Transcriptome sequencing and analysis of the coconut leaf beetle, *Brontispa longissima*

W. Yan[1,2], L. Liu[1], C.X. Li[1], S.C. Huang[1], Z.L. Ma[3], W.Q. Qin[1] and Z.Q. Peng[4]

[1]Coconut Research Institute, Chinese Academy of Tropical Agricultural Sciences, Hainan, China
[2]Key Laboratory for Silviculture and Conservation of Ministry of Education, Beijing Forestry University, Beijing, China
[3]Institute of Tropical Bioscience and Biotechnology, Chinese Academy of Tropical Agricultural Sciences, Hainan, China
[4]Environment and Plant Protection Institute, Chinese Academy of Tropical Agricultural Sciences, Hainan, China

Corresponding author: W.Q. Qin
E-mail: qwq268@126.com

ABSTRACT. The coconut leaf beetle, *Brontispa longissima*, is a destructive pest of palm plants. Although its ecological and biological characteristics are well understood, its genetic information remains largely unknown. To advance our understanding of its molecular basis of biology and ecology, we sequenced and analyzed its whole transcriptome by using high-throughput Illumina paired-end sequencing technology. Approximately 8.08 Gb of clean reads were generated in a single run, which were assembled by using Trinity into 41,652 unigenes with an average length of 932 bp. By sequence similarity searches for known proteins, 23,077 (55.4%) unigenes were annotated by BLASTx searches against the NCBI non-redundant protein database. Of the unigenes assembled, 18,153 and 13,733 were assigned to Gene Ontology and Clusters of Orthologous Groups of proteins, respectively. In addition, 10,415 unigenes were mapped onto 247 pathways using the

Kyoto Encyclopedia of Genes and Genomes Pathway database. These transcriptomic resources will facilitate gene identification and elucidate the molecular mechanisms of biological and ecological aspects underlying this palm pest, in order to design a new control strategy.

Key words: *Brontispa longissima*; Transcriptome; Assembly

INTRODUCTION

The coconut leaf beetle, *Brontispa longissima* (Coleoptera: Chrysomelidae), has been one of the most damaging pests of coconut (*Cocos nucifera*) and other palms in Southeast Asia and the Pacific, particularly over the last three decades (Rethinam and Singh, 2007). Both larvae and adults feed on the soft tissues of young unfolded coconut leaflets that later become brown, dry, and die, resulting in the stunting of the palm, reduced nut production (as much as 50-70% in Samoa), and the death of the plant (Voegele, 1989; Acevedo et al., 2014). Due to the pest's high adaptability to environmental stressors and changing cultural practices, and its robustness and high reproductive rate, there is no effective strategy against it. To protect coconut palms from damage, the control of this pest relies mainly on chemical insecticides, and frequent chemical treatments to the invasive area are necessary to eradicate its populations (Lin et al., 2012). Large quantities of pesticides are not only expensive, but also cause environmental pollution and insecticide resistance. To develop alternative control methods it is critical to obtain a considerable scientific knowledge of this pest (Zhu et al., 2013a). Although its biological and ecological characteristics have been widely reported (Zhong et al., 2005; Takano et al., 2012, 2013), it has been little investigated at the molecular level. By August 20, 2014 only 264 *B. longissima* nucleotide sequences had been deposited in the National Center for Biotechnology Information (NCBI) database. These data are insufficient for elucidating the molecular mechanisms related to this pest. In order to improve our understanding of its genetic background, in this study the *B. longissima* transcriptome was deeply profiled using next generation, high-throughput sequencing, and functionally annotated using a systematic bioinformatic strategy.

MATERIAL AND METHODS

B. longissima colonies were collected from a coconut palm in Sanya, Hainan Province, China. After emergence, adults were maintained at $25° \pm 2°C$, 70-80% relative humidity, and a photoperiod of 12:12 h (L:D). They were reared on coconut palm leaves. Adult females and males of different ages were used for RNA extraction. Total RNA was extracted by using TRIzol reagent (Invitrogen), following the manufacturer protocol. Residual DNA in total RNA was removed by using DNase I (Promega). Total RNA quantity and purity were determined using a BioPhotometer® plus (Eppendorf), and RNA integrity was visualized by electrophoresis on ethidium bromide-stained 1% agarose gel. mRNA was isolated from 20 μg total RNA using magnetic beads with Oligo (dT), and digested into short fragments by using a fragmentation buffer. The short fragments were used as templates for first-strand cDNA synthesis with random-hexamer primers, followed by second-strand cDNA synthesis using a reaction system of buffer, dNTPs, RNase H, and DNA polymerase I. Double-stranded cDNA was then purified for end repair, dA tailing, adaptor ligation, and DNA fragment enrichment, in order to create

