An Integrated Approach to Identify Cytochrome P450 Superfamilies in Plant Species within the Malvids

Taikui Zhang
[1]Co-Innovation Center for Sustainable Forestry in Southern China, Nanjing Forestry University
[2]College of Forestry, Nanjing Forestry University
taikuizhang@126.com

Cuiyu Liu
[1]Co-Innovation Center for Sustainable Forestry in Southern China, Nanjing Forestry University
[2]College of Forestry, Nanjing Forestry University
ankar_liu@163.com

Hanyao Zhang
[3]Key Laboratory for Forest Genetic and Tree Improvement and Propagation in Universities of Yunnan Province, Southwest Forestry University
hanyaoz@163.com

Zhaohe Yuan[*]
[1]Co-Innovation Center for Sustainable Forestry in Southern China, Nanjing Forestry University
[2]College of Forestry, Nanjing Forestry University
zhyuan88@hotmail.com

ABSTRACT

The cytochrome P450 (CYP) monooxygenases present an enzyme superfamily contributing diverse biological functions and thus are crucial for survival in the plant. Although the CYP members in several plant species have been clarified, little is known about the evolutionary relationships between species within the Malvids clade, that includes cabbage (*Brassica rapa*), cacao (*Theobroma cacao*), cotton (*Gossypium arboreum*), eucalypts (*Eucalyptus grandis*), orange (*Citrus sinensis*), papaya (*Carica papaya*), pomegranate (*Punica granatum*), and other key species widely cultivated in agricultural. Here we *de nove* assembled the high-quality pomegranate transcripts with N50 of 1791bp using RNA-seq. Using an integrated HMM-search and InterProScan-verification as well as CYPED-annotation approach we identified 2350 putative CYP candidate proteins in pomegranate and the other related species within the Malvids clade. Four identified motifs were responsible for the conserved sequence structure and diverse enzyme functions. Phylogenetic analysis showed the distinct expansion of CYP families between species. Eight test species within the Malvids clade shared distinct CYP paralogs from their common ancestor. All paralogous pairs in each pomegranate CYP family showed a low ratio of Ka/Ks (<1), indicating that some replacement substitutions inner CYP family have been purified by natural selection, presumably because of their deleterious effects. Our findings provide a resource for the taxonomic and evolutionary study on plant CYP families.

CCS Concepts

• Applied computing~Computational genomics

Keywords

Cytochrome P450; comparative genomics; transcriptomics

1. INTRODUCTION

The cytochrome P450 (CYP) monooxygenases represent a large and important enzyme superfamily in plants[1]. Most plant CYPs are associated with the membranes of the endoplasmic reticulum and require an NADPH-dependent P450 reductase[2]. CYP enzymes catalyze a wide variety of monooxygenation/hydroxylation reactions in biochemical pathways, and defend organisms from endogenous and noxious environmental compounds[3]. CYP nomenclature system is based on a hierarchical clustering of genes into families and subfamilies: CYP families are named by number, the subfamilies by capital letters, and the specific genes by a second number. To date, 317 CYP families have been identified in the CYtochrome P450 Engineering Database (CYPED, https://cyped.biocatnet.de) according to the sequence similarity. Figments in flowers or fruits are crucial for attracting pollinators or gatherers to spread. CYP75A and CYP75B subfamilies play important roles in the biosynthesis of flavonoids and anthocyanins, both of which are major pigments[4]. Additionally, the CYP82 family genes specifically reside in dicots and are usually induced by distinct environmental stresses[1]. CYP monooxygenases contribute a broad array of biological functions in living organisms and thus are crucial for survival[5].

Plant CYP families belong to the E-Class[6] and were divided into ten clans[7]. Although the identification of CYP members in several plant species[1, 7-9] has been clarified, little is known about the evolutionary relationships between species. The increasing data of whole genome sequence offered an opportunity to address it. The Malvidis clade in Angiosperm Phylogeny Group (APG) IV system[10] is one of the main clades of Rosides, and contains numerous key species with the releasing of whole genome sequence, such as cabbage (*Brassica rapa*), papaya (*Carica papaya*), orange (*Citrus sinensis*), eucalypts (*Eucalyptus grandis*), cotton (*Gossypium arboreum*) and cacao (*Theobroma cacao*). Pomegranate (*Punica granatum*) is an ancient medicinal fruit tree and cultivated worldwide[11]. It is a member of the family Lythraceae, the clade Malvidis[12]. To study the CYP families of species within the Malvids clade can recover the relationships between the large members and their diverse enzyme functions.

Mining superfamilies in numerous plant genomes through traditional method is hard because hundreds of members need to verify on line one by one. HMM-search is a traditional approach

ICBCB '17, January 06-08, 2017, Hong Kong, Hong Kong
© 2017 ACM. ISBN 978-1-4503-4827-0/17/01$15.00
DOI: http://dx.doi.org/10.1145/3035012.3035021

[1,2] 159 Longpan Rd., Nanjing 210037, China.

