Comparative transcriptome provides strategy for phylogenetic analysis and SSR marker development in *Chaenomeles*

Wenhao Shao
Chinese Academy of Forestry
Shiqing Huang
Longshan Forest Farm of Anji County
Yongzhi Zhang
Longshan Forest Farm of Anji County
Jingmin Jiang
Chinese Academy of Forestry
Hui Li (cerclihui@caf.ac.cn)
Guangzhou Institute of Forestry and Landscape Architecture

Research Article

Keywords: Chaenomeles, transcriptome, phylogenetic relationships, selective pressure, SSR marker

Posted Date: April 12th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-393262/v1

Version of Record: A version of this preprint was published at Scientific Reports on August 12th, 2021. See the published version at https://doi.org/10.1038/s41598-021-95776-z.

Abstract

The genus *Chaenomeles* has long been considered as an important ornamental, herbal and cash plant and widely cultivated in East Asian. Traditional researches of *Chaenomeles* mainly focus on evolutionary relationships on phenotypic level. In this study, we conducted RNA-seq for 10 *Chaenomeles* germplasms supplemented with one related species *Docynia delavayi* (*D. delavay*) by Illumina HiSeq2500 platform. After de novo assemblies, we have generated unigenes for each germplasm with numbers from 40 084 to 48 487. By pairwise comparison of the orthologous sequences, 9 659 othologus within the 11 germplasms were obtained, with 6 154 othologous genes identified as single-copy genes. The phylogenetic tree was visualized to reveal evolutionary relationship for these 11 germplasms. GO and KEGG analyses were performed for these common single-copy genes to compare the functional similarities and differences. Selective pressure analysis based on 6 154 common single-copy genes reveals that 45 genes were under positive pressure selection. Most of them involved in plant disease defense system building process. 292 genes containing simple sequence repeats (SSRs) were used to develop SSR markers and compare their function in secondary metabolism pathways. Finally, 10 primers were chosen as SSR markers candidates for *Chaenomeles* germplasms by comprehensive screening. Our research provides new methodology and reference for future related research in *Chaemomeles* and is also useful for improvement, breeding and selection project in other related species.

Introduction

Chaenomeles, a genus within the subfamily Maloideae (Rosaceae) comprising four diploid (2n = 34) species [1], is widely planted in East Asian and has important ecologic, ornamental and economic value. In China, *Chaenomeles* has more than 3000 years cultivated history. About 700 years ago, people began to recognize *Chaenomeles*' medicinal values: their effect on curing rheumatoid arthritis and hypopepsia [2]. Now, pharmacologists find their leaves and flowers contain substantial prunasin; their seeds have a large content of amygdalin; their fruits contain abundance chemical ingredients such as oleanolic acid, malic acid, pectinic acid, Chinese wolfberry citric acid, tartaric acid, citric acid, tannin, flavonoids saponin and proanthocyanidins which acts as antioxidant and anticancer agents [3]. Rest of these ingredients also have different pharmacological activities [4].

In the *Rosaceae* family, *Chaenomeles* has close phylogeny relationships with *Cydonia*, *Malus* and *Pyrus* genus. Lindley established *Chaenomeles* genus in 1822. Subsequently, some taxonomists began to reclassify some species into this genus. In 1890, Koehne reclassified *Cydonia Sinensis (C. Sinensis)* as *Chaenomeles sinensis* (Thouin) Koehne (*C. sinensis*) [5]. In 1906, Schneider reclassified *Cydonia cathayensis* (*C. cathayensis*) as *Chaenomeles cathayensis* (Hemsl.) Schneid.(*C. cathayensis*) [6]. In 1929, Nakai also reclassified *Cydonia speciosa* (*C. speciosa*) as *Chaenomeles speciosa* (Sweet) Nakai (*C. speciosa*) [7]. All of these three species had been thought the members of *Cydonia* before that time. However, Koehne's classification result is still controversial and not be in agreement [8]. Some taxonomists established independent genera *Pseudochaenomeles* Carr. (1882) and *Pseudocydonia* Schneid. (1906). Phipps [9,10] considered that *Pseudocydonia* as an independent genus, in addition, it may be an intermediate type between *Chaenomeles* and *Cydonia*. Due to these controversies, the number of *Chaenomeles* now still is ambiguous. According to I. et al[1], there are only four species in this genus. However, there are still five species identified in China, including *Chaenomeles speciosa* (Sweet) Nakai (*C. speciosa*), *Chaenomeles cathayensis* (Hemsl) Schneider (*C. cathayensis*), *Chaenomeles japonica* (Thunb) Lindley (*C. japonica*), *Chaenomeles thibetica* Yu (*C. thibetica*) and *Chaenomeles sinensis* (Thouin) Koehne (*C. sinensis*) [2,11].

Efficient methods to clarify the taxonomic status of both the wild and the cultivated germplasms are much needed [1]. In present, both morphological traits and various molecular markers are used to solve the taxonomic confusion [12–14]. In the past several years, Restriction fragment length polymorphism (RFLP), Random Amplified Polymorphic DNA (RAPD), Amplified fragment length polymorphism (AFLP), Simple Sequence Repeats (SSR), EST-based microsatellite (EST-SSR), Inter-simple sequence repeat (ISSR), Sequence characterized amplified region (SCAR) and single nucleotide polymorphism (SNP) have developed into the mainstream method to detect evolutionary relationship in relative species on genomic level. Lots of related researches were reported. In *Chaenomeles*, Barthish et al [15] employed RADP to analysis offspring families of *C. cathayensis* and *C. speciosa*. Their result showed the RAPD-based proportion of between-family variability is greatly higher in the hybrid populations than in pure species. He et al [16] evaluated genetic relationships among 52 *C. specisa* accessions grown in China by combining AFLP with leaf morphology characters. 208 polymorphic markers were identified. Zhang et al, used EST-SSR markers of *Malus* to research genetic diversity of 33 *Chaenomeles* germplasms and obtain exciting results [17]. With the development the next generation sequencing (NGS) technologies and algorithms for data processing, RNA-seq has become an efficient and economical approach for genomic and trascriptomic resources digging, and has been increasingly considered as an efficient tool to identify transcript-wide molecule markers and to solve the phylogeny relationships innon-model species [18–21]. Although few studies about transcriptome datasets for individual species, increasing attention has been given to comparative transcriptomes in genus *Chaenomeles*.

Here, we sequenced transcriptomes for 10 *Chaenomeles* germplasm which belong to five *Chaenomeles* species and one *D. delavayi* germplasm which is near-source species of *Chaenomeles* by Illumina HiSeq 2500 platform. After denovo assemble and functional annotation, common orthologous genes among the 11 germplasms were obtained. We selected the single-copy genes from all common orthologous genes to conduct phylogenetic tree construction to confirm their evolutionary relationship. By pairwise comparison of the orthologous sequences, speciation differentiation genes under positive selection were obtained to illustrate some key genes during evolution. A plenty of SSR sites were identified to be potential molecular markers for *Chaenomeles*. Common genes containing SSR sites among the 10 *Chaenomeles* germplasms were selected to detect gene expression model in secondary metabolism pathway. Finally, we developed some SSR marker for these 10 *Chaenomeles* germplasms. This study displays the first comparative exploration of *Chaenomeles* transcriptomes using high-throughput RNA-seq and provides an important methodology and database resources for facilitating further studies on the phylogeny of *Chaenomeles* and its related genus.

Results

De novo assembly and unigene qualification result of the 11 germplasms

After assembly of unigenes for the 11 germplams, totally we obtained 42 169 unigenes in *D. delavayi* (DY), 42 868 unigenes in *C. sinensis* (GP), 48 011 unigenes in *C. speciosa* (ZPCA), 48 487 unigenes in *C. speciosa* (ZPXC), 47 160 unigenes in *C. speciosa* (ZPLP), 46 193 unigenes in *C. speciosa* (ZPLY), 43 875 unigenes in *C. thibetica* (XZ), 40 084 unigenes in *C. thibetica* (XZSS), 49 571 unigenes in *C. thibetica* (MYXZ), 41 091 unigenes in *C. cathayensis* (MY), and 47 856 unigenes in *C. japonica* (RB) respectively (Fig. 2A). N50 numbers were 7 584 in *D. delavayi* (DY), 8 069 in *C. sinensis* (GP), 8 273 in *C. speciosa* (ZPCA), 8 304 in *C. speciosa* (ZPXC), 8 174 in *C. speciosa* (ZPLP), 8 374 in *C. speciosa* (ZPLY), 8 077 in *C. thibetica* (XZ), 7 508 in *C. thibetica* (XZSS), 8 369 in *C. thibetica* (MYXZ), 7 552 in *C. cathayensis* (MY), and 8 338 in *C. japonica* (RB) respectively (Fig. 2B).

