SCIENCE CHINA
Life Sciences

• RESEARCH PAPER •

May 2012 Vol.55 No.5: 452–466
doi: 10.1007/s11427-012-4319-6

De novo characterization of the root transcriptome of a traditional Chinese medicinal plant *Polygonum cuspidatum*

HAO DaCheng[1*], MA Pei[2], MU Jun[1], CHEN ShiLin[2*], XIAO PeiGen[2*], PENG Yong[2], HUO Li[3], XU LiJia[2] & SUN Chao[2]

[1]*Biotechnology Institute, School of Environment, Dalian Jiaotong University, Dalian 116028, China;*
[2]*Key Laboratory of Bioactive Substances and Resources Utilization of Chinese Herbal Medicine of Ministry of Education, Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences, Beijing 100193, China;*
[3]*School of Software, Dalian Jiaotong University, Dalian 116028, China*

Received January 23, 2012; accepted April 9, 2012

Various active components have been extracted from the root of *Polygonum cuspidatum*. However, the genetic basis for their activity is virtually unknown. In this study, 25600002 short reads (2.3 Gb) of *P. cuspidatum* root transcriptome were obtained via Illumina HiSeq 2000 sequencing. A total of 86418 unigenes were assembled *de novo* and annotated. Twelve, 18, 60 and 54 unigenes were respectively mapped to the mevalonic acid (MVA), methyl-D-erythritol 4-phosphate (MEP), shikimate and resveratrol biosynthesis pathways, suggesting that they are involved in the biosynthesis of pharmaceutically important anthraquinone and resveratrol. Eighteen potential UDP-glycosyltransferase unigenes were identified as the candidates most likely to be involved in the biosynthesis of glycosides of secondary metabolites. Identification of relevant genes could be important in eventually increasing the yields of the medicinally useful constituents of the *P. cuspidatum* root. From the previously published transcriptome data of 19 non-model plant taxa, 1127 shared orthologs were identified and characterized. This information will be very useful for future functional, phylogenetic and evolutionary studies of these plants.

***Polygonum cuspidatum*, root, transcriptome, HiSeq 2000 sequencing, secondary metabolism, repetitive sequence, ortholog**

Citation: Hao D C, Ma P, Mu J, *et al*. *De novo* characterization of the root transcriptome of a traditional Chinese medicinal plant *Polygonum cuspidatum*. Sci China Life Sci, 2012, 55: 452−466, doi: 10.1007/s11427-012-4319-6

Polygonum cuspidatum, also known as Huzhang, Japanese knotweed or Mexican bamboo, is a large, herbaceous perennial plant of the eudicot family Polygonaceae. *P. cuspidatum* is native to eastern Asia in China, Japan and Korea. In North America and in several countries in Europe, *P. cuspidatum* has been classified as an invasive species because of its large underground network of roots and strong growth [1]. There are many medicinal plants in the Polygonaceae family, for example, *Rheum palmatum*, *Rheum officinale*, *Rumex acetosa*, *Fagopyrum cymosum*, *Polygonum multiflorum*, *Polygonum aviculare*, *Polygonum bistorta*, and *Polygonum orientale*. The roots (including rhizomes) of these plants contain various but overlapping active components with medicinal utility. The underground part of *P. cuspidatum* has been used in traditional Chinese medicine for thousands of years as a painkiller, antipyretic, diuretic, and a remedy for cough, arthralgia, recurring bronchitis, jaundice, amenorrhea and hypertension [2]. Many secondary metabolites with therapeutic efficacy, such as resveratrol, anthraquinone and their glycosides [3], have been found to be present in large amounts in the roots. Resveratrol, a kind of polyphenolic stilbene, has been reported to

*Corresponding author (email: hao@djtu.edu.cn; xiaopg@public.bta.net.cn; slchen@implad.ac.cn)

have a variety of anti-inflammatory, anti-carcinogenic, anti-HIV [4], anti-fungal, neuroprotective and anti-platelet effects [5], and is used in the treatment of cardiovascular disease, infection, depression, stress-related and aging-related diseases. Currently, *P. cuspidatum* is the most important concentrated source of resveratrol, replacing grape byproducts. Anthraquinones, such as physcion, emodin, citreorosein, rhein, chrysophanol and anthraglycoside B, possess antioxidant, neuroprotective and antityrosinase activity [6]. Glycosides of stilbene and anthraquinone, such as polydatin, piceid (trans-resveratrol glucoside) and emodin-8-O-β-D-glucoside, have various health-promoting effects. Polydatin inhibits the activation of neurohormone, attenuates ventricular remodeling and has a lipid-lowering effect [7]. Piceid is a promising skin-lightening agent [8].

The well-characterized resveratrol pathway consists of four enzymes: phenylalanine ammonia lyase (PAL), cinnamic acid 4-hydroxylase (C4H), 4-coumarate: CoA ligase (4CL) and stilbene synthase (STS) [9]. PAL, C4H and 4CL are members of the common phenylpropanoid pathway in plants. STS is a member of the type III polyketide synthase family and is only found in species that accumulate resveratrol and related compounds. However, whether or not members of the *Polygonum* genus have these enzymes is, as yet, unknown. The MVA (mevalonic acid), MEP (2-C-methyl-D-erythritol 4-phosphate), and shikimate pathways are involved in the biosynthesis of anthraquinone [10]; however, the genes that are involved in these pathways have not been explored in *Polygonum*. There is also a lack of knowledge on the glycosylation of secondary metabolites in *Polygonum*. The absence of this kind of data for *Polygonum* hampers the development of improvements in cost-effective drug production from these plants.

The Illumina HiSeq 2000 second-generation sequencing platform uses paired-end 90 bp (PE90) sequencing and is better than the Genome Analyzer (GA) IIx platform, with 76 bp paired-end reads (PE76) [11], in sequencing throughput and data generation rate. The HiSeq 2000 platform has been used in human genome sequencing [12]. Sequencing data generated from the HiSeq and GAIIx platforms have been found to be of comparable quality but the HiSeq 2000 reads cover the genome more uniformly [12]. The HiSeq 2000 platform has also been used in the Earth Microbiome Project (www.earthmicrobiome.org) where it generated more than 250 billion base pairs of genetic information in eight days [13]. To assess microbial diversity, Zhou *et al*. [14] developed a barcoded Illumina PE sequencing method that sequenced each 16S rRNA-V6 tag sequence from both ends on the HiSeq 2000. The paired-end (PE) reads were then overlapped to obtain the V6 tag, which Zhou *et al*. reported significantly increased the sequencing accuracy to 99.65% by verifying the 3' end of each single end (SE) in which the sequencing quality was degraded [14].

In spite of its economic importance, very little molecular genetic and genomic research has been targeted at the family Polygonaceae. In recent years, RNA sequencing has revolutionized the exploration of gene expression. Logacheva *et al*. [15] performed *de novo* sequencing and characterization of the floral transcriptome in two Polygonaceae, *Fagopyrum esculentum* and *F. tataricum*, using 454 pyrosequencing technology but genes involved in secondary metabolism, as well as molecular markers and repetitive sequences, were not studied. To date, the HiSeq 2000 platform has not been used to sequence the transcriptomes of the medicinal plant. In the present study, we performed the first *de novo* sequencing and characterization of a medicinal plant transcriptome using HiSeq 2000. The genes that are potentially involved in the biosynthesis of health-promoting stilbene, anthraquinone and their glycosides were identified in the root transcriptome. A surprisingly large number of transposable elements (TEs) and simple sequence repeats (SSRs) were detected and characterized. We also identified orthologs in the transcriptomes of 19 non-model plants. This study illustrates the utility of Illumina HiSeq 2000 sequencing technology in the identification of novel genes and SSR markers in non-model organisms.

1 Materials and methods

1.1 Transcriptome sequencing

A total of 107.5 μg of RNA was extracted from the roots (including rhizomes) of three two-year-old cultivated *P. cuspidatum*. The experimental pipeline that was used is shown in Appendix Figure 1A in the electronic version. RNA integrity was confirmed using the Agilent 2100 Bioanalyzer with a minimum integrity number of eight. Beads with Oligo(dT) were used to isolate poly(A) mRNA after the total RNA was collected. Fragmentation buffer was added to break the mRNA into short fragments. Taking these short fragments as templates, random hexamer-primers were used to synthesize the first-strand cDNA. The second-strand cDNA was synthesized using GEX Second Strand buffer 10 μL, 25 mmol L^{-1} dNTPs 1.2 μL, RNaseH 1 μL and DNA polymerase I 5 μL. The short fragments of double-strand cDNA were purified with a QiaQuick PCR extraction kit and resolved with EB buffer for end repair and adding of a poly(A). Next, the short fragments were connected with sequencing adapters and purified by agarose gel electrophoresis. Suitable fragments were selected as the templates for the PCR amplification. Finally, the library was sequenced using Illumina HiSeq™ 2000.

1.2 Short reads assembly

The pipeline used for the bioinformatic analysis is shown in Appendix Figure 1B in the electronic version. Sequencing-received raw image data was transformed by base calling into sequence data which was called raw data or raw

