GMR Genetics and Molecular Research
Online Journal
ISSN:1676-5680

Transcriptome profiling of the crofton weed gall fly *Procecidochares utilis*

X. Gao[1], J.Y. Zhu[2], S. Ma[1], Z. Zhang[3], C. Xiao[1], Q. Li[1], Z.Y. Li[1] and G.X. Wu[1]

[1]College of Plant Protection, Yunnan Agricultural University, Kunming, China
[2]Key Laboratory of Forest Disaster Warning and Control of Yunnan Province, Southwest Forestry University, Kunming, China
[3]Department of Pathogen Biology, Taishan Medical University, Tai'an, China

Corresponding author: G.X. Wu
E-mail: wugx1@163.com

Genet. Mol. Res. 13 (2): 2857-2864 (2014)
Received December 12, 2012
Accepted May 16, 2013
Published March 19, 2014
DOI http://dx.doi.org/10.4238/2014.March.19.1

ABSTRACT. *Procecidochares utilis* is a tephritid gall fly, which is known to be an effective biological agent that can be used to control the notoriously widespread crofton weed *Eupatorium adenophorum*. Despite its importance, genetic resources for *P. utilis* remain scarce. In this study, 1.2 Gb sequences were generated using Illumina paired-end sequencing technology. *De novo* assemblies yielded 491,760 contigs, 90,474 scaffolds, and 58,562 unigenes. Among the unigenes, 34,809 (59.44%) had a homologous match against the National Center for Biotechnology Information non-redundant protein database by translated Basic Local Alignment Search Tool (BlastX) with a cut-off E-value of 10^{-5}. Among the unigenes, 57,627 were classified in the Gene Ontology database, 15,910 were assigned to Clusters of Orthologous Groups, and 38,565 were found in Kyoto Encyclopedia of Genes and Genomes. In addition, 5723 simple sequence repeats (SSRs) were discovered based on the unigene sequences. The transcriptome sequences and SSRs obtained represent a major molecular resource for *P. utilis*, which will extend our knowledge of the comparative and

functional genomics of this organism and enable population genomic and gene-based association studies of the gall fly.

Key words: Gall fly; *Procecidochares utilis*; Transcriptome; Simple sequence repeat marker; Illumina sequencing

INTRODUCTION

Procecidochares utilis (Diptera: Trypetidae) is a tephritid stem gall-forming fly, which is restricted to a single host plant, the composite *Eupatorium adenophorum* known as crofton weed belonging to the Asteraceae family (Haseler, 1965). Both *P. utilis* and its host are natives of Central America, mainly Mexico. However, *E. adenophorum*, a hazardous invading species, has successfully invaded many regions on the globe in a wide variety of natural and anthropogenic ecosystems that range from forest and grassland to farmland (Sang et al., 2010) and is now one of the most noxious invasive plants worldwide, causing serious economic losses and environmental damages. Since *P. utilis* was first introduced in Hawaii in 1945 to combat *E. adenophorum*, it has been highly successful in the control of this plant in some localities (Bess and Frank, 1959). Much attention has been paid to manage *E. adenophorum* using this biological control method because it often grew in inaccessible areas. This fly was subsequently used in many other countries such as New Zealand, Australia, and China, and was proposed to be a suitable agent for the control of *E. adenophorum* (Erasmus et al., 1992; Ma et al., 2012). With respect to its prominence as a bio-control agent, a great deal of research has been conducted on the basic ecological and biological characteristics of *P. utilis* (Li et al., 2006), while the mechanisms behind molecular regulations in this species remain poorly understood. At present, only 28 nucleotide sequences are available in the National Center for Biotechnology Information (NCBI) database (prior to October 2012). Obviously, these genetic data are scarce and insufficient for elucidating the molecular mechanism of *P. utilis* in ecological systems.

Over the past several years, the introduction of novel next-generation high-throughput sequencing, such as 454 and Illumina pyrosequencing, has been an efficient means to gain functional genomic level data for non-model organisms, has provided fascinating opportunities in the life sciences, and has dramatically improved the efficiency and speed of gene and genetic marker discovery (Xue et al., 2010; Seal et al., 2012; Zhu et al., 2012). In light of its advantages, the transcriptome data (over one billion bases of high-quality DNA sequence) of *P. utilis* were generated with the Illumina technology in this study. Additionally, a great number of simple sequence repeats (SSRs) (or microsatellites) were obtained. The large-scale transcriptome sequence data and genetic markers are undoubtedly valuable for molecular studies of the gall fly.

