Genomics Data 12 (2017) 71–73

Contents lists available at ScienceDirect

Genomics Data

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

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

GalarzaJuan A. [a,*,1], Kishor Dhaygude [b,1], Johanna Mappes [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

ARTICLE INFO

Article history:
Received 6 March 2017
Accepted 19 March 2017
Available online 21 March 2017

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.

1. Direct link to deposited data

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

2. Specifications

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

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 −20C° until RNA extraction.

* Corresponding author.
E-mail address: juan.galarza@jyu.fi (J.A. Galarza).
[1] These authors contributed equally to the realization of this paper.

http://dx.doi.org/10.1016/j.gdata.2017.03.008

Table 1
Properties and statistic of the Final_Assembly transcriptome for the wood tiger moth (*Parasemia plantaginis*).

Total unigenes	54,657
Unigenes after ribosomal filtering	54,346
N50(bp)	10,747
Mean coverage	39.12×
No. mapped reads	366,046,742(98.44%)
Annotated in nr	17,800
Annotated in Swiss-Prot	6309
Annotated in GO	16,936
Annotated in Inter-Pro	20,020

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/pro-jects/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 filter 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

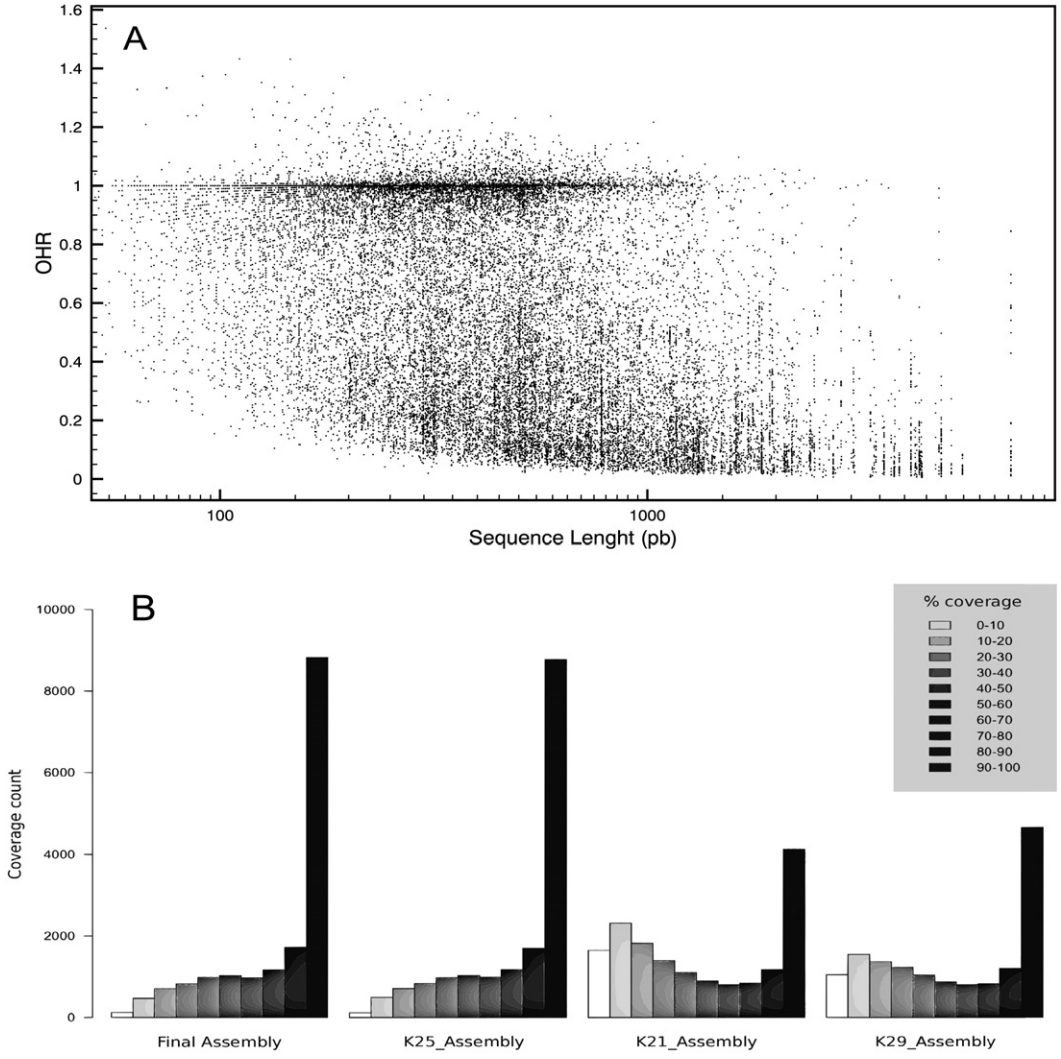

Fig. 1. (A) Ortholog hit ratios (OHR) of unigenes from Final_Assembly against silkworm (*Bobyx mori*) genome. (B) Coverage obtained from the different assemblies with varying K-mer length from 21 to 29.

(Final_Assembly) was constructed by combining the K25_Assembly with the other two K21 and K25 assemblies using CAP3 [7]. The Scaffolding software was run with default parameters making sure that the minimum overlap between two contigs was at least 100 bp with a 95% sequence similarity for building supercontigs.

