Genome-wide transcriptome analysis of triterpene biosynthetic genes of *Anoectochilus roxburghii* plant

**CURRENT STATUS:** POSTED

Hongzhen Wang  
Zhejiang A and F University

Haishun Xu  
Zhejiang A and F University

Peter E. Brodelius  
Linneuniversitet

Xueqian Wu  
Zhejiang A and F University

qingsong Shao  
Zhejiang A and F University

✉ sqszjfc@126.com **Corresponding Author**

Juan Xu  
Zhejiang A and F University

Bingsong Zheng  
Zhejiang A and F University

**DOI:**  
10.21203/rs.2.20025/v1

**SUBJECT AREAS**  
Epigenetics & Genomics

**KEYWORDS**  
*Anoectochilus roxburghii, RNA-seq, qPCR, gene-to-metabolites correlation, triterpenes biosynthetic gene*
Abstract

Background: Anoectochilus roxburghii is a medicinal plant and contains a variety of bioactive components, including triterpene, which exhibits important pharmacological properties with low toxicity. However, little is known about the biosynthetic pathway of triterpene or about the genome and transcriptome in A. roxburghii. Results: In order to analyze transcriptional determinants related to the biosynthesis of the bioactive components, we performed transcriptome sequencing in A. roxburghii (SRX1818644, SRX1818642 and SRX1818641) and annotated the sequences from three samples. In total, 137,679,059 clean reads were obtained, corresponding to 12.20 Gb of total nucleotides. They were then assembled into 86,382 contigs and 68,938 unigenes, which were further annotated according to sequence similarity with known genes in COG, EST, Nr, Pfam and Uniprot databases, leading to 10,040·29,442·39,551·34,991 and 28,082 unigenes, respectively. GO analysis classified all unigenes into three functional categories, i.e. biological processes (43,206 unigenes in 22 categories), molecular functions (46,978 unigenes in 15 categories) and cellular components (20,951 unigenes in 18 categories). Candidate triterpenes biosynthetic genes ArHMGR1 in MEV pathway, ArDXS1, ArDXS4 ArDXS5, ArDXS8-10, ArDXR1-2 and ArHDR1-2 in MEP pathway and ArFDS1, ArSM and ArOCS were selected based on RNA-seq and gene-to-metabolites correlation analysis. Conclusion: The transcriptomes of A. roxburghii plant include 86,382 contigs and 68,938 unigenes. The assembled dataset allowed identification of genes encoding enzymes in the biosynthesis of bioactive components in A. roxburghii plant. Candidate genes that encode enzymes being important in triterpenes biosynthetic pathway were selected. This will facilitate the study of expression and regulation in the biosynthesis of bioactive component in A. roxburghii.

Background

The genus Anoectochilus (Orchidaceae) is mainly distributed in tropical areas of Asia to Oceania and comprised of 43 species (http://www.theplantlist.org/1.1/browse/A/Orchidaceae/Anoectochilus/). Among them, species A. formosanus, A. koshunensis, and A. roxburghii are utilized in traditional Chinese medicine for treatment of diabetes, liver and cardiovascular diseases, cancer, malnutrition in children and other conditions [1, 2]. The wild plants are now in endangered condition mainly because
of excessive excavation and low seeds germination rate. Liu et al. have sequenced the transcriptome of *A. roxburghii* seeds to investigate the subject of seed germination [3]. Various triterpenes have been identified in *Anoectochilus* plant, such as, sorghumol,friedelin,oleanolic acid and ursolic acid [4, 5]. All these triterpenoids exhibit certain pharmacological benefits, such as antibacterial, antiviral [6], antipyretic effects [7].

In plant cells, triterpenoids are synthesized in cytosol from isopentenyl diphosphate (IDP), which is generated from mevalonate pathway (MVA) and 2-C-methyl-D-erythritol 4-phosphate (MEP) pathway (Figure 1A) [8]. In MVA pathway, 3-hydroxy-3-methyl-glutaryl coenzyme A reductase (*HMGR*) is a rate-limiting enzyme [9]. In MEP pathway, despite of 1-Deoxy-D-xylulose 5-phosphate synthase (*DXS*) plays a major role, 1-deoxy-D-xylulose 5-phosphate reductase (*DXR*) and hydroxy-2-methyl-2-(E)-butenyl 4-diphosphate reductase (*HDR*) may be involved too[10]. Using IDP and DMADP (dimethylallyl diphosphate) as substrates (Figure 1A), farnesyl diphosphate synthase (*FDS*) catalyzes the formation of farnesyl diphosphate (FDP), a central precursor for sesquiterpenes, triterpenes and sterols biosynthesis in cytosol. Early studies also indicated that there exists a small gene family encoding *FDS* in *Arabidopsis thaliana* [11]. As illustrated in Figure 1A, FDP is used to produce squalene, which is further converted by squalene monoxygenase (*SM*) to 2,3-oxidosqualene, a common precursor of triterpenoid biosynthesis (Figure 1A). Oxidosqualene cyclases (*OSC*s), which are product-specific, then cyclize 2,3-Oxidosqualene into different triterpenoid skeletons, customizing the first committed step of triterpenoid biosynthesis. The cyclization of 2,3-Oxidosqualene is initiated with the formation of triterpene skeleton, which undergoes several rearrangements to form different types of triterpenes. *OSC*s contain two groups of enzymes, monofunctional and multifunctional synthases [12, 13].

Although triterpenes are widely distributed in plant kingdom and known for their pharmaceutical benefits, understanding of their biosynthesis genes in *Anoectochilus* plants is still limited. Thus, transcriptome sequencing of *Anoectochilus* plant was carried out to reveal enzymes related to triterpenoids biosynthesis.

**Results**

Three cDNA libraries generated from RNAs of the whole *A. roxburghii* plants were subjected to paired-
end sequencing on an Illumina platform. A de novo assembly strategy was used to construct *A. roxburghii* transcripts. In total, 165,643,100 raw reads with nucleotides of 15.43 Gb were generated from the sequencing platform (Table 1). After stringent quality check and data cleaning, 137,679,059 quality reads were obtained. The number of clean reads is 83.12% of raw reads (Table 1). The average length of the clean reads is 95.21 bp (Table 1) with a high Q20 value of 99.24%. All clean reads were assembled into 86,382 contigs, corresponding to 68,938 unigenes using Trinity. 74.23 % of the clean reads could be matched to assembled unigenes, and 87.13% of these clean reads are unique and match to unigenes (Table 1). BUSCO tool [14, 15] was used to assess the quality and completeness of *A. roxburghii* plant transcriptome, the results showed that 1008 transcripts (70 %) of 1440 BUSCO genes were complete, among these complete genes, 978 were single copy and 30 were duplicated, 144 (10 %) were fragmented and 288 (20 %) were missing. Further analysis showed that the average length of the contigs and unigenes is 421 bp and 404 bp separately, and N50 lengths of contigs and unigenes are 464 bp and 434 bp, respectively (Table 1). All contigs and unigenes are longer than 200 bp, among those, 4012 contigs and 2822 unigenes are between 1000 and 2000 bp, and 245 contigs and 160 unigenes are longer than 2500 bp (supplementary Figure 1s). *A. roxburghii* transcriptome is 27874926 bp (around 27Mb) with a GC content of 46.81 %, which is higher than that of *Arabidopsis* (42.5 %), soybean (40.9 %) and chickpea (40.3 %) but lower than that of rice (55 %) [16].