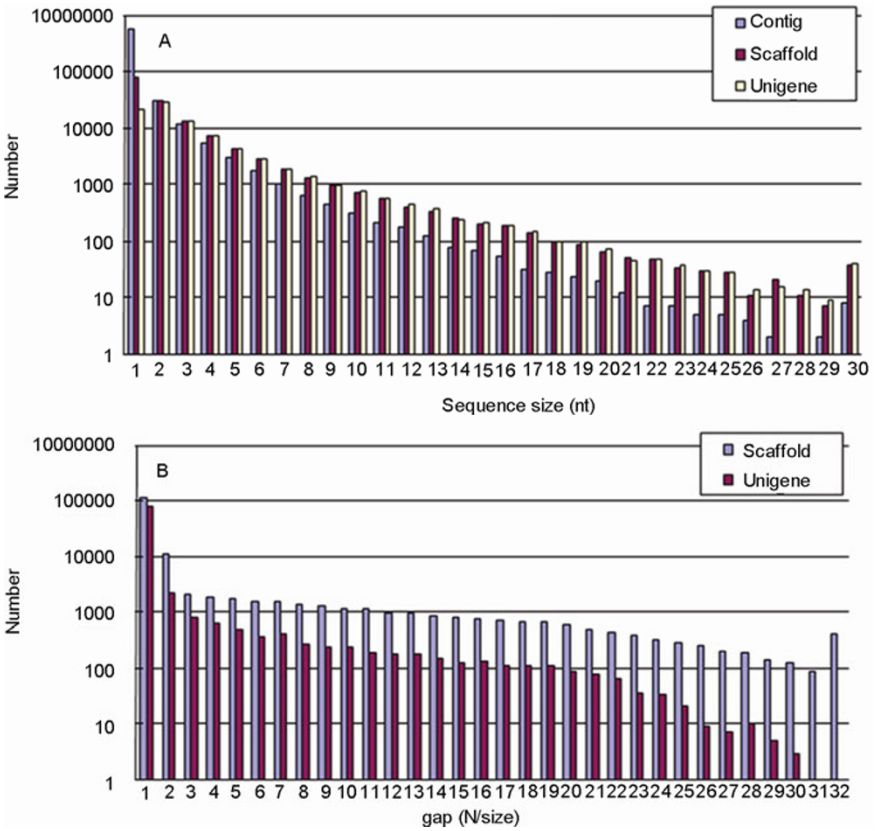

Figure 1 Statistics of Illumina short read assembly quality. A, The length distribution of the *de novo* assembly for contigs, scaffolds and unigenes is shown: 1, 200; 2, 300; 3, 400; 4, 500; 5, 600; 6, 700; 7, 800; 8, 900; 9, 1000; 10, 1100; 11, 1200; 12, 1300; 13, 1400; 14, 1500; 15, 1600; 16, 1700; 17, 1800; 18, 1900; 19, 2000; 20, 2100; 21, 2200; 22, 2300; 23, 2400; 24, 2500; 25, 2600; 26, 2700; 27, 2800; 28, 2900; 29, 3000; 30, >3000. B, The gap distribution: 1, 0; 2, 0.01; 3, 0.02; 4, 0.03; 5, 0.04; 6, 0.05; 7, 0.06; 8, 0.07; 9, 0.08; 10, 0.09; 11, 0.1; 12, 0.11; 13, 0.12; 14, 0.13; 15, 0.14; 16, 0.15; 17, 0.16; 18, 0.17; 19, 0.18; 20, 0.19; 21, 0.2; 22, 0.21; 23, 0.22; 24, 0.23; 25, 0.24; 26, 0.25; 27, 0.26; 28, 0.27; 29, 0.28; 30, 0.29; 31, 0.3; 32, >0.3.

reads and stored in FASTQ format. All the raw reads have been submitted to the NCBI Sequence Read Archive under the accession number SRA038892.1. The sequencing quality values for the bases range from 2 to 35. If E is the sequencing error rate and sQ is the sequencing quality value, then $sQ=-10\log E$.

Raw reads that only had 3' adaptor fragments were removed from the dataset before data analysis. The clean reads are those that remained after dirty raw reads were filtered out and the clean reads form the dataset on which the following analyses were based. The clean reads were mapped to the *Arabidopsis thaliana* and *Vitis vinifera* genomes and gene sequences using SOAP2 [16]. The program allows at most two mismatches in the alignments. The number of clean reads that could be mapped back to the two reference genomes and genes provided an overall assessment of the *P. cuspidatum* sequence data and gave an insight into the comparative genomics in these plants.

If the randomness of the fragmentation that was performed to break the mRNA into the short sequences was poor, then the reads would be more frequently generated from specific regions of the original transcripts and the analyses would be affected. The randomness of mRNA fragmentation was evaluated by the distribution of reads in the reference genes. The total number of reads that were aligned to the reference genes was counted and their relative positions were located. The ratio of the reads location in the reference gene to gene length and the distribution of reads in the reference genes were used to determine whether or not the randomness of fragmentation was good. If the fragmentation was random, then the distribution should be homogeneous [17].

The transcriptome *de novo* assembly was performed using short reads assembling program SOAPdenovo [16] and the contigs, scaffolds and unigenes were assembled sequentially. First, reads with certain lengths of overlap are formed into longer fragment contigs without gaps (N). Then, the reads are mapped back to the contigs; the paired-end reads can be used to detect contigs from the same transcript as well as the distances between these contigs. Next, the contigs are connected to form scaffolds using N to represent gaps of unknown length between two contigs. The paired-end reads are used to fill so that sequences with least number of Ns that cannot be extended on either end are obtained. These sequences were defined as the unigenes. The following SOAPdenovo parameters were used for the as-

sembly: –*K*=29, –*M*=2, –pair_num_cutoff=4, –*p*=27, –kmer = 29.

1.3 Unigene function annotation, GO classification, and metabolic pathway analysis

Unigene sequences were searched against NCBI's nr, Swiss-Prot, KEGG and COG protein databases using BLASTX (*E*-value<0.00001) to identify the proteins that had the highest sequence similarity with the given unigenes along and to retrieve their functional annotations. When the search results obtained from different databases conflicted a priority order of GenBank's nr, Swiss-Prot, KEGG and COG database was followed to decide the direction of the unigenes sequences. When no matches were found for the unigene sequences in these databases, ESTScan [18] was introduced to predict the coding region and to decide the sequence direction. The Gene Ontology (GO; www.geneontology.org) is a standardized gene functional classification system that offers a dynamic-updated controlled vocabulary and a strictly defined concept to comprehensively describe properties of genes and their products in any organism. GO has three ontologies: molecular function, cellular component and biological process. Every GO-term belongs to a type of ontology. For the unigenes that had matches to proteins in the nr database, we use the Blast2GO program [19] to assign GO annotations to the unigenes. Then the WEGO software [20] was used to functionally classify the GO terms to plot the distribution of the gene functions of *P. cuspidatum* at the macro level. The KEGG database (www.genome.jp/kegg) contains metabolic pathways that represent molecular interactions and reaction networks [21]. We used the KEGG annotation to assign pathway annotations to the unigenes. The COG (www.ncbi.nlm.nih.gov/COG) is a database in which orthologous gene products are classified. The unigenes were aligned to the COG database to predict and classify the possible functions of the unigenes.

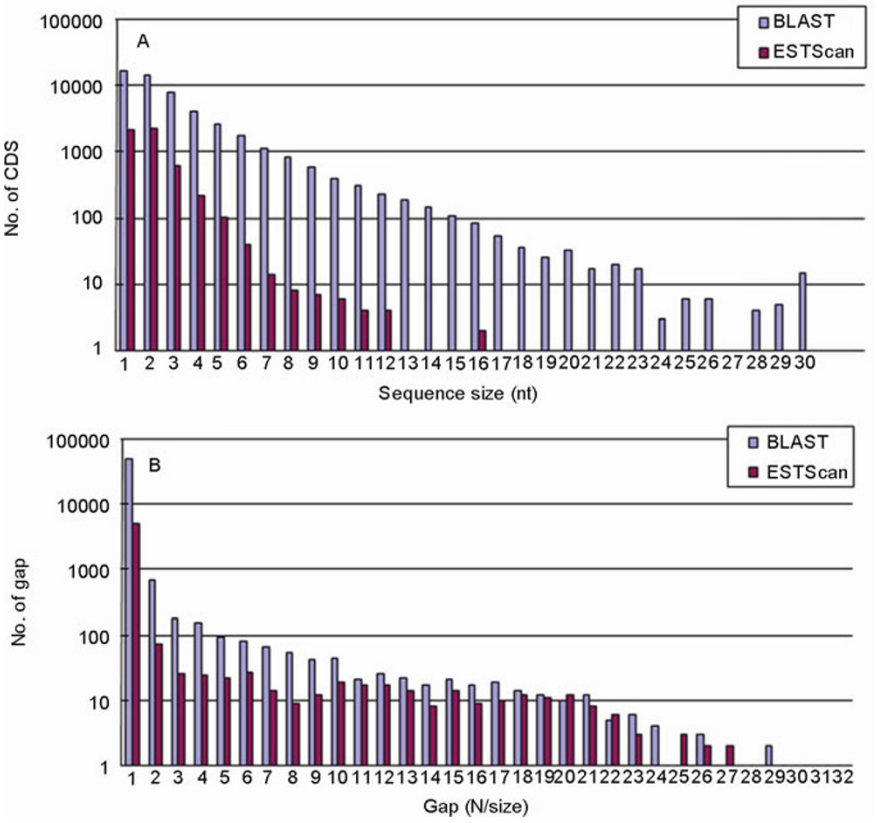

Figure 2 Prediction of protein coding sequence (CDS) for the assembled *P. cuspidatum* unigenes. Unigenes were first aligned by BLASTX (*E*-value<0.00001) to protein databases in the priority order of nr, Swiss-Prot, KEGG and COG. Unigenes aligned to the higher priority databases do not proceed to the next circle. The searches end when all circles are finished. Proteins with the highest ranks in the BLAST results are taken to decide the coding region sequences of the unigenes. The coding region sequences are translated into amino acid sequences with the standard codon table. Thus, both the nucleotide sequences (5'→3') and the amino acid sequences of the unigene coding region were acquired. Unigenes that cannot be aligned to any database were scanned by ESTScan (http://www.ch. embnet.org/software/ESTScan.html) to get the nucleotide sequence (5'→3') and the putative amino acid sequence of the coding regions. A, Length distribution of CDS predicted from BLAST results and by ESTScan. 1, 200; 2, 300; 3, 400; 4, 500; 5, 600; 6, 700; 7, 800; 8, 900; 9, 1000; 10, 1100; 11, 1200; 12, 1300; 13, 1400; 14, 1500; 15, 1600; 16, 1700; 17, 1800; 18, 1900; 19, 2000; 20, 2100; 21, 2200; 22, 2300; 23, 2400; 24, 2500; 25, 2600; 26, 2700; 27, 2800; 28, 2900; 29, 3000; 30, >3000. B, The gap (N) distribution of CDS predicted from BLAST results and by ESTScan. 1, 0; 2, 0.01; 3, 0.02; 4, 0.03; 5, 0.04; 6, 0.05; 7, 0.06; 8, 0.07; 9, 0.08; 10, 0.09; 11, 0.1; 12, 0.11; 13, 0.12; 14, 0.13; 15, 0.14; 16, 0.15; 17, 0.16; 18, 0.17; 19, 0.18; 20, 0.19; 21, 0.2; 22, 0.21; 23, 0.22; 24, 0.23; 25, 0.24; 26, 0.25; 27, 0.26; 28, 0.27; 29, 0.28; 30, 0.29; 31, 0.3; 32, >0.3.