The lengths of N50 were 1 655bp in *D. delavayi* (DY), 1 744bp in *C. sinensis* (GP), 1 699bp in *C. speciosa* (ZPCA), 1 716bp in *C. speciosa* (ZPXC), 1 706bp in *C. speciosa* (ZPLP), 1 696bp in *C. speciosa* (ZPLY), 1 807bp in *C. thibetica* (XZ), 1 857bp in *C. thibetica* (XZSS), 1 724bp in *C. thibetica* (MYXZ), 1 712bp in *C. cathayensis* (MY), and 1 651bp in *C. japonica* (RB) respectively (Fig. 2C). These results suggested our assembly results are good enough for further analyses. In addition, we displayed the gene expression level of the 11 germplasms (Fig. 2D). After standardization by $\log_{10}$ (RPKM) algorithm, the expression level ranged from − 0.5 to 2. Except *C. thibetica* (XZ) and *C. thibetica* (MYXZ), the rest gernplams had higher proportion of gene expressed at 0.1–0.2 after standardization. The median of gene expression level was greater in *C. thibetica* (XZSS) than those in the rest germplasms (Fig. 2D) indicating different gene expression distributions in *C. thibetica* (XZ), *C. thibetica* (XZSS) and *C. thibetica* (MYXZ).

Screening of othologous genes and single-copy genes

To better understand the similarities and differences among the 11 germplasms, comparative analysis of genes among different germplasms is the most efficient methods. Genes from different germplasms with the best blast hits are considered as orthologs. All these orthologs were commonly used to generate orthologous pairs. We identified orthologous groups for all predicted nucleic acid and proteins sequence from the 11 matierals by using OrthoMCL software. After merging same family gene, finally, we obtained 28 181 OrthoMCL genes in *C. thibetica* (XZ), 28 908 OrthoMCL genes in *D. delavayi* (DY), 29 882 OrthoMCL genes in *C. speciosa* (ZPLP), 28 725 OrthoMCL genes in *C. sinensis* (GP), 31 682 OrthoMCL genes in *C. thibetica* (MYXZ), 27 989 OrthoMCL genes in *C. cathayensis* (MY), 31 472 OrthoMCL genes in *C. japonica* (RB), 26 214 OrthoMCL genes in *C. thibetica* (XZSS), 30 404 OrthoMCL genes in *C. speciosa* (ZPCA), 30 410 OrthoMCL genes in *C. speciosa* (ZPXC), and 29 960 OrthoMCL genes in *C. speciosa* (ZPLY) (Supplementary Table S1). Then we found the common othologs among 11 germplasms by venn diagram. A total of 9659 othologous genes were obtained (Fig. 3A). Single-copy genes are valuable resources for phylogenetic analysis and SSR marker selection, which only have one copy in genome of corresponding species. We selected single-copy gene within 9 659 othologs also by using venn graph. And 6 154 othologous single-copy genes were obtained in our result (Fig. 3B, Supplementary Table S2).

Phylogenetic analysis based on single-copy orthologous families

In our study, the single-copy gene families were used to conduct phylogenetic analysis. We finally identified 6 154 orthologous groups in the 11 germplasms. Using these orthologous families, we constructed an evolutionary tree via iqtree software based on the Maximum likelihood method. According to the evolutionary tree, the 11 germplasms were divided into different clades (Fig. 4).

C. sinensis (GP) and *D. delavayi* (DY) are in the same branch and are relatively far away from other species, indicating that they have close phylogenetic relationship. Compared with other species in the genus, *C. japonica* (RB) was classified as a single branch, showing interspecific differentiation, which is also supported by AFLP studies [22]. *C. speciosa* (ZPLY), commonly known as 'Yizhoumugua' in Linyi City of Shandong Province, is generally considered to be originated from interspecific hybridization with other species in the genus [23]. Our study showed that it had a relatively distant genetic relationship with the other three *Chaenomeles* germplasms.

Three germplasms of *C. thibetica*, namely XZ, XZSS and MYXZ, clustered in the same branch with *C. cathayensis* (MY), showing a very close phylogenetic relationship. Previous study [23] suggested that *C. thibetica* might be a hybrid species of *C. cathayensis* and *D. delavayi* (DY). However, this view still needs more sufficient evidence to judge.

C. speciosa is the most widely distributed species of the genus in China. Due to its outstanding pharmacological effects, it has been domesticated and cultivated for a long time and formed three genuine producing areas, corresponding to the distribution of three local varieties, namely 'Cunmugua' (ZPCA), 'Xuanmugua' (ZPXC) and 'Ziqiumugua' (ZPLP). Our phylogenetic analysis proved their close relationship.

GO and KEGG analyses result for single-copy genes in 11 germplasms

In order to understand the function of 6 154 common single-copy genes in 11 germplasms, we conducted GO and KEGG analyses (Fig. 5). According to the GO and KEGG result, 6 154 common single-copy genes shared similar metabolism process. After merging the top 15 GO terms in each germplasm, we totally obtained 27 common GO terms (Fig. 5A). Common single-copy genes were significantly enriched in these 27 GO metabolic processes in all the 10 *Chaenomeles* germplasms except *D. delavayi* (DY). In *D. delavayi* (DY), common single-copy genes were insignificantly enriched in DNA metabolic process (GO: 0006259) with *p.* adjust 0.294 and chromosome organization (GO: 0051276) with *p.* adjust 0.564 respectively indicating its different metabolic process from *Chaenomeles* (Fig. 5A Supplementary Table S3). All of the 11 germplasms especially enriched in nitrogen compound metabolic process (GO: 0006807), organic cyclic compound metabolic process (GO: 1901360), heterocycle metabolic process (GO: 0046483) cellular nitrogen and compound metabolic process (GO: 0034641), and nucleic acid metabolic process (GO: 0090304) (Fig. 5A, Supplementary Table S3).

In KEGG analysis, after merging the top 15 pathways in each germplasm, finally we obtained 31 common pathways (Fig. 5B). The common single-copy genes were significantly enriched in Non-homologous end-joining (ko03450), Ribosome biogenesis in eukaryotes (ko03008) and Purine metabolism (ko00230) in all of 11 germplasms (Fig. 5B Supplementary Table S4). Except these two pathways, *D. delavayi* (DY) also enriched in Biosynthesis of amino acids (ko01230) with gene number of 97 out of 297 (*Q* value 0.032); *C. sinensis* (GP) also enriched in Spliceosome (ko03040) and Biosynthesis of amino acids (ko01230) with gene number of 89 out of 263 (*Q* value 0.008) and 96 out of 296 (*Q* value 0.0158) respectively; *C. cathayensis* (MY) enriched in Aminoacyl-tRNA biosynthesis (ko00970) with gene number of 31 out of 70 (0.006); *C. thibetica* (MYXZ) enriched in Base excision repair (ko03410) with gene number of 24 out of 52 (*Q* value 0.0006); *C. japonica* (RB) also enriched in N-Glycan biosynthesis (ko00510) with gene number of 24 out of 57(*Q* value 0.0257); *C. thibetica* (XZ) also significantly enriched in Biosynthesis of amino acids (ko01230), Nucleotide excision repair (ko03420) with gene number of 97 out of 290 (*Q* value 0.0097) and 36 out of 87 (*Q* value 0.0097) respectively; *C. thibetica* (XZSS) enriched in Base excision repair (ko03410) and Spliceosome (ko03040) with number of 24 out of 48(*Q* value 0.0204) and 88 out of 247(*Q* value 0.0204) respectively; *C. speciosa* (ZPCA) significantly Nucleotide excision repair (ko03420) with gene number of 36 out of 89 (*Q* value 0.0014); Both of *C. speciosa* (ZPLP) and *C. speciosa* (ZPLY) enriched in Base excision repair (ko03410) with gene number of 25 out of 54 (*Q* value 0.0016) and 25 out of 46 (*Q* value 0.00096) respectively. *C. speciosa* (ZPXC) enriched in Aminoacyl-tRNA biosynthesis (ko00970) and Spliceosome (ko03040) with gene number of 32 out of 70 (*Q* value 0.0004) and 88 out of 290 (*Q* value 0.0034) respectively (Fig. 5B, Supplementary Table S4).

