Data in Brief 40 (2022) 107708

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Exploration of RNA-Seq data to identify a potential pathogen of the leaf-mining moth, *Stomphastis thraustica* (Meyrick, 1908) (Lepidoptera: Gracillariidae)

Kayvan Etebari[a,*], Dianne B.J. Taylor[b], Md. Mahbubur Rahman[b], Kunjithapatham Dhileepan[b], Michael J. Furlong[a], Sassan Asgari[a]

[a] School of Biological Sciences, The University of Queensland, St. Lucia, QLD 4072, Australia
[b] Biosecurity Queensland, Department of Agriculture and Fisheries, Ecosciences Precinct, Boggo Road, Dutton Park, QLD 4102, Australia

ARTICLE INFO

Article history:
Received 13 October 2021
Revised 3 December 2021
Accepted 9 December 2021
Available online 11 December 2021

Keywords:
Insect transcriptome
Insect pathology
Saccharomyces cerevisiae
Yeast
Bellyache bush
Weed biological control
Insect viruses

ABSTRACT

The leaf-mining moth, *Stomphastis thraustica* (Meyrick, 1908) was imported to Australia as a potential biological control agent of an exotic weed, bellyache bush (*Jatropha gossypiifolia*), from Peru. The insect colony has been maintained in the quarantine facility for over eight years but recently, significant mortality was observed in the culture. The larvae demonstrated swollen intersegments with a fragile integument. The infected larvae are cloudy muted green or yellowish whereas a healthy late instar larva is a vivid green. They slowly dehydrate and eventually die, at which point the larval body becomes rubbery and turns to black. We used next generation sequencing to identify the cause of mortality in the insects. Total RNA was extracted from 20 larvae in two cohorts, one with and one without apparent symptoms of disease, for deep sequencing on NovaSeq platform after eukaryote ribosomal RNA depletion. We identified several non-insect sequences belonging to viruses, bacteria, and fungi, but none of those showed significant abundance or enrichment in the infected dataset. The sequences related to

* Corresponding author at: School of Biological Sciences, Goddard Building, The University of Queensland, St. Lucia, QLD 4072, Australia.
E-mail address: k.etebari@uq.edu.au (K. Etebari).
Social media: (K. Etebari), (S. Asgari)

https://doi.org/10.1016/j.dib.2021.107708

a unicellular yeast, *Saccharomyces cerevisiae*, and they were among the highly expressed non-insect contigs; more than 5% of reads in both libraries mapped to the genome of this opportunistic microorganism.

Specifications Table

Subject	Agricultural and Biological Sciences
Specific Subject Area	Insect transcriptomics and Pathogen discovery
Type of data	RNA-Seq Data, Tables and Figure
How data were acquired	High-throughput strand specific RNA sequencing after rRNA depletion on NovaSeq PE150 platform by Genewiz sequencing facility in China
Data format	Raw: FASTQ files Analysed: Assembled contigs in FASTA format
Description of data collection	Total RNA was extracted from 20 *Stomphastis thraustica* larvae in two cohorts, with and without apparent symptoms of unknown disease in the laboratory population. CLC Genomic Workbench v21.5 was used for RNA-Seq analysis. We also used de novo assemble metagenome tool in CLC Microbial Genomics module for metatranscriptomic analysis of this data.
Data source location	A small leaf-mining moth, *S. thraustica* (Lepidoptera: Gracillariidae) was imported into Australia from Peru in 2014. The insect colony has been maintained at the Department of Agriculture and Fisheries quarantine facility in Brisbane, Queensland, Australia.
Data accessibility	Repository name: Deep sequencing data have been deposited in the National Centre for Biotechnology Information's (NCBI's) Gene Expression Omnibus (GEO) and are accessible through GEO series accession numbers GSE185938. Direct link to the dataset: https://www.ebi.ac.uk/ena/browser/view/PRJNA771399

Value of the Data

- The RNA-Seq data is the first transcriptome of *Stomphastis thraustica* larvae which facilitate future genomic study of this species.
- Our data is useful for biologists and insect pathologists who investigate the pathogenic role of *Saccharomyces cerevisiae* in insect populations.
- The transcriptomic analysis of this data provides a list of the microbial community of *S. thraustica* larvae.
- This data provides essential information and knowledge for future work on this under described insect species.

