RESEARCH ARTICLE

A first draft genome of the Sugarcane borer, *Diatraea saccharalis*. [version 1; peer review: 2 approved with reservations]

Lucas Borges dos Santos ¹, João Paulo Gomes Viana²,
Fabricio José Biasotto Francischini³, Sofia Victoria Fogliata⁴, Andrea L. Joyce⁵,
Anete Pereira de Souza^{1,6}, María Gabriela Murúa⁷, Steven J. Clough^{2,8},
Maria Imaculada Zucchi⁹

¹Center of Molecular Biology and Genetic Engineering (CBMEG), University of Campinas (UNICAMP), Campinas, São Paulo, Brazil

²Department of Crop Sciences, University of Illinois, Urbana, Illinois, USA

³Syngenta Agro SA., Uberlândia, Minas Gerais, Brazil

⁴Syngenta Agro S.A., Santa Isabel, Santa Fe, Argentina

⁵Department of Public Health, University of California, Merced, California, USA

⁶Department of Plant Biology, University of Campinas (UNICAMP), Campinas, São Paulo, Brazil

⁷Institute of Agroindustrial Technology of the Argentine Northwest, Estación Experimental Agroindustrial Obispo Colombes, National Council for Scientific and Technical Research (ITANOA-EEAOC-CONICET), Las Talitas, Tucumán, Argentina

⁸US Department of Agriculture - Agricultural Research Service, Urbana, Illinois, USA

⁹Laboratory of Conservation Genetics and Genomics, Agribusiness Technological Development of São Paulo (APTA), Piracicaba, São Paulo, Brazil

V1 First published: 23 Oct 2020, 9:1269
<https://doi.org/10.12688/f1000research.26614.1>
Latest published: 23 Oct 2020, 9:1269
<https://doi.org/10.12688/f1000research.26614.1>

Abstract

Background: The sugarcane borer (*Diatraea saccharalis*), a widely distributed moth throughout the Americas, is a pest that affects economically important crops such as sugarcane, sorghum, wheat, maize and rice. Given its significant impact on yield reduction, whole-genome information of the species is needed. Here, we report the first draft assembly of the *D. saccharalis* genome.

Methods: The genomic sequences were obtained using the Illumina HiSeq 2500 whole-genome sequencing of a single adult male specimen. We assembled the short-reads using the SPAdes software and predicted protein-coding genes using MAKER. Genome assembly completeness was assessed through BUSCO and the repetitive content by RepeatMasker.

Results: The 453 Mb assembled sequences contain 1,445 BUSCO gene orthologs and 1,161 predicted gene models identified based on homology evidence to the domestic silk moth, *Bombyx mori*. The repeat content composes 41.18% of the genomic sequences which is in the range of other lepidopteran species.

Conclusions: Functional annotation reveals that predicted gene

Open Peer Review

Approval Status ? ?

	1	2
version 1		
23 Oct 2020	view	view

1. **Jay Ghurye** , Verily Life Sciences, Cambridge, USA

2. **Steven M Van Belleghem** , University of Puerto Rico, Rio Piedras, Puerto Rico

Any reports and responses or comments on the article can be found at the end of the article.

models are involved in important cellular mechanisms such as metabolic pathways and protein synthesis. Thus, the data generated in this study expands our knowledge on the genomic characteristics of this devastating pest and provides essential resources for future genetic studies of the species.

Keywords

Diatraea saccharalis, sugarcane borer, draft genome, assembly, Lepidoptera

This article is included in the **Genomics and Genetics** gateway.

Corresponding author: Maria Imaculada Zucchi (mizucchi@gmail.com)

Author roles: **Borges dos Santos L:** Formal Analysis, Visualization, Writing – Original Draft Preparation; **Paulo Gomes Viana J:** Methodology, Supervision, Writing – Review & Editing; **José Biasotto Francischini F:** Conceptualization, Writing – Review & Editing; **Victoria Fogliata S:** Investigation, Writing – Review & Editing; **L. Joyce A:** Resources, Writing – Review & Editing; **Pereira de Souza A:** Funding Acquisition, Writing – Review & Editing; **Gabriela Murúa M:** Supervision, Writing – Review & Editing; **J. Clough S:** Resources, Supervision, Writing – Review & Editing; **Imaculada Zucchi M:** Conceptualization, Funding Acquisition, Project Administration, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This study was funded by the Brazilian National Council for Scientific and Technological Development (CNPq) [311194/2012-5], the São Paulo Research Foundation (FAPESP) [Process number 2012/50848-3] and the Coordination for the Improvement of Higher Education Personnel (Capes) [88881.161041/2017-01].

Copyright: © 2020 Borges dos Santos L *et al.* This is an open access article distributed under the terms of the **Creative Commons Attribution License**, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Borges dos Santos L, Paulo Gomes Viana J, José Biasotto Francischini F *et al.* **A first draft genome of the Sugarcane borer, *Diatraea saccharalis*. [version 1; peer review: 2 approved with reservations]** F1000Research 2020, 9:1269 <https://doi.org/10.12688/f1000research.26614.1>

First published: 23 Oct 2020, 9:1269 <https://doi.org/10.12688/f1000research.26614.1>

Introduction

Diatraea saccharalis (Fabricius), commonly known as the sugarcane borer, is an important moth pest of the family Crambidae (Lepidoptera; Crambidae) that is distributed throughout the Americas, including South America, the Caribbean, Central America, and the southeastern United States (Box, 1931; CAB International, 1989; Dyar & Heinrich, 1927). *D. saccharalis* host plants include important crops such as sugarcane (*Saccharum* spp.), sorghum (*Sorghum bicolor*), wheat (*Triticum* spp.), maize (*Zea mays* L.), and rice (*Oryza sativa* L.) (Myers, 1932; Rodríguez-del-Bosque *et al.*, 1988; Roe, 1981). The damage caused by the *D. saccharalis* larvae in young plants produce a characteristic damage, that is the appearance in series of holes across the leaves that are still rolled up, and posteriorly begin to feed off the apical meristem that is killed, resulting in a condition known as “dead heart”. In the developed plants, the larvae, after hatching, start to feed on the leaves and migrate to the sheath region where they are protected, and feed on their sheath and start to scrape the stalk. As the borer develops, generally third instar and older, it feeds almost exclusively within tunnels in stalks (Flynn *et al.*, 1984). As a result of the borer behavior, the physical strength of the mature plant stalk is reduced, decreasing the plant biomass and sugar content in more developed crops, and have an increased susceptibility to plant pathogens due to the holes made by the larvae (Cruz, 2007; Wilson *et al.*, 2017). In maize fields, a new behavior of the *D. saccharalis* was observed, where the larva may burrow into corn ears (Rodríguez-del-Bosque *et al.*, 1990).

