Gene 575 (2016) 655–666

Contents lists available at ScienceDirect

Gene

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

Research paper

De novo transcriptome sequencing of *Cryptotermes domesticus* and comparative analysis of gene expression in response to different wood species

Wenjing Wu [a,b,c], Zhenyou Huang [a], Zhiqiang Li [a,b,c,*], Shijun Zhang [a,b,c], Xiaolin Liu [a,d], Daifei Gu [a,e]

[a] Guangdong Entomological Institute, Xingangxi Road 105, Guangzhou, Guangdong 510260, China
[b] Guangdong Public Laboratory of Wild Animal Conservation and Utilization, Xingangxi Road 105, Guangzhou, Guangdong 510260, China
[c] Guangdong Key Laboratory of Integrated Pest Management in Agriculture, Xingangxi Road 105, Guangzhou, Guangdong 510260, China
[d] Institute of Entomology and State Key Laboratory of Biocontrol, Sun Yat-sen University, Xingangxi Road 135, Guangzhou, Guangdong 510275, China
[e] College of Forestry, Northeast Forestry University, Hexing Road 26, Harbin, Heilongjiang 150040, China

ARTICLE INFO

Article history:
Received 24 July 2015
Accepted 18 September 2015
Available online 26 September 2015

Keywords:
Drywood termite
Casuarina equisetifolia
Koompassia excelsa
Myristica sp.
Feeding response

ABSTRACT

The drywood termite *Cryptotermes domesticus* is an important worldwide pest with limited genomic resources that causes substantial damage to dry timber and structural lumber. Here, we performed transcriptome sequencing for *Cr. domesticus* pseudergate using Illumina paired-end sequencing technology. A total of 108,745,470 clean reads were collected and assembled into 302,979 contigs with an average length of 648 bp and an N50 length of 893 bp. A total of 185,248 unigenes and 100,680 proteins were identified among the assembled contigs. Of these, there were 152,317 (50.27%) contigs with significant similarity to publicly available databases. To understand how the termites respond to phylogenetically diverse wood species, variations in gene expression were examined among pseudergates feeding on three wood species from different plant families, *Casuarina equisetifolia* (CE), *Koompassia excelsa* (KE) and *Myristica* sp. (MS). A total of 417 (118 up-regulated/299 down-regulated), 599 (148 up-regulated/451 down-regulated) and 505 (223 up-regulated/282 down-regulated) differentially expressed genes were detected in KE vs. CE, KE vs. MS and CE vs. MS, respectively. Digital gene expression analysis indicated that different wood species played an important role in the expression of termite genes, such as genes involved in carbohydrate metabolism, and proteins with catalytic activity and hydrolase activity. Additionally, the genes encoding cellulase were identified and analyzed. This study provides the first primary transcriptome of *Cr. domesticus* and lays a foundation for future functional genomics studies in the feeding responses.

1. Introduction

Drywood termite *Cryptotermes domesticus* (Blattodea: Kalotermitidae) is a related lower termite species, and the termite foragers are called a "pseudergate" without a true worker caste. *Cr. domesticus* is capable of living in a piece of dry wood without free water for years, so it is a serious pest causing substantial damage to dry timber or structural lumber (Huang et al. 2000; Rust and Su 2012). Moreover, *Cr. domesticus* is an important worldwide invasive termite species, spreading from Southeast Asia to the Pacific islands and Central America (Evans et al. 2013). In China, *Cr. domesticus* has been found in Hainan, Guangdong, Guangxi, Yunnan and Taiwan, and it is one of the most economically important termite pests (Huang et al. 2000). This species has a great capacity to feed on wood from different plant families and in which their offspring can grow normally (Qian et al. 2005; Huang et al. 2007; Huang et al. 2011). *Myristica* sp. (Myristicaceae) and *Koompassia excelsa* (Caesalpiniaceae) are common timbers imported from Southeast Asia, and *Casuarina equisetifolia* (Casuarinaceae) is an important shelter-forest tree species in South China. *Cr. domesticus* colonies can survive in the specimens from *K. excelsa* and *Myristica* sp. (Qian et al. 2005; Huang et al. 2007) as well as *Ca. equisetifolia* (based on our observations). Additionally, we observed in our laboratory that the alates fly was also associated with these tree species. However, after an approximately ten-year observation period, we found that the fitness of the *Cr. domesticus* colonies

