De novo transcriptome assembly and its annotation for the aposematic wood tiger moth (Parasemia plantaginis)

Galarza, Juan

2017-03-21


http://hdl.handle.net/10138/218279
https://doi.org/10.1016/j.gdata.2017.03.008

Downloaded from Helda, University of Helsinki institutional repository.

This is an electronic reprint of the original article.

This reprint may differ from the original in pagination and typographic detail.

Please cite the original version.
De novo transcriptome assembly and its annotation for the aposematic wood tiger moth (*Parasemia plantaginis*)

Galarza Juan A. a,⁎, Kishor Dhaygude b,1, Johanna Mappesa a

a Centre of Excellence in Biological Interactions, Dept. of Biological and Environmental Sciences, University of Jyväskylä, Finland

b Centre of Excellence in Biological Interactions, University of Helsinki, Finland

**Abstract**

In this paper we report the public availability of transcriptome resources for the aposematic wood tiger moth (*Parasemia plantaginis*). A comprehensive assembly methods, quality statistics, and annotation are provided. This reference transcriptome may serve as a useful resource for investigating functional gene activity in aposematic Lepidopteran species. All data is freely available at the European Nucleotide Archive (http://www.ebi.ac.uk/ena) under study accession number: PRJEB14172.

© 2017 The Authors. 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/).

1. Direct link to deposited data

http://www.ebi.ac.uk/ena. Study accession number: PRJEB1417.

2. Specifications

<table>
<thead>
<tr>
<th>Organism/cell line/tissue</th>
<th>Wood tiger moth (<em>Parasemia plantaginis</em>)/whole larva</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>Undetermined</td>
</tr>
<tr>
<td>Sequencer or array type</td>
<td>Illumina HiScanSQ</td>
</tr>
<tr>
<td>Data format</td>
<td>FASTQ</td>
</tr>
<tr>
<td>Experimental factors</td>
<td>De novo assembly, completeness assessment, and annotation</td>
</tr>
<tr>
<td>Experimental features</td>
<td>RNA-seq from whole larvae (<em>n</em> = 16) of Wood tiger moth (<em>Parasemia plantaginis</em>) from different developmental stages.</td>
</tr>
<tr>
<td>Consent</td>
<td>N/A</td>
</tr>
<tr>
<td>Sample source location</td>
<td>Jyväskylä, Central Finland</td>
</tr>
</tbody>
</table>

3. Introduction

Many plants and animals advertise unpalatability through conspicuous colouration as a form of warning signals to potential predators (i.e. aposematism). This defensive strategy is used by many Lepidopterans (butterflies and moths), and while its ecological and evolutionary consequences are relatively well-studied [2,14] scarce information is available about their molecular bases. The wood tiger moth (*Parasemia plantaginis*) is a diurnal aposematic species that shows considerable colour variation throughout its distribution range [6]. It has been recently classified as *Arctia plantaginis* [13]. Two male colour morphs and one female colour morph co-exist within local populations. It has been shown that the two male colour morphs differ in their warning signal efficacy, one being better protected than the other against avian predators [9]. Furthermore, in southern Finland, the genetic composition of the populations fluctuates between generations [4]. Thus, this species provides a valuable opportunity to investigate frequency-dependent selection processes in nature. However, the warning signals displayed by adults are pre-determined during the larval stage, when bodily resources are allocated to determine their shape and pattern. After metamorphosing into the adult stage, no further development or adaptations take place at the phenotypic level. Thus, functional gene activity during early life-stages must be investigated to gain insight about its possible effects on the adult phenotype.

Here we report a de novo transcriptome assembly of the wood tiger moth at its larval stage. Our aim was to obtain a high-coverage, high-quality reference transcriptome representative of different developmental stages. The data presented here are the first transcriptome resources available for the wood tiger moth.