[3] 300 Bailong Rd., Kunming 650224, China.

[*] Corresponding author. Tel: +86-25-85427056.

to mine based on HMM model[13]. InterProScan-verification search putative candidates against several databases such as SMART, PFAM, GENE3D and PRINTS[14]. For CYP superfamily, CYPED-annotation is a special approach based on the CYPED database. HMM-search and InterProScan-verification as well as CYPED-annotation approaches could be integrated to identify CYP superfamilies.

Here we *de nove* assembled the pomegranate transcriptome and identified putative CYP candidate proteins in pomegranate and the other related species within the Malvids clade, and analyzed characteristic motif in CYP proteins. Then the phylogenetic analysis on 972 CYP proteins was employed, and the distinct expansion of CYP families was observed. The selective evolution analysis on pomegranate CYP families was performed. Our work introduced an integrated approach to mine superfamilies in the plant genome and provided a resource for the taxonomic and evolutionary study on plant CYP families.

2. MATERIAL AND METHODS

2.1 Pomegranate Transcriptome

The peel of the mature fruit of pomegranate was collected, and total RNA was extracted using TRI Reagent (Sigma Life Science, USA) according to manufacturer's instructions. Paired-end RNA-Seq libraries were constructed and sequenced using an Illumina HiSeq 4000 platform according to the manufacture's protocol (Illumina, USA). The raw reads cleaned by using NGSQCToolkit v2.3.3[15] were *de nove* assembled through Velvet v1.2.10[16] and Oases v0.2.08[17]. Open reading frame (ORF) of the transcripts were further predicted by TransDecoder (http://transdecoder.github.io).

2.2 Genome Sequences

Genome sequences of plant species within the Malvids clade including cabbage, papaya, orange, eucalypts, cotton and cacao and outgroup species of peach (*Prunus persica*) were collected. Sequences of cabbage, papaya, and cotton were downloaded from PLAZ (http://bioinformatics.psb.ugent.be/plaza/), which of orange, eucalypts, cacao and peach were achieved from JGI (https://phytozome.jgi.doe.gov/). We made the perl scripts to filter out the asterisk (*) in protein sequences, and to remove the incomplete sequences such as containing poly N and X.

2.3 Identification of CYP Family

An integrated HMM-search and InterProScan-verification as well as CYPED-annotation approach was applied to identify the putative CYP families in the plant. The CYP family HMM model was built through HMMER v3.1[13] with the alignments downloaded from Pfam (http://pfam.xfam.org, Accession: PF00067). The model was used to predict putative CYP members in local protein set of pomegranate and other seven species with cutoff E-value of 1e-4. We made a program (SelectHMM, https://github.com/Redpome/SelectHMM) based on Perl language to extract large-scale candidate members from the result file of hmmsearch computing. All the putative candidate proteins were annotated and classified in SMART, PFAM, GENE3D and PRINTS databases by InterProScan v5.20[14]. The incredible sequences without CYP domains were deleted. Then, the filtered sequences were further blasted against the CYPED database using the blast program[18] with the cutoff E-value of 1e-100. These sequences annotated with CYP members were finally collected. Additionally, the program MEME (http://meme-suite.org/index.html) was used to predict conserved motifs outside of CYP domains. All putative motifs with expected values of >1E-100 were discarded.

2.4 Sequence Alignment and Phylogenetic Analysis

Sequence alignment was performed by MAFFT v7.305b[19]. Short sequences were excluded from the alignment to optimize our ability to recover meaningful overall phylogenetic patterns. Neighbor joining (NJ) trees were built by MEGA-CC v7.0[20] with 1000 bootstrap replicates. Maximum likelihood (ML) trees were built using MEGA-CC v7.0 with 1000 bootstrap replicates. Trees were visualized using the program ggtree[21].

2.5 Selective Evolution Analysis

Selective evolution analysis of pomegranate CYP superfamily was performed using the PAML program[22]. PgCYP protein alignments and the corresponding cDNA sequences were converted to codon alignments using PAL2NAL (http://www.bork.embl.de/pal2nal/) with auto-removing gaps, inframe stop codons, and mismatched codons between protein and DNA. Based on a rate of 6.161029 substitutions per site per year, divergence time (T) was calculated using the Ks value with the formula: $T = Ks/(2\times 6.1\times 10^{-9})\times 10^{-6}$ MYA[23].

3. RESULTS AND DISCUSSIONS

3.1 Unigenes of Pomegranate Transcripts

Assembling *de nove* pomegranate transcriptome by using Velvet and Oases with a kmer 31 yielded 76445 unigenes with the N50 of 1791 bp and the Maximum sequence length of 14926 bp. Further prediction using TransDecoder generated 49312 ORFs, about 51% (25314) of which belongs to the complete ORF type. Here we assembled a better pomegranate transcript, and N50 was more longer than that in 'Black' (709 bp) and 'Nana' (701 bp) as reported previously[24]. The transcripts are responded to the coding sequences of the whole genome. The pomegranate transcripts were enough to identify the CYP superfamily, although the pomegranate genome sequences were unknown.