Detecting genes under selective pressure

The non-synonymous (dN) substitutions and synonymous (dS) substitutions are used to estimate the change of coding protein sequences. The magnitude of the Ka/Ks ratio provides evidence of genes under strong functional constraints (Ka/Ks < 1) or undergoing adaptive evolution (Ka/Ks > 1) [24–26]. The ratio of non-synonymous substitution rate (Ka) and synonymous substitution rate (Ks) are used to assess whether genes under selection. We generated pairwise comparison of different transcriptome datasets. According to the blast results, a total of 6 057 single-copy genes underwent selective process (Fig. 6A, Supplementary Table S5A). Among them, 467 out of 6 057 genes underwent purification selection indicating those genes are disadvantage during evolution process. 693 out of 6 057 genes experienced weak positive selection and 45 out of 6 057 genes experienced strong positive selection indicating those genes are domain genes that determine evolution direction (Fig. 6B, Supplementary Table S5B-D). Further, we blasted these 45 positive selected genes to Arabidopsis (Supplementary Table S5D). 19 out of 45 genes can be found homologous genes in Arabidopsis. ORTHOMCL10413 (*RGA2*) mainly acts as a repressor of the gibberellin (GA) signaling pathway through transcription coactivator of the zinc finger transcription factors GAF1/IDD2 and ENY/IDD1 to regulate gibberellin homeostasis and signaling [27]. ORTHOMCL9348 (*RPL5*) participates in cell proliferation and plays a role for translation in leaf dorsoventral patterning to specify leaf adaxial identity [28]. ORTHOMCL8960 (*HEL*) has a modular structure consisting of an N-terminal hevein-like domain (CB-HEL) and a C-terminal domain (CD-HEL) that is posttranslationally processed. Both domains show a strong antifungal activity indicating this gene is charge for defending plant against pathogen attack [29]. ORTHOMCL9718 (*FIP2*) can delay flowering by repress expression of FLOWERING LOCUS C [30]. ORTHOMCL4564 (*CSA1*) also play important role in disease defense [31]. ORTHOMCL9773 (*MYB27*) is member of MYB family, lots of studies indicates that MYB family members can regulate secondary metabolism, control cell shape and defense disease resistance and increase hormone responses [32]. ORTHOMCL8338 (*BZIP60*) involves in controlling endoplasmic reticulum pressure to enhance immune response in both animal and plants [33]. In addition, although in plants, ORTHOMCL5656 (*CYS6*), ORTHOMCL6947 (*HPR3*) and ORTHOMCL6257 (*F11P17.9*) still have no reports for their function, according GO analysis result, these three genes main participated in disease defense process (Supplementary Table S3). All of these indicate most of the positive selection genes involves in plant disease defense system building. That is why those plants can survive during long period evolution and selection process.

Gene expression model in biosynthesis of secondary metabolisms pathway for 10 germplasms of *Chaenomeles*

In all 6 154 common single-copy genes, we predicted SSR sites by using MISA software. A total of 1 210, 1 151, 1 128, 1 253, 1 186, 1 225, 1 178, 1 181, 1 170, 1 165 genes contained SSR sites in *C. sinensis* (GP), *C. japonica* (RB), *C. cathayensis* (MY), *C. thibetica* (XZ), *C. thibetica* (XZSS), *C. thibetica* (MYXZ), *C. speciosa* (ZPCA), *C. speciosa* (ZPLP), *C. speciosa* (ZPXC), and *C. speciosa* (ZPLY) respectively (Supplementary Table S6). We picked out the common genes which contain SSR markers for 10 germplasms. Totally, 292 common single-copy genes were selected (Fig. 7, Supplementary Table S7). Further, we conducted KEGG analyses for these 292 common single-copy genes. All of the 10 germplasms were significantly enriched in the biosynthesis of secondary metabolisms (Supplementary Table S8). So we applied this pathway to check the gene expression model in each germplasm. Although all of the 10 germplasms had genes participated same process in the pathway, there still existed differences on the gene expression level. For examples, from phosphatidylethanolamine to 1, 2 diacyl-sn-glycerol, and 1, 2 diacyl-sn-glycerol to phosphatidylcholine, *NPC6* expressed lower in *C. thibetica* (MYXZ) than the rest gremplams. From Magnesium protoporphyrin to Protoporphyrin, *CHLD* expressed lower in *C. thibetica* (XZ) than the rest germplasms. During transformation from 3-oxoacyl-CoA to acyl-CoA and Malonyl-CoA, *KCSLL* expressed higher in *C. sinensis* (GP), C. japonica (RB), *C. thibetica* (XZSS) and *C. speciosa* (ZPXC) than the rest germplasms (Fig. 8, Supplementary Table S9). However, there were similarities of gene expression in 10 germplasms. For instances, during transformation of 6-Geranylgerany-2- methylbenzene-1, 4-diol to 6-Geranylgeranyl − 2, 3 dimetheyl benzene-1, 4-diol, shikimate 3-phosphate to O5-(1-Carboxyvinyl) -3- phosphoshikimate, (3S) -3- methyl − 2- oxopentanoate to L - Isoleucine, 3-methyl-2- oxobutanoate to L-valine and L-leucine to 2-Oxoisocaproate, we detected *VTE3*, *EPSPS-1* and *BACT2* had high expression level in all of 10 *Chaenomeles* germplasms (Fig. 8, Supplementary Table S9).

Global identification of simple sequence repeats (SSRs)

Then we designed primers for genes containing SSR sites by primer3 following mentioned rules in methods, totally we obtained 373, 357, 362, 366, 355, 354, 352, 365, 370 and 372 primers in *C. sinensis* (GP), *C. cathayensis* (MY), *C. thibetica* (MYXZ), *C. japonica* (RB), *C. thibetica* (XZ), *C. thibetica* (XZSS),

C. speciosa (ZPCA), *C. speciosa* (ZPLP), *C. speciosa* (ZPLY) and *C. speciosa* (ZPXC) respectively (Supplementary Table S10A). The number of common genes used for primer design in 10 *Chaenomeles* germplasms was 110 (Supplementary Table S10B).

A total of 110 pairs of SSR primers were synthesized, and the 10 *Chaenomeles* germplasms were used to verify the PCR amplification ability. Based on this, a total of 45 pairs of primers were screened out (Supplementary Table S11). Then, capillary electrophoresis was used to detect the polymorphism and specificity of the 45 primers. The results showed that only 10 pairs of primers had good polymorphism. Primers from ORTHOMCL6247, ORTHOMCL8473, ORTHOMCL7263 ORTHOMCL4541, ORTHOMCL9476, ORTHOMCL9834 and ORTHOMCL8700 had the best polymorphism relatively (Supplementary Fig. S1). The primers from ORTHOMCL4735, ORTHOMCL8947 and ORTHOMCL6535 also had relatively good polymorphism, and poor amplification result for individual germplasm (Supplementary Fig. S1). Table 1 displayed all selected 10 primers information including tandem repeats unites, annealing temperature and detailed forward and reverse primers (Table 1).