MATERIAL AND METHODS

Insects

Galls of *E. adenophorum* were collected in the suburbs of Kunming, China. All galls were kept in a cage, were covered with gauze, and were maintained at 25°C with 75% relative humidity and a 14/10-light/dark cycle. After emergence, *P. utilis* adults were fed a 20% honeydew solution. Then, the *E. adenophorum* that was cultured in the laboratory was used as the host to rear *P. utilis*.

cDNA preparation and sequencing

The total RNA of *P. utilis* was extracted using Trizol reagent (Invitrogen, USA). The RNA quality and quantity were verified by the 2100 Bioanalyzer (Agilent Technologies, USA) with a minimum RNA integrity number value of 8. According to the Illumina manufacturer instructions, mRNA was purified from 20 µg 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 RNA fragments were used for first-strand cDNA synthesis using reverse transcriptase and random primers. Then, double-stranded cDNA was synthesized using DNA polymerase I and RNaseH. After end repair and ligation of adaptors, the products were amplified by polymerase chain reaction (PCR) and purified using the QIAquick PCR Purification Kit to create a cDNA library. The cDNA library was sequenced on the Illumina sequencing platform (HiSeq2000) according to manufacturer instructions. The raw data have been deposited in the Sequence Read Archive (DRA) of DNA Data Bank of Japan (DDBJ).

Assembly and annotation

The raw reads from the images were generated by Illumina Genome Analyzer Pipeline 1.3. Prior to data analysis, raw reads were first cleaned by removing adaptor sequences and low-quality sequences (reads with unknown sequences). After the removal of low-quality reads, the processed reads were assembled using the SOAP *de novo* software and clustered with TGI Clustering tools (Pertea et al., 2003; Li et al., 2009). All assembled unigenes were annotated based on the protein sequences in the NCBI non-redundant (Nr) protein database by translated Basic Local Alignment Search Tool (BlastX) with an E-value of 1E-5. The BLAST2GO software (Aparicio et al., 2006) was employed to deal with the BlastX results in XML format and then to perform the functional annotation by Gene Ontology (GO), Cluster of Orthologous Groups (COG), and Kyoto Encyclopedia of Genes and Genomes (KEGG) metabolism pathways.

Marker identification

SSRs were identified with the microsatellite identification tool (MISA) (http://pgrc.ipk-gatersleben.de/misa/). The analysis was run with the length of the repeat motifs to be searched set between 1 and 6 and the minimum number of repeats set equal to 6. Minimum repeat numbers of 12, 6, 5, 4, 4, and 4 were set for mono-, di-, tri-, tetra-, penta-, and hexanucleotide microsatellites, respectively. Primer3 v2.2.2 (http://primer3.sourceforge.net) was used to design the primer pairs with default settings. Then, the following selection criteria were applied: primers cannot contain an SSR (2-6 bases repeated more than 4 times), and primers have only 1 matched unigene when mapped against all unigenes (allow 3-base mismatch at the 5'-end and 1-base mismatch at the 3'-end) (Wang et al., 2012).

RESULTS AND DISCUSSION

Illumina sequencing and assembly

In a single run, 13,333,334 raw reads with an average sequence length of 90 bp were

generated from the library, which encompassed 1.2 Gb of sequence data (Table 1). The GC content was 39.21%, which is similar to that of insects from previous transcriptome studies (Price et al., 2011). After quality filtering, assembly and clustering, a total of 491,760 contigs and 90,474 scaffolds were obtained, representing 58,562 unigenes. The average length of contigs, scaffolds, and unigenes were 128, 328, and 427 bp, respectively. Their length distributions are shown in **Figures S1** and **S2** and Figure 1. Their distribution patterns were similar. These results suggested that the transcriptome sequencing data from *P. utilis* were effectively assembled. The majority of them were between 75 and 500 bp in length. The short length resulted from the sequencing capacity of the Illumina technology or the low coverage of the transcriptome that was represented in this dataset (Yang et al., 2010; Zhu et al., 2012).

Table 1. Summary of the short reads and the assemblies.