Repeats, including retroelements, DNA transposons, simple sequence repeats (SSRs) and tandem repeats in the *P. cuspidatum* unigenes were analyzed using the protein-based RepeatMasker (http://www.repeatmasker.org). The Tandem Repeat Finder (TRF) 4.04 (http://tandem.bu.edu/trf/trf.basic.submit.html) was used with the default parameters to detect tandem repeats and to confirm the results of RepeatMasker. The six classes of SSRs, mono-, di-, tri-, tetra-, penta-, and hexa-nucleotide tandem repeats, were scanned with WebSat (http://wsmartins.net/websat/) [22]. In the unigene dataset monomers with at least 18 repeats, dimers with nine repeats, trimers with six repeats, tetramers and pentamers with four repeats, and hexamers with three repeats were found.

1.4 Ultra-performance liquid chromatography (UPLC)

Reference compounds of polydatin (111575–200502), resveratrol (111535–200502), emodin (110756–200110), physcion (110758–201013) (purity>98%) were obtained from the National Institute for the Control of Pharmaceutical and Biological Products (Beijing, China). Emodin-8-glucoside (purity>98%) was purchased from the Weikeqi Biological Technology Co., Ltd. (Sichuan, China). For each of the 38 *P. cuspidatum* samples (Appendix Figure 3 in the electronic version), the air-dried root material was ground to fine particles and sieved through an 80-mesh screen. 0.1 g of fine particles was weighed accurately and extracted with 80% methanol (10 mL) for 2 h with the assistance of ultrasonication for 30 min. The sample solution was filtered through a 0.22-μm filter before UPLC analysis. All analyses were performed on a Waters ACQUITY ultra-performance liquid chromatography (UPLC™) system equipped with a binary solvent manager, a sample manager, a column manager and a TUV detector. A Waters UPLC™ BEH C_{18} column (1.7 μm; 100 mm × 2.1 mm i.d.) was used as the solid phase. The mobile phase consisted of 0.5% acetic acid (A) and CH_3CN (B). Gradient elution was carried out with the following profile: 0–3.5 min, 8%–20%B; 3.5–6 min, 20%–40%B; 6–8 min, 40%–60%B; 8–10 min, 60%–65%B; 10–12 min, 65%–95%B; 12–13 min, balance to 95%B. The flow rate was 0.3 mL min^{-1} and the column temperature was kept at 35°C; the injection volume was 0.6 μL and scan wavelength was set at 290 nm.

1.5 Orthologous clustering

The OrthoMCL database [23] is a scalable method for constructing orthologous groups across multiple eukaryotic taxa that uses a Markov Cluster algorithm to group putative orthologs and paralogs. The OrthoMCL algorithm was applied to generate orthologous groups for the transcriptome datasets of 19 non-model plants, *P. cuspidatum* (this study), *Fagopyrum esculentum*, *F. tataricum* [15], *T. mairei* [11], Korea *T. cuspidata* [24], China *T. cuspidata* [25], *Ginkgo biloba* [26], *Huperzia serrata*, *Phlegmariurus carinatus* [27], *Pteridium aquilinum* [28], *Panax quinquefolius* [29], *Panax ginseng* [30], *Salvia miltiorrhiza* [31], *Camptotheca acuminata* [32], *Artemisia annua* [33], *Cucurbita pepo* [34], *Glycyrrhiza uralensis* [35], *Eucalyptus* hybrid [36], and *Oryza longistaminata* [37]. These plants cover a broad range of ferns, gymnosperms, monocots, and eudicots. All the putative proteins from the transcriptomes of these plants were compared (all against all) using BLASTP, and a score for each pair of proteins (u, v) with significant BLASTP hits was assigned (E=1×10^{-5}; with at least 50% of similarity). Based on the scores, orthologous groups of genes from different plant transcriptomes were identified using OrthoMCL with the default parameters. Among the identified groups, only the groups with one-to-one orthologous relationships were considered for further analyses. The functional category of each orthologous group was obtained by BLASTing the sequences against the COG database (http://www.ncbi.nih.gov/COG/) with an E-value of 1×10^{-5}. The KEGG database was used to assign pathway annotations to the orthologs.

2 Results

2.1 Transcriptome sequencing (mRNA-seq) output, assembly, and expression annotation

The Illumina HiSeq 2000 second generation sequencing generated 25600002 reads with a total of 2304000180 (2.30 Gb) nucleotides. The average read size, Q20 percentage (sequencing error rate<1%), and GC percentage were 90 bp, 91.13%, and 48.74%, respectively. These short reads were assembled into 624460 contigs with a mean length of 132 bp and a contig N50 of 118 bp (Figure 1). From these contigs, 148723 scaffolds were built using SOAPdenovo, with a mean length of 262 bp and an N50 of 318 bp (Figure 1). Because all the annotations and bioinformatic analyses in this study were based on the unigenes, the N50 sizes of the contigs and scaffolds are not very important. The results from an assembly are related to the assembly software used as well as to the sequencing depth; the more the sequencing data, the longer the assembled contigs. The 86418 scaffolds were *de novo* assembled to obtain unigenes with mean a length of 365 bp and an N50 size of 408 bp (Figure 1; the unigene sequences are available on request). Protein coding sequence (CDS) predictions were performed based on the assembled unigenes (Figure 2). The BLAST searches identified 51897 potential CDSs, 39.3% (20406) of which were more than 300 nt long and 96.9% (50273) of which had no gaps in their sequence alignments. The CDSs of the unigenes that had no hits in the BLAST searches were predicted by ESTScan (Figure 2). Of the 5394 CDSs predicted by ESTScan, 19.1% (1030) had sequence lengths of more than 300 nt, and 93.1% (5020) had no gaps in their sequence

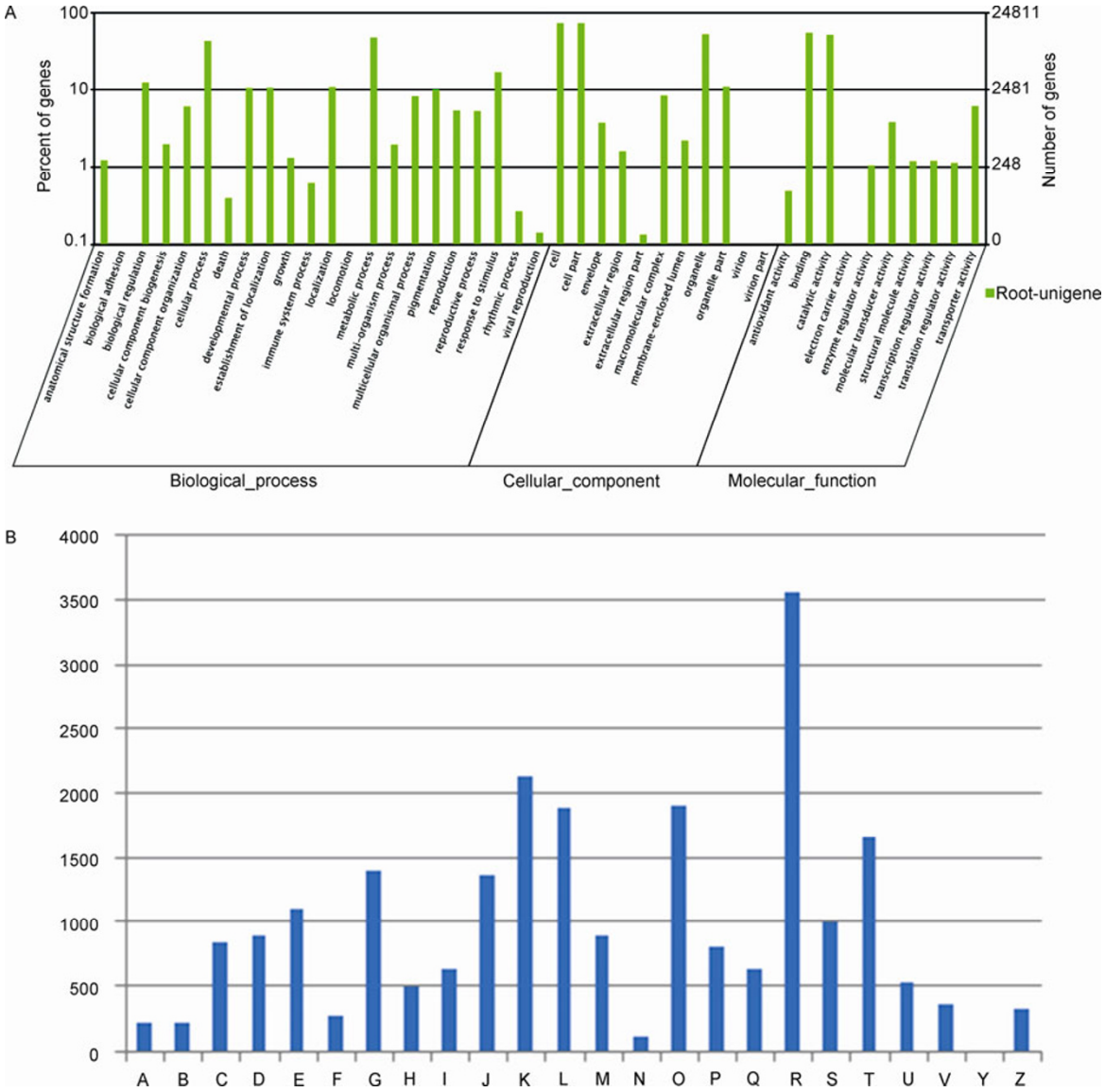

Figure 3 Gene ontology and COG classification assigned to the *P. cuspidatum* unigenes. The *x*-axis represents the categories the unigenes belong to and the *y*-axis the unigene numbers. A, Histogram presentation of the gene ontology classification. The results are summarized in the three main GO categories: biological process, cellular component and molecular function. The right *y*-axis indicates the number of genes in a category. The left *y*-axis indicates the percentage of a specific category of genes in that main category. B, Histogram presentation of clusters of orthologous groups (COG) classification. Out of 52752 nr hits, 23319 sequences have a COG classification among the 24 categories: RNA processing and modification (A); chromatin structure and dynamics (B); energy production and conversion (C); cell cycle control, cell division, chromosome partitioning (D); amino acid transport and metabolism (E); nucleotide transport and metabolism (F); carbohydrate transport and metabolism (G); coenzyme transport and metabolism (H); lipid transport and metabolism (I); translation, ribosomal structure and biogenesis (J); transcription (K); replication, recombination and repair (L); cell wall/membrane/envelope biogenesis (M); cell motility (N); posttranslational modification, protein turnover, chaperones (O); inorganic ion transport and metabolism (P); secondary metabolites biosynthesis, transport and catabolism (Q); general function prediction only (R); function unknown (S); signal transduction mechanisms (T); intracellular trafficking, secretion, and vesicular transport (U); defense mechanisms (V); nuclear structure (Y); cytoskeleton (Z).

alignments. These results suggested that the transcript assemblies were robust and that the 57291 potential CDSs and their respective 5′ and/or 3′ untranslated regions were successfully assembled. The distribution of the reads in the assembled unigenes was largely homogeneous (Appendix Figure 2A in the electronic version), suggesting good sequencing randomness. Currently, no *Polygonum* genome is publicly available. Therefore, we mapped the *P. cuspidatum* reads to the *A. thaliana* and *V. vinifera* genomes and reference genesto obtain an initial impression of similarities and differences between these three angiosperm genomes and to identify highly conserved gene sequences. Totally 609180 reads were mapped to the *A. thaliana* genome but only 6282 reads were mapped to the *V. vinifera* genome, possibly because of the relatively low quality of the *V. vinifera* sequencing and assembly. Only 24 reads were "perfect matches" to sequences in the *V. vinifera* genome; in the *A. thaliana* genome, 186076 reads were "perfect matches".