Table 1
Primer and fragment length for selected genes

Othology ID	Gene symbol	Tandem repeats	Expected length (bp)	Annealing Temperature (°C)	Forward primer (5'->3')	Reverse primer (5'->3')
ORTHOMCL6247	*T19L5.4*	$(ACC)_8$	272	60	CTGAACAAACTCACCCCCAT	ATTGAACGCTTGGATAACCG
ORTHOMCL8425	*UGT89B2*	$(CTC)_7$	137	60	CCTCACTCACCAACTCGTCA	AGGTTTGGATGGGATTAGGG
ORTHOMCL8473	*DOF5.3*	$(CAC)_8$	123	60	GGACCATGGGCAATAACAAC	GAGAGCCATATGATCCGGG
ORTHOMCL7263	--	$(GGC)_5$	171	60	CCGTACACAAAACAAGCCCT	ATATTCCCGGAAACTGACCC
ORTHOMCL4541	*ORP3C*	$(AG)_{11}$	104	60	TCTTCCCTTTCATTTTCCGA	TCTGACCCTTCTCTGGGCTA
ORTHOMCL4735	*CRK25*	$(TCC)_5$	272	60	AAGGGGGACGAGTTCTGTTT	GTCTCACCACTCGGTTCGTT
ORTHOMCL9476	*SDAD1*	$(TGA)_5$	162	60	AAGATGACGGCAATGAGGAC	CATCTTCGCTTCCACTGTCA
ORTHOMCL9834	*KRI1*	$(GGAGAG)_4$	256	60	ATTTTGAATCGGACGACGAC	CCATCCTCCATCAAATGCTT
ORTHOMCL8700	*BBR*	$(AGC)_5$	184	60	GAGGACGAAGAAATTGGCAG	ATTTGCAGGTGTAGGGATGC
ORTHOMCL8947	*trmB*	$(CAGCAA)_4$	211	60	GCTGCATTACCCAGAAGAGC	GGTTGAGATTAAGGTCGGCA
ORTHOMCL6535	--	$(CT)_9$cactctctcc$(CT)_6$	112	60	TCACTTTGGTCCATGTCTGC	CGATATGTGTGTGCTCGGAC

Discussion

Several approaches including morphological markers, cytological markers, and biochemical markers have been explored to classify plant resources [22]. However, each technique is imperfect. For example, great phenotypic difference individuals from geographical isolation create obstacle for taxonomists. Subjective factors of taxonomists lead to misclassification. Limited number of isoenzyme and low resolution of cytology prevent taxonomists from recognizing the plants. Due to the flaws, molecular markers have become a prevalent way for evolutionary studies in *Chaenomeles* and the related species [34–37]. Various molecular markers providing reliable and convictive evidence for related researches promote phytowaxonomy research, whereas, some technologies are still unaffordable because of high cost. Nowadays, RNA-seq has become a economical and efficient tool for tapping molecular information.

The phylogeny and classification of *Chaenomeles* have been controversial for a long time, especially whether *C. sinensis* is a legally effective species. Robertson et al [38] conducted phylogenetic relationships among 88 genera of Rosaceae, and the results show it is far from *Chaenomeles* group, indicating that *C. sinensis* is not the member of *Chaenomeles*. In our research, we found *C. sinensis* had closer evolutionary relationship with *D. delavayi* confirming Robertson's result. There was obvious differentiation between *C. sinensis* (GP) and all other species of *Chaenomeles*. In terms of phenotype, *C. sinensis* (GP) with the characters of branches unarmed, flowers solitary, coetaneous, sepals reflexed, stipules ovate-lanceolate and margin glandular serrate is also significant different from the other *Chaenomeles* species which have the characters of branches armed, flowers fascicled, precocious or coetaneous, sepals erect, rarely reflexed, stipules herbaceous, reinform or auriculate and margin serrate[23]. Pollen morphology also shows that it is distant from other species [39]. Therefore, based on the phylogenetic results and phenotypic characteristics, we support that *C. sinensis* should not belong to *Chaenomeles*, but should be an independent genus *Pseudocydonia*. This view was also supported by our another study on the phylogeny of *Chaenomeles* based on chloroplast genome sequencing [40].

The genetic characteristics of populations should be dictated by the interplay of genetic drift, gene flow and natural selection. These processes may be strongly influenced by the demography and spatial distribution of populations [41]. However, their extrinsic feature have some disparities, even main

chemical components also appears differences, this lead genuine regional drug in doctor of traditional Chinese medicine. From molecular marker perspective, although above same *Chaenomeles*, they have formed gene-diversity offsprings leading to different bands (Fig. 8), this phenomenon has been reported in Bartish's research, which discriminate 42 *Chaenomeles* by RAPD method [42].

For closely related taxa, the advantages of homologous single copy genes for phylogenetic and phylogeographic analysis are clear because of their rapid evolutionary rates and clear avoidance of paralogy [43–46]. According past researches, there are a lot of single copy genes identified in many plants, including *Euasterids*, *Rosaceae*, *Poaceae*, *Cycads* and other 29 angiosperms [47–51]. For example, Duarte et al. totally identified a series of 959 single-copy genes in *A.* thaliana, *P.* trichocarpa, *V.* vinifera and *O.* sativa, and used 18 single-copy genes to improve resolution of *Brassicaceae* phylogeny [43]. In this study, we identified 6154 single-copy genes according to the strict filtering criterion. All these detected single-copy genes proved high effectiveness in phylogeny construction of *Chaenomeles* species, which provided a rough phylogenetic framework for the whole genus (Fig. 4). Therefore, this result indicated that all these candidate single copy genes from the transcriptomes of *Chaenomeles* species have good application in molecular phylogeny studies of *Chaenomeles* genus.

In our result, 10 out of 45 SSR markers were finally confirmed to distinguish the 10 *Chaenomeles* germplasms. They have good appearance in both polymorphism and amplification. The efficiency of primer developing rate was 22.2% is a little lower than Wang et al's 28.6% who develop SSR marker in Pineaaple. For the developing rate, many factors can influence the result, such as the selected base number of primers, genetic diversities of tested species, and developing methods. For instance, Wang [52] compared four methods including ISSR, SSR, SCoT and RAPD by using 47 germplasms of *Ananas comosus* (L.) Merr. Among them, the developing rate of RAPD can reach 47.06% and significantly higher than rest methods. However, SCo T and SSR method have good comprehensive performance comparing two other methods. Taken together, our results have proved SSR marker developing based on transcriptome is promising.

Although classification studies in *Chaenomeles* have been made in great progress, system of nomination and classification still existing some problem. The results of this study support the idea of *C. sinensis* as an independent genus, *Pseudochaenomeles*. With increasing exploration of edible and medicinal value, it is important to distinguish *Chaenomeles* assortment and avoid misusing. So the authoritative criterion is urgent. Due to delay of *Chaenomeles* genome, it's still hard to explored SSR markers efficiently. Because of the low cost for transcriptome sequencing, it will be optimum selection, thus, our research provides new methodology and reference for future related research in *Chaemomeles*. The large genetic diversity in the genus *Chaenomeles* as inferred from molecular markers is also useful for improvement, breeding and selection project in related species.

Methods

Plant material, total RNA extraction and Illumina sequencing

All five species of *Chaenomeles* and one related species (*D. delavayi*) were collected (Fig. 1). *C. sinensis* (GP) was collected in Anji County of Zhejiang Provice. According to the local common name, *C. speciosa* was divided into four types, including *C. speciosa* 'Chunmugua' (ZPCA), *C. speciosa* 'Ziqiumugua' (ZPLP), *C. speciosa* 'Xuanmugua' (ZPXC) and *C. speciosa* 'Yizhoumugua' (ZPLY). Three local types of *C. thibetica* were also collected, which distribute in Lasha (XZ), Bomi (XZSS) and Motuo (MYXZ) of Xizang Province respectively. *C. cathayensis* (MY) was collected from Shiping County, Yunnan Province. *C. japonica* (RB) was introduced from Japan and collected from the Germplasm Resource Bank of *Chaenomeles* in Anji County of Zhejiang Province. *D. delavayi* (DY) was collected from Lancang County of Yunnan Province. The voucher specimens were deposited in the Herbarium of Research of Institute of Subtropical Forestry, Chinese Academy of Forestry with specimens ID showed in Fig. 1.

A total of 15 duplicates leaves were collected from each germplasm resource. According to the manufacturer's instruction, total RNA was isolated from mature leaf samples as described in Owczarek et al [53]. After testing the qualification of extracted RAN including concentration, RIN value, ratio of 28S to 18S, OD_{260} to OD_{280} by using Agilent 2100 Bioanalyzer and NanoDrop, then the RNAs from same germplasm were pooled to prepare cDNA library by using cDNA Synthesis Kit (Illumina Inc., San Diego, CA) following the manufacturer's recommendations [54]. Finally, the constructed cDNA libraries were sequenced using Illumina HiSeq 2500 by paired-end 150 bp strategy.