**Function annotations** All unigenes were aligned to five public protein and gene databases, i.e. COG, EST, Nr, Pfam and Uniprot. Blastx was performed against these databases, and proteins and genes sharing highest similarities (E≤10^{-5}) to query unigenes were adopted to annotate unigenes. All 68,938 unigenes were annotated in this manner, among which 10,040 (14.56%), 29,442 (42.71%), 39,551 (57.37%) of 34,991 (50.76%) and 28,082 (40.73%) unigenes were annotated via COG, EST, Nr, Pfam and Uniprot, respectively.

Based on species distribution, the highest proportion of matching sequences among the unigenes annotated via Nr is from *Vitis vinifera* (19.60%), and the number of matched unigenes is 7,753, followed by *Oryza sativa* Japonica Group (7.62%), *Populus trichocarpa* (6.84%), *Ricinus communis*
(5.62%), Sorghum bicolor (5.00%), Glycine max (4.48%), Brachypodium distachyon (4.19%), Zea mays (3.99%), Oryza sativa Indica Group (3.55%), Hordeum vulgare subsp. vulgare (2.49%) and Medicago truncatula (2.25%). The number of matched unigenes to these species are 3,014, 2,707, 2,225, 1,979, 1,772, 1,660, 1,579, 1,406, 987, and 890, respectively. The lowest proportion of matching sequences is originated from Arabidopsis thaliana (1.67%) with matched sequences of 663.

**GO function categories** GO annotation classified all 68,938 unigenes to three categories, i.e. biological process, molecular function and cellular function. The biological process contains 22 categories, among these, 43,206 unigenes were assigned to metabolic processes, corresponding to the highest number of unigenes. Genes related to secondary metabolites biosynthesis, nucleic acid metabolic process, protein metabolic process and cellular metabolic process were annotated by GO. Cellular function group includes 15 categories in which catalytic processes hold the highest number of 46,978 unigenes. Cellular components can be divided into 18 categories, where cell part has the most unigenes of 20,951 (Figure 2). These GO annotations provide valuable information to investigate specific processes, molecular functions and cellular structures of A. roxburghii transcriptome.

**KEGG pathways** All unigenes were assigned to 316 KEGG pathways. These containing the most unigenes are metabolic pathways (2,693 unigenes, 5.35%, KEGG pathway ID: map01100). Within the KEGG metabolite pathway classification, secondary metabolite (such as glycoside, flavonoids, terpenoids and terpenoid backbones) biosynthetic pathways are of special interest due to their bioactivities in A. roxburghii plant. Terpenoid backbone biosynthesis (KEGG ID: map00900), diterpenoid biosynthesis (map00904), and sesquiterpenoid and triterpenoid biosynthesis (map00909) contain 61, 20 and 1 unigenes respectively. Here we mainly focus on terpenoid backbone biosynthesis, including the putative triterpenes biosynthetic pathways shown in Figure 1A. All the putative triterpenes biosynthetic genes have E value ≤1e-74 (supplementary Table S1), when blasted with best matching sequences in NCBI, except ArOSC3 (comp324881_c2) matches best to Ananas comosus, all putative triterpenes biosynthetic genes showed best matches to orchid plant, such as Dendrobium officinale, D. huoshanense, D. catenatum and Phalaenopsis equestris (supplementary Table S1).
Gene-to-metabolites correlation analysis

Triterpenoid accumulation

Triterpene accumulation was analyzed using oleanolic acid as a general evaluation index. Triterpenoid content was determined by test oleanolic acid using HPLC. Triterpene accumulation in root, stem and leaf showed significant difference (P<0.05). Triterpenoid accumulation is the highest in root (3.75 ± 0.03 mg/g, fresh weight, FW), lowest in the leaves (1.42 ±0.05 mg/g, FW). The content is medium in the stems (1.60±0.04mg/g, FW).

Relative expression level of triterpene biosynthetic genes in A. roxburghii

RT-qPCR was used to analyze relative expression of putative triterpene biosynthetic genes, including ArHMGR in MEV pathway, ArDXS, ArDXR and ArHDR in MEP pathway and ArFDS, ArSS and ArOCS of triterpene biosynthesis, respectively. All tested genes expression showed significant difference (P≤0.05) or extremely significant difference (P≤0.01) in root, stem and leaf.

Relative expressions of HMGR, DXS, DXR and HDR

HMGR is a rate-limiting enzyme in MVA pathway [9]. In A. roxburghii, two unigenes were annotated as ArHMGR1-2 (Supplementary 1, Supplementary Table S1-2). As shown in Figure 1B, ArHMGR1 showed highest expression in root, moderate expression in stem and scarcely expressed in leaf, while ArHMGR2 showed different expression pattern (Figure 1B). Variable correlation coefficients between ArHMGR1-2 expression and triterpene accumulation are 0.877 and 0.051 respectively (Table 2). The data suggest that, ArHMGR1 may play roles in triterpene biosynthesis as its relative expression level correlates with triterpene accumulation, while ArHMGR 2 may be irrelevant with triterpene biosynthesis.

High plants retain both MVA and MEP pathways [17]. They have similar functions in ginsenoside biosynthesis in ginseng root but MEP expressed a much higher level than MEV in ginseng Leaf [18]. To understand whether MEP pathway has any influence on triterpene biosynthesis in A. roxburghii, DXS, DXR and HDR in MEP pathway were analyzed. There are ten unigenes annotated as ArDXS1-10. Among them, ArDXS1, ArDXS4 ArDXS5 and ArDXS10 were highly expressed in root, but scarcely expressed in stem and Leaf. (Figure 1B). The relative expression levels of ArDXS 8-9 are very low in
root, stem and leaf. However, root contains more ArDXS 8-9 than in stem, least in leaf. Variable correlation coefficients between the individual gene expression level of ArDXS1, ArDXS4 ArDXS5, ArDXS8-10 and triterpene accumulation, are 0.930, 0.987, 0.981, 0.950, 0.921 and 0.929, respectively (Table 2). These high correlation coefficients suggest a strong relationship of triterpene accumulation with all the six ArDXS genes, and these ArDXSs are candidate genes in MEP pathway for triterpene biosynthesis. The levels of the remaining four genes ArDXS2, ArDXS3, ArDXS6 and ArDXS7, however, only exhibit a moderate correlation with triterpene accumulation. The calculated Pearson correlation coefficients are 0.497, 0.596, 0.715 and 0.732, respectively (Table 2).

DXR and HDR are also rate-limiting enzymes in MEP pathway in plant, but their roles vary with plants and conditions [10]. In this study, two unigenes were annotated as ArDXR1 and ArDXR2 and another two unigenes as ArHDR1 and ArHDR2 (Supplementary 1, Supplementary Table S1-2). ArDXR1-2 and ArHDR1-2 showed relative higher expression in root than in stem and leaf. Variable correlation coefficients between triterpene accumulation and the individual gene expression of ArDXR1-2 and ArHDR1-2 are 0.941, 0.944, 0.944 and 0.861 respectively, suggesting that ArDXR1-2 and ArHDR1-2 are positively correlated with triterpene accumulation. Therefore, we believe that ArDXR1-2 and ArHDR1-2 are candidate genes contributing in triterpene biosynthesis.