1. Data Description

Data described in this article originated from cDNA sequencing of two cohorts of *Stomphastis thraustica* late instar larvae, one with and one without apparent symptoms of disease. The first obvious sign of affected larvae is a decline in the feeding activity of late instar larvae in the leaf mine; they eventually die in the leaf. Dead larvae are more likely to be found in leaves located lower on the plant than leaves located near the top of the plant. If they exit the leaf,

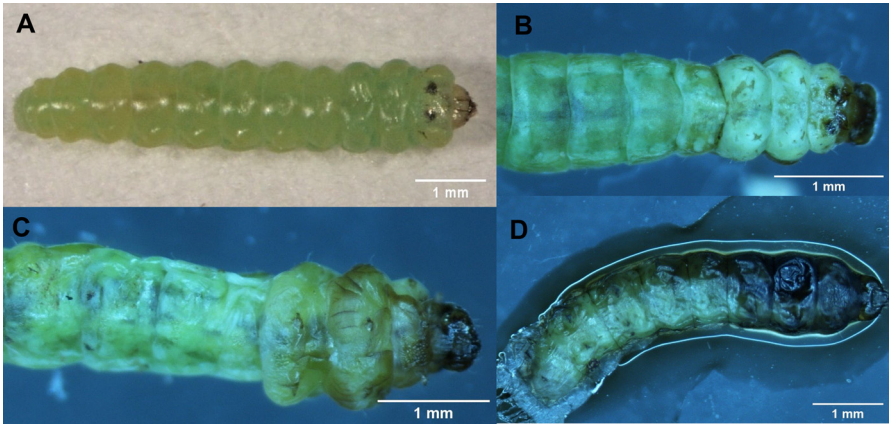

Fig. 1. The progression of disease in *Stomphastis thraustica* larvae. A) healthy prepupa with vivid green colouration, B) cloudy, muted green colour with a fragile integument, C) dehydrated larva that is less active, and D) rubbery larva, body turns black after death.

Table 1
Summary statistics of the *de novo* assembly of *S. thraustica* larvae.

Measurement	*de novo* assembly in transcriptome mode (Length or count)	*de novo* assembly in metagenome mode (Length or count)
Number of contigs	26,816	17,315
Number of contigs > 1kb	4361	1783
Total length of contigs	19,257,964	10,335,551
Total length of contigs > 1kb	7,329,740	2,719,167
Minimum contig length	300	300
Maximum contig length	25,029	10,377
Mean contig length	718	597
Median contig length	523	462
N25	1427	1036
N50	773	629
N75	495	425
N90	371	342

affected larvae are lethargic and some fall from the plant to the cage floor. Only a few of these affected individuals can successfully pupate. Affected larvae/prepupae are more delicate and they become injured easily. While healthy late instar larvae are a vivid green colour, affected larvae are cloudy muted green or yellowish and have swollen intersegments with a fragile integument. They slowly dehydrate and eventually die, when the larval body becomes rubbery and turns to black (Fig. 1).

In total, 207,772,228 paired-end reads were generated from two RNA-Seq libraries. We *de novo* assembled 26,816 contigs using CLC genomic workbench v21.0.5 from 191,110,344 clean and trimmed reads (Table 1). We also used unmapped reads to the proxy genomes as input for further *de novo* assembly of metagenome in CLC Microbial Genomics module from which 17,315 contigs were produced. More than 97% of trimmed reads mapped to the proxy genome references (Table 2). The assembled contigs are available in FASTA format (these files are accessible through Gene Expression Omnibus (GEO) series accession numbers GSE185938 at NCBI website). The outcome of BLASTx search for 8926 contigs from transcriptomic *de novo* assembly is available in Supplementary Table S1.

We identified several partial sequences of insect-specific viruses from family *Rhabdoviridae, Metaviridae*, and *Chuviridae* but this data is not conclusive to consider those viruses as the cause

Table 2
Summary statistics of the mapping to the proxy genome reference.

	Amyelois transitella		*Conopomorpha cramerella*	
	Read Count	Percentage	Read Count	Percentage
Mapped reads	185,533,100	97.08	187,504,643	98.11
Not mapped reads	5,577,244	2.92	3,605,701	1.88
Reads in pairs	183,285,598	95.91	160,371,752	83.91
Broken paired reads	2,247,502	1.18	27,132,891	14.19

Table 3
The list of identified non-insects' sequences from *S. thraustica* transcriptome data.