We mapped back all reads to the Final_Assembly using bowtie as above to calculate the overall expression profile for all transcripts. We then removed any miss-assemblies or assembly errors by manual inspection of transcripts that showed RPKM (Reads Per Kilobase of transcript per million mapped reads) values of <1. Finally, we used Vmatch (http://www.vmatch.de) to filter out possible redundant unigenes present in this Final_Assembly. The complete assembly workflow is presented in Supplementary file 1.

4.3. Transcriptome validation

Transcriptome validation for non-model organisms is especially challenging because of the noisy nature of transcripts and lack of a reference genome to perform a guided assembly. Orthology check with the genome of the closest reference species available is one way to assess the quality of the de novo assembly. Hence, we computed the ortholog hit ratio [11] of the Final_Assembly using the silkworm moth (*Bombyx mori*; taxa ID: 17,701 Uniref90), which is the closest species with a thoroughly annotated genome. Ratios close to 1 are indicative that contigs of transcripts (i.e. unigene) matching the ortholog locus have been fully assembled.

4.4. Transcriptome annotation

For annotating the transcriptome, the Final_Assembly was first checked for possible ribosomal contamination that may have survived the RNA purification phase. The transcriptome was blasted [1] (BLASTn; e-value ≤ 10^{-5}) against publicly available ribosomal databases for archaea, bacteria, and eukaryote domains (SILVA_123_SSUParc_Taxa_Trunc & SILVA_123_LSUParc_Taxa_Trunc - Release 123; [12]. All unigenes that showed significant hits were removed (<3%). The remaining unigenes were compared against a non-redundant protein database (nr) (NCBI; last updated 30-05-2016) and the Swiss-Prot (last updated 26–05-2016) to retrieve basic annotation using BLASTx. After blasting, all hits that showed <70% amino acid identity, sequence length of <200 bp, and e-value ≤ 10^{-5} were filtered out using custom scripts. Gene ontology terms (GO) and information of the protein family was obtained using Blast2Go v.4.0 [3].

5. Results

A total of 372,037,058 pairs of reads were obtained from the sequencing runs (Phred +33; ASCII range "!" to "J"). After trimming and filtering, 371,847,565 pairs of reads (%GC 45) with a length of 85 bp were used for the transcriptome assemblies. The basic descriptors and quality metrics of the Final_Assembly are presented in Table 1 and Fig. 1. To evaluate the best transcript assembly method (TA), we compared the completeness (coverage) of TA's produced from each K-mer assembly to the closest annotated genome (*Bombyx mori*), as proposed by [10]. The results showed that merging assemblies of different K-mers yield the highest coverage (Fig. 1).

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

[1] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool. J. Mol. Biol. 215 (1990) 403–410.
[2] C.A. Barnett, M. Bateson, C. Rowe, State-dependent decision making: educated predators strategically trade off the costs and benefits of consuming aposematic prey. Behav. Ecol. 18 (2007) 645–651.
[3] A. Conesa, S. Gotz, J.M. Garcia-Gomez, et al., Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21 (2005) 3674–3676.
[4] J.A. Galarza, O. Nokelainen, R. Ashrafi, R.H. Hegna, J. Mappes, Temporal relationship between genetic and warning signal variation in the aposematic wood tiger moth (*Parasemia plantaginis*). Mol. Ecol. 23 (2014) 4939–4957.
[5] M.G. Grabherr, B.J. Haas, M. Yassour, et al., Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29 (2011) 644–652.
[6] R.H. Hegna, J.A. Galarza, J. Mappes, Global phylogeography and geographical variation in warning coloration of the wood tiger moth (*Parasemia plantaginis*). J. Biogeogr. 42 (2015) 1469–1481.
[7] X. Huang, A. Madan, CAP3: a DNA sequence assembly program. Genome Res. 9 (1999) 868–877.
[8] B. Langmead, C. Trapnell, M. Pop, S.L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10 (2009) R25.
[9] O. Nokelainen, R.H. Hegna, J.H. Reudler, C. Lindstedt, J. Mappes, Trade-off between warning signal efficacy and mating success in the wood tiger moth. Proc. R. Soc. B Biol. Sci. 279 (2012) 257–265.
[10] S. O'Neil, S. Emrich, Assessing de novo transcriptome assembly metrics for consistency and utility. BMC Genomics 14 (2013) 465.
[11] S.T. O'Neil, J.D.K. Dzurisin, R.D. Carmichael, et al., Population-level transcriptome sequencing of nonmodel organisms *Erynnis propertius* and *Papilio zelicaon*. BMC Genomics 11 (2010) 310.
[12] C. Quast, E. Pruesse, P. Yilmaz, et al., The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 41 (2013) D590–D596.
[13] K. Rönkä, J. Mappes, L. Kaila, N. Wahlberg, Putting Parasemia in its phylogenetic place: a molecular analysis of the subtribe Arctiina (Lepidoptera). Syst. Entomol. 41 (2016) 844–853.
[14] P.J. Weldon, G.M. Burghardt, Evolving détente: the origin of warning signals via concurrent reciprocal selection. Biol. J. Linn. Soc. 116 (2015) 239–246.