Abbreviations: BG, β-glucosidase; CE, pseudergates living in *Casuarina equisetifolia* woods; COG, Cluster of Orthologous Groups; DEG, differentially expressed gene; DGE, digital gene expression; GHF, glycosyl hydrolase family; GO, gene ontology; KE, pseudergates living in *Koompassia excelsa* woods; KEGG, Kyoto Encyclopedia of Genes and Genomes; KOG, euKaryotic Orthologous Groups of proteins; MS, pseudergates living in *Myristica* sp. woods; NR, NCBI non-redundant protein; NT, NCBI nucleotide; ORF, open reading frame; PFAM, protein family; RPKM, reads per kilobase per million; SRA, NCBI Sequence Read Archive; SSR, Simple Sequence Repeat; TrEMBL, a computer-annotated supplement of Swiss-Prot that contains all the translations of EMBL nucleotide sequence entries not yet integrated in Swiss-Prot.

* Corresponding author at: Guangdong Entomological Institute, Xingangxi Road 105, Guangzhou, Guangdong 510260, China.
E-mail address: lizq@gdei.gd.cn (Z. Li).

http://dx.doi.org/10.1016/j.gene.2015.09.055
0378-1119/

was different among the tree species, and the consumption of *Myristica* sp. was higher in laboratory conditions.

At present, most studies of drywood termites are primarily focused on their distribution (Scheffrahn et al. 2000; Li et al. 2008; Evans et al. 2013), resistance (Qian et al. 2005; Garcia et al. 2011; Hadi et al. 2012) and control (Sbeghen-Loss et al. 2011; Rust and Su 2012) because the termites spread rapidly and are difficult to control. Currently, the termite genome of the lower termite *Zootermopsis nevadensis* (Termopsidae) has been reported (Terrapon et al. 2014). However, adaptation mechanisms of drywood termites in different wood species have not been reported at the molecular level. Although the feeding habits of drywood termites have been well studied, it is very difficult to investigate drywood termites as cryptic and single-piece nesters. To understand the feeding responses of *Cr. domesticus* to phylogenetically distant wood species, we assembled the transcriptome of *Cr. domesticus* and compared the gene expression profiles among different wood species. Here, we analyze *Cr. domesticus* colonies reared in *Ca. equisetifolia*, *K. excelsa* and *Myristica* sp. This work may lay the foundation to reveal molecular fitness mechanisms and to prohibit the continued spread of drywood termites.

2. Materials and methods

2.1. Termite materials

Cr. domesticus colonies were collected from Zhanjiang City, Guangdong Province, China, and the termites from infested wood species were reared in a glass jar (90 cm long, 50 cm wide, 70 cm high) until primary reproductives occurred in the laboratory at the Guangdong Entomological Institute, Guangzhou City, Guangdong Province, China. From 2002 to 2003, the new primary reproductives as well as their offspring were collected to establish a new colony in three wood species, *Ca. equisetifolia* (CE), *K. excelsa* (KE) and *Myristica* sp. (MS) at a constant temperature of 27 °C and relative humidity of 80%. In 2014, *Cr. domesticus* pseudergates were collected from CE, KE and MS and were immediately frozen and stored in liquid nitrogen until used.

For each of the termite colonies, 5 individual pseudergates were randomly chosen and then pooled to generate one mixed sample per colony. Three RNA samples (five pseudergates each from CE, KE and MS woods) were extracted using TRIzol reagent (Invitrogen, Carlsbad, CA, USA) according to the manufacturer's instructions. Total RNA integrity was checked with an Agilent Technologies 2100 Bioanalyzer (Agilent Technologies, Inc., Santa Clara, CA, USA) and a 1% agarose gel. All samples had a RNA integrity number (RIN) of more than 6.

