Transcriptome characterization and screening of molecular markers in ecologically important Himalayan species (Rhododendron arboreum)

Journal:	<i>Genome</i>
Manuscript ID	gen-2017-0143.R2
Manuscript Type:	Article
Date Submitted by the Author:	23-Jan-2018
Complete List of Authors:	Choudhary, Shruti; Central University of Punjab, Plant Sciences Thakur, Sapna; Central University of Punjab, Plant Sciences Najar, Raoof; Central University of Punjab, Plant Sciences Majeed, Aasim; Central University of Punjab, Plant Sciences Singh, Amandeep; Central University of Punjab, Plant Sciences Bhardwaj, Pankaj; Central University of Punjab, Plant Sciences
Is the invited manuscript for consideration in a Special Issue? :	N/A
Keyword:	de novo transcriptome, Rhododendron arboreum, microsatellite, SNP, annotation

SCHOLARONE™
Manuscripts

7 **¹Corresponding author:** Dr. Pankaj Bhardwaj, Asstt. Prof., Molecular Genetics Laboratory, Centre for Plant
8 Sciences, Central University of Punjab, City Campus, Mansa Road, Bathinda-
9 151001, India. pankajihbt@gmail.com, pankajbhardwaj@cup.edu.in

10 **Abstract**

11 *Rhododendron arboreum*, an ecologically prominent species, also lends commercial and medicinal
12 benefits in the form of palatable juices and useful herbal-drugs. The local abundance and survival of the species
13 under highly fluctuating climate make it an ideal model for genetic structure and functional analysis. However, a
14 lack of genomic data has hampered the additional research. In the present study, cDNA libraries from floral and
15 foliar tissues of the species were sequenced to provide a foundation for understanding the functional aspects of
16 the genome and to construct an enriched repository that will promote genomics studies in the genera. Illumina's
17 platform facilitated the generation of ~100 million high-quality paired-end reads. *De novo* assembly, clustering,
18 and filtering out shorter transcripts predicted 113,167 non-redundant transcripts with an average length of
19 1164.6 bases. 71,961 transcripts were categorized based on the functional annotations in the Gene Ontology
20 database, whereby 5,710 were grouped into 141 pathways and 23,746 encoded for different transcription factors.
21 Transcriptome screening further identified 35,419 microsatellite regions, of which, 43 polymorphic loci were
22 characterized on 30 genotypes. Seven hundred and nineteen transcripts had 811 high-quality single-nucleotide
23 polymorphic variants with a minimum coverage of 10, a total score of 20, and SNP% of 50.

24 **Keywords:** *Rhododendron arboreum*, *De novo* transcriptome, Microsatellite, SNP, Annotation

25 **Introduction**

26 A limited genetic information has confined the functional and population genomics, especially in the
27 non-model organisms. Sequencing and characterizing the transcriptome is a suitable approach in exploring the
28 genetic richness of a species that lacks a reference genome. The transcriptome is the depiction of gene-rich
29 slices of the genome, existing across tissues, cells, or a developmental stage. The next-generation sequencing
30 (NGS) technology has enhanced the magnitudes of fundamentals in the species, which are naïve of the genetic
31 resource availability (Ellegren 2014). Modern era’s sequencing platforms have allowed the generation of
32 massive sequence information at relatively lesser expenses. It has provided an all-around access to the molecular
33 dimensions facilitating gene elucidation, annotation, marker screening, and conservation genetics.

34 The Himalayas are among the longest altitude-temperature gradients. The ranges are home to roughly
35 two-thirds of the world’s *Rhododendron* species (subgenus *Hymenanthus*, family Ericaceae) (Singh et al. 2009).
36 The high-altitude flora has specific requirements for humidity, temperature, precipitation, and photoperiod
37 (Gugger et al. 2013; Parmesan and Hanley 2015; Komac et al. 2016). Abiotic factors and their interactions with
38 the genotype of an organism primarily affect its growth and distribution pattern. There are unique microclimates
39 and noticeable topographical gradients in the mountainous regions; even a minor variation in the climate could
40 prove vulnerable to the endemic biodiversity (Thuiller et al. 2005). The resultant upward shift in the snowline
41 has already altered the habitat perimeters and phenological characteristics of many species, counting
42 *Rhododendron* genus as well (Xu et al., 2009). A rise in annual temperature not only influences the population
43 structure of a single species but also threatens the ecosystem as a whole, necessitating a timely evaluation of the
44 associated risks (Mommott et al. 2007; Urban 2015).

45 *Rhododendron arboreum* (2n=26) is dispersed from India to China throughout the Himalayas,
46 inhabiting the widest temperature range (4.4-19.3°C) (Vetaas 2002). The genus *Rhododendron* is a source of
47 useful phytochemicals having potential health benefits (Jaiswal et al. 2012). In addition, the species holds
48 ecological, aesthetic, and commercial eminence in India (Singh et al. 2009; Kumar 2012). Due to the economic
49 and medicinal values, the plant is being exploited at a faster rate for harvesting flowers, leaves, and wood.
50 Furthermore, a decrease in land cover as a result of anthropogenic activities and global climate changes can
51 cause a decline in the diversity. Despite the varying environmental cues, *R. arboreum* exploited the earlier
52 trends in climatic conditions, which favored its dominance in the landscapes (Ranjitkar et al. 2013). The
53 inherent genetic variability is very well appreciated for the survival of a species in a specific environment.

Diversity assessment can be accomplished with the help of molecular markers, which can also be utilized to explain the differences in selection pressure on a population by interpreting the level of polymorphism in transcribed loci.

One of the frequently employed marker system in genetic diversity and genotype identification studies is the simple sequence repeat (SSR or microsatellite) marker. SSRs are 1-6 base pairs long tandem repeats present both in coding as well as noncoding regions, and are usually characterized by a high degree of polymorphism. The co-dominant inheritance, genome-wide distribution and multiallelic characteristics deliver a high information content at an increased rate of reproducibility (Metzgar et al. 2000; Morgante et al. 2002). These features have enhanced the utility of SSR markers in population genetics and in the characterization or management of plant genetic resources (Ekblom and Galindo 2011; Strickler et al. 2012). SSRs derived from expressed sequence tags (ESTs), also known as EST-SSRs, can act as functional markers once associated with a target trait (Varshney et al. 2005). Another preferred sequence-based marker in ecology and population studies is based on single-nucleotide polymorphism (SNP). SNP markers are co-dominant, bi-allelic, highly polymorphic, and reproducible in nature. Another advantage of SNP markers is their frequent occurrence across the genome (Project IRGS, 2005), making them more informative than other marker systems (Bielenberg et al. 2015). Additionally, a low-cost automation for genotyping with high multiplex ratio has facilitated to work with a large number of individuals (Singh et al. 2013).

The genome-wide screening of molecular markers from high-throughput sequencing data has played a significant role in studying the genetic basis of variability at a locus (Stinchcombe and Hoekstra 2007). Microsatellite markers have been earlier employed to assess genetic diversity in a threatened *Rhododendron* species (Bruni et al. 2012). We reported the development of polymorphic genomic SSRs in *R. arboreum* (Choudhary et al. 2014). Similarly, another study utilized random primers for diversity assessment in the species (Kuttapetty et al. 2014). The previous population genetic studies based on the tedious, expensive, and time-consuming marker isolation and genotyping procedures could engage a relatively smaller number of genetic loci. However, the accurate and precise estimate of genetic structure necessitates the implementation of more number of markers.

The plants growing at high altitude ranges, as *Rhododendron*, have a sound, but an unexplored genetic foundation, which offers them physiological benefits over other species. However, a negligible genetic data is a constraint to the genomics research in this Himalayan species. Here, we have applied the full-length cDNA

sequencing approach as a fast yet effective course to improve the genetic resource of the species. Flower and leaf tissues were collected from the temperate zone of the Indian Himalayan Region during the spring season. With the objectives of (i) constructing a *de novo* transcriptome assembly, (ii) providing functional and pathway annotations, (iii) creating a molecular marker database and (iv) validating a set of SSR loci, the present study supplements the genetic literature of the species. Our study would prove worthy of capturing the functional gene networks embedded in the species and would assist evolutionary or ecological studies in future.

Materials and Methods:

Plant material and nucleic acid extraction

Healthy flowers and leaves were collected from mature plants growing in their natural habitats in Jhatingri (31.945910°N 76.894685°E) and Ropru (31.726562°N 76.862830°E), Himachal Pradesh (HP), India during the month of February-March. The tissues were frozen in liquid nitrogen and stored at -80°C until use. RNA isolated from floral and leaf tissues using the Total RNA Isolation kit (Bangalore GeneiTM) was treated with DNase and stored until use. Nanodrop spectrophotometer and electrophoresis on 2% agarose gel under denaturing conditions assessed the quality and quantity of RNA. RNA samples of optimum quality and free from DNA contamination, were used for further processing. RNA integrity was also evaluated using Agilent RNA Bioanalyzer chip.

