Whole genome assembly and annotation of the lucerne weevil *Sitona discoideus*

Mandira Katuwal[1], Upendra R. Bhattarai[1,2], Craig B. Phillips[3], Neil J. Gemmell[1*], Eddy Dowle[1*]

[1]Department of Anatomy, University of Otago, Dunedin 9016, New Zealand
[2]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA
[3]Pests, Weeds & Biosecurity, AgResearch, Lincoln, Christchurch 8140, New Zealand

*Corresponding authors

Abstract

Weevils are a diverse insect group that includes many economically important invasive pest species. Despite their importance and diversity, only nine weevil genomes have been sequenced, representing a tiny fraction of this heterogeneous taxon. The genus *Sitona* consists of over 100 species, including *Sitona discoideus* (Coleoptera: Curculionidae: Entiminae), commonly known as lucerne (or alfalfa root) weevil. *Sitona discoideus* is an important pest of forage crops, particularly *Medicago* species. Using a dual sequencing approach with Oxford Nanopore MinION long-reads and 10x Genomics linked-read sequencing, we generated a high-quality hybrid genome assembly of *S. discoideus*. Benchmarks derived from evolutionarily informed expectations of gene content for near-universal single-copy orthologs comparison (BUSCO) scores are above 96% for single-copy orthologs derived from eukaryotes, arthropods, and insects. With a *de novo* repeat library, Repeatmasker annotated 81.45% of the genome as various repeat elements, of which 22.1% were unclassified. Using the MAKER2 pipeline, we annotated 10,008 protein-coding genes and 13,611 mRNAs. Furthermore, 68.84% of total predicted mRNAs and 67.90% of predicted proteins were functionally annotated to one or more of InterPro, gene ontology, and Pfam databases. This high-quality genome assembly and annotation will enable the development of critical novel genetic pest control technologies and act as an essential reference genome for broader population genetics and weevil comparative genetic studies.

Keywords: lucerne weevil, alfalfa root weevil, whole-genome sequencing, hybrid assembly, annotation, forage pest, Curculionidae

Introduction

Beetles (Coleoptera) are among the most diverse group of metazoans, with over 350,000 described species representing about one-fourth of all described species on the planet (Hunt, Bergsten et al. 2007, Stork, McBroom et al. 2015). Among beetles, the family Curculionidae ("true" weevils) contains over 60,000 described species, including many economically important invasive agricultural pests (Oberprieler, Marvaldi et al. 2007, McKenna, Sequeira et al. 2009). They present an excellent model for studying species diversity and evolution (Hunt, Bergsten et al. 2007) and insect–microbe association studies (Toju, Tanabe et al. 2013, Morera-Margarit, Pope et al. 2021). Despite their importance and striking diversity, only nine Curculionidae genomes are publicly available to date (Mei, Jing et al. 2022) limiting our understanding of this highly diversified taxon.

The lucerne weevil or alfalfa root weevil *S. discoideus* Gyllenhal 1834 (Coleoptera: Curculionidae: Entiminae) feeds mainly on *Medicago* species and occasionally on *Trifolium* species (Vink and Phillips 2007). They are strong flyers and are highly dispersive (Brockerhoff, Barratt et al. 2010). Although they are originally from southern Europe and northern Africa, they are currently found in many parts of the world, including Australia (Chadwick 1978), New Zealand (Esson 1975), the United States (O'Brien Charles 1982), Chile (Elgueta 1993), Argentina (Del Río, Lanteri et al. 2019), and South Africa (Geertsema and Volschenk 1993). *Sitona discoideus* populations in New Zealand are likely derived from a single introduction from Australia (Esson 1975) as both populations appeared closely related when compared to a Norfolk Island population (Vink and Phillips 2007). Because of its invasiveness, *S. discoideus* has established itself as a significant pest of lucerne in countries like New Zealand and Australia, costing millions of dollars annually (Hopkins 1982, Goldson and Muscroft-Taylor 1988). Adult *S. discoideus* feed on plant foliage, whereas larvae feed on the roots and root nodules. The latter stage is the more damaging as they can destroy the root nodules (Sue, Ferro et al. 1980) and significantly reduce plant productivity (Goldson, Dyson et al. 1985).

Like with many other pests, chemical control of *S. discoideus* is economically and ecologically unsustainable (Geertsema and Volschenk 1993) driving the preference for biological control strategies. Biological control has been a widely applicable tool for controlling invasive species worldwide as one of the most economical and long-term effective strategies (Clout and Williams 2009). However, heavy reliance on classical biological control of invasive species without integrating it into a more complete integrated pest management approach, increases the chances of failure because of imbalances caused by dramatic swings in pest populations (Street 2015). Hence, incorporating other control methods and information such as mechanical, cultural, ecological and genetic technologies is vital for the sustainability and effectiveness of biological control strategies (DiTomaso, Van Steenwyk et al. 2017). Furthermore, novel genetic tools possess a great potential to advance our understanding and enhance the precision and predictability of biological control (Goldson, Bourdôt et al. 2015, Street 2015).

Sequencing a pest species' genome holds a myriad of opportunities, from the development of novel biocontrol strategies through genetic modification (Teem, Alphey et al. 2020) to comparative genomic studies to understand the underlying genetic traits of interest such as parasite or pesticide resistance (Chilana, Sharma et al. 2012). However, the lack of reference genomes for the genus *Sitona*, including *S. discoideus*, hinders our genetic understanding of its biology, ecology, and evolution. Therefore, creating a reference genome for this pernicious weevil will be an essential addition to the genomic resources available for weevils and the subfamily Entiminae. The rapid development of sequencing technologies like long-reads and linked-reads and assembly algorithms can be utilized to generate reference genomes with high quality and contiguity as they reflect gene content and genome structure (Whibley, Kelley et al. 2021). Furthermore, these technologies allow us to resolve haplotype issues, particularly for creating de novo assemblies of a heterozygous diploid organism (Zhang, Wu et al. 2020).

