Marine Genomics xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Marine Genomics

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

De novo assembly and annotation of the *Avicennia officinalis* L. transcriptome

Haomin Lyu [1], Xinnian Li [1], Zixiao Guo, Ziwen He *, Suhua Shi

State Key Laboratory of Biocontrol and Guangdong Key Laboratory of Plant Resources, Sun Yat-sen University, Guangzhou 510275, China

ARTICLE INFO

Article history:
Received 29 April 2017
Received in revised form 3 July 2017
Accepted 10 July 2017
Available online xxxx

Keywords:
Avicennia officinalis
RNA-seq data
De novo assembly
Annotation
Salt glands
Candidate genes

ABSTRACT

Avicennia officinalis L. is a typical mangrove species, inhabiting inhospitable environments in the interface between sea and land. In this study, we generated RNA-seq data to *de novo* assemble the *A. officinalis* transcriptome. Starting with 36.24 million 100 bp paired-end reads, 38,576 high-confidence transcripts with an average length of 834 bp were produced after filtration of weakly expressed and redundant transcripts. We found known protein homologs for 22,254 of these transcripts, and assigned them to at least one of 119 gene ontology (GO) terms. In addition, we identified different copies and isoforms of three candidate genes, *AoPIP*, *AoTIP* and *AoDHN1*, which might be involved in salt excretion via salt glands. All these genes were highly expressed in leaf tissue of *A. officinalis*, suggestive of a complicated mechanism of response to salt stress. We further identified 613 microsatellite markers for the assessment of genetic diversity and population differentiation in *A. officinalis*. Genomic resources generated in this study would be an important foundation for future research into molecular mechanisms underlying salt and other stress tolerance, as well as the evolutionary history of this mangrove species and its relatives.

1. Introduction

Mangroves are woody plants inhabiting intertidal zones of tropical or subtropical coasts (Tomlinson, 1986). They have evolved both morphological and physiological strategies to thrive in the inhospitable conditions prevalent in their habitats, such as high salinity, hypoxia, high sedimentation and muddy anaerobic soils (Giri et al., 2011). *Avicennia officinalis* L. is a widespread mangrove species distributed throughout the Indo-West Pacific region (Duke, 1991). It exhibits some typical adaptive traits of mangroves: breathing roots, salt tolerance, and viviparity (Tomlinson, 1986). While this mangrove species provides important ecological and economic benefits for the area it inhabits, it has lost about 24% of its habitat since 1980 and is at risk from coastal development, overcutting, and global climate change (Duke et al., 2010).

Strategies mangroves taken to tolerate high salinity include ultrafiltration, salt secretion and ion sequestration (Liang et al., 2008). Of all the ~70 mangrove species, only several mangrove species (8 in *Avicennia* and 7 in other genera) use the unique mechanism of salt secretion that employs multicellular salt glands (Tomlinson, 1986). Among them, *A. officinalis* possesses almost the most significant features of salt secreting by salt glands in leaf tissue, and deposits obvious salt crystals (Tomlinson, 1986; Parida and Jha, 2010). These structures are evolved in only a few orders of halophytes and absent in the model species *Arabidopsis thaliana* (Flowers et al., 2010; Flowers and Colmer, 2008). Although transcriptomes of some other mangrove species have been sequenced (He et al., 2015; Li et al., 2017; Yang et al., 2015), most of them were weakly salt-secreting or non-secreting (Tomlinson, 1986). Thus, *A. officinalis* can be a good system to study the salt-gland mechanism of salt tolerance. Compared with other salt-secreting species with transcriptome data like *A. marina* (Huang et al., 2014), a more solid foundation has been laid by structural and functional examination of these organs in *A. officinalis* (Jyothi-Prakash et al., 2014; Tan et al., 2013). Two aquaporin genes (*AoPIP*, *AoTIP*) and one dehydrin gene (*AoDHN1*) are preferentially expressed in *A. officinalis* salt glands and might be associated with response to increasing salt concentration (Jyothi-Prakash et al., 2014; Tan et al., 2013). Genomic resources, such as expression profiles of functional genes in *A. officinalis* leaves, could enable more comprehensive molecular examination of excess salt secretion by salt glands.