	Reads	Contigs	Scaffolds	Unigenes
Number of sequences	13,333,334	491,760	90,474	58,562
Mean length (bp)	90	128	328	427
Total length (bp)	1,200,000,060	62,710,986	29,660,670	25,013,511

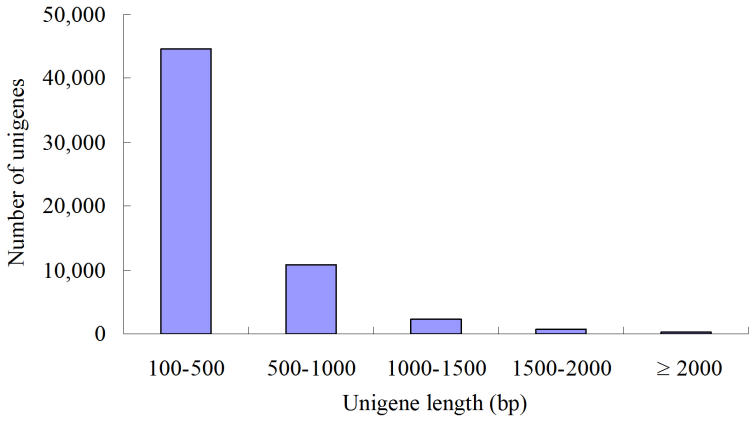

Figure 1. Length distribution of unigenes.

Functional annotation

By BlastX search with a cut-off E-value of 10^{-5}, 34,809 unigenes (59.44% of all unigenes) returned above cut-off Blast results when searched against the Nr nucleotide database (**Table S1**). However, the other 23,753 unigenes (40.56% of all unigenes) could not be matched to known genes, which is relatively common in transcriptome sequence analysis. This might be due to several possible factors, including the propensity of short read lengths to hinder assembly, incomplete incidence of genes corresponding to low abundance transcripts in current sequence databases, and very little sequence information from closely related species (Kaur et al., 2011). The E-value distribution of the top hits in the Nr database showed that 10.55% of the mapped sequences have strong homology (E-value was less than 1E-99), whereas 20.41 and 69.03% of the homologous sequences ranged from 1E-50 to 1E-99 and

1E-5 to 1E-49 (Figure 2A). For species distribution, 53.95% of the distinct sequences have top matches (first hit) that were trained with sequences from the dipteran species fruit fly (*Drosophila*), followed by *Nasonia vitripennis* (8.66%) and *Apis mellifera* (6.50%) (Figure 2B). This could result because *P. utilis* and *Drosophila* belong to the same order and the *Drosophila* genome is fully sequenced and currently represents the vast majority of dipteran sequences available in GenBank, which is similar to that of *Trialeurodes vaporariorum* (Karatolos et al., 2011).

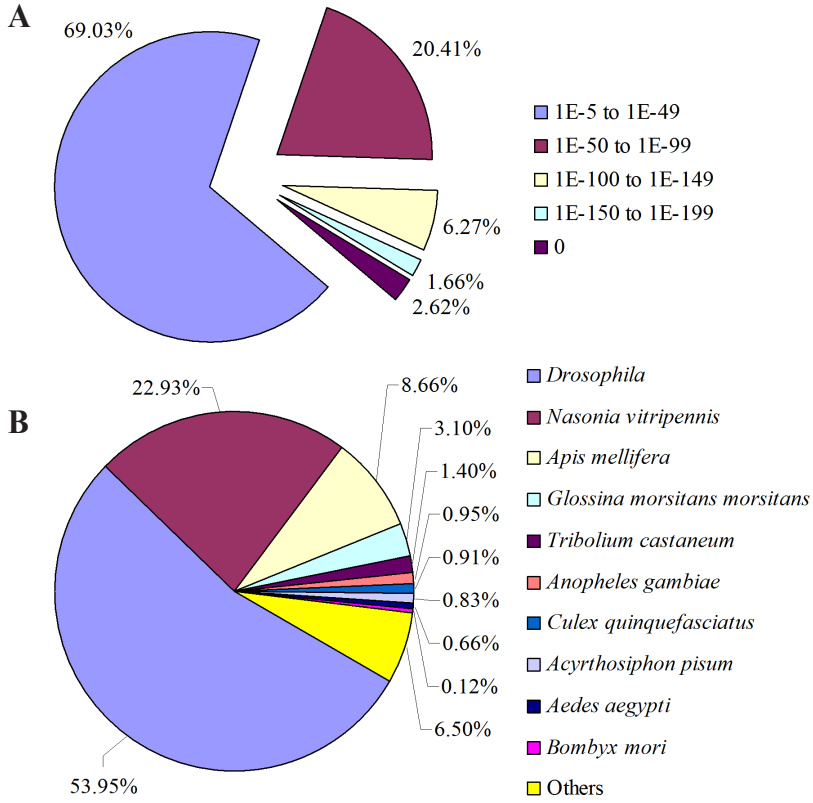

Figure 2. Characteristics of homology search of unigenes against the Nr database. **A.** E-value distribution. **B.** Species distribution. The first hit of each sequence with a cut-off E-value of 1.0E-5 was used for analysis.