Here, utilizing the dual sequencing approach with 10x Genomics linked-reads and Oxford nanopore long-reads, we present high-quality genome assembly and annotation of *S. discoideus.*

Materials and methods

Weevil sampling and pre-processing

Sitona discoideus specimens were collected from three sites: Lincoln (-43.64230, 172.47090), Hindon (-45.68701, 170.22198), and Grassmere (-43.055663, 171.759499), across the South Island in New Zealand, because of their availability. The adult weevils were collected from mixed grass/legume paddocks containing lucerne (*Medicago sativa*) using a modified leaf blower. Collected weevils were identified, and snap-frozen in liquid nitrogen and stored individually in a tube at -80 ˚C until further use. *Sitona discoideus* in New Zealand are parasitized by an introduced endoparasitoid wasp, *Microctonus aethiopoides*, so they were dissected under a dissection microscope to determine their parasitization status, and tissues from a single non-parasitized weevil were used for each nucleotide extraction. The weevils were washed with double-distilled water before dissection and were dissected using 1x PBS buffer (Goldson and Emberson 1981).

High Molecular Weight (HMW) genomic DNA extraction

The high molecular weight genomic DNA from the tissue of adult weevils was extracted following the 10x Genomics recommended protocol for single insect DNA purification (https://support.10xgenomics.com/permalink/7HBJeZucc80CwkMAmA4oQ2). Extracted DNA was subjected to bead clean-up using AMPure XP beads before the library preparation. DNA was quantified in Qubit, and quality was checked in nanodrop and then stored at -20 °C. Only high-quality DNA was further used for library preparation.

10X Genomics library preparation and sequencing

DNA was size selected to remove fragments shorter than 40kb using the Blue Pippin (Sage Science, USA). After size selection, 5.96 ng/µl of HMW DNA was used for Chromium 10x linked read (10x genomics, USA) library preparations following the manufacturer's protocol at the Genetic Analysis Service (GAS), University of Otago (Dunedin, New Zealand). The library was sequenced to generate 2 x 151bp paired-end reads on the Nova-seq platform (Illumina) at Garvan Institute, Australia. The Nova-seq yielded only about 30x the coverage of the estimated genome size of *S. discoideus* compared to the recommended 56x for the standard assembly coverage required by the Supernova assembler. Therefore, we further sequenced the same libraries on a single lane of rapid flowcell in Hi-seq 2500 (Illumina) at Otago Genomics Facility (OGF), University of Otago, Dunedin, New Zealand.

Oxford MinION library preparation and sequencing

Before the nanopore library preparation, extracted DNA was sheared five times with a 26-gauge needle (Terumo, Japan). We prepared four long-read sequencing libraries using DNA from three males and a female adult. Libraries were prepared using a ligation sequencing kit (SQK-LSK109) (Oxford Nanopore Technologies, Oxford, UK) following the manufacturer's protocol. The libraries thus prepared were individually loaded onto four R9 chemistry flowcells (FLO-MIN106) and sequenced for 72 hours or till pore exhaustion.

mRNA sequencing

We extracted total RNA from the different tissues and sexes of *S. discoideus* using a Direct-zol RNA MicroPrep kit (Zymo Research), using the on-filter DNAse treatment as per the manufacturer's protocol. Samples for mRNA sequencing included different tissues (head, abdomen, and gonads) from adult males and females. We extracted total RNA from each individual and tissue type separately. The extracted total RNA was accessed for quantity and purity using Qubit 2.0 Fluorometer (Life Technologies, USA) and nanodrop; high-quality samples were stored at -80°C until further processing.

RNA integrity was evaluated using a Fragment Analyzer (Advanced Analytical Technologies Inc., USA) at the OGF. The report yielded RNA quality number (RQN) values that ranged from 5.7 to 8.4, where seven out of 12 samples produced an RQN value of 7 or above. However, the collapse of the 28S peak, a widespread phenomenon for RNA extracted from insects (Winnebeck, Millar et al. 2010), might be the reason behind lower RQN values; thus, we determined the quality via the trace instead of relying on the RQN value. After the quality control step, 12 RNA samples were used for Truseq stranded mRNA libraries preparation. A single equimolar RNA library pool was generated and a single lane of 2x150bp paired-end sequencing on the HiSeq 2500 V2 Rapid Sequencing flowcell was carried out in at the OGF.

Transcriptome assembly

mRNA-seq reads were quality filtered using Trimmomatic (v.0.39) with options: SLIDINGWINDOW:4:20 LEADING:5 TRAILING:5 MINLEN:36. The filtered reads were *de novo* assembled using Trinity (v.2.8.6) with all the default options.

Genome size estimation

We performed flow cytometry analysis using a single head of *S. discoideus* with two biological replicates at Flowjoanna (Palmerston North, NZ), following the standard procedures described in (Bhattarai, Katuwal et al. 2022). Rooster red blood cells (RRBC) obtained from a domestic rooster were used as a reference sample. The raw data of nuclei peaks were analyzed using Flowjo (BD BioSciences, USA) followed by the sample's calculation of pg/nuclei.

Assembly strategies and bioinformatics pipeline

We sequenced and assembled the genome of the lucerne weevil *S. discoideus* at a total coverage depth of approximately ~100x using linked and long-read strategies. We tried several assemblers and pipelines; however, the final pipeline optimized several criteria, including the BUSCO scores (Seppey, Manni et al. 2019) for gene completeness and reference-free metrics (total length, number of contigs/scaffold, number of N's per 100kbp, N50 values and ortholog completeness). The assembly pipeline is described below (Figure 1).