Similarly, more reads mapped to *A. thaliana* genes than to *V. vinifera* genes (288172 vs. 235798). The mapped *P. cuspidatum* reads are evenly distributed cross the *A. thaliana* reference genes (Appendix Figure 2B in the electronic version). Although more *P. cuspidatum* reads mapped to the *A. thaliana* genome compared to the *V.* genome, most *P. cuspidatum* reads could not be mapped to either *A. thaliana* or *V. vinifera* (data not shown).

Gene coverage is defined as the percentage of a reference gene that is covered by reads. This value is equal to the ratio of the number of bases in a gene covered by unique mapping reads, to the number of total bases in that gene. A total of 146 *V. vinifera* genes and 91 *A. thaliana* genes were covered by *the P. cuspidatum* reads (Appendix Tables 1 and 2 in the electronic version) and, of these, 81 *A. thaliana* genes and 140 *V. vinifera* genes were covered over less than 50% of their lengths. The frequency distribution was not significantly different between of the *A. thaliana* and *V. vinifera* genes (chi square test, P=0.075). Many of the genes that matched the *P. cuspidatum* reads were from the chloroplast, mitochondria, cell wall, nucleolus, or other organelles (Appendix Tables 1 and 2 in the electronic version), and may represent conserved/homologous genes that might have potential phylogenetic utility in the three lineages.

The RPKM method was used to eliminate the influence of different gene lengths and sequencing levels on the calculation of gene expressions [38]. The mean RPKM value of all the unigenes was 32.28; the maximal value was 1371.63 (unigene 46709). Many of the 59 unigenes that had RPKM values of more than 600 (Appendix Table 3 in the electronic version) may be involved in various physiological and metabolic processes. There were 177 unigenes in the dataset that had RPKM values of less than 0.2, implying that Illumina HiSeq 2000 could potentially detect genes with extremely low expression levels.

2.2 Functional annotation

The unigene sequences were first searched against the NCBI non-redundant (nr) database using BLASTX with a cut-off E-value of 1×10^{-5}. A total of 52752 unigenes (61.04% of all Unigenes) returned hits above the cut-off value; however, 39% of the unigenes returned no matches probably because of the lack of publicly available genomic and EST information for species in the *Polygonum* genus. Similarly, 53690 unigenes (62.1% of the total) had no matches to the protein sequences in Swiss-Prot.

For the unigenes that did find matches in SwissProt, the gene ontology (GO) annotation was used to classify the functions of the predicted *P. cuspidatum* genes. Based on the sequence homology results, 24811 unigene sequences were categorized into 43 functional groups (Figure 3A). In each of the three main GO categories (biological process, cellular component and molecular function), the "metabolic process" (11803 unigenes), "cell part" (18063 unigenes) and "binding" (13572 unigenes) terms were dominant. A high percentage of the unigenes were also annotated with the "cellular process" (10651 unigenes), "organelle" (12971 unigenes) and "catalytic activity" (12764 unigenes) terms while only a few unigenes were classified as "biological adhesion" (14 unigenes), "virion" (12 unigenes) and "electron carrier activity" (6 unigenes) (Figure 3A). To further evaluate the completeness of our transcriptome library and the effectiveness of our annotation process, we used the annotated unigene sequences to search for the genes in the COG classifications. Out of the 52752 nr hits, 23319 had a corresponding COG classification (Figure 3B). Among the 24 COG categories that were assigned to unigenes, the "general function prediction" cluster represented the largest group (3563, 15.3%) followed by "transcription" (2139, 9.2%) and "posttranslational modification, protein turnover, chaperones" (1902, 8.2%) while, nuclear structure (10, 0.04%), cell motility (107, 0.46%) and chromatin structure and dynamics (216, 0.93%) represented the smallest groups (Figure 3B). To identify the biological pathways that were active in *P. cuspidatum*, we mapped the 52752 annotated sequences to the reference canonical pathways in KEGG. A total of 22572 unigene sequences were assigned to 119 KEGG pathways. The pathways most represented were "metabolic pathways" (5632 unigenes); "biosynthesis of secondary metabolites" (2912 unigenes) and "plant-pathogen interaction" (1795 unigenes) (Appendix Table 4 in the electronic version). In the "secondary metabolism" subclass, the MVA, MEP, and shikimate pathways are involved in the biosynthesis of pharmaceutically active component anthraquinone such as emodin, rhein, and physcion [10] (Appendix Figure 3 in the electronic version); resveratrol (stilbene) biosynthesis branches from the phenylpropanoid pathway [9]. Twelve, 18, 71 and 54 unigenes mapped to these four metabolic pathways respectively. All the genes in these pathways were found in the transcriptome dataset and their expression levels are shown in Figure 4. These results imply that, in *P. cuspidatum* the roots, the genes involved in anthraquinone and resveratrol biosyntheses are actively expressed. These unigene sequences and their annotations will provide a valuable resource for investigating specific processes, functions and pathways in *P. cuspidatum* and related species, and will allow for the identification of novel genes involved in the secondary metabolite synthesis pathways.

2.3 Detection of sequences related to the glycoside biosynthetic pathway and metabolism

Glycosyltransferases (GTs) are enzymes (EC 2.4) that act as catalysts for the transfer of a monosaccharide unit from an activated nucleotide sugar (glycosyl donor) to a glycosyl acceptor molecule, usually an alcohol. Family 1 glycosyltransferases (GT1s), the UDP glycosyltransferases (UGTs), catalyze the transfer of a glycosyl moiety from UDP sugars to a wide range of acceptor molecules. UGTs play important

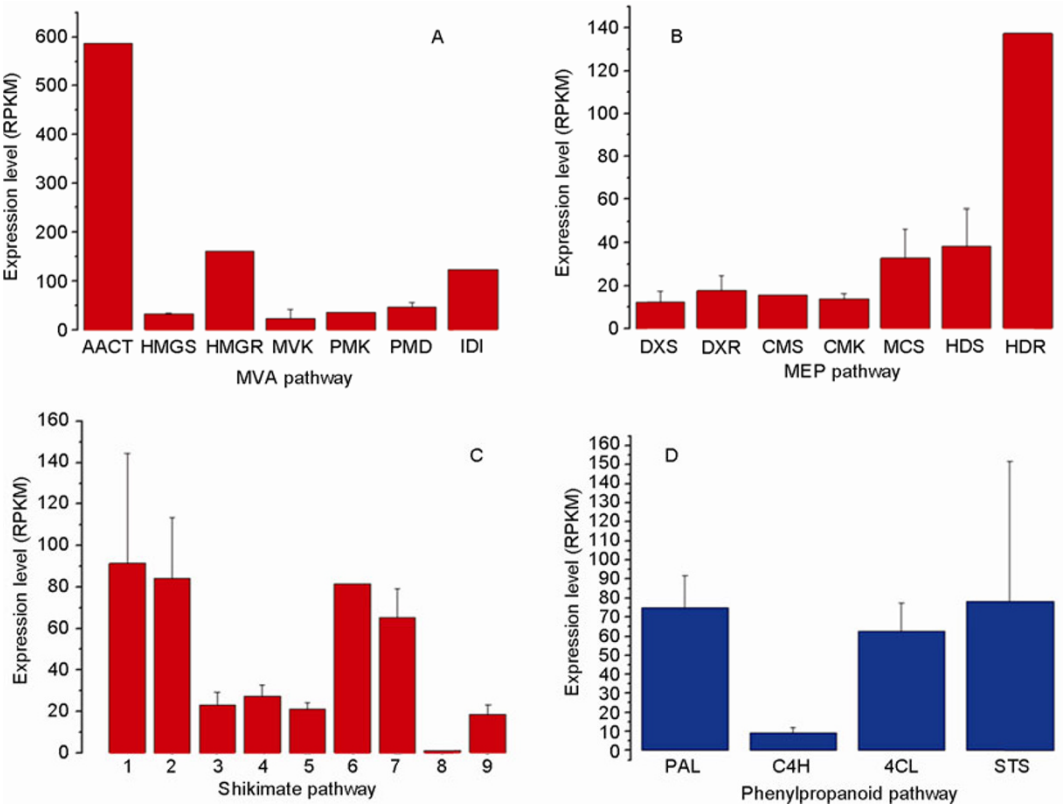

Figure 4 Expression levels (RPKM value) of the *P. cuspidatum* unigene in four KEGG metabolic pathways. A, Unigene expression levels in the MVA pathway. AACT, acetoacetyl CoA thiolase; HMGS, HMG-CoA synthase; HMGR, HMG-CoA reductase; MVK, MVA kinase; PMK, MVP kinase; PMD, MVPP decarboxylase; IDI, IPP isomerase. B, Unigene expression levels in the MEP pathway. DXS, 1-deoxyxylulose 5-phosphate synthase; DXR, DXP reductoisomerase; CMS, MEP cytidyltransferase; CMK, 4-(cytidine-5′-diphospho)-2-C-methyl-d-erythritol kinase; MCS, 2-C-methyl-d-erythritol 2,4-cyclodiphosphate synthase; HDS, 4-hydroxy- 3-methylbut 2-en-yl-diphosphate synthase; HDR, 1-hydroxy-2-methyl- butenyl 4-diphosphate reductase. C, Unigene expression levels in the shikimate pathway; 1, 3-deoxy-D-arabino-heptulosonate-7-phosphate (DAHP) synthase; 2, 3-dehydroquinate synthase; 3, shikimate 5-dehydrogenase; 4, shikimate: NADP oxidoreductase; 5, shikimate kinase; 6, EPSP synthase; 7, chorismate synthase; 8, isochorismate synthase; 9, chorismate mutase. D, Unigene expression levels in the phenylpropanoid pathway. PAL, phenylalanine ammonia lyase; C4H, cinnamic acid 4-hydroxylase; 4CL, 4-coumarate: CoA ligase; STS, stilbene synthase. Bars represent the standard error of the average.