the final cDNA library. The library was sequenced using the Illumina® Hi-Seq2000 platform. The raw data were deposited in the DNA Data Bank of Japan (DDBJ) sequence-read archive (SRA) database. After sequencing, raw image data were transformed by base calling into raw reads, and were then filtered to remove low-quality reads. Clean reads were assembled into contigs and unigenes using the Trinity short-read assembling software (version release 05-10-2012) in order to obtain sequences that could not be extended at either end (Grabherr et al., 2011). The resulting unigenes were analyzed by searching the Nr, Nt, SwissProt, PFAM, GO (Gene Ontology), COG (Clusters of Orthologous Groups of proteins), and KEGG (Kyoto Encyclopedia of Genes and Genomes) databases using the BLAST algorithm with an E-value cutoff of 1 x 10^{-5}, and the transcripts obtained were functionally annotated as the retrieved protein or nucleic acid with the highest sequence similarity. The HMMER 3.0 software was used to search for eukaryotic protein domains in the transcriptome (Finn et al., 2010). GO terms were extracted from the best hits obtained by BLASTx against the specific databases, using the Blast2GO® software (Götz et al., 2008). COG and KEGG annotations were analyzed by using the Blastall program.

RESULTS AND DISCUSSION

The data generated by the Illumina® sequencing contained 82,865,068 raw reads (Table 1). After a cleaning step for removing adapters, low-quality bases, and contaminants, 80,849,500 clean reads, corresponding to 8.08 Gb, were produced. The GC percentage of the reads was 38.93%, which is comparable with the genome sequences of other insects (Price et al., 2011). Over 98% of the clean reads had quality scores higher than the Q20 level. These reads were assembled into 58,815 contigs and 41,652 unigenes. The number of unigenes was in a range similar to that reported for other beetles, such as *Holotrichia parallela*, *Meligethes aeneus*, and *Tenebrio molitor* (Zhu et al., 2013b; Ju et al., 2014; Zimmer et al., 2014). The average lengths of the contigs and unigenes were 1449 and 932 bp, respectively. Unigene lengths varied from 201 to 25893 bp, with an N50 of 1805 bp (**Figure S1**). Most unigenes were shorter than 500 bp (55.48% of total), and only 26.21% of them were longer than 1000 bp, which is a similar result to that obtained in previous transcriptome analyses in different insects using Illumina® sequencing (Zhu et al., 2012; Kumar et al., 2014). The large number of short-length unigenes generated might be due to the short length of the sequencing-read ability associated with the Illumina® platform, and coverage of individual reads (Carvalho et al., 2010; Zhu et al., 2013b).

In total, 25,162 of the unigenes (60.41%) were annotated in at least one database searched (Table 1). According to the Nr database, over half of the unigenes (23,077) had at least one putative homologue from an existing gene model. Due to a lack of genetic information, the remaining 44.6% of the unigenes did not match with any known sequences, and were probably novel. The unigenes without hits probably belonged to untranslated regions, noncoding RNA, short sequences not containing a protein domain, or assembly mistakes (Fu et al., 2013). Regarding species distribution, most of the BLASTx hits were matched to *Tribolium castaneum* sequences with a high identity (**Table S1**). This result is consistent with our knowledge of the evolutionary relationship between *B. longissima* and *T. castaneum* (that belong to the same order), and is due to the availability of the *T. castaneum* genome, which accounts for the vast majority of deposited coleopteran sequences (Richards et al., 2008).

Functional annotation is an essential requirement for understanding the transcriptomic data of non-model organisms, and provides important information for the investigation of *B.*

longissima. GO is an international, standardized functional gene classification system, which offers a dynamically updated controlled vocabulary and a strictly defined concept to comprehensively describe the properties of genes and their products in any organism (Ye et al., 2006; Li et al., 2012). Based on the sequence homology, 18,153 (43.58%) unigenes were assigned to three functional groups based on their GO terminology: biological processes, molecular functions, and cellular components (Figure 1 and **Table S2**).

Table 1. Sequencing, assembling, and functional annotation.