The sequencing reads from 10x Genomics Chromium linked reads sequencing were assembled using Supernova assembler v.2.1.1 with default parameters (Weisenfeld, Kumar et al. 2017). Multiple criteria, including gene completeness BUSCO scores (Seppey, Manni et al. 2019), Quast (Gurevich, Saveliev et al. 2013) and reference-free metrics (discussed above), were applied to assess the assembly quality. We used the "pseudohap" style of the Supernova "mkoutput" function to export the sequence in Fasta format.

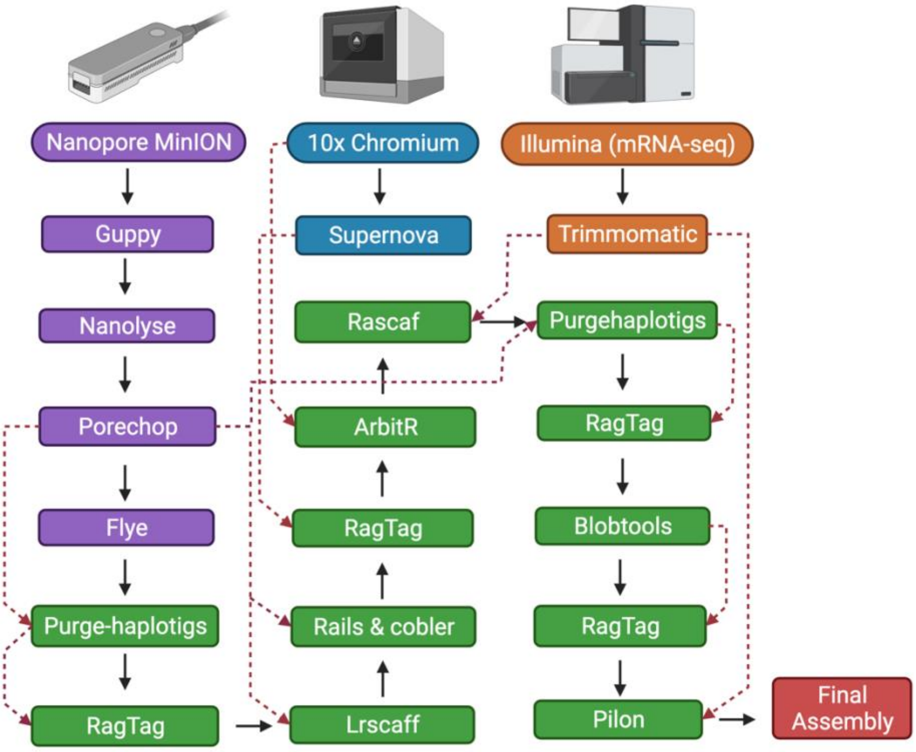

Figure 1: Schematic representation of the assembly pipeline for the *Sitona discoideus* genome. The black arrow represents the workflow, and the red dotted line represents the additional input data in the pipeline (Created with Biorender.com).

The raw Nanopore reads from four MinION runs were base-called using Guppy v.5.0.7 (Oxford Nanopore Technologies) and adapter sequences were removed with Porechop v.0.2.4 (Wick, Judd et al. 2017). The filtered reads were assembled using Flye v.2.7.1 with default parameters (Kolmogorov, Yuan et al. 2019). The assembly statistics and the BUSCO percentage were better for the long-read Flye assembly than the linked-read Supernova assembly, so the Flye assembly was used as the primary assembly. Redundant and duplicated contigs were removed using Purgehalotigs (Roach, Schmidt et al. 2018).

Using the raw-filtered Nanopore reads as input, we scaffolded and gap-closed the purged assembly using Ragtag v.2.1.0 (Alonge, Soyk et al. 2019) Lrscaff v.1.1.11 (Qin, Wu et al. 2019), Rails v.1.5.1, and Cobbler v.0.6.1 (Warren 2016). The resultant assembly was again scaffolded with the Supernova assembly using Ragtag v.2.1.0 before being further scaffolded with ArbitR v.0.2 (Hiltunen, Ryberg et al. 2021) using the raw linked-read data. LongRanger (v.2.2.2) was used to align the linked-read data for ArbitR. We then used the mRNA-sequencing data to scaffold the genome further using Rascaf (Song, Shankar et al. 2016). Redundant and duplicated haplotigs of the genome were removed using the second round of Purgehaplotigs, with the discarded haplotigs being used for scaffolding the genome through Ragtag. The assembly was quality filtered to remove contaminants with Blobtools2 (Kumar, Jones et al. 2013, Laetsch and Blaxter 2017). We removed the contigs categorized as being from the bacterial and viral superkingdom, as well as contigs with less than five times coverage and less than 1000bp length. The reads discarded as shorter length (<1000bp) and low coverage (<5x) were used for the final scaffolding step via RagTag. The final assembly underwent two polishing rounds with Pilon (v1.24) (Walker, Abeel et al. 2014) using the Illumina mRNA-seq data.

Repeat content analysis

We generated a custom repeat library to aid with annotation for *S. discoideus* using multiple de novo repeat and homology-based identifiers, including LTRharvest (Ellinghaus, Kurtz et al. 2008), LTRdigest (Steinbiss, Willhoeft et al. 2009), RepeatModeler (Flynn, Hubley et al. 2020), TransposonPSI (Haas 2007) and SINEBase (Vassetzky and Kramerov 2013). We removed the redundant reads by concatenating the individual libraries and merging the sequences with over 80 % similarity using usearch v.11.0.667 (Edgar 2010) and then classified them with Repeat Classifier. We also mapped the sequences with unknown categories present in the library against the UniProtKB/Swiss-Prot database (e-value <1e-01), where the un-annotated repeat sequences were eliminated from the library. RepeatMasker v.4.1.2 (Chen 2004) was used with the final repeat library to produce a report for genome repeat content. Because of the time and the computational resources needed, we ran RepeatMasker with the quickest run option (-qq) and skipped the bacterial insertion element check option (-no_is). The repeat library was used to input the Maker2 (v.2.31.9) pipeline (Holt and Yandell 2011) during annotation.