For SSR characterization, leaves were sampled from thirty individuals, ten from each of the three regions- HP, Kashmir (KA), and Uttarakhand (UK), India. Sampling from HP included the areas of Barot, Dharamshala, Jhatingri, Kotmorse, and Ropru; from KA covered Anantnag, Kulgam, Tchittergul, and Uri; and from UK comprised Dehradun, Gharwal, Uttarkashi, and Yamunotri. Genomic DNA was isolated from leaves using CTAB method (Doyle 1987) and later, purified using RNase treatment followed by phenol: chloroform: isoamyl extraction. The quality and quantity of DNA were determined by Nanodrop spectrophotometer and 0.8% agarose gel electrophoresis.

cDNA library preparation and sequencing

For paired-end sequencing on Illumina's NextSeq500 platform, cDNA libraries were prepared using the TruSeq RNA library preparation protocol. Briefly, the fragmented mRNA was reverse transcribed followed by second-strand cDNA synthesis, adapter ligation, and amplification. In summary, four cDNA libraries

prepared from two tissues (leaf-L, flower-F) of two regions- Jhatingri (L2 and F2) and Ropru (L4 and F4) were barcoded separately and sequenced.

De novo transcriptome assembly

Raw reads (fastq) generated from sequencing were trimmed off the adapter and low-quality (Phred score >10) bases using the Trimmomatic tool (Bolger et al. 2014). FastQC program determined the base quality check (QC), sequencing read length, and associated parameters. The pre-processed reads from each sample were individually assembled using Trinity (Grabherr et al. 2011) on a web server hosted at Indiana University (<https://galaxy.ncgas-trinity.indiana.edu/>). Trinity is capable of generating isoforms, which arise as a result of alternate splicing or gene duplication events. Based on shared sequence content, several transcripts could be grouped into a single cluster, each of which, was denoted as a 'Trinity gene' (Haas et al. 2013). All the transcripts collected from the four libraries were concatenated to construct a master or reference assembly. CD-HIT-EST algorithm clustered the nucleotide dataset (Li and Godzik 2006) and removed identical sequences based on 95% sequence homology. This step grouped the transcripts assembled from the four libraries to obtain a non-redundant dataset and the sequences, which could not be extended further were finally referred as unigenes. Only those unigenes with >500bp length were kept for downstream analysis.

For the quality of assembly, the parameters assessed were: (i) Nx statistics (Nx is the minimum length of contigs represented by at least x% of the assembled bases), (ii) prediction of full-length transcripts, and (iii) evaluating the read content of the assembled data by the Bowtie algorithm (Langmead et al. 2009). N50 and N10 values indicated that 50% or 10% of the assembled transcript nucleotides are found in the contigs, which are of $\geq$ N50 or N10 length, respectively (Haas et al. 2013). Secondly, the number of full-length genes were estimated to examine the extent of genetic composition and transcript occurrence by aligning the assembly against a reference data set. Such an index based on a reference is more appropriate and is preferred over N50 stats to determine the quality of a transcriptome. Since no annotated reference is available for *R. arboreum*, BLAST search against Swiss-Prot database served for full-length transcript analysis. An assessment of the completeness of assembly and annotation was implemented using BUSCO v2 (Benchmarking Universal Single-Copy Orthologs; Simão et al. 2015). The tool analyzes the assembly relative to the already available transcriptomes of a lineage and searches for a list of conserved orthologs. The plant lineage datasets were utilized for the present study. Only the longest isoform of the 'Trinity gene' was kept for the analysis as suggested for transcriptome assemblies.

139 *Unigene Annotation*

140 The open reading frames (ORFs) were predicted from unigenes by TransDecoder v2.1 (Haas and
141 Papanicolaou 2012) to ascertain the coding regions within the transcripts. The nucleotide, as well as
142 TransDecoder-predicted peptide sequences, were aligned using standalone BLAST+ v2.4 tool, against Swiss-
143 Prot/UniProt, nt (non-redundant nucleotide database), and nr (non-redundant protein) database at the E-value
144 cutoff of 1.00E-05. The peptide sequences were searched for protein family domains against the Pfam database
145 by HMMER (Finn et al. 2011). The occurrence and site of signal peptide cleavage and transmembrane regions
146 were predicted by signal v4.1 (Petersen et al. 2011) and tmhmm server v2.0 (Krogh et al. 2001), respectively.

147 After primary annotations, gene ontology (GO) and pathway numbers were assigned by GO database
148 (Ashburner et al. 2000) and KAAS (KEGG Automatic Annotation Server) (Kanehisa et al. 2011), respectively.
149 GO system classified the genes and their products to assign equivalent terms based on their ontologies under
150 biological process, molecular function, and cellular component categories. The transcripts with similar functions
151 were categorized under a single group. *Arabidopsis lyrata*, *Arabidopsis thaliana*, *Citrus sinensis*, *Cucumis*
152 *sativus*, *Fragaria vesca*, *Glycine max*, *Oryza sativa japonica*, *Solanum lycopersicum*, *Theobroma cacao*, and
153 *Vitis vinifera* were the reference for pathway analysis using KAAS server (<http://www.genome.jp/tools/kaas/>).
154 Unigenes with cutoff values of 1.00E-05 were assigned the enzyme commission (EC) numbers, based on which,
155 KEGG mapped the unigenes to a biochemical pathway with manually curated ortholog groups (KEGG genes)
156 and assigned the KO identifiers. Additionally, the transcription factors (TFs) encoded by the assembled
157 transcripts of *R. arboreum* were identified by aligning against the Plant Transcription Factor Database (TFDB)
158 v3.0 at the cutoff of 1.00E-06 (Jin et al. 2013).

159 *Molecular marker prediction and characterization*

160 MicroSatellite Identification tool (MISA) was used to identify 1-6 nucleotide long SSR motifs repeated
161 at least five times. Primer designing using Batch Primer3 was limited to di-nucleotide and tri-nucleotide repeats
162 with the expected product size range of 100-300 bp, primer length of 18-21 bp, melting temperature between 50-
163 70 °C, and GC content of 30-50%. Randomly selected 70 loci were characterized on thirty genotypes sampled
164 from KA, HP, and UK. Each amplification reaction of 20 µl comprised of template DNA (25 ng/µl), 1.5 U Taq
165 DNA polymerase (DSS-Takara), 1X Taq buffer (supplemented with 1 mM Tris pH 9.0, 50 mM KCl 0.01%
166 gelatin, and 1.5 mM MgCl₂), 2.5 mM of dNTP mix (Genei™), and 5 ng of primer pair. The thermal profile of

PCR was as follows: an initial denaturation for 3 min at 94°C; followed by 35 cycles of denaturation (94°C), annealing at a specific temperature, and elongation (72 °C) for 1 min each; and a final elongation of 8 min at 72°C. The amplified products were mixed with loading buffer, denatured at 95°C for 5 min, snap-cooled, and electrophoresed on 6% denaturing polyacrylamide gel. The bands were visualized by silver staining protocol (Creste et al. 2001) following a prior exposure to formaldehyde. A preliminary statistics including, polymorphic information content (PIC), number of alleles, effective number of alleles, range of allele length, observed (H_o)/expected (H_e) heterozygosity, and allele frequency were assessed using GENALEX 6 (Peakall and Smouse 2006). The homology search was performed using BLAST with a cutoff of 1.00E-05 and >50% identity against UniProt database entries.

Lasergene SeqMan Pro's (DNASTAR® Inc., v12.2.0.82) variant identifying feature enumerated the transcriptome-wide occurrence of SNPs. Since the 'Trinity unigenes' cannot be directly utilized as an input in SeqMan Pro, we followed an indirect approach. The trimmed high-quality reads obtained from four cDNA libraries (representing two tissues from two individuals) were first processed by the *de novo* assembly construction protocol of SeqMan Pro. Only non-redundant contigs were kept for further screening and were renamed as per their sequence similarity with the Trinity transcripts by a script developed in-house. SNP Discovery Parameter for the neighborhood quality score was set as per Altshuler et al. 2000. Keeping the 'Neighborhood Window' of 5, five bases up and downstream of the SNP base were considered to obtain a 'minimum neighborhood score' of 20. The putative SNP was rejected if the specified window contains one or more mismatches with respect to the reference sequence. Depth or number of reads containing the SNP in an aligned column was kept at 4. SNP%, the proportion of the most prevalent non-reference base in the aligned column, was also taken as a criterion to remove the non-significant variants. Following the above parameters, SNPs with a minimum depth of 4 and with the SNP% of 25 were obtained. To further increase the stringency and to rule out sequencing errors, the variants were filtered at minimum depth and SNP% of 10 and 50, respectively.

Results

RNA Sequencing and de novo Assembly

The RNA isolated from flower and leaf samples of *R. arboreum* collected from two different locations was sequenced. The four cDNA libraries delivered a total of 105.3 million paired-end reads (sample-wise details

195 given in Table1). Since the genome is not available, the pre-processed reads from each sample were assembled
196 by the *de novo* approach of Trinity pipeline with the K-mer value of 25; statistics are summarized in Table 1.
197 Then, clustering (at 95% identity) with CD-HIT-EST and filtering of shorter transcripts (<500bp) subsequently
198 generated a total of 115,672 non-redundant unigenes with a mean length of around 1,164 bp and a maximum
199 transcript length of 15,900 bp. The N50 value was 1,387 bases (Table 2) concluding that half of the assembled
200 nucleotides were found in contigs having that much length. Figure 1 displays the length distribution graph for
201 the transcripts of final assembly.

