Data Article

Transcriptome assembly dataset of anthelmintic response in *Fasciola hepatica*

Estefan Miranda-Miranda\textsuperscript{a,*}, Raquel Cossio-Bayugar\textsuperscript{a}, Hugo Aguilar-Díaz\textsuperscript{a}, Verónica Narváez-Padilla\textsuperscript{b}, Bernardo Sachman-Ruíz\textsuperscript{a}, Enrique Reynaud\textsuperscript{b}

\textsuperscript{a}Centro Nacional de Investigaciones Disciplinarias en Salud Animal e Inocuidad, Instituto Nacional de Investigaciones Forestales Agrícolas y Pecuarias (INIFAP), Cuernavaca, Morelos, Mexico
\textsuperscript{b}Departamento de Genética del Desarrollo y Fisiología Molecular, Instituto de Biotecnología, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, México

\textbf{A R T I C L E   I N F O}

\textit{Article history:}
Received 1 September 2020
Revised 13 January 2021
Accepted 26 January 2021
Available online 4 February 2021

\textbf{Keywords:}
Liver fluke
Anthelmintic resistance
Triclabendazole
Transcriptome
High-throughput sequencing

\textbf{A B S T R A C T}

*Fasciola hepatica* is a worldwide distributed zoonotic parasitic trematode, which causes a severe liver disease clinically known as fasciolasis in a large number of wild animals, several livestock species as well as humans, prevention and control of fasciolasis is made by massive use of anthelmintic compounds on livestock and inevitably this practice has led to the emergence of anthelmintic resistant *Fasciola hepatica* and there is a great scientific effort to elucidate the molecular basis of anthelmintic resistance of parasitic helminths in general and of Fasciola hepatica in particular that may lead to improved anthelmintic compounds. In our project, we sequenced the transcriptomes obtained from the anthelmintic response to Triclabendazole and Albendazole on four samples from sensitive and resistant strains of *Fasciola hepatica* on Illumina HiSeq 4000 Platform and generated about 10.03 Gb per sample. The average genome-mapping rate is 81.29% and the average gene-mapping rate is 62.81%. 30,105 genes were identified in which 28,669 of them are known genes and 1,237 of them are novel genes from novel coding transcripts without any known features, 20,743 novel RNA transcripts were identified of which 14,293 of them are

\* Corresponding author.

\textit{E-mail address:} miranda.estefhan@inifap.gob.mx (E. Miranda-Miranda).

Social media: \textsuperscript{a} (R. Cossio-Bayugar)

https://doi.org/10.1016/j.dib.2021.106808

2352-3409/© 2021 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)
previous unknown splicing event for known genes but no alternative splicing was detected, the remaining 5,213 transcripts were found to be long noncoding RNA.

© 2021 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

---

### Specifications Table

| Subject                  | Parasitology                  |
|--------------------------|-------------------------------|
| Specific subject area    | Parasite transcriptomics      |
| Type of data             | Tables, figures, text file, RNASeq data and raw sequencing data |
| How data was acquired    | Illumina HiSeq 4000 sequencing platform |
| Data format              | RAW, Filtered, Analyzed       |
| Parameters of data collection | RNA was isolated from a multisample set of two strains of anthelmintic resistant *Fasciola hepatica* under in vitro anthelmintic treatment, compared to susceptible untreated controls, cDNA libraries were obtained by reverse transcription and sequenced using the illumina HiSeq 4000 platform |
| Description of data collection | The collection contains a NCBI-Bioproject accession PRJNA679050, which contains the transcriptome information for NCBI-Biosamples: SAMN16822856 (Adult *Fasciola hepatica* susceptible to Triclabendazole), SAMN16822857 (Juvenile *Fasciola hepatica* susceptible to Triclabendazole), SAMN16822858 (Adult *Fasciola hepatica* resistant to Triclabendazole), SAMN16822859 (Adult *Fasciola hepatica* resistant to Albendazole), NCBI-Biosamples are complemented with the information related to parasite isolation, location, culture and propagation of the parasites and experimental conditions. RNA was collected from *Fasciola hepatica* and a cDNA library was built and high-throughput sequenced using the Illumina HiSeq 4000 platform, raw reads were recorded in a FASTQ files and uploaded to NCBI as Sequence Reads Archives (SRA): SRS77702012, SRS77702011, SRS77702012 and SRS77702013. Reads containing adapter or reads of low quality were filtered, and the remaining clean reads were mapped to the reference *Fasciola hepatica* genome GCA_000302435.1. Gene expression, estimation and annotation were then carried out and the RNA-seq dataset submitted to the NCBI were made to coincide to Biosamples as follows: SRX9523044: RNAseq data Transcriptome Adult susceptible, SRX9523045: RNAseq data Transcriptome Juvenile susceptible, SRX9523046: RNAseq data Transcriptome Adult TCBZ resistant, SRX9523047: RNAseq data Transcriptome Adult ABZ resistant. Supplemental information containing statistics and gene distribution of the Transcriptome is provided as links within the manuscript in the form of raw sequencing readings (novel_coding_transcript.gtf), and tables that includes differential expression for each biosample in FPKM, GO and KEGG numbers as well as GENBANK accession numbers for each sequence, details of raw read generated, assembly and annotation information, overall transcriptomic annotation information such as mapping rate, number of known and unknown transcript identified, splicing events and long noncoding RNA transcripts as well as the annotated gene ontology divided in number of genes found as cellular components or fulfilling a biological process or molecular function are depicted (supplementary data 2, 3). |

