Plant Omics Journal POJ

POJ 8(2):89-95 (2015) ISSN:1836-3644

Transcriptomics analysis of Chinese hawthorn (*Crataegus pinnatifida*) provides insights into the biosynthesis of polyphenolic compounds

Mingxia Yang[1,2], Zhigang Dong[2], Qiufen Cao[2], Mingchang Chen[1,3*]

[1]The Institute of Loess Plateau, Shanxi University, Taiyuan, People's Republic of China
[2]Pomology Institute of Shanxi Academy of Agricultural Sciences, Taigu, People's Republic of China
[3]Department of Agriculture Shanxi Province, Taiyuan, People's Republic of China

***Corresponding author: mcchensx@sohu.com**

Abstract

Due to its high polyphenolic compound content, Chinese hawthorn, which belongs to the Rosaceae family, is a popular fruit consumed as food and medicine source in China. However, little genetic studies have been performed on hawthorn, partially because of the lack of genomic resources. To this end, we performed *de novo* transcriptomics analysis of hawthorn using the cost-effective Illumina paired-end RNA sequencing technology. Briefly explain how study was conducted and what measured then. 6.5 Gbp data were obtained and assembled into 83,817 transcripts. Approximately 98% of the transcripts had homologous in the available database, and 97% had homologous hits in the apple, pear and peach genomes. A total of 208 candidate key enzymes potentially involved in polyphenolic compound biosynthetic pathways were identified and compared to their homologous in related species. Approximately 77.12% of transcripts (64,636) were predicted to contain CDS (coding sequences). Using gene family comparative analysis, 14,800 gene families and 20,943 specific genes of hawthorn were identified. These specific genes were enriched in 3 secondary metabolite pathways (stilbenoid, diarylheptanoid and gingerol biosynthesis, phenylpropanoid biosynthesis and plavone and flavonol biosynthesis) that are related to phenolic compound biosynthesis and other basic metabolic processes. In addition, 10,472 simple sequence repeats (SSRs) were detected.

Keywords: *De novo*; Hawthorn / *Crataegus pinnatifida*; polyphenolic compounds; gene family comparative.
Abbreviations: SSRs_simple sequence repeats; PE_pair-end; CDS_coding sequence; KEGG_Kyoto Encyclopedia of Genes and Genomes; EST_expressed sequence tags.

Introduction

Hawthorn (Crataegus) is a large genus of shrubs in the Rosaceae family and is native to temperate regions of the Northern Hemisphere in Europe, Asia and North America (Rigelsky and Sweet, 2002). The Chinese hawthorn is commonly called "Shanlihong" in China, of which *Crateagus pinnatifida* Bge. and *C. pinnatifida* Bge. var. *major* N.E.Br. are the most important varieties, due to their large and delicious fruits with a characteristic acidic taste (Zhao and Tian, 1996; Liu et al., 2010). The fruit has been consumed for 2,500 years in China and may be eaten raw or processed into sauce, juice, wine and other sweet food (Dai et al., 2007). Chinese hawthorn is commonly considered to comprise 18 species, but only fruits of *C. pinnatifida* and *C. pinnatifida* var. *major* are included in the Chinese Pharmacopoeia (Yang et al., 2012). Hawthorn is considered one of the oldest pharmaceutical plants in China, and the fruit has primarily been used to improve digestion or decrease food stasis, to lower blood cholesterol and to reduce the risk of cardiovascular diseases (Rigelsky and Sweet, 2002; Dai et al., 2007; Kao et al., 2005; Chang et al., 2005; Pittler et al., 2003), making it among the most significant of lesser known fruit species(Gazdik et al., 2008; Jurikova et al., 2012; Rop et al., 2012). It is also well known that all these beneficial health-promoting activities are connected with polyphenolic compounds and triterpene acids (Yang et al., 2012). Chinese hawthorn (*Crataegus pinnatifida* Bge.) fruit has been confirmed to have high polyphenolic content (Cui et al., 2006), primarily including phenolic acids, flavonoids and procyanidins. The quantification of polyphenolic compounds have been well-studied, and more than 50 flavonoids have been isolated from *Crateagus spp.*(Yang et al., 2012; Jurikova et al., 2012). However, there have been few reports on the genetic structure of polyphenolic acid biosynthesis and its regulation in hawthorn. Although transcript assembly and quantification by RNA-seq of differentially expressed genes between soft-endocarp and hard-endocarp hawthorns has been performed, and 52,673 high-quality ESTs were generated (Dai et al., 2013), a comprehensive description of its transcriptome remains unavailable. In this study, we used *de novo* mRNA-seq analysis and obtained 83,817 transcripts. Numerous candidate genes involved in the biosynthesis of polyphenolic compounds were identified. The resulting annotated sequences extend the genomic resources available and may provide a basic resource for future studies of the molecular mechanisms of polyphenolic acid biosynthesis and regulation in hawthorn.