**Relative expression of FDS, SM and OSC**

Farnesyl diphosphate synthase (FDS) catalyzes the formation of FDP (Figure 1A), a central precursor for biosynthesis of sesquiterpenes, triterpenes and sterols. Plants contain a small gene family encoding FDS [11] and A. roxburghii has only one. ArFDS showed similar expression in tested tissues. Relative to the reference genes, the levels are 1.7 in root, 1.2 in stem and 1.5 (Figure 1B) in leaf, respectively.

Cyclization of 2,3-oxidosqulene is the committed step of triterpene biosynthesis to produce different triterpene skeletons by OSC [19]. In this study, one gene was annotated as ArSM (Supplementary 1, Supplementary Table S1-2), which showed highest expression in root, moderate expression in stem and lowest expression in leaf. The correlation coefficient between ArSM expression levels and triterpene contents was up to 0.955 (Table 2), suggesting that ArSM plays a role in the accumulation
of triterpene and is a candidate gene catalyzing the formation of 2,3-oxidosquelen.

Three genes were annotated as ArOCS1-3 (Figure 1B, Supplementary 1, Supplementary Table S1-2).

Although ArOCS1 showed higher relative expression in root, moderate in stem and lower in leaf, its expression level is significantly correlated with triterpene content. Therefore, we believe ArOCS1 is a candidate enzyme that catalyzes triterpene skeletons formation.

**Discussion**

RNA sequencing, is attractive for study of non-model medicinal pant without genomic sequence information [20]. With this technology, 15.43 Gb of raw reads of A. roxburghii transcriptome were obtained. Trinity achieved the best overall metric score of 95.9 comparing with the other assembly tools, such as SOAPdenovo-Trans and Trans-ABySS, and performed generally well in constructing full-length transcripts [21]. 86,382 contigs, corresponding to 68,938 unigenes were assembled using Trinity. A. roxburghii transcriptome showed 70 % completeness, transcriptome studies of Arabidopsis thaliana and Noccaea caerulescens showed higher completeness percentage, which was 99.3% [22, 23] and 90 % [24], respectively. However, moderate BUSCO completeness of A. roxburghii transcriptome is comparable to other plants, such as Camellia nitidissima flower transcriptome, which has a moderate to high level completeness [25], Aethionema Arabicum seeds transcriome (63.1 %) [22] and Euphorbia pulcherrima (77 %) and E. pekinensis transcriptome (77 -81 %) [26].

Meanwhile, different BUSCO completeness levels were observed in different growth stage of same tissue and different tissues of the same species [26, 27]. The moderate BUSCO completeness level in this study is expected, since A. roxburghii plant represents a specific stage that lacks of the transcription which is active in flower or seeds.

The average length of contigs and unigenes (421 bp and 406 bp, respectively) are longer than the average transcript size from OrchidBase (347 bp) [28]. All the assembled genes are longer than 200 bp and annotated. The analysis results suggest that the transcriptome has good quality and can be used for deeper analysis [29]. Comparing A. roxburghii plant transcriptome with seeds transcriptome [3], genes related to nuclear-transcribed mRNA catabolic process, such as genes encoding ribosomal proteins and RNA polymerase α-subunit were enriched in whole plants, and these genes are silent in
seeds [3]. Leaf is an active tissue in photosynthesis. Compared with the dataset of seeds, associated with photosynthetic process is significantly enriched in our dataset. Genes that are important for growth and development, especially these related to carbohydrate, energy, and translation metabolism, as well as other metabolic pathways, are also present in the dataset. The observation was consistent with Aethionema arabicum seeds which lack of transcripts in photosynthetically/developmentally active tissue [30]. In the function annotation section, the top-hit specie is Vitis vinifera, a dicot, which is also found to be the best matches with other orchid plant, such as Phalaenopsis [31, 32], Cymbidium faberi Rolfe and Chiloglottis trapeziformis transcriptome [33, 34]. Dicot plant transcriptome, such as Panax japonicas also exhibits highest matches with Vitis vinifera [35]. Despite many species have a much higher number of uniGenes or ESTs similar to grapevine, the biological sense of this correlation remains unclear.

**Gene-to-metabolites correlation analysis**

Triterpene and triterpene saponin are mainly accumulated in root in some plants [36, 37]. Triterpene accumulation in A. roxburghii aligns with this observation. This study also analyzed the genes based on RNA-seq and gene-to-metabolites correlation. The number of HMGR genes varies from plant to plant [38]. There are two HMGR genes reported in Panax ginseng ‘Meyer’ plant, which are PgHMGR1 and PgHMGR2 [39]. HMGR1 activity is positively correlated with ginsenoside production. Overexpression or inhibited expression of PgHMGR1 will result in enhanced or reduced triterpenes production in ginseng. PgHMGR1 plays a major role in providing triterpenes in ginseng, whereas PgHMGR2 follows a different regulatory pathway [39]. HMGR represents a key enzyme in the synthetic pathway of Alisol B 23-acetate in Alisma orientale (Sam.) Juz. Alisol B 23-acetate, a kind of triterpene, is the main active component and the general evaluation index for the quality of A. orientale (Sam.) Juz. HMGR expression showed significant positive correlation with Alisol B 23-acetate content in A. orientale (Sam.) Juz plant [40]. Here two unigenes were annotated as ArHMGR1-2. They showed different expression patterns, suggesting existence of at least two functionally different HMGR copies in A. roxburghii genome. ArHMGR1 may play some roles in triterpene biosynthesis while ArHMGR2 may have some other roles.
Contribution of MEP pathway to ginsenoside biosynthesis differs with tissues of ginseng [18]. DXS, DXR and HDR are rate-limiting enzymes in MEP pathway [10]. These three enzymes are encoded by small gene families in plants [10]. The presence of multiple genes encoding DXS appears to be widespread in plant [10]. The number depends on the plants and ranges 2 to 13 [10, 41]. In this study, 10 DXSs were found. ArDXS1, ArDXS4 ArDXS5 and ArDXS8-10 showed similar expression level which is significantly and positively correlated to triterpene biosynthesis. This indicates that all these six DXSs genes may belong to DXS2 clade, and play some roles in secondary terpenoid biosynthesis [10]. The other four DXSs may have some other roles. DXR is encoded by two genes in Hevea brasiliensis [42], HDR is encoded by two genes in Ginkgo biloba and Pinus taeda plant [10]. In this study, DXR and HDR are also encoded by two genes, which shows similar expression.

There are two FDS genes were respectively identified in Arabidopsis [43], maize [44] and wheat [45]. FDS1 expresses in all tissues at all developmental stages and provides FPP for isoprenoids serving basic plant cell functions [11, 43-45]. FDS2 expresses in specific tissues at particular stages of the development and provides FPP for isoprenoids with particular functions [11, 43-45]. In the study only one FDS gene was found, it showed similar expression in different tissues. It is still possible that two FDS genes exist in the plant but only one was annotated presumably due to their similarity in sequence. It is speculated that FDS1 constitutes the main part of the annotated gene and FDS2 expression in different tissues is masked, similar to the pattern of FPS expression in Artemisia annua [46].

Number of SM genes varies from 1 to 6 in plants [47-49]. Their over expressions can be utilized to increase triterpene production in plants and yeast [47]. In the study, one ArSM was annotated. It may play important role in triterpene biosynthesis because it’s expression is correlated with triterpene biosynthesis.

