Integrated omics approach to understand *Pistacia* - aphid gall development

Pavithra Narendran
The University of Trans-disciplinary Health Sciences and Technology https://orcid.org/0000-0003-1613-5340
Noorunnisa Begum
The University of Trans-disciplinary Health Sciences and Technology
Amit Bhatt
Dabur India Ltd
Subrahmanya Kumar Kukkupuni
The University of Trans-disciplinary Health Sciences and Technology
Padma Venkatasubramanian
SRM Institute of Science and Technology
Thottethodi Subrahmanya Keshava Prasad
Yenepoya (Deemed to be University)
Mohd Altaf Najar
Yenepoya (Deemed to be University)
Kallare P Arunkumar
Central Muga Eri Research and Training Institute
Sastry JLN
Dabur India Ltd
Badrinarayan S
Dabur India Ltd
Malali Gowda (heredityd ecoder@gmail.com)

Research article

Keywords: Traditional medicine, Ayurveda, Host, Peptides, Karkatashringi, Baizongia pistaciae L

Posted Date: January 24th, 2020

DOI: https://doi.org/10.21203/rs.2.21772/v1

Abstract

Background: *Pistacia chinensis subsp. integerrima* (J. L. Stewart ex Brandis) Rech. f. belongs to family Anacardiaceae, is well known for usage of galls in traditional medicine. The galls in the apical meristem of plants are induced by aphid *Baizongia pistaciae* L infestation. The Pistacia galls are used in Ayurvedic formulations for the treatment of cough and respiratory diseases, loss of appetite, dyspeptic vomiting and dysentery. 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. There is not much work carried out at molecular level to understand the gall development. In this study, we obtained molecular insight on insect and plant through genomic, transcriptomic and proteomic approaches. Results: We sequenced the whole genome of *Pistacia* genome using Illumina sequencing platform. The assembly genome size of *Pistacia* was 549 Mb with 80688 scaffolds at N50 = 12607 nts. A total of 51290 genes were from *Pistacia* genome annotation. The transcriptome analysis revealed of 76186 and 46327 transcripts from gall and leaf, respectively. The ethylene responsive transcription factor families were induced abundantly in galls including GATA, bHLH, MYB, BZIP, Trihelix, MADS, B3 domain. In addition, we have also obtained highly expressed genes in gall for biosynthesis of secondary metabolites, plant-aphid interactions, stress responses, phytohormone signal transduction and terpene biosynthesis. We have identified 21 proteins against *Cajanus cajan* , 10 proteins against Arabidopsis thaliana and 14 proteins against *Drosophila melanogaster* . We also identified abundant peptides for actin and tubulin of aphid in the gall tissue. Conclusion: This study provides a basis for the understanding of the genes expressed in gall when compared to leaf and also genes required for gall induction and development. We hypothesed that actin and microtubule proteins are from aphids origin as we found that transcript TRINITY_DN180182_c0_g1_i1 mapped to *Acyrthosiphon pisum* tubulin beta-1 (Tub1), mRNA with 75 % homology.

Background

The coexistence of plants and insects has been for more than 350 million years and they have developed a strategy to overcome each other’s defence system (1). Pistacia chinensis subsp. integerrima (J. L. Stewart ex Brandis) Rech. f. (here after called as Pistacia) is well known for formation of galls on apical buds. The Pistacia plant belongs to the family Anacardiaceae and is a dioecious tree native to Asia. Pistacia is widely distributed in East Afghanistan, Pakistan, North West & West Himalaya (2). These galls are horn shaped and formed due to infestation of insect Baizongia pistaciae L., which follows two year life cycle. These galls are called by different names such as karkatashringi, kakroi, kakring, kakra, kakkar and kakarsinghi (3). The galls are of source of secondary metabolites which is used for the treatment of cough and respiratory ailments, loss of appetite, dyspeptic vomiting, and dysentery. Karkatashringi is one of the ingredients in many ayurvedic formulations such as shringiadi chura, karkatadi chura, balachaturbhadra churna, brihat talisadi churna, devadarvayadi kwatha churna, shatavaryadi ghrita, chayanprasha avaleha, dashmularista, siva gutika, khadiradi gutika etc (4). The estimated annual demand for Karkatashringi is about 150–200 MT in India (5).

The molecular mechanisms that are responsible for induction and development of gall is not known. The gall requirement is high, but supply has reduced dramatically. When timing of gall-aphid mortality have significantly reduce the quality of gall development (6, 7). There is no genomic information available for Pistacia. In this study, we aimed to understand the molecular mechanisms of gall development (plant-insect) through the genome, transcriptome and proteome analysis of Pistacia galls. These molecular resources will help in understanding of Pistacia genes and pathways responsible for gall formation.

Results

Genome assembly

Pistacia Whole genome was sequenced using HiSeq 2500 with paired end (2X150nts) chemistry. A total of 543.6 million reads with 82.5 Gb Illumina was generated. The contigs (≥ 500 nts) of 85633, scaffolds without gapfilling of 81143 and scaffolds without gapfilling of 80688 were obatined. The N50 contig length was 10240 nts, N50 scaffold was 12607 nts. The total genome assembly length of scaffolds was 386.8 Mb based on our analysis. The GC % of Pistacia was 35.82 (**Table 1**). To test the genome assembly completeness, BUSCO was used. BUSCO analysis represented 90.9% (1929 out of 2121) single-copy orthologs in the genome (**Table 2**). The core eukaryotic genes mapped to Pistacia genome is described in **Additional file 1**. As about 90 % of the core eukaryotic genes being mapped to the Pistacia contigs with more than 95% homology which implies that genome assembly is complete and sequencing quality was good.