Data source location

Cuernavaca, Morelos. Mexico 18.49 53.90 N 99.13 41.80 W

Data source location

Repository name: Mendeley Direct URL to data to supplementary tables 1 to 4 and text file containing supplementary tables description: https://data.mendeley.com/datasets/d4ptznrrzt/1

NCBI-BioProject Data identification number: PRJNA679050:
https://www.ncbi.nlm.nih.gov/bioproject/PRJNA679050/

NCBI-BioSample SAMN16822856 (Adult *Fasciola hepatica* susceptible to Triclabendazole): https://www.ncbi.nlm.nih.gov/biosample/16822856

NCBI-BioSample SAMN16822857 (Juvenile *Fasciola hepatica* susceptible to Triclabendazole): https://www.ncbi.nlm.nih.gov/biosample/16822857

NCBI-BioSample SAMN16822858 (Adult *Fasciola hepatica* resistant to Triclabendazole).
Triclabendazole): [https://www.ncbi.nlm.nih.gov/biosample/16822858](https://www.ncbi.nlm.nih.gov/biosample/16822858). NCBI-BioSample SAMN16822859 (Adult Fasciola hepatica resistant to Albendazole): [https://www.ncbi.nlm.nih.gov/biosample/16822859](https://www.ncbi.nlm.nih.gov/biosample/16822859). NCBI-SRA RNAseq datasets:
SRX9523044: RNAseq data Transcriptome Adult susceptible: [https://www.ncbi.nlm.nih.gov/sra/SRX9523044](https://www.ncbi.nlm.nih.gov/sra/SRX9523044) SRX9523045: RNAseq data Transcriptome Juvenile susceptible: [https://www.ncbi.nlm.nih.gov/sra/SRX9523045](https://www.ncbi.nlm.nih.gov/sra/SRX9523045) SRX9523046: RNAseq data Transcriptome Adult TCBZ resistant: [https://www.ncbi.nlm.nih.gov/sra/SRX9523046](https://www.ncbi.nlm.nih.gov/sra/SRX9523046) RNAseq data Transcriptome Adult ABZ resistant: [https://www.ncbi.nlm.nih.gov/sra/SRX9523047](https://www.ncbi.nlm.nih.gov/sra/SRX9523047)

Value of the Data

- The RNA-seq dataset from anthelmintic sensitive and resistant strains of the zoonotic parasitic trematode *Fasciola hepatica* may be useful for scientific research on the field of parasitology around the world, looking for the molecular basis of anthelmintic resistance in all kind of parasitic helminths that represent a health treat to humans and animals.
- The dataset from anthelmintic sensitive and resistant strains of *Fasciola hepatica* may be useful for Molecular Biologist, Health Scientist, Parasitologists and Veterinarians that may find useful to have access to those genes that are differentially expressed in sensitive and resistant parasites as a complement of their own research on anthelmintic resistance.
- This RNA-seq dataset contains those genes involved in the response to anthelmintic compounds and in the generation of liver flukes resistant to these compounds and should be detectable by comparative analysis of the dataset.
- The molecular mechanism controlled by the genes involved in anthelmintic resistance, will be elucidated and also enable the characterization of anthelmintic-resistance genes and the corresponding proteins that are involved leading to possible solutions for anthelmintic remediation.
- The use of the dataset may produce better anthelmintic formulations capable to control anthelmintic-resistant *Fasciola hepatica*, originating a healthier livestock and a lower public health risk for the society.

