De novo Transcriptome and tissue specific expression analysis of gene associated with biosynthesis of medicinally active metabolites in a high valued medicinal plant *Operculina turpethum* (L.)

Bhagyashree Biswal
Siksha O Anusandhan University School of Pharmaceutical Sciences
Biswajit Jena
Siksha O Anusandhan University School of Pharmaceutical Sciences
Alok Kumar Giri
Siksha O Anusandhan University School of Pharmaceutical Sciences
Laxmikanta Acharya (laxmikantaacharya@soa.ac.in)
Siksha O Anusandhan University School of Pharmaceutical Sciences https://orcid.org/0000-0002-8434-8500

Research Article

Keywords: Transcriptome, Operculina turpethum, Transcription Factor, SNP, SSR, Illumina Sequencing

Posted Date: April 1st, 2021

DOI: https://doi.org/10.21203/rs.3.rs-312726/v1

Abstract

***Operculina turpethum* (L.) Silva Manso** (Nisoth), an important medicinal plant whose root and stem tissues are found to be key ingredient in more than 135 herbal formulation in both Unani and Ayurvedic medicine system and is used for the treatment of various health disorders. It has a very high demand in the Indian Pharmaceutical industry with an annual average consumption of about 660 metric tonne (dry weight). The plant because of its wide usability is also exported to countries like Sri Lanka, Singapore, Netherlands, Japan, United States, etc. leading to a total export value of approximately 2.4 million USD. Terpenoids, steroid glycosides, resin glycosides, dammarane type triterpenoid saponins, flavonoids, coumarins are the major bioactive principles present in *O. turpethum*. However, overexploitation of root/root bark over the other part of the plant has caused depletion of its germplasm resources from the wild while bringing it to Near Threatened category which subsequently lead to unavailability of the requisite plant material followed by adulteration. Additionally, the limited genomics and transcriptomic resource is a major hindrance in the genetic and molecular research including elucidation of biosynthetic pathway, identification of enhanced traits and proper authentication and genetic diversity study of the plant. Hence, de novo transcriptome sequencing of root and stem tissues of *O. turpethum* was performed using Illumina HiSeq platform which generated a total of 64259 unigenes and 20870 CDS (coding sequence) with a mean length of 449bp and 571bp respectively. Further, 20218 and 16458 unigenes showed significant similarity with the identified proteins of NR(non-redundant) and Uniprot database respectively. The homology search carried out against publicly available database found the best match with Ipomoea nil sequences (82.6%). The KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway analysis identified 6538 unigenes functionally assigned to 378 modules with phenylpropanoid biosynthesis pathway as the most enriched among the secondary metabolite biosynthesis pathway followed by terpenoid biosynthesis. Moreover, transcription factors, SSRs (Simple sequence repeats) and SNPs (single nucleotide polymorphism) were also identified in this study. Apparently, the present investigation reported the first ever transcriptome analysis of the root and stem tissues of the non-model plant *O. turpethum* of family Convolvulaceae, which will provide a baseline for further molecular research.

Introduction

The pantropical genus *Operculina* is a major genus of the family Convolvulaceae comprising of 15 species. *Operculina turpethum* (L.) Silva Manso, popularly known as Indian jalap/ Turpeth/ Nisoth/ Trivrit, is one of the most industrially and therapeutically important medicinal herb of the morning glory family. This perennial vine is native to temperate and tropical region of Asia (India, Nepal, Bangladesh, Pakistan, Sri Lanka, China, Taiwan and Myanmar) but also found to grow in some part of Australia, Africa (Somalia, Kenya, Tanzania, Zimbabwe, Mozambique, Comoros, Madagascar and Mauritius), Pacific islands and southern America (West Indies). The plant is commonly found in the moist deciduous and tropical dry region of peninsular and central India (Western Ghats, Kerala, Karnataka and Tamil Nadu). *Operculina turpethum* root and stem are key ingredients in more than 135 herbal formulations in both Unani and Ayurvedic medicine system which are used to treat diverse ailments including obesity,

constipation, gastric ulcer, diarrhoea, asthma, uterine problem, cough splenomegaly, jaundice, anaemia, hyperlipidaemia, tumours, joint and muscle pain, paralysis and rheumatoid arthritis and tuberculosis. In addition, extensive pharmacological studies of different extracts of *O. turpethum* with animal models have demonstrated the antibacterial (Kiran et al. 2017; Kiran et al. 2018), analgesic (Ezeja et al. 2015), antioxidant (Sharma and Singh 2012), anti-inflammatory, anti-cancer (Arora et al. 2017), anti-diabetic (Pulipaka et al. 2012), hepato-protective(Prabhakaran and Ranganayakulu 2014), anti-ulcer (Ignatius et al. 2013), anti-arthritic (Tamizhmozhi and Nagavalli 2016), immune-modulatory (Tamizhmozhi and Nagavalli 2017) and anti-nephrotoxic, antispasmodic, bronchodilator (Shareef et al. 2014), laxative (Onoja et al. 2015) and larvicidal potential (Bhattacharya and Chandra, 2015) ascribed to their bioactive constituents including flavonoids, coumarin, scopoletin, Coumaric acid derivatives (N-p-coumaryl tyramine), triterpenoid (lanosta-5-ene, cycloartenol and 24-methylene-δ-5-lanosterol) dammarane type triterpenoid saponin (operculinosides A, B, C, D) (Ding et al. 2011), resin glycoside (turpethosides A, B) (Ding et al. 2012), glycosidic acid (turpethic acids A-C), acrylamide, phytosterol (daucosterol and β-sitosterol), betulin, lupeol, α- and β-turpethein, and steroid glycoside etc. Apart from the pharmacological application, the seeds of the plant are reported to be potential source of commercial gum (Singh et al. 2003). However excessive exploitation of root/root bark over the other part of the plant has caused depletion of its germplasm resources from the wild while bringing it to Near Threatened category (Ved et al. 2016) which subsequently lead to unavailability of the requisite plant material followed by adulteration.

Though *O. turpethum* has been extensively explored in phytochemical and pharmacological field, no reports related to transcriptomic and genetic studies of the plant is available in public database which are requisite for elucidating the metabolic pathways of the active ingredients and further improvement of its germplasm along with study of genetic diversity and proper identification of the immense medicinal plant.

In the current era, the advent of next generation sequencing techniques (NGS) has made the transcriptomic research of non-model organism very rapid and feasible. The high throughput RNA-sequencing (RNA-seq) approach has been widely employed to measure expression of gene across the transcriptome along with the detection of functional gene, alternative splicing, single nucleotide variant, post-transcriptional alternation, gene fusion and gene involved in the secondary metabolite biosynthetic pathway and development of molecular marker along with “Single nucleotide polymorphism” (SNPs) and “Simple sequence repeats (SSRs)” which can considerably play a pivotal role in the phylogenetic and genetic diversity study of medicinal plants.(Wang et al. 2009; Ekblom and Galindo 2011)

In this study, the Illumina 2 x150 paired end platform was adopted for the comprehensive transcriptome profile analysis of root and stem tissues of *Operculina turpethum.* The resulted sequence data were assembled and then annotated in multiple databases and the genes related to secondary metabolite biosynthesis were identified. All together the transcriptome data will serve as a foundation to explore the plant at genomic and transcriptomic level.

Material And Methods

Plant materials and RNA extraction:

The root and stem of *Operculina turpethum* (Fig. 1), collected from the local medicinal plant garden (Bhubaneswar, Odisha, India), were quick frozen in liquid nitrogen and subsequently preserved at -80^{0}C till further analysis. The total RNA of two tissues (root and stem) of *O. turpethum* was isolated using RNeasy Plant Mini Kit. 1% Formaldehyde Denaturing Agarose gel and Qubit® 2.0 Fluorometer was used for checking the quantity and quality of the isolated RNA.

Illumina 2 X 150 paired end library preparation and sequencing:

The total RNA isolated was subjected to oligo dT magnetic beads for the enrichment of mRNA. The purified mRNA was then subjected to fragmentation at appropriate temperature. The first strand cDNA of the fragmented mRNA was generated using RT-PCR (Reverse-Transcription PCR) which was then followed by synthesis of the second strand cDNA, A-base addition and adaptor-index ligation and lastly amplification. The libraries of amplified cDNA were screened through HS (High Sensitivity) DNA chip on Bioanalyzer 2100 (Agilent Technologies) according to the instructions provided by manufacturer. After the quality assessment, the library was subjected to 2x150bp PE chemistry on Illumina platform for cluster generation and sequencing. Sequencing of the template from both forward and reverse direction is facilitated by the paired end sequencing.

De novo assembly and CDS prediction:

The raw sequence data acquired from sequencing was subjected to quality control screening which include elimination of low-quality reads and adaptor or primer sequences containing reads. The clean reads generated after data processing were assembled using Trinity software (Version 2.1.1) with a fixed k-mer size of 25. The non-redundant clustered transcripts(unigenes) were predicted from the assembled transcripts using CD-hit software(version-4.6.1). finally all the assembled unigenes of the stem and root tissues were further processed for the prediction of CDS regions using TransDecoder tool (http://transdecoder.github.io) at default parameters with a minimum length of 100 amino acid of the encoded protein plus homology search with Pfam and UniProt databases .

Homology search and functional annotation of Unigenes:

The *O. turpethum* unigenes were searched against NCBI's "Non-redundant (NR)" database using BLASTX with an E-value cut off of 1E^{-5} to identify the sequence conservation. Further, for the functional characterization, the unigenes were submitted to BLASTX search against Uniport, Pfam, KOG/COG database. The GO mapping of the NR annotated sequences was obtained through Blast2GO command line V-1.4.1. For the assignment of orthologs and prediction of metabolic pathways in *O. turpethum,* the unigenes were compared against the KEGG database through KEGG automatic annotation server (KAAS) using BLASTX with threshold bit-score value of 60 (default).

Identification of Transcription factor:

The transcription factor families were identified by homology searches of the assembled unigenes against the plant transcription factor database (Plant TFDB v5.0) using BLASTX with an *E-value* cut off of $1E^{-5}$.

SSR and variant identification:

All the assembled unigenes were subjected to SSR detection by using a Perl script programme MISA (Microsatellite finder tool) (http://pgrc.ipk-gatersleben.de/misa/misa.html). For search parameter, the minimum number of repetitions for di-, tri-, tetra-, penta-, hexa-nucleotide was set to 6,5,3,3,3 respectively. For variant (SNPs and In/Dels) identification, the reads of root sample were considered as reference. The high-quality clean reads of stem sample were mapped on to the de novo assembled root unigenes using bwa v0.7.12-r1039. Then the conversion of resulting SAM file into BAM file, sorting and removal of duplicate reads achieved by using **picard-tools v1.119. GATK v3.5** (Genome Analysis Toolkit) pipeline was used for SNPs and Indel calling. Further, the resulting raw variants were filtered using vcf tools (vcf_utils packge) from SAM tool with mapping read depth 5 and quality cut off 20.

Differential Gene Expression Analysis:

Master de novo assembly generated using combined reads of Root and Stem samples was used for DEG (Differentially Expressed Gene) analysis. The reads of Root and Stem samples were mapped separately to unigenes sequences obtained from master assembly using bwa v0.7.12-r1039. Finally, these mapped reads with Root (control) vs Stem (Treated) combination were given as input to DESseq Bioconductor package in R, which consequently provides normalized values regarded as “basemean” which was further utilized for logFC and pvalue evaluation. In-house R-script (Xcelris proprietary Script) was employed to delineate the distribution and graphical representation of differentially expressed genes found in Root-vs-Stem samples. The criteria for the identification of DEGs were described in the table below **(Table 1)**.

Further the Top 50 significantly expressed genes (i.e. highly up and highly downregulated genes) were depicted in form of heatmap through MeV (Multiple Experiment Viewer) using hierarchical clustering approach.

The complete workflow for Illumina Sequencing de novo Assembly, Annotation and other Bioinformatics Analysis carried out in the Root- Stem Transcriptome of Operculina turpethum is provided in the **Fig.S1**.