Sequencing	Raw reads	82,865,068
	Clean reads	80,849,500
	Q20 percentage of clean reads	98.26%
	GC percentage of clean nucleotides	38.93%
Assembling	Contigs	58,815
	Length of total contigs (bp)	85,241,743
	Mean length of contigs (bp)	1,449
	Unigenes	41,652
	Length of total unigenes (bp)	38,829,793
	Mean length of unigenes (bp)	932
Annotation	Nr	23,077 (55.4%)
	Nt	4,933 (11.84%)
	SwissProt	17,523 (42.07%)
	PFAM	17,333 (41.61%)
	GO	18,153 (43.58%)
	KOG	13,733 (32.97%)
	KEGG	10,415 (25.00%)
	All databases	1,492 (3.58%)
	At least one database	25,162 (60.41%)
	Total	41,652 (100%)

GO, Gene Ontology; COG, Clusters of Orthologous Groups of proteins; KEGG, Kyoto Encyclopedia of Genes and Genomes.
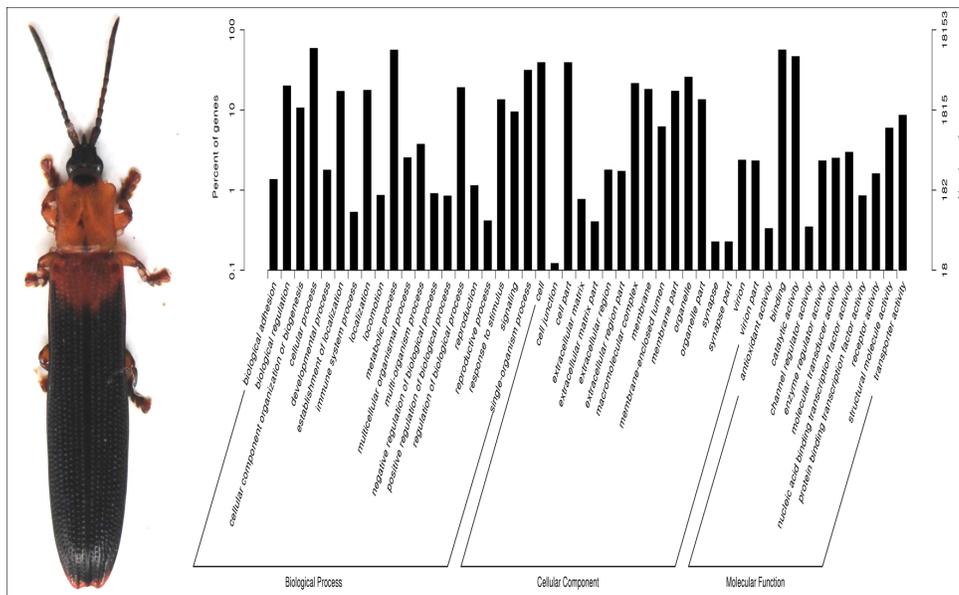
Figure 1. Gene Ontology (GO) categories.

However, some of these unigenes were involved in several GO terms. The majority of the GO terms represented biological processes (48,483, 45.67%), followed by cellular components (34,453, 32.45%) and molecular functions (23,224, 21.88%). High percentages of the unigenes were from the cellular processes, metabolic processes, binging, catalytic activity, cells, and cell part categories, but there were only a few unigenes from the metallochaperone activity, protein binding transcription factor activity, nucleoids, rhythmic processes, growth, or cell-killing categories. Cellular and metabolic processes were prominently represented, indicating that important metabolic activities and cell processes occur during this developmental stage. To further evaluate the completeness of our transcriptome library and the effectiveness of our annotation process, COG (that was designed as an attempt to classify proteins from completely sequenced genomes on the basis of the orthology concept) was used to predict and classify possible functions (Tatusov et al., 2000). In all, 13,733 unigenes were assigned to 26 function classes (Figure 2 and **Table S3**).
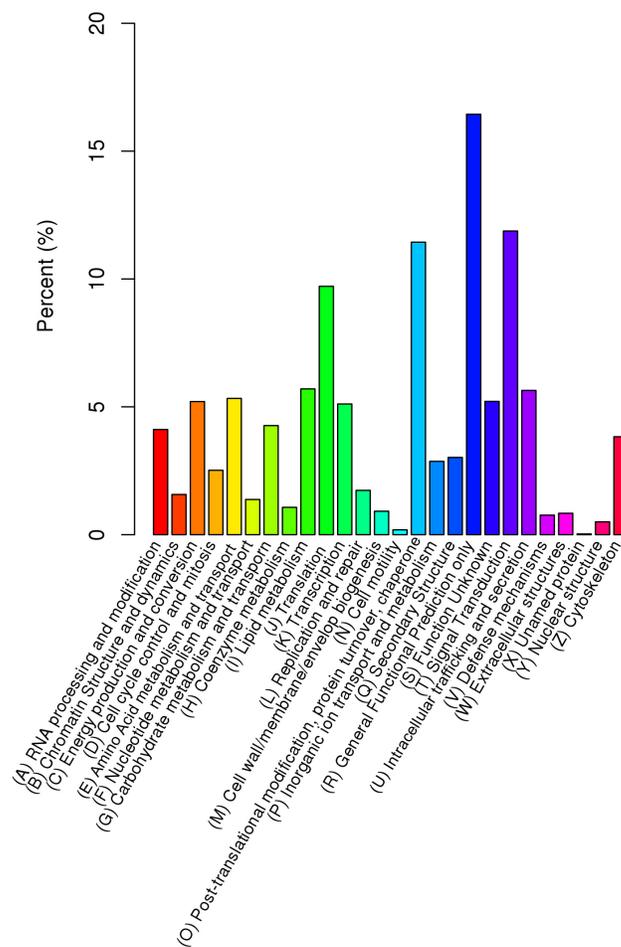
Figure 2. Functional assignments according to the Clusters of Orthologous Groups of proteins (COG) classification.