roles in the stabilization, enhancement of water solubility and deactivation/detoxification of natural products, leading to regulation of metabolic homeostasis, detoxification of xenobiotics, and the biosynthesis, storage and transport properties of secondary metabolites [39]. In plants, UGTs are generally localized in the cytosol, and are involved in the biosynthesis of plant natural products such as flavonoids, phenylpropanoids, terpenoids and steroids, and in the regulation of plant hormones [40]. A total of 391 GT sequences were found in the *P. cuspidatum* transcriptome dataset, including 14 GT2s, 53 GT8s, 16 GT14s, one GT28, four GT37s, 15 GT47s, 18 UGTs, and 270 other GTs. Among the 18 UGTs were two UGT71s, one UGT72, three UGT73s, one UGT74, one UGT75, one UGT76, two UGT89s, two UGT90s, and five UGT95s. The expression of UGT75 (RPKM 145.47; Figure 5) was the highest, followed by UGT71 (mean RPKM 68.76) and UGT73 (mean RPKM 59.21). UGTs 79, 80, 84, 85, and 88 were not identified in the root transcriptome of *P. cuspidatum*. The main pharmaceutical components of *P. cuspidatum* root, anthraquinone and stilbene, undergo glycosylation via the cataly-

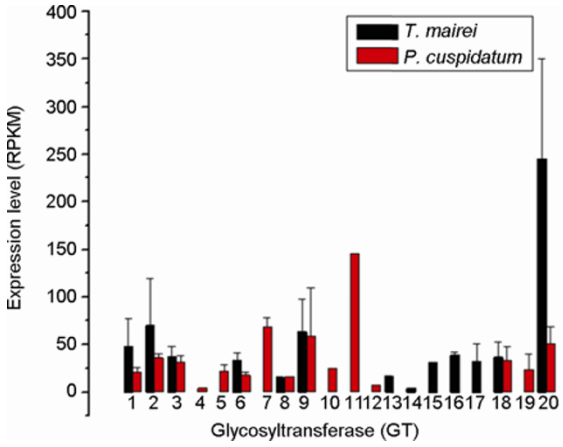

Figure 5 Unigene expression levels (RPKM value) of glycosyltransferases (including UGTs) in *P. cuspidatum*. 1, GT2; 2, GT8; 3, GT14; 4, GT28; 5, GT37; 6, GT47; 7, UGT71; 8, UGT72; 9, UGT73; 10, UGT74; 11, UGT75; 12, UGT76; 13, UGT79; 14, UGT80; 15, UGT84; 16, UGT85; 17, UGT88; 18, UGT89; 19, UGT90; 20, UGT95. Bars represent the standard error of the average. To highlight the GTs that were not expressed in the *Polygonum* root, the GTs expressed in *T. mairei* are also shown; this data is not for comparing the two species.

sis of UGTs and the resulting glycosides were found to be highly abundant in the UPLC analysis (Appendix Figure 3 in the electronic version; data not shown); however, which UGTs are responsible for these reactions is unclear. The identification of the relevant genes could have important implications in eventually increasing the yields of the pharmaceutically active glycosides. The 18 candidate UGT unigenes found that were identified in the root transcriptome may play a role in the biosynthesis of anthraquinone glycoside and stilbene glycoside and will be the subject of further study.

2.4 Assessment of transposable element (TE) and simple sequence repeat (SSR) abundance

The BLAST searches indicated that there was an abundance of TE-like sequences in our unigene dataset. Therefore, the overall status of repetitive elements in the transcriptome of *P. cuspidatum* was assessed by a combination of repeat-masking, TRF, and WebSat analyses. Totally, 900 TEs were detected in the *P. cuspidatum* unigenes. The most common TE type that was found is shown in Appendix Table 5 in the electronic version. The TEs were present at a low frequency in the *P. cuspidatum* unigenes (1.04%). Among the detected TEs, 548 transcriptionally active retroelements were found. This observed frequency for the TEs in *P. cuspidatum* is less than the frequency reported for TEs in the transcriptomes of *A. thaliana* and *Pinus contorta* [41]. In *P. contorta*, 6.2% of the raw 454 reads were estimated to represent transcriptionally active retroelements. LTR/Copia (208 unigenes) is the predominant type of retroelement identified in the *P. cuspidatum* dataset (Appendix Table 5 in the electronic version), followed by LTR/Gypsy (123 unigenes) and LINE/L1 (83 unigenes). Among the detected TEs, 323 (0.37% of unigenes) transcriptionally active DNA transposons were found. DNA/MuDR (105) was the predominant type of DNA transposon in *P. cuspidatum* (Table S5), followed by DNA/hAT-Ac (46) and DNA/hAT-Tag1 (44). We identified 1286 SSRs and 49 tandem repeats in the *P. cuspidatum* unigenes (Appendix Table 5 and Appendix Figure 4 in the electronic version). Tri-nucleotide repeats (807) were by far the most common SSRs, followed by mono-nucleotide (144), hexa-nucleotide (108), and penta-nucleotide (85) repeats. Among the tri-nucleotide repeats, 6–8 (233) and 9–11 (198) repeat units were the most common, followed by repeat unit numbers >26 (106) and 12–14 (101).

2.5 Identification of orthologous groups by OrthoMCL

We used OrthoMCL to identify gene orthologs in the unigene dataset. When the unigenes were searched against the OrthoMCL database, 50 ortho-groups from 19 plant taxa were found. These groups possessed a strict single-copy orthologous relationship (i.e., clusters contained exactly one member per species; Table 1). Among 14 COG categories the unigenes belong to, the cluster for "carbohydrate transport and metabolism" represented the largest group (8 unigenes, 21.6%) followed by "posttranslational modification, protein turnover, chaperones" (6 unigenes, 16.2%) (Figure 6(A)). When 11 taxa were used, 210 orthologs were identified (Figure 6(B)). Among 22 COGs categories the unigenes belong to, the cluster for "carbohydrate transport and metabolism" represented the largest group (27 unigenes, 16.5%) followed by "posttranslational modification, protein turnover, chaperones" (20 unigenes, 12.2%) and "signal transduction mechanisms" (18 unigenes, 11.0%). When only six taxa, *P. cuspidatum*, *F. esculentum*, *F. tataricum*, *T. mairei*, China *T. cuspidata*, and Korea *T. cuspidata*, were used, 1127 orthologs were identified. Among these orthologs, 861 of the unigenes sequences had COG classifications that covered 24 COG categories (Figure 6(C)). The cluster for "general function prediction only" represented the largest group (124 unigenes, 14.4%) followed by "posttranslational modification, protein turnover, chaperones" (102 unigenes, 11.8%) and "carbohydrate transport and metabolism" (84 unigenes, 9.8%).

3 Discussion and conclusion

In recent years, plant genomics has developed rapidly with the application of next-generation sequencing technology. However, very few studies have been carried out on the genomics of medicinal plants [42]. Such studies are imperative to build a foundation for the development of natural medicines and for the selection of cultivars with good agricultural traits, as well as to raise the study of traditional Chinese medicine to the frontiers of the modern life sciences. Using Roche 454 pyrosequencing, transcriptome datasets of nine medicinal plants, *T. cuspidata* [25], *Ginkgo biloba* [26], *Huperzia serrata*, *Phlegmariurus carinatus* [27], *Panax quinquefolius* [29], *Panax ginseng* [30], *Salvia miltiorrhiza* [31], *Camptotheca acuminata* [32], and *Glycyrrhiza uralensis* [35] have been obtained. Illumina second generation sequencing was used previously to obtain the transcriptome dataset of *T. mairei* [11]. In the present study, we have focused on the functional genomics of an important medicinal plant, *P. cuspidatum*, to promote the development of natural medicines such as resveratrol, anthraquinone and glycoside, and the selection of cultivars with good medicinal and agricultural traits. Here, the experimental high-throughput sequencing data of 19 plant taxa have been combined and insightful data mining has been performed based on the short reads and assembled unigene sequences obtained by *P. cuspidatum* sequencing.

3.1 HiSeq 2000 sequencing: PE 76 vs. PE 90

For HiSeq 2000, the PE module enables paired-end se-

Table 1 Fifty shared orthologs inferred using OrthoMCL from the transcriptome data of 19 plant taxa[a)]

