DE NOVO ASSEMBLY AND CHARACTERIZATION OF RHODODENDRON HYBRIDUM HORT. (ERICACEAE) GLOBAL TRANSCRIPTOME USING ILLUMINA SEQUENCING

SHIPING CHENG*, YUXIA ZONG, MINGHUI CHEN, JIANSHENG WANG, MENGJIE LIAO AND FEI LIU

Pingdingshan University, Pingdingshan, 467000, Henan Province, China
*Corresponding author’s email: shipingcheng@163.com

Abstract

Rhododendron hybridum Hort. (Ericaceae) is an important ornamental species with striking continuous flowering features. However, few genomic resources are currently available for the species; thus, breeding programs are blocked by a lack of genetic information. Here, we document our transcriptomic profiling of four different R. hybridum tissues using whole transcriptome shotgun sequencing (RNA-Seq) to gain insight on functional genes, and to isolate expressed sequence tag-simple sequence repeat (EST-SSR) markers for breeding and conservation purposes. In total 38,050,296 high-quality sequence reads were obtained, and 56,120 unigenes (contiguous sequence assemblies representing transcripts, with N50 = 1,236 bp) were assembled. Of these, 32,580 (58.05%) and 8,788 (15.66%) were annotated to the GO and KEGG databases, respectively. Additionally, 38,775 (69.09%) and 37,409 (66.66%) of the R. hybridum unigenes aligned to the Arabidopsis thaliana and Oryza sativa genomes, respectively. A total of 21,103 SSR motifs were identified in 15,050 of the unigenes. Among them, dinucleotide repeats account for the largest proportion (49.27%), followed by mono- (35.94%) and trinucleotide repeats (21.5%). This is the first comprehensive transcriptome dataset for R. hybridum, and RNA-Seq technology is proven to be useful approach for EST-SSR development. Such a vast quantity of sequence data and microsatellite markers should prove to be robust tools for subsequent genomic research and breeding programs in R. hybridum and related species.

Key words: Rhododendron hybridum; Transcriptome; Function annotation; EST-SSR.

Introduction

Rhododendron hybridum Hort. (Ericaceae) is highly valued for its ornamental and economic importance, and is widely distributed in the world. R. hybridum originated from natural hybridization within a bank of R. indicum, R. sinistii, and R. macrostomum, but the detailed breeding history is not clear (Huang, 1999). R. hybridum is well known for its very beautiful and colorful flowers with a variety of designs. The most striking distinction is continuous flowering, and the flowering time can be controlled. Previous studies of R. hybridum have focused primarily on its morphological characteristics and tissue culture (Zhang & Chen, 2005; Wu, 2007). However, very few reports exist on the diversity of germplasm resources and molecular markers for this species, except for ISSR analysis on the cultivars of Rhododendron hybridum (Zheng et al., 2011). This lack of genomic resource coupled with the species’ broad ornamental application, necessitates a need for the generation of genetic information. The total transcriptome, recognized as a valuable genetic resource, plays a vital role in functional marker development and other genomic studies, and in turn facilitates efficient breeding efforts.

Recent advances in bioinformatics and genetic technologies have generated a wave of transcriptome profiles for many organisms (Xiang et al., 2015). Transcriptomics is considered a powerful approach for identifying key genes for important traits in any tissue of interest (Cheng et al., 2017; Wilhelm and Landry, 2009). Transcriptomics can also assist in the isolation and characterization of molecular markers. Microsatellite (e.g. simple sequence repeat, SSR) has been widely used in population genetic studies, due to codominance and high polymorphism levels. However, the traditional methods for developing SSRs are costly and laborious. Recently, transcriptome mining has proven to be an effective method for developing SSRs, particularly for less well-studied species with little basic genetic information (Keeling et al., 2011; Magbanua et al., 2011; Ritland, 2012). A limited number of R. hybridum SSR markers have been developed (Chen et al., submitted paper), however, this dataset is too sparse for marker-assisted selective breeding purposes. To better address questions in population and evolutionary genetics studies, a large number of SSR markers are a prerequisite. With the development of next-generation sequencing technologies, many EST-SSRs have been found and evaluated in multiple species, including sweet potato (Wang et al., 2010), chickpea (Garg et al., 2011), Epimedium sagittatum (Zeng et al., 2010), and Siberian wildrye (Zhou et al., 2016). The results of these studies have shown that SSRs vary widely between different plant species.

We used the Illumina HiSeq X TEN (San Diego, CA, USA) NGS platform to construct a reference transcriptome of R. hybridum, in the present study. A normalized cDNA library was generated derived from flower, leaf, stem, and root of an adult individual. Then a large number of EST-SSR markers were developed based on the transcriptome information obtained. To our knowledge, this is the first comprehensive transcriptome of R. hybridum, and the large-scale sequence data and EST-SSR markers will be valuable for further population genetic studies and marker-assisted breeding programs in R. hybridum and other Rhododendron species.