Although the sugarcane borer has a wide distribution across the Western Hemisphere (Dyar & Heinrich, 1927), it is treated as a single species and its identification relies on morphological characteristics despite the high variability shown in these traits between populations (Dyar & Heinrich, 1927; Francischini *et al.*, 2017; Pashley *et al.*, 1990). Several studies have developed molecular markers to provide a solid differentiation from other species of the same genus (Bravo *et al.*, 2008; Francischini *et al.*, 2017; Joyce *et al.*, 2016; Pavinato *et al.*, 2013; Pavinato *et al.*, 2017) and to investigate the genetic variation among populations of *D. saccharalis* (Joyce *et al.*, 2014; Joyce *et al.*, 2016; Silva-Brandão *et al.*, 2015). Geographical comparisons using individuals from the southeastern United States and Argentina have also suggested that the species might be treated as a cryptic species complex due to the significant genetic divergence among lineages (Fogliata *et al.*, 2019; Joyce *et al.*, 2014). The high genetic variation and a wide plant host range have also supported the success of this pest in colonizing regions with different environmental conditions, and human migration may likely be a major factor that contributed to its spread throughout the Americas (Francischini *et al.*, 2019). Furthermore, the great increase of genetic diversity of the sugarcane borer in Brazil coincides with the expansion of agricultural production of sugarcane and maize crops in the Brazilian landscape (Pavinato *et al.*, 2018). This makes *D. saccharalis* an interesting organism to study a wide range of questions including ecological adaptation and evolutionary dynamics. However, such studies require a significant number of molecular resources which are currently missing. A reference

mitochondrial genome (Li *et al.*, 2011) and a transcriptome assembly have been described (Merlin & Cônsoli, 2019); however, a well-assembled reference genome of *D. saccharalis* is not yet available.

Across the approximate 170,000 lepidopteran species, there are about 80 published genomes to date (Triant *et al.*, 2018), however only a few of them have scaffolds assigned to the chromosome level [*Bombyx mori* (Kawamoto *et al.*, 2019), *Heliconius melpomene* (Heliconius Genome Consortium, 2012), *Plutella xylostella* (You *et al.*, 2013), *Melitaea cinxia* (Ahola *et al.*, 2014), *Trichoplusia ni* (Chen *et al.*, 2019), and *Cydia pomonella* (Wan *et al.*, 2019)]. Recently, a few *de novo* genome assemblies of moth species have been released, such as *Spodoptera frugiperda* (Gouin *et al.*, 2017; Kakumani *et al.*, 2014), *Thaumatopoea pityocampa* (Gschloessl *et al.*, 2018b) and *Achroia grisella* (Koseva *et al.*, 2019). Due to the complexity of insect genomes, the draft assemblies have poor continuity and contain many gaps, which may result in a loss of a large number of genomic regions which relate to biological features of the species (Li *et al.*, 2019). The lack of genetic information of this clade makes it difficult to understand some evolutionary processes such as those underlying the adaptation to plant defense systems under different environments (Davey *et al.*, 2016). The rise of next generation sequencing technologies and the improvement in assembly algorithms have been providing more resources to assemble large and more complex genomes of non-model organisms at a lower cost (Wachi *et al.*, 2018). Nevertheless, many challenges still remain when selecting the best assembly software and setting the best parameters in order to get the most accurate genome assembly.

In this study, we report the first genome sequence of *D. saccharalis* which was assembled using paired-end (PE) short reads generated by the Illumina HiSeq platform with high sequencing depth coverage. This draft genome assembly reveals important aspects of the genetics for this species such as genome size and heterozygosity levels. In addition, we identified hundreds of protein-coding gene models using a homology-based approach using *B. mori* protein information and investigated the functional annotation of these proteins. However, the difficulties in assembling long tandem-repeat regions using a short-read technology also resulted in fragmented sequences within the assembly. These data provide a vastly improved genome that will serve as a valuable resource for further genomic studies of the species in order to develop more effective approaches for the management of this pest.

Methods

DNA isolation and sequencing

D. saccharalis moths used for the DNA extraction were from an inbred colony which originated in Houma, Louisiana, and was reared on an artificial diet (Southland Products, Lake Village Arkansas). The insects were raised for about 70 generations, with new insects occasionally added to the colony. The adult moths were placed in 100% ethanol and shipped to the lab for DNA extraction. Genomic DNA was

isolated from thorax and leg tissues derived from a single adult male specimen. The tissues were ground up to become a fine powder in a 10 mM Tris-HCl (Sigma-Aldrich, USA), 400 mM NaCl (Fisher Scientific, USA) and 2 mM EDTA (Fisher Scientific, USA) solution with the addition of RNase A (New England BioLabs, USA), and the high molecular weight DNA was precipitated using 120 µl 5M NaCl (Fisher Scientific, USA) (Miller *et al.*, 1988). Prior to sequencing, we assessed the DNA concentration (620 ng) using a NanoDrop ND-1000 spectrometer (v3.8.1, Thermo Scientific, USA) and fragment size (18.2 kb) in a 0.7% agarose gel (Fisher Scientific, USA). A PCR-free shotgun strategy was used to prepare the genomic library with insert sizes ranging from 200 to 500 bp. Raw sequence data from paired-end libraries with read lengths of 2x250 were generated by an Illumina HiSeq 2500 sequencer at the University of Illinois at Urbana-Champaign.