Genome annotation

We predicted 51290 genes from Pistacia genome assembly (**Additional File 2**), and 46591 genes were functionally annotated using Uniprot database. The Gene Ontology (GO) method was used to classify Pistacia genes into three categories: molecular function (16513 genes), cellular component (16483 genes) and biological process (16081) in **Figure 1**. We mined genes for binding (12142 genes) and catalaytic activity (9957 genes) which were highly represented in the genome. For biological function, 15059 genes for cellular component and 12791 genes for cellular process 10096 genes for metabolic process were identified from this study. Further we classified genes responsible for enzymes (**Figure 2)**, among these oxidoreductases (1358), transferases (3442), hydrolases (2494), lyases (518), isomerases (361), ligases (615) and translocases (233) were annotated. This study is of its first kind where we identified metabolic as well as enzyme classes in Pistacia.

Transcriptome analysis

The gall and leaf samples of *Pistacia* were collected from different locations in Uttarakhand. The RNA from leaf and gall samples were sequenced Illumina sequencing with paired-end reads chemistry (2x150 nts). We obtained gall ~37.3 and 33.1 million for gall and leaf samples, respectively (**Table 3**). The GC percentage was varied from 55.57–61.70% in these tissues. Illumina short reads were assembled for each tissue and also carried out combined assembly using Trinity for differential gene expression

analysis. Assembly yielded 76186 gall and 46327 transcripts from gall and leaf respectively. We observed N50 of 467 and 448 bases for gall and leaf, respectively. We identified 62801 transcripts that are specific to gall, 34300 to leaf and 13385 were commonly present in both the tissues (**Figure 3**).

BLAST2GO program was used to identify the enzymatic classes in gall and leaf tissues. About six classes; oxidoreductases (2017 transcripts), transferases (2955), hydrolases (3940), lyases (745), isomerases (546) and ligases (896) were highly expressed in the gall when compared to leaf (**Figure 4)**. The Gene Ontology (GO) distribution in gall and leaf were represented in the **Figure 5**.

Transcription factors

The potential transcription factors such as ethylene responsive, GATA, bHLH, MYB, BZIP, Trihelix, MADS, B3 domain, nitrogen assimilation, WRKY, steA, UNE12, NAC, MIZ, AS1, C6, fungal specific domain and GRAS are the most abundant transcription factor families identified in gall. Ethylene-responsive, CarD, bHLH, TCP, NAC domain, myb, pca, POU domain and Fe-S cluster assembly found to be identified from leaf. Ethylene responsive (32 transcripts) was found to be highly expressed in gall when compared to leaf (**Figure 6**). In the present study, we identified the putative genes (62801 transcripts) that are specifically expressed in gall (**Additional File 3**). GA receptor and scarecrow-like protein 8 - DELLA proteins were highly expressed in gall with an FPKM value of 9.28 and 26.97 respectively (**Table 4**). We identified 535 genes that were differentially expressed between gall and leaf tissues (**Additional File 4**). Among these genes coding for biosynthesis of secondary metabolites, plant-aphid interactions, stress responses, phytohormone signal transduction and terpene biosynthesis were highly expressed in gall than in leaf (**Table 5**).

Transcription factors from gall transcripts were only expressed in gall compared to leaf (**Table 6**). Gibberellin receptor GID1B and DELLA proteins have functional role in gall formation (8). They have been elucidated in the signal transduction pathway of Giberellic acid (GA) (**Table 7**).

Proteins analysis of gall tissue

Dry gall contains plant and aphid tissues that are used in preparation Ayurvedic formulations. Peptide sequencing of dry gall powder was done. Among these, 84 peptides matched to 21 proteins of *Cajanaus cajan* (**Table 8**) and 13 peptides matched to 10 proteins of *Arabidopsis thaliana* with **(Table 9)** and 34 peptides matched to 14 proteins against *Drosophila melanogaster* **(Table 10).** Interestingly, tubulin peptides from dry gall mapped to proteins of pea aphid and *D. melanogaster* **(Figure 7).** This observation shown that tubulin proteins from the dead skeletal aphids in dry gall. When we further analysed RNA transcripts against peptides, we found that TRINITY_DN191393_c0_g1_i1 matched to actin 5C, isoform D (*Drosophila melanogaster*) with 77.6% homology whereas transcript TRINITY_DN180182_c0_g1_i1 from transcriptome mapped to *Acyrthosiphon pisum* tubulin beta-1 (Tub1), mRNA with 75 % homology. This indicates that actin and tubulin from transcriptome as well as determined for peptides are specific to insects. We found the homology for plant tubulin (*Vitis vinifera*) with about 83 % and fungal tubulin

(*Bortyris cinerea*) with about 76 %. Even though these proteins are present in plant as well as fungus, we found the peptide isolated form gall mapped to high homology to aphids.

Simple Sequence Repeat Markers

A total of 231624 SSRs were identified (**Table 11**) from *Pistacia* genome. As expected, mononucleotide motifs were the most abundant type of repeats (74.1%) in genome of Pistacia followed by di- (17.3%), tri- (6.9%) (**Figure 8)**. The SSR distribution types are shown in **Additional File 5**. The most abundant repeat types were (A/T) which accounted to 96.25% followed by AT/AT repeats of 56.52 %.

Discussion

Modified plant structures are termed as galls that are induced by other organisms such as bacteria, viruses, and insects. Gall forming insects often obtain easy food access and gain protection from harsh environmental conditions and natural enemies. Gall formation is a highly adaptive and evolutionary advantageous phenomenon (9). Pistacia from the Anacardiaceae family known for its leaf galls induced by Baizongia pistaciae L are of ethno medicinal importance. Pistacia gall has the ability to resolve several health issues like asthma, dysentery, psoriasis, snakebite etc (10). A very little is known about the mechanism of gall formation and the way insect and plant recruit plant biochemical, physiological and developmental pathways for their own benefit.