OrthoMCL family ID	KO ID	*E*-value	KO definition (annotation)	*T. mairei*	Korea *T. cuspidata*	*P. cuspidatum*	*F. esculentum*	*F. tataricum*
ORTHOMCL24	K05391	0	cyclic nucleotide gated channel	Unigene20149	contig04521	Unigene126	c3596	GAA4HQR02FE6TG
ORTHOMCL28	NF	7×10^{-66}	transmembrane nine 1	Unigene15778	contig25526	Unigene20771	c1553	GAA4HQR02FYPZ9
ORTHOMCL12	K11000	0	callose synthase	Unigene10459	contig04275	Unigene1863	GAA4HQR01DG05U	GAA4HQR02GNW3H
ORTHOMCL412	K00924	0	glycogen synthase kinase 3β	Unigene35375	contig02178	Unigene83345	c1074	c480
ORTHOMCL139	K06689	5×10^{-84}	ubiquitin-conjugating enzyme E2 D/E	Unigene14692	contig10529	Unigene32577	c1267	c5123
ORTHOMCL30	K13648	0	α-1,4-galacturonosyltransferase	Unigene21195	contig02766	Unigene20662	c16561	c13031
ORTHOMCL121	NF	3×10^{-87}	Endosomal P24A protein precursor	Unigene16994	contig19414	Unigene1912	GAA4HQR01AH8L5	GAA4HQR02F48LM
ORTHOMCL40	K08959	1×10^{-49}	casein kinase 1, δ	Unigene24599	contig18040	Unigene41533	c15274	c21278
ORTHOMCL224	K01078	1×10^{-17}	acid phosphatase	Unigene3067	contig07470	Unigene5882	c12726	c2793
ORTHOMCL44	K06067	2×10^{-134}	histone deacetylase 1/2	Unigene14298	contig01593	Unigene67381	c24297	GAA4HQR02F3XTC
ORTHOMCL104	K00368	4×10^{-66}	nitrite reductase	Unigene7812	contig03544	Unigene10946	c5166	GAA4HQR02F7GDP
ORTHOMCL68	K01288	4×10^{-59}	carboxypeptidase D	Unigene20106	contig07118	Unigene82502	GAA4HQR01DIMPF	c12853
ORTHOMCL132	NF	2×10^{-175}	transporter	Unigene21131	contig18076	Unigene15064	GAA4HQR01B5PBJ	GAA4HQR02F8MIP
ORTHOMCL57	K08332	2×10^{-14}	vacuolar protein 8	Unigene10388	contig08747	Unigene13837	GAA4HQR01E3OJE	GAA4HQR02G2SH6
ORTHOMCL13	K04371	5×10^{-159}	extracellular signal-regulated kinase 1/2	Unigene17514	contig03817	Unigene10805	GAA4HQR01C7X4P	c12281
ORTHOMCL186	K03696	0	ATP-dependent Clp protease ATP-binding subunit ClpC	Unigene31895	contig22258	Unigene42453	GAA4HQR01BOAZF	GAA4HQR02IQL5E
ORTHOMCL125	K00700	0	1,4-α-glucan branching enzyme	Unigene8922	contig03147	Unigene28731	c11684	c22143
ORTHOMCL83	K05592	2×10^{-49}	ATP-dependent RNA helicase DeaD	Unigene15475	contig09075	Unigene13488	c6309	c19592
ORTHOMCL100	K00924	0	Phosphotransferase	Unigene14146	contig09733	Unigene5048	c5835	c17580
ORTHOMCL22	K08023	8×10^{-7}	latent transforming growth factor β binding protein	Unigene29197	contig15933	Unigene24373	GAA4HQR01AK2JH	GAA4HQR02GZYMK
ORTHOMCL112	K00130	0	betaine-aldehyde dehydrogenase	Unigene33104	contig02071	Unigene34444	GAA4HQR01BOFU5	c2166
ORTHOMCL409	NF	8×10^{-32}	gland development related protein 4-like	Unigene14040	contig01124	Unigene45329	c3194	GAA4HQR02IG3QG
ORTHOMCL160	K01176	3×10^{-86}	α-amylase	Unigene11457	contig08096	Unigene35202	c10464	c14161
ORTHOMCL39	K00850	0	6-phosphofructokinase	Unigene14127	contig11873	Unigene2308	GAA4HQR01CWDKA	c11212
ORTHOMCL188	K03798	0	cell division protease FtsH	Unigene35567	contig03253	Unigene19724	GAA4HQR01BGSGW	c19126
ORTHOMCL240	K09480	0	digalactosyldiacylglycerol synthase	Unigene11979	contig21538	Unigene154	c14061	c1072
ORTHOMCL75	K00036	0	glucose-6-phosphate 1-dehydrogenase	Unigene31658	contig01623	Unigene34989	c15853	GAA4HQR02H362G
ORTHOMCL94	K05692	7×10^{-103}	actin β/γ	Unigene15638	contig23816	Unigene20782	c1175	c1550
ORTHOMCL20	K13463	2×10^{-26}	coronatine-insensitive protein 1	Unigene2180	contig20966	Unigene13297	GAA4HQR01A7BRJ	c10368
ORTHOMCL219	K01895	1×10^{-105}	acetyl-CoA synthetase	Unigene17004	contig21007	Unigene2955	GAA4HQR01DBH55	GAA4HQR02JAM3J
ORTHOMCL117	K12450	8×10^{-88}	UDP-glucose 4,6-dehydratase	Unigene21835	contig09894	Unigene12369	c16067	GAA4HQR02I4OVA

(*To be continued on the next page*)

(*Continued*)

OrthoMCL family ID	KO ID	*E*-value	KO definition (annotation)	*T. mairei*	Korea *T. cuspidata*	*P. cuspidatum*	*F. esculentum*	*F. tataricum*
ORTHOMCL32	K07195	6×10^{-131}	exocyst complex component 7	Unigene20997	contig06968	Unigene12	GAA4HQR01DUW3J	GAA4HQR02FSQY0
ORTHOMCL54	K00083	3×10^{-97}	cinnamyl-alcohol dehydrogenase	Unigene27476	contig35877	Unigene29654	GAA4HQR01AM23I	GAA4HQR02HIVUG
ORTHOMCL97	K05658	1×10^{-155}	ATP-binding cassette, subfamily B (MDR/TAP), member 1	Unigene4614	contig05091	Unigene21678	c19130	GAA4HQR02FQ3XA
ORTHOMCL230	K01623	2×10^{-132}	fructose-bisphosphate aldolase, class I	Unigene22612	contig12990	Unigene56848	GAA4HQR01DOCR1	GAA4HQR02IWU70
ORTHOMCL41	K00799	8×10^{-35}	glutathione S-transferase	Unigene15715	contig19888	Unigene24945	GAA4HQR01C7R74	c12073
ORTHOMCL151	K01102	1×10^{-23}	pyruvate dehydrogenase phosphatase	Unigene18079	contig04257	Unigene15145	GAA4HQR01CQJA6	c14726
ORTHOMCL27	K00121	7×10^{-174}	S-(hydroxymethyl)glutathione dehydrogenase/alcohol dehydrogenase	Unigene13935	contig18006	contig18006	c340	GAA4HQR02IQNOI
ORTHOMCL51	K03767	4×10^{-70}	peptidyl-prolyl cis-trans isomerase A	Unigene13609	contig01156	Unigene51119	GAA4HQR01CESIS	GAA4HQR02HDE4U
ORTHOMCL55	K00517	7×10^{-79}	monooxygenase	Unigene11948	contig06251	Unigene26798	GAA4HQR01B1V4T	c7456
ORTHOMCL16	NF	5×10^{-35}	cellulose synthase 1A	Unigene13291	contig25175	Unigene16367	GAA4HQR01B17OO	GAA4HQR02FYCKG
ORTHOMCL73	NF	3×10^{-132}	Ubiquitin ligase	Unigene20905	contig15129	Unigene13107	c1472	c166
ORTHOMCL447	NF	6×10^{-78}	monodehydroascorbate reductase	Unigene23050	contig27199	Unigene53300	c4358	c9925
ORTHOMCL173	NF	7×10^{-39}	cryptochrome 2	Unigene19202	contig14271	Unigene1434	GAA4HQR01COR8O	GAA4HQR02HP058
ORTHOMCL152	NF	5×10^{-68}	Sec61 transport protein	Unigene29250	contig04792	Unigene27647	c1618	c319
ORTHOMCL46	NF	7×10^{-35}	ATP binding protein	Unigene20569	contig05601	Unigene1667	c15355	GAA4HQR02FW0YV
ORTHOMCL84	NF	6×10^{-69}	pleiotropic drug resistance like protein	Unigene33028	contig15169	Unigene76078	GAA4HQR01DCSH9	c8420
ORTHOMCL87	NF	6×10^{-48}	LIM domain-containing protein	Unigene22138	contig00607	Unigene68124	GAA4HQR01CP0LK	c15939
ORTHOMCL48	NF	3×10^{-86}	neutral invertase 5	Unigene36251	contig04818	Unigene2967	GAA4HQR01B3RQT	GAA4HQR02FTM9N
ORTHOMCL354	NF	2×10^{-101}	protein serine/threonine phosphatase	Unigene19484	contig14007	Unigene30984	c24530	c378

a) The contig sequences of Korea *T. cuspidata* can be found at http://www.nature.com/nbt/journal/v28/n11/abs/nbt.1693.html. The contig and singleton sequences of *F. esculentum* and *F. tataricum* can be found at http://www.biomedcentral.com/1471-2164/12/30. The unigene sequences of *P. cuspidatum* and *T. mairei* are available upon request. NF, not found in KEGG database. In this case, *E*-value and search result of BLASTN are shown.

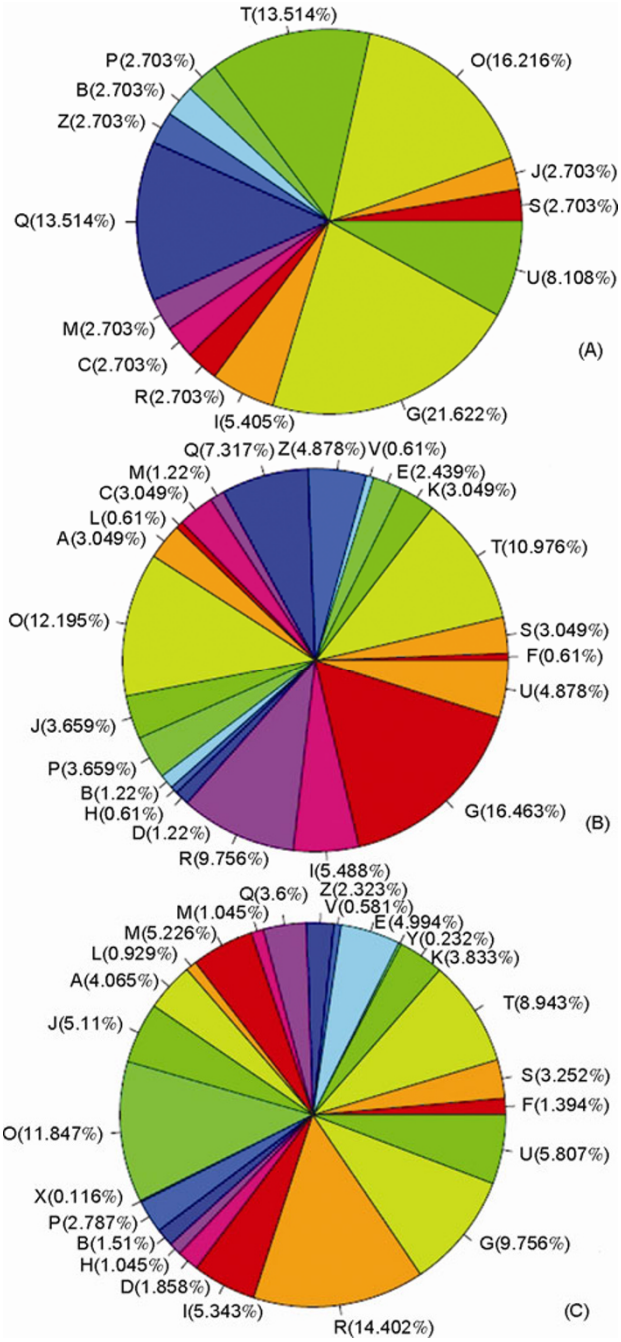

Figure 6 COG classification of orthologs in non-model plant transcriptomes. (A) COG classification in 19 plant transcriptomes. (B) COG classification in 11 plant transcriptomes. (C) COG classification in 6 plant transcriptomes. A, RNA processing and modification; B, chromatin structure and dynamics; C, energy production and conversion; D, cell cycle control, cell division, chromosome partitioning; E, amino acid transport and metabolism; F, nucleotide transport and metabolism; G, carbohydrate transport and metabolism; H, coenzyme transport and metabolism; I, lipid transport and metabolism; J, translation, ribosomal structure and biogenesis; K, transcription; L, replication, recombination and repair; M, cell wall/membrane/envelope biogenesis; N, cell motility; O, posttranslational modification, protein turnover, chaperones; P, inorganic ion transport and metabolism; Q, secondary metabolites biosynthesis, transport and catabolism; R, general function prediction only; S, function unknown; T, signal transduction mechanisms; U, intracellular trafficking, secretion, and vesicular transport; V, defense mechanisms; X, extracellular structure; Y, nuclear structure; Z, cytoskeleton.