Microorganism name*	# Sequence	Group
Saccharomyces cerevisiae	62	Fungi
Ogataea polymorpha	1	Fungi
Talaromyces stipitatus	1	Fungi
Mixia osmundae	1	Fungi
Hanseniaspora opuntiae	1	Fungi
Kluyveromyces marxianus	1	Fungi
Histoplasma capsulatum	1	Fungi
Hubei lepidoptera virus 4	2	Viruses
Lambdina fiscellaria nucleopolyhedrovirus	2	Viruses
Trichoplusia ni TED virus	2	Viruses
Xenotropic MuLV-related virus	1	Viruses
Scaldis River bee virus	1	Viruses
Xenotropic murine leukemia virus	1	Viruses
Orgi virus	3	Viruses
Gata virus	1	Viruses
Hubei odonate virus 11	1	Viruses
Spodoptera frugiperda rhabdovirus	1	Viruses
Candidatus Woesearchaeota	1	Archaea
Escherichia coli	3	Bacteria
Acinetobacter baumannii	3	Bacteria
Salmonella enterica subsp. enterica serovar Typhi	1	Bacteria
Streptococcus pneumoniae	1	Bacteria
Actinomyces odontolyticus	1	Bacteria
Propionibacterium acnes	1	Bacteria
Vibrio anguillarum	1	Bacteria
Streptococcus mutans	1	Bacteria
Salmonella enterica	1	Bacteria
Streptococcus mutans	1	Bacteria
Anaerotignum lactatifermentans	1	Bacteria
Pseudomonas sp.	1	Bacteria
Bacteroidetes bacterium	1	Bacteria
Piscirickettsia salmonis	1	Bacteria
Pseudomonas syringae	1	Bacteria
Haemophilus influenzae	1	Bacteria
Pseudomonas amygdali pv. mori	1	Bacteria
Stylonychia lemnae	1	Eukaryota
Trichomonas vaginalis G3	1	Eukaryota
Trypanosoma brucei brucei	1	Eukaryota
Brugia malayi	2	Nematoda

* More information about these sequences can be found in supplementary Table S1.

of disease in the larvae (Tables 3 and S1). Some of those sequences might represent Endogenous Viral Elements (EVEs) which could be part of the *S. thraustica* genome [1]. These *S. thraustica* larval RNA-Seq data provided no evidence for the presence of detectable Microsporidia-like *Nosema sp.*, which is a well-known cause of mortality in other lepidopteran larvae [2,3]. The sequences related to a unicellular yeast, *Saccharomyces cerevisiae*, are among the highly expressed non-insect contigs and more than 5% of reads in both libraries (symptomatic and asymptomatic)

mapped to the genome of this opportunistic microorganism. Although previous studies have shown that *S. cerevisiae* can be pathogenic to insects [4], further investigation is required to confirm this potential cause of disease in *S. thraustica* larvae. We assume that *S. cerevisiae* causes mortality when the rearing or food conditions is not optimum for their insect host. The symptom of this infection is similar to a previously described case in *Galleria mellonella,* when the larvae injected with a lethal dose of *S. cerevisiae* turned black (consistent with melanization) within 30 min post injection [4].

2. Experimental Design, Materials and Methods

2.1. Insect collection and sample preparation

The leaf-mining moth (*S. thraustica*) was imported from Peru to Australia in 2014 as a potential biological control agent of the weed, bellyache bush (*Jatropha gossypiifolia*). The insect colony has been maintained in the Department of Agriculture and Fisheries quarantine facility in Brisbane since that time. Bellyache bush is a serious weed of rangelands and riparian zones of northern Australia, and it has the potential to expand its range significantly in the region [5]. *S. thraustica* larvae enter the leaf and remain within the leaf until pupation. Prepupae are highly mobile, though most pupate on the leaf that they emerge from [6]. Recently, significant mortality was observed in the laboratory colony.

We presume the microorganism associated with larval disease has been well established in the *S. thraustica* laboratory colony and most of the individuals are already infected with this potential pathogen and selecting a healthy or non-infected larvae was not easily possible. We used a next generation sequencing approach to identify the cause of mortality in the insects. Twenty larvae with and 20 larvae without obvious signs of disease (symptomatic and asymptomatic individuals) were preserved in an RNA stabilization reagent (RNAprotect®, QIAGEN Cat No.:76104) for further RNA extraction and sequencing.

The whole larvae were transferred to Qiazol lysis reagent for RNA extraction according to the manufacturer's instructions (QIAGEN; Cat No.: 79306). The RNA samples were treated with DNase I for 1 h at 37°C and then their concentrations were measured using a spectrophotometer and integrity was ensured through analysis of RNA on a 1% (w/v) agarose gel. After checking the RNA quality, total RNA from two samples (symptomatic and asymptomatic individuals) were submitted to the Genewiz sequencing facility (Jiangsu, China) for library preparation (after eukaryote ribosomal RNA depletion) and strand specific total RNA sequencing on NovaSeq platform.