3.2 Identification of CYP Family

Identifying superfamilies in several plant genomes using the traditional HMM-search method is a huge project, the integrated approach to mine superfamilies is more effectively. Plant genomes contain hundreds of CYP proteins that contribute to essential functions and species-specific metabolism[8]. More than one hundred putative CYP candidates were identified in each test species using the integrated identification approach (Table 1). Pomegranate transcriptome had 174 CYP members, the most of which were inferred to be CYP E-Class proteins (Table 1). Although the related plant species in the Malvidis clade had different CYP members, they had similar percent of CYP E-Class: cabbage (94.7 %), cacao (96.7 %), cotton (94.9 %), eucalypts (96.2 %), orange (94.3 %), papaya (99 %), peach (96.6 %) and pomegranate (96.0 %) (Table 1). Plant genomes of the species within the Malvids clade contained CYP superfamilies with more than 100 members and the majority of which are E-Class as reported previously[6].

To investigate the motif characteristic of CYP superfamilies in plant species within the Malvids clade, we performed the local alignment of the putative CYP candidates. In agreement with the CYP logo on PFAM (http://pfam.xfam.org/family/PF00067#tabview=tab4), the structure analysis showed a few well-conserved sequence regions (Figure 1). Four significant regions (a, b, c and d; $p < 0.01$) in CYP proteins suggested the high conserved motif (Figure 1), which approved of the plant CYP motifs as reported previously[5]. The most characteristic motif was FGXGXRXCXG (Figure 1), based on all test sites (972) with very low E-values (1e-6333). These identified motifs were responsible for the conserved sequence structure and diverse enzyme functions. Motifs in a, c and d regions (Figure 1) illustrated the CYP significant domains of oxygen activation, perf and heme binding respectively[8].

Table 1 CYP family in plant species within the Malvids clade

Species	CYP	B-class	E-class I	E-class IV
cabbage	342	4	307	17
cacao	273	0	251	13
cotton	350	1	322	10
eucalypts	521	5	488	13
orange	298	4	281	0
papaya	102	0	99	2
peach	290	1	275	5
pomegranate	174	2	163	4

CYP column contains all candidate proteins which are annotated and classified in IPR001128, IPR002397, IPR002401 and IPR002403. CYP members classified into IPR002397 are annotated with “CYP B-Class (B-class)”. These clustered in IPR002401 and IPR002403 are annotated with “CYP E-Class group I (E-class I)” and “CYP E-Class group IV (E-class IV)”, respectively.

According to the taxonomic criterion of the plant CYP superfamily[7], we classed the annotated proteins against CYPED into five clans and assessed the motifs of clans that contained at least 30 sequences. CYP71 Clan and CYP85 Clan contained three high conserved regions despite a considerable variation in sequence.

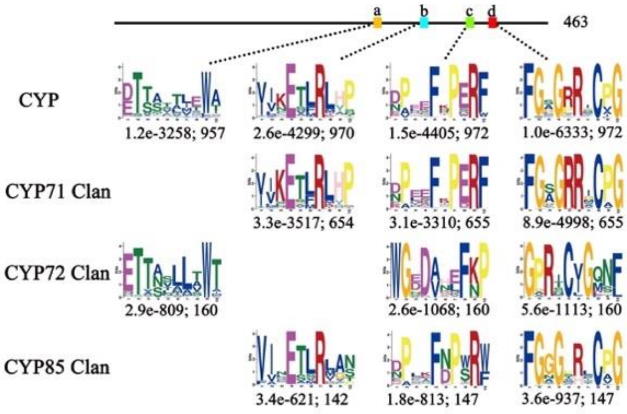


Figure 1 Sequence logos of the conserved CYP motifs.

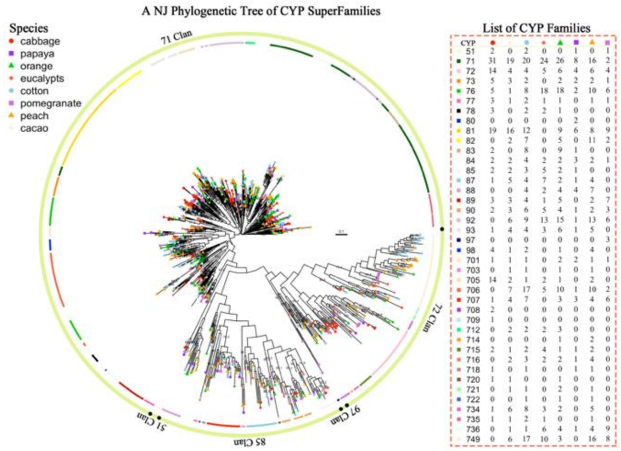
The CYP contains 972 CYP proteins from eight species. CYP71 Clan contains 655 CYP proteins which belong to CYP71, CYP73, CYP75, CYP76, CYP77, CYP78, CYP80, CYP81, CYP82, CYP83, CYP84, CYP89, CYP92, CYP93, CYP98, CYP701, CYP703, CYP705, CYP706, CYP712, and CYP736 families. CYP72 Clan contains 160 CYP proteins which belong to CYP72, CYP709, CYP714, CYP715, CYP721, CYP734, CYP735 and CYP749 families. CYP85 Clan contains 147 CYP proteins which belong to CYP85, CYP87, CYP88, CYP90, CYP707, CYP708, CYP716, CYP718, CYP720 and CYP724 families. Data below the logo indicate the statistical significance of the motif and the number of sites contributing to the construction of the motif.

