Journal of Genetics (2022)101:51
https://doi.org/10.1007/s12041-022-01391-w

RESEARCH ARTICLE

De novo genome assembly and annotation of gall-forming medicinal plant *Pistacia chinensis* subsp. *integerrima* (J. L. Stewart ex Brandis) Rech. f.

SANTHOSH N. HEGDE[1], NOORUNNISA BEGUM[1], AMIT BHATT[2], SUBRAHMANYA KUMAR KUKKUPUNI[1], PADMA VENKATASUBRAMANIAN[3], J. L. N. SASTRY[2], S. BADRINARAYAN[2], MALALI GOWDA[1]* and PAVITHRA NARENDRAN[1]*

[1]*Center for Functional Genomics and Bioinformatics, The University of Trans-Disciplinary Health Sciences and Technology (TDU), 74/2, Post Attur via Yelahanka, Jarakabande Kaval, Bengaluru 560 064, India*
[2]*Dabur India Limited and Dabur Research and Development Centre, 22, Site-IV, Sahiabad, Ghaziabad 201 010, India*
[3]*SRM Institute of Science and Technology, SRM Nagar, Kattankulathur, Chengalpattu 603 203, India*
*For correspondence. E-mail: Malali Gowda, malali.gene@gmail.com; Pavithra Narendran, pavithranarendran@gmail.com.

Received 13 June 2022; revised 25 July 2022; accepted 28 July 2022

Abstract. *Pistacia chinensis* subsp. *integerrima* is one of the medicinal plants, well known for gall formation and popularly used in Ayurveda to treat various systemic diseases such as chronic disorders, respiratory problems, etc. *P. integerrima* genome characterization will aid in the study of *Pistacia* genes and pathways involved in therapeutic application. To understand the biological characteristics of this plant and to gain the genetic insight into the biosynthesis of its natural compounds, the whole genome of *P. integerrima* and its leaf transcriptome was sequenced using Illumina sequencing technology. The sequenced genome was functionally annotated, and gene prediction was performed with integrated genome annotation workflow. The pathway analysis was carried out using KEGG database. We obtained a draft genome assembly of 462 Mb with N50 16,145 bp. A total of 39,452 genes were found, and 18,492 of these contained RNA or protein evidence. We characterized the genes involved in biosynthetic pathways of different plant secondary metabolites such as flavonoids and terpenoids. Also, we identified miR397 and miR828 family noncoding RNA; which mainly targets the laccase (LCA) and MYB protein functioning respectively. Phylogeneic analysis showed that *P. integerrima* is genetically more closer to *P. vera.* In this study, we attempt to explore the whole genome information of *P. integerrima* which will provide a genomic insight in the future for omics studies as well as serves as valuable resource for the molecular characterization of medicinal compounds.

Keywords. medicinal plant; Ayurveda; genome; *Pistacia integerrima*.

Introduction

Pistacia chinensis subsp. *integerrima* (J. L. Stewart ex Brandis) Rech. f. (hereafter called as *P. integerrima*) is well known for its medicinal value as antibacterial and antifungal activity in traditional medicine (Bibi *et al.* 2015). This species belongs to the family Anacardiaceae and is a dioecious tree native to Asia. The *P. integerrima* belonging to genus *Pistacia* consists of at least 11 species and is estimated to be about 80 million years old (Ziya *et al.* 2016). *Pistacia* is widely distributed in east Afghanistan, Pakistan, northwest and west Himalayas. *P. integerrima* is well known for the formation of rough, horn-shaped hollow structures known as 'galls' on apical buds due to insect infestation (Orwa *et al.* 2009; Pant and Samant 2010; Zahoor *et al.* 2018).

P. integerrima galls are used as an important ingredient in many medicines for treatment of cough and respiratory problems, lack of appetite, dyspeptic vomiting and dysentery. These galls are referred by various names across India,

Supplementary Information: The online version contains supplementary material available at https://doi.org/10.1007/s12041-022-01391-w.

Published online: 02 November 2022

such as karkatshringi, kakroi, kakring, kakra, kakkar, and kakarsinghii (Orwa *et al.* 2009). This plant is mentioned as 'Karkatashringi' in Ayurveda books and prescribed to manage respiratory tract diseases including cough, dyspnoea and hiccups (Orwa *et al.* 2009). It is used in the preparation of compound medicinal formulations such as Shringyadi churna, Balachaturbhadra churna, Brihat talisadi churna, Devadarvayadi kwatha churna, Shatavaryadi ghrita, Chyawanprash avaleha, Dashmularista, Siva gutika, Khadiradi gutika, etc (Barinder and Saurabh 2015). It has an antispasmodic effect on muscles at low doses, inhibiting repetitive peristaltic intestinal movements (Rauf 2019). The in-depth phytochemical analysis of the roots, leaves, barks and fruits of this plant have been reported by researchers (Zahoor *et al.* 2018). *P. integerrima* contains several phytoconstituents like flavonoids, tannins, triterpenes, alkaloids, lipophilic elements (tanshinone), phenolic acids, anthocyanins, purpurogallin, etc. of commercial value and therapeutic potential (Rauf *et al.* 2015). *P. integerrima* extracts have yielded a number of terpenoids, sterols, and phenolic compounds. Because of its chemical contents, the plant possesses a wide range of biological effects, including antimicrobial, antioxidant, analgesic, cytotoxicity, and phytotoxicity (Rauf *et al*. 2015). Many flavonoid glycosides such as Pistacides A and B have been reported from this plant. Along with flavonoid glycosides, carotenoids, triterpenoids, and catechins were discovered in the leaves of *P. integerrima* (Ullah *et al*. 2012).

Today with the advent of next-generation sequencing technology, we can explore the medicinal plant species at the genome level. Understanding the biochemical pathways of valuable bio-compounds from a genomics viewpoint would offer essential tools by synthetic biology for the large-scale development and production of novel chemicals. Genes implicated in some biochemical pathways of many medicinal plants could be identified using new technologies such as next-generation sequencing (Zhang *et al.* 2019). Genome study offers information on a plant species' genome composition, including genome size estimates, degrees of heterozygosity, and repeat information (Ziya *et al.* 2016). There is no genomic information available for *P. integerrima*. In this study, we aimed to characterise the genome of *P. integerrima* which will help in understanding the *Pistacia* genes and pathways responsible for therapeutic use.