OSC is a key enzyme in triterpene biosynthetic pathway [12, 13, 50]. Copy of OSC varies with plants. There are four monofunctional and one multifunctional OSCs in Banaba [12]. However, only two multifunctional OSCs exist in Apple [13]. Taraxacum coreanum, instead contains two monofunctional and two multifunctional OCSs [50]. In the study, three OSCs genes were identified, ArOCS1 expression
profile is positively corrected with the accumulation of oleanolic acid. Therefore, it may play important role in the formation of oleanolic acid skeleton.

In summary, above discussed candidate genes may function in triterpene biosynthesis in the plant. They are positively correlated each other and showed similar expression profile, confirming the idea that genes that participate in same component biosynthesis are co-expressed, and simultaneously up or down regulated under same endogenous or exogenous stimuli [51].

Conclusion
The transcriptomes of A. roxburghii plant include 86,382 contigs and 68,938 unigenes. The assembled dataset allowed identification of genes encoding enzymes in the biosynthesis of bioactive components in A. roxburghii plant. Candidate genes that encode enzymes being important in triterpenes biosynthetic pathway were selected. This will facilitate the study of expression and regulation in the biosynthesis of bioactive component in A.roxburghii.

Methods
Plant materials
A. roxburghii plants were grown in pots containing soil specific for orchid plant under 16 h light and 8 h darkness in a greenhouse at Baicaoyuan of Zhejiang A and F University, Hangzou, China. roxburghii seeds were collected from the Baicaoyuan, their identity were confirmed by Professor Runhuai Hu at Zhejiang A & F University. Based on national regulations and Convention on the Trade in Endangered Species of Wild Fauna and Flora: (http://www.cites.org.cn/), no permits were required for the collection of A. roxburghii. The plant were grown at 25 ± 2 °C and with 80-100 % humidity for about three months to a height of around 10 cm. The whole plant was harvested and frozen in liquid nitrogen, then grounded into fine powder for RNA extraction.

RNA extraction and RNA-seq
Total RNA was extracted from A. roxburghii plant using Purelink™ Plant RNA Reagent (Invitrogen) according to the manufacture’s instruction. A pool from five A. roxburghii plants was used as one sample. Each sample has triplicates. Genomic DNA was removed by the treatment with DNase I (Fermentas). RNA integrity was checked on 1 % agarose gel, and RNA concentration and purity were
determined using NanoDrop 2000C spectrophotometer (Thermo Scientific, USA). mRNA was purified from total RNA using oligo (dT) magnetic beads and fragmented into small pieces in fragmentation at 37 °C for 10 min. First stranded cDNA was synthesized using Superscript II (Promega, USA) and random primers. The second stranded cDNA synthesis was subsequently performed using DNA polymerase I and RNaseH (Promega). Double-stranded cDNA was purified with AMPure XP system (Beckman Coulter, Beverly). The purified double-stranded cDNA was end-repaired using T4 DNA polymerase (NEB), Klenow enzyme (NEB), and T4 polynucleotide kinase (NEB) followed by a single A base addition using Klenow exo-polymerase, then ligated with adapter using DNA ligase (NEB). To preferentially select cDNA fragments of 300bp in length, library fragments were purified with AMPure XP system (Beckman Coulter, Beverly) and amplified by PCR to create a cDNA library. The cDNA library was sequenced using Illumina 2000 sequencing platform and the average length of sequenced reads was 100 nt. Transcriptome sequencing was performed by staff at Zhejiang Tianke High technology development Co. Ltd (Hangzhou, China).

**Analysis of Illumina sequencing results**

Raw reads generated from sequencing machine were filtered to obtain clean reads by removing the adapter sequences, repetitive, redundant and low-quality sequences from the raw reads. Predicted errors of raw reads were corrected using Rcorrector under default settings [52] and adapter sequences of all raw reads were removed using ILLUMINACLIP: adapters/TruSeq3-PE-2.fa:2:30. The quality of the trimmed and filtered reads were assessed using FastQC v0.10.1, low-quality reads with more than 40% of quality value ≤ 20 bases or the reads with unknown nucleotide percentage > 4% using NGS QC Toolkit v2.3. Clean read quality was evaluate with quality score (Q score) higher than 20 (Q20), the higher percentage of Q20 clean reads means the better quality. The data was uploaded to NCBI SRA (ID: SRP076090).

Clean reads were assembled into contigs with Inchworm with parameters of k value was 25 with step size 5 [53, 54]. The assembled contigs were clustered using Chrysalis [54], all overlapped contigs were connected into components using a de Bruijn graph approach. In a final step, butterfly simplifies
all the generated graphs with parameters ‘min kmer cov 2 –min contig length 100 –run’ to generate unigenes [54]. Clean reads were aligned to assembled unigenes using SOAP2 [55], duplicated reads and multi-mapped reads were filtered from the alignment results to eliminate the PCR interference and ambiguous mapping. All unigenes were aligned by BLASTx (E-value \(1 \times 10^{-5}\)) against COG, EST, non-redundant (Nr), Pfam and Uniprot databases. Best-alignment results were used to determine unigenes sequence. Unigenes showed best matches with fungal species were removed.

Benchmarking Universal Single-Copy Orthologs (BUSCO) with default parameters was applied to assess the completeness of assembled transcriptome [14, 15], and the embryophyta_odb9 database, which contains 1440 total BUSCO groups was used [56]. With Uniprot annotation, all unigenes were annotated according to biological process, cellular component, and molecular function by Gene Ontology (GO) database [57, 58]. The predicted protein sequences were submitted to the KEGG (http://www.genome.jp/tools/kaas/) method. KEGG pathway annotation was performed by BLAST analysis against automatic annotation server (KAAS) KAAS (http://www.genome.jp/tools/kaas/) with the bi-directional best-hit (BBH), Arabidopsis thaliana was used as reference species [59].

**Analysis candidate triterpenes biosynthetic genes**

The candidate triterpenes biosynthetic genes were obtained based on KEGG pathways and functional annotations. Unigenes involved in triterpene biosynthesis were aligned with plant protein sequences from public databases protein databases: Non-redundant (Nr) protein database, the Swiss-Prot protein database, the Kyoto Encyclopedia of Genes and Genomes (KEGG) database and the Cluster of Orthologous Groups of proteins (COG) database using BLASTx (E-value \(1 \times 10^{-5}\)). Sequence hits in a database would not a search against the next. In the study, RNA-seq, assembling with trinity collected in August, 2016. Go annotation and KEGG pathway constantly updated to August, 2019. BUSCO data collected in 8, 2019. Triterpene accumulation and QPCR data in this version collected in August, 2019.

**Quantitative analysis of triterpene**

Oleanolic acid was used as evaluation index for quantity analysis of triterpene. Quantitative analysis of oleanolic acid was carried out according to the literature [60] with minor modifications. 0.5 g of fresh leaves was homogenized in 5 mL methanol and treated with ultrasonic waves for 90 min (60W,
After extraction, the sample was filtered through a 0.45 µm filter membrane before being injection for analysis. The quantitation analysis was done by means of standard curves. HPLC analyses were performed using a Waters™ 2695 HPLC pump separation module (Milford, MA, USA) equipped with an auto-injector system (100 µL). The mobile phase was acetonitrile and 0.5% acetic acid as a ratio of 85:15. Flow rate was set to 0.8 mL/min. Agilent TC-C18 column (250 mm×4.6 mm, 5 µm) was used and the column temperature maintained at 25°C. The effluent was measured at a wavelength of 210 nm for the detection of ursolic acid. Acid content was calculated using a standard equation $y = 71820x - 32554$, $R^2 = 0.9995$, $x$ was the concentration of extract while $y$ was the absorbance of triterpenoids, linear range from 10 - 400 µg / ml. Triterpene content was calculated according to the formula $C = 5x/0.5$, $C$ was triterpene content. Each sample has three biological repeats and each biological repeat was analyzed in triplicates.