The most abundantly represented groups were general functional prediction only, signal transduction, post-translational modification, protein turnover, chaperone, and translation, which comprised 1334, 1571, 1631, and 2258 unigenes, respectively. The nuclear structure (69), cell motility (26), and unnamed protein (4) clusters were the smallest groups. Pathway-based analysis is important in understanding biological functions and gene interactions, and based on a comparison against the KEGG database, 10,415 assembled unigenes in 247 pathways were assigned to 31 groups (Figure 3 and **Table S4**). Of these, the best-represented groups were translation (1276), signal transduction (1000), carbohydrate metabolism (838), transport and catabolism (807), and folding, sorting, and degradation (800). These annotations provide a valuable resource for investigating specific processes, structures, functions, and pathways that will guide research on *B. longissima*.
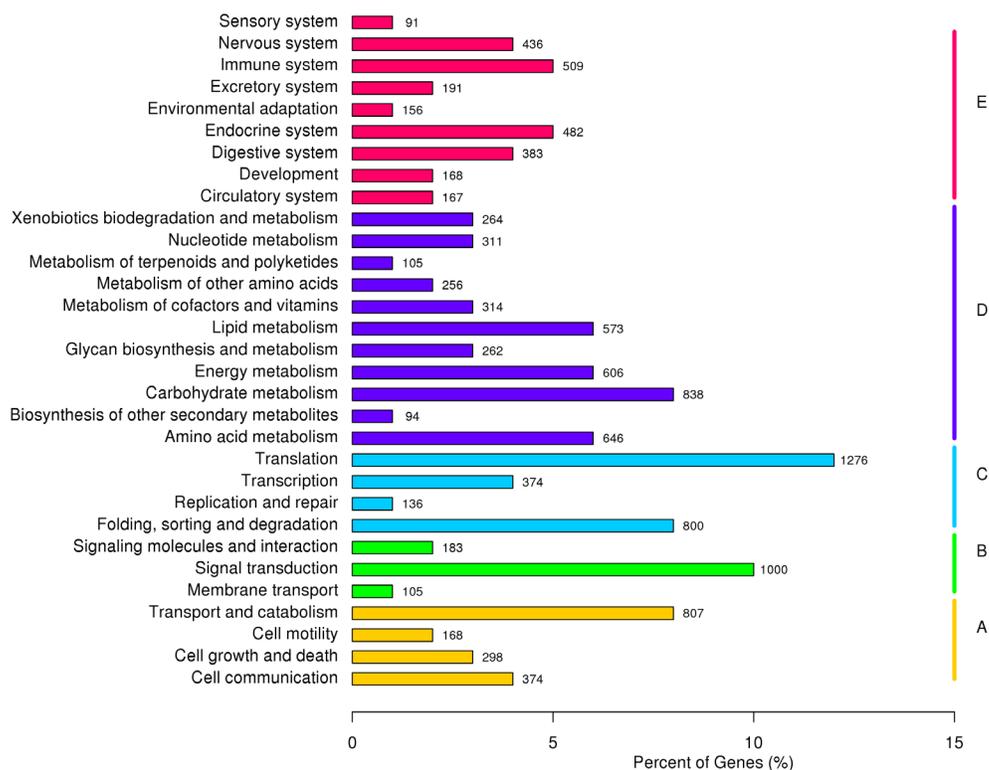
Figure 3. Kyoto Encyclopedia of Genes and Genomes (KEGG) annotation of unigenes.

ACKNOWLEDGMENTS

Conflicts of interest

The author declare no conflict of interest.

Supplementary material

REFERENCES

Acevedo AMT, Torres MAJ, Manting MME, Sabado E, et al. (2014). Sex ratios of *Brontispa longissima* (Gestro) infesting coconuts in selected provinces in the Philippines. *Ann. Biol. Res.* 5: 111-116.

