et al.* 2021) with no external evidence (ab initio) and with arthropod orthologous protein sequences as external evidence. The Arthropod protein sequences were downloaded from the orthologous sequence database OrthoDB v11 for genome annotation with protein sequence evidence, which included 4,307,558 sequences (Kuznetsov *et al.* 2023). To map the protein sequences to the soft-masked genome, ProtHint v2.6.0 was used with default parameters (Brůna *et al.* 2020). The resulting output of ProtHint was then used as the input for BRAKER, together with the soft-masked genome for annotation with protein sequence evidence.

The ab initio annotation and the protein sequence-based annotation were combined to generate the final gene transcripts using EVM v2.0.0 (Haas *et al.* 2008). ProtHint uses Spaln to generate a splice-aware genome annotation of protein sequences (Gotoh 2008). The annotation produced by Spaln was combined as protein sequence alignment for EVM. Finally, the functional domains of the final set of gene transcripts were annotated using InterProScan v5.60-92.0 (Jones *et al.* 2014). We assessed the quality of the annotation using BUSCO (endopterygota_odb10 data set) and web version of the OMArk (https://omark.omabrowser.org/home/). We generated annotation summary statistics using AGAT v1.2.1 (Dainat 2023).

Results

Assembly statistics

The *C. gloriosa* (Cglo_1.0: JAYRCI000000000) genome was built using Nanopore long-reads and scaffolded with Omni-C data.

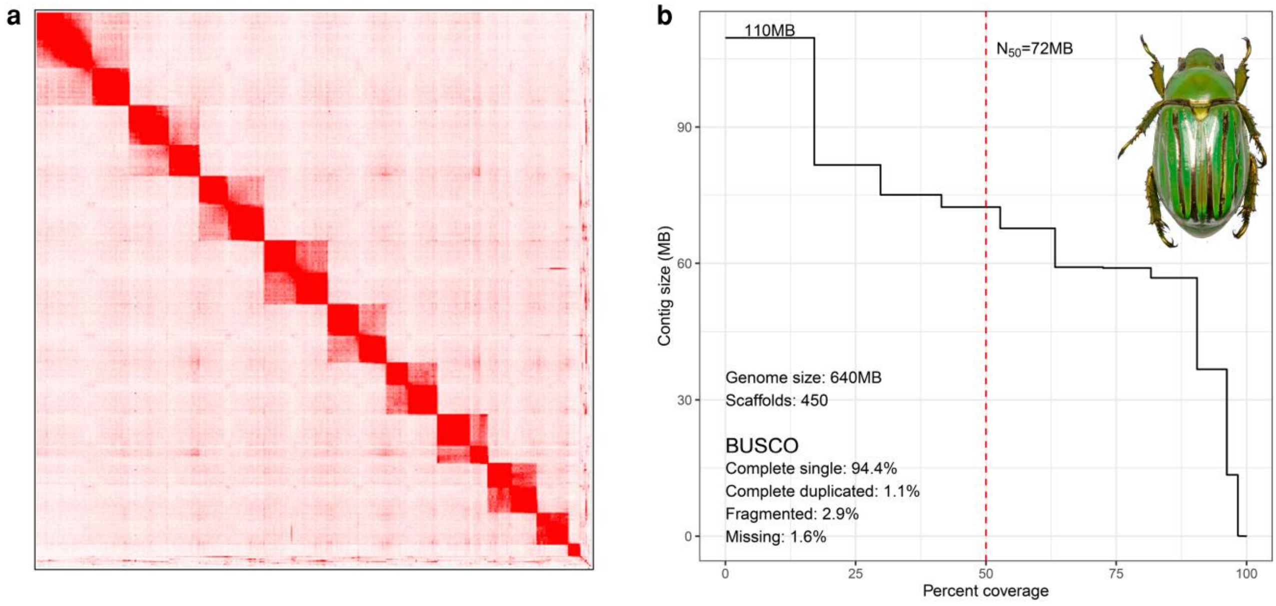

Fig. 1. a) HiC contact map of *C. gloriosa*. b) Assembly statistics and completeness of the *C. gloriosa* genome (genome size is 640 MB, the longest scaffold is 110 MB, scaffoldN50 is 72 MB, and GC composition is 35.9%). In BUSCO analysis, 95.5% of the orthologs are complete single or duplicated, and 3.5% are fragmented or missing.