Result And Discussion

The advent of high throughput next generation sequencing platform has dramatically reformed the perception of the complex and dynamic character of transcriptome by providing a better comprehensive and quantitative aspects of gene expression, allele specific expression and alternative splicing. Besides, the de novo transcriptome approach is a predominantly used cost effective technique for the non-model

plant where the good-quality reference genome information is not available and thereby enabling us to identify all the expressed transcripts. Currently the Illumina NGS platform has become the main workhouse for the generation of massive sequence information with respect to genomics and transcriptomics data for various non -model plant and animal species. Here a comprehensive report was provided on transcriptome profiling and differentially expressed genes of *O. turpethum* root and stem tissues for the first time which will be useful in exploring the biosynthetic pathway of pharmacologically active compound followed by genetic engineering of the immense medicinal plant.

In the present study the transcriptome sequencing of only root and stem tissues of *O. turpethum* was carried out as both these tissues have been reported to contain a number of bioactive phytochemicals which account for the tremendous pharmacological activities of the potent medicinal plant.

RNA Sequencing and Assembly:

For RNA sequencing, about 1µg RNA was used as input and the libraries construction was carried out through TruSeq Stranded mRNA Library Preparation Kit (Illumina). After enrichment, the mRNA fragments were subjected to cDNA synthesis followed by amplification using required PCR cycles. The mean sizes of the libraries are 416bp and 415bp respectively for samples Root and Stem. The libraries sequenced using 2x150bp PE chemistry on Illumina platform generated ~5 GB data per sample. After filtering, a total of 28622974 (root) and 26898420 (stem) raw reads were obtained having a Q20 base value (base quality more than 20) of 96.83%. A de novo assembly was done as no reference genome data was available for *Operculina turpethum.* Master assembly was performed taking reads of Root and Stem samples together using Trinity (at default parameters, kmer 25). The raw reads upon assembling, a total of 76790 transcripts for both root and stem taken together. The total transcript size was 35332145 bp with an average transcript length of 460 bp. The maximum transcript length was 4104 bp and a N50 length of 583 bp **(Table 2**). CD-HIT-EST executable was used to eliminate the shorter redundant sequences which have more than 90% identity with 100% coverage for any other transcripts and the clustered non-redundant transcripts thus obtained were designated as unigenes. This was done as low-quality bases and the presence of adapters in reads may hamper the assembly process resulting in mis assembly or truncated contigs. The total number of Unigenes obtained for both root and stem was 64259 with a total of 28856611 bases in unigenes. The mean unigene length was 449bp with a maximum length of 4104bp and N50 length of 564bp. **(Table 2**), where the length of majority of unigenes ranges from 200 to 500bp (**Fig.2, Table S1**). The higher N50 value further approves of a better-quality assembly. The CDS prediction was done from these unigenes using Transdecoder at default parameters a minimum length of 100 amino acid of the encoded protein plus homology search with Pfam and UniProt databases. A total of 20870 CDS were obtained having a total of 11929458 bases. The mean CDS length tends to be 571bp with maximum length of 2745bp.The maximum number of CDS belonged to 300 to 400 bp length and minimum to the account of 200 to less than 300 bp. (**Fig.2, Table S2**)

Homology search and Functional annotation:

A total of 20870 unigenes (32.5%) out of 64259 assembled unigenes were annotated functionally by searching against NCBI non-redundant (Nr) protein sequence database, UniProt, Clusters of Orthologous group of protein (KOG/COG), and Pfam database using BLASTX with an E-value threshold of $1E^{-5}$. On doing the similarity search, it was found that 20218 unigenes could be annotated to the nr database whereas, 16458 unigenes had similarity with UniPort, 10450 with KOG and 9760 with Pfam database. The comparative count of unigenes annotation in different databases was depicted in the form of Venn diagram using (http://www.interactivenn.net), which showed that a total of 6975 unigenes were co-annotated in four databases **(Fig. 3**).

The top-hit species distribution analysis revealed that most of the *O. turpethum* unigenes (16717, 82.6%) had significant homology with *Ipomea nil* sequences followed by *Cuscuta australis* (374, 1.8%) (**Fig. 4a)**. This was the case because both the plant species, *O. turpethum and I. nil* belongs to the Convolvulaceae family and morphologically look quite similar to each other except the flower color and size, whereas *Cuscuta australis* although belonged to the same family showed significant morphological difference and belonged to plant parasite group. Furthermore, the E-value distribution of the top-hits showed that 89% of the mapped sequences had significantly high scores for homology (E-value<10^{-50}), whereas 11% of the annotated sequences exhibited homology with e-value ranging from E^{-5} to E^{-50} (**Fig..4b**). Likewise, around 93% of the annotated sequences were found to have similarity above 80% **(Fig. 4c)**. These outcomes indicate the high similarity of the annotated sequences with the known sequences available in the public database, implying good quality assembly. However, there exists a large number of sequences (43389,67.52%) without any BLAST hits which may be due to the presence of new/novel genes performing function related to specific plants or due to the presence of untranslated regions or the short sequence lacking the conserved protein domain.

KOG classification produced hits for 10450 unigenes which were further classified into 25 KOG functional categories **(Fig. 5**). The highly enriched KOG category for Root-Stem sample was “Signal transduction mechanisms (T)” with 1478 unigenes followed by “Posttranslational modification, protein turnover, chaperones (O)” (1306) and General function prediction only (R)” (1289).

In Pfam analysis, the most abundant domains identified were representing “Pkinase” with 310 unigenes followed by “Pkinase_Tyr” (248) and “P450” (131). The top 10 most abundant Pfam domains and its counts were shown in the table below (**Table 2).**

GO classification of NR annotated Unigenes:

Out of 20218 NR annotated unigenes, a total of 1209 unigenes were assigned at least one GO terms using BLAST GO. The GO classification system comprises of 3 main domains: Biological process (BP), cellular component (CC) and Molecular function (MF), which were further divided into 40 subcategories in level 2 GO term annotation. Among all genes with GO annotation, 917 unigenes belongs to biological process category, 686 unigenes belongs to cellular component whereas the molecular function was the highly represented category with 977 unigenes (**Table S3**). Among biological process, metabolic process

accounts for the largest proportion followed by cellular process, localization and biological regulation. Under the cellular component domain, cell followed by plasma membrane were the most enriched sub categories. In molecular function, binding was the highly represented category followed by catalytic activity and transporter activity (**Fig. 6a)**.

Each of the three GO domains were further categorized into level 3 and level 4 GO terms for the extensive function analysis basing on GO database. For example, Binding was the most enriched level 2 term for molecular function and in the level 3 term, the binding activity was assigned to some specific function such as 'heterocyclic compound binding' (GO:1901363, 386 unigenes), 'organic cyclic compound binding' (GO:0097159, 386 unigenes), 'ion binding' (GO:0043167, 330 unigenes)(**Fig. 6b**). Furthermore, in level 4 term, the organic cyclic compound binding(belonging to binding activity) was subcategorized into 'nucleoside phosphate binding' (GO:1901265, 179 unigenes), 'nucleotide binding' (GO:0000166,), ribonucleotide binding (GO:0032553, 154 unigenes), and 'nucleic acid binding' (GO:0003676, 141 unigenes)(**Fig. 6b**).

Pathway analysis Using KEGG:

Among the sequences examined against KEGG database, 6518 unigenes were functionally assigned to 378 KEGG modules belonging to five main pathway categories, of which Metabolism was the most abundant category with 2751 unigenes (42.20%) followed by genetic information processing (1466 unigenes, 22.49%), environmental information processing (963 unigenes, 14.7%), cellular process (944 unigenes, 14.4%) and organismal system (394 unigenes, 6.02%). Among metabolism, most of the unigenes were involved in carbohydrate metabolism (21.7%) and amino acid metabolism (14.6%) followed by lipid metabolism (12.4%), energy metabolism (10.5%) and biosynthesis of other secondary metabolites (10.5%) (**Table S4**). **Fig. 7a** represents the distribution of top 20 most enriched pathways according to KEGG database with Signal transduction" as the most abundant pathway comprising of 946 unigenes followed by "Translation" (602) and "Carbohydrate metabolism" (595), "Transport and catabolism" (527). In this study the unigenes involved in Terpenoid and polyketide metabolism and other secondary metabolite biosynthesis were detected which support the presence of diverse secondary metabolite in *O. turpethum*.(**Fig.7b**) Along with above data, this study also identified some of the biosynthesis pathways related to antimicrobial compound including Streptomycin biosynthesis (PATH:ko005210), Neomycin, kanamycin and gentamicin biosynthesis (PATH:ko00524), Novobiocin biosynthesis (PATH:ko00401), which were also reported to be present in *Phyllanthus amarus and Plumbago zeylanica* transcriptomes (Bose Mazumdar and Chattopadhyay 2015; Karpaga Raja Sundari et al. 2020). The functional characterization of the non-model plant *O. turpethum* revealed that, the *de novo* transcriptome analysis based on RNA-seq will promote further research on the biochemistry, molecular genetics and physiology of *O. turpethum* or related species.

Differential Gene Expression Analysis in Root vs Stem sample:

Overall, 17444 DEGs (Differentially expressed genes) were identified out of which 8722 genes were upregulated and 8722 genes were downregulated in root Vs stem system (**Table S5**). The statistical

analysis of the expression of tissue specific unigenes revealed that 451 and 2975 unigenes were exclusively expressed in the root and stem tissues of *O. turpethum* respectively. The Scatter Plot and Volcano Plot **(Fig. 8)** represent the upregulated and downregulated unigenes in root and stem tissues. Additionally, the hierarchical clustering approach was used to represent the top 50 highly upregulated and highly downregulated genes in the form of heatmap (**Fig. 9)** .

Identification of gene involved in phenylpropanoid biosynthesis in O. turpethum:

In the present study it was found that the phenylpropanoid biosynthesis pathway contained maximum number of unigenes (190) encoding 16 key enzymes (**Table S6)** associated with phenylpropanoid biosynthesis (**Fig.S2**) and hence taken into further consideration. Phenylpropanoids are the diverse class of natural products whose biosynthesis is known to begin from the deamination of an aromatic amino acid Phenyl alanine by Phenyl alanine ammonia lyase (PAL, EC:4.3.1.24, 6unigenes) to form Cinnamic acid which is then hydroxylated by cinnamate-4-hydroxylase (EC:1.14.14.91, 1unigene) to form *p*-coumaric acid. *The* addition of a CoA thioester to *p*-coumaric acid by the enzyme 4-coumarate-Coenzyme A ligase (EC: 6.2.1.12, 8 unigenes) enzyme gives rise to *p*-coumaroyl CoA which serves as a high energy intermediate in the biosynthesis of lignin (cell wall component), flavonoids (pigments), pest resistance and UV protected compound (Isoflavonoids, flavonoids, stilbenes, furanocoumarins and coumarins) (Vogt 2010).The digital gene expression analysis revealed that gene encoding enzyme such as PAL(6 unigenes), CYP73A(1 unigene), 4CL(5 unigenes), HCT(2 unigenes), C3'H(1 unigene), COMT(1 unigene), F6H, involved in biosynthesis of Scopoletin (coumarin), were found to be significantly upregulated in stem tissues with some of them showing stem specific expression. Phenylpropanoids are reported to support the plant growth and survival by protecting the plant from UV- radiation and photo-oxidative effect, strengthening of specialized cell wall thereby providing vascular integrity, structural support and pathogen resistance to plants and stimulation of symbiotic nitrogen fixation (Korkina 2007). Furthermore, the phenylpropanoids and their derivatives were also known to possess several biological activities including antioxidant, antimicrobial, anticancer, antidiabetic, neuroprotective activity. The compounds also exhibit significant application in cosmetics, food and cosmetics industry due to their antimicrobial, antioxidant and photoprotective activity (Neelam et al. 2020).