quencing up to 2 × 100 bp for fragments ranging from 200 bp to 5 kb. PE sequencing data offers long-range positional information, empowers easy detection of structural variations such as chromosomal rearrangements, copy number variations, and indels; and simplifies the *de novo* assembly. In sequencing the *Taxus* transcriptome [11], the sequencing strategy PE 74+7+76 was used on the Illumina GAIIx platform; that is, the read lengths of the two ends of the same sequence was different, one was 73 bp, the other 75 bp. Therefore, the merged read length was 73+75=148 bp (total nt 2033154144/total reads 13737528). In the present study, the sequencing strategy PE 91+8+91 was applied; that is, the read length of the two ends of the same sequence was the same, both reads were 90 bp long and were counted as two reads. Thus, the read length is 90 bp (total nt 2304000180/total reads 25600002). The longer the sequenced read, the lower the quality value at the sequencing end. This relationship might explain the lower Q20 percentage (91.13%) for *P. cuspidatum* compared to that for *T. mairei* (96.48%). Q30 (sequencing error rate<0.1%) values for *P. cuspidatum* were 81.87% and 86.29% for two ends, well above the criteria for Q20 (>80%). In this study, PE 90 was combined with HiSeq 2000 in the *de novo* sequencing of the traditional Chinese medicinal plant *P. cuspidatum*. The high-quality of the sequencing and the high-quality of the unigene assembly that was obtained suggest that the HiSeq 2000 combined with PE90 provided a reliable platform for high-throughput transcriptome sequencing.

3.2 Genes involved in secondary metabolism

Sequence similarity searches against public databases identified 52752 unigenes that could be annotated with gene descriptions, conserved protein domains, and/or gene ontology terms. Some of the unigenes were assigned to putative metabolic pathways. Targeted searches using these annotations identified most of the genes that are associated with several primary metabolic pathways and natural product pathways. These genes, such as those that code for resverateol, anthraquinone and enzymes in the glycoside biosynthesis pathways, are important for the quality of *P. cuspidatum* as a medicinal plant. This is the first time that the novel candidate genes of these secondary pathways have been discovered in Caryophyllales including *Polygonum*. The MEP and MVA pathways play important roles in the biosynthesis of terpenoids [11,27,43], anthraquinone, ginsenosides [29], and glycyrrhizin [35]. The number of EST and/or unigene sequences that were associated with these two pathways have been recorded in previous 454-based transcriptome studies [27,29,35,43], but the expression levels of the unigenes is unknown. In our *P. cuspidatum* dataset, the expression levels of the unigenes are represented as RPKM values based on the data generated by the Illumina platform; thus, the expression levels of the genes encoding the different enzyme in the same pathway could be

compared quantitatively. In the MVA pathway, the first enzyme AACT had the highest RPKM value (Figure 4A). In contrast, the most highly expressed enzyme in the MEP pathway was the terminal enzyme HDR (1-hydroxy-2-methyl-butenyl 4-diphosphate reductase), which catalyzes the formation of isopentenyl diphosphate (IPP) and its isomer, dimethylallyl diphosphate (DMAPP). Anthraquinone production is increased by the overexpression of 1-deoxy-D: -xylulose-5-phosphate synthase (DXS) in *Morinda citrifolia* cells [44]. In *Polygonum*, DXS had the lowest expression (RPKM 12.2) compared to the genes coding the downstream enzyme (Figure 4B). Whether or not the overexpression of DXS could enhance anthraquinone production in *P. cuspidatum* is worthy of further study. In the shikimate pathway, DAHP (3-deoxy-d-arabino-heptu- losonate 7-phosphate) synthase catalyses the first step and has the highest mRNA levels (Figure 4C), while the expression level of the downstream isochorismate synthase, which channels chorismate to the production of anthraquinones and which is involved in the regulation of anthraquinone biosynthesis [45], had the lowest (RPKM 1.01). In the future, it will be interesting to investigate whether or not the expressions of the genes encoding these pathway enzymes are up-regulated upon elicitation.

KEGG pathway analysis suggested that transcripts of the phenylpropanoid biosynthesis pathway and stilbenoid, diarylheptanoid and gingerol biosynthesis pathways were highly abundant in *P. cuspidatum* (Appendix Table 5 in the electronic version). Besides their medicinal utility, phenylpropanoids such as resveratrol can function as inducible antimicrobial compounds, and can act as signal molecules in plant-microbe interactions [46]. PAL (phenylalanine ammonia lyase), C4H (cinnamic acid 4-hydroxylase) and 4CL (4-coumarate:CoA ligase) are involved in the resveratrol, flavonoid and coumarin biosynthesis pathways [9,27,45,47], but the transcript abundance of these genes is different in *P. cuspidatum*, *Huperzia serrata*, *Phlegmariurus carinatus*, *Camellia sinensis* and *Artemisia tridentate*. Furthermore, the expression levels of STS (stilbene synthase), the final enzyme in the resveratrol biosynthesis pathway, were more than eight times higher than those of the upstream C4H (cinnamic acid 4-hydroxylase, a cytochrome P450) in *P. cuspidatum* (Figure 4D). Surprisingly, more unigenes representing *PAL* (phenylalanine ammonia lyase, 26 unigenes) *and 4CL* (4-coumarate: CoA ligase, 24 unigenes) were found in the root transcriptome data. These results reveal the complexity in the regulation of transcription and in the control of metabolic flux, as well as the complexity of species-specific and tissue-specific transcriptomes. Transcriptional studies of the genes encoding the enzymes of biosynthetic pathways such as PAL, C4H, 4CL and STS could partially explain the different levels of resveratrol in the different plants. The results have provided hints for selecting markers for the development of cultivars with high phenolic content, which can be of value to the drug industry.

UGTs catalyze the transfer of sugars to various acceptor molecules including flavonoids, phytohormone, lignin, sterol [39], ginsenoside backbone [29], and aglycone glycyrrhetic acid [35]. In *P. cuspidatum*, it was supposed that stilbene, anthraquinone and torachryson are the acceptor molecules of UGTs, because many glycosides, including resveratroloside, polydatin, piceid, emodin-8-O-glucoside, emodin-1-O-glucoside, physcion-8-O-glucoside, physcion-8-O-(6'-acetyl) glucoside, torachryson-8-O-glucoside, and torachryson-8-O-(6'-acetyl) glucoside, have been detected (Figure S3 and data not shown) [3]. It has been suggested that UGT74/75/84 that belong to orthologous group 14 of GT1s are involved in auxin, anthranilate, anthocyanin and phenylpropanoids metabolism [39]. One UGT74 (unigene 83067) and one UGT75 (unigene 83611) were found in our transcriptome dataset. UGT84 was not found in the *P. cuspidatum* transcriptome but was present in the previously reported *T. mairei* transcriptome (unigene 14589) [11]. Considering the complete coverage of our transcriptome dataset, we have proposed that the *P. cuspidatum* root might not express the UGT84, UGT79, UGT80, UGT85 and UGT88 genes that are present in the *T. mairei* transcriptome (Figure 5). As mentioned, resveratrol biosynthesis branches from the phenylpropanoid pathway. It is therefore worth investigating whether or not unigene 83067 and/or unigene 83611 might encode a bona fide UGT that is responsible for the glycosylation of stilbene (resveratrol). Because the shikimate/MEP/MVA pathway and the phenylpropanoid pathway are closely linked, it may be possible to discover the UGT that is responsible for the glycosylation of other putative acceptor molecules based on a screen study of the 18 UGT unigenes that were identified in the *P. cuspidatum* transcriptome dataset.

3.3 Transcriptome-based ortholog identification

With the progress of next generation sequencing efforts, comparative genomic approaches have increasingly been employed to facilitate both evolutionary and functional analyses. Conserved sequences can be used to infer evolutionary history, while homology implies conserved biochemical and physiological functions, which could be used to facilitate genome annotation. A number of low copy nuclear genes have been previously identified in flowering plants, including the phytochromes, ADH, TPI, GAP3DH, LEAFY, ACCase, PGK, petD, GBSSI, GPAT, ncpGS, GIGANTEA, GPA1, AGB1, PPR and RBP2, that have primarily been used as phylogenetic markers [48]. However, none of these genes were in our identified orthologous groups, implying that they may not be single-copy or low-copy genes. Therefore, these genes might not be ubiquitous in plants and might not be useful for the phylogenetic and evolutionary study of a broad range of plant taxa. On the other hand, current molecular systematics in flowering

plants has been dominated by the use of phylogenetic markers derived from the plastid genome (for example, rbcL, matK, psbA-trnH, trnL-F) or ribosomal DNA (18S, ITS) [49]. However, the chloroplast rbcL and matK genes were found to undergo positive selection in many plant lineages [50], while ITS pseudogenes re often troublesome because of incorrect assumptions of orthology in the phylogenetic reconstructions [51]. It is slowly being realized that the alternative target regions in the nuclear genome (low-copy nuclear genes, LCNGs) are burdened with similar problems, and developing useful LCNGs for non-model organisms requires investments of time and effort that hinder its use as a real practical alternative [51].Our study, for the first time, has provided a wealth of shared single copy nuclear genes based on 19 transcriptome datasets obtained from Illumina- and 454-based high-throughput sequencing. Using the gene clustering algorithm Tribe-MCL, Duarte *et al*. [48] identified 959 shared single-copy genes in the genomes of the model plants *A. thaliana*, *Populus trichocarpa*, *V. vinifera* and *Oryza sativa*. However, how many of these genes are shared by other lineages such as fern, moss, and gymnosperm is, as yet, unknown. In contrast, the single-copy nuclear genes identified by OrthoMCL are promising markers for phylogenetics (Appendix Figure 5 in the electronic version), and might contain more phylogenetically-informative sites than the commonly used markers from the chloroplast or mitochondria genomes.