Results and Discussion

RNA-seq data filtering and assembly

Using Illumina sequencing, each fragment yielded a pair-end (PE) read with 100 bp at each end. After data filtering, we generated 32.3 million clean reads, a total of 6.5 billion bases with more than 98% Q20 bases (base calling quality more than 20 and an error rate of less than 0.01), and these data were used for de novo assembly (Table 1) by Trinity (Grabherr et al., 2011). A total of 83,817 transcripts of more than 200 bp in length were obtained, including 11,551 clusters (containing a total of 39,952 transcripts) and 43,865 singletons. Within a cluster, the similarity region between transcripts was greater than 70%, and most of these transcripts were derived from alternative splice and few transcripts from paralogous genes. The total length and N50 length were 78,620,139 bp and 1,615 bp, respectively (Table 1). A substantial number of large transcripts were produced; 27,720 of transcripts (33.1%) were larger than 1000 bp in length. the average sequencing depth is approximate 83X and the coverage of transcripts exhibited a positive relationship with the length of the given sequences (Figure 1 and Figure 2).

Annotation and CDS prediction

For annotation, a homology search of the transcripts was conducted in Nr, Swiss-Prot, KEGG, COG and Nt using BLAST with an E-value threshold of 1×10^{-5}. A total of 61,985 (73.95%) transcripts were annotated to at least one of the five databases (Table 2 and Table S1). Among them, 57,038 transcripts could be annotated to the Nr database, in which 70.73% of transcripts showed a best hit belonging to peach (*Amygdalus persica*). To identify the homologous genes in three related plant species, the transcripts were realigned to the apple, pear and peach genomes and their whole proteome. A total of 97% transcripts (81,301) had at least one homolog in the three species. Apple has the most homologous genes of hawthorn among the three, and 96% of transcripts (80,643) were realigned to the genome (77,627 using blat and 60,857 using blastx). Pear has the second most gene homologs to hawthorn: 77,099 transcripts (92%) were realigned to the pear genome (71,471 using blat and 58,094 using blastx). And 69% of transcripts (58,242) could be realigned to the peach genome (22,793 using blastn and 55,792 using blastx) (Table 2). These data suggested that hawthorn is more closely related to apple than to pear and peach at the genetic level. The overlap ratio was analyzed against the hawthorn (Crataegus pinnatifida accessions H8 and S7) ESTs database using blastn. A total of 18,316 and 5,845 unique transcripts were found in our data and previously published data, respectively. This result may be due to the different sample source. Our sample contained several tissues, and the unique transcripts represent more genes in the whole plant; the unique transcripts from former EST data contained more expression genes for fruit because the samples were derived from fruit with several stages. Above all, approximately 98% of transcripts had homologs in this study, and 97% had homologous hits in apple, pear and peach. Only 2% transcripts lacked homologs, which may be novel transcripts of hawthorn or errors in sequence or assembly. A total of 64,636 transcripts (77.12%) were predicted to have CDS of no less than 60 bp in length, based on blastx and ESTscan, and the distribution of length is shown in Figure 1. Nearly all the CDS (63,452) were identified with homologous matches in the protein data. Only 1,184 CDS were predicted with ESTScan. In this study, 77.12% transcripts were predicted to contain CDS, in agreement with chili pepper and with other reports (Liu et al., 2013). The transcripts without identified coding regions may be too short to meet the criteria for CDS prediction or may be non-coding RNAs. The putative non-coding RNAs need to be validated in a future study.

Table 1. Summary of sequencing and assembly for Hawthorn.

Reads	
Length (bp)	100+100
Total Number	32,332,207
Total Bases (bp)	6,530,452,306
GC percentage	48.51%
Q20 percentage	98.38%
Transcripts (≥200 bp)	
Total Number	83,817
Total Length (bp)	78,620,139
Clusters/contigs	11,551/39,952
Singletons	43,865
N50 Length (bp)	1,615
Mean Length (bp)	938
Average depth(X)	83

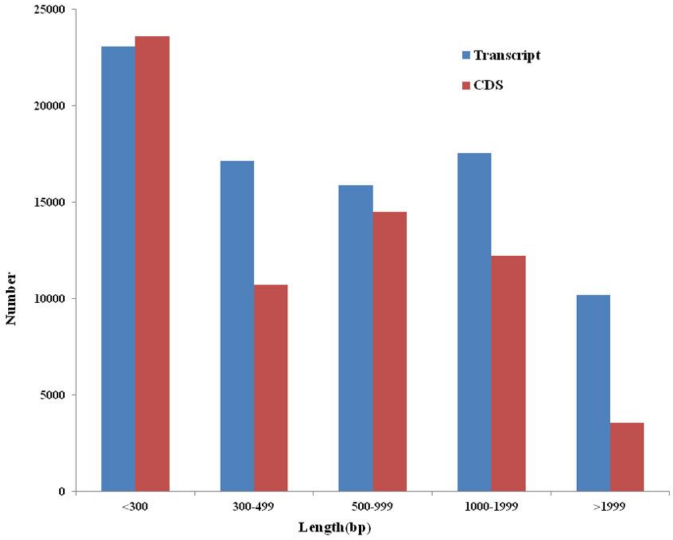


Fig 1. Length distribution of hawthorn transcripts and CDS.