Identification of unigenes involved in flavonoid biosynthetic pathway:

The Flavonoids(flavonols, flavandiols, flavones, chalcones, anthocyanins) are synthesized via phenylpropanoid biosynthesis pathway and known to be accountable for the coloration of flowers, fruits and seeds, plant reproduction and fertility, auxin transport, nodulation and also involved in defense mechanism by protecting the plant against UV-radiation, pathogen infection, herbivore attack, metal toxicity etc (Falcone Ferreyra et al. 2012). Interestingly, the present study identified 17 unigenes encoding 9 key enzymes (**Table S7**) involved in flavonoid biosynthesis (PATH: ko00941)(**Fig.S3**). Chalcone synthase (EC :2.3.1.74,2 unigenes), the first enzyme specific for flavonoid biosynthesis pathway which convert 4-coumaryl CoA to Chalcone, was found to be upregulated in the stem tissues based on the digital gene expression analysis. The isomerization of Chalcone to Naringenin is catalysed by the enzyme

Chalcone isomerase (EC:5.5.1.6, 3 unigenes) which was also found to be highly expressed in stem tissues (18 folds) as compared with root tissues. Naringenin then enters into the late step of flavonoid biosynthesis from which all other flavonoids are derived. Again, the pathway annotation revealed that the unigenes encoding anthocyanidin 3-O-glucosyltransferase (EC:2.4.1.115) and kaempferol 3-O-beta-D-galactosyltransferase (EC:2.4.1.234), were found to be stem specific transcripts which are known to be involved in anthocyanin and flavonol glycoside biosynthesis respectively.

Previous studies have reported the presence of Flavonoids like quercetin, kaempferol and the flavonoid glycoside, Formononetin 7-O-*β*-D-glucopyranoside in the aerial parts of *O. turpethum,* which also reported to exhibit anti-arthritic, immunomodulatory, anti-oxidant and anti-inflammatory activity (Tamizhmozhi and Nagavalli 2016; Tamizhmozhi and Nagavalli 2017) . From the comparative gene expression analysis, it was found that the most of the unigenes encoding CHS, CHI, F3H, FLS, DFR, BZ1, CYP81E were highly expressed in stem tissues indicating that they might be the key gene in regulating the biosynthesis of flavonoids which require further functional characterization.

Identification of key gene involved in Terpenoid biosynthesis pathway:

Similarly, terpenoids comprise the largest group of structurally diverse natural compounds and which are known to biosynthesize via two routes: '2-C-methyl-D-erythritol 4-phosphate (MEP)' pathway and 'mevalonate acid (MVA)' pathway (Sandeep and Ghosh 2020). The isoprene unit (C_5) synthesized from MEP pathway are engaged in the formation of mono-(C10), Di- (C20), and some polyterpenoids (Zhao et al. 2013) whereas the isoprene unit from the MVA pathway are used in the synthesis of triterpene(C30) and Sesquiterpene(C15) (Schilmiller et al. 2009). In the dataset, 86 genes were found to encode 42 key enzymes (**Table S8**). The identified 43 unigenes of terpenoid backbone biosynthesis (PATH: ko00900, 43 unigenes)(**Fig. S4**) encoded 7 enzymes for each of the MEP and MVA pathway and the MEP pathway related gene showed high expression in stem tissues as compared with the root tissue. 4 unigenes were found to be involved in monoterpenoid biosynthesis (PATH: ko00902)(**Fig.S5**) encoding Neomenthol dehydrogenase (EC: 1.1.1.208, 1unigene) and 8-Hydroxygeraniol dehydrogenase (EC: 1.1.1.324, 3 unigenes) and 7 unigenes were predicted to be associated with diterpenoid biosynthesis (PATH: ko00904) (**Fig.S6**) including ent-kaurene synthase(EC: 4.2.3.19, 1 unigene), ent-kaurene oxidase(EC: 1.14.14.86, 1unigene), gibberellin-44 dioxygenase(EC: 1.14.11.12, 1 unigenes), gibberellin 2beta-dioxygenase(EC: 1.14.11.13, 3unigenes) and trimethyltridecatetraene /dimethylnonatriene synthase (EC: 1.14.14.58 /1.14.14.59). Furthermore ,7 unigenes were identified as Sesquiterpenoid and triterpenoid biosynthesis (PATH: ko00909) (**Fig.S7**) related gene encoding Squalene synthase (farnesyl-diphosphate farnesyltransferase) (EC:2.5.1.21, 2 unigenes), squalene monooxygenase (EC: 1.14.14.17, 1 unigene), NAD+-dependent farnesol dehydrogenase(EC: 1.1.1.354, 2 unigenes) and (3S,6E)-nerolidol synthase (EC: 4.2.3.48,1 unigene) and germacrene D synthase (EC: 4.2.3.75, 1 unigene). According to the digital gene expression analysis, the one unigene of each of isoprene synthase (CDS_3411_Unigene_16235), prenyl protein peptidase (CDS_11648_Unigene_3123), NAD+-dependent farnesol dehydrogenase (CDS_5038_Unigene_19017), Squalene synthase (CDS_13671_Unigene_4006) and (3S,6E)-nerolidol synthase (CDS_7899_Unigene_23587) were found to be exclusively present in stem tissues of *O.*

turpethum. The differential expression patterns of the genes may be responsible for the differential accumulation of previously reported terpenoids such as Carvacrol, Thymol, Cycloartenol, Lanosta-5-ene, 24-methylene-δ-5-lanosterol, lupeol, betulin, linalool in *O. turpethum.*

Identification of Transcription factors (TFs):

Transcription factors are sequence specific DNA binding trans-regulatory protein and reported to mediate the gene expression with reference to numerous developmental and environmental stimuli by recognizing specific cis-regulatory DNA sequences at the promoter region if their target gene. Moreover, the transcription factors are found to play a significant function in the regulation of several pathway related to secondary metabolism by controlling the metabolic flux and cellular differentiation to a large extent. Here a total of 1079 unigenes encoding putative transcription factors were identified and further grouped into 46 different TFs families in the *O. turpethum* transcriptome, out of which WRKY constitute the most abundant TFs family with 173 unigenes (15.7%) followed by BHLH(150, 13.6%), MYB(123,11.2%), ERF(88, 8.02%), bZIP (50,4.5%), GRAS(43, 3.9%) (**Fig. 10a**). WRKY TF is one of the major and largest group of plant specific TFs family which are previously reported to be involved in regulating several biological processes such as development of plant (Ramachandran et al. 1994), responses to pathogen entry (Cheong et al. 2002), nutritional deficiency, endosperm, seed, embryo and micropyle development and senescence (Bakshi and Oelmüller 2014), responses to different biotic and abiotic stress, phytohormone signalling pathway and also demonstrated to play major role in regulating the expression of gene engaged in biosynthesis of various secondary metabolites like flavanols, phenolic compounds including lignin and tannins (Guillaumie et al. 2009; Phukan et al. 2016). Besides WRKY, other TFs families such as bHLH, MYB, C2H are also involved in secondary metabolism pathway (Patra et al. 2013).

Among the TFs encoding unigenes (1097) identified in *O. turpethum* transcriptome, 597 unigenes classified in 24 TFs families exhibited differential expression with 291 upregulated unigene in root versus stem comparison. 55 unigenes of WRKY TFs family followed by ERF (46), bHLH (30) was found to be highly upregulated in the stem tissues whereas, the most frequently upregulated genes in root tissues belongs to TFs family MYB (45) followed by C3H (36) and bHLH (31) (**Fig. 10b**). The high expression level of WRKY transcription factor in stem tissue might be regulating the biosynthesis of terpenoids as reported in the *Gossypium arboreum* and *Taxus chinensis* where GaWRKY1 and TaWRKY1 were found to regulate the synthesis of gossypol and paclitaxel biosynthesis. (Xu et al. 2004; Li et al. 2013).

Identification of SSR:

Simple sequence repeat (SSR) or Microsatellite are stretches of DNA consisting of short tandem repeat motifs of 1-6 nucleotides. SSRs were first applied to plant science by (Akkaya et al. 1992) and over the past 30 years, it has been extensively used in plant genotyping as they are highly reproducible and transferable among related species, informative, multiallelic in nature and exhibit co-dominant inheritance. Basically, the SSR markers are favourable in genetic diversity study, estimation of gene flow and rate of crossing over and construction of linkage map, QTL mapping, study of genetic relatedness

and population structure, cultivar identification and DNA fingerprinting (Lörz and Wenzel 2005; Nadeem et al. 2018). The emergence of high throughput next generation sequencing approach has come up with a new framework for identifying microsatellites. Currently there are no studies addressing the genetic diversity and classification of germplasm resource of *O. turpethum* based on SSR marker as they have not been discovered so far. In this study, the genic SSRs are identified for the first time using the largescale transcriptome data which could be helpful for the genetic and breeding studies. For the identification of SSRs, the *O. turpethum* transcripts were searched with MISA software. Of the 64,259 transcripts of *O. turpethum*, a total of 8585 potential SSR loci were discovered. The number of SSRs containing sequences in *O. turpethum* was 6970.The frequency of SSRs is 13.4% and the average distribution distance is 3361bp (**Table 4**).The number of transcripts containing more than one SSR was 1248 and the number of SSRs present in compound form was 346,where the maximum number of bases interrupting 2 SSRs in a compound microsatellite is 10. Analysis of the data showed that the most abundant motif type in *O. turpethum* was the trinucleotide repeats(4174:48.30%), also recorded in studies of other plant species such as *I. nil ,I. batatas, T. cordifolia, P. ovata.*(Wang et al. 2010; Wei et al. 2015; Singh et al. 2016; Kotwal et al. 2016) The next class of repeat motifs observed frequently was tetranucleotide repeats (1840), followed by dinucleotide repeat (1320), pentanucleotide (653) and hexanucleotide (614) repeat motifs (**Fig.11a**). Among all identified SSRs, the AAG/CTT trinucleotide motif was found to be most abundant accounting for a largest fraction (11.4% :977) off SSRs followed by AG/CT dinucleotide repeat motif with a frequency of 7.4% (**Fig.11b**). This frequency of distribution appears to contradict the previous findings in most plant genomes such as *I. batatas*, *I. nil* where the AG/CT dinucleotide motif repeat was found to be most abundant which may be due the different genetic makeup of different species and the different standards used for the search of SSRs. Furthermore, the current study reported the occurrence of CCG/CGG trinucleotide motif, a most predominant motif in monocot plant species, which accounts for 6.4% of total SSRs detected. But the recent findings do not support the notion of rare appearance of CCG/CGG motif in most dicot plant species.

This study reports the discovery of genome wide SSRs for the first time in *O. turpethum,* which will enrich the molecular marker resource of *O. turpethum* and will also be helpful in further research related to genetic diversity studies, genetic linkage mapping, and marker assisted selection to trigger the traditional plant breeding.

Variant analysis:

SNPs (Single Nucleotide polymorphism) were defined as nucleotide variation at a single base in the DNA sequence whereas indels were referred to insertion and/or deletion of one or more nucleotides in the sequence. Presently, the SNPs and indel markers have become significantly popular in the field of plant molecular genetics and breeding with applicability in the identification of cultivar and plant variety, construction of genetic map QTL analysis, marker assisted selection, etc. as they are highly abundant in the genome and compatible for high throughput detection platforms (Mammadov et al. 2012). The next generation sequencing technology has now been extensively applied for the development of vast genotyping arrays leading to fast and efficient detection of SNP and indel markers(Kumar et al. 2012).

In this study, a total of 5863 SNPs were detected in 951 unigenes. Of these 3505 (59.78%) were transition and 2358(40.21%) were transversion mutation with a Ts/Tv ratio of 1.48. Among the transition mutation, T/C occurred at highest frequency of 17.5% followed by C/T with a frequency of 6.3%. Likewise, T/A followed by A/T variation were most abundant with a frequency of 7.5% and 6.3% respectively, among the transversion mutation.(**Fig. 12**)

Similarly, in comparison with the reference sequences, a total of 401 Indels were identified from the *Operculina turpethum* stem sample, out of which 261(65.08%) were deletions and 140(34.91%) were insertions. On an average one deletion was found per 45.07kb of de novo assembled sequence where as one insertion was found per 85.21kb. The deletions were located in 248 unigenes. The average size of a deletion was 7.3bp with a variation of 1-27bp. The longest deletion of 27bp was detected in the gene encoding B3-domain containing REM5 protein involved in transcription and transcriptional regulation. The average size of an insertion was 9.3bp varying in the range of 1-92bp. The longest insertion of 92 bp was found in the gene encoding protein serine-threonine kinase.