Filtering and assembling of reads, quantitative calculation and functional annotation for unigenes

Firstly, raw reads were filtered with fastp software by removing adaptor and low quality reads [55]. Then, the cleaned reads were de novo assembled using Trinity with the default parameters for each individual sample [56]. Expression level of unigenes was calculated by RSEM [57] software, and was standardized by RPKM. Unigenes were queried against the non-redundant protein (Nr) database, the Swiss-Prot protein database, the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database and the Cluster of Orthologous Groups (COG) database using BLASTx with an *E*-value cut-off of 1e-5 [58].. According to the blast results, the coding region sequences (CDS) of unigenes were extracted and translated into peptides. For the unigenes without matching with the above four databases, we predicted the coding region sequences by using TransDecoder. Gene Ontology (GO) annotation of the unigenes were obtained using Blast2GO [59]. Enrichment analyses in GO term and KEGG pathway were conducted as described in Li et al [60].

Identification of orthologs, single-copy genes and multiple-copy genes

Diamond[61] and OrthoMCL[62] softwares were used to identify the orthologs among the 11 germplasms. The first step, multiple sequences comparisons among different germplasms were conducted with cutoff of *E*-value ≤ 1e-5, query cover ≥ 30% by using Diamond. The second step, the orthologs were merged into different families by using OrthoMCL. These obtained orthologous genes can be divided into single-copy genes and multiple-copy genes according to their copy numbers.

Phylogenetic analysis and selective pressure analysis

Based on the above analyses, we firstly constructed sequence alignment by using MUSCLE (multiple sequence alignment with high accuracy and high throughput) for single-copy genes. Then, the phylogenetic tree was performed in iqtree software [63] by Maximum likelihood method [64]. To detect whether genes were affected by nature condition, paml-codeml was used to compute non synonymous substitution rate (Ka), synonymous substitution rate (Ks) and ratio Ka to Ks. We aligned orthologous pairs [65]. Orthologs genes with Ka/Ks > 1, which were considered as genes under strong positive selection; orthologous genes with 0.5 < Ka/Ks < 0.1, which were considered as affected by weak positive selection; orthologous genes with Ka/Ks < 0.1, which were considered as affected by negative selection. The significance tests were measured with *p* values, which were corrected via false discovery rate (FDR).

Expression model of single-copy genes in secondary metabolites pathway

After performing KEGG analysis, the significant enriched pathway-secondary metabolites pathway was selected for detecting gene expression model among the 10 *Chaenomeles* germplasms. The pathway was depicted by Adobe Illustrator CS5 software [66], and heatmaps were visualized by TBtools 0.6669 [67].

Identification and screening of SSRs among 10 *Chaenomeles* germplasms

We identified SSRs motifs in all unigenes using MISA software [68]. According to the tandem repeat including AAC, ACA, CAA, GTT, TGT and TTG, all these SSRs motifs were divided into mono-, di-, tri-, tetra-, penta- and hexanucleotide type. PCR primers were designed for the SSRs using the program Primer3, which have more than 150 bp flanking sequences [69]. All of the primers used for SSR development were showed in Supplementary Table S11.

DNA was extracted from leaves of the 11 germplasms as described in Bartish et al [70]. Fluorescently labeled primers were synthesized for amplifying gene fragment as following rules: the primer should have 2, 3, 4, or 5 tandem repeat units; the length of PCR products range from 150bp to 300 bp; the position of genes should not focus on one site; the polymorphic sites should be chosen firstly; different combination of tandem repeat units should be chosen averagely; the repeat base in primer should be less than four; length of primers should be located at about 20 ~ 23bp; TM value of primers should be at 60 °C; consecutive A or T at 5′ or 3′ region of primer should be less than two; repeat sequence within primer was forbidden. The PCR products were detected by capillary electrophoresis.

Declarations

Acknowledgements

This research was funded by the Fundamental Research Funds for the Central Non-profit Research Institution of Chinese Academy of Forestry （CAFYBB2018SY016）. We thank Guangzhou Genedenovo Biotechnology for assisting in sequencing and bioinformatics analysis.

Author contribution

W.H.S.、H.L. and J.M.J. conceived and designed the study; W.H.S., S.Q.H. and Y.Z.Z. collected the plant germplasms; W.H.S. and H.L. performed bioinformatics analyses and data visualizing; H.L. and W.H.S. wrote the manuscript; all of the authors have revised and approved this manuscript.

Additional Information

All sequencing data was deposited on NCBI (https://www.ncbi. nlm.nih.gov/) at BioProject ID: PRJNA718952. Supplementary dataset 1-11 are available online.

Competing financial interests: The authors declare no competing financial interests.

Ethics approval and consent to participate：The 11 plant materials in this research were collected with the permission of the local authorities. The research on plants and collection of plant material are complied with institutional, national, or international guidelines. We also comply with IUCN Policy Statement on Research Involving Species at Risk of Extinction and the Convention on the Trade in Endangered Species of Wild Fauna and Flora.

References

1 I. *et al*. Phylogenetic relationships and differentiation among and within populations of *Chaenomeles* Lindl. (*Rosaceae*) estimated with RAPDs and isozymes. *TAG Theoretical&Applied Genetics* (2000).

2 Song-jie, Y. Research Advances on Plant Germplasm Resources of *Chaenomeles* [J]. *Hubei Agricultural Sciences* **20** (2011).

3 Strek, M., Gorlach, S., Podsedek, A., Sosnowska, D. & Hrabec, E. Procyanidin Oligomers from Japanese Quince (*Chaenomeles japonica*) Fruit Inhibit Activity of MMP-2 and MMP-9 Metalloproteinases. *Journal of Agricultural and Food Chemistry* **55**, 6447-6452 (2007).

4 Chen, R.-l., Wu, T.-j. & Dai, Y.-j. Studies on the Chemical Constiuents of Four Species of *Chaenomeles*. *West China Journal of Pharmaceutical Sciences* **15**, 39-39 (2000).

5 Koehne, E. Gattungen der Pomaceen. (1890).

6 Morley, B. D. Augustine Henry: his botanical activities in China, 1882-1890. *Glasra* **3**, 21-81 (1979).

7 Galan, R. & Palmer, J. The occurrence of the rare Ciboria aestivalis in Europe. *CZECH MYCOLOGY* **52**, 277-288 (2000).

8 Potter, D. *et al.* Phylogeny and classification of Rosaceae. *Plant systematics and evolution* **266**, 5-43 (2007).

9 Phipps, J. Mespilus canescens, a new Rosaceous endemic from Arkansas. *Systematic botany*, 26-32 (1990).

10 Phipps, J. B., Robertson, K. R., Rohrer, J. R. & Smith, P. G. Origins and Evolution of Subfam. Maloideae (Rosaceae). *Systematic Botany* **16**, 303 (1991).

11 Rumpunen, K., Bartish, I., Garkavagustavsson, L. & Nybom, H. Molecular and morphological diversity in the plant genus *Chaenomeles*. (2003).

12 da Silva, J. A. T. *et al.* Santalum molecular biology: molecular markers for genetic diversity, phylogenetics and taxonomy, and genetic transformation. *Agroforestry Systems* **92**, 1301-1315 (2018).

13 Chrungoo, N. *et al.* Establishing taxonomic identity and selecting genetically diverse populations for conservation of threatened plants using molecular markers. *Current Science* **114**, 539 (2018).

14 Sharma, V. & Salwal, R. in *Molecular Markers in Mycology* 37-52 (Springer, 2017).

15 Bartish, I. V., Rumpunen, K. & Nybom, H. Combined analyses of RAPDs, cpDNA and morphology demonstrate spontaneous hybridization in the plant genus *Chaenomeles*. *Heredity* **85**, 383-392 (2010).

16 He, J. *et al.* Genetic variability of cultivated *Chaenomeles speciosa* (Sweet) Nakai based on AFLP analysis. *Biochemical Systematics & Ecology* **57**, 445-450 (2014).

17 Yan-Yan, Z. *et al.* Analysis of Genetic Diversity in *Chaenomeles* Using Apple EST-SSRs. *Biotechnology Bulletin* (2016).

18 Julio, E. *et al.* RNA-Seq analysis of Orobanche resistance in Nicotiana tabacum: development of molecular markers for breeding recessive tolerance from 'Wika'tobacco variety. *Euphytica* **216**, 6 (2020).