Quantitative real-time PCR analysis

Total RNA extraction and first DNA synthesis and quantitative real-time PCR (qPCR) was performed according to published protocols [61, 62]. Each sample has three independent biological repeats, and technical triplets and a negative control of each sample were run. qPCR was performed using specific primers (Supplementary Table S1-2), the cycle threshold (CT) values were used for further analysis. The primer efficiency of each amplicon was calculated by the program Linreg v. 12.1 based on the log linear phase of the amplification curve [63]. The Best-Keeper software [64] was used to search for stable reference genes among all genes tested. Based on Best-Keeper and related report, β-actin (Genbank ID: JF825425) and β-elongation factor (Genbank ID: JF825420.1) genes were used as reference [65], The relative expression level of triterpene biosynthetic genes in root, stem and leaf were compared with reference genes using the $2^{-\Delta\Delta Ct}$ method [66], relative expression of reference genes was set as 1.0. Expression profile of candidate triterpene biosynthetic genes in root, stem and leaf were shown as heatmap using the TBTools software (https://github.com/CJ-Chen/TBtools/releases).

Gene-to-metabolite correlation analysis

The expression of triterpene biosynthetic genes and the triterpene accumulation in different tissues
were correlative analysis using Pearson’s correlation analysis in IBM SPSS Statistics 24 software (IBM, New York, USA). Gene-to-metabolite correlation was constructed to identify triterpene biosynthetic genes in A. roxburghii plant.

Statistics analysis
Data are presented as mean ± standard error of biological triplicates. Statistical significance was determined by using one-way ANOVA in IBM SPSS Statistics 24 software (IBM, New York, USA).

Abbreviations
a-AS: a-amyrin synthase; β-AS: β-amyrin synthase; FDS: farnesyl diposphate synthase; CYP: CytochromeP450 monooxygenase; FS: fridelin synthase; HMG: 3-hydroxy-3-methyl-glutaryl coenzyme A reductase; IDI: isopentenyl diphosphate isomerase; OSC: oxidosqualene cyclases; SM: squalene monoxygenase; SOS: Sorghumol synthase; SS: squalene synthase. Plastid: DXR: 1-deoxy-Dxylulose-5-phosphate reductoisomerase DXS: 1-deoxy-D-xylulose-5-phosphate synthase; HDR: hydroxy-2-methyl-2-(E)-butenyl 4-diphosphate reductase; IDI: isopentenyl diphosphate isomerase;

Declarations
Ethics approval and consent to participate
Not applicable.

Consent for publication
The authors have full consent to publish. There are no person’s data in the manuscript to report.

Availability of data and materials
De novo transcriptome sequencing result was deposited in NCBI SRA with ID of SRP076090.

Sequences of genes encoding enzymes of triterpene biosynthesis were provided in the supplementary part.

Complemeting of interests
All authors declare there are no conflict of interests

Funding
This work was supported by National Natural Science Foundation of China (Grant No. 81673531, 81303167). Natural Science Foundation of Zhejiang province (Grant No. LY14H280008) and talent
starting funding from Zhejiang Agriculture and Forestry University (Grant No. 2014FR062). Science and Technology planning project of Zhejiang province (Grant No. 2017C02012). These funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Authors’ contributions
HW and QS initiated this project. HW collected and analyzed the data, performed all experiments. HW, XW, HX, JX, ZB wrote the manuscript. HW and PB reviewed and edited the paper. All authors read and approved the final manuscript.

Acknowledgements
The authors would like to acknowledge Zhejiang Tianke High technology development Co. Ltd for RNA-seq and assistance for data analysis.

References
1. Hsieh CC, Hsiao HB, Lin WC: A standardized aqueous extract of Anoectochilus formosanus modulated airway hyperresponsiveness in an OVA-inhaled murine model. Phytomedicine: international journal of phytotherapy and phytopharmacology 2010, 17(8-9):557-562.

2. Tseng CC, Shang HF, Wang LF, Su B, Hsu CC, Kao HY, Cheng KT: Antitumor and immunostimulating effects of Anoectochilus formosanus Hayata. Phytomedicine: international journal of phytotherapy and phytopharmacology 2006, 13(5):366-370.

3. Liu SS, Chen J, Li SC, Zeng X, Meng ZX, Guo SX: Comparative Transcriptome Analysis of Genes Involved in GA-GID1-DELLA Regulatory Module in Symbiotic and Asymbiotic Seed Germination of Anoectochilus roxburghii (Wall.) Lindl. (Orchidaceae). International journal of molecular sciences 2015, 16(12):30190-30203.

4. He CN, Wang C. L., Guo S. X., Yang J. S., Xiao P. G., : Study on chemical
constituents of *Anoectochilus roxburghii* (III). *Natural Product Research and Development* 2005, 17:259-262.

5. Huang L, Chen T, Ye Z, Chen G: *Use of liquid chromatography-atmospheric pressure chemical ionization-ion trap mass spectrometry for identification of oleanolic acid and ursolic acid in Anoectochilus roxburghii (wall.) Lindl.* *Journal of mass spectrometry : JMS* 2007, 42(7):910-917.

6. Szakiel A, Ruszkowski D, Grudniak A, Kurek A, Wolska KI, Doligalska M, Janiszowska W: *Antibacterial and antiparasitic activity of oleanolic acid and its glycosides isolated from marigold (Calendula officinalis).* *Planta medica* 2008, 74(14):1709-1715.

7. Antonisamy P, Duraipandiyan V, Ignacimuthu S: *Anti-inflammatory, analgesic and antipyretic effects of friedelin isolated from Azima tetracantha Lam. in mouse and rat models.* *The Journal of pharmacy and pharmacology* 2011, 63(8):1070-1077.

8. Steliopoulos P, Wust M, Adam KP, Mosandl A: *Biosynthesis of the sesquiterpene germacrene D in Solidago canadensis: 13C and (2)H labeling studies.* *Phytochemistry* 2002, 60(1):13-20.

9. Chappell J, Wolf F, Proulx J, Cuellar R, Saunders C: *Is the Reaction Catalyzed by 3-Hydroxy-3-Methylglutaryl Coenzyme A Reductase a Rate-Limiting Step for Isoprenoid Biosynthesis in Plants?* *Plant physiology* 1995, 109(4):1337-1343.

10. Cordoba E, Salmi M, Leon P: *Unravelling the regulatory mechanisms that modulate the MEP pathway in higher plants.* *Journal of experimental botany* 2009, 60(10):2933-2943.

11. Cunillera N, Boronat A, Ferrer A: *Spatial and temporal patterns of GUS expression directed by 5' regions of the Arabidopsis thaliana farnesyl*
diphosphate synthase genes FPS1 and FPS2. Plant molecular biology 2000, 44(6):747-758.

12. Sandeep, Misra RC, Chanotiya CS, Mukhopadhyay P, Ghosh S: Oxidosqualene cyclase and CYP716 enzymes contribute to triterpene structural diversity in the medicinal tree banaba. The New phytologist 2019, 222(1):408-424.