Conclusion

Presumably, the current investigation for the first time reported the transcriptome analysis of root and stem tissues of *O. turpethum* which was carried out without any reference genome using Illumina HiSeq platform. The study generated a total numbers of 76790 transcripts and 64259 unigenes with an average sequence length of 449bp, pooled from the root and stem tissues together. About 32.5% of the identified unigenes were annotated and functionally classified in 6 databases. The KEGG pathway analysis provides an overview of important pathway along with identification of a large number of candidate genes encoding key enzymes involved in secondary metabolites including phenylpropanoid, flavonoids and terpenoid biosynthesis. Various conserved transcription factor families were also predicted for the advancement of future genomic research of the plant. Furthermore, a total number of 8585 potential genomic SSRs and 5836 SNPs were identified in this study representing the first report of its type which would be a cost-efficient resource for genetic diversity assessment and marker assisted breeding by developing functional molecular marker. Altogether this finding will further prosper the knowledge on the biosynthesis, regulation, tissue-specific accumulation of important bioactive compound and their enhancement through genetic engineering along with the selection of superior allele governing the desired trait for breeding of the potential medicinal plant *O. turpethum* in future.

Abbreviations

CDS, Coding Sequence; NR, non-redundant; KEGG, Kyoto Encyclopedia of Genes and Genomes; SNP, Single nucleotide polymorphism; SSR, Short sequence repeats; NGS, next generation sequencing; GO, gene ontology; KOG, EuKaryotic Orthologous Groups; DEG, Differentially Expressed Gene; kmer, ;bp, basepair; UV, ultraviolet; QTL, quantitative trait loci.

Declarations

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflicts of interest: The authors declare that they have no conflict of interest.

Ethics approval: Not applicable

Consent to participate: Not applicable

Consent for publication: All authors provide their consent for publication.

Availability of data and material: The generated raw sequence data were deposited to NCBI Sequence Read Archive(SRA) database under the Bio project accession No PRJNA655823 and SRA accession No SRX8904980 (for root tissue) and SRX8904981(for stem tissue). The authors confirm that the data supporting the findings of this study are available within the article and its supplementary materials. Any other relevant data are available upon request from the corresponding author.

Code availability: Not applicable

Authors' contributions: Laxmikanta Acharya designed the experiment. Alok Kumar Giri contributed to collection of plant materials. Bhagyashree Biswal performed the experiment. Biswajit Jena contributed to data analysis and Bhagyashree Biswal prepared the manuscript. Laxmikanta Acharya supervised the entire study and revised the manuscript. All author reviewed and approved the article for publication.

Acknowledgments: All authors thank Prof. (Dr.) Manojranjan Nayak, Founder & President, Siksha ‘O’ Ansusandhan (Deemed to be University), Bhubaneswar, Odisha, India for providing the required infrastructure for the study. The authors are grateful to the Head, Centre for Biotechnology and Dean, School of Pharmaceutical Sciences, Siksha O Anusandhan(Deemed to be University) for the support.

References

Akkaya MS, Bhagwat AA, Cregan PB (1992) Length polymorphisms of simple sequence repeat DNA in soybean. Genetics 132:1131–1139

Arora R, Bharti V, Gaur P, Aggarwal S, Mittal M, Das SN (2017) Operculina turpethum extract inhibits growth and proliferation by inhibiting NF-κB, COX-2 and cyclin D1 and induces apoptosis by up regulating P53 in oral cancer cells. Arch Oral Biol 80:1–9 . https://doi.org/10.1016/j.archoralbio.2017.03.015

Bakshi M, Oelmüller R (2014) WRKY transcription factors: Jack of many trades in plants. Plant Signal Behav 9:e27700 . https://doi.org/10.4161/psb.27700

BHATTACHARYA K, CHANDRA G (2015) Biocontrol efficacy of Operculina turpethum (L.) (Convolvulaceae) leaf extractives against larval form of malarial mosquito Anopheles stephensi (Liston

1901). Int J Pharm Bio Sci 6:460–468

Bose Mazumdar A, Chattopadhyay S (2015) Sequencing, De novo Assembly, Functional Annotation and Analysis of Phyllanthus amarus Leaf Transcriptome Using the Illumina Platform. Front Plant Sci 6:1199 . https://doi.org/10.3389/fpls.2015.01199

Cheong YH, Chang H-S, Gupta R, Wang X, Zhu T, Luan S (2002) Transcriptional Profiling Reveals Novel Interactions between Wounding, Pathogen, Abiotic Stress, and Hormonal Responses in Arabidopsis. Plant Physiol 129:661 LP – 677 . https://doi.org/10.1104/pp.002857

Ding W, Jiang Z-H, Wu P, Xu L, Wei X (2012) Resin glycosides from the aerial parts of Operculina turpethum. Phytochemistry 81:165–174 . https://doi.org/10.1016/j.phytochem.2012.05.010

Ding W, Zeng F, Xu L, Chen Y, Wang Y, Wei X (2011) Bioactive dammarane-type saponins from Operculina turpethum. J Nat Prod 74:1868–1874 . https://doi.org/10.1021/np200274m

Ekblom R, Galindo J (2011) Applications of next generation sequencing in molecular ecology of non-model organisms. Heredity (Edinb) 107:1–15 . https://doi.org/10.1038/hdy.2010.152

Ezeja M, Onoja S, Omeh Y, Chibiko C (2015) Analgesic and antioxidant activities of the methanolic extract of Operculina turpethum leaves in mice. Int J Basic Clin Pharmacol 4:453–457 . https://doi.org/10.18203/2319-2003.ijbcp20150015

Falcone Ferreyra ML, Rius S, Casati P (2012) Flavonoids: biosynthesis, biological functions, and biotechnological applications. Front Plant Sci 3:222 . https://doi.org/10.3389/fpls.2012.00222

Guillaumie S, Mzid R, Méchin V, Léon C, Hichri I, Destrac-Irvine A, Trossat-Magnin C, Delrot S, Lauvergeat V (2009) The grapevine transcription factor WRKY2 influences the lignin pathway and xylem development in tobacco. Plant Mol Biol 72:215 . https://doi.org/10.1007/s11103-009-9563-1

Ignatius V, Narayanan M, Subramanian V, Periyasamy BM (2013) Antiulcer Activity of Indigenous Plant *Operculina turpethum* Linn. Evidence-Based Complement Altern Med 2013:272134 . https://doi.org/10.1155/2013/272134

Karpaga Raja Sundari B, Budhwar R, Dwarakanath BS, Thyagarajan SP (2020) De novo transcriptome analysis unravels tissue-specific expression of candidate genes involved in major secondary metabolite biosynthetic pathways of Plumbago zeylanica: implication for pharmacological potential. 3 Biotech 10:271 . https://doi.org/10.1007/s13205-020-02263-9

Kiran B, Chauhan B, N P (2018) Antibacterial Activity of Different Solvent Extract of Operculina Turpethum (L .) Silva (Root) Against Important Species of Bacteria. World J Pharm Res 7:410–415 . https://doi.org/10.20959/wjpr201813-12771

Kiran B, Padmini N, Chauhan JB (2017) In-vitro evaluation antibacterial potentiality of aqueous extract of Operculina turpethum silva (root). Eur J Pharm Med Res 4:261–263

Korkina LG (2007) Phenylpropanoids as naturally occurring antioxidants: from plant defense to human health. Cell Mol Biol (Noisy-le-grand) 53:15–25

Kotwal S, Kaul S, Sharma P, Gupta M, Shankar R, Jain M, Dhar MK (2016) De Novo Transcriptome Analysis of Medicinally Important Plantago ovata Using RNA-Seq. PLoS One 11:e0150273

Kumar S, Banks TW, Cloutier S (2012) SNP Discovery through Next-Generation Sequencing and Its Applications. Int J Plant Genomics 2012:831460 . https://doi.org/10.1155/2012/831460

Li S, Zhang P, Zhang M, Fu C, Yu L (2013) Functional analysis of a WRKY transcription factor involved in transcriptional activation of the DBAT gene in Taxus chinensis. Plant Biol (Stuttg) 15:19–26 . https://doi.org/10.1111/j.1438-8677.2012.00611.x

Lörz H, Wenzel G (2005) Molecular Marker Systems in Plant Breeding and Crop Improvement. Springer Nature

Mammadov J, Aggarwal R, Buyyarapu R, Kumpatla S (2012) SNP Markers and Their Impact on Plant Breeding. Int J Plant Genomics 2012:728398 . https://doi.org/10.1155/2012/728398

Nadeem MA, Nawaz MA, Shahid MQ, Doğan Y, Comertpay G, Yıldız M, Hatipoğlu R, Ahmad F, Alsaleh A, Labhane N, Özkan H, Chung G, Baloch FS (2018) DNA molecular markers in plant breeding: current status and recent advancements in genomic selection and genome editing. Biotechnol Biotechnol Equip 32:261–285 . https://doi.org/10.1080/13102818.2017.1400401

Neelam, Khatkar A, Sharma KK (2020) Phenylpropanoids and its derivatives: biological activities and its role in food, pharmaceutical and cosmetic industries. Crit Rev Food Sci Nutr 60:2655–2675 . https://doi.org/10.1080/10408398.2019.1653822

Onoja SO, Madubuike GK, Ezeja MI, Chukwu C (2015) Investigation of the Laxative Activity of Operculina turpethum Extract in Mice. Int J Pharm Clin Res 7:275–279

Patra B, Schluttenhofer C, Wu Y, Pattanaik S, Yuan L (2013) Transcriptional regulation of secondary metabolite biosynthesis in plants. Biochim Biophys Acta 1829:1236–1247 . https://doi.org/10.1016/j.bbagrm.2013.09.006

Phukan UJ, Jeena GS, Shukla RK (2016) WRKY Transcription Factors: Molecular Regulation and Stress Responses in Plants. Front Plant Sci 7:760 . https://doi.org/10.3389/fpls.2016.00760

Prabhakaran V, Ranganayakulu D (2014) Hepatoprotective activity of hydroalcoholic extract of Operculina turpethum Linn. against d-galactosamine induced hepatotoxicity in rats. Int J Phytopharm 5:358–365

Pulipaka S, Challa S, Pingili R (2012) Comparative antidiabetic activity of methanolic extract of Operculina turpethum stem and root against healthy and streptozotocin induced diabetic rats. Int Curr Pharm J 1: . https://doi.org/10.3329/icpj.v1i9.11618

Ramachandran S, Hiratsuka K, Chua NH (1994) Transcription factors in plant growth and development. Curr Opin Genet Dev 4:642–646 . https://doi.org/10.1016/0959-437x(94)90129-q

Sandeep, Ghosh S (2020) Chapter 12 - Triterpenoids: Structural diversity, biosynthetic pathway, and bioactivity. In: Atta-ur-Rahman BT-S in NPC (ed) Bioactive Natural Products. Elsevier, pp 411–461

Schilmiller AL, Schauvinhold I, Larson M, Xu R, Charbonneau AL, Schmidt A, Wilkerson C, Last RL, Pichersky E (2009) Monoterpenes in the glandular trichomes of tomato are synthesized from a neryl diphosphate precursor rather than geranyl diphosphate. Proc Natl Acad Sci 106:10865 LP – 10870 . https://doi.org/10.1073/pnas.0904113106

Shareef H, Rizwani GH, Mandukhail SR, Watanabe N, Gilani AH (2014) Studies on antidiarrhoeal, antispasmodic and bronchodilator activities of Operculina turpethum Linn. BMC Complement Altern Med 14:479 . https://doi.org/10.1186/1472-6882-14-479