3.3 Phylogenetic Analysis of CYP Families

Although up to 2350 putative CYP candidate proteins were identified in pomegranate and other related species within the Malvids clade, only 972 conserved CYP members were selected to reconstruct a phylogenetic tree. The huge difference inferred that CYP monooxygenases are super-families with numerous divergent families as reviewed in references[3, 5, 9].

Comparative genomics analysis showed that cabbage genome contained the most CYP71 members, and only two putative CYP71 genes were predicted in pomegranate (Figure 1). Coincidently, cabbage also had the majority of CYP72, CYP81 and CYP705 proteins in the conserved CYP alignments (Figure 1). Orange genome contained the most members of CYP83, CYP92 and CYP93. By contrast, pomegranate had the smaller members of the most CYP families. We concluded that cabbage CYP71, CYP72, CYP81 and CYP705 expanded in the Malvids clade. Orange genome obtained the expanded CYP families during the evolution in Malvids clade. Notably, eucalypts had the most CYP candidates (Table 1) but shared the low conserved CYP members (Figure 1). The distinct expansion of CYP families supported that the paralogs loss occurs following polyploid formation in eukaryotes[25-27]. Genome amplification through duplication has its counterpart in genome reduction, by the elimination of paralogs[27]. The eucalypts genome undergone a palaeotetraploidy event, superimposed on the earlier palaeohexaploidy event shared by all eudicots[26]. Thus the eucalypts genome lost substantially conserved CYP paralogs shared by common ancestor after ploidy event, compared to the other species.

To understand the relationship of 972 putative CYP candidates, an NJ phylogenetic tree was reconstructed. They were annotated in 40 families using the CYPED database (Figure 2). Based on the topologies and clade support values, candidates were classified into five clans including 51 Clan, 71 Clan, 72 Clan, 85 Clan, and 97 Clan (Figure 2). Our findings were consistent with the taxonomy of CYP clans in rice and Arabidopsis[7], although only five clans were observed here. The difference might be related to the more species investigated than the previous work[7]. In each clade of the phylogenetic tree, CYP members from each species tended to cluster together (Figure 2), demonstrating that CYP families expanded after divergence from their common ancestor.



CYP	●	●	●	◆	▲	■	▲	■
51	2	0	2	0	0	1	0	1
71	31	19	20	24	26	8	16	2
72	14	4	4	5	6	4	6	4
73	5	3	2	0	2	2	2	1
76	5	1	8	18	18	2	10	6
77	3	1	2	1	1	0	1	1
78	3	0	2	2	1	0	0	0
80	0	0	0	0	0	2	0	0
81	19	16	12	0	9	6	8	9
82	0	2	7	0	5	0	11	2
83	2	0	8	0	9	1	0	0
84	2	2	4	2	2	3	2	1
85	2	2	3	5	2	1	0	0
87	1	5	4	7	2	1	4	0
88	0	0	4	2	4	4	7	0
89	3	3	4	1	5	0	2	7
90	2	3	6	5	4	1	2	3
92	0	6	9	13	15	1	13	6
93	1	4	4	3	6	1	5	0
97	0	0	0	0	0	0	0	3
98	4	1	2	0	1	0	4	0
701	1	1	1	0	2	2	1	1
703	0	1	1	0	1	0	1	0
705	14	2	1	2	1	0	2	0
706	0	7	17	5	10	1	10	2
707	1	4	7	0	3	3	4	6
708	2	0	0	0	0	0	0	0
709	1	0	0	0	0	0	0	0
712	0	2	2	2	3	0	0	0
714	0	0	0	0	1	0	2	0
715	2	1	2	4	1	1	2	0
716	0	2	3	2	2	1	4	0
718	1	0	1	0	0	1	1	0
720	1	1	0	0	1	0	0	0
721	0	1	1	0	2	0	1	0
722	0	0	1	0	1	0	1	0
734	1	6	8	3	2	0	5	0
735	1	1	2	1	0	0	1	0
736	0	1	1	6	4	1	4	9
749	0	6	17	10	3	0	16	8

Figure 2 The phylogenetic tree and classification of 972 CYP genes.