Next Generation Sequencing technology was applied to sequence non-model species such as Pistacia. The estimated genome size of Pistacia is ~ 549 MB based on whole genome assembly. The SSR (di, tri, and tetra) predicted from Pistacia genome. SSRs are useful tools to identify genetic hotspot and diversity in natural populations, gene bank. These markers will be useful in fingerprinting and authentication of Pistacia Ayurvedic formulations. They can be used for genetic linkage mapping and evolutionary studies and have a high level of transferability between closely related species (11).

From the transcriptome analysis, we found that organic substrate metabolic process (14338 transcripts) representing biological process was higher in gall compared to leaf (5932 transcripts). Similarly intrinsic component of membrane (8096 transcripts) was higher in gall representing cellular component. Higher organic cyclic binding (11367 transcripts) were higher molecular functions in gall when compared to leaf (4982 transcripts). We also found an abundance of enzyme types in gall when compared to leaf. As enzymes are associated with various biological processes, the identification of important enzyme codes might provide us information on important functional pathways and metabolic activities between gall and leaf tissue of Pistacia (12, 13). These results suggest that higher metabolic function and enzyme activity in gall when compared to leaf.

In our studies we found that genes coding for terpene synthesis such as beta-amyrin 28-oxidase (72%), hydroxymethylglutaryl-CoA synthase, Farnesyl pyrophosphate synthase, beta-amyrin synthase 1-like, gamma-terpinene synthase and squalene synthase were highly expressed in gall tissue as compared to leaf. Among these, terpene synthesis genes such as geranylgeranyl pyrophosphate synthase family

protein and geranyl diphosphate synthase are highly expressed in gall compared to leaf. It was reported that Baizgonia pistaciae - induced galls showed high accumulation of monoterpenes levels (14).

Terpenes being the most structurally diverse class of plant metabolites play a vital role as pollinator attractants, growth regulators, and stabilizers of membrane structure as well as in mediating direct and indirect plant defences. The key enzymes responsible for monoterpene formation in plants are monoterpene synthases. In the previous studies it has been reported that total monoterpene synthase activity is tenfold higher in galls as compared to leaves. The biosynthesis machinery has helped aphids to manipulate the enzymatic system of their host plant and recruit for their own benefit (9).

We observed ethylene responsive, GATA, bHLH, MYB, BZIP, Trihelix, MADS, B3 domain, nitrogen assimilation, WRKY, steA, UNE12, NAC, MIZ, AS1, C6, fungal specific domain and GRAS transcription factor families identified in the gall tissue. Studies WHO have been reported that transcription factors play an important role in regulation of metabolic process. MYB proteins are associated with regulation of secondary metabolism, control of cellular morphogenesis and regulation of meristem formation and the cell cycle (15). The bHLH (basic helix-loop-helix) family have been reported in controlling cell proliferation and development of specific cell lineage (16).

We have also found major facilitator superfamily (MFS) transcripts, multidrug transporter and ABC transporter permease. Environmental stresses are the major factors that contribute to plant growth and development. It has been found that many transporter families of plants participate in responding to pesticides, toxic chemicals and infection of pathogens. The transporters from plants are classified into five families such as ATP binding cassette (ABC), multidrug and toxic compound exporters (MATE), small multidrug resistance (SMR), resistance-nodulation-division proteins (RND), and major facilitator superfamily (MFS). Multidrug transporters are the group of membrane proteins in MFS known to have a role in the regulatory network of responding to the stress by which they act as drug/H + transporters to reduce the accumulation of toxin in the body (17).

In this study we also observed that GA receptor and scarecrow-like protein 8 - DELLA proteins have highly expressed in gall. Wang et al 2016., have reported that exogenous Gibberellic Acid (GA) application promote gall development. Giberellin receptors and DELLA proteins have been reported to have major role in signal transduction of GA in plant development (8).

In the current study identified serine/threonine-protein kinase transcripts highly expressed in the gall tissue. Gall development through aphid infestation on Pistacia apical meristem is complex physiological process which involves active attack by aphids, passive defence of plant, confrontation between the host and the aphid, and mutual coordination between the two. Therefore it is expected that transcripts linked to plant-aphid interaction would be enriched in the gall tissue. It has been reported WHO that protein phosphorylation is one of the most important events in the disease resistance pathway. Serine/threonine-protein kinases in plant cells are referred to ‘central processor unit’, as they receive information from receptors in terms of environmental stress, phytohormones, and other external factors, and bring a changes in metabolism, gene expression, and cell growth and division (18).

Pistacia gall has been used in Ayurveda formulation. White cottony layer present inside the gall may be the insect excretion and might also be the possible metabolites synthesised from plant-insect interaction. In order to identify whether any insect proteins were present inside the gall, the white cottony layer inside gall was analysed for proteins. We also found that one of the transcripts in RNA seq data was also mapped to tubulin of pea aphid (GCA_000142985.2Acyr_2.0). Tubulin is known to express differently at different stages of development in insects. Tubulins are the building blocks of microtubules and these tubulins have major role in maintaining cell shape, movement and intracellular transport of substances. Tubulins are the essential components of cytoskeleton and spindles (19). Therefore, the study on tubulin and its information helps to understand the expression analysis of the other genes in Pistacia. We hypothesize that the actin filaments and microtubules are secreted from the insects and deposited in the galls. In the future, it would be interesting to profile metabolites from the gall and elucidate the molecular interactions. This study provides a basis for the understanding of the genes expressed in gall when compared to leaf and also genes required for gall induction and development.