Materials and methods

Sample collection

In this study, trees of *P. integerrima* were identified at their natural habitat and authenticated by the expert taxonomists and the field botanists. The leaf samples were collected from the state of Uttarakhand, India with the guidance of botanists. Plant sample for RNA isolation was immediately rinsed with normal saline and stored in RNALater to prevent degradation of RNA. These samples were shipped to the lab and stored in −80°C. Voucher specimens of the gall and leaf were deposited at FRLH Herbarium, Bengaluru.

DNA isolation, library preparation and sequencing

Genomic DNA was extracted using CTAB method (Novaes *et al.* 2009). The quality of Genomic DNA was confirmed using gel electrophoresis and Nanodrop. The genomic DNA was enzymatically fragmented to generate ∼250 bp fragments. About 50 ng of fragmented DNA was used to generate a sequencing library using NEBNext® Ultra™ II DNA FS Library Prep kit for Illumina. The library was quantified using Qubit DNA High Sensitivity quantitation assay and library quality was checked on Bioanalyzer 2100 using Agilent 7500 DNA kit. DNA sequencing was done on the HiSeq 2500 platform.

RNA library preparation and sequencing

Total RNA was extracted from mature leaf samples. The quality of RNA was checked on Bioanalyzer and quantified using QUBIT dsRNA HS kit. The library was prepared using 'NEBNext® Ultra™ RNA Library Prep Kit for Illumina®' with Illumina standardized protocol. The final enriched libraries were further validated for quality on Agilent Bioanalyser using DNA High Sensitivity chip and for quantification on real-time PCR (KAPA Library Quantification kit). The quality and quantity of the prepared library met the Illumina standards required for further sequencing and hence the library was normalized. The library was denatured using NaOH followed by neutralizing the pH conditions by adding 0.2 N Tris, pH 7 and was taken further for cluster generation and sequencing. About 40 million paired-end reads were generated on Illumina NextSeq500 platform.

Genome size estimation and de novo assembly

Sequencing data of *P. integerrima* from the Illumina library were used to perform initial quality of raw reads using FastQC v0.11.6 (Andrews 2010) and trim galore v0.4.4_dev (Krueger 2015) was used with default parameters to remove low-quality reads. The filtered reads were further taken for k-mer analysis to estimate the genome size using Jellyfish v2.2.10 (Marçais and Kingsford 2011). The distribution of 21-kmer showed a major peak at 69×. The *P. integerrima* genome size was calculated based on the total number of k-mers and the corresponding k-mer depth using the formula: genome size = k-mer number/peak depth.

The high quality reads were used to predict the best k-mer using KmerGenie v1.7048 (Chikhi and Medvedev 2014).

The assembly size was predicted using different k-mers from 21 to 121 with interval of two in diploid mode. The best k-mer which gave the best assembly size was selected for *de novo* assembly. The assembly was performed using two software, SOAPdenovo v2.04 (Luo *et al.* 2012) and SPAdes v3.11.1 (Bankevich *et al.* 2012).

Postprocessing and validation of genome assembly

The assembly was further used for scaffolding using SSPACE v5.26.2 (Boetzer *et al.* 2011) and GapClosure v1.12 (Luo *et al.* 2012) module from SOAPdenovo2. The chloroplast, mitochondrial and vector sequences were downloaded from NCBI database. The chloroplast, mitochondrial and vector sequences were identified using Blastn v2.6.0+ (Camacho *et al.* 2009) (blastn parameter used for mitochondria and chloroplast; e-value is 1e−10 and for vector sequences; reward: 1, penalty: −5, gapopen: 3, gapextend: 3, dust: yes, soft_masking: true, searchsp: 1750000000000) and removed from the assembly. After the removal of chloroplast, mitochondria and vector sequences, genome assembly of *P. integerrima* was subjected to benchmarking universal single-copy orthologs (BUSCO) v3.0.2 (Sima *et al.* 2015) to assess the genome completeness.

Transcriptome assembly and annotation

Data quality was checked using FastQC v0.11.613 (Andrews 2010). Reads with Phred quality score < 30 were removed using Trimgalore v0.4.4_dev14 (Krueger 2015). The trimmed reads from leaf were used to assemble into transcripts using Trinity v2.4.031 (Grabherr *et al.* 2011). TransDecoder-v5.0.2 (https://transdecoder.github.io/) was used to identify candidate coding regions within transcript sequences of the leaf. These sequences were searched against the Viridiplantae sequences from UniProtKB database.

Genome annotation

***Repeats identification*:** The repeat library building and *de novo* repeat identification was done using Repeat modeller v1.0.11 (https://www.repeatmasker.org/RepeatModeler/). Repeat annotation was carried out using Repbase v23_8 (Bao *et al.* 2015) with *A. thaliana* as reference. The RepeatMasker (https://www.repeatmasker.org) was used to mask the identified and annotated repeats in the assembly. The simple sequence repeats (SSRs) were identified using MIcroSAtellite (MISA) identification tool (Thiel *et al.* 2003).

***Protein-coding gene prediction*:** The repeat masked assembly was carried forward for annotation using different methods such as *ab initio* based predictions, homology-based predictions and evidence-based predictions. In *ab initio* based predictions, Augustus v3.3.2 (Stanke and Stephan 2003), GlimmerHMM v3.0.4 (Majoros *et al.* 2004), GeneID v1.4.4 (Blanco *et al.* 2007) and SNAP (Korf 2004) were used with parameters trained from *A. thaliana*. Predicted homologous protein sequences from *P. vera* (GCF_008641045.1), *A. thaliana* (GCF_000001735.4), *Theobroma cacao* (GCF_000208745.1), *Citrus sinensis* (GCA_000695605.1), *Vitis vinifera* (GCF_000003745.3), *Beta vulgaris* subsp. *vulgaris* (GCF_000511025.2), *Dorcoceras hygrometricum* (GCA_001598015.1), *Ricinus communis* (castor bean; GCF_000151685.1), *Oryza sativa* (Nipponbare;IRGSP-1.0; GCF_001433935.1) were retrieved from NCBI refseq database and used for homology-based gene prediction by aligning these genomes to our assembly to identify the homologous genes with GeMoMa v1.7 (Keilwagen *et al.* 2016). Using Trinity v2.4.031 (Grabherr *et al.* 2011), the RNA-Seq reads were assembled into *de novo* contigs into unigenes and the resulting unigenes were aligned using BLAT (Kent 2002) to the repeat-masked assemblies and then the gene structures of BLAT alignment results were modelled using program to assemble spliced alignments (PASA v2.4.1) (Haas *et al.* 2003). The RNA seq reads were also aligned to repeat masked genome assembly using Tophat2 (Kim *et al.* 2013) with minimum and maximum intron lengths of 50 and 500,000 bp, respectively. And assembled using cufflinks v2.2.1 (Trapnell *et al.* 2010) and were used as evidence in PASA. Predictions from all the three methods were combined with EVidenceModeler v1.1.1 (Haas *et al.* 2008) (score > 1000) to produce a consensus gene set (Bi *et al.* 2019). The functional annotation of protein-coding genes was carried out by aligning protein sequences to the Swissprot (Boeckmann *et al.* 2003) database. The protein motifs and domains were annotated by searching against the Pfam (Mistry *et al.* 2021) database. Pathway mapping of genes was done using the Kyoto Encyclopaedia of Genes and Genomes (KEGG) Automatic Annotation Server (KASS) (Moriya *et al.* 2007).