Note：27,720 of transcripts (33.1%) were larger than 1000 bp in length.

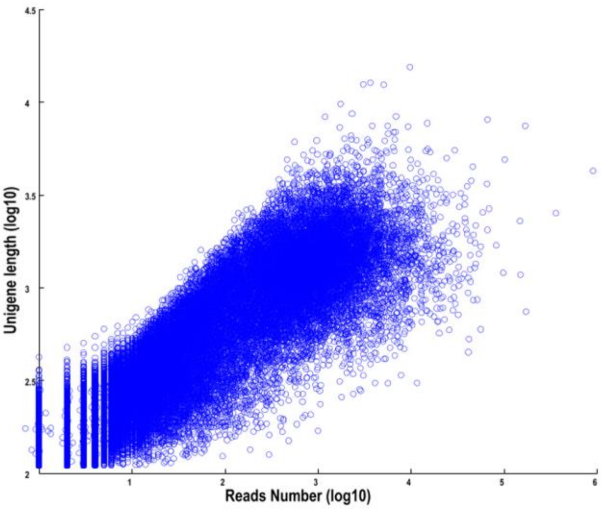
Polyphenolic compounds biosynthetic genes identified by KEGG annotation

Currently, hawthorn is well known to lower blood cholesterol and reduce the risk of cardiovascular disease because its fruit is rich in polyphenolic acid (Chang et al., 2005; Pettler et al., 2003). Flavonoids, one of the most prevalent polyphenolic acids, are also considered to be among the most important bioactive compounds in Chinese hawthorn *C. pinnatifida* Bge fruits (Liu et al., 2010). Phenylpropanoid biosynthesis was reported to be responsible for the biosynthesis of lignin, flavonoids and other metabolites in vascular plants (Li et al., 2010). Flavonoids are synthesized from phenylpropanoid derivatives by condensation with malonyl-CoA. Flavones and flavonols (3-hydroxyflavones) are common flavonoids in the plant kingdom. They are synthesized as part of the flavonoid modification pathways in aglycone and glucoside forms. To understand the key enzymes in the biosynthesis of these compounds, we identified the key genes, which were annotated as involved in the phenylpropanoid biosynthesis (map00940), flavonoid biosynthesis (map00941) and flavone and flavonol biosynthesis (map00944) pathways. Wu et al., (2012) identified 352 and 541 key phenylpropanoid biosynthesis genes in pear and apple, respectively (Table 3). In this study, we identified 143 and 115 transcribed genes in two hawthorn transcriptome and 212 genes in the whole

Table 2. Summary of annotations of Hawthorn transcripts.

Annotated databases	NO. of hits
Public database	61,985
NT	55,149
NR	57,038
UniProt/Swiss-Prot	36,061
KEGG	32,271
COG	20,425
GO	41,361
Apple	80,643
blastx[1]	60,857
blat[2]	77,627
Pear	77,099
blastx[1]	58,094
blat[2]	71,471
Peach	58,242
blastx[1]	55,792
blat[2]	22,793
EST data[3]	65,500
Total	82,221

Note：1 presents the transcripts were annotated by whole protein sequences using blastx, 2 presents the transcripts were annotated by whole genome using blat. 3 the data from NCBI (the accession number: GALU00000000) (Dai et al., 2013).



Note：the average sequencing depth is approximate 83X and the coverage of transcripts exhibited a positive relationship with the length of the given sequences

Fig 2. Log-log plot showing the dependence of unigene lengths on the number.

peach genome. We also identified 25, 15, 93, 42 and 42 peptides involved in the flavone and flavonol biosynthesis pathway and 98, 67, 401, 235 and 126 peptides involved in the flavonoid biosynthesis pathway in hawthorn, hawthorn EST, apple, pear and peach, respectively (Table 3 and Table S2). It is not surprising that fewer genes for major enzymes have been identified in hawthorn transcriptome than in apple, pear and peach, as only transcribed hawthorn genes are included in these data. Notably, there have two enzymes, F5H and FLS, Transcribed more copies in the hawthorn EST (NCBI; *Crataegus pinnatifida* accessions H8 and S7) than present study (Table 3). F5H is a key enzyme of S-lignin biosynthesis, and this data maybe suggestion that S-lignin is the dominant type in hawthorn fruit. FLS is a downstream enzyme in Flavonoid biosynthesis and it maybe favor to other polyphenolic compounds biosynthesis. These data corresponds to the source of sample, more Polyphenolic compounds biosynthesis in fruit.