4. Experimental design, materials and methods

Larvae originated from populations of Central Finland. A total of 16 larvae from instar 1 to instar 5 were selected for sequencing. All larvae were fed dandelion (*Taraxacum spp.*) and reared individually in petri dishes before immersion in RNAlater stabilising buffer. All samples were kept at −20°C until RNA extraction.
4.1. RNA extraction

Total RNA was extracted using RNeasy Mini Kit (Qiagen) according to manufacturer’s instructions with additional TriReagent (MRC, Inc.) and DNase (Qiagen, Valencia, US.A.), treatments. The quality and quantity of total RNA was inspected in a BioAnalyzer 2100 using RNA 6000 Nano Kit (Agilent). Subsequently, mRNA was isolated by means of two isolation cycles using Dynabeads mRNA purification kit (Ambion®) and quantified using RNA 6000 Pico Kit in a BioAnalyzer 2100 (Agilent). Pair-end (2 × 100 pb) cDNA libraries were constructed for each sample according to Illumina’s TruSeq Stranded HT protocol. The libraries were individually indexed and sequenced in an Illumina HiScanSQ sequencer at the DNA sequencing and genomics laboratory, Institute of Biotechnology of the University of Helsinki, Finland.

4.2. Transcriptome assembly

The quality of the raw reads was first inspected with FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Based on this initial quality check, we used the FASTX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) to remove low quality bases and sequencing artifacts. Bases with a Phred quality score of <25 were filtered out, and reads shorter than 85 bases after trimming were removed. Pair-end reads were then sorted and synchronized using custom scripts.

We used the high-quality reads from all samples obtained after FastQC and FASTX to perform an initial assembly (K25_Assembly) using Trinity (trinityrnaseq_r2013-02-25) software [5] with the following parameters: 4 CPUs for Inchworm and Butterfly, a maximum memory 200 GB, a minimum contig length of 200 bp, and K-mer = 25. The default K-mer of 25 recovered most full-length transcripts across a broad range of expression levels. To identify any unassembled reads, we mapped back the reads to the K25_Assembly using bowtie v. 0.12.7 [8]. The unassembled reads were then used to construct two further assemblies namely; K21_Assembly and K29_Assembly using two different K-mer settings of 21 and 29 respectively. A fourth assembly

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Properties and statistic of the Final_Assembly transcriptome for the wood tiger moth (Parasemia plantaginis).</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total unigenes</td>
<td>54,657</td>
</tr>
<tr>
<td>Unigenes after ribosomal filtering</td>
<td>54,346</td>
</tr>
<tr>
<td>N50(bp)</td>
<td>10,747</td>
</tr>
<tr>
<td>Mean coverage</td>
<td>39.12 ×</td>
</tr>
<tr>
<td>No. mapped reads</td>
<td>366,046,742(98.44%)</td>
</tr>
<tr>
<td>Annotated in nr</td>
<td>6309</td>
</tr>
<tr>
<td>Annotated in GO</td>
<td>16,936</td>
</tr>
<tr>
<td>Annotated in Inter-Pro</td>
<td>20,020</td>
</tr>
</tbody>
</table>

Fig. 1. (A) Ortholog hit ratios (OHR) of unigenes from Final_Assembly against silkworm (Bombyx mori) genome. (B) Coverage obtained from the different assemblies with varying K-mer length from 21 to 29.
yield the highest coverage (Fig. 1). The results showed that merging assemblies of different K-mers sequencing runs (Phred +33; ASCII range using custom scripts. Gene ontology terms (GO) and information of Release 123; All unigenes that showed signi
chae, bacteria, and eukaryote domains (SILVA_123_SSUParc_Taxa_Trunc & SILVA_123_LSUParc_Taxa_Trunc -
chaea, bacteria, and eukaryote domains (SILVA_123_SSUParc_Taxa_Trunc & SILVA_123_LSUParc_Taxa_Trunc -
5.1. Annotation results
A total of 17,800 unigenes returned a blast hit with e-value ≤ 10−5 and <70% amino acid identity. Of these unigenes, 16,036 had gene ontology (GO) annotation available with a mean GO level of 6.2 across biological processes (P), molecular (F) function and cellular components (C) categories. The main P,F,C after removing redundant GO terms are summarized in Supplementary file 1. The number of unigenes annotated to the different databases is shown in Table 1 and its functional annotation is provided in the Supplementary file 2.

Acknowledgements
This project was funded by the Centre of Excellence in Biological Interaction, via the Academy of Finland (Project No.252411). The authors are thankful to the Finish Centre of Science (CSC-IT) for access to computational resources, and to Petri Auvinen and Lars Paulin from the DNA sequencing and genomics laboratory, Institute of Biotechnology, of the University of Helsinki, Finland.

Appendix A. Supplementary data
Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.gdata.2017.03.008.

References