De novo genome assembly

Prior to the assembly, the quality of raw read data from sequencing was assessed using FastQC v.0.11.8 (Andrews, 2010) and preprocessed to remove adapter sequences, base-calling duplicates, as well as the elimination of low-quality reads (Phred Score < 30) using the Trimmomatic tool v.0.39 (Bolger *et al.*, 2014). Based on assembly statistics, an optimal setup was found using ~40 million reads randomly selected which corresponds to a coverage close to 60X.

Genome size and heterozygosity rate estimations were made using a k-mer frequency distribution approach (Li & Waterman, 2003). Initially, the occurrence of unique 21-mers (where $k=21$) was counted and a distribution histogram generated using Jellyfish software v.2.0 (Marçais & Kingsford, 2011). The results were plotted using the Genome Scope tool v.0.1 (Vurture *et al.*, 2017) with recommended parameters.

The genome assembly was constructed using SPAdes v. 3.11.1 (Bankevich *et al.*, 2012). Briefly, the assembly process starts with the selection of multiple k-mer sizes ($k = 21, 33, 55, 77, 99$, and 127), followed by the construction of individual De Bruijn graphs for each size of k . Afterward, graphs were combined to calculate the distances between the k-mers and to map the edges of the final assembly graph. A set of contiguous DNA sequences, defined as contigs, were generated as output from SPAdes. Scaffolds were produced by combining contigs using read pair information and gap filling.

Gene prediction and functional annotation

Following the genome assembly, we used an established MAKER v.2 pipeline (Holt & Yandell, 2011) to predict gene structure within the *D. saccharalis* scaffolds longer than 50 kb. First, repeat sequences were masked using RepeatMasker v.4.1.0 (Smit *et al.*, 2013–2015), based on known repeat elements in RepBase (Bao *et al.*, 2015). A total of 22,510 protein sequences from the lepidopteran silkworm *B. mori* reference assembly (Duan *et al.*, 2010) were used for the homology-based prediction approach. A Python script (Santos, 2020) was made to filter out gene models having low-similarity scores

(Aligned identity < 50%) when compared to the reference protein.

The predicted protein dataset was imported into Blast2GO Basic software v.5.2.5 (Götz *et al.*, 2008) for functional annotation analysis. Blast alignments were made using the NR database (NCBI, 1988) with Taxonomy Filter parameter set to “moths” and E-value threshold $\geq 1.0e-03$. GO annotations were made in accordance with the Blast2GO protocols.

Repeat annotation

To access the fraction of the genomic sequences from *D. saccharalis* that corresponded to repetitive elements we first identified repeated families using RepeatModeler v.1.0.11 (Smit & Hubley, 2010) with the parameter “-engine ncbi” set. The custom library generated in this step was then used to mask the genome on RepeatMasker v.4.1.0 (Smit *et al.*, 2013) with default parameters. The same procedures were also applied to the *B. mori* reference genome (Kawamoto *et al.*, 2019) to compare the proportion of repeat content between these two species and verify the accuracy of this approach.

Genome assembly evaluation

To assess the accuracy of the genomic sequences, we mapped the total set of Illumina raw paired-end reads to the assembled scaffolds using the BBMap software v.38.36 (Bushnell, 2014) to analyze the accordance rate between the primary dataset and the final assembly. Subsequently, we searched for the presence and completeness of highly conserved single copy orthologs from the “Insecta” database using BUSCO v.3.0 (Simão *et al.*, 2015). The results were then compared to *B. mori* reference assembly results.

Results

Short read de novo sequencing and assembly

The genome of *D. saccharalis* was sequenced using the Illumina HiSeq 2500 system with paired-end 2x250 nt reads. We obtained over 306 Megabase pairs (Mb) of paired-end sequence data, representing 212X average genome coverage.

The haploid genome size estimate was made using the abundance of unique k-mers ($k = 21$), which corresponds to the second peak, highlighted by the dotted line in the plot (Figure 1). GenomeScope estimated a haploid genome size of 359 Mb. Additionally, this analysis revealed 0.69% of variation across the genomic sequences, indicating that the *D. saccharalis* genome of the selected individual had low heterozygosity properties, most likely due to the inbred nature of the colony used in this experiment.

An optimal coverage size of input data was defined as 60X, corresponding to 46 million randomly selected reads obtained by a Python script (Santos, 2020). The assembly of the reduced data set using SPAdes resulted in 50,460 scaffolds larger than 1 kb having a cumulative length of 453 Mb, and a scaffold N50, an indication of assembly contiguity, of 16.3 kb. Additionally, the genomic GC content corresponds to 33.75%

Figure 1. GenomeScope result of *D. saccharalis* genome. The first peak corresponds to the heterozygous and the second, the homozygous peak. Estimate of the heterozygous portion is 0.69%. Values evidenced in the subtitle correspond to the inferred total genome length (len), genome unique length percent (uniq), overall heterozygosity rate (het), mean k-mer coverage for heterozygous bases (kcov), read error rate (err), average rate of read duplications (dup) and k-mer size (k).

of the bases. Summary statistics of the assembly are described in [Table 1](#).

Genome assembly quality and completeness assessment

The alignment of raw input short-read data resulted in a total of 96.40% of the reads mapped to the assembly, of which 89.64% were properly paired to the scaffolds. To access the completeness of our *D. saccharalis* genome assembly, we used BUSCO ([Simão et al., 2015](#)) and a set of 1,658 conserved, single-copy Insecta genes. We found that our assembly was highly complete, with 87.1% (1,445 out of 1,658) of these BUSCO genes being present in the genomic sequences (86.5% single copy, 0.6% duplicated genes). The *B. mori* reference genome contains 98.5% complete BUSCO genes ([Figure 2](#)).