The GO classification provides the ontology of defined terms that represent gene product properties, such as biological processes, cellular components, and molecular functions (Xie et al., 2012). Under the GO annotation, the unigenes mapped to 57,627 GO terms, which were categorized into 48 functional groups (**Table S2**). Regarding the biological process classification, the main groups were involved in cellular process (19.22%) and metabolic process (14.85%). Within the cellular component category, the vast majority was cell (29.45%), cell part (29.45%), and organelle (16.23%). In the category of molecular function, binding (45.16%) and catalytic (35.01%) showed the highest abundance. Among all of the terms, cell killing, viral reproduction, metallochaperone, nutrient reservoir, proteasome regulator, and protein tag were lowly represented, and no more than 5 unigenes were assigned to each term. To further evaluate the effectiveness of

the annotation process, the COG assignments were used. Overall, 15,910 unigenes were classified into 25 categories that involved different processes, suggesting that the assembled unigenes represented a wide diversity of transcripts in the *P. utilis* genome (**Figure S3**). The terms of general function prediction only (16.29%); translation, ribosomal structure and biogenesis, transcription, and replication (9.05%); posttranslational modification, protein turnover, and chaperones (8.19%); and recombination and repair (7.39%) were the most represented. Extracellular structures and nuclear structure were the smallest groups. There were only 9 and 13 unigenes assigned to them, respectively. To survey genes that were involved in important pathways, annotated unigenes were mapped to KEGG pathways. In total, 38,565 unigenes were grouped into 202 known metabolic or signaling pathways (**Table S3**). The major pathways were metabolic pathways (9.33%), spliceosome (2.19%), pathways in cancer (2.10%), focal adhesion (1.75%), purine metabolism (1.68%), endocytosis (1.55%), ubiquitin-mediated proteolysis (1.55%), and regulation of actin cytoskeleton (1.54%). Following the function of GO and COG terms and KEGG pathways, some categories were of particular interest to help discover important genes for further investigation.

SSR discovery

Based on the unigene sequences, 5723 SSRs were identified within 58,562 unigenes. Of the SSRs, 21.09, 31.71, and 42.53% were mononucleotide, dinucleotide, and trinucleotide repeats, respectively (Figure 3). The results demonstrated that trinucleotide repeats were the most abundant microsatellites in this transcriptome data set, which is consistent with other recent reports in insects (Xu et al., 2012). Among all of the SSR types, AAC/GTT (21.58%) represented the dominant type, followed by A/T (21.02%), AG/CT (14.29%), AT/AT (8.53%), and AC/GT (8.32%) (**Table S4**). Of the 58,562 unigenes, 809 contained more than 1 SSR. Using Primer3 v2.2.2, 4666 primer pairs were successfully designed according to the identified SSRs (**Table S5**). Molecular markers are widely applied in the study of insect evolution and differentiation. The results would be helpful for investigating the functional diversity in *P. utilis* natural populations.

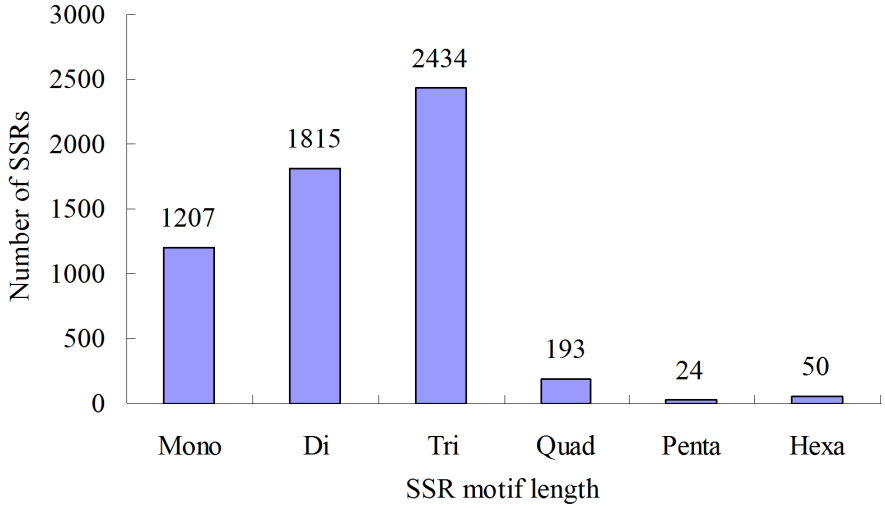

Figure 3. Summary statistics for the SSRs identified in the transcriptome.