19 Thakur, O. & Randhawa, G. S. Identification and characterization of SSR, SNP and InDel molecular markers from RNA-Seq data of guar (Cyamopsis tetragonoloba, L. Taub.) roots. *BMC genomics* **19**, 951 (2018).

20 Wu, N. *et al.* RNA-seq facilitates development of chromosome-specific markers and transfer of rye chromatin to wheat. *Molecular breeding* **38**, 6 (2018).

21 Li, H., Ruan, C.-J., Wang, L., Ding, J. & Tian, X.-J. Development of RNA-Seq SSR Markers and Application to Genetic Relationship Analysis among Sea Buckthorn Germplasm. *Journal of the American Society for Horticultural Science* **142**, 200-208 (2017).

22 Chen, H. (Master degree dissertation, Shandong Agricultural University, Tai'an.(in …, 2008).

23 Arnold, J. & Zhuge, R. Flora of China. (2007).

24 Carbone, I., Ramirez-Prado, J. H., Jakobek, J. L. & Horn, B. W. Gene duplication, modularity and adaptation in the evolution of the aflatoxin gene cluster. *BMC Evolutionary Biology* **7**, 1-12 (2007).

25 Hurst, L. D. The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends in genetics: TIG* **18**, 486-486 (2002).

26 Carbone, I., Jakobek, J. L., RAMIREZ-PRADO, J. H. & Horn, B. W. Recombination, balancing selection and adaptive evolution in the aflatoxin gene cluster of Aspergillus parasiticus. *Molecular Ecology* **16** (2007).

27 Fukazawa, J. *et al.* DELLAs function as coactivators of GAI-ASSOCIATED FACTOR1 in regulation of gibberellin homeostasis and signaling in Arabidopsis. *The Plant cell* **26**, 2920-2938 (2014).

28 Fujikura, U., Horiguchi, G., Ponce, M. R., Micol, J. L. & Tsukaya, H. Coordination of cell proliferation and cell expansion mediated by ribosome-related processes in the leaves of Arabidopsis thaliana. *Plant Journal* **59**, 499-508 (2010).

29 Bertini, L. *et al.* Modular structure of HEL protein from Arabidopsis reveals new potential functions for PR-4 proteins. *Biological Chemistry* **393**, 1533-1546 (2012).

30 Geraldo, N., Bäurle, I., Kidou, S.-i., Hu, X. & Dean, C. FRIGIDA delays flowering in Arabidopsis via a cotranscriptional mechanism involving direct interaction with the nuclear cap-binding complex. *Plant Physiology* **150**, 1611-1618 (2009).

31 Faigon-Soverna, A. *et al.* A Constitutive Shade-Avoidance Mutant Implicates TIR-NBS-LRR Proteins in Arabidopsis Photomorphogenic Development. *The Plant cell* **18**, 2919-2928 (2006).

32 Kranz, H. D. *et al.* Towards functional characterisation of the members of theR2R3-MYBgene family fromArabidopsis thaliana. *Plant Journal* **16**, 263-276 (2010).

33 Moreno, A. A. *et al.* IRE1/bZIP60-Mediated Unfolded Protein Response Plays Distinct Roles in Plant Immunity and Abiotic Stress Responses. *PloS one* **7**, e31944 (2012).

34 Yu, Y. *et al.* Genome structure of cotton revealed by a genome-wide SSR genetic map constructed from a BC 1 population between Gossypium hirsutum and G. barbadense. *BMC genomics* **12**, 15 (2011).

35 Nie, X. *et al.* Genome-wide SSR-based association mapping for fiber quality in nation-wide upland cotton inbreed cultivars in China. *BMC genomics* **17**, 352 (2016).

36 Liu, Q. *et al.* Genetic diversity and population structure of pear (Pyrus spp.) collections revealed by a set of core genome-wide SSR markers. *Tree genetics* **11**, 128 (2015).

37 Khan, M. K. *et al.* Genome wide SSR high density genetic map construction from an interspecific cross of Gossypium hirsutum× Gossypium tomentosum. *Frontiers in plant science* **7**, 436 (2016).

38 Robertson, K. R., Phipps, J. B. & Smith, R. A Synopsis of Genera in Maloideae (Rosaceae). *Systematic Botany* **16**, 376 (1991).

39 Lei, Z. L. C. H. Z. & Dekui, Z. Pollen Morphology and Cultivar Classification of the Genus *Chaenomeles* [J]. *Scientia Silvae Sinicae* **5** (2008).

40 Shao, W. & Jiang, J. The complete chloroplast genome sequences of two *Chaenomeles* species (*Chaenomeles cathayensis* and *Chaenomeles thibetica*). *Mitochondrial DNA Part B* **5**, 3191-3192 (2020).

41 Eckert, C. G., Samis, K. E. & Lougheed, S. C. Genetic variation across species' geographical ranges: the central-marginal hypothesis and beyond. *Molecular Ecology* **17**, 1170-1188 (2010).

42 Bartish, I. V., Rumpunen, K. & Nybom, H. Genetic diversity in *Chaenomeles* (*Rosaceae*) revealed by RAPD analysis. *Plant Systematics & Evolution* **214**, 131-145 (1999).

43 Duarte, J. M. *et al.* Identification of shared single copy nuclear genes in Arabidopsis, Populus, Vitis and Oryzaand their phylogenetic utility across various taxonomic levels. *BMC Evolutionary Biology* **10**, 61 (2010).

44 Feau, N., Decourcelle, T., Husson, C., Desprez-Loustau, M.-L. & Dutech, C. Finding single copy genes out of sequenced genomes for multilocus phylogenetics in non-model fungi. *PLoS One* **6** (2011).

45 Li, Z. *et al.* Single-copy genes as molecular markers for phylogenomic studies in seed plants. *Genome Biology Evolution* **9**, 1130-1147 (2017).

46 Teasdale, L. C., Köhler, F., Murray, K. D., O'hara, T. & Moussalli, A. Identification and qualification of 500 nuclear, single-copy, orthologous genes for the Eupulmonata (Gastropoda) using transcriptome sequencing and exon capture. *Molecular ecology resources* **16**, 1107-1123 (2016).

47 Wu, F., Mueller, L. A., Crouzillat, D., Pétiard, V. & Tanksley, S. D. Combining bioinformatics and phylogenetics to identify large sets of single-copy orthologous genes (COSII) for comparative, evolutionary and systematic studies: a test case in the euasterid plant clade. *Genetics* **174**, 1407-1420 (2006).

48 Cabrera, A. *et al.* Development and bin mapping of a Rosaceae Conserved Ortholog Set (COS) of markers. *BMC Genomics* **10**, 562 (2009).

49 Fan, X. *et al.* Phylogeny and evolutionary history of Leymus (Triticeae; Poaceae) based on a single-copy nuclear gene encoding plastid acetyl-CoA carboxylase. *BMC Evolutionary Biology* **9**, 247 (2009).

50 Salas-Leiva, D. E. *et al.* Phylogeny of the cycads based on multiple single-copy nuclear genes: congruence of concatenated parsimony, likelihood and species tree inference methods. *Annals of Botany* **112**, 1263-1278 (2013).

51 Han, F., Peng, Y., Xu, L. & Xiao, P. Identification, characterization, and utilization of single copy genes in 29 angiosperm genomes. *BMC genomics* **15**, 504 (2014).

52 Wang, J. S., Jun-Hu, H. E., Chen, H. R., Chen, Y. Y. & University, P. Comparison on the Detection Efficiency of Different Types of Molecular Markers in Pineapple. *Hubei Agricultural Sciences* (2015).

53 Owczarek, K. *et al.* Flavanols from Japanese quince (*Chaenomeles japonica*) fruit suppress expression of cyclooxygenase-2, metalloproteinase-9, and nuclear factor-kappaB in human colon cancer cells. *Acta Biochimica Polonica* **64**, 567–576-567–576 (2017).

54 Zhang, M., Mo, H., Sun, W., Guo, Y. & Li, J. Systematic isolation and characterization of cadmium tolerant genes in tobacco: A cDNA library construction and screening approach. *PLoS One* **11**, e0161147 (2016).

55 Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, 884-890 (2018).