***Noncoding RNAs identification*:** tRNAscan-SE v.2.0.7 (Chan and Lowe 2019) algorithm with default parameters was applied to predict tRNA genes. MiRNA and small nucleolar RNA (SnRNA) genes were predicted using INFERNAL v1.1.2 software (Nawrocki and Eddy 2013) with the Rfam database (Griffiths-Jones *et al.* 2003).

Gene family construction and orthology detection

OrthoVenn2 (Xu *et al.* 2019) was used for the clustering of protein sequences from the six species including *A. thaliana*, *Citrus sinensis*, *P. vera*, *Solanum tuberosum*, *Vitis vinifera* and *P. integerrima*. Comparative data from OrthoVenn2 analysis was further classified into a list of potential orthologs, co-orthologs and paralogs. Besides classifying, the

programme has also grouped the proteins into specific groups by clustering the gene-pairs.

Phylogenetic analysis of P. integerrima

The proteomes of 13 sequenced plant species including *P. integerrima*, *A. thaliana*, *Oryza sativa*, *Vitis vinifera*, *Glycine max*, *Theobrama cocao*, *Beta vulgaris*, *Citrus sinesis*, *Ricinus communis*, *P. vera*, *Solanum tuberosum*, *Solanum lycopersicum*, *Eucalyptus grandis* were used to search for homologues and unique genes using ProteinOrtho v6.0.27 (Lechner *et al.* 2011) tool. Phylogenetic tree was constructed using ProteinOrtho.

Results

Genome sequencing and de novo assembly of P. integerrima

The paired end (2 × 150 bp) DNA libraries were produced for *P. integerrima* using Illumina HiSeq 2500. In total, 551 million reads (83 Gb) were generated (table 1 in electronic supplementary material. The high-quality reads (phred score >30) were used to predict genome size with k-mer size 21. The peak for k-mer depth vs k-mer coverage was predicted at 69. The genome size was calculated using k-mer coverage / k-mer depth, which gave a haploid genome size 452,531,485 bp (452 Mbp), where we predicted a 55.4% (251,140,882 bp) single copy region. Previous studies have shown that all pistachio cultivars are diploid with chromosome numbers $2n = 24$, 28 and 30 (Ghaffari *et al.* 2005). The closest organism of *P. integerrima* is *P. vera* L. and its chromosome size and genome size was found to be $2n$=30 and 513 Mb (Basr *et al.* 2003; Ziya *et al.* 2016).

The high-quality reads were then assembled using two software, namely SOAPdenovo2 and SPAdes. Before assembly, the best k-mer was checked using KmerGenie. Also the assembly size was predicted using different k-mers from 21 to 121 with intervals of two. Both gave the best k-mer size as 115. Among SOAPdenovo2 and SPAdes, SPAdes has given the best assembly with k-mer size 115. We obtained 205,922 contigs (>200 bp) with total assembly size 463,342,136 bp (463.3 Mb) and the largest contig size 238,122 bp. The N50 value was found to be 16,145 bp and the average size of scaffolds is 2250.1 bp. We have made an attempt of scaffolding of contig assembly using SSPACE (Boetzer *et al.* 2011) and GapClosure (Luo *et al.* 2012) using the same set of reads used for the assembly. Since the same set of reads were used, we did not find the difference in the primary contig assembly and the scaffold assembly. The assembly was further used for removal of chloroplast, mitochondria and vector sequences. After the removal of these sequences, we got an assembly of 462,064,448 bp (462 Mb) and the average contig size was found to be 2243.9 bp with largest contig size 238,122 bp (table 1). The assembled sequences were validated using benchmarking universal single-copy orthologs (BUSCO) for assembly completeness with eudicots_odb10 lineage. We searched a total 2121 BUSCO groups and obtained 1938 complete BUSCOs (1907 single copy and 31 duplicated), 108 fragmented BUSCOs and 75 missing BUSCOs. Hence 91.4% of complete BUSCOs predicted shows the good assembly (table 2 in electronic supplementary material).

Transcriptome assembly

The RNA seq data was generated for leaf sample using Illumina NextSeq500 platform (table 3 in electronic supplementary material. *De novo* transcriptome assembly was performed for leaf sample of *P. integerrima* using trinity. The best assembly resulted in 46,571 transcripts with N50 467 bp and mean transcript length as 492 bp. Also we identified 38,098 trinity genes (table 2). All the transcripts were aligned against swissprot database to identify the putative biological function. Of the 46,571 transcripts, 26,950 transcripts were aligned to swissprot database.

P. integerrima genome annotation

The final assembly was used to identify the repetitive sequences including simple sequence repeats (SSRs) and complex repeats. We have identified 206,323 SSRs and 27,214 scaffolds with more than one SSR. We classified SSRs into mono, di, tri, tetra, penta and hexa. Mono repeats were more in the assembly (152,221) followed by di (36,566), tri (14,282), tetra (2184), penta (577) and hexa (493) (figure 1 in electronic supplementary material. Complex repeat analysis showed 57.31% *P. integerrima* genome consisting of transposable elements (TEs), the majority being unclassified and long terminal repeat (LTR) retroelements. We identified a small portion of SINEs and DNA transposons (table 4 in electronic supplementary material. The identified repeats were masked for further analysis.