202 Another parameter to assess the transcriptome quality is RNA-seq read representation, where short
203 reads were mapped to the transcripts and were segregated based on the proper-improper pairing. Bowtie's
204 comprehensive capturing of alignments estimated that ~81% of the assembled reads mapped back to the
205 assembly (gen-2017-0143.R2Suppla). For demonstrating the full-length transcripts, BLAST hits against Swiss-
206 Prot database were assessed. Percent length coverage distribution displayed that at least 9,156 proteins matched
207 to the corresponding transcript by >80% of their lengths (gen-2017-0143.R2Supplb). The BUSCO analysis
208 examined the completeness of the assembly based on the orthologous gene content available in plants. Of the
209 1,440 single copy Arabidopsis records, the present transcriptome, with 89,885 genes, was expected to be 82.9%
210 complete (874 full single-copy and 319 duplicated BUSCOs) whereas, 6.6% and 10.5% transcripts were found
211 fragmented (95) and missing (152), respectively (gen-2017-0143.R2Supple).

212 *Annotation and functional classification of unigenes*

213 TransDecoder predicted the likely coding regions to extract >100 amino acid long ORFs. To enhance
214 the sensitivity of detecting biologically significant ORFs, BLAST and Pfam homologies were kept as the
215 retention criteria. All the unigenes were translated to obtain a total of 96,872 peptides, with around 1.6%
216 transcripts contributing to more than one peptide. 55%, 56%, 63%, and 38% unigenes showed identical matches
217 in Swiss-Prot, nr, UniProt, and nt databases (of Viridiplantae Kingdom), respectively (Table 3). The protein
218 domains, signal peptide cleavage sites, and transmembrane helices were identified in 1%, 4.3%, and 26%
219 unigenes, respectively. The full-length proteins retrieved from alignment against Swiss-Prot exhibited the
220 highest sequence similarities to the known plant genes. A species distribution chart of the top BLAST hits
221 enumerated the efficiency of gene discovery by the assembly protocol. It indicated that 20% of the annotated *R.*
222 *arboreum* transcripts showed homology to the genera *Vitis*, followed by *Coffea*, *Solanum*, *Theobroma*, *Citrus*,

and *Prunus* species in decreasing order. 0.1% and 0.3% of the annotated transcripts were also found homologous to other species of *Rhododendron* and Ericaceae family, respectively (gen-2017-0143.R2Suppld).

Combining BLAST hits from all databases, we obtained GO terms for 71,961 unigenes. The remaining unannotated that did not match to the already known sequences may either be considered novel or a constituent of untranslated region or non-coding RNA. The functional annotations of unigenes from the sequence similarity search are enlisted in the supplementary information (gen-2017-0143.R2SupplII). GO terms were divided into three broad categories- biological processes (22.84%), molecular functions (50.8%), and cellular components (23.34%). The major sets within each of the three sub-categories are summarized in Figure 2. Under molecular function, *binding activity* (GO:0005488) was the most prominent category (48%), followed by *catalytic activity* (GO:0003824; 43%), *transporter* (GO:0005215; 4%), *structural molecule* (GO:0005198; 3%), and *translation regulation* (GO:0045182; 1%) (Figure 2A). Likewise, for the biological processes, *metabolic process* (GO:0008152; 39%), *cellular process* (GO:0009987; 31%), *localization* (GO:0051179; 12%), *cellular component organization or biogenesis* (GO:0071840; 7%), and *response to stimulus* (GO:0050896; 6%) were the top categories (Figure 2B). Furthermore, under the broad category of cellular components, *cell part* (GO:0044464) occupied the majority (44%), followed by *organelle* (GO:0043226; 26%), *macromolecular complex* (GO:0032991; 17%), and *membrane* (GO:0016020; 17%) (Figure 2C).

To ascertain the active metabolic pathways in floral and leaf transcriptome, KEGG pathways were assigned to the unigenes having GO terms (Figure 3). KO Ids were obtained for 5,710 transcripts, which encoded for the enzymes of 141 diverse pathways falling under six major groups (gen-2017-0143.R2Supple). Besides ribosome-associated pathways, transcripts related to spliceosome complexes, carbohydrate metabolism, amino acid biosynthesis, protein processing, plant hormone signaling, RNA transport, purine/pyrimidine metabolism, etc. were found.

The transcripts assembled from RNA-seq data of *R. arboreum* were also aligned against transcription factors (TFs) from 17 monocot and 49 eudicot species in the plant transcription factor database (TFDB). TFs are proteins which bind to DNA in a sequence-specific manner to regulate the transcription of corresponding genes. The plant TFDB, hosted by Centre for Bioinformatics (Peking University, China; <http://planttfdb.cbi.pku.edu.cn/>), is a web-resource of such factors identified from different species of green plants. Around 21,488 peptides translated from 20,513 transcripts (18% of the total) were found to encode for 3,292 unique TFs, which belonged to 59 different families (gen-2017-0143.R2Supplf). The basic/helix-loop-

helix (bhlh) was the most represented family followed by MYB-related, NAC, WRKY, FAR1, and B3 families, which correspond to myeloblastosis, NAM/ATAF/CUC (no apical meristem, Arabidopsis transcription activation factor, cup-shaped cotyledon), W-box cis-element, far-red impaired response, and ARF/ABI3/RAV (Auxin Response Factors, Abscissic acid Insensitive3, Related to ABI3/VP1) related domains, respectively.

Microsatellite prediction and identification

SSR-containing sequences were mined using MISA to identify 1-6 nucleotide long motifs having ≥ 5 contiguous repeats. Of 113,167 unigenes (>500 bp length) kept for SSR identification, 27,333 (24%) sequences contained SSR region and 6,423 (5.7%) had more than one motif (Table 4). Among 35,419 repeats located, dinucleotide repeats dominated, accounting to 19,425 (54.8%), followed by mono (10,215; 28.8%), tri- (5,529; 15.6%), tetra- (183; 0.5%), penta (39; 0.1%), and hexanucleotide (28; 0.07%) repeats and 2,641 were compound SSRs. For the dimer repeat, AG/TC (90.4%) was the most frequent type followed by AC/GT (6.5%) and AT (3.1%) repeats. Among the trinucleotides, AAG/CTT dominated, followed by AGG/CCT, ACC/TTG, CTC/GAG, and so on (details not shown).

From the 70 loci (only di- and tri-repeats) selected for experimental validation, 51 (73%) showed adequate amplifications, of which, 43 polymorphic loci were further used for variability analysis among thirty genotypes representing the three Indian states: HP, UK, and KA. Of the total 201 alleles (including null alleles) amplified by all the loci, the allele number varied from 3 to 9 with the average of 5 alleles per locus. Amplification profiles for selected loci are shown in supplementary data (gen-2017-0143.R2Supplg). H_o ranged from 0.000 to 0.800 (average: 0.364) and was significantly lower than H_e which ranged from 0.380-0.798 (mean: 0.650). All the loci showed significant deviations from Hardy-Weinberg equilibrium (HWE). Allele number and its frequency within a population are used to determine Polymorphic Information Content (PIC), which act as a criterion to assess the usefulness of a marker in revealing polymorphism. The PIC value for the present set of loci varied from 0.343 to 0.773 with an average of 0.6.

Homology searches using BLAST against protein database classified the loci according to their functions. The proteins encoded by these unigenes included exoribonuclease, 3-ketoacyl-CoA synthase, aldehyde dehydrogenase, arginine decarboxylase, auxin-responsive protein, ascorbate peroxidase, glycosyltransferase, flavonoid hydroxylase, geranyl diphosphate synthase, microsomal oleate desaturase, peptidyl-prolyl cis-trans isomerase, and shikimate kinase. GO terms indicated their role in transcription;

regulation of cell cycle, cell division, cell differentiation, and flower development; photomorphogenesis; protein ubiquitination and folding; response to freezing, auxin, and stress; meristem transition from vegetative to reproductive phase; and transport (refer to gen-2017-0143.R2SupplII). Table 5 enlists primer sequence and other locus characteristics including BLAST hits reported for these transcripts.

Screening SNPs in the transcriptome

The transcripts from four cDNA libraries (prepared from flower and leaf of two individuals) were reconstructed *de novo* using Lasergene SeqMan Pro (DNASTAR® Inc.). The reconstructed contigs that were homologous to ‘Trinity unigenes’ were concatenated to generate a reference and were compared with the pre-processed reads from each sample for SNP detection. The supplementary information (gen-2017-0143.R2SupplIII) summarizes different parameters of a variant including, SNP position, reference and called base, SNP%, depth, SNP type, and probable function. We located 36,921 SNPs across 7,518 contigs with a minimum SNP% and depth of 25 (average: 35.5%) and 4 (average: 67.6), respectively (gen-2017-0143.R2SupplIII). The variants were categorized into indels and SNPs, which were further classified into transitions and transversions. Base transitions (purine to purine or pyrimidine to pyrimidine) were the highest SNP type (60.2%), followed by transversions (32.5%), indels (7.2%), and a minor proportion of multi-allelic forms. Stringent filtering at a minimum depth of 10 (average: 27.3) with SNP% of 50 and allowing only a strict match in the neighborhood bases at a minimum score of 20, yielded 811 high-quality SNPs in 719 contigs (~1 SNP per contig). This included 55.2% transitions, 33.8% transversions, and 9.2% indels (Table 6).

