Marine Genomics 4 (2011) 129–136

Contents lists available at ScienceDirect

Marine Genomics

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

Transcriptome sequencing of a highly salt tolerant mangrove species *Sonneratia alba* using Illumina platform

Sufang Chen [a,1], Renchao Zhou [a,1], Yelin Huang [a], Meng Zhang [a], Guili Yang [a], Cairong Zhong [b], Suhua Shi [a,*]

[a] State Key Laboratory of Biocontrol and Guangdong Key Laboratory of Plant Resources, Sun Yat-sen University, Guangzhou 510275, China
[b] Hainan Dongzhai Harbor National Nature Reserve, Haikou 571129, China

ARTICLE INFO

Article history:
Received 24 January 2011
Received in revised form 19 March 2011
Accepted 27 March 2011

Keywords:
Mangroves
De novo assemble
Codon usage bias
Adaptation
High throughput sequencing

ABSTRACT

Mangroves are critical and threatened marine resources, yet few transcriptomic and genomic data are available in public databases. The transcriptome of a highly salt tolerant mangrove species, *Sonneratia alba*, was sequenced using the Illumina Genome Analyzer in this study. Over 15 million 75-bp paired-end reads were assembled into 30,628 unique sequences with an average length of 581 bp. Of them, 2358 SSRs were detected, with di-nucleotide repeats (59.2%) and tri-nucleotide repeats (37.7%) being the most common. Analysis of codon usage bias based on 20,945 coding sequences indicated that genes of *S. alba* were less biased than those of some microorganisms and *Drosophila* and that codon usage variation in *S. alba* was due primarily to compositional mutation bias, while translational selection has a relatively weak effect. Genome-wide gene ontology (GO) assignments showed that *S. alba* shared a similar GO slim classification with *Arabidopsis thaliana*. High percentages of sequences assigned to GO slim category 'mitochondrion' and four KEGG pathways, such as carbohydrates and secondary metabolites metabolism, may contribute to salt adaptation of *S. alba*. In addition, 1266 unique sequences matched to 273 known salt responsive genes (gene families) in other species were screened as candidates for salt tolerance of *S. alba*, and some of these genes showed fairly high coverage depth. At last, we identified four genes with signals of strong diversifying selection ($K_a/K_s > 1$) by comparing the transcriptome sequences of *S. alba* with 249 known ESTs from its congener *S. caseolaris*. This study demonstrated a successful application of the Illumina platform to de novo assembly of the transcriptome of a non-model organism. Abundant SSR markers, salt responsive genes and four genes with signature of natural selection obtained from *S. alba* provide abundant sequence sources for future genetic diversity, salt adaptation and speciation studies.

1. Introduction

Mangroves are trees that grow in intertidal salty environments and major contributors to the marine resources. Their root systems protect the reefs from terrestrial sediments and other forms of pollution, and provide attachment surfaces for various marine organisms (Tomlinson, 1986). Despite increasing awareness of their ecological importance, the molecular mechanisms underlying their adaptation to extremely high saline intertidal habitats remain elusive. One main limitation for further research at the genomic level is a lack of genomic and transcriptomic resources for these non-model organisms.

Recently, the development of novel high-throughput DNA sequencing methods has provided an opportunity to address this question by de novo assembly or mapping and quantification of transcriptomes (Wang et al., 2009b). However, particular attention has been paid to sequencing the transcriptomes of model species such as yeast (Wilhelm et al., 2008), *Arabidopsis thaliana* (Lister et al., 2008), mouse (Cloonan et al., 2008) and human (Morin et al., 2008), and few transcriptome studies have involved non-model organisms so far. As an early step in sequencing the transcriptomes of non-model species, these high-throughput sequencing technologies were used to sequence the transcriptome of *Pinus contorta* (Parchman et al., 2010), *Eucalyptus grandis* (Novaes et al., 2008), *Ipomoea batatas* (Wang et al., 2010), and two mangrove species, *Rhizophora mangle* and *Heritiera littoralis* (Dassanayake et al., 2009).