The repeat masked assembly was further used for protein-coding gene prediction. We predicted 39,425 gene models, of which 18,492 genes have RNA or protein evidence and 20,933 genes were *de novo* identified. Also we found that the genes with mean gene length and mean CDS length had

Table 1. *P. integerrima* genome assembly statistics.

Attributes	Assembly statistics
Total assembly size (> 200 bp)	462,029,679 bp
Mean contig / scaffold size	2253.3 bp
Largest contig / scaffolds	238,122 bp
Total contigs / scaffolds (> 200 bp)	205,044
N50	16,109 bp
L50	6843

Table 2. RNA seq assembly statistics of leaf.

Contents	Leaf
Transcripts	46,571
Unigenes	38,098
N50 (bp)	467
Mean transcript length (bp)	492.75
Assembled bases (bp)	22,947,973

2107.6 nts and 1113.5 nts, respectively. We identified average three introns per gene with average length of 327.9 nts (table 3). The predicted genes were searched against different databases such as Swissprot and Pfam. Of the 39,425 genes, 38,466 genes have homology with Swissprot database. The genes were assigned with gene ontology (GO) terms. The GO annotation revealed that genes are equally distributed among molecular function, biological process and cellular component (figure 1). Majority of genes involved in binding, catalytic activity, molecular function regulator, transporter activity, etc. under molecular function. Similarly, more genes are involved in cellular anatomical entities, protein-containing complexes under cellular components. In the case of biological processes, more genes are involved in cellular processes, metabolic processes, biological regulation, response to stimulus, localization, developmental process, growth, reproductive process, etc (figure 2 in electronic supplementary material. *P. integerrima* gene models were also searched against the Pfam database. Of the 39,425 genes, 34,850 genes have been identified with Pfam domain.

We identified 488 tRNAs in *P. integerrima* genome with an average length of 75 nts. We also identified 25 tRNAs with introns. There were 1602 small nucleolar RNA (snRNA), of which 1372 were plant-specific, 191 microRNA (miRNA), 99 spliceosomal RNA, 108 eukaryotic large subunit ribosomal RNA (rRNA) and 35 eukaryotic small subunit ribosomal RNA (rRNA) in *P. integerrima* genome (table 5 in electronic supplementary material.

Table 3. *P. integerrima* genome annotation information.

Annotation	Protein coding genes
Number	39,425
Transcript/protein evidence	18,492
Ab initio	20,933
Single-exon gene count	13612
Mean gene length (bp)	2107.6
Mean CDS length (bp)	1113.5
Mean introns per gene (bp)	3.0
Mean intron length (bp)	327.9
Mean intergenic region length (bp)	1826.3
SwissProt hits	38,426
Pfam hits	290,400

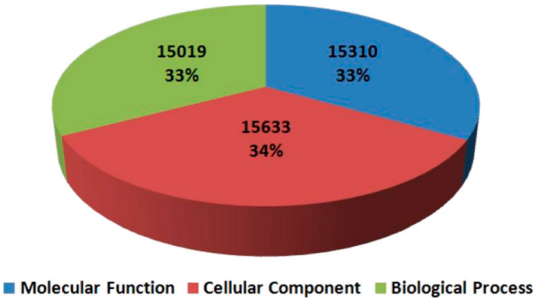

Figure 1. GO of *P. integerrima.*

Pathway analysis of P. integerrima

The gene models were aligned to KEGG database for pathway analysis. Totally, 7893 gene models were assigned to 3895 KO IDs which forms 398 pathways. The genes are categorized into three different levels. Level I includes metabolism, genetic information processing, environmental information processing, cellular processes, human diseases and organismal systems. More number of genes from *P. integerrima* are involved in metabolism (figure 2). At level II, global and overview maps involved 1997 genes followed by neurodegenerative disorder (713), signal transduction (438), carbohydrate metabolism (334) etc (figure 3). At level III, metabolic pathways, biosynthesis of secondary metabolites, microbial metabolism in diverse environments, etc are the top pathways with more genes (table 6 in electronic supplementary material. We identified 103 genes involved in biosynthesis of different secondary metabolites. The phenylpropnoid biosynthesis and flavonoid biosynthesis pathway have more number of genes involved (figure 4a). Also, we identified 92 genes involved in the metabolism of terpenoids and polyketides, where we identified terpenoid biosynthesis (31) and carotenoid biosynthesis (21) pathways (figure 4b). Terpenoids involves two different biosynthetic pathways; meva-lonic acid (MVA) in the cytoplasm and methylerythritol 4-phosphate (MEP) in the plastid. The biosynthetic pathway for terpenoids is primarily composed of three stages: the first stage involves the formation of intermediates, such as isopentenyl-P (IPP) and dimethylallyl-PP (DMAPP), which are common precursors of terpenoids; the second stage involves the synthesis of three direct precursors, geranyl-PP (GPP), farnesyl-PP (FPP), and geranyl geranyl-PP (GGPP); and the third stage involves the modification of various enzymes. All of the genes linked to MVA and MEP pathways of terpenoid backbone biosynthesis were successfully assigned (file 1 in electronic supplementary material).

Orthology detection and phylogenetic analysis of P. integerrima

The protein sequences for 39,425 genes of *P. integerrima* were compared with proteomes of five (*A. thaliana, Citrus*