Conclusion

We had approximately 130 X coverage data for Pistacia genome and estimated genome size was 549 Mb. We found a total of 51290 genes and 231624 SSRs in Pistacia genome. Transcriptome sequencing and de-novo assembly of transcriptome of gall and leaf from Pistacia generated 76186 transcripts from gall and 46327 from leaf. Transcription factors and enzyme codes in gall and leaf identified from this study. We have identified differentially expressed genes coding for biosynthesis of secondary metabolites, plant-aphid interactions, stress responses, phytohormone signal transduction and terpene biosynthesis that are highly expressed in gall when compared to leaf. Peptide analysis from dry gall of Pistacia has identified most abundant aphid peptides such as actin and tubulin. This is the first multi-omic study that used to identify differentially expressed genes in gall of Pistacia.

Methods

Sample Collection

The gall and leaf samples of *Pistacia* were collected from different locations in Uttarakhand with the guidance of expert botanists and taxonomists. The plant was identified by Dr. Noorunnisa Begum, TDU. Plant sample was immediately rinsed with normal saline and stored in RNA Later to prevent degradation. These samples were shipped to lab and stored at -80°C freezer. Voucher specimens of the gall and leaf were deposited at the National Herbarium and Raw Drug Repository at *Foundation for Revitalisation of Local Health Traditions* (*FRLHT*), Bangalore.

DNA and RNA isolation

Genomic DNA was extracted using CTAB method (20). Genomic DNA quality was checked using gel electrophoresis and Nanodrop. RNA was extracted from samples using Bengaluru Genomics Centre

(Bengaluru) company in-house protocol. The quality of RNA was checked on Bioanalyzer and quantified using QUBIT dsRNA HS kit.

DNA Library Preparation and Sequencing

The genomic DNA was enzymatically fragmented to generate ~ 250bp fragments. About 50ng of fragmented DNA was used to generate sequencing library using NEBNext® Ultra™ II DNA FS Library Prep Kit for Illumina. The library was quantitated 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 HiSeq 2500 platform.

RNA Library Preparation and Sequencing

Total RNA was extracted from 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 0.2N 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 Assembly and Annotation

The quality of raw reads of Illumina sequencing was checked for the ambiguous bases, Phred score >Q30, read length, nucleotide base content and other parameters using FASTQC. Adapter and overrepresented sequences was removed using TrimGalore software. High quality paired end reads were assembled using The SPAdes assembler (21). Contigs were further scaffolded using SSPACE (22), and gaps in the scaffolds were closed by the GapFiller module (23). Scaffolds were then subjected to Benchmarking Universal Single-Copy Orthologs (BUSCO) to assess genome completeness (24). SSR markers were identified from the assembled sequences using MIcroSAtellite tool (MISA) (http://pgrc.ipk-gatersleben.de/misa/) as per the protocol described by Kuravadi et al (25). AUGUSTUS software was used for gene prediction (26). Functional annotation of genes was carried out using Uniprot database.

Transcriptome Assembly and Annotation

The quality of raw reads of Illumina sequencing was checked for the ambiguous bases, Phred score >Q30, read length, nucleotide base content and other parameters using FASTQC. Adapter and overrepresented sequences was removed using TrimGalore software. All the clean reads of samples (gall and leaf) were assembled individually as well as both the samples combined together using the Trinity software (version: 2.3.2) (http://trinityrnaseq.github.io/) with default parameters and a K-mer size of 25 (27). Assembly of the high quality clean reads results in longer fragments known as contigs. Transcripts

with more than 300 bp were retained for further analysis. The paired-end reads were then mapped back to the contigs using RSEM (RNASeq by Expectation Maximization) for abundance estimation. We searched all transcript sequences against protein Non-redundant (Nr) databases using BLASTX. BLAST2GO analysis was done to obtain Gene Ontology (GO) annotation describing cellular component, molecular function and biological process (12).

Proteome Sequencing and Analysis

White cottony layer obtained from the inner hollow part of the galls was dissolved in 100 µL of 1X SDS sample buffer and boiled for 3min. Post run gel was stained with CBB-R-350. Gel band with the protein of interest was isolated. The band was further excised into smaller pieces of 1 mm these pieces were then destained using 40% acetonitrile (ACN) solution. Proteins were then reduced using 5mM Dithiothreitol (DTT) at 60°C for 45 min, reduced proteins were then alkylated using Iodoacetic acid (IAA) at room temperature for 10 min. Reduced and alkylated proteins were subjected to enzymatic digestion using sequencing grade modified trypsin (Promega, Madison, WI. Cat #: V5111) overnight at 37ºC. The digested peptides were extracted from the gel by vigorous vortexing using the buffer containing 40% ACN in 1% formic acid. The peptide extracts were vacuum dried and C_{18} clean up performed prior to storage at −80°C until further use.

The digested peptides obtained from gel band was cleaned using C18 StageTip based method using solvent-A (0.1% formic acid) for equilibrating column and cleaning peptides after loading on the column and were eluted using solvent (40% Acetonitrile and 0.1% formic acid) and then dried using SpeedVac, Thermo Scientific Orbitrap Fusion Tribrid mass spectrometer (Thermo Fischer Scientific, Bremen, Germany) connected to Easy-nLC-1200 nanoflow liquid chromatography system (Thermo Scientific). The peptides were reconstituted in 0.1% formic acid and loaded onto a 2 cm trap column (nanoViper, 3 µm C18 Aq) (Thermo Fisher Scientific). Peptides were separated using a 15 cm analytical column (nanoViper, 75 µm silica capillary, 2 µm C18 Aq) at a flow rate of 300nl/min. The solvent gradients were set as linear gradient of 5-35% solvent B (80% acetonitrile in 0.1% formic acid) for 90 min. The total run time for each fraction was 120 minutes. Global MS survey scan was carried out at a scan range of 400-1600 m/z mass range (120,000 mass resolution at 200 m/z) in a data-dependent mode using an Orbitrap mass analyzer. The maximum injection time was 5 ms. Only peptides with charge state 2-6 were considered for analysis, and the dynamic exclusion rate was set to 30 s. For MS/MS analysis, data was acquired at top speed mode with 3s cycles and subjected to higher collision energy dissociation with 34% normalized collision energy. MS/MS scans were carried out at a range of 100-1600 m/z using Orbitrap mass analyzer at a resolution of 30,000 at 200 m/z. Maximum injection time was 120ms. Internal calibration was carried out using lock mass option (m/z 445.1200025) from ambient air.