The initial assembly using Nanopore data yielded 239 contigs spanning 642 MB at a coverage of 44×, contig N50 of 8.7 MB, and the largest contig of 30.1 MB. The fact that our assembly size is smaller than the estimate based on flow cytometry suggests that some repetitive content was not assembled in the genome. The completeness of the assembly was assessed using BUSCO v5.2.2 and the endopterygota_odb10 data set as the core set of 2,124 single-copy orthologs genes. BUSCO results show that 95.8% of the single-copy orthologs are present in the assembled genome (94.6% complete single copy, 1.2% complete duplicated, 2.9% fragmented, and 1.3% missing). Our contamination screening pipeline identified a single contig matched with a prokaryotic species. This contig was removed from the genome, and the new filtered genome was used for subsequent assembly processes.

The genome scaffolding pipeline initially identified 492 scaffolds, with 19 scaffolds exceeding 1 MB in size. Manual curation using JBAT placed these 19 scaffolds into 10 scaffolds representing the 10 chromosomes in a typical scarab beetle karyotype (Fig. 1a). Smaller scaffolds were also arranged when the connections were clear. This process reduced the total number of scaffolds to 454. The final assembly had a scaffold N50 of 72 MB, and the largest fragment size was 109 MB (before curation, the N50 was 37 MB, and the largest fragment size was 75 MB). The largest 10 scaffolds covered 98.3% of the genome. When assessed for the completeness of the scaffolded assembly using BUSCO v5.2.2, we did not observe a significant change in the BUSCO scores when compared with the contig level assembly (94.4% complete single copy, 1.1% complete duplicated, 2.9% fragmented, and 1.6% missing) (Fig. 1b).

Repetitive sequence and genome annotation

Annotation of repeats identified 52.94% of the genome as repetitive sequence. Most repeats were DNA transposons (20.58% of the genome), followed by retro elements (15.60% of the genome), and satellites and simple sequence repeats (0.70% of the genome). Accumulation of repetitive sequences along the scaffolds shows enrichment of repetitive elements toward the center of the chromosomes, indicating the centromeric regions (Supplementary Fig. 1). The divergence of repetitive sequences was assessed, and 3 distinct peaks were suggestive of 2 separate repetitive sequence expansion events in the past (one more recent and one in the distant past) (Fig. 2).

The genome annotation pipeline identified 19,421 gene transcripts (Supplementary Table 1). BUSCO analysis using the endopterygota_odb10 core gene set identified 77.5% of the single-copy orthologs are present in the genome annotation (76.1% complete single copy, 1.4% complete duplicated, 7.6% fragmented, and 14.9% missing). Assembly completeness using OMArk identified 86.55% of the conserved hierarchical orthogroups (4,840 at the level of Endopterygota) are present indicating that we have captured a large portion of the gene space in our structural annotation (Supplementary Fig. 2). Functional assignment using InterProScan assigned 16,257 gene transcripts with functional domains.

Genome synteny

The *C. gloriosa* assembly was compared with *T. castaneum* and *T. dichotumus genomes*. The comparison of *T. castaneum* indicates that linkage group (LG) 2 of *T. castaneum* shows synteny with 2 scaffolds of *C. gloriosa* (Fig. 3a). Furthermore, LG10 of *T. castaneum* matches several scaffolds in *C. gloriosa*. The comparison with *T. dichotumus* indicates that each of the *C. gloriosa* scaffolds has a clear 1-to-1 orthology with scaffolds in the *T. dichotumus* assembly (Fig. 3a). Despite this structural conservation, frequent inversions are documented within chromosomes. The LGX (the X chromosome) of *T. castaneum* has a clear orthology with scaffold 10 of *C. gloriosa*, providing potential evidence for the X chromosome scaffold. Mapping short-read sequences back to the genome assembly shows that all scaffolds except for scaffold 10 have similar coverage across all specimens (Supplementary Fig. 3). Some specimens have a normalized coverage near 1×, while others have a coverage near 0.5×. This further confirms that scaffold 10 is the X chromosome of *C. gloriosa*.