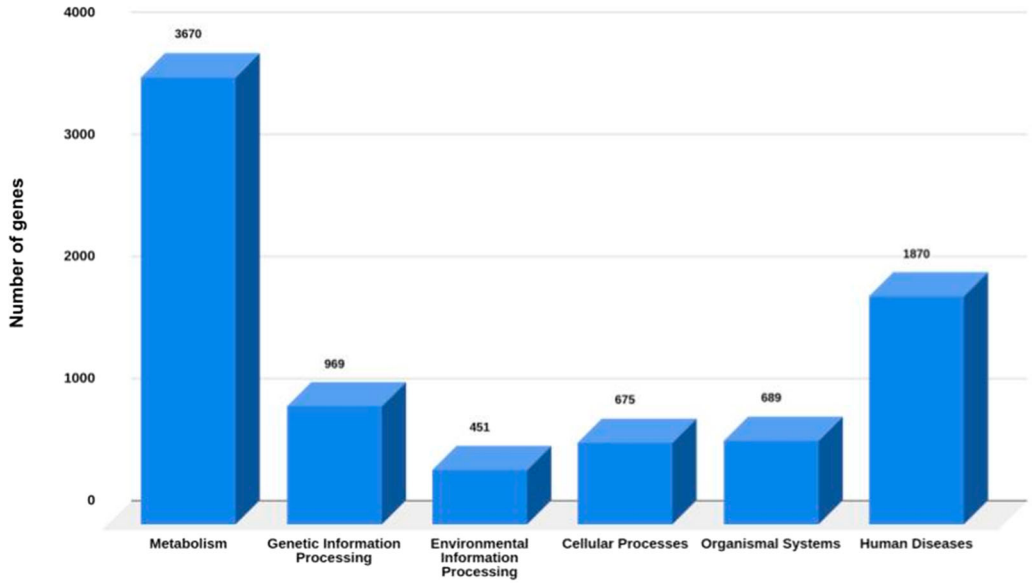

Figure 2. Predicted pathways of *P. integerrima* through predicted genes from genome. More number of genes are identified in the metabolism pathway which is followed by human diseases, genetic information processing, etc.

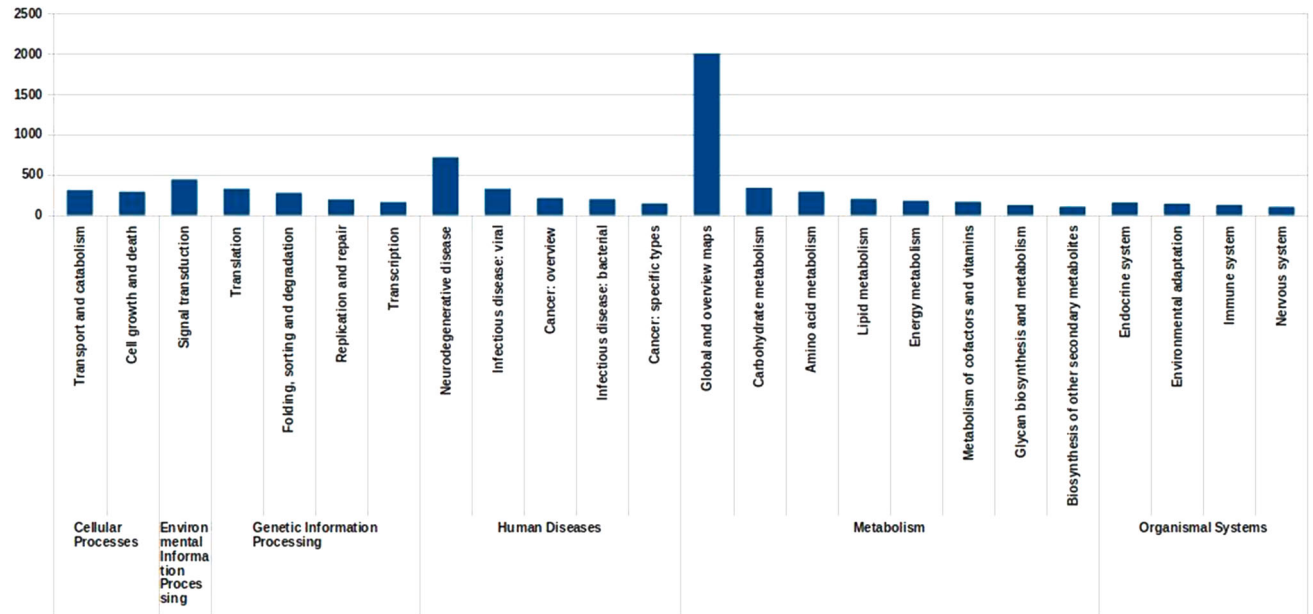

Figure 3. Characterization of *P. integerrima* genes involved in different pathways.

sinensis, P. vera, Solanum tuberosum and *Vitis vinifera)* sequenced plant species. The six species form 26,174 clusters, 23,944 orthologous clusters (at least two species) and 2230 single-copy gene clusters (table 4). *P. integerrima* forms 19,932 (76%) clusters and 10,673 singletons (proteins are not in any cluster). The singletons of *P. integerrima* were aligned to KEGG database for pathway analysis. Of the 10,673 proteins, 93 were assigned with KO IDs. More number of genes were identified in metabolic pathways and biosynthesis of secondary metabolites. There were 9065 core orthologous groups (COGs) gene families shared by all six species which includes 10,142 (13.07%) genes from *P. integerrima* (figure 3 in electronic supplementary material. The similarity matrix showed that the *P. integerrima* shares more clusters (18,325) with *P. vera* (figures 4 & 5 in electronic supplementary material. Along with the Orthovenn2, the proteome of 13 plant species were compared and a phylogenetic tree was built using proteinortho software based on similarity of the proteome among the plant species. Phylogeny among 13 species revealed that *P. integerrima* is closer to *P. vera* followed by *Solanum* species (figure 5).

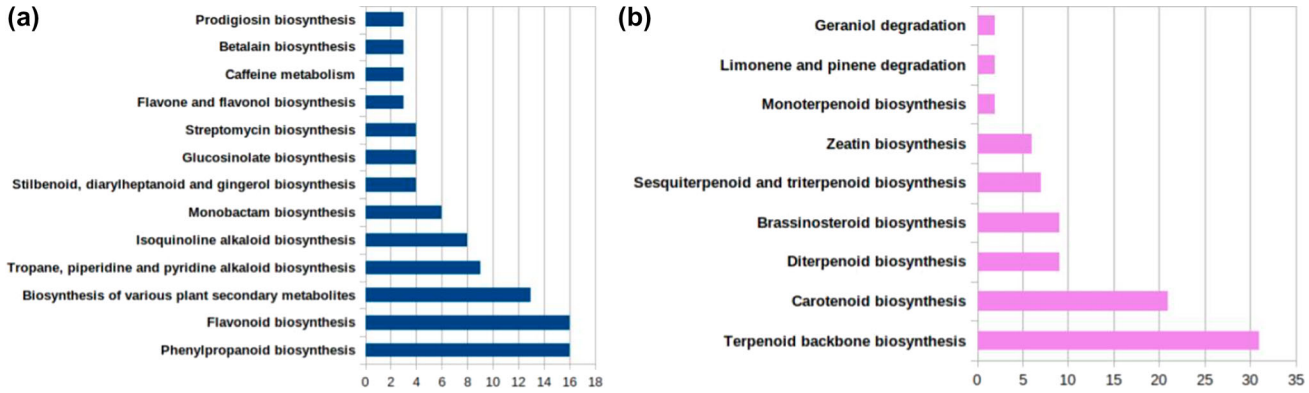

Figure 4. Pathways of biosynthesis of different secondary metabolites identified in *P. integerrima.* (a) Biosynthesis of different secondary metabolites pathway (phenylpropanoids, terpenoids, etc). (b) Biosynthesis of terpenoids.