Discussion

Paired-end sequencing and de novo assembly

The *de novo* assembly of short-read sequences aimed to provide a reference transcriptome, incorporating the datasets from leaf and floral tissues of *R. arboreum*. The unigenes reconstructed from over 100 million high-quality reads had a mean length of 1,164 bp with a good proportion being >500 bp in length. A considerable percentage of reads re-aligned back to the assembly as evident from 80% of the reads contributing towards contig formation by the Bowtie analysis. All the above facts along with the amplification rate of >70% displayed in case of microsatellite screening indicated the accuracy of the assembling strategy. An N50 value of 1,387 bp and the supplementary stats reported for the existing transcriptome were comparable to those obtained

using trinity in other non-model species (Chen et al. 2015; Zhang et al. 2015; González et al. 2016; Li et al. 2016).

As genomic data on *R. arboreum* is scarce, the overall transcriptome coverage was assessed by homology searches against the already available protein databases. Here, each transcript received a single best match and the alignment length across its top hit was calculated. As concluded from gen-2017-0143.R2Supplb, 10% of the entire transcriptome had alignment lengths of >70% to their respective proteins, which was rather low. Based on the full-length transcript count, we concluded that the transcriptome depth could be upgraded further by a larger scale sequencing effort. Another level of assessment using BUSCO with plant orthologs revealed that the existing unigene set was nearly 83% complete. The results were in agreement with the Trinity reconstituted *de novo* assemblies of other plants (Blande et al. 2017; Babineau et al. 2017). Overall the first transcriptome assembly of *R. arboreum* can form a sound basis to support future research on the species.

Annotations

The second objective of the study was a functional classification of the transcriptome, which was accomplished by following sequence similarity searches against the known protein and nucleotide databases. From the ORFs identified in 78,810 unigenes, 57,352 (72%) exhibited alignments against Swiss-Prot database. Similarly, more than half of the unigenes aligned to the nr database. The sequence search statistics, as well as the species distribution of the top BLAST hits, in our study, followed a similar trend as in *Vaccinium macrocarpon* (Sun et al. 2015). A high proportion of contig set showing homology to the already known proteins indicates the quality of the assembly algorithm followed. Among the annotated transcripts, only 0.2% contigs demonstrated homology with the isogenic or close relatives, reflecting a limited genomic data availability in the Ericaceae family.

GO terms reported for around 63% of the assembled transcripts were categorized at the levels of biological process, cellular component, and molecular function. The remaining unannotated transcripts would likely correspond to non-coding RNAs, sequences with the unknown domain, untranslated regions, or species-specific genes. 'Binding activity' and 'catalytic activity', 'cellular process' and 'metabolic process', and 'organelle' and 'cell part' were among the major groups in the molecular function, biological process, and cellular component, respectively. Functional classifications described for the transcriptome of *R. arboreum* was congruent to the trend reported for chilling tolerant *Chorispora bungeana* (Zhao et al. 2012). Pathway analysis

supplemented the knowledge of gene functions and improved the understanding of a biological process. The pathway annotations allocated 5,710 unigenes to 141 metabolic pathways in the KEGG database (gen-2017-0143.R2Supple). Overall, the assignment of a substantial number of transcripts under different GO categories endorsed their broad functional diversity as well as confirmed the efficiency of Illumina's platform. Consistent with the accounts of individual *de novo* assemblies (Huang et al. 2012; Die and Rowland 2014; Bastías et al. 2016), the present results also depicted the implementation of bioinformatics tools in assigning biological significance to the RNA-seq data.

Molecular marker identification and characterization

NGS technology provided a rich sequence resource for the bulk development of genic microsatellites in *R. arboreum*. SSRs were classified by the type and length of the units (Table 4), mono (≥ 10), di (≥ 6), tri (≥ 5), tetra (≥ 5), penta (≥ 5), and hexa (≥ 6) nucleotide repeats. Excluding the mononucleotide repeats, dinucleotides followed by trinucleotides were the predominant type, where AG and AAG motif, respectively, held the largest share. The findings resembled the accounts in *V. macrocarpon* (Schlautman et al. 2015), *Pterospira andromedea* (Grubisha et al. 2014), and other plant genomes. SSR mining in the transcriptome exhibited that 24% of *R. arboreum* transcripts contained at least one microsatellite region. The results differed from *V. macrocarpon* and other cultivars of the species (Liu et al. 2014; Schlautman et al. 2015), which might be due to the smaller dataset and different SSR-selection conditions in the present study. Overall, the transcriptome sequencing can be recommended as a valuable resource for high-throughput microsatellite identification in *R. arboreum*.

Thirty genotypes representing three natural habitats of *R. arboreum* were selected to validate a set of microsatellite loci. Of the 70 primer pairs designed, 51 could generate high-quality amplicons. SSR amplification rate of 73% suggested a high quality of the assembled unigenes. Of these, 43 (61%) pairs displayed polymorphism and the amplified product was observed to be within the expected size range for nearly all the loci. Other parameters such as PIC value, N_a , N_e , H_o , and H_e were enumerated to evaluate the informativeness and efficiency of each marker. Following the scheme described by Schlautman et al. 2015, the PIC value >0.5 exhibited by the majority (88%) of loci emphasized their higher informative value and good resolving power (Table 5). We obtained 201 alleles with an average of 4.674 alleles per locus and average effective allele number of 3. In terms of these characteristics of the reported loci, our results differed from those

in *Vaccinium* species (Liu et al. 2014) due to different selection parameters such as origin, location and type of repeat, and number and nature of primer or populations which were chosen for the study.

The overall H_o of 0.3634 was found to be lesser than $H_e=0.650$, accompanied by significant departures from HWE ($p>0.05$) due to loss of heterozygosity for nearly all the locus (except two). Similar results were reported for the genic microsatellite loci characterized in *Prunus* species (Dettori et al. 2015), pine (Lesser et al. 2012), and other eukaryotic populations (Coppe et al. 2012; Liu et al. 2016). SSR density across the genome is affected by direct selection pressure and evolutionary constraints as a microsatellite mutation can be harmful for gene function (Ellis and Burke 2007; Kalia et al. 2011; Merritt et al. 2015). It should be noted here that the loci under study exhibited significant homology to distinct protein classes- enzymes and other regulatory factors. The involvement of these transcripts in fundamental processes and cellular responses could explain the lesser observed heterozygosity or allelic diversity that led to significant HW deviations. Such an event of lower diversity is expected for the transcriptome owing to its highly conserved nature than the rest of the genome. Functional annotations offered the representation of the SSR loci in vital physiological activities namely, response to cold, metal ion binding, protein folding, and flower development (Table 5 and gen-2017-0143.R2SupplI). Above all, the developed markers have displayed sufficient polymorphism to acquire further insights on functional genomics and population structure analysis in *R. arboreum*.

Enriching the record of molecular tools, SNPs and indels were also identified in the transcript set of *R. arboreum* with SeqMan Pro, which creates a reference before detecting variants in reconstructed contigs. The SNP calling for the existing contig set coincided with other reports, where no reference genome was previously available (Salazar et al. 2015). 36,921 putative SNPs with a minimum depth of 4, SNP% of $\geq 25\%$, and SNP score of 20 were obtained in 7,518 contigs. The variants were filtered to rule out base call errors in order to reduce the selection of false positives and a high assay failure probability. Increasing the coverage limits to the factor of 10 and those of SNP% to 50, we attained 811 SNPs among 719 contigs. The resulting SNPs were mainly biallelic with a high number of transitions, followed by transversions, indels, and a minor number of other allelic forms or multiple base substitutions. Our report agreed with the assertions on other tree species (Geraldes et al. 2011; Koepke et al. 2012; Cokus et al. 2015). Stringent selection and the presence of SNPs in the annotated transcripts have enhanced their significance by offering comprehensive analyses of molecular variation. Also, an estimation of their function will benefit the estimates of polymorphic signatures in a particular trait. To exemplify, the identification of variation in genic region, directed towards climate

responsiveness, has been demonstrated in ecologically important species (Miguel et al. 2015; González et al. 2016; Gugger et al. 2016; Sork et al. 2016; Xu et al. 2016). However, it requires further investigations to ascertain these facts in *R. arboreum*. With a bigger marker dataset, such studies can prove more conclusive in characterizing the variation at the population level and will assist genomics studies in *R. arboreum*.

Acknowledgements

The work is supported by University Grants Commission, India (BSR-UGC 30-13/2014) and RSM (CUPB/CC/14/OO/4507). The study presented here is a part of the project aiming for diversity and genetic structure analysis in *R. arboreum*. ST and SC want to thank Indian Council for Medical Research for the fellowship towards Ph.D. The authors also acknowledge Sh. Devi Singh Thakur and Gagan Sharma for additional support during sampling and Vicky Kumar for his suggestions with the scripts in filtering the SNPs. The authors are sincerely thankful to anonymous reviewers for their valuable comments that were very productive and contributed significantly in the further refining of this manuscript.

Authors' contribution

PB conceived the study and designed and organized all the experiments. SC and ST collected the samples for RNA isolation and library preparation, collated the sequence data, and performed SSR characterization in the lab. SC carried out the bioinformatics portion, compiled and analyzed the results, and wrote the manuscript. RAN, AM, and AS did the sampling for leaf tissues and supported nucleic acid isolation and marker characterization. PB and ST coordinated in improving the work with their valuable suggestions. All the authors have read and approved the final manuscript.

Data archiving statement

The raw sequence data of the libraries are available from the NCBI SRA with the accession No. SRR4449163, SRR4449164, SRR4449165 and SRR4449166 under the study with accession No.: SRP092027. The master transcript file is submitted to DRYAD database, and the reviewer's link will be made available when required.