Sonneratia (Lythraceae sensu lato), a typical mangrove genus consisting of 5–6 diploid species and 4 interspecific hybrids (Duke, 1987; Zhou et al., 2005; Qiu et al., 2008), is widely distributed on tropical and subtropical coasts in the Indo-West Pacific region. Species in this genus differ in their habitats with respect to salinity (Duke et al., 1998). *S. alba*, the most widespread species of this genus and one of the most salt tolerant mangrove species, is always found in the low intertidal zones of downstream estuarine areas. It grows best in salinities ranging from 5 to 50% sea water (Ball and Pidsley, 1995). In contrast, *S. lanceolata*, a less salt tolerant congener, occurs only in the

* Corresponding author. Tel.: +86 20 84113677; fax: +86 20 34022356.
E-mail address: lssssh@mail.sysu.edu.cn (S. Shi).
[1] These authors contributed equally to this work.

1874-7787/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.margen.2011.03.005

upstream estuarine region. It grows best in salinities ranging from 0 to 5% sea water (Ball and Pidsley, 1995). The adaptation of *Sonneratia* species to their habitats with different salinities may play an important role in species diversification within the genus, which makes *Sonneratia* an ideal system for studying adaptation to saline intertidal habitats.

In this study, we sequenced the transcriptome of *S. alba* using the Illumina platform. After de novo assembly of the transcriptome, simple sequence repeats (SSRs) were identified and codon usage bias was analyzed. Based on known salt responsive genes, we searched their homologous sequences in *S. alba* and calculated their coverage depth. By comparing the transcriptome sequences of *S. alba* with the ESTs from *S. caseolaris*, genes under positive selection were identified and we wish to gain insights into the role of natural selection in adaptation of mangroves to their environments.

2. Materials and methods

2.1. Total RNA extraction, cDNA library construction and Illumina sequencing

Seeds of *S. alba* were collected from mature fruits in Qionghai, Hainan, China. After two months since seed germination, the seedlings were transferred to Hoagland's solution (Hoagland and Arnon, 1950) plus NaCl to a final concentration of 250 mM, under which *S. alba* grew best compared to other concentrations tested in our preliminary study (S. Chen, unpublished data). Four weeks later, the fresh roots of these seedlings were harvested and pooled, and total RNAs were isolated using a modified CTAB method (Fu et al., 2004). mRNAs were extracted from these total RNAs using Oligotex^{TM}-dT30 (TaKaRa, Dalian, China) and then fragmented ultrasonically and converted to double-stranded cDNAs. A nucleotide "A" was added at the 3′-end of cDNAs, and then adapters were ligated to both ends. The QIAquick Gel Extraction Kit was used to purify and collect cDNAs of approximately 215 bp in length. Each amplified molecule was then sequenced using Illumina GA technology to obtain short sequences of 75 bp from both ends.

2.2. De novo assembly of the transcriptome

Three short read assemblers, ABySS 1.2.1, Velvet 1.12 and Edena 2.1.1 (Hernandez et al., 2008; Zerbino and Birney, 2008; Birol et al., 2009; Simpson et al., 2009), were selected to de novo assemble the transcriptome of *S. alba*. These assemblers are all based on a de Bruijn graph representation of sequence neighborhoods, differing in how to treat errors, resolve ambiguities and use read-pair information (Flicek and Birney, 2009). The assembled sequences less than 200 bp in length were discarded due to a low annotation rate (Gotz et al., 2008).

Poor base qualities in the tail of the reads could have negative effects on the assembly. Therefore, we trimmed the reads from the end and obtained nine groups of reads with different lengths (51, 54, 57, 60, 63, 66, 69, 72 and 75 bp) and then de novo assembled the transcriptome using the three assemblers. The parameter k-mer was optimized and selected according to $k = \text{integ}\ (1/2L + 1)$, where L is the length of the reads. Assemblies from different length of reads were then compared to determine the effects of poor base qualities on the assembly. Additionally, the assemblies from different assemblers were also compared to each other using CD-hit 3.1.2 with a global sequence identity 0.99 (Li and Godzik, 2006). To obtain a database as complete as possible, assemblies from reads with different lengths (51–75 bp) by the three assemblers were pooled. CD-hit was also used to remove the redundancy, and CAP3 (Huang and Madan, 1999) was used to reassemble the sequences by setting the minimal overlap length as 100 bp and the minimal percent identity of the overlap as 99%.