Table 4. Summary of proteins of *P. integerrima* shared with different species.

Species	Proteins	Clusters	Singletons
Arabidopsis thaliana	27,413	14,425	5211
Solanum tuberosum	37,475	13,990	9593
Vitis vinifera	26,556	14,709	7426
P. integerrima	39,425	19,932	10,673
P. vera	41,299	19,935	7033
Citrus sinensis	39,056	17,267	2548

Discussion

Galls of *P. integerrima* are highly valued for its medicinal properties and traditionally used to treat common diseases such as cough, dyspepsia, vomiting, phthisis, asthma and dysentery (Chopra *et al.* 1986; Aggarwal *et al.* 2006; Munir *et al.* 2011). The drastic decrease in the gall formation in recent decades is possibly due to climate change or human interventions, making it an important species from conservation point of view. Hence, the relevance of our attempt to generate whole genome data and other genetic resources for *P. integerrima* will have massive significance. The development in NGS has enabled many scientists to study extensively, mainly in nonmodel species (Lu *et al.* 2016). The genome and transcriptome sequencing of *P. integerrima* is essential for gaining insights into its genetic information, and functional analysis of pathways associated with the secondary metabolites production responsible for its medicinal properties. Here, for the first time, we made an effort to sequence the whole genome of *P. integerrima*.

Many of the genome assemblers use de Bruijn graph framework, where reads are chopped up into *k*-mers (substrings of length *k*) and the size of these *k*-mers is the most significant parameter (Chikhi and Medvedev 2014). Here we used different k-mers (21–121, with an interval of two) to obtain the best k-mer size by assessing the assembly size for each k-mer. Also we used two different assemblers, namely SPAdes and SOAPdenovo2, with the k-mer size 115 to get the best assembly using short reads. Further, the k-mer technique has been successfully used to estimate genome size using NGS reads without any prior information of genome size (Lu *et al.* 2016). Here, for the first time, we report the genome size of *P. integerrima.* The k-mer method predicted the genome size of *P. integerrima* is about 450 Mb, which is close to the genome size of *P. vera* L. (513 Mb) (Basr *et al.* 2003; Ziya *et al.* 2016). The k-mer analysis also revealed that *P. integerrima* has a high amount of heterozygosity, which is likely attributable to the genus dicotyledonous mating mechanism.

For the prediction of more accurate gene models, it is necessary to use a comprehensive approach that involves information intrinsic to the genome sequence (*ab initio*) and any extrinsic data such as proteins, transcripts (Haas *et al.* 2008). Here, we used the integration of three approaches including *ab initio*, homology-based and evidence-based methods (RNA seq) to find the high accurate gene models (figure 6 in electronic supplementary material. The genes were functionally annotated using Swissprot database and 97% of the predicted genes were functionally annotated. Alkaloids, flavonoids, tannins, saponins, sterols, and essential oils are among the phytochemicals found in *P. integerrima*. *P. integerrima* leaf phytochemical examination revealed the presence of carotenoids, triterpenoids, catechins, and flavonoids (Bibi *et al.* 2015). Pathway analysis of *P. integerrima* showed that more number of genes were involved in metbolic pathways and biosynthesis of secondary metabolites genes (table 6 in electronic supplementary material. Among them, number of genes involved in phenylpropanoid biosynthesis and flavonoids pathway was observed to be more (figure 5). The metabolism of phenylpropanoid produces a large number of secondary metabolites such as lignin or flavonoid and also all elements of plant responses to biotic and abiotic stimuli are influenced by phenylpropanoids (Vogt 2010). Terpenoids are a group of natural chemicals with a wide range of structures and biological roles (for example, tocopherol, brassinolide, and gibberellin are involved in cell development and defence) (Wang *et al.* 2019). We used KEGG analysis to find most of

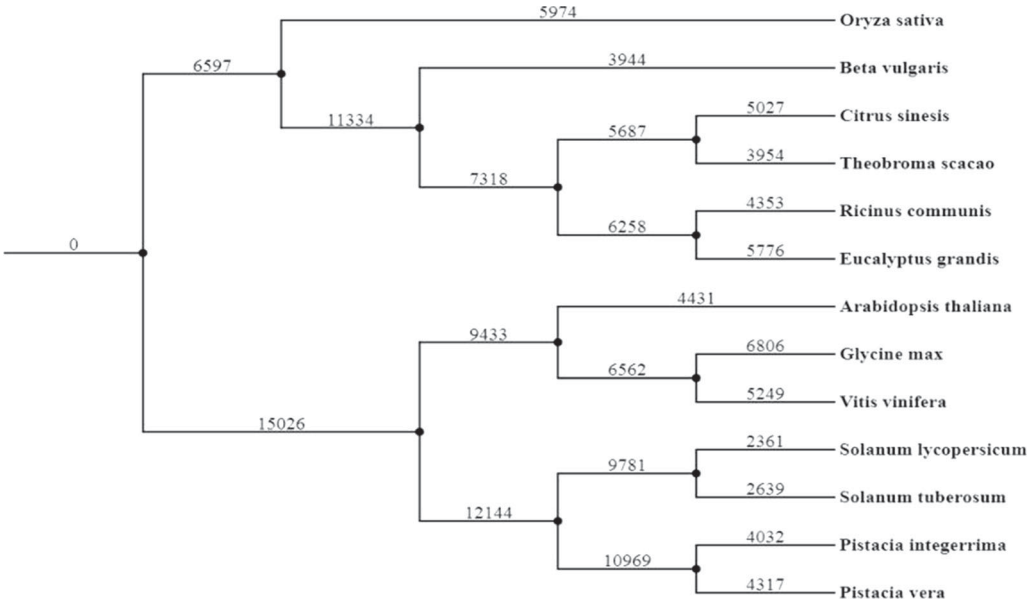

Figure 5. Phylogenetic tree of *P. integerrima* with other species. *P. integerrima* is more closer to *P. vera* which is another species of the same genus *Pistacia.*

the genes expected to encode terpenoid-backbone biosynthesis enzymes among various significant secondary metabolites in *P. integerrima*. The FPP is the principal precursor for both primary (sterols, dolichols, ubiquinone, brassinosteroids, and protein prenylation) and secondary (triterpenes) terpene metabolism, and genes expected to encode FPP synthesising enzymes have been effectively mined. Similarly, genes for GGPS, which contributes to the synthesis of GGPP, a key precursor for both primary and specialised isoprenoid chemicals (carotenoids, GA, ABA, tocopherol, diterpenes) were identified.