The raw files obtained after the data acquisition were searched using Proteome Discoverer software suite version 2.1 (Thermo Fisher Scientific) the data was searched against *Drosophila melanogaster*, *Arabidopsis thaliana* and *Cajanus cajan* databases. Protein sequences were downloaded from NCBI, and MS/MS data were searched against the protein database along with known mass spectrometer

contaminates using SEQUEST and Mascot algorithms. Search parameters included carbamidomethylation of cysteine as a static modification, oxidation of methionine, N-terminal acetylation at peptide N-terminus and Deamidation N and Q as a dynamic modification; and minimum peptide length was 7 amino acids selected with 2 missed cleavages allowed. Mass tolerance was set to 10 ppm at MS level and 0.05 Da for MS/MS., and false discovery rate was set to 1% at the PSM level.

Comparative analysis of transcriptome and proteome

RNA sequenced data from gall and leaf was compared with the identified proteins from *Pistacia* dry gall powder using standalone BLAST with the following parameters set as 90 % identity, 60 % query coverage and e-value of 0.001.

Declarations

Ethics approval and consent to participate: Not applicable

Consent for publication: Not applicable

Availability of data and materials: The whole genome and transcriptome sequencing data of gall and leaf samples used for this study has been deposited in NCBI Sequence Read Archive (SRA). Bio-project accession number is PRJNA480376.

Competing interests: The authors declare that they have no competing interests

Funding: This study was funded by Dabur India Limited. The funding body periodically reviewed the study design, experiments and provided their inputs for the best results.

Authors' contributions: PN and MG were involved in sample collection for RNA and protein studies. NB and AB involved in authentic sample collection. TSK, GD, MAN and KPA were involved in proteome analysis. JS and BS played critical role in designing the study. PV was involved in initiating and heading the project. PN, MG, SK, NB and PV were involved in editing the manuscript and providing inputs. All authors have read and approved the final manuscript

Acknowledgements

Dr. NB Brindavanam for his support at every stage in the project, right from inception; Bengaluru Genomics Centre for their support in sequencing and data analysis. We also thank Yenepoya (Deemed to be University) for access to proteomic facility; Dr. Yashwanth Subbannayya, Center for Systems Biology and Molecular Medicine, Yenepoya Research Centre Yenepoya (Deemed to be University), for his help in coordinating the proteome analysis. We also thank Mr. Santhosh Hegde, Research Fellow, The University of Trans-Disciplinary Health Sciences and Technology for his help in bioinformatics.

Abbreviations

BUSCO
Benchmarking Universal Single-Copy Orthologs
GO
Gene Ontology
MFS
Major Facilitator Superfamily
ABC
ATP binding cassette
MATE
Multidrug and Toxic Compound Exporters
SMR
Small Multidrug Resistance
RND
Resistance-Nodulation-Division Proteins
MFS
Major Facilitator Superfamily
GA
Gibberellic Acid
MISA
MIcroSAtellite tool
RSEM
RNASeq by Expectation Maximization
ACN
Acetonitrile
DTT
Dithiothreitol
IAA
Iodoacetic acid

References

1. Chen H, Liu J, Cui K, Lu Q, Wang C, Wu H, et al. Molecular mechanisms of tannin accumulation in Rhus galls and genes involved in plant-insect interactions. Scientific reports. 2018;8(1):9841.
2. Pant S, Samant S. Ethnobotanical observations in the Mornaula reserve forest of Komoun, West Himalaya, India. Ethnobotanical Leaflets. 2010;2010(2):8.
3. Orwa C, Mutua A, Kindt R, Jamnadass R, Simons A. Agroforestree database: a tree species reference and selection guide version 4.0. World Agroforestry Centre ICRAF, Nairobi, KE. 2009.
4. Barinder K, Saurabh S. A review on gall karkatshringi. Journal of Medicinal Plants Research. 2015;9(21):636-40.
5. Ved D, Goraya G. Demand and supply of medicinal plants in India. NMPB, New Delhi & FRLHT, Bangalore, India. 2007;18.
6. Diamond SE, Blair CP, Abrahamson WG. Testing the nutrition hypothesis for the adaptive nature of insect galls: does a non-adapted herbivore perform better in galls? Ecological Entomology. 2008;33(3):385-93.
7. Mapes CC, Davies P. Plant hormones and gall formation by Eurosta solidaginis on Solidago altissima. UNITED STATES DEPARTMENT OF AGRICULTURE FOREST SERVICE GENERAL TECHNICAL REPORT NC. 1998:161-72.
8. Wang H, Liu J, Cui K, Chen H, Yang Z, Wu H, et al. Gibberellic acid is selectively downregulated in response to aphid-induced gall formation. Acta physiologiae plantarum. 2016;38(9):214.
9. Rand K, Bar E, Ari MB, Davidovich-Rikanati R, Dudareva N, Inbar M, et al. Differences in Monoterpene Biosynthesis and Accumulation in Pistacia palaestina Leaves and Aphid-Induced Galls. J Chem Ecol. 2017;43(2):143-52.
10. Patel S, Rauf A, Khan H. The relevance of folkloric usage of plant galls as medicines: Finding the scientific rationale. Biomedicine & pharmacotherapy = Biomedecine & pharmacotherapie. 2018;97:240-7.
11. Ziya Motalebipour E, Kafkas S, Khodaeiaminjan M, Coban N, Gozel H. Genome survey of pistachio (Pistacia vera L.) by next generation sequencing: Development of novel SSR markers and genetic diversity in Pistacia species. BMC genomics. 2016;17(1):998.
12. Wu G, Zhang L, Yin Y, Wu J, Yu L, Zhou Y, et al. Sequencing, de novo assembly and comparative analysis of Raphanus sativus transcriptome. Frontiers in plant science. 2015;6:198.
13. Zhang X, Allan AC, Li C, Wang Y, Yao Q. De Novo Assembly and Characterization of the Transcriptome of the Chinese Medicinal Herb, Gentiana rigescens. International journal of molecular sciences. 2015;16(5):11550-73.
14. Rand K, Bar E, Ben-Ari M, Lewinsohn E, Inbar M. The mono-and sesquiterpene content of aphid-induced galls on Pistacia palaestina is not a simple reflection of their composition in intact leaves. Journal of chemical ecology. 2014;40(6):632-42.