Sharma V, Singh M (2012) In vitro radical scavenging activity and phytochemical screening for evaluation of the antioxidant potential of Operculina turpethum root extract. J Pharm Res 5:783–787

Singh R, Kumar R, Mahato AK, Paliwal R, Singh AK, Kumar S, Marla SS, Kumar A, Singh NK (2016) De novo transcriptome sequencing facilitates genomic resource generation in Tinospora cordifolia. Funct Integr Genomics 16:581–591 . https://doi.org/10.1007/s10142-016-0508-x

Singh V, Srivastava V, Pandey M, Sethi R, Sanghi R (2003) Ipomoea turpethum seeds: a potential source of commercial gum. Carbohydr Polym 51:357–359 . https://doi.org/https://doi.org/10.1016/S0144-8617(02)00186-8

Tamizhmozhi M, Nagavalli D (2016) Curative effect of formonoetin-7-o-b-d-glucopyranoside from methanolic extract of Operculina turpethum in freund's complete adjuvant induced arthritis. Indo Am J Pharm Sci 3:1415–1423 . https://doi.org/doi.org/10.5281/zenodo.208204

Tamizhmozhi M, Nagavalli D (2017) Immunomodulatory activity of Formonoetin-7- O - β -D-glucopyranoside isolated from Methanolic Extract of Operculina turpethum. Int J ChemTech Res 10:356–364

Ved DK, Sureshchandra ST, Barve V, Srinivas V, Sangeetha S, Ravikumar K, R. K, Kulkarni V, Kumar AS, Venugopal SN, Somashekhar BS, Sumanth MV, Begum N, Rani S, K.V. S, Desale N (2016) Conservation concerned species. envis.frlht.org / frlhtenvis.nic.in

Vogt T (2010) Phenylpropanoid Biosynthesis. Mol Plant 3:2–20 . https://doi.org/https://doi.org/10.1093/mp/ssp106

Wang Z, Fang B, Chen J, Zhang X, Luo Z, Huang L, Chen X, Li Y (2010) De novo assembly and characterization of root transcriptome using Illumina paired-end sequencing and development of cSSR markers in sweet potato (Ipomoea batatas). BMC Genomics 11:726 . https://doi.org/10.1186/1471-2164-11-726

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10:57–63 . https://doi.org/10.1038/nrg2484

Wei C, Tao X, Li M, He B, Yan L, Tan X, Zhang Y (2015) De novo transcriptome assembly of Ipomoea nil using Illumina sequencing for gene discovery and SSR marker identification. Mol Genet Genomics 290:1873–1884 . https://doi.org/10.1007/s00438-015-1034-6

Xu Y-H, Wang J-W, Wang S, Wang J-Y, Chen X-Y (2004) Characterization of GaWRKY1, a cotton transcription factor that regulates the sesquiterpene synthase gene (+)-delta-cadinene synthase-A. Plant Physiol 135:507–515 . https://doi.org/10.1104/pp.104.038612

Zhao L, Chang W, Xiao Y, Liu H, Liu P (2013) Methylerythritol phosphate pathway of isoprenoid biosynthesis. Annu Rev Biochem 82:497–530 . https://doi.org/10.1146/annurev-biochem-052010-100934

Tables

Table 1. Criteria for the identification of DEGs

Condition	Status
log2FC > 0	*Up regulated*
log2FC < 0	*Down regulated*
log2FC > 0 and p-value <0.05	*Significantly up regulated*
log2FC < 0 and P-value <0.05	*Significantly down regulated*

Table 2: Statistics of Assembled Transcripts and Unigenes

Description	Number of Transcripts	Number of Unigenes
Length >= 200 && <= 500bp	54979	47043
Length >= 500 && <= 1000bp	15136	11804
Length >= 1000 && <= 5000bp	6675	5412
Total no	76790	64259
Total number of bases(bp)	35332145	28856611
Average length (bp)	460	449
Maximum length	4104	4104
N50(bp)	583	564

Table 3. Top Pfam domain in Root-Stem

Domain	Count
Pkinase	310
Pkinase_Tyr	248
p450	131
RRM_1	128
zf-RING_2	83
Ras	77
COX1	75
PP2C	71
WRKY	65

Table 4: SSR Statistics

Results of Microsatellite Search	Root-Stem
Total number of sequences examined	64,259
Total size of examined sequences (bp)	2885611
Total number of identified SSRs	8585
Number of SSR containing sequences	6970
Number of sequences containing more than 1 SSR	1248
Number of SSRs present in compound formation	346

Supporting Information

Figure S1. Workflow for Illumina Sequencing de novo Assembly, Annotation and other Bioinformatics Analysis carried out in the Root- Stem Transcriptome of Operculina turpethum.

Figure S2. The color-coded map corresponds to map ko00940 in the KEGG database. The Red-color box represents the unigenes encoding key enzyme involved in Phenylpropanoid Biosynthesis pathway in *O. turpethum.*

Figure S3. The color-coded map corresponds to map ko00941 in the KEGG database. The Red-color box represents the unigenes encoding key enzyme involved in Flavonoid Biosynthesis pathway in *O. turpethum.*

Figure S4. The color-coded map corresponds to map ko00900 in the KEGG database. The Red-color box represents the unigenes encoding key enzyme involved in Terpenoid backbone Biosynthesis pathway in *O. turpethum.*

Figure S5. The color-coded map corresponds to map ko00902 in the KEGG database. The Red-color box represents the unigenes encoding key enzyme involved in Monoterpenoid Biosynthesis pathway in *O. turpethum.*

Figure S6. The color-coded map corresponds to map ko00904 in the KEGG database. The Red-color box represents the unigenes encoding key enzyme involved in Diterpenoid Biosynthesis pathway in *O. turpethum.*

Figure S7. The color-coded map corresponds to map ko00909 in the KEGG database. The Red-color box represents the unigenes encoding key enzyme involved in Sesquiterpenoid and triterpenoid Biosynthesis pathway in *O. turpethum.*

Table S1:Unigene Length Distribution

Table S2: CDS Length Distribution

Table S3: Level 2 Gene Ontology Distribution Of NR annotated Unigenes, divided into 3 main Domain Biological process (BP), cellular component (CC) and Molecular function (MF)

Table S4: KEGG pathway annotation of the unigenes. These unigenes were divided into five branches (Metabolism; Genetic Information Processing; Environmental Information Processing; Cellular Processes; Organismal System.)

Table S5: Differentially Expressed Unigenes in Root-Stem

Table S6: Unigenes Encoding the Key Enzymes Associated With Phenylpropanoid Biosynthesis

Table S7: Unigenes Encoding the Key Enzymes Associated With Flavonoid Biosynthesis

Table S8: Unigenes Encoding the Key Enzymes Associated With Terpenoid Biosynthesis