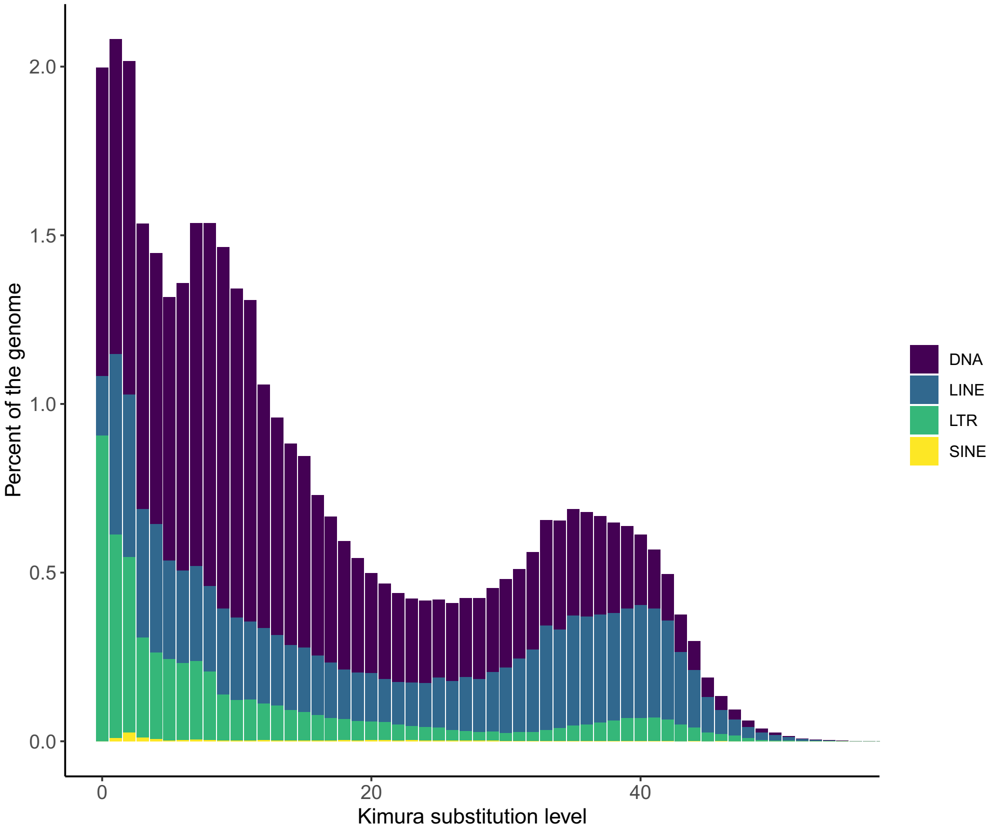

Fig. 2. Kimura substitution level of the repetitive sequences identified in the *C. gloriosa* genome.

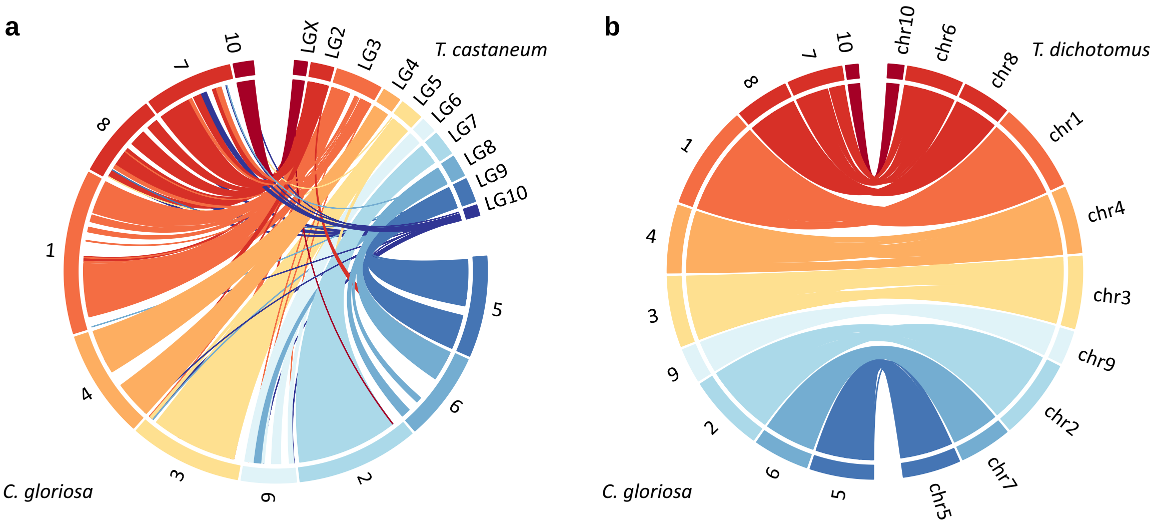

Fig. 3. Circos plots comparing the *C. gloriosa* genome with the (a) *T. castaneum* genome and the (b) *T. dichotomus* genome.

Assembly and annotation of the mitochondrial genome

The mitochondrial genome was assembled using Unicycler and annotated using the MITOS 2 web server. The *C. gloriosa* mitochondrial genome assembly was 17,644 bases in size and consisted of 13 protein-coding genes, 2 ribosomal genes, and 22 tRNA-coding genes (Supplementary Fig. 4).