1. Data Description

The data set contains transcriptome analysis of a multisample RNA-seq of *Fasciola hepatica* exhibiting different phenotypes of anthelmintic resistance; the data includes differential expression in FPKM, GO and KEGG numbers as well as GENBANK accession numbers for each sequence.

Table 1 details of raw read generated assembly and annotation information.

Table 2 describes the overall transcriptomic annotation information such as mapping rate, number of known and unknown transcript identified, splicing events and long noncoding RNA transcripts.

Fig. 1 shows the annotated gene ontology divided in number of genes found as cellular components or fulfilling a biological process or molecular function.

Fig. 2 is a relation of known and novel genes found in the sequenced *Fasciola hepatica* samples.

Fig. 3 shows the results of functional enrichment gene ontology sub classified 56 biological functions.

Bioproject and BioSample data has been uploaded to NCBI according to instructions [1], assigning accession Numbers PRJNA679050, SAMN16822856, SAMN16822857, SAMN16822858, SAMN16822859.
Table 1
Transcriptome alignment, assembly to reference genome and expression estimation.

| BioSample Accessions (Phenotype) | Total Row Reads (M) | Total Clean Reads (M) | Total Clean Bases (Gb) | Clean Reads Q20 (%) | Clean Reads Q30 (%) | Clean Reads ratio (%) |
|----------------------------------|---------------------|-----------------------|------------------------|---------------------|---------------------|----------------------|
| SAMN16822856 (TCBZ susceptible) | 104.08              | 100.07                | 10.1                   | 99.36               | 97.74               | 96.15                |
| SAMN16822857 (TCBZ susceptible) | 106.34              | 100.32                | 10.3                   | 98.67               | 95.22               | 94.34                |
| SAMN16822856 (TCBZ resistant)   | 106.34              | 100.64                | 10.6                   | 99.28               | 97.62               | 94.64                |
| SAMN16822856 (ABZ resistant)    | 105.14              | 100.72                | 10.8                   | 99.37               | 96.31               | 95.32                |

Bio samples entries at the NCBI
Total Raw Reads (Mb): The reads amount before filtering, Unit: Mb
Total Clean Reads (Mb): The reads amount after filtering, Unit: Mb
Total Clean Bases (Gb): The total base amount after filtering, Unit: Gb
Clean Reads Q20 (%): The Q20 value for the clean reads
Clean Reads Q30 (%): The Q30 value for the clean reads
Clean Reads Ratio (%): The ratio of the amount of clean reads
TCBZ: Triclabendazole
ABZ: Albendazole

Table 2
Overall transcriptomic annotation information. A number of novel genes and transcripts were identified under anthelmintic treatment of *F. hepatica*, as well as unknown splicing events that may be relevant for the study of anthelmintic resistance in parasitic trematodes.

|                      |                   |
|----------------------|-------------------|
| Genome mapping rate  | 81.29%            |
| Gene mapping rate    | 62.81%            |
| Genes identified     | 30105             |
| Known genes          | 28669             |
| Novel genes          | 1237              |
| Novel transcripts    | 14293             |
| Unknown splicing events | 14293          |
| Long noncoding RNA transcripts | 5213          |

Fig. 1. Annotated gene ontology (GO) found in the *Fasciola hepatica* transcriptome. GO terms assigned to each numbers were classified into three major categories and this graph represents how many genes were found for each category.
Fig. 2. Relation of known and novel genes found in the sequenced Biosamples. Different strains of *F. hepatica* strains exhibit a different number on novel genes expressed under anthelmintic treatment.