15. Jin H, Martin C. Multifunctionality and diversity within the plant MYB-gene family. Plant molecular biology. 1999;41(5):577-85.
16. Heim MA, Jakoby M, Werber M, Martin C, Weisshaar B, Bailey PC. The basic helix–loop–helix transcription factor family in plants: a genome-wide study of protein structure and functional diversity. Molecular biology and evolution. 2003;20(5):735-47.
17. Peng H, Han S, Luo M, Gao J, Liu X, Zhao M. Roles of multidrug transporters of MFS in plant stress responses. International Journal of Bioscience, Biochemistry and Bioinformatics. 2011;1(2):109.
18. Cao A, Xing L, Wang X, Yang X, Wang W, Sun Y, et al. Serine/threonine kinase gene Stpk-V, a key member of powdery mildew resistance gene Pm21, confers powdery mildew resistance in wheat. Proceedings of the National Academy of Sciences. 2011;108(19):7727-32.
19. Liu P, Yang Z-X, Chen X-M, Chen H. Full-length cloning, sequence analysis and expression detection of the β-tubulin gene from the Chinese gall aphid (Schlechtendalia chinensis). Scientific reports. 2017;7(1):6459.
20. Novaes R, Rodrigues J, Lovato M. An efficient protocol for tissue sampling and DNA isolation from the stem bark of Leguminosae trees. Genet Mol Res. 2009;8(1):86-96.
21. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of computational biology. 2012;19(5):455-77.
22. Boetzer M, Henkel C, Jansen H, Butler D, Pirovano W. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics27: 578–579. 2011.
23. Boetzer M, Pirovano W. Toward almost closed genomes with GapFiller. Genome biology. 2012;13(6):R56.
24. Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics. 2007;23(9):1061-7.
25. Kuravadi NA, Yenagi V, Rangiah K, Mahesh H, Rajamani A, Shirke MD, et al. Comprehensive analyses of genomes, transcriptomes and metabolites of neem tree. PeerJ. 2015;3:e1066.
26. Mahesh HB, Subba P, Advani J, Shirke MD, Loganathan RM, Chandana SL, et al. Multi-omics driven assembly and annotation of the sandalwood (Santalum album) genome. Plant physiology. 2018;176(4):2772-88.
27. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature biotechnology. 2011;29(7):644.

Supplementary Material

Additional File 1: List of core eukaryotic genes mapped to Pistacia genome

Additional File 2: Functional annotation of Pistacia genome

Additional File 3: Top 50 highly expressed transcripts in gall and leaf

Additional File 4: Differential expressed transcripts in gall and leaf

Additional File 5: SSR distribution types

Additional File 6: List of plant hormonal signal transcripts expressed in gall

Additional File 7: Transcription factors transcripts in gall and leaf

Additional File 8: Functional annotation of genes in gall and leaf

Tables

Due to technical limitations, the tables are only available as a download in the supplemental files section.

Figures

Figure 1: Functional annotation of genes based on Gene Ontology (GO) categories from *Pistacia* genome

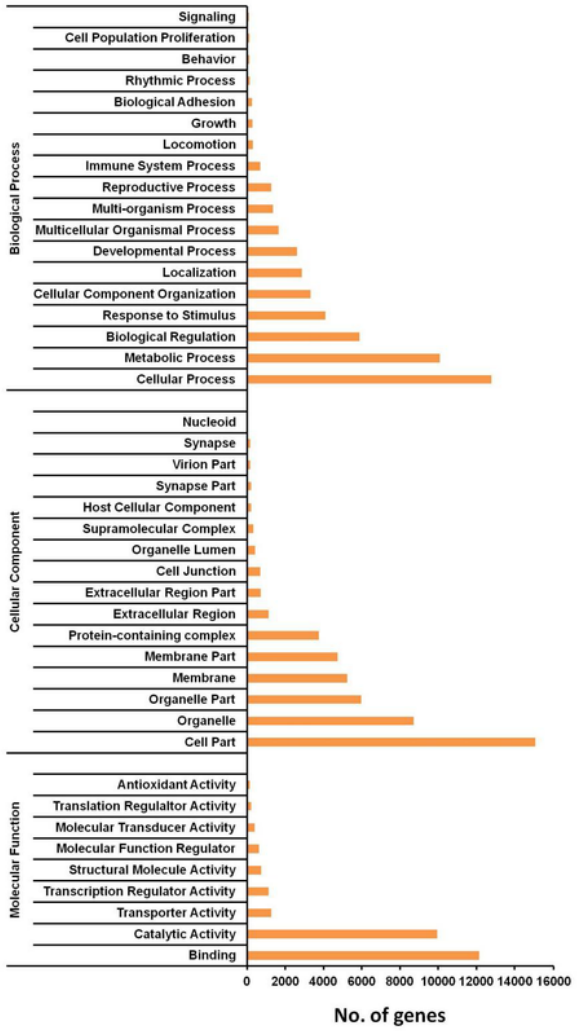

Figure 1

Functional annotation of genes based on Gene Ontology (GO) categories from Pistacia genome