References

Ashburner, M., Ball, C.A., Blake J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., and Eppig, J.T. 2000. Gene Ontology: tool for the unification of biology. *Nature Genet.* **25**: 25-29.

- Altshuler, D., Itshuler, D., Pollara, V.J., Cowles, C.R., and van Etten, W.J. 2000. A SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature*, **407**: 513-516.
- Babineau, M., Mahmood, K., Mathiassen, S.K., Kudsk, P., and Kristensen, M. 2017. *De novo* transcriptome assembly analysis of weed *Apera spica-venti* from seven tissues and growth stages. *BMC Genomics*, **18**:128.
- Bastías, A., Correa, F., Rojas, P., Almada, R., Muñoz, C., and Sagredo, B. 2016. Identification and Characterization of microsatellite loci in maqui (*Aristotelia chilensis* [Molina] Stunz.) using next-generation sequencing (NGS). *PloS one*, **11**: e0159825.
- Bielenberg, D.G., Rauh, B., Fan, S., Gasic, K., Abbott, A.G., Reighard, G.L., Okie, W.R., and Wells, C.E. 2015. Genotyping by sequencing for SNP-based linkage map construction and QTL analysis of chilling requirement and bloom date in peach [*Prunus persica* (L.) Batsch]. *PloS one*, **10**: e0139406.
- Blande, D., Halimaa, P., Tervahauta, A.I., Aarts, M.G.M., and Kärenlampi, S.O. 2017. *De novo* transcriptome assemblies of four accessions of the metal hyperaccumulator plant *Noccaea caerulescens*. *Sci. Data*, **4**: 160131.
- Bolger, A.M., Lohse, M., and Usadel, B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **btu170**.
- Bruni, I., de Mattia, F., Labra, M., Grassi, F., Fluch, S., Berenyi, M., and Ferrari, C. 2012. Genetic variability of relict *Rhododendron ferrugineum* L. populations in the Northern Apennines with some inferences for a conservation strategy. *Plant Biosystems*, **146**(1): 24-32.
- Chen, S.F., Li, M.W., Jing, H.J., Zhou, R.C., Yang, G.L., Wu, W., Fan, Q., and Liao, W.B. 2015. *De novo* transcriptome assembly in *Firmiana danxiaensis*, a tree species endemic to the Danxia Landform. *PloS one*, **10**: e0139373.
- Choudhary, S., Thakur, S., Saini, R.G., and Bhardwaj, P. 2014. Development and characterization of genomic microsatellite markers in *Rhododendron arboreum*. *Conserv. Genet. Res.* **6**(4): 937-940.
- Cokus, S.J., Gugger, P.F., and Sork, V.L. 2015. Evolutionary insights from *de novo* transcriptome assembly and SNP discovery in California white oaks. *BMC Genomics*, **16**(1): 552.

- 443 Coppe, A., Bortoluzzi, S., Murari, G., Marino, I.A.M.M., Zane, L., and Papetti, C. 2012. Sequencing and
444 characterization of striped venus transcriptome expand resources for clam fishery genetics. PloS one, 7(9):
445 e44185.
- 446 Creste, S., Neto, A.T., and Figueira, A. 2001. Detection of single sequence repeat polymorphisms in denaturing
447 polyacrylamide sequencing gels by silver staining. Plant Mol. Biol. Rep. 19: 299.
- 448 Dettori, M.T., Micali, S., Giovinnazzi, J., Scalabrin, S., Verde, I., and Cipriani, G. 2015. Mining microsatellites
449 in the peach genome: development of new long-core SSR markers for genetic analyses in five Prunus
450 species. SpringerPlus, 4: 337.
- 451 Die, J.V., and Rowland, L.J. 2014. Elucidating cold acclimation pathway in blueberry by transcriptome
452 profiling. Environ. Exper. Bot. 106: 87-98.
- 453 Doyle, J.J. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem. Bull. 19:
454 11-15.
- 455 Ekblom, R., and Galindo, J. 2011. Applications of next generation sequencing in molecular ecology of non-
456 model organisms. Heredity, 107: 1-15.
- 457 Ellegren, H. 2014. Genome sequencing and population genomics in non-model organisms. Trends Ecol. Evolut.
458 29: 51-63.
- 459 Ellis, J.R., and Burke, J.M. 2007. EST-SSRs as a resource for population genetic analyses. Heredity, 99:125-
460 132.
- 461 Finn, R.D., Clements, J., and Eddy, S.R. 2011. HMMER web server: interactive sequence similarity searching.
462 Nucleic Acids Res. gkr367.
- 463 Geraldes, A., Pang, J., Thiessen, N., Cezard, T., Moore, R., Zhao, Y., Tam, A., Wang S., Friedmann, M., Jones,
464 S.J.M., Cronk, Q.C.B., Douglas, C.J., and Birol, I. 2011. SNP discovery in black cottonwood (*Populus*
465 *trichocarpa*) by population transcriptome resequencing. Mol. Ecol. Res. 11(s1): 81-92.
- 466 González, M., Maldonado, J., Salazar, E., Silva, H., and Carrasco, B. 2016. *De novo* transcriptome assembly of
467 'Angeleno' and 'Lamoon' Japanese plum cultivars (*Prunus salicina*). Genomics Data, 9: 35-36.

- 468 Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L.,
469 Raychowdhury, R., and Zeng, Q. 2011. Full-length transcriptome assembly from RNA-Seq data without a
470 reference genome. *Nature Biotechnol.* **29**: 644-652.
- 471 Grubisha, L.C., Nelson, B.A., Dowie, N.J., Miller, S.L., and Klooster, M.R. 2014. Characterization of
472 microsatellite markers for pinedrops, *Pterospira andromedea* (Ericaceae), from Illumina MiSeq sequencing.
473 *App. Plant Sci.* **2**(11):1400072.
- 474 Gugger, P.F., Fitz-Gibbon, S., Pellegrini, M., and Sork, V.L. 2016. Species-wide patterns of DNA methylation
475 variation in *Quercus lobata* and its association with climate gradients. *Mol. Ecol.* **25**: 1665-1680.
- 476 Gugger, P.F., Ikegami, M., and Sork, V.L. 2013. Influence of late Quaternary climate change on present patterns
477 of genetic variation in valley oak, *Quercus lobata* Née. *Mol. Ecol.* **22**: 3598-3612.
- 478 Haas, B., and Papanicolaou, A. 2012. TransDecoder (Find Coding Regions within Transcripts).
- 479 Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li,
480 B., Lieber, M., Macmanes, M.D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R.,
481 William, T., Dewey, C.N., Henschel, R., Leduc, R.D., Friedman, N., and Regev, A. 2013. *De novo* transcript
482 sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat.*
483 *Protoc.* **8**(8): 1494-512.
- 484 Huang, J., Lu, X., Yan, H., Chen, S., Zhang, W., Huang, R., and Zheng, Y. 2012. Transcriptome
485 characterization and sequencing-based identification of salt-responsive genes in *Millettia pinnata*, a semi-
486 mangrove plant. *DNA Res.* **19**: 195-207.
- 487 Jaiswal, R., Jayasingheb, L., and Kuhnerta, N. 2012. Identification and characterization of proanthocyanidins of
488 16 members of the *Rhododendron* genus (Ericaceae) by tandem LC–MS. *J. Mass Spectrom.* **47**(4): 502-515.
- 489 Jin, J., Zhang, H., Kong, L., Gao, G., and Luo, J. 2013. PlantTFDB 3.0: a portal for the functional and
490 evolutionary study of plant transcription factors. *Nucleic Acids Res.* **gkt1016**.
- 491 Kalia, R.K., Rai, M.K., Kalia, S., Singh, R., and Dhawan, A.K. 2011. Microsatellite markers: an overview of the
492 recent progress in plants. *Euphytica*, **177**: 309–334.

- 493 Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. 2011. KEGG for integration and interpretation
494 of large-scale molecular data sets. *Nucleic Acids Res.* **gkr988**.
- 495 Koepke, T., Schaeffer, S., Krishnan, V., Jiwan, D., Harper, A., Whiting, M., Oraguzie, N., and Dhingra, A.
496 2012. Rapid gene-based SNP and haplotype marker development in non-model eukaryotes using 3'UTR
497 sequencing. *BMC Genomics*, **13**: 18.
- 498 Komac, B., Esteban, P., Trapero, L., and Caritg, R. 2016. Modelization of the current and future habitat
499 suitability of *Rhododendron ferrugineum* using potential snow accumulation. *PloS one*, **11**: e0147324.
- 500 Krogh, A., Larsson, B., von Heijne, G., Sonnhammer, E.L. 2001. Predicting transmembrane protein topology
501 with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305**: 567-580.
- 502 Kumar, P. 2012. Assessment of impact of climate change on Rhododendrons in Sikkim Himalayas using
503 Maxent modelling: limitations and challenges. *Biodivers. Conserv.* **21**: 1251-1266.
- 504 Kuttapetty, M., Pillai, P., Varghese, R., and Seeni, S. 2014. Genetic diversity analysis in disjunct populations of
505 *Rhododendron arboreum* from the temperate and tropical forests of Indian subcontinent corroborate Satpura
506 hypothesis of species migration. *Biologia*, **69**(3): 311-322.
- 507 Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. 2009. Ultrafast and memory-efficient alignment of
508 short DNA sequences to the human genome. *Genome Biol.* **10**(3).
- 509 Lesser, M.R., Parchman, T.L., and Buerkle, C.A. 2012. Cross-species transferability of SSR loci developed
510 from transcriptome sequencing in lodgepole pine. *Mol. Ecol. Res.* **12**(3): 448-455.
- 511 Li, M., Dong, X., Peng, J., Xu, W., Ren R., Liu, J., Cao, F., and Liu, Z. 2016. *De novo* transcriptome sequencing
512 and gene expression analysis reveal potential mechanisms of seed abortion in dove tree (*Davidia involucrata*
513 Baill.). *BMC Plant Biol.* **16**: 82.
- 514 Li, W., and Godzik, A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or
515 nucleotide sequences. *Bioinformatics*, **22**: 1658-1659.