The inner circle is the neighbor-joining (NJ) tree including 972 CYP proteins from 8 species within the Malvids clade. Eight distinct shapes with different colors show their taxonomic groups, as indicated in the legend. The middle circle is the corresponding CYPs, which are covered by forty different colors to show their annotated groups. The outer numbers indicate the five clans derived in this study. The right list counts the number of CYP families in each species in the tree. Bootstrap values >75% are shown in the tree.

To infer the similarity and evolutionary ancestry of pomegranate CYP families, we reconstructed an unrooted ML phylogenetic tree based on 83 putative PgCYP candidates (Figure 3). According to high bootstrap values (>75%), the pomegranate CYP superfamily was categorized into five clades. The CYP81 and CYP82 families consisted of the clade A, indicating these two families shared the most recent common ancestor (MRCA). The CYP71 and CYP736 families made up the clade C, and clans of A, B and C shared the MRCA. Thus we inferred that CYP71 and CYP82 shared the distinct MRCAs, although the CYP82 family belonged to the CYP71 clan[1, 7].

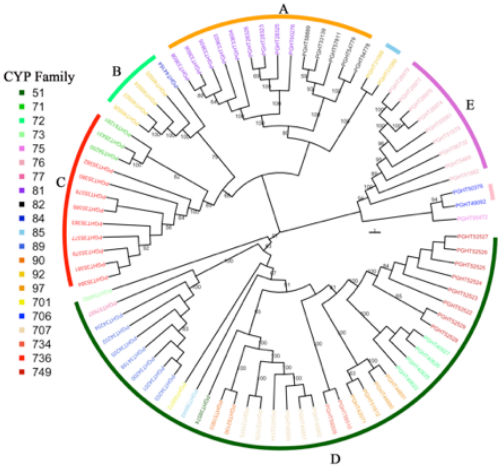


Figure 3 The phylogenetic tree of pomegranate CYP genes.

The inner circle is the maximum likelihood (ML) tree including 83 pomegranate CYP proteins. Tip-labels with twenty distinct colors show their CYP family groups, as indicated in the legend. Bootstrap values >50% are shown in the tree.

3.4 Selective Evolution Analysis of CYP Families

To identify the putative selective evolution events of pomegranate CYP families, we calculated Ka and Ks substitution rates for each gene influenced by nonsynonymous SNPs/InDels. All paralogous pairs in each CYP family showed a low ratio of Ka/Ks (<1) (Figure 4), indicating that some replacement substitutions inner CYP family have been purified by natural selection, presumably because of their deleterious effects[28]. CYP81 families clustered into two distinct groups based on a creditable bootstrap value of 100% (Figure 3), and the divergence time between groups traced back to 360-200 MYA ago (Figure 4). Additionally, paralogous CYP gene pairs belonging to two distinct families also showed a low ratio of Ka/Ks (Figure 5). Each two CYP families were divided into from their ancestor family before 100 MYA or early (Figure 5). Our findings suggested an understanding of selective analysis on CYP paralogs in pomegranate.

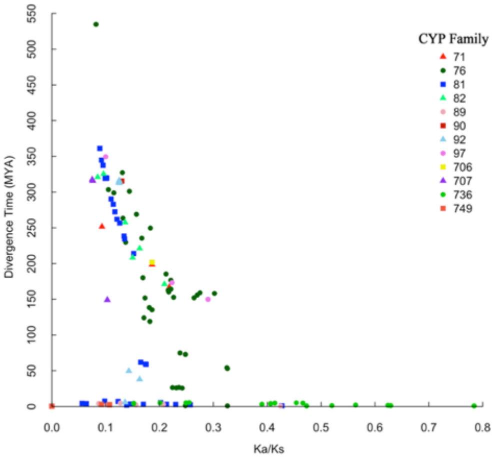


Figure 4 Divergence between paralogous CYP gene pairs belonging to the same CYP family in pomegranate.

Points represent paralogous PgCYP gene pairs. Different colors and shapes show their CYP family as indicated in the legend. The horizontal axis means non-synonymous substitutions per nonsynonymous site (Ka/Ks). Vertical axis presents the divergence time (MYA).

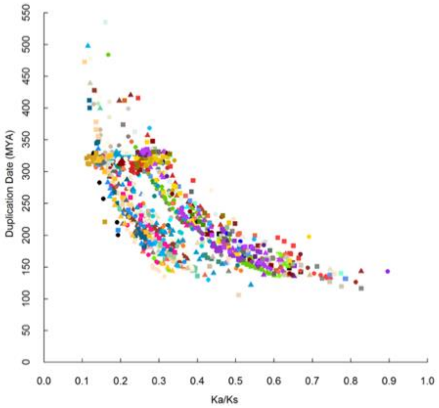


Figure 5 Divergence between paralogous CYP gene pairs belonging to different CYP families in pomegranate.