Figures

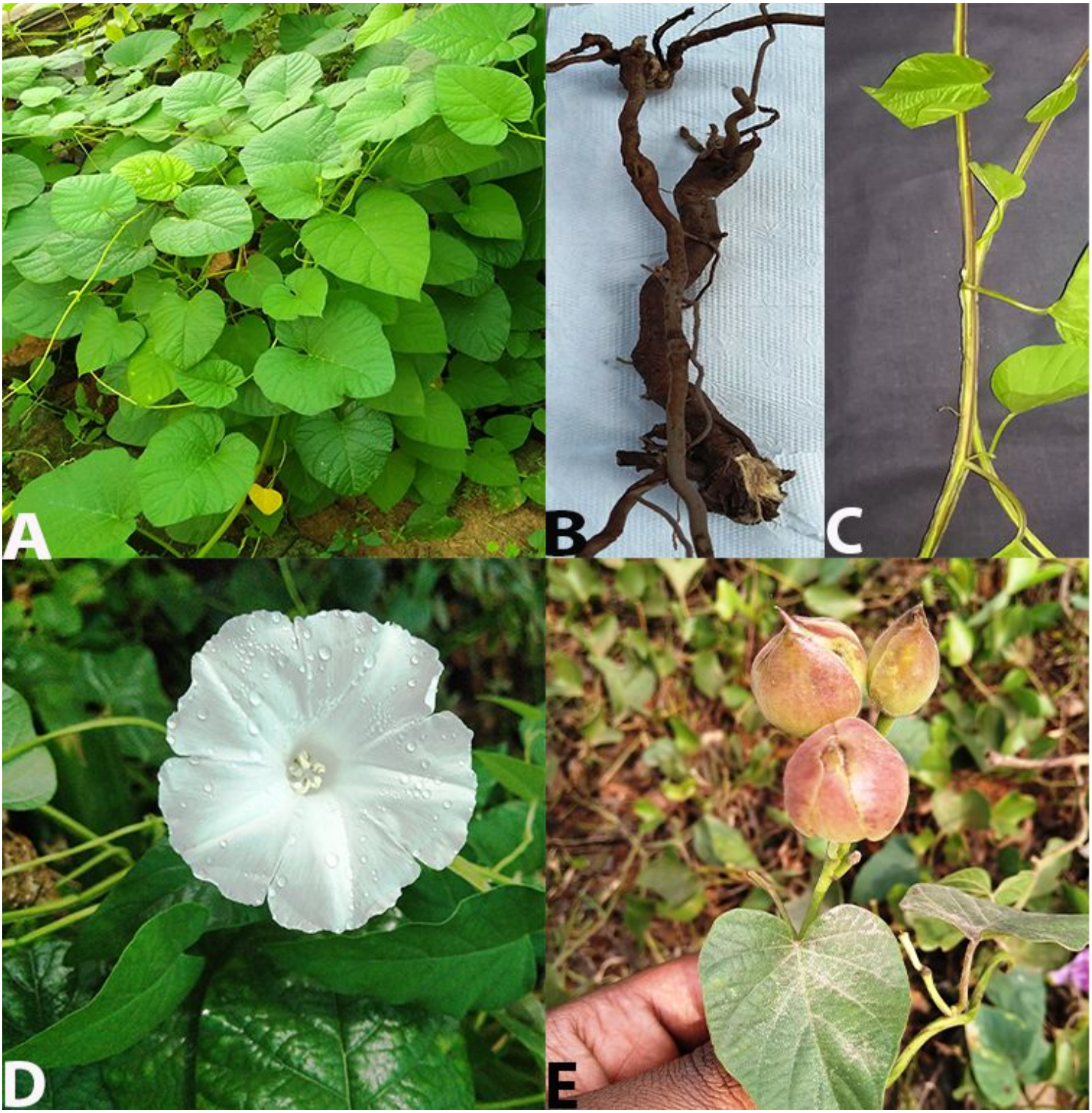

Figure 1

A) O. turpethum plant, B) Root, C) Stem, D) Flower and E) Fruit

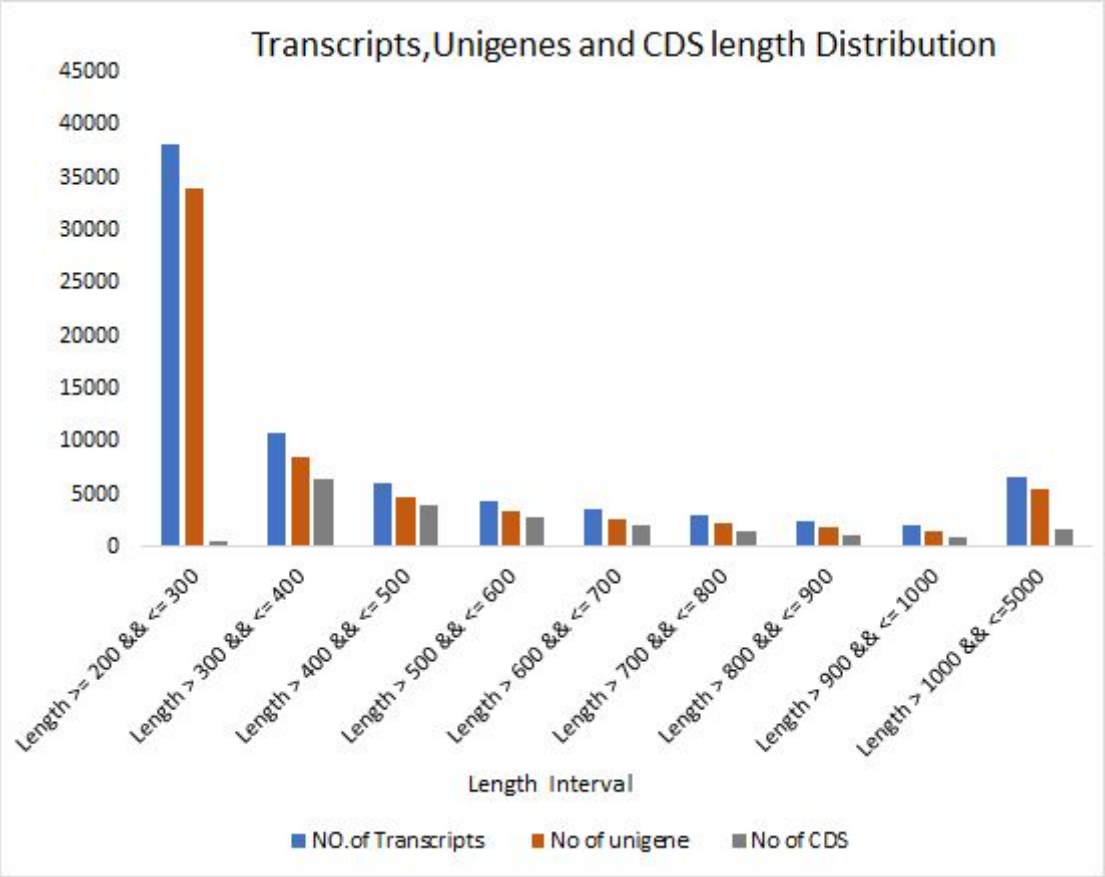

Figure 2

Transcripts, unigenes and CDS length distribution

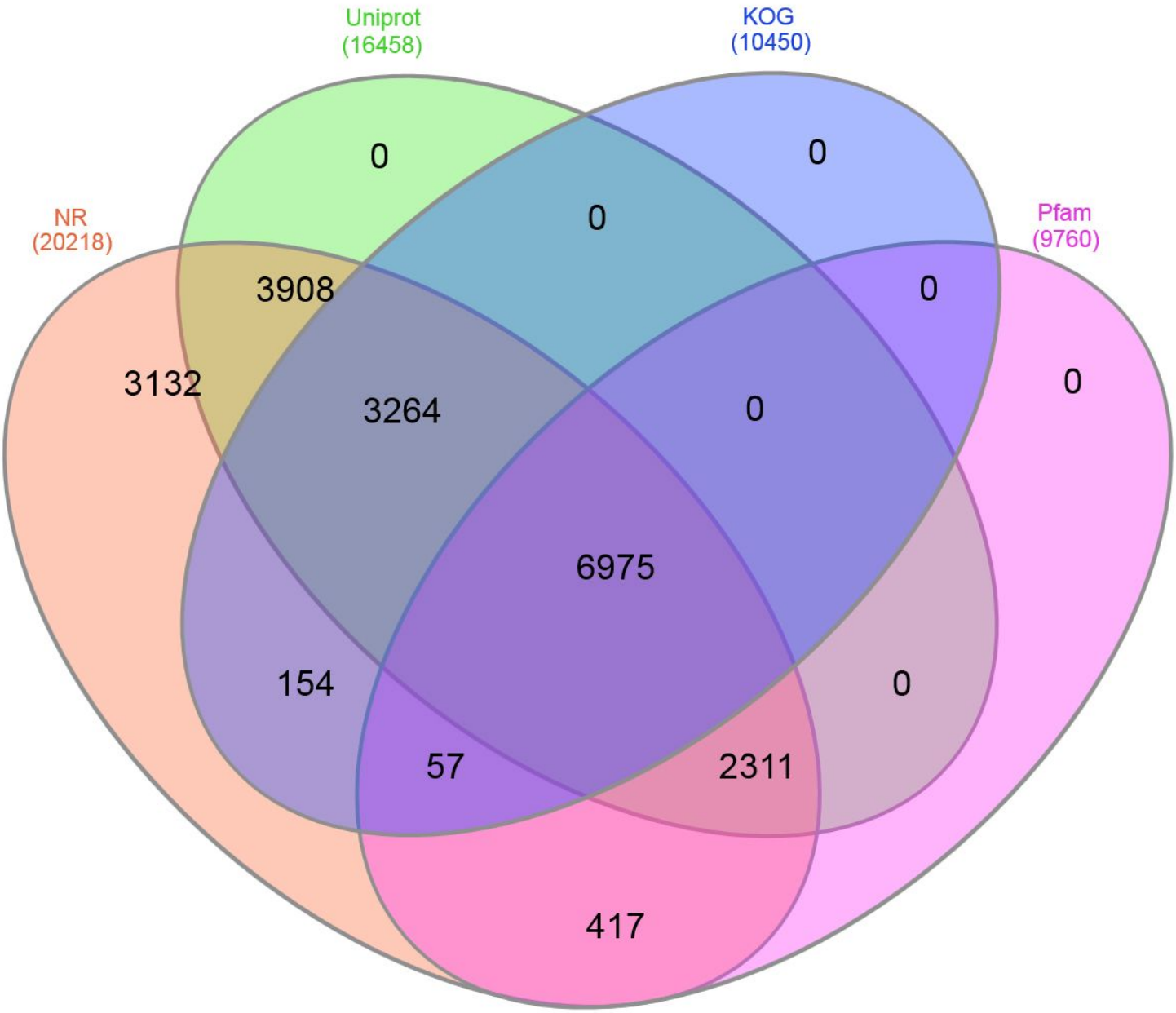

Figure 3

Venn diagram for Root-Stem annotated Unigenes in different databases

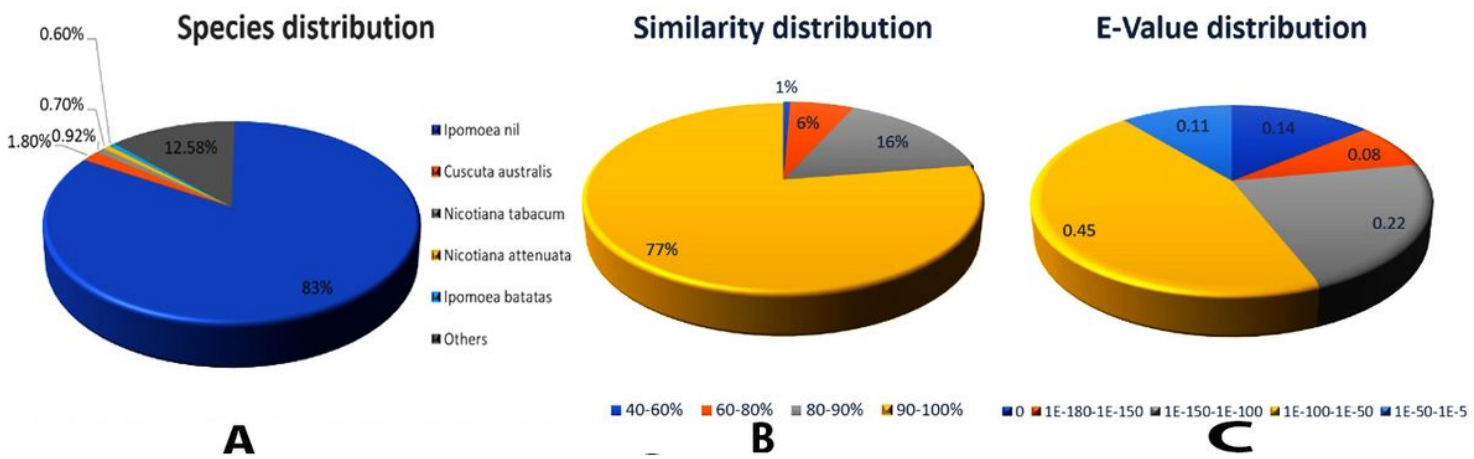

Figure 4

a) Species Distribution, b) E-value distribution and c) Similarity distribution

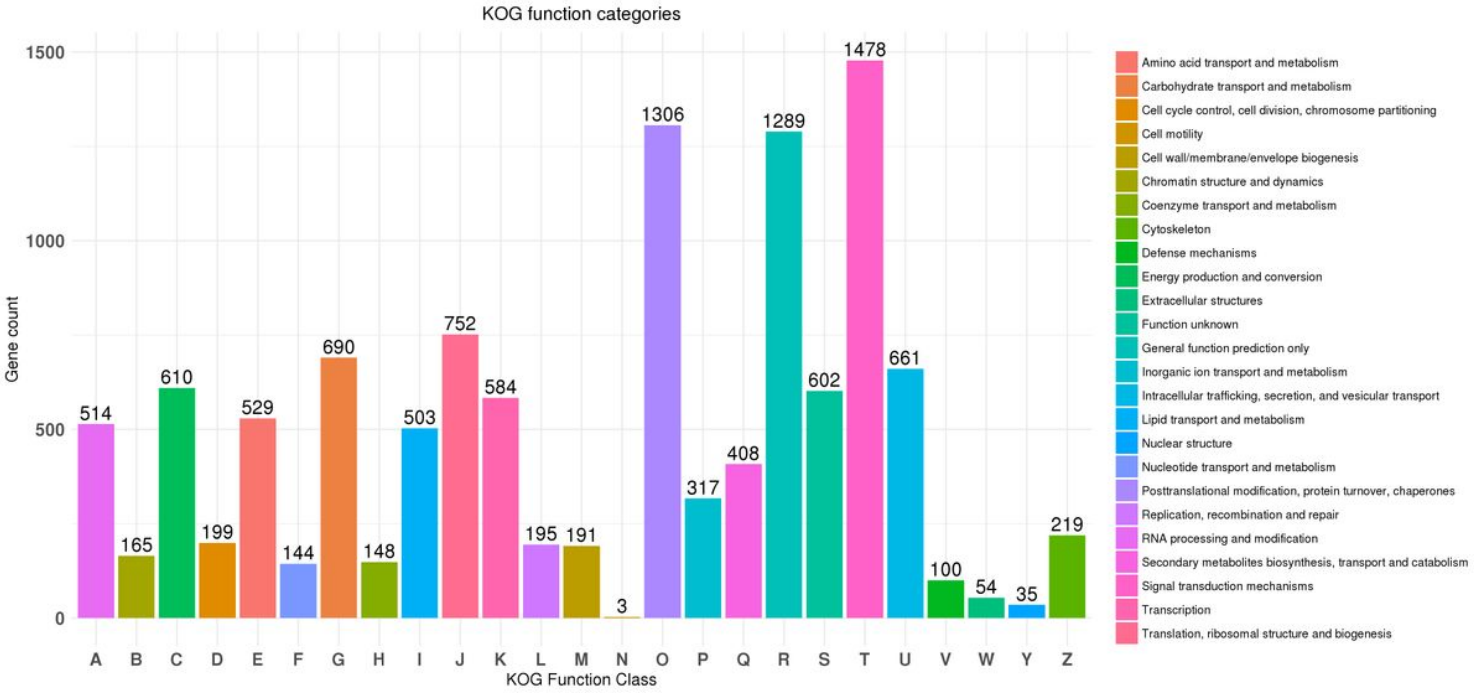

Figure 5

KOG classification for Root-Stem CDS

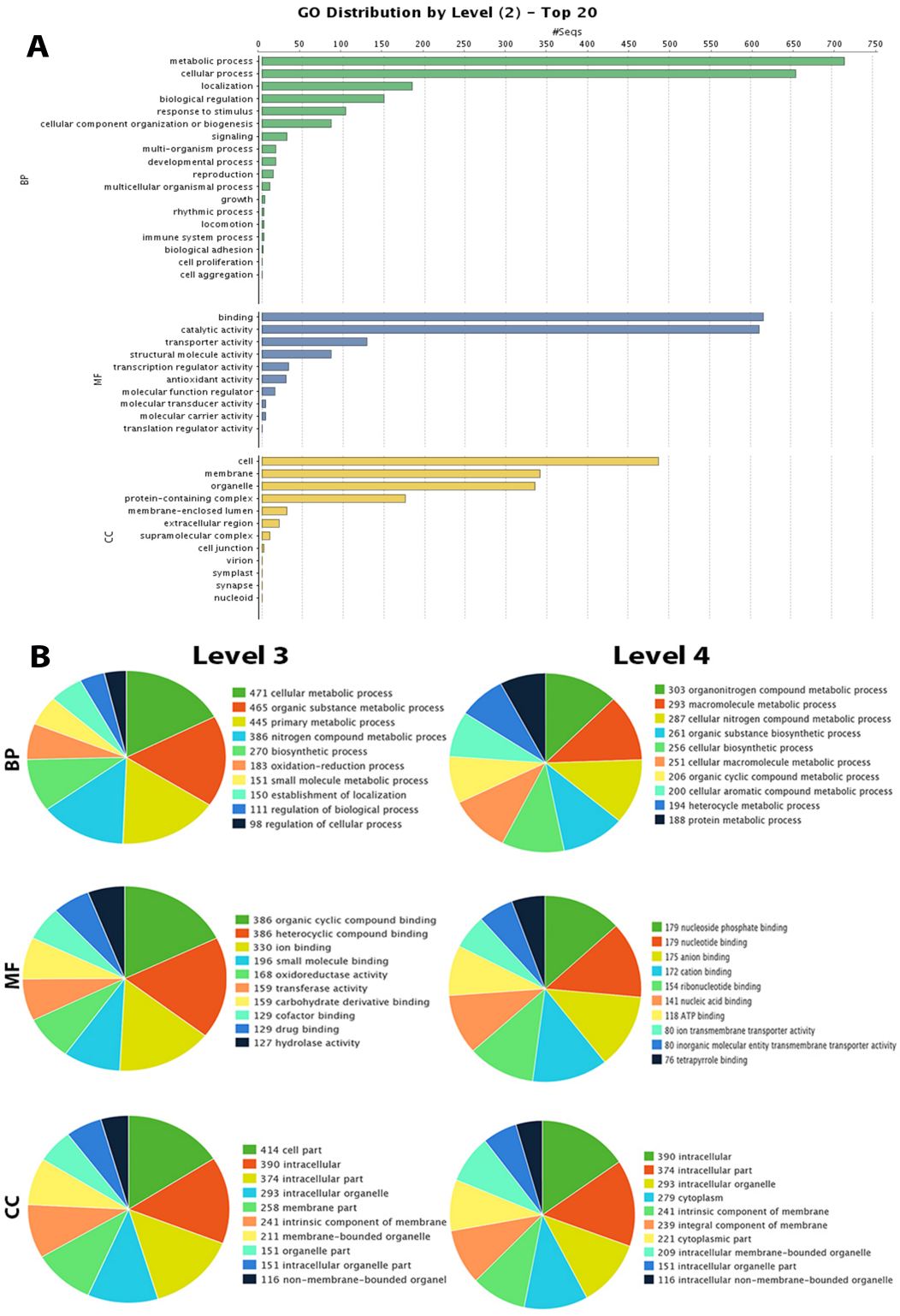

Figure 6

Gene Ontology distribution for Root-Stem sample. a) level 2, b) level 3 and 4. The Unigenes were assigned to three main categories: BP-Biological Processes, MF-Molecular Functions and CC- Cellular Components.

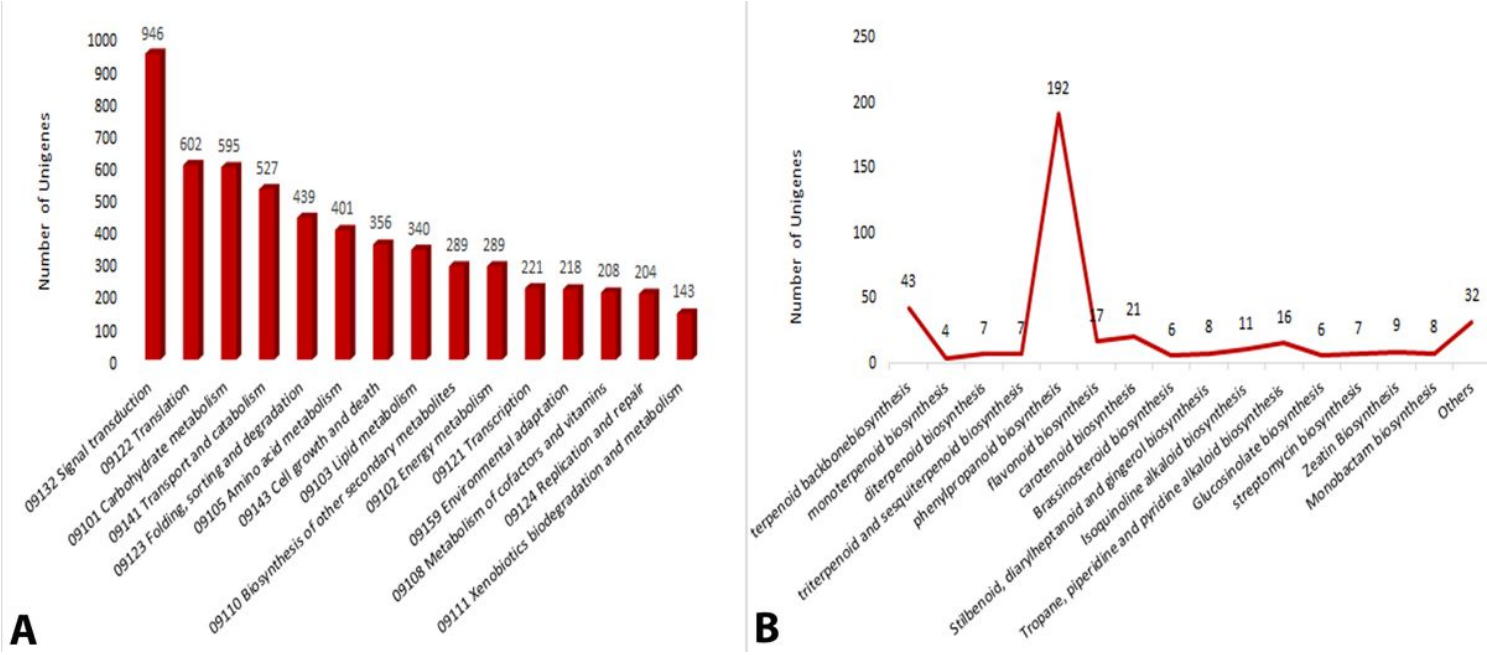

Figure 7

The top 15 largest KEGG pathways