56 Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology* **29**, 644 (2011).

57 Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics* **12**, 1-16 (2011).

58 Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Research* **12**, 656-664 (2002).

59 Conesa, A. *et al.* Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674-3676 (2005).

60 Li, H. *et al.* MicroRNA comparison between poplar and larch provides insight into the different mechanism of wood formation. *Plant Cell Reports* **39**, 1199-1217 (2020).

61 Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nature methods* **12**, 59-60 (2015).

62 Li, L., Stoeckert, C. J. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome research* **13**, 2178-2189 (2003).

63 Nguyen, L.-T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution* **32**, 268-274 (2015).

64 Retief, J. D. Phylogenetic analysis using PHYLIP. *Methods Mol Biol* **132**, 243-258 (2000).

65 Wang, D., Zhang, Y., Zhang, Z., Zhu, J. & Yu, J. KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genomics, proteomics bioinformatics* **8**, 77-80 (2010).

66 Team, A. C. *Adobe Illustrator CS5 Classroom in a Book: ADOBE ILLUST CS5 CLASSROOM_p1*. (Pearson Education, 2010).

67 Chen, C., Chen, H., He, Y. & Xia, R. TBtools, a toolkit for biologists integrating various biological data handling tools with a user-friendly interface. *BioRxiv*, 289660 (2018).

68 Beier, S., Thiel, T., Münch, T., Scholz, U. & Mascher, M. MISA-web: a web server for microsatellite prediction. *Bioinformatics* **33**, 2583-2585 (2017).

69 Untergasser, A. *et al.* Primer3—new capabilities and interfaces. *Nucleic acids research* **40**, e115-e115 (2012).

70 Bartish, I., Garkava, L., Rumpunen, K. & Nybom, H. Phylogenetic relationships and differentiation among and within populations of *Chaenomeles* Lindl.(*Rosaceae*) estimated with RAPDs and isozymes. *Theoretical and Applied Genetics* **101**, 554-563 (2000).

Figures

Figure 1

Distribution of 11 germplasms used for the research in China. Different symbols represent different germplasms. Each germplasm is labled with specimens ID as follow: C. sinensis (GP) (SWH190512); C. speciosa (ZPLY) (SWH190535); C. speciosa (ZPXC) (SWH190621); C. speciosa (ZPCA) (SWH190657); C. speciosa (ZPLP) (SWH190645); C. thibetica (XZ) (SWH190805); C. thibetica (XZSS) (SWH190823); C. thibetica (MYXZ) (SWH190837); C. cathayensis (MY) (SWH190711); C. japonica (RB) (SWH190527); D. delavayi (DY) (SWH190726). All of the germplasms were identified by Professor J.M.J..Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

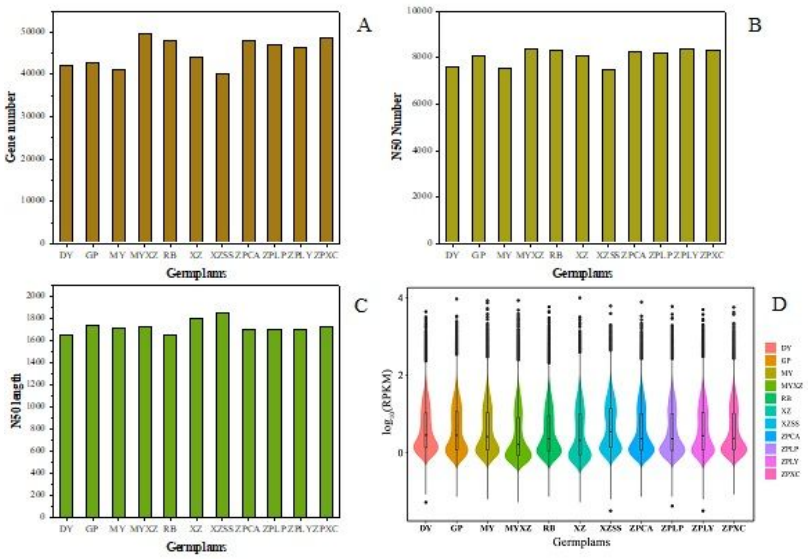

Figure 2

Assembling results and gene expression distribution in 10 germplasms of Chaenomeles and Docynia delavayi (D. delavay). In this graph, A is gene numbers in 11 materials; B is N50 numbers in 11 materials; C is N50 length in 11 materials; D is violin graph for gene expression distribution in 11 materials. In X-axis, DY represents D. delavayi (DY); GP represents C.sinensis (GP); MY represents C. cathayensis （MY）; XZ represents C. thibetica（XZ）; MYXZ represents C. thibetica (MYXZ); XZSS represents C. thibetica (XZSS); RB represents C. japonica (RB); ZPCA represents C. speciosa (ZPCA); ZPLP represents C. speciosa (ZPLP); ZPLY represents C. speciosa (ZPLY); ZPXC represents C.speciosa (ZPXC). In violin graph, different colors represent different species. Area and wide of violin graph represent expression abundance of expressed genes in corresponding species. Expression level in vertical coordinate are standalidized by log10 (RPKM) algorithm.

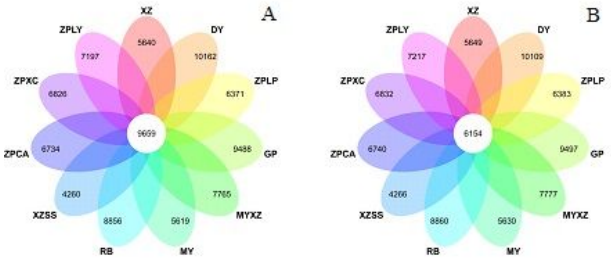

Figure 3

Venn graphs for selection of common genes and common single copy genes among 10 germplasms of Chaenomeles and D. delavay. A depcits selection of common genes in 11 materials, B depcits selection of common single copy genes in 11 materials. Only Common genes among 11 materials and specific genes in each germplasm are showed in this graph.

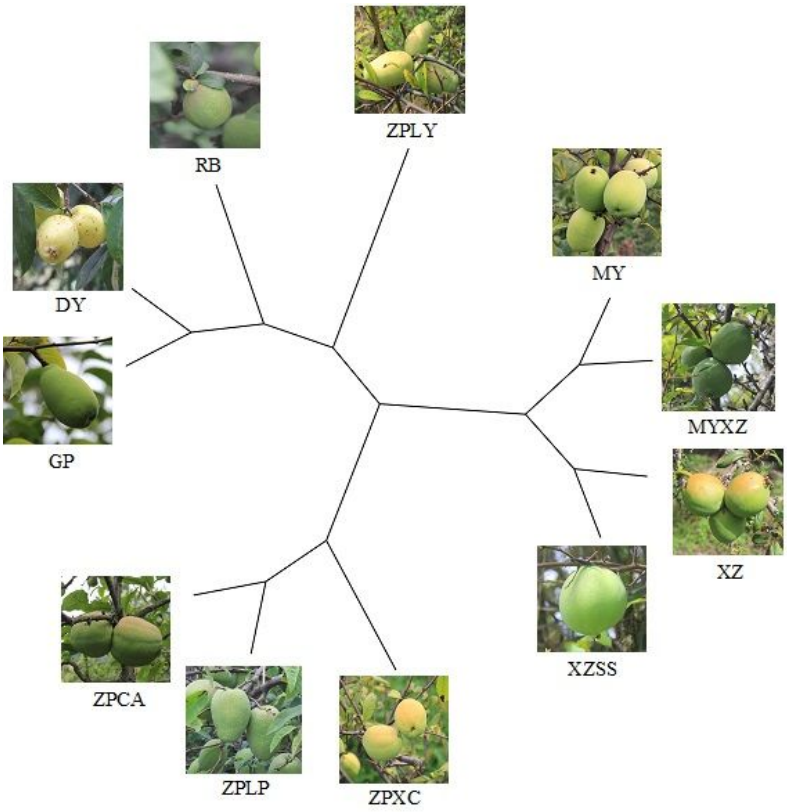

Figure 4

Phylogenetic tree based on common single copy genes for 10 germplasms of Chaenomeles and D. delavay. A consensus phylogenetic tree, created from individual trees originating from OrthoMCL groups with one representative per germplasm. The numbers of branches indicate number of the germplasm have been partitioned into three sets.