Genome annotation

The weevil genome annotation was performed following the MAKER2 pipeline with three iterations, including both evidence-based and ab initio gene models. The evidence-based models were used for the first round of Maker, whereas the latter two rounds used ab initio gene model predictions. For the first round with the MAKER2 pipeline, 260,683 mRNA transcripts assembled through the Trinity pipeline (Grabherr, Haas et al. 2011), along with 5,281 mRNA and 13,621 Entiminae subfamily protein sequences downloaded from NCBI, were used as inputs. Snap and BUSCO trained Augustus was used for the latter two ab initio gene prediction rounds.

Results and discussion

Genome size estimates

The estimated genome size of *S. discoideus* from the flow cytometry was 946.215 ± 31.119 MB (mean ± SD). This is the first genome size estimation for the genus *Sitona*. This genome size estimation is within the range of those reported for other Curculionidae (162.6 to 2,025 MB) in InsectBase 2.0 (Mei, Jing et al. 2021).

Transcriptome assembly

The Trinity pipeline produced an assembly of 205,961,878 bp length, with 260,683 contigs in total and 219 contigs with lengths more than 10,000 bp. The assembly has a GC ratio of 38.96% and N50 of 4512 bp. A BUSCO (v.5.2.2) analysis using the insecta_odb10 database found a complete BUSCO score of 96%.

Genome assembly

The 10x Genomics Chromium linked read library from a single individual weevil resulted in ~311 million reads with coverage of ~50x of the estimated genome size of *S. discoideus*. The Supernova assembler gave the assembly size of 340.11 MB, 144,095 total contigs with an N50 of 0.0068 MB and L50 of 6,063. We used Quast to estimate genome quality from the eukaryotic database, which reported a complete BUSCO of 35.97% and a partial BUSCO of 4.62% for the supernova assembly.

Similarly, sequencing with Nanopore MinION generated 30.4 Gb bases across ~ 7.6 million reads. The N50 of the read length and the median read length were obtained from pycoQC (v 2.5.2) and were 13,500 and 1,230, respectively, with a median PHRED score of 13.23 (Supplementary Table 1). As the primary assembler, Flye generated an assembly length of 1,889 MB. The assembly resulted in 86,442 contigs with an N50 of 0.0785 MB and an L50 value of 6,634. A complete BUSCO score of 96.70% and partial BUSCO score of 1.32 % were reported with Quast using the eukaryotic database. The long-read assembly from Flye provided better contiguity and gene completeness (Table 1); therefore, we considered it our primary assembly to process further.

Table 1. Assembly statistics of *Sitona discoideus* genome assembly. Quast BUSCO scores are to its default Eukaryota database.

	Assembly length	No. of scaffolds	N50	L50	Ns per 100 kbp	BUSCO % (Quast)	
						Complete	Partial
Supernova assembly	340,110,374	144,095	6,788	6,063	1,224.43	35.97	4.62
Flye assembly	1,889,302,767	86,442	78,533	6,634	1.56	96.70	1.32
Final hybrid assembly	1,172,662,393	6,835	297,589	952	3,338.46	96.04	1.32

Our *de novo* hybrid assembly resulted in a draft genome size of 1,172.66 MB spanning 6,835 contigs with N50 and L50 of 0.29 and 952 MB respectively, suggesting a contiguous assembly (Table 1). Furthermore, the assembly has 1,320 complete BUSCO genes using the insect database, representing 96.04% completeness, this includes 11.2% duplicated genes. The fragmented BUSCO was 1.32%. Similarly, BUSCO analysis against the Eukaryota database with Quast showed a complete and partial BUSCO of 96.04% and 1.32% respectively (Table 1). The contiguity and the completeness statistics shows that the assembly is of high quality (Figure 2). The assembly statistics after each round of processing are given in Supplementary Table 2.

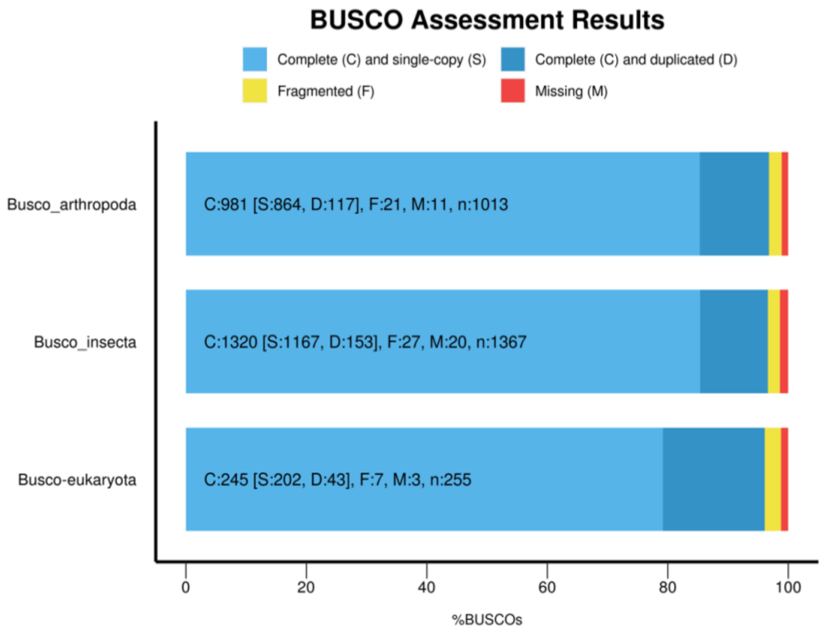

Figure 2. The BUSCO v.5 reports for the final hybrid assembly of the *Sitona discoideus* genome. BUSCO percentage (x-axis) from Arthropoda, Eukaryota, and Insecta (odt10) databases (y-axis) is shown in the bar plot. The light blue portion of the bar represents complete and single-copy orthologs, dark blue represents complete and duplicated orthologs, yellow represents fragmented BUSCO genes and red represents missing BUSCO genes.