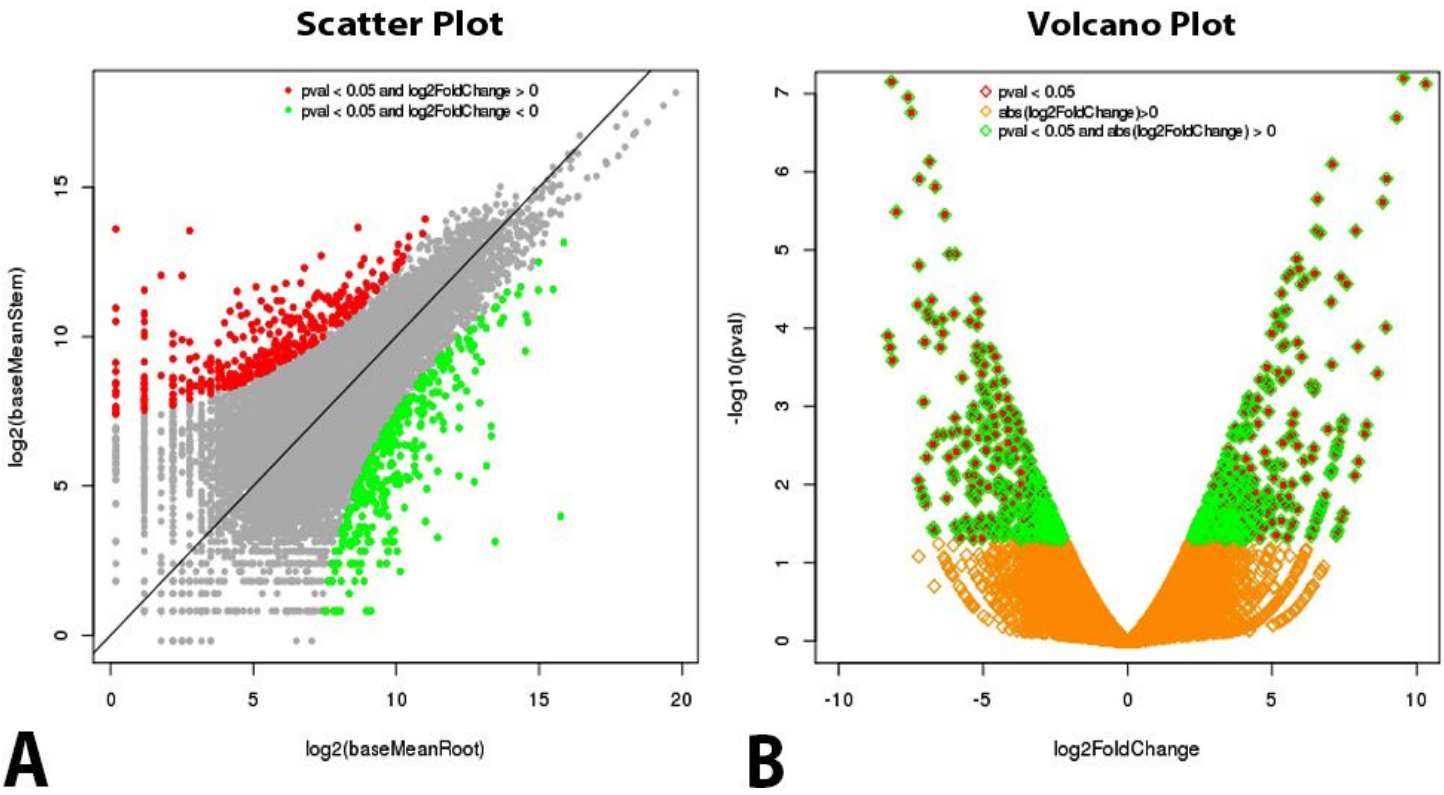

Figure 8

a) Scatter plot for normalized values obtained through DESeq basemean values of all differentially expressed genes in Root-vs-Stem. Each dot indicates one gene. The dots above the black diagonal line indicate up-regulated genes and below this line are down-regulated genes. Where red dots represent

significantly upregulated genes with Pvalue<0.05 and log2FC >0 and green dots represents significantly downregulated genes with Pvalue<0.05 and log2FC < 0. The horizontal coordinates represent the log2(basemean)values of Root and the vertical coordinates the Stem log2(basemean) values b) Volcano plots of the distribution of gene expression for Root-Vs-Stem samples. DESeq was performed to show the differentially expressed genes. Red, green, and orange correspond to genes with p-value < 0.05, absolute log2FC <0 and (p-value < 0.05 and absolute log2FC >0) respectively.