Repeat content

Our analysis revealed that approximately 186 Mb of the *D. saccharalis* genome were identified to be various nucleotide repeat elements of which represent over 41% of the assembled sequences ([Table 2](#)). Although the majority of these elements were unclassified (35.44%), the repeats are composed by 14.8 Mb of long interspersed nuclear elements (LINEs), 4.5 Mb of simple repeats, 3.9 Mb of DNA transposons, as

Table 1. Summarized results of the *D. saccharalis* genome assembly and parameters used.

Property	<i>D. saccharalis</i> assembly
Input paired-end reads (million)	45 M
Estimated coverage	60x
Total contigs (>1kb)	56777
Contig N50 (bases)	14183
Total scaffolds	50455
Total scaffolds > 50kb	548
Total scaffolds > 100kb	26
Total scaffold length (bases)	222605
Scaffold N50 (bases)	16315
Assembly size	453,235,217
G+C content (%)	33.75%

Figure 2. Completeness assessment of the *D. saccharalis* genome assembly with comparison to the *B. mori* reference genome using the BUSCO platform. Bars show the proportions of conserved single-copy genes found in each assembly as a proportion of the total gene set (n).

Table 2. Comparison of the genomic repetitive content between the sugarcane borer (*D. saccharalis*) and the silk moth (*B. mori*) generated with libraries from *de novo* repeat identification using RepeatModeler.

Repeat type ¹	<i>D. saccharalis</i>		<i>B. mori</i>	
	Length (Mb)	% genome	Length (Mb)	% genome
LINEs	14.88	3.28	70.66	15.87
LTR	0.36	0.08	4.83	0.97
DNA Transposons	3.9	0.86	11.22	2.52
Unclassified	160.61	35.44	112.69	25.32
Simple Repeats	4.5	0.99	6.16	1.39
Low Complexity	0.7	0.16	0.56	0.13
Total	186.66	41.18	206.16	46.32

¹LINE, long interspersed nuclear element; LTR, long terminal repeat retrotransposon.

well as retrotransposons and low complexity elements. The consistency of our findings to the sugarcane borer assembly was assessed by analyzing the repeat content present in *B. mori* chromosomes through the same pipeline. *B. mori* assembly has shown a similar pattern of high repeat genomic content,

having 46.32% of its genome composed of repeated elements. Moreover, the number of unclassified elements also compose a significant part of the repeats (112.6 Mb), followed by LINEs (70 Mb), DNA transposons (11.2 Mb), retrotransposons (4.8 Mb) and simple repeats (6.1 Mb).

Functional annotation of gene models

A total of 548 *D. saccharalis* genomic sequences longer than 50 kb were used in the MAKER homology-based gene prediction pipeline along with a library of repeated elements generated using RepeatModeler v.1.0.11 and the *B. mori* protein dataset as reference for building the models. The analysis resulted in a total of 1,394 predicted gene models among the *D. saccharalis* scaffolds, of which contain a certain level of similarity to a given *B. mori* protein sequence (Additional table S1, extended data (Santos, 2020)). Subsequently, we selected 1,161 gene models presenting high similarity to its respective reference protein (Identity $\geq 50\%$). In total, 1,094 proteins are related to the selected gene models and a small fraction of these proteins (4.6%) are the product of multiple gene models. The functional annotation pipeline was further applied using the 1,094 protein sequences, which were further assigned into different categories of their respective Gene Ontology (GO) terms using the Blast2GO software (Götz et al., 2008) (Additional table S2, extended data (Santos, 2020)). The NCBI “moth” protein database identified 1,002 proteins (91.5%) with information, and 826 (75.5%) having a gene ontology assignment. The most abundant GO categories of these proteins are shown in Figure 3.

Discussion

The rise of low cost third generation sequencing technologies has been motivating the development of innovative genomic studies involving non-model species (Ellegren, 2014). Consequently, we notice an emergent increase in the demand for genome assemblies of more organisms. Although initial “draft” assemblies might not be contiguous enough to elucidate more complex events such as whole genome duplications, these projects have been demonstrated to be essential in the investigation of important genomic characteristics of poorly studied organisms, such as repetitive regions, genome size estimation, occurrence of variation, synteny with other species, gene size evolution and protein sequence evolution (da Fonseca et al., 2016; Fournier et al., 2017; Liyanage et al., 2019).

The key objective of this study was to provide the first genome representation of the sugarcane borer, *D. saccharalis*, using a high throughput sequencing approach to produce *de novo* genomic sequences. Our analyses indicate that the species has an estimated genome size of 359 Mb and produced a total length of 453 Mb of assembled genomic sequences representing the *D. saccharalis* genome. The genomic GC content was 33.74%, which is in the range of other lepidopteran

Figure 3. Gene ontology functional classification of *B. mori* proteins identified in the *D. saccharalis* genome assembly. The results are categorized in the three main categories: cellular component (orange), molecular function (green), and biological process (purple). The y-axis indicates the number of hits identified for each category.

species such as *Ostrinia scapularis* (37.4%) (Gschloessl *et al.*, 2018a), *S. frugiperda* (32.9%) (Kakumani *et al.*, 2014), and *B. mori* (38.8%) (Kawamoto *et al.*, 2019).

Repetitive element estimates from the *de novo* assembled sequences revealed that the *D. saccharalis* genome is over 40% composed of repetitive sequences. This observation, in addition to the discrepancy between the estimated genome size and the total assembly length, suggests that the assembly of repetitive-rich regions is more difficult when using short-reads, leading to the formation of an accurate but more fragmented assembly with fairly small contigs (Peona *et al.*, 2018; Treangen & Salzberg, 2012).