Gene family comparative and Hawthorn-specific genes

To understand the gene family and hawthorn-specific genes, the gene family was identified among hawthorn, apple, pear and peach. Based on the hawthorn genes derived from mRNA-seq, putative alternative splice variants were filtered based on sequence similarity, using the criteria of an overlap ratio of no less than 70% between any sequences. The longest protein isoform was retained for each cluster. A total of 37,909 proteins were selected for gene family analysis. Using OrthoMCL methods, we obtained 14,800 gene families and 387 unique gene families of hawthorn among these four species (Figure 3). The majority of families were shared with apple (86.73%), followed by pear (82.55%) and peach (80.29%) (Figure 3), which suggest that hawthorn is more closely related to apple than to pear and peach at the genetic level. A significant percentage of transcripts (55.25% / 20943) in hawthorn weren't found to be from a conserved lineage. This result could be attributed to the presence of novel families. Alternatively, the derived transcripts may be from chimeric sequences (assembly errors) or non-conserved areas of proteins where homology is not detected, in agreement with several other studies (Wang et al., 2004; Liang et al., 2008; Mittapalli et al., 2010; Bai et al., 2011). The KEGG enrichment in these transcripts contained biosynthesis of other secondary metabolites (stilbenoid, diarylheptanoid and gingerol biosynthesis, phenylpropanoid biosynthesis, flavone and flavonol biosynthesis), terpenoid and polyketide metabolism (zeatin biosynthesis, brassinosteroid biosynthesis and limonene and pinene degradation), carbohydrate metabolism, lipid metabolism, amino acid metabolism and plant-pathogen interaction (Table 4 and Table S3). The biosynthesis of secondary metabolites was enriched in hawthorn compared to the related fruits, which may suggest that hawthorn possessed more (or special) genes for secondary metabolitesthat are associated with the characteristic acidic flavor as well as widely reported health benefits ofChinese hawthorn fruit, (Liu et al., 2010; Chang et al., 2005; Pittler et al., 2003). In contrast, apple, peach and pear are worldwide fruits with sweet flavors and little polyphenolic content compared to hawthorn fruit. The other various enriched pathways were mainly related to basic metabolism, carbohydrate metabolism, lipid metabolism, and amino acid metabolism.

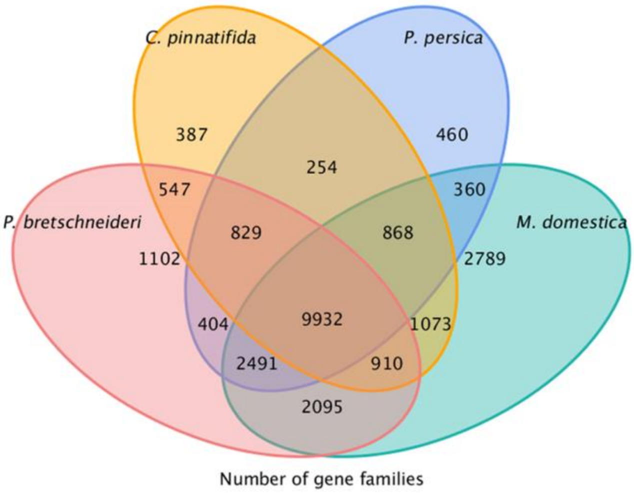
Putative molecular markers

Transcriptomes are important resources for the rapid and cost-effective development of genetic markers (Du et al., 2011; Garg et al., 2011). The molecular markers derived from the transcribed regions are more conservative, providing a greater potential for identifying functional genes. To detect SSR markers, all of the transcripts were scanned by the MISA Perl script (http://pgrc.ipk-gatersleben.de/misa/). In total, 9,180 (10.95%) transcripts contained 10,472 cDNA-SSR markers (Table 5). In these cDNA-SSRs, di-nucleotide (6,707) and tri-nucleotide (3,118) repeat motifs had the highest frequencies, followed by hexa-nucleotide repeats (332), quad-nucleotide repeats (159) and penta-nucleotide repeats (156). After designing and filtering primers, 827 cDNA-SSR markers were found to have at least one primer (Table S4), providing a platform for future understanding of hawthorn polymorphisms.

Table 3. The key enzymes of polyphenolic compounds biosynthesis identified by KEGG annotation in four fruits.