- 516 Liu, F., Hu, Z., Liu, W., Li, J., Wang, W., Liang, Z., Wang, F., and Sunb, X. 2016. Distribution, function and
517 evolution characterization of microsatellite in *Sargassum thunbergii* (Fucales, Phaeophyta) transcriptome
518 and their application in marker development. *Sci. Rep.* **6**: 18947.
- 519 Liu, Y.C., Liu, S., Liu, D.C., Wei, Y.X., Liu, C., Yang, Y.M., Tao, C.G., and Liu, W.S. 2014. Exploiting EST
520 databases for the development and characterization of EST-SSR markers in blueberry (*Vaccinium*) and their
521 cross-species transferability in *Vaccinium* species. *Sci. Hort.* **176**: 319-329.
- 522 Memmott, J., Craze, P.G., Waser, N.M., and Price, M.V. 2007. Global warming and the disruption of plant-
523 pollinator interactions. *Ecol. Lett.* **10**: 710-717.
- 524 Merritt, B.J., Culley, T.M., Avanesyan, A., Stokes, R., and Brzyski, J. 2015. An empirical review:
525 characteristics of plant microsatellite markers that confer higher levels of genetic variation. *App. Plant Sci.*
526 **3**(8):1500025.
- 527 Metzgar, D., Bytof, J., and Wills, C. 2000. Selection against frameshift mutations limits microsatellite
528 expansion in coding DNA. *Genome Res.* **10**: 72-80.
- 529 Miguel, A., de Vega-Bartol, J., Marum, L., Chaves, I., Santo, T., Leitão, J., Varela, M.C., and Miguel, C.M.
530 2015. Characterization of the cork oak transcriptome dynamics during acorn development. *BMC Plant Biol.*
531 **15**: 158.
- 532 Morgante, M., Hanafey, M., and Powell, W. 2002. Microsatellites are preferentially associate with nonrepetitive
533 DNA in plant genomes. *Nature Genet.* **30**: 194-200.
- 534 Parmesan, C., and Hanley, M.E. 2015. Plants and climate change: complexities and surprises. *Ann. Bot.* **116**:
535 849-864.
- 536 Peakall, R., and Smouse, P.E. 2006. GENALEX 6: genetic analysis in Excel. Population genetic software for
537 teaching and research. *Mol Ecol. Notes*, **6**: 288-295.
- 538 Project IRGS. 2005. The map-based sequence of the rice genome. *Nature*, **436**(7052): 793-800.
- 539 Petersen, T.N., Brunak, S., von Heijne, G., and Nielsen, H. 2011. SignalP 4.0: discriminating signal peptides
540 from transmembrane regions. *Nature Methods*, **8**: 785-786.

- 541 Ranjitkar, S., Luedeling, E., Shrestha, K.K., Guan, K., and Xu, J. 2013. Flowering phenology of tree
542 rhododendron along an elevation gradient in two sites in the Eastern Himalayas. *Int. J. Biometeorol.* **57**: 225-
543 240.
- 544 Salazar, J.A., Rubio, M., Ruiz, D., Tartarini, S., Martínez-Gómez, P., and Dondini, L. 2015. SNP development
545 for genetic diversity analysis in apricot. *Tree Genet. Genomes*, **11**: 15.
- 546 Schlautman, B., Fajardo, D., Bougie, T., Wiesman, E., Polashock, J., Vorsa, N., Steffan, S., and Zalapa, J. 2015.
547 Development and validation of 697 novel polymorphic genomic and EST-SSR markers in the American
548 cranberry (*Vaccinium macrocarpon* Ait.). *Molecules*, **20**: 2001-2013.
- 549 Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. 2015. BUSCO: assessing
550 genome assembly and annotation completeness with single-copy orthologs *Bioinformatics*, **31**(19): 3210-
551 3212.
- 552 Singh, K.K., Rai, L.K., Gurung, B. 2009. Conservation of rhododendrons in Sikkim Himalaya: an overview.
553 *World J. Agric. Sci.*, **5**: 284-296.
- 554 Singh, N., Choudhury, D.R., Singh, A.K., Kumar, S., Srinivasan, K., Tyagi, R., Singh, N.K., and Singh, R.
555 2013. Comparison of SSR and SNP markers in estimation of genetic diversity and population structure of
556 Indian rice varieties. *PloS one*, **8**(12): e84136.
- 557 Stinchcombe, J.R., and Hoekstra, H.E. 2007. Combining population genomics and quantitative genetics: finding
558 the genes underlying ecologically important traits. *Heredity*, **100**: 158-170.
- 559 Sork, V.L., Squire, K., Gugger, P.F., Steele, S.E., Levy, E.D., and Eckert, A.J. 2016. Landscape genomic
560 analysis of candidate genes for climate adaptation in a California endemic oak, *Quercus lobata*. *Am. J. Bot.*
561 **103**: 33-46.
- 562 Strickler, S.R., Bombarely, A., and Mueller, L.A. 2012. Designing a transcriptome next-generation sequencing
563 project for a nonmodel plant species. *Am. J. Bot.* **99**: 257-26.
- 564 Sun, H., Liu, Y., Gai, Y., Geng, J., Chen, L., Liu, H., Kang, L., Tian, Y., and Li, Y. 2015. *De novo* sequencing
565 and analysis of the cranberry fruit transcriptome to identify putative genes involved in flavonoid
566 biosynthesis, transport and regulation. *BMC Genomics*, **16**: 652.

- 567 Thuiller, W., Lavorel, S., Araújo, M.B., Sykes, M.T., and Prentice, I.C. 2005. Climate change threats to plant
568 diversity in Europe. P.N.A.S. USA, **102**: 8245-8250.
- 569 Urban, M.C. 2015. Accelerating extinction risk from climate change. Science, **348**: 571-573.
- 570 Varshney, R.K., Granera, A., and Sorrells, M.E. 2005. Genic microsatellite markers in plants: features and
571 applications. Trends Biotechnol. **23**(1): 48-55.
- 572 Vetaas, O.R. 2002. Realized and potential climate niches: a comparison of four Rhododendron tree species. J.
573 Biogeogr. **29**: 545-554.
- 574 Xu, J., Grumbine, R.E., Shrestha, A., Eriksson, M., Yang, X., Wang, Y., and Wilkes, A. 2009. The melting
575 Himalayas: cascading effects of climate change on water, biodiversity, and livelihoods. Conserv. Biol. **23**(3):
576 520-530.
- 577 Xu, Q., Zhu, C., Fan, Y., Song, Z., Xing, S., Liu, W., Yan, J., and Sang, T. 2016. Population transcriptomics
578 uncovers the regulation of gene expression variation in adaptation to changing environment. Sci. Rep. **6**:
579 25536.
- 580 Zhang, H.B., Xia, E.H., Huang, H., Jiang, J.J., Liu, B.Y., and Gao, L.Z. 2015. *De novo* transcriptome assembly
581 of the wild relative of tea tree (*Camellia taliensis*) and comparative analysis with tea transcriptome identified
582 putative genes associated with tea quality and stress response. BMC Genomics, **16**: 298.
- 583 Zhao, Z., Tan, L., Dang, C., Zhang, H., Wu, Q., and An, L. 2012. Deep-sequencing transcriptome analysis of
584 chilling tolerance mechanisms of a subnival alpine plant, *Chorispora bungeana*. BMC Plant Biol. **12**: 222-
585 239.

Tables Legends:

Table 1: A summary of statistics for the sequences obtained after sequencing of the four cDNA libraries of flower and leaf of *R. arboreum*

Table 2: Statistics for the final master assembly generated after concatenation and clustering of the four transcript libraries of flowers and leaves of *R. arboreum*

Table 3: Annotation statistics of the assembled unigenes of *R. arboreum*

Table 4: Statistics for Simple Sequence Repeat (SSR) search in the transcriptome of *R. arboreum*

Table 5: Characteristics of the 43 Simple Sequence Repeat (SSR) loci of *R. arboreum*

Table 6: A summary of the Single-Nucleotide Polymorphism (SNP) search using SeqMan Pro (DNASTAR, Inc.) in the transcriptome of *R. arboreum*.