As there are no publicly available genomes for the genus *Sitona*, we considered another forage pest weevil, the Argentine stem weevil (*Listronotus bonariensis*) to compare the genome with. The genome size of *L. bonariensis* is 1,112.4 MB with an N50 of 0.12 MB with a BUSCO completeness of 83.9% (Harrop, Le Lec et al. 2020). The *S. discoideus* genome size and N50 value are similar; however, the higher complete BUSCO score of 96.04% and a low partial BUSCO score of only 1.32% show that the assembly of *S. discoideus* is of high quality and contiguity in terms of its gene completeness.

Genome repeat contents

The Repeat Masker masked 81.45% of the genome as repeats. It reported 66.21% as the total interspersed repeats. This includes 25.06% as Retroelements, 19.95% as DNA transposons, 13.12% as Rolling-circles, and 21.2% as unclassified elements. Other repeat categories included Small RNA (0.35%), Satellites (0.36%), Simple repeats (1.37%), and Low complexity (0.04%) (Table 2).

Table 2. Repeat content analysis of *Sitona discoideus* genome

Total number of sequences:	6,835
Total length:	1,172,662,393 bp (1,133,536,485 bp excl N/X-runs)
GC level:	33.02%
Total bases masked:	955,108,331 bp (81.45 %)

	Number of elements*	**Length occupied (bp)**	**Percentage of sequence**
Retroelements	**1,085,606**	**293,820,055**	**25.06%**
SINEs:	6,879	956,131	0.08%
Penelope	61,490	15,370,643	1.31%
LINEs:	728,354	159,143,100	13.57%
CRE/SLACS	18,570	2,914,287	0.25%
L2/CR1/Rex	168,324	39,116,800	3.34%
R1/LOA/Jockey	11,294	3,863,102	0.33%
R2/R4/NeSL	3,875	1,266,357	0.11%
RTE/Bov-B	193,034	43,938,680	3.75%
L1/CIN4	6,295	1,262,565	0.11%
LTR elements:	350,373	133,720,824	11.40%
BEL/Pao	35,849	16,176,470	1.38%
Ty1/Copia	15,353	4,943,344	0.42%
Gypsy/DIRS1	210,601	92,271,804	7.87%
Retroviral	78,520	15,941,914	1.36%
DNA transposons	**1,125,058**	**233,922,632**	**19.95%**
hobo-Activator	53,056	9,640,266	0.82%
Tc1-IS630-Pogo	289,780	59,798,713	5.10%
En-Spm	-	-	0.00%
MuDR-IS905	-	-	0.00%
PiggyBac	4,994	1,304,891	0.11%
Tourist/Harbinger	12,332	2,502,841	0.21%
Other (Mirage P-element Transib)	18,072	3,965,663	0.34%
Rolling-circles	**852,338**	**153,842,660**	**13.12%**
Unclassified:	**1,220,300**	**248,624,883**	**21.20%**
Total interspersed repeats:		**776,367,570**	**66.21%**
Small RNA:	**26,400**	**4,137,760**	**0.35%**
Satellites:	**19,963**	**4,194,640**	**0.36%**
Simple repeats:	**114,977**	**16,051,131**	**1.37%**
Low complexity:	**10,922**	**514,570**	**0.04%**

Note: * most repeats fragmented by insertions or deletions were counted as one element

Genome annotation

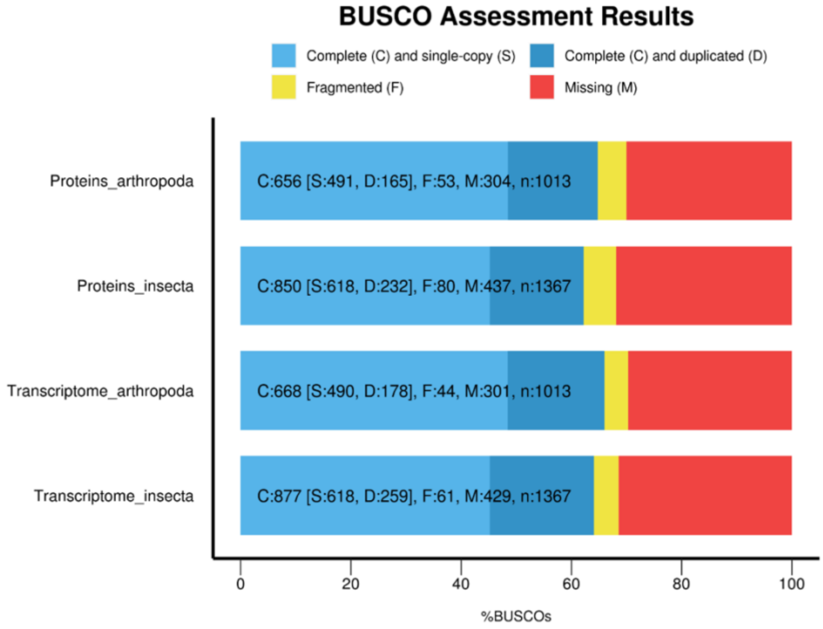

Figure 3. Annotation completeness through BUSCO database. The plot shows the BUSCO percentage (x-axis) for the annotated Proteins and Transcriptomes using Insecta_odb10 and Arthropda_odb10 database as indicated on the y-axis. BUSCO version 5 is used for the analysis.