Material and Methods

Plant materials and RNA extraction: Fresh root, leaf, flower, and stem tissues were collected from one adult R. hybridum tree located at the Baotianman National Nature Reserve (Henan, China). Total RNA was extracted from the four tissues using an RNeasy Kit (QIAGEN, Hilden, Germany) and quantified with an Agilent 2100 Bioanalyzer RNA Nanochip Kit (QIAGEN, Hilden, Germany). An equal amount of RNA from each sample

was pooled, and the mixed RNA (~10 μg) was subjected to Illumina HiSeq X TEN (San Diego, CA, USA) sequencing.

**RNA-seq for Illumina X TEN sequencing:** The whole transcriptome shotgun sequencing (RNA-seq) mRNA library was constructed using a mRNASeq Sample Preparation Kit (Illumina Inc., San Diego, CA, USA). Isolated Poly-(A) mRNA fragments were used to synthesize the first-strand cDNA and the second strand cDNA was synthesized using DNA polymerase I. The polymerase chain reaction (PCR) was performed to selectively enrich and amplify the cDNA fragments using the PCR primers PE 1.0 and PE 2.0. The paired-end library was synthesized using a Genomic Sample Prep Kit, following the manufacturer’s instructions. The 150–200 bp cDNA fragments were selected for downstream enrichment. The mixed cDNA library was then sequenced by pair-end on an Illumina HiSeq™ X TEN (Illumina, San Diego, CA, USA) at Novel Bioinformatics Ltd., Co., Ltd (Shanghai, China).

**De novo transcriptome assembly and annotation analysis:** FastQC Toolkit was used to evaluate the quality of the raw sequencing data. This evaluation metrics can help us better understand the nature of the data, before subsequent variant evaluation. By removing the adapter sequences, ambiguous and low-quality reads, clean reads were obtained for further analysis. Then Trinity software (Grabherr et al., 2011) was used to de novo assemble the high-quality transcriptome, and then unigenes were obtained. We then mapped the R. hybridum unigenes to the Arabidopsis thaliana and Oryza sativa genomes, to evaluate model genome homologies, using the genomic mapping and alignment program with default parameters (Wu and Watanabe, 2005). The unigenes were subsequently subjected to BlastX against the non-redundant (nr) protein database, and corresponding annotations were compiled from nr, as well as from appropriate entries from the protein family database (Pfam, http://pfam.xfam.org/). In addition, gene ontology (GO) terms were annotated using the Blast2GO program. Furthermore, the Kyoto Encyclopedia of Genes and Genomes (KEGG, http://www.genome.jp/kegg) was used for pathway assignments.

**EST-SSR isolation and marker development:** SSR mining was conducted from our reference unigenes using Micro SATellite (MISA, http://pgrc.ipkgs.tuleben.de/misa/) with six unit_size-min_repeate parameters: 1-10, 2-6, 3-6, 4-6, 5-6, 6-5. The SSR primer pair was designed using Primer3 (Rozen & Skaltsky, 2000) with default settings and a PCR product length ranging from 100 to 250 bp (Thiel et al., 2003). Primers (20 to 24 bp long) for amplification of SSR-containing fragments (100 to 400 bp long) were designed using Primer 3, with 56°C as the optimal annealing temperature. The newly developed markers were given unique names starting with RH, representing R. hybridum. The marker sequences contain SSR start and end sequences.

**Results**

**Sequence data and assembly:** Total RNA was extracted from four different tissues to investigate the *R. hybridum* transcriptome. A total of 45,597,914 raw reads, each 150 bp long, were obtained (NCBI SRA accession GSE97630). After a strict filtration step that removed adaptor and low-quality sequences, 38,050,296 clean reads, with 85.1% Q30 bases, and 49.14% GC content, were obtained for downstream analysis. These high-quality reads were assembled into 56,120 unigenes, with an average length of 904 bp. The median and N50 length were 549 and 1,236 bp, respectively, ranging from a minimum of 295 to a maximum of 112,240 bp.

**Functional annotation:** Unigene primary functions were ascertained with GO analysis and belonged into three primary categories: 86,605 to biological process, 76,487 to cellular component, and 32,678 to molecular function. These unigenes are further distributed 30 categories. In the biological process category, ‘metabolic processes’ is the most frequently assigned GO term, followed by ‘biological process’, ‘cellular processes’ and ‘biosynthetic process’. A large number of the unigenes are assigned to ‘nucleus’ and ‘membrane’, followed by ‘cytoplasm’ and ‘plasma membrane’, in the cellular component category (Fig. 1). These results suggested that a broad diversity of transcripts was sampled by pooling total RNA from four different tissues.