Table 1: A summary of statistics for the sequences obtained after sequencing of the four cDNA libraries of flower and leaf of *R. arboreum*

Summary	Jhatingri Leaf (L2).fa	Ropru Leaf (L4).fa	Jhatingri Flower (F2).fa	Ropru Flower (F4).fa
Number of Raw reads (in million)	30.4	32.6	14.4	45.1
Percent GC	48	49	44.71	45.62
Number of Reads left after pre-processing (in million)	28.7	28.6	13.6	34.5
Number of Transcripts Generated	80,064	111,483	63,117	106,527
Maximum Transcript Length (bp)	11,520	14,816	15,860	10,008
Minimum Transcript Length (bp)	300	300	300	300
Average Transcript Length (bp)	804.1	824.4	884.30	836.61
Total Transcripts Length (bp)	64,381,317	91,907,227	55,814,544	89,121,107
Transcripts >= 1 Kbp	18,209	26,210	18,036	26,980
N50 value (bp)*	1,117	1,138	1,257	1,172

*At least 50% of the assembled transcript nucleotides are found in contigs that are at least of N50 length

Table 2: Statistics for the final master assembly generated after concatenation and clustering of the four transcript libraries of flowers and leaves of *R. arboreum*

<i>Attributes</i>	<i>Stats</i>
<i>Total number of 'trinity genes'*</i>	89,885
<i>Total number of contigs generated</i>	113,167
<i>Percent GC</i>	44.86
<i>Maximum transcripts length (bp)</i>	15,900
<i>Minimum transcripts length (bp)</i>	500
<i>Average transcripts length (bp)</i>	1164.6
<i>Median transcripts length (bp)</i>	628
<i>Total transcripts length (bp)</i>	1,31,789,145
<i>Total number of non-ATGC characters</i>	0
<i>Total number of Transcripts ≥ 500 b</i>	113,167
<i>Total number of Transcripts ≥ 1 Kb</i>	46,542
<i>Total number of Transcripts ≥ 10 Kb</i>	20
<i>N50 value** (bp)</i>	1,387
<i>N10 value*** (bp)</i>	3,374

*On the basis of shared sequence content, the Trinity algorithm groups different transcripts into a single cluster, each of which is individually denoted as a 'Trinity gene'

**At least 50% of the assembled transcript nucleotides are found in contigs that are at least of N50 length

***At least 10% of the assembled transcript nucleotides are found in contigs that are at least of N10 length

Table 3: Annotation statistics of the assembled unigenes of *R. arboreum*

<i>Attribute</i>	<i>Number of unigenes</i>
Total unigenes* available for annotation	113,167
Number of translated unigenes	78,810
Transdecoder predicted peptides	96,872
BLASTp hits against Swiss-Prot	57,352
BLASTx hits against Swiss-Prot	62,716
BLASTx hits against nr (viridiplantae) database	62,959
BLASTn hits against nt database	43,220
P-fam hits	1,118
tmhmm predictions	29,233
signalP predictions	4,857
GO annotations	71,961
KEGG annotations	5,710
BLASTp hits against plant TF database	20,513
Number of unannotated unigenes	41,206

*Contigs are generated by adjoining K-mers of definite length, those sequences that could not be extended further are referred as unigenes

Table 4: Statistics for Simple Sequence Repeat (SSR) search in the transcriptome of *R. arboreum*

<i>Attribute</i>	<i>Stats</i>	<i>Motif length (bp)</i>
<i>Number of sequences examined</i>	113,167	-
<i>Size of examined sequences (bp)</i>	131,789,145	-
<i>Number of SSRs located</i>	35,419	-
<i>Number of SSR-containing sequences</i>	27,333	-
<i>Number of sequences with >1 SSR</i>	6,423	-
<i>Number of compound SSRs</i>	2,641	-
<i>Mono-nucleotide Repeats</i>	10,215	>= 10
<i>Di-nucleotide Repeats</i>	19,425	>= 6
AC/GT	14,809 (90.4%)	6-12
AG/CT	1,067 (6.5%)	6-12
AT/TA	503 (3.1%)	6-11
GC/CG	24 (0.1%)	6-7
<i>Tri-nucleotide Repeats</i>	5,529	>= 5
<i>Tetra-nucleotide Repeats</i>	183	>= 5
<i>Penta-nucleotide Repeats</i>	39	>= 5
<i>Hexa-nucleotide Repeats</i>	28	>= 5

Table 5: Characteristics of the 43 Simple Sequence Repeat (SSR) loci of *R. arboreum*

#	Locus ID	Primer Sequence	Motif	T _a	N _a	N _c	Observed length (bp)	H _o	H _e	PIC	Probable function
1	F2_c10213_g1_i1	F: CAGTCCTCTCTCTCCTTCG R: CACGTCGATAATCTGAAGGTA	(CTT) ₆	57	4	3.383	144-160	0.367***	0.704	0.650	Transcription initiation factor TFIIB
2	F2_c12172_g1_i1	F: ATCATTGTCATCATCTTTCA R: AAAGTCCTCCTAATCGAAAAA	(GA) ₇	57	8	2.985	144-160	0.400***	0.665	0.635	Hydroxyacylglutathione hydrolase
3	F2_c15191_g1_i1	F: AAACCGTTTTGTCTCGTAAG R: CCAGGAAATCAACCTTCTTTA	(CT) ₁₀	57	7	3.727	138-150	0.433**	0.732	0.691	U3 small nucleolar ribonucleoprotein
4	F2_c17711_g1_i4	F: ATCACAGACACCGTTGTAATC R: ATCTATAGCCATCGTTGAACA	(AAT) ₄	52	3	2.799	150-160	0.433***	0.643	0.568	Gluconokinase
5	F2_c18454_g1_i1	F: GTCTTGGGTTAGGATCACCT R: CATTTCCGTTCAATCAATTT	(GCG) ₅	52	4	2.936	152-156	0.433***	0.659	0.598	Ribose 5-phosphate isomerase A
6	F2_c19464_g1_i1	F: CAGTAGGTTTAAGGTGAGCAG R: GATAGAGCGAGAGAGAAGAGG	(TC) ₇	57	9	4.959	150-170	0.267***	0.798	0.773	RP-S20; small subunit ribosomal protein S20
7	F2_c19472_g1_i2	F: ACCCAATCCAAATGATTTAT R: GCGGAGAGATTTTATTCTTT	(AC) ₆	52	6	3.965	160-176	0.333***	0.748	0.706	5'-AMP-activated protein kinase, regulatory gamma subunit
8	F2_c19855_g1_i1	F: AGAGAGAGGGAGAGAAATTGA R: ATTATTCTTCAACCCATGAAT	(GGA) ₅	52	3	2.469	160-170	0.167***	0.595	0.510	auxin-responsive protein IAA
9	F2_c20104_g1_i1	F: ACATGAAGAAGATCGCTTGT R: CTTAGTTCAAGACCAAAGCAA	(GCC) ₄	57	4	2.517	164-174	0.067***	0.603	0.535	ABCBI; ATP-binding cassette, subfamily B
10	F4_c34834_g1_i2	F: TCTGCTGAGATTGAAAAGAAG R: CCCAAGAGAGAGAGAAGAGAC	(AGA) ₆	57	9	3.141	154-174	0.433***	0.682	0.661	CHMP4;charged multivesicular body protein 4
11	F4_c33966_g1_i1	F: GGGCTAAAGAAGGATTCTAAA R: TTACTCTGCATCTCACATTCC	(GAG) ₄	57	4	3.000	162-170	0.200***	0.667	0.611	SAUR family protein
12	F4_c1114_g1_i1	F: TTACGAGAACTCCCTCTAAGC R: CTCGAGAGAGTAGAGGAAGAGA	(TCT) ₄	57	3	2.830	144-171	0.467***	0.647	0.571	Aldehyde dehydrogenase (NAD+)
13	F4_c119403_g1_i1	F: GATGCTCTCCATTCTGTAGC R: CTACATCCACACTTGCTTCTC	(AG) ₇	52	5	3.186	150-160	0.500***	0.686	0.623	Aldehyde dehydrogenase (NAD+)
14	F4_c120793_g1_i1	F: ATCTCAAAGTCACGGATAACA R: TTTGACGTACTTGAGGTTTAC	(TC) ₈	55	3	1.998	174-190	0.100***	0.499	0.432	3-ketoacyl-CoA synthase
15	F4_c120959_g1_i1	F: GTCAGATGCAAACTCCAAG R: TAATTGCAAGACAAGAACCAT	(CTG) ₄	57	3	2.711	194-200	0.667***	0.631	0.556	RNF5; E3 ubiquitin-protein ligase
16	F4_c12175_g1_i1	F: CTTATTACCACCGACCTTTTC R: AGGTGAGGGTTTTAATAGGC	(CT) ₇	52	5	2.791	150-160	0.367***	0.642	0.586	Adenylate kinase
17	F4_c142711_g1_i1	F: AATTATGAGGGGAGAGAGAGA R: GCACATAATTTGCGTACACC	(GA) ₇	55	6	4.724	140-166	0.800***	0.788	0.755	ADP-ribosylation factor GTPase-activating protein 1
18	F4_c14322_g1_i1	F: TTTACCATGACGTCTAAAGGA R: TCTTTCAAATACAACAACCTGGA	(CA) ₈	55	6	2.936	136-160	0.367***	0.659	0.624	Syntaxin of plants SYP7
19	F4_c17802_g1_i1	F: GAGCCATCCTGGTAACATCT R: ATTTATCTCCGCTCAGTCG	(AG) ₈	57	8	4.489	140-156	0.633***	0.777	0.748	U4/U6 U5 tri-snRNP component SNU23
20	F4_c20375_g1_i1	F: AAACACCATATGTTGAAGAGC R: GCGCCTCTGATTAGAATACAA	(GCT) ₆	55	7	4.651	140-170	0.333***	0.785	0.753	RP-S5e;small subunit ribosomal protein S5e
21	F4_c21402_g1_i1	F: GAACCTCTTCGAGCTTCTTG R: AACCCTTTTAGTCCGAGTTTT	(CCG) ₄	55	3	2.605	248-266	0.300***	0.616	0.547	Inosine triphosphate pyrophosphatase
22	F4_c21545_g1_i1	F: CTCCTTGTCTGTGTAGTCG R: CAGTACTTAAACCTGGCAGA	(TC) ₆	55	3	2.830	156-160	0.000***	0.647	0.571	Exosome complex component MTR3