An interesting additional feature of *A. officinalis* is the among-population differentiation in leaf morphology (Duke et al., 1998; Duke, 1991). Development of abundant genomic sequences and genetic markers can open the way to molecular population genetic examination of genetic variation within the species and elucidate the significance of this polymorphism. Since a high-confidence phylogenetic relationship among eight species in genus *Avicennia* has already been estimated (Li et al., 2016), these population-genetic approaches will fit into a robust overall evolutionary framework.

* Corresponding author.
E-mail address: heziwen@mail.sysu.edu.cn (Z. He).
[1] HL and XL contributed equally to this paper.

http://dx.doi.org/10.1016/j.margen.2017.07.002
1874-7787/

In this study, we present a *de novo* assembly and annotation of the *A. officinalis* transcriptome. This genomic resource can be leveraged in future mechanistic studies of a variety of aspects of mangrove biology, including the unusual salt tolerance adaptations using salt glands. As an example of the potential use of this data set in population genetics, we identified a set of simple sequence repeats (SSR) that can be used for quick and inexpensive genotyping of *A. officinalis* population samples.

2. Data description

2.1. Sampling and sequencing

We collected *Avicennia officinalis* L. viviparous propagules from Cilacap (Java Island, Indonesia). The propagules were first germinated in Petri dishes in the laboratory, and then cultivated in pots with nutrient soil for two weeks. Seedlings were transported and planted in a mangrove forest natural habitat in Dongzhai Harbor Mangrove Natural Reserve, Hainan, China. MIxS information is presented in Table 1.

After about one year of healthy growth in this natural environment, fresh leaves of *A. officinalis* were sampled and stored at −80 °C before RNA extraction. Total RNA was extracted using a modified CTAB method (Fu et al., 2004). Library preparation and sequencing of RNA samples were carried out at BGI (Beijing Genomics Institute), using the Illumina HiSeq 2000 platform. Sequencing was performed with an insert size of 200 base pairs (bp). The adaptor sequences were removed from the resulting data using an in-house Perl script.

2.2. De novo transcriptome assembly

We generated 36.24 million paired-end reads with a read length of 100 bp. FastQC v0.10.1 was used to evaluate read quality (Andrews, 2010). Raw paired-end reads were filtered using an in-house Perl script according to three criteria: average read score > 20; number of low quality sites (score < 10) in one read ≤ 20; missing data ('N') count per read ≤ 5. This filtering step excluded <1000 reads from the raw dataset (0.003%). The transcriptome of *A. officinalis* was first assembled using Trinity v2.0.6 using the high-quality sequencing data with default parameters (Grabherr et al., 2011). A total of 50,654 initial transcripts were produced from the initial assembly (Supplementary Table 1). We then filtered this set to remove weakly expressed and redundant transcripts, maximizing the confidence in our transcriptome. Realignment of the original short reads to the newly-assembled 50,654 transcripts was performed using the Burrows-Wheeler Aligner (BWA-v0.7.4-r385) (Li and Durbin, 2009). RPKM (Reads Per Kilobase per Million reads) values were calculated from the alignment results. By applying an RPKM cut-off of 1, transcripts with low expression were filtered out. We removed redundant transcripts using clustering tools implemented in CD-HIT with sequence identity threshold of 1.0 and word size of 5 ("-c 1.0 -n 5") (Fu et al., 2012). The final filtered assembly comprises 38,576 non-redundant transcripts (unigenes), with an average length of 834 bp and N50 of 606 bp (Supplementary Table 1). Approximately 28% of all the filtered transcripts were longer than 1 kb (Supplementary Table 1).

Table 1
Avicennia officinalis MIxS information.