We aligned the unigenes to KEGG database discover relevant biological pathways, and got 8,788 (15.66%) produced significant matches and 125 KEGG pathways. Of these, 29.61% were classified in metabolism pathway systems, with most being involved in the biosynthesis of secondary metabolites and glycolysis. Genetic information processing pathways accounted for 20.71% of the 8,788, with most being involved with ‘ribosomes’, ‘protein processing’, ‘spliceosomes’, and ‘RNA transport’. Environmental information processing accounted for 19.59% of the total, and was subdivided into ‘plant hormone signal transduction’, ‘the phosphatydlinositol signaling system’, and ‘ABC transporters’. Cellular processes accounted for 18.64% of the total, and included ‘endocytosis’, ‘phagosomes’, and ‘peroxisomes’. Organismal systems, including ‘plant-pathogen interactions’, ‘circuitar rhythm’, and ‘mediated cytotoxicity’, accounted for 14.28% of the total (Fig. 2).

All 56,120 *R. hybridum* unigenes were filtered with MISA to explore for SSR markers. Six types of SSRs were found; the percentages of mono-, di-, tri-, tetra-, penta-, and hexanucleotides were 35.94, 49.27, 13.77, 0.53, 0.18, and 0.30%, respectively. Statistical analysis revealed that 15,050 of our unigenes contain SSRs, 4,544 of which contain more than one identical type of SSR. Of the 15,050 SSRs containing unigenes, 2,275 contain mixed combinations of several different types of SSRs (Table 1). The mononucleotide repeat frequency for A/T (35.42%) is significantly larger than for C/G (5.16%). The most abundant dinucleotide repeat motif is AG/CT, while the rarest is CG/GC. Of the 10 trinucleotide categories, the AAG/CTT motif is the most common.
ASSEMBLY AND CHARACTERIZATION OF GLOBAL TRANSCRIPTOME USING ILLUMINA SEQUENCING

Fig. 1. GO classification of the *R. hybridum* unigenes.

![GO-Analysis-Annotation](image)

**Biological Process**  
- Metabolic pathways  
- Biosynthesis of secondary metabolites  
- Carbon fixation in Photosynthesis  
- Posttranslational modification  
- Posttranslational modification of protein  
- RNA transport  
- Protein processing in endoplasmic reticulum  
- RNA degradation  
- RNA splicing  
- Spliceosomal RNA metabolism  
- Spliceosomal RNA metabolism  
- Spliceosomal RNA metabolism

**Molecular Function**  
- Biological process  
- Cell  
- Cellular component  
- Molecular function  
- Monolayer  
- Multicellular organismal process  
- Nucleic acid binding  
- Protein binding  
- Signal transduction  
- Signaling molecule binding  
- Small molecule binding  
- Transferase activity  
- Transferase activity

**Cellular Component**  
- Genetic information processing  
- Environmental information processing  
- Cellular processes  
- Organismal systems

Fig. 2. KEGG classification of *R. hybridum* unigenes.

![Log2 (DiffGene)](image)

**Table 1. SSRs identified in the *R. hybridum* transcriptome.**

<table>
<thead>
<tr>
<th>Number</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of sequences examined</td>
<td>56120</td>
</tr>
<tr>
<td>Total size of examined sequences</td>
<td>50772160</td>
</tr>
<tr>
<td>Total number of identified SSRs</td>
<td>21103</td>
</tr>
<tr>
<td>Number of SSR containing sequences</td>
<td>15050</td>
</tr>
<tr>
<td>Number of sequences containing more than 1 SSR</td>
<td>4544</td>
</tr>
<tr>
<td>Number of SSRs present in compound formation</td>
<td>2275</td>
</tr>
<tr>
<td>mono-</td>
<td>7585</td>
</tr>
<tr>
<td>di-</td>
<td>10398</td>
</tr>
<tr>
<td>tri-</td>
<td>2906</td>
</tr>
<tr>
<td>tetra-</td>
<td>112</td>
</tr>
<tr>
<td>penta-</td>
<td>38</td>
</tr>
<tr>
<td>hex-</td>
<td>64</td>
</tr>
<tr>
<td>total</td>
<td>21103</td>
</tr>
</tbody>
</table>

**Discussion**

High-throughput RNA-Seq is a widely proven and powerful approach for obtaining comprehensive transcripts from species of interest. Recently, increasing less-studied species have been sequenced using NGS technologies. We combined RNA NGS with advanced bioinformatics analyses to construct the *R. hybridum* transcriptome. *R. hybridum* is a widespread ornamental tree of great economic value, yet it severely lacks genomic resources. The vast quantity of transcriptome data that we have generated helps to fulfill this need in this species. And the availability of a large set of EST-SSR markers will be useful for population genetic studies and breeding programs of *R. hybridum*. 