Phenylpropanoid biosynthesis	Hawthorn	Hawthorn EST	Apple	Pear	Peach
4CL	9	6	18(16)	13	8
C3'H	5	3	16(9)	7	7
C4H	8	1	24(19)	6	2
CAD	13	12	54(50)	33	40
CCOMT	5	4	21(18)	9	7
CCR	23	15	59(47)	34	26
COMT	1	1	2(2)	1	1
F5H	6	9	60(52)	39	13
HCT	24	22	195(171)	108	34
PAL	4	1	10(8)	3	2
POD	45	41	165(149)	99	72
Total	143	115	624(541)	352	212
Flavone and flavonol biosynthesis					
flavonol 3-O-methyltransferase	8	7	61	24	13
flavonoid 3'-monooxygenase	16	6	26	14	23
flavonol 3-O-glucosyltransferase	0	1	0	1	0
CYP75A	1	1	6	3	6
Total	25	15	93	42	42
Flavonoid biosynthesis					
naringenin 3-dioxygenase	2	2	6	3	1
CYP73A	8	1	26	10	2
caffeoyl-CoA O-methyltransferase	5	4	22	9	7
CHS	9	4	17	18	9
chalcone isomerase	4	3	22	9	3
leucoanthocyanidin dioxygenase	8	3	27	15	10
FLS	4	9	20	10	11
flavonoid 3'-monooxygenase	16	6	26	14	23
ANR	5	3	16	8	6
C3'H	5	3	16(9)	7	7
HCT	24	22	195(171)	108	34
LAR	2	2	8	3	2
DFR	5	4	25	15	8
CYP75A	1	1	6	6	3
Total	98	67	432(401)	235	126

Note 1 the apple and pear genes in Phenylpropanoid biosynthesis were identified by Wu et al (2012). For apple, the numbers in parenthesis show gene count after filtering the overlapped genes.



Note：14,800 gene families and 387 unique gene families of hawthorn among these four species were analysed. The majority of families were shared with apple (86.73%), followed by pear (82.55%) and peach (80.29%), which suggest that hawthorn is more closely related to apple than to pear and peach at the genetic level.

Fig 3. Hawthorn similarity comparison with apple, pear and peach

Table 4. Summary of KEGG pathway enrichment for hawthorn-specific peptides.

Pathway name	NO.	Q value	Pathway ID
Metabolism of terpenoids and polyketides			
Zeatin biosynthesis	111	6.65E-14	ko00908
Limonene and pinene degradation	79	2.40E-05	ko00903
Brassinosteroid biosynthesis	32	6.56E-05	ko00905
Biosynthesis of other secondary metabolites			
Stilbenoid, diarylheptanoid and gingerol biosynthesis	92	1.45E-05	ko00945
Phenylpropanoid biosynthesis	151	2.95E-05	ko00940
Flavone and flavonol biosynthesis	48	7.87E-04	ko00944
Carbohydrate metabolism			
Pyruvate metabolism	102	5.35E-05	ko00620
Propanoate metabolism	56	3.02E-04	ko00640
Glycolysis / Gluconeogenesis	114	4.71E-04	ko00010
Lipid metabolism			
Linoleic acid metabolism	42	9.70E-05	ko00591
Fatty acid metabolism	65	9.80E-05	ko00071
Amino acid metabolism			
Tryptophan metabolism	43	1.92E-04	ko00380
Valine, leucine and isoleucine degradation	66	3.74E-04	ko00280
Phenylalanine metabolism	70	3.74E-04	ko00360
Environmental adaptation			
Plant-pathogen interaction	643	4.35E-07	ko04626

Table 5. Length distribution of cSSRs based on the number of repeat units.

Number of repeats	Di-nucleotide repeats	Tri-nucleotide repeats	Quad-nucleotide repeat	Penta-nucleotide repeats	Hexa-nucleotide repeats
4	0	0	0	138	332
5	0	1,796	121	18	0
6	1,591	816	38	0	0
7	1,009	453	0	0	0
8	942	53	0	0	0
9	1,454	0	0	0	0
10	1,390	0	0	0	0
11	313	0	0	0	0
12	8	0	0	0	0
SubTotal	6,707	3,118	159	156	332

Materials and Methods

Plant materials and RNA extraction

‘Zezhouhong’ hawthorn, a hawthorn variety which is famous for it’s big size, good flavor and long history was obtained from Jincheng hawthorn garden, Shanxi Province. Fruits and other tissues including leaves, stem, shoots and root were dissected from a hundred-year-old tree, and immediately frozen and stored in liquid nitrogen until analysis. Total RNA was extracted from these materials using the Norgan RNA Purification Kit (Norgan Biotek Corp., Ontario, Canada). The quality and quantity of total RNA was analyzed using an UltrasecTM 2100 pro UV/Visible Spectrophotometer (Amer-sham Biosciences, Uppsala, Sweden) and gel electrophoresis. Equal quantities of high-quality RNA from each material were pooled for cDNA synthesis.

Construction of the mRNA-seq Library for Illumina Sequencing

The mRNA-seq library was constructed following the manufacturer’s instructions for the mRNA-Seq Sample Preparation Kit (Cat# RS-930-1001, Illumina Inc., San Diego, CA). Briefly, the poly-(A) mRNA was isolated from the total RNA samples using magnetic oligo (dT) beads. The mRNA was then fragmented and transcribed into first-strand cDNA using random hexamer-primers, followed by second-strand cDNA synthesis and the addition of adapters. Quality control analysis of the library was performed to quantify the DNA concentration and insert size. After quality control of the cDNA libraries, pair-end sequencing analysis was conducted via Illumina HiSeq™ 2000 according to the Illumina manufacturer’s protocol.