We identified 10,008 genes and 13,611 mRNAs in the assembled genome by combining evidence-based and ab initio gene models in the MAKER2 pipeline. The total gene length is 84.63 MB constituting 7.2% of the whole genome, and the mean gene length is 8,456 bp. Similarly, the longest gene annotated is 192,956 bp, and the longest CDS is 21,423 bp (Table 3). We also functionally annotated 68.84% of total predicted mRNAs and 67.90% of predicted proteins through either one or more of the InterPro, gene ontology, and Pfam databases (Supplementary Table 3). We got 64.1% and 62.2% of complete BUSCO scores for the annotated transcriptome and annotated proteins compared with the Insecta_odb10 database (Figure 3). We found that 99% of the gene models have an AED score of 0.6 or less, indicating highly confident gene prediction (Supplementary Figure 1). This assembly is comparable to a recently curated genome of the weevil *Pissodes strobi*, where 11,382 high confidence genes were reported, and 42.9% complete BUSCO genes were identified from Endopterygota_odb10 datasets (Gagalova, Whitehill et al. 2022). The number of annotated genes in *S. discoideus* is fewer than those reported in other coleopteran genomes, like the Easter egg weevil *(Pachyrhynchus sulphureomaculatus)* with 18,741 genes (Van Dam, Cabras et al. 2021), the pine beetle *Dendroctonus ponderosae* genome with 14,342 reported genes (Keeling, Yuen et al. 2013) and the Colorado potato beetle *Leptinotarsa decemlineata* genome with 16,533 genes

(Schoville, Chen et al. 2018). A comparative study would shed more light on the difference in gene numbers between the beetles.

Table 3. Genome annotation summary for *Sitona discoideus*

Total sequence length	1,172,662,393
Number of genes	10,008
Number of mRNAs	13,611
Number of exons	88,311
Number of introns	74,700
Number of CDS	13,611
Total gene length	84,629,294
Total mRNA length	122,057,115
Total exon length	122,057,115
Total intron length	97,921,646
Total CDS length	16,937,688
Shortest gene	108
Shortest mRNA	108
Shortest exon	3
Shortest intron	5
Shortest CDS	18
Longest gene	192,956
Longest mRNA	192,956
Longest exon	11,865
Longest intron	162,917
Longest CDS	21,423
mean gene length	8,456
mean mRNA length	8,968
mean exon length	275
mean intron length	1,311
mean CDS length	1,244
% of genome covered by genes	7.2
% of genome covered by CDS	1.4
mean mRNAs per gene	1
mean exons per mRNA	6
mean introns per mRNA	5

Here, we report a high-quality assembled and annotated reference genome of *S. discoideus* using a dual sequencing approach, linked and long reads. This genome will aid in a wide range of genetic, genomic, and phylogenetic studies, particularly for the genus *Sitona* and other weevils of the subfamily Entiminae. More crucially this high-quality genome will guide our understanding of an economically important insect pest for which no management methods except biological control are available.

Conflict of Interest

The authors have no conflict of interest to disclose.

Funder Information

This study was supported by the Ministry of Business, Innovation and Employment of New Zealand via CONT-46253-CRFSI-AGR and the University of Otago.

Acknowledgments

We thank Dr. Scott Hardwick (AgResearch, Lincoln), Colin Ferguson (AgResearch, Invermay) for making the weevil samples available. We also thank the support team at New Zealand eScience Infrastructure (NeSI) for providing their HPC platform for the analysis.

References:

Alonge, M., S. Soyk, S. Ramakrishnan, X. Wang, S. Goodwin, F. J. Sedlazeck, Z. B. Lippman and M. C. Schatz (2019). "RaGOO: fast and accurate reference-guided scaffolding of draft genomes." Genome biology **20**(1): 1-17.

Bhattarai, U. R., M. Katuwal, R. Poulin, N. J. Gemmell and E. Dowle (2022). "A high-quality genome assembly and annotation of the European earwig *Forficula auricularia*." bioRxiv.

Brockerhoff, E. G., B. I. Barratt, J. R. Beggs, L. L. Fagan, K. Malcolm, C. B. Phillips and C. J. Vink (2010). "Feathers to Fur." New Zealand Journal of Ecology **34**(1): 158-174.

Chadwick, C. (1978). "Distribution and food plants of certain Curculionoidea (Coleoptera) with special reference to New South Wales." General and Applied Entomology: The Journal of the Entomological Society of New South Wales **10**: 3-38.

Chen, N. (2004). "Using Repeat Masker to identify repetitive elements in genomic sequences." Current protocols in bioinformatics **5**(1): 4.10. 11-14.10. 14.

Chilana, P., A. Sharma and A. Rai (2012). "Insect genomic resources: status, availability and future." Current Science: 571-580.

Clout, M. N. and P. A. Williams (2009). Invasive species management: a handbook of principles and techniques, Oxford University Press.

Del Río, M. G., A. A. Lanteri, C. G. Gittins Lopez, V. I. Rivero, D. A. González and M. F. Lopez Armengol (2019). "Primer registro de la plaga exótica *Sitona discoideus* Gyllenhal 1834 (Coleoptera: Curculionidae) en Argentina." Revista de la Sociedad Entomológica Argentina **78**(2): 1-4.

DiTomaso, J. M., R. A. Van Steenwyk, R. M. Nowierski, J. L. Vollmer, E. Lane, E. Chilton, P. L. Burch, P. E. Cowan, K. Zimmerman and C. P. Dionigi (2017). "Enhancing the effectiveness of biological control programs of invasive species through a more comprehensive pest management approach." Pest management science **73**(1): 9-13.