Figure 2: Classification of potential enzyme genes in *Pistacia* genome

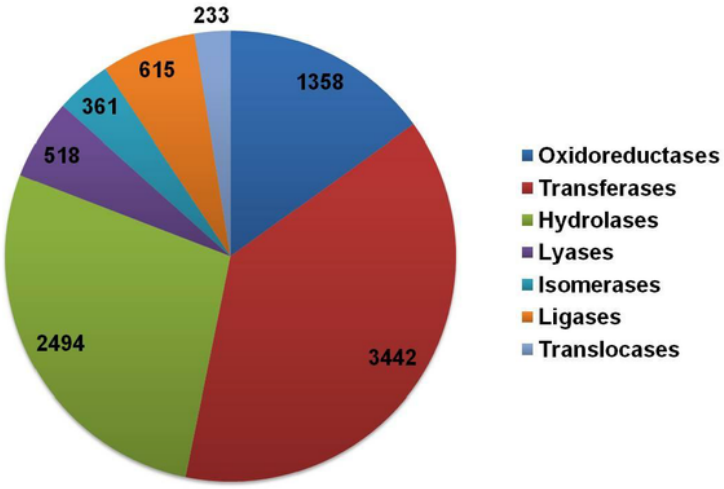

Figure 2

Classification of potential enzyme genes in Pistacia genome

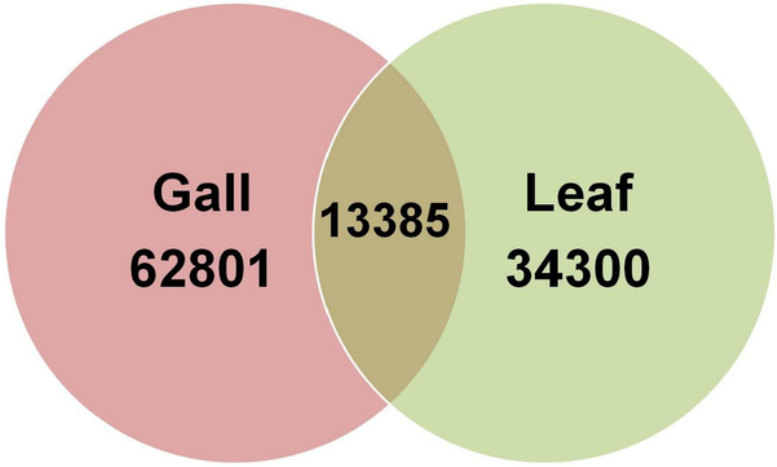

Figure 3: Specific transcripts in the gall and leaf of Pistacia. The figure represents the overlapping transcripts present in both gall and leaf

Figure 3

Specific transcripts in the gall and leaf of Pistacia. The figure represents the overlapping transcripts present in both gall and leaf

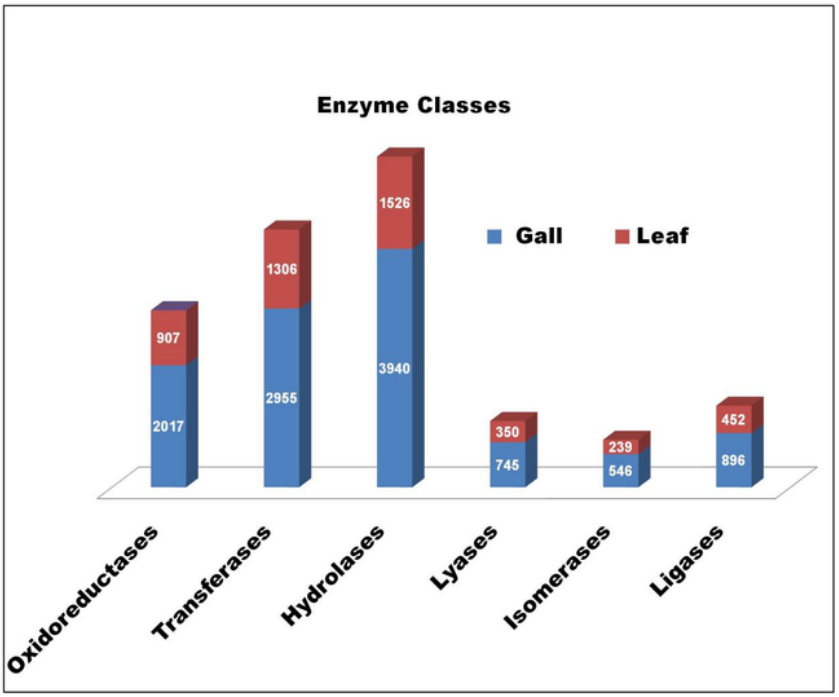

Figure 4: Functional characterization and classification of enzyme gall and leaf of Pistacia.