Supplemental information is provided at https://data.mendeley.com/datasets/j3v92bnkss/draft?a=03fb8a9e-376c-44dc-ac9c-73495ad8bade

2. Experimental Design, Materials and Methods

2.1. Sample preparation

Four samples of total RNA were obtained from as many *Fasciola hepatica* strains exhibiting different phenotypes of resistance and susceptibility response to anthelmintic compounds, anthelmintic resistant and susceptible parasites strains were maintained as anthelmintic bioassay reference at the National Center for Disciplinary Research in Veterinary Parasitology (INIFAP, Mexico), juvenile and adult Parasites were obtained from experimentally infested rabbits, and described as BioSamples as required by the NCBI for registration and accession number assignment [1].

2.2. RNA isolation

Ten living parasites from each BioSample were maintained in 20 ml Minimal Essential Medium Eagle (MEME Sigma chemicals) supplemented with 10% Fetal Bovine Serum (FBS GIBCO) The medium containing Biosamples SAMN15668962 (TCBZ resistant parasites) and SAMN15668963 (ABZ resistant parasites) were additionally supplemented with 10 ug/ml of Tri-clabendazol and Albendazole respectively. All biosamples were incubated at 37 °C overnight and preserved in RNAlater® until RNA extraction. One parasite from each sample was frozen at -196 °C and macerated to a fine powder; RNA was extracted using the phenol-chloroform procedure as described previously [2]. RNA quality, integrity, 28S/18S RNA ratio and fragment length
distribution, were assessed by capillary electrophoresis using an Agilent 2100 Bioanalyzer, estimating an average of 400 ng/ul for all samples and a RNA Integrity Value of 5. For additional estimation of the total RNA sample QC such as RNA concentration and RIN value, a Nanodrop® was used.

2.3. High-throughput sequencing

Transcriptome sequencing was carried out by BGI Genomics Co. Ltd. using Illumina HiSeq 4000 platform with a reading length of 2 × 150 bp. Adapter contamination and low-quality regions (Q < 20) towards 3’ end was trimmed out using Cutadapt program [3]. The final quality of processed reads was accessed using FastQC tool [4].

2.4. Transcriptome assembly

Reference genome GCA_900302435.1 was used for transcriptome assembly, removal of redundant transcripts and expression estimation was carried out via cufflinks [5] and displayed in Table 1. Assembly raw data is available at supplemental data section.
2.5. Transcriptome analysis

The BLAST similarity analysis was performed using the assembled transcriptome against Uniprot database [5]. Gene Ontology terms associated with the transcripts were extracted from the Uniprot database and integrated with the BLAST search results. We performed Gene Ontology (GO) Analysis classification and functional enrichment. GO was represented by three ontologies categories: molecular biological function, cellular component and biological process and the GO classification results are shown as Figs. 1, 2.

Ethics statement

Animal management was performed accordingly to the ethical guidelines of our institutions Animal care and use was according to the Mexican norm NOM-062-ZOO-1999, and its technical specifications for production, care and use of laboratory animals.

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 research received funding from the National Council for Science and Technology of Mexico (CONACyT) grants CIPN 2015-021, 255478 and by funds from DGAPA/UNAM PAPIIT-IN204214, PAPIIT-IN206517.

Supplementary Materials

Supplementary material associated with this article can be found in the online version at: https://data.mendeley.com/datasets/d4ptznrznt/1.

References

[1] T. Barrett, BioSample. 2013 Nov 14, The NCBI Handbook [Internet], 2nd edition, National Center for Biotechnology Information (US), Bethesda (MD), 2013. https://www.ncbi.nlm.nih.gov/books/NBK143764/.
[2] J.F. Sambrook, D.W. Russell, in: Molecular Cloning: A Laboratory Manual, 3rd ed., Cold Spring Harbor Laboratory Press, 2001, p. 2100.
[3] M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads, EMBnet J. 17 (2011) 10. doi:10.14806/ej.17.1.200.
[4] S. Andrews, F. Krueger, A. Seconds-Pichon, F. Biggins, S. Wingett, QC Fast, A quality control tool for high throughput sequence data, Babraham Bioinform. (2015) Babraham Bioinformatics, Babraham Inst.
[5] C. Trapnell, et al., Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks, Nat. Protoc. 7 (2012) 562–578, doi:10.1038/nprot.2012.016.