Edgar, R. C. (2010). "Search and clustering orders of magnitude faster than BLAST." Bioinformatics **26**(19): 2460-2461.

Elgueta, M. (1993). Las especies de Curculionoidea (Insecta: Coleoptera) de interés agrícola en Chile, Museo Nacional de Historia Natural.

Ellinghaus, D., S. Kurtz and U. Willhoeft (2008). "LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons." BMC bioinformatics **9**(1): 1-14.

Esson, M. (1975). Notes on the biology and distribution of three recently discovered exotic weevil pests in Hawkes Bay. Proceedings of the Twenty-eighth New Zealand Weed and Pest Control Conference, Angus Inn, Hastings, August 5 to 7, 1975., New Zealand Weed and Pest Control Society, Inc.

Flynn, J. M., R. Hubley, C. Goubert, J. Rosen, A. G. Clark, C. Feschotte and A. F. Smit (2020). "RepeatModeler2 for automated genomic discovery of transposable element families." Proceedings of the National Academy of Sciences **117**(17): 9451-9457.

Gagalova, K. K., J. G. Whitehill, L. Culibrk, D. Lin, V. Lévesque-Tremblay, C. I. Keeling, L. Coombe, M. M. Yuen, I. Birol and J. Bohlmann (2022). "The genome of the forest insect pest *Pissodes strobi* reveals genome expansion and evidence of a Wolbachia endosymbiont." G3 **12**(4): jkac038.

Geertsema, H. and E. Volschenk (1993). "First record of *Sitona discoideus* Gyllenhal (Coleoptera: Curculionidae): a pest of lucerne, in South Africa." PHYTOPHYLACTICA-PRETORIA- **25**: 275-275.

Goldson, S., G. Bourdôt, E. Brockerhoff, A. Byrom, M. Clout, M. McGlone, W. Nelson, A. Popay, D. Suckling and M. Templeton (2015). "New Zealand pest management: current and future challenges." Journal of the Royal Society of New Zealand **45**(1): 31-58.

Goldson, S., C. Dyson, J. Proffitt, E. Frampton and J. Logan (1985). "The effect of *Sitona discoideus* Gyllenhal (Coleoptera: Curculionidae) on lucerne yields in New Zealand." Bulletin of entomological research **75**(3): 429-442.

Goldson, S. and R. Emberson (1981). "Reproductive morphology of the Argentine stem weevil, *Hyperodes bonariensis* (Coleoptera: Curculionidae)." New Zealand journal of zoology **8**(1): 67-77.

Goldson, S. and K. Muscroft-Taylor (1988). "Inter-seasonal variation in Sitona discoideus Gyllenhal (Coleoptera: Curculionidae) larval damage to lucerne in Canterbury and the economics of insecticidal control." New Zealand journal of agricultural research **31**(3): 339-346.

Grabherr, M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury and Q. Zeng (2011). "Full-length transcriptome assembly from RNA-Seq data without a reference genome." Nature biotechnology **29**(7): 644-652.

Gurevich, A., V. Saveliev, N. Vyahhi and G. Tesler (2013). "QUAST: quality assessment tool for genome assemblies." Bioinformatics **29**(8): 1072-1075.

Haas, B. (2007). "TransposonPSI: an application of PSI-Blast to mine (retro-) transposon ORF homologies." Broad Institute, Cambridge, MA, USA.

Harrop, T. W., M. F. Le Lec, R. Jauregui, S. E. Taylor, S. N. Inwood, T. van Stijn, H. Henry, J. Skelly, S. Ganesh and R. L. Ashby (2020). "Genetic diversity in invasive populations of argentine stem weevil associated with adaptation to biocontrol." Insects **11**(7): 441.

Hiltunen, M., M. Ryberg and H. Johannesson (2021). "ARBitR: an overlap-aware genome assembly scaffolder for linked reads." Bioinformatics **37**(15): 2203-2205.

Holt, C. and M. Yandell (2011). "MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects." BMC bioinformatics **12**(1): 1-14.

Hopkins, D. (1982). Establishment and spread of the sitona weevil parasite microctonus aethiopoides in south Australia [Sitona discoideus]. Proceedings of the 3rd Australasian

Conference on Grassland Invertebrate Ecology/KE Lee, editor, Plympton, SA: Government Printing Division,[1982].

Hunt, T., J. Bergsten, Z. Levkanicova, A. Papadopoulou, O. S. John, R. Wild, P. M. Hammond, D. Ahrens, M. Balke and M. S. Caterino (2007). "A comprehensive phylogeny of beetles reveals the evolutionary origins of a superradiation." Science **318**(5858): 1913-1916.

Keeling, C. I., M. M. Yuen, N. Y. Liao, T. R. Docking, S. K. Chan, G. A. Taylor, D. L. Palmquist, S. D. Jackman, A. Nguyen and M. Li (2013). "Draft genome of the mountain pine beetle, *Dendroctonus ponderosae* Hopkins, a major forest pest." Genome biology **14**(3): R27.

Kolmogorov, M., J. Yuan, Y. Lin and P. A. Pevzner (2019). "Assembly of long, error-prone reads using repeat graphs." Nature biotechnology **37**(5): 540-546.

Kumar, S., M. Jones, G. Koutsovoulos, M. Clarke and M. Blaxter (2013). "Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots." Frontiers in genetics **4**: 237.

Laetsch, D. R. and M. L. Blaxter (2017). "BlobTools: Interrogation of genome assemblies." F1000Research **6**(1287): 1287.

McKenna, D. D., A. S. Sequeira, A. E. Marvaldi and B. D. Farrell (2009). "Temporal lags and overlap in the diversification of weevils and flowering plants." Proceedings of the National Academy of Sciences **106**(17): 7083-7088.