Discussion

C. gloriosa is 1 of only 4 *Chrysina* species that have a distribution reaching the United States. Until now, no whole-genome data for a *Chrysina* species in the United States existed, making it difficult to understand their evolution and assess the need to conserve these charismatic beetles. As a first step in understanding the biology and evolution of this beetle, a chromosome-scale genome assembly for *C. gloriosa* was built using a combination of long-read sequencing and Omni-C data. The genome assembly consists of 10 large scaffolds capturing 98.3% of the genome, representing a typical Scarabaeidae 10 chromosomes (9 autosome pairs and an XY chromosome pair). The X chromosome scaffold of the *C. gloriosa* genome assembly was identified through synteny analysis with the *T. castaneum* genome and further validation using individual read-depth data.

To our knowledge, this is the only genome assembled to chromosome-level for the genus *Chrysina*. The only other genome assembly available in this genus is for *C. resplendens* which used short-read sequencing data. The *C. gloriosa* assembly size is slightly larger at 642 MB than *C. resplendens* assembly, which has a size of 611 MB spread across 113,068 scaffolds. However, these species have smaller genomes than *Chrysina woodii* (another *Chrysina* species that reaches the United States), whose genome size is estimated to be 856 MB using flow cytometry (Hanrahan and Johnston 2011). Compared with the *C. resplendens* genome, the completeness of our genome is relatively high, with 94.4% of the BUSCO genes present as complete single copies. In contrast, the *C. resplendens* genome has a BUSCO single-copy score of just 68.6% (Feron and Waterhouse 2022).

In summary, the *C. gloriosa* genome assembly is the first step in a long-term project to monitor population connectivity and demography of *Chrysina* species in the sky islands of the southwestern United States. This genome will serve as a starting point to answer fundamental questions about the resilience and connectivity of populations and help to determine the need to conserve this species in the future. Finally, this assembly will serve as a community resource in understanding the evolutionary dynamics of genomes in Scaraboidea and Coleoptera.

Data availability

Sequencing data are available at NCBI GenBank under the BioProject PRJNA1043134. Other data, results, and scripts used for data processing and generating figures are available at the following GitHub repository: https://github.com/Tsylvester8/Cglo-genome.

Supplemental material available at G3 online.

Funding

This work was supported by the National Institute of General Medical Sciences, National Institutes of Health grant R35 GM138098 to H.B.

Conflicts of interest

The authors declare no conflict of interest.

Literature cited

Andrews S. 2010. FastQC: A quality control tool for high throughput sequence data. [accessed 2022 Jul 12]. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

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.

Bernt M, Donath A, Jühling F, Externbrink F, Florentz C, Fritzsch G, Pütz J, Middendorf M, Stadler PF. 2013. MITOS: improved de novo metazoan mitochondrial genome annotation. Mol Phylogenet Evol. 69(2):313–319. doi:10.1016/j.ympev.2012.08.023.

Brady P, Cummings M. 2010. Differential response to circularly polarized light by the jewel scarab beetle *Chrysina gloriosa*. Am Nat. 175(5):614–620. doi:10.1086/651593.

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.

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

Cazier MA. 1951. The genera Chrysina and Plusiotis of north central Mexico (Coleoptera, Scarabaeidae). American Museum novitates; no. 1516. [accessed 2022 Jun 10]. https://digitallibrary.amnh.org/bitstream/handle/2246/2371/N1516.pdf?sequence=1.

Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 34(17):i884–i890. doi:10.1093/bioinformatics/bty560.

Dainat J. 2023. AGAT: Another Gff Analysis Toolkit to handle annotations in any GTF/GFF format. (version v1.2.0). Zendo. doi:10.5281/zenodo.3552717.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, *et al.* 2021. Twelve years of SAMtools and BCFtools. GigaScience 10(2):giab008. doi:10.1093/gigascience/giab008.

Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS, Machol I, Lander ES, Aiden AP, *et al.* 2017. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science 356(6333):92–95. doi:10.1126/science.aal3327.

Dudchenko O, Shamim MS, Batra SS, Durand NC, Musial NT, Mostofa R, Pham M, St Hilaire BG, Yao W, Stamenova E, *et al.* 2018. The Juicebox Assembly Tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under $1000. bioRxiv 254797. https://doi.org/10.1101/254797, preprint: not peer reviewed. [accessed 2022 Jul 14].

Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES, Aiden EL. 2016. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst. 3(1):95–98. doi:10.1016/j.cels.2016.07.002.

Ewels P, Magnusson M, Lundin S, Käller M. 2016. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 32(19):3047–3048. doi:10.1093/bioinformatics/btw354.