RNA-Seq data filter

To ensure the accuracy of subsequent analysis, the following filtering criterion was used to minimize the effects of sequencing error during gene assembly. Reads were removed if they contained the sequencing adaptor, more than 5% unknown nucleotides or more than 20% bases of low quality (quality scores in Phred scale less than 10). All reads were deposited in the National Center for Biotechnology Information (NCBI) and can be accessed in the Short Read Archive (SRA) under accession number SRP041853.

Transcript assembly

Trinity (trinityrnaseq_r2012-05-18; http://trinityrnaseq.sourceforge.net/) was used for *de novo* assembly to generate a set of transcripts (Grabherr et al., 2011). The following parameters were used in Trinity:

min_glue = 2, V = 10, edge-thr = 0.05, min_kmer_cov = 2, path_reinforcement_distance = 75, group_pairs_distance = 250. Then, any redundant fragments were removed using TGICL (TGI Clustering tools) and the Phrap assembler (Pertea et al., 2003). The following parameters were used to ensure a high quality of assembly: -l 40 -v 25 -O '-repeat_stringency 0.95 -minmatch 35 -minscore 35'. Finally, based on sequence similarity, the transcripts were divided into two classes: cluster (prefixed with 'CL') and singleton (prefixed with 'unigene'). Within a cluster, the similarity between transcripts was greater than 70%.

Annotation and CDS prediction

The transcripts were annotated using public plant databases, whole protein sets of related species and hawthorn EST data. First, the transcripts were aligned to four public protein databases, the NCBI non-redundant (Nr) database, Swiss-Prot protein database (Swiss-Prot), Kyoto Encyclopedia of Genes and Genomes (KEGG) and Clusters of Orthologous Groups of proteins (COG). Based on Nr annotation, the GO annotation was analyzed by Blast2GO software (v2.5.0) (Conesa et al., 2005). In addition, transcripts were annotated with the NCBI non-redundant nucleotide (Nt) database using blastn. Second, they were aligned to the whole protein sets and genomes of three related species (WPRS; apple, pear and peach) by blastx with a cut-off E-value of 1e-5 and blat (Jun et al., 2012; Velasco et al., 2010; Verde et al., 2013). Third, using blastn with the cut-off E-value of 1e-10, a homology analysis was performed against the public hawthorn EST (NCBI, accession number GALU00000000) (Dai et al., 2013). The CDS was predicted using blastx and ESTscan. The best hit from WPRS and the four public protein databases was used to determine the sequence direction and CDS (coding sequences) of the transcripts. When different databases conflicted with each other, the results were prioritized in the following order: WPRS, nr, Swiss-Prot and KEGG. When a transcript was not covered in blastx, it was predicted by ESTScan. The shortest CDS were at least 60 bp, and the peptide sequences were translated using standard codons.

Gene family comparative and Hawthorn-specific Genes

To identify the hawthorn-specific genes and gene families, we selected the following related species to represent sequenced species: apple (*Malus* x *domestica*), peach (*Prunus persica*) and pear (*Pyrus bretschneideri*). For comparative analysis, we used the following pipeline to cluster individual genes into gene families: 1) we collected protein sequences longer than 30 amino acids from these four species, with the longest protein isoform being retained for each gene or cluster; 2) blastp was used to align all protein sequences against a database containing a protein dataset of all species with an e-value of 1e-5; 3) the gene families were extracted by the stand-alone OrthoMCL (Li et al., 2003; Li et al., 2003) program using a default MCL inflation parameter of 1.5.

Pathway enrichment analysis for the hawthorn-specific genes was performed based on the algorithm presented by KOBAS (Xie et al., 2011), with the whole hawthorn transcriptome set as the background. The p-value was approximated by the hypergeometric distribution test and multiple testing corrections using FDR. The enriched cutoff is Q-value less than 0.001.

Discovery of cDNA-SSR (cSSR)

The cSSRs were identified using a Perl script from MIcroSAtellite (MISA), with transcripts for reference. The di-, tri-, tetra-, penta- and hexa-nucleotide sequences, with minimum repeat numbers of 6, 5, 5, 4 and 4, respectively, were applied as the search criteria (http://pgrc.ipk-gatersleben.de/misa/). Primer3-2.3.4 was used to design PCR primers with default settings. Primers were filtered based on the following criteria: (1) no SSRs in the primer; (2) three mismatches at the 5’ -end and one mismatch at the 3’ -end were allowed when aligning primers to transcripts; (3) each primer could only map to one transcript (Untergasser et al., 2012).

Conclusion

Our data provide the most comprehensive transcriptomic resource currently available for Chinese hawthorn (*Crataegus pinnatifida*). Candidate genes potentially involved in polyphenolic compound biosynthesis were identified, and the genes predicted to be unique to Chinese hawthorn are expected to yield better insight into Chinese hawthorn gene diversity. Many of the characterized SSRs contributed to marker development. These data constitute a new valuable resource for genomic studies on Chinese hawthorn.