2.2. cDNA library construction and Illumina sequencing for transcriptome analysis

To obtain complete gene expression information, a pooled RNA sample obtained from CE, KE and MS was used to construct a reference library for *de novo* transcriptome sequencing. According to the Illumina manufacturer's protocols, poly $(A)^+$ RNA was purified from 5 μg of total RNA using oligo (dT) magnetic beads and fragmented into short sequences in the presence of divalent cations at 94 °C for 5 min. The cleaved poly $(A)^+$ RNA was transcribed, and then second-strand cDNA synthesis was performed. After the end-repair and ligation of adapters, the products were amplified by PCR and purified using the QIAquick PCR Purification Kit (QIAGEN, Inc., Valencia, CA, USA) to create a cDNA library. The cDNA library was sequenced on the Illumina HiSeq2000 platform at the Annoroad Gene Technology Corporation (Beijing, China), which generated approximately 100 bp paired-end raw reads.

2.3. Data filter and de novo assembly of reads

Raw sequencing data were evaluated by the quality control tool FastQC (version 0.11.2) to characterize sequence quality, GC content, base N content, and adapter content. After quality control testing, the raw reads were cleaned by removing the adapter sequences, low quality sequences (more than 15% bases with quality value under 19), and any reads with more than 5% of unknown sequences designated as 'N'. The clean reads were then used for transcriptome *de novo* assembly using the Trinity program with its default parameter values (Grabherr et al. 2011). In consideration of gut symbionts, we filtered the sequences of all of the known protozoa of *Cr. domesticus*: *Devescovina axiomaculata*, *Foaina acontophira*, *Foaina grassii*, and *Stephanympha helumbium* (Huang et al. 2000). After filtering, the assembled transcriptome of *Cr. domesticus* was used as a reference for the following analyses.

2.4. Analysis and functional annotation of sequences

All assembled contigs were searched against NCBI nucleotide (NT), NCBI non-redundant protein (NR), Swiss-Prot, Kyoto Encyclopedia of Genes and Genome (KEGG), Cluster of Orthologous Groups (COG), euKaryotic Orthologous Groups (KOG), and TrEMBL (a computer-annotated supplement of Swiss-Prot that contains all the translations of EMBL nucleotide sequence entries not yet integrated in Swiss-Prot) databases using the BLAST algorithm (Korf et al. 2003) with an E-value cut-off of 10^{-5}. If there were conflicting results between the different databases, priority was given, in descending order, to NR, Swiss-Prot, KEGG and COG. The best alignments from the four databases were used to determine the direction of the contigs. Contigs with NR annotation were further analyzed with Blast2GO (version 3.0.8) to obtain their gene ontology (GO) annotations (Conesa et al. 2005). The Web Gene Ontology (WEGO) annotation software was used to classify all of the contigs according to GO functions and to understand the distribution of gene functions of the species from the macro level (Ye et al. 2006). The TransDecoder (v2.0.1) program was used to predict the open reading frame (ORF) from contig sequences and translate them into peptide sequences. TransDecoder identifies likely coding sequences based on the following criteria: a minimum length ORF of 300 bp; a log-likelihood score > 0; the above coding score is greatest when the ORF is scored in the 1st reading frame as compared to scores in the other 5 reading frames; if a candidate ORF is found fully encapsulated by the coordinates of another candidate ORF, the longer one is reported. The peptide sequences were aligned with Pfam (Protein family) using HMMER3 with an E-value cut-off of 10^{-10}.

2.5. SSR detection

The assembled sequences were used to identify the signatures of SSRs (Simple Sequence Repeat). FASTA files containing all of the assembled sequences were used as an input file in the Batch Primer3 software (http://probes.pw.usda.gov/batchprimer3/). The parameters were adjusted for identification of perfect di-, tri-, tetra-, penta-, and hexanucleotide motifs with a minimum of 9, 6, 5, 4, and 4 repeats, respectively. Mononucleotide repeats were ignored because it was difficult to distinguish genuine mononucleotide repeats from polyadenylation products and single nucleotide stretch errors generated by sequencing.

2.6. Digital gene expression (DGE) library preparation and sequencing

DGE library preparation for the three different colony samples (CE, KE, MS) was performed in parallel using Illumina Digital Gene Expression-Tag Profiling with *Nla*III kit (Illumina Inc., San Diego, CA, USA). Briefly, mRNA was purified from 2 μg of total RNA by binding the mRNA to a magnetic oligo (dT) bead. First and second strand cDNAs were directly synthesized and bead-bound cDNA was subsequently digested with *Nla*III. The cDNA fragments with 3′ ends were