13. Andre CM, Legay S, Deleruelle A, Nieuwenhuizen N, Punter M, Brendolise C, Cooney JM, Lateur M, Hausman JF, Larondelle Y et al: Multifunctional oxidosqualene cyclases and cytochrome P450 involved in the biosynthesis of apple fruit triterpenic acids. The New phytologist 2016, 211(4):1279-1294.

14. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM: BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics (Oxford, England) 2015, 31(19):3210-3212.

15. Seppey M, Manni M, Zdobnov EM: BUSCO: Assessing Genome Assembly and Annotation Completeness. Methods in molecular biology (Clifton, NJ) 2019, 1962:227-245.

16. Garg R, Patel RK, Tyagi AK, Jain M: De novo assembly of chickpea transcriptome using short reads for gene discovery and marker identification. DNA research : an international journal for rapid publication of reports on genes and genomes 2011, 18(1):53-63.

17. Mendoza-Poudereux I, Kutzner E, Huber C, Segura J, Eisenreich W, Arrillaga I: Metabolic cross-talk between pathways of terpenoid backbone biosynthesis in spike lavender. Plant physiology and biochemistry : PPB 2015, 95:113-120.

18. Xue L, He Z, Bi X, Xu W, Wei T, Wu S, Hu S: Transcriptomic profiling reveals MEP pathway contributing to ginsenoside biosynthesis in Panax ginseng. BMC genomics 2019, 20(1):383.
19. Thimmappa R, Geisler K, Louveau T, O'Maille P, Osbourn A: **Triterpene biosynthesis in plants.** *Annual review of plant biology* 2014, **65:**225-257.

20. Ungaro A, Pech N, Martin JF: **Challenges and advances for transcriptome assembly in non-model species.** 2017, **12**(9):e0185020.

21. Holzer M, Marz M: **De novo transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-Seq assemblers.** *GigaScience* 2019, **8**(5).

22. Wilhelmsson PKI: **Usability of reference-free transcriptome assemblies for detection of differential expression: a case study on Aethionema arabicum dimorphic seeds.** *BMC genomics* 2019, **20**(1):95.

23. Berardini TZ, Reiser L, Li D, Mezheritsky Y, Muller R, Strait E, Huala E: **The Arabidopsis information resource: Making and mining the "gold standard" annotated reference plant genome.** *Genesis (New York, NY : 2000)* 2015, **53**(8):474-485.

24. Blande D, Halimaa P, Tervahauta AI, Aarts MG, Karenlampi SO: **De novo transcriptome assemblies of four accessions of the metal hyperaccumulator plant Noccaea caerulescens.** *Scientific data* 2017, **4:**160131.

25. Zhou X, Li J, Zhu Y, Ni S, Chen J, Feng X, Zhang Y, Li S, Zhu H, Wen Y: **De novo Assembly of the Camellia nitidissima Transcriptome Reveals Key Genes of Flower Pigment Biosynthesis.** *Front Plant Sci* 2017, **8:**1545.

26. Vilperte V, Lucaciu CR, Halbwirth H, Boehm R, Rattei T, Debener T: **Hybrid de novo transcriptome assembly of poinsettia (Euphorbia pulcherrima Willd. Ex Klotsch) bracts.** *BMC genomics* 2019, **20**(1):900.

27. Babineau M, Mahmood K, Mathiassen SK, Kudsk P, Kristensen M: **De novo transcriptome assembly analysis of weed Apera spica-venti from seven tissues and growth stages.** *BMC genomics* 2017, **18**(1):128.
28. Tsai WC, Fu CH, Hsiao YY, Huang YM, Chen LJ, Wang M, Liu ZJ, Chen HH: OrchidBase 2.0: comprehensive collection of Orchidaceae floral transcriptomes. *Plant & cell physiology* 2013, 54(2):e7.

29. Hsiao YY, ChenYW, Huang SC, Pan ZJ, Fu CH, Chen WH, Tsai WC, Chen HH: Gene discovery using next-generation pyrosequencing to develop ESTs for *Phalaenopsis* orchids. *BMC genomics* 2011, 12.

30. Wilhelmsson PKI, Chandler JO: Usability of reference-free transcriptome assemblies for detection of differential expression: a case study on *Aethionema arabicum* dimorphic seeds. *BMC genomics* 2019, 20.

31. Huang JZ, Lin CP, Cheng TC, Chang BC, Cheng SY, Chen YW, Lee CY, Chin SW, Chen FC: A de novo floral transcriptome reveals clues into Phalaenopsis orchid flower development. *PloS one* 2015, 10(5):e0123474.

32. Su CL, Chao YT, Chang YCA, Chen WC, Chen CY, Lee AY, Hwa KT, Shih MC: De Novo Assembly of Expressed Transcripts and Global Analysis of the *Phalaenopsis aphrodite* Transcriptome. *Plant and Cell Physiology* 2011, 52(9):1501-1514.

33. Sun Y, Wang G, Li Y, Jiang L, Yang Y, Guan S: De novo transcriptome sequencing and comparative analysis to discover genes related to floral development in *Cymbidium faberi Rolfe*. *SpringerPlus* 2016, 5(1):1458.

34. Wong DCJ, Amarasinghe R, Rodriguez-Delgado C, Eyles R, Pichersky E, Peakall R: Tissue-Specific Floral Transcriptome Analysis of the Sexually Deceptive Orchid Chiloglottis trapeziformis Provides Insights into the Biosynthesis and Regulation of Its Unique UV-B Dependent Floral Volatile, Chiloglottone 1. *Frontiers in Plant Science* 2017, 8.

35. Rai A, Yamazaki M, Takahashi H, Nakamura M, Kojoma M, Suzuki H, Saito K: RNA-seq Transcriptome Analysis of *Panax japonicus*, and Its Comparison with Other
Panax Species to Identify Potential Genes Involved in the Saponins Biosynthesis. *Front Plant Sci* 2016, 7:481.

36. Ling Y, Lin Z, Zha W, Lian T, You S: Rapid Detection and Characterisation of Triterpene Saponins from the Root of *Pulsatilla chinensis* (Bunge) Regel by HPLC-ESI-QTOF-MS/MS. *Phytochemical analysis: PCA* 2016, 27(3-4):174-183.

37. Zhang EF, Ling Y, Yin Z, Zhang Q: Identification and structural characterisation of triterpene saponins from the root of *Ardisia mamillata* Hance by HPLC-ESI-QTOF-MS/MS. *Natural product research* 2018, 32(8):918-923.

38. Li W, Liu W, Wei H, He Q, Chen J, Zhang B, Zhu S: Species-specific expansion and molecular evolution of the 3-hydroxy-3-methylglutaryl coenzyme A reductase (HMGR) gene family in plants. *PloS one* 2014, 9(4):e94172.

39. Kim YJ, Lee OR, Oh JY, Jang MG, Yang DC: Functional analysis of 3-hydroxy-3-methylglutaryl coenzyme a reductase encoding genes in triterpene saponin-producing ginseng. *Plant physiology* 2014, 165(1):373-387.

40. Gu W, Geng C, Xue W, Wu Q, Chao J, Xu F, Sun H, Jiang L, Han Y, Zhang S: Characterization and function of the 3-hydroxy-3-methylglutaryl-CoA reductase gene in *Alisma orientale* (Sam.) Juz. and its relationship with protostane triterpene production. *Plant physiology and biochemistry: PPB* 2015, 97:378-389.