The consistency of our pipeline was confirmed by analyzing the repetitiveness present in *B. mori* chromosomes. The percentage of repetitive elements found in this species (46.32%) was significantly similar to the findings from Kawamoto *et al.*, 2019 (46.45%). Additionally, the high proportion of repetitiveness also appears in other lepidopteran members such as *H. vespertilio* (50.3%) (Pippel *et al.*, 2020), *T. pityocampa* (45.3%) (Gschloessl *et al.*, 2018b) and *S. frugiperda* (29%) (Gouin *et al.*, 2017).

In terms of completeness of genic regions, our *D. saccharalis* assembly showed the presence of a significant number of conserved insect single-copy genes. Out of 1,658 Insecta BUSCO gene orthologs, only 3.9% were missing in the assembly. When compared to a well-studied species, *B. mori*, which has its genome assembled at a chromosome level, the number of missing genes is 1%, suggesting that some of these genes are likely absent from the genomes of some lepidopteran species.

Genetic similarities between the two species were also evidenced through functional annotation of predicted genes using a homology-based approach. The analyses identified over 1,000 gene models on *D. saccharalis* genome using protein information from *B. mori*. Further gene ontology annotation revealed that proteins with high identity levels between the two species are related to important cellular components such as the nucleus, plasma membranes and cytoplasm, as well as being involved in molecular mechanisms essential for cell maintenance such as the transcription process, protein synthesis and metabolic pathways. However, future additional transcriptome evidence would improve the annotation and

validate these gene models, providing more solid information about the genes present in the *D. saccharalis* genome.

To this end, the combination of a *de novo* assembly algorithm and several predictive models has successfully identified important features of the *D. saccharalis* genome for the first time, and generated sequence data that can be immediately used at several levels: the discovery of additional genomic features, the identification of SNP markers (Boutet *et al.*, 2016) and improvements on the genome annotation by the combination with transcriptomic data (Holt & Yandell, 2011). Furthermore, this will provide valuable resource for the investigation of potential novel techniques to control this pest on crops (Kirk *et al.*, 2013).

Data availability Underlying data

Whole-genome sequencing project of *Diatraea saccharalis*. BioProject, Accession number: PRJNA647758. <https://www.ebi.ac.uk/ena/browser/view/PRJNA647758>.

Raw gDNA Illumina reads of *Diatraea saccharalis*. Biosample, Accession number: SAMN15594284. <https://identifiers.org/biosample:SAMN15594284>

The complete draft genome assembly of *Diatraea saccharalis*. GenBank, Accession number: JACGTY000000000. <https://www.ncbi.nlm.nih.gov/nuccore/JACGTY000000000>.

Extended data

Zenodo: bsLucas98/Dsaccharalis_genomeassembly: A first draft genome of the sugarcane borer, *Diatraea saccharalis*. <https://doi.org/10.5281/zenodo.4067084> (Santos, 2020).

This project contains the following extended data:

- Select_subset_reads.py (script used to select a subset of raw reads)
- SupplementaryTables_DsacGenome.xlsx (Additional Table 1 and 2)
- protein_identity.py (Script used to select protein gene models having >=50% similarity to the *B. mori* model)

This data is licensed under the terms of the Creative Commons Attribution 0 1.0 Universal (CC0).

References

Ahola V, Lehtonen R, Somervuo P, *et al.*: **The *Glanville fritillaria* genome retains an ancient karyotype and reveals selective chromosomal fusions in Lepidoptera.** *Nat Commun.* 2014; 5(1): 4737–9.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Andrews S: **FASTQC. A quality control tool for high throughput sequence data.** 2010.
[Reference Source](#)

Bankevich A, Nurk S, Antipov D, *et al.*: **SPAdes: a new genome assembly**

algorithm and its applications to single-cell sequencing. *J Comput Biol.* 2012; **19**(5): 455–477.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Bao W, Kojima KK, Kohany O: **Repbased Update, a database of repetitive elements in eukaryotic genomes.** *Mob DNA.* 2015; **6**(1): 11.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Bolger AM, Lohse M, Usadel B: **Trimmomatic: a flexible trimmer for Illumina sequence data.** *Bioinformatics.* 2014; **30**(15): 2114–2120.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Boutet G, Carvalho SA, Falque M, *et al.*: **SNP discovery and genetic mapping using genotyping by sequencing of whole genome genomic DNA from a pea RIL population.** *BMC Genomics.* 2016; **17**(1): 121.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Box HE: **The Crambine genera *Diatraea* and *Xanthopherne* (Lep., Pyral.).** *Bull Entomol Res.* 1931; **22**(1): 1–50.

[Publisher Full Text](#)

Bravo JP, Silva JLC, Munhoz REF, *et al.*: **DNA barcode information for the sugar cane moth borer *Diatraea saccharalis*.** *Genet Mol Res.* 2008; **7**(3): 741–748.

[PubMed Abstract](#) | [Publisher Full Text](#)

Bushnell B: **BBMap: a fast, accurate, splice-aware aligner.** (No. LBNL-7065E) Lawrence Berkeley National Lab (LBNL), Berkeley, CA (United States). 2014.

[Reference Source](#)

CAB International: ***Diatraea saccharalis* [Distribution map]. Distribution maps of plant pests.** *Agricultural Map 5* (revised). London: CABI 1989.

[Reference Source](#)

Chen W, Yang X, Tetreau G, *et al.*: **A high-quality chromosome-level genome assembly of a generalist herbivore, *Trichoplusia ni*.** *Mol Ecol Resour.* 2019; **19**(2): 485–496.

[PubMed Abstract](#) | [Publisher Full Text](#)

Cruz I: **A broca da cana-de-açúcar, *Diatraea saccharalis*, em milho, no Brasil.** *Embrapa Milho e Sorgo-Circular Técnica (INFOTECA-E).* 2007.

[Reference Source](#)

da Fonseca RR, Albrechtsen A, Themudo GE, *et al.*: **Next-generation biology: sequencing and data analysis approaches for non-model organisms.** *Mar Genomics.* 2016; **30**: 3–13.