Carvalho RA, Azeredo-Espin AM and Torres TT (2010). Deep sequencing of new world screw-worm transcripts to discover genes involved in insecticide resistance. *BMC Genomics* 11: 695.

Finn RD, Mistry J and Tate J (2010). The Pfam protein families database. *Nucleic Acids Res.* 38: D211-222.

Fu N, Wang Q and Shen HL (2013). *De novo* assembly, gene annotation and marker development using Illumina paired-end transcriptome sequences in celery (*Apium graveolens* L.). *PLoS One* 8: e57686.

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

Grabherr MG, Haas BJ, Yassour M, Levin JZ, et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29: 644-652.

Ju Q, Li X, Jiang XJ, Qu MJ, et al. (2014). Transcriptome and tissue-specific expression analysis of OBP and CSP genes in the dark black chafer. *Arch. Insect Biochem. Physiol.* doi: 10.1002/arch.21188.

Kumar A, Congiu L, Lindström L, Piiroinen S, et al. (2014). Sequencing, *de novo* assembly and annotation of the Colorado potato beetle, *Leptinotarsa decemlineata*, transcriptome. *PLoS One* 9: e86012.

Li C, Weng S, Chen Y, Yu X, et al. (2012). Analysis of *Litopenaeus vannamei* transcriptome using the next-generation DNA sequencing technique. *PLoS One* 7: e47442.

Lin YY, Jin T, Jin QA, Wen HB, et al. (2012). Differential susceptibilities of *Brontispa longissima* (Coleoptera: Hispidae) to insecticides in Southeast Asia. *J. Econ. Entomol.* 105: 988-993.

Price DP, Nagarajan V, Churbanov A, Houde P, et al. (2011). The fat body transcriptomes of the yellow fever mosquito *Aedes aegypti*, pre- and post-blood meal. *PLoS One* 6: e22573.

Rethinam P and Singh SP (2007). Current status of the coconut beetle outbreaks in the Asia-Pacific region. In: Developing an Asia-Pacific strategy for forest invasive species: The coconut beetle problem-bridging agriculture and forestry (Appanah S, Sim HC and Sankaran KV, eds.). RAP, Bangkok, 1-23.

Richards S, Gibbs RA, Weinstock GM, Brown SJ, et al. (2008). The genome of the model beetle and pest *Tribolium castaneum*. *Nature* 452: 949-955.

Takano S, Takasu K, Fushimi T, Ichiki RT, et al. (2012). Life history traits and damage potential of an invasive pest *Brontispa longissima* (Coleoptera: Chrysomelidae) on *Satakentia liukiuensis*. *Entomol. Sci.* 15: 238-245.

Takano S, Takasu K, Murata AM, Huong NT, et al. (2013). Comparative developmental and reproductive biology of geographical populations from two cryptic species in *Brontispa longissima* (Coleoptera: Chrysomelidae). *Entomol. Sci.* 16: 335-340.

Tatusov RL, Galperin MY, Natale DA and Koonin EV (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* 28: 33-36.

Voegele J (1989). Biological control of *Brontispa longissima* in Western Samoa: An ecological and economic evaluation. *Agri. Ecosyst. Environ.* 27: 315-329.

Ye J, Fang L, Zheng H, Zhang Y, et al. (2006). WEGO: a web tool for plotting GO annotations. *Nucleic Acids Res.* 34: W293-297.

Zhong Y, Li H, Liu K, Wen H, et al. (2005). Effects of temperature on *Brontispa longissima* population growth. *Chin. J. Appl. Ecol.* 16: 2369-2372.

Zhu JY, Zhao N and Yang B (2012). Global transcriptome profiling of the pine shoot beetle, *Tomicus yunnanensis* (Coleoptera: Scolytinae). *PLoS One* 7: e32291.

Zhu JY, Li YH, Yang S and Li QW (2013a). *De novo* assembly and characterization of the global transcriptome for *Rhyacionia leptotubula* using Illumina paired-end sequencing. *PLoS One* 8: e81096.

Zhu JY, Yang P, Zhang Z, Wu GX, et al. (2013b). Transcriptomic immune response of *Tenebrio molitor* pupae to parasitization by *Scleroderma guani*. *PLoS One* 8: e54411.

Zimmer CT, Maiwald F, Schorn C, Bass C, et al. (2014). A *de novo* transcriptome of European pollen beetle populations and its analysis, with special reference to insecticide action and resistance. *Insect Mol. Biol.* 23: 511-526.