41. Zheng X, Xu H, Ma X, Zhan R, Chen W: Triterpenoid saponin biosynthetic pathway profiling and candidate gene mining of the *Ilex asprella* root using RNA-Seq. *International journal of molecular sciences* 2014, 15(4):5970-5987.

42. Seetang-Nun Y, Sharkey TD, Suvachittanont W: Molecular cloning and characterization of two cDNAs encoding 1-deoxy-D-xylulose 5-phosphate reductoisomerase from *Hevea brasiliensis*. *Journal of plant physiology* 2008,
43. Closa M, Vranova E, Bortolotti C, Bigler L, Arro M, Ferrer A, Gruissem W: The
Arabidopsis thaliana FPP synthase isozymes have overlapping and specific
functions in isoprenoid biosynthesis, and complete loss of FPP synthase
activity causes early developmental arrest. The Plant journal: for cell and
molecular biology 2010, 63(3):512-525.

44. Richter A, Seidl-Adams I, Kollner TG, Schaff C, Tumlinson JH, Degenhardt J: A small,
differentially regulated family of farnesyl diphosphate synthases in maize
(Zea mays) provides farnesyl diphosphate for the biosynthesis of herbivore-
induced sesquiterpenes. Planta 2015, 241(6):1351-1361.

45. Zhang Y, Li ZX, Yu XD, Fan J, Pickett JA, Jones HD, Zhou JJ, Birkett MA, Caulfield J,
Napier JA et al: Molecular characterization of two isoforms of a farnesyl
pyrophosphate synthase gene in wheat and their roles in sesquiterpene
synthesis and inducible defence against aphid infestation. The New phytologist
2015, 206(3):1101-1115.

46. Olofsson L, Engstrom A, Lundgren A, Brodelius PE: Relative expression of genes of
terpene metabolism in different tissues of Artemisia annua L. BMC Plant Biol
2011, 11:45.

47. Dong L, Pollier J, Bassard JE, Ntallas G, Almeida A, Lazaridi E, Khakimov B, Arendt P,
de Oliveira LS, Lota F et al: Co-expression of squalene epoxidases with
triterpene cyclases boosts production of triterpenoids in plants and yeast.
Metabolic engineering 2018, 49:1-12.

48. Zhou J, Zhang Y, Hu T, Su P, Zhang Y, Liu Y, Huang L, Gao W: Functional
characterization of squalene epoxidase genes in the medicinal plant
Tripterygium wilfordii. International journal of biological macromolecules 2018,
49. Rasbery JM, Shan H, LeClair RJ, Norman M, Matsuda SP, Bartel B: *Arabidopsis thaliana* squalene epoxidase 1 is essential for root and seed development. *The Journal of biological chemistry* 2007, 282(23):17002-17013.

50. Han JY, Jo HJ, Kwon EK, Choi YE: Cloning and Characterization of Oxidosqualene Cyclases Involved in Taraxasterol, Taraxerol and Bauerenol Triterpene Biosynthesis in *Taraxacum coreanum*. *Plant & cell physiology* 2019, 60(7):1595-1603.

51. Herbgenomics. China: Science Press.

52. Song L, Florea L: Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads. *GigaScience* 2015, 4:48.

53. Peng Y, Leung HC, Yiu SM, Lv MJ, Zhu XG, Chin FY: IDBA-tran: a more robust de novo de Bruijn graph assembler for transcriptomes with uneven expression levels. *Bioinformatics (Oxford, England)* 2013, 29(13):i326-334.

54. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M et al: De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* 2013, 8(8):1494-1512.

55. Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J: SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics (Oxford, England)* 2009, 25(15):1966-1967.

56. Li FD, Tong W, Xia EH, Wei CL: Optimized sequencing depth and de novo assembler for deeply reconstructing the transcriptome of the tea plant, an economically important plant species. *BMC genomics* 2019, 20(1):553.

57. Gotz S, Garcia-Gomez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talon
M, Dopazo J, Conesa A: High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic acids research* 2008, **36**(10):3420-3435.

58. Aparicio G, Gotz S, Conesa A, Segrelles D, Blanquer I, Garcia JM, Hernandez V, Robles M, Talon M: Blast2GO goes grid: developing a grid-enabled prototype for functional genomics analysis. *Studies in health technology and informatics* 2006, **120**:194-204.

59. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M: KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic acids research* 2007, **35**(Web Server issue):W182-185.

60. Xia EQ, Yu YY, Xu XR, Deng GF, Guo YJ, Li HB: Ultrasound-assisted extraction of oleanolic acid and ursolic acid from *Ligustrum lucidum* Ait. *Ultrasonics sonochemistry* 2012, **19**(4):772-776.

61. Wang H, Han J, Kanagarajan S, Lundgren A, Brodelius PE: Trichome-specific expression of the amorpha-4,11-diene 12-hydroxylase (cyp71av1) gene, encoding a key enzyme of artemisinin biosynthesis in *Artemisia annua*, as reported by a promoter-GUS fusion. *Plant molecular biology* 2013, **81**(1-2):119-138.

62. Wang H, Kanagarajan S, Han J, Hao M, Yang Y, Lundgren A, Brodelius PE: Studies on the expression of linalool synthase using a promoter-beta-glucuronidase fusion in transgenic *Artemisia annua*. *Journal of plant physiology* 2014, **171**(2):85-96.

63. Ruijter JM, Ramakers C, Hoogaars WMH, Karlen Y, Bakker O, van den Hoff MJB, Moorman AFM: Amplification efficiency: linking baseline and bias in the analysis of quantitative PCR data. *Nucleic acids research* 2009, **37**(6):e45.

64. Pfaffl MW, Tichopad A, Prgomet C, Neuvians TP: Determination of stable
housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper--Excel-based tool using pair-wise correlations. Biotechnology letters 2004, 26(6):509-515.

65. Zhang G, Zhao MM, Song C, Luo AX, Bai JF, Guo SX: Characterization of reference genes for quantitative real-time PCR analysis in various tissues of Anoectochilus roxburghii. Molecular biology reports 2012, 39(5):5905-5912.

66. Livak KJ, Schmittgen TD: Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods (San Diego, Calif) 2001, 25(4):402-408.