[PubMed Abstract](#) | [Publisher Full Text](#)

Davey JW, Chouteau M, Barker SL, *et al.*: **Major improvements to the *Heliconius melpomene* genome assembly used to confirm 10 chromosome fusion events in 6 million years of butterfly evolution.** *G3 (Bethesda).* 2016; **6**(3): 695–708.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Duan J, Li R, Cheng D, *et al.*: **SilkDB v2.0: a platform for silkworm (*Bombyx mori*) genome biology.** *Nucleic Acids Res.* 2010; **38**(Database issue): D453–456.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Dyar HG, Heinrich C: **The American moths of the genus *Diatraea* and allies.** *Proceedings of the United States National Museum.* 1927; **71**(2691): 1–48.

[Publisher Full Text](#)

Ellegren H: **Genome sequencing and population genomics in non-model organisms.** *Trends Ecol Evol.* 2014; **29**(1): 51–63.

[PubMed Abstract](#) | [Publisher Full Text](#)

Flynn JL, Reagan TE, Ogunwolu EO: **Establishment and damage of the sugarcane borer (Lepidoptera: Pyralidae) in corn as influenced by plant development.** *J Econ Entomol.* 1984; **77**(3): 691–697.

[Publisher Full Text](#)

Fogliata SV, Herrero MI, Vera MA, *et al.*: **Host plant or geographic barrier? Reproductive compatibility among *Diatraea saccharalis* populations from different host plant species and locations in Argentina.** *Entomol Exp Appl.* 2019; **167**(2): 129–140.

[Publisher Full Text](#)

Fournier T, Gounot JS, Freil K, *et al.*: **High-quality *de novo* genome assembly of the *Dekkera bruxellensis* yeast using nanopore MinION sequencing.** *G3 (Bethesda).* 2017; **7**(10): 3243–3250.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Francischini FJ, Cordeiro EM, de Campos JB, *et al.*: ***Diatraea saccharalis* history of colonization in the Americas. The case for human-mediated dispersal.** *PLoS One.* 2019; **14**(7): e0220031.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Francischini FJ, De Campos JB, Alves-Pereira A, *et al.*: **Morphological and molecular characterization of Brazilian populations of *Diatraea saccharalis* (Fabricius, 1794) (Lepidoptera: Crambidae) and the evolutionary relationship among species of *Diatraea* Guilding.** *PLoS One.* 2017; **12**(11): e0186266.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Gouin A, Breteau A, Nam K, *et al.*: **Two genomes of highly polyphagous lepidopteran pests (*Spodoptera frugiperda* Noctuidae) with different host-plant ranges.** *Sci Rep.* 2017; **7**(1): 11816.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Gschloessl B, Dorkeld F, Audiot P, *et al.*: ***De novo* genome and transcriptome resources of the Adzuki bean borer *Ostrinia scapulalis* (Lepidoptera: Crambidae).** *Data Brief.* 2018a; **17**: 781–787.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Gschloessl B, Dorkeld F, Berges H, *et al.*: **Draft genome and reference transcriptomic resources for the urticating pine defoliator *Thaumetopoea***

***pitoyocampa* (Lepidoptera: Notodontidae).** *Mol Ecol Resour.* 2018b; **18**(3): 602–619.

[PubMed Abstract](#) | [Publisher Full Text](#)

Götz S, García-Gómez JM, Terol J, *et al.*: **High-throughput functional annotation and data mining with the Blast2GO suite.** *Nucleic Acids Res.* 2008; **36**(10): 3420–3435.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Heliconius Genome Consortium: **Butterfly genome reveals promiscuous exchange of mimicry adaptations among species.** *Nature.* 2012; **487**(7405): 94.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Holt C, Yandell M: **MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects.** *BMC Bioinformatics.* 2011; **12**: 491.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Joyce AL, White WH, Nuessly GS, *et al.*: **Geographic population structure of the sugarcane borer, *Diatraea saccharalis* (F.) (Lepidoptera: Crambidae), in the southern United States.** *PLoS One.* 2014; **9**(10): e110036.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Joyce AL, Sermeno Chicas M, Serrano Cervantes L, *et al.*: **Host-plant associated genetic divergence of two *Diatraea* spp. (Lepidoptera: Crambidae) stemborers on novel crop plants.** *Ecol Evol.* 2016; **6**(23): 8632–8644.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Kakumani PK, Malhotra P, Mukherjee SK, *et al.*: **A draft genome assembly of the army worm, *Spodoptera frugiperda*.** *Genomics.* 2014; **104**(2): 134–143.

[PubMed Abstract](#) | [Publisher Full Text](#)

Kawamoto M, Jouraku A, Toyoda A, *et al.*: **High-quality genome assembly of the silkworm, *Bombyx mori*.** *Insect Biochem Mol Biol.* 2019; **107**: 53–62.

[PubMed Abstract](#) | [Publisher Full Text](#)

Kirk H, Dorn S, Mazzi D: **Molecular genetics and genomics generate new insights into invertebrate pest invasions.** *Evol Appl.* 2013; **6**(5): 842–856.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Koseva BS, Hackett JL, Zhou Y, *et al.*: **Quantitative genetic mapping and genome assembly in the lesser wax moth *Achroia grisella*.** *G3 (Bethesda).* 2019; **9**(7): 2349–2361.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Li F, Zhao X, Li M, *et al.*: **Insect genomes: progress and challenges.** *Insect Mol Biol.* 2019; **28**(6): 739–758.

[PubMed Abstract](#) | [Publisher Full Text](#)

Li W, Zhang X, Fan Z, *et al.*: **Structural characteristics and phylogenetic analysis of the mitochondrial genome of the sugarcane borer, *Diatraea saccharalis* (Lepidoptera: Crambidae).** *DNA Cell Biol.* 2011; **30**(1): 3–8.