Mei, Y., D. Jing, S. Tang, X. Chen, H. Chen, H. Duanmu, Y. Cong, M. Chen, X. Ye and H. Zhou (2022). "InsectBase 2.0: a comprehensive gene resource for insects." Nucleic acids research **50**(D1): D1040-D1045.

Morera-Margarit, P., T. W. Pope, C. Mitchell and A. J. Karley (2021). "Could bacterial associations determine the success of weevil species?" Annals of Applied Biology **178**(1): 51-61.

O'Brien Charles, W. (1982). "Annotated checklist of the weevils (Curculionidae sensu lato) of North America, Central America, and the West Indies (Coleoptera, Curculionidea)[Caribbean]." American Entomological Institute. Memoirs (USA). no. 34.

Oberprieler, R. G., A. E. Marvaldi and R. S. Anderson (2007). "Weevils, weevils, weevils everywhere." Zootaxa **1668**(1): 491–520-491–520.

Qin, M., S. Wu, A. Li, F. Zhao, H. Feng, L. Ding and J. Ruan (2019). "LRScaf: improving draft genomes using long noisy reads." BMC genomics **20**(1): 1-12.

Roach, M. J., S. A. Schmidt and A. R. Borneman (2018). "Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies." BMC bioinformatics **19**(1): 1-10.

Schoville, S. D., Y. H. Chen, M. N. Andersson, J. B. Benoit, A. Bhandari, J. H. Bowsher, K. Brevik, K. Cappelle, M.-J. M. Chen and A. K. Childers (2018). "A model species for agricultural pest genomics: the genome of the Colorado potato beetle, *Leptinotarsa decemlineata* (Coleoptera: Chrysomelidae)." Scientific reports **8**(1): 1-18.

Seppey, M., M. Manni and E. M. Zdobnov (2019). BUSCO: assessing genome assembly and annotation completeness. Gene prediction, Springer**:** 227-245.

Song, L., D. S. Shankar and L. Florea (2016). "Rascaf: improving genome assembly with RNA sequencing data." The plant genome **9**(3): plantgenome2016.2003.0027.

Steinbiss, S., U. Willhoeft, G. Gremme and S. Kurtz (2009). "Fine-grained annotation and classification of de novo predicted LTR retrotransposons." Nucleic acids research **37**(21): 7002-7013.

Stork, N. E., J. McBroom, C. Gely and A. J. Hamilton (2015). "New approaches narrow global species estimates for beetles, insects, and terrestrial arthropods." Proceedings of the National Academy of Sciences **112**(24): 7519-7523.

Street, C. (2015). "Enhancing the Effectiveness of Biological Control Programs of Invasive Species by Utilizing an Integrated Pest Management Approach." from https://www.doi.gov/sites/doi.gov/files/uploads/isac_biocontrols_white_paper_rev.pdf.

Sue, K., D. Ferro and R. Emberson (1980). "A rearing method for *Sitona humeralis* Stephens (Coleoptera: Curculionidae), and its development under controlled conditions." Bulletin of entomological research **70**(1): 97-102.

Teem, J. L., L. Alphey, S. Descamps, M. P. Edgington, O. Edwards, N. Gemmell, T. Harvey-Samuel, R. L. Melnick, K. P. Oh and A. J. Piaggio (2020). "Genetic biocontrol for invasive species." Frontiers in Bioengineering and Biotechnology **8**: 452.

Toju, H., A. S. Tanabe, Y. Notsu, T. Sota and T. Fukatsu (2013). "Diversification of endosymbiosis: replacements, co-speciation and promiscuity of bacteriocyte symbionts in weevils." The ISME journal **7**(7): 1378-1390.

Van Dam, M. H., A. A. Cabras, J. B. Henderson, A. J. Rominger, C. Pérez Estrada, A. D. Omer, O. Dudchenko, E. Lieberman Aiden and A. W. Lam (2021). "The Easter Egg Weevil (Pachyrhynchus) genome reveals syntenic patterns in Coleoptera across 200 million years of evolution." PLoS genetics **17**(8): e1009745.

Vassetzky, N. S. and D. A. Kramerov (2013). "SINEBase: a database and tool for SINE analysis." Nucleic acids research **41**(D1): D83-D89.

Vink, C. J. and C. B. Phillips (2007). "First record of *Sitona discoideus* Gyllenhal 1834 (Coleoptera: Curculionidae) on Norfolk Island." New Zealand Journal of Zoology **34**(4): 283-287.

Walker, B. J., T. Abeel, T. Shea, M. Priest, A. Abouelliel, S. Sakthikumar, C. A. Cuomo, Q. Zeng, J. Wortman and S. K. Young (2014). "Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement." PloS one **9**(11): e112963.

Warren, R. L. (2016). "RAILS and Cobbler: Scaffolding and automated finishing of draft genomes using long DNA sequences." Journal of Open Source Software **1**(7): 116.

Weisenfeld, N. I., V. Kumar, P. Shah, D. M. Church and D. B. Jaffe (2017). "Direct determination of diploid genome sequences." Genome research **27**(5): 757-767.

Whibley, A., J. L. Kelley and S. R. Narum (2021). The changing face of genome assemblies: Guidance on achieving high-quality reference genomes, Wiley Online Library.

Wick, R. R., L. M. Judd, C. L. Gorrie and K. E. Holt (2017). "Completing bacterial genome assemblies with multiplex MinION sequencing." Microbial genomics **3**(10).

Winnebeck, E. C., C. D. Millar and G. R. Warman (2010). "Why does insect RNA look degraded?" Journal of Insect Science **10**(1): 159.

Zhang, X., R. Wu, Y. Wang, J. Yu and H. Tang (2020). "Unzipping haplotypes in diploid and polyploid genomes." Computational and structural biotechnology journal **18**: 66-72.