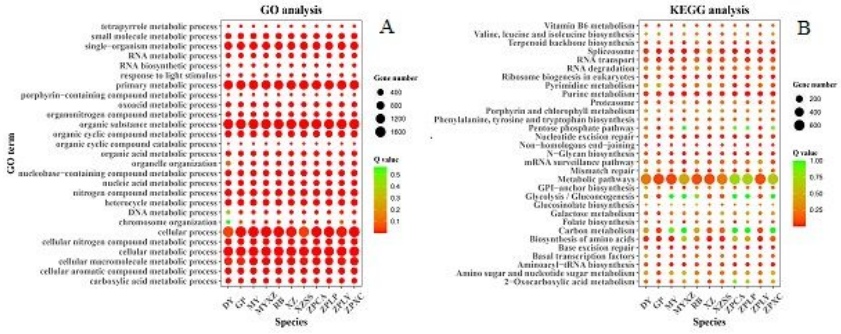

Figure 5

GO and KEGG analyses for common single-copy genes among 10 germplasms of Chaenomeles and D. delavay. In this graph, pot size means enriched gene numbers; from red to green, the degree of significance decreases. In both of GO and KEGG analysis result, only top15 GO terms and pathways are showed.

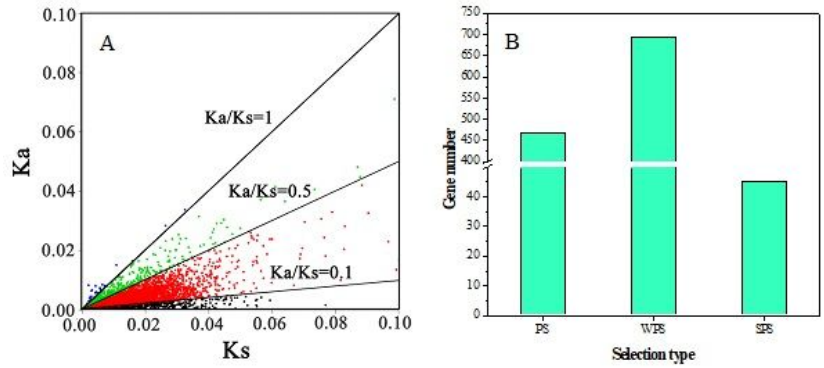

Figure 6

Selective pressure analyses for single-copy genes in common genes of 10 germplasms of Chaenomeles and D. delavay. In graph A, each color dot represents one single-copy gene family. Diagonal cater corner represents Ka/Ks=1; the first dot line represents Ka/Ks=0.5; the second dot line represents Ka/Ks=0.1. In graph B, PS is short name of purification selection; WPS is short name of weak positive selection; SPS is short name of strong positive selection.

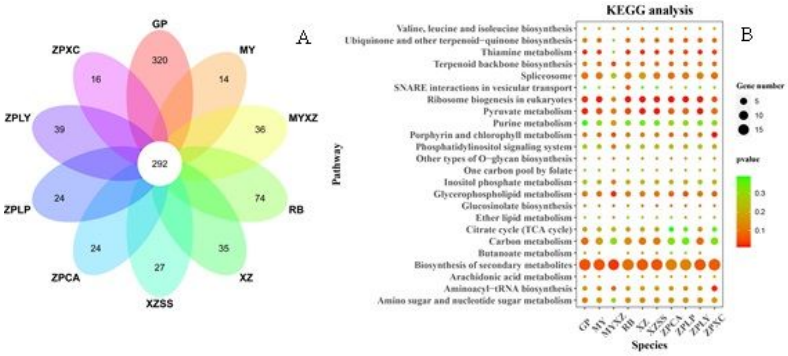

Figure 7

Venn graph and KEGG analysis for common single copy genes in 10 germplasms of Chaenomeles containing SSR sites. In graph B, pot size means enriched gene numbers; from red to green, the degree of significance decreases.

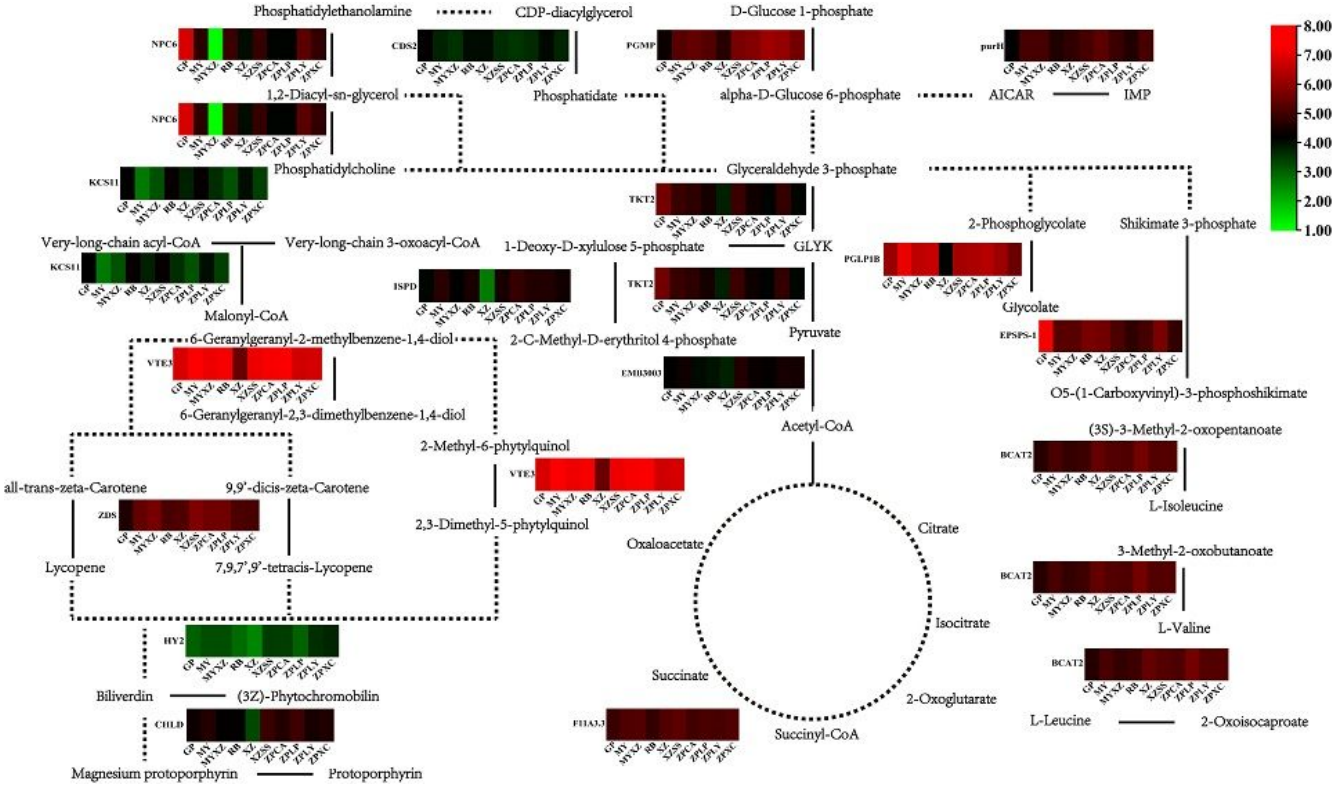

Figure 8

Gene expression models in biosynthesis of secondary metabolites pathway. This graph display part of biosynthesis of secondary metabolite. Dot lines describe indirect metabolism flow; solid lines descript direct metabolism flow. Color bar describes gene expression level in corresponding position. RPKMs in heatmaps are standardized by log2 (RPKM) algorithm, Green means low expression level, red means high expression level.

Supplementary Files

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

- SupplementaryFigure.docx
- SupplementaryTableS1.xlsx

- SupplementaryTableS2.xlsx
- SupplementaryTableS3.xlsx
- SupplementaryTableS4.xlsx
- SupplementaryTableS5.xlsx
- SupplementaryTableS6.xlsx
- SupplementaryTableS7.xlsx
- SupplementaryTableS8.xlsx
- SupplementaryTableS9.xlsx
- SupplementaryTableS10.xlsx
- SupplementaryTableS11.xlsx