2.2. RNA-Seq data analysis

The CLC Genomics Workbench version 21.0.5 was used for bioinformatics analyses. Both libraries were trimmed from any vector or adapter sequences remaining. Low quality reads (quality score below 0.05) and reads with more than two ambiguous nucleotides were discarded. In the absence of a reference genome, we used a *de novo* assembly approach (word size 25, bubble size 50 and minimum contig length 300 bp) to process these data. The contigs were corrected by mapping all reads against the assembled sequences (min. length fraction, maximum mismatch, insertion, and deletion cost of 0.8, 2, 3 and 3 respectively). The Read Per Kilobase of transcript per Million mapped reads (RPKMs) value was calculated for each of the assembled contigs. To search for a potential pathogen, we retained all contigs above 600 bp and all contigs with RPKM above 10 (1446 contigs), regardless of their size, for downstream analysis.

Due to lack of biological replicates, the nature of preparation of this RNA-Seq library, and the possibility of asymptomatic infection of the control group, these datasets are not suitable for assessment of differentially expressed genes.

We also mapped the trimmed reads to the genomes of *Conopomorpha cramerella* (GCA_012932125.1) and *Amyelois transitella* (GCA_001186105.1) as proxy genome references to

discard insect-specific reads. The unmapped reads were retained for metatranscriptome *de novo* assembly using CLC Microbial Genomics module. BLASTx was used to identify sequence similarity of all assembled contigs with protein database (nr). We consider a sequence as a potential candidate for further analysis, if it meets more than one of these criteria: (1) The sequence belongs to one of the well-known insect pathogens, (2) The sequence is among highly expressed contigs in the dataset, (3) The complete genome sequence has been identified, or (4) More than 5% of reads in the library mapped to that microorganism genome.

CRediT Author Statement

Kayvan Etebari: Conceptualization, Investigation, Methodology, Visualization, Data curation, Writing – original draft preparation; **Dianne B. J. Taylor**: Investigation, Resources, Writing – reviewing & editing; **Md Mahbubur Rahman**: Investigation, Resources, Writing – reviewing & editing; **Kunjithapatham Dhileepan**: Investigation, Resources, Writing – reviewing & editing; **Michael Furlong**: Resources, Writing – reviewing & editing; **Sassan Asgari**: Conceptualization, Methodology, Supervision, Writing – reviewing & editing.

Ethics Statement

This article is an original work of the authors. All procedures performed in studies involving animals were in accordance with the ethical standards of the institution or practice at which the studies were conducted and no human participants were involved in this article. Compliance with Ethical Standards.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships which have or could be perceived to have influenced the work reported in this article.

Acknowledgments

This work was supported by Meat and Livestock Australia and the Queensland Department of Agriculture and Fisheries.

Supplementary Materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.dib.2021.107708.

References

[1] F.Z. Dezordi, C.R.d.S. Vasconcelos, A.M. Rezende, G.L. Wallau, In and outs of Chuviridae endogenous viral elements: origin of a potentially new retrovirus and signature of ancient and ongoing arms race in mosquito genomes, Front. Genet. 11 (2020) 542437, doi:10.3389/fgene.2020.542437.
[2] L.F. Solter, D.K. Pilarska, C.F. Vossbrinck, Host specificity of microsporidia pathogenic to forest lepidoptera, Biol. Control. 19 (1) (2000) 48–56 doi.org/, doi:10.1006/bcon.2000.0845.
[3] N. Kermani, Z.A. Abu-hassan, H. Dieng, N.F. Ismail, M. Attia, I. Abd Ghani, Pathogenicity of *Nosema sp.* (microsporidia) in the Diamondback moth, *Plutella xylostella* (Lepidoptera: Plutellidae), PLoS ONE 8 (5) (2013) e62884 doi.org/, doi:10.1371/journal.pone.0062884.

[4] S.S. Phadke, C.J. Maclean, S.Y. Zhao, E.A. Mueller, L.A. Michelotti, K.L. Norman, et al., Genome-wide screen for *Saccharomyces cerevisiae* genes contributing to opportunistic pathogenicity in an invertebrate model host, G3 8 (1) (2018) 63–78, doi:10.1534/g3.117.300245.
[5] K. Dhileepan, S. Neser, J. De Prins, Biological control of bellyache bush (Jatropha gossypiifolia) in Australia: South America as a possible source of natural enemies, in: F.A.C. Impson, C.A. Kleinjan, J.H. Hoffmann (Eds.), Proceedings of the XIV International Symposium on Biological Control of Weed, South Africa, Kruger National Park, 2014, pp. 5–10. 2-7 March.
[6] D.B.J. Taylor, E.L. Snow, K. Moore, K. Dhileepan, At last, biological control of Bellyache bush, 14th Queensland Weed Symposium, 2017 4-7 December.