Tables

Table 1 Data summary of the sequencing results of A. roxburghii plant

| Sample | Sample 1 | Sample 2 |
|--------|----------|----------|
| No. of raw reads | 54518572 | 61617472 |
| Raw reads nucleotides (G) | 5.08 | 5.74 |
| No. of clean read (%) | 45245266 (99) | 5109937 (82.93) |
| Nucleotides (G) | 4.01 | 4.52 |
| Average length (bp) | 95.26 | 95.00 |
| No of clean reads matched to unigene (%) | 33510861 (74.06) | 40617195 (79.49) |
| No. of unique matched clean reads (%) | 28908261 (86.27) | 35671871 (87.82) |
|                | No.    | Nucleotides (bp) | N50 length (bp) | Th
|----------------|--------|-----------------|-----------------|---|
| Contigs        | 86382  | 36450496        | 464             |   |
| Unigenes       | 68938  | 27874926        | 434             |   |

Table 2 Pearson’s correlations between expression of pupative triterpene biosynthetic genes and triterpene accumulation

|                | Contignt | ArHMG R1 | ArHM GR2 | ArDX R1 | ArDX R2 | ArHD R1 | ArHD R2 | ArFDS M | ArS C1 | ArOS C2 | ArOS C3 | ArDX S1 | ArDX S2 |
|----------------|----------|----------|----------|---------|---------|---------|---------|---------|-------|---------|---------|---------|---------|
| Content        |          |          |          |         |         |         |         |         |       |         |         |         |         |
|                |          | .877**   | .053     | .941*   | .944*   | .944*   | .861*   | .567    | .95    | .774    | .512    | -.457   | .930*   |
|                |          |          |          |         |         |         |         |         |       |         |         |         |         |
|                |          |          |          |         |         |         |         |         |       |         |         |         |         |
|                | ArHMG    | .877*    | .216     | .967*   | .983*   | .983*   | .921*   | .762*   | .97    | .940    | .513    | -.418   | .964*   |
|                | R1       |          |          |         |         |         |         |         |       |         |         |         |         |
|                |          |          |          |         |         |         |         |         |       |         |         |         |         |
|                | ArHMG    | .053     | .216     | -.001   | .112    | .112    | -.179   | .449    | .14    | .334    | -.552   | -.861*  | .057    |
|                | R2       |          |          |         |         |         |         |         |       |         |         |         |         |
|                | ArDX R1  | .941*    | .967**   | -.001   | .993*   | .993*   | .974*   | .657    | .98    | .870    | .636    | -.296   | .982*   |
|                |          |          |          |         |         |         |         |         |       |         |         |         |         |
|                | ArDX R2  | .944*    | .983**   | .112    | .993*   | 1.00    | .946*   | .723*   | .99    | .912    | .582    | -.395   | .980*   |
|                |          |          |          |         |         |         |         |         |       |         |         |         |         |
|                | ArHDR    | .944*    | .983**   | .112    | .993*   | 1.00    | .946*   | .723*   | .912   | .582    | -.395   | .980*   | .556    |
|                | 1        |          |          |         |         |         |         |         |       |         |         |         |         |
|                | ArHDR    | .861*    | .921**   | -.179   | .974*   | .946*   | .946*   | .610    | .92    | .824    | .756*   | -.080   | .945*   |
|                | 2        |          |          |         |         |         |         |         |       |         |         |         |         |
|                | ArFDS    | .567     | .762*    | .449    | .657    | .723*   | .723*   | .610    | 717    | .936    | .392    | -.483   | .666    |
|                |          |          |          |         |         |         |         |         |       |         |         |         |         |
|                | ArSM     | .955*    | .979**   | .043    | .988*   | .998*   | .998*   | .929*   | 717    | .907    | .054    | -.436   | .977*   |
|                |          |          |          |         |         |         |         |         |       |         |         |         |         |
|                | ArOSC1   | .774*    | .940**   | .334    | .870*   | .912*   | .912*   | .824*   | .936*  | .90     | .487    | -.461   | .872*   |
|                |          |          |          |         |         |         |         |         |       |         |         |         |         |
|                | ArOSC2   | .512     | .513     | -.552   | .636    | .582    | .582    | .756*   | .392   | .54     | .487    | .316    | .591    |
|                |          |          |          |         |         |         |         |         |       |         |         |         |         |

* indicates p < 0.05, ** indicates p < 0.01.
| Protein  | ArOSC3 | ArDXS1 | ArDXS2 | ArDXS3 | ArDXS4 | ArDXS5 | ArDXS6 | ArDXS7 | ArDXS8 | ArDXS9 | ArDXS10 |
|----------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
|          | -.457  | .930*  | .497   | .596   | .987*  | .981*  | .715*  | .732*  | .950*  | .950*  | .921*  |
|          | -.418  | .964** | .582   | .792*  | .903** | .893** | .950** | .663   | .912** | .917** | .921** |
|          | -.861* | .057   | .573   | .392   | .068   | .164   | .387   | .398   | .033   | .106   | .315   |
|          | -.296  | .982*  | .485   | .700*  | .955*  | .928*  | .851*  | .645   | .940*  | .940*  | .908*  |
|          | -.395  | .980*  | .556   | .755*  | .959*  | .948*  | .891*  | .696*  | .949*  | .949*  | .944*  |
|          | -.395  | .980*  | .556   | .755*  | .959*  | .948*  | .891*  | .696*  | .949*  | .949*  | .944*  |
|          | -.080  | .945*  | .378   | .664   | .884*  | .839*  | .808*  | .524   | .914*  | .914*  | .809*  |
|          | -.483  | .666   | .674*  | .930*  | .598   | .677*  | .889*  | .608   | .701*  | .691*  | .764*  |
|          | -.43   | .97    | .56    | .74    | .96    | .95    | .88    | -.70   | .96    | .96    | .95    |
|          | -.46   | .872   | .643   | .906   | .800   | .836   | .978   | .657   | .860   | .833   | .892   |
|          | .316   | .591   | .293   | .537   | .579   | .528   | .418   | .398   | .693*  | .652   | .480   |
|          | -.341  | .341   | -.714* | -.473  | -.457  | -.544  | -.452  | -.716* | -.387  | -.440  | -.627  |
|          | -.714* | -.341  | -.43   | -.46   | .505   | .942   | .922   | .858   | .941   | .928   | .908   |

**"*" means P ≤ 0.05, "**" means P ≤ 0.01.**

Additional Files

**Additional file 1:**
Genes encoding enzymes that participate pentacyclic biosynthesis of *A. roxburghii* plant.

**Additional file 2:**

Figure S1
Additional file 3:

Table S1-S2

Figures
Summary of pentacyclic triterpene biosynthesis (A) and Relative expression of related genes involved in A. roxburghii (B). A: Cytosol: a-AS: a-amyrin synthase; β-AS: β-amyrin synthase; FDS: farnesyl diphosphate synthase; CYP: CytochromeP450 monooxygenase; FS: fridelin synthase; HMGRT: 3-hydroxy-3-methyl-glutaryl coenzyme A reductase; IDI: isopentenyl diphosphate isomerase; OSC: oxidosqualene cyclases; SM: squalene monoxygenase; SOS: Sorghumol synthase; SS: squalene synthase. Plastid: DXR: 1-deoxy-Dxylulose-5-phosphate reductoisomerase DXS: 1-deoxy-D-xylulose-5-phosphate synthase; HDR: hydroxy-2-methyl-2-(E)-butenyl 4-diphosphate reductase; IDI: isopentenyl diphosphate isomerase; B: qPCR based relative expression of putative triterpene biosynthetic genes in root, stem and leaf were shown as heatmap. Primers are listed in supplementary Table S2. The relative expression ratios of transcripts of each transcript was calculated relative to the reference genes of β-actin and β-elongation factor. Their relative expression was set as 1.0.
Figure 2

Functional classification of assembled sequences of A. roxburghii based on gene ontology
(GO) categorization. A total of 68,938 unigenes were assigned into three main categories, i.e. biological process, molecular function and cellular function. Among these, biological process contains 22 categories, cellular function includes 15 categories and cellular component 18 categories. The y-axis ‘Frequency’ indicates the number of genes in a specific functional cluster. The x-axis shows a description of functional categories.

Supplementary Files
This is a list of supplementary files associated with this preprint. Click to download.
4-Genes encoding the enzymes of triterpene biosynthesis .docx
5-Figure 1s.doc
6- Table S1.docx