Points represent paralogous PgCYP gene pairs. The horizontal axis means Ka/Ks. Vertical axis presents the divergence time (MYA)

4. CONCLUSIONS

CYP family is an important superfamily and has large members in the plant. We *de nove* assembled high-quality pomegranate transcripts. Putative CYP candidate proteins in pomegranate and the other related species within the Malvids clade were obtained using our integrated HMM-search and InterProScan-verification as well as CYPED-annotation approach. This enzyme superfamily contributes diverse biological functions and thus is crucial for survival in the plant. Four major significant motifs in CYP proteins were identified for responding to their conserved structures and multi-functions. Eight test species within the Malvids clade shared distinct CYP paralogs from their common ancestor. Selective evolution analysis indicated that some replacement substitutions inner CYP family have been purified by natural selection. Our findings provided a resource for the taxonomic and evolutionary study on plant CYP families.

5. ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (31272143), the Initiative Project for Talents of Nanjing Forestry University (GXL2014070), the Priority Academic Program Development of Jiangsu High Education Institutions (PAPD), the Research Fund for Postgraduate Innovation Project of Jiangsu Province (KYLX16_0857), and the Doctorate Fellowship Foundation of Nanjing Forestry University.

6. REFERENCES

[1] Yan, Q., Cui, X., Lin, S., Gan, S., Xing, H. and Dou, D., 2016. *GmCYP82A3*, a Soybean Cytochrome P450 Family Gene Involved in the Jasmonic Acid and Ethylene Signaling Pathway, Enhances Plant Resistance to Biotic and Abiotic Stresses. *PLoS One* 11, 9 (Sep. 2016), e0162253. DOI= http://dx.doi.org/10.1371/journal.pone.0162253.

[2] Seitz, C., Ameres, S., Schlangen, K., Forkmann, G. and Halbwirth, H., 2015. Multiple evolution of flavonoid 3',5'-hydroxylase. *Planta* 242, 3 (Sep. 2015), 561-573. DOI= http://dx.doi.org/10.1007/s00425-015-2293-5.

[3] Almeida, D., Maldonado, E., Khan, I., Silva, L., Gilbert, M. T. P., Zhang, G., Jarvis, E. D., O'Brien, S. J., Johnson, W. E. and Antunes, A., 2016. Whole-Genome Identification, Phylogeny, and Evolution of the Cytochrome P450 Family 2 (CYP2) Subfamilies in Birds. *Genome Biol. Evol.* 8, 4 (Apr. 2016), 1115-1131. DOI= http://dx.doi.org/10.1093/gbe/evw041.

[4] Tanaka, Y., 2006. Flower colour and cytochromes P450. *Phytochem. Rev.* 5, 2 (Jun. 2006), 283-291. DOI= http://dx.doi.org/10.1007/s11101-006-9003-7.

[5] Chen, W., Lee, M.-K., Jefcoate, C., Kim, S.-C., Chen, F. and Yu, J.-H., 2014. Fungal Cytochrome P450 Monooxygenases: Their Distribution, Structure, Functions, Family Expansion, and Evolutionary Origin. *Genome Biol. Evol.* 6, 7 (Jul. 2014), 1620-1634. DOI= http://dx.doi.org/10.1093/gbe/evu132.

[6] Degtyarenko, K. N. and Archakov, A. I., 1993. Molecular evolution of P450 superfamily and P450-containing monooxygenase systems. *FEBS Lett.* 332, 1-2 (Oct.1993), 1-8. DOI= http://dx.doi.org/10.1016/0014-5793(93)80470-F.

[7] Nelson, D. R., Schuler, M. A., Paquette, S. M., Werck-Reichhart, D. and Bak, S., 2004. Comparative Genomics of Rice and Arabidopsis. Analysis of 727 *Cytochrome P450* Genes and Pseudogenes from a Monocot and a Dicot. *Plant Physiol.* 135, 2 (Jun. 2004), 756-772. DOI= http://dx.doi.org/10.1104/pp.104.039826.

[8] Prall, W., Hendy, O. and Thornton, L. E., 2016. Utility of a Phylogenetic Perspective in Structural Analysis of CYP72A Enzymes from Flowering Plants. *PLoS One* 11, 9 (Sep. 2016), e0163024. DOI= http://dx.doi.org/10.1371/journal.pone.0163024.

[9] Warren, R. L., Keeling, C. I., Yuen, M. M. S., Raymond, A., Taylor, G. A., Vandervalk, B. P., Mohamadi, H., Paulino, D., Chiu, R., Jackman, S. D., Robertson, G., Yang, C., Boyle, B., Hoffmann, M., Weigel, D., Nelson, D. R., Ritland, C., Isabel, N., Jaquish, B., Yanchuk, A., Bousquet, J., Jones, S. J. M., MacKay, J., Birol, I. and Bohlmann, J., 2015. Improved white spruce (*Picea glauca*) genome assemblies and annotation of large gene families of conifer terpenoid and phenolic defense metabolism. *The Plant Journal* 83, 2 (Jun. 2015), 189-212. DOI= http://dx.doi.org/10.1111/tpj.12886.