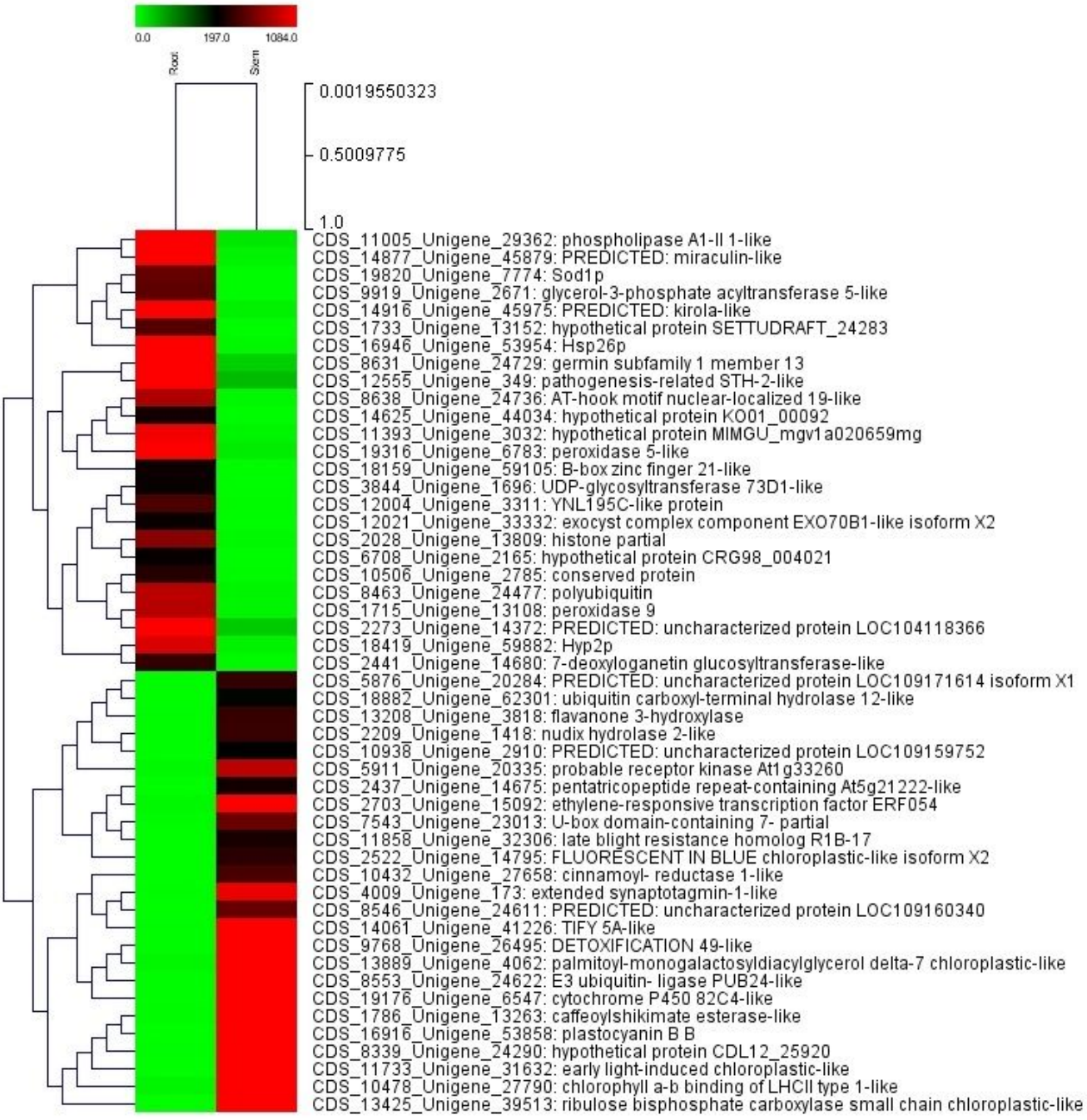

Figure 9

Heat map representing top 50 Highly upregulated and highly downregulated unigenes in Root-Vs- Stem.

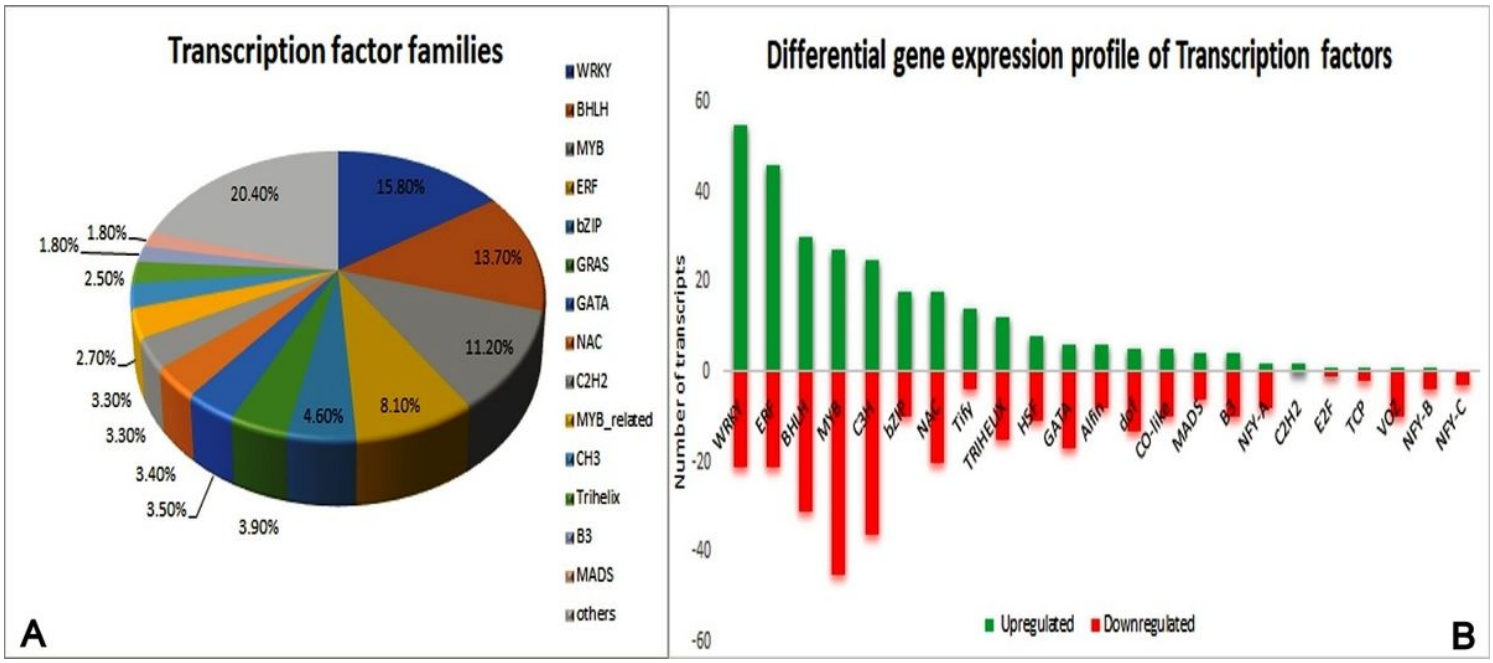

Figure 10

a) Transcription factor families and b) Differential gene expression profile of TFs

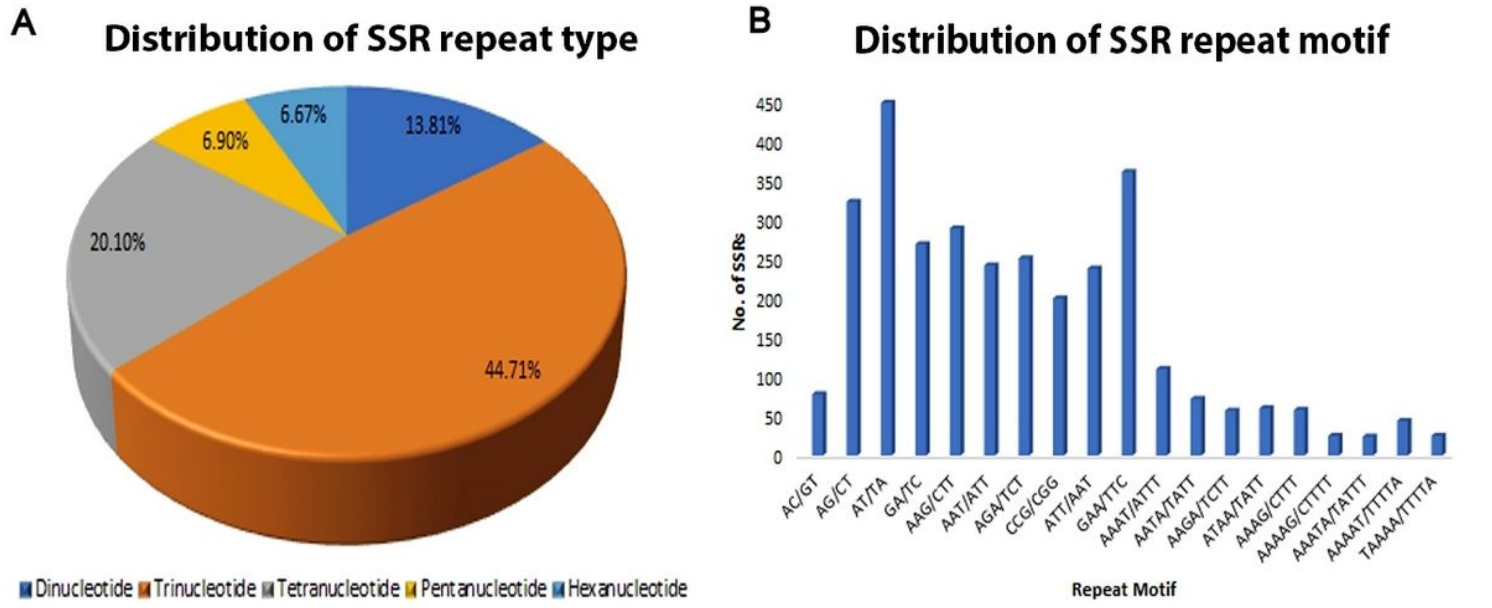

Figure 11

a) Distribution of SSR repeat type and b) Distribution of SSR repeat motif

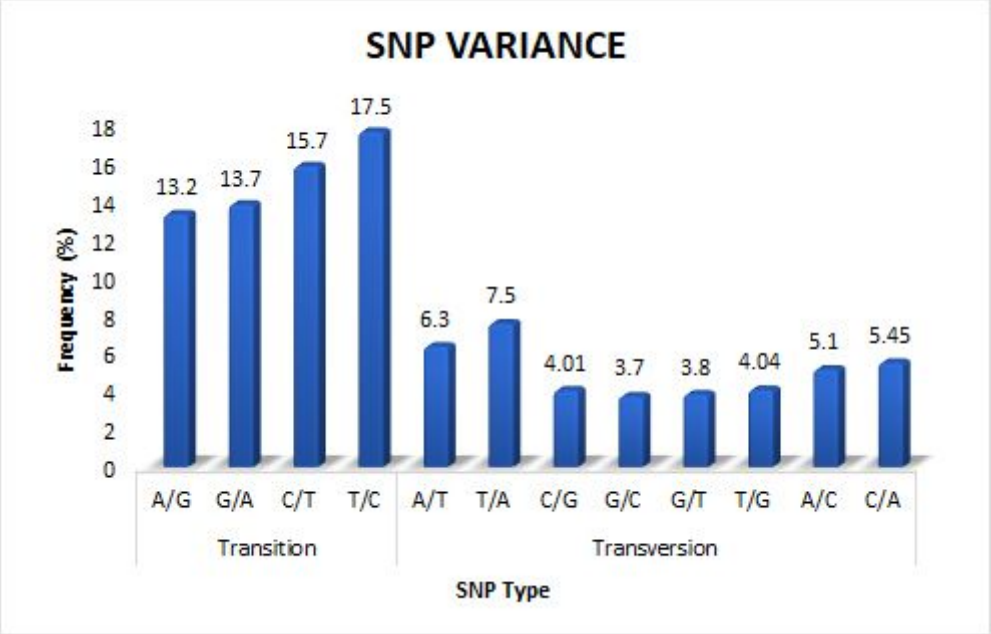

Figure 12

SNP Variance Chart

Supplementary Files

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

- SupportingInformationFiguresS1S7.pdf
- SupportingInformationTablesS1S8.xlsx