Noncoding RNAs (ncRNAs), such as microRNAs (miRNAs), small interfering RNAs (siRNAs), and long noncoding RNAs (lncRNAs), play vital regulatory functions in plant growth and secondary metabolism, as well as biotic and abiotic stress responses. More than 30 miRNAs and five lncRNAs have been predicted to regulate bioactive compound production through various regulatory modules and pathways (Li *et al.* 2021). We identified miRNAs from miR397 family which mainly target the laccase (LAC) genes functioning in lignin synthesis (Huang *et al.* 2021) and miR828, target coding sequences of specific helix motifs in the mRNA sequences of MYB proteins which cause decay of MYB RNA and the production of a cascade of secondary siRNAs that depend on RNA-dependent RNA polymerase 6 which lead to the promote anthocyanin biosynthesis (Tirumalai *et al.* 2019).

Conclusion

In this study, we present a draft genome of *P. integerrima*, a medicinal plant known to produce galls. This is the first genome study of *P. integerrima*. Pathway information from this study aid in the exploration of biosynthesis and metabolism of various secondary metabolites. This study will also help with comparative genomic research and serve as a reference for future sequencing studies of this species. It will also aid in phylogenetic analysis and evolutionary investigations for this species. The availability of genomic information from this research is likely to allow the detection of alleles, genetic mapping and recognition of candidate genes involved in gall formation.

This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession JAGTWX000000000. The raw sequence reads of whole genome and RNA sequencing is deposited in NCBI SRA database under bioproject PRJNA480376.

Acknowledgements

Authors acknowledge Dr N. B. Brindavanam for his support at every stage in this study, right from inception; Bengaluru Genomics Centre for their support in sequencing and data analysis; Dabur India Ltd. and Dr Sasibhushan Vedula for the support at various strata.

Authors' contributions

MG designed field experiments and data generation and analysis. PN and MG were involved in sample collection. NB and AB were involved in authentic sample collection. MG and PV was involved in initiating and heading the project. SNH was involved in data analysis, interpretation and drafting manuscript. MG, PN, SK, NB and PV were involved in editing the manuscript and providing inputs.

References

Aggarwal B. B., Ichikawa H., Garodia P., Weerasinghe P., Sethi G., Bhatt I. D. *et al.* 2006 From traditional Ayurvedic medicine to modern medicine: Identification of therapeutic targets for