---

*Fig. 1. GO classification of the *R. hybridum* unigenes.*

*Fig. 2. KEGG classification of *R. hybridum* unigenes.*
In the present study, a cDNA library containing four tissues RNA was successfully created and paired-end sequenced. After strict filtration, about 38.05 million clean reads were obtained and assembled into 56,120 unigenes, of ~50,772,160 bp total length. It is well known that the longer the N50 and the shorter the N90 suggests the better the quality of the transcriptome data. Here, the N50 was 1,236 and the N90 was 259, which is relatively better than many other published transcriptomes. Our results showed that the *R. hybridum* unigenes were generally longer than the average lengths of unigenes previously reported in other species, including *Colilia nasus* (580 bp) (Fang et al., 2015), *Eriocheir sinensis* (382 bp) (He et al., 2012), *Acropora millepora* (440 bp) (Meyer et al., 2009), *Pinus contorta* (500 bp) (Parchman et al., 2010), *Chlamydomonas spp.* (665 bp) (Kim et al., 2013), and *Camelina sinensis* (733 bp) (Wu et al., 2013). Comparisons of our *R. hybridum* transcriptome to other published transcriptomes show our results to be of high overall quality. These results will provide high quality sequence data for future gene cloning and transgenic engineering studies in Rhododendron.

32,580 unigenes were assigned GO terms belonging to three main categories: biological process, cellular component, and molecular function. In which 12,061 related to metabolic processes, including secondary metabolite biosynthesis, transport, and catabolism. A high number of unigenes also link to ‘response to stimulus’ and ‘signaling’, suggesting that *R. hybridum* can rapidly adapt to changes in the natural environment. These findings should prove important to future breeding programs.

The KEGG pathway annotation can help further elucidate the biological functions of genes and molecular interaction networks (Kanehisa et al., 2008). We found 8,788 of our unigenes grouped into 126 pathways. Representative KEGG pathways are shown in Fig. 2. The most frequent are ‘metabolic’, ‘biosynthesis of secondary metabolites’, ‘ribosome’, and ‘RNA transport’. Genes involved in ‘carbohydrate metabolism’, such as ‘starch and sucrose metabolism’ and ‘glycolysis/glucogenesis’, are found, which is consistent with adequate transcriptome sampling. We do note, though, that while the present study succeeded in providing functional annotation, more studies are still required for further validation. Overall, our results indicate that high-throughput RNA-Seq technology is a cost-effective method for transcriptome analysis in plants with limited basic genetic resources.

EST-SSRs have been widely proven to be highly efficient in molecular breeding and pedigree tracking studies. In the present study, NGS transcriptome data enabled the efficient development of EST-SSR markers. A total of 21,103 SSRs were found distributed in 15,050 unigenes, at an average density of one SSR per 2.41 kb. Previous studies have identified dinucleotide repeats to be the most abundant SSR, followed by tri- or mononucleotide repeats, in the majority of the dicotyledonous species. In our study, dinucleotide repeats (49.17% of all SSRs) were the most frequent motif and AG/CT (46.42% of all dinucleotide) was the most common one. These results are consistent with previous studies in peanut, sugar beet, and canola. While among trinucleotide repeats, AAG/CTT was the dominant motif (3.44% of all SSRs), which is consistent with results in *Ammopiptanthus mongolicus* (Wu et al., 2014). Among the 7,585 mononucleotide repeats, 7,476 were A/T motif accounting for 98.6%. A/T motif has been reported to provide a means of filling gaps in linkage maps constructed with higher order SSRs (Kumpatla & Mukhopadhyay, 2005). Obviously, the distribution and frequency of each SSR motif varies greatly depending on species (Toth et al., 2000). Moreover, the abundance of different SSRs depend on various factors, such as the size of analyzed dataset, the applied SSR identification criteria, and the applied mining tools (Varshney et al., 2005).

To our knowledge, this is the first comprehensive transcriptome dataset for *R. hybridum*. In this study, RNA-Seq technology is proven to be useful approach for EST-SSR development. The vast transcripts and large number of EST-SSR markers we generated will provide useful genomic information for population genetics and breeding programs for *R. hybridum* and closely related species.

Acknowledgments

This research was supported by Scientific Research Foundation for the introduction of talent of Pingdingshan University (No. PXYSQD2016009), Science and Technology Plan Projects of Henan Province (172102110111), Key Research Project of Colleges and Universities of Henan Province (16A220004).

References


ASSEMBLY AND CHARACTERIZATION OF GLOBAL TRANSCRIPTOME USING ILLUMINA SEQUENCING


(Received for publication 25 May 2017)