Item	Description
Classification	Plantae; Angiosperms; Eudicots; Asterids; Lamiales; Acanthaceae; *Avicennia officinalis*
Investigation type	Eukaryote
Project name	*Avicennia officinalis* transcriptome
Geographic location	Cilacap; Java Island; Indonesia
Latitude, longitude	7°40′59.11″S, 108°49′42.91″E
Collector	Xinnian Li
Collection date	Nov-2012
Environment	Mangrove forest
Biome	ENVO: 01000181
Feature	ENVO: 00000316
Material	ENVO: 00002230
Plant height	5–10 m
Sequencing method	Illumina HiSeq 2000
Assembly method	Trinity 2.0.6
Finishing strategy	High quality transcriptome assembly

2.3. Functional annotation

First, all filtered transcripts were compared to the NCBI non-redundant (nr) protein database for functional annotation using BLASTX with an e-value cutoff of 1e-5 (Camacho et al., 2009). Of the 38,576 non-redundant transcripts, 22,254 (57.69%) showed significant sequence similarity to known proteins in the database (Supplementary Table 1). Genomes of eight Asterid species were the most frequent matches, as expected (Supplementary Fig. S1). More than half of the transcripts (17,274/22,254) matched proteins from *Sesamum indicum*, the most closely related species with available whole-genome data in the nr protein database (Supplementary Fig. S1). In addition, 2082 (9.4%) of the transcripts were annotated to *Erythranthe guttatus*, formerly known as *Mimulus guttatus* (Supplementary Fig. S1).

We then attempted to assign gene ontology (GO) terms to the transcripts with blast hits using Blast2go software (Conesa et al., 2005). All 22,254 transcripts were assigned to at least one of the 119 level-2 GO terms (Supplementary Table 2). Of the three level-1 ontologies, Molecular Function contained the largest number of categories (63), while Cellular Component was assigned 36 and Biological Process 20 categories (Supplementary Table 2). We further plotted the distribution of the 37 GO terms that contained >1% of all the 22,254 annotated transcripts (Fig. 1). The functional annotations can be used to further explore molecular mechanisms of salt excretion by salt glands, initially focusing on genes associated with transporter activity in the Molecular Function set (Fig. 1, Supplementary Table 3).

Given that another mangrove of *A. marina* was the closest species with transcriptome data of leaf tissue (Huang et al., 2014), we further identified the orthologous genes between this species and *A. officinalis* by conducting the reciprocal best hit Blast with an e-value cutoff of 1e-5. Approximately 60.17% (23,320/38,756) transcripts in *A. officinalis* were found orthologous in the transcriptome of *A. marina*. Of the 22,254 transcripts with functional annotations, only 12,951 (58.20%) orthologous genes were expressed in the leaf tissue of *A. officinalis*.

2.4. Identification of copies and isoforms of AoPIP, AoTIP, and AoDHN1

We used previously-published (Jyothi-Prakash et al., 2014; Tan et al., 2013) cDNA sequences of three known genes (*AoPIP*, *AoTIP* and *AoDHN1*) with potential functional roles in salt-gland-mediated salinity tolerance to search our set of 38,572 *A. officinalis* transcripts. We required that putative orthologs have BLAST *E*-values less than 1E-5, cover >30% of the original cDNA, and be at least half as long. This filtering left us with eight copies and isoforms of *AoPIP*, one copy but two isoforms of *AoTIP*, and only one copy of *AoDHN1* (Fig. 2, Supplementary Table 4). To get the full length of each *AoPIP* gene, two bioinformatic procedures were taken (1) comparing the *AoPIP* genes with their orthologs in *A. marina* (Huang et al., 2014); (2) concatenating the paired-reads with specific polymorphisms (SNPs and indels) in this region to re-assemble and prolong each uncomplete *AoPIP* sequences. Then, we got the complete sequences of each AoPIP genes (Supplementary Fig. 2).

Comparing with the mean (29.07) and median (8.32) RPKM value of all transcripts, the *AoDHN1* (4936.86) and both isoforms of *AoTIP* (296.54 of *AoTIP_I1* of, 305.02 of *AoTIP_I2*) are highly expressed in *A. officinalis* leaf tissue, supporting their potential role in salt-gland salinity tolerance (Supplementary Table 4). Among the *AoPIP* isoforms and homologs, copy 5 are the highest-expressed, suggesting that it

Please cite this article as: Lyu, H., et al., *De novo* assembly and annotation of the *Avicennia officinalis* L. transcriptome, Mar. Genomics (2017), http://dx.doi.org/10.1016/j.margen.2017.07.002