suppression of inflammation and cancer. *Expert Opin. Ther. Targets* **10**, 87–118.
Andrews S. 2010 FastQC: a quality control tool for high throughput sequence data (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/).
Bankevich A., Nurk S., Antipov D., Gurevich A. A., Dvorkin M., Kulikov A. S. *et al*. 2012 SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477.
Bao W., Kojima K. K. and Kohany O. 2015 Repbase update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 11.
Barinder K. and Saurabh S. 2015 A review on gall karkatshringi. *J. Med. Plants Res.* **9**, 636–640.
Basr I. H., Kafkas S. and Topaktas M. 2003 Chromosome numbers of four Pistacia (Anacardiaceae) species. *J. Hortic. Sci. Biotechnol.* **78**, 35–38.
Bi Q., Zhao Y., Du W., Lu Y., Gui L., Zheng Z. *et al*. 2019 Pseudomolecule-level assembly of the Chinese oil tree yellowhorn (*Xanthoceras sorbifolium*) genome. *Gigascience* **8**, 1–11.
Bibi Y., Zia M. and Qayyum A. 2015 Review-An overview of *Pistacia integerrima* a medicinal plant species: Ethnobotany, biological activities and phytochemistry. *Pak. J. Pharm. Sci.* **28**, 1009–1013.
Blanco E., Parra G. and Guigó R. 2007 Using geneid to Identify Genes. In *Current protocols in bioinformatics*. **4,** unit 4.3.
Boeckmann B., Bairoch A., Apweiler R., Blatter M., Estreicher A., Gasteiger E. *et al*. 2003 The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 365–370.
Boetzer M., Henkel C. V., Jansen H. J. and Butler D. 2011 Scaffolding pre-assembled contigs using SSPACE Summary. *Bioinformatics (Oxford, England).* **27**, 578–579.
Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., Bealer K. *et al*. 2009 BLAST+: architecture and applications. *BMC Bioinformatics* **9**, 1–9.
Chan P. and Lowe T. 2019 tRNAscan-SE: searching for tRNA genes. *Gene Predict.* **1962**, 1–21.
Chikhi R. and Medvedev P. 2014 Informed and automated k-mer size selection for genome assembly. *Bioinformatics* **30**, 31–37.
Chopra R. N., Nayar S. L. and Chopra I. 1986 *Glossary of Indian medicinal plants* (Including the Supplement). Council of Scientific and Industrial Research, New Delhi.
Ghaffari S. M., Shabazaz M. and Behboodi B. S. 2005 Chromosome variation in Pistacia genus. *Options Mediterraneennes Serie A.* **63**, 347–354.
Grabherr M. G., Haas B. J., Yassour M., Levin J. Z., Thompson D. A., Amit I. *et al*. 2011 Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652.
Griffiths-Jones S., Bateman A., Marshall M., Khanna A. and Eddy S. R. 2003 Rfam: an RNA family database. *Nucleic Acids Res.* **31**, 439–441.
Haas B. J., Delcher A. L., Mount S. M., Wortman Jr. J. R. *et al*. 2003 Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666.
Haas B. J., Salzberg S. L., Zhu W., Pertea M., Allen J. E., Orvis J. *et al*. 2008 Automated eukaryotic gene structure annotation using evidence modeler and the program to assemble spliced alignments. *Genome Biol.* **9**, 1–22.
Huang S., Zhou J., Gao L. and Tang Y. 2021 Plant miR397 and its functions. *Funct. Plant Biol.* **48**, 361–370.
Keilwagen J., Wenk M., Erickson J. L., Schattat M. H., Grau J. and Hartung F. 2016 Using intron position conservation for homology-based gene prediction. *Nucleic Acids Res.* **44**, e89.
Kent W. J. 2002 BLAT - The BLAST -like alignment tool. *Genome Res.* **12**, 656–664.
Kim D., Pertea G., Trapnell C., Pimentel H., Kelley R. and Salzberg S. L. 2013 TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36.
Korf I. 2004 Gene finding in novel genomes. *BMC Bioinformatics* **9**, 1–9.
Krueger F. 2015 Trim galore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/).
Lechner M., Findeiß S., Steiner L., Marz M., Stadler P. F. and Prohaska S. J. 2011 Proteinortho: detection of (Co-) orthologs in large-scale analysis. *BMC Bioinformatics* **12**, 124.
Li C., Wang M., Qiu X., Zhou H. and Lu S. 2021 Noncoding RNAs in medicinal plants and their regulatory roles in bioactive compound production. *Curr. Pharm. Biotechnol.* **22**, 341–359.
Lu M., An H. and Li L. 2016 Genome survey sequencing for the characterization of the genetic background of rosa roxburghii tratt and leaf ascorbate metabolism genes. *PLoS One* **11**, 1–17.
Luo R., Liu B., Xie Y., Li Z., Huang W., Yuan J. *et al*. 2012 SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience.* **1**, 2047-217X-1-18.
Majoros W. H., Pertea M. and Salzberg S. L. 2004 TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics (Oxford, England)* **20**, 2878–2879.
Marçais G. and Kingsford C. 2011 A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics.* **27**, 764–770.
Mistry J., Chuguransky S., Williams L., Qureshi M., Salazar G. A., Sonnhammer E. L. L. *et al*. 2021 Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419.
Moriya Y., Itoh M., Okuda S., Yoshizawa A. C. and Kanehisa M. 2007 KAAS: An automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.* **35**, 182–185.
Munir M., Khan M. A., Ahmed M., Bano A., Ahmed S. N., Tariq K. *et al*. 2011 Foliar epidermal anatomy of some ethnobotanically important species of wild edible fruits of northern Pakistan. *J. Med. Plants Res.* **5**, 5873–5880.
Nawrocki E. P. and Eddy S. R. 2013 Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935.
Novaes R. M. L., Rodrigues J. G. and Lovato M. B. 2009 An efficient protocol for tissue sampling and DNA isolation from the stem bark of Leguminosae trees. *Genet. Mol. Res.* **8**, 86–96.
Orwa C., Mutua A., Kindt R., Jamnadass R. and Simons A. 2009 Agroforestree database: a tree reference and selection guide version 4.0. *World Agroforestry Centre*, Kenya. (http://apps.worldagroforestry.org/treedb2/).
Pant S. and Samant S. S. 2010 Ethnobotanical observations in the mornaula reserve forest of Komoun, West Himalaya, India. *Ethnobot. Leafl.* **14**, 193–217.
Rauf A. 2019 A Mini Review on a *Pistacia integerrima* well-known medicinal plant: its active phytochemicals with exciting pharmacological profile. *Act. Sci. Nutr. Health.* **3**, 45–48.
Rauf A., Saleem M., Uddin G., Siddiqui B. S., Khan H., Raza M. *et al*. 2015 Phosphodiesterase-1 Inhibitory Activity of Two Flavonoids Isolated from *Pistacia integerrima* J. L. Stewart Galls. *Evid. Based Complement Alternat. Med.* **2015**, 506564.
Sima F. A., Waterhouse R. M., Ioannidis P., Kriventseva E. V. and Zdobnov E. M. 2015 Genome analysis BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212.
Stanke M. and Stephan W. 2003 Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics (Oxford, England).* **19**, 215–225.
Thiel T., Michalek W., Varshney R. K. and Graner A. 2003 Exploiting EST databases for the development and

characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theor. Appl. Genet.* **106**, 411–422.

Tirumalai V., Swetha C., Nair A., Pandit A. and Shivaprasad P. V. 2019 MiR828 and miR858 regulate VvMYB114 to promote anthocyanin and flavonol accumulation in grapes. *J. Exp. Bot.* **70**, 4775–4791.

Trapnell C., Williams B. A., Pertea G., Mortazavi A., Kwan G., van Baren M. J. *et al*. 2010 Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol.* **28**, 511–515.

Ullah Z., Mehmood R., Imran M., Malikb A. and Afzal R. A. 2012 Flavonoid constituents of *Pistacia integerrima*. *Nat Prod Commun.* **7**, 1011–1014.

Vogt T. 2010 Phenylpropanoid biosynthesis. *Mol. Plant* **3**, 2–20.

Wang Q., Quan S. and Xiao H. 2019 Towards efficient terpenoid biosynthesis: manipulating IPP and DMAPP supply. *Bioresour. Bioprocess* **6**, 6.

Xu L., Dong Z., Fang L., Luo Y., Wei Z., Guo H. *et al*. 2019 OrthoVenn2: a web server for whole-genome comparison and annotation of orthologous clusters across multiple species. *Nucleic Acids Res.* **47**, 52–58.

Zahoor M., Zafar R. and Rahman N. U. 2018 Isolation and identification of phenolic antioxidants from *Pistacia integerrima* gall and their anticholine esterase activities. *Heliyon* **4**, e01007.

Zhang Y., Zheng L., Zheng Y., Zhou C., Huang P., Xiao X. *et al*. 2019 Assembly and annotation of a draft genome of the medicinal plant *Polygonum cuspidatum*. *Front. Plant Sci.* **10**, 1274.

Ziya M. E., Kafkas S., Khodaeiaminjan M., Çoban N. and Gözel H. 2016 Genome survey of pistachio (*Pistacia vera* L.) by next generation sequencing: development of novel SSR markers and genetic diversity in Pistacia species. *BMC Genomics* **17**, 998.

Corresponding editor: Durgadas P. Kasbekar