Figure 4

Functional characterization and classification of enzyme gall and leaf of Pistacia

Figure 5: Gene Ontology (GO) annotation of gall and leaf

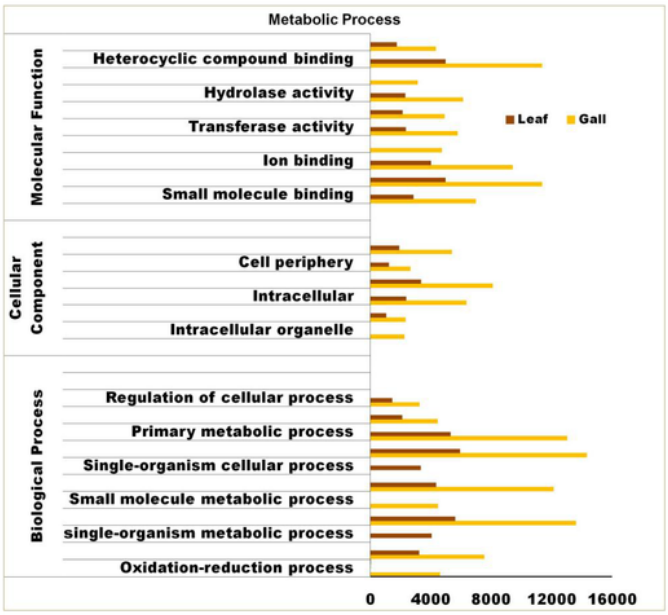

Figure 5

Gene Ontology (GO) annotation of gall and leaf

Figure 6: Transcription factors distribution in (A) Gall and (B) Leaf

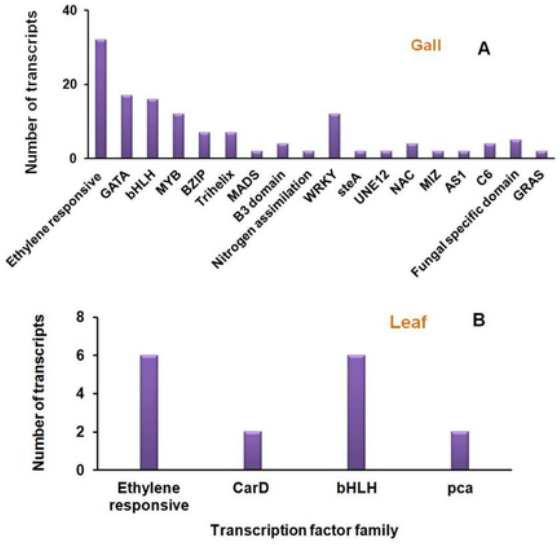

Figure 6

Transcription factors distribution in (A) gall and (B) Leaf

>XP_001951158.2 PREDICTED: tubulin beta chain-like [*Acyrthosiphon pisum*]
MLGGKYVPRAILVDLEPGTMDSVRSGPFGQIFRPDNFVFGQSGAGNNWAKGHYTEGAELVDSVLDVVRKE
AESCDCLQGFQLTHSLGGGTGSGMGTLLISKIREEYPDRIMTTYSIVPSPKVSDTVVEPYNSTLSVHQLV
ENTDKTFCIDNEALYDICFRTLKLTTPTYGDLNHLVSAAMCGVTTCFRFPGQLNSDLRKLAVNMVPFPRL
HFFITGFAPLTSRGSQQYRALTVPELVQQMFDAKNMMAACDPRHGRYLTAASIFRGRMSMKEVDEQMLNV
QTKNSSYFVEWIPNNTKTAVCDIPPRGLKMSATFIGNTTAIQEMFKRVSEQFTSMFRRKAFLHWYTGEGM
DEMEFTEAESNMNDLVSEYQQYQDATVDEDGEGDDEEDADA

>AAX52480.1 actin 5C, isoform D [*Drosophila melanogaster*]
MCDEEVAALVVDNGSGMCKAGFAGDDAPRAVFPSIVGRPRHQGVMVGMGQKDSYVGDEAQSKRGILTLKY
PIEHGIVTNWDDMEKIWHHTFYNELRVAPEEHPVLLTEAPLNPKANREKMTQIMFETFNTPAMYVAIQAV
LSLYASGRTTGIVLDSGDGVSHTVPIYEGYALPHAILRLDLAGRDLTDYLMKILTERGYSFTTTAEREIV
RDIKEKLCYVALDFEQEMATAASSSSLEKSYELPDGQVITIGNERFRCPEALFQPSFLGMEACGIHETTY
NSIMKCDVDIRKDLYANTVLSGGTTMYPGIADRMQKEITALAPSTMKIKIIAPPERKYSVWIGGSILASL
STFQQMWISKQEYDESGPSIVHRKCF

Figure 7: Peptides of Tubulin and Actin from Pistacia gall map to insect proteins. (color represent peptides identified)

Figure 7

Peptides of Tubulin and Actin from Pistacia gall map to insect proteins

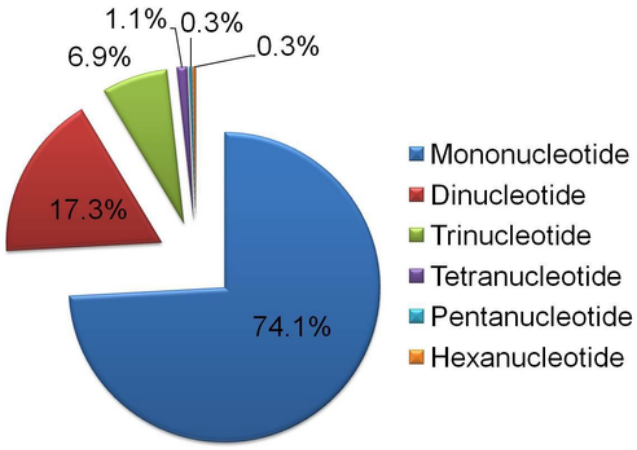

Figure 8: Distribution of 231624 SSRs in the Pistacia genome based on the repeat type

Figure 8

Distribution of 231624 SSRs in the Pistacia genome based on the repeat type

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AdditionalFile2.xls
- Tables.pdf
- AdditionalFile4.xls
- AdditionalFile7.xls
- AdditionalFile1.xls
- AdditionalFile3.xls
- AdditionalFile6.xls
- AdditionalFile8.xls
- AdditionalFile5.xls