In conclusion, the sequences generated in this work represent the largest collection of *Polygonum* sequences deposited in public databases. Novel genes involved in the biosynthesis of pharmaceutically active components, transcriptionally active TEs and SSRs were identified for the first time in Caryophyllales including *Polygonum*. The clustering of orthologous genes has provided the first framework for integrating information from multiple transcriptomes, highlighting the divergence and conservation of gene families and biological processes. These results can be used to produce genetically improved varieties of *Polygonum* with increased secondary metabolite yields, different compound compositions and better medicinal and agronomic characteristics.

This work was supported by the National Science and Technology Major Program (Grant No. 2008ZX10005-004).

1 Grimsby J L, Kesseli R. Genetic composition of invasive Japanese knotweed s. l. in the United States. Biol Invasions, 2010, 12: 1943–1946
2 Yan S, Li L, Yu S, *et al*. Effect of Tabellae Polygoni Cuspidati on Blood Lipids and Rheologiacl Property in Rats. China J Chin Mat Med (Zhongguo Zhongyao Zazhi), 1993, 18: 617–619
3 Dong J, Wang H, Wan L, *et al*. Identification and determination of major constituents in *Polygonum cuspidatum Sieb. et Zucc*. by high performance liquid chromatography/electrospray ionization-ion trap-time-of-flight mass spectrometry. Se Pu, 2009, 27: 425–430
4 James J S. Resveratrol: why it matters in HIV. AIDS Treat News, 2006, 3–5
5 Vang O, Ahmad N, Baile C A, *et al*. What is new for an old molecule? Systematic review and recommendations on the use of resveratrol. PLoS ONE, 2011, 6: e19881
6 Leu Y L, Hwang T L, Hu J W, *et al*. Anthraquinones from *Polygonum cuspidatum* as tyrosinase inhibitors for dermal use. Phytother Res, 2008, 22: 552–556
7 Gao J P, Chen C X, Gu W L, *et al*. Effects of polydatin on attenuating ventricular remodeling in isoproterenol-induced mouse and pressure-overload rat models. Fitoterapia, 2010, 81: 953–960
8 Jeong E T, Jin M H, Kim M S, *et al*. Inhibition of melanogenesis by piceid isolated from *Polygonum cuspidatum*. Arch Pharm Res, 2010, 33: 1331–1338
9 Halls C, Yu O. Potential for metabolic engineering of resveratrol biosynthesis. Trends Biotechnol, 2008, 26: 77–81
10 Perassolo M, Quevedo C V, Busto V D, *et al*. Role of reactive oxygen species and proline cycle in anthraquinone accumulation in *Rubia tinctorum* cell suspension cultures subjected to methyl jasmonate elicitation. Plant Physiol Biochem, 2011, 49: 758–763
11 Hao D C, Ge G, Xiao P G, *et al*. The first insight into the tissue specific *Taxus* transcriptome via Illumina second generation sequencing. PLoS ONE, 2011, 6: e21220
12 Ajay S S, Parker S C, Ozel Abaan H, *et al*. Accurate and comprehensive sequencing of personal genomes. Genome Res, 2011, 21: 1498–1505
13 Gilbert J A, Meyer F, Antonopoulos D, *et al*. Meeting report: the terabase metagenomics workshop and the vision of an Earth microbiome project. Stand Genomic Sci, 2010, 3: 243–248
14 Zhou H W, Li D F, Tam N F, *et al*. BIPES, a cost-effective high-throughput method for assessing microbial diversity. ISME J, 2011, 5: 741–749
15 Logacheva M D, Kasianov A S, Vinogradov D V, *et al*. *De novo* sequencing and characterization of floral transcriptome in two species of buckwheat (*Fagopyrum*). BMC Genomics, 2011, 12: 30
16 Li R, Zhu H, Ruan J, *et al*. *De novo* assembly of human genomes with massively parallel short read sequencing. Genome Res, 2010, 20: 265–272
17 Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet, 2009, 10: 57–63
18 Iseli C, Jongeneel C V, Bucher P. ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. Proc Int Conf Intell Syst Mol Biol, 1999, 138–148
19 Conesa A, Götz S, García-Gómez J M, *et al*. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics, 2005, 21: 3674–3676
20 Ye J, Fang L, Zheng H, *et al*. WEGO: a web tool for plotting GO annotations. Nucleic Acids Res, 2006, 34: W293–297
21 Kanehisa M, Araki M, Goto S, *et al*. KEGG for linking genomes to life and the environment. Nucleic Acids Res, 2008, 36: D480–484
22 Martins W S, Lucas D C, Neves K F, *et al*. WebSat—a web software for microsatellite marker development. Bioinformation, 2009, 3: 282–283
23 Chen F, Mackey A J, Stoeckert C J Jr, *et al*. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. Nucleic Acids Res, 2006, 34: D363–368
24 Lee E K, Jin Y W, Park J H, *et al*. Cultured cambial meristematic cells as a source of plant natural products. Nat Biotechnol, 2010, 28: 1213–1217
25 Wu Q, Sun C, Luo H, *et al*. Transcriptome analysis of *Taxus cuspidata* needles based on 454 pyrosequencing. Planta Med, 2011, 77: 394–400
26 Lin X, Zhang J, Li Y, *et al*. Functional genomics of a living fossil tree *Ginkgo* based on next generation sequencing technology. Physiol Plant, 2011, 143: 207–218
27 Luo H, Li Y, Sun C, *et al*. Comparison of 454-ESTs from *Huperzia serrata* and *Phlegmariurus carinatus* reveals putative genes involved in lycopodium alkaloid biosynthesis and developmental regulation. BMC Plant Biol, 2010, 10: 209
28 Der J P, Barker M S, Wickett N J, *et al*. *De novo* characterization of

the gametophyte transcriptome in bracken fern, *Pteridium aquilinum*. BMC Genomics, 2011, 12: 99
29 Sun C, Li Y, Wu Q, *et al*. *De novo* sequencing and analysis of the American ginseng root transcriptome using a GS FLX Titanium platform to discover putative genes involved in ginsenoside biosynthesis. BMC Genomics, 2010, 11: 262
30 Chen S L, Luo H, Li Y, *et al*. 454 EST analysis detects genes putatively involved in ginsenoside biosynthesis in *Panax ginseng*. Plant Cell Rep, 2011, 30: 1593–1601
31 Li Y, Sun C, Luo H M, *et al*. Transcriptome characterization for *Salvia miltiorrhiza* using 454 GS FLX. Yao Xue Xue Bao, 2010, 45: 524–529
32 Sun Y, Luo H, Li Y, *et al*. Pyrosequencing of the *Camptotheca acuminata* transcriptome reveals putative genes involved in camptothecin biosynthesis and transport. BMC Genomics, 2011, 12:533
33 Wang W, Wang Y, Zhang Q, *et al*. Global characterization of *Artemisia annua* glandular trichome transcriptome using 454 pyrosequencing. BMC Genomics, 2009, 10: 465
34 Blanca J, Cañizares J, Roig C, *et al*. Transcriptome characterization and high throughput SSRs and SNPs discovery in *Cucurbita pepo* (Cucurbitaceae). BMC Genomics, 2011, 12: 104
35 Li Y, Luo H M, Sun C, *et al*. EST analysis reveals putative genes involved in glycyrrhizin biosynthesis. BMC Genomics, 2010, 11: 268
36 Mizrachi E, Hefer C A, Ranik M, *et al*. *De novo* assembled expressed gene catalog of a fast-growing Eucalyptus tree produced by Illumina mRNA-Seq. BMC Genomics, 2010, 11: 681
37 Yang H, Hu L, Hurek T, *et al*. Global characterization of the root transcriptome of a wild species of rice, *Oryza longistaminata*, by deep sequencing. BMC Genomics, 2010, 11: 705
38 Mortazavi A, Williams B A, McCue K, *et al*. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods, 2008, 5: 621–628
39 Yonekura-Sakakibara K, Hanada K. An evolutionary view of functional diversity in family 1 glycosyltransferases. Plant J, 2011, 66: 182–193
40 Bowles D, Lim E K, Poppenberger B, *et al*. Glycosyltransferases of lipophilic small molecules. Annu Rev Plant Biol, 2006, 57: 567–597
41 Parchman T L, Geist K S, Grahnen J A, *et al*. Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery. BMC Genomics, 2010, 11: 180
42 Chen S L, Xiang L, Guo X, *et al*. An introduction to the medicinal plant genome project. Front Med, 2011, 5: 178–184
43 Bajgain P, Richardson B A, Price J C, *et al*. Transcriptome characterization and polymorphism detection between subspecies of big sagebrush (*Artemisia tridentata*). BMC Genomics, 2011, 12: 370
44 Quevedo C, Perassolo M, Alechine E, *et al*. Increasing anthraquinone production by overexpression of 1-deoxy-D: -xylulose-5-phosphate synthase in transgenic cell suspension cultures of *Morinda citrifolia*. Biotechnol Lett, 2010, 32: 997–1003
45 Stalman M, Koskamp A M, Luderer R, *et al*. Regulation of anthraquinone biosynthesis in cell cultures of *Morinda citrifolia*. J Plant Physiol, 2003, 160: 607–614
46 Naoumkina M A, Zhao Q, Gallego-Giraldo L, *et al*. Genome-wide analysis of phenylpropanoid defence pathways. Mol Plant Pathol, 2010, 11: 829–846
47 Shi C Y, Yang H, Wei C L, *et al*. Deep sequencing of the *Camellia sinensis* transcriptome revealed candidate genes for major metabolic pathways of tea-specific compounds. BMC Genomics, 2011, 12: 131
48 Duarte J M, Wall P K, Edger P P, *et al*. Identification of shared single copy nuclear genes in *Arabidopsis*, *Populus*, *Vitis* and *Oryza* and their phylogenetic utility across various taxonomic levels. BMC Evol Biol, 2010, 10: 61
49 Hao D C, Xiao P G, Huang B, *et al*. Interspecific relationships and origins of Taxaceae and Cephalotaxaceae revealed by partitioned Bayesian analyses of chloroplast and nuclear DNA sequences. Plant Syst Evol, 2008, 276: 89–104
50 Hao D C, Chen S L, Xiao P G. Molecular evolution and positive Darwinian selection of the chloroplast maturase matK. J Plant Res, 2010, 123: 241–247
51 Nieto Feliner G, Rosselló J A. Better the devil you know? Guidelines for insightful utilization of nrDNA ITS in species-level evolutionary studies in plants. Mol Phylogenet Evol, 2007, 44: 911–919