CONCLUSION

In this study, the generation of a substantial transcriptome dataset of *P. utilis* was described. The 58,562 unigenes represent a major genomic level resource for *P. utilis* and will be useful for comparative and functional genomic studies in the gall fly. We identified 5723 SSRs from this dataset, which will be useful for future studies on biodiversity, molecular taxonomy, population genetics, and genetic linkage mapping.

ACKNOWLEDGMENTS

Research supported by funding from the Natural Science Foundation of China (#30960221, #31340012 and #81071390), the Yunnan Provincial Science and Technology Innovation Team Plan of China (#2011HC005) and the Program for Innovative Research Team (in Science and Technology) in University of Yunnan Province.

Supplementary material

REFERENCES

Aparicio G, Götz S, Conesa A, Segrelles D, et al. (2006). Blast2GO goes grid: developing a grid-enabled prototype for functional genomics analysis. *Stud. Health Technol. Inform.* 120: 194-204.

Bess HA and Frank HH (1959). Biological control of Pamakani, *Eupatorium adenophorum*, in Hawaii by a tephritid gall fly, *Procecidochares utilis*. 2. Population studies of the weed, the fly and the parasites of the fly. *Ecology* 40: 244-249.

Erasmus DJ, Bennett PH and van Staden J (1992). The effect of galls induced by the gall fly *Procecidochares utilis* on vegetative growth and reproductive potential of crofton weed, *Ageratina adenophora*. *Ann. Appl. Biol.* 120: 173-181.

Haseler WH (1965). Life history and behavior of the crofton weed gall fly *Procecidochares utilis* Stone (Diptera: Trypetidae). *J. Entomol. Soc. Queensl.* 4: 27-32.

Karatolos N, Pauchet Y, Wilkinson P, Chauhan R, et al. (2011). Pyrosequencing the transcriptome of the greenhouse whitefly, *Trialeurodes vaporariorum* reveals multiple transcripts encoding insecticide targets and detoxifying enzymes. *BMC Genomics* 12: 56.

Kaur S, Cogan NO, Pembleton LW, Shinozuka M, et al. (2011). Transcriptome sequencing of lentil based on second-generation technology permits large-scale unigene assembly and SSR marker discovery. *BMC Genomics* 12: 265.

Li AF, Gao XM, Dang WG, Huang RX, et al. (2006). Parasitism of *Procecidochares utilis* and its effect on growth and reproduction of *Eupatorium adenophorum*. *Chin. J. Plant Ecol.* 30: 496-503.

Li R, Yu C, Li Y, Lam TW, et al. (2009). SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* 25: 1966-1967.

Ma S, Gao X, Zhu JY, Wu GX, et al. (2012). Effects of temperature and supplementary nutrients on the life span of adult *Procecidochares utilis* (Diptera: Tephritidae). *J. Biaosafety* 21: 236-239.

Pertea G, Huang X, Liang F, Antonescu V, et al. (2003). TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics* 19: 651-652.

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.

Sang WG, Zhu L and Axmacher JC (2010). Invasion pattern of *Eupatorium adenophorum* Spreng in southern China. *Biol. Invasions* 12: 1721-1730.

Seal S, Patel MV, Collins C, Colvin J, et al. (2012). Next generation transcriptome sequencing and quantitative real-time PCR technologies for characterisation of the *Bemisia tabaci* asia 1 mtCOI phylogenetic clade. *J. Integr. Agr.* 11: 281-292.

Wang S, Wang X, He Q, Liu X, et al. (2012). Transcriptome analysis of the roots at early and late seedling stages using Illumina paired-end sequencing and development of EST-SSR markers in radish. *Plant Cell Rep.* 31: 1437-1447.

Xie F, Burklew CE, Yang Y, Liu M, et al. (2012). *De novo* sequencing and a comprehensive analysis of purple sweet potato (*Impomoea batatas* L.) transcriptome. *Planta* 236: 101-113.

Xu Y, Zhou W, Zhou Y, Wu J, et al. (2012). Transcriptome and comparative gene expression analysis of *Sogatella furcifera* (Horvath) in response to southern rice black-streaked dwarf virus. *PLoS One* 7: e36238.

Xue J, Bao YY, Li BL, Cheng YB, et al. (2010). Transcriptome analysis of the brown planthopper *Nilaparvata lugens*. *PLoS One* 5: e14233.

Yang H, Hu L, Hurek T and Reinhold-Hurek B (2010). Global characterization of the root transcriptome of a wild species of rice, *Oryza longistaminata*, by deep sequencing. *BMC Genomics* 11: 705.

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