Feron R, Waterhouse RM. 2022. Assessing species coverage and assembly quality of rapidly accumulating sequenced genomes. GigaScience 11:giac006. doi:10.1093/gigascience/giac006.

Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. 2020. RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci U S A. 117(17):9451–9457. doi:10.1073/pnas.1921046117.

Genoways HH, Baker RJ. 1979. Biological investigations in the Guadalupe Mountains National Park, Texas: Proceedings of a Symposium Held at Texas Tech University, Lubbock, Texas, April 4–5, 1975. National Park Service.

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

Greiner S, Lehwark P, Bock R. 2019. OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. Nucleic Acids Res. 47(W1): W59–W64. doi:10.1093/nar/gkz238.

Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. 9(1):R7. doi:10.1186/gb-2008-9-1-r7.

Hanrahan SJ, Johnston JS. 2011. New genome size estimates of 134 species of arthropods. Chromosome Res. 19(6):809–823. doi:10.1007/s10577-011-9231-6.

Hawks DC. 2001. Taxonomic and nomenclatural changes in Chrysina and synonymic checklist of species (Scarabaeidae: Rutelinae). Occasional Papers Consortium Coleopterorum. 4: 1–8.

Howe K, Chow W, Collins J, Pelan S, Pointon D-L, Sims Y, Torrance J, Tracey A, Wood J. 2021. Significantly improving the quality of genome assemblies through curation. GigaScience 10(1): giaa153. doi:10.1093/gigascience/giaa153.

Hu J, Fan J, Sun Z, Liu S. 2020. NextPolish: a fast and efficient genome polishing tool for long-read assembly. Bioinformatics. 36(7): 2253–2255. doi:10.1093/bioinformatics/btz891.

Hu J, Wang Z, Sun Z, Hu B, Ayoola AO, Liang F, Li J, Sandoval JR, Cooper DN, Ye K, *et al.* 2023. An efficient error correction and accurate assembly tool for noisy long reads. bioRxiv 531669. https://doi.org/10.1101/2023.03.09.531669, preprint: not peer reviewed. [accessed 2024 Jan 8]

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.

Johnston JS, Bernardini A, Hjelmen CE. 2019. Genome size estimation and quantitative cytogenetics in insects. Methods Mol Biol. 1858: 15–26. doi:10.1007/978-1-4939-8775-7_2.

Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, *et al.* 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics. 30(9):1236–1240. doi:10.1093/bioinformatics/btu031.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: an information aesthetic for comparative genomics. Genome Res. 19(9):1639–1645. doi:10.1101/gr.092759.109.

Kuznetsov D, Tegenfeldt F, Manni M, Seppey M, Berkeley M, Kriventseva EV, Zdobnov EM. 2023. OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. Nucleic Acids Res. 51(D1):D445–D451. doi:10.1093/nar/gkac998.

Laetsch DR, Blaxter ML. 2017. BlobTools: interrogation of genome assemblies. F1000Res. 6(1287):1287. doi:10.12688/f1000research.12232.1.

LeConte JL. 1854. Descriptions of new Coleoptera collected by Thos. H. Webb, MD, in the years 1850–51 and 52, while Secretary to the US and Mexican Boundary Commission. Proc Acad Nat Sci Philadelphia. 7(1854–1855):220–225.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv [q-bioGN] 3997. https://doi.org/10.48550/arXiv.1303.3997, preprint: not peer reviewed.

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

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 25(14):1754–1760. doi:10.1093/bioinformatics/btp324.

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.

Nakabayashi R, Morishita S. 2020. HiC-Hiker: a probabilistic model to determine contig orientation in chromosome-length scaffolds with Hi-C. Bioinformatics. 36(13):3966–3974. doi:10.1093/bioinformatics/btaa288.

Ritcher PO. 1966. White grubs and their allies: a study of North American scarabaeoid larvae. [accessed 2022 Jul 20]. https://ir.library.oregonstate.edu/downloads/nk322k13b.

Sharma V, Crne M, Park JO, Srinivasarao M. 2009. Structural origin of circularly polarized iridescence in jeweled beetles. Science 325(5939):449–451. doi:10.1126/science.1172051.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 31(19): 3210–3212. doi:10.1093/bioinformatics/btv351.

Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol. 13(6):e1005595. doi:10.1371/journal.pcbi.1005595.

Young FN. 1957. Notes on the habits of *Plusiotis gloriosa* Le conte (Scarabaeidae). Coleopt Bull. 11(3/4):67–70. https://www.jstor.org/stable/3999012.

Editor: K. Vogel