[PubMed Abstract](#) | [Publisher Full Text](#)

Li X, Waterman MS: **Estimating the repeat structure and length of DNA sequences using k -tuples.** *Genome Res.* 2003; **13**(8): 1916–1922.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Liyanage DS, Oh M, Omeka WKM, *et al.*: **First draft genome assembly of redlip mullet (*Liza haematocheila*) from Family Mugilidae.** *Front Genet.* 2019; **10**: 1246.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

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

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Merlin BL, Cónsoli FL: **Regulation of the larval transcriptome of *Diatraea saccharalis* (Lepidoptera: Crambidae) by maternal and other factors of the parasitoid *Cotesia flavipes* (Hymenoptera: Braconidae).** *Front Physiol.* 2019; **10**: 1106.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Miller SA, Dykes DD, Polesky HF: **A simple salting out procedure for extracting DNA from human nucleated cells.** *Nucleic Acids Res.* 1988; **16**(3): 1215.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Myers JG: **The original habitat and hosts of three major sugar-cane pests of tropical America (*Diatraea*, *Castnia* and *Tomaspis*).** *Bulletin of Entomological Research.* 1932; **23**(2): 257–271.

[Publisher Full Text](#)

National Center for Biotechnology Information (NCBI). Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 1988 –[cited 2017 Apr 06].

[Reference Source](#)

Pashley DP, Hardy TN, Hammond AM, *et al.*: **Genetic evidence for sibling species within the sugarcane borer (Lepidoptera: Pyralidae).** *Ann Entomol Soc Am.* 1990; **83**(6): 1048–1053.

[Publisher Full Text](#)

Pavinato VAC, Margarido GRA, Wijeratne AJ, *et al.*: **Restriction site associated DNA (RAD) for *de novo* sequencing and marker discovery in sugarcane borer, *Diatraea saccharalis* Fab (Lepidoptera: Crambidae).** *Mol Ecol Resour.* 2017; **17**(3): 454–465.

[PubMed Abstract](#) | [Publisher Full Text](#)

Pavinato VAC, Silva-Brandão KL, Montei M, *et al.*: **Development and characterization of microsatellite loci for genetic studies of the sugarcane borer, *Diatraea saccharalis* (Lepidoptera: Crambidae).** *Genet Mol Res.* 2013; **12**(2): 1631–5.

[PubMed Abstract](#) | [Publisher Full Text](#)

Pavinato VA, Michel AP, De Campos JB, *et al.*: **Influence of historical land use and modern agricultural expansion on the spatial and ecological divergence of sugarcane borer, *Diatraea saccharalis* (Lepidoptera: Crambidae) in Brazil.** *Heredity (Edinb)*. 2018; **120**(1): 25–37.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Peona V, Weissensteiner MH, Suh A: **How complete are “complete” genome assemblies? —An avian perspective.** *Mol Ecol Res*. 2018; **18**(6): 1188–1195.

[PubMed Abstract](#) | [Publisher Full Text](#)

Pippel M, Jebb D, Patzold F, *et al.*: **A highly contiguous genome assembly of the bat hawkmoth *Hyles vespertilio* (Lepidoptera: Sphingidae).** *Gigascience*. 2020; **9**(1): gaaa001.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Rodríguez-del-Bosque LA, Smith JW, Browning HW: **Bibliography of the neotropical cornstalk borer, *Diatraea lineolata* (Lepidoptera: Pyralidae).** *Fla Entomol*. 1988; **71**: 176–186.

[Publisher Full Text](#)

Rodríguez-del-Bosque LA, Smith JW, Browning HW: **Feeding and pupation sites of *Diatraea lineolata*, *D. saccharalis*, and *Eoreuma loftini* (Lepidoptera: Pyralidae) in relation to corn phenology.** *J Econ Entomol*. 1990; **83**(3): 850–855.

[Publisher Full Text](#)

Roe RM: **A bibliography of the sugarcane borer, *Diatraea saccharalis* (Fabricius), 1887–1980.** Agricultural Research Service (Southern Region), US Department of Agriculture. 1981; **20**.

[Reference Source](#)

Santos LB: **bslucas98/Dsaccharalis_genomeassembly: A first draft genome of the sugarcane borer, *Diatraea saccharalis* (Version v1.0.0) [Data set].** *Zenodo*. 2020.

Silva-Brandão KL, Santos TV, Cônsoli FL, *et al.*: **Genetic diversity and structure of Brazilian populations of *Diatraea saccharalis* (Lepidoptera: Crambidae): Implications for pest management.** *J Econ Entomol*. 2015; **108**(1): 307–316.

[PubMed Abstract](#) | [Publisher Full Text](#)

Simão FA, Waterhouse RM, Ioannidis P, *et al.*: **BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.**

Bioinformatics. 2015; **31**(19): 3210–3212.

[PubMed Abstract](#) | [Publisher Full Text](#)

Smit AF, Hubley R: **RepeatModeler Open-1.0.** 2010.

[Reference Source](#)

Smit AF, Hubley R, Green P: **RepeatMasker Open-4.0.** 2013–2015.

Treangen TJ, Salzberg SL: **Repetitive DNA and next-generation sequencing: computational challenges and solutions.** *Nat Rev Genet*. 2012; **13**(1): 36–46.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Triant DA, Cinel SD, Kawahara AY: **Lepidoptera genomes: current knowledge, gaps and future directions.** *Curr Opin Insect Sci*. 2018; **25**: 99–105.

[PubMed Abstract](#) | [Publisher Full Text](#)

Vurtture GW, Sedlazeck FJ, Nattestad M, *et al.*: **GenomeScope: fast reference-free genome profiling from short reads.** *Bioinformatics*. 2017; **33**(14): 2202–2204.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Wachi N, Matsubayashi, KW, Maeto K: **Application of next-generation sequencing to the study of non-model insects.** *Entomol Sci*. 2018; **21**(1): 3–11.