[10] Byng, J. W., Chase, M. W., Christenhusz, M. J. M., Fay, M. F., Judd, W. S., Mabberley, D. J., Sennikov, A. N., Soltis, D. E., Soltis, P. S., Stevens, P. F., Briggs, B., Brockington, S., Chautems, A., Clark, J. C., Conran, J., Haston, E., Mo€ller, M., Moore, M., Olmstead, R., Perret, M., Skog, L., Smith, J., Tank, D., Vorontsova, M. and Weber, A., 2016. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Bot. J. Linn. Soc.* 181, 1, 1-20. DOI= http://dx.doi.org/10.1111/boj.12385.

[11] Zhao, X., Yuan, Z., Feng, L. and Fang, Y., 2015. Cloning and expression of anthocyanin biosynthetic genes in red and white pomegranate. *J. Plant Res.* 128, 4 (Jul. 2015), 687-696. DOI= http://dx.doi.org/10.1007/s10265-015-0717-8.

[12] Teixeira da Silva, J. A., Rana, T. S., Narzary, D., Verma, N., Meshram, D. T. and Ranade, S. A., 2013. Pomegranate biology and biotechnology: A review. *Sci. Hortic.* 160(Aug. 2013), 85-107. DOI= http://dx.doi.org/http://dx.doi.org/10.1016/j.scienta.2013.05.017.

[13] Finn, R. D., Clements, J. and Eddy, S. R., 2011. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* 39, suppl_2 (May. 2011), W29-W37. DOI= http://dx.doi.org/10.1093/nar/gkr367.

[14] Mitchell, A., Chang, H.-Y., Daugherty, L., Fraser, M., Hunter, S., Lopez, R., McAnulla, C., McMenamin, C., Nuka, G., Pesseat, S., Sangrador-Vegas, A., Scheremetjew, M., Rato, C., Yong, S.-Y., Bateman, A., Punta, M., Attwood, T. K., Sigrist, C. J. A., Redaschi, N., Rivoire, C., Xenarios, I., Kahn, D., Guyot, D., Bork, P., Letunic, I., Gough, J., Oates, M., Haft, D., Huang, H., Natale, D. A., Wu, C. H., Orengo, C., Sillitoe, I., Mi, H., Thomas, P. D. and Finn, R. D., 2014. The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res.* 43, D1 (Nov. 2014), D213-D221. DOI= http://dx.doi.org/10.1093/nar/gku1243.

[15] Patel, R. K. and Jain, M., 2012. NGS QC Toolkit: A Toolkit for Quality Control of Next Generation Sequencing Data. *PLoS One* 7, 2 (Feb. 2012), e30619. DOI= http://dx.doi.org/10.1371/journal.pone.0030619.

[16] Zerbino, D. R. and Birney, E., 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome*

Res. 18, 5 (May. 2008), 821-829. DOI= http://dx.doi.org/10.1101/gr.074492.107.

[17] Schulz, M. H., Zerbino, D. R., Vingron, M. and Birney, E., 2012. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* 28, 8 (Feb. 2012), 1086-1092. DOI= http://dx.doi.org/10.1093/bioinformatics/bts094.

[18] Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J., 1990. Basic local alignment search tool. *J. Mol. Biol.* 215, 3 (Oct. 1990), 403-410. DOI= http://dx.doi.org/http://dx.doi.org/10.1016/S0022-2836(05)80360-2.

[19] Katoh, K. and Standley, D. M., 2016. A simple method to control over-alignment in the MAFFT multiple sequence alignment program. *Bioinformatics* 32, 13 (Feb. 2016), 1933-1942. DOI= http://dx.doi.org/10.1093/bioinformatics/btw108.

[20] Kumar, S., Stecher, G. and Tamura, K., 2016. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol. Biol. Evol.* 33, 7 (Jul. 2016), 1870-1874. DOI= http://dx.doi.org/10.1093/molbev/msw054.

[21] Yu, G., Smith, D. K., Zhu, H., Guan, Y. and Lam, T. T.-Y., 2016. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.*(Sep. 2016). DOI= http://dx.doi.org/10.1111/2041-210X.12628.

[22] Yang, Z., 2007. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 8 (Aug. 2007), 1586-1591. DOI= http://dx.doi.org/10.1093/molbev/msm088.

[23] Feng, L., Chen, Z., Ma, H., Chen, X., Li, Y., Wang, Y. and Xiang, Y., 2014. The IQD Gene Family in Soybean: Structure, Phylogeny, Evolution and Expression. *PLoS One* 9, 10 (Oct. 2014), e110896. DOI= http://dx.doi.org/10.1371/journal.pone.0110896.