23	F4_c2893_g1_i2	F: GTTGAAAATACTGGGACCTCT R: CTTACATCAAGAGGCAAATA	(AAG) ₄	57	3	2.817	144-146	0.500***	0.645	0.572	Oligosaccharyltransferase complex subunit alpha (ribophorin I)
24	F4_c29135_g1_i2	F: TCCATCATCTCTTCTTCTCA R: ATCAAAGCCTCCTTTGTATC	(ATG) ₆	57	4	3.109	144-154	0.233***	0.678	0.629	Myb proto-oncogene protein
25	F4_c29390_g1_i1	F: TCCATGTCGTGGAGTAGG R: CCATCAGCTGTCATGAATAA	(TCT) ₄	57	3	1.613	170-180	0.133***	0.380	0.343	DNA polymerase delta subunit 1
26	F4_c27506_g1_i1	F: CTGACGGAGAACACAAATCTA R: GTTGGTGTTCGTTTCAGTTAC	(GA) ₆	57	3	1.965	140-150	0.200***	0.491	0.433	Arginine decarboxylase
27	F4_c27873_g1_i1	F: GTACGATCTCAAGTCGTTCAA R: GAAAGTCCAGCAGATCCAG	(CGT) ₄	57	8	2.557	140-160	0.333***	0.609	0.584	THI4; thiamine thiazole synthase
28	F4_c24918_g1_i2	F: GTAGACGATGGGTCCATATTT R: CTCTAGTCTTTTCTCCACCAG	(GAG) ₅	57	4	2.510	180-200	0.233***	0.602	0.523	CYP75A; flavonoid 3',5'-hydroxylase
29	F4_c25637_g1_i1	F: GAACACGATTGGGTCTTTAG R: AAGGTCTCGTGTTTTGAGTT	(TGC) ₆	57	6	4.091	144-160	0.367***	0.756	0.719	Peptidyl; prolyl isomerase H (cyclophilin H)
30	F4_c25211_g1_i1	F: AGGTCTAGGCTTACACCATCT R: CAAAGAATTCCGACACAATTT	(TC) ₈	57	6	3.352	134-150	0.300***	0.702	0.655	DCTPP1; dCTP diphosphatase
31	F4_c32694_g1_i1	F: CAGAACACTCACTCTCACCAT R: CTCCCATTTTAAGCTTCAGTC	(AG) ₇	57	6	4.306	146-158	0.367***	0.768	0.732	Small subunit ribosomal protein S15Ae
32	F4_c32889_g1_i1	F: AGAGAGAGAGAGACCTGCCTA R: TGATCTCAAGGAAGATTCAGA	(GA) ₈	57	3	2.985	110-120	0.633***	0.665	0.592	FAD2; omega-6 fatty acid desaturase
33	F4_c31337_g1_i1	F: TGGAGTACAACAACCTCTCAC R: TCTGGTTCAACAACAACCTACC	(TCA) ₄	57	3	2.378	154-164	0.433***	0.579	0.540	Jasmonate ZIM domain-containing protein
34	F4_c32077_g1_i1	F: GGAAAATATATCGAGAGACAAAA R: GATTGAACTTTTGAGGGAAT	(CT) ₇	57	3	1.867	156-160	0.267***	0.464	0.419	IST1-vacuolar protein sorting-associated protein
35	L2_c13782_g1_i1	F: GATCACTATCCCAACAATGG R: GTGTAGAACCTCAGCACGTT	(CCT) ₄	57	4	2.174	190-200	0.333***	0.540	0.490	Protein transport protein SEC61 subunit beta
36	L2_c14487_g1_i1	F: ATCAGACCCTCAAATCTTCT R: ATGATTCCAGGTCGATAGAT	(AGA) ₄	57	4	3.719	120-136	0.667***	0.731	0.682	psaF; photosystem I subunit III
37	L2_c16633_g1_i1	F: GGAATACTCCTCTCTGAGAC R: GAAAGCTCTCTTCTGATTCG	(TC) ₉	57	3	2.605	150-160	0.367***	0.616	0.542	beta-amylase
38	L2_c17056_g2_i1	F: CAGAGTTGTCCAGTAAGTTCTG R: TGGTTTCTCAATGGCTAATA	(AAG) ₅	57	3	2.582	160-170	0.233***	0.613	0.532	Geranylgeranyl diphosphate synthase, type II
39	L2_c18244_g1_i1	F: GAGGCCACTGTGTGTGAAC R: CTCTACTCCATCTCCTCCTTC	(AGA) ₄	57	6	4.401	150-156	0.367***	0.773	0.743	L-ascorbate peroxidase
40	L2_c21578_g1_i1	F: GCAAACTGGGAATATAAAACC R: GGAAAAGCCTATTCTAGCACT	(GCT) ₄	57	3	1.737	148-158	0.200***	0.424	0.386	3-heterogeneous nuclear ribonucleoprotein A1
41	L4_c28280_g1_i2	F: AACAGCCTCAAGTCATAATCA R: TGACAGTAAAGGAAAGACAGC	(GAA) ₄	52	4	2.740	150-160	0.433***	0.635	0.584	SYVN1; E3 ubiquitin-protein ligase synoviolin
42	L4_c26046_g1_i1	F: ACCCTAATCGTGTCTTCTTCTC R: GAACTGTTGCAGAGATTGAAC	(AG) ₇	52	5	3.261	154-164	0.467***	0.693	0.634	U3 small nucleolar RNA-associated protein 13
43	L4_c24337_g1_i2	F: ACGCTTAGACCTTACAGTGAA R: TGAAGAGTTGCTTACTGATCC	(GAA) ₅	52	4	3.614	156-170	0.533***	0.723	0.673	cdc20; cofactor of APC complex
Total				201		132.018					
Mean				4.674		3.070		0.364	0.650	0.6	

T_a- Annealing temperature (°C); bp: base pair; N_a- Total number of alleles; N_e- effective number of alleles; H_o- Observed heterozygosity; H_e- Expected heterozygosity; PIC- Polymorphic information content; Significant deviations from Hardy-Weinberg equilibrium at *p<0.05, **p<0.01, ***p<0.001.

Table 6: A summary of the Single-Nucleotide Polymorphism (SNP) search using SeqMan Pro (DNASTAR, Inc.) in the transcriptome of *R. arboreum*.

<i>Type of SNP</i>	<i>Total</i>	<i>Average depth</i>	<i>Average SNP (%)</i>
Total	36,921	67.6	35.5
# Indels	2,667 (7.2%)	41.2	34.8
# Transition	22,229 (60.2%)	73.3	35.6
# Transversion	11,992 (32.5%)	62.9	35.3
# Misc	33		
# Contigs	7,518 (5 SNP/ contig)		
Filtered[#]	811	27.3	50.5
# Indels	75 (9.2%)	37.6	51.9
# Transition	448 (55.2%)	26.1	50.2
# Transversion	274 (33.8%)	24.8	50.3
# Misc	14		
# Contigs	719 (1 SNP/ contig)		

[#] Results after removing the SNPs with <10 depth and with SNP% of <50%

Draft

Figure Captions:

Figure 1: The length distribution of the contigs of the master transcriptome of *R. arboreum*

Figure 2: Categories for Top 10 Gene Ontologies under each subcategory **A.** Molecular Function; **B.** Biological Process; and **C.** Cellular Components, as represented by the transcriptome of *R. arboreum*

Figure 3: Pathway annotations of transcripts of *R. arboreum* representing various functional categories

Draft

Legend for Supplementary data:**gen-2017-0143.R2Suppla:**

- Alignment statistics to assess the read content of the transcriptome of *R. arboreum*
- A distribution characteristic of the percent length coverage for the top matching Swiss-Prot database hit
- A summary of completeness of *R. arboreum* transcriptome as per the comparison with BUSCOs from plant lineage dataset
- Major species represented by GO terms of the annotated transcripts in *R. arboreum*
- A summary of KEGG pathway annotations
- A summary of proportion of the 59 different TF families in transcriptome of *R. arboreum*
- Amplification profile generated for the loci, namely (A.) F4_c142711_g1_i1, and (B.) F4_c20375_g1_i1). Lanes 1- 30 represent sampled individuals of *R. arboreum*; Ld: 100 bp DNA ladder (Bangalore GeneiTM) as size standard

gen-2017-0143.R2Supplb: A brief summary of GO annotations

gen-2017-0143.R2Supplc: A list of all the contigs identified with SNPs and Indels from the transcriptome library of *R. arboreum*

Draft

Figure 1: The length distribution of the contigs of the master transcriptome of *R. arboreum*

Figure 2: Categories for Top 10 Gene Ontologies under each subcategory **A.** Molecular Function; **B.** Biological Process; and **C.** Cellular Components, as represented by the transcriptome of *R. arboreum*

Figure 3: Pathway annotations of transcripts of *R. arboreum* representing various functional categories