Conclusions

This study aimed to provide basic transcriptomics information on Chinese hawthorn *C. pinnatifida*. Firstly, a large number of candidate genes potentially involved in polyphenolic compound biosynthesis pathways were identified and compared to related species, which are worthy of further investigation; Secondly, a significant number of SSRs were identified, which could facilitate the identification of polymorphisms in hawthorn populations; In addition, orthologous sequences and unigenes unique to hawthorn were preliminary classified. These datasets will improve our understanding of the molecular mechanisms of polyphenolic compound biosynthesis in hawthorn, and these characteristic sequences will provide a new platform for hawthorn molecular studies.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

MCC QFC conceived and designed the experiments, MXY ZGD performed the experiments, QFC anlyzed the data, MXY ZGD QFC constructed the database platform and drafted the manuscript. All authors have read and approved the final manuscript.

Acknowledgements

We gratefully acknowledge the Beijing Biomarker Technologies co., LTD for the data analysis, MXY is supported by the Department of Science and Technology of Shanxi in China (Grant No, 20110311013-2). But the funding had no role in study design, data collection and analysis, the decision to publish, or the preparation of the manuscript.

References

Bai X, Mamidala P, Rajarapu SP, Jones SC, Mittapalli O (2011) Tr anscriptomics of the Bed Bug (*Cimex lectularius*). PLoS One. 6(1): e16336.

Chang WT, Dao J, Shao ZH (2005) Hawthorn: Potential roles in cardiovascular disease. Am J Chin Med. 33: 1–10.

Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M (2005) Blast2go: a universal tool for annotation,visualization and analysis in functional genomics research. Bioinformatics. 21(18):3674-3676.

Cui T, Nakamura K, Tian S, Kayahara H, Tian YL (2006) Polyphenolic content and physiological activities of Chinese hawthorn extracts. Biosci Biotechnol Biochem. 70:2948–2956.

Dai H, Zhang Z, Guo X (2007) Adventitious bud regeneration from leaf and cotyledon explants of Chinese hawthorn (*Crataegus pinnatifida* Bge. var. *major* N.E.Br.). *In Vitro* Cell Dev Biol. 43:2–8.

Dai H, Han G, Yan Y, Zhang F, Liu Z, et al. (2013) Transcript assembly and quantification by RNA-Seq reveals differentially expressed genes between soft-endocarp and hard-endocarp hawthorns. PLoS ONE. 8(9): e72910.

Du H, Bao Z, Hou R, Wang S, Su H, Yan J, Tian M, Li Y, Wei W, Lu W, Hu XL, Wang S, Hu JJ (2012) Transcriptome sequencing and characterization for the sea cucumber Apostichopus japonicus (Selenka, 1867). PLoS One. 7: e33311.

Garg R, Patel RK, Tyagi AK, Jain M (2011) De novo assembly of chickpea transcriptome using short reads for gene discovery and marker identification. DNA Res. 18(1): 53-63.

Gazdik Z, Krska B, Adam V, Saloun J, Pokorna T, Reznicek V, Horna A, Kizek R (2008) Electrochemical determination of the antioxidant potential of some less common fruit species. Sensors. 8:7564–7570.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 29: 644–652.

Wu J, Wang Z, Shi Z, Zhang S, Ming R, Zhu S, Khan MA, Tao S, Korban SS, Wang H, Chen NJ, Nishio T, Xu X, Cong L, Qi K, Huang X, Wang Y, Zhao X, Wu J, Deng C, Gou C, Zhou W, Yin H, Qin G, Sha Y, Tao Y, Chen H, Yang Y, Song Y, Zhan D, Wang J, Li L, Dai M, Gu C, Wang Y, Shi D, Wang X, Zhang H, Zeng L, Zheng D, Wang C, Chen M, Wang G, Xie L, Sovero V, Sha S, Huang W, Zhang S, Zhang M, Sun J, Xu L, Li Y, Liu X, Li Q, Shen J, Wang J, Paull RE, Bennetzen JL, Wang J, Zhang S (2012) The genome of pear (*Pyrus bretschneideri* Rehd.) Genome Res. 23 (2): 396-408

Jurikova T, Sochor J, Rop O, Mlcek J, Balla S, Szekeres L, Adam V, Kizek R (2012) "Polyphenolic profile and biological activity of chinese hawthorn (*Crataegus pinnatifida* BUNGE) Fruits." Molecules. 17(12): 14490-14509.

Jurikova T, Rop O, Mlcek J, Sochor J, Balla S, Szekeres L, Hegedusova A, Hubalek J, Adam V, Kizek R (2012) Phenolic profile of edible honeysuckle berries (*Genus Lonicera*) and their biological effects. Molecules. 17: 61–79.

Kao ES, Wang CJ, Lin WL, Yin YF, Wang CP, Tseng TH (2005) Anti-inflammatory potential of flavonoid contents from dried fruit of Crataegus pinnatifida in vitro and in vivo. J Agric Food Chem. 53: 430–436.