[Publisher Full Text](#)

Wan F, Yin C, Tang R, *et al.*: **A chromosome-level genome assembly of *Cydia pomonella* provides insights into chemical ecology and insecticide resistance.** *Nat Commun*. 2019; **10**(1): 1–14.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Wilson BE, VanWeelden MT, Beuzelin JM, *et al.*: **Efficacy of insect growth regulators and diamide insecticides for control of stem borers (Lepidoptera: Crambidae) in sugarcane.** *J Econ Entomol*. 2017; **110**(2): 453–463.

[PubMed Abstract](#) | [Publisher Full Text](#)

You M, Yue Z, He W, *et al.*: **A heterozygous moth genome provides insights into herbivory and detoxification.** *Nat Genet*. 2013; **45**(2): 220–225.

[PubMed Abstract](#) | [Publisher Full Text](#)

Open Peer Review

Current Peer Review Status: ? ?

Version 1

Reviewer Report 25 May 2021

<https://doi.org/10.5256/f1000research.29379.r85607>

© 2021 Van Belleghem S. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Steven M Van Belleghem

Department of Biology, University of Puerto Rico, Rio Piedras, Puerto Rico

Lucas Borgos dos Santos and colleagues describe the first draft genome assembly of the sugarcane borer *D. sachharalis*, a common pest species throughout the Americas. They used high-coverage short-read sequencing (Illumina 2x250 bp) to assemble a ~453 Mb genome fragmented over 50,460 scaffolds with an N50 of 16.3 kb and 87% completeness based on BUSCO analysis.

I believe the manuscript is well-written and provides a helpful guide into understanding this genomics resource, its quality and its usability.

These are some remarks I hope can improve the manuscript:

Abstract:

- Results: Please provide estimated completeness (in %) based on BUSCO analysis. N50 to get a sense of the fragmentation would also be interesting here.
- Conclusion: "Functional annotation reveals that predicted gene models are involved in important cellular mechanisms such as metabolic pathways and protein synthesis." I personally see very little point in reporting functional annotation results if it is not for making comparisons. Metabolic pathways and protein synthesis are common to all organisms.

Introduction:

- Suggested edit: "... that is the appearance OF A series of holes across the leaves that are still rolled up..."

Results:

- "We obtained over 306 Megabase pairs (Mb)" I assume you mean 306 million paired reads, for a total of $2 \times 250 \times 306,000,000 = 153 \text{ Gb}$?
- "...this analysis revealed 0.69% of variation across the genomic sequences...". Please explain

by giving comparative examples to demonstrate that this is on the low side.

- Please comment on why the cumulative length of the assembly (453 Mb) is longer than the genome size estimation (359 Mb). Could it also include haplotypes that are split up as two scaffolds due to regions of high heterozygosity? Using 'haplomerger' could improve this.
- "The alignment of raw input short-read data resulted in a total of 96.40% of the reads mapped to the assembly, of which 89.64% were properly paired to the scaffolds." Please comment on what this means about the assembly. Does this mean scaffolds are 89.64% accurate?
- BUSCO genes: How many were completely identified and how many were identified as fragments?
- Table 1: Please use commas to separate 1,000 everywhere.
- "...of which contain a certain level of similarity to a given B. mori protein sequence..." Please specify "certain level".
- 1,394 predicted gene models seem rather low. Is this due to the fragmentation of the genome? The stringency levels?

Discussion:

- "The consistency of our pipeline was confirmed by analyzing the repetitiveness present in B. mori chromosomes" It was unclear to me that you performed this analysis yourself.
- "Further gene ontology annotation revealed that proteins with high identity levels between the two species are related to important cellular components such as the nucleus, plasma membranes and cytoplasm, as well as being involved in molecular mechanisms essential for cell maintenance such as the transcription process, protein synthesis and metabolic pathways." I am not convinced this analysis provides a biologically meaningful result, given the low number of genes identified and there not being an alternative dataset to compare the GOs to. I think this type of analysis makes more sense in RNA-seq and differential gene expression studies.
- I find the discussion slightly repetitive with the results and would suggest merging the results and discussion and having a short conclusion paragraph.

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

Yes

Is the study design appropriate and is the work technically sound?

Yes

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

Yes

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

Yes

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

Yes

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Population genomics

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

Reviewer Report 28 January 2021

<https://doi.org/10.5256/f1000research.29379.r78017>

© 2021 Ghurye J. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Jay Ghurye

Verily Life Sciences, Cambridge, MA, USA

I enjoyed the manuscript describing the assembly of the first ever *D. sachharalis* genome. Overall, the text is easy to follow.

Here are the suggestions:

1. Please specify exact parameters and commands for all the tools used in the manuscript. That would help with the reproducibility of the results.
2. The description of de novo genome assembly is quite convoluted. Just saying the genome was assembled and scaffolded using SPAdes assembler would be sufficient.
3. Genome assembly evaluation can be made more thorough. Please consider reporting metrics reported by paired end reads based evaluation tools such as ALE (<https://academic.oup.com/bioinformatics/article/29/4/435/1992221>¹).
4. Please report the base level accuracy of the genome assembly (Assembly QV). That would help evaluate the how good assembly is at a single base pair resolution.
5. Can you use closely related species of *D. sachharalis* to scaffold your draft assembly to

chromosome level? There are tools to do that if you can use such a genome.

6. In table 1, the total scaffold length is much smaller than the assembly size. Why is that the case?
7. In table 1, please consider formatting the number with commas.
8. In both the tables, please make numbers right formatted.

References

1. Clark SC, Egan R, Frazier PI, Wang Z: ALE: a generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies. *Bioinformatics*. 2013; **29** (4): 435-43 [PubMed Abstract](#) | [Publisher Full Text](#)

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

Yes

Is the study design appropriate and is the work technically sound?

Yes

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

Partly

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

Not applicable

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

Yes

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Genome assembly, scaffolding, metagenomics

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

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research