[24] Ophir, R., Sherman, A., Rubinstein, M., Eshed, R., Sharabi Schwager, M., Harel-Beja, R., Bar-Ya'akov, I. and Holland, D., 2014. Single-Nucleotide Polymorphism Markers from De-Novo Assembly of the Pomegranate Transcriptome Reveal Germplasm Genetic Diversity. *PLoS One* 9, 2 (Feb. 2014), e88998. DOI= http://dx.doi.org/10.1371/journal.pone.0088998.

[25] Wang, X., Wang, H., Wang, J., Sun, R., Wu, J., Liu, S., Bai, Y., Mun, J.-H., Bancroft, I., Cheng, F., Huang, S., Li, X., Hua, W., Wang, J., Wang, X., Freeling, M., Pires, J. C., Paterson, A. H., Chalhoub, B., Wang, B., Hayward, A., Sharpe, A. G., Park, B.-S., Weisshaar, B., Liu, B., Li, B., Liu, B., Tong, C., Song, C., Duran, C., Peng, C., Geng, C., Koh, C., Lin, C., Edwards, D., Mu, D., Shen, D., Soumpourou, E., Li, F., Fraser, F., Conant, G., Lassalle, G., King, G. J., Bonnema, G., Tang, H., Wang, H., Belcram, H., Zhou, H., Hirakawa, H., Abe, H., Guo, H., Wang, H., Jin, H., Parkin, I. A. P., Batley, J., Kim, J.-S., Just, J., Li, J., Xu, J., Deng, J., Kim, J. A., Li, J., Yu, J., Meng, J., Wang, J., Min, J., Poulain, J., Wang, J., Hatakeyama, K., Wu, K., Wang, L., Fang, L., Trick, M., Links, M. G., Zhao, M., Jin, M., Ramchiary, N., Drou, N., Berkman, P. J., Cai, Q., Huang, Q., Li, R., Tabata, S., Cheng, S., Zhang, S., Zhang, S., Huang, S., Sato, S., Sun, S., Kwon, S.-J., Choi, S.-R., Lee, T.-H., Fan, W., Zhao, X., Tan, X., Xu, X., Wang, Y., Qiu, Y., Yin, Y., Li, Y., Du, Y., Liao, Y., Lim, Y., Narusaka, Y., Wang, Y., Wang, Z., Li, Z., Wang, Z., Xiong, Z. and Zhang, Z., 2011. The genome of the mesopolyploid crop species *Brassica rapa*. *Nat. Genet.* 43, 10 (Aug. 2011), 1035-1039. DOI= http://dx.doi.org/10.1038/ng.919.

[26] Myburg, A. A., Grattapaglia, D., Tuskan, G. A., Hellsten, U., Hayes, R. D., Grimwood, J., Jenkins, J., Lindquist, E., Tice, H., Bauer, D., Goodstein, D. M., Dubchak, I., Poliakov, A., Mizrachi, E., Kullan, A. R. K., Hussey, S. G., Pinard, D., van der Merwe, K., Singh, P., van Jaarsveld, I., Silva-Junior, O. B., Togawa, R. C., Pappas, M. R., Faria, D. A., Sansaloni, C. P., Petroli, C. D., Yang, X., Ranjan, P., Tschaplinski, T. J., Ye, C.-Y., Li, T., Sterck, L., Vanneste, K., Murat, F., Soler, M., Clemente, H. S., Saidi, N., Cassan-Wang, H., Dunand, C., Hefer, C. A., Bornberg-Bauer, E., Kersting, A. R., Vining, K., Amarasinghe, V., Ranik, M., Naithani, S., Elser, J., Boyd, A. E., Liston, A., Spatafora, J. W., Dharmwardhana, P., Raja, R., Sullivan, C., Romanel, E., Alves-Ferreira, M., Kulheim, C., Foley, W., Carocha, V., Paiva, J., Kudrna, D., Brommonschenkel, S. H., Pasquali, G., Byrne, M., Rigault, P., Tibbits, J., Spokevicius, A., Jones, R. C., Steane, D. A., Vaillancourt, R. E., Potts, B. M., Joubert, F., Barry, K., Pappas, G. J., Strauss, S. H., Jaiswal, P., Grima-Pettenati, J., Salse, J., Van de Peer, Y., Rokhsar, D. S. and Schmutz, J., 2014. The genome of *Eucalyptus grandis*. *Nature* 510, 7505 (Jun. 2014), 356-362. DOI= http://dx.doi.org/10.1038/nature13308.

[27] Sankoff, D., Zheng, C. and Zhu, Q., 2010. The collapse of gene complement following whole genome duplication. *BMC Genomics* 11, 1 (May. 2010), 313. DOI= http://dx.doi.org/10.1186/1471-2164-11-313.

[28] Sillo, F., Garbelotto, M., Friedman, M. and Gonthier, P., 2015. Comparative Genomics of Sibling Fungal Pathogenic Taxa Identifies Adaptive Evolution without Divergence in Pathogenicity Genes or Genomic Structure. *Genome Biol. Evol.* 7, 12 (Dec. 2015), 3190-3206. DOI= http://dx.doi.org/10.1093/gbe/evv209.