2.3. Identifying SSR markers from the transcriptome of S. alba

Simple Sequence Repeats were detected in the 30,628 sequences of *S. alba* with MISA perl script (http://pgrc.ipk-gatersleben.de/misa). We defined the microsatellites here if the minimum number of repeats was larger than 5 for bi-nucleotide repeats, and larger than 4 for tri-, tetra-, penta- and hex-nucleotide repeats.

2.4. Measures of synonymous codon usage bias

Open reading frames (ORFs) were determined by searching against the nonredundant NCBI database using a cutoff value $1e^{-6}$ and the coding regions were extracted according to the top BLASTX hit with a custom perl script. Sequences encoded by chloroplast and mitochondrial genomes were recognized and removed by searching against the chloroplast and mitochondrial genome sequences of several plant species using a cutoff value $1e^{-6}$ (chloroplast genomes: *A. thaliana*, *Brassica napus*, *Beta vulgaris*, and *Nicotiana tabacum*; and mitochondrial genomes: *A. thaliana*, *Acorus calamus*, *Eucalyptus globulus*, and *Vitis vinifera*, http://megasun.bch.umontreal.ca/ogmp/projects/other/all_list.html). Relative synonymous codon usage (RSCU) (Sharp and Tuohy, 1986), GC content, GC content of the synonymous third codon position (GC_{3s}), and the effective number of codons (Nc) (Wright, 1990) were calculated using the program CodonW 1.4.2 (http://codonw.sourceforge.net/). The expected value of Nc (ENc) under the assumption that codon bias is only due to mutation bias was calculated using the following formula: $ENc = 2 + s\{29/[s^2 + (1-s)^2]\}$; where $s = GC_{3s}$. The Codon Adaptation Index (CAI) is always used to assess the extent to which selection has been effective in moulding the pattern of codon usage (Sharp and Li, 1987). With the EMBOSS 5.0.0 package (Rice et al., 2000), 200 genes encoding ribosomal proteins which were always highly expressed in most organisms, and 200 genes with the highest coverage depth were used as references to calculate CAI, respectively (we denote them CAI_1 and CAI_2). To avoid the influence of GC_{3s} on the calculation of CAI, we also calculated CAI (we denote it CAI_3 here) by selecting 200 reference genes with the highest ENc values (>59) and the lowest Nc values (<42). Correspondence analysis (COA) (Greenacre, 1984) was carried out on RSCU values using the program CodonW. The correlation analysis was done using the Spearman's rank correlation analysis method wrapped in the multianalysis software SPSS version 13.0 (http://spss.com). Genes of *S. alba* containing intron regions were downloaded from NCBI (GenBank accession numbers: GQ121922–GQ121992). Exons and introns were extracted with a custom perl script. GC_{3s} of the exons and GC content of the introns were calculated using CodonW.

2.5. Functional annotation

The gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) annotation for each sequence of *S. alba* were performed using the automatic annotation tool Blast2GO with a cutoff e-value $1e^{-6}$ (Conesa et al., 2005; Conesa and Gotz, 2008). GO slim classification of the transcriptome of *S. alba* was then compared with that of *A. thaliana* (www.arabidopsis.org). As genome-wide KEGG annotations for *A. thaliana* are not available in the TAIR database, we selected KEGG pathways of *Populus trichocarpa* for comparison with *S. alba* (http://genome.jgi-psf.org/cgi-bin/metapathways?db=Poptr1_1).

2.6. Identifying candidate salt responsive genes

Munns (2005) listed many candidate genes for salt tolerance, and classified them into three main functional groups. Genes in each group were described by a series of comprehensive reviews and updates (Rhodes et al., 2002; Zhu, 2002; Blokhina et al., 2003). According to these literature, sequences annotated as these salt