Li L, Stoeckert Jr CJ, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13: 2178-2189.

Li X, Bonawitz ND, Weng JK, Chapple C (2010) The growth reduction associated with repressed lignin biosynthesis in Arabidopsis thaliana is independent of flavonoids. Plant Cell. 22: 1620–1632.

Liang H, Carlson JE, Leebens-Mack JH, Wall PK, Mueller LA, Buzgo M, Landherr LL, Hu Y, DiLoreto DS, Hut DC, Field D, Tanksley SD, Ma H, dePamphilis CW (2008) An EST database for Liriodendron tulipifera L. floral buds: the first EST resource for functional and comparative genomics in Liriodendron. Tree Genet Genomes. 4: 419–433.

Liu PZ, Kallio H, Lu DG, Zhou CS, Ou SY, Yang BR (2010) Acids, sugars, and sugar alcohols in chinese hawthorn (*Crataegus spp.*) fruits. J Agric Food Chem. 58:1012–1019.

Liu S, Li W, Wu Y, Chen C, Lei J (2013) De novo transcriptome assembly in chili pepper (*Capsicum frutescens*) to identify genes involved in the biosynthesis of capsaicinoids. PLoS One. 8(1): e48156.

Mittapalli O, Bai X, Mamidala P, Rajarapu SP, Bonello P, Herms DA(2010) Tissue-specific transcriptomics of the exotic invasive insect pest emerald ash borer (*Agrilus planipennis)*. PLoS One. 5(10): e13708.

Pertea G, Huang XQ, Liang F, Antonescu V, Sultana R, Karamycheva S, Lee Y, White J, Cheung F, Parvizi B, Tsai J, Quackenbush J (2003) TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics. 19: 651–652.

Pittler MH, Schmidt K, Ernst E (2003) Hawthorn extract for treating chronic heart failure: Meta-analysis of randomized trials. Am J Med. 114:665–674.

Rigelsky JM, Sweet BV (2002) Hawthorn: Pharmacology and therapeutic uses. Am J Health-Syst Pharm. 59: 417–422.

Rop O, Posolda M, Mlcek J, Reznicek V, Sochor J, Adam V, Kizek R, Sumczynski D (2012) Qualities of native apple cultivar juices characteristic of central europe. Not Bot Horti Agrobot Cluj Na. 40:222–228.

Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG (2012) Primer3--new capabilities and interfaces. Nucleic Acids Res. 40(15): e115.

Velasco R, Zharkikh A, Affourtit J, Dhingra A, Cestaro A, Kalyanaraman A, Fontana P, Bhatnagar SK, Troggio M, Pruss D, Salvi S, Pindo M, Baldi P, Castelletti S, Cavaiuolo M, Coppola G, Costa F, Cova V, Dal Ri A, Goremykin V, Komjanc M, Longhi S, Magnago P, Malacarne G, Malnoy M, Micheletti D, Moretto M, Perazzolli M, Si-Ammour A, Vezzulli S, Zini E, Eldredge G, Fitzgerald LM, Gutin N, Lanchbury J, Macalma T, Mitchell JT, Reid J, Wardell B, Kodira C, Chen Z, Desany B, Niazi F, Palmer M, Koepke T, Jiwan D, Schaeffer S, Krishnan V, Wu C, Chu VT, King ST, Vick J, Tao Q, Mraz A, Stormo A, Stormo K, Bogden R, Ederle D, Stella A, Vecchietti A, Kater MM, Masiero S, Lasserre P, Lespinasse Y, Allan AC, Bus V, Chagné D, Crowhurst RN, Gleave AP, Lavezzo E, Fawcett JA, Proost S, Rouzé P, Sterck L, Toppo S, Lazzari B, Hellens RP, Durel CE, Gutin A, Bumgarner RE, Gardiner SE, Skolnick M, Egholm M, Van de Peer Y, Salamini F, Viola R (2010) The genome of the domesticated apple (*Malus* × *domestica* Borkh.). Nat Genet. 42: 833-839.

Verde, Ignazio, Abbott, Albert G (2013) The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. Nat Genet. 45(5):487-94

Wang J-PZ, Lindsay BG, Leebens-Mack J, Cui L, Wall K, Miller WC, dePamphilis CW (2004) EST clustering error evaluation and correction. Bioinformatics. 20: 2973–2984.

Xie C, Mao X, Huang J, Ding Y, Wu J, Dong S, Kong L, Gao G, Li C, Wei L (2011) KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. Nucleic Acids Res. 39:W316-322.

Yang BR, Liu PZ (2012) Composition and health effects of phenolic compounds in hawthorn (*Crataegus spp.*) of different origins. J. Sci. Food Agric. 92:1578–1590.

Zhao HC, Tian BF (1996) China Fruit-Plant Monograph-Hawthorn Flora